Chapitre 1 Formalisations et définitions

Afin d'être cohérent tout au long de ce manuscrit, nous allons, dans un premier temps, exposer la formalisation que nous utiliserons et introduire un ensemble de notations et de notions. Ce chapitre constitue une base de théorie qui nous sera nécessaire dans les chapitres suivants.

1 Formalisation

Soient Π une population étudiée et Ω un sous-ensemble de n objets observés : $\Omega = \{\omega_1,...,\omega_i,...,\omega_n\}$, chaque élément de Π est caractérisé par un ensemble $\Gamma = \{X_1,...,X_k,...,X_p,Y\}$ de p variables exogènes ou explicatives et d'une variable à expliquer ou endogène Y.

Soit X l'ensemble des variables exogènes. $X_k(\omega_i) \in O_k$ est la valeur prise par ω_i pour la variable X_k. Si X_k est une variable qualitative, O_k est un ensemble fini de valeurs observables pour X_k appelé espace des modalités de X_k : $O_k = \{m_1^k,...,m_v^k,...,m_{q^k}^k\}$.

m_v^k est la v$^{\text{ième}}$ modalité de la variable X_k et q^k est le nombre de modalités de la variable X_k.

$O = O_1 \times ... \times O_k \times ... \times O_p$ représente l'espace de travail et l'application T décrit un objet ω_i de X. T est de la forme suivante :

$$T : \Pi \to O$$
$$\omega_i \mapsto T(\omega_i) = (X_1(\omega_i),...,X_k(\omega_i),...,X_p(\omega_i)).$$

Ce formalisme sera complété au cours des chapitres qui suivent en fonction des besoins.

2 Les données brutes

Les données brutes sont les données acquises lors de la phase de compréhension du domaine. Ces données n'ont connu aucun traitement, elles sont telles que l'utilisateur les a trouvées. Aussi elles ne sont pas forcément constituées des variables les plus aptes à décrire le problème étudié.

Les variables appartenant aux données brutes peuvent être classées en fonction de leur nature. Elles peuvent être :

- Redondantes : C'est à dire qu'il existe deux variables identiques. On parle également de redondance si plusieurs variables prises ensembles jouent le même rôle.
- Corrélées : Des variables sont considérées comme corrélées si leur combinaison est capable de déterminer les classes induites par la variable endogène.
- Bruitées : Ce sont des variables qui ne permettent pas de distinguer des individus appartenant à deux classes différentes.
- Pertinentes : Ce sont des variables qui sont utiles pour la discrimination de l'ensemble des individus. Ce sont ces variables que nous voulons lors de la phase de sélection de variables sélectionnées.

La notion de pertinence est primordiale au sein de la phase de prétraitement des données. En effet, le but même de cette phase de l'ECD est de ne conserver que les variables pertinentes. Aussi il nous semble important d'en préciser la définition.

3 Pertinences

Les processus de sélection et de construction de variables doivent permettre de sélectionner et respectivement de construire des variables pertinentes. Cependant, il existe de nombreuses définitions de la pertinence d'une variables. Nous allons tout d'abord présenter ces différentes définitions puis expliciter celle que nous emploierons.

3.1 La pertinence selon Blum et Langley

Blum et Langley, [5], ont classifié les définitions de la pertinence d'une variable en fonction de la question suivante : nous désirons qu'une variable soit pertinente par rapport à quoi ?

3.1.1 Pertinence par rapport à une variable endogène

Une variable X_k est pertinente par rapport à Y si et seulement si il existe un couple $(\omega_i, \omega_j) \in \Pi$ tel que ω_i et ω_j ne différent que par la valeur de la variable X et $Y(\omega_i) \neq Y(\omega_j)$. La variable est dite non pertinente sinon. Cette définition rencontre des problèmes en présence de variables redondantes qui sont considérées comme non pertinentes quelle que soit la situation. Et, le calcul s'effectue sur l'échantillon donc la variable n'est pas forcément pertinente sur l'ensemble de toutes les données. Elle rencontre également des problèmes en présence d'un grand nombre de variables : il devient alors improbable que deux individus ne diffèrent que par la valeur d'une seule variable. Cependant, cette définition est toujours utilisée pour l'analyse théorique des algorithmes d'apprentissage, quand la notion de pertinence est utilisée pour prouver la convergence d'un algorithme.

Pour palier aux désavantages liés aux variables redondantes, John, Kohavi et Pfleger [6] ont défini deux notions de pertinences par rapport à l'échantillon.

3.1.2 Pertinence forte par rapport à l'échantillon

X est une variable fortement pertinente si et seulement si :

$$\exists(\omega_i, \omega_j) \in \Omega^2 / Y(\omega_i) \neq Y(\omega_j) \text{ et } X_t(\omega_i) = X_t(\omega_j); t = 1, \ldots, p; t \neq k \text{ et } X_k(\omega) \neq X_k(\omega)$$

Cette définition rajoute la contrainte suivante : les deux objets ω_i et ω_j doivent appartenir à l'échantillon d'apprentissage des objets observés.

3.1.3 Pertinence faible par rapport à l'échantillon

Une variable X est faiblement pertinente par rapport à l'échantillon Ω si et seulement si il est possible de retirer un sous-ensemble de variables afin de rendre X fortement pertinente.

Ces deux notions de pertinence sont utiles dans la décision de garder ou d'ignorer une variable. Une variable fortement pertinente sera gardée, tandis que le sort d'une variable faiblement pertinente dépendra des autres variables ignorées.

3.1.4 Pertinence par rapport à un algorithme d'apprentissage

Etant donnés Ω un échantillon de données, L un algorithme d'apprentissage et un ensemble de variables X, la variable X est utile de manière incrémentale par rapport à L si l'exactitude de l'hypothèses, selon laquelle L est performant à partir de l'ensemble de variables $\{X\} \cup \mathrm{X}$, est meilleure que l'exactitude de l'hypothèse selon laquelle L est performant à partir de l'ensemble de variables. Cette méthode est particulièrement intéressante pour les algorithmes de sélection de variables qui ajoutent ou enlèvent de manière incrémentale des variables à leur ensemble courant.

3.1.5 Pertinence par rapport à une classe

Une variable X est pertinente pour une classe si X apparaît dans chaque formule booléenne représentant la classe et non pertinente sinon, [7].

X est pertinente si et seulement si il existe une modalité m_v^k de X et une modalité y_u de Y telle que $P(X_k = m_v^k) > 0 \quad et \quad P(Y = y_u / X_k = m_v^k) \neq P(Y = y_u)$. Il existe un problème quand les classes sont équiprobables : chaque variable est considérée comme non-pertinente car les probabilités sont égales.

3.2 Pertinence selon Kohavi

En 1997, Kohavi, [8], définit les notions de pertinence faible et de pertinence forte et déclare qu'une variable est pertinente si elle l'est soit fortement soit faiblement et non pertinente sinon.

3.2.1 Pertinence forte

Une variable X est fortement pertinente si la suppression de X détériore la qualité de l'apprentissage.

3.2.2 Pertinence faible

Une variable X est faiblement pertinente si elle n'est pas fortement pertinente et s'il existe un sous-ensemble X' de variables tel que la performance du classifieur obtenue avec X' est moins bonne que celle du classifieur obtenue avec $\mathrm{X}' \cup \{X_k\}$.

La multitude des définitions de pertinence entraîne une notion vague de ce terme. Nous décidons d'adopter comme définition celle de Kohavi, [8]. Cette définition nous paraît relativement générale et appropriée au cadre de la sélection et de la construction de variables.