Une méthode supervisée pour initialiser les centres des K-moyennes

Oumaima Alaoui Ismaili*,**, Vincent Lemaire*, Antoine Cornuèjols**

*Orange Labs, AV. Pierre Marzin 22307 Lannion cedex France
(oumaima.alaouiismaili, vincent.lemaire)@orange.com

**AgroParisTech 16, rue Claude Bernard 75005 Paris
antoine.cornuejols@agroparistech.fr

Résumé. Au cours des dernières années, la classification à base de clustering s'est imposée comme un sujet de recherche important. Cette approche vise à décrire et à prédire un concept cible d'une manière simultanée. Partant du fait que le choix des centres pour l'algorithme des K-moyennes standard a un impact direct sur la qualité des résultats obtenus, cet article vise alors 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 des K-moyennes.

1 Introduction

Depuis quelques années, l'étude d'un nouvel aspect de l'apprentissage connu sous le nom de la *classification à base de clustering* (ou *Supervised clustering en anglais*) a suscité beaucoup d'intérêt (*e.g.,* (Eick et al., 2004) et (Cevikalp et al., 2007)). Les approches appartenant à ce type d'apprentissage visent à décrire et à prédire d'une manière simultanée (Alaoui Ismaili et al., 2015a). Dans ce cadre d'étude, on suppose que la classification à base de clustering est étroitement liée à l'estimation de la distribution des données conditionnellement à une variable cible. A partir d'une base de données étiquetée, ces approches cherchent à découvrir la structure interne de la variable cible afin de pouvoir prédire ultérieurement la classe des nouvelles instances.

La figure 1 illustre la différence entre les trois types d'apprentissage : le clustering standard (a), la classification supervisée (b) et la classification à base de clustering (c). Dans la classification supervisée, la compacité des classes apprises (dans la phase d'apprentissage) n'est pas une condition importante (*e.g,* le groupe B2 de (b)). Le clustering regroupe les instances homogènes sans tenir compte de leur étiquetage (*e.g,* le groupe A4 de la a)). La classification à base de clustering vise à former des groupes compacts et purs en termes de classes (*e.g,* les 6 groupes de (c)).

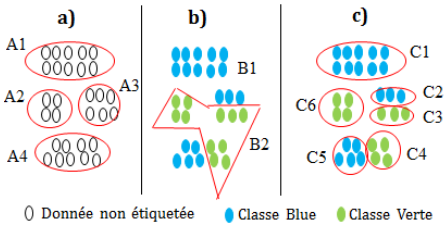

FIG. 1: Types d'apprentissage

La classification à base de clustering est très utile dans les domaines critiques où l'interprétation des résultats fournis par un système d'apprentissage est une condition primordiale. Elle permet à l'utilisateur de découvrir les différentes voies qui peuvent mener à une même prédiction : par exemple de découvrir que deux instances de même classe peuvent être très hétérogènes (*e.g.,* les instances appartenant au groupe $C1$ et au groupe $C5$ de (c)).

La classification à base des K-moyennes est une version modifiée de l'algorithme des K-moyennes standard. Elle cherche à générer des partitions ayant un bon compromis entre la compacité des groupes formés et leur pureté en termes de classes (voir la partie (c) de la figure ci-dessus). La prédiction de la classe des nouvelles instances se réalise par la suite en se basant sur la structure interne découverte lors de la phase d'apprentissage.

De par sa nature, l'algorithme standard des K-moyennes converge rarement vers un optimum global. La qualité de l'optimum local atteint et le temps requis par l'algorithme dépendent entre autres du choix des centres initiaux. L'utilisation d'une mauvaise méthode d'initialisation peut générer plusieurs effets indésirables tels que : $i)$ des clusters vides, $ii)$ une convergence plus lente, et $iii)$ une grande probabilité de tomber sur un mauvais optimum local et donc la nécessité d'exécuter l'algorithme plusieurs fois (Jain, 2010). Afin d'éviter ce risque, les centres choisis au départ doivent fournir une bonne couverture de l'espace de données. Ceci permet à l'algorithme d'obtenir un bon résultat sans recourir à de nombreuses excécutions, voire même à l'aide d'une seul exécution si la méthode est déterministe.

A partir de ce constat, il devient 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 au sens de la classification à base de clustering. Spécifiquement, une bonne méthode d'initialisation supervisée devrait réussir à capter les points appartenant aux régions denses et pures en termes de classes. L'obtention de ces points candidats peut faciliter la tâche à l'algorithme des K-moyennes standard puisqu'il va débuter avec une "bonne" partition initiale. Le but de ce travail est alors de répondre à la question : *"À quel point une méthode d'initialisation supervisée pourrait aider l'algorithme des K-moyennes standard à former des groupes compacts, homogènes et purs au sens de la classification à base de clustering ? "*

Cet article est consacré exclusivement à l'étude de l'impact d'une méthode d'initialisation supervisée sur la performance prédictive de l'algorithme des K-moyennes. On suppose ici que la compacité des groupes est garantie par l'algorithme des K-moyennes [1]. Pour atteindre cet objectif, nous proposons une nouvelle méthode supervisée d'initialisation des centres dans la section 2. Les résultats obtenus grâce à l'algorithme des K-moyennes en utilisant cette nouvelle méthode et les méthodes issues de la littérature sont présentés et discutés dans la section 3. Finalement, une conclusion générale sur la validité de l'hypothèse fait l'objet de la section 4.

2 Contribution : Rocchio-and-Split

L'intérêt de l'utilisation d'une méthode supervisée pour initialiser les centres dans le cadre de la classification à base des K-moyennes peut être vu clairement dans le cas de déséquilibre des classes à prédire. Au cours de la phase d'initialisation, 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, en terme de classe à prédire,

1. Dans cette étude, l'algorithme des K-moyennes est exécuté 100 fois. Ensuite, la partition optimale est considérée comme étant celle qui minimise la MSE (Erreur Quadratique Moyennes).

des clusters serait introduite. A partir de ce constat, on pense que l'intégration de l'information contenue dans la variable cible dans le processus d'initialisation peut s'avérer très utile.

La méthode d'initialisation proposée dans cet article est appelée "***Rochio-And-Split***" (RS). Dans le cas où le nombre de clusters (K) est égal au nombre de classes (C), cette méthode associe à chaque classe un centre qui est son centre de gravité (*i.e.,* la méthode Rocchio (Manning et al., 2008)). Dans le cas contraire, *i.e.,* lorsque le nombre de clusters est supérieur au nombre de classes, cette méthode suit une division hiérarchique descendante : elle part de C groupes où chaque groupe représente une classe. Ensuite, à chaque itération, le groupe qui vérifie un critère déterminé est divisé en deux : partant du fait que les groupes formés doivent être les plus homogènes et les plus compacts possible, le groupe à diviser est donc le groupe le plus dispersé. Pour mesurer cette dispersion, l'inertie intra-clusters est utilisée. Ce processus est répété jusqu'à ce que le nombre de groupes formés soit égal au nombre de centres désirés. Finalement, les centres sont obtenus en calculant les centres de gravité de chacun de ces groupes.

En se basant sur l'objectif finale de la classification à base de clustering [2], la méthode RS cherche à identifier les régions denses dans la classe la plus dispersée. Pour ce faire, la méthode commence par sélectionner la classe ou le groupe ayant une dispersion élevé. Ce groupe est considéré comme étant le groupe candidat à diviser. Pour le diviser, RS commence par sélectionner l'instance la plus éloignée du centre de gravité de ce groupe, notée $X_{i_{max}}$. Notons d_1 cette distance maximale. Ensuite, l'approche regroupe toutes les instances qui sont distantes de à $X_{i_{max}}$ de moins de d_1. Cela correspond à diviser le cercle de rayon d_1 en deux. Les différentes étapes de l'approche SB sont présentées dans l'algorithme 1.

Algorithm 1 Rochio-And-Split

Entrées: $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$ *: Le jeu de données d'apprentissage.*
K : Le nombre de clusters. et *C : Le nombre de classes.*
- Calculer le centre de gravité de chaque classe
Si $C = K$ **alors**
Sortie: les C Centres de gravité
fin Si
Si $K > C$ **alors**
Tant que *le nombre de clusters n'est pas atteint* **faire**
- Calculer la dispersion dans chaque cluster en sens de l'inertie intra
- Diviser le cluster le plus dispersé C_k en deux sous-clusters C_{k1} et C_{k2} de la manière suivante :
Sélectionner l'instance $X_{i_{max}}$ de C_k telle que $i_{max} = argmax_{\{i \in C_k\}} d(X_i, \mu_k)$ et $d_1 = d(X_{i_{max}}, \mu_k)$ avec μ_k est le centre de gravité du cluster k
Pour j allant de 1 jusqu'à N_k (le nombre d'instance dans C_k) **faire**
Si $d_1 \geq d(X_{i_{max}}, X_j)$ **alors**
$X_j \in C_{k1}$
Sinon
$X_j \in C_{k2}$
fin Si
fin Pour
- Supprimer le centre du groupe sélectionné et calculer les centres de gravité des deux clusters C_{k1} et C_{k2}
fin Tant que
Sortie: Les K centres initiaux
fin Si

2. Discerner des groupes compacts et purs en termes de classe afin de prédire ultérieurement la classe des nouveaux exemples

L'un des points fort de l'approche RS est qu'elle est une approche déterministe. Cela permet de réduire le temps de calcul puisqu'une seule exécution est nécessaire. La complexité de cette approche est linéaire en N : $\mathcal{O}(CN_jd + KN_kd(K - C))$.

De plus, la technique suivie pour diviser les groupes dispersés a pour effet de capter les deux points appartenant aux régions denses et pures en termes de classe. L'inconvénient majeur de cette approche est qu'elle est sensible à la présence de données bruitées, ou "outliers", en raison de l'utilisation de la distance maximale lors de la division. Néanmoins, ce point peut être très atténué voir éliminé en utilisant par exemple un bon prétraitement capable d'éliminer le bruit existant.

3 Expérimentation

Pour vérifier la validité de notre hypothèse, à savoir : "*les méthodes d'initialisation supervisées aident-elles l'algorithme des K-moyennes standard à fournir de bons résultats en termes de prédiction ?*", une comparaison des performances prédictives moyennes de l'algorithme des K-moyennes, précédé par différentes méthodes d'initialisation (supervisées et non supervisées) est effectuée. La méthodologie utilisée est la suivante :

- Les méthodes d'initialisation : Dans cette étude comparative, on s'intéresse aux méthodes d'initialisation communément utilisées dans la littérature. Pour les méthodes non supervisée, nous avons choisi d'utiliser : **Forgy** (Random), **Sample**, **MaxiMin** (déterministe (MM) et non déterministe (MM(Rand)), **Variance Partitioning** (Var-Part) et **K-means++** (K++). Le lecteur pourra trouver une description plus détaillée de ces méthodes dans (Celebi et Kingravi, 2015). Pour les méthodes supervisées, nous avons choisi d'utiliser la méthode **K-means++R** (K++R) (Lemaire et al., 2015) et la méthode proposée ici : **Rocchio-and-Split** (RS).

- Le prétraitement : Le prétraitement utilisé dans cette étude expérimentale est un prétraitement supervisé nommé : "*Conditional Info*". Ce choix fait suite à l'étude menée dans (Alaoui Ismaili et al., 2015b) où les auteurs ont montré que l'utilisation de ce prétraitement aide l'algorithme des K-moyennes standard à atteindre une bonne performance prédictive (le processus de prédiction est expliqué ci-dessous).

Nom	M_n	M_c	N	C	$K*$	**Nom**	M_n	M_c	N	C	$K*$
Iris (33)	4	0	150	3	4	Australian (56)	14	0	690	2	64
Hepatitis (79)	6	13	155	2	28	Pima (65)	8	0	768	2	27
Wine (40)	13	0	178	3	51	Vehicle (26)	18	0	846	4	76
Glass (36)	10	0	214	6	46	Tictactoe (65.34)	0	9	958	2	20
Heart (56)	10	3	270	2	60	LED (11)	7	0	1000	10	46
Horsecolic (63)	7	20	368	2	10	German (70)	24	0	1000	2	13
Soybean (14)	0	35	376	19	40	Segmentation (14)	19	0	2310	7	64
Breast (65)	9	0	683	2	60	Abalone (16)	7	1	4177	28	64

TAB. 1: Liste des jeux de données utilisés - Nom ($\approx$ pourcentage classe majoritaire)

- Les jeux de données : Les jeux de données utilisés dans cette étude sont tirés du répertoire de l'UCI (Lichman, 2013). Ces jeux de données ont été choisis afin d'avoir des bases de données diverses en termes de nombre de classes C, de variables (continues M_n et/ou catégorielles M_c) et d'instances N (voir Tableau 1).

- Les critères d'évaluation : Dans le cadre d'apprentissage supervisé, le critère d'évaluation communément utilisé est *Adjusted Rand Index* (*ARI*) (Hubert et Arabie, 1985).

- Nombre de clusters : Il varie de C jusqu'à $K*$. Pour chaque jeu de données, $K*$ a été déterminé au préalable de manière à ce que la partition obtenue, avec K=$K*$ permette d'obtenir un ratio (inertie inter) / (inertie totale) de 95% (voir Tableau 1).

- Affectation des classes aux clusters : A la fin du processus d'apprentissage, chaque groupe appris prend j comme étiquette si la majorité des exemples qui le forme sont de la classe j (*i.e.,* l'utilisation du vote majoritaire).

- La prédiction : la prédiction d'un nouvel exemple se fait selon son appartenance à un des groupes appris. Autrement dit, l'exemple reçoit j comme prédiction s'il est plus proche du centre de gravité du groupe de classe j (*i.e.,* l'utilisation du 1 plus proche voisin).

Il est à rappeler que les approches de classification à base de clustering cherchent à décrire et à prédire d'une manière simultanée. L'objectif est alors de trouver au cours de la phase d'apprentissage, le meilleur compromis entre la compacité et la pureté des groupes appris. Il s'agit donc de découvrir la structure interne de la variable cible. La prédiction de la classe des nouvelles instances est réalisée en se basant sur cette structure. Puisqu'il n'existe pas un critère global permettant de mesurer ce compromis, dans ce qui suit les performances prédictives des méthodes seront évaluées en utilisant l'ARI. En ce qui concerne la compacité, on suppose dans cet article qu'elle est garantie par l'algorithme des K-moyennes. Après de nombreuses exécutions [3] l'algorithme choisit celle qui minimise l'Erreur Quadratique Moyenne).

Pour cet axe d'évaluation, nous commençons par tracer pour chaque jeu de données et pour chaque méthode d'initialisation la courbe d'ARI en fonction du nombre de clusters (voir la méthodologie ci-dessus). L'aire sous cette courbe est ensuite calculée (ALC-ARI : Area under the Learning Curve de l'ARI). Enfin muni des valeurs d'ALC-ARI, qui synthétisent les résultats de chaque méthode d'initialisation, nous appliquons le test de Friedman couplé au test post-hoc de Nemenyi pour un seuil de significativité $\alpha = 0.05$.

La figure 2 présente le résultat des deux tests statistiques effectués sur les 16 jeux de données. Les méthodes dans cette figure sont classées par ordre décroissant selon leurs performance prédictives : plus la méthode est proche de 1 plus elle est meilleure en prédiction. Le test de Friedman montre qu'il existe une différence significative ($p_{value} = 1.38^{-6} \ll 0.05$) entre les méthodes, tandis que le test de Nemenyi partitionne les méthodes en deux groupes. Cette figure nous montre également que la méthode RS est la méthode qui fournit de bons résultats en termes de prédiction par rapport aux autres méthodes. Il est important de rappeler que la méthode RS est une méthode déterministe. De ce fait, l'algorithme des K-moyennes est exécuté qu'une seule fois. Cela veut dire que la méthode RS est capable d'obtenir de bonne performance prédictive mais elle permet également de réduire notablement le temps de calcul vis-à-vis des autres méthodes.

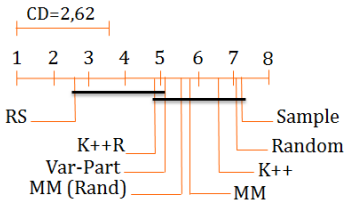

FIG. 2: Test de Friedman couplé au test post-hoc de Nemenyi

3. Dans cette étude comparative, on exécute l'algorithme des K-moyennes 100 fois

4 Conclusion

Cet article a présenté l'influence d'une étape d'initialisation supervisée sur la qualité des résultats générés par l'algorithme des K-moyennes standard. Nous avons pu montrer qu'une bonne méthode d'initialisation supervisée a la capacité d'aider cet algorithme à atteindre l'objectif de la classification à base des K-moyennes. Les résultats expérimentaux ont montré que l'algorithme des K-moyennes précédé par la méthode RS parvient à obtenir de bons résultats en termes de prédiction et ce en une seule exécution.

Références

Alaoui Ismaili, O., V. Lemaire, et A. Cornuéjols (2015a). Classification à base de clustering ou comment décrire et prédire simultanément. In *Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA)*.

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

Celebi, M. E. et H. A. Kingravi (2015). Linear, deterministic, and order-invariant initialization methods for the k-means clustering algorithm. In *Partitional Clustering Algorithms*, pp. 79–98. Springer.

Cevikalp, H., D. Larlus, et F. Jurie (2007). A supervised clustering algorithm for the initialization of rbf neural network classifiers. In *Signal Processing and Communications Applications, 2007. SIU 2007. IEEE 15th*, pp. 1–4. IEEE.

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

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

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

Lemaire, V., O. A. Ismaili, et A. Cornuéjols (2015). An initialization scheme for supervized k-means. In *International Joint Conference on Neural Networks (IJCNN), IEEE, Ireland*.

Lichman, M. (2013). UCI machine learning repository.

Manning, C. D., P. Raghavan, H. Schütze, et al. (2008). *Introduction to information retrieval*, Volume 1. Cambridge university press Cambridge.

Summary

Over the last few years, researchers have focused their attention on a new approach, supervised clustering, that combines the main characteristics of both traditional clustering and supervised classification tasks. Motivated by the importance of the initialization step in the traditional clustering context, this paper explores to what extent a supervised initialization step could help traditional clustering to obtain better performances on supervised clustering tasks.