Classification de données binaires via l'introduction de mesures de similarités dans les modèles de mélange

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

[1] *Inria Grenoble Rhône-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. 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. La méthode d'autopsie verbale permet de disposer des causes probables de décès. Cette communication présente une méthode de classification sur des données binaires de diagnostics par autopsie verbale dans les zones de Niakhar, Bandafassi et Mlomp (Sénégal). Cette méthode combine l'utilisation de mesures de similarités avec une méthode de classification récente basée sur l'introduction d'un noyau dans le modèle de mélange gaussien.

Mots-clés. Modèle de mélange, données binaires, méthode à noyau, mesure de similarité, aide au diagnostic.

Abstract. Evaluations in the health field are more and more based on data on causes of death from verbal autopsies in countries which do not keep civil registers or with incomplete registers. The application of the verbal autopsy method enables to estimate the probable causes of death. This communication presents a classification method for binary data issued from verbal autopsy diagnosis in the areas of Niakhar, Bandafassi and Mlomp (Senegal). This method combines similarity measures with a recent classification method based on the introduction of a kernel in the Gaussian mixture model.

Keywords. Mixture model, binary data, kernel method, similarity measure, aided diagnostic.

1 Contexte de l'étude

Le manque de données fiables sur les niveaux et les causes de mortalité dans les régions défavorisées de la planète constitue un frein aux efforts déployés pour établir une base de données solide sur laquelle s'appuie la politique, la planification, la surveillance et l'évaluation sanitaire. La mortalité reste très élevée dans nombre de pays africains au sud du Sahara, et particulièrement 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]. L'élaboration d'un tel système de diagnostic se réalise généralement en trois phases [1] : la phase d'enquête qui comprend l'interrogatoire du patient (ou de ses proches) et le prétraitement des données, la phase de choix d'une méthode de décision (discrimination entre classes) et la phase d'exploitation où l'objectif est d'associer aux nouveaux indidividus leurs groupes (maladie, causes de décès)

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 un individu donné. Les variables aléatoires binaires $X = (X_j, j = 1, \ldots, p)$ définissent les symptômes et variables socio-démographiques. 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 ainsi que leur appartenance à l'un des L groupes (variable Y). Le jeu de données considéré ici comporte $n = 2500$ individus répartis dans $L = 22$ classes et caractérisés par $p = 30$ variables.

2 Classification grâce à une fonction noyau

Considérons un échantillon d'apprentissage $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ constitué de réalisations indépendantes d'un vecteur aléatoire X binaire et où les étiquettes $\{y_1, \ldots, y_n\}$ sont des réalisations indépendantes d'une variable aléatoire $Y \in \{1, \ldots, L\}$ indiquant l'appartenance des observations aux L classes, ie $y_i = k$ signifie que x_i appartient à la $k$ème classe C_k.

Soit K une fonction noyau définie par $K : \{0,1\}^2 \to \mathbb{R}^+$ satisfaisant les conditions de Mercer. Pour tout $k = 1, \ldots, L$, on introduit également la fonction $\rho_k : \{0,1\}^2 \to \mathbb{R}$ définie par

$$\rho_k(x, y) = K(x, y) - \frac{1}{n_k} \sum_{x_\ell \in C_k} (K(x_\ell, y) + K(x, x_\ell)) + \frac{1}{n_k^2} \sum_{x_\ell, x_{\ell'} \in C_k} K(x_\ell, x_{\ell'}),$$

où n_k est le cardinal de la classe C_k. On définit alors la matrice M_k par $(M_k)_{\ell,\ell'} = \rho_k(x_\ell, x_{\ell'})/n_k$ pour tout $(\ell, \ell') \in \{1, \ldots, n_k\}^2$. On définit la règle de classification suivante : $x \to C_i$ si et seulement si $i = \arg\min_{k=1,\ldots,L} D_k(x)$ où D_k est la fonction de classification proposée par [2] :

$$D_k(x) = \frac{1}{n_k} \sum_{j=1}^{d_k} \frac{1}{\lambda_{kj}} \left(\frac{1}{\lambda_{kj}} - \frac{1}{\lambda} \right) \left(\sum_{x_\ell \in C_k} \beta_{kj\ell} \rho_k(x, x_\ell) \right)^2 + \frac{1}{\lambda} \rho_k(x, x)$$
$$+ \sum_{j=1}^{d_k} \log(\lambda_{kj}) + (d_{\max} - d_k) \log(\lambda) - 2 \log(\pi_k). \quad (1)$$

On a noté $\pi_k = n_k/n$, λ_{kj} la $j$ème plus grande valeur propre de la matrice M_k , β_{kj} le vecteur propre associé ($\beta_{kj\ell}$ représente sa ℓème coordonnée), $j = 1, \ldots, d_k$ et $d_{max} = \max(d_1, \ldots, d_L)$. Enfin, on a

$$\lambda = \sum_{k=1}^{L} \pi_k (\text{trace}(M_k) - \sum_{j=1}^{d_k} \lambda_{kj}) \Big/ \sum_{k=1}^{L} \pi_k (r_k - d_k)$$

et les paramètres d_k et r_k représentent respectivement la dimension intrinsèque de la classe C_k (à estimer) et une caractéristique (connue, cf [2], Table 2) du noyau K calculé sur C_k, $k = 1, \ldots, L$. Nous renvoyons le lecteur à [2] pour plus de détails sur les fondements statistiques de cette règle de classification. Formellement, elle s'apparente à la règle de classification quadratique obtenue à partir d'un modèle de mélange gaussien dans laquelle les produits scalaires sont remplacés par une fonction de similarité non-linéaire. La mise en œuvre de cette méthode nécessite le choix d'une fonction noyau K mesurant la similarité entre deux vecteurs binaires. Nous présentons dans le paragraphe suivant une brève synthèse des mesures existantes dans la littérature.

3 Mesures de similarité et dissimilarité

Soient x_ℓ et $x_{\ell'}$ deux individus dans $\{0,1\}^p$ que l'on souhaite comparer. De nombreuses mesures de similarité (ou dissimilarité) ont été élaborées depuis plus de cent ans et utilisées dans divers domaines. L'article de synthèse [3] liste 76 exemples de telles mesures. On note $a =< x_\ell, x_{\ell'} >$ et $d =< \mathbf{1} - x_\ell, \mathbf{1} - x_{\ell'} >$ respectivement le nombre de 1 et 0 communs entre les deux vecteurs ($\mathbf{1}$ désigne le vecteur de $\mathbb{R}^p$ dont toutes les composantes sont égales à 1). De même, on introduit $b =< \mathbf{1} - x_\ell, x_{\ell'} >$ et $c =< x_\ell, \mathbf{1} - x_{\ell'} >$ avec $a + b + c + d = p$. Nous proposons de synthétiser une partie des mesures de [3] en introduisant la mesure de similarité suivante :

$$S(x_\ell, x_{\ell'}) = \frac{\alpha a - \theta(b+c) + \beta d}{\alpha' a + \theta'(b+c) + \beta' d} \tag{2}$$

où α, α', β, β', θ et θ' sont des coefficients réels quelconques. Dans le paragraphe suivant, nous considérerons également le cas particulier

$$S_{\text{Sylla \& Girard}}(x_\ell, x_{\ell'}) = \alpha a + (1-\alpha) d.$$

La Table 1 présente 28 mesures de similarité issues de [3] réécrites dans le formalisme (2).Certaines mesures proposées dans des contextes différents se révèlent équivalentes pour des données binaires. En outre, les mesures "Sokal & Michener" et "Innerproduct" sont des cas particuliers de la mesure $S_{\text{Sylla \& Girard}}$ obtenus avec $\alpha = 1/2$. De même, les mesures "Intersection" et "Russell & Rao" sont des cas particuliers de la mesure $S_{\text{Sylla \& Girard}}$ obtenus avec $\alpha = 1$.

4 Construction de noyaux associés à des observations binaires

Noyaux linéaires. Le noyau linéaire est la plus simple des fonctions noyaux. Elle est donnée par $K_{\text{linéaire}}(x_\ell, x_{\ell'}) =< x_\ell, x_{\ell'} >= a$, ce qui, dans le cas binaire, revient à comptabiliser le nombre de 1 communs entre x_ℓ et $x_{\ell'}$. On peut montrer ([2], Proposition 3) que la règle de classification associée (1) est linéaire. Remarquons que le noyau $K_{\text{linéaire}}(x_\ell, x_{\ell'}) =< \mathbf{1} - x_\ell, \mathbf{1} - x_{\ell'} >= d$ comptabilisant le nombre de 0 conduit à la même règle de classification. Ainsi, pour des données binaires, la règle de classification linéaire est indépendante du codage des observations.

Noyaux exponentiels. Le noyau de type exponentiel le plus connu est le noyau RBF :

$$K_{\text{RBF}}(x_\ell, x_{\ell'}) = \exp\left(-\frac{\|x_\ell - x_{\ell'}\|^2}{2\sigma^2}\right),$$

où σ est un paramètre positif. Pour des observations binaires, on a les simplifications suivantes :

$$K_{\text{RBF}}(x_\ell, x_{\ell'}) = \exp\left(-\frac{\|x_\ell - x_{\ell'}\|}{2\sigma^2}\right) = \exp\left(-\frac{(b+c)}{2\sigma^2}\right) = \exp\left(\frac{S_{\text{Hamming}}(x_\ell, x_{\ell'})}{2\sigma^2}\right).$$

Il apparaît ainsi que le noyau RBF peut être retrouvé en utilisant le mesure de similarité de Hamming (voir Table 1). Nous proposons d'étendre cette construction à toutes les mesures de similarité S de la Table 1 et plus généralement aux 76 mesures recensées dans [3] en posant :

$$K(x_\ell, x_{\ell'}) = \exp\left(\frac{S(x_\ell, x_{\ell'})}{2\sigma^2}\right).$$

Clairement, deux mesures de similarité ne différant que par des constantes multiplicatives ou additives donneront des règles de classification équivalentes.

5 Résultats

Les performances de la méthode de classification associée aux différents noyaux sont évaluées par double validation croisée. L'échantillon de taille $n = 2500$ est découpé aléatoirement en un ensemble d'apprentissage (comprenant approximativement 80% des individus) et un ensemble de test (comprenant approximativement 20% des individus). Les paramètres α, σ du noyau et d_i sont choisis par validation croisée sur le jeu d'apprentissage (la dimension d_i est déterminée par l'intermédiaire d'un `seuil` sur la variance cumulée, voir [2] pour plus de détails) : 5 fois consécutivement, 100 individus sont retirés aléatoirement de l'échantillon d'apprentissage, et les paramètres (α, σ, `seuil`) sont estimés par maximisation du taux de bien classés sur les 100 individus retirés. Le taux de bien classés global est estimé sur l'échantillon test en répétant l'ensemble du procédé 50 fois. La Table 2 résume les paramètres ainsi estimés et les pourcentages de classification corrects associés aux 12 noyaux ayant donné les meilleurs résultats. Il apparaît que ces 12 noyaux donnent de très bons résultats. A titre de comparaison, une classification par modèle de mélange multinomial sous hypothèse d'indépendance conditionnelle donne un taux de classification correcte de l'ordre de 50% seulement [4].

References

[1] J. P. Chippaux. Conception, utilisation et exploitation des autopsies verbales. *Médecine Tropicale*, **69**, 143–150, 2010.

[2] C. Bouveyron, M. Fauvel & S. Girard. Kernel discriminant analysis and clustering with parsimonious Gaussian process models. *Statistics and Computing*, à paraître, 2015.

[3] C. Seung-Seok, C. Sung-Hyuk & C. Tappert. A survey of binary similarity and distance measures *Systemics, Cybernetics and Informatics*, **8**, 43–48, 2010.

[4] S. Sylla, S. Girard, A. Diongue, A. Diallo & C. Sokhna. Classification supervisée par modèle de mélange: Application aux diagnostics par autopsie verbale, *46èmes Journées de Statistique organisées par la Société Française de Statistique*, Rennes, 2014.

Nom	α	θ	β	α'	θ'	β'	**équation**
Jaccard	1	0	0	1	1	0	(1)
Tanimoto	-	-	-	-	-	-	(65)
Dice	2	0	0	2	1	0	(2)
Czekanowski	-	-	-	-	-	-	(3)
Nei & li	-	-	-	-	-	-	(5)
3w-Jaccard	3	0	0	3	1	0	(4)
Sokal & Sneath-I	1	0	0	1	2	0	(6)
Sylla & Girard	α	0	$1-\alpha$	1	1	1	
Sokal & Michener	1	0	1	1	1	1	(7)
Innerproduct	-	-	-	-	-	-	(13)
Sokal & Sneath-II	2	0	2	2	1	2	(8)
Gower & Legendre	-	-	-	-	-	-	(11)
Roger & Tanimoto	1	0	1	1	2	1	(9)
Faith	1	0	0.5	1	1	1	(10)
Intersection	1	0	0	1	1	1	(12)
Russell & Rao	-	-	-	-	-	-	(14)
Hamming*	0	1	0	1	1	1	(15)
Squared-Euclid*	-	-	-	-	-	-	(17)
Canberra*	-	-	-	-	-	-	(18)
Manhattan*	-	-	-	-	-	-	(19)
Mean-Manhattan*	-	-	-	-	-	-	(20)
Cityblock*	-	-	-	-	-	-	(21)
Minkowski*	-	-	-	-	-	-	(22)
Vari*	-	-	-	-	-	-	(23)
Lance & Williams*	0	1	0	2	1	0	(27)
Bray & Curtis*	-	-	-	-	-	-	(28)
Sokal & Sneath-III	-1	0	-1	0	1	0	(56)
Kulczynski-I	-1	0	0	0	1	0	(64)
Hamann	1	1	1	1	1	1	(67)

Table 1: Mesures de similarité. Les mesures marquées (*) sont obtenues en prenant l'opposé de la mesure de dissimilarité associée. La dernière colonne fait référence au numéro de l'équation dans [3].

Noyau	α	σ	`seuil`	**Taux bien classés (apprentissage)**	**Taux bien classés (test)**	**équation**
Euclid		4	0.60	83.82	87.99	(16)
Pearson		10	0.95	83.25	87.72	(51)
Hellinger		6	0.60	83.21	87.68	(29,30)
Dice		2	0.60	83.00	87.32	(2,3,5)
3w-Jaccard		2	0.75	82.87	87.21	(4)
Ochia1		2	0.60	82.77	87.15	(33,38)
Gower & Legendre		4	0.80	82.64	86.61	(8,11)
Roger & Tanimoto		2	0.65	82.39	85.89	(9)
Sylla & Girard	0.1	1.9	0.90	81.50	85.81	
Sylla & Girard	0.3	2.2	0.85	81.46	85.47	
Sylla & Girard	0.5	1.4	0.80	81.35	85.05	(15,17,...,23)
Sylla & Girard	0.2	1.8	0.80	81.11	85.58	

Table 2: Classement des noyaux taux de classification correcte décroissants (en %). La dernière colonne fait référence au numéro de l'équation dans [3].