Chapitre 3 Construction de variables

La qualité d'un apprentissage est entre autres choses liée à la présence de variables discriminantes. Or, dans le cas d'une qualité d'apprentissage médiocre et en l'absence de nouvelles informations disponibles, il est nécessaire de trouver un moyen qui, à partir de l'information disponible, nous permettra de re-décrire le données d'entrée du problème d'apprentissage considéré et éventuellement d'obtenir de nouvelles variables discriminantes. Les méthodes de construction de variables résolvent ce problème. En effet, comme nous l'avons précisé précédemment, la construction de variables permet, lors de la phase de prétraitement des données, la création de nouvelles variables synthétiques. Ces variables synthétiques sont issues de la découverte de relations entre variables initiales. Les méthodes de construction de variables entraînent une augmentation de l'espace de représentation des données, dans la mesure où de nouvelles variables sont construites. Cependant, aucune information extérieure aux données initiales n'est ajoutée lors du processus de construction.

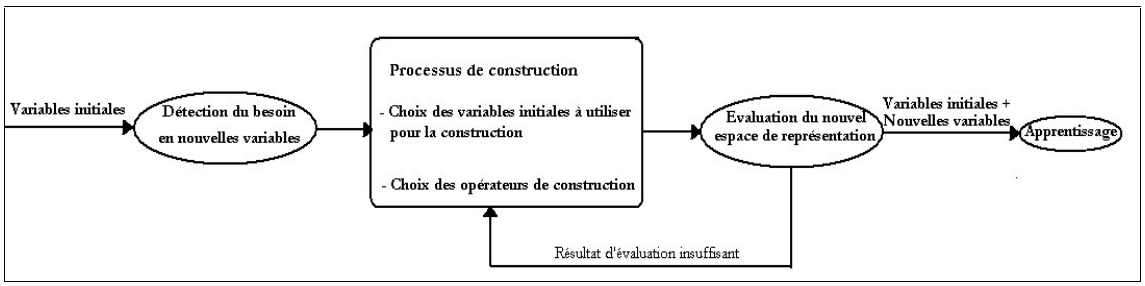

Figure 19 Processus de construction de variables.

Avant le lancement du processus de construction de variables à proprement dit, figure 19, nous devons savoir s'il est nécessaire de construire de nouvelles variables. Donc, tout processus de construction devrait être précédé d'une phase de détection des besoins en nouvelles variables. Il est utile de répondre à la question : les variables initiales sont elles suffisantes pour discriminer au mieux la variables endogènes ? Cette phase de détection est effectuée sans apport de nouvelles informations ; on peut ainsi parler de détection a priori. Si de nouvelles variables sont nécessaires, alors le processus de construction de variables commence. Selon la méthode de construction choisie, les variables initiales utilisées et les opérateurs de construction qui leurs seront appliqués ne sont pas les mêmes : ils

peuvent être imposés ou laissés au choix de l'utilisateur. Le nombre de variables construites dépend également de la méthode de construction choisie.

Le nouvel espace de représentation est composé des variables exogènes initiales et des variables exogènes synthétiques. Ce nouvel espace est alors évalué. L'évaluation est représentée par l'analyse des résultats d'apprentissage obtenus sur le nouvel espace de représentation. En se basant sur cette évaluation, le système décide d'arrêter le processus de construction si l'espace de représentation est satisfaisant, dans ce cas on passe à la phase d'apprentissage, ou de continuer à chercher un meilleur espace de représentation sinon, dans ce cas le processus de construction est réitéré.

La construction de variables doit permettre l'accroissement de la qualité prédictive. Cet accroissement doit être effectif non seulement sur l'ensemble d'apprentissage mais aussi et en priorité sur l'ensemble total des données. Pour ce faire, la construction de variables doit également permettre la simplification du modèle d'apprentissage. En effet, si la complexité du modèle d'apprentissage n'est pas prise en compte, l'amélioration de la qualité prédictive sur les données d'apprentissage risque d'entraîner un sur-ajustement (phénomène d'over-fitting) sur l'ensemble d'apprentissage et ainsi une réduction de la qualité prédictive du modèle sur l'ensemble total des données.

La première section est consacrée à décrire l'ensemble des méthodes de construction existantes. Ensuite, nous présentons la méthode de construction que nous proposons. Enfin, nous présentons l'ensemble des expérimentations que nous avons réalisé.

1 Méthodes de construction de variables

1.1 Définitions

La construction de variables est la construction de nouvelles variables synthétiques dans le but de re-décrire les données d'entrée d'un problème d'apprentissage. Cependant, elle peut être caractérisée par de nombreuses définitions :

- Murphy l'a décrite comme toute forme d'induction générant de nouvelles descriptions non présentées dans les données d'entrées initiales, [87] ;
- Mitchell l'a définie comme le procédé d'accroissement de l'espace de représentation, basé sur les connaissances du domaine, [88] ;

- Rendell l'énonce comme la création de concepts utiles et n'existant pas dans la description initiale des données, [89] ;
- Matheus considère qu'elle est l'application d'un ensemble d'opérateurs sur les variables existantes qui peuvent être des variables initiales ou des variables construites [90] ;
- Michalski l'oppose à l'apprentissage empirique sur le point de vue de l'espace de représentation : « Pour l'apprentissage inductif empirique, l'espace de description est le même pour les individus et pour les hypothèses, alors que pour la construction de variables ces espaces sont différents », [91].

1.2 Classification des méthodes de construction

Il existe différentes taxonomies des méthodes de construction de variables. Nous nous intérressons à deux de ces taxonomies : celle de Bloedorn, [92] et celle de Fawcett, [93].

1.2.1 Taxonomie de Bloedorn

Bloedorn, [92], détermine quatre catégories au sein des méthodes de construction, tableau 40 :

- L'induction constructive dirigée par les hypothèses (Hypothesis Driven Constructive Induction HCI) : L'extraction de variables utiles dans les hypothèses est utilisée ici comme base pour la construction de variables ;
- L'induction constructive dirigée par les données (Data Driven Constructive Induction DCI) : Les résultats obtenus par l'application des méthodes d'analyse des données sont utilisés pour construire des variables ;
- L'induction constructive dirigée par la connaissance (Knowledge Driven Constructive Induction KCI) : La connaissance fournie par un expert du domaine est utilisée pour transformer l'espace de représentation ;
- L'induction constructive Multi stratégique (Multistrategy Driven Constructive Induction MCI) : Elle combine tous les types d'approches citées préalablement pour modifier l'espace de représentation.

1.2.2 Taxonomie de Fawcett

De manière générale, la modification de l'espace de représentation vise l'obtention d'une qualité prédictive meilleure. Ce but peut alors être atteint en créant de nouvelles variables dans une optique particulière.

Ainsi, selon [93], les travaux existant en construction de variables peuvent être divisés en trois catégories selon le but pour lequel sont créées les nouvelles variables, tableau 41 :

- Changer le biais de recherche : Dans ce cas les variables créées utilisent la même représentation que les variables exogènes ; ainsi, l'utilisation de la construction de variables ne permet pas réellement d'étendre l'espace de représentation pour l'apprentissage, mais peut changer l'ordre dans lequel sont considérer les classes de la variable endogène. Cela constitue l'idée de base des algorithmes CITRE [90] et FRINGE [94], [95] ;
- Accroître la concision : certains systèmes construisent de nouvelles variables non pas dans l'optique d'améliorer directement la qualité de la prédiction du point de vue du taux de réussite en apprentissage mais dans l'optique d'accroître la concision de la théorie du domaine développée par l'algorithme d'apprentissage. La concision peut alors être vue comme un but en soi ou encore un moyen d'améliorer l'efficacité de l'algorithme d'apprentissage. L'algorithmes RINCON [96] a été créé dans ce but ;
- Supporter les représentations complémentaires : Les variables construites utilisent une représentation différente de celle de la variable endogène ; ainsi, l'espace recherché par construction de variables est fondamentalement différent de celui recherché par apprentissage. On peut citer par exemple les systèmes STABB [97], [98], STAGGER [99], [100], MIRO [101].

Ces deux taxonomies nous permettent une vision sous des angles différents des méthodes de construction. La taxonomie de Bloedorn considère essentiellement les méthodes de construction utilisant un processus d'induction et les classe en fonction de leur mode de fonctionnement. La taxonomie de Fawcett, quant à elle, s'intéresse essentiellement au but principal de la méthode de construction.

Méthode	Dirigée par les hypothèses	Dirigée par les données	Dirigée par la connaissance	Multi-stratégique
	AQ17-HCI, [102] FRINGE, [95] LFC,	AQ17-DCI, [103] BACON, [105] PLS0, [108] STAGGER, [99], [100] FCE, [110]	CITRE, [90] AQ15, [106] MIRO, [101]	AQ-BC, [104] AQ17, [107] INDUCE-1, [109]

Tableau 40 Taxonomie de Bloedern, [92]

Buts :	Modifier le biais de recherche	Accroître la concision	Supporter les représentations complémentaires
	CITRE, [90] FRINGE, [94] LFC	RINCON, [96]	STAGGER, [100], [99] MIRO, [101]

Tableau 41 Taxonomie de Fawcett, [93]

Nous proposons de classifier les méthodes en quatre catégories, inspirées fortement de la taxonomie de Bloedorn, de la manière suivante : les méthodes de construction par analyse topologique des arbres, les méthodes de construction par analyse et exploration des données, les méthodes de construction basée sur l'utilisation des connaissances du domaine ou d'un expert et les méthodes multi-stratégiques. Le tableau 42 présente cette classification et les différentes méthodes de construction de variables que nous avons recensées. Les méthodes écrites en caractères gras correspondent à des sous-catégories de méthodes. Les méthodes les plus nombreuses sont celles appartenant à la catégorie des méthodes par analyse et exploration des données.

Cependant, cette classification des méthodes n'est qu'une indication sur la manière de trier et de présenter ces méthodes. En aucun cas, elle ne tient lieu de référence en terme de taxonomie des méthodes de construction de variables.

Méthodes par :	Analyse topologique des arbres	Analyse et exploration des données	Utilisation des connaissances du domaine ou d'un expert	Méthodes multi-stratégiques
	AQ17-HCI, [102]	BACON, [105]	AQ15, [106]	AQ-BC, [104]
	STRUCT	PLS0, [108]	IB3-CI, [111]	INDUCE-1, [109]
	PRAX	STAGGER, [99], [100]	RINCON, [96]	AQ17, [107]
	FRINGE, [94]	FCE, [110]	MIRO, [101]	
	LFC	AQ17-DCI, [103]		
	GALA, [112]	Méthode de Zheng, [113]		
	CITRE, [90]	**Méthodes utilisant la programmation génétique, [51, 52, 114-116]**		
		Méthodes d'analyse des données, [117, 118]		
		Construction par décomposition de fonctions, [119]		
		Méthode utilisant les treillis, [120]		

Tableau 42 Classification des méthodes de construction de variables.

1.3 Construction par analyse et exploration des données

Ces méthodes analysent et explorent les données, les relations existantes entre variables exogènes, individus, et variable endogène afin de déterminer de nouvelles variables. Il existe de nombreuses méthodes de ce type. Nous n'étudierons en détail qu'un petit nombre de ces méthodes. Les autres méthodes seront citées et brièvement présentées.

1.3.1 BACON

La méthode BACON, [105], base sa procédure de construction sur les interdépendances existantes entre des variables numériques.

1.3.2 PLS0

La méthode PLS0, [108], crée de nouvelles variables à partir des variables initiales en utilisant une forme conceptuelle de clustering.

1.3.3 STAGGER

L'algorithme STAGGER, [99] et [100], génère de nouvelles variables à l'aide de combinaisons booléennes de variables numériques. STAGGER associe l'apprentissage par pondération et l'apprentissage de fonctions booléennes. Les travaux qui ont suivi, [99] y ont ajouté le partitionnement de variables numériques. Les nouvelles variables sont formées en appliquant les opérateurs booléens ET, OU et NON aux variables existantes selon un procédé composé de plusieurs heuristiques. Ainsi, STAGGER peut apprendre des combinaisons linéaires de variables.

1.3.4 FCE

L'algorithme FCE, [110], prend comme point de départ un ensemble d'espaces de représentation. Après avoir détecté les parties inconsistantes de chaque espace, un nouvel espace de représentation est construit. Cet espace est un produit des espaces déjà existants : sa taille est donc supérieure à celle des espaces initiaux.

1.3.5 AQ17-DCI, [103]

Ce système est basé sur l'algorithme AQ classique [121], mais il inclut également un algorithme d'induction qui génère de nouvelles variables. La qualité de chaque nouvelle variable générée est évaluée selon une fonction de qualité. Si cette fonction est supérieure à un certain seuil alors la variable est sélectionnée. Cet algorithme fonctionne en deux parties. La première partie correspond au traitement des variables numériques et la seconde partie au traitement des variables binaires. Ces deux parties sont caractérisées par une fonction de qualité. Ainsi, la qualité d'une variable est mesurée conjointement par deux critères :

- Le gain d'information qu'elle apporte, ceci est mesuré à l'aide du Khi2,
- Le coût, qui ne doit pas dépasser une certaine valeur fixée par l'utilisateur. Il est égal à la somme des coûts des variables utilisées plus le coût de chaque opérateur utilisé.

1.3.5.1 Le traitement des variables numériques

Détection de l'ensemble des variables numériques
Répéter pour toutes les combinaisons possibles de variables
Répéter pour tous les opérateurs de construction de variables (tableau 43)
Calcul des valeurs de la nouvelle variable
Fin Répéter
Evaluation du pouvoir discriminant de cette nouvel variable grâce à la fonction de qualité
Si la fonction de qualité est supérieure au seuil fixé Alors
Conservation de la variable
Sinon
    Suppression de la variable
    Fin Si
Fin Répéter
Si l'ensemble des règles obtenues est satisfaisant du point de vue du critère de précision et de la complexité des règles fixés par l'utilisateur Alors
Arrêt du processus
Sinon
Recherche d'un meilleur espace de représentation
Fin Si

Algorithme 15 AQ17-DCI – variables numériques.

Opérateurs	**Arguments**	**Notation**	**Interprétation**
Equivalence	Variables x, y	x=y	Si x=y alors 1, sinon 0
Supérieur à	Variables x, y	x>y	Si x>y alors 1, sinon 0
Supérieur ou égal	Variables x, y	x>=y	Si x>=y alors 1, sinon 0
Addition	Variables x, y	x+y	Somme de x et y
Soustraction	Variables x, y	x-y	Différence entre x et y
Différence	Variables x, y	|x-y|	Différence absolue entre x et y
Multiplication	Variables x, y	X*y	Produit de x et y
Division	Variables x, y	x/y	Quotient de x divisé par y
Maximum	Ensemble de variables S	Max(S)	Valeur maximale de l'ensemble S
Minimum	Ensemble de variables S	Min(S)	Valeur minimale de l'ensemble S
Moyenne	Ensemble de variables S	Moy(S)	Moyenne des valeurs de l'ensemble S
Compte	Ensemble de variables	{Attr(S,C)	Nombre de variables de S satisfaisant C

Tableau 43 Opérateurs numériques de construction de variables

L'algorithme permettant de traiter les variables numériques possède une liste par défaut d'opérateurs (tableau 43) mais autorise l'utilisateur à modifier la liste pour l'ajuster au problème traité.

1.3.5.2 Le traitement des variables binaires

Recherche d'un meilleur espace de représentation

Identification des variables binaires

Recherche des symétries ou des symétries rapprochées entre deux variables, en se basant sur les travaux décrits dans [Michalski, 69]

Pour chaque candidat aux symétries, création d'un nouvel variable : cet variable est la somme arithmétique des variables du groupe

Détermination la valeur de la fonction de qualité de chacun des variables nouvellement créés et sélection du meilleur variable

Amélioration de la base de données avec ce nouvel variable

Apprentissage

Si l'ensemble de règles obtenues est satisfaisant selon les critères de précision et de complexité des règles fixés par l'utilisateur Alors

Arrêt du processus

Sinon

Recherche d'un meilleur espace de représentation

Fin Si

Algorithme 16 AQ17-DCI - variables binaires.

1.3.5.3 Contraintes supplémentaires

Le nombre de combinaisons possibles, que ce soit pour les variables numériques ou pour les variables binaires, est trop important pour pouvoir être toutes traiter. Aussi, il existe des contraintes supplémentaires permettant à la fois de réduire le nombre de combinaisons à tester et de créer des variables significatives et facilement interprétables. Ces contraintes sont de deux ordres :

- Contrainte d'unité : un opérateur ne peut être appliqué qu'à deux variables ayant la même unité ;
- Contrainte de seuil : un nombre maximal de variables générées est fixé par l'utilisateur.

Cette méthode permet d'exprimer des fonctions booléennes symétriques ou partiellement symétriques, tout comme des fonctions plus complexes qui dépendent de la présence d'un certain nombre de variables.

1.3.6 Méthode de Zheng

Zheng [113], construit des variables du type M-of-N et du type X-of-N. Les variables initiales sont des variables continues qui sont transformées en variables binaires ou nominales par discrétisation. Nous reprenons ici la formalisation utilisée par Zheng.

Soit $\{A_i | 1 \le i \le MaxAttr\}$ l'ensemble des variables du domaine, MaxAttr est le nombre de variables. Et, pour chaque A_i, $\{V_{ij} | 1 \le j \le MaxAttrVal\}$ est son sous-ensemble de valeurs. MaxAttrVal est le nombre de valeurs différentes de A_i. N_+ est le nombre de paires attributs/valeurs dans la représentation c'est à dire la taille de la représentation. N est le nombre de variables différentes de la représentation. AV_k est vrai si la variable A_i de l'exemple k a la valeur V_{ij}.

Les quatre types de nouvelles variables sont les suivants :

- X-of-N est composé d'un ensemble de paires attributs/valeurs défini comme suit :

$$X-of-\left\{ AV_k \middle| \begin{matrix} AV_k \quad est \quad une \quad paire \quad attribut-valeur \quad telle \quad que \quad A_i = V_{ij} \\ 1 \le k \le N_+, N \le N_+, \quad 1 \le N \le MaxAttr \end{matrix} \right\}.$$

  La valeur d'une représentation X-of-N peut être un nombre entre 0 et N. Elle est égale à X si et seulement si X exemples de AV_k sont vrais ;

- M-of-N est composé d'un ensemble de paires attributs/valeurs et d'un nombre M définis comme suit :

$$M-of-\left\{ AV_k \middle| \begin{matrix} AV_k \quad est \quad une \quad paire \quad attribut-valeur \quad telle \quad que \quad A_i = V_{ij} \\ 1 \le k \le N_+, N \le N_+, \quad 1 \le N \le MaxAttr, \quad 1 \le M \le N \end{matrix} \right\}$$

  Sa valeur est soit vrai soit faux. Sa valeur sera vraie si et seulement si au moins M exemples de AV_k sont vrais ;

- La conjonction est vraie si toutes les paires de la conjonction de paires d'attributs-valeurs sont vraies ;
- La disjonction est vraie si au moins une des paires de la disjonction de paires d'attributs-valeurs est vraie.

L'auteur suppose que la représentation X-of-N s'effectue sur des variables nominales. Si l'ensemble d'apprentissage est petit et si le concept cible est complexe, dans les arbres de décision, cette représentation entraînera un grand nombre de petits sous-ensembles pour l'ensemble d'apprentissage.

1.3.7 Construction de variables par décomposition de fonctions

Zupan, Bohanec, Demsar et Bratko [119], créent une méthode de décomposition de fonction qu'il est possible d'appliquer à la construction de variables. Cette méthode se décompose en trois partie : la décomposition de la fonction étudiée, découverte de la meilleur partition de l'ensemble des variables et construction des nouvelles variables.

1.3.7.1 Décomposition d'une fonction

Soit $Y = F(\mathrm{X})$ avec $\mathrm{X} = X_1,\ldots,X_p$ les variables initiales. Le but est de décomposer $Y = F(\mathrm{X})$ en $Y = G\big(A, H(B)\big)$ avec A et B des sous-ensembles de variables de X tels que $A \cup B = \mathrm{X}$. A est appelé l'ensemble libre, et, B l'ensemble lié. G et H sont des fonctions tabulées non prédéfinies mais découvertes lors de la décomposition. Lors de cette décomposition, une nouvelle variable $c = H(B)$ est découverte. Récursivement, la décomposition peut être appliquée à H. Il est donc possible de créer une hiérarchie de variables. Chaque nouvelle variable créée peut être décrite par une hiérarchie de variables. Les auteurs utilisent des fonctions avec des variables nominales. Le tout est implémenté dans le système HINT (Hierarchy INduction, Tool).

1.3.7.2 La meilleure partition de X

Il existe de nombreuses partitions différentes de X. La meilleure partition de X est celle dont le nombre de valeurs requises par la nouvelle variable c est minimum. Lors de la décomposition de F, un ensemble de partitions candidates sont examinées et celle qui entraîne une variable c avec le plus petit nombre de valeurs possibles est choisie pour la décomposition.

1.3.7.3 La construction de variables

Pour ajouter une seule variable, on procède de la manière suivante : pour chaque partition, une nouvelle variable est créée. La nouvelle variable conservée sera celle avec le moins de valeurs possibles. Pour ajouter L variables, avec $L > 1$, un ensemble candidat de combinaisons de variables initiales est examiné. Seront ajoutées les L nouvelles variables qui ont le plus petit nombre de valeurs.

1.3.8 Méthode utilisant la théorie des treillis

Pour construire des variables, Nguifo et Njiwoua, dans [120] utilisent la théorie des treillis. L'algorithme IGLUE, développé par les deux auteurs procède en deux étapes : la construction du treillis pour les exemples positifs de l'ensemble d'apprentissage et la construction de nouvelles variables.

Lors de la première étape, les variables qui n'appartiennent pas au treillis sont considérées comme non pertinentes. Lors de la deuxième étape, les nouvelles variables sont construites. Pour cela, ils attribuent à chaque variable X_k présente dans le treillis une nouvelle variable $d_k = \sum_{i=1}^{n} d_{ik}$. d_{ik} correspond au nombre de fois où la variable X_k est liée à l'individu ω_i dans le treillis.

1.3.9 Construction de variables basée sur la programmation génétique

Le système Genetic Constructive Induction (GCI) [114] utilise une approche par programmation génétique pour construire de nouvelles variables. Un algorithme d'apprentissage par rétropropagation réalise la tâche d'apprentissage traditionnelle puis utilise l'espace de représentation modifié. L'espace de recherche du système de programmation génétique consiste en l'ensemble des combinaisons possibles des opérateurs de construction. La qualité du programme génétique spécifique est mesurée par les performances d'un algorithme d'apprentissage sélectif apprenant avec l'espace de représentation amélioré. Pour éviter le problème de l'overfitting, un paramètre est utilisé pour mesurer la complexité attendue des variables nouvellement définies en se basant sur la taille du nouveau programme génétique. Le besoin en nouvelles variables est détecté par l'analyse des performances de l'algorithme d'apprentissage sélectif. Si les performances ne sont pas suffisamment bonnes, le système de programmation génétique recherchera des nouvelles variables qui amélioreront peut être les performances. De nombreux autres algorithmes utilisent la programmation génétique tels que GABIL [115], GA-SMART [27], ou [52].

1.3.10 Méthodes d'analyse de données

Dans la plupart des situations, on dispose de plusieurs observations sur chaque individu constituant la population d'étude. On a donc à prendre en compte *p* variables par individu, *p* étant strictement supérieur à 1. L'étude séparée de chacune de ces variables donne quelques informations mais est insuffisante car elle laisse de côté les liaisons entre variables, ce qui est pourtant souvent ce que l'on

veut étudier. C'est le rôle de la statistique multifactorielle, [117, 118], que d'analyser les données dans leur ensemble, en prenant en compte toutes les variables. Nous allons présenter les trois méthodes principales d'analyse de données qui permettent de construire de nouvelles variables. Il convient de noter que ces méthodes peuvent effectuer de la construction de variable en non supervisé.

1.3.10.1 Analyse en composantes principales (ACP)

L'ACP est une méthode qui permet d'étudier les données multidimensionnelles, lorsque toutes les variables observées sont de type numérique, de préférence dans les mêmes unités, et dans le but de découvrir les liens existants entre ces variables. Cette méthode permet également :

- de mettre en évidence des groupes d'individus homogènes vis-à-vis de l'ensemble des variables,
- de révéler des différences entre individus relativement à l'ensemble des variables,
- de repérer les individus ayant un comportement atypique,
- de réduire la quantité d'information nécessaire à la description de la position d'un individu dans l'ensemble de la population, et, de construire des variables statistiques synthétiques définies à partir des variables étudiées. Ces nouvelles variables artificielles permettent d' « expliquer » l'ensemble des variables statistiques prises en compte dans le processus de l'ACP. Il est important de noter que l'ACP permet de construire de nouvelles variables de manière non supervisée.

1.3.10.2 Analyse Factorielle des Correspondances (AFC)

Cette méthode ne traite que les variables qualitatives. Elle est destinée à extraire de l'information des tableaux de contingence et ainsi, étudier la correspondance entre les deux variables qualitatives constituant le tableau de contingence. Les objectifs de cette méthode sont doubles :

- Elle cherche à obtenir une représentation satisfaisante des individus contenus dans le tableau de contingence à l'aide d'une « carte des modalités » pour chacune des deux variables, le problème étant alors de choisir la « meilleure carte ». La carte des modalités est composée par la superposition des modalités-lignes et des modalités-colonnes du tableau de contingence. En ce sens, c'est un appareil à décrire très puissant ;
- Elle cherche à découvrir les liens pouvant exister entre les modalités de chaque variable. En ce sens, elle permet de créer de nouvelles variables synthétiques qui sont une combinaison des variables originales.

1.3.10.3 Analyse Factorielle des Correspondances Multiples (AFCM)

L'AFCM est une généralisation de l'AFC, quand il y a plus de deux variables qualitatives et en présence de variables quantitatives. L'intérêt principal de l'AFCM réside dans le fait qu'elle révèle les proximités existantes entre modalités d'une même variable, entre modalités de deux variables différentes, entre individus et entre individus et modalités. Pour cela, est défini un ensemble de distance caractérisant chaque type de proximités. La découverte de ces proximités peut ainsi conduire à la construction de variables artificielles.

1.4 Construction par utilisation des connaissances du domaine ou d'un expert

Ces méthodes utilisent les connaissances du domaine ou les connaissances fournies par un expert dans le but de construire et tester de nouvelles variables.

1.4.1 MIRO

MIRO, [101], utilise les connaissances du domaine sous la forme d'un ensemble de règles spécifiées par un expert afin de construire un nouvel espace de représentation. Le nouvel espace de représentation obtenu est de taille plus importante que l'espace initial et il lui sera appliqué un processus d'induction afin de construire de nouvelles variables.

1.4.2 IB3-CI

La méthode IB3-CI, [111], génère des variables booléennes à partir d'opérateurs de conjonction et combine l'apprentissage à partir d'instance de [122] avec la construction de variables incrémentale de [99].

1.4.3 AQ15

Michalski et al., [106], propose la méthode AQ15. Un expert du domaine définit un ensemble de règles qui sont fournies à l'algorithme AQ15. Ces règles sont soit sous une forme arithmétique soit sous une forme logique. AQ15 utilise cet ensemble des règles pour la construction de nouvelles variables.

1.4.4 RINCON

Le système RINCON a pour but explicite d'accroître la concision de la théorie du domaine, et est incrémental. Dans RINCON, la théorie du domaine est représentée par un graphe acyclique de concepts, organisés du général au spécifique.

1.5 Construction de variables par analyse topologique des arbres

Les méthodes de ce type déterminent de nouvelles variables par l'analyse des règles issues d'arbres d'induction.

1.5.1 CITRE

CITRE, [123], est un système basé sur les arbres de décision. Il effectue de la construction de variables en sélectionnant les relations des nouvelles variables dans les branches positives de l'arbre. Pour fournir à CITRE un problème d'apprentissage, l'ensemble de variables initiales $X = \{X_1,...,X_p\}$ et un ensemble d'individus d'apprentissage $\Omega = \{\omega_1,...,\omega_n\}$ décrit par les variables initiales lui sont fournis. A partir des variables initiales et des individus, un arbre de décision initial, basé sur un critère d'information est construit. CITRE sélectionne des paires de relations booléennes à partir des nœuds des branches positives de l'arbre , comme par exemple la relation booléenne suivante : $couleur = rouge\ et\ taille = grand$ où $couleur$ et $taille$ sont deux variables initiales. La sélection des relations booléennes se fait grâce à l'une des méthodes suivantes :

- Root : Sélection des relations dans les deux premiers nœuds de chaque branche positive ;
- Fringe : Sélection des relations dans les deux derniers nœuds de chaque branche positive ;
- Root-fringe : Sélection des relations dans les deux derniers et les deux premiers nœuds de chaque branche positive ;
- Adjacente : Toutes les paires adjacentes le long de chaque branche positive ;
- All : Toutes les combinaisons de paires de variables le long de chaque branche positive.

A partir de ces relations booléennes, CITRE forme de nouvelles variables booléennes : pour l'exemple précédent, la variable créée sera de la forme suivante : si l'individu ω_i possède à la fois la modalité $rouge$ pour la variable $couleur$ et la modalité $grand$ pour la variable $taille$ alors la valeur de la

nouvelle variable pour cet individu sera *VRAI* sinon sa valeur sera *FAUX* . Cependant, une contrainte est rajoutée lors de la sélection des paires de relations : ces dernières doivent passer au travers d'un filtre de connaissance du domaine qui élimine les paires ne satisfaisant pas les contraintes imposées par le domaine. Si l'on prend comme exemple la base Tic Tac Toe, l'une des contraintes liées au domaine est l'information sur l'adjacence des pièces. Le filtre permet ainsi de diminuer l'espace des nouvelles variables qui, ainsi, ne deviendra pas surchargé.

1.5.2 AQ17-HCI

Cette méthode de construction de variables, [102], génère des nouvelles variables à partir des règles obtenues à partir de l'algorithme AQ, [121].

```
AQ17-HCI
Début
    Induction des règles en utilisant l'algorithme AQ à partir d'un sous-ensemble de l'échantillon
d'apprentissage
    Identification des variables de l'ensemble initial non présentes dans les règles et élimination de ces
variables
    Pour chaque classe de la variable endogène Faire
        Génération d'une nouvelle variable représentant la règle de meilleure qualité
    Fin Pour
    Modification de l'ensemble d'apprentissage en y adjoignant les nouvelles variables et en gardant celles
considérées comme utiles
    Induction à partir de ce nouvel ensemble d'apprentissage
    Test des règles sur le reste des données d'apprentissage
    Si les performances ne sont pas satisfaisantes Alors
        Recommencer
     Sinon
        Extension de l'ensemble des données d'apprentissage avec les nouvelles variables
        Induction finale sur cet ensemble
      Fin Si
Fin
```

Algorithme 17 AQ17-HCI.

Les données d'entrées de AQ17–HCI sont constituées par l'ensemble d'apprentissage. AQ est appliqué sur l'ensemble d'apprentissage et génère un ensemble de règles. Ces règles sont ensuite évaluées par l'intermédiaire d'un critère d'évaluation, et les meilleures règles sont combinées en de nouvelles variables. Ces variables sont incorporées dans l'ensemble d'apprentissage et le processus d'apprentissage est répété. Le processus continue jusqu'à satisfaction d'un critère d'arrêt lié aux performances des nouvelles variables.

Les algorithmes STRUCT, [124], et PRAX, [125], fonctionnent de la même manière que AQ17-HCI : ils utilisent les règles d'apprentissage pour représenter et créer de nouvelles variables.

1.5.3 FRINGE

La méthode FRINGE, [95] et [94], s'applique initialement sur des problèmes à deux classes avec des variables exogènes booléennes. Cependant, la présence de variables qualitatives ne pose pas de problème. En effet, ces dernières subissent alors un codage disjonctif complet. C'est un processus itératif qui se déroule en trois étapes :

- Construction d'un arbre de décision,
- Analyse de cet arbre,
- Détermination des variables synthétiques candidates qui seront introduites dans la liste des variables prédictives.

Un nouvel arbre est construit et ainsi de suite, jusqu'à ce qu'aucune nouvelle variable ne soit construite. Cette méthode est relativement simple.

FRINGE
Données d'entrée : M le nombre maximal de variables prédictives
X l'ensemble initial des variables exogènes
Début
$i = 0$
fixer M
$X^1 = X$
Répéter
Construire un arbre à partir de l'échantillon d'apprentissage et de l'ensemble X^i
Construire l'ensemble des nouvelles variables F
$X^{i+1} = X^i \cup F$
jusqu'à ce que $X^{i+1} = X^I$ ou $\left|X^{i+1}\right| \geq M$
Fin

Algorithme 18 FRINGE

L'essentiel de la stratégie se situe dans l'extraction des nouvelles variables à partir de l'arbre c'est à dire la construction de F. Dans l'algorithme initial, on se contente de construire de nouvelles variables constituées de conjonctions de propositions des deux derniers nœuds précédant les feuilles, menant à une conclusion positive.

Plusieurs raisons expliquent le choix de cette stratégie :

- Les variables sont construites par combinaisons de variables deux à deux,
- Les nœuds situés sur la frange sont les moins sûrs, puisqu'ils couvrent très peu d'individus de la base d'apprentissage,
- La répétition des séquences de sous-arbres a lieu, le plus souvent, dans la partie basse du modèle.

D'autres études ont permis, par la suite, d'enrichir la liste des formes détectées avec le système FRINGE. Les études de [95], et [94] sont essentiellement basées sur les conjonctions de propositions, le terme proposition correspondant à une variable ou à sa négation. Les travaux [126] ont permis d'y ajouter les disjonctions, et les travaux [127], la forme XOR. L'algorithme de base ne change pas, ces perfectionnements touchant uniquement au pouvoir de représentation des nouvelles variables construites.

Toutefois, ces algorithmes ne se comportent correctement que pour un faible nombre de problèmes, ils ne sont pas adaptés à la prise en compte du bruit, et les formes détectées dépendent du caractère glouton de l'algorithme.

1.5.4 LFC

L'association de la construction de variables intermédiaires et de la forward selection se révèle selon [128], [129] plus efficace qu'une technique isolée, en termes de concision et de précision du classifieur. L'algorithme LFC met en œuvre ce principe et construit un arbre de décision en appliquant une recherche de type forward selection contrainte, combinée à la construction de variables synthétiques sur les nœuds.

LFC est un algorithme assez compliqué qui utilise plusieurs "astuces" pour limiter la complexité de la forward selection. La base de cet algorithme est la stratégie "divide and conquer". La sélection de la variable qui permettra la segmentation sur un nœud s'effectue par une recherche de type forward selection traduite à travers la composition de nouvelles variables formées de conjonctions de propositions. La limitation des investigations est réalisée à l'aide de différentes heuristiques qui rendent le parmétrage de LFC complexe.

Même si les expérimentations menées ont conduit à des résultats positifs sur des bases de données synthétiques, l'algorithme LFC, de par son paramétrage complexe et largement influent sur les résultats, se confronte à des difficultés lorsqu'il est utilisé sur des bases de données réelles.

GALA, [112], est un algorithme de construction qui ressemble à LFC de par son ensemble d'opérateurs.

1.6 Construction de variables Multi-stratégique

L'une des méthodologies les plus importantes en construction de variables est l'intégration de multiples stratégies d'apprentissage coopérant afin d'obtenir des résultats de bonne qualité.

1.6.1 AQ-BC

AQ-BC combine induction supervisée et classification bayésienne non supervisée. L'utilisation de la classification bayésienne par l'intermédiaire du système AUTOCLASS, [130], a pour but la création d'un espace de représentation plus adapté à l'apprentissage. En effet, AUTOCLASS découvre des relations et des caractéristiques intéressantes dans les données. Ainsi, grâce à la construction de variables, ces relations et ces caractéristiques permettent la modification de l'espace de représentation de manière à ce que l'algorithme d'apprentissage AQ15c, [131], puisse apprendre des concepts utiles.

1. AUTOCLASS permet d'extraire des classes décrivant des concepts abstraits, qui sont décrits par utilisation du système d'apprentissage AQ15c. Le rôle d'AQ15c est en fait de généraliser les descriptions des classes extraites par AUTOCLASS et d'exprimer ces descriptions en termes logiques.

2. Ces descriptions des concepts abstraits sont alors utilisées pour étendre et améliorer l'espace de représentation originel des données.

3. Ce nouvel espace de représentation des données est alors employé pour l'apprentissage supervisé final réalisé par AQ15c.

Algorithme 19 AQ-BC

Les descriptions de concepts abstraits appris lors de la première étape peuvent illustrer et associer les descriptions de concepts appris lors de la deuxième phase qui génère un ensemble de règles descriptives. De cette manière, les structures hiérarchisées d'hypothèses découvertes grâce aux

classifications imbriquées fournissent des informations intéressantes qui ne pourraient être obtenues à partir de l'un de ces systèmes utilisé seul.

1.6.2 INDUCE-1

La méthode INDUCE-1, [109], est une méthode dirigée à la fois par les données et par les connaissances du domaine. Elle utilise une variété de règles et de procédures pour générer de nouvelles variables, nommées méta-variables. Ceci s'effectue grâce à une description structurelle des exemples d'apprentissage, qui correspond à la partie dirigée par les connaissances du domaine, associée à la détermination des dépendances qualitative entre les variables, qui correspond à la partie dirigée par les données.

1.6.3 AQ17

La méthode AQ17, [107], est une méthode à la fois dirigée par les données, les hypothèses et les connaissances du domaine. Elle intègre de manière synergique les systèmes INDUCE-1, AQ15, AQ17-DCI et AQ17-HCI.

1.7 Conclusions

Il existe un grand nombre de méthodes de construction de variables. Cependant, les méthodes ne traitant qu'un seul type de variables sont difficilement applicables sur des problèmes réels. C'est le cas des méthodes BACON, STAGGER et la méthode de Zheng qui ne traitent que des variables numériques. C'est également le cas de la méthode de FRINGE qui ne travaillent que sur des variables booléennes. L'utilisation de cette méthode est également restreinte par le fait qu'elle ne traite que des problèmes à deux classes. La méthode LFC, quant à elle, est trop difficile à paramétrer.

Le type de méthodes qui nous paraît le plus attrayant sont les méthodes utilisant l'analyse topologique des arbres, telles que FRINGE. En effet, ces méthodes, grace à l'utilisation d'un arbre d'induction pour générer les règles qui serviront à la construction de nouvelles variables, tiennent comptent des relations existantes entre les variables exogènes et plus particulièrement entre certaines modalités de différentes variables exogènes. Ces méthodes sont de plus intéressantes car elles permettent de prendre en compte les liens entre les règles générées et les classes de la variable endogène.

2 Notre Méthode de Construction

La méthode que nous proposons appartient à la catégorie des méthodes de construction de variables par analyse topologique des arbres.

2.1 Cadre d'analyse

Notre méthode utilisant l'analyse topologique des arbres pour construire de nouvelles variables, nous avons choisi comme arbre d'induction pour générer les règles servant à la construction de nouvelles variables l'arbre ID3. Nous avons éliminé les arbres d'induction tels que Sipina ou ChAID, à cause du fait qu'ils autorisent les fusions. En effet, comme nous le verrons par la suite, nous cherchons à obtenir des règles vérifiées par un groupe d'individus ayant tous la même valeur pour la variable endogène.

Notre méthode travaille uniquement sur des variables qualitatives, ceci dans le but de traiter l'ensemble des variables de la même manière. Ainsi, les variables quantitatives sont discrétisées à l'aide de la méthode Fusinter, [132] et [71]. Dès lors, toutes les variables peuvent être considérées comme qualitatives.

Notre cadre de travail, inspirée de [133], reste, comme pour la partie sélection de variables, supervisé.

2.1.1 Formalisation

Soient Π une population étudiée et Ω un sous-ensemble de n individus observés : $\Omega = \{\omega_1,...,\omega_i,...,\omega_n\}$. La variable endogène Y possède un ensemble C de modalités représentant ses classes. Aussi, nous pouvons définir la variable endogène : $Y : \Omega \mapsto C = \{C_1,...,C_r,...,C_{\mathrm{R}}\}$. Chaque individu ω_i, avec $i \in [1,...,n]$, est décrit par p variables exogènes $X_1,...,X_k,...,X_p$ et appartient à la classe C_r avec $r \in [1,...,\mathrm{R}]$.

Soit O_k le domaine d'application de la variable X_k. Le cardinal de O_k est égal au nombre de modalités de X_k : $|O_k| = q_k$. Aussi, il est possible de définir la variable X_k de la manière suivante :

$$X_k : \Omega \mapsto O_k = \{m_1^k,...,m_j^k,...,m_{q^k}^k\}.$$

2.1.2 Définitions

Après la formalisation de notre problème, nous pouvons maintenant introduire un certain nombre de notions que nous utiliserons par la suite.

Notion 1 : Soit $P = \{p_1,...,p_l,...,p_L\}$ l'ensemble des prémisses, la prémisse $p_l \in P$ appartenant à P est une conjonction de plusieurs propositions logiques : $p_l = \bigcap Propositions\ logiques$. Une propositions logiques est une paire attribut-valeur. Dans notre cas, le travail avec les arbres d'induction implique qu'une prémisse $p_l \in P$ soit la conjonction de plusieurs modalités de différentes variables : $p_l = \bigcap_{\substack{j \\ k}} m_j^k$.

Notion 2 : Le résultat fourni par un arbre d'induction est un ensemble de règles $R = \{R_1,...,R_t,...,R_T\}$. Chaque règle $R_t = (p_l, C_r)$ est une combinaison unique entre une prémisse et une classe. La classe est considérée comme la conclusion de la règle et la prémisse comme sa condition.

Notion 3 : On appelle regroupement de règles l'ensemble $\Re_{g_r} = \{R_{g_r} | \forall l, R_{g_r} = (p_l, C_r)\}$. Cet ensemble regroupe toutes les règles qui ont pour conclusion la classe C_r et comme condition une prémisse quelconque.

Notion 4 : La base de construction BC d'un problème d'apprentissage est le support d'information à partir duquel seront construites les nouvelles variables. La base de construction regroupe l'ensemble des règles obtenues à partir de l'arbre d'induction et de l'ensemble initial des variables.

Notion 5 : Les données de construction DC d'un problème d'apprentissage sont constituées de l'ensemble des variables initiales et de la part des individus allouée au processus de construction de variables.

2.2 Point de départ et base méthodologique

Avant de présenter à proprement parler notre méthode, nous désirons exposer un certain nombre de remarques qui constituent le point de départ et les bases de notre méthodologie.

2.2.1 Les liens inter-variables

La première remarque concerne les liens existants non pas entre modalités d'une même variable mais entre modalités de plusieurs variables différentes. C'est ce que nous appelons les liens inter-variables. Lors du processus d'apprentissage, une variable est sélectionnée parce que c'est elle qui apporte le plus d'information. Or, ce fait est souvent induit par la présence d'une ou plusieurs modalités de cette variable mais, est très rarement induit par la totalité des modalités de la variable.

Afin d'illustrer au mieux cette remarque, nous allons présenter une exemple. La figure 20 nous montre le fragment d'un arbre de décision et en particulier, les t^{ième} et t+1^{ième} itérations du processus d'apprentissage. A l'itération t+1, la variable X_k est supposée être la variable qui apporte le plus d'information. Cette variable possède trois modalités m_1^k, m_2^k et m_3^k. La nouvelle partition inférée par le choix de la variable X_k entraîne la disparition de la feuille F_1 et l'apparition de trois nouvelles feuilles F_2, F_3 et F_4.

Le calcul du gain d'incertitude $\Im_{t+1}$ nous donne : $\Im_{t+1} = I_t - I_{t+1}$. I_t est l'incertitude liée à la partition obtenue en t et I_{t+1} est l'incertitude liée à la partition obtenue en $t+1$.

Dans notre exemple, à l'itération $t+1$, l'incertitude I_{t+1} se calcule de la manière suivante :

$$I_{t+1} = I_{t+1}(F_2, F_3, F_4) = \sum_{g=2}^{4} f_{g.}\left(-\sum_{j=1}^{3}\frac{f_{gj}}{f_{g.}}\log_2\frac{f_{gj}}{f_{g.}}\right)$$

$$= \underbrace{-\frac{25}{60}\left(\frac{5}{25}\log_2\frac{5}{25}+\frac{20}{25}\log_2\frac{20}{25}\right)}_{F_2} \underbrace{-\frac{5}{60}\left(\frac{5}{5}\log_2\frac{5}{5}+\frac{0}{5}\log_2\frac{0}{5}\right)}_{F_3} \underbrace{-\frac{30}{60}\left(\frac{10}{30}\log_2\frac{10}{30}+\frac{20}{30}\log_2\frac{20}{30}\right)}_{F_4}$$

$$= \underbrace{0{,}6016}_{F_2} + \underbrace{0}_{F_3} + \underbrace{0{,}4591}_{F_4}$$

avec $f_{gj} = \frac{n_{gj}}{n}$.

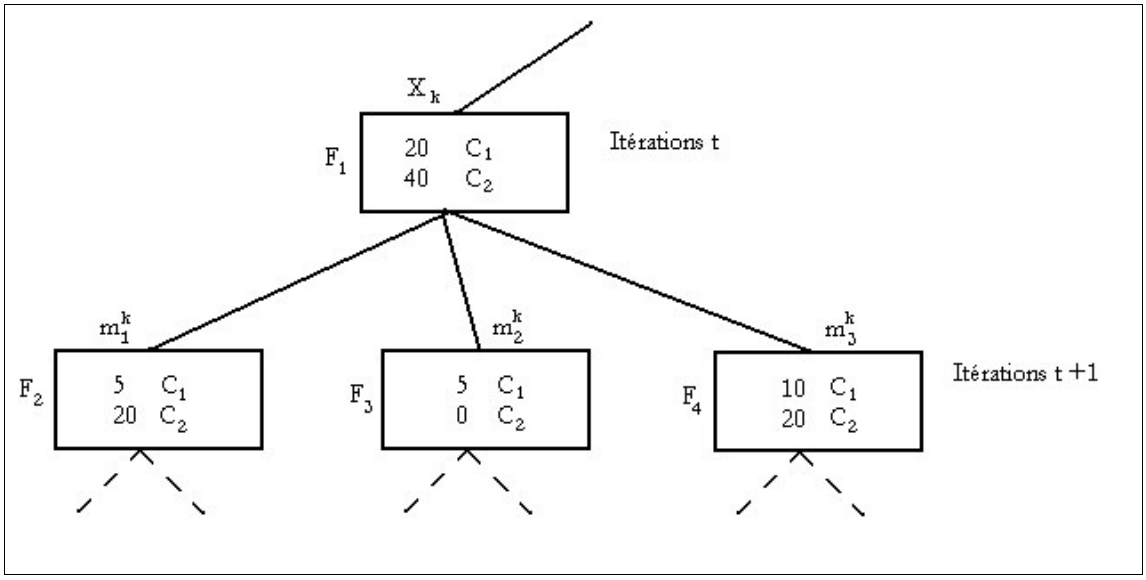

Figure 20 Fragment d'un arbre de décision – Exemple.

L'incertitude I_t étant déterminée à l'itération précédente, sa valeur est connue et constante en $t+1$ et la variable X_k a été sélectionnée car elle entraînait l'incertitude I_{t+1} la plus faible et permettait ainsi de maximiser le gain d'incertitude $\Im_{t+1}$. Or, on peut remarquer que la feuille F_3, liée à la modalité m_2^k, possède l'entropie la plus faible. C'est donc cette modalité qui discrimine le mieux et le plus la variable endogène. Ainsi, il existe une inégalité de discrimination au sein des modalités d'une même variable exogène.

Une variable endogène est donc discriminée par une modalité caractéristique plutôt que par une variable particulière. Aussi, il paraît judicieux de s'intéresser à une suite de modalités de variables qui discriminent au mieux la variable endogène.

Cette observation se retrouve également dans la méthode d'analyse de données l'AFCM, [117]. En effet, l'AFCM s'intéresse également aux liens ou proximités existants entre des modalités d'une même variable ou de deux variables différentes. Cependant, elle s'intéresse aux liens existants entre uniquement deux modalités et travaille en non supervisé. Notre méthode pourra travailler sur les liens existants entre deux ou plusieurs modalités et travaille en supervisé. Pour cette raison, nous un effectuerons tout au long de notre exposé un parallèle entre notre méthode et l'AFCM.

Toutefois, nous ne prétendons pas que notre méthode et l'AFCM se ressemble : l'AFCM fonctionne en non supervisé et notre méthode fonctionne en supervisé. Les liens découverts par notre méthode sont orientés en fonction de la variable endogène tandis que les liens découverts par l'AFCM correspondent à la structure même des données sans information sur la variable endogène. Nous tenons à effectuer un parallèle entre l'AFCM et notre méthode pour l'unique raison qu'elles s'intéressent l'une comme l'autre aux liens existants entre modalités de variables différentes. Pour cela, nous allons présenter la définition de la distance entre deux modalités d'une même variable et la distance entre deux modalités de deux variables différentes utilisées dans le processus de l'AFCM.

2.2.1.1 La distance entre deux modalités d'une même variable au sein de l'AFCM

Considérons les modalités j et j' de la variable X_k, soient m_j^k et $m_{j'}^k$. La distance entre ces deux modalités est définie comme suit :

$$d_k^2\left(m_j^k, m_{j'}^k\right) = \frac{1}{p}\sum_{i=1}^{n}\left(\frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_j^k \\ 0 \text{ sinon}\end{cases}}{\begin{array}{l}\text{Nbre d'ind. possédant} \\ \text{la modalités } m_j^k\end{array}} - \frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_{j'}^k \\ 0 \text{ sinon}\end{cases}}{\begin{array}{l}\text{Nbre d'ind. possédant} \\ \text{la modalités } m_{j'}^k\end{array}}\right)^2 = \frac{1}{p}\sum_{i=1}^{n}\left(Q_i^k\right)^2,$$

p est le nombre de variables et, i représente les individus. Nous posons :

$$Q_i^k = \frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_j^k \\ 0 \text{ sinon}\end{cases}}{\begin{array}{l}\text{Nbre d'ind. possédant} \\ \text{la modalités } m_j^k\end{array}} - \frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_{j'}^k \\ 0 \text{ sinon}\end{cases}}{\begin{array}{l}\text{Nbre d'ind. possédant} \\ \text{la modalités } m_{j'}^k\end{array}}.$$

Il existe trois cas possibles, pour un individu donné :

- Soit l'individu ne possède ni l'une ni l'autre des deux modalités, $card_1^k$ est le nombre d'individus appartenant à Ω ne possédant aucune des deux modalités ;
- Soit l'individu possède la modalité m_j^k, $card_2^k$ est le nombre d'individus appartenant à Ω possédant la modalité m_j^k ;

- Soit l'individu possède la modalité $m_{j'}^k$, $card_3^k$ est le nombre d'individus appartenant à Ω possédant la modalité $m_{j'}^k$.

La distance peut se réécrire de la manière suivante :

$$d_k^2 = \frac{1}{p}\sum_{i=1}^{n}\left(\left(\frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_j^k\\ 0 \text{ sinon}\end{cases}}{\text{Nbre d' ind. possédant la modalités } m_j^k}\right)^2 + \left(\frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_{j'}^k\\ 0 \text{ sinon}\end{cases}}{\text{Nbre d' ind. possédant la modalités } m_{j'}^k}\right)^2 - 2\underbrace{\left(\frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_j^k\\ 0 \text{ sinon}\end{cases}}{\text{Nbre d' ind. possédant la modalités } m_j^k}\right)\left(\frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_{j'}^k\\ 0 \text{ sinon}\end{cases}}{\text{Nbre d' ind. possédant la modalités } m_{j'}^k}\right)}_{\text{= 0 ; Un individu ne peut pas posséder les deux modalités d' une même variables.}}\right)$$

$$d_k^2 = \frac{1}{p}\left(\frac{1}{card_2^k} + \frac{1}{card_3^k}\right).$$

Il est maintenant facile de déterminer les cas où deux modalités d'une même variable sont considérées comme proches :

- Deux modalités d'une même variable seront considérées comme proches si un grand nombre d'individus les possèdent, dans ce cas la distance devient minimale et tend vers 0 ;
- La distance sera maximale pour des modalités rares, c'est à dire pour des modalités qui ne se retrouveront qu'une fois chacune, chez deux individus distincts. La distance sera alors égale à : $d_k^2 = \frac{2}{p}$.

2.2.1.2 La distance entre deux modalités de deux variables distinctes au sein de l'AFCM

Considérons les modalités j et j' de X_k et X_h, deux variables distinctes, soient m_j^k et $m_{j'}^k$. La distance entre ces deux modalités est définie comme suit :

$$d^2\left(m_j^k, m_{j'}^k\right) = \frac{1}{p}\sum_{i=1}^{n}\left(\frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_j^k \\ 0 \text{ sinon}\end{cases}}{\begin{array}{l}\text{Nbre d'ind. possédant} \\ \text{la modalités } m_j^k\end{array}} - \frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_{j'}^k \\ 0 \text{ sinon}\end{cases}}{\begin{array}{l}\text{Nbre d'ind. possédant} \\ \text{la modalités } m_{j'}^k\end{array}}\right)^2 = \frac{1}{p}\sum_{i=1}^{n}\left(Q_i\right)^2 ,$$

p est le nombre de variables et, i représente les individus. Nous posons :

$$Q_i = \frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_j^k \\ 0 \text{ sinon}\end{cases}}{\begin{array}{l}\text{Nbre d'ind. possédant} \\ \text{la modalités } m_j^k\end{array}} - \frac{\begin{cases}1 \text{ si } \omega_i \text{ possède } m_{j'}^k \\ 0 \text{ sinon}\end{cases}}{\begin{array}{l}\text{Nbre d'ind. possédant} \\ \text{la modalités } m_{j'}^k\end{array}} .$$

Plus cette distance est proche de 0 et plus il y a de concordances entre les deux modalités ; ce qui signifie qu'il y a un grand nombre d'individus chez lesquels les deux modalités on été observées simultanément. Si cette distance s'éloigne de 0, cela signifie que nous retrouverons rarement ces deux modalités ensemble chez les individus.

Pour un individu donné, quatre situations sont possibles, ce qui nous permet de découper l'ensemble des individus Ω en quatre sous-ensembles :

- Ω_1 qui regroupe tous les individus ne possédant aucune des deux modalités. Pour l'individu $\omega_i \in \Omega_1$ qui appartient à ce sous-ensemble, la valeur de Q_i est égale à 0, $Q_i = 0$. Ainsi, ces individus n'auront aucune influence sur la valeur de la distance entre les deux modalités. Le cardinal de Ω_1 est noté $card_1$;
- Ω_2 qui regroupe les individus possédant la modalité m_j^k et ne possédant pas la modalité $m_{j'}^k$. Pour les individus appartenant à ce sous-ensemble, Q_i sera strictement positif, $Q_i > 0$. Ainsi, ces individus contribueront à maximiser cette distance. Le cardinal de Ω_2 est noté $card_2$;
- Ω_3 qui regroupe les individus possédant la modalité $m_{j'}^k$ et ne possédant pas la modalité m_j^k. Pour les individus appartenant à ce sous-ensemble, Q_i sera strictement négatif, $Q_i < 0$. Ainsi, ces individus contribueront à maximiser cette distance. Le cardinal de Ω_3 est noté $card_3$;

- Ω_4 qui regroupe les individus possédant la modalité $m_{j'}^k$ et la modalité m_j^k. Pour les individus appartenant à ce sous-ensemble, la valeur de Q_i sera soit positive soit négative et sa valeur sera inférieure à celles des cas où ω_i appartient à Ω_2 ou Ω_3, $|Q_i| \geq 0$. Ainsi, ces individus contribueront à minimiser cette distance. Le cardinal de Ω_4 est noté $card_4$.

Nous remarquons que : $card_1 + card_2 + card_3 + card_4 = n$, n étant la nombre total d'individus.

A l'aide de cette formulation, il est possible de réécrire la formule de la distance entre deux modalités de variables différentes :

$$d^2\left(m_j^k, m_{j'}^k\right) = \frac{1}{K}\left(\frac{card_2}{(card_2 + card_4)^2} + \frac{card_3}{(card_3 + card_4)^2}\right)$$

d^2 aura une valeur maximale lorsque $card_4 = 0$, c'est à dire dans le cas où aucun individu ne possède à la fois les deux modalités. Cette situation est équivalente à la situation de deux modalités d'une même variable. d^2 sera alors égale à $d^2 = \frac{1}{K}\left(\frac{1}{card_2} + \frac{1}{card_4}\right)$, et sera maximale lorsque $card_2$ et $card_3$ seront les plus petits possible. Ceci est le cas quand $card_2 = card_3 = 1$ et donc $card_1 = n - 2$.

d^2 aura une valeur minimale lorsque $card_2 = card_3 = 0$, c'est à dire dans le cas où aucun individu ne possède que l'une ou l'autre des deux modalités, la distance est alors nulle : $d^2 = 0$. Dans cette situation, $card_4 \neq 0$, sinon cela signifierait qu'aucun individu ne possède ces modalités : ce cas est exclu. Ainsi, il n'y a ici que des concordances entre les deux modalités, tous les individus possèdent simultanément ces deux modalités. Si $card_2$ et $card_3$ sont non nuls alors la distance sera d'autant plus faible que $card_4$ sera élevé, c'est à dire qu'un grand nombre d'individus possèdent à la fois les deux modalités.

Nous retrouvons donc au sein même du processus de l'analyse de données, cette notion de proximité entre modalités de variables différentes dont il nous paraît important de tenir compte.

Nous désirons que notre méthode tienne compte de ce lien entre modalités inter-variables. C'est pour cette raison que les variables que nous allons créer seront sous la forme de conjonction de modalités de plusieurs variables différentes. Bien sûr, nous nous intéressons uniquement aux modalités de variables distinctes. Car, le long d'une branche d'un arbre d'induction, en particulier de ID3, les individus ne pourront pas posséder deux modalités d'une même variable. Ces conjonctions nous permettront de discriminer au mieux la variable endogène et seront obtenues à l'aide des règles issues de l'arbre de décision.

2.2.2 Intérêt des franges ou parties basses des arbres d'induction

Lors du processus d'apprentissage, les parties basses des arbres d'induction sont toujours élaguées à cause du trop petit nombre d'individus composant leurs différentes feuilles. Ce nombre minime d'individus entraîne un gain d'information faible pour ces parties basses. Ainsi, elles ne satisfont pas le critère du gain d'information fixé par l'utilisateur et sont éliminées. Cependant, ces franges peuvent contenir de l'information et de ce fait être pertinentes. C'est le cas en particulier pour les bases dont la variable endogène possède une classe moins bien représentée par rapport aux autres.

En conséquence, nous tenons à prendre en compte les parties basses des arbres d'induction et en l'occurrence de ID3. Afin d'atteindre les différentes feuilles composant les franges, nous supprimons l'une des contraintes de l'arbre de décision ID3 : la contrainte liée au gain d'incertitude minimal. La croissance de ID3 ne sera uniquement limitée par :

- Soit le fait que le nœud terminal ne contienne que des individus d'une même classe ;
- Soit par le fait que toutes les variables exogènes soient, à l'itération en cours, indépendantes de la variable endogène. Afin de déterminer cette indépendance, nous appliquerons le test du χ^2.

Cette remarque nous conduit à définir une notion supplémentaire :

Notion 6 : L'arbre pour lequel la contrainte liée au gain d'information minimal est supprimée est nommé arbre non contraint. Les règles issues de cet arbre non contraint constitueront notre base de construction.

2.3 Présentation et déroulement de notre méthode

Les caractéristiques de notre méthode sont les suivantes :

- Elle fonctionne en apprentissage supervisé,
- Elle appartient à la catégorie des méthodes de construction de variables par analyse topologique des arbres,
- Elle traite uniquement des variables qualitatives,
- Elle utilise pour la création de nouvelles variables les opérateurs booléens ET et OU,
- Les variables créées sont de type booléen.

Notre méthode se décompose en trois étapes distinctes et successives : Génération de la base de construction, Classement de la base de construction, et Construction des nouvelles variables (figure 21).

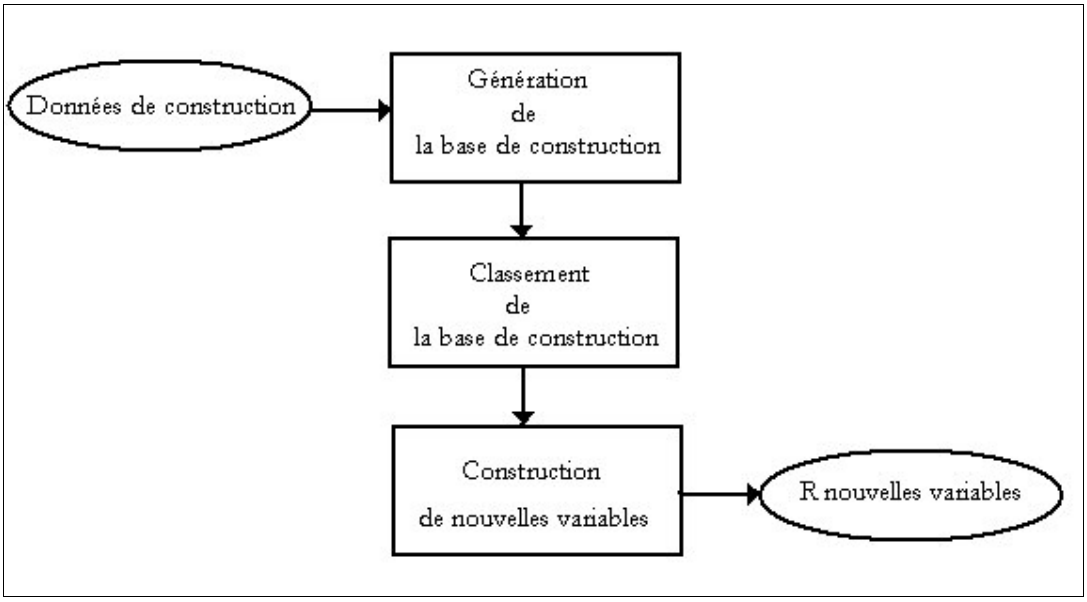

Figure 21 Processus de construction.

Afin de visualiser le processus de construction, un exemple artificiel illustrera toutes les étapes de notre méthode. Cet exemple très simple est composé de 10 individus, de 3 variables exogènes à valeur dans $\{C, D\}$ et d'une variable endogène à deux classes $C = \{A, B\}$. Le tableau 44 reprend les données consacrées à la construction de variables.

	X_1	X_2	X_3	Y
1	D	D	D	A
2	D	C	D	A
3	D	C	D	A
4	D	D	D	A
5	D	D	C	B
6	C	D	D	B
7	D	D	C	B
8	C	D	D	B
9	D	D	C	B
10	C	D	D	B

Tableau 44 Echantillon de données.

2.3.1 Génération de la base de construction

Le premier élément nécessaire pour construire de nouvelles variables est la base de construction. Notre base de construction est constituée d'un ensemble de règles, figure 22. L'arbre d'induction non contraint est lancé sur les données initiales réservées à la phase de construction. Nous obtenons un ensemble de règles. Ces règles sont de la forme suivante :

$Si\ p_l \quad Alors\ Y = C_r$ avec $1 \le l \le L$ et $1 \le r \le R$.

Chaque règle est caractérisée par :

- La modalité de la variable endogène qui lui est associée ;
- Le nombre d'individus vérifiant cette règle ;
- Un support : le support d'une règle indique le pourcentage d'individus vérifiant cette règle ;
- Et, une confiance : la confiance d'une règle mesure sa validité. C'est le pourcentage d'individus vérifiant la conclusion de la règle parmi ceux qui vérifient sa prémisse.

La contrainte liée au gain d'incertitude minimal ayant été supprimée, le nombre de règles générées est assez important. Ne seront conservées que les règles :

- Qui auront un nombre d'individus supérieur à 0 ; c'est à dire dont le support et la confiance (de part leur définition) seront non nuls ;
- Pour lesquelles il sera possible d'attribuer une modalité de la variable endogène : le nombre d'individus vérifiant la conclusion de la règle doit être supérieur au nombre d'individus vérifiant la prémisse de la règle et non sa conclusion ; c'est à dire dont la confiance sera supérieure ou égale à $0{,}5$.

L'ensemble des règles restantes forme notre base de construction. Ces règles sont des règles pour lesquelles les individus les vérifiant appartiennent tous à la même classe : Quelque soit l'individu ω_i vérifiant la règle $R_t = (p_l, C_r)$, la valeur de la variable endogène pour cette individu sera C_r, $Y(\omega_i) = C_r$.

Aussi, si nous poursuivons notre parallèle avec l'AFCM, les individus vérifiant la règle $R_t = (p_l, C_r)$, lors du calcul de la distance entre les modalités, constituant R_t prises deux à deux, appartiennent au sous-ensemble Ω_4 c'est à dire au sous-ensemble contenant tous les individus possédant à la fois les deux modalités considérées. Donc, du point de vue de l'AFCM, ce sont des individus qui contribueront à rendre la distance entre modalités minime. Plus ils seront nombreux et plus la distance sera faible.

Pour notre exemple, la figure 23 nous montre le résultat de l'Arbre Non Contraint. La base de construction est composée d'un ensemble de 4 règles .

Les règles sont décrites dans le tableau 45. Chaque règle possède un nombre d'individus et un support supérieur à 0 ainsi qu'une confiance supérieure ou égale à 0,5. Dans cet exemple illustratif, comme chaque variable exogène ne comporte que deux modalités, toutes les règles ont une confiance supérieure à 0,5. Ainsi, elles seront toutes sélectionnées pour la construction de nouvelles variables.

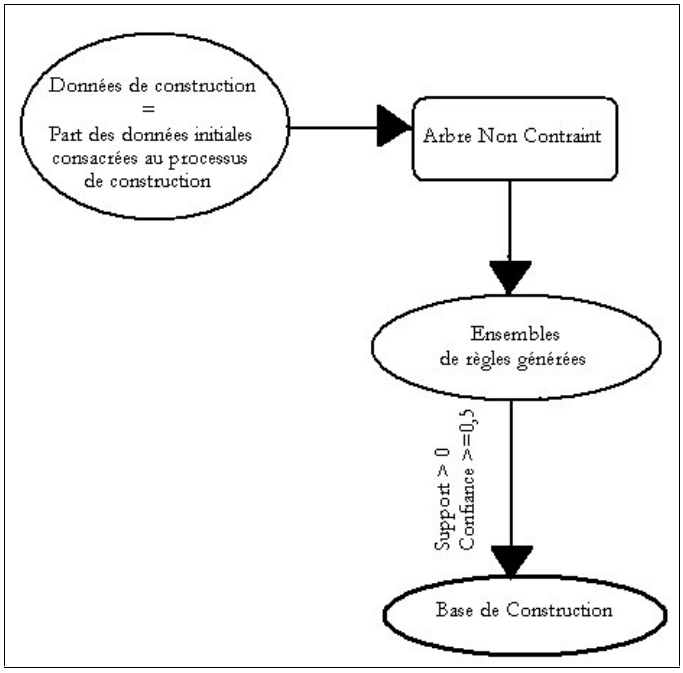

Figure 22 Génération de la base de construction.

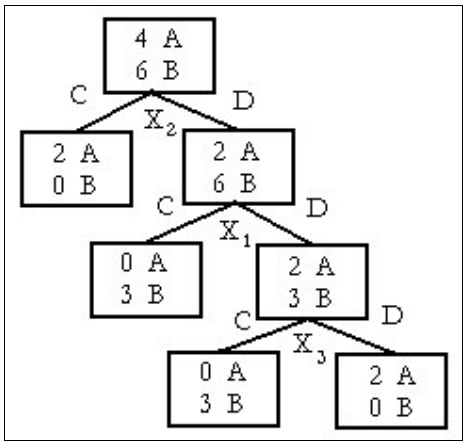

Figure 23 Résultat de l'arbre non contraint.

Règle	Classe	Nombre d'individus	Support	Confiance
Si $X_2 = D$ et $X_1 = D$ et $X_3 = D$ Alors $\mathrm{Y} = \mathrm{A}$	A	2	0,2	1
Si $X_2 = C$ Alors $\mathrm{Y} = \mathrm{A}$	A	3	0,3	1
Si $X_2 = D$ et $X_1 = C$ Alors $\mathrm{Y} = \mathrm{B}$	B	3	0,3	1
Si $X_2 = D$ et $X_1 = D$ et $X_3 = C$ Alors $\mathrm{Y} = \mathrm{B}$	B	2	0,2	1

Tableau 45 Base de construction.

Pour la règle présentée dans la troisième ligne du tableau 45, les individus la vérifiant sont les individus 6, 8 et 10. Si nous continuons le parallèle avec l'AFCM, ces trois individus font partie du sous-ensemble Ω_4 lors du calcul de la distance entre la modalité C de la variable X_1 et la modalité D de la variable X_2. D'ailleurs, dans ce cas précis, le cardinal de Ω_4 est égal à 3, $card_4 = 3$. Le calcul de la distance entre ces deux modalités donne comme résultat : $d^2 = \frac{1}{3}\left(\frac{3}{(3+3)^2} + \frac{8}{(8+3)^2} \right) = 0{,}034$.

2.3.2 Classement de la base de construction

Cette deuxième étape, figure 24, consiste à classer l'ensemble des règles sélectionnées à l'étape précédente. Les règles sont regroupées en fonction de la classe de la variable endogène qui leur est associée. Subséquemment, les règles de la forme $Si\ p_l\ \ Alors\ Y = C_r$ appartiennent au regroupement de règles nommé $\Re_{g_r}$. Il y aura autant de regroupements de règles que de classes de la variable endogène.

Dans notre exemple, la variable endogène ayant deux modalités A et B, deux regroupements de règles sont formés. Chaque regroupement est constitué de deux règles, tableau 46.

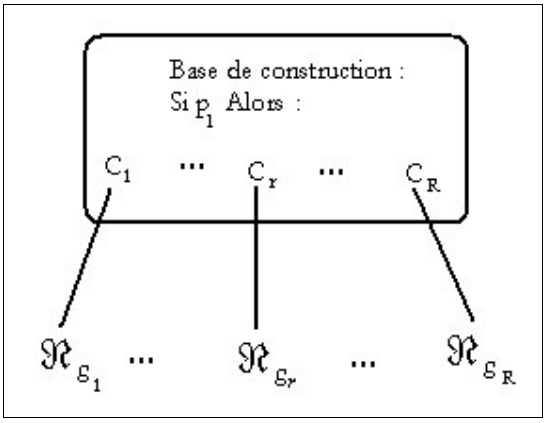

Figure 24 Classement de la base de construction.

Regroupement de règles	Variable intermédiaire	Nouvelle variable
Si $X_2 = D$ et $X_1 = D$ et $X_3 = D$ Alors Y = A	x	X_A
Si $X_2 = C$ Alors Y = A	y	
Si $X_2 = D$ et $X_1 = C$ Alors Y = B	z	X_B
Si $X_2 = D$ et $X_1 = D$ et $X_3 = C$ Alors Y = B	v	

Tableau 46 Regroupements de règles.

2.3.3 Construction des nouvelles variables

La dernière étape, figure 25 est la plus importante : c'est elle qui va permettre la création à proprement dite des nouvelles variables.

Pour chaque regroupement de règles, une nouvelle variable sera construite. Le nombre de nouvelles variables sera donc égal au nombre de modalités de la variable endogène. Cela permet de limiter le nombre de variables créées et de ne pas surcharger inutilement l'espace de représentation.

Toutes les règles sont des conjonctions de modalités de variables. Et, à partir de chaque regroupement de règles, la construction d'une nouvelle variable s'effectue en deux phases successives.

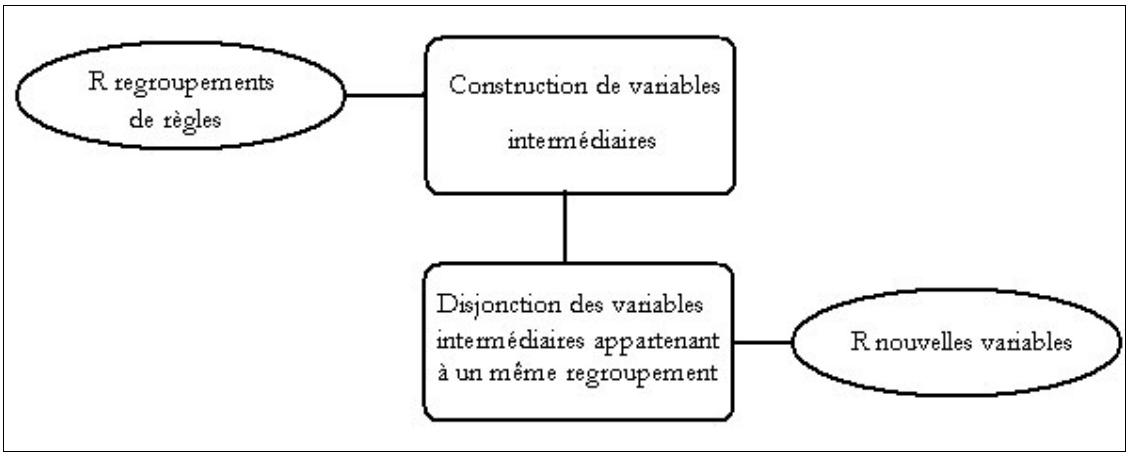

Figure 25 Construction de nouvelles variables.

La première phase concerne la création de variables intermédiaires x_{il} pour chaque regroupement de règles :

$$\forall r \in [1,...,R] \quad \forall R_t = (p_l, C_r) \in \Re_{g_r},$$
$$x_{il} : (R_t, \omega_i) \mapsto \{Vrai, Faux\}$$
$$x_{il} = \begin{cases} Vrai \ si \ \omega_i \ vérifie \ p_l \ et \ C_r \\ Faux \ Sinon \end{cases}$$

Une variable intermédiaire est créée pour chaque règle de la base de construction. Si un individu satisfait la prémisse de la règle et sa conclusion alors sa valeur pour la variable x_{il} associée à cette règle sera Vrai sinon sa valeur sera égale à Faux.

Les nouvelles variables peuvent maintenant être construites. Elles sont au nombre de R et de la forme suivante :

$$\forall r \in [1,...,R], \ X_{ik} = \bigcup_{R_l \in \Re_{g_r}} x_{il} \ .$$

Chaque nouvelle variable est la disjonction des variables intermédiaires d'un même regroupement c'est à dire la disjonction des règles d'un même regroupement. Ainsi, les nouvelles variables construites sont au nombre de R, le nombre de classes du problème étudié et sont de nature booléenne.

Le tableau 47 nous présente à la fois les variables intermédiaires créées ainsi que les nouvelles variables construites pour notre exemple. Il y a donc quatre variables intermédiaires (autant que de règles générées par l'Arbre Non Contraint) et deux nouvelles variables (autant que de classes de la variable endogène). Les variables intermédiaires sont obtenues de la manière suivante :

- $x = (X_2 = D) \cap (X_1 = D) \cap (X_3 = D)$,
- $y = (X_2 = C)$,
- $z = (X_2 = D) \cap (X_1 = C)$,
- $v = (X_2 = D) \cap (X_1 = D) \cap (X_3 = C)$.

X_A et X_B sont construites de la manière suivante :

- $X_A = x \cup y$,
- $X_B = z \cup v$.

Notre méthode tient compte de l'algorithme d'apprentissage, de ses caractéristiques et de son influence sur les données étudiées. C'est une méthode qui prend en compte les liens existants entre les variables, et ce par le fait que les variables construites sont des conjonctions de modalités de différentes variables. De plus le nombre de variables créées est faible, donc l'espace de représentation ne devient pas surchargé.

2.4 Etude comparative entre l'AFCM et notre méthode

Afin d'effectuer une comparaison entre l'AFCM et notre méthode, nous avons décidé de traiter un même exemple par les deux méthodes. Nous savons déjà que notre méthode, à l'instar de l'AFCM, cherche à découvrir les liens existants entre plusieurs modalités de plusieurs variables différentes, cet exemple nous permettra de vérifier la concordance entre les liaisons découvertes par l'AFCM et celle remarquées par notre méthode.

Les données de cet exemple artificiel sont constituées de quatre variables exogènes et une variable endogène. Les variables exogènes sont de type qualitatif. Elles représentent la taille, le poids, l'âge et le sexe d'un ensemble de 26 enfants. Ces enfants ont entre 4 et 6 ans. Nous cherchons à étudier la croissance de ces enfants ; ils ont été répartis en deux classes : ceux qui ont eu une croissance rapide et

ceux qui ont eu une faible croissance. Le tableau 48 nous montre l'ensemble de l'échantillon étudié. Pour l'application de notre méthode la variable endogène sera le type de croissance.

	Variables initiales			Variables intermédiaires		**Nouvelle variable**	Variables intermédiaires		**Nouvelle variable**
	X_1	X_2	X_3	x	y	X_A	z	v	X_B
1	D	D	D	Vrai	Faux	**Vrai**	Faux	Faux	**Faux**
2	D	C	D	Faux	Vrai	**Vrai**	Faux	Faux	**Faux**
3	D	C	D	Faux	Vrai	**Vrai**	Faux	Faux	**Faux**
4	D	D	D	Vrai	Faux	**Vrai**	Faux	Faux	**Faux**
5	D	D	C	Faux	Faux	**Faux**	Faux	Vrai	**Vrai**
6	C	D	D	Faux	Faux	**Faux**	Vrai	Faux	**Vrai**
7	D	D	C	Faux	Faux	**Faux**	Faux	Vrai	**Vrai**
8	C	D	D	Faux	Faux	**Faux**	Vrai	Faux	**Vrai**
9	D	D	C	Faux	Faux	**Faux**	Faux	Vrai	**Vrai**
10	C	D	D	Faux	Faux	**Faux**	Vrai	Faux	**Vrai**

Tableau 47 Variables intermédiaires et nouvelles variables.

Pour pouvoir appliquer l'AFCM, nous avons transformé notre tableau de données en tableau de disjonctif complet, tableau 49. La variable endogène a été supprimée, elle ne sera pas utilisée pour l'AFCM.

2.4.1 Résultats obtenus avec notre méthode

Un ensemble de dix règles est obtenu avec l'arbre non contraint : seules neuf sont sélectionnées pour la base de construction. Deux variables sont créées.

Individu	Sexe	Taille	Poids	Âge	Type de croissance
1	Garçon	112 - 120 cm	15 - 21 kg	4 à 5 ans et demi	Croissance Rapide
2	Garçon	104 - 112 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Rapide
3	Fille	104 - 112 cm	15 - 21 kg	4 à 5 ans et demi	Croissance Rapide
4	Garçon	Plus de 120 cm	21 - 28 kg	5 ans et demi à 6 ans	Croissance Rapide
5	Garçon	112 - 120 cm	21 - 28 kg	5 ans et demi à 6 ans	Croissance Rapide
6	Garçon	112 - 120 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Rapide
7	Fille	104 - 112 cm	15 - 21 kg	4 à 5 ans et demi	Croissance Rapide
8	Garçon	Plus de 120 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Rapide
9	Fille	Plus de 120 cm	21 - 28 kg	5 ans et demi à 6 ans	Croissance Rapide
10	Fille	112 - 120 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Rapide
11	Fille	112 - 120 cm	21 - 28 kg	5 ans et demi à 6 ans	Croissance Rapide
12	Fille	104 - 112 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Rapide
13	Fille	112 - 120 cm	15 - 21 kg	4 à 5 ans et demi	Croissance Rapide
14	Garçon	112 - 120 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
15	Fille	104 - 112 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
16	Garçon	112 - 120 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
17	Garçon	112 - 120 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
18	Garçon	112 - 120 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
19	Fille	112 - 120 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
20	Garçon	104 - 112 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
21	Fille	104 - 112 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
22	Garçon	104 - 112 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
23	Garçon	112 - 120 cm	15 - 21 kg	5 ans et demi à 6 ans	Croissance Lente
24	Garçon	104 - 112 cm	15 - 21 kg	4 à 5 ans et demi	Croissance Lente
25	Garçon	112 - 120 cm	15 - 21 kg	4 à 5 ans et demi	Croissance Lente
26	Garçon	104 - 112 cm	15 - 21 kg	4 à 5 ans et demi	Croissance Lente

Tableau 48 Données de l'échantillon étudié.

Individu	Sexe 1	Sexe 2	Taille 1	Taille 2	Taille 3	Poids 1	Poids 2	Age 1	Age 2	Somme
	Garçon	Fille	104 - 112 cm	112 - 120 cm	Plus de 120 cm	15 - 21 kg	21 - 28 kg	4 à 5 ans et demi	5 ans et demi à 6 ans	
1	1	0	0	1	0	1	0	1	0	4
2	1	0	1	0	0	1	0	0	1	4
3	0	1	1	0	0	1	0	1	0	4
4	1	0	0	0	1	0	1	0	1	4
5	1	0	0	1	0	0	1	0	1	4
6	1	0	0	1	0	1	0	0	1	4
7	0	1	1	0	0	1	0	1	0	4
8	1	0	0	0	1	1	0	0	1	4
9	0	1	0	0	1	0	1	0	1	4
10	0	1	0	1	0	1	0	0	1	4
11	0	1	0	1	0	0	1	0	1	4
12	0	1	1	0	0	1	0	0	1	4
13	0	1	0	1	0	1	0	1	0	4
14	1	0	0	1	0	1	0	0	1	4
15	0	1	1	0	0	1	0	0	1	4
16	1	0	0	1	0	1	0	0	1	4
17	1	0	0	1	0	1	0	0	1	4
18	1	0	0	1	0	1	0	0	1	4
19	0	1	0	1	0	1	0	0	1	4
20	1	0	1	0	0	1	0	0	1	4
21	0	1	1	0	0	1	0	0	1	4
22	1	0	1	0	0	1	0	0	1	4
23	1	0	0	1	0	1	0	0	1	4
24	1	0	1	0	0	1	0	1	0	4
25	1	0	0	1	0	1	0	1	0	4
26	1	0	1	0	0	1	0	1	0	4
Somme	16	10	10	13	3	22	4	7	19	104

Tableau 49 Tableau disjonctif complet

Les liens inter-variables découverts sont les suivants :

- Règle 1 liant Poids 1, Sexe 2 et Âge 1 ;

- Règle 2 liant Poids 1, Sexe 1 et Taille 3 ;
- Règle 3 liant Poids 1, Sexe 2, Âge 2 et Taille 2 ;
- Règle 4 liant Poids 1, Sexe 2, Âge 2 et Taille 1 ;
- Règle 5 liant Poids 1, Sexe 2, Âge 2 et Taille 1 ;
- Règle 6 liant Poids 1, Sexe 1, Âge 1 et Taille 2 ;
- Règle 7 liant Poids 1, Sexe 1, Âge 2 et Taille 2 ;
- Règle 8 liant Poids 1, Sexe 1, Âge 1 et Taille 1 ;
- Règle 9 liant Poids 1, Sexe 1, Âge 2 et Taille 1.

Chacune de ces règles a entraîné la création d'une variable intermédiaire. Les règles 1 et 2 ont permis de construire la variable synthétique correspondante à la classe « Croissance Rapide » et les autres règles la nouvelle variable associée à la classe « Croissance Lente ».

2.4.2 Résultats de l'AFCM

Comme résultat de l'AFCM, nous ne donnons en détail que le tableau récapitulant les distances entre modalités de variables. Le meilleur plan sur lequel projeter le nuage des 26 modalités-lignes et 9 modalités-colonnes est engendré par les vecteurs $\vec{u_1}$ et $\vec{u_2}$ suivant de $\mathbf{R^9}$:

$$\vec{u_1} = (0{,}031; -0{,}039; -0{,}343; 0{,}045; 0{,}533; -0{,}244; 0{,}571; -0{,}390; 0{,}237)$$
$$\vec{u_2} = (0{,}434; -0{,}549; -0{,}396; 0{,}466; -0{,}246; 0{,}086; -0{,}202; -0{,}142; 0{,}086).$$

Les valeurs propres qui leur sont associées sont $\lambda_1 = 0{,}436$ et $\lambda_2 = 0{,}297$.

Dans le tableau 50, nous avons grisé un certain nombre de règles utilisées pour construire de nouvelles variables par notre méthode. La règle 8 est avec un fond noir, la règle 1 est avec un fond gris clair et la règle 3 est en gris foncé. Bien sûr, l'AFCM ne calculant que des distances entre deux modalités, chaque règle est représenté par un ensemble de distance : la règle 1, par exemple, est caractérisée par les distances Poids 1-Sexe 2, Poids 1-Age 1 et Sexe 2-Age 1. Les distances caractérisant chaque règle sont pour la plupart parmi les plus faibles et proches de 0. Les règles qui serviront à la construction de variables n'entraînent pas forcément les distances les plus faibles entre les modalités, mais des distances parmi les faibles pour chaque modalité.

Si l'on considère la règle 1, la modalité Poids 1 est liée avec Sexe 1, Taille 1 et Age 2. Or les distances entre Poids 1-Sexe 1, Poids 1-Taille 1 et Poids 1-Age 2 font bien partie des distances les plus faibles entre Poids 1 et toutes les autres modalités des autres variables.

	Sexe 1	**Sexe 2**	**Taille 1**	**Taille 2**	**Taille 3**	**Poids 1**	**Poids 2**	**Age 1**	**Age 2**
Sexe 1	0,0000	0,0406	0,0250	0,0132	0,0781	0,0071	0,0625	0,0335	0,0090
Sexe 2	0,0406	0,0000	0,0250	0,0288	0,0917	0,0182	0,0625	0,0393	0,0090
Taille 1	0,0250	0,0250	0,0000	0,0442	0,1083	0,0136	0,0875	0,0321	0,0224
Taille 2	0,0132	0,0288	0,0442	0,0000	0,1026	0,0114	0,0625	0,0385	0,0121
Taille 3	0,0781	0,0917	0,1083	0,1026	0,0000	0,0871	0,0625	0,1190	0,0702
Poids 1	0,0071	0,0182	0,0136	0,0114	0,0871	0,0000	0,0739	0,0244	0,0066
Poids 2	0,0625	0,0625	0,0875	0,0625	0,0625	0,0739	0,0000	0,0982	0,0493
Age 1	0,0335	0,0393	0,0321	0,0385	0,1190	0,0244	0,0982	0,0000	0,0489
Age 2	0,0090	0,0197	0,0224	0,0121	0,0702	0,0066	0,0493	0,0489	0,0000

Tableau 50 Distances entre modalités.

Le principe même de l'AFCM, et en particulier l'utilisation de tableau de disjonctif complet, montre l'importance de la modalité d'une variable et non pas de l'ensemble des modalités d'une même variable. Ce qui se retrouve également dans notre méthode. De plus, les liens inter-variables découverts par notre méthode sont confirmés par le calcul des distances entre modalités de l'AFCM.

2.5 Conclusions

Notre méthode de construction de variables appartient à la catégorie des méthodes par analyse topologique des arbres. En effet, notre méthode permet la création de variables grâce à l'analyse des règles issues d'un arbre d'apprentissage en l'occurrence ID3. Notre méthode travaille en apprentissage supervisé et ne traite que les variables qualitatives. Elle s'intéresse et prend en compte les liens existants entre les variables exogènes par le biais de conjonctions et disjonctions de modalités de variables exogènes différentes. En ce sens, elle se rapproche de l'AFCM.

Après avoir généré un ensemble de règles à partir de l'arbre non contraint, ces différentes règles sont classées en fonction de leur conclusion, c'est à dire en fonction de la modalité de la variable endogène

qui leur est associée. Une variable intermédiaire est créée pour chaque règle générée. Les variables intermédiaires sont des conjonctions des modalités constituant les prémisses des règles. Les modalités formant une variable intermédiaire sont des modalités appartenant à des variables différentes les unes des autres. Comme les règles, les variables intermédiaires sont regroupées en fonction de la classe de la variable endogène qui leur est associée. Les nouvelles variables peuvent maintenant être construites : elles sont des disjonctions des variables intermédiaires appartenant à un même groupe. Ainsi, les opérateurs de construction que nous utilisons sont au nombre de deux : ET et OU. Les variables synthétiques sont de type booléen. Il y en a autant qu'il y a de classes pour la variable endogène. Cette limitation du nombre de variables construites permet de ne pas surcharger inutilement l'espace de représentation des données.

3 Expérimentations

3.1 Présentation des expérimentations

Nous avons testé notre méthode sur un ensemble de quatorze bases issues de la collection de l'UCI [70]. Les variables quantitatives ont été discrétisées à l'aide Fusinter [71]. Pour toutes les bases, nous avons procédé de la même manière : l'ensemble des données a été divisé aléatoirement en deux parties. La première partie contient 30% des individus tout en gardant la répartition initiale des classes et nous servira pour la construction de variables. Les 70% restants sont utilisés pour les tests avant et après construction. Notre méthode de construction est basée sur la construction d'un arbre d'induction de type ID3. Pour cette raison, il nous paraît intéressant de tester l'efficacité de notre méthode en amont de différents arbres d'induction.

- ID3, [12], permet la construction de graphes d'induction arborescents à l'aide d'un critère basé sur le gain d'informations. Le seuil minimal du gain d'information a été fixé dans nos expérimentations à 0,05.
- C4.5, [134], est le descendant de ID3 et utilise comme critère le ratio de gain. C4.5 développe l'arbre au maximum et applique ensuite une procédure d'élagage qui supprime les sous-arbres ne vérifiant pas une condition liée au taux d'erreur.
- Sipina, [84] et [85], est une méthode conduisant à un graphe d'induction non arborescent. Sipina produit une succession de partitions par fusion et/ou éclatement des nœuds du graphe. Le nombre

minimum d'individus que doit posséder chaque sommet est fixé à 5 et le paramètre contrôlant le développement du graphe est fixé à 1.

- ChAID, [135], effectue des regroupements entre certaines modalités du prédicteur en vue construire des sous-arbres ayant un nombre optimal de branches. ChAID réalise ainsi des fusions entre sommets mais contrairement à Sipina qui généralise ce principe à tous les sommets terminaux de la partition, ChAID ne fusionne que les sommets terminaux issus d'un même père. Le seuil de segmentation est 0,00001 et le seuil de fusion est fixé à 0,05.
- Le modèle bayésien naïf, [86], utilise le théorème de Bayes pour estimer les probabilités a posteriori de toutes les classes. Pour chaque individu, la classe avec la probabilité a posteriori la plus élevée est choisi comme prédiction.

Nous avons utilisé différentes procédures de validation : un bootstrap, une 10-Cross-Validation et cinq 2-Cross-Validation.

3.2 Analyse de l'évaluation expérimentale

Les résultats des expérimentations sont regroupés dans les tableaux 52 à 67. Ces résultats sont également présentés sous forme graphique dans les figures 26 à 40.

Le tableau 52 permet d'évaluer le comportement général des diverses méthodes d'apprentissage testées lorsqu'elles sont associées à notre méthode de construction de variables. En effet, il présente la valeur moyenne du rapport « taux d'erreur avec construction/taux d'erreur sans construction » pour chaque algorithme d'apprentissage utilisé et dans le cadre soit d'une 10-Cross-Validation, soit de cinq 2-Cross-Validation, soit dans le cadre d'un bootstrap. La moyenne de ce rapport est calculée sur l'ensemble des quatorze jeux de données considérés. Les résultats permettent de conclure que, de manière générale, notre méthode de construction de variables impliquent l'obtention d'un taux d'erreur inférieur lorsque l'on rajoute à l'ensemble des variables les variables construites. Ainsi, quelle que soit la méthode d'apprentissage, les taux d'erreur obtenus sont corrects et souvent inférieurs lors de l'utilisation de notre méthode de construction. Toutefois, on peut noter que les moins bons résultats sont obtenus avec la méthode Sipina dans le cadre d'une 10 Cross-Validation.

Le tableau 51 nous présente les caractéristiques des différents jeux de données étudiés ainsi que la taille de la base de construction (c'est à dire le nombre de règles que nous utilisons lors de la construction des variables synthétiques).

Les tableaux 53 à 67 et les figures 26 à 40 permettent d'appréhender de manière plus précise les résultats obtenus. Les tableaux présentent, pour chaque jeu de données le taux d'erreur moyen en validation, l'écart type de ce taux d'erreur, l'écart et l'écart relatif existants entre les taux d'erreur avant et après construction ainsi que la différence qu'il existe entre les écarts type avant et après construction. Les figures se contentent de présenter le taux d'erreur moyen en validation.

L'étude de ces résultats nous permet de tirer un certain nombre de remarques qui sont les suivantes :

- La tendance générale d'un taux d'erreur plus faible après le processus de construction se vérifie localement pour la plupart des jeux de données,
- Quelque soit l'algorithme d'apprentissage, les résultats après construction sont très bons pour la base Monks-1. Ceci est du au concept sous-jacent aux données de Monks-1 (variable 1 = variable 2 ou variable 5 = 1).
- Les meilleurs résultats après construction sont obtenus avec la méthode d'apprentissage ChAID. La majorité des jeux de données connaissent une amélioration de leur taux de succès après le processus de construction et ce, quelle que soit la procédure de validation. Notre méthode paraît relativement efficace avec ChAID.
- Les résultats obtenus après la construction de nouvelles variables avec C4.5, ID3 et les Bayésiens Naïfs sont plutôt bons. Les meilleurs résultats sont obtenus pour les jeux de données : Diabetes, German, et Vehicle.
- Notre méthode de construction ne paraît pas vraiment adapté pour la méthode Sipina. En effet, les résultats obtenus après le processus de construction sont décevants. Pour un certain nombre de bases la qualité d'apprentissage se dégrade après l'application de notre méthode de construction : l'augmentation du taux d'erreur atteint 7,61 points pour la base Monks-2 avec une 10-Cross-Validation et 47,14% pour la base Iono avec cinq 2-Cross-Validation. Cependant nous avons remarquer que si nous contraignons Sipina de manière à ce qu'il n'effectue plus de fusion, alors les taux d'erreur près construction diminuent fortement. Ainsi, les variables créées par notre méthode posent des problèmes lorsque nous sommes en présence d'arbre d'induction comportant une procédure de fusion.
- Le calcul des différences entre écart type avant et après construction nous montre que la stabilité des apprentissages est quasiment similaire pour les apprentissages sur un même jeu de données quelque soit la méthode d'apprentissage employée et la procédure de validation utilisée.

- La qualité d'apprentissage sera d'autant plus élevée que la taille de la base de construction sera importante.
- Il convient cependant de noter une dégradation des résultats après le processus de construction de variables pour les bases Iris et Monks-3 dans de nombreux cas. Nous pouvons penser ici à un problème lié à l'over-fitting dû au fait que les nouvelles variables sont construites à l'aide d'ID3 libéré de sa contrainte de gain d'information minimum. Cependant, nous avons appliqué le processus de construction de variables à tous les jeux de données sans tenir compte du fait que ce processus n'est pas toujours nécessaire. Nous verrons dans le prochain chapitre que certaine base dont Monks-3 ne nécessitent pas de processus de construction lors de la phase de prétraitement.

En définitive, selon nous, cette étude expérimentale nous permet de conclure que notre méthode est relativement bien adaptée à différentes méthodes d'apprentissage que ces méthodes soient de type arbre d'induction ou les bayésiens naïfs. Cependant, elle n'est pas adaptée à des méthodes du type de Sipina qui comportent une procédure de fusion. L'application du processus permet de garder la stabilité de l'apprentissage constante. Le fait que ce sont presque toujours les mêmes jeux de données qui obtiennent des taux d'erreur plus faibles après le processus de construction nous permet de penser que la construction de variables n'est pas nécessaire et ni même efficace pour tous jeux de données.

Base	Taille	Taille échantillon pour la construction	Taille échantillon pour l'apprentissage	Nbre de Variables	Nbre de Classes	Taille base de construction
Austra	690	207	483	14	2	52
Breast	699	210	489	9	2	25
Cleve	303	91	212	13	2	32
CRX	690	207	483	15	2	58
German	1000	300	700	20	2	117
Heart	270	81	189	13	2	18
Iono	351	105	246	34	2	22
Iris	150	45	105	4	3	6
Monks-1	556	167	389	6	2	35
Monks-2	601	180	421	6	2	75
Monks-3	554	166	388	6	2	14
Pima	768	230	538	8	2	89
Tic Tac	958	287	671	9	2	93
Vehicle	846	254	592	18	4	116

Tableau 51 Caractéristiques des jeux de données.

	Bayésiens Naïfs	**C4.5**	**ChAID**	**Sipina**	**ID3**
10-Cross-Validation	0,9031	0,8920	0,8826	0,9563	0,9131
Cinq 2-Cross-Validation	0,9190	0,9237	0,8790	0,9362	0,9112
Bootstrap	0,9248	0,8634	0,8861	0,8835	0,9245

Tableau 52 Evaluation de notre méthode de construction de variables.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	16,60	4,57	16,19	5,03	-0,41	-2,47%	0,46
Breast	5,95	1,95	5,32	3,31	-0,63	-10,59%	1,36
Cleve	18,53	8,68	20,39	11,29	1,86	10,04%	2,61
CRX	14,73	5,68	13,92	5,86	-0,81	-5,50%	0,18
German	31,86	7,53	22,14	2,65	-9,72	-30,51%	-4,88
Heart	27,05	10,29	17,6	6,87	-9,45	-34,94%	-3,42
Iono	21,37	8,39	17,95	6,89	-3,42	-16,00%	-1,5
Iris	3,73	4,57	7,55	10,03	3,82	102,41%	5,46
Monks-1	25,22	8,3	1,03	1,7	-24,19	-95,92%	-6,6
Monks-2	34,91	6,79	34,92	7,33	0,01	0,03%	0,54
Monks-3	1,28	1,28	1,29	2,07	0,01	0,78%	0,79
Pima	26,11	5,43	26,11	3,85	0	0,00%	-1,58
Tic Tac Toe	33,43	5	15,97	4,77	-17,46	-52,23%	-0,23
Vehicle	34,24	4,96	35,76	4,58	1,52	4,44%	-0,38

Tableau 53 Evaluation de notre méthode de construction avec ID3 pour une 10-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	15,91	2,58	15,7	2,45	-0,21	-1,32%	-0,13
Breast	5,7	1,89	6,31	2,44	0,61	10,70%	0,55
Cleve	32,23	5,68	26,15	6,23	-6,08	-18,86%	0,55
CRX	14,66	2,43	14,66	2,84	0	0,00%	0,41
German	28,57	4,58	25,14	3,79	-3,43	-12,01%	-0,79
Heart	28,95	4,71	24,74	3,16	-4,21	-14,54%	-1,55
Iono	13,39	3,62	12,56	5,96	-0,83	-6,20%	2,34
Iris	3,81	4,67	3,77	3,55	-0,04	-1,05%	-1,12
Monks-1	25,19	5,7	1,03	0,96	-24,16	-95,91%	-4,74
Monks-2	34,92	3,57	34,89	9,77	-0,03	-0,09%	6,2
Monks-3	1,29	0,81	1,29	1,15	0	0,00%	0,34
Pima	24,3	2,48	25,59	3,22	1,29	5,31%	0,74
Tic Tac Toe	22,8	3,94	22,46	4,4	-0,34	-1,49%	0,46
Vehicle	29,41	3,49	30,08	4,49	0,67	2,28%	1

Tableau 54 Evaluation de notre méthode de construction avec ID3 pour cinq 2-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	16,24	2,64	16,18	3,12	-0,06	-0,37%	0,48
Breast	5,19	1,57	5,36	2,87	0,17	3,28%	1,3
Cleve	24,38	5	23,75	3,87	-0,63	-2,58%	-1,13
CRX	16,16	2,95	17,22	2,92	1,06	6,56%	-0,03
German	29,15	3,53	25,11	2,88	-4,04	-13,86%	-0,65
Heart	25,33	4,09	23,15	4,92	-2,18	-8,61%	0,83
Iono	10,63	3,09	11,41	4,34	0,78	7,34%	1,25
Iris	3,29	2,41	3,84	3,11	0,55	16,72%	0,7
Monks-1	23,37	5,02	1,02	2,31	-22,35	-95,64%	-2,71
Monks-2	36,45	3,93	36,26	3,73	-0,19	-0,52%	-0,2
Monks-3	1,39	0,9	1,34	0,78	-0,05	-3,60%	-0,12
Pima	25,68	2,52	23,9	3,13	-1,78	-6,93%	0,61
Tic Tac Toe	20,26	4,32	19,41	3,1	-0,85	-4,20%	-1,22
Vehicle	30,16	3,44	27,34	3,46	-2,82	-9,35%	0,02

Tableau 55 Evaluation de notre méthode de construction avec ID3 pour un bootstrap (20 réplications).

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	15,87	5,26	13,83	2,69	-2,04	-12,85%	-2,57
Breast	8,13	4,14	4,68	2,24	-3,45	-42,44%	-1,9
Cleve	24,7	11,39	21,06	9,35	-3,64	-14,74%	-2,04
CRX	15,06	3,52	15,48	5,42	0,42	2,79%	1,9
German	28	5,47	28,43	5,21	0,43	1,54%	-0,26
Heart	23,68	9,78	28,95	9,19	5,27	22,26%	-0,59
Iono	9,7	5,98	8,17	7,16	-1,53	-15,77%	1,18
Iris	4,64	6,17	5,45	6,03	0,81	17,46%	-0,14
Monks-1	13,64	4,53	0,52	1,04	-13,12	-96,19%	-3,49
Monks-2	41,12	8,08	39,9	6,7	-1,22	-2,97%	-1,38
Monks-3	1,29	1,29	1,55	1,26	0,26	20,16%	-0,03
Pima	24,51	5,33	22,08	4,2	-2,43	-9,91%	-1,13
Tic Tac Toe	15,93	3,85	14,88	4,25	-1,05	-6,59%	0,4
Vehicle	31,42	6,92	24,55	6,36	-6,87	-21,87%	-0,56

Tableau 56 Evaluation de notre méthode de construction avec C4.5 pour une 10-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	14,47	3,18	16,12	2,41	1,65	11,40%	-0,77
Breast	6,11	2,12	5,5	1,99	-0,61	-9,98%	-0,13
Cleve	23,81	3,89	24,77	3,52	0,96	4,03%	-0,37
CRX	13,65	3,22	13,63	2,27	-0,02	-0,15%	-0,95
German	30,57	5,91	27,43	3,88	-3,14	-10,27%	-2,03
Heart	24,74	6,53	25,79	5,1	1,05	4,24%	-1,43
Iono	11,36	7,11	9,69	3,62	-1,67	-14,70%	-3,49
Iris	4,68	3,01	4,63	2,88	-0,05	-1,07%	-0,13
Monks-1	15,39	7,64	0,52	0,63	-14,87	-96,62%	-7,01
Monks-2	37,52	2,64	37,53	2,24	0,01	0,03%	-0,4
Monks-3	1,29	1,15	1,55	1,25	0,26	20,16%	0,1
Pima	24,31	3,11	22,07	3,07	-2,24	-9,21%	-0,04
Tic Tac Toe	15,18	2,76	16,22	2,7	1,04	6,85%	-0,06
Vehicle	32,77	5,04	26,72	4,77	-6,05	-18,46%	-0,27

Tableau 57 Evaluation de notre méthode de construction avec C4.5 pour cinq 2-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	14,59	2,39	13,97	2,29	-0,62	-4,25%	-0,1
Breast	5,47	2,27	4,73	1,53	-0,74	-13,53%	-0,74
Cleve	20,84	3,6	19,89	4,07	-0,95	-4,56%	0,47
CRX	13,94	2,32	14,8	2,54	0,86	6,17%	0,22
German	26,48	2,33	23,48	2,27	-3	-11,33%	-0,06
Heart	20,57	5,59	20,27	5,77	-0,3	-1,46%	0,18
Iono	10,8	3,02	8,88	3,49	-1,92	-17,78%	0,47
Iris	3,68	2,75	3,45	3,1	-0,23	-6,25%	0,35
Monks-1	12,06	6,9	0,34	0,65	-11,72	-97,18%	-6,25
Monks-2	32,71	2,96	31,52	3,92	-1,19	-3,64%	0,96
Monks-3	1,84	1,39	1,46	0,71	-0,38	-20,65%	-0,68
Pima	22,15	2,33	21,28	2,43	-0,87	-3,93%	0,1
Tic Tac Toe	14,44	3,21	12,89	2,15	-1,55	-10,73%	-1,06
Vehicle	27,67	3,96	24,61	2,99	-3,06	-11,06%	-0,97

Tableau 58 Evaluation de notre méthode de construction avec C4.5 pour un Bootstrap (20 réplications).

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	16,73	3,95	16,93	5,17	0,2	1,20%	1,22
Breast	7,13	2,29	3,88	2,32	-3,25	-45,58%	0,03
Cleve	21,47	8,57	26,8	10,02	5,33	24,83%	1,45
CRX	16,3	6,22	18,41	4,48	2,11	12,94%	-1,74
German	28,14	5,5	27,86	2,95	-0,28	-1,00%	-2,55
Heart	23,16	10,04	23,68	8,64	0,52	2,25%	-1,4
Iono	7,73	6,95	10,22	6,55	2,49	32,21%	-0,4
Iris	4,64	6,17	5,55	6,09	0,91	19,61%	-0,08
Monks-1	20,11	4,89	5,91	3,04	-14,2	-70,61%	-1,85
Monks-2	38,24	7	45,85	6,5	7,61	19,90%	-0,5
Monks-3	1,79	2,58	1,29	1,29	-0,5	-27,93%	-1,29
Pima	24,3	4,46	24,3	4,32	0	0,00%	-0,14
Tic Tac Toe	20,67	3,77	20,08	3,47	-0,59	-2,85%	-0,3
Vehicle	47,26	6,24	22,7	5,47	-24,56	-51,97%	-0,77

Tableau 59 Evaluation de notre méthode de construction avec Sipina pour une 10-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	17,36	1,58	15,29	1,48	-2,07	-11,92%	-0,1
Breast	7,13	2,27	5,69	2,96	-1,44	-20,20%	0,69
Cleve	21,46	8,43	25,66	8,1	4,2	19,57%	-0,33
CRX	16,31	3,96	17,35	3,75	1,04	6,38%	-0,21
German	29,29	4,74	27,71	3,87	-1,58	-5,39%	-0,87
Heart	21,58	4,53	26,32	8,32	4,74	21,96%	3,79
Iono	7,7	4,35	11,33	4,72	3,63	47,14%	0,37
Iris	4,68	3,01	5,54	4,44	0,86	18,38%	1,43
Monks-1	19,76	11,8	5,91	2,64	-13,85	-70,09%	-9,16
Monks-2	42,28	1,19	40,85	3,95	-1,43	-3,38%	2,76
Monks-3	2,31	2,2	1,29	1,15	-1,02	-44,16%	-1,05
Pima	23,19	3,29	23,18	5,86	-0,01	-0,04%	2,57
Tic Tac Toe	24,12	4,26	22,31	4,38	-1,81	-7,50%	0,12
Vehicle	47,56	1,96	23,87	3,26	-23,69	-49,81%	1,3

Tableau 60 Evaluation de notre méthode de construction avec Sipina pour cinq 2-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	14,48	1,83	15,13	1,89	0,65	4,49%	0,06
Breast	6,15	1,32	5,32	1,85	-0,83	-13,50%	0,53
Cleve	23,6	6,5	21,08	3,98	-2,52	-10,68%	-2,52
CRX	14,61	2,23	15,27	2,35	0,66	4,52%	0,12
German	27,22	2,86	24,71	2,73	-2,51	-9,22%	-0,13
Heart	20,69	6,19	18,86	4,13	-1,83	-8,84%	-2,06
Iono	8,43	2,57	9,79	3,62	1,36	16,13%	1,05
Iris	4,29	1,87	4,84	4,35	0,55	12,82%	2,48
Monks-1	13,51	8,19	5,54	2,32	-7,97	-58,99%	-5,87
Monks-2	34,31	4,58	34,55	3,17	0,24	0,70%	-1,41
Monks-3	2,63	2,75	1,31	0,76	-1,32	-50,19%	-1,99
Pima	21,68	2,53	21,13	2,4	-0,55	-2,54%	-0,13
Tic Tac Toe	19,15	3,89	17,45	2,84	-1,7	-8,88%	-1,05
Vehicle	46,31	3,02	24,52	3,2	-21,79	-47,05%	0,18

Tableau 61 Evaluation de notre méthode de construction avec Sipina pour un Bootstrap (20 réplications).

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	15,29	2,66	15,66	6,59	0,37	2,42%	3,93
Breast	5,48	3,92	4,47	3,33	-1,01	-18,43%	-0,59
Cleve	31,75	12,61	32,68	8,36	0,93	2,93%	-4,25
CRX	16,51	6,4	16,9	5,18	0,39	2,36%	-1,22
German	31,86	5,79	25	5,2	-6,86	-21,53%	-0,59
Heart	23,68	10,33	22,11	7,74	-1,57	-6,63%	-2,59
Iono	8,72	4,34	8,12	3,17	-0,6	-6,88%	-1,17
Iris	4,64	4,64	4,64	4,64	0	0,00%	0
Monks-1	25,2	5,28	6,17	2,04	-19,03	-75,52%	-3,24
Monks-2	34,93	5,52	34,91	4,54	-0,02	-0,06%	-0,98
Monks-3	1,29	1,73	1,29	1,73	0	0,00%	0
Pima	25,41	4,51	24,5	4,09	-0,91	-3,58%	-0,42
Tic Tac Toe	28,12	2,92	21,14	2,77	-6,98	-24,82%	-0,15
Vehicle	30,58	5,07	27,21	5,25	-3,37	-11,02%	0,18

Tableau 62 Evaluation de notre méthode de construction avec ChAID pour une 10-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	15,3	2,72	15,29	2,21	-0,01	-0,07%	-0,51
Breast	4,48	1,5	4,28	0,75	-0,2	-4,46%	-0,75
Cleve	31,76	2,17	29,45	2,44	-2,31	-7,27%	0,27
CRX	16,94	2,78	16,51	5,35	-0,43	-2,54%	2,57
German	30	3,44	24,14	1,46	-5,86	-19,53%	-1,98
Heart	27,37	6,98	23,68	4,99	-3,69	-13,48%	-1,99
Iono	10,54	3,52	8,91	2,76	-1,63	-15,46%	-0,76
Iris	5,58	6,84	5,58	3,49	0	0,00%	-3,35
Monks-1	25,17	5,46	5,92	2,29	-19,25	-76,48%	-3,17
Monks-2	34,91	5,2	34,9	4,63	-0,01	-0,03%	-0,57
Monks-3	1,29	1,4	1,29	1,42	0	0,00%	0,02
Pima	25,61	1,94	23,93	4,6	-1,68	-6,56%	2,66
Tic Tac Toe	28,12	4,71	22,18	3,12	-5,94	-21,12%	-1,59
Vehicle	30,59	4,14	29,24	4,55	-1,35	-4,41%	0,41

Tableau 63 Evaluation de notre méthode de construction avec ChAID pour cinq 2-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	15,93	2,72	16,05	2,58	0,12	0,75%	-0,14
Breast	4,99	1,28	4,67	1,29	-0,32	-6,41%	0,01
Cleve	25,76	2,79	27,38	5,47	1,62	6,29%	2,68
CRX	16,3	1,8	16	2,35	-0,3	-1,84%	0,55
German	29,79	2,78	24,82	2,52	-4,97	-16,68%	-0,26
Heart	25,29	4,8	24,4	4,24	-0,89	-3,52%	-0,56
Iono	8,28	3,06	7,17	3,36	-1,11	-13,41%	0,3
Iris	4,16	3,99	4,17	2,9	0,01	0,24%	-1,09
Monks-1	25,62	2,95	5,76	1,92	-19,86	-77,52%	-1,03
Monks-2	35,57	2,8	35,83	4,22	0,26	0,73%	1,42
Monks-3	1,52	1,4	1,27	0,87	-0,25	-16,45%	-0,53
Pima	25,1	2,85	24,01	2,31	-1,09	-4,34%	-0,54
Tic Tac Toe	28,08	3,7	21,91	1,82	-6,17	-21,97%	-1,88
Vehicle	30,74	3,59	26,21	3,26	-4,53	-14,74%	-0,33

Tableau 64 Evaluation de notre méthode de construction avec ChAID pour un Bootstrap (20 réplications).

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	14,26	4,58	12,79	3,5	-1,47	-10,31%	-1,08
Breast	2,65	1,31	3,87	3,59	1,22	46,04%	2,28
Cleve	21	6,63	19,18	6,13	-1,82	-8,67%	-0,5
CRX	14,67	3,14	17,35	5,42	2,68	18,27%	2,28
German	23,71	6,58	23	4,4	-0,71	-2,99%	-2,18
Heart	17,37	7,46	18,42	8,89	1,05	6,04%	1,43
Iono	6,83	5,06	6,07	3,23	-0,76	-11,13%	-1,83
Iris	6,45	7,14	7,45	9,14	1	15,50%	2
Monks-1	25,22	6	5,93	4,04	-19,29	-76,49%	-1,96
Monks-2	38,94	4,14	41,31	7,06	2,37	6,09%	2,92
Monks-3	3,88	2,9	1,28	1,72	-2,6	-67,01%	-1,18
Pima	21,14	5,42	20,59	3,99	-0,55	-2,60%	-1,43
Tic Tac Toe	29,61	5,15	21,28	4,17	-8,33	-28,13%	-0,98
Vehicle	34,27	5,52	26,39	6,5	-7,88	-22,99%	0,98

Tableau 65 Evaluation de notre méthode de construction avec le modèle de bayésien naïf pour une 10-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	14,47	3,52	13,02	2,13	-1,45	-10,02%	-1,39
Breast	2,86	1,98	4,08	1,45	1,22	42,66%	-0,53
Cleve	20,08	2,25	19,67	5,02	-0,41	-2,04%	2,77
CRX	15,09	3,37	16,32	4,23	1,23	8,15%	0,86
German	23,86	3,9	22,71	3,36	-1,15	-4,82%	-0,54
Heart	17,37	6,36	17,86	5,1	0,49	2,82%	-1,26
Iono	7,71	2,42	7,68	2,33	-0,03	-0,39%	-0,09
Iris	4,72	4,26	7,49	2,32	2,77	58,69%	-1,94
Monks-1	25,69	6,98	5,91	1,31	-19,78	-76,99%	-5,67
Monks-2	38,95	3,81	39,43	4,8	0,48	1,23%	0,99
Monks-3	3,87	1,64	1,3	1,16	-2,57	-66,41%	-0,48
Pima	21,89	4,07	20,59	2,6	-1,3	-5,94%	-1,47
Tic Tac Toe	30,34	3,82	21,27	2,81	-9,07	-29,89%	-1,01
Vehicle	36,3	1,24	27,56	4,03	-8,74	-24,08%	2,79

Tableau 66 Evaluation de notre méthode de construction avec le modèle de bayésien naïf pour cinq 2-Cross-Validation.

Base	Avant Construction		Après Construction		Ecart entre taux d'erreur	Ecart relatif entre taux d'erreur	Différence entre Ecarts type
	Erreur	Ecart Type	Erreur	Ecart Type			
Austra	14,73	2,57	12,87	1,85	-1,86	-12,63%	-0,72
Breast	2,76	1,21	3,92	0,98	1,16	42,03%	-0,23
Cleve	19,98	3,58	20,21	2,69	0,23	1,15%	-0,89
CRX	13,92	1,91	15,9	2,61	1,98	14,22%	0,7
German	22,73	1,38	21,49	1,47	-1,24	-5,46%	0,09
Heart	16,85	4,23	18,48	3,63	1,63	9,67%	-0,6
Iono	7,04	3,5	6,56	3	-0,48	-6,82%	-0,5
Iris	4,77	3,18	6,86	3,18	2,09	43,82%	0
Monks-1	26,49	2,9	5,91	1,5	-20,58	-77,69%	-1,4
Monks-2	38,51	2,96	38,79	3,08	0,28	0,73%	0,12
Monks-3	3,58	1,61	1,28	0,73	-2,3	-64,25%	-0,88
Pima	21,23	2,4	20,42	1,83	-0,81	-3,82%	-0,57
Tic Tac Toe	27,88	2,98	21	2,5	-6,88	-24,68%	-0,48
Vehicle	31,49	2,38	25,75	2,99	-5,74	-18,23%	0,61

Tableau 67 Evaluation de notre méthode de construction avec le modèle de bayésien naïf pour un Bootsrap (20 réplications).

Pour l'ensemble des graphiques qui suivent la légende est la suivante :

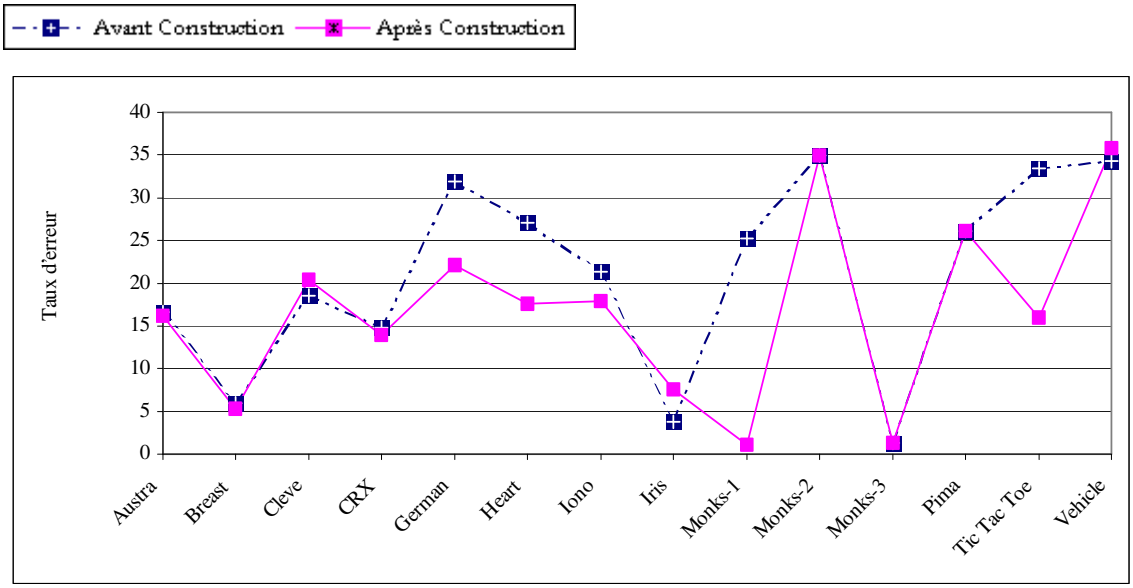

Figure 26 Taux d'erreur avec une 10 Cross-Valisation avec ID3.

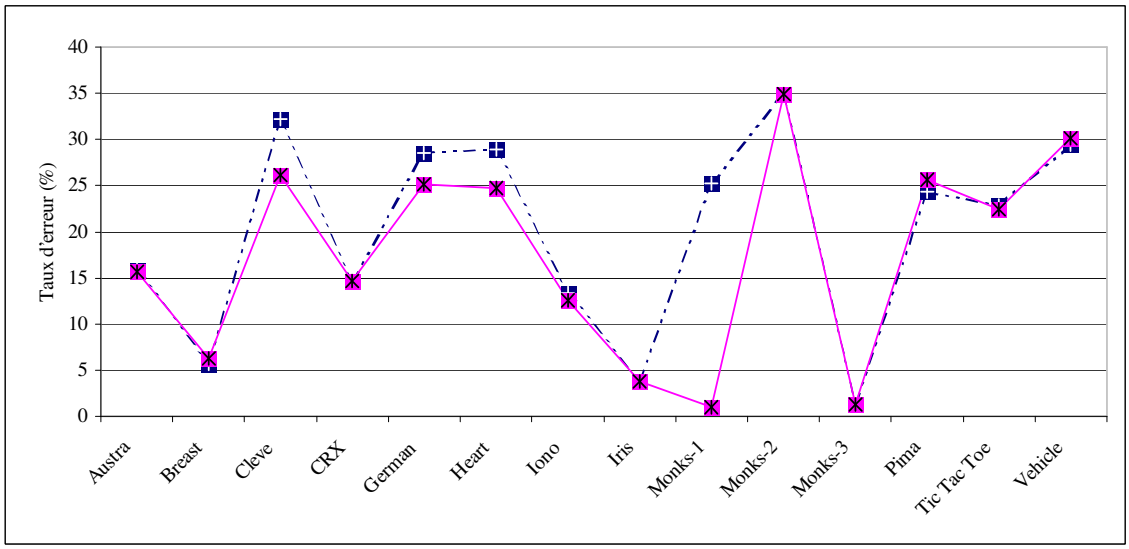

Figure 27 Taux d'erreur pour cinq 2 Cross-Validation avec ID3

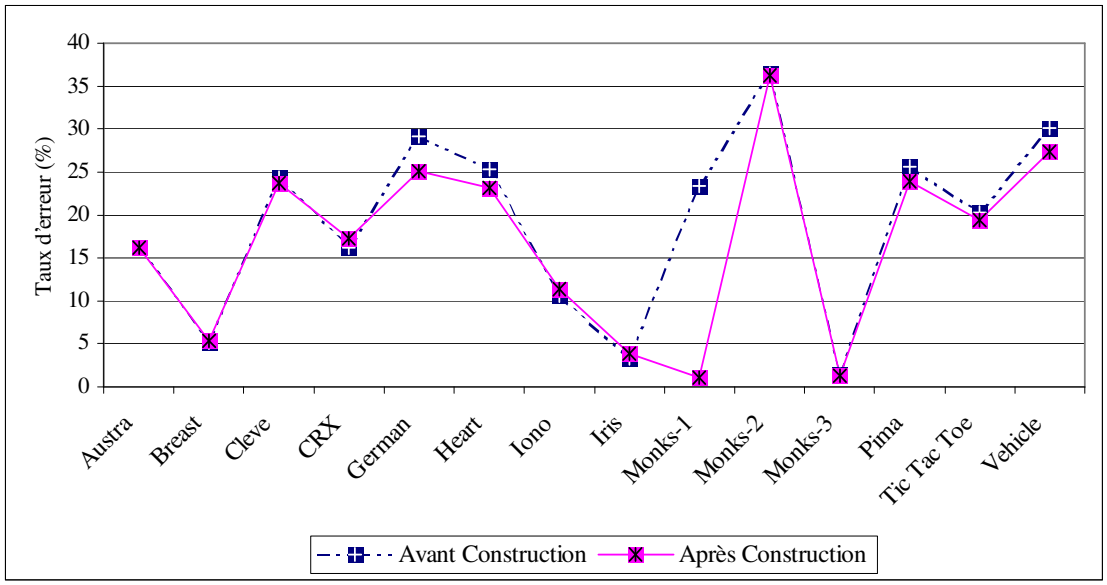

Figure 28 Taux d'erreur pour un Bootstrap avec ID3.

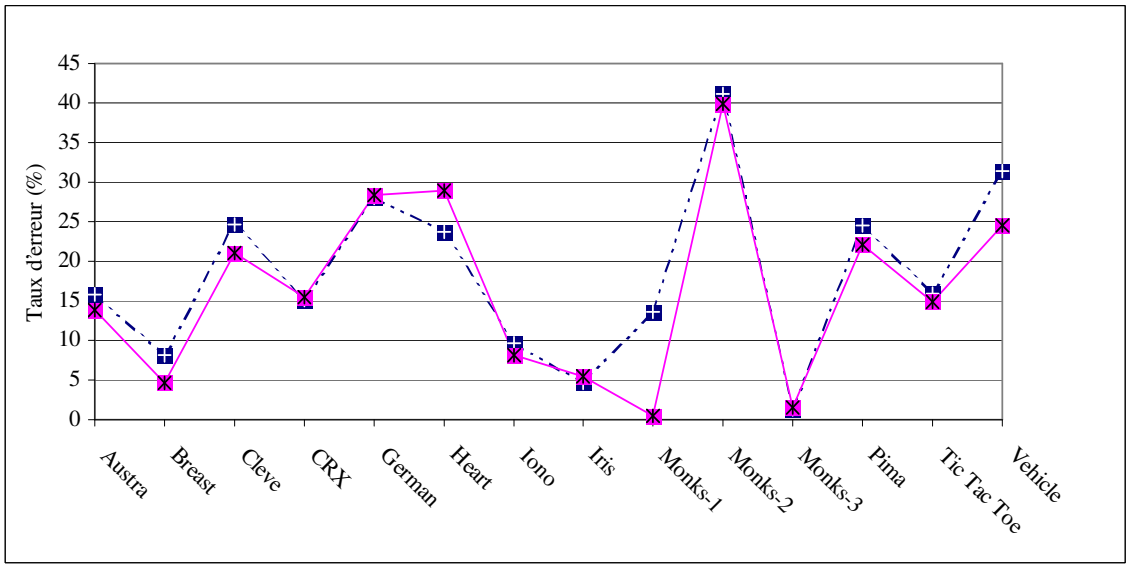

Figure 29 Taux d'erreur pour une 10 Cross-Validation avec C4.5

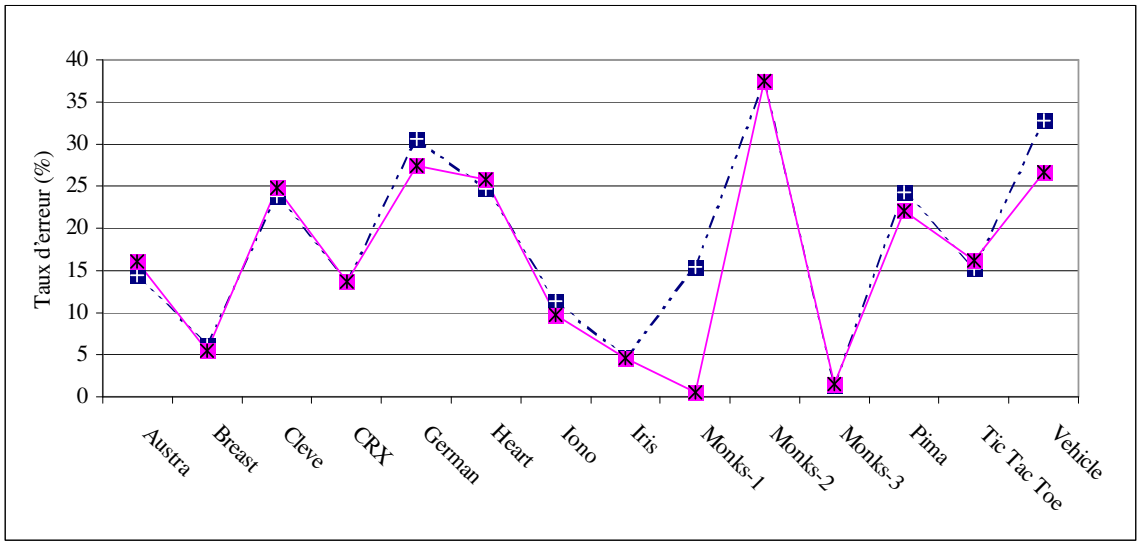

Figure 30 Moyenne du taux d'erreur pour cinq 2 Cross-Validation avec C4.5

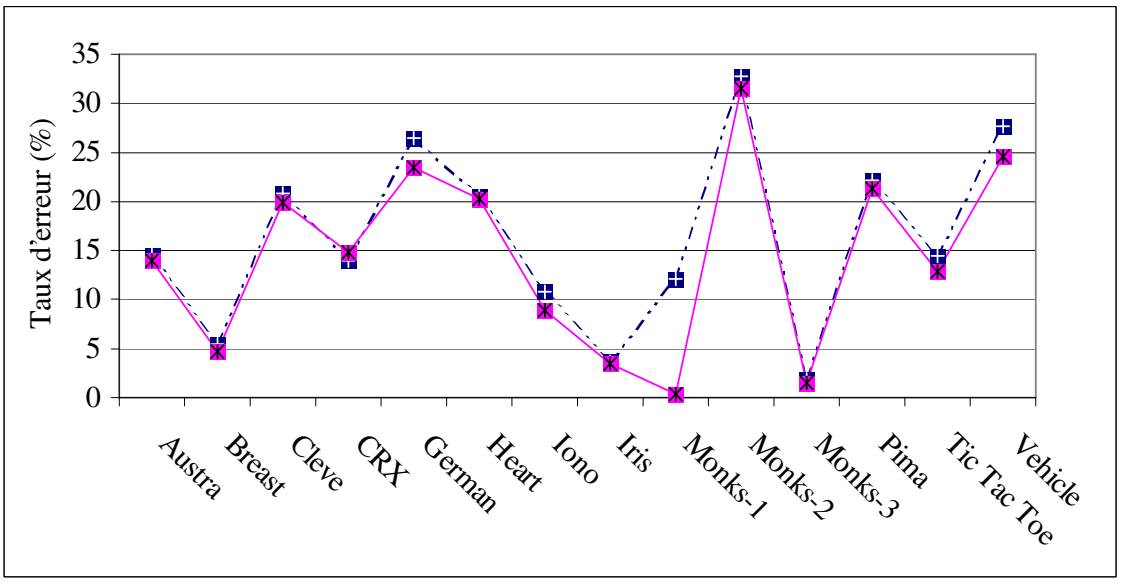

Figure 31 Moyenne du taux d'erreur pour un Bootstrap avec C4.5.

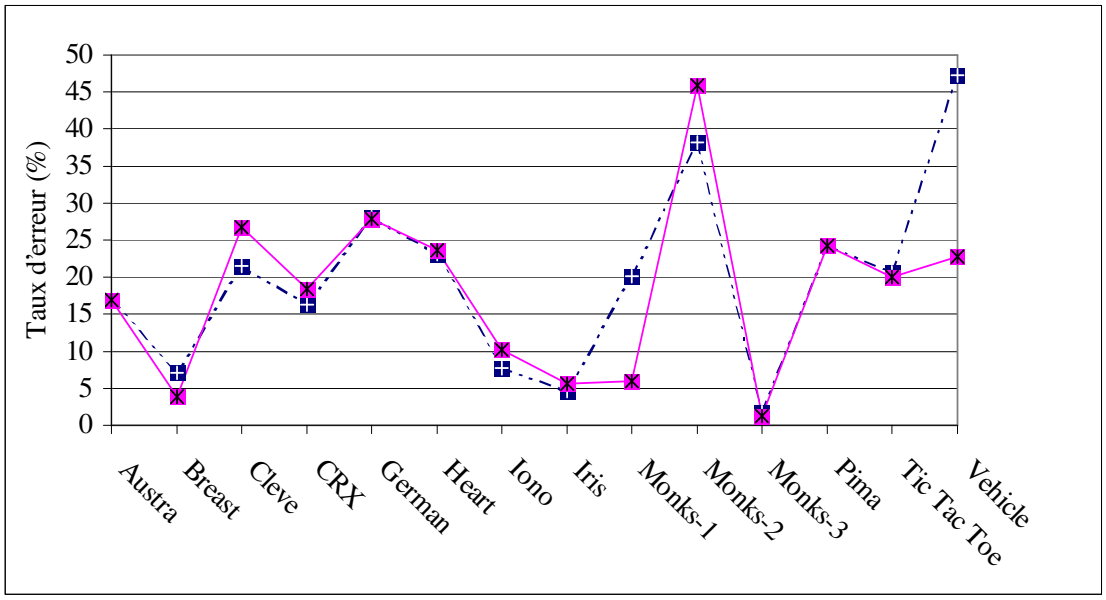

Figure 32 Moyenne du taux d'erreur pour une 10 Cross-Validation avec Sipina.

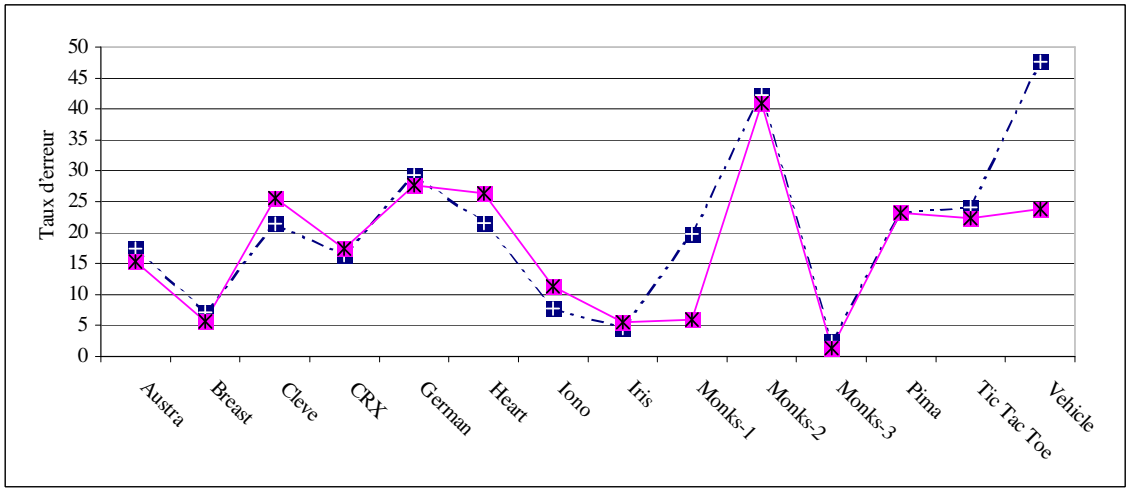

Figure 33 Moyenne du taux d'erreur pour cinq 2 Cross-Validation avec Sipina

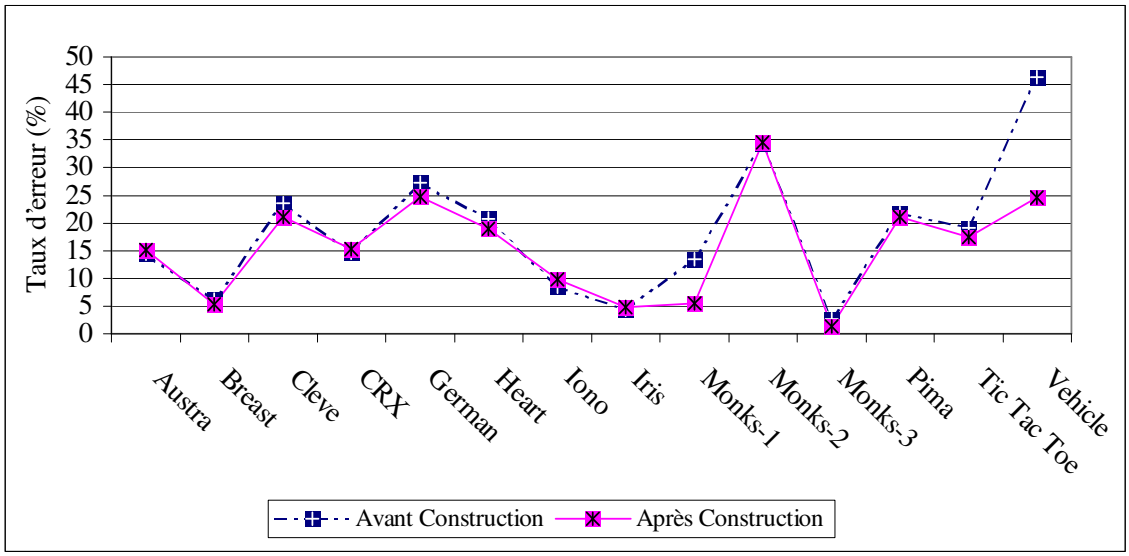

Figure 34 Moyenne du taux d'erreur pour un Bootstrap avec Sipina.

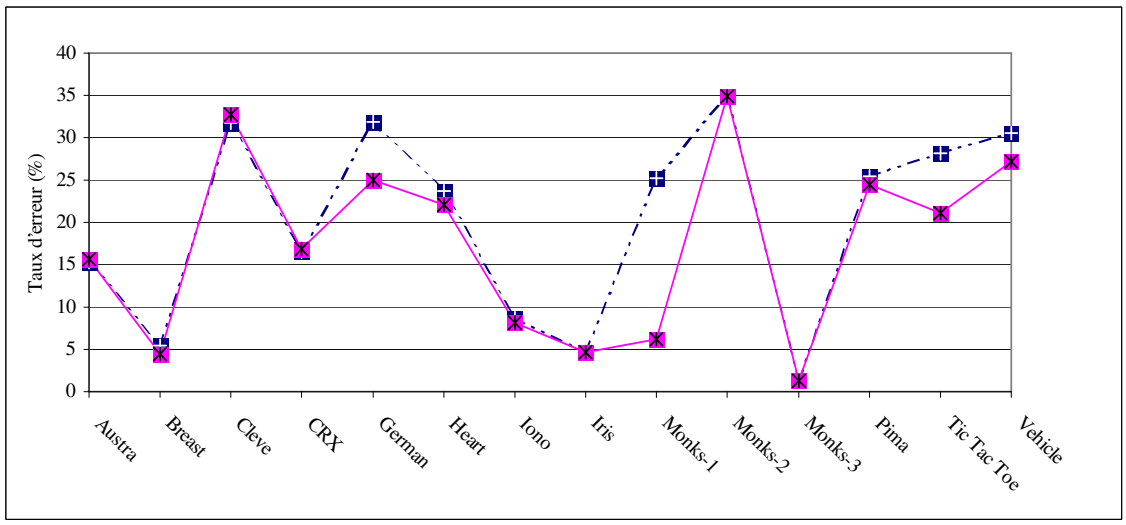

Figure 35 Moyenne du taux d'erreur pour une 10 Cross-Validation avec CHAID.

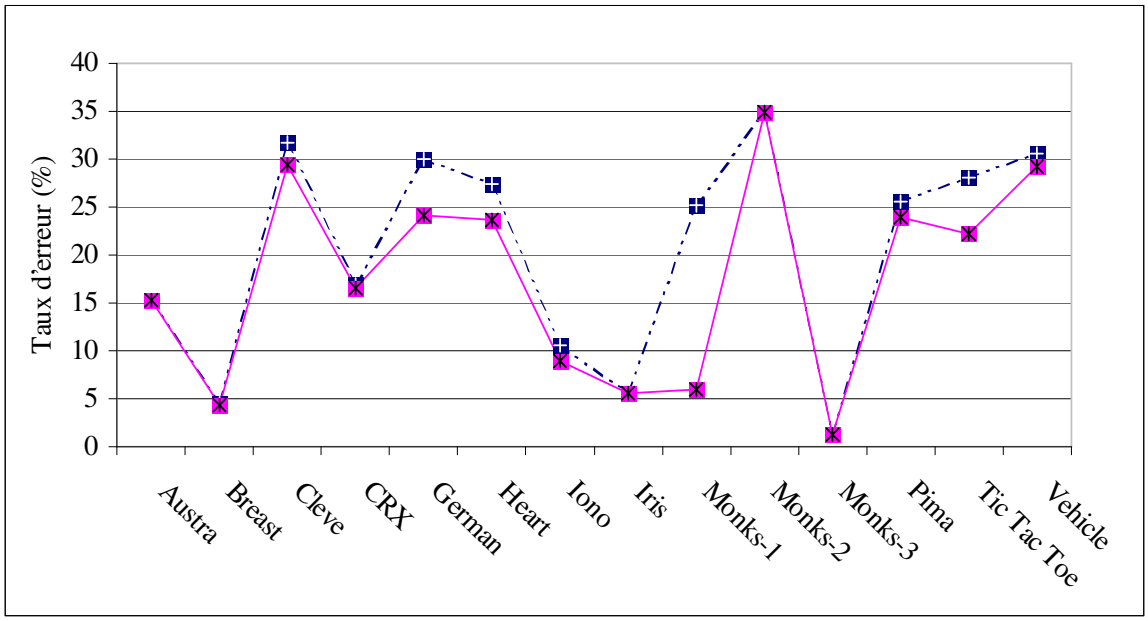

Figure 36 Moyenne du taux d'erreur pour cinq 2 Cross-Validation avec CHAID

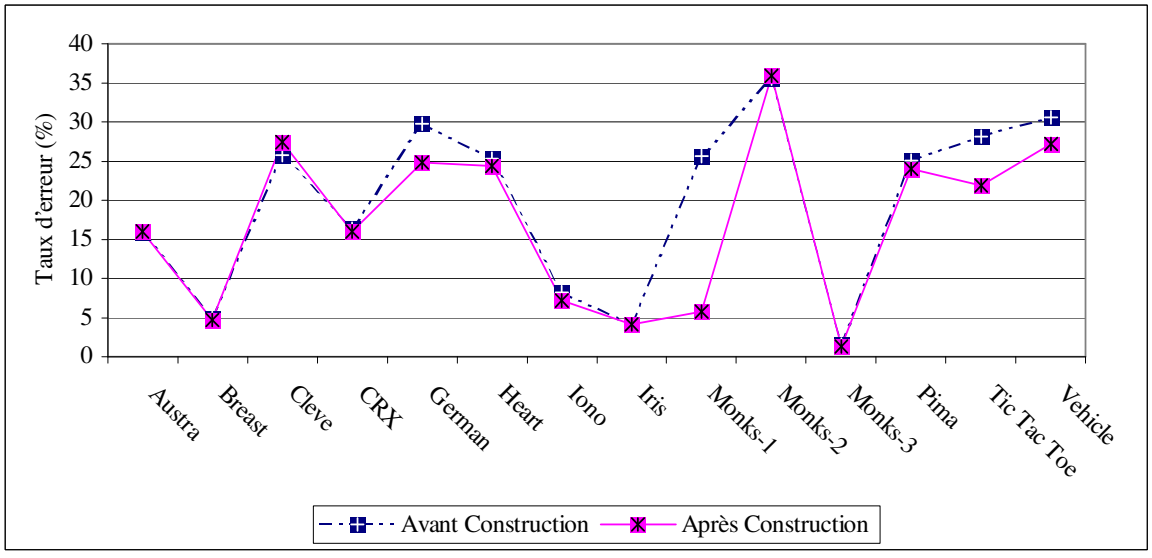

Figure 37 Moyenne du taux d'erreur pour un Bootstrap avec CHAID.

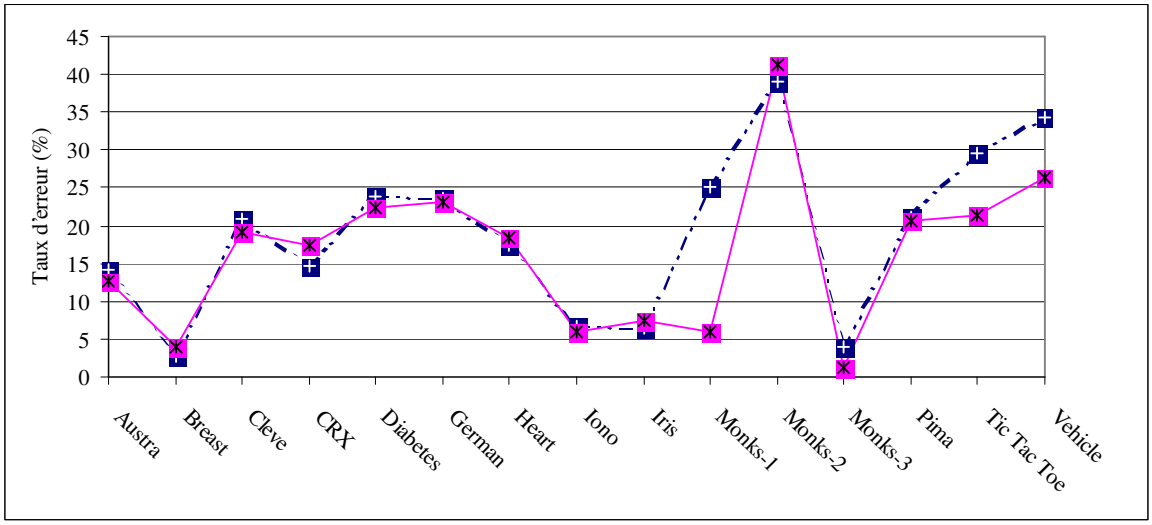

Figure 38 Moyenne du taux d'erreur pour une 10 Cross-Validation avec les Bayésiens Naïfs.

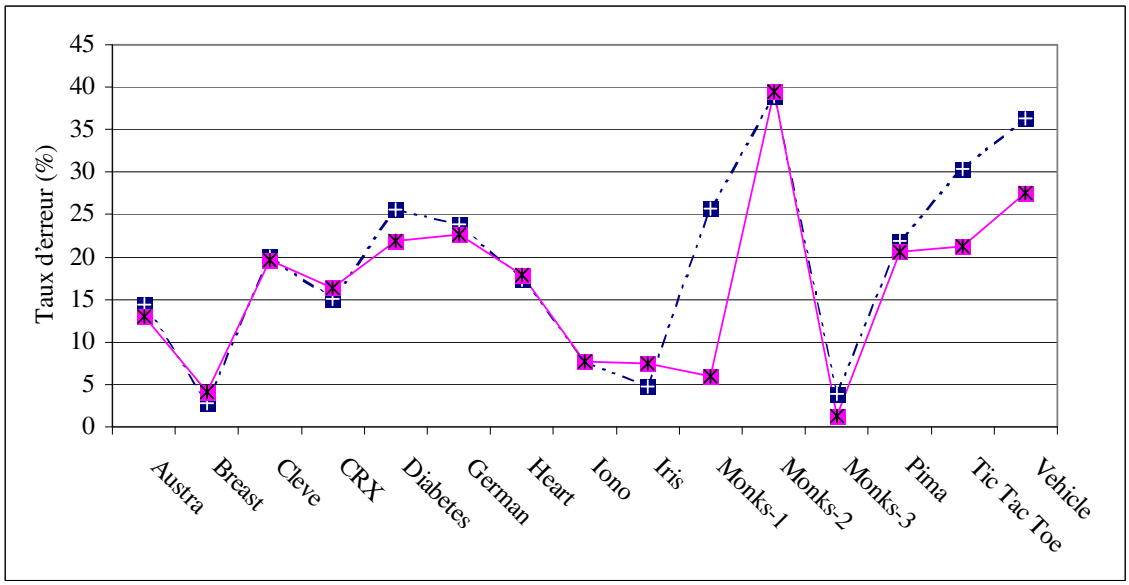

Figure 39 Moyenne du taux d'erreur pour cinq 2 Cross-Validation avec les Bayésiens Naïfs.

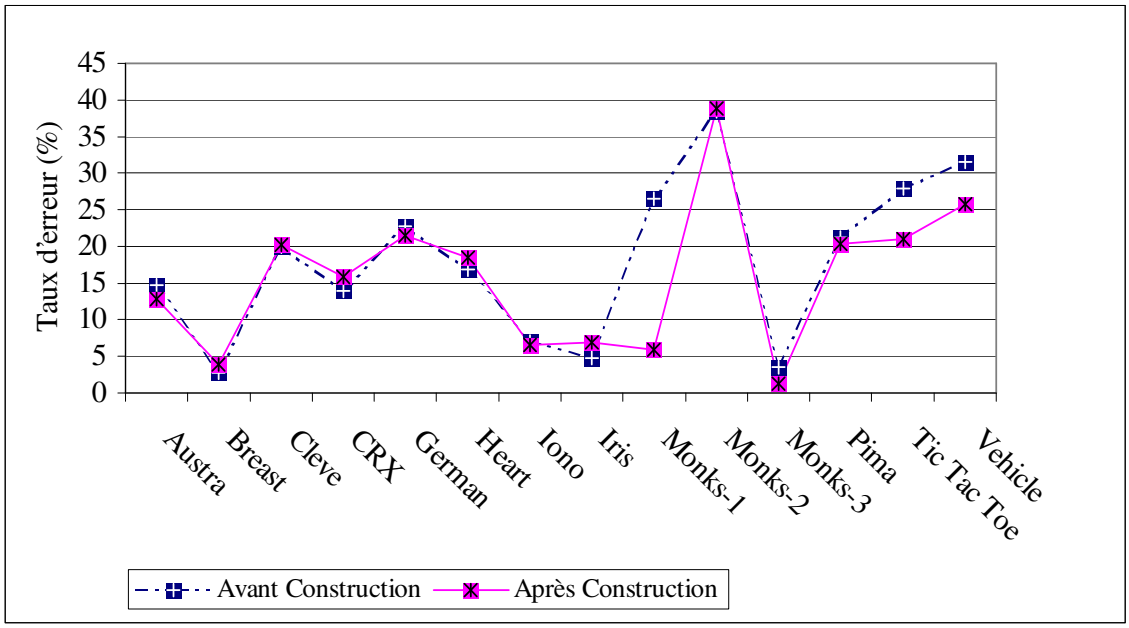

Figure 40 Moyenne du taux d'erreur pour un Bootstrap avec les Bayésiens Naïfs.

4 Conclusion

La méthode que nous proposons est une méthode de construction par analyse topologique des arbres. Elle ne traite que les variables qualitatives. Elle utilise comme opérateurs de construction les opérateurs booléens ET et OU. Le nombre de variables qu'elle crée est connu avant l'application du processus de construction : Il est égal au nombre de classes de la variable endogène. Donc peu de variables synthétiques sont construites, ce qui permet de ne pas surcharger l'espace de représentation des données.

Catégorie de méthode	Méthode par analyse topologique des arbres
Type de variables traitées	Variables qualitatives
Opérateurs de construction	ET et OU
Nombre de variables créées	Correspond au nombre de classes de la variable endogène
Type des variables créées	Variables booléennes

Tableau 68 Caractéristiques de notre méthode de construction de variables.

C'est une méthode qui prend en compte les liens existants entre les variables et ce par le fait que les variables construites sont des conjonctions de modalités de différentes variables. Les liens découverts

par notre méthode se retrouvent lors de l'étude du même problème par l'AFCM. Notre méthode fonctionne en trois étapes :

- La première étape consiste en la génération des règles qui vont former la base de construction c'est à dire qui seront utilisées pour construire les nouvelles variables. Ces règles sont générées à l'aide de l'arbre non contraint.
- La deuxième étape consiste en la création de variables intermédiaires par la conjonction des modalités constituant chaque règles. Chaque variable intermédiaire est associée à une règle. Nous obtenons ainsi autant de variables intermédiaires que de règles. Ensuite, les variables intermédiaires sont regroupées en fonction de la conclusion de la règle qui leur est associée.
- La dernière étape consiste en la construction des nouvelles variables. Ces dernières sont la disjonction des variables intermédiaires au sein d'un même groupe.

Notre méthode permet de conserver la stabilité du modèle et d'améliorer la qualité d'apprentissage. Elle est relativement efficace avec des arbres tels que ChAID, C4.5 et ID3 ou avec les bayésiens naïfs.

Il nous semble intéressant à l'avenir de permettre à l'utilisateur d'agir sur le choix des règles formant la base de construction : un expert du domaine peut nous permettre de mieux nous adapter au problème traité en pondérant, par exemple, les règles qu'il pense être essentielles.

Il est également indispensable de détecter le moment où il est utile d'appliquer le processus de construction afin de ne pas dégrader la qualité d'apprentissage par la construction de nouvelles variables. Le chapitre suivant propose un indice permettant de détecter ce moment.