Méthodes clusterwise pour données complexes

Ndèye Niang, Giorgio Russolillo

30 mars 2021

Préambule

1. Université d'inscription : Conservatoire national des arts et métiers
2. Ecole doctorale : SMI
3. Laboratoire d'accueil : CEDRIC
4. Encadrement :
    1. Directeur de thèse : Ndèye Niang
    2. Co-directeur de thèse : Giorgio Russolillo

Cadre général

Le clustering, aussi appelé classification non-supervisée, consiste à identifier des groupes d'individus partageant des caractéristiques similaires dans une population qui elle est hétérogène. Par exemple, dans le domaine du marketing, on peut avoir une population de clients dans laquelle il existe des groupes inconnus de clients avec un profil d'achat similaire. Un clustering permettra de les identifier afin de leur proposer une campagne de publicité adaptée. Au-delà du marketing, l'hétérogénéité liée à structure en groupes des individus est rencontrée dans un grand nombre d'applications, par exemple en sciences sociales, sciences de l'éducation, environnement, études cliniques, ou plus généralement dans le domaine scientifique. Les méthodes de clustering permettent dans tous ces domaines de mieux comprendre la population en la résumant à la description des différents groupes qui la composent. Elles peuvent aussi être utilisées comme méthodes de pré-traitement afin de constituer des groupes homogènes d'individus sur lesquels il est ensuite possible d'effectuer des analyses. En définitive, elles font partie des méthodes fondamentales de la science des données pour gérer l'hétérogénéité des observations.

Les méthodes de clustering restent néanmoins des méthodes limitées pour gérer l'hétérogénéité dans des problématiques de régression. En effet, le clustering étant effectué indépendamment de l'objectif de régression, il n'y a pas de garanties sur son caractère optimal vis-à-vis du critère de régression. Les méthodes de régression typologique ou clusterwise permettent d'apporter une réponse à ce problème travers la recherche simultanée d'une partition des données en classes et du modèle de régression local associé à ces classes. Les premiers travaux relatifs à ces méthodes sont attribués à Späth (1979) selon DeSarbo and Cron (1988). Mais on peut aussi citer ceux de Bock (1969) et Diday (1976) puis ceux de Charles (1977). DeSarbo and Cron (1988) proposent une méthode de régression linéaire typologique fondée sur un modèle de mélange de gaussiennes avec des estimateurs du maximum de vraisemblance et l'algorithme EM.

Une des limites des méthodes clusterwise est qu'elles sont rapidement sur-paramétrées dès lors que le nombre de classes est grand devant le nombre de variables. En effet, les modèles de régression locaux ne peuvent alors plus être ajustés. Différentes stratégies ont été envisagées pour gérer cette difficulté. Preda and Saporta (2005) ont notamment

utilisé des méthodes PLS dans le cadre spécifique de données fonctionnelles, tandis que Suk and Hwang (2010), Bougeard et al. (2018) ont envisager des approches clusterwise parcimonieuses.

Une autre difficulté, inhérente à l'application sur données réelles, est la gestion des données incomplètes. Les données manquantes posent un vrai problème pour l'analyse car les méthodes de clusterwise ne sont pas prévues pour être appliquées sur ce type de données. Si dans certains contextes on peut parfois se limiter à une suppression des individus incomplets, cela devient impossible dès lors que le nombre de variables est grand, ne serait-ce que parce qu'aucun individu n'est alors complet.

Parmi les solutions offertes pour gérer ce problème, les techniques d'imputation multiple sont sûrement les plus populaires (Rubin (1987)). Le principe est de compléter plusieurs fois le tableau incomplet selon un modèle (dit modèle d'imputation), puis d'ajuster sur chacun de ces tableaux imputés le modèle souhaité (dit modèle d'analyse) et enfin d'agréger les résultats selon des règles bien définies. Toutefois, définir un modèle d'imputation n'est pas trivial, celui-ci doit ajuster correctement l'ensemble des données alors que celles si sont dans ce cas hétérogènes.

Ainsi, le cadre général de cette thèse est d'étendre les méthodes clusterwise à un contexte plus large qui est celui de la grande dimension et des données incomplètes.

Objectifs de la thèse

Méthodes clusterwise sur données incomplètes

Parmi les méthodes d'imputation multiple, les approches dites séquentielles (van Buuren (2018)) sont très populaires. Pour un tableau donné, elles consistent à imputer chacune des variables selon un modèle de régression. Les variables sont imputées tour à tour, fournissant ainsi un jeu de données complété sans avoir eu à définir explicitement la distribution de l'ensemble des variables. Cette opération est répétée plusieurs fois de façon à produire plusieurs tableaux imputés.

Les modèles utilisés pour imputer chacune des variables ne sont généralement pas adaptés à l'imputation d'individus structurés en groupe inconnus. Pour parvenir au développement d'une méthode d'imputation séquentielle dédiée aux données hétérogènes, nous proposons naturellement de nous appuyer sur les méthodes de régression clusterwise.

Le premier objectif de cette thèse sera donc de gérer les données manquantes en régression clusterwise en proposant une méthode d'imputation multiple basée elle-même sur les méthodes clusterwise.

Méthodes clusterwise en grande dimension

Les méthodes clusterwise peuvent être rapidement limitées en grande dimension. La difficulté ici étant que les modèles locaux sont rapidement sur-paramétrés quand le nombre de variables est grand. Il s'agira ici de proposer une approche clusterwise régularisée afin de gérer la grande dimension. On pourra notamment s'appuyer sur des approches PLS telles que proposées dans (Chun et al., 2010).

Méthodes clusterwise sur données incomplètes en grande dimension

Sur la base développements précédents, la suite consistera à proposer une nouvelle méthode d'imputation multiple séquentielle gérant à la fois l'hétérogénéité et la grande dimension, offrant ainsi une solution pour l'application des méthodes clusterwise aux données complexes.

Application

Tous les apports méthodologiques précédemment présentés seront systématiquement évalués par simulation en confrontant les méthodes proposées à l'état de l'art le plus récent. Par ailleurs, ils se concluront par une mise en application sur des données réelles provenant de l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (ANSES).

Echancier

Le travail de thèse se déroulera selon les étapes suivantes

- Etape 1 (3 mois) : effectuer une étude bibliographique détaillée sur des méthodes de régression typologique dont le but est de dégager des axes de recherches permettant l'extension de la régression clusterwise à la grande dimension ainsi qu'aux données incomplètes.
- Etape 2 (9 mois) : développer des méthodes clusterwise parcimonieuses et les évaluer. Publication
- Etape 3 (9 mois) : développer une nouvelle méthode d'imputation multiple séquentielle pour des hétérogènes et l'appliquer à la gestion des données manquantes en régression clusterwise. Publication.
- Etape 4 (9 mois) : sur la base des travaux précédents, développer une méthodologie pour appliquer les méthodes clusterwise sur des données à la fois incomplètes et en grande dimension. Publication

Les 6 derniers mois seront consacrés à la rédaction de la thèse.

Moyens consacrés

Les moyens tant matériels (informatiques) qu'humains (compétences propres) du laboratoire d'accueil et des autres partenaires (ANSES et CEDRIC/CNAM) seront mis à disposition du doctorant.

Références

Bock, HH. 1969. "The Equivalence of Two Extremal Problems and Its Application to the Iterative Classification of Multivariate Data." *Mathematisches Forschungsinstitut*, 10.

Bougeard, Stéphanie, Véronique Cariou, Gilbert Saporta, and Ndèye Niang. 2018. "Prediction for regularized clusterwise multiblock regression." *Applied Stochastic Models in Business and Industry* 34 (6): 852–67. doi:10.1002/asmb.2335.

Charles, Christian. 1977. "Régression Typologique et Reconnaissance Des Formes." PhD thesis, Université Paris IX.

Chun, H. and Keles, S. 2010. “Sparse partial least squares regression for simultaneous dimension reduction and variable selection”. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, 72, pp.3–25.

DeSarbo, Wayne S, and William L Cron. 1988. “A Maximum Likelihood Methodology for Clusterwise Linear Regression.” *Journal of Classification* 5 (2). Springer: 249–82.

Diday, E. 1976. “Classification et Sélection de Paramètres Sous Contraintes.” *Rapport de Recherche IRIA-LABORIA*, no. 188.

Preda, Cristian, and Gilbert Saporta. 2005. “Clusterwise PLS regression on a stochastic process.” *Computational Statistics and Data Analysis* 49: 99–108. doi:10.1016/j.csda.2004.05.002.

Rubin, D. 1987. *Multiple Imputation for Non-Response in Survey*. New-York: Wiley.

Späth, H. 1979. “Clusterwise Linear Regression.” *Computing* 22: 367–73.

Suk, Hye Won, and Heungsun Hwang. 2010. “Regularized Fuzzy Clusterwise Ridge Regression.” *Advances in Data Analysis and Classification* 4 (1). Springer: 35–51.

van Buuren, S. 2018. *Flexible Imputation of Missing Data (Chapman & Hall/Crc Interdisciplinary Statistics)*. Hardcover; Chapman; Hall/CRC.