Evaluation supervisée de métrique : application à la préparation de données séquentielles

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. De nos jours, le statisticien n'a plus nécessairement le contrôle sur la récolte des données. Le besoin d'une analyse statistique vient dans un second temps, une fois les données récoltées. Par conséquent, un travail est à fournir lors de la phase de préparation des données afin de passer d'une représentation informatique à une représentation statistique adaptée au problème considéré. Dans cet article, nous étudions un procédé de sélection d'une bonne représentation en nous basant sur des travaux antérieurs.
Nous proposons un protocole d'évaluation de la pertinence d'une représentation par l'intermédiaire d'une métrique, dans le cas de la classification supervisée. Ce protocole exploite une méthode de classification non paramétrique régularisée, garantissant l'automaticité et la fiabilité de l'évaluation. Nous illustrons le fonctionnement et les apports de ce protocole par un problème réel de préparation de données de consommation téléphonique. Nous montrons également la fiabilité et l'interprétabilité des décisions qui en résultent.

1 Préparation de données

Avec l'émergence des systèmes d'information au tournant des années 90, la récolte des données brutes a été rendue complètement indépendante de toute finalité statistique. L'analyse de ces données est un objectif qui intervient dans un second temps. La phase de préparation, dont le but est de construire à partir des données brutes une table de données pour modélisation, est donc devenue une partie critique et souvent coûteuse en temps du processus de fouille de données (Chapman et al., 2000).

L'analyste se trouve dans la situation suivante. D'une part, il dispose d'un entrepôt de données mis en place et alimenté dans un autre but que celui d'une quelconque analyse statistique. D'autre part, le propriétaire de l'entrepôt envisage d'exploiter ses données afin de compléter ses connaissances et pose une question à l'analyste. Celui-ci doit alors tourner la question en un problème d'analyse statistique, extraire de l'entrepôt les données susceptibles d'être pertinentes vis-à-vis de la question posée, les mettre sous forme d'une table, procéder à la modélisation et interpréter les résultats afin de répondre à la question initiale.