Stabilité de la sélection de variables sur des données haute dimension : une application à l'expression génique

David Dernoncourt

THESE DE DOCTORAT DE

L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité : Informatique Biomédicale

ÉCOLE DOCTORALE PIERRE LOUIS DE SANTÉ PUBLIQUE À PARIS : ÉPIDÉMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMÉDICALE

Présentée par

David DERNONCOURT

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Stabilité de la sélection de variables sur des données haute dimension : une application à l'expression génique

soutenue le 15 octobre 2014

devant le jury composé de :

M Antoine Cornuéjols, Professeur, AgroParisTech ……Rapporteur
Mme Barbara Heude, Chargée de recherches, INSERM ……..…….Rapporteur
M Nicolas Bredèche, Professeur, Université Paris 6 ……..……Examinateur
M Blaise Hanczar, Maître de conférences, LIPADE ……..……...Encadrant
Mme Karine Clément, Professeur, INSERM ……..……..……..…Co-directrice
M Jean-Daniel Zucker, Directeur de recherches, IRD……..……..……..……Directeur

Résumé

Les biotechnologies modernes sont capables, via les technologies dites « haut débit », de mesurer de très grandes quantités de variables à l'échelle de chaque individu : séquence ADN, expressions des gènes, profil lipidique... L'extraction de connaissances à partir de ces données peut se faire par l'utilisation de techniques d'apprentissage automatique, par exemple par des méthodes de classification. Cependant, ces données contiennent un très grand nombre de variables mais ne sont mesurées que sur quelques centaines de patients, dans les meilleurs des cas. Cette association d'une dimensionnalité élevée à une petite taille d'échantillon fait de la sélection de variables une étape préalable indispensable pour réduire le risque de surapprentissage, diminuer les temps de calcul, et améliorer l'interprétabilité des modèles.

Lorsque le nombre d'observations est faible, cette sélection a tendance à être instable, au point qu'il est courant d'observer que sur deux jeux de données différents mais traitant d'un problème similaire, les variables sélectionnées ne se recoupent presque pas. Pourtant, obtenir une sélection stable semble crucial si l'on veut avoir confiance dans la pertinence effective des variables sélectionnées à des fins d'extraction de connaissances. Dans le cadre de ce travail de thèse, nous avons tout d'abord cherché à déterminer quels sont les facteurs, au niveau des données, qui influencent le plus la stabilité de la sélection, sur tout de type de données (biologiques mais aussi d'autres domaines). Puis nous avons proposé une approche, spécifique aux données puces à ADN, faisant appel aux annotations fonctionnelles (Gene Ontology) pour assister les méthodes de sélection habituelles, en enrichissant les données avec des connaissances *a priori*. Cette approche, focalisée sur la stabilité fonctionnelle, n'a cependant pas permis d'améliorer la stabilité (aussi bien fonctionnelle qu'au niveau des gènes). Nous avons ensuite travaillé sur des méthodes d'ensemble, en nous focalisant tout d'abord sur l'influence de la méthode d'agrégation de l'ensemble sur la stabilité de la sélection, puis sur des ensembles hybrides. Si les méthodes d'ensemble peuvent permettre une amélioration de la stabilité parfois importante, ces résultats sont variables selon les jeux de données et s'obtiennent parfois au détriment de la précision du classifieur. Dans un dernier chapitre, nous appliquons les méthodes étudiées à un problème de prédiction de la reprise de poids suite à un régime, à partir de données puces, chez des patients obèses.

Publications en lien avec la thèse

Publications

• Dernoncourt D, Hanczar B, Zucker JD. Experimental Analysis of Feature Selection Stability for High-Dimension and Low-Sample Size Gene Expression Classification Task. *Proceedings of the12th IEEE International Conference on BioInformatics and BioEngineering (BIBE), IEEE*, 2012, pp. 350-355. doi: 10.1109/BIBE.2012.6399649

• Dernoncourt D., Hanczar B., Zucker J.D. Analysis of feature selection stability on high dimension and small sample data. *Computational Statistics & Data Analysis*, 2014, 71(C), pp. 681-693. doi: 10.1016/j.csda.2013.07.012

Communications, Posters

• Dernoncourt D, Hanczar B, Zucker JD. Évolution de la stabilité de la sélection de variables en fonction de la taille d'échantillon et de la dimension. *CAp (Conférence Francophone sur l'Apprentissage Automatique)*, Nancy, 23-25 mai 2012 [Communication orale]

• Dernoncourt D, Hanczar B, Zucker JD. Stability of Ensemble Feature Selection on High-Dimension and Low-Sample Size Data: Influence of the Aggregation Method. *ICPRAM 2014 - Proceedings of the 3nd International Conference on Pattern Recognition Applications and Methods*, Angers, France, 6-8 March, 2014, pp. 325-330 [Poster + short paper]

Table des matières

Table des figures

Liste des tableaux

Chapitre 1 :
Introduction

Les biotechnologies modernes sont capables, via les technologies dites « haut débit », de mesurer de très grandes quantités de variables à l'échelle de chaque individu : séquence ADN, expressions des gènes, profil lipidique... On assiste ainsi au développement de diverses sciences « omiques » (génomique, transcriptomique, lipidomique, protéomique...) exploitant ces techniques et les données qu'elles produisent. Parmi ces technologies, les biopuces (ou puces à ADN, en anglais *microarrays*), qui permettent de mesurer les niveaux d'expression de plusieurs dizaines de milliers de gènes (le transcriptome), sont une des méthodes les plus répandues, bien qu'en perte de vitesse face à l'émergence des technologies de *Next Generation Sequencing* (NGS).

L'extraction de connaissances à partir de ces données peut se faire par l'utilisation de techniques d'apprentissage automatique (Hastie et al., 2009). Cependant, ces données contiennent un très grand nombre de variables mais ne sont mesurées que sur quelques centaines de patients, dans les meilleurs des cas. Cette dimensionnalité élevée associée à une petite taille de l'échantillon, souvent appelé « fléau de la dimension » (*curse of dimensionality*) (Simon, 2003), représente un défi pour les techniques de classification, car toutes deux augmentent le risque de surapprentissage et diminuent la précision des classifieurs (Jain & Chandrasekaran, 1982). En outre, la dimensionnalité élevée peut augmenter le temps de calcul de façon excessive, car les classifieurs ne s'adaptent généralement pas très bien à un très grand nombre de variables. Enfin, un classifieur basé sur un très grand nombre de variables sera plus difficile à interpréter qu'un classifieur qui serait basé sur un nombre réduit de variables. Pour faire face à ces problèmes, une étape préalable de sélection de variables est utilisée pour réduire la dimensionnalité des données.

La sélection de variables consiste à retirer les variables non pertinentes ou redondantes de l'ensemble de variables initial $\mathcal{F} = \{f_1, f_2, ..., f_{|\mathcal{F}|=D}\}$, de façon à conserver un sous-

ensemble $S \subset \mathcal{F}$ ne contenant que des variables informatives et utiles pour la classification. Cette étape de sélection est critique, car d'une part, si les variables sélectionnées ne sont pas pertinentes, le classifieur ne pourra pas être bon, et d'autre part, si la sélection de variables est instable, cela réduira la confiance qu'on aura dans la pertinence effective des variables sélectionnées à des fins d'extraction de connaissances. Il est donc important d'avoir une sélection stable, ce qui en pratique n'est souvent pas le cas. Par exemple, dans (Miecznikowski et al., 2010), cinq tâches de classification portant sur un problème semblable (pronostic de cancer du sein à partir de données d'expression géniques) sont réalisées sur cinq jeux de données différents, résultant en des sélections de gènes presque sans recouvrement. Dans cette thèse, nous nous intéressons précisément à cette (in)stabilité de la sélection de variables.

Dans cette première partie, nous décrivons d'abord rapidement les principales données « omiques » haute dimension, et plus en détail le principe des puces à ADN, qui constituent la plupart des données que nous avons utilisées. Puis nous présentons les méthodes de classification supervisée fréquemment utilisées sur les données hautes dimensions, et enfin les méthodes de sélection de variables que nous utiliserons. La problématique du choix de la mesure de stabilité, ainsi que le détail des mesures de stabilité utilisées, sera abordée dans le chapitre 2.2.

1.1 Données « omiques », puces à ADN

1.1.1 Données « omiques »

Le terme de sciences « omiques » recouvre l'ensemble des sciences faisant appel aux technologies de biologie haut-débit. On distingue ainsi, par exemple, selon le niveau auquel on se place (Figure 1) :

- la génomique, qui s'intéresse à l'étude du génome (gènes) des individus ou des espèces, des interactions entre les gènes, et entre les gènes et l'environnement
- l'épigénomique, qui s'intéresse à l'ensemble des modifications épigénétiques (méthylation de l'ADN...)
- la transcriptomique, qui s'intéresse au transcriptome, c'est-à-dire les gènes transcrits (ARN). C'est dans ce contexte que sont utilisées les puces à ADN (présentées plus loin),

et plus récemment les techniques de RNA-seq (*RNA Sequencing* ou *Whole Transcriptome Shotgun Sequencing*)

- la protéomique, qui étudie l'ensemble des protéines et polypeptides, leurs structures, leurs interactions
- la lipidomique, qui étudie l'ensemble des lipides, et de leurs réseaux et pathways
- la métagénomique, qui étudie le matériel génétique d'un échantillon issu d'un environnement complexe (par exemple l'intestin, le sol...) : par exemple, le projet MetaCardis[1], auquel participe notre laboratoire, vise à étudier les liens entre le microbiote intestinal et les maladies cardiométaboliques.

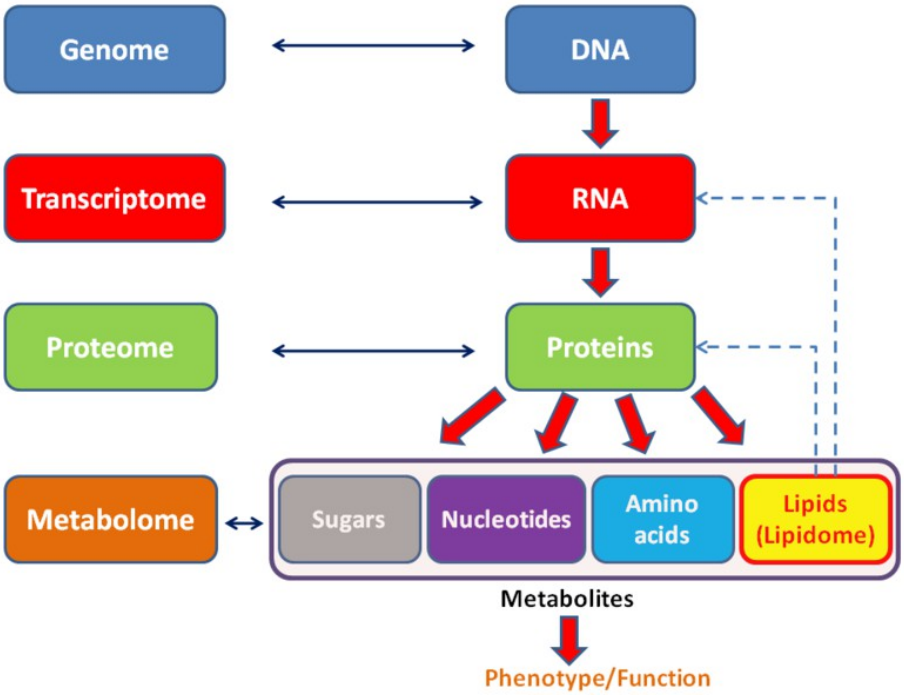

Figure 1: Schéma général indiquant les relations entre génome, transcriptome, protéome et métabolome (Wikipedia & Lmaps, 2009)

Ces champs d'études ont pour point commun de produire des données avec un grand nombre de variables (ensemble des gènes, ensemble des lipides...), le plus souvent sur un petit nombre de patients. D'un point de vue d'apprentissage automatique, ces données présentent donc des questions et problématiques relativement similaires : si l'on veut construire un classifieur

1 http://www.metacardis.eu

pour expliquer ou prédire la réponse à un traitement ou la survenue d'une pathologie à partir de données d'expression génique (transcriptome) ou à partir du lipidome, on se trouvera dans les deux cas face à une variable à expliquer y et une matrice d'observations X avec un grand nombre (plusieurs milliers) de colonnes (variables/gènes/lipides) et un petit nombre (quelques dizaines à quelques centaines, dans le meilleur des cas) de lignes (observations/individus).

Dans cette thèse, nous avons choisi, pour traiter du problème de l'apprentissage automatique sur données biologiques haute dimension, de nous focaliser sur les données puces, qui sont un type de données haute dimension produit dans notre laboratoire. Elles ont l'avantage d'être utilisées depuis déjà de nombreuses années, avec donc de nombreux jeux de données publiques disponibles.

1.1.2 Puces à ADN

Les puces à ADN (Gomase et al., 2008), également appelées biopuces, ou encore puces à gènes, permettent de mesurer les niveaux d'expression simultanés de plusieurs dizaines de milliers de gènes dans un prélèvement. Cette technologie a été publiée pour la première fois en 1995 (Schena et al., 1995), et s'est par la suite très rapidement répandue.

Techniquement, son principe de base est proche de celui du Southern blot (Southern, 1975), fondé sur l'hybridation entre deux séquences complémentaires d'acides nucléiques, mais réalisé sur un support beaucoup plus dense. La puce à ADN est une lame de verre de quelques cm², sur laquelle sont positionnées plusieurs dizaines de milliers de sondes d'ADN complémentaire (ADNc). Les différentes étapes d'acquisition de données d'expression à partir d'un échantillon sont schématisées sur la Figure 2. L'ARN messager (ARNm) est isolé à partir du prélèvement d'intérêt. Il subit ensuite une transcription inverse et une amplification en ADNc. Puis l'ADNc du prélèvement est marqué par des fluorochromes, par exemple par cyanine 3 (Cy3, vert) ou cyanine 5 (Cy5, rouge), avant d'être hybridé à l'ADNc de la sonde. La puce est ensuite lavée et lue par un scanner, qui numérise les intensités de couleurs, proportionnelles aux niveaux d'expression. Ces données brutes subissent ensuite divers prétraitements (filtrages des spots non exploitables, normalisation...) avant d'être exploitables.

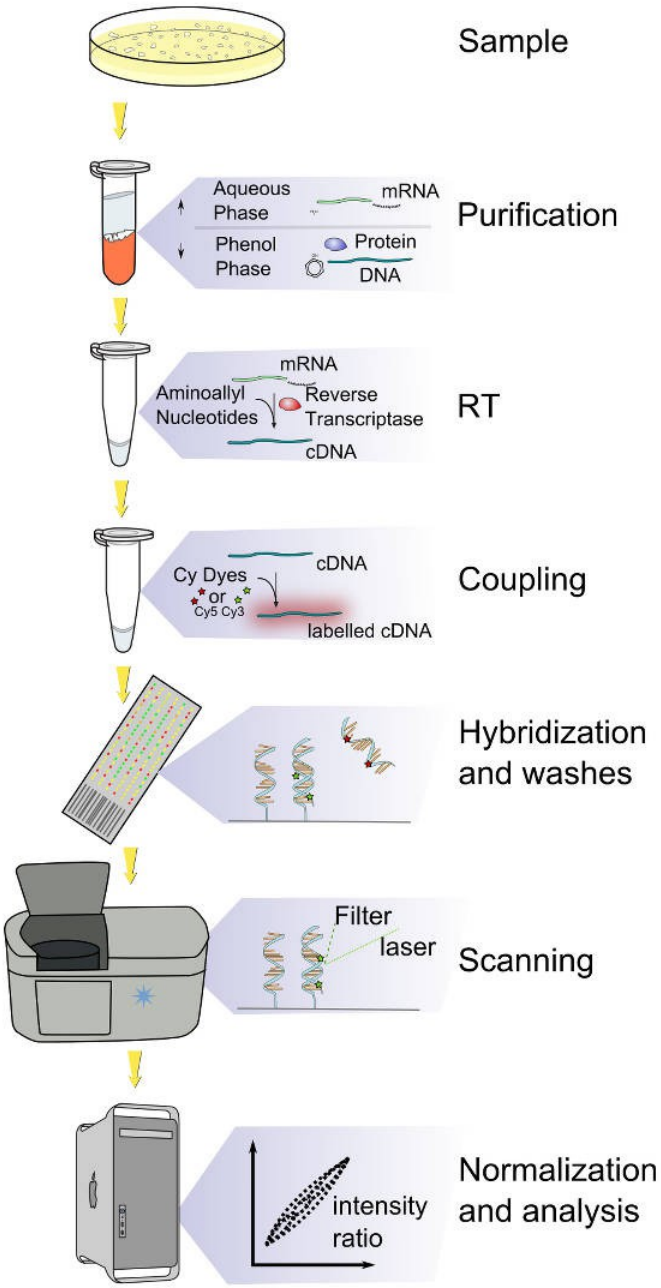

Figure 2: Vue schématique de la mesure de données d'expression par biopuce. Source: https://commons.wikimedia.org/wiki/File:DNA_microarray_experiment.svg

1.1.2.1 Puces spottées vs puces à oligonucléotides

Les sondes ADNc placées sur la puce peuvent y être installées via diverses méthodes. On distingue en particulier les puces « spottées » et les puces à oligonucléotides.

Dans les puces spottées, les sondes peuvent être des séquences d'ADNc, des produits de

PCR ou des oligonucléotides synthétisés avant d'être placés sur la puce. Ces sondes sont déposées à l'aide d'un micropipetteur robotisé. Ce type de puce a pour avantage de pouvoir être facilement personnalisable.

Dans les puces dites à oligonucléotides, les sondes sont des fragments nucléotidiques synthétisés directement sur la plaque, cette synthèse pouvant être réalisée par exemple par photolithographie ou par « jet d'encre » (Goldmann & Gonzalez, 2000). Les sondes de ce type de puce peuvent contenir de l'ordre de 25 à 70 paires de bases. Cette méthode est utilisée par des sociétés spécialisées telles qu'Agilent Technologies et Affymetrix, Inc., et est la plus fréquemment rencontrée.

1.1.2.2 Puces à une couleur vs puces à deux couleurs

Figure 3: Analyse d'expression par puces à ADN d'après (Stears et al., 2003). a) La puce à une couleur utilise un seul fluorochrome. Deux puces sont nécessaires pour générer les profils d'expression de deux échantillons. Les gènes sur- et sous-exprimés (respectivement verts et rouges) sont identifiés en superposant les images de différentes puces. b) Marquage deux couleurs, utilisant deux fluorochromes, pour marquer deux échantillons différents sur une même puce. Les gènes sur- et sous-exprimés sont identifiés en superposant les images obtenues par une double lecture de la puce à des longueurs d'onde différentes.

Sur une puce à deux couleurs (Figure 3b), on hybride deux échantillons simultanément (par exemple, un sujet malade vs un sujet témoin, ou un sujet vs un pool de référence) sur la même puce, chacun étant marqué différemment (typiquement, vert et rouge). Le ratio d'expression entre les deux échantillons sera ensuite déterminé à partir de l'intensité de chaque fluorochrome. Afin de corriger un éventuel biais lié aux fluorochromes, une procédure de dye-swap peut être réalisée (Churchill, 2002; Yang & Speed, 2002) : celle-ci consiste à renouveler l'expérience en inversant les fluorochromes entre les échantillons. Les résultats des deux expériences sont ensuite combinés afin de n'obtenir qu'une mesure synthétique.

Sur une puce à une couleur (Figure 3a), un seul échantillon est hybridé sur chaque puce, marqué par un seul fluorochrome. La comparaison de deux échantillons nécessite donc deux puces. La méthode a pour avantage de rendre plus facile la comparaison à des données provenant d'autres expériences, une fois les effets batch pris en compte.

1.1.2.3 Prétraitement des données

Le résultat de l'acquisition d'une puce ADN par le scanner est une image dans laquelle chaque spot coloré correspond à l'expression d'un gène détectée par une sonde. Ces données brutes doivent subir des étapes de prétraitement avant de pouvoir être analysées, car de nombreuses sources possibles de perturbation peuvent introduire une variabilité technique qui masquerait la variabilité biologique qu'on cherche à mettre en évidence. Par exemple, parmi les sources de variabilité on peut citer la fabrication des puces, les étapes de préparation des échantillons (purification, transcription inverse, amplification...), l'étape d'hybridation, qui est influencée par les conditions de température et d'humidité, le lavage, qui peut rendre certains spots hétérogènes (marques de lavage), et même l'acquisition elle-même (par exemple, la grille de lecture peut être décalée - Figure 4).

Parmi les traitements réalisés, on notera en particulier le filtrage des spots non exploitables, qui fait que toutes les sondes de la puce ne se retrouveront pas dans le jeu de données final, et une normalisation en plusieurs étapes : correction du bruit de fond (qui est mesuré pour chaque spot), mise à l'échelle des données, et log transformation des données. Cette mise à l'échelle peut être réalisée sur la base de la moyenne globale des intensités, sur des gènes « de ménage » (*house-keeping genes*), ou encore sur des sondes de contrôles.

Figure 4: Vérification de la détection des centroïdes des spots (Agilent Technologies, 2009)

1.1.3 Données biopuces utilisées

Dans cette thèse, nous utilisons cinq jeux de données biopuces publics pour le développement et la comparaison des méthodes de sélection. Ces données sont présentées dans le Tableau 1. Sur les données cancer du côlon (Alon et al., 1999), la tâche de classification consiste à différencier des échantillons de tissu colique sain et cancéreux (adénocarcinome). Sur les données cancer du côlon (Golub et al., 1999), la tâche de classification consiste à différencier des échantillons de moelle osseuse de leucémie aiguë myéloïde et de leucémie aiguë lymphoblastique. Sur les deux jeux de données cancer du sein (Pawitan et al., 2005 ; van de Vijver et al., 2002), la tâche de classification consiste à séparer les "bons" et les "mauvais" pronostics, le mauvais pronostic étant défini dans (Pawitan et al., 2005) comme une rechute ou un décès (toutes causes confondues) dans les 5 ans. Enfin, sur les données cancer du poumon (Bhattacharjee et al., 2001), la tâche de classification consiste à différencier les adénocarcinomes d'autres tumeurs pulmonaires ou de tissu pulmonaire sain.

Table 1: Jeux de données biopuces publics utilisés

Nom	*N*	*D*	*N/D*	*Source*
Cancer du colon	62	2000	0.03	Alon et al., 1999
Leucémie	72	7129	0.01	Golub et al., 1999
BK Pawitan	159	8112	0.02	Pawitan et al., 2005
Cancer du poumon	203	2000	0.10	Bhattacharjee et al., 2001
BK Vijver	294	2000	0.15	van de Vijver et al., 2002

Nous nous attacherons également, dans le dernier chapitre, à appliquer l'ensemble de ces méthodes et à étudier plus en détail, au niveau biologique cette fois, les sélections de variables obtenues sur un jeu de données biopuces ayant déjà fait l'objet de travaux dans notre laboratoire, DiOGenes. DiOGenes (Diet, Obesity and Genes) (Larsen et al., 2010 ; Mutch et al., 2011) est un projet européen visant entre autres à étudier, chez des sujets obèses et chez des sujets de poids normal, les déterminants (aussi bien génétiques que diététiques et comportementaux) de la prise de poids. Dans le cadre de ce projet, un groupe de 932 patients a suivi un régime hypocalorique de 8 semaines. Des prélèvements de tissu adipeux sous-cutané abdominal ont été obtenus par aspiration sous anesthésie locale au début (J0) et à la fin (S8) du régime hypocalorique. Les 596 sujets qui avaient perdu plus de 8% de leur poids ont poursuivi par un régime contrôlé normocalorique pendant 6 mois, et ont été classés en « repreneurs » (50-100% de reprise de poids) et « non-repreneurs » (0-10% de reprise de poids) de poids. Dans chacun de ces 2 groupes, 20 femmes ont été sélectionnées au hasard, en appariant les 2 groupes sur le poids, l'indice de masse corporelle, l'apport énergétique total, les taux sanguins de cholestérol, triglycérides, cholestérol HDL, adiponectine, C-reactive protein (CRP), glucose et insuline, ainsi que la résistance à l'insuline mesurée par HOMA-IR (Matthews et al., 1985). À noter que cet appariement n'a pas été réalisé au niveau individuel, mais au niveau des groupes dans leur ensemble (en utilisant les valeurs moyennes de chaque variable).

Les données d'expression issues des prélèvements de tissu adipeux ont été mesurées via des biopuces Agilent 4x44K *whole human genome*. Nous nous sommes intéressés ici à la prédiction de la reprise de poids après le régime à partir des données puces de J0. Ces données ont été normalisées en utilisant le package R goulphar (Lemoine et al., 2006), puis les sondes de contrôles ou sans identification, ainsi que gènes ayant des valeurs manquantes ont été retirés. À la fin de ces prétraitements, ce jeu de données contient D=13078 variables, pour N=40 observations.

1.2 Apprentissage automatique, méthodes de classification

L'apprentissage automatique (en anglais *machine learning*) est une branche de l'intelligence artificielle qui s'intéresse à l'extraction d'informations par des méthodes automatisées qui réalisent un « apprentissage » à partir de données. On distingue deux grandes

classes de méthodes d'apprentissage automatique (Figure 5) : l'apprentissage supervisé (classification et régression) et l'apprentissage non supervisé (clustering).

Figure 5: Le cycle de l'apprentissage automatique, tiré de (Kuncheva 2004)

Dans l'apprentissage non supervisé, on dispose de données contenant des variables observées, mais on ne dispose pas d'étiquettes permettant de regrouper les observations (individus) d'une manière ou d'une autre dans des classes préétablies. L'algorithme d'apprentissage non supervisé a pour but de découvrir la structure cachée des données : quelles

observations peuvent être regroupées, en combien de groupes (classes), d'après quelles variables.

Dans l'apprentissage supervisé, on dispose de données contenant des informations à la fois sur les D variables explicatives x_i (par exemple pour les biopuces, les expressions des gènes) et sur la variable à expliquer y (la classe, par exemple la réponse à un traitement, bonne ou mauvaise). Les observations consistent en des couples $((x_{1k}, x_{2k},...,x_{Dk}),y_k)$ et le rôle de l'algorithme de classification est alors de réaliser une correspondance entre les variables explicatives et la variable à expliquer, soit de rechercher un modèle $f()$ tel que $y=f(x_1, x_2,...,x_D)$. Cette correspondance doit éviter le surapprentissage, car son objectif est d'être généralisable : appliqué sur de nouveaux échantillons, dont la classe est cette fois inconnue, il doit permettre de prédire la classe de ces échantillons (Wessels et al., 2005).

On distingue également d'autres méthodes d'apprentissage automatique, notamment : l'apprentissage semi-supervisé, qui correspond au cas, intermédiaire entre classification et clustering, où certains échantillons d'apprentissage ont une classe connue et d'autres ont une classe inconnue ; l'apprentissage partiellement supervisé, dans lequel la classe des échantillons est partiellement connue (par exemple, pour un échantillon on sait qu'il n'appartient pas à la classe A, mais il appartient peut-être à la classe B ou à la classe C) ; et l'apprentissage par renforcement, dans lequel l'algorithme agit itérativement sur l'environnement et apprend des conséquences de ses actions.

Les problèmes qui nous intéressent dans le cadre de cette thèse font appel à de l'apprentissage supervisé : les données que nous utilisons disposent toutes d'une classe y connue à l'avance, telle que la réponse au traitement ou le type de tumeur sur les données puces. Ce type de problème est fréquent en biologie, car il est souvent possible d'obtenir l'information sur la classe soit directement (par exemple, examen histologique pour déterminer le type de cancer), soit en attendant suffisamment pour obtenir la réalisation de l'évènement (par exemple, réponse au traitement). Dans ce qui suit, nous présentons les principales méthodes de classification utilisées dans le cadre des données biopuces.

1.2.1 k plus proches voisins (kNN)

La méthode des k plus proches voisins (kNN pour *k-nearest neighbors*) (Cover, 1967) est

une méthode non paramétrique dans laquelle les nouveaux échantillons sont classifiés en fonction de leur proximité aux échantillons d'apprentissage. Face à un nouvel échantillon à classifier, la méthode des kNN va rechercher les k échantillons les plus proches dans l'ensemble d'apprentissage ; et faire un vote majoritaire sur leurs classes pour déterminer la classe du nouvel échantillon (Figure 6). Le seul paramètre du modèle est k, nombre de voisins à considérer, de préférence impair pour éviter les ex aequo. Dans les applications aux données avec un faible nombre d'observations, les valeurs de k sont classiquement faibles (3 ou 5). Il est également possible de pondérer les contributions des voisins pour donner plus de poids aux plus proches, par exemple en donnant aux voisins un poids inverse de leur distance à l'échantillon à classifier.

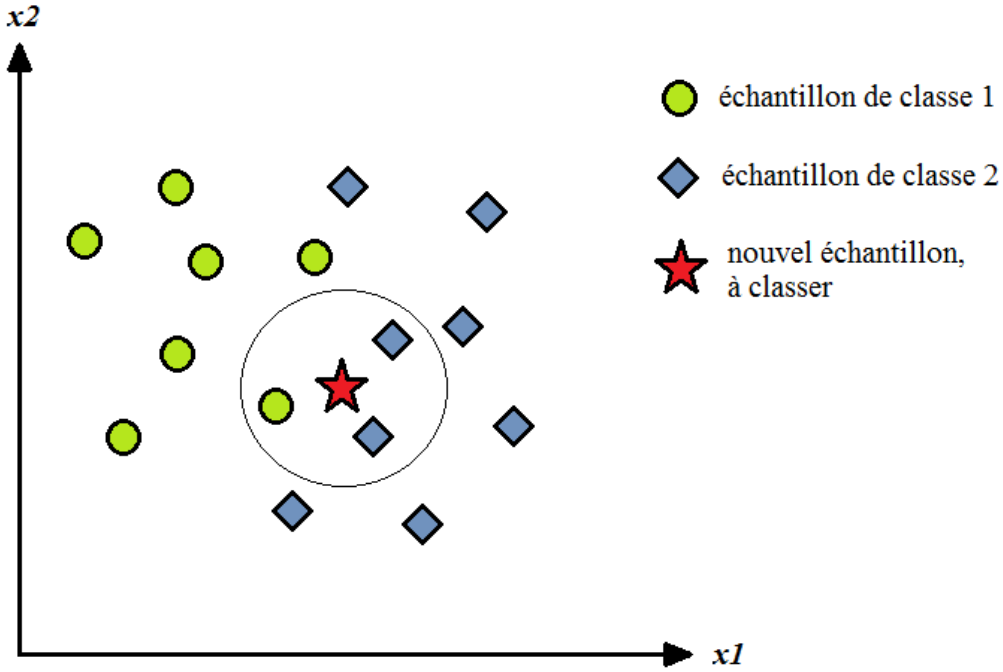

Figure 6: Méthode des k plus proches voisins. Le nouvel échantillon (étoile) est affecté à la classe majoritaire parmi ses k (ici k=3) plus proches voisins, ici la classe 2.

1.2.2 Analyse discriminante

L'analyse discriminante quadratique, l'analyse discriminante linéaire (*LDA* pour *linear discriminat analysis*) et l'analyse discriminante linéaire diagonale (*DLDA* pour *diagonal linear discriminat analysis*) sont basées sur l'analyse discriminante linéaire de Fisher (Fisher, 1936) et décrites dans (Dudoit et al., 2002). L'analyse discriminante linéaire de Fisher recherche une

combinaison linéaire $x\beta$ des variables $x = (x_1,...,x_D)$ qui maximise le ratio entre la différence intergroupe de $x\beta$ et sa variance intragroupe. La classe prédite y_k pour une observation $x_k = (x_{1k},...,x_{Dk})$ est la classe dont la moyenne des $x\beta$ observés est la plus proche de $x_k\beta$ et donc pour laquelle $P(X=x_k/Y=y_k)$ est maximum. Dans sa description initiale, l'analyse discriminante linéaire de Fisher est une approche non paramétrique.

Dans l'analyse discriminante quadratique, on ajoute l'hypothèse que les variables sont distribuées normalement dans chaque classe. Dans la LDA, on ajoute en plus l'hypothèse que la matrice de covariance Σ des variables est identique dans les deux classes : $\Sigma = \Sigma_{classe1} = \Sigma_{classe2}$. Dans la DLDA, on ajoute en plus l'hypothèse que cette matrice de covariance Σ, identique dans les deux classes, est diagonale : $\Sigma = \Delta = \text{diag}(\sigma_1,...,\sigma_p)$. Bien que ces hypothèses soient sans doute excessives dans le cas de données biopuces, en pratique cela n'empêche pas LDA et DLDA d'obtenir généralement de bonnes performances sur ce type de données (Dudoit et al., 2002 ; Dettling & Bühlmann, 2004 ; Shieh et al., 2006). Dans la plupart de nos expériences, nous utiliserons un classifieur LDA.

1.2.3 Machines à vecteurs de support (SVM)

Les machines à vecteurs de support (ou séparateurs à vaste marge, *SVM* pour *support vector machine*) (Cortes & Vapnik, 1995) sont basées sur la recherche d'un hyperplan qui sépare les deux classes avec une marge (distance entre l'hyperplan et les observations les plus proches) maximale. Les SVMs ont été conçus pour les tâches de classification binaires, mais il existe des variantes permettant de traiter les cas multicatégoriels.

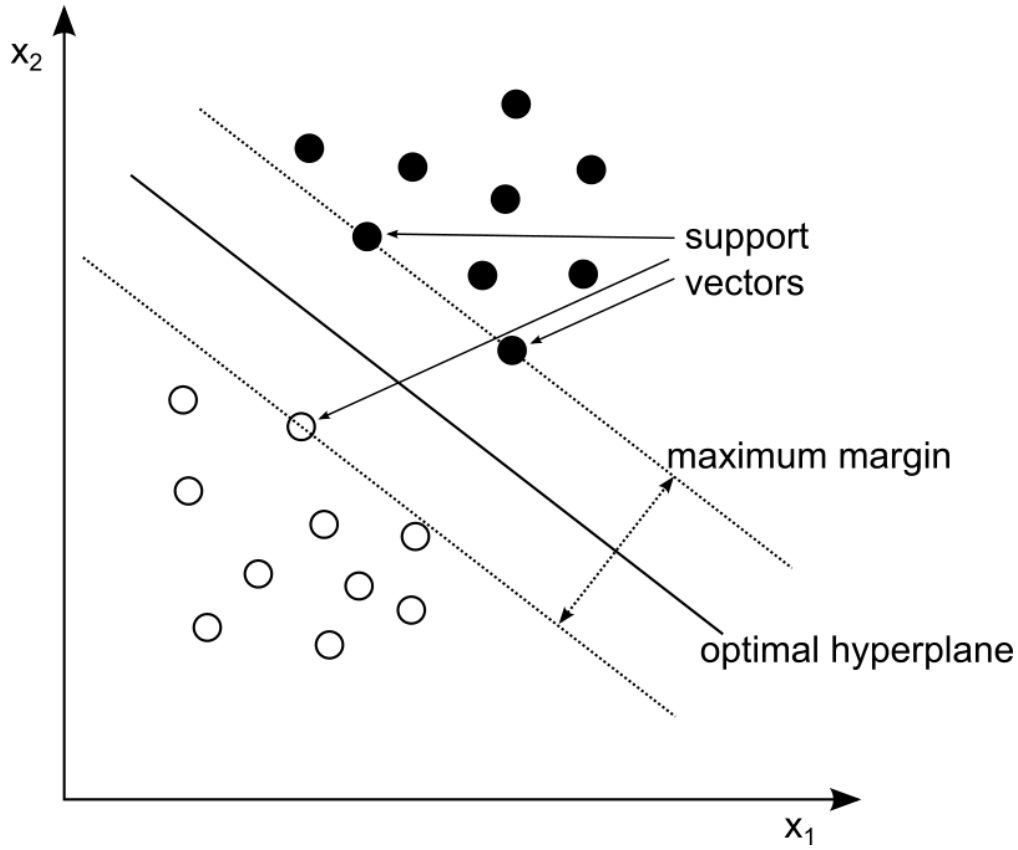

Figure 7: SVM: exemple d'un problème séparable dans un espace à 2 dimensions. Les vecteurs de support définissent les marges de la plus large séparation entre les 2 classes. D'après (Cortes & Vapnik, 1995)

Leur principe est plus facile à illustrer dans le cas d'un SVM binaire et linéaire : sur la Figure 7, les observations sont caractérisées par deux variables x_1 et x_2, et par leur classe 1 (noir) ou 2 (blanc) Le SVM linéaire calcul l'équation de l'hyperplan (dans cet exemple une droite) qui sépare les 2 classes avec la marge maximale et qui est sous la forme:

$$0 = w \cdot x + w_0$$

où w est un vecteur de poids, w_0 une constante, et x le vecteur des variables explicatives (dans cet exemple, $x=(x_1,x_2)$). Sur la Figure 7, il existe une infinité de séparateurs possible. Il est néanmoins intuitif que l'hyperplan choisi, passant le plus loin possible des observations les plus proches et donc maximisant la marge, est un meilleur séparateur. Les observations les plus proches de l'hyperplan, délimitant la marge, sont appelées vecteurs de support.

La plupart des problèmes réels ne sont pas linéairement séparables. Pour que les SVMs

soient utilisables sur ces problèmes, on peut alors utiliser deux techniques : la redescription dans un espace de dimension supérieure et la marge souple (*soft margin*).

Passer dans une dimension supérieure peut permettre de trouver un séparateur linéaire. On utilise pour ce changement de dimension une fonction non linéaire φ, qui permet de convertir les observations x de l'espace d'origine en $\varphi(x)$ dans l'espace de redescription. On recherchera alors dans cet espace l'hyperplan $0 = w \cdot \varphi(x) + w_0$ séparateur avec la marge maximale. Il n'est pas nécessaire de connaître φ pour trouver cet hyperplan. En effet, le calcul de l'hyperplan nécessite seulement la connaissance du produit scalaire entre les points de l'espace de redescription. Il suffit donc de définir une fonction noyau vérifiant $K(x_1,x_2) = \varphi(x_1) \cdot \varphi(x_1)$ pour être capable de trouver l'hyperplan. Cette "astuce du noyau" (*kernel trick*) a aussi pour avantage de réduire le temps de calcul, car le calcul des produits scalaires dans l'espace de redescription, de dimension plus élevée, serait plus long. Les fonctions noyaux les plus classiquement utilisées sont (Burges, 1998) :

- le noyau linéaire : $K(x_1,x_2) = x_1 \cdot x_1$, ce qui correspond en fait à une absence de changement d'espace et nous ramène au cas du SVM linéaire décrit plus haut,
- le noyau polynomial : $K(x_1,x_2) = (x_1 \cdot x_1+1)^d$, $d \in R_+^*$, et
- le noyau gaussien (ou radial) : $K(x_1, x_2) = e^{(-\gamma\|x_1-x_2\|^2)}$, $\gamma \in R_+^*$.

La marge souple introduit une tolérance pour des observations mal classées lors du calcul de l'hyperplan séparateur. La méthode ne recherchera alors plus un hyperplan parfaitement séparateur, mais un hyperplan qui minimise le coût lié aux erreurs de classement. Cela permet de gérer les cas non linéairement séparables, mais aussi de limiter le risque de surapprentissage, en mettant en balance un coût lié aux poids du modèle et un coût lié aux erreurs de classification. Les SVMs sont fréquemment utilisés sur les données biopuces et obtiennent alors généralement de bonnes performances, par exemple dans (Ramaswamy et al., 2001 ; Sayes et al., 2008).

1.2.4 Réseaux de neurones artificiels

Les réseaux de neurones artificiels sont basés sur une formalisation des réseaux biologiques, initialement décrite afin de mieux comprendre l'activité nerveuse (McCulloch &

Pitts, 1943). Puis, des études ont suggéré que le cerveau utilise le même algorithme pour traiter de multiples problèmes différents (hypothèse de l'algorithme d'apprentissage unique) : par exemple, des cellules ganglionnaires rétiniennes ont été connectées au thalamus somatosensoriel (Métin & Frost, 1989) ou au thalamus auditif (Roe et al., 1992), et les réponses obtenues, enregistrées respectivement dans le cortex somatosensoriel et le cortex auditif primaire étaient dans les deux cas assez proches de celles observées normalement dans le cortex visuel. Ainsi, imiter les méthodes de traitement de l'information trouvées dans le cerveau pourrait fournir un algorithme d'apprentissage automatique très puissant et polyvalent. C'est ce que les réseaux de neurones artificiels tentent d'accomplir.

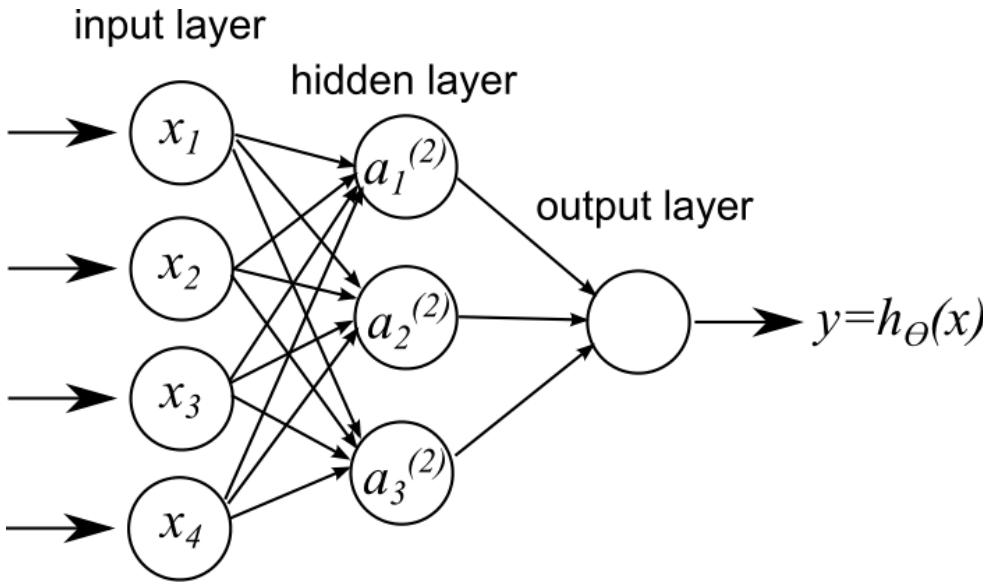

Figure 8: Un perceptron multicouche simple avec une couche d'entrée (4 neurones), une couche intermédiaire (3 neurones), et une sortie

La Figure 8 présente un perceptron multicouche, qui est un type classique et simple de réseau de neurones. Les neurones sont modélisés comme des unités ayant des entrées (équivalents de dendrites) et une sortie (équivalent de l'axone). Ces unités utilisent une fonction d'activation (souvent une fonction logistique) pour déterminer la valeur de sortie. Ils sont organisés en couches, avec une couche d'entrée qui reçoit les données (variables explicatives), une couche de sortie qui produit la prédiction (avec autant de neurones que de modalités dans la variable à prédire moins un), et une ou plusieurs couches cachées entre les deux. Cette structure permet aux réseaux de neurones de modéliser des interactions complexes et des relations non

linéaires (Geman et al.,1992). Cependant, il faut suffisamment de données pour pouvoir estimer les différents paramètres du modèle quand le réseau choisi est complexe (neurones ou liaisons nombreuses). Sur des données haute dimension avec petit échantillon comme les données puces, une étude comparative leur a trouvé des performances de classification plutôt inférieures à celles des SVMs ou des kNN (Statnikov et al., 2005).

1.2.5 Boosting

Le boosting est une méthode qui consiste à combiner des classifieurs faibles, c'est-à-dire dont la seule garantie est qu'ils soient un peu meilleurs que le hasard, pour créer un classifieur plus performant (Cornuéjols et al., 2002) (nous nous intéressons ici à la classification, mais les méthodes de boosting peuvent aussi être appliquées à la régression). Les classifieurs faibles sont entraînés par itérations successives, qui bénéficient de l'expérience des itérations précédentes par une pondération des observations (typiquement, ajout de poids aux observations mal classifiées et retrait de poids aux observations bien classifiées). Ces classifieurs sont ajoutés au modèle final avec en général une pondération relative à leur précision.

Figure 9: L'algorithme Adaboost, tel que présenté dans (Freund & Shapire, 1997).

AdaBoost (pour *Adaptive Boosting*) (Freund & Shapire, 1997) est l'une des méthodes de boosting les plus connues et est détaillé dans la Figure 9. Dans AdaBoost, à chaque itération *t*,

on recherche un classifieur faible $h_t(x)$ qui minimise le taux d'erreur ε_t pondéré par le poids relatif $p^t = {w^t}/{\sum_{i \in N} w_i^t}$ des observations. On enregistre $\beta_t = \varepsilon_t/(1-\varepsilon_t)$, qui servira à pondérer les classifieurs dans l'ensemble final, et pour l'itération suivante les poids w_i^t de chaque observation sont multipliés par $\beta_t^{1-|h_t(x_i)-y_i|}$. Le classifieur final réalise une somme pondérée par $log(1/\beta_t)$ des sorties des classifieurs faibles le constituant. Sur des données puces, des variantes telles que LogitBoost obtiennent généralement de meilleures performances qu'AdaBoost (Dettling & Bühlmann, 2003 ; Yang et al., 2010).

1.2.6 Forêts aléatoires

Les forêts aléatoires (*random forests*) (Breiman, 2001) sont une méthode non paramétrique basée sur l'agrégation d'un ensemble d'arbres de décision, l'idée étant qu'un tel ensemble s'avère plus robuste (moins de sensibilité au bruit, moins de surapprentissage) qu'un arbre de décision unique. Cette méthode s'appuie sur la « sagesse de la foule », qui nécessite, autant que possible, une diversité et une indépendance entre les arbres (Rokach, 2010). Afin de générer une telle diversité dans les arbres de décisions, chaque arbre est construit sur un rééchantillonnage du jeu d'apprentissage (tirage avec remise) et en n'utilisant qu'un sous-ensemble aléatoire des variables. Chaque arbre est déroulé jusqu'au bout (pas d'élagage, contrairement aux arbres de décisions simples dans lesquels au contraire l'élagage permet de limiter le surapprentissage). Une fois la forêt construite, pour l'appliquer aux données de validation on applique simplement un à un chaque arbre de décision aux données de validation. Chaque arbre produit une classification (on dit qu'il « vote »), et donc pour chaque observation de validation on obtient autant de classifications que d'arbres. On calcule ensuite pour chaque observation la proportion d'arbres ayant voté pour chaque classe, et cette proportion correspond à la probabilité de l'observation d'appartenir à une classe ou l'autre.

Bien que la publication d'origine (Breiman, 2001) se soit intéressée à des problèmes de basse dimensionnalité, les forêts aléatoires ont été utilisées avec de bons résultats sur des données hautes dimensions telles que des données biopuces (Díaz-Uriarte & Alvarez de Andrés, 2006 ; Sayes et al., 2008) ou des données protéomiques (Izmirlan, 2004). Les forêts aléatoires peuvent également être utilisées pour la sélection de variables (Genuer et al., 2010).

1.2.7 Estimation des performances du classifieur

Pour évaluer la pertinence du classifieur sans que la mesure soit biaisée par un surapprentissage éventuel, on mesure la précision (ou le taux d'erreur) de la classification qu'il réalise sur une base test qui n'a pas servi à la construction du modèle. En pratique, le nombre d'observations est trop faible pour qu'on puisse se contenter de diviser l'échantillon en une base d'apprentissage et une base de test (il y aurait trop de variabilité sur le taux d'erreur). Pour pallier cela, la méthode la plus classiquement utilisée est la validation croisée (*cross-validation*) (Kohavi, 1995).

Dans une validation croisée à k-fold (classiquement, k=10), les données sont divisées en une partition aléatoire de k blocs de même taille et ayant la même proportion d'individus de chaque classe que l'échantillon total (validation croisée dite « stratifiée », qui améliore l'estimation du taux d'erreur par rapport à une validation croisée non stratifiée (Boulesteix et al., 2008)). Puis chaque « fold » consiste à entraîner le classifieur sur un échantillon d'apprentissage constitué de k-1 blocs, puis à le tester sur le bloc restant, afin d'obtenir des valeurs prédites sur ce bloc. L'opération est répétée k fois, chaque fois avec un bloc différent comme échantillon de validation et les k-1 autres comme échantillon d'apprentissage. La validation croisée peut être répétée sur plusieurs partitions différentes, afin de réduire la variance inhérente à la validation croisée (Braga-Neto & Dougherty, 2004). L'estimation de l'erreur est obtenue en faisant la moyenne de toutes les erreurs observées sur les échantillons de validation.

1.3 Méthodes de sélection de variables

Les méthodes de sélection de variables peuvent être divisées en trois catégories : les méthodes filtres, les méthodes enveloppes (*wrapper*) et les méthodes intégrées (*embedded*) (Saeys et al., 2007 ; Bolón-Canedo et al., 2014) (Figure 10). Dans les méthodes intégrées, le processus de sélection de variables est réalisé en même temps que la construction du classifieur, par le système d'apprentissage lui-même, ce qui présente donc un risque de surapprentissage lorsque le nombre d'observations est faible comparé au nombre de variables. Dans les méthodes enveloppes, un sous-ensemble de variables est sélectionné en fonction des performances d'un classifieur construit sur ces variables. Du fait de cette exploration de l'espace des sous-ensembles de variables, avec pour chaque sous-ensemble plusieurs constructions de classifieur

pour une validation croisée, elles ont un coût computationnel élevé. Les méthodes filtres sélectionnent les variables indépendamment du classifieur. Elles sont donc rapides, aptes à traiter des jeux de données haute dimension.

Figure 10: Trois approches de la sélection de variables, de gauche à droite : méthodes filtres, méthodes enveloppes, méthodes intégrées. Tiré de (Saitta & Zucker, 2013).

Certains auteurs considèrent que les méthodes enveloppes devraient être préférées quand cela est techniquement possible (Pudil & Somol, 2008). Cependant, sur des données de très haute dimensionnalité, les filtres restent la méthode de choix pour des raisons de faisabilité, c'est pourquoi nous nous focaliserons sur eux, à l'exception du SVM-RFE (Guyon et al., 2002), qui est une méthode enveloppe rapide, développée initialement pour la sélection de gènes, et fréquemment utilisée. Une taxonomie des filtres est présentée dans (Lazar et al., 2012), et distingue en particulier les filtres univariés, qui ne prennent pas en compte les interactions entre les variables, et les filtres bi- ou multivariés, qui tiennent compte des interactions entre variables deux à deux ou dans leur globalité. Nous nous limiterons ici essentiellement aux méthodes que nous avons mises en oeuvre dans nos expériences. Le choix des méthodes utilisées s'est basé sur la fréquence de leur usage et/ou sur leur stabilité plutôt supérieure à la moyenne dans la littérature (cas du t-score et de SVM-RFE) (Haury et al., 2011; Tapia et al., 2011; Sayes et al., 2008), ainsi que, pour le t-score, sur son adéquation avec nos modèles de données artificielles.

Dans les formules suivantes, sauf mention contraire, l’ensemble de variables initial est

noté $\mathcal{F} = \{f_1, f_2, ..., f_{|\mathcal{F}|=D}\}$, et le sous-ensemble de variables sélectionnées est $S \subset \mathcal{F}$. x_{ik} est la réalisation de la variable *i* pour l'individu *k*, μ_{i1} est la moyenne de la variable *i* pour l'ensemble des individus de la classe *1*, σ^2_{i1} sa variance, μ_i est la moyenne de la variable *i* pour l'ensemble des *n* individus, et n_1 est le nombre d'individus dans la classe 1 ($n=n_1+n_2$). $\hat{\mu}_{i1}$ et $\hat{\sigma}_{i1}$ sont respectivement les moyennes et variances observées pour la variable *i* pour l'ensemble des individus de la classe *1*.

1.3.1 Filtres univariés

1.3.1.1 t-score

Le filtre t-score, ou t-test, est basé sur le test t de Welch (Welch, 1947), utilisé en statistiques pour tester la significativité de la différence entre les moyennes d'une variable dans deux groupe. Ce test fait l'hypothèse d'une distribution normale de la variable testée, ou d'un échantillon "suffisamment" grand (classiquement, n>30), et il tient compte, contrairement au t de Student, du cas où la variance de la variable testée est inégale entre les deux groupes. Sa formule est :

$$t_{Welch}(i) = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{\frac{\hat{\sigma}_{i1}}{n_1} + \frac{\hat{\sigma}_{i2}}{n_2}}}$$

Dans une utilisation pour la sélection de variables, ce t-score est réalisé sur toutes les variables entre les deux groupes à classifier. Puis les variables sont classées selon la valeur absolue de leur score, et le filtre garde les *d* meilleures variables (ce d étant un paramètre soit à définit *a priori*, soit à optimiser par une validation croisée imbriquée), ou encore toutes les variables ayant un score supérieur à un seuil à définir lui aussi. Bien que d'une part il soit très simple, et que d'autre part l'hypothèse de normalité de la distribution des variables n'est pas nécessairement réalisée sur les données puces, ce filtre a permis, sur données puces, d'obtenir des sélections parmi les plus stables (ou les moins instables !), tout en permettant de bonnes performances de classification (Haury et al., 2011).

Cette approche consistant à calculer un score, puis à l'utiliser pour classer les variables pour enfin sélectionner les *d* meilleures variables (ou des variables ayant un score supérieur à

une valeur seuil) est commune à la plupart des filtres, et en particulier aux filtres univariés (Lazar et al., 2012).

1.3.1.2 Rapport signal sur bruit

Le rapport signal sur bruit (en anglais *signal to noise ratio*, *SNR* ou encore *S2N*) est une mesure très proche du t-test, qui fait le ratio entre les différences des moyennes (signal) entre les deux classes et la somme des variances (bruit) sur les deux classes. Il est défini par :

$$S2N(i) = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\hat{\sigma}_{i1} + \hat{\sigma}_{i2}}$$

Comme avec le t-score, la sélection de variables est ensuite réalisée en conservant les d variables ayant le *S2N* le plus élevé. Cette méthode de sélection est moins fréquemment utilisée que le t-score, mais tout comme lui elle fait partie des méthodes de sélection les moins instables sur données haute dimension (Wald, 2013).

1.3.1.3 Information mutuelle

L'information mutuelle est une mesure de la dépendance statistique entre deux variables. Dans le contexte de la sélection de variables, il s'agit donc de mesurer la dépendance entre chaque variable explicative (gène...) et la variable d'intérêt (classe) : plus cette dépendance est forte, plus le filtre considère la variable pertinente. L'information mutuelle *I(X,Y)* entre une variable *X* et une variable *Y* est définit par :

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(X = x, Y = y) \cdot \log \left(\frac{p(X = x, Y = y)}{p(X = x) \cdot p(Y = y)} \right)$$

Cette définition considère le cas où les deux variables sont discrètes, ce qui est le cas pour *Y* mais pas pour les *X*. Il existe une définition continue, exploitant les densités de probabilité, et une définition mixte, correspondant au cas nous intéressant, avec une variable discrète (*Y*) et l'autre continue (*X*) :

$$I(X,Y) = \sum_{y \in Y} \int_{x} p(X = x, Y = y) \cdot \log \left(\frac{p(X = x, Y = y)}{p(X = x) \cdot p(Y = y)} \right) dx$$

Cependant, sur les petits échantillons, ces densités de probabilités sont difficilement estimables (Chow & Huang, 2005), et on utilise souvent la version discrète, avec des données d'expression discrétisées spécifiquement pour l'étape de sélection. Cette discrétisation peut cependant conduire à une perte d'information plus ou moins importante (Kotsiantis & Kanellopoulos, 2006) – certains algorithmes de discrétisation sont d'ailleurs basés sur la minimisation de cette perte d'information (Shi et al., 2014). Dans (Zhou et al., 2006), l'information mutuelle est utilisée comme première étape de sélection avant une seconde étape d'élimination des variables redondantes.

1.3.2 Filtres multivariés

1.3.2.1 CAT-score

Le t-score ajusté pour les corrélations (*CAT-score* pour *correlation-adjusted t-score*) est un filtre qui, comme son nom l'indique, est proche du t-score mais y ajoute la prise en compte des interactions entre les variables. Il a été décrit dans (Zuber & Strimmer, 2009), pour une utilisation sur des données biopuces. La Figure 11, tirée de ce papier, présente la relation entre cette mesure et le t-score. Pour cela on définit $P=(\rho i,j)$, matrice des corrélations entre les variables (gènes) et $V=(\sigma_1,...,\sigma_p)$, vecteur des variances des variables.

Figure 11: Relation entre fold-change, t-score et CAT-score, tiré de (Zuber & Strimmer, 2009)

Ou plus concisément, si τ est le vecteur des t-score des variables, τ_{adj} vecteur des CAT score des variables est défini par :

$$\tau_{adj} = P^{-1/2}\,\tau$$

Ainsi, le CAT score est par construction un t-score décorrélé, que l'on utilise de la même façon pour réaliser la sélection. Sur les données haute dimension et petit échantillon, ses auteurs recommandent de remplacer, dans son calcul, les corrélations et les variances par leur *shrinkage estimators* (Schäfer & Strimmer, 2005 ; Opgen-Rhein & Strimmer, 2007). C'est cette version,

implémentée dans R dans le package CMA, que nous utiliserons.

1.3.2.2 ReliefF

La méthode Relief a été initialement décrite par (Kira & Rendell, 1992) pour évaluer la pertinence des variables dans les problèmes de classification binaire. Il s'agit d'un algorithme itératif, qui part de poids nuls pour toutes les variables, puis tire au hasard une observation et 2 de ses proches voisins (un dans chaque classe), met à jour les poids des variables en fonction de la distance relative entre l'observation et ses 2 voisins pour chaque variable (le poids de la variable augmente si cette variable diffère plus entre le near miss et l'observation qu'entre le near hit et l'observation, et diminue dans le cas contraire), et répète cette opération *m* fois. Plus formellement, l'algorithme de Relief est le suivant :

```
diviser l'ensemble X des observations en
      X1 (observations de la classe 1)
      et X2 (observations de la classe 2)
initialiser le vecteur poids des variables W0=(0, ..., 0)
pour i=1 à m {
      prendre une observation aléatoire xi ∈ X
      choisir un near hit Hi au hasard parmi les observations
            les plus proches de x et de la même classe que xi
      choisir un near miss Mi au hasard parmi les observations
            les plus proches de x et de classe différente de xi
      mettre à jour W selon la formule Wi=Wi-1-(xi-Hi)²+(xi-Mi)²
}
```

Le paramètre *m* représente le nombre d'observations utilisées dans l'algorithme. En pratique, sur de petits échantillons, on le choisit généralement égal au nombre d'observations total de l'échantillon d'apprentissage. Dans la description d'origine, le *near hit* et le *near miss* sont choisis au hasard parmi des voisins « suffisamment proches ». Dans la version ReliefF proposée par (Kononenko, 1994), l'algorithme utilise les k *near hits* et *near misses* les plus proches et dont il moyenne les contributions afin d'obtenir une meilleure robustesse au bruit. ReliefF contient par ailleurs une adaptation au cas multiclasse, dans laquelle les *near misses* sont pondérés par la fréquence de leurs classes respectives. La fonction de mise à jour des poids devient alors :

$$W_i = W_{i-1} - (x_i - H_i)^2 + \sum_{C \neq classe(x_i)} P(C) \cdot (x_i - M(C)_i)^2$$

On utilise parfois également une version ReliefF-W (pour *weighted*), dans laquelle les *near hits* et *near misses* sont pondérés en fonction de leur distance à l'observation d'origine. Sur des données haute dimension, ReliefF fournit une sélection stable (ReliefF-W est moins stable), bien que sensible à l'ordre des observations (Yang et al., 2011), sans toutefois surpasser le rapport signal sur bruit (Wald et al., 2013). Les classifieurs construits sur les sélections aussi bien par ReliefF que par ReliefF-W obtiennent des taux d'erreur dans la moyenne (Bolón-Canedo et al., 2012 ; Hulse et al., 2012).

1.3.2.3 Couverture de Markov

Dans un réseau bayésien, la couverture de Markov (Figure 12) de la variable T correspond à l'ensemble de variables constitué de ses parents, enfants et coparents. Les parents P sont les variables qui ont un arc dirigé vers T (P cause T), les enfants C sont les variables qui ont un arc venant de T (T cause C), et le coparents Sp sont les variables qui ont un arc dirigé vers un des enfants de T (Sp et T causent C). Ainsi, la couverture de Markov de T est un ensemble de variables qui contiennent autant d'information utile à la prédiction de T que l'ensemble de toutes les variables (hors T).

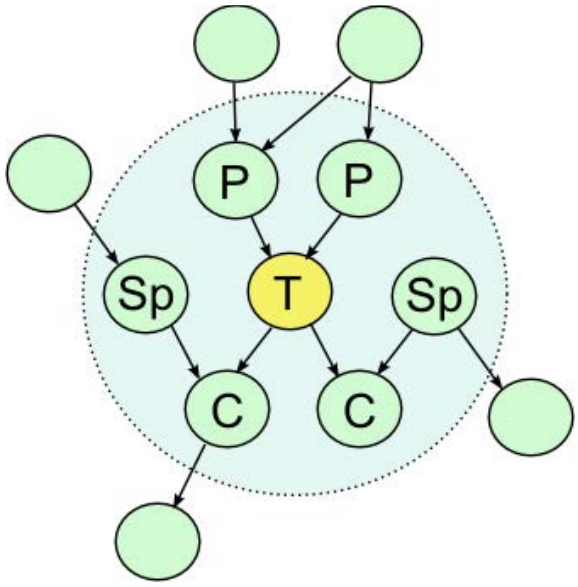

Figure 12: Exemple de couverture de Markov. L'ensemble des noeuds, liés par des arcs dirigés, correspond au réseau bayésien. La couverture de Markov de T, indiqué par un cercle pointillé, est constituée de ses parents P, ses enfants C et ses co-parents ou époux Sp.

Si l'on considère que les variables explicatives et la classe interagissent entre elles sous la

forme d'un réseau bayésien, la couverture de Markov correspond à la solution théorique optimale au problème de la sélection de variables (Tsamardinos & Aliferis, 2003). Mais en pratique, elle est difficile à obtenir.

De nombreux algorithmes ont été proposés, nécessitant généralement la discrétisation des variables. Par exemple, un des plus récents, IPC-MB (Fu, 2010), repose sur le G-test, qui est une statistique plus proche du χ^2 théorique que le χ^2 de Pearson sur de petits échantillons. Ces algorithmes ont pour point commun de réaliser de très nombreux tests d'indépendance conditionnelle, et perdent donc en fiabilité sur de très petits échantillons avec beaucoup de variables. Une variante ne nécessitant pas de discrétisation des variables, IPC-MBUVR/MWW (Dernoncourt et al., 2011), permet de meilleures performances en terme de précision de classification sur la sélection, mais reste inférieure à la plupart des autres filtres sur les très petits échantillons (n=~50). Par ailleurs, la couverture de Markov a tendance à produire une sélection instable (Trajdos et al., 2014).

1.3.3 SVM-RFE

La *recursive feature elimination* (RFE) (ou *backward feature elimination*) (Kohavi & John, 1997) consiste à éliminer progressivement les variables les moins discriminantes. Pour cela, on part de l'ensemble de variables initial, sur lequel on construit un classifieur, puis on élimine une certaine proportion des variables les moins discriminantes (on peut les éliminer une à une, mais sur des données hautes dimensions on prend généralement des paliers plus larges, par exemple 10% dans (Han & Yu, 2010)). Cette opération est répétée jusqu'à avoir éliminé un nombre suffisant des variables.

Dans le cas du SVM-RFE (Guyon et al., 2002), comme son nom l'indique, le classifieur utilisé est un SVM, généralement linéaire. Comme nous l'avons vu précédemment, le SVM produit un vecteur de poids *w* associé aux différentes variables. C'est ce vecteur de poids (mis au carré ou en valeur absolue) qui permet de classer les variables à chaque itération du SVM-RFE. Pour le scoring final, afin d'obtenir un score pour toutes les variables et non uniquement pour celles conservées dans la dernière itération, les variables éliminées conservent leur poids d'élimination. La sélection par SVM-RFE est relativement souvent utilisée sur les données biopuces, avec de bonnes performances à la fois en termes de stabilité et de précision du

classifieur construit sur la sélection (Han & Yu, 2012 ; Sayes et al., 2008 ; Tapia et al., 2011).

1.3.4 Évaluation de la méthode de sélection

L'évaluation de la méthode de sélection est double. En effet, comme nous l'avons vu en début d'introduction, une bonne sélection doit, dans l'idéal, à la fois permettre de bonnes performances de classification et être reproductible (donc stable).

La validation croisée, permettant d'évaluer les performances du classifieur, a déjà été décrite en 1.2.6. Elle s'applique exactement de la même façon dans le cas où une sélection de variables précède la classification. Il faut alors être attentif à bien réaliser autant de sélections que de classifications, à l'intérieur de la validation croisée, et non une sélection unique en amont de la validation croisée, qui surestimerait les performances.

La stabilité de la sélection peut s'évaluer par diverses mesures de stabilité, qui seront détaillées dans le chapitre suivant. Nous noterons juste ici que la validation croisée habituelle, divisant l'échantillon en un gros échantillon d'apprentissage et un petit échantillon de validation, n'est pas adaptée. En effet, les mesures de stabilité fonctionnent en comparant les similitudes entre deux sélections. Contrairement au cas de la performance de la classification, où les échantillons d'apprentissage et de validation ont des rôles asymétriques (le classifieur est entraîné sur l'échantillon d'apprentissage puis testé sur l'échantillon de validation dont la petitesse n'a que peu d'importance, cf le cas extrême de la *leave one out cross-validation* qui consiste à réaliser une validation croisée avec une observation unique dans l'échantillon de validation), la stabilité est étudiée en utilisant ces deux échantillons de façon symétrique. Une sélection est réalisée sur l'échantillon d'apprentissage A, une autre sélection est réalisée sur l'échantillon de validation B, puis la mesure de stabilité compare le degré de similitude entre ces deux sélections.

Si l'une des deux sélections est réalisée sur un trop petit échantillon, elle sera dégradée et la stabilité mesurée sera sous-évaluée, potentiellement de beaucoup. C'est ce qui est illustré sur la Figure 13, qui présente en ordonnée plusieurs mesures de stabilité de la sélection et en abscisse la taille n_A du premier échantillon A utilisé pour mesurer la stabilité. Le second échantillon B est constitué des observations non utilisées dans le premier, soit $n_B=N-n_A$ observations (avec dans cet exemple N=203). Quand n_A est très petit, la sélection réalisée sur A

est de mauvaise qualité, et la stabilité est basse. Quand n_A est très grand (proche de N), n_B est très petit et la stabilité est basse également : cela ne correspond pas à une mauvaise stabilité sur l'échantillon d'apprentissage, mais à une mauvaise stabilité sur l'échantillon de validation. La stabilité la plus élevée et la plus juste est obtenue pour $n_A=n_B=N/2$, qui correspond au cas où aucune des 2 sélections comparées ne pénalise la mesure, d'où la forme en cloche des courbes de stabilité en fonction de n_A.

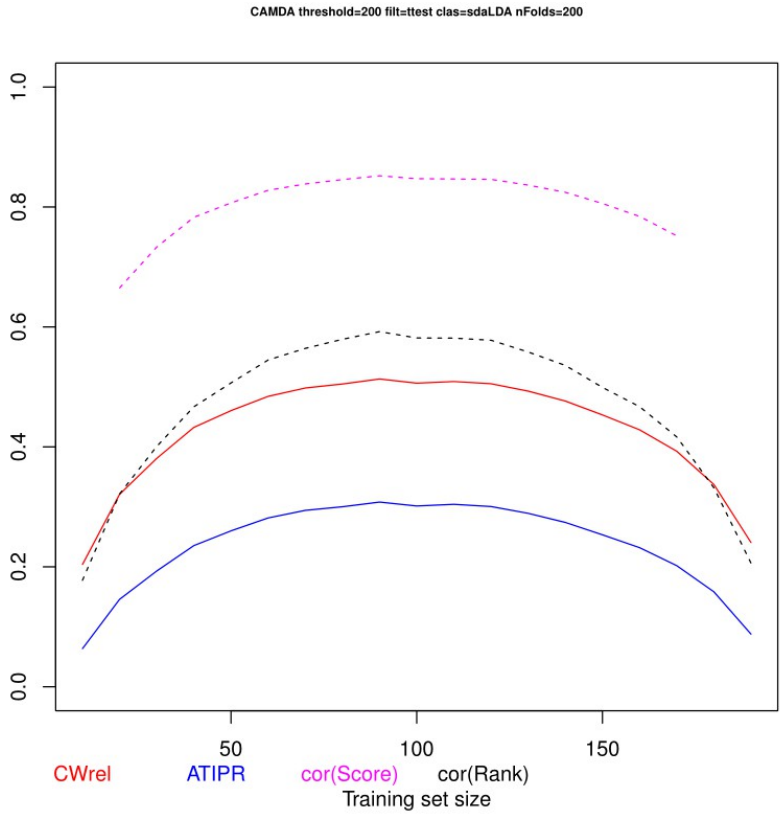

Figure 13: Stabilité de la sélection de variables entre 2 échantillons en fonction du déséquilibre de taille entre les deux échantillons. En abscisse, taille de l'échantillons A (l'autre est de taille *203 - [taille de A]*), en ordonnée, valeur des différentes mesures de stabilité (CW_{rel}, ATI_{PA}, corrélation des scores et corrélation des rangs, cf chapitre 2). Les données utilisées sont tirées de (Bhattacharjee et al., 2001)

On pourrait penser qu'il suffirait alors de comparer des sélections entre plusieurs jeux d'apprentissage successifs, et c'est d'ailleurs ce qui est fait dans un certain nombre de publications. Cependant, du fait du grand nombre d'observations communes entre les échantillons d'apprentissage que cela entraîne, la stabilité mesurée sera alors surévaluée (Haury et al., 2011). La Figure 14 présente les mêmes mesures de stabilité, sur le même jeu de données,

et en faisant varier la taille n_A de l'échantillon A comme sur la Figure 13, mais cette fois on veille à avoir toujours $n_A=n_B$ et on autorise, pour cela, les superpositions aléatoires entre l'échantillon A et l'échantillon B. Par exemple, pour deux échantillons de $n_A=n_B=N/2$ observations, il est probable qu'il y ait autour de N/4 observations communes entre les deux échantillons. On observe qu'en dehors des tailles d'échantillons les plus faibles, où l'instabilité due au petit échantillon prédomine sur la légère surévaluation de la stabilité, liée à quelques rares observations communes, les stabilités mesurées sont nettement plus élevées que dans le cas sans superpositions.

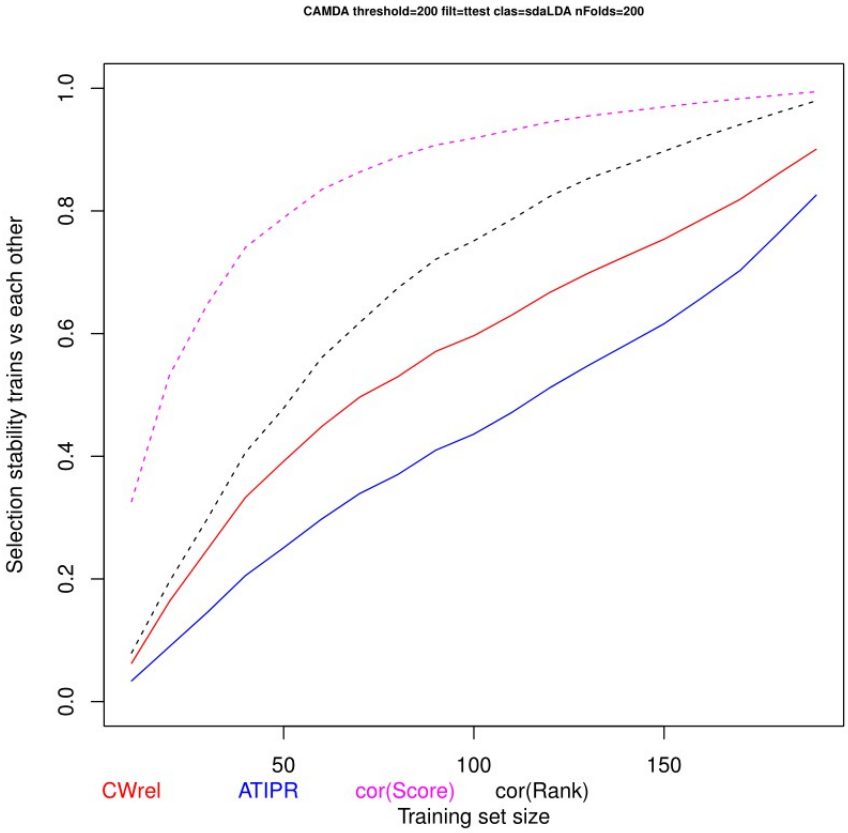

Figure 14: Stabilité de la sélection de variables entre des échantillons d'apprentissage ayant des observations communes. En abscisse, la taille des échantillons. Le nombre total d'observations est de 203 (données de (Bhattacharjee et al., 2001)), la proportion attendue d'observations communes entre deux échantillons est donc par exemple de ~50% pour des échantillons de taille 100.

La méthode que nous avons donc retenue consiste à diviser l'échantillon en deux échantillons de taille égale, ce qui nous permet à la fois de mesurer la stabilité et de mesurer la

performance de classification, en utilisant l'un de ces échantillons comme échantillon d'apprentissage et l'autre comme échantillon de validation. Comme dans une "validation croisée" où k serait égal à 2. Lors de la constitution de ces échantillons, nous avons également veillé à conserver les proportions de chaque classe telles que dans l'échantillon d'origine. Afin d'éviter une trop grande variabilité, cette opération est répétée un certain nombre de fois (dans nos expériences, généralement 50 fois), à la manière de la validation croisée répétée (Braga-Neto & Dougherty, 2004).

Nous obtenons alors un taux d'erreur de classification un peu surestimé, mais nous avons considéré ce biais comme acceptable car :

- il dégrade la mesure de la performance de classification mais optimise la mesure de la stabilité, qui est notre centre d'intérêt ici : nos mesures de performances de classification sont là avant tout pour nous assurer que les méthodes améliorant la stabilité ne le font pas au détriment de la précision de la classification
- et surtout, il s'agit du meilleur compromis possible : on a une réduction de la taille d'échantillon de 45% sur l'évaluation de la classification par rapport à une validation croisée 10-fold, mais on évite soit une réduction de la taille d'échantillon de 89% sur l'évaluation de la sélection de variables (cas des bords de la Figure 13), soit une quantité de 90% d'observations communes entre deux sélections (cas où n=~180 sur la Figure 14).

Chapitre 2 :
Analyse de la stabilité de la sélection de variables sur des données haute dimension et petit échantillon

Ce chapitre a été publié sous un format légèrement différent dans *Computational Statistics and Data Analysis* (Dernoncourt et al., 2014). Nous y abordons dans un premier temps comment mesurer la stabilité de la sélection de variables, avec une taxonomie des mesures de stabilité, un choix de mesures pertinentes via cette taxonomie, et la description d'une nouvelle mesure de stabilité ajustée sur la taille de la sélection. Nous utilisons ensuite ces mesures pour étudier comment les différentes caractéristiques des données (nombre d'observations, nombre de variables, distribution des variables) ainsi que le choix du seuil de sélection peuvent influencer cette stabilité.

L'analyse est réalisée en trois étapes : la première est théorique, à l'aide d'un modèle mathématique simple, la seconde est empirique, à partir de données artificielles sans puis avec corrélations entre les variables, et la dernière est basée sur des jeux de données réels. Nous nous intéressons principalement aux petits échantillons et aux données puces, néanmoins des jeux d'autres domaines et avec plus d'observations et moins de variables ont également été utilisés.

Les trois analyses obtiennent des résultats convergents : tous les facteurs étudiés ont une influence plus ou moins importante sur la stabilité de la sélection de variables, mais le facteur qui semble prépondérant est le ratio N/D du nombre d'observations sur le nombre de variables. Ceci nous suggère qu'une des pistes les plus intéressantes pour améliorer la stabilité de la sélection serait des méthodes capables d'améliorer ce ratio, par exemple l'augmentation de N par la combinaison de jeux de données traitant d'un problème identique ou très proche, ou la diminution de D via l'utilisation de données *a priori*, qui fera l'objet du chapitre suivant.

Abstract

Feature selection is an important step when building a classifier on high dimensional data. As the number of observations is small, the feature selection tends to be unstable. It is common that two feature subsets, obtained from different datasets but dealing with the same classification problem, do not overlap significantly. Although it is a crucial problem, few works have been done on the selection stability. The behavior of feature selection is analyzed in various conditions, not exclusively but with a focus on t-score based feature selection approaches and small sample data. The analysis is in three steps: the first one is theoretical using a simple mathematical model; the second one is empirical and based on artificial data; and the last one is based on real data. These three analyses lead to the same results and give a better understanding of the feature selection problem in high dimension data.

2.1 Introduction

Classification tasks in which the number of features D is much larger than the number of samples N are an increasingly frequent problem and became recently a research area of its own (Hastie et al., 2009). For instance, in computational biology, microarray data contain the simultaneous expression of tens of thousands of genes, and metagenomic data contain in the order of a few millions of genes... usually measured on (at most) a few hundreds patients. High dimensionality and small sample size pose a challenge to classification techniques, since they both increase the risk of overfitting and decrease the accuracy of classifiers (Jain & Chandrasekaran, 1982). More-over, high dimensionality can increase computation time beyond reasonable limits, as classifiers usually do not scale too well to huge numbers of features. To deal with these problems, feature selection is used to reduce data dimensionality.

Feature selection refers to the process of removing irrelevant or redundant features from the original set of features $\mathcal{F} = \{f_1, f_2, ..., f_{|\mathcal{F}|=D}\}$, so as to retain a subset $S \subset \mathcal{F}$ containing only informative features useful for classification. Feature selection methods can be broken down into three categories: filter, wrapper and embedded methods (Saeys et al., 2007). It is generally agreed that wrappers or embedded methods should be preferred if technically feasible (Pudil and Somol, 2008), however, on very high dimensional data, filters remain the method of choice for tractability reasons, which is why we will focus on them.

Beyond classification performance, the other main objective of feature selection is to obtain a reliable and robust list of predictive variables (signature). A good signature must not overfit the available data and be exportable to other datasets related to the same classification problem. These conditions can not be respected if the subset of selected features is highly variable. A lot of examples in the literature show that in small-sample or high dimension settings, the feature selection is not stable. For instance, in (Miecznikowski et al., 2010), five classification tasks dealing with a similar problem (breast cancer prognosis prediction from gene expression data) were performed on five different datasets (among which the van de Vijver and Pawitan datasets), leading to highly variable results of the individual gene analysis. Several other studies, such as (Ioannidis, 2005), (Michiels et al., 2005), (Ein-Dor et al., 2006) and (Haury et al., 2011), emphasized the difficulty to obtain a reproducible gene signature on high-dimension small-sample data. This difficulty to find a common subset of predictors between such different but similar datasets, or even between different sample subsets from a same dataset, raises the problem of feature selection stability.

Few studies have already dealt with this problem, and most of them have focused on comparing the stability of different, pre-existing or new feature selection methods, without exploring how different types of variations in the training sets affect this stability (for instance, Kalousis et al. (2005), Somol et al. (2009) and (Yao & Wang, 2013)). Moreover, they most often used stability measures which could be biased by the proportion of selected features (most stability measures artificially increase when the proportion of selected features increases) or by the amount of non-selected features (some stability measures take into account the stability of both selected and unselected features, so can be excessively high on datasets containing a large proportion of easy to exclude, irrelevant features). In this work, we investigate the behavior of the feature selection stability and its impact on the classifiers. We first present the main measures of selection stability used in machine learning and propose corrections of some of them that are biased. Then we present our analysis of the behavior of feature selection in three steps. In the first step we present a theoretical analysis of the performance and stability of feature selection on a simple Gaussian model. The second step is an empirical analysis performed on a large number of simulations based on artificial data. In the last step we present results of selection stability on real data. These three analyses lead to the same conclusions: in

high dimensions feature selection is not stable and the probability for relevant features to be selected can be very low.

2.2 Stability measures

The stability of a feature selection method was defined in (Kalousis et al., 2007) as *the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution*. To evaluate this robustness, quite a few different stability measures have already been described. We follow the taxonomy presented by Somol and Novovicová (Somol & Novovicová, 2010), who distinguished:

- *feature-focused* versus *subset-focused* measures: the former evaluate feature selection frequencies over all feature subsets considered together as a whole, while the latter evaluate similarities within every pairs of selected feature subsets. Both types provide complementary information, so we want to have at least one of each.
- *selection-registering* versus *selection-exclusion-registering* measures: the first only considers the stability of selected features while the latter also measures the stability of excluded features. On large datasets where a huge number of features are irrelevant and easy to exclude, selection-exclusion-registering measures will be strongly upward biased, so we will only be interested in selection-registering measures here.
- *subset-size-biased* versus *subset-size-unbiased* measures: the first yield values bounded more tightly than [0;1], with most notably the lower bound strongly increasing with the proportion of selected features, the latter are adjusted to be actually bounded by [0;1]. Obviously, for better generalization, we want to use *subset-size-unbiased* measures.

2.2.1 Relative weighted consistency, an unbiased feature-focused measure

Among the stability measures sorted in the above-mentioned taxonomy, only one was both selection-registering and subset-size-unbiased: the relative weighted consistency CW_{rel} (Somol & Novovicová, 2010). It was based on a subset-size-biased measure, the weighted consistency CW, corrected to be actually bounded by [0;1] no matter the proportion of selected features. A value of 0 indicates the highest possible instability, while a value of 1 indicates the highest

possible stability, i.e., if all feature subsets have the same cardinality, all subsets are identical.

Let $\mathcal{S} = \{S_1, S_2, ..., S_\omega\}$ be a system of ω feature subsets obtained from ω runs of the feature selection routine on different samplings, $\Omega = \sum_{i=1}^{\omega} |S_i|$ be the total number of occurrences of any feature in $\mathcal{S}$ and F_f be the number of occurrences of feature $f \in \mathcal{F}$ in system $\mathcal{S}$. *CW* was defined as follow:

$$CW(\mathcal{S}) = \sum_{f \in X} \frac{F_f}{\Omega} \cdot \frac{F_f - 1}{\omega - 1}$$

and CW_{rel} was then derived by adjusting *CW* on its minimal and maximal possible values CW_{min} and CW_{max}:

$$CW_{rel}(\mathcal{S}, \mathcal{F}) = \frac{CW(\mathcal{S}) - CW_{min}(\Omega, \omega, \mathcal{F})}{CW_{max}(\Omega, \omega) - CW_{min}(\Omega, \omega, \mathcal{F})}$$

2.2.2 Partially adjusted average Tanimoto index, an unbiased subset-focused measure

CW_{rel} is a *feature-focused* measure, so we looked for a *subset-focused* measure to complement it. Kuncheva's stability index (Kuncheva, 2007) and the stability measure defined in (Krizek et al., 2007) are both *subset-focused*, but they can only be used on subsets of equal cardinality. We retained the Average Tanimoto Index *ATI*, also introduced in (Somol & Novovicová, 2010). ATI is a generalization based on Kalousis's similarity measure S_S between two sets S_i and S_j (Kalousis et al., 2005):

$$S_S(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$$

This similarity index is computed over all subset pairs, then averaged:

$$ATI(\mathcal{S}) = \frac{2}{\omega(\omega - 1)} \sum_{i=1}^{\omega-1} \sum_{j=i+1}^{\omega} S_S(S_i, S_j)$$

ATI is *subset-focused* and *selection-registering*, but it is also *subset-size-biased*. We

propose a correction of this index, the partially adjusted average Tanimoto index ATI_{PA}. It is defined as follow:

$$ATI_{PA}(\mathcal{S}) = Max(\frac{ATI(\mathcal{S}) - ATI_{exp}(\mathcal{S})}{ATI_{max}(\mathcal{S}) - ATI_{exp}(\mathcal{S})}, 0)$$

where ATI_{max} is the maximal possible value of *ATI* and ATI_{exp} is the expected value of *ATI* when feature subsets are randomly defined. Because we will use a feature selection method which outputs a subset of predefined size, $ATI_{max} = CW_{max} = 1$ (when all feature subsets are identical). To obtain ATI_{exp}, we used an experimentally-determined approximation, computed as a function of the proportion of selected features. It should be noted that the correction we perform in ATI_{PA} slightly differs from the one performed in CW_{rel}: CW_{rel} is adjusted on the smallest possible value, while ATI_{PA} is adjusted on the expected value. We preferred that second approach to adjust *ATI* because we think it makes it easier to show a stability score higher than random, but for CW_{rel} we kept the original adjustement published by Somol in order to avoid an unnecesssary multiplication of incomparable metrics. The max operator ensures that ATI_{PA} is within the [0;1] interval and not negative as it could happen for the first argument of the max if the stability happens to be worse than random.

2.2.3 Correlation-based measures

Both *ATI* and *CW* focus on the stability of selected features. While this aspect is important for knowledge discovery, for the purpose of evaluating feature selection methods, the stability of the score over all features may be an interesting information, too. *Selection-exclusion-registering* measures – such as the average normalized Hamming index, derived from the Hamming index introduced in (Dunne et al., 2002) – will be too biased when the proportion of excluded features is too high. Correlations of features scores and ranks, on the other hand, provide a more balanced overview, even though the latter will be penalized when lots of features have a similar relevance, which occurs for example when lots of features are equally irrelevant in a very high-dimensional dataset. So, we used the average score (or weight) correlation $\overline{S}_W$ and the average rank correlation $\overline{S}_R$, as described in (Kalousis et al., 2005). The average score correlation is:

$$\overline{S_W} = \frac{2}{\omega(\omega-1)} \sum_{i=1}^{\omega-1} \sum_{j=i+1}^{\omega} \frac{\sum_{k=1}^{D}(w_i^{(k)} - \mu_{wi})(w_j^{(k)} - \mu_{wj})}{\sqrt{\sum_{k=1}^{D}(w_i^{(k)} - \mu_{wi})^2 \sum_{k=1}^{D}(w_j^{(k)} - \mu_{wj})^2}}$$

where $w^{(k)}_i$ is the score attributed to feature f_k in the ith run of the feature selection procedure (used to obtain S_i), and μ_{wi} is the average score attributed to features in the ith run of the feature selection procedure. The average rank correlation is:

$$\overline{S_R} = \frac{2}{\omega(\omega-1)} \sum_{i=1}^{\omega-1} \sum_{j=i+1}^{\omega} \left(1 - 6\frac{\sum_{k=1}^{D}(r_i^{(k)} - r_j^{(k)})^2}{D(D^2-1)}\right)$$

where $r^{(k)}_i$ is the rank attributed to feature f_k in the ith run of the feature selection procedure (obtained by ranking $|w^{(k)}_i|$).

2.3 Analysis on the mathematical model

The objective of this analysis is, by using a very simple model, to compute theoretically the performance and stability of the feature selection depending on the data parameters

Let us consider a classification problem in D dimensions. The two classes C_1 and C_2 are equally likely. All features are independent. On each feature f_i, the two classes follow respectively a Gaussian distribution $\mathcal{N}(-\mu_i, 1)$ and $\mathcal{N}(\mu_i, 1)$. We consider two types of features: informative features and non-informative features. For all informative features $\mu_i = \mu^*$ and for all non-informative features $\mu_i = 0$. Which means only informative features can discriminate the classes and be useful for classification. Let us consider a set of D_g informative features called F_g and a set of D_b non-informative features called F_b. We have $D = D_g + D_b$. A perfect feature selection will keep the D_g informative features and drop the D_b non-informative features. Let us consider a dataset containing N examples drawn from this model. We analyze analytically the behavior of the feature selection performed on this dataset. We express in particular the probability of selecting an informative feature and the stability of the selection.

On our model (Gaussian and independent features), a natural choice for the feature selection method is the t-score, as it precisely assumes Gaussian and independent features. This method computes a score for each feature as follows:

$$\hat{S}c(f_i) = \frac{(\hat{\mu}_{i,C_1} - \hat{\mu}_{i,C_2})^2}{\hat{\sigma}^2_{i,C_1} + \hat{\sigma}^2_{i,C_2}}$$

where $\hat{\mu}_{i,C_k}$ and $\hat{\sigma}_{i,C_k}$ represent the estimated mean and standard deviation of feature f_i in class C_k. The higher the score, the more discriminating the feature is. The selection keeps the d features with the highest scores. $(\hat{\mu}_{i,C_1} - \hat{\mu}_{i,C_2})$ follows a Gaussian distribution $\mathcal{N}(2\mu_i, 2/N)$. The probability distribution of $\hat{S}c(f_i)$ can be expressed in our model using the noncentral χ^2 distribution with one degree of freedom:

$$p_{\hat{S}c(f_i)}(x) = F_i\left(\frac{N}{2}x\right) \;\; where \; F_i \rightsquigarrow \chi^2_1(N\mu_i^2)$$

We can define two distributions: one for the informative features p_g and one for the non-informative features p_b . Let P_g and P_b be their cumulative distributions. From these probability distributions, we can express the probability for an informative feature to be selected. Let $f_g \in F_g$ be an informative feature and let $x = \hat{S}c(f_g)$ be its score, we have:

$$p_{select}(f_g) = \int_0^{+\infty} p_g(x) \cdot p_{select}(f_g|x)dx$$

If d features are selected, the conditional probability for a feature to be selected corresponds to the probability to be ranked in the d first features, i.e. the probability that the score of f_g is higher than the scores of at least $D - d$ features:

$$p_{select}(f_g|x) = p(\#(x > \hat{S}c(f_i)) \geq D - d)$$

This can be expressed as the probability that the score of f_g is higher than the scores of i informative features times the probability that the score of f_g is higher than at least the scores of $D - d - i$ non-informative features:

$$p_{select}(f_g|x) = \sum_{i=0}^{D_g-1} p(\#(x > \hat{S}c(f_i)|f_i \in F_g) = i) \cdot p(\#(x > \hat{S}c(f_i)|f_i \in F_b) \geq D - d - i)$$

All those values follow a binomial distribution, so the probability of f_g to be selected can also be written as:

$$p_{select}(f_g|x) = \sum_{i=0}^{D_g-1} b(i, D_g - 1, P_g(x)) \cdot B(D - d - i, D_b, P_b(x))$$

where b and B are respectively the binomial and cumulative binomial distributions. The probability of a non-informative feature to be selected $p_{select}(f_b)$ can be express in the same way.

We can also express the expected stability CW_{rel} of a system $\mathcal{S}$ contain ing ω feature selections. The expected numbers of occurrences of informative and non-informative features are respectively $p_{select}(f_g)\ \omega$ and $p_{select}(f_b)\ \omega$. When we include those values in the formula of CW_{rel} we obtain the expected stability $E[CW_{rel}(\mathcal{S})]$. Annexe 3 of the thesis provides a detailed calculation of CW_{rel} when the number of selections is large enough.

In Figure 15, we show the probability for an informative feature to be selected (left panel) and the expected stability $E[CW_{rel}(\mathcal{S})]$ (right panel) in function of N and μ*. We fixed D = 1000, D_g = 100, D_b = 900, and d = 100. N varies from 25 to 2000. μ* varies from 0.01 to 0.15, which corresponds to classification problems where the Bayes error goes from to 0.01 to 0.49.

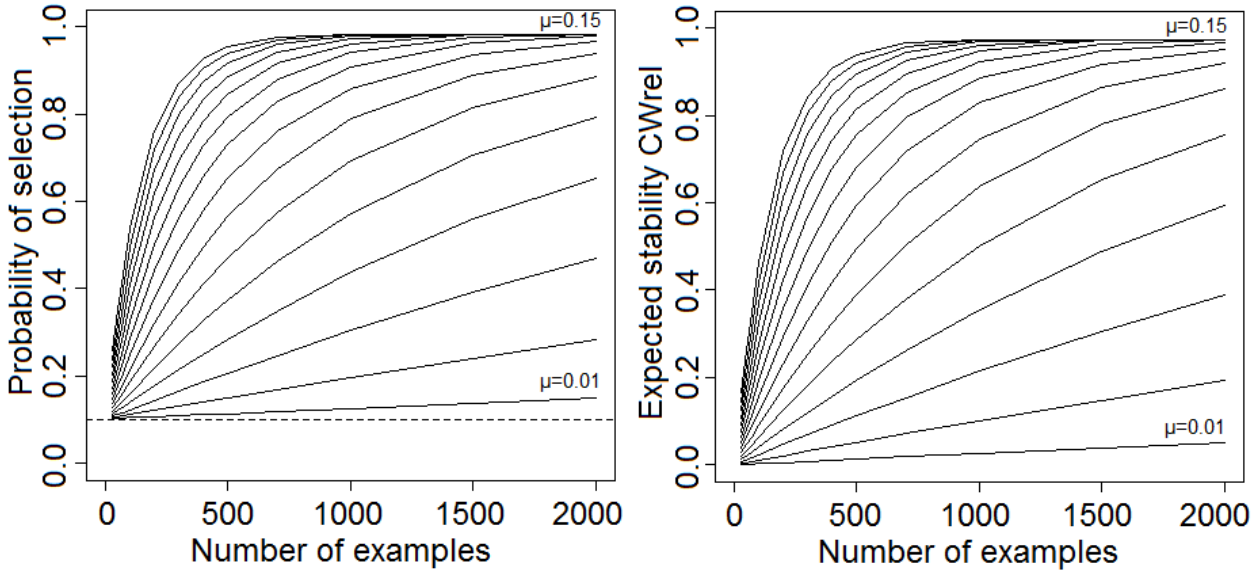

Figure 15: Left panel: Probability for an informative feature to be selected in function of N and μ*. Right panel: Expected stability $E[CW_{rel}(S)]$ in function of N and μ*.

We can see that the probability for informative features to be selected increases substantially with N. However, for the hardest problems (μ*<0.05), even for a sample size of

2000 the probability for informative features to be selected does not reach 1 ($p_{select}(f_g)$ = 0.8 for μ^* = 0.05; 0.15 for μ^* = 0.01). For the easiest problems, $p_{select}(f_g)$ increases rapidly, almost reaching 1 for N = 500 in the case where μ^* = 0.15. Nonetheless, even on those easiest problems, $p_{select}(f_g)$ is low for very small sample sizes: for N = 25,$p_{select}(f_g)$ < 0.3 and for N = 100, $p_{select}(f_g)$ < 0.5. Note that, since we always select the 10% top variables, in the worst cases with the smallest sample size, the selection is random and every variable has an equal probability to be selected, so $p_{select}(f_g)$ tends to 0.1.

The expected stability $E[CW_{rel}(\mathcal{S})]$ follows similar patterns. Notably $E[CW_{rel}(\mathcal{S})]$ is below 0.2 for the smallest sample size (N = 25) and below 0.5 for N = 100, for the easiest problems. Just like the probability for informative features to be selected, on easy problems it increases fast, almost reaching 1 for N = 500, but for hard problems it remains low even for N = 2000. It should be noted that this setting is ideal for our feature selection procedure not only because t-score assumes, as a hypothesis, that features have a Gaussian distribution, which matches our model, but also because we know the number of relevant features and configure the filter accordingly to keep this number of features ($d = D_g$), which optimizes the stability. On datasets where the real number of relevant features is unknown, stability will likely be lower.

2.4 Analysis on artificial data

2.4.1 Generation of artificial data

We used two different artificial data structures. The first data distributions used to generate training and test sets consisted of a variable number ($D \in [50;10000]$) of random, independent features. Training sets consisted of $N \in [25;10000]$ examples, so that the *N/D* ratio goes from 0.0025 to 200, exceeding the range of *N/D* seen in our real datasets. Figure 16 shows the N/D ratios of our artificial data and compares them with the ratios of our real data. We see that the artificial data cover the whole range of the real data.

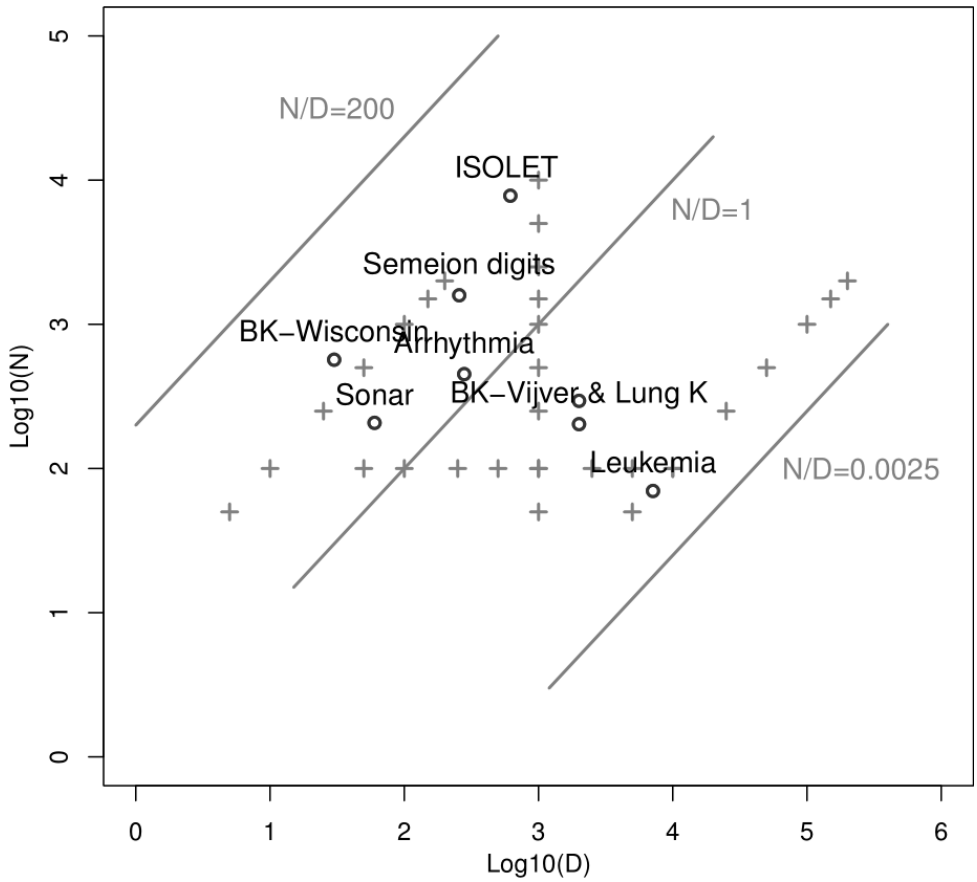

Figure 16: *N/D* ratios of some of our artificial datasets (crosses) and real datasets (circles).

Each of the two classes follows a normal distribution defined respectively by $\mathcal{N}(\mu,\sigma^2)$ and $\mathcal{N}(-\mu,\sigma^2)$, where μ is a vector of means such that $|\mu| = D$ and the standard deviation is $\sigma = 1$ for all features. The elements μ_i of μ were drawn from a triangular distribution with a lower limit and mode equal to 0 (probability density function: $f(x) = 2 - 2x$ for $x \in [0;1]$). To obtain various shapes of strictly decreasing probability densities, simulating varying feature dispersion and relevance, we then raised μ to a power of γ ($\mu_i = \mu_i^{\gamma}$, $\gamma \in [1;10]$). Finally, μ was scaled down so that either $\mathcal{F}$ would yield a specified Bayes error (ε_{Bayes}) or so that the largest μ_i had a specific value $\mu_{i\,max}$. In our experiments we chose $\varepsilon_{Bayes} = 0.10$ or $\mu_{i\,max} = 0.15$.

In order to study some more realistic data, we also generated data with similar characteristics, with a fixed number of variables (D=1000) and with covariance matrix:

$$\Sigma = \begin{vmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{100} \end{vmatrix}$$

where Σ is a block diagonal matrix and Σ_i is a 10 × 10 square matrix with elements 1 along its diagonal and 0.5 off its diagonal, similarly to the method used in (Han & Yu, 2012). We created two kinds of correlated data using this covariance matrix: some data with the same μ as with the uncorrelated data, and some data with $\mu \approx 0.1$ for 100 informative variables and $\mu = 0$ for the 900 non-informative variables.

The score used to rank features on the training data was the absolute value of the t-score. Then the top d features with the highest scores were selected.

Picking the right feature selection threshold is still an open problem. In practice we can either fix a threshold on the relevance score, or on a predetermined number of features. In our stability study, these two approaches are related to the same problem: the impact of the size of the selected features set on the stability. In our simulations, we want to study the impact of each parameter one by one. If we use a threshold on the score, the number of selected features will change with the other parameters, making it more difficult to interpret the results. Thus, we chose to fix the number of selected features in order to compare selections of the same size. We did run some simulations with a threshold on the p-value, they are presented in Figure 24 of supplementary materials.

We focused on the t-score because it assumes independent and normally distributed features, which matches our artificial data, and because it was shown to be as or more stable than other selection methods on small sample microarray data (Haury et al., 2011). In order to assess the influence of the choice of the selection method, we also performed some selections using shrinkage correlation-adjusted t-score (”CAT score”) (Zuber & Strimmer, 2009), mutual information as implemented in R package SlimPLS, and recursive feature elimination based on support vector machine (SVM-RFE) (Guyon et al., 2002).

For various combinations of parameters *N*, *D*, *d* and γ, 100 training sets were generated.

For each of them, feature selection was performed and a linear discriminant analysis (LDA) classifier was trained. Each classifier was then applied to a test set consisting of 10000 samples. Besides the stability measures described in section 2, we measured the average classification error rate and the frequency with which each feature was selected, and computed two Bayes errors: $\varepsilon_{BayesOptimal}$ and $\varepsilon_{BayesObs}$. $\varepsilon_{BayesOptimal}$ is the error rate of the Bayes classifier in the best feature subset of size d. It represents the best classification error rate if we select the true d best features and then build an ideal classifier on them. This value can be used as a measure of the problem difficulty, as well as a base point to evaluate the feature selection and classification. $\varepsilon_{BayesObs}$ is the error rate of the Bayes classifier in a given feature selection. This value can be used as a measure of the feature subset quality. The closer $\varepsilon_{BayesObs}$ is to $\varepsilon_{BayesOptimal}$, the better the selection.

2.4.2 Results on the artificial data

In this set of simulations, we present the performance and stability of the feature selection depending on dataset parameters.

Figure 17 provides an intuitive overview of feature scoring stability in two extreme settings: one with a very small sample (N = 50, left column), the other one with a large sample (N = 5000, right column). In the small sample case, feature scores (Figure 17a)) do not vary much with feature μ_i, and even though the most relevant features have a slightly higher score than the least relevant ones on average, their scores vary approximately on the same range. This contrasts with the large sample case, where the most relevant features have scores in the [8;12] range, far away from the least relevant ones, which stay in the [0;3] range and are thus easy to tell apart. The correlation between feature scores and μ_i decreases with N.

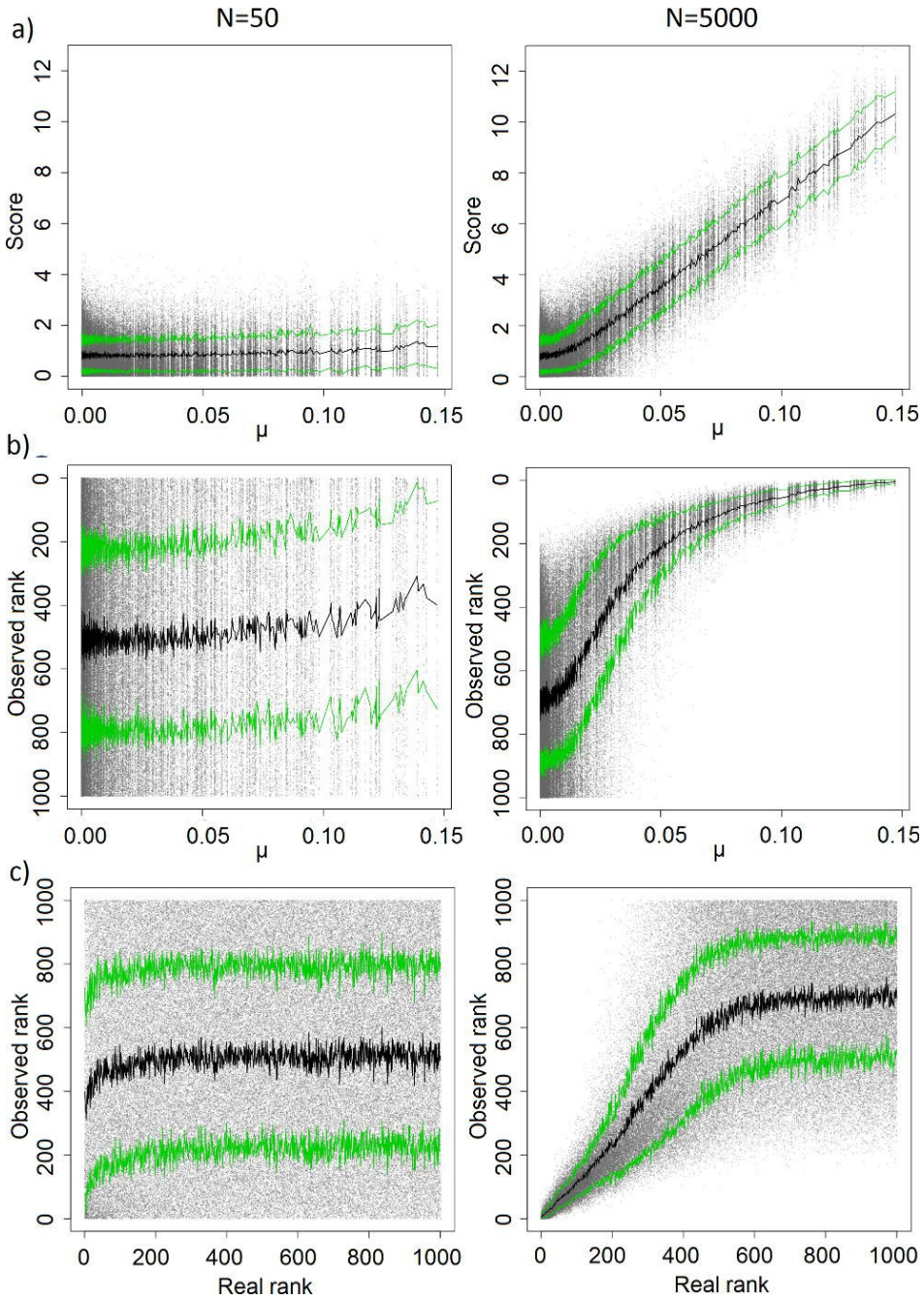

Figure 17: On artificial data (first model, μ based on triangular distribution – for other data see Figures A2.1 and A2.2 in Annexe 2 of the thesis), with D = 1000, d = 100, γ = 2 and N = 50 (left) or N = 5000(right). a) observed score (in absolute value) given real μ_i , b) observed rank given real μ_i ,c) observed rank given real rank. One point per feature and per training set, the blackcurve is the average per feature, the green curves the average ± standard deviation.

The resulting ranks reflect the inconsistency of the scores. Figure 17b) represents observed feature ranks given feature μ_i, Figure 17c) provides a slightly different visualization

(observed feature ranks given true feature ranks). Due to the way our model was conceived, the least relevant features can be considered as noise, even though technically they do have some very tiny relevance (all μ_i are >0, but many of them are only slightly higher than 0). So having the worst features poorly ranked among each other is not a bad result. However, in the small sample case, even the true best features only have a slightly better average rank than the other features. For over 90% of the remaining features, the assigned rank is pure noise, as illustrated by scatter plot and standard deviation lines (green curves) on Figure 17. In the large sample case, the true best features are ranked much more accurately, even though some noise remains among them, and only the worst half of features are assigned a mostly noisy rank. The results show that in small-sample data there is no correlation between the feature score and ranking obtained from the selection methods and the actual quality of the features.

From our simulations, we computed empirically the probability of each feature to be selected. Figure 18 presents the evolution of this probability given μ_i. We can see that in the small sample case (Figure 18a)), the probability for the most relevant features to actually be selected does not reach 35%, while even the least relevant features have a non negligible probability to be selected. In the large sample case (Figure 18b)), the selection is much more accurate: all features with $\mu_i > 0.10$ have a probability to be selected close to 1 and all features with $\mu_i < 0.05$ are almost never selected. Figure 18c) shows the evolution of the regression curve from $N = 25$ to $N = 10000$: as the sample size increases, the logistic shape increasingly stands out, illustrating how the selection progressively becomes more accurate. But only when the sample size reaches around 1000 observations is the feature selection algorithm able to select the most relevant features with a good sensitivity. In small sample data, the probability to reliably select good features is therefore very low.

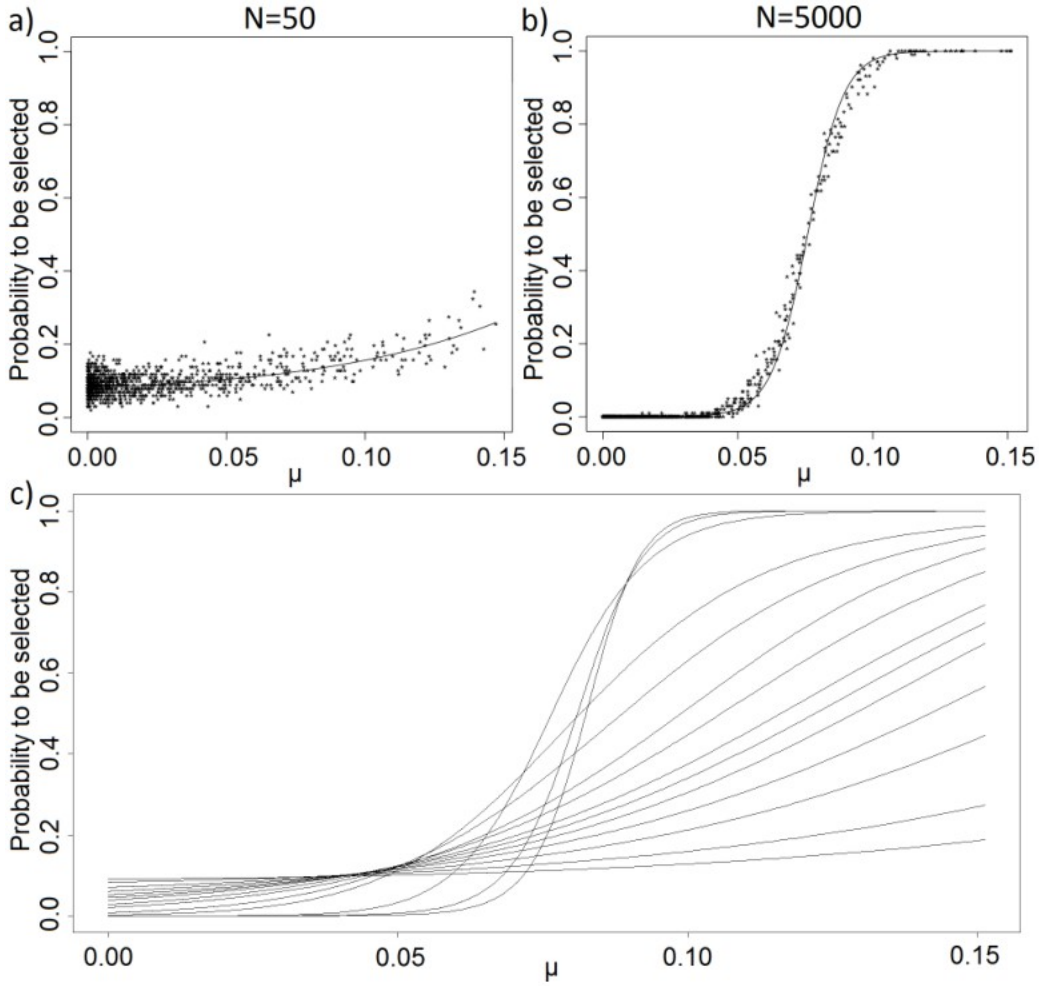

Figure 18: Observed probability for a feature to be selected given its real μ_i. On artificialdata, with $D = 1000$, $d = 100$ and $\gamma = 2$. a) $N = 50$ b) $N = 5000$, one point per feature and the curve was obtained via logistic regression. c) N varying from 25 (curve with thelowest value at $\mu_i = 0.15$) to 10000 (first curve to reach 1). As the sample size grows, thelogistic shape increasingly stands out.

Figure 19 presents results obtained on similar simulations but with other filters: CAT score and mutual information. Mutual information performed a bit worse than t-score and CAT score. SVM-RFE (results in Figure 25 of supplementary materials) performed worse than mutual information and was too heavy to compute on $N \geq 5000$. Still, the results all have a similar evolution with sample size.

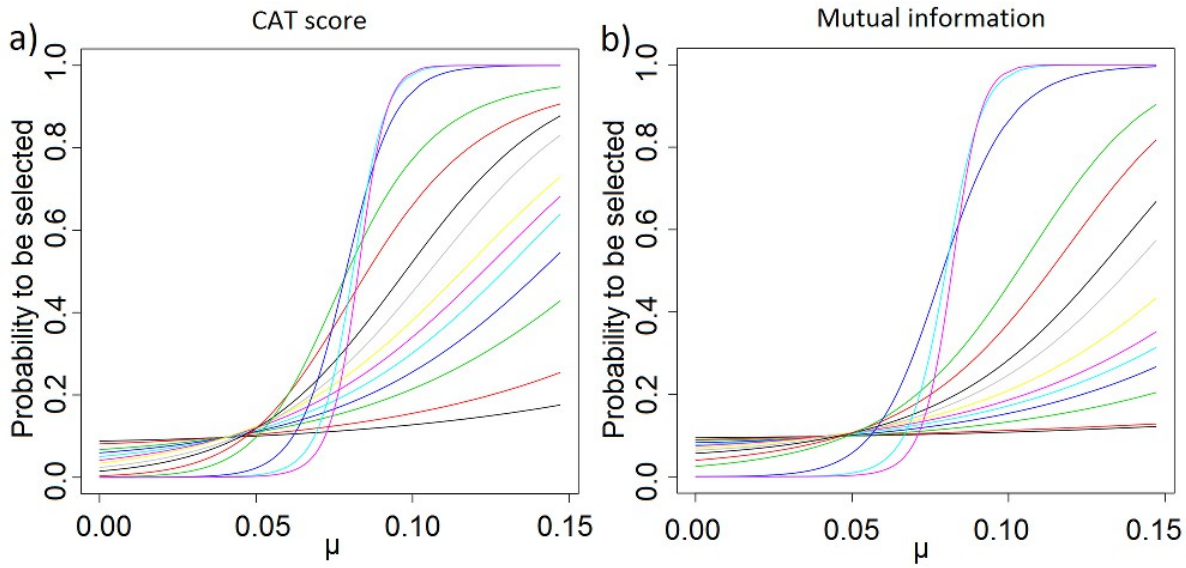

Figure 19: Observed probability for a feature to be selected given its real μ i . On artificialdata, with D = 1000, d = 100, γ = 2 and N varying from 25 (curve with the lowest valueat μ i = 0.15) to 10000 (first curve to reach 1). a) CAT score filter. b) Mutual informationfilter.

Figure 20 presents the evolution of stability measures under varying dataset parameters. The stability is much influenced by the sample size N, with stability measures close to zero when the sample size is around 100 and increasing a lot when additional samples are added to the training set, up to 0.6+ for ATI_{PA} and almost 0.9 for $\overline{S}_W$. It is also much influenced by the total number of variables D, with fairly high values (0.4 to 0.6) when the dataset only contains 100 samples and 50 variables, quickly reaching close to zero with so few as 1000 variables.

To a lesser extent, stability is also influenced by the selection threshold d. In this case, CW_{rel} is minimal when we select very few variables, then it increases to reach a maximum when we select around 150-180 variables, finally it slowly but regularly decreases as we add more unreliable variables. The shape of this curve illustrates the difficulty to reliably identify even the most relevant variables: trying to keep just the 2 best features will yield highly unstable results, while trying to keep 50 best features will much more likely include maybe the 5 or 10 best features with a very high reliability, leading to a higher stability even though the rest of the selection is not as stable. Note that in this setting, obviously $\overline{S}_W$ and $\overline{S}_R$ do not vary, as they do not take into account the fact that a feature was selected or not.

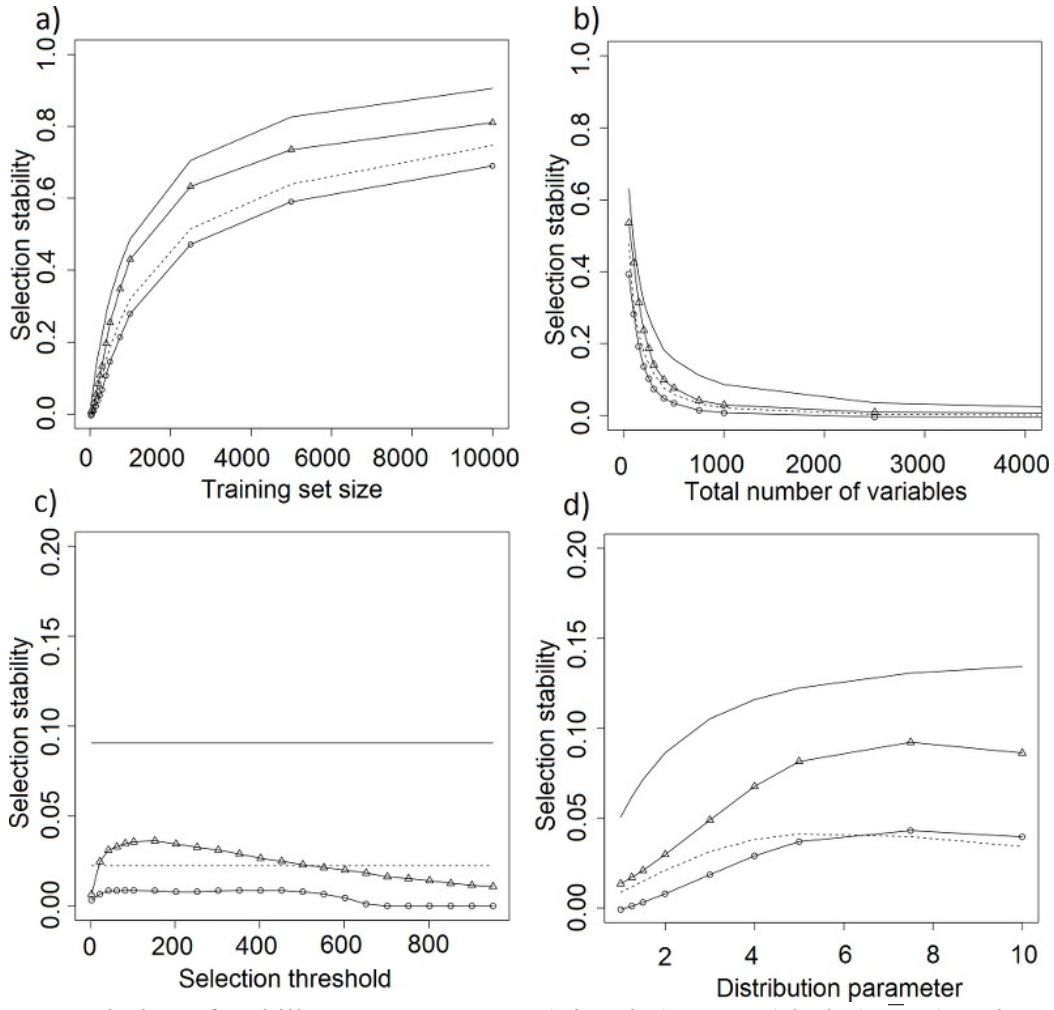

Figure 20: Evolution of stability measures CW_{rel} (triangles), ATI_{PA} (circles), $\overline{S}_W$ (continuous line) and $\overline{S}_R$ (dashes) given: a) $N \in [25;10000]$ b) $D \in [50;10000]$ c) $d \in [2;1000]$ and d) $\gamma \in [1;10]$. When the were not the were not the one being iterated on, parameter values were: $N = 100$, $D = 1000$, $d = D \cdot 10\%$, $\gamma = 2$.

Variable distribution γ also has some influence on stability: stability measures are minimal when variables are distributed on the triangular distribution, and increase with γ, but following different patterns. $\overline{S}_W$ always increases: this measure is not penalized by ranking difficulties or by instability in the final selection, and only benefits from variables taking extreme values: variables with initial μ_i close to zero do not really lose much score correlation when they get squished even closer to zero, while variables with higher μ_i do benefit from getting more isolated farther away from zero. $\overline{S}_R$ increases at first, but then starts decreasing after reaching a maximum at $\gamma = 5$: this measure first benefits from the increased dispersion of variables with high or intermediate μ_i, but at some point this effect is overcome by the increased difficulty to

rank variables with intermediate μ_i (because we kept a constant, realistic Bayes error, the more we streched the distribution the harder intermediately relevant variables became to identify), which eventually get too close to zero. CW_{rel} and ATI_{PA}, which perform their selection based on a cutoff in the ranks, evolve as a consequence of $\overline{S}_R$, only their decrease is somewhat delayed because they are only affected by the top d rankings. It is likely that a subset-size optimizing feature selection method would see a higher influence of data distribution over selection stability, because it would probably drop the decreasingly relevant variables while keeping the ones increasingly easier to identify.

Figure 21 presents the evolution of stability measures with sample size using t-score and CAT score filters on non correlated and correlated datasets. Non correlated and correlated data (or the two kinds of correlated data) cannot really be formally compared since a Bayes error cannot be computed on correlated data, and introducing the correlations modifies the distribution of μ, but we can still observe that correlation does not fundamentally change the stability nor its evolution with sample size. CAT score performed worse than t-score on correlated data, even on large samples. This could be explained by the fact that this method tries not to select correlated variables, so it tends to keep fewer variables per block even in the most significant blocks.

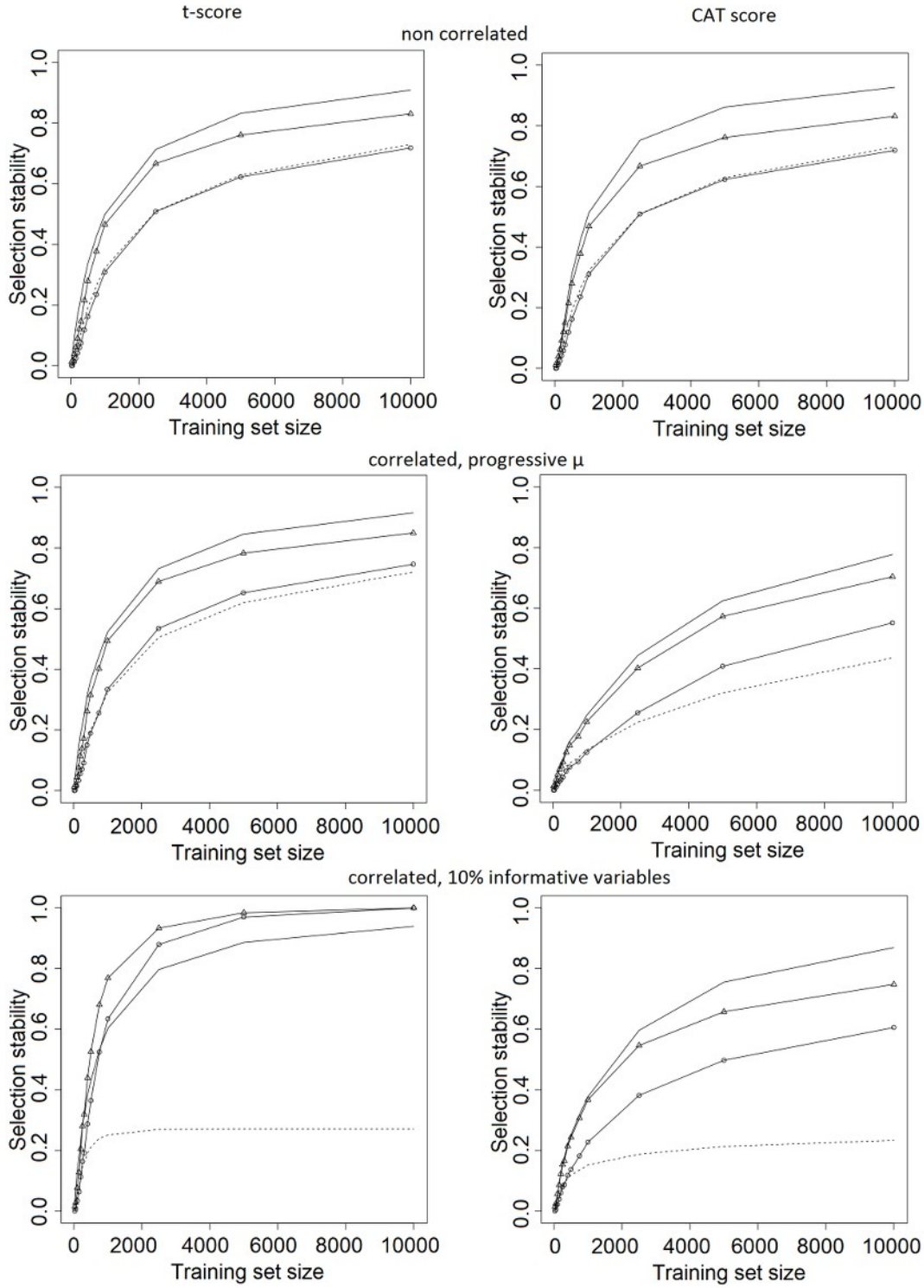

Figure 21: Evolution of stability measures CW_{rel} (triangles), ATI_{PA} (circles), $\overline{S}_W$ (continuous line) and $\overline{S}_R$ (dashes) when using t-score (left) and CAT score (right) filter, on non correlated data (top), correlated data with progressive μ (middle), and correlateddata with 100 informative variables vs 900 non-informative variables (bottom). $N = 100$, $D = 1000$, $d = 100$, $\gamma = 2$.

Figure 22 reports the classification error rates obtained from the same selections as in Figure 20. The dashed curves indicate $\varepsilon_{BayesOptimal}$, the best possible classification error rate on the dataset with d features (selecting the true d best features then building an ideal classifier on it), the grey curves indicate $\varepsilon_{BayesObs}$, the best possible classification error rate if building an ideal classifier on the selected features, the black curve indicate the observed error rate. The error rate and Bayes error on the selected features increase when the training sample size decreases. For small sample, this increase is very strong and we see large differences between $\varepsilon_{BayesOptimal}$ and $\varepsilon_{BayesObs}$, meaning that the feature selections have bad performance.

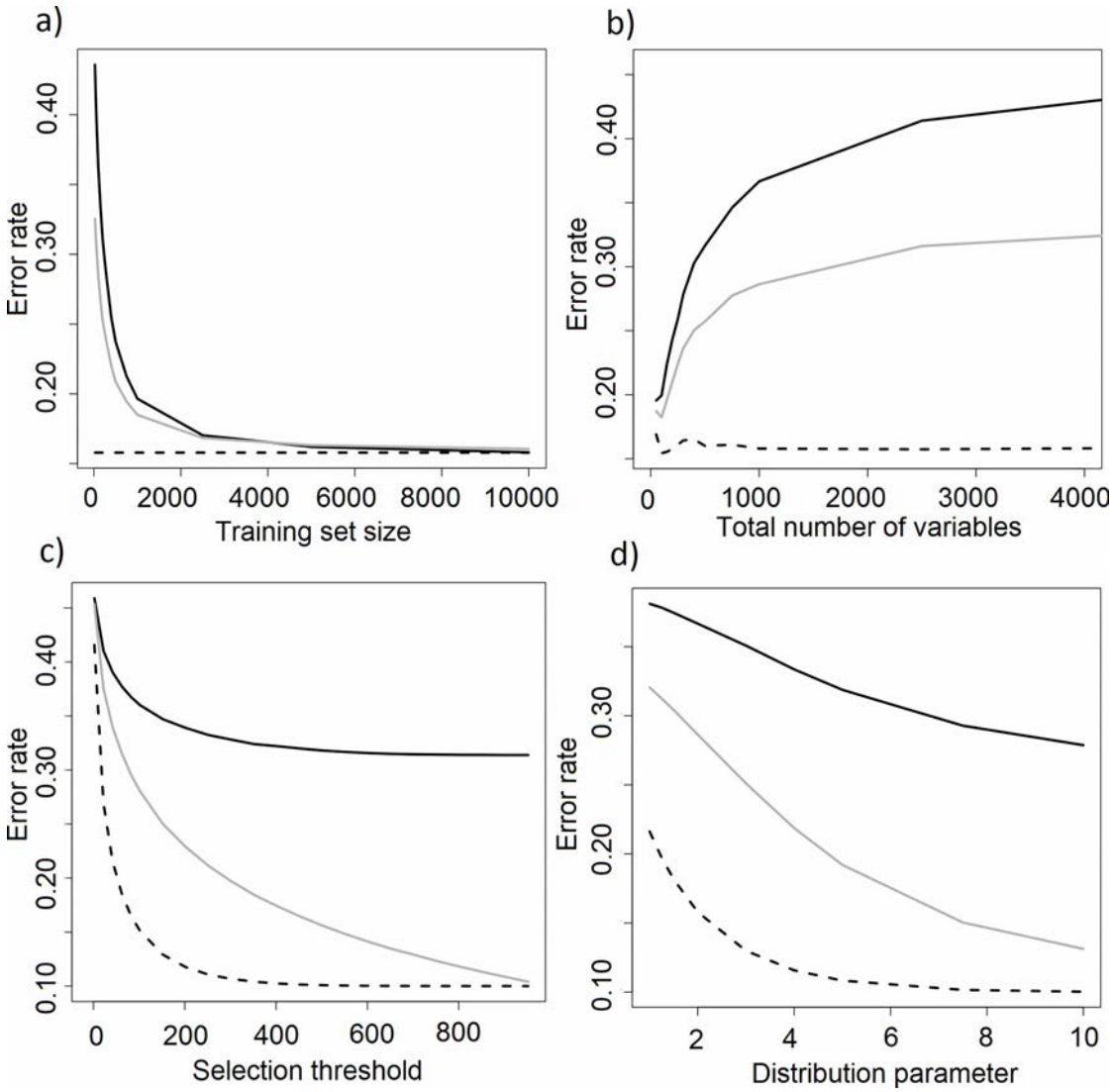

Figure 22: Evolution of error rate (black), $\varepsilon_{BayesObs}$ (grey) and $\varepsilon_{BayesOptimal}$ (dashes) given: a) $N \in [25;10000]$ b) $D \in [50;10000]$ c) $d \in [2;1000]$ and d) $\gamma \in [1;10]$. Whenthe were not the one being iterated on, parameter values were: $N = 100$, $D = 1000$, $d = D \cdot 10\%$, $\gamma = 2$.

The error rate is also influenced by an increase in the total number of variables (Figure 22b)): on the small tested sample ($N = 100$), with only 50 variables the classification error rate is lower than 20% (for an optimal Bayes error of over 16%), but when we reach 2500 variables the error rate is over 40%, with a Bayes error on the selected features over 30%. This is a consequence of the dilution of the information: since we kept a constant $\varepsilon_{BayesOptimal}$ (to account for the fact that real datasets do not necessarily contain all the variables needed to do a perfect prediction even if we had an unlimited supply of samples), when the dimension increases, the information gets spread over weak features, so $\varepsilon_{BayesObs}$ becomes higher. These weak features do not contain a lot of information, so the classification error rate increases.

The evolution of the error rate with the selection threshold d (Figure 22c)) might seem a little more surprising. Particularly, the regular decrease of the Bayes error on the selection may give the impression that, as we increase the threshold, the feature selection keeps including relevant variables in the proper order. This, of course, is not the case: as we loosen the threshold for inclusion, we include more and more slightly relevant variables, more and more randomly (as can be deducted from the stability measures seen previously). Even for small thresholds, the selection does not contains necessarily the most relevant variables. This point is illustrated by the rapid growth of the distance between the $\varepsilon_{BayesOptimal}$ and the $\varepsilon_{BayesObs}$ curves. When the $\varepsilon_{BayesObs}$ curve finally reaches the $\varepsilon_{BayesOptimal}$ curve, it is not because the selection is good but because all variables are included. That is why the classification error does not decrease much when $d > 200$ (computation time, though, increases substantially): the information contained in the most relevant variables is flooded by the poorly relevant but selected variables. Similar results were obtained when increasing the sample size to $N = 1000$, only with lower error rates and higher stability values (results not shown).

Distribution parameter γ seems to have a higher influence on error rate than on stability. The higher γ, the lower the error rate (Figure 22d)): as γ increases, the number of highly relevant features decreases but their discrimination power increases and allow for a better selection (this is diluted by the increased instability on the other features when we observe stability measures, but this stands out when we observe how $\varepsilon_{BayesObs}$ gets closer and closer to $\varepsilon_{BayesOptimal}$) and a better classification accuracy. However, even though the classification error rate does improve, it does so slower than $\varepsilon_{BayesObs}$. An explanation to this difference is that, despite an improvement in the

most relevant variables, the classifier is still hampered by the remaining not-so-relevant variables.

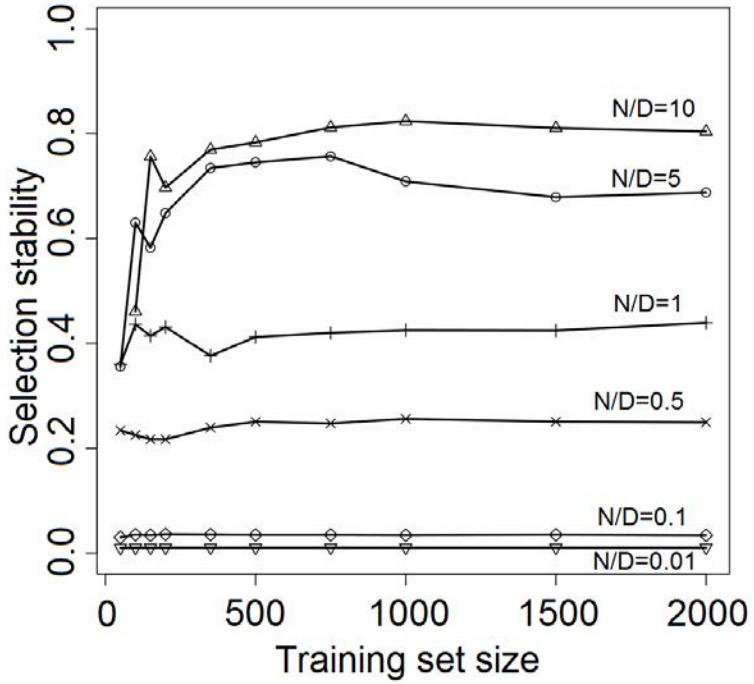

Figure 23: Evolution of CW_{rel} with the number of training examples for a constant *N/D* ratio. The different curves correspond to different *N/D* ratios (lowest: 0.01; highest: 10).

Figure 23 shows the stability CW_{rel} as a function of the number of training examples for a constant N/D ratio. The different curves correspond to different *N/D* ratios (from 0.01 to 10). We see the stability is constant for a fixed *N/D* ratio, except for some variations in the lowest dimension values for $N/D \geq 5$, caused by random variations in the problem difficulty (very few selected variables on those specific points). For small-sample problem, where $N/D << 1$, the stability depends on the *N/D* ratio, not on *N* or *D* alone. In gene expression data, the *N/D* ratio typically ranges from 0.001 to 0.1, which leads to a maximum stability of 0.2 in our simulations. Note that those simulations are based on Gaussian, uncorrelated features, which is one of the easiest classification problems. Moreover we use a selection based on the t-score, which assumes features are Gaussian, hence matches our model and should perform well on it. In real data, the distribution of the classes is much more complex than Gaussian, and we do not have a feature selection method that knows the real feature distributions. So, the stability on real data should be lower than the stability on artificial data: the values reported on Figure 23 could be

considered as upper bounds. Results on correlated data are presented in Annexe 2 of the thesis.

2.5 Analysis on real data

2.5.1 Description of the real data

We experimented with eight publicly available datasets, presented in Table 2. When the original dataset dealt with a multiclass classification task, we reduced the problem to a 2-class, 1-vs-all classification. For each datasets, for different values of N, 200 training sets were generated by randomly drawing samples from the dataset (without replacement). For each of them, feature selection was performed and a classifier was trained (using the same methods as with the artificial data). Each classifier was then applied to a test set consisting of the samples not included in the corresponding training set. We measured the stability of the feature selection across the training sets and the average classification error rate.

Table 2: Characteristics of the real datasets.

Name	N	D	N/D	Source
Leukemia	72	7129	0.01	(Golub et al., 1999)
Lung cancer	203	2000	0.10	(Bhattacharjee et al., 2001)
Breast cancer Vijver	295	2000	0.15	(van de Vijver et al., 2002)
Arrhythmia	452	279	1.6	UCI repository (Frank and Asuncion, 2010)
Sonar	208	60	3.5	UCI repository
Semeion digits	1593	256	6.2	UCI repository
ISOLET	7797	617	12.6	UCI repository
Breast cancer Wisconsin	569	32	17.7	UCI repository

2.5.2 Result on the real data

Table 3 presents stability measures and error rates observed on real datasets for various training set sizes. Because of varying underlying problem difficulties, differences in absolute values between datasets cannot be attributed only to differences in dimensions. Nonetheless, those results provide an overview of the kind of values which can be expected, and of their dependency on sample size. Globally, error rates decrease when N increases, with a faster evolution with N when N/D is low, which is quite similar to what we observed on our artificial data.

Table 3: Classification error rate and selection stability on the real datasets.

	Breast cancer Vijver ($D = 2000$)		Lung cancer ($D = 2000$)		Leukemia ($D = 7129$)	
	$N = 50$ $N/D = 0.025$	$N = 100$ $N/D = 0.05$	$N = 50$ $N/D = 0.025$	$N = 100$ $N/D = 0.05$	$N = 20$ $N/D = 0.003$	$N = 35$ $N/D = 0.005$
Error rate (%)	38.2	37.4	8.0	6.1	9.2	4.0
CW_{rel}	0.20	0.26	0.46	0.51	0.24	0.30
ATI_{PA}	0.06	0.10	0.26	0.30	0.13	0.18
$\overline{S_R}$	0.09	0.14	0.51	0.58	0.22	0.26
$\overline{S_W}$	0.33	0.41	0.81	0.85	0.55	0.60
	Breast cancer Wisconsin ($D = 30$)		Sonar ($D = 60$)		Semeion digits ($D = 256$)	
	$N = 50$ $N/D = 1.67$	$N = 284$ $N/D = 9.48$	$N = 50$ $N/D = 0.83$	$N = 100$ $N/D = 1.67$	$N = 100$ $N/D = 0.39$	$N = 796$ $N/D = 3.11$
Error rate (%)	9.5	9.0	32.5	31.0	4.0	3.6
CW_{rel}	0.75	0.97	0.28	0.30	0.54	0.65
ATI_{PA}	0.62	0.95	0.14	0.15	0.34	0.46
$\overline{S_R}$	0.91	0.97	0.51	0.59	0.73	0.92
$\overline{S_W}$	0.93	0.97	0.68	0.74	–	0.96
	ISOLET ($D = 618$)			Arrhythmia ($D = 279$)		
	$N = 100$ $N/D = 0.16$	$N = 618$ $N/D = 1$	$N = 3900$ $N/D = 6.31$	$N = 50$ $N/D = 0.18$	$N = 100$ $N/D = 0.36$	$N = 226$ $N/D = 0.81$
Error rate (%)	4.3	2.7	2.4	34.0	30.3	27.5
CW_{rel}	0.32	0.71	0.86	0.34	0.41	0.48
ATI_{PA}	0.15	0.53	0.75	0.16	0.22	0.28
$\overline{S_R}$	0.42	0.81	0.93	0.42	0.46	0.51

Stability measures, although different in absolute value, share a same trend, opposite to the error rate. Stability increases with training set size, but globally it remains rather low as long as $N/D < 1$, although higher than on our artificial data with similar dimensions. Notably, the Vijver breast cancer dataset has a higher stability but a similar error rate compared to our non-correlated artificial dataset. This discrepancy might be caused by the different correlation structure in our artificial datasets compared to those in the real datasets: in our second set of correlated artificial data, we observed a higher stability with a higher error rate. With an easier problem, we would have a higher stability with a lower error rate. So by adjusting both the problem difficulty (μ distribution) and the amount of correlation, it might be possible to obtain results which get closer to the Vijver dataset, as well as the others. Another explanation could be the differences in the experimental design between artificial and real data. In artificial data, the stability is computed from a set of 100 independent datasets. In real data, we have only one dataset that is split in two subsets, this process is repeated 200 times. The 200 splits are not independent, there is therefore a bias that artificially increases the measured stability.

Nonetheless, in a similar way to the error rate, stability increases with N, with a faster

evolution with N when N/D is low. And stability values are the lowest when the N/D ratio is the lowest. Performing a simple regression on the data presented in Table 3 shows an independent correlation between, on one hand, stability (CW_{rel}) and problem difficulty (measured as the smallest error rate obtained on the dataset), and on the other hand, stability and the N/D ratio.

2.6 Discussion and Conclusion

In this paper, we analyzed the performance of feature selection and especially its stability in small-sample high-dimension settings. We used existing measures of stability and we introduced ATI_{PA}, a modification of the ATI stability measure adjusted to avoid a bias on the number of selected features. We investigated the behavior of stability depending on the number of examples, features, selected features and distribution of the discrimination power of features. We show that in small sample problems the probability to select the best features is very low even with a feature selection method which works based on accurate assumptions on the features' distributions. We show empirically that for Gaussian data, the stability depends on the N/D ratio. The results of our simulations on artificial data show that the stability is dramatically low (almost 0) for $N/D \leq 0.1$, but then raises quite rapidly as N/D reaches ~5. It could be interesting to study this further, to see if a phase transition, which is frequent in machine learning (Saitta et al., 2011), can be identified on both artificial and real data. In real data, where variable distributions are unknown but certainly more complex than Gaussian models, the used feature selection methods are not perfectly adapted, so we cannot expect better results. It is even highly likely that the situation is worse, so we can consider our simulation results as an upper bound of stability for real data in function of their size (N and D). This leads to the conclusion that for any high-dimension small-sample data, it is not possible to obtain a stable feature selection for a classification task. Although we mainly explored the use of the t-test filter with a fixed threshold on the number of selected variable, this conclusion also holds for the other selection methods we tested, and the method of choice of the feature selection threshold (a predetermined number of variables versus a threshold on the relevance score) did not make much difference either. These conclusions could explain a lot of results published in various domains. For example, in medicine, several gene expression signatures of a given cancer have been identified on different microarray datasets, but there is almost no overlap between

signatures when the results on a given dataset are verified on the other datasets (Miecznikowski et al., 2010).

To improve the stability of feature selection, the first option is "simply" to increase the number of examples in the datasets. While research projects are necessarily limited in that respect, it seems hopeless to try to construct a stable classifier based only on a few tens of examples. A second way would be to find reliable methods to reduce the dimensionality of the data prior to applying usual filters. For instance, a priori knowledge and unsupervised methods could be used to filter out some of the irrelevant variables. It could also be interesting to exploit the redundancy of the feature to compress the dimensions of the data. As future work, it would be interesting to perform similar tests with more complex data and on other feature selection methods, particularly on those specifically designed to improve stability, such as Consensus Group Stable feature selection or Complementary Pairs Stability Selection (Shah & Samworth, 2011).

Supplementary materials

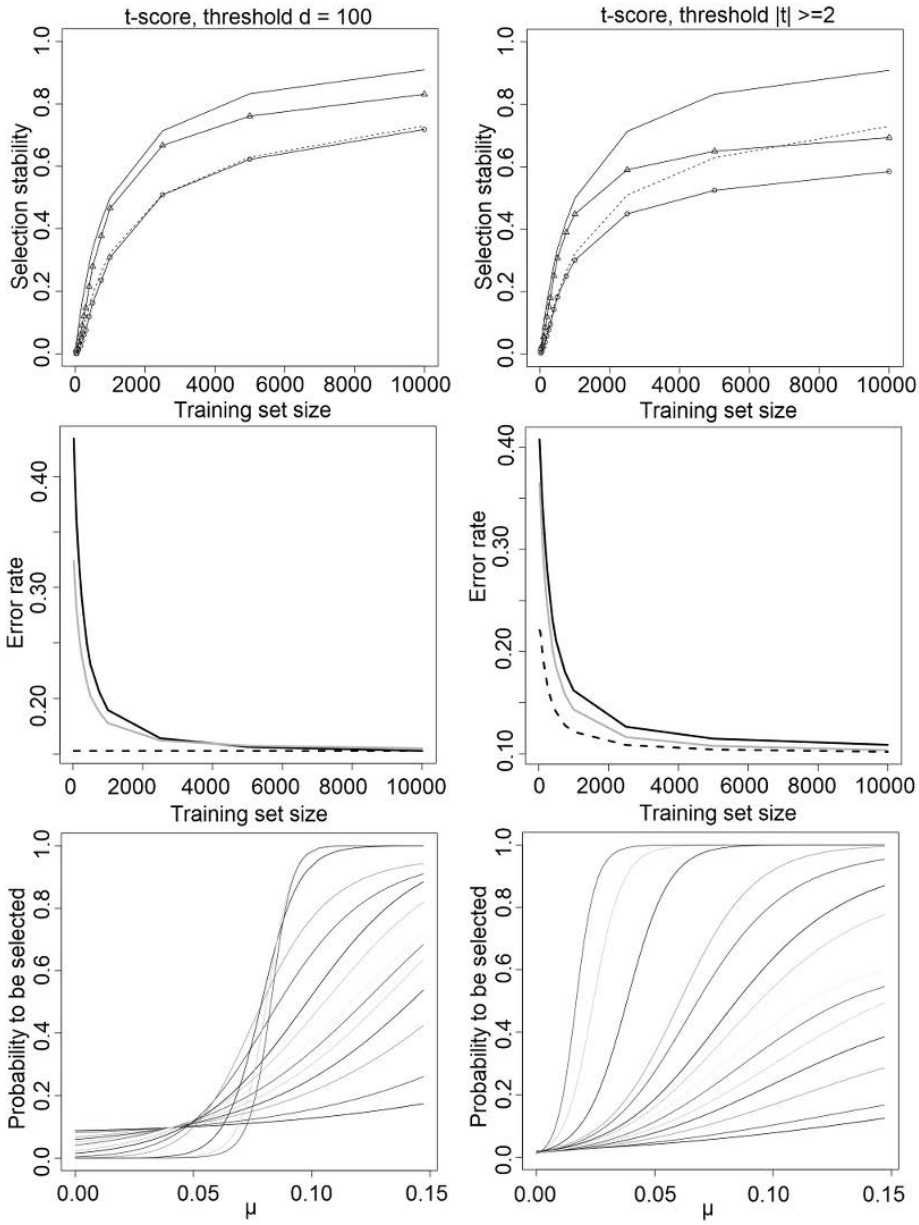

Figure 24: Feature selection using t-test with a threshold based on a number of variables (d=100, left column), versus a threshold on the p-value ($p<0.05$, right column), on non-correlated artificial data. Top row: stability given sample size; middle row: error rate given sample size; bottom row: probability for a feature to be selected given its real μ.

Compared to the threshold on the number of variables, the threshold on the p-value leads to a slight improvement in stability and error rate for small samples, a degradation of stability and an improvement on error rate in large samples (due to a large increase in the number of selected variables: around 400 variables when N=10000 versus around 40 variables when N=25).

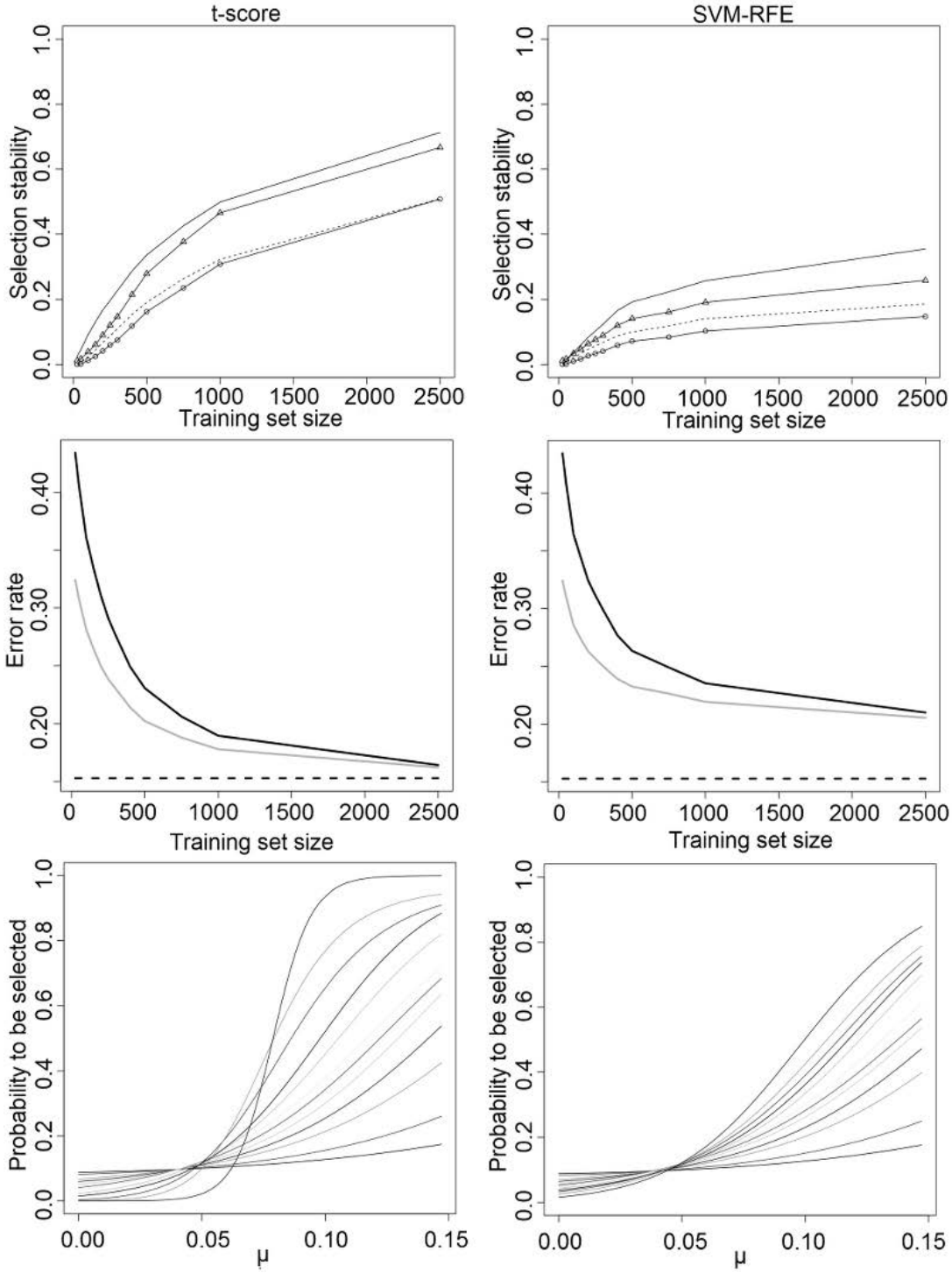

Figure 25: Feature selection using t-test (left column), versus SVM-RFE (right column), on non-correlated artificial data. Top row: stability given sample size; middle row: error rate given sample size; bottom row: probability for a feature to be selected given its real μ. The highest sample size is only 2500 because of the slow speed of SVM-RFE when N becomes too high. SVM-RFE resulted in lower stability and higher error rate than t-test on those data.

Chapitre 3 :
Présélection par regroupements fonctionnels :
Correlation-Gene Ontology (CoGO)

Dans le chapitre précédent, nous avons vu que, si une multitude de facteurs peuvent influencer la stabilité de la sélection de variables, l'un des plus importants est le ratio *N/D* du nombre d'observations sur le nombre de variables. Malheureusement, en pratique, le nombre d'observations est limité par les conditions expérimentales (capacités de recrutement, coût des produits et techniques utilisées, contraintes de temps), et les dizaines ou quelques centaines de patients constituant l'échantillon sont, le plus souvent, la taille d'échantillon maximale possible.

Pour augmenter *N*, certaines études se sont intéressées à la combinaison de plusieurs jeux de données. Une première méthode, le *transfer learning* (Helleputte & Dupont, 2009), utilise des données d'expériences précédentes, relevant de tâches de classification similaires ou très proches du nouveau problème à traiter, pour augmenter le score d'un petit ensemble de gènes observés comme pertinents sur ces données antérieures. D'autres combinent directement les données d'autres expériences avec les données à analyser, avec toutes les problématiques de normalisation qui en découlent, afin de rendre comparables et combinables les observations issues de plusieurs expériences (Shen et al., 2004 ; Warnat et al., 2005 ; Xu et al., 2005). Ce type de technique pose des questions vis-à-vis de la mesure de la stabilité : si la normalisation est effectuée sur l'ensemble des données, en amont de la validation croisée, la stabilité (et la précision de classification) mesurée pourrait être surestimée, ce qui est un problème fréquemment observé dans l'étude des méthodes de pondération (Yu et al., 2012, Wang et al., 2013). Si la normalisation est effectuée à l'intérieur de chaque run de validation croisée, stabilité et précision risquent au contraire d'être sous-estimées. Par ailleurs, et surtout, ces méthodes nécessitent, par définition, au moins 2 jeux de données : elles ne sont pas applicables à l'étude d'un nouveau problème sur lequel on ne dispose pas d'autres données d'expériences très proches à intégrer.

Pour cette dernière raison, nous avons choisi ici d'aborder l'autre côté du ratio : la diminution de *D*. Nous proposons dans ce chapitre d'exploiter des données a priori afin de réaliser une présélection, en amont d'une méthode de sélection de variables supervisée habituelle. Cependant, il semble hasardeux de réaliser une présélection basée uniquement sur les données a priori : une telle méthode risquerait de restreindre l'extraction de connaissances à des connaissances... déjà connues. La méthode que nous avons développée, CoGO pour *Correlation-Gene Ontology*, a donc retenu une approche hybride, combinant à la fois des données *a priori*, tirée de Gene Ontology (Ashburner et al., 2000), et les données d'expression utilisées de manière non supervisée.

Plusieurs méthodes d'enrichissement à partir de bases de connaissances pour la sélection de variables ont déjà été décrites (He & Yu, 2010). Par exemple, une première méthode, la sélection "partiellement supervisée" (Helleputte & Dupont, 2009b), propose d'utiliser des connaissances *a priori* sur la pertinence des gènes pour la tâche de classification en cours, en boostant le scoring des gènes connus *a priori* comme pertinents, de la même manière dans le *transfer learning*. Cette approche a permis une amélioration de la stabilité et, dans une moindre mesure, des performances de classification sur la plupart des données testées, mais son prérequis d'avoir une connaissance *a priori* des gènes pertinents pour la tâche de classification en cours ne la rend pas utilisable dans les cas où l'on ne dispose pas de connaissances a priori aussi précises. Une autre approche (Cun & Fröhlich, 2012) a exploité un réseau d'interactions protéines-protéines construit à partir de KEGG (Kanehisa & Goto, 2000) et de la base de données Pathway Commons (Cerami et al., 2011) avec plusieurs méthodes intégrant un réseau à la sélection de variables, telles que les network-based SVMs (Zhu et al., 2009). Les résultats ont été mitigés, avec une amélioration de l'interprétabilité de la sélection mais pas de gain en terme de précision. Par exemple, les networks-based SVMs ont permis des gains de stabilité mais au prix d'une importante dégradation de la précision.

Une autre étude (Staiger et al., 2012) a comparé différentes méthodes de sélection, basées sur la création de variables composites d'après des données *a priori* puis sur l'application de méthodes de sélections sur ce jeu de variables composites en nombre réduit, sur 6 jeux de données puces, et conclu à l'absence d'amélioration de la précision ou de la stabilité par rapport aux méthodes de sélection simples usuelles. Cette étude remarque en particulier que les

performances des méthodes intégrant des données *a priori* sont inchangées lorsqu'on randomise les données *a priori*, et que l'amélioration de la stabilité qui a pu être observée dans les publications d'origine disparaît si on corrige la mesure de stabilité par le nombre de variables.

Figure 26: Utilisation de données a priori (ici un réseau d'interaction protéine-protéine) pour réduire la dimension par création de méta-gènes. Tiré de (He & Yu, 2010).

Si on résume ces approches, les premières exploitent directement données d'expression et données *a priori* en une seule étape de sélection, et les secondes réalisent d'abord une étape de construction de variables composites (ou méta-gènes) (Figure 26), puis une étape de sélection avec une méthode habituelle appliquée aux méta-gènes. Avec CoGO, nous proposons une approche, à notre connaissance non exploitée pour l'instant, en deux étapes avec une première étape de présélection de gènes, utilisant à la fois des données a priori (les annotations GO des gènes) et les données observées (les données puces à analyser), produisant un sous-ensemble de gènes non modifiés (par opposition aux méta-gènes de certaines méthodes), et une seconde étape de sélection, utilisant des méthodes de sélection habituelles sur ce sous-ensemble de

gènes.

3.1 CoGO, une méthode de pré-sélection de gène combinant données a priori et données observées

3.1.1 Gene Ontology

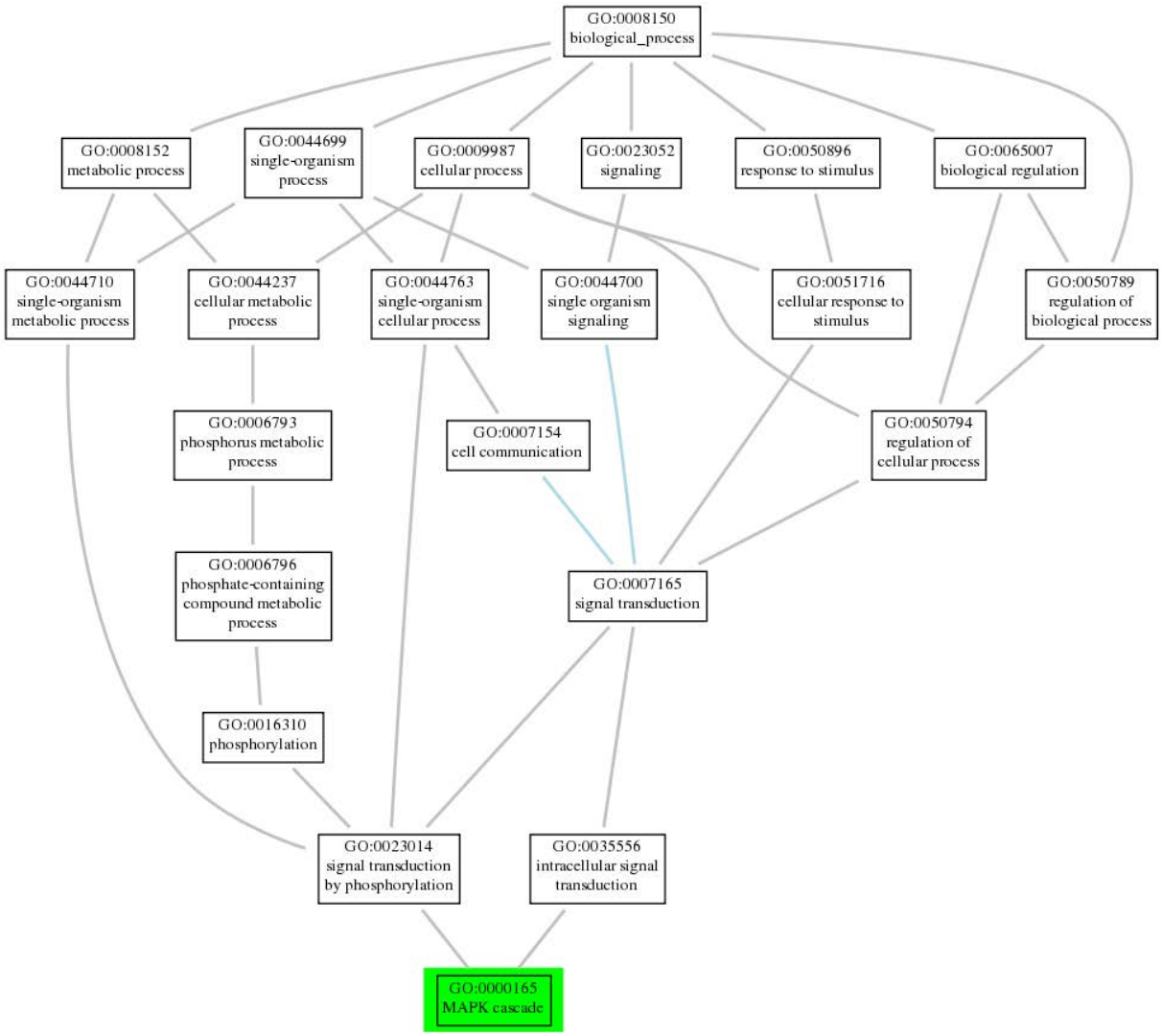

Figure 27: Graphe orienté acyclique des processus biologiques parents de l'annotation GO 0000165 "MAPK cascade". Obtenu via le package R RamiGO (Schröder et al., 2013).

Gene Ontology (GO) est une ontologie qui a pour but de structurer la description des gènes et des produits géniques. Elle est structurée en un graphe orienté acyclique et divisée en trois domaines :

- composants cellulaires (*cellular components*, ~3600 termes)
- fonctions moléculaires (*molecular functions*, ~10600 termes)
- processus biologiques (*biological process*, ~26600 termes)

Les relations utilisées sont globalement simples pour une ontologie (*is a*, *part of*, *has part*, *regulates*, (Gene Ontology, 2014)), n'en faisant pas (pour l'instant) une ontologie appropriée à des fins de raisonnement automatique. Cependant, ces relations sont plus que suffisantes pour l'usage qui nous intéresse, qui consiste à observer la proximité entre les termes (Figure 27), et GO est généralement considérée comme la plus importante ressource d'annotations génomiques, avec plus de 40000 termes définis en juin 2014.

Les annotations des différents gènes peuvent être obtenues automatiquement via différents logiciels. Nous avons utilisé le package R GOsim et sa fonction getGOInfo() (Fröhlich et al., 2007).

3.1.2 Mesure de la similarité des fonctions et des gènes dans GO

Afin de regrouper les gènes, nous avons besoin d'une mesure de leur similarité fonctionnelle. Nous avons utilisé pour cela la mesure de *relevance similarity* décrite dans (Schlicker et al., 2006). Cette mesure est elle-même basée sur une combinaison de la mesure de similarité de Lin (Lin, 1998) et de la mesure de contenu d'information de Resnik (Resnik, 1995), utilisée comme pondération.

3.1.2.1 Probabilité d'un terme GO

La *relevance similarity* considère que la probabilité d'un terme GO est proportionnelle à la fréquence avec laquelle il apparaît dans les annotations des gènes, soit en tant qu'annotation directe, soit en tant qu'ancêtre d'une annotation. La fréquence du terme T peut alors s'écrire :

$$freq(T) = anno(T) + \sum_{D \in descendants(T)} freq(D)$$

où *anno(T)* est le nombre de gènes annotés par le terme T et *descendants(T)* est l'ensemble des descendants de T. La probabilité de T est alors définie par :

$$p(T) = \frac{freq(T)}{freq(racine)}$$

où *freq(racine)* est la fréquence du terme racine.

3.1.2.2 Contenu d'information (information content) de Resnik

Suivant la définition de la probabilité d'un terme GO T, Resnik définit l'*information content* de T, IC(T), comme moins sa logvraisemblance : *IC(T) = -log(p(T))*. L'interprétation intuitive de ce concept est que plus un terme est globalement probable, moins l'information qu'il apporte est importante.

Il utilise ensuite cet *information content* pour définir une mesure de similarité, que nous n'utiliserons pas ici : la similarité de 2 termes correspond à l'*information content* de leur ancêtre commun ayant l'*information content* le plus élevé :

$$sim_{Resnik}(T_1, T_2) = \max_{T \in ancêtres(T_1,T_2)} (-\log\,(p(T))$$

où *ancêtres(T_1,T_2)* est l'ensemble des ancêtres communs des termes T_1 et T_2. Cette mesure a pour inconvénient de ne pas prendre en compte la distance séparant T_1 et T_2 de leur ancêtre commun, et de ne pas avoir de maximum.

3.1.2.3 Mesure de similarité de Lin

Lin définit la similarité entre deux termes T_1 et T_2 comme le ratio de deux fois l'*information content* de leur ancêtre commun ayant l'*information content* le plus élevé sur la somme des *information contents* de T_1 et T_2. Intuitivement, cela correspond au ratio entre l'information apportée par l'ancêtre commun qui décrit le plus T_1 et T_2 sur l'information de la description complète (somme de leurs informations) de T_1 et T_2 :

$$sim_{Lin}(T_1, T_2) = \max_{T \in ancêtres(T_1,T_2)} \left(\frac{2 \cdot \log\,(p(T))}{\log\big(p(T_1)\big) + \log\,(p(T_2))}\right)$$

Contrairement à la mesure de similarité de Resnik, cette mesure est bornée entre 0 et 1. Mais elle a pour inconvénient de peu prendre en compte la distance à la racine : les termes

abstraits, hauts placés dans l'ontologie, sont peu pénalisés par rapport aux termes plus précis.

3.1.2.4 Relevance similarity

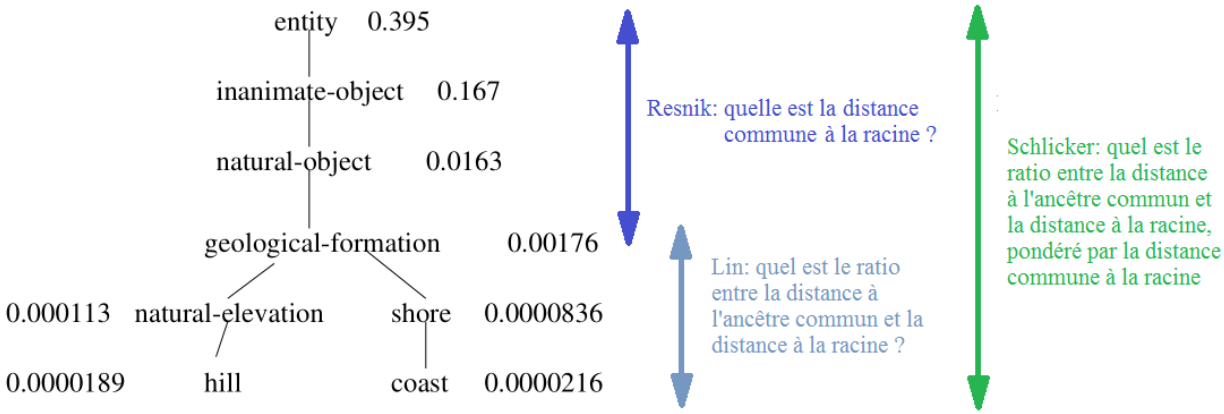

Figure 28: Relations entre les mesures de similarité définies par Resnik, Lin et Schlicker, sur une ontologie tirée de (Lin, 1998).

Afin de prendre en compte l'*information content* de Resnik dans la mesure de stabilité de Lin (Figure 28), Schlicker définit la *relevance similarity*, sim_{Rel}, dans laquelle la mesure de similarité de Lin est pondérée par $1\text{-}exp(\text{-}sim_{Resnik})$: cette pondération permet de prendre en compte qu'un terme plus fréquent ($p(T)$ élevée) est moins informatif. Sim_{Rel} est définie par :

$$sim_{Rel}(T_1, T_2) = \max_{T \in ancêtres(T_1,T_2)} \left(\frac{2 \cdot \log(p(T))}{\log(p(T_1)) + \log(p(T_2))} \cdot (1 - p(T)) \right)$$

Cette mesure est bornée entre 0 et 1.

3.1.2.5 Similarité fonctionnelle entre deux gènes

Une fois la mesure de similarité entre les termes définie, il reste à la transposer aux gènes, qui sont le plus souvent associés à plusieurs annotations. Nous avons utilisé la méthode de calcul de similarité fonctionnelle entre deux gènes décrite par (Wang et al., 2007). Pour cela, on définit d'abord la similarité fonctionnelle $Sim(go,GO)$ entre un terme go et un ensemble de termes $GO=\{go_1, go_2, ..., go_k\}$. $Sim(go,GO)$ est définie comme la similarité maximale entre le terme go et n'importe lequel des termes go_i de l'ensemble GO :

$$Sim(go, GO) = \max_{i \in [1;k]} (sim_{Rel}(go, go_i))$$

À partir de cette mesure, on définit la similarité fonctionnelle $Sim(G_1,G_2)$ entre deux gènes G_1 et G_2, annotés par les ensembles de termes $GO_1=\{go_{11}, go_{12}, ..., go_{1k}\}$ et $GO_2=\{go_{21}, go_{22}, ..., go_{2m}\}$ comme la moyenne des similarités entre chaque terme du premier avec le second et des similarités entre chaque terme du second avec le premier :

$$Sim(G_1, G_2) = \frac{\sum_{i\in[1,k]} Sim(go_{1i}, GO_2) + \sum_{i\in[1,m]} Sim(go_{2i}, GO_1)}{k+m}$$

Tout comme Sim_{Rel} et $Sim(go,GO)$, cette mesure est bornée entre 0 et 1.

3.1.3 Méthode CoGO

Un des premiers choix qui nous est venu à l'esprit pour combiner données *a priori* et données observées a été de combiner la mesure de similarité GO de Wang avec une autre mesure de similarité, observationnelle cette fois : la corrélation de Spearman entre les niveaux d'expression des gènes. La corrélation entre les variables a déjà été utilisée à des fins de sélection de variables par clustering, sur des données protéomiques, dans (Shin et al., 2008). La méthode mise en oeuvre dans cette publication a utilisé les corrélations, en valeur absolue, entre les variables, sans intégrer des données *a priori*, pour définir des clusters. Puis un représentant unique était conservé par cluster, et la sélection finale était obtenue par un filtre t-score appliqué sur le sous-ensemble constitué par les représentants uniques de chaque cluster. Les analyses réalisées ne se sont pas intéressées à la stabilité, mais ont en revanche montré une amélioration modérés des performances de classification par régression logistique par rapport à une sélection par t-score seul.

Afin de nous assurer de la pertinence de combiner similarité GO et corrélations en une métrique de similarité composite, nous avons d'abord cherché à nous assurer qu'il existe bien un lien entre similarité fonctionnelle et corrélation du niveau d'expression. Pour cela, nous avons regroupé les gènes par fonction, et calculé, pour chaque fonction, la corrélation de Spearman moyenne, en valeur absolue, entre les niveaux d'expression des gènes annotés par cette fonction. Le choix de la valeur absolue de la corrélation est justifié par le fait que les annotations ne précisent pas, par exemple, si un gène annoté est associé positivement ou négativement à la fonction, ainsi on peut s'attendre à trouver aussi bien des corrélations négatives que positives : c'est l'intensité de ces corrélations, et non leur signe, qui nous intéresse. Sur les données de

(Pawitan et al., 2005), la corrélation de Spearman entre les niveaux d'expression des gènes dans les groupes fonctionnels était de 0.32, alors qu'elle n'était que de 0.13 sur le jeu de données en général. Sur les données Diogenes (présentées dans le chapitre suivant), on observe une différence semblable : corrélation de 0.31 à l'intérieur des groupes fonctionnels vs 0.15 sur l'ensemble des gènes. Ainsi, la corrélation en valeur absolue des niveaux d'expressions entre gènes partageant une même fonction, bien que modeste, est notablement plus élevée que la corrélation entre gènes tirés au hasard : combiner un regroupement par corrélation et un regroupement par similarité fonctionnel pourrait donc s'avérer à la fois compatible et complémentaire.

La méthode CoGO que nous avons développée peut être divisée en 3 étapes (Figure 29) :

1. construction d'une matrice de similarité composite entre les gènes,
2. clustering des gènes à partir de cette matrice, puis on ne garde que le centroïde de chaque cluster, réduisant *D*,
3. application d'une méthode de sélection de variable habituelles sur cet ensemble restreint.

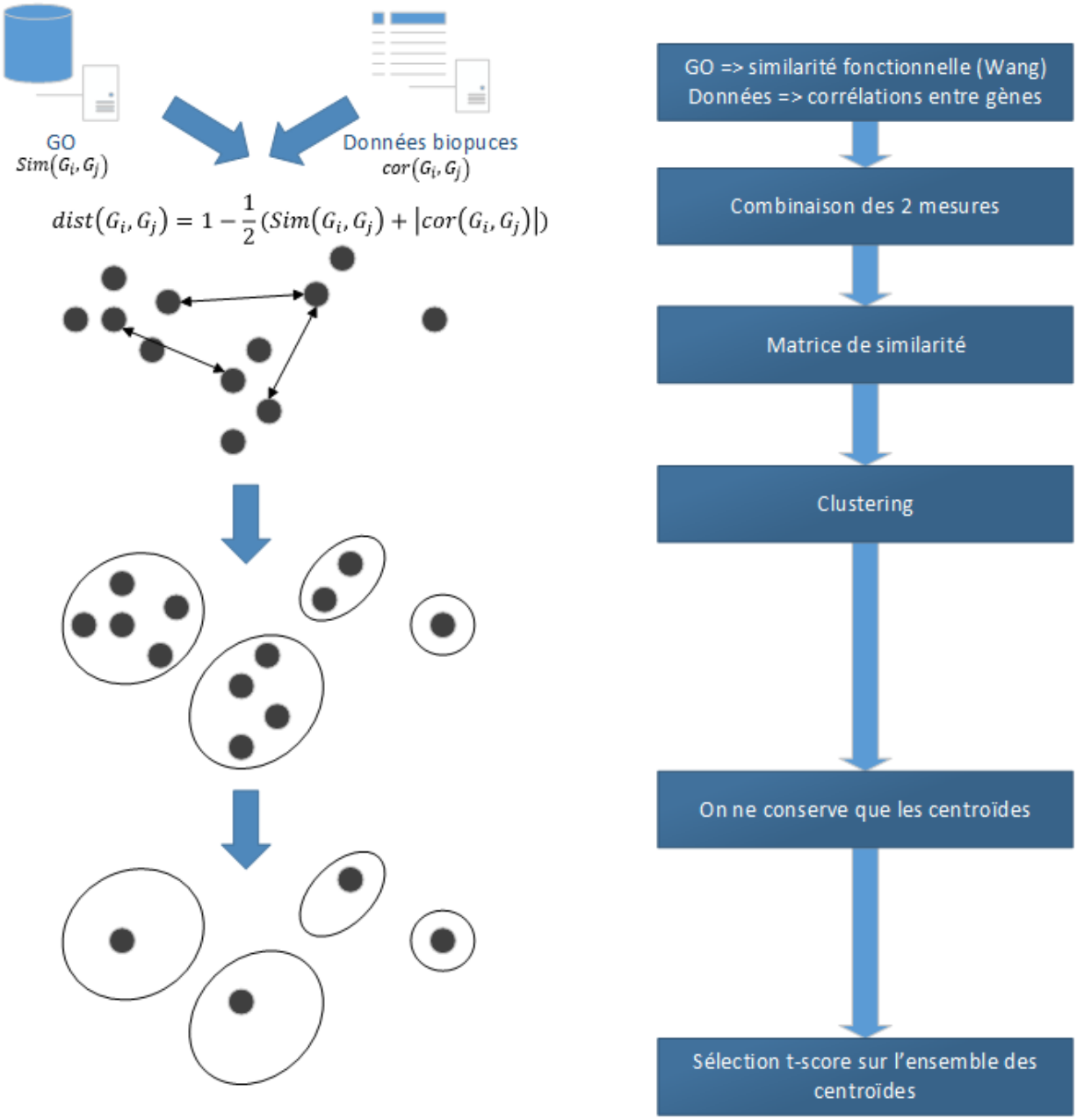

Figure 29: Pipeline de la méthode CoGO

Dans la première étape, une première matrice est construite à partir de GO, en calculant la similarité fonctionnelle de Wang entre chacun des gènes. Cette matrice, très longue à calculer (~50h), est calculée une fois pour toutes pour le jeu de données à analyser (ou plus généralement, elle est propre à chaque biopuce). Une seconde matrice est construite à partir des données d'expression, en calculant les corrélations entre chacun des gènes. Cette matrice est, heureusement, bien plus rapide à calculer. Pour éviter le surapprentissage, elle doit être

recalculée, lors de la validation croisée, sur chaque échantillon d'apprentissage (et non globalement sur l'ensemble des observations du jeu de données).

Ces deux matrices sont ensuite combinées, simplement en réalisant une moyenne, qui peut être pondérée ou non selon qu'on veuille donner plus d'importance à l'une des deux sources de données. Dans nos expériences, nous leur avons donné un poids égal, et la matrice de similarité Sim_{CoGO} peut être définie comme :

$$Sim_{CoGO} = \frac{1}{2}\begin{bmatrix} Sim(G_1,G_1) + |cor(G_1,G_1)| & \cdots & Sim(G_D,G_1) + |cor(G_D,G_1)| \\ \vdots & \ddots & \vdots \\ Sim(G_1,G_D) + |cor(G_1,G_D)| & \cdots & Sim(G_D,G_D) + |cor(G_D,G_D)| \end{bmatrix}$$

On remarquera qu'il s'agit d'une matrice symétrique (car $Sim(G_1,G_2)=Sim(G_2,G_1)$ et $cor(G_1,G_2)=cor(G_2,G_1)$), et que la diagonale est égale à $(1,...,1)$ (car $Sim(G_1,G_1)=1$ et $cor(G_1,G_1)=1$).

Dans la seconde étape, cette matrice Sim_{CoGO} est utilisée pour réaliser un clustering des gènes. La méthode de clustering que nous avons utilisée est le *Partitioning Around Medoids (PAM)* (Kaufman & Rousseeuw, 1990 ; Theodoridis & Koutroumbas, 2008). On crée ainsi des groupes de gènes fonctionnellement "proches" d'après à la fois les données connues (similarité fonctionnelle des annotations Gene Ontology) et les données observées (corrélation entre les niveaux d'expression). Pour chacun de ces groupes, on ne conserve qu'un représentant le plus pertinent : le centroïde du groupe. On réalise ainsi une présélection qui ne devrait pas causer beaucoup de perte d'information pour la classification, chaque paquet fonctionnel identifié étant conservé.

La principale limite de cette étape est son temps de calcul, ainsi que le choix du nombre k de clusters, ces 2 problèmes étant liés. Un trop grand nombre de clusters et de gènes augmente considérablement le temps de calcul. Étant donné que sur les données biopuces, une grande partie des variables ne sont pas pertinentes, nous avons commencé par ne conserver que les 25% des gènes les plus associés à la classe, selon un filtre t-test. Puis nous avons réalisé le clustering sur cet ensemble de gènes réduit, avec $k=200$ clusters soit le double du nombre $d=100$ de variables à conserver dans la sélection finale, ce nombre ayant été choisi car il s'agit d'un ordre de grandeur approprié, fréquemment utilisé sur données biopuces (Dittman et al., 2011). On

obtient ainsi une présélection de 200 centroïdes.

Dans la troisième étape, un filtre t-score est appliqué à cette présélection.

3.2 Mesure de la stabilité fonctionnelle

Il est attendu que les centroïdes des clusters soient variables. En effet, à l'intérieur d'un paquet de gènes très proches, il est vraisemblable que des variations, même mineures, au niveau des données puissent fréquemment déplacer le centroïde du cluster d'un gène à un autre. Si l'on se place au niveau des gènes, il est donc probable que la présélection soit instable, ce qui affecterait également la stabilité de la sélection finale. En revanche, quand le centroïde change, on s'attend à ce que son "remplaçant" soit fonctionnellement très proche, puisqu'il est, *a priori*, issu du même cluster. Afin de mesurer la stabilité en nous affranchissant des variations de centroïde, nous avons donc ajouté aux mesures de stabilité de la sélection des gènes décrites dans le Chapitre 2 (CW_{rel}, et ATI_{PA}) une mesure de stabilité de la sélection des fonctions.

$$S_1 = (G_{1,1}, G_{1,2}, \dots, G_{1,x}) \qquad S_2 = (G_{2,1}, G_{2,2}, \dots, G_{2,y})$$

Conversion en termes GO

$$S_1 = (n_{1,1}go_{1,1}, n_{1,2}go_{1,2}, \dots, n_{1,x}go_{1,x}) \qquad S_2 = (n_{2,1}go_{2,1}, n_{2,2}go_{2,2}, \dots, n_{2,y}go_{2,y})$$

Suppression des termes non sur-exprimés

$$S_1 = (n_{1,1}go_{1,1}, n_{1,2}go_{1,2}, \dots, n_{1,a}go_{1,a}) \qquad S_2 = (n_{2,1}go_{2,1}, n_{2,2}go_{2,2}, \dots, n_{2,b}go_{2,b})$$

Calcul de la stabilité fonctionnelle

$$Sim(S_1, S_2) = \frac{\sum_{i\in[1,a]} n_{1,i} Sim(go_{1,i}, S_2) + \sum_{i\in[1,b]} n_{2,i} Sim(go_{2i}, S_1)}{\sum_{i\in[1,a]} n_{1,i} + \sum_{i\in[1,b]} n_{2,i}}$$

Figure 30: Méthode de calcul de la stabilité fonctionnelle. $n_{i,j}$ correspond au nombre d'occurences du terme $go_{i,j}$ dans la sélection S_i.

La mesure de stabilité fonctionnelle que nous avons utilisée se base sur la similarité

fonctionnelle entre deux gènes de (Wang et al., 2007) décrite précédemment. Les deux sélections à comparer sont considérées chacune comme l'ensemble des annotations GO de leurs gènes. Les termes GO peuvent être sélectionnés plusieurs fois (si plusieurs gènes de la sélection partagent la même annotation) : le nombre d'occurrences de chaque terme dans les sélections est calculé. Nous appliquons ensuite un test hypergéométrique, tel que décrit dans (Fury et al., 2006) pour déterminer, dans chaque sélection, quels sont les termes surreprésentés par rapport à l'ensemble des annotations des gènes de la biopuce. Nous avons retenu un seuil de significativité de 5% avec ajustement de Bonferroni. La similarité fonctionnelle de Wang est ensuite appliquée sur les ensembles de termes restants, en pondérant chaque terme par son nombre d'occurrences. La procédure est détaillée sur la Figure 30.

Enfin, afin de tenir compte, comme dans les mesures de stabilité ATI_{PA} et CW_{rel}, de la stabilité liée au hasard, la mesure est ajustée par sa valeur attendue sur une sélection aléatoire, déterminée expérimentalement sur chaque jeu de données. Cette stabilité fonctionnelle ajustée SF_A est donc définie comme :

$$SF_A(S_1, S_2) = \frac{Sim(S_1, S_2) - Sim_{exp}(.)}{Sim_{max}(.) - Sim_{exp}(.)}$$

où $Sim_{max}(.) = 1$ est la stabilité maximale possible et $Sim_{exp}(.)$ est la stabilité attendue sur une sélection aléatoire, déterminée expérimentalement sur 100 sélections aléatoires.

3.3 Expérimentation de CoGO sur les données DiOGenes et Golub

3.3.1 Design expérimental

En raison de la lenteur des méthodes mises en oeuvre, aussi bien pour la mesure de la stabilité fonctionnelle que pour la sélection de variables, nous avons limité nos expériences initiales à deux jeux de données : les données DiOGenes, présentées en introduction, et les données leucémie (Golub et al., 1999), déjà utilisées précédemment.

D'une manière similaire à la méthode utilisée dans le Chapitre 2, 50 échantillons d'apprentissage ont été générés par tirage aléatoire sans remise de la moitié des observations. Sur chacun d'entre eux, une sélection de variables à été réalisée en utilisant l'une des méthodes de sélections testées : t-test, SVM-RFE, CoGO, information mutuelle, CAT-score, ReliefF,

ensemble t-score et ensemble SVM-RFE (l'agrégation des ensembles a été réalisée par le score moyen, cf Chapitre 4). Puis un classifieur LDA et un classifieur forêt aléatoire ont été construits sur la sélection. Pour chaque échantillon d'apprentissage, l'échantillon de validation correspondant était constitué de la moitié d'observations restante. Le taux d'erreur moyen a été mesuré en appliquant les deux classifieurs à l'échantillon de validation correspondant. La stabilité de la sélection a été mesurée en comparant 2 à 2 les sélections réalisées sur chaque paire échantillon test – échantillon de validation, avec les mesures CW_{rel}, ATI_{PA} et SF_A.

3.3.2 Résultats

3.3.2.1 Comparaison de la mesure de stabilité fonctionnelle aux mesures de stabilité des gènes

La Figure 31 présente les évolutions d'ATI_{PA} et SF_A en fonction de CW_{rel}. CW_{rel} et ATI_{PA} évoluent globalement de manière fortement similaire, à l'exception d'un point (correspondant à la sélection par ReliefF sur les données leucémie). La stabilité fonctionnelle SF_A est un peu moins liée à la stabilité CW_{rel}. La corrélation entre les mesures reste tout de même nette dans les deux cas (0.972 entre CW_{rel} et ATI_{PA}, 0.916 entre CW_{rel} et SF_A).

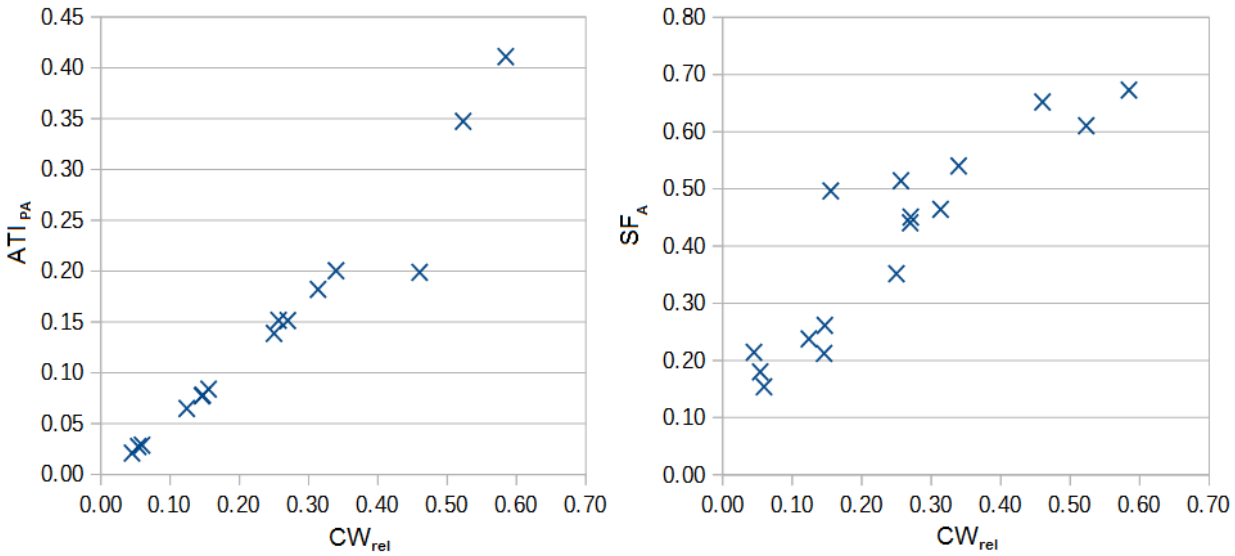

Figure 31: Mesures de stabilité ATI_{PA} (à gauche) et SF_A (à droite) en fonction de CW_{rel}.

3.3.2.2 Performances du filtre CoGO

Le Tableau 4 présente les stabilités et taux d'erreur (moyenne des deux classifieurs) en

fonction de la méthode de sélection utilisée sur les 2 jeux de données. Dans les deux cas, la méthode la plus stable est l'ensemble SVM-RFE, suivi par le SVM-RFE simple, et ce quelle que soit la mesure de stabilité utilisée. La sélection SVM-RFE, a fortiori dans sa version ensemble, aboutit à des classifieurs parmi les plus précis sur les données leucémie, mais aux classifieurs les moins performants sur les données DiOGenes. ReliefF obtient, sur les deux jeux de données, un des meilleurs taux d'erreur, mais des résultats hétérogènes en matière de stabilité : sur DiOGenes, il a l'une des pires stabilités CW_{rel}, à peine supérieure à celle de l'information mutuelle et du CAT-score, et la stabilité fonctionnelle la plus faible, en revanche il a la seconde stabilité fonctionnelle et la 3e stabilité CW_{rel} sur les données leucémie.

Table 4: Stabilité et taux d'erreur moyen en fonction de la méthode de sélection sur les données DiOGenes et leucémie.

Données	Méthode de sélection	Taux d'erreur	CWrel	ATI_{PA}	SF_A
DiOGenes	t-score	**0.108**	0.124	0.065	0.238
	SVM-RFE	0.177	0.156	0.084	0.496
	CoGO	0.120	0.126	0.065	0.212
	Information mutuelle	0.120	0.054	0.028	0.180
	CAT-score	0.120	0.045	0.021	0.214
	ReliefF	0.115	0.060	0.029	0.154
	ensemble t-score	0.166	0.147	0.078	0.261
	ensemble SVM-RFE	0.189	**0.257**	**0.152**	**0.514**
Leucémie (Golub)	t-score	0.045	0.314	0.182	0.464
	SVM-RFE	0.047	0.523	0.347	0.610
	CoGO	0.067	0.250	0.139	0.352
	Information mutuelle	0.047	0.270	0.151	0.441
	CAT-score	0.047	0.271	0.152	0.451
	ReliefF	**0.043**	0.460	0.199	0.652
	ensemble t-score	0.091	0.340	0.200	0.540
	ensemble SVM-RFE	**0.043**	**0.585**	**0.411**	**0.673**

La sélection CoGO obtient une stabilité dans la moyenne sur les données DiOGenes, mais la plus basse sur les données leucémie. En particulier, la stabilité fonctionnelle de CoGO n'est pas meilleure que sa stabilité CW_{rel}, et on observe même une tendance inverse : sur les données DiOGenes, CoGO obtient la 4e stabilité CW_{rel} et ATI_{PA}, mais la 6e stabilité fonctionnelle, égalée en particulier par la stabilité fonctionnelle du CAT-score, qui a pourtant une stabilité CW_{rel} bien moindre (0.045 vs 0.126). Sur les données leucémie, la stabilité de CoGO est la plus basse, aussi bien au niveau des gènes qu'au niveau des fonctions, et les classifieurs construits sur la sélection

CoGO sont moins précis que ceux construits sur les autres sélections (la seule sélection moins précise étant alors l'ensemble t-score).

3.3.2.3 Performances du filtre CoGO avec une matrice de similarité opposée

Suite à une erreur dans notre implémentation initiale, nous avons également réalisé des mesures en exécutant CoGO sur une matrice de similarité égale à l'opposée de la véritable matrice de similarité. Les résultats obtenus sont présentés dans le Tableau 5. Par rapport à la version correcte de CoGO, les taux d'erreur sont proches bien qu'un peu augmentés. En revanche, les stabilités sont très diminuées, et cette diminution est particulièrement marquée sur la stabilité fonctionnelle, qui est négative, c'est-à-dire inférieure à la stabilité d'une sélection aléatoire, sur les deux jeux de données.

Table 5: Stabilité et taux d'erreur moyen de la méthode CoGO avec une matrice de similarité opposée.

Données	Méthode de sélection	Taux d'erreur	CWrel	$AT\Pi_{PA}$	SF_A
DiOGenes	CoGO opposée	0.135	0.035	0.016	-0.155
Leucémie	CoGO opposée	0.069	0.155	0.079	-0.106

3.4 Discussion et conclusion

La méthode CoGO, dans cette version, n'a pas permis une amélioration de la stabilité et a même tendance à produire des sélections peu stables. Une première hypothèse pour expliquer ces mauvaises performances est que la correspondance entre d'une part les corrélations observées sur les données biopuces, et d'autre part les groupes de fonctions GO serait insuffisante pour en faire une combinaison pertinente.

En effet, bien que les corrélations entre gènes partageant une même annotation soient, en valeur relative, plus du double de celles entre gènes pris au hasard, en valeur absolue la différence reste modérée et les valeurs des corrélations à l'intérieur des fonctions restent faibles (~0.32 comme nous l'avons indiqué dans la présentation de la méthode). Cependant, les expériences réalisées avec la méthode CoGO "opposée" suggèrent que les regroupements basés sur les fonctions et les corrélations ont bien une certaine pertinence : quand on réalise des sélections basées sur des clusters les plus fonctionnellement hétérogènes possibles, sur la base

de la métrique de similarité définie dans CoGO, la stabilité est très basse en général et au niveau fonctionnel en particulier. Ce qui indique que bien que certains gènes soient communs entre les sélections ($CW_{rel}>0$), au-delà de l'intersection liée au hasard ($ATI_{PA}>0$), les gènes non communs sont par contre tellement différents d'une sélection à l'autre que la stabilité fonctionnelle mesurée est inférieure à la stabilité fonctionnelle liée au hasard ($SF_A<0$). La métrique de similarité de CoGO est donc bien capable, dans une certaine mesure, d'identifier quels gènes sont les plus distants, et donc inversement lesquels sont les plus proches. Pour aller plus loin, il serait intéressant d'étudier une version utilisant uniquement la similarité fonctionnelle GO et une version utilisant uniquement les corrélations.

Plusieurs autres pistes peuvent être avancées pour expliquer les mauvaises performances de CoGO. Tout d'abord, GO inclut des annotations de niveaux de preuve variés. Par exemple, en octobre 2007, sur 16 millions d'annotations GO, plus de 95% étaient des annotations obtenues automatiquement, et chez l'humain, moins de 30% des gènes avaient une annotation de source expérimentale (Rhee et al., 2008). Il semble difficile de se passer des annotations non expérimentales, mais leur prise en compte implique une certaine quantité d'erreurs, qui peuvent réduire la pertinence de nos groupements fonctionnels. Peut-être que la mise en place d'une pondération des annotations, en fonction de leur niveau de preuve, dans le calcul de la matrice de similarité pourrait constituer un compromis intéressant, bien que difficile à calculer, posant en particulier une nouvelle problématique de choix d'un paramètre (quelle pondération donner aux différents niveaux de preuve ?).

D'autre part, une partie des gènes (moins de 5% sur les données leucémie, plus de 20% sur les données DiOGenes) n'ont aucune annotation. Afin de ne pas les éliminer de l'analyse, nous leur avons assigné une similarité fonctionnelle *a priori* (les corrélations ont ensuite été mesurées normalement) avec les autres gènes égale à la similarité fonctionnelle *a priori* moyenne sur l'ensemble des gènes analysés. Cette approximation a pu nuire également à la qualité des clusters. Cependant, les performances de CoGO sont moins bonnes sur les données leucémie que sur DiOGenes, alors que les données leucémie ont plus de quatre fois moins d'annotations manquantes.

Les simplifications réalisées au niveau de l'implémentation de la méthode, nécessaires afin d'alléger les calculs, ont également pu dégrader les performances. Le préfiltrage, par un filtre t-

score, pour ne garder que 25% des gènes les plus associés à la classe avant de les clusteriser n'est sans doute pas une limitation majeure, car sur des données puces, pour un problème donné un grand nombre de variables ne sont pas pertinentes et il est relativement facile d'en éliminer une grande partie de manière suffisamment fiable (Somol & Novovicova, 2010). Cela dit, les clusters des gènes les moins pertinents, à proximité de ce seuil de 25%, sont vraisemblablement rendus instables par l'élimination occasionnelle d'une partie de leurs membres. Il serait donc intéressant de mettre en oeuvre la méthode sans préfiltrage... ce qui augmenterait d'autant plus les temps de calcul qu'il faudrait également augmenter le nombre de clusters. La fixation arbitraire du nombre de clusters pourrait être plus problématique, mais optimiser automatiquement le nombre de clusters semble difficilement réalisable vus les temps de calcul observés pour réaliser une sélection avec un nombre de clusters prédéterminé (~30 minutes pour une sélection avec 200 clusters). Nous avons réalisé des expériences avec un autre nombre de clusters (150), les résultats obtenus ont été similaires. Cela suggère que le nombre de clusters n'a pas une trop grande importance, en tout cas dans cette plage de valeurs, mais il n'empêche qu'une recherche plus exhaustive d'un nombre de clusters optimal, intégrée à la méthode, serait un plus si elle ne s'accompagnait pas d'une augmentation déraisonnable du temps de calcul. Il pourrait aussi être intéressant de s'intéresser à l'homogénéité des clusters, plutôt que leur nombre, comme critère d'arrêt du clustering.

Enfin, une dernière piste pour expliquer la stabilité moindre de CoGO est liée à son fonctionnement théorique même : CoGO essaye de regrouper des gènes fonctionnellement proches, via leurs annotations et leurs corrélations observées, afin de ne conserver qu'un représentant pertinent par groupe fonctionnel. Cela permet de réaliser la sélection finale à partir d'un sous-ensemble réduit de gènes, conservant le plus possible l'information contenue dans le jeu de données complet, ce qui devrait donc être plus stable par augmentation du ratio N/D. Mais, même en mettant de côté l'instabilité de l'étape de création des clusters et d'identification de leur centroïde, le fait de ne conserver qu'un seul représentant par groupe fonctionnel pourrait de par lui-même diminuer la stabilité. En effet, en ne conservant qu'un seul représentant par groupe fonctionnel, on élimine des variables redondantes. Or, des travaux suggèrent que la stabilité est liée à la redondance (Haury, 2012) : il est donc possible qu'en diminuant trop la redondance, on diminue la stabilité. De plus, conserver un certain nombre (restreint !) de

variables pertinentes redondantes peut permettre plus de robustesse de la classification en cas d'aléa de mesure sur l'une d'entre elles. Enfin, on peut imaginer que certains groupes fonctionnels soient mieux représentés, indépendamment des aléas de mesures, par plusieurs gènes pltôt que par un gène unique, ce que notre méthode ne permet pas de détecter. Cela pourrait expliquer les performances de classification un peu moindres sur les sélections CoGO, et peut-être serait-il pertinent de réaliser une version qui conserverait non plus un mais 2 ou 3 représentants par cluster, ou, pour tenir compte de leur taille, une certaine proportion de gènes par cluster.

Chapitre 4 :
Apport des méthodes d'ensemble pour la stabilité de la sélection de variables

Dans ce chapitre, nous revenons à la sélection basée uniquement sur les données observées, et nous nous intéressons cette fois à l'apport des méthodes d'ensemble pour l'amélioration de la stabilité de la sélection de variables. Le concept de la sélection par méthodes d'ensemble est de produire plusieurs sélections indépendantes puis de les combiner de manière à obtenir une sélection agrégée plus performante que chacune d'entre elles. Ce concept est basé sur celui de la « sagesse des foules », qui décrit que, *sous certaines conditions contrôlées, l'agrégation d'informations de plusieurs sources permet d'obtenir des règles de décisions souvent supérieures à celles qui auraient pu être produites par un individu seul, même expert* (Surowiecki, 2004 ; Rokach, 2010).

Cependant, toutes les foules ne sont pas « sages », et Surowiecki précise que pour être sage, une foule doit remplir les critères suivants :

- diversité des opinions : chaque membre doit avoir de l'information personnelle, même si cette information n'est qu'une interprétation subjective des faits objectifs
- indépendance : l'opinion de chaque membre ne doit pas être influencée par les opinions des autres
- décentralisation : les membres peuvent se spécialiser et raisonner en se basant sur des connaissances locales
- agrégation : on dispose d'un mécanisme pour agréger les opinions et décisions individuelles en une décision collective.

Les méthodes d'ensemble tentent de reproduire ces conditions dans l'analyse du jeu de données à traiter. Leur première application en apprentissage supervisé remonte aux années

1970, quand (Tukey, 1977) propose de combiner deux modèles de régression linéaire, l'un entraîné sur les données d'origine et l'autre sur les résidus du premier. Quelques années plus tard, (Dasarathy & Sheela, 1979) présentent un classifieur « composite », constitué de 2 ou plus classifieurs « composants » (dans la publication d'origine, un classifieur linéaire et un classifieur kNN).

Depuis, les méthodes d'ensemble ont été largement appliquées aux classifieurs (Peng, 2006 ; Chen & Zhao, 2008 ; Reboiro-Jato et al., 2014), et puisqu'elles ont souvent permis d'en améliorer la précision ou la robustesse au bruit, nous pouvons supposer qu'elles pourraient apporter des bénéfices similaires aux méthodes de sélection de variables (Saeys et al., 2008 ; Boulesteix & Slawski, 2009 ; Yang et al., 2010). Pour l'instant, la plupart des travaux sur la sélection de variables par méthodes d'ensemble se sont plutôt focalisés sur la précision de la classification, mettant en avant, sur les données haute dimension et petit échantillon, des gains (ou des pertes) de précision problème-dépendant (Han et al., 2013) et "filtre-dépendant" (Wald et al., 2013b). Les travaux qui ont aussi étudié la stabilité de la sélection ont obtenu des résultats variables également, laissant l'impression générale que les gains (ou pertes), aussi bien en stabilité de la sélection qu'en précision de la classification, obtenus par les méthodes d'ensemble sont problème-dépendants (Saeys et al., 2008). Cependant, ces travaux ont souvent utilisé des mesures de stabilité sur des rééchantillonnages non disjoints, ce qui augmente fortement la stabilité mesurée (Haury et al., 2011) et pourrait influencer les variations relatives de stabilité. Par ailleurs, à notre connaissance, aucune étude ne s'est intéressée spécifiquement à l'influence de la méthode d'agrégation utilisée dans l'ensemble sur la stabilité (Awada et al., 2012).

4.1 Influence de la méthode d'agrégation sur la stabilité

Dans cette section, publiée dans (Dernoncourt et al., 2014b), nous étudions si les méthodes d'ensemble améliorent la stabilité de la sélection de variables, avec un focus sur l'influence de la méthode d'agrégation. Dans un premier temps, nous présentons les étapes clés des méthodes d'ensemble, puis nous réalisons une étude empirique de la stabilité de la sélection sur des données artificielles et réelles.

4.1.1 Sélection de variables par méthodes d'ensemble

La création d'une sélection de variables par ensemble peut être divisée en deux étapes. La

première étape, qui correspond aux trois premiers critères de l'obtention d'une « foule sage » décrits par (Surowiecki, 2004), consiste à générer un ensemble de sélecteurs diversifiés. La seconde étape, tout aussi importante (Wald et al., 2012), correspond au dernier critère de Surowiecki et consiste à les agréger.

4.1.1.1 Génération de la diversité

La diversité des sélecteurs de variables peut être obtenue via différentes méthodes, dont une taxonomie a été proposée par (Brown et al., 2005) dans le cadre de la classification. Transposées à la sélection de variables, nous retiendrons en particulier les méthodes suivantes, qui peuvent être utilisées seules ou associées :

- la manipulation de l'échantillon d'apprentissage : les observations fournies à la méthode de sélection varient à chaque itération. Typiquement, on réalise un rééchantillonnage aléatoire et chaque sélection de variables contenue dans l'ensemble est ainsi obtenue à partir d'un échantillon d'apprentissage différent.
- la manipulation de la méthode de sélection : par exemple, utilisation de différentes valeurs pour les paramètres de la méthode de sélection, si celle-ci en a.
- le partitionnement de l'espace de recherche : chaque sélection de variables est réalisée à partir d'un espace de recherche différent. Par exemple, dans une forêt aléatoire, chacun des arbres est appris à partir d'un sous-ensemble de variables différent et aléatoire.
- et enfin l'hybridation, qui consiste à utiliser des méthodes de sélection différentes à l'intérieur de l'ensemble.

Dans cette partie, nous nous focalisons sur la manipulation de l'échantillon d'apprentissage, qui est la méthode la plus couramment utilisée. La diversité a ainsi été obtenue en rééchantillonnant les données par bootstrap $B = 40$ fois. Ce nombre a été choisi afin de limiter les temps de calcul et parce que dans nos expériences prélimiaires, nous avons constaté qu'au dessus de $B > 40$ les résultats ne changaient pas de façon significative.

4.1.1.2 Agrégation

Nous avons testé les méthodes d'agrégation suivantes :

- score moyen : sur chaque rééchantillonnage, la méthode de sélection produit un score $s_{f_i,j}$ pour chaque gène f_i. Le score final W_{f_i} du gène f_i sur l'ensemble est simplement obtenu en faisant la moyenne des scores obtenus sur chacun des rééchantillonnages :

$$W_{f_i} = \frac{\sum_{j=1}^{B} s_{f_i,j}}{B}$$

- rang moyen (Abeel et al., 2010) : sur chaque rééchantillonnage, le score $s_{f_i,j}$ de chaque gène est converti en un rang $r_{f_i,j}$. Le score final W_{f_i} est obtenu en faisant la moyenne des rangs obtenus sur chacun des rééchantillonnages :

$$W_{f_i} = \frac{\sum_{j=1}^{B} r_{f_i,j}}{B}$$

- *stablity selection* (Meinshausen and Bhlmann, 2010) : le score final W_{f_i} de chaque gène est obtenu en mesurant la fréquence avec laquelle le gène est classé parmi les d gènes les mieux classés sur chacun des rééchantillonnages :

$$W_{f_i} = \frac{\sum_{j=1}^{B} I(f_{i,j})}{B}$$

où $I(f_{i,j}) = 1$ si le gène f_i est classé dans le top d sur la sélection de variables réalisée sur le $j^{ème}$ rééchantillonnage de l'ensemble, et $I(f_{i,j}) = 0$ sinon.

Pour chacune de ces méthodes, les d gènes avec le score W_{f_i} le plus élevé sont retenus dans la sélection finale produite par l'ensemble.

4.1.2 Design expérimental

Afin d'évaluer l'impact des méthodes d'ensemble sur la stabilité de la sélection de variables, nous avons réalisé des expériences sur des données artificielles et sur des données réelles. Les méthodes de sélections de base testées ont été : le t-score, les forêts aléatoires, SVM-RFE et l'information mutuelle. Pour chacune de ces méthodes, une sélection a été réalisée sans et avec méthode d'ensemble, et les méthodes d'ensemble ont utilisé chacune des méthodes d'agrégation décrites dans la section précédente.

La stabilité a été mesurée entre des échantillons sans recouvrement : pour les données artificielles, d'autres échantillons générés, pour les données artificielles, des échantillons constitués à chaque fois des observations non utilisées pour l'apprentissage. Nous avons utilisé la *relative weighted consistency*, *CWrel*, décrite en 2.2.1, comme mesure de stabilité unique. Ce choix a été fait afin de ne pas multiplier les critères de mesure, parce qu'il s'agit d'une méthode adaptée aux données haute dimension contenant de nombreuses variables non pertinentes (pas de surévaluation de la stabilité quand de nombreuses variables sont faciles à éliminer de manière constante), et parce que CW_{rel} est ajustée pour tenir compte de la stabilité liée au hasard. Par ailleurs, nous avons pu constater dans le Chapitre 2 que les 4 mesures de stabilités qui y sont présentées évoluent globalement de la même manière. L'erreur de classification a été mesurée sur un classifieur LDA appliqué aux échantillons tests.

4.1.2.1 Données artificielles

Trois types de données artificielles ont été utilisés, semblables à celles utilisées dans le Chapitre 2, basées sur un modèle gaussien avec $D=1000$ variables. Chacune des deux classes suit une distribution normale définie respectivement par $\mathcal{N}(\mu,\Sigma)$ et $\mathcal{N}(-\mu,\Sigma)$, où μ est un vecteur de moyennes tel que $|\mu| = D$ et Σ est la matrice de covariance.

Dans la première structure, "*NC100*" (non corrélé, 100 variables significatives), toutes les variables étaient indépendantes (Σ est la matrice identité) et le jeu de données contenait $d=100$ variables associées (toute avec la même intensité : $\mu_i = 1$) à la classe et $D - d = 900$ variables non informatives ($\mu_i = 0$). μ a ensuite été recalibré pour assurer une erreur de Bayes de 10%.

Dans la deuxième structure, "*NC*", toutes les variables étaient indépendantes et les éléments μ_i of μ étaient tirés d'une distribution triangulaire de densité de probabilité $f(x)=2-2x$ sur [0;1], puis élevés au carré. μ a ensuite été ensuite recalibré pour assurer une erreur de Bayes de 10%.

Dans la troisième structure, "CB", les μ utilisés sont les mêmes que dans "NC" mais la matrice de covariance Σ est telle que les variables sont corrélées par blocs de 10 :

$$\Sigma = \begin{vmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{100} \end{vmatrix}$$

où Σ est une matrice de blocs diagonale et Σ_i est une matrice 10×10 avec des éléments valant 1 sur sa diagonale et 0.5 ailleurs.

Sur ces modèles, 50 échantillons d'apprentissage, contenant chacun $N = 50$ observations, ont été générés, sur lesquels ont été appliquées les méthodes de sélection avant apprentissage d'un classifieur LDA. Ce classifieur a été appliqué sur un jeu test contenant 10000 observations.

4.1.2.1 Données réelles

Cinq jeux de données biopuces, présentés dans le Tableau 6, ont été utilisés. Pour chaque jeu de données, 50 échantillons d'entraînement ont été générés par tirage aléatoire, sans remise, de la moitié des observations. Les méthodes de sélection ont été appliquées sur chacun d'entre eux, avant apprentissage d'un classifieur LDA. Chaque classifieur a ensuite été testé sur un échantillon de validation constitué de la moitié des observations non incluse dans l'échantillon d'apprentissage. La stabilité de la sélection a été mesurée sur chaque paire l'échantillon d'apprentissage - l'échantillon test, de façon à ne pas avoir d'observations communes, et la mesure finale de la stabilité correspond à la moyenne de ces 50 mesures.

Table 6: Jeux de données biopuces utilisés

Nom	*N*	*D*	*N/D*	*Source*
Cancer du colon	62	2000	0.03	Alon et al., 1999
Leucémie	72	7129	0.01	Golub et al., 1999
BK Pawitan	159	8112	0.02	Pawitan et al., 2005
Cancer du poumon	203	2000	0.10	Bhattacharjee et al., 2001
BK Vijver	294	2000	0.15	van de Vijver et al., 2002

4.1.3 Résultats

4.1.3.1 Résultats sur les données artificielles

La Figure 32 présente la stabilité de la sélection et le taux d'erreur sur le classifieur en résultant sur les données *NC100*, la Figure 33 sur les données *NC*, et la Figure 34 sur les

données *CB*. D'une manière générale, on observe que la sélection par méthode d'ensemble fournit des résultats soit autant soit plus stables que lorsque la méthode de sélection est utilisée hors ensemble (single), avec un taux d'erreur inférieur ou semblable.

Sur les 3 types de données, le t-score obtient, globalement, la stabilité la plus haute et le taux d'erreur le plus bas. Il est de plus amélioré par sa version ensemble, mais uniquement lorsque l'agrégation est réalisée par le score moyen : l'agrégation par rang moyen ou par *stability selection* a produit des résultats très proches de la méthode sans ensemble, aussi bien sur la stabilité que sur le taux d'erreur.

	t-score		**SVM-RFE**		**Forêt aléatoire**	
	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité
Single	0.338	0.075	0.345	0.065	*0.383*	*0.031*
Score moyen	**0.29**	**0.135**	0.339	0.071	0.347	0.059
Rang moyen	0.34	0.072	0.34	0.072	0.347	0.061
Stability selection	0.339	0.071	0.344	0.068	0.345	0.062

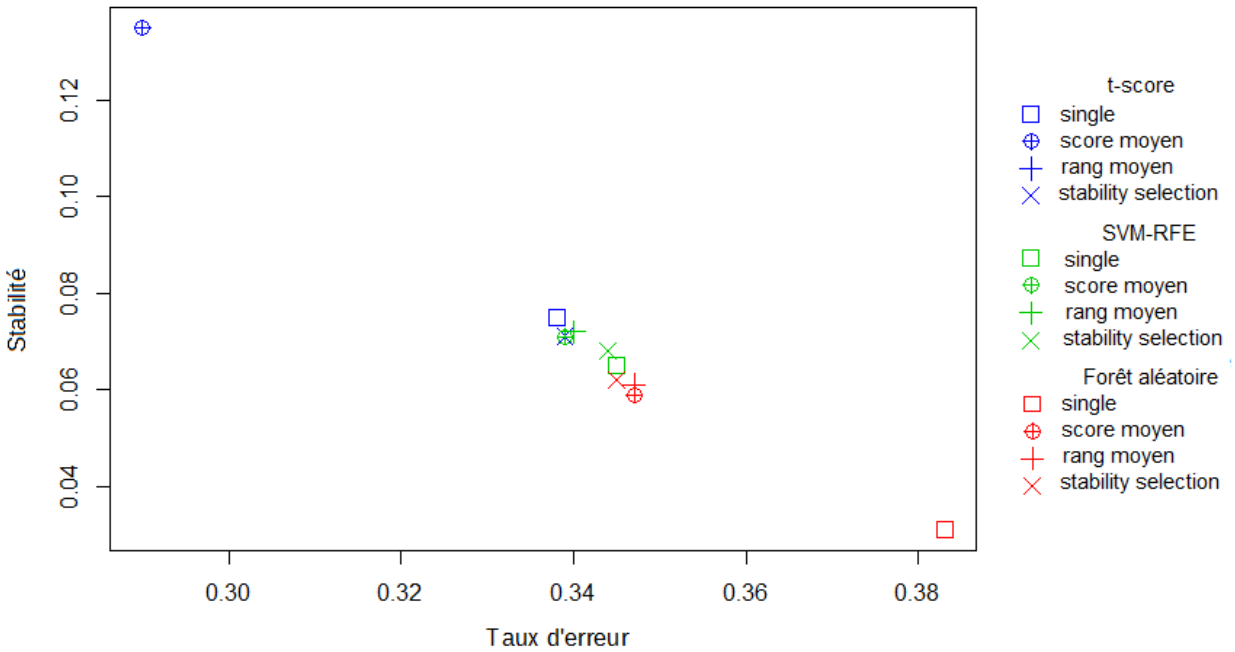

Figure 32: Stabilité et taux d'erreur en fonction de la méthode de sélection et du type d'aggrégation dans la méthode d'ensemble, sur les données NC100. En gras : meilleur point. En italique : pire point.

	t-score		SVM-RFE		Forêt aléatoire	
	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité
Single	0.36	0.05	0.36	0.049	*0.391*	*0.027*
Score moyen	**0.32**	**0.071**	0.364	0.045	0.365	0.042
Rang moyen	0.361	0.045	0.364	0.044	0.363	0.045
Stability selection	0.366	0.043	0.363	0.045	0.368	0.04

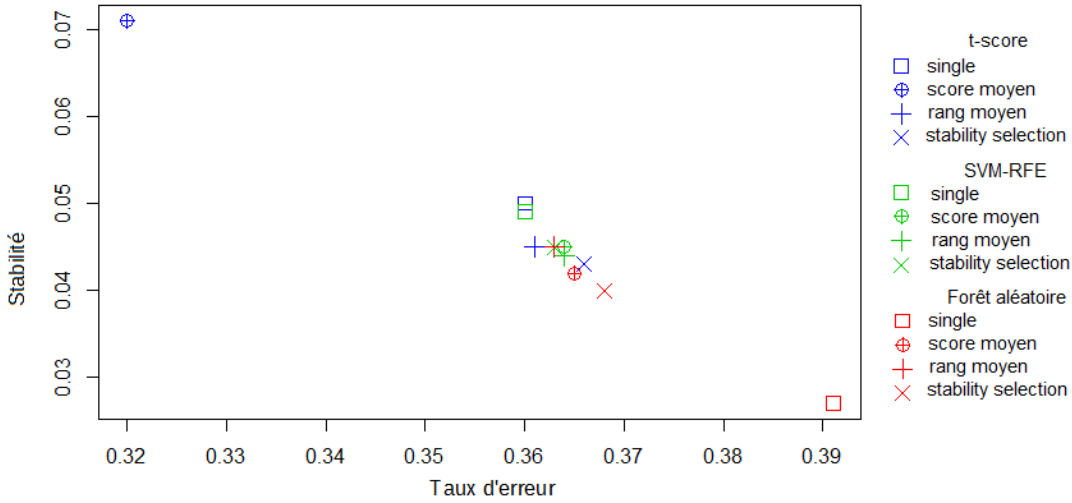

Figure 33: Stabilité et taux d'erreur en fonction de la méthode de sélection et du type d'aggrégation dans la méthode d'ensemble, sur les données NC.

	t-score		SVM-RFE		Forêt aléatoire	
	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité
Single	0.239	0.235	*0.242*	*0.113*	*0.244*	*0.099*
Score moyen	**0.223**	**0.265**	0.238	0.152	0.232	0.193
Rang moyen	0.235	0.226	0.238	0.154	0.23	0.195
Stability selection	0.237	0.217	0.239	0.15	0.233	0.19

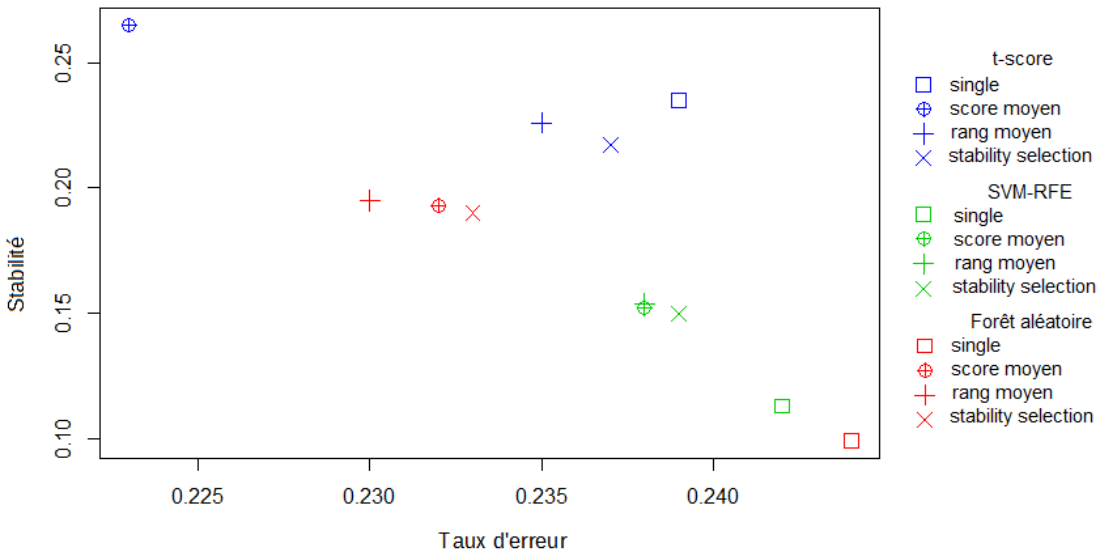

Figure 34: Stabilité et taux d'erreur en fonction de la méthode de sélection et du type d'aggrégation dans la méthode d'ensemble, sur les données CB.

Par rapport au t-score, SVM-RFE obtient une stabilité semblable sur les données non-corrélées, mais de moitié moindre sur les données corrélées. Avec, dans tous les cas, un taux d'erreur proche du t-score. L'utilisation des méthodes d'ensemble n'a pas amélioré SVM-RFE sur les données non corrélées, par contre elle a amélioré sa stabilité sur les données corrélées, de façon semblable quelle que soit la méthode d'agrégation.

Les forêts aléatoires ont obtenu les moins bons résultats, avec une stabilité de moitié moindre que celle du t-score sur l'ensemble des jeux de données, et un taux d'erreur supérieur, surtout sur les données non corrélées. Elles sont en revanche fortement améliorées par les méthodes d'ensemble, qui leur permettent d'atteindre des stabilités et taux d'erreur comparables au SVM-RFE sur les données non corrélées, et meilleurs que les ensembles SVM-RFE sur les données corrélées. Comme avec SVM-RFE, le choix de la méthode d'agrégation utilisée dans l'ensemble n'a pas eu d'influence sur ces améliorations.

4.1.3.2 Résultats sur les données réelles

Le Tableau 7 présente la stabilité de la sélection de variables et le taux d'erreur sur les classifieurs en résultant sur les jeux de données biopuces. D'une manière générale, les méthodes d'ensemble produisent des résultats soit autant soit plus stables que les sélections de variables single, au moins pour l'une des méthodes d'agrégation, à l'exception du t-score sur les données leucémie. En revanche, à la différence de ce que nous observons sur les données artificielles, le taux d'erreur est parfois augmenté par les méthodes d'ensemble.

Table 7: Stabilité et taux d'erreur en fonction de la méthode de sélection et du type d'agrégation dans la méthode d'ensemble, sur les données biopuces (résultats détaillés)

Données	Méthode d'aggrégation	t-score		SVM-RFE		Forêt aléatoire		Information mutuelle	
		Taux d'erreur	Stabilité	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité
Cancer côlon	Single	0.188	0.31	0.182	0.448	0.196	0.163	0.181	0.14
	Score moyen	0.305	0.327	0.203	**0.588**	0.179	0.206	0.217	0.149
	Rang moyen	0.215	0.277	0.203	0.494	0.199	0.163	0.209	0.149
	Stability selection	0.21	0.262	0.206	0.568	**0.177**	0.21	0.213	0.145
Leucémie	Single	0.049	0.322	0.044	0.525	**0.042**	0.22	0.05	0.263
	Score moyen	0.117	0.315	0.051	**0.581**	0.043	0.265	0.052	0.3
	Rang moyen	0.054	0.269	0.047	0.517	0.067	0.094	0.053	0.297
	Stability selection	0.058	0.246	0.047	0.565	0.046	0.269	0.049	0.294
BK Pawitan	Single	0.342	0.071	**0.283**	0.129	0.309	0.011	0.313	0.023
	Score moyen	0.328	0.085	**0.283**	**0.18**	0.32	0.012	0.343	0.024
	Rang moyen	0.344	0.065	0.298	0.095	0.324	0.015	0.335	0.024
	Stability selection	0.317	0.047	0.289	**0.18**	0.314	0.011	0.329	0.022
Cancer poumon	Single	**0.054**	0.515	0.084	0.377	0.06	0.342	0.061	0.372
	Score moyen	0.076	**0.536**	0.083	0.498	0.064	0.398	0.063	0.379
	Rang moyen	0.058	0.417	0.067	0.444	0.063	0.389	0.064	0.377
	Stability selection	0.058	0.445	0.082	0.487	0.063	0.394	0.064	0.376
BK Vijver	Single	0.382	0.254	0.377	0.159	0.359	0.078	0.36	0.077
	Score moyen	0.364	**0.345**	0.373	0.221	0.368	0.107	**0.357**	0.091
	Rang moyen	0.371	0.237	0.368	0.158	0.371	0.105	0.36	0.092
	Stability selection	0.374	0.237	0.376	0.215	0.371	0.106	0.359	0.088

Le t-score obtient la stabilité la plus élevée sur 2 des 5 jeux de données. Sur 4 jeux de données, sa stabilité est améliorée par les méthodes d'ensemble, mais uniquement avec l'agrégation par score moyen. Par ailleurs, sur les données cancer du côlon et cancer du poumon, cette amélioration se fait au prix d'un taux d'erreur largement augmenté. Sur les données leucémie, les méthodes d'ensemble dégradent sa stabilité et son taux d'erreur. En moyenne (Tableau 34 et Figure 35), la méthode d'ensemble avec agrégation par score moyen améliore la stabilité du t-test mais dégrade les performances de classification, et les autres méthodes d'agrégation dégradent la stabilité.

Table 8: Stabilité et taux d'erreur en fonction de la méthode de sélection et du type d'agrégation dans la méthode d'ensemble : en haut, moyenne pondérée sur l'ensemble des données biopuces, en bas, moyenne sur l'ensemble des données et des méthodes de sélection, par type d'agrégation

	t-score		**SVM-RFE**		**Forêt aléatoire**		**Info mutuelle**	
	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité	Taux d'erreur	Stabilité
Single	0.210	0.305	0.206	0.306	0.194	0.152	0.195	0.169
Score moyen	0.230	0.340	0.207	0.405	0.201	0.189	0.205	0.181
Rang moyen	0.210	0.256	0.200	0.320	0.209	0.159	0.204	0.180
Stability selection	0.206	0.253	0.209	0.395	0.201	0.189	0.202	0.177

	Moyenne	
	Taux d'erreur	Stabilité
Single	0.201	0.233
Score moyen	0.211	0.279
Rang moyen	0.206	0.229
Stability selection	0.205	0.253

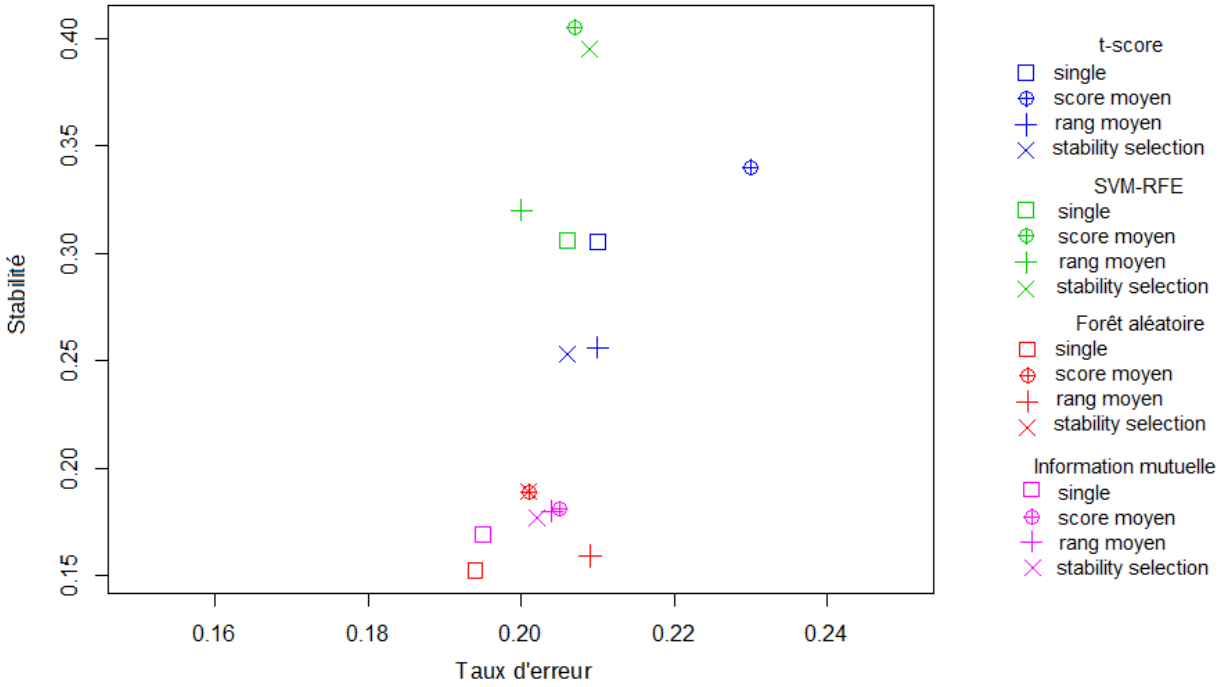

Figure 35: Moyenne des stabilités et des taux d'erreur en fonction de la méthode de sélection et du type d'ensemble.

Le SVM-RFE obtient la stabilité la plus élevée sur 3 des 5 jeux de données. Sur tous les jeux de données, sa stabilité est améliorée par les méthodes d'ensemble. L'amélioration la plus marquée est observée avec l'agrégation par score moyen, suivie de près par la *stability selection*. L'agrégation par rang moyen produit par contre des résultats moins intéressants : elle obtient les plus faibles gains de stabilité sur les données cancer du côlon et cancer du poumon, et conduit à

une stabilité inchangée voire dégradée sur les autres jeux de données. Contrairement au t-score, le taux d'erreur est généralement inchangé par les méthodes d'ensemble avec SVM-RFE, en dehors d'une augmentation de 10% sur les données cancer du côlon (quelle que soit la méthode d'agrégation), et d'une diminution de 20% sur les données cancer du poumon (uniquement pour l'agrégation par le rang moyen). En moyenne, les méthodes d'ensemble améliorent la stabilité du SVM-RFE sans dégrader les performances de classification. Ces gains de stabilité sont variables selon les méthodes d'agrégation : le score moyen permet d'obtenir le gain le plus important (+33%), suivi de près par la *stability selection* (+30%), alors que l'agrégation par rang moyen n'obtient qu'un gain de 5%.

Les forêts aléatoires obtiennent la stabilité la moins bonne sur 4 des 5 jeux de données, parmi lesquels elles sont à égalité avec l'information mutuelle sur les données Vijver. En revanche, elles obtiennent des taux d'erreur compétitifs (meilleur taux d'erreur sans ensemble sur les données Vijver et leucémie, meilleur taux d'erreur avec ensemble sur les données cancer du côlon), bien que les différences de taux d'erreur entre les différentes méthodes de sélection soient globalement faibles, à l'exception des données Pawitan. Comme pour SVM-RFE, leur stabilité est généralement améliorée par les méthodes d'ensemble. Les agrégations par score moyen et par *stability selection* améliorent la stabilité sur tous les jeux de données, l'agrégation par rang moyen est plus inconstante : amélioration similaire aux autres agrégations sur les données Vijver et cancer du poumon, amélioration plus importante sur les données Pawitan (où les forêts aléatoires ont une stabilité largement inférieure aux autres méthodes), pas d'amélioration sur les données cancer du colon, et une forte dégradation sur les données leucémie. En moyenne, les méthodes d'ensemble avec agrégation par score moyen ou par *stability selection* améliorent la stabilité des forêts aléatoires (+24%), pour un coût modeste sur le taux d'erreur (+3%), mais cette stabilité reste largement inférieure à celle du t-score et du SVM-RFE (-50%).

L'information mutuelle obtient des stabilités proches des forêts aléatoires, et la plus faible stabilité sur les données cancer du côlon. Cependant, comme les forêts aléatoires, elle obtient de bons taux d'erreur. Ses versions ensemble obtiennent une stabilité similaire (données cancer du côlon, Pawitan et cancer du poumon) ou modérément augmentée (données leucémie et Vijver), sans différence notable entre les méthodes d'agrégation. Les méthodes d'ensemble augmentent le

taux d'erreur sur les données cancer du côlon et Pawitan, également sans différence notable en fonction de la méthode d'agrégation. En moyenne, les méthodes d'ensemble augmentent très modestement (+7%) la stabilité de l'information mutuelle, avec un coût proche (+5%) sur le taux d'erreur.

En moyenne sur l'ensemble des données biopuces et des méthodes single (Tableau 34), les ensembles agrégés par score moyen permettent d'augmenter la stabilité de 20%, au prix d'une augmentation du taux d'erreur de 5%. L'apport le plus intéressant semble se situer au niveau de la sélection par SVM-RFE, dont la stabilité est augmentée en moyenne de 33% par l'ensemble avec agregation par score moyen, sans conséquence négative (en moyenne) sur la précision de classification.

4.1.4 Discussion et conclusion

Dans cette section, nous avons étudié l'influence des méthodes d'ensemble sur la stabilité de la sélection de variables et, dans une moindre mesure, le taux d'erreur du classifieur en résultant, sur des données biopuces et des données artificielles de dimensions semblables. Nous nous sommes focalisés sur la méthode d'agrégation car c'est une étape importante de la construction de l'ensemble, qui a notre connaissance a très peu été étudiée dans de telles conditions. Nous avons trouvé, comme dans (Haury et al., 2011), que le rang moyen obtient généralement une stabilité moindre que les autres méthodes d'agrégation, voire moindre que les méthodes utilisées sans ensemble. À cela nous avons ajouté que l'agrégation par score moyen obtient, à de rares exceptions près, la meilleure stabilité. L'agrégation par stability selection obtient des stabilités intermédiaires.

Nous avons également observé, dans certains cas, un compromis entre stabilité et taux d'erreur : une stabilité améliorée s'accompagne parfois d'un taux d'erreur dégradé. Des études précédentes sur la sélection de variables en général (Lausser et al., 2013) ou plus particulièrement sur la sélection par méthodes d'ensemble (Saeys et al., 2008) ont déjà suggéré un tel compromis problème-dépendant. Ici, nous trouvons de plus que la méthode d'agrégation a aussi un rôle dans ce compromis, puisque la stabilité et le taux d'erreur sont affectés différemment, selon les données, par les différentes méthodes d'agrégation. Ce compromis ne semble par contre pas s'appliquer à nos données artificielles : sur celles-ci, un meilleur taux

d'erreur s'est toujours accompagné d'une meilleure stabilité. Cette différence pourrait s'expliquer par des différences structurelles avec les données réelles (ces dernières sont vraisemblablement bien plus complexes et avec de plus nombreuses interactions), ou par un manque de diversité dans nos données artificielles : le compromis n'étant pas observé sur tous les jeux de données, peut-être que par chance il n'a pas lieu sur notre modèle artificiel de base et que ses variantes sont trop proches pour qu'un compromis y apparaisse.

Malgré ces différences de compromis erreur-stabilité entre les jeux de données que nous avons étudiés, un résultat apparaît constant : la méthode la plus stable en single a toujours pu voir sa stabilité améliorée par une méthode d'ensemble (par une ou plusieurs méthodes d'agrégation, le plus souvent au moins par l'agrégation par le score moyen), avec ou sans compromis sur la précision de classification.

4.2 Ensembles hybrides

Dans la section précédente, nous avons constaté que les méthodes de sélection, avec ou sans ensemble, à base de SVM-RFE et t-score sont alternativement les plus stables. Ce qui nous a amenés à nous demander s'il est possible de construire une méthode qui permettrait d'obtenir le meilleur des deux. Nous avons listé, toujours dans la section précédente, une topologie des techniques de génération de la diversité, puis nous avons choisi la technique habituelle (Yang et al., 2010) du rééchantillonnage multiple seul. Pour combiner SVM-RFE et t-score, la solution qui vient alors à l'esprit est de réaliser un ensemble en y ajoutant une autre technique de génération de diversité : l'hybridation de méthodes.

Si les ensembles hybrides sont utilisés depuis longtemps en classification (Dasarathy & Sheela, 1979 ; Dietterich, 2000 ; Tang et al., 2012), leur application à la sélection de variables est bien plus récente, et s'est focalisée jusqu'ici sur la précision de la classification et non la stabilité de la sélection (Awada et al., 2012), avec des résultats plutôt neutres (pas d'amélioration ni détérioration) sur cette précision, que l'ensemble soit réalisé au niveau de la sélection seule (Dittman et al., 2012), ou au niveau de blocs englobant sélection et classification (Bolón-Canedo et al., 2012). Les ensembles hybrides pourraient donc être intéressants si, sans changer le taux d'erreur, ils permettaient d'améliorer la stabilité.

La méthode d'agrégation choisie dans le cadre des ensembles hybrides est volontiers le

rang moyen (Dittman et al., 2012), car il n'est pas soumis à des plages de valeurs différentes selon les méthodes. Cependant, comme nous avons observé précédemment que les ensembles agrégés par rang moyen résultent en une stabilité non améliorée (voire diminuée), pour une précision non améliorée également, alors qu'au contraire les ensembles agrégés par score moyen permettent une amélioration notable de la stabilité, il nous semble pertinent de mettre en oeuvre une agrégation par score moyen. Afin de focaliser la méthode sur l'amélioration de la stabilité, nous avons, dans une partie des expériences, ajouté une pondération des scores de chaque rééchantillonnage par la stabilité obtenue par la méthode de sélection (sans ensemble) sur le rééchantillonnage.

4.2.1 Méthodes

Comme précédemment, les données ont été rééchantillonnées par bootstrap $B = 40$ fois. Sur la moitié de ces rééchantillonnages, un classement des variables par t-score a été réalisé. Sur l'autre moitié, un classement par SVM-RFE a été réalisé. On obtient ainsi deux groupes de $B/2 = 20$ scoring des variables, avec des scores sur des échelles de valeurs différentes entre ces deux groupes. Afin de remédier à cette incomparabilité des scores produits par les deux méthodes, ceux-ci ont été normalisés en les divisant par le score maximum : soit $s_{f_i,j}$ le score du gène f_i. dans le rééchantillonnage j, le score normalisé $s'_{f_i,j}$ est :

$$s'_{f_i,j} = \frac{s_{f_i,j}}{\max\limits_{i \in [1;D]} (s_{f_i,j})}$$

On obtient ainsi, pour chaque rééchantillonnage de l'ensemble, des scores compris entre 0 et 1. Pour obtenir le score final W_{f_i} du gène f_i sur l'ensemble, on utilise ensuite le score moyen comme décrit précédemment, mais avec trois variantes possibles :

- score moyen simple : W_{f_i} correspond à la moyenne des scores obtenus sur chacun des rééchantillonnages :

$$W_{f_i} = \frac{1}{B} \cdot \sum_{j=1}^{B} s'_{f_i,j}$$

- score moyen pondéré par la stabilité : W_{f_i} correspond à la moyenne des scores obtenus

sur chacun des rééchantillonnages, pondérés par la stabilité de chaque scoring, déterminée par une validation croisée imbriquée .

$$W_{f_i} = \frac{1}{B} \cdot \sum_{j=1}^{B} CW_{rel}(j) \cdot s'_{f_i,j}$$

où $CW_{rel}(j)$ correspond à la stabilité (*relative weighted consistency*) de la méthode de sélection single utilisée (donc, soit t-score, soit SVM-RFE) mesurée à l'intérieur du j^{ème} rééchantillonnage.

- score moyen pondéré par méthode : W_{f_i} correspond à la moyenne des scores obtenus sur chacun des rééchantillonnages, pondérés par un poids unique pour chaque méthode : λ pour le t-score et $(1-\lambda)$ pour SVM-RFE :

$$W_{f_i} = \frac{1}{B} \cdot \sum_{j=1}^{B} \lambda^{I_j(t)} \cdot (1-\lambda)^{I_j(RFE)} \cdot s'_{f_i,j}$$

où $I_j(t) = 1$ si la méthode de sélection utilisée pour calculer $s'_{f_i,j}$ est le t-score et 0 sinon, et $I_j(RFE) = 1$ si la méthode de sélection utilisée pour calculer $s'_{f_i,j}$ est SVM-RFE et 0 sinon.

4.2.2 Design expérimental

Les expériences réalisées sur ces méthodes ont été semblables à celles décrites en 4.2.1. Pour chacune de ces méthodes (t-score single et ensemble, SVM-RFE single et ensemble, ensemble hybride non pondéré, ensemble hybride pondéré par la stabilité, ensemble hybride pondéré par méthode), la stabilité de la sélection (*CWrel*) a été mesurée entre des échantillons sans recouvrement et le taux d'erreur de classification a été mesuré sur un classifieur LDA appliqué aux échantillons tests. Dans le cas de l'ensemble hybride pondéré par méthode, nous avons fait varier entre λ 0 et 1, par incrément de 0.05. Ainsi cette méthode recouvre également des variantes normalisées de l'ensemble t-score et de l'ensemble SVM-RFE, ainsi que l'ensemble hybride non pondéré (dans les cas où λ vaut respectivement 1, 0 et 0.5).

Les méthodes ont été testées sur les données artificielles NC100, avec *N=100*

observations, ainsi que sur les 5 jeux de données biopuces présentés précédemment.

4.2.3 Résultats

Le Tableau 9 présente la stabilité de la sélection et le taux d'erreur sur le classifieur en résultant sur les données *NC100*. Sur ces données, l'utilisation de la méthode d'ensemble permet de gagner en stabilité et précision pour la sélection par t-score, mais pas pour la sélection par SVM-RFE. L'ensemble hybride t-score + SVM-RFE obtient des performances semblables à celle de l'ensemble t-score, aussi bien sans pondération (poids identique donné aux runs t-score et aux runs SVM-RFE) qu'avec une pondération par la stabilité.

Table 9: Stabilité et taux d'erreur en fonction de la méthode de sélection sur les données *NC100*.

Méthode d'aggrégation	Taux d'erreur	Stabilité
t-score	0.312	0.137
SVM-RFE	0.308	0.123
ensemble t-score	0.267	0.192
ensemble SVM-RFE	0.313	0.126
ensemble hybride non-pondéré	0.263	0.196
ensemble hybride pondéré/stab	**0.261**	**0.200**

La Figure 36 présente l'évolution de la stabilité et du taux d'erreur dans un ensemble hybride t-score + SVM-RFE pondéré manuellement, quand λ évolue entre 0 (ce qui correspond à un ensemble SVM-RFE avec des scores normalisés) et 1 (ce qui correspond à un ensemble t-score avec des scores normalisés). On remarque que la stabilité de l'ensemble SVM-RFE normalisé est légèrement plus élevée que celle de l'ensemble SVM-RFE non normalisé (0.138 vs 0.123). L'ensemble SVM-RFE reste néanmoins nettement moins performant que l'ensemble t-score sur ces données. L'ajout de poids au t-score permet de rapidement améliorer la stabilité et diminuer le taux d'erreur, mais aucune pondération ne permet de dépasser notablement les performances de l'ensemble t-score simple. Stabilité et taux d'erreur semblent toutefois légèrement meilleurs autour de $\lambda \in [0.4;0.5]$, avec des valeurs alors très proches de celles obtenues par les sélections par ensemble hybride pondéré par la stabilité ou non pondéré (ce cas correspondant à $\lambda=0.5$).

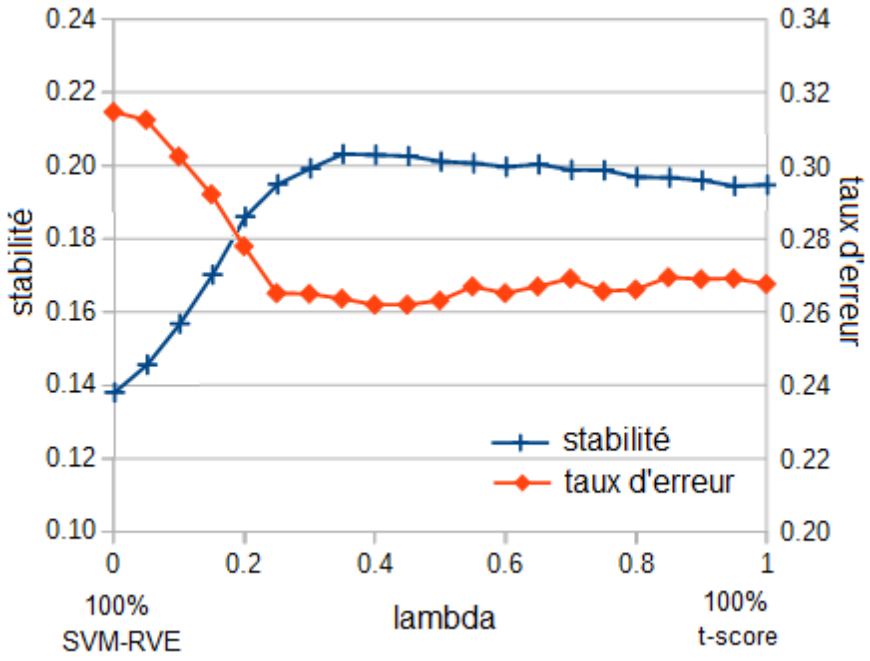

Figure 36: Stabilité et taux d'erreur de l'ensemble hybride t-score – SVM-RFE en fonction de la pondération λ, sur les données NC 100.

Le Tableau 10 présente la stabilité de la sélection de variables et le taux d'erreur sur les classifieurs en résultant sur les jeux de données biopuces. Les meilleures stabilités sont obtenues, sur tous les jeux de données, par les méthodes d'ensemble "simples" : comme précédemment, l'ensemble SVM-RFE obtient la meilleure stabilité sur les données cancer du côlon, leucémie et BK Pawitan, l'ensemble t-score sur les données cancer du poumon et BK Vijver. Il est plus difficile de dégager une tendance au niveau des taux d'erreur : le meilleur taux d'erreur est obtenu alternativement par chacune des méthodes non hybrides.

L'ensemble hybride t-score + SVM-RFE sans pondération obtient dans l'ensemble de mauvaises performances, avec le taux d'erreur le plus élevé sur les données cancer du côlon, le second taux d'erreur le plus élevé sur les données leucémie, pour une stabilité généralement équivalente ou légèrement inférieure à l'ensemble t-score. L'ensemble hybride pondéré par la stabilité obtient de meilleurs résultats, avec notamment :

- un taux d'erreur nettement moindre que l'ensemble hybride non pondéré sur les données cancer du côlon (26.3% vs 32.9%), le plaçant à mi-chemin entre l'ensemble t-test (31.2%) et l'ensemble SVM-RFE (20.5%)

- un taux d'erreur moindre que l'ensemble hybride non pondéré sur les données leucémie (7.5% vs 9.4%), le rapprochant cette fois plus de l'ensemble SVM-RFE (5.4%) que de l'ensemble t-test (11.1%), avec une stabilité supérieure aux ensembles t-score et hybride non pondéré (0.398 vs respectivement 0.331 et 0.344)

Table 10: Stabilité et taux d'erreur en fonction de la méthode de sélection sur les données biopuces

Données	Méthode d'agrégation	Taux d'erreur	Stabilité
Cancer côlon (Alon)	t-score	0.208	0.297
	SVM-RFE	0.188	0.442
	ensemble t-score	0.312	0.341
	ensemble SVM-RFE	0.205	0.591
	ensemble hybride non-pondéré	0.329	0.342
	ensemble hybride pondéré/stab	0.263	0.355
Leucémie (Golub)	t-score	0.050	0.314
	SVM-RFE	0.043	0.518
	ensemble t-score	0.111	0.331
	ensemble SVM-RFE	0.054	0.585
	ensemble hybride non-pondéré	0.094	0.344
	ensemble hybride pondéré/stab	0.075	0.398
BK Pawitan	t-score	0.342	0.073
	SVM-RFE	0.291	0.125
	ensemble t-score	0.326	0.092
	ensemble SVM-RFE	0.282	0.191
	ensemble hybride non-pondéré	0.302	0.076
	ensemble hybride pondéré/stab	0.309	0.109
Cancer poumon (CAMDA)	t-score	0.055	0.524
	SVM-RFE	0.082	0.382
	ensemble t-score	0.077	0.565
	ensemble SVM-RFE	0.077	0.521
	ensemble hybride non-pondéré	0.078	0.561
	ensemble hybride pondéré/stab	0.082	0.560
BK Vijver	t-score	0.380	0.259
	SVM-RFE	0.376	0.143
	ensemble t-score	0.366	0.368
	ensemble SVM-RFE	0.375	0.225
	ensemble hybride non-pondéré	0.372	0.345
	ensemble hybride pondéré/stab	0.371	0.353

Les tendances générales s'observent mieux dans le Tableau 11, qui présente les moyennes

des taux d'erreurs et stabilité par méthode sur l'ensemble des données : l'ensemble hybride non pondéré obtient une précision semblable à celle de l'ensemble t-score, avec une stabilité légèrement moindre (0.336 vs 0.346). L'ensemble hybride avec pondération par la stabilité obtient quant à lui un taux d'erreur légèrement meilleur que celui de l'ensemble t-score (22.1% vs 23%), avec une stabilité légèrement supérieure (0.36). Les deux types d'ensemble hybride obtiennent des stabilités supérieures aux méthodes single. Cependant, l'ensemble SVM-RFE reste en moyenne la meilleure méthode, avec à la fois un taux d'erreur moindre et une meilleure stabilité.

Table 11: Moyennes pondérées des stabilités et taux d'erreur en fonction de la méthode de sélection sur les données biopuces

Méthode d'agrégation	Taux d'erreur	Stabilité
t-score	0.211	0.298
SVM-RFE	0.208	0.315
ensemble t-score	0.230	0.346
ensemble SVM-RFE	0.206	0.422
ensemble hybride non-pondéré	0.228	0.336
ensemble hybride pondéré/stab	0.221	0.360

La Figure 37 présente l'évolution de la stabilité et du taux d'erreur dans un ensemble hybride t-score + SVM-RFE pondéré manuellement, quand λ évolue entre 0 et 1, sur les différents jeux de données biopuces. On constate que sur les données cancer du côlon et leucémie, l'évolution de la stabilité entre l'ensemble 100% SVM-RFE et l'ensemble 100% t-score est monotone : aucune valeur de lambda ne permet à l'ensemble hybride de dépasser la stabilité de l'un au l'autre. Sur les autres données, l'évolution de la stabilité n'est pas strictement monotone, et quelques points sortent de l'intervalle [stabilité SVM-RFE ; stabilité t-score], mais ces écarts sont très faibles et correspondent soit à une stabilité inférieure (cas des données BK Pawitan pour $\lambda \in [0.3;0.8]$), soit à des fluctuations ponctuelles (cas des données Vijver pour $\lambda = 0.05$ ou des données cancer du poumon pour $\lambda \in [0.45;0.55]$).

On observe un phénomène similaire avec l'évolution du taux d'erreur en fonction de λ : on réalise une transition progressive, monotone à quelques fluctuations près, de la performance d'un ensemble non hybride à l'autre, sans surclasser le meilleur des deux.

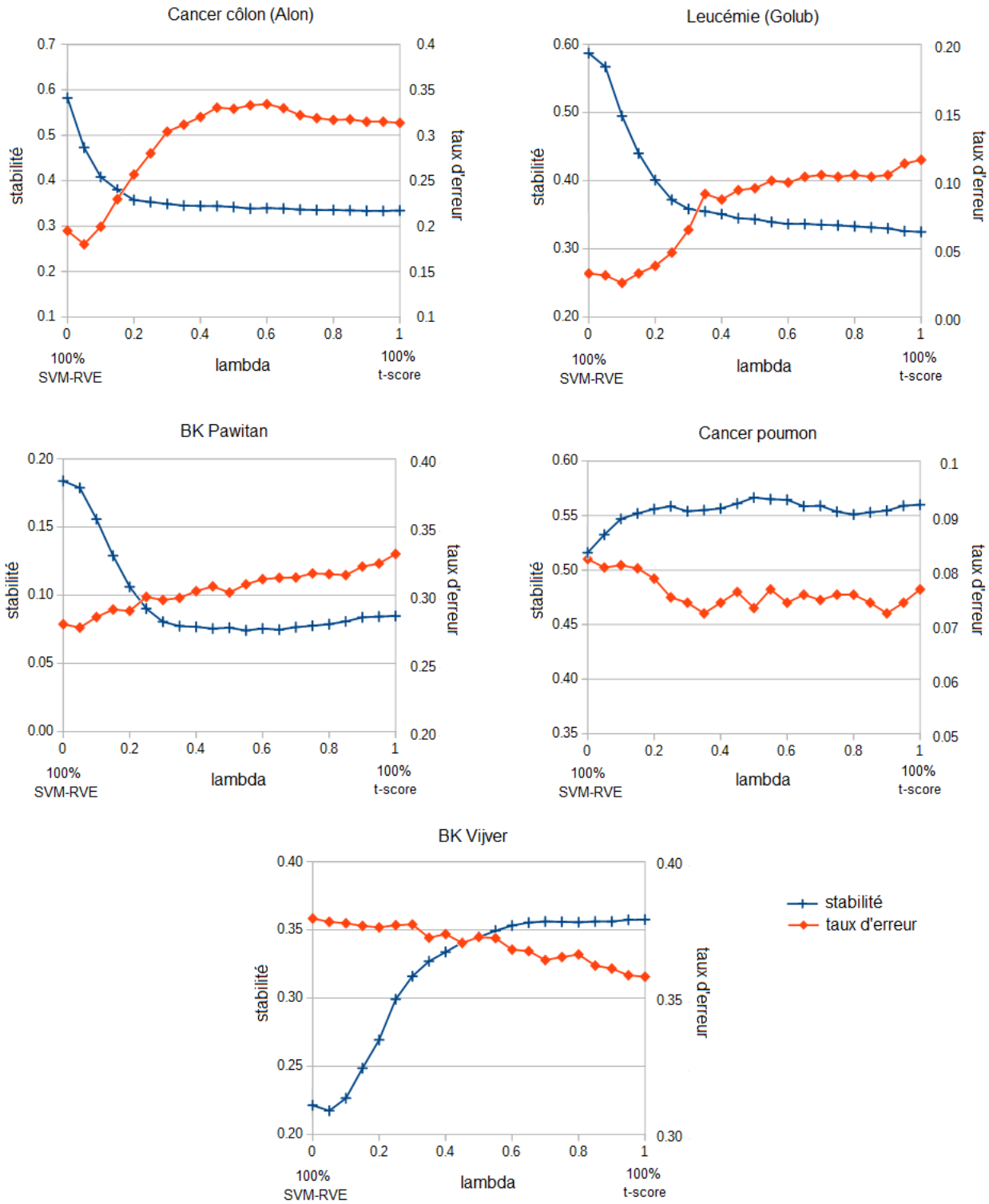

Figure 37: Stabilité et taux d'erreur de l'ensemble hybride t-score – SVM-RFE en fonction de la pondération λ, sur les données biopuces.

4.2.4 Discussion et conclusion

Dans cette section, nous avons cherché à combiner deux méthodes de sélection de variables performantes, t-score et SVM-RFE, dans des ensembles hybrides afin d'améliorer la stabilité de la sélection sans trop dégrader le taux d'erreur du classifieur en résultant, sur des données biopuces et des données artificielles de dimensions semblables. Nous avons choisi d'explorer les ensembles hybrides, car les ensembles non hybrides à base de l'une ou l'autre de ces méthodes se sont montrés prometteurs pour l'amélioration de la stabilité, mais de façon variable selon les données, et avec parfois un compromis entre stabilité et taux d'erreur. Nous avons observé que la sélection par ensemble hybride permet d'obtenir des performances de classification globalement intermédiaires à celles obtenues par des ensembles simples basés sur chacune des méthodes le constituant, ce qui est comparable aux résultats de la seule étude que nous avons trouvée sur de telles méthodes (Dittman et al., 2012). En revanche, la stabilité de la sélection par ensemble hybride non pondéré est légèrement moindre que la stabilité du moins bon ensemble. Ce défaut est corrigé dans la version pondérée par la stabilité, cependant les gains en stabilité et en précision par rapport à la méthode la moins performante restent modestes.

Les courbes d'évolution de la stabilité et du taux d'erreur en fonction de λ sur les ensembles hybrides pondérés manuellement permettent de mieux comprendre les résultats obtenus sur les ensembles hybrides non pondérés ou pondérés sur la stabilité. Aucune pondération ne permet de surpasser à la fois le point de "départ" ($\lambda = 0$) et le point "d'arrivée" ($\lambda = 1$), aussi bien en stabilité qu'en précision. L'absence de pondération ($\lambda = 0.5$) conduit à utiliser le point de milieu de courbe, qui n'est généralement pas le meilleur compromis, d'où les mauvaises performances de l'ensemble hybride non pondéré. La pondération par la stabilité permet de trouver un meilleur compromis, d'où l'amélioration de la stabilité par rapport à l'ensemble non pondéré et à l'ensemble t-score. Cependant, l'évolution rapide de la stabilité et du taux d'erreur dès de petites pondérations rend plus difficile cette recherche du bon poids, et l'évolution quasi monotone et en miroir des courbes de taux d'erreur et de stabilité font que la recherche de cette pondération, plutôt que la recherche d'un compromis optimal, consiste plutôt en la recherche du "bon côté" (ensemble SVM-RFE seul ou ensemble t-score seul) propre à chaque jeu de données.

Il est ainsi possible que les pondérations que nous avons utilisées soient trop faibles. Pour

améliorer cette méthode d'ensemble hybride, il nous semblerait donc pertinent de tester des pondérations plus sensibles, par exemple en mettant les poids au carré ou plus. Il serait même envisageable, puisqu'il s'agit plus du choix du "bon côté" que d'un compromis, de pondérer de façon binaire, en comparant sur chaque rééchantillonnage de la validation croisée la stabilité de chaque méthode (ici, t-score vs SVM-RFE, mais la méthode pourrait être étendue à d'autres méthodes de sélection identifiées comme ayant un bon compromis taux d'erreur – stabilité) et ne gardant pour la construction de l'ensemble que la méthode la plus stable sur le rééchantillonnage en question. Les travaux de (Cornuéjols & Martin, 2014) explorent une méthode de pondération intermédiaire entre ces deux extrêmes, dans le cadre d'un ensemble hybride avec une quantité bien plus importante de méthodes de sélection et une seule itération par méthode. La pertinence de chaque méthode est évaluée par une mesure de "surcorrélation" entre cette méthode et chaque autre, proche d'une mesure de stabilité ATI_{PA}. L'agrégation est ensuite réalisée en 2 étapes : d'abord, l'élimination des méthodes de sélections les moins pertinentes, puis parmi les méthodes retenues, une pondération de chaque méthode selon l'exponentielle de sa surcorrélation avec les autres méthodes non éliminées.

Ces perspectives sont encourageantes pour la sélection par ensemble hybride, qui mériterait d'être mieux explorée. Néanmoins, il semble probable que les seules améliorations qui pourront être apportées par les ensembles hybrides seront un meilleur compromis ou choix automatique entre les méthodes mises en oeuvre dans l'ensemble. Cela permettrait par exemple de se libérer du caractère problème-dépendant des performances des différentes méthodes de sélection, mais pas une amélioration de ces performances au-delà de ce qui peut être fait par les méthodes d'ensemble simples.

Par ailleurs, dans nos expériences, notre focus sur la stabilité n'a pas posé de problème au niveau du taux d'erreur du fait que le compromis erreur – stabilité que nous avons observé dans certains cas (et en particulier dans la section précédente) n'a pas concerné les deux méthodes (ensemble t-score et ensemble SVM-RFE) mises en balance dans notre ensemble hybride sur les jeux de données utilisés : sur chaque jeu de données, la méthode la plus stable était aussi celle obtenant la meilleure précision. La mise en oeuvre d'un ensemble hybride avec des pondérations plus sensibles telles que suggérées plus haut (poids au carré, ou a fortiori pondération binaire), dans le cas où il existe un compromis erreur – stabilité entre les méthodes utilisées dans

l'ensemble, risque d'aboutir à une sélection stable mais moins performante sur le taux d'erreur, et d'être alors peut-être moins intéressante qu'une pondération plus modérée telle que celle que nous avons utilisée.

Chapitre 5 :
Application des méthodes aux données DiOGenes : à la recherche de gènes prédictifs de la reprise de poids après un régime hypocalorique

Dans ce chapitre, nous mettons en application les méthodes de sélection et de classification abordées précédemment sur les données DiOGenes pour la recherche de gènes prédictifs de la reprise de poids suite à un régime hypocalorique. Après une rapide présentation générale de l'obésité, nous détaillerons les performances de classification et stabilité de différentes méthodes, puis proposerons une interprétation biologique des variables sélectionnées dans les modèles de prédiction.

5.1 Éléments cliniques et épidémiologies sur l'obésité

5.1.1 Définition

L'obésité correspond à une surcharge pondérale liée à une accumulation excessive et anormale de graisse corporelle. Sa définition clinique se base sur l'indice de masse corporelle (IMC, en anglais *BMI* pour *body mass index*), mesuré en kg/m² et dont la formule est :

$$\text{IMC} = \text{Poids}/\text{Taille}^2$$

où le poids est mesuré en kilogrammes et la taille en mètres. Une personne adulte est considérée comme obèse lorsque son IMC est supérieur ou égal à 30 kg/m². Pour un IMC supérieur ou égal à 40 kg/m², on parle d'obésité morbide ou encore d'obésité de classe 3 (Tableau 12).

L'IMC est une mesure simple, qui peut recouvrir des réalités différentes : deux individus ayant un même IMC peuvent avoir une répartition des graisses et une composition corporelle largement différentes (Wells et al., 2006). Cependant, c'est une mesure peu coûteuse et facile à obtenir, et donc idéale pour une première évaluation de la corpulence, ou une évaluation de la

prévalence de l'obésité au niveau de la population. De plus, malgré son lien imparfait avec la composition corporelle, des études suggèrent qu'il reste un bon marqueur de risque des complications liées à l'obésité, tel que le risque cardio-vasculaire (Wells, 2014).

Table 12: Classification en surpoids et sous-poids selon l'IMC, tiré de (World Obesity Federation / Policy & Prevention)

5.1.2 Épidémiologie

L'obésité est restée relativement peu fréquente jusqu'au siècle dernier (Haslam, 2007). Aux États-Unis, dès les années 1930, les compagnies d'assurance utilisent le poids comme données pour établir les primes, ayant identifié un lien entre un poids excessif et la mortalité précoce. Plus tard, (Breslow, 1952) observe que parmi les patients « bien portants » examinés à la *Boston Health Protection Clinic*, 1 sur 6 ont un surpoids de plus de 20%, et que parmi un groupe de travailleurs de force de San Francisco, 2 sur 5 ont un tel surpoids. Dans ce même papier, il fait le lien entre l'augmentation de la prévalence du surpoids et l'augmentation de la prévalence des maladies cardiovasculaires. Depuis, la prévalence du surpoids et de l'obésité a continué d'augmenter, particulièrement durant les dernières décennies, et touchant le monde entier

(Figures 38 et 39) : en 1997, l'OMS reconnaît l'obésité comme une épidémie mondiale (Caballero, 2007). En 2013, on estime que, dans le monde, 2.1 milliards de personnes sont en surpoids ou obèses (comparé à 857 millions en 1980), parmi lesquelles on compte 671 millions d'obèses (Ng et al., 2014).

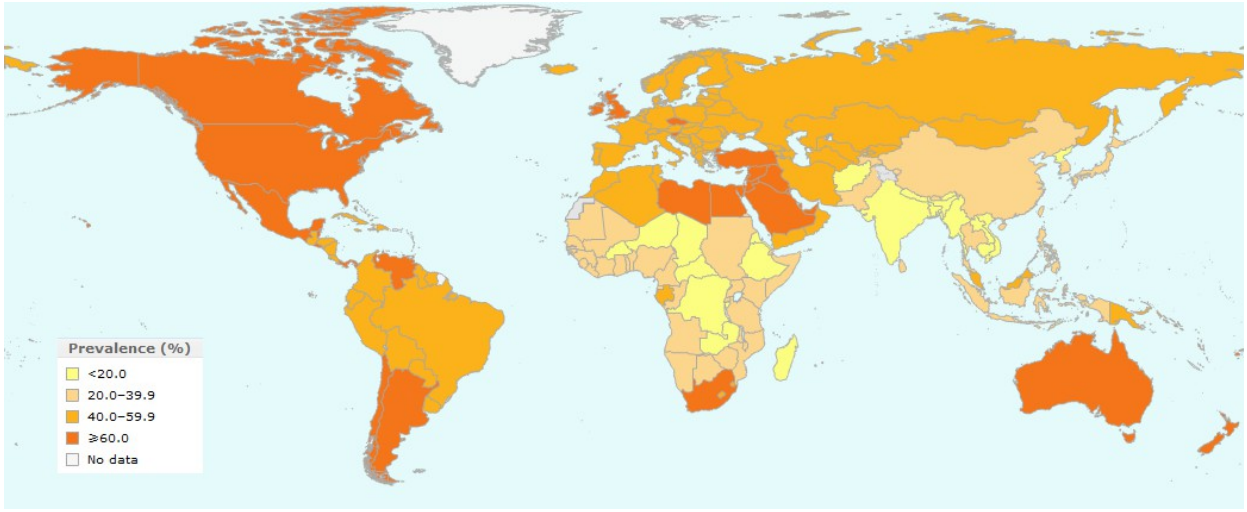

Figure 38: Prévalence du surpoids dans le monde (données OMS 2008).

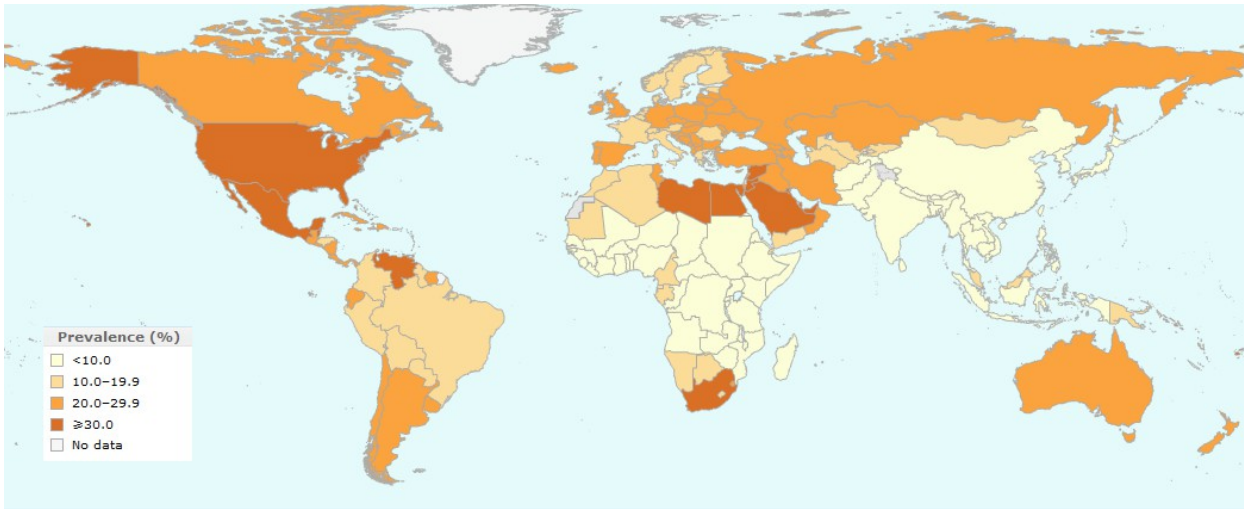

Figure 39: Prévalence de l'obésité dans le monde (données OMS 2008). Visualisation : http://gamapserver.who.int/gho/interactive_charts/ncd/risk_factors/overweight_obesity/atlas.html

Les causes de l'obésité sont multiples. Dans de (très) rares cas, l'obésité peut être secondaire à une autre pathologie (hypothyroïdie, syndrome de Cushing, syndrome de Prader-Willi...), iatrogène (corticoïdes, psychotropes...) ou liée à une mutation génétique unique

(obésité dite monogénique, environ 200 cas répertoriés en 2006 (Mutch & Clément, 2006)). Mais dans la très large majorité des cas, il s'agit d'une obésité dite primaire ou essentielle, ou encore polygénique, pour laquelle les causes sont hautement multifactorielles, associant un terrain génétique favorable à une multitude d'influences environnementales qui se sont développées dans nos sociétés modernes : sédentarité, mauvaises habitudes alimentaires, stress (Kiecolt-Glaser et al., 2014), pression sociale, etc. (Figure 40). Dans le reste de ce chapitre, nous ne nous intéresserons qu'à ce type d'obésité.

Figure 40: Influence de l'environnement sur le développement de l'obésité. Tiré de (Mutch & Clément, 2006).

5.1.3 Complications

L'obésité est capable d'entraîner de très nombreuses complications, qui en font toute la gravité et dont beaucoup peuvent réduire l'espérance de vie : en 2010, on estime que dans le monde l'obésité et le surpoids sont responsables de 3.4 millions de décès par an (Ng et al., 2014). Les complications de l'obésité sont notamment cardiovasculaires (hypertension, insuffisance cardiaque...), métaboliques (diabètes, dyslipidémie...), respiratoires (syndrome restrictif, apnée du sommeil...), ostéoarticulaires, et sont détaillées dans le Tableau 13.

Table 13: Principales complications de l'obésité. Tiré de (Basdevant, 2006).

5.1.4 Tissu adipeux et obésité

Le tissu adipeux a longtemps été considéré comme un simple lieu de stockage et de mobilisation de réserves énergétiques sous la forme de lipides, et en effet il constitue le principal lieu de stockage énergétique chez l'homme. La mise en évidence, dans les années 1990, de la sécrétion de diverses hormones par le tissu adipeux, en tête desquelles « l'hormone de la satiété », la leptine (Neill, 2010), et un an plus tard l'adiponectine (Scherer et al., 1995), va radicalement faire évoluer cette vision.

Le tissu adipeux est maintenant considéré comme un organe sécrétagogue à part entière, sécrétant de très nombreuses protéines, peptides et facteurs lipidiques agissant de façon autocrine, paracrine ou endocrine, contrôlant non seulement son activité, mais également le métabolisme énergétique dans sa globalité (Pégorier, 2007). Le Tableau 14 liste les principales adipokines sécrétées par le tissu adipeux et leurs rôles biologiques.

Table 14: Sources et fonctions des principales adipokines. Tiré de (Ouchi et al., 2011).

L'obésité entraîne une augmentation de la taille (hypertrophie) et du nombre (hyperplasie) des adipocytes. L'hypertrophie est le mécanisme généralement prépondérant, et l'hyperplasie est d'autant plus marquée que l'obésité est importante : elle correspond à la production de nouveaux adipocytes, via le recrutement de préadipocytes, quand les adipocytes existants ne sont plus capables de grossir pour stocker l'excès de lipides (Hirtsch & Batchelor, 1976).

Figure 41: Altérations du tissu adipeux dans l'obésité. Tiré de (Ouchi et al., 2011).

L'obésité modifie également la sécrétion des adipokines, avec en particulier une augmentation de la sécrétion de molécules pro-inflammatoires et une diminution de la production de molécules anti-inflammatoires telles que l'adiponectine (Figure 41), aboutissant à une inflammation chronique de bas grade (Itoh et al., 2011 ; Ouchi et al., 2011), qui jouerait un rôle important dans l'insulino-résistance (Divoux, 2010). Il a par ailleurs été montré que, chez des patients obèses, la perte de poids induite par une restriction calorique diminue l'expression des gènes liés à l'inflammation dans le tissu adipeux sous-cutané (Clément et al., 2004). Le tissu adipeux est également le siège d'une accumulation macrophagique, qui n'est que partiellement réversible à la perte de poids (Aron-Wisnewsky et al., 2009).

Le tissu adipeux de l'obèse présente une fibrose à des degrés variés, dont les causes sont encore discutées : expression de facteurs fibrotiques induite par l'hypoxie du tissu adipeux, et/ou conséquence de l'état inflammatoire chronique du tissu adipeux (Abdennour, 2013). Cette fibrose pourrait être un marqueur du dysfonctionnement métabolique du tissu adipeux (Khan et al., 2009), et elle serait un facteur prédictif de la perte de poids après chirurgie bariatrique (Divoux et al., 2010).

5.1.5 Principes thérapeutiques

Le traitement de l'obésité repose sur des mesures diététiques (diminution des apports

caloriques, éducation alimentaire), associées à la mise en place d'une activité physique modérée régulière, et au dépistage et au traitement des complications éventuelles, tout en assurant un soutien psychologique approprié (visant notamment à prévenir ou traiter d'éventuels troubles du comportement alimentaire). La chirurgie bariatrique est envisagée en dernier recours, après échec de la prise en charge hygiénodiététique initiale. En 2010, 25000 opérations de chirurgie bariatriques sont réalisées chaque année en France (Bout et al., 2010).

Les mesures diététiques et/ou l'activité physique permettent généralement d'obtenir une bonne perte de poids à court terme, en revanche le maintien à long terme de ces résultats est plus difficile (Barte et al., 2010). Différents facteurs favorisant le maintien de la perte de poids ont été identifiés, notamment l'activité physique, les habitudes alimentaires, la durée de maintenance de la perte de poids, et la balance des macronutriments (répartition des glucides, protides et lipides) (Capel et al., 2008 ; Delbridge et al., 2009). Il est largement admis que le poids et la composition corporelle sont influencés par une composante génétique, en revanche le rôle des déterminants génétiques dans le maintien de la perte de poids reste méconnu (Mutch et al., 2011). C'est dans ce contexte que se situe notre analyse des données DiOGenes pour la prédiction de la reprise de poids, suite à un régime, à partir de données d'expression du tissu adipeux sous-cutané mesurées au début du régime.

5.2 Design expérimental

Les données DiOGenes ont été présentées en introduction (1.1.3). Pour mémoire, elles consistent en 13078 variables (puce Agilent 4x44K whole human genome) mesurées chez 40 patientes (20 dans chaque groupe) à J0 d'un régime hypocalorique de 8 semaines, et la variable d'intérêt est la reprise ou non de poids après 6 mois de régime normocalorique, parmi les patientes ayant initialement perdu plus de 8% de leur poids.

Les méthodes de sélection utilisées ont été les principales méthodes utilisées dans cette thèse, à l'exception des ensembles hybrides : le t-score, le CAT-score, SVM-RFE, l'information mutuelle, ReliefF, CoGO, l'ensemble t-score et l'ensemble SVM-RFE (les ensembles ont été réalisés avec agrégation par score moyen). Les 100 meilleures variables ont été retenues. Sur ces différentes sélections, des classifieurs LDA, SVM avec kernel linéaire et radial, kNN, et forêt aléatoire ont été entraînés. La stabilité de chaque méthode de sélection a été évaluée sur 50

paires jeu d'apprentissage – jeu test, chacun constitué de 20 observations respectant la proportion des classes du jeu de données initial. Le taux d'erreur de classification a été évalué sur ces mêmes paires.

À partir des différentes sélections réalisées, nous avons mesuré à quelle fréquence chaque gène était sélectionné, et nous avons extrait une liste des gènes sélectionnés par la moitié ou plus des méthodes de sélection. Ce seuil a été choisi afin d'une part de se focaliser sur des gènes sélectionnés de manière relativement stable, et d'autre part d'avoir une liste de gènes de taille suffisamment modérée pour être interprétable. Les informations (fonctions, profils d'expression…) relatives à ces gènes ont ensuite été recherchées en utilisant le portail S.O.U.R.C.E. de Stanford[2] (Demeter et al., 2007), l'outil Gene du NCBI[3], UniProt[4] (UniProt Consortium, 2014), genenames.org, et BioGPS[5] (Wu et al., 2009).

5.3 Résultats

5.3.1 Stabilité et performance de prédiction

Le Tableau 15 présente les stabilités et taux d'erreur pour chaque classifieur en fonction de la méthode de sélection utilisée. L'ensemble SVM-RFE, méthode la plus stable, confirme ses mauvaises performances de classification observées sur DiOGenes dans le chapitre 3 sur les classifieurs ajoutés ici. Le t-score obtient le meilleur taux d'erreur moyen, avec une stabilité correcte : seuls SVM-RFE et les méthodes d'ensembles obtiennent une meilleure stabilité, mais ils produisent des classifieurs moins précis. Le classifieur le plus performant est la forêt aléatoire, avec un taux d'erreur moyen de 12.5%. 3 combinaisons obtiennent un taux d'erreur inférieur à 10% : t-score + forêt aléatoire, t-score + kNN, et ReliefF + forêt aléatoire. Si l'on augmente la taille de l'échantillon d'apprentissage à 30 observations, l'erreur de classification avec une forêt aléatoire construite sur une sélection t-score descend à 7.7%. Si en plus on double la taille de la sélection, ces mêmes méthodes obtiennent un taux d'erreur de 8.8%. Si on double la taille de la sélection en conservant cette fois une taille d'échantillon d'apprentissage de 20 observations, la stabilité de la sélection passe de 0.124 à 0.178 (et le taux d'erreur est inchangé).

2 http://smd.princeton.edu/cgi-bin/source/sourceResult
3 http://www.ncbi.nlm.nih.gov/gene
4 http://www.uniprot.org
5 http://biogps.gnf.org

Table 15: Stabilité et taux d'erreur moyen en fonction de la méthode de sélection et du classifieur.

Méthode de sélection	Stabilité CWrel	Taux d'erreur par classifieur					
		LDA	RF	SVM l	SVM r	kNN	Moyenne
t-score	0.124	0.123	0.094	0.110	0.123	0.099	0.110
SVM-RFE	0.156	0.180	0.174	0.145	0.186	0.290	0.195
CoGO	0.126	0.113	0.128	0.132	0.158	0.129	0.132
Information mutuelle	0.054	0.135	0.105	0.120	0.140	0.290	0.158
CAT-score	0.045	0.125	0.115	0.123	0.128	0.124	0.123
ReliefF	0.060	0.140	0.090	0.120	0.155	0.121	0.125
ensemble t-score	0.147	0.203	0.130	0.149	0.157	0.151	0.158
ensemble SVM-RFE	0.257	0.215	0.163	0.208	0.234	0.335	0.231
Moyenne	0.121	0.154	0.125	0.138	0.160	0.192	0.154

5.3.2 Gènes communs aux différentes sélections

Le Tableau 16 liste les 30 sondes sélectionnées par au moins la moitié des 8 méthodes de sélection, ainsi que les 29 gènes et protéines correspondants, une de ces sondes n'étant pas annotée. On remarque qu'aucun gène n'est sélectionné 8 fois, et seuls 2 gènes sont sélectionnés 7 fois et 6 fois. Le score correspond au score moyen (en valeur absolue) normalisé sur 7 méthodes de sélection (toutes sauf CoGO, qui produit les mêmes scores que le t-score mais uniquement sur les centroïdes des clusters de gènes). Le score maximum possible est de 1, le score minimum possible est de 0, le score maximum moyen observé est de 0.72. Les gènes surexprimés dans le groupe de patients qui ont repris du poids ont leur score coloré en rouge, et les gènes sous-exprimés en vert.

Table 16: Liste des gènes sélectionnés par au moins la moitié des méthodes de sélection.

# sél	Score	ProbeID	GeneID	Gène	Protéine	Cytoband
7	↓0.69	A_23_P135494	25932	CLIC4	chloride intracellular channel 4	1p36.11
7	↓0.57	A_23_P42802	9601	PDIA4	protein disulfide isomerase family A, member 4	7q35
6	↓0.72	A_23_P114740	3078	CFHR1	complement factor H-related 1	1q32
6	↑0.72	A_24_P375360	122592	KRT18P6	keratin 18 pseudogene 6	14q13.2
5	↑0.64	A_23_P165180	8625	RFXANK	regulatory factor X-associated ankyrin-containing protein	19p12
5	↑0.62	A_23_P307525	122416	ANKRD9	ankyrin repeat domain 9	14q32.31
5	↑0.61	A_23_P138881	89	ACTN3	actinin, alpha 3	11q13.1

# sél	Score	ProbeID	GeneID	Gène	Protéine	Cytoband
5	↑0.57	A_23_P101332	65095	KRI1	KRI1 homolog (S. cerevisiae)	19p13.2
5	↓0.53	A_23_P215634	3486	IGFBP3	insulin-like growth factor binding protein 3	7p13-p12
5	↑0.52	A_24_P70183	4629	MYH11	myosin, heavy chain 11, smooth muscle	16p13.11
5	↓0.51	A_23_P5761	4780	NFE2L2	nuclear factor (erythroid-derived 2)-like 2	2q31
5	↑0.50	A_23_P5115	57418	WDR18	WD repeat domain 18	19p13.3
5	↓0.47	A_32_P95397	3688	ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	10p11.2
4	↓0.59	A_23_P408095	11034	DSTN	destrin (actin depolymerizing factor)	20p12.1
4	↑0.59	A_23_P340890	90326	THAP3	THAP domain containing, apoptosis associated protein 3	1p36.31
4	↓0.57	A_23_P14946	8720	MBTPS1	membrane-bound transcription factor peptidase, site 1	16q24
4	↑0.55	A_23_P133799	89953	KLC4	kinesin light chain 4	6p21.1
4	↓0.55	A_32_P202057	10971	YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta polypeptide	2p25.1
4	↑0.54	A_23_P97442	84065	TMEM222	transmembrane protein 222	1p36.11
4	↓0.53	A_23_P208090	4152	MBD1	methyl-CpG binding domain protein 1	18q21
4	↑0.52	A_23_P130886	84954	MPND	MPN domain containing	19p13.3
4	↓0.52	A_23_P126474	6746	SSR2	signal sequence receptor, beta (translocon-associated protein beta)	1q21-q23
4	↑0.52	A_23_P26476	84219	WDR24	WD repeat domain 24	16p13.3
4	↓0.51	A_23_P259054	57231	SNX14	sorting nexin 14	6q14.3
4	↓0.50	A_23_P88381	8650	NUMB	numb homolog (Drosophila)	14q24.3
4	↓0.49	A_23_P60816	10015	PDCD6IP	programmed cell death 6 interacting protein	3p22.3
4	↑0.48	A_24_P204084				

# sél	Score	ProbeID	GeneID	Gène	Protéine	Cytoband
4	↓0.46	A_23_P200560	998	CDC42	cell division cycle 42 (GTP binding protein, 25kDa)	1p36.1
4	↓0.43	A_23_P18078	5918	RARRES 1	retinoic acid receptor responder (tazarotene induced) 1	3q25.32
4	↓0.42	A_23_P207020	56270	WDR45B	WD repeat domain 45B	17q25.3

Le Tableau 17 présente les informations connues dans la littérature sur ces gènes. On remarque en particulier parmi les gènes sélectionnés la présence de :

- CLIC4, sous-exprimé dans le groupe repreneur, qui pourrait jouer un rôle dans la régulation du volume des organelles intracellulaires (Suginta et al., 2001) et qui joue un rôle dans l'angiogenèse (Chalothorn et al., 2009 ; Ulmasov et al., 2009). Une molécule proche de CLIC4, CLIC5, a été liée à la prise de poids chez des souris soumises à un régime hypercalorique (Bradford et al., 2010).
- CFHR1, sous-exprimé dans le groupe repreneur, qui code une protéine de la famille du facteur H, qui a un rôle dans la régulation du complément (Skerka et al., 2013).
- RFXANK, surexprimé dans le groupe repreneur, qui a un rôle dans la production du complexe majeur d'histocompatibilité de classe II (CMH II) (Nagarajan et al., 1999).
- IGFBP3, sous-exprimé dans le groupe repreneur, qui aurait un rôle de stimulation de l'apoptose cellulaire (Rajah et al., 2002) et de diminution de la prolifération cellulaire sur des cellules cancéreuses (Muhlbradt et al., 2009).
- NFE2L2, sous-exprimé dans le groupe repreneur, qui est impliqué dans un pathway anti-inflammatoire et anti-apoptotique dans les cellules endothéliales (Kim et al., 2012).
- MBTPS1, sous-exprimé dans le groupe repreneur, qui a un rôle dans la production des lysosomes (Marschner et al., 2011). Une altération des fonctions lysosomales peut conduire à une accumulation cellulaire de lipides (Walkey & Vanier, 2009).
- WDR24, NUMB et PDCD6IP, qui auraient des rôles, encore mal déterminés, dans la prolifération, la croissance ou l'apoptose cellulaire. WDR24, surexprimé chez les repreneurs, favoriserait la croissance cellulaire via mTORC1. NUMB, sous-exprimé

chez les repreneurs, régulerait la croissance cellulaire (rôle suppresseur de tumeur).

- DSTN et CDC42, sous-exprimé dans le groupe repreneur, qui auraient notamment un rôle dans l'entretien du cytosquelette (Xu et al., 2014 ; Kim et al., 2014)

Table 17: Littérature relative aux gènes listés dans le Tableau 16.

Rang	Gène	Protéine	Littérature
↓1	CLIC4	chloride intracellular channel 4	Chloride channels are a diverse group of proteins that regulate fundamental cellular processes including stabilization of cell membrane potential, transepithelial transport, maintenance of intracellular pH, and regulation of cell volume. Chloride intracellular channel 4 (CLIC4) protein is a member of the p64 family; the gene is expressed in many tissues and exhibits an intracellular vesicular pattern in Panc-1 cells (pancreatic cancer cells)
↓2	PDIA4	protein disulfide isomerase family A, member 4	A study implicated ERp72 (alias of PDIA4) in the signal transduction pathway for priming human neutrophils (Weisbart, 1992)
↓3	CFHR1	complement factor H-related 1	Encodes a secreted protein belonging to the complement factor H protein family. Involved in complement regulation. Mutations in this gene are associated with an increased risk of atypical hemolytic-uremic syndrome.
↑4	KRT18P6	keratin 18 pseudogene 6	
↑5	RFXANK	regulatory factor X-associated ankyrin-containing protein	Major histocompatibility (MHC) class II molecules are transmembrane proteins that have a central role in development and control of the immune system. The protein encoded by this gene forms a complex that binds to the X box motif of certain MHC class II gene promoters and activates their transcription. This protein contains ankyrin repeats involved in protein-protein interactions. Mutations in this gene have been linked to bare lymphocyte syndrome type II, complementation group B.
↑6	ANKRD9	ankyrin repeat domain 9	

Rang	Gène	Protéine	Littérature
↑7	ACTN3	alpha-actinin-3	This gene encodes a member of the alpha-actin binding protein gene family. The encoded protein is primarily expressed in skeletal muscle and functions as a structural component of sarcomeric Z line. About 18% of the world population lack a functional actn3 due to a stop codon polymorphism at position 577. The absence of a functional actn3 expression is not correlated with a disease state.
↑8	KRI1	KRI1 homolog (S. cerevisiae)	This gene overlaps with the gene for cysteine endopeptidase AUT-like 4 in a head-to-tail orientation
↓9	IGFBP3	insulin-like growth factor binding protein 3	Member of the insulin-like growth factor binding protein (IGFBP) family. The protein forms a ternary complex with insulin-like growth factor acid-labile subunit (IGFALS) and either insulin-like growth factor (IGF) I or II. In this form, it circulates in the plasma, prolonging the half-life of IGFs and altering their interaction with cell surface receptors.
↑10	MYH11	myosin heavy chain 11	The protein encoded by this gene is a smooth muscle myosin belonging to the myosin heavy chain family. Defects in MYH11 are the cause of familial aortic aneurysm thoracic type 4 (Zhu et al., 2006)
↓11	NFE2L2	nuclear factor erythroid 2-related factor 2	Encodes a transcription factor which is a member of a small family of basic leucine zipper (bZIP) proteins. The encoded transcription factor regulates genes which contain antioxidant response elements (ARE) in their promoters; many of these genes encode proteins involved in response to injury and inflammation which includes the production of free radicals. Important for the coordinated up-regulation of genes in response to oxidative stress.
↑12	WDR18	WD repeat-containing protein 18	May play a role during development. Functions as a component of the Five Friends of Methylated CHTOP (5FMC) complex; the 5FMC complex is recruited to ZNF148 by methylated CHTOP, leading to transactivation of ZNF148 target genes. WDR18 works together with TopBP1 to promote DNA damage checkpoint signaling (Yan & Willis, 2013).

Rang	Gène	Protéine	Littérature
↓13	ITGB1	integrin, beta 1 (fibronectin receptor, subunit beta, glycoprotein IIa, CD29)	Integrins are heterodimeric proteins made up of alpha and beta subunits. Integrin family members are membrane receptors involved in cell adhesion and recognition in a variety of processes including embryogenesis, hemostasis, tissue repair, immune response and metastatic diffusion of tumor cells. This gene encodes a beta subunit.
↓14	DSTN	destrin (actin depolymerizing factor)	Belongs to the actin-binding proteins ADF family. This family of proteins is responsible for enhancing the turnover rate of actin in vivo. Role in the regulation of the actin cytoskeleton (Xu et al., 2014). This gene encodes the actin depolymerizing protein that severs actin filaments (F-actin) and binds to actin monomers (G-actin).
↑15	THAP3	THAP domain containing, apoptosis associated protein 3	Component of a THAP1/THAP3-HCFC1-OGT complex that is required for the regulation of the transcriptional activity of RRM1 (Ribonucleoside-diphosphate reductase large subunit).
↓16	MBTPS1	membrane-bound transcription factor peptidase, site 1	Serine protease that catalyzes the first step in the proteolytic activation of the sterol regulatory element-binding proteins (SREBPs). Other known substrates are BDNF, GNPTAB and ATF6.Mediates the protein cleavage of GNPTAB into subunit alpha and beta, thereby participating in biogenesis of lysosomes.
↑17	KLC4	kinesin light chain 4	Kinesin is a microtubule-associated force-producing protein that may play a role in organelle transport. The light chain may function in coupling of cargo to the heavy chain or in the modulation of its ATPase activity
↓18	YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta polypeptide	Adapter protein implicated in the regulation of a large spectrum of both general and specialized signaling pathways. Binds to a large number of partners, usually by recognition of a phosphoserine or phosphothreonine motif. Binding generally results in the modulation of the activity of the binding partner. Negatively regulates the kinase activity of PDPK1. Up-regulated in the lumbar spinal cord from patients with sporadic amyotrophic lateral sclerosis (ALS) compared with controls, with highest levels of expression in individuals with predominant lower motor neuron involvement.
↑19	TMEM222	transmembrane protein 222	

Rang	Gène	Protéine	Littérature
↓20	MBD1	methyl-CpG binding domain protein 1	The protein encoded by this gene is a member of a family of nuclear proteins related by the presence of a methyl-CpG binding domain (MBD). These proteins are capable of binding specifically to methylated DNA, and some members can also repress transcription from methylated gene promoters.
↑21	MPND	MPN domain containing	Probable protease
↓22	SSR2	signal sequence receptor, beta (translocon-associated protein subunit beta)	TRAP proteins are part of a complex whose function is to bind calcium to the ER membrane and thereby regulate the retention of ER resident proteins
↑23	WDR24	WD repeat domain 24	Component of the GATOR2 subcomplex, made of MIOS, SEC13, SEH1L, WDR24 and WDR59 . GATOR2 inhibits GATOR1. GATOR1 inhibits mTORC1. mTORC1 promotes cell growth (Bar-Peled et al., 2013).
↓24	SNX14	sorting nexin 14	May be involved in several stages of intracellular trafficking.
↓25	NUMB	numb homolog (Drosophila)	Plays a role in the process of neurogenesis. Required throughout embryonic neurogenesis to maintain neural progenitor cells, also called radial glial cells (RGCs), by allowing their daughter cells to choose progenitor over neuronal cell fate. Also involved postnatally in the subventricular zone (SVZ) neurogenesis by regulating SVZ neuroblasts survival and ependymal wall integrity. May also mediate local repair of brain ventricular wall damage. Interacts with RALBP1 in a complex also containing EPN1 and TFAP2A during interphase and mitosis. Tumor suppressive role of NUMB isoform 1 in esophageal squamous cell carcinoma (Hong et al., 2014). RBM10 mutations identified in lung cancer cells disrupt NUMB splicing regulation to promote cell growth (Bechara et al., 2013).
↓26	PDCD6IP	programmed cell death 6 interacting protein	May play a role in the regulation of both apoptosis and cell proliferation. Bind to the products of the PDCD6 gene, a required protein in apoptosis. PDCD6IP insertion/deletion polymorphism may be associated with non-small cell lung cancer risk (Liu et al., 2014).
↑27			

Rang	Gène	Protéine	Littérature
↓28	CDC42	cell division cycle 42 (GTP binding protein, 25kDa)	The protein encoded by this gene is a small GTPase of the Rho-subfamily, which regulates signaling pathways that control diverse cellular functions including cell morphology, migration, endocytosis and cell cycle progression. Involved in epithelial cell polarization processes. Important for reorganization of actin cytoskeleton (Kim et al., 2014). Regulates the bipolar attachment of spindle microtubules to kinetochores before chromosome congression in metaphase. Plays a role in the extension and maintenance of the formation of thin, actin-rich surface projections called filopodia. Overexpressed in NK and myeloid cells, and more generally in leukocytes.
↓29	RARRES1	retinoic acid receptor responder protein 1 (tazarotene induced gene 1 protein) (TIG1)	Inhibitor of the cytoplasmic carboxypeptidase AGBL2, may regulate the alpha-tubulin tyrosination cycle. The expression of this gene is upregulated by tazarotene as well as by retinoic acid receptors. The expression of this gene is found to be downregulated in prostate cancer, hepatocellular carcinoma (Chen et al., 2014), and acute leukemia (Wang et al., 2014). Correlated with shorter survival in patients with inflammatory breast cancer (Wang et al., 2013b). Linked to fibrogenesis in liver, but it is still unclear if it is an independent and superior driver of fibrogenesis or a kind of defense mechanism to interfere with fibrogenic actions (Teufel et al., 2012).
↓30	WDR45B	WD repeat domain 45B, WD repeat domain phosphoinositide-interacting protein 3 / WIPI49-like protein	Ubiquitously expressed. Highly expressed in heart, skeletal muscle and pancreas. Part of a group of WD-repeat protein Interacting with PhosphoInosides (WIPI). WIPI genes were found aberrantly expressed in a variety of matched tumor tissues, which could be linked to starvation-induced autophagy. WIPI proteins may share an evolutionary conserved function in autophagy and that autophagic capacity may be compromised in human cancers (Proikas-Cezanne et al., 2004).

5.4 Discussion et conclusion

Dans cette section, les classifieurs que nous avons construits pour prédire la reprise de poids à partir de données d'expression du tissu adipeux avant régime ont atteint une précision supérieure à 90%, et même 92% en utilisant un échantillon d'apprentissage aussi grand que possible. Il s'agit donc d'une tâche de classification relativement « facile », pourtant la stabilité

des sélections est mauvaise : à taux d'erreur de classification comparable, la stabilité de la sélection sur les données cancer du poumon mesurée dans le Chapitre 4 est plus du double. Cela pourrait témoigner d'un grand nombre de variables pertinentes, toutefois redondantes pour la prédiction : augmenter la taille de la sélection de 100 à 200 variables ne permet pas d'améliorer les performances de classification, mais augmente la stabilité de 40% (alors que la mesure de stabilité est ajustée sur la taille des sélections).

La maintenance à long terme de la réduction de poids est un challenge important dans la prise en charge de l'obésité. À ce titre, la connaissance des mécanismes impliqués dans le contrôle du poids peut être utile pour prévenir la reprise de poids (Márquez-Quiñones et al., 2010). L'étude de la littérature relative aux gènes les plus souvent sélectionnés par les différentes méthodes de sélection confirme, pour bon nombre d'entre eux, une pertinence possible ou probable.

On note tout d'abord un certain nombre de gènes impliqués dans l'inflammation : RFXANK, producteur de CMH II, est surexprimé dans le groupe repreneur, NFE2L2 et CFHR1, probablement anti-inflammatoires ou régulateurs de l'inflammation, y sont sous-exprimés. On observe également une expression différentielle de WDR24, IGFBP3, NUMB et PDCD6IP, impliqués dans la prolifération et/ou l'apoptose cellulaire. Une surexpression des pathways de l'inflammation et de la prolifération cellulaire a été associée à la reprise de poids par (Márquez-Quiñones et al., 2010), et une augmentation de l'apoptose du tissu adipeux sous-cutané a été associée au maintien de la perte de poids (Mutch et al., 2011). Par ailleurs, l'inflammation favoriserait la prolifération des préadipocytes (Keophiphath et al., 2009).

MBTPS1 est sous-exprimé dans le groupe repreneur et a un rôle dans la production des lysosomes : cela pourrait témoigner d'une certaine dysfonction lysosomale chez les repreneurs, entraînant une difficulté à mobiliser les lipides. La sous-expression de DSTN et CDC42 suggère quant à elle une altération de l'entretien du cytosquelette chez les repreneurs. Le cytosquelette aurait une influence sur l'adipogenèse : une inhibition du cytosquelette par cytochalasine D ou blebbistatine entraîne une augmentation de l'adipogenèse (Schiller et al., 2013), et la fonction adipocitaire est affectée par des facteurs mécaniques, via la tension du cytosquelette (Pellegrinelli et al., 2014). Cela pourrait également rentrer dans le cadre d'une altération plus générale des adipocytes : il a été mis en évidence, chez les sujets qui reprennent du poids, une

surexpression de gènes codant pour des protéines ribosomales qui pourrait témoigner d'une dérégulation de l'adipogenèse (Márquez-Quiñones et al., 2010). On note également une sous-expression de RARRES1, qui pourrait être lié à la fibrose et/ou à la prolifération cellulaire.

En conclusion, le transcriptome du tissu adipeux sous-cutané semble être un relativement bon prédicteur du maintien ou de la reprise de poids. Bien qu'il soit difficile, vue la forte instabilité de la sélection, d'identifier précisément quels gènes sont les plus pertinents pour la prédiction, et *a fortiori* quels gènes sont une cause directe ou indirecte ou de simples marqueurs de risque, les gènes communs entre les différentes sélections sont souvent impliqués dans des grandes fonctions déjà connues pour jouer un rôle dans l'obésité : inflammation, prolifération cellulaire, fibrose. L'efficacité des classifieurs semble donc avoir des bases biologiques cohérentes. Pour aller plus loin dans la prédiction, il serait intéressant de mettre en compétition ces modèles avec d'autres modèles reposant sur des données cliniques et biologiques, et surtout de réaliser des modèles combinant données transcriptomique et données clinico-biologiques, pour déterminer si les données transcriptomiques apportent un plus dans la précision. De tels modèles ont précédemment été réalisés pour la prédiction de la perte de poids suite à un régime (Temanni, 2009), avec des résultats supérieurs aux modèles utilisant uniquement un des deux types de données.

Chapitre 6 :
Conclusion générale

L'arrivée des technologies haut débit a transformé la recherche biomédicale. Quand nous n'étions pas capables de mesurer, en quelques manipulations, l'ensemble du transcriptome, du génome, ou du profil lipidique, etc., la recherche était centrée sur les hypothèses : avant d'étudier le lien entre un gène et un phénotype, il était indispensable d'avoir préalablement une suspicion de ce lien, afin de réaliser les mesures pertinentes ciblées sur ce gène (ou sur un nombre limité de gènes d'intérêt). Des technologies telles que les biopuces ont, pour ainsi dire, inversé ce raisonnement : on mesure d'abord (tout !), et on extrait l'information ensuite.

Ce nouveau mode de raisonnement a permis de nombreuses avancées, par exemple pour la classification de différents types de tumeurs (Anglesio et al., 2008), pour la prédiction des résultats cliniques (Shi et al., 2010), ou la détection précoce de maladies (Chon & Lancaster, 2011 ; Mok et al, 2001). Plus généralement, on assiste à l'apparition d'une « médecine génomique » (Manolio & Green, 2014), dans laquelle l'analyse de données haute dimension (génomiques, transcriptomiques...) prend une importance croissante pour la compréhension biologique des maladies, la science médicale et l'efficacité des soins (Table 18), et permet le développement d'une « médecine personnalisée » (Hamburg & Collins, 2010). Mais ces données haute dimension posent un problème qui était autrefois inhabituel : le nombre de variables surpasse très largement le nombre d'observations.

Sur des problèmes dans lesquels l'information utile n'est pas concentrée sur quelques gènes très significatifs, il est possible de réaliser des signatures exportables sur d'autres jeux de données traitant du même problème. Mais si l'on construit un classifieur à partir de chacun de ces jeux de données, les gènes sélectionnés seront très variables d'un jeu de données à l'autre, et pas toujours meilleurs qu'une sélection aléatoire de taille suffisamment grande (Lauss et al., 2010 ; Venet et al., 2011). Il est donc délicat d'interpréter la (ou plutôt, les) sélection(s). Pour avoir une meilleure confiance dans les gènes sélectionnés, il semble important que la sélection

soit stable, et c'est l'objectif que nous avons poursuivi dans ce travail.

Table 18: Recherches génomiques relatives aux maladies dans trois domaines de recherche du National Human Genome Research Institute (NHRI). Tiré de (Manolio & Green, 2014).

Dans un premier temps, nous avons étudié dans quelle mesure les différentes caractéristiques des jeux de données peuvent influencer la stabilité de la sélection. Il en est ressorti que tous les aspects étudiés (nombre d'observations N, nombre de variables D, seuil de sélection, distribution de la significativité des variables, difficulté du problème de classification sous-jacent) ont une influence, mais qu'en particulier le facteur le plus important et modifiable est le ratio N/D.

Suite à cela, nous avons proposé dans le Chapitre 3 une approche originale pour réduire le nombre de variables par une présélection, en amont d'une méthode de sélection habituelle (ici le t-score), à partir de données *a priori* issues de Gene Ontology. Les résultats ont

malheureusement été décevants, ce qui pourrait être lié à l'action conjuguée des imprécisions de GO, de la pertinence imparfaite de regrouper les gènes par leurs annotations et de ne garder qu'un représentant unique par groupe, et des simplifications que nous avons dû réaliser pour produire une procédure raisonnablement rapide (en particulier, le choix d'un nombre arbitraire de clusters). Il serait probablement intéressant de réessayer cette méthode dans quelques années, quand les annotations GO (ou d'autres bases de connaissances telles que KEGG) se seront encore enrichies et que les performances computationnelles permettront de réaliser une version de cette méthode non (ou moins) simplifiée. Cependant, de nombreuses autres approches combinant données observées et données *a priori* ont été tentées, et ont conduit à des résultats mitigés : les signatures obtenues ont été considérées comme plus interprétables car constituées de gènes mieux connectés sur le réseau fonctionnel utilisé (Haury et al., 2010), mais ni la stabilité ni le taux d'erreur ne sont améliorés (Cun & Fröhlich, 2012 ; Staiger et al., 2012).

Dans le Chapitre 4, nous sommes revenus à des méthodes plus classiques, en nous intéressant cette fois à la stabilité de la sélection par méthodes d'ensemble. Nous avons alors observé que les performances de ces méthodes sont grandement influencées par la méthode d'agrégation choisie, l'agrégation par score moyen étant généralement et assez nettement la meilleure sur les données utilisées. En revanche, comme d'autres études, nous avons constaté dans plusieurs cas un compromis entre taux d'erreur et stabilité. Ce compromis a d'ailleurs conduit certains auteurs à proposer un score d'évaluation de la performance d'un classifieur combinant stabilité de la sélection et précision du classifieur (Davis et al., 2006). Suite aux méthodes d'ensemble simples, nous avons testé une méthode d'ensemble hybride, combinant t-score et SVM-RFE, mais celle-ci n'a fourni que des performances intermédiaires entre ensemble t-score et ensemble SVM-RFE. Au vu de nos résultats, les ensembles hybrides ne semblent pas en mesure d'apporter une meilleure stabilité ou précision que les ensembles simples. Mais nos analyses se sont limitées aux ensembles t-score et SVM-RFE, et cela mériterait d'être exploré avec plus de méthodes de sélection, et peut-être également en mettant en oeuvre une agrégation par *stability selection*, qui garantirait que les différentes méthodes de sélection contribuent également : l'agrégation par score moyen, bien adaptée au cas d'un ensemble à une méthode, n'est finalement peut-être pas aussi appropriée dans le cas des ensembles hybrides.

Enfin, nous avons appliqué dans le Chapitre 5 les différentes méthodes étudiées à un jeu

de données transcriptomiques du laboratoire, pour la prédiction de la reprise de poids après un régime réussi chez des patientes obèses, à partir du transcriptome du tissu adipeux mesuré au début du régime. Les performances de classification obtenues sont bonnes (précision >90%), la stabilité de la sélection beaucoup moins (~0.10-0.15), ce qui pourrait rendre compte du caractère multicausal de l'obésité : de multiples grandes fonctions sont impliquées, et au niveau de chacune un grand nombre de gènes ont des fonctions proches ou redondantes, et ont donc une pertinence équivalente. L'étude de la littérature relative aux gènes communs entre les sélections a souvent retrouvé une implication dans des mécanismes connus pour être en lien avec l'obésité, notamment l'inflammation, la prolifération cellulaire et la fibrose. Il semble cependant impossible de faire la part entre causes directes, causes indirectes, ou simples marqueurs de risque sur une sélection si instable. Mais est-il vraiment réaliste d'espérer obtenir une sélection stable sur de si petites données ?

S'il est peut-être possible d'améliorer encore un peu la stabilité via des méthodes basées uniquement sur le jeu de données biopuces étudié, les progrès réalisables sont limités par le "mur" du rapport N/D. Et ce d'autant plus que, sur de très petits échantillons, les méthodes de sélection (et même de classification) les plus simples sont constamment parmi les meilleures (Haury, 2012) : il y a peu d'information comparé à toutes les variables à traiter, ainsi les méthodes plus complexes font trop d'erreurs pour être meilleures. C'est par exemple ce que l'on observe sur les couvertures de Markov, qui reposent sur la réalisation d'un très grand nombre de tests d'indépendance conditionnelle dont beaucoup sont erronés sur de petits échantillons (Dernoncourt et al., 2011). Pour franchir cette limite, les méthodes basées sur l'enrichissement à l'aide de données *a priori* semblent prometteuses, mais les résultats sont pour l'instant décevants. Il est vraisemblable qu'ils s'amélioreront au fur et à mesure que les bases de connaissances s'enrichiront et se corrigeront, mais on peut se demander à quel point ces méthodes seront efficaces tant que les bases de connaissances utilisées ne contiendront pas déjà... une bonne partie des connaissances qu'on cherche à extraire des données. La combinaison de plusieurs jeux de données traitant d'un problème très proche ou identique, que nous n'avons ici qu'évoquée, pourrait également s'avérer efficace, par la simple augmentation de N, mais avec une utilité limitée précisément par la nécessité de disposer de plusieurs jeux de données très proches.

Cette difficulté à obtenir une bonne stabilité est d'autant plus problématique qu'elle

s'accompagne souvent d'un compromis avec le taux d'erreur, qui a été patent dans certaines de nos expériences sur les méthodes d'ensemble, et que l'on retrouve d'une manière plus générale dans la littérature (Han & Yu, 2012 ; Lausser et al. 2013 ; Sayes et al., 2008). Certains sont allés jusqu'à proposer une métrique d'évaluation de modèle combinant stabilité et précision (Davis et al., 2006). Mais est-ce vraiment pertinent ?

Comme indiqué en introduction, on voudrait une sélection stable afin d'avoir confiance en la pertinence des variables sélectionnées, afin de pouvoir se dire que les gènes utilisés dans le modèle sont "les bons". Cependant, la précision du classifieur est encore plus essentielle : une sélection, aussi stable soit-elle, qui produit un mauvais classifieur est selon toute vraisemblance moins intéressante qu'une sélection moins stable mais donnant de bonnes performances de classification.

Dans ce travail, malgré une vigilance sur la précision de la classification, nous nous sommes fortement focalisés sur la stabilité. *A posteriori*, il nous apparaît important de plus insister sur la précision, même (surtout !) dans les travaux centrés sur l'amélioration de la stabilité. L'objectif premier de la sélection et de la classification, s'il faut choisir, doit être la précision de la classification. La stabilité de la sélection n'intervient alors que secondairement, pour estimer la robustesse de l'extraction de connaissances qui peut être faite à partir des modèles, et pour estimer si les gènes de notre sélection ont une chance d'être directement associés à la variable à expliquer (Woodward, 2010) ou s'ils sont plutôt quelques-uns parmi les membres des grands pathways liés cette variable. Ce qui est sans doute le cas des gènes que nous avons isolés sur les données DiOGenes, qui sont remarquablement instables mais semblent pourtant, d'après la précision de la classification et leur littérature, appartenir à des pathways pertinents.

Références

Abdennour, M. (2013). Nouvelles Méthodes pour le diagnostic de pathologies associées à l'inflammation systémique des tissus adipeux et aux pathologies hépatiques, chez les patients obèses. Thèse de doctorat, Université Pierre et Marie Curie.

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398. doi: 10.1093/bioinformatics/btp630.

Agilent Technologies (2009). *Agilent Feature Extraction Software (v10.7) Reference Guide*.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the USA*, 96(12), 6745–6750. doi: 10.1073/pnas.96.12.6745

Anglesio, M.S., Arnold, J.M., George, J., Tinker, A.V., Tothill, R., Waddell, N., Simms, L., Locandro, B., Fereday, S., Traficante, N., Russell, P., Sharma, R., Birrer, M. J., deFazio, A., Chenevix-Trench, G., Bowtell, D.D.L. (2008). Mutation of ERBB2 provides a novel alternative mechanism for the ubiquitous activation of RAS-MAPK in ovarian serous low malignant potential tumors. *Molecular Cancer Research*, 6(11), 1678–1690. doi: 10.1158/1541-7786.MCR-08-0193

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25-9. doi: 10.1038/75556

Aron-Wisnewsky, J., Tordjman, J., Poitou, C., Darakhshan, F., Hugol, D., Basdevant, A., Aissat, A., Guerre-Millo, M., Clément, K. (2009). Human adipose tissue macrophages: m1 and m2 cell surface markers in subcutaneous and omental depots and after weight loss. *Journal of Clinical Endocrinology & Metabolism*, 94(11), 4619-4623. doi: 10.1210/jc.2009-0925

Awada, W., Khoshgoftaar, T., Dittman, D., Wald, R., Napolitano, A. (2012). A Review of the Stability of Feature Selection Techniques for Bioinformatics Data. *IEEE 13th International Conference on Information Reuse and Integration (IRI)*, 356-363. doi: 10.1109/IRI.2012.6303031

Bar-Peled, L., Chantranupong, L., Cherniack, A.D., Chen, W.W., Ottina, K.A., Grabiner, B.C., Spear, E.D., Carter, S.L., Meyerson, M., Sabatini, D.M. (2013). A Tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. *Science*, 340(6136), 1100-1106. doi: 10.1126/science.1232044

Barte, J.C.M., Ter Bogt, N.C.W., Bogers, R.P., Teixeira, P.J., Blissmer, B., Mori, T.A., Bemelmans, W.J. (2010). Maintenance of weight loss after lifestyle interventions for overweight and obesity, a systematic review. *Obesity Reviews*, 11(12), 899–906. doi: 10.1111/j.1467-789X.2010.00740.x

Basdevant, A. (2006). L'obésité : origines et conséquences d'une épidémie. *Comptes Rendus Biologies*, 329(8), 562-569. doi: 10.1016/j.crvi.2006.03.018

Bechara, E.G., Sebestyén, E., Bernardis, I., Eyras, E., Valcárcel, J. (2013). RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. *Molecular Cell*, 52(5), 720-33. doi: 10.1016/j.molcel.2013.11.010

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America,* 98, 13790–13795. doi:10.1073/pnas.191502998

Bradford, E.M., Miller, M.L., Prasad, V., Nieman, M.L., Gawenis, L.R., Berryman, M., Lorenz, J.N., Tso, P., Shull, G.E. (2010) CLIC5 mutant mice are resistant to diet-induced obesity and exhibit gastric hemorrhaging and increased susceptibility to torpor. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, 298, R1531-42. doi: 10.1152/ajpregu.00849.2009

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A. (2012). An ensemble of filters and classifiers for microarray data classification. *Pattern Recognition*, 45(1), 531–539. doi: 10.1016/j.patcog.2011.06.006

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J.M, Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282, 111-135. doi: 10.1016/j.ins.2014.05.042

Boulesteix, A.-L., Strobl, C., Augustin, T., & Daumer, M. (2008). Evaluating microarray-based classifiers: An overview. *Cancer Informatics*, 6, 77–97

Boulesteix, A.-L. & Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 59, 556–568

Bout, B. (2010). L'organisation de la recherche et ses perspectives en matière de prévention

et de traitement de l'obésité. *Office parlementaire d'évaluation des choix scientifiques et technologiques*.

Braga-Neto, U. M., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20, 374–380. doi: 10.1093/bioinformatics/btg419

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

Breslow, L. (1952). Public health aspects of weight control. *Am J Public Health*, 42, 1116-20

Brown, G., Wyatt, J., Harris, R., Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Journal of Information Fusion*, 6(1):5–20. doi: 10.1016/j.inffus.2004.04.004

Burges, C.J.C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2, 121–167.

Caballero, B. (2007). The Global Epidemic of Obesity: An Overview. *Epidemiologic Reviews*, 29(1), 1-5. doi: 10.1093/epirev/mxm012

Capel, F., Viguerie, N., Vega, N., Dejean, S., Arner, P., Klimcakova, E., Martinez, J.A., Saris, W.H., Holst, C., Taylor, M., Oppert, J.M., Sørensen, T.I., Clément, K., Vidal, H., Langin, D. (2008). Contribution of energy restriction and macronutrient composition to changes in adipose tissue gene expression during dietary weight-loss programs in obese women. *The Journal of Clinical Endocrinology & Metabolism*, 93(11), 4315–4322. doi: 10.1210/jc.2008-0814

Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(Database issue):D685-90. doi: 10.1093/nar/gkq1039

Chalothorn, D., Zhang, H., Smith, J.E., Edwards, J.C., Faber, J.E. (2009). Chloride intracellular channel-4 is a determinant of native collateral formation in skeletal muscle and brain. *Circulation Research*, 105, 89–98. doi: 10.1161/CIRCRESAHA.109.197145

Chen, Y., Zhao, Y. (2008). A novel ensemble of classifiers for microarray data classification. Applied Soft Computing, 8(4), 1664–1669. doi: 10.1016/j.asoc.2008.01.006

Chen, X.H., Wu, W.G., Ding, J. (2014). Aberrant TIG1 methylation associated with its decreased expression and clinicopathological significance in hepatocellular carcinoma. *Tumor Biology*, 35(2), 967-71. doi: 10.1007/s13277-013-1129-9

Chon, H.S., Lancaster, J.M. (2011). Microarray-based gene expression studies in ovarian

cancer. Cancer Control, 18(1), 8–15.

Chow, T. W S; Huang, D. (2005). Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information. *Neural Networks, IEEE Transactions on*, 16(1), 213–224 doi: 10.1109/TNN.2004.841414

Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* 32, 490 – 495. doi:10.1038/ng1031

Clément, K., Viguerie, N., Poitou, C., Carette, C., Pelloux, V., Curat, C. A., Sicard, A., Rome, S., Benis, A., Zucker, J.-D., Vidal, H., Laville, M., Barsh, G. S., Basdevant, A., Stich, V., Cancello, R., Langin, D. (2004) Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. *The FASEB Journal*, 18, 1657–1669. doi: 10.1096/fj.04-2204com

Cornuéjols, A., Miclet, L., Kodratoff, Y. (2002). Le boosting d'un algorithme d'apprentissage. In *Apprentissage Artificiel: Concepts et algorithmes*. Eyrolles. pp 354-362

Cornuéjols, A. & Martin, C. (2014). Une méthode d'ensemble en apprentissage non supervisé quand on ne connaît rien sur la performance des experts ? In *Proc. AAFD-2014 (Journées Apprentissage Artificiel et Fouille de Données)*, Université Paris 13, France

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

Cover, T.; Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on* vol.13, no.1, pp.21,27, January 1967. doi:10.1109/TIT.1967.1053964

Cun, Y., Fröhlich, F.H. (2012). Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*, 13:69. doi:10.1186/1471-2105-13-69

Dasarathy, B.V., Sheela, B.V. (1979). A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 708–713. doi: 10.1109/PROC.1979.11321

Davis, C.A., Gerick, F., Hintermair, V., Friedel, C.C., Fundel, K., Küffner, R., Zimmer, R. (2006). Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19), 2356-2363. doi: 10.1093/bioinformatics/btl400

Delbridge, E.A., Prendergast, L.A., Pritchard, J.E., Proietto, J. (2009). One-year weight maintenance after significant weight loss in healthy overweight and obese subjects: does diet composition matter? *American Journal of Clinical Nutrition*, 90(5), 1203–14. doi: 10.3945/ajcn.2008.27209

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., Matese,

J.C., Nitzberg, M., Wymore, F., Zachariah, Z.K., Brown, P.O., Sherlock, G., Ball, C.A. (2007). The stanford microarray database: implementation of new analysis tools and open source release of software. *Nucleic Acids Research*, 35, D766–D770. doi: 10.1093/nar/gkl1019

Dernoncourt, D., Hanczar, B., Zucker, J.D. (2011). An Empirical Analysis of Markov Blanket Filters for Feature Selection on Microarray Data. *Proceedings of the Fifth International Workshop of Machine Learning in Systems Biology*, 19–23.

Dernoncourt, D., Hanczar, B., Zucker, J.D. (2014). Analysis of feature selection stability on high dimension and small sample data. *Computational Statistics & Data Analysis*, 71(C), 681-693. doi:10.1016/j.csda.2013.07.012

Dernoncourt, D., Hanczar, B., Zucker, J.D. (2014b). Stability of Ensemble Feature Selection on High-Dimension and Low-Sample Size Data: Influence of the Aggregation Method. *Proceedings of the 3nd International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 325-330.

Dettling, M., Bühlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics*, 12(9), 1061-1069. doi: 10.1093/bioinformatics/btf867

Dettling, M., Bühlmann, P. (2004). Finding Predictive Gene Groups from Microarray Data. *Journal of Multivariate Analysis*, 90(1), 106-131. doi: 10.1016/j.jmva.2004.02.012

Díaz-Uriarte, R. & Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. 7:3. doi: 10.1186/1471-2105-7-3

Dietterich, T.G. (2000). Ensemble methods in machine learning. *Proceedings of the first International Workshop on Multiple classifier systems, Lecture Notes in Computer Science*, 1857, 1–15. doi: 10.1007/3-540-45014-9_1

Dittman, D., Khoshgoftaar, T., Wald, R., Napolitano, A. (2011). Random Forest: A Reliable Tool For Patient Response Prediction. *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 289–296. doi: 10.1109/BIBMW.2011.6112389

Dittman, D., Khoshgoftaar, T., Wald, R., Napolitano, A. (2012). Determining the Number of Iterations Appropriate for Ensemble Gene Selection on Microarray Data. *11th International Conference on Machine Learning and Applications (ICMLA)*, 1, 82–85 doi: 10.1109/ICMLA.2012.23

Divoux, A. (2010). Origines, rôles et conséquences de l'inflammation du tissu adipeux chez le sujet obèse : de nouvelles hypothèses. Thèse de doctorat, Université Pierre et Marie Curie.

Divoux, A., Tordjman, J, Lacasa, D., Veyrie, N., Hugol, D., Aissat, A., Basdevant, A., Guerre-Millo, M., Poitou, C., Zucker, J.D., Bedossa, P., Clément, K. (2010). Fibrosis in human adipose tissue: composition, distribution, and link with lipid metabolism and fat mass loss. *Diabetes*, 59(11), 2817–2825. doi: 10.2337/db10-0585

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97, 77–87.

Dunne K., Cunningham P. & Azuaje F. (2002). Solutions to instability problems with sequential wrapper-based approaches to feature selection. *Machine Learning*, (TCD-CS-2002-28), 1–22

Ein-Dor, L., Zuk, O., & Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103, 5923–5928.

Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7, 179-188. doi:10.1111/j.1469-1809.1936.tb02137.x

Frank, A., & Asuncion, A. (2010). UCI machine learning repository.

Freund, Y., Shapire, R.E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. doi: 10.1006/jcss.1997.1504

Fröhlich, H., Speer, N., Poustka, A., Beißbarth, B. (2007). GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products BMC Bioinformatics, 8:166. doi:10.1186/1471-2105-8-166

Fu S. (2010). Efficient Learning of Markov Blanket and Markov Blanket Classifier. Thèse de doctorat, École Polytechnique de Montréal, Canada, 2010.

Fury, W., Batliwalla, F., Gregersen, P.K., Li, W. (2006). Overlapping Probabilities of Top Ranking Gene Lists, Hypergeometric Distribution, and Stringency of Gene Selection Criterion. *Proceedings of 28th Annual International Conference of the Engineering in Medicine and Biology Society*, IEEE, 5531–5534. doi:10.1109/IEMBS.2006.260828

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4, 1–58.

Gene Ontology (2014). Extended GO Relations. http://www.geneontology.org/GO.ontology-ext.relations.shtml

Genuer, R., Poggi, J.M., Tuleau-Malot, C. (2010). Variable Selection using Random Forests.

Pattern Recognition Letters, 31(14), 2225-2236. doi: 10.1016/j.patrec.2010.03.014

Goldmann, T., Gonzalez, J.S. (2000). DNA-printing: utilization of a standard inkjet printer for the transfer of nucleic acids to solid supports. *Journal of Biochemical and Biophysical Methods*, 42(3):105-110. doi: 10.1016/S0165-022X(99)00049-4

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537. doi: 10.1126/science.286.5439.531

Gomase VS1, Tagore S, Kale KV (2008). Microarray: an approach for current drug targets. *Curr Drug Metab*. 2008 Mar;9(3):221-31. doi:10.2174/138920008783884795

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46, 389–422.

Hamburg, M.A., Collins, F.S. (2010). The Path to Personalized Medicine. *New England Journal of Medicine*, 363, 301-304. doi: 10.1056/NEJMp1006304

Han, Y. and Yu, L. (2012). A variance reduction framework for stable feature selection. *Statistical Analysis and Data Mining*, 5(5):428–445. doi: 10.1002/sam.11152

Han, Y., Yang, Y., and Zhou, X. (2013). Co-regularized ensemble for feature selection. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, IJCAI'13*, AAAI Press, 1380–1386.

Haslam, D. (2007). Obesity: a medical history. *Obesity Reviews*, 8(s1), 31-36. doi: 10.1111/j.1467-789X.2007.00314.x

Hastie, T., Tibshirani, R., & Friedman, J. (2009). High-Dimensional Problems: p ≫ N. In *The Elements of Statistical Learning*, 2nd ed.. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. pp 649-698

Haury, A.C., Jacob, L., Vert, J.P. (2010). Increasing stability and interpretability of gene expression signatures. *ArXiv e-prints*. arXiv:1001.3109

Haury, A.C., Gestraud, P., Vert, J.P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 6(12), e28210.

Haury, A.C. (2012) Sélection de variables à partir de données d'expression, Signatures moléculaires pour le pronostic du cancer du sein et inférence de réseaux de régulation génique. Thèse de doctorat, École nationale supérieure des mines de Paris.

He, Z., Yu, W. (2010). Stable feature selection for biomarker discovery. *Computational*

Biology and Chemistry, 34(4), 215-225. doi: 10.1016/j.compbiolchem.2010.07.002

He, L., Cao, Z., Wang, Y., Du, W., Liang, Y. (2014). An Ensemble Feature Selection Method Based on mRMR for Paired Microarray Data. *Journal of Computational Information Systems* 10(11), 4875–4882.

Helleputte, T., Dupont, P. (2009). Feature selection by transfer learning with linear regularized models. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Artificial Intelligence. 533–547. doi: 10.1007/978-3-642-04180-8_52

Helleputte, T., Dupont, P. (2009b). Partially supervised feature selection with regularized linear models. In *International Conference on Machine Learning (ICML)*. 409-416. doi: 10.1145/1553374.1553427

Hirsch, J., Batchelor, B. (1976). Adipose tissue cellularity in human obesity. *Clinics in Endocrinology and Metabolism*, 5(2), 299-311. doi: 10.1016/S0300-595X(76)80023-0

Hong, J., Liu, Z. Zhu, H., Zhang, X., Liang, Y., Yao, S., Wang, F., Xie, X., Zhang, B., Tan, T., Fu, L., Nie, J., Cheng, C. (2014). The tumor suppressive role of NUMB isoform 1 in esophageal squamous cell carcinoma. *Oncotarget*, 5, In Press.

Hulse, J.V, Khoshgoftaar, T.M., Napolitano, A., Wald, R. (2012). Threshold-based feature selection techniques for high-dimensional bioinformatics data. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 1(1), 47-61. doi: 10.1007/s13721-012-0006-6

Ioannidis, J. P. (2005). Microarrays and molecular research: noise discovery? *Lancet*, 365, 454–455.

Itoh, M., Suganami, T., Hachiya, R., Ogawa, Y. (2011). Adipose Tissue Remodeling as Homeostatic Inflammation. International Journal of Inflammation, vol. 2011, Article ID 720926, 8 pages. doi: 10.4061/2011/720926

Izmirlian, G. (2004). Application of the Random Forest Classification Algorithm to a SELDI-TOF Proteomics Study in the Setting of a Cancer Prevention Trial. *Annals of the New York Academy of Sciences*, 1020: 154–174. doi:10.1196/annals.1310.015

Jain, A. K., & Chandrasekaran, B. (1982). 39 dimensionality and sample size considerations in pattern recognition practice. *Handbook of Statistics*, 2, 835–855.

Kalousis, A., Prados, J., & Hilario, M. (2005). Stability of feature selection algorithms. In *ICDM* (pp. 218–225). IEEE Computer Society.

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: a

study on high-dimensional spaces. *Knowl. Inf. Syst.*, 12,95–116.

Kanehisa, M., Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27-30. doi:10.1093/nar/28.1.27

Kaufman, L., Rousseeuw, P.J. (1990). Partitioning Around Medoids (Program PAM). In *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York. pp 68–126

Keophiphath, M., Achard, V., Henegar, C., Rouault, C., Clement, K., Lacasa, D. (2009). Macrophage-secreted factors promote a profibrotic phenotype in human preadipocytes. *Molecular Endocrinology*, 23(1), 11-24. doi: 10.1210/me.2008-0183

Khan, T., Muise, E.S., Iyengar, P., Wang, Z.V., Chandalia, M., Abate, N., Zhang, B.B., Bonaldo, P., Chua, S., Scherer, P.E. (2009). Metabolic dysregulation and adipose tissue fibrosis: Role of collagen VI. *Molecular and Cellular Biology*, 29, 1575–1591. doi: 10.1128/MCB.01300-08

Kiecolt-Glaser, J.K., Habash D.L., Fagundes, C.P., Andridge, R., Peng, J., Malarkey, W.B., Belury, M.A. (2014). Daily Stressors, Past Depression, and Metabolic Responses to High-Fat Meals: A Novel Path to Obesity. *Biological Psychiatry*, In Press. doi: 10.1016/j.biopsych.2014.05.018

Kim, I.H., Wang, H., Soderling, S.H., Yasuda, R. (2014). Loss of Cdc42 leads to defects in synaptic plasticity and remote memory recall. *Elife*, 8,e02839. doi: 10.7554/eLife.02839

Kim, M., Kim, S., Lim, J.H., Lee, C., Choi, H.C., Woo, C.H. (2012). Laminar flow activation of ERK5 protein in vascular endothelium leads to atheroprotective effect via NF-E2-related factor 2 (Nrf2) activation. *Journal of Biological Chemistry*, 287, 40722-40731. doi: 10.1074/jbc.M112.381509

Kira; K.; Rendell, L.A. (1992). The feature selection problem: Traditional methods and a new solution. *AAAI92: Proc. 10th Nat. Conf. on Artificial Intelligence*. John Wiley & Sons, Ltd., pp. 129–134

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence* (pp. 1137–1143). Morgan Kaufmann

Kohavi, R., John, G.H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1), 273–324. doi: 10.1016/S0004-3702(97)00043-X

Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief. *Proceedings of the European conference on machine learning on Machine Learning*. Springer-Verlag New York, Inc., 1994, pp. 171–182.

Kotsiantis, S., Kanellopoulos, D. (2006). Discretization Techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1), 47-58

Krizek, P., Kittler, J., & Hlav´ ac, V. (2007). Improving stability of feature selection methods. In W. Kropatsch, M. Kampel, & A. Hanbury (Eds.), *Computer Analysis of Images and Patterns* (pp. 929–936). Springer Berlin / Heidelberg volume 4673 of *Lecture Notes in Computer Science*.

Kuncheva, L. I. (2004). Combining Pattern Classifiers: Methods and Algorithms. Wiley. ISBN: 978-0-471-21078-8

Kuncheva, L. I. (2007). A stability index for feature selection. In V. Devedzic (Ed.), *Artificial Intelligence and Applications* (pp. 421–427). IASTED/ACTA Press.

Larsen, T.M., Dalskov, S., van Baak, M., Jebb, S., Kafatos, A., Pfeiffer, A., Martinez, J.A., Handjieva-Darlenska, T., Kunesová, M., Holst, C., Saris, W.H., Astrup, A. (2010). The Diet, Obesity and Genes (Diogenes) Dietary Study in eight European countries - a comprehensive design for long-term intervention. *Obesity Reviews*, 11(1), 76-91. doi: 10.1111/j.1467-789X.2009.00603.x

Lauss, M., Ringnér, M., Mattias Höglund, M. (2010). Prediction of Stage, Grade, and Survival in Bladder Cancer Using Genome-wide Expression Data: A Validation Study. *Clinical Cancer Research*, 16, 4421-4433. doi: 10.1158/1078-0432.CCR-10-0606

Lausser, L., Müssel, C., Maucher, M., Kestler, H.A. (2013). Measuring and visualizing the stability of biomarker selection techniques. *Computational Statistics*, 28:51–65. doi: 10.1007/s00180-011-0284-y

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., & Nowé, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4), 1106–1119.

Lemoine, S., Combes, F., Servant, N., Le Crom, S. (2006). Goulphar: rapid access and expertise for standard two-color microarray normalization methods. *BMC Bioinformatics*, 7:467. doi:10.1186/1471-2105-7-467

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the Fifteenth International Conference on Machine Learning* (ICML-98).

Liu, H., Liu, L., and Zhang, H. (2010). Ensemble gene selection by grouping for microarray data classification. *Journal of biomedical informatics*, 43(1):81–87. doi: 10.1016/j.jbi.2009.08.010

Liu, S.G., Yuan, S.H., Wu, H.Y., Huang, C.S., Liu, J. (2014). The programmed cell death 6

interacting protein insertion/deletion polymorphism is associated with non-small cell lung cancer risk in a Chinese Han population. *Tumor Biology*, In Press. doi: 10.1007/s13277-014-2081-z

Manolio, T.A., Green, E.D. (2014). Leading the way to genomic medicine. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics - Special Issue: Implementation of Genomic Medicine*, 166(1), 1-7. doi: 10.1002/ajmg.c.31384

Márquez-Quiñones, A., Mutch, D.M., Debard, C., Wang, P., Combes, M., Roussel, B., Holst, C., Martinez, J.A., Handjieva-Darlenska, T., Kalouskova, P., Jebb, S., Babalis, D., Pfeiffer, A.F., Larsen, T.M., Astrup, A., Saris, W.H., Mariman, E., Clément, K., Vidal, H., Langin, D., Viguerie, N., DiOGenes Project (2010). Adipose tissue transcriptome reflects variations between subjects with continued weight loss and subjects regaining weight 6 mo after caloric restriction independent of energy intake. *American Journal of Clinical Nutrition*, 92(4), 975-984. doi: 10.3945/ajcn.2010.29808

Marschner, K., Kollmann, K., Schweizer, M., Braulke, T., Pohl, S. (2011). A key enzyme in the biogenesis of lysosomes is a protease that regulates cholesterol metabolism. Science, 333(6038), 87-90. doi: 10.1126/science.1205677

Matthews, D.R., Hosker, J.P., Rudenski, A.S., Naylor, B.A., Treacher, D.F., Turner, R.C. (1985). Homeostasis model assessment: insulin resistance and β-cell function from fasting glucose and insulin concentrations in man. *Diabetologia*, 28, 412-419. doi: 10.1007/BF00280883

McCulloch, W., Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115–133. doi: 10.1007/BF02478259

Meinshausen, N., Bhlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473. doi: 10.1111/j.1467-9868.2010.00740.x

Métin, C., & Frost, D. O. (1989). Visual responses of neurons in somatosensory cortex of hamsters with experimentally induced retinal projections to somatosensory thalamus. *Proceedings of the National Academy of Sciences*, 86, 357–361.

Michiels, S., Koscielny, S., Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365, 488–492. doi: 10.1016/S0140-6736(05)17866-0

Miecznikowski, J. C., Wang, D., Liu, S., Sucheston, L., & Gold, D. (2010). Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways. *BMC Cancer*, 10, 573.

Mok, S.C, Chao, J., Skates, S., Wong, K.K. Yiu, G.K., Muto, M.G., Berkowitz, R.S., Cramer,

D.W. (2001). Prostasin, a potential serum marker for ovarian cancer: identification through microarray technology. *Journal of the National Cancer Institute*, 93(19), 1458–1464. doi: 10.1093/jnci/93.19.1458

Muhlbradt, E., Asatiani, E., Ortner, E., Wang, A., Gelmann, E.P. (2009). NKX3.1 activates expression of insulin-like growth factor binding protein-3 to mediate insulin-like growth factor-I signaling and cell proliferation. *Cancer Research*, 69(6), 2615-22. doi: 10.1158/0008-5472.CAN-08-3022

Mutch, D.M., Pers, T.H., Temanni, M.R., Pelloux, V., Marquez-Quiñones, A., Holst, C., Martinez, J.A., Babalis, D., van Baak, M.A., Handjieva-Darlenska, T., Walker, C.G., Astrup, A., Saris, W.H., Langin, D., Viguerie, N., Zucker, J.D., Clément, K., DiOGenes Project (2011). A distinct adipose tissue gene expression response to caloric restriction predicts 6-mo weight maintenance in obese subjects. *American Journal of Clinical Nutrition*, 94(6), 1399-1409. doi: 10.3945/ajcn.110.006858

Mutch, D.M., Clément, K. (2006). Unraveling the Genetics of Human Obesity. *PLoS Genet*, 2(12): e188. doi: 10.1371/journal.pgen.0020188

Nagarajan, U.M., Louis-Plence, P., DeSandro, A., Nilsen, R., Bushey, A., Boss, J.M. (1999). RFX-B is the gene responsible for the most common cause of the bare lymphocyte syndrome, an MHC class II immunodeficiency. *Immunity*, 10(2), 153–62. doi: 10.1016/S1074-7613(00)80016-3

Neill, U.S. (2010). Leaping for leptin: the 2010 Albert Lasker Basic Medical Research Award goes to Douglas Coleman and Jeffrey M. Friedman. *Journal of Clinical Investigation*, 120 (10), 3413–3418. doi: 10.1172/JCI45094

Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., Mullany, E. C., Biryukov, S., Abbafati, C., Abera, S. F., Abraham, J. P., Abu-Rmeileh, N. M. E., Achoki, T., AlBuhairan, F. S., Alemu, Z. A., Alfonso, R., Ali, M. K., Ali, R., Guzman, N. A., Ammar, W., Anwari, P., Banerjee, A., Barquera, S., Basu, S., Bennett, D. A., Bhutta, Z., Blore, J., Cabral, N., Nonato, I. C., Chang, J.-C., Chowdhury, R., Courville, K. J., Criqui, M. H., Cundiff, D. K., Dabhadkar, K. C., Dandona, L., Davis, A., Dayama, A., Dharmaratne, S. D., Ding, E. L., Durrani, A. M., Esteghamati, A., Farzadfar, F., Fay, D. F. J., Feigin, V. L., Flaxman, A., Forouzanfar, M. H., Goto, A., Green, M. A., Gupta, R., Hafezi-Nejad, N., Hankey, G. J., Harewood, H. C., Havmoeller, R., Hay, S., Hernandez, L., Husseini, A., Idrisov, B. T., Ikeda, N., Islami, F., Jahangir, E., Jassal, S. K., Jee, S. H., Jeffreys, M., Jonas, J. B., Kabagambe, E. K., Khalifa, S. E. A. H., Kengne, A. P., Khader, Y. S., Khang, Y.-H., Kim, D., Kimokoti, R. W., Kinge, J. M., Kokubo, Y., Kosen, S., Kwan, G., Lai, T., Leinsalu, M., Li, Y., Liang, X., Liu, S., Logroscino, G., Lotufo, P. A., Lu, Y., Ma, J., Mainoo, N. K., Mensah, G. A., Merriman, T. R., Mokdad, A. H., Moschandreas, J., Naghavi, M., Naheed, A., Nand, D., Narayan, K. M. V., Nelson, E. L., Neuhouser, M. L., Nisar, M. I., Ohkubo, T., Oti, S. O., Pedroza, A., Prabhakaran, D., Roy, N., Sampson, U., Seo, H., Sepanlou, S. G., Shibuya, K., Shiri, R., Shiue, I., Singh, G. M.,

Singh, J. A., Skirbekk, V., Stapelberg, N. J. C., Sturua, L., Sykes, B. L., Tobias, M., Tran, B. X., Trasande, L., Toyoshima, H., van de Vijver, S., Vasankari, T. J., Veerman, J. L., Velasquez-Melendez, G., Vlassov, V. V., Vollset, S. E., Vos, T., Wang, C., Wang, S. X., Weiderpass, E., Werdecker, A., Wright, J. L., Yang, Y. C., Yatsuya, H., Yoon, J., Yoon, S.-J., Zhao, Y., Zhou, M., Zhu, S., Lopez, A. D., Murray, C. J. L., & Gakidou, E. (2014). Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the global burden of disease study 2013. *The Lancet*, In Press. doi: 10.1016/S0140-6736(14)60460-8

Opgen-Rhein, R., & Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical Applications in Genetics and Molecular Biology*, 6(1). doi: 10.2202/1544-6115.1252

Ouchi, N., Parker, J.L., Lugus, J.J.,Walsh, K. (2011) Adipokines in inflammation and metabolic disease. *Nature Reviews Immunology*, 11, 85-97. doi: 10.1038/nri2921

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., Liu, E.T., Miller, L., Nordgren, H., Ploner, H., Sandelin, K., Shaw, P.M., Smeds, J., Skoog, L., Wedrén, S., Bergh, J. (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, 7:R953-R964. doi:10.1186/bcr1325

Pégorier, J.P. (2007). Le tissue adipeux: Son rôle dans les maladies métaboliques. *Traité de nutrition artificielle de l'adulte*, 341-352. doi: 10.1007/978-2-287-33475-7_25

Pellegrinelli, V., Heuvingh, J., du Roure, O., Rouault, C., Devulder, A., Klein, C., Lacasa, M., Clément, E., Lacasa, D., Clément, K. (2014). Human adipocyte function is impacted by mechanical cues. *Journal of Pathology*, 233(2), 183-95. doi: 10.1002/path.4347

Peng, Y. (2006). A novel ensemble machine learning for robust microarray data classification. *Computers in Biology & Medicine*, 36(6), 553–573. doi: 10.1016/j.compbiomed.2005.04.001

Proikas-Cezanne, T., Waddell, S., Gaugel, A., Frickey, T., Lupas, A., Nordheim, A. (2004). WIPI-1alpha (WIPI49), a member of the novel 7-bladed WIPI protein family, is aberrantly expressed in human cancer and is linked to starvation-induced autophagy. *Oncogene*, 23, 9314–9325. doi: 10.1038/sj.onc.1208331

Pudil, P., & Somol, P. (2008). Identifying the most informative variables for decision-making problems - a survey of recent approaches and accompanying problems. *Acta Oeconomica Pragensia*, 2008, 37–55.

Rajah, R., Lee, K.W., Cohen, P. (2002). Insulin-like growth factor binding protein-3 mediates tumor necrosis factor-alpha-induced apoptosis: role of Bcl-2 phosphorylation. *Cell Growth & Differentiation*, 13(4), 163-171.

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26), 15149-54

Reboiro-Jato, M., Díaz, F., Glez-Peña, D., Fdez-Riverola, F. (2014). A novel ensemble of classifiers that use biological relevant gene sets for microarray classification. Applied Soft Computing, 17, 117–126. doi: 10.1016/j.asoc.2014.01.002

Resnik, P (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. 448-453. arXiv:cmp-lg/9511007

Rhee, S.Y., Wood,, V., Dolinski, K., Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7), 509–515. doi: 10.1038/nrg2363

Roe, A., Pallas, S., Kwon, Y., & Sur, M. (1992). Visual projections routed to the auditory pathway in ferrets: receptive fields of visual neurons in primary auditory cortex. *The Journal of Neuroscience*, 12, 3651–3664.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33, 1–39. doi: 10.1007/s10462-009-9124-7

Saeys, Y., Inza, I., & Larraaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23, 2507–2517.

Saeys, Y., Abeel, T., and Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In Daelemans, W., Goethals, B., and Morik, K., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5212 of *Lecture Notes in Computer Science*, pages 313–325. Springer Berlin Heidelberg.

Saitta, L., Giordana, A., Cornuéjols A. (2011). Phase Transitions in Machine Learning. Cambridge University Press. ISBN-13: 978-0521763912

Saitta, L., Zucker J.D. (2013). Feature Selection. *In Abstraction in Artificial Intelligence and Complex Systems*. New York Inc., Springer-Verlag, pp 278-283

Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270 (5235): 467–470. doi:10.1126/science.270.5235.467

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.*, 4:32.

Scherer, P.E., Williams, S., Fogliano, M., Baldini, G., Lodish, H.F. (1995). A novel serum protein similar to C1q, produced exclusively in adipocytes. *Journal of Biological Chemistry*, 270(45), 26746-267469. doi: 10.1074/jbc.270.45.26746

Schiller, Z.A., Schiele, N.R., Sims, J.K., Lee, K., Kuo, C.K. (2013). Adipogenesis of adipose-derived stem cells may be regulated via the cytoskeleton at physiological oxygen levels in vitro. *Stem Cell Research & Therapy*, 4(4), 79. doi: 10.1186/scrt230

Schlicker, A., Domingues, F.S., Rahnenführer, J., Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7:302. doi:10.1186/1471-2105-7-302

Schröder, M.S., Gusenleitner, D., Quackenbush, J., Culhane, A.C., Haibe-Kains, B. (2013). RamiGO: an R/Bioconductor package providing an AmiGO Visualize interface. *Bioinformatics*, 29, 666-668. doi:10.1093/bioinformatics/bts708

Shah, R. D., & Samworth, R. J. (2011). Variable selection with error control: Another look at Stability Selection. *ArXiv e-prints*.

Shen, R., Ghosh, D., Chinnaiyan, A. (2004). Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*, 5:94. doi:10.1186/1471-2164-5-94

Shieh, G.S, Jiang, Y.C., Shih, Y.S. (2006). Comparison of Support Vector Machines to Other Classifiers Using Gene Expression Data. Communications in Statistics - Simulation and Computation, 35(1), 241–256. doi: 10.1080/03610910500416215

Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L, Shaughnessy JD Jr, Oberthuer A, Thomas RS, Paules RS, Fielden M, Barlogie B, Chen W, Du P, Fischer M, Furlanello C, Gallas BD, Ge X, Megherbi DB, Symmans WF, Wang MD, Zhang J, Bitter H, Brors B, Bushel PR, Bylesjo M, Chen M, Cheng J, Cheng J, Chou J, Davison TS, Delorenzi M, Deng Y, Devanarayan V, Dix DJ, Dopazo J, Dorff KC, Elloumi F, Fan J, Fan S, Fan X, Fang H, Gonzaludo N, Hess KR, Hong H, Huan J, Irizarry RA, Judson R, Juraeva D, Lababidi S, Lambert CG, Li L, Li Y, Li Z, Lin SM, Liu G, Lobenhofer EK, Luo J, Luo W, McCall MN, Nikolsky Y, Pennello GA, Perkins RG, Philip R, Popovici V, Price ND, Qian F, Scherer A, Shi T, Shi W, Sung J, Thierry-Mieg D, Thierry-Mieg J, Thodima V, Trygg J, Vishnuvajjala L, Wang SJ, Wu J, Wu Y, Xie Q, Yousef WA, Zhang L, Zhang X, Zhong S, Zhou Y, Zhu S, Arasappan D, Bao W, Lucas AB, Berthold F, Brennan RJ, Buness A, Catalano JG, Chang C, Chen R, Cheng Y, Cui J, Czika W, Demichelis F, Deng X, Dosymbekov D, Eils R, Feng Y, Fostel J, Fulmer-Smentek S, Fuscoe JC, Gatto L, Ge W, Goldstein DR, Guo L, Halbert DN, Han J, Harris SC, Hatzis C, Herman D, Huang J, Jensen RV, Jiang R, Johnson CD, Jurman G, Kahlert Y, Khuder SA, Kohl M, Li J, Li L, Li M, Li QZ, Li S, Li Z, Liu J, Liu Y, Liu Z, Meng L, Madera M, Martinez-Murillo F, Medina I, Meehan J, Miclaus K, Moffitt RA, Montaner D, Mukherjee P, Mulligan GJ, Neville P, Nikolskaya T, Ning B, Page GP,

Parker J, Parry RM, Peng X, Peterson RL, Phan JH, Quanz B, Ren Y, Riccadonna S, Roter AH, Samuelson FW, Schumacher MM, Shambaugh JD, Shi Q, Shippy R, Si S, Smalter A, Sotiriou C, Soukup M, Staedtler F, Steiner G, Stokes TH, Sun Q, Tan PY, Tang R, Tezak Z, Thorn B, Tsyganova M, Turpaz Y, Vega SC, Visintainer R, von Frese J, Wang C, Wang E, Wang J, Wang W, Westermann F, Willey JC, Woods M, Wu S, Xiao N, Xu J, Xu L, Yang L, Zeng X, Zhang J, Zhang L, Zhang M, Zhao C, Puri RK, Scherf U, Tong W, Wolfinger RD; MAQC Consortium (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28, 827–838. doi: 10.1038/nbt.1665

Shi, Z., Xia, Y., Wu, F., Dai, J. (2014). The Discretization Algorithm for Rough Data and Its Application to Intrusion Detection. *Journal of Networks*, 9(6), 1380-1387. doi:10.4304/jnw.9.6.1380-1387

Shin, H., Sheu, B., Joseph, M., Markey, M.K. (2008) Guilt-by-association feature selection: Identifying biomarkers from proteomic profiles. *Journal of Biomedical Informatics*, 41(1), 124–136. doi: 10.1016/j.jbi.2007.04.003

Simon, R. (2003). Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *SIGKDD Explorations Special Issue on Microarray Data Mining*, 5(2):31–36.

Skerka, C., Chen, Q., Fremeaux-Bacchi, V., Roumenina, L.T. (2013). Complement factor H related proteins (CFHRs). *Molecular Immunology*, 56(3), 170-180. doi: 10.1016/j.molimm.2013.06.001

Somol, P., Grim, J., & Pudil, P. (2009). Criteria ensembles in feature selection. In J. A. Benediktsson, J. Kittler, & F. Roli (Eds.), *MCS* (pp. 304–313). Springer volume 5519 of *Lecture Notes in Computer Science*.

Somol, P., & Novovicova, J. (2010). Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans. Pattern Anal. Mach. Intell*., 32, 1921–1939.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis, *Journal of Molecular Biology*, 98(3), 503–508, doi: 10.1016/S0022-2836(75)80083-0.

Staiger, C., Cadot, S., Kooter, R., Dittrich, M., Müller, T., Klau, G., Wessels, L. (2012). A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. PloS one, 7(4): e34796. doi: 10.1371/journal.pone.0034796

Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, S. (2005). A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis. *Bioinformatics*, 21(5):631-43

Stears RL, Martinsky T, Schena M. (2003) Trends in microarray analysis. *Nat Med,* 9, 140-5. doi: 10.1038/nm0103-140

Suginta, W., Karoulias, N., Aitken, A., Ashley, R.H. (2001). Chloride intracellular channel protein CLIC4 (p64H1) binds directly to brain dynamin I in a complex containing actin, tubulin and 14-3-3 isoforms. *Biochemical Journal*, 359, 55–64.

Surowiecki, J. (2004). The Wisdom of Crowds: Why the Many are Smarter Than the Few and how Collective Wisdom Shapes Business, Economies, Societies, and Nations. Doubleday. ISBN-13: 978-0739311950

Tang, L., Yu, L., Wang, S., Li, J., Wang, S. (2012). A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting. *Applied Energy*, 93, 432–443. doi: 10.1016/j.apenergy.2011.12.030

Tapia, E , Bulacio, P, Angelone, L. (2011) Sparse and stable gene selection with consensus SVM-RFE, *Pattern Recognition Letters*, 33, 164–172. doi: 10.1016/j.patrec.2011.09.031

Temanni, M.R. (2009). Combinaison de sources de données pour l'amélioration de la prédiction en apprentissage: une application à la prédiction de la perte de poids chez l'obèse à partir de données transcriptomiques et cliniques. Thèse de doctorat, Université Pierre et Marie Curie.

Teufel, A., Becker, D., Weber, S.N., Dooley, S., Breitkopf-Heinlein, K., Maass, T., Hochrath, K., Krupp, M., Marquardt, J.U., Kolb, M., Korn, B., Niehrs, C,, Zimmermann, T., Godoy, P., Galle, P.R., Lammert, F. (2012). Identification of RARRES1 as a core regulator in liver fibrosis. *Journal of Molecular Medicine*, 90(12), 1439-47. doi: 10.1007/s00109-012-0919-7

Theodoridis, S., Koutroumbas, K. (2008). The PAM Algorithm. *Pattern Recognition, Fourth Edition*. pp. 746–748. ISBN-13: 978-1597492720

Trajdos, P., Kamizelich, A., Kurzynski, M. (2014) Three-Step Framework of Feature Selection for Data of DNA Microarray Experiments. *Advances in Intelligent Systems and Computing*, 283, 409-420. doi: 10.1007/978-3-319-06593-9_36

Tsamardinos, I., Aliferis, C.F. (2003). Towards principled feature selection: relevancy, filters and wrappers. *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.

Tukey, J.W. (1977). Exploratory data analysis. Pearson. ISBN-13: 978-0201076165

Ulmasov, B., Bruno, J., Gordon, N., Hartnett, M.E., Edwards, J.C. (2009). Chloride intracellular channel protein-4 functions in angiogenesis by supporting acidification of vacuoles along the intracellular tubulogenic pathway. *American Journal of Pathology*,

174(3), 1084–96. doi: 10.2353/ajpath.2009.080625

UniProt Consortium (The) (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 42(D1), D191-D198. doi: 10.1093/nar/gkt1140

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., & Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347, 1999–2009. doi: 10.1056/NEJMoa021967

Venet, D., Dumont, J., Detours, V. (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS computational biology*, 7(10):e1002240. doi: 10.1371/journal.pcbi.1002240

Wald, R., Khoshgoftaar, T. M., and Dittman, D. J. (2012). Mean Aggregation Versus Robust Rank Aggregation For Ensemble Gene Selection. 1*1th International Conference on Machine Learning and Applications - ICMLA* (1) 63-69. doi: 10.1109/ICMLA.2012.20

Wald, R., Khoshgoftaar, T.M., Abu Shanab, A., Napolitano, A. (2013) Comparative Analysis on the Stability of Feature Selection Techniques Using Three Frameworks on Biological Datasets. *12th International Conference on Machine Learning and Applications - ICMLA* (1) 2013: 418-423. doi: 10.1109/ICMLA.2013.85

Wald, R., Khoshgoftaar, T. M., and Dittman, D. J. (2013b). Ensemble gene selection versus single gene selection: Which is better? *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society (FLAIRS) Conference*. In Boonthum-Denecke, C. and Youngblood, G. M., editors. AAAI Press.

Walkey, S.U., Vanier, M.T. (2009). Secondary lipid accumulation in lysosomal disease. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1793(4), 726-736. doi: 10.1016/j.bbamcr.2008.11.014

Wang, J.Z., Du, Z., Payatttakool, R., Yu, P.S, Chen, C.F. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23 (10), 1274–1281. doi: 10.1093/bioinformatics/btm087

Wang, G; Gao, J.; Hu, F. (2013) A stable gene selection method based on sample weighting. *26th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 1–4. doi: 10.1109/CCECE.2013.6567792

Wang, X., Saso, H., Iwamoto, T., Xia, W., Gong, Y., Pusztai, L., Woodward, W.A., Reuben, J.M., Warner, S.L., Bearss, D.J., Hortobagyi, G.N., Hung, M.C., Ueno, N.T. (2013b). TIG1 promotes the development and progression of inflammatory breast cancer through

activation of Axl kinase. *Cancer Research*, 73(21), 6516-25. doi: 10.1158/0008-5472.CAN-13-0967

Wang, Y., Zhang, J.Y., Lin, F.R., Niu, Z.Y., Zhou, J.H. (2014) Effects of DNA Methylation on Expression of TIG1 Gene in Acute Leukemia. *Zhongguo Shi Yan Xue Ye Xue Za Zhi (Journal of experimental hematology / Chinese Association of Pathophysiology)*, 22(3), 648-52. doi: 10.7534/j.issn.1009-2137.2014.03.014. PMID: 24989270

Warnat, P., Eils, R., Brors, B. (2005). Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6:265. doi: 10.1186/1471-2105-6-265

Weisbart, R.H. (1992). An antibody that binds a neutrophil membrane protein, ERp72, primes human neutrophils for enhanced oxidative metabolism in response to formyl-methionyl-leucyl-phenylalanine. Implications for ERp72 in the signal transduction pathway for neutrophil priming. *Journal of Immunology*, 148(12), 3958-63.

Welch B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34 (1/2), 28–35. doi: 10.1093/biomet/34.1-2.28

Wells, J.C.K., Fewtrell, M.S., Williams, J.E., Haroun, D., Lawson, M.S., Cole, T.J. (2006) Body composition in normal weight, overweight and obese children:matched case-controlanalyses oftotal andregional tissue masses, and body composition trends in relation to relative weight. *International Journal of Obesity*, 30, 1506–1513. doi: 10.1038/sj.ijo.0803402

Wells, J.C.K. (2014) Commentary: The paradox of body mass index in obesity assessment: not a good index of adiposity, but not a bad index of cardio-metabolic risk. *International Journal of Epidemiology*, 43(3), 672-674. doi: 10.1093/ije/dyu060

Wessels, L.F.A., Reinders M.J.T., Hart A.A.M., Veenman, C.J., Dai, H., He, Y.D., van't Veer, L.J. (2005). A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, 21, 3755–3762. doi: 10.1093/bioinformatics/bti429

Wikipedia & Lmaps 2009 : https://en.wikipedia.org/wiki/File:Metabolomics_schema.png

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, 25(3), 287-318. doi: 10.1007/s10539-010-9200-z

World Obesity Federation / Policy & Prevention : http://www.worldobesity.org

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C., Haase, J., Janes, J., Huss, J., Su, A. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology*, 10, R130. doi:

10.1186/gb-2009-10-11-r130

Xu, L., Tan, A.C., Naiman, D.Q., Geman, D., Winslow, R.L. (2005). Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, 21 (20): p. 3905-3911. doi: 10.1093/bioinformatics/bti647

Xu, J.N., Liu, X., Wang, H., Hu, C.M. Luo, Q.H., Zhou, Q.Q. (2014).Effects of destrin pathway mutations on the gene expression profile. *Genetics and Molecular Research*, 13 (2): 2628 – 2637. doi: 10.4238/2014.April.8.5

Yan, S., Willis, J. (2013). WD40-repeat protein WDR18 collaborates with TopBP1 to facilitate DNA damage checkpoint signaling. *Biochemical and Biophysical Research Communications*, 431(3), 466-471. doi: 10.1016/j.bbrc.2012.12.144

Yang, Y. H. & Speed, T. (2002) Design issues for cDNA microarray experiments. *Nature Reviews Genetics* 3, 579-588. doi: 10.1038/nrg863

Yang, P., Yang, Y.H., Zhou, B.B., Zomaya, A.Y. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308. doi: 10.2174/157489310794072508

Yang, P., Ho, J.W.K., Yang, Y.H., Zhou, B.B. (2011). Gene-gene interaction filtering with ensemble of filters. *BMC Bioinformatics*, 12(Suppl 1):S10. doi: 10.1186/1471-2105-12-S1-S10

Yao, W., & Wang, Q. (2013). Robust variable selection through MAVE. *Computational Statistics & Data Analysis*, 63, 42 – 49. doi: 10.1016/j.csda.2013.01.021

Yu, L., Han, Y., Berens, M.E. (2012). Stable Gene Selection from Microarray Data via Sample Weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1), 262-272. doi: 10.1109/TCBB.2011.47

Zhu, L., Vranckx, R., Khau Van Kien, P., Lalande, A., Boisset, N., Mathieu, F., Wegman, M., Glancy, L., Gasc, J.M., Brunotte, F., Bruneval, P., Wolf, J.E., Michel, J.B., Jeunemaitre, X. (2006). Mutations in myosin heavy chain 11 cause a syndrome associating thoracic aortic aneurysm/aortic dissection and patent ductus arteriosus. *Nature Genetics*, 38, 343 – 349. doi: 10.1038/ng1721

Zhou, W., Zhou, C., Liu, G., Zhu, H. (2006). Feature Selection for Microarray Data Analysis Using Mutual Information and Rough Set Theory. *3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI)*, June 7–9, 2006, Athens, Greece. 492-499. doi: 10.1007/0-387-34224-9_57

Zhu, Y., Shen, X., Pan, W. (2009). Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics*, 10(Suppl 1):S21. doi: 10.1186/1471-2105-

10-S1-S21

Zuber, V., & Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25, 2700–2707. doi: 10.1093/bioinformatics/btp460

Annexes

Annexe 1 : classification sur données transcriptomiques

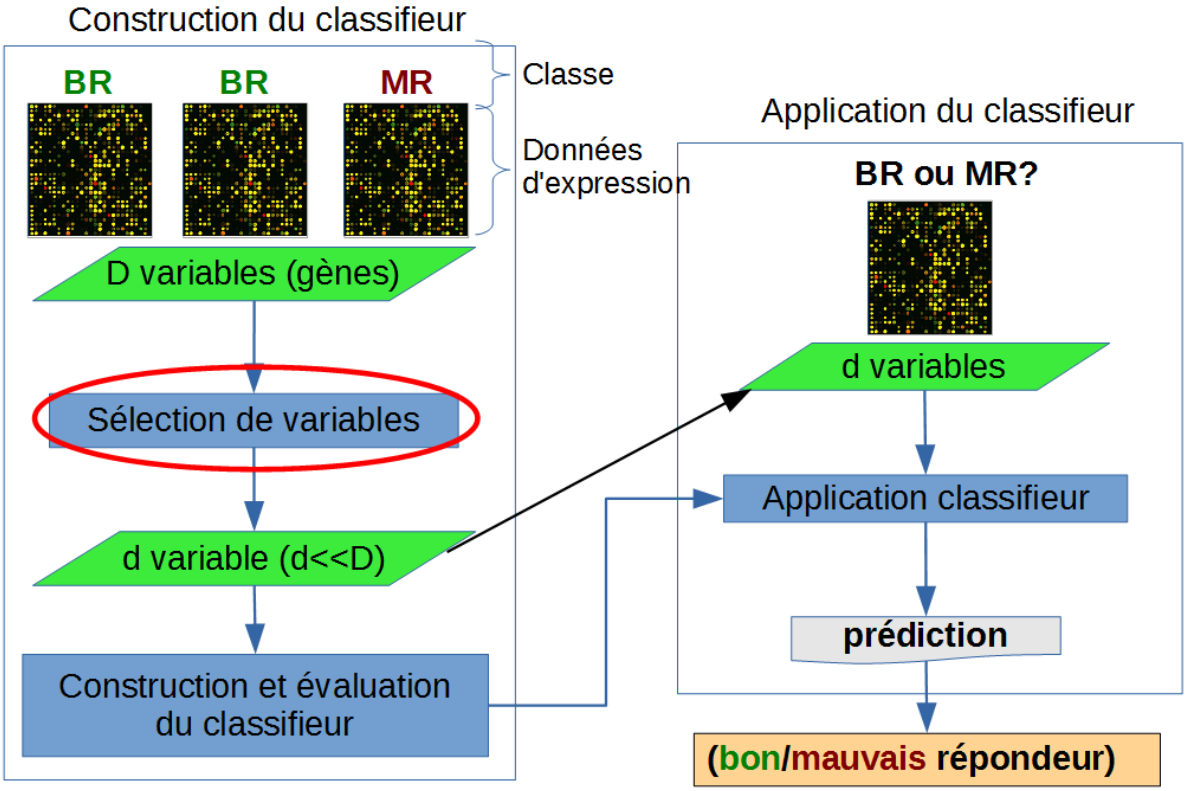

Figure A1.1 : Cas particulier de la Figure 5 appliquée à la classification sur données biopuces dans l'exemple de la prédiction d'une réponse thérapeutique. À gauche, une base de données (données biopuces + réponse connue classée en bon versus mauvais répondeur) est utilisée pour la sélection de gènes pertinents puis la construction d'un classifieur, dont les performances sont évaluées par validation croisée. À droite, sur un nouveau patient, on ne dispose que de données puces et on utilise le classifieur pour prédire la réponse (bon versus mauvais répondeur).

Annexe 2 : figures complémentaires au chapitre 2 : résultats additionnels sur données artificielles

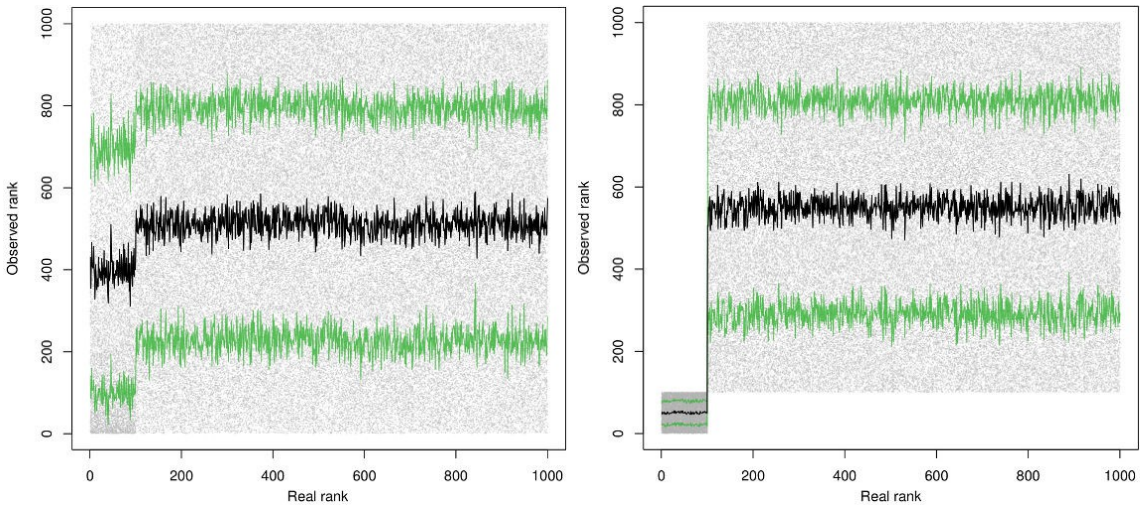

Figure A2.1 : Rang observé en fonction du rang réel sur un modèle de données artificielles avec 100 variables significatives (μ_i identiques, pour une erreur de Bayes de 10%) et 900 variables non significatives (μ_i=0). À gauche N=50, à droite N=5000. Courbe noire : moyenne, courbes vertes : moyenne +/- déviation standard. Filtre t-score.

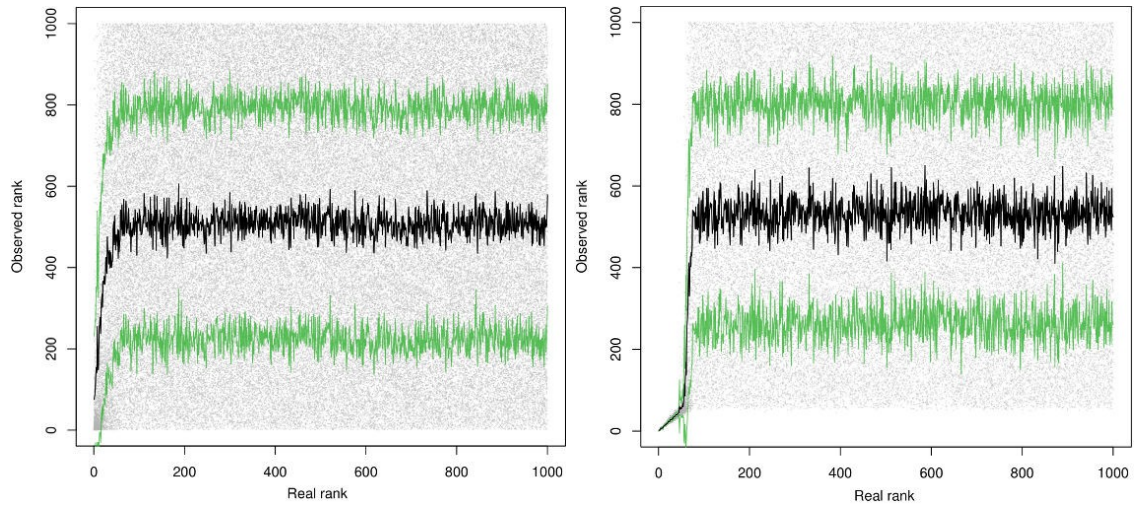

Figure A2.2 : Rang observé en fonction du rang réel sur un modèle de données artificielles avec 100 variables significatives (μ_i progressifs, pour une erreur de Bayes de 10%) et 900 variables non significatives (μ_i=0). À gauche N=50, à droite N=5000. Filtre t-score.

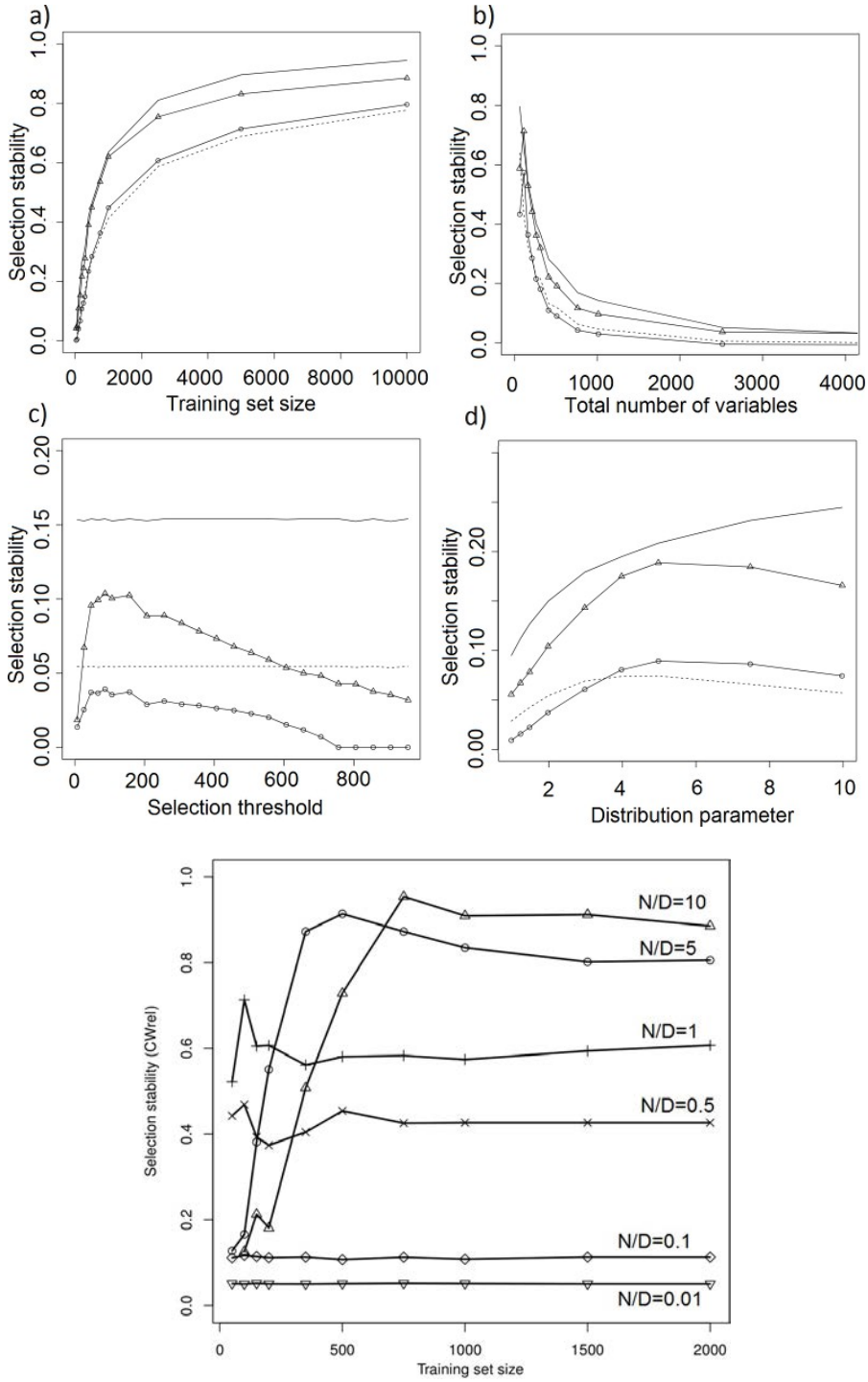

Figure A2.3 : Sur des données artificielles avec corrélation entre les variables, évolution des mesures de stabilité CW_{rel} (triangles), ATI_{PA} (cercles), S_W (lignes continues), S_R en fonction de : a) N sur [25;10000] b) D sur [50;10000] c) d sur [2;1000] et d) gamma sur [1;10]. En bas : stabilité CW_{rel} à ratio N/A constant. Filtre t-score.

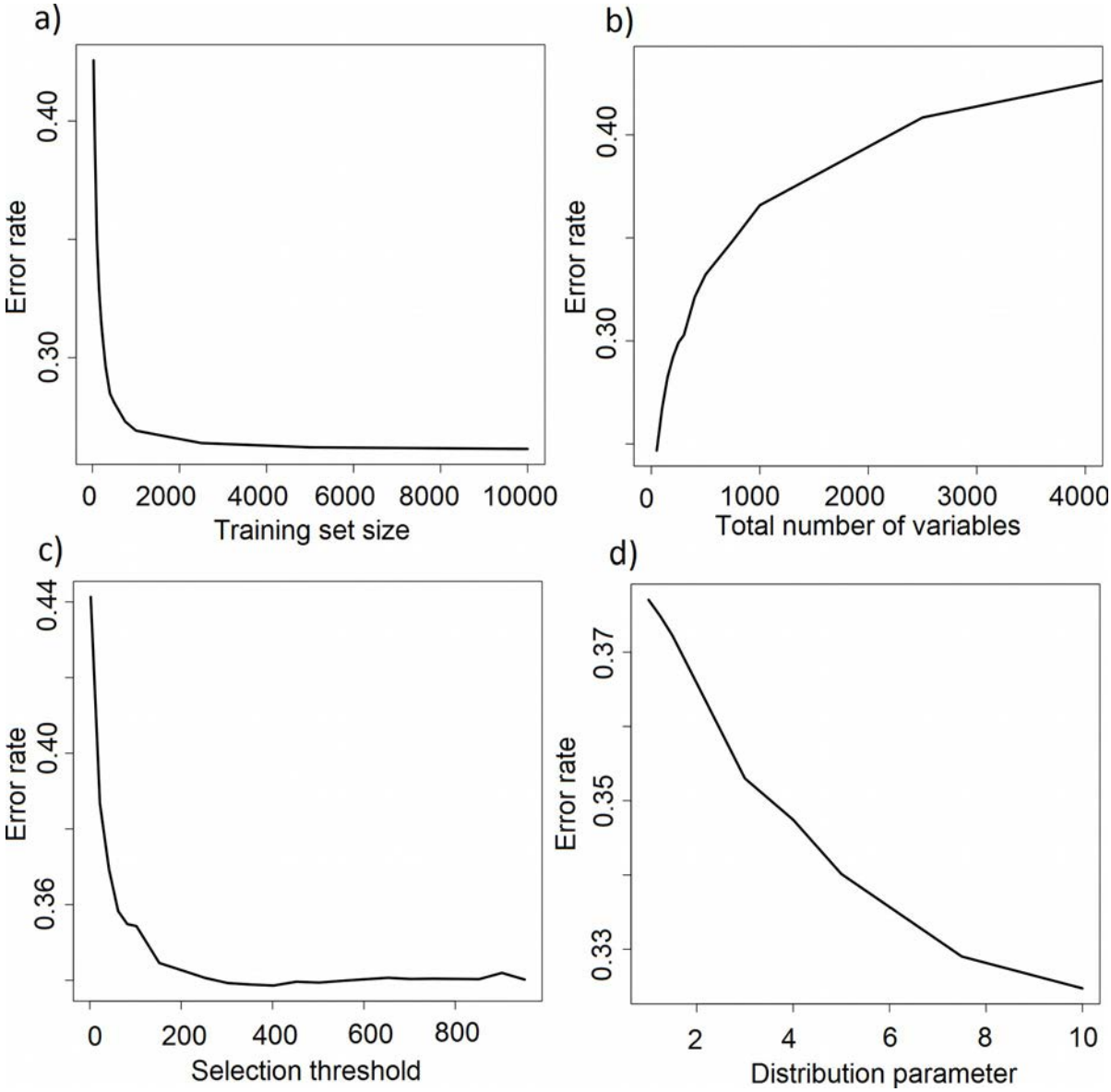

Figure A2.4 : Sur des données artificielles avec corrélation entre les variables, évolution du taux d'erreur en fonction de : a) N sur [25;10000] b) D sur [50;10000] c) d sur [2;1000] et d) gamma sur [1;10]. Filtre t-score, classifieur LDA.

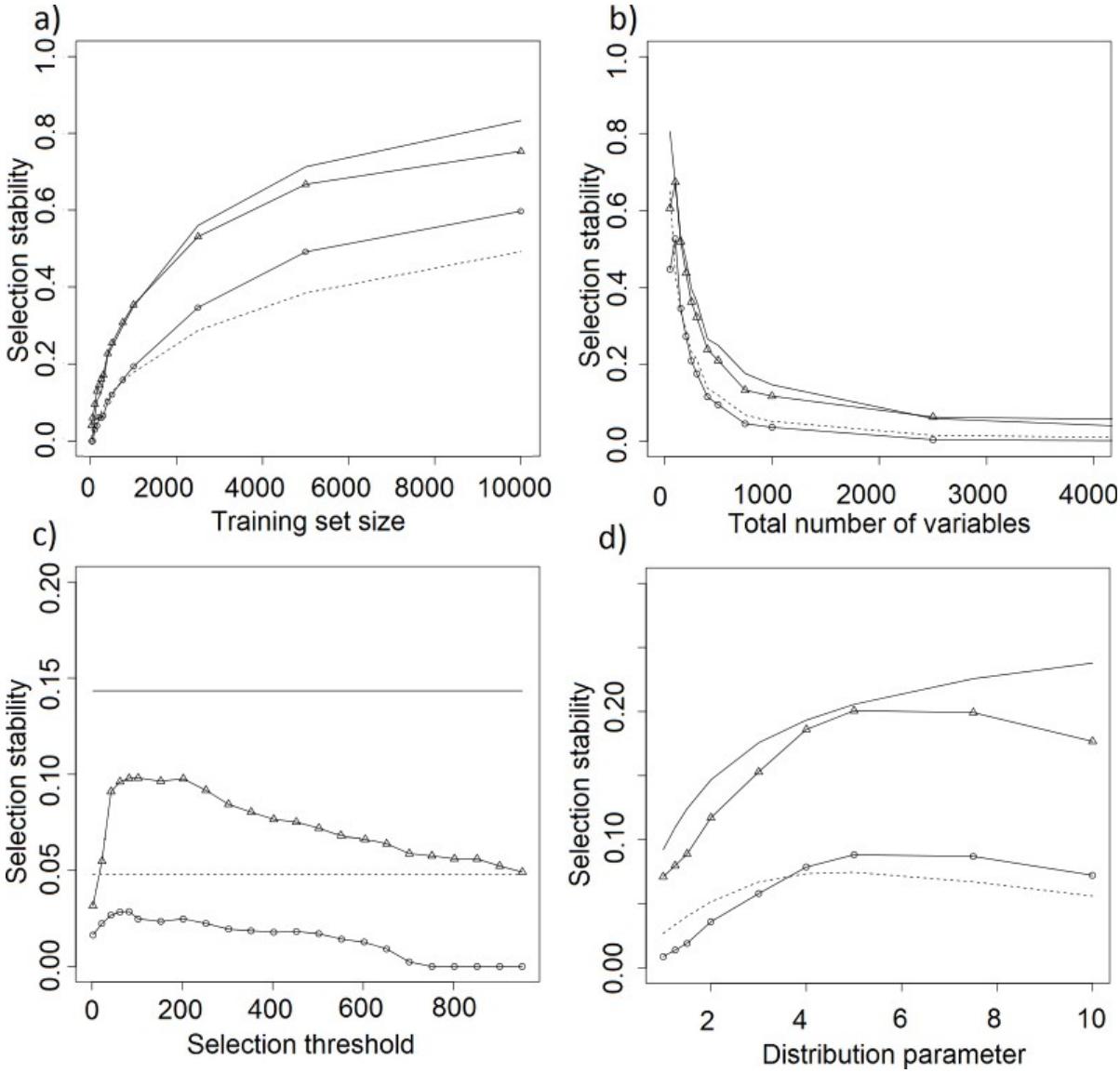

Figure A2.5 : Sur des données artificielles avec corrélation entre les variables, évolution des mesures de stabilité CW_{rel} (triangles), ATI_{PA} (cercles), S_W (lignes continues), S_R en fonction de : a) N sur [25;10000] b) D sur [50;10000] c) d sur [2;1000] et d) gamma sur [1;10]. Filtre CAT-score.

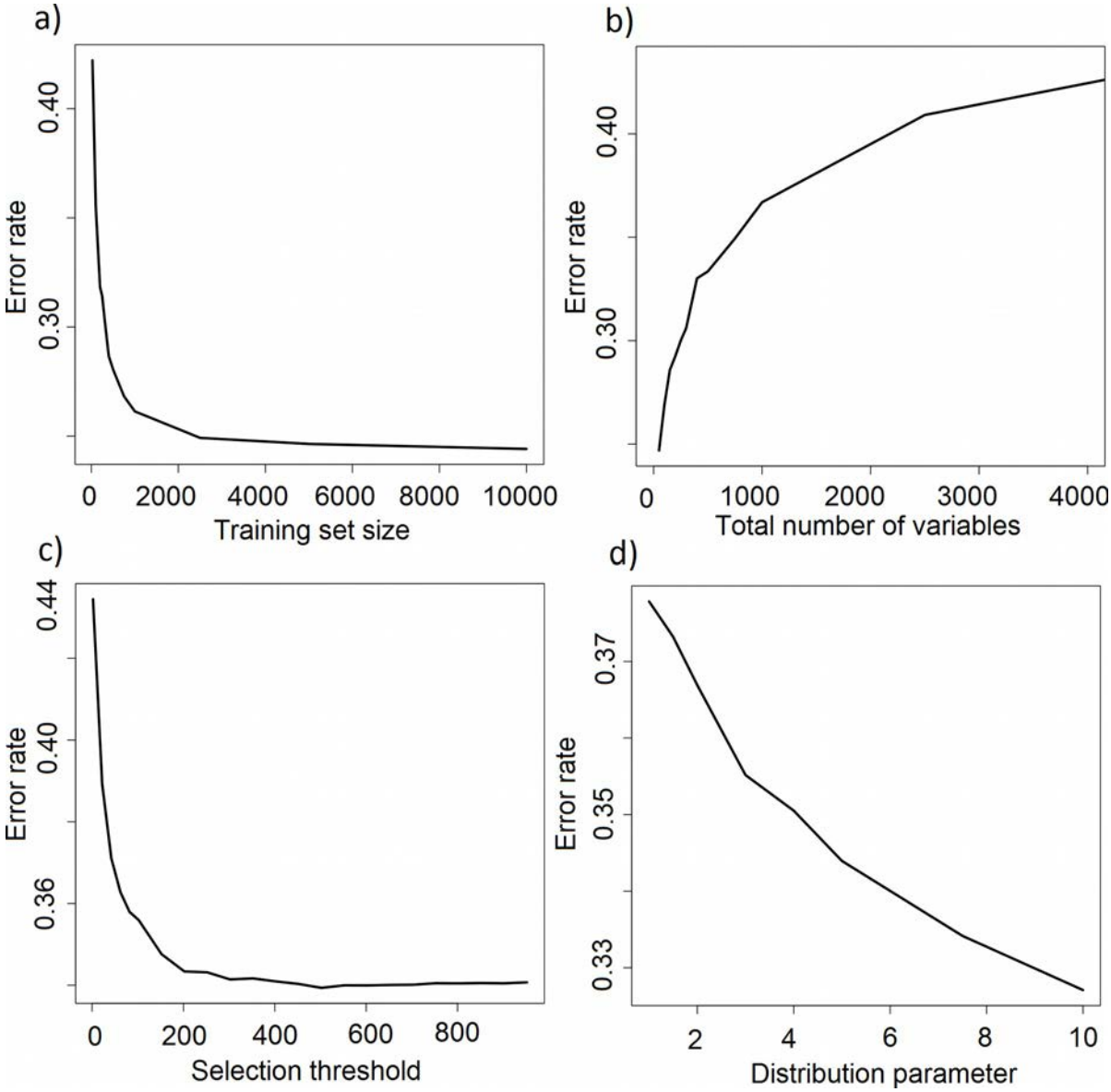

Figure A2.6 : Sur des données artificielles avec corrélation entre les variables, évolution du taux d'erreur en fonction de : a) N sur [25;10000] b) D sur [50;10000] c) d sur [2;1000] et d) gamma sur [1;10]. Filtre CAT-score, classifieur LDA.

Annexe 3 : calcul de CWrel sur le modèle théorique présenté en 2.3

On calcule d'abord CW.

Partant de la formule initiale :

$$CW(\mathcal{S}) = \sum_{f \in \mathcal{F}} \frac{F_f}{\Omega} \cdot \frac{F_f - 1}{\omega - 1}$$

On sépare les bonnes et les mauvaises variables :

$$CW(\mathcal{S}) = \sum_{f \in \mathcal{F}_g} \frac{F_f}{\Omega} \cdot \frac{F_f - 1}{\omega - 1} + \sum_{f \in \mathcal{F}_b} \frac{F_f}{\Omega} \cdot \frac{F_f - 1}{\omega - 1}$$

On remplace F_f par $p_{select}(f_g)\ \omega$ ou $p_{select}(f_b)\ \omega$:

$$CW(\mathcal{S}) = \sum_{f \in \mathcal{F}_g} \frac{p_{select}(f_g) \cdot \omega}{\Omega} \cdot \frac{(p_{select}(f_g) \cdot \omega) - 1}{\omega - 1} + \sum_{f \in \mathcal{F}_b} \frac{p_{select}(f_b) \cdot \omega}{\Omega} \cdot \frac{(p_{select}(f_b) \cdot \omega) - 1}{\omega - 1}$$

$$CW(\mathcal{S}) = D_g \cdot \frac{p_{select}(f_g) \cdot \omega}{\Omega} \cdot \frac{p_{select}(f_g) \cdot \omega - 1}{\omega - 1} + D_b \cdot \frac{p_{select}(f_b) \cdot \omega}{\Omega} \cdot \frac{p_{select}(f_b) \cdot \omega - 1}{\omega - 1}$$

Si on considère un nombre de selections ω très grand (ω→∞), on peut simplifier :

$$CW(\mathcal{S}) = D_g \cdot \frac{p_{select}(f_g) \cdot \omega}{\Omega} \cdot \frac{p_{select}(f_g) \cdot \omega}{\omega} + D_b \cdot \frac{p_{select}(f_b) \cdot \omega}{\Omega} \cdot \frac{p_{select}(f_b) \cdot \omega}{\omega}$$

$$CW(\mathcal{S}) = D_g \cdot \frac{p_{select}(f_g)^2 \cdot \omega}{\Omega} + D_b \cdot \frac{p_{select}(f_b)^2 \cdot \omega}{\Omega}$$

On remarque que Ω = d · ω pour obtenir enfin :

$$CW(\mathcal{S}) = \frac{1}{d} \cdot (D_g \cdot p_{select}(f_g)^2 + D_b \cdot p_{select}(f_b)^2)$$

Toutes les sélections étant de taille égale, $CW_{max} = 1$

La formule de CW_{min} est donnée dans (Somol & Novovicova, 2010)

$$CW_{min}(\Omega, \omega, \mathcal{F}) = \frac{\Omega^2 - D \cdot (\Omega - \Delta) - \Delta^2}{D \cdot \Omega(\omega - 1)}$$

avec : $\Delta = \Omega \cdot \mathrm{mod(D)}$

Une fois de plus, si on considère un nombre de selections ω très grand (ω→∞), on peut simplifier :

$$\underset{\omega \to \infty}{CW_{min}(\Omega, \omega, \mathcal{F})} = \frac{\Omega^2 - D \cdot \Omega}{D \cdot \Omega \cdot \omega} = \frac{\Omega \cdot (\Omega - D)}{D \cdot \Omega \cdot \omega}$$

$$\underset{\omega \to \infty}{CW_{min}(\Omega, \omega, \mathcal{F})} = \frac{\Omega}{D \cdot \omega}$$

On remarque encore que $\Omega = d \cdot \omega$ et on obtient :

$$\underset{\omega \to \infty}{CW_{min}(\Omega, \omega, \mathcal{F})} = \frac{d}{D}$$

On peut enfin obtenir CW_{rel} en substituant ses différents éléments par les formules calculées ci-dessus :

$$CW_{rel}(\mathcal{S}, \mathcal{F}) = \frac{CW(\mathcal{S}) - CW_{min}(\Omega, \omega, \mathcal{F})}{CW_{max}(\Omega, \omega) - CW_{min}(\Omega, \omega, \mathcal{F})}$$

Stabilité de la sélection de variables sur des données haute dimension : une application à l'expression génique

Résumé : Les technologies dites « haut débit » permettent de mesurer de très grandes quantités de variables à l'échelle de chaque individu : séquence ADN, expressions des gènes, profil lipidique… L'extraction de connaissances à partir de ces données peut se faire par exemple par des méthodes de classification. Ces données contenant un très grand nombre de variables, mesurées sur quelques centaines de patients, la sélection de variables est une étape préalable indispensable pour réduire le risque de surapprentissage, diminuer les temps de calcul, et améliorer l'interprétabilité des modèles. Lorsque le nombre d'observations est faible, la sélection tend à être instable, et on observe souvent que sur deux jeux de données différents mais traitant d'un même problème, les variables sélectionnées ne se recoupent presque pas. Pourtant, obtenir une sélection stable semble crucial si l'on veut avoir confiance dans la pertinence effective des variables sélectionnées à des fins d'extraction de connaissances. Dans ce travail, nous avons d'abord cherché à déterminer quels sont les facteurs qui influencent le plus la stabilité de la sélection. Puis nous avons proposé une approche, spécifique aux données puces à ADN, faisant appel aux annotations fonctionnelles pour assister les méthodes de sélection habituelles, en enrichissant les données avec des connaissances *a priori*. Nous avons ensuite travaillé sur deux aspects des méthodes d'ensemble : le choix de la méthode d'agrégation et les ensembles hybrides. Dans un dernier chapitre, nous appliquons les méthodes étudiées à un problème de prédiction de la reprise de poids suite à un régime, à partir de données puces, chez des patients obèses.

Mots-clés : *Sélection de variables, stabilité, données biopuces, données haute dimension, extraction de connaissances, obésité*

Feature selection stability on high dimensional data: an application to gene expression data

Abstract: High throughput technologies allow us to measure very high amounts of variables in patients: DNA sequence, gene expression, lipid profile… Knowledge discovery can be performed on such data using, for instance, classification methods. However, those data contain a very high number of variables, which are measured, in the best cases, on a few hundreds of patients. This makes feature selection a necessary first step so as to reduce the risk of overfitting, reduce computation time, and improve model interpretability. When the amount of observations is low, feature selection tends to be unstable. It is common to observe that two selections obtained from two different datasets dealing with the same problem barely overlap. Yet, it seems important to obtain a stable selection if we want to be confident that the selected variables are really relevant, in an objective of knowledge discovery. In this work, we first tried to determine which factors have the most influence on feature selection stability. We then proposed a feature selection method, specific to microarray data, using functional annotations from Gene Ontology in order to assist usual feature selection methods, with the addition of *a priori* knowledge to the data. We then worked on two aspects of ensemble methods: the choice of the aggregation method, and hybrid ensemble methods. In the final chapter, we applied the methods studied in the thesis to a dataset from our lab, dealing with the prediction of weight regain after a diet, from microarray data, in obese patients.

Keywords : *Feature selection, stability, microarray data, high dimensional data, knowledge discovery, obesity*