Réduction de dimension par partition de variables en estimation de densité pour des problèmes de dimension modérée et application à la segmentation de données biologiques

Louis Pujol, M. Glisse, P. Massart
Université Paris-Saclay, Inria Saclay, Université Paris-Saclay

Email : louis.pujol@universite-paris-saclay.fr

Mots Clés : Classification automatique, estimation de densité, réduction de dimension

Biographie – Titulaire du master 2 mathématique de l'aléatoire, spécialité statistique et machine learning de l'université Paris-Saclay, obtenu en 2019, j'ai commencé une thèse en novembre 2019 sous la direction de Pascal Massart (université Paris-Saclay) et de Marc Glisse (Inria) et en collaboration avec une entreprise évoluant dans le domaine de la biotechnologie, Metafora Biosystems. Cette thèse est financé par le DIM MathInnov via le projet Paris Region PhD^2.

Resumé :

Les progrès techniques intervenus ces dernières décennies dans le domaine des biotechnologies ont permis le développement de techniques d'acquisition de données de plus en plus sophistiquées. Ces données représentent potentiellement une mine d'information pour le biologiste mais leur complexité rend nécessaire le développement d'outils spécifiques afin d'exploiter pleinement leur potentiel.

La classification non supervisée, ou clustering est une technique d'importance primordiale pour la compréhension fine de données complexes. Elle permet de segmenter un jeu de données en différentes classes homogènes. De nombreuses méthodes ont été proposées pour effectuer en pratique cette classification. Nous nous appuyons ici une méthode basée sur des techniques d'analyse topologique des données [1]. Cette technique à l'avantage d'être hiérarchique et donc de délivrer une information riche à l'utilisateur et est basée sur une première étape d'estimation de densité. C'est cette étape d'estimation de densité qui nous intéresse d'un point de vue mathématique.

L'estimation de densité est un problème classique en statistique non paramétrique. De manière générale on se pose la question de notre capacité à correctement approcher une fonction de densité inconnue à partir d'un échantillon de données indépendantes et identiquement distribuées tirées selon cette densité.

Ce problème a été bien étudié du point de vue minimax et il ressort que la difficulté d'estimation vient de deux paramètres : la régularité de la densité à estimer et la dimension de l'espace sur lequel elle est définie. La dépendance en la dimension rend le problème difficile dès lors que l'on dépasse la dimension 3 ou 4 pour des de jeux de données de taille inférieure au million.

Ce problème est connu sous le nom de fléau de la dimension. Différentes techniques de réduction de dimension sont envisageables pour le dépasser. Nous présenterons l'approche développée par Lepski [2] et étudiée par Rebelles [4] consistant à proposer une hypothèse d'indépendance par blocs entre les variables observées. Cette approche consiste à partitionner l'ensemble des variables et à calculer un estimateur de la densité comme produit d'estimateurs de densité marginales sur chacun des blocs de la partition.

Cette approche présente une vertu théorique majeure : si la partition considérée coïncide avec une décomposition en composantes indépendantes de la vraie densité alors la difficulté du problème n'est plus liée à la dimension ambiante mais à la taille du plus gros bloc de la partition. Cela suggère en pratique de se limiter à des partitions dont les sous-ensembles ont une taille maximale

k fixée en fonction du nombre de données à disposition. La mise en œuvre pratique d'une méthode permettant de simultanément sélectionner la partition et d'estimer la densité n'est pas évidente à cause de la combinatoire du nombre de partitions, même en se limitant à des sous-ensembles de taille au plus k.

Nous présenterons une méthodologie permettant d'effectuer l'estimation de densité via sélection de partition en temps raisonnable pour des jeux de données de dimension modérées (de l'ordre de quelques dizaines). Alors que les travaux cités étaient concentrés sur l'analyse en perte $L_p, 1 \leq p \leq \infty$, nous utilisons une perte de Kullback-Leibler et son pendant empirique, la log-vraisemblance empirique, pour estimer partition des variables et densité.

L'avantage majeur de cette reformulation est de bénéficier du bon comportement du logarithme vis-à-vis d'une structure produit.

Nous montrerons comment nous tirons bénéfice de cette propriété pour réécrire le problème de sélection de partition comme un problème d'optimisation linéaire entière pour lequel des algorithmes de résolution efficaces existent. Nous mettrons en avant le gain en temps de calcul par rapport à une approche naïve.

Nous montrerons également une inégalité oracle pour la sélection de partition. De manière classique, le logarithme du nombre de modèles apparaît comme une constante dans une telle borne [3]. Dans notre cas nous mettrons en avant le fait que notre hypothèse structurelle nous permet de voir apparaître le logarithme du nombre de sous-ensemble impliqués dans les partitions considérées, de l'ordre de $k \log d$ plutôt que le logarithme du nombre de partitions lui-même, d'ordre supérieur à $(d/2) \log d$ et ainsi de réduire l'impact de la dimension sur la précision de l'inégalité oracle.

Références

[1] Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):1–38, 2013.

[2] Oleg Lepski et al. Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. *Annals of Statistics*, 41(2):1005–1034, 2013.

[3] Pascal Massart. Concentration inequalities and model selection. 2007.

[4] Gilles Rebelles et al. Lp adaptive estimation of an anisotropic density under independence hypothesis. *Electronic journal of statistics*, 9(1):106–134, 2015.