Distance dépendante de la classe

Sommaire

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

[10] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols. Supervised pre-processings are useful for supervised clustering. In *Springer Series Studies in Classification, Data Analysis, and Knowledge Organization*, Bremen, 2015.

3.1 Introduction

La question évoquée dans cette thèse "*comment décrire et prédire d'une manière simultanée ?*" conduit à un nouvel aspect de l'apprentissage. Ce dernier est appelé *"le clustering prédictif"* (voir Algorithme 2 du Chapitre 2). Les algorithmes appartenant à ce type d'apprentissage peuvent être catégorisés en fonction des besoins de l'utilisateur. La première catégorie regroupe l'ensemble des algorithmes de clustering "modifiés" permettant de prédire correctement la classe des nouvelles instances sous contrainte d'avoir un nombre minimal de clusters dans la phase d'apprentissage (voir la partie gauche de la figure 3.1). Cette catégorie met l'accent sur l'axe de prédiction tout en "ignorant" l'axe de description. La deuxième catégorie regroupe l'ensemble des algorithmes permettant tout d'abord de découvrir la structure interne *complète* de la variable cible, puis munie de cette structure prédire correctement la classe des nouvelles instances (voir la partie droite de la figure 3.1). Contrairement aux algorithmes de la première catégorie, ces algorithmes cherchent à réaliser un compromis entre la description et la prédiction sans privilégier un axe par rapport à l'autre.

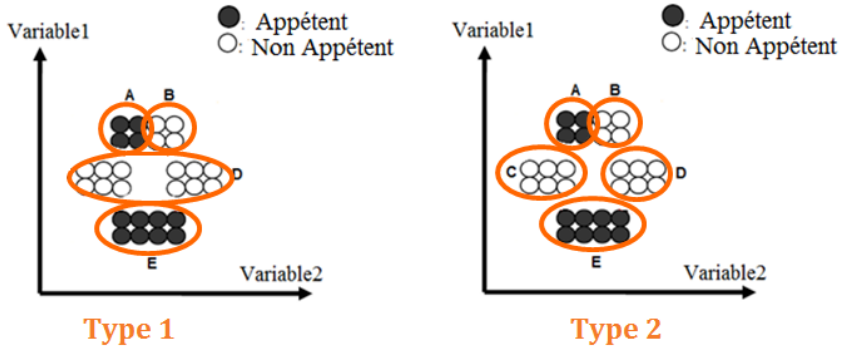

FIGURE 3.1 – Les deux types des algorithmes de clustering prédictif

Les algorithmes classiques de clustering visent à subdiviser l'ensemble des instances en un certain nombre de groupes (ou clusters) de telle sorte que les instances au sein de chaque cluster doivent être similaires entre elles et dissimilaires des instances des autres clusters. Pour l'algorithme des K-moyennes, cette notion de similarité/dissimilarité est représentée par une distance ou une métrique. Deux instances sont considérées comme similaires si et seulement si elles sont proches en termes de distance (la distance la plus utilisée est la distance Euclidienne). En revanche, dans le cadre des K-moyennes prédictives, la similarité entre deux instances n'est pas uniquement liée à leur proximité en termes de distance mais elle est également liée à leur ressemblance en termes de leur classe d'appartenance : deux instances sont similaires si et seulement si elles sont proches en termes de distance **et** appartiennent à la même classe. De ce fait, l'utilisation d'une distance ne prenant pas en compte l'information de la classe reste insuffisante : deux instances proches en termes de distance vont être considérées comme similaires indifféremment à leur classe d'appartenance et donc la probabilité que l'algorithme les regroupe ensemble sera élevée. À titre d'exemple, les instances appartenant aux deux groupes A et B de la figure 3.1.

Pour permettre à l'algorithme de K-moyennes classique de prendre en considération l'information donnée par la variable cible et ainsi améliorer sa performance prédictive, deux voies peuvent être exploitées. La première voie consiste à modifier la fonction du coût de l'algorithme des K-moyennes afin de proposer une nouvelle fonction objectif capable d'établir une certaine relation entre la similarité classique pour les instances et leur classe d'appartenance. À titre d'exemple, Peralta et al. ont défini dans [85] une nouvelle fonction objectif qui s'écrit sous forme

d'une combinaison convexe entre la fonction objectif usuelle de l'algorithme de K-moyennes et sa version supervisée. Cependant, cette fonction nécessite un paramètre utilisateur pour équilibrer les deux scores ce qui requière une phase d'ajustement (validation croisée pour trouver la bonne valeur du paramètre). La deuxième voie consiste à incorporer l'information donnée par la variable cible dans les données sans modifier la fonction du coût de l'algorithme des K-moyennes classique. Dans ce chapitre, on s'intéresse exclusivement à l'étude de la deuxième voie.

La démarche suivie pour incorporer l'information donnée par la variable cible dans les données doit impérativement respecter le point crucial qu'un algorithme des K-moyennes prédictives doit posséder, à savoir, l'interprétabilité des résultats (voir la section 2.6.1 du chapitre 2). L'intérêt de cette démarche est d'une part de rendre la tâche d'interprétation des résultats plus aisée pour l'algorithme des K-moyennes standard. D'autre part, elle permet de modifier indirectement la fonction du coût de l'algorithme des K-moyennes standard dans le but de l'aider à atteindre l'objectif des K-moyennes prédictives. Pour atteindre cet objectif, la méthode recherchée doit être capable de générer une distance, dite supervisée, permettant d'établir une certaine relation entre la similarité classique entre les instances et leur classe d'appartenance.

Dans ce chapitre, on suppose qu'une estimation des distributions uni-variées conditionnelles aux classes $P(\mathcal{X}|C)$ pourrait aider l'algorithme des K-moyennes standard à atteindre l'objectif mentionné ci-dessus. Parmi les méthodes permettant cette estimation probabiliste, on choisit de s'orienter vers les méthodes les plus interprétables, à savoir : les méthodes supervisées de discrétisation pour les variables continues et les méthodes supervisées de groupage en modalités pour les variables catégorielles. Ces méthodes cherchent à trouver la partition des valeurs ou des modalités qui donne le maximum d'informations sur la répartition des J classes connaissant les intervalles de discrétisation ou les groupes de modalités. A titre d'exemple, la figure 3.2 présente la discrétisation supervisée des deux variables continues (Variable 1 et Variable 2) pour un jeu de données caractérisé par la présence de deux classes. Pour cet exemple illustratif, la variable 1 est divisée en 3 intervalles tandis que la variable 2 est divisée en 4 intervalles.

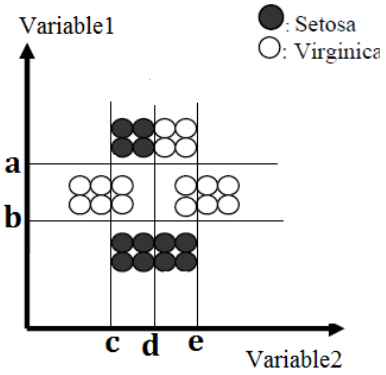

FIGURE 3.2 – Discrétisation supervisée

Une façon intuitive permettant d'exploiter les informations générées par les méthodes de discrétisation et de groupage en modalités est de considérer pour chaque instance l'appartenance de ses valeurs à un intervalle ou à un groupe de modalités (voir, Figure 3.3). Suivant cette démarche, chaque instances X_i sera transformée en $\sum_{l=1}^{d} t_l$ (t_l est le nombre d'intervalles ou groupes de modalité issu de la variable l et d est le nombre des variables descriptives) variables booléennes. Dans ce cas, pour mesurer la similarité entre les instances, la distance de Hamming est utilisée. Cette distance vérifie la propriété suivante : *deux instances proches en termes de distance sont également proches en termes de leurs appartenances aux mêmes intervalles de*

discrétisation (ou aux groupes de modalités selon la nature des variables descriptives). Cette méthode est appelée par la suite **Binarization** (BIN-BIN). Pour plus de détails, voir la section 3.2.2 de ce chapitre.

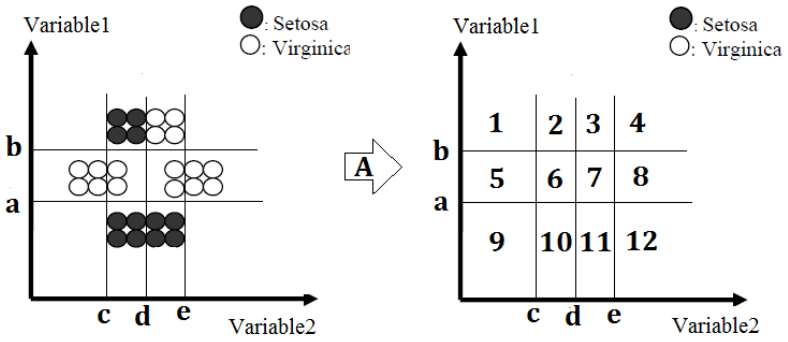

FIGURE 3.3 – L'exploitation des données à travers l'appartenance aux intervalles de discrétisation

Une autre façon permettant d'exploiter les informations générées par les méthodes de discrétisation et de groupage en modalités est de considérer la quantité d'informations contenue dans les intervalles ou dans les groupes de modalités conditionnellement aux classes "$P(X \in I|C_j)$", $j \in \{1, \ldots, J\}$" où I est le nombre d'intervalles ou de groupes de modalités pour une variable donnée (voir Figure 3.4). Suivant cette démarche, chaque instance X_i sera transformée en $J \times d$ variables numériques : chacune des d variables d'origines est transformée en J synthétiques variables $(log(P(X_i \in I|C_1)), \ldots, log(P(X_i \in I|C_J)))$. Dans ce cas, pour mesurer la similarité entre les instances, la distance bayésienne est utilisée. Cette distance vérifie la propriété suivante : *deux instances proches en termes de distance sont également proches en termes de leur probabilité d'appartenir à la même classe.* Cette méthode est appelée par la suite **Conditional Info** (CI-CI). Pour plus de détails voir la section 3.2.3 de ce chapitre.

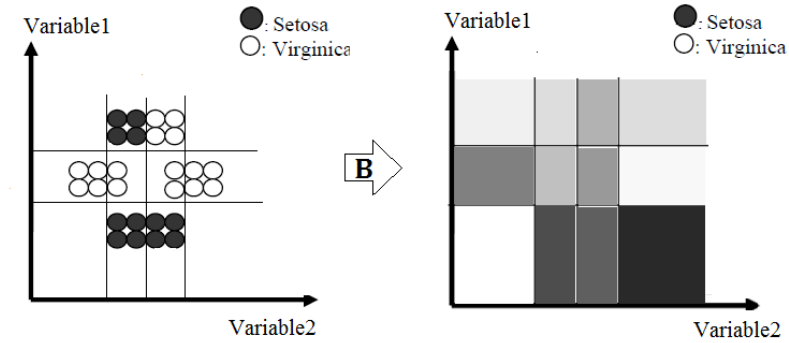

FIGURE 3.4 – L'exploitation des données à travers le calcul de la quantité d'information contenue dans les intervalles

Le reste de ce chapitre est organisé comme suit : la section 3.2.1 définit le principe de la discrétisation pour les variables continues et du groupage en modalités pour les variables catégorielles. Afin d'exploiter les informations générées par ces méthodes, les deux sections 3.2.2 et 3.2.3 présentent respectivement l'approche Binarization (BIN-BIN) et l'approche Conditional Info (CI-CI). Pour étudier l'impact de l'utilisation d'une distance supervisée (à travers l'utilisation du prétraitement supervisé BIN-BIN ou CI-CI) sur la qualité (au sens du clustering prédictif) des résultats fournis par l'algorithme des K-moyennes standard, une étude expérimen-

tale sera menée dans la section 3.4. L'objectif de cette section est de chercher à répondre à la question suivante : *les méthodes de prétraitement supervisées pourraient-elles aider l'algorithme des K-moyennes standard à atteindre l'objectif du clustering prédictif ?* Finalement, et avant de conclure dans la section 3.6, les deux axes de description et d'interprétation seront discutés dans la section 3.5.

3.2 Distance dépendante de la classe

L'incapacité de l'algorithme des K-moyennes standard à atteindre l'objectif des algorithmes du clustering prédictif s'illustre essentiellement dans le cas de la non corrélation entre les classes et les clusters. C'est le cas où au moins l'une des régions denses est caractérisée par la présence d'au moins deux classes. La figure 3.5 présente un exemple illustratif de la présence d'une région dense ($\{A \cup B\}$) contenant deux classes. En effet, l'algorithme des K-moyennes standard considère deux instances proches en termes de distance comme similaires indifféremment de leur classe d'appartenance. Dans le cadre du clustering prédictif, ceci peut générer une détérioration au niveau de la performance prédictif du modèle.

FIGURE 3.5 – Cas de la non corrélation entre les classes et les clusters

Pour surmonter ce problème, l'incorporation de l'information donnée par la variable cible dans les données s'avère nécessaire. Cette incorporation pourrait aider l'algorithme des K-moyennes standard à prendre en considération l'appartenance des instances aux classes. Dans ce chapitre, on suppose que l'estimation des distributions uni-variées conditionnelles aux classes ($P(\mathcal{X}|C)$) pourrait aider cet algorithme à atteindre l'objectif souhaité.

3.2.1 Estimation des densités conditionnelles aux classes

Comme signalé dans la section 2.6.1 du chapitre 2, les algorithmes du clustering prédictif sont des algorithmes qui fournissent des résultats facilement interprétables par l'utilisateur. De ce fait, lors de la recherche de la méthode permettant d'insérer l'information donnée par la classe cible dans les données, la contrainte d'interprétabilité doit impérativement être respectée.

Parmi les méthodes les plus "*interprétables*" permettant une estimation des densités conditionnellement aux classes, on trouve les méthodes de discrétisation supervisées pour les variables continues et le groupage en modalités pour les variables catégorielles.

Pour les variables continues, la discrétisation supervisée consiste à diviser le domaine de chaque variable en un nombre fini d'intervalles identifiés chacun par un code $I_l, l \in \{1, \ldots, t\}$. Elle vise à trouver la partition des valeurs qui donne le maximum d'informations sur la répartition des J classes connaissant l'intervalle de discrétisation $I_l, l \in \{1, \ldots, t\}$. Cette partition optimale est décrite par la table de contingence comme illustrée dans le tableau 3.1.

	I_1	I_2	...	I_t	Somme
C_1	n_{11}	n_{12}	...	n_{1t}	$n_{1.}$
C_2	n_{21}	n_{22}	...	n_{1t}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_J	n_{J1}	n_{J2}	...	n_{Jt}	$n_{J.}$
Somme	$n_{.1}$	$n_{.2}$	...	$n_{.t}$	N

TABLE 3.1 – Table de contingence pour une variable continue

Dans la littérature, il existe une variété de méthodes de discrétisation ou de groupage en modalités selon la nature des variables descriptives. Les méthodes les plus répandues sont notamment MODL [23], ChiSplit [21], BalancedGain [67] et l'arbre de décision C4.5 ou CART [87]. Le choix de la méthode qu'on va utiliser sera conditionnée par certains points, à savoir : la robustesse, la rapidité, la précision et finalement la minimisation des connaissances *a priori* (*i.e.,* pas (ou très peu) de paramètres utilisateur). Parmi les méthodes de discrétisation et de groupages en modalités qui respectent l'ensemble de ces points, on trouve la méthode MODL, proposé par Boullé dans [23]. Il est important de noter que ce choix n'est pas une obligation. D'autres méthodes peuvent également être utilisées à condition qu'elles respectent les points cités ci-dessus.

L'approche MODL considère la discrétisation comme un problème de sélection de modèle. Ainsi, une discrétisation est considérée comme un modèle paramétré par le nombre d'intervalles, leurs bornes et les effectifs des classes cibles sur chaque intervalle. La famille de modèles considérée est l'ensemble des discrétisations possibles. Cette famille est dotée d'une distribution a priori hiérarchique et uniforme à chaque niveau. Pour plus de détails sur cette approche voir [23].

Après cette étape dite étape de préparation supervisée des données, chaque variable est représentée par une table de contingence. Pour pouvoir exploiter les connaissances utiles existant dans ces tables, la recherche d'une méthode de recodage s'avère nécessaire. Cette méthode doit être capable de générer une distance dépendante de la classe permettant d'établir une certaine relation entre la similarité usuelle et le comportement des instances vis-à-vis de la classe cible.

Une façon permettant une exploitation aisée des tables de contingence est de considérer l'appartenance des valeurs de chaque instance aux intervalles de discrétisation ou aux groupes de modalité selon la nature des variables descriptives. Cette approche est appelée dans ce qui suit **Binarization** (BIN-BIN).

3.2.2 Binarization (BIN-BIN) - Distance de Hamming

L'approche la plus intuitive permettant d'exploiter les informations données par les tables de contingences est la considération de l'appartenance des valeurs de chaque instance aux intervalles de discrétisation ou aux groupes de modalités (à titre d'exemple, voir la figure 3.3). Il s'agit ici de transformer chaque variable en t variables booléennes ; où t est le nombre d'intervalles (ou groupes de modalités) généré par la méthode de discrétisation (ou de groupage en modalités). Cette opération de transformation est basée sur un recodage disjonctif complet : la variable synthétique prend 1 comme valeur si la valeur de la variable d'origine appartient à l'intervalle en question et elle prend zéro comme valeur dans le cas contraire. À titre d'exemple, la variable 1 du jeu de données présenté dans la figure 3.6 est transformée en 3 (nombre d'intervalles : $]-\infty, a]$,$]a, b]$ et $]b, +\infty]$) variables booléennes ($\{1, 0, 0\}$, $\{0, 1, 0\}$ et $\{0, 0, 1\}$) avec $\{1, 0, 0\}$ signifie que la valeur de la variable d'origine appartient au premier intervalle. Suivant ce processus, chaque instance

X_i sera transformée en un vecteur booléen de dimension $\sum_{l=1}^{d} t_l$. Cette approche est nommée **Binarization** (BIN-BIN).

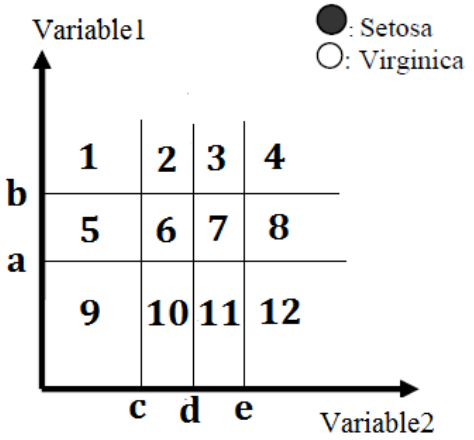

FIGURE 3.6 – Exemple illustratif d'appartenance aux intervalles après une opération discrétisation

Étant donné que les nouvelles variables synthétiques après l'opération de recodage sont des variables booléennes, une distance entre les individus appropriée à utiliser à ce stade est la distance de Hamming. On définit la distance de Hamming sur les attributs recodés comme le nombre d'attributs recodés de façon différentes. La formule mathématique de cette distance pour deux instances X_1 et X_2 de dimension p est donnée par l'équation 3.1.

$$\forall X_{1j}, X_{2j} \in \mathcal{D}, (j \in \{1, \ldots, d\}) \quad dist_H(X_{1j}, X_{2j}) = \#\{j : X_{1j} \neq X_{2j}\} \tag{3.1}$$

À partir de ce recodage, on constate que deux instances ayant une distance d_H nulle seront alors associées à la même prédiction de classe cible. À titre d'exemple, les instances appartenant au block 5 de la figure 3.6 sont quasiment de la même classe en raison de l'utilisation d'une méthode de discrétisation supervisée et elles sont présentées sous forme d'un vecteur de dimension $7 = 3+4$ prenant la forme suivante : $\{0, 1, 0, 1, 0, 0, 0\}$. De ce fait, la conclusion qui peut être tirée est que les distances faibles sont corrélées avec des instances ayant des comportements similaires vis-à-vis de la classe cible.

Avantages

— La méthode Binarization est une méthode capable de distinguer les instances selon leurs appartenances aux intervalles de discrétisation ou aux groupes de modalités de telle sorte que deux instances proches en termes de distance vont être proches en termes de leurs appartenances aux intervalles. On remarque que plus la distance est petite plus le comportement des instances vis-à-vis de la classe sera proche.
— La distance Hamming utilisée pour la méthode Binarization a l'avantage de se calculer simplement sous forme d'une distance L1 suite à un recodage binaire disjonctif complet sur chacun des attributs.
— Elle a également l'avantage d'une normalisation par rapport à la distance euclidienne.

Limites

— La qualité de la corrélation entre la distance de Hamming et le comportement en prédiction est difficile à évaluer.
— La distance de Hamming est peu discriminante : deux recodages différents peuvent correspondre à des comportements très ou peu différents vis-à-vis de la classe cible. À titre d'exemple, la figure 3.7 présente une discrétisation en 6 intervalles de la 7^{eme} variable "V7" de la base de données Waveform [2] pour un problème à trois classes (0, 1 et 2). L'axe des abscisses de la figure 3.7 représente l'ensemble des valeurs que peut prendre la variable V7. L'axe des ordonnées quant à lui, représente l'ensemble des probabilités d'appartenir à une des classes conditionnellement aux intervalles de discrétisation conditionnellement (par exemple, la courbe orange correspond à la classe 2). Les barres verticales désignent les intervalles de discrétisation (au nombre de 6). Clairement, on souhaiterait que deux instances recodées sur le $3ième$ et le $4ième$ intervalle (mélanges presque homogènes des trois classes) soient considérées comme plus proches que deux instances recodées sur le 1er intervalle (classe 2 largement majoritaire) et le $6ième$ intervalle (classe 2 absente) ce qui n'est pas le cas lorsque le recodage Binarization est utilisé.

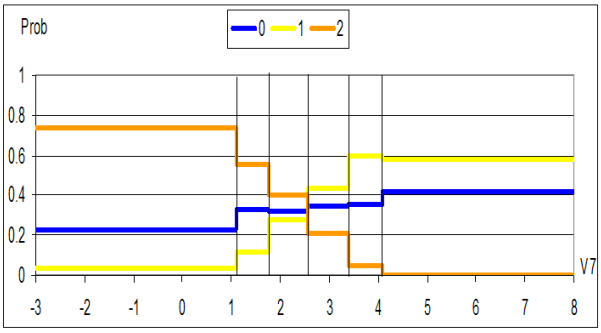

FIGURE 3.7 – Discrétisation de la variable V7 de la base Waveform via l'approche MODL

Pour surmonter les difficultés rencontrées par l'approche Binarization, la section 3.2.3 présente une nouvelle approche de recodage qui prend en considération la quantité d'information existant dans les intervalles de discrétisation (ou groupes de modalités) conditionnellement aux classes. Cette approche est appelée **Conditional Info** (CI-CI).

3.2.3 Conditional Info (CI-CI) - Distance bayésienne

Après la préparation supervisée des données, une seconde étape est mise en place dans le but d'exploiter les informations données par la discrétisation et le groupage en modalités. Cette étape est une étape de recodage où chaque variable de X_i est recodée en une variable qualitative contenant I_J valeurs de recodage. Chaque instance X_i ($i \in \{1, \ldots, N\}$) des données est alors recodée sous forme d'un vecteur de modalités discrètes $\hat{X}_i = X_{i1_1}, \ldots, X_{i1_J}, \ldots, X_{id_1}, \ldots, X_{id_J}$ où X_{id_J} représente la valeur du recodage de la variable d pour la classe J ($X_{id_J} = log(P(X_{id}|C_J))$). Ainsi, les variables de départ sont alors toutes représentées sous une forme numérique, sur un vecteur de $d \times J$ composantes : $log(P(X_{im}|C_J)), i \in \{1, \ldots, N\}, m \in \{1, \ldots, d\}$. Les deux étapes (discrétisation supervisée et recodage) forment une méthode de prétraitement supervisé des données que l'on appelle '*Conditional Info*' (CI-CI).

2. La base de données Waveform est une base de données de l'UCI caractérisée par la présence de 21 variables descriptives et une variable cible contenant 3 classes (0, 1 et 2).

Par exemple, pour un problème de classification binaire (*i.e.,* $J = 2$), la méthode de prétraitement Conditional Info transforme chaque instance X_i ($i \in \{1, \ldots, N\}$) en un vecteur $\hat{X}_i$ de $2 \times d$ composantes de la manière suivante :

$$\hat{X}_i = (log(P(X_{i1} \in I_k|C_1), log(P(X_{i1} \in I_k|C_2), \ldots, log(P(X_{id} \in I_k|C_1), log(P(X_{id} \in I_k|C_2))$$

avec I_k ($k \in \{1, \ldots, t\}$) représente l'intervalle de discrétisation obtenu dans la première étape de Conditional Info.

Soit $D = \{(X_i, Y_i)\}_1^N$ un ensemble d'apprentissage de taille N, avec $X_i = \{X_{i1}, \ldots, X_{id}\}$ est un vecteur de d variables et $Y_{i \in \{1,\ldots,N\}} \in \{C_1, \ldots, C_J\}$ est une variable cible composée de J classes. Le prétraitement Conditional Info permet d'écrire une distance bayésienne dépendante de la classe, notée $dist_B^p$, pour la norme ℓ_p. La formule mathématique de cette distance entre deux instances $\hat{X}_1$ et $\hat{X}_2$ est définie de la manière suivante :

$$dist_B^p(\hat{X}_1, \hat{X}_2) = \sum_{j=1}^{J} \left\| log(P(\hat{X}_1|C_j)) - log(P(\hat{X}_2|C_j)) \right\|_p \tag{3.2}$$

avec $\|.\|_p$ est la distance de Minkowski :

$$\left\| log(P(\hat{X}_1|C_j)) - log(P(\hat{X}_2|C_j)) \right\|_p = \sqrt[p]{\sum_{m=1}^{d} \left| log(P(X_{1m}|C_j)) - log(P(X_{2m}|C_j)) \right|^p} \tag{3.3}$$

Il est facile de montrer que $dist_B^p$ est bien une distance. Elle vérifie les trois propriétés, à savoir, la séparation, la symétrie et l'inégalité triangulaire :

1. La séparation : $\forall(\hat{X}_1, \hat{X}_2) \in \mathcal{D}\ dist_B^p(\hat{X}_1, \hat{X}_2) = 0 \Leftrightarrow \hat{X}_1 = \hat{X}_2$
2. La symétrie : $\forall(\hat{X}_1, \hat{X}_2) \in \mathcal{D}\ dist_B^p(\hat{X}_1, \hat{X}_2) = dist_B^p(\hat{X}_2, \hat{X}_1)$
3. L'inégalité triangulaire : $\forall(\hat{X}_1, \hat{X}_2, \hat{X}_3) \in \mathcal{D}\ dist_B^p(\hat{X}_1, \hat{X}_3) \leq dist_B^p(\hat{X}_1, \hat{X}_2) + dist_B^p(\hat{X}_2, \hat{X}_3)$

• Pour **la norme** ℓ_1, la distance $dist_B^1(\hat{X}_1, \hat{X}_2)$, s'écrit comme suit :

$$dist_B^1(\hat{X}_1, \hat{X}_2) = \sum_{j=1}^{J} \sum_{m=1}^{d} \left| log(P(X_{1m}|C_j)) - log(P(X_{1m}|C_j)) \right| \tag{3.4}$$

• Pour **la norme** ℓ_2, la distance $dist_B^2(\hat{X}_1, \hat{X}_2)$, s'écrit comme suit :

$$dist_B^2(\hat{X}_1, \hat{X}_2) = \sum_{j=1}^{J} \sqrt{\sum_{m=1}^{d} \left| log(P(X_{1m}|C_j)) - log(P(X_{1m}|C_j)) \right|^2} \tag{3.5}$$

Pour chaque variable, la distance $dist_B^p$, peut s'interpréter comme une distance entre les ratios des probabilités ($log(P) - log(P') = log(\frac{P}{P'})$) en donnant une plus grande importance aux faibles différences de ratio (en raison de l'utilisation du logarithme).

Après avoir défini la distance dépendante de classe, on cherche à ce stade, à savoir si la distance obtenue permet de vérifier que "deux instances proches au sens de leur distribution (similarité) sont également proches au sens de leur comportement vis-à-vis de la classe à prédire". Il s'agit ici de montrer que la distance entre les distributions des classes conditionnellement aux données est

inférieure ou égale à la distance $dist_B^p$ entre les instances. Par conséquent, plus deux instances sont proches en termes de la distance $dist_B^p$ plus la probabilité qu'elles appartiennent de la même classe sera grande.

Adoptons le principe du prédicteur Bayésien, la distance entre les distributions des classes prédites pour deux instances $\hat{X}_1$ et $\hat{X}_2$ peut s'écrire selon la formule suivante :

$$\Delta^p(\hat{X}_1, \hat{X}_2) = \sum_{j=1}^{J} \left\| log(P(C_j|\hat{X}_1)) - log(P(C_j|\hat{X}_2)) \right\|_p \tag{3.6}$$

avec, $\forall i \in \{1, \ldots, N\}$

$$P(C_j|\hat{X}_i) = \frac{P(C_j)P(\hat{X}_i|C_j)}{P(\hat{X}_i)} = \frac{P(C_j)P(\hat{X}_{i1}, \ldots, \hat{X}_{id}|C_j)}{P(\hat{X}_i)} \tag{3.7}$$

En tenant compte de l'hypothèse d'indépendance des variables explicatives conditionnellement à la variable cible, l'équation 3.7 peut s'écrire de la manière suivante :

$$P(C_j|\hat{X}_i) = \frac{P(C_j)P(\hat{X}_i|C_j)}{P(\hat{X}_i)} = \frac{P(C_j)\prod_{m=1}^{d} P(X_{im}|C_j)}{P(\hat{X}_i)} \tag{3.8}$$

Par conséquent, on a :

$$log(P(C_j|\hat{X}_i)) = \sum_{m=1}^{d} log(P(X_{im}|C_j)) + log(P(C_j)) - log(P(\hat{X}_i)) \tag{3.9}$$

Partant de la définition donnée par l'équation 3.6 de la distance entre les distributions de classes prédites, on trouve la majoration suivante :

$$\Delta^p(\hat{X}_1, \hat{X}_2) \leq \left[dist_B^p(\hat{X}_1, \hat{X}_2) + J \left\| log(P(\hat{X}_2)) - log(P(\hat{X}_1)) \right\|_p \right] \tag{3.10}$$

Démonstration

$\Delta^p(\hat{X}_1, \hat{X}_2) = \sum_{j=1}^{J} \left\| log(P(C_j|\hat{X}_1)) - log(P(C_j|\hat{X}_2)) \right\|_p$

D'après l'équation 3.7, on trouve que

$$= \sum_{j=1}^{J} \left\| log(\tfrac{P(\hat{X}_1|C_j)P(C_j)}{P(\hat{X}_1)}) - log(\tfrac{P(\hat{X}_2|C_j)P(C_j)}{P(\hat{X}_2)}) \right\|_p$$

$$= \sum_{j=1}^{J} \left\| log(P(\hat{X}_1|C_j)) + log(P(C_j)) - log(P(\hat{X}_1)) - log(P(\hat{X}_2|C_j)) - log(P(C_j)) + log(P(\hat{X}_2)) \right\|_p$$

D'où

$$\leq \sum_{j=1}^{J} \left[\left\| log(P(\hat{X}_1|C_j)) - log(P(\hat{X}_2|C_j)) \right\|_p + \left\| log(P(\hat{X}_2)) - log(P(\hat{X}_1)) \right\|_p \right]$$

$$= \sum_{j=1}^{J} \left\| log(P(\hat{X}_1|C_j)) - log(P(\hat{X}_2|C_j)) \right\|_p + \sum_{j=1}^{J} \left\| log(P(\hat{X}_2)) - log(P(\hat{X}_1)) \right\|_p$$

$$= \sum_{j=1}^{J} \left\| log(P(\hat{X}_1|C_j)) - log(P(\hat{X}_2|C_j)) \right\|_p + J \left\| log(P(\hat{X}_2)) - log(P(\hat{X}_1)) \right\|_p$$
$$= dist_B^p(\hat{X}_1, \hat{X}_2) + J \left\| log(P(\hat{X}_2)) - log(P(\hat{X}_1)) \right\|_p$$

D'où

$$\Delta^p(\hat{X}_1, \hat{X}_2) \leq dist_B^p(\hat{X}_1, \hat{X}_2) + J \left\| log(P(\hat{X}_2)) - log(P(\hat{X}_1)) \right\|_p$$

Cette majoration signifie que deux instances de même probabilité globale proches au sens de $dist_B^1$ seront proches au sens de la prédiction des probabilités par classe cible. Elles seront également proches attribut par attribut, ce qui accroît la validité sémantique de la notion de proximité sous-jacente. Cette majoration est vraie aussi dans le cadre de la régression linéaire.

Note : Au cours de cette thèse, l'algorithme des K-moyennes est exécuté en utilisant la distance $dist_B^2$ en norme ℓ_2 et en utilisant la moyenne pour la mise à jour des centres.

Exemple illustratif

La figure 3.8 (partie gauche) présente le jeu de données Mouse qui est caractérisé par la présence de trois classes à prédire (noire, rouge et verte), deux variables descriptives continues (Variable 1 et Variable 2) et 490 instances. La première étape du prétraitement Conditional Info consiste à découper le domaine des deux variables descriptives en un nombre fini d'intervalles (5 intervalles pour la variable 1 et 3 intervalles pour la variable 2 comme le montre le tableau 3.2). Pour l'étape de recodage, chaque variable de X_i est transformée en 3 nouvelles variables synthétiques (puisque $|J| = 3$). Par exemple, la variable 1 d'une instance X_i ayant sa valeur dans l'intervalle $]0.27; 0.32]$ est transformée ainsi : $log(P(X_{i1} \in]0.27; 0.32]|noire), log(P(X_{i1} \in]0.27; 0.32]|rouge), log(P(X_{i1} \in]0.27; 0.32]|verte)$.

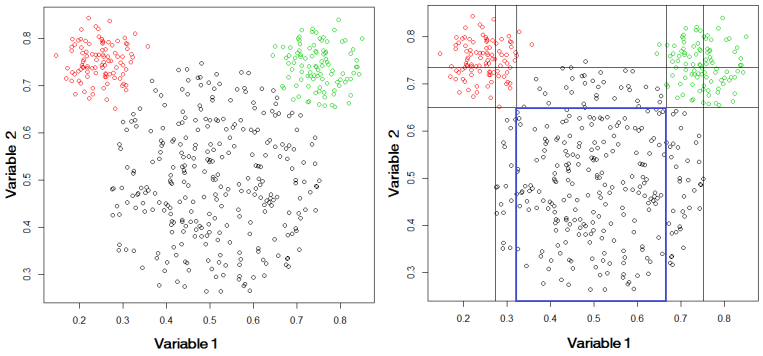

FIGURE 3.8 – Le jeu de données Mouse

A partir de ce type de recodage, on constate que toutes les instances ayant des valeurs dans les mêmes intervalles de discrétisation vont recevoir la même quantité d'information. A titre d'exemple, les instances entourées par le carré bleu dans la figure 3.8 (partie droite) ont bien des valeurs appartenant à l'intervalle de discrétisation $]0.32; 0.67]$ pour la variable 1 et des valeurs appartenant à l'intervalle de discrétisation $]-\inf; 0.65]$ pour la variable 2. Ces instances appartenant à la classe noire vont donc recevoir la même quantité d'information variable par variable. La figure 3.9 présente une illustration de la répartition des valeurs de la variable 1

Variable 1	Classe 'noire'	Classe 'rouge'	Classe 'verte'
$]-\inf;0.27]$	0	73	0
$]0.27;0.32]$	16	25	0
$]0.32;0.67]$	238	2	2
$]0.67;0.75]$	36	0	49
$]0.75;+\inf]$	0	0	49

Variable 2	Classe 'noire'	Classe 'rouge'	Classe 'verte'
$]-\inf;0.65]$	257	0	0
$]0.65;0.74]$	32	34	45
$]0.74;+\inf]$	1	66	55

TABLE 3.2 – La discrétisation des deux variables descriptives en utilisant l'approche MODL

avant et après le prétraitement. La partie gauche de cette figure présente la répartition des valeurs de départ de la variable 1. Tandis que les trois graphiques restant présentent la répartition des valeurs des trois variables synthétique obtenues après le prétraitement. Par exemple, le graphique présenté dans la partie droite de la figure 3.9 présente la répartition des valeurs de la variable $log(P(X_{i1} \in I_k|la\ classe\ verte))\ \forall i \in N$ avec I_k présente le nombre d'intervalles de discrétisation obtenu. d'après le tableau de droite présenté dans Table 3.2, on constate que, pour la variable 1, les instances de la classes verte sont réparties seulement dans 3 intervalles comme le prouve le graphique présenté dans la partie droite de la figure 3.9. Ceci prouve que les instances proches en termes de la distance $dist^1_B$ proposée seront proches au sens de la prédiction des probabilités par classe cible.

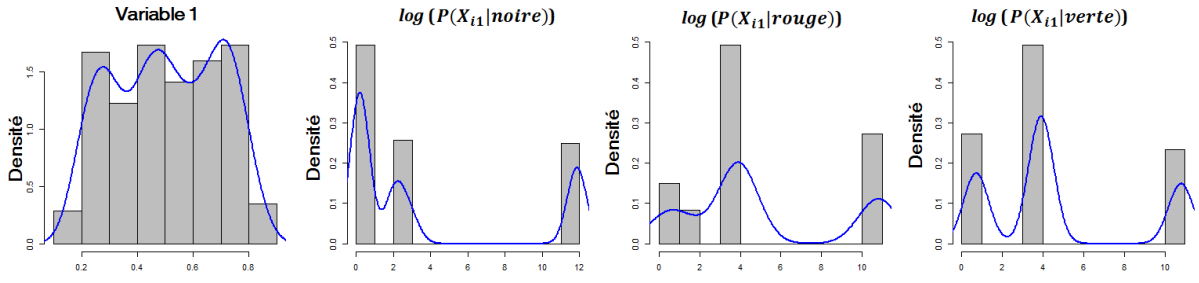

FIGURE 3.9 – La répartition de la variable 1 avant et après le prétraitement

Avantages

— Le premier point remarquable du prétraitement Conditional Info est sa capacité à construire une distance bayésienne qui vérifie que deux instances proches en termes de cette distance sont également proches en termes de leur probabilité d'appartenir à la même classe. Ce point pourrait être très utile pour l'algorithme des K-moyennes standard : l'incorporation de l'information cible dans les données sans la nécessité de modifier la fonction de coût de l'algorithme qui pourrait augmenter sa complexité algorithmique.

— Lors de la présence du bruit dans les données, après l'étape de discrétisation ou de groupage supervisé des variables, il existe deux possibilités : *i*) soit les exemples aberrants les plus proches reçoivent la même quantité d'information en formant un ou des groupes compacts (les points aberrants blues encadrés par le cadre magenta de la figure 3.10). Il est à signaler que ces points aberrants bleus sont bien de la classe rouge). *ii*) soit les exemples aberrants reçoivent la même quantité d'information que les instances qui forment le block de discrétisation ou de groupage (les points aberrants et les points rouges encadré par le cadre cyan de la figure 3.10). Par conséquent, l'effet de ces derniers sur ce groupe sera éliminé. Ceci montre que le prétraitement supervisé Conditional Info diminue en quelque sorte l'impact du bruit sur la qualité des résultats.

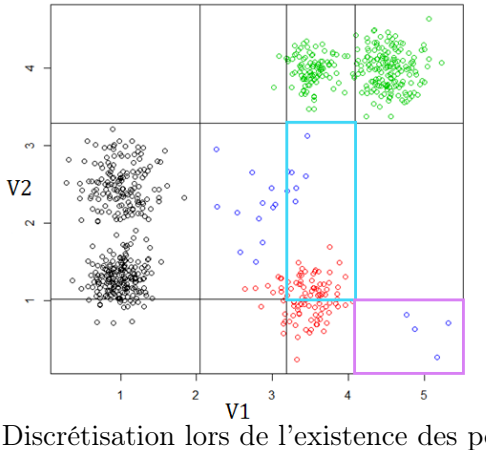

FIGURE 3.10 – Discrétisation lors de l'existence des points aberrants

Limites

Le prétraitement Conditional Info ne conserve pas la notion d'instance, c'est-à-dire que les instances peuvent recevoir la même quantité d'information (variable par variable) si toutes les variables de ces instances ont des valeurs qui appartiennent aux mêmes intervalles de discrétisation (par exemple, les instances entourées par le carré bleu dans la partie droite de la figure 3.8). Or ceci n'est pas un problème dans notre cas puisqu'on s'intéresse plutôt à faire des interprétations locales des données au lieu d'une interprétation individuelle (en se basant sur le prototype de chaque groupe). Les intervalles de discrétisation et les groupages en modalités ont dans notre cas une grande importance pour la réalisation de ces interprétations locales (voir l'exemple illustratif dans la section 3.5.3).

3.3 Protocole expérimental

3.3.1 Protocole

Comme il est connu, l'algorithme des K-moyennes est l'un des algorithmes de partitionnement qui converge rarement vers un optimum global. De ce fait, les résultats présentés dans cette section sont obtenus lorsque l'algorithme des K-moyennes est exécuté 100 fois avec différentes initialisations des centres en utilisant l'algorithme K-means++ [15]. La partition finale générée par l'algorithme des K-moyennes est choisie parmi les 100 partitions en utilisant un critère prédéterminé (voir le choix de la meilleure partition ci-dessous). Le reste des points à prendre en considération pour pouvoir aboutir aux résultats présentés dans la section 3.4 sont présentés ci-dessous.

- Les méthodes de prétraitement : Afin d'étudier l'impact de l'utilisation des méthodes de prétraitements supervisés (Conditional Info et Binarization) sur la qualité des résultats issus de l'algorithme des K-moyennes classique en termes de prédiction, nous allons les comparer aux méthodes de prétraitements usuels (non supervisés) pour les K-moyennes. Dans cette étude expérimentale, nous n'avons pas choisi de comparer les méthodes proposées avec les méthodes de prétraitement supervisées telles que l'analyse en composantes principales (ACP) en raison de la nécessité d'avoir des méthodes interprétables.

Pour les variables continues, selon notre connaissance, la normalisation des données est le pré-

traitement le plus communément utilisé dans la littérature pour l'algorithme des K-moyennes. Il permet d'ajuster une série de valeurs suivant une fonction de transformation pour les rendre comparables. La normalisation est nécessaire quand l'incompatibilité des unités de mesures entre les variables peut affecter les résultats sans apporter d'interprétations claires. Pour une comparaison équitable entre les variables ce prétraitement s'avère nécessaire. Les trois types de normalisation les plus répandus dans la littérature sont : *Min-Max Normalization (NORM)*, *Centrer et Réduire (CR)* et *Rank Normalization (RN)*.

1. **Min-Max Normalisation (NORM)** effectue une transformation linéaire sur les valeurs originelles des données. Si le minimum et le maximum de la variable u sont connus, alors cette dernière peut être transformée en une nouvelle variable qui prend ses valeurs dans $[0, 1]$. Cette transformation est effectuée en utilisant la formule suivante : $X'_{iu} = \frac{X_{iu} - min_{i=1,\dots,N} X_{iu}}{max_{i=1,\dots,N} X_{iu} - min_{i=1,\dots,N} X_{iu}}$ avec X_{iu} est la valeur d'origine de la variable u pour l'instance i.
2. **Centrer et réduire (CR)** fait référence à la transformation de données en soustrayant à chaque valeur la moyenne et en la divisant par l'écart-type. Cette transformation rendra toutes les valeurs en unités compatibles avec une distribution de moyenne 0 et d'écart-type 1. La formule qui permet cette transformation est : $X'_{iu} = \frac{X_{iu} - \mu}{\sigma}$ avec μ et σ sont respectivement la moyenne et l'écart type de la variable u.
3. **Rank Normalization (RN)** fait référence à la transformation de données en des groupes équitables de valeurs en respectant la répartition des valeurs de la variable u dans l'espace. Cette transformation commence par trier les valeurs de la variable u en ordre croissant. Ensuite, le vecteur résultant est divisé en H intervalles, où H représente le nombre d'intervalles. Suivant cet ordre, l'approche assigne à chaque intervalle un label $r \in \{1, \dots, H\}$. Finalement, pour la valeur X_{iu} appartient à l'intervalle r, alors celle-ci est recodée de la manière suivante $X'_{iu} = \frac{r}{H}$.

Pour les variables catégorielles, l'approche **Basic-Grouping-Binarization (BGB)** est la méthode de prétraitement la plus répandue dans la littérature. Cette approche vise à transformer les modalités des variables catégorielles en des valeurs booléennes. Les différentes étapes de **BGB** sont : *i*) grouper les modalités de chaque variable en g groupes de même fréquence où g est un paramètre utilisateur. *ii*) assigner à chaque groupe un label $v \in \{1, \dots, g\}$. *iii*) utiliser le codage disjonctif complet.

Le tableau 3.3 présente l'ensemble des prétraitements supervisés et non supervisés utilisé dans la partie expérimentale (Section 3.4).

Les prétraitements non supervisés			**Les prétraitements supervisés**		
Nom	variables numériques	variables catégorielles	Nom	variables numériques	variables catégorielles
RN-BGB	RN	BGB	BIN-BIN	BIN	BIN
CR-BGB	CR	BGB	CI-CI	CI	CI
NORM-BGB	NORM	BGB			

TABLE 3.3 – Liste des prétraitements utilisés

- Les jeux de données : Pour évaluer et comparer les différentes méthodes de prétraitements en fonction de leur capacité à aider l'algorithme des K-moyennes standard à atteindre l'objectif du clustering prédictif (première type), nous allons effectuer des tests sur différents jeux de données de l'UCI [1]. Ces jeux de données ont été choisis afin d'avoir des bases de données diverses en termes de nombre de classes J, de variables (continues M_n et/ou catégorielles M_c) et

d'instances N. Le tableau 3.5 présente l'ensemble des jeux de données utilisé dans la première partie d'expérimentation, tandis que, le tableau 3.4 présente l'ensemble des jeux de données utilisé dans la deuxième partie d'expérimentation. Ces derniers sont les jeux de données utilisés par Eick et al. dans [46] et par Al-Harbi et al. dans [6]. Pour une comparaison équitable entre les performances des algorithmes, ces jeux de données sont modifiés de la même façon que dans [46] et [6].

ID	**Nom**	M_n	M_c	N	J	J_{maj}
18	Iris	4	0	150	3	33
19	Pima	8	0	768	2	65
20	Auto-Import	15	11	205	2	60
21	Contraceptive	2	7	1473	2	61
22	Heart	10	3	270	2	56

TABLE 3.4 – Liste des jeux de données utilisés dans la deuxième partie d'expérimentation - (J_{maj} représente ≈ pourcentage classe majoritaire)

ID	**Nom**	M_n	M_c	N	J	J_{maj}
1	Wine	13	0	178	3	40
2	Glass	10	0	214	6	36
3	Horsecolic	7	20	368	2	63
4	Soybean	0	35	376	19	14
5	Breast	9	0	683	2	65
6	Australian	14	0	690	2	56
7	Vehicle	18	0	846	4	26
8	Tictactoe	0	9	958	2	65
9	LED	7	0	1000	10	11
10	German	24	0	1000	2	70
11	Segmentation	19	0	2310	7	14
12	Abalone	7	1	4177	28	16
13	Waveform	21	0	5000	3	34
14	Adult	7	8	48842	2	76
15	Mushroom	0	22	8416	2	53
16	PenDigits	16	0	110992	10	10
17	Phoneme	256	0	2254	5	26

TABLE 3.5 – Liste des jeux de données utilisés dans la première partie d'expérimentation - (J_{maj} représente ≈ pourcentage classe majoritaire)

- Le choix de la meilleure partition : Lorsque l'algorithme des K-moyennes est exécuté plusieurs fois (100 fois dans cette étude expérimentale), le choix de la meilleure partition est alors une étape cruciale. Dans cette étude expérimentale, la meilleure partition est définie comme étant la partition qui minimise l'erreur quadratique moyenne (MSE). La formule mathématique utilisée pour la MSE est donnée comme suit :

$$MSE = \frac{1}{N}\frac{1}{Z}\frac{1}{K}\sum_{i=1}^{N}\sum_{z=1}^{Z}\sum_{t=1}^{K}(XR_i^z - k_t^z)^2 \tag{3.11}$$

— N est le nombre d'instances dans l'ensemble de données.

— Z est le nombre de variable après le processus de prétraitement. Par exemple, pour conditional Info, $Z = (M_n + M_c) \times J$.
— K est le nombre de clusters.
— XR est le nouveau vecteur d'instance après le processus de prétraitement utilisé. Par exemple, pour conditional Info, ce nouveau vecteur XR est de dimension $(M_n + M_c) \times J$.
— k_t^z est le centre de gravité du cluster t, représenté sous forme d'un un vecteur de dimension Z.

- Le nombre de clusters (K) : Étant donné un prétraitement i, le nombre de clusters varie de J (nombre de classes) jusqu'à K_i. Pour chaque jeu de données, K_i a été déterminé au préalable de manière à ce que la partition obtenue, avec $K{=}K_i$ permette d'obtenir un ratio (inertie inter / inertie totale) de 80%. La valeur de K_i ($i \in \{CI-CI, BIN-BIN, RN-BGB, CR-BGB, NORM-BGB\}$) pour chaque jeu de données et pour chaque prétraitement i est indiquée dans le tableau 3.6. Il est à noter que dans cette étude, le nombre de clusters K ne doit pas être inférieur à C puisqu'on suppose que la variable cible a une structure interne à découvrir. Pour plus de détails voir Annexe A de ce mémoire.

Données	**CI-CI**	**BIN-BIN**	**RN-BGB**	**CR-BGB**	**NORM-BGB**
Wine	12	47	38	35	33
Glass	15	21	25	17	12
Horsecolic	6	7	200	11	14
Soybean	20	20	49	49	49
Breast	4	56	12	15	12
Australian	22	74	210	58	126
Vehicle	11	126	24	17	16
Tictactoe	12	13	496	499	500
LED	17	19	19	17	19
German	7	10	363	280	217
Segmentation	23	64	21	15	8
Abalone	29	29	29	29	29
Waveform	86	64	64	64	64
Adult	12	64	64	64	64
Mushroom	8	78	64	250	64
PenDigits	73	64	33	28	22
Phoneme	64	64	64	64	64
Iris	3	6	4	4	4
Pima	10	22	74	69	61
Auto-imports	5	9	33	19	35
Contraceptive	6	18	92	56	73
Heart	23	29	90	67	53

TABLE 3.6 – Détermination de K_i pour chaque prétraitement i

- L'attribution des classes aux groupes appris : 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 : A la présence d'une nouvelle instance, l'algorithme lui affecte l'étiquette du cluster qui lui est plus proche [3] (*i.e.,* l'utilisation du 1 plus proche voisin).

3. Une instance i est plus proche au cluster C_1 que au cluster C_2 si et seulement $dist(i, g_1) < dist(i, g_2)$ avec

- La cross validation : Pour la première partie d'expérimentation (Section 3.4.1), pour pouvoir comparer les résultats obtenus, un 2×5 folds cross validation a été effectué sur chaque jeu de données. Les résultats sont donc présentés comme une moyenne de 10 tests. Pour la deuxième partie expérimentale (Section 3.4.2), pour être en mesure de comparer nos résultats avec ceux de Eick [46] et de Al-Harbi [6], nous effectuons : *i*) un 20×5 folds cross validation (comme les expérimentations effectuées par Al-Harbi dans [6]) pour les jeux de données Auto-import, Breast, Contraceptive et Pima. Ces jeux de données sont également modifiés de la même façon que dans [6]. *ii*) un 10×10 folds cross validation (comme les expérimentations effectuées par Eick dans [46]) pour les jeux de données Glass, Heart, Vehicle et Iris.

3.3.2 Évaluation de la qualité du clustering prédictif

Les algorithmes du **clustering prédictif du premier type** privilégient principalement l'axe de description. Comme dans la classification supervisée, ces algorithmes cherchent à prédire correctement la classe des nouvelles instances. La seule différence ici est que les algorithmes du clustering prédictif du premier type génèrent des clusters souvent supérieur au nombre des classes. Pour l'évaluation des performances prédictives de ces algorithmes, on cherche souvent à comparer deux partitions ayant un nombre différent de groupes : la première partition est la partition contenant les varies classes des instances et la deuxième partition est celle qui contient les ID-clusters générés par les algorithmes du clustering prédictif.

Pour ce cadre d'étude, il est clair que l'utilisation d'un critère tel que l'accuracy s'avère insuffisant. Parmi les nombreux critères existant dans la littérature permettant de comparer deux partitions ayant un nombre de clusters différents, on avons choisi d'utiliser : l'indice de rand ajusté (ARI) [57] et la variation d'information [76]. Ce choix a été basé d'après l'étude réalisée dans [102].

Indice de Rand Ajusté (ARI)

Soit $D = \{(X_i, Y_i)\}_1^N$ un ensemble d'apprentissage de taille N et $\mathcal{C}_1 = \{s_1, \ldots, s_{K_1}\}$ et $\mathcal{C}_2 = \{u_1, \ldots, u_{K_2}\}$ deux partitions ayant respectivement K_1 et K_2 clusters tel que $D = \cup_{i=1}^{K_1} s_i = \cup_{j=1}^{K_2} u_j$. L'indice de rand ajusté (ARI) [57] permettant de comparer les deux partitions $\mathcal{C}_1$ et $\mathcal{C}_2$ est donné par la formule suivante :

$$ARI = \frac{\sum_{i,j}\binom{N_{ij}}{2} - \left[\sum_i \binom{N_{i.}}{2} \sum_j \binom{N_{.j}}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{N_{i.}}{2} + \sum_j \binom{N_{.j}}{2}\right] - \left[\sum_i \binom{N_{i.}}{2} \sum_j \binom{N_{.j}}{2}\right] \binom{N}{2}} \tag{3.12}$$

avec

— N_{ij} représente le nombre d'instances appartenant à la fois au cluster s_i et au cluster u_j .

— $N_{i.}$ représente le nombre d'instances appartenant au cluster s_i $\forall i \in \{1, \ldots, K_1\}$.

— $N_{.j}$ représente le nombre d'instances appartenant au cluster u_j $\forall j \in \{1, \ldots, K_2\}$.

L'indice de rand ajusté (ARI) est compris entre 0 et 1. Il est égal à 1 lorsque deux partitions sont exactement identiques. c'est critère à maximiser

g_1 (respectivement g_2) est le centre de gravité du cluster C_1 (respectivement C_2).

Variation d'Information (VI)

Le critère de comparaison Variation d'Information (VI) issu de la théorie de l'information, quantifie l'information apportée par la connaissance d'une partition $\mathcal{C}$ sur une partition $\mathcal{C}'$. Soit N_j le cardinal de la classe c_j. Soit :

- $P(j) = \frac{N_j}{N}$ la probabilité d'une observation X_i choisie au hasard appartienne à la classe c_j.
- $P(j,l) = \frac{|c_j \cap c'_j|}{N}$ la probabilité que les observations appartiennent aux classes $c_j \in \mathcal{C}$ et $c'_j \in \mathcal{C}'$.

La Variation d'Information entre deux partitions $\mathcal{C}$ et $\mathcal{C}'$ est la somme de l'information sur $\mathcal{C}$ que l'on perd et de l'information sur $\mathcal{C}'$ que l'on gagne lorsqu'on passe de la partition $\mathcal{C}$ à la partition $\mathcal{C}'$. Formellement, on a :

$$VI(\mathcal{C},\mathcal{C}') = \mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{C}') - 2\mathcal{I}(\mathcal{C},\mathcal{C}') \tag{3.13}$$

$\mathcal{H}$ et $\mathcal{I}$ présentent respectivement l'entropie et l'information mutuelle. Pour plus de détails sur cette mesure, voir [76]. Une version normalisée de VI, notée VIn a été proposé dans [102]. Cette dernière est donnée par l'équation 3.14 :

$$VIn(\mathcal{C},\mathcal{C}') = 1 - \frac{2\ \mathcal{I}(\mathcal{C},\mathcal{C}')}{\mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{C}')} \tag{3.14}$$

VIn est compris entre 0 et 1. Elle est égale à zéro si et seulement si $\mathcal{C}$ et $\mathcal{C}'$ sont identiques. Au cours de cette thèse, la version normalisée VI est utilisée que l'en note également VI.

Pour évaluer la performance prédictive des modèles utilisés dans ce chapitre, nous allons utiliser l'indice de rand ajusté (ARI).

Dans le cadre du **clustering prédictif du deuxième type**, on cherche à réaliser un compromis entre la prédiction et la description. À notre connaissance, il n'existe pas dans la littérature de critère global qui permette de mesurer ce compromis. Une possibilité pour évaluer ce compromis est d'utiliser le Front de Pareto (pour plus de détails, voir [20]). Cependant, pour l'axe de description, les résultats issus de l'algorithme des K-moyennes en utilisant les différents prétraitements ne sont pas comparables. En effet, le nombre de variables ainsi que la plage de variation diffèrent d'un prétraitement à l'autre. Par conséquent, l'utilisation d'un critère d'évaluation interne tel que Davies-Bouldin [38] ou la MSE (équation 4.2) ne permet pas de réaliser une telle comparaison. De ce fait, on suppose dans cette étude expérimentale que la partie "description" est garantie par l'algorithme des K-moyennes [4] et on évalue seulement la partie "prédiction" en utilisant le critère d'évaluation communément utilisé dans la littérature à savoir, *indice de rand ajusté ARI* (ou Adjusted Rand Index).

En dehors de ces deux critères, d'autres critères ont également été utilisés dans cette thèse tels l'erreur quadratique moyenne "MSE" (voir page 60) , la précision "ACC" (voir page 69) et BACC (Balanced Accuracy) (voir page 67).

3.4 Résultats

Pour être en mesure de répondre à la question posée dans la section 4.1, à savoir : "*les méthodes de prétraitement supervisées pourrait-elles aider l'algorithme des K-moyennes standard*

4. Dans la phase d'apprentissage, la partition finale générée par l'algorithme des K-moyennes est définie comme étant la partition ayant la meilleure MSE (parmi les 100 partitions, voir le choix de la meilleure partition).

à fournir de bons résultats au sens du clustering prédictif ?", nous allons diviser notre étude expérimentale en deux grandes parties.

Dans la première partie (Section 3.4.1), nous allons comparer les méthodes de prétraitement usuelles (non supervisées) pour les K-moyennes avec les deux méthodes de prétraitements supervisées (Conditional Info et Binarization) proposées dans les deux sections 3.2.2 et 3.2.3. Cette partie a pour but d'étudier l'impact de l'utilisation des méthodes de prétraitements supervisées sur la qualité des résultats issus par l'algorithme classique des K-moyennes. Il est à rappeler que la qualité discutée dans ce chapitre est définie comme étant le pouvoir prédictif de l'algorithme à bien prédire la classe des nouvelles instances. Ce pouvoir est calculé à l'aide de l'indice de rand ajusté ARI (équation 3.12).

Dans la deuxième partie (Section 3.4.2), nous allons comparer les performances prédictives de l'algorithme des K-moyennes précédé à chaque fois par un prétraitement supervisé avec les deux algorithmes les plus répandus dans le cadre du clustering supervisé. Ces algorithmes sont notamment, l'algorithme de Eick et al. proposé dans [46] et l'algorithme de Al-Harbi et al. proposé dans [6]. Cette partie a pour but d'étudier le degré de compétitivité de l'algorithme classique des K-moyennes précédé par les deux prétraitements supervisés (Conditional Info, et Binarization) avec ces deux algorithmes de clustering supervisés.

3.4.1 Distances supervisées *Vs.* distances non supervisées

Le but de cette première étude expérimentale est de vérifier si l'incorporation de l'information cible dans les données via un prétraitement supervisé pourrait aider l'algorithme des K-moyennes standard à atteindre l'objectif du clustering prédictif. Comme défini dans la section 1.6 du Chapitre 1, le clustering prédictif traite principalement trois axes, à savoir : la description, la prédiction et l'interprétation. Pour les différents prétraitements utilisés dans cette partie (voir Section 3.3), il s'avère difficile de comparer leurs performances suivant l'axe de description. En effet, À notre connaissance, les critères d'évaluation internes proposés dans le cadre du clustering sont tous basés sur une mesure de similarité. Or, ces méthodes de prétraitement n'ont pas forcément la même plage de variation ni le même nombre de variables. Pour cette raison, on suppose dans cette étude expérimentale que l'axe de description est garanti par l'algorithme des K-moyennes [5] et on évalue uniquement la performance des méthodes suivant l'axe de prédiction. Concernant l'axe d'interprétation, il sera discuté dans la section 3.5.3. Cette première partie d'expérimentation cherche donc à savoir si l'algorithme classique des K-moyennes précédé par les prétraitements supervisés parvient à bien prédire la classe des nouvelles instances comparé aux prétraitements non supervisés.

Pour l'algorithme des K-moyennes, le choix du nombre de clusters (K) est un problème en soi : il n'est pas évident de connaître à l'avance pour chaque jeu de données le nombre de clusters convenable. Cette étude expérimentale est donc divisée en deux selon la façon de choisir le nombre de clusters K. Dans la première partie, ce dernier est considéré comme un paramètre utilisateur : K est égal, pour chaque jeu de données, au nombre de classes (J) à prédire. Dans la deuxième partie le nombre de clusters est considéré comme une sortie de l'algorithme : pour chaque prétraitement i, l'algorithme des K-moyennes est exécuté avec différent nombre de clusters (de J jusqu'à K_i, pour plus de détails, voir le choix du nombre de clusters dans la section 3.3), ensuite, le nombre de clusters optimal $K_{opti} \in \{J, \ldots, K_i\}$ est considéré comme étant la

5. La partition finale générée par l'algorithme des K-moyennes est définie comme étant la partition qui optimise l'erreur quadratique moyenne (MSE) parmi les 100 partitions (voir le choix de la meilleure partition dans Section 3.3).

partition qui optimise l'indice de rand ajusté.

A. Le nombre de clusters est une entrée

Dans cette partie, on se limite au cas où le nombre de clusters K est égal au nombre de classes à prédire J. Dans ce cas, le problème du clustering prédictif devient un problème de classification supervisée. Le but ici est de connaître la capacité de l'algorithme classique des K-moyennes précédé par les méthodes de prétraitements supervisés à prédire correctement la classe des nouvelles instances.

La figure 3.11 présente les performances prédictives moyennes (en termes d'ARI) de l'algorithme des K-moyennes précédé par les différentes méthodes de prétraitement (supervisées et non supervisées) pour 17 jeux de données de l'UCI (voir tableau 3.5 de la section 3.3).

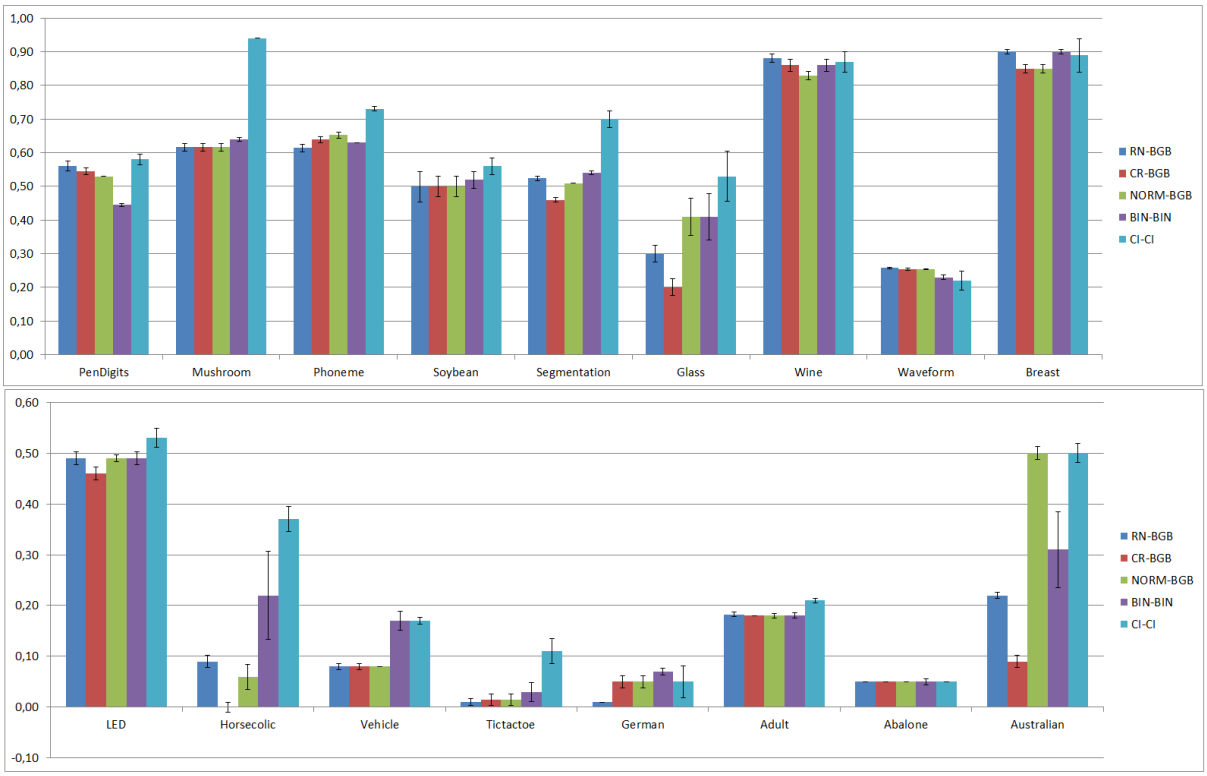

FIGURE 3.11 – Les performances prédictives moyennes des K-moyennes précédé par les différentes méthodes de prétraitement en utilisant l'ARI

Dans cette figure, on observe que la méthode de prétraitement supervisée "Conditional Info" a une performance prédictive soit : *i*) meilleure de celles de prétraitement non supervisés (12 jeux de données sur 17), *ii*) compétitive avec les performances de ces derniers (5 jeux de données sur 17). L'ensemble des tableaux contenant les résultats détaillés (en apprentissage et en test) qui servent à obtenir ces résultats synthétiques présentés dans cette partie sont situés dans la section B.2.1 l'annexe B.

À ce stade, pour être en mesure de classer les différentes méthodes de prétraitement selon leur pouvoir prédictif sur les 17 jeux de données, nous allons utiliser le test de Friedman couplé au test post-hoc de Nemenyi [41] (voir Section B.1 de l'annexe B). La figure 3.12 présente les résultats des comparaisons des performances prédictives en termes d'ARI en apprentissage (partie gauche de la figure) et en test (partie droite de la figure) de l'algorithme des K-moyennes en utilisant à chaque fois une méthode de prétraitement. Les méthodes sont classées par ordre décroissant selon leurs performances prédictives en se basant sur la moyenne des rangs : plus le rang de la méthode est proche de 1 meilleure est la prédiction.

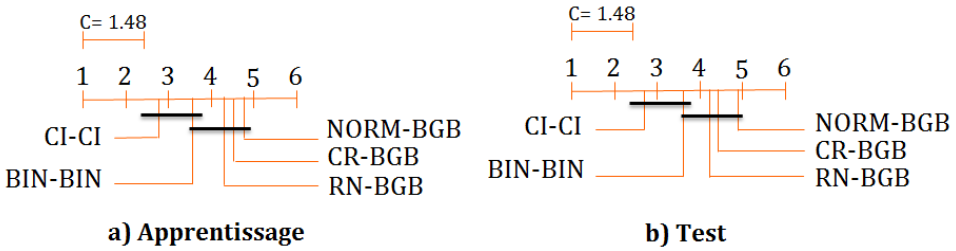

FIGURE 3.12 – Le test de Friedman couplé au test post-hoc de Nemenyi pour les 21 jeux de données en utilisant l'ARI en apprentissage a) et en test b)

D'après les résultats de test de Friedman, il existe une différence significative entre les 4 méthodes de prétraitement ($p_{value} < 10^{-4} \ll 0.05$). Que ce soit en apprentissage ou en test, d'après les résultats du test de Nemenyi, on constate que les deux méthodes supervisées sont celles qui ont une bonne performance en termes de prédiction tandis que la méthode Normalization est celle qui fournit des résultats moins bons en termes de prédiction.

B. Le nombre de clusters est une sortie

Dans le cadre du clustering prédictif, on s'attend à ce que le nombre de clusters soit supérieur au nombre de classes du fait qu'on souhaite découvrir à ce stade la structure interne de la variable cible (on suppose qu'au moins une des classes contient une structure sous-jacente à découvrir). Dans cette partie, on considère que le nombre de clusters K comme une sortie de l'algorithme des K-moyennes : pour chaque jeu de données et pour chaque prétraitement i, l'algorithme des K-moyennes est exécuté avec différentes valeurs de K (de J jusqu'à K_i) tout en effectuant une validation croisée en 10 folds. Ensuite, à la fin de la phase d'apprentissage, le nombre de clusters considéré est celui qui correspond à la partition ayant une bonne performance en termes de l'indice de rand ajusté (*i.e.,* celle qui optimise l'ARI). Puisque le critère d'ARI est utilisé pour sélectionner le nombre optimal de clusters, la qualité prédictive de l'algorithme en question précédé par les différentes méthodes de prétraitements est mesurée dans cette partie en utilisant "*Balanced Accuracy*"(BACC).

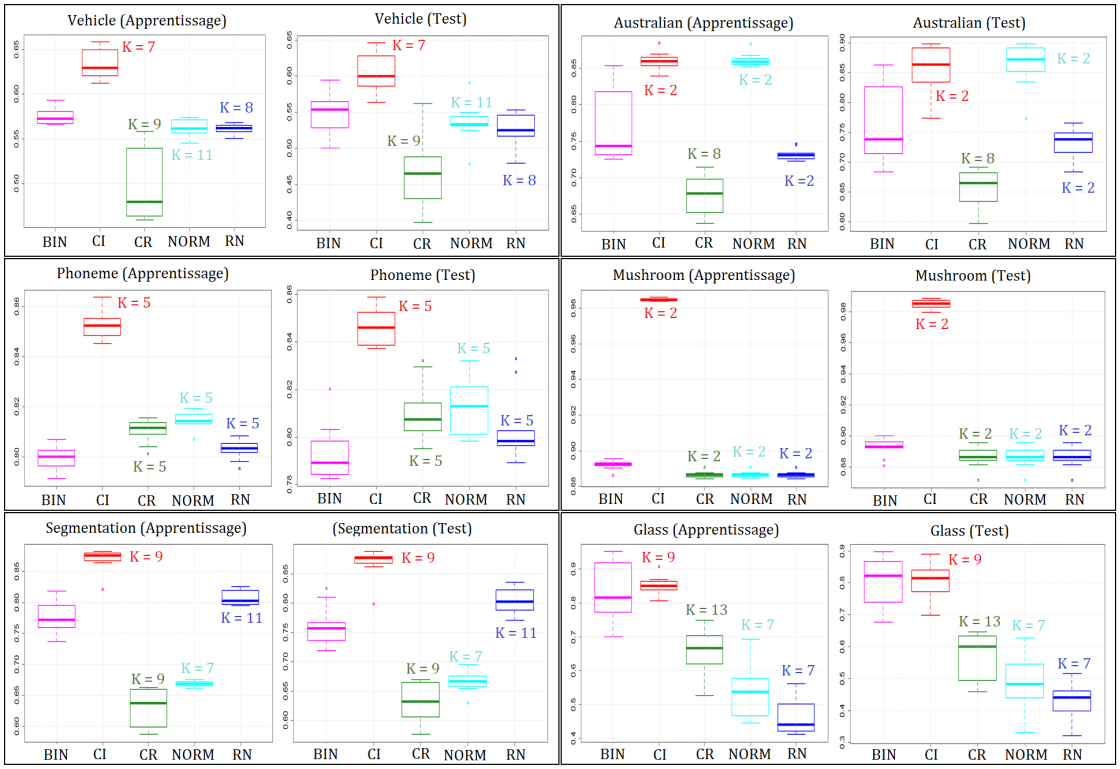

FIGURE 3.13 – La performances moyenne (en termes de BACC en test) de l'algorithme des K-moyennes standard précédé par les différentes méthodes de prétraitement dans le cas où K est une sortie.

La figure 3.13 et le tableau 3.7 présentent les performances prédictives moyennes (en termes de BACC) de l'algorithme classique des K-moyennes précédé par les différentes méthodes de prétraitement lorsque le nombre de clusters est considéré comme une sortie. Ces résultats montrent clairement que Conditional Info est la méthode qui fournit de bons résultats en termes de prédiction tout en gardant un nombre minimal de clusters.

Données	Méthodes	K	BACC (A)	BACC (T)	Données	Méthodes	K	BACC (A)	BACC (T)
German	RN-BGB	6	0.5 ± 0	0.5 ± 0	Horsecolic	RN-BGB	3	0.45 ± 0.01	0.58 ± 0.08
	NORM-BGB	2	0.5 ± 0	0.5 ± 0		NORM-BGB	6	0.77 ± 0.08	0.59 ± 0.19
	CR-BGB	2	0.5 ± 0	0.5 ± 0		CR-BGB	11	$\mathbf{0.94 \pm 0.04}$	0.67 ± 0.19
	BIN-BIN	2	0.5 ± 0	0.5 ± 0		BIN-BIN	2	0.50 ± 0.02	0.62 ± 0.07
	CI-CI	5	$\mathbf{0.56 \pm 0.02}$	$\mathbf{0.54 \pm 0.03}$		CI-CI	2	0.53 ± 0.01	$\mathbf{0.70 \pm 0.09}$
LED	RN-BGB	11	$\mathbf{0.71 \pm 0.02}$	0.7 ± 0.02	Soyeban	RN-BGB	22	0.74 ± 0.03	0.71 ± 0.04
	NORM-BGB	11	0.7 ± 0.02	0.69 ± 0.03		NORM-BGB	22	0.74 ± 0.03	0.71 ± 0.04
	CR-BGB	10	0.66 ± 0.02	0.66 ± 0.02		CR-BGB	22	0.74 ± 0.03	0.71 ± 0.04
	BIN-BIN	11	0.7 ± 0.02	0.69 ± 0.03		BIN-BIN	22	0.77 ± 0.02	0.75 ± 0.03
	CI-CI	10	$\mathbf{0.71 \pm 0.02}$	$\mathbf{0.71 \pm 0.02}$		CI-CI	20	$\mathbf{0.79 \pm 0.01}$	$\mathbf{0.79 \pm 0.02}$
Tictactoe	RN-BGB	17	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	Wine	RN-BGB	3	0.98 ± 0.01	0.97 ± 0.02
	NORM-BGB	17	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$		NORM-BGB	3	0.96 ± 0.00	0.95 ± 0.02
	CR-BGB	19	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$		CR-BGB	3	0.97 ± 0.01	0.96 ± 0.02
	BIN-BIN	8	0.66 ± 0.02	0.62 ± 0.06		BIN-BIN	3	0.97 ± 0.01	0.96 ± 0.02
	CI-CI	2	0.62 ± 0.08	0.62 ± 0.08		CI-CI	3	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.97 \pm 0.01}$
Adult	RN-BGB	2	0.5 ± 0.00	0.5 ± 0.00	Waveform	RN-BGB	5	0.74 ± 0.00	0.74 ± 0.01
	NORM-BGB	2	0.5 ± 0.00	0.5 ± 0.00		NORM-BGB	5	0.74 ± 0.00	0.74 ± 0.01
	CR-BGB	5	0.5 ± 0.00	0.5 ± 0.00		CR-BGB	5	0.74 ± 0.00	0.74 ± 0.01
	BIN-BIN	2	0.5 ± 0.00	0.5 ± 0.00		BIN-BIN	5	$\mathbf{0.75 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$
	CI-CI	4	$\mathbf{0.54 \pm 0.01}$	$\mathbf{0.54 \pm 0.01}$		CI-CI	4	0.59 ± 0.02	0.58 ± 0.02
PenDigits	RN-BGB	12	0.82 ± 0.01	0.82 ± 0.01	Breast	RN-BGB	2	0.98 ± 0.00	0.98 ± 0.01
	NORM-BGB	12	0.82 ± 0.00	0.81 ± 0.01		NORM-BGB	2	0.95 ± 0.01	0.95 ± 0.01
	CR-BGB	13	0.83 ± 0.00	0.83 ± 0.00		CR-BGB	2	0.95 ± 0.01	0.95 ± 0.01
	BIN-BIN	11	$\mathbf{0.97 \pm 0.00}$	$\mathbf{0.97 \pm 0.00}$		BIN-BIN	2	$\mathbf{0.98 \pm 0.01}$	$\mathbf{0.98 \pm 0.00}$
	CI-CI	12	0.76 ± 0.03	0.75 ± 0.03		CI-CI	2	$\mathbf{0.98 \pm 0.02}$	$\mathbf{0.98 \pm 0.00}$
Abalone	RN-BGB	28	0.08 ± 0.01	0.09 ± 0.00					
	NORM-BGB	28	0.11 ± 0.01	0.12 ± 0.01					
	CR-BGB	28	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.13 \pm 0.01}$					
	BIN-BIN	29	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.13 \pm 0.01}$					
	CI-CI	28	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.13 \pm 0.01}$					

TABLE 3.7 – Les performances moyennes (en termes de BACC) de l'algorithme des K-moyennes standard précédé par les différents méthodes de prétraitement dans le cas où K est une sortie (A : Apprentissage, T : Test)

3.4.2 Distances supervisées *Vs.* Clustering supervisé

Dans la littérature, plusieurs algorithmes de clustering ont été modifiés (comme l'algorithme des K-moyennes) dans le but de les adapter au problème de la classification supervisée. Ces algorithmes sont connus sous le nom de "clustering supervisé". Dans cette deuxième partie d'expérimentation, nous allons comparer les performances prédictives moyennes de l'algorithme classique des K-moyennes précédé à chaque fois par une méthode de prétraitement supervisée (Conditional Info et Binarization) avec les performance moyennes des deux algorithmes de clustering supervisé les plus répandus dans la littérature (algorithme de Eick *et al.* et algorithme de AL-Harbi *et al.*). Le but de cette partie est de savoir si la modification d'une seule étape de l'algorithme classique des K-moyennes (*i.e.,* l'étape de prétraitement des données, voir la section 1.6.3 du Chapitre 1) aboutit à être compétitif avec les algorithmes de clustering supervisé communément utilisés.

L'algorithme de AL-Harbi prend le nombre de clusters en entrée de l'algorithme. Pour chaque jeu de données, ce nombre est défini comme étant le nombre de classes à prédire. Les jeux de données utilisés sont modifiés de la même façon que AL-Harbi dans [6] pour une comparaison équitable des résultats. Le tableau 3.8 (partie en bas) présente les performances prédictives moyennes en termes d'accuracy (ACC) de l'algorithme de AL-Harbi et de l'algorithme des K-

Comparaison avec l'algorithme de Eick : (K en une sortie)						
	Glass		Heart		Iris	
	K	ACC en test	K	ACC en test	K	ACC en test
Eick algorithm	34	0.636	2	0.745	3	**0.973**
K-means with BIN	7	$\mathbf{0.677 \pm 0.091}$	2	$\mathbf{0.813 \pm 0.076}$	4	0.933 ± 0.064
K-means with C.I	6	0.620 ± 0.093	2	0.808 ± 0.079	3	0.902 ± 0.083
Comparaison avec l'algorithme de Al-Harbi : (K est une entrée)						
	Auto-import		Breast		Pima	
	K	ACC en test	K	ACC en test	K	ACC en test
Algorithme de Al-Harbi	2	**0.925**	2	**0.976**	2	**0.746**
K-moyennes avec BIN	2	0.831 ± 0.054	2	0.974 ± 0.012	2	0.699 ± 0.043
K-moyennes avec C.I	2	0.814 ± 0.102	2	0.969 ± 0.020	2	0.740 ± 0.033

TABLE 3.8 – Comparaison des prétraitements supervisés avec les deux algorithmes de Eick et de Al-Harbi

moyennes standard précédé à chaque fois par une méthode de prétraitement supervisé (CI et BIN). Ces résultats montrent que les performances de l'algorithme de K-moyennes précédé par les méthodes de prétraitement sont compétitives à celle obtenues par l'algorithme de Al-Harbi. Cependant, il est important de rappeler que ce dernier intègre un algorithme génétique dans le fonctionnement de l'algorithme des K-moyennes afin d'optimiser une fonction objectif prédéfinie l'auteur. Ceci augmente la complexité algorithmique de l'algorithme. De ce fait, l'utilisation de l'algorithme de K-moyennes standard précédé par une étape de prétraitement reste préférable.

Pour l'algorithme de Eick, le nombre de clusters est considéré comme une sortie de l'algorithme. Le tableau 3.8 (partie en haut) présente les performances prédictives moyennes en termes d'accuracy (ACC) de l'algorithme de Eick et de l'algorithme des K-moyennes standard précédé à chaque fois par une méthode de prétraitement supervisé (CI et BIN). Ces résultats montrent que les performances de l'algorithme de K-moyennes précédé par les méthodes de prétraitement supervisées sont compétitives à celle obtenues par l'algorithme de Eick. L'algorithme des K-moyennes avec les prétraitements supervisés conserve un nombre faible de clusters comparé à l'algorithme de Eick. De plus, l'algorithme de Eick est un algorithme qui nécessite beaucoup d'efforts de calcul (voir section 1.5.2 du Chapitre 1). De ce fait, l'utilisation de l'algorithme des K-moyennes standard précédé par une étape de prétraitement reste préférable.

3.4.3 Conclusion

Dans cette première partie d'expérimentation nous avons pu montrer que le prétraitement supervisé Conditional Info parvient à aider l'algorithme classique des K-moyennes dans le contexte supervisé comparé aux méthodes de prétraitement non supervisées usuelles pour les K-moyennes. De plus, avec seulement la modification d'une seule étape de l'algorithme classique des K-moyennes, ce dernier arrive à avoir des résultats très compétitifs en termes de prédiction avec les deux algorithmes les plus répandus dans le cadre du clustering supervisé (algorithme de Eick et algorithme de Al-Harbi).

3.5 Discussion

Comme nous l'avons évoqué précédemment, il est difficile de comparer la qualité des résultats (en termes de description) générés par les différentes méthodes de prétraitement : les données

résultants des différentes méthodes de prétraitement n'ont pas la même plage de variation ni le même nombre de variables. Or, la majorité des critères existant dans la littérature permettant d'évaluer l'axe de description se basent principalement sur une mesure de similarité.

Cette section a pour but d'évaluer principalement les deux axes non traités dans la section précédente, à savoir, l'axe de description et l'axe d'interprétation. Dans un premier temps, nous allons étudier dans la section 3.5.1 la capacité des différentes méthodes de prétraitement à avoir des présentations des données pertinentes (évaluée par le degré de chevauchement dans les valeurs des attributs de différentes classes, la séparabilité linéaire des données et la séparabilité entre les différentes classes). Dans un deuxième temps, nous allons étudier dans la section 3.5.2 la capacité des prétraitements à construire de bonnes matrices de Gram relativement à la variable cible. Finalement et avant de conclure, nous allons présenter dans la section 3.5.3 un exemple qui illustre la facilité d'interprétation des résultats de l'algorithme des K-moyennes quand il est précédé par le prétraitement supervisé Conditional Info.

3.5.1 La complexité des données

La complexité des problèmes de la classification supervisée est attribuée à plusieurs sources [18]. Parmi ces sources, on trouve l'ambiguïté des classes. Cette dernière fait référence à la situation où les instances de différentes classes ne peuvent pas être distinguées. Ceci pourrait être du à l'incapacité des variables de départ à décrire le concept cible : il se peut que ces variables ne sont pas très discriminantes pour le problème de la classification supervisée. Ce type de complexité ne peut être résolu au niveau du modèle d'apprentissage. Dans ce cas, un prétraitement des données s'avère nécessaire pour désambiguïser les classes. Un bon prétraitement est donc celui qui permet une description plus pertinente du concept cible.

Dans cette section, on cherche à comparer le comportement des différentes méthodes de prétraitement (supervisées et supervisées) présentées ci-dessus vis-à-vis d'un problème de classification supervisée. Il s'agit ici d'évaluer la capacité de chaque méthode de prétraitement à mieux décrire le concept cible par rapport à la description de départ. Une meilleure description permettra de facilité la tâche à l'algorithme des K-moyennes standard dans le but d'atteindre l'objectif des algorithmes de la classification supervisée. Par conséquent, une meilleure performances prédictive sera introduite. Cette problématique est connue sous le nom de la complexité des données.

Il existe dans la littérature des mesures permettant d'évaluer la complexité des données après un prétraitement des données. Ces mesures sont divisées en trois catégories. La première catégorie mesure le degré de chevauchement dans les valeurs des variables de différentes classes. La deuxième catégorie estime dans quelle mesure les classes sont séparables en examinant la longueur et la linéarité de la frontière de décision. La troisième catégorie mesure la compacité des groupes des différentes classes. Dans cette partie, nous allons utiliser la même mesure utilisée dans l'article [82], à savoir, les mesures F1, L2, N1 et N3. Pour plus de détails sur ces mesures, voir [81].

• **Le ratio discriminant maximal de Fisher (F1)** mesure la puissance discriminative maximale de chaque variable. La formule mathématique de F1 est donnée comme suit :

$$F1 = max_{j=1}^{d} FDR_j \tag{3.15}$$

Quand le problème de la classification est binaire (2 classes), le ratio pour chaque variable est alors calculé de la façon suivant :

$$FDR_j = \frac{(\mu_1^j - \mu_2^j)^2}{(\sigma_1^j)^2 + (\sigma_2^j)^2} \tag{3.16}$$

La mesure F1 est compris entre 0, et μ_k, avec μ_k est la moyenne de la variable j pour la classe k ($k \in \{1, 2\}$). Une valeur élevée de F1 signifie qu'au moins l'une des variables permet à l'algorithme d'apprentissage de séparer les instances de classes différentes en des partitions parallèles à un axe de l'espace des variables. Cependant, une faible valeur de F1 ne signifie pas que les classes ne sont pas linéairement séparables, mais plutôt qu'elles ne peuvent pas être discriminées par des hyperplans parallèles à l'un des axes de l'espace des variables.

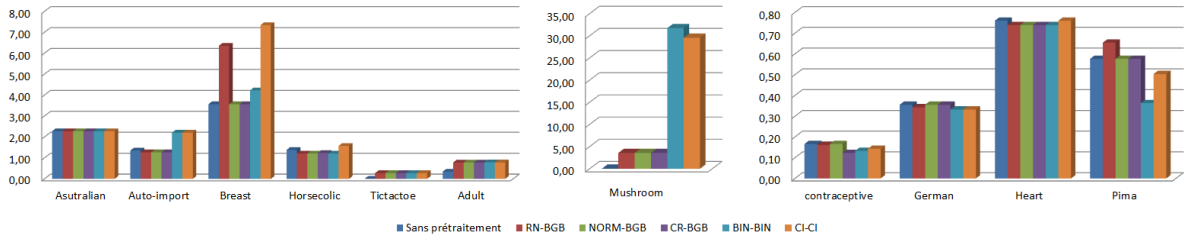

FIGURE 3.14 – La mesure F1 (à maximiser)

La figure 3.14 présente les performances des différentes méthodes de prétraitement en utilisant la mesure F1. Les résultats de cette figure montrent que les prétraitements supervisés CI-CI et BIN-BIN sont meilleurs ou très compétitifs avec les méthodes de prétraitement non supervisées. Ceci signifie que les méthodes de prétraitement supervisées parviennent à construire des variables synthétiques très discriminante pour le problème de la classification supervisée par rapport aux méthodes de prétraitement non supervisées et aux données de départ.

• **L'erreur d'apprentissage d'un classificateur linéaire (L2)** mesure le degré de la linéarité dans les données d'apprentissage. La mesure L2 commence par apprendre un algorithme d'apprentissage linéaire. Dans ce cas, l'algorithme des machines à vecteurs de support (SVM) [Vapnik, 1995] ayant un noyau linéaire et intégrant l'algorithme SMO [Platt, 1999] (Optimisation séquentielle minimale) est utilisé. Ensuite, la mesure renvoie l'erreur d'apprentissage sous forme de pourcentage des instances mal-classées. La mesure L2 est comprise entre 0 et 1. Une faible valeur de L2 signifie que les deux classes sont bien séparées.

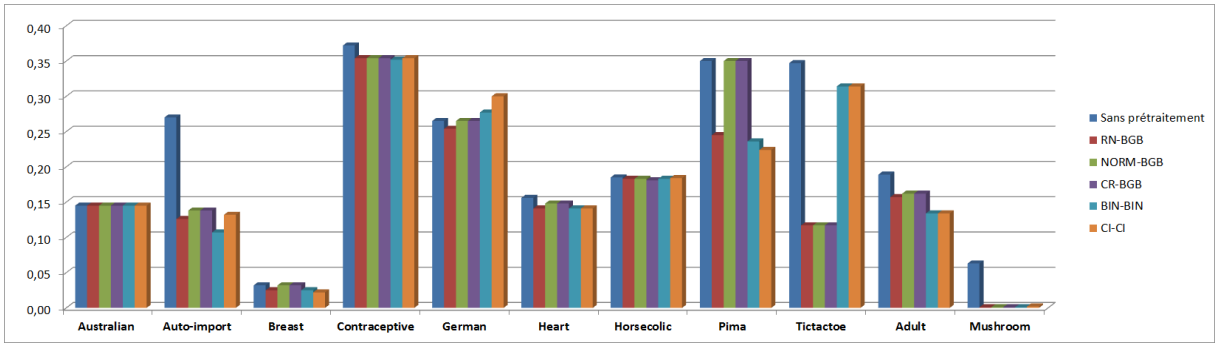

FIGURE 3.15 – La mesure L2 (à minimiser)

La figure 3.15 présente les performances des différentes méthodes de prétraitement en utilisant la mesure L2. Les résultats de cette figure montrent que les prétraitements supervisés CI-CI et

BIN-BIN sont meilleurs ou très compétitifs avec les méthodes de prétraitement non supervisées. Ceci signifie que les méthode de prétraitement supervisées parviennent à construire des données dont le degré de la linéarité est dans la plupart des temps, meilleur que celui des données du départ et des données générées par les méthodes de prétraitement non supervisées.

• **La proportion des points sur la frontière des classes (N1)** estime la longueur de la frontière des classes. Cette mesure est inspirée par le test proposé par Friedman et Rafsky [1979]. Elle commence par construire un arbre de recouvrement minimal (MST) sur l'ensemble des données en connectant tout d'abord tous les points à l'aide de la distance Euclidienne. Ensuite, elle retourne le rapport entre le nombre des nœuds de l'arbre de recouvrement qui relient les instances de différentes classes et le nombre total des instances dans l'ensemble de données. La mesure N1 est comprise entre 0 et 1. Une valeur élevée de cette mesure indique la majorité des points sont situés près de la frontière. Ceci peut rendre la tâche d'apprendre la frontière avec une bonne précision très difficile pour l'algorithme d'apprentissage.

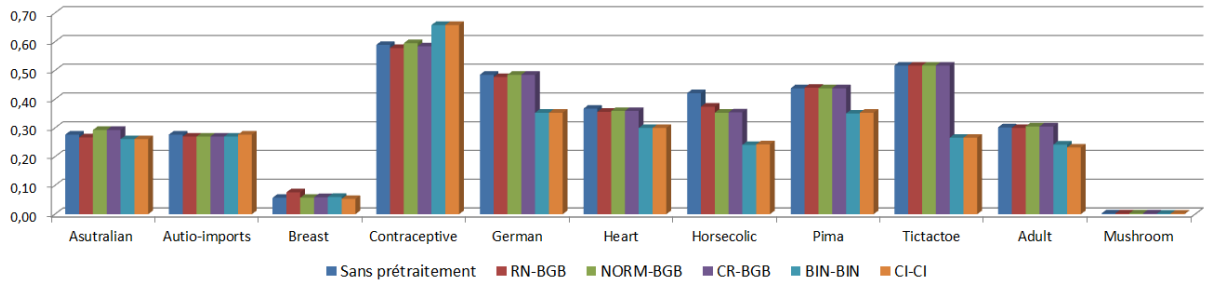

FIGURE 3.16 – La mesure N1 (à minimiser)

La figure 3.16 présente les performances des différentes méthodes de prétraitement en utilisant la mesure N1. Les résultats de cette figure montrent que les prétraitements supervisés CI-CI et BIN-BIN sont dans la plupart des temps, meilleurs que les méthodes de prétraitement non supervisées ((8/11) succès, (2/11) égalités et (1/11) échec). Ceci signifie que les méthodes de prétraitement supervisé parviennent à construire des données dont les classes sont bien séparées par rapport aux données du départ et aux données générées par les méthodes de prétraitement non supervisées.

• **Le taux d'erreur leave-one-out du classifieur '$1 - ppv$' (N3)** indique à quel point les instances de classes différentes sont proches. Elle renvoie le taux d'erreur de la validation leave-one-out pour le prédicteur K-plus proches voisins, avec K est fixé à 1. La mesure N3 varie dans l'intervalle $[0, 1]$. Les faibles valeurs de cette mesure indiquent qu'il existe un écart important dans la frontière des classes.

La figure 3.17 présente les performances des différentes méthodes de prétraitement en utilisant la mesure N3. Les résultats de cette figure montrent que les prétraitements supervisés CI-CI et BIN-BIN sont quasiment meilleurs que les méthodes de prétraitement non supervisées ((7/11) succès, (1/11) égalité et (3/11) échecs). Ceci signifie que les méthodes de prétraitement supervisé parviennent à construire des données dont les classes sont bien séparées par rapport aux données du départ et aux données générées par les prétraitements non supervisés.

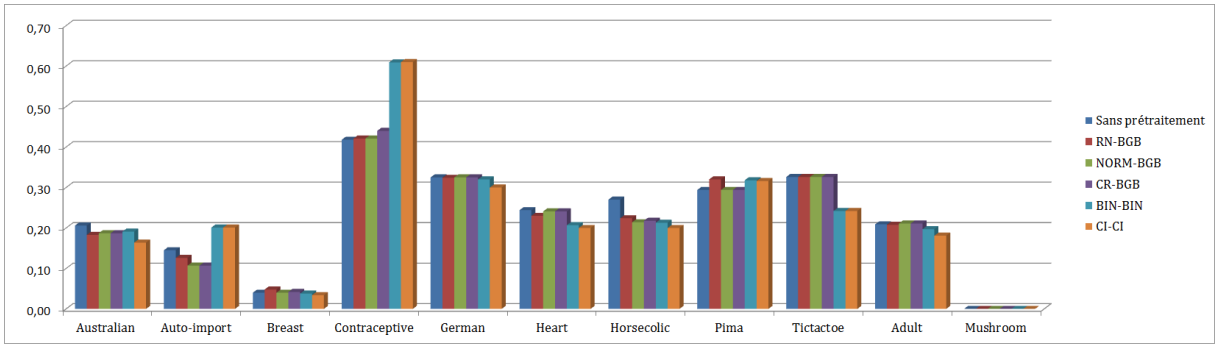

FIGURE 3.17 – La mesure N3 (à minimiser)

Discussion : Cette première étude expérimentale montre que nos méthodes de prétraitement parviennent dans la plupart des temps à construire des variables synthétiques discriminantes pour le problème de la classification supervisée par rapport aux variables de départ. Ensuite, les résultats expérimentaux montrent que nos méthodes de prétraitement supervisées des données parviennent dans la plupart des temps à générées des données dont les classes sont bien séparées par rapport aux données de départ et aux données générées par les méthodes de prétraitement non supervisées.

3.5.2 La similarité

Au niveau de l'axe de description, la comparaison de la qualité des résultats issus de l'algorithme des K-moyennes précédé par les différentes méthodes de prétraitements est une question très difficile. En effet, Ces méthodes n'ont pas la même plage de variation ni le même nombre de variables. Ceci rend l'utilisation des critères internes proposés dans le cadre du clustering inutile pour ce type de comparaison : ces critères se basent essentiellement sur une mesure de similarité pour évaluer la ressemblance entre les paires d'instances.

Un moyen pour surmonter ce problème, et de pouvoir comparer les différentes méthodes de prétraitement est d'évaluer, pour chaque jeu de données, la capacité des méthodes à construire de bonnes matrices de Gram relativement à la variable cible. Une matrice de Gram est une mesure de similitude entre les instances relativement à la variable cible. Une bonne matrice de Gram est donc celle qui produit une description plus concise de l'étiquetage des instances. Suivant ce contexte, la meilleure méthode de prétraitement est celle qui permet de construire des matrices de Gram pertinentes. Ceci reflète la capacité de la méthode à produire une bonne description des données vis-à-vis de la variable cible.

Il existe dans la littérature, un critère d'évaluation nommé EVA [48] permettant d'avoir une indication sur la capacité de chaque méthode de prétraitement à bien construire de bonnes matrices de Gram. La mesure EVA est une méthode qui prend en entrée une matrice de Gram et une variable cible catégorielle et renvoie un gain de compression. EVA mesure le gain qu'une partition établissant un compromis entre le nombre de groupes et la répartition des étiquettes peut apporter par rapport à la partition ayant un seul groupe. Celui-ci quantifie la capacité de la matrice de Gram à produire une description concise de l'étiquetage des instances. Elle évalue en général l'intérêt d'une matrice de Gram relativement à un problème de classification supervisée. EVA est une mesure à maximiser qui prend ses valeurs entre 0 et 1. Une valeur proche de 1 indique que la mesure de similitude relativement à la variable cible est très pertinente.

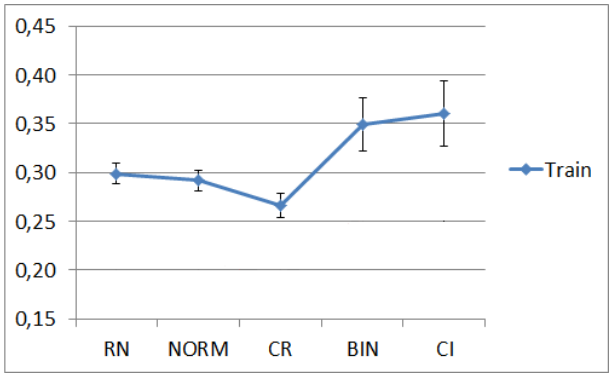

FIGURE 3.18 – La moyenne des performances en termes d'EVA sur 11 jeux de données (en test).

La figure 3.18 présente la moyenne des performances en termes d'EVA pour chaque méthode de prétraitement sur 11 jeux de données de l'UCI. Ces résultats sont obtenus en utilisant 100% des données en phase d'apprentissage. L'algorithme utilisé est l'algorithme des K-moyennes standard précédé par les différentes méthodes de prétraitement et la méthode d'initialisation des centres K-means++. Le nombre de clusters K considéré dans cette étude expérimentale est le nombre de classes J associé à chaque jeu de données.

Ces résultats montrent que la méthode de prétraitement Conditional Info (CI) suivie par la méthode Binarization (BIN) sont les deux méthodes de prétraitement qui produisent une meilleure description des données relativement à la variable cible et donc produisent des matrices de Gram pertinentes par rapport à celles produit par les méthodes de prétraitement non supervisées.

3.5.3 L'interprétation

L'interprétation des résultats issus de l'algorithme des K-moyennes prédictives est une condition incontournable dans cadre d'étude (voir la section 2.6.1 du chapitre 2). L'algorithme proposé dans cette thèse prend en entrée un biais de langage B permettant de bien décrire les données (voir l'algorithme générique du clustering prédictif (Algorithme 2) présenté dans la section 2.6.1 Chapitre 2). Ce biais de langage peut être vu par exemple comme des histogrammes permettant de connaître la répartition des variables dans chaque cluster appris.

Les méthodes de prétraitements proposées dans les deux sections 3.2.2 et 3.2.3 de ce chapitre sont en général des méthodes faciles à interpréter. En effet, la discrétisation supervisée des variables continues et le groupage supervisé en modalités des variables catégorielles rend l'interprétation "locale" des résultats issus de l'algorithme des K-moyennes plus facile. Cette interprétation locale permet à l'utilisateur de comprendre pour chaque cluster en particulier les facteurs les plus importants qui contribuent à sa construction. Pour illustrer la facilité d'interprétation des résultats issus de l'algorithme des K-moyennes standard précédé par les méthodes de prétraitement proposées (par exemple, Conditional Info), nous allons utiliser le jeu de données Adult de l'UCI. C'est un jeu de données caractérisé par la présence de 48842 instances, 15 variables descriptives et une variable cible possédant deux classes ("more", "less").

Pour ce cas illustratif, nous avons fixé le nombre de clusters à quatre. Les deux figures 3.19 et 3.20 dégagent quelques informations pertinentes concernant les instances du premier cluster de cette partition. Ces figures présentent pour chacune des variables (Relationship, Marital status, ..., etc) la proportion des instances appartenant à un groupe de modalité ou à un intervalle (selon la nature de la variable traitée) connaissant l'ensemble des données.

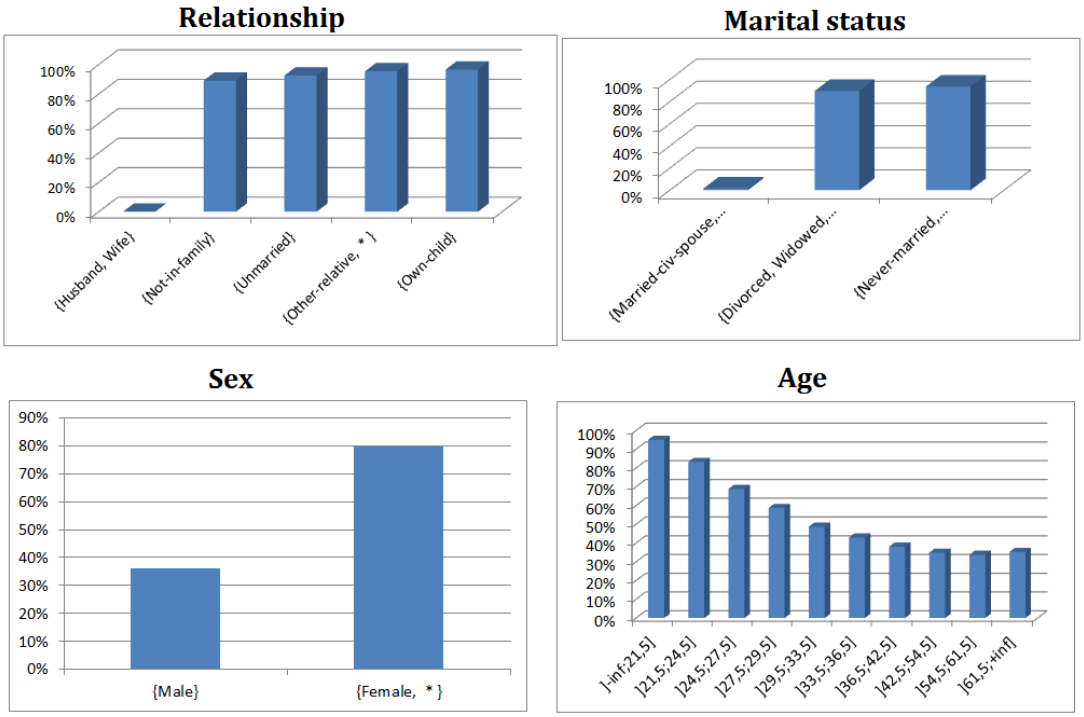

FIGURE 3.19 – L'interprétation locale d'un cluster d'une partition issus de l'algorithme des K-moyennes standard précédé par le prétraitement Conditional Info

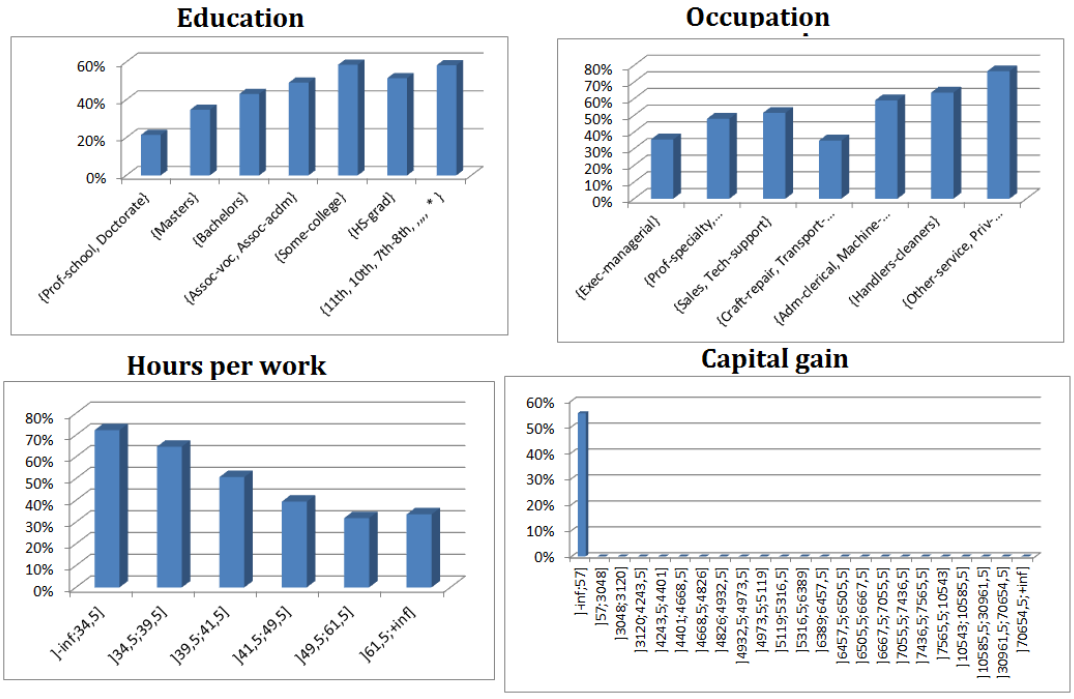

FIGURE 3.20 – L'interprétation locale d'un cluster d'une partition issue de l'algorithme des K-moyennes standard précédé par le prétraitement Conditional Info (Suite)

D'après les graphiques présentés dans la figure 3.19, On constate que :
— Pour la variable **Relationship** : le cluster ne contient pas de gens mariés. Par contre, il possède 92% des gens non mariés, 95% des gens ayant d'autres relations et 89% de gens sans famille existant dans de la base Adult.
— Pour la variable **Marital status** : les gens qui forment le cluster en question sont notamment soit divorcés, veufs ou leurs conjoints sont absents (90% de l'ensemble des données) ou bien séparés ou jamais mariés (94% de l'ensemble des données).
— Pour la variable **Sex** : Le cluster est formé des hommes et des femmes. Cependant, 80% des femmes existant dans l'ensemble des données Adult se trouvent dans ce cluster et seulement 36% des hommes.
— Pour la variable **Age** : le cluster contient tous les âges. L'information la plus importante ici est que 95% des gens de l'ensemble de données Adult ayant moins de 21.5 ans forment Ce cluster.

D'après les graphiques présentés dans la figure 3.20, On constate que :
— Pour la variable **Education** : Les gens de ce cluster ont des niveaux d'études variés. Par exemple, 95% des lycéens existant dans l'ensemble des données se trouvent dans ce cluster.
— Pour la variable **Occupation** : Les gens de ce cluster occupent des postes différents. À titre d'exemple, 51% des gens de l'ensemble des données travaillant dans le domaine de la vente se trouvent dans ce cluster.
— Pour la variable **Hours per work** : Puisque ce cluster possède des gens qui travaillent dans des postes divers, les heures de travail pendant la semaine varient également. Par exemple, 72% des gens travaillant moins de 34.5 heures par semaine existant dans l'ensemble de données se trouvent dans ce cluster.
— Pour la variable **Capital gain** : Ce cluster ne contient que des gens qui ont un capital gain inférieur à 57.

3.6 Bilan et synthèse

Ce chapitre a présenté l'influence d'une étape de prétraitement supervisé sur la qualité des résultats (au sens du clustering prédictif) générés par l'algorithme classique des K-moyennes. Tout d'abord, nous avons pu montrer que l'utilisation d'une distance dépendante de la classe construite à l'aide d'un prétraitement supervisé a la capacité d'aider l'algorithme des K-moyennes à atteindre l'objectif de clustering prédictif (du premier type) comparé aux méthodes non supervisées de prétraitement. En se basant sur l'ensemble des résultats obtenus dans la partie expérimentale, nous constatons que :

— *Pour l'axe de prédiction :* L'algorithme des K-moyennes standard précédé par le prétraitement supervisé Conditional Info parvient à fournir de bonnes performances prédictives en termes d'ARI par rapport aux performances obtenues lorsque l'algorithme est précédé par les méthodes non supervisées de prétraitement. La figure 3.21 présente la performance prédictive en termes d'ALC-ARI (en moyenne) obtenue sur 21 jeux de données lorsque l'algorithme des K-moyennes standard est précédé par Conditional Info et Rank Normalisation et/ou Basic Grouping . Cette figure confirme la conclusion tirée ci-dessus.

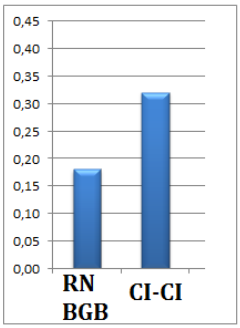

FIGURE 3.21 – la performance prédictive en termes d'ALC-ARI (en moyenne) obtenue sur 21 jeux de données lorsque l'algorithme des K-moyennes est précédé par CI-CI ou par RN-BGB

À travers l'utilisation d'un certain nombre de mesures de complexité des données, nous avons également pu montrer que nos méthodes de prétraitement supervisées des données (Conditional Info et Binarization) parviennent à construire dans la majorité des temps des variables synthétiques très discriminantes pour le problème de la classification supervisée. Les données générées par ces méthodes sont en général des données caractérisées par la présence des classes bien séparées les unes des autres par rapport aux données générées par les prétraitements non supervisées ou aux données de départ.

— *Pour l'axe de description :* Puisque les méthodes de prétraitements n'ont pas le même nombre de variables ni la même plage de variation, nous avons évalué l'axe de description en utilisant un critère permettant de mesurer la capacité des méthodes de prétraitements à construire de bonnes matrices de Gram relativement à la variable cible. Suivant ce contexte, nous avons pu montrer que le prétraitement supervisé Conditional Info suivi par le prétraitement Binarization parviennent à fournir de bonnes matrice de Gram par rapport aux prétraitements non supervisés.

— *Pour l'axe d'interprétation :* Dans la partie expérimentale, nous avons pu vérifier que la discrétisation supervisée des variables continues et le groupage en modalités des variables catégorielles permettent une interprétation aisée et plus concise des résultats issus de l'algorithme des K-moyennes.

Ensuite, nous avons pu montrer qu'avec la modification d'une seule étape de l'algorithme classique des K-moyennes (prétraitement des données), nous avons pu être compétitif (en termes de prédiction) face aux deux algorithmes de clustering supervisé les plus répandus dans la littérature et avec les arbres de décision tout en gardant une complexité algorithmique intéressante.

L'étape qui suit le prétraitement des données et qui semble avoir un impact direct sur la qualité des résultats issus des K-moyennes est l'étape d'initialisation des centres. Dans le cas de déséquilibre des classes à prédire (existence d'une classe majoritaire et d'une classe minoritaire), l'utilisation d'une méthode d'initialisation des centres non supervisée semble inappropriée. Par exemple, dans le cas où le nombre de clusters est égal au nombre de classes, la probabilité d'avoir plus d'un centre dans la classe majoritaire et de n'avoir aucun centre dans la classe minoritaire est élevée. Par conséquent, une détérioration au niveau de la prédiction va se produire. De ce fait, l'utilisation d'une étape d'initialisation supervisée des centres pourrait augmenter la performance de l'algorithme des K-moyennes en termes de prédiction. Cette étape d'initialisation fera donc l'objet du chapitre suivant.