Une Approche Filtre pour la Sélection de Variables en Apprentissage Non Supervisé

Pierre-Emmanuel JOUVE *, Nicolas NICOLOYANNIS *
*LABORATOIRE ERIC, Université Lumière - Lyon2, http://eric.univ-lyon2.fr
Bâtiment L, 5 av. Pierre Mendès-France
69 676 BRON cedex FRANCE
pierre.jouve@eric.univ-lyon2.fr, nicoloyannis@univ-lyon2.fr

Résumé. La Sélection de Variable (SV) constitue une technique efficace pour réduire la dimension des espaces d'apprentissage et s'avère être une méthode essentielle pour le pré-traitement de données afin de supprimer les variables bruitées et/ou inutiles. Peu de méthodes de SV ont été proposées dans le cadre de l'apprentissage non supervisé, et, la plupart d'entre elles, sont des méthodes dites "enveloppes" nécessitant l'utilisation d'un algorithme d'apprentissage pour évaluer les sous ensembles de variables. Or, l'approche "enveloppe" est largement mal adaptée à une utilisation lors de cas "réels". En effet, d'une part ces méthodes ne sont pas indépendantes vis à vis des algorithmes d'apprentissage non supervisé qui nécessitent le plus souvent de fixer un certain nombre de paramètres ; mais surtout, il n'existe pas de critères bien adaptés à l'évaluation de la qualité d'apprentissage non supervisé dans des sous espaces différents. Nous proposons et évaluons dans ce papier une méthode "filtre" et donc indépendante des algorithmes d'apprentissage non supervisé. Cette méthode s'appuie sur deux indices permettant d'évaluer l'adéquation entre deux ensembles de variables (entre deux sous espaces).

1 Introduction

La grande dimensionnalité de l'espace de représentation des données est un problème commun en apprentissage. La Sélection de Variables (SV) permet de déterminer quelles sont les variables pertinentes et constitue ainsi une technique efficiente pour la réduction de la dimension. Une variable pertinente pour une tâche d'apprentissage peut être définie comme une variable dont la suppression dégrade de manière significative la qualité de l'apprentissage réalisé. La suppression des variables non pertinentes permet donc la réduction de dimensionnalité, et, peut simultanément impliquer un accroissement de la précision et de la compréhensibilité des modèles bâtis. Il existe deux contextes principaux pour l'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé (clustering). S'il existe nombre de méthodes pour la SV dans le contexte supervisé (Dash et al. 1997), il n'existe que peu de méthodes (la plupart étant récentes) pour le contexte non supervisé. Cela peut être expliqué par le fait qu'il est plus aisé de sélectionner des variables pour l'apprentissage supervisé que pour le clustering. Dans le cadre supervisé, ce qui doit être appris est "connu a priori" alors que cela n'est pas le cas pour le clustering, dès lors, déterminer les variables pertinentes pour cette tâche peut être ardu. Le processus de SV pour le clustering peut être vu comme le processus de

sélection des variables pertinentes pour la structure à découvrir sous jacente (Dash et al. 2000) (cette structure correspond dans la majorité des cas à une partition des objets du jeu de données considéré). Parmi les méthodes de SV proposées pour l'apprentissage non supervisé (Dash et al. 2000, 2002, Devaney et al. 1997, Dy et al. 2000, Kim et al. 2000, Talavera 2000) la plupart correspond à une approche de type "enveloppe". Ces méthodes évaluent les sous ensembles de variables (sous espaces) sélectionnées au moyen d'un algorithme de clustering (cet algorithme utilisant plus tard le sous espace finalement sélectionné pour effectuer la tâche d'apprentissage non supervisé) (par exemple, les Kmeans sont utilisés dans (Dash et al. 2000, Kim et al. 2000) l'algorithme EM dans (Dy et al. 2000). Dans le cadre supervisé, si les méthodes de type "enveloppe" présentent différents désavantages (coût calculatoire élevé, manque de robustesse selon les algorithmes employés), elles sont toutefois intéressantes lorsque la précision de l'apprentissage est très importante. Cependant, contrairement à l'apprentissage supervisé pour lequel il existe un consensus sur la façon d'évaluer la qualité d'un apprentissage, il n'existe pas d'unanimité sur le critère à employer pour évaluer la qualité d'un clustering (de plus, ce critère devrait fonctionner correctement dans différents sous espaces). Ces éléments rendent problématique l'utilisation de l'approche "enveloppe" pour la SV dans le cadre non supervisé. Nous proposons et évaluons donc dans cet article une méthode "filtre" pour la SV. Une méthode filtre est, par définition, indépendante des algorithmes de clustering, et permet donc d'éviter le problème de l'absence d'un critère consensuel pour déterminer la qualité d'un clustering. La méthode proposée se base sur deux indices permettant d'évaluer l'adéquation entre deux ensembles de variables (i.e. permettant de déterminer si deux sous ensembles "véhiculent" la même information).

2 Concepts et Formalismes Introductifs

Cette section, essentielle pour la présentation de notre méthode de SV, consiste en la présentation de deux indices permettant d'évaluer dans quelle mesure deux ensembles de variables véhiculent la même information (par la suite, cette évaluation est nommée évaluation de l'adéquation entre deux ensembles de variables). Nous considérons un problème d'apprentissage non supervisé impliquant un jeu de données DS composé par un ensemble (O) de n objets décrits par un ensemble (SA) de l variables.

Notation 1 *$O = \{o_i, i = 1..n\}$ ensemble de n objets*
$SA = \{A_1,..,A_l\}$ ensemble des l variables décrivant les objets de O.
$o_i = [o_{i_1},..,o_{i_l}]$ un objet de O, o_{i_j} correspond à la valeur prise par o_i pour la variable A_j (cette valeur peut être numérique ou catégorielle)

2.1 Notion de Lien

Dans le cadre des données catégorielles, la notion de similarité entre objets d'un jeu de données est utilisée ; dans cet article, nous lui substituons une extension de cette notion pouvant être appliquée à plusieurs types de données (catégorielle ou numérique). Cette notion est appelée lien selon une variable et se définit comme suit :

Définition 1 ***Lien entre 2 objets:*** *Nous associons à chaque variable A_i une function notée $lien_i$ définissant un lien (une sorte de similarité) ou un non-lien (une sorte de*

dissimilarité) selon la variable A_i entre deux objets de O:

$$lien_i(o_{a_i},o_{b_i}) = \begin{cases} 1 \textit{ si une condition particulière déterminant un lien} \\ \textit{(selon } A_i\textit{) entre les objets } o_a \textit{ et } o_b \textit{ est vérifiée} \\ 0 \textit{ sinon (non-lien)} \end{cases} \quad (1)$$

EXEMPLES :
- Pour une variable catégorielle A_i, on peut naturellement définir $lien_i$ comme suit :
$lien_i(o_{a_i},o_{b_i}) = \begin{cases} 1 \text{ si } o_{a_i} = o_{b_i} \\ 0 \text{ sinon} \end{cases}$
- Pour une variable numérique A_i, on peut, par exemple, définir $lien_i$ comme suit :
$lien_i(o_{a_i},o_{b_i}) = \begin{cases} 1 \text{ si } |o_{a_i} - o_{b_i}| \leq \delta, \text{ avec } \delta \text{ un seuil fixé par l'utilisateur} \\ 0 \text{ sinon} \end{cases}$
- Pour une variable numérique A_i, on peut également envisager une discrétisation et appliquer ensuite la définition de $lien_i$ proposée pour les variables catégorielles.

2.2 Évaluation de l'adéquation entre un ensemble de variables SA et un sous-ensemble $SA_\star$ de SA ($SA_\star \subseteq SA$)

Pour évaluer l'adéquation entre $SA = \{A_1,..,A_l\}$ (l'ensemble des variables décrivant les objets du jeu de données DS) et $SA_\star = \{A_{\star_1},..,A_{\star_m}\}$ un sous-ensemble de SA ($SA_\star \subseteq SA$) nous utilisons quatre indices définis dans (Jouve 2003). Ces indices permettent d'évaluer dans quelle mesure deux ensembles de variables sont en adéquation (i.e. "dans quelle mesure ils véhiculent la même information concernant les objets du jeu de données"). Ils sont présentés de manière relativement intuitive ci-dessous, leur formulation mathématique est donnée page suivante.

Considérons les couples $((o_a,o_b),(A_{\star_j},A_i))$ composés par :
- un couple d'objets (o_a,o_b) tel que $a < b$;
- et un couple de variables $(A_{\star_j},A_i)$ constitué d'une variable $A_{\star_j} \in SA_\star$ et d'une variable $A_i \in SA$ telles que $A_{\star_j} \neq A_i$.

Les indices sont alors les suivants :

- $\widetilde{\mathbf{LL}}(\mathbf{SA_\star},\mathbf{SA})$ qui correspond au nombre de couples $((o_a,o_b),(A_{\star_j},A_i))$ tels que :
  1. (o_a,o_b) est caractérisé par un **lien** selon $A_{\star_j} : lien_{\star_j}(o_{a_{\star_j}},o_{b_{\star_j}}) = 1$,
  2. (o_a,o_b) est caractérisé par un **lien** selon $A_i : lien_i(o_{a_i},o_{b_i}) = 1$.
- $\widetilde{\widetilde{\mathbf{LL}}}(\mathbf{SA_\star},\mathbf{SA})$ qui correspond au nombre de couples $((o_a,o_b),(A_{\star_j},A_i))$ tels que :
  1. (o_a,o_b) est caractérisé par un **non-lien** selon $A_{\star_j} : lien_{\star_j}(o_{a_{\star_j}},o_{b_{\star_j}}) = 0$,
  2. (o_a,o_b) est caractérisé par un **non-lien** selon $A_i : lien_i(o_{a_i},o_{b_i}) = 0$.
- $\widetilde{\mathbf{L}}\widetilde{\widetilde{\mathbf{L}}}(\mathbf{SA_\star},\mathbf{SA})$ qui correspond au nombre de couples $((o_a,o_b),(A_{\star_j},A_i))$ tels que :
  1. (o_a,o_b) est caractérisé par un **lien** selon $A_{\star_j} : lien_{\star_j}(o_{a_{\star_j}},o_{b_{\star_j}}) = 1$,
  2. (o_a,o_b) est caractérisé par un **non-lien** selon $A_i : lien_i(o_{a_i},o_{b_i}) = 0$.
- $\widetilde{\widetilde{\mathbf{L}}}\widetilde{\mathbf{L}}(\mathbf{SA_\star},\mathbf{SA})$ qui correspond au nombre de couples $((o_a,o_b),(A_{\star_j},A_i))$ tels que :
  1. (o_a,o_b) est caractérisé par un **non-lien** selon $A_{\star_j} : lien_{\star_j}(o_{a_{\star_j}},o_{b_{\star_j}}) = 0$,
  2. (o_a,o_b) est caractérisé par un **lien** selon $A_i : lien_i(o_{a_i},o_{b_i}) = 1$.

$$\widetilde{LL}(SA_\star,SA) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{\substack{j=1..m \\ j \text{ tel que } A_{\star_j} \neq A_i}} lien_i(o_{a_i},o_{b_i}) \times lien_j(o_{a_j},o_{b_j}) \tag{2}$$

$$\overline{\widetilde{LL}}(SA_\star,SA) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{\substack{j=1..m \\ j \text{ tel que } A_{\star_j} \neq A_i}} (1-lien_i(o_{a_i},o_{b_i})) \times (1-lien_j(o_{a_j},o_{b_j})) \tag{3}$$

$$\widetilde{L\overline{L}}(SA_\star,SA) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{\substack{j=1..m \\ j \text{ tel que } A_{\star_j} \neq A_i}} lien_i(o_{a_i},o_{b_i}) \times (1-lien_j(o_{a_j},o_{b_j})) \tag{4}$$

$$\overline{\widetilde{L}\widetilde{L}}(SA_\star,SA) = \sum_{a=1..n} \sum_{b=a+1..n} \sum_{i=1..l} \sum_{\substack{j=1..m \\ j \text{ tel que } A_{\star_j} \neq A_i}} (1-lien_i(o_{a_i},o_{b_i})) \times lien_j(o_{a_j},o_{b_j}) \tag{5}$$

Nous avons montré dans (Jouve 2003) que le niveau d'adéquation entre $SA_\star$ et SA peut être caractérisé par les indices précédemment définis ($\widetilde{LL}$, $\overline{\widetilde{LL}}$, $\widetilde{L\overline{L}}$, $\overline{\widetilde{L}\widetilde{L}}$) et qu'une forte adéquation entre SA et $SA_\star$ est associée à de fortes valeurs pour $\widetilde{LL}$, $\overline{\widetilde{L}\widetilde{L}}$. Cependant, la signification de "fortes valeurs" n'est pas complètement intuitive, donc nous avons également déterminé dans (Jouve 2003) les lois statistiques (deux lois binomiales différentes) suivies par les indices $\widetilde{LL}$ et $\overline{\widetilde{LL}}$ sous l'hypothèse de non-adéquation. Cela nous a alors permis de dériver (grâce à une approximation normale suivie d'un centrage réduction) deux indices $Aq_1(SA_\star,SA)$ et $Aq_2(SA_\star,SA)$ qui caractérisent respectivement dans quelle mesure les valeurs de $\widetilde{LL}$ et $\overline{\widetilde{LL}}$ sont significativement fortes.

Sous l'hypothèse de non adéquation, ces deux indices suivent une loi normale centrée réduite ($N(0,1)$) (*moyenne = 0 ; écart type = 1*), ces indices sont définis comme suit :

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

$$Aq_2(SA_\star,SA) = \frac{\widetilde{\widetilde{LL}} - \frac{\widetilde{\widetilde{(LL+LL)}}\,\widetilde{\widetilde{(LL+LL)}}}{\widetilde{\widetilde{LL+LL+LL+LL}}}}{\sqrt{\frac{\widetilde{\widetilde{(LL+LL)}}\,\widetilde{\widetilde{(LL+LL)}}}{\widetilde{\widetilde{LL+LL+LL+LL}}} \times \left(1 - \frac{\widetilde{\widetilde{LL+LL}}}{\widetilde{\widetilde{LL+LL+LL+LL}}}\right)}}, \; Aq_2(SA_\star,SA) \hookrightarrow N(0,1)$$

Conséquemment, nous pouvons dire que l'adéquation entre $SA_\star$ et SA est forte si les valeurs pour $Aq_1(SA_\star,SA)$ et $Aq_2(SA_\star,SA)$ sont simultanément significativement élevées. Pour simplifier, plus les valeurs pour $Aq_1(SA_\star,SA)$ et $Aq_2(SA_\star,SA)$ sont simultanément élevées, plus cela signifie que l'adéquation entre $SA_\star$ et SA est forte.

Remarque Importante :
Nous avons montré dans (Jouve 2003) que les indices $Aq_1(SA_\star,SA)$ et $Aq_2(SA_\star,SA)$ possèdent une propriété spécifique très intéressante concernant leur calcul :

SI [$\frac{l(l-1)}{2}$ tables de contingences particulières croisant chaque variable de SA sont construites] *(ce qui ne requiert qu'une passe sur le jeu de données, et* $O(\frac{l(l-1)}{2}n)$ *(resp.* $O(\frac{l(l-1)}{2}n^2)$*) comparaisons si toutes les variables de SA sont catégorielles ou numériques discrétisées (resp. si des variables de SA sont numériques non discrétisées))*

ALORS : il est possible de calculer $Aq_1(SA_\star,SA)$ et $Aq_2(SA_\star,SA)$ pour tout sous-ensemble de SA sans accéder au jeu de données. Cela étant réalisé avec une complexité en $o(\frac{l(l-1)}{2})$ en accédant simplement aux $\frac{l(l-1)}{2}$ tables de contingences particulières.

3 Une Nouvelle Méthode de Sélection de Variables de Type Filtre pour le Clustering

La méthode que nous proposons est basée sur les indices $Aq_1(SA_\star,SA)$ et $Aq_2(SA_\star,SA)$. L'idée de base est de découvrir le sous-ensemble de SA le plus en adéquation avec SA (i.e. le sous-ensemble qui semble véhiculer au mieux l'information inclue dans SA). Pour ce faire, nous utilisons les indices $Aq_1(SA_\star,SA)$ et $Aq_2(SA_\star,SA)$ pour dériver une unique nouvelle mesure qui caractérise l'adéquation entre SA et $SA_\star$ ($SA_\star \subseteq SA$). Puis, l'objectif est de découvrir le sous-ensemble de SA qui optimise cette mesure.

La nouvelle mesure d'adéquation nommée $fit(SA,SA_\star)$, est basée sur le fait qu'une forte adéquation entre $SA_\star$ et SA est caractérisée par des valeurs simultanément fortes pour $Aq_1(SA_\star,SA)$ et $Aq_2(SA_\star,SA)$. Elle est définie comme suit :

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

Nous pouvons voir que, d'une certaine manière, cette fonction correspond à une distance entre deux sous-ensembles de variables du point de vue de l'adéquation avec l'ensemble des variables SA. Plus précisément, on peut voir cette mesure comme la distance du point de vue de l'adéquation avec l'ensemble des variables SA entre : un sous-ensemble virtuel de variables (pour lequel les valeurs pour Aq_1 et Aq_2 seraient respectivement $a\tilde{q}_1$ et $a\tilde{q}_2$) et le sous-ensemble de variables $SA_\star$.

En fait, nous fixons $a\tilde{q}_1 = a\tilde{q}_2 = fortes\ valeurs$ de manière à conférer au sous-ensemble virtuel de variables l'aspect d'un ensemble de variables idéal du point de vue de l'adéquation avec SA. Ainsi, plus la valeur pour cette mesure est faible (en quelque sorte, plus la distance est faible du point de vue de l'adéquation avec SA entre le sous-ensemble virtuel de variable et le sous-ensemble de variables $SA_\star$), plus l'adéquation entre SA et $SA_\star$ peut être considérée comme forte.

La méthode filtre de sélection de variables que nous proposons est basée sur l'utilisation de cette mesure : elle consiste en la recherche du sous-ensemble de SA qui minimise la fonction $fit(SA,SA_\star)$.

La recherche pourrait être exhaustive mais cela impliquerait un coût calculatoire bien trop important, afin de limiter ce coût nous avons utilisé un algorithme génétique (AG) de manière à ne réaliser qu'une exploration partielle de l'espace composé des sous-ensembles de SA [1]. L'AG employé est défini de la manière suivante :
(1) un chromosome correspond (code) un sous-ensemble de SA ;
(2) chaque gène du chromosome correspond à une variable de SA (donc, il y a l gènes) ;
(3) chaque gène d'un chromosome à une valeur binaire : le gène vaut 1 si la variable qui lui est associée est présente dans le sous-ensemble de SA codé par le chromosome auquel il appartient ; le gène vaut 0 si la variable qui lui est associée n'est pas présente dans le sous-ensemble de SA codé par le chromosome auquel il appartient.

L'algorithme de la méthode de SV est donné ci dessous, notons que :
(1) il ne requiert qu'une seule passe sur le jeu de données ;
(2) il nécessite le stockage de plusieurs tables de contingences mais que cela ne correspond qu'à un faible coût en terme de mémoire ;
(3) sa complexité est faible (quadratique selon le nombre de variables du jeu de données et complètement indépendante du nombre d'objets une fois que les tables de contingences nécessaires ont été bâties) ;
(4) il peut traiter indifféremment des données catégorielles, numériques ou mixtes. (notons que le coût calculatoire nécessaire à la création des tables de contingence peut cependant paraître excessif dans le cas de variables numériques (complexité quadratique selon le nombre d'objets) et qu'il est préférable de travailler sur des données catégorielles ou numériques discrétisées (la complexité de la création des tables de contingences étant alors linéaire selon le nombre d'objets).

Algorithme, Méthode de SV "filtre" pour le Clustering :
1. En une passe sur les données, bâtir les $\frac{l(l-1)}{2}$ tables de contingence nécessaires au calcul des indices d'adéquation présentés.
2. Utiliser l'AG avec pour fonction objectif à minimiser la fonction $fit(SA,SA_\star)$.
3. Sélectionner le meilleur sous-espace découvert par l'AG.

1. Notons que nous aurions pu opter pour d'autres méthodes d'optimisation et qu'il ne s'agit ici que d'un choix arbitraire discutable. En effet, l'emploi d'autres approches gloutonnes permettrait de limiter plus encore le coût calculatoire. Toutefois le choix définitif de la méthode d'optimisation à employer est ici hors de notre propos.

4 Evaluations Expérimentales

Afin d'évaluer cette méthode 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 (Merz et al. 1996).

4.1 Evaluation expérimentale sur jeux de données synthétiques

Description

L'objectif est de tester dans quelle mesure notre méthode détecte les variables pertinentes. Pour cela nous avons bâti des jeux de données synthétiques comprenant 1000 objets caractérisés par 9 variables (A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9) véritablement vecteur d'information et par un ensemble de $(l-9)$ variables correspondant à du bruit. Plus précisément : les objets o_1 à o_{250} (resp. o_{251} à o_{500}) (resp. o_{501} à o_{750}) (resp. o_{751} à o_{1000}) possèdent tous la même valeur D pour les variables A_1, A_2, A_3 (resp. A_3, A_4, A_5) (resp. A_5, A_6, A_7) (resp. A_7, A_8, A_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 sur la figure 1, la composition des jeux de données. On visualise ainsi 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.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	...	A_l	...	A_l
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. 1 – *Jeu de données synthétique*

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 A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9 ainsi que par respectivement 9 (resp. 18)(resp. 27)(resp. 36) (resp. 81) (resp. 171) variables "bruit".

Soient des jeux de données composés respectivement de 18 variables dont 50% sont sources d'informations (resp. 27 variables dont $\frac{1}{3}$ sont sources d'informations) (resp. 36 variables dont 25% sont sources d'informations) (resp. 45 variables dont 20% sont

sources d'informations) (resp. 90 variables dont 10% sont sources d'informations) (resp. 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 SV, chacune des séries étant caractérisée par le nombre de générations de l'AG utilisé. Ainsi pour la première série, le nombre de générations valait *50*, ce nombre valait respectivement *100*, *500*, *1000* et *2500* pour les deuxième, troisième, quatrième et cinquième séries. Les autres paramètres de l'AG étant : *nombre de chromosomes par génération=30; proba. de croisement=0.98 ; proba. de mutation=0,4 ; élitisme=oui.*

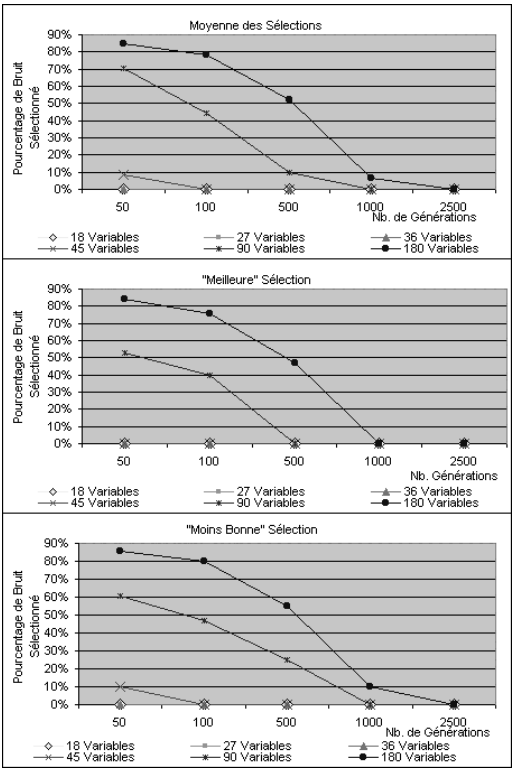
Analyse des Résultats

Les résultats sont présentés sur la figure page suivante, ils nécessitent toutefois des explications... Notons tout d'abord que chacun des $6 \times 5 \times 5 = 150$ processus de SV réalisés a mené à l'obtention d'un sous-espace de variables comprenant les 9 variables pertinentes (A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_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 SV. Elles détaillent pour chaque série : la moyenne du pourcentage de variables "bruit" sélectionnées par les 5 processus de SV de la série ; le pourcentage le plus faible de variables "bruit" sélectionnées (i.e. le pourcentage de variables "bruit" sélectionnées par le processus de SV 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 SV 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).

Concernant le pourcentage de variables non pertinentes (variables "bruit") 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, pour les jeux de données comportant 10% ou moins de 10% de variables pertinentes, la sélection de l'ensemble optimal de variables ($SA_\star = \{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9\}$) est obtenue pour des nombres de générations supérieurs ou égaux à 1000.

La méthode apparaît donc comme excellente, car, les indices ainsi que la fonction objectif 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 sans impliquer pas un temps de calcul démesuré (voir (Jouve 2003) pour des informations sur le temps de calcul). A titre indicatif, pour le jeu de données



Expérimentations : jeux de données synthétiques

comportant 180 variables, le nombre de sous ensembles non vides de l'espace de représentation des données 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 clustering 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.

4.2 Evaluation Expérimentale sur Jeux de Données de l'UCI

Description

L'objectif de ces expérimentations est de déterminer si les clusterings obtenus en considérant un sous-ensemble de variables (un sous-ensemble de SA) sélectionné par notre méthode de SV possèdent un niveau de qualité équivalent ou meilleur que les clusterings obtenus en considérant l'ensemble de variables (SA) dans son intégralité.

Pour cela, nous avons utilisé deux jeux de données classiques provenant de la collection de l'UCI (Merz et al. 1998) : les jeux Mushrooms et Small Soybean Diseases.

Plus précisément, nous avons appliqué notre méthode de SV sur ces deux jeux de données, il en a résulté la sélection des sous-ensembles de variables suivants :

- 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.

Puis, nous avons exécuté des processus de clusterings sur le jeu de données Small Soybean Diseases (resp. Mushrooms) en considérant soit l'intégralité des 35 (resp. 22) variables ou en ne considérant que les 9 (resp. 15) variables sélectionnées par notre méthode de SV. Nous présentons ici les résultats obtenus avec la méthode de clustering pour données catégorielles K-Modes (Huang 1997) (qui correspond à une adaptation de la méthode K-Means dans le cadre de données catégorielles). Différents "paramètrages" (différents nombres de classes) ont été utilisés de manière à générer des clusterings en divers nombre de classes [2].

Pour résumer, pour le jeu de données Mushrooms (resp. Small Soybean Diseases), nous avons réalisé des clusterings (en utilisant la méthode K-Modes) en 2, 3, 4, ..., 24, 25 (resp. 2, 3, ..., 9, 10) classes soit en considérant l'intégralité des variables ou en considérant le sous-ensemble sélectionné.

Afin d'évaluer la qualité des clusterings obtenus, nous avons utilisé une mesure de validité interne. Cette mesure étant en fait le critère QKM, critère devant être minimisé par la méthode K-Modes. Évidemment, nous avons calculé la valeur de ce critère en prenant en compte l'intégralité des variables du jeu de données considéré, et ce, même si le clustering était obtenu en ne considérant qu'un sous-ensemble de variables. (Ainsi, nous pouvons dire que nous avons évalué chaque clustering par le biais du critère QKM calculé en tenant compte de l'intégralité des variables du jeu de données considéré.)

Analyse des Résultats

Jeu de Données Mushrooms On peut observer (figure 2) que la qualité des clusterings (selon le critère QKM) obtenus soit en considérant l'intégralité des variables soit le sous-ensemble de variables sélectionnées est quasiment similaire. Cela montre que les clusterings obtenus en incluant l'étape de pré-traitement de SV sont aussi bons (au sens du critère QKM) que ceux obtenus sans cette étape. Cela montre la pertinence des variables du sous-ensemble sélectionné et donc l'efficacité de notre méthode.

Jeu de Données Small Soybean Diseases Les résultats (figure 3) sont similaires à ceux obtenus pour le jeu de Données Mushrooms, cela montre également que le sous ensemble de variables sélectionnées est bon et que notre méthode est véritablement efficace. Notons également que le nombre de variables sélectionnées est ici relativement faible ($\simeq 25\%$ du nombre initial de variables).

Remarque : Des expérimentations plus complètes (impliquant différents critères et méthodologies pour évaluer la qualité des clusterings -telle que l'utilisation d'une mesure de validité externe- et impliquant différentes méthodes de clusterings telle que la méthode Kerouac (Jouve et al. 2003)) sont présentées dans (Jouve 2003). Ces expérimentations supplémentaires confirment également l'efficacité de notre méthode de SV. De plus amples expérimentations, concernant notamment la composition des classes des clusterings obtenus, montrent que, pour un nombre donné de classes, les

2. Pour chaque "paramètrage" (chaque nombre de classes), nous avons exécuté 10 processus de clustering différents et conservé finalement le clustering possédant la meilleure valeur pour le critère QKM (critère devant être optimisé au sein de la méthode K-Modes). Ceci étant réalisé de manière à minimiser l'effet d'initialisation de la méthode K-Modes.

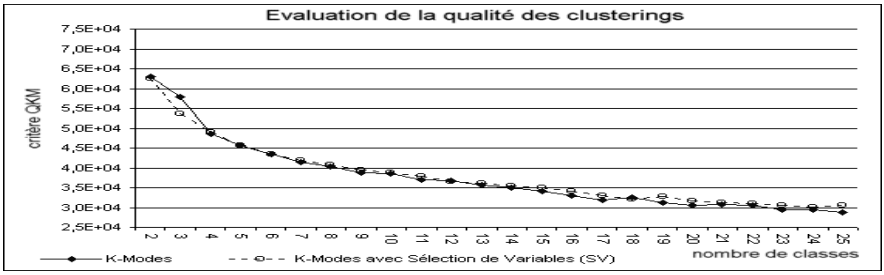


FIG. 2 – *Expérimentation sur le jeu de données Mushrooms*

clusterings obtenus en tenant compte de l'ensemble des variables et ceux obtenus en ne tenant compte que du sous ensemble sélectionné possèdent des compositions similaires (i.e. les compositions des classes -en terme d'objets inclus- sont très proches).

5 Discussion Finale et Conclusion

Nous proposons donc une nouvelle méthode "filtre" pour la SV dans le cadre de l'apprentissage non supervisé. En complément des avantages classiques des méthodes "filtre" (indépendance vis à vis des algorithmes d'apprentissage non supervisé et non assujettissement au problème de l'absence de critère consensuel pour l'estimation de la qualité d'un clustering), cette méthode possède plusieurs caractéristiques intéressantes : **(1)** elle ne requiert qu'une unique passe sur les données contrairement aux autres approches filtres (telle (Dash et al. 2002)), ce qui lui confère un temps d'exécution relativement faible ; **(2)** elle nécessite le stockage de tables de contingence mais cela correspond à un très faible coût en terme de mémoire utilisée ; **(3)** sa complexité algorithmique est faible (quadratique selon le nombre de variables du jeu de données et complètement indépendante du nombre d'objets une fois que les tables de contingences nécessaire ont été bâties) ; **(4)** elle peut indifféremment traiter des données catégorielles ou numériques contrairement par exemple à l'approche filtre proposée dans (Dash et al. 2002) ; **(5)** de manière similaire à la méthode proposée dans (Dash et al. 2002) cette méthode permet de sélectionner des variables du jeu de données et non une sélection de nouvelles variables comme le font les approches basées sur l'analyse factorielle ou le multi-dimensionnal scaling. Ce point est particulièrement important si l'on désire bâtir un modèle aisément interprétable.

Les expérimentations ont montré que : **(1)** les clusterings obtenus avec une étape de SV réalisé par notre méthode sont de bonne qualité ; **(2)** le nombre de variables sélectionnées peut parfois être très faible ; **(3)** les données fortement bruitées peuvent être traitées par notre méthode ; **(4)** le temps d'exécution est faible.

Il existe différentes améliorations possibles, telles que : **(1)** amélioration du point de vue du coût calculatoire en substituant à l'AG une méthode d'optimisation gloutonne ; **(2)** modification de la structure de l'AG afin de ne pas rechercher le sous ensemble "optimal" mais le "meilleur" sous ensemble incluant un nombre fixé de variables...

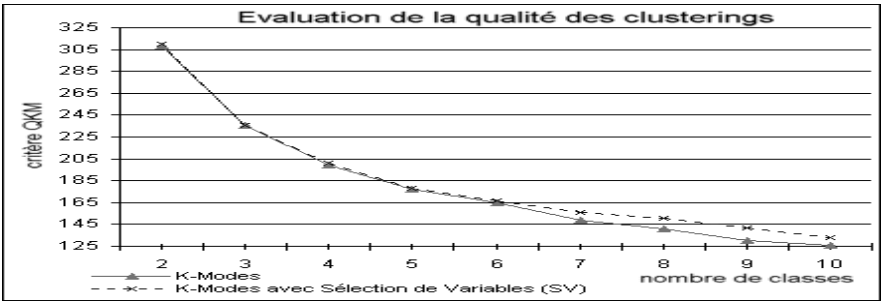


FIG. 3 – *Experimentation sur le jeu de données Small Soybean Diseases*

6 Références

Dash M., Liu H.(1997) : Feature selection for classification. International Journal of Intelligent Data Analysis, 1(3)

Dash M., Liu H.(2000) : Feature selection for clustering. In Proc. of Fourth PacificAsia Conference on Knowledge Discovery and Data Mining, (PAKDD).

Dash M. , Choi K., Scheuermann P., Liu H.(2002) : Feature Selection for Clustering - A Filter Solution. In Proc. of International Conference on Data Mining (ICDM02), 115–122

Devaney M., Ram A.(1997) : Efficient feature selection in conceptual clustering. In Proc. of the International Conference on Machine Learning (ICML), 92–97

Dy J. G., Brodley C. E.(2000) :Visualization and interactive feature selection for unsupervised data. In Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD), 360–364

Huang Z.(1997) : A Fast Clustering Algorithm to Cluster Very Large Categorical Data Ensembles in Data Mining. In Research Issues on Data Mining and Knowledge Discovery.

Jouve P. E.(2003) : Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données. Thèse de Doctorat, Lab. ERIC, Université Lyon II.

Jouve P.E., Nicoloyannis N.(2003) : KEROUAC, an Algorithm for Clustering Categorical Data Ensembles with Practical Advantages. In. Proc. of International Workshop on Data Mining for Actionable Knowledge.

Kim Y. S., Street W. N., Menczer F.(2000) : Feature selection in unsupervised learning via evolutionary search. In Proc. of ACM SIGKDD International Conference on Knowledge and Discovery, 365–369

Merz, C., Murphy, P.(1996) : UCI repository of machine learning databases. http://www.ics.uci.edu/#mlearn/mlrepository.html.

Talavera L.(2000) : Feature selection and incremental learning of probabilistic concept hierarchies. In Proc. of Int. Conference on Machine Learning (ICML).