Détection en ligne de multiples changements dans un panel de données catégorielles

Milad Leyli-Abadi, Allou Same, Latifa Oukhellou

Détection en ligne de multiples changements dans un panel de données catégorielles

Milad Leyli-abadi*, Allou Samé*, Latifa Oukhellou*

*Université Paris-Est, IFSTTAR, COSYS, GRETTIA, F-77447 Marne-la-Vallée, France
{milad.leyli-abadi, allou.same, latifa.oukhellou}@ifsttar.fr

Résumé. Cet article présente une méthode de détection de changements communs à un ensemble de séquences catégorielles. La méthode proposée est basée sur un test séquentiel de rapport de vraisemblance généralisé fondé lui-même sur des chaînes de Markov non homogènes modélisant les données avant et après les changements.

1 Introduction

De nos jours, l'usage des données longitudinales devient de plus en plus répandu dans de nombreux domaines. Par exemple, dans les réseaux urbains (électricité ou eau), les compteurs intelligents permettent la collecte de telles données sur la consommation de multiples usagers. Dans ce travail, nous nous intéressons plus spécifiquement à l'analyse conjointe de multiples séquences catégorielles où chaque catégorie correspond à un mode d'usage dans le réseau.

Cet article propose une méthode en ligne basée sur le test séquentiel du rapport de vraisemblance généralisé (Basseville et al., 1993) pour détecter des changements dans de multiples séquences catégorielles. Ces changements pourront être interprétés comme des modifications du comportement des usagers du réseau. Nous proposons un seuil adaptatif permettant de décider d'éventuels changements de comportement. Dans le domaine des réseaux urbains notamment, la détection de changement permettra aux gestionnaires de réseau, de mieux répondre aux besoins évolutifs des usagers.

D'autre part, le comportement des usagers étant très souvent lié à des facteurs exogènes (température, précipitations, etc.) (House-Peters et al., 2010), nous modélisons les séquences catégorielles avant et après changement par une chaîne de Markov non homogène.

L'article est organisé de la manière suivante : les sections 2 et 3 décrivent respectivement les données et la méthodologie adoptée. Les résultats expérimentaux et l'évaluation de la méthode proposée sont détaillés dans la section 4.

2 Données

Le panel de données catégorielles analysé dans cet article, noté $(z_{it})_{1\leq i\leq n, 1\leq t\leq T}$, est relatif à n entités (ex. compteurs communicants) observées durant T instants (jours ou semaines). La figure 1 illustre ce type de données, chaque couleur faisant référence à une catégorie. Dans notre cas, les données exogènes associées à ces séquences catégorielles, qui seront notée $\mathbf{u} = (u_{it})$, comprennent généralement la température, la précipitation et les données calendaires. Dans la suite de l'article, on utilisera de manière équivalente les notations $(\mathbf{z}_1, ..., \mathbf{z}_T)$, avec $\mathbf{z}_t = (z_{it})_{i=1,...,n}$.

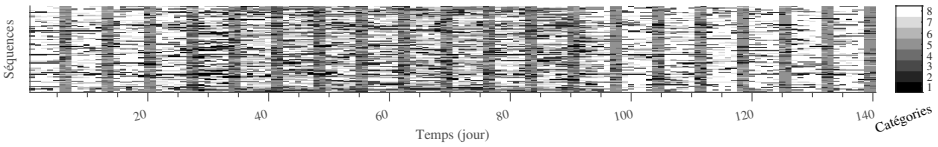


FIG. 1: Séquences catégorielles correspondant au comportement de 100 usagers d'un réseau durant 140 jours. Chaque ligne correspond à l'évolution du comportement d'un usager et chaque catégorie correspond à un profil d'usage journalier

3 Méthode proposée

L'approche proposée pour la détection de changement dans un ensemble de séquences catégorielles est basée sur le test séquentiel du rapport de vraisemblance généralisé. Les hypothèses du test sont définies comme suit :

$$\begin{cases} H_0 : & (\mathbf{z}_1, ..., \mathbf{z}_T) & \sim & P_{\tilde{\theta}_0} \\ H_A : & (\mathbf{z}_1, ..., \mathbf{z}_{\tau-1}) & \sim & P_{\theta_0} \\ & (\mathbf{z}_\tau, ..., \mathbf{z}_T) & \sim & P_{\theta_1} \end{cases} \tag{1}$$

où H_0 est l'hypothèse sous laquelle la séquence entière de données est distribuée suivant la même loi $P_{\tilde{\theta}_0}$ et l'hypothèse H_1 considère qu'à partir d'un instant τ, les données ne suivent plus le même modèle qu'avant cet instant. La distribution P_θ d'une séquence $(\mathbf{z}_a, \ldots, \mathbf{z}_b)$ donnée, avec $a < b$, est supposée être celle d'une chaîne de Markov non homogène dont les probabilités initiales et de transition sont définies comme suit :

$$P_\theta(z_{i,1} = k|\mathbf{u}_1) = \frac{e^{\boldsymbol{\alpha}_k^\top \mathbf{u}_{i,1}}}{\sum_{\ell=1}^K e^{\boldsymbol{\alpha}_\ell^\top \mathbf{u}_{i,1}}}, \tag{2}$$

$$P_\theta(z_{i,t} = k|z_{i,t-1} = \ell, \mathbf{u}_t) = \frac{e^{\boldsymbol{\beta}_{k,\ell}^\top \mathbf{u}_{i,t}}}{\sum_{h=1}^K e^{\boldsymbol{\beta}_{h,\ell}^\top \mathbf{u}_{i,t}}}, \tag{3}$$

où $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ désignent respectivement les paramètres associés aux probabilités initiales et de transition. Afin de décider entre les deux hypothèses, nous nous appuyons sur le logarithme du rapport de vraisemblance, défini par :

$$\Lambda_1^T(\tau) = \log \left(\frac{\left(\prod_{i=1}^n P_{\theta_0}(z_{i1}|\mathbf{u}_{i1}) \prod_{t=2}^{\tau-1} P_{\theta_0}(z_{it}|z_{it-1}, \mathbf{u}_{it}) \times \right.}{\left. \prod_{i=1}^n P_{\theta_1}(z_{i\tau}|\mathbf{u}_{i\tau}) \prod_{t=\tau+1}^T P_{\theta_1}(z_{it}|z_{it-1}, \mathbf{u}_{it}) \right)} \right) \tag{4}$$

En développant cette équation, on obtient :

$$\begin{aligned} \Lambda_1^T(\tau) &= \sum_{i=1}^n \log P_{\theta_0}(z_{i1}|\mathbf{u}_{i1}) + \sum_{i=1}^n \sum_{t=2}^{\tau-1} \log P_{\theta_0}(z_{it}|z_{it-1}, \mathbf{u}_{it}) \\ &+ \sum_{i=1}^n \log P_{\theta_1}(z_{i\tau}|\mathbf{u}_{i\tau}) + \sum_{i=1}^n \sum_{t=\tau+1}^T \log P_{\theta_1}(z_{it}|z_{it-1}, \mathbf{u}_{it}) \\ &- \sum_{i=1}^n \log P_{\tilde{\theta}_0}(z_{i1}|\mathbf{u}_{i1}) - \sum_{i=1}^n \sum_{t=2}^T \log P_{\tilde{\theta}_0}(z_{it}|z_{it-1}, \mathbf{u}_{it}). \end{aligned} \tag{5}$$

Les paramètres $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_{0\ell}, \boldsymbol{\alpha}_1, \boldsymbol{\beta}_{1\ell}, \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\beta}}_{0\ell}, \tau)$ sont estimés en utilisant la méthode de maximum de vraisemblance :

$$\Lambda_T = \max_{\tau, (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1), \tilde{\boldsymbol{\theta}}_0} \Lambda_1^T(\tau). \tag{6}$$

La règle de décision suivante permet finalement de décider entre les deux hypothèses :

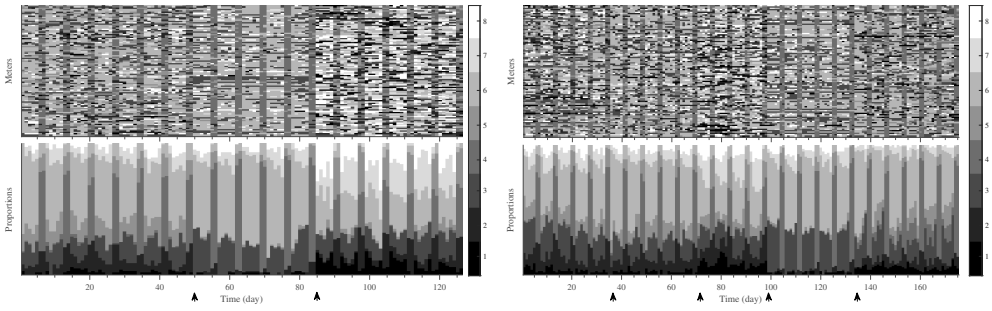
$$d = \{0 \text{ si } \Lambda_T < h;\ 1 \text{ si } \Lambda_T \geq h\}, \tag{7}$$

où h est le seuil de décision. Si la statistique de test dépasse cette valeur, un changement est détecté. La stratégie qui vient d'être décrite permet la détection d'un unique point de changement dans une séquence. Pour adapter cette stratégie à la détection de multiples changements, nous l'appliquons de manière séquentielle sur des fenêtres de taille croissante. Tant qu'aucun point de changement n'est détecté, la méthode de détection est appliquée sur une nouvelle fenêtre de taille plus grande que la précédente. Dès qu'un point de changement est détecté, la fenêtre est réinitialisée de même que le seuil de détection.

Estimation d'un seuil adaptatif. Pour avoir une estimation de la valeur du seuil, nous avons effectué des simulations de type Monte Carlo. À partir d'une séquence de données initiales ne présentant pas de changement (comportement nominal), un modèle de Markov non homogène est d'abord estimé par la méthode du maximum de vraisemblance. Ensuite, plusieurs séquences sont générées à partir de ce modèle et la statistique de test est évaluée pour chacune de ces séquences. Le fractile $(\mathcal{Q}_{1-p})$ de la distribution des statistiques de test est considéré comme la valeur du seuil. Le seuil est estimé de nouveau après chaque détection.

4 Expérimentation

Afin d'évaluer la performance de la méthode proposée, nous avons conçu deux bases de données en considérant un nombre de changements différent (voir Figure 2). La méthode proposée est comparée à deux approches, l'une basée sur un modèle de Markov homogène (MM) et l'autre ne faisant pas l'hypothèse markovienne (équivalent à un modèle de régression logistique). Cette dernière méthode sera donc notée LR.

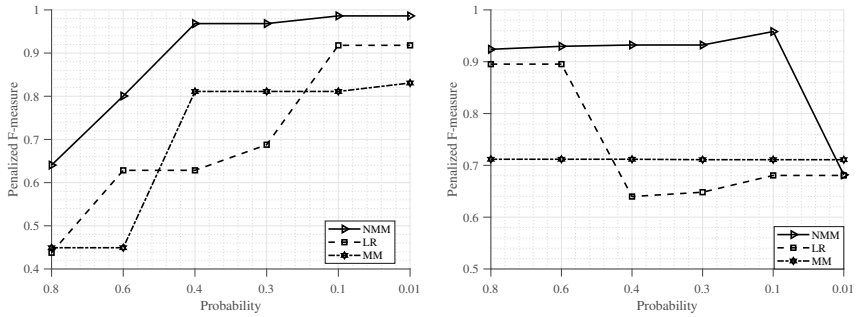


(a) Base 1 (2 changements) (b) Base 2 (4 changements)

FIG. 2: Représentation graphique des jeux données. Les figures du dessus représentent les bases de données générées, et les figures du dessous représentent la proportion des catégories au fils du temps. Les changements sont indiqués par des flèches sur l'axe des abscisses.

La comparaison est effectuée en termes de la F-mesure (voir la Figure 3) et trois autres critères (voir le tableau 1) qui sont l'aire sous la courbe ROC (AUC), le taux de vrais positifs

(TPR) et le délai de détection (DD). La figure 3 montre pour chacune des bases, le calcul de la F-mesure en fonction de différentes valeurs de probabilité (p) associées au fractile (différentes valeurs de seuil). En observant ces deux graphiques et le tableau 1, on remarque que les meilleures performances sont obtenues pour la méthode proposée.



(a) Base 1 (b) Base 2

FIG. 3: Calcul de la F-mesure pour chacune des bases et en utilisant les 3 méthodes

5 Conclusion

Dans cet article, une méthode de détection de changement basée sur la rapport de vraisemblance généralisé est proposée. Un seuil adaptatif est calculé permettant d'ajuster ce dernier aux différents types de changements et de réduire le nombre de fausses alarmes. Les expérimentations réalisées sur deux bases de données montrent de bonnes performances de la méthode proposée.

TAB. 1: Tableau de comparaison des méthodes évaluées : MM : modèle de Markov homogène ; LR : régression logistique ; NMM : modèle de Markov non-homogène (modèle proposé). Les différents critères sont : AUC : aire sous la courbe ROC ; TPR : taux de vrais positifs ; DD : Délai de détection.

Modèle	**MM**			**LR**			**NMM**		
Critère	**AUC**	**TPR**	**DD**	**AUC**	**TPR**	**DD**	**AUC**	**TPR**	**DD**
Base 1	0.77	0.78	7.5	0.82	0.73	**5**	**0.92**	**0.96**	7.5
Base 2	0.70	0.65	**2.6**	0.83	0.76	6.3	**0.91**	**0.96**	8.7

Références

Basseville, M., I. V. Nikiforov, et al. (1993). *Detection of abrupt changes : theory and application*, Volume 104. Prentice Hall Englewood Cliffs.

House-Peters, L., B. Pratt, et H. Chang (2010). Effects of urban spatial structure, sociodemographics, and climate on residential water consumption in hillsboro, oregon 1. *JAWRA Journal of the American Water Resources Association 46*(3), 461–472.

Summary

This article presents a change detection method performing on a set of categorical sequences. The proposed method is based on a generalized sequential likelihood ratio test, where the data are supposed to be distributed following a non homogeneous Markov model before and after potential change points.