Appariement de descripteurs évoluant en temps

Anne-Lise Bedenel, Christophe Biernacki, Laetitia Jourdan

Appariement de descripteurs evoluant en temps Application a la comparaison d'assurance en ligne

Anne-Lise Bedenel [1,2,3], Christophe Biernacki [2] et Laetitia Jourdan [3]

[1] *PIXEO* [1], *France – anne-lise.bedenel@mercihenri.com*
[2] *Université Lille 1, Inria, France – christophe.biernacki@math.univ-lille1.fr*
[3] *Université Lille 1 CRIStAL, UMR 9189, Inria, France – laetitia.jourdan@inria.fr*

Résumé. Dans le domaine du web, et plus particulièrement de la comparaison d'assurance en ligne, les données évoluent constamment ce qui rend leur exploitation difficile. Par exemple, la plupart des méthodes d'apprentissage standards nécessitent des descripteurs de données identiques pour les échantillons d'apprentissage et test. Cependant, afin de répondre aux attentes métier, les formulaires en ligne d'où proviennent les données sont régulièrement modifiés. Cela implique une modification régulière des variables et des descripteurs de données qui complexifie les analyses. Dans ce travail, nous proposons une méthode permettant d'estimer et de comprendre les liens qui se forment lors de la modification des descripteurs de données afin de les apparier. Cette étape est préliminaire à l'application de nombreuses méthodes d'apprentissage ultérieures.

Mots-clés. Réseaux bayèsiens dynamiques, choix de modèles, agrégation de modèles.

Abstract. In the web domain, and in particular for insurance comparison, data constantly evolve, implying that it is difficult to directly exploit them. For example, to do a classification, performing standard learning processes require data descriptor equal for both learning and test samples. Indeed, for answering to web surfer expectation, online forms whence data come from are regularly modified. So, features and data descriptors are also regularly modified. In this work, we introduce a process to estimate and understand connections between transformed data descriptors. This estimated matching between descriptors will be a preliminary step before applying later classical learning methods.

Keywords. Bayesian dynamic network, model selection, model averaging.

1 Introduction

L'objectif d'un comparateur d'assurance est de proposer à ses internautes l'offre la plus adaptée à leurs attentes, selon leurs profils. Pour la plupart des comparateurs d'assurance en ligne, la comparaison se fait sur un seul critère : le prix. Afin d'affiner la comparaison

1. http://groupe-pixeo.com/

des internautes, nous souhaitons créer un modèle permettant de prédire l'offre la plus en adéquation avec les attentes de l'internaute, indépendamment du prix. Cet objectif en statistique est plutôt classique mais le fonctionnement d'un comparateur d'assurance en ligne a des contraintes qui nous empêchent d'appliquer directement les méthodes standards. En effet, pour faire une comparaison en ligne, un internaute doit remplir un formulaire de questions. Ce formulaire reprend les questions des systèmes de tarifications des assureurs partenaires du comparateur. Ainsi, à l'aide d'un web service, les assureurs partenaires peuvent renvoyer à l'internaute le prix réel de l'offre selon le profil qu'il a renseigné et de ses attentes. Or, la particularité d'un comparateur d'assurance vient du fait que les formulaires proposés aux internautes changent constamment, car le comparateur d'assurance adapte son formulaire en permanence :

— *Pour les assureurs* : chaque assureur ayant son propre système de tarification, les questions ne sont pas homogènes entre tous les assureurs partenaires.
— *Pour les internautes* : les questions sont adaptées régulièrement afin de viser plus de clarté et/ou de simplicité.

Ces changements peuvent être des ajouts, des suppressions, des fusions ou des modifications de questions, donc de descripteurs au sens statistique.

Créer par exemple un modéle de classification supervisée avec ces variables devient alors complexe. En effet, la plupart des méthodes statistiques standards requièrent des échantillons conséquents pour l'apprentissage et doivent aussi s'appuyer sur des descripteurs identiques entre les échantillons d'apprentissage et de test. Ces conditions ne sont alors pas remplies avec les données provenant du comparateur d'assurance.

Dans ce travail, nous nous intéressons à la modification des descripteurs d'une variable. Nous proposons une méthode pour estimer et comprendre les liens qui se forment (appariement) lors de la modification des descripteurs d'une variable à l'aide d'un cas d'usage (section 2). Puis nous présentons les résultats obtenus sur des données simulées et réelles (section 3).

2 Modélisation du problème

2.1 Formalisation

Pour introduire une modélisation générale, nous sommes partis d'un cas d'usage, où la question posée est la suivante : Comment les internautes réagissent-ils lorsque les descripteurs d'une variable changent ?

Pour répondre à cette question, nous nous focalisons sur la variable « Niveau de garantie souhaité » que les internautes doivent renseigner. Cette variable avait initialement quatre choix de réponses possibles (ou descripteurs) qui étaient {Tiers (T), Tiers++ (T++), Intermédiaire (Inter), Tous Risques (TR)}. Nous noterons $X \in \{1, \ldots, I\}$ cette variable, I désignant le nombre de descripteurs ($I = 4$ dans notre exemple). Dans un

second temps, suite à une redéfinition du site web, cette variable a été décomposée en sept nouveaux descripteurs qui étaient {Tiers (T), Tiers+ (T+), Tiers++ (T++), Intermédiaire (Inter), Tous Risques (TR), Tous Risques+ (TR+), Tous Risques++ (TR++)}. Nous noterons symétriquement $Y \in \{1, \ldots, J\}$ cette nouvelle variable, J désignant le nouveau nombre de descripteurs ($J = 7$ dans notre exemple).

Contrairement à la plupart des sites e-commerce, lorsqu'un internaute vient comparer un produit d'assurance, il est rare qu'il revienne sur le site. Cela amène comme spécificité que les observations disponibles sur X et Y ne sont jamais appariées. Plus précisèment cette propriété peut s'écrire de la façon suivante.

Période avant la redéfinition du site web N^- internautes ont renseigné la variable X, produisant des réalisations *observées* $\mathbf{X}^- = (X_1^-, \ldots, X_{N^-}^-)$. Comme on vient de le préciser, la variable Y n'a jamais été renseignée par contre, produisant des réalisations *non observées* $\mathbf{Y}^- = (Y_1^-, \ldots, Y_{N^-}^-)$.

Période après la redéfinition du site web De façon symétrique, N^+ internautes ont renseigné la variable Y, produisant des réalisations *observées* $\mathbf{Y}^+ = (Y_1^+, \ldots, X_{N^+}^+)$. Comme attendu, la variable X n'a jamais été renseignée par contre, produisant des réalisations *non observées* $\mathbf{X}^+ = (X_1^+, \ldots, X_{N^+}^+)$.

2.2 Modélisation probabiliste et estimation

Nous supposons maintenant que chaque couple (X_n^*, Y_n^*) est une réalisation i.i.d du couple (X, Y), avec $n = 1, \ldots, N^*$ et $* \in \{-, +\}$. La loi du couple (X, Y) peut s'écrire trivialement de la façon suivante :

$$P(X = i, Y = j) = p_{ij} p_i \tag{1}$$

où $p_{ij} = P(Y = j | X = i)$ et $p_i = P(X = i)$, avec $i = 1, \ldots, I$ et $j = 1, \ldots, J$. Cependant cette écriture a comme intérêt de faire apparaître la probabilité p_{ij} de transition (ou d'appariement) entre les descripteurs de X et Y, ce qui est l'objectif principal de ce premier travail. Notre problématique est donc maintenant d'estimer l'ensemble de toutes les transitions $\mathbf{p}_{\cdot\cdot} = (p_{ij})$. Nous noterons aussi $\mathbf{p}_{\cdot} = (p_i)$, qui lui aussi est un paramêtre inconnu.

L'ensemble des transitions $\mathbf{p}_{\cdot\cdot}$ est représenté dans le cas de notre exemple par des arcs orientés sur la figure 1. Nous remarquons alors que les paramètres $\mathbf{p}_{\cdot\cdot}$ à estimer sont beaucoup plus nombreux que le nombre de paramètres de la loi de Y, ce qui conduit à un modèle surparamétré donc non identifiable. Plus précisément, toujours pour notre exemple, il y a 28 probabilités d'appariements (soit 24 paramètres libres dans $\mathbf{p}_{\cdot\cdot}$) alors que la loi de Y a seulement 6 paramètres libres ($J - 1$). Le modèle n'est alors pas identifiable. Afin d'avoir un modèle identifiable, il est donc nécessaire d'imposer des contraintes sur $\mathbf{p}_{\cdot\cdot}$ qui restreignent le nombre de paramètres libres à 6 ou moins ($\dim(\mathbf{p}_{\cdot\cdot}) \leq J - 1$).

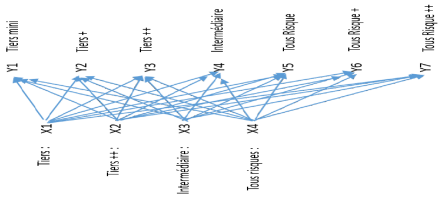

FIGURE 1 – Graphe des appariements possibles entre X et Y.

Nous proposons des contraintes très simples consistant à fixer à zéro certaines valeurs de $\mathbf{p}_{..}$, ceci afin de respecter la contrainte d'identifiabilité $\dim(\mathbf{p}_{..}) \leq J-1$. Une contrainte de ce type correspond à un modèle que nous noterons $\mathbf{m}$, et il y aura au total un ensemble de modèles $\mathcal{M} = \{\mathbf{m}\}$ que nous mettrons plus tard en compétition.

À modèle $\mathbf{m}$ fixé pour le moment, l'estimation des paramètres $\mathbf{p}_{..}$ et $\mathbf{p}_{.}$ peut se faire en maximisant la log-vraisemblance des données observées définie par

$$\ell_{\mathbf{m}}(\mathbf{p}_{..}, \mathbf{p}_{.}; \mathbf{X}^-, \mathbf{Y}^+) = \sum_{j=1}^{J} N_j^+ \ln \left(\sum_{i=1}^{I} p_{ij} p_i \right) + \sum_{i=1}^{I} N_i^- \ln p_i \tag{2}$$

où $N_i^- = \#\{X_n^- = i, n = 1, \ldots, N^-\}$ et $N_j^+ = \#\{Y_n^+ = j, n = 1, \ldots, N^+\}$. Comme il s'agit d'un problème à données manquantes (les données $\mathbf{X}^+$ et $\mathbf{Y}^-$), un algorithme EM [1, 4] est approprié pour maximiser cette log-vraisemblance. Cependant, à ce stade du travail, nous proposons de faire plus simple, même si sous-optimal, en maximisant la vraisemblance de façon séquentielle. Ainsi, dans un 1er temps, nous maximisons la log-vraisemblance marginale $\ell_{\mathbf{m}}(\mathbf{p}_{.}; \mathbf{X}^-) = \sum_{i=1}^{I} N_i^- \ln p_i$ pour estimer $\mathbf{p}_{.}$. Un estimateur $\hat{\mathbf{p}}_{.}$ est donc obtenu classiquement à partir des fréquences observées de $\mathbf{X}^-$. Ensuite, nous maximisons la log-vraisemblance conditionnelle $\ell_{\mathbf{m}}(\mathbf{p}_{..}, \hat{\mathbf{p}}_{.}; \mathbf{Y}^+|\mathbf{X}^-) = \sum_{j=1}^{J} N_j^+ \ln \left(\sum_{i=1}^{I} p_{ij} \hat{p}_i \right)$, qui est une fonction concave en $\mathbf{p}_{..}$, pour obtenir un estimateur $\hat{\mathbf{p}}_{..}(\mathbf{m})$, que nous noterons plus simplement $\hat{\mathbf{p}}_{..}$ en absence d'ambiguïté. Cette optimisation est réalisée par un algorithme de maximisation sous contrainte standard, la contrainte étant imposée par le modèle $\mathbf{m}$.

2.3 Sélection et agrégation de modèles

Le choix entre les différents modéles $\mathbf{m}$ de l'ensemble $\mathcal{M}$ se fait à l'aide du critère BIC conditionnel [3] donné par

$$\mathrm{BIC}_{\mathbf{m}} = -2\ell_{\mathbf{m}}(\hat{\mathbf{p}}_{..}, \hat{\mathbf{p}}_{.}; \mathbf{Y}^+|\mathbf{X}^-) + \nu_{\mathbf{m}} \ln N^+ \tag{3}$$

où $\nu_{\mathbf{m}} = \dim(\mathbf{m})$ correspond au nombre de paramètres libres du modèle $\mathbf{m}$. Le modèle $\hat{\mathbf{m}}$ conduisant à la valeur la plus faible du critère BIC est retenu.

Une autre façon d'utiliser BIC est de faire l'agrégation de modèle plutôt qu'une sélection en procédant à du *Bayesian model averaging* [2]. Un estimateur moyen $\bar{\mathbf{p}}_{..}$ est alors obtenu de la façon suivante :

$$\bar{\mathbf{p}}_{..} = \sum_{\mathbf{m}\in\mathcal{M}} \hat{\mathbf{p}}_{..}(\mathbf{m})\hat{P}(\mathbf{m}|\mathbf{X}^-, \mathbf{Y}^+) \tag{4}$$

où $\hat{P}(\mathbf{m}|\mathbf{X}^-, \mathbf{Y}^+) \propto \exp(-\mathrm{BIC}_{\mathbf{m}})$. Cette option permet d'obtenir une estimation de $\mathbf{p}_{..}$ qui au final permet de sortir de la contrainte possiblement trop forte des modèles de $\mathcal{M}$.

3 Expériences numériques

3.1 Données simulées

Nous évaluons tout d'abord notre méthode d'estimation et de choix de modèles sur un jeu de données où $N^- = 8\,052$ et $N^+ = 8\,052$. Ces valeurs d'effectifs ont été sélectionnées car s'approchant des valeurs constatées dans certains cas réels. Cependant, nous avons simplifié ici le nombre de modèles possibles en retenant $I = 3$ et $J = 4$.

Le tableau 1 indique sur la partie gauche le vrai jeu de paramètres et de données utilisé pour la simulation, et à droite les valeurs estimées par la vraisemblance et le critère BIC. Nous constatons alors que la méthode utilisée a permis de sélectionner le vrai modéle et de fournir une estimation assez précise des paramètres.

$X = i \backslash Y = j$	1	2	3	4
1	1	.	.	.
2	.	.	0.25	0.75
3	.	1	.	.

$X = i \backslash Y = j$	1	2	3	4
1	1	.	.	.
2	.	.	0.27	0.73
3	.	1	.	.

TABLE 1 – Probabilités d'appariements vraies (tableau de gauche) et estimées par le critère BIC (tableau de droite).

3.2 Données réelles

Nous considérons maintenant un jeu de données réelles provenant du comparateur en ligne de la société PIXEO, avec $N^- = 11\,441$, $N^+ = 8\,668$, $I = 4$ et $J = 7$. Au total, $\#\mathcal{M} = 4\,095$ modèles différents sont testés, correspondant au nombre exhaustif de modèles admissibles.

Le modèle estimé par BIC (la meilleure valeur de BIC est égale à 28 994.57), ainsi que la valeur des paramètres $\hat{\mathbf{p}}_{..}$ correspondante, sont donnés dans le tableau 2. Cette solution

indique par exemple que les internautes choisissant du Tiers (T) avant la modification rechoisiraient en majorité du Tiers s'ils revenaient avec les nouveaux descripteurs. Une plus petite partie choisiraient du Tiers++ (T++) néanmoins. De même, 100% des internautes choisissant du Tous Risques(TR), choisiraient de nouveau du Tous risques. Concernant les modalités ajoutées, le Tiers+(T+) serait privilégié par ceux qui auparavant choisissaient du Tiers++ (T++), le Tous risques+ (TR+) par ceux qui auparavant choisissaient de l'internédiaire (Inter).

$X \backslash Y$	T	T+	T++	Inter	TR	TR+	TR++
T	0.64	.	0.36	.	.	.	.
T++	.	0.90	.	.	.	.	0.10
Inter	.	.	.	0.74	.	0.26	.
TR	.	.	.	.	1.00	.	.

TABLE 2 – Probabilités d'appariements estimées selon le modèle ayant le plus petit BIC.

Comme plusieurs modèles ont des valeurs de BIC très proches, il a été décidé de procéder également à du *Bayesian model averaging*. Le tableau 3 présente les solutions estimées de cette manière. On remarque par exemple que la modalité intermédiaire (Inter) se répartit cette fois un peu plus sur les niveaux inférieurs et supérieurs assez proches.

$X \backslash Y$	T	T+	T++	Inter	TR	TR+	TR++
T	0.64	.	0.34	0.02	.	.	.
T++	.	0.90	.	.	.	.	0.10
Inter	.	.	0.03	0.71	.	0.26	.
TR	.	.	.	.	1.00	.	.

TABLE 3 – Probabilités d'appariements estimées selon le *Bayesian model averaging*.

Références

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, pages 1–38, 1977.

[2] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging : A tutorial. *Statistical Science*, 14(4) :382–401, 1999.

[3] Emilie Lebarbier and Tristan Mary-Huard. Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la Société française de statistique*, 147(1) :39–57, 2006.

[4] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.