Illustration d'une méthode d'évaluation supervisée par un problème de classification de courbes

Sylvain Ferrandiz, Marc Boullé

France Télécom R&D
2, avenue Pierre Marzin, 22300 Lannion
sylvain.ferrandiz@francetelecom.com
marc.boulle@francetelecom.com

RÉSUMÉ. *La récolte des données est de moins en moins contrainte par l'aspect technique de sa mise en œuvre. En conséquence, il est aujourd'hui possible de suivre dans le temps toute caractéristique mesurée. Lors de la préparation d'une table de données et de la construction d'un modèle, le statisticien doit ainsi compter avec la présence d'un nombre croissant de variables dynamiques. A côté des questions usuellement traitées en phase de préparation, comme la sélection des variables pertinentes, toute variable dynamique soulève de plus un problème de représentation. Afin d'automatiser la prise de décision, aujourd'hui basée sur la connaissance métier, nous appliquons une méthode d'évaluation supervisée pour quantifier la pertinence d'une variable dynamique. L'évaluation est automatique et régularisée, ce qui profite à la qualité des choix opérés. Le propos est illustré sur un problème de classification de courbes de consommation téléphonique.*

MOTS-CLÉS : *Classification supervisée, préparation de données, variables dynamiques.*

1. Préparation de données et variables dynamiques

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. 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 [CHA 00].

L'évolution des moyens techniques le permettant, il est aujourd'hui possible de suivre dans le temps une caractéristique, et ce sur une longue période. A côté des variables usuelles, qu'on qualifie ici 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, mesure de la propagation des ondes en sismologie. Une variable dynamique est ainsi formée par une suite de mesures et se distingue d'une variable multivariée par son caractère séquentiel.

De par la décorrélation entre la récolte des données et la modélisation, la précision des mesures et l'échelle des temps de mesure sont sans aucun rapport avec les besoins d'une étude statistique, car seulement limitées par les contraintes techniques. Dès lors, la préparation de données dynamiques passe nécessairement par une phase de recherche de représentation : d'une représentation brute (les données observées) il faut passer à une représentation cohérente pour modélisation subséquente, et ce pour chacune des variables dynamiques. Au cours de cette recherche de représentation, d'autres problèmes que celui de l'échelle des temps de mesure sont à traiter, comme le bruit sur les mesures, le non alignement des temps de mesure, le facteur d'échelle entre les individus, etc.

En phase de préparation, hormis le problème de représentation, 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, ainsi que la sélection de variables explicatives.

En pratique, la sélection d'une représentation pour chaque variable dynamique et la sélection de variables dynamiques sont basées sur la connaissance métier : en reconnaissance de la parole, il est "usuel" de travailler avec les log-périodogrammes des signaux ; en téléphonie, il est "usuel" de travailler avec un découpage en tranches horaires prédéterminé ; en médecine, il est "usuel" de considérer à la fois un électroencéphalogramme, un électro-oculogramme et un électromyogramme afin d'étudier la phase de sommeil paradoxal. La connaissance métier porte sur un phénomème particulier et s'accumule au fur et à mesure qu'on valide de nouvelles hypothèses sur ce phénomène.

On applique ici une méthode d'évaluation non paramétrique et générique jugeant la qualité d'une représentation à l'étude d'une variable dynamique. Ses qualités favorisent l'automatisation de la prise de décision et la rendent indépendante du domaine d'application. Le contexte est celui de la classification supervisée. Le critère d'évaluation c est introduit dans [FER 06b] et l'algorithme d'optimisation dans [FER 06a]. On se propose dans ce papier d'illustrer l'apport de la méthode sur un problème d'évaluation supervisée de variables dynamiques. La section 2 dérive du critère c une méthode d'évaluation de la pertinence d'une représentation. La section 3 montre son apport à travers une expérimentation sur un problème de classification de profils de consommation téléphonique.

2. Une approche informationnelle de l'évaluation

Dans le cas d'une variable statique continue, [BOU 06] aborde la question de l'évaluation de la pertinence vis-à-vis d'une variable cible symbolique comme un problème de modélisation. Les modèles considérés sont les partitions de la variable continue en intervalles. Une approche informationnelle permet de définir un critère s'interprétant comme la probabilité que le modèle explique les données. La sélection du modèle le plus probable conduit à une méthode de discrétisation d'une variable statique continue. La probabilité que ce modèle explique les données s'utilise alors comme un indicateur de pertinence de la variable descriptive relativement à la variable cible.

Dans [FER 06b], l'approche est adaptée afin de traiter le cas où l'on dispose d'une mesure de similitude entre les instances de l'échantillon. Toute partition de Voronoi induite par un sous-ensemble d'instances constitue un modèle. Le partitionnement d'une variable en intervalles est ainsi généralisé en un partitionnement de l'espace en cellules. La probabilité qu'un modèle explique les données est explicitée et la sélection du modèle le plus probable conduit à une méthode de sélection d'instances. Là encore, la probabilité associée au modèle sélectionné constitue un indicateur supervisé de pertinence. La méthode est évaluée dans [FER 06a] en tant que méthode de sélection d'instances pour la classification par le plus proche voisin.

En pratique, la représentation des données dynamiques conduit à définir une matrice de similitude. Par exemple, une fois la transformée de Fourier appliquée, il est usuel d'utiliser une distance euclidienne pondérée. On considère donc qu'une représentation R n'est autre qu'une matrice de similitude. Ainsi, on utilise le critère $c(M, R)$ présenté dans [FER 06b], qui mesure de manière supervisée l'intérêt d'une partition de Voronoi M relative à un sous-ensemble de l'échantillon et définie à l'aide de la matrice R. Dès lors, si on note

$$c^*(R) = min_M\, c(M, R),$$

la fonction c^* fournit une évaluation de la qualité de la représentation R et permet ainsi de comparer différentes représentations. Pour une matrice de similitude R donnée, il suffit d'appliquer un algorithme d'optimisation combinatoire et d'attribuer à R la valeur rencontrée optimale du critère $c(M, R)$. Une heuristique d'optimisation efficace est présentée dans [FER 06a]. Le critère c est le suivant :

$$c(M, R) = \log N + \log \binom{N + K - 1}{K} + \sum_{k=1}^{K} \log \binom{N_k + J - 1}{J - 1} + \sum_{k=1}^{K} \log \frac{N_k!}{N_{k1}! \ldots N_{kJ}!}.$$

où N désigne le nombre d'instances de l'échantillon, J le nombre de classes cibles, K le nombre de groupes de la partition, N_k le nombre d'instances dans la k^{eme} cellule et N_{kj} le nombre d'instances dans la k^{eme} cellule portant la j^{eme} étiquette.

Le critère c est non paramétrique et régularisé. Il quantifie le compromis entre le nombre de groupes de la partition (les deux premiers termes) et la distribution de la cible (les deux derniers termes), ce qui correspond à un compromis entre complexité du modèle et ajustement du modèle aux données de l'échantillon. La régularisation est un moyen sûr d'endiguer le phénomène de sur-apprentissage. Etant de surcroît non paramétrique, l'évaluation se passe de validation ou de validation croisée. On dispose ainsi de plus d'instances pour ajuster le modèle, ce qui augmente sa qualité.

Afin de travailler avec un indicateur normalisé, on considère la transformation suivante de c^* :

$$g^*(R) = 1 - \frac{c^*(R)}{c_0(R)},$$

où $c_0(R)$ est la valeur du critère $c(M, R)$ pour le modèle M constitué par un seul groupe. Comme $c(M, R)$ n'est autre que l'opposé du logarithme d'une probabilité et comme une telle quantité s'interprète comme une longueur de codage (d'après les travaux de [SHA 48]), $g^*(R)$ mesure un gain de compression. Le gain de compression $g^*(R)$ est supérieur à 0 (dès lors que la partition en un unique groupe est évaluée durant l'optimisation) et inférieur à 1. Si $g^*(R) = 0$, la représentation R n'apporte aucune information sur la variable cible. Plus la valeur de $g^*(R)$ est proche de 1, plus les classes cibles sont séparées. Il est à noter que ce critère est générique et ne se limite pas à l'évaluation de représentations de données dynamiques.

3. Classification de profils de consommation

On illustre les apports de notre méthode par une expérimentation sur des données de consommation en téléphonie fixe. C'est un problème de classification de 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 et ce pour 2636 instances. On applique la méthode à la variable dynamique constituée par les 168 variables descriptives. La mesure de similitude adoptée est la métrique L_1.

L'évaluation fournit un gain de compression de 0.051, ce qui est très faible et caractérise un fort mélange des classes cibles. En plus de quantifier la pertinence d'une variable, elle fournit un support à la discrimination réalisée : une distribution des classes cibles et un prototype accompagnent chaque groupe. La méthode partitionne ici les instances en 7 groupes et les caractéristiques de trois d'entre eux sont reportées sur la figure 1. Les distributions relatives à chacun de ces groupes sont représentées par des histogrammes groupés. Dans chaque groupe, en calculant la valeur moyenne de chacune des 168 variables, on obtient un profil de consommation moyen caractéristique de ce groupe. Ce profil est plus parlant que le simple profil de consommation du prototype car il tient compte de toutes les instances du groupe.

Le modèle étant visualisable, il est facilement interprétable. Par exemple, on voit que les individus du groupe 7 sont en grand nombre (35% des instances), qu'ils une consommation moyenne plus élevée que la moyenne globale, et que ce comportement est majoritairement caractéristique de la classe A (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 $(16\%, 20\%, 57\%, 6\%)$) avec un profil de consommation atypique (pics de consommation élevés), mais est de taille réduite (4% des instances). Le groupe 4 discrimine lui aussi la classe C, moins fortement tout de même que le groupe 1, et se différentie par une consommation moyenne très faible.

Il est à noter que ce n'est pas la visualisation en elle-même qui est nouvelle. Elle peut en effet être utilisée conjointement à toute méthode fournissant un ensemble de prototypes. La nouveauté réside dans le fait que la méthode d'évaluation proposée ici optimise exactement les paramètres de cette visualisation. En effet, les ca-

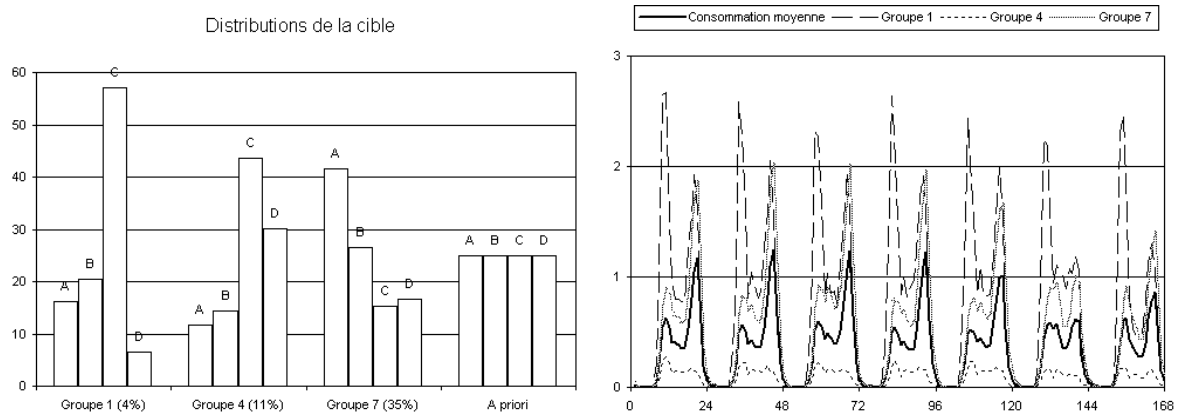

FIG. 1. *Caractéristiques des groupes 1, 4 et 7. A chaque profil est associée une distribution des classes cibles. Les groupes 1 et 4 contiennent majoritairement des instances de classe C. Les instances du groupe 1 correspondent à de très fortes consommations, avec des pics très marqués. Celles du groupe 4 correspondent aux faibles consommations.*

ractéristiques visualisées (taille des groupes, distribution des classes cibles dans les groupes, prototypes) sont exactement celles apparaissant dans le critère et les modèles. La visualisation n'en est que plus pertinente.

4. Conclusion

A cours d'un processus de fouille de données, le statisticien traite aujourd'hui aussi bien des variables statiques que des variables dynamiques. Ces dernières soulèvent avec plus de force la question du choix d'une représentation. Nous avons appliqué dans ce papier un procédé d'évaluation supervisée de la pertinence d'une représentation.

En évitant de passer par une phase de validation, on utilise plus de données pour l'apprentissage. Ceci assure une pertinence plus forte de l'hypothèse validée. En gérant le sur-apprentissage, l'hypothèse sélectionnée ne "colle" pas aux données, ce qui assure sa robustesse. C'est l'adoption d'une approche informationnelle qui conduit à une telle évaluation.

Le bon comportement attendu a été illustré sur un jeu de données réel, dans le but de classer de manière supervisée des courbes de consommation téléphonique. Les expérimentations ont montré l'apport explicatif de la structuration des données : la méthode optimise ce que l'utilisateur voit. De par sa généricité, l'évaluation n'est pas limitée au problème de représentation des données dynamiques.

5. Bibliographie

[BOU 06] BOULLÉ M., MODL : a bayes optimal discretization method for continuous attributes, *Machine learning*, , A paraître en 2006.

[CHA 00] CHAPMAN P., CLINTON J., KERBER R., KHABAZA T., REINARTZ T., SHEARER C., WIRTH R., CRISP-DM 1.0 : step-by-step data mining guide, 2000.

[FER 06a] FERRANDIZ S., BOULLÉ M., Sélection supervisée d'instances : une approche descriptive, *Actes de la conférence sur l'extraction et la gestion des connaissances*, vol. 2, 2006, p. 421–432.

[FER 06b] FERRANDIZ S., BOULLÉ M., Supervised evaluation of Voronoi partitions, *Journal of intelligent data analysis*, , A paraître en 2006.

[SHA 48] SHANNON C., A mathematical theory of communication, rapport, 1948, Bell systems technical journal.