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.

Entre autres, la préparation passe par la définition et la sélection des individus et variables constituant la table qui va servir à la modélisation. Les possibilités qui s'offrent à l'analyste lors de cette étape sont, virtuellement, limitées uniquement par son imagination. En pratique, l'extraction et la mise en forme sont soumises à deux contraintes : celle sur les ressources et celle sur le passage à l'échelle des méthodes de modélisation.

D'une part, le temps alloué à une étude est nécessairement limité, souvent très contraint. Les données brutes ne sont pas toujours accessibles facilement et rapidement. D'autre part, les algorithmes de modélisation sont rarement linéaires en le nombre de lignes et colonnes de la table. De manière plus insidieuse, l'analyste doit éviter de tomber dans le piège de la dimension, qui conduit à considérer un nombre de variables trop élevé pour le nombre d'individus à disposition. L'information portée par les individus est noyée dans un espace de représentation de taille inadaptée et de nombreuses techniques de modélisation ont les pires difficultés à produire un modèle pertinent.

Nous laissons de côté la question de la définition des individus et nous intéressons à la définition des colonnes de la table. L'analyste doit capturer à l'aide d'un ensemble de variables l'information pertinente pour la question posée, sans autre aide que son intuition sur les variables "a priori susceptibles" d'expliquer le phénomène étudié et sous contrainte de ressource et de passage à l'échelle des procédés de modélisation. Dans cet article, nous nous proposons d'aider l'analyste dans cette tâche.

L'article est organisé comme suit. La section 2 présente le cadre de la sélection de variable, pour mieux circonscrire le lieu où se situe notre contribution. Notamment, nous montrons l'intérêt de disposer d'une méthode d'évaluation automatique et fiable d'une métrique. La section 3 déduit des travaux de Ferrandiz et Boullé (2006b) une telle méthode, dans le cadre de la classification supervisée. Pour la convenance du lecteur, la technique de modélisation introduite dans Ferrandiz et Boullé (2006b) est brièvement décrite dans la section 4. Enfin, la section 5 illustre notre propos par des expérimentations sur un problème de préparation de profils de consommation en téléphonie fixe dans un contexte supervisé.

2 Sélection de variables

Le problème de la sélection de variables ne se pose pas uniquement en préparation et sa résolution dans différents contextes a donné lieu à la publication de nombreux articles. Nous posons ici les définitions utiles pour la suite et présentons une taxonomie simplifiée des méthodes de sélection, puis nous montrons l'intérêt d'avoir à disposition une méthode d'évaluation de la pertinence d'une métrique.

Nous notons $\mathcal{I} = [\![1, N]\!]$ l'ensemble des N individus. Une *variable* est une fonction de $\mathcal{I}$ dans un ensemble $\mathbb{X}$. Nous appelons cet ensemble *espace de représentation* de la variable.

Si $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_D$ avec $D \in \mathbb{N}^*$, X est dite *statique*. Elle est dite *multidimensionnelle* si $D > 1$ et *unidimensionnelle* sinon. Si les $\mathbb{X}_d$ sont des alphabets, X est dite *catégorielle*. Si les $\mathbb{X}_d$ sont des parties de $\mathbb{R}$, on dit que X est *numérique*.

Enfin, si $\mathbb{X}$ est un ensemble de suites à valeurs dans un ensemble $\mathbb{X}'$, X est dite *séquentielle*. Autrement dit, X est séquentielle lorsque chaque individu se voit associé une suite de mesures dans un même ensemble $\mathbb{X}'$.

En phase de préparation des données d'un processus de fouille, construire une variable, c'est définir une succession de transformations à appliquer aux données brutes. A partir des

données brutes, une variable $X : \mathcal{I} \rightarrow \mathbb{X}$ est construite en choisissant un espace de représentation $\mathbb{X}$ et en définissant un procédé de mesure X projetant chaque individu dans $\mathbb{X}$. Une fois la variable $X : \mathcal{I} \rightarrow \mathbb{X}$ obtenue, se pose la question de son intérêt relativement à la question étudiée. On entre alors dans le cadre de la sélection de variables.

De nombreux articles permettent de cerner les pratiques de la sélection, notamment Blum et Langley (1997), Kohavi et John (1997) et Guyon et Elisseeff (2003). Dans l'article Kohavi et John (1997), une distinction est opérée entre approche *enveloppe* (de l'anglais wrapper) et approche *filtre* (de l'anglais filter).

L'approche enveloppe consiste à évaluer l'impact d'une modification de l'ensemble de variables sur la performance d'un modèle. Une technique de modélisation étant spécifiée, elle est appliquée à différents ensembles de variables et l'ensemble de variables conduisant au modèle le plus performant est conservé. Chaque évaluation nécessite l'ajustement d'un modèle, ce qui se révèle coûteux en temps. Cette approche, en faisant intervenir le modèle dans l'évaluation, est plus adaptée à la phase modélisation qu'à la phase de préparation d'une analyse.

Dans l'approche filtre, l'intérêt des variables est estimé indépendamment d'une quelconque modélisation. Dans Blum et Langley (1997), il est proposé de construire les méthodes de filtrage en deux temps. Tout d'abord, l'analyste définit un critère de pertinence d'une variable ou d'un ensemble de variables puis il adopte une heuristique de sélection. Comme précisé dans Guyon et Elisseeff (2003), on parle d'évaluation *multivariée* lorsque le critère d'évaluation porte sur un ensemble de variables, et d'évaluation *univariée* lorsque chaque variable est évaluée individuellement.

L'approche filtre est naturellement adaptée à la phase de préparation d'un processus de fouille, plus particulièrement la version univariée. L'analyste définit un certain nombre de variables, les évalue individuellement et élimine celles déclarées non pertinentes. La préparation des variables est ainsi plus rapide et plus facile à remettre en question. A contrario, l'approche multivariée est d'autant plus coûteuse que le nombre de variables croît.

L'approche filtre univariée de la sélection de variables repose sur l'emploi d'une méthode d'évaluation de l'intérêt d'une variable quelconque $X : \mathcal{I} \rightarrow \mathbb{X}$. Non seulement en pratique mais aussi sur le plan formel, l'espace de représentation $\mathbb{X}$ est souvent muni d'une métrique (ou : distance) $\delta : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}_+$. C'est notamment le cas lorsque X est multidimensionnelle numérique, mais aussi lorsque X est séquentielle.

L'information contenue dans une variable peut donc être observée à travers le prisme d'une matrice de distance sur les individus. Une telle matrice est obtenue en calculant les distances au sens de δ entre tout couple d'individus. Dans ce cas, l'évaluation de la métrique induit une évaluation de la pertinence de la variable X.

3 Evaluation probabiliste d'une métrique

Nous proposons ici un protocole d'évaluation de la pertinence d'une métrique exploitant des travaux antérieurs. Le contexte est celui de la classification supervisée : une variable cible catégorielle unidimensionnelle est à expliquer.

Dans le cas d'une variable statique numérique unidimensionnelle, Boullé (2006) aborde la question de l'évaluation de la pertinence vis-à-vis d'une variable cible catégorielle comme un problème de maximisation de la probabilité a posteriori. Les modèles considérés sont les partitions de la variable numérique en intervalles. Une approche bayésienne permet de définir

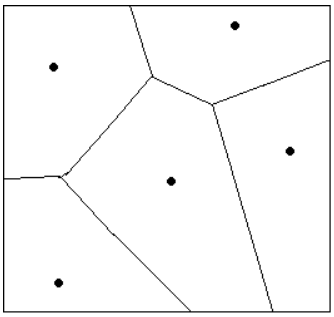

FIG. 1 – *Exemples de partitions de Voronoi pour la métrique euclidienne.*

un critère s'interprétant comme la probabilité que le modèle explique la variable cible catégorielle. La sélection du modèle le plus probable conduit à une méthode de discrétisation d'une variable statique numérique. La probabilité du modèle le plus probable s'utilise alors comme un indicateur de pertinence de la variable descriptive relativement à la variable cible.

Dans Ferrandiz et Boullé (2006b), l'approche est adaptée afin de traiter le cas où l'espace de représentation est muni d'une métrique. A l'aide de la métrique, une partition de Voronoi est associée à tout sous-ensemble d'individu (*c.f.* fig.1 pour des exemples). Le partitionnement d'une variable en intervalles est ainsi généralisé en un partitionnement de l'espace en cellules. L'ensemble des partitions obtenues constitue l'ensemble des modèles. La probabilité qu'un modèle explique la variable cible catégorielle 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 de la métrique. La méthode est évaluée dans Ferrandiz et Boullé (2006a) en tant que méthode de sélection d'instances pour la classification par le plus proche voisin.

Soyons plus formels. L'ensemble des individus est noté $\mathcal{I}$. Pour une métrique δ, tout sous-ensemble H de $\mathcal{I}$ définit une partition de $\mathcal{I}$, dite de Voronoi : chaque individu est associé au plus proche élément de H au sens de δ. Notons $\mathcal{H}_\delta(\mathcal{I})$ l'ensemble des partitions de Voronoi.

Si on note Y la variable cible catégorielle, nous disposons d'un critère $c_{\delta,Y} : \mathcal{H}_\delta(\mathcal{I}) \rightarrow \mathbb{R}$ qui associe à chaque partition la probabilité que cette partition explique la cible Y. Nous proposons alors d'évaluer la métrique δ par la probabilité du modèle le plus probable :

$$c^*(\delta) = \max_{H \in \mathcal{H}_\delta(\mathcal{I})} c_{\delta,Y}(H).$$

La fonction c^* fournit une évaluation supervisée de la qualité de la métrique δ et permet ainsi de comparer différentes métriques et, à travers elles, différents espaces de représentation et différentes variables. Pour une métrique δ donnée, on applique un algorithme d'optimisation combinatoire et on attribue la valeur rencontrée optimale du critère $c_{\delta,Y}$. Le critère $c_{\delta,Y}$ et l'heuristique d'optimisation proposés dans Ferrandiz et Boullé (2006a) et Ferrandiz et Boullé (2006b) sont décrits plus en détail dans la prochaine section.

Le critère $c_{\delta,Y}$ est non paramétrique et régularisé. Il quantifie le compromis entre le nombre de groupes de la partition et la discrimination de la cible, ce qui correspond à un compromis

entre complexité du modèle et ajustement du modèle aux données de l'échantillon. La régularisation est un moyen d'endiguer le phénomène de sur-apprentissage et d'assurer ainsi la fiabilité de la décision. Etant non paramétrique, l'évaluation se passe de validation ou de validation croisée. On dispose ainsi de plus d'individus pour ajuster le modèle, ce qui augmente sa qualité.

Afin de travailler avec un indicateur normalisé, nous considérons la transformation suivante de c^* :

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

où $c_0(\delta)$ est la valeur du critère $c_{\delta,Y}$ pour le modèle constitué par un seul groupe. D'après les travaux de Shannon (1948), l'opposé du logarithme d'une probabilité s'interprète comme une longueur de codage. L'indicateur $g^*(\delta)$ s'interprète alors comme un gain de compression. Il est supérieur à 0 et inférieur à 1. Si $g^*(\delta) = 0$, la métrique δ n'apporte aucune information sur la variable cible. Plus la valeur de $g^*(\delta)$ est proche de 1, plus les classes cibles sont séparées, et plus la métrique δ est pertinente.

4 Sélection d'instances : critère et algorithme

L'évaluation de la qualité d'une métrique introduite ci-dessus repose sur la recherche de la meilleure partition de Voronoi induite par un sous-ensemble de l'échantillon. Pour la convenance du lecteur, nous décrivons dans cette section le critère et l'heuristique d'optimisation utilisés, déjà proposés et étudiés dans Ferrandiz et Boullé (2006a) et Ferrandiz et Boullé (2006b).

4.1 Evaluation bayésienne d'une partition

Posons les notations. Soit $\mathcal{I}$ un ensemble de N individus. Soit $Y : \mathcal{I} \to \mathbb{L}$ une variable cible catégorielle, $\mathbb{L}$ étant un alphabet de taille J. Soit $X : \mathcal{I} \to \mathbb{X}$ une variable descriptive, l'espace de représentation $\mathbb{X}$ étant muni d'une métrique δ.

Soit H un ensemble de K individus. La partition de Voronoi $V(H) = (V(k))_{k \in H}$ associée à H est définie par :

$$\forall k \in H,\ V(k) = \left\{ i \in \mathcal{I};\ k = \arg\min_{k' \in H} \delta(X(i), X(k')) \right\}.$$

Pour $k \in H$, la *cellule de Voronoi* $V(k)$ contient les individus i dont k est l'élément de H le plus proche, relativement à δ. L'élément k est appelé *prototype* de la cellule $V(k)$. La fig.1 donne des exemples de telles partitions.

Si les éléments de H sont indexés de 1 à K, N_k ($1 \leq k \leq K$) est le nombre d'individus dans la cellule du k^{eme} élément de H et N_{kj} désigne le nombre de tels individus de la j^{eme} classe. Ainsi, $N = N_1 + \cdots + N_K$ et $N_k = N_{k1} + \cdots + N_{kJ}$.

L'approche adoptée dans Ferrandiz et Boullé (2006b) conduit à évaluer H par :

$$-\log c_{\delta,Y}(H) = \log N + \log \binom{N+K-1}{K} + \sum_{k=1}^{K} \log \binom{N_k+J-1}{J-1} + \sum_{k=1}^{K} \log \frac{N_k!}{N_{k1}! \ldots N_{kJ}!}.$$

Le premier terme quantifie la probabilité d'apparition du nombre K de cellules de $V(H)$, le second terme quantifie la probabilité d'apparition des K prototypes de H, et les derniers termes quantifient cellule par cellule la probabilité d'apparition de la variable cible. Ils résultent de l'adoption de l'approche bayésienne de l'évaluation : la somme des deux premiers termes correspond à l'a priori sur les modèles et la somme des deux derniers termes à la vraisemblance de la cible.

La dernière somme, d'après la formule de Stirling $\log x! \approx x \log x - x + O(\log x)$, se comporte asymptotiquement comme N fois l'entropie conditionnelle de la variable cible Y en connaissance de la partition :

$$\sum_{k=1}^{K} \log \frac{N_k!}{N_{k1}! \ldots N_{kJ}!} \approx -N \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{N_{kj}}{N} \log \frac{N_{kj}}{N_k}.$$

Intuitivement, le critère quantifie la discrimination des distributions par un terme entropique et la pondère par un coût structurel mesurant la complexité de la partition. Pour cela, il prend en compte diverses caractéristiques, comme le nombre de groupes, la répartition des instances dans les groupes (*i.e* les coefficients N_k), la répartition des instances dans les classes (*i.e.* les coefficients N_{kj}).

4.2 Heuristique d'optimisation

Nous disposons d'un critère d'évaluation des ensembles de prototypes H. Il reste à proposer un algorithme d'optimisation. Nous reprenons celui introduit dans Ferrandiz et Boullé (2006a), qui consiste à encapsuler une optimisation gloutonne d'un ensemble de prototypes dans une méta-heuristique. La complexité algorithmique de l'optimisation gloutonne est réduite en exploitant les propriétés du critère et des partitions de Voronoi. La méta-heuristique permet de remettre en question le résultat de la recherche gloutonne afin d'optimiser encore un peu plus le critère.

L'heuristique gloutonne Glouton(H) s'applique à tout ensemble H de K prototypes. Elle comporte K étapes. A chaque étape, tout sous-ensemble obtenu par élimination d'un prototype est évalué. Parmi ceux-ci, celui minimisant la valeur du critère est déclaré vainqueur. Cette étape gloutonne est itérée et appliquée à chaque vainqueur successif, jusqu'à évaluation d'un singleton. Le meilleur sous-ensemble rencontré est retourné.

Cette méthode considère $O(K^2)$ sous-ensembles et chaque évaluation nécessite la recherche du plus proche prototype pour chaque instance. Une implantation directe de Glouton(H) possède une complexité en $O(NK^3)$. L'exploitation des propriétés du critère et des partitions de Voronoi conduit à une implantation de complexité un $O(NK \log K)$ nécessitant un espace mémoire en $O(NK)$.

L'heuristique gloutonne effectue rapidement un grand nombre d'évaluations. Il est naturel d'envisager une application répétée mais limitée de cet algorithme. Pour cela, la méta-heuristique de recherche à voisinage variable est adoptée (Hansen et Mladenovic (2001)). Elle consiste à appliquer l'heuristique de base (*i.e.* l'algorithme Glouton) à un modèle proche de la solution considérée. Si la nouvelle solution n'est pas meilleure, on considère un voisinage plus grand. Sinon, la méta-heuristique repart de la nouvelle meilleure solution avec une taille de voisinage minimale. Ce procédé est contrôlé par une taille maximale du voisinage à explorer.

5 Classification de profils de consommation

Nous illustrons les apports de notre méthode par des expérimentations 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 sur l'échantillon est uniforme. On dispose de 168 variables descriptives numériques, chacune mesurant la consommation téléphonique sur une tranche horaire de la semaine. Nous répartissons uniformément les 3516 individus de l'échantillon entre un ensemble d'apprentissage (75% des individus) et un ensemble de test (25% des individus), de manière stratifiée (*i.e.* en respectant la distribution a priori des classes cibles).

5.1 Evaluation d'une variable séquentielle

La méthode d'évaluation d'une métrique proposée dans cet article, en plus de quantifier la pertinence d'une métrique, fournit un support explicatif : la partition la plus probable. Ainsi, on dispose d'une distribution des classes cibles et d'un prototype pour chaque groupe, ce qui autorise une explication du résultat purement numérique.

Nous illustrons cet aspect en appliquant notre méthode à la variable séquentielle constituée par les 168 variables descriptives du problème de classification de profils de consommation, l'espace de représentation étant muni de la métrique L_1. Autrement dit, chaque individu se voit associer une suite de 168 mesures de consommation et la distance entre deux profils est mesurée à l'aide de la métrique L_1.

L'évaluation de la métrique fournit un gain de compression de 0.051, ce qui est très faible et caractérise un fort mélange des classes cibles. Mais il n'est pas nul et la méthode partitionne les individus en 7 groupes. Les distributions relatives à chacun des groupes sont représentées par des histogrammes groupés sur cette la fig.2. En calculant la valeur moyenne de chacune des 168 variables dans chaque groupe, on obtient 7 profils de consommation caractéristiques. Trois de ces profils sont reportés sur la fig.2, ainsi que le profil moyen de consommation (*i.e.* celui calculé sur tout l'ensemble d'apprentissage).

Le résultat étant visualisable, il est facilement interprétable. Par exemple, on voit que les individus du groupe 7 sont en grand nombre (35% de l'échantillon d'apprentissage), qu'ils ont 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 individus). 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.

Afin de vérifier de visu la fiabilité du découpage effectué, nous calculons les distributions en test (reportées sur la fig.2). Bien que l'ensemble de test soit trois fois plus petit que l'ensemble d'apprentissage, on constate que la distribution des individus dans les groupes est stable ($(4\%, 5\%, 6\%, 11\%, 19\%, 21\%, 35\%)$ en apprentissage et $(3\%, 4\%, 6\%, 12\%, 17\%, 24\%, 34\%)$ en test), que les classes majoritaires dans chaque groupe en test sont les mêmes que celles observées en apprentissage, etc.

Notre méthode n'est pas la première à fournir de telles informations. Ainsi, l'analyse discriminante, linéaire ou quadratique (Hastie et al., 2001), ou toute méthode supervisée construisant

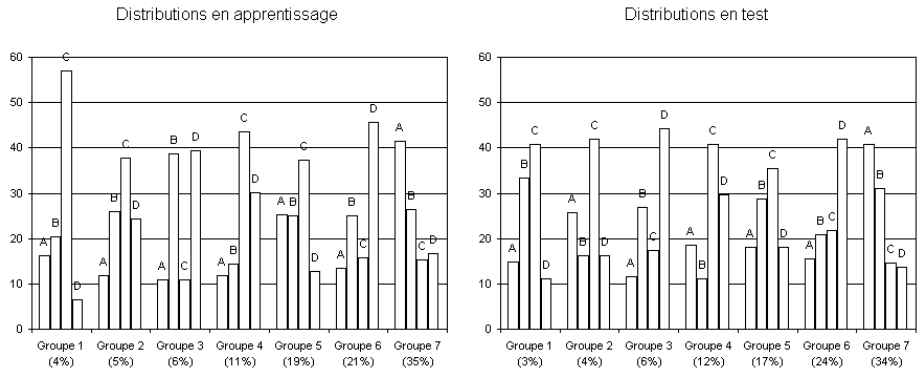

FIG. 2 – *Distributions des étiquettes dans chaque groupe, en apprentissage et en test. Le support de chaque groupe est reporté en abscisse. Par exemple, le groupe 5 contient* 19% *des individus de l'ensemble d'apprentissage,* 17% *des individus de l'ensemble de test, et ses éléments se répartissent dans les classes cibles suivant la distribution* $(25\%, 25\%, 37\%, 13\%)$ *en apprentissage et* $(18\%, 29\%, 35\%, 18\%)$ *en test.*

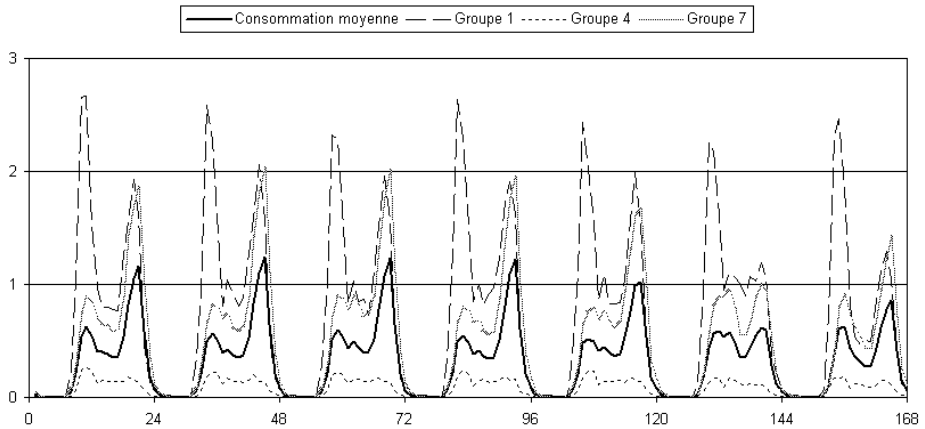

FIG. 3 – *Consommation moyenne sur l'ensemble d'apprentissage et consommations moyennes dans chacun des groupes 1, 4 et 7. Les individus du groupe 1 correspondent à de très fortes consommations, avec des pics très marqués. Ceux du groupe 7 correspondent aux faibles consommations.*

des prototypes, comme les méthodes de quantification (Kohonen, 2001), conduisent à de telles visualisations. Mais c'est la seule à évaluer strictement les informations qui sont visualisées : le calcul du gain de compression prend en compte la répartition des individus dans les groupes, les répartitions des individus dans les classes cibles groupe par groupe. La visualisation n'en est que plus adaptée.

L'analyse discriminante suppose les individus d'une classe toutes générées par une même gaussienne. Les modèles considérés sont tous de même capacité, ce qui se traduit en pratique par un nombre de groupes égal au nombre J de classes. L'exemple étudié ici montre que contraindre la capacité revient à limiter la richesse de l'information extraite. De toute façon, les paramètres des J gaussiennes sont ajustés en maximisant la vraisemblance complète et la vraisemblance ne tient pas compte des différences de capacité : on ne peut utiliser la mesure de vraisemblance pour comparer un modèle d'analyse discriminante linéaire avec un modèle d'analyse discriminante quadratique, et encore moins avec un modèle d'analyse de mélange (qui autorise un nombre quelconque de gaussiennes par classe, *c.f.* Hastie et al. (2001)). On est ramené à appliquer un second critère. En pratique, c'est le risque empirique qu'on utilise, avec ses limites.

Les techniques de quantification, hautement paramétriques, nécessitent entre autre de fixer le nombre de prototypes. Le choix d'un "bon" nombre de prototypes repose donc sur un critère alternatif. En pratique, là encore, on estime le risque empirique. L'idée sous-jacente aux techniques de quantification étant de repousser les prototypes en cas de mauvais étiquetage et de les rapprocher dans le cas contraire, la présence de prototypes "morts" à la fin de l'optimisation constitue de plus un effet secondaire peu désirable. En effet, de nombreux prototypes sont déplacés au point de ne plus être sollicités par la suite. En terme de partition de Voronoi associée, cela signifie que plusieurs cellules finales ne contiennent aucun élément de l'ensemble d'apprentissage. Le résultat perd de sa pertinence et la visualisation associée est rendue caduque.

S'il est usuel de mettre de côté un ensemble d'individus, dit de validation, pour ajuster certains paramètres ou contrôler la fiabilité de l'estimation, c'est inutile lorsqu'on utilise notre méthode. Elle est en effet non paramétrique et la fiabilité est intrinsèquement assurée par l'usage d'un critère régularisé. Tous les individus servent à la prise décision, ce qui profite nécessairement à la qualité de celle-ci.

5.2 Comparaison et sélection de variables séquentielles

La définition d'une variable séquentielle nécessite la définition d'un espace de représentation. Celui-ci est souvent muni d'une métrique. Nous avons proposé dans ce qui précède une méthode d'évaluation supervisée de la pertinence d'une métrique. A travers le prisme d'une métrique, nous sommes donc en mesure d'évaluer la pertinence d'une variable séquentielle par le gain de compression que mesure notre méthode.

Nous illustrons son utilité par un problème de sélection de variables séquentielles. Pour cela, nous considérons les 24 variables séquentielles définies par les tranches horaires du problème de classification de profils de consommation. Chaque variable est composée de 7 mesures, correspondant à la consommation sur chaque jour de la semaine pour une tranche horaire fixée. Pour chacune, nous munissons l'espace de représentation de la métrique L_1 et appliquons notre méthode d'évaluation sur l'ensemble d'apprentissage. Le gain de compression et le taux de bonne prédiction en test sont reportés pour comparaison sur la fig.4.

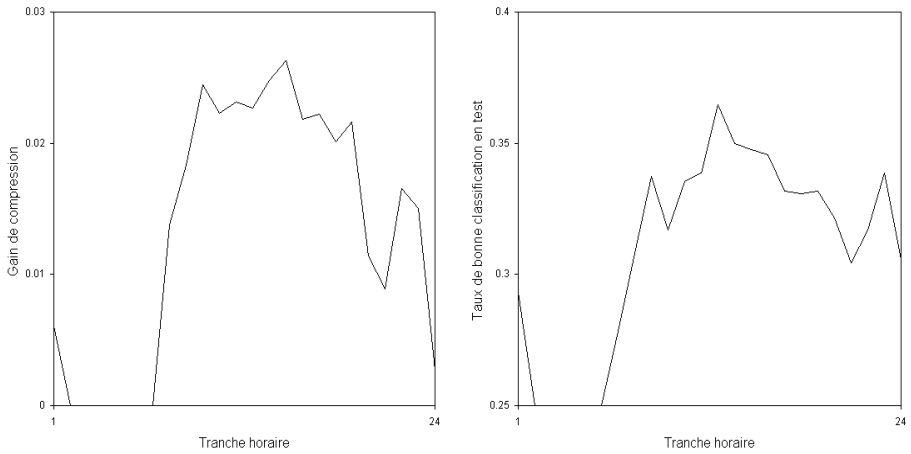

FIG. 4 – *Gain de compression et taux de bonne prédiction en test. Le gain de compression mesure la répartition des classes cibles dans les cellules de la partition, le taux de bonne prédiction mesure le caractère majoritaire d'une classe. Un bon gain de compression implique un bon taux de bonne prédiction.*

Utilisé pour de la sélection, le gain de compression conduit à choisir la tranche horaire 14, ou toute variable dont le gain de compression mesuré dépasse un certain seuil fixé a priori par l'analyste. A l'opposé, le gain de compression est nul pour les tranches horaires de fin de nuit. Ceci signifie qu'un seul groupe est constitué et que les classes cibles sont mélangées. Considérées isolément, ces variables ne sont d'aucun intérêt. C'est la présence d'une régularisation, qui consiste à contrôler la discrimination opérée par la capacité de la partition, couplée avec le fait que les partitions considérées fournissent des capacités allant d'un minimum (un seul groupe) à un maximum (autant de groupes que d'instances), qui rend possible une telle conclusion.

5.3 Sélection d'une métrique

Dans l'expérience précédente, nous avons utilisé la métrique L_1 pour mesurer la distance séparant deux profils. Nous reproduisons cette expérience et considérons deux métriques supplémentaires : la métrique euclidienne et un noyau gaussien (définissant une métrique euclidienne dans un espace implicite). Les courbes de gain de compression sont reportées sur la fig.5.

Certains comportements des courbes sont analogues. Par exemple, quelle que soit la métrique ici considérée, les tranches horaires de fin de nuit sont déclarées non pertinentes relativement à la cible. Mais c'est l'utilisation de la métrique L_1 qui conduit aux meilleurs gains de compression, quasiment pour toutes les tranches horaires. Pour ces variables séquentielles et cette cible, l'analyste est conduit automatiquement et de manière fiable à choisir cette métrique au détriment des deux autres. S'il dispose de temps, il peut même s'aider de la visualisation proposée précédemment pour expliquer les différences de comportement sur chaque tranche horaire.

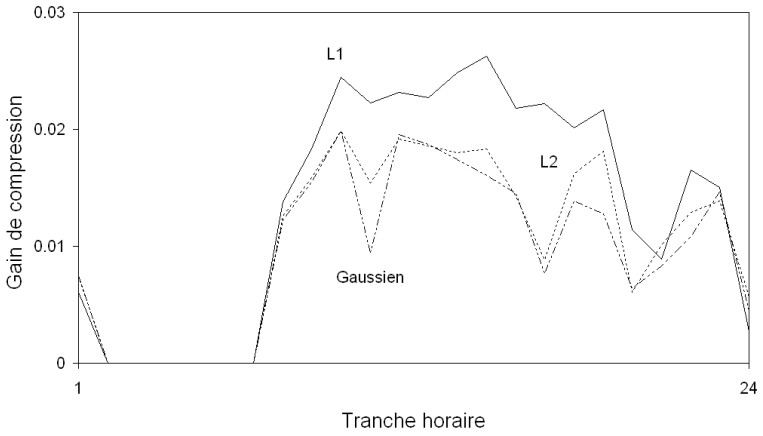

FIG. 5 – *Gain de compression pour chaque tranche horaire et pour trois métriques. La métrique L_1 est la plus pertinente des trois.*

6 Conclusion

En fouille de données, dès lors que la récolte des données n'est pas orientée dans le sens de l'analyse, un travail de préparation est à mener. Une table doit d'abord être construite pour ensuite procéder à une modélisation statistique qui réponde à la question posée par le propriétaire des données. A priori, de nombreuses variables sont susceptibles d'expliquer le phénomène étudié et il s'agit d'inclure dans la table les plus pertinentes d'entre elles.

L'approche adoptée en préparation est l'approche filtre univariée, indépendante d'un modèle particulier et plus à même de faire face à un nombre élevé de variables. Dans ce cadre, la qualité de la méthode d'évaluation utilisée pour juger de l'intérêt d'une variable est cruciale. En exploitant des travaux antérieurs, nous avons ici proposé une méthode automatique et fiable d'évaluation d'une métrique, dans le cas de la classification supervisée. Nous avons illustré son apport sur un problème réel de classification de profils de consommation téléphonique.

Cet exemple d'application montre l'apport de notre méthode en préparation de données. L'analyste dispose grâce à elle d'un outil pour mener à bien la sélection filtre univariée des variables qu'il supposent a priori pertinentes. Cet outil permet d'évaluer la pertinence a posteriori (après observation des données) de variables dont l'espace de représentation est muni d'une métrique. Cette évaluation est automatique, fiable et se passe d'un ensemble de validation. Plus généralement, l'outil permet de sélectionner la métrique la plus adaptée.

Références

Blum, A. et P. Langley (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence 97*(1-2), 245–271.

Boullé, M. (2006). MODL : a bayes optimal discretization method for continuous attributes. *Machine learning 65*(1), 131–165.

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 (2006a). Sélection supervisée d'instances : une approche descriptive. In *Actes de la conférence sur l'extraction et la gestion des connaissances*, Volume 2, pp. 421–432.

Ferrandiz, S. et M. Boullé (2006b). Supervised evaluation of Voronoi partitions. *Journal of intelligent data analysis 10*(3), 269–284.

Guyon, I. et A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of machine learning research 3*, 1157–1182.

Hansen, P. et N. Mladenovic (2001). Variable neighborhood search : principles and applications. *European journal of operational research 130*, 449–467.

Hastie, T., R. Tibshirani, et J. Friedman (2001). *The elements of statistical learning*. Springer.

Kohavi, R. et G. John (1997). Wrappers for feature selection. *Artificial intelligence 97*(1-2), 273–324.

Kohonen, T. (2001). *Self-organizing maps*. Springer.

Shannon, C. (1948). A mathematical theory of communication. Technical report, Bell systems technical journal.

Summary

In data mining, the statistical analysis is secondary. A great amount of work is to be done during the preparation step in order to build a statistical representation from a computer representation. In this paper, we propose a new method of selection of a representation by addressing the problem of the evaluation of a metric and applying a former work.

In the supervised context, we make use of a regularized and non parametric classification method to evaluate the relevance of a metric. The evaluation is thus automatic and reliable. Experiments are carried out on a real dataset, in order to illustrate the properties of the method. The task aims at classifying curves summarizing call detailed records. We illustrate the reliability and the comprehensibility of the resulting decisions as well.