Entre la prédiction et la description

Sommaire

Ce chapitre a fait l'objet des publications suivantes :

[9] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols. **Classification à base de clustering ou comment décrire et prédire simultanément ?**. Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA), Rennes, pages 7-12, 2015.

[12] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols : **Clustering prédictif : décrire, prédire et interpréter simultanément** à venir sur invitation suite à la conférence RJCIA, in Revue d'intelligence Artificielle (RIA).

2.1 Introduction

Au cours de ces dernières décennies, le monde a connu une véritable explosion du volume des données. La multiplication des systèmes et d'appareils capables de générer et de transmettre automatiquement des données est l'un des principaux facteurs à l'origine de ce phénomène. Chaque individu peut générer quotidiennement une multitude d'informations diverses et variées (*e.g.,* images, films, textes, sons, *etc.*) via le web, les réseaux sociaux et les appareils nomades. L'innovation continue des techniques de stockage figure également parmi les principaux facteurs de cette croissance exponentielle du volume des données. Par exemple, les grandes entreprises comme *Orange* et *Amazon* récoltent et stockent quotidiennement une avalanche de données concernant les comportements de leurs clients. Les résultats d'analyses médicales et les mesures effectuées un peu partout dans le monde comme les mesures météorologiques remplissent aussi d'importantes bases de données numériques.

Les données récoltées par ou pour les entreprises sont devenues un atout important. Les informations présentes, mais à découvrir au sein des grands volumes de données, sont devenues pour ces entreprises un facteur de compétitivité et d'innovation. Par exemple, à travers la connaissance des comportements des consommateurs, les entreprises peuvent avoir un aperçu de leurs attentes et de leurs besoins. L'étude des résultats médicaux peut également aider à mieux identifier les patients à risque, permettant ainsi de prévenir plutôt que de guérir. De ce fait, il existe un grand intérêt à développer des techniques permettant d'utiliser au mieux les gisements de données afin d'en extraire un maximum de connaissances utiles.

Dans la littérature, de nombreuses techniques d'analyse issues de diverses disciplines scientifiques (*e.g.,* statistique, Intelligence Artificielle, Informatique) ont été proposées. Par exemple, l'analyse multivariée [88] regroupe l'ensemble des méthodes statistiques qui s'attachent à l'observation et au traitement simultané de plusieurs variables en vue d'en dégager une information synthétique pertinente. Les deux grandes catégories de méthodes d'analyse statistique multivariées sont, d'une part, les méthodes dites *descriptives* et, d'autre part, les méthodes dites *prédictives*.

Les méthodes descriptives ont pour objectif d'organiser, de simplifier et d'aider à comprendre les phénomènes existant dans un ensemble important de données non étiquetées. Cet ensemble est organisé en instances constituées de plusieurs variables descriptives, où aucune des variables n'a d'importance particulière par rapport aux autres. Toutes les variables sont donc prises en compte au même niveau. Les trois grandes catégories de méthodes descriptives sont : *la description*, *la segmentation* et *l'association*.

1. *La description* [88] consiste à dégager les aspects les plus intéressants de la structure des données. Par exemple, les techniques d'analyse factorielles consistent à dégager des variables cachées dites "*facteurs*" à partir d'un ensemble de mesures. L'utilité de ces facteurs réside dans le fait qu'un nombre réduit de ces derniers explique aussi bien les données que l'ensemble des variables descriptives. Parmi les techniques factorielles, on citera celles les plus connues : Analyse en Composantes Principales (ACP) pour les variables quantitatives, Analyse des Correspondances Multiples (ACM) pour les variables qualitatives, Analyse Factorielle des Correspondances (AFC) pour les variables qualitatives et Analyse Factorielle Multiple (AFM) pour des groupes de variables quantitatives et/ou qualitatives.
2. *La segmentation* (le clustering ou la classification non supervisée) [2, 88, 62] cherche à discerner une structure dans un ensemble de données non étiquetées. L'objectif est de trouver une typologie ou une répartition des individus en groupes distincts. Chaque groupe (ou

cluster) doit contenir les individus les plus homogènes possible. Il s'agit donc de construire un modèle permettant de mieux présenter les observations de manière à la fois précise et compacte (voir section 2.2). Parmi les méthodes permettant d'atteindre cet objectif, on trouve par exemple : l'algorithme des K-moyennes, la classification hiérarchique ascendante/descendante et les réseaux de Kohonen, *etc.*

3. *L'association* consiste à mesurer le degré d'association entre deux ou plusieurs variables. Les relations découvertes sont exprimées sous forme de règles d'association. Cette analyse est appelée aussi analyse d'affinité. Elle est très utile par exemple pour détecter les produits achetés simultanément, dans une grande surface, par un très grand nombre de clients. Cette information sert à mieux fixer les assortiments et les offres promotionnelles. Les algorithmes utilisés dans ce cadre ont comme principe de détecter les propriétés qui reviennent fréquemment dans l'ensemble des données afin d'en déduire une catégorisation. Dans ce cadre d'étude, l'algorithme Apriori [3] est l'algorithme le plus utilisé.

Les méthodes prédictives permettent de prévoir et d'expliquer à partir d'un ensemble de données étiquetées un ou plusieurs phénomènes observables. Dans ce cadre, deux types de techniques se distinguent : *la régression* et *la classification supervisée.*

1. *La régression* a pour but de trouver à partir d'un ensemble de données, le lien entre les prédicteurs et une variable cible "numérique" à prédire. Parmi les méthodes permettant d'atteindre cet objectif, on trouve par exemple : la régression linéaire simple, la régression multiple, la régression logistique et le modèle linéaire généralisé (GLM) [88, 35], *etc.*
2. *La classification supervisée* est une estimation qui consiste à découvrir le lien entre une variable cible "catégorielle" et des variables descriptives. L'idée de base est de proposer un modèle permettant de prévoir l'appartenance des nouveaux individus à des classes prédéterminées. Les méthodes les plus répandues dans ce cadre sont : les réseaux de neurones (ANN), les machines à vecteurs de support (SVM) et forêts aléatoires (RF) [35, 68].

Dans la littérature sur le sujet d'extraction des connaissances utiles, le terme d'**apprentissage automatique** est souvent utilisé. Comme l'indique son nom, cette technique consiste à programmer la machine pour qu'elle apprenne à effectuer des tâches difficiles à travers des moyens algorithmiques. L'idée de base est de construire un modèle à partir d'un jeu de données, duquel les performances peuvent être évaluées en utilisant des méthodes de validation. Ces méthodes diffèrent selon le type d'apprentissage suivi (*e.g.,* la précision pour la classification supervisée et l'inertie intra\inter clusters pour le clustering). L'apprentissage automatique se décline en plusieurs variantes en fonction de la nature des données dont on dispose (supervisé, non supervisé, *etc.*). On peut donc placer la classification supervisée dans le domaine de l'apprentissage supervisé et le clustering dans le domaine de l'apprentissage non supervisé.

Dans cette thèse, nous nous intéressons exclusivement à la classification supervisée et non supervisée qui ont historiquement permis d'extrapoler de nouvelles informations à partir des informations présentes ou bien de découvrir et d'expliquer certains phénomènes existants mais noyés dans le volume de données.

Depuis quelques années, les chercheurs ont concentré leur attention sur l'étude d'un nouvel aspect d'apprentissage. Ce dernier fusionne à la fois les caractéristiques de la classification supervisée (la prédiction) et du clustering (la description). Les algorithmes appartenant à ce type d'apprentissage cherchent à **décrire et à prédire simultanément**. Il s'agit ici de découvrir la structure interne de la variable cible. Puis, munis de cette structure, de prédire la classe des nouvelles instances. Cette technique permet à l'utilisateur d'améliorer sa compréhension vis-à-vis des données. En effet, contrairement à la classification supervisée, les algorithmes descriptifs et prédictifs à la fois permettent à l'utilisateur de connaître les différentes voies qui peuvent mener à une même prédiction : deux instances très différentes peuvent avoir la même prédiction de classe. L'obtention d'une telle information est très utile dans plusieurs domaines d'application, notamment, dans les domaines critiques où l'interprétation des résultats issus des algorithmes d'apprentissage est une condition primordiale. A titre d'exemple, dans le domaine médical, deux patients X_1 et X_2 ayant comme prédiction un test positif (la classe $\{+\}$ de la figure 2.1) pour l'AVC (*i.e.,* une grande probabilité d'avoir un Accident Vasculaire Cérébral) n'ont pas forcément les mêmes causes et\ou les mêmes symptômes de l'AVC : il se peut que le patient X_1 soit une personne âgée, qui souffrait de la fibrillation auriculaire et qui par conséquent a eu des maux de têtes et des difficultés à apprendre (par exemple, X_1 appartient au groupe A de la figure 2.1). Tandis que le patient X_2, pourrait être une jeune personne qui consommait de l'alcool d'une manière excessive et, par conséquent a perdu l'équilibre (par exemple, X_2 appartient au groupe B de la figure 2.1).

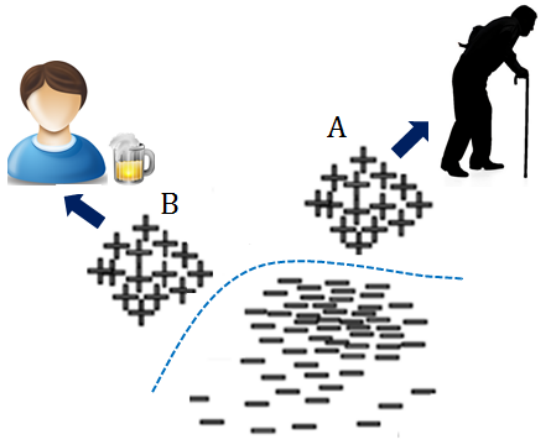

FIGURE 2.1 – Problème d'une classification binaire

L'objectif de cette thèse est la recherche d'un algorithme d'apprentissage "interprétable" permettant de décrire et de prédire d'une manière simultanée. Il s'agit ici de trouver un modèle capable d'équilibrer les trois axes (description, prédiction et interprétation) comme le montre la figure 2.2.

Pour atteindre cet objectif, il existe deux voies principales, à savoir : 1) rendre les méthodes descriptives plus prédictives ou 2) rendre les méthodes prédictives plus descriptives. Ceci est effectué en respectant l'axe d'interprétation. Avant d'entamer cette problématique dans la section 2.5, il est intéressant tout d'abord d'avoir une vision globale sur la classification supervisée, le clustering et l'interprétation des résultats issus d'un modèle d'apprentissage.

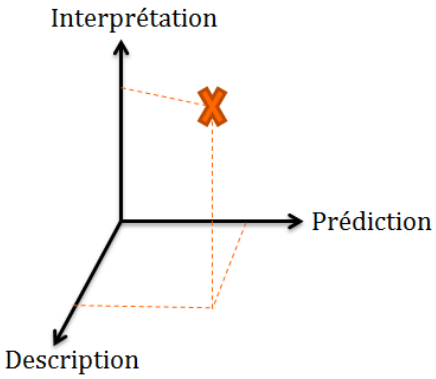

FIGURE 2.2 – Les trois axes traités dans cette thèse

Le reste de ce chapitre est donc organisé comme suit : La section 2.2 présente les trois étapes principales qui forment le processus du clustering. La section 2.3 met l'accent sur la capacité des modèles de classification supervisée à générer des résultats facilement interprétable par l'utilisateur. Cette section est divisée en deux parties principales. La première partie est dédiée aux modèles naturellement interprétables (voir Section 2.3.1). La deuxième partie (voir Section 2.3.2), quant à elle, présente l'ensemble des techniques qui peuvent aider à interpréter facilement les résultats générés par les modèles boîtes noires (i.e., les modèles fournissant des résultats incompréhensibles par l'être humain). Puisque l'interprétation est une notion clé dans cette thèse, la section 2.4 présentera d'une manière générale ses différents aspects. La section 2.5, quant à elle, met l'accent sur un nouveau type d'apprentissage qui consiste à fusionner les caractéristiques de la classification supervisée et du clustering. Finalement, la section 2.6 se focalise sur la présentation de notre problématique, de nos objectifs et des propositions préliminaires.

2.2 Approche descriptive : *La classification non supervisée*

Les approches descriptives désignent l'ensemble des méthodes permettant d'organiser et d'identifier des tendances dans les données. Loin de la volonté de faire un état de l'art exhaustif de toutes les méthodes descriptives existantes, on s'intéresse dans cette section uniquement à l'une des moyennes utilisée pour décrire les données, à savoir, *le clustering*. Cette section en présente ses concepts clefs.

Le clustering consiste à trouver la distribution sous-jacente des exemples dans leur espace de description. Autrement dit, à partir d'une base de données non étiquetées, cette approche vise à former des groupes (ou clusters) homogènes en fonction d'une certaine notion de similarité. Les observations qui sont considérées similaires sont associées au même groupe alors celles qui sont considérées comme différentes sont associées à des groupes différents. Plus formellement, dans les problèmes de clustering, les données $\mathcal{D} = \{X_i\}_{i=1}^{N}$ sont composées de N observations sans étiquette (ou classe), chacune décrite par plusieurs variables. On notera $X_i = \{X_i^1, \ldots, X_i^d\}$, l'ensemble de d variables décrivant l'observation i ($i \in [1, N]$). L'objectif ici est donc de partitionner l'espace d'entré en K clusters. Chaque cluster S_k ($k \in \{1, \ldots, K\}$) doit être, d'une part, différent des autres clusters et d'autre part, doit contenir des observations similaires.

D'une manière générale, le processus du clustering se divise en trois étapes principales (voir Figure 2.3) : (1) La préparation des données, (2) Le choix de l'algorithme de clustering et (3) La validation et l'interprétation des résultats.

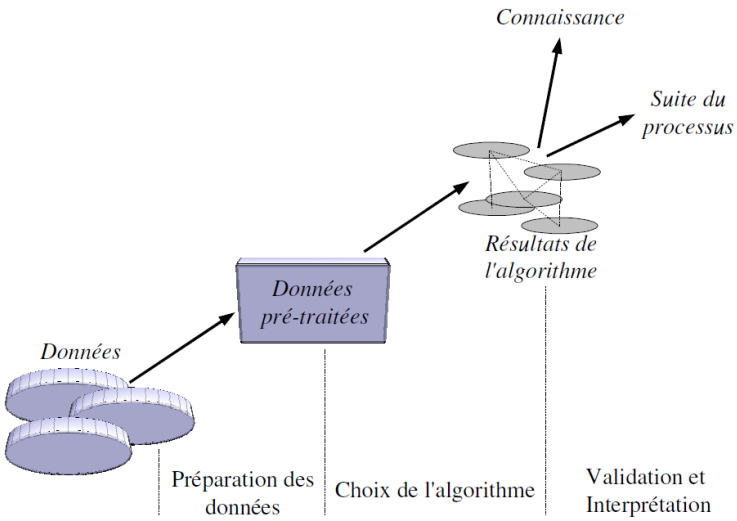

FIGURE 2.3 – Les différentes étapes du processus de clustering

2.2.1 La préparation des données

La phase de préparation des données est particulièrement importante dans le processus de la classification non supervisée (*e.g.,* [77] [30]). Bien souvent, une mauvaise représentation produit un clustering complexe et difficilement exploitable. Les données initiales peuvent contenir du bruit, des outliers, provenir de variables natives ou construites, *etc.* L'objectif du prétraitement est donc de chercher une représentation des données performante dans le contexte d'étude. La phase de préparation des données comprend des étapes de sélection, nettoyage, construction, intégration et recodage des données. Généralement, un bon prétraitement peut permettre au modèle d'identifier des clusters intéressants. Cependant, il se peut que celui-ci empêche l'interprétation ultérieure des résultats. Un état de l'art bien détaillé de ce sujet sera présenté dans le chapitre 2 de ce mémoire.

2.2.2 Le choix de l'algorithme de clustering

Le choix de l'algorithme de clustering dépend en général de la nature des variables (*e.g.,* quantitative et qualitative) dans les données et des clusters attendus (*e.g.,* nombre, forme, densité, *etc*). Généralement, les critères de décision peuvent être :

- *Les connaissances a priori* définissent l'ensemble d'informations concernant le nombre de clusters désiré, la distance minimale entre les clusters disjoints, *etc.*
- *La présentation des résultats* définit le type de la sortie de l'algorithme (*e.g.,* une hiérarchie de clusters ou une partition de l'ensemble des exemples).
- *La complexité* représente le temps de calcul nécessaire à la résolution d'un problème. Il est un critère important à prendre en compte lors du choix de l'algorithme. En particulier, il est admis que la complexité algorithme doit être linéaire en fonction du nombre d'exemples dans le cas des bases de données volumineuses.
- *Déterministe* définit la capacité des algorithmes à fournir les mêmes résultats (sans aucun changement) en utilisant les mêmes données en entrée.
- *Incrémental* définit la manière dont les données sont intégrées dans l'algorithme. Dans ce cas, les données sont intégrées au fur et à mesure de leur arrivée.
- *Prise en compte du contexte* définit la capacité de l'algorithme à prendre en compte ou

non la problématique du contexte.
— *La tolérance au bruit* définit la capacité de l'algorithme à gérer ou non le bruit qui peut exister dans les données.
— *La tolérance aux clusters de tailles variées* définit la capacité de l'algorithme à détecter des clusters ayant des tailles différentes (Figure 2.4).
— *La tolérance aux clusters de densités variées* définit la capacité de l'algorithme à réaliser des clusters ayant des densités différentes (Figure 2.4).

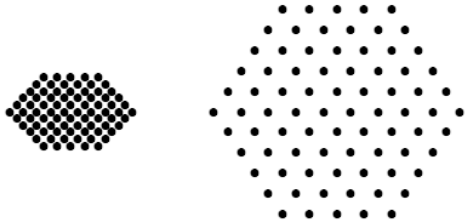

FIGURE 2.4 – Clusters de tailles et de densités différentes

— *La tolérance aux clusters de formes quelconques* définit la capacité de l'algorithme à réaliser des clusters ayant des formes différentes (Figure 2.5).
— *La tolérance aux clusters concentriques* définit la capacité de l'algorithme à réaliser des clusters concentriques, c'est-à-dire, inscrits les uns dans les autres (Figure 2.5).

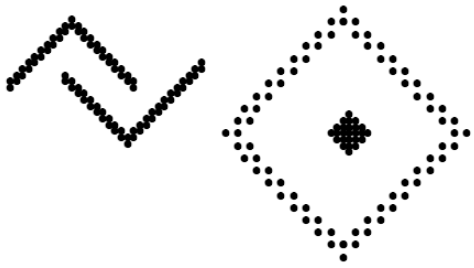

FIGURE 2.5 – Clusters de formes variées et concentriques

Le clustering se catégorise en plusieurs familles de méthodes selon la stratégie suivie pour construire les clusters. Parmi ces méthodes on trouve :

• **Le clustering hiérarchique**, comme l'indique son nom, cette approche consiste à former une hiérarchie de clusters : plus on descend dans la hiérarchie, plus les groupes sont spécifiques à un certain nombre d'exemples considérés comme similaires. Le clustering hiérarchique se catégorise en deux grandes familles : les méthodes '*ascendantes*' et les méthodes '*descendantes*'.

1. *Les méthodes ascendantes* commencent par une solution spécifique aux données pour arriver à une autre plus générale. Les méthodes de cette catégorie démarrent avec autant d'exemples que de clusters. Ensuite, elles fusionnent à chaque étape des clusters selon un critère donné jusqu'à l'obtention d'un seul cluster contenant ainsi l'ensemble de données.
2. *Les méthodes descendantes* partent d'une solution générale vers une autre plus spécifique. Les méthodes de cette catégorie démarrent avec un seul cluster contenant la totalité

des données, ensuite, elles divisent à chaque étape les clusters selon un critère jusqu'à l'obtention d'un ensemble de clusters différents stockés aux feuilles de la hiérarchie.

Il existe différentes approches pour mesurer la distance entre les clusters. On citera à titre d'exemple :

— L'approche *single-link* définit la distance entre deux clusters comme étant le minimum des distances pour toutes les paires d'exemples appartenant à des clusters différents [2].
— L'approche *complete-link* définit la distance entre deux clusters comme étant le maximum des distances pour toutes les paires d'exemples appartenant à des clusters différents [2].
— L'approche *average-link* définit la distance entre deux clusters comme étant la moyenne des distances pour toutes les paires d'exemples appartenant à des clusters différents [2].

• **Le clustering par partitionnement** consiste à diviser de manière optimale l'ensemble des instances en un groupe fini de groupes (K). L'objectif est ici de minimiser une mesure de la dissemblance intra-groupe pour k groupes. Le problème étant lié à l'optimisation d'une combinatoire, la solution trouvée sera rarement l'optimum global mais plutôt un des nombreux optimums locaux. Parmi ces méthodes, on trouve : les K-moyennes, les K-médoïdes ou le partitionnement autour des médoïdes (PAM) et la Carte Auto-Organisatrice.

• **Le clustering spectral** est considéré également comme un clustering de partitionnement. Par rapport à des algorithmes classiques comme celui des K-moyennes, cette technique offre l'avantage de classer des ensembles de données de structure « non-globulaire » dans un espace de représentation adéquat.

• **Le clustering basé sur la densité** consiste à identifier dans l'espace de description des objets les régions de forte densité, entourées par des régions de faible densité pour former les clusters.

• **Le clustering basé sur les grilles** consiste à partitionner l'espace en différentes cellules à l'aide d'une grille, puis à identifier les ensembles de cellules denses connectées pour la formation des clusters. Les méthodes appartenant à cette catégorie nécessitent deux paramètres à savoir : la taille de la grille et la densité minimum déterminant si une cellule de la grille est considérée comme dense ou non.

• **Le clustering basé sur les graphes** considère les clusters comme étant des ensembles de nœuds connectés dans un graphe. L'objectif est donc de former le graphe qui connecte les ensembles entre eux de telle manière que la somme des valeurs des arcs correspondant aux distances entre les exemples soit minimale.

Des états de l'art plus détaillés sont disponibles dans la littérature. Le lecteur souhaitant une description plus avancée des méthodes pourra s'y référer ([2],[88], [62]). Le tableau 2.1 présente à titre illustratif un ensemble de caractéristiques associées à certaines méthodes de clustering. Il est à noter qu'aucune méthode de clustering n'est intrinsèquement meilleure que les autres sur l'ensemble des problèmes envisageables.

Caractéristiques	**Hiérarchique**	**K-moyennes**
Connaissances a priori	Nombre de clusters K	Nombre de clusters K
Présentation des résultats	Hiérarchie	K centroïdes
Complexité	$O(M \times N^2)$	$O(M \times N \times K)$
Déterministe	oui	non
Incrémental	non	oui
Prise en compte du contexte	non	non
Tolérance au bruit	non	non
Tolérance aux clusters de tailles variées	oui	non
Tolérance aux clusters de densités variées	oui	oui
Tolérance aux clusters de forme quelconque	oui	non
Tolérance aux clusters concentriques	oui	non

Caractéristiques	**Basé sur la densité**	**Basé sur la grille**
Connaissances a priori	Critère de densité du voisinage	Taille de grille et critère de densité
Présentation des résultats	Partition	Ensemble de cellules connectées
Complexité	$O(M \times N^2)$	$O(M\times$ taille de grille$)$
Déterministe	oui	oui
Incrémental	non	non
Prise en compte du contexte	non	non
Tolérance au bruit	oui	oui
Tolérance aux clusters de tailles variées	oui	oui
Tolérance aux clusters de densités variées	non	non
Tolérance aux clusters de forme quelconque	oui	oui
Tolérance aux clusters concentriques	oui	oui

TABLE 2.1 – Caractéristiques associées aux différentes méthodes de clustering.
N = Le nombre d'observations.
d = Le nombre de variables.
K = Le nombre de clusters.

2.2.3 Validation et Interprétation des résultats

L'évaluation de la pertinence de la partition générée par le clustering est un domaine de recherche très actif. La difficulté de ce problème réside dans le fait que l'évaluation des résultats du clustering est en partie subjective [53]. En effet, pour un même jeu de données, il existe souvent un grand nombre de partitions possibles. De plus, il est impossible de définir un critère universel permettant d'évaluer sans biais les résultats obtenus par l'ensemble des algorithmes de clustering. Pour atteindre cet objectif, de nombreuses techniques ont été développées pour identifier la « meilleure » partition générée par un algorithme de clustering. Cette identification est souvent liée à la méthode utilisée.

Dans la littérature, les critères analytiques permettant de mesurer la qualité des résultats issus des algorithmes de clustering peuvent être catégorisés en trois grandes familles : *interne, externe* et *relatif.*

1. *Les mesures de qualité interne* [71] se calculent uniquement à partir des informations contenues dans les données, sans avoir recours à des connaissances a priori. Ces mesures sont en général des approches non supervisées qui se basent sur des informations internes au clustering. Ces mesures se basent souvent sur la définition intuitive et la plus simple du clustering : les groupes d'instances doivent être les plus compacts possibles (*i.e.,* la similarité intra clusters) et différents les uns des autres (*i.e.,* la similarité inter clusters).

Les mesures internes permettent donc d'évaluer *la compacité* et *la séparabilité* des clusters. Les mesures de qualité internes les plus connues dans la littérature sont : l'indice *Davies Bouldin* [38], *indice SD* [55], *indice Silhouette* [91] *etc.*

2. *Les mesures de qualité externe* [45] s'appuient sur une connaissance a priori des caractéristiques d'un bon clustering. Ces mesures sont en général des approches supervisées qui consistent à mesurer le degré de correspondance entre la partition générée par l'algorithme de clustering et une partition connue des données. De nombreuses mesures ont été proposées pour atteindre cet objectif. Citons par exemple, l'indice Adjusted Rand Index (ARI) [57], l'information mutuelle normalisée (NMI) [96] et l' entropie conditionnelle [45].
3. *Les mesures de qualité relative* permettent de comparer plusieurs partitionnements obtenus à partir d'un même jeu de données. Il s'agit tout simplement de l'utilisation des critères internes ou externes pour choisir la meilleure partition générée par le même algorithme sur le même jeu de données.

Selon les besoins de l'utilisateur, le clustering peut être utilisé pour deux différentes raisons :

1. La tâche de clustering peut être inscrite comme une étape intermédiaire dans un traitement d'apprentissage (voir Figure2.6) : il s'agit ici de considérer le clustering comme une étape de prétraitement utilisée dans une autre tâche telle que la classification supervisée. Une description des clusters (*i.e.,* l'interprétation des résultats générés par le clustering) n'est pas nécessaire dans cette situation. On cherche uniquement dans ce cas à obtenir l'appartenance des observations à l'un des clusters (ID-cluster) sans avoir besoin d'interpréter les résultats issus de l'algorithme. Le point le plus important ici est de s'assurer que la qualité des résultats fournis par l'algorithme est bonne. Pour se faire, les critères d'évaluation cités ci-dessus peuvent être utilisés (*e.g.,* Davies-Boulin "DB", Silhouette, etc).

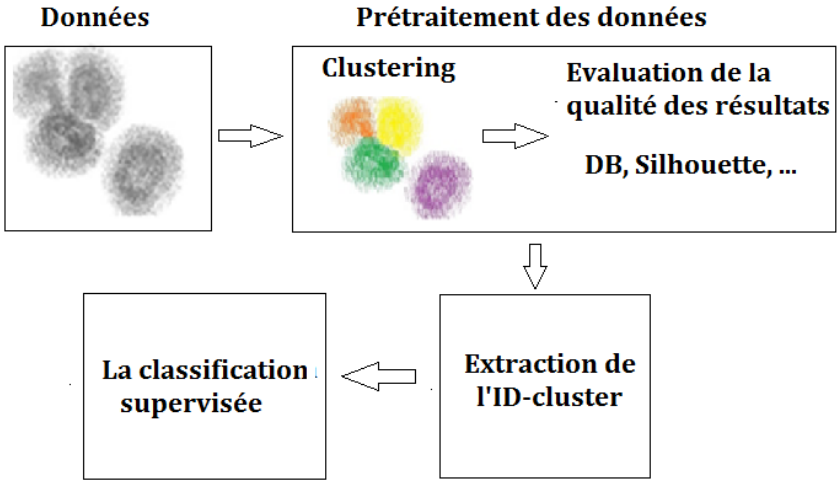

FIGURE 2.6 – Le clustering est considéré comme une étape de prétraitement.

2. Les clusters générés par le clustering constituent un résultat final (Voir Figure2.7). Dans ce cas, le clustering constitue à lui seul un processus global de découverte de groupes. L'exploitation des clusters pour une application donnée passe alors par une description de ces derniers.

FIGURE 2.7 – Le clustering constitue un résultat final

2.3 Approche prédictive : *La classification supervisée*

La classification supervisée cherche à prédire la classe des nouvelles instances en se basant sur des informations connues *a priori*. Elle est un processus à deux étapes : une étape d'apprentissage et une étape de classification.

Dans l'étape d'apprentissage, un modèle est construit en analysant un jeu de données dit *"d'apprentissage"* dans lequel la classe de chaque instance est supposée prédéfinie. Soit $\mathcal{D} = \{(X_i, Y_i), i \in \{1, ..., N\}\}$ un jeu de données d'apprentissage composé de N instances. Chaque instance $(X_i = \{X_i^1, ..., X_i^d\}, Y_i \in \{1, ..., J\})$ est représentée par un vecteur de variables de dimension d et d'une variable cible Y_i indiquant son appartenance à une des J classes. Soit χ et κ respectivement l'espaces des valeurs d'entré et de sortie. D'une manière plus formelle, l'étape d'apprentissage a pour but d'apprendre, à partir des données d'apprentissage, une fonction $f : \chi \rightarrow \kappa$ de telle sorte que $f(X)$ est un "bon" prédicteur de la valeur correspondante à Y.

Dans l'étape de classification, le modèle construit dans la première étape est utilisé pour classer les nouvelles instances.

Le modèle construit par un algorithme d'apprentissage doit en général remplir un certain nombre de critères. Citons à titre d'exemple :

- Le taux d'erreur doit être le plus bas possible. Ce point peut être mesuré en utilisant plusieurs critères d'évaluation. A titre d'exemple, la précision (ACC), l'air sous la courbe de ROC (AUC), l'indice ARI (Adjusted Rand Index),..., *etc.*
- Il doit être aussi peu sensible que possible aux fluctuations aléatoires des données d'apprentissage.
- les décisions de classification doivent autant que possible être explicites et compréhensibles.

Le tableau 2.2 présente une comparaison de quelques modèles de la classification supervisée en se basant sur quelques critères de pertinence. Des états de l'art plus détaillés sont disponibles, le lecteur souhaitant une description plus avancée de ces modèles pourra s'y référer ([35, 68]).

Loin de vouloir donner une description détaillée des différentes méthodes de la classification supervisée, cette section traite exclusivement l'aspect interprétable de quelques méthodes. Dans

Critère	Arbre de décision	SVM	Plus proche voisin	Bayésien naïf
Rapidité d'apprentissage	+	- -	+ +	+
Rapidité et facilité de mise à jour	- -	-	+ +	+ +
Précision	+ +	+ +	+	+
Simplicité (nombre de paramètres)	-	-	+ +	+ +
Rapidité de classement	+ +	-	-	+ +
Interprétabilité	++	-	+ +	+ +
Généralisation - Sensibilité au bruit	-	+	-	+ +

TABLE 2.2 – Comparaison de quelques méthodes de classification suivant quelques critères de pertinence

la littérature, les modèles de la classification supervisée se catégorisent en deux grandes familles : les modèles transparents et les modèles boîtes noires. Les modèles transparents désignent tous les algorithmes d'apprentissage qui fournissent des résultats facilement interprétables par l'utilisateur. Contrairement aux modèles boîtes noires (ou opaque) qui désignent les algorithmes d'apprentissage fournissant des résultats non difficilement compréhensibles par l'utilisateur.

2.3.1 Les modèles transparents

Dans cette section, nous présentons les facteurs principaux permettant aux modèles tels les arbres de décision de générer des résultats compréhensibles par l'utilisateur.

1. **Les arbres de décision** ([87],[25]) figurent parmi les modèles naturellement interprétables. Ils fournissent des règles faciles à interpréter par l'être humain. En effet, l'arbre de décision est en général un classifieur présenté sous forme d'une structure arborescente (*e.g.,* voir la figure 2.8). Chaque nœud de l'arbre est : *i*) soit un nœud "de décision" où des tests ont été effectués sur les valeurs d'une seule variable. *ii*) Soit un nœud "feuille" qui détient la prédiction de la classe. Par conséquent, des règles inductives sont crées pour tous les chemins possibles de la racine à une des feuilles.

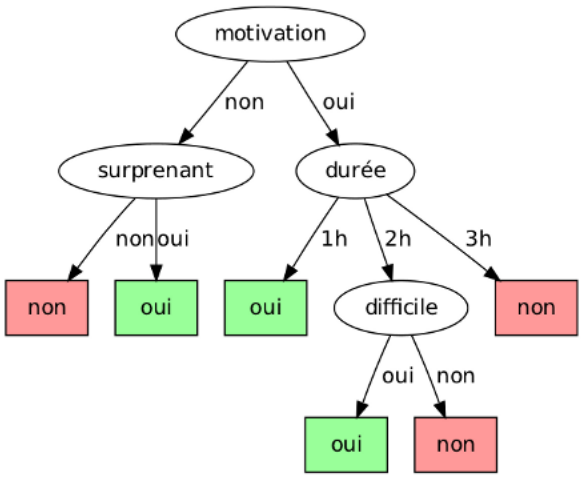

FIGURE 2.8 – Exemple d'arbre de décision pour la question "la présentation est-elle intéressante ?"

2. **Les K plus proches voisins** [37] est un classifieur qui ne nécessite pas d'apprendre une fonction précise d'apprentissage pour qu'il prédise la classe des nouvelles instances. Dans le cas des jeux de données de petites dimensions, ce modèle a la capacité de fournir à l'utilisateur un certain type explication concernant la classification de chaque nouvelle instance. Ces explications sont obtenues à l'aide d'une analyse simple des K plus proches voisins utilisés pour classer une instance. Il est à signaler que dans le cas des jeux de données de grandes dimensions, l'algorithme des K-plus proches voisins devient un modèle boîte noire.

Pour plus de détails sur les différentes raisons permettant à ces modèles d'être interprétables, le lecteur peut se référer à l'article [50].

2.3.2 Interprétation des modèles boîtes noires

Dans certains cas d'étude, l'interprétation des résultats d'un classifieur reste une question secondaire. La performance prédictive du modèle est dans cette situation le point clé pour la résolution des problématiques. Les algorithmes utilisés dans ce cadre privilégient plus le critère de performance prédictive que celui de l'interprétation. Ces modèles sont connus sous le nom *des modèles boîtes noires*. Parmi ces méthodes, on pourra citer notamment les réseaux de neurones (ANN) et les machines à vecteurs de support (SVM).

Les résultats fournis par les modèles boîtes noires sont incompréhensibles et ne conduisent donc pas à des interprétations informatives. Seule une étude des erreurs de prévisions permet de se faire une idée de la qualité du modèle en question. À titre d'exemple :

— *Les réseaux de neurones* (ANN) reçoivent les informations sur une couche réceptrice de "neurones". Ils traitent tout d'abord ces informations avec ou sans l'aide d'une ou plusieurs couches "cachées" contenant un ou plusieurs neurones. Ensuite, ils produisent un (ou plusieurs) signaux de sortie. Généralement, ces sorties sont des vecteurs de lien de connexion qui ne donnent aucune indication supplémentaire sur la contribution des variables lors de la classification supervisée.
— *Le modèle SVM* a pour objectif de trouver l'hyperplan optimal qui sépare au mieux les données dans l'espace d'entré. Cependant, les seules informations fournies par ce dernier sont en général soit les vecteurs de support sans aucune autre information, soit les coefficients de l'hyperplan de séparation et éventuellement le taux de bonnes classifications. L'utilisateur trouve donc une difficulté d'expliquer ce qui fait qu'un individu est dans une classe plutôt que dans une autre.

Pour une interprétation aisée des résultats fournis par ces modèles, plusieurs techniques ont vu le jour. Les techniques proposées peuvent être soit dédiées ou généralistes. Dans le premier cas, les techniques sont fondées sur le fonctionnement interne du modèle en question. Par conséquent, ces techniques ne peuvent être utilisées que pour ce modèle. Au contrario, les techniques généralistes sont des techniques utilisables pour tous les modèles de classification. Le tableau 2.3 présente quelques méthodes de sélection des variables généralisées et dédiées permettant de faciliter la tâche de l'interprétation des résultats issus des ANN et des SVM .

Les méthodes dédiées					les méthodes généralisées			
Méthodes	classifieur	Année	Techniques	Source	Méthodes	Année	Techniques	Source
SVM-RFE	SVM	2002	embedded	[54]	Robnik et al.	2008	filter	[90]
Féraud et al.	ANN	2002	embedded	[47]	Oh et al.	2004	embedded	[83]
MOI	SVM	2004	wrapper	[94]	Penget al.	2005	filter	[84]

TABLE 2.3 – Techniques de sélection des variables pour les ANN et les SVM

La technique d'extraction des règles est également l'une des techniques permettant de résoudre la problématique d'interprétation des résultats générés par les modèles boîtes noires. Le principe de cette technique est d'extraire un nombre réduit de règles qui imitent le fonctionnement des modèles opaques. Elle peut être une technique dédiée si l'approche suivie est «décompositionnelle» et elle peut être généraliste si l'approche suivie est «pédagogique». Les deux notions «décompositionnelle» et «pédagogique» seront discutées dans ce qui suit.

L'importance de la technique d'extraction des règles réside essentiellement dans le fait qu'elle :

1. Fournit un nombre réduit de règles qui imitent *fidèlement* le comportement du modèle boîte noire en termes de prédiction. Cet ensemble de règle est souvent compréhensible par l'utilisateur.
2. Améliore les performances des techniques d'induction des règles en supprimant par exemple le bruit présent dans les données. Cela peut être fait en remplaçant la variable cible par les prédictions faites par le modèle opaque. On constate donc qu'un modèle boîte noire performant peut être utilisé dans une étape de prétraitement pour nettoyer les données [75].
3. Étend l'utilisation du modèle *boîte noire* à des domaines "critiques" (*i.e.*, les domaines où l'interprétation est un critère crucial comme par exemple la médecine).

Lors de la construction d'un algorithme d'extraction des règles, plusieurs questions peuvent apparaitre :

— *Quelle logique faut-il suivre pour former les règles ?*
— *A quel niveau d'apprentissage faut-il extraire les règles ?*
— *Comment mesurer la cohérence entre les règles extraites et les prédictions faites par les modèles opaques ?*
— *... etc.*

Pour répondre à ces différentes questions, Andrews et al. [14] ont proposé un système de classification (basé sur cinq critères) pour l'extraction des règles à partir des réseaux de neurones. Ce système peut être peut être étendu à d'autres modèles opaques tels que les SVM ; Ces critères sont : (1) la transparence de l'algorithme d'extraction par rapport au modèle sous-jacent, (2) la puissance expressive des règles, (3) la qualité des règles extraites, (4) la scalabilité de l'algorithmique et (5) la consistance de l'algorithme.

(1) La transparence de l'algorithme par rapport au modèle : Ce critère définit la relation entre les algorithmes d'extraction des règles et l'architecture interne du modèle sous-jacent. Selon la taxonomie présentée par Andrews et al. [14], il existe trois manières différentes pour

extraire les règles à partir d'un modèle boîte noire à savoir les techniques *décompositionnelles*, les techniques *pédagogiques* et les techniques *éclectiques* :

— *Les techniques décompositionnelles* sont étroitement liées au fonctionnement interne du modèle sous-jacent. Par exemple, pour les réseaux de neurones, ces techniques s'intéressent au fonctionnement interne de chaque neurone du réseau. Ensuite, les règles extraites sont agrégées afin d'avoir une relation globale. Pour les SVM, ces techniques utilisent une approximation locale de la frontière de décision en utilisant des hyper-rectangles ou des ellipsoïdes comme régions dans l'espace d'entrée (voir Figure 2.9 (b)).

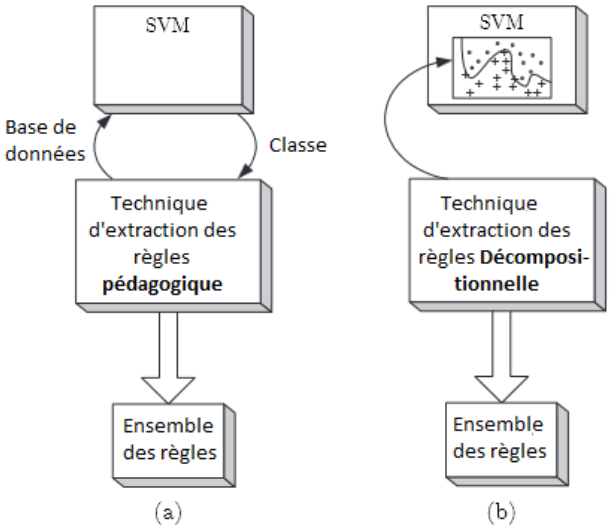

FIGURE 2.9 – Exemple illustratif de la technique décompositionnelle (b) et la technique pédagogique (a) dans le cadre des SVM

— *Les techniques pédagogiques* Les algorithmes qui suivent la technique pédagogique ne sont pas des algorithmes d'extraction des règles au sens strict du mot. En effet, ils extraient directement les connaissances utiles à partir des données d'apprentissage. Pour interpréter les résultats du modèle opaque en utilisant les algorithmes qui suivent cette technique, il suffit donc de remplacer la variable cible par les prédictions générées par celui-ci. À titre d'exemple, voir la figure 2.9 (a) dans le cadre des SVM.

— *Les techniques éclectiques* combinent les deux techniques décompositionnelle et pédagogique à la fois.

(2) La puissance expressive des règles : dépend généralement de la langue utilisée par l'utilisateur pour les exprimer. Autrement dit, elle dépend de la forme de celles-ci. Les règles les plus utilisées dans la littérature, à notre connaissance, sont les règles *propositionnelles*, les règles *logiques de forme* ***M*** *de* ***N*** *conditions*, les règles *floues*, les règles *obliques* et les règles *équations* :

— *Les règles propositionnelles* sont les règles les plus répandues dans la littérature en raison de leur simplicité. Ces règles prennent la forme suivante : "Si *condition* Alors *expression*". Dans le cadre d'extraction des connaissances, la plupart des algorithmes veilleront à ce que les règles soient mutuellement exclusives dans le but de ne pouvoir utiliser qu'une seule règle pour la prise de décision quand une nouvelle observation est présentée. Néanmoins,

il existe aussi des algorithmes qui permettent d'extraire plusieurs règles pour une seule observation, ce qui nécessite l'utilisation d'un mécanisme supplémentaire pour combiner les différentes prédictions. Par exemple, dans [33], Chen associe un facteur de confiance à chacune des règles de telle sorte que les règles tirées avec une grande confiance auraient un grand impact sur la décision finale.

— *Les règles logiques de forme M de N conditions* sont les règles de la forme : "Si *au moins* M de N conditions Alors *expression*" où M est un entier et N est un ensemble de conditions. Ces règles sont étroitement liées aux règles propositionnelles puisqu'elles peuvent facilement être transformées en celles-ci.

— *Les règles floues* sont les règles qui prennent la forme suivante : " Si X est faible et Y est moyen Alors *expression*". Généralement, les règles floues sont à la fois compréhensibles et faciles puisqu'elles sont exprimées par des concepts linguistiques faciles à interpréter par l'utilisateur.

— *Les règles obliques* sont basées sur des fonctions discriminantes par morceaux : "Si $(\alpha x_1 + \beta x_2 > c_1)$ et $(\sigma x_1 + \rho x_2 > c_2)$ et ... Alors *expression*". Ces règles sont généralement plus difficiles à comprendre par rapport aux règles propositionnelles. Toutefois, elles se caractérisent par leurs capacités à créer des frontières qui ne sont pas forcément parallèles aux axes de l'espace d'origine d'entrée. Par conséquent, elles nécessitent moins de conditions que les règles propositionnelles.

— *Les règles équations* sont les règles qui contiennent une fonction polynomiale dans la partie condition : "Si $\alpha x_1^2 + \beta x_2^2 + \sigma x_1 x_2 + \rho x_1 + \varphi x_2$ Alors *expression*". La façon dont ces règles ont été construites les rend plus difficiles à comprendre et par conséquent elles contribuent peu à l'interprétation des modèles opaques.

(3) La qualité des règles extraites : Andrews et al.[14] ont proposé un ensemble de trois critères pour évaluer la qualité des règles à savoir : *La précision*, *la fidélité* et *la compréhensibilité* :

— *La précision* mesure la capacité des règles extraites à prédire correctement les classes des nouvelles instances dans l'ensemble des données de test. Elle est définie généralement comme le pourcentage des instances bien classées.

— *La fidélité* est étroitement liée à la précision. Elle mesure la capacité des règles à imiter la prédiction du modèle d'apprentissage à partir duquel elles ont été extraites.

— *La compréhensibilité* est mesurée par le nombre des règles extraites et le nombre des antécédents par règle (i.e. nombre des variables).

(4) La scalabilité de l'algorithmique : La scalabilité définit généralement la capacité de l'algorithme à faire face aux problèmes de grandes dimensions (un très grand nombre de variables d'entrées) et/ou de grande taille (nombre d'exemples élevé) de la même façon que les problèmes jouets. Bien évidemment, ce critère dépend du temps d'exécution et de la performance de l'algorithme. Cependant, dans le cadre d'extraction des règles, à côté du temps d'exécution de l'algorithme, les règles extraites devraient rester compréhensibles quel que soit la dimension ou la taille de l'ensemble d'apprentissage. La scalabilité mentionne la façon dont le temps d'exécution de l'algorithme et la compréhensibilité des règles extraites varient en fonction de différents facteurs tels que le modèle opaque, la taille de l'ensemble d'apprentissage et le nombre des variables d'entrées [36].

(5) La consistance de l'algorithme : La consistance d'un algorithme peut prendre plusieurs définitions ; Par exemple, elle peut être définie comme étant la capacité à générer, sous différentes sessions d'apprentissage, des règles avec les mêmes degrés d'accuracy. En outre, dans

[63], Johansson *et al.* définissent la consistance de l'algorithme comme étant sa capacité à extraire des règles similaires à chaque fois qu'il est appliqué à un même ensemble de données. Cependant, les auteurs soulignent immédiatement la difficulté associée à cette définition, puisqu'il n'y a pas de définition simple de similitude des règles (à notre connaissance).

Les tableaux 2.4 et 2.5 présentent respectivement quelques méthodes d'extraction des règles (décompositionnelle et pédagogique) permettant de faciliter la tâche d'interprétation des résultats issus des ANN et des SVM .

Les réseaux de neurones (ANN)				
Méthode	**Année**	**Technique**	**Type**	**Source**
RN2	2002	Crée des règles de forme polynomiale, RN2 garantit de produire des unités dans les unités cachées et regroupe les valeurs d'activation des unités cachées en utilisant un algorithme de clustering.	D	[92]
REFANN	2002	Approxime la fonction d'activation des ANN par des fonctions linéaires par morceaux	D	[93]
GEX	2004	Algorithme génétique, fournit des règles propositionnelles	P	[74]
BUR	2004	Basé sur " Gradient Boosting Machines"	P	[33]
ITER	2006	Basé sur une augmentation itérative des hypercubes	P	[58]
Coalition-opposition	2007	Extrait des coalitions et des oppositions minimaux du neurone à partir d'un arbre.	D	[16]

TABLE 2.4 – Techniques décompositionnelles (D) et pédagogiques (P) d'extraction des règles à partir des ANN

Les machines à vecteur de support (SVM)				
Méthode	**Année**	**Technique**	**Type**	**Source**
CART	1984	Arbre de décision	P	[25]
CN2	1989	Induction par règles, algorithme de recouvrement séquentiel	P	[34]
C4.5	1993	Arbre de décision	P	[87]
SVM + prototype	2002	Clustering, basé sur des régions ellipsoïdes et hyper-rectangulaires	D	[79]
RulExSVM	2004	Fournit des règles propositionnelles, cet algorithme est basé sur des régions hyper-rectangulaires	D	[103]
HRE	2005	SVC (Support Vector Clustering), basé sur des règles hyper-rectangulaires	D	[104]
Fung et al.	2005	Algorithme itératif pour les SVM linéaire, basé sur une programmation linéaire	D	[51]
Hien et al.	2014	Fournit des règles floues	D	[80]

TABLE 2.5 – Techniques décompositionnelles (D) et pédagogiques (P) d'extraction des règles à partir des SVM

2.4 Interprétation

L'interprétation des résultats issus d'un modèle (descriptif ou prédictif) est un acte très subjectif. Il dépend généralement de l'utilisateur et du domaine d'application. En effet, les connaissances utiles doivent être exprimées dans la langue et la sémantique de celui-ci. De plus, l'interprétation des résultats du modèle diffère d'un domaine d'application à l'autre en fonction du niveau de détails demandé (voir Section 2.4.3).

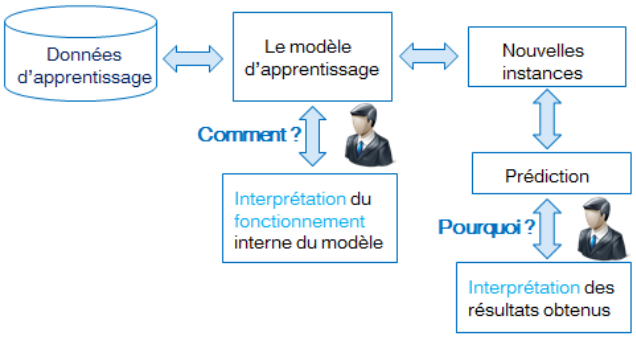

FIGURE 2.10 – Emplacement des deux types d'interprétation au cours du processus de la classification

Dans le cadre de l'apprentissage automatique, deux types d'interprétation se distinguent : l'interprétation du modèle et l'interprétation des résultats issus de ce dernier. Dans le premier cas, l'utilisateur cherche à expliquer comment le modèle fonctionne afin d'obtenir ses résultats. Il cherche à comprendre la logique suivie par le modèle pour atteindre l'objectif désiré. Dans le deuxième cas, l'utilisateur s'intéresse à expliquer pourquoi un tel résultat a été obtenu par le modèle. Il s'agit donc de connaître les différents facteurs qui ont un grand impact sur l'obtention de ce résultat. A titre d'exemple, l'interprétation des machines à vecteurs de support (SVM) se fait à travers la compréhension de son fonctionnement interne (e.g l'obtention des vecteurs supports, de l'hyperplan de séparation des classes, etc.). L'interprétation des résultats des SVM, quant à elle, se fait à l'aide de la compréhension des relations présentes entre les entrées et la sortie de l'algorithme. Il s'agit alors de détecter les causes des prédictions (voir Section 2.4.1). La figure 2.10 illustre schématiquement l'emplacement de ces deux types d'interprétation au cours du processus de classification.

Généralement, on dit d'un modèle qu'il est interprétable s'il est capable de fournir à l'utilisateur des résultats compréhensibles : si, à partir des résultats obtenus par le modèle, l'utilisateur peut extraire facilement les connaissances utiles en se basant sur les variables natives. A titre d'exemple :

1. **Dans le cadre du clustering** : On dit qu'un modèle est interprétable s'il permet à l'utilisateur de comprendre les causes de la formation des clusters. On parle ici d'une interprétation locale (voir Section 2.4.3). Il s'agit donc d'identifier les variables qui contribuent le plus à la formation de chaque cluster. En effet, certaines variables peuvent être discriminantes pour la formation d'un cluster et s'avérer peu révélatrices pour la formation d'autres. Dans le cas où on s'intéresse au traitement global du processus de clustering, la description des clusters n'est pas nécessaire. Seule une analyse de qualité peut être

suffisante. On parle d'ici d'une interprétation globale (voir Section 2.4.3). Pour plus de détails sur l'interprétation des résultats de clustering voir Section 2.2.3.

2. **Dans le cadre de la classification** : On dit qu'un modèle est interprétable s'il permet de répondre aux deux questions suivantes : (1) Quelles sont les causes des prédictions ? (2) Quelle est la fiabilité d'une prédiction ?. Les réponses à ces deux questions sont présentées respectivement dans les deux sections 2.4.1 et 2.4.2.

2.4.1 Les raisons d'une prédiction

Intuitivement, la façon qui permet de comprendre les causes d'une prédiction est de réaliser à quel point les variables ont contribué au résultat du modèle. Généralement, les données d'apprentissage dont on dispose ne contiennent pas forcément que des variables pertinentes. Il est possible que certaines variables correspondent à du bruit ou qu'elles soient corrélées, redondantes, peu informatives ou même inutiles au problème de la classification. Les variables pertinentes sont souvent celles qui ont un fort impact sur la prédiction par rapport à celles non informatives. De ce fait, on cherchera à **"mesurer l'importance des variables"** en fonction de leur contribution dans le processus d'apprentissage.

Il est à noter que la mesure d'importance des variables peut être réalisée à plusieurs niveaux : avant, après ou au cours du processus d'apprentissage.

— *Avant l'apprentissage* : les informations disponibles à ce stade sont seulement les données. Cette technique est indépendance du modèle sous-jacent. L'utilisateur cherche donc à identifier les variables les plus importantes en s'appuyant sur les propriétés des données. C'est-à-dire celles qui séparent aux mieux les données.
— *Au cours de l'apprentissage* : Cette technique est étroitement liée au modèle d'apprentissage. Elle intègre les performances prédictives du classifieur dans la procédure du calcul d'importance des variables. La mesure de cette importance s'effectue dans ce cas suivant une procédure itérative : elle compare plusieurs résultats issus du modèle en utilisant le même jeu de données afin d'évaluer l'importance des variables.
— *Après l'apprentissage* : s'appuie sur une connaissance *a priori* des caractéristiques d'un bon modèle et du résultat obtenu par ce dernier. A ce stade, le résultat final généré par le modèle est considéré comme une référence afin de mesurer l'importance des variables.

Le calcul de l'importance d'une variable permet de mesurer à quel point l'information qu'elle contient a été décisive dans l'obtention du résultat. Dans le cas où l'on dispose d'un nombre important de variables ou lorsqu'on utilise un modèle boîte noire (voir Section 2.3.2), plusieurs techniques ont vu le jour permettant ainsi une interprétation aisée des résultats. A travers l'utilisation de ces techniques, l'utilisateur peut concentrer son attention sur les variables les plus pertinentes :

• **Le tri des variables en fonction de leur importance** consiste à classer les variables en fonction de leur contribution à la formation : *i*) des groupes d'instances homogènes si on est dans le cadre du clustering ou *ii*) des groupes d'instances ayant la même classe si on est dans le cadre de la classification supervisée. Cette technique permet à l'utilisateur de se concentrer sur les variables importantes pour une interprétation aisée. L'importance des variables peut être mesurée à l'aide de plusieurs critères. A titre d'exemple, on citera l'information mutuelle et le critère de Fisher pour la classification supervisée et l'indice Davies Bouldin et la Silhouette pour le clustering.

• **La pondération** consiste à affecter des poids aux variables qui évoluent au cours de l'apprentissage en fonction de leur importance. Il s'agit donc de donner un rôle plus important

aux variables contenant l'information intéressante pendant l'apprentissage. Ces poids servent ensuite lors de la phase d'interprétation des résultats.

• **La sélection des variables** a pour objectif de trouver un sous-ensemble *pertinent* de variables parmi celles de l'ensemble de départ. L'avantage de cette méthode réside dans le fait qu'elle permet : (1) d'améliorer souvent la performance prédictive du modèle, (2) de faciliter l'interprétation des résultats, et (3) d'étendre l'utilisation des modèles *boîtes noires* à des domaines "critiques". Le processus de sélection des variables passe généralement par trois étapes différentes, à savoir : (1) un algorithme de recherche, (2) un critère d'évaluation, et (3) un critère d'arrêt.

(1) L'algorithme de recherche a pour objectif d'explorer l'espace de combinaison des variables. Il peut être *Exhaustif*, *Heuristique* ou *Aléatoire* :

— *Exhaustif* : Cette catégorie consiste à sélectionner le meilleur sous-ensemble des variables parmi tous les sous-ensembles existant en faisant une recherche exhaustive. Cependant, le problème majeur de cette stratégie est que le nombre de combinaisons des variables possibles croit exponentiellement quand le nombre des variables augmentent. Ceci rend la recherche exhaustive quasiment impossible.

— *Heuristique* : Les algorithmes qui appartiennent à cette catégorie sont généralement les algorithmes itératifs ; A chaque itération, une ou des variables peuvent être ajoutées ou rejetées selon leurs importances. Ces algorithmes sont connus généralement par leur simplicité et leur rapidité. Dans la littérature, cette catégorie se divise en trois types de procédure de recherche à savoir *Forward*, *Backward* et *Stepwise*.

1. *Forward* : L'objectif de cette procédure est de partir d'un ensemble vide de variables et d'ajouter successivement, à chaque itération, une ou des variables pertinentes.
2. *Backward* : Le principe de cette procédure de recherche est de partir de l'ensemble global de toutes les variables et de supprimer séquentiellement, à chaque itération, une ou des variables (les moins pertinentes).
3. *Stepwise* : L'idée centrale de cette approche est d'ajouter ou de rejeter une ou des variables au sous-ensemble de variables courant.

— *Aléatoire* : L'idée derrière cette catégorie est de générer aléatoirement un nombre de sous-ensembles de variables afin de sélectionner le 'meilleur' sous-ensemble parmi ces derniers. Cette catégorie est appelée aussi, l'approche stochastique.

(2) Le critère d'évaluation permet de mesurer la qualité d'un sous-ensemble de variables suivant que l'on utilise l'approche *filter*, *wrapper* ou *embedded* :

— *L'approche 'filter'* : Cette approche est indépendante du modèle sous-jacent. Autrement dit, elle présélectionne les variables, puis elle utilise ces variables sélectionnées dans le modèle d'apprentissage. Cependant, cette approche peut être considérée comme une étape de prétraitement (avant la phase d'apprentissage). L'approche *filter* repose sur le calcul d'un score qui permet de calculer la pertinence de chaque variable en s'appuyant sur les propriétés des données d'apprentissage. Plusieurs scores ont été proposés dans la littérature permettant ainsi de mesurer la pertinence d'une variable. Ces scores peuvent aussi être utilisés comme un critère d'évaluation. Parmi ces scores, on trouve *le critère de Fisher*, *le critère de corrélation* et *l'information mutuelle*.

— *L'approche 'wrapper'* : Contrairement à l'approche *filter*, les méthodes appartenant à l'approche *wrapper* intègrent les performances prédictives du classifieur dans la procédure de recherche pour évaluer la qualité des sous-ensembles de variables.

— *L'approche 'embedded'* : Le principe de cette approche est d'incorporer la sélection lors du processus d'apprentissage. La sélection des variables s'appuie donc sur un critère propre à la méthode. Les arbres de décision sont l'illustration la plus emblématique. Dans les méthodes de sélection de type "*wrapper*", la base d'apprentissage est divisée en deux parties : une base d'apprentissage et une base de validation pour valider le sous-ensemble de caractéristiques sélectionné. En revanche, les méthodes "*embedded*" peuvent se servir de tous les exemples d'apprentissage pour établir le système.

(3) Les critères d'arrêt les plus utilisés, pour les trois approches cités ci-dessus ('*filter*', '*wrapper*' et '*embedded*') sont :

— *Pour l'approche filter* : Une fois les variables triées selon leur importance dans le processus, l'utilisateur peut sélectionner les variables les plus pertinentes afin de les utiliser par un classifieur. Le nombre de variables sélectionné est donc fixé a priori par celui-ci.

— *Pour l'approche wrapper* : Le critère d'arrêt est basé sur une fonction d'évaluation et dépend de deux faits : *i*) l'ajout ou la suppression d'une variable ne produit aucun sous ensemble plus performant, et *ii*) le sous ensemble obtenu est, d'après la fonction d'évaluation, le sous ensemble optimal. Le processus continue jusqu'à ce que le critère d'arrêt soit satisfait.

— *Pour l'approche embedded* : Le processus de recherche s'arrête lorsque la précision dépasse un certain seuil fixé a priori par l'utilisateur.

2.4.2 La fiabilité d'une prédiction

Les prédictions générées par les modèles d'apprentissage sont souvent susceptibles de contenir des erreurs. Les critères permettant de mesurer la qualité prédictive des modèles d'apprentissage (e.g., l'accuracy) fournissent souvent à l'utilisateur une information "globale" sur leurs capacités à bien classer des nouvelles instances. Néanmoins, ces critères ne fournissent pas une information "locale" sur l'erreur de prédiction prévue pour chaque instance en particulier. D'un autre côté, les modèles les plus performants comme par exemple les SVM sont des modèles complexes qui ne permettent pas à l'utilisateur de comprendre facilement pourquoi une prédiction particulière a été faite. À partir de ce constat, on trouve qu'il est utile que l'utilisateur soit informé de la fiabilité de chaque prédiction.

Pour répondre à cette problématique, il semble important de comprendre la relation entre les variables descriptives et la valeur prédite. Plus précisément, il s'agit ici de détecter l'effet du changement de la valeur de chaque variable sur la valeur prédite. La mesure de tel effet permet d'une part de connaître l'importance de chaque variable et d'autre part de déterminer les valeurs qui sont les seuils du changement de la valeur prédite. Plusieurs travaux ont été effectués dans ce cadre, on pourra citer notamment ceux proposés par Robnik *et al.* [90] et par Briesemeister *et al.* [27].

2.4.3 La granularité d'une interprétation

Selon le besoin de l'utilisateur et du domaine d'application auquel il s'intéresse, le concept d'interprétation peut prendre plusieurs formes : *individuelle, locale et globale.*

Interprétation individuelle : Dans ce cas, on s'intéresse à l'interprétation de la prédiction prévue pour une instance en particulier. Il s'agit donc de découvrir les différents facteurs qui influent la sortie de cette instance. Pour atteindre cet objectif, l'utilisateur doit donc déterminer

la relation entre les valeurs des variables descriptives et la valeur prédite pour cette instance. Ceci peut être réalisé à titre d'exemple : *i*) en mesurant l'importance des variables selon leur contribution à l'obtention de cette estimation. *ii*) en déterminant la fiabilité de cette prédiction individuelle (voir Section 2.4.2), *etc.*. Ce genre d'interprétation est très utile dans plusieurs domaines d'application. Par exemple, dans le domaine de la finance : lors de la prise de la décision d'approbation des crédits dans les banques, l'agent doit être capable d'expliquer au client les raisons du refus d'un crédit.

Interprétation locale : Dans ce cas, on s'intéresse à la description d'un groupe en particulier. Il s'agit ici de comprendre pourquoi les instances appartenant à un même groupe ont la même sortie. Ceci revient à réaliser à quel point les valeurs de chaque variable influent sur la décision prédictive du groupe en question. Que ce soit dans le cadre de la classification supervisée ou le cadre du clustering, plusieurs techniques ont été proposées pour atteindre cet objectif. Parmi ces méthodes, on trouve l'importance des variables, la pondération, la sélection des variables, *etc.* Ce genre d'interprétation est très utile dans plusieurs domaines d'application. Par exemple, dans le cadre de la Gestion de la Relation Client, l'agent doit connaître les besoins et les attentes des clients en fonction de leurs comportements afin qu'il puisse adapter les actions marketing vers ces clients.

Interprétation globale : Dans ce cas, la description des groupes ou d'une instance en particulier n'est pas nécessaire. Seul un traitement global des résultats est suffisant. L'utilisateur cherche : *i*) soit à découvrir l'ensemble des facteurs qui influent les décisions prises. On parle ici de la contribution des variables dans l'obtention des résultats. *ii*) soit à orienter le traitement vers l'analyse globale de la qualité des résultats obtenus. Cette qualité dépend de l'apprentissage suivi. Par exemple, dans le cadre de la classification supervisée, les critères de qualité les plus utilisés sont : l'accuracy (ACC), AUC, la courbe de Lift, l'ARI (Adjusted Rand Index),..., *etc.* Dans le cadre du clustering, les critères de qualité les plus utilisés sont : le critère MSE, l'inertie intra\inter clusters, la Silhouette [91], l'indice SD [55], l'indice Davies-Bouldin [38],..., *etc.*

Le tableau 2.6 présente les différents moyens permettant d'aboutir à ces formes d'interprétation.

	Classification supervisée			Clustering			
	Importance	Sélection	Qualité	Importance	Sélection	Qualité	Profil moyen
individuelle	✓	✓	–	✓	✓	–	–
Locale	✓	✓	–	✓	✓	–	✓
Globale	✓	✓	✓	✓	✓	✓	✓

TABLE 2.6 – Techniques permettant de réaliser les différents types d'interprétation

2.5 Approche descriptive et prédictive simultanément

2.5.1 Contexte

Depuis quelques années, les chercheurs ont concentré leur attention sur l'étude d'un nouvel aspect d'apprentissage. Ce dernier fusionne à la fois les caractéristiques de la classification supervisée (la prédiction) et du clustering (la description). Les algorithmes appartenant à ce type

d'apprentissage cherchent à décrire et de prédire d'une manière simultanée. Autrement dit, ces algorithmes visent à découvrir la structure interne de la variable cible. Puis, munis de cette structure, ils cherchent à prédire la classe des nouvelles instances.

Dans le domaine de l'apprentissage automatique, il existe principalement deux axes à traiter, à savoir l'axe de prédiction et l'axe de description. Dans le premier axe, on cherche à prédire une valeur (pour la régression) ou une classe (pour la classification supervisée) pour une nouvelle donnée à partir d'un ensemble de données d'apprentissage (voir Section 2.3). Contrairement au deuxième axe où l'on cherche à découvrir la structure sous-jacente d'un ensemble de données (voir Section 2.2) à partir de la distinction des groupes d'instances homogènes.

À côté de ces deux axes, on peut ajouter un troisième axe qui est l'axe d'interprétation (voir Section 2.4). Dans certains domaines appelés "domaines critiques" (*e.g.,* la médecine, les services bancaires, *etc*), la compréhension (ou la description) des résultats issus d'un modèle d'apprentissage est une condition aussi importante que sa performance prédictive. Dans ces domaines, l'utilisateur doit avoir un certain niveau de confiance vis-à-vis des hypothèses générées par le modèle. L'accident nucléaire de Three Mile Island [1] est l'un des exemples concrets qui montre la nécessité d'utiliser un modèle à la fois performant et interprétable. Le facteur majeur derrière cet incident est que la personne n'a pas eu confiance dans les recommandations faites par la machine. Le domaine médical est aussi l'un des domaines critiques où la vie de l'être humain est mise en jeu. Lors de la prise d'une décision en se basant sur un modèle d'apprentissage, le médecin doit être précis et convaincant. Par exemple, si cette décision conduit à un préjudice majeur pour le patient, alors le médecin doit être capable de défendre sa décision s'il est accusé de négligence médicale. Dans de telles situation, la qualité des modèles utilisés réside dans leurs capacités à fournir des résultats étant à la fois performants en prédiction et compréhensibles.

Les algorithmes de la classification supervisée traitent principalement l'axe de prédiction. Le but majeur de ces algorithmes est d'apprendre, à partir d'un ensemble de données, un modèle permettant de prédire ultérieurement la classe de nouvelles instances. Pour certains algorithmes de la classification supervisée, ce processus est effectué sans prendre en considération la probabilité d'existence d'une structure sous-jacente au sein d'au moins une des classes (*i.e.,* la description). À titre d'exemple, le jeu de données présenté dans la figure 2.11 est caractérisé par la présence de trois sous-groupes différents pour la classe Virginica et de deux sous-groupes différents pour la classe Setosa.

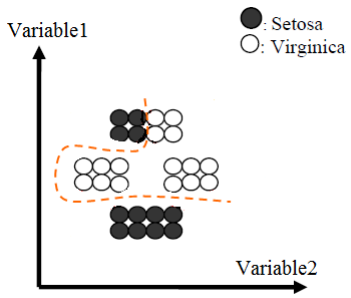

FIGURE 2.11 – Principe de la classification supervisée via les SVM

Pour cet exemple illustratif, un algorithme d'apprentissage tel que les machines à vecteurs de support (SVM) va construire une frontière de décision non linéaire séparant les deux classes sans accorder la moindre importance à la structure interne de la variable cible. On déduit que

1. http ://en.wikipedia.org/wiki/Three_Mile_Island_accident

certains d'algorithmes de la classification supervisée peuvent avoir du mal à décrire l'ensemble des données. Concernant l'axe d'interprétation, les algorithmes de la classification supervisée peuvent être divisés en deux grandes catégories. La première catégorie englobe l'ensemble des algorithmes d'apprentissage performants mais qui fournissent des résultats difficilement immédiatement compréhensibles par l'utilisateur. C'est le cas des modèles appelés "boîtes noires" (*e.g.,* les ANN et les SVM, voir Section 2.3.2). La deuxième catégorie, quant-à-elle, englobe l'ensemble des algorithmes d'apprentissages qui sont naturellement plus interprétables. Ces derniers sont souvent des algorithmes moins performants par rapport aux modèles boîtes noires. C'est le cas des modèles transparents (*e.g.,* les arbres de décision, voir Section 2.3.1).

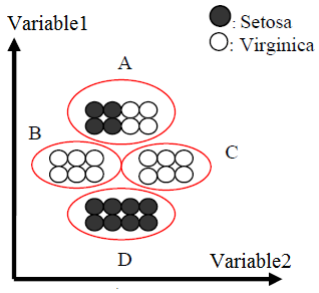

FIGURE 2.12 – Principe du clustering

Les algorithmes de clustering, quant à eux, traitent principalement l'axe de description. En effet, ces algorithmes cherchent à subdiviser l'ensemble des données en un certain nombre de groupes (ou clusters) de manière à ce que les instances soient similaires au sein de chaque groupe et dissimilaires d'un groupe à l'autre. Cette notion de similarité/dissimilarité entre les individus est définie dans le cadre non supervisé (c'est-à-dire, l'absence d'une classe à prédire). Cependant, deux instances proches en termes de distance peuvent appartenir à des classes différentes. Par exemple, le groupe A de la partie gauche de la figure 2.12. A partir de ce constat, on déduit que les algorithmes classiques de clustering peuvent avoir du mal à prédire la classe des nouvelles instances.

À partir de ce constat, on déduit que les algorithmes de la classification supervisée et du clustering ont du mal à décrire et à prédire d'une manière simultanée sous la contrainte d'interprétation. Pour tenter de résoudre cette problématique, deux grandes voies existent (voir la figure 2.13).

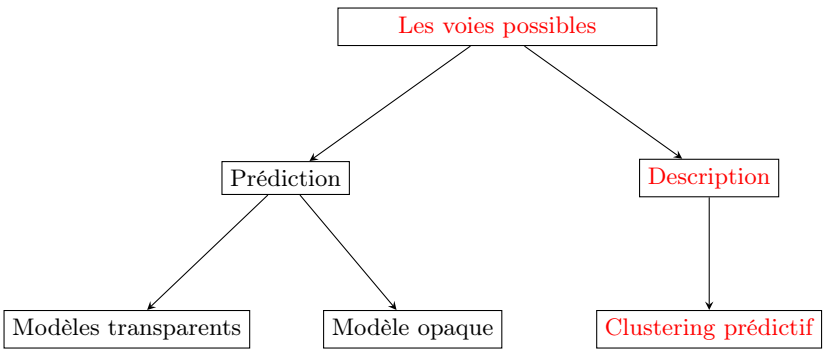

FIGURE 2.13 – Les différentes voies possibles pour la description et la prédiction simultanées

La première voie consiste à adapter les algorithmes de la classification supervisée au problème de la description des données. Dans ce cas, on se trouve face à deux possibilités :

1. rendre les algorithmes transparents plus performants en prédiction et en description sans perdre leur faculté d'intelligibilité des résultats.
2. modifier les modèles boîtes noires pour permettre une bonne description des données. Ces algorithmes doivent posséder également une technique permettant une interprétation facile des résultats.

La deuxième voie, quant-à-elle, consiste à rendre les algorithmes de clustering plus performant en prédiction sans perdre leur faculté à bien décrire les données. Ce type d'apprentissage est appelé *le clustering prédictif*. Dans cette thèse, nous nous intéressons exclusivement à la problématique traitée par cette deuxième voie.

2.5.2 Clustering prédictif

A. Définition

Le clustering prédictif englobe principalement l'ensemble des algorithmes de clustering soumise à des modifications dans le but de les adapter au problème de la classification supervisée. Ceci est effectué en préservant la faculté de l'algorithme à bien décrire les données. L'objectif majeur des algorithmes du clustering prédictif est de découvrir dans la phase d'apprentissage la structure interne de la variable cible. Puis, munie de cette structure, ces algorithmes cherchent à prédire la classe des nouvelles instances. Dans la littérature, il existe deux grandes catégories du clustering prédictif (voir les deux figures 2.14 et 2.15). **La première catégorie** privilège l'axe de prédiction par rapport aux deux autres axes (*i.e.,* l'interprétation et la description) tout en exigeant de minimiser le nombre de groupes appris dans la phase d'apprentissage (voir la figure 2.14). Par contre, les algorithmes de **la deuxième catégorie** cherchent dans la phase d'apprentissage à réaliser le compromis entre la description et la prédiction en découvrant la structure interne *complète* de la variable cible (voir la figure 2.15).

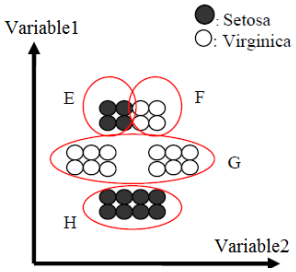

FIGURE 2.14 – Premier type du clustering prédictif

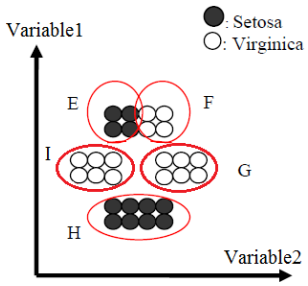

FIGURE 2.15 – Deuxième type du clustering prédictif

B. État de l'art

Dans la littérature, il existe plusieurs variations du clustering prédictif, à savoir, la décomposition des classes, l'arbre de décision, le clustering supervisé, le clustering prédictif de Dzeroski *et al.* et le clusterwise.

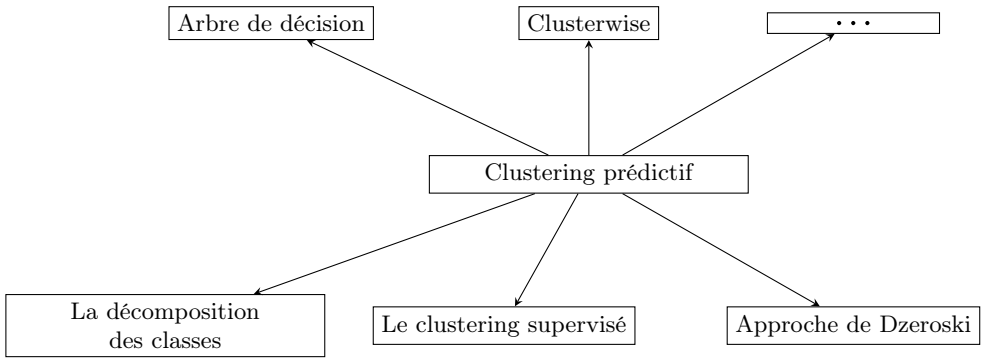

FIGURE 2.16 – Les différentes variations du clustering prédictif

B.1. la décomposition des classes

La technique de décomposition des classes consiste à décrire chaque classe individuellement en utilisant un algorithme de clustering. Chaque classe est donc décrite par un certain nombre de clusters. Ce nombre peut être différent d'une classe à l'autre. À la fin de la phase d'apprentissage, on obtient alors P clusters résultants ($P = \sum_{j=1}^{J} K_j$) avec J est le nombre de classes dans le jeu de données. Dans la littérature, la décomposition des classes est utilisée souvent pour améliorer la performance des classifieurs linéaires simples comme les SVM linéaires (*e.g.,* Vilalta et al. et Wu et al.). Cette technique se résume en deux étapes principales : (1) réalisation d'un clustering de type k-moyennes (où K_j est selon les auteurs une entrée ou une sortie de l'algorithme) par groupe d'exemples qui appartiennent à la même classe j, (2) entraînement d'un classifieur sur les P classes résultantes et interprétation des résultats.

La technique de la décomposition des classes engendre dans la phase d'apprentissage des groupes totalement purs en termes de classes (traitement individuel de chaque classe). De plus, chaque groupe formé contient probablement les instances les plus homogènes possibles et qui diffèrent des instances des autres groupes en raison de l'utilisation d'un algorithme de clustering pour décrire les classes. Cependant, dans le cadre du clustering prédictif, cette technique risque de générer, dans la phase de test, des prototypes (si un algorithme de partitionnement est utilisé) virtuels en raison de forte proximité entre deux prototypes de classes différentes dans la phase d'apprentissage. À titre d'exemple, si le jeu de données contient du bruit comme illustré dans la figure 2.17, les clusters de différentes classes formés dans la phase d'apprentissage peuvent être proches les uns des autres (par exemple, les deux clusters A et B de la figure 2.17). Par conséquent, la probabilité d'affecter les nouvelles instances à seulement l'un des clusters est importante.

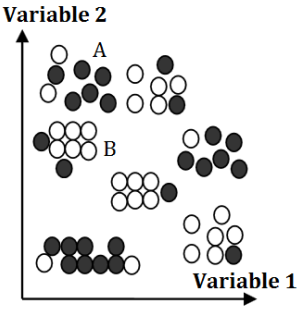

FIGURE 2.17 – Problème d'une classification binaire

B.2. Les arbres de décision

Un arbre de décision peut être considéré comme une hiérarchie de clusters, où chaque nœud représente un cluster. Un tel arbre est appelé un arbre de clustering. Sa structure récursive contient une combinaison de nœuds et de feuilles internes (Figure 2.18). Chaque nœud spécifie un test à effectuer sur une seule variable et ses branches indiquent les résultats possibles du test. Une instance peut alors être classée suivant l'un des chemins de la racine vers un nœud de feuille.

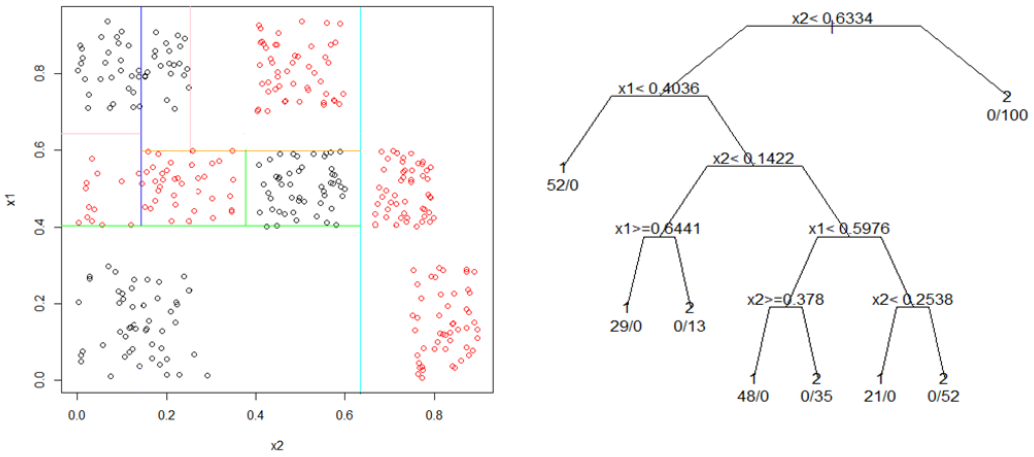

FIGURE 2.18 – Résolution d'un problème de classification binaire par le biais d'un arbre de décision. Les feuilles étant pures en termes de classes, l'arbre ne se développe plus.

Grâce à sa structure arborescente, cet algorithme a la capacité de fournir à l'utilisateur des résultats compréhensibles (sous formes de règles) qui semblent donner une structure à la variable cible apprise. Les instances obtenues suivant un certain chemin ont normalement la même classe et partagent ainsi les mêmes caractéristiques. Cet algorithme semble donc approprié pour atteindre l'objectif de la première catégorie des approches prédictives et descriptives d'une manière simultanée. Cependant, les arbres de décision sont dans certains cas incapables d'atteindre l'objectif de la deuxième catégorie des approches prédictives et descriptives d'une manière simultanée. En effet, selon la distribution des données dans l'espace d'entrée, l'arbre de décision crée naturellement des polytopes (fermés et ouverts) à l'aide des règles. La présence des polytopes ouverts empêche l'algorithme de découvrir la structure *complète* du concept cible Y. La figure 2.18 présente un exemple illustratif du fonctionnement de l'arbre de décision sur un jeu de données caractérisé par la présence de deux classes ('rouge' et 'noire'), 350 instances et deux variables descriptives $x1$ et $x2$. A partir de ce résultat, on constate que cet algorithme fusionne les deux sous-groupes de classe rouge (situés à la droite de la figure), bien que les exemples du premier sous-groupe ont des caractéristiques différentes de celles du deuxième sous-groupe. Dans le cas extrême, ces deux sous-groupes peuvent même être très éloignés et donc être de caractéristiques assez différentes.

Dans la littérature, les améliorations apportées sur la performance prédictive du modèle sont effectuées en ignorant la contrainte d'interprétation. A titre d'exemple, la présence des méthodes d'ensembles (*e.g.,* Boosting, les forêts d'arbres aléatoires, *etc*).

B.3. Le clustering supervisé

Dans la littérature, plusieurs algorithmes de clustering standard ont été soumis à des modification afin qu'ils soient adaptés au problème supervisé. Ces algorithmes sont connus sous le nom de *clustering supervisé* (ou en anglais *supervised clustering*). La différence entre le clustering standard (non supervisé) et le clustering supervisé est donnée dans la figure 2.19. Les algorithmes de clustering supervisé visent à former des clusters purs en termes de classe tout en minimisant le nombre de clusters K (la première catégorie du clustering prédictif). Cette contrainte sur K va empêcher ces algorithmes de clustering supervisé de découvrir la structure complète du concept cible. De ce fait, un seul cluster peut donc contenir un certain nombre de sous-groupes distincts (voir le cluster G de la figure 2.19 b)).

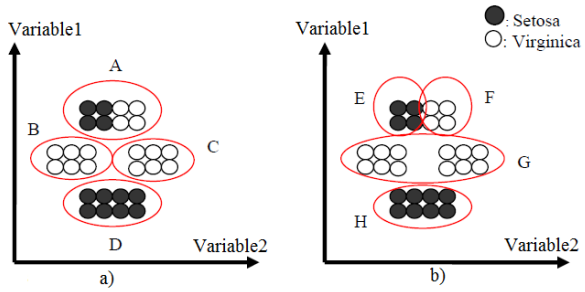

FIGURE 2.19 – La différence entre le clustering standard a) et le clustering supervisé b)

Les algorithmes de clustering supervisé les plus répandus dans la littérature sont :

• Al-Harbi *et al.* [6] proposent des modifications au niveau de l'algorithme des K-moyennes. Ils remplacent la distance euclidienne usuelle par une distance euclidienne pondérée. Le vecteur de poids est choisi de telle sorte que la confiance des partitions générées par l'algorithme des K-moyennes soit maximisée. Cette confiance est définie comme étant le pourcentage d'objets classés correctement par rapport au nombre total d'objets dans le jeu de données. Dans cet algorithme, le nombre de clusters est une entrée.

• Aguilar *et al.* [4] et Slonim *et al.* [95] ont proposé des méthodes basées sur l'approche agglomérative ascendante. Dans [4], les auteurs ont proposé un nouvel algorithme de clustering hiérarchique (S-NN) basé sur les techniques du plus proche voisin. Cet algorithme commence par N clusters où N est le nombre d'objets du jeu de données. Ensuite, il fusionne successivement les clusters ayant des voisins identiques (*i.e.,* objets proches ayant la même étiquette). Par conséquent, tous les voisins ayant les distances plus courtes que le premier ennemi (*i.e.,* l'objet qui n'a pas la même étiquette) seront collectés. Tishby *et al.* ont introduit dans [99] la méthode 'information bottleneck'. Basée sur cette méthode, ils ont proposé une nouvelle méthode de clustering (agglomérative) [95] qui maximise d'une manière explicite, l'information mutuelle entre les données et la variable cible par cluster.

• Dans [32], Cevikalp *et al.* ont proposé une méthode, nommée HC, qui crée des clusters homogènes. Ces travaux sont effectués dans le but de trouver le nombre et l'emplacement initial des couches cachées pour un réseau RBF. Cevikalp *et al.* supposent que les classes sont séparables puisqu'ils cherchent des clusters purs en classe. Le nombre et l'emplacement des clusters sont déterminés en fonction de la répartition des clusters ayant des chevauchements entre les classes. L'idée centrale de l'algorithme HC est de partir d'un nombre de clusters égal au nombre de classes

puis de diviser les clusters qui se chevauchent en tenant compte de l'information supplémentaire donnée par la variable cible.

• Eick *et al.* [46] proposent quatre algorithmes de clustering supervisés, basés sur des exemples représentatifs. Ce genre d'algorithme a pour but de trouver un sous-ensemble de représentants dans l'ensemble d'entrées de telle sorte que le clustering généré en utilisant ce dernier minimise une certaine fonction de pertinence. Dans [46], les auteurs utilisent une nouvelle fonction pour mesurer la qualité de ces algorithmes. Cette fonction remplit les deux critères suivants : *i*) minimisation de l'impureté de classe dans chaque cluster *ii*) minimisation du nombre de clusters ?

SPAM, le premier algorithme proposé par Eick et al, est une variation de l'algorithme de clustering PAM (Partitioning Around Medoids). Le deuxième algorithme proposé par les auteurs dans [46] est SRIDHCR (Single Representative Insertion/Deletion Steepest Decent Hill Climbing with Randomized Restart). Cet algorithme est un algorithme itératif. Il commence par initialiser aléatoirement un certain nombre d'exemples représentatifs. Les clusters sont alors créés en attribuant les exemples au cluster ayant le représentant le plus proche. Par la suite, l'algorithme vise à améliorer la qualité du clustering par l'ajout ou la suppression d'un exemple de l'ensemble des représentants. L'algorithme s'arrête lorsqu'aucune amélioration au niveau de la fonction de pertinence ne peut être réalisée. Le troisième algorithme proposé par ces auteurs est TDS (Top Down Splitting). Cet algorithme suit une approche descendante. Il commence par un seul cluster (*i.e.,* le cluster racine qui contient tous les exemples). Ensuite, il divise d'une manière récursive les clusters (si cette division n'entraîne pas une augmentation de la valeur de la fonction de pertinence) en remplaçant le médoïde du cluster par deux médoïdes : Le premier (respectivement le deuxième) médoïde correspond au medoid de la classe la plus fréquente (respectivement la deuxième classe fréquente) dans le cluster. Le dernier algorithme proposé par Eick et al. dans [46] est l'algorithme SCEC (Supervised Clustering using Evolutionary Computing). Cet algorithme utilise les techniques évolutionnistes pour trouver l'ensemble optimal des représentants .

Au-delà du fait que les algorithmes de clustering supervisé ne permettent pas une découverte complète du concept cible, chacun de ces algorithmes a des points faibles. A titre d'exemple :

— Les quatre algorithmes proposés par Eick *et al.* [46] sont basés sur l'optimisation d'une fonction de pertinence qui nécessite un paramètre de régularisation β. L'algorithme SPAM est une variation de l'algorithme PAM. Cet algorithme est coûteux en temps de calcul. Sa complexité est en $O(K \times (N-K)^2 \times t)$ avec N est le nombre d'instances, K est le nombre de clusters et t est le nombre d'itérations. De plus, le nombre de clusters est un paramètre utilisateur. L'algorithme SRIDHCR est aussi un algorithme coûteux en temps de calcul. Dans chaque itération, pour décider d'ajouter ou de retirer un exemple de l'ensemble des représentants, N partitions doivent être construites et évaluées. De plus, le nombre de clusters et la qualité de la partition générée par cet algorithme dépendent de l'ensemble de représentants choisi au départ. Pour choisir la meilleure partition (celle qui minimise la fonction de pertinence), SRIDHCR est exécuté r fois. Le paramètre r est à définir par l'utilisateur.
— L'algorithme de Cevikalp traite en particulier les classes ayant des chevauchements. Si une classe n'a pas de chevauchement avec les autres classes, celle-ci sera considérée comme un seul cluster bien qu'elle contienne une structure sous-jacente. De plus, cette méthode est très sensible à la présence de bruit.
— L'algorithme S-NN est basé sur l'approche agglomérative ascendante qui est une approche coûteuse en temps de calcul.
— L'algorithme de Al-Harbi *et al.* est un K-moyennes modifié. Généralement, l'algorithme des K-moyennes est caractérisé par sa complexité linéaire. Pour l'optimisation des poids,

l'algorithme de Al-Harbi *et al.* utilise un algorithme génétique. L'utilisation de sa méta-heuristique augmente le coût du modèle. En ce qui concerne le nombre de clusters, l'utilisateur doit le définir *a priori*.

B.4. Clustering prédictif basé sur les arbres (approche de Dzeroski)

Le clustering prédictif basé sur les arbres (ou en anglais predictif clustering trees "PCT") proposé par Dzeroski *et al.* dans [43], peut être présenté comme une généralisation des arbres de décision. Il peut être utilisé pour une variété de tâches d'apprentissage, y compris la prédiction et la description. Le PCT considère un arbre de décision comme une hiérarchie de clusters : la racine d'un PCT correspond à un cluster contenant l'ensemble des données qui est récursivement partitionné en des petits sous-groupes tout en se déplaçant vers le bas de l'arbre. Les feuilles représentent les clusters au niveau le plus bas de la hiérarchie et chaque feuille est étiquetée avec le prototype du cluster correspondant. L'heuristique (h) qui est utilisé pour sélectionner les tests (t) est la réduction de la variance causée par le partitionnement (P) des instances. En maximisant la réduction de la variance, l'homogénéité du cluster est maximisée et la performance prédictive est ainsi améliorée.

La principale différence entre l'algorithme de PCT et d'autres algorithmes d'apprentissage basés sur les arbres de décision est que celui-ci considère la fonction de variance et la fonction prototype (qui calcule une étiquette pour chaque feuille) en tant que paramètres qui peuvent être instanciés pour une tâche d'apprentissage donnée. L'algorithme du clustering prédictif selon Dzeroski *et al.* est présenté dans l'algorithme 1.

Entrée :
- Un ensemble d'exemple E, où chaque exemple prend la forme suivante : $O = (A, Y)$ (A est un vecteur contenant d variables descriptives et Y est une classe cible)
- Un biais de langage B qui permet de décrire les données.
- Une distance *dist* permettant de mesurer la proximité entre deux exemples donnés.
- Une fonction p, dite prototype, permettant d'affecter à chaque exemple une étiquette.

Sortie :
- Chaque cluster est associé avec une description exprimée par le biais de langage B.
- Chaque cluster a une prédiction exprimée par le prototype.
- Inertie intra cluster est minimale (similarité maximale).
- Inertie inter-clusters est maximale (similarité minimale).

Algorithme 1 – Le clustering prédictif selon Dzeroski *et al.*

B.5. Le clusterwise

L'objectif de la régression linéaire typologique (ou clusterwise) est de déterminer une partition d'un ensemble de N instances en K clusters obtenus selon un modèle de régression linéaire reliant une variable y à un ensemble de variables explicatives $\{x_j, j = 1, \ldots, d\}$. On note X la matrice des données associée aux variables explicatives. Cela revient à supposer l'existence d'une variable latente qualitative C à K modalités telle que $E(y|x) = b_0^k + b_1^k x_1 + b_2^k x_2 + \ldots + b_d^k x_d$ où les b_j^k sont les coefficients de la régression de y sur les x_j restreinte aux N_k observations de la classe k.

décrites par y^k, X^k ; avec $N_k \geq d$ pour garantir l'existence d'une solution pour les b^k. La régression typologique (ou clusterwise) revient donc à chercher simultanément une partition en K clusters et et le vecteur b^k des coefficients b_j^k correspondant minimisant le critère $Z = \sum_{k=1}^{K} \parallel X^k b^k - y^k \parallel^2$

Diverses méthodes et algorithmes ont été proposés pour l'estimation de ces coefficients. On peut par exemple citer les travaux de DeSarbo et Cron [42] qui utilisent une méthode du maximum de vraisemblance et l'algorithme EM pour estimer les paramètres du modèle.

La régression linéaire typologique (ou clusterwise) fait l'objet de nombreuses publications, en association avec des données fonctionnelles (Preda et Saporta, [86]), des données symboliques (de Carvalho et al., [39]), dans des cas multiblocs (De Roover et al., [40]).

2.6 Conclusion : notre objectif

2.6.1 Objectif

Dans cette thèse, nous nous sommes fixés comme objectif de développer un algorithme d'apprentissage "interprétable" qui permet de décrire et de prédire d'une manière simultanée. Pour ce faire, nous proposons d'adapter un algorithme de clustering au problème de la classification supervisée. Autrement dit, l'idée est de modifier un algorithme de clustering afin qu'il soit un bon prédicteur tout en gardant sa faculté à bien décrire les données et donc le concept cible à apprendre. Cet algorithme doit également fournir des résultats facilement interprétables par l'utilisateur. Ce type d'algorithme est connu sous le nom de clustering prédictif.

Le modèle d'apprentissage recherché au cours cette thèse est un modèle qui traite principalement trois différents axes, à savoir, la description, l'interprétation et la prédiction (voir la figure 2.20). D'une part, l'utilisation d'un algorithme de clustering donne une garantie immédiate sur l'axe de description. Il s'agit de décrire l'ensemble des données à l'aide de la découverte de la structure sous-jacente existante dans celui-ci. Cependant, le concept de similarité utilisé dans le cadre du clustering traditionnel ne prend pas en considération l'appartenance des instances à des classes différentes. Par conséquent, deux instances similaires de classes différentes peuvent être fusionnées ensemble. Ceci produit une détérioration au niveau de la performance prédictive du modèle en question. D'où la nécessité d'incorporer l'information donnée par la variable cible dans le processus du clustering afin d'assurer l'axe de prédiction et donc la découverte de la structure interne de la variable cible. D'autre part, les algorithmes de partitionnement tels que les K-moyennes fournissent en général des partitions dont chaque groupe est représenté par un prototype. L'utilisation dans ce cas d'un biais de langage est nécessaire pour rendre les résultats issus de cet algorithmes plus interprétables par l'utilisateur.

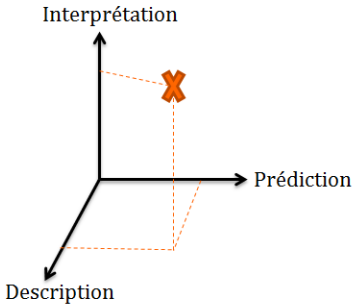

FIGURE 2.20 – Les trois axes traités dans cette thèse

D'une manière générale, le modèle du clustering prédictif recherché au cours de cette thèse doit prendre en considération les points suivants :

1. **La maximisation de la performance prédictive du modèle** : Comme dans le cadre de la classification supervisée, le but majeur des algorithmes de clustering prédictif est de prédire correctement la classe des nouvelles instances.
2. **La découverte de la structure interne de la variable cible** : Dans la phase d'apprentissage, ce modèle doit être capable de découvrir la structure sous-jacente de l'ensemble des données tout en tenant compte de l'étiquetage des instances.
3. **La facilité de l'interprétation des résultats** : Dans les domaines critiques (e.g., le service marketing à Orange), l'interprétation des résultats générés par un système d'apprentissage est une question incontournable. Pour cette raison, on souhaite avoir un modèle interprétable même par des utilisateurs non experts.
4. **La minimisation des connaissances,** *a priori* **requises, de la part de l'utilisateur (***i.e.,* **pas ou peu de paramètres utilisateur)** : Le modèle recherché doit être un algorithme qui contient très peu ou aucun paramètres utilisateur.
5. **La minimisation de la taille (complexité) du modèle prédictif - descriptif** : Le modèle recherché doit être efficace, rapide en termes de calcul et facile à implémenter.

En tenant compte de tous les points cités ci-dessus et en adoptant la définition donnée par Dzeroski *et al.* dans [43], le clustering prédictif peut être présenté par l'algorithme 2.

Entrée :

- Un ensemble de données $\mathcal{D}$, où chaque instance X_i est décrite par un vecteur de d dimensions et par une classe $Y_i \in \{1, \ldots, J\}$.
- Un ensemble de prototypes initiaux (ou centres initiaux) qui forment la partition initiale.
- Une distance *dist* qui mesure la proximité entre deux instances.
- Un biais de langage B permettant de décrire les données.

Sortie :

- Chaque cluster est représenté par un prototype qui possède la même étiquette de classe.
- Chaque cluster est associé avec une description donnée par le biais de langage B.

Si (*l'algorithme est dédié à la première catégorie*) **Alors**
- Le nombre de clusters (K) est minimal.
- Le taux de bonnes classifications est maximal.

Fin Si

Si (*l'algorithme est dédié à la deuxième catégorie*) **Alors**
- 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.

Fin Si

Algorithme 2 – Algorithme générique du clustering prédictif

Les approches issues de la littérature qui semblent adéquates pour résoudre notre problématique sont les approches cités dans la section 2.5.2. Cependant, ces algorithmes ne prennent pas en considération tous les points cités ci-dessus. Dans cette thèse, nous avons choisi de modifier l'algorithme de clustering le plus répandu dans la littérature, à savoir l'algorithme des K-moyennes. La version modifiée de celui-ci sera nommée *les K-moyennes prédictives*.

2.6.2 K-moyennes prédictives

La méthode des centres mobiles (ou les K-moyennes) due à Forgy [49] permet de subdiviser l'ensemble des individus en un certain nombre de classes en employant une stratégie d'optimisation itérative dont le principe général est de générer une partition initiale, puis de chercher à l'améliorer en réattribuant les données d'une classe à l'autre. Cet algorithme recherche des maxima locaux en optimisant une fonction objectif traduisant le fait que les individus doivent être similaires au sein d'une même classe, et dissimilaires d'une classe à une autre. Les classes de la partition finale, prises deux à deux, sont d'intersection vide et chacune est représentée par un noyau (ou prototype).

L'algorithme des *K-moyennes prédictives* consiste à prédire la classe d'une nouvelle instance en se basant sur sa proximité à un des groupes formés dans la phase d'apprentissage. Plus précisément, dans le problème du clustering prédictif, les données d'apprentissage $\mathcal{D} = \{X_i\}_{i=1}^{N}$ sont composées de N instances. Chaque instance i est décrite par d variables descriptives $\{X_i^1, \ldots, X_i^d\}$ et une variable cible Y_i contenant l'information de la classe. On notera $Y_i \in \{1, \ldots, J\}$ où J est le nombre de classes. L'objectif de l'algorithme des K-moyennes prédictives est donc de former, à partir des données d'apprentissage, K clusters purs et homogènes : les instances appartenant à un cluster $k \in \{1, \ldots, K\}$ doivent, d'une part, avoir la même classe j et d'autre part être différentes des instances appartenant à d'autres clusters. A la fin du processus d'apprentissage, une technique est utilisée pour étiqueter chaque cluster appris (*e.g.,* l'utilisation du vote majoritaire). Au final, la prédiction d'une nouvelle instance se fait selon son appartenance à un des clusters appris. Autrement dit, cette nouvelle instance reçoit j comme prédiction si elle est plus proche du centre de gravité du cluster de classe j (*i.e.,* utilisation du 1 plus proche voisin). Les différentes étapes de l'algorithme des K-moyennes prédictives sont présentées dans l'algorithme 3.

Étape 1 : Prétraitement des données
Étape 2 : Initialisation des centres
Étape 3 : Répéter un certain nombre de fois (R) jusqu'à convergence
  3.1 Cœur de l'algorithme
Étape 4 : Choix de la meilleure convergence
Étape 5 : Mesure d'importance des variables (après la convergence et sans réapprendre le modèle)
Étape 6 : Affectation des classes aux clusters appris.
Étape 7 : Prédiction de la classe des nouveaux exemples.

Algorithme 3 – Les étapes des K-moyennes prédictives.

Pour aboutir à notre objectif, chaque étape de l'algorithme des K-moyennes (Algorithme 3) pourrait être traitée individuellement. L'idée est de tester à quel point la supervision de chaque

étape pourrait aider l'algorithme des K-moyennes standard à remplir la tâche du clustering prédictif. Au final, on pourrait obtenir un algorithme des K-moyennes supervisé à chaque étape. Cet algorithme sera comparé par la suite aux approches potentielles décrites dans la section 2.5.2 B. Dans cette thèse on ne s'intéresse qu'à la modification de quatre étapes de l'algorithme des K-moyennes standard. Ces étapes sont présentées dans ce qui suit :

L'étape 1 des K-moyennes prédictives (voir Algorithme 3) fait l'objet du **chapitre 3** de ce mémoire. L'étape du prétraitement des données est une étape primordiale que ce soit dans le cadre du clustering classique ou dans le cadre du clustering prédictif. L'intérêt de cette étape est d'incorporer l'information donnée par la variable cible dans les données dans le but de permettre à l'algorithme des K-moyennes standard de la prendre en considération. En effet, lorsqu'on dispose d'un ensemble de données où les instances de différentes classes sont proches les unes des autres en termes de distance, l'application de l'algorithme classique des K-moyennes sur ces données va entrainer une détérioration au niveau de la performance prédictive (ces instances proches vont être fusionnées ensemble indifféremment à leur classe d'appartenance). Le but de ce chapitre est donc de définir une distance dépendante de la classe cible qui vérifie que deux instances proches en termes de distances sont également proches en termes de leur comportement vis-à-vis de la variable cible. Cette distance peut être écrite à l'aide d'un prétraitement supervisé des données. Pour respecter les points imposés dans la section 2.6.1, le prétraitement proposé doit impérativement être interprétable (point 3.), robuste (point 1.), rapide (point 5.), et sans paramètres utilisateur (point 4.). Avec ce genre de prétraitement de données on espère augmenter la performance prédictive de l'algorithme classique des K-moyennes comparant à sa performance prédictive en utilisant des prétraitements non supervisés.

L'étape 2 des K-moyennes prédictives (voir Algorithme 3) fait l'objet du **chapitre 4** de ce mémoire. L'un des inconvénients de l'algorithme des K-moyennes standard réside dans sa sensibilité envers le choix des centres initiaux. En effet, l'étape d'initialisation influence la qualité de la solution trouvée ainsi que le temps d'exécution [29]. Lors de déséquilibre des classes à prédire (par exemple, l'existence d'une classe majoritaire et d'une classe minoritaire) dans l'ensemble des données, l'utilisation d'une méthode d'initialisation non supervisé s'avère insuffisante. En effet, la probabilité de choisir plus d'un centre dans la classe majoritaire et de ne choisir aucun centre dans la classe minoritaire est très grande. Par conséquent, une détérioration au niveau de la performance prédictive du modèle peut être produite. A partir de ce constat, il est naturel de se demander si l'utilisation d'une méthode d'initialisation supervisée pourrait aider l'algorithme des K-moyennes standard à remplir la tâche du clustering prédictif. Plus précisément, cette étape traite principalement le premier point cité dans la section 2.6.1 qui est la maximisation de la performance prédictive du modèle.

L'étape 4 des K-moyennes prédictives (voir Algorithme 3) fait l'objet du **Chapitre 5** de ce mémoire. L'algorithme des K-moyennes n'assure pas de trouver un minimum global. Il est souvent exécuté plusieurs fois (on parle de "réplicates") et la meilleure solution en termes d'erreur quadratique moyennes est alors choisie. Dans le cadre du clustering prédictif, la notion de "meilleure solution" diffère de celle connue dans le cadre du clustering standard. Pour la première catégorie du clustering prédictif où on l'impose d'avoir un nombre faible de clusters (voir la figure 2.14 de la section 2.5.2), l'axe privilégié dans ce cas est l'axe de prédiction. Un critère supervisé tel que l'indice de rand ajusté peut être utilisé pour mesurer la qualité des résultats et donc choisir la meilleure partition. Pour la deuxième catégorie du clustering prédictif où l'on cherche à découvrir la structure complète de la variable cible, l'utilisation des critères proposés

dans la littérature s'avère insuffisant. En effet, dans ce cadre d'étude, on cherche à réaliser le bon compromis entre la prédiction et la description et à notre connaissance, il n'existe pas dans la littérature un critère analytique permettant d'évaluer ce compromis. Le but de ce chapitre est donc de proposer un critère analytique permettant de mesurer la qualité des résultats générés par la deuxième catégorie du clustering prédictif. Cette étape cherche à traiter les deux premiers points cités dans la section 2.6.1, à savoir, la maximisation de la performance prédictive du modèle et la découverte de la structure interne de la variable cible.

L'étape 5 des K-moyennes prédictives (voir Algorithme 3) fait l'objet de l'annexe E de ce mémoire. Pour une interprétation aisée des résultats générés par l'algorithme des K-moyennes prédictives, on cherchera dans cette partie de la thèse à proposer une méthode supervisée permettant de mesurer l'importance des variables selon leurs contributions dans le processus d'apprentissage. Partant d'une partition de référence (partition à interpréter), l'importance de chaque variable sera définie alors comme le pouvoir prédictif de celle-ci à bien prédire cette partition de référence. Ce pouvoir prédictif est mesuré à l'aide de un arbre de décision. Cette étape traite principalement le troisième point cité dans la section 2.6.1 à savoir, la facilité de l'interprétation des résultats. Puisque le travail effectué au cours de cette thèse sur ce sujet n'est pas encore achevé, nous avons choisi donc de le placer dans un annexe au lieu de le considéré comme un chapitre.

Le chapitre 6 présente d'une part une synthèse des résultats obtenus dans les chapitres précédents et d'autre part, il présente une comparaison de l'algorithme des K-moyennes prédictives proposé dans cette thèse (intégrant les différentes étapes supervisées) avec d'autres méthodes issus de la littérature. Cette partie expérimentale est divisée en deux grandes parties. La première partie se focalise sur le côté prédictif du modèle. Tandis que la deuxième partie se focalise sur l'aspect interprétable du modèle et la capacité de celui-ci à bien découvrir la structure interne de la variable cible. Pour finir, une conclusion dresse le bilan des trois années de thèse, des travaux réalisés et des travaux futurs. Nous rappelons les différentes notions introduites dans la thèse, ainsi que les résultats obtenus.