Journées SMAI MODE 2018

Clustering avec sélection de variables par minimisation alternée et application à la biologie

Cyprien GILET
I3S, Université Côte d'Azur, CNRS gilet@i3s.unice.fr

Marie DEPREZ
IPMC, Université Côte d'Azur, CNRS deprez@ipmc.cnrs.fr

Jean-Baptiste CAILLAU
LJAD, Université Côte d'Azur, CNRS/Inria, Sophia-Antipolis, caillau@unice.fr

Michel BARLAUD
I3S, Université Côte d'Azur & CNRS (barlaud@i3s.unice.fr)

Résumé. Ce travail traite de classification non supervisée avec sélection de variables. Le problème est d'estimer à la fois les labels ainsi qu'une matrice creuse de poids pour réduire la dimension des données. Pour traiter ce problème combinatoire non convexe tout en maintenant le contrôle de la parcimonie de la matrice de poids, nous proposons une minimisation alternée de la norme de Frobenius de notre critère. L'algorithme qui en résulte, "k-sparse", alterne k-means et gradient projeté. L'étape de gradient projeté est de type splitting, avec une projection exacte sur la boule ℓ^1 pour assurer la parcimonie. La convergence de l'étape de gradient projeté est établie. La norme de Frobenius de notre critère converge lorsque le nombre d'itérations de notre algorithme tend vers l'infini. Plusieurs tests sur des bases de données Single Cell RNA-seq montrent que notre méthode améliore de façon significative les résultats d'autres méthodes de clustering [3, 4]. La complexité de notre algorithme est linéaire avec le nombre d'observations (cellules), de sorte que la méthode est applicable sur de grandes bases de données.

Mots-clefs : Clustering, gradient projeté, optimisation alternée, sélection de variables.

En grande dimension le clustering est peu efficace : à mesure que la dimension augmente les vecteurs deviennent indiscernables et le pouvoir prédictif des méthodes mentionnées ci-dessus est considérablement réduit. Le clustering est devenu un outil très important pour l'analyse de données génomiques et de nouvelles machines RNA-seq permettent aujourd'hui de mesurer l'expression de 20000 gènes pour un grand nombre de cellules. Dans le but de réduire la dimension, une approche populaire est de projeter les données dans un espace de plus petite dimension avant de réaliser le clustering. Des méthodes telles que la *Principal Component Analysis (PCA)*, la *Linear Discriminant Analysis (LDA)*, le *Spectral Clustering* ou les *Kernel methods* sont souvent utilisées pour projeter les données dans un espace de dimension réduite mais ne permettent pas d'écarter les variables considérées comme bruit (qui peuvent être source d'erreurs). Une approche populaire pour sélectionner les variables pertinentes est la méthode *Least Absolute Shrinkage and Selection Operator* (LASSO). Cependant le paramètre utilisé dans la formulation LASSO pour promouvoir la parcimonie n'est pas simple à estimer.

Plutôt qu'une pénalité, nous proposons de définir une contrainte en norme ℓ^1 et tirons parti de l'existence d'une projection exacte ℓ^1 [2].

Notre approche peut être décrite comme suit [1]. Soit $X \in \mathbb{R}^{m\times d}$ la base de données à analyser contenant l'expression de d variables pour m observations. Nous définissons $Y \in \{0,1\}^{m\times k}$ (où k représente le nombre de clusters) la matrice de présence telle que :

$$y_{ij} = \begin{cases} 1 & \text{si l'observation } i \text{ appartient au cluster } j \\ 0 & \text{sinon} \end{cases} \qquad \text{et} \qquad \sum_{j=1}^{k} y_{ij} = 1\ , \quad i = 1,\dots,m \tag{1}$$

Nous définissons de plus $W \in \mathbb{R}^{d\times \bar{d}}$ matrice de projection dans un espace de dimension $\bar{d} \ll d$, et $\mu \in \mathbb{R}^{k\times \bar{d}}$ matrice des coordonnées des centroïdes dans l'espace projeté XW :

$$\mu(j,:) := \frac{1}{\sum_{i=1}^{m} y_{ij}} \sum_{i \text{ t.q. } y_{ij}=1} (XW)(i,:)\ , \qquad j = 1,\dots,k. \tag{2}$$

Notre approche cherche à minimiser le critère suivant :

$$\min_{W,Y} \ \frac{1}{2}\|Y\mu - XW\|_F^2 \quad \text{t.q} \quad \|W\|_1 \leq \eta \tag{3}$$

Pour cela nous cherchons à résoudre les problèmes **1** et **2** de façon alternée.

Problème 1 *Pour Y donnée, $\eta > 0$, nous cherchons à résoudre le critère* (3) *en W :*

$$\min_{W} \ \frac{1}{2}\|Y\mu - XW\|_F^2 \quad \text{t.q} \quad \|W\|_1 \leq \eta$$

Pour cette étape nous utilisons un algorithme de gradient-projeté avec une projection exacte sur la boule ℓ^1 [2]*. Cette étape a pour objectif d'attribuer du poids seulement aux variables pertinentes pour discriminer les clusters.*

Problème 2 *Pour W donnée, nous cherchons à résoudre le critère* (3) *en Y :*

$$\min_{Y} \ \frac{1}{2}\|Y\mu - XW\|_F^2$$

Après avoir mis à jour la matrice de projection W (et donc écarté une partie des variables considérées comme bruit), cette étape a pour objectif de reconstruire les clusters de sorte que l'on minimise l'inertie intra-classe dans le nouvel espace projeté XW. Pour cela nous utilisons la fonction k-means dans l'espace XW.

Algorithm 1 Algorithme k-sparse clustering.

Input: $X, Y_0, \mu_0, W_0, L, N, k, \gamma, \eta$
$Y \leftarrow Y_0$, $\mu \leftarrow \mu_0$, $W \leftarrow W_0$
for $l = 0, \ldots, L$ **do**
  for $n = 0, \ldots, N$ **do**
    $V \leftarrow W - \gamma X^T(XW - Y\mu)$
    $W \leftarrow P_\eta^1(V)$
  end for
  $Y \leftarrow \texttt{kmeans}(XW, k)$
  $\mu \leftarrow \texttt{centroids}(Y, XW)$
end for
Output: Y, W

Proposition 1 *La norme de Frobenius de notre critère* (3) *converge lorsque le nombre d'itérations dans l'algorithme k-sparse tend vers l'infini.*

Pour évaluer les performances de notre algorithme k-sparse, nous l'avons testé sur quatre bases de données génomiques réelles [3] pour lesquelles la vérité terrain (les classes) était connue. Nous avons comparé nos résultats avec d'autres méthodes de clustering [3, 4] et montrons que notre approche améliore significativement les résultats de celles-ci en termes de précision, de silhouette et de temps de calcul.

Références

[1] C. Gilet, M. Deprez, J.-B. Caillau and M. Barlaud. *Clustering with feature selection using alternating minimization. Application to computational biology*. arXiv:1711.02974v2.

[2] L. Condat. *Fast projection onto the simplex and the ℓ^1 ball*. Math. Program. A, **158** (2016), no. 1, 575–585.

[3] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti and S. Batzoglou. *Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning*. Nature Methods **14** (2017), no. 4, 414–416.

[4] D. M. Witten and R. Tibshirani. *A framework for feature selection in clustering*. J. Am. Stat. Assoc. **105** (2010), no. 490, 713–726.

[5] P. L. Combettes and J.-C. Pesquet. *Proximal splitting methods in signal processing*. In Fixed-point algorithms for inverse problems in science and engineering, 185–212, Springer, 2011.

[6] P.-L. Lions and B. Mercier. *Splitting algorithms for the sum of two nonlinear operators*. SIAM J. Numer. Analysis, **16** (1979), no. 6, 964–979.