Préparation supervisée de données dynamiques

Sylvain Ferrandiz[*,**], Marc Boullé[*]

[*]France Télécom R&D,
2, avenue Pierre Marzin, 22300 Lannion
sylvain.ferrandiz@francetelecom.com,
marc.boulle@francetelecom.com,
[**]GREYC, Université de Caen,
boulevard du Maréchal Juin, BP 5186, 14032 Caen Cedex,

Résumé. L'aspect technique représentant de moins en moins une contrainte, il est aujourd'hui possible de suivre dans le temps les caractéristiques mesurées. Le statisticien est en conséquence confronté à la présence d'un nombre croissant de variables dynamiques. Dans le cadre de la préparation de données pour la classification supervisée, nous développons une méthode d'évaluation de la pertinence d'une variable dynamique. Le propos est illustré sur un problème de classification de courbes de consommation téléphonique.

1 Evaluation supervisée de variables dynamiques

Avec l'émergence des systèmes d'information au tournant des années 90, la récolte des données brutes a été rendue indépendante de toute finalité statistique. Modéliser directement de telles données est devenu impossible. La phase de préparation des données, dont l'objectif est de construire une table de données pour modélisation à partir des données brutes, est donc devenue une partie critique et souvent coûteuse en temps du processus de fouille de données Chapman et al. (2000).

L'évolution des moyens techniques le permettant, il est aujourd'hui possible de mesurer l'évolution dans le temps d'une caractéristique, et ce sur une longue période. A côté des variables usuelles, qu'on qualifie de *statiques*, sont donc de plus en plus présentes des variables *dynamiques* : mesure de l'activité cardiaque en médecine, mesure de la pression en météorologie, mesure de la consommation téléphonique en télécommunication,...

En phase de préparation, les variables dynamiques sont placées dans le même contexte que les variables statiques et sont naturellement amenées à subir les mêmes traitements. Si on cherche à expliquer une variable cible symbolique, les tâches principales de préparation sont l'évaluation de la dépendance entre variable(s) explicative(s) et variable cible, et la sélection de variables explicatives.

On se propose dans ce papier d'illustrer l'apport de la méthode introduite par Ferrandiz et Boullé (2006) à l'évaluation supervisée de variables dynamiques. Pour cela, on conduit une expérimentation sur un problème de classification de profils de consommation en téléphonie fixe.

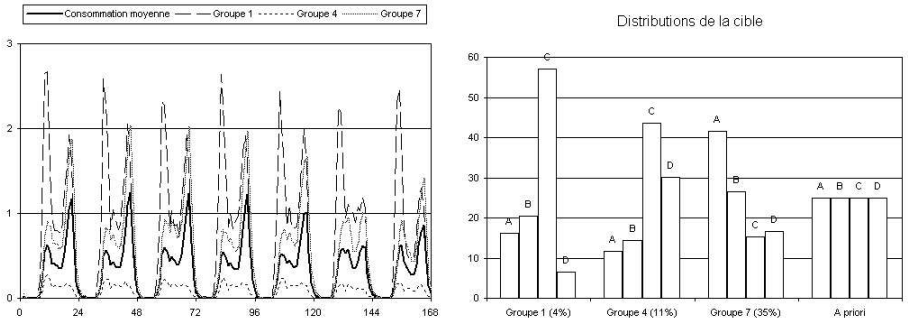

FIG. 1 – *Caractéristiques des groupes 1, 4 et 7. Les instances du groupe 1 correspondent à de très fortes consommations, avec des pics très marqués. Le groupe 1 contient majoritairement des instances de classe C.*

2 Classification de profils de consommation

On illustre les apports de la méthode décrite dans Ferrandiz et Boullé (2006) pour l'évaluation de variables dynamiques. On considère un problème de classification de 2636 profils de consommation suivant 4 classes cibles A, B, C et D. La distribution des classes cibles est uniforme sur l'échantillon. On dispose de 168 variables descriptives continues, chacune mesurant la consommation téléphonique sur une tranche horaire de la semaine, constituant ainsi une variable dynamique.

L'évaluation fournit une valeur du critère de 0.051, ce qui est très faible et caractérise un fort mélange des classes cibles. La méthode partitionne les instances en 7 groupes. En calculant la valeur moyenne de chacune des 168 variables et ce dans chaque groupe, on obtient 7 profils de consommation caractéristiques. Trois de ces profils sont reportés sur la Figure 1, ainsi que le profil moyen de consommation (*i.e.* celui calculé sur tout l'ensemble d'apprentissage). De plus, à chaque groupe est associée une distribution des classes cibles, qu'on représente par des histogrammes groupés sur cette même figure.

Le résultat est facilement interprétable. Par exemple, le profil moyen des instances du groupe 7 se distingue du profil moyen global par une consommation plus élevée. Ces instances sont en grand nombre (35% des instances) et la répartition dans les classes cibles A, B, C, D est $(41\%, 26\%, 15\%, 17\%)$. Le groupe 1 est quant à lui plus discriminant (la répartition dans les classes cibles est $(16\%, 20\%, 57\%, 6\%)$), mais de taille réduite (4% des instances) : il caractérise un comportement atypique mais discriminant.

Références

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, et R. Wirth (2000). *CRISP-DM 1.0 : step-by-step data mining guide*.

Ferrandiz, S. et M. Boullé (2006). Sélection supervisée d'instances : une approche descriptive. In *Actes de la conférence EGC*, Volume 2, pp. 421–432.