Classification supervisée par modèle de mélange: Application aux diagnostics par autopsie verbale

Seydou N. SYLLA [1,2,3], Stéphane GIRARD [1], Abdou Ka DIONGUE [2]
Aldiouma DIALLO [3] & Cheikh SOKHNA [3]

[1] *Inria Grenoble Rhone-Alpes & LJK, France, stephane.girard@inria.fr*
[2] *LERSTAD-UGB, Saint-Louis, Sénégal, abdou.diongue@edu.ugb.sn*
[3] *URMITE-IRD, Dakar, Sénégal, seydou-nourou.sylla@ird.fr, cheikh.sokhna@ird.fr, aldiouma.diallo@ird.fr*

Résumé. La surveillance et les évaluations dans le domaine sanitaire font de plus en plus appel aux données relatives aux causes de décès provenant des autopsies verbales dans les pays ne tenant pas de registres d'état civil ou disposant de registres incomplets. L'application de la méthode d'autopsie verbale permet de disposer des causes probables de décès. L'autopsie verbale est devenue la principale source d'information sur les causes de décès dans ces populations. Cette communication présente un modèle de mélange multinomial hiérarchique sous l'hypothèse d'indépendance conditionnelle appliqué à des données de diagnostics par autopsie verbale dans les zones de Niakhar, Bandafassi et Mlomp (Sénégal).

Mots-clés. Modèle de mélange, classification, aide au diagnostic.

Abstract. Health monitoring and evaluation make more and more use of data on causes of death from verbal autopsies in countries which do not keep records of civil status or with incomplete records. The application of verbal autopsy method allows to discover probable cause of death. Verbal autopsy has become the main source of information on causes of death in these populations. This talk will presents a hierarchical multinomial mixture model applied to diagnostic data by verbal autopsy in areas of Niakhar, Bandafassi and Mlomp (Sénégal).

Keywords. Mixture model, classification, aided diagnostic.

1 Contexte de l'étude

La mortalité reste trés élevée dans nombre de pays africains au sud du Sahara, et particuliérement en milieu rural. Une meilleure connaissance des causes de décès permettrait, d'une part l'évaluation de l'impact des programmes dirigés vers la réduction de la mortalité, et d'autre part l'allocation de ressources dans ces domaines. L'application d'une méthode dite d'autopsie verbale permet de disposer des causes probables de décès. L'autopsie verbale est devenue la principale source d'information sur les causes de décès dans les populations pour lesquelles il n'existe ni système d'état civil ni certificat médical de décès [1].

2 Modèle de mélange hiérarchique sous indépendance conditionnelle

Données et notations. Les données dont on dispose consistent en la présence de plusieurs symptômes et la déclaration de plusieurs causes probables de décès mesurées sur des personnes décédées durant la période de 1985 à 2010 dans les trois sites de l'IRD (Niakhar, Bandafassi et Mlomp) au Sénégal. Ces variables binaires représentent la présence (1) ou l'absence (0) des symptômes et des variables non symptomatiques sur l'individu i donné. Pour chaque symptôme donné, un ensemble de sous items est collecté. Les variables aléatoires binaires $X = (X_j, j = 1, \ldots, p)$ définissent les symptômes et variables socio-démographiques. Les variables aléatoires $Z = (Z_j^\ell, j = 1, \ldots, p, \ell = 1, \ldots, p_j)$ représentent les p_j sous-items pour chaque variable X_j. La variable aléatoire Y est la variable à expliquer représentant le groupe (diagnostics des médecins). On dispose d'un échantillon de n individus décrits par les variables explicatives X et Z ainsi que leur appartenance à l'un des K groupes (variable Y). Nous proposons deux modèles:

- Un modèle de mélange sous l'hypothèse d'indépendance conditionnelle prenant en compte seulement les variables aléatoires X représentant les données symptomatiques et non symptomatiques (age, saison, sexe,...)
- Un modèle de mélange hiérarchique permettant de prendre en compte les variables aléatoires Z représentant les sous-items de chaque variable X.

Dans ces modèles, les classes considérées sont les diagnostics établis par les médecins à la lecture de la fiche d'enquête (autopsie verbale). Le jeu de données considéré ici comporte $n = 2500$ individus répartis dans $K = 22$ classes et caractérisés par plus de $\sum_{j=1}^{p}(1 + q_j) = 30$ variables.

Modèle de mélange sous hypothèse d'indépendance conditionnelle. Dans cette partie de l'étude, on ne considére que l'échantillon issu des variables X et Y. Les variables explicatives sont supposées indépendantes à l'intérieur de chaque groupe $X_i \perp X_j | Y$ pour tout $i \neq j$:

$$P(X = x|Y = k) = \prod_{j=1}^{p} P(X_j = x_j|Y = k).$$

Le théorème des probabilités totales permet d'obtenir la loi marginale de X :

$$P(X = x) = \sum_{k=1}^{K} P(Y = k) \prod_{j=1}^{p} P(X_j = x_j|Y = k),$$

et selon le théorème de Bayes, les probabilités a posteriori d'appartenance aux classes s'écrivent

$$P(Y = k|X = x) \propto P(Y = k) \prod_{j=1}^{p} P(X_j = x_j|Y = k).$$

La variable Y est modélisée par une loi multinomiale à k niveaux de probabilités $p_1, \ldots, p_K$, c'est à dire $P(Y = k) = p_k$ pour $k = 1, \ldots, K$. Conditionnellement à $Y = k$, X_j est modélisée

par une loi de Bernoulli de paramètre $\theta_{j,k}$, pour tout $j = 1, \ldots, p$ et $k = 1, \ldots, K$. L'estimation des paramètres du modèle s'effectue par la méthode du maximum de vraisemblance. Ainsi, $p_{j,k}$ est estimé par la proportion d'individus présentant le symptôme j parmi les individus qui ont été diagnostiqués porteurs de la maladie k. L'estimation des probabilités a priori p_k a été réalisée dans deux cadres différents : hypothèse d'égalité des probabilités et probabilités libres. Dans le premier cas, p_k est estimé par $1/K$ et dans le second cas par n_k/n ou n_k est le nombre d'individus diagnostiqués pour la cause k et n est le le nombre total d'individus $n = \sum_{k=1}^{K} n_k$. L'affection d'un individu à l'une des classes est faite par la règle du maximum a posteriori : x est affecté au groupe ℓ si et seulement si

$$\ell = \arg \max_{k=1,\ldots,K} p_k \prod_{j=1}^{p} \theta_{j,k}^{x_j} (1 - \theta_{j,k})^{1-x_j}.$$

Le taux de classification correcte est alors la proportion d'accords entre les diagnostics des médecins et les résultats du modèle.

Modèle de mélange hiérarchique. Dans cette partie de l'étude, on considére l'échantillon complet des variables X, Y et Z. Les probabilités a posteriori d'appartenance aux classes font appel au calcul de $P(Y = y|X = x, Z = z)$ qui n'est pas détaillé ici.

Sélection de variables. Nous avons utilisé une procédure de sélection de variables sous l'hypothèse d'indépendance conditionnelle. Les variables sont ordonnées selon leur performance de classification lorqu'elles sont utilisées seules. On parcourt ensuite cet ensemble ordonné et l'on retient le nombre de variables qui donne le meilleur taux d'erreur par validation croisée. Les résultats sont représentés Figure 1 dans le cas du modèle de mélange non-hiérarchique. La sélection de 15 variables permet d'atteindre un taux de classification de l'ordre de 50%.

Réduction du nombre de classes. Afin de réduire la complexité du probléme, nous avons étudié dans quelle mesure il était possible de réduire le nombre de classes. Pour ceci, chaque groupe k est caractérisé par le vecteur des probabilités $\Pi_k \in \mathbb{R}^n$ des probabilités a posteriori d'appartenance des individus. Les classes sont alors regroupées entre elles par l'algorithme des k-medoids [2]. L'évolution du pourcentage d'individus bien classés en fonction du nombre de classes retenues est présenté Figure 2 pour le modèle non-hiérarchique. Il apparait tout d'abord qu'il est préférable de ne pas se restreindre à des probabilités a priori de classes égales. Sans surprise, plus le nombre de classes considéré est faible plus le taux de bien classés est important. Cependant, on remarque une stabilisation des courbes pour 6, 7 et 8 groupes avec un taux de bien classés de l'ordre de 60%. Les partitions correspondantes en 6, 7 et 8 groupes des causes de décès ont été soumises à des experts du domaine pour validation.
La classification utilisant le modèle hiérarchique est en cours d'implémentation, les résultats devraient être disponibles prochainement.

References

[1] J. P. Chippaux. (2009) Conception, utilisation et exploitation des autopsies verbales. *Médecine Tropicale*, **69**, 143–150.

[2] T. Velmurugan and T. Santhanam. (2010) Computational complexity between k-means and k-medoids clustering algorithm for normal and uniform distributions of data points. *Journal of Computer Science*, **6**, 363–368.

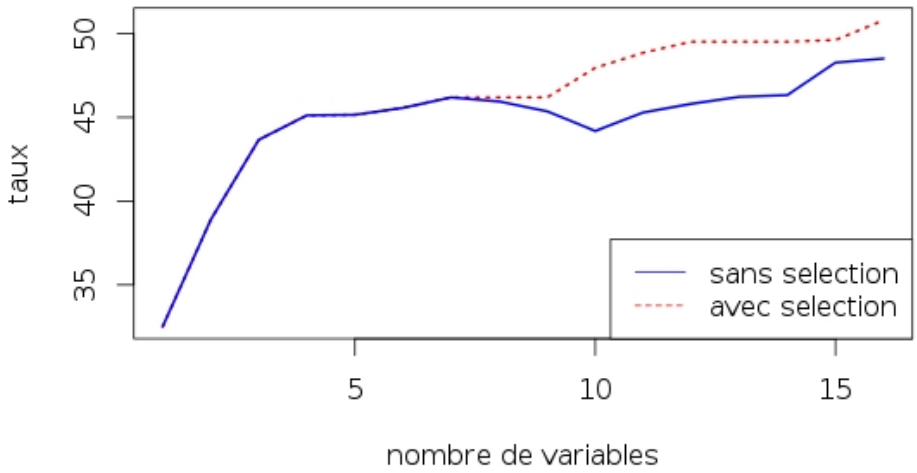

Figure 1: Taux de bien classés en fonction du nombre de variables. La courbe bleue présente les résultats sans sélection, *i.e.* toutes les variables sont entrées séquentiellement dans le modèle. La courbe rouge présente les résultats avec sélection, *i.e.* une variable est ajoutée au modèle seulement si elle entraine une augmentation du taux de bien classés.

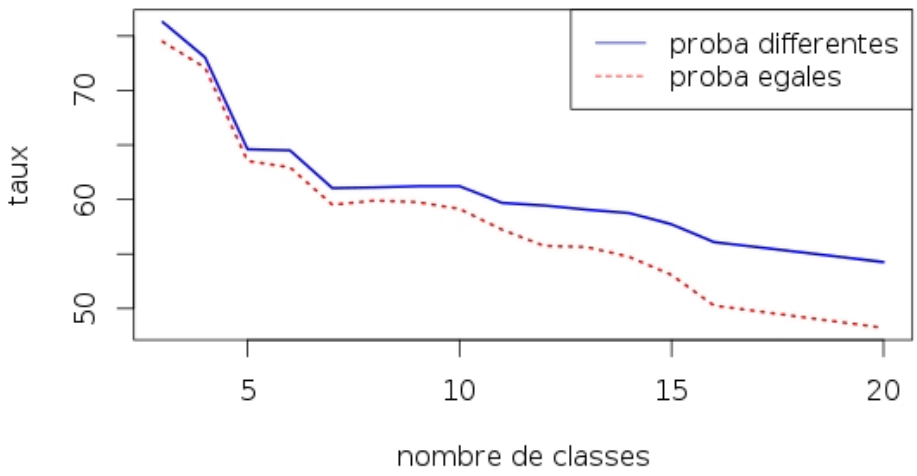

Figure 2: Taux de bien classés en fonction du nombre de classes selon que les probabilités a priori des classes sont supposées égales (courbe rouge) ou différentes (courbe bleue).