Évaluation de la qualité de l'algorithme des K-moyennes prédictives

Sommaire

Ce chapitre a fait l'objet de la publication suivante :

[13] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols : « Evaluation of predictive clustering quality », in MBC2, on Model Based clustering and classification (MBC2,2016).

5.1 Introduction

Comme évoqué précédemment dans le chapitre 2 Section 2.5.2, il existe deux types de clustering prédictif. **Le premier type** consiste à discerner un nombre *minimal* de groupes d'instances purs en termes de classes dans le but de prédire ultérieurement la classe des nouvelles instances (voir la figure 5.1). Il s'agit de découvrir *partiellement* la structure interne de la variable cible. Comme pour la classification supervisée, le but majeur de ces algorithmes est de prédire correctement la classe des nouvelles instances. De ce fait, pour évaluer la qualité des résultats issus de ce type d'algorithme, l'un des critères supervisés dédiés à la classification supervisée, tels l'indice de rand Ajusté (ARI) [57] ou la variation d'information (VI) [76], peut être utilisé.

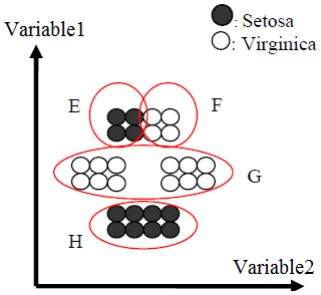

FIGURE 5.1 – Premier type du clustering prédictif

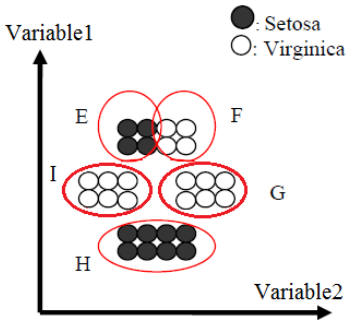

FIGURE 5.2 – Deuxième type du clustering prédictif

Le deuxième type, quant à lui, a pour but de discerner des groupes d'instances compacts, purs en termes de classe et éloignés les uns des autres (voir la figure 5.2). Contrairement à la classification supervisée, les algorithmes appartenant à ce type d'apprentissage cherchent à découvrir la structure interne *complète* de la variable cible. Puis, munie de cette structure, ils cherchent à prédire la classe des nouvelles instances. Dans ce cadre d'étude, aucun axe n'est privilégié par rapport à l'autre (*i.e.*,, la description et la prédiction). Une bonne partition au sens du deuxième type du clustering prédictif est donc celle qui réalise un bon compromis entre la description et la prédiction. Le critère choisi pour mesurer la qualité des résultats issus par les algorithmes du deuxième type du clustering prédictif doit impérativement équilibrer les trois points suivant :

1. Inertie intra-clusters minimale.
2. Inertie inter-clusters maximale.
3. Taux de bonnes classifications maximal.

Ce chapitre traite exclusivement la problématique d'évaluation de la qualité pour l'algorithme des K-moyennes prédictives du deuxième type. Dans la phase d'apprentissage, l'algorithme des K-moyennes prédictives nécessite une évaluation de la qualité à deux niveaux : 1) *Pour le choix de la meilleure partition à K fixé.* En effet, l'algorithme des K-moyennes prédictives converge rarement vers un optimum global. Pour cette raison, pour un nombre fixe de clusters, cet algorithme doit être exécuté plusieurs fois dans le but de choisir, via un critère analytique, la meilleure partition au sens du deuxième type du clustering prédictif, 2) *Pour la sélection du nombre optimal de clusters (K_{opti}).* En effet, l'algorithme des K-moyennes prédictives nécessite une connaissance *a priori* du nombre optimal de clusters ce qui n'est pas une tâche aisée dans la réalité. Pour surmonter ce problème, l'algorithme peut être exécuté plusieurs fois avec différents nombres de

clusters dans le but de sélectionner le nombre optimal de clusters permettant ainsi de mieux décrire la structure interne de la variable cible. Le critère utilisé dans ce cas doit impérativement être capable de comparer deux partitions ayant des nombres de clusters différents ce qui n'est pas une obligation pour le critère utilisé au premier niveau d'évaluation.

Les critères qui peuvent être utilisés pour choisir la meilleure partition (à K fixé) sont notamment les critères supervisés tel que l'indice de Rand Ajusté "ARI" ou les critères non supervisés tel que l'erreur quadratique moyenne "MSE". Le choix du critère à utiliser aura donc un impact direct sur les résultats : la meilleure partition va donc en dépendre. Pour pouvoir effectuer ce choix, il est important tout d'abord de connaître l'influence de l'utilisation d'un des critères (supervisé ou non supervisé) sur les résultats obtenus. Dans ce contexte d'étude, le bon critère à utiliser est celui qui conduit soit à une amélioration significative au niveau des deux axes (*i.e.,* la description et la prédiction) vis-à-vis des résultats obtenus par les autres critères ou soit à une amélioration significative sur l'un des deux axes (par exemple, l'axe de prédiction si le critère supervisé ARI est utilisé) et à une détérioration très légère (voire aucune détérioration) de l'autre axe (voir Figure 5.3). La section 5.2.1 de ce chapitre présente une étude expérimentale permettant de répondre à cette question.

Pour le deuxième niveau d'évaluation, les critères supervisés et non supervisés existant dans la littérature n'arrivent pas dans tous les cas à sélectionner le nombre optimal de clusters permettant de mieux découvrir la structure interne de la variable cible. Par exemple, pour les critères non supervisés, cette incapacité s'illustre essentiellement dans le cas de la présence des régions denses possédant au moins deux classes. En effet, la majorité de ces critères se basent sur une métrique permettant de mesurer la proximité entre les instances indifféremment de leurs classes d'appartenance. Par conséquent, des instances de différentes classes peuvent être vues comme similaires si elles sont proches en termes de distance. Pour tenter de résoudre cette problématique, ce chapitre propose une version supervisée de l'indice de Davies-Bouldin, noté **SDB**. Cet indice est basé sur une nouvelle mesure de similarité 'supervisée' permettant de relier la proximité des instances en termes de distance à leur classe d'appartenance : *deux instances sont considérées comme similaires si et seulement si elles sont proches en termes de distance et appartiennent à la même classe.* Le lecteur pourra trouver une description plus détaillée de cette problématique dans la section 5.3 de ce chapitre.

Le reste de ce chapitre est organisé comme suit : la section 5.2.1 a pour but de tester l'impact de l'utilisation d'un critère supervisé ou non supervisé sur les résultats obtenus lors du choix de la meilleure partition. Cette étude expérimentale donne une indication de la capacité de ces critères à réaliser un bon compromis entre la description et la prédiction (pour un nombre fixe de clusters). La section 5.2.2 présente une discussion sur les cas où les critères supervisés et non supervisés se montrent incapables de sélectionner le nombre optimal de cluster. Ceci peut être vu comme un point de départ vers la recherche d'un nouveau critère d'évaluation pour le clustering prédictif. Dans ce contexte, la section 5.3 présente la nouvelle version supervisée de l'indice de Davies-Bouldin, notée SDB (Supervised Davies-Bouldin). Cet indice est basé sur une nouvelle mesure de similarité supervisée présentée dans la section 5.3.2. Finalement, avant de conclure ce chapitre dans la section 5.5, quelques études expérimentales seront menées dans la section 5.4 afin de prouver la capacité du critère modifié à bien mesurer le compromis entre la description et la prédiction.

5.2 Évaluation de la qualité du deuxième type du clustering prédictif

5.2.1 Influence du choix de la meilleure partition

Dans la phase d'apprentissage, l'algorithme des K-moyennes prédictive converge rarement vers un optimum global. De ce fait, pour un nombre fixe de clusters, cet algorithme doit être exécuté $R > 1$ fois (voir la boucle **Pour** de l'algorithme 7) dans le but de choisir la meilleure partition au sens du clustering prédictif du deuxième type.

Entrée
- Un ensemble de données D, où chaque instance X_i est décrite par un vecteur de d dimensions et par une classe $Y_i \in \{1, \ldots, J\}$.
- Le nombre de clusters souhaité, noté K.

Début
1) Prétraitement des données.
2) Initialisation des centres.
Pour un nombre fixé de partitions, noté **R faire**
Répéter
3) *Affectation :* générer une nouvelle partition en assignant chaque instance X_i au groupe dont le centre est le plus proche.

$$X_i \in C_k \forall j \in 1, \ldots, K \qquad k = min_j \parallel X_i - \mu_j \parallel$$

avec μ_k est le centre de gravité du cluster C_k.
4) *Représentation :* calculer les centres associés à la nouvelle partition

$$\mu_k = \frac{1}{N_k} \sum_{X_i \in C_k} X_i$$

jusqu'à ce que (convergence de l'algorithme)
Fin Pour
5) Choix de la meilleure partition parmi les R partitions.
6) Attribution des classes aux clusters formés.
7) Prédiction de la classe des nouvelles instances.
Fin

Sortie
- Chaque cluster est représenté par un prototype qui possède la même prédiction de classe.
- Chaque cluster est associé à une description donnée par le biais de langage B.
- L'inertie intra-clusters est minimale (l'homogénéité des instances est maximale).
- L'inertie inter-clusters est maximale (la similarité entre les clusters est minimale).
- Le taux de bonnes classifications est maximal.

Algorithme 7 – Algorithme des K-moyennes prédictives du deuxième type

Il est à rappeler qu'une bonne partition au sens du deuxième type du clustering prédictif est celle qui réalise un bon compromis entre la description et la prédiction. Les trois points à respecter lors de l'évaluation de la qualité des résultats sont notamment la compacité, la séparabilité et la pureté des clusters en termes de classe. Dans ce contexte d'étude, les critères existant dans la littérature permettant de choisir la meilleure partition sont les critères supervisés tels que l'ARI ou les critères non supervisés tels que la MSE. Cependant, les critères supervisés privilégient principalement l'axe de prédiction tandis que les critères non supervisés privilégient principalement l'axe de description. Le choix du critère à utiliser aura donc un impact direct sur la qualité des résultats au sens du clustering prédictif : la meilleure partition va en dépendre. Suivant ce raisonnement, il est naturel de se demander quel est le critère (supervisé ou non supervisé) le plus adéquat à utiliser dans notre cadre d'étude ?

Dans ce contexte d'étude, un bon critère (supervisé ou non supervisé) est défini comme celui qui conduit :

— soit à une amélioration significative au niveau des deux axes vis-à-vis des résultats obtenus par les autres critères d'évaluation (*i.e.,* les résultats obtenus via ce critère ont tendance à suivre la flèche 3 de la figure 5.3 si les deux critères utilisés pour évaluer l'axe description et l'axe de prédiction sont à maximiser).

— soit à une amélioration significative au niveau d'un axe (par exemple, au niveau de l'axe de prédiction si le critère supervisé ARI est utilisé) et à une détérioration très légère (voire aucune détérioration) au niveau de l'autre axe : les résultats obtenus pour la meilleure partition via le critère ont tendance à suivre la flèche 1 de la figure 5.3 si un critère supervisé tel que l'ARI est utilisé pour le choix, ou bien de suivre la flèche 2 de la figure 5.3 si un critère non supervisé tel que MSE est utilisé.

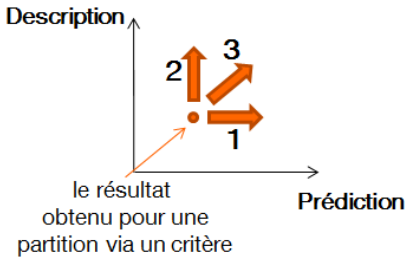

FIGURE 5.3 – L'influence de l'utilisation d'un critère supervisé ou non supervisé sur les résultats.

Pour être en mesure de connaître l'influence du choix du critère d'évaluation sur la qualité des résultats et donc connaître le critère (ARI ou MSE) le plus adéquat à utiliser pour choisir la meilleure partition, une étude expérimentale est alors menée en utilisant plusieurs jeux de données de l'UCI. Cette étude expérimentale nous permet également d'avoir une idée sur le degré de la corrélation entre les deux critères ARI et MSE. Dans cette étude, pour chaque jeu de données, le nombre de clusters est varié entre J et $J+10$, J étant le nombre de classes à prédire. Pour un nombre fixe de clusters, l'algorithme des K-moyennes est exécuté 100 fois dans le but de choisir la meilleure partition en utilisant soit l'ARI soit la MSE. La méthode d'initialisation utilisée à ce stade est la méthode qui garantit un bon compromis entre la prédiction et la description à savoir, S-Bisecting (voir Chapitre 4 Section 4.6).

La partie gauche (respectivement droite) des figures 5.4, 5.5, 5.6, 5.7 et 5.8 présente respectivement les valeurs de l'indice ARI (respectivement MSE) lorsque le choix de la meilleure partition est effectué à l'aide de la MSE (les courbes rouges de la figure) ou à l'aide de l'ARI (voir les courbes bleues de la figure) pour les jeux de données Wine, Hepatitis, Breast, Horsecolic et Segmentation. Le lecteur pourra trouver les résultats sur d'autres jeux de données dans l'annexe D de ce mémoire.

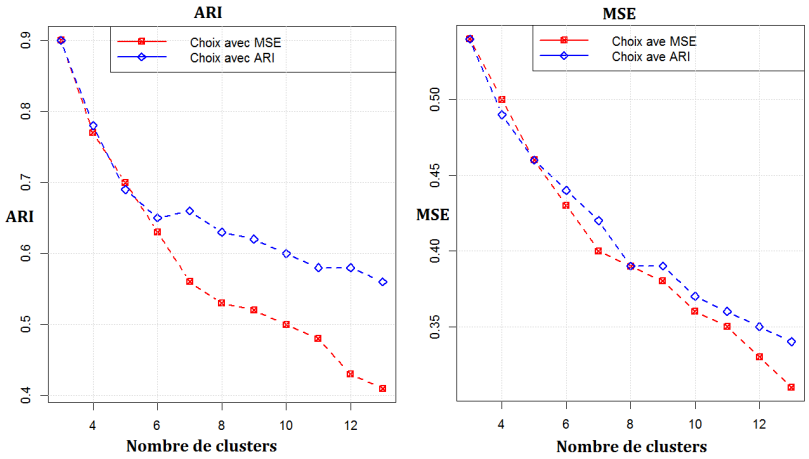

FIGURE 5.4 – L'évolution des courbes de l'ARI (partie gauche) et de la MSE (partie droite) selon le critère utilisé pour choisir la meilleure partition pour le jeu de données **Wine**

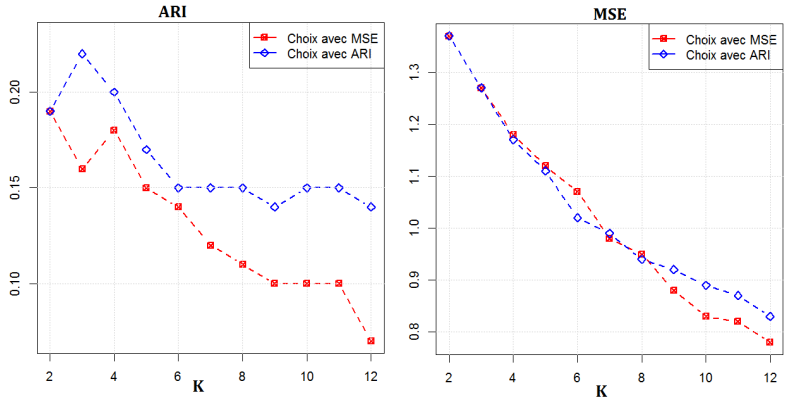

FIGURE 5.5 – L'évolution des courbes de l'ARI (partie gauche) et de la MSE (partie droite) selon le critère utilisé pour choisir la meilleure partition pour le jeu de données **Hepatitis**

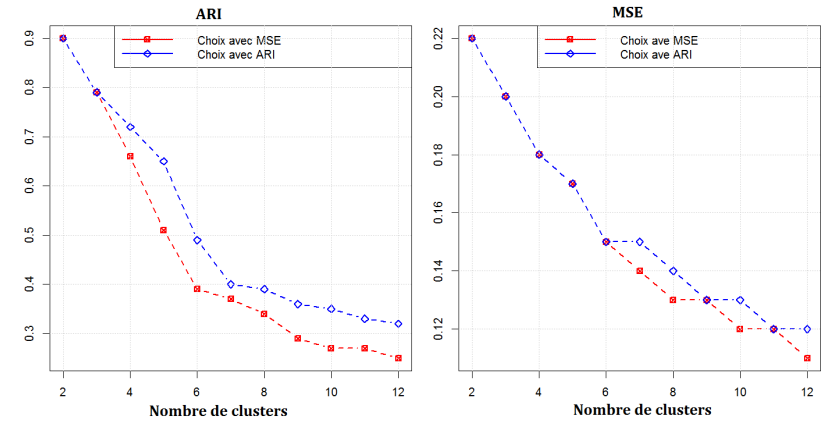

FIGURE 5.6 – L'évolution des courbes de l'ARI (partie gauche) et de la MSE (partie droite) selon le critère utilisé pour choisir la meilleure partition pour le jeu de données **Breast**.

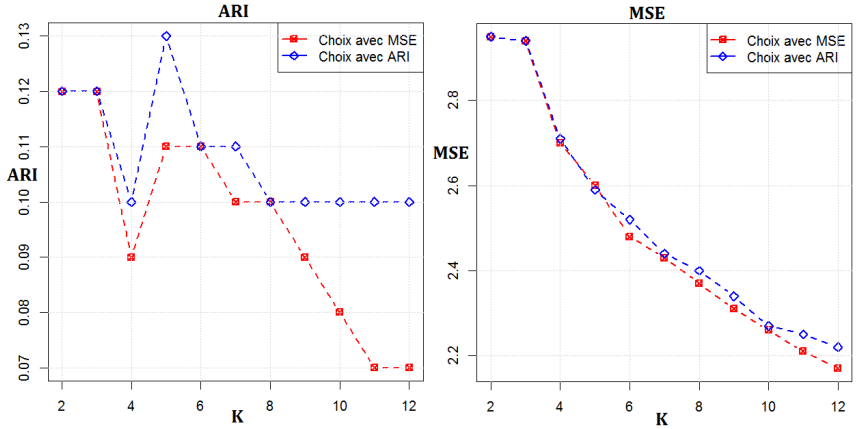

FIGURE 5.7 – L'évolution des courbes de l'ARI (partie gauche) et de la MSE (partie droite) selon le critère utilisé pour choisir la meilleure partition pour le jeu de données **Horsecolic**.

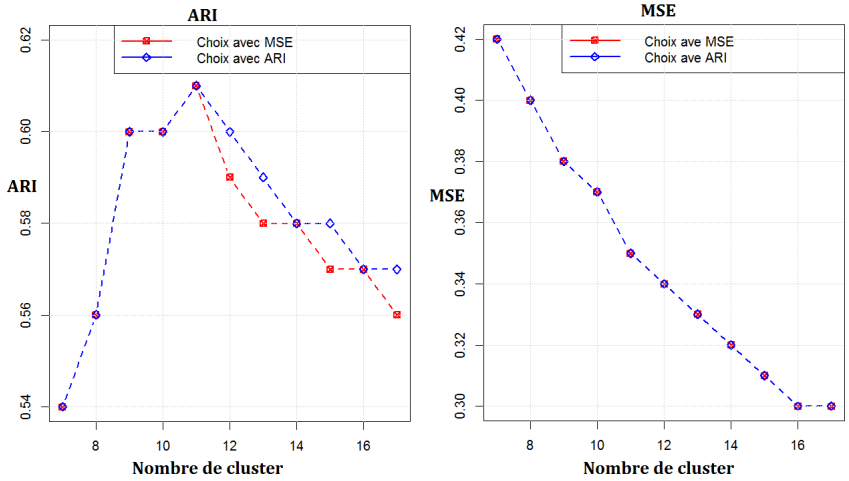

FIGURE 5.8 – L'évolution des courbes de l'ARI (partie gauche) et de la MSE (partie droite) selon le critère utilisé pour choisir la meilleure partition pour le jeu de données **Segmentation**

Les résultats expérimentaux présentés dans ces figures 5.4, 5.5, 5.6, 5.7 et 5.8 prouvent que les deux critères ARI et la MSE ne sont pas forcément corrélés : une amélioration au niveau d'un critère n'implique pas forcement de détérioration au niveau de l'autre critère. À titre d'exemple, la figure 5.5 montre que le choix de la meilleure partition en utilisant l'ARI engendre une amélioration significative au niveau de l'ARI avec une amélioration très légère au niveau de la MSE dans le cas où $k \in \{2, \ldots, 8\}$.

Le critère le plus adéquat parmi l'ARI et la MSE pour choisir la meilleure partition dans le cadre du clustering prédictif du deuxième type semble être (sur ces jeux de données) le critère ARI. En effet, celui-ci permet d'améliorer significativement les performances prédictives du modèle vis-à-vis des résultats obtenus en utilisant le critère non supervisé MSE (à titre d'exemple, pour le jeu de données Wine dans le cas où $K \in \{7, \ldots, 13\}$, voir la courbe bleue de la partie droite de la figure 5.4). De plus, aucune détérioration (par exemple, pour le jeu de données Segmentation, voir la partie droite de la figure 5.8) ou bien une détérioration très légère (par exemple, voir la courbe rouge pour le jeu de données Breast présenté dans la partie droite de la figure 5.6) au niveau de l'axe de description est introduite.

5.2.2 Choix du nombre optimal de clusters

L'algorithme des K-moyennes prédictives nécessite une connaissance a priori du nombre optimal du cluster (voir Algorithme 7). Or, il est très difficile dans la réalité de connaître à l'avance, pour chaque jeu de données, ce nombre optimal. Pour remédier à ce problème, l'algorithme des K-moyennes prédictives doit être exécuté plusieurs fois avec différents nombres de clusters dans le but de sélectionner le nombre de clusters permettant ainsi de découvrir au mieux la structure interne de la variable cible.

Pour le choix du nombre optimal de clusters, le critère recherché doit être capable de sélectionner la partition découvrant la structure interne "complète" de la variable cible. Il s'agit ici de pouvoir comparer deux partitions ayant un nombre de clusters différents au sens du deuxième type du clustering prédictif. Les critères supervisés et non supervisés existant dans la littérature permettant de comparer des partitions avec différents nombre de clusters n'arrivent pas toujours à détecter le nombre de clusters optimal. En effet, les critères proposés dans le cadre de la classification supervisée privilégient principalement l'axe de prédiction par rapport à l'axe de description. D'une part, l'utilisation des critères tels que la précision (ACC) et l'aire sous la courbe de ROC (AUC) pour mesurer la qualité des résultats s'avère inappropriée dans ce cadre d'étude : l'augmentation du nombre de clusters induit souvent une amélioration, ou une stagnation, au niveau de la performance prédictive du modèle. Par conséquent, la partition optimale sélectionnée par ce type de critère peut contenir un nombre de clusters très grand par rapport au "réel" nombre.

D'autre part, l'utilisation des critères d'évaluation tels que l'indice de rand ajusté (ARI) ou les critères basés sur la théorie d'information comme la variation d'information (VI), s'avère également insuffisante : lorsque deux sous-groupes différents d'instances de même classe sont proches comme illustré dans la figure 5.9 (les deux sous-groupes de la classe rouge situés au milieu de la figure), ces deux critères cherchent à les fusionner ensemble. Dans le cas extrême, ces deux sous-groupes peuvent être également fusionnés même s'ils sont très éloignés l'un de l'autre.

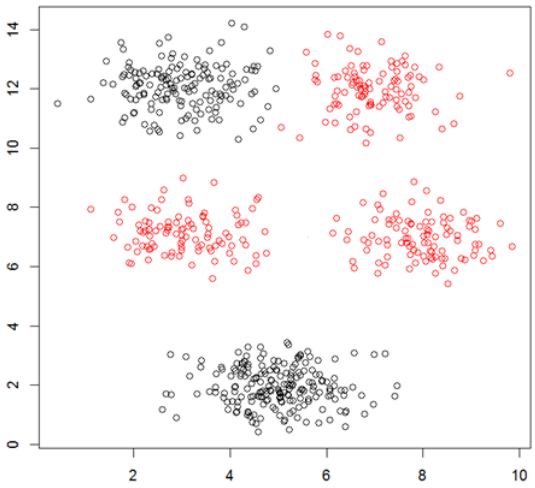

FIGURE 5.9 – Exemple illustratif d'incapacité des critères supervisés à mesurer le compromis description/prédiction

Enfin, les critères non supervisés proposés dans la littérature pour évaluer la qualité des résultats issus des algorithmes du clustering traditionnel privilégient principalement l'axe de description. L'utilisation de tels critères dans le cadre du clustering prédictif s'avère inappropriée. En effet, l'incapacité des critères non supervisés à mesurer la qualité des résultats issus des algorithmes du clustering prédictif s'illustre principalement dans le cas de la non corrélation entre les clusters et les classes. C'est le cas de la présence de plus de deux classes dans au moins une des régions denses. À titre d'exemple, le jeu de données présenté dans la figure 5.10 possède deux régions denses caractérisées par la présence des deux classes (rouge et noire). Pour cet exemple, il est clair que la partition optimale suivant ces critères est celle qui contient le nombre de régions denses comme nombre de clusters (*i.e.,* 4 clusters). Par conséquent, deux groupes formés dans ce cas vont contenir des instances de classes différentes ce qui conduit à une détérioration au niveau de la prédiction.

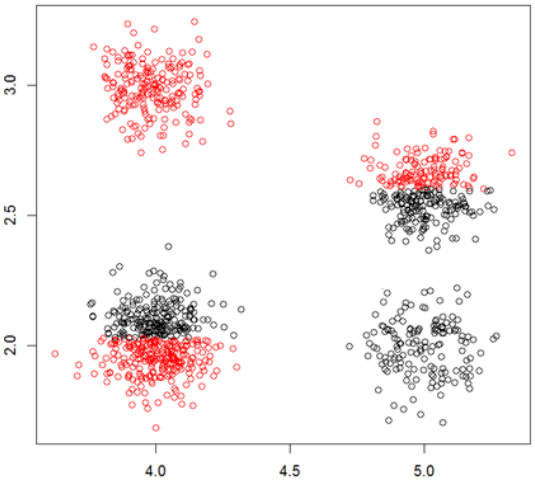

FIGURE 5.10 – Exemple illustratif d'incapacité des critères non supervisés à mesurer le compromis description/prédiction

Ceci est dû au fait que la majorité de ces critères se basent sur une mesure de similarité (ou une distance) qui évalue la proximité entre les instances indifféremment de leurs classes d'appartenance. De ce fait, deux instances proches en termes de distance vont être considérées comme similaires bien qu'elles appartiennent à des classes différentes. On en conclut que, la recherche d'un critère permettant de mesurer le compromis entre la description et la prédiction s'avère nécessaire.

5.2.3 Vers la recherche d'un critère d'évaluation

Pour mesurer le compromis entre la description et la prédiction, le front de Pareto [20] peut être utilisé. C'est une technique qui permet de résoudre les problèmes d'optimisation multi-critères. Elle consiste à chercher un ensemble de solutions dites non dominées, parmi lesquelles on ne peut pas décider si l'une est meilleure que l'autre, aucune ne permet systématiquement de trouver l'optimum pour tous les critères. Cet ensemble est nommé l'ensemble de Pareto.

Soient $f_1, \ldots, f_Z$ des critères à minimiser et x_1 et x_2 deux solutions potentielles au problème multi-critères. La solution x_1 est dite dominée x_2 si :

$$\forall i \quad f_i(x_1) \leq f_i(x_2)$$
$$\text{avec au moins un } i \text{ tel que } f_i(x_1) < f_i(x_2)$$

- La solution x_1 est dite faiblement non dominée, s'il n'existe pas de solution x_2 telle que : $\forall i, f_i(x_1) \leq f_i(x_2)$.

- La solution x_1 est dite fortement non dominée, s'il n'existe pas de solution x_2 telle que : $f_i(x_1) \leq f_i(x_2)$ avec au moins un i tel que $f_i(x_1) < f_i(x_2)$

La figure 5.11 illustre le concept de dominance dans le cas d'un problème d'optimisation bicritère, un problème où on cherche à minimiser deux fonctions f_1 et f_2. Dans ce cas, les solutions représentées par les points A et B dominent la solution représentée par le point C. Par contre, les solutions représentées par les points A , B et D ne sont dominées par aucune solution. Les points représentatifs de ces solutions non dominées constituent le front de Pareto.

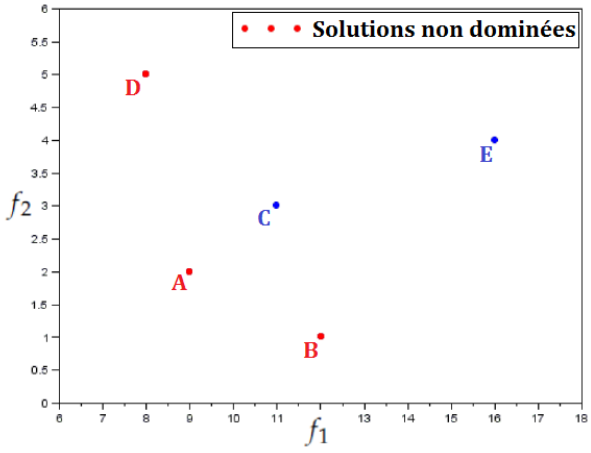

FIGURE 5.11 – Illustration de la notion de dominance pour un cas bicritère, f_1 et f_2 sont deux fonctions à minimiser. La solution A domine les solutions C et E

Dans notre cadre d'étude, les deux critères à optimiser sont notamment un critère supervisé (*e.g.,* ARI, VI) qui mesure la performance prédictive du modèle en question et un critère non

supervisé (*e.g.,* Indice Davies-Bouldin [38]) qui mesure la capacité de celui-ci à discerner des groupes d'instances compacts et éloignés les uns des autres.

À notre connaissance, il n'existe pas dans la littérature de critère analytique qui permette de mesurer la qualité des résultats issus du deuxième type du clustering prédictif. Cependant, certains chercheurs ont intégré dans leurs approches de clustering prédictif des fonctions objectifs cherchant à combiner deux quantités différentes (supervisée et non supervisée). Ces fonctions peuvent donc être un point de départ vers la recherche d'un nouveau critère d'évaluation.

Par exemple, Eick *et al.* [46] ont proposé une fonction objectif à minimiser pour l'algorithme des K-modes supervisé. Cette fonction se compose d'un critère supervisé ($Impurity(X)$) évaluant la capacité du modèle à bien classer les instances et d'un critère non supervisé ($penalty(K)$) pénalisant l'obtention d'un nombre maximal de clusters. La formule mathématique de cette fonction est donnée par l'équation 5.1 :

$$q(X) = Impurity(X) + \beta * penalty(K) \tag{5.1}$$

avec

$$Impurity(X) = \frac{Nombre\ d'instances\ mal\ classes}{N}$$

et

$$penalty(K) == \begin{cases} \sqrt{\frac{K-J}{n}}, & K \geq J \\ 0 & K < J \end{cases}$$

β est un paramètre utilisateur compris entre 0 et 2. Une grande valeur de β implique une large pénalité pour un nombre élevé de clusters. Si on se place dans le cadre du deuxième type du clustering prédictif, l'utilisation de cette fonction pour mesurer le compromis prédiction/description reste insuffisant. En effet, la proximité entre les paires d'instances en termes de distance n'est pas introduite dans celle-ci. De plus, les résultats obtenus via ce critère vont dépendre du choix de la valeur de *beta* qui n'est pas une tâche facile.

Peralta et al. ont proposée dans [85] une fonction objectif pour l'algorithme des K-moyennes supervisé écrite sous forme d'une combinaison convexe entre deux quantités différentes : l'une représente la fonction objectif usuelle de l'algorithme des K-moyennes standard et l'autre représente sa version supervisée. La formule mathématique de cette fonction est donnée par l'équation 5.2

$$J = \sum_{n=1}^{N} \left[\alpha \sum_{k=1}^{K} \sum_{l=1}^{J} \delta_{nk}^{l} \parallel x_n - u_k^l \parallel^2 \rho_k^l + (1-\alpha) \sum_{k=1}^{K} \delta_{nk} \parallel x_n - u_k \parallel^2 \right] \tag{5.2}$$

avec δ_{nk}^l est une fonction indicatrice supervisée qui assigne l'instance x_n au centre u_k^l désignant le centre de gravité du cluster k ayant comme classe l. ρ_k^l est un facteur défini pour les instances de classe l appartenant au cluster k. δ_{nk} est fonction indicatrice non supervisée qui assigne l'instance x_n au cluster k. Finalement, α est un paramètre utilisateur compris entre 0 et 1 gérant l'équilibre entre les deux scores (supervisé et non supervisé) du clustering. Comme pour la fonction proposée par Eick *et al.*, les résultats obtenus via cette fonction vont dépendre du choix de la valeur de α qui n'est pas une tâche facile.

5.3 Proposition d'un indice pour le clustering prédictif (Type 2)

5.3.1 Motivation

Pour mesurer le compromis entre la description et la prédiction, trois points doivent être respectés, à savoir : 1) la minimisation de l'inertie intra-clusters, 2) la maximisation de l'inertie inter-clusters et 3) la maximisation du taux de bonnes classifications. Un bon critère au sens du clustering prédictif est donc celui qui équilibre ces trois points et qui vérifie les contraintes, à savoir :

1. *L'interprétabilité* : le critère doit être facile à interpréter
2. *Le nombre de clusters* : le critère ne doit pas être trop biaisé par le nombre de clusters. En effet, ce critère doit être capable de comparer deux partitions ayant un nombre de clusters différent.
3. *Le nombre d'instances* : le critère ne doit pas être trop biaisé par le nombre d'instances dans chaque cluster. En effet, ce critère doit être capable de mesurer le degré de la compacité, la séparabilité et la pureté dans des clusters de différents effectifs. Par exemple, en se basant sur des proportions.
4. *La complexité* : le critère ne doit pas avoir une complexité trop supérieure à celle de l'algorithme des K-moyennes prédictives utilisé.
5. *Le bruit* : le critère doit être relativement stable en cas de perturbations aléatoires.

Lors de la recherche d'un nouveau critère pour le deuxième type du clustering prédictif, deux voies intuitives peuvent être exploitées, à savoir : *i*) la modification d'un critère supervisé à travers l'intégration d'une mesure qui évalue la proximité des paires d'instances, et *ii*) la modification d'un critère non supervisé à travers l'intégration d'une mesure qui relie la proximité des instances en termes de distance à leurs classes d'appartenance. Dans cette thèse, nous nous intéressons à l'étude de la deuxième voie. Il s'agit ici de modifier le critère d'évaluation Davies-Bouldin (DB) communément utilisé dans le cadre du clustering standard.

La raison de l'incapacité de l'indice de Davies-Bouldin à sélectionner le nombre optimal des clusters dans le cas de la non corrélation entre les classes et les clusters revient au fait que celui-ci utilise une métrique (ou une distance) non supervisée pour mesurer la ressemblance entre les instances. Cette métrique évalue la ressemblance entre instances en se basant sur leur proximité en termes de distance et sans accorder aucune importance à leurs classes d'appartenance. Par conséquent, des instances de classes différentes peuvent être considérées comme similaires si elles sont proches en termes de distance (cas d'une région dense possédant au moins deux classes). Pour permettre à l'indice Davies-Bouldin de surmonter ce problème, une intégration d'une mesure de similarité supervisée s'avère nécessaire. La section suivante propose donc une nouvelle mesure de similarité permettant de prendre en considération l'appartenance des instances aux différentes classes lors de l'évaluation de leur proximité.

5.3.2 Proposition d'une nouvelle mesure de similarité supervisée

Une similarité ou dissimilarité est définie comme étant toute application à valeurs numérique qui permet de mesurer le lien entre les individus d'un même ensemble. Pour une similarité, le lien entre deux individus sera d'autant plus fort que sa valeur est grande. Pour une dissimilarité le lien sera d'autant plus fort que sa valeur de la dissimilarité sera petite.

Définition :

Un opérateur de ressemblance $s : \mathcal{D}\times\mathcal{D} \rightarrow [0,1]$ défini sur l'ensemble d'individus $\mathcal{D} = \{X_i\}_{i=1}^{N}$ est un indice de similarité (ou similarité), s'il vérifie les propriétés suivantes :

1. *Symétrie* : $\forall X_i, X_j \in \mathcal{D} \quad s(X_i, X_j) = s(X_j, X_i)$
2. *Positivité* : $\forall X_i, X_j \in \mathcal{D} \quad s(X_i, X_j) \geq 0$
3. *Maximalité* : $\forall X_i, X_j \in \mathcal{D} \quad s(X_i, X_i) = s(X_j, X_j) \geq s(X_i, X_j)$

Il convient de noter ici que le passage de l'indice de similarité s à la notion duale d'indice de dissimilarité (que nous noterons Sim), est trivial. Étant donné s_{max} la similarité d'une instance avec elle-même ($s_{max} = 1$), il suffit de poser :

$$\forall X_i, X_j \in \mathcal{D} \quad Sim(X_i, X_j) = s_{max} - s(X_i, X_j) \tag{5.3}$$

Proposition d'une nouvelle mesure de dissimilarité

Dans le contexte supervisé, chaque instance $X_i = \{X_{i1}, \ldots, X_{id}\}_{i=1}^{N}$ possède d variables descriptives et une variable qualitative décrivant sa classe d'appartenance $Y_i = f(X_i)$. Pour évaluer la ressemblance entre deux paires d'instances étiquetées, quatre scénarios possibles peuvent être définis :

— **Scénario 1** : Proches (en termes de distance) et de même classe (Figure 5.12 A).
— **Scénario 2** : Proches (en termes de distance) et de classes différentes (Figure 5.12 C).
— **Scénario 3** : Éloignées (en termes de distance) et de même classe (Figure 5.12 B).
— **Scénario 4** : Éloignées (en termes de distance) et de classes différentes (Figure 5.12 D).

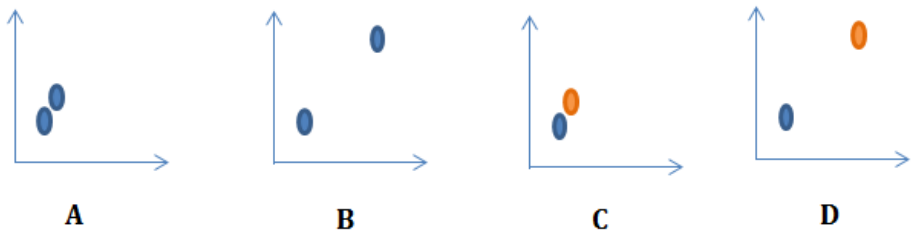

FIGURE 5.12 – Les 4 scénarios illustrant la ressemblance entre deux instances étiquetées

Suivant ces quatre scénarios, il est clair que la forte similarité « $s(X_i, X_j) = 1$ » ou la faible dissimilarité « $Sim(X_i, X_j) = 0$ » va correspondre au premier scénario (Figure 5.12 A) et la faible similarité (ou la forte dissimilarité) va correspondre au quatrième scénario (Figure 5.12 D). La mesure de similarité/dissimilarité proposée dans ce contexte doit prendre en compte les quatre scénarios cités ci-dessus.

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

$$\forall X_i, X_j \in \mathcal{D} \quad Sim(X_i, X_j) = 1 - \frac{exp(-\delta(X_i, X_j))}{1 + dist(X_i, X_j)^2} \tag{5.4}$$

avec δ est une fonction indicatrice présentée comme suit :

$$\delta(X_i, X_j) = \begin{cases} 0 & \text{si} \quad f(X_i) = f(X_j) \\ 1 & \text{si} \quad f(X_i) \neq f(X_j) \end{cases} \tag{5.5}$$

Il est à noter que la vraie classe $f(X_i)$ de l'instance X_i peut être remplacée par la classe prédite $\hat{f}(X_i)$ selon le besoin.

$dist(X_i, X_j)$ est la distance Euclidienne donnée par la formule suivante :

$$dist(X_i, X_j) = \|X_i - X_j\|_2 = \sqrt{\sum_{l=1}^{d}(X_{il} - X_{jl})^2}, \quad \forall X_i, X_j \in \mathcal{D} \tag{5.6}$$

Il est à noter que dans le cas de grandes dimensions, la quantité $dist(X_i, X_j)^2$ sera très grande. Ceci peut induire une stagnation au niveau de la mesure de similarité $s(X_i, X_j) = \frac{exp(-\delta(X_i, X_j))}{1+dist(X_i, X_j)^2}$. Pour pallier ce problème, l'utilisation des données normalisées sera utile pour une diminution de la quantité $dist(X_i, X_j)^2$. De plus, la mesure suivante peut être utilisée.

$$dist(X_i, X_j)^2 = \sum_{l=1}^{d} \frac{(X_{il} - X_{jl})^2}{d} \tag{5.7}$$

La quantité $s(X_i, X_j) = \frac{exp(-\delta(X_i, X_j))}{1+dist(X_i, X_j)^2}$ est bien une mesure de similarité qui vérifie les trois propriétés citées ci-dessus, à savoir la symétrie, la positivité et la maximalité. Elle prend ses valeurs dans l'intervalle $[0, 1]$ tout comme la mesure de dissimilarité $Sim(X_i, X_j)$ présentée dans l'équation 5.4 :

— $\forall X_i, X_j \in \mathcal{D} \quad Sim(X_i, X_j) = 0 \Leftrightarrow dist(X_i, X_j) = 0 \quad \textbf{ET} \quad X_i$ et X_j ont la même classe.
— $\forall X_i, X_j \in \mathcal{D} \quad Sim(X_i, X_j) = 1 \Leftrightarrow dist(X_i, X_j) \to \infty$.

À l'aide de cette nouvelle mesure de dissimilarité, deux instances sont considérées comme similaires, si et seulement si, elles sont proches en termes de distance et appartiennent à la même classe. Cependant, lorsqu'elles appartiennent à des classes différentes, leur proximité en termes de distance $\frac{1}{dist(A, X_i)^2+1}$ est pénalisée par le terme $exp(-1)$. À titre d'exemple, la partie milieu de la figure 5.13 présente l'influence de la classe sur les résultats obtenus par la mesure proposée. La courbe noire (*respectivement*, la courbe bleue) présente les valeurs de la mesure de dissimilarité proposée entre une instance A et d'autres instances de l'espace (voir la partie gauche de la figure 5.13) lorsque toutes les instances appartiennent à la même classe (respectivement, les instances ont une classe différente de celle de l'instance A).

Cette figure montre clairement l'impact de la classe sur les résultats obtenus par la mesure proposée. La partie droite de la figure 5.13, quant à elle, présente la distance euclidienne entre l'instance A et les autres instances de l'espace présenté dans la partie gauche de la figure 5.13. Visuellement, on remarque que la courbe obtenue par la distance euclidienne a la même allure que les courbes obtenues par la mesure proposée.

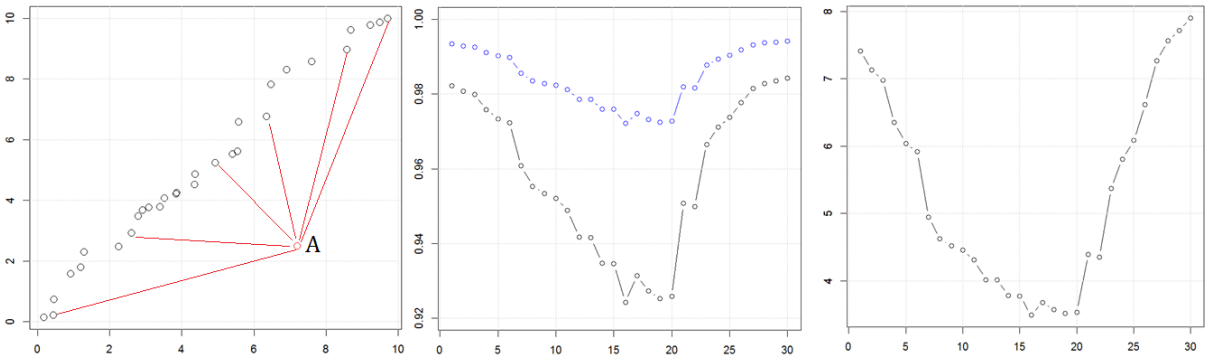

FIGURE 5.13 – Différence entre la mesure proposée (figure du milieu) et la distance Euclidienne (figure à droite).

5.3.3 La version supervisée de l'indice de Davies-Bouldin (SDB)

Avant d'intégrer la nouvelle mesure de dissimilarité proposée ci-dessus dans l'indice de Davies-Bouldin , il est important de le définir dans son contexte.

Rappel

L'indice Davies-Bouldin (DB) [38] traite chaque cluster individuellement et cherche à mesurer à quel point il est similaire au cluster qui lui est le plus proche. L'indice DB est décrit par la formule suivante :

$$DB = \frac{1}{K}\sum_{k=1}^{K} max_{1\leq k\neq t\leq K}\left\{\frac{S_k + S_t}{M_{kt}}\right\} \tag{5.8}$$

Pour chaque cluster $k \in \{1, \ldots, K\}$ de la partition, l'indice DB cherche le cluster t $(t \neq k)$ qui maximise la quantité R_{kt}, décrite par la formule suivante :

$$R_{kt} = \frac{S_k + S_t}{M_{kt}} \tag{5.9}$$

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. La formule mathématique de S_k est donnée par l'équation 5.10.

$$S_k = \left(\frac{1}{N_k}\sum_{i=1}^{N_k} \parallel X_i - G_k \parallel_p\right)^{\frac{1}{p}} \tag{5.10}$$

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 (voir équation 5.11).

$$M_{kt} = \parallel G_k - G_t \parallel_p \tag{5.11}$$

La mesure R_{kt} vérifie les trois propriétés suivantes :

1. $R_{kt} \geq 0$
2. **Si** $S_k \geq S_t$ **et** $M_{kt} = M_{tm}$ **alors** $R_{kt} > R_{tm}$

3. **Si** $S_k = S_m$ **et** $M_{kt} \leq M_{tm}$ **alors** $R_{kt} > R_{tm}$

À partir de ces propriétés, on constate que plus l'indice DB est minimal, plus les clusters formés sont compacts et éloignés les uns des autres.

Dans le cadre du clustering prédictif, une bonne partition au sens du clustering prédictif est celle qui fournit des groupes d'instances compacts, purs en termes de classe et éloignés les uns des autres. La nouvelle version du critère Davies-Bouldin, nommée **SDB** (Supervised Davies-Bouldin) doit être capable d'équilibrer les trois points suivants :

— L'inertie intra-clusters minimale (compacité)
— L'inertie inter-clusters maximale (séparabilité)
— Le taux de bonnes classifications est maximal (prédiction)

L'algorithme des K-moyennes prédictives cherche à former dans la phase d'apprentissage des groupes d'instances compacts, purs en termes de classes et éloignés les uns des autres dans le but de prédire ultérieurement la classe des nouvelles instances. 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 donnée par l'équation dans la quantité S_k comme suit :

$$S_k = \frac{1}{N_k} \sum_{i=1}^{N_k} Sim(X_i, G_k) \tag{5.12}$$

avec

$$Sim(X_i, G_k) = 1 - \frac{exp(-\delta_1(X_i, G_k))}{1 + dist(X_i, G_k)^2} \tag{5.13}$$

Le score S_k mesure le degré de ressemblance des instances du cluster k avec leur centre de gravité G_k comme le montrent l'équation 5.12 et la figure 5.14. Dans le cadre supervisé, cette ressemblance est évaluée en respectant les 4 scénarios discutés dans la section 5.3.2.

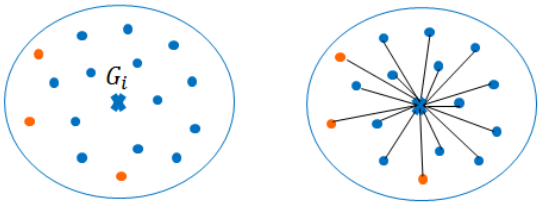

FIGURE 5.14 – Evaluation de la compacité et de la pureté pour un cluster donné.

L'utilisation de la nouvelle mesure de dissimilarité nécessite une connaissance des étiquettes des instances en question. Pour évaluer la pureté en termes de classes dans le cluster k, la vraie classe $f(X_i)$ de chaque instance X_i est comparée à la classe prédite (ou induite) pour le cluster k. Cette classe prédite est associé au centre de gravité G_k du cluster k, notée $\hat{f}(G_k)$. La fonction indicatrice δ_1 qui compare les étiquettes dans la nouvelle mesure de dissimilarité est donnée par l'équation 5.14.

$$\delta_1(X_i, G_k) = \begin{cases} 0 & \text{si} \quad f(X_i) = \hat{f}(G_k) \\ 1 & \text{si} \quad f(X_i) \neq \hat{f}(G_k) \end{cases} \tag{5.14}$$

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

Pour la séparabilité des clusters, la nouvelle mesure de dissimilarité est utilisée dans le but d'évaluer la ressemblance entre les centres de gravité comme le montre la figure 5.15. Cette ressemblance est mesurée en utilisant la nouvelle mesure de dissimilarité donnée par l'équation 5.15. Il est à rappeler que les étiquettes associées aux centres de gravité sont éventuellement les classes prédites pour les clusters.

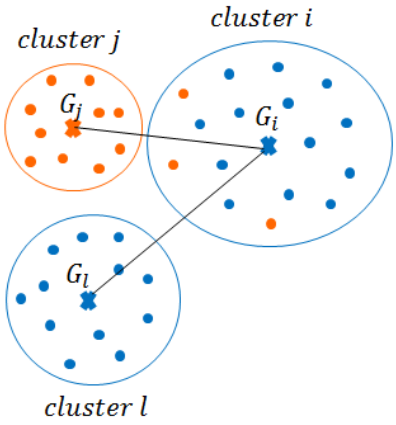

FIGURE 5.15 – Evaluation d'éloignement d'un cluster par rapport aux autres.

$$M_{kt} = Sim(G_k, G_t) = 1 - \frac{exp(-\delta_2(G_k, G_t))}{1 + dist(G_k, G_t)^2} \tag{5.15}$$

avec

$$\delta_2(G_k, G_t) = \begin{cases} 0 & \text{si} \quad \hat{f}(G_k) = \hat{f}(G_t) \\ 1 & \text{si} \quad \hat{f}(G_k) \neq \hat{f}(G_t) \end{cases} \tag{5.16}$$

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.

La version supervisée de l'indice Davies-Bouldin, nommée SDB prend ses valeurs dans l'intervalle $[0, +\infty[$ et est définie comme suit :

$$SDB = \frac{1}{K} \sum_{k=1}^{K} max_{1 \leq k \neq t \leq K} \left\{ \frac{S_k + S_t}{M_{kt}} \right\} \tag{5.17}$$

avec les scores S_k et M_{kt} sont donnés respectivement par les équations 5.14 et 5.15. Comme l'indice de Davies-Bouldin standard, 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.

Récapitulatif

Davies- Bouldin (DB)	**Davies-Bouldin supervisé (SDB)**
L'indice DB s'écrit sous la forme suivant : $DB = \frac{1}{K}\sum_{k=1}^{K} max_{1\leq k\neq t\leq K}\{R_{kt}\}$	L'indice SDB s'écrit sous la forme suivant : $SDB = \frac{1}{K}\sum_{k=1}^{K} max_{1\leq k\neq t\leq K}\{R_{kt}\}$
avec	avec
$R_{kt} = \frac{S_k+S_t}{M_{kt}}$	$R_{kt} = \frac{S_k+S_t}{M_{kt}}$
La compacité	**La compacité**
$S_k = \left(\frac{1}{N_k}\sum_{i=1}^{N_k} \parallel X_i - G_k \parallel_p\right)^{\frac{1}{p}}$	$S_k = \frac{1}{N_k}\sum_{i=1}^{N_k} Sim(X_i, G_k)$
	avec
X_i est une instance de dimension d	$Sim(X_i, G_k) = 1 - \frac{exp(-\delta_1(X_i,G_k))}{1+dist(X_i,G_k)^2}$
G_k est le centre de gravité du cluster k	$\delta_1(X_i, G_k) = \begin{cases} 0 & \text{si} \quad f(X_i) = \hat{f}(G_k) \\ 1 & \text{si} \quad f(X_i) \neq \hat{f}(G_k) \end{cases}$
La séparabilité	**La séparabilité**
$M_{kt} = \parallel G_k - G_t \parallel_p$	$M_{kt} = Sim(G_k, G_t) = 1 - \frac{exp(-\delta_2(G_k,G_t))}{1+dist(G_k,G_t)^2}$
	avec
	$\delta_2(G_k, G_t) = \begin{cases} 0 & \text{si} \quad \hat{f}(G_k) = \hat{f}(G_t) \\ 1 & \text{si} \quad \hat{f}(G_k) \neq \hat{f}(G_t) \end{cases}$
	$\hat{f}(G_k)$ est la classe prédite pour le cluster k
La plage de variation	**La plage de variation**
$[0; +\infty[$	$[0; +\infty[$

5.4 Expérimentation

Afin de vérifier la capacité de la version supervisée de l'indice Davies-Bouldin à mesurer le compromis entre la description et la prédiction et donc mesurer la qualité des résultats issus par les algorithmes de clustering prédictif, nous allons utiliser différents jeux de données : *i*) des jeux

de données contrôlés de petite dimension. Le nombre de variables descriptives dans ce cas est fixé à 2. Pour ces jeux de données, la structure interne de la variable cible est connue à l'avance. L'objectif ici est donc de connaître la capacité du critère modifié (SDB) à bien sélectionner le nombre optimal de clusters dans le cas de la corrélation et la non corrélation entre les clusters et les classes, *ii*) des jeux de données simulés de grandes dimensions. La structure interne de la variable cible pour chacun de ces jeux de données est également connue a priori. L'objectif ici est de mesurer la capacité du critère SDB à fournir de bons résultats (*i.e.,* compromis entre la description et la prédiction) y compris le cas de la grande dimensionnalité, *iii*) un jeu de données de grande dimension de l'UCI dont la structure interne de la variable cible n'est pas connue. Dans ce cas, on cherche à tirer des conclusions sur la capacité de SDB vis-à-vis du critère non supervisé DB et du critère supervisé ARI.

5.4.1 Sur des jeux de données contrôlés

Cas 1 : Non corrélation des classes et des clusters

Comme évoqué précédemment, dans le cas de la non corrélation entre les classes et les clusters, les critères non supervisés tel que DB n'arrivent pas à détecter le nombre de clusters optimal au sens du clustering prédictif. À titre d'exemple, le jeu de données présenté dans la partie gauche de la figure 5.16 est caractérisé par la présence de deux régions denses possédant deux classes (rouge et noire). Visuellement, pour ce jeu de données, la partition optimale au sens du clustering prédictif est celle qui contient 6 groupes. La partie milieu de la figure présente les valeurs des deux critères DB et SDB en fonction du nombre de clusters. Il est à signaler que les partitions sont obtenues en utilisant l'algorithme des K-moyennes standard précédé par la méthode d'initialisation S-Bisecting. Le critère supervisé SDB arrive à détecter la partition optimale pour ce jeu de données jouet tandis que le critère non supervisé DB n'arrive pas à détecter le nombre exacte des régions denses (*i.e.,* 4 régions).

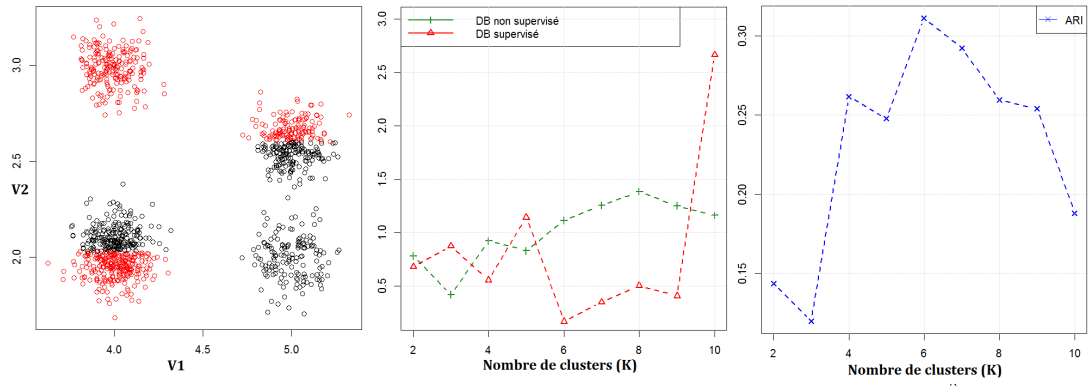

FIGURE 5.16 – Premier jeu de données jouet

L'une des propriétés importantes que le critère d'évaluation proposé doit posséder est la résistance au bruit qui peut exister dans les données. Le jeu de données présenté dans la partie gauche de la figure 5.17 est caractérisé par la présence de deux classes dont chacune possède deux sous-groupes différents. Un bon critère doit pouvoir les détecter malgré le bruit d'étiquetage existant. La partie milieu (respectivement la partie gauche) de la figure 5.17 présente les valeurs du critère SDB (respectivement, DB) lorsqu'on ajoute dans chaque classe 5%, 10%,20%,30 et 40% de bruits. Les résultats obtenus montrent que le critère modifié SDB arrive facilement à

détecter, pour ce jeu de données, le nombre optimal de clusters quel que soit la quantité du bruit intégrée. Cependant le critère DB n'arrive pas à atteindre l'objectif recherché (*i.e.,* la détection du nombre optimal de clusters qui est dans ce cas 4 clusters).

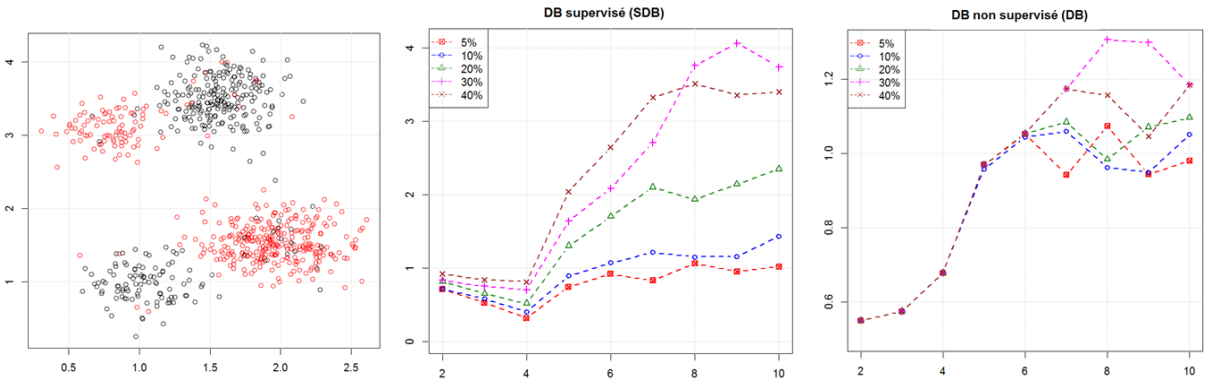

FIGURE 5.17 – Cas d'existence du bruit dans les données.

Cas 2 : Corrélation entre les classes et les clusters

Dans le cas de la corrélation entre les classes et les clusters, les critères non supervisés deviennent alors plus adaptés pour détecter la meilleure partition. Cependant, dans le cas où des sous-groupes différents de même classe sont proches, les critères supervisés tel que l'ARI n'arrivent pas à détecter le nombre optimale de clusters au sens du clustering prédictif puisqu'ils cherchent plutôt à optimiser l'axe de prédiction tout en ignorant l'axe de description. À titre d'exemple, pour les deux jeux de données situés dans la partie gauche des deux figures 5.18 et 5.19, le critère ARI n'arrive pas à détecter le nombre optimal de clusters au sens du clustering prédictif (voir respectivement la partie droite des deux figures 5.18 et 5.19). Le critère non supervisé DB arrive à détecter le nombre optimal du premier jeu de données (voir la courbe verte du graphique situé au milieu des deux figures). Le critère modifié SDB, quant-à-lui, arrive à sélectionner le nombre de clusters optimal dans les deux cas (voir la courbe rouge du graphique au milieu des deux figures).

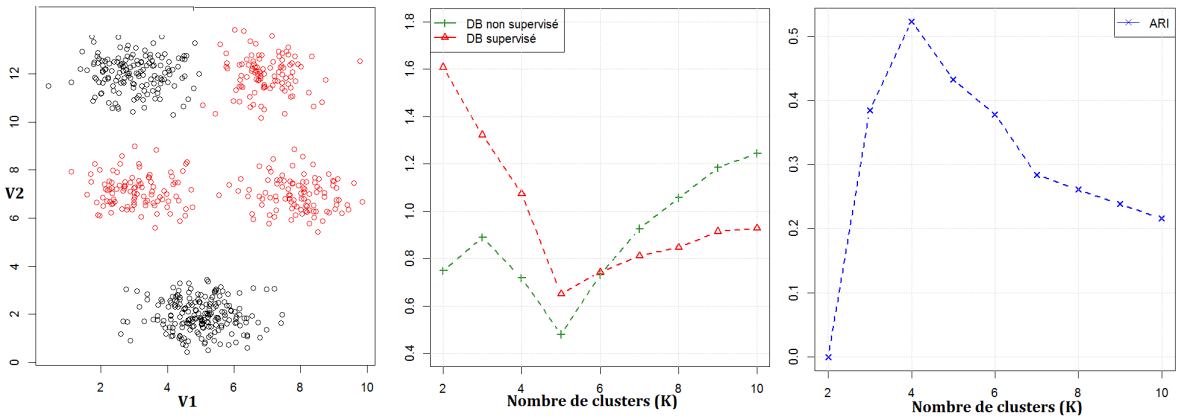

FIGURE 5.18 – Deuxième jeu de données jouet

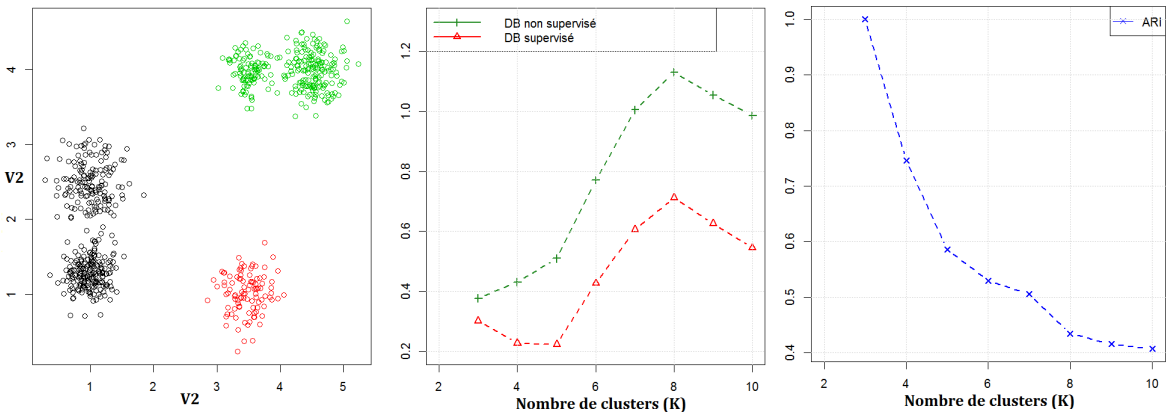

FIGURE 5.19 – Troisième jeu de données jouet

5.4.2 Sur des bases de données simulées de grandes dimensions

Pour valider le comportement du critère modifié vis-à-vis de l'évaluation du compromis description/prédiction dans le cas de la grande dimensionnalité, nous allons simuler quelques jeux de données de façon à obtenir une connaissance *a priori* sur la structure interne de la variable cible. La méthodologie suivie pour obtenir les jeux de données présentés dans ce qui suit est la suivante :

1. générer K_{opti} centres provisoires dans l'espace.
2. Pour chaque centre k_i, générer N_k instances de façon à ce qu'elles soient très proches de leur centre de gravité provisoire.
3. Répéter l'étape 2. pour le reste des centres provisoires.
4. Pour chaque groupe, attribuer une classe.

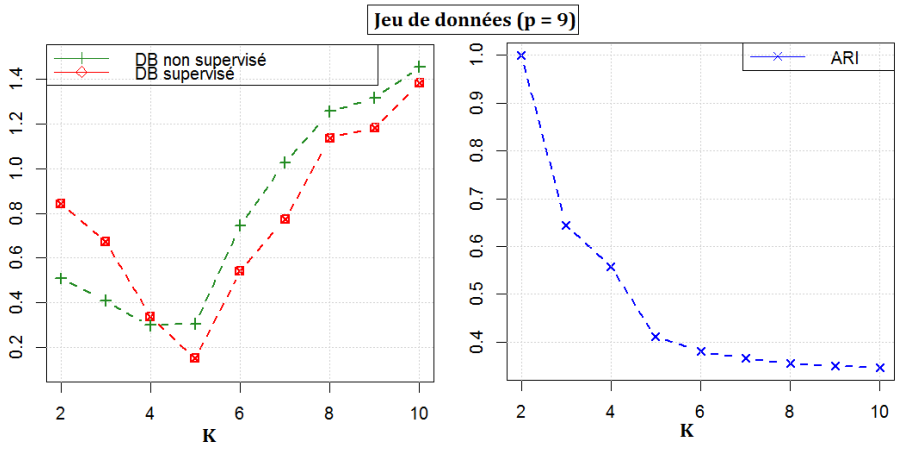

FIGURE 5.20 – Jeu de données contenant 9 variables descriptives, 3 classes et 5 sous-groupes à découvrir

Le premier jeu de données simulé est un jeu de données caractérisé par la présence de 825 instances, 9 variables descriptives et une variable à prédire contenant trois classes à prédire. La première classe contient 2 sous-groupes différents tandis que la deuxième classe contient 3 sous-groupes (*i.e.*, $Kopti = 5$). La figure 5.20 présente les valeurs des critères en fonction du nombre

de clusters. On constate que le critère supervisé ARI n'arrivent pas à détecter le nombre optimal de clusters tandis que les deux critères DB et SDB arrivent à le détecter.

Le deuxième jeu de données simulé est un jeu de données 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.,* $Kopti = 5$). Les résultats présentés dans la figure 5.21 montrent que les deux critères ARI et DB n'arrivent pas à détecter le nombre optimal de clusters, tandis que le critère modifié SDB arrive facilement à le détecter.

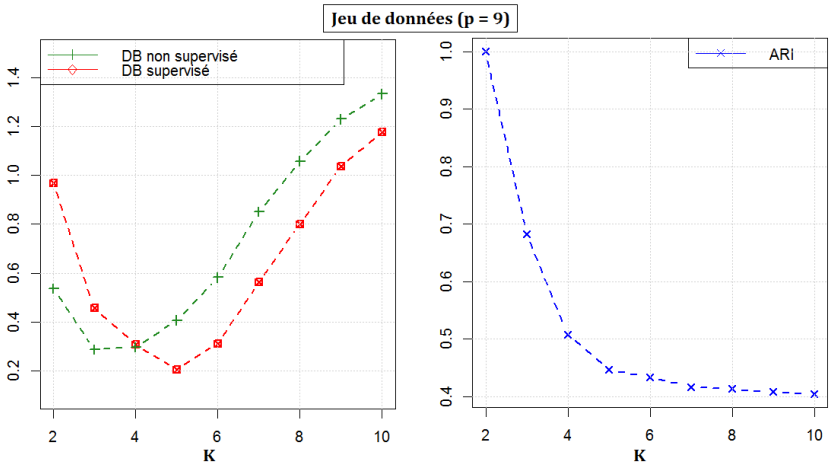

FIGURE 5.21 – Jeu de données contenant 9 variables descriptives, 2 classes et 5 sous-groupes à découvrir

Pour le troisième jeu de données, nous avons augmenté la dimensionnalité. Ce dernier est 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.,* $Kopti = 4$). La figure 5.22 montre que le critère ARI est incapable de détecter le nombre optimal de clusters tandis que les critères DB et SDB y arrivent.

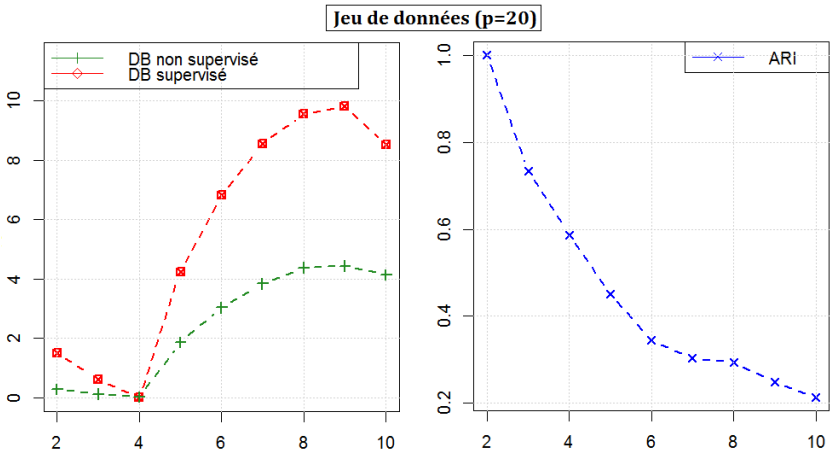

FIGURE 5.22 – Jeu de données contenant 20 variables descriptives, 2 classes et 4 sous-groupes à découvrir

5.4.3 Sur des données de l'UCI

D'après les résultats expérimentaux obtenus dans les sections précédentes, on remarque que le nombre optimal de clusters détecté par la version supervisée de l'indice de Davies-Bouldin, ne doit pas être inférieur au nombre optimal détecté par le critère supervisé ARI **et** par l'indice standard de Davies-Bouldin. En effet, le critère ARI peut fusionner deux sous-groupes différents s'ils sont de la même classe et le critère DB peut fusionner deux sous-groupes proches de classes différentes.

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 la base de données Adult de l'UCI. Cette base est constituée de 48842 instances, 15 variables descriptives et une variable à prédire contenant 2 classes ("more" et "less").

La figure 5.23 présente les valeurs des critères SDB (partie gauche), DB (partie milieu) et ARI (partie droite) en fonction du nombre de clusters. Dans cette étude expérimentale, le nombre de clusters varie de $J = 2$ (J : nombre de classes) jusqu'à 10. Les partitions sont obtenues en utilisant toujours l'algorithme des K-moyennes précédé par le prétraitement Rank normalization pour les variables continues et Basic Grouping pour les variables catégorielles et par la méthode d'initialisation S-Bisecting. On constate que SDB et ARI détectent le même nombre de clusters tandis que l'indice DB indique qu'il n'existe pas une structure interne à découvrir dans la variable cible. Ceci peut être expliqué par le cas de la non corrélation entre les clusters et les classes comme déjà évoqué.

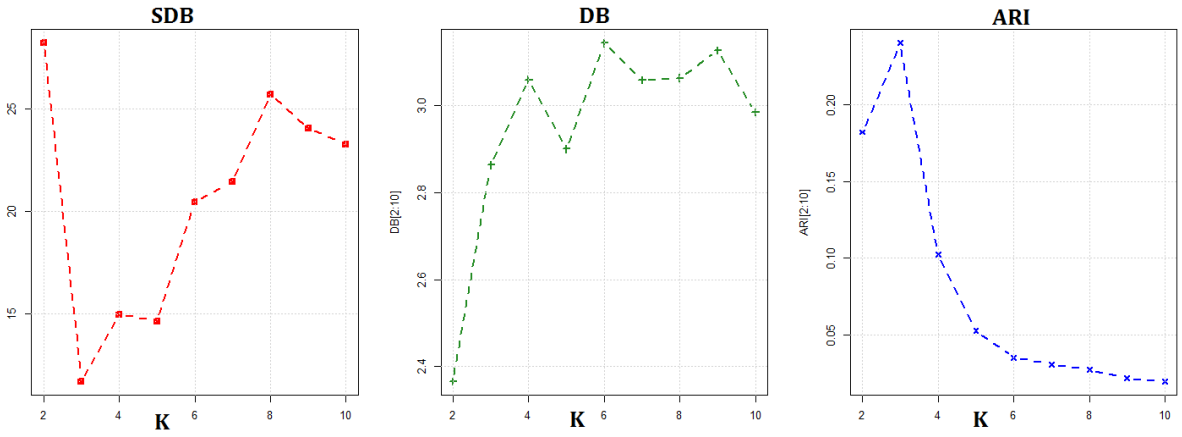

FIGURE 5.23 – Les valeurs des trois critères d'évaluation pour le jeu de données Adult en utilisant Rank Normalization et Basic grouping

5.5 Bilan

Ce chapitre a présenté une version supervisée de l'indice Davies-Bouldin, nommée SDB (Supervised Davies-Bouldin). Cet indice est basé sur une nouvelle mesure de similarité supervisée permettant d'établir une certaine relation entre la proximité des instances en termes de 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