Chapitre 4 Gestion de la phase de prétraitement

La sélection et la construction de variables permettent un prétraitement des données. Ce prétraitement est relativement important. C'est cette phase qui conditionne la qualité des modèles établis en fouille de données et qui permet de faire émerger l'information contenue au sein des données. La sélection de variables permet de supprimer à la fois le bruit engendré par certaines variables et les variables redondantes. Ainsi, après un processus de sélection, seules les variables apportant de l'information pertinente sont conservées, et la taille de l'espace de représentation des données est réduite. La construction de variables permet de re-décrire les données d'entrée du problème d'apprentissage grâce à la création de variables synthétiques. Ainsi, après le processus de construction, la taille de l'espace de représentation des données augmente et les données peuvent être étudiées sous un autre angle.

Une question reste, cependant en suspens : quand l'utilisation de la sélection et/ou de la construction est elle nécessaire ? La réponse à cette question dépend de l'étude de la structure des données. Il est possible de distinguer quatre cas de structures de données :

- Toutes les variables présentes sont nécessaires pour discriminer la variable endogène mais leur présence n'est pas suffisante. Il manque de l'information et seule la construction de variables peut palier à ce manque.
- L'information nécessaire pour discriminer la variable endogène est contenue dans les variables initiales. Mais parmi ces variables certaines sont soit redondantes soit du bruit et empêchent les autres variables à discriminer au mieux la variable endogène. Pour éliminer ces variables indésirables, la sélection de variables est suffisante.
- Une partie des variables permet de discriminer la variable endogène mais leur présence n'est pas suffisante. Les variables restantes sont soit redondantes soit du bruit. A ce niveau, les processus de construction et de sélection sont nécessaires conjointement.
- La variable endogène est parfaitement discriminée par les variables initiales. Ces dernières sont nécessaires et suffisantes, les processus de construction et de sélection sont inutiles dans ce cas. C'est la situation la plus favorable, cependant il est très rare qu'elle se produise.

Nous pensons qu'il est plus efficace d'appliquer le processus de sélection de variables dans tous les cas. En effet, la sélection permet de réduire la taille de l'espace de représentation des données tout en améliorant la qualité d'apprentissage ou, dans le pire des cas en la gardant constante grâce à la suppression des variables bruitées, redondantes et/ou non pertinentes. De plus, si l'ensemble des variables sont pertinentes, situation très rare, alors la méthode de sélection sélectionnera la totalité des variables. La question initiale se transforme alors en la question suivante : est il nécessaire d'appliquer le processus de construction de variables ? La construction de nouvelles variables permettra t'elle d'améliorer la qualité d'apprentissage ?

Nous proposons d'utiliser le coefficient Kappa de Cohen, [136], qui permet de chiffrer la qualité de l'accord réel entre des jugements qualitatifs appariés. Dans la pratique, ces jugements correspondent au jugement réel et au jugement observé. Nous allons utiliser le coefficient Kappa dans deux situations afin d'obtenir un double indice qui sera l'initiateur d'un système permettant de décider, lorsque nous sommes dans la phase de prétraitement des données, quel processus de prétraitement de variables utilisé.

1 Double indice

Nous désirons que notre indice réponde à la question suivante : pour un jeu de données particulier, est il indispensable de construire de nouvelles variables à partir des variables initiales ? Ce qui revient à se demander si les données contiennent en elles mêmes l'information nécessaire et suffisante pour discriminer au mieux la variable endogène, ou s'il manque de l'information pour apprendre au mieux le concept cible. Pour répondre à ces questions, il est donc nécessaire, avant tout, d'étudier la structure des données. Cette dernière sera étudiée sous deux angles différents :

- En premier lieu, nous voulons obtenir une « mesure » de l'information que les variables exogènes apportent vis à vis de la variable endogène. C'est à dire, les résultats fournis par l'algorithme d'apprentissage sont ils en accord avec la variable endogène ?
- Ensuite, nous voulons apprécier la qualité de la structure inhérente aux données. C'est à dire sans avoir accès à l'information apportée par la variable endogène, comment se regroupent naturellement les données ? Les classes induites par la variable endogène peuvent elles se retrouver au sein des données sans information sur ces classes ou sur la variable endogène elle même ?

1.1 Le coefficient Kappa

Le coefficient de Kappa permet de chiffrer l'accord entre deux ou plusieurs observateurs ou techniques en l'absence de référence. C'est un outil pratique, relativement simple et très utilisé en pratique, en particulier dans le milieu médical. Il peut révéler des désaccords cachés, une divergence systématique ou non entre des juges. Si l'on prend un exemple dans le domaine médical où deux ou plusieurs praticiens examinant le même patient proposent des diagnostics différents ou des décisions thérapeutiques différentes. En l'absence d'une référence, cette multiplication des avis n'apporte pas la sécurité attendue d'un parfait accord diagnostique ou thérapeutique pour le patient. Le coefficient Kappa nous permet d'estimer le taux d'accord entre les praticiens et d'étudier leurs désaccords.

Appliquer dans notre cas, le coefficient Kappa va nous permettre de mesurer l'accord entre d'une part les valeurs de la variable endogène et les résultats d'un apprentissage supervisé et d'autre part les valeurs de la variable endogène et les résultats d'un apprentissage non supervisé.

L'accord entre des jugements est défini comme la conformité de deux ou plusieurs informations qui se rapportent au même objet. Cette notion implique un appariement des jugements. On obtient un appariement des jugements si ces derniers portent tous sur les mêmes objets.

1.1.1 Calcul de K

Kappa est le pourcentage de l'accord maximum corrigé de ce qu'il serait sous le simple effet du hasard. Dans le cas d'une étude entre deux observateurs qui émettent chacun un jugement possédant $r \geq 2$ modalités de jugement, le coefficient Kappa se calcul de la manière suivante :

$$\mathrm{K} = \frac{p_o - p_e}{1 - p_e}$$

avec $p_o = \frac{\sum_{r=1}^{R} n_{rr}}{n}$ et $p_e = \frac{1}{n^2} \sum_{r=1}^{R} n_{r.} n_{.r}$, tableau 69.

Il est possible de noter un certain nombre de propriétés de K :

- p_o représente la proportion d'accord observée et p_e la proportion de concordance attendue lorsque les jugements sont indépendants ;

- $K \in R$: K appartient à l'ensemble des réels ;
- Les valeurs de K varient entre 1 et –1 : $-1 \le K \le 1$;
- K atteint sa valeur maximale 1 lorsque $p_o = 1$ et $p_e = 0,5$: dans cette situation, l'accord entre les deux observateurs est maximal. Les deux observateurs possèdent les mêmes jugements ;
- K atteint sa valeur minimale 1 lorsque $p_o = 0$ et $p = 0,5$: dans cette situation, les deux observateurs sont en complet désaccord. Les deux observateurs possèdent des jugements opposés ;
- Plus K est proche de 1 et plus l'accord entre les jugements est élevé.

Obs. 2 / Obs. 1	$C_1^{'}$	. . .	$C_r^{'}$	. . .	$C_R^{'}$	Total
C_1	n_{11}	. . .	n_{1r}	. . .	n_{1R}	$\sum_{r=1}^{R} n_{1r}$
. . .	. . .	. . .	. . .	. . .	. . .	. . .
C_r	n_{r1}	. . .	n_{rr}	. . .	n_{rR}	$n_{r.}$
. . .	. . .	. . .	. . .	. . .	. . .	. . .
C_R	n_{R1}	. . .	n_R	. . .	n_{RR}	$\sum_{r=1}^{R} n_{Rr}$
Total	$\sum_{r=1}^{R} n_{r1}$	. . .	$n_{.r}$	. . .	$\sum_{r=1}^{R} n_{rR}$	n

Tableau 69 Tableau croisé des jugements des deux observateurs.

Si l'on se trouve dans le cas particulier où les jugements ne possèdent que deux modalités (tableau 70) alors le coefficient Kappa se définit de la manière suivante :

$$K = \frac{2(ad - bc)}{n_{1.}n_{.2} + n_{2.}n_{.1}}.$$

Obs. 2 / Obs. 1	$C_1^{'}$	$C_2^{'}$	total
C_1	a	b	$n_{1.}$
C_2	c	d	$n_{2.}$
total	$n_{.1}$	$n_{.2}$	n

Tableau 70 Tableau croisé pour deux observateurs.

Nous effectuons le calcul du coefficient Kappa sur un exemple comportant deux observateurs dont les jugements possèdent trois modalités. Cet exemple est présenté dans le tableau 71.

		Observateur 2			
		$C_1^{'}$	$C_2^{'}$	$C_3^{'}$	Total
Observateur 1	C_1	25	3	5	33
	C_2	6	32	8	46
	C_3	1	0	20	21
	Total	32	35	33	100

Tableau 71 Exemple de tableau croisé.

Ainsi, le coefficient Kappa se calcule de la manière suivante :

$$K = \frac{((25+32+20)/100) - ((1/100^2)\cdot(32\cdot 30 + 35\cdot 46 + 33\cdot 21))}{1 - ((1/100^2)\cdot(32\cdot 30 + 35\cdot 46 + 33\cdot 21))} = \frac{0{,}77 - 0{,}336}{1 - 0{,}336}$$

$$K = 0{,}654$$

1.1.2 Valeurs seuil

[137] ont proposé un classement de l'accord en fonction de la valeur du coefficient Kappa. Il est ainsi possible de qualifier l'accord entre les deux jugements considérés en fonction de la valeur du coefficient Kappa, (tableau 72) :

- Si $0{,}75 \leq K \leq 1$ alors l'accord entre les deux jugements est excellent.
- Si $K = 1$ alors l'accord est parfait : il n'y a aucune différence entre les deux jugements.
- Si $K = -1$ alors les deux jugements sont opposés : il n'y a aucune concordance entre les deux avis.
- Si $K = 0$ alors les deux jugements sont considérés comme indépendants.

Valeur de K	Accord entre les jugements
$0{,}75 \leq K \leq 1$	Excellent
$K > 0{,}41$	Bon
$K \leq 0{,}41$	Faible
$K = 0$	Jugements indépendants
$-1 < K < 0$	Désaccord > accord
$K = -1$	Désaccord total

Tableau 72 Valeurs seuil pour le coefficient Kappa

1.2 « Coefficient Kappa supervisé »

Nous voulons savoir si, sans effectuer de construction de variables, les variables exogènes permettent de bien discriminer la variable endogène. La condition d'utilisation du coefficient Kappa est d'être en présence de deux échantillons appariés. Nous allons utiliser le coefficient Kappa de la manière suivante :

- L'observateur 1 est représenté par les valeurs de la variable endogène (i.e. ses classes) pour l'ensemble des individus ;

- L'observateur 2 est représenté par les valeurs assignées aux individus par l'algorithme d'apprentissage choisi.

Nous sommes bien en présence de deux échantillon appariés. En effet, le même échantillon d'individus est jugé par deux observateurs différents. Nous pouvons ainsi déterminer si la classification des individus induite par l'algorithme d'apprentissage correspond à la classification de la variable endogène.

		Algorithme d'apprentissage supervisé			
		$C_1^{'}$	. . .	$C_R^{'}$	Total
Variable endogène	C_1	n_{11}	. . .	n_{1R}	$\sum_{r=1}^{R} n_{1r}$
	. . .	. . .	. . .	. . .	. . .
	C_R	n_{R1}	. . .	n_{RR}	$\sum_{r=1}^{R} n_{Rr}$
	Total	$\sum_{r=1}^{R} n_{r1}$	. . .	$\sum_{r=1}^{R} n_{rR}$	n

Tableau 73 Application du coefficient Kappa en supervisé.

Ainsi, les C_r et $C_r^{'}$ correspondent à la même classe de la variable endogène. Nous noterons le coefficient Kappa obtenu dans cette situation K_S et nous le nommerons « Kappa supervisé ».

Afin de calculer la valeur de K_S, nous devons dans un premier temps lancer l'algorithme d'apprentissage sur l'ensemble des individus qui seront utilisés par la suite pour le processus de construction de variables. Ensuite, il est nécessaire de construire le tableau croisé, tableau 73. Enfin, nous pouvons calculer K_S.

1.3 Structure inhérente aux données

Afin de déterminer si les données contiennent en elles-mêmes l'information nécessaire et suffisante pour retrouver le concept inféré par la variable endogène, nous effectuons une comparaison entre le résultat d'un apprentissage en non supervisé appliqué sur nos données privées de leur variable endogène et la classification induite par la variable endogène.

Afin de pouvoir effectuer une comparaison entre les résultats obtenus en supervisé et ceux obtenus en non supervisés, nous utilisons également le coefficient Kappa pour l'évaluation de la quantité d'information apportée par les variables exogènes.

1.3.1.1 Phase de classification non supervisée

Nous voulons obtenir un classement des individus en R groupes, R étant égal au nombre de classes de la variable endogène. Nous avons choisi d'utiliser une méthode « partitionnelle » qui nous permettra de déterminer une partition des objets du jeu de données. Ce type de méthodes procède à une recherche itérative de la partition optimisant un critère particulier. Si nous travaillons sur des variables qualitatives, nous utiliserons la méthode des K-Means [138], si nous travaillons avec des variables quantitatives, nous utiliserons la méthode des K-Modes, [139].

Nous avons choisi la méthode des K-Means car elle est la plus connue et la plus populaire des méthodes de classification non supervisée. Cette méthode vise à déterminer la partition de l'ensemble des individus du jeu de données considéré telle qu'elle optimise une fonction objectif basée sur le carré de la distance euclidienne entre l'individu ω_i et le centre de la classe à laquelle il appartient. L'objectif de cet algorithme est donc de déterminer la partition, possédant un nombre de classes fixé a priori, telle que l'homogénéité des classes soit la plus forte possible. Plus spécifiquement, l'algorithme débute par l'initialisation du centre des R classes, puis assigne séquentiellement chaque individu à la classe dont le centre est le plus proche. Ce dernier processus est réitéré tant que des mouvements d'objets d'une classe à une autre ont lieu. L'algorithme des K-Modes constitue l'adaptation des K-Means au cadre des données qualitatives.

Bien sûr, nous aurions pu utiliser d'autres méthodes de classification non supervisée. Ces méthodes ont été retenues parce qu'elles sont connues et nous donnent la possibilité de choisir le nombre de classes de la partition qu'elles effectuent.

1.3.1.2 Proportion liée à l'ensemble des individus

Nous considérons que les individus bien classés par l'algorithme d'apprentissage non supervisé le sont au regard de la variable endogène. Pour pouvoir appliquer le coefficient Kappa, nous devons être en présence de deux échantillons appariés. Or, le même échantillon d'individus est jugé par deux observateurs différents : La variable endogène et l'algorithme d'apprentissage non supervisé. Nous pouvons donc utiliser le coefficient Kappa de la manière suivante :

- L'observateur 1 est représenté par les valeurs de la variable endogène (i.e. ses classes) pour l'ensemble des individus ;
- L'observateur 2 est représenté par les valeurs assignées aux individus par l'algorithme d'apprentissage non supervisé, à savoir les K-Modes.

Nous pouvons ainsi déterminer si la classification des individus induite par l'algorithme d'apprentissage correspond à la classification de la variable endogène.

		Algorithme d'apprentissage non supervisé			
		$C_1^{'}$	. . .	$C_R^{'}$	Total
Variable endogène	C_1	n_{11}	. . .	n_{1R}	$\sum_{r=1}^{R} n_{1r}$
	. . .	. . .	. . .	. . .	. . .
	C_R	n_{R1}	. . .	n_{RR}	$\sum_{r=1}^{R} n_{Rr}$
	Total	$\sum_{r=1}^{R} n_{r1}$	. . .	$\sum_{r=1}^{R} n_{rR}$	n

Tableau 74 Application du coefficient Kappa en non supervisé.

Ainsi, les C_r et $C_r^{'}$ doivent correspondre à la même classe de la variable endogène. Nous noterons le coefficient Kappa obtenu dans cette situation K_{NS} et nous le nommerons « Kappa non supervisé ».

Afin de calculer la valeur de K_{NS}, nous devons dans un premier temps lancer l'algorithme d'apprentissage non supervisé sur l'ensemble des individus qui seront utilisés par la suite pour le processus de construction de variables. Ensuite, il est nécessaire de construire le tableau croisé, tableau 74. Enfin, nous pouvons calculer K_{NS}.

1.4 Procédure de choix

Construire ou ne pas construire de nouvelles variables va dépendre des valeurs des deux coefficients Kappa. Les valeurs seuil définies dans le cadre général d'utilisation du coefficient Kappa sont ici conservées. Ainsi si les valeurs de K_{NS} et K_S sont strictement supérieures à 0,41 alors la construction de variables est considérée comme inutile. Si ces valeurs sont comprises entre 0,41 et 0 alors la construction de variables est conseillée. Par contre, si ces valeurs sont inférieures ou égales à 0, alors la construction de variables apparaît comme indispensable.

K_S / K_S	$]0{,}41;1]$	$]0;0{,}41]$	$[-1;0]$
$]0{,}41;1]$	Construction inutile	Construction conseillée	Construction recommandée
$]0;0{,}41]$	Construction conseillée	Construction recommandée	Construction indispensable
$[-1;0]$	Construction recommandée	Construction indispensable	Construction indispensable

Tableau 75 Tableau décisionnel.

Le tableau 75 nous permet de prendre une décision au sujet de la construction en fonction des valeurs de K_{NS} et K_S. Dans ce tableau, la notion «recommandée» est considérée comme plus forte que la notion «conseillée».

Ainsi, si K_{NS} et K_S sont inférieurs à 0 alors la construction de variables est jugée indispensable. En effet, cela signifie que d'une part la structure intrinsèque des données ne correspond pas du tout à la

répartition des individus induite par la variable endogène et d'autre part que l'algorithme d'apprentissage supervisé n'arrive pas non plus à retrouver la classification de la variable endogène.

Si K_{NS} ou K_S sont inférieures à 0 ou si K_{NS} et K_S sont comprises entre 0 et 0,41, alors la construction de variables est recommandée à l'utilisateur. Dans ces situations, la construction peut apportée un plus à la qualité d'apprentissage. Cependant la décision doit dépendre de :

- La taille de l'espace de représentation : l'utilisateur veut il obtenir un espace de représentation relativement réduit ?
- La qualité d'apprentissage : l'utilisateur veut il améliorer la qualité d'apprentissage sans tenir compte de la taille de l'espace de représentation ?
- Le coût calculatoire : l'application de la construction de variable nécessite un certain temps de calcul, l'utilisateur est il prêt subir ce coût calculatoire supplémentaire ?

Si K_{NS} et K_S sont supérieure à 0,41 alors la construction paraît inutile. Cela signifie que la structure intrinsèque des données correspond presque parfaitement à la structure de données induite par la variable endogène et que l'algorithme d'apprentissage supervisé conduit à des résultats en terme de classification semblable à la classification de la variable endogène.

Afin d'obtenir une représentation visuelle claire et compréhensible par tous, nous utilisons un graphique dont l'axe des abscisses représente K_S et l'axe des ordonnées représente K_{NS}. Sur la figure 41, les parties grisées sont les situations pour lesquelles l'utilisateur n'a pas d'influence. Les parties blanches correspondent aux situations où l'utilisateur peut choisir s'il désire ou non construction en fonction des caractéristiques du problème et de ces priorités en terme de taille de l'espace de représentation, de qualité d'apprentissage et de coût calculatoire.

Chaque jeu de données traité est représenté dans le graphe décisionnel par un point ayant pour abscisse la valeur du K_S qui lui est associée et pour ordonnée la valeur du K_{NS}. Selon la région dans laquelle se situe le point, l'utilisateur peut aisément connaître s'il doit ou non appliquer une méthode de construction.

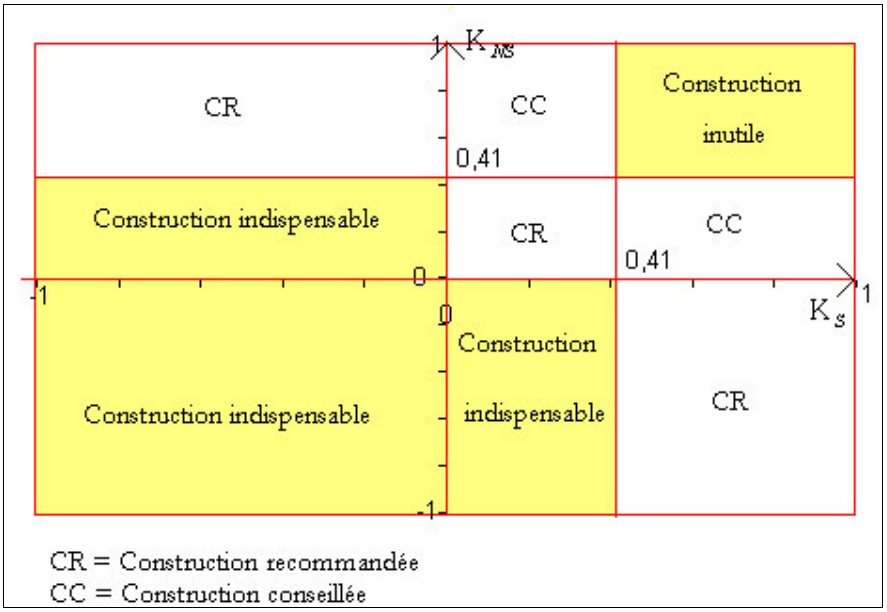

Figure 41 Graphique décisionnel.

2 Système de gestion de la phase de prétraitement des variables

Nous proposons de structurer la phase de prétraitement des variables, figure 42. Nous sommes en présence, lors de la phase de prétraitement des variables initiales, aussi, en fonction des résultats fournis par le double indice, un type de prétraitement est appliqué. Deux situations peuvent se rencontrer :

- Appliquer uniquement le processus de sélection ;
- Appliquer à la fois la construction et la sélection de variables.

L'ensemble des variables initiales sont fournies au système de prétraitement des variables. La première action consiste en l'application de la méthode de sélection de variables afin de « nettoyer » les variables exogènes en supprimant les variables bruitées, redondantes et/ou non pertinentes. Le sous-ensemble de variables obtenu est alors soumis au double indice. Si ce dernier conclut que la construction de variables est nécessaire alors une méthode de construction est appliquée au sous-ensemble de variables. Le sous-ensemble de variables voit sa taille augmenter. Si le double indice conclut que la phase de construction est inutile alors le sous-ensemble de variables sélectionnées est directement fourni à l'algorithme d'apprentissage.

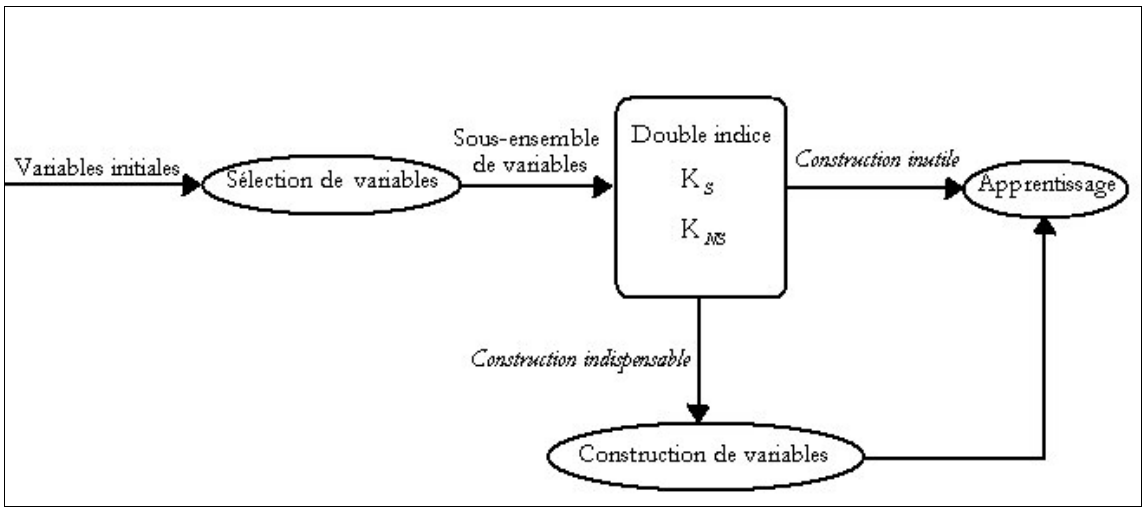

Figure 42 Système de prétraitement des variables.

Lors de nos expérimentations, nous avons bien sûr utilisé nos méthodes de sélection et de construction de variables. Etant donné qu'elles sont relativement peu coûteuses, nous avons dans la plupart des cas appliqué à la fois la sélection et la construction de variables. Cependant, l'utilisateur peut employer ce système de prétraitement avec tous les types de méthodes de construction et de sélection.

3 Expérimentations

Nous avons effectué l'étude expérimentale sur un ensemble de neuf jeux de données issus de la collection de l'UCI [70]. Les variables quantitatives ont été discrétisées à l'aide Fusinter [71]. Le découpage de l'ensemble des données a été effectué comme suit : la totalité des individus a été partagée aléatoirement en deux parties, tout en gardant la répartition initiale des classes. Le premier sous-ensemble d'individus contient 30% des individus et nous servira pour appliquer à la fois le processus de sélection de variables, le calcul du double, et le processus de construction de variables, si ce dernier est nécessaire. Les 70% d'individus restants sont utilisés pour les tests avant et après la phase de prétraitement. La méthode de sélection utilisée est celle développée au chapitre 2 et la méthode de construction utilisée est la méthode présentée au chapitre 3.

Sur chaque jeu de données, nous avons d'abord appliqué la méthode de sélection. Ensuite, nous avons calculer le double indice pour l'ensemble des jeux. Les tableaux 76 et 77 nous présentent les valeurs

de K_{NS} et K_S. La figure 43 nous montre le graphe décisionnel qui possède comme abscisses les valeurs de K_S et comme ordonnées les valeurs de K_{NS}. La répartition des jeux de données s'effectue comme suit :

- Peu de bases ne nécessitent pas de construction. Seules Monks-3, Iono et Vehicle ne nécessitent pas le processus de construction de variables. Cependant, Vehicle est à la limite de cette zone. Pour la 10 Cross-Validation, on peut remarquer que le fait de n'appliquer que la méthode de sélection de variables entraîne une amélioration de la qualité d'apprentissage. Pour Monks-3, le processus de sélection entraîne une dégradation de la qualité d'apprentissage, toutefois le nombre de variables constituant l'espace de représentation est relativement réduit. Et, comme nous l'avons vu au chapitre précédent, l'application du processus de construction sur Monks-3 entraîne une plus grande dégradation de la qualité d'apprentissage.
- La construction de variables est considérée comme indispensable pour la base German.
- La construction de variables est conseillée pour les bases Austra, Breast et CRX.
- La construction de variables est recommandée pour les bases Pima et Tic Tac Toe.
- Nous avons décidé d'appliquer le processus de construction de variables pour tous les jeux pour lesquels la construction est conseillée, recommandée ou indispensable.
- Les figures 44 et 45 et les tableaux 78 et 79 nous présentent les taux d'erreur moyens en Cross-Validation avant et après la phase de prétraitement. La figure 46 nous présente l'évolution de la taille de l'espace de représentation des données.
- Dans la grande majorité des cas, la phase de prétraitement est suivie d'une amélioration de la qualité d'apprentissage. De plus cette amélioration est accompagnée par une réduction de la taille de l'espace de représentation des données. Même si la phase de prétraitement est constituée par le processus de sélection et par le processus de construction.

	Austra	Breast	CRX	Iono	German	Monks-3	Pima	Tic Tac Toe	Vehicle
p_o	0,5580	0,6810	0,5556	0,9038	0,5526	0,9819	0,6043	0,6760	0,5787
p_e	0,5548	0,6321	0,5502	0,5361	0,5970	0,5030	0,6376	0,5294	0,2495
K	0,0072	0,1329	0,0118	0,7927	-0,1100	0,9636	-0,0918	0,3114	0,4387

Tableau 76 Calcul de K .

S	Austra	Breast	CRX	Iono	German	Monks-3	Pima	Tic Tac Toe	Vehicle
p_o	0,8738	0,9567	0,9078	0,9808	0,7633	0,9940	0,7860	0,6538	0,7171
p_e	0,4962	0,5487	0,5033	0,5361	0,6160	0,5030	0,5336	0,5357	0,2480
K	0,7495	0,9041	0,8143	0,9585	0,3837	0,9879	0,5413	0,2544	0,6238

Tableau 77 Calcul de K .

	Sans prétraitement		Avec prétraitement		Ecart	Ecart relatif	Taille espace de représentation	Prétraitement
	Erreur	σ	Erreur	σ				
Austra	16,60	4,57	15,72	5,6	-0,88	-5,30%	3	Sélection +Construction
Breast	5,95	1,95	4,29	2,13	-1,66	-27,90%	5	Sélection +Construction
CRX	14,73	5,68	14,06	4,73	-0,67	-4,55%	5	Sélection +Construction
German	31,86	7,53	25,57	7,16	-6,29	-19,74%	7	Sélection +Construction
Iono	21,37	8,39	11,73	5,59	-9,64	-45,11%	2	Sélection
Monks-3	1,28	1,28	3,88	2,69	2,60	203,13%	2	Sélection
Pima	26,11	5,43	25,45	7,86	-0,66	-2,53%	4	Sélection +Construction
Tic Tac Toe	33,43	5	23,08	5,59	-10,35	-30,96%	9	Sélection +Construction
Vehicle	34,24	4,96	28,75	5,44	-5,49	-16,03%	14	Sélection

Tableau 78 Evaluation du système avec ID3 pour une 10 Cross-Validation.

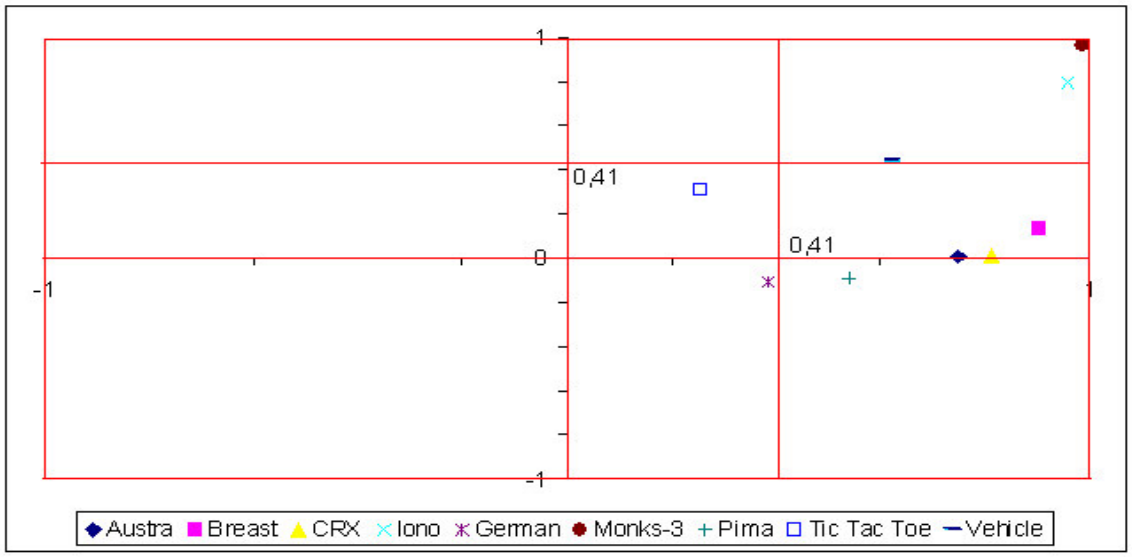

Figure 43 Graphique décisionnel pour les bases de l'UCI Irvine.

	Sans prétraitement		Avec prétraitement		Ecart	Ecart relatif	Taille de l'espace de représentation	Processus de prétraitement
	Erreur	σ	Erreur	σ				
Austra	15,91	2,58	15,49	4,63	-0,42	-2,64%	3	Sélection +Construction
Breast	5,7	1,89	5,29	0,99	-0,41	-7,19%	5	Sélection +Construction
CRX	14,66	2,43	16,11	2,36	1,45	9,89%	5	Sélection +Construction
German	28,57	4,58	24,14	2,23	-4,43	-15,51%	7	Sélection +Construction
Iono	13,39	3,62	11,72	2,91	-1,67	-12,47%	2	Sélection
Monks-3	1,29	0,81	3,86	3,34	2,57	199,22%	2	Sélection
Pima	24,3	2,48	23,74	5,22	-0,56	-2,30%	4	Sélection +Construction
Tic Tac Toe	22,8	3,94	24,1	2,2	1,30	5,70%	9	Sélection +Construction
Vehicle	29,41	3,49	32,1	2,78	2,69	9,15%	14	Sélection

Tableau 79 Evaluation du système avec ID3 pour cinq 2 Cross-Validation.

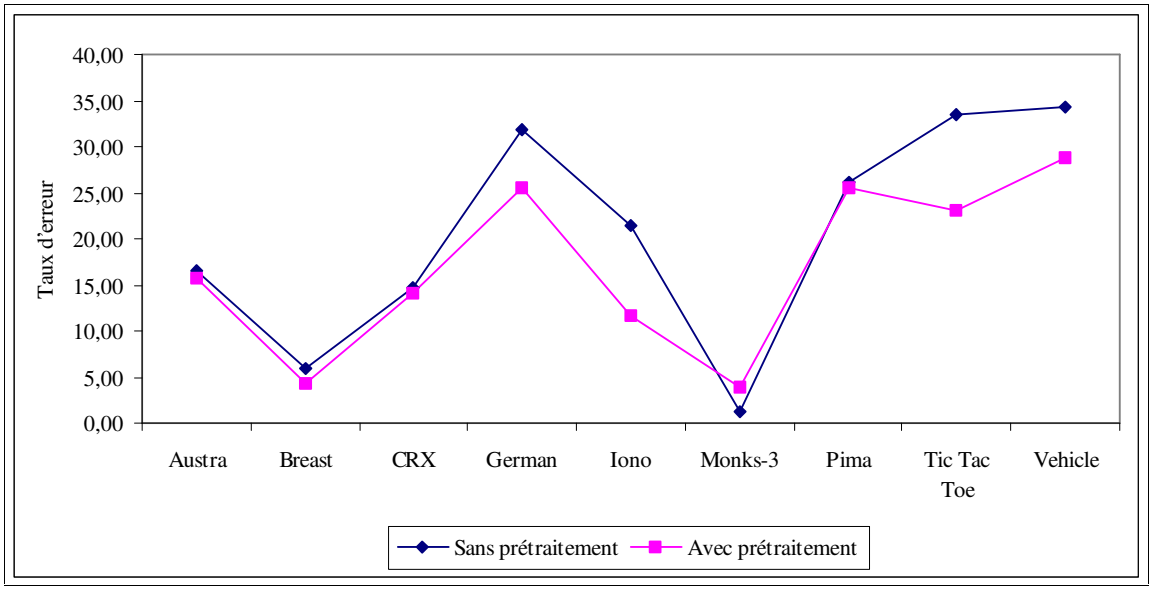

Figure 44 Taux d'erreur avec ID3 pour un 10 Cross-Validation.

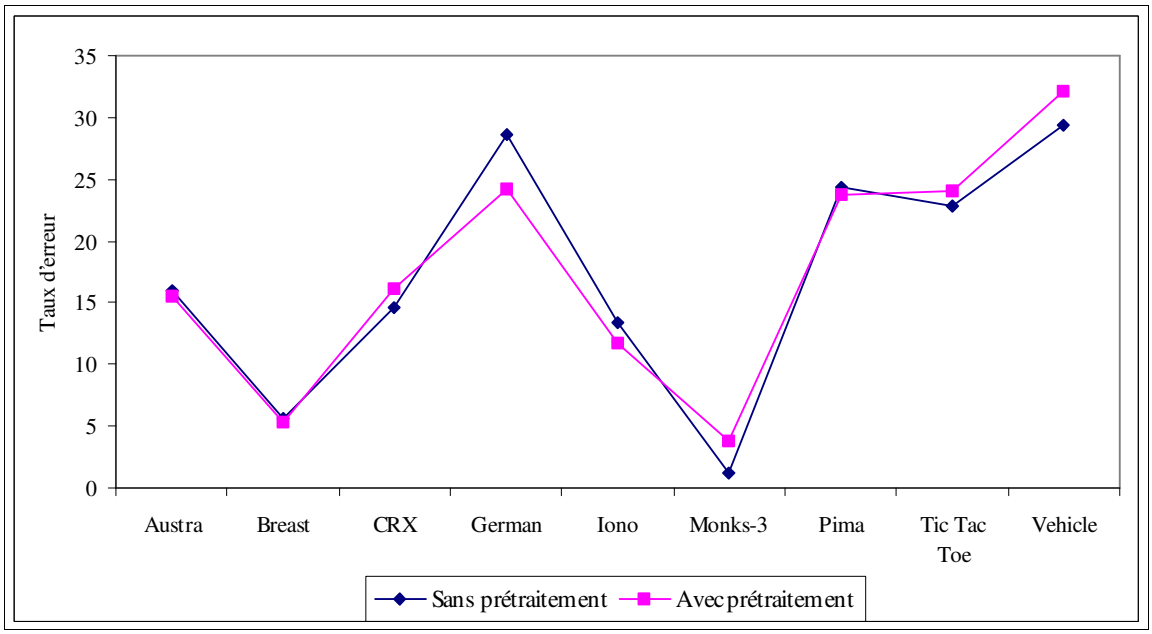

Figure 45 Taux d'erreur avec ID3 pour cinq 2 Cross-Validation.

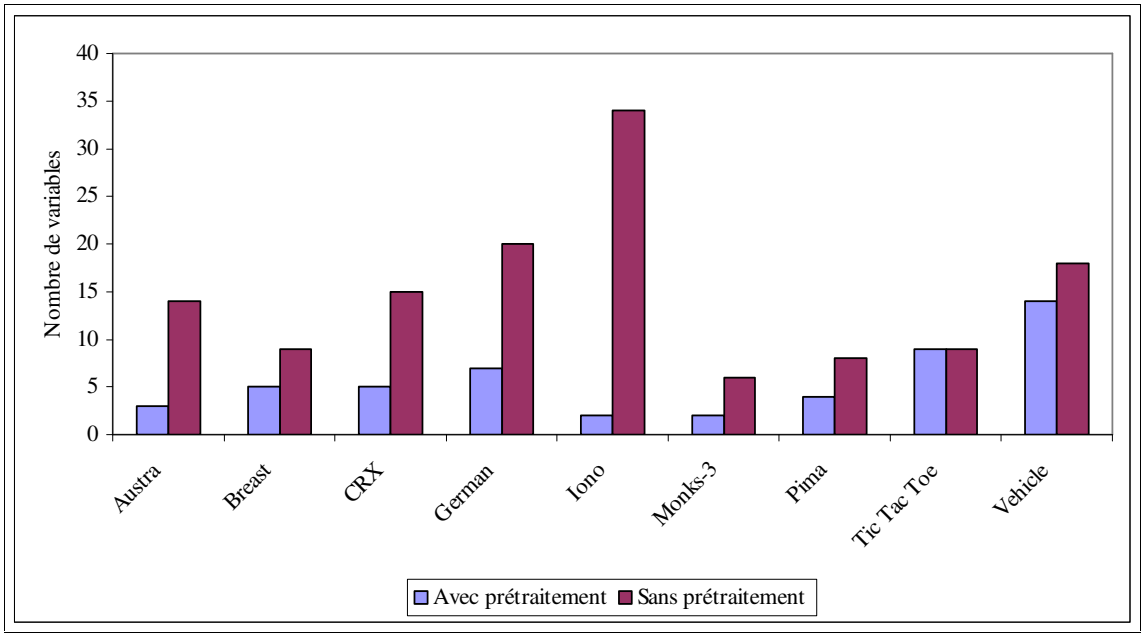

Figure 46 Evaluation de la taille de l'espace de représentation.

4 Conclusion

Nous avons mis en place un système permettant de gérer et d'«optimiser» la phase de prétraitement des variables. Après l'application de la méthode de sélection de variables, nous utilisons un double indice basé sur le coefficient de Kappa. Ce double indice nous permet de déterminer si l'application du processus de construction de variables est nécessaire. Le double indice se compose du calcul de deux coefficients de Kappa :

- K_S est le coefficient Kappa appliqué pour comparer la classification obtenue suite à l'application d'un algorithme d'apprentissage supervisé et la classification induite par la variable endogène ;
- K_{NS} est le coefficient Kappa appliqué pour comparer la classification non supervisé des données et la classification induite par la variable endogène.

Les expérimentations nous ont montré qu'en effet le processus de construction de variables n'est pas toujours indispensable. La qualité de l'apprentissage, après la phase de prétraitement que celle-ci soit composée des processus de sélection et de construction ou seulement du processus de sélection, se retrouve améliorer.

L'utilisateur peut influencer la décision liée au processus de construction. En effet, selon les valeurs du double indice, il n'est fourni à l'utilisateur qu'une indication sur le choix à effectuer. Ainsi, l'utilisateur est libre d'appliquer le processus de construction de variables en fonction de ces priorités (taille de l'espace de représentation, amélioration de la qualité d'apprentissage…).

Nous avons choisi d'appliquer dans tous les cas le processus de sélection de variables. En effet, ce processus de prétraitement est important : il permet de supprimer les éléments non pertinents de l'ensemble des variables. Cependant, l'application de ce processus entraîne un coût calculatoire supplémentaire et peut selon les cas (en particulier Monks-3) entraîner une dégradation de la qualité d'apprentissage. Il serait intéressant de créer un indice qui puisse arbitrer entre sélection et construction. Quatre situations seraient alors possibles :

- La construction et la sélection de variables sont nécessaires pour améliorer la qualité de l'apprentissage ;
- La sélection de variables est elle seule indispensable ;
- La construction de variables est elle seul indispensable ;
- Ni la construction ni la sélection ne sont utiles.

Ainsi, la phase de prétraitement serait alors précédée du calcul d'un indice jouant le rôle de décisionnaire au sein de la phase de prétraitement.