Un critère d'évaluation pour les K-moyennes prédictives

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. L'algorithme des K-moyennes prédictives est un des algorithmes de clustering prédictif visant à décrire et à prédire d'une manière simultanée. Contrairement à la classification supervisée et au clustering traditionnel, la performance de ce type d'algorithme est étroitement liée à sa capacité à réaliser un bon compromis entre la description et la prédiction. Or, à notre connaissance, il n'existe pas dans la littérature un critère analytique permettant de mesurer ce compromis. Cet article a pour objectif de proposer une version modifiée de l'indice Davies-Bouldin, nommée SDB, permettant ainsi d'évaluer la qualité des résultats issus de l'algorithme des K-moyennes prédictives. Cette modification se base sur l'intégration d'une nouvelle mesure de dissimilarité permettant d'établir une relation entre la proximité des observations en termes de distance et leur classe d'appartenance. Les résultats expérimentaux montrent que la version modifiée de l'indice DB parvient à mesurer la qualité des résultats issus de l'algorithme des K-moyennes prédictives.

1 Introduction

L'algorithme des K-moyennes prédictives (Ismaili et al., 2015, 2016; Dimitrovski et al., 2014) est une version modifiée de l'algorithme des K-moyennes traditionnel (MacQueen, 1967). L'objectif de ce type d'algorithme est de *décrire et de prédire simultanément*. Contrairement à l'algorithme des K-moyennes traditionnel, l'algorithme des K-moyennes prédictives cherche à discerner à partir d'une base de données étiquetées, des groupes d'instances compacts, éloignés les uns des autres et purs en termes de classe dans le but de prédire ultérieurement la classe des nouvelles instances (voir la figure 1).

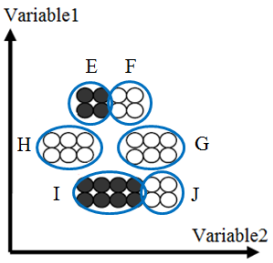

FIG. 1: Objectif du clustering prédictif

Pour mesurer la qualité des résultats issus de l'algorithme des K-moyennes prédictives, trois points doivent être pris en considération : i) le taux de bonnes prédiction, ii) la compacité et iii) la séparabilité des clusters. Il s'agit ici de réaliser un compromis entre la prédiction et la description.

Dans ce cadre d'étude, l'utilisation des critères usuels de la classification supervisée et du clustering pour évaluer la performance de ce type d'algorithme s'avère insuffisante. D'une part, les critères supervisés privilégient principalement l'axe de prédiction et ils n'accordent aucune importance à la compacité et à la séparabilité des clusters. D'autre part, la majorité des critères non supervisés se basent sur une mesure de similarité qui évalue la proximité entre les instances en termes de distance sans accorder une importance à leur classe d'appartenance. Par conséquent, deux instances d'étiquettes différentes vont être considérées comme similaires si elles sont proches en termes de distance.

À notre connaissance, il n'existe pas dans la littérature un critère analytique permettant de mesurer la qualité des résultats issus de l'algorithme des K-moyennes prédictives. Seule une technique permettant de résoudre des problèmes multi-critères tel le Front de Pareto peut être utilisée. Dans le problème de la sélection du nombre optimal de clusters (i.e, le choix du K), cette technique fournit dans la majorité du temps plusieurs solutions possibles (*i.e.,* optimums non dominés) pour un seul problème. L'objectif de cet article est de proposer un seul critère analytique permettant de sélectionner le nombre optimal de clusters dans le cadre des K-moyennes prédictives. Ce critère sera considéré comme pertinent s'il fournit, pour chaque jeu de données, un résultat très proche d'un optimum non dominé du Front de Pareto.

Le reste de cet article est organisé comme suit : le section 2 propose une version supervisée de l'indice de qualité non supervisé Davies-Bouldin (DB). Cette version notée SDB, intègre une nouvelle mesure de dissimilarité permettant d'évaluer la ressemblance entre deux instances étiquetées. Finalement, avant de conclure dans la section 4, plusieurs études expérimentales sont menées dans la section 3 afin de d'étudier la capacité de SDB à atteindre l'objectif souhaité.

2 Proposition d'une version supervisée de l'indice Davies-Bouldin (SDB)

Lors de la recherche d'un nouveau critère permettant de mesurer le compromis entre la description et la prédiction, deux voies peuvent être exploitées, à savoir : i) la modification d'un critère dédié à la classification supervisée, ii) la modification d'un critère dédié au clustering. Dans cet article, nous nous intéressons exclusivement à l'étude de la deuxième voie.

L'incapacité des critères non supervisés à mesurer le compromis entre la description et la prédiction s'illustre essentiellement dans le cas où certaines régions denses du jeu de données contiennent plusieurs classes. En effet, la plupart de ces critères sont basés sur une métrique qui permet d'évaluer la proximité entre les observations sans accorder d'importance à leur classe d'appartenance. Dans ce cas, deux observations ayant des étiquettes différentes vont être considérées comme similaires si elles sont proches en termes de distance. Pour surmonter ce problème, il est important de proposer nouvelle mesure permettant d'établir une relation entre la proximité des observations en termes de distance et leur classe d'appartenance.

Définition : Soit X_i et X_j deux observations de dimension d dans $\mathcal{D}$ appartenant respectivement à la classe $f(X_i)$ et $f(X_j)$. La nouvelle mesure de dissimilarité $DSim(X_i, X_j)$ qui relie la proximité de X_i et X_j à leurs classes d'appartenance est définie comme suit :

$$\forall X_i, X_j \in \mathcal{D} \quad DSim(X_i, X_j) = 1 - \frac{exp(-\delta(X_i, X_j))}{1 + dist(X_i, X_j)^2} \tag{1}$$

avec δ est la fonction indicatrice suivante : $\delta(X_i, X_j) = \begin{cases} 0 & \text{si} \quad f(X_i) = f(X_j) \\ 1 & \text{si} \quad f(X_i) \neq f(X_j) \end{cases}$

Il est à noter que la vraie classe $f(X_i)$ de l'observation X_i peut être remplacée par la classe prédite $\hat{f}(X_i)$ selon le besoin. La distance $dist(X_i, X_j)$ utilisée est une distance normalisée. La mesure de dissimilarité $DSim(X_i, X_j)$ prend ses valeurs entre 0 et 1 :

— $\forall X_i, X_j \in \mathcal{D} \;\; DSim(X_i, X_j) = 0 \Leftrightarrow dist(X_i, X_j) = 0$ **ET** X_i et X_j ont la même classe.

— $\forall X_i, X_j \in \mathcal{D} \;\; DSim(X_i, X_j) = 1 \Leftrightarrow dist(X_i, X_j) \to \infty$.

La mesure de dissimilarité proposée utilise un paramètre intrinsèque qui permet de pénaliser la distance entre deux observations de classes différentes et qui sont proches en termes de distance. Ce paramètre a un impact direct sur les résultats. Dans notre cadre d'étude, nous avons constaté que l'utilisation de l'exponentielle nous permet d'obtenir des résultats qui sont très proches du Front de Pareto.

Afin d'être en mesure d'évaluer le compromis entre la description et la prédiction, la mesure de dissimilarité présentée ci-dessus peut être intégrée dans un critère non supervisé dédié au clustering traditionnel. Ce critère doit impérativement être basé sur la notion d'inertie intra/inter clusters afin de pouvoir évaluer le compromis compacité-prédiction. Parmi ces critères, on trouve l'indice Davies-Bouldin (DB) (Davies et Bouldin, 1979). DB traite chaque cluster individuellement et cherche à mesurer à quel point il est similaire au cluster qui lui est le plus proche. La version supervisée de DB, notée SDB, est donnée par la formule suivante :

$$SDB = \frac{1}{K} \sum_{k=1}^{K} max_{1 \leq k \neq t \leq K} \left\{ \frac{S_k + S_t}{M_{kt}} \right\} \tag{2}$$

S_k mesure le degré de la compacité du cluster k. Elle représente la moyenne des distances entre les observations du cluster k et leur centre de gravité G_k. Dans notre cadre d'étude, la compacité et la pureté en termes de classes peuvent être évaluées simultanément en intégrant la nouvelle mesure de similarité dans la quantité $S_k = \frac{1}{N_k} \sum_{i=1}^{N_k} Sim(X_i, G_k)$ avec $Sim(X_i, G_k) = 1 - \frac{exp(-\delta_1(X_i, G_k))}{1 + dist(X_i, G_k)^2}$ et $\delta_1(X_i, G_k) = \begin{cases} 0 & \text{si} \quad f(X_i) = \hat{f}(G_k) \\ 1 & \text{si} \quad f(X_i) \neq \hat{f}(G_k) \end{cases}$

La mesure de compacité S_k prend ses valeurs dans l'intervalle $[0, 1]$. $S_k = 0$ si le cluster k est formé d'une seule observation et $S_k = 1$ si les observations qui le forment sont très éloignées les unes des autres et appartiennent à des classes différentes. De ce fait, on constate que plus S_k est petite plus le cluster k est compact et pur en termes de classe.

La quantité M_{kt}, quant à elle, mesure le degré de la séparabilité entre les deux clusters k et t. Elle représente donc la distance entre le centre de gravité des deux clusters :

$M_{kt} = Sim(G_k, G_t) = 1 - \frac{exp(-\delta_2(G_k, G_t))}{1 + dist(G_k, G_t)^2}$, $\delta_2(G_k, G_t) = \begin{cases} 0 & \text{si} \quad \hat{f}(G_k) = \hat{f}(G_t) \\ 1 & \text{si} \quad \hat{f}(G_k) \neq \hat{f}(G_t) \end{cases}$

La mesure de séparabilité M_{kt} prend ses valeurs dans l'intervalle $[0, 1]$. Elle est égale à zéro si $dist(G_k, G_t) = 0$ et les deux clusters ont la même classe prédite ($\hat{f}(G_t) = \hat{f}(G_k)$). Elle est égale à 1 si et seulement si $dist(G_k, G_t) \to \infty$. De ce fait, plus M_{kt} est grande plus les deux clusters sont éloignés les uns des autres.

SDB est un critère à minimiser. Plus SDB est proche de 0 plus les groupes appris sont compacts, purs en termes de classes et éloignés les uns des autres.

3 Expérimentation

Afin de vérifier la capacité de SDB à mesurer le compromis entre la description et la prédiction, l'algorithme des K-moyennes prédictives proposé dans (Ismaili et al., 2015, 2016) est utilisé. Dans cette étude expérimentale, nous allons étudier le problème de la sélection du nombre optimal de clusters dans le cadre des K-moyennes prédictives en utilisant des jeux de données contrôlés et des jeux de données de l'UCI.

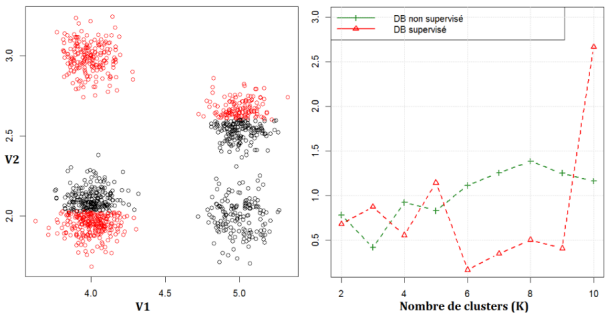

FIG. 2: Premier jeu de données contrôlé

Sur des jeux de données contrôlés : ces jeux de données permettent de mieux évaluer la performance des algorithmes puisque l'on connaît par construction la structure sous-jacente des données. Le premier jeu de données présenté dans cette étude expérimentale contient des régions denses contenant 2 classes (voir la partie gauche de la figure 2). C'est le cas où les critères non supervisés tels DB ont du mal à sélectionner le nombre optimal des clusers. Visuellement, pour ce jeu de données, on constate que la partition optimale au sens du clustering prédictif est celle qui contient 6 groupes. Les résultats présentés dans la partie droite de la figure 2 montrent que le critère SDB (courbe rouge) parvient à détecter le nombre exacte des clusters tandis que le critère DB (courbe verte) ne parvient pas à le détecter.

Le deuxième jeu de données simulé (voir la partie gauche de la figure 3) est caractérisé par la présence de 765 instances, 9 variables descriptives et une variable possédant deux classes à prédire dont la première contient 3 sous-groupes et la deuxième contient deux sous-groupes (*i.e.,* $K_{opti} = 5$). Les résultats présentés dans la partie gauche de la figure 3 montrent que le critère SDB (courbe rouge) parvient à sélectionner le nombre réel des clusters contrairement à l'indice DB (courbe verte).

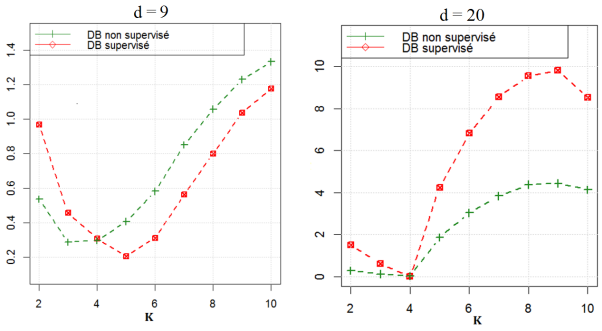

FIG. 3: Deuxième et troisième jeux de données contrôlés

Le troisième jeu de données (voir la partie droite de la figure 3) est lui caractérisé par la présence de 2376 instances, 20 variables descriptives et une variable à prédire contenant 2 classes dont chacune possède deux sous-groupes (*i.e.,* $K_{opti} = 4$). La partie droite de la figure 3 montre que les deux critères SDB (courbe rouge) et DB (courbe verte) arrivent à détecter le nombre optimal de clusters.

Sur des jeux de données de l'UCI : afin de montrer davantage la capacité du critère SDB à bien détecter le nombre optimal de clusters (au sens du clustering prédictif) et donc détecter la partition qui réalise le bon compromis entre la description et la prédiction, nous allons mener une étude sur 6 jeux de données de l'UCI (voir les 5 premières colonnes du tableau 1). Les résultats obtenus par le critère SDB pour chaque jeu de données seront comparés aux résultats obtenus par le Front de Pareto en utilisant le critère non supervisé DB pour évaluer la description et le critère supervisé indice de Rand Ajusté (ou ARI) (Hubert et Arabie, 1985) pour évaluer la prédiction. Pour avoir deux critères à minimiser 1-ARI est utilisé. Dans cette étude expérimentale, pour le problème de la sélection du nombre optimal de clusters pour les K-moyennes prédictives, le critère supervisé SDB sera considéré comme pertinent s'il fournit, pour chaque jeu de données, un résultat très proche d'un optimum non dominé obtenu par le Front de Pareto.

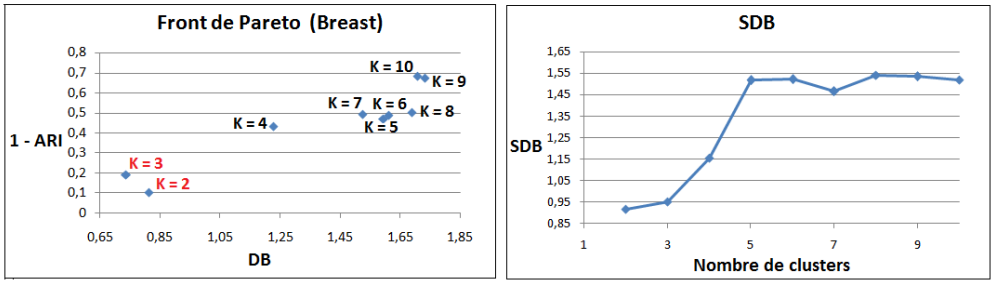

FIG. 4: Comparaison des résultats obtenus par le Front de Pareto et par SDB pour le jeu de données Breast.

La figure 4 et tableau 1 présentent respectivement l'évaluation de la performance de l'algorithme des K-moyennes prédictives pour différents nombres de clusters en utilisant le Front de Pareto et le critère SDB pour le jeu de données Breast et d'autres jeux de l'UCI. Ces résultats montrent que le critère SDB parvient à sélectionner, pour chaque jeu de données, la partition qui établie un bon compromis entre la description et la prédiction : les résultats obtenus par SDB appartiennent aux Fronts de Pareto obtenus pour les 6 jeux de données comme le montre la figure 4 ($K_{opti} = 2$) et les deux dernières colonnes du tableau 1.

ID	Données	# Instances	# Variables	# Classes	Font de Pareto	SDB
1	Breast	683	9	2	K2, K3	K2
2	Wine	178	13	3	K3	K3
3	German	1000	24	2	K4, K5, K10	K10
4	Adult	48842	15	2	K3, K4, K5	K3
5	Mushroom	8416	22	2	K2, K3, K10	K10
6	Waveform	5000	21	3	K3, K4, K5, K6	K3

TAB. 1: Comparaison des résultats obtenus par le Front de Pareto avec ceux obtenus par le critère SDB pour 6 jeux de l'UCI.

4 Conclusion

Cet article a présenté une version supervisée de l'indice Davies-Bouldin, nommée SDB permettant de mesurer la qualité des résultat issus de l'algorithme des K-moyennes préditives. Cet indice est basé sur une nouvelle mesure de dissimilarité permettant d'établir une relation entre la proximité des instances (distance) et leurs classes d'appartenance. Deux instances sont considérées comme similaires suivant cette nouvelle mesure, si et seulement si, elles sont proches en termes de distance **et** appartiennent à la même classe. Grâce à cette nouvelle mesure, la version supervisée de l'indice de Davies-Bouldin arrive à surmonter le problème de la non corrélation entre les clusters et les classes. Les résultats expérimentaux ont montré que l'indice SDB arrive à bien détecter le nombre optimal de clusters (sous forme d'un scalaire) permettant de mieux découvrir la structure interne de la variable cible par rapport au critère DB.

Références

Davies, D. L. et D. W. Bouldin (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell. 1*(2), 224–227.

Dimitrovski, I., D. Kocev, S. Loskovska, et S. Dzeroski (2014). Fast and efficient visual codebook construction for multi-label annotation using predictive clustering trees. *Pattern Recognition Letters 38*, 38–45.

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

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

Ismaili, O. A., V. Lemaire, et A. Cornuéjols (2016). Une méthode supervisée pour initialiser les centres des k-moyennes. In *16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016, 18-22 Janvier 2016, Reims, France*, pp. 147–152.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam et J. Neyman (Eds.), *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - Vol. 1*, pp. 281–297.

Summary

Predictive K-means is a predictive clustering algorithm which allows to describe and predict simultaneously. Unlike supervised classification and traditional clustering, the performance of this type of algorithm is closely related to its ability to achieve a good tradeoff between both the prediction and the description. Yet, to our knowledge, an analytical criterion to measure this compromise does not exist. In this paper, we propose SDB a modified version of Davies-Bouldin index to evaluate the performance quality of the predictive K-means. This modification is based on the integration of a new dissimilarity measure to build a relationship between the closeness of observations in terms of distance and their class membership. The experimental results has shown that our proposed criterion allows to measure the description/prediction compromise from the results obtained by the predictive K-means approach.