Modélisation statistique de la mortalité maternelle et néonatale pour l'aide à la planification et à la gestion des services de santé en Afrique Sub-Saharienne

Cheikh Ndour

THÈSE

Pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR & DE L'UNIVERSITÉ GASTON BERGER DE SAINT-LOUIS

Diplôme National - Arrêté du 7 août 2006

DOMAINE DE RECHERCHE : Mathématiques Appliquées - Option Statistique

Présentée par

Cheikh Ndour

Modélisation statistique de la mortalité maternelle et néonatale pour l'aide à la planification et à la gestion des services de santé en Afrique Sub-Saharienne

Directeurs de thèse : **Mme Sophie Mercier, M. Aliou Diop**

Co-direction : **M. Simplice Dossou Gbété, M. Alexandre Dumont**

Soutenue le 19 Mai 2014

Devant la Commission d'Examen

JURY

Philippe BESSE	Professeur des Universités, INSA TOULOUSE	Rapporteur
Gilbert SAPORTA	Professeur des Universités, CNAM de Paris	Rapporteur
Richard D. DE VEAUX	Professeur, WILLIAMS COLLEGE	Rapporteur
Ivan KOJADINOVIC	Professeur des Universités, U.P.P.A	Examinateur
Ahmadou ALIOUM	Professeur des Universités, UNIVERSTE BORDEAUX 2	Examinateur
Simplice DOSSOU-GBETE	Maître de Conférences, U.P.P.A	Codirecteur
Sophie MERCIER	Professeur des Universités, U.P.P.A	Directrice
Aliou DIOP	Professeur, UNIVERSITE GASTON BERGER	Directeur

Thèse préparée au sein du Laboratoire de Mathématiques et de leurs Applications - Pau (L.M.A.P), dans l'école doctorale des sciences exactes et leurs Applications (ED211) de l'UPPA (France) en collaboration avec le Laboratoire d'Etudes et de Recherches en Statistiques et Développement (LERSTAD) de l'Université Gaston Berger de Saint-Louis (Sénégal).

Remerciement

Juste quelques lignes pour exprimer toute ma gratitude envers les personnes qui ont contribué à la réalisation de cette thèse.

Je tiens à remercier les membres de mon jury de thèse pour l'intérêt qu'ils ont porté à mon travail. J'exprime ma reconnaissance à Ivan KOJADINOVIC de m'avoir fait l'honneur d'examiner mon travail et de présider mon jury, ainsi qu'à Ahmadou ALIOUM pour avoir examiner mes travaux. J'adresse mes plus sincères remerciements à mes rapporteurs, Gilbert SAPORTA, Phillipe BESSE et Richard D. DE VEAUX, pour l'honneur qu'ils m'ont fait en acceptant de rapporter mes travaux. Je tiens à les remercier pour l'intérêt qu'ils ont porté à mon manuscrit, pour leurs lectures minutieuses, leurs remarques et leurs suggestions.

Mes premiers remerciements vont naturellement à la personne à qui cette thèse doit le plus, et envers qui je suis profondément reconnaissant. Je veux nommer Simplice DOSSOU-GBETE. C'est une chance et une grande fierté pour moi d'avoir travaillé avec lui pendant ces trois dernières années. Je le remercie de m'avoir fait bénéficier de l'étendue de ses connaissances et de son savoir faire, de son encadrement au quotidien, toujours teinté de bonne humeur.
Mes chaleureux remerciements vont à Sophie MERCIER d'avoir accepté de diriger ma thèse et de son soutien sans faille pour la réalisation de cette thèse. Mes remerciements vont également à Noëlle BRU. J'adresse mes vives remerciements à Aliou DIOP, car c'est en partie lui qui m'a guidé vers les Proba-Stat. Je tiens à le remercier sincèrement pour la confiance qu'il m'a accordé en me donnant la chance de mener cette thèse, pour son soutien sans limite et pour sa disponibilité. A travers sa personne, je tiens à remercier l'ensemble du corps enseignant et personnel administratif de l'UFR Sciences Appliquées et Technologie de l'université Gaston Berger.

Cette thèse a été motivée par des données réelles collectées au cours de la phase pré-intervention du projet QUARITE (qualité des soins, gestion du risque et de techniques obstétricales). C'est l'oc-

casion pour moi de remercier chaleureusement Alexandre DUMONT pour la confiance qu'il a accordé à ma petite personne, pour avoir guidé mes premiers pas dans le monde de la recherche et pour son soutien. A travers lui je voudrais remercier l'Institut de Recherche pour le Développement (I.R.D) en particulier tous les membres de UMR 216 de l'I.R.D. Mes remerciements vont également à la Service de Coopération et d'Actions Culturelles (S.C.A.C) de l'Ambassade de France au Sénégal qui a accepté de financer mes trois premiers séjours d'étude à l'Université de Pau et des Pays de l'Adour (UPPA). Je remercie aussi le Laboratoire de Mathématiques et de leurs Applications de Pau (L.M.A.P) pour le financement de mon dernier séjour d'étude à l'UPPA qui m'a permis de finaliser mes travaux. Je remercie également le Laboratoire d'Etudes et de Recherches en Statistiques et Développement (LERSTAD) de l'Université Gaston Berger qui a pris en charge les frais de transport de mon dernier séjour à Pau.

Toute ma profonde gratitude à tous les docteurs et doctorants du LMAP avec qui j'ai partagé de très bon moments, à Johng-ay TAMATORO avec qui j'ai partagé mes galères mais aussi des soirées que je ne suis pas prêt à oublier. Quelques mots de remerciement à tout le personnel du LMAP, particulièrement à Marie-Claire Hummel, Sylvie Berton, Marie-Laure Rius, Chantal Blanchard, Bruno Demoisy et Lina Goncalves, grâce à qui mes séjours et mes déplacements ont été sans difficultés. Je n'oublie pas le personnel de l'école doctorale des sciences en particulier Jacqueline Petitbon, Annemarie Venancio, Olivier Autexier, Stéphan Leborgne et Marc Odunlami.
Je réserve une mention spéciale aux doctorants du Laboratoire d'Etudes et de Recherches en Statistiques et Développement (LERSTAD) de l'Université Gaston Berger (Sénégal). Il m'est impossible d'oublier leur bonne humeur et les 30 mm de pause que nous passions à la cafétéria et surtout mes deux grands Ibou FAYE et Lucien GNING. Mes pensées les plus chères vont à mes ami(e)s et mes proches qui ont cru en moi, qui m'ont soutenu moralement et avec qui j'ai partagé mes angoisses et mes rêves.

C'est un merci du fond du cœur que je lance à ma famille, à mes chers frères Mamadou, Babou et Ngor, à mes aimables sœurs Ndella, Diouldé et Khady et ma chère bien-aimée Aminata DIOP. Je réserve le plus grand des mercis à mes parents en particulier à ma très chère mère surtout pour sa patience. Ils ont été, sont et resteront les piliers qui proposent un appui sans limite.

Merci encore à tous et aux courageux qui veulent continuer, je souhaite une très bonne lecture!!

Cheikh Ndour.

Table des matières

Table des figures

Liste des tableaux

Chapitre I

Introduction Générale

1 Motivation

En Afrique Sud du Sahara (ASS), la mortalité maternelle (MM) est parmi les plus élevées au monde. La réduction du taux de MM de trois quart entre 1990 et 2015, constitue le cinquième Objectif du Millénaire pour le Développement (OMD5) ; malheureusement, les progrès sont lents et l'atteinte des objectifs fixés est très hypothétique [20]. Depuis 1990, certains pays en Asie et en Afrique du Nord ont fait baisser de plus de moitié la mortalité maternelle. Il y a eu aussi des progrès en Afrique subsaharienne (5%). Mais, sur ce continent et contrairement aux pays développés où le risque à la naissance pour une femme de mourir pendant une grossesse ou peu de temps après est de 1 sur 3800, le risque de mortalité maternelle reste très élevé à 1 sur 39[1]. A l'échelle mondiale, pour la période 1995-1998, on a enregistré 430 décès maternels pour 100 000 naissances vivantes. En Afrique subsaharienne, le taux de décès maternel est estimé à 975 pour 100 000 naissances vivantes contre 13 pour les pays industrialisés (WHO, 2000).

Des études dans différents pays d'Afrique subsaharienne ont identifié plusieurs facteurs de risque indépendants qui diffèrent sensiblement entre les auteurs, probablement en raison des différences entre les populations d'étude, l'environnement, les variables recueillies et les méthodes statistiques utilisées. Ainsi, il reste difficile de fournir aux professionnels de la santé des pays d'Afrique subsaharienne des recommandations pour identifier les signes ou symptômes cliniques qui pourraient aider le personnel à détecter les patients à haut risque de décès à l'hôpital. Pourtant, ces critères pourront aider le personnel à décider si un patient doit être traité comme un cas de haute priorité par les professionnels de santé qualifiés dans les soins obstétricaux d'urgence complets [3]. Bien que les critères de complications obstétricales graves sont proposés par l'Organisation mondiale de la Santé (OMS) comme facteurs prédictifs appropriés de mortalité maternelle [21], des difficultés subsistent dans leur identification, et il y a peu d'expérience avec l'utilisation de ces critères dans les pays à faible revenu [19].

Du point de l'apprentissage statistique supervisé, tout ensemble de données où la distribution a priori de la variable réponse est significativement différente de la distribution uniforme est considéré comme un jeu de données déséquilibrées. Cependant, la compréhension commune de la communauté

statistique est que les données déséquilibrées correspondent à des ensembles de données présentant un déséquilibre significatif, et dans certains cas un déséquilibre extrême. Plus précisément, cette forme de déséquilibre est considérée comme un déséquilibre entre les classes de la variable d'intérêt. On rencontre très souvent des déséquilibres d'ordre 1/100, 1/1 000 et 1/10 000 entre les classes, où dans chaque cas, une classe domine sévèrement une autre. De ce point de vue, on peut aborder l'analyse de la mortalité maternelle sous l'angle de données déséquilibrées. On dit que le décès maternel, considéré ici comme étant l'événement d'intérêt, est un événement rare par rapport à l'événement non-décès.

2 Classification supervisée

La classification supervisée consiste à classer de nouveaux objets en se basant sur l'observation d'exemples similaires. Elle est l'une des tâches typiques du domaine du data mining. Ici chaque objet est décrit par un couple (X, Y) où X est une vecteur de p variables aléatoires pouvant être numériques, discrètes ou catégorielles. La variable X prend ses valeurs dans un domaine $\mathcal{X}$ produit de p domaines numériques, discrètes et catégoriels. La variable réponse Y prend ses valeurs dans le domaine catégoriel $\mathcal{Y} = \{y_1, \ldots, y_s\}$.
Lorsqu'on traite un problème de classification supervisée, on considère la réalisation $\{(x_i, y_i), i = 1 : n\}$ d'un échantillon $T_n = (X_1, Y_1), \cdots, (X_n, Y_n)$ pour construire une règle $\mathbf{R}_{T_n} : \mathcal{X} \longrightarrow \mathcal{Y}$ qui permet une prédiction future de la variable réponse Y, en se basant sur l'observation de X seulement.

En général on considère que T_n est une suite d'éléments aléatoires indépendantes et identiquement distribuées suivant une loi F inconnue définie sur $\mathcal{X} \times \mathcal{Y}$. De plus la règle de classement qui réalise le minimum d'erreur de classement est la règle de Bayes définie par

$$\mathbf{R}_{T_n}(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \Pr(Y = y | X = x) \qquad \forall x \in \mathcal{X}$$

Ce qui correspond à

$$\mathbf{R}_{T_n}(x) = y \text{ si } \frac{\Pr(Y = y | X = x)}{\Pr(Y = y' | X = x)} > 1 \qquad \forall y \neq y' \tag{I.1}$$

où $\Pr(Y = y | X = x)$ est la probabilité conditionnelle estimée d'appartenance à y.
Parmi les méthodes statistiques de classement on peut distinguer des approches non paramétriques comme la méthode des plus proches voisins et les arbres binaires de classement et des approches paramétriques comme l'analyse discriminante, la régression logistique et les réseaux de neurones. Ces approches sont basées sur une évaluation implicite ou explicite de la distribution conditionnelle $\Pr(Y = y | X = x)$. Par exemple, dans une analyse discriminante, on suppose une loi a priori $[Y]$ sur la variable Y. Puis on s'intéresse donc à loi de probabilité conditionnelle $\Pr(Y|X)$ définie par :

- $\Pr(Y = y | X = x) = \dfrac{\Pr(X = x | Y = y) \Pr(Y = y)}{\sum_{i=1}^{s} \Pr(X = x | Y = y_i) \Pr(Y = y_i)}$ si X est une variable discrète

– $\Pr(Y=y|X=x) = \dfrac{f(x|Y=y)\Pr(Y=y)}{\sum_{i=1}^{s} f(x|Y=y_i)\Pr(Y=y_i)}$ si la loi de X conditionnellement à $Y=y$ admet une densité

L'analyse est basée sur la possibilité d'estimer la loi de probabilité conditionnelle $\Pr(X|Y)$ à partir des données. Tandis que les arbres de classement affectent les objets dans les différentes classes en fonction de l'estimation non paramétrique de la loi de probabilité conditionnelle $\Pr(Y|X)$.

A partir de la relation (I.1), on considère la famille des règles de classement indexée par $t \in]0,+\infty[$ et définie par

$$\mathbf{R}^t_{T_n}(x) = y \text{ si } \frac{\Pr(Y=y|X=x)}{\Pr(Y=y'|X=x)} > t \qquad \forall y \neq y' \tag{I.2}$$

La relation

$$x_1 \mathcal{R} x_2 \Leftrightarrow \frac{\Pr(Y=y|X=x_1)}{\Pr(Y=y'|X=x_1)} > t \text{ et } \frac{\Pr(Y=y|X=x_2)}{\Pr(Y=y'|X=x_2)} > t \qquad (x_1,x_2) \in \mathcal{X}\times\mathcal{X}$$

est une relation d'équivalence. Et donc $\mathbf{R}^t_{T_n}(x)$ est constante sur la classe $\dot{\hat{x}}$. Il en résulte que la règle $\mathbf{R}^t_{T_n}$ produit une partition de $\mathcal{X}$ en sous-ensembles dont chacune d'eux est associée à une classe de Y.

Dans ce travail, nous abordons le problème de la classification supervisée lorsque la variable réponse est binaire et que la distribution a priori de ses classes est déséquilibrée. On rencontre cette situation dans plusieurs domaines tels que la finance (identification de transactions de cartes de crédit frauduleuses ou demande de crédits défaillants), l'épidémiologie (diagnostic de cellules cancéreuses par la radiographie ou toute maladie rare), les sciences sociales (détection de comportement anormal), l'informatique (reconnaissance de la forme dans des données d'image ou catégorisation de textes), la bio-statistique (affectation d'un objet à sa famille d'appartenance). Ce problème n'est pas nouveau dans le domaine du data mining. Il a été rapporté plusieurs fois dans la littérature que la distribution déséquilibrée des classes de la variable réponse affaiblit lourdement le processus d'apprentissage, puisque le classifieur tend à se focaliser sur la classe prévalente en ignorant la classe rare.

3 L'état de l'art

Le problème de la classification supervisée dans une situation où l'événement d'intérêt est considéré comme un événement a priori rare n'est pas un problème nouveau dans le domaine du data mining. Dans un passé récent, plusieurs méthodes d'apprentissage d'un classifieur sur des données déséquilibrées ont été proposées dans la littérature [8]. La situation à ce jour semble fournir des méthodes multiples, chacune d'entre elles améliorant les méthodes existantes en ce qui concerne certains aspects, mais présentant des limites par rapport à d'autres aspects. Dans de nombreux cas, on ne sait pas clairement pourquoi une technique doit être préférée aux autres, et seules des raisons heuristiques

sont données pour justifier les propositions suggérées.

3.1 Quelques méthodes de classement standard et leurs limites

Dans une telle situation, le but de l'analyse est de produire un classifieur qui offrira une grande précision pour la classe minoritaire sans pour autant compromettre gravement l'exactitude de la classe majoritaire. Lorsqu'un algorithme d'apprentissage standard, paramétrique ou non-paramétrique, est appliqué aux données déséquilibrées, les règles d'induction qui décrivent la classe minoritaire sont souvent rares et plus faibles que celles de la classe majoritaire, puisque la classe minoritaire est souvent à la fois en infériorité numérique et sous-représentée.

– La régression logistique, traditionnellement connue comme étant l'une des méthodes paramétriques les plus usuelles pour une classification supervisée binaire, a pour objectif de modéliser le paramètre de la distribution de la variable réponse Y qui, pour une unité i donnée, prend la valeur 1 avec une probabilité π et 0 avec la probabilité $1-\pi$. Il est supposé que

$$logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) = X^t\beta$$

où X est un vecteur de variables aléatoires et β un vecteur de paramètres.
Le classement de nouvelles unités pourrait être obtenu par l'estimation de π par

$$\widehat{\pi} = \frac{e^{X^t\widehat{\beta}}}{1+e^{X^t\widehat{\beta}}}$$

où $\widehat{\beta}$ est l'estimation du paramètre β.
On classe en $Y=1$, les unités qui ont estimé π supérieur à un seuil (0.5 par défaut). Et lorsque nous sommes dans une situation où la probabilité de la classe d'intérêt ($Y=1$) de la variable réponse tend vers zéro, alors le paramètre π est sous-estimé.
La régression logistique est inefficace lorsqu'il s'agit de traiter des données déséquilibrées car la probabilité conditionnelle de la classe rare est sous-estimée [11].

– Le but de l'analyse discriminante linéaire consiste à chercher $\text{argmax}_{j\in\mathcal{D}om(Y)} f(x|y=j)\Pr(Y=j)$, où $f(x|y=j)$ est la densité d'une loi gaussienne de moyenne μ_j pour le groupe j et de matrice de covariance $\Sigma=\Sigma_1=\Sigma_0$. Lorsque les paramètres de la distribution sont connus, la fonction de discrimination déduite de la règle de décision de Bayes est donnée par

$$g_j(X) = -\frac{1}{2}(X-\mu_j)^t\Sigma^{-1}(X-\mu_j) - \frac{1}{2}ln|\Sigma| - \frac{d}{2}ln2\pi + ln\Pr(y=j)$$

où d est la dimension de $\mathcal{D}om(X)$.
Pour estimer μ_j et Σ_j, on utilise habituellement la moyenne empirique $\widehat{\mu}_j$ et la matrice de

covariance de l'échantillon $\widehat{\Sigma}_j$. La matrice de covariance de l'échantillon est donnée par :

$$\widehat{\Sigma}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{j_i} - \widehat{\mu}_j)(X_{j_i} - \widehat{\mu}_j)^t \qquad j = 0, 1$$

où n_j est le nombre d'observations de la classe $y = j$.
En effet Xie et Qiu ont montré dans [22] que l'ensemble d'apprentissage déséquilibré n'a pas d'effets sur la matrice de projection si les deux matrices de covariance de l'échantillon sont identiques. Mais si les deux matrices de covariance d'échantillonnage sont différentes, l'énorme déséquilibre dans la répartition des classes est très problématique pour l'analyse discriminante linéaire parce que la probabilité a priori de la classe majoritaire éclipse les différences dans les termes de la matrice de covariance d'échantillonnage. Toutefois, l'hypothèse de l'égalité des matrices de covariance d'échantillonnage est limitée à des cas particuliers dans des scénarios de la vie réelle. Par conséquent, nous devons considérer l'effet de l'ensemble d'apprentissage déséquilibré sur la performance de l'analyse discriminante linéaire dans la pratique. Et par conséquent on peut dire que la distribution a priori déséquilibrée de la variable Y nuit à la performance de l'analyse discriminante linéaire.

– Le classifieur bayésien naïf donné par :

$$h(X) = \text{argmax}_{y \in \{0,1\}} \Pr(Y = y) \prod_{i=1}^{d} \Pr(X_i | Y = y)$$

est fortement dominé par la classe $y = 0$ puisque l'estimation empirique de la probabilité $\Pr(y = 1)$ à partir de l'ensemble d'apprentissage $\mathcal{D}_n$ est très faible. Donc l'utilisation du classifieur bayésien naïf n'est pas envigeasable puisqu'il produit une sous-estimation explicite de la probabilité conditionnelle $\Pr(Y = 1 | X = x)$.

– L'objectif des arbres de décision est de prédire la valeur d'une variable qualitative en fonction d'un ensemble de variables explicatives de nature quelconque. L'algorithme détermine la règle de classement en deux temps : (1) On commence par partitionner les données selon les modalités de l'attribut le plus discriminant, puis on répète l'opération localement sur chaque nœud ainsi obtenu jusqu'à la réalisation d'un critère d'arrêt. (2) On dérive la règle de classement en choisissant dans chaque nœud la modalité majoritaire de la variable à prédire, en général simplement la plus probable, dans chaque feuille (nœud terminal) de l'arbre. Le principal problème de cette procédure en présence de données déséquilibrées est que le partitionnement successif de l'espace des données résulte sur l'observation de moins en moins d'exemples de la classe rare occasionnant moins de feuilles décrivant la classe minoritaire et successivement des estimations plus faibles de la confiance. Ils fournissent ainsi une sous-estimation implicite de la probabilité

conditionnelle $\Pr(Y = 1|X = x)$ via la distribution des classes au niveau des feuilles terminales [4]. Les arbres de décision ne sont donc pas appropriés pour construire une règle de classement sur des données déséquilibrées.

– Les réseaux de neurones ne sont pas non plus adaptables puisqu'ils produisent une estimation de la distribution a posteriori $\Pr(Y|X = x)$ qui est fortement dominée par la classe $y = 0$.

Le principe de classement des différentes méthodes énumérées ci-dessus consiste à calculer un score prédictif pour chaque nouvelle observation puis comparer ce score avec un seuil t fixé a priori. Ce pendant les scores calculés à partir de données déséquilibrées sont très proches de zéro. Par conséquent ils dépassent rarement le seuil t fixé.

3.2 Quelques solutions proposées pour la prise en charge des données déséquilibrées

Plusieurs travaux ont été consacrés au problème de classement pour données déséquilibrées et même dans un passé récent, que ce soit du point de vue statistique conventionnelle en tant que telle ainsi que de l'apprentissage automatique. Certaines œuvres parmi eux envisageront l'amélioration de l'ajustement des modèles de régression pour produire une fonction de classification avec un faible biais de prédiction sans perdre des fonctionnalités intéressantes des méthodes classiques comme la capacité à évaluer la contribution de chaque variable dans les variations de la probabilité de la classe cible (méthodes de régression) ou de l'identification du motif de risque (arbre de décision). Dans l'ensemble, les méthodes visant à s'attaquer au problème de classification sur données dont la distribution de la variable réponse est déséquilibrée peuvent être divisées en deux grandes catégories : les méthodes préconisant un prétraitement des données et les méthodes intervenant au niveau du processus d'apprentissage.

– Le prétraitement des données proposé par certaines méthodes pour traiter des données déséquilibrées consiste à simuler un ensemble d'apprentissage non déséquilibré conditionnellement aux données observées. Les techniques de simulation (ré-échantillonnage) proposées dans la littérature sont nombreuses et variées. On peut citer la méthode du sur-échantillonnage avec remplacement qui consiste à dupliquer les observations de la classe rare et le sous-échantillonnage sans remplacement qui consiste à supprimer des observations de la classe dominante. La plupart des méthodes actuelles sont basées sur ces deux techniques d'échantillonnage. Elles permettent de réduire le degré de déséquilibre de l'échantillon d'apprentissage et par conséquent améliorer la précision globale du classifieur. Cependant le sous-échantillonnage peut conduire à supprimer des données capitales pour la construction du classifieur. De même le sur-échantillonnage augmente la vraisemblance du modèle ajusté puisqu'il crée des doublons dans l'échantillon d'apprentissage. Il faut noter aussi que le classifieur obtenu à partir de ces deux techniques est fortement dé-

pendant de l'ensemble d'apprentissage. Pour parer à cette éventualité, de nouvelles stratégies de sélection de nouvelles observations ont été proposées dans un passé récent. Pour plus de détails on peut consulter les travaux de Lee (1999,2000)[12, 13], les travaux de Chawal et al. (2002) qui ont proposé la méthode SMOTE (Synthetic Minority Oversampling Technique) ou bien les travaux de Menardi et al. (2012) qui ont proposé l'algorithme ROSE (Random Over-Sampling Exemples)[17].

– Les solutions préconisant un algorithme d'apprentissage sont nombreuses et variées. Parmi les plus utilisées figure celle qui consiste à modifier le processus d'apprentissage en tenant compte des coûts de mauvais classements différents. Cette approche permet de donner plus de poids aux observations de la classe rare. Cette approche est utilisée lorsque la distribution déséquilibrée des classes est associée à des coûts de mauvais classement. Dans ce cas, une règle de classification minimisant le coût de mauvais classement moyen est établie. Certains classifieurs tels que les réseaux de neurones, les méthodes de régression, etc., produisent un score représentant le degrés d'appartenance d'une observation du domaine des covariables à une classe de la variable réponse. La règle de classement est définie par la spécification d'un seuil λ [6]. On peut faire varier le seuil λ de manière à ce que la règle de classification soit sensible par rapport à la classe faiblement représentée. D'autres approches consistent à des techniques d'agrégation comme bagging, boosting ou forêts aléatoires (random forest), qui combinent plusieurs fonctions de classification avec un grand taux d'erreur individuel pour produire une nouvelle fonction de classification avec un plus petit taux d'erreur [2]. A ces dernières, on peut ajouter les méthodes consistant à associer le ré-échantillonnage de l'ensemble d'apprentissage avec la combinaison des classifieurs [9].

Ces différents processus d'apprentissage, bien qu'ils aient la faculté d'améliorer les performances des classifieurs en présence de la distribution déséquilibrée des classes de la variable réponse, ont le désavantage d'être lourds, et en plus le classifieur obtenu est sous forme d'une boîte noire. Il est difficile (impossible) d'identifier les profils qui ont contribué à la construction du classifieur.

4 Classification supervisée et règles d'association

Notre objectif est de proposer une méthode d'apprentissage statistique qui fournit un classifieur efficace et permettant d'identifier les profils pertinents corrélés avec la classe cible de la variable réponse. Pour atteindre cet objectif, nous nous sommes tournés vers l'apprentissage des règles d'association qui est une méthode bien connue dans le domaine du data mining. Il est utilisé pour le traitement de grandes bases de données pour la découverte non supervisée de modèles locaux qui expriment des relations précieuses cachées et potentielles entre les variables d'entrée. En examinant les règles d'association d'un point de vue de l'apprentissage statistique supervisé, un ensemble pertinent de classifieurs faibles est obtenu à partir duquel on tire une règle de classification qui fonctionne bien. Une telle approche n'est pas réellement nouvelle puisqu'elle a déjà été prise en compte dans la littérature de

l'apprentissage automatique [14].

Des études récentes dans le domaine du data mining ont proposé une nouvelle approche de classement appelé "classement associatif" qui a montré des taux d'erreur plus faibles que les algorithmes traditionnels tels que les arbres de décision. Cependant, parce que le nombre de règles d'association possibles en général est très grand, les algorithmes sont complexes et sujettes à un sur-ajustement.
Lorsqu'il s'agit de traiter un problème de classification supervisée, on se focalise sur un sous-ensemble particulier de règles d'association communément appelé "Class Association Rules" (CAR). Quand on utilise les CARs pour classer un nouvel objet (i.e un objet ou individu qui n'a pas participé à la construction du sous-ensemble), il arrive que plus d'une règle soit éligible. C'est pour cette raison qu'une relation d'ordre est définie dans l'ensemble des CARs. Parmi les algorithmes de classement associatif, l'algorithme **CBA** : ("Classification Based on Associations")[16], l'algorithme **CMAR** :("Classification based on Multiple Association Rules")[15] et l'algorithme **CPAR** :("Classification base on Predictive Association Rules")[23] sont les plus utilisés dans la littérature.
L'algorithme CBA génère premièrement un ensemble de règles d'association candidates à l'aide d'un seuil de support minimum et d'un seuil de confiance minimum. Ensuite il définit la relation d'ordre suivante sur l'ensemble des règles candidates. La règle r_i précède la règle r_j si

- r_i a une confiance plus élevée que celle de r_j ; ou bien
- si leurs confiances sont égales, r_i a un support plus élevé que celui de r_j ; ou bien
- si leurs confiances et leurs supports sont égaux deux à deux, r_i est généré avant r_j.

Pour prédire la classe d'un nouvel objet, la première règle vérifiée par l'objet est choisie pour la prédiction.
L'algorithme CMAR est similaire à l'algorithme CBA par la méthode de générer l'ensemble des règles candidates mais aussi par la relation d'ordre établie sur ce dernier. Leur différence majeure se situe au niveau de la procédure d'élagage et le principe de classement d'un nouvel objet. Au niveau de la procédure d'élagage, l'algorithme CMAR utilise une structure d'arbre plus efficace [7] et un test du chi2 (χ^2) pour élaguer les règles redondantes et les informations bruyantes. Au niveau du principe de classement, CMAR sélectionne le sous-ensemble de règles vérifiées par le nouvel objet.

- si toutes les règles du sous-ensemble ont la même classe, l'objet est affecté à cette classe.
- sinon, on divise les règles en groupe selon la classe correspondante et on affecte l'objet à la classe la plus représentée [15].

L'algorithme CPAR combine les avantages du classement associatif et des algorithmes précédents. Au lieu de générer les règles de la même façon que les deux algorithmes précédents, CPAR adopte un algorithme plus général (FOIL) [18] pour générer des règles à partir des données d'apprentissage. En outre, CPAR génère et teste plus de règles que les algorithmes CBA et CMAR pour éviter de manquer des règles importantes. Pour éviter aussi le sur-ajustement, CPAR calcule la précision attendue appélée l'estimation d'erreur attendue de Laplace pour évaluer la précision de chaque règle [5]. La précision attendue est définie par

$$\text{LaplaceAccuracy} = (n_c + 1)/(n_{tot} + k)$$

où k est le nombre des classes de la variable réponse, n_c est le nombre d'observations dans la classe c prédite par la règle et n_{tot} est le nombre total d'observations.
L'algorithme CPAR utilise les k meilleures règles pour la prédiction de la classe d'une observation.

La procédure que nous proposons dans cette thèse s'inspire des algorithmes précédemment cités. Nous adoptons la méthode utilisée dans les algorithmes CBA et CMAR pour générer les règles, mais aussi le test d'indépendance pour élaguer les règles qui ne sont pas corrélées avec la variable réponse. Nous avons utilisé la mesure de l'entropie pour discrétiser les variables numériques au lieu de l'utiliser dans la procédure de génération des règles telle qu'elle a été utilisée dans CPAR.
La différence majeure entre la procédure que nous proposons et les algorithmes précédents se situe au niveau de la recherche de l'ensemble optimal des profils qui seront combinés pour construire un classifieur. L'idée principale de la procédure consiste à utiliser les outils de la statistique inférentielle pour sélectionner un ensemble réduit et optimal de profils. Dans la procédure, nous avons utilisé des mesures statistiques telles que la sensibilité et la spécificité pour réduire l'ensemble des profils candidats et ensuite nous avons utilisé la valeur prédictive positive pour sélectionner l'ensemble réduit et optimal de profils qui définiront un classifieur.

Bibliographie

[1] OMS | OMD 5 : améliorer la santé maternelle. 1

[2] Breiman, L. Bagging predictors. *Machine Learning 24*, 2 (1996), 123–140. 7

[3] Campbell, O. M. R., Graham, W. J., and Lancet Maternal Survival Series steering group. Strategies for reducing maternal mortality : getting on with what works. *Lancet 368*, 9543 (2006), 1284–1299. PMID : 17027735. 1

[4] Chawla, N. V. C4.5 and imbalanced data sets : investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *Proceedings of the ICML'03 Workshop on Class Imbalances* (2003). 6

[5] Clark, P., and Boswell, R. Rule induction with CN2 : some recent improvements. Springer-Verlag, pp. 151–163. 8

[6] Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett. 27*, 8 (2006), 861–874. 7

[7] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. *SIGMOD Rec. 29*, 2 (2000), 1–12. 8

[8] He, H., and Garcia, E. A. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng. 21*, 9 (2009), 1263–1284. 3

[9] Hualin Wang, X. S. Bagging probit models for unbalanced classification. *IGI Global ch017* (2010), 290–296. 7

[10] Japkowicz, N., and Stephen, S. The class imbalance problem : A systematic study. *Intell. Data Anal. 6*, 5 (2002), 429–449.

[11] King, G., and Zeng, L. Logistic regression in rare events data. *Political Analysis 9* (2001), 137 – 163. 4

[12] Lee, S. S. Regularization in skewed binary classification. *Computational Statistics 14*, 2 (1999), 277–292. 7

Bibliographie

[13] Lee, S. S. Noisy replication in skewed binary classification. *Computational Statistics & Data Analysis 34*, 2 (2000), 165–191. 7

[14] Li, J., Fu, A. W.-c., and Fahey, P. Efficient discovery of risk patterns in medical data. *Artificial intelligence in medicine 45*, 1 (2009), 77–89. 8

[15] Li, W., Han, J., and Pei, J. CMAR : accurate and efficient classification based on multiple class-association rules. In *ICDM 2001, Proceedings IEEE International Conference on Data Mining, 2001* (2001), pp. 369–376. 8

[16] Liu, B., Hsu, W., and Ma, Y. Integrating classification and association rule mining. In *Proceedings of the 4th international conference on Knowledge Discovery and Data mining (KDD'98)* (1998), AAAI Press, pp. 80–86. 8

[17] Menardi, G., and Torelli, N. Training and assessing classification rules with unbalanced data. Working papers DEAMS, DEAMS - Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti", 2010. 7

[18] Quinlan, J. R., and Cameron-jones, R. M. FOIL : a midterm report. In *In Proceedings of the European Conference on Machine Learning* (1993), Springer-Verlag, pp. 3–20. 8

[19] Ronsmans, C. Severe acute maternal morbidity in low-income countries. *Best practice & research. Clinical obstetrics & gynaecology 23*, 3 (2009), 305–316. PMID : 19201657. 1

[20] Salama, P., Lawn, J., Bryce, J., Bustreo, F., Fauveau, V., Starrs, A., and Mason, E. Making the countdown count. *Lancet 371* (2008), 1219–1221. PMID : 18406842. 1

[21] Say, L., Souza, J. P., Pattinson, R. C., and WHO working group on Maternal Mortality and Morbidity classifications. Maternal near miss–towards a standard tool for monitoring quality of maternal health care. *Best practice & research. Clinical obstetrics & gynaecology 23*, 3 (2009), 287–296. PMID : 19303368. 1

[22] Xie, J., and Qiu, Z. The effect of imbalanced data sets on LDA : a theoretical and empirical analysis. *Pattern Recogn. 40*, 2 (2007), 557–562. 5

[23] Yin, X., and Han, J. CPAR : classification based on predictive association rules. In *Proceedings of 2003 SIAM International Conference on Data Minng (SDM'03). San Fransisco, CA* (2003). 8

Chapitre II

Apprentissage d'un classifieur binaire par règles d'association

1 Introduction

Dans le présent travail, nous proposons une approche consistant à déterminer les profils, expression d'interactions entre les covariables, corrélés avec la variable réponse pour construire une fonction de classement. Cette approche est en étroite liaison avec la notion de règles d'association. Des approches similaires ont été proposées dans la littérature du domaine de l'intelligence artificielle ces dernières années [6–8]. L'idée principale consiste à rechercher un ensemble optimal de profils à partir d'un ensemble de profils fréquents. La stratégie consiste à élaguer les profils redondants et les profils de faible performance en se basant essentiellement sur les mesures statistiques suivantes : la sensibilité, la spécificité et les valeurs prédictives. Le présent travail vise à insérer cette approche dans le cadre de la statistique traditionnelle et à montrer la pertinence de son application dans un problème réel.

2 Profils et classement basé sur un profil

On considère un couple de variables aléatoires (Y, X), où Y est une variable de Bernoulli et $X = (X_j)_{j=1:p}$ est une suite finie de p variables aléatoires où chaque X_j est une variable non numérique à q_j modalités $m_{h(j)}, h(j) = 1 : q_j; j = 1 : p$.

2.1 Profil

Définition 1. On appelle profil toute suite finie d'événements $\left(X_j = m_{h(j)}\right)_{j\in J}$, où $J \subseteq 1 : p$ et $m_{h(j)}$ est une modalité de la variable X_j.

La longueur du profil $\left(X_j = m_{h(j)}\right)_{j\in J}$ est égale à la taille (cardinal) de l'ensemble $J \subset 1 : p$. Pour simplifier les notations dans la suite, on écrit $m_h^{X_j}$ pour désigner la modalité $m_{h(j)}$ de la variable X_j et on note $\left(m_h^{X_j}\right)_{j\in J}$ pour désigner le profils $\left(X_j = m_{h(j)}\right)_{j\in J}$.

Un profil peut être vu comme la réalisation conjointe de $|J|$ variables $(X_j)_{j\in J}$. Plus la taille du profil est grande, plus le nombre de variables conjointement réalisées augmente. Dans le domaine de l'intelligence artificielle et de l'apprentissage automatique, un profil est plus connu sous le nom d'item-set. Un profil de taille k est un k-itemset. Un profil $\left(m_h^{X_j}\right)_{j\in J}$ peut être compris comme l'expression d'une interaction entre les différentes variables non numériques $(X_j)_{j\in J}$ qui le définissent. La taille d'un profil est équivalente à la complexité d'une interaction dans un modèle paramétrique tel que la régression logistique. La gestion des interactions existant entre les covariables est l'un des avantages d'un profil par rapport aux modèles paramétriques. Un profil est pertinent lorsque sa probabilité d'occurrence est significative.

Définition 2. Soient $\left(m_h^{X_l}\right)_{l\in L}$ et $\left(m_h^{X_j}\right)_{j\in J}$ deux profils. On dit que $\left(m_h^{X_j}\right)_{j\in J}$ est emboîté dans $\left(m_h^{X_l}\right)_{l\in L}$ si les conditions suivantes sont vérifiées.

a) $L \subset J$

b) $\forall\, l \in L, \forall h \in \{1 : q_l\} \quad \exists\,!\, j \in J, \exists\,!\, k \in \{1 : q_j\}$ tel que $m_h^{X_l} = m_k^{X_j}$

Ils sont disjoints si $L \cap J = \emptyset$.

2.2 Classement associé à un profil et paramètres de performance

On peut associer à tout profil $U = \left(m_h^{X_j}\right)_{j\in J}$ une fonction indicatrice $\phi(\cdot, U)$ définie par :

$$\phi(X, U) = \prod_{j\in J} \mathbb{1}_{\left(X_j = m_h^{X_j}\right)}(X)$$

Par définition $\phi(\cdot, U)$ est un classifieur binaire. $\phi(X, U) = 1$ si tous les événements $[X_j = m_h^{X_j}]$ sont conjointement réalisés. Dans le domaine de l'intelligence artificielle, on appelle couverture du profil $U = \left(m_h^{X_j}\right)_{j\in J}$ la probabilité $\Pr\{\phi(X, U) = 1\}$ et on appelle support du profil $U = \left(m_h^{X_j}\right)_{j\in J}$ la probabilité $\Pr\{\phi(X, U) = 1, Y = 1\}$.

Dans cette analyse, nous nous plaçons dans le cadre de la statistique pour aborder le problème. A chaque profil U, un seul classifieur $\phi(X, U)$ lui est associé. Par la suite, on peut remarquer que la pertinence d'un profil est étroitement liée avec la performance du classifieur qui lui est associé. Ainsi on peut donc utiliser les indicateurs de performance des classifieurs associés pour sélectionner un ensemble réduit de profils pertinents dont on se servira pour construire une règle de classement efficace. Cependant plusieurs indicateurs de performance ont été proposés dans la littérature pour évaluer les performances d'un classifieur donné. Parmi les plus utilisés figure l'erreur de classement. L'erreur de classement $Err(U, Y)$ d'un classifieur $\phi(X, U)$ engendré par un profil U est définie par :

$$Err(U, Y) = \Pr\{\phi(X, U) \neq Y\} = \Pr\{\phi(X, U) = 1, Y = 0\} + \Pr\{\phi(X, U) = 0, Y = 1\}$$

On peut en déduire alors l'expression suivante :

$$Err(U,Y) = \Pr\{Y=1\} + \Pr\{\phi(X,U)=1\} - 2\Pr\{Y=1,\phi(X,U)=1\}$$

On constate que l'erreur de classement est gouverné par le support $\Pr\{Y=1,\phi(X,U)=1\}$ du profil U. L'erreur de classement est une fonction décroissante du support du profil. Pour deux profils de même couverture, l'erreur de classement décroît avec le support des profils. Par conséquent, plus le support du profil est élevé meilleur est le profil. On s'intéressera alors aux profils pour lesquels les classifieurs associés réalisent des probabilités $\Pr(Y=1,\phi(X,U)=1)$ supérieurs à un seuil s_0.

Pour un classifieur binaire, on considère en particulier la sensibilité et la spécificité définie par

$$\text{Sensib}(U,Y) = \frac{\Pr(\phi(X,U)=1,Y=1)}{\Pr(Y=1)}$$

$$\text{Spécif}(U,Y) = \frac{\Pr(\phi(X,U)=0,Y=0)}{\Pr(Y=0)}$$

On observe que la sensibilité croît avec la probabilité $\Pr(\phi(X,U)=1,Y=1)$. Deux autres paramètres pourront aider à l'évaluation de la qualité du classifieur $\phi(X,U)$ donc à la sélection du classifieur dans un ensemble de classifieurs : la valeur prédictive positive (VPP) et la valeur prédictive négative (VPN).

$$VPP(U,Y) = \frac{\Pr(\phi(X,U)=1,Y=1)}{\Pr(\phi(X,U)=1)}$$

$$VPN(U,Y) = \frac{\Pr(\phi(X,U)=0,Y=0)}{\Pr(\phi(X,U)=0)}$$

On peut établir les relations suivantes :

$$\text{Sensib}(U,Y) = VPP(U,Y)\frac{\Pr(\phi(X,U)=1)}{\Pr(Y=1)}$$

$$\text{Spécif}(U,Y) = 1 - [1-VPP(U,Y)]\,\frac{\Pr(\phi(X,U)=1)}{1-\Pr(Y=1)}$$

$$VPN(U,Y) = [1-VPP(U,Y)]\,\frac{\Pr(\phi(X,U)=1)}{\Pr(\phi(X,U)=0)}$$

Pour deux profils U_1 et U_2 de même probabilité d'occurrence (couverture), la spécificité croît avec la valeur prédictive positive du classifieur. Il en résulte que parmi les profils U de même couverture $\Pr(\phi(X,U)=1)$, on pourra s'intéresser à ceux pour lesquels les valeurs prédictives positives des classifieurs associés sont au dessus d'un seuil c_0 .

La valeur prédictive positive d'un profil est communément appelée confiance dans le domaine de

l'intelligence artificielle et de l'apprentissage automatique. En plus de la valeur prédictive positive (VPP) et de la valeur prédictive négative (VPN), on peut aussi baser la sélection des profils sur les paramètres suivants :
Le rapport de vraisemblance positif du profil U que nous notons par $RVP()$ est défini par :

$$RVP(U,Y) = \frac{\Pr(\phi(X,U)=1|Y=1)}{\Pr(\phi(X,U)=1|Y=0)}$$

on a alors

$$\begin{aligned} RVP(U,Y) &= \frac{\Pr\{Y=0\}}{\Pr\{Y=1\}} \frac{\Pr\{\phi(X,U)=1, Y=1\}}{\Pr\{\phi(X,U)=1, Y=0\}} \\ &= \frac{VPP(U,Y)}{1-VPP(U,Y)} \frac{\Pr\{Y=0\}}{\Pr\{Y=1\}} \end{aligned}$$

Le rapport de vraisemblance négatif du profil U que nous notons par $RVN()$ est défini par :

$$RVN = \frac{\Pr(\phi(X,U)=0|Y=1)}{\Pr(\phi(X,U)=0|Y=0)}$$

on a alors

$$\begin{aligned} RVN(U,Y) &= \frac{\Pr\{Y=0\}}{\Pr\{Y=1\}} \frac{\Pr\{\phi(X,U)=0\} - \Pr\{\phi(X,U)=0, Y=0\}}{\Pr\{\phi(X,U)=0, Y=0\}} \\ &= \frac{1-VPN(U,Y)}{VPN(U,Y)} \frac{\Pr\{Y=0\}}{\Pr\{Y=1\}} \end{aligned}$$

On a aussi $RVP(U,Y) = \frac{Sensibilite(U,Y)}{1-Specificite(U,Y)}$. Et donc $RVP(U,V) > 1$ entraîne que le classifieur $\phi(X,U)$ a de meilleurs indicateurs de performance que le classifieur de même sensibilité qui consiste à classer positive au hasard toute nouvelle observation. C'est à dire sur une courbe ROC [2], la courbe du classifieur $\phi(X,U)$ se situe au dessus de la première bissectrice.

Le risque relatif du profil U que nous notons $RR()$ est défini par :

$$RR(U,Y) = \frac{\Pr(Y=1|\phi(X,U)=1)}{\Pr(Y=1|\phi(X,U)=0)}$$

On peut établir que

$$RR(U,Y) = VPP(U,Y) \frac{1-\Pr\{\phi(X,U)=1\}}{\Pr\{Y=1\} - \Pr\{Y=1, \phi(X,U)=1\}}$$

Notons par G_1 le groupe d'objets vérifiant le profil U et G_0 le groupe d'objets ne vérifiant pas le profil

U. Le risque relatif est une mesure statistique qui permet de comparer la probabilité d'occurrence de l'événement $[Y = 1]$ dans G_1 par rapport à G_0. Le profil $U = \left(m_h^{X_l}\right)_{l \in L}$ est un profil à risque [4] pour Y si le risque relatif excède un seuil τ donné. La probabilité d'occurrence de l'événement $[Y = 1]$ dans G_1 est τ fois plus importante que la probabilité d'occurrence de l'événement $[Y = 1]$ dans G_0. Dans la suite, nous nous intéresserons alors aux profils U pour lesquels la probabilité d'occurrence de $[Y = 1]$ dans G_1 est τ fois plus importante que la probabilité d'occurrence de $[Y = 1]$ dans G_0. Par défaut le paramètre τ est supérieur à un ($\tau > 1$). Le sous-ensemble de profils U pour lequel la probabilité conditionnelle $\Pr([Y = 1]|[\phi(X_i, U) = 1])$ est plus élevée que la probabilité conditionnelle $\Pr([Y = 1]|[\phi(X_i, U) = 0])$ constitue un ensemble potentiel pour construire un bon classifieur.

Les critères conventionnels d'évaluation utilisés, tels que la précision globale ou le taux d'erreur, ne fournit pas suffisamment d'informations dans le cas de l'apprentissage déséquilibré. En effet, des mesures d'évaluation plus performantes, telles que les courbes ROC (receiver operating characteristc), les courbes de précision-sensibilité et les courbes de coûts, sont nécessaires à l'évaluation concluante d'un classifieur en présence de données déséquilibrées. L'expression de l' aire en dessous de la courbe ROC d'un classifieur généré par un profil U est donnée par :

$$AUC(U, Y) = \begin{cases} \frac{1}{2}(\text{Sensib}(U, Y) + \text{Spécif}(U, Y)) & \text{si Sensib}(U, Y) + \text{Spécif}(U, Y) \geq 1 \\ 1 - \frac{1}{2}(\text{Sensib}(U, Y) + \text{Spécif}(U, Y)) & \text{si Sensib}(U, Y) + \text{Spécif}(U, Y) < 1 \end{cases}$$

L'aire sous la courbe ROC (AUC) est une mesure utile pour évaluer la performance d'un profil. La comparaison des AUC de différents profils peut établir une relation de domination entre les profils. On peut l'utiliser alors pour la sélection d'un sous ensemble optimal de profils.

A partir de cette section, il apparaît clairement que les principaux paramètres d'apprentissage d'un classifieur basé sur un ensemble optimal de profils sont le support et la valeur prédictive positive. Ils permettent de gérer à la fois l'erreur de classement et la sensibilité du classifieur. Dans toute la suite, nous nous intéressons aux profils dont le support est supérieur à un seuil s_0 et la valeur prédictive positive (confiance) est supérieure à c_0.

Dans la littérature de la classification associative (classification supervisée basée sur les règles d'association), plusieurs mesures de performance ont été proposées pour l'extraction de règles d'association [3]. Une étude comparative exhaustive de plusieurs mesures de performance a été menée dans [9]. La plupart des mesures de performance sont destinées à découvrir les profils les plus fréquents. Raison pour laquelle la majeur partie d'entre elles ne sont pas appropriées lorsqu'il s'agit de traiter un problème de classification supervisée sur des données déséquilibrées. Le support et la confiance restent les mesures de performance les plus utilisées dans les algorithmes d'extraction des règles d'association basés sur la sélection des profils fréquents. Dans ces algorithmes, généralement le support est utilisé pour trouver les profils fréquents suivant sa propriété d'anti-monotonicité [1, 4]. Quant à la confiance elle est utilisée pour générer les règles à partir des profils fréquents et à les filtrer à l'aide d'un seuil de confiance minimum.

Selon ses propriétés, chaque mesure est utile pour certaines applications, mais pas pour d'autres [12]. Ces mesures peuvent produire des informations contradictoires sur l'intérêt et la pertinence d'un profil. Un exemple bien connu d'une telle mesure controversée est le support. D'une part, il est grandement utilisé à des fins de filtrage dans les algorithmes d'extraction [1, 10], puisque sa propriété d'anti-monotonicité simplifie le vaste ensemble de profils qui doit être exploré. D'autre part, il a presque tous les défauts qu'un utilisateur souhaite éviter par exemple la variabilité de la valeur sous l'hypothèse d'indépendance [11].

A notre connaissance, seuls la sensibilité connue sous le nom de support local, le risque relatif et l'odds ratio ont été utilisés pour la recherche d'un ensemble optimal de profil dans le cadre d'un problème de classification supervisée sur des données déséquilibrées par Li et al.[5]. En présence de données déséquilibrées, le support $\Pr\{\phi(X,U)=1, Y=1\}$ d'un profil U serait guère fréquent lorsque la classe d'intérêt $\{Y=1\}$ est rare. C'est pourquoi Li et al. ont définit le support local (sensibilité) $\Pr\{\phi(X,U)=1 \,|\, Y=1\}$ comme étant le support d'un profil dans le groupe d'observations vérifiant $\{Y=1\}$ puisque le support local vérifie la propriété d'anti-monotonicité du support. Ainsi un profil U est fréquent lorsque son support local est supérieure à un seuil minimum fixé. Leurs résultats ont montré que la sensibilité et le risque relatif sont des mesures statistiques pertinentes pour la sélection de profils optimaux lorsqu'on traite des données déséquilibrées.

Les algorithmes d'extraction de règles d'association basés sur les profils fréquents produisent en général un vaste ensemble de règles d'association dont la majeur partie sont triviales et sans intérêts. Pour construire un classifieur performant à partir du vaste ensemble de règles d'association explorées, nous allons donc établir une stratégie d'élagage des profils redondants et une stratégie de réduction de l'ensemble des profils fréquents et non redondants.

2.3 Profils redondants et sélection de profils

A l'instar des méthodes standards de classement, la procédure de sélection de profils que nous proposons s'intéressera en particulier aux profils qui sont corrélés avec la variable réponse.

Proposition 1. Soient $U = \left(m_h^{X_l}\right)_{l\in L}$ et $U' = \left(m_h^{X_j}\right)_{j\in J}$ deux profils. Si U' est emboîté dans U alors :

1. $\Pr\{\phi(X,U)=1, Y=1\} \geq \Pr\{\phi(X,U')=1, Y=1\}$
2. $\Pr\{\phi(X,U)=0, Y=0\} \leq \Pr\{\phi(X,U')=0, Y=0\}$

Preuve. Pour simplifier les expressions, on note $\phi(X,U)$ par ϕ_U et $\phi(X,U')$ par $\phi_{U'}$.
Par hypothèse U' est emboité dans U donc on a

$$\{\phi_U=1\} \supset \{\phi_{U'}=1\} \quad \text{et} \quad \{\phi_U=0\} \subset \{\phi_{U'}=0\}$$

On en déduit que :

$$\{\phi_U=1\} \supset \{\phi_{U'}=1\} \Rightarrow \{\phi_U=1, Y=1\} \supset \{\phi_{U'}=1, Y=1\} \Rightarrow \Pr\{\phi_U=1, Y=1\} \geq \Pr\{\phi_{U'}=1, Y=1\}$$

$$\{\phi_U=0\} \subset \{\phi_{U'}=0\} \Rightarrow \{\phi_U=0, Y=0\} \subset \{\phi_{U'}=0, Y=0\} \Rightarrow \Pr\{\phi_U=0, Y=0\} \leq \Pr\{\phi_{U'}=0, Y=0\}$$

□

Définition 3. Soient $U = \left(m_h^{X_l}\right)_{l\in L}$ et $U' = \left(m_h^{X_j}\right)_{j\in J}$ deux profils tels que U' soit emboîté dans U. On dit que le profil U' est redondant par rapport à U, si le(s) indicateur(s) de performance de la fonction de classement $\phi(\cdot, U)$ générée par le profil U est (sont) plus élevé(s) que le(s) indicateur(s) de performance de la fonction de classifieur $\phi(\cdot, U')$ générée par le profil U'.

Proposition 2. Soient $U = \left(m_h^{X_l}\right)_{l\in L}$ et $U' = \left(m_h^{X_j}\right)_{j\in J}$ deux profils. Si U' est emboîté dans U alors la valeur prédictive positive du classifieur généré par le profil U est comprise entre

$$Min\left\{VPP(U',Y), \frac{\Pr(Y=1,\phi(X,U)=1) - \Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1) - \Pr(\phi(X,U')=1)}\right\}$$

et

$$Max\left\{VPP(U',Y), \frac{\Pr(Y=1,\phi(X,U)=1) - \Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1) - \Pr(\phi(X,U')=1)}\right\}$$

Preuve. On a

$$\begin{aligned}
VPP(U,Y) &= \frac{\Pr(Y=1,\phi(X,U)=1)}{\Pr(\phi(X,U)=1)} \\
VPP(U,Y) &= \frac{\Pr(Y=1,\phi(X,U)=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} + \frac{\Pr(Y=1,\phi(X,U)=1,\phi(X,U')=0)}{\Pr(\phi(X,U)=1)} \\
&= \frac{\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} + \frac{\Pr(Y=1,\phi(X,U)=1) - \Pr(Y=1,\phi(X,U)=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} \\
&= VPP(U',Y)\frac{\Pr(\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} + \frac{\Pr(Y=1,\phi(X,U)=1) - \Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1) - \Pr(\phi(X,U')=1)}\left[1 - \frac{\Pr(\phi(X,U')=1)}{\Pr(\phi(X,U)=1)}\right]
\end{aligned}$$

On obtient une combinaison convexe de $VPP(U,Y)$ et $\frac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)-\Pr(\phi(X,U')=1)}$ par rapport à $\frac{\Pr(\phi(X,U')=1)}{\Pr(\phi(X,U)=1)}$. On en déduit que $VPP(U,Y$ est comprise entre

$$Min\left\{\frac{\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U')=1)},\frac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)-\Pr(\phi(X,U)=1)}\right\}$$

et

$$Max\left\{\frac{\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U')=1)},\frac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)-\Pr(\phi(X,U)=1)}\right\}$$

□

En résumé de la proposition 2, on a

1. Si $VPP(Y,U') < \dfrac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)-\Pr(\phi(X,U')=1)}$

   alors $VPP(Y,U') < VPP(Y,U)$. C'est à dire que U' est redondant par rapport à U. Par conséquent, on peut éliminer le plus long puisque sa sensibilité est plus faible et son erreur de classement est plus forte. Par contre

2. Si $VPP(Y,U') > \dfrac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)-\Pr(\phi(X,U')=1)}$

   alors $VPP(Y,U') > VPP(Y,U)$. Il est préférable de garder le profil U' au profit du profil U, puisque les indicateurs de performance du profil U sont meilleurs que ceux du profil U.

Proposition 3. Soient $U=\left(m_h^{X_l}\right)_{l\in L}$ et $U'=\left(m_h^{X_j}\right)_{j\in J}$ deux profils tels que U' soit emboîté dans U. Alors $\Pr\{\phi(X,U)=1\}=\Pr\{\phi(X,U')=1\}$ si et seulement si

1. $\Pr\{\phi(X,U)=1,Y=1\}=\Pr\{\phi(X,U')=1,Y=1\}$
2. $\Pr\{\phi(X,U)=0,Y=0\}=\Pr\{\phi(X,U')=0,Y=0\}$

Preuve. Supposons que $\Pr\{\phi_U=1\}=\Pr\{\phi_{U'}=1\}$. On a

$$\Pr\{\phi_U=1\}=\Pr\{\phi_U=1,Y=1\}+\Pr\{\phi_U=1,Y=0\}$$

$$\Pr\{\phi_{U'}=1\}=\Pr\{\phi_{U'}=1,Y=1\}+\Pr\{\phi_{U'}=1,Y=0\}$$

On obtient

$$\Pr\{\phi_U=1,Y=1\}+\Pr\{\phi_U=1,Y=0\}=\Pr\{\phi_{U'}=1,Y=1\}+\Pr\{\phi_{U'}=1,Y=0\} \qquad (a)$$

Puisque $[\phi_{U'}=1]\subset[\phi_U=1]$ alors

$$\Pr\{\phi_U=1,Y=1\}-\Pr\{\phi_{U'}=1,Y=1\}\geq 0 \qquad (b)$$

$$\Pr\{\phi_U=1,Y=0\}-\Pr\{\phi_{U'}=1,Y=0\}\geq 0 \qquad (c)$$

On peut déduire de (a), (b) et (c) les égalités suivantes :

$$\Pr\{\phi_U=1,Y=1\}=\Pr\{\phi_{U'}=1,Y=1\} \qquad (1)$$

$$\Pr\{\phi_{U'}=1, Y=0\} = \Pr\{\phi_U=1, Y=0\} \qquad (*)$$

Par ailleurs on a

$$\Pr\{\phi_{U'}=1, Y=0\} = \Pr\{Y=0\} - \Pr\{\phi_{U'}=0, Y=0\} \qquad (**)$$

$$\Pr\{\phi_U=1, Y=0\} = \Pr\{Y=0\} - \Pr\{\phi_U=0, Y=0\} \qquad (***)$$

En faisant la différence membre à membre des égalités $(**)$ et $(***)$ et en tenant compte de l'égalité $(*)$, on obtient

$$\Pr\{\phi_{U'}=0, Y=0\} = \Pr\{\phi_U=0, Y=0\} \qquad (2)$$

Supposons maintenant que les égalités suivantes soient vraies :

$$\Pr\{\phi_U=1, Y=1\} = \Pr\{\phi_{U'}=1, Y=1\} \qquad (1)$$

$$\Pr\{\phi_{U'}=0, Y=0\} = \Pr\{\phi_U=0, Y=0\} \qquad (2)$$

De l'égalité (2) on déduit

$$\Pr\{Y=0\} - \Pr\{\phi_U=1, Y=0\} = \Pr\{Y=0\} - \Pr\{\phi_{U'}=1, Y=0\}$$

On obtient alors les égalités suivantes

$$\Pr\{\phi_U=1, Y=1\} = \Pr\{\phi_{U'}=1, Y=1\}$$

$$\Pr\{\phi_U=1, Y=0\} = \Pr\{\phi_{U'}=1, Y=0\}$$

En faisant les sommes membres à membres des deux égalités on obtient :

$$\Pr\{\phi_U=1, Y=1\} + \Pr\{\phi_U=1, Y=0\} = \Pr\{\phi_U=1\}$$

$$\Pr\{\phi_{U'}=1, Y=1\} + \Pr\{\phi_{U'}=1, Y=0\} = \Pr\{\phi_{U'}=1\}$$

D'où

$$\Pr\{\phi_U=1\} = \Pr\{\phi_{U'}=1\}$$

□

Lorsqu'on divise par $\Pr(Y=1)$ les deux termes de l'égalité 1 de la proposition 3, on obtient que le profil U' est redondant par rapport au profil U selon la définition 3. Le même résultat est obtenu en divisant les deux termes de l'égalité 2 par $\Pr(Y=0)$.

Corollaire 1. Soient $U = \left(m_h^{X_l}\right)_{l\in L}$ et $U' = \left(m_h^{X_j}\right)_{j\in J}$ deux profils tels que U' soit emboîté dans U. Les propositions suivantes sont équivalentes :

1.

$$\Pr\{\phi(X,U)=1\} = \Pr\{\phi(X,U')=1\}$$

2.

$$\begin{cases} VPP(U,Y) = VPP(U',Y) \\ \Pr\{\phi(X,U)=1, Y=1\} = \Pr\{\phi(X,U')=1, Y=1\} \end{cases}$$

3.

$$\begin{cases} VPN(U,Y) = VPN(U',Y) \\ \Pr\{\phi(X,U)=0, Y=0\} = \Pr\{\phi(X,U')=0, Y=0\} \end{cases}$$

4.

$$\begin{cases} RVP(U,Y) = RVP(U',Y) \\ \Pr\{\phi(X,U)=1, Y=1\} = \Pr\{\phi(X,U')=1, Y=1\} \end{cases}$$

5.

$$\begin{cases} RVN(U,Y) = RVN(U',Y) \\ \Pr\{\phi(X,U)=1, Y=1\} = \Pr\{\phi(X,U')=1, Y=1\} \end{cases}$$

6.

$$\begin{cases} Err(U,Y) = Err(U',Y) \\ \Pr\{\phi(X,U)=1, Y=1\} = \Pr\{\phi(X,U')=1, Y=1\} \end{cases}$$

7.

$$\begin{cases} RR(U,Y) = RR(U',Y) \\ \Pr\{\phi(X,U)=1, Y=1\} = \Pr\{\phi(X,U')=1, Y=1\} \end{cases}$$

Preuve. Pour simplifier les expressions, on note par $\phi(X,U)$ par ϕ_U et $\phi(X,U')$ par $\phi_{U'}$.

1)Montrons que 1. est équivalent à 2.
Supposons que 1. est vrai
D'après la proposition 3, si $\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$ alors

$$\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$$

Si on divise les termes respectives de cette dernière égalité par $\Pr\{\phi_U = 1\}$ et $\Pr\{\phi_{U'} = 1\}$ respectivement, on obtient

$$\frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 1\}} = \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 1\}}$$

Réciproquement : supposons que 2. soit vrai
Si 2. est vrai alors on a $VPP(U, Y) - VPP(U', Y) = 0$
D'où

$$\frac{\Pr\{\phi_U = 1, Y = 1\}\left[\Pr\{\phi_{U'} = 1\} - \Pr\{\phi_U = 1\}\right]}{\Pr\{\phi_U = 1\}\Pr\{\phi_{U'} = 1\}} = 0$$

On en déduit que

$$\Pr\{\phi_{U'} = 1\} - \Pr\{\phi_U = 1\} = 0$$

2) Montrons que 1. est équivalent à 3.
Supposons que 1. soit vrai
On a $\Pr\{\phi_U = 1\} = 1 - \Pr\{\phi_U = 0\}$ et $\Pr\{\phi_{U'} = 1\} = 1 - \Pr\{\phi_{U'} = 0\}$ Donc

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\} \quad \Rightarrow \quad \Pr\{\phi_U = 0\} = \Pr\{\phi_{U'} = 0\}$$

D'après la proposition 3, on a aussi $\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$ entraîne que

$$\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$$

En divisant les termes respectives de l'égalité ci-dessus par $\Pr\{\phi_U = 0\}$ et $\Pr\{\phi_{U'} = 0\}$ respectivement, on obtient

$$\frac{\Pr\{\phi_U = 0, Y = 0\}}{\Pr\{\phi_U = 0\}} = \frac{\Pr\{\phi_U = 0, Y = 0\}}{\Pr\{\phi_{U'} = 0\}}$$

Réciproquement : supposons que 2. soit vrai
Si 3.1 est vrai alors $VPN(U, Y) - VPN(U', Y) = 0$. On obtient donc

$$\frac{\Pr\{\phi_U = 0, Y = 0\}\left[\Pr\{\phi_{U'} = 0\} - \Pr\{\phi_U = 0\}\right]}{\Pr\{\phi_U = 0\}\Pr\{\phi_{U'} = 0\}} = 0$$

Il en résulte que

$$\begin{aligned} \Pr\{\phi_{U'} = 0\} &= \Pr\{\phi_U = 0\} \\ 1 - \Pr\{\phi_{U'} = 1\} &= 1 - \Pr\{\phi_U = 1\} \end{aligned}$$

d'où

$$\Pr\{\phi_{U'} = 1\} = \Pr\{\phi_U = 1\}$$

3) Montrons que 1. est équivalent à 4.
On suppose que 1. est vrai
Par définition on a

$$RVP(U, Y) = \frac{\Pr\{Y = 0\}}{\Pr\{Y = 1\}} \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{Y = 0\} - \Pr\{\phi_U = 0, Y = 0\}}$$

Et d'après la proposition 3, on a $\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$ entraîne que

$$\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\} \qquad \text{et} \qquad \Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$$

Donc si on remplace $\Pr\{\phi_U = 1, Y = 1\}$ par $\Pr\{\phi_{U'} = 1, Y = 1\}$ et $\Pr\{\phi_U = 0, Y = 0\}$ par $\Pr\{\phi_{U'} = 0, Y = 0\}$ dans l'expression de $RVP(U, Y)$, on obtient

$$RVP(U,Y) = \frac{\Pr\{Y=0\}}{\Pr\{Y=1\}} \frac{\Pr\{\phi_{U'}=1, Y=1\}}{\Pr\{Y=0\} - \Pr\{\phi_{U'}=0, Y=0\}}$$

D'où

$$RVP(U,Y) = RVP(U',Y)$$

Réciproquement : supposons que 4. soit vrai
Si 4. est vrai alors $RVP(U,Y) - RVP(U',Y) = 0$. Il en résulte que

$$\frac{\Pr\{\phi_U=1, Y=1\}}{\Pr\{Y=0\} - \Pr\{\phi_U=0, Y=0\}} - \frac{\Pr\{\phi_{U'}=1, Y=1\}}{\Pr\{Y=0\} - \Pr\{\phi_{U'}=0, Y=0\}} = 0$$

Puisque $\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$ on en déduit donc que $\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_U = 0, Y = 0\}$.
D'après la proposition 3, on a donc

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$$

4) Montrons que 1. est équivalent à 5.
On suppose que 1. est vrai

Par définition on a

$$RVN(U,Y) = \frac{\Pr\{Y=0\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\} - \Pr\{\phi_U=1, Y=1\}}{\Pr\{\phi_U=0, Y=0\}}$$

Et d'après la proposition 3, on a $\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$ entraîne que

$$\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\} \qquad \text{et} \qquad \Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$$

Donc si on remplace $\Pr\{\phi_U = 1, Y = 1\}$ par $\Pr\{\phi_{U'} = 1, Y = 1\}$ et $\Pr\{\phi_U = 0, Y = 0\}$ par $\Pr\{\phi_{U'} = 0, Y = 0\}$ dans l'expression de $RVN(U, Y)$, on obtient

$$RVN(U,Y) = \frac{\Pr\{Y=0\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\} - \Pr\{\phi_{U'}=1, Y=1\}}{\Pr\{\phi_{U'}=0, Y=0\}}$$

D'où

$$RVN(U,Y) = RVN(U',Y)$$

Réciproquement : supposons que 5. soit vrai
Si 5. est vrai alors $RVN(U,Y) - RVN(U',Y) = 0$. On peut en déduire que

$$\frac{\Pr\{Y=1\} - \Pr\{\phi_U=1, Y=1\}}{\Pr\{\phi_U=0, Y=0\}} - \frac{\Pr\{Y=1\} - \Pr\{\phi_{U'}=1, Y=1\}}{\Pr\{\phi_{U'}=0, Y=0\}} = 0$$

Puisque $\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$, on obtient donc que $\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$. D'où

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$$

d'après la proposition 3

5) Montrons que 1. est équivalent à 6.
On suppose que 1. est vrai
On a

$$Err(U,Y) = \Pr\{Y=1\} + \Pr\{\phi_U=1\} - 2\Pr\{\phi_U=1, Y=1\} \qquad (1)$$

$$Err(U',Y) = \Pr\{Y=1\} + \Pr\{\phi_{U'}=1\} - 2\Pr\{\phi_{U'}=1, Y=1\} \qquad (2)$$

Si $\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$ alors $\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$ (proposition 3)

Il en résulte des égalités précédentes que

$$Err(U, Y) = Err(U', Y)$$

Réciproquement : supposons que 6. soit vrai
Si on les différences membre à membres des égalités (1) et (2) ci-dessus, on obtient

$$\Pr\{\phi_U = 1\} - \Pr\{\phi_{U'} = 1\} = Err(U, Y) - Err(U', Y) + 2(\Pr\{\phi_U = 1, Y = 1\} - \Pr\{\phi_{U'} = 1, Y = 1\})$$

donc si 6. est vrai alors $\Pr\{\phi_U = 1\} - \Pr\{\phi_{U'} = 1\} = 0$

6) Montrons que 1. est équivalent à 7.
On suppose que 1. est vrai
On a

$$\begin{aligned} \Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\} \quad &\Leftrightarrow \quad 1 - \Pr\{\phi_U = 1\} = 1 - \Pr\{\phi_{U'} = 1\} \\ &\Leftrightarrow \quad \Pr\{\phi_U = 0\} = \Pr\{\phi_{U'} = 0\} \end{aligned}$$

alors

$$\frac{\Pr\{\phi_U = 0\}}{\Pr\{\phi_U = 1\}} = \frac{\Pr\{\phi_{U'} = 0\}}{\Pr\{\phi_{U'} = 1\}} \qquad (1)$$

D'autre part, on a $\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$ entraîne que

(a) $\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$

(b) $\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$

d'aprés la proposition 3. Puisque $\Pr\{\phi_U = 0\} = \Pr\{\phi_{U'} = 0\}$ et $\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$ alors

$$\Pr\{\phi_U = 0, Y = 1\} = \Pr\{\phi_{U'} = 0, Y = 1\}$$

On en déduit que

$$\frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 0, Y = 1\}} = \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 0, Y = 1\}} \qquad (2)$$

En faisant les produit membre à membre des égalités (1) et (2) on obtient

$$\frac{\Pr\{\phi_U = 0\}}{\Pr\{\phi_U = 1\}} \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 0, Y = 1\}} = \frac{\Pr\{\phi_{U'} = 0\}}{\Pr\{\phi_{U'} = 1\}} \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 0, Y = 1\}}$$

Il en résulte que

$$RR(U, Y) = RR(U', Y)$$

Réciproquement : supposons que 7. soit vrai alors $RR(U, Y) - RR(U', Y) = 0$. Donc

$$\frac{VPP(U, Y)}{1 - VPN(U, Y)} - \frac{VPP(U', Y)}{1 - VPN(U', Y)} = 0$$

d'où $VPP(U, Y) = VPP(U', Y)$.
On a donc $VPP(U, Y) = VPP(U', Y)$ et $\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$. Il en résulte que

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$$

□

Proposition 4. Soient $U = \left(m_h^{X_l}\right)_{l \in L}$ et $U' = \left(m_h^{X_j}\right)_{j \in J}$ deux profils tels que U' soit emboîté dans U. Si $\Pr\{\phi(X,U) = 1, Y = 1\} = \Pr\{\phi(X,U') = 1, Y = 1\}$ alors

1. $VPP(U,Y) \leq VPP(U',Y)$
2. $VPN(U,Y) \leq VPN(U',Y)$
3. $RVP(U,Y) \leq RVP(U',Y)$
4. $RVN(U,Y) \geq RVN(U',Y)$
5. $Err(U,Y) \geq Err(U',Y)$
6. $RR(U,Y) \leq RR(U',Y)$

Preuve. Pour simplifier les expressions, on note par $\phi(X,U)$ par ϕ_U et $\phi(X,U')$ par $\phi_{U'}$.
1) Montrons que $VPP(U,Y) \leq VPP(U',Y)$
On a U' emboîté dans U entraîne que $\{\phi_U = 1\} \supset \{\phi_{U'} = 1\}$. Donc

$$\frac{1}{\Pr\{\phi_U = 1\}} \leq \frac{1}{\Pr\{\phi_{U'} = 1\}}$$

Si l'égalité $\Pr\{\phi_U = 1,Y = 1\} = \Pr\{\phi_{U'} = 1,Y = 1\}$ est vérifiée alors

$$\begin{aligned} \frac{\Pr\{\phi_U = 1,Y = 1\}}{\Pr\{\phi_U = 1\}} &= \frac{\Pr\{\phi_{U'} = 1,Y = 1\}}{\Pr\{\phi_U = 1\}} \\ &\leq \frac{\Pr\{\phi_{U'} = 1,Y = 1\}}{\Pr\{\phi_{U'} = 1\}} \qquad (1) \end{aligned}$$

On obtient donc

$$VPP(U,Y) \leq VPP(U',Y)$$

2) Montrons que $VPN(U,Y) \leq VPN(U',Y)$
On a U' emboîté dans U entraîne que $\{\phi_U = 1\} \supset \{\phi_{U'} = 1\}$. Donc

$$\frac{1}{\Pr\{\phi_{U'} = 0\}} \leq \frac{1}{\Pr\{\phi_U = 0\}}$$

Par ailleurs si on a $\Pr\{\phi_U = 1,Y = 1\} = \Pr\{\phi_{U'} = 1,Y = 1\}$ alors

$$\Pr\{\phi_U = 0,Y = 1\} = \Pr\{\phi_{U'} = 0,Y = 1\}$$

On en déduit que

$$\Pr\{\phi_U = 0\} - \Pr\{\phi_U = 0,Y = 0\} = \Pr\{\phi_{U'} = 0\} - \Pr\{\phi_{U'} = 0,Y = 0\}$$

donc

$$\begin{aligned} 1 - \frac{\Pr\{\phi_U = 0,Y = 0\}}{\Pr\{\phi_U = 0\}} &\geq 1 - \frac{\Pr\{\phi_{U'} = 0,Y = 0\}}{\Pr\{\phi_{U'} = 0\}} \\ \frac{\Pr\{\phi_U = 0,Y = 0\}}{\Pr\{\phi_U = 0\}} &\leq \frac{\Pr\{\phi_{U'} = 0,Y = 0\}}{\Pr\{\phi_{U'} = 0\}} \end{aligned}$$

On obtient donc

$$VPN(U,Y) \leq VPV(U',Y)$$

3) Montrons que $RVP(U,Y) = RVP(U',Y)$

Par définition on a

$$\begin{aligned} RVP(U,Y) &= \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{\phi_U=1, Y=1\}}{\Pr\{Y=0\} - \Pr\{\phi_U=0, Y=0\}} \\ &= \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{\phi_U=1, Y=1\}}{\Pr\{\phi_U=1, Y=0\}} \\ &= \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{\phi_U=1, Y=1\}}{\Pr\{\phi_U=1\} - \Pr\{\phi_U=1, Y=1\}} \end{aligned}$$

donc si $\Pr\{\phi_U=1, Y=1\} = \Pr\{\phi_{U'}=1, Y=1\}$ et U' emboîté dans U alors

$$\Pr\{\phi_U=1\} - \Pr\{\phi_U=1, Y=1\} \geq \Pr\{\phi_{U'}=1\} - \Pr\{\phi_{U'}=1, Y=1\}$$

donc

$$RVP(U,Y) \leq \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{\phi_{U'}=1, Y=1\}}{\Pr\{\phi_{U'}=1\} - \Pr\{\phi_{U'}=1, Y=1\}}$$

d'où

$$RVP(U,Y) \leq RVP(U',Y)$$

4) Montrons que $RVN(U,Y) \geq RVN(U',Y)$

Par définition on a

$$RVN(U,Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\} - \Pr\{\phi_U=1, Y=1\}}{\Pr\{\phi_U=0, Y=0\}}$$

par hypothèse $\Pr\{\phi_U=1, Y=1\} = \Pr\{\phi_{U'}=1, Y=1\}$ on a alors

$$RVN(U,Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\} - \Pr\{\phi_{U'}=1, Y=1\}}{\Pr\{\phi_U=0, Y=0\}}$$

Par ailleurs U' emboîté dans U entraîne que $\Pr\{\phi_U=0, Y=0\} \leq \Pr\{\phi_{U'}=0, Y=0\}$. On en déduit que

$$RVN(U,Y) \geq \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\} - \Pr\{\phi_{U'}=1, Y=1\}}{\Pr\{\phi_{U'}=0, Y=0\}}$$

d'où

$$RVN(U,Y) \geq RVN(U',Y)$$

5) Montrons que $Err(U,Y) \geq Err(U',Y)$

Par définition on a

$$Err(U,Y) = \Pr\{Y=1\} + \Pr\{\phi_U=1\} - 2\Pr\{\phi_U=1, Y=1\}$$

$$Err(U',Y) = \Pr\{Y=1\} + \Pr\{\phi_{U'}=1\} - 2\Pr\{\phi_{U'}=1, Y=1\}$$

Par hypothèse on a $\Pr\{\phi_U=1, Y=1\} = \Pr\{\phi_{U'}=1, Y=1\}$ donc

$$Err(U,Y) - Err(U',Y) = \Pr\{\phi_U=1\} - \Pr\{\phi_{U'}=1\}$$

et puisque U' est emboité dans U alors $\Pr\{\phi_U = 1\} \geq \Pr\{\phi_{U'} = 1\}$. On obtient donc

$$Err(U,Y) \geq Err(U',Y)$$

6) Montrons que $RR(U,Y) \leq RR(U',Y)$

On a

$$\Pr\{\phi_U = 0,Y = 1\} = \Pr\{Y = 1\} - \Pr\{\phi_U = 1,Y = 1\}$$

Puisqu'on a $\Pr\{\phi_U = 1,Y = 1\} = \Pr\{\phi_{U'} = 1,Y = 1\}$, on obtient alors implique aussi

$$\Pr\{\phi_U = 0,Y = 1\} = \Pr\{Y = 1\} - \Pr\{\phi_{U'} = 1,Y = 1\}$$

D'où

$$\Pr\{\phi_U = 0,Y = 1\} = \Pr\{\phi_{U'} = 0,Y = 1\}$$

Puisque que U' est emboîté dans U, on en déduit que

$$\frac{\Pr\{\phi_U = 0,Y = 1\}}{\Pr\{\phi_U = 0\}} \quad \geq \quad \frac{\Pr\{\phi_{U'} = 0,Y = 1\}}{\Pr\{\phi_{U'} = 0\}} \qquad (2)$$

Si on fait le rapport membre à membre des inégalités (1) et (2), il en résulte que

$$RR(U,Y) \leq RR(U',Y)$$

□

Il découle de la proposition 4 que lorsque les fonctions de classification générées par deux profils emboîtés ont la même sensibilité et des spécificités différentes alors la fonction de classification générée par le profil le plus long a une erreur de classement plus faible, une valeur prédictive positive (confiance) plus élevée, un rapport de vraisemblance positif plus élevé, un rapport de vraisemblance négatif plus faible et un risque relatif plus élevé que celui de la fonction de classification générée par le profil le plus court. De plus U' emboîté dans U implique que la fonction de classification générée par U' a une spécificité plus élevée que celle de la fonction de classification générée par U. Par conséquent on préférera le profil le plus long puisque ses indicateurs de performance (sensibilité, spécificité et erreur de classement) sont meilleurs.

Proposition 5. Soient $U = \left(m_h^{X_l}\right)_{l \in L}$ et $U' = \left(m_h^{X_j}\right)_{j \in J}$ deux profils tels que U' soit emboîté dans U. Si $\Pr\{\phi(X,U) = 0, Y = 0\} = \Pr\{\phi(X,U') = 0, Y = 0\}$ alors

1. $VPP(U,Y) \geq VPP(U',Y)$
2. $VPN(U,Y) \geq VPN(U',Y)$
3. $RVP(U,Y) \geq RVP(U',Y)$
4. $RVN(U,Y) \leq RVN(U',Y)$
5. $Err(U,Y) \leq Err(U',Y)$
6. $RR(U,Y) \geq RR(U',Y)$

Preuve. 1) Montrons que $VPP(U,Y) \geq VPP(U',Y)$
Par définition

$$\frac{\Pr\{\phi_U = 1,Y = 1\}}{\Pr\{\phi_U = 1\}} = \frac{\Pr\{\phi_U = 1,Y = 1\}}{\Pr\{\phi_U = 1,Y = 1\} + \Pr\{\phi_U = 1,Y = 0\}}$$

et

$$\frac{\Pr\{\phi_{U'} = 1,Y = 1\}}{\Pr\{\phi_{U'} = 1\}} = \frac{\Pr\{\phi_{U'} = 1,Y = 1\}}{\Pr\{\phi_{U'} = 1,Y = 1\} + \Pr\{\phi_{U'} = 1,Y = 0\}}$$

Comme $\Pr\{\phi_U = 0,Y = 0\} = \Pr\{\phi_{U'} = 0,Y = 0\}$, on sait que $\Pr\{\phi_U = 1,Y = 0\} = \Pr\{\phi_{U'} = 1,Y = 0\}$

et en plus si a, b, c sont des réels positifs et $a \geq c$ on a $\frac{a}{a+b} \geq \frac{c}{c+b}$. On peut déduire de ces deux conditions que

$$\frac{\Pr\{\phi_U = 1,Y = 1\}}{\Pr\{\phi_U = 1\}} \geq \frac{\Pr\{\phi_{U'} = 1,Y = 1\}}{\Pr\{\phi_{U'} = 1\}} \qquad (1)$$

On obtient donc

$$VPP(U,Y) \geq VPP(U',Y)$$

2) Montrons que $VPN(U,Y) \geq VPN(U',Y)$
Par définition

$$\frac{\Pr\{\phi_U = 0,Y = 0\}}{\Pr\{\phi_U = 0\}} = \frac{\Pr\{\phi_U = 0,Y = 0\}}{\Pr\{\phi_U = 0,Y = 1\} + \Pr\{\phi_U = 0,Y = 0\}}$$

et

$$\frac{\Pr\{\phi_{U'} = 0,Y = 0\}}{\Pr\{\phi_{U'} = 0\}} = \frac{\Pr\{\phi_{U'} = 0,Y = 0\}}{\Pr\{\phi_{U'} = 0,Y = 1\} + \Pr\{\phi_{U'} = 0,Y = 0\}}$$

Puisque U' est emboîté dans U alors $\Pr\{\phi_{U'} = 0,Y = 1\} \geq \Pr\{\phi_U = 0,Y = 1\}$.
d'où

$$\frac{\Pr\{\phi_U = 0,Y = 0\}}{\Pr\{\phi_U = 0\}} \geq \frac{\Pr\{\phi_{U'} = 0,Y = 0\}}{\Pr\{\phi_{U'} = 0\}}$$

puisque $\Pr\{\phi_{U'} = 0,Y = 0\} = \Pr\{\phi_U = 0,Y = 0\}$ donc

$$VPN(U,Y) \geq VPN(U',Y)$$

3) Montrons que $RVP(U,Y) \geq RVP(U',Y)$
Par définition on a

$$RVP(U,Y) = \frac{1 - \Pr\{Y = 1\}}{\Pr\{Y = 1\}} \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{Y = 0\} - \Pr\{\phi_U = 0, Y = 0\}}$$

et

$$RVP(U',Y) = \frac{1 - \Pr\{Y = 1\}}{\Pr\{Y = 1\}} \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{Y = 0\} - \Pr\{\phi_{U'} = 0, Y = 0\}}$$

par hypothèse on $\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$ donc le signe de $RVP(U,Y) - RVP(U',Y)$ dépend du signe $\Pr\{\phi_{U1} = 1, Y = 1\} - \Pr\{\phi_{U'} = 1, Y = 1\}$
or on a le profil U' emboîte dans le profil U. Ceci entraîne que

$$\Pr\{\phi_{U1} = 1, Y = 1\} - \Pr\{\phi_{U'} = 1, Y = 1\} \geq 0$$

d'où

$$RVP(U,Y) \geq RVP(U',Y)$$

4) Montrons que $RVN(U,Y) \leq RVN(U',Y)$
Par définition

$$RVN(U,Y) = \frac{1 - \Pr\{Y = 1\}}{\Pr\{Y = 1\}} \frac{\Pr\{Y = 1\} - \Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 0, Y = 0\}}$$

et

$$RVN(U',Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\}-\Pr\{\phi_{U'}=1,Y=1\}}{\Pr\{\phi_{U'}=0,Y=0\}}$$

par hypothèse on $\Pr\{\phi_U=0,Y=0\} = \Pr\{\phi_{U'}=0,Y=0\}$ donc le signe de $RVN(U,Y) - RVN(U',Y)$ dépend du signe $\Pr\{\phi_{U'}=1,Y=1\} - \Pr\{\phi_U=1,Y=1\}$
or on a le profil U' emboîte dans le profil U. Ceci entraîne que

$$\Pr\{\phi_{U'}=1,Y=1\} - \Pr\{\phi_U=1,Y=1\} \leq 0$$

d'où

$$RVN(U,Y) \leq RVP(U,Y)$$

5) Montrons que $Err(U,Y) \leq Err(U',Y)$
On a

$$\Pr\{\phi_U=0,Y=0\} = 1 - \Pr\{Y=1\} - \Pr\{\phi_U=1\} + \Pr\{\phi_U=1,Y=1\}$$

$$\Pr\{\phi_U=1,Y=1\} = \Pr\{\phi_U=0,Y=0\} + \Pr\{\phi_U=1\} + \Pr\{Y=1\} - 1$$

si on remplace $\Pr\{\phi_U=1,Y=1\}$ par son expression dans $Err(U,Y)$, on obtient

$$Err(U,Y) = -2\Pr\{\phi_U=0,Y=0\} - \Pr\{\phi_U=1\} - \Pr\{Y=1\} - 2$$

de même on a

$$Err(U',Y) = -2\Pr\{\phi_{U'}=0,Y=0\} - \Pr\{\phi_{U'}=1\} - \Pr\{Y=1\} - 2$$

et puisque on a par hypothèse que $\Pr\{\phi_U=0,Y=0\} = \Pr\{\phi_{U'}=0,Y=0\}$ alors

$$Err(U,Y) - Err(U',Y) = -\Pr\{\phi_U=1\} + \Pr\{\phi_{U'}=1\}$$

par ailleurs $-\Pr\{\phi_U=1\} + \Pr\{\phi_{U'}=1\} \leq 0$ puisque U' est emboité dans U. d'où

$$Err(U,Y) - Err(U',Y) \leq 0$$

6) Montrons que $RR(U,Y) \geq RR(U',Y)$
On a

$$\frac{\Pr\{\phi_U=0,Y=1\}}{\Pr\{\phi_U=0\}} = \frac{\Pr\{\phi_U=0\} - \Pr\{\phi_U=0,Y=0\}}{\Pr\{\phi_U=0\}}$$

et

$$\frac{\Pr\{\phi_{U'}=0,Y=1\}}{\Pr\{\phi_{U'}=0\}} = \frac{\Pr\{\phi_{U'}=0\} - \Pr\{\phi_{U'}=0,Y=0\}}{\Pr\{\phi_{U'}=0\}}$$

en tenant compte que $\Pr\{\phi_U=0,Y=0\} = \Pr\{\phi_{U'}=0,Y=0\}$ et $\Pr\{\phi_U=0\} \leq \Pr\{\phi_{U'}=0\}$, on a

$$\frac{\Pr\{\phi_U=0,Y=0\}}{\Pr\{\phi_U=0\}} \geq \frac{\Pr\{\phi_{U'}=0,Y=0\}}{\Pr\{\phi_{U'}=0\}}$$

et il s'en suit que

$$\frac{\Pr\{\phi_U=0,Y=1\}}{\Pr\{\phi_U=0\}} \leq \frac{\Pr\{\phi_{U'}=0,Y=1\}}{\Pr\{\phi_{U'}=0\}}$$

d'où

$$\frac{\Pr\{\phi_U=0\}}{\Pr\{\phi_U=0,Y=1\}} \geq \frac{\Pr\{\phi_{U'}=0\}}{\Pr\{\phi_{U'}=0,Y=1\}} \qquad (2)$$

en faisant le produit membre à membre des inégalités (1) et (2) on obtient que $\mathrm{RR}(U',Y) \leq \mathrm{RR}(U,Y)$

□

Il résulte de la proposition 5 que si on a deux profils U et U' emboîtés tels que les fonctions de classification qui leurs sont associées ont des spécificités égales alors non seulement la sensibilité de la fonction de classification générée par U est plus élevée à cause de l'emboîtement mais aussi son erreur de classement est plus faible, sa valeur prédictive positive (confiance) est plus forte, son rapport de vraisemblance positif est plus élevé, son rapport de vraisemblance négatif est plus faible et son risque relatif est plus élevé que ceux de la fonction de classification générée par U'. On peut élaguer le profil U' qui est de plus grande taille. Cette proposition a été utilisée par Jiuyong Li et al [4] en premier en se basant sur la propriété anti-monotone du support.

3 Règles d'association binaires et classifieur associé à un profil

3.1 Règle d'association

Définition 4. Considérons $U = \left(m_h^{X_l}\right)_{l\in L}$ et $U' = \left(m_h^{X_j}\right)_{j\in J}$ deux profils disjoints. Une règle d'association est l'expression d'une implication de la forme $U \to U'$ signifiant que les probabilités $\Pr\left\{\left[\prod_{k\in L\cup J} \mathbb{1}\left(X_k = m_h^{X_k}\right) = 1\right]\right\}$ et $\Pr\left\{\left[\prod_{j\in J} \mathbb{1}\left(X_j = m_h^{X_j}\right) = 1\right] \mid \left[\prod_{l\in L} \mathbb{1}\left(X_l = m_h^{X_l}\right) = 1\right]\right\}$ sont significatives (supérieurs aux seuil s_0 et c_0 respectivement). On appelle U l'antécédent de la règle et U' la conséquence de la règle.

Une règle d'association $U \to U'$ exprime le fait que non seulement il y a une forte probabilité que les événements $\left[\prod_{j\in J} \mathbb{1}\left(X_j = m_h^{X_j}\right) = 1\right]$ et $\left[\prod_{l\in L} \mathbb{1}\left(X_l = m_h^{X_l}\right) = 1\right]$ aient lieu simultanément mais aussi que l'événement $\left[\prod_{j\in J} \mathbb{1}\left(X_j = m_h^{X_j}\right) = 1\right]$ ait une forte probabilité d'occurrence conditionnellement à l'événement $\left[\prod_{l\in L} \mathbb{1}\left(X_l = m_h^{X_l}\right) = 1\right]$.

Définition 5. Considérons une règle d'association $U \to U'$ où $U = \left(m_h^{X_l}\right)_{l\in L}$ et $U = \left(m_h^{X_j}\right)_{j\in J}$. La probabilité $\Pr\left\{\left[\prod_{k\in L\cup J} \mathbb{1}\left(X_k = m_h^{X_k}\right) = 1\right]\right\}$ est appelé le support de la règle d'association et la probabilité conditionnelle $\Pr\left\{\left[\prod_{j\in J} \mathbb{1}\left(X_j = m_h^{X_j}\right) = 1\right] \mid \left[\prod_{l\in L} \mathbb{1}\left(X_l = m_h^{X_l}\right) = 1\right]\right\}$ est sa confiance.

Il apparaît que le classifieur associé à un profil est une implication de la forme $[\phi(X,U) = 1] \to [Y = 1]$ dès lors qu'on exige que $\Pr(\phi(X,U) = 1, Y = 1) > s_0$ et $\Pr(Y = 1|\phi(X,U) = 1) > c_0$. Une telle règle d'association est dite binaire.

3.2 Classifieur basé sur un ensemble de profils

Dans un apprentissage statistique par règles d'association binaires, l'apprentissage automatique se résume en deux étapes. La première consiste à générer l'ensemble des profils $\mathcal{U}_\lambda$ défini par :

$$\mathcal{U}_\lambda = \left\{ U = \left(m_h^{X_j} \right)_{j \in J} ; \Pr(Y = 1, \phi(X, U) = 1) > s_0, \Pr(Y = 1 | \phi(X, U) = 1) > c_0 \right\}$$

où $\lambda = (s_0, c_0)$ est le paramètre qui spécifie l'ensemble $\mathcal{U}_\lambda$.
Le paramètre c_0 représente le seuil de confiance minimum et le paramètre s_0 représente le seuil de support minimum. Dans la pratique, on pourra étendre le paramètre λ en ajoutant le paramètre r_0 représentant le seuil de risque relatif minimum et le paramètre l_0 représentant la longueur ou taille maximale d'un profil.
La deuxième étape consiste à implémenter l'ensemble des fonctions indicatrices G_λ défini par :

$$G_\lambda = \{ \phi(X, U); U \in \mathcal{U}_\lambda \}$$

Lorsque la probabilité de la classe d'intérêt tend vers zero, la sensibilité du classifieur associé à un profil U (i.e., $\phi(X, U)$) peut être faible. En considérant un ensemble de profils, on peut espérer aboutir à un classifieur avec une meilleure sensibilité sans trop détériorer le niveau de spécificité. Etant donné un ensemble de profils G_λ pour un λ fixé, la fonction

$$\phi(X, \lambda, k) = \mathbb{1}\left(\sum_{U \in \mathcal{U}_\lambda} \phi(X, U) > k \right) \qquad k \in \{1, \cdots, |\mathcal{U}_\lambda|\}$$

définit également un classifieur.

4 Conclusion

L'objectif de cette analyse est de défendre une méthodologie permettant de mettre en place une fonction de classement binaire lorsqu'il s'agit d'une tâche de classification supervisée où la classe cible est un événement rare. Cet objectif est atteint par le recours à des règles d'association pour explorer les données afin d'identifier les profils qui sont corrélés avec la classe cible. Des profils pertinents sont sélectionnés sur la base de leurs sensibilités et spécificités, de leurs valeurs prédictives positives ou négatives, de leurs rapports de vraisemblance positifs ou négatifs et de leurs risques relatifs pour constituer un ensemble optimal de profils.
Dans la suite, nous allons mettre en place un algorithme d'apprentissage statistique pour établir une règle de classement (classifieur) basé sur un ensemble optimal de profils lorsque : (1) nous disposons d'un ensemble d'observations indépendantes et identiquement distribuées ; (2) les observations ne sont pas indépendantes et identiquement distribuées.

Bibliographie

[1] Agrawal, R., and Srikant, R. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases* (San Francisco, CA, USA, 1994), VLDB '94, Morgan Kaufmann Publishers Inc., pp. 487–499. 18

[2] Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett. 27*, 8 (2006), 861–874. 16

[3] Lenca, P., Meyer, P., Vaillant, B., and Lallich, S. On selecting interestingness measures for association rules : User oriented description and multiple criteria decision aid. *European Journal of Operational Research 184*, 2 (2008), 610–626. 18

[4] Li, J., Fu, A. W.-c., and Fahey, P. Efficient discovery of risk patterns in medical data. *Artificial intelligence in medicine 45*, 1 (2009), 77–89. 17, 18, 31

[5] Li, J., Fu, A. W.-c., He, H., Chen, J., Jin, H., McAullay, D., Williams, G., Sparks, R., and Kelman, C. Mining risk patterns in medical data. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (New York, NY, USA, 2005), KDD '05, ACM, pp. 770–775. 18

[6] Li, W., Han, J., and Pei, J. CMAR : accurate and efficient classification based on multiple class-association rules. In *ICDM 2001, Proceedings IEEE International Conference on Data Mining, 2001* (2001), pp. 369–376. 13

[7] Liu, B., Hsu, W., and Ma, Y. Integrating classification and association rule mining. pp. 80–86. 13

[8] Liu, B., Ma, Y., and Wong, C.-k. Classification using association rules : Weaknesses and enhancements. *In Grossman, R. L., et al (eds), Data Mining for Scientific and Engeneering Applications. Kluwer Academic Publishers* (2001), 591–601. 13

[9] Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., and Yamaguchi, T. Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In *Knowledge Discovery in Databases : PKDD 2004*, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds., no. 3202 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, pp. 362–373. 18

[10] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th International Conference on Database Theory* (1999), Springer-Verlag, pp. 398–416. 18

[11] Piatetsky-Shapiro, G. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*. pp. 229–248. 18

[12] Tan, P.-N., Kumar, V., and Srivastava, J. Selecting the right objective measure for association analysis. *Inf. Syst. 29*, 4 (2004), 293–313. 18

Appendices

Annexe A

Annexe Chapitre II

A Preuve de la proposition 1

Preuve. Par hypothèse U' est emboité dans U donc on a :

$$\Pr\{\phi_U = 1\} \geq \Pr\{\phi_{U'} = 1\} \quad \text{et} \quad \Pr\{\phi_U = 0\} \leq \Pr\{\phi_{U'} = 0\}$$

De ces deux inégalités, on déduit que :

$$\Pr\{\phi_U = 1\} \geq \Pr\{\phi_{U'} = 1\} \Rightarrow \Pr\{\phi_U = 1, Y = 1\} \geq \Pr\{\phi_{U'} = 1, Y = 1\}$$

$$\Pr\{\phi_U = 0\} \leq \Pr\{\phi_{U'} = 0\} \Rightarrow \Pr\{\phi_U = 0, Y = 0\} \leq \Pr\{\phi_{U'} = 0, Y = 0\}$$

□

B Preuve de la proposition 3

Preuve. Supposons que $\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$. On a

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_U = 1, Y = 1\} + \Pr\{\phi_U = 1, Y = 0\}$$
$$\Pr\{\phi_{U'} = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\} + \Pr\{\phi_{U'} = 1, Y = 0\}$$

On obtient

$$\Pr\{\phi_U = 1, Y = 1\} + \Pr\{\phi_U = 1, Y = 0\} = \Pr\{\phi_{U'} = 1, Y = 1\} + \Pr\{\phi_{U'} = 1, Y = 0\}$$

$$\Pr\{\phi_U = 1, Y = 1\} - \Pr\{\phi_{U'} = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 0\} - \Pr\{\phi_U = 1, Y = 0\}$$

Puisque $[\phi_{U'} = 1] \subset [\phi_U = 1]$ alors

$$\Pr\{\phi_U = 1, Y = 1\} - \Pr\{\phi_{U'} = 1, Y = 1\} \geq 0 \qquad (a)$$

$$\Pr\{\phi_U = 1, Y = 0\} - \Pr\{\phi_{U'} = 1, Y = 0\} \geq 0 \qquad (b)$$

On peut déduire de (a) et (b) les égalités suivantes :

$$\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$$

$$\Pr\{\phi_{U'} = 1, Y = 0\} = \Pr\{\phi_U = 1, Y = 0\} \Leftrightarrow \Pr\{\phi_{U'} = 0, Y = 0\} = \Pr\{\phi_U = 0, Y = 0\}$$

Supposons maintenant que les égalités suivantes soient vraies :

$$\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$$

$$\Pr\{\phi_{U'} = 0, Y = 0\} = \Pr\{\phi_U = 0, Y = 0\}$$

Puisque $\Pr\{\phi_{U'} = 0, Y = 0\} = \Pr\{\phi_U = 0, Y = 0\} \Leftrightarrow \Pr\{\phi_U = 1, Y = 0\} = \Pr\{\phi_{U'} = 1, Y = 0\}$, on a alors les égalités suivantes

$$\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$$

$$\Pr\{\phi_U = 1, Y = 0\} = \Pr\{\phi_{U'} = 1, Y = 0\}$$

En faisant les sommes membres à membres des deux égalités on obtient :

$$\Pr\{\phi_U = 1, Y = 1\} + \Pr\{\phi_U = 1, Y = 0\} = \Pr\{\phi_U = 1\}$$

$$\Pr\{\phi_{U'} = 1, Y = 1\} + \Pr\{\phi_{U'} = 1, Y = 0\} = \Pr\{\phi_{U'} = 1\}$$

D'où

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\}$$

□

C Preuve du Corollaire 1

Preuve. 1) Montrons que $VPP(U, Y) = VPP(U', Y)$
Par définition on a :

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\} \quad \Rightarrow \quad \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 1\}} = \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_{U'} = 1\}}$$

Et d'après la proposition 2 on a

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\} \Rightarrow \Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$$

Donc on a

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\} \quad \Rightarrow \quad \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 1\}} = \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 1\}}$$

2) Montrons que $VPN(U, Y) = VPN(U', Y)$

On a $\Pr\{\phi_U = 1\} = 1 - \Pr\{\phi_U = 0\}$ et $\Pr\{\phi_{U'} = 1\} = 1 - \Pr\{\phi_{U'} = 0\}$ Donc

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\} \quad \Rightarrow \quad \Pr\{\phi_U = 0\} = \Pr\{\phi_{U'} = 0\}$$

$$\Rightarrow \quad \frac{\Pr\{\phi_U = 0, Y = 0\}}{\Pr\{\phi_U = 0\}} = \frac{\Pr\{\phi_U = 0, Y = 0\}}{\Pr\{\phi_{U'} = 0\}}$$

Et d'après la proposition 2 on a

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\} \Rightarrow \Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$$

On en déduit que

$$\Pr\{\phi_U = 1\} = \Pr\{\phi_{U'} = 1\} \quad \Rightarrow \quad \frac{\Pr\{\phi_U = 0, Y = 0\}}{\Pr\{\phi_U = 0\}} = \frac{\Pr\{\phi_{U'} = 0, Y = 0\}}{\Pr\{\phi_{U'} = 0\}}$$

3) Montrons que $RVP(U, Y) = RVP(U', Y)$

Par définition on a

$$RVP(U, Y) = \frac{1 - \Pr\{Y = 1\}}{\Pr\{Y = 1\}} \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{Y = 0\} - \Pr\{\phi_U = 0, Y = 0\}}$$

Et d'après la proposition 2 on a

$$RVP(U, Y) = \frac{1 - \Pr\{Y = 1\}}{\Pr\{Y = 1\}} \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{Y = 0\} - \Pr\{\phi_{U'} = 0, Y = 0\}}$$

D'où

$$RVP(U, Y) = RVP(U', Y)$$

4) Montrons que $RVN(U, Y) = RVN(U', Y)$

Par définition on a

$$RVN(U,Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\} - \Pr\{\phi_U=1, Y=1\}}{\Pr\{\phi_U=0, Y=0\}}$$

Et d'après la proposition 2, on a

$$RVN(U,Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\} - \Pr\{\phi_{U'}=1, Y=1\}}{\Pr\{\phi_{U'}=0, Y=0\}}$$

D'où

$$RVN(U,Y) = RVN(U',Y)$$

5) Montrons que $Err(U,Y) = Err(U',Y)$

Par définition on a

$$Err(U,Y) = \Pr\{Y=1\} + \Pr\{\phi_U=1\} - 2\Pr\{\phi_U=1, Y=1\}$$

$$Err(U',Y) = \Pr\{Y=1\} + \Pr\{\phi_{U'}=1\} - 2\Pr\{\phi_{U'}=1, Y=1\}$$

Si $\Pr\{\phi_U=1\} = \Pr\{\phi_{U'}=1\}$ alors $\Pr\{\phi_U=1, Y=1\} = \Pr\{\phi_{U'}=1, Y=1\}$ (proposition 2)

Il en résulte des égalités précédentes que

$$Err(U,Y) = Err(U',Y)$$

6) Montrons que $RR(U,Y) = RR(U',Y)$

On a par hypothèse

$$\begin{aligned} \Pr\{\phi_U=1\} = \Pr\{\phi_{U'}=1\} &\Leftrightarrow 1-\Pr\{\phi_U=1\} = 1-\Pr\{\phi_{U'}=1\} \\ &\Leftrightarrow \Pr\{\phi_U=0\} = \Pr\{\phi_{U'}=0\} \end{aligned}$$

alors

$$\frac{\Pr\{\phi_U=0\}}{\Pr\{\phi_U=1\}} = \frac{\Pr\{\phi_{U'}=0\}}{\Pr\{\phi_{U'}=1\}} \qquad (1)$$

D'aprés la proposition 1, si on a $\Pr\{\phi_U=1\} = \Pr\{\phi_{U'}=1\}$ alors

(a) $\Pr\{\phi_U=1, Y=1\} = \Pr\{\phi_{U'}=1, Y=1\}$

(b) $\Pr\{\phi_U=0, Y=0\} = \Pr\{\phi_{U'}=0, Y=0\}$

Puisque $\Pr\{\phi_U = 0\} = \Pr\{\phi_{U'} = 0\}$ alors

$$\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\} \quad \Leftrightarrow \quad \Pr\{\phi_U = 0, Y = 1\} = \Pr\{\phi_{U'} = 0, Y = 1\}$$

On en déduit que

$$\frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 0, Y = 1\}} = \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 0, Y = 1\}} \qquad (2)$$

En faisant le produit les produit membre à membre des égalités (1) et (2) on obtient

$$\frac{\Pr\{\phi_U = 0\}}{\Pr\{\phi_U = 1\}} \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 0, Y = 1\}} = \frac{\Pr\{\phi_{U'} = 0\}}{\Pr\{\phi_{U'} = 1\}} \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 0, Y = 1\}}$$

Il en résulte que

$$RR(U, Y) = RR(U', Y)$$

□

D Preuve de la proposition 4

Preuve. 1) Montrons que $VPP(U, Y) \leq VPP(U', Y)$

Par hypothèse on a :

$$\frac{1}{\Pr\{\phi_U = 1\}} \leq \frac{1}{\Pr\{\phi_{U'} = 1\}}$$

Si l'égalité $\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_{U'} = 1, Y = 1\}$ est vérifiée alors

$$\begin{aligned} \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 1\}} &= \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_U = 1\}} \\ &\leq \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 1\}} \qquad (1) \end{aligned}$$

On obtient donc

$$VPP(U, Y) \leq VPP(U', Y)$$

2) Montrons que $VPN(U, Y) \leq VPN(U', Y)$

Par hypothèse on a :

$$\frac{1}{\Pr\{\phi_{U'} = 0\}} \leq \frac{1}{\Pr\{\phi_U = 0\}}$$

par ailleurs on a

$$\Pr\{\phi_U = 1,Y = 1\} = \Pr\{\phi_{U'} = 1,Y = 1\} \Rightarrow \Pr\{\phi_U = 0,Y = 1\} = \Pr\{\phi_{U'} = 0,Y = 1\}$$

$$\Rightarrow \Pr\{\phi_U = 0\} - \Pr\{\phi_U = 0,Y = 0\} = \Pr\{\phi_{U'} = 0\} - \Pr\{\phi_{U'} = 0,Y = 0\}$$

donc

$$\begin{aligned} 1 - \frac{\Pr\{\phi_U = 0,Y = 0\}}{\Pr\{\phi_U = 0\}} &\geq 1 - \frac{\Pr\{\phi_{U'} = 0,Y = 0\}}{\Pr\{\phi_{U'} = 0\}} \\ \frac{\Pr\{\phi_U = 0,Y = 0\}}{\Pr\{\phi_U = 0\}} &\leq \frac{\Pr\{\phi_{U'} = 0,Y = 0\}}{\Pr\{\phi_{U'} = 0\}} \end{aligned}$$

On obtient donc

$$VPN(U,Y) \leq VPV(U',Y)$$

3) Montrons que $RVP(U,Y) = RVP(U',Y)$

Par définition on a

$$\begin{aligned} RVP(U,Y) &= \frac{1 - \Pr\{Y = 1\}}{\Pr\{Y = 1\}} \frac{\Pr\{\phi_U = 1,Y = 1\}}{\Pr\{Y = 0\} - \Pr\{\phi_U = 0,Y = 0\}} \\ &= \frac{1 - \Pr\{Y = 1\}}{\Pr\{Y = 1\}} \frac{\Pr\{\phi_U = 1,Y = 1\}}{\Pr\{\phi_U = 1,Y = 0\}} \\ &= \frac{1 - \Pr\{Y = 1\}}{\Pr\{Y = 1\}} \frac{\Pr\{\phi_U = 1,Y = 1\}}{\Pr\{\phi_U = 1\} - \Pr\{\phi_U = 1,Y = 1\}} \end{aligned}$$

donc si $\Pr\{\phi_U = 1,Y = 1\} = \Pr\{\phi_{U'} = 1,Y = 1\}$ et U' emboîté dans U alors

$$\Pr\{\phi_U = 1\} - \Pr\{\phi_U = 1,Y = 1\} \geq \Pr\{\phi_{U'} = 1\} - \Pr\{\phi_{U'} = 1,Y = 1\}$$

donc

$$RVP(U,Y) \leq \frac{1 - \Pr\{Y = 1\}}{\Pr\{Y = 1\}} \frac{\Pr\{\phi_{U'} = 1,Y = 1\}}{\Pr\{\phi_{U'} = 1\} - \Pr\{\phi_{U'} = 1,Y = 1\}}$$

d'où

$$RVP(U,Y) \leq RVP(U',Y)$$

4) Montrons que $RVN(U,Y) \geq RVN(U',Y)$

Par définition on a

$$RVN(U,Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\}-\Pr\{\phi_U=1,Y=1\}}{\Pr\{\phi_U=0,Y=0\}}$$

par hypothèse $\Pr\{\phi_U=1,Y=1\} = \Pr\{\phi_{U'}=1,Y=1\}$ on a alors

$$RVN(U,Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\}-\Pr\{\phi_{U'}=1,Y=1\}}{\Pr\{\phi_U=0,Y=0\}}$$

par ailleurs U' emboîté dans U entraîne que $\Pr\{\phi_U=0,Y=0\} \leq \Pr\{\phi_{U'}=0,Y=0\}$. On en déduit que

$$RVN(U,Y) \geq \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\}-\Pr\{\phi_{U'}=1,Y=1\}}{\Pr\{\phi_{U'}=0,Y=0\}}$$

d'où

$$RVN(U,Y) \geq RVN(U',Y)$$

5) Montrons que $Err(U,Y) \geq Err(U',Y)$
Par définition on a

$$Err(U,Y) = \Pr\{Y=1\} + \Pr\{\phi_U=1\} - 2\Pr\{\phi_U=1,Y=1\}$$

$$Err(U',Y) = \Pr\{Y=1\} + \Pr\{\phi_{U'}=1\} - 2\Pr\{\phi_{U'}=1,Y=1\}$$

Par hypothèse on a $\Pr\{\phi_U=1,Y=1\} = \Pr\{\phi_{U'}=1,Y=1\}$ donc

$$Err(U,Y) - Err(U,Y) = \Pr\{\phi_U=1\} - \Pr\{\phi_{U'}=1\}$$

et puisque U' est emboité dans U alors $\Pr\{\phi_U=1\} \geq \Pr\{\phi_{U'}=1\}$. On obtient donc

$$Err(U,Y) \geq Err(U'Y)$$

6) Montrons que $RR(U,Y) \leq RR(U',Y)$
L'égalité $\Pr\{\phi_U=1,Y=1\} = \Pr\{\phi_{U'}=1,Y=1\}$ implique aussi

$$\begin{aligned}
\Pr\{\phi_U=0,Y=1\} &= \Pr\{Y=1\} - \Pr\{\phi_U=1,Y=1\} \\
&= \Pr\{Y=1\} - \Pr\{\phi_{U'}=1,Y=1\} \\
&= \Pr\{\phi_{U'}=0,Y=1\}
\end{aligned}$$

d'où

$$\begin{aligned} \frac{\Pr\{\phi_U = 0, Y = 1\}}{\Pr\{\phi_U = 0\}} &= \frac{\Pr\{\phi_{U'} = 0, Y = 1\}}{\Pr\{\phi_U = 0\}} \\ &\geq \frac{\Pr\{\phi_{U'} = 0, Y = 1\}}{\Pr\{\phi_{U'} = 0\}} \end{aligned} \qquad (2)$$

Si on fait le rapport membre à membre des inégalités (1) et (2), il en résulte que

$$RR(U, Y) \leq RR(U', Y)$$

□

E Preuve de la proposition 5

Preuve. 1) Montrons que $VPP(U, Y) \geq VPP(U', Y)$

Par définition

$$\frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 1\}} = \frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 1, Y = 1\} + \Pr\{\phi_U = 1, Y = 0\}}$$

et

$$\frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 1\}} = \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 1, Y = 1\} + \Pr\{\phi_{U'} = 1, Y = 0\}}$$

On sait que $\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\} \Leftrightarrow \Pr\{\phi_U = 1, Y = 0\} = \Pr\{\phi_{U'} = 1, Y = 0\}$

et en plus si a, b, c sont des réels positifs et $a \geq c$ on a $\frac{a}{a+b} \geq \frac{c}{c+b}$. On peut déduire de ces deux conditions que

$$\frac{\Pr\{\phi_U = 1, Y = 1\}}{\Pr\{\phi_U = 1\}} \geq \frac{\Pr\{\phi_{U'} = 1, Y = 1\}}{\Pr\{\phi_{U'} = 1\}} \qquad (1)$$

On obtient donc

$$VPP(U, Y) \geq VPP(U', Y)$$

2) Montrons que $VPN(U, Y) \geq VPN(U', Y)$

Par définition

$$\frac{\Pr\{\phi_U = 0, Y = 0\}}{\Pr\{\phi_U = 0\}} = \frac{\Pr\{\phi_U = 0, Y = 0\}}{\Pr\{\phi_U = 0, Y = 1\} + \Pr\{\phi_U = 0, Y = 0\}}$$

et

$$\frac{\Pr\{\phi_{U'} = 0, Y = 0\}}{\Pr\{\phi_{U'} = 0\}} = \frac{\Pr\{\phi_{U'} = 0, Y = 0\}}{\Pr\{\phi_{U'} = 0, Y = 1\} + \Pr\{\phi_{U'} = 0, Y = 0\}}$$

Puisque U' est emboîté dans U alors $\Pr\{\phi_{U'}=0,Y=1\} \geq \Pr\{\phi_U=0,Y=1\}$.
d'où

$$\frac{\Pr\{\phi_U=0,Y=0\}}{\Pr\{\phi_U=0\}} \geq \frac{\Pr\{\phi_{U'}=0,Y=0\}}{\Pr\{\phi_{U'}=0\}}$$

puisque $\Pr\{\phi_{U'}=0,Y=0\} = \Pr\{\phi_U=0,Y=0\}$ donc

$$VPN(U,Y) \geq VPN(U',Y)$$

3) Montrons que $RVP(U,Y) \geq RVP(U',Y)$
Par définition on a

$$RVP(U,Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{\phi_U=1,Y=1\}}{\Pr\{Y=0\}-\Pr\{\phi_U=0,Y=0\}}$$

et

$$RVP(U',Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{\phi_{U'}=1,Y=1\}}{\Pr\{Y=0\}-\Pr\{\phi_{U'}=0,Y=0\}}$$

par hypothèse on $\Pr\{\phi_U=0,Y=0\} = \Pr\{\phi_{U'}=0,Y=0\}$ donc le signe de $RVP(U,Y)-RVP(U',Y)$ dépend du signe $\Pr\{\phi_{U1}=1,Y=1\} - \Pr\{\phi_{U'}=1,Y=1\}$
or on a le profil U' emboîte dans le profil U. Ceci entraîne que

$$\Pr\{\phi_{U1}=1,Y=1\} - \Pr\{\phi_{U'}=1,Y=1\} \geq 0$$

d'où

$$RVP(U,Y) \geq RVP(U',Y)$$

4) Montrons que $RVN(U,Y) \leq RVN(U',Y)$
Par définition

$$RVN(U,Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\}-\Pr\{\phi_U=1,Y=1\}}{\Pr\{\phi_U=0,Y=0\}}$$

et

$$RVN(U',Y) = \frac{1-\Pr\{Y=1\}}{\Pr\{Y=1\}} \frac{\Pr\{Y=1\}-\Pr\{\phi_{U'}=1,Y=1\}}{\Pr\{\phi_{U'}=0,Y=0\}}$$

par hypothèse on $\Pr\{\phi_U=0,Y=0\} = \Pr\{\phi_{U'}=0,Y=0\}$ donc le signe de $RVN(U,Y)-RVN(U',Y)$ dépend du signe $\Pr\{\phi_{U'}=1,Y=1\} - \Pr\{\phi_U=1,Y=1\}$
or on a le profil U' emboîte dans le profil U. Ceci entraîne que

$$\Pr\{\phi_{U'}=1,Y=1\} - \Pr\{\phi_U=1,Y=1\} \leq 0$$

d'où

$$RVN(U,Y) \leq RVP(U,Y)$$

5) Montrons que $Err(U,Y) \leq Err(U',Y)$

On a

$$\Pr\{\phi_U = 0, Y = 0\} = 1 - \Pr\{Y = 1\} - \Pr\{\phi_U = 1\} + \Pr\{\phi_U = 1, Y = 1\}$$

$$\Pr\{\phi_U = 1, Y = 1\} = \Pr\{\phi_U = 0, Y = 0\} + \Pr\{\phi_U = 1\} + \Pr\{Y = 1\} - 1$$

si on remplace $\Pr\{\phi_U = 1, Y = 1\}$ par son expression dans $Err(U,Y)$, on obtient

$$Err(U,Y) = -2\Pr\{\phi_U = 0, Y = 0\} - \Pr\{\phi_U = 1\} - \Pr\{Y = 1\} - 2$$

de même on a

$$Err(U',Y) = -2\Pr\{\phi_{U'} = 0, Y = 0\} - \Pr\{\phi_{U'} = 1\} - \Pr\{Y = 1\} - 2$$

et puisque on a par hypothèse que $\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$ alors

$$Err(U,Y) - Err(U',Y) = -\Pr\{\phi_U = 1\} + \Pr\{\phi_{U'} = 1\}$$

par ailleurs $-\Pr\{\phi_U = 1\} + \Pr\{\phi_{U'} = 1\} \leq 0$ puisque U' est emboité dans U. d'où

$$Err(U,Y) - Err(U',Y) \leq 0$$

6) Montrons que $RR(U,Y) \geq RR(U',Y)$

On a

$$\frac{\Pr\{\phi_U = 0, Y = 1\}}{\Pr\{\phi_U = 0\}} = \frac{\Pr\{\phi_U = 0\} - \Pr\{\phi_U = 0, Y = 0\}}{\Pr\{\phi_U = 0\}}$$

et

$$\frac{\Pr\{\phi_{U'} = 0, Y = 1\}}{\Pr\{\phi_{U'} = 0\}} = \frac{\Pr\{\phi_{U'} = 0\} - \Pr\{\phi_{U'} = 0, Y = 0\}}{\Pr\{\phi_{U'} = 0\}}$$

en tenant compte que $\Pr\{\phi_U = 0, Y = 0\} = \Pr\{\phi_{U'} = 0, Y = 0\}$ et $\Pr\{\phi_U = 0\} \leq \Pr\{\phi_{U'} = 0\}$, on a

$$\frac{\Pr\{\phi_U = 0, Y = 0\}}{\Pr\{\phi_U = 0\}} \geq \frac{\Pr\{\phi_{U'} = 0, Y = 0\}}{\Pr\{\phi_{U'} = 0\}}$$

et il s'en suit que

$$\frac{\Pr\{\phi_U = 0, Y = 1\}}{\Pr\{\phi_U = 0\}} \leq \frac{\Pr\{\phi_{U'} = 0, Y = 1\}}{\Pr\{\phi_{U'} = 0\}}$$

d'où

$$\frac{\Pr\{\phi_U = 0\}}{\Pr\{\phi_U = 0, Y = 1\}} \geq \frac{\Pr\{\phi_{U'} = 0\}}{\Pr\{\phi_{U'} = 0, Y = 1\}} \qquad (2)$$

en faisant le produit membre à membre des inégalités (1) et (2) on obtient que $\mathrm{RR}(U',Y) \leq \mathrm{RR}(U,Y)$

□

F Preuve de la proposition 2

Preuve. On a

$$\begin{aligned}
\frac{\Pr(Y=1,\phi(X,U)=1)}{\Pr(\phi(X,U)=1)} &= \frac{\Pr(Y=1,\phi(X,U)=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} + \frac{\Pr(Y=1,\phi(X,U)=1,\phi(X,U)=0)}{\Pr(\phi(X,U)=1)} \\
&= \frac{\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} + \\
&\quad \frac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U)=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} \\
&= \frac{\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U')=1)}\frac{\Pr(\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} + \\
&\quad \frac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} \\
&= \frac{\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U')=1)}\frac{\Pr(\phi(X,U')=1)}{\Pr(\phi(X,U)=1)} + \\
&\quad \frac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)-\Pr(\phi(X,U)=1)}\left[1-\frac{\Pr(\phi(X,U')=1)}{\Pr(\phi(X,U)=1)}\right]
\end{aligned}$$

On obtient une combinaison convexe de $\frac{\Pr(Y=1,\phi(X,U)=1)}{\Pr(\phi(X,U)=1)}$ par rapport à $\frac{\Pr(\phi(X,U')=1)}{\Pr(\phi(X,U)=1)}$. On en déduit que $\frac{\Pr(Y=1,\phi(X,U)=1)}{\Pr(\phi(X,U)=1)}$ est compris entre

$$Min\left\{\frac{\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U')=1)}, \frac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)-\Pr(\phi(X,U)=1)}\right\}$$

et

$$Max\left\{\frac{\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U')=1)}, \frac{\Pr(Y=1,\phi(X,U)=1)-\Pr(Y=1,\phi(X,U')=1)}{\Pr(\phi(X,U)=1)-\Pr(\phi(X,U)=1)}\right\}$$

□

Chapitre III

Classifieur basé sur un ensemble de profils lorsque les données sont indépendantes et identiquement distribuées

1 Introduction

Dans cette thèse, nous proposons une méthode de classement basée sur les règles d'association binaire dans le but d'améliorer les performances d'une règle de classement lorsque la classe cible de la variable réponse binaire est faiblement représentée. Généralement dans une telle situation, la règle de classement a une forte spécificité. Donc pour améliorer les performances de la règle de classement, nous nous intéressons plus aux profils de classement dont les classifieurs associés ont des sensibilités fortes.

A travers les indices de performances présentés au chapitre II, on peut affirmer que l'apprentissage du classifieur associé à un profil est fortement dépendant de la valeur prédictive positive (VPP). Généralement on estime ce dernier par le maximum de vraisemblance. Mais dans une situation où le support (la couverture) du profil est trop faible, il est recommandé d'estimer la VPP par une forme corrigée de Laplace [11] définie par

$$VPP(U,Y) = \frac{\Pr\{\phi(X,U)=1, Y=1\} + 1}{\Pr\{\phi(X,U)=1, Y=1\} + \Pr\{\phi(X,U)=1, Y=0\} + |\mathrm{Dom}(Y)|}$$

Dans la suite, nous verrons qu'il est possible d'avoir une interprétation Bayésienne de la formule de Laplace.

Soit $\mathcal{D}_n = (y_i, x_i)$ un ensemble fini d'éléments générés de façon aléatoire par la loi du couple (Y, X), où Y est une variable binaire et $X = (X_j)_{j=1:p}$ est un vecteur de variables aléatoires, où la variable X_j peut être numérique ou catégorielle. A l'aide des outils statistiques présentés dans le chapitre II, nous présentons un algorithme d'apprentissage dont les performances sont comparables avec d'autres méthodes très connues pour un classement binaire.

2 Algorithme d'apprentissage d'un classifieur basé sur un ensemble de profils

Dès qu'un phénomène, qu'il soit physique, biologique ou autre, est trop complexe ou encore trop bruité pour accéder à une description analytique débouchant sur une modélisation déterministe, un ensemble d'approches est élaboré afin d'en décrire au mieux le comportement à partir d'une série d'observations. On appelle apprentissage statistique l'ensemble d'approches élaboré [5]. C'est une combinaison à la fois de l'apprentissage automatique et de la statistique [26]. L'apprentissage automatique consiste à utiliser des ordinateurs pour optimiser un modèle de traitement de l'information selon certains critères de performance à partir d'observations. Tandis que la statistique permet de formaliser le processus, de garantir sa qualité et éventuellement de suggérer de nouvelles techniques. Cependant le principe de l'apprentissage reste le même, mais la démarche est différente selon que la taille du jeu de données est grande ou petite.

2.1 Présentation de l'algorithme de construction du classifieur

Lorsque la taille des données est suffisamment grande, on adoptera l'approche Apprentissage/Validation/Test pour la sélection d'un ensemble optimal de profils. Cette approche consiste à subdiviser les données de manière aléatoire en trois ensembles : un ensemble d'apprentissage, un ensemble de validation et un ensemble test. L'apprentissage statistique que nous proposons peut être résumée par les différentes étapes suivantes :

1. Discrétiser toutes les variables numériques par une méthode de discrétisation (au choix)
2. A partir d'un ensemble d'apprentissage :
   (a) Spécifier le paramètre d'apprentissage $\lambda = (s_0, c_0, l_0)$
   (b) Générer un ensemble $\mathcal{U}_\lambda$ de profils
   (c) Elaguer les profils redondants dans $\mathcal{U}_\lambda$ pour constituer un petit ensemble

$$\mathcal{U}_\lambda^1 = \{[\phi(X, U) = 1] \rightarrow [Y = 1]; U \in \mathcal{U}_\lambda\}$$

3. A partir d'un ensemble de validation :
   (a) Réévaluer l' indicateur de performance VPP (ou RVP ou RVN) de toutes les règles dans $\mathcal{U}_\lambda^1$
   (b) Supprimer les profils dont le RVP est inférieur à un (1)
   (c) Parmi les profils dans $\mathcal{U}_\lambda^1$ qui sont emboîtés, ne retenir que le profil dont le VPP (ou le RVP ou le RVN) est le plus significatif.

4. Au sortir de l'étape 3, on dispose alors d'un ensemble de profils $\mathcal{U}_\lambda^2$ tel que $|\mathcal{U}_\lambda^2| \leq |\mathcal{U}_\lambda^1|$.
5. Définir la règle de classement (classifieur) ϕ d'une observation X par

$$\phi(X,\lambda) = \begin{cases} 1 & \text{si} \quad \sum_{m=1}^{|\mathcal{U}_\lambda^2|} \phi(X, U_m) > 0 \\ 0 & \text{sinon} \end{cases}$$

Le classifieur $\phi(X,\lambda)$ est un cas particulier du classifieur défini au chapitre II à la section 3.2 où on a choisi k égal à zéro. On choisit alors de classer positive une observation X lorsqu'elle vérifie au moins un profil parmi ceux qui sont dans l'ensemble $\mathcal{U}_\lambda^2$.

Dans tout ce qui suit, on fixe à un le nombre minimum de profils à vérifier pour qu'une observation soit classée positive.

3 Prétraitement des données : discrétisation des covariables numériques

Un ensemble de données pour un classement est normalement sous la forme d'un tableau de données qui est décrit par un ensemble de variables distinctes. La plupart des applications réelles (données réelles) pour une classification supervisée comportent à la fois des variables numériques (continues) et des variables nominales (catégorielles). Certaines méthodes de classement, particulièrement l'algorithme d'apprentissage des règles d'association, exigent que toutes les covariables soient nominales. Ainsi il est nécessaire de convertir les variables continues en des variables discrètes. L'idée consiste à transformer chaque variable numérique X_j en une variable catégorielle X_j^*. La variable X_j^* est obtenue en subdivisant le domaine des valeurs de X_j en q_j intervalles $m_h^{X_j}, h = 1 : q_j$. La variable X_j^* sera utilisée à la place de X_j pour construire le classifieur.

En général une variable continue est une variable dont le domaine de définition est totalement ordonné. La discrétisation doit être choisie de manière à apporter des informations de classification utiles sans modifier les classes auxquelles les observations du domaine de la variable appartiennent. En général, une discrétisation est simplement une condition logique, en termes d'une ou plusieurs valeurs évaluées, qui sert à partitionner les données en au moins deux sous-ensembles. Supposons que X_j soit une variable numérique et l'intervalle $[a,b]$ soit son domaine. Une partition π_{X_j} sur $[a,b]$ est définie comme le sous-ensemble des k intervalles suivants

$$\pi_{X_j} = \{[x_{j0}, x_{j1}), [x_{j1}, x_{j2}), \ldots, [x_{j(k-1)}, x_{jk}]\}$$

où $x_{j0} = a$, $x_{j(i-1)} < x_{ji}$ pour $i = 1 : k$ et $x_{jk} = b$. Ainsi la discrétisation est le processus qui produit

une partition π_{X_j} sur $[a, b]$.

Plusieurs méthodes de discrétisation des variables numériques ont été étudiées dans la littérature. On peut, par exemple, considérer des combinaisons linéaires de plusieurs variables et comparer le résultat avec un seuil (Breiman et al., 1984)[7]. Il est aussi possible d'éviter le seuillage en formant une condition qui compare les valeurs de deux ou plusieurs variables directement. Cependant le nombre de telles expressions possibles rend l'espace de recherche très vaste.
La méthode de discrétisation d'une variable numérique la plus simple reste la méthode de largeur d'intervalle égale (Equal Interval Width Method). Elle consiste à partitionner son domaine en intervalles de largeur égales.

Une méthode de discrétisation de variable numérique par la discrétisation adaptative a été proposée dans [8]. La méthode consiste à diviser d'abord le domaine de la variable en deux intervalles de largeur égale et un processus d'apprentissage est lancé pour générer les règles. Ensuite, la qualité des règles est testée en évaluant les performances des règles. Si la mesure de performance est inférieure à un seuil fixe, l'un des intervalles est subdivisé en outre, et le processus est répété. Le principal inconvénient de cette méthode, cependant, est la répétition du processus d'apprentissage jusqu'à ce que le niveau de performance finale soit atteint.

Une discrétisation basée sur l'entropie marginale maximale a été introduite dans [30]. Ce procédé consiste à diviser le domaine de la variable numérique de telle sorte que la fréquence d'échantillonnage dans chaque intervalle soit approximativement égale. Ce procédé est généralement appelé la méthode par intervalle de fréquence égale (Equal Frequency per Interval Method). Le seul paramètre fourni par l'utilisateur est le nombre d'intervalles à induire sur le domaine d'origine. La discrétisation par la mesure de l'entropie utilise les bornes du domaine de la variable pour induire les intervalles souhaités. Cette méthode de sélection d'un point de coupure est utilisée dans l'algorithme ID3 [23], dans l'algorithme CART [6], et d'autres [15].

Lorsque nous traitons un problème de classification supervisée, il est naturel de penser à discrétiser les variables numériques en fonction de la variable réponse. Ceci constitue l'un des points faibles des différentes méthodes de discrétisation citées précédemment. Ce concept est pris en compte par la méthode de discrétisation avec la classe-entropie comme critère pour sélectionner le meilleur point de coupure [13]. Dans tout ce qui suit, nous avons utilisé la méthode de discrétisation dont le critère d'arrêt est basé sur le principe de la longueur de description minimum plus connu sous le nom de MDLP (Minimum Description Length Principle). Cette méthode est initiée par Fayyad et Irani [13, 14]. La méthode est présentée comme une méthode efficace pour la discrétisation pour l'apprentissage des arbres de décision et du classifieur de Bayes Naïf [2] (voir l'annexe B pour plus de détails).

4 Extraction d'un ensemble initial de profils

L'ensemble des profils $\mathcal{U}_\lambda$, généré au départ pour l'apprentissage du classifieur, est caractérisé par c_0, une estimation de la VPP, et s_0, une estimation du support. L'un des plus connus algorithmes d'exploration des règles d'association, utilisant c_0 et s_0 pour l'extraction des règles les plus fréquentes, reste l'algorithme *"apriori"*. Il est l'un des algorithmes d'extraction de règles d'association qui a utilisé en premier l'élagage basé sur le support pour contrôler systématiquement la croissance exponentielle des règles candidates. C'est la raison pour laquelle, nous avons choisi de l'utiliser pour la suite. On pouvait utiliser d'autres algorithmes d'extraction de règles fréquentes existant dans la littérature par exemple l'algorithme *"FP-Growth"*(*FPtree structure*) [17]. Un choix de l'algorithme d'extraction est laissé à l'utilisateur. Ci-après (Tableau III.1), nous présentons un pseudo code de la partie de génération des profils fréquents par l'algorithme *"apriori"*. Soit C_k l'ensemble des profils de longueur k candidats, $\mathcal{D}$ l'ensemble de toutes les observations et F_k l'ensemble des profils fréquents et de longueur k.

Algorithme : Génération de règles fréquentes par l'algorithme *"apriori"*

– Entrées : $\mathcal{D}$ un ensemble d'observations, s_0 un support minimum et c_0 une confiance minimum
– Sorties : $\mathcal{U}_\lambda$ un ensemble de profils fréquents

1 : k=1
2 : $F_k =$ {Trouver tous les 1-itemsets fréquents}
3 : **répéter**
4 : k=k+1
5 : C_k = apriori-gen(F_{k-1}). {Générer les profils candidats}
6 : **pour** chaque observation $t \in D$ **faire**
7 : C_t = subset(C_k , t). {Identifier tous les candidats contenus dans t}
8 : **pour** chaque profil candidat $c \in C_t$ **faire**
9 : $supp(c) = supp(c) + 1$. {Incrémenter le compte du support}
10 : **si** $t.class = c.class$ **faire** { $t.class$: la classe associée à l'observation t}
11 : $conf(c) = conf(c) + 1$. {Incrémenter le compte de la confiance}
12 : **fin si**
13 : **fin pour**
14 : **fin pour**
15 : F_k ={c $\in C_k$ | supp(c) $\geq s_0$; conf(c)/supp(c) $\geq c_0$ }
{Extraire les profils fréquents de taille k}
16 : **jusqu'à** $F_k = \emptyset$
17 : Retourner : $\mathcal{U}_\lambda = \bigcup_k F_k$

Tableau III.1 – Algorithme de génération des règles fréquentes (*"apriori"*)

Pour la suite, nous nous intéresserons aux profils générés à partir de l'algorithme *"apriori"* qui sont corrélés avec la variable réponse et qui vérifient les conditions d'apprentissages suivantes : support $\geq s_0$,

confiance $\geq c_0$, risque relatif $\geq r_0$, taille $\leq l_0$. Cette étape de l'algorithme est élaborée sur l'échantillon d'apprentissage. Au sortir de cette phase, un vaste ensemble $\mathcal{U}_\lambda$, $\lambda = (s_0, c_0, r_0, l_0)$ contenant à la fois des profils redondants et des profils de faibles performances, est généré. Il est donc nécessaire d'élaborer une procédure d'élagage des profils redondants pour réduire le vaste ensemble $\mathcal{U}_\lambda$ à un ensemble $\mathcal{U}_\lambda^1$ ne contenant que des profils fréquents et non redondants.

5 Elagage des profils redondants

Dans cette section, nous nous intéressons aux profils qui sont liés à la variable réponse. La suppression des profils qui ne sont pas corrélés à la variable réponse et des profils redondants permettra de sélectionner un ensemble réduit de profils dont on pourra se servir pour construire un classifieur performant.

Soient $U_1 = \left(m_h^{X_j}\right)_{j\in J}$ et $U_2 = \left(m_h^{X_l}\right)_{l\in L}$ deux profils tels que U_2 soit emboîté dans U_1. L'application des résultats théoriques précédents nécessite de faire un test d'hypothèse sur l'égalité des couvertures, sur l'égalité des supports ou sur l'égalité des spécificités de deux profils emboîtés. Pour cela, il est possible de faire un test stochastique

5.1 Test stochastique (randomisé) pour la sélection entre deux profils emboîtés

En principe, si l'égalité n'est pas vérifiée sur un échantillon donné, on peut affirmer qu'elle n'est pas vérifiée sur la population dont est issu l'échantillon. Par contre on ne peut pas en dire autant lorsqu'elle est vraie sur un échantillon. C'est la raison pour laquelle un test stochastique (ou test randomisé) est nécessaire.

On note par $\phi(X, U_1) = \prod_{j\in J} \mathbb{1}\left(X_j = m_h^{X_j}\right)$ et $\phi(X, U_2) = \prod_{l\in L} \mathbb{1}\left(X_l = m_h^{X_l}\right)$ les fonctions de classement générées respectivement par U_1 et U_2. Puisque U_2 est emboîté dans U_1, on a $[\phi(X, U_2) = 1] \subset [\phi(X, U_1) = 1]$.

(a) Soit le paramètre θ_1 défini par $\theta_1 = \Pr(\phi(X, U_1) = 1) - \Pr(\phi(X, U_2) = 1)$. Nous voulons tester si oui ou non θ_1 est nulle i.e décider entre les deux hypothèses

$$H_0^1 : \theta_1 = 0 \quad vs \quad H_1^1 : \theta_1 \neq 0$$

Nous allons considérer la variable aléatoire définie par

$$Z_1(X) = \phi(X, U_1) - \phi(X, U_2)$$

Puisque $[\phi(X,U_2)=1] \subset [\phi(X,U_1)=1]$, on peut écrire

$$Z_1(X) = \begin{cases} 1 & \text{si } \phi(X,U_1)=1 \text{ et } \phi(X,U_2)=0 \\ \\ 0 & \text{si } \phi(X,U_1)=\phi(X,U_2) \end{cases}$$

(b) Pour tester l'égalité des sensibilités de U_1 et U_2, on considère le paramètre θ_2 défini par $\theta_2 = \Pr([\phi(X,U_1)=1, Y=1]) - \Pr([\phi(X,U_2)=1, Y=1])$. Les hypothèses à tester sont :

$$H_0^2 : \theta_2 = 0 \quad vs \quad H_1^2 : \theta_2 \neq 0$$

On peut associer au test la variable aléatoire $Z_2(X)$ définie par

$$Z_2(X) = \mathbb{1}\left([\phi(X,U_1)=1, Y=1]\right) - \mathbb{1}\left([\phi(X,U_2)=1, Y=1]\right)$$

Puisque $[\phi(X,U_2)=1, Y=1] \subset [\phi(X,U_1)=1, Y=1]$, on peut écrire

$$Z_2(X) = \begin{cases} 1 & \text{si } \mathbb{1}\left(\phi(X,U_1)=1, Y=1]\right)=1 \text{ et } \mathbb{1}\left(\phi(X,U_2)=1, Y=1]\right)=0 \\ \\ 0 & \text{si } \mathbb{1}\left(\phi(X,U_1)=1, Y=1]\right)=\mathbb{1}\left([\phi(X,U_2)=1, Y=1]\right) \end{cases}$$

(c) Pour tester l'égalité des spécificités de U_1 et U_2, on considère le paramètre θ_3 défini par $\theta_3 = \Pr([\phi(X,U_2)=0, Y=0]) - \Pr([\phi(X,U_1)=0, Y=0]$. L' hypothèse nulle et son alternative sont données par :

$$H_0^3 : \theta_3 = 0 \quad vs \quad H_1^3 : \theta_3 \neq 0$$

La variable aléatoire $Z_3(X)$ associée au test est définie par

$$Z_3(X) = \mathbb{1}\left([\phi(X,U_2)=0, Y=0]\right) - \mathbb{1}\left([\phi(X,U_1)=0, Y=0]\right)$$

Puisque $[\phi(X,U_2)=0, Y=0] \supset [\phi(X,U_1)=0, Y=0]$, on peut écrire

$$Z_3(X) = \begin{cases} 1 & \text{si } \mathbb{1}\left([\phi(X,U_1)=0, Y=0]\right)=0 \text{ et } \mathbb{1}\left([\phi(X,U_2)=0, Y=0]\right)=1 \\ \\ 0 & \text{si } \mathbb{1}\left([\phi(X,U_1)=0, Y=0]\right)=\mathbb{1}\left([\phi(X,U_2)=0, Y=0]\right) \end{cases}$$

Les variables $(Z_k(X))_{k=1:3}$ sont donc des variables aléatoires de Bernoulli de paramètre $(\theta_k)_{k=1:3}$.

On considère une suite d'éléments aléatoires $\mathcal{D}_n = (X_i, Y_i)_{i\in 1:n}$ indépendants et identiquement distribués, où Y_i est une réalisation d'une variable de Bernoulli Y et X_i est une suite finie de p réalisations d'un vecteur de variables aléatoires non numériques $(X_j)_{j=1:p}$ à q_j modalités $m_h^{X_j}; h = 1 : q_j, j = 1 : p$. Puisque les observations $(X_i)_{i=1:n}$ sont indépendantes alors les $Z_k(X_i)_{i=1:n}$ constituent des réalisations indépendantes. Donc la somme $\sum_{i=1}^n Z_k(X_i)$ est une réalisation d'une variable aléatoire suivant la loi binomiale $\mathcal{BN}(n, \theta_k)$. Nous considérons le test stochastique défini comme suit : Pour tout $k = 1 : 3$

$$\varphi_k(\mathcal{D}_n) = \begin{cases} 1 & \text{si } \sum_{i=1}^n Z_k(X_i) > 0 \\ \\ 1-\gamma_k & \text{si } \sum_{i=1}^n Z_k(X_i) = 0 \quad et \quad 0 < \gamma_k \leq 1 \end{cases}$$

On tire un nombre μ uniformément réparti entre 0 et 1. Si $\mu \geq 1-\gamma_k$ on rejette H_0^k et si $\mu < 1-\gamma_k$ on accepte H_0^k avec $0 < \gamma_k \leq 1$. L'application du test stochastique s'effectue comme suit :

- Si $\varphi_k(\mathcal{D}_n) = 1$: rejeter H_0^k
- Si $\varphi_k(\mathcal{D}_n) = 1-\gamma_k$: rejeter H_0^k avec une probabilité γ_k i.e. on génère une valeur μ uniforme sur 0 et 1. Si $\mu \geq 1-\gamma_k$, on rejette H_0^k, sinon on accepte.

Le niveau du test est obtenu en calculant

$$\begin{aligned}
\Pr\left(\text{rejeter } H_0^k | H_0^k\right) &= \Pr\left(\varphi_k(\mathcal{D}_n) = 1 \,|\, H_0^k\right) + \Pr\left(\varphi_k(\mathcal{D}_n) = 1-\gamma_k, \mu \geq 1-\gamma_k \,|\, H_0^k\right) \\
&= \Pr\left(\varphi_k(\mathcal{D}_n) = 1 \,|\, H_0^k\right) + \Pr\left(\varphi_k(\mathcal{D}_n) = 1-\gamma_k | \, H_0^k\right) \Pr\left(\mu \geq 1-\gamma_k \,|\, H_0^k\right) \\
&= \Pr\left(\sum_{i=1}^n Z_k(X_i) > 0 \,|\, H_0^k\right) + \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_0^k\right) \Pr\left(\mu \geq 1-\gamma_k\right) \\
&= 1 - \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_0^k\right) + \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_0^k\right) \left(1 - \Pr\left(\mu < 1-\gamma_k\right)\right) \\
&= 1 - \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_0^k\right) + \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_0^k\right) \gamma_k \\
&= \gamma_k \qquad \text{puisque} \quad \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_0^k\right) = 1
\end{aligned}$$

Et on obtient la puissance du test en calculant

$$\begin{aligned}
\Pr\left(\text{rejeter } H_0^k | H_1^k\right) &= \Pr\left(\varphi_k(\mathcal{D}_n) = 1 \,|\, H_1^k\right) + \Pr\left(\varphi_k(\mathcal{D}_n) = 1-\gamma_k, \mu \geq 1-\gamma_k \,|\, H_1^k\right) \\
&= \Pr\left(\varphi_k(\mathcal{D}_n) = 1 \,|\, H_1^k\right) + \Pr\left(\varphi_k(\mathcal{D}_n) = 1-\gamma_k | \, H_1^k\right) \Pr\left(\mu \geq 1-\gamma_k \,|\, H_1^k\right) \\
&= \Pr\left(\sum_{i=1}^n Z_k(X_i) > 0 \,|\, H_1^k\right) + \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_1^k\right) \Pr\left(\mu \geq 1-\gamma_k\right) \\
&= 1 - \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_1^k\right) + \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_1^k\right) \left(1 - \Pr\left(\mu < 1-\gamma_k\right)\right) \\
&= 1 - \Pr\left(\sum_{i=1}^n Z_k(X_i) = 0 \,|\, H_1^k\right) \Pr\left(\mu < 1-\gamma_k\right) \\
&= 1 - (1-\theta_k)^n (1-\gamma_k)
\end{aligned}$$

5.2 Algorithme de la procédure d'élagage

En se basant sur les résultats présentés dans la section précédente, on peut proposer une procédure d'élagage des profils redondants comme suit.

Algorithme : Procédure d'élagage des profils redondants

– Entrées : $\mathcal{R}$ un ensemble de profils
– Sorties : $\mathcal{R}'$ un ensemble de profils non redondants

1 : On se donne $\mathcal{R}$ un ensemble de profils
2 : **pour** tout profil $U \in \mathcal{R}$ **faire**

3 : $\mathcal{S}_U = \text{subset}(U, \mathcal{R})$ {le sous-ensemble de profils de $\mathcal{R}$ emboîtés dans U}

4 : **pour** tout profil $U' \in \mathcal{S}_U$ **faire**

5 : Tester $H_0^1 : \Pr\{\phi(X,U) = 1\} = \Pr\{\phi(X,U') = 1\}$ vs $H_1^1 : \Pr\{\phi(X,U) = 1\} \neq \Pr\{\phi(X,U') = 1\}$

6 : **Si** H_0^1 est vraie, $\mathcal{S}'_U = \text{delete}(U', \mathcal{S}_U)$ {supprimer U' de $\mathcal{S}_U$ en vertu de la proposition 3.}

7 : **Sinon**
8 : Tester $H_0^2 : \Pr\{\phi(X,U) = 1, Y = 1\} = \Pr\{\phi(X,U') = 1, Y = 1\}$ contre son opposée H_1^2

9 : Si H_0^2 est vraie, $\mathcal{S}'_U = \text{delete}(U, \mathcal{S}_U)$ {supprimer U de $\mathcal{S}_U$ en vertu de la proposition 4.}

10 : Tester $H_0^3 : \Pr\{\phi(X,U) = 0, Y = 0\} = \Pr\{\phi(X,U') = 0, Y = 0\}$ contre son opposée H_1^3

11 : Si H_0^3 est vraie, $\mathcal{S}'_U = \text{delete}(U', \mathcal{S}_U)$ {supprimer U' de $\mathcal{S}_U$ selon la proposition 5.}

12 : **fin si**
13 : **fin pour** U'
14 : **fin pour** U
15 : Retourner $\mathcal{R}' = \bigcup_{U \in \mathcal{R}} \mathcal{S}'_U$

Tableau III.2 – Algorithme d'élagage des profils redondants

Le test stochastique présenté ci-dessus est applicable quelle que soit la taille des données d'analyse. Habituellement, l'ensemble $\mathcal{U}_\lambda^1$ contient un grand nombre de profils, certainement plus qu'il en faut pour construire une fonction de classification qui est efficace et facile à mettre en œuvre.

6 Détermination d'un ensemble optimal de profils

6.1 Lorsque les données sont de grande taille

D'une manière générale, on peut utiliser un test comparant les valeurs prédictives positives de deux profils emboîtés pour sélectionner le profil le plus adéquat. Ce test est basé sur la normalité asymptotique du logarithme de rapport des valeurs prédictives positives des deux profils emboîtés.

6.1.1 Test d'hypothèse asymptotique pour la sélection d'un ensemble optimal de profils

Proposition 6. Soient $U_1 = \left(m_h^{X_j}\right)_{j\in J}$ et $U_2 = \left(m_h^{X_j}\right)_{j\in L}$ deux profils emboîtés tels que $J \subset L$. Soient $\widehat{VPP}(U_1, Y)$ et $\widehat{VPP}(U_2, Y)$ les estimateurs empiriques de $VPP(U_1, Y)$ et $VPP(U_2, Y)$ respectivement. La variable aléatoire $log\left(\frac{\widehat{VPP}(U_1,Y)}{\widehat{VPP}(U_2,Y)}\right)$ est asymptotiquement distribuée suivant une loi normale centrée de variance

$$\Sigma = \sum_{i=1}^{6} p_i \nabla_i^2 - \left(\sum_{i=1}^{6} p_i \nabla_i\right)^2$$

où

$$\begin{pmatrix} \nabla_1 \\ \nabla_2 \\ \nabla_3 \\ \nabla_4 \\ \nabla_5 \\ \nabla_6 \end{pmatrix} = \begin{pmatrix} \frac{1}{p_1+p_4} + \frac{1}{p_1+p_2} - \frac{1}{p_1} - \frac{1}{p_1+p_2+p_4+p_5} \\ \frac{1}{p_1+p_2} - \frac{1}{p_1+p_2+p_4+p_5} \\ 0 \\ \frac{1}{p_1+p_4} - \frac{1}{p_1+p_2+p_4+p_5} \\ -\frac{1}{p_1+p_2+p_4+p_5} \\ 0 \end{pmatrix}$$

Preuve. Soit le vecteur aléatoire $(Y, \phi(X, U_1), \phi(X, U_2))$. On considère les événements suivants :

$$E_1 = \{Y = 1, \phi(X, U_1) = 1, \phi(X, U_2) = 1\} \qquad E_2 = \{Y = 1, \phi(X, U_1) = 1, \phi(X, U_2) = 0\}$$
$$E_3 = \{Y = 1, \phi(X, U_1) = 0, \phi(X, U_2) = 0\} \qquad E_4 = \{Y = 0, \phi(X, U_1) = 1, \phi(X, U_2) = 1\}$$
$$E_5 = \{Y = 0, \phi(X, U_1) = 1, \phi(X, U_2) = 0\} \qquad E_6 = \{Y = 0, \phi(X, U_1) = 0, \phi(X, U_2) = 0\}$$

dont les probabilités de réalisation sont p_1, p_2, p_3, p_4, p_5 et p_6 respectivement avec

$$\sum_{i=1}^{6} p_i = 1$$

Compte tenu du fait que U_2 soit emboîté dans U_1, on a

$$p_1 = \Pr(Y = 1, \phi(X, U_2) = 1) \qquad p_4 = \Pr(Y = 0, \phi(X, U_2) = 1)$$
$$p_3 = \Pr(Y = 1, \phi(X, U_1) = 0) \qquad p_6 = \Pr(Y = 0, \phi(X, U_1) = 0)$$

On note par $I_{E_k}, k = 1 : 6$ la fonction indicatrice de l'événement E_k. La distribution de Bernoulli généralisée de paramètres $\theta = (p_1, \ldots, p_6)$ de la variable aléatoire $Z = (I_{E_1}, \ldots, I_{E_6})$ admet comme matrice de variance covariance la matrice

$$\Lambda(\theta) = diag(\theta) - \theta^T\theta$$

Soit $(Z_i)_{i=1:n}$ une suite indépendante de distribution la Bernoulli généralisée. Si on considère

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

l'estimateur empirique de θ, le théorème central limite permet de dire que

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Lambda(\theta))$$

Par ailleurs nous avons

$$VPP(U_1, Y) = \frac{\Pr\{Y = 1, \phi(X, U_1) = 1\}}{\Pr\{\phi(X, U_1) = 1\}} = \frac{p_1 + p_2}{p_1 + p_2 + p_4 + p_5}$$

$$VPP(U_2, Y) = \frac{\Pr\{Y = 1, \phi(X, U_2) = 1\}}{\Pr\{\phi(X, U_2) = 1\}} = \frac{p_1}{p_1 + p_4}$$

d'où

$$\frac{VPP(U_1, Y)}{VPP(U_2, Y)} = \frac{(p_1 + p_4)(p_1 + p_2)}{p_1(p_1 + p_2 + p_4 + p_5)}$$

Soit la fonction

$$g(\theta) = log\left(\frac{VPP(U_1, Y)}{VPP(U_2, Y)}\right)$$

On a

$$\nabla g(\theta) = \begin{pmatrix} \frac{1}{p_1+p_4} + \frac{1}{p_1+p_2} - \frac{1}{p_1} - \frac{1}{p_1+p_2+p_4+p_5} \\ \frac{1}{p_1+p_2} - \frac{1}{p_1+p_2+p_4+p_5} \\ 0 \\ \frac{1}{p_1+p_4} - \frac{1}{p_1+p_2+p_4+p_5} \\ -\frac{1}{p_1+p_2+p_4+p_5} \\ 0 \end{pmatrix}$$

En utilisant la Méthode Delta Multivariée, on démontre que

$$\sqrt{n}\left(g(\widehat{\theta}_n) - g(\theta)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, {}^T\nabla g(\theta)\Lambda(\theta)\nabla g(\theta)\right)$$

où

$$\begin{aligned} {}^T\nabla g(\theta)\Lambda(\theta)\nabla g(\theta) &= {}^T\nabla g(\theta) diag(\theta)\nabla g(\theta) - (\theta\nabla g(\theta))^T (\theta\nabla g(\theta)) \\ &= \sum_{i=1}^{6} p_i \nabla_i^2 - \left(\sum_{i=1}^{6} p_i \nabla_i\right)^2 \end{aligned}$$

avec

$$\nabla g(\theta) = \begin{pmatrix} \nabla_1 \\ \vdots \\ \nabla_6 \end{pmatrix}$$

Etant donné que $\sum_{i=1}^{6} p_i = 1$, alors ${}^T\nabla g(\theta)\Lambda(\theta)\nabla g(\theta) > 0$ puisque c'est une variance du vecteur $(\nabla_1, \ldots, \nabla_6)$ qui n'est pas colinéaire avec le vecteur $\mathbb{1} = (1, \ldots, 1)$. □

L'application : $\theta \longmapsto \nabla g(\theta)$ est continue de même que l'application : $\theta \longmapsto \Lambda(\theta)$. Et puisque $\widehat{\theta}_n$ converge en presque sûrement vers θ, on obtient alors

$${}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n) \xrightarrow{p.s} {}^T\nabla g(\theta)\Lambda(\theta)\nabla g(\theta)$$

Grâce au théorème de Slutsky, on peut conclure que

$$\frac{\sqrt{n}\left(g(\widehat{\theta}_n) - g(\theta)\right)}{\sqrt{{}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Sous l'hypothèse que $\dfrac{VPP(U_1, Y)}{VPP(U_2, Y)} = 1$, si la taille de l'échantillon est suffisamment grande alors

$$\frac{\sqrt{n}\left(g(\widehat{\theta}_n)\right)}{\sqrt{{}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Ce qui nous permet de construire une stratégie de sélection du profil le plus adéquat. Si on note par $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite, on peut effectuer les tests suivants.

1. **Test 1** :

(a) Sélectionner le profil U_1 si

$$g(\widehat{\theta}_n) \geq q_{1-\alpha/2}\sqrt{\frac{{}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n)}{n}}$$

(b) Sélectionner le profil U_2 si

$$g(\widehat{\theta}_n) \leq -q_{1-\alpha/2}\sqrt{\frac{{}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n)}{n}}$$

(c) Choisir au hasard entre U_1 et U_2 si

$$g(\widehat{\theta}_n) \in \left]\ -q_{1-\alpha/2}\sqrt{\frac{{}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n)}{n}}, q_{1-\alpha/2}\sqrt{\frac{{}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n)}{n}}\ \right[$$

Cette troisième étape du test utilise le principe du test stochastique (test randomisé) où on génère une réalisation b d'une variable de Bernoulli de paramètre 1/2. On sélectionne U_1 si $b = 1$ sinon on sélectionne U_2.

2. **Test 2** :

(a) Sélectionner le profil U_2 si

$$g(\widehat{\theta}_n) < -q_{1-\alpha/2}\sqrt{\frac{{}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n)}{n}}$$

(b) sinon Sélectionner le profil U_1

Le Test 2 permet de favoriser les profils les plus courts. Les résultats présentés dans cette analyse sont obtenus en utilisant le Test 2.

6.1.2 Algorithme

A partir d'un ensemble de validation, nous cherchons à réduire l'ensemble $\mathcal{U}_\lambda^1$ en utilisant la valeur prédictive positive comme paramètre de comparaison. Les indicateurs de performance tels que les rapports de vraisemblance positifs (RVP) ou les rapports de vraisemblance négatifs (RVN) peuvent également être utilisés.

Algorithme : Réduction de l'ensemble $\mathcal{U}_\lambda^1$

– Entrées : $\mathcal{D}$ un ensemble d'observation ; $\mathcal{U}_\lambda^1$ un ensemble de règles non redondantes
– Sorties : $\mathcal{U}_\lambda^2$ un ensemble optimal de profils

1 : **pour** tout profil $C \in \mathcal{U}_\lambda^1$ **faire**
2 : $S = is.subset(C, \mathcal{U}_\lambda^1)$ {le sous-ensemble des profils emboîtés dans C}
3 : **pour** tout profil $C' \in S$ **faire**
4 : Evaluer les indicateurs suivants
5 : $\widehat{\theta}_n = (p_1, \ldots, p_6|\mathcal{D})$
6 : $g(\widehat{\theta}_n) = log(VPP(C, Y|\widehat{\theta}_n)) - log(VPP(C', Y|\widehat{\theta}_n))$
7 : $\Lambda(\widehat{\theta}_n) = diag(\widehat{\theta}_n) - \widehat{\theta}_n^t \widehat{\theta}_n$
8 : $\nabla_n = \nabla g(\widehat{\theta}_n)$
9 : **fin pour**
10 : **Si** il existe $C' \in S$ tel que $g(\widehat{\theta}_n) < -q_{1-\alpha/2}\sqrt{\frac{\nabla_n^t \Lambda(\widehat{\theta}_n)\nabla_n}{n}}$ **faire**
11 : $\mathcal{U}_\lambda^2 = delete(C, \mathcal{U}_\lambda^1)$ {Supprimer le profil C}
12 : **Sinon**
$\mathcal{U}_\lambda^2 = delete(S, \mathcal{U}_\lambda^1)$ {Supprimer le sous-ensemble S}
13 : **fin si**
14 : **fin pour**
15 : **Résultat** $\mathcal{U}_\lambda^2$

Tableau III.3 – Algorithme de réduction de l'ensemble non redondant

Le processus d'apprentissage, tel qu'il a été décrit jusqu'ici requiert une grande base de données qu'il faudra échantillonner en trois sous-ensembles (apprentissage, validation et test) de tailles suffisamment grandes. Habituellement dans la tâche de l'apprentissage automatique, il est courant que le nombre d'observations disponibles ne permettent pas une subdivision des données en trois échantillons, un pour l'apprentissage, un pour la validation et un pour le test. Le recours à l'échantillon de validation permet d'évaluer les paramètres de performance sur un échantillon différent mais issu de la même distribution que l'échantillon d'apprentissage. On peut envisager alors une procédure bootstrap.

6.2 Lorsque les données sont de taille petite

Lorsqu'on ne dispose pas de données suffisantes pour une subdivision en trois sous-ensembles : apprentissage, validation et test, on peut recourir à une procédure de bootstrap pour la validation du

classifieur. En effet lorsque n,la taille de l'échantillon, est petite, la condition

$$S = \frac{\sqrt{n}\left(g(\widehat{\theta}_n) - g(\theta)\right)}{\sqrt{{}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

n'est plus assurée. D'où la nécessité de recourir à un test d'hypothèse bootstrap.

6.2.1 Test d'hypothèse boostrap pour la sélection d'un ensemble optimal de profils

Le bootstrap est une technique de ré-échantillonnage bien connue dans la littérature [9, 10]. Le principe fondamental du bootstrap est de substituer à la distribution inconnue F, dont est issu l'échantillon d'apprentissage, la distribution empirique F_n qui donne un poids $1/n$ à chaque réalisation. Ainsi on obtient un échantillon de taille n dit échantillon bootstrap selon la distribution empirique F_n par n tirages aléatoires avec remise parmi les n observations initiales.

La statistique d'intérêt S a une distribution d'échantillonnage notée F_S. Cette distribution dépend de la distribution G_Z de la variable aléatoire Z dont les valeurs observées sont $z_1, \ldots, z_n$. On écrit $F_S(s, G_Z)$, où G_Z est la distribution de Bernoulli généralisée de la variable Z. La distribution G_Z, quant à elle, dépend de la distribution F_X de la variable aléatoire X dont les observations sont $x_1, \ldots, x_n$. On note $G_Z(z, F_X)$. En résumé, la distribution F_S dépend de la réalisation z de la variable Z et de la distribution F_X de la variable X. On écrit $F_S(s, z, F_X)$.
Puisque F_X est inconnue, on travaille avec une estimation de F_X que l'on note $\widehat{F}_X$ et qui est la distribution empirique F_n des données $\{x_1, \ldots, x_n\}$. Le fait de remplacer F_X par F_n va donner une distribution d'échantillonnage F_S également modifiée. On écrit $F_S(s, z, F_n)$ au lieu de $F_S(s, z, F_x)$. Remplacer F_X par F_n et générer un échantillon de taille n selon la distribution F_n revient de même que de tirer avec remise n éléments de l'ensemble de données originales $\{x_1, \ldots, x_n\}$.
On a $g(\widehat{\theta}_n)$ un estimateur de la quantité $g(\theta)$ et $\widehat{\sigma}_n = \sqrt{\frac{1}{n}\left({}^T\nabla g(\widehat{\theta}_n)\Lambda(\widehat{\theta}_n)\nabla g(\widehat{\theta}_n)\right)}$ un estimateur de l'écart type de $g(\widehat{\theta}_n) - g(\theta)$. On note par $g(\widehat{\theta}^*_n)$ une estimation de $g(\theta)$ et $\widehat{\sigma}^*_n$ une estimation de l'écart type de $g(\widehat{\theta}^*_n) - g(\widehat{\theta}_n)$ toutes deux calculées à partir d'un échantillon bootstrap. En particulier $\widehat{\sigma}^*_n$ est l'estimation empirique bootstrap de l'écart type de $g(\widehat{\theta}^*_n) - g(\widehat{\theta}_n)$. Alors la distribution bootstrap de $\left(g(\widehat{\theta}^*_n) - g(\widehat{\theta}_n)\right)/\widehat{\sigma}^*_n$ estime la distribution bootstrap de $\left(g(\widehat{\theta}_n) - g(\theta)\right)/\widehat{\sigma}_n$ sous l'hypothèse nulle [16]. Baser le test d'hypothèse sur la distribution bootstrap de $\left(g(\widehat{\theta}^*_n) - g(\widehat{\theta}_n)\right)/\widehat{\sigma}^*_n$ permet d'améliorer la précision du niveau du test sans modifier la puissance du test [4, 16].
Pour appliquer le test bilatéral bootstrap de $H_0 : g(\theta) = 0$ au niveau α, on effectue les instructions suivantes : commence par

1. Calculer la valeur de la statistique S pour l'échantillon de départ : soit s_0 la valeur observée.
2. Simuler B échantillons de taille n observations tirées de façon aléatoire avec remise à partir de

l'ensemble de données originales, et obtenir ainsi B valeurs simulées de s_b^* de S :

$$s_b^* = \frac{g(\widehat{\theta}_n^b) - g(\widehat{\theta}_n)}{\widehat{\sigma}_n^b}, \qquad b = 1, \ldots, B$$

3. Calculer la $p - value\ bootstrap$

$$p^* = \frac{1}{B} \sum_{b=1}^{B} I(s_b^* > s_0)$$

On peut formuler alors la règle de décision suivante :

1. **Test 1.**

   (a) Sélectionner le profil U_2 si $p^* < \alpha/2$

   (b) Sélectionner le profil U_1 si $p^* > 1 - \alpha/2$

   (c) Choisir au hasard entre U_1 et U_2 si $p^* \in [\,\alpha/2,\, 1 - \alpha/2\,]$
   Cette troisième étape du test utilise le principe du test stochastique (test randomisé) où on génère une réalisation b d'une variable de Bernoulli de paramètre 1/2. On sélectionne U_1 si $b = 1$ sinon on sélectionne U_2.

2. **Test 2** :

   (a) Sélectionner le profil U_2 si $p^* < \alpha/2$

   (b) sinon Sélectionner le profil U_1

Le Test 2 permet de favoriser les profils les plus courts. Les résultats présentés dans cette analyse sont obtenus en utilisant le Test 2.

6.2.2 Algorithme

L'algorithme d'apprentissage statistique, adapté au bootstrap, est le suivant :

Algorithme : Réduction de l'ensemble $\mathcal{U}_\lambda^1$

– Entrées : $\mathcal{D}$ un ensemble d'observation ; $\mathcal{U}_\lambda^1$ un ensemble de règles non redondantes, $\alpha = 0.05$ le niveau du test et B le nombre d'échantillon bootstrap (20 par défaut).
– Sorties : $\mathcal{U}_\lambda^2$ un ensemble optimal de profils

pour tout profil $C \in \mathcal{U}_\lambda^1$ **faire**
  $S = is.supset(C, \mathcal{U}_\lambda^1)$ {le sous-ensemble des profils emboîtés dans C}
  pour tout profil $C' \in S$ **faire**
    Evaluer les indicateurs suivants
    $\widehat{\theta}_n = (p_1, \ldots, p_6 | \mathcal{D})$
    $g(\widehat{\theta}_n) = log(VPP(C, Y|\widehat{\theta}_n)) - log(VPP(C', Y|\widehat{\theta}_n))$
    $\Lambda(\widehat{\theta}_n) = diag(\widehat{\theta}_n) - \widehat{\theta}_n^t \widehat{\theta}_n$
    $\nabla_n = \nabla g(\widehat{\theta}_n)$
    $\widehat{\sigma}_n = \sqrt{\frac{1}{n}\left(\nabla_n^t \Lambda(\widehat{\theta}_n) \nabla_n\right)}$
    $s_0 = g(\widehat{\theta}_n)/\widehat{\sigma}_n$
    pour tout échantillon boostrap $\mathcal{D}^b$ **faire**
      $\widehat{\theta}_n^b = (p_1, \ldots, p_6 | \mathcal{D}^b)$
      $g(\widehat{\theta}_n^b) = log(VPP(C, Y|\widehat{\theta}_n^b)) - log(VPP(C', Y|\widehat{\theta}_n^b))$
      $\Lambda(\widehat{\theta}_n^b) = diag(\widehat{\theta}_n^b) - (\widehat{\theta}_n^b)^t \widehat{\theta}_n^b$
      $\nabla_n = \nabla\left(g(\widehat{\theta}_n^b) - g(\widehat{\theta}_n)\right)$
      $\widehat{\sigma}_n^b = \sqrt{\frac{1}{n}\left(\nabla_n^t \Lambda(\widehat{\theta}_n^b) \nabla_n\right)}$
      $s_b^* = \left(g(\widehat{\theta}_n^b) - g(\widehat{\theta}_n)\right)/\widehat{\sigma}_n^b$
    fin pour
    Calculer la $p - value$

$$p^* = \frac{1}{B}\sum_{b=1}^{B} I(s_b^* > s_0)$$

    si $p^* < \alpha/2$ **faire**
      $\mathcal{U}_\lambda^2 = delete(C, \mathcal{U}_\lambda^1)$ {Supprimer le profil C}
    sinon
      $\mathcal{U}_\lambda^2 = delete(C', \mathcal{U}_\lambda^1)$ {Supprimer le profil C'}
    fin si
  fin pour
fin pour
Résultat $\mathcal{U}_\lambda^2$

Tableau III.4 – Algorithme de réduction de l'ensemble non redondant lorsque l'échantillon d'apprentissage est de petite taille

7 Application à des données de la littérature

Toutes les données que nous avons utilisé pour l'application de l'algorithme d'apprentissage sont issues du répertoire d'apprentissage automatique UCI (UCI Machine Leraning Repository) [3]. Toutes les analyses relatives à la méthode de classement proposée ont été réalisées dans l'environnement de programmation R [25]. L'exploration des règles d'association a été faite en utilisant le package arules [1]. Nous avons également utilisé le package rpart [28], la package partykit [18], le package $e1071$ [22] et le package DMwR [29] pour comparer notre approche avec celles existantes dans la littérature.

7.1 Données Adult Data Set

Les données d'application sont extraites de la base de données du bureau de recensement de 1994 [19]. Elles contiennent essentiellement des sujets âgés de plus de 16 ans et ayant à la fois un revenu brut ajusté supérieur à 1 et un volume horaire de travail positif. Au total, elles contiennent 45222 sujets hormis les données manquantes. Les sujets sont échantillonnés sur deux ensembles : un ensemble d'apprentissage de 30162 sujets (2/3 de données totales) et un ensemble test de 15060 sujets. Les données contiennent 14 covariables dont 5 sont continues et 8 sont nominales dont une variable réponse binaire indexant le revenu annuel d'un sujet à plus de $ 50K ou moins. L'objectif visé dans cette analyse est de trouver un profil prédictif du niveau de revenu d'un sujet donné.
Pour évaluer la procédure d'apprentissage des règles d'association binaire, nous allons effectuer plusieurs expériences en sur-échantillonnant ou en sous-échantillonnant le jeu de données census. Pour obtenir un ensemble de données déséquilibrées, on commence par sélectionner toutes les observations de la classe prévalente ; ensuite on se fixe une proportion α de la classe rare. Soit n le nombre d'observations de la classe prévalente. On sélectionne $n' = n\alpha/(1-\alpha)$ observations de la classe rare. On obtient ainsi, un échantillon de $n + n'$ observations avec une proportion α de la classe rare.
Dans tout ce qui suit, nous avons fixé le paramètre de la taille maximale des règles à 4, le paramètre du risque relatif minimal égal à 1 et le paramètre de la p-value minimale associée au test exact de Fisher égale à 0.05. Après avoir construit notre échantillon déséquilibré, on se fixe un seuil de support minimale (minsup) et un seuil de valeur prédictive positive minimale (minconf). Ces derniers nous permettront de générer l'ensemble de règles d'association fréquentes $\mathcal{R}$. Pour chaque expérience, on subdivise aléatoirement l'échantillon en deux parties : apprentissage et validation. Un ensemble test est utilisé pour évaluer les performances du classifieur. Cependant, on peut évaluer deux types d'erreurs de classement : l'erreur de classement lorsque la distribution de l'ensemble d'apprentissage est différente de la distribution de l'ensemble test et l'erreur de classement lorsque la distribution de l'ensemble d'apprentissage est identique à la distribution de l'ensemble test.

7.1.1 Performances du classifieur lorsque la distribution de l'échantillon test est identique à celui de l'échantillon d'apprentissage

Proportions	Nb profils dans $\mathcal{U}_\lambda$	Erreur.cl $\mathcal{U}_\lambda$	Nb profils dans $\mathcal{U}_\lambda^2$	Sensibilité	Spécificité	Erreur.clt	Minsup Minconf
	76	0.22	12	0.68	0.81	0.19	
	129	0.28	10	0.69	0.78	0.22	
	110	0.25	15	0.70	0.78	0.23	
	69	0.19	14	0.60	0.83	0.17	
	92	0.24	12	0.72	0.80	0.20	
<=50K >50K	101	0.27	16	0.71	0.81	0.19	0.001 0.028
0.993 0.007	130	0.32	12	0.80	0.77	0.23	
	145	0.30	11	0.74	0.81	0.19	
	126	0.35	17	0.74	0.74	0.26	
	101	0.24	06	0.62	0.83	0.17	
	110	0.22	13	0.74	0.81	0.19	
	104	0.23	11	0.60	0.81	0.19	
	61	0.19	10	0.67	0.83	0.17	
	62	0.19	10	0.67	0.85	0.16	
	69	0.21	11	0.72	0.82	0.18	
	34	0.08	04	0.49	0.93	0.08	
	91	0.23	09	0.71	0.83	0.17	
<=50K >50K	81	0.21	09	0.61	0.85	0.15	0.002 0.06
0.985 0.015	70	0.19	10	0.71	0.83	0.17	
	59	0.22	15	0.80	0.78	0.22	
	67	0.21	08	0.72	0.84	0.16	
	91	0.24	11	0.70	0.80	0.20	
	69	0.21	09	0.72	0.83	0.18	
	92	0.23	07	0.60	0.89	0.12	

Tableau III.5 – Performance prédictive sur 12 expériences : (0.7% & 1.5%)

Proportions	Nb profils dans $\mathcal{U}_\lambda$	Erreur.cl $\mathcal{U}_\lambda$	Nb profils dans $\mathcal{U}_\lambda^2$	Sensibilité	Spécificité	Erreur.clt	Minsup Minconf
	56	0.23	22	0.79	0.77	0.23	
	64	0.25	19	0.77	0.79	0.21	
	43	0.19	15	0.68	0.84	0.17	
	55	0.26	09	0.68	0.83	0.17	
	35	0.19	06	0.48	0.92	0.10	
<=50K >50K	44	0.20	10	0.67	0.86	0.14	0.005 0.10
0.97 0.03	35	0.22	09	0.70	0.83	0.17	
	66	0.25	16	0.71	0.81	0.20	
	51	0.20	11	0.75	0.83	0.18	
	59	0.24	11	0.68	0.81	0.19	
	58	0.24	16	0.80	0.77	0.23	
	50	0.26	13	0.81	0.77	0.23	
	67	0.20	21	0.76	0.80	0.20	
	83	0.22	24	0.77	0.78	0.22	
	69	0.20	15	0.73	0.83	0.18	
	74	0.16	20	0.66	0.86	0.16	
	73	0.20	14	0.70	0.83	0.18	
<=50K >50K	50	0.20	14	0.71	0.83	0.17	0.01 0.23
0.93 0.07	50	0.16	16	0.63	0.88	0.14	
	63	0.18	20	0.67	0.83	0.18	
	50	0.16	19	0.64	0.86	0.16	
	55	0.20	16	0.72	0.83	0.17	
	67	0.20	17	0.73	0.83	0.18	
	75	0.18	18	0.67	0.83	0.18	

Tableau III.6 – Performance prédictive sur 12 expériences : (3% & 7%)

Proportions	Nb profils dans $\mathcal{U}_\lambda$	Erreur.cl $\mathcal{U}_\lambda$	Nb profils dans $\mathcal{U}_\lambda^2$	Sensibilité	Spécificité	Erreur.clt	Minsup Minconf
	62	0.23	19	0.67	0.86	0.17	
	56	0.22	20	0.78	0.78	0.22	
	62	0.23	17	0.63	0.83	0.20	
	60	0.22	18	0.68	0.83	0.19	
	49	0.23	22	0.75	0.78	0.23	
<=50K >50K	40	0.20	10	0.58	0.88	0.16	0.025 0.4
0.85 0.15	54	0.23	21	0.70	0.83	0.19	
	59	0.23	18	0.61	0.86	0.18	
	33	0.19	13	0.64	0.86	0.18	
	44	0.21	18	0.73	0.80	0.21	
	65	0.23	20	0.67	0.86	0.17	
	46	0.23	11	0.64	0.86	0.18	
	58	0.20	17	0.65	0.88	0.16	
	66	0.20	22	0.68	0.83	0.20	
	62	0.20	21	0.67	0.86	0.18	
	66	0.20	21	0.68	0.83	0.20	
	46	0.18	18	0.61	0.88	0.18	
<=50K >50K	64	0.20	23	0.65	0.88	0.16	0.03 0.5
0.80 0.20	53	0.18	18	0.65	0.88	0.17	
	75	0.22	19	0.68	0.83	0.20	
	57	0.18	19	0.65	0.88	0.16	
	49	0.19	19	0.68	0.84	0.19	
	58	0.18	20	0.67	0.86	0.18	
	67	0.22	20	0.74	0.81	0.20	

Tableau III.7 – Performance prédictive sur 12 expériences : (15% & 20%)

Avec un diagramme-boîtes en parallèle, nous avons représenté, pour chaque série de 100 valeurs des différentes mesures de performances (sensibilité, spécificité et erreur de classement), la distribution de celles-ci de manière très simplifiée avec la médiane (trait épais), une boîte qui s'étend du quartile 0.25 au quartile 0.75, et des moustaches qui s'étendent par défaut jusqu'à la valeur distante d'au maximum 1.5 fois la distance inter-quartile.

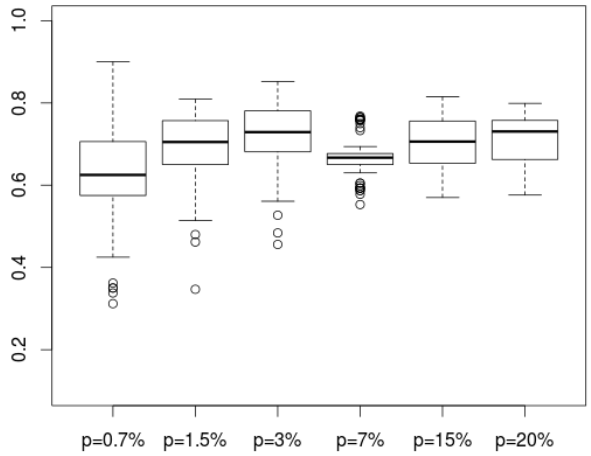


Figure III.1 – Distribution de la sensibilité estimée sur 100 échantillons

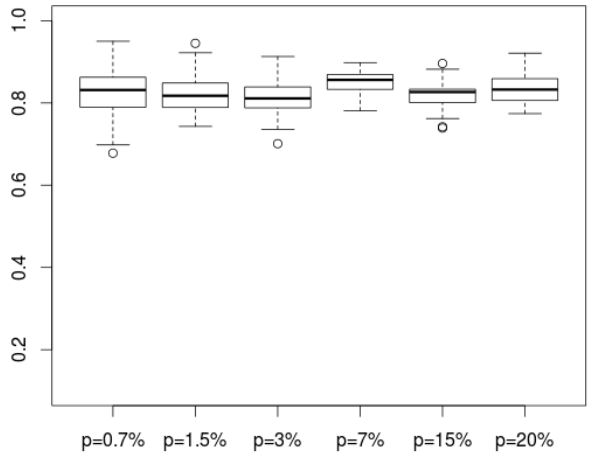


Figure III.2 – Distribution de la spécificité estimée sur 100 échantillons

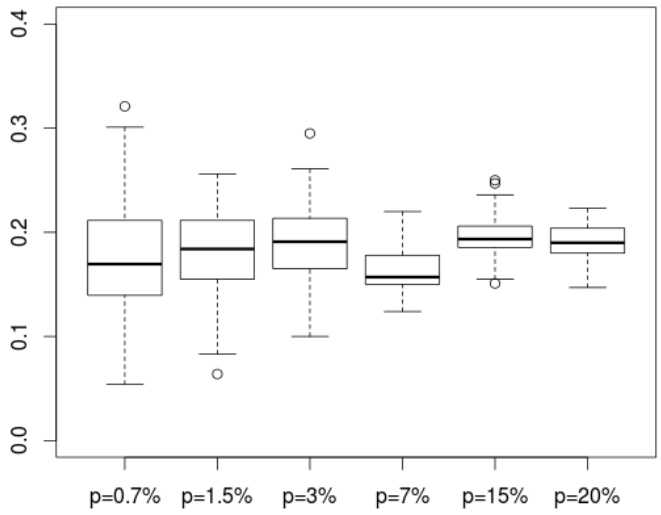


Figure III.3 – Distribution de l'erreur de classement estimée sur 100 échantillons

7.1.2 Performances du classifieur lorsque la distribution de l'échantillon test est différente de celui de l'échantillon d'apprentissage

Proportions	Nb profils dans $\mathcal{U}_\lambda$	Erreur.cl $\mathcal{U}_\lambda$	Nb profils dans $\mathcal{U}_\lambda^2$	Sensibilité	Spécificité	Erreur.clt	Minsup	Minconf
	80	0.24	12	0.71	0.77	0.24		
	101	0.26	14	0.78	0.73	0.25		
	80	0.23	11	0.74	0.78	0.23		
	47	0.20	13	0.70	0.82	0.20		
	93	0.24	07	0.57	0.83	0.23		
<=50K >50K	46	0.21	14	0.73	0.83	0.20	0.001	0.028
0.993 0.007	113	0.27	14	0.72	0.73	0.27		
	71	0.21	12	0.55	0.83	0.23		
	94	0.23	13	0.74	0.79	0.22		
	51	0.20	08	0.54	0.85	0.22		
	102	0.23	04	0.46	0.91	0.19		
	53	0.19	06	0.49	0.90	0.20		
	53	0.19	08	0.58	0.89	0.18		
	67	0.20	16	0.66	0.83	0.21		
	40	0.19	10	0.64	0.84	0.21		
	59	0.18	16	0.66	0.86	0.18		
	100	0.23	16	0.74	0.77	0.23		
<=50K >50K	53	0.19	13	0.69	0.84	0.19	0.002	0.06
0.985 0.015	46	0.17	08	0.57	0.93	0.16		
	64	0.18	14	0.65	0.86	0.18		
	73	0.19	12	0.59	0.85	0.21		
	60	0.18	11	0.61	0.90	0.17		
	74	0.18	14	0.65	0.89	0.17		
	97	0.21	13	0.74	0.83	0.19		

Tableau III.8 – Performance prédictive sur 12 expériences : (0.7% & 1.5%)

Proportions	Nb profils dans $\mathcal{U}_\lambda$	Erreur.cl $\mathcal{U}_\lambda$	Nb profils dans $\mathcal{U}_\lambda^2$	Sensibilité	Spécificité	Erreur.clt	Minsup Minconf
	74	0.26	20	0.66	0.81	0.23	
	75	0.24	18	0.82	0.75	0.24	
	63	0.22	16	0.76	0.81	0.20	
	68	0.24	12	0.67	0.79	0.24	
	61	0.21	11	0.58	0.88	0.19	
<=50K >50K	61	0.22	16	0.77	0.80	0.21	0.005 0.1
0.97 0.03	51	0.21	10	0.70	0.82	0.21	
	74	0.27	20	0.75	0.76	0.25	
	51	0.21	11	0.70	0.82	0.21	
	85	0.25	14	0.72	0.80	0.22	
	62	0.24	15	0.77	0.77	0.23	
	71	0.24	21	0.80	0.75	0.24	
	73	0.20	25	0.75	0.81	0.20	
	71	0.22	23	0.74	0.83	0.19	
	76	0.20	24	0.75	0.81	0.20	
	85	0.22	24	0.75	0.81	0.20	
	75	0.20	20	0.66	0.86	0.18	
<=50K >50K	73	0.20	20	0.74	0.83	0.19	0.01 0.23
0.93 0.07	71	0.18	19	0.66	0.86	0.18	
	71	0.21	23	0.76	0.80	0.21	
	71	0.18	22	0.66	0.86	0.18	
	67	0.20	20	0.73	0.84	0.18	
	76	0.20	20	0.66	0.86	0.18	
	79	0.22	18	0.68	0.83	0.21	

Tableau III.9 – Performance prédictive sur 12 expériences : (3% & 7%)

Proportions	Nb profils dans $\mathcal{U}_\lambda$	Erreur.cl $\mathcal{U}_\lambda$	Nb profils dans $\mathcal{U}_\lambda^2$	Sensibilité	Spécificité	Erreur.clt	Minsup Minconf
	56	0.20	16	0.66	0.86	0.18	
	65	0.22	19	0.76	0.81	0.20	
	62	0.22	17	0.66	0.86	0.18	
	59	0.22	14	0.65	0.89	0.17	
	55	0.22	17	0.66	0.86	0.18	
<=50K >50K	52	0.21	21	0.70	0.81	0.21	0.025 0.4
0.85 0.15	60	0.22	21	0.74	0.81	0.21	
	56	0.22	22	0.75	0.81	0.20	
	52	0.23	15	0.60	0.86	0.20	
	58	0.22	20	0.77	0.79	0.22	
	64	0.22	18	0.66	0.86	0.18	
	56	0.22	17	0.75	0.81	0.20	
	62	0.21	20	0.72	0.84	0.18	
	65	0.21	21	0.76	0.81	0.20	
	59	0.21	21	0.74	0.82	0.20	
	75	0.22	22	0.74	0.83	0.19	
	54	0.18	17	0.65	0.89	0.17	
<=50K >50K	62	0.20	22	0.66	0.87	0.18	0.03 0.5
0.80 0.20	54	0.20	20	0.68	0.84	0.20	
	58	0.20	22	0.75	0.81	0.20	
	54	0.18	19	0.65	0.89	0.17	
	46	0.18	18	0.64	0.89	0.17	
	56	0.20	19	0.74	0.82	0.20	
	70	0.22	25	0.74	0.83	0.19	

Tableau III.10 – Performance prédictive sur 12 expériences : (15% & 20%)

Avec un diagramme-boîtes en parallèle, nous avons représenté, pour chaque série de 100 valeurs des différentes mesures de performances (sensibilité, spécificité et erreur de classement), la distribution de celles-ci de manière très simplifiée avec la médiane (trait épais), une boîte qui s'étend du quartile 0.25 au quartile 0.75, et des moustaches qui s'étendent par défaut jusqu'à la valeur distante d'au maximum 1.5 fois la distance inter-quartile.

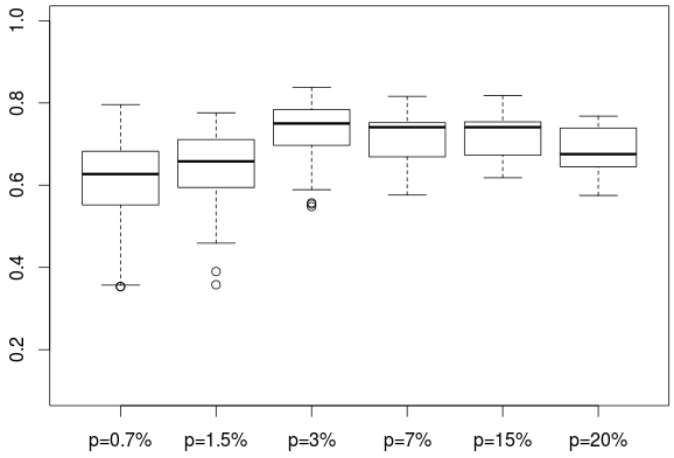


Figure III.4 – Distribution de la sensibilité estimée sur 100 échantillons

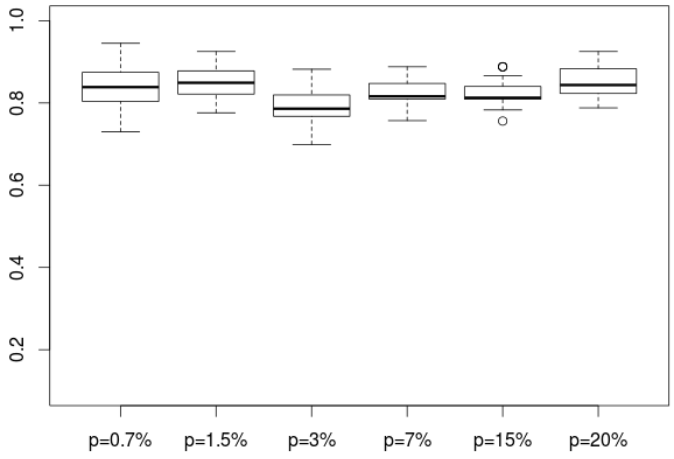


Figure III.5 – Distribution de la spécificité estimée sur 100 échantillons

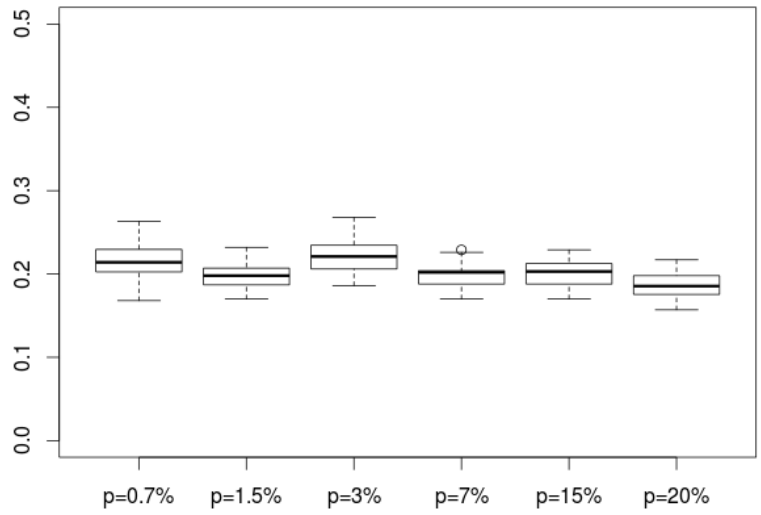


Figure III.6 – Distribution de l'erreur de classement estimée sur 100 échantillons

7.2 Comparaison de la méthode d'apprentissage avec des méthodes alternatives

Le classement binaire basé sur la régression logistique ou les arbres binaires de régression implique l'ajustement d'un modèle paramétrique ou non paramétrique aux probabilités conditionnelles $\Pr(Y = y|X = x)$ où $y \in \mathcal{D}om(Y)$ et $x \in \mathcal{D}om(X)$. Notons par $\Pr(Y = y|X = x, \mathcal{D})$ la probabilité ajustée aux données $\mathcal{D}$ et considérée comme un score. Dans ces cas, le classifieur ϕ est alors défini par la donnée d'un seuil $\lambda \in]0, 1[$ par

$$\phi(x|\lambda) = \begin{cases} 1 & \text{si } \Pr(Y = y|X = x, \mathcal{D}) > \lambda \\ 0 & \text{sinon} \end{cases}$$

Dans le cas de l'analyse discriminante ou des réseaux bayésiens comme le réseau bayésien naïf on considère une loi a priori π pour la distribution de probabilité des classes et on ajuste un modèle paramétrique ou non paramétrique aux lois conditionnelles de X sachant que $Y = y$. Notons par $\Pr(X = x|Y = y)$ la densité conditionnelle de X sachant $Y = y$ selon que X est discrète ou non. Le classifieur est obtenu à partir de la loi a posteriori de Y sachant que $X = x$ qui est définie par $\dfrac{\Pr(x|Y = y, \mathcal{D})\pi(y)}{\Pr(x|\mathcal{D})}$ considérée comme un score où $\Pr(x|Y = y, \mathcal{D})$ est la loi ajustée en utilisant les données $\mathcal{D}$ et $\Pr(x|\mathcal{D})$ est la loi marginale de X correspondant au couple $(\Pr(x|y, \mathcal{D}), \pi(y))$. Ce classifieur est alors défini, pour $\lambda > 0$ fixé, par

$$\phi(x|\lambda) = \begin{cases} 1 & \text{si } \Pr(Y = y|X = x, \mathcal{D})\pi(y) > \lambda \\ 0 & \text{sinon} \end{cases}$$

Il se pose alors la question de sélectionner un classifieur optimal sur la base d'un compromis sur des mesures de performance comme la sensibilité , la spécificité, le taux d'erreur, etc. La courbe ROC et la mesure AUC sont généralement utilisées pour réaliser cet objectif. Cette démarche peut être étendue aux méthodes d'agrégation de classifieur comme le boosting d'arbre binaire de classement ou le random forest. Généralement ces méthodes utilisent un seuil $\lambda = 0.5$ par défaut. Très souvent le classifieur $\phi(x|\lambda)$ associé au seuil $\lambda = 0.5$ ne fournit pas de meilleurs performances. Ainsi pour comparer notre méthode de classement à ces différentes méthodes, nous considérons la stratégie suivante :

1. Nous identifions le seuil optimal pour chaque méthode associant un score à une observation. C'est à dire le seuil qui produit le classifieur dont les mesures de performance fournit le meilleur compromis.
2. Nous comparons alors les classifieurs ainsi obtenus à notre classifieur. Les résultats obtenus sont présentés dans les tableaux ci-dessous.

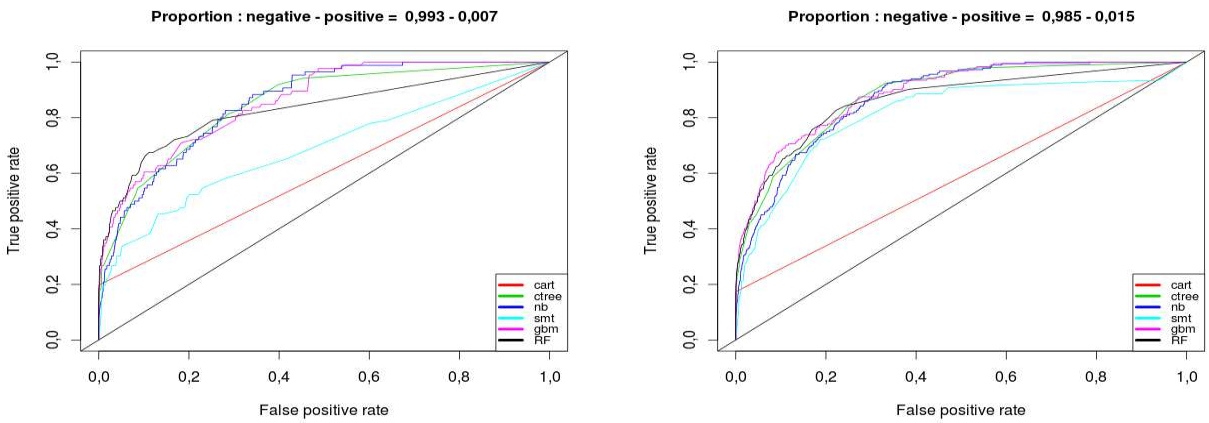
Les résultats présentés ci-dessous sont obtenus en utilisant le package **caret**[20] (**c**lassification and **re**gression **t**raining) dans l'environnement de programmation **R**. Ce dernier contient un riche ensemble de fonctions de modélisation à la fois pour la classification et la régression. Le package **caret** permet d'éliminer la différence syntaxique située entre un grand nombre d'algorithme pour la construction et la prédiction de modèles. Il contient un ensemble d'approches raisonnables semi-automatisées pour l'optimisation des valeurs des paramètres d'apprentissage. A l'aide du package **caret**, on peut donc trouver, pour la plus part des méthodes (classification ou régression), le classifieur optimal qui ajuste le mieux les données d'apprentissage grâce à sa fonction *train*. La fonction *train* est utilisée pour sélectionner les valeurs du(des) paramètre(s) d'apprentissage du modèle et/ou d'estimer les performances du modèle en utilisant une méthode d'échantillonnage. En utilisant une méthode d'échantillonnage telle que le bootstrap ou la validation croisée, un ensemble d'observations est simulé conditionnellement aux données d'apprentissage. A chaque ensemble échantillonné correspond un classifieur. Pour chaque combinaison de paramètres d'apprentissage candidats, un modèle est ajusté aux données échantillonnées et ensuite est utilisé pour la prédiction. La performance du modèle est estimée en agrégeant les prédictions du modèle sur les données échantillonnées. Ces performances estimées sont utilisées pour évaluer laquelle des combinaisons des paramètres d'apprentissage est appropriée. Pour des données de grande taille telles que les données "Adult Dataset" nous avons choisi la validation croisée comme méthode d'échantillonnage et pour les données de petite taille, par exemple les données "Credit Approval Dataset", nous avons utilisé le bootstrap comme méthode de ré-échantillonnage.

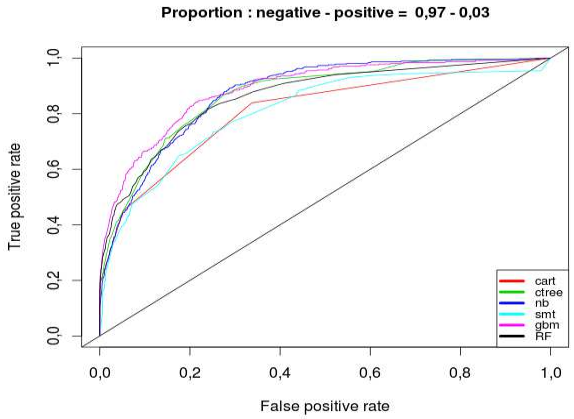
Le taux d'erreur de classement est la mesure de performance généralement associée aux algorithmes d'apprentissage automatique. Dans le contexte des ensembles de données symétriques et des ensembles de données avec des coûts de mauvais classement égaux, il est raisonnable d'utiliser le taux d'erreur comme mesure de performance. Par contre lorsque les données sont déséquilibrées ou lorsqu'elles sont associées à des coûts d'erreur inégaux, il est plus approprié d'utiliser la courbe ROC ou d'autres

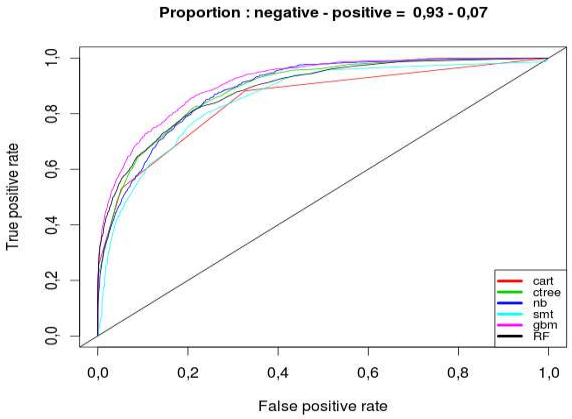
techniques similaires (Ling et Li, 1998 ; Drummond & Holte, 2000 ; Provost & Fawcett, 2001 ; Bradley, 1997 ; Turney 1996). L'aire sous la courbe ROC (AUC) est une mesure utile de la performance du classificateur car elle est indépendante du critère de décision choisi et aux changements de la distribution des classes [12]. La comparaison des AUC peut établir une relation de domination entre les classifieurs.
Le score de Pierce constitue aussi une mesure de performance conçue pour la prévision d'événements climatiques rares afin de pénaliser les modèles ne prévoyant jamais ces événements ou encore générant trop de fausses alertes. Le modèle idéal prévoit tous les événements rares sans fausse alerte. Le score de Pierce : $Sensibilité + Spécificité - 1$, compris entre -1 et 1, évalue la qualité d'un modèle de prévision. Si ce score est supérieur à 0, le taux de bonnes prévisions est supérieur à celui des fausses alertes et plus il est proche de 1, meilleur est le modèle.
Dans la suite, nous avons choisi de comparer notre méthode à des méthodes alternatives qui associent un score à chaque observation. Pour ces méthodes il est donc possible de construire leurs courbes ROC. Pour chaque méthode alternative, on peut produire un ensemble de classifieurs et puis sélectionner le classifieur le plus pertinent suivant un critère de sélection à l'aide de la fonction *train* du package **caret**. Dans cette analyse nous avons choisi la précision (taux de bien classés) comme critère de sélection. Par la suite, nous allons comparer les performances des meilleurs classifieurs sélectionnés avec les performances de notre classifeur. Les résultats sont présentés sous forme de tableaux.

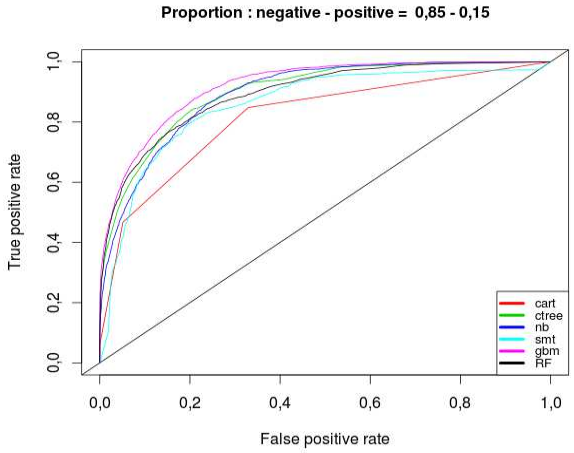
7.2.1 Données Adult Data Set

1. **Lorsque la distribution de l'échantillon test est identique à celui de l'échantillon d'apprentissage**









On peut constater à partir des graphes ci-dessus que lorsque la proportion d'observations positives devient de plus en plus grande, les courbes ROC se rapprochent de plus en plus.

Distributions "0" - "1"	ARM					CART					CTREE				
	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	0,700	0,779	0,222	0,740	0,479	0,105	1,000	0,007	0,552	0,105	0,593	0,878	0,124	0,736	0,471
0.985 - 0.015	0,671	0,821	0,181	0,746	0,492	0,152	1,000	0,014	0,576	0,152	0,793	0,777	0,222	0,785	0,570
0.970 - 0.030	0,644	0,807	0,198	0,726	0,451	0,450	0,948	0,067	0,699	0,398	0,812	0,766	0,232	0,789	0,578
0.930 - 0.070	0,754	0,799	0,204	0,776	0,553	0,530	0,948	0,083	0,739	0,478	0,797	0,805	0,196	0,801	0,602
0.850 - 0.150	0,791	0,774	0,223	0,782	0,565	0,467	0,949	0,127	0,708	0,416	0,813	0,818	0,183	0,815	0,631

Distributions "0" - "1"	ARM					Naive Bayes					SMOTE				
	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	0,700	0,779	0,222	0,74	0,479	0,814	0,719	0,279	0,766	0,533	0,547	0,770	0,231	0,659	0,317
0.985 - 0.015	0,671	0,821	0,181	0,746	0,492	0,799	0,761	0,238	0,780	0,560	0,696	0,821	0,181	0,758	0,517
0.970 - 0.030	0,644	0,807	0,198	0,726	0,451	0,842	0,750	0,247	0,796	0,592	0,716	0,755	0,247	0,736	0,471
0.930 - 0.070	0,754	0,799	0,204	0,776	0,553	0,850	0,760	0,233	0,805	0,610	0,783	0,772	0,227	0,778	0,555
0.850 - 0.150	0,791	0,774	0,223	0,782	0,565	0,832	0,785	0,207	0,808	0,617	0,805	0,792	0,207	0,798	0,597

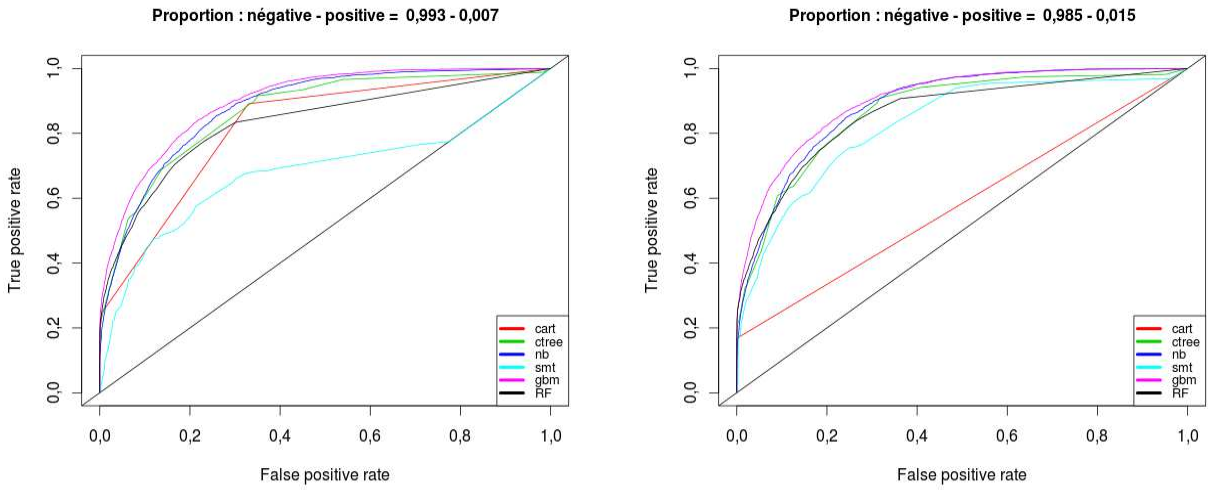
Distributionss "0" - "1"	ARM					Boosting					Random forests				
	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	0,700	0,779	0,222	0,740	0,479	0,756	0,793	0,208	0,774	0,549	0,698	0,851	0,150	0,774	0,549
0.985 - 0.015	0,671	0,821	0,181	0,746	0,492	0,766	0,814	0,187	0,790	0,580	0,799	0,794	0,205	0,796	0,593
0.970 - 0.030	0,644	0,807	0,198	0,726	0,451	0,823	0,801	0,199	0,812	0,624	0,791	0,780	0,220	0,786	0,571
0.930 - 0.070	0,754	0,799	0,204	0,776	0,553	0,842	0,804	0,193	0,823	0,646	0,804	0,794	0,204	0,799	0,598
0.850 - 0.150	0,791	0,774	0,223	0,782	0,565	0,836	0,828	0,171	0,832	0,664	0,780	0,833	0,176	0,806	0,613

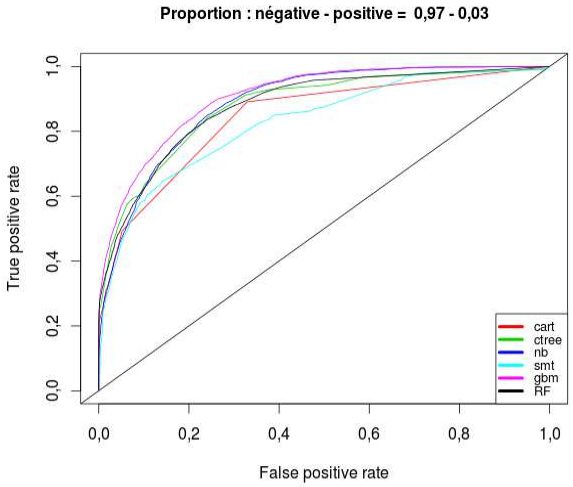
Tableau III.11 – Performances prédictives des méthodes alternatives

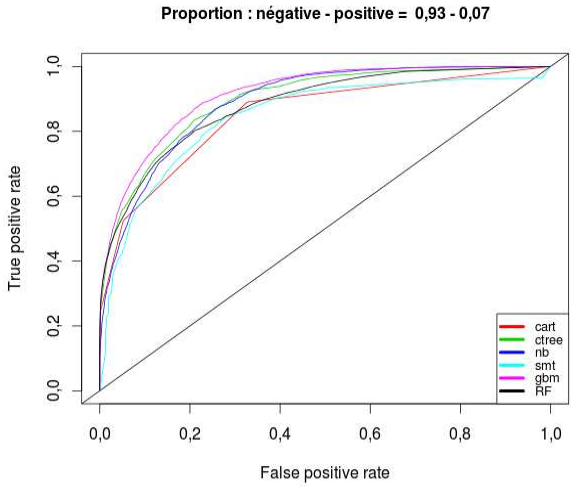
On peut constater que notre méthode d'apprentissage (ARM) est plus performante que la méthode CART. Du point de vue de l'aire en dessous de la courbe ROC (AUC) et du score de Pierce (PSS), la méthode ARM enregistre des valeurs largement au dessus des valeurs de la méthode CART. Elle produit également des sensibilités plus élevées variant entre 62% et 80% tandis que la méthode CART enregistre des sensibilités entre 10% et 50%. Par contre la méthode CART est plus spécifique (95%-100%) et admet des erreurs de classement plus faibles (7%-12%) contre (77%-81%) et (18%-22%) respectivement pour la méthode ARM.
Le classifieur naïf de Bayes, malgré qu'il produit des sensibilités, des AUC et des PSS plus élevés que ceux produits par la méthodes ARM, enregistre de forts taux d'erreurs de classement entre 21% et 28% avec des spécificités plus petites que celles de la méthodes ARM.
Les résultats présentés dans le tableau III.11 ci-dessus montrent une forte équivalence entre la méthode ARM et les méthodes SMOTE, Boosting et forêts aléatoires. Réputées d'être les meilleurs méthodes de classement en terme de performance, la méthode boosting et la méthode des forêts aléatoires présentent des performances sensiblement égales aux performances de la méthode ARM.

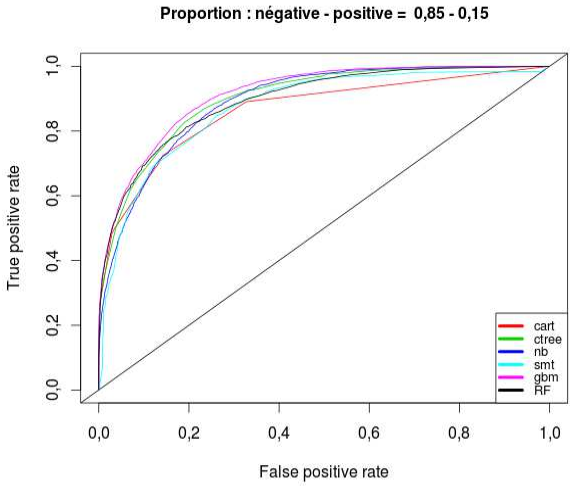
2. **Lorsque la distribution de l'échantillon test est différente de celui de l'échantillon d'apprentissage**

A ma connaissance, les performances d'un classifeur binaire sont généralement évaluées à partir d'un ensemble test dont la distribution est identique à celle de l'ensemble d'apprentissage qui a servis à construire le classifieur. Nous voulons évaluer les performances de la méthode d'apprentissage statistique et de les comparer avec les performances des méthodes alternatives lorsque la distribution de l'échantillon d'apprentissage est différente de la distribution de l'ensemble test.









De même on peut constater aussi, à partir des graphes ci-dessus, que lorsque la proportion d'observations positives devient de plus en plus grande, les courbes ROC se rapprochent de plus en plus.

Distributions "0" - "1"	ARM					CART					CTREE				
	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	0,729	0,763	0,245	0,746	0,492	0,248	0,995	0,189	0,621	0,243	0,555	0,922	0,168	0,738	0,477
0.985 - 0.015	0,594	0,866	0,201	0,730	0,46	0,168	0,999	0,205	0,584	0,167	0,637	0,874	0,184	0,756	0,511
0.970 - 0.030	0,697	0,750	0,263	0,724	0,447	0,493	0,948	0,164	0,720	0,441	0,840	0,761	0,220	0,800	0,601
0.930 - 0.070	0,752	0,800	0,212	0,776	0,552	0,525	0,948	0,156	0,736	0,473	0,811	0,804	0,194	0,808	0,615
0.850 - 0.150	0,754	0,799	0,212	0,776	0,553	0,724	0,858	0,175	0,791	0,582	0,819	0,816	0,183	0,817	0,635

Distributions "0" - "1"	ARM					Naive Bayes					SMOTE				
	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	0,729	0,763	0,245	0,746	0,492	0,814	0,775	0,216	0,794	0,589	0,649	0,705	0,309	0,677	0,354
0.985 - 0.015	0,594	0,866	0,201	0,730	0,460	0,829	0,773	0,213	0,801	0,602	0,728	0,776	0,236	0,752	0,504
0.970 - 0.030	0,697	0,750	0,263	0,724	0,447	0,831	0,776	0,211	0,804	0,607	0,649	0,855	0,196	0,752	0,504
0.930 - 0.070	0,752	0,800	0,212	0,776	0,552	0,835	0,770	0,214	0,802	0,605	0,793	0,768	0,226	0,780	0,561
0.850 - 0.150	0,754	0,799	0,212	0,776	0,553	0,825	0,784	0,206	0,804	0,609	0,825	0,754	0,229	0,790	0,579

Distributions "0" - "1"	ARM					Boosting					Random forests				
	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	0,729	0,763	0,245	0,746	0,492	0,806	0,807	0,193	0,806	0,613	0,733	0,809	0,210	0,771	0,542
0.985 - 0.015	0,594	0,866	0,201	0,730	0,460	0,820	0,807	0,190	0,814	0,627	0,793	0,775	0,220	0,784	0,568
0.970 - 0.030	0,697	0,750	0,263	0,724	0,447	0,823	0,812	0,185	0,818	0,635	0,800	0,794	0,205	0,797	0,594
0.930 - 0.070	0,752	0,800	0,212	0,776	0,552	0,839	0,817	0,177	0,828	0,656	0,788	0,800	0,203	0,794	0,588
0.850 - 0.150	0,754	0,799	0,212	0,776	0,553	0,831	0,832	0,169	0,831	0,663	0,808	0,808	0,191	0,808	0,616

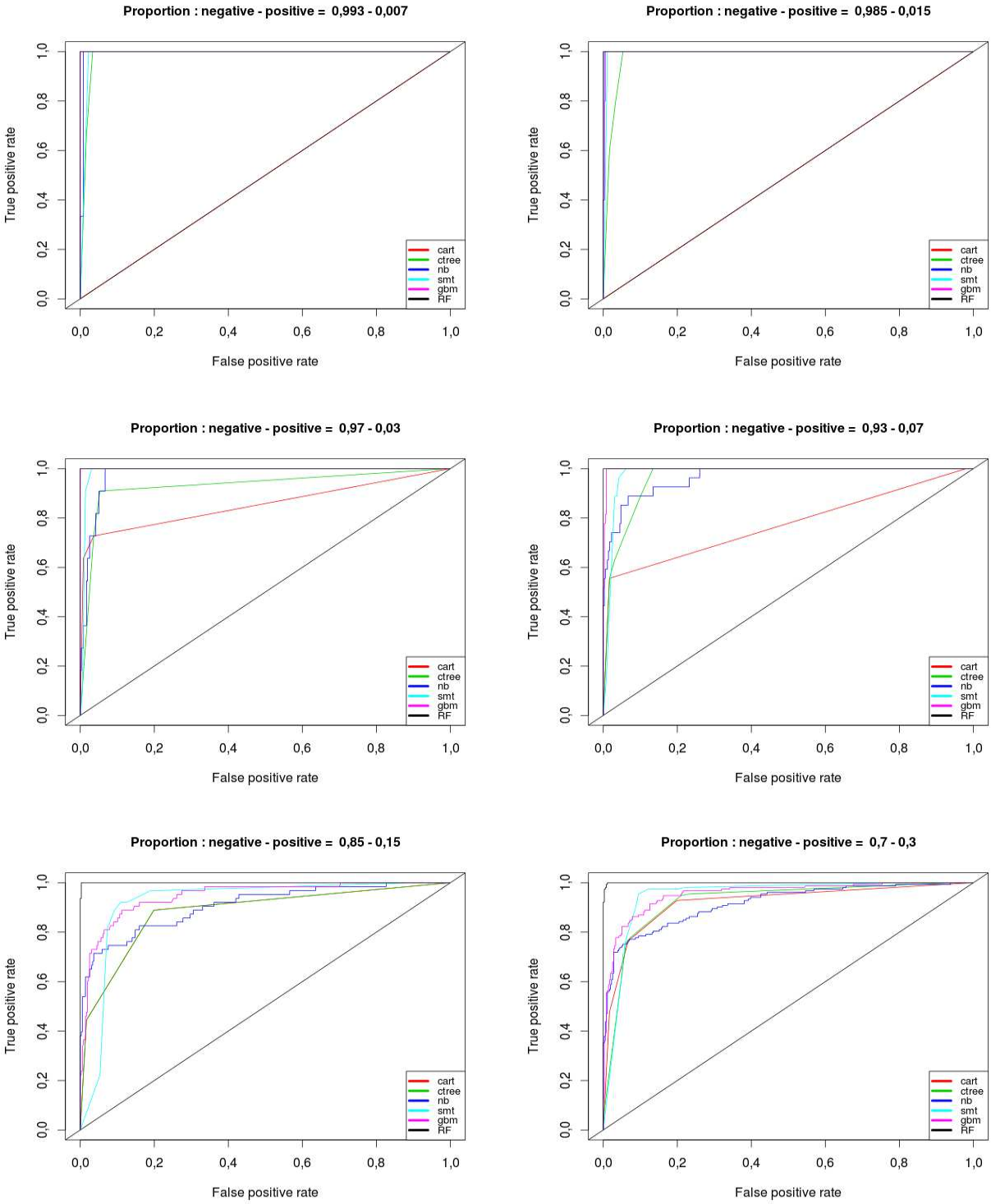
Tableau III.12 – Performances prédictives des méthodes alternatives

Ici aussi on obtient des résultats analogiques aux résultats obtenus lorsque la distribution de l'ensemble d'apprentissage est identique à la distribution de l'ensemble test. On observe que la méthode d'apprentissage ARM est plus performante que la méthode CART. Du point de vue de l'aire en dessous de la courbe ROC (AUC) et du score de Pierce (PSS), la méthode ARM enregistre des valeurs largement au dessus des valeurs de la méthode CART. Elle produit également des sensibilités plus élevées variant entre 59% et 75% tandis que la méthode CART enregistre des sensibilités entre 16% et 72%. Par contre la méthode CART est plus spécifique et admet des erreurs de classement plus faibles sur tous les échantillons simulés.
Dans le cas où la distribution d'apprentissage est différente de la distribution test, les indices de performances du classifieur naïf de Bayes sont meilleurs que les indices de performance de la méthode d'apprentissage ARM sur tous les échantillons simulés sauf au niveau de la spécificité où on a enregistré des taux sensiblement égaux. On peut constater aussi que la méthode Boosting domine largement la méthode ARM sur tous les échantillons en plus elle enregistre des taux d'erreur inférieurs à 20% des scores de Pierce supérieurs à 61% . Tandis que la méthode des forêts aléatoires enregistre des taux d'erreurs inférieurs à 22% et des scores de Pierce compris entre $54 - 61\%$. Là où la méthode ARM enregistre des taux d'erreurs supérieurs à 20% et des scores de Pierce inférieurs à 55%.

– ARM : Association Rules Mining ; CART : Classification And Regression Tree ; CTREE : Conditional tree ; Naive Bayes : Naive Bayes Classifier ; SMOTE : Synthetic Minority Oversampling Technique,

7.2.2 Données Credit Approval Data Set

Le jeu de données "credit approval" concerne des demandes de carte de crédit [24]. Tous les noms et valeurs des variables ont été modifiés pour protéger la confidentialité des données. Les données contiennent au total 690 observations incluant les données manquantes. Elles sont constituées d'un mélange de 6 variables numériques, de 9 variables non-numériques et d'une variable réponse binaire ("+","-"). L'objectif visé dans cette analyse est de trouver un profil prédictif d'approbation d'une carte crédit à un sujet donné.



On constate également que lorsque la proportion d'observations positives devient de plus en plus grande, les courbes ROC se rapprochent de plus en plus.

Distributions	ARM					CART					CTREE				
"-" "+"	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	1.000	0.852	0.147	0.926	0.852	-	-	-	-	-	1.000	0,966	0,033	0,983	0,966
0.985 - 0.015	1.000	0.832	0.166	0.916	0.832	-	-	-	-	-	1.000	0,947	0,052	0,974	0,947
0.970 - 0.030	0.909	0.714	0.280	0.811	0.632	0,727	0,964	0,043	0,845	0,691	0,909	0,947	0,055	0,928	0,856
0.930 - 0.070	0.889	0.818	0.177	0.853	0.707	0,556	0,983	0,047	0,770	0,539	1,000	0,866	0,125	0,933	0,866
0.850 - 0.150	0.857	0.765	0.221	0.811	0.622	0,889	0,801	0,186	0,845	0,690	0,889	0,801	0,186	0,845	0,690
0.700 - 0.300	0.935	0.625	0.283	0.780	0.560	0,928	0,801	0,161	0,864	0,729	0,948	0,790	0,163	0,869	0,738

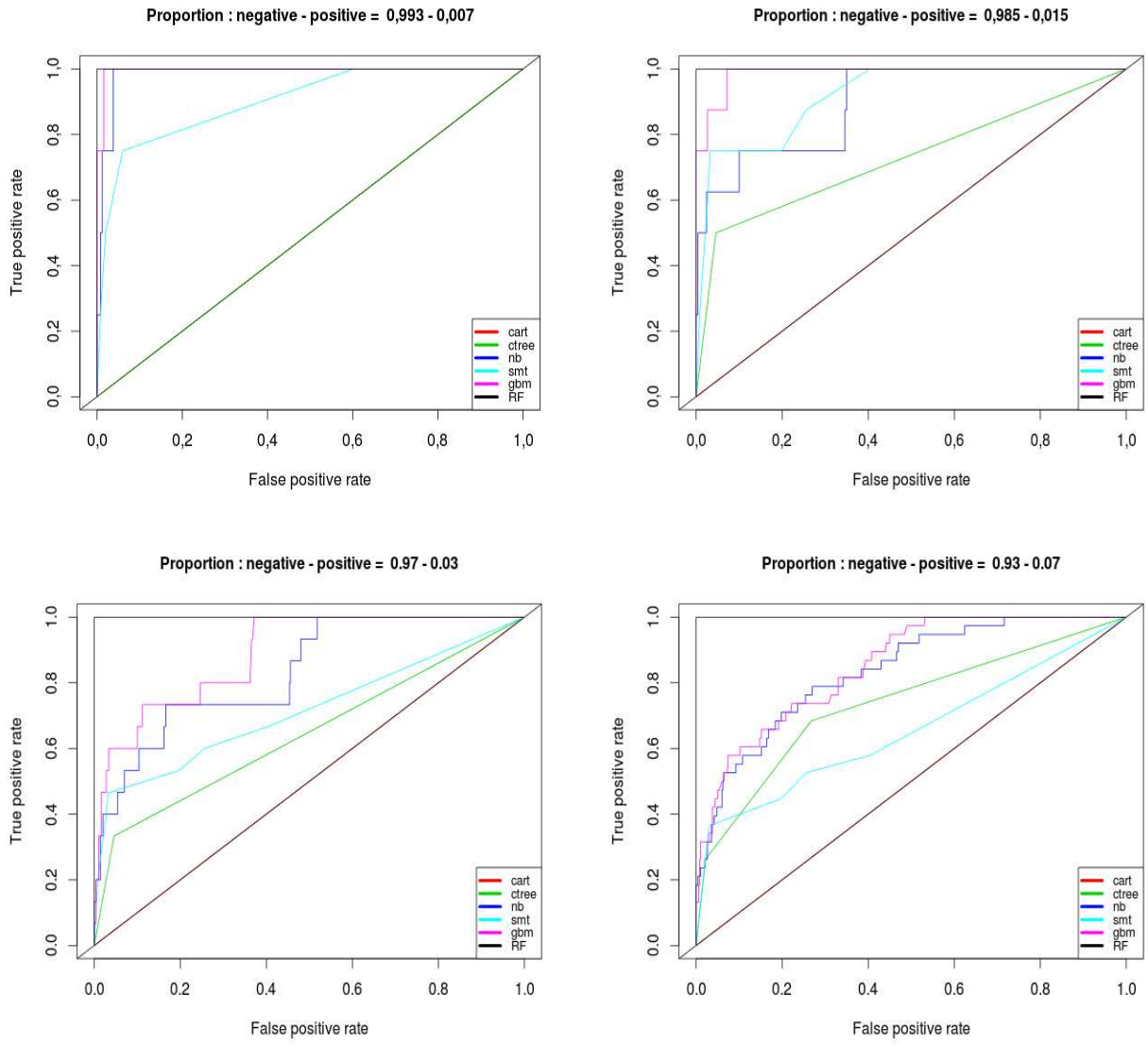
Distributions	ARM					Naive Bayes					SMOTE				
"-" "+"	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	1.000	0.852	0.147	0.926	0.852	1.000	0,992	0,008	0,996	0,992	1.000	0,992	0,008	0,996	0,992
0.985 - 0.015	1.000	0.832	0.166	0.916	0.832	1.000	0,997	0,003	0,998	0,997	1.000	0,997	0,003	0,998	0,997
0.970 - 0.030	0.909	0.714	0.280	0.811	0.632	1,000	0,933	0,065	0,966	0,933	1,000	0,933	0,065	0,966	0,933
0.930 - 0.070	0.889	0.818	0.177	0.853	0.707	0,889	0,933	0,070	0,911	0,822	0,889	0,933	0,070	0,911	0,822
0.850 - 0.150	0.857	0.765	0.221	0.811	0.622	0,825	0,840	0,162	0,832	0,665	0,825	0,840	0,162	0,832	0,665
0.700 - 0.300	0.935	0.625	0.283	0.780	0.560	0,784	0,905	0,132	0,844	0,689	0,784	0,905	0,132	0,844	0,689

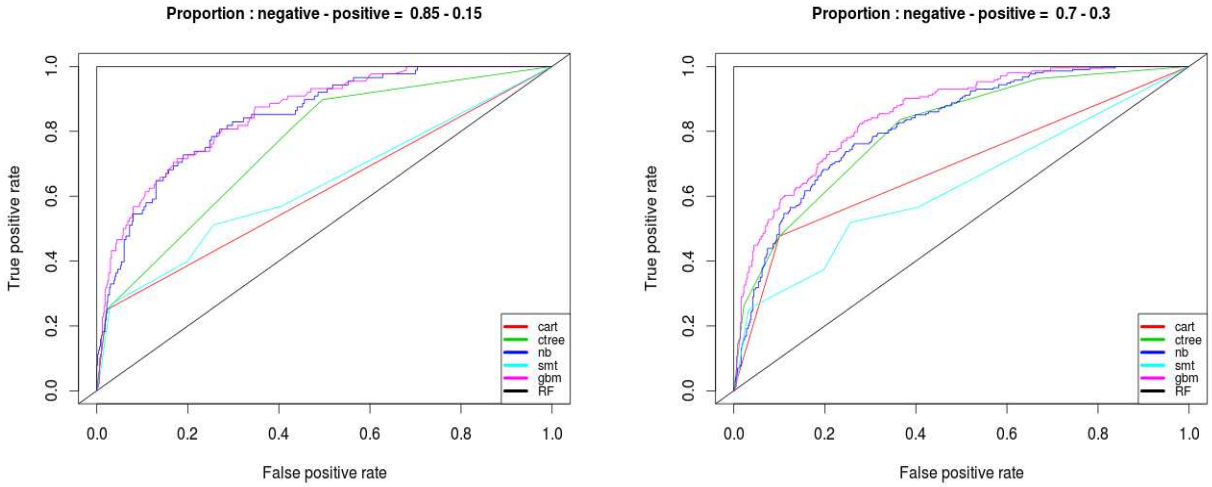
Distrbutions	ARM					Boosting					Random Forests				
"-" "+"	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	1.000	0.852	0.147	0.926	0.852	1.000	1,000	0,000	1,000	1,000	1.000	1,000	0,000	1,000	1,000
0.985 - 0.015	1.000	0.832	0.166	0.916	0.832	1.000	0,994	0,006	0,997	0,994	1.000	1,000	0,000	1,000	1,000
0.970 - 0.030	0.909	0.714	0.280	0.811	0.632	1,000	1,000	0,000	1,000	1,000	1,000	1,000	0,000	1,000	1,000
0.930 - 0.070	0.889	0.818	0.177	0.853	0.707	1,000	0,992	0,008	0,996	0,992	1,000	1,000	0,000	1,000	1,000
0.850 - 0.150	0.857	0.765	0.221	0.811	0.622	0,889	0,888	0,112	0,889	0,777	1,000	0,997	0,002	0,998	0,997
0.700 - 0.300	0.935	0.625	0.283	0.780	0.560	0,915	0,874	0,113	0,895	0,789	0,993	0,992	0,008	0,992	0,985

Tableau III.13 – Performances prédictives des méthodes alternatives par bootstrap

7.2.3 Données Pima Indians Diabetes Data Set

Le jeu de données "pima-indians-diabetes" est constitué par des femmes d'au moins 21 ans d'origine indienne Pima auxquelles on a administré un test pour le diabète [27]. L'échantillon est constitué de 8 variables numériques et d'une variable réponse binaire qui prend la valeur 1 si le test est positif. Il contient au total 768 observations. L'objectif de l'analyse est de déterminer si oui ou non la patiente présente des signes de diabète selon les normes de l'organisation mondiale de la santé.





On constate de même que lorsque la proportion d'observations positives devient de plus en plus grande, les courbes ROC se rapprochent de plus en plus.

Distributions		ARM					CART					CTREE				
"-"	"+"	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993	0.007	1.000	0.824	0.175	0.912	0.824	-	-	-	-	-	-	-	-	-	-
0.985	- 0.015	0.875	0.846	0.154	0.860	0.721	-	-	-	-	-	0,50	0,954	0,053	0,727	0,454
0.970	- 0.030	0.933	0.786	0.210	0.859	0.719	-	-	-	-	-	0,333	0,954	0,064	0,644	0,287
0.930	- 0.070	0.711	0.782	0.223	0.746	0.492	-	-	-	-	-	0,684	0,732	0,271	0,708	0,416
0.850	- 0.150	0.784	0.602	0.370	0.693	0.482	0,250	0,978	0,131	0,614	0,228	0,807	0,574	0,391	0,690	0,381
0.700	- 0.300	0.785	0.682	0.287	0.733	0.466	0,477	0,900	0,227	0,688	0,377	0,836	0,634	0,305	0,735	0,470

Distributions		ARM					Naive Bayes					SMOTE				
"-"	"+"	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993	- 0.007	1.000	0.824	0.175	0.912	0.824	1.000	0,962	0,038	0,981	0,962	1.000	0,962	0,038	0,981	0,962
0.985	- 0.015	0.875	0.846	0.154	0.860	0.721	0,750	0,900	0,102	0,825	0,650	0,750	0,900	0,102	0,825	0,650
0.970	- 0.030	0.933	0.786	0.210	0.859	0.719	0,733	0,834	0,169	0,784	0,567	0,733	0,834	0,169	0,784	0,567
0.930	- 0.070	0.711	0.782	0.223	0.746	0.492	0,789	0,730	0,266	0,760	0,519	0,789	0,730	0,266	0,760	0,519
0.850	- 0.150	0.784	0.602	0.370	0.693	0.482	0,784	0,748	0,246	0,766	0,532	0,784	0,748	0,246	0,766	0,532
0.700	- 0.300	0.785	0.682	0.287	0.733	0.466	0,762	0,736	0,256	0,749	0,498	0,762	0,736	0,256	0,749	0,498

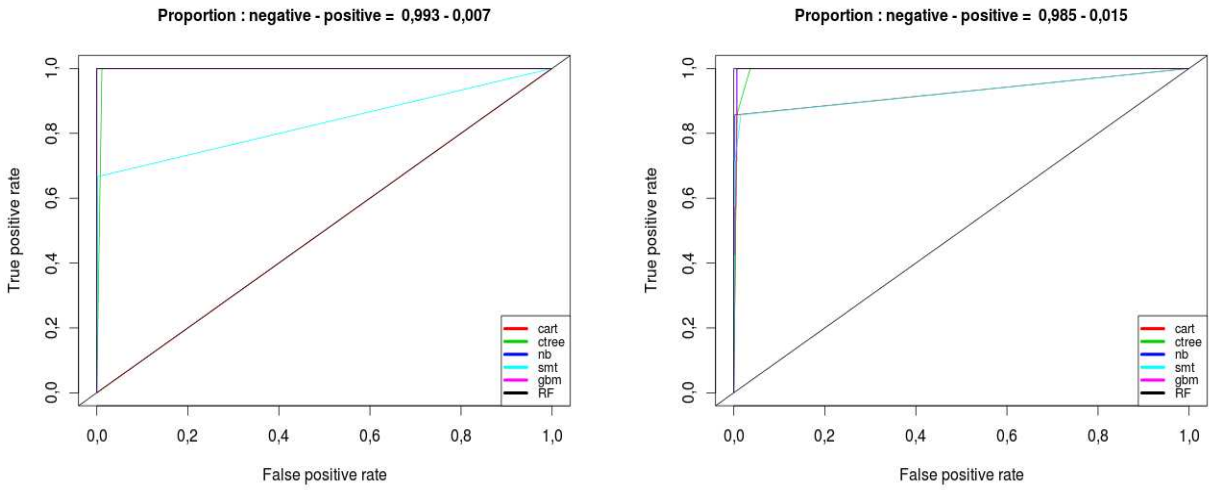
Distributions		ARM					Boosting					Random Forests				
"-"	"+"	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993	- 0.007	1.000	0.824	0.175	0.912	0.824	1.000	0,984	0,016	0,992	0,984	1.000	1,000	0,000	1,000	1,000
0.985	- 0.015	0.875	0.846	0.154	0.860	0.721	1,000	0,928	0,071	0,964	0,928	1,000	1,000	0,000	1,000	1,000
0.970	- 0.030	0.933	0.786	0.210	0.859	0.719	0,800	0,758	0,241	0,779	0,558	1,000	1,000	0,000	1,000	1,000
0.930	- 0.070	0.711	0.782	0.223	0.746	0.492	0,737	0,778	0,225	0,758	0,515	1,000	1,000	0,000	1,000	1,000
0.850	- 0.150	0.784	0.602	0.370	0.693	0.482	0,716	0,824	0,193	0,770	0,540	1,000	1,000	0,000	1,000	1,000
0.700	- 0.300	0.785	0.682	0.287	0.733	0.466	0,822	0,724	0,246	0,773	0,546	1,000	1,000	0,000	1,000	1,000

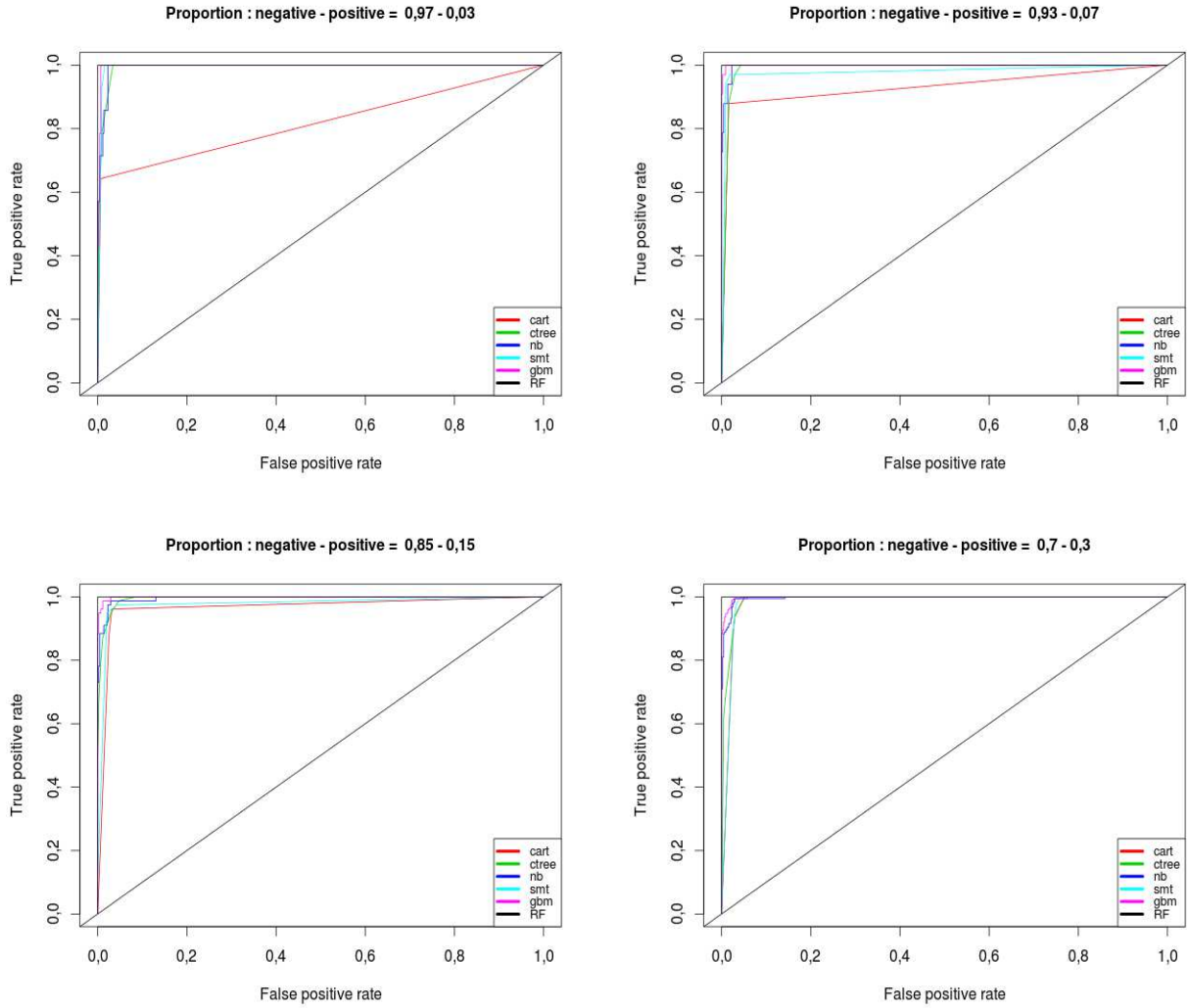
Tableau III.14 – Performances prédictives des méthodes alternatives par bootstrap

Les résultats obtenus à partir des données "Pima Indians Diabetes Dataset" et "Credit Approval Dataset" montrent que, même en présence d'un jeux de données de petite taille, la méthode ARM reste toujours meilleur que la méthode CART de même que la méthode CTREE dont, pour les données "Credit Approval Dataset", les taux d'erreur peuvent aller jusqu'à 39% et les scores de Pierce inférieurs à 47% tandis que la méthode ARM enregistre des scores supérieurs à 47%. Ce pendant elles enregistrent des scores de même ordre de grandeur pour les données "Pima Indians Diabetes Dataset" mais avec des taux d'erreur plus élevés pour la méthode ARM. Il faut noter aussi que pour les deux jeux de données les indicateurs de performance (sensibilité, spécificité, AUC et PSS) décroissent et le taux d'erreur croît au fure et à mesure que la proportion d'observations positives augmente.

7.2.4 Données Breast Cancer Data Set

Les données obtenues à partir du diagnostique de Wisconsin du cancer du sein (WDBC), fourni par le Centre Hospitalier Universitaire de Wisconsin, a été dérivé d'un groupe d'images par aspiration à l'aiguille fine (FNA) de la poitrine [21]. Une programmation génétique avec différentes tailles de la population a été utilisée pour cette étude. L'objectif est d'identifier la classe "benign" ou "malignant" de chaque numéro. Les échantillons arrivent périodiquement comme le Dr Wolberg rapporte ses cas cliniques. La base de données reflète donc ce regroupement chronologique des données. Chaque variable à l'exception de la première a été convertie en 11 attributs numériques primitifs avec des valeurs allant de 0 à 10. Il y a 16 valeurs manquantes. Les données contiennent 699 observations sur 11 variables, l'une étant une variable de caractère, 9 étant ordonnées ou nominales, et une classe cible.





On peut remarquer également que lorsque la proportion d'observations positives devient de plus en plus grande, les courbes ROC se rapprochent de plus en plus.

Distributions	ARM					CART					CTREE				
"-" "+"	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	1,000	0,964	0,036	0,982	0,964	-	-	-	-	-	1.000	0,989	0,011	0,994	0,989
0.985 - 0.015	1,000	0,908	0,091	0,954	0,908	0,857	0,993	0,009	0,925	0,850	1.000	0,964	0,035	0,982	0,964
0.970 - 0.030	1,000	0,883	0,114	0,942	0,883	0,643	0,993	0,018	0,818	0,636	1.000	0,966	0,033	0,983	0,966
0.930 - 0.070	1,000	0,858	0,132	0,929	0,858	0,879	0,984	0,023	0,932	0,863	0,97	0,971	0,029	0,970	0,941
0.850 - 0.150	0,987	0,858	0,123	0,922	0,845	0,962	0,968	0,033	0,965	0,930	0,987	0,953	0,042	0,970	0,940
0.700 - 0.300	0,995	0,858	0,101	0,926	0,853	1,000	0,948	0,036	0,974	0,948	1,000	0,948	0,036	0,974	0,948

Distributions	ARM					Naive Bayes					SMOTE				
"-" "+"	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0,993 - 0,007	1,000	0,964	0,036	0,982	0,964	1,000	1,000	0,000	1,000	1,000	1,000	1,000	0,000	1,000	1,000
0,985 - 0,015	1,000	0,908	0,091	0,954	0,908	1,000	0,993	0,007	0,996	0,993	1,000	0,993	0,007	0,996	0,993
0,970 - 0,030	1,000	0,883	0,114	0,942	0,883	1,000	0,977	0,022	0,988	0,977	1,000	0,977	0,022	0,988	0,977
0,930 - 0,070	1,000	0,858	0,132	0,929	0,858	1,000	0,977	0,021	0,988	0,977	1,000	0,977	0,021	0,988	0,977
0,850 - 0,150	0,987	0,858	0,123	0,922	0,845	0,987	0,971	0,027	0,979	0,958	0,987	0,971	0,027	0,979	0,958
0,700 - 0,300	0,995	0,858	0,101	0,926	0,853	0,995	0,971	0,023	0,983	0,966	0,995	0,971	0,023	0,983	0,966

Distributions	ARM					Boosting					Random Forests				
"-" "+"	sensib	spécf	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss	sensib	spécif	err.cl	auc	pss
0.993 - 0.007	1,000	0,964	0,036	0,982	0,964	1,000	1,000	0,000	1,000	1,000	1,000	0,000	1,000	1,000	1,000
0.985 - 0.015	1,000	0,908	0,091	0,954	0,908	1,000	0,993	0,007	0,996	0,993	1,000	1,000	0,000	1,000	1,000
0.970 - 0.030	1,000	0,883	0,114	0,942	0,883	1,000	0,993	0,007	0,996	0,993	1,000	1,000	0,000	1,000	1,000
0.930 - 0.070	1,000	0,858	0,132	0,929	0,858	1,000	0,991	0,008	0,996	0,991	1,000	1,000	0,000	1,000	1,000
0.850 - 0.150	0,987	0,858	0,123	0,922	0,845	0,987	0,989	0,012	0,988	0,976	1,000	1,000	0,000	1,000	1,000
0.700 - 0.300	0,995	0,858	0,101	0,926	0,853	0,989	0,977	0,019	0,983	0,966	1,000	1,000	0,000	1,000	1,000

Tableau III.15 – Performances prédictives des méthodes alternatives à partir de 20 échantillons bootstrap

Les résultats obtenus à partir des données "Breast Cancer Dataset" confirment donc que en présence de données de petite taille et déséquilibrées, la méthode ARM domine la méthode CART et enregistre des performances sensiblement équivalentes aux performances obtenues à partir des méthodes de classement telles que la méthode Boosting et la méthode des forêts aléatoires.

Il ressort de cette analyse que notre méthode d'apprentissage est largement plus performante que la méthode CART. Ce pendant elle est comparable à la méthode CTREE, le classifieur naif de Bayes, la méthode SMOTE, le boosting d'arbres de classement et la méthode random forest. Du point de vue de la sensibilité, de la spécificité, de l'aire en dessous de la courbe ROC et du score de Pierce, notre méthode d'apprentissage à les même ordres de valeur que les méthodes citées précédemment. Par contre elle enregistre une erreur de classement supérieur à celles des autres méthodes de l'ordre de 10^{-1} à 10^{-2}.
Par ailleurs, on peut remarquer que si CART et CTREE permettent de fournir un outil d'aide à la décision (arbre de décision) permettant de visualiser des profils pertinents cela n'est pas le cas des méthodes comme le boosting et les forêts aléatoires qui parfois ont des performances supérieures à ceux obtenues par la méthode d'apprentissage étudiée dans la thèse. D'où l'avantage de cette dernière sur les autres car elle permet d'avoir des performances sensiblement égales aux méthodes comme le boosting et les forêts aléatoires mais aussi elle permet de visualiser les profils les plus pertinents pour construire une règle de classement.

8 Conclusion

La procédure permet de surmonter l'impuissance des méthodes de régression qui sous-estiment les probabilités conditionnelles de l'apparition de la classe cible lorsque la fréquence des instances qui appartiennent à cette classe est très faible. De plus les interactions d'attributs qui sont fortement corrélées avec la classe cible sont spécifiées, ainsi la fonction de classification n'apparaît pas comme une boîte noire. Néanmoins il faut remarquer qu'une étape de prétraitement des données est nécessaire avant d'effectuer la procédure car il est supposé que les variables soient évaluées sur une échelle non numérique.

Bibliographie

[1] Agrawal, R., and Srikant, R. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases* (San Francisco, CA, USA, 1994), VLDB '94, Morgan Kaufmann Publishers Inc., pp. 487–499. 66

[2] An, A., and Cercone, N. Discretization of continuous attributes for learning classification rules. In *Methodologies for Knowledge Discovery and Data Mining*, N. Zhong and L. Zhou, Eds., no. 1574 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1999, pp. 509–514. 52

[3] Bache, K., and Lichman, M. UCI machine learning repository [http ://archive.ics.uci.edu/ml]. Tech. rep., University of California, Irvine, School of Information and Computer Sciences, 2013. 66

[4] Becher, H., Hall, P., and Wilson, S. R. Bootstrap hypothesis testing procedures. *Biometrics 49*, 4 (1993), 1268–1272. 63

[5] Besse, P. Apprentissage statistique & data mining. [*http : //www.math.u − psud.fr/ stafav/IMG/pdf/Apprens*$_{t}$*at.pdf*], Mis en ligne Octobre 2006, consulté en Décembre 2013. 50

[6] Breiman, L. *Classification and regression trees*. Chapman and Hall, 1984. 52

[7] Breiman, L. Bagging predictors. *Machine Learning 24*, 2 (1996), 123–140. 52

[8] Chan, C.-C., Batur, C., and Srinivasan, A. Determination of quantization intervals in rule based model for dynamic systems. In *, 1991 IEEE International Conference on Systems, Man, and Cybernetics, 1991. 'Decision Aiding for Complex Systems, Conference Proceedings* (1991), vol. 3, pp. 1719–1723. 52

[9] Efron, B. *The jackknife, the bootstrap, and other resampling plans.* Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1982. 63

[10] Efron, B., and Tibshirani, R. J. *An Introduction to the Bootstrap.* Taylor & Francis, 1994. 63

[11] Fawcett, T. Using rule sets to maximize ROC performance. IEEE Computer Society, pp. 131–138. 49

[12] Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett. 27*, 8 (2006), 861–874. 78

[13] Fayyad, U. M., and Irani, K. B. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning 8* (1992), 87–102. 52

[14] Fayyad, U. M., and Irani, K. B. Multi-interval discretization of continuous-valued attributes for classification learning. *Artificial Intelligence 13* (1993), 1022–1027. 52

[15] Gelfand, S., Ravishankar, C. S., and Delp, E. An iterative growing and pruning algorithm for classification tree design. In *, IEEE International Conference on Systems, Man and Cybernetics, 1989. Conference Proceedings* (1989), vol. 2, pp. 818–823. 52

[16] Hall, P., and Wilson, S. R. Two guidelines for bootstrap hypothesis testing. *Biometrics 47*, 2 (1991), 757–762. 63

[17] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. *SIGMOD Rec. 29*, 2 (2000), 1–12. 53

[18] Hothorn, T., and Zeileis, A. *partykit : A Toolkit for Recursive Partytioning*, 2013. R package version 0.1-6. 66

[19] Kohavi, R. Scaling up the accuracy of naive-bayes classifiers : a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), p. to appear. 66

[20] Kuhn, M. Building predictive models in r using the caret package. *Journal of Statistical Software 28*, 05 (2008). 77

[21] Mangasarian, O., and Wolberg, W. Cancer diagnosis via linear programming. *SIAM News 23*, 5 (1990), 1–18. 90

[22] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. *e1071 : Misc Functions of the Department of Statistics (e1071), TU Wien*, 2012. R package version 1.6-1. 66

[23] Quinlan, J. R. Induction of decision trees. *Machine Learning 1*, 1 (1986), 81–106. 52

[24] Quinlan, J. R. Simplifying decision trees. *Int. J. Hum.-Comput. Stud 51* (1999), 497. 84

[25] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. 66

[26] Rossi, F. *Introduction à l'apprentissage statistique*. Projet AxIS, INRIA Paris Rocquencourt, 2008. 50

[27] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., and Johannes, R. S. Using the adap learning algorithm to forecast the onset of diabetes mellitus. *Johns Hopkins APL Technical Digest 10* (1988), 262–266. 87

[28] Therneau, T., Atkinson, B., and Ripley, B. *rpart : Recursive Partitioning*, 2013. R package version 4.1-3. 66

[29] Torgo, L. *Data Mining with R, learning with case studies*. Chapman and Hall/CRC, 2010. 66

[30] Wong, A. K. C., and Chiu, D. Synthesizing statistical knowledge from incomplete mixed-mode data. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-9*, 6 (1987), 796–805. 52

Appendices

Annexe B

Annexe Chapitre III

A.1 Discrétisation par la Méthode de largeur d'intervalle égale

La technique de discrétisation la plus simple est celle dont le domaine de la variable discrétisée est le plus petit possible. i.e., $|\pi_A|$=2. La plus simple discrétisation, pas nécessairement la meilleure, est la discrétisation binaire. Bien qu'il existe une infinité de discrétisations binaires pour n'importe quel intervalle, toute variable numérique dans un ensemble de n observations peut seulement prendre au plus n valeurs distinctes. Ainsi, au plus $n-1$ discrétisations binaires sont pratiquement possibles. La méthode la plus simple pour discrétiser une variable numérique consiste à partitionner son domaine en intervalles de largeur égale. On l'appelle la Méthode de largeur d'intervalle égal (Equal Interval Width Method).

A.2 Discrétisation par la méthode par intervalle de fréquence égale

Plusieurs algorithmes de discrétisation basés sur la méthode de largeur d'intervalle égale (Equal Interval Width Method) ou sur la méthode par intervalle de fréquence égale (Equal Frequency per Interval Method) ont été étudiés dans plusieurs papiers. Parmi ces derniers on peut citer l'algorithme ChiMerge de Kerber [**?**] qui utilise la statistique du χ^2 pour discrétiser une variable numérique. On peut citer aussi l'algorithme $Chi2$ [**?**] qui est une amélioration de l'algorithme ChiMerger sur le choix du critère d'arrêt α.

A.3 Discrétisation par la mesure de l'entropie

Supposons que nous avons un ensemble S de N observations. Pour discrétiser une variable numérique A, nous choisissons le « meilleur » point de coupure T_A de son domaine de définition en évaluant tous les points de coupure candidats. Premièrement il faut ordonner les observations dans l'ordre croissant des valeurs de la variable A et le point milieu entre chaque paire successive d'observations dans la séquence ordonnée est considéré comme un point de coupure potentiel. Ainsi pour chaque variable

numérique, on aura $N-1$ points de coupure potentiels (si on suppose que les observations n'ont pas des valeurs de A identiques). Pour chaque point de coupure T, les données sont partitionnées en deux ensembles et l'entropie de la partition obtenue peut être alors calculée.

L'ensemble S peut être vu comme un ensemble d'événements réalisé par une ou plusieurs variables. A chaque événement E_i est associée une probabilité $P(E_i, S)$. En général ces probabilités sont non-uniformes, à l'événement E_i on associe la probabilité $P(E_i, S)$, mais de somme égale à 1 car toutes les réalisations possibles sont prises en comptes. La quantité d'information I_i d'un événement simple E_i est définie comme le logarithme de base 2 de la probabilité de l'événement $P(E_i, S)$:

$$I_i = log_2 P(E_i, S)$$

L'entropie $Ent(E_i, S)$ de l'événement E_i est l'opposé de I_i ($Ent(E_i, S) = -I_i$). L'entropie peut être vue comme l'"incertitude". Obtenir une quantité d'informations d'un événement c'est perdre la même quantité d'incertitudes de l'événement, ainsi I_i et $Ent(E_i, S)$ ne différent que par le signe. Par définition I_i est toujours négative. Elle varie entre $-\infty$ et 0 puisque $P(E_i, S)$ est une probabilité. Intuitivement, plus l'événement est improbable, plus l'incertitude augmente. A partir de la définition précédente, on peut alors définir l'entropie d'un ensemble d'événements. L'entropie d'un ensemble S est l'entropie moyenne de tous les événements de l'ensemble. Elle est calculée en pondérant chacune des entropies $Ent(E_i, S)$ par la probabilité $P(E_i, S)$ de l'événement.

$$Ent(S) = -\sum_i P(E_i, S) Ent(E_i, S) = -\sum_i P(E_i, S) log_2 P(E_i, S)$$

Le choix de la mesure logarithmique est justifié par le désir d'une entropie additive. Nous voulons que l'algèbre de notre mesure reflète les règles de probabilité. C'est à dire que lorsque nous recevons un ensemble d'événements indépendants, nous aimerions pouvoir dire que l'entropie totale reçue est la somme des entropies individuelles. Mais la probabilité conjointe d'événements indépendants est le produit des probabilités des événements, et donc nous devons prendre le logarithme afin que la probabilité conjointe des événements indépendants puisse contribuer de façon additive à l'entropie acquise.

A.4 Discrétisation par la méthode MDLP

Ici les événements d'intérêt sont spécialement les classes des observations d'un ensemble S. Supposons qu'il y ait k classes : $C_1, \ldots, C_k$ et notons par $P(C_i, S)$ la proportion d'observation dans S de classe C_i. Pour calculer l'entropie d'une classe donnée après que l'ensemble S est partitionné en deux sous-ensembles S_1 et S_2, nous prenons la moyenne pondérée des entropies des partitions.

Définition 6. Pour un ensemble S d'observations, une variable A, et une valeur de coupure T. Supposons $S_1 \subset S$ le sous-ensemble des observations dans S dont les valeurs correspondantes de A

sont plus petites que T et $S_2 = S - S_1$. L'entropie de la partition indicée par T, notée par $E(A, T, S)$, est définie par

$$E(A, T, S) = \frac{|S_1|}{N} Ent(S_1) + \frac{|S_2|}{N} Ent(S_2)$$

où $N = |S|$ est le nombre d'observations dans l'ensemble S.

Le meilleur point de coupure parmi tous les points de coupure candidats est le point de coupure T_A pour lequel

$$T_A = \underset{T}{\operatorname{argmim}}\, E(A, T, S)$$

Ceci détermine une discrétisation binaire de la variable A. Fayyad et Irani [**?**] ont montré que la valeur T_A de la variable A qui minimise la classe-entropie $E(A, T_A, S)$ pour un ensemble d'apprentissage S doit toujours être une valeur (une borne) entre deux observations de classes différentes dans la séquence des observations ordonnées. L'ensemble S est alors subdivisé en deux sous-ensembles par le point de coupure T_A. Une suite de points de coupure est obtenue en appliquant de manière récursive la même méthode de discrétisation binaire pour chacun des sous-ensembles nouvellement produits jusqu'à ce que la condition suivante soit réalisée :

$$Gaint(A, T, S) \leq \frac{log_2(N-1)}{N} + \frac{\Delta(A, T; S)}{N}$$

où $Gaint(A, T; S) = Ent(S) - E(A, T; S)$, $\Delta(A, T; S) = log_2(3^k - 2) - [k_1 Ent(S_1) - k_2 Ent(S_2)]$, et k,k_1 et k_2 sont les nombres de classes représentées dans les ensembles S,S_1 et S_2 respectivement[**?**]. Cette méthode de discrétisation d'une variable numérique est généralement appelée le principe de la longueur de description minimal (Minimal Description Length Principle).

Chapitre IV

Classifieur basé sur un ensemble de profils lorsque les observations ne sont pas identiquement distribuées

1 Introduction

L'apprentissage statistique présenté dans la première partie de cette analyse (voir chapitre III) est élaboré sous l'hypothèse d'indépendance et de distribution identique (i.i.d) des éléments aléatoires $(Y_i, X_i)_{i=1,\dots,n}$ qui ont généré les observations. Dans le présent chapitre, nous cherchons à adapter notre procédure d'apprentissage dans une situation où les données, en plus d'être déséquilibrées, sont réparties entre m clusters (groupes ou blocs) tirés aléatoirement à partir d'une population donnée. On suppose que chaque cluster admet une distribution $[Y, X]_h; h \in \{1, \cdots, m\}$ indépendantes des autres. Etant donné que l'indicateur de performance au tour duquel la procédure d'apprentissage a été élaborée est la valeur prédictive positive, nous proposons un estimateur Bayésien de la valeur prédictive positive de tout profil U conditionnellement à la distribution $[Y, X]_h$ des observations dans un cluster h donné. Cette approche nous permet de tenir en compte l'effet du cluster dans les résultats de l'analyse.

Les méthodes d'analyse classiques permettant de traiter des données groupées (essais multicentriques) introduisent en général la variable d'échantillonnage (groupe, cluster ou centre) comme variables explicatives en autorisant les interactions. Cependant elles ont des limites : (1) Lorsque le nombre de groupes est important, les introduire tous dans le modèle devient problématique. (2) Puisque l'un des groupes est utilisé comme groupe de référence, on ignore les écarts de chaque groupe à la moyenne. (3) Les groupes participant à l'essai constituent un échantillon d'une population plus large de groupes, on peut souhaiter faire des prédictions pour un groupe n'ayant pas participé à l'essai. (4) On peut aussi souhaiter avoir une mesure d'hétérogénéité entre les groupes.

Le modèle Bêta-binomiale figure parmi les méthodes alternatives les plus utilisées dans la littéra-

ture. Ce dernier permet à la fois d'estimer l'espérance de la probabilité de succès conditionnellement à un profil $U(X)$ dans la population et sa variabilité d'un groupe à un autre. De plus, il permet d'inférer sur la probabilité de succès conditionnellement à l'événement $[U(X) = 1]$ dans n'importe quel groupe, pas seulement ceux échantillonnés.

2 Modèle hiérarchique pour le calcul des valeurs prédictives positives

Nous étudions dans ce chapitre un modèle statistique correspondant au cas où les données sont générées par une suite $(Y_i, X_i)_{i=1:n}$ d'éléments aléatoires non identiquement distribués. Il en résulte alors une hétérogénéité des données dont il faudrait tenir compte dans le modèle statistique sur lequel l'analyse du classifieur sera basée.
Nous considérons ici la situation particulière où la suite $(Y_i, X_i)_{i=1:n}$ est structurée suivant une partition de m sous-ensembles $(Y_{ih}, X_{ih})_{\substack{h=1:m \\ i=1:n_h}}$ telles que les éléments de la suite $(Y_{ih}, X_{ih})_{i=1:n_h}$ soient indépendants et de même loi $[Y, X]_h$. Nous supposons que les éléments de la suite $[Y, X]^{\mathcal{L}} = \{[Y, X]_h, h = 1 : m\}$ sont générés de façon indépendante suivant une loi μ sur l'ensemble $Prob(Y, X)$ des lois de probabilités sur $Dom(Y) \times Dom(X)$ muni de la tribu associée à la topologie de la convergence faible. Si on se donne $U(X)$, un profil défini par X, on a alors

- $[Y|\theta_h^U, [Y, X]_h] = \text{Bernoulli}(\theta_h^U)$, où $\theta_h^U = \Pr(Y = 1|U(X) = 1, [Y, X]^{\mathcal{L}} = [Y, X]_h)$
- la suite $\left(\theta_h^U\right)_{h=1:m}$ est un échantillon iid.

On considère désormais que la suite $\theta^U = \left(\theta_h^U\right)_{h=1:m}$ est issue de la loi Bêta de paramètres (α_U, β_U). On désigne par $\left[Y, \theta^U, [Y, X]^{\mathcal{L}}\right]$ et $\left[\theta^U, [Y, X]^{\mathcal{L}}\right]$ les lois de probabilité respectives de $\left(Y, \theta^U, [Y, X]^{\mathcal{L}}\right)$ et $\left(\theta^U, [Y, X]^{\mathcal{L}}\right)$. Le principe de la factorisation permet d'écrire

$$\begin{aligned}
\left[Y, \theta^U, [Y, X]^{\mathcal{L}}\right] &= \left[Y|\theta^U, [Y, X]^{\mathcal{L}}\right] \left[\theta^U, [Y, X]^{\mathcal{L}}\right] \\
\left[Y, \theta^U, [Y, X]^{\mathcal{L}}\right] &= \left[Y|\theta^U, [Y, X]^{\mathcal{L}}\right] \left[\theta^U|[Y, X]^{\mathcal{L}}\right] \left[[Y, X]^{\mathcal{L}}\right] \\
\prod_{h=1}^{m} \left[Y, \theta_h^U, [Y, X]_h\right] &= \prod_{h=1}^{m} \left(\left[Y|\theta_h^U, [Y, X]_h\right] \left[\theta_h^U|[Y, X]_h\right] [[Y, X]_h]\right)
\end{aligned}$$

On peut remplacer la loi $\left[\theta_h^U|[Y, X]_h\right]$ par la loi $\left[\theta_h^U|\alpha_U, \beta_U\right]$ dans l'expression précédente puisqu'il s'agit de la même distribution. Pour réduire la complexité du problème, nous allons nous intéresser pour la suite à la distribution $\left[Y|\theta_h^U, [Y, X]_h\right]$ et à la distribution $\left[\theta_h^U|\alpha_U, \beta_U\right]$. Le modèle hiérarchique à étudier est alors le suivant :

$$\begin{cases}
\left[Y|\theta_h^U, [Y, X]_h\right] &= Bernoulli\left(\theta_h^U\right) \\
\left[\theta_h^U|\alpha_U, \beta_U\right] &= Beta(\alpha_U, \beta_U)
\end{cases}$$

Ce modèle permet d'estimer la probabilité $\Pr(Y=1|U(X),[Y,X]_h)$ qui n'est rien d'autre que la valeur prédictive positive (VPP) du profil $U(X)$ sous la contrainte $[Y,X]_h$.

Proposition 7. Si Y est une variable binaire telle que $Y|\theta_h^U,[Y,X]_h \sim Bernoulli\left(\theta_h^U\right)$ où $\theta_h^U|\alpha_U,\beta_U$ est une variable aléatoire de loi $Beta(\alpha_U,\beta_U)$ alors on a

$$\Pr(Y=1|U(X)) = \frac{\alpha_U}{\alpha_U+\beta_U} \tag{IV.1}$$

$$\left[\theta_h^U|Y=y,\alpha_U,\beta_U\right] = Beta(\alpha_U+y,\beta_U+1-y) \tag{IV.2}$$

Preuve.
On a

$$\begin{aligned}\Pr(Y=1|U(X)) &= \mathbb{E}\left(Y|\theta_h^U\right)\\ &= \mathbb{E}\left(\mathbb{E}\left(Y|\theta_h^U,[Y,X]_h\right)\right)\\ &= \mathbb{E}\left(\theta_h^U\right)\end{aligned}$$

Par ailleurs, on a

$$\left[\theta_h^U|Y=y,\alpha_U,\beta_U\right] = \frac{\left[Y=y|\theta_h^U,\alpha_U,\beta_U\right]\left[\theta_h^U|\alpha_U,\beta_U\right]}{\int_0^1\left[Y=y|\theta_h^U,\alpha_U,\beta_U\right]\left[\theta_h^U|\alpha_U,\beta_U\right]d\theta_h^U}$$

On en déduit alors que

$$\left[\theta_h^U|Y=y,\alpha_U,\beta_U\right] = \frac{\Gamma(\alpha_U+y)\Gamma(\beta_U-y+1)}{\Gamma(\alpha_U+\beta_U+1)}\theta_h^{\alpha_U+y-1}(1-\theta_h)^{\beta_U-y}$$

□

On a $\mathbb{E}(\theta_h^U|\alpha_U,\beta_U) = \dfrac{\alpha_U}{\alpha_U+\beta_U}$ et $\mathbb{V}ar(\theta_h^U|\alpha_U,\beta_U) = \dfrac{\alpha_U}{\alpha_U+\beta_U}\dfrac{\beta_U}{(\alpha_U+\beta_U)(\alpha_U+\beta_U+1)}$

L'application $(\alpha_U,\beta_U) \longrightarrow \begin{pmatrix}\pi_U = \frac{\alpha_U}{\alpha_U+\beta_U}\\ \gamma_U = \frac{1}{\alpha_U+\beta_U+1}\end{pmatrix}$ étant injective, on peut envisager de reparamétrer la famille de loi Bêta par la moyenne π_U et le paramètre γ_U appelé paramètre de dispersion. Pour π_U fixé, le paramètre γ_U détermine la forme de la densité. Nous retiendrons dans la suite du travail cette paramétrisation de la famille des lois Bêta.

3 Lois a posteriori des paramètres relatifs aux clusters : approche Bayésienne empirique

Pour alléger les notations dans cette section, on pose $\tau_U = 1/\gamma_U - 1$. Dans la suite, nous avons choisi d'écrire le modèle en fonction des paramètres $\{\pi_U, \tau_U\}$. Cependant les résultats seront présentés en fonction des paramètres $\{\pi_U, \gamma_U\}$. On pose le modèle suivant :

$$\begin{cases} \left[Y|\theta_h^U, [Y,X]_h\right] = \prod_{k=1}^{m} \left(\theta_h^U\right)^{\mathbb{1}_{[Y=1]}(y)\delta_{\{1,[Y,X]_h\}}(U(X),[Y,X]_h)} \left(1-\theta_h^U\right)^{(1-\mathbb{1}_{[Y=1]}(y))\delta_{\{1,[Y,X]_h\}}(U(X),[Y,X]_k)} \\ \left[\theta_h^U|\pi_U,\tau_U\right] = \frac{\Gamma(\tau_U)}{\Gamma(\pi_U\tau_U)\Gamma((1-\pi_U)\tau_U)} \left(\theta_h^U\right)^{\pi_U\tau_U-1} \left(1-\theta_h^U\right)^{(1-\pi_U)\tau_U-1} \mathbb{1}_{[0,1]}\left(\theta_h^U\right) \end{cases}$$

3.1 Détermination de la loi a posteriori du paramètre θ_h^U par une approche Bayésienne empirique

Dans une approche bayésienne complète, la détermination de la loi a posteriori de θ_h^U nécessite la spécification d'une loi a priori pour le couple (π_U,γ_U). En défaut de la spécification d'une telle loi a priori, on peut adopter une approche empirique pour la détermination a posteriori du vecteur $(\theta_h^U)_{h=1:m}$ et de ses éléments marginaux.

3.2 Loi a posteriori : approche bayésienne empirique

La méthode de Bayes empirique est très souvent utilisée lorsqu'il s'agit d'un problème d'estimation de paramètres multiples où les relations connues (*i.i.d.*) entres les composantes du vecteur de paramètres inconnus $\left(\theta_h^U\right)_{h=1:m}$ suggèrent de partager les informations entre les différentes réalisations similaires du couple (Y,X) pour obtenir une meilleure estimation de chaque paramètre θ_h^U. L'approche de Bayes empirique a été classée en deux catégories par Morris, C.N.[1983][7] dont : le cas non paramétrique (voir [8] pour plus de détails) et le cas paramétrique.
Dans le cas paramétrique, on suppose que la loi a priori du paramètre θ_h^U est dans une classe paramétrique $\left[\theta_h^U|\pi_U,\gamma_U\right]$, où les hyperparamètres π_U et γ_U sont inconnus. L'idée principale consiste à estimer les hyperparamètres d'abord et de les replacer dans la loi a priori avant d'estimer la loi a posteriori (pour plus de détails, consulter [2, 3]).

On considère, $(Y_i, X_i)_{i=1:n_h}$, une suite de n_h réalisations indépendantes de $[Y,X]_h$. On note $n_{hU} = \sum_{i=1}^{n_h} \mathbb{1}(U(X_i)=1)$ le nombre d'observations i telles que $U(X_i)=1$. On suppose que n_{hU} est un entier connu et supérieur strictement à un. On note $S_{hU} = \sum_{i=1}^{n_{hU}} \mathbb{1}(Y_i=1, U(X_i)=1)$ une variable aléatoire qui détermine le nombre d'observations i telles que $U(X_i)=1$ et $Y_i=1$. On suppose que

$(S_{hU}|\theta_h^U)_{h=1:m}$ est une suite de variables aléatoires indépendantes mais pas nécessairement identiquement distribuées. Pour tout cluster h donné, on suppose que

$$\left[S_{hU}|\theta_h^U\right] = \text{Binomiale}(n_{hU}, \theta_h^U)$$

L'objectif est de trouver une estimation ponctuelle pour θ_h^U à partir des observations S_{hU}. On commence par déterminer la loi a posteriori de $\theta_h^U|\pi_U, \gamma_U$ qui dépend des données par S_{hU}. La loi a posteriori est donnée par :

$$\left[\theta_h^U|S_{hU}, \pi_U, \gamma_U\right] = \frac{\left[S_{hU}|\theta_h^U\right]\left[\theta_h^U|\pi_U, \gamma_U\right]}{[S_{hU}|\pi_U, \gamma_U]}$$

En supposant que les hyperparamètres π_U et γ_U sont inconnus, nous les estimerons à partir de la distribution marginale de toutes les données, $[S_{hU}|\pi_U, \gamma_U]$. On obtient la distribution a posteriori estimée :

$$\left[\theta_h^U|S_{hU}, \widehat{\pi}_U, \widehat{\gamma}_U\right]$$

où $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ sont des fonctions de S_{hU} (i.e., $\widehat{\pi}_U(S_{hU})$ et $\widehat{\gamma}_U(S_{hU})$). Ces estimateurs sont habituellement obtenus par la méthode du maximum de vraisemblance (MLE) ou la méthode des moments (MOM) à partir de la distribution marginale $[S_{hU}|\pi_U, \gamma_U]$. Une fois les estimateurs $\{\widehat{\pi}_U, \widehat{\gamma}_U\}$ obtenus, nous pouvons estimer alors $\widehat{\theta}_h^U$ comme étant la moyenne de la distribution a posteriori estimée. Notons que, $\widehat{\theta}_h^U$ dépend de toutes les données par le biais de $\widehat{\pi}_U$ et $\widehat{\gamma}_U$. Dans cette analyse, nous proposons d'estimer les hyperparamètres $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ par la méthode des moments.

4 Estimation des hyperparamètres π_U et γ_U

4.1 Estimation par la méthode des moments

Le principe de la méthode des moments consiste à estimer les paramètres recherchés en égalisant certains moments théoriques (qui dépendent de ces paramètres) avec leurs contreparties empiriques. L'égalisation se justifie par la loi des grands nombres qui implique que l'on peut "approcher" une espérance mathématique par une moyenne empirique. On est donc amené à résoudre un système d'équations.

4.1.1 Moments des variables S_{hU} et θ_h^U

Etant donné que la loi a priori de $\theta_h|\pi_U, \gamma_U$ est connue (la loi Bêta), il est possible de déterminer les expressions explicites de ses moments d'ordre un et deux. Nous commencerons par donner l'expression

des moments d'ordre n. Ensuite nous en déduirons les moments d'ordre un, deux, trois et quatre.

$$\mathbf{E}\left(\left(\theta_h^U\right)^n|\pi_U,\gamma_U\right) = \left[\frac{\Gamma\left[\frac{1-\gamma_U}{\gamma_U}\right]}{\Gamma\left[\frac{\pi_U(1-\gamma_U)}{\gamma_U}\right]}\right]\left[\frac{\Gamma\left[\frac{\pi_U(1-\gamma_U)}{\gamma_U}+n\right]}{\Gamma\left[\frac{1-\gamma_U}{\gamma_U}+n\right]}\right]$$

On obtient alors :

$$\mathbb{E}\left(\theta_h^U|\pi_U,\gamma_U\right) = \pi_U$$

$$\mathbb{E}\left(\left(\theta_h^U\right)^2|\pi_U,\gamma_U\right) = \pi_U^2+\gamma_U\pi_U(1-\pi_U)$$

Nous déduisons des moments de θ_h^U les moments suivants :

$$\begin{aligned}\mathbb{E}\left(S_{hU}\right) &= \mathbb{E}\left[\mathbb{E}\left(S_{hU}|\theta_h^U,\pi_U,\gamma_U\right)\right]\\ &= \mathbb{E}\left(\mathbb{E}\left(S_{hU}|\theta_h^U\right)|\pi_U,\gamma_U\right)\\ &= n_{hU}\pi_U\end{aligned}$$

$$\begin{aligned}\mathbb{V}ar(S_{hU}) &= \mathbb{E}[\mathbb{V}ar(S_{hU}|\pi_U,\gamma_U)]+\mathbb{V}ar[\mathbb{E}(S_{hU}|\pi_U,\gamma_U)]\\ &= n_{hU}\pi_U(1-\pi_U)+\gamma_U\pi_U(1-\pi_U)n_{hU}(n_{hU}-1)\end{aligned}$$

Nous supposons que les observations de n_{hU} sont strictement supérieures à 1 (i.e. $n_{hU}>1$). On obtient alors

$$\begin{cases}\mathbb{E}(S_{hU}) = n_{hU}\pi_U\\ \\ \mathbb{E}[(S_{hU})^2] = n_{hU}\pi_U(1-\pi_U+n_{hU}\pi_U)+\gamma_U\pi_U(1-\pi_U)n_{hU}(n_{hU}-1)\end{cases}$$

En faisant la différence membre à membre des deux égalités ci-dessus, on obtient les égalités suivantes

$$\begin{cases}\mathbb{E}\left(\frac{S_{hU}}{n_{hU}}\right) = \pi_U\\ \\ \mathbb{E}\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right) = \pi_U^2+\gamma_U\pi_U(1-\pi_U)\end{cases}$$

4.1.2 Estimation de π_U et γ_U

Dans ses travaux, Griffiths a montré que lorsque les n_{hU} sont inégaux, l'estimation des paramètres π_U et γ_U par des moments empiriques pondérés produit de meilleurs estimateurs que lorsque on utilise des moments empiriques non pondérés [4].

Si on suppose que les variables $\left(\frac{S_{hU}}{n_{hU}}\right)$ sont indépendantes et de variances non nulles de même que les variables $\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)$, il est alors désirable d'utiliser leurs moments empiriques pondérés dans le but d'obtenir de meilleurs estimateurs de π_U et de γ_U.
Soit

$$W_U = \sum_{h=1}^{m} \frac{w_{hU}}{w_U}\left(\frac{S_{hU}}{n_{hU}}\right), \qquad w_U = \sum_{h=1}^{m} w_{hU} \tag{IV.3}$$

et

$$S_U = \sum_{h=1}^{m} \frac{v_{hU}}{v_U}\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right), \qquad v_U = \sum_{h=1}^{m} v_{hU} \tag{IV.4}$$

les moments empiriques respectifs de $\left(\frac{S_{hU}}{n_{hU}}\right)$ et $\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)$, où w_{hU} et v_{hU} représente les coefficients de pondération respectifs. Dans la suite, on verra comment ils sont choisis.
En définissant les statistiques des équations (IV.3) et (IV.4) égales à leurs valeurs théoriques et en résolvant les équations qui en résultent par rapport à π_U et γ_U, nous obtenons les estimateurs suivants :

$$\widehat{\pi}_U = \sum_{h=1}^{m} \frac{w_{hU}}{w_U}\frac{S_{hU}}{n_{hU}} \tag{IV.5}$$

$$\widehat{\gamma}_U = \frac{\sum_{h=1}^{m} \frac{v_{hU}}{v_U}\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right) - \widehat{\pi}_U^2}{\widehat{\pi}_U(1-\widehat{\pi}_U)} \tag{IV.6}$$

Les estimateurs des moments pondérés dépendent du choix des poids $\{w_{hU},v_{hU}\}$. Il est très connu de la littérature que si $\{w_{hU},v_{hU}\}$ sont choisis proportionnellement aux variances respectives de $\left(\frac{S_{hU}}{n_{hU}}\right)$ et $\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)$ alors W_U et S_U ont les plus petites variances parmi tous les estimateurs linéaires sans biais de π_U et γ_U respectivement. Si nous pondérons chaque variable $\frac{S_{hU}}{n_{hU}}$ et chaque variable $\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}$ par l'inverse de sa variance (supposée être connue) alors W_U et S_U sont les estimateurs linéaires sans biais et de variance minimum de π_U et de $\pi_U^2+\gamma_U\pi_U(1-\pi_U)$ respectivement. Les poids correspondants sont :

$$\begin{aligned} \mathbb{V}ar\left(\frac{S_{hU}}{n_{hU}}\right) &= \frac{\pi_U(1-\pi_U)}{n_{hU}} + \gamma_U\pi_U(1-\pi_U)\left(1-\frac{1}{n_{hU}}\right) \\ \left[\mathbb{V}ar\left(\frac{S_{hU}}{n_{hU}}\right)\right]^{-1} &= \frac{n_{hU}}{\pi_U(1-\pi_U)+\gamma_U\pi_U(1-\pi_U)(n_{hU}-1)} \end{aligned}$$

$$\begin{aligned}\mathbb{V}ar\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right) &= \mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] - \left[\mathbb{E}\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)\right]^2 \\ &= \mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] - \left[\pi_U^2+\gamma_U\pi_U(1-\pi_U)\right]^2\end{aligned}$$

avec

$$\mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] = \frac{1}{[n_{hU}(n_{hU}-1)]^2}\left[\mathbb{E}\left(S_{hU}^4\right)-2\mathbb{E}\left(S_{hU}^3\right)+\mathbb{E}\left(S_{hU}^2\right)\right]$$

$$\begin{aligned}\mathbb{E}\left(S_{hU}^2|\theta_h^U\right) &= n_{hU}\theta_h^U + n_{hU}\left(n_{hU}-1\right)\left(\theta_h^U\right)^2 \\ \mathbb{E}\left(S_{hU}^3|\theta_h^U\right) &= n_{hU}\theta_h^U + 2n_{hU}(n_{hU}-1)\left(\theta_h^U\right)^2 + n_{hU}(n_{hU}-1)(n_{hU}-2)\left(\theta_h^U\right)^3 \\ \mathbb{E}\left(S_{hU}^4|\theta_h^U\right) &= n_{hU}\theta_h^U + 4n_{hU}(n_{hU}-1)\left(\theta_h^U\right)^2 + 4n_{hU}(n_{hU}-1)(n_{hU}-2)\left(\theta_h^U\right)^3 \\ &+ n_{hU}(n_{hU}-1)(n_{hU}-2)(n_{hU}-3)\left(\theta_h^U\right)^4\end{aligned}$$

donc

$$\mathbb{E}\left[\left(S_{hU}^2-S_{hU}\right)^2|\theta_h^U\right] = n_{hU}(n_{hU}-1)\left[\left(\theta_h^U\right)^2+2(n_{hU}-2)\left(\theta_h^U\right)^3+(n_{hU}-2)(n_{hU}-3)\left(\theta_h^U\right)^4\right]$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] &= \frac{1}{n_{hU}(n_{hU}-1)}\left[\mathbb{E}\left(\left(\theta_h^U\right)^2|\pi_U,\gamma_U\right)+2(n_{hU}-2)\mathbb{E}\left(\left(\theta_h^U\right)^3|\pi_U,\gamma_U\right)\right. \\ &\left.+(n_{hU}-2)(n_{hU}-3)\mathbb{E}\left(\left(\theta_h^U\right)^4|\pi_U,\gamma_U\right)\right]\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] = \frac{\mathbb{E}\left(\left(\theta_h^U\right)^2|\pi_U,\gamma_U\right)}{n_{hU}(n_{hU}-1)}&\left[1+2(n_{hU}-2)\frac{\pi_U(1-\gamma_U)+2\gamma_U}{1+\gamma_U}\right. \\ &\left.+(n_{hU}-2)(n_{hU}-3)\frac{\pi_U(1-\gamma_U)+2\gamma_U}{1+\gamma_U}\frac{\pi_U(1-\gamma_U)+3\gamma_U}{1+2\gamma_U}\right]\end{aligned}$$

Puisque $\pi_U(1-\pi_U)$ est constant (indépendant de h), alors nous considérons pour w_{hU} la valeur suivante :

$$w_{hU} = \frac{n_{hU}}{1+\gamma_U(n_{hU}-1)} \tag{IV.7}$$

et pour v_{hU} la valeur suivante :

$$v_{hU} = \frac{1}{\mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] - \left[\pi_U^2 + \gamma_U \pi_U (1-\pi_U)\right]^2} \tag{IV.8}$$

Cependant l' estimation du paramètre w_{hU} et du paramètre v_{hU} est compliquée par le fait que tous les deux paramètres dépendent des paramètres π_U et γ_U inconnus. Une manière de les estimer consisterait à remplacer π_U et γ_U par leurs estimations respectives $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ dans les équations (IV.7) et (IV.8). Cependant, lorsque m le nombre de cluster n'est pas suffisamment grand, la loi des grands nombres ne s'applique pas et par conséquent, les moments empiriques W_U et S_U n'approchent pas suffisamment bien les moments théoriques. En plus le signe de $\widehat{\gamma}_U$ dépend de la suite (S_{hU}, n_{hU}). Les estimateurs ainsi obtenus peuvent avoir tendance à sortir du support des paramètres (voir annexe C).
Pour parer à cette difficulté, une méthode de pondération empirique a été proposée en premier par Kleinman en [1973][6] puis améliorée par Tchuang-Stein en [1993][1] pour l'estimation de w_{hU}. A partir de cet algorithme, une estimation de $\widehat{\pi}_U$ a été déduite. Nous nous sommes inspirés de cette méthode pour établir l' algorithme d'estimation de π_U et de γ_U décrit ci-dessous.

4.1.3 Algorithme de la méthode de Pondération Empirique

On propose de choisir une valeur initiale $\gamma_0 = 0$ ou $\gamma_0 = 1$ du paramètre γ_U pour obtenir les valeurs initiales w_0 et v_0 de w_{hU} et v_{hU} respectivement. Ensuite on utilise les équations (IV.5) et (IV.6) pour obtenir les estimations de π_U et de γ_U. A partir de cette estimation de γ_U, notée $\widehat{\gamma}_U$, on calcule le couple $\{w_{hU}, v_{hU}\}$ à partir des équations (IV.7) et (IV.8). Et enfin on utilisera ces poids empiriques pour former de nouvelles estimations de W_U et S_U. On répète cette itération jusqu'à ce que les différences entre deux itérations consécutives d'estimations W_U, S_U et $\widehat{\gamma}_U$ soient à la fois plus petites qu'une certaine valeur prédéterminée, par défaut 10^{-6}. Pour des soucis de programmation, nous proposons de réinitialiser à 10^{-6} les estimations négatives de γ_U au lieu de 0 comme proposé par Kleinman.
Pour des raison de programmation, nous avons ajouté la masse de Dirac au point 0 de n_{hU} dans le calcul des statistiques W_U et S_U. Dans la simulation, il n'est pas évident d'avoir toutes les statistiques $(n_{hU})_{k=1}^k$ supérieures strictement à 1. En utilisant cette astuce, nous nous assurons que les dénominateurs de $S_{hU}/[n_{hU} + \delta_0(n_{hU})]$ et $S_{hU}(S_{hU}-1)/[n_{hU}(n_{hU}-1)\delta_0(n_{hU}(n_{hU}-1)) + \delta_0(n_{hU}(n_{hU}-1))]$ soient toujours égaux à 1 si n_{hU} est égale à un ou zéro. Dans le cas où $n_{hU} = 0$, on sait que S_{hU} est presque sûrement nulle. Ceci nous permet de pouvoir faire des estimations de π_U et de γ_U même s'il existe des réalisations $(Y_i, X_i)_{i=1:n_h}$ de $[Y, X]_h$ pour lesquelles le profil $U(X)$ n'a pas été observé $(U(X) = 0)$.

Algorithme : Méthode de pondération empirique

on suppose avoir observé les statistiques suivantes : $(S_{hU})_{h=1:H}$ et $(n_{hU})_{h=1:H}$
on commence par donner une valeur initiale $\gamma_U = 0$ ou $\gamma_U = 1$ et le nombre d'itérations maximum de la procédure : $maxiter = 100$ (par défaut)
on initialise
- $W_U = \frac{1}{K}\sum_{h=1}^{H} \frac{S_{hU}}{n_{hU}+\delta_0(n_{hU})}$
- $S_U = \frac{1}{K}\sum_{h=1}^{H} \frac{S_{hU}(S_{hU}-1)}{n_{hU}(n_{hU}-1)\delta_0(n_{hU}(n_{hU}-1))+\delta_0(n_{hU}(n_{hU}-1))}$

Déclarer une variable booléenne *cond.arret* (condition d'arrêt) initialisée à *vrai* et une variable t initialisée à 0.
Tant que *cond.arret* est toujours vrai **faire** :
initialiser : $t = t + 1$; $\gamma_U^t = \gamma_U$; $\pi_U^t = W_U$ et $S_U^t = S_U$
calculer en fonction de π_U^t et γ_U^t le couple $\{w_{hU}, v_{hU}\}$
En suite calculer les statistiques :
- $W_U = \sum_{h=1}^{m} \frac{w_{hU}}{w_U}\left(\frac{S_{hU}}{n_{hU}+\delta_0(n_{hU})}\right)$
- $S_U = \sum_{h=1}^{m} \frac{v_{hU}}{v_U}\left(\frac{S_{hU}(S_{hU}-1)}{n_{hU}(n_{hU}-1)\delta_0(n_{hU}(n_{hU}-1))+\delta_0(n_{hU}(n_{hU}-1))}\right)$

Puis on associe $\pi_U = W_U$ et $\gamma_U = \frac{S_U - \pi_U^2}{\pi_U(1-\pi_U)}$
- si $\gamma_U < 0 \quad \Rightarrow \quad \gamma_U = 10^{-6}$

cond.arret $= \{|\gamma_U - \gamma_U^t| > 10^{-6}, |\pi_U - \pi_U^t| > 10^{-6}, |S_U - S_U^t| > 10^{-6}, t < maxiter\}$
fin tant que

Tableau IV.1 – Algorithme de la méthode de pondération empirique

4.2 Estimation des hyperparamètres par la méthode du maximum de vraisemblance

Pour simplifier les notations, on a choisi d'omettre l'indice U sur les paramètres π et γ. De plus on considère le changement de paramètre $\tau = \frac{1}{\gamma} - 1$.

4.2.1 Vraisemblance des paramètres

On a

$$\begin{aligned}
[(S_h)_{h=1:m}|\pi,\tau] &= \prod_{h=1}^{m}[S_h|\pi,\tau] \\
&= \prod_{h=1}^{m}\binom{s_h}{n_h}\frac{\Gamma(\tau)}{\Gamma(\tau+n_h)}\frac{\Gamma(\pi\tau+s_h)}{\Gamma(\pi\tau)}\frac{\Gamma((1-\pi)\tau+n_h-s_h)}{\Gamma((1-\pi)\tau)} \\
&= \prod_{h=1}^{m}\left\{\binom{s_h}{n_h}\left[\prod_{j=0}^{n_h-1}\frac{1}{\tau+j}\right]\left[\prod_{k=0}^{s_h-1}(\pi\tau+k)\right]\left[\prod_{l=0}^{n_h-s_h-1}((1-\pi)\tau+l)\right]\right\}
\end{aligned}$$

La vraisemblance des paramètres π et τ est donnée par :

$$L(\pi,\tau) = \sum_{h=1}^{m}\left\{log\left(\binom{s_h}{n_h}\right) - \sum_{j=0}^{n_h-1} log(\tau+j) + \sum_{k=0}^{s_h-1} log(\pi\tau+k) + \sum_{l=0}^{n_h-s_h-1} log((1-\pi)\tau+l)\right\}$$

L'optimisation de la vraisemblance $L(\pi,\tau)$ est très compliquée à implémenter. Il n'est pas possible de trouver une solution analytique. Cependant plusieurs algorithmes itératifs ont été proposés dans la littérature pour venir à bout cette difficulté. Dans cette analyse, nous proposons d'utiliser un algorithme MM.

4.2.2 Présentation du principe et des éléments d'un algorithme MM

Nous allons utiliser l'algorithme MM (Minimisation-Maximisation) pour estimer les paramètres π et τ. Les algorithmes MM ont pour objectif de substituer à un problème d'optimisation numérique d'une fonction f compliquée à implémenter par celui de l'optimisation d'une fonction auxiliaire g dont l'optimum correspond à un optimum local de f. La fonction auxiliaire g est telle que

$$\begin{aligned} f(x) &\geq g(x|x') \qquad x \in \Delta \times \mathbf{D} \\ f(x) &= g(x|x) \end{aligned}$$

On observe que si pour x_0 fixé et $x_1 = \underset{x}{\text{argmax}}\, g(x|x_0)$, alors on a $f(x_1) \geq g(x_1|x_0) \geq g(x_0|x_0) = f(x_0)$ Il en résulte que les algorithmes MM sont des algorithmes monotones. Les algorithmes MM procèdent en deux étapes. La première étape consiste à trouver la fonction g telle que

$$L(\pi,\tau) \geq g(\pi,\tau|\pi',\tau') \tag{IV.9}$$
$$L(\pi,\tau) = g(\pi,\tau|\pi,\tau) \quad \forall\ (\pi,\tau) \tag{IV.10}$$

La deuxième étape consiste à trouver un couple $(\widehat{\pi},\widehat{\tau})$ qui maximise la fonction $g(\pi,\tau|\pi',\tau')$.

$$(\widehat{\pi},\widehat{\tau}) = \underset{(\pi,\tau)}{\text{argmax}}\, g(\pi,\tau|\pi',\tau')$$

4.2.3 Proposition de la fonction auxiliaire et ses propriétés

Proposition 8. Soient $L(\pi,\tau)$ la log-vraisemblance du couple des paramètres (π,τ) et (π',τ') une valeur connue des paramètres (π,τ). La fonction auxiliaire $g(\pi,\tau|\pi',\tau')$ définie par

$$g(\pi,\tau|\pi',\tau') = A(\pi',\tau')[log(\pi)+log(\tau)] + B(\pi',\tau')[log(1-\pi)+log(\tau)] - (\tau-\tau')C(\tau') + D(\pi',\tau')$$

vérifie les conditions (IV.9) et (IV.10).

Preuve. Puisque la fonction $-log(x)$ est convexe, on a :

$$-log(\tau+j) \geq -log(\tau'+j) - \frac{(\tau-\tau')}{\tau'+j}$$

En utilisant la concavité de la fonction $log(x)$, on obtient

$$log(\pi\tau+k) \geq \frac{\pi'\tau'}{\pi'\tau'+k} log\left(\frac{\pi'\tau'+k}{\pi'\tau'}\pi\tau\right) + \frac{k}{\pi'\tau'+k} log\left(\frac{\pi'\tau'+k}{k}k\right)$$

$$log((1-\pi)\tau+l) \geq \frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l} log\left(\frac{(1-\pi')\tau'+l}{(1-\pi')\tau'}(1-\pi)\tau\right) + \frac{l}{(1-\pi')\tau'+l} log\left(\frac{(1-\pi')\tau'+l}{l}l\right)$$

On peut donc poser

$$\begin{aligned}
g(\pi,\tau|\pi',\tau') &= \sum_{h=1}^{m} log\left(\binom{s_h}{n_h}\right) - \sum_{h=1}^{m}\sum_{j=0}^{n_h-1} log(\tau'+j) - (\tau-\tau')\sum_{h=1}^{m}\sum_{j=0}^{n_h-1}\frac{1}{\tau'+j} \\
&+ \sum_{h=1}^{m} \mathbb{1}(s_h \geq 1)\left(\sum_{k=0}^{s_h-1}\frac{\pi'\tau'}{\pi'\tau'+k}\left[log\left(\frac{\pi'\tau'+k}{\pi'\tau'}\pi\tau\right)\right] + \sum_{h=1}^{m}\sum_{k=0}^{s_h-1}\frac{k}{\pi'\tau'+k} log\left(\pi'\tau'+k\right)\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(n_h > s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l}\left[log\left(\frac{(1-\pi')\tau'+l}{(1-\pi')\tau'}(1-\pi)\tau\right)\right]\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(n_h > s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{l}{(1-\pi')\tau'+l} log\left((1-\pi')\tau'+l\right)\right)
\end{aligned}$$

On peut réécrit la fonction $g(\pi,\tau|\pi',\tau')$ de telle sorte que les paramètres π et τ soient séparés. On obtient

$$\begin{aligned}
g(\pi,\tau|\pi',\tau') &= \sum_{h=1}^{m} log\left(\binom{s_h}{n_h}\right) - \sum_{h=1}^{m}\sum_{j=0}^{n_h-1} log(\tau'+j) - (\tau-\tau')\sum_{h=1}^{m}\sum_{j=0}^{n_h-1}\frac{1}{\tau'+j} \\
&+ \sum_{h=1}^{m} \mathbb{1}(s_h \geq 1\left(\sum_{k=0}^{s_h-1}\frac{\pi'\tau'}{\pi'\tau'+k}\left[log\left(\frac{\pi'\tau'+k}{\pi'\tau'}\right) + log(\pi) + log(\tau)\right]\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(s_h \geq 1)\left(\sum_{k=0}^{s_h-1}\frac{k}{\pi'\tau'+k} log\left(\pi'\tau'+k\right)\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(n_h > s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l}\left[log\left(\frac{(1-\pi')\tau'+l}{(1-\pi')\tau'}\right) + log(1-\pi) + log(\tau)\right]\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(n_h > s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{l}{(1-\pi')\tau'+l} log\left((1-\pi')\tau'+l\right)\right)
\end{aligned}$$

Si on pose

$$\begin{aligned}
A(\pi',\tau') &= \sum_{h=1}^{m} \mathbb{1}\left(s_h \geq 1\right)\left(\sum_{k=0}^{s_h-1}\frac{\pi'\tau'}{\pi'\tau'+k}\right) \\
B(\pi',\tau') &= \sum_{h=1}^{m} \mathbb{1}\left(n_h \geq s_h\right)\left(\sum_{l=0}^{n_h-s_h-1}\frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l}\right) \\
C(\tau') &= \sum_{h=1}^{m}\sum_{j=0}^{n_h-1}\frac{1}{\tau'+j}
\end{aligned}$$

$$
\begin{aligned}
D(\pi',\tau') &= \sum_{h=1}^{m} log\left(\binom{s_h}{n_h}\right) - \sum_{h=1}^{m}\sum_{j=0}^{n_h-1} log(\tau'+j) + \sum_{h=1}^{m}\mathbb{1}(s_h\geq 1)\left(\sum_{k=0}^{s_h-1}\frac{\pi'\tau'}{\pi'\tau'+k} log\left(\frac{\pi'\tau'+k}{\pi'\tau'}\right)\right) \\
&+ \sum_{h=1}^{m}\mathbb{1}(s_h\geq 1)\left(\sum_{k=0}^{s_h-1}\frac{k}{\pi'\tau'+k} log\left(\pi'\tau'+k\right)\right) + \sum_{h=1}^{m}\mathbb{1}(n_h> s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l} log\left(\frac{(1-\pi')\tau'+l}{(1-\pi')\tau'}\right)\right) \\
&+ \sum_{h=1}^{m}\mathbb{1}(n_h> s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{l}{(1-\pi')\tau'+l} log\left((1-\pi')\tau'+l\right)\right)
\end{aligned}
$$

Il en résulte que

$$g(\pi,\tau|\pi',\tau') = Cste - (\tau-\tau')C(\tau') + A(\pi',\tau')[log(\pi)+log(\tau)] + B(\pi',\tau')[log(1-\pi)+log(\tau)]$$

On a

$$L(\pi,\tau)\geq g(\pi,\tau|\pi',\tau')$$

En plus lorsque on pose $\pi=\pi'$ et $\tau=\tau'$, on obtient

$$L(\pi,\tau) = g(\pi,\tau|\pi,\tau)$$

□

Les couples candidats sont l'ensemble des couples annulant les dérivées partielles de la fonction $g(\pi,\tau|\pi',\tau')$.

$$
\begin{aligned}
\frac{\delta}{\delta\pi}g(\pi,\tau|\pi',\tau') &= \frac{1}{\pi}A(\pi',\tau') - \frac{1}{1-\pi}B(\pi',\tau') \\
\frac{\delta}{\delta\tau}g(\pi,\tau|\pi',\tau') &= -C(\tau') + \frac{1}{\tau}\left[A(\pi',\tau')+B(\pi',\tau')\right]
\end{aligned}
$$

Il en résulte que

$$\widehat{\pi} = \frac{A(\pi',\tau')}{A(\pi',\tau')+B(\pi',\tau')} \tag{IV.11}$$

$$\widehat{\tau} = \frac{A(\pi',\tau')+B(\pi',\tau')}{C(\tau')} \tag{IV.12}$$

En plus on a

$$\frac{\delta^2}{\delta^2\pi}g(\pi,\tau|\pi',\tau') = -\frac{1}{\pi^2}A(\pi',\tau') - \frac{1}{(1-\pi)^2}B(\pi',\tau') \tag{IV.13}$$

$$\frac{\delta^2}{\delta^2\tau}g(\pi,\tau|\pi',\tau') = -\frac{1}{\tau^2}\left[A(\pi',\tau')+B(\pi',\tau')\right] \tag{IV.14}$$

Par conséquent on a

$$\frac{\delta^2}{\delta^2\tau}g(\widehat{\pi},\widehat{\tau}|\pi',\tau') = -C(\tau')\leq 0$$

et

$$\frac{\delta^2}{\delta^2\pi}g(\widehat{\pi},\widehat{\tau}|\pi',\tau') = -\left(\frac{(1-\widehat{\pi}^2)A(\pi',\tau') + \widehat{\pi}^2 B(\pi',\tau')}{\widehat{\pi}^2(1-\widehat{\pi})^2}\right) \leq 0$$

Le couple $(\widehat{\pi},\widehat{\tau})$ donnée par les équations (IV.11) et (IV.12) est donc un maximum local de la fonction $g(\pi,\tau|\pi',\tau')$. En se servant des équations (IV.9) et (IV.10), on obtient $L(\widehat{\pi},\widehat{\tau}) \geq L(\pi',\tau')$. Le couple $(\widehat{\pi},\widehat{\tau})$ maximisant la vraisemblance est atteint lorsque la condition d'arrêt (**??**) est obtenue.

4.2.4 Algorithme

La phase de maximisation consiste à maximiser la fonction $g(\pi,\tau|\pi',\tau')$. Cette dernière partie correspond à l'algorithme numérique itératif de newton pour optimiser la fonction g. Le principe de l'algorithme est le suivant :

Algorithme : MM (Minimisation-Maximisation)

- Entrées : $\mathcal{D} = \{(s_h, n_h); h = 1 : m\}$ un ensemble d'observations ; (π^0, τ^0) valeurs initiales des paramètres à estimer et *maxiter* le nombre d'itération maximum.
- Sortie : le couple $(\widehat{\pi},\widehat{\tau})$

Variables déclarées :

- *cond.arret* : une variable booléenne initialisée à vrai
- t : étape itérative initialisée à 0
- $(\pi^t, \tau^t) \leftarrow (\pi^0, \tau^0)$

Tant que *cond.arret* est vrai **faire** :

On itère $t \leftarrow t+1$ et $\left(\pi^{(t-1)}, \tau^{(t-1)}\right) \leftarrow \left(\pi^{(t)}, \tau^{(t)}\right)$

- $\pi^{(t)} \leftarrow \dfrac{A\left(\pi^{(t-1)}, \tau^{(t-1)}\right)}{A\left(\pi^{(t-1)}, \tau^{(t-1)}\right) + B\left(\pi^{(t-1)}, \tau^{(t-1)}\right)}$
- $\tau^{(t)} \leftarrow \dfrac{A\left(\pi^{(t-1)}, \tau^{(t-1)}\right) + B\left(\pi^{(t-1)}, \tau^{(t-1)}\right)}{C\left(\tau^{(t-1)}\right)}$

$$cond.arret \leftarrow \left\{\left(\left(\pi^{(t)} - \pi^{(t-1)}\right)^2 + \left(\tau^{(t)} - \tau^{(t-1)}\right)^2 + 1 \neq 1\right) \;\&\; (t < maxietr)\right\}$$

fin tant que

résultats : $\left(\pi^{(t)}, \tau^{(t)}\right)$

Tableau IV.2 – Algorithme MM (Minimisation-Maximisation)

5 Eléments pour la formulation d'un classifieur individuel pour les groupes

Soit $U(X)$ un profil donné. Nous observons S_{hU} co-occurrences dans n_{hU} observations pertinentes pour un cluster h donné. Nous modélisons le nombre de co-occurrences par une loi $Binomiale(n_{hU}, \theta_h^U)$ et θ_h^U par une loi $Beta(\pi_U(1-\gamma_U)/\gamma_U, (1-\pi_U)(1-\gamma_U)/\gamma_U)$ de manière hiérarchique pour partager l'information entre les clusters similaires. De manière plus formelle, nous proposons le modèle suivant :

$$\begin{aligned} S_{hU} &\sim \text{Binom}(n_{hU}, \theta_h^U) \\ \theta_h^U &\sim \text{Beta}\left(\frac{\pi_U(1-\gamma_U)}{\gamma_U}, \frac{(1-\pi_U)(1-\gamma_U)}{\gamma_U}\right) \end{aligned}$$

Sous ces hypothèses, on a

$$\begin{aligned} \mathbb{E}\left(\theta_h^U | [Y,X]_h, \pi_U, \gamma_U\right) &= \int_0^1 \Pr\left(Y=1 | U(X)=1, [Y,X]_h\right) \left[\theta_h^U | [Y,X]_h, \pi_U, \gamma_U\right] d\theta_h^U \\ \mathbb{E}\left(\theta_h^U | [Y,X]_h, \pi_U, \gamma_U\right) &= \frac{S_{hU} + \pi_U\left(\frac{1}{\gamma_U} - 1\right)}{n_{hU} + \left(\frac{1}{\gamma_U} - 1\right)} \end{aligned}$$

La valeur prédictive positive a posteriori est donnée par

$$VPP(U,Y,h) = \mathbb{E}\left(\mathbb{E}\left(\theta_h^U | Y, \widehat{\pi}_U, \widehat{\gamma}_U\right)\right)$$

Puisqu'on n'a pas supposé une loi a priori sur les hyperparamètres π_U et γ_U, alors leurs estimations sont faites à partir des données $([Y,X]_h)_{h=1:m}$. Par conséquent la valeur prédictive positive a posteriori obtenue est un estimateur empirique de Bayes de la valeur prédictive positive du classifieur $\phi(X,U)$ généré par le profil $U(X)$.

$$VPP(U,Y,h) = \frac{S_{hU} + \widehat{\pi}_U\left(\frac{1}{\widehat{\gamma}_U} - 1\right)}{n_{hU} + \left(\frac{1}{\widehat{\gamma}_U} - 1\right)}$$

Pour chaque profil $U(X)$ fixé, on a une suite $(VPP(U,Y,h))_{h=1:m}$ dont chaque $VPP(U,Y,h)$ dépend des observations de la loi $[Y,X]_h$. $VPP(U,Y,h)$ est une estimation de la valeur prédictive positive du profil $U(X)$ dans le cluster h en tenant compte de ses fréquences dans les autres clusters. On peut écrire $VPP(U,Y,h)$ sous la forme d'une combinaison linéaire convexe de $\dfrac{S_{hU}}{n_{hU}}$ et de π_U. :

$$VPP(U,Y,h) = \frac{S_{hU}}{n_{hU}}\left(1 - \frac{(1-\widehat{\gamma}_U)/\widehat{\gamma}_U}{n_{hU} + (1-\widehat{\gamma}_U)/\widehat{\gamma}_U}\right) + \widehat{\pi}_U\left(\frac{(1-\widehat{\gamma}_U)/\widehat{\gamma}_U}{n_{hU} + (1-\widehat{\gamma}_U)/\widehat{\gamma}_U}\right)$$

La statistique $\frac{S_{hU}}{n_{hU}}$ représente la valeur prédictive positive du profil $U(X)$ dans le cluster h indépendamment des autres clusters. Tandis que π_U représente la valeur prédictive positive du profil $U(X)$ dans la population.

$$VPP(U,Y) = \widehat{\pi}_U$$

Pour prédire la classe d'une observation dans un cluster h spécifié, on pourra utiliser la statistique $VPP(U,Y,h)$. Par contre, lorsqu'il s'agira de prédire la classe d'une observation dant le cluster n'est pas spécifié ou n'a pas participé à l'estimation des paramètres π_U et γ_U, on pourra se servir de la statistique $VPP(U,Y)$.

Pour adapter la procédure d'apprentissage étudiée dans le chapitre II à une analyse hiérarchique, nous allons construire l'algorithme de la recherche de l'ensemble optimal au tour de la valeur prédictive positive $VPP(U,Y,h)$ du classifieur $U(X)$ pour un cluster h donné.

Si on note par $\phi_h(U,X) = \delta_h(C)U(X)$ le classifieur généré par le profil U pour le cluster h et par $\mathcal{D} = \{(y_i, x_i, c_i); i = 1 : n\}$ l'ensemble des observations du triplet de variables (Y,X,C). On peut interpréter la sensibilité du classifieur $\phi_h(U,X)$ pour le cluster h, $\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\}$, comme une fonctionnelle de la loi a posteriori de $\phi_h(U,X)$ conditionnellement aux données $\mathcal{D}$ et à $Y = 1$. Tenant compte que

$$\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\} = VPP(U,Y,h)\frac{\Pr\{\phi_h(U,X) = 1 \mid \mathcal{D}\}}{\Pr\{Y = 1 \mid \mathcal{D}\}}$$

on a

$$\frac{\Pr\{\phi_h(U',X) = 1 \mid Y = 1,\mathcal{D}\}}{\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\}} = \left[\frac{\Pr\{\phi_h(U',X) = 1 \mid \mathcal{D}\}}{\Pr\{\phi(U,X) = 1 \mid \mathcal{D}\}}\right]\left[\frac{VPP(U',Y,h)}{VPP(U,Y,h)}\right]$$

D'où l'interprétation du quotient $\frac{VPP(U',Y,h)}{VPP(U,Y,h)}$ comme un facteur de Bayes. Comme $U' \prec U$ alors $\frac{\Pr\{\phi_h(U',X) = 1 \mid Y = 1,\mathcal{D}\}}{\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\}} \leq 1$. Plus grand est le facteur de Bayes, donc en faveur du classifieur $\phi(U',X)$, plus proche de 1 sera le quotient $\frac{\Pr\{\phi_h(U',X) = 1 \mid Y = 1,\mathcal{D}\}}{\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\}}$. Suivant le point de vue exprimé par Kass & Raftery (1995) [5] à savoir, "Le facteur de Bayes est un résumé des preuves fournies par les données en faveur d'une théorie scientifique par un modèle statistique, par opposition aux théories alternatives", on considère la grille ci-dessous pour interpréter le facteur de Bayes en faveur ou non du classifieur associé au profil le plus détaillé $U' \prec U$:

Facteur de Bayes	Interprétation
1-3.2	on ne peut pas soutenir que le profil U' est un meilleur classifieur que U
3.2-10	on peut soutenir que U' est un meilleur classifieur que U
10-100	On peut fortement soutenir que U' est un meilleur classifieur que U
≥ 100	il n'y pas de doute que U' est un meilleur classifieur que U

6 Algorithme de la procédure d'apprentissage

L'adoption de l'algorithme d'apprentissage au cas où les données sont hétérogènes nécessite au préalable un prétraitement des données. En premier lieu, il faut discrétiser les variables numériques, si il en existe, en utilisant l'une des méthodes étudiées au chapitre III. En deuxième lieu, il faut subdiviser les données en trois sous-ensembles : un ensemble d'apprentissage, un ensemble de validation et un ensemble test. La procédure de construction du classifieur peut être résumée en deux grandes étapes. Une fois que nous avons fini de construire le classifieur, il nous reste à évaluer ses performances sur l'ensemble test. Ceci constitue la troisième étape de la procédure d'apprentissage.

1. **Etape 1** : A partir d'un ensemble d'apprentissage
   (a) Générer un ensemble de profils fréquents $\mathcal{U}_\lambda$, en utilisant le paramètre d'apprentissage $\lambda = (s_0, c_0, l_0)$
   (b) Elaguer les profils redondants dans l'ensemble $\mathcal{U}_\lambda$
   (c) Sélectionner les profils qui sont significativement corrélés avec la variable réponse (test fisher)
2. **Etape 2** : A partir d'un ensemble de validation
   (a) Pour chaque profil U : Estimer $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ (par MOM ou MLE)
   (b) Pour chaque cluster h
      i. Estimer la valeur prédictive positive a posteriori de chaque profil U

$$VPP(U, Y, h) = \frac{\sum_{i=1}^{n} Y_i \delta_h(c_i)\, \phi(U, x_i) + \widehat{\pi}_U \left(\frac{1}{\widehat{\gamma}_U} - 1\right)}{\sum_{i=1}^{n} \delta_h(c_i)\, \phi(U, x_i) + \left(\frac{1}{\widehat{\gamma}_U} - 1\right)}$$

      ii. Si il existe deux profils U et U' tels que U' soit emboîté dans U :
         A. Calculer le facteur de Bayes

$$BF(U', U) = \frac{VPP(U', Y, h)}{VPP(U, Y, h)}$$

         B. On supprime le profil U si $BF(U', U) \geq 100$. Sinon on supprime le profil U'.
   (c) fin pour

   Au sortir des étapes 1 et 2, on obtient un ensemble optimal de profils $\mathcal{U}_\lambda^h$.
3. **Etape 3** : A partir d'un ensemble test
   (a) Pour chaque cluster h

i. Définir la règle de classement (classifieur) ϕ d'une observation X par

$$\phi(X,\lambda) = \begin{cases} 1 & \text{si} \quad \sum_{j=1}^{|\mathcal{U}_\lambda^h|} \phi(X,U_j) > 0 \\ \\ 0 & \text{sinon} \end{cases}$$

Le classifieur $\phi(X,\lambda)$ est un cas particulier du classifieur défini au chapitre II à la section 3.2 où on a choisi k égale à zéro. On choisit alors de classer positive une observation X lorsqu'elle vérifie au moins un profil parmi ceux qui sont dans l'ensemble $\mathcal{U}_\lambda^h$.

La première étape consiste à générer $\mathcal{U}_\lambda$ un ensemble de profils à la fois fréquents et significativement corrélés avec la variable réponse, où λ est un paramètre d'apprentissage à spécifier par l'utilisateur. D'ailleurs c'est pour des raisons d'insuffisance de mémoire que le paramètre λ est utilisé. Sinon l'idéal est de générer tous profils existant dans l'ensemble d'apprentissage. Dans la deuxième étape, il est aussi question d'estimation les paramètres π et γ pour chaque profil appartenant à $\mathcal{U}_\lambda$ et de construire un ensemble $\mathcal{U}_\lambda^h$ spécifique à chaque cluster h.
Cette procédure nécessite de subdiviser des données en trois sous-ensembles : apprentissage, validation et test. Il faut subdiviser les données de telle sorte que tous les clusters soient représentés dans chaque sous-ensemble avec la même proportion que dans l'ensemble de départ.

Bibliographie

[1] Chuang-Stein, C. An application of the beta-binomial model to combine and monitor medical event rates in clinical trials. *Drug Information Journal 27*, 2 (1993), 515–523. 113

[2] Efron, B., and Morris, C. Empirical bayes on vector observations : An extension of stein's method. *Biometrika 59*, 2 (1972), 335. 108

[3] Efron, B., and Morris, C. N. Multivariate empirical bayes and estimation of covariance matrices, 1974. 108

[4] Griffiths, D. A. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics 29*, 4 (1973), 637. 110

[5] Kass, R., and Raftery, A. Bayes factors. *Journal of the American Statistical Association 90*, 430 (1995), 773–795. 120

[6] Kleinman, J. C. Proportions with extraneous variance : Single and independent sample. *Journal of the American Statistical Association 68*, 341 (1973), 46. 113

[7] Morris, C. N. Parametric empirical bayes inference : Theory and applications. *Journal of the American Statistical Association 78*, 381 (1983), 47. 108

[8] Robbins, H. Asymptotically subminimax solutions of compound statistical decision problems. In *Second Berkeley Symposium on Mathematical Statistics and Probability* (1951), vol. -1, pp. 131–149. 108

Appendices

Annexe C

Annexe Chapitre IV

B.1 Existence de l'estimation des moments des paramètres d'une Bêta-Binomiale

Généralement on pose

$$\frac{S_k}{n_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} Y_{ki}$$

On a

$$\mathbb{V}ar\left(\left[\frac{S_k}{n_k}\right]^2\right) < 1 \Rightarrow \frac{1}{m}\sum_{k=1}^{m}\left(\left[\frac{S_k}{n_k}\right]^2 - \mathbb{E}\left(\left[\frac{S_k}{n_k}\right]^2\right)\right) \xrightarrow{p.s} 0$$

$$\begin{aligned}
\mathbb{E}\left(\left[\frac{S_k}{n_k}\right]^2\right) &= \frac{1}{n_k^2}\mathbb{E}\left(\mathbb{E}\left(S_k^2|\theta_k\right)\right) \\
&= \frac{1}{n_k^2}\mathbb{E}\left(n_k\theta_k(1-\theta_k) + n_k^2\theta_k^2\right) \\
&= \frac{1}{n_k}\pi + \frac{n_k-1}{n_k}\left(\pi(1-\pi)\gamma + \pi^2\right)
\end{aligned}$$

On obtient par la suite

$$\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left(\left[\frac{S_k}{n_k}\right]^2\right) = \pi\left(\frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k}\right) + \left[\pi(1-\pi)\gamma + \pi^2\right]\left(\frac{1}{m}\sum_{k=1}^{m}\left(1-\frac{1}{n_k}\right)\right)$$

Si on remplace le terme à gauche de l'équation par sa valeur empirique, on obtient

$$\widehat{\gamma} = \frac{\frac{1}{m}\sum_{k=1}^{m}\left(\frac{S_k}{n_k}\right)^2 - \widehat{\pi}\left(\frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k}\right) - \widehat{\pi}^2\left(1-\frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k}\right)}{\widehat{\pi}(1-\widehat{\pi})\left[\frac{1}{m}\sum_{k=1}^{m}\left(1-\frac{1}{n_k}\right)\right]}$$

Par ailleurs, on a

$$\widehat{\pi}(1-\widehat{\pi})\left[\frac{1}{m}\sum_{k=1}^{m}\left(1-\frac{1}{n_k}\right)\right] \geq 0$$

Donc le signe de $\widehat{\gamma}$ dépend de son numérateur. Or si on pose

$$\begin{aligned} a &= \left(\frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k}\right) \geq 0 \\ b &= \frac{1}{m}\sum_{k=1}^{m}\left(\frac{S_k}{n_k}\right)^2 \geq 0 \end{aligned}$$

on obtient

$$\widehat{\gamma} = \frac{b-(\widehat{\pi}a+\widehat{\pi}^2(1-a))}{\widehat{\pi}(1-\widehat{\pi})\left[\frac{1}{m}\sum_{k=1}^{m}\left(1-\frac{1}{n_k}\right)\right]}$$

On a $\widehat{\pi}a+\widehat{\pi}^2(1-a) \in [\widehat{\pi}^2, \widehat{\pi}]$ car c'est une combinaison linéaire convexe. A l'aide de l'inégalité de la variance, on a aussi

$$\left(\frac{1}{m}\sum_{k=1}^{m}\frac{S_k}{n_k}\right)^2 \leq b \leq \frac{1}{m}\sum_{k=1}^{m}\frac{S_k}{n_k}$$

Puisque

$$\frac{1}{m}\sum_{k=1}^{m}\frac{S_k}{n_k} = \widehat{\pi}$$

alors $b \in [\widehat{\pi}^2, \widehat{\pi}]$.
Le signe de $\widehat{\gamma}$ dépend donc de la suite (S_k, n_k). Cette équation des moments, comme d'autres proposées dans la littérature, n'admettent pas toujours une solution dans $]0,1[\times]0,1[$; d'où le recourt à une méthode de pondération empirique.

B.2 Estimation par simulation des performances des estimateurs obtenus par la méthode de pondération empirique

B.2.1 Organisation des simulations

Avant d'étudier les propriétés statistiques des estimateurs, nous allons décrire la simulation d'un échantillon Bêta-binomial. Nous simulons un échantillon Bêta-binomial de la manière suivante :

1. On se donne n_U, l'ensemble des observations d'étude vérifiant le profil $U(X)$. Nous supposons avoir disposé de n_U observations constituées à partir de m réalisations de la variable $[Y,X]^{\mathcal{L}}$, où chaque réalisation $[Y,X]_h$ de $[Y,X]^{\mathcal{L}}$ est une suite d'observations indépendantes $(Y_i,X_i)_{i=1:n_h}$ de taille n_h.
2. On génère m réalisations $(\theta_h^U)_{h=1:m}$ d'une loi Bêta de paramètres α_U et β_U donnés. Ensuite on construit une suite $(n_{hU})_{h=1:m}$ telle que $\sum n_{hU} = n_U$.

3. Pour chaque h, on simule n_{hU} observations d'une loi de Bernoulli de probabilité de succès θ_h^U. Ainsi pour chaque couple (α_U, β_U), nous pouvons disposer des statistiques $(S_{hU})_{h=1:m}$ et $(n_{hU})_{h=1:m}$.

On appelle l'échantillon $(S_{hU}, n_{hU})_{h=1:m}$ un échantillon Bêta-Binomial puisqu'il est obtenu à partir d'une combinaison d'une loi Bêta et d'une loi Binomiale.

B.2.2 Présentation et analyse des résultats

Pour étudier des propriétés statistiques des estimations, on suppose avoir $n_U = 100000$ observations constituées à partir de $m = 50$ réalisations de $[Y, X]^{\mathcal{L}}$. On se fixe une valeur de 0.007 pour le paramètre π_U et on fait varier le paramètre γ_U avec les valeurs suivantes :0.005, 0.0075, 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75. Nous avons fait le choix de ces valeurs pour simuler des données semblables à nos données réelles. Par exemple, pour le couple $\pi_U = 0.007$ et $\gamma_U = 0.01$, un aperçu de la forme de la densité de la loi Bêta associée est représentée ci dessous.

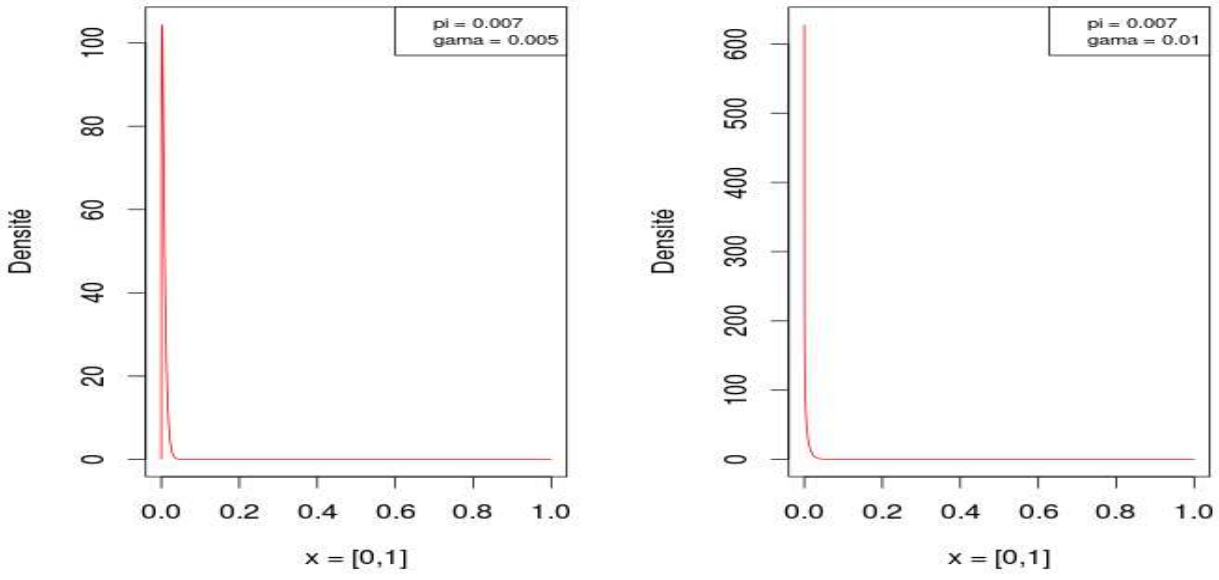


Figure A.1 – Forme de la densité de Bêta

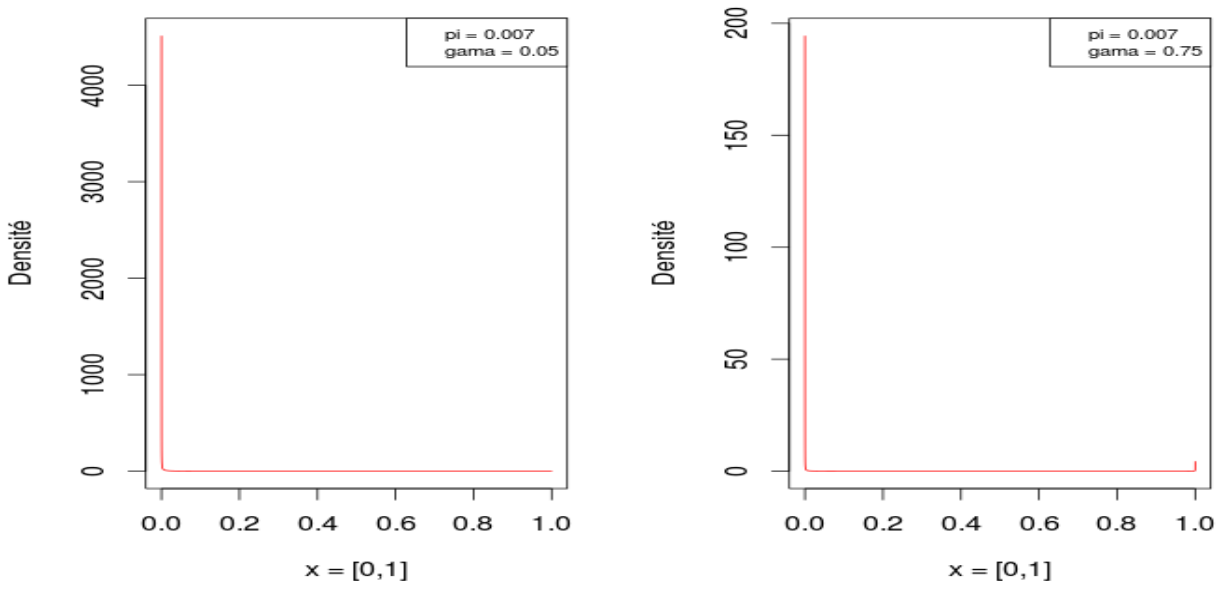


Figure A.2 – Forme de la densité de Bêta

Pour chaque combinaison (π_U, γ_{Uj}) ; $j = 1 : 10$, on en déduit un couple (α_U, β_U) à partir duquel un échantillon Bêta-Binomial $(S_{hU}, n_{hU})_{h=1:m}$ est généré. Ainsi à chaque couple (π_U, γ_U) correspond un échantillon Bêta-Binomial. En combinant les valeurs de π_U et de γ_U, nous simulons 10 échantillons Bêta-Binomial sur lesquels les paramètres π_U et γ_U seront estimés. Dans le tableau A.1, nous présentons les estimations obtenues à partir des équations des moments proposées par Kleinman que nous notons MOMK, les estimations obtenues à partir des équations des moments proposées dans cette analyse que nous notons par MOMG et les estimations obtenues par la méthode du maximum de vraisemblance notées EMV, pour des valeurs de π_U et γ_U fixées.

		MOMK		MOMG		EMV	
π_U	γ_U	$\widehat{\pi}_U$	$\widehat{\gamma}_U$	$\widehat{\pi}_U$	$\widehat{\gamma}_U$	$\widehat{\pi}_U$	$\widehat{\gamma}_U$
0.007	0.0050	0.0062	0.0029	0.0062	0.0029	0.0061	0.0045
0.007	0.0075	0.0088	0.0055	0.0088	0.0054	0.0088	0.0058
0.007	0.0100	0.0080	0.0086	0.0080	0.0090	0.0080	0.0073
0.007	0.0250	0.0079	0.0205	0.0079	0.0202	0.0079	0.0213
0.007	0.0500	0.0091	0.0522	0.0091	0.0511	0.0090	0.0636
0.007	0.0750	0.0045	0.0532	0.0045	0.0526	0.0045	0.0437
0.007	0.1000	0.0077	0.2118	0.0077	0.2072	0.0080	0.1455
0.007	0.2500	0.0070	0.3013	0.0070	0.2946	0.0064	0.3395
0.007	0.5000	0.0018	0.0707	0.0018	0.0697	0.0017	0.1860
0.007	0.7500	0.0197	0.9866	0.0197	0.9666	0.0156	0.8187

Tableau A.1 – Valeurs estimées des paramètres π et γ

A travers ce tableau, on constate que, pour les deux méthodes MOMK et MOMG, nous avons la même estimation de $\widehat{\pi}$ quelque soient les valeurs du couple (π,γ). Ceci est justifié par le fait que nous avons utilisé le même estimateur de π dans les deux méthodes. On constate aussi que la valeur estimée de $\widehat{\pi}$ par la méthode EMV est peu différente de la valeur estimée de $\widehat{\pi}$ par les deux premières méthodes. Cependant on note une différence entre les trois approches aux niveaux des estimations de $\widehat{\gamma}$. Les résultats présentés dans le tableau ci-dessus ne nous permettent pas de départager les trois méthodes. Par contre, on peut comparer les trois approches en calculant les racines carrées des erreurs quadratiques moyennes des estimateurs en procédant par simulation.

Nous considérons les valeurs d'apprentissage suivantes : $\pi_U = 0.007$ et $\gamma_U = (0.005, 0.05)$. Pour chaque couple (π_U, γ_U) fixé, nous porterons nos simulations sur les couples suivants : ($n_U = 20\,000$, $m = 10$), ($n_U = 50\,000$, $m = 50$), ($n_U = 100\,000$, $m = 100$), ($n_U = 200\,000$, $m = 150$), ($n_U = 300\,000$, $m = 200$), ($n_U = 400\,000$, $m = 250$), ($n_U = 500\,000$, $m = 300$), ($n_U = 600\,000$, $m = 350$) et ($n_U = 700\,000$, $m = 400$). Pour chaque couple (n_U, m) fixé, on simule $B = 250$ échantillons Bêta-Binomial sur lesquels on estime π et γ pour chaque échantillon. Et à la fin on calcule la racine carrée de l'erreur quadratique moyenne correspondante de chaque paramètre dans chaque méthode. Les résultats obtenus sont présentés dans les tableaux ci-dessous.

B=250

	n=20.000, m=10			n=50.000, m=50			n=100.000, m=100		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00193	0.00193	0.00193	0.00105	0.00106	0.00105	0.00075	0.00075	0.00075
RMSE ($\widehat{\gamma}$)	0.00407	0.00337	0.00315	0.00238	0.00204	0.00152	0.00155	0.00147	0.00121

B=250

	n=200.000, m=150			n=300.000, m=200			n=400.000, m=250		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00055	0.00055	0.00055	0.00048	0.00048	0.00049	0.00041	0.00041	0.00041
RMSE ($\widehat{\gamma}$)	0.00126	0.00114	0.00087	0.00095	0.00088	0.00067	0.00091	0.00091	0.00064

B=250

	n=500.000, m=300			n=600.000, m=350			n=700.000, m=400		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00037	0.00037	0.00037	0.00040	0.00040	0.00040	0.00037	0.00037	0.00036
RMSE ($\widehat{\gamma}$)	0.00088	0.00075	0.00060	0.00072	0.00070	0.00056	0.00067	0.00065	0.00049

Tableau A.2 – Racines carrées des erreurs quadratiques moyennes des estimateurs de $\pi = 0.007$ et $\gamma = 0.005$

B=250

	n=20.000, m=10			n=50.000, m=50			n=100.000, m=100		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00557	0.00560	0.00542	0.00248	0.00248	0.00244	0.00185	0.00186	0.00184
RMSE ($\widehat{\gamma}$)	0.03650	0.03622	0.04707	0.02239	0.02210	0.02024	0.01952	0.01936	0.01538

B=250

	n=200.000, m=150			n=300.000, m=200			n=400.000, m=250		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00147	0.00147	0.00147	0.00129	0.00130	0.00129	0.00122	0.00122	0.00122
RMSE ($\widehat{\gamma}$)	0.01686	0.01640	0.01134	0.01700	0.01540	0.01077	0.01403	0.01385	0.00971

B=250

	n=500.000, m=300			n=600.000, m=350			n=700.000, m=400		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00115	0.00115	0.00115	0.00104	0.00104	0.00104	0.00091	0.00091	0.00092
RMSE ($\widehat{\gamma}$)	0.01286	0.01255	0.00948	0.01057	0.01056	0.00833	0.01115	0.01112	0.00762

Tableau A.3 – Racines carrées des erreurs quadratiques moyennes des estimateurs de $\pi = 0.007$ et $\gamma = 0.05$

Les résultats présentés dans le tableau A.2 et le tableau A.3 montrent une convergence des erreurs quadratiques moyennes de $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ vers zéro pour toutes les trois méthodes. On peut constater aussi que la méthode d'estimation par le maximum de vraisemblance (EMV) est meilleur que les deux autres méthodes puisqu'elle enregistre la plus petite erreur quadratique moyenne sur les neufs échantillons simulés. Elle est suivie par la méthode MOMG qui a la deuxième plus petite erreur quadratique moyenne. En pratique, on suggère donc d'estimer les hyperparamètres par la méthode du maximum de vraisemblance.

B.3 Loi conditionnelle de θ_h^U

D'après le théorème de Bayes, on peut déterminer la distribution conditionnelle $\left[\left(\theta_h^U\right)_{h=1:m}|Y,\pi_U,\tau_U\right]$ par :

$$\begin{aligned}\left[\left(\theta_h^U\right)_{h=1:H}|Y,\pi_U,\tau_U\right] &= \frac{\left[\left(\theta_h^U\right)_{h=1:H},Y|\pi_U,\tau_U\right]}{[Y|\pi_U,\tau_U]}\\ \left[\left(\theta_h^U\right)_{h=1:m}|Y,\pi_U,\tau_U\right] &= \frac{\left[Y|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right]}{[Y|\pi_U,\tau_U]}\end{aligned}$$

On en déduit que

$$\left[\left(\theta_h^U\right)_{h=1:m}|Y,\pi_U,\tau_U\right] = \frac{\left[Y|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right]}{\int_{[0,1]^m}\left[Y,|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right]\otimes d\theta_h^U}$$

Par ailleurs, on a

$$\left[Y|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right] = \left[Y|U(X),[Y,X]^{\mathcal{L}}\right]$$

En plus nous avons $\left[\left(\theta_h^U\right)_{h=1:m}|Y,\pi_U,\tau_U\right] = \prod_{h=1}^{m}\left[\theta_h^U|Y,\pi_U,\tau_U\right]$ puisque la suite $\left(\theta_h^U\right)_{h=1:m}$ est un échantillon iid. Pour simplifier les expressions, nous posons

$$\begin{aligned}\mathbf{a} &= \delta_{(1,[Y,X]_h)}\left(U(X),[Y,X]^{\mathcal{L}}\right)\mathbb{1}_{[Y=1]}(y)\\ \mathbf{b} &= \delta_{(1,[Y,X]_h)}\left(U(X),[Y,X]^{\mathcal{L}}\right)(1-\mathbb{1}_{[Y=1]}(y))\end{aligned}$$

et

$$\Gamma(A) = \frac{\Gamma(\widehat{\tau}_x)}{\Gamma(\widehat{\pi}_x\widehat{\tau}_x)\Gamma((1-\widehat{\pi}_x)\widehat{\tau}_x)}$$

Par la suite, on obtient

$$\begin{aligned}\left[Y|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right] &= \left\{\prod_{h=1}^{m}\left[Y|\theta_h^U,[Y,X]_h\right]\right\}\left\{\prod_{h=1}^{m}\left[\theta_h^U|\pi_U,\gamma_U\right]\right\}\\ &= \left\{\prod_{h=1}^{m}\left(\theta_h^U\right)^{\mathbf{a}}\left(1-\theta_h^U\right)^{\mathbf{b}}\right\}\left\{\prod_{h=1}^{m}\Gamma(A)\left(\theta_h^U\right)^{\pi_U\tau_U-1}\left(1-\theta_h^U\right)^{(1-\pi_U)\tau_U-1}\right\}\\ &= [\Gamma(A)]^m\prod_{h=1}^{m}\left\{\left(\theta_h^U\right)^{\mathbf{a}+\pi_U\tau_U-1}\left(1-\theta_h^U\right)^{\mathbf{b}+(1-\pi_U)\tau_U-1}\right\}\end{aligned}$$

Par ailleurs, on a

$$\begin{aligned}&\int_{[0,1]^m}\left[Y,|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right]\otimes d\theta_h^U\\ &\qquad= [\Gamma(A)]^m\prod_{h=1}^{m}\left\{\int_0^1\left(\theta_h^U\right)^{\mathbf{a}+\pi_U\tau_U-1}\left(1-\theta_h^U\right)^{\mathbf{b}+(1-\pi_U)\tau_U-1}\right\}\\ &\qquad= [\Gamma(A)]^m\prod_{h=1}^{m}\frac{\Gamma\left[\mathbf{a}+\pi_U\tau_U\right]\Gamma\left[\mathbf{b}+(1-\pi_U)\tau_U\right]}{\Gamma\left[\mathbf{a}+\mathbf{b}+\tau_U\right]}\end{aligned}$$

Puisque les variables $\left(\theta_h^U\right)_{h=1:m}$ sont indépendantes et identiquement distribuées, on obtient

$$\prod_{h=1}^{m}\left[\theta_h^U|Y,\pi_U,\tau_U\right]=\prod_{h=1}^{m}\frac{\Gamma\left[\mathbf{a}+\mathbf{b}+\tau_U\right]\left(\theta_h^U\right)^{\mathbf{a}+\pi_U\tau_U-1}\left(1-\theta_h^U\right)^{\mathbf{b}+(1-\pi_U)\tau_U-1}}{\Gamma\left[\mathbf{a}+\pi_U\tau_U\right]\Gamma\left[\mathbf{b}+(1-\pi_U)\tau_U\right]}$$

donc

$$\prod_{h=1}^{H}\left[\theta_h^U|Y,\pi_U,\tau_U\right] \;=\; \prod_{h=1}^{H}\text{Beta}\left(\mathbf{a},\mathbf{b}\right)$$

On en déduit que la loi conditionnelle de θ_h^U est une loi Bêta définie par :

$$\left[\theta_h^U|Y,\pi_U,\tau_U\right]=\text{Beta}\left(\mathbf{a}+\pi_U\tau_U,\mathbf{b}+(1-\pi_U)\tau_U\right)$$

Chapitre V

Application à la Mortalité Maternelle dans les hôpitaux de référence au Sénégal et au Mali

1 Introduction

Selon l'Organisation Mondiale de la Santé (OMS), chaque année 585 000 femmes meurent dans le monde suite à des complications liées à la grossesse, à l'accouchement ou au post-partum [15]. Pour réduire cette mortalité, les politiques de santé adoptées par de nombreux pays d'Afrique subsaharienne reposent en grande partie sur la disponibilité des services de Soins Obstétricaux d'Urgence (SOU), incluant la césarienne et la transfusion sanguine, dans les hôpitaux de référence au niveau des districts ou régions sanitaires. Par contre, l'accès à ces services est très variable d'une région à une autre, avec une grande disparité entre milieu rural et urbain (Starrs, 1987). Des études réalisées en Afrique de l'Ouest, dans le cadre du suivi et de l'évaluation des interventions, ont révélé des taux de Mortalité Maternelle (MM) élevés et variables d'un hôpital à un autre au sein d'un même pays, mais aussi d'un pays à un autre [7, 9, 11–13, 20, 21].

Les résultats des études concernant les causes de la MM dans les pays en développement montre que, de tous les décès maternels qui surviennent en Afrique, 75% seraient dus à des complications obstétricales directes qui sont : les hémorragies (cause principale de la mortalité maternelle reconnue mondialement), les infections puerpérales, les dystocies, les troubles hypertensifs de la grossesse et les avortements clandestins [17]. Les causes indirectes les plus couramment rencontrées en Afrique subsaharienne sont essentiellement l'anémie, le paludisme, l'hépatite virale et le sida. Un facteur de risque de la MM se définit comme une caractéristique plus fréquente chez les mères qui meurent que celles qui ne meurent pas (OMS, 1991). Les facteurs qui prédisposent aux événements mortels de la maternité peuvent être regroupés en deux grandes catégories : les facteurs individuels liés aux femmes et les facteurs reliés au système de santé ou facteurs institutionnels.

Facteurs individuels : De nombreuses études dans les pays en développement ont montré que la primi-parité, d'autant plus qu'elle concerne une femme plus jeune, et la grande multi-parité sont des facteurs de risque importants de complication sévère, indépendamment de l'âge [4, 14]. Ce dernier est un facteur de risque majeur chez les patientes d'âges extrêmes (inférieurs à 16 ou

supérieurs à 35 ans) identifié depuis longtemps. Même si le rôle d'un espace inter génésique court (inférieur à 2 ans) sur la Mortalité Maternelle a été peu étudié, il représente un facteur de risque retrouvé très présent chez les femmes de l'Afrique de l'Ouest.

Facteurs institutionnels : Les études qui traitent des facteurs liés aux services de santé sont pour la plupart observationnelles et limitées à des comparaisons entre pays [19]. Elles révèlent cependant que le niveau de MM est plus élevé dans les pays où les femmes ont le moins accès aux services de santé équipés et de bonne qualité [2]. Parmi les femmes qui utilisent les services de santé, la mortalité reste élevée dans certains hôpitaux. Peu d'études ont été réalisées dans ce contexte. La seule étude recensée qui analyse la relation possible entre les données institutionnelles et la MM, a été réalisée dans un pays développé : les états Unis d'Amérique [18]. La particularité de cette étude réside dans l'utilisation d'une analyse multivariée de la famille des modèles linéaires généralisés, la régression de poisson, pour estimer le risque relatif, entre la disponibilité des Soins Obstétricaux d'Urgence (SOU), des médecins spécialisés en SOU et le taux de Mortalité Maternelle Humaine.

Les grandes stratégies qui devraient permettre de réduire le taux de mortalité maternelle sont connues : le recours aux soins obstétricaux essentiels tels que l'accouchement assisté par du personnel qualifié et le recours à des services offrant des soins obstétricaux d'urgence en cas de complication obstétricale sont les principales mesures recommandées [3]. Plusieurs pays ont adopté des feuilles de route qui constituent un cadre national structuré de la planification des programmes et des activités qui visent à réduire la mortalité. Leur mise en œuvre se heurte à des problèmes structurels qui affectent les systèmes de santé de la plupart des pays de l'ASS et en premier lieu le problème récurrent du financement. La question des ressources humaines est en passe de devenir le défi majeur qui limite déjà la capacité de ces systèmes de santé de faire face à des problèmes de santé déjà existants et à d'autres à venir.

Des études dans différents pays d'Afrique subsaharienne ont identifié plusieurs facteurs de risque indépendants qui diffèrent sensiblement entre les auteurs, probablement en raison des différences entre les populations d'étude, l'environnement, les variables recueillies et les méthodes statistiques utilisées. Ainsi, il reste difficile de fournir aux professionnels de la santé des pays d'Afrique subsaharienne des recommandations pour identifier les signes ou symptômes cliniques qui pourraient aider le personnel à détecter les patients à haut risque de décès à l'hôpital. C'est dans ce contexte que le projet QUARITE a été mis en place.

Le projet QUARITE est un essai randomisé par grappes multicentrique international destiné à évaluer l'efficacité d'un programme d'amélioration de la qualité des soins au Sénégal et au Mali, comparé avec un groupe contrôle (soins habituels) sans intervention extérieure [6]. Le critère d'évaluation primaire de l'essai est la mortalité maternelle en milieu hospitalier. Avec environ 80 000 naissances survenant chaque année dans 46 hôpitaux de référence, QUARITE est l'un des plus grands essais randomisés par grappes dans la santé maternelle et périnatale jamais entrepris dans les pays à faibles revenus.

Ainsi, le processus expérimental donne une occasion unique d'évaluer la mortalité maternelle en milieu hospitalier à partir d'un grand nombre de centres, dans une variété de contextes, et en tenant compte des différentes caractéristiques de la mère et de l'hôpital. Dans cette analyse, il est question de mesurer la mortalité maternelle dans les hôpitaux de référence au Mali et au Sénégal avant la mise en œuvre du programme d'amélioration de la qualité des soins et d'évaluer les prédicteurs de mortalité à l'hôpital chez les patients qui fréquentent ces établissements de santé.

2 Présentation des données et objectifs de l'étude

Les données de l'étude ont été recueillies au cours de l'exécution du projet QUARITE dans sa phase de pré-intervention qui s'est déroulée du 1^{er} Octobre 2007 au 30 Septembre 2008 au Sénégal et du 1^{er} Novembre 2007 au 31 Octobre 2008 au Mali. Les données considérées sont issues d'un échantillonnage à deux niveaux : un niveau hôpital et un niveau patiente. Les hôpitaux qui ont participé à la collecte des données ont été tirés au hasard parmi ceux de leurs pays : (1) disposant d'un bloc opératoire fonctionnel, (2) pratiquant au moins 800 accouchements par an, (3) ayant un consentement signé par le chef de service de la maternité et le directeur de l'établissement. Au total 46 hôpitaux de référence, dont 24 au Sénégal et 22 au Mali, ont été enrôlés dans l'étude. La population ciblée est l'ensemble des femmes enceintes qui sont prises en charge dans les hôpitaux de référence. Sont incluses, les femmes admises pour un accouchement et les patientes dirigées secondairement vers un des hôpitaux concernés par l'étude. Elles sont exclues : les femmes admises après un accouchement à domicile et les femmes prises en charge dans une autre structure. Au total 89 518 patientes sont incluses parmi lesquelles 617 sont décédées. Soit un taux de 0.7%. L'hôpital constitue l'unité de randomisation et d'intervention pendant que la patiente admise pour un accouchement représente l'unité d'analyse.

Seules les données patientes sont analysées dans ce travail. L'échantillon d'étude est constitué de 89518 patientes décrites par 24 variables explicatives réparties en trois groupes : un premier groupe de sept variables décrivant l'état de la patiente avant la grossesse en cours, un deuxième groupe de onze variables portant sur l'état d'avancement de la grossesse et un troisième groupe de six variables relatant le cours de l'accouchement. Plus une variable réponse binaire. Elle prend la valeur 1 si la patiente décède avant d'être autorisée à quitter l'hôpital (617 patientes) et 0 sinon (88 901 patientes). L'analyse des données a deux objectifs. Dans un premier temps, on cherche à : (1) Identifier les profils caractéristiques des patientes décédées sans tenir compte de l'échantillonnage au niveau hôpital ; (2) Elaborer une règle de classification performante et facile à comprendre comme un arbre de décision ou une régression logistique. Et dans un deuxième temps, on cherche à : (1) Identifier les profils caractéristiques des patientes décédées sachant que les hôpitaux de références sont échantillonnés à partir d'un ensemble d'hôpitaux éligibles ; (2) Elaborer une règle de classification performante et facile à comprendre comme un arbre de décision ou une régression logistique selon l'hôpital.

3 Prétraitement des données

Parmi les 24 variables explicatives de l'échantillon, nous avons 21 variables catégorielles et 3 variables numériques dont l'âge de la patiente, la parité (le nombre d'accouchements précédents la grossesse en cours) et le nombre de consultations prénatales pour la grossesse en cours. Pour se mettre dans les conditions d'application de l'algorithme d'apprentissage du chapitre II, nous avons discrétisé les variables numériques en utilisant la méthode du principe de la longueur de description minimal ou "Minimal Description Length Principle" (voir annexe B). Ci-dessous, nous présentons la liste des variables explicatives et leurs modalités respectives selon les groupes d'appartenance.

Variables	**modalités**
Historique des antécédents médicaux	
Groupe d'âge (en années)	< 30 ≥ 30
Parité (nombre d'accouchements)	< 5 ≥ 5
Hypertension artérielle chronique	0 1
Cardiaque chronique / Insuffisance rénale	0 1
Broncho-pneumopathie chronique	0 1
Drépanocytose	0 1
Antécédent césarienne	0 1

Tableau V.1 – Liste des variables : historique des antécédents médicaux

Variables	**modalités**
Grossesse en cours	
Hypertension gestationnelle	0 1
Pré-éclampsie/éclampsie	0 1
Saignement vaginal (près du terme)	0 1
Anémie chronique Sévère	0 1
Diabète gestationnel	0 1
Rupture prématurée des membranes	0 1
Tractus urinaire infection / pyélonéphrite	0 1
VIH / SIDA	0 1
Paludisme	0 1
Grossesse multiple	0 1
Nombre de consultations prénatales	< 3, $=3$, ≥ 4

Tableau V.2 – Liste de variables : Grossesse en cours

Variables	modalités
Travail et accouchement	
Evacuer par un autre établissement de santé	0 1
Induction du travail	0 1
Mode d'accouchement	
voie vaginale	0
forceps / ventouse	1
urgence avant l'accouchement césarienne	2
intrapartum accouchement par césarienne	3
césarienne élective	4
Hémorragie post-partum antécédent ou immédiat	0 1
Travail prolongé / dystocique	0 1
Rupture utérine	0 1

Tableau V.3 – Liste des variables : Travail et accouchement

4 Analyse des données sous l'hypothèse que la population est homogène

4.1 Echantillonnage des données

L'apprentissage statistique que nous proposons dans cette analyse nécessite de subdiviser la base de données en trois échantillons de même taille : *Apprentissage*, $Validation$ et $Test$. Les échantillons sont obtenus de manière à ce qu'une partie des $Hopitaux$ serve à l'apprentissage et à la validation du modèle et l'autre partie des clusters soit utilisée pour tester la performance du modèle. On note n le nombre total des patientes inclues dans l'étude et $m = 46$ le nombre total d'hôpitaux.

En fonction de la valeur m_0 donnée, soit l'échantillon $Test$ est constitué exclusivement d'hôpitaux qui n'ont pas servi à l'élaboration du classifieur ; soit il contient un faible taux d'observations des hôpitaux qui ont participé à la construction du classifieur. Ce procédé permet, à l'aide du classifieur, de faire des prédictions plus tard sur des hôpitaux qui n'ont pas participé à l'étude. Pour la suite, nous nous sommes fixés de manière arbitraire une valeur m_0 égale à 36.

Algorithme : Echantillonnage des données

- Entrées : $\mathcal{D}$ un ensemble d'observation ; m le nombre de clusters dans $\mathcal{D}$, m_0 un entier supérieur à $m/2$
- Sorties : Echantillons : $Train$, $Valid$, $Test$

1 : n : le nombre d'observations dans $\mathcal{D}$
2 : $n_0 = \lfloor n/2 + 0.5 \rfloor$
3 : k = tirage aléatoire sans remise de m_0 clusters parmi les m clusters dans $\mathcal{D}$
4 : n_1 : le nombre d'observations dans les k clusters
5 : $Test$ = les observations qui ne sont pas dans les k clusters
6 : **Si** $(n_1 > 2 * n_0)$ **faire**
7 : Sub = tirage aléatoire sans remise de $(2 * n_0)$ observations parmi les n_1 observations
8 : $Test$ = Ajouter dans l'échantillon Test les $(n_1 - 2 * n_0)$ observations restantes
9 : **sinon** Sub : permuter les n_1 observations
10 : **Fin si**
11 : $Train$ = la première moitié des observations dans Sub constitue l'ensemble d'apprentissage
12 : $Valid$ = la deuxième moitié des observations dans Sub constitue l'ensemble validation

Tableau V.4 – Algorithme d'échantillonnage

4.2 Construction du classifieur

A partir de l'ensemble d'apprentissage, on a appliqué la procédure *"apriori"* (algorithme III.1) du package "arules" avec le paramètre d'apprentissage $\lambda = (s_0, c_0, l_0)$. Au support minimum s_0, on a affecté les valeurs suivantes : 9.10^{-4}, 1.10^{-3}, 2.10^{-3} et 3.10^{-3}. A la valeur prédictive positive minimale c_0 on a alloué les valeurs suivantes : 5%, 4%, 3% et 2%. Et on a fixé la longueur maximale l_0 à 5.
Pour chaque combinaison des trois paramètres, on génère un ensemble de profils fréquents ($\mathcal{U}_\lambda$). Puis à l'aide de la procédure d'élagage (algorithme III.2), on supprime tous les profils redondants dans $\mathcal{U}_\lambda$ pour obtenir un ensemble de profils $\mathcal{U}_\lambda^1$ de taille plus petite. En général, l'ensemble $\mathcal{U}_\lambda^1$ est très vaste au point qu'on ne peut pas s'en servir pour construire un classifieur efficace. On se sert alors de l'algorithme III.3 pour réduire l'ensemble $\mathcal{U}_\lambda^1$. Cette étape de la procédure permet de supprimer tous les profils dans $\mathcal{U}_\lambda^1$ de faibles performances. On obtient alors un ensemble $\mathcal{U}_\lambda^2$ contenant les profils de meilleurs performances et non redondants. De cet ensemble, on pourra alors déduire un classifieur ϕ performant. L'ensemble Test servira à calculer les performances du classifieur (sensibilité, spécificité et erreur de classement).

La combinaison des différents paramètres conduit à la construction de 16 règles de classement (classifieurs). Le meilleur classifieur est sélectionné à partir de cet ensemble (voir tableau ci-dessous **??**) en variant la sensibilité de chaque classifieur par rapport à sa spécificité. Toutes les analyses relatives à la méthode de classification proposée ont été réalisées dans l'environnement de programmation R

[16]. L'exploitation des règles d'association a été faite en utilisant le package arules [1, 10].

4.3 Recherche d'un classifieur optimal

Les 16 classifieurs considérés sont des classifieurs binaires discrets [8]. Ils produisent chacun un seul point dans l'espace ROC (Receiver Operating Characteristics). Officiellement, un point dans l'espace ROC est meilleur qu'un autre si il est au nord-ouest (sensibilité élevée, 1-spécificité faible) par rapport à l'autre. Relativement au classifieur, plus l'aire en-dessous de la courbe ROC est élevée, meilleur est le classifieur.

Habituellement, pour comparer des classifieurs, on compare les taux d'erreur de classement associés. Cependant, dans le contexte où la distribution des classes de la variables réponse est déséquilibrée, il est plus approprié d'utiliser l'aire en-dessous de la courbe ROC. L'aire sous la courbe ROC, communément notée AUC, a une propriété statistique importante. Le AUC d'un classifieur peut être traduit comme suit : la probabilité de classer une observation positive choisie de manière aléatoire est plus élevée que celle d'une observation négative choisie au hasard. A ces deux indicateurs de performance on a associé le score de Pierce afin de disqualifier les classifieurs générant trop de fausses alertes. Pour choisir le meilleur, on compare en premier lieu les scores de Pierce (PSS). On choisit les cinq meilleurs classifieurs. Ensuite on compare leurs AUC, puis leurs erreurs de classement, leurs sensibilités et leurs spécificités avant de comparer les tailles de leurs ensembles optimaux de profils $\mathcal{U}_\lambda^2$.

Paramètres d'apprentissage		Profils		Performances				
Support min	VPP (conf) min	Taille $\mathcal{U}_\lambda$	Taille $\mathcal{U}_\lambda^2$	Sensibilité	Spécificité	Erreur	AUC	PSS
9.10^{-04}	0.03	5988	44	0.81	0.79	0.21	0.80	0.60
9.10^{-04}	0.04	3971	34	0.78	0.85	0.15	0.81	0.63
9.10^{-04}	0.05	2957	18	0.66	0.92	0.09	0.79	0.58
1.10^{-03}	0.03	5054	40	0.84	0.78	0.22	0.81	0.62
1.10^{-03}	0.04	3373	27	0.75	0.86	0.14	0.80	0.61
1.10^{-03}	0.05	2518	15	0.61	0.92	0.08	0.77	0.53
2.10^{-03}	0.03	1522	13	0.79	0.80	0.20	0.79	0.59
2.10^{-03}	0.04	1152	03	0.39	0.98	0.03	0.69	0.37
2.10^{-03}	0.05	1050	03	0.39	0.98	0.03	0.69	0.37
3.10^{-03}	0.03	725	04	0.65	0.86	0.14	0.76	0.51
3.10^{-03}	0.04	610	02	0.46	0.94	0.06	0.70	0.40
3.10^{-03}	0.05	610	02	0.46	0.94	0.06	0.70	0.40

Tableau V.5 – Tableau des performances des 12 ensembles optimaux obtenus à partir du test asymptotique

Le classifieur dont les performances sont représentées à la ligne 07 du tableau V.5 est le meilleur classifieur selon les critères de sélection énumérés précédemment. Dans le tableau V.6, nous représentons la matrice de confiance qui lui est associée.

	Observations			sensibilité	spécificité	erreur clmt
Prédictions		non	oui			
	non	23610	45	0.789	0.797	0.204
	oui	6017	168			
	total	29627	213			

Tableau V.6 – Matrice de confusion du classifieur optimal par test asymptotique

4.4 Structure de l'arbre constitué par les profils de risque composant le classifieur optimal

Une structure d'arbre peut être utilisée pour visualiser les règles de l'ensemble optimal ($\mathcal{U}_\lambda^2$) des profils à risque qui constituent la règle de classement (classifieur optimal). Cette arborescence permet de présenter la règle de classement sous une forme facile à comprendre comme un arbre de décision. Chaque branche de l'arbre constitue un profil à risques dont le risque relatif associé est donné au niveau de la feuille terminale de la branche.

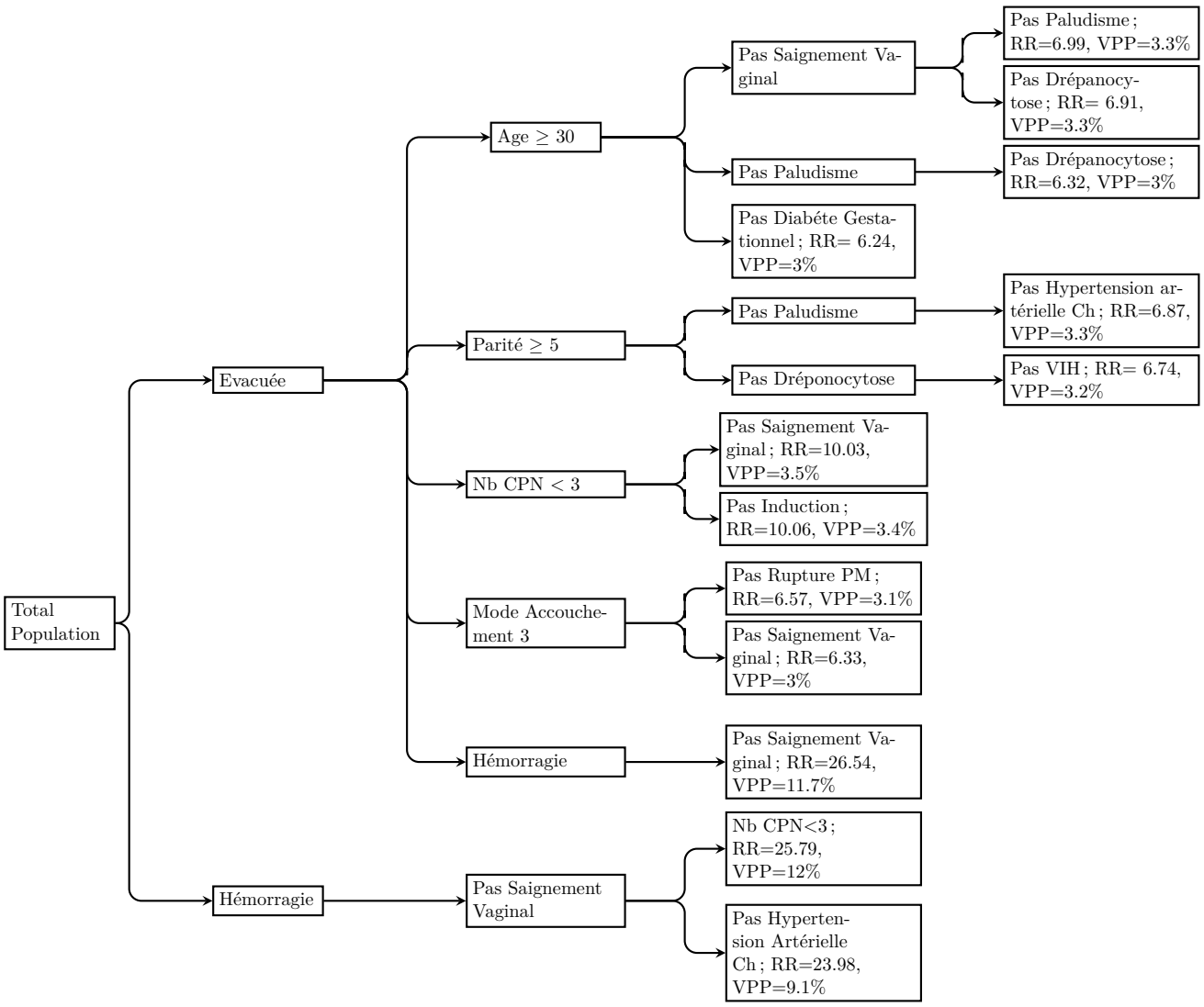


Figure V.1 – Représentation de l'arbre des profils à risque

Dans cette analyse, nous avons posé une hypothèse forte à savoir que les observations sont toutes indépendantes et identiquement distribuées. Ce n'est pas tout à fait exact puisque les données de l'étude (QUARITE) sont obtenues à partir d'un essai multicentrique randomisé. Supposer donc que les distributions des observations dans les différents hôpitaux de référence sont identiques peut avoir une forte influence sur les résultats de l'analyse. Il serait donc judicieux d'analyser les données en tenant compte de l'effet aléatoire au niveau des hôpitaux.

5 Analyse des données sous l'hypothèse que la population est hétérogène

Dans cette deuxième partie de l'analyse, nous considérons la méthode de sélection de l'ensemble optimal en utilisant le facteur de Bayes. On choisit 40 hôpitaux de manière aléatoire dont il faudra subdiviser en trois sous ensembles : Apprentissage, Validation et Test, de tels sorte que chaque sous-ensemble contient les 40 hôpitaux avec la même proportion. Les 6 hôpitaux restants seront ajoutés à l'ensemble test construit précédemment.
Pour générer l'ensemble des profils non redondants et significativement corrélés avec la variable réponse, nous avons considéré les paramètres d'apprentissage suivant : $s_0 = 2.10^{-3}$, $c_0 = 0.03$ et $l_0 = 5$. Ce choix est dû au fait que ces paramètres ont fourni le meilleur classifieur dans le cas iid. Au sortir de la première étape de la procédure d'apprentissage, on a obtenu un ensemble de 1522 profils non redondants et significativement corrélés avec la variable réponse.

5.1 Présentation des résultats pour les hôpitaux ayant participé à l'estimation des hyperparamètres

Pour chaque hôpital participant à l'élaboration du classifieur (à l'estimation des hyperparamètres de la Bêta-Binomiale), la sélection de l'ensemble optimal de profils est effectuée en fonction de la valeur prédictive positive a posteriori. Dans les tableaux ci-dessous (Tableau V.7 et Tableau V.8), on a présenté les résultats du classement avec les valeurs prédictives positives a posteriori.

Hopital	taille $\mathcal{U}_\lambda^h$	sensib	spécif	erreur	auc
01	09	1.00	0.98	0.02	0.99
03	09	0.92	0.78	0.22	0.85
04	09	0.92	0.54	0.45	0.73
05	09	-	0.97	0.03	-
06	09	0.00	0.92	0.08	0.54
07	09	-	0.97	0.03	-
08	09	0.67	0.88	0.12	0.78
09	09	0.60	0.79	0.21	0.69
11	09	1.00	0.67	0.32	0.84
14	09	0.71	0.59	0.41	0.65

Tableau V.7 – Tableau de performance pour les hôpitaux ayant participé à l'estimation des hyperparamètres

Hopital	taille $\mathcal{U}_\lambda^h$	sensib	spécif	erreur	auc
15	09	1.00	0.73	0.27	0.86
16	09	0.83	0.69	0.31	0.76
17	09	0.80	0.81	0.19	0.81
18	09	1.00	0.61	0.39	0.80
19	09	0.75	0.96	0.04	0.85
20	09	0.71	0.66	0.34	0.68
23	09	0.78	0.72	0.28	0.75
24	09	1.00	0.77	0.23	0.88
25	09	0.78	0.83	0.17	0.80
26	09	1.00	0.69	0.30	0.85
27	09	0.33	0.67	0.34	0.50
28	09	1.00	0.82	0.18	0.91
29	09	0.33	0.71	0.29	0.52
30	09	1.00	0.52	0.48	0.76
31	09	1.00	0.84	0.15	0.92
32	09	1.00	0.77	0.23	0.88
33	09	0.60	0.50	0.50	0.55
34	09	1.00	0.80	0.20	0.90
35	09	1.00	0.83	0.17	0.92
36	09	0.50	0.76	0.24	0.63
37	09	1.00	0.66	0.34	0.83
38	09	1.00	0.81	0.19	0.90
39	09	0.75	0.61	0.38	0.68
40	09	0.50	0.59	0.41	0.54
41	09	1.00	0.57	0.43	0.78
42	09	1.00	0.82	0.17	0.91
43	09	1.00	0.92	0.09	0.96
44	09	1.00	0.88	0.12	0.94
45	09	1.00	0.87	0.13	0.94
46	09	-	0.80	0.20	-

Tableau V.8 – Tableau de performance pour les hôpitaux ayant participé à l'estimation des hyperparamètres

Si on prend le seuil de sélection du facteur de Bayes égale à 3, on distingue trois classifieurs pour tous les hôpitaux : un classifieur $\mathcal{C}_1$ de 08 profils pour les hôpitaux 03, 04, 11, 16, 17, 18, 20, 23, 24 ; un classifeur $\mathcal{C}_2$ de 09 profils dont les 08 profils du classifieur $\mathcal{C}_1$ plus le profils "$\{Hemorragie = 1, SaignementV = 0\}$" pour le hôpital 09 et un classifieur $\mathcal{C}_3$ de 09 profils dont les 08 profils du classifieur $\mathcal{C}_1$ plus le profils "$\{Hemorragie = 1\}$" pour le reste des hôpitaux. Par contre, lorsqu'on prend un seuil de sélection supérieur ou égale à 10, on a un classifieur unique de 09 profils pour tous les hôpitaux qui ont participé a l'estimation des hyperparamètres. Il s'agit du classifieur $\mathcal{C}_3$. A la Figure V.2, nous avons une présentation sous forme d'arbre de l'ensemble des profils optimaux qui constituent le classifieur $\mathcal{C}_3$.

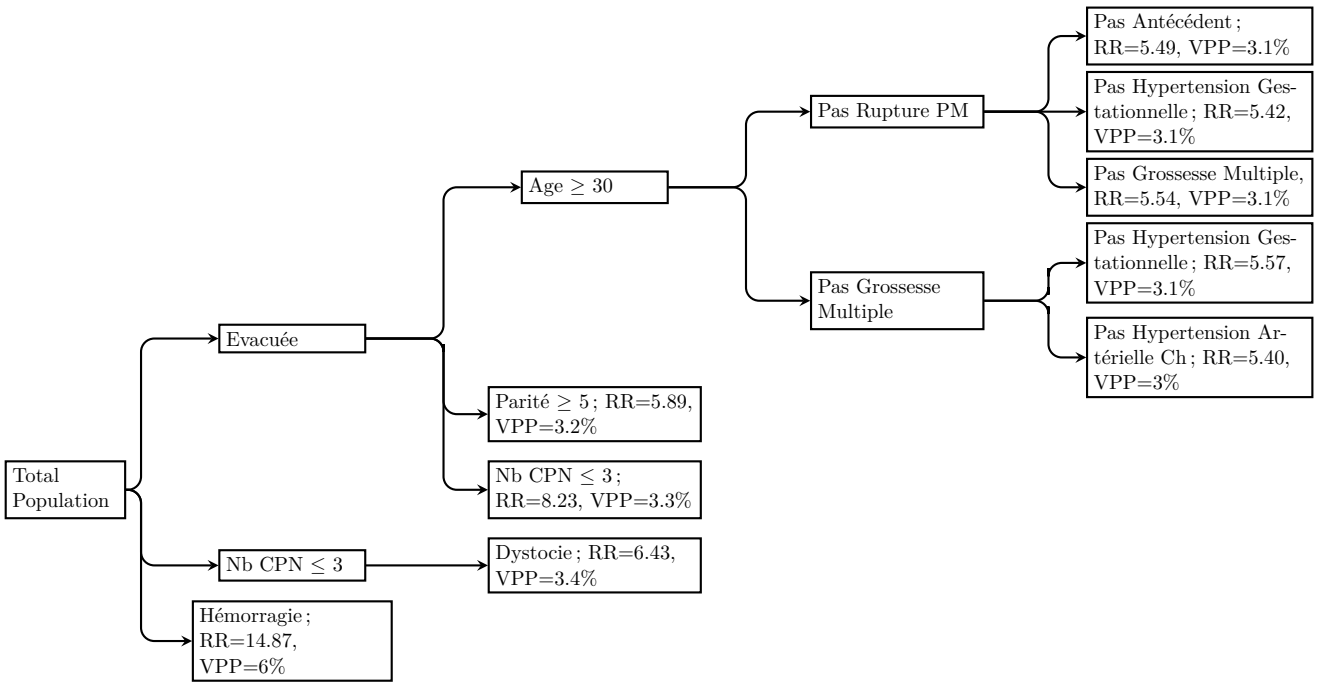


Figure V.2 – Représentation sous forme d'arbre des profils à risque communs à tous les hôpitaux

5.2 Présentation des résultats pour les hôpitaux n'ayant pas participé à l'estimation des hyperparamètres

Pour les hôpitaux qui n'ont pas participé à l'élaboration du classifieur, il s'agira de construire un classifieur moyen dont on pourra utiliser pour faire le classement dans tout nouveau hôpital. La sélection de l'ensemble optimal de profils se fera en fonction de l'estimation de l'hyperparamètre π. Dans le tableau ci-dessous (Tableau V.9), on a présenté les résultats du classement avec les valeurs prédictives positives en tenant compte de l'hétérogénéité des données. Le classifeur moyen obtenu est identique au classifieur $\mathcal{C}_3$ présenté à la figure V.2.

Hopital	taille $\mathcal{U}_\lambda^2$	sensib	spécif	erreur	auc
02	09	0.91	0.88	0.12	0.89
10	09	0.87	0.70	0.29	0.79
12	09	0.60	0.88	0.13	0.74
13	09	1.00	0.88	0.12	0.94
21	09	1.00	0.95	0.05	0.97
22	09	0.73	0.78	0.23	0.75

Tableau V.9 – Tableau de performance pour les hôpitaux n'ayant pas participé à l'estimation des hyperparamètres

6 Discussion

Pour un usage clinique, les structures d'arbre de la figure V.1 et de la figure V.2 sont utilisées pour visualiser les profils explorés. Chaque branche de l'arbre constitue un profil à risque dont le risque relatif associé est donné au niveau du nœud terminal de la branche. Dans chaque nœud une paire "variable-modalité" est représentée. Chaque branche détermine une partition de la sous-population à risque. Par exemple, selon la figure V.2, les patientes qui présentent une hémorragie sont 14.87 fois plus susceptibles de mourir que la moyenne de la population. Pour les patientes qui ont une dystocie et qui ont effectué moins de trois consultations prénatales présente un risque relatif de 6.43. Le RR passe à 8.23 pour les patientes qui ont été évacuée à partir d'un autre établissement et qui ont moins de trois consultations prénatales pendant la grossesse. Des interprétations similaires de l'arborescenceV.1 peuvent être faites pour les branches qui identifient respectivement les patientes évacuées et âgées de plus de 30 ans ou ayant une parité supérieur à 5.

La règle de classement établie dans cette étude confirme que les patientes présentant une hémorragie, un accouchement prolongé ou une parité supérieure ou égale à 5, doivent être gérées avec une haute priorité par les professionnels de santé qualifiés dans les services SOU complets [3], en particulier si la patiente est évacuée par un autre établissement de santé. Compte tenu de la crise des ressources humaines au Mali et au Sénégal, la disponibilité de personnel qualifié (sages-femmes et médecins) est problématique et de nombreuses tâches sont déléguées au personnel de santé moins qualifié (étudiants, matrones, sages-assistants). Ces professionnels peuvent jouer un rôle crucial dans l'amélioration des résultats maternels dans les hôpitaux de référence s'ils sont impliqués dans des tâches appropriées et reçoivent une formation adéquate. Plus précisément, nos résultats indiquent qu'ils devraient être formés pour détecter les ruptures utérines et les hémorragies. Les tâches et les actions requises sont assez précises et simples : poser des questions sur la douleur et les contractions, ainsi que des saignements vaginaux, mesurer la pression artérielle, jauge de protéines, détecter une perte de sang excessive et des convulsions. Même le personnel de santé non qualifié pourrait détecter, à l'admission ou pendant le travail (accouchement), les signes d'alarme suivants : douleur aiguë et perte de contractions, la pression artérielle> 140/90 mmHg, protéinurie> 1, l'hémorragie, et ils doivent alors alerter immédiatement le personnel qualifié si l'un de ces signes est détecté. La détection précoce de ces signes de complication, et la gestion immédiate des patientes par des sages-femmes ou les médecins permettrait d'améliorer les résultats maternels [3, 5, 13].
Les profils définis par le modèle de classement basé sur les règles d'association apportent des connaissances utiles aux professionnels des soins de santé dans les hôpitaux de référence au Mali et au Sénégal. Ils peuvent servir de référence dans leur décision de traiter les patients qui accouchent dans les établissements de santé.

7 Conclusion

Un effet important de l'apprentissage statistique établie dans cette étude pourrait être une identification rapide par les professionnels de la santé qualifiés ou non-qualifiés des mères à haut risque de mortalité à l'hôpital qui doivent être ensuite offert des soins obstétricaux d'urgence de haute priorité. Cette stratégie devrait viser toutes les femmes enceintes fréquentant les hôpitaux de référence au Sénégal et au Mali. Elle devrait aussi viser à détecter et à gérer les complications mortelles par des interventions fondées sur des preuves avec un suivi intensif des femmes qui ont un ante-partum césarienne d'urgence. Dans d'autres contextes, d'autres études sont nécessaires pour évaluer l'impact de cette stratégie sur la réduction des taux de mortalité maternelle, du temps d'accès au soins, de légalité globale et de la mortalité maternelle à l'hôpital. Cette stratégie offrira encore plus d'avantages si elle est combinée avec des interventions améliorant le système de référence maternelle.

Bibliographie

[1] Agrawal, R., and Srikant, R. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases* (San Francisco, CA, USA, 1994), VLDB '94, Morgan Kaufmann Publishers Inc., pp. 487–499. 143

[2] Bhatia, J. C. *A study of maternal mortality in Anantapur district, Andhra Pradesh, India.* Indian Insititute of Management, Bangalore, 1988. 138

[3] Campbell, O. M. R., Graham, W. J., and Lancet Maternal Survival Series steering group. Strategies for reducing maternal mortality : getting on with what works. *Lancet 368*, 9543 (2006), 1284–1299. PMID : 17027735. 138, 149

[4] Chen, L., Rohde, J., and Jolly, R. A looming crisis : health in the central asian republics. *The Lancet 339*, 8807 (1992), 1465–1467. 137

[5] Dogba, M., Fournier, P., Dumont, A., Zunzunegui, M.-V., Tourigny, C., and Berthe-Cisse, S. Mother and newborn survival according to point of entry and type of human resources in a maternal referral system in kayes (Mali). *Reproductive Health 8*, 1 (2011), 13. PMID : 21569276. 149

[6] Dumont, A., Fournier, P., Fraser, W., Haddad, S., Traore, M., Diop, I., Gueye, M., Gaye, A., Couturier, F., Pasquier, J.-C., Beaudoin, F., Lalonde, A., Hatem, M., and Abrahamowicz, M. QUARITE (quality of care, risk management and technology in obstetrics) : a cluster-randomized trial of a multifaceted intervention to improve emergency obstetric care in Senegal and Mali. *Trials 10* (2009), 85. PMID : 19765280. 138

[7] Dumont, A., Gaye, A., de Bernis, L., Chaillet, N., Landry, A., Delage, J., and Bouvier-Colle, M.-H. Facility-based maternal death reviews : effects on maternal mortality in a district hospital in Senegal. *Bulletin of the World Health Organization 84*, 3 (2006), 218–224. PMID : 16583081. 137

[8] Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett. 27*, 8 (2006), 861–874. 143

[9] Fournier, P., Dumont, A., Tourigny, C., Dunkley, G., and Dramé, S. Improved access to comprehensive emergency obstetric care and its effect on institutional maternal mortality in rural Mali. *Bulletin of the World Health Organization 87*, 1 (2009), 30–38. PMID : 19197402. 137

[10] Hahsler, M., Grün, B., and Hornik, K. A computational environment for mining association rules and frequent item sets. *JOURNAL OF STATISTICAL SOFTWARE* (2005), 2005. 143

[11] Igberase, G. O., and Ebeigbe, P. N. Maternal mortality in a rural referral hospital in the niger delta, Nigeria. *Journal of obstetrics and gynaecology : the journal of the Institute of Obstetrics and Gynaecology 27*, 3 (2007), 275–278. PMID : 17464810. 137

[12] Kampikaho, A., and Irwig, L. M. Risk factors for maternal mortality in five kampala hospitals, 1980-1986. *International journal of epidemiology 19*, 4 (1990), 1116–1118. PMID : 2083999. 137

[13] Mbola Mbassi, S., Mbu, R., and Bouvier-Colle, M. H. Delay in the management of obstetric complications : study in 7 maternity units in Cameroon. *Médecine tropicale : revue du Corps de santé colonial 69* (2009), 480–484. PMID : 20025179. 137, 149

[14] Murphy, M. Social consequences of vesico-vaginal fistula in northern Nigeria. *Journal of biosocial science 13*, 2 (1981), 139–150. PMID : 7287771. 137

[15] Prual, A. Pregnancy and delivery in western africa. high risk motherhood. *Santé publique (Vandoeuvre-lès-Nancy, France) 11*, 2 (1999), 155–165. 137

[16] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. 143

[17] Ronsmans, C., and Graham, W. J. Maternal mortality : who, when, where, and why. *The Lancet 368*, 9542 (2006), 1189–1200. 137

[18] Sullivan, S. A., Hill, E. G., Newman, R. B., and Menard, M. K. Maternal-fetal medicine specialist density is inversely associated with maternal mortality ratios. *American Journal of Obstetrics and Gynecology 193*, 3 (2005), 1083–1088. 138

[19] Sundari, T. K. The untold story : how the health care systems in developing countries contribute to maternal mortality. *International journal of health services : planning, administration, evaluation 22*, 3 (1992), 513–528. PMID : 1644513. 138

[20] Teguete, I., Traore, Y., Dennis, N., Mounkoro, N., Traore, M., and Dolo, A. A 19-year retrospective investigation of maternal mortality at point G national hospital, Bamako,

Mali. *International journal of gynaecology and obstetrics : the official organ of the International Federation of Gynaecology and Obstetrics 108*, 3 (2010), 194–198. PMID : 19944419. 137

[21] Thonneau, P. F., Matsudai, T., Alihonou, E., De Souza, J., Faye, O., Moreau, J.-C., Djanhan, Y., Welffens-Ekra, C., and Goyaux, N. Distribution of causes of maternal mortality during delivery and post-partum : results of an african multicentre hospital-based study. *European journal of obstetrics, gynecology, and reproductive biology 114*, 2 (2004), 150–154. PMID : 15140507. 137

Chapitre VI

Conclusion générale et perspectives

1 Conclusion générale

Dans cette analyse, nous avions comme objectifs : l'identification des profils caractéristiques des patientes décédées, la modélisation de la probabilité de décès en tenant compte de l'effet hôpital et la mis en place d'une règle de classement efficace permettant de trier les patientes à haut risque. Pour atteindre ces objectifs, nous avons choisi une approche basée sur les règles d'association dans le but de contourner les difficultés liées à la faible occurrence de la modalité d'intérêt de la variable réponse. Jusque là, les différentes méthodes statistiques proposées dans la littérature pour l'analyse de données déséquilibrées dans le cadre d'une classification supervisée produisent : soit un classifieur fortement dépendant de l'ensemble d'apprentissage, soit un classifieur efficace mais sous forme d'une boîte noire. Dans le domaine de l'intelligence artificielle, des algorithmes basés sur les règles d'association, tels que CBA (Classification Based on Association), CMAR (Classification Based on Multiples Association Rules) et CPAR (Classification Based on Prédictive Association Rules), ont été développés pour identifier les profils corrélés avec la modalité d'intérêt de la variable cible. Cependant ces algorithmes produisent un classifieur représenté par un vaste ensemble de profils dont la plus part d'entre eux ne sont pas pertinents. Dans certains domaines tels que la médecine, le classifieur produit est difficile à manipuler voire inutilisable. La procédure d'apprentissage statistique que nous avons présenté dans cette analyse permet de prendre en compte les avantages des méthodes d'analyse qui lui ont précédé. La procédure permet de construire un classifieur performant à partir d'un ensemble réduit et optimal de profils. En effet l'une des grandes difficultés avec les règles d'association reste la production d'un vaste ensemble de profils. Dans le chapitre III, nous avons proposé deux tests d'hypothèse : un test stochastique et un test asymptotique pour l'élagage des profils redondants. Ceci permet à la fois de supprimer une bonne partie des profils qui ne sont pas pertinents et de réduire considérablement l'ensemble des profils candidats pour constituer la règle de classement.
Pour sélectionner l'ensemble optimal de profils, nous avons proposé les algorithmes III.3 et III.4 selon la taille du jeux de données dont on dispose. Si la taille des données est suffisamment grande, on propose d'utiliser l'algorithme III.3 sur un ensemble de validation indépendant de l'ensemble d'apprentissage. Pour un jeux données de petite taille, on peut utiliser l'algorithme III.4 qui, à partir d'un

nombre fini d'échantillons bootstrap des données, sélectionne un ensemble optimal de profils. Ce qui nous permet de réduire la forte dépendance du classifieur de l'ensemble d'apprentissage.
L'indicateur de performance principal pour la sélection des profils candidats reste la valeur prédictive positive. Et pour tenir compte de l'effet hôpital dans la modélisation, nous avons effectué une estimation bayésienne empirique de la valeur prédictive positive pour partager l'information entre les hôpitaux. A ce niveau, nous avons proposé deux méthodes d'estimation des hyperparamètres : la méthode d'estimation des moments combinée avec un algorithme de pondération empirique et la méthode d'estimation du maximum de vraisemblance combinée avec un algorithme MM (Minimisation-Maximisation).
En combinant les profils de l'ensemble réduit et optimal de profils, on construit un classifeur performant et facile à interpréter par tout agent de santé maternelle. Il peut être affiché sous forme de tableau ou de poster dans les salles d'accouchement dans les hôpitaux en Afrique subsaharienne pour aider les sages femmes dans une prise de décision rapide.

2 Perspectives

1. **Introduire des covariables observables sur les clusters :** Dans des travaux à venir, nous allons étendre notre modèle en introduisant une matrice de covariables $\mathbf{M}$ de dimension $n \times q$, où $n = \sum_{h=1}^{m} n_h$ est le nombre d'observations et q le nombre de caractéristiques observables sur tous les clusters. Le nombre de covariables est supposé être strictement supérieur à un ($q > 1$).

$$\begin{cases} \left[Y|\theta_h^U, [Y,X]_h\right] = \prod_{k=1}^{m} \left(\theta_h^U\right)^{\mathbb{1}_{[Y=1]}(y)\delta_{\{1,[Y,X]_h\}}(U(X),[Y,X]_h)} \left(1-\theta_h^U\right)^{(1-\mathbb{1}_{[Y=1]}(y))\delta_{\{1,[Y,X]_h\}}(U(X),[Y,X]_k)} \\ \left[\theta_h^U|\pi_U,\tau_U\right] = \frac{\Gamma(\tau_U)}{\Gamma(\pi_U\tau_U)\Gamma((1-\pi_U)\tau_U)} \left(\theta_h^U\right)^{\pi_U\tau_U-1} \left(1-\theta_h^U\right)^{(1-\pi_U)\tau_U-1} \mathbb{1}_{[0,1]}\left(\theta_h^U\right) \\ logit\left(\pi_U\right) = \mathbf{W}^t\beta_U \end{cases}$$

   où $\mathbf{W} = (U(X), \mathbf{M}_i)_{n\times(q+1)}$ est un vecteur de dimension $(q+1)$, $\mathbf{M}_i$ désignant un vecteur ligne de la matrice $\mathbf{M}$ et β_U est le vecteur des coefficients de régression associés au profil $U(X)$.
   Le paramètre τ_U est le paramètre qui gouverne la corrélation entre les observations du même cluster h vérifiant le profil $U(X)$. On montre que pour deux observations i et j vérifiant $U(X)$ dans un cluster h, on a $corr(Y_{hi}, Y_{hj}) \geq \dfrac{1}{\tau_U + 1}$.

2. **Etude de la stabilité du classifieur :** Dans un futur proche, nous nous intéresserons à la stabilité du classifieur lorsque les données d'apprentissage subissent des modifications. Ceci semble être un point important pour la généralisation des résultats obtenus sur l'ensemble des centres de santé en Afrique Sub-Saharienne.

3. **Améliorer les performances du classifieur :** Dans ce travail, nous avons choisi de classer positive une observation t lorsqu'elle vérifie au moins un profil optimal $(U_i)_{i=1}^{M}$.

$$\phi(t) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^{M} \phi_{U_i}(t) > 0 \\ 0 & \text{sinon} \end{cases}$$

Les résultats obtenus montrent un taux de faux positifs très élevé. Ceci pourrait être justifié par le fait que la règle de classement $\phi()$ est une fonction des profils corrélés avec la classe rare.
Par exemple les résultats obtenus à partir des données du projet QUARITE révèlent un taux de faux positifs supérieur à 20% des patientes vivantes et représentant plus de 90% du taux d'erreur de classement (voir tableau V.6). La prise en charge de ce groupe de patientes peut entraîner des coûts très élevés qui risquent de contrarier le bon fonctionnement de la structure.

Résumé

L'objectif de cette thèse est de proposer une méthodologie statistique permettant de formuler une règle de classement capable de surmonter les difficultés qui se présentent dans le traitement des données lorsque la distribution a priori de la variable réponse est déséquilibrée. Notre proposition est construite autour d'un ensemble particulier de règles d'association appelées "class association rules".

Dans le chapitre II, nous avons exposé les bases théoriques qui sous-tendent la méthode. Nous avons utilisé les indicateurs de performance usuels existant dans la littérature pour évaluer un classifieur. A chaque "class association rule" est associée un classifieur faible engendré par l'antécédent de la règle que nous appelons profils. L'idée de la méthode est alors de combiner un nombre réduit de classifieurs faibles pour constituer une règle de classement performante.

Dans le chapitre III, nous avons développé les différentes étapes de la procédure d'apprentissage statistique lorsque les observations sont indépendantes et identiquement distribuées. On distingue trois grandes étapes : (1) une étape de génération d'un ensemble initial de profils, (2) une étape d'élagage de profils redondants et (3) une étape de sélection d'un ensemble optimal de profils. Pour la première étape, nous avons utilisé l'algorithme "apriori" reconnu comme l'un des algorithmes de base pour l'exploration des règles d'association. Pour la deuxième étape, nous avons proposé un test stochastique. Et pour la dernière étape un test asymptotique est effectué sur le rapport des valeurs prédictives positives des classifieurs lorsque les profils générateurs respectifs sont emboîtés. Il en résulte un ensemble réduit et optimal de profils dont la combinaison produit une règle de classement performante.

Dans le chapitre IV, nous avons proposé une extension de la méthode d'apprentissage statistique lorsque les observations ne sont pas identiquement distribuées. Il s'agit précisément d'adapter la procédure de sélection de l'ensemble optimal lorsque les données ne sont pas identiquement distribuées. L'idée générale consiste à faire une estimation bayésienne de toutes les valeurs prédictives positives des classifieurs faibles. Par la suite, à l'aide du facteur de Bayes, on effectue un test d'hypothèse sur le rapport des valeurs prédictives positives lorsque les profils sont emboîtés.

Dans le chapitre V, nous avons appliqué la méthodologie mise en place dans les chapitres précédents aux données du projet QUARITE concernant la mortalité maternelle au Sénégal et au Mali.

Mots clés : apprentissage statistique, classement, données déséquilibrées, estimation Bayésienne empirique, mortalité maternelle, profils, règles d'association, sélection de profils, test d'hypothèse

Abstract

The aim of this thesis is to design a supervised statistical learning methodology that can overcome the weakness of standard methods when the prior distribution of the response variable is unbalanced. The proposed methodology is built using class association rules. Chapter II deals with theorical basis of statistical learning method by relating various classifiers performance metrics with class association rules. Since the classifier corresponding to a class association rules is a weak classifer, we propose to select a small number of such weak classifiers and to combine them in the aim to build an efficient classifier.

In Chapter III, we develop the different steps of the statistical learning method when observations are independent and identically distributed. There are three main steps : In the first step, an initial set of patterns correlated with the target class is generated using "apriori" algorithm. In the second step, we propose a hypothesis test to prune redondant patterns. In the third step, an hypothesis test is performed based on the ratio of the positive predictive values of the classifiers when respective generating patterns are nested. This results in a reduced and optimal set of patterns whose combination provides an efficient classifier.

In Chapter IV, we extend the classification method that we proposed in Chapter III in order to handle the case where observations are not identically distributed. The aim being here to adapt the procedure for selecting the optimal set of patterns when data are grouped data. In this setting we compute the estimation of the positive predictive values as the mean of the posterior distribution of the target class probability by using empirical Bayes method. Thereafter, using Bayes factor, a hypothesis test based on the ratio of the positive predictive values is carried out when patterns are nested.

Chapter V is devoted to the application of the proposed methodology to process a real world dataset. We studied the QUARITE project dataset on maternal mortality in Senegal and Mali in order to provide a decision making tree that health care professionals can refer to when managing patients delivering in their health facilities.

Keywords : association rules, classification, empirical Bayesian estimation, hypothesis testing, maternal mortality, patterns, selection profiles, statistical learning, unbalanced data