Efficacité de classification de la méthode des k-moyennes tronquées

Christel Ruwet [1]

[1] *Université de Liège - Département de mathématiques*
Grande Traverse, 12 B-4000 Liège, Belgium
cruwet@ulg.ac.be

Résumé. La méthode des k-moyennes est utilisée en classification afin de regrouper les observations les plus similaires en k groupes. Lorsqu'un second échantillon est disponible pour tester la qualité des regroupements ainsi obtenus, le taux de mauvaise classification peut être calculé. Si les échantillons proviennent d'un mélange de deux distributions homogènes et à symétrie sphérique, alors le taux de mauvaise classification atteint celui obtenu avec la règle de Bayes. Cela étant, la méthode des k-moyennes est optimale sous ce modèle de mélange. Elle n'est cependant pas robuste aux points aberrants qui pourraient se trouver dans l'échantillon servant à construire les groupements. Pour enrayer ce problème, le méthode des k-moyennes a été adaptée de diverses façons. Cette présentation s'intéresse à la méthode des k-moyennes tronquées définie en écartant une certaine proportion des observations. L'avantage de cette méthode, outre sa résistance aux points aberrants, est que le caractère optimal de la classification obtenue est conservé. Il est cependant bien connu que la suppression d'une partie des observations conduit à une perte d'efficacité de classification. Celle-ci peut être mesurée à l'aide de la fonction d'influence du taux de mauvaise classification.

Mots-clés. Classification, Efficacité, Fonction d'influence, Optimalité, Robustesse

Abstract. The k-means method is used in classification to group similar observations in k groups. When a second sample is available to test the obtained groupings, the rate of misclassification can be computed. If the samples are generated from a mixture of two homoscedastic and spherically symmetric distributions, the rate of misclassification equals that of the Bayes rule. Therefore, the k-means method is optimal under such a mixture model. However, it is not robust with respect to outliers in the dataset used to construct the groups. To avoid this problem, the k-means procedure has been adapted in many ways. This presentation focuses on the trimmed k-means method defined by trimming some of the observations. The advantage of this method, besides its resistance to outliers, is that optimality is preserved. However, it is well known that trimming observations leads to a loss in classification efficiency. The latter can be measured by means of the influence function of the misclassifiation rate.

Keywords. Classification, Efficiency, Influence function, Optimality, Robustness

1 Définitions et résultats existants

La méthode des k-moyennes est une méthode de classification qui permet de mettre au jour une éventuelle structure de groupes dans un ensemble de données. Cette méthode n'est pas récente (les premiers articles traitant d'aspects théoriques remontent aux années cinquante, avec notamment les travaux de Cox, 1957, et de Fisher, 1958) mais elle est toujours d'actualité. Bock (2007) en propose d'ailleurs une revue de la littérature. Cette méthode se base sur le critère des moindres carrés. Pour un nombre fixé k, elle cherche à rassembler les observations les plus similaires autour de k centres $\{T_1, \ldots, T_k\} \subset \mathbb{R}^d$ (aussi appelés les *k-moyennes*) définis, au niveau de la population, par les fonctionnelles statistiques $T_j : F \mapsto T_j(F)$, $j = 1, \ldots, k$ telles que

$$\{T_1(F), \ldots, T_k(F)\} = \arg\min_{t_1,\ldots,t_k} \int \left(\inf_{1\leq j\leq k} \|x - t_j\|^2 \right) dF(x). \tag{1}$$

Une fois les centres définis, les groupes peuvent être construits. Le j-ème groupe rassemble les observations plus proches du j-ème centre que des autres centres. En terme de fonctionnelle statistique, il est donc défini par

$$R_j(F) = \left\{ x \in \mathbb{R}^d : (T_j(F) - T_l(F))^t x - \frac{1}{2}\left(\|T_j(F)\|^2 - \|T_l(F)\|^2 \right) \geq 0 \, \forall \, l \neq j \right\} \tag{2}$$

pour $j = 1, \ldots, k$. Les observations à la limite entre deux groupes sont assignées au groupe avec le plus petit indice (par convention).

Quand la classification est basée sur un modèle ("model-based clustering"), il est habituellement supposé qu'une méthode de classification est appliquée à une distribution F qui est un modèle de mélange, c'est-à-dire que $F = \sum_{j=1}^{k} \pi_j(F) F_j$, où chaque composante du modèle de mélange représente une sous-population dénotée par $P_j, j = 1, \ldots, k$. Ces ensembles P_j peuvent être caractérisés par une variable aléatoire latente, Y, donnant les appartenances, c'est-à-dire $P_j = \{x : Y(x) = j\}$. Avec cette notation, $\pi_j(F)$ est la probabilité que X appartienne à P_j et F_j est la distribution conditionnelle de X sous P_j. Dans ce cas, il semble naturel d'espérer que les groupes construits par la méthode de classification reflètent les différentes sous-populations. Pour s'en assurer, il est possible de calculer la probabilité de mauvaise classification définie par

$$\mathrm{ER}(F) = \sum_{j=1}^{k} \pi_j(F) \mathbb{P}_F[X \notin R_j(F) | X \in P_j]. \tag{3}$$

Dans la pratique, pour éviter de sous-estimer cette probabilité, il est courant d'utiliser deux échantillons différents: l'un pour construire les groupes (échantillon de travail) et l'autre pour vérifier la qualité de ces groupes (échantillon test). C'est pourquoi la probabilité de mauvaise classification est souvent définie en utilisant deux distributions, une

distribution de travail, F, et une distribution test, F_m, ce qui donne alors

$$\mathrm{ER}(F, F_m) = \sum_{j=1}^{k} \pi_j(F_m) \mathbb{P}_{F_m}[X \notin R_j(F) | X \in P_j]. \tag{4}$$

Ruwet et Haesbroeck (2011) ont démontré que la méthode des 2-moyennes (cas particulier pour $k = 2$) atteint la probabilité de mauvaise classification minimale (définie par la règle de classification de Bayes) lorsque le modèle F est un mélange balancé ($\pi_1 = \pi_2$) de deux distributions homogènes et à symétrie sphérique. Sous de tels modèles, elle donne ainsi un résultat aussi bon que d'autres méthodes qui tiennent compte des appartenances dans la construction des groupes, comme par exemple la discrimination linéaire de Fisher.

Cependant, comme toute procédure basée sur la somme des carrés de toutes les observations, cette procédure n'est pas résistante aux points aberrants pouvant affecter l'échantillon. Cela a été formellement montré par García-Escudero et Gordaliza (1999) qui ont calculé les fonctions d'influence des 2-moyennes ainsi que leur point de rupture. La fonction d'influence d'une fonctionnelle statistique T est définie par

$$\mathrm{IF}(x; T, F) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon \Delta_x) - T(F)}{\varepsilon} \tag{5}$$

où Δ_x représente la distribution de Dirac en x (Hampel *et al.*, 1986). Cette fonction mesure l'impact sur la fonctionnelle d'une contamination de masse infinitésimale située en x. Le point de rupture est quant-à-lui une mesure empirique, c'est-à-dire qu'il se calcule sur un échantillon et non plus sur une distribution. Il mesure la proportion de contamination nécessaire dans un échantillon pour que l'estimateur ne soit plus fiable. La définition formelle du point de rupture dépend du type d'estimateur considéré. Comme les k-moyennes sont des estimateurs de position, prenons l'exemple particulier d'un estimateur T_n de position. Dans ce cas, la fiabilité est définie à l'aide du biais et le point de rupture calculé sur l'échantillon X_n est défini comme

$$\varepsilon(T_n, X_n) = \min_{0 \leq m \leq n} \left\{ \frac{m}{n} : \max_{X_n^*} \|T_n(X_n) - T_n(X_n^*)\| = \infty \right\} \tag{6}$$

où X_n^* est un échantillon obtenu en modifiant m observations de X_n (Donoho et Huber, 1983). En ce qui concerne les 2-moyennes, García-Escudero et Gordaliza (1999) ont montré que les fonctions d'influence ont un comportement linéaire en x et que le point de rupture vaut $1/n$ quelque soit l'échantillon. Comme les fonctions d'influence des 2-moyennes sont non bornées, l'impact d'un point aberrant, si petit soit il par rapport à la taille de l'échantillon, peut être dramatique pour l'estimateur. De plus, comme le point de rupture est asymptotiquement nul, il suffit d'une observation contaminée pour détruire totalement l'estimation des 2-moyennes.

Pour remédier à cela, une idée est de ne pas baser la fonction objective du problème de minimisation (1) sur toutes les observations, c'est-à-dire de tronquer une partie des

données. La méthode des k-moyennes tronquées, introduite par Cuesta-Albertos *et al.* (1997) cherche donc un sous-ensemble X_α de l'échantillon X_n qui contiendrait $n - \lfloor n\alpha \rfloor$ observations et k centres $\{T_1, \ldots, T_k\} \subset \mathbb{R}^d$ de telle sorte à minimiser la somme des carrés des distances au centre le plus proche

$$\{T_1, \ldots, T_k\} = \min_{X_\alpha} \min_{t_1,\ldots,t_k} \sum_{x_i \in X_\alpha} \inf_{1 \leq j \leq k} \|x_i - t_j\|^2. \tag{7}$$

Dans leur article, García-Escudero et Gordaliza (1999) montrent que cette façon de faire mène à des centres pour lesquels les fonctions d'influence sont bornées et le point de rupture est strictement positif (pour autant que l'échantillon soit bien composé de deux sous-populations).

2 Nouveaux résultats

Dans ce travail, la fonction d'influence de la probabilité de mauvaise classification obtenue par la méthode des 2-moyennes tronquées sera calculée. On verra qu'elle hérite du caractère borné des fonctions d'influence des 2-moyennes tronquées. Par ailleurs, le caractère optimal (dans le sens de minimiser la probabilité de mauvaise classification) de cette procédure sera également prouvé.

Cela étant, il sera possible de comparer la méthode des k-moyennes tronquées à la méthode des k-moyennes à l'aide de leur efficacité de classification. Suivant l'approche déjà utilisée par Croux *et al.* (2008), le calcul de l'efficacité de classification sera basé sur la fonction d'influence du taux de mauvaise classification. Comme attendu, nous verrons que le gain de robustesse acquis grâce aux 2-moyennes tronquées entraîne malheureusement une perte d'efficacité de classification. Il s'agit là du prix à payer afin d'atteindre une certaine résistance envers les points aberrants. Cette perte d'efficacité augmente avec α et est, par exemple, toujours inférieur à 50% sous le modèle $F_m = 0.5\Phi(\mu_1, 1) + 0.5\Phi(\mu_2, 1)$ ($\Phi(\mu, \sigma^2)$ représente la fonction de répartition d'une loi normale de moyenne μ et de variance σ^2) lorsque $\alpha = 0.05$ et $\mu_1, \mu_2 \in \mathbb{R}$.

Ces résultats théoriques seront accompagnés de résultats de simulations comparant les taux de mauvaise classification des deux méthodes. Nous verrons ainsi que l'augmentation du taux mauvaise classification de la méthode des k-moyennes tronquées par rapport à la méthode classique est fort relative par rapport à la protection contre les points aberrants qu'elle permet d'atteindre.

Bibliographie

[1] Bock H.-H. (2007) Clustering methods: a history of k-means algorithms, *Selected contributions in data analysis and classification*, 161–172. Springer, Berlin.

[2] Cox D.R. (1957) Note on grouping, *Journal of the American Statistical Association* 52:543–547.
[3] Croux C., Filzmoser P., Joossens K. (2008) Classification efficiencies for robust linear discriminant analysis, *Statistica Sinica*, 18:581–599.
[4] Cuesta-Albertos J.A., Gordaliza A., Matrán C. (1997) Trimmed k-means: an attempt to robustify quantizers, *Annals of Statistics*, 25:553–576.
[5] Donoho D., Huber P.J. (1983) The notion of breakdown point. *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., 157–184.
[6] Fisher W.D. (1598) On grouping for maximum heterogeneity, *Journal of the American Statistical Association*, 53:789–798.
[7] García-Escudero L.A., Gordaliza A. (1999) Robustness properties of k means and trimmed k means, *Journal of the American Statistical Association*, 94:956–969.
[8] Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986) *Robust statistics. The approach based on influence functions*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.
[9] Ruwet C., Haesbroeck G. (2011) Classification performance resulting from a 2-means, *Submitted for publication.* Available at `http://hdl.handle.net/2268/81354`