Université Lumière Lyon2
Année 2003

Thèse
pour obtenir le grade de
Docteur
en
Informatique

présentée et soutenue publiquement par

Pierre-Emmanuel JOUVE
le 10 décembre 2003

Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données

préparée au sein du laboratoire ERIC
Equipe de Recherche en Ingénierie des Connaissances

sous la direction de
Nicolas Nicoloyannis

devant le jury, composé de:

Jean-Paul Rasson, Rapporteur	Professeur, Facultés Universitaires N.D. de la Paix, Namur
Gilles Venturini, Rapporteur	Professeur, Université de Tours
Mohand-Saïd Hacid, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Michel Lamure, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Gilbert Ritschard, Examinateur	Professeur, Université de Genève
Nicolas Nicoloyannis, Directeur de thèse	Professeur, Université Lumière-Lyon 2

Table des matières

1 Introduction, Préambule **1**

2 Concepts, Notions et Notations Utiles **7**
2.1 Données Catégorielles 7
2.1.1 Domaines et Attributs Catégoriels 8
2.1.2 Objets Catégoriels 9
2.1.2.1 Similarités, Dissimilarités entre Objets Catégoriels 10
2.1.3 Ensemble d'Objets Catégoriels 11
2.1.3.1 Mode d'un Ensemble d'Objets Catégoriels . . . 11
2.1.3.2 Similarités et Dissimilarités entre Ensembles d'Objets Catégoriels 12
2.1.3.3 Similarités et Dissimilarités au sein d'un Ensemble d'Objets Catégoriels 12
2.1.3.4 Voisinage d'une Partition d'un Ensemble d'Objets Catégoriels 13
2.2 Le Nouveau Critère de Condorcet 13

3 Classification Non Supervisée **15**
3.1 Introduction 15
3.1.1 Méthodologie Générale de la Classification Non Supervisée 16
3.1.2 Applications de la Classification Non Supervisée 16
3.1.3 Taxonomies des Méthodes de Classification Non Supervisée 17
3.1.4 Méthodes de Classification Non Supervisée pour Données Catégorielles 19
3.1.5 Challenges Actuels en Classification Non Supervisée . . 22
3.2 Une Nouvelle Méthode de Classification Non Supervisée "Orientée Utilisateur" 24
3.2.1 Critère d'Évaluation de l'Aspect Naturel d'une Partition d'Objets 24
3.2.2 La Méthode de Classification Non Supervisée "Orientée Utilisateur" 26
3.2.2.1 Travaux Liés et Spécificités du Travail 26
3.2.2.2 L'Algorithme de Classification Non Supervisée 27

3.2.2.3 Complexité de l'Algorithme 29
3.2.2.4 Qualités de la Méthode pour l'Utilisateur . . . 30
3.2.2.5 Illustration du Fonctionnement de l'Algorithme 30
3.2.3 Evaluation de l'Algorithme de Classification non Supervisée 31
3.2.3.1 Evaluation de la Validité des Classifications . . 31
3.2.3.2 Evaluation de la Stabilité 37
3.2.3.3 Evaluation de l'Efficacité Algorithmique 40
3.2.4 Eléments Additionnels 42
3.2.4.1 Valeurs Spécifiques pour le Domaine des Variables Catégorielles 42
3.2.4.2 Gestion des Valeurs Manquantes : 44
3.2.4.3 Introduction de Contraintes : 44
3.2.4.4 De l'Apprentissage Non Supervisé à l'Apprentissage Supervisé : l'Apprentissage Non Supervisé sous Contraintes 50
3.3 Conclusion 54

4 Validité en Apprentissage Non Supervisé 57
4.1 Validité d'une Classification Non Supervisée : Définition et Evaluation 58
4.1.1 Mode d'Evaluation par Critères Externes 59
4.1.1.1 Méthode de Monte Carlo 59
4.1.1.2 Mesures Statistiques 60
4.1.2 Mode d'Evaluation par Critères Internes 61
4.1.3 Modes d'Evaluation Relatifs 63
4.1.3.1 Cas 1 : Le nombre final de classes, nc, n'est pas contenu dans P_{alg}. 63
4.1.3.2 Cas 2 : Le nombre final de classes, nc, est contenu dans P_{alg}. 64
4.1.3.3 Indices 64
4.1.4 Autres Modes d'Evaluation 67
4.2 Nouveaux Indices et Nouvelle Méthodologie pour l'Evaluation et la Comparaison de la Validité de Classifications Non Supervisées 68
4.2.1 Concepts et Formalismes Introductifs 69
4.2.1.1 Evaluation de l'homogénéité interne des classes d'une cns 71
4.2.1.2 Evaluation de la séparation entre classes d'une cns (ou hétérogénéité entre classes) 72
4.2.1.3 Notions Additionnelles 73
4.2.1.4 Remarques importantes concernant l'aspect calculatoire 73

4.2.2 La nouvelle méthodologie pour l'évaluation et la comparaison de validité de cns 75
4.2.2.1 Caractérisation statistique des valeurs de : *LM* et *NLD* 76
4.2.2.2 Méthodologie 77
4.2.2.3 Expérimentations 82
4.2.2.4 Expérimentations sur le jeu de données Small Soybean Disease 82
4.2.3 Expériences sur le jeu de données Mushrooms 92
4.2.3.1 Description 92
4.2.3.2 Analyse des Résultats 95
4.2.4 Résumé et Informations Supplémentaires 96

5 Sélection de Variables, Contributions pour l'apprentissage supervisé et non supervisé 105
5.1 Sélection de Variables pour l'Apprentissage Supervisé 107
5.1.1 Caractéristiques de la Sélection de Variables 107
5.1.2 Les Types de Méthodes 107
5.1.3 Directions de Recherche 108
5.1.3.1 Forward Selection (FS) (Ajout de variables) . . 108
5.1.3.2 Backward Elimination (BE) (Suppression de variables) 109
5.1.3.3 Méthodes Bidirectionnelles 109
5.1.4 Stratégie de Recherche 109
5.1.5 Fonction d'Evaluation 110
5.1.6 Critère d'Arrêt 111
5.1.7 Approches Filtres 111
5.1.8 Approches Enveloppes 114
5.1.9 Autres Approches 115
5.2 Contribution à la Sélection de Variables pour l'Apprentissage Supervisé : Une Nouvelle Méthode Efficiente et Rapide 118
5.2.1 Hypothèses et Idées Fondamentales 118
5.2.2 Evaluation de la Validité d'une Partition dans un Sous-Espace de l'ERD 119
5.2.3 La Nouvelle Méthode de Sélections de Variables 120
5.2.3.1 La Méthode de Base : une Méthode Exhaustive 121
5.2.3.2 Réduction de la Complexité par Introduction d'un AG 124
5.2.4 Evaluation Expérimentale 126
5.2.4.1 Présentation de l'Evaluation Expérimentale . . 126
5.2.4.2 Analyse de l'Evaluation Expérimentale 127
5.2.5 Conclusion 131
5.3 Contribution à la Sélection de Variables pour l'Apprentissage Non Supervisé : Une Nouvelle Méthode Efficiente et Rapide . . 143

5.3.1 Evaluation de l'Adéquation entre deux Ensembles de Variables . . . 144
5.3.2 Remarques Importantes Concernant l'Aspect Calculatoire 145
5.3.3 Evaluation de l'adéquation entre **EV** un Ensemble de Variables et $\mathbf{EV}_\star$ un Sous Ensemble de **EV** ($\mathbf{EV}_\star \subseteq \mathbf{EV}$) 146
5.3.4 Evaluation/Comparaison de l'Adéquation entre un Ensemble de Variables (**EV**) et des Sous Ensembles de **EV** 148
5.3.5 La Nouvelle Méthode de Sélection de Variables . . . 148
5.3.6 Evaluations Expérimentales . . . 149
5.3.6.1 **Expérience #1** : Evaluation expérimentale sur jeux de données synthétiques . . . 149
5.3.6.2 **Expérience #2** : Evaluation Expérimentale sur Jeux de Données de l'UCI . . . 154
5.3.7 Conclusion . . . 162

6 Agrégation de Classifications Non Supervisées : La Problématique "Cluster Ensembles" 165
6.1 Introduction . . . 165
6.1.1 Illustration de la Problématique "Cluster Ensembles" . . . 167
6.1.2 Motivations, Objectifs de la Problématique "Cluster Ensembles" . . . 168
6.1.2.1 Réutilisation de Connaissances . . . 169
6.1.2.2 Calcul Distribué pour la cns . . . 169
6.1.3 Travaux Liés . . . 172
6.1.4 Principaux Challenges pour la Problématique "Cluster Ensembles" . . . 175
6.2 Mesures d'Adéquation . . . 175
6.2.1 Adéquation entre Classifications Non Supervisées . . . 176
6.2.2 Adéquation pour un Couple de Classification Non Supervisée . . . 176
6.2.3 Adéquation entre une Classification Non Supervisée et un Ensemble de Classifications Non Supervisées . . . 177
6.3 Contribution à la Problématique "Cluster Ensembles" : Trois Méthodes pour l'Agrégation de Classifications Non Supervisées . . . 177
6.3.1 Première Méthode pour l'Agrégation de cns : Une Méthode Intuitive . . . 178
6.3.2 Seconde Méthode pour l'Agrégation de Classifications Non Supervisées : Utilisation de la Méthode KEROUAC . . . 179
6.3.2.1 Utilisation de KEROUAC pour la cns en considérant des Méta-Variables . . . 179
6.3.2.2 Relation entre $P_\star$ and P_β . . . 180
6.3.2.3 Conclusion . . . 181
6.3.2.4 Illustration . . . 181

6.3.2.5 Propriétés de la Méthode 182
6.3.3 Troisième Méthode pour l'Agrégation de Classifications Non Supervisées : Utilisation de la Méthode K-Modes . . 183
6.3.3.1 Illustration 184
6.3.3.2 Propriétés de la Méthode 184
6.3.4 Evaluations Expérimentales 184
6.3.4.1 Evaluations, Comparaisons et Discussions Préliminaires 184
6.3.4.2 Evaluations, Comparaisons et Discussions Complémentaires 191
6.3.4.3 Comportement de la méthode KEROUAC face à des cns à agréger possédant des nombre de classes très différents 204
6.4 Conclusion 207

7 Conclusion 211

8 Données Utilisées pour les Expérimentations 217
8.1 Jeu de Données ADULT 217
8.2 Jeu de Données MUSHROOMS 218
8.3 Jeu de Données BREAST CANCER 220
8.4 Jeu de Données CAR 222
8.5 Jeu de Données : ADULT 224
8.6 Jeu de Données Contraceptive Method Choice 225
8.7 Jeu de Données FLAGS 226
8.8 Jeu de Données GERMAN 227
8.9 Jeu de Données HOUSE VOTES 84 229
8.10 Jeu de Données IONOSPHERE 230
8.11 Jeu de Données MONKS 231
8.12 Jeu de Données NURSERY 232
8.13 Jeu de Données PIMA 234
8.14 Jeu de Données SICK 235
8.15 Jeu de Données SMALL SOYBEAN DISEASES 236
8.16 Jeu de Données VEHICLE 237
8.17 Jeu de Données WINE 240
8.18 Jeu de Données SPAM 241

Bibliographie 243

Table des figures 254

Liste des tableaux 257

5 Sélection de Variables, Contributions pour l'apprentissage supervisé et non supervisé

"Less is more..."

- Huan Liu & Hiroshi Motoda -
"Feature Extraction, Construction, and Selection : A Data Mining Perspective", Kluwer Academic, Boston MA (1998)

La tendance actuelle d'un accroissement toujours plus fort de la taille des bases de données rend la problématique de l'amélioration de l'espace de représentation des données (ERD) de plus en plus critique en ECD. Une des difficultés majeures associées à la problématique de l'amélioration de la qualité de l'ERD est celle de la dimension de cet espace[1]. Ce problème se traduit par le nombre de variables (descripteurs) caractérisant chaque objet (par exemple, le nombre de variables exogènes dans le cadre de l'apprentissage supervisé)[2]. Un nombre élevé de descripteurs peut en effet s'avérer pénalisant pour un traitement pertinent et efficace des données, d'une part par les problèmes algorithmiques que cela peut entraîner (liés au coût calculatoire et à la capacité de stockage nécessaire), et d'autre part car parmi les descripteurs certains peuvent être non-pertinents, inutiles et/ou redondants perturbant ainsi le bon traitement des données. Or, il est très souvent difficile voire impossible de distinguer les descripteurs pertinents des descripteurs non-pertinents.

Le problème de la dimension des données peut ainsi être résumé par l'aphorisme de Liu et Motoda "Less is more" [LM98] qui met en exergue la nécessité de supprimer l'ensemble des portions non pertinentes des données de manière préalable à tout traitement si on désire en extraire des informations utiles et compréhensibles.

La sélection de variables (SdV) constitue une solution à ce problème. Ce processus vise en effet à la détermination d'un sous ensemble optimal de descripteurs[3] et donc à la réduction du nombre de variables par élimination des variables non pertinentes. Cela implique alors une accélération des traitements

1. l'autre difficulté majeure étant la quantité de données
2. le problème de la quantité de données se caractérise par le nombre d'objets à traiter

postérieurs et peut engendrer une amélioration de la qualité de ces mêmes traitements (la précision prédictive de modèles d'apprentissage supervisé peut, par exemple, être accrue). Enfin, le bruit généré par certaines variables peut être réduit.

La SdV suit généralement un processus itératif (cf. figure 5.1) menant au choix d'un sous-ensemble optimal de variables, ce processus se décompose en trois étapes (Génération de sous-ensembles de variables → Evaluation des sous-ensembles → Test sur le critère d'arrêt) qui s'enchaînent séquentiellement et constituent la partie itérative du mécanisme de SdV. Cette partie itérative s'achève lorsqu'un critère d'arrêt est satisfait, elle est alors suivie par une phase de validation du sous ensemble optimal par l'intermédiaire d'un algorithme d'apprentissage. Plus rarement, la SdV est constituée par un unique processus séquentiel permettant d'établir un classement des différentes variables de l'ERD selon leur intérêt pour le processus ultérieur d'apprentissage.

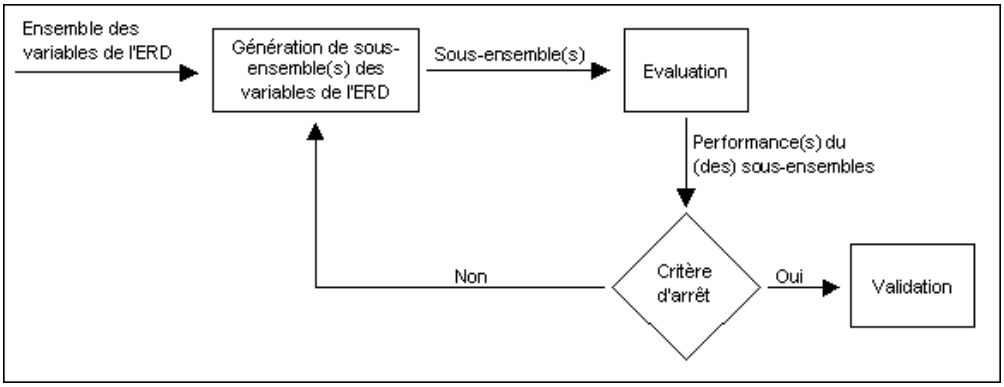

FIG. 5.1 –: *schéma du processus de sélection de variables*

L'évaluation de la qualité du processus de SdV s'effectue en prenant en compte :

- le différentiel de qualité des processus d'apprentissage réalisés respectivement sur l'ERD complet et sur le sous espace de l'ERD constitué des variables sélectionnées (par exemple, le différentiel de précision prédictive en apprentissage supervisé entre un modèle bâti à partir de l'ERD complet et un modèle bâti à partir d'un sous-espace de l'ERD) ;
- le différentiel de dimension de l'ERD complet et du sous espace de l'ERD constitué des variables sélectionnées et conséquemment l'accélération du processus d'apprentissage impliquée par la SdV.

3. l'optimalité du sous ensemble s'entend ici comme l'optimalité du sous ensemble par rapport à un critère particulier

5.1 Sélection de Variables pour l'Apprentissage Supervisé

Pour l'apprentissage supervisé, l'objectif de la SdV est de déterminer quelles sont les variables exogènes de l'ERD pertinentes pour la prédiction de la variable endogène. La réduction de la dimension de l'ERD doit alors permettre une accélération de la phase d'apprentissage et/ou de généralisation, ainsi que l'obtention d'une capacité prédictive du modèle équivalente ou supérieure à celle du modèle bâti à partir l'ensemble complet des variables de l'ERD.

5.1.1 Caractéristiques de la Sélection de Variables

- **Intérêts** :
  - permettre l'élimination de variables inutiles et redondantes,
  - accélérer le processus d'apprentissage
  - améliorer la précision prédictive des algorithmes d'induction.
  - permettre d'effectuer des recherches sur un sous-ensemble optimal de variables (il est alors possible de prendre en compte les interactions qui existent entre les variables).

- **Forme des résultats**, deux formes possibles :
  - une liste ordonnée de variables classées selon un critère d'évaluation. (Ce type de résultats ne fournit des renseignements que sur la pertinence d'une variable par rapport aux autres et le nombre de variables constituant l'ensemble final doit être connu/déterminé.)
  - un sous-ensemble optimum de variables au sein duquel aucune différence ne peut être faîte quant à la pertinence des variables.

- **Caractéristiques des Méthodes** :
  - Un type d'approche : filtre ou enveloppe,
  - Une direction de recherche dans l'espace des sous-ensemble de variables de l'ERD : par élimination ou par ajout ou par élimination/ajout de variable(s),
  - Une stratégie de recherche : complète ou heuristique ou aléatoire,
  - Une fonction d'évaluation des sous ensembles de variables (l'algorithme de recherche doit maximiser ou minimiser cette fonction),
  - Un critère d'arrêt.

5.1.2 Les Types de Méthodes

Il existe deux familles d'algorithmes visant à sélectionner un sous-ensemble optimal de variables : les méthodes "enveloppe" (Wrapper Approach) et les

méthodes "filtre" (Filter Approach). Ces deux familles d'approches s'opposent de part l'utilisation ou la non-utilisation de l'algorithme d'induction : les méthodes enveloppe utilisent l'algorithme d'induction contrairement aux méthodes filtre. Ainsi, les méthodes enveloppe évaluent les différents sous-ensembles générés par l'intermédiaire de l'algorithme d'apprentissage (cf. figure 5.2 à droite) ; les méthodes filtre, quant à elles, n'utilisent absolument pas l'algorithme d'apprentissage dans leur processus de recherche du sous-ensemble optimal (cf. figure 5.2 à gauche).

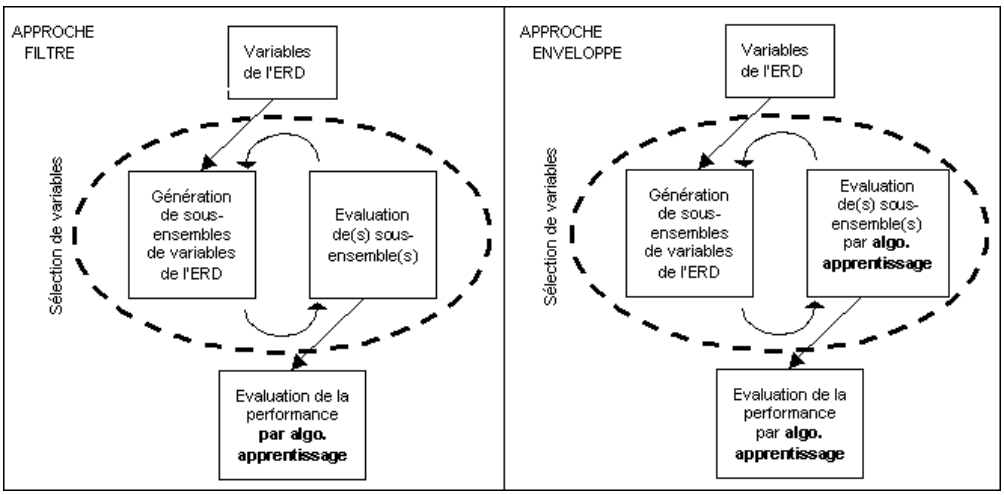

FIG. 5.2 –: *Approches Filtre et Enveloppe pour la Sélection de Variables*

5.1.3 Directions de Recherche

La sélection de variables est un problème de recherche où chaque état de l'espace de recherche spécifie un des 2^p sous-ensemble de variables de l'ERD (p le nombre de variables de l'ERD). Le passage de l'état initial à l'état final peut être schématisé par un graphe partiellement ordonné où chaque état enfant possède un ensemble de variables différents de ses parents. Les méthodes de sélection de variables utilisent donc l'ordre partiel des variables pour organiser leur recherche d'un sous-ensemble optimal de variables. Cet ordre partiel correspond à l'agencement des variables dans le temps, c'est à dire à leur utilisation lors du processus de sélection. Les directions de recherche peuvent être de trois types : Ajout de variables, Suppression de variables, Ajout/Suppression de variables.

5.1.3.1 Forward Selection (FS) (Ajout de variables)

Cette stratégie débute avec l'ensemble vide, puis, à chaque itération, la variable optimale suivant un certain critère est ajoutée. Le processus s'arrête

quand il n'y a plus de variable à ajouter, ou quand un certain critère est satisfait.

5.1.3.2 Backward Elimination (BE) (Suppression de variables)

Cette stratégie débute avec l'ensemble de toutes les variables, puis, à chaque itération, une variable est enlevée de l'ensemble. Cette variable est telle que sa suppression donne le meilleur sous-ensemble selon un critère particulier. Le processus s'arrête quand il n'y a plus de variable à supprimer, ou quand un certain critère est satisfait.

5.1.3.3 Méthodes Bidirectionnelles

Il est également possible d'utiliser une variation de l'ordre partiel des variables : Devijver et Kittler [DK82] définissent un opérateur qui ajoute k variables et en enlève une. La première décision à prendre est donc le point de départ de la recherche :

- Avec un ensemble vide : il s'agit de la Forward Stepwise Selection
- Avec un ensemble complet : Backward Stepwise Elimination
- Avec un ensemble d'attributs choisis aléatoirement.

REMARQUE : Les méthodes bidirectionelles permettent de pallier au problème de l'irrévocabilité de la suppression ou de l'ajout d'une variable. En effet, l'importance d'une variable peut se voir modifiée au cours des différentes itérations du processus de SdV. Ces méthodes autorisent l'ajout et la suppression d'une variable de l'ensemble des variables à n'importe quelle étape de la recherche (autre que la première) contrairement à la FS (resp. BE) pour laquelle une fois qu'une variable a été ajoutée (resp. supprimée) il est impossible de la retirer (resp. réintégrer).

5.1.4 Stratégie de Recherche

La stratégie de recherche dépend de la taille de l'espace de recherche. Si la taille de l'ensemble initial de variables est p, alors le nombre de sous-ensembles candidats est 2^p. Une recherche exhaustive n'est donc que rarement envisageable, ainsi, pour atteindre les objectifs de la SdV trois catégories de méthodes sont applicables :

- **Les stratégies de recherche complète** : une recherche complète des sous-ensembles optimaux est effectuée en tenant compte de la fonction d'évaluation utilisée. Cette méthode n'est pas forcément exhaustive [SJL90] : différentes fonctions heuristiques peuvent être utilisées afin de réduire l'espace de recherche sans compromettre les chances de trouver le sous-ensemble optimal.

- **Les stratégies de recherche heuristique** : on ajoute (ou on ôte) pas à pas des variables à l'ensemble des variables sélectionnées (ou restantes) jusqu'à ce que le sous-ensemble ne puisse plus être amélioré. Cette méthode revient à parcourir le chemin reliant l'état initial à l'état final du graphe des états. A chaque itération, toutes les variables restantes jusque là peuvent être sélectionnées. Il y a plusieurs variantes de ce simple processus mais la génération des sous-ensembles est fondamentalement incrémentale (soit croissante soit décroissante). La taille de l'espace de recherche est p^2 (il existe cependant des exceptions : [KR92a], [Car93]). Ces procédures sont simples à implémenter, peu coûteuses et donnent un résultat très rapidement parce que l'espace de recherche est seulement quadratique en terme de nombre de variables. Mais elles ne permettent pas d'obtenir un sous-ensemble optimal.
- **Les stratégies de recherche aléatoire** : la recherche d'un bon sous-ensemble se fait sur l'espace réduit aux sous-ensembles possibles. La taille de cet espace (inférieure à 2^p) est définie en fixant le nombre d'itérations. L'optimalité de la solution dépend à la fois des ressources disponibles et des valeurs assignées aux paramètres liés à la procédure de génération. Pour obtenir une solution, il n'est pas nécessaire d'attendre la fin de la recherche. Il n'est cependant pas possible de savoir si le sous-ensemble obtenu à l'instant t est optimal mais seulement si il est meilleur que les précédents.

5.1.5 Fonction d'Evaluation

L'objectif associé à la fonction d'évaluation est de mesurer la capacité d'une variable, ou d'un ensemble de variables, à discriminer les classes de la partition impliquée par la variable endogène. L'optimalité d'un sous-ensemble est relative à la fonction d'évaluation utilisée. Dash et Liu [DL97] considèrent que ces fonctions peuvent être regroupées en cinq catégories qui sont les suivantes :

- **Information** : fonctions quantifiant l'information apportée par une variable sur la variable à prédire. La variable, ayant le gain d'information le plus élevé, est préférée aux autres variables. (Le gain d'information étant la différence entre l'incertitude a priori et l'incertitude a posteriori.)
- **Distance** : fonctions s'intéressant au pouvoir discriminant d'une variable. Elles évaluent la séparabilité des classes en se basant sur les distributions de probabilités des classes. Une variable est préférée à une autre si elle induit une plus grande séparabilité.
- **Dépendance** : fonctions mesurant la corrélation ou l'association. Elles permettent de calculer le degré avec lequel une variable exogène est associée à une variable endogène.
- **Consistance** : fonctions liées au biais des variables minimum (min-features bias [AD91]). Ces méthodes recherchent l'ensemble de variables le plus petit qui satisfait un pourcentage d'inconsistance minimum défini par

l'utilisateur. (Deux objets sont dits inconsistants si leurs modalités sont identiques et s'ils appartiennent à deux classes différentes.) Ces mesures peuvent permettre de détecter les variables redondantes.

– **Précision** : ces méthodes utilisent le classifieur comme fonction d'évaluation. Le classifieur choisit, parmi tous les sous-ensembles de variables, celui qui est à l'origine de la meilleure précision prédictive.

5.1.6 Critère d'Arrêt

Il existe deux types de critères d'arrêt selon que celui ci est associé à la stratégie de recherche ou à la fonction d'évaluation :

– Un critère d'arrêt associé à une stratégie de recherche se base soit :
  – sur un nombre pré-défini de variables à sélectionner,
  – sur un nombre d'itérations pré-fixé.
– Un critère d'arrêt associé à une fonction d'évaluation se base soit sur le fait que:
  – l'ajout ou la suppression d'une variable ne produit aucun sous-ensemble plus performant,
  – le sous-ensemble obtenu est, d'après certaines fonctions d'évaluation, le sous-ensemble optimal.

5.1.7 Approches Filtres

Le filtrage est un processus de pré-traitement des données par filtrage des variables non pertinentes avant que n'intervienne la phase d'induction. Il utilise les caractéristiques générales de l'ensemble d'apprentissage pour sélectionner certaines variables et en exclure d'autres. Le schéma le plus simple est d'évaluer individuellement chaque variable grâce à une fonction d'évaluation et de sélectionner les variables possédant les plus grandes valeurs. La fonction d'évaluation est, la plupart du temps, sous la forme d'un critère nommé critère de sélection.

1. ***Critères de Sélection***
   Il convient de distinguer les critères issus de l'approche statistique de ceux basés sur la comparaison par paires d'objets.

   Approche statistique :
   Il existe deux types de mesures : les mesures myopes et les mesures contextuelles.

   – **Les mesures myopes** sont des estimateurs de la qualité d'une variable hors du contexte des autres variables explicatives. Elles sont

inadaptées pour les algorithmes traitant des données contenant des variables corrélées. Il en existe 3 catégories :

- Les mesures liées à l'information : L'entropie de Shannon [Sha48], le gain d'information , le ratio du gain [Qui86], le gain normalisé [JKSK97], la distance de De Mantaras [DM91],
- les mesures liées au critère de distance (la distance existante entre les distributions de probabilités des classes est considérée et ainsi la séparabilité des classes est évaluée) : le critère de Gini [BFOS84], le critère ORT [FI92].
- Les mesures liées au critère de dépendance, ces mesures calculent l'écart à l'indépendance de deux variables d'un tableau de contingence, le critère du khi2 (Pearson 1904), le critère de Tschuprow [Har84][Min87], le coefficient de Cramer.

– **Les mesures contextuelles** estiment la qualité d'une variable descriptive dans le contexte des autres variables descriptives. Ces mesures sont plus coûteuses mais permettent de découvrir des dépendances indécelables par les mesures myopes, la mesure la plus connue est le critère heuristique-statistique τ de Zhou [ZD91].

Comparaisons par Paires
L'idée fondamentale des comparaisons par paires est à attribuer à Condorcet dès 1785 et consiste en la comparaison des partitions induites par deux variables catégorielles, paires d'objets à paires d'objets [Ken39]. Ces mesures sont moins nombreuses que les mesures statistiques. Elles se décomposent également en mesures myopes et mesures contextuelles.

- **Mesures myopes** : Critère de Condorcet [Mic82], critère de Zahn [Zha64], critère de l'écart à l'indépendance [Mar84a][Mar84b].
- **Mesures contextuelles** : le mérite contextuel [Hon94], Relief [KR92a] [KR92b] (ici, le critère à lui seul n'est pas contextuel car une variable est estimée en dehors du contexte des autres variables. Mais, l'algorithme dans lequel il s'intègre le rend contextuel.).

Critères Liés aux Paires de Concepts :
Ces critères, décrit dans [ZKV94], sont basés sur les paires de concepts. Ils travaillent sur les représentants respectifs des concepts et tentent de les discriminer.

2. ***Présentation des Méthodes Filtre***
Afin de répertorier les différents algorithmes de filtrage, deux axes ont été utilisés : la stratégie de recherche et le critère de sélection utilisés. Nous listons maintenant diverses méthodes de SdV, cette présentation est organisée selon les deux axes précités.

Approches Complètes

- **Critère de distance :** message de description de longueur minimale (MDML)[SJL90]
- **Critère de consistance :** PRESET [Mod93], FOCUS [AG92], FOCUS2 [AG94], les méthodes Branch and Bound [NF77], ces dernières utilisent un critère de sélection caractérisé par une propriété de monotonicité (tout sous ensembles de variables possède une valeur du critère de sélection moins bonne ou similaire à celle des sous ensembles l'incluant) qui permet l'élimination, a priori, de certains sous ensembles de variables par utilisation des méthodes Branch and Bound. ABB [LMD98], par exemple, utilise ce principe et un critère d'inconsistance, sa complexité est en $O(2^p n)$.

Approches Heuristiques:

- **Critère d'information :** GIM [AG94], algorithme basé sur la couverture de markov [KS96], Cardie [Car93], MIFS [Bat94]. Cette dernière méthode utilise l'information mutuelle pour évaluer l'information contenue dans chaque variable pour sélectionner le sous ensemble de variable le plus informatif par détermination du sous ensemble de variables qui possède la plus grande information mutuelle avec la variable endogène tout en minimisant l'information mutuelle existant entre les variables exogènes sélectionnées. La sélection des variables est ici effectuée séquentiellement (processus de forward sélection).
- **Critère de distance** : GS [AG94],
- **Critère d'indépendance :** algorithme du khi2 [LS95], CFS [Hal00b] qui utilise le coefficient de Kvalseth (incertitude symétrique) pour mesurer la liaison entre variables associé classiquement à une recherche du type Best First, G3 [LR00],
- **Critère de consistance :** GS pondéré [AG94], Relief [KR92a] qui est un algorithme itératif basé sur la pondération des variables et inspiré des algorithmes d'apprentissage par instances. Relief possèdent plusieurs évolutions (Relief A, Relief B, Relief C, Relief D, Relief E, Relief F), la plus intéressante étant certainement Relief F [Kon94], qui permet de traiter des problèmes multi-classes. La complexité de Relief (et de ces différentes versions) est en $O(Ipn)$, avec n le nombre d'objets et I le nombre d'itérations.

Approches Aléatoires :
LVF [LS96] est une version filtre des algorithmes « Las Vegas ». Ces derniers font des choix probabilistes qui les mènent plus rapidement vers

une solution correcte. Un certain type de ces algorithmes utilise la stratégie aléatoire pour guider leur recherche de telle manière qu'une solution correcte est garantie même si des choix non judicieux ont été réalisés. LVF utilise un critère de sélection basé sur l'inconsistance qui spécifie jusqu'à quel point la réduction de la dimension des données peut être acceptée, le seuil d'acceptation fixé par l'utilisateur, constitue le critère d'arrêt. Cet algorithme trouve rapidement une solution proche de l'optimum.

Méthodes à base d'algorithmes génétiques
Les algorithmes génétiques (AG) sont des stratégies de recherche basées sur le principe de sélection naturelle. Une population de solutions possibles nommées chromosomes est maintenue. Les chromosomes sont sélectionnés, croisés et mutés dans le but de faire évoluer une nouvelle population. Le processus est répété jusqu'à ce qu'une condition d'arrêt soit atteinte pour l'individu le plus adapté de la population ou quand un certain nombre de générations a été produit. La capacité à effectuer des recherches dans des espaces très grands et sans connaissance sur le domaine, ainsi que la relative insensibilité bruit des AGs sont connus. Ils apparaissent donc idéaux pour des usages où les connaissances du domaine et les théories sont difficiles voire impossibles à obtenir. Lors de l'utilisation des AGs, la chose la plus importante est de choisir une représentation bien appropriée et une fonction d'évaluation adéquate. Les AGs ont été utilisés, dans le cadre de la SdV pour l'apprentissage supervisé, par de nombreux auteurs tels que Guerra-Salcedo [GSCWS99], Freitas [Fre02], Whitley et Smith [GSCWS99], Vafaie et De Jong [VJ92], [VDJ93], [VDJ94], Yang [YPH97], [YH98]... Notons que, dans la majorité des cas d'utilisation d'AG, l'approche de SdV n'est pas de type filtre mais de type enveloppe.
Le schéma de la figure 5.4 (page 125) résume le processus de sélection de variables par les AG pour une approche filtre.

L'avantage des approches filtres est leur capacité à être utilisées en amont de n'importe quel algorithme d'induction due à leur indépendance vis à vis de celui ci. Cependant, elles ignorent totalement les effets du sous-ensemble de variables sélectionnées sur les performances de cet algorithme.

5.1.8 Approches Enveloppes

Ces approches ont été introduites par John, Kohavi et Plfeger [JKP94]. Pour ces auteurs, les algorithmes de filtrage ne sont pas toujours efficaces car ils ignorent totalement l'influence de l'ensemble de variables sélectionnées sur

les performances de l'algorithme d'induction. Pour résoudre ce problème, ils proposent une approche différente qui utilise le résultat de l'algorithme d'apprentissage comme fonction d'évaluation : « les méthodes enveloppes ». L'algorithme d'induction appliqué aux données pré-traitées est utilisé comme un sous-programme et considéré comme une boite noire par cet ensemble de méthodes. L'algorithme d'induction travaille sur l'ensemble de données avec différents sous-ensembles de variables et fournit pour chacun d'eux la précision estimée sur le classement de nouvelles instances (cf. figure 5.2 à droite). Le sous-ensemble induisant le classifieur le plus précis est ensuite retenu pour la tache d'induction. Les méthodes enveloppe consistent donc à estimer le taux de succès (par cross-validation) en utilisant uniquement les variables du sous-ensemble à évaluer. D'autres auteurs ont repris cette méthode en utilisant d'autres approches. Parmi ces méthodes, on peut citer : La méthode Oblivion de Langley et Sage [LS94], la méthode BEAM de Aha et Bankert [AB95], CAP de Caruana et Freitag [CF94], NLC [RRJ03], Doak [Doa92], Race [ML94].

Le désavantage majeur des méthodes enveloppe est le coût calculatoire important dû à l'appel de l'algorithme d'induction pour chaque sous-ensemble considéré.

5.1.9 Autres Approches

D'autres approches plus ou moins usitées peuvent être employées nous pouvons notamment citées les méthodes issues de l'économétrie et de l'analyse factorielle. Enfin, les méthodes basées sur les support vectors machines constituent une approche suscitant actuellement un intérêt très vif, le très récent numéro spécial de la revue *Journal of Machine Learning Research* [GE03] témoigne parfaitement de cet intérêt et introduit cette problématique tout en proposant plusieurs méthodes participant des approches filtre et/ou enveloppe. Méthodes de sélection de variables utilisant les SVM (Méthode de Stoppiglia [SD03] (classement des variables selon leur pertinence), l'algorithme SVM-RFE [Rak03], l'algorithme VS-SSVM [BBMES03] (algorithmes de Backward Sélection)...).

Méthodes	Auteur et Année	Types	Dir. de Recherche	Strat. de Recherche	Critère de Sélection	Critère d'Arrêt
Branch and Bound	Narandra et Fukunaga, 1977	Filtre	Backward	Complète	Consistance	Nb. Itérations
MDLM	Scheinvald, 1990	Filtre	Forward	Complète	Distance	Nb. Itérations
Focus	Almuallim et Dietterich, 1991	Filtre	Forward	Complète	Consistance	Sous-Ens. Optimal
Relief	Kira et Rendell, 1992	Filtre	Autre	Heuristique	Consistance	Seuil Seuil
Focus 2	Almuallim et Dietterich, 1992	Filtre	Forward	Complète	Consistance	Sous-Ens. Optimal
Preset	Modrzejewski, 1993	Filtre	Autre	Complète	Consistance	Sous-Ens. Optimal
Sélection et AG	Vafaie et De Jong, 1993	Filtre	Autre	Aléatoire	Consistance	Nb. Itérations
GIM	Almuallim, 1994	Filtre	Forward	Heuristique	Information	Sous-Ens. Optimal
MIFS	Battiti, 1994	Filtre	Forward	Heuristique	Information	Sous-Ens. Optimal
GS	Almuallim, 1995	Filtre	Forward	Heuristique	Distance	Pas d'amélioration
GP	Almuallim, 1996	Filtre	Forward	Heuristique	Consistance	Sous-Ens. Optimal
Relief-F	Kononenko, 1994	Filtre	Autre	Heuristique	Consistance	Seuil
Khi2	Liu, 1995	Filtre	Autre	Heuristique	Indépendance	Sous-Ens. Optimal
LVF	Liu, 1996	Filtre	Autre	Aléatoire	Consistance	Nb. Itérations
Sél. & couv. de Markov	Koller, 1996	Filtre	Backward	Heuristique	Information	Nb. Itérations
CFS	Hall, 2000 2000	Filtre	Forward	Heuristique	Indépendance	Sous-Ens. Optimal
G3	Rakotomalala et Lallich, 2000	Filtre	Forward	Heuristique	Indépendance	Sous-Ens. Optimal

TAB. 5.1 –: *Tableau récapitulatif (Partie 1) inspiré de l'exposé de l'article [LEB02]*

Méthodes	Auteur et Année	Types	Dir. de Recherche	Strat. de Recherche	Critère de Sélection	Critère d'Arrêt
Méthode de Doak	Doak, 1992	Env.	Autre	Heuristique	Précision	Sous-Ens. Optimal
CAP	Caruana et Freitag, 1994	Env.	Autre	Complète	Précision	Sous-Ens. Optimal
RACE	Moore et Lee	Env.	Autre	Complète	Précision	Sous-Ens. Optimal
Méthode de John	John, 1994	Env.	Forward	Complète	Précision	Pas d'amélioration
Oblivious	Langley et Sage, 1994	Env.	Backward	Complète	Précision	Sous-Ens. Optimal
Régressions	Johnston, 1988	Env.	Forward	Complète	Consistance	
Backward Elimination	Johnston, 1989	Env.	Backward	Heuristique	Consistance	Seuil
Forward Selection	Johnston, 1990	Env.	Forward	Heuristique	Consistance	Seuil Seuil
Stepwise Regression	Johnston, 1991	Env.	Forward	Heuristique	Consistance	Seuil Seuil
Stagewise Regression	Johnston, 1992	Env.	Forward	Heuristique	Consistance	Seuil Seuil
BEAM	Aha et Bankert, 1996	Env. .	Autre	Heuristique	Précision	Sous-Ens. Optimal
NLC	Ruiz, 2003	Env.	Forward	Heuristique	Précision	Sous-Ens. Optimal
SVM SVM	Stoppiglia, 2003	Filtre	Forward	Heuristique	Distance	Seuil

TAB. 5.2 –: *récapitulatif (Partie 2) inspiré de l'exposé de l'article [LEB02]*

5.2 Contribution à la Sélection de Variables pour l'Apprentissage Supervisé : Une Nouvelle Méthode Efficiente et Rapide

Nous proposons maintenant une nouvelle méthode efficiente et rapide pour la sélection de variables dans le cadre de l'apprentissage supervisé sur variables catégorielles[4]. Cette méthode de type filtre ne requiert qu'une unique passe sur le jeux de données (ce qui lui confère un avantage calculatoire important sur la plupart des méthodes). Elle utilise en fait un algorithme génétique (AG) et la méthodologie d'évaluation/comparaison de la validité de cns (cf. chapitre 4) au sein d'un processus itératif pour la sélection d'un sous-ensemble de variables .

Dans les sections suivantes nous considérons un problème d'apprentissage caractérisé par un ensemble $O = \{o_1,...,o_n\}$ de n objets décrits par :

- un espace de représentation des données EV comprenant p variables catégorielles $EV = \{V_1,...,V_p\}$. ($o_i = \{o_{i_1},...,o_{i_p}\}$)
- une variable catégorielle V_A qui représente le concept à apprendre (variable endogène) possédant k modalités.

Nous utilisons un problème d'apprentissage synthétique pour illustrer nos développements (voir table 5.3) : $O = \{o_1,o_2,o_3,o_4,o_5,o_6,o_7\}$, $EV = \{V_1,V_2,V_3,V_4\}$, V_A a 3 modalités a,b,c.

	V_1	V_2	V_3	V_4	V_A
o_1	o	o	o	o	a
o_2	o	o	n	o	a
o_3	o	n	o	o	a
o_4	n	o	n	o	b
o_5	n	o	o	o	b
o_6	n	o	n	n	c
o_7	n	n	o	n	c

TAB. 5.3 –: *Jeu de données synthétiques*

5.2.1 Hypothèses et Idées Fondamentales

Nous donnons maintenant les idées et hypothèses qui constituent les bases de la méthode que nous proposons :

1. ***Hypothèse** : Si l'ERD, EV, d'un problème d'apprentissage est tel que le concept à apprendre implique une structure naturelle de l'ensemble des objets O dans cet ERD, alors cela doit permettre un bon processus d'apprentissage.*

4. Cette méthode peut également être employée dans le cadre de données quantitatives mais son coût calculatoire s'accroît alors grandement la rendant moins attrayante...

2. ***Hypothèse*** *: Une cns valide de l'ensemble des objets O correspond à une structure naturelle de O.*
3. ***Hypothèse*** *: Sur la base de 1 et 2, on peut admettre que si l'ERD EV d'un problème d'apprentissage est tel que le concept à apprendre implique une organisation des objets de O selon une cns valide dans cet espace EV, alors l'ERD EV doit autoriser un bon processus d'apprentissage.*
4. ***Idée*** *: Dans le cadre de la sélection de variables, nous pouvons considérer que l'ERD $EV_\star \subseteq EV$, constitué des variables sélectionnées pour l'apprentissage, doit être tel que le concept à apprendre implique une organisation des objets de O selon une cns valide dans l'espace $EV_\star$.*
5. ***Hypothèse*** *: La cns de l'ensemble d'objets O impliquée par le concept à apprendre est composée d'autant de classes qu'il existe de modalités du concept à apprendre, et chaque classe est exclusivement composée d'objets correspondant à la même modalité du concept à apprendre.(Cette cns est notée par la suite P)*
   Pour le problème d'apprentissage servant d'exemple :
   $P = \{\{o_1,o_2,o_3\},\{o_4,o_5\},\{o_6,o_7\}\}$.
6. ***Idée*** *: Dans le cadre de la sélection de variables, si nous considérons l'ensemble de tous les ERDs potentiels (ces espaces sont les sous espaces non vides de EV), l'ERD que l'on sélectionne finalement (i.e. l'ERD constitué des variables sélectionnées pour l'apprentissage) doit être tel que la cns P apparaît comme la plus valide au sein de ce sous espace de EV.*

5.2.2 Evaluation de la Validité d'une Partition dans un Sous-Espace de l'ERD

La notion de validité d'une partition $P = \{C_1,...,C_k\}$ dans un sous espace de l'ERD $EV_\star$ est définie de manière identique à la validité d'une partition dans l'ERD complet EV : on utilise la méthode présentée au chapitre précédent. Le point important est de ne prendre en compte non pas l'ensemble des variables de EV mais uniquement celles de $EV_\star$. Nous utilisons donc les indices $LM_{EV_\star}(P)$, $NLM_{EV_\star}(P)$, $LD_{EV_\star}(P)$, $NLD_{EV_\star}(P)$:

$$LM_{EV_\star}(P) = \sum_{g=1..k} \sum_{\substack{o_a \in C_g, o_b \in C_g, \\ a<b}} \sum_{\substack{i \text{ tel que} \\ V_i \in EV_\star}} (lien_i(o_{a_i}, o_{b_i}))$$

$$NLM_{EV_\star}(P) = \sum_{g=1..k} \sum_{\substack{o_a \in C_g, o_b \in C_g, \\ a<b}} \sum_{\substack{i \text{ tel que} \\ V_i \in EV_\star}} (1 - lien_i(o_{a_i}, o_{b_i}))$$

$$LD_{EV_\star}(P) = \sum_{\substack{f=1..k, g=1..k \\ f<g}} \sum_{o_a \in C_f, o_b \in C_g,} \sum_{\substack{i \text{ tel que} \\ V_i \in EV_\star}} (lien_i\,(o_{a_i}, o_{b_i}))$$

$$NLD_{EV_\star}(P) = \sum_{\substack{f=1..k, g=1..k \\ f<g}} \sum_{o_a \in C_f, o_b \in C_g,} \sum_{\substack{i \text{ tel que} \\ V_i \in EV_\star}} (1 - lien_i\,(o_{a_i}, o_{b_i}))$$

Ce qui nous permet de définir les indices supplémentaires :

- $L_{EV_\star}(O) + NL_{EV_\star}(O) = \frac{n\times(n-1)}{2} \times card(EV_\star)$
- $D_{EV_\star}(P) + M_{EV_\star}(P) = \frac{n\times(n-1)}{2} \times card(EV_\star)$
- $M_{EV_\star}(P) = NLM_{EV_\star}(P) + LM_{EV_\star}(P)$
- $D_{EV_\star}(P) = NLD_{EV_\star}(P) + LD_{EV_\star}(P)$
- $L_{EV_\star}(O) = LM_{EV_\star}(P) + LD_{EV_\star}(P)$
- $NL_{EV_\star}(O) = NLM_{EV_\star}(P) + NLD_{EV_\star}(P)$

Ainsi, nous utiliserons les indices $xv_1^{EV_\star}(P)$ et $xv_2^{EV_\star}(P)$ pour caractériser la validité de P au sein de $EV_\star$:

- $xv_1^{EV_\star}(P) = \frac{LM_{EV_\star}(P) - M_{EV_\star}(P) \times \frac{L_{EV_\star}(O)}{L_{EV_\star}(O)+NL_{EV_\star}(O)}}{\sqrt{M_{EV_\star}(P) \times \frac{L_{EV_\star}(O)}{L_{EV_\star}(O)+NL_{EV_\star}(O)} \times (1 - \frac{L_{EV_\star}(O)}{L_{EV_\star}(O)+NL_{EV_\star}(O)})}}$
  $xv_1^{EV_\star}(P) \hookrightarrow N\,(0,1)$
- $xv_2^{EV_\star}(P) = \frac{NLD_{EV_\star}(P) - D_{EV_\star}(P) \times \frac{NL_{EV_\star}(O)}{L_{EV_\star}(O)+NL_{EV_\star}(O)}}{\sqrt{D_{EV_\star}(P) \times \frac{NL_{EV_\star}(O)}{L_{EV_\star}(O)+NL_{EV_\star}(O)} \times (1 - \frac{NL_{EV_\star}(O)}{L_{EV_\star}(O)+NL_{EV_\star}(O)})}}$
  $xv_2^{EV_\star}(P) \hookrightarrow N\,(0,1)$

Ainsi, on peut comparer la validité de P dans différents sous espaces de EV, et ce en suivant un mode de comparaison identique à celui présenté dans le chapitre 4.

5.2.3 La Nouvelle Méthode de Sélections de Variables

Nous montrons maintenant comment utiliser la méthodologie de comparaison de validité de cns (présentée au chapitre 4) en l'associant à un AG pour

bâtir une méthode de sélection de variables pour l'apprentissage supervisé basée sur les idées et hypothèses émises précédemment.

5.2.3.1 La Méthode de Base : une Méthode Exhaustive

L'idée de base de cette méthode est de considérer la partition P et de tester la validité de cette cns dans chaque sous espace de EV. Étant données les hypothèses précédemment émises, l'ERD sélectionné (ou l'ensemble des sous-espaces sélectionnés) est le sous espace (ou l'ensemble des sous-espaces) impliquant la plus forte validité pour P. La méthode peut être traduite par l'algorithme 4.

Ce processus requiert une unique passe sur les données pour obtenir toutes les tables de contingence utiles (qui ne nécessitent qu'une faible capacité de stockage), $2^p - 1$ calculs pour tester chaque sous espace non vide de EV, et $2^p - 1$ comparaisons pour déterminer le meilleur sous espace (ou l'ensemble des meilleurs sous espaces). (voir le chapitre précédent et l'exemple suivant l'algorithme 4 pour des informations complémentaires sur le coût calculatoire.) Si le nombre de variables p est faible, l'utilisation de cette méthode est envisageable (car réalisable du point de vue calculatoire), mais pour des nombres de variables un peu plus élevés l'utilisation de cette méthode n'est pas envisageable du point de vue calculatoire. Nous devons alors adopter une heuristique pour déterminer le meilleur sous espace, ou au moins, un bon sous espace, sans pour autant utiliser une phase de test exhaustive et ainsi limiter le coût calculatoire de la méthode. Nous avons choisi d'adopter les algorithmes génétiques (AGs) qui sont connus comme une solution efficace pour la résolution de problèmes combinatoires.

Algorithme 4 SdV pour l'Apprentissage Supervisé : Méthode Exhaustive

1. Données : la partition P, l'ERD EV

2. En une unique passe sur les données bâtir les tables de contingence nécessaires aux calculs des mesures de validité nécessaires à la méthodologie d'évaluation/comparaison de la validité de cns présentée préalablement (i.e. les tables de contingence croisant la variable endogène et exogènes).

3. En utilisant les informations de ces tables de contingence, calculer les valeurs des 2 mesures de validité $xv_1^{EV_\star}$ et $xv_2^{EV_\star}$ pour la cns P dans chaque sous espace non vide $EV_\star$ de EV puis comparer la validité de la cns P dans chacun de ces sous espaces.

4. Cette comparaison permet la sélection du sous-espace dans lequel (ou de l'ensemble des sous-espaces) dans lequel (ou lesquels) P apparaît la plus valide. Ce sous espace (ou cet ensemble de sous espaces) constitue alors le sous espace sélectionné (ou l'ensemble des sous-espaces sélectionnés).

EXEMPLE : *Pour le jeu de données synthétique illustratif, appliquer cette méthode exhaustive revient tout d'abord à calculer les valeurs pour* $xv_1^{EV_\star}(P)$ *et* $xv_2^{EV_\star}(P)$ *dans chaque sous-espace de* $EV = \{V_1,V_2,V_3,V_4\}$*. Ces valeurs sont reportées dans la table 5.4 et la figure 5.3 présente graphiquement ces valeurs.*[5]

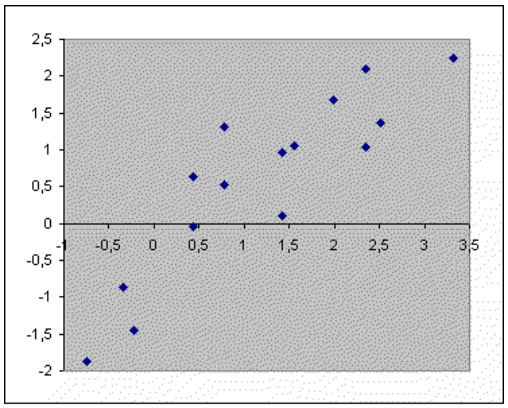

FIG. 5.3 –: *Valeurs* $xv_1^{EV_\star}(P)$ *et* $xv_2^{EV_\star}(P)$ *dans chaque sous-espace de* $EV = \{V_1,V_2,V_3,V_4\}$

Nous rappelons une nouvelle fois que le calcul de l'ensemble de ces valeurs n'a nécessité qu'une seule passe sur les données :

– *cette unique passe permet d'obtenir les tables de contingence croisant la variable endogène* V_A *avec les variables exogènes.*

$V_A \setminus V_1$	o	n	$V_A \setminus V_2$	o	n	$V_A \setminus V_3$	o	n	$V_A \setminus V_4$	o	n
a	3	0	a	2	1	a	2	1	a	3	0
b	0	2	b	2	0	b	1	1	b	2	0
c	0	2	c	1	1	c	1	1	c	0	2

Tables de Contingence croisant la variable endogène V_A *et les variables exogènes*

– *le calcul de chaque valeur est alors réalisé à partir de ces tables : si la table de contingence pour une variable* V_i *est notée:*

$V_A \backslash V_i$	V_{i_1}	...	$V_{i_{m_i}}$	
V_{A_1}	α_{1i1}	...	α_{1im_i}	$\alpha_{1i.}$
...	...	...	...	...
V_{A_k}	α_{ki1}	...	α_{kim_i}	$\alpha_{ki.}$
	$\alpha_{.i_1}$	...	$\alpha_{.im_i}$	n

V_A la variable endogène à k modalités,
V_i une variable exogène à m_i modalités notées $V_{i_j} (j = 1..m_i)$.
α_{lih} le nombre d'objets ayant la valeur V_{i_h} pour V_i et la valeur V_{A_l} pour V_A.
$\alpha_{.i_j} = \sum_{h=1..k} \alpha_{hi_j}$
$\alpha_{hi_.} = \sum_{j=1..m_i} \alpha_{hi_j}$

5. Attention, cet exemple vise essentiellement à illustrer les différentes étapes de la méthodologie, en effet, étant donné le faible nombre d'individus du jeu de données l'approximation normale est ici douteuse...

	$xv_1(P)^{EV_\star}$	$xv_2(P)^{EV_\star}$	fit_1	rang fit_1	fit_2	rang fit_2	f_1	f_2
$V_\star = \{V_1\}$	2,35	2,09	3,14	2	11	2	11	11
$V_\star = \{V_2\}$	-0,34	-0,87	0	12	14,14	12	0	14,14
$V_\star = \{V_3\}$	-0,22	-1,46	0	12	14,14	12	0	14,14
$V_\star = \{V_4\}$	2,35	1,03	2,57	5	11,79	5	11,58	11,79
$V_\star = \{V_1,V_2\}$	1,42	0,96	1,71	7	12,46	7	12,43	12,46
$V_\star = \{V_1,V_3\}$	0,78	1,32	1,53	8	12,66	8	12,61	12,66
$V_\star = \{V_1,V_4\}$	3,32	2,24	4	1	10,24	1	10,14	10,24
$V_\star = \{V_2,V_3\}$	-0,75	-1,87	0	12	14,14	12	0	14,14
$V_\star = \{V_2,V_4\}$	1,42	0,11	1,42	9	13,09	9	12,72	13,09
$V_\star = \{V_3,V_4\}$	0,78	0,53	0,94	10	13,22	10	13,2	13,22
$V_\star = \{V_1,V_2,V_3\}$	0,44	0,64	0,78	11	13,38	11	13,37	13,38
$V_\star = \{V_1,V_2,V_4\}$	2,51	1,37	2,86	3	11,43	3	11,28	11,43
$V_\star = \{V_1,V_3,V_4\}$	1,99	1,67	2,6	4	11,56	4	11,54	11,56
$V_\star = \{V_2,V_3,V_4\}$	0,44	-0,05	0	12	14,14	12	0	14,14
$V_\star = \{V_1,V_2,V_3,V_4\}$	1,56	1,05	1,88	6	12,3	6	12,26	12,3

TAB. 5.4:

on peut alors calculer :

$$LM(P) = \sum_{\substack{i=1..p \text{ tel que} \\ V_i \in EV_\star}} \sum_{j=1..m_i} \sum_{z=1..k} \frac{\alpha_{zi_j}(\alpha_{zi_j}-1)}{2}$$

$$M(P) = card(EV_\star) \times \sum_{z=1..k} \frac{card(C_z)(card(C_z)-1)}{2}$$

$$L(O) = \sum_{\substack{i=1..p \text{ tel que} \\ V_i \in EV_\star}} \sum_{j=1..m_i} \frac{\alpha_{.i_j}(\alpha_{.i_j}-1)}{2}$$

$$NLM(P) = M(P) - LM(P)\,;\, LD(P) = L(O) - LM(P)$$

$$D(P) = \frac{n(n-1)}{2} \times card(EV_\star) - M(P)\,;\, NLD(P) = NV(P) - LD(P)$$

– *Puis, on utilise la méthodologie du chapitre 4 pour déterminer le sous-espace de EV dans lequel P apparaît comme la plus naturelle, il s'agit ici de $V_\star = \{V_1,V_4\}$ (notons au passage que la conjonction de ces deux variables correspond effectivement au concept le plus simple permettant une discrimination parfaite des 3 modalités de la variable endogène AL).*

5.2.3.2 Réduction de la Complexité par Introduction d'un AG

Le problème auquel nous sommes confronté, la découverte d'un bon sous-espace sans pour autant pratiquer une recherche exhaustive, peut effectivement être résolu efficacement par utilisation d'un AG de la manière suivante :

- chaque chromosome de l'AG correspond à un sous espace de EV qui est caractérisé par la présence/absence de variables de EV ;
- chaque chromosome possède p gènes, chaque gène correspond à l'une des p variables de EV, un gène a une valeur binaire (un gène est codé sur un seul bit) qui code la présence/absence de la variable dans le sous espace de EV codé par le chromosome ;
- la fonction de fitness de l'AG est basée sur la méthodologie pour l'évaluation/comparaison de la validité de cns. Toutefois dans la mesure où cette fonction de fitness doit permettre de comparer tout couple de sous espaces (i.e. selon cette fonction il n'existe pas de couple de sous espaces tel que la fonction de fitness ne puisse comparer la validité de P dans ces 2 sous espaces), la fonction de fitness correspond à une adaptation de la méthodologie proposée préalablement.
- pour le reste, l'AG est utilisé et défini de manière classique.

L'algorithme 5 ainsi que la figure 5.4 illustrent le fonctionnement de la méthode de sélection de variables que nous proposons.

Algorithme 5 Sélection de Variables pour l'Apprentissage Supervisé : Utilisation d'un AG

---

1. Données : la cns P, l'ERD EV

2. En une unique passe sur les données bâtir les tables de contingence nécessaires aux calculs des mesures de validité nécessaires à la méthodologie d'évaluation/comparaison de la validité de cns présentée préalablement.

3. Fixer les paramètres de l'AG : *nombre de générations, taille de la population, Probabilité de Croisement, Probabilité de mutation*

4. Lancer l'AG utilisant la fonction de fitness spécifique définie par la suite.

5. Sélectionner le meilleur sous-espace déterminé par l'AG

Concernant la définition de la fonction de fitness de l'AG, un problème est effectivement soulevé : la méthodologie utilisée pour l'évaluation/comparaison de la validité de cns implique une optimisation multi-objectif (elle nécessite la comparaison de couples de valeurs et peut mener à des comparaisons impossibles de sous espaces) qui s'avèrent problématique pour l'utilisation d'AGs. L'utilisation d'AGs multi-objectif qui impliquent pour la plupart des mécanismes de sélection coûteux du point de vue calculatoire, constitue une solution éventuelle, nous lui préférons la solution consistant à dériver une fonc-

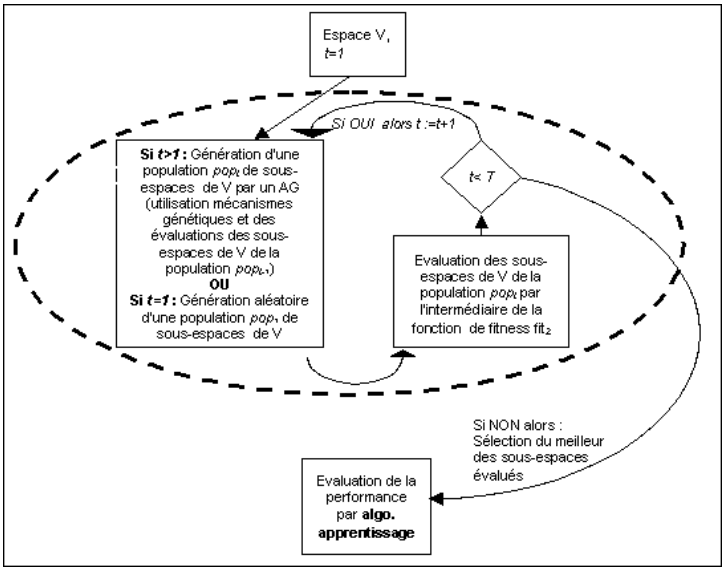

FIG. 5.4 –: *schéma fonctionnel de la méthode proposée*

tion de fitness telle qu'elle intègre en une unique fonction objectif les deux fonctions objectifs.

Nous proposons maintenant deux fonctions basées sur l'observation que P doit présenter de fortes valeurs pour $xv_1^{EV_\star}$ et $xv_2^{EV_\star}$ pour être considérée comme valide dans $EV_\star$. Ces fonctions $fit1(P,EV_\star)$ et $fit2(P,EV_\star)$ (avec $EV_\star$ le sous espace considéré) sont les suivantes :

1. $fit1(P,EV_\star) = \begin{cases} \sqrt{(xv_{1_P}^{EV_\star})^2 + (xv_{2_P}^{EV_\star})^2} \text{ si } xv_{1_P}^{EV_\star} > 0 \text{ et } xv_{2_P}^{EV_\star} > 0 \\ 0 \text{ sinon} \end{cases}$
   qui correspond en quelque sorte à une distance du point de vue de la validité entre une structure ne constituant pas une cns valide (ou encore la cns P dans un sous espace de EV tel que cette cns n'apparaisse pas comme valide) et la cns P dans le sous espace $EV_\star$. Cette fonction de fitness doit être maximisée.

2. $fit2(P,EV_\star) = \begin{cases} \sqrt{(\tilde{x_1} - xv_{1_P}^{EV_\star})^2 + (\tilde{x_2} - xv_{2_P}^{EV_\star})^2} \text{ si } xv_{1_P}^{EV_\star} > 0 \text{ et } xv_{2_P}^{EV_\star} > 0 \\ +\infty \text{ sinon} \end{cases}$
   qui correspond en quelque sorte à une distance du point de vue de la validité entre une cns virtuelle particulière (dont les valeurs xv_1 et xv_2 seraient respectivement $\tilde{x_1}$ et $\tilde{x_2}$) et la cns P. En fait, dans ce cas, nous fixons $\tilde{x_1} = \tilde{x_2} = très\ forte\ valeur$ de manière à conférer à la cns virtuelle particulière l'aspect d'une sorte de cns idéale du point de vue de la validité (ou encore la validité de P dans un espace tel qu'il confère à P une validité idéale). Ainsi, cette dernière fonction de fitness correspond en somme à une distance du point de vue de la validité entre une cns

virtuelle idéale du point de vue de la validité et la cns P. Cette fonction de fitness doit donc être minimisée.

Les tests réalisés ont montré que la seconde fonction de fitness est la plus intéressante car elle mène à des cns possédant des valeurs équilibrées pour xv_1 et xv_2 contrairement à la première fonction de fitness qui peut mener à des valeurs non équilibrées (i.e. une très forte valeur pour l'une des 2 valeurs et faible pour l'autre). Cela s'explique notamment par la forme de ces deux fonctions, la figure 5.5 présentant respectivement les surfaces impliquées par la fonction $f_1(x,y) = \left(max_{x\in[0;10],y\in[0;10]}(\sqrt{x^2+y^2})\right) - \sqrt{x^2+y^2}$ et par la fonction $f_2(x,y) = \sqrt{(10-x)^2+(10-y)^2}$ permet d'appréhender cela de manière intuitive.

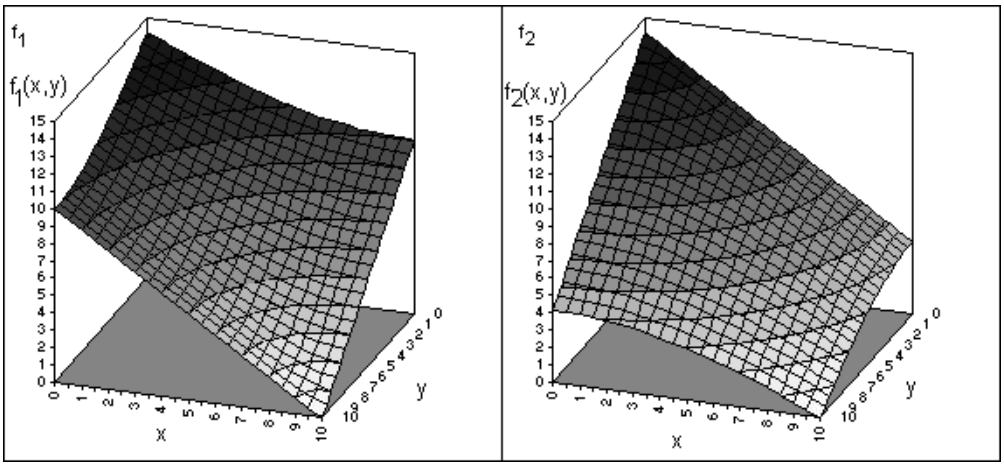

FIG. 5.5 –: *fonctions f_1 et f_2*

5.2.4 Evaluation Expérimentale

5.2.4.1 Présentation de l'Evaluation Expérimentale

L'évaluation expérimentale, réalisée sur 17 jeux de données de la collection de l'université de Californie à Irvine [MM96][6], a mis en jeu cinq méthodes d'apprentissage différentes : ID3, C4.5, Sipina, 1-plus proche voisins et bayésiens naïfs et utilisée des espaces de représentation respectivement issus de processus de SdV préalables réalisés par les algorithmes ReliefF, CFS, MIFS

6. Ces jeux de données sont les jeux : GERMAN, MUSHROOMS (noté MUSH.), SICK, VEHICLE, ADULT, MONKS 3, FLAGS, BREAST CANCER (noté BREAST), ZOO, WINE, CANCER, PIMA, WAVE, CONTRACEPTION (noté CONTRA.), ION, SPAM, HOUSE-VOTES84 (noté HVOTES). Ils sont présentés en annexe (voir page 217). De plus, toutes les variables numériques ont subi un processus de discrétisation supervisée par l'intermédiaire de la méthode FUSINTER [ZRR98].

(ces 3 méthodes constituant des méthodes de référence du domaine), par notre algorithme de SdV, ou encore sans sélection préalable. Divers apprentissages ont permis la réalisation d'une étude comparative concernant d'une part le taux d'erreur des divers apprentissages selon la méthode de SdV employée et d'autre part le nombre de variables sélectionnées par chaque méthode de SdV. Cette évaluation est menée pour une 10-cross-validation ainsi que pour cinq 2-cross-validations. Notons de plus que :

- La version de CFS utilisée est telle que le critère employé est bien le critère classique (voir [Hal00b]) et la stratégie de recherche est basée sur un AG et non sur une simple approche de type best first.
- La version de MIFS employée est la version classique (critère classique, voir [Bat94], et stratégie de recherche gloutonne classique).
- La version de ReliefF employée est telle que : le critère employé est bien le critère à la fois de consistance et contextuel classique ; la stratégie de recherche utilise quant à elle un échantillon d'objets de la taille de l'ensemble des objets du jeu de données.
- CFS, MIFS et notre méthode fournissent quant à elles le sous-ensemble optimal de variables (ou un sous-ensemble l'approchant).
- ReliefF fournit la liste des variables classées selon leur pertinence (nous avons ensuite étudié cette liste de valeur afin de déterminer le sous-ensemble de variables apparemment le plus intéressant).
- l'AG utilisé pour CFS et notre méthode est une version élitiste des AGs de base, il est paramètré de la manière suivante : *nb de générations =2000, taille de la population = 30, Proba. Croisement = 0.98, Proba. mutation = 0.3* .

5.2.4.2 Analyse de l'Evaluation Expérimentale

Les résultats des expériences sont regroupés au sein des tableaux 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17. Les résultats présentés dans les 11 derniers tableaux sont présentés de manière graphique dans les figures 5.6, 5.7, 5.8, 5.9, 5.10, 5.11.

Les tableaux 5.5, 5.6, 5.7 (voir page 130) ainsi que la figure 5.6 (voir page 131) présentent des résultats généraux :

- Les tableaux 5.5, 5.6 permettent d'évaluer le comportement général des diverses méthodes d'apprentissage utilisées lorsqu'elles sont associées aux méthodes de SdV. En effet, ils présentent la valeur moyenne du rapport "taux de succès avec sélection / taux de succès sans sélection" pour chaque méthode d'apprentissage associée à chacune des méthodes de SdV, et ce, soit dans le cadre d'une 10-cross-validation (tableau 5.5), soit dans le cadre de cinq 2-cross-validations (tableau 5.6) (la moyenne est calculée sur l'ensemble des 17 jeux de données). Les résultats permettent de conclure que, de manière générale, l'ensemble des méthodes de SdV impliquent l'obtention de taux de succès quasi-équivalent lorsqu'on utilise les variables fournies par ces méthodes ou l'ensemble complet des

variables. Ainsi, quelle que soit la méthode d'apprentissage utilisée et quelle que soit la méthode de SdV utilisée, les taux de succès sont corrects et quasiment similaires. On peut toutefois noter un très léger déficit de qualité d'apprentissage pour la méthode d'apprentissage Sipina lorsqu'elle est associée à la méthode de CFS. On peut ainsi conclure que de manière générale ces 4 méthodes de SdV sont presque équivalentes du point de vue de la qualité des apprentissages qu'elles impliquent.

- Le tableau 5.7 et la figure 5.6 permettent l'évaluation de la réduction de la taille de l'ERD impliquée par l'utilisation des méthodes de SdV. Ainsi, il apparaît clairement que l'ensemble de ces méthodes permettent une réduction significative de la taille de l'ERD. De plus, il existe ici des distinctions claires entre les méthodes de SdV :
  - CFS réduit de manière générale très significativement la taille de cet espace puisqu'en moyenne elle ne conserve que 41,4% des variables. Elle constitue la méthode la plus efficace pour la réduction de l'ERD : son apparente plus grande capacité à réduire cet espace n'est mise en défaut que sur quelques rares jeux de données.
  - Notre méthode et MIFS permettent également, en général, de réduire significativement la taille de cet espace puisqu'en moyenne elles ne conservent respectivement que 56,9% et 62,6% des variables. Elles constituent, derrière CFS les méthodes les plus efficaces pour la réduction de l'ERD. Leur proximité en moyenne sur leur capacité à réduire l'ERD ne reflète cependant pas leurs comportements largement différents : selon le jeu de données, il peut arriver que l'une surpasse fortement l'autre dans sa capacité à réduire l'ERD. On peut ainsi conclure que si notre méthode semble légèrement plus efficace que MIFS de ce point de vue, il est par contre clair que ponctuellement ce résultat peut être inversé.
  - La méthode ReliefF, même si elle permet de réduire l'ERD (74,4% des variables conservées en moyenne), semble cependant en retrait par rapport aux autres méthodes.
- Du point de vue du coût calculatoire, MIFS, CFS et notre méthode nécessitent un temps de calcul proche avec un avantage toutefois à MIFS qui utilise une stratégie de recherche gloutonne contrairement aux 2 autres méthodes (temps de calcul de l'ordre de quelques secondes à la minute selon les jeux de données). En effet, les AGs sont, en principe, plus lents que les méthodes d'optimisation gloutonnes telle que celle employée dans MIFS. En fait, CFS et notre méthode pourraient être plus rapides si nous remplacions l'AG par une telle méthode d'optimisation (bien que dans ce cas nous pourrions obtenir des résultats de moindre qualité du point de vue de la correction en prédiction, nous ne pensons pas que la réduction de qualité associée soit réellement significative, et envisageons actuellement de tester cette approche...). ReliefF, par contre, implique un

temps de calcul plus important (parfois plusieurs minutes) ce qui s'explique par les multiples passes sur le jeu de données que cette méthode implique contrairement aux 3 autres méthodes.

Les tableaux 5.8, 5.9, 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 5.16, 5.17, ainsi que les figures 5.7, 5.8, 5.9, 5.10, 5.11 permettent d'appréhender de manière plus précise (sur des cas particuliers) les résultats. Les tableaux présentent sur chaque cas isolé (réalisation d'une 10-cross-validation ou de cinq 2-cross-validations pour une méthode d'apprentissage particulière) le taux d'erreur moyen en validation ainsi que l'écart-type de ce taux d'erreur. Les figures se contentent de présenter le taux d'erreur moyen en validation.

Les points les plus intéressants que l'on peut extraire d'une analyse détaillée de ces résultats sont :

- que la tendance générale de taux de correction proche pour les apprentissages réalisés avec et sans SdV est vérifiée localement,
- que CFS semble impliquer parfois des déficits importants en terme de correction et notamment lorsqu'elle sélectionne un nombre faible de variables (le cas du jeu de données MONKS 3 par exemple),
- que, tout comme pour la réduction de l'ERD, la méthode MIFS et la notre sont en général proches mais il arrive ponctuellement que l'une surpasse plus fortement l'autre.
- que, la stabilité des apprentissages est quasiment similaire pour les apprentissages sur un même jeu de données que l'on ait utilisé ou non la SdV et quelle que soit la méthode de SdV employée.

En définitive, selon nous, cette étude expérimentale tend à privilégier l'utilisation de CFS par rapport à MIFS et notre méthode et que l'on peut rejeter l'idée d'employer ReliefF sans trop de soucis. Toutefois, le coût calculatoire faible de CFS, MIFS et notre méthode, associé à l'unique passe sur les données qu'elles nécessitent, ainsi que la variabilité "ponctuelle" des résultats (déficit en terme de correction parfois important pour CFS, différentiel en terme de correction et de nombre de variables sélectionnées parfois significatif entre MIFS et notre méthode), semblent plaider en faveur d'une utilisation simultanée de ces 3 méthodes.

	ID3	C4.5	Sipina	B. Naïfs	1-PPV
Notre Méthode	0.9987	1.0001	0.9842	0.9951	1.0121
MIFS	1.0044	1.0086	0.9961	1.0030	1.0070
CFS	0.9951	0.9935	0.9679	0.9957	0.9955
ReliefF	0.9966	0.9999	0.9936	1.0011	1.0055

TAB. 5.5 –: *Evaluation des Méthodes de SdV pour une 10-Cross-Validation*

	ID3	C4.5	Sipina	B. Naïfs	1-PPV
Notre Méthode	0.9960	1.0046	1.0074	1.0193	1.0042
MIFS	1.0030	1.0086	1.0024	1.0118	1.0078
CFS	0.9863	0.9988	0.9879	1.0351	1.0199
ReliefF	0.9928	1.0014	0.9997	1.0102	1.0107

TAB. 5.6 –: *Evaluation des Méthodes de SdV pour cinq 2-Cross-Validations*

	sans SdV	Notre méthode	MIFS	ReliefF	CFS
GERMAN	20	$6^{30\%}$	$3^{15\%}$	$14^{70\%}$	$5^{25\%}$
MUSH.	22	$8^{36.36\%}$	$1^{4.55\%}$	$17^{77.27\%}$	$3^{13.64\%}$
SICK	28	$6^{21.43\%}$	$9^{32.14\%}$	$12^{42.86\%}$	$1^{3.57\%}$
VEHICLE	18	$12^{66.67\%}$	$6^{33.33\%}$	$18^{100\%}$	$10^{55.56\%}$
ADULT	14	$7^{50\%}$	$5^{35.71\%}$	$6^{42.86\%}$	$5^{35.71\%}$
MONKS 3	6	$3^{50\%}$	$6^{100\%}$	$2^{33.33\%}$	$1^{16.67\%}$
FLAGS	28	$14^{50\%}$	$21^{75\%}$	$27^{96.43\%}$	$3^{10.71\%}$
BREAST	9	$8^{88.89\%}$	$9^{100\%}$	$4^{44.44\%}$	$9^{100\%}$
ZOO	16	$12^{75\%}$	$16^{100\%}$	$14^{87.5\%}$	$9^{56.25\%}$
WINE	13	$11^{84.62\%}$	$13^{100\%}$	$11^{84.62\%}$	$9^{69.23\%}$
CANCER	9	$8^{88.89\%}$	$9^{100\%}$	$9^{100\%}$	$9^{100\%}$
PIMA	8	$2^{25\%}$	$4^{50\%}$	$7^{87.5\%}$	$3^{37.5\%}$
WAVE	21	$15^{71.43\%}$	$21^{100\%}$	$19^{90.48\%}$	$15^{71.43\%}$
CONTRA.	9	$2^{22.22\%}$	$2^{22.22\%}$	$2^{22.22\%}$	$5^{55.56\%}$
ION	34	$25^{73.53\%}$	$13^{38.24\%}$	$33^{97.06\%}$	$9^{26.47\%}$
SPAM	57	$25^{43.86\%}$	$51^{89.47\%}$	$57^{100\%}$	$12^{21.05\%}$
HVOTES	16	$10^{62.5\%}$	$11^{68.75\%}$	$14^{87.5\%}$	$1^{6.25\%}$
moyenne	19.29	$10.24^{56.9\%}$	$11.76^{62.6\%}$	$15.65^{74.4\%}$	$6.41^{41.4\%}$

TAB. 5.7 –: *Evaluation des Méthodes de SdV sur 17 jeux de données de la collection de l'UCI : Nombre de variables sélectionnées*$^{\%\ de\ variables\ sélectionées}$

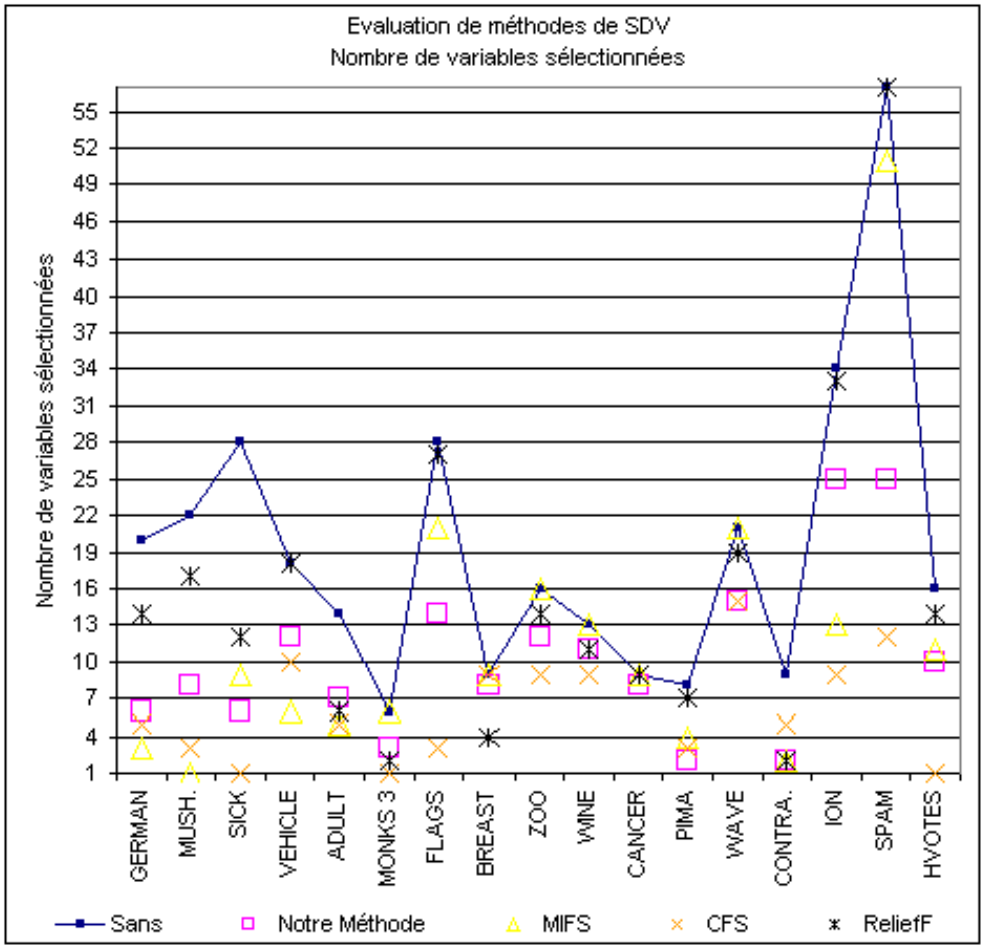

FIG. 5.6 –: *Evaluation Expérimentale de Méthodes de SdV*

5.2.5 Conclusion

En résumé, nous proposons, une méthode :

- basée sur l'hypothèse que l'espace de représentation des données doit être tel que le concept à apprendre doit impliquer qu'une cns représentant ce concept soit valide dans cet espace ;
- ne nécessitant qu'une unique passe sur le jeu de données, et une complexité algorithmique faible ce qui lui confère une rapidité très intéressante ;
- utilisant la méthodologie préalablement introduite pour la comparaison de la validité de cns ;
- utilisant un AG et une nouvelle fonction de fitness particulière afin de résoudre le problème combinatoire de la recherche du sous espace de EV

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$31.2^{3.37}$	$31.2^{3.37}$	$31.7^{3.63}$	$31.7^{4.56}$	$30.8^{3.66}$
MUSH.	$0.07^{0.13}$	$0.1^{0.09}$	$1.48^{0.32}$	$1.03^{0.29}$	0^{0}
SICK	$2.36^{1.11}$	$3.25^{1.06}$	$2.93^{1.1}$	$3.25^{0.85}$	$2.79^{0.94}$
VEHICLE	$33.7^{4.64}$	$32.51^{3.34}$	$31.92^{3.34}$	$32.74^{4.48}$	$33.7^{4.64}$
ADULT	$15.03^{0.89}$	$14.9^{0.61}$	$17.83^{2.84}$	$14.81^{0.08}$	$15.61^{0.63}$
MONKS 3	0^{0}	$2.76^{2.66}$	0^{0}	$19.44^{5.18}$	$2.78^{2.9}$
FLAGS	$49.53^{7.23}$	$48.37^{8.71}$	$45.26^{7.59}$	$42.18^{10.71}$	$51.13^{9.29}$
BREAST	$9.3^{2.23}$	$9.01^{2.22}$	$9.3^{2.23}$	$9.3^{2.23}$	$5.58^{3.35}$
ZOO	$26.64^{10.62}$	$26.76^{10.93}$	$26.64^{10.62}$	$26.73^{7.75}$	$26.82^{14.96}$
WINE	$8.46^{3.83}$	$8.43^{6.42}$	$8.46^{3.83}$	$8.4^{7.16}$	$9.51^{6.07}$
CANCER	$7.47^{2.73}$	$8.48^{2.48}$	$7.47^{2.73}$	$7.47^{2.73}$	$7.47^{2.73}$
PIMA	$25.26^{3.22}$	$25.27^{2.59}$	$25.26^{3.93}$	$25.28^{6.14}$	$25.26^{4.42}$
WAVE	$27.36^{1.37}$	$27.66^{2.19}$	$27.36^{1.37}$	$27.74^{2.17}$	$28.14^{1.55}$
CONTRA.	$52^{3.83}$	$51.94^{4.2}$	$51.8^{2.25}$	$53.7^{3.73}$	$52.01^{2.31}$
ION	$10.83^{6.98}$	$10.5^{5.96}$	$8.26^{5.92}$	$8.56^{3.14}$	$11.13^{6.58}$
SPAM	$11.52^{1.38}$	$13.17^{1.86}$	$12.19^{1.16}$	$12.45^{1.32}$	$11.52^{1.38}$
HVOTES	$4.37^{3.62}$	$4.38^{2.64}$	$4.39^{2.84}$	$4.35^{2.79}$	$5.05^{3.36}$

TAB. 5.8 –: *Evaluation des Méthodes de SdV avec ID3 pour une 10-Cross-Validation*
Légende : a^y a = moy. taux d'erreur pour une 10-cross-validation, y écart type taux d'erreur

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$30.54^{0.47}$	$30.52^{0.44}$	$30.3^{0.25}$	$30.5^{0.28}$	$30.8^{0.53}$
MUSH.	$0.26^{0.05}$	$0.41^{0.11}$	1.48^{0}	$1.04^{0.03}$	$0.23^{0.04}$
SICK	$3.19^{0.17}$	3.25^{0}	3.25^{0}	3.25^{0}	3.25^{0}
VEHICLE	$37.64^{0.62}$	$38.06^{1.54}$	$37.3^{1.99}$	$37.02^{1.63}$	$37.64^{0.62}$
ADULT	$15.34^{0.16}$	$15.23^{0.08}$	$16.92^{0.4}$	$15.06^{0.08}$	$16.36^{0.06}$
MONKS 3	$5^{0.06}$	$4.72^{0.24}$	$5^{0.06}$	19.44^{0}	$5.09^{0.33}$
FLAGS	$52.99^{1.71}$	$53.3^{3.52}$	$52.59^{2.55}$	$53.4^{4.41}$	$55.05^{4.25}$
BREAST	$8.56^{0.48}$	$8.38^{0.5}$	$8.56^{0.48}$	$8.56^{0.48}$	$8.61^{0.55}$
ZOO	$27.88^{0.95}$	$31.07^{5.62}$	$27.88^{0.95}$	$31.71^{7.02}$	$27.11^{0.81}$
WINE	$14.49^{1.15}$	$18.2^{0.98}$	$14.49^{1.15}$	$19.33^{1.04}$	$20.22^{3.52}$
CANCER	$8.37^{0.55}$	$7.82^{0.65}$	$8.37^{0.55}$	$8.37^{0.55}$	$8.37^{0.55}$
PIMA	$26.43^{1.44}$	$26.35^{1.34}$	25.26^{0}	$25.78^{1.04}$	25.26^{0}
WAVE	$29.84^{1.38}$	$30.14^{0.66}$	$29.84^{1.38}$	$30.04^{0.52}$	$29.65^{0.56}$
CONTRA.	$53.22^{0.39}$	$53.51^{0.84}$	$53.4^{0.93}$	$53.33^{1.09}$	$53.81^{0.59}$
ION	$18.52^{1.39}$	$20.06^{3.99}$	$17.67^{1.47}$	$19.43^{2.23}$	$18.98^{1.32}$
SPAM	$17.45^{0.2}$	$13.06^{1.06}$	$13.66^{0.34}$	$13.67^{0.35}$	$17.45^{0.2}$
HVOTES	$4.64^{0.55}$	$4.92^{0.69}$	$4.87^{0.63}$	$5.24^{0.72}$	$5.15^{1.08}$

TAB. 5.9 –: *Evaluation des Méthodes de SdV avec ID3 pour cinq 2-Cross-Validations*
Légende : a^y a = moy. taux d'erreur pour cinq 2-cross-validation, y écart type taux d'erreur

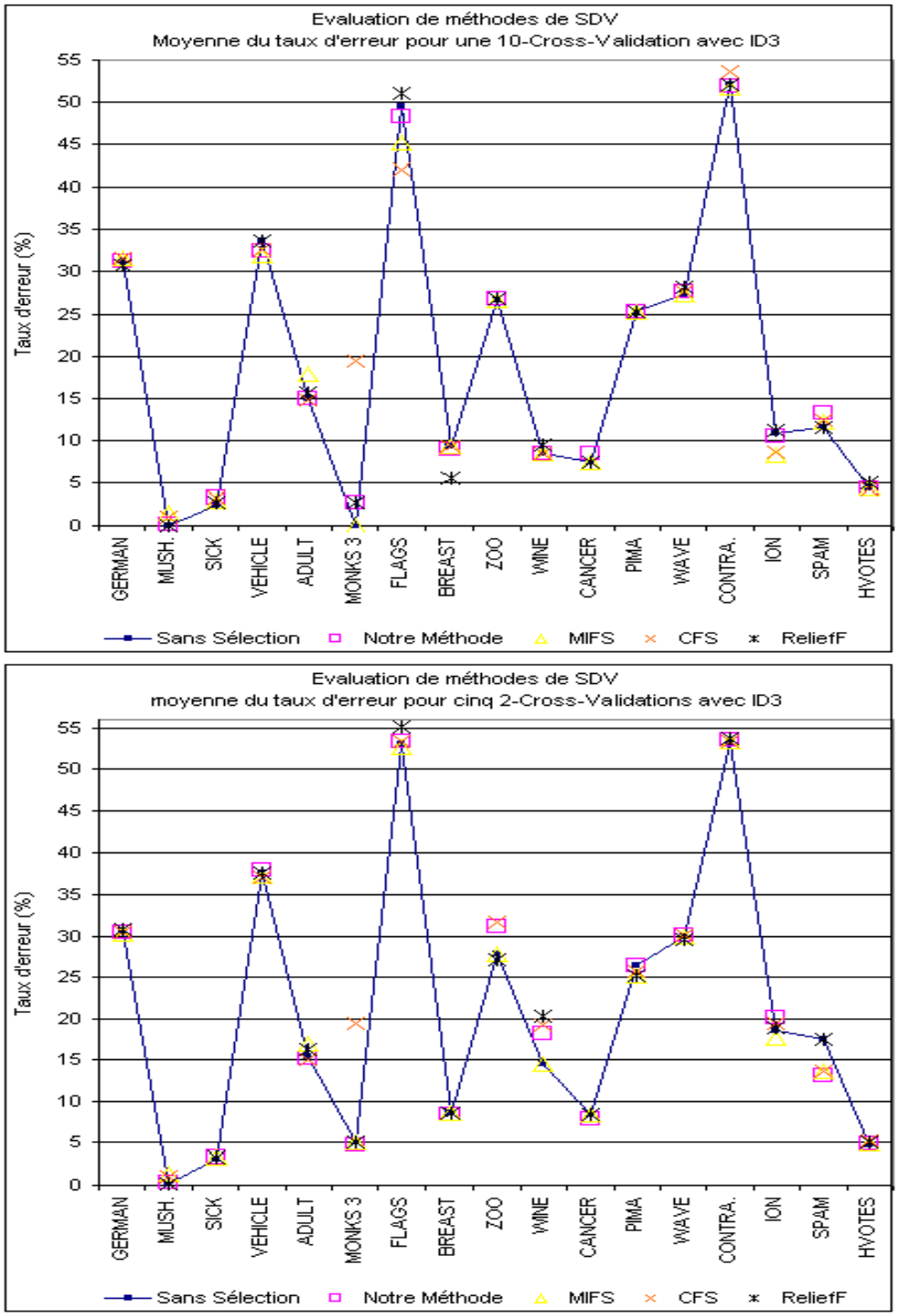

FIG. 5.7 –: *Evaluation Expérimentale de Méthodes de SdV*

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$29^{4.1}$	$27^{5.04}$	$26.1^{3.24}$	$25.7^{4.05}$	$28.3^{4.31}$
MUSH.	0^{0}	0^{0}	$1.48^{0.43}$	$0.98^{0.34}$	0^{0}
SICK	$2.14^{0.81}$	$3.25^{0.91}$	$2.32^{1.01}$	$3.25^{0.82}$	$2.57^{0.69}$
VEHICLE	$32.4^{4.67}$	$32.15^{4.42}$	$32.5^{1.87}$	$34.17^{5.2}$	$32.4^{4.67}$
ADULT	$14.4^{0.59}$	$14.23^{0.64}$	$14.24^{0.49}$	$14.42^{0.54}$	$15.07^{0.37}$
MONKS 3	0^{0}	$2.77^{2.02}$	0^{0}	$19.47^{6.23}$	$2.77^{2.71}$
FLAGS	$34.03^{10.56}$	$27.71^{10.12}$	$28.84^{10.31}$	$28.37^{11.41}$	$31.97^{9.03}$
BREAST	$5.43^{2.28}$	$5.43^{2.1}$	$5.43^{2.28}$	$5.43^{2.28}$	$3.72^{1.31}$
ZOO	$7^{6.4}$	$10.73^{9.97}$	$7^{6.4}$	$8.91^{9.43}$	$5.82^{7.69}$
WINE	$6.14^{6.31}$	$4.54^{5.67}$	$6.14^{6.31}$	$6.14^{3.89}$	$8.4^{7.53}$
CANCER	$5.27^{2.64}$	$4.82^{1.94}$	$5.27^{2.64}$	$5.27^{2.64}$	$5.27^{2.64}$
PIMA	$26.3^{5.07}$	$26.04^{3.26}$	$22.4^{3.98}$	$22.13^{4.57}$	$24.99^{6.03}$
WAVE	$25.5^{1.88}$	$26.12^{1.62}$	$25.5^{1.88}$	$25.98^{1.78}$	$25.92^{1.63}$
CONTRA.	$50.71^{3.6}$	$51.87^{3.83}$	$51.26^{4.06}$	$52.61^{2.6}$	51.53^{3}
ION	$8.55^{3.13}$	$8.54^{3.59}$	$8.29^{5.02}$	$8.27^{4.7}$	$9.42^{4.26}$
SPAM	$7.44^{0.68}$	$11.8^{1.64}$	$7.65^{0.6}$	$8.8^{0.83}$	$7.44^{0.68}$
HVOTES	$6.22^{3.88}$	$5.74^{3.25}$	$6.19^{2.89}$	$4.38^{3.34}$	$5.83^{3.81}$

TAB. 5.10 –: *Evaluation des Méthodes de SdV avec C4.5 pour une 10-Cross-Validation*

Légende : a^y a = moy. taux d'erreur pour la 10-cross-validation, y écart type taux d'erreur

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$30.54^{0.47}$	28.42^{1}	$27.48^{1.3}$	$26.88^{0.89}$	$29.24^{1.35}$
MUSH.	$0.26^{0.05}$	0^{0}	1.48^{0}	$1^{0.04}$	$0.03^{0.04}$
SICK	$3.19^{0.17}$	3.25^{0}	$2.42^{0.11}$	3.25^{0}	$2.63^{0.07}$
VEHICLE	$37.64^{0.62}$	$34.66^{1.96}$	$33.83^{1.92}$	$34.6^{1.55}$	$34.18^{1.97}$
ADULT	$15.34^{0.16}$	$14.34^{0.08}$	$14.25^{0.02}$	$14.47^{0.01}$	$15.27^{0.06}$
MONKS 3	$5^{0.06}$	2.78^{0}	0^{0}	19.44^{0}	$3.29^{1.02}$
FLAGS	$52.99^{1.71}$	$34.85^{1.58}$	$36.91^{3.27}$	$33.4^{3.05}$	$35.46^{3.5}$
BREAST	$8.56^{0.48}$	$5.32^{0.82}$	$5.41^{0.7}$	$5.41^{0.7}$	$5.38^{0.79}$
ZOO	$27.88^{0.95}$	$13.29^{3.23}$	$13.85^{2.44}$	$13.29^{1.48}$	$11.68^{2.69}$
WINE	$14.49^{1.15}$	$8.99^{2.22}$	$8.43^{2.59}$	$7.64^{3.07}$	$7.08^{2.48}$
CANCER	$8.37^{0.55}$	$5.13^{0.71}$	$6.76^{0.3}$	$6.76^{0.3}$	$6.76^{0.3}$
PIMA	$26.43^{1.44}$	$25.16^{0.46}$	$23.2^{1.14}$	$23.83^{1.11}$	24.24^{1}
WAVE	$29.84^{1.38}$	$27.36^{0.27}$	$27.82^{0.48}$	$27.59^{0.61}$	$27.56^{0.66}$
CONTRA.	$53.22^{0.39}$	$52.75^{1.69}$	$51.31^{0.55}$	53.21^{1}	$53.22^{1.12}$
ION	$18.52^{1.39}$	$9.12^{1.23}$	$9.17^{0.83}$	$9.12^{0.74}$	$10.03^{2.44}$
SPAM	$17.45^{0.2}$	$12.3^{0.36}$	$8.77^{0.22}$	$9.71^{0.36}$	$15.38^{0.26}$
HVOTES	$4.64^{0.55}$	$4.87^{0.53}$	$6.67^{0.99}$	$4.6^{0.46}$	$5.29^{0.69}$

TAB. 5.11 –: *Evaluation des Méthodes de SdV avec C4.5 pour cinq 2-Cross-Validations*

Légende : a^y a = moy. taux d'erreur pour cinq 2-cross-validations, y écart type taux d'erreur

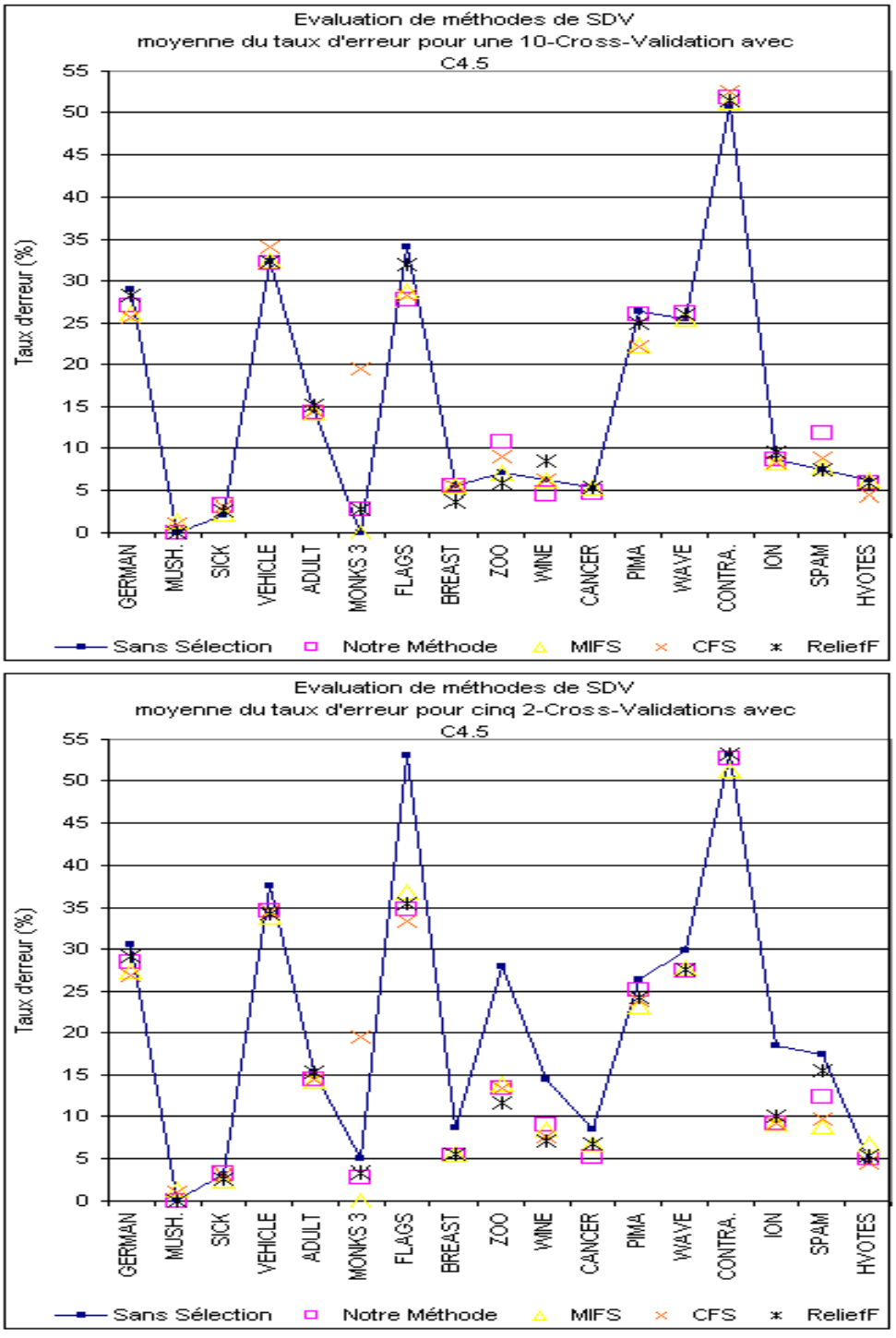

FIG. 5.8 –: *Evaluation Expérimentale de Méthodes de SdV*

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$29.9^{4.5}$	$28.9^{6.01}$	$25.3^{2.83}$	$26.2^{3.49}$	$30.8^{4.21}$
MUSH.	$0.62^{0.17}$	$0.1^{0.12}$	$1.48^{0.28}$	$0.98^{0.35}$	$0.57^{0.21}$
SICK	$2.64^{0.99}$	$3.25^{1.18}$	$2.96^{1.21}$	$3.25^{1.03}$	$2.64^{0.8}$
VEHICLE	$43.51^{1.81}$	$47.86^{6.89}$	$44.8^{4.93}$	$48.57^{6.84}$	$43.51^{1.81}$
ADULT	$17.21^{0.64}$	$17.78^{0.61}$	$17.55^{0.48}$	$17.94^{0.02}$	$20.82^{0.81}$
MONKS 3	0^{0}	$2.77^{1.72}$	0^{0}	$19.46^{6.28}$	$2.78^{2.7}$
FLAGS	$46.37^{11.36}$	$49.58^{9.83}$	$50.11^{7.97}$	$48.92^{10.39}$	$49.97^{12.7}$
BREAST	$6.87^{3.07}$	$5.58^{3.58}$	$6.87^{3.07}$	$6.87^{3.07}$	$5.58^{2.59}$
ZOO	$13.82^{7.88}$	$16.64^{12.06}$	$13.82^{7.88}$	$26.82^{9.16}$	$10.82^{11.23}$
WINE	$7.22^{7.88}$	$7.94^{6.94}$	$7.22^{7.88}$	$8.43^{6.23}$	$6.7^{4.84}$
CANCER	$4.68^{3.44}$	$4.98^{2.54}$	$4.68^{3.44}$	$4.68^{3.44}$	$4.68^{3.44}$
PIMA	$26.05^{5.41}$	$25.26^{5.94}$	$23.04^{4.4}$	$23.44^{5.54}$	$24.73^{3.08}$
WAVE	$23.76^{1.48}$	$24.54^{1.3}$	$23.76^{1.48}$	$27.49^{0.76}$	$24.08^{1.72}$
CONTRA.	$56.29^{5.52}$	$58.32^{3.84}$	$58.86^{2.32}$	$59.67^{3.2}$	$57.29^{2.55}$
ION	$12.26^{5.14}$	$12.24^{3.56}$	$12.26^{5.29}$	$10.55^{5.59}$	$12.52^{4.56}$
SPAM	$10.68^{1.21}$	$13.78^{0.95}$	$11.13^{1.72}$	$11.76^{1.67}$	$10.68^{1.21}$
HVOTES	$4.38^{2.18}$	$4.38^{3.91}$	$4.36^{3.14}$	$4.36^{3.14}$	$4.35^{2.43}$

TAB. 5.12 –: *Evaluation des Méthodes de SdV avec Sipina pour une 10-Cross-Validation*

Légende : a^y a = moy. taux d'erreur pour la 10-cross-validation, y écart type taux d'erreur

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$29.86^{0.71}$	$30.34^{0.73}$	$26.7^{0.7}$	$27.24^{0.73}$	$29.68^{1.12}$
MUSH.	$0.48^{0.16}$	0.1^{0}	1.48^{0}	$1.3^{0.1}$	$0.42^{0.13}$
SICK	$2.41^{0.04}$	3.25^{0}	$3.31^{0.29}$	3.25^{0}	$2.66^{0.1}$
VEHICLE	$46.48^{0.96}$	$49.6^{0.88}$	$50.47^{3.02}$	$47.57^{1.2}$	$46.48^{0.96}$
ADULT	$17.5^{0.15}$	$17.83^{0.05}$	$17.68^{0.06}$	17.92^{0}	$20.84^{0.06}$
MONKS 3	$3.19^{1.13}$	$3.1^{0.65}$	$3.19^{1.13}$	19.44^{0}	2.78^{0}
FLAGS	$54.85^{1.06}$	$49.69^{1.68}$	$54.64^{3.56}$	$50.72^{3.51}$	$54.95^{3.82}$
BREAST	$6.32^{0.6}$	$6.44^{0.45}$	$6.32^{0.6}$	$6.32^{0.6}$	$5.95^{0.47}$
ZOO	$18.23^{0.8}$	$18.41^{1.36}$	$18.23^{0.8}$	$26.92^{0.4}$	$19.4^{1.18}$
WINE	$18.76^{2.09}$	$18.43^{1.79}$	$18.76^{2.09}$	$18.2^{0.98}$	$18.43^{2.9}$
CANCER	$6.53^{0.73}$	$5.89^{0.19}$	$6.53^{0.73}$	$6.53^{0.73}$	$6.53^{0.73}$
PIMA	$26.12^{1.09}$	$25.03^{0.46}$	$23.26^{1.17}$	$24.64^{0.79}$	$25.36^{0.64}$
WAVE	$26.64^{1.08}$	$27.08^{1.26}$	$26.64^{1.08}$	$26.78^{0.79}$	$27.89^{0.69}$
CONTRA.	$58.38^{0.62}$	$57.56^{0.2}$	$58.75^{0.75}$	$59.23^{0.75}$	57.3^{0}
ION	$12.25^{1.62}$	$12.31^{0.38}$	$12.25^{1.1}$	$10.94^{0.74}$	$11.11^{0.31}$
SPAM	$16.93^{0.33}$	$13.87^{0.22}$	$11.46^{0.61}$	$12.87^{0.18}$	$16.93^{0.33}$
HVOTES	$7.22^{3.54}$	$6.11^{2.13}$	$7.95^{1.77}$	$15.22^{8.9}$	$6.66^{3.26}$

TAB. 5.13 –: *Evaluation des Méthodes de SdV avec Sipina pour cinq 2-Cross-Validations*

Légende : a^y a = moy. taux d'erreur pour cinq-2-cross-validations, y écart type taux d'erreur

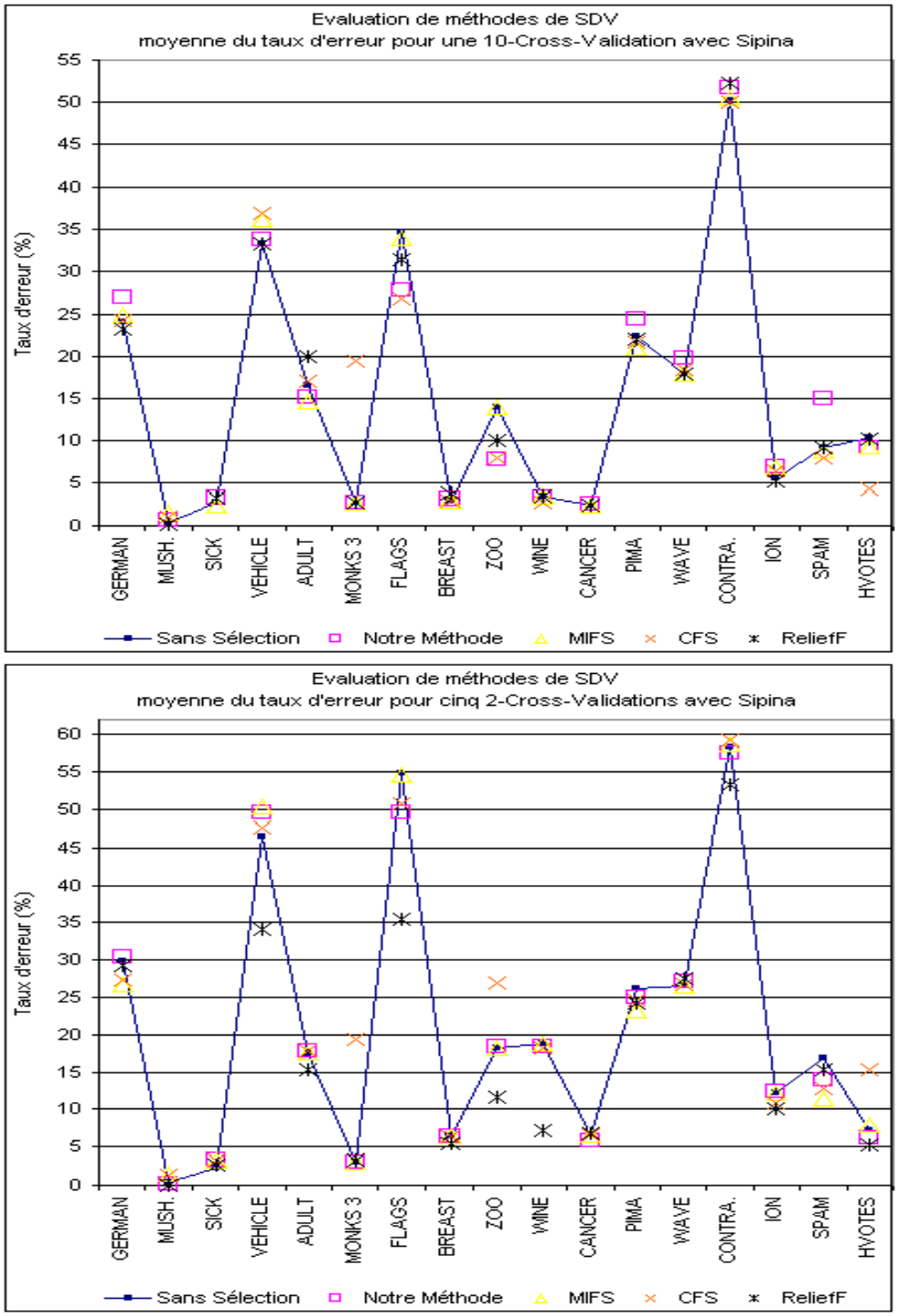

FIG. 5.9 –: *Evaluation Expérimentale de Méthodes de SdV*

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$24^{3.6}$	$26.9^{2.39}$	$25^{4.65}$	$24.2^{2.71}$	$23.2^{4.31}$
MUSH.	$0.31^{0.19}$	$0.65^{0.58}$	$1.48^{0.37}$	$1.08^{0.3}$	$0.23^{0.21}$
SICK	$2.82^{1.45}$	$3.29^{0.98}$	$2.46^{1.18}$	$3.25^{1.25}$	$3.32^{0.89}$
VEHICLE	$33.32^{4.56}$	$33.8^{5.09}$	$36.28^{5.96}$	$37.01^{3.71}$	$33.32^{4.56}$
ADULT	$16.45^{0.72}$	$15.11^{0.9}$	$14.55^{0.57}$	$17.08^{0.54}$	$19.86^{0.78}$
MONKS 3	$2.79^{2.71}$	$2.78^{0.94}$	$2.79^{2.71}$	$19.46^{4.22}$	$2.79^{3.42}$
FLAGS	$34.73^{11.8}$	$27.82^{5.99}$	$34.05^{6.66}$	$26.76^{10.27}$	$31.37^{14.21}$
BREAST	$3^{2.16}$	$3.14^{2.7}$	$3^{2.16}$	$3^{2.16}$	$3.86^{2.48}$
ZOO	$14^{11.14}$	$7.73^{9.2}$	$14^{11.14}$	$7.91^{9.78}$	$9.91^{7.75}$
WINE	$3.4^{4.56}$	$3.33^{3.69}$	$3.4^{4.56}$	$2.78^{3.73}$	$3.4^{2.78}$
CANCER	$2.49^{2.08}$	$2.63^{2.04}$	$2.49^{2.08}$	$2.49^{2.08}$	$2.49^{2.08}$
PIMA	$22.26^{4.13}$	$24.34^{6.39}$	$20.96^{3.11}$	$21.74^{4.05}$	$22.01^{3.23}$
WAVE	$17.8^{1.49}$	$19.78^{1.43}$	$17.8^{1.49}$	$18.44^{2.59}$	$17.82^{2.01}$
CONTRA.	$50.16^{4.49}$	$51.74^{4.35}$	$50.51^{4.26}$	$50.1^{3.57}$	$52.32^{0.72}$
ION	$5.42^{2.99}$	$6.85^{3.19}$	$6.83^{3.65}$	$6.29^{5.54}$	$5.14^{3.08}$
SPAM	$9.06^{0.69}$	$15.02^{1.47}$	$8.76^{1.31}$	$7.87^{1.06}$	$9.06^{0.69}$
HVOTES	$10.34^{4.15}$	$9.18^{4.59}$	$9.2^{2.5}$	$4.37^{2.4}$	$10.2^{3.9}$

TAB. 5.14 –: *Evaluation des Méthodes de SdV avec B.Naïfs pour une 10-Cross-Validation*

Légende : a^y a = moy. taux d'erreur pour la 10-cross-validation, y écart type taux d'erreur

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$24.44^{0.74}$	$26.74^{0.3}$	$26.4^{1.17}$	$24.78^{0.53}$	$23.98^{0.7}$
MUSH.	$0.32^{0.02}$	$0.66^{0.04}$	1.48^{0}	1.08^{0}	$0.31^{0.05}$
SICK	$3.28^{0.24}$	3.32^{0}	$2.59^{0.09}$	3.25^{0}	$3.21^{0.14}$
VEHICLE	$34.54^{1.01}$	$36.03^{1.97}$	$36.78^{0.5}$	$38.53^{2.18}$	$34.54^{1.01}$
ADULT	$16.44^{0.05}$	$15.09^{0.03}$	$14.7^{0.19}$	$17.08^{0.01}$	$19.78^{0.03}$
MONKS 3	2.78^{0}	2.78^{0}	2.78^{0}	19.44^{0}	2.78^{0}
FLAGS	$44.43^{2.77}$	$36.8^{2.58}$	$43.4^{4.34}$	$25.77^{1.63}$	$40^{5.16}$
BREAST	$3^{0.37}$	$2.75^{0.06}$	$3^{0.37}$	$3^{0.37}$	$4.32^{0.11}$
ZOO	$25.7^{4.39}$	$15.67^{0.71}$	$25.7^{4.39}$	$11.9^{3.3}$	$14.29^{3.79}$
WINE	$9.55^{1.88}$	$7.19^{4.53}$	$9.55^{1.88}$	$8.31^{2.11}$	$8.09^{1.76}$
CANCER	$2.9^{0.32}$	$2.64^{0.16}$	$2.9^{0.32}$	$2.9^{0.32}$	$2.9^{0.32}$
PIMA	$22.5^{0.1}$	$24.53^{0.23}$	$21.02^{0.34}$	$21.54^{0.68}$	$21.2^{0.54}$
WAVE	$18.16^{0.19}$	$19.74^{0.1}$	$18.16^{0.19}$	$18.52^{0.1}$	$18.06^{0.17}$
CONTRA.	$49.75^{0.87}$	$52.19^{1.14}$	$51.84^{1.26}$	$50.63^{0.22}$	$53.21^{0.43}$
ION	$9.91^{1.78}$	$8.38^{1.26}$	$7.98^{1.03}$	$7.07^{0.73}$	$8.03^{0.8}$
SPAM	$24.97^{0.32}$	$15.3^{0.09}$	$9.17^{0.09}$	$7.91^{0.09}$	$24.97^{0.32}$
HVOTES	$10.12^{1.21}$	$8.83^{0.42}$	$8.69^{0.49}$	4.37^{0}	$10.02^{1.05}$

TAB. 5.15 –: *Evaluation des Méthodes de SdV avec B.Naïfs pour cinq 2-Cross-Validations*

Légende : a^y a = moy. taux d'erreur pour cinq 2-cross-validations, y écart type taux d'erreur

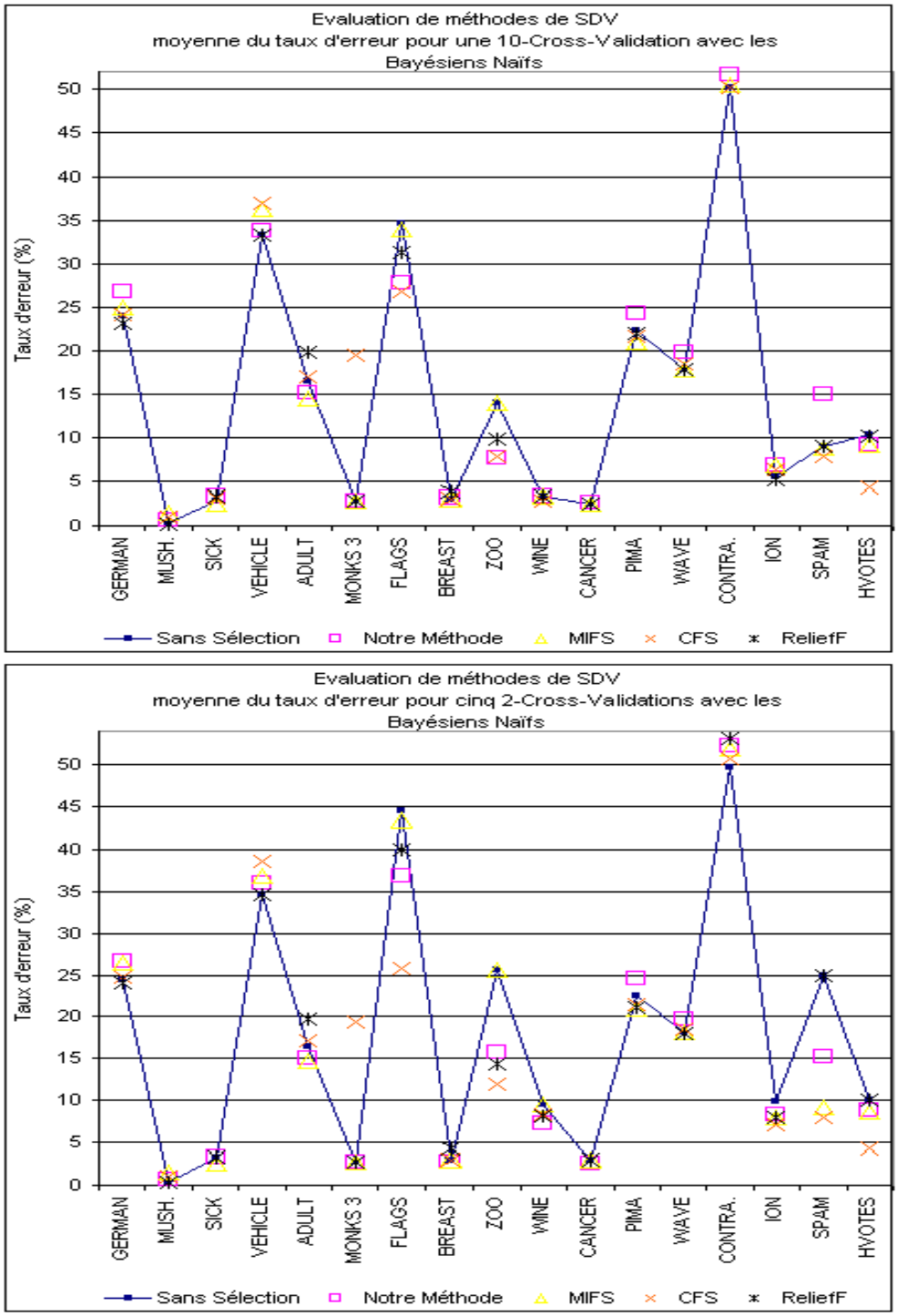

FIG. 5.10 –: *Evaluation Expérimentale de Méthodes de SdV*

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$31.6^{6.07}$	$35.1^{3.78}$	$34.9^{4.25}$	$33^{4.92}$	$31.8^{2.89}$
MUSH.	0^{0}	0^{0}	$1.48^{0.27}$	$0.98^{0.21}$	0^{0}
SICK	$2.79^{0.91}$	$3.68^{1.73}$	$6.11^{1.76}$	$3.25^{1.11}$	$2.39^{0.8}$
VEHICLE	$33.21^{2.73}$	$35.46^{2.88}$	$35.22^{5.72}$	$36.05^{3.21}$	$33.21^{2.73}$
ADULT	$20.19^{0.74}$	$20.02^{0.63}$	$20.05^{0.39}$	$22.16^{0.94}$	$21.66^{0.55}$
MONKS 3	$13.66^{4.68}$	$3.01^{1.48}$	$13.66^{4.68}$	$37.72^{4.39}$	$5.78^{3.31}$
FLAGS	$46.95^{10.21}$	$40.29^{10.92}$	$40.68^{13.85}$	$39.74^{8.25}$	$46.95^{7.73}$
BREAST	$5.58^{3.1}$	$5.29^{2.79}$	$5.58^{3.1}$	$5.58^{3.1}$	$5.72^{2.22}$
ZOO	$3.91^{6.56}$	$3^{4.58}$	$3.91^{6.56}$	$4^{4.9}$	$2.91^{4.45}$
WINE	$4.48^{4.17}$	$4.51^{3.36}$	$4.48^{4.17}$	$2.22^{3.69}$	$4.44^{6.94}$
CANCER	$5.42^{2.54}$	$7.03^{2.68}$	$5.42^{2.54}$	$5.42^{2.54}$	$5.42^{2.54}$
PIMA	$31.65^{4.64}$	$30.98^{4.13}$	$29.7^{5.54}$	$34.38^{5.6}$	33.32^{7}
WAVE	$40.48^{2.22}$	$38.4^{1.32}$	$40.48^{2.22}$	$38.42^{1.32}$	$40.04^{2.43}$
CONTRA.	$56.82^{4.49}$	$61.71^{2.7}$	$56.22^{3.12}$	$55.34^{4.84}$	$55.4^{3.73}$
ION	$13.97^{5.35}$	$10.83^{3.81}$	$14.56^{6.47}$	$12.52^{5.38}$	$12.79^{7.25}$
SPAM	$9^{0.91}$	$11^{1.37}$	$9.89^{0.87}$	$9.82^{1.09}$	$9^{0.91}$
HVOTES	$13.76^{4.37}$	$4.38^{2.21}$	$5.06^{5.21}$	$4.38^{2.84}$	$15.41^{4.5}$

TAB. 5.16 –: *Evaluation des Méthodes de SdV avec 1-PPV pour une 10-Cross-Validation*

Légende : a^y a = moy. taux d'erreur pour la 10-cross-validation, y écart type taux d'erreur

	sans SdV	Notre méthode	MIFS	CFS	ReliefF
GERMAN	$32.92^{1.49}$	$35^{2.83}$	$34.06^{1.15}$	$31.42^{1.33}$	$33.64^{0.39}$
MUSH.	0^{0}	0^{0}	$5.57^{8.19}$	$7.41^{7.87}$	$0.02^{0.04}$
SICK	$3.09^{0.43}$	$4.02^{0.67}$	$3.78^{0.78}$	$4.78^{0.92}$	$3.37^{1.96}$
VEHICLE	$36.41^{0.91}$	$37.92^{1.99}$	$36.41^{2.01}$	$37.71^{0.78}$	$36.41^{0.91}$
ADULT	$20.92^{0.13}$	$20.98^{1.18}$	$19.8^{1.28}$	$20.89^{1.01}$	$20.99^{0.27}$
MONKS 3	$12.13^{1.45}$	4.21^{1}	$12.13^{1.45}$	$27.55^{5.29}$	$4.54^{1.06}$
FLAGS	$48.14^{3.29}$	$46.08^{3.57}$	$48.87^{2.89}$	$32.68^{3.82}$	$50.31^{3.95}$
BREAST	$5.58^{0.6}$	$6.01^{0.68}$	$5.58^{0.6}$	$5.58^{0.6}$	$6.21^{1.07}$
ZOO	$7.13^{1.6}$	$7.15^{4.97}$	$7.13^{1.6}$	$4.56^{2.03}$	$6.53^{2.55}$
WINE	$8.88^{1.96}$	$4.94^{1.3}$	$8.88^{1.96}$	$3.15^{0.57}$	$4.38^{0.42}$
CANCER	$5.04^{0.48}$	$6.85^{0.98}$	$5.04^{0.48}$	$5.04^{0.48}$	$5.04^{0.48}$
PIMA	$31.87^{1.23}$	$36.38^{3.69}$	$26.93^{1.51}$	$28.65^{1.45}$	$30.89^{2.48}$
WAVE	$40.94^{0.51}$	$39.55^{0.43}$	$40.94^{0.51}$	$39.5^{0.62}$	$40.72^{0.72}$
CONTRA.	$59.02^{0.34}$	$60.45^{1.73}$	$57.8^{1.04}$	$56.13^{0.45}$	$55.84^{0.54}$
ION	$13.62^{0.75}$	$13.56^{0.84}$	$12.88^{0.49}$	$12.14^{1.16}$	$13.11^{0.75}$
SPAM	$10.55^{0.25}$	$14.44^{2.92}$	$10.82^{0.36}$	$10.95^{0.35}$	$10.55^{0.25}$
HVOTES	$13.89^{1.12}$	$4.6^{0.46}$	$4.83^{0.52}$	$8.82^{6.96}$	$13.84^{1.88}$

TAB. 5.17 –: *Evaluation des Méthodes de SdV avec 1-PPV pour cinq 2-Cross-Validations*

Légende : a^y a = moy. taux d'erreur pour cinq 2-cross-validations, y écart type taux d'erreur

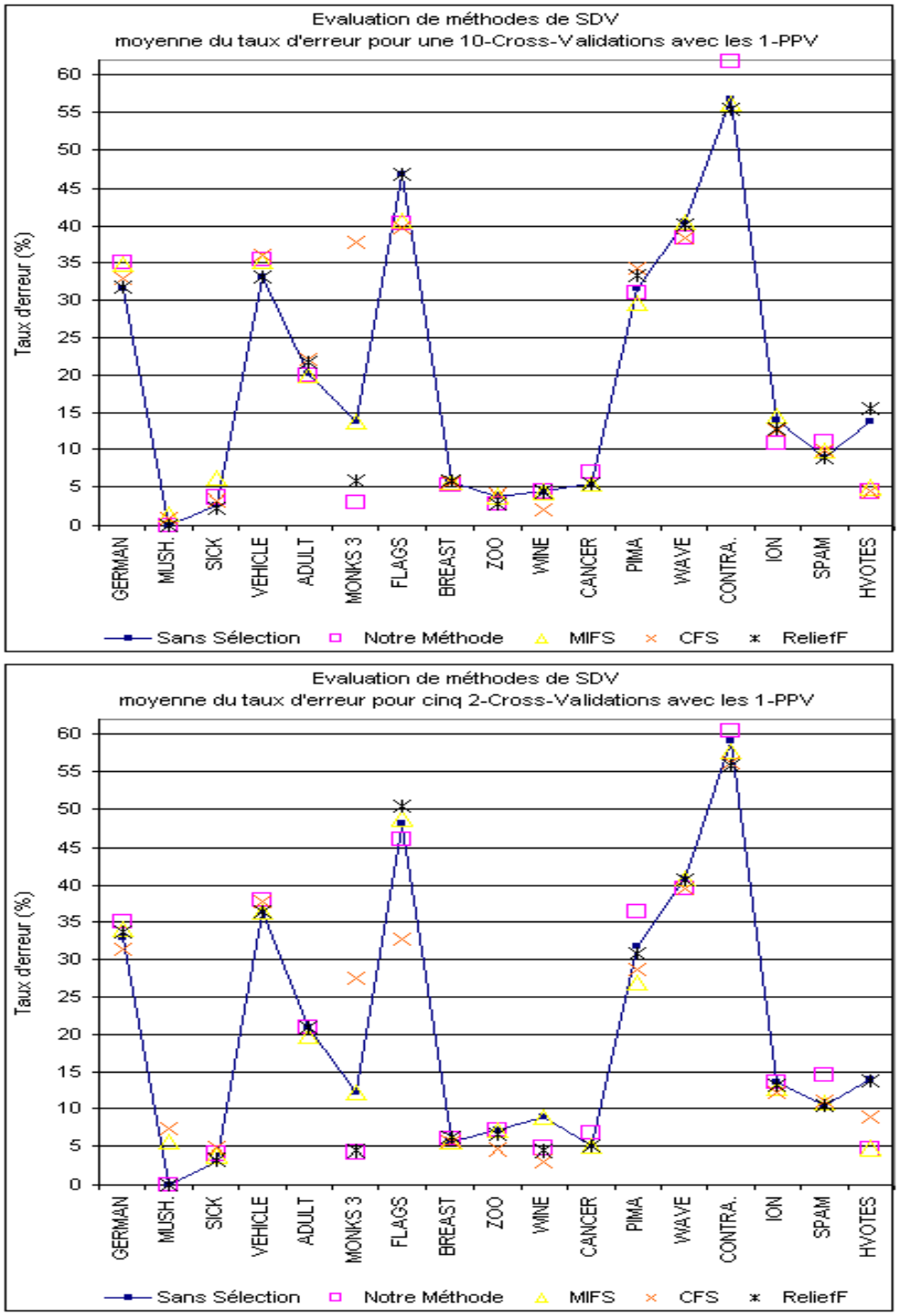

FIG. 5.11 –: *Evaluation Expérimentale de Méthodes de SdV*

impliquant la validité la plus forte pour P.

Les évaluations expérimentales ont montré que :

- concernant la précision prédictive, notre méthode se comporte, en général, comme les 3 autres méthodes testées (qui constituent des méthodes de référence du domaine),
- concernant le nombre de variables sélectionnées, la réduction du nombre de variables due à notre méthode est réelle même si elle est inférieure à celle impliquée par CFS,
- concernant le temps de calcul, notre méthode est un peu plus lente que MIFS qui est une méthode de sélection extrêmement rapide,
- qu'une utilisation simultanée des méthodes CFS, MIFS et la notre semble réalisable et judicieuse.

Nous pouvons également conclure que :

- le paradigme de sélection sous-jacent à notre méthode est relativement différent de ceux de MIFS et CFS et peut être mieux adapté à certains jeux de données ;
- notre méthode peut être améliorée du point de vue du coût calculatoire (voir ci dessous) ;
- on peut aisément modifier la structure de l'AG de manière à pouvoir rechercher non pas l'ensemble "optimal" de variables mais le meilleur ensemble de variables tel qu'il comprenne au plus un nombre fixé de variables, afin de réduire le nombre de variables sélectionnées.

Enfin, bien que l'hypothèse 5 (une classe par modalité du concept à apprendre, cf. page 119) soit forte, notre méthode fournit des résultats de qualité proche ou supérieure à ceux des méthodes existantes. Les travaux futurs seront dirigés vers :

- une amélioration de la méthode, par relaxation de l'hypothèse 5, la relaxation de cette hypothèse pouvant éventuellement être réalisée par modification de la fonction de fitness utilisée et en donnant notamment plus d'importance à l'aspect séparation des classes ($xv_2(P)^{EV_\star}$) par rapport à l'aspect homogénéité interne des classes ($xv_1(P)^{EV_\star}$);
- une réduction du temps de calcul associé par substitution d'une méthode d'optimisation gloutonne à l'AG.

REMARQUE : La méthode que nous venons de proposer peut également être employée si des variables quantitatives sont présentes, cependant, si elle ne nécessite également qu'une unique passe sur les données, sa complexité calculatoire est alors en $O(n^2)$.

5.3 Contribution à la Sélection de Variables pour l'Apprentissage Non Supervisé : Une Nouvelle Méthode Efficiente et Rapide

Si dans le cadre de l'apprentissage supervisé, la problématique de la SdV a été l'objet de nombreux travaux, le constat est radicalement opposé pour l'apprentissage non supervisé. En effet, si l'on exclut les méthodes d'analyse factorielle et celles issues du "multidimensional scaling", seules quelques rares approches ont été proposées [DL03], [DCSL02], [LJF02] (ces approches étant de type enveloppe à l'exception de [DCSL02]). Cela s'explique sans doute notamment par la relative difficulté à déterminer clairement ce que doit permettre la SdV dans ce cadre puisqu'en apprentissage non supervisé il n'existe pas, à proprement parlé et contrairement à l'apprentissage supervisé, de structure objectif connue et devant être extraite.

Selon nous, l'objectif de la SdV dans le cadre de l'apprentissage non supervisé est de réduire l'ERD de manière à ce que la validité de la meilleure cns obtenue par l'intermédiaire d'un algorithme donné appliqué sur un jeu de données complet (ERD complet), et, la validité de la meilleure cns obtenue par l'intermédiaire du même algorithme sur ce jeu de données réduit (ERD réduit) soient proches, ou, que la cns obtenue dans le second cas présente une meilleure validité. Ainsi, d'un point de vue pragmatique, et puisque les algorithmes de cns possèdent un coût calculatoire sensible à la dimension de l'ERD, l'objectif de la SdV sera principalement d'accélérer le temps de traitement nécessaire à la cns tout en assurant le maintien ou l'accroissement de la qualité (validité) de la structure extraite.

Nous proposons ici une méthode de SdV pour l'apprentissage non supervisé, cette méthode dérive d'une certaine manière de la méthode proposée pour l'apprentissage supervisé. En effet, dans le chapitre précédent nous recherchions le sous espace de l'ERD tel que la partition impliquée par la variable endogène apparaisse la plus valide possible, cette fois, tout se passe comme si il n'y avait pas une unique variable endogène mais p variables endogènes (les p variables de l'ERD) pour lesquelles nous devons rechercher le sous espace de l'ERD tel que les partitions impliquées par ces p variables apparaissent dans leur ensemble comme les plus valides. On considère donc ici l'apprentissage non supervisé comme un apprentissage supervisé de p concepts de manière simultanée.

Nous proposons donc une extension de la notion de la validité d'une partition : l'adéquation entre deux ensembles de variables, qui, dans le cadre de données catégorielles, correspond à l'adéquation entre deux ensembles de partitions (car les variables catégorielles définissent des partitions sur un ensemble d'objets). Cette notion permet de définir deux indices pour l'évaluation de l'adéquation entre un ensemble de variables et un sous ensemble de ce dernier. La caractérisation statistique de ces indices, similairement à celle des indices xv_1 et xv_2, mène à une méthodologie d'évaluation/comparaison de l'adéqua-

tion entre sous ensembles de variables et un ensemble particulier de variables. Cela nous permet, de manière identique à la méthodologie d'évaluation de la validité de cns, de dériver une méthode de SdV pour l'apprentissage non supervisé.

5.3.1 Evaluation de l'Adéquation entre deux Ensembles de Variables

Afin d'évaluer l'adéquation entre deux ensembles de variables catégorielles $EV_1 = \{V_{1_i}, i = 1..l\}$ et $EV_2 = \{V_{2_j}, j = 1..m\}$ nous utilisons un ensemble de 4 indices :

- $\mathbf{LL}(\mathbf{EV_1}, \mathbf{EV_2})$: qui comptabilise le nombre de couples formés d'un couple d'objets (o_a, o_b) et d'un couple de variables (V_{1_i}, V_{2_j}) formé d'une variable V_{1_i} de EV_1 et d'une variable EV_{2_j} de V_2 tels que :
  le couple d'objets (o_a, o_b) est caractérisé simultanément par un **lien** selon V_{1_i} et par un **lien** selon V_{2_j}.

$$LL(EV_1, EV_2) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{j=1..m} lien_i(o_{a_i}, o_{b_i}) \times lien_j(o_{a_j}, o_{b_j}) \tag{5.1}$$

- $\overline{\mathbf{LL}}(\mathbf{EV_1}, \mathbf{EV_2})$: qui comptabilise le nombre de couples formés d'un couple d'objets (o_a, o_b) et d'un couple de variables (V_{1_i}, V_{2_j}) formé d'une variable V_{1_i} de EV_1 et d'une variable V_{2_j} de EV_2 tels que :
  le couple d'objets (o_a, o_b) est caractérisé simultanément par un **non-lien** selon V_{1_i} et par un **non-lien** selon V_{2_j}.

$$\overline{LL}(EV_1, EV_2) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{j=1..m} (1 - lien_i(o_{a_i}, o_{b_i})) \times (1 - lien_j(o_{a_j}, o_{b_j})) \tag{5.2}$$

- $\mathbf{L\overline{L}}(\mathbf{EV_1}, \mathbf{EV_2})$: qui comptabilise le nombre de couples formés d'un couple d'objets (o_a, o_b) et d'un couple de variables (V_{1_i}, V_{2_j}) formé d'une variable V_{1_i} de EV_1 et d'une variable V_{2_j} de EV_2 tels que :
  le couple d'objets (o_a, o_b) est caractérisé simultanément par un **lien** selon V_{1_i} et par un **non-lien** selon V_{2_j}.

$$L\overline{L}(EV_1, EV_2) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{j=1..m} lien_i(o_{a_i}, o_{b_i}) \times (1 - lien_j(o_{a_j}, o_{b_j})) \tag{5.3}$$

- $\overline{\mathbf{L}}\mathbf{L}(\mathbf{EV_1}, \mathbf{EV_2})$: qui comptabilise le nombre de couples formés d'un couple d'objets (o_a, o_b) et d'un couple de variables (V_{1_i}, V_{2_j}) formé d'une variable V_{1_i} de EV_1 et d'une variable V_{2_j} de EV_2 tels que :

le couple d'objets (o_a,o_b) est caractérisé simultanément par un **non-lien** selon V_{1_i} et par un **lien** selon V_{2_j}.

$$\overline{L}L(EV_1,EV_2) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{j=1..m} (1-lien_i(o_{a_i},o_{b_i}))\times lien_j(o_{a_j},o_{b_j}) \tag{5.4}$$

REMARQUES :

- Par la suite nous nous contentons d'écrire (sauf indication contraire) LL, $\overline{L}\overline{L}$, $\overline{L}L$, $L\overline{L}$ en lieu et place respective de $LL(EV_1,EV_2)$, $\overline{L}\overline{L}(EV_1,EV_2)$, $\overline{L}L(EV_1,EV_2)$, $L\overline{L}(EV_1,EV_2)$.
- $LL + L\overline{L}$ correspond à m fois le nombre de liens au sein des variables de EV_1
- $\overline{L}L+\overline{L}\overline{L}$ correspond à m fois le nombre de non-liens au sein des variables de EV_1
- $LL + \overline{L}L$ correspond à l fois le nombre de liens au sein des variables de EV_2
- $L\overline{L} + \overline{L}\overline{L}$ correspond à l fois le nombre de liens au sein des variables de EV_2
- Une adéquation forte entre EV_1 et EV_2 se caractérise par de fortes valeurs pour LL et $\overline{L}\overline{L}$.
- L'ensemble de ces relations peuvent être résumées au sein d'une table de contingence :

	Liens dans EV_2	Non-Liens dans EV_2	$Total$
Liens dans EV_1	LL	$L\overline{L}$	$LL + L\overline{L}$
Non-Liens dans EV_1	$\overline{L}L$	$\overline{L}\overline{L}$	$\overline{L}L + \overline{L}\overline{L}$
$Total$	$LL + \overline{L}L$	$L\overline{L} + \overline{L}\overline{L}$	$\frac{n\times(n-1)}{2} \times l \times m$

5.3.2 Remarques Importantes Concernant l'Aspect Calculatoire

Bâtir cette table de contingence ne nécessite qu'une seule passe sur le jeu de données. Dans le cas de données catégorielles uniquement, cela ne requiert que $O(nlm)$ comparaisons[7], ce nombre de comparaisons peut atteindre $O(n^2lm)$ dans le cas de présence de variables quantitatives et d'utilisation de fonctions $lien_i$ telles que définies dans le cas 2 de l'exemple illustratif du chapitre précédent (page 69).

Du point de vue de l'utilisation mémoire, quel que soit la nature des données, le stockage de lm tables de contingence est nécessaire, ce qui correspond à un encombrement mémoire relativement faible et surtout totalement indépendant du nombre d'objets du jeu de données considéré.

7. Intuitivement, les définitions formelles de LL, $\overline{L}L$, $L\overline{L}$ et $\overline{L}\overline{L}$ semblent impliquer $O(n^2lm)$ comparaisons mais des astuces de calcul permettent de réduire ce nombre de comparaisons, ces astuces sont similaires à celles présentées précédemment) afin de bâtir $l \times m$ tables de contingence (croisant les l variables de EV_1 avec les m variables de EV_2)

5.3.3 Evaluation de l'adéquation entre $\mathbf{EV}$ un Ensemble de Variables et $\mathbf{EV}_\star$ un Sous Ensemble de $\mathbf{EV}$ ($\mathbf{EV}_\star \subseteq \mathbf{EV}$)

Afin d'évaluer l'adéquation entre EV un ensemble de variables et $EV_\star$ un sous ensemble de EV ($EV_\star \subseteq EV$) nous utilisons une adaptation des 4 indices que nous venons de présenter :

- $\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}(\mathbf{EV}_\star,\mathbf{EV})$: qui comptabilise le nombre de couples formés d'un couple d'objets (o_a,o_b) et d'un couple de variables (V_{1_i},V_{2_j}) formé d'une variable V_{1_i} de $EV_\star$ et d'une variable V_{2_j} de EV telles que :
  $V_{1_i} \neq V_{2_j}$, le couple d'objets (o_a,o_b) est caractérisé simultanément par un **lien** selon V_{1_i} et par un **lien** selon V_{2_j}.

$$\widetilde{L}\widetilde{L}(EV_\star,EV) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{\substack{j\ =\ 1..m \\ j \text{ tel que } V_{1_i} \neq V_{2_j}}} lien_i(o_{a_i},o_{b_i}) \times lien_j(o_{a_j},o_{b_j}) \tag{5.5}$$

- $\overline{\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}}(\mathbf{EV}_\star,\mathbf{EV})$: qui comptabilise le nombre de couples formés d'un couple d'objets (o_a,o_b) et d'un couple de variables (V_{1_i},V_{2_j}) formé d'une variable V_{1_i} de $EV_\star$ et d'une variable V_{2_j} de EV telles que :
  $V_{1_i} \neq V_{2_j}$, le couple d'objets (o_a,o_b) est caractérisé simultanément par un **non-lien** selon V_{1_i} et par un **non-lien** selon V_{2_j}.

$$\overline{\widetilde{L}\widetilde{L}}(EV_\star,EV) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{\substack{j\ =\ 1..m \\ j \text{ tel que } V_{1_i} \neq V_{2_j}}} (1-lien_i(o_{a_i},o_{b_i})) \times (1-lien_j(o_{a_j},o_{b_j})) \tag{5.6}$$

- $\widetilde{\mathbf{L}}\overline{\widetilde{\mathbf{L}}}(\mathbf{EV}_\star,\mathbf{EV})$: qui comptabilise le nombre de couples formés d'un couple d'objets (o_a,o_b) et d'un couple de variables (V_{1_i},V_{2_j}) formé d'une variable V_{1_i} de $EV_\star$ et d'une variable V_{2_j} de EV telles que :
  $V_{1_i} \neq V_{2_j}$, le couple d'objets (o_a,o_b) est caractérisé simultanément par un **lien** selon V_{1_i} et par un **non-lien** selon V_{2_j}.

$$\widetilde{L}\overline{\widetilde{L}}(EV_\star,EV) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{\substack{j\ =\ 1..m \\ j \text{ tel que } V_{1_i} \neq V_{2_j}}} lien_i(o_{a_i},o_{b_i}) \times (1-lien_j(o_{a_j},o_{b_j})) \tag{5.7}$$

– $\overline{\widetilde{\mathbf{LL}}}(\mathbf{EV}_\star,\mathbf{EV})$ qui comptabilise le nombre de couples formés d'un couple d'objets (o_a,o_b) et d'un couple de variables (V_{1_i},V_{2_j}) formé d'une variable V_{1_i} de $EV_\star$ et d'une variable V_{2_j} de EV telles que :
$V_{1_i} \neq V_{2_j}$, le couple d'objets (o_a,o_b) est caractérisé simultanément par un **non-lien** selon V_{1_i} et par un **lien** selon V_{2_j}.

$$\overline{\widetilde{L}}\widetilde{L}(EV_\star,EV) = \sum_{a=1..n}\sum_{b=a+1..n}\sum_{i=1..l}\sum_{\substack{j\ =\ 1..m \\ j \text{ tel que } V_{1_i} \neq V_{2_j}}} (1-lien_i(o_{a_i},o_{b_i}))\times lien_j(o_{a_j},o_{b_j}) \tag{5.8}$$

REMARQUES :

– Par la suite nous nous contentons d'écrire (sauf indication contraire) $\widetilde{L}\widetilde{L}$, $\overline{\widetilde{L}}\overline{\widetilde{L}}$, $\overline{\widetilde{L}}\widetilde{L}$, $\widetilde{L}\overline{\widetilde{L}}$ en lieu et place respective de $\widetilde{L}\widetilde{L}(EV_1,EV_2)$, $\overline{\widetilde{L}}\overline{\widetilde{L}}(EV_1,EV_2)$, $\overline{\widetilde{L}}\widetilde{L}(EV_1,EV_2)$, $\widetilde{L}\overline{\widetilde{L}}(EV_1,EV_2)$.
– Une adéquation forte entre $EV_\star$ et EV se caractérise par de fortes valeurs pour LL et $\overline{\widetilde{L}}\overline{\widetilde{L}}$.
– L'ensemble de ces relations peuvent être résumées au sein d'une table de contingence :

	$\widetilde{L}$	$\overline{\widetilde{L}}$	$Total$
$\widetilde{L}$	$\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}$	$\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}$	$\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}$
$\overline{\widetilde{L}}$	$\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}$	$\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}$	$\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}$
$Total$	$\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}$	$\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}$	

Ainsi, le niveau d'adéquation entre $EV_\star$ et EV peut être caractérisé par les indices $\widetilde{L}\widetilde{L}$, $\overline{\widetilde{L}}\widetilde{L}$, $\widetilde{L}\overline{\widetilde{L}}$, $\overline{\widetilde{L}}\overline{\widetilde{L}}$. Une forte adéquation étant associée à de fortes valeurs pour $\widetilde{L}\widetilde{L}$, $\overline{\widetilde{L}}\overline{\widetilde{L}}$. Cependant, la signification de fortes valeurs n'étant pas totalement intuitive nous proposons, de manière similaire aux indices concernant l'évaluation de la validité de cns (partitions), de déterminer les lois statistiques suivies par les indices $\widetilde{L}\widetilde{L}$ et $\overline{\widetilde{L}}\overline{\widetilde{L}}$ en cas de non adéquation. Cela permet alors de dériver deux indices $Aq_1(EV_\star,EV)$ et $Aq_2(EV_\star,EV)$ caractérisant respectivement la significativité de $\widetilde{L}\widetilde{L}$ et $\overline{\widetilde{L}}\overline{\widetilde{L}}$ et suivant, dans les conditions de non adéquation, la loi normale centrée réduite :

$$Aq_1(EV_\star,EV) = \frac{\widetilde{\widetilde{L}}\widetilde{\widetilde{L}} - \frac{(\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}})(\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}})}{\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}}}{\sqrt{\frac{(\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}})(\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}})}{\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}} \times \left(1-\frac{\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}}{\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}}\right)}}, \; Aq_1(EV_\star,EV) \hookrightarrow N(0,1)$$

$$Aq_2(EV_\star,EV) = \frac{\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}} - \frac{(\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}})(\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}})}{\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}}}{\sqrt{\frac{(\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}})(\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}})}{\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}} \times \left(1-\frac{\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}}{\widetilde{\widetilde{L}}\widetilde{\widetilde{L}}+\overline{\widetilde{\widetilde{L}}}\widetilde{\widetilde{L}}+\widetilde{\widetilde{L}}\overline{\widetilde{\widetilde{L}}}+\overline{\widetilde{\widetilde{L}}}\overline{\widetilde{\widetilde{L}}}}\right)}}, \; Aq_2(EV_\star,EV) \hookrightarrow N(0,1)$$

5.3.4 Evaluation/Comparaison de l'Adéquation entre un Ensemble de Variables (EV) et des Sous Ensembles de EV

Afin d'évaluer/comparer l'adéquation entre un ensemble de variables (EV) et des sous ensembles de EV, nous utilisons la méthodologie présentée dans le cadre de l'évaluation/comparaison de la validité de cns pour laquelle on substitue toutefois les indices Aq_1 et Aq_2 aux indices xv_1 et xv_2.

Ainsi, la comparaison de l'adéquation de deux sous ensembles ($EV_1 \subseteq EV$ et $EV_2 \subseteq EV$) peut être réalisée par comparaison des couples de valeurs $(Aq_1(EV_\star,EV_1),Aq_2(EV_\star,EV_1))$ et $(Aq_1(EV_\star,EV_2),Aq_2(EV_\star,EV_2))$. Cette comparaison mène à 4 situations différentes :

- EV_1 est considéré comme plus en adéquation avec EV que EV_2 ssi $(Aq_1(EV_\star,EV_2) < Aq_1(EV_\star,EV_1)$ et $Aq_2(EV_\star,EV_2) < Aq_2(EV_\star,EV_1))$ ou $(Aq_1(EV_\star,EV_2) \leq Aq_1(EV_\star,EV_1)$ et $Aq_2(EV_\star,EV_2) < Aq_2(EV_\star,EV_1))$ ou $(Aq_1(EV_\star,EV_2) < Aq_1(EV_\star,EV_1)$ et $Aq_2(EV_\star,EV_2) \leq Aq_2(EV_\star,EV_1))$
  nous notons cette relation : $EV_1 < b > EV_2$
- EV_2 est considéré comme plus en adéquation avec EV que EV_1 ssi $(Aq_1(EV_\star,EV_1) < Aq_1(EV_\star,EV_2)$ et $Aq_2(EV_\star,EV_1) < Aq_2(EV_\star,EV_2))$ ou $(Aq_1(EV_\star,EV_1) \leq Aq_1(EV_\star,EV_2)$ et $Aq_2(EV_\star,EV_1) < Aq_2(EV_\star,EV_2))$ ou $(Aq_1(EV_\star,EV_1) < Aq_1(EV_\star,EV_2)$ et $Aq_2(EV_\star,EV_1) \leq Aq_2(EV_\star,EV_2))$
  nous notons cette relation : $EV_2 < b > EV_1$
- EV_1 et EV_2 sont considérés comme équivalents du point de vue de l'adéquation avec EV ssi $(Aq_1(EV_\star,EV_1) = Aq_1(EV_\star,EV_2)$ et $Aq_2(EV_\star,EV_1) = Aq_2(EV_\star,EV_2))$
  nous notons cette relation : $EV_2 < s > EV_1$
- EV_1 et EV_2 sont considérés comme incomparables du point de vue de l'adéquation avec EV ssi $(Aq_1(EV_\star,EV_1) < Aq_1(EV_\star,EV_2)$ et $Aq_2(EV_\star,EV_1) > Aq_2(EV_\star,EV_2))$ ou $(Aq_1(EV_\star,EV_1) > Aq_1(EV_\star,EV_2)$ et $Aq_2(EV_\star,EV_1) < Aq_2(EV_\star,EV_2))$
  nous notons cette relation : $EV_2 <? > EV_1$

5.3.5 La Nouvelle Méthode de Sélection de Variables

La méthode est ici en tout point identique à celle proposée pour l'apprentissage supervisé à l'unique exception du calcul des indices de validité des sous espaces de l'ERD que nous venons de présenter.

Concernant la définition de la fonction de fitness de l'AG, nous proposons une fonction basée sur l'observation que $Aq_1(EV_\star,EV)$ et $Aq_2(EV_\star,EV)$ doivent être les plus élevées possibles pour que $EV_\star$ et EV soient considérés comme étant en adéquation. Cette fonctions $fit(EV,EV_\star)$ est la suivante :

$$fit(EV,EV_\star) = \begin{cases} \sqrt{(a\tilde{q}_1 - Aq_1(EV_\star,EV))^2 + (a\tilde{q}_2 - Aq_2(EV_\star,EV))^2}, \\ \text{si } Aq_1(EV_\star,EV) > 0 \text{ et } Aq_2(EV_\star,EV) > 0 \\ 0 \text{ sinon} \end{cases}$$

qui correspond en quelque sorte à une distance du point de vue de la validité entre un ensemble virtuel particulier de variables (dont les valeurs Aq_1

Algorithme 6 Sélection de Variables pour l'Apprentissage Non Supervisé

1. Données : l'ERD EV

2. En une unique passe sur les données bâtir les $\frac{p(p-1)}{2}$ tables de contingence nécessaires aux calculs des mesures d'adéquation entre ensembles de variables présentées préalablement.

3. Fixer les paramètres de l'AG : *nombre de générations, taille de la population, Probabilité de Croisement, Probabilité de mutation*

4. Lancer l'AG utilisant la fonction de fitness spécifique définie ci-dessous.

5. Sélectionner le meilleur sous espace déterminé par l'AG

et Aq_2 seraient respectivement $a\tilde{q}_1$ et $a\tilde{q}_2$) et l'espace $EV_\star$. En fait, nous fixons $a\tilde{q}_1 = a\tilde{q}_2 = très\ forte\ valeur$ de manière à conférer à l'espace virtuel particulier l'aspect d'une sorte d'espace idéal du point de vue de l'adéquation avec EV. Ainsi, cette fonction de fitness correspond en somme à une distance du point de vue de la validité entre un espace virtuel idéal du point de vue de l'adéquation avec EV. Cette fonction de fitness doit donc être minimisée.

5.3.6 Evaluations Expérimentales

Afin d'évaluer la qualité et l'intérêt de la méthode que nous proposons nous présentons ici deux types d'expérimentations : l'une sur des jeux de données synthétiques, l'autre sur des jeux de données provenant de la collection de l'UCI.

5.3.6.1 Expérience #1 : Evaluation expérimentale sur jeux de données synthétiques

Description L'objectif est ici de tester dans quelle mesure notre méthode détecte les variables pertinentes (vecteur d'une véritable source d'informations), pour cela nous avons bâti le jeu de données synthétique suivant :
Ce jeu de données comprend 1000 objets caractérisés par 9 variables véritablement porteuses d'information V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9, et par un ensemble d'autres variables correspondant à du bruit.

En définitive, les 250 premiers objets possèdent tous la même valeur D pour les variables V_1, V_2, V_3 ; quant aux variables restantes une valeur parmi A, B et C leur est assignée de manière aléatoire (la probabilité d'assignation de chaque valeur est $\frac{1}{3}$).

Pour les 250 objets suivants, ils possèdent tous la même valeur D pour les variables V_3, V_4, V_5 ; quant aux variables restantes une valeur parmi A, B et C leur est assignée de manière aléatoire (la probabilité d'assignation de chaque valeur est $\frac{1}{3}$).

Les 250 objets suivants possèdent tous la même valeur D pour les variables V_5, V_6, V_7 ; quant aux variables restantes une valeur parmi A, B et C leur est

assignée de manière aléatoire (la probabilité d'assignation de chaque valeur est $\frac{1}{3}$).

Finalement, les 250 objets restants possèdent tous la même valeur D pour les variables V_7, V_8, V_9 ; quant aux variables restantes une valeur parmi A, B et C leur est assignée de manière aléatoire (la probabilité d'assignation de chaque valeur est $\frac{1}{3}$).

Nous illustrons dans la figure 5.12, la composition du jeu de données. On peut ainsi se rendre compte que seules les 9 premières variables sont sources d'informations et que la structure des données est donc une partition des objets en 4 classes.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	...	V_i	...	V_p
o_1	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_a	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_{250}	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_{251}	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_b	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_{500}	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_{501}	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_c	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_{750}	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_{751}	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_d	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C
o_{1000}	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C	D	D	D	A, B, C	A, B, C	A, B, C	A, B, C	A, B, C

FIG. 5.12 –: *Jeu de données synthétiques*

Les expérimentations menées sont les suivantes : nous avons exécuté plusieurs processus de SdV pour 6 jeux de données composés des 1000 objets caractérisés par les variables V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9 ainsi que par respectivement :

- 9 variables "bruit" pour le premier jeu de données (soit un jeu de données composé de 18 variables dont 50% sont sources d'informations)
- 18 variables "bruit" pour le deuxième jeu de données (soit un jeu de données composé de 27 variables dont $\frac{1}{3}$ sont sources d'informations)
- 27 variables "bruit" pour le troisième jeu de données (soit un jeu de données composé de 36 variables dont 25% sont sources d'informations)
- 36 variables "bruit" pour le quatrième jeu de données (soit un jeu de données composé de 45 variables dont 20% sont sources d'informations)
- 81 variables "bruit" pour le cinquième jeu de données (soit un jeu de données composé de 90 variables dont 10% sont sources d'informations)

– 171 variables "bruit" pour le sixième jeu de données (soit un jeu de données composé de 180 variables dont 5% sont sources d'informations).

Pour chacun des 6 jeux de données, nous avons ensuite lancé 5 séries de 5 processus de SdV :

– la première série étant caractérisée par un nombre de générations valant *50* pour l'AG utilisé ;
– la deuxième série étant caractérisée par un nombre de générations valant *100* pour l'AG utilisé ;
– la troisième série étant caractérisée par un nombre de générations valant *500* pour l'AG utilisé ;
– la quatrième série étant caractérisée par un nombres de générations valant *1000* pour l'AG utilisé ;
– la cinquième série étant caractérisée par un nombres de générations valant *2500* pour l'AG utilisé.

Les autres paramètres de l'AG ont été fixés à : *nombre de chromosomes par génération = 30* ; *probabilité de croisement = 0,98* ; *probabilité de mutation = 0,4* ; *élitisme = oui*.

Analyse des Résultats Les résultats sont présentés dans la figure 5.13 (voir page 153), ils nécessitent toutefois des explications... Notons tout d'abord que chacun des $6 \times 5 \times 5 = 150$[8] processus de SdV réalisés a mené à l'obtention d'un sous-espace de variables comprenant les 9 variables porteuses d'informations (V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9). Ainsi, les différentes courbes décrivent combien de variables "bruit" ont été simultanément sélectionnées avec les 9 variables pertinentes pour chaque série de 5 processus de SdV, elles détaillent pour chaque série :

– la moyenne du pourcentage de variables "bruit" sélectionnées par les 5 processus de SdV de la série;
– le pourcentage le plus faible de variables "bruit" sélectionnées (i.e. le pourcentage de variables "bruit" sélectionnées par le processus de SdV que l'on peut qualifier de "meilleur") ;
– le pourcentage le plus fort de variables "bruit" sélectionnées (i.e. le pourcentage de variables "bruit" sélectionnées par le processus de SdV que l'on peut qualifier de "moins bon").

Le premier point intéressant réside dans la capacité de la méthode à ne pas omettre de variables pertinentes dans la sélection qu'elle effectue, et ce, même lorsque la portion des variables pertinentes est très faible (5%) et que, simultanément, le nombre de générations de l'AG est très faible (50) (pour des nombres si faibles de générations on peut réellement considérer que le processus d'optimisation associé à l'utilisation de l'AG n'est pas arrivé à terme).

8. $=$ *nombre de jeux de données différents*$\times$*nombre de séries de processus de SdV différentes* $\times$ *nombre de processus de SdV par séries de processus de SdV*

Notons que le temps de calcul associé à ces traitements n'a été au maximum que d'une quinzaine de minutes pour les processus les plus longs (i.e. ceux impliquant le plus de variables et le plus grand nombre de générations) (temps de calcul obtenu pour un logiciel développé en Pascal Objet sous Delphi et exécuté sur un PC 128 Mo Ram, 600 Mhz). De plus, la complexité algorithmique de la méthode est indépendante du nombre d'individus une fois la passe sur le jeu de données réalisée et les tables de contingence dérivées.

Concernant le pourcentage de variables non pertinentes (donc "indésirables") introduites dans les sous ensembles de variables sélectionnées, on observe :

- qu'il est nul (resp. quasi nul) pour les jeux de données composés d'au moins 25% (resp. 20%) de variables pertinentes; et ce même pour des nombres de générations très faibles (50) ;
- que, concernant les jeux de données comportant 10% ou moins de 10% de variables pertinentes, la sélection de l'ensemble optimal de variables ($EV_\star = \{V_1,V_2,V_3,V_4,V_5,V_6,V_7,V_8,V_9\}$) est obtenue pour des nombres de générations supérieurs ou égaux à 1000.

La méthode apparaît donc comme excellente ici, car, les indices ainsi que la fonction de fitness utilisés rendent réellement compte de ce qu'est un bon sous ensemble de variables, et de plus, le processus d'optimisation utilisé permet la découverte du sous ensemble optimal tout en n'impliquant pas un temps de calcul démesuré (une quinzaine de minutes pour le cas le moins favorable). A titre indicatif, pour le jeu de données comportant 180 variables, le nombre de sous ensembles non vides de l'ERD est $2^{180} - 1 = 1{,}53 \times 10^{54}$, le nombre maximal de sous ensembles testés (dans le cas de 2500 générations et en admettant qu'un sous espace n'est évalué qu'une seule fois par l'AG) est $2500 \times 30 = 75000$, la comparaison entre ces deux valeurs montre bien l'efficacité du processus de recherche...

Ainsi, sur ces exemples synthétiques (certes relativement simplistes) la méthode que nous proposons semble d'une efficacité redoutable. Notons enfin que l'application d'algorithmes de cns sur le jeu de données "réduit" mènerait bien à la découverte de la structure en 4 classes et que le temps de calcul associé serait réduit d'un facteur allant de 2 à 20 (resp. 4 à 400) dans le cas d'algorithme possédant une complexité linéaire (resp. quadratique) selon le nombre de variables.

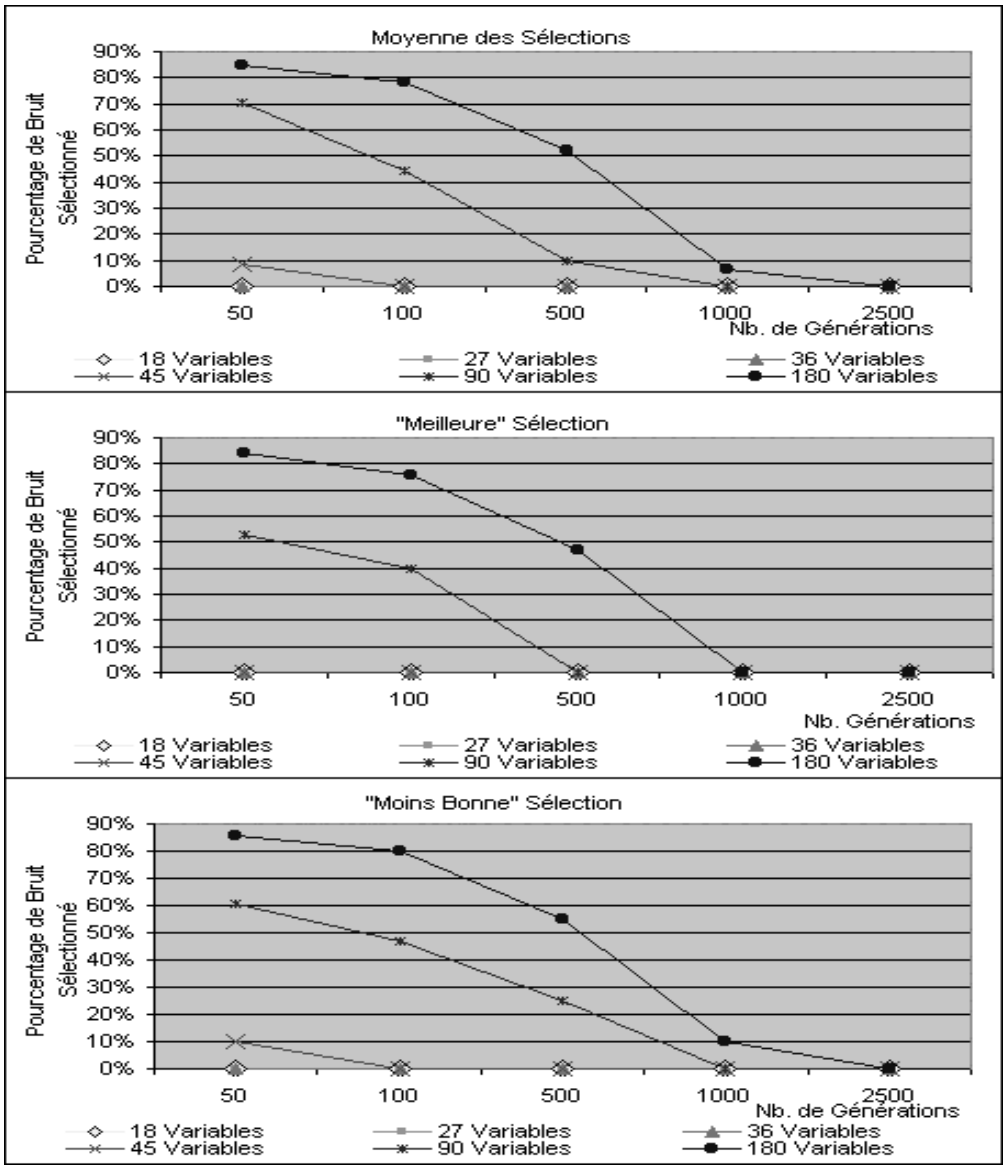

FIG. 5.13 –: *Résultats des expériences sur jeux de données synthétiques pour l'évaluation de la méthode de SdV en apprentissage non supervisé*

5.3.6.2 Expérience #2 : Evaluation Expérimentale sur Jeux de Données de l'UCI

Description Nous avons mené ici les mêmes expériences qu'au chapitre précédent : les jeux de données Small Soybean Diseases ainsi que Mushrooms sont utilisés pour réaliser diverses cns puis tester la validité de ces cns en considérant les jeux de données (les ERDs) dans leur intégralité.

De plus, nous avons évidemment mené les mêmes expérimentations en ne considérant pour chaque jeu de données que les variables sélectionnées par notre méthodologie de SdV :

- Pour le jeu de données : Small Soybean Diseases : seules 9 variables (plant-stand, precip., temp, area-damaged, stem-cankers, canker-lesion, int-discolor, sclerotia, fruit-pods) ont été sélectionnées parmi les 35 variables du jeu de données ;
- pour le jeu de données Mushrooms seules 15 variables (bruises?, odor, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, spore-print-color, population, habitat) ont été sélectionnées parmi les 22 variables.

La question est alors de savoir si la validité des cns obtenues par application des algorithmes de cns (KEROUAC et K-Modes) sur le jeu de données "réduit" (ERD "réduit") est aussi bonne ou meilleure que celle des cns obtenues par application des algorithmes de cns (KEROUAC et K-Modes) sur le jeu de données "complet" (ERD "complet").

Nous avons donc utilisé les méthodes de cns pour données catégorielles KEROUAC et K-Modes avec des paramètres différents (des valeurs différentes pour le facteur de granularité pour KEROUAC et des nombres de classes différents pour les K-Modes) de manière à générer des cns possédant des nombres de classes différents (les paramètres ont été fixés de manière à obtenir pour le jeu de données Small Soybean des cns en 2, 3, 4, 5, 6, 7, 8, 9, 10 classes et des cns en 2, 3, 4, 5, 6, 7, 8, 9 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 classes pour le jeu de données Mushrooms). Ces expériences ont donc été menées à la fois sur le jeu de données "complet" et sur le jeu de données "réduit".

REMARQUE : Pour la méthode des K-Modes nous avons réalisé pour chaque nombre de classes 10 expériences différentes et conservé la cns correspondant à la meilleure valeur pour le critère QKM (critère sous-jacent à cette méthode).

Les résultats sont exposés (à titre indicatif), pour le jeu de données Mushrooms, dans les tableaux des figures 5.16 et 5.17 (page 159 et 160). Ces tableaux

donnent les informations suivantes pour chaque partition (cns):

- le nombre de classes (#Cl.),
- la valeur à laquelle a été fixée le facteur de granularité (α) (si la cns a été obtenue grâce à KEROUAC),
- les valeurs de xv_1 et xv_2 calculées dans l'ERD "complet" (pour l'ensemble des cns qu'elles aient été obtenues par application d'un algorithme de cns sur l'intégralité du jeu de données (ERD "complet") ou non (ERD "réduit")),
- les valeurs de xv_1 et xv_2 calculées dans l'ERD "réduit" (si la cns a été obtenue par application d'un algorithme de cns sur le jeu de données "réduit" (ERD "réduit")),
- la valeur du critère à minimiser sous-jacent à la méthode K-Modes (QKM) calculée dans l'ERD "complet", mais également la valeur calculée dans l'ERD "réduit" si la cns a été obtenue par application d'un algorithme de cns sur le jeu de données "réduit" (ERD "réduit"),
- la valeur du critère à minimiser sous-jacent à la méthode KEROUAC (NCC) calculée dans l'ERD "complet", mais également la valeur calculée dans l'ERD "réduit" si la cns a été obtenue par application d'un algorithme de cns sur le jeu de données "réduit" (ERD "réduit"),
- le taux de correction (T.C.) de chaque partition pour le concept "pathologie".

Analyse des Résultats

Jeu de Données Mushrooms Considérons tout d'abord les résultats associés aux expérimentations sur le jeu de données Mushrooms (voir tableaux des figures 5.16 et 5.17 ainsi que les figures 5.14 et 5.15).

Concernant l'évaluation de la validité des cns par l'intermédiaire de la **méthodologie présentée au chapitre 4** (évaluation réalisée en considérant l'ERD "complet"), on observe sur la figure 5.14 que, du point de vue de la validité, les cns réalisées sur le jeu de données considéré dans son intégralité (ERD "complet") ou partiellement (ERD "réduit") sont très proches voire équivalentes que la méthode employée soit KEROUAC ou les K-Modes. On peut même observer que la validité des cns obtenues sur l'ERD "réduit" semble de manière générale meilleure. Concernant la cns la plus valide, la cns la plus valide obtenue dans l'ERD "complet" est celle comprenant 19 classes obtenue par KEROUAC pour le jeu de donnée "complet" et elle est du point de vue de la validité quasi-équivalente à celle comprenant 17 classes obtenue par KEROUAC sur le jeu de données réduit. Pour les cns obtenues par l'intermédiaire des K-Modes, l'utilisation de l'ERD "réduit" mène à des cns équivalentes ou meilleures du point de vue de la validité.

On peut également observer que l'évaluation de la validité des cns obtenues dans l'ERD "réduit" donne un profil de courbes très similaire à celui observé dans l'ERD "complet" (voir courbes Kerouac, Kerouac + SdV, K-Modes (Meilleur), K-Modes + SdV (Meilleur) et les courbes Kerouac + SdV (dans l'ERD "réduit"), K-Modes + SdV (Meilleur) (dans l'ERD "réduit") de la figure 5.14). De plus, les valeurs des indices xv_1 et xv_2 sont proches. Ce dernier point est très intéressant car il semble montrer que l'évaluation de la validité réalisée dans l'ERD "réduit" est conforme à celle réalisée dans l'ERD "complet" ce qui implique que l'on peut se contenter de procéder à l'évaluation de la validité dans l'ERD "réduit" (et ainsi limiter le coût calculatoire nécessaire à la validation).

D'après ce mode d'évaluation de la validité de cns, notre méthodologie de sélection de variables fournit de très bons ensembles de variables puisque la validité des cns bâties sur l'ERD "réduit" est quasi-équivalente à celle des cns obtenues à partir de l'ERD "complet"...

Si on considère maintenant les valeurs des **critères** QKM **et** NCC ainsi que celle du **taux de correction** (T.C.) pour chaque cns, on observe là encore (voir figure 5.15) la presque parfaite adéquation des résultats obtenus lors de l'utilisation de l'ERD "complet" et lors de l'utilisation de l'ERD "réduit". Ces résultats étayent, eux aussi, la conclusion que les cns obtenues par utilisation de l'ERD "réduit" présentent un niveau de validité équivalent à celles obtenues en accédant à l'ERD complet.

Les résultats obtenus sur ce jeu de données sont frappants et semblent démontrer la très grande efficacité de notre méthodologie de sélection de variables pour réduire la dimension de l'ERD tout en conservant un niveau de validité des cns d'excellente qualité. Notons de plus, que la réduction de l'ERD permet à la fois la réduction des coûts calculatoire et de stockage associés au processus de cns et la réduction de ces mêmes coûts pour le processus d'évaluation de la validité des cns.

Jeu de Données Small Soybean Disease Nous ne détaillons pas l'analyse des résultats obtenus sur le jeu K-Modes (voir figure 5.18) qui tendent à apporter les mêmes conclusions. De plus, cette fois-ci, la réduction de la dimension de l'ERD est plus importante : seulement 25,7% des variables sont conservées (68,2% des variables étaient conservées pour le jeu de données Mushrooms).

Notons cependant l'intégration d'un élément supplémentaire pour l'évaluation de la validité des cns obtenues par application des algorithmes de cns sur l'ERD "réduit" : un des graphiques de la figure 5.18 permet l'évaluation de l'adéquation entre les couples de cns possédant le même nombre de classes[9].

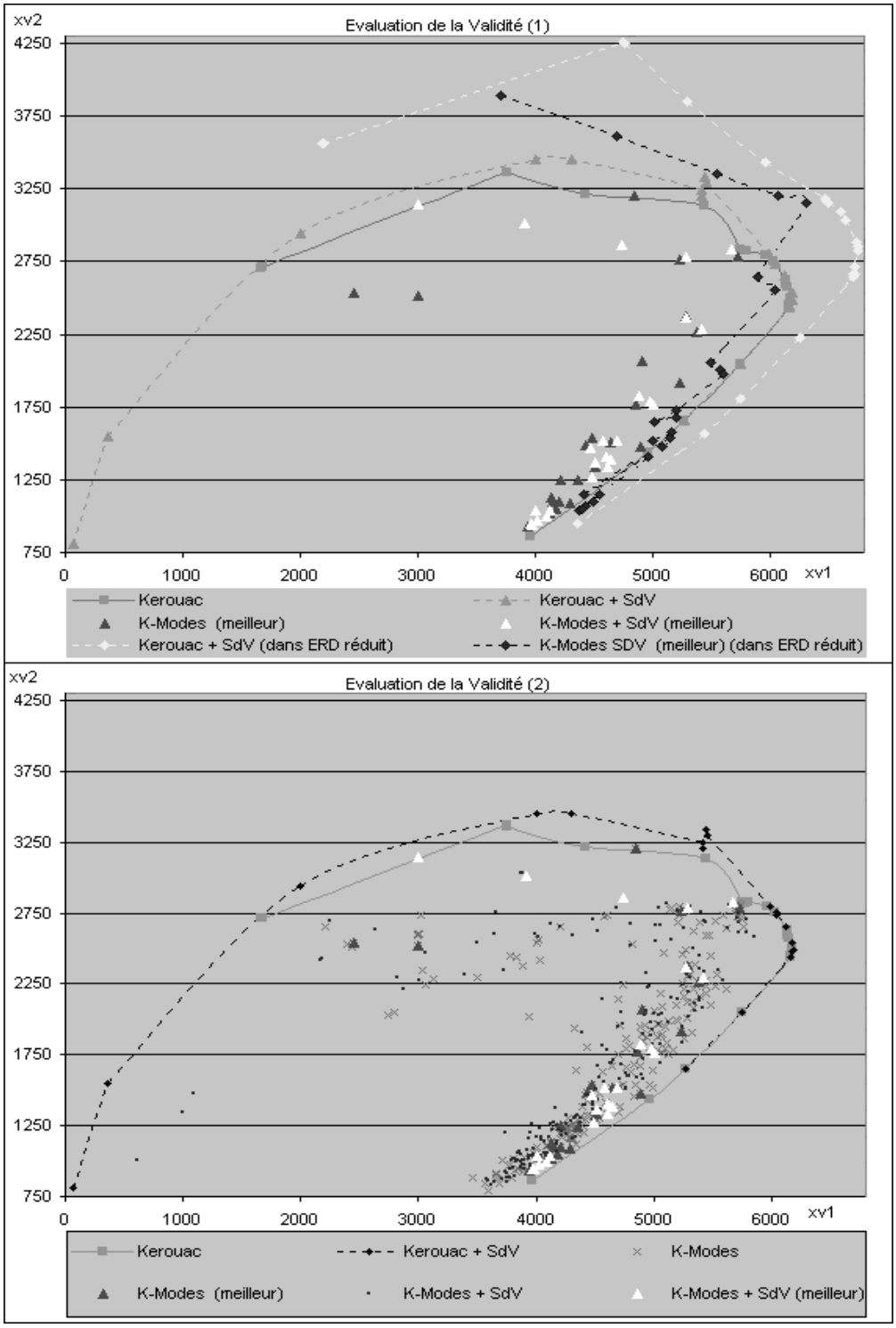

FIG. 5.14 –: *Résultats des expériences sur le jeu de données Mushrooms pour l'évaluation de la méthode de SdV en apprentissage non supervisé*

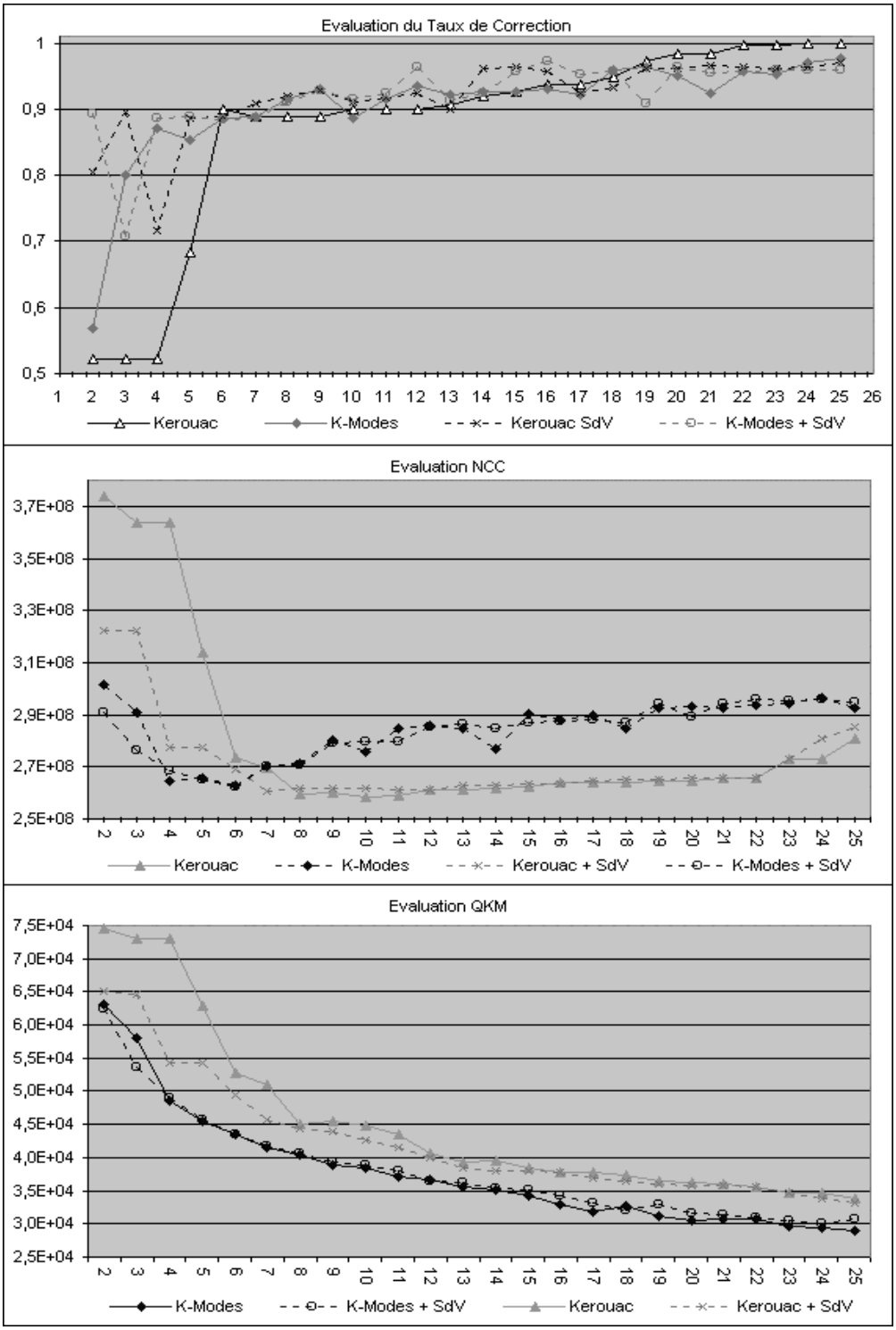

FIG. 5.15 –: *Résultats des expériences sur le jeu de données Mushrooms pour l'évaluation de la méthode de SdV en apprentissage non supervisé*

Algo.	SdV	#Cl.	α	*xv1* (ERD complet)	*xv2* (ERD complet)	*xv1* (ERD réduit)	*xv2* (ERD réduit)	*QKM* (ERD complet)	*QKM* (ERD réduit)	*NCC* (ERD complet)	*NCC* (ERD réduit)	T.C.
Kerouac	Non	2	0,5	65,28	814,63	x	x	74555	x	3,74E+08	x	52,1%
Kerouac	Non	3	0,55	364,19	1548,93	x	x	73003	x	3,64E+08	x	52,1%
Kerouac	Non	4	0,6	364,21	1548,96	x	x	72909	x	3,64E+08	x	52,1%
Kerouac	Non	5	0,65	1998,05	2942,31	x	x	62858	x	3,14E+08	x	68,3%
Kerouac	Non	6	0,7	4005,27	3455,12	x	x	52758	x	2,74E+08	x	90,0%
Kerouac	Non	7	0,8	4302,08	3451,16	x	x	51066	x	2,69E+08	x	88,8%
Kerouac	Non	8	0,85	5416,25	3247,15	x	x	45102	x	2,60E+08	x	88,8%
Kerouac	Non	9	0,865	5414,86	3206,88	x	x	45466	x	2,60E+08	x	88,8%
Kerouac	Non	10	1	5440,97	3334,72	x	x	44790	x	2,59E+08	x	89,9%
Kerouac	Non	11	1,075	5455,89	3294,50	x	x	43414	x	2,59E+08	x	89,9%
Kerouac	Non	12	1,1	5990,08	2799,89	x	x	40746	x	2,61E+08	x	89,9%
Kerouac	Non	13	1,2	6038,25	2750,24	x	x	39308	x	2,61E+08	x	90,5%
Kerouac	Non	14	1,225	6041,13	2731,23	x	x	39536	x	2,62E+08	x	91,9%
Kerouac	Non	15	1,25	6115,69	2655,86	x	x	38456	x	2,62E+08	x	92,5%
Kerouac	Non	16	1,475	6179,03	2544,00	x	x	37748	x	2,64E+08	x	93,7%
Kerouac	Non	17	1,482	6179,07	2543,76	x	x	37692	x	2,64E+08	x	93,7%
Kerouac	Non	18	1,5	6180,72	2535,83	x	x	37260	x	2,64E+08	x	94,9%
Kerouac	Non	19	1,825	6181,50	2486,96	x	x	36540	x	2,65E+08	x	97,2%
Kerouac	Non	20	1,85	6178,93	2480,13	x	x	36300	x	2,65E+08	x	98,4%
Kerouac	Non	21	2	6163,76	2443,12	x	x	35916	x	2,65E+08	x	98,4%
Kerouac	Non	22	2,5	6158,23	2433,17	x	x	35628	x	2,66E+08	x	99,6%
Kerouac	Non	23	2,85	5748,99	2045,88	x	x	34764	x	2,73E+08	x	99,6%
Kerouac	Non	24	3	5747,60	2044,67	x	x	34700	x	2,73E+08	x	100,0%
Kerouac	Non	25	3,05	5264,11	1652,58	x	x	33836	x	2,81E+08	x	100,0%
Kerouac	Oui	2	0,4	1665,16	2709,24	2189,10	3561,69	65034	53328	3,23E+08	2,39E+08	68,2%
Kerouac	Oui	3	0,408	1674,90	2715,43	2196,79	3561,54	64680	53010	3,22E+08	2,38E+08	68,2%
Kerouac	Oui	4	0,5	3753,67	3365,00	4746,65	4255,17	54350	42766	2,78E+08	1,72E+08	88,8%
Kerouac	Oui	5	0,55	3758,18	3359,97	4756,24	4252,28	54192	42608	2,77E+08	1,72E+08	89,1%
Kerouac	Oui	6	0,6	4423,19	3218,01	5297,42	3854,04	49350	38826	2,69E+08	1,60E+08	89,4%
Kerouac	Oui	7	0,7	5441,62	3129,19	5963,08	3429,05	45782	35834	2,60E+08	1,49E+08	89,4%
Kerouac	Oui	8	0,725	5752,20	2829,95	6464,36	3180,32	44294	34054	2,62E+08	1,45E+08	89,2%
Kerouac	Oui	9	0,8	5781,35	2821,14	6485,33	3164,67	43934	33742	2,62E+08	1,44E+08	89,9%
Kerouac	Oui	10	0,85	5805,33	2822,65	6493,72	3157,36	42574	32862	2,62E+08	1,44E+08	89,9%
Kerouac	Oui	11	0,9	5961,70	2792,17	6595,79	3089,15	41480	32290	2,61E+08	1,44E+08	91,6%
Kerouac	Oui	12	0,915	6026,27	2758,08	6635,10	3036,73	40040	31138	2,61E+08	1,43E+08	91,6%
Kerouac	Oui	13	1	6131,39	2629,16	6733,76	2887,45	38400	29794	2,63E+08	1,43E+08	93,6%
Kerouac	Oui	14	1,01	6134,28	2626,20	6735,96	2883,79	38096	29538	2,63E+08	1,43E+08	94,2%
Kerouac	Oui	15	1,05	6129,08	2591,43	6746,70	2852,57	37984	29382	2,63E+08	1,43E+08	97,4%
Kerouac	Oui	16	1,3	6144,52	2575,06	6746,34	2827,27	37872	29242	2,63E+08	1,44E+08	97,8%
Kerouac	Oui	17	1,4	6173,85	2496,00	6725,89	2719,18	36976	28494	2,65E+08	1,44E+08	96,9%
Kerouac	Oui	18	1,5	6166,37	2450,71	6713,71	2668,24	36356	27942	2,65E+08	1,44E+08	96,8%
Kerouac	Oui	19	1,75	6163,88	2443,91	6710,33	2660,57	36116	27702	2,65E+08	1,45E+08	98,0%
Kerouac	Oui	20	1,825	6158,34	2433,96	6702,07	2648,86	35828	27510	2,66E+08	1,45E+08	99,2%
Kerouac	Oui	21	2	6158,16	2433,40	6701,53	2648,12	35684	27398	2,66E+08	1,45E+08	99,6%
Kerouac	Oui	22	2,1	6158,23	2433,17	6701,17	2647,69	35628	27358	2,66E+08	1,45E+08	99,6%
Kerouac	Oui	23	2,5	5748,99	2045,88	6250,31	2224,29	34764	26494	2,73E+08	1,49E+08	99,6%
Kerouac	Oui	24	2,4	5265,61	1653,76	5749,70	1805,80	33900	25630	2,81E+08	1,54E+08	99,6%
Kerouac	Oui	25	2,75	4964,02	1433,71	5437,69	1570,52	33236	24950	2,85E+08	1,57E+08	99,9%

FIG. 5.16 –: *Résultats des expériences sur le jeu de données Mushrooms pour l'évaluation de la méthode de SdV en apprentissage non supervisé*

Algo.	SdV	#Cl.	α	$xv1$ (ERD complet)	$xv2$ (ERD complet)	$xv1$ (ERD réduit)	$xv2$ (ERD réduit)	QKM (ERD complet)	QKM (ERD réduit)	NCC (ERD complet)	NCC (ERD réduit)	T.C.
K-Modes	Non	2	x	2446,97	2536,24	x	x	63002	x	3,01E+08	x	56,8%
K-Modes	Non	3	x	2997,53	2518,52	x	x	57950	x	2,91E+08	x	80,1%
K-Modes	Non	4	x	4843,13	3203,64	x	x	48639	x	2,65E+08	x	87,1%
K-Modes	Non	5	x	5224,51	2761,32	x	x	45558	x	2,66E+08	x	85,3%
K-Modes	Non	6	x	5718,20	2783,41	x	x	43395	x	2,63E+08	x	88,7%
K-Modes	Non	7	x	5279,42	2372,88	x	x	41509	x	2,70E+08	x	88,8%
K-Modes	Non	8	x	5373,55	2262,90	x	x	40382	x	2,71E+08	x	91,5%
K-Modes	Non	9	x	4859,74	1771,83	x	x	38830	x	2,80E+08	x	92,9%
K-Modes	Non	10	x	4902,09	2068,90	x	x	38479	x	2,76E+08	x	88,7%
K-Modes	Non	11	x	4475,37	1542,21	x	x	37016	x	2,85E+08	x	91,6%
K-Modes	Non	12	x	4425,79	1486,61	x	x	36708	x	2,86E+08	x	93,4%
K-Modes	Non	13	x	4645,16	1512,24	x	x	35587	x	2,85E+08	x	92,2%
K-Modes	Non	14	x	5230,93	1914,67	x	x	35026	x	2,77E+08	x	92,7%
K-Modes	Non	15	x	4207,03	1249,32	x	x	34157	x	2,90E+08	x	92,5%
K-Modes	Non	16	x	4508,57	1334,15	x	x	32918	x	2,88E+08	x	93,0%
K-Modes	Non	17	x	4356,21	1243,66	x	x	31816	x	2,90E+08	x	92,1%
K-Modes	Non	18	x	4888,91	1474,66	x	x	32615	x	2,85E+08	x	95,9%
K-Modes	Non	19	x	4202,00	1099,96	x	x	31235	x	2,93E+08	x	96,3%
K-Modes	Non	20	x	4141,25	1097,39	x	x	30545	x	2,93E+08	x	95,1%
K-Modes	Non	21	x	4127,52	1125,23	x	x	30717	x	2,93E+08	x	92,5%
K-Modes	Non	22	x	4173,15	1046,59	x	x	30620	x	2,94E+08	x	95,7%
K-Modes	Non	23	x	4123,91	1015,15	x	x	29524	x	2,94E+08	x	95,3%
K-Modes	Non	24	x	3933,08	932,79	x	x	29412	x	2,96E+08	x	97,0%
K-Modes	Non	25	x	4289,07	1092,48	x	x	28878	x	2,93E+08	x	97,6%
K-Modes	Oui	2	x	2995,12	3146,26	3708,20	3895,33	62524	50820	2,91E+08	1,93E+08	89,2%
K-Modes	Oui	3	x	3910,78	3012,38	4694,96	3616,42	53594	42440	2,76E+08	1,69E+08	70,9%
K-Modes	Oui	4	x	4737,88	2859,79	5548,91	3349,33	49055	37948	2,69E+08	1,54E+08	88,7%
K-Modes	Oui	5	x	5283,26	2786,86	6063,65	3198,50	45755	35518	2,65E+08	1,48E+08	88,9%
K-Modes	Oui	6	x	5662,85	2830,27	6301,53	3149,48	43528	33588	2,62E+08	1,46E+08	88,4%
K-Modes	Oui	7	x	5275,70	2366,85	5890,17	2642,53	41818	31883	2,70E+08	1,50E+08	88,6%
K-Modes	Oui	8	x	5410,69	2290,32	6045,29	2558,94	40725	30886	2,71E+08	1,49E+08	91,2%
K-Modes	Oui	9	x	4878,43	1827,13	5499,81	2059,85	39425	29420	2,79E+08	1,54E+08	92,8%
K-Modes	Oui	10	x	4979,68	1791,95	5567,07	2003,33	38844	29000	2,80E+08	1,54E+08	91,5%
K-Modes	Oui	11	x	5006,49	1765,97	5598,28	1974,72	37990	28176	2,80E+08	1,54E+08	92,3%
K-Modes	Oui	12	x	4575,60	1515,89	5205,80	1724,68	36494	26726	2,85E+08	1,57E+08	96,3%
K-Modes	Oui	13	x	4469,83	1467,21	5013,00	1645,51	36182	26619	2,86E+08	1,59E+08	91,2%
K-Modes	Oui	14	x	4687,35	1513,87	5194,15	1677,55	35411	26126	2,85E+08	1,58E+08	92,4%
K-Modes	Oui	15	x	4602,39	1407,95	5166,33	1580,47	35056	25624	2,87E+08	1,58E+08	95,6%
K-Modes	Oui	16	x	4509,15	1365,88	5005,07	1516,10	34145	25165	2,88E+08	1,59E+08	97,1%
K-Modes	Oui	17	x	4608,15	1337,92	5075,67	1473,66	33063	24279	2,88E+08	1,59E+08	95,3%
K-Modes	Oui	18	x	4641,12	1385,40	5152,63	1538,09	32072	23228	2,87E+08	1,58E+08	95,7%
K-Modes	Oui	19	x	4005,76	1037,15	4418,26	1143,95	32850	23820	2,94E+08	1,64E+08	90,8%
K-Modes	Oui	20	x	4481,29	1268,39	4966,05	1405,59	31577	22792	2,89E+08	1,60E+08	96,3%
K-Modes	Oui	21	x	4117,64	1041,47	4550,10	1150,85	31303	22476	2,94E+08	1,63E+08	95,5%
K-Modes	Oui	22	x	3972,32	942,44	4378,35	1038,78	30987	22247	2,96E+08	1,64E+08	95,6%
K-Modes	Oui	23	x	4013,55	970,78	4421,22	1069,38	30608	21884	2,95E+08	1,64E+08	95,8%
K-Modes	Oui	24	x	3964,29	945,64	4395,04	1048,40	30040	21174	2,96E+08	1,64E+08	95,9%
K-Modes	Oui	25	x	4091,70	999,62	4496,15	1098,43	30664	21970	2,95E+08	1,64E+08	96,0%

FIG. 5.17 –: *Résultats des expériences sur le jeu de données Mushrooms pour l'évaluation de la méthode de SdV en apprentissage non supervisé*

Nous évaluons en effet l'adéquation entre :

- les couples de cns composés de cns possédant le même nombre de classes et tels que l'une des cns corresponde à celle obtenue par la méthode KEROUAC sur l'ERD "complet", l'autre à celle obtenue par la méthode KEROUAC sur l'ERD "réduit" (courbe nommée KEROUAC) ;
- les couples de cns composés de cns possédant le même nombre de classes et tels que l'une des cns correspond à la meilleure cns[10] obtenue par la méthode K-Modes sur l'ERD "complet", l'autre à la meilleure cns par la méthode K-Modes sur l'ERD "réduit" (courbe nommée K-Modes + SdV);
- les couples de cns composés de cns possédant le même nombre de classes tels que l'une des cns correspond à la meilleure cns obtenue par la méthode K-Modes sur l'ERD "complet", l'autre à la cns obtenue par la méthode K-Modes sur l'ERD "réduit" et impliquant la meilleure adéquation (i.e. la plus faible valeur de l'indice)(courbe nommée min. K-Modes + SdV);
- les couples de cns composés de cns possédant le même nombre de classes tesl que l'une des cns correspond à la meilleure cns obtenue par la méthode K-Modes sur l'ERD "complet", l'autre à la cns obtenue par la méthode K-Modes sur l'ERD "réduit" et impliquant la moins bonne adéquation (i.e. la plus forte valeur de l'indice)(courbe nommée max. K-Modes + SdV);
- les couples de cns composés de cns possédant le même nombre de classes tels que l'une des cns correspondant à la meilleure cns obtenue par la méthode K-Modes sur l'ERD "complet", l'autre à la cns obtenue par la méthode K-Modes sur l'ERD "complet" et impliquant la meilleure adéquation (i.e. la plus faible valeur de l'indice)(courbe nommée min. K-Modes);
- les couples de cns composés de cns possédant le même nombre de classes tels que l'une des cns correspond à la meilleure cns obtenue par la méthode K-Modes sur l'ERD "complet", l'autre à la cns obtenue par la méthode K-Modes sur l'ERD "complet" et impliquant la moins bonne adéquation (i.e. la plus forte valeur de l'indice)(courbe nommée max. K-Modes);

La dernière série de valeur sert de témoin : elle montre les valeurs maximales de l'indice d'adéquation pour un couple de cns (ayant le même nombre de classes) lors de l'application du même algorithme de cns (les K-Modes) sur l'intégralité du jeu de données. Cette série montre donc la variabilité maximale pour l'adéquation en considérant des processus de cns appliqués sur le même ERD : l'ERD "complet". Les valeurs obtenues pour l'ensemble des autres

9. l'indice utilisé ici pour l'évaluation de l'adéquation entre cns est l'indice Adq présenté au chapitre suivant, page 176. Plus sa valeur est proche de 0 plus l'adéquation est forte.

10. meilleure cns signifie ici qu'il s'agit de la cns ayant la plus faible valeur pour le critère QKM (critère sous-jacent à la méthode K-Modes)

courbes sont inférieures ou très proches. Cela prouve que les cns obtenues par application des algorithmes de cns sur l'ERD "réduit" présentent une excellente adéquation avec celles obtenues par application des algorithmes de cns sur l'ERD "complet". (Dans le cas de la méthode KEROUAC la valeur de l'indice d'adéquation vaut même 0 pour les couples de cns en 2, 3 et 4 classes ce qui signifie que ces couples de cns sont composés de cns identiques). Cette dernière expérience constitue une nouvelle indication de la qualité de notre méthodologie de SdV pour l'apprentissage non supervisé.

5.3.7 Conclusion

En résumé, nous proposons, une méthode de sélection de variables pour l'apprentissage non supervisé:

- ne nécessitant qu'une unique passe sur le jeu de données, et une complexité algorithmique faible (dans le cas de données catégorielles, la complexité est linéaire selon le nombre d'objets du jeu de données et quadratique selon le nombre de variables du jeu de données) ce qui lui confère une rapidité très intéressante ;
- possédant un coût de stockage faible
- utilisant une extension de la méthodologie préalablement introduite pour la comparaison de la validité de cns et un AG ;

Les évaluations expérimentales ont montré que :

- les cns obtenues sur l'ERD "réduit" sont d'excellente qualité ;
- l'ERD peut être parfois extrêmement "réduit" et la présence d'un fort bruit ne semble pas handicaper cette méthode ;
- concernant le temps de calcul, notre méthode est bonne.

Les remarques concernant les éventuelles améliorations de la méthode sont les mêmes que celles apportées pour la méthode de SdV pour l'apprentissage supervisé :

- notre méthode peut être améliorée du point de vue du coût calculatoire (une réduction du temps de calcul associé par substitution d'une méthode d'optimisation gloutonne à l'AG).
- on peut aisément modifier la structure de l'AG de manière à pouvoir rechercher non pas l'ensemble "optimal" de variables mais le meilleur ensemble de variables tel qu'il comprenne au plus un nombre fixé de variables.

Notons également que cette méthode de SdV pour l'apprentissage non supervisé permet une sélection des variables de l'ERD initial et non une sélection de variables correspondant à une transformation des variables de l'ERD initial comme le permet, par exemple, les approches basées sur l'analyse factorielle.

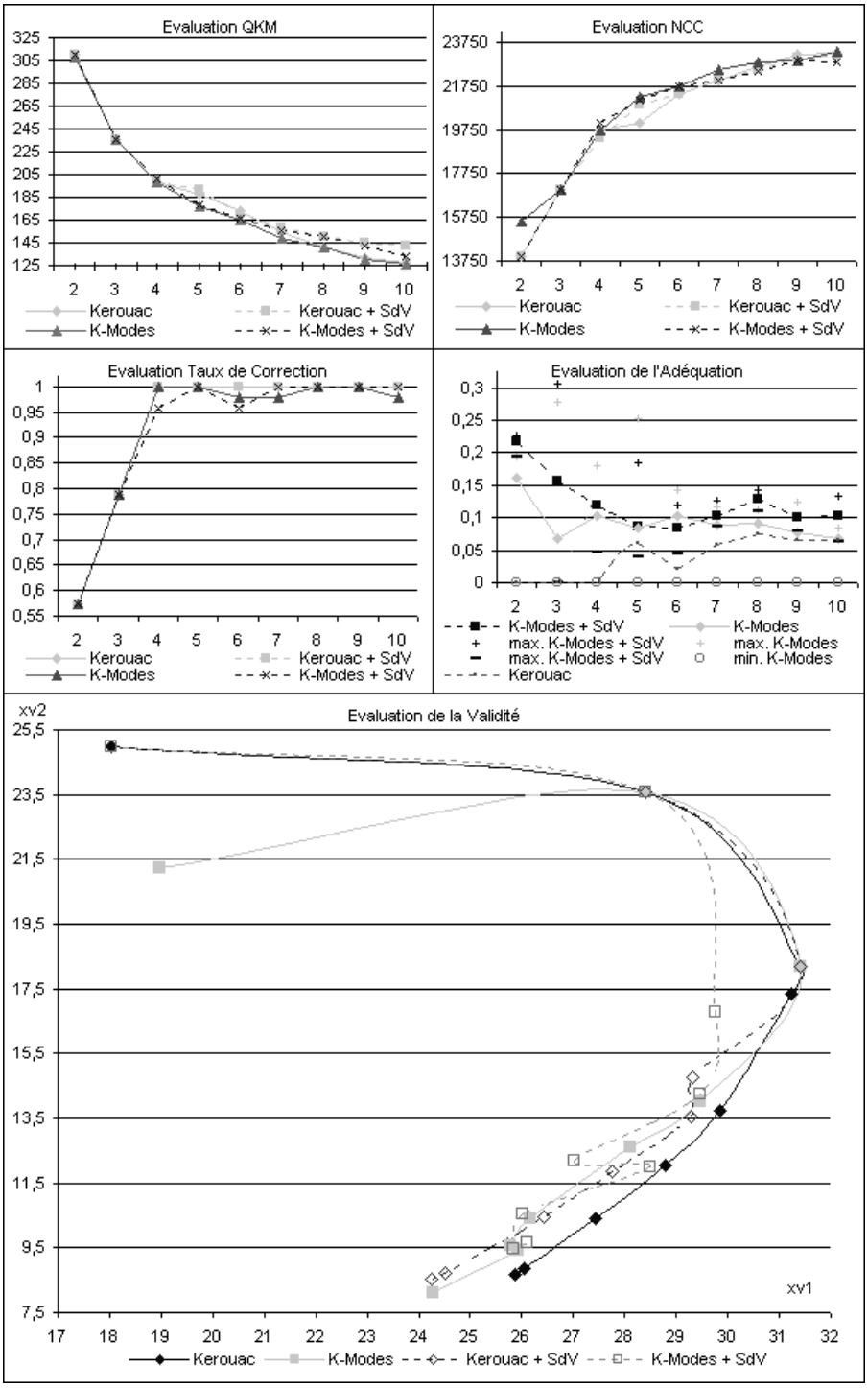

FIG. 5.18 –: *Résultats des expériences sur le jeu de données Small Soybean Disease pour l'évaluation de la méthode de sélection de variables en apprentissage non supervisé*

Ce point est particulièrement intéressant si on veut bâtir un modèle aisément interprétable.

REMARQUE : La méthode que nous venons de proposer peut également être employée si des variables quantitatives sont présentes, cependant, si elle ne nécessité alors qu'une unique passe sur les données, et implique un coût de stockage équivalent, sa complexité quadratique selon le nombre d'objets du jeu de données et linéaire selon le nombre de variables du jeu de données est handicapante du point de vue du coût calculatoire.