Classification de variables : une approche dynamique en grande dimension

Christian Derquenne

Electricité de France - Recherche et Développement - 7, boulevard Gaspard Monge - 91120 Palaiseau - christian.derquenne@edf.fr

Résumé. La recherche de structures dans les données représente une aide essentielle pour comprendre les phénomènes à analyser. Les méthodes de classification de variables permettent de répondre à cette problématique, mais elles peuvent être pénalisées par un trop grand nombre de variables. Nous proposons une nouvelle approche de type "Diviser pour Régner" fondée sur le principe MapReduce pour pallier ce problème. La table de données est divisée en plusieurs sous-tableaux traités en parallèle, puis réconciliés à l'aide de l'Analyse des Correspondances Multiples. Cette approche est appliquée sur des données simulées et fournit de très bons résultats.

Mots-clés. Classification, grande dimension, MapReduce, apprentissage non supervisé.

Abstract. The search for structures in the data represents an essential help to understand the phenomena to be analyzed. The methods of clustering of numerical variables make it possible to answer this problem, but they can be penalized by too many variables. We propose a new "Divide and Conquer" approach based on the MapReduce principle to overcome this problem. The data table is divided into several sub-tables processed in parallel, then reconciled using the Multiple Correspondence Analysis. This approach is applied to simulated data and provides very good results.

Keywords. Clustering, high dimension, MapReduce, unsupervised learning.

1 Contexte - objectif

La recherche exploratoire de structures dans les données est essentielle dans de nombreuses applications (biologie, environnement, finance, management de l'énergie, ...) afin de comprendre les comportements des individus, les liens entre les variables, ... Les outils de visualisation, de réduction de dimension, de recherche de patterns permettent de répondre efficacement à ce type de problématiques. Nous nous plaçons dans cadre de classification non supervisée et plus particulièrement dans le domaine de la classification de variables numériques ayant des liens quelconques (linéaires ou non linéaires). Plusieurs approches ont été proposées pour répondre à cette problématique. Les principales reposent sur la réduction de l'espace factoriel en associant au mieux les variables initiales à de nouvelles composantes (Sarle, 1990, Vigneau et al., 2003, Chavent et al., 2011, Bühlmann et al., 2013, Chen M., 2014, Chen Y. et al., 2016). Nous avons développé une méthode nommée "double critère contrôlé dynamique" disponible pour des liens linéaires (Derquenne, 2016). Celle-ci est fondée simultanément sur un test d'indépendance linéaire simple entre les variables initiales et/ou des variables latentes (première composante principale de l'ACP) et un test d'unidimensionnalité sur les classes obtenues afin de construire une typologie de façon dynamique au moyen du contrôle du nombre de groupes et de leur qualité.

Cette méthode a été étendue pour des relations quelconques (Derquenne, 2017). Cette approche est fondée sur des transformations polynomiales entre couple de variables initiales et/ou variables latentes (première composante principale issue d'une ACP non linéaire). Les résultats obtenus à l'aide de ces deux approches sont très satisfaisants comparées à d'autres méthodes qui peinent à détecter le "bon" nombre de classes et le "bon" contenu de celles-ci validés à partir de tests sur des données simulées.

En effet, les méthodes existantes sont plus ou moins performantes, en termes de qualité de la typologie obtenue (compacité, isolation, "bon" nombre de classes et "bon" contenu des groupes). Il en est de même pour le temps de calcul et la taille mémoire requise. Pour des nombres raisonnables d'observations et de variables ($n < 10000$ et $p < 100$), les algorithmes fonctionnent bien sur ces deux aspects. Par contre, si n et p deviennent très grands, alors le temps de calcul et la capacité mémoire deviennent trop élevés. Pour pallier ce problème, une stratégie du type "diviser pour mieux régner" peut alors être adéquate pour traiter cet aspect "grande dimension". Nous posons tout d'abord deux postulats et nous proposons une méthode pour résoudre ce problème. Puis, nous appliquons celle-ci sur des jeux de données simulées en grande dimension afin d'évaluer ses performances. Enfin, nous concluons sur les améliorations à apporter, les applications potentielles et les voies futures.

2 Problème, postulat et proposition

Comme indiqué dans la section précédente, une trop grande taille de la table des données (nombre de variables et nombre d'individus) peut affecter la performance des algorithmes et donc pénaliser la qualité des résultats obtenus. L'objectif de ce papier est de proposer une approche statistique afin de répondre à la question suivante : "Comment tenir compte de la grande dimension lorsque que nous appliquons une méthode statistique classique ?". Pour cela, nous posons deux postulats.

Postulat 1 : Si une méthode fournit de bons résultats sur un échantillon tiré de la table de données entière, il n'y a pas de raison que cette méthode donne de mauvais résultats sur un autre échantillon tiré de la même grande base de données.

Postulat 2 : Si nous combinons les résultats d'un nombre d'échantillons issus de la grande table de données à l'aide d'un processus adéquat, alors les résultats agrégés devraient être comparables au résultat global provenant de l'application de la méthode sur l'ensemble de la base de données.

Ce processus est fondé sur le principe : "Diviser et Conquérir" (DCP). Le principe général est le suivant : (i) la table de données entière est découpée en S échantillons ; (ii) chaque échantillon est traité en parallèle à l'aide de la méthode choisie ; (iii) les résultats sont combinés et ils sont traités selon une procédure adéquate ; (iv) le résultat final est obtenu. Ce principe est fondé sur l'approche **MapReduce** qui est une méthode générique pour traiter des bases de données massives distribuées sur de nombreux fichiers systèmes. Elle a été développée par Google^{TM} (Dean et al., 2004). Après avoir présenté brièvement la démarche de la méthode de classification de variables introduite en 2016 et 2017, nous développerons l'approche DCP associée à celle-ci.

2.1 Une approche dynamique pour la classification de variables

Soient $X_1, ..., X_q$, q variables numériques dont on suppose que les relations sont linéaires ou absentes, alors la première étape consiste à agréger les deux variables les plus corrélées linéairement pour constituer la première classe. Pour cela, on fixe un seuil critique du test de corrélation (par exemple, $\alpha_\rho = 0,05$), alors si la plus petite p-valeur parmi les $q(q-1)/2$ couples de variables est inférieure à α_ρ, on regroupera ces deux variables. Puis la première composante principale est calculée sur celles-ci, soit Z_1. De nouvelles corrélations sont calculées entre Z_1 et les $q-2$ variables restantes. Trois cas peuvent se présenter : soit une classe de trois variables, soit deux classes de deux variables, soit aucune corrélation significative est trouvée, alors l'algorithme s'arrête. Dans ce dernier cas, il y aura un groupe de deux variables et $q-2$ classes singleton. Si le processus continue, dès qu'un groupe possède au moins trois variables, un test d'unidimensionnalité est pratiqué sur la deuxième valeur propre, tel que $H_0 : \lambda_2 \leq 1$ (Saporta, 1999). Si l'hypothèse nulle d'unidimensionnalité est rejetée, alors on recherche si parmi les p-valeurs restantes issues des tests de corrélations, la plus petite est inférieure au seuil fixé. Si c'est le cas, les trois possibilités indiquées précédemment se représenteront. Le processus de constitution des classes se poursuit jusqu'à ce que plus aucune p-valeur de corrélation est inférieure à α_ρ et que le test d'unidimensionnalité pour chaque classe est rejeté. A la fin de ce processus, nous obtenons M classes. Cette approche a été généralisée pour des liens quelconques entre variables (Derquenne, 2017).

2.2 Extension de la classification de variables en grande dimension

Nous nous plaçons dans le contexte suivant. Soient $X_1, ..., X_q$, q variables numériques, telles que $X_j \in \mathbb{R}^n$ où $n \gg 10000$ est le nombre d'individus contenus dans la grande base de données E.

Soient $E_1, ..., E_s, ..., E_S$, S échantillons aléatoires tirés sans remise de n individus, tel que $E = \cup_{s=1}^{S} E_s$ où $\text{card}(E_s) = n_s$ et $\sum_{s=1}^{S} = n$.

Chaque échantillon E_s est découpé en L sous-échantillons aléatoires sans remise de variables tirés parmi les p variables initiales, tels que $Q_1, ..., Q_l, ..., Q_L$ où $\text{card}(Q_l) = p_l$ et $\sum_{l=1}^{L} = p$.

Remarque : Le découpage en L échantillons aléatoires de variables peut être différent ou non pour chaque échantillon d'individus : E_s.

Enfin, T_{sl} correspond à la sous-table de données de n_s individus et de p_l variables.

Le processus détaillé DCP adapté à la classification de variables se déroule de la façon suivante.

(i) Sur chaque sous-ensemble de données T_{sl}, l'approche dynamique de classification de variables est appliquée, alors M_{sl} classes sont obtenues, ainsi que M_{sl} premières composantes principales associées : $Y_1^{(sl)}, ...Y_{M_{sl}}^{(sl)}$. Cette étape (i) est réalisée en parallèle sur chaque T_{sl} de E_s. Il s'agit de la phase **MAP**. Cela permet d'obtenir : $Y_1^{(s_1)}, ..., Y_{M_{s_1}}^{(s_1)}, ..., Y_1^{(s_l)}, ..., Y_{M_{s_l}}^{(s_l)}, ..., Y_1^{(s_L)}, ..., Y_{M_{s_L}}^{(s_L)}$ premières composantes principales pour l'échantillon E_s.

(ii) Sur les premières composantes principales obtenues précédemment, l'approche dynamique de classification de variables est appliquée, et fournit alors M_s nouveaux groupes : $(C_1^{(s)}, ..., C_k^{(s)}, ..., C_{M_s}^{(s)})$

et M_s nouvelles premières composantes principales associées : $Z_1^{(s)}, ..., Z_k^{(s)}, ..., Z_{M_s}^{(s)}$. Cette nouvelle classification permet à chaque variable initiale de départ X_j d'appartenir à une classe $C_k^{(s)}$ parmi les M_s classes construites précédemment. Ces attributions de variables à des classes permettent de construire une nouvelle variable V_s contenant pour chaque variable X_j le numéro de sa classe parmi les M_s groupes. Cette étape (ii) est réalisée en parallèle sur chaque E_s. Il s'agit de la phase **REDUCE**.

(iii) L'objectif de cette étape et la suivante est de réconcilier l'ensemble les résultats issus de (i) et (ii) afin d'obtenir une classification globale pour l'ensemble des S échantillons E_s d'individus et l'ensemble des p variables X_j. En effet, lorsque les étapes (i) et (ii) sont terminées, nous disposons de S variables catégorielles : $V_1, ..., V_s, ..., V_S$ contenant les numéros de classes de chaque variable X_j. Si, comme on le suppose, les résultats de classification issus de chaque échantillon E_s d'individus se ressemblent alors les comportements devraient être similaires. En d'autres termes, les variables $V_1, ..., V_s, ..., V_S$ devraient être dépendantes. Un moyen de mesurer cette dépendance est d'appliquer une analyse des correspondances multiples (ACM) sur ces variables pour lesquelles les individus sont simplement les variables initiales $X_1, ..., X_j, ..., X_q$. Les résultats de l'ACM fournissent R composantes principales $U_1, ..., U_r, ...U_R$. S'il y a une structure de groupes de variables X_j dans les données, alors cela devraient se retrouver dans l'espace des individus (les variables initiales) de l'ACM.

(iv) Cette ultime étape permet de voir s'il y a ou pas une structure de groupes. Pour cela, nous classifions les individus (les variables X_j) à partir des composantes principales de l'ACM. Toutes celles-ci peuvent être retenues ou il est possible de sélectionner celles qui rassemblent par exemple, 80% de l'inertie expliquée. Par ailleurs, nous avons choisi d'utiliser une approche de classification hiérarchique ascendante au moyen du critère de Ward. Les résultats obtenus fournissent M classes : $G_1, ..., G_m, ..., G_M$ contenant respectivement $p_1, ..., p_m, ..., p_M$ variables initiales $X_1, ..., X_j, ..., X_q$, tel que $\sum_{m=1}^{M} p_m = p$ et contenant les n individus.

2.3 Evaluation de la classification en grande dimension

Les résultats obtenus à l'aide de l'approche proposée précédemment doivent être évalués. Pour cela, nous utilisons deux niveaux de validation.

Le premier niveau évalue la qualité de reconstitution de la classification observée sur l'ensemble de la table des données. Pour cela, nous comparons l'inertie expliquée de la classification observée (OCI) de l'équation (1) et l'inertie expliquée de la classification estimée (ECI) de l'équation (2) à l'aide de l'approche proposée.

$$OCI = \frac{1}{p} \sum_{m=1}^{\tilde{M}} \sum_{X_j \in G_m} \rho^2(X_j, \tilde{Z}_m) \quad (1) \qquad ECI = \frac{1}{p} \sum_{m=1}^{M} \sum_{X_j \in \tilde{G}_m} \rho^2(X_j, Z_m) \quad (2)$$

où $\tilde{Z}_m$ et Z_m sont les $\tilde{M}$, respectivement les M premières composantes principales observées et estimées. Généralement ECI est inférieure à OCI. Plus elle proche, plus la qualité de reconstitution de la classification observée est bonne.

Le second niveau permet d'évaluer la qualité de la classification estimée. Pour cela nous comparons les contenus des typologies observée et estimée à partir de leur tableau de contigence dont les lignes et les colonnes correspondent aux numéros des classes. Les indices de Rand, de

Jaccard, γ, le T de Tchuprow, le V de Cramer et le pourcentage de bien classés permettent d'évaluer la qualité de la typologie estimée. Ces indices varient entre 0 et 1, plus la valeur obtenue est proche de l'unité, plus l'adéquation est bonne.

3 Application de l'approche de classification de variables en grande dimension et comparaisons

Afin d'évaluer la qualité de l'approche proposée, nous avons simulé un jeu de données possédant 100000 observations et 1000 variables. Celui-ci est découpé en 9 classes (6 avec une seule variable chacune $X_2, X_3, X_4, X_6, X_7, X_8$, les 3 autres avec, respectivement, 301, 492 et 201 variables), tels que : $X_j = X_1 + 2\epsilon_t$ pour $j = 501, ..., 800$ où $X_1 \rightsquigarrow \mathcal{N}(0,1)$; $X_j = 2X_9 + 2\epsilon_t$ pour $j = 10, ..., 500$ où $X_9 \rightsquigarrow \mathcal{N}(0,1)$; $X_j = 0,1X_5 + \epsilon_t$ pour $j = 801, 1000$ où $X_5 \rightsquigarrow \mathcal{N}(0,1)$ et $\epsilon_t \rightsquigarrow \mathcal{N}(0,1)$.

La construction des sous-tables de variables et d'individus se déroule de la façon suivante. 10 échantillons issus de tirage sans remise E_s de 10000 individus sont constitués, puis chacun d'eux est découpé en 10 sous-échantillons Q_l de variables de taille 100 sont également tirés au hasard sans remise. Signalons que chaque Q_l pour $l = 1, 10$ possède les mêmes variables (cf. remarque de 2.2). Par conséquent, le processus **MapReduce** calcule 100 classifications initiales en parallèle, puis 10 classifications globales en parallèle, une ACM et une classification finale. Cela correspond à 112 tâches.

Le tableau de contigence (table 1) croise les typologies observée (en colonne) et estimée (en ligne). Les classes 1 et 2 de la classification estimée regroupent exactement les classes 1 et 2 de la typologie observée. Il en est de même pour la classe 3 calculée qui correspond à la classe 4 observée. Les classes 4 et 6 regroupent des variables séparées à l'origine. Les classes 5 et 7 sont relatives à la classe 3 observée. Par ailleurs, les résultats des indices sont satisfaisants. En effet, ECI est très légèrement inférieure à OCI : 0,3177 *vs* 0,3226. Les indices d'adéquation confirment ce bon résultat : $T = 0,6576$, $V = 0,7593$, $\gamma = 0,9867$, $Rand = 0,9936$ et $Jaccard = 0,9829$. Enfin, le pourcentage de bien classés vaut 79,4%.

Observée	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Estimée	$(X_1, X_{501}, ..., X_{800})$	$(X_9, ..., X_{500})$	$(X_5, X_{801}, ..., X_{1000})$	(X_2)	(X_3)	(X_4)	(X_6)	(X_7)	(X_8)
1	301	0	0	0	0	0	0	0	0
2	0	492	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	1	1	0	0
5	0	0	185	0	1	0	0	0	0
6	0	0	1	0	0	0	0	0	1
7	0	0	15	0	0	0	0	1	0

Table 1: Tableau de contingence des typologies observée et estimée

En termes de performance, nous divisons par 6, la CPU pour un ordinateur à deux coeurs seulement, entre l'application de la méthode classification dynamique de variables sur l'ensemble

du jeu de données et l'application de la méthode à grande dimension proposée. Ce qui produit un gain de 200%.

4 Apports, applications et voies futures

La méthode en grande dimension pour la classification de variables est fondée sur l'approche dynamique qui offrait déjà une bonne qualité de résultats (2016, 2017). L'extension de l'approche dynamique à la grande dimension préserve l'inertie globale et le contenu des classes (études de simulation). L'approche dynamique pour grande dimension fournit de bonnes performances en termes de temps de calcul, même dans le cas de processus séquentiel. Des fonctions R ont été développées pour l'approche dynamique de classification de variables en grande dimension. D'autres applications sur données simulées et réelles ont fourni des résultats de même bonne qualité (prix spot avec commodités, demandes résiduelles pour classifier les zones géographiques, données statiques). Les améliorations et voies futures sont les suivantes : intégration d'autres méthodes de classification (CLV, ClustOfVar, ...) en DCP, comparaison avec des méthodes de classification en grande dimension, plus de simulations à l'aide d'ordinateurs plus puissants, traitements sur des données plus complexes et développement d'une méthode de co-clustering pour des données en grande dimension.

Bibliographie

[1] Bühlmann P., Rütimann P., van de Geer S., and Zhang C-H, (2013): Correlated in regression: Clustering and sparse estimation. *Journal of Stat. Planning and Inference*, **143**(11), 1835-1858.
[2] Chavent M., Kuentz V., Liquet B. et Saracco J., (2011): Classication de variables : le package ClustOfVar, *43ièmes Journées de Statistique*, Tunis, Tunisie.
[3] Chen M., (2014): *Classification de variables autour de variables latentes avec filtrage de l'information : application à des données en grande dimension*, Thèse de doctorat, Université de Nantes, Ecole VENAM.
[4] Chen Y. and Yang U., (2016): A Novel Information-Theoretic Approach for Variable Clustering and Predictive Modeling Using Dirichlet Process Mixtures, www.nature.com/scientificreports.
[5] Dean, J. and Ghemawat, S., (2004): MapReduce: simplified data processing on large clusters. In Proceedings of Sixth Symposium on Operating System Design and Implementation.
[5] Derquenne Ch., (2016): Classification de variables : une approche à double critères contrôlés dynamiques, *48ièmes Journées de Statistique*, Montpellier, France.
[6] Derquenne Ch., (2017): Classification de variables avec des relations non linéaires, *49ièmes Journées de Statistique*, Avignon, France.
[7] Saporta G., (1999): Some Simple Rules for interpreting Outputs of Principal Components and Correspondence Analysis, *IXth International Symposium on ASMDA*, Lisbon, Portugal.
[8] Sarle W., (1990): *The VARCLUS Procedure. SAS/STAT 9.2 User's Guide.* Cary, NC: SAS Institute, Inc. **93**, 7453-7484.
[9] Vigneau E. and Qannari E.M., (2003): Clustering of Variables Around Latent Components. *Communications in Statistics - Simulation and Computation*, **32**(4), 1131-1150.