Initialisation des k-moyennes à l'aide d'une décomposition supervisée des classes

Oumaima Alaoui Ismaili[*,**], Vincent Lemaire[*], Antoine Cornuéjols[**]

[*]Orange Labs, AV. Pierre Marzin 22307 Lannion cedex France
oumaima.alaouiismaili@orange.com,
vincent.lemaire@orange.com
[**]AgroParisTech 16, rue Claude Bernard 75005 Paris
antoine.cornuejols@agroparistech.fr

Résumé. Au cours des dernières années, les chercheurs ont concentré leur attention sur une nouvelle approche nommée *la classification à base de clustering*. Cette approche vise à décrire et à prédire le concept cible d'une manière simultanée. Motivé par l'importance de l'étape d'initialisation des centres pour les algorithmes basés sur le partitionnement (e.g les K-moyenne), cet article vise à tester à quel point une méthode d'initialisation supervisée pourrait aider l'algorithme des K-moyennes standard à remplir la tâche de la classification à base de clustering.

1 Introduction

Au cours de ces dernières décennies, on assiste à une augmentation significative du volume des données. L'innovation continue des techniques de stockage est l'un des principaux facteurs à l'origine de ce phénomène. Par exemple, les grandes entreprises comme Orange et Amazon récoltent et stockent durant leurs opérations quotidiennes, une avalanche de données concernant les comportements de leurs clients. Cependant, les données seules n'ont aucune valeur. Il existe une grande différence entre le fait de les obtenir, de les comprendre et de pouvoir les exploiter. Les connaissances utiles sont souvent cachées par la quantité des données et elles ne sont obtenues qu'à travers la compréhension de ces dernières. De ce fait, il existe un grand intérêt à développer des techniques permettant d'utiliser au mieux le gisement de données afin d'en extraire un maximum de connaissances utiles.

Dans la littérature, de nombreuses techniques d'analyse issues de diverses disciplines scientifiques (e.g Statistique, Intelligence Artificielle, Informatique) ont été proposées. Par exemple, l'analyse multivariée regroupe l'ensemble des méthodes statistiques qui s'attachent à l'observation et au traitement simultané de plusieurs variables en vue d'en dégager une information synthétique pertinente. Les deux grandes catégories de méthodes d'analyse statistique multivariées sont, d'une part, les méthodes dites *descriptives* et, d'autre part, les méthodes dites *prédictives*. Les méthodes descriptives ont pour objectif d'organiser, de simplifier et d'aider à comprendre les phénomènes existant dans un ensemble important de données. Cet ensemble est organisé en instances de plusieurs variables descriptives, dans lequel aucune des variables

n'a d'importance particulière par rapport aux autres. Parmi ces méthodes, on trouve le clustering (Aggarwal et Reddy (2013)) qui vise à trouver une typologie ou une répartition des individus en groupes distincts où les instances dans chaque groupe (ou cluster) doivent être les plus homogènes possible. Les méthodes prédictives ont, quant à elles, pour objectif de prévoir et d'expliquer à partir d'un ensemble de données étiquetées un ou plusieurs phénomènes observables. Parmi ces méthodes, on trouve la classification (Cornuéjols et Miclet (2010)) supervisée qui vise à prévoir l'appartenance des nouveaux individus à des classes prédéterminées.

Récemment, des chercheurs ont concentré leur attention sur l'étude d'une nouvelle méthode d'apprentissage connue sous le nom de *la classification à base de clustering* ou *Supervised clustering* en anglais (Al-Harbi et Rayward-Smith (2006); Eick et al. (2004)). Cette technique a pour objectif de *décrire* et de *prédire* le concept cible d'une manière simultanée. L'idée est donc découvrir la structure de la variable cible (si elle existe), puis muni de cette structure de pouvoir prédire l'appartenance au concept cible. Dans cette étude, nous nous intéressons à la classification à base des K-moyennes (Jain (2010)). La classification à base de clustering vise à modifier l'algorithme des K-moyennes afin qu'il soit un bon classifieur.

L'un des inconvénients de l'algorithme des K-moyennes standard réside dans sa sensibilité envers le choix des centres initiaux. En effet, l'étape d'initialisation influence la qualité de la solution trouvée ainsi que le temps d'exécution (Celebi et Kingravi (2014)). A partir de ce constat, il est naturel de se demander si l'utilisation d'une méthode d'initialisation supervisée peut aider l'algorithme des K-moyennes standard à obtenir de bons résultats (i.e avoir, dans la phase d'apprentissage, un bon compromis entre l'inertie intra clusters et la pureté des classes dans chaque cluster) suivant le principe de la classification à base de clustering. Le but du travail présenté dans cet article est donc de chercher à savoir si le fait d'introduire l'information contenue dans la variable cible dans l'étape d'initialisation de l'algorithme des K-moyennes peut aider cet algorithme à mieux remplir la tâche de la classification à base de clustering.

2 Proposition

L'intérêt d'utiliser une méthode d'initialisation supervisée peut être vu dans le cas de déséquilibre des classes à prédire. La probabilité de choisir plus qu'un centre dans la classe majoritaire et de ne choisir aucun centre dans la classe minoritaire est très élevée. Par conséquent, une détérioration au niveau de la pureté des classes dans les clusters serait introduite.

Dans cet article, nous proposons une nouvelle méthode supervisée d'initialisation des centres pour les K-moyennes. Cette méthode, appelée *S-Bisecting*, est basée d'une part sur le principe de la décomposition des classes (Ocegueda-Hernandez et Vilalta (2013)) où chacune des classes est traitée individuellement et d'autre part sur le Bisecting K-means (Steinbach et al. (2000)). Les différentes étape de cette méthode sont comme suit :

— Calculer le centre de gravité de chaque classe.
— Si le nombre de clusters K est égale au nombre de classes C, alors
  — **Sortie** : Les C centres de gravité.
— Si K est supérieure à C, alors

1. Diviser le cluster le plus dispersé au sens de l'inertie intra-cluster en deux.
2. Recalculer le centre de gravité de chaque sous cluster.
3. Répéter 1 et 2 jusqu'à atteindre le nombre de clusters désiré.

— **Sortie** : Les K centres de gravité.

Il est à signaler que la méthode utilisée pour diviser les classes dispersées est l'algorithme 2-moyennes.

3 Expérimentation

Pour évaluer le comportement de la méthode d'initialisation proposée, nous allons comparer les performances moyennes des résultats obtenus par l'algorithme des K-moyennes standard en utilisant d'une part des méthodes connues d'initialisation non supervisées (e.g la sélection aléatoire, Sample (Maitra et al. (2010)) et K-means++ (Arthur et Vassilvitskii (2007))) et d'une autre part la méthode proposée. Avant l'application de l'algorithme des K-moyennes, une discrétisation supervisée pour les variables continues et un groupage supervisé de modalités pour les variables catégorielles a été effectué (voir Alaoui Ismaili et al. (2015) pour plus de détail). Pour mesurer la qualité de ces résultats, deux points doivent pris en compte :

(1) La pureté des classes dans chaque clusters est mesurée dans cette étude par le critère ARI (Adjusted Rand Index) en train (e.g la figure 1) et en test (e.g la figure 2).

(2) L'homogénéité des instances dans chaque cluster est mesurée par le ratio Inertie Inter/Inertie Intra. La figure 3 donne à titre illustratif les performances moyennes des résultats obtenus par les k-moyennes en utilisant l'une des méthodes d'initialisation et en variant le nombre de clusters ($K \in \{2, 4, 8, 16, 32, 64\}$) sur le jeu de données Mushroom.

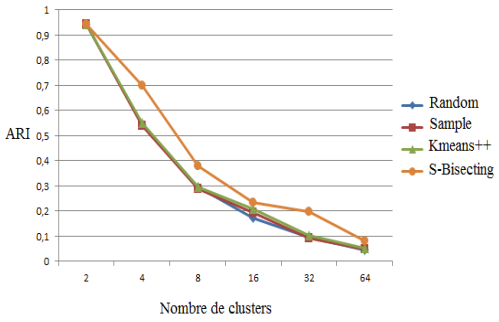

FIG. 1 – *Les performances moyennes en termes d'ARI en apprentissage pour le jeu de données Mushroom*

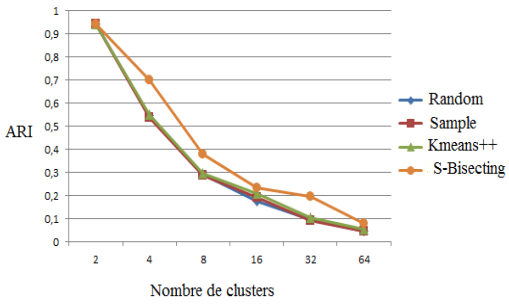

FIG. 2 – *Les performances moyennes en termes d'ARI en test pour le jeu de données Mushroom*

Nous montrerons lors des journées de la SFC l'ensemble des résultats obtenus sur 9 jeux de données de l'UCI. Ces résultats nous ont permis d'évaluer l'influence d'une méthode d'initialisation supervisée sur la performance prédictive de l'algorithme des K-moyennes standard. Pour le critère d'homogénéité, les résultats obtenus par la méthode d'initialisation supervisée sont compétitifs avec ceux obtenus par les méthodes non supervisées.

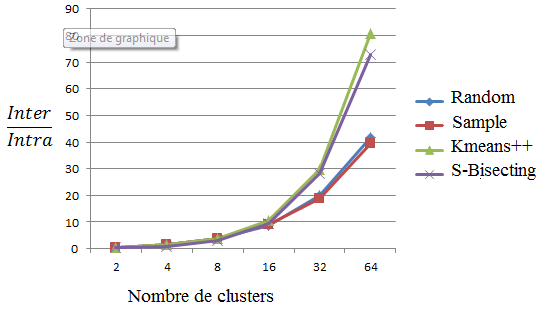

FIG. 3 – *Les performances moyennes en termes d'inertie en apprentissage pour le jeu de données Mushroom*

Références

Aggarwal, C. C. et C. K. Reddy (2013). *DATA CLUSTERING Algorithms and Applications*. Data Mining and Knowledge Discovery Series.

Al-Harbi, S. H. et V. J. Rayward-Smith (2006). Adapting k-means for supervised clustering. *Applied Intelligence 24*(3), 219–226.

Alaoui Ismaili, O., V. Lemaire, et C. Antoine (2015). Supervised preprocessings are useful for supervised clustering. *Springer Series Studies in Classification, Data Analysis, and Knowledge Organization*.

Arthur, D. et S. Vassilvitskii (2007). K-means++ : The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, USA, pp. 1027–1035. Society for Industrial and Applied Mathematics.

Celebi, M. E. et H. A. Kingravi (2014). Linear, deterministic, and order-invariant initialization methods for the k-means clustering algorithm. *CoRR abs/1409.3854*.

Cornuéjols, A. et L. Miclet (2010). *Apprentissage artificiel : Concepts et algorithmes*. Eyrolles.

Eick, C. F., N. Zeidat, et Z. Zhao (2004). Supervised clustering-algorithms and benefits. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pp. 774–776. IEEE.

Jain, A. K. (2010). Data clustering : 50 years beyond k-means. *Pattern Recogn. Lett. 31*(8), 651–666.

Maitra, R., A. D. Peterson, et A. P. Ghosh (2010). A systematic evaluation of different methods for initializing the k-means clustering algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 522–537.

Ocegueda-Hernandez, F. et R. Vilalta (2013). An empirical study of the suitability of class decomposition for linear models : When does it work well ? In *SDM*, pp. 432–440. SIAM.

Steinbach, M., G. Karypis, et V. Kumar (2000). A comparison of document clustering techniques. In *In KDD Workshop on Text Mining*.