HAL open science

Extraction de frontière de Markov dans les données de grande dimension

Grégory Thibault, Alexandre Aussem, Stéphane Bonnevay

To cite this version:

Grégory Thibault, Alexandre Aussem, Stéphane Bonnevay. Extraction de frontière de Markov dans les données de grande dimension. 5èmes Journées Francophones sur les Réseaux Bayésiens (JFRB2010), May 2010, Nantes, France. hal-00467544

HAL Id: hal-00467544

https://hal.science/hal-00467544

Submitted on 19 Apr 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Extraction de frontière de Markov dans les données de grande dimension

Une méthode d'apprentissage incrémentale pour la sélection de variables

Grégory Thibault* — **Alexandre Aussem*** — **Stéphane Bonnevay****

* *Laboratoire LIESP, Université de Lyon*
43 Boulevard du 11 Novembre 1918
F-69622 Villeurbanne Cedex

gregory.thibault@bat710.univ-lyon1.fr
aaussem@univ-lyon1.fr

** *Laboratoire ERIC, Université de Lyon*
5 avenue Pierre Mendès-France
F-69676 Bron Cedex

bonnevay@univ-lyon1.fr

RÉSUMÉ. *Notre but est de résoudre le problème de sélection de sous-ensembles de variables pour des données constituées de milliers de variables. Notre solution est une méta-procédure qui utilise une méthode d'apprentissage de structure de réseau bayésien en combinant incrémentalement ses résultats. Nous montrons l'efficacité d'une telle démarche en utilisant une méthode à base de score correcte sur de petits ensembles de données mais ne passant pas à l'échelle, puis en mesurant la capacité de passage à l'échelle, les performances et la stabilité de l'algorithme. Les expériences utilisent des données synthétiques et réelles, allant jusqu'à* 139351 *variables.*

ABSTRACT. *Our aim is to solve the feature subset selection problem with thousands of variables using an incremental procedure. The procedure combines incrementally the outputs of non-scalable search-and-score Bayesian network structure learning methods that are run on much smaller sets of variables. We assess the scalability, the performance and the stability of the procedure through several experiments on synthetic and real databases scaling up to 139 351 variables. Our method is shown to be efficient in terms of both running time and accuracy.*

MOTS-CLÉS : *réseaux Bayésiens, sélection de variables, frontière de Markov, passage à l'échelle*

KEYWORDS: *Bayesian networks, features selection, Markov boundary, scalability*

1. Introduction

La sélection de sous-ensembles de variables (en Anglais *feature subset selection* ou FSS) est un composant essentiel de la modélisation, de la construction de modèles d'aide à la décision ou encore de la découverte assistée par ordinateur. La FSS permet de diminuer la complexité des modèles tout en conservant une efficacité au moins aussi bonne, puisque seules les variables informatives et non-redondantes en information seront conservées. En outre, la FSS permet également à un expert d'avoir une vue des variables pertinentes de son problème.

Récemment, la taille des bases de données a augmenté considérablement, tant en nombre d'instances qu'en nombre d'attributs considérés. Adapter ou créer de nouvelles méthodes de sélection efficaces à ces nouvelles échelles est un nouveau défi. Ce défi présente de bonnes opportunités et challenges pour la recherche en extraction de connaissances et apprentissage machine.

Il existe un grand nombre d'études comparatives pour la sélection de variables, mais peu concernent les méthodes passant à l'échelle jusqu'à 100000 variables. Les forêts aléatoires (*random forest* ou RF) et les algorithmes d'élimination récursive de variables pour les machines à vecteur de support (Nilsson *et al.*, 2006) (*support vector machine - recursive feature elimination* ou SVM-RFE) sont très largement utilisés, mais souffrent de nombreux problèmes. Par exemple, dans l'analyse de données des micro-puces à ADN, il est commun d'utiliser des tests statistiques pour contrôler la précision en maximisant le rappel afin d'obtenir des ensembles de gènes de bonne qualité. (Nilsson *et al.*, 2006) a montré qu'aucune méthode basée sur les SVM ne permettait ce type de contrôle. De plus, la robustesse du processus de sélection de variables est aussi une donnée importante qui doit être prise en compte (Saeys *et al.*, 2008). Or, (Tang *et al.*, 2007) a montré expérimentalement que SVM-RFE est sensible au problème dit de filtrage sortant (*"filter-out factor"*), et par conséquent n'est pas un algorithme stable.

Le but des méthodes de sélection de variables est d'éliminer les variables qui n'apportent aucune information supplémentaire à l'ensemble de variables restantes, mais les méthodes basées sur les RF sélectionnent les variables dès lors qu'elles ont un pouvoir prédictif, sans prendre en considération la redondance de cette information. Enfin, (Ma *et al.*, 2008; Hua *et al.*, 2009) ont montré à travers un nombre important de comparaisons sur des données génétiques de grande dimension qu'aucun des algorithmes de sélection de variables considéré n'était meilleur dans tous les scénarios. Tous ces travaux montrent qu'il reste du travail dans ce domaine récent.

Une solution idéale à ce problème peut être trouvée grâce aux réseaux bayésiens. Dans un réseau bayésien qui représente fidèlement une loi de probabilité, il est facile de trouver l'ensemble minimal de variables, appelé frontière de Markov, qui est la solution optimale au problème de sélection. Mais en pratique, il n'est pas possible de construire le réseau qui représente fidèlement une loi de probabilité à très forte dimension. Récemment, de puissantes méthodes d'extraction de frontière de Markov, correctes et qui ne nécessitent pas la construction du réseau complet (donc passant à

l'échelle) ont été crées (Peña *et al.*, 2007; Rodrigues de Morais *et al.*, 2008; Tsamardinos *et al.*, 2006; Yaramakala *et al.*, 2005). Ces méthodes dites sous-contraintes (ou CB pour *constraint-based* en anglais) sont basées sur la recherche d'indépendances conditionnelles dans les données ce qui permet d'extraire la MB d'une cible sans avoir à connaître toute la structure des données.

Dans notre travail, nous proposons une méthode permettant d'utiliser n'importe quel algorithme d'apprentissage de structure de réseau bayésien, y compris des méthodes à base de score, très consommatrices en temps afin d'extraire la MB. Nous montrons que la procédure est capable de résoudre le problème de sélection parmi des milliers de variables en temps quasi-linéaire tout en conservant une grande efficacité et une grande robustesse.

2. Sélection de variables

Les techniques de sélection sont divisibles en trois catégories, selon la manière dont elles interagissent avec le classificateur. Les méthodes filtres (*filter methods*) opèrent directement sur le jeu de données et fournissent une pondération, un classement ou un ensemble de variables en sortie. Ces méthodes ont l'avantage d'être rapides et indépendantes du modèle de classification, mais au prix de résultats inférieurs. Les méthodes *wrapper* effectuent une recherche dans l'espace de sous-ensembles de variables, guidée par le résultat du modèle — par exemple les performances en validation croisée sur les données d'apprentissage. Elles font souvent de meilleurs résultats que les méthodes de filtrage, mais au prix d'un temps de calcul plus important. Enfin, les méthodes embarquées (*embedded*) utilisent l'information interne du modèle de classification (par exemple, le vecteur de poids dans les SVM (*support vector machines*). Elles offrent souvent un bon compromis entre performance et temps de calcul.

Trouver l'ensemble minimal de variables requiert une recherche exhaustive parmi tous les sous-ensembles de variables pertinents, ce qui est un problème NP-complet, et peut avoir une multitude de solutions. Les méthodes de découverte de la frontière de Markov peuvent être considérés comme étant des méthodes hybrides, entre les méthodes filtres et les méthodes embarquées. Elles permettent de résoudre la sélection et, dans un même temps, construire un réseau bayésien local autour de la variable cible qui peut être utilisé ensuite comme classifieur probabiliste.

3. Couverture et frontière de Markov dans les réseaux bayésiens

Une couverture de Markov (*Markov blanket*) $\mathbf{M}_T$ de T est un ensemble quelconque de variables qui rend T indépendant de toutes les variables restantes sachant $\mathbf{M}_T$. Une frontière de Markov $\mathbf{MB}_T$ de T est une couverture de Markov minimale telle qu'aucun sous-ensemble n'est aussi une couverture de Markov de T.

Soient $\mathbf{PC}_T$ l'ensemble de parents et d'enfants de T dans $\mathcal{G}$ — c'est-à-dire les variables directement liées à T — et $\mathbf{SP}_T$ l'ensemble d'épouses de T, alors on peut énoncer le résultat suivant[1] :

Théorème 1 *Supposons que $<\mathcal{G}, \mathcal{P}>$ soit un réseau bayésien qui satisfasse la condition de fidélité. Alors, pour tout X, les ensembles $\boldsymbol{PC}_X$ et $\boldsymbol{SP}_X$ sont uniques. L'ensemble $\boldsymbol{SP}_X \cup \boldsymbol{SP}_X$ est l'unique frontière de Markov de X.*

4. Une méthode incrémentale pour la sélection à grande échelle

Algorithme 1 Procédure générique et incrémentale de FSS par recherche de MB

1: **fonction** IFSS($\mathcal{D}, cible, taillesel, Vars$)
2: $MB \leftarrow \emptyset$
3: **répéter**
4: $ensembleTest \leftarrow \{cible\} \cup MB$
5: $ensembleTest \leftarrow ensembleTest \cup$ SELECTION($taillesel$)
6: $\mathcal{G} \leftarrow$ APPRENTISSAGE_RB($\mathcal{D}, ensembleTest$)
7: $MB \leftarrow$ EXTRAIRE_MB($\mathcal{G}$) ▷ extraction de MB
8: **jusqu'à** critère_stop
9: **retourne** MB ▷ sortie = variables de la MB
10: **fin fonction**

L'idée maîtresse de ce papier est qu'une procédure incrémentale pourrait contribuer à atténuer l'obstacle lié à la complexité du problème en agrégeant les sorties de plusieurs algorithmes de sélection de variables fonctionnant sur un nombre plus faible de variables. Plus précisément, un même modèle de sélection est appliqué sur de petits sous-ensembles de variables de façon incrémentale. Notre méthode, nommée IFSS pour *Incremental Feature Subset Selection*, est basée sur l'apprentissage de structure de réseau bayésien. Les arguments en entrée sont :

- $\mathcal{D}$: données utilisées pour l'apprentissage supervisé,
- $cible$: la variable cible,
- $taillesel$: nombre de nouvelles variables à chaque itération,
- $Vars$: ensemble des variables, hormis la $cible$.

$apprentissage_RB$ peut être implémenté par n'importe quel algorithme d'apprentissage de structure. Dans cette étude, il est implémenté avec l'algorithme de recherche gloutonne GES (méthode à base de score (Chickering, 2002)).

Au départ, l'ensemble des variables $ensembleTest$ est initialisé aléatoirement. Une première MB est extraite à partir de $\mathcal{G}$. À chaque itération les variables de MB sont conservées dans $ensembleTest$ et quelques autres variables sont ajoutées par une sélection aléatoire uniforme et sans remplacement. La taille de cette sélection

1. on trouvera notamment une preuve dans (Neapolitan, 2004)

$taillesel$ est adaptée en fonction de la taille supposée de la frontière de Markov MB. Dans cette étude, nous choisissons le critère d'arrêt de telle sorte que chaque variable soit introduite dans $ensembleTest$ deux fois. Lors du premier passage, chaque variable de $Vars$ est sélectionnée une fois dans l'ensemble de test puis le processus est réitéré avec l'ensemble des variables $Vars$.

Si les données sont représentables fidèlement et si $apprentissage_RB$ est un algorithme correct sur des sous-ensembles de variables, à la fin du premier passage, l'ensemble MB — l'ensemble candidat — contient PC_{cible} car par définition, aucune autre variable n'est en mesure de d-séparer cet ensemble (ou une sous-partie) de la variable $cible$. De même, lors du second passage, c'est l'ensemble SP_{cible} qui viendra s'ajouter. Sous cette condition de fidélité, IFSS renvoie la frontière de Markov correcte. Dans la pratique, notre espoir est de découvrir les variables qu'aurait trouvé l'algorithme d'induction $apprentissage_RB$ sur la base de données complète, s'il en avait eu le temps.

5. Expérimentations

Dans cette section, nous évaluons la précision, la capacité de passage à l'échelle et la robustesse d'IFSS à travers plusieurs expériences empiriques sur les jeux de données de référence. Nous utilisons un algorithme d'apprentissage de structure de RB de la littérature (GES) paramétré pour maximiser le score bayésien de Dirichlet[2] :

$$BD(\mathcal{B} \mid \mathcal{D}) = p(\mathcal{B}) \cdot \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})}$$

Notons qu'aucune information *a priori* sur la structure n'est fourni pour les tests sur les données synthétiques, c'est-à-dire $p(\mathcal{B})$ est uniforme. De plus, pour toutes les expériences, l'*a priori* sur les paramètres est non-informatif, c'est-à-dire que le paramètre α est un *a priori* uniforme de Dirichlet dont le paramètre de taille d'échantillonnage pour l'équivalence (ESS) a pour valeur la plus grande modalité d'attribut (voir (Neapolitan, 2004) pour plus de détails).

La validation de l'algorithme est effectuée sur des critères classiques de performance.

– taux de vrais positifs [3](TPR), ou recall : le ratio *vrais positifs / positifs*,

– taux de faux positifs [4](FPR), ou $1-$spécificité : le ratio *faux positifs / négatifs*,

– indice Kappa (κ),

– précision (Acc), à savoir le taux de bonnes prédictions,

– précision équilibrée (WAcc), la moyenne des précisions sur toutes les classes,

2. nous désignerons l'algorithme ainsi implémenté par $IFSS_{GES}$
3. un positif est une valeur dans la classe à prédire : il est vrai s'il est correctement classé
4. un négatif est une valeur en dehors de la classe : il est vrai s'il est correctement rejeté

– le temps en secondes.

Kappa est une mesure qui évalue l'amélioration apportée par rapport au hasard. Les plages suivantes sont suggérées par la littérature pour la statistique Kappa : faible $K < 0.4$, bon $0.4 \leq K \leq 0.75$ et excellent $K > 0.75$.

5.1. *Précision sur des petits réseaux*

Benchmark	# var	# edges	target	MB size	# samples
ASIA	8	8	OR	5	10 000
ASIA8	64	64	OR	5	10 000
ALARM	37	46	HR	8	30 000
INSULIN	35	52	IPA	18	50 000
INSURANCE	27	52	Accident	10	30 000
HAILFINDER	56	66	Scenario	17	50 000

Tableau 1. *Description des* benchmarks *synthétiques utilisés dans nos expériences pour comparer la précision d'extraction de MB par* $IFSS_{GES}$ *et* GES.

Nous rapportons ici les résultats de nos expériences sur six *benchmarks* décrits dans le tableau 1 : ASIA, ASIA8 (ASIA dupliqué 8 fois), ALARM, INSULIN, INSURANCE et HAILFINDER (voir (Tsamardinos *et al.*, 2006) et références citées). Pour chaque réseau, 10 bases de données indépendantes et identiquement distribuées ont été générées par la méthode *logic sampling*. La quantité de données a été choisi suffisamment élevée pour éviter les biais dus à un manque de données. La tâche est d'apprendre la MB de la variable qui apparaît dans la troisième colonne du tableau 1.

Le tableau 2 résume les indices comparés de performance moyenne obtenus par $IFSS_{GES}$ et GES sur 10 exécutions pour chaque réseau. $IFSS_{GES}$ obtient des résultats semblables à ceux de GES, sauf sur INSURANCE où $IFSS_{GES}$ dépasse largement GES. C'est une grande surprise puisque $IFSS_{GES}$ n'a pas été conçu pour surpasser la méthode d'apprentissage utilisée dans la méta-procédure mais seulement pour passer à l'échelle.

	GES					IFSS				
	κ	TPR	FPR	WAcc	Time	κ	TPR	FPR	WAcc	Time
ASIA	0.959	1.000	0.050	0.975	0.10	0.959	1.000	0.050	0.975	0.06
ASIA8	0.867	1.000	0.024	0.988	27.87	0.834	1.000	0.031	0.984	1.29
ALARM	0.916	0.875	0.000	0.938	6.79	0.916	0.875	0.000	0.938	2.56
INSULIN	0.840	0.933	0.094	0.920	15.16	0.870	0.933	0.063	0.935	7.83
INSURANCE	0.663	0.700	0.063	0.819	5.81	0.858	0.860	0.019	0.921	2.60
HAILFINDER	0.589	0.571	0.037	0.767	48.71	0.517	0.471	0.016	0.727	6.19

Tableau 2. *Performance moyenne comparée d'*$IFSS_{GES}$ *et de* GES

5.2. *Performance comparée aux algorithmes de la littérature*

Dans cette section, nous comparons les performances d'$IFSS_{GES}$ contre MBOR (Rodrigues de Morais *et al.*, 2010), PCMB (Peña *et al.*, 2007), MMMB (Tsamardinos

et al., 2006) et Inter-IAMB (IIAMB) (Yaramakala *et al.*, 2005). PCMB et IIAMB sont deux des plus précis algorithmes basés sur les contraintes de la littérature. PCMB utilise une stratégie diviser-pour-régner, alors que Inter-IAMB est incrémental. Pour PCMB, nous utilisons le code source donné par (Peña *et al.*, 2007). Pour MMMB et IIAMB, nous utilisons le logiciel proposé dans (Tsamardinos *et al.*, 2006). Pour MBOR, nous utilisons le code MATLAB® de l'auteur (Rodrigues de Morais *et al.*, 2010).

Benchmark	# var	# edges	target	MB size
INSULIN	35	52	1	18
LINK	724	1125	424	31
GENE	801	972	40	13
PIGS	441	592	435	68

Tableau 3. *Description des réseaux utilisés pour comparer $IFSS_{GES}$ avec les autres algorithmes de la littérature.*

Les bases de données sont générées à partir des réseaux déjà utilisés comme *benchmarks* pour les algorithmes ci-dessus[5], à savoir GENE, INSULIN, LINK et PIGS, décrits dans le tableau 3. Mis à part INSULIN, ces réseaux sont plus gros que ceux utilisés dans les expériences précédentes, GES ne s'adaptant pas à plus de 400 variables avec notre code. Pour chaque réseau, nous avons échantillonné des bases de données contenant 100, 500, 1000 et 5000 instances. Pour évaluer la performance, nous combinons la valeur de précision positive et le rappel grâce à la distance euclidienne d'avec la prédiction parfaite, tel que proposé dans (Peña *et al.*, 2007) : $\sqrt{(1-precision)^2+(1-recall)^2}$. Les résultats sont montrés figure 1.

$IFSS_{GES}$ semble compétitif par rapport aux algorithmes de la littérature. Toutefois, en inspectant les résultats, il est difficile de conclure. Par conséquent, nous effectuons un test non-paramétrique de Friedman sur les résultats de ces cinq algorithmes, en prenant comme mesure de performance la distance médiane euclidienne entre le résultat et la prédiction parfaite au cours des 16 expériences (4 benchmarks et 4 formats de données), tel que recommandé dans (Demsar, 2006). Le test indique que les algorithmes peuvent être distingués (cf. (Thibault *et al.*, 2009) pour plus de détails sur le calcul). Ensuite, nous procédons à un test de Nemenyi (voir encore (Demsar, 2006)) pour comparer les algorithmes entre eux. Les performances de MBOR et $IFSS_{GES}$ sont significativement meilleures que Inter-IAMB.

5.3. *Passage à l'échelle*

Dans cette section, nous utilisons la base de données THROMBIN qui a été fournie par DuPont Pharmaceuticals pour la KDD-Cup 2001. C'est un exemple réel de conception de médicaments (Cheng *et al.*, 2002). L'ensemble d'apprentissage contient

5. Tous les réseaux et les données sont disponibles dans le *bayesian networks repository* http ://www.cs.huji.ac.il/labs/compbio/Repository

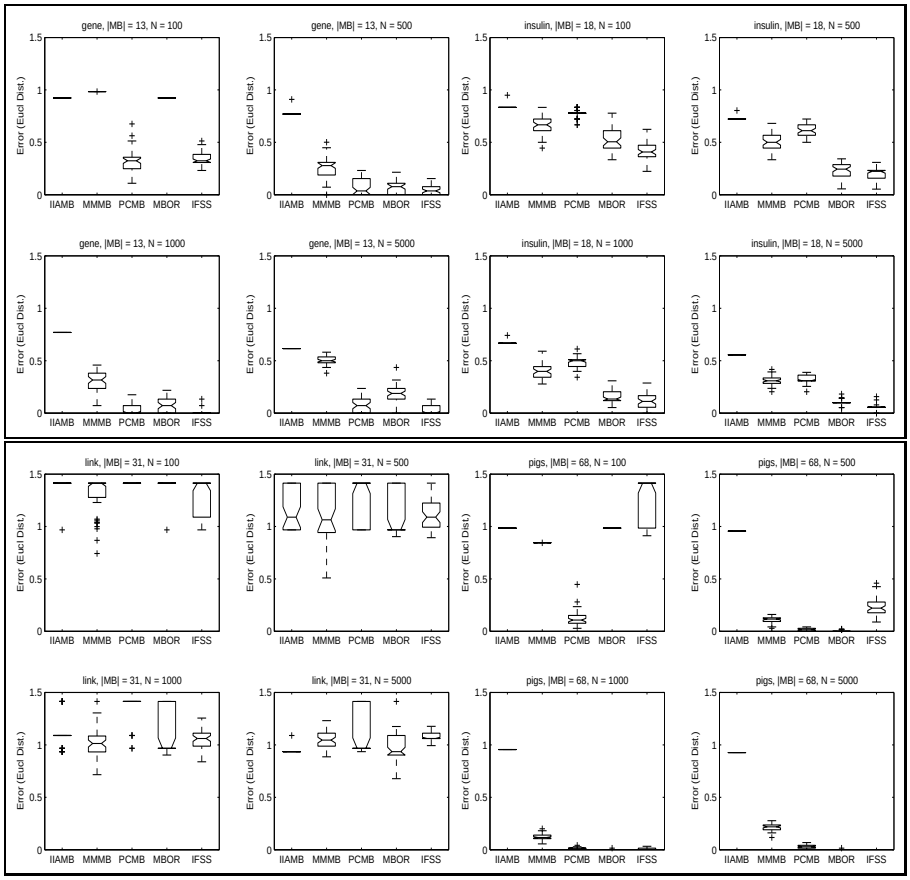

Figure 1. *Distributions des distances euclidiennes sur 100 bases. Résultats pour 100, 500, 1000 et 5000 instances en forme de boîtes à moustaches pour* $IFSS_{GES}$*, MBOR, PCMB, MMMB et Inter-IAMB sur les réseaux GENE, PIGS, INSULIN et LINK.*

1909 instances caractérisées par 139351 variables binaires. Les variables décrivent les propriétés tri-dimensionnelles des composés. Chaque composé est étiqueté avec une des deux classes, selon qu'il se fixe au site cible ou non. La tâche de KDD-Cup 2001 était d'apprendre un classifieur afin de prédire l'affinité de fixation et, par conséquent, le potentiel d'un composé comme agent anti-coagulation. Les classifieurs soumis à KDD-Cup 2001 ont été évalués sur les 634 composés (données de test) par la moyenne équilibrée (WAcc) de classification. La base THROMBIN est difficile pour trois raisons. Premièrement, elle a un très grand nombre de variables. Deuxièmement, les données d'apprentissage sont extrêmement déséquilibrées : seuls 42 des 1909 composés se fixent. Troisièmement, les données de test ne possèdent pas la même distribution de probabilité que les données d'apprentissage, car les composés dans les données de test ont été synthétisés sur la base des résultats enregistrés dans les données d'apprentis-

sage. Obtenir une précision supérieure à 60% est impressionnant comme il est indiqué dans (Peña *et al.*, 2007).

$IFSS_{GES}$ a été exécuté 61 fois dans le temps dont nous disposions pour nos expériences, avec *a priori* sur les structures arbitrairement fixé à $10^{-16\times f}$, où f est le nombre de paramètres libres dans le DAG. Les sorties ont été utilisées comme entrées d'un classifieur bayésien naïf et une classification sur les données d'essai a été réalisée. Comme le montre la figure 2, IFSS obtient des résultats entre 36% (très mauvais) et 71%, avec une moyenne de 55% et seulement 46 lancements de IFSS obtiennent plus de 50% en précision équilibrée, c'est-à-dire mieux que le classifieur aléatoire. Ces résultats sont comparables à MBOR (Rodrigues de Morais *et al.*, 2008) et IAMB (Tsamardinos *et al.*, 2003) qui permettent d'atteindre respectivement 53% — sur 10 lancements — et 54% — sur 114 lancements. C'est par contre un peu moins bon que PCMB (Peña *et al.*, 2007) qui obtient 63% (sur 114 lancements). Notons que dans cette expérience, chaque résultat d'$IFSS_{GES}$ a été obtenu en 220 minutes environ sur une machine de bureau classique, valeur comparable aux autres algorithmes.

Néanmoins, la meilleure MB obtient un score de 71.1%, ce qui est impressionnant par rapport à (Cheng *et al.*, 2002; Peña *et al.*, 2007). Il convient de mentionner que J. Cheng, le vainqueur de la KDD-cup 2001, obtient 68.4%, à partir des quatre variables obtenues par son algorithme (10695, 16794, 79651 et 91839) en utilisant un classifieur bayésien. Le tableau 4 compare ces deux MB en utilisant plusieurs classifieurs.

	IFSS					Cheng				
	κ	TPR	FPR	Acc	WAcc	κ	TPR	FPR	Acc	WAcc
Output model	0.420	0.467	0.085	0.809	0.691	0.316	0.633	0.264	0.711	0.684
NaiveBN	0.437	0.547	0.120	0.801	0.713	0.297	0.600	0.258	0.708	0.671
SVM	0.464	0.500	0.076	0.823	0.712	0.312	0.313	0.056	0.795	0.629
RForest	0.439	0.513	0.099	0.809	0.707	0.312	0.313	0.056	0.795	0.629

Tableau 4. *Résultats de classification en utilisant la meilleure MB avec celle de Cheng*

5.4. *Robustesse*

Lors d'une sélection sur des ensembles de données avec un grand nombre de variables et un petit nombre d'échantillons, la robustesse du processus de FSS devient aussi important que la performance du modèle obtenu. Par exemple, dans l'analyse de puces à ADN, les experts du domaine préfèrent nettement une sélection de gènes stables que dans la plupart des cas, car ces gènes sont ensuite analysés avec de multiples expériences qui exigent beaucoup de temps et d'effort (Aussem *et al.*, 2009). Avec de telles bases de données de grande dimension, tous les algorithmes de FSS sont soumis à une certaine variabilité. Comme indiqué dans (Saeys *et al.*, 2008), la robustesse peut être considérée de différents points de vue : la perturbation au niveau des instances (par exemple en supprimant ou en ajoutant des échantillons), au niveau des variables (par exemple par l'ajout de bruit à des variables), la variation des paramètres de l'algorithme de FSS, ou une combinaison d'entre eux. Ici, nous mettons l'accent

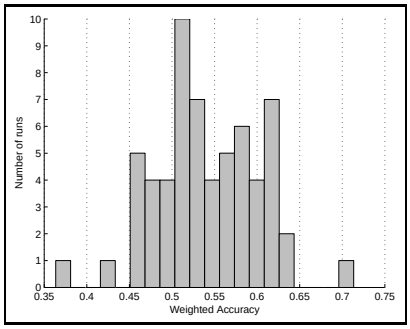

Figure 2. *Précision des résultats des* 61 *lancements d'*$IFSS_{GES}$.

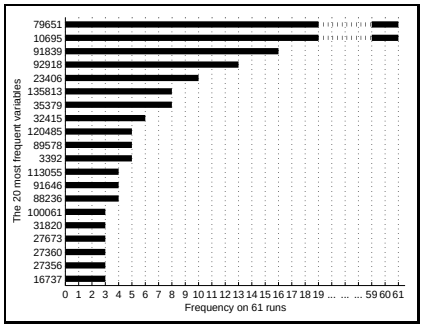

Figure 3. *Fréquences des* 20 *variables les plus fréquemment sélectionnées.*

sur la robustesse de l'algorithme de FSS définie par la variation du résultat lors d'une permutation aléatoire des variables.

On considère à nouveau les 61 lancements de IFSS sur des données THROMBIN. Les 20 variables les plus fréquemment sélectionnées sont présentées dans la figure 3 en ordre décroissant de fréquence. Comme nous pouvons le constater, les variables 79651 et 10695 ont toujours été sélectionnées. Ces variables sont également présentes parmi les quatre variables du gagnant de la KDD-cup en 2001, et la variable 79651 est toujours présente dans les 10 meilleures MB obtenues par KIAMB (Peña *et al.*, 2007). La troisième variable la plus fréquente, à savoir 91839, est également l'une des quatre variables du gagnant de la KDD-cup.

À la suite de (Kalousis *et al.*, 2007), nous adoptons une approche fondée sur la similitude d'ensembles pour mesurer la stabilité des variables en comparant les 61 sorties de IFSS. Nous utilisons l'indice de Jaccard qui est une mesure de similarité entre deux sous-ensembles S_1 et S_2. Plus les ensembles sont similaires, plus l'index de Jaccard est grand. La stabilité globale peut être définie comme la moyenne des similitudes paire-à-paire entre les $n = 61$ MB :

$$I_{tot} = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} I(S_i, S_j)}{n(n-1)} \quad \text{avec} \quad I(S_i, S_j) = \frac{|S_i \bigcap S_j|}{|S_i \bigcup S_j|}$$

Une moyenne de 0.336 (avec un écart-type de 0.116) a été obtenue pour $IFSS_{GES}$.

5.5. *Mesure empirique de la complexité*

Dans cette section, nous allons évaluer empiriquement la complexité de IFSS. Pour ce faire, nous considérons à nouveau la base de données THROMBIN. Lancer IFSS sur les parties de plus en plus grandes d'une base de données réelle fournit des informations sur la complexité empirique de l'algorithme dans un contexte réel. La tâche

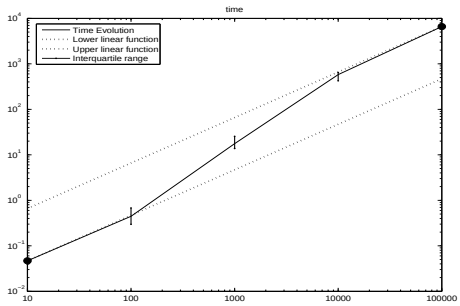

Figure 4. *Temps moyens d'exécution en échelle log-log selon le nombre de variables*

ici est encore d'extraire la MB à partir de 1909 composés. Nous choisissons au hasard des sous-ensembles de 0.01%, 0.1%, 1% et 10% de toutes les variables (dont la cible) et utilisons ces sous-ensembles comme entrée d'IFSS. L'ensemble du processus a été répété 20 fois. Nous rapportons dans la figure 4 l'évolution du temps (en échelle log-log) en fonction du nombre de variables. À chaque variation du nombre de variable, le temps d'exécution varie d'un facteur multiplicatif. Le graphique montre que que ce facteur est empiriquement en $O(n^{1,3})$. La complexité en temps quasi-linéaire de IFSS s'avère prometteuse pour une gamme d'applications dans le domaine de la biologie qui fait intervenir des centaines de milliers de variables.

6. Conclusion

Nous avons présenté et mesuré la performance d'un processus de sélection de variables ayant la capacité de passer à l'échelle. Cette procédure associe incrémentalement les résultats de méthodes d'apprentissage de structure de réseaux bayésiens qui fonctionnent uniquement sur des ensembles réduits de variables. La méthode s'est avérée très efficace en termes de temps de fonctionnement et de précision. De nouvelles expériences, plus intensives, et utilisant d'autres algorithmes d'apprentissage seront bientôt réalisées.

7. Bibliographie

Aussem A., Rodrigues de Morais S., Perraud F., Rome S., « Robust gene selection from microarray data with a novel Markov boundary learning method : Application to diabetes analysis », *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty ECSQARU'09*, vol. 5590 of *Lecture Notes in Artificial Intelligence*, Springer-Verlag Berlin Heidelberg, Verona, Italy, p. 724-735, 2009.

Cheng J., Hatzis C., Hayashi H., Krogel M., Morishita S., Page D., Sese J., « KDD Cup 2001 Report », *ACM SIGKDD Explorations*, p. 1-18, 2002.

Chickering D. M., « Optimal Structure Identification with Greedy Search », *Journal of Machine Learning Research*, vol. 3, p. 507-554, 2002.

Demsar J., « Statistical comparisons of claissifiers over multiple data sets », *Journal of Machine Learning Research*, vol. 7, p. 1-30, 2006.

Hua J., Tembe W., Dougherty E., « Performance of feature-selection methods in the classification of high-dimension data », *Pattern Recognition*, vol. 42, p. 409-424, 2009.

Kalousis A., Prados J., Hilario M., « stability of feature selection algorithms : a study on high-dimensional spaces », *Knowl. Inf. Syst.*, vol. 12, p. 95-116, 2007.

Ma S., Huang J., « Penalized feature selection and classification in bioinformatics », *Briefings in Bioinformatics*, vol. 5, p. 392-403, 2008.

Neapolitan R. E., *Learning Bayesian Networks*, Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

Nilsson R., Peña J., Björkegren J., Tegnér J., « Evaluating Feature Selection for SVMs in High Dimensions », *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD*, 2006.

Peña J., Nilsson R., Björkegren J., Tegnér J., « Towards Scalable and Data Efficient Learning of Markov Boundaries », *International Journal of Approximate Reasoning*, vol. 45, n° 2, p. 211-232, 2007.

Rodrigues de Morais S., Aussem A., « A Novel Scalable and Data Efficient Feature Subset Selection Algorithm », *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML-PKDD'08*, vol. 5212 of *Lecture Notes in Computer Science*, Springer-Verlag Berlin Heidelberg, Antwerp, Belgium, p. 298-312, 2008.

Rodrigues de Morais S., Aussem A., « A Novel Markov Boundary Based Feature Subset Selection Algorithm », *Neurocomputing*, vol. 73, p. 578-584, 2010.

Saeys Y., Abeel T., de Peer Y. V., « Robust Feature Selection Using Ensemble Feature Selection Techniques », *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD*, Antwerp, Belgium, p. 313-325, 2008.

Tang Y., Zhang Y., Huang Z., « Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis », *IEEE-ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, p. 365-381, 2007.

Thibault G., Aussem A., Bonnevay S., « Incremental Bayesian Network Learning for Scalable Feature Selection », *8th International Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France, 2009, Proceedings*, vol. 5772 of *Lecture Notes in Computer Science*, Springer-Verlag Berlin Heidelberg, p. 202-212, 2009.

Tsamardinos I., Aliferis C., Statnikov A., « Algorithms for Large Scale Markov Blanket Discovery. », *Florida Artificial Intelligence Research Society Conference FLAIRS'03*, p. 376-381, 2003.

Tsamardinos I., Brown L. E., Aliferis C. F., « The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. », *Machine Learning*, vol. 65, n° 1, p. 31-78, 2006.

Yaramakala S., Margaritis D., « Speculative Markov Blanket Discovery for Optimal Feature Selection. », *IEEE International Conference on Data Mining*, p. 809-812, 2005.