Département de Mathématiques et Informatique

Master de Recherche
Méthodes Numériques et Statistiques Appliquées

Option : Méthodes statistiques pour la discrimination et le scoring

Rapport de Projet de fin d'études

Sélection supervisée de variables basée sur l'association et la redondance.

Réalisé par :
ISLAH HAMZA

Encadré par :
Pr H.CHAMLAL

Devant le jury :
Pr. H. CHAMLAL, Faculté des Sciences Ain Chock Casa
Pr. T. OUADERHMAN Faculté des Sciences Ain Chock Casa
Pr. S. MOUSSATEN Faculté des Sciences Ain Chock Casa

La date : 19/07/2022

REMERCIEMENT

Je remercie Dieu, le tout puissant, je rende grâce pour m'avoir donné santé, patience, volonté et surtout raison. Ainsi. J'exprime toute ma reconnaissance à madame CHAMLAL HASNA, Prof à l'université Hassan II Faculté des Sciences Ain Chock Casa, pour avoir assuré l'encadrement de ce travail. Je la remercie pour son soutien, son orientation et ses consignes. Son expérience et sa connaissance ont contribué à ma formation scientifique. Nous tenons également à remercier l'ensemble des membres du jury Prof T. OUADERHMAN, Prof S. MOUSSATEN et Prof H. CHAMLAL pour avoir bien voulu examiner et juger mon travail.

J'exprime ma profonde reconnaissance à mes parents, mes frères, ma soeur ainsi que toute ma famille pour leurs encouragements et prières qui m'ont permis de finaliser ce travail.

Enfin, j'exprime ma gratitude à tous ceux qui ont contribué d'une manière ou d'une autre à l'élaboration de ce travail.

À mes chers parents,

À toute ma famille

À tous mes proches

À tous mes amis

À tous mes professeurs

Je dédie le fruit de mes années d'études.

ABSTRACT :

Currently, variable selection is an active research area where various mathematical and computational disciplines intersect to find a set of explanatory variables that are important for learning models. A primary task of selection is to eliminate the redundancy that exists between the variables. In this report, we will examine how to identify redundancy between variables based on their associations. Then, we will study some methods of selections.

RÉSUMÉ :

Actuellement, la sélection de variables est un domaine de recherche actif, où diverses disciplines mathématiques et informatiques se croisent pour trouver un sous ensemble de variables explicatives importantes pour les modèles d'apprentissage. Une tâche principale de la sélection, c'est d'éliminer la redondance qui existe entre les variables explicatives. Dans ce rapport, nous allons examiner comment identifier la redondance entre les variables en fonction de leurs associations. Ensuite, nous étudierons quelques méthodes de sélection.

Table des matières

INTRODUCTION GÉNÉRALE

L'apprentissage supervisé est une branche de l'apprentissage automatique (machine learning) qui vise à prédire une variable Y à partir des variables $X = \{X_1, ..., X_p\}$ appelées variables explicatives. Pour cela, nous avons besoin d'un échantillon dont on connaît les valeurs de X et Y (training data set en anglais) pour entraîner un modèle d'apprentissage automatique. Ensuite, nous appliquerons le modèle obtenu pour prédire la valeur de Y à des nouveaux individus pour lesquels nous ne connaissons que les valeurs de X. Ainsi, les variables utilisées dans la phase d'apprentissage contrôlent directement la qualité des résultats de Y obtenus. Cependant, l'ensemble des variables X peut contenir des variables de bruit et/ou des variables redondantes. En effet, nous pouvons classer les variables explicatives en 4 catégories : les variables pertinents (A), les variables faiblement pertinentes et redondantes (B), les variables faiblement pertinentes et non redondantes(C) et finalement les variables non pertinentes(D).

La sélection des variables est un processus qui consiste à extraire un sous ensemble de variables (A) et (C) de l'ensemble X. En général, on peut distinguer entre trois types de méthodes de sélection :

• **Les méthodes filtre** qui utilisent des mesures pour évaluer un sous ensemble de variables explicatives indépendamment aux algorithmes d'apprentissage. Ces mesures sont choisies pour être rapides à calculer, tout en capturant l'utilité de l'ensemble des variables.

• **Les méthodes wrapper** qui utilisent le modèle d'apprentissage pour évaluer les sous ensembles de variables. Chaque sous ensemble est utilisé pour construire le modèle d'apprentissage et la précision des prédictions sera utilisé comme le critère d'évaluation. Comme les méthodes wrapper forment un nouveau modèle pour chaque sous ensemble, elles sont très gourmandes en calcul, mais fournissent généralement l'ensemble de variables le plus performant.

• **Les méthodes embedded** qui effectuent la sélection des variables comme une étape de la construction du modèle, les variables qui contribuent le plus à chaque itération du processus de la construction du modèle seront sélectionnées. La sélection basée sur les arbres de décision, les forêts aléatoires, la régression LASSO sont les méthodes embedded le plus utilisé.

Dans ce rapport, nous nous concentrerons sur les méthodes filtre pour la sélection de variables. En identifiant les règles d'association pour éliminer les variables redondantes et les variables non pertinentes.

Le reste de ce travail sera présenté comme suit : Dans le chapitre 1, nous verrons quelques critères utilisés pour quantifier les associations entre les variables. Le chapitre 2 sera consacré aux méthodes de la sélection dans le cas où chaque individu a une seule modalité de Y, ce problème de classification est appelé (single-label) en anglais et on va l'appeler dans ce rapport par problème de classification avec une seule réponse. Le chapitre 3 traite le cas où Y est non disjonctif (multi-réponse ou multi-lables en an-

glais). Finalement, dans les deux chapitres 4 et 5 on va faire une étude expérimentale aux méthodes de sélections traitées pour tester leurs efficacités.

PARTIE I.

ÉTUDE THÉORIQUE

CHAPITRE 1.

CRITÈRES D'ASSOCIATION

1.1 Introduction

Généralement, on peut définir la redondance entre deux variables par l'information qu'elles contiennent. On dit alors qu'une variable X_1 est redondante à X_2 et réciproquement si elles fournissent une quantité considérable d'informations communes. Pour définir le concept de la redondance en statistique, nous devons quantifier l'information partagée par les variables. Cela peut se faire par l'étude des critères d'association et des relations qui existent entre ces variables.

1.2 Corrélation de Pearson

1.2.1 Rappel de covariance

Définition 1

Soient X et Y deux variables aléatoire. La covariance entre eux est définie par :

$$cov(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{1.1}$$

La covariance représente la variation conjointe des deux variables X et Y, tel que si :

- $cov(X,Y) \gg 0$ signifie qu'en moyenne les variables X et Y ont le même sens de variation par rapport à leurs espérances
- $cov(X,Y) \ll 0$ alors en moyenne, X et Y ont des directions de variation opposées par rapport à leurs espérances.
- $cov(X,Y)$ vaut zéro, implique qu'il n'y a aucun effet de la variation de X par rapport à sa moyenne sur la variation de Y.

Avec ces résultats, nous pouvons utiliser la covariance pour déterminer l'existence

d'une relation linéaire entre les variables X et Y.

Propriété 1

Soit X,X'et Y trois variables aléatoires, alors nous avons les résultats suivants :

- $Cov(X, X) = \mathbb{V}(X)$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(\lambda X + X', Y) = \lambda Cov(X, Y) + Cov(X', Y)$.
- $Cov(X + c, Y) = Cov(X, Y)$ *avec* $c \in \mathbb{R}$

Proposition 1

Soient X et Y deux variables aléatoire alors :

$$|Cov(X, Y)| \leq \sqrt{\mathbb{V}(X).\mathbb{V}(Y)}$$

En effet :
On applique l'Inégalité de Cauchy-Schwarz dans l'espace $L_2(\Omega)$ (L'espace des variables aléatoires de variance finie)

$$\begin{aligned} |Cov(X, Y)| &= |\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))| \\ &\leq \sqrt{\mathbb{E}((X - \mathbb{E}(X)))^2 \mathbb{E}((Y - \mathbb{E}(Y)))^2} \\ &\leq \sqrt{\mathbb{V}(X)\mathbb{V}(Y)} \end{aligned} \tag{1.2}$$

Définition 2 (Covariance empirique)

Sur un échantillon de taille n, nous estimons la Covariance par :

$$\widehat{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{1.3}$$

Où $\bar{x}$ (resp $\bar{y}$) correspond à la moyenne arithmétique de X (resp de Y)

1.2.2 Corrélation de Pearson

Définition 3

Le coefficient de corrélation linéaire simple, dit de Bravais-Pearson (ou de Pearson), C'est une normalisation de la covariance par le produit des écarts-type des deux variables.

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{\mathbb{V}(X).\mathbb{V}(Y)}}$$

Propriété 2

$0 \leq \rho_{X,Y} \leq 1$ *(cela est clair à partir de la proposition 1)*

Proposition 2

Si X et Y soient indépendants alors $\rho_{X,Y} = 0$ la réciproque est fausse sauf dans le cas de Gaussiens

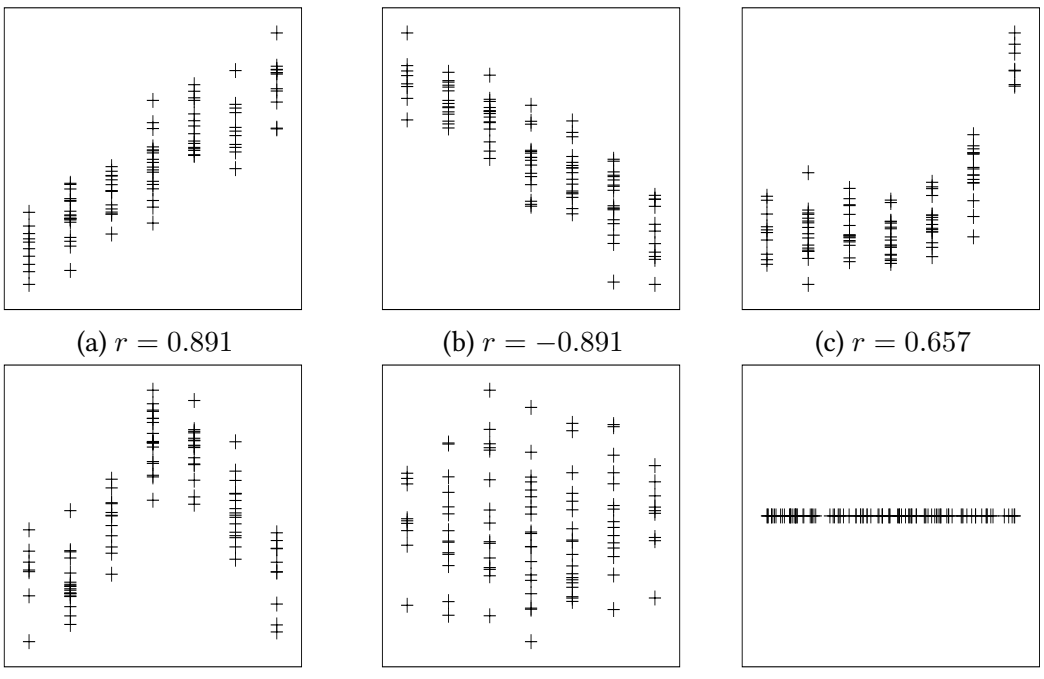
Définition 4 (Coefficient de corrélation empirique)

Pour un échantillon de taille, n en peut estimer l'indice de corrélation de Pearson par :

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \tag{1.4}$$

Où $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ sont les estimateurs des écarts-types de X et Y respectivement

L'indice de corrélation de Pearson est un véritable outil pour capturer les relations linéaires les deux figures (a) et (b). Cependant, si la relation entre les variables n'est pas linéaire, soit il ne fournit que des informations sur l'existence de la relation figure(c), soit, il donne des résultats erronés figures (d), (e).



(a) $r = 0.891$ (b) $r = -0.891$ (c) $r = 0.657$

(d) $r = 0.29$ (e) $r = -0.0006$ (f) $r = \text{Na}$

FIGURE 1.1 – le coefficient de corrélation dans des différents cas

1.2.3 Signification

Pour étudier la signification des valeurs de r obtenues et sous l'hypothèse de normalité des deux variables aléatoires, on effectue un test statistique qui a pour hypothèse nulle

$$H_0 : r = 0 \text{ contre } H_1 : r \neq 0$$

.
Sous H_0 la statistique de test (voir [1] page 131)

$$\frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \rightsquigarrow T_{n-2}$$

Corrélation partielle

Dans la pratique, la relation entre les variables en général n'est pas simple. En effet, la corrélation qui apparaît entre deux variables X_1 et X_2 peut être due à une troisième variable X_3,la corrélation entre X_1 et X_2 est dit fausse corrélation. Pour éviter ce problème, on doit étudier la corrélation partielle entre deux variables en éliminant l'effet d'une ou plusieurs variables

Définition 5

soient X_1, X_2, X_3 trois variables issues d'une loi normale, alors le coefficient de corrélation partielle entre X_1 et X_2 en éliminant l'effet de X_3 est défini par :

$$\rho_{X_1X_2,X_3} = \frac{\rho_{X_1X_2} - \rho_{X_1X_3}\rho_{X_2X_3}}{\sqrt{1-\rho^2_{X_1X_3}} \times \sqrt{1-\rho^2_{X_2X_3}}} \tag{1.5}$$

Remarque 1

Avec les mêmes conditions que dans le cas simple l'étude de la signification de l'indice de corrélation partiel se fait la statistique de test : $\frac{r}{\sqrt{1-r^2}}\sqrt{n-d-2} \rightsquigarrow T_{n-d-2}$ avec d c'est le nombre de variables fixés (voir [1])

Corrélation multiple

Définition 6

la corrélation multiple R c'est la valeur maximale de corrélation entre Y et la combinaison linéaire des X_i s'écrit comme :

$$R = sup_{a_1,\ldots,a_p} r(Y; \sum_{i=1}^{n} a_i X_i) \tag{1.6}$$

• Si Y est une combinaison linéaire des X_i alors $R = 1$

• Si les variables X_i et Y sont centrées, on peut calcule R^2 par la formule

$$R^2 = \frac{Y'X(X'X)^{-}X'Y}{Y'Y}$$

avec X, c'est la matrice des $X_i, i = 1, ..p$
la formule reliant corrélation multiple et corrélation partielle [1] est :

$$1 - R^2_{y.x_1\ldots x_p} = (1 - r^2_{yx_1})(1 - r^2_{yx_2.x_1})\ldots(1 - r^2_{yx_p.x_1\ldots x_{p-1}})$$

1.3 Indice de KENDALL

À partir des variables quantitatives et des variables qualitatives ordinales, on peut créer des variables dites de rang. Par exemple, à parti de la variable note d'un examen, qui a pour modalités (passable, assez bien, bien, très bien). Nous construisons une autre variable qui possède les valeurs (1 ≈ passable) (2 ≈ assez bien), (3 ≈ bien) et (4 ≈ très bien). Pour étudier les associations entre deux variables, nous pouvons le faire en étudiant leurs variables de rang.

Remarque 2

Dans ce qui suit, dans cette section, les variables X et Y représentent des variables de rang.

Définition 7

Soient (X_1, Y_1) et (X_2, Y_2) deux réalisations indépendantes des deux variables X et Y. L'indice de KENDALL [2] est défini par :

$$\tau = 2\mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0) - 1 \tag{1.7}$$

L'indice de KENDALL permet de savoir si les variables X et Y varient dans le même sens ou dans des sens contraires.

- $\tau = 1 \Rightarrow 2\mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0) - 1 = 1 \Rightarrow \mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0) = 1$ donc les deux variables ont presque sûr la même variation.
- $\tau = -1$ de même en trouve que les deux variables ont presque sûr une variation de sens contraire.

Absence d'ex æquo

Remarque 3

ex æquo signifie qu'il existe des individus avec le même rang.

Définition 8

Pour un échantillon, on estime le taux de KENDALL par :

$$\tau = \frac{S}{M} \tag{1.8}$$

avec $S = P - Q$,

où P *représente le nombre de paires* (i, j) *tel que* $(X_i - X_j)(Y_i - Y_j) > 0$

Q *le nombre de paires* (i, j) *qui vérifier* $(X_i - X_j)(Y_i - Y_j) < 0$

et M *c'est le nombre maximal qui peut prend P c'est* $M = \frac{n(n-1)}{2}$

Exemple 1

Soient X et Y deux variables de rang suivant

X	10	4	16	5	13	14
Y	17	14	20	8	11	23

comparaison en X	comparaison en Y	P	Q
10>4	17>14	T	
10>16	17<20	T	
10>5	17>8	T	
10<13	17>11		T
10<14	17>23	T	
4<16	14<20	T	
4<5	14>8		T
4<13	14>11		T
4<14	14<23	T	
16>5	20>8	T	
16>13	20>11	T	
16>14	20<23		T
5<13	8<11	T	
5<14	8<23	T	
13 <14	11<23	T	
		P=11	Q=4

$$\tau = \frac{(11-4)}{C_6^2} = \frac{7}{15} = 0.467$$

Proposition 3

Soient A et B le codage des comparaisons par paires des variables rangs X et Y

$$A_{ij} = \begin{cases} 1 \text{ si } X(i) < X(j) \\ -1 \text{ si } X(i) > X(j) \end{cases} \quad B_{ij} = \begin{cases} 1 \,\text{si}\, Y(i) < Y(j) \\ -1 \,\text{si}\, Y(i) > Y(j) \end{cases} \quad A_{ii} = B_{ii} = 0$$

alors $\tau(X,Y) = \text{cor}(A,B) = \text{cov}(A,B)$

En effet :

$$\sum_{i \neq j} A_{ij}B_{ij} = 2S = 2(P-Q) \text{ et } \sum_{i \neq j} A_{ij}^2 = \sum_{i \neq j} B_{ij}^2 = n(n-1)$$

Présence d'ex æquo

Dans le cas de présence d'ex æquo, on retient le codage suivant pour les variables rangs X et Y :

$$A_{ij} = \begin{cases} 1 \text{ si } X(i) & < X(j) \\ 0 \text{ si } X(i) & = X(j) \\ -1 \text{ si } X(i) & > X(j) \end{cases} \quad B_{ij} = \begin{cases} 1 \text{ si } Y(i) < Y(j) \\ 0 \text{ si } Y(i) = Y(j) \\ -1 \text{ si } Y(i) > Y(j) \end{cases}$$

Définition 9

Soient X (resp Y) deux variables de rang qui possèdent n_1 (resp n_2) groupes d'ex æquo, désignons par $u_i (1 \leq i \leq n_1)$(resp. $vj, (1 \leq j \leq n_2)$) le nombre d'individus ex æquo du $ime(resp.jme)groupe$.

$$\tau_1(X,Y) = \frac{S}{\sqrt{\left(\left(\frac{n(n-1)}{2} - V\right)\left(\frac{n(n-1)}{2} - U\right)\right)}},$$

$$V = \frac{\sum_j v_j \left(v_j - 1\right)}{2}, U = \frac{\sum_i u_i \left(u_i - 1\right)}{2}$$

l'autre expression, τ_2, du coefficient τ s'écrit sous la forme :

$$\tau_2(X,Y) = \frac{S}{\frac{n(n-1)}{2}}$$

Proposition 4

$\tau_1(X,Y) = cor(A,B)$ *et* $\tau_2 = cov(A,B)$

En effet :

$$\sum_{i \neq j} A_{ij} B_{ij} = 2S, \text{ et } \sum_{i \neq j} A_{ij}^2 = n(n-1) - 2U et \sum_{i \neq j} B_{ij}^2 = n(n-1) - 2V$$

Remarque 4

Grâce aux deux propositions (3) (4) *, on peut définir le Taux de KENDALL partielle et multiple.*

1.4 Rapport de corrélation

En exploitant le théorème de variance totale.

$$\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y/X)) + \mathbb{E}(\mathbb{V}(Y/X)) \tag{1.9}$$

On définit une mesure d'association suivante.

Définition 10

le rapport de corrélation entre deux variables statistique est définie par :

$$\eta^2_{Y/X} = \frac{\mathbb{V}(\mathbb{E}(Y/X))}{\mathbb{V}(Y)} \tag{1.10}$$

Il est clair que $0 \leq \eta^2_{Y/X} \leq 1$

Si le $\eta^2_{Y/X}$ =1 donc $\mathbb{E}(\mathbb{V}(Y/X)) = 0 \Rightarrow \mathbb{V}(Y/X) = 0$ presque sûr.Cela implique que pour X fixée, la variance de Y est nulle. alors Y est en fonction de X. $(ie\ Y = f(X))$

Si le $\eta^2_{Y/X}$ =0 $\mathbb{V}(\mathbb{E}(Y/X)) = 0 \Rightarrow \mathbb{E}(Y/X) = 0(p.s)$

Remarque 5

Contrairement à la corrélation de Pearson et le tau de KENDALL, le rapport de corrélation est une mesure d'association none symétrique.

$$\eta^2_{Y/X} \neq \eta^2_{X/Y}$$

Définition 11

Soit X un variable avec K catégorie, on notera $n_1, n_2, ..., n_k$ les effectifs observés et $\bar{y_1}, \bar{y_2}, ..., \bar{y_k}$; les moyennes de Y pour chaque catégorie.

Le coefficient de corrélation empirique s'écrit :

$$e^2 = \frac{\frac{1}{n}\sum_{i=1}^{k} n_i(\bar{y_i} - \bar{y})^2}{s_y^2} \tag{1.11}$$

Test de signification

Pour étudier la signification de e on fait un test statistique tel que

$$H_0 : e = 0 \text{ contre } H_1 : e \neq 0$$

sous l'hypothèse nulle, on définit la statistique de test

$$\frac{\frac{e^2}{k-1}}{\frac{1-e^2}{n-k}} \rightsquigarrow F(k-1, n-k) \tag{1.12}$$

1.5 Théorie d'information

Exemple Introductif

À l'aide des résultats de la théorie d'information, on peut définir des métriques pour étudier les associations entre les variables statistiques.
Tout d'abord, on donne un exemple introductif de la théorie d'information. Supposons que nous avons une urne contenant 4 boules, une blanche et les autres noires. On tire une boule sans remise, on a donc deux possibilités :

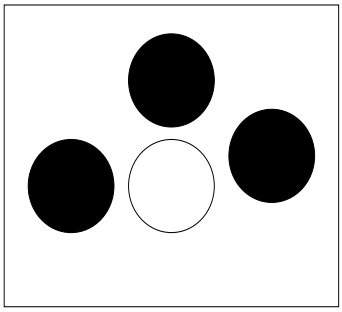


FIGURE 1.2 – Urne avec 4 boules

- obtenir la boule blanche avec 1/4 de chance,
- obtenir une boule noire avec 3/4 de chance.

Si au début nous obtenons la boule blanche, alors si on répète le tirage une deuxième fois, on va obtenir sûrement une boule noire. Mais si la première boule était noire, alors à la deuxième fois, on va soit obtenir une boule noire avec chance de 2/3 ou la boule blanche avec probabilité de 1/3. On conclut donc que l'événement "tirer la boule blanche " est plus informatif que l'autre événement, on quantifie cette notion d'information par la définition suivante

Définition 12

Soit X une variable aléatoire à valeur dans $x_1, ..., x_n$*, et on définit* $p_1, ..., p_n$
par $P(X = x_i) = p_i$*. l'information portée par une réalisation de l'événement* $X = x_i$ *:*

$$Info(X = x_i) = log(\frac{1}{p_i})$$

Propriété 3

- *Info* > 0
- *Soient A et B deux événements indépendants alors*

$$Info(A \cap B) = Info(A) + Info(B)$$

On dispose d'une mesure de l'information portée par une réalisation d'une variable, il est donc naturel d'introduire une mesure de l'information portée par la variable elle-même.

Définition 13

Soit X et Y deux variables aléatoires discrets, on définit alors l'entropie de X par :

$$H(X) = -\sum_{y \in Y} p(y) log(p(y))$$

Et l'entropie conjointe entre deux variables

$$H(X, Y) = -\sum_{y \in Y} \sum_{x \in X} p(y, x) log(p(y, x)) \tag{1.13}$$

Remarque 6

on peut également définir l'entropie de Shannon dans le cas continu comme :

$$H(X) = -\int_X f(x) log(f(x)) dx$$

Mais dans cette section, on va se focaliser seulement sur le cas discret.

Théorème 1

Soient X et Y deux variables aléatoires alors :

$$H(X,Y) \leq H(Y) + H(X) \tag{1.14}$$

avec l'égalité dans le cas d'indépendance entre X et Y

Définition 14 (entropie conditionnelle)

Étant donné une distribution de probabilité conditionnelle p(Y|X), on peut définir une entropie conditionnelle,

$$H(X|Y) = -\sum_{y \in Y} \sum_{x \in X} p(x,y) log(p(y|x)) \tag{1.15}$$

cette quantité permet de calculer l'incertitude de la variable X sachant que Y est déterminée

Théorème 2

Soient les variables aléatoires $X_1, ..., X_n$ alors :

$$H(X1, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1) \tag{1.16}$$

Théorème 3 ([3])

$$H(Y|X) \leq H(Y) \tag{1.17}$$

Pour étudier les interactions entre deux processus, il nous faut être capable de mesure l'information commune à deux variables aléatoires. Pour cela, on étend la notion d'entropie à la notion d'information mutuelle

Définition 15

Soit X, Y deux variables aléatoires de lois P_X et P_Y , et de loi conjointe, $P_{X,Y}$ alors on

définit l'information mutuelle de ces variables par :

$$\begin{aligned} I(X,Y) &= \sum_{x\in X}\sum_{y\in Y} p(x,y)log(\frac{p(x,y)}{p(x)p(y)} \\ &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \end{aligned} \tag{1.18}$$

Propriété 4

Pour toutes variables aléatoires X et Y

$$I(X,Y) \geq 0$$

on peut également définir l'information mutuelle conditionnelle.

Remarque 7

à partir de l'information mutuelle, on peut définir des mesures normalisées de l'information :

- *(asymmetric uncertainty)*

$$AU_Y(X,Y) = \frac{I(X;Y)}{H(Y)}$$

- *(Symmetric Uncertainty)*

$$SU(X;Y) = \frac{2I(X;Y)}{H(X)+H(Y)}$$

- *(Symmetric Relevance)*

$$SR(X;Y) = \frac{I(X;Y)}{H(X,Y)}$$

CHAPITRE 2.
LA SÉLÉCTION DE VARIABLES : CAS DISJONCTIF

2.1 Introduction

Comme mentionné dans l'introduction, les méthodes filtres prennent l'association entre les variables comme critère pour la sélection. Dans ce travail, nous allons concentrer nos études sur certaines méthodes basées sur la théorie de l'information pour quantifier la pertinence et la redondance entre les variables. Pour deux raisons : la première est que les mesures de la théorie de l'information sont indépendantes du type de variables, (qualitatif, quantitatif). La seconde est que ces mesures ne tiennent pas compte de la distribution des variables explicatives.

2.2 La méthode mRMR

La dépendance maximale

Pour construire une méthode de sélection basée sur l'information mutuelle comme mesure d'association, nous pouvons simplement chercher l'ensemble S_m de m variables explicatives qui vérifient l'association maximale avec la variable Y, ce qui est équivalent à trouver les variables $X_1, ...X_n$ qui vérifient :

$$\max \left[D(S_m; Y) = I\left(\{X_i, i = 1, \ldots, m\}; Y \right) \right] \tag{2.1}$$

L'information mutuelle multiple dans le cas discret, (respectivement continu) s'écrit :

$$I(S_m; Y) = \sum_{x_1,...,x_m,y} P(x_1, ...x_m, y) \log \frac{P(x_1, ..., x_m, y)}{P(x1, ..., x_m)P(y)}$$

$$I\left(S_m; Y\right) = \int \cdots \int p\left(x_1, \cdots, x_m, y\right) \log \frac{p\left(x_1, \cdots, x_m, y\right)}{p\left(x_1, \cdots, x_m\right) p(y)} dx_1 \cdots dx_m dy$$

Quand S contient m $\gg$1 variables, Souvent, il est difficile d'obtenir une estimation

précise pour la densité multiple p alors pratiquement l'implémentation de cette méthode est couteuse.

Pertinence maximale

Au lieu de maximiser la formule (2.1) on peut utiliser une alternative qui consiste à maximiser la moyenne d'information mutuelle individuelle entre chaque variable explicative et le variable cible Y donc le problème devient :

$$\max D(S,c), \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i;c)$$

Redondance minimale

Le dernier schéma ne prend pas en considération les relations entre les variables explicatives, cela signifie que l'ensemble sélectionné peut contenir des variables redondantes, alors on doit introduire une mesure pour quantifier la redondance

Définition 16

Soit S un ensemble de variables, on définit la quantité de la redondance pour S par :

$$R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j)$$

le critère de la redondance minimale vise à chercher parmi les ensembles candidats lequel qui minimise R.

La pertinence maximale et la redondance minimale

Finalement, la méthode MRMR [4] consiste à combiner simultanément entre la maximisation de la pertinence et la minimisation de la redondance, cela revient donc à maximiser l'écart entre les deux. On obtient alors le schéma d'optimisation suivant :

$$max\ \phi(S) = max(D(S) - R(S))$$

Mise en œuvre

En pratique, les méthodes de recherche ascendantes peuvent être utilisées pour trouver un optimal local défini par la fonction ϕ. Tous d'abord, on commence par sélectionner la variable qui possède le maximum d'information mutuelle avec Y. Puis à l'état $i > 1$ S contient déjà $i - 1$ éléments, on cherche le variable qui maximise la formule :

$$\max_{x_j \in X - S_{i-1}} \left[I(x_j; y) - \frac{1}{i-1} \sum_{x_i \in S_{i-1}} I(x_j; x_i) \right]$$

On répète ce processus jusqu'à obtenir un ensemble S avec m variable. Le pseudo code qui représente cette méthode est le suivant :

Algorithme

Entré : L'ensemble des variables explicatives X , la variable à prédire Y,

le nombre des éléments K

Sorti : Ensemble des Variables sélectionnées S

$S \leftarrow \emptyset$

Pour $x \in X$ faire

calculer $I(x, Y)$ l'information mutuelle entre x et Y

fin Pour

$S \leftarrow arg\ max_{x \in X} I(x, Y)$

$m \leftarrow 1$

TanQue (m $<$ K) faire

pour $x \in X - S$ faire

J(x) $\leftarrow I(x; y) - \frac{1}{\|S\|} \sum_{x_i \in S} I(x; x_i)$

fin Pour

$S \leftarrow S \cup arg\ max_{x \in X - S}\ J(x)$

$m \leftarrow m + 1$

fin TanQue

Discussion

Le score calculé par la méthode MRMR prend en considération une portion de la redondance entre les variables explicative indépendamment de la variable à expliquer qui s'appelle la redondance non pertinente. Cela affectera le choix des variables. Pour illustrer ce problème, on prend l'exemple suivant :
Soient X_1, X_2 deux variables explicatives et Y la variable cible.

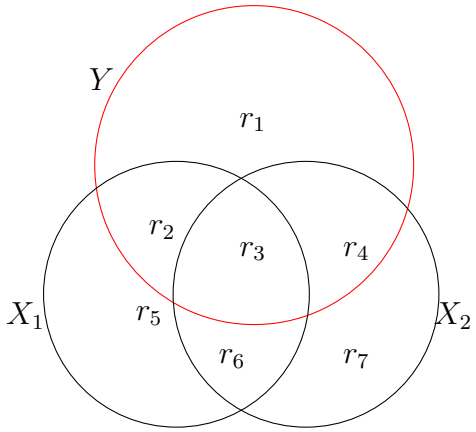


FIGURE 2.1 – Redondance non pertinente

Supposons que X_1 a été sélectionné dans la première étape. le score de X_2 par la méthode MRMR

$$MRMR\_Score(X_2) = I(X_2, Y) - I(X_2, X_1) = (r_3 + r_4) - (r_3 + r_6) = r_4 - r_6,$$

la quantité r_6 n'a aucun effet sur la pertinence de X_2 sur Y sachant que X_1 est déjà sélectionnée, mais il peut impacter le choix des variables. r_6 représente la redondance entre $X_1 et X_2$ indépendamment à Y. Il s'appelle la redondance non pertinente.

2.3 La méthode MRMSR

Pour surmonter le risque de la confusion rencontré dans la méthode MRMR. Cette méthode [5] prendra les mêmes idées de maximiser la pertinence et de minimiser la redondance, mais cette fois avec des différentes mesures.

Pertinence maximale

Définition 17

On mesure La pertinence entre les X_i et la variable cible Y par :

$$R(X_i, Y) = \frac{I(X_i, Y)}{H(Y)}$$

tel que I c'est l'information mutuelle et $H(Y)$ représente l'entropie de Shannon.

Proposition 5

Avec les mêmes hypothèses précédentes, on a :

$$0 \leq R(X_i, Y) = \frac{I(X_i, Y)}{H(Y)} \leq 1$$

***En effet**,*

on exploite les résultats de la théorie d'information, on a

$$0 \leq I(X_i, Y) = H(Y) - H(Y|X_i) \leq H(Y)$$

D'où

$$0 \leq R(X_i, Y) = \frac{I(X_i, Y)}{H(Y)} \leq 1$$

Redondance minimale

Définition 18 (redondance supervisée)

Pour mesurer la redondance, on définit un indice de similarité entre les variables explicatives X_i par :

$$S_{i,j} = 1 - \varphi_{i,j} = 1 - \frac{I\left(X_i; C \mid X_j\right) + I\left(X_j; C \mid X_i\right)}{2H(C)}$$

Propriété 5

$\forall X_i, X_j$. *l'indice $S_{i,j}$ Vérifie :*

- $0 \leq S_{i,j} \leq 1$

- *les variables* X^i, X^j *sont complétement corrélées si* $S_{i,j} = 1$
- $S_{i,j} = S_{j,i}$

pertinence maximale et la redondance minimale

On peut formuler le critère maximisation de la pertinence et minimisation de la redondance comme suivant :

$$\max \left\{ J\left(X_k\right) = R\left(X_k, Y\right) - \frac{1}{k-1} \sum_{X_l \in \mathcal{X}} S_{l,k} \right\}$$

$$\max \left\{ \frac{I\left(X_k; Y\right)}{H(Y)} - \frac{1}{k-1} \sum_{X_l \in \mathcal{X}} \left(1 - \frac{I\left(X_l; C \mid X_k\right) + I\left(X_k; Y \mid X_l\right)}{2H(Y)} \right) \right\}$$

Mise en œuvre

De même pour la première méthode, on va utiliser une méthode ascendante pour construire l'ensemble de k éléments. En effet, à la première étape, l'ensemble S ne contient aucune variable donc on va sélectionner la première variable seulement selon sa pertinence à la variable Y. et puis à l'état i on cherche les X_i qui maximisent le critère précédant. Le pseudo code pour implémenter cette méthode :

Algorithme

Entrés : X l'ensemble des variables explicatif. Y le variable à expliquer

Sortis : ensemble des variables sélection S.

$S \leftarrow \emptyset$; $m \leftarrow 0$;

calculer $H(Y)$ entrepie

Pour i de 1 à p faire :

Calculer $I(X_i, Y)$;

$R(X_i, Y) = \frac{I(X_i,Y)}{H(Y)}$

fin Pour

TanQue($m < q$) faire :

Si $m == 0$ faire :

sélection $X_l \leftarrow arg\ \max_{X_i \in X}\{R(X_i, Y\}$

$m \leftarrow m + 1; S \leftarrow S \cup X_l; X = X - X_l$

fin si

Pour $x_i \in X$:

calculer $I(X_l; Y|X_i)$ et $I(X_i; Y|X_l)$

$S_{l,i} \leftarrow 1 - \frac{I(X_l;Y|x_i)+I(x_i;Y|X_l)}{2H(Y)}$

$T_i \leftarrow T_i + S_{l,i}; J(F_i) \leftarrow R(F_i, C) - \frac{1}{|\overline{\mathcal{F}}|}T_i$

fin Pour

selection $X_l \leftarrow arg\ max_{X_i \in X}\{R(x_i, Y\}$

$m \leftarrow m + 1; S \leftarrow S \cup X_l; X = X - X_l$

fin TanQue

return S

2.4 La méthode OMICFS

Il s'agit d'une autre méthode construite pour surmonter le problème de la redondance non pertinente observée dans le MRMR, mais cette fois-ci nous utiliserons une autre stratégie basée sur la construction de variables indépendantes.

Pour quantifier le degré d'association entre les variables, nous utiliserons le coefficient d'information maximal MIC.

Définition 19

Soient X et Y deux variables, le MIC score [6] est définie comme :

$$\mathrm{MIC}(X,Y) = \max_{x_n y_n < N^{0.6}} \left[\frac{I_{x_n,y_n}(X,Y)}{\log_2 \min\{X_n, Y_n\}} \right] \tag{2.2}$$

avec N est la taille de l'échantillon, x_n et y_n désignent le nombre de groupes imposé sur les axes X et Y .

Dans le cas où les variables explicatives sont issus d'une loi normale centrée, alors s'il existe une relation entre les variables, elle sera linéaire. À partir de ces variables, on peut créer des variables orthogonales par l'algorithme de Gram-Schmidt GSO. Donc le critère de la maximisation de la pertinence et de la minimisation de la redondance sera optimiser indirectement par le calcule de MIC score entre ces variables indépendantes et le variable à prédire Y. La fonction score de cette méthode sera comme suit :

Définition 20

Soient X l'ensemble des variables explicatives et Y le variable cible. Supposons qu'on ait sélectionnés, $S_j = \{X^1, ..., X^j\}$ variables, on définit le score OMICFS de la variable X_i par :

$$J_{S_j}(X_i) = MIC(GSO(X_i; S_j), Y)$$

D'abord, le GSO consacré à calculer la variable orthogonale d'une variable candidate par rapport à d'autres variables, pour éliminer la redondance entre elles.

$$GSO\,(x_i, S_j) = \frac{u_i}{\|u_i\|}, u_i = x_i - \frac{\langle x_i, q_1\rangle}{\langle q_1, q_1\rangle} q_1 - \ldots - \frac{\langle x_i, q_{m-1}\rangle}{\langle q_{m-1}, q_{m-1}\rangle} q_{m-1}$$

où $q_i = GSO(s_i, S_{i-1})$, s_j c'est le variable sélectionnée à l'étape i
Et puis le MIC score entre ces variables orthogonaux et le variable cible permet de sélectionner les variables qui ont la pertinence maximale et la redondance minimale.

Mise en œuvre

En première étape, on calcule MIC score entre les variables candidates, et on sélectionne le variable avec la valeur maximale. Puis, chaque fois, on va calculer l'orthogonalisation de Gram-Schmidt des variables candidats par rapport aux variables sélectionnées pour éliminer la redondance avant calculer la pertinence de ces nouvelles variables pour sélectionner la candidate qui correspondant au max de pertinence des variables orthogonaux

---

Algorithme : OMICFS

Entrés : variables explicatives X, le variable cible Y, D le nombre de variables sorties

Sorties : l'ensemble S

$S \leftarrow \emptyset$

Pour i de 1 jusqu'à P faire

$MRelS(X_i) \leftarrow MIC(X_i, Y)$

fin Pour

Si $P >> N$ alors

$X_r \leftarrow$ ordonné $[MRelS_(X_i)]$ dans l'ordre décroissant

$X' \leftarrow X_r[1, .., \lambda N/log(N)],$ tel que $D < \lambda N/log(N) < P$

Sinon

$X' \leftarrow X$

fin Si

$s_1 \leftarrow arg\ max_{X^i \in X'}[MRelS_{(X_i)}]$

Pour m de 2 jusqu'à D faire

$s_m \leftarrow arg\ max_{X_i \in X' - S}[MIC(GSO(X_i, S), Y)]$

$S \leftarrow S \cup s_m$

fin Pour

return S

2.5 La méthode GBFS

Les méthodes de la sélection qu'on a vue jusqu'à maintenant sont basées sur l'inclusion ou l'exclusion individuelle des variables. L'utilisation de telles approches signifie

que des informations telles que la contribution collaborative où la corrélation entre ces variables peuvent être perdues. Donc le sous ensemble final de variables sélectionnées peut contenir des niveaux élevés de redondance. Pour exploiter l'effet collaboratif entre les variables, on peut les grouper selon des critères comme la pertinence ou la redondance et puis appliquer une stratégie de sélection basée sur ces groupes.

Featurs groupping

En effet, il existe plusieurs méthodes de groupement des variables, à titre d'exemple la méthode Kmeans [7]. Dans ce rapport, on va utiliser une approche basée sur la théorie des graphes. Pour mesurer l'interaction entre les variables. on utilise l'information mutuelle triple comme métrique.

Définition 21 (le gain d'interaction)

Soient X_i , X_j et Y trois variables aléatoires, L'information mutuelle triple on appelleé aussi gain d'interaction est définie par :

$$I\left(X_i, X_j, Y\right) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} \sum_{y \in Y} p\left(x_i, x_j, y\right) \log \frac{p\left(x_i, x_j, y\right) p\left(x_i\right) p\left(x_j\right) p(y)}{p\left(x_i, x_j\right) p\left(x_i, Y\right) p\left(x_j, y\right)}$$

- Une valeur de gain d'interaction positive indique que les deux variables fournissent ensemble plus d'informations sur la variable Y qu'ils ne le font individuellement. Plus la valeur positive est élevée, plus la collaboration est forte.
- Une interaction négative implique que deux variables sont redondantes.

Proposition 6

le gain d'interaction vérifie :

$$-\left[H\left(X_i\right) + H\left(X_j\right)\right] \leq I\left(X_i, X_j, Y\right) \leq \left[H\left(X_i\right) + H\left(X_j\right)\right]$$

La construction des groupes passe par les trois étapes suivantes[8] :

1- Construiction de graphe

Tous d'abord, on construit un graphe non orienté qui représente les associations entre les variables, tel que les nœuds de ce graphe sont les variables et les arrêts sont les gains d'interaction entre les variables.

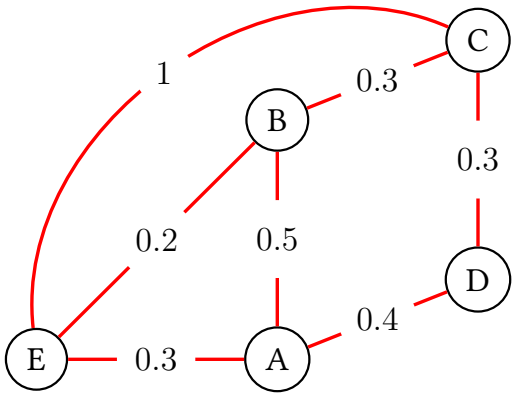


FIGURE 2.2 – Graphe représente les associations entre les variables

Remarque 8

Pour construire le graphe, on peut utiliser d'autre mesure que le gain d'interaction. à titre d'exemple le taux de KENDALL, corrélation de Pearson.

2- Construction de l'arbre couvrant de poid minimal

Ensuite, on utilise l'algorithme de KRUSCAL [9] pour construire un arbre couvrant de poids minimal

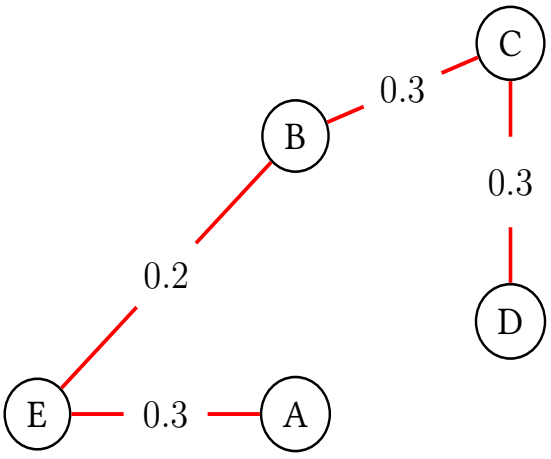


FIGURE 2.3 – Arbre couvrant de poids minimal

3- Construction des groupes

Finalement, on construit les groupes en éliminant les arrêts de poids maximaux, donc on obtient plusieurs composants connexes.

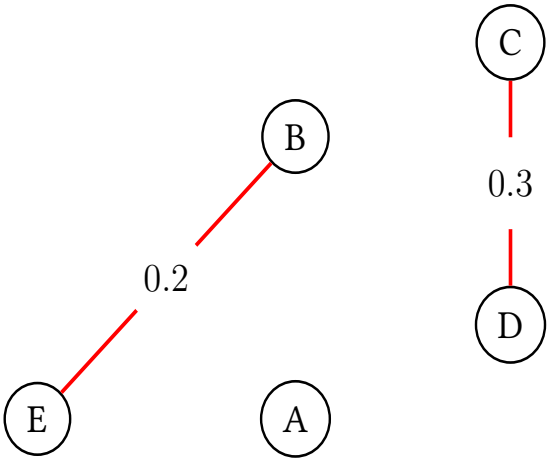


FIGURE 2.4 – les groupes de variables

Mise en œuvre de la sélection

Après la construction des groupes, on va exploiter ses groupes pour trouves un ensemble de variables S. tous d'abord on pend le père de variables le plus collaboratif celui qui maximise le gain d'interaction. Puis pour chaque groupe qui n'a aucun élément sélectionné, on va choisir un représentant le plus collaboratif avec l'ensemble sélectionné. ceci est valable à l'aide de la métrique suivante :

$$\mathrm{Col}\left(X_i, R\right) = \Sigma_{X_j \in R \cap 0 \leqslant I(X_i, X_j, Y)} I\left(X_i, X_j, Y\right)$$

Pour juger de la qualité de ce regroupement, la consistance probabiliste est adoptée comme métrique de qualité.

$$f(S,Y) = 1 - \sum_{j=1}^{k} \left(\sum_{y \in Y} p\left(N_S^j \mid y\right) p(y) - \sup_{y \in Y} \left(p\left(N_S^j, y\right) \right) \right)$$

avec N_S^j représente toutes les combinaisons des variables explicatives dans le tableau de données.

On résume ces étapes avec le shéma suivant :

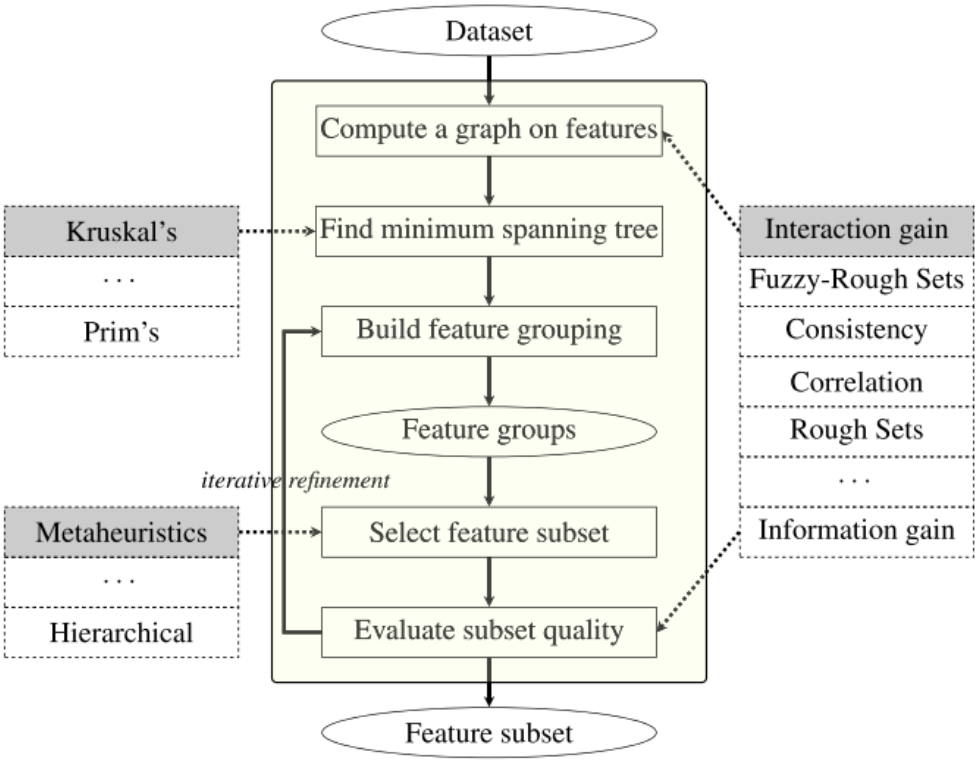


FIGURE 2.5 – schéma résume GBFS algorithme

Algorithme

Entré : Variables explicatives X , Cible Y, K

Sorties : Ensemble des Variables sélectionnées S et F les groupe de variables

// étape 1 la construction du graphe en utilisant le gain d'information

$G \leftarrow \{V, E\}$: graphe non orienté avec $V \leftarrow A$ et $E \leftarrow \emptyset$

TanQue$a_i, a_j \in V$

Si $< a_i, a_j > \neq \in E$

$Poids(< a_i, a_j >) \leftarrow \frac{I(a_i,a_j,Y)}{H(A_i)+H(A_j)}$

$E \leftarrow E \cup \{< a_i, a_j >\}$

// étape 2 construction d'arbre de poids minimal T en utilisant l'algorithme de Kruskal

$T \leftarrow Kruskall(G)$ qui possède les nodes V' et les arcs E'

// étape 3 générer les groupes et sélection l'ensemble S

Répéter :

éliminer les arrêts avec le maximum poids dans E' et diviser l'arbre à des forets.

$F \leftarrow \{V', E'\}$

$R \leftarrow \emptyset$

sélectionner un arrêt avec le poids maximal dans E et introduire ces nœuds dans R

Pour $f \in F$ & $f' \cap R == \emptyset$

$S \leftarrow R$

Pour chaque $a' \in f'$

Si $Col(a', R)$ c'est le maximum alors

Temp $\leftarrow R \cup \{a'\}$

fin Si fin Pour textbffin Pour

Évaluer les résultats sélectionnés

si $f(Temp, Y) > f(R, Y)$ alors

$R \leftarrow Temp$

sinon

$S \leftarrow R$

return S et F

fin Répéter

2.6 La méthode KIF

la méthode KIF [10] sélectionne les variables à travers les interactions entre elle par rapport à Y. pour mesurer ces interactions, on va se baser sur le taux de KENDALL pour construire une fonction de score.

Définition 22

Pour un variable Y avec K classe KENDALL Interaction Filtre score (KIF) entre variables X_i, X_j *est défini par :*

$$\omega_{ij} = \sum_{k=1}^{K} \pi_k |\tau_k(X_i, X_j) - \tau(X_i, X_j)| \tag{2.3}$$

avec $\pi_k = \mathbb{P}(Y = k)$ *et* $\tau_k = 2\mathbb{P}((X_j\tilde{X}_j)(X_i\tilde{X}_i) > 0; Y = k; \tilde{Y} = k) - 1$ *représente le taux de Kendall conditionnelle*

Définition 23

l'estimateur de l'indice KIF ω *sur un échantillon de taille n est définie par*

$$\hat{\omega}_{ij} = \sum_{k=1}^{K} \hat{\pi}_k |\hat{\tau}_k(X_i, X_j) - \hat{\tau}(X_i, X_j)| \tag{2.4}$$

où $\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} 1\{Y_i = k\}$ *et*
$\hat{\tau}(X_j, X_l) = \frac{4}{n(n-1)}\sum_{i<t=1}^{n} \mathbf{1}\{(X_{ij} - X_{tj})X_{il} - X_{tl}) > 0\} - 1$
$\hat{\tau_k}(X_j, X_l) = \frac{4}{n_k(n_k-1)}\sum_{i<t=1}^{n} \mathbf{1}\{(X_{ij} - X_{tj})X_{il} - X_{tl}) > 0, Y_i = k, Y_t = k\} - 1$

Mise en oeuvre

On exploite l'indice kif pour sélectionner les variables, on va procéder comme suit. tout d'abord, on va calculer les interactions entre tous les variables

pour trouver l'ensemble S de variables pertinentes, on commence tout d'abord avec le calcule des KIF score pour chaque couple de variables, si le nombre de variables est p alors, on doit calculer p(p-1)/2 score. après, on ordonne les paires de variables par le KIF score et on sélectionne les d premières couples les plus pertinents à la variable Y.
le pseudo code pour réaliser cette procédure est représenté dans le schéma suivant :

Algorithme

Entré : X les variables explicative, Y variables cible, d nombre de paires à sélectionner

Sorties : S l'ensemble des paires sélectionner

$S \leftarrow \emptyset$

$K \leftarrow$ nombre de modalités de Y

Pour chaque paire (X_i, X_j)

calculons $\tau(X_i, X_j)$

Pour $k = 1 : K$ faire

calculons $\hat{\tau}_k(X_i, X_j)$ par la formule précédent

calculons π_k

end Pour

$KIF_{i,j} \leftarrow \sum_{k=1}^{K} \hat{\pi}_k |\hat{\tau}_k(X_i, X_j) - \hat{\tau}(X_i, X_j)|$

classer les variables en ordre descendant par le Kif score

end Pour

$S \leftarrow$ d premières variables

CHAPITRE 3.
LA SÉLECTION DE VARIABLES : CAS NON DISJONCTIF

3.1 Introduction

les données multi-réponses

Dans plusieurs problèmes d'apprentissage automatique tels que le multimédia, l'analyse de données biologiques, l'exploration de textes, les variables cibles peuvent simultanément prendre plusieurs classes. Dans ce cas, on dit que le problème est disjonctif ou multi-réponses. [11]. Nous pouvons définir ce problème à travers l'exemple suivant : On considère un problème de classification des images tel que la variable à prédire Y représente les animaux domicile apparu dans ces images, les modalités de Y sont {Chat, Chien, Hamster, Perroquet}. L'image ci-dessous contient à la fois un chat et un chien, alors pour le même individu (image dans cet exemple), on peut avoir simultanément plusieurs classes (animaux).

FIGURE 3.1 – Photo by Tran Mau TriTam on Unsplash

On dit alors qu'il s'agit d'un problème non Disjonctif, multi-réponses ou bien multi-lables si chaque $Y_i \subset L$ où $L = \{l_1, ..., l_k\}$ c'est l'ensemble de lables

La classification multi-réponses

D'après [12], on peut diviser les méthodes de sélections multi-réponses en deux groupes : Méthodes basées sur la transformation et les méthodes d'adaptation. Pour illustrer ces deux méthodes, on va utiliser le tableau suivant :

	Y			
	Chat	Chien	Hamster	perroquet
E_1	1	1	0	0
E_2	1	0	0	1
E_3	0	0	1	0
E_4	1	0	0	1

Méthodes de transformation

Ces méthodes transforment le problème de classification multi-réponses en un ou plusieurs problèmes de classification ou de régression simple et puis appliquer les méthodes de classification classique pour chaque problème. Parmi ces méthodes, on trouve :

- (dubbet PT1)[12] qui consiste à sélectionner subjectivement ou aléatoirement

l'une des modalités pour chaque instance et rejette les autres, on obtient donc un problème multiclasse. On applique la méthode PT1 sur le tableau des données précédent, on obient

	Chat	Chien	Hamster	perroquet
E_1	1	0	0	0
E_2	1	0	0	0
E_3	0	0	1	0
E_4	0	0	0	1

L'inconvénient de cette méthode est qu'elle supprime une grande partie du contenu informatif de l'ensemble de données original.

- la méthode (dubbet PT3) qui utilise les combinaisons entre les modalités pour créer

des nouveaux de sort que chaque individue appartient seulement à une classe

	Chat et Chien	Chat et perroquet	Hamster
E_1	1	0	0
E_2	0	1	0
E_3	0	0	1
E_4	0	1	0

- la méthode (dubbed PT4) c'est la méthode le plus utilisée qui consiste à construire k problème de classification binaire tel que k représente le nombre de modalités.(voir [12])

Les méthodes adaptatives

Ces méthodes modifient les méthodes existantes dans le cas classique pour être applicable dans le cas multi-lables comme exemples On trouve la méthode arbre de décision $c4.5$ multi-lables, ML KNN

3.2 La sélection Multi-labels

la méthode "Fast Methode Featurs selection" [13] également basée sur la théorie de l'information pour mesurer la signification des variables en fonction des diverses associations entre les variables explicatives et les labels de variable Y.

Définition 24

Soient S un ensemble de variables explicatives et L l'ensemble des labels, on mesure la dépendance entre ces deux ensembles par :

$$M(S;L) = \sum_{k=2}^{|L|+n} \sum_{p=1}^{k-1} (-1)^k V_k \left(S'_p \times L'_{k-p}\right) \tag{3.1}$$

tel que $V_k(S') = \sum_{X \in S'_k} I(X)$ *I(X) représente l'information d'interaction qui définie par*

$$I(X) = -\sum_{Y \in X'} (-1)^{|Y|} H(Y) \tag{3.2}$$

Puisque dans cette section, on va considérer seulement le score individuel des variables, donc S va contenir un seul élément, f d'où la formule (3.1) devient

$$M(S;L) = M(f;L) = \sum_{k=2}^{|L|+1} (-1)^k V_k \left(f \times L'_{k-1}\right) \tag{3.3}$$

Remarque 9

la formule calcule l'interaction entre chaque variable et toutes les combinaisons possibles des Labels. l'inconvénient de cette formule est que la complexité va augmenter exponentiellement avec le nombre des modalités de variable Y.

Pour cela on va approcher cette dernière formule par :

$$M_b(f;L) = \sum_{k=2}^{b+1}(-1)^k V_k\left(f \times L'_{k-1}\right) \tag{3.4}$$

si b=2

$$\begin{aligned} M_2(f;L) &= \sum_{k=2}^{3}(-1)^k V_k\left(f \times L'_{k-1}\right) = V_2\left(f \times L'_1\right) - V_3\left(f \times L'_2\right) \\ &= \sum_{l_i \in L} I\left(f, l_i\right) - \sum_{l_i, l_j \in L} I\left(f, l_i, l_j\right) \end{aligned}$$

3.3 Mise en oeuvre

Pour évaluer les variables, on prend tout d'abord un sous ensembles Q des labels les plus informatives (ie ayant les plus grande entropie) de l'ensemble L puis à l'aide des résultats précédants, on construit une fonction score pour évaluer chaque variable explicative :

$$J(f) = V_2(f, L'_1) + \sum_{k=2}^{b+1}(-1)^k V_k(f \times Q'_{k-1}) \tag{3.5}$$

le pseudo code qui résume cette méthode est :

Algorithme
Entré : X ,Y ,n :nombre de variables à sélectionner, q : nombre de modalités considérer , b Sortie : S l'ensemble de variables sélectionnés
$S \leftarrow \emptyset$ **Pour** $f \in X$ faire : $H_f \leftarrow$ entropie de f **fin Pour** $F^* \leftarrow$ les variables explicatives avec le score le plus élevé de H **Pour** $l \in L$ faire : $H_l \leftarrow H(l)$ **end Pour** $Q \leftarrow$ les variables avec le score le plus élvé de H_l **pour all** $f \in F^*$ faire $J(f) \leftarrow$ la formule (3.5) **end Pour** Ordonné F^* on se bason sur J $S \leftarrow$ le meuilleur n variables de F^*

PARTIE II.

ÉTUDE EXPÉRIMENTALE

CHAPITRE 4.

EXPÉRIENCE SUR DES DONNÉES Á RÉPONSE SIMPLE

4.1 Introduction

Pour étudier l'efficacité des algorithmes mis en œuvre dans les chapitres précédents, nous allons réaliser une série d'expériences sur des données simulées et réelles.
Les tests sont effectués sous une machine avec les caractéristiques suivantes :
Intel® Core™ i3-2350M CPU @ 2.30GHz × 4 .
Vous pouvez trouver le code source des algorithmes implimentés sous R et aussi les tests effectués, dans mon dépôt GitHub
https://github.com/ISLAH-Hamza/supervised-Fs.

4.2 Étude sur des données simulées

Le tableau de données Toys

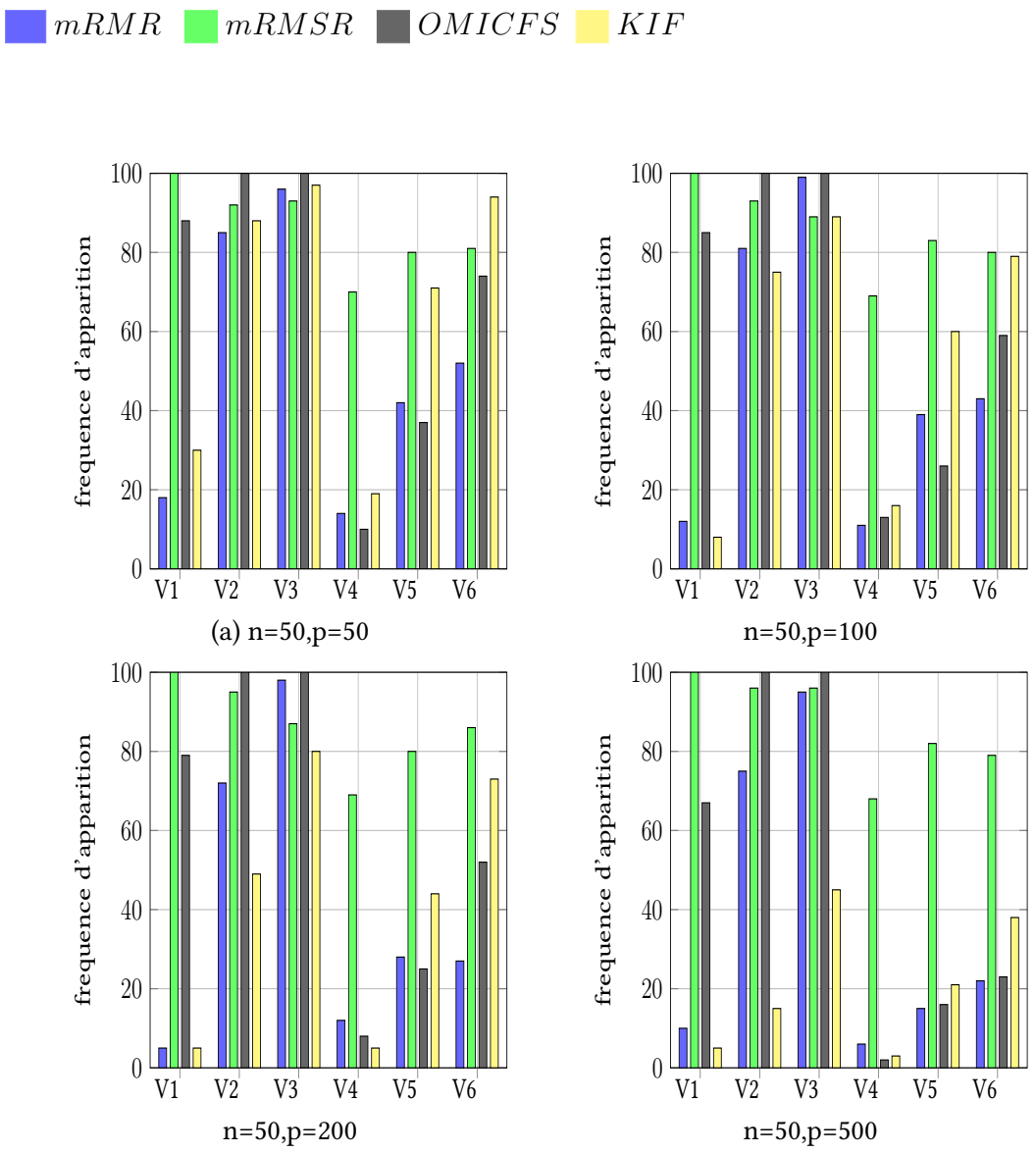
Le tableau de données **TOYS** simule un problème de classification binaire telle que les modalités de la variable à expliquer Y sont $\{-1, 1\}$. Cette simulation est construite de sorte que les six premières variables soient les plus pertinentes avec l'ordre d'importance suivant V_3, V_2, V_1, V_6, V_5, V_4 et que les autres constituent un bruit. Le modèle de simulation est le suivant :

- avec la probabilité de 0.7 $V_j \rightsquigarrow N(y.j, 1)$ pour j=2,3,4 et $V_j \rightsquigarrow N(0, 1) pour j = 4, 5, 6$
- avec la probabilité de 0.3 $V_j \rightsquigarrow N(0, 1)$ pour j=2,3,4 et $V_j \rightsquigarrow N(y.(j - 3), 1) pour j = 4, 5, 6$
- les autres $V_j \rightsquigarrow N(0, 20)$ pour $j = 7, ..., p$

2-Protocole expérimental

L'expérience se déroule comme suit. On va simuler 100 tableaux de données avec n lignes et p colonnes, Puis on applique les méthodes mRMR, mRMSR, OMICFS et

KIF, pour sélectionner un ensemble S de 6 variables, ensuite on calcule la fréquence d'apparition des six premières variables indépendantes $V_1, ...V_6$ parmi cet ensemble. On répète ce process pour des différentes valeurs de n et p. les résultats obtenus sont représentés dans les graphes suivants :

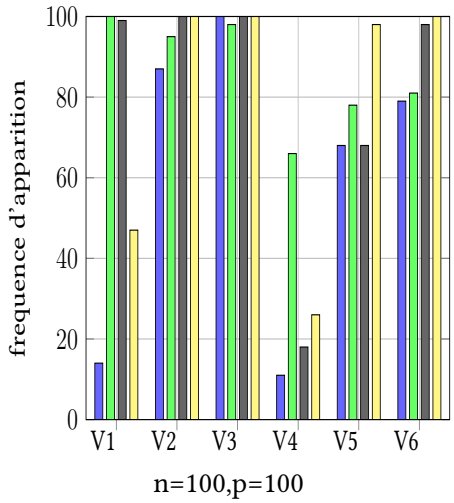


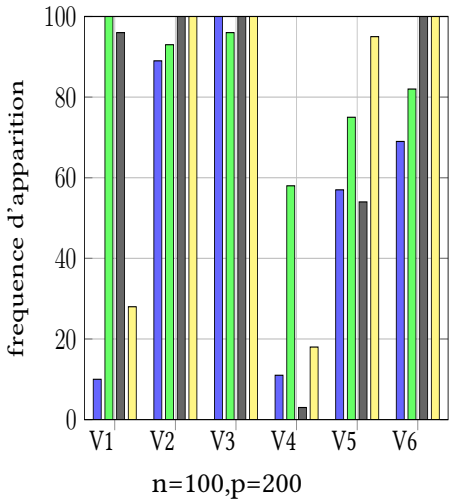
(a) n=50,p=50

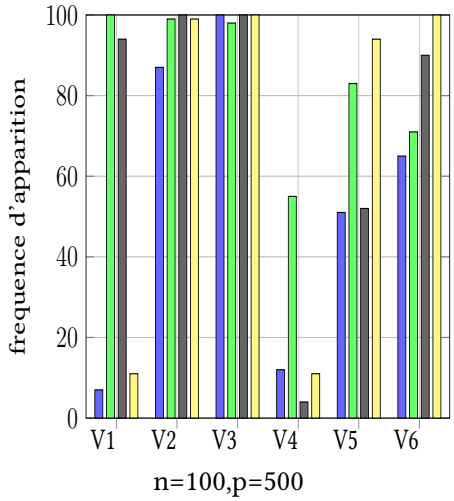
n=50,p=100

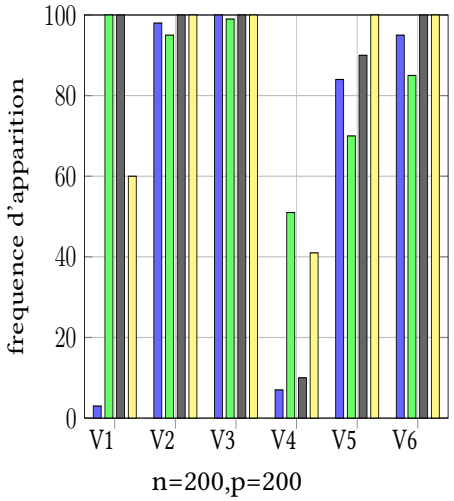
n=50,p=200

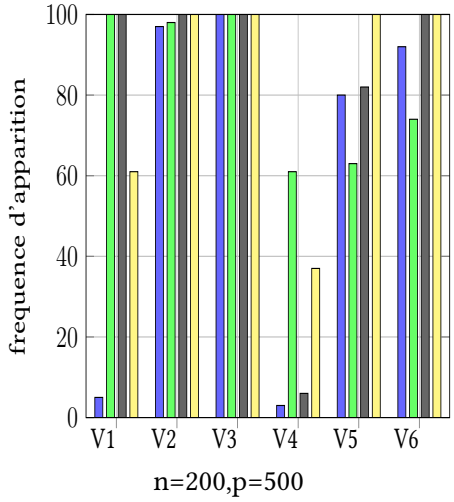
n=50,p=500

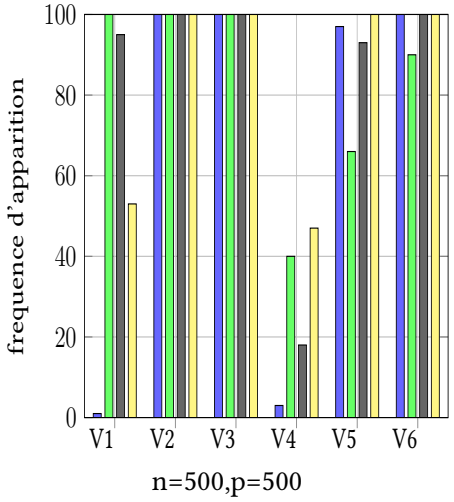

frequence d'apparition
0
20
40
60
80
100
V1
V2
V3
V4
V5
V6
n=100,p=100


frequence d'apparition
0
20
40
60
80
100
V1
V2
V3
V4
V5
V6
n=100,p=200


frequence d'apparition
0
20
40
60
80
100
V1
V2
V3
V4
V5
V6
n=100,p=500


frequence d'apparition
0
20
40
60
80
100
V1
V2
V3
V4
V5
V6
n=200,p=200


frequence d'apparition
0
20
40
60
80
100
V1
V2
V3
V4
V5
V6
n=200,p=500


frequence d'apparition
0
20
40
60
80
100
V1
V2
V3
V4
V5
V6
n=500,p=500

3-Discussion

Dans tous les cas, les résultats obtenus par les deux méthodes OMICFS et MRMRSR sont meilleurs que ceux de la méthode mrmr ce qui confirme que la contribution de la redondance non pertinente dans la fonction score de mrmr donne des résultats perturbés. Et généralement, les méthodes fonctionnent bien lorsque n et p sont suffisamment grands.

Étude sur des données réelles

Les données utilisées dans cette étude sont représentées dans le tableau suivant, avec n désigne le nombre de lignes et p le nombre de colonnes.

Data	Source	n	p	classes
Covid	github	863	15	2
Diabetic	UCI	1151	20	2
Scadi	UCI	70	206	6
Divorce	UCI	170	55	2

TABLE 4.1 – Description des données

Les sources de ces données sont UCI https://archive.ics.uci.edu/ml/index.php et Github https://github.com/AtharvaPeshkar/Covid-19-Patient-Health-Analytics

4.2.1 Mesures d'évaluation

Pour juger les résultats de la classification, nous allons utiliser des mesures [14] basées sur le tableau de confusion.
Dans le cas d'un problème de classification binaire, le tableau de confusion s'écrit :

valeurs prédit	**valeurs réelles**	
	Positive	Négative
Positive	vrai Positif	Faux Positif
Négative	vrai Négatif	Faux Négatif

tel que :

- VP (vrai positif) : représente le nombre des éléments qui sont positifs et ils sont prédits positifs
- VN (vrai négatif) : le nombre des éléments qui sont négatifs et ils sont prédits négatifs
- FP (faux positif) : les éléments qui sont positifs, mais le modèle les prédits négatifs
- FN (faux négatif) : le nombre des éléments qui sont négatifs et ils sont prédits positifs.

Alors, on définit :

$$accuracy = \frac{VP}{VP + VN + FP + FN}$$

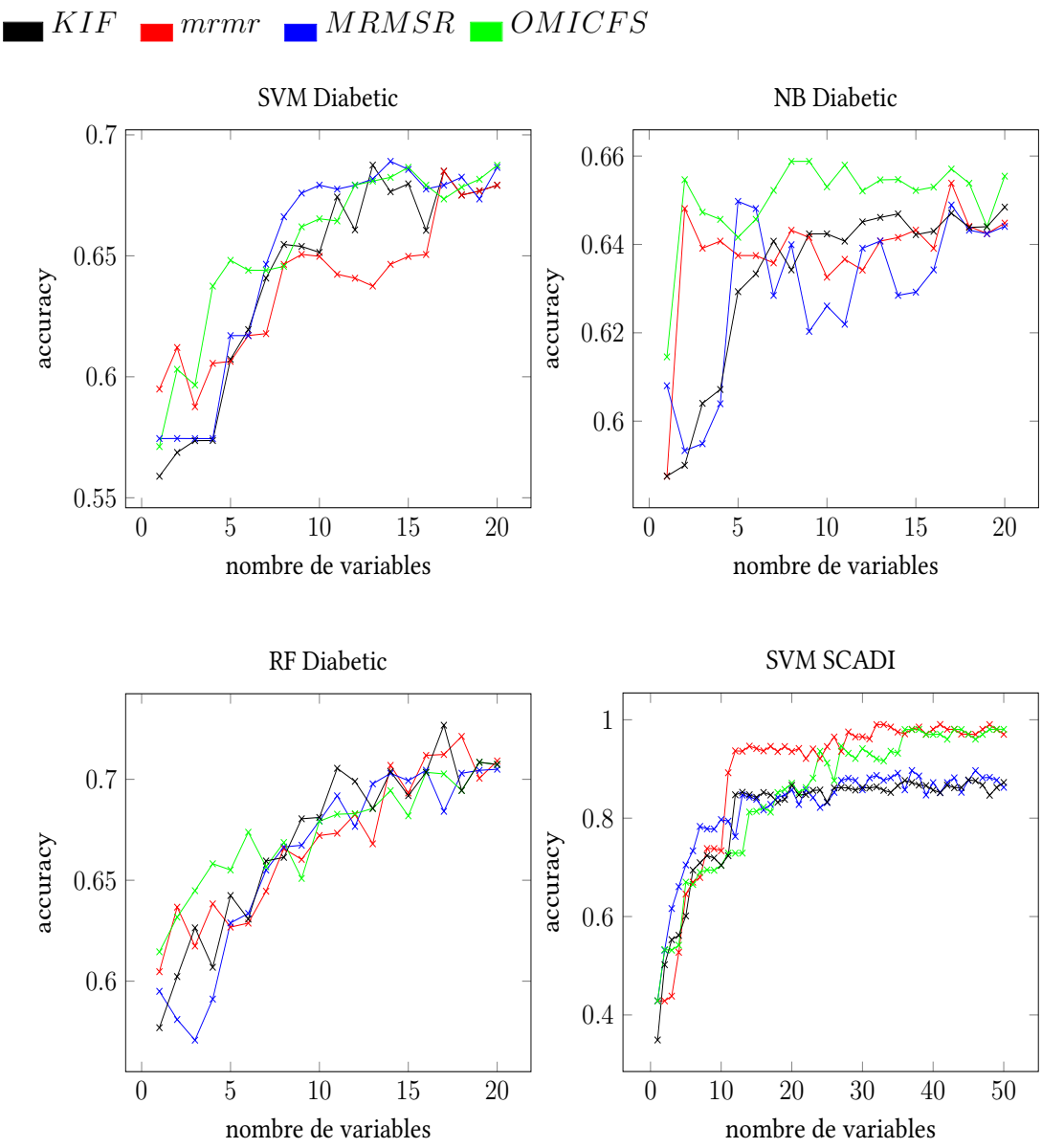
$$Recall = \frac{VP}{VP + FN}$$

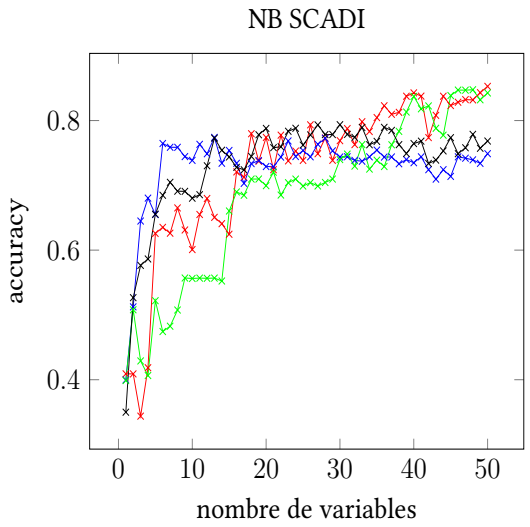
$$Presision = \frac{VP}{VP + FP}$$

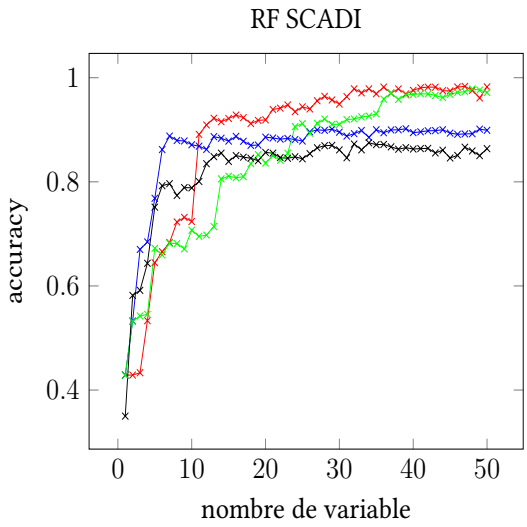
$$F1 = 2\frac{Presision \times recall}{presesion + recall}$$

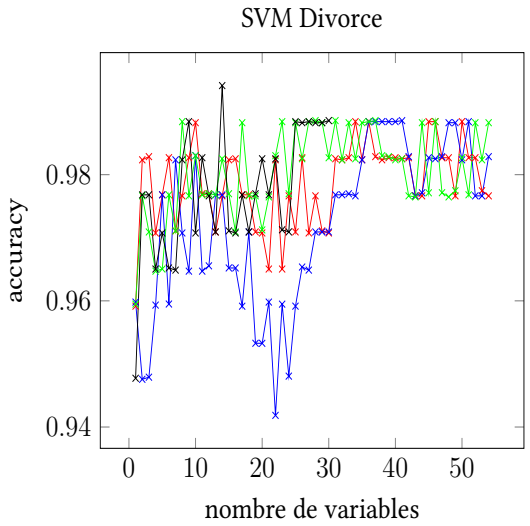
Protocole expérimental

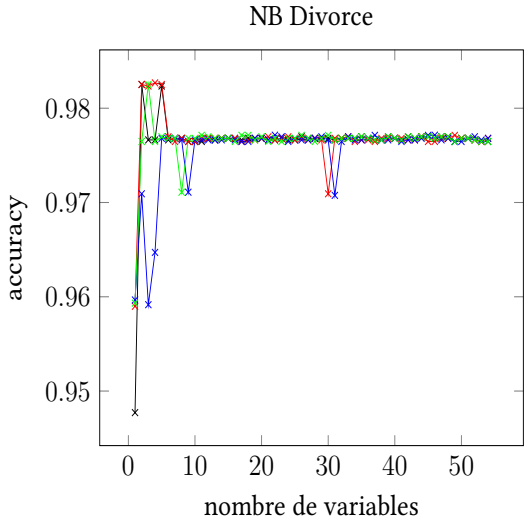
Avant de commencer les tests, il faut tout d'abord équilibrer les tableaux de données de sorte que le nombre des individus qui possèdent le même y_k soit le même pour toutes les classes de Y (voir A) Puis, nous construisons des modèles de classification suivante : Foret aléatoir avec 100 arbres, Naïve Bayes et classificateur SVM linéaire. Pour évaluer les méthodes de sélection, on utilise la validation croisée "5-fold cross validation". Le premier test sera comme suit, on va sélectionner chaque fois un nombre i de variables i de 1 jusqu'à P le nombre totale des variables explicatives chaque fois, on va entrainer les classifiers avec les i variables et obtenir l'accuracy. Les résultats sont représentés dans les graphes suivants

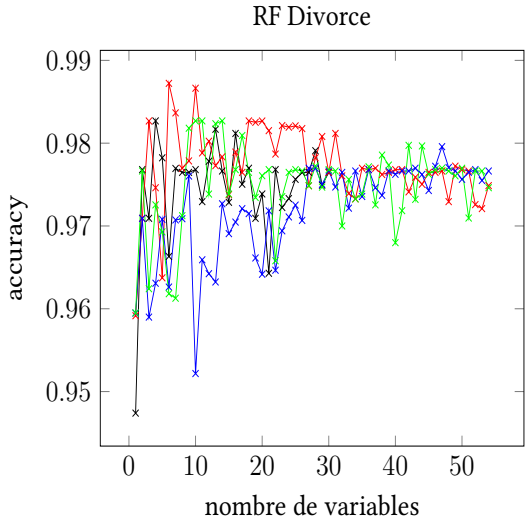

KIF
mrmr
MRMSR
OMICFS
SVM Diabetic
NB Diabetic
RF Diabetic
SVM SCADI
accuracy
nombre de variables

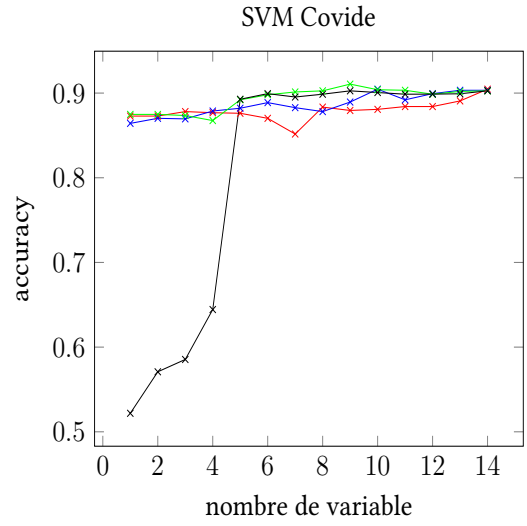

NB SCADI
accuracy
0.8
0.6
0.4
0
10
20
30
40
50
nombre de variables


RF SCADI
accuracy
1
0.8
0.6
0.4
0
10
20
30
40
50
nombre de variable


SVM Divorce
accuracy
0.98
0.96
0.94
0
10
20
30
40
50
nombre de variables


NB Divorce
accuracy
0.98
0.97
0.96
0.95
0
10
20
30
40
50
nombre de variables


RF Divorce
accuracy
0.99
0.98
0.97
0.96
0.95
0
10
20
30
40
50
nombre de variables


SVM Covide
accuracy
0.9
0.8
0.7
0.6
0.5
0
2
4
6
8
10
12
14
nombre de variable

Comparaison des points de pic sur les courbes d'accuracy correspondant à mRMR,mRMSR,OMICFS,KIF utilisons svm

Data	P	MRMR			MRMSR			OMICFS			KIF			GBS		
		N_S	F1	recall	N_S	F1	recall	N_S	F1	recall	N_S	F1	recall	N_S	F1	recall
Diabetic	20	17	0.70	0.77	14	0.70	0.76	20	0.70	0.77	13	0.70	0.76	8	0.61	0.62
Divorce	55	33	0.98	0.98	35	0.97	0.98	16	0.97	0.97	14	0.95	0.96	10	0.95	0.96
Covide	15	14	0.91	0.89	10	.90	0.89	9	0.90	0.88	11	0.91	0.89	7	0.87	0.87

Comparaison des points de pic sur les courbes d' accuracy correspondant à mRMR,mRMSR,OMICFS,KIF utilisons NB

Data	P	MRMR			MRMSR			OMICFS			KIF			GBS		
		N_S	F1	recall	N_S	F1	recall	N_S	F1	recall	N_S	F1	recall	N_S	F1	recall
Diabetic	20	17	0.71	0.79	5	0.61	0.62	9	0.67	0.72	10	0.68	0.72	8	0.81	0.76
Divorce	55	4	0.98	0.99	1	0.95	0.93	3	0.97	0.96	2	0.97	0.98	10	0.97	0.98
Covide	15	8	0.9	0.99	1	0.95	0.92	3	0.97	0.96	5	0.97	0.97	7	0.83	0.80

N_s représente le nombre de variables sélectionnés par la méthode

On calcule la moyenne de temps d'exécution de chaque algorithme et on obtient les résultats suivants : Pour sélectionner 6 variables

Data	MRMR	MRMSR	OMICFS	KIF
Diabetic	0.66	0.49	0.69	1.2
SCADI	5.54	3.92	1.28	66.4
Divorce	1.37	0.97	0.418	2.1
Covide	0.38	0.34	0.32	0.198

TABLE 4.2 – Temps de calcule des algorithmes de sélection

À travers les résultats obtenus, nous pouvons conclure que les méthodes étudiées performent bien toujours, chaque méthode présente des inconvénients,la méthode KIF est très couteuse en calcule, Tanit que la méthode OMICFS suppose la normalité des variables gaussiens et la méthode MRMR peut donner des résultats moins pertinents que les autres à cause du problème de la redondance non pertinente et finalement la méthode GBS sélectionne un petit nombre de variables

CHAPITRE 5.

EXPÉRIENCE SUR DES DONNÉES Á RÉPONSE MULTIPLE

Pour étudier l'algorithme de la sélection FMFS, sous la même configuration du chapitre précédent, on va faire deux expériences, l'une sur un tableau simulé et l'autre sur un problème réel.

5.1 Étude sur des données simulées

5.1.1 Description du tableau de données

on simule un tableau de données avec 15 variables explicatives et une variable a expliqué Y avec 4 modalités.[15] Tous d'abord on simule les variables explicatives par les étapes suivantes :

- $V_j \rightsquigarrow U([0,1])\ j = 1, ..., 10$
- $V_{11} = (V_1 - V_2)/2$
- $V_{12} = (V_1 + V_2)/2$
- $V_{13} = V_3 + 0.1$
- $V_{14} = V_4 - 0.2$
- $V_{15} = 2V_5$
- les autres $Vj \rightsquigarrow N(0, 20)$ pour j = 16, ..., p

les modalités du variable Y sont définis de sort que les variables pertinents seront V_{11}(ou V_1 et V_2), V_3 (ou V_{13}), V_4(ou V_{14} et $V_5(ouV_{15})$:

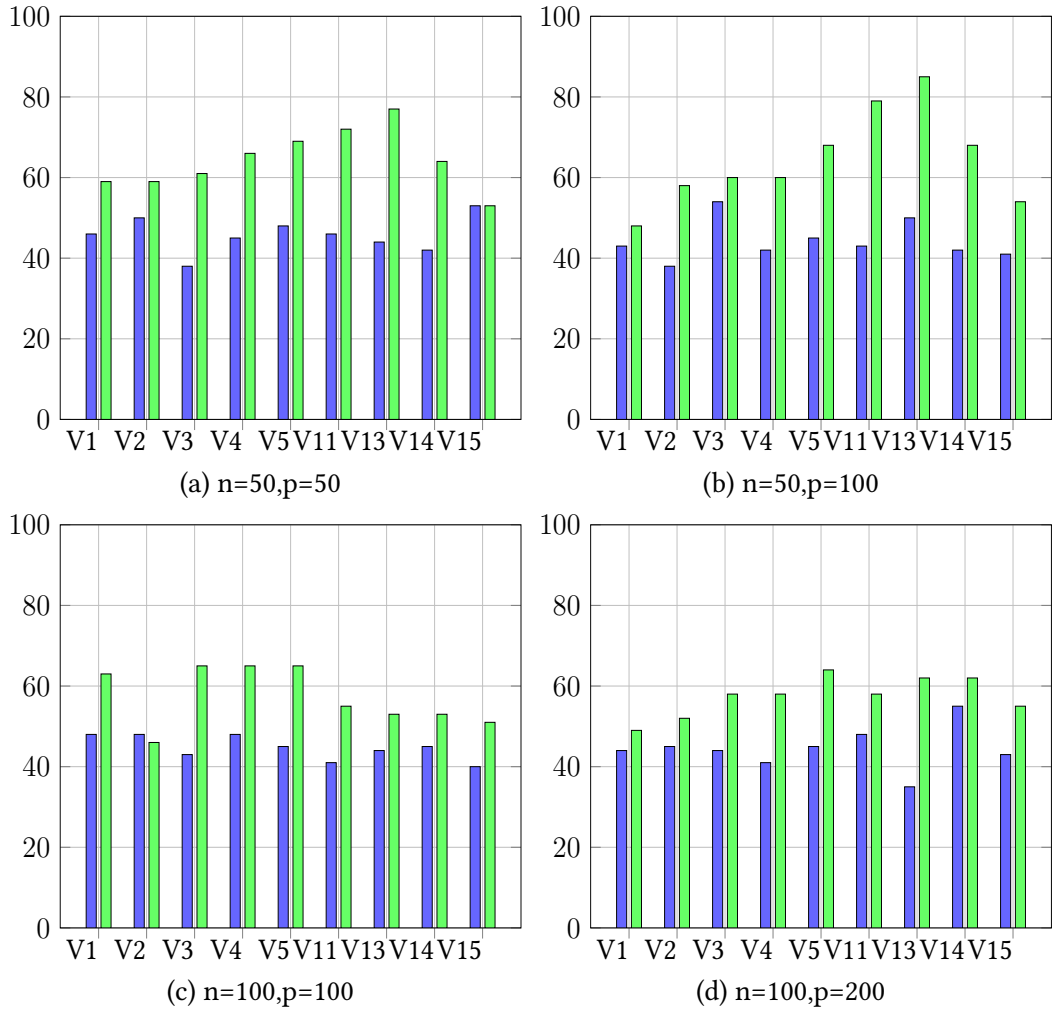
- $Y_1 = 1$ si $V1 > V_2$
- $Y_2 = 1$ si $V_4 > V_3$
- $Y_3 = 1$ si $Y_1 + Y_2 = 1$
- $Y_4 = 1$ si $V_5 > 0.8$
- $Y_i = 0$ sinon $(i = 1, 2, 3, 4)$

Protocole expérimental

on sélectionne les 9 premières variables à partir de 100 tableaux simuler et on calcule la fréquence d'apparitions des variables suivantes V1, V2, V3, V4, V5, V11, V13, V14,

V15 dans deux cas lorsque b=1 et lorsque b=2. les résultats obtenus sont représentés dans les figures suivantes :

$b = 1$ $b = 2$



(a) n=50,p=50 (b) n=50,p=100

(c) n=100,p=100 (d) n=100,p=200

La méthode FMFS est efficace pour sélectionner les variables pertinentes, mais il ne peut pas éliminer les variables redondantes, par exemple dans la figure (a) la variable V14 apparus à presque 80% des cas et la variable et que presque 65% des cas, la variable V_4 et V_{14} sont redondantes.

5.2 Étude sur des données réelles

le tableau de données qu'on va utiliser dans cette section est de IMDB Movies Dataset

Mesures d'évaluation

Contrairement à la classification à un seul label où la classification d'une nouvelle instance n'a que deux résultats possibles, correct ou incorrect. La classification multi-labels devrait également prendre en compte les résultats de chaque label. Donc, il faut utiliser d'autre mesures de qualités de prédiction pour le cas multi-lables [16] à tire d'exemple :

$$HammingLoss = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i \Delta Z_i}{|L|}$$

$$Subsetaccuracy = \frac{1}{N}\sum_{[}i = 1]^{N} I(Z_i = Y_i)$$

$$F - Measure = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap Z_i|}{|Y_i + Z_i|}$$

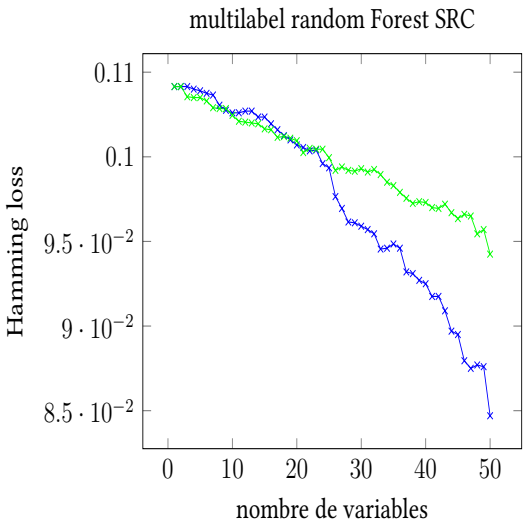
$$Accuracy(H, D) = \frac{1}{N}\sum_{i=1}^{N}|\frac{Y_i \cap Z_i}{Y_i \cup Z_i}|$$

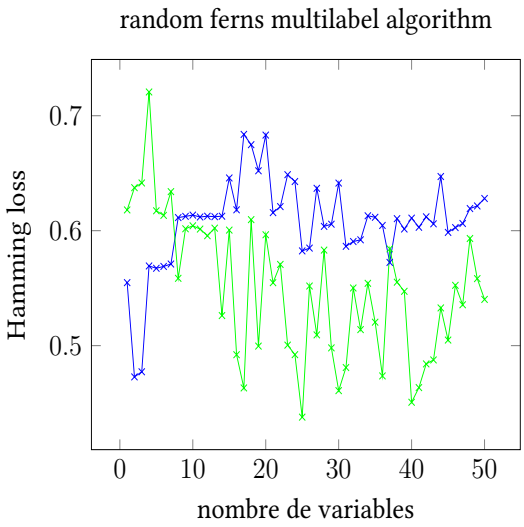
tel que Δ c'est la différence symétrique où $A\Delta B = (A \cup B) - (A \cap B)$
Y_i c'est la vraie valeur de Y pour chaque individu et Z_i c'est la valeur estimée de Y_i avec le modèle d'apprentissage. On note ici que Y_i est Z_i sont des vecteurs La fonction I prend deux valeurs, I(vraie)=1 et I(faux)=0

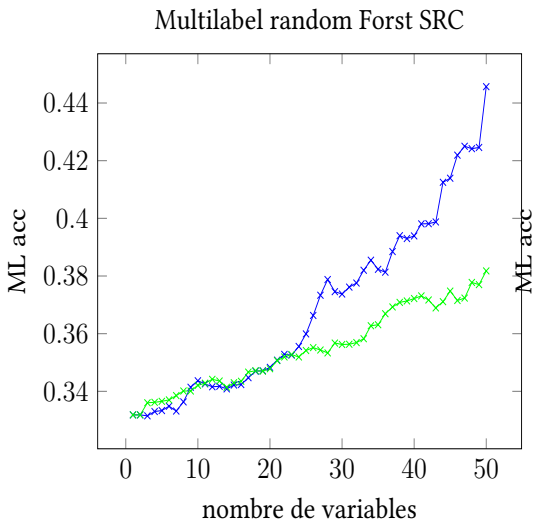
Protocole expérimental

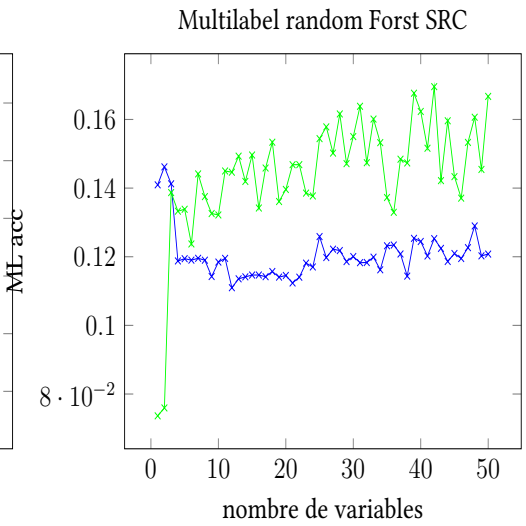
Après prétraitement et construction des variables prédictives à partir de text, on sélectionne les variables et on effectue un classifieur multi-labed puis on calculer les mesures de performance.

Les graphiques suivants confirment ce que nous avons dit dans la première simulation et le résultat théorique, la méthode FMFS peut sélectionner les variables pertinentes principales, mais elle ne prend pas en compte la redondance entre les variables


multilabel random Forest SRC
Hamming loss
0.11
0.1
$9.5 \cdot 10^{-2}$
$9 \cdot 10^{-2}$
$8.5 \cdot 10^{-2}$
0
10
20
30
40
50
nombre de variables


random ferns multilabel algorithm
Hamming loss
0.7
0.6
0.5
0
10
20
30
40
50
nombre de variables


Multilabel random Forst SRC
ML acc
0.44
0.42
0.4
0.38
0.36
0.34
0
10
20
30
40
50
nombre de variables


Multilabel random Forst SRC
ML acc
0.16
0.14
0.12
0.1
$8 \cdot 10^{-2}$
0
10
20
30
40
50
nombre de variables

CONCLUSION

Dans ce rapport, nous avons abordé quelques méthodes récentes utilisées pour la sélection des variables qui sont basées sur les critères d'association pour identifier les variables pertinentes et/ou redondantes. Dans le chapitre 1 nous avons présenté quelques critères d'association comme l'indice de kendall et l'information mutuelle. Puis le chapitre 2 été consacré aux méthodes de la sélection dans le cas où Y contient une seule réponse (single-label) tel que la méthode mRMR qui utilise une fonction score pour maximiser la pertinence et minimiser la redondance. Les deux méthodes mRMSR et OMICFS sont deux améliorations différentes de la méthode mRMR, pour surmonter le problème de la redondance la méthode GBS construit des groupes avec les variables similaires et sélection des variables à travers ces groupes. Finalement la méthode kif qui étudie les interactions entre les variables explicatives sous la base du taux de KENDALL au lieu d'information mutuelle. Le Chapitre 3 été dédié à la sélection dans le cas où le variable de Y est non disjonctive (multi-labels). Finalement, on a terminé par une étude expérimentale des méthodes ci-dessus.

L'étude de la redondance entre chaque paire de variables est utile pour la sélection, mais elle n'est pas suffisante, car il néglige l'effet collaboratif des ensembles de variables.

Pour le cas non disjonctive, il existe des difficultés d'identifier les variables redondantes, car deux variables peuvent être redondantes à une modalité de la variable Y (Label) et non redondantes par rapport aux autres.

PARTIE III.

ANNEXES

ANNEXE A.

LES DONNÉES DÉSÉQUILIBRÉES

A.1 Introduction

La plupart des classifieurs sont efficaces lorsque la distribution des classes de la variable de réponse Y dans l'ensemble de données est bien équilibrée. Cependant, ce n'est pas toujours le cas dans le monde réel, où une classe peut être représentée par un grand nombre d'exemples, tandis que l'autre n'est représentée que par quelques-uns seulement. C'est ce qu'on appelle le problème des données déséquilibrées.

A.1.1 La difficulté des données non équilibrées

Afin de mieux comprendre ce problème, on considère les deux problèmes de classification illustrée dans les figures A.1 A.2 :

La figure A.1 représente le cas d'un problème de classification binaire équilibré tandis que Dans le A.2 il y a un grand déséquilibre entre la classe majoritaire (+) et la classe minoritaire (-). Dans une situation similaire à celle illustrée à la Fig. A.2, les cas épars de la classe minoritaire peuvent confondre un classifieur comme le k-Nest Neighbor (k-NN). Par exemple, 5-NN peut classer incorrectement de nombreux cas de la classe minoritaire parce que les plus proches voisins de ces cas sont des exemples appartenant à la classe (+) à la classe majoritaire. le point noir dans le cas équilibré est classé dans (-), contrairement au cas non équilibré

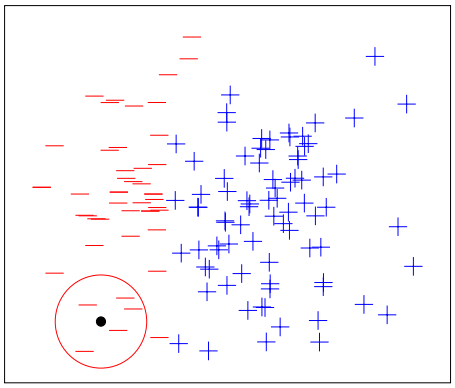


FIGURE A.1

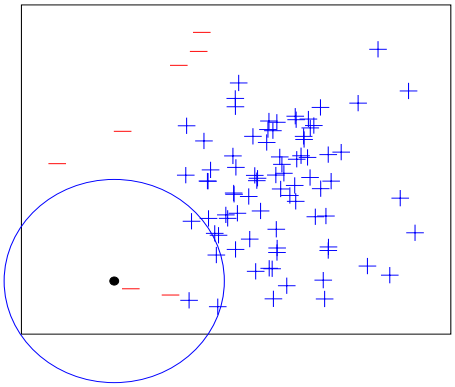
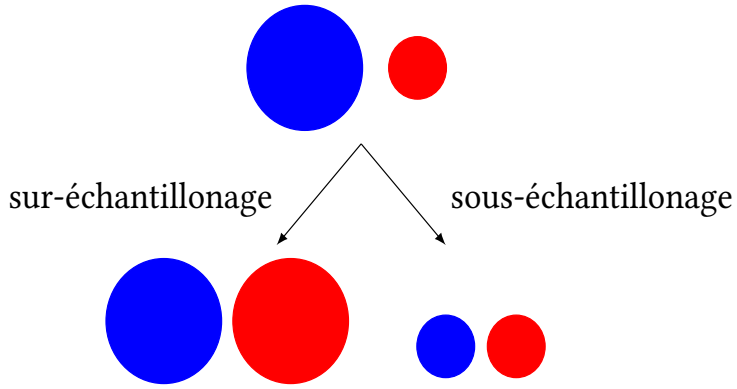


FIGURE A.2

A.2 Traitement des données déséquilibrées

A.2.1 Ré-échantillonnage

C'est une stratégie qui vise à échantillonner à nouveau l'ensemble de données original, soit par :
• le sur-échantillonnage de la classe minoritaire à travers une sélection aléatoire et avec répétition des éléments de la classe minoritaire jusqu'à ce que le nombre d'éléments des deux classes soit proche l'un de l'autre. Puisque cette stratégie duplique les éléments de la classe minoritaire, alors le modèle formé avec ces données peut être sur-ajusté (c'est-à-dire, le modèle fonction bien avec les données d'entraînement, mais donne des faux résultats pour les données de test).
• le sous-échantillonnant la classe majoritaire par l'extraction d'un sous ensemble de la classe majoritaire avec la même taille que la classe minoritaire par une sélection aléatoire. L'inconvénient de cette méthode est que beaucoup d'informations importantes dans les éléments négligés est perdue.



A.2.2 apprentissage sensible au poids

L'apprentissage sensible au poids consiste à ajuster les coûts des différentes classes au niveau des algorithmes d'apprentissage afin de contrer le déséquilibre des classes. Par exemple, pour le classifieur SVM [17] le Lagrangien dans le cas d'équilibre s'écrit :

$$L = \frac{\|w\|^2}{2} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i(w.x_i + b) - 1 + \xi_i\right] - \sum_{i=1}^{n} r_i\xi_i$$

avec $\alpha_i \geq 0$ et $r_i \geq 0$, La constante de pénalité C représente le compromis entre. Pour satisfait les conditions de KKT, il faut que $0 \leq \alpha_i \leq C$ et $\sum_{i=1}^{n}\alpha_i y_i = 0$ l'erreur empirique et la marge. comme le classifieur KNN dans le cas déséquilibre A.2 le SVM donnes des résultats erronées pour la classification. Nous pouvons surmonter ce problème en ajoutant un biais au modèle SVM. En utilisant des coûts d'erreur différents pour les deux classes, le Lagrangien devient alors :

$$L = \frac{\|w\|^2}{2} + C^{(+)}\sum_{i|y_i=1}^{n} + \xi_i C^{(-)}\sum_{i|y_i=-1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i(w.x_i + b) - 1 + \xi_i\right] - \sum_{i=1}^{n} r_i\xi_i$$

Les contraintes sur α_i devient : $0 \leq \alpha_i \leq C^{(+)}$ si $y_i = 1$ et $0 \leq \alpha_i \leq C^{(-)}$ si $y_i = -1$

Bibliographie

[1] G. Saporta, "Probabilités, analyse des données et statistique," *Editions technip*, 2006.

[2] O. Essaadouni, *Quelques généralisations des mesures d'association de Kendall et de Spearman pour des données discrètes multivariées.* PhD thesis, Université du Québec à Trois-Rivières, 2007.

[3] M. Thomas and A. T. Joy, *Elements of information theory.* Wiley-Interscience, 2006.

[4] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," in *2019 IEEE international conference on data science and advanced analytics (DSAA)*, pp. 442–452, IEEE, 2019.

[5] Y. Wang, X. Li, and R. Ruiz, "Feature selection with maximal relevance and minimal supervised redundancy," *IEEE Transactions on Cybernetics*, 2022.

[6] H. Lyu, M. Wan, J. Han, R. Liu, and C. Wang, "A filter feature selection method based on the maximal information coefficient and gram-schmidt orthogonalization for biomedical data mining," *Computers in biology and medicine*, vol. 89, pp. 264–274, 2017.

[7] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716–80727, 2020.

[8] L. Zheng, F. Chao, N. Mac Parthaláin, D. Zhang, and Q. Shen, "Feature grouping and selection : a graph-based approach," *Information Sciences*, vol. 546, pp. 1256–1272, 2021.

[9] E. La Chance, *Algorithmes pour le problème de l'arbre couvrant minimal.* PhD thesis, Université du Québec à Chicoutimi, 2014.

[10] Y. Anzarmou, A. Mkhadri, and K. Oualkacha, "The kendall interaction filter for variable interaction screening in high dimensional classification problems," *Journal of Applied Statistics*, pp. 1–19, 2022.

[11] G. Tsoumakas and I. Katakis, "Multi-label classification : An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[12] G. Tsoumakas and I. Katakis, "Multi-label classification : An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[13] J. Lee and D.-W. Kim, "Fast multi-label feature selection based on information-theoretic feature ranking," *Pattern Recognition*, vol. 48, no. 9, pp. 2761–2771, 2015.

[14] D. M. Powers, "Evaluation : from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv :2010.16061*, 2020.

[15] S. Singha and P. P. Shenoy, "An adaptive heuristic for feature selection based on complementarity," *Machine Learning*, vol. 107, no. 12, pp. 2027–2071, 2018.

[16] N. Spolaôr, E. A. Cherman, M. C. Monard, and H. D. Lee, "A comparison of multi-label feature selection methods using the problem transformation approach," *Electronic notes in theoretical computer science*, vol. 292, pp. 135–151, 2013.

[17] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *European conference on machine learning*, pp. 39–50, Springer, 2004.