

Les méthodes d'Induction d'Arbres

CHAID - CART - C4.5 et les autres...

Ricco RAKOTOMALALA

Différenciation des méthodes

Mesures d'Evaluation de la Segmentation -- Impact

- Mesures statistiques
- Mesures issues de la théorie de l'information

Regroupement des modalités

- 1 modalité = 1 branche
- Arbre Binaire
- Arbre m-aire

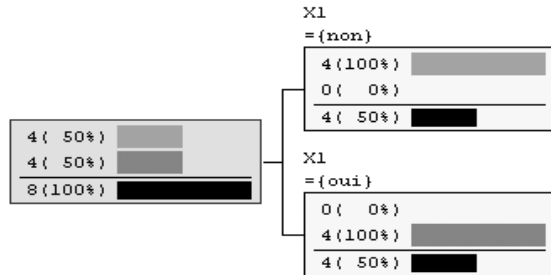
Détermination de la taille « optimale »

- Pré-pruning
- Post-pruning

Autres subtilités : coûts, graphes, arbres obliques, arbres flous

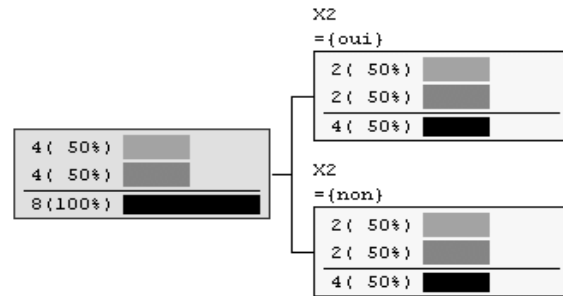
Evaluer une segmentation -- Impact

Comment les caractériser



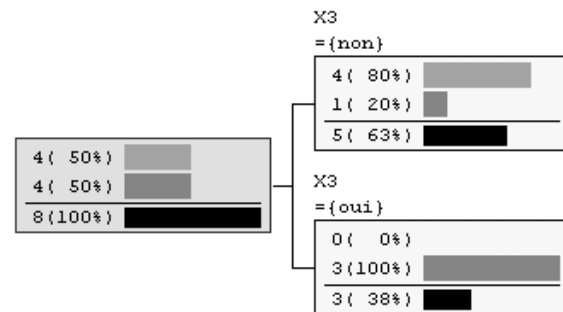
S1 : Maximalité

Distribution « pure » dans les feuilles



S2 : Minimalité

Pas de modification des distributions



S3 : Intermédiaire

Modification des distributions, association de certaines valeurs de X avec celles de Y

Impact

Mesures de liaison statistique – CHI-2 et ses normalisations (CHAID)

Tableau de calcul

Caractériser : la connaissance de X améliore la connaissance des valeurs de Y

Y / X	x_1	x_l	x_L	Σ
y_1		\vdots		
y_k	\dots	n_{kl}	\dots	$n_{k.}$
y_K		\vdots		
Σ		$n_{.l}$		n

Principe

*Comparer les valeurs observées avec les valeurs théoriques lorsque Y et X sont indépendants (produit des marges)
CHI-2 varie entre 0 et +oo*

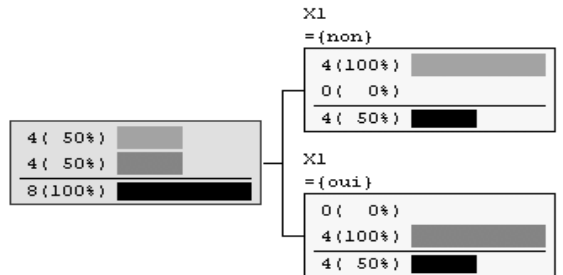
$$\chi^2 = \sum_{k=1}^K \sum_{l=1}^L \frac{\left(n_{kl} - \frac{n_{k.} \times n_{.l}}{n} \right)^2}{\frac{n_{k.} \times n_{.l}}{n}}$$

T de Tschuprow est une normalisation par les degrés de libertés. Il varie entre 0 et 1.

$$t^2 = \frac{\chi^2}{n \times \sqrt{(K-1) \times (L-1)}}$$

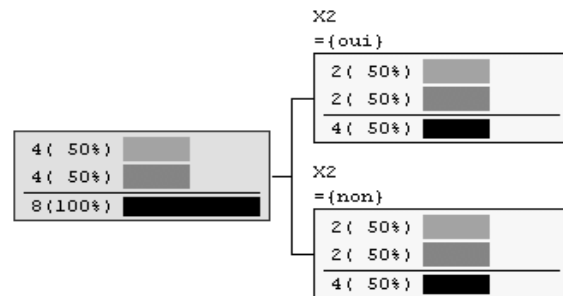
Impact

Exemple pour le t de Tschuprow -- CHAID



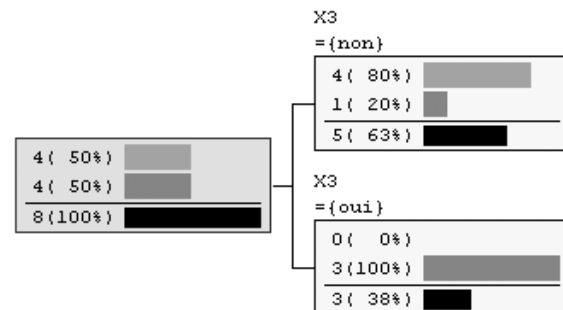
S1 : 1.0

Distribution enfant « pure »



S2 : 0.0

Pas de modification des distributions



S3 : 0.7746

Modification des distributions, association des valeurs de certaines valeurs de X avec celles de Y

Impact

Théorie de l'information – Le gain informationnel (C4.5)

Entropie de Shannon

Quantité d'information pour connaître les valeurs de Y

$$E(Y) = -\sum_{k=1}^K \frac{n_{k.}}{n} \times \log_2 \left(\frac{n_{k.}}{n} \right)$$

Entropie Conditionnelle

*Quantité d'information pour connaître les valeurs de Y
Sachant les valeurs de X*

$$E(Y / X) = -\sum_{l=1}^L \frac{n_{.l}}{n} \sum_{k=1}^K \frac{n_{kl}}{n_{.l}} \times \log_2 \left(\frac{n_{kl}}{n_{.l}} \right)$$

Gain d'entropie

$$G(Y / X) = E(Y) - E(Y / X)$$

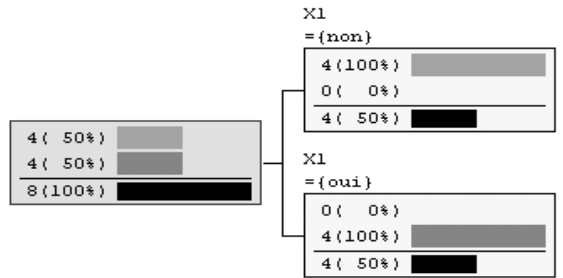
Gain d'entropie normalisée

*Gain Ratio - Tenir compte de
la distribution marginale de X*

$$GR(Y / X) = \frac{E(Y) - E(Y / X)}{E(X)}$$

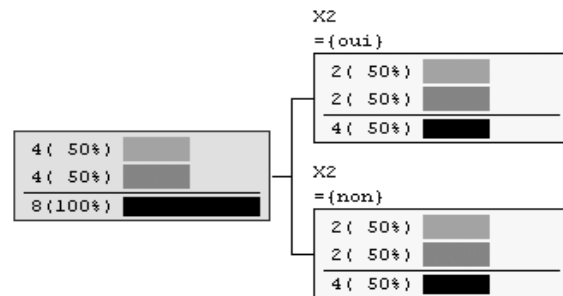
Impact

Exemple pour le gain ratio – C4.5



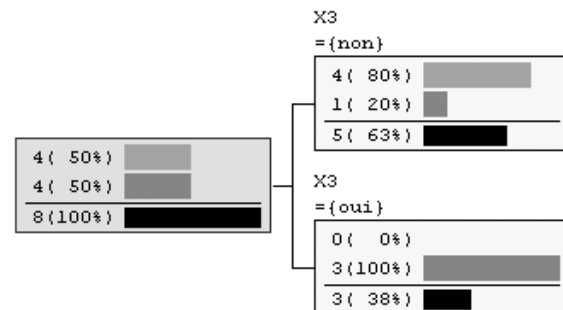
S1 : 1.0

Distribution « pure » dans les feuilles



S2 : 0.0

Pas de modification des distributions



S3 : 0.5750

Modification des distributions, association des valeurs de certaines valeurs de X avec celles de Y

Impact

Indice de concentration (CART)

Indice de Gini

Concentration des valeurs de Y

$$I(Y) = - \sum_{k=1}^K \frac{n_{k.}}{n} \times \left(1 - \frac{n_{k.}}{n} \right)$$

Indice de Gini conditionnel

*Concentration de Y
sachant les valeurs de X*

$$I(Y / X) = - \sum_{l=1}^L \frac{n_{.l}}{n} \sum_{k=1}^K \frac{n_{kl}}{n_{.l}} \times \left(1 - \frac{n_{kl}}{n_{.l}} \right)$$

Amélioration de la concentration

$$D(Y / X) = I(Y) - I(Y / X)$$

Indice de Gini = Entropie Quadratique

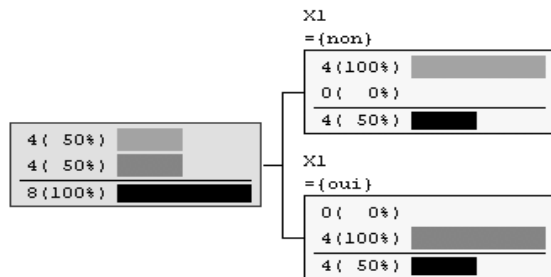
On peut aussi interpréter D comme un gain informationnel

Indice de Gini = Variance sur variables catégorielles

On peut aussi interpréter D comme une variance inter-classes = variance totale - variance intra

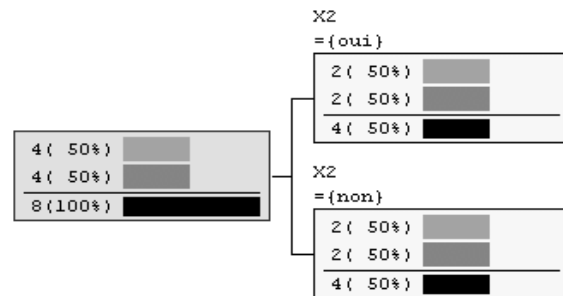
Impact

Exemple pour l'indice de Gini – CART



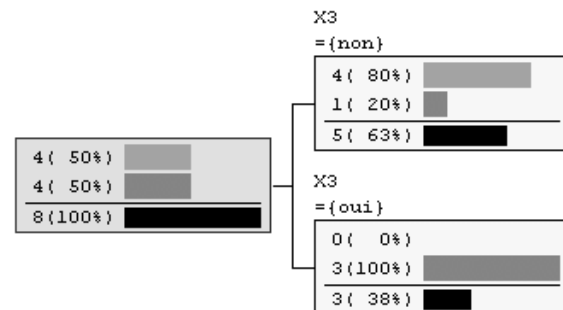
S1 : 0.5

Distribution « pure » dans les feuilles



S2 : 0.0

Pas de modification des distributions



S3 : 0.3

Modification des distributions, association des valeurs de certaines valeurs de X avec celles de Y

Impact -- Le rôle de la normalisation

Éviter la fragmentation des données – La propriété de Fusion des mesures

Y / X1	A1	B1	C1	D1	Total
positif	2	3	6	3	14
négatif	4	4	8	0	16
Total	6	7	14	3	30

CHI-2	3.9796
T Tschuprow	0.0766

Segmentation en 4 modalités avec la variable X1

Y / X2	A2	B2	D2	Total
positif	2	9	3	14
négatif	4	12	0	16
Total	6	21	3	30

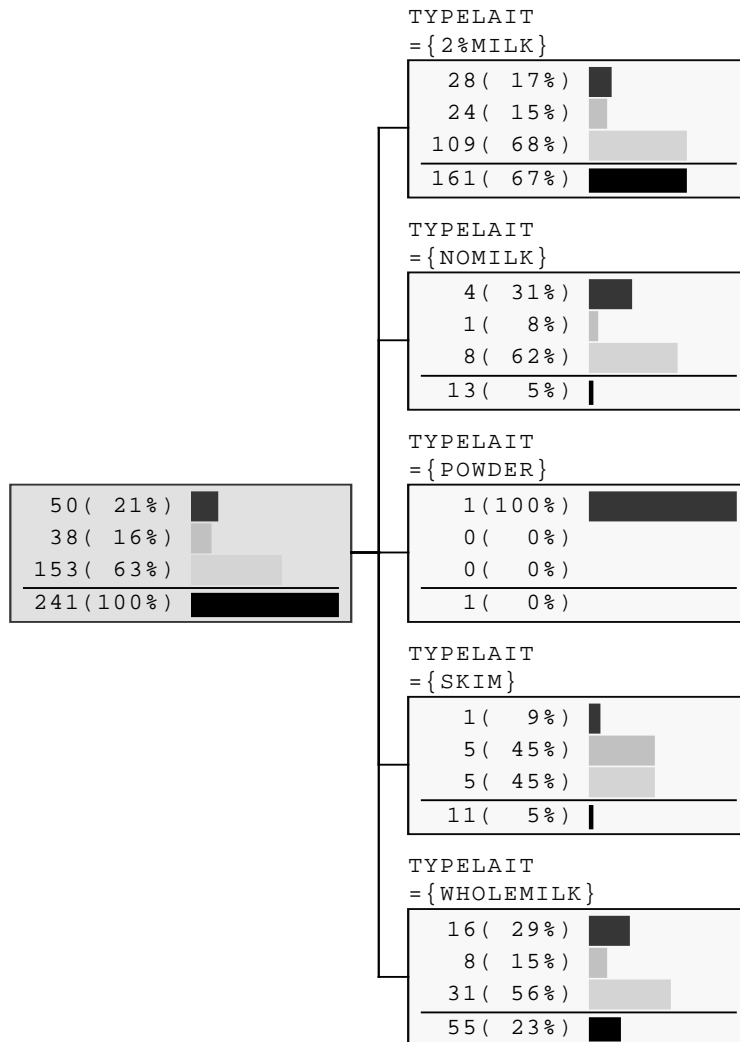
CHI-2	3.9796
T Tschuprow	0.0938

Segmentation en 3 modalités avec la variable X2

- Le t de Tschuprow normalise le CHI-2
 - Le Gain Ratio normalise le gain informationnel
 - Le Gain de Gini n'est pas normalisé
- (mais on s'affranchit autrement de cette limitation dans CART)*

Regroupement des modalités

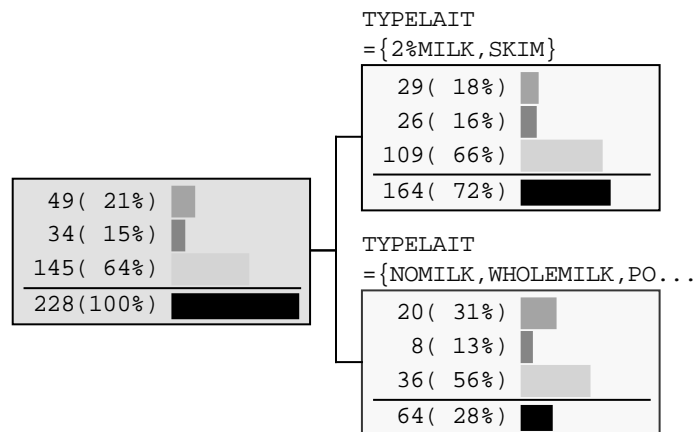
1 modalité = 1 branche de l'arbre – C4.5



- Simplicité du calcul et d'interprétation
- Danger de fragmentation, surtout sur les petits effectifs
- Arbres « larges »
- La mesure est chargée de favoriser les variables ayant peu de modalités

Regroupement des modalités

L'arbre binaire -- CART

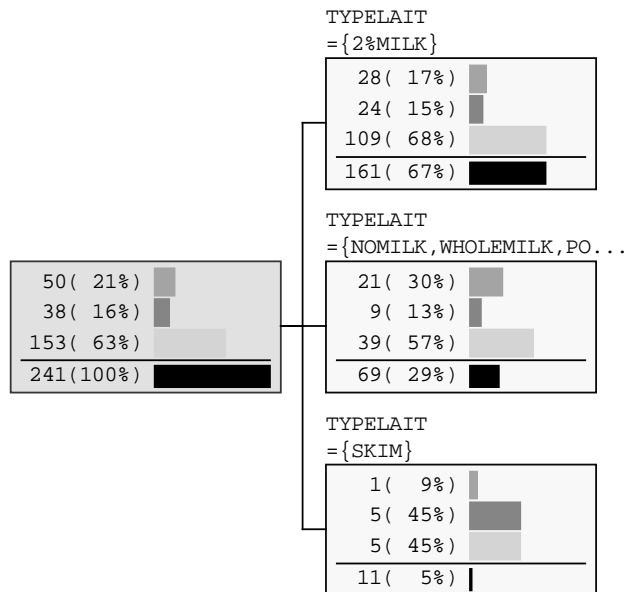


- Regroupement de manière à optimiser l'impact
- Moins de fragmentation
- Arbres « profonds »
- La binarisation compense l'absence de normalisation du gain de Gini
- La binarisation n'est pas toujours pertinente

Regroupement des modalités

L'arbre m-aire -- CHAID

- Regroupement des feuilles ayant le même « profil »
- Moins de fragmentation
- Difficulté à régler le paramètre de fusion



Principe : test d'équivalence distributionnelle
*Fusionner les feuilles issues de la segmentation
 Tant que les profils ne sont pas significativement
 différents*

	NoMilk, Powder	WholeMilk
High	5	16
Low	1	8
Normal	8	31
Total	14	55

$$\chi^2 = 14 \times 55 \times \left[\frac{(5/14 - 16/55)^2}{5+16} + \frac{(1/14 - 8/55)^2}{1+8} + \frac{(8/14 - 31/55)^2}{8+31} \right]$$

$$= 0.6309$$

$$p\text{-value}_{\chi^2_{[(3-1) \times (2-1)]}} = 0.73$$

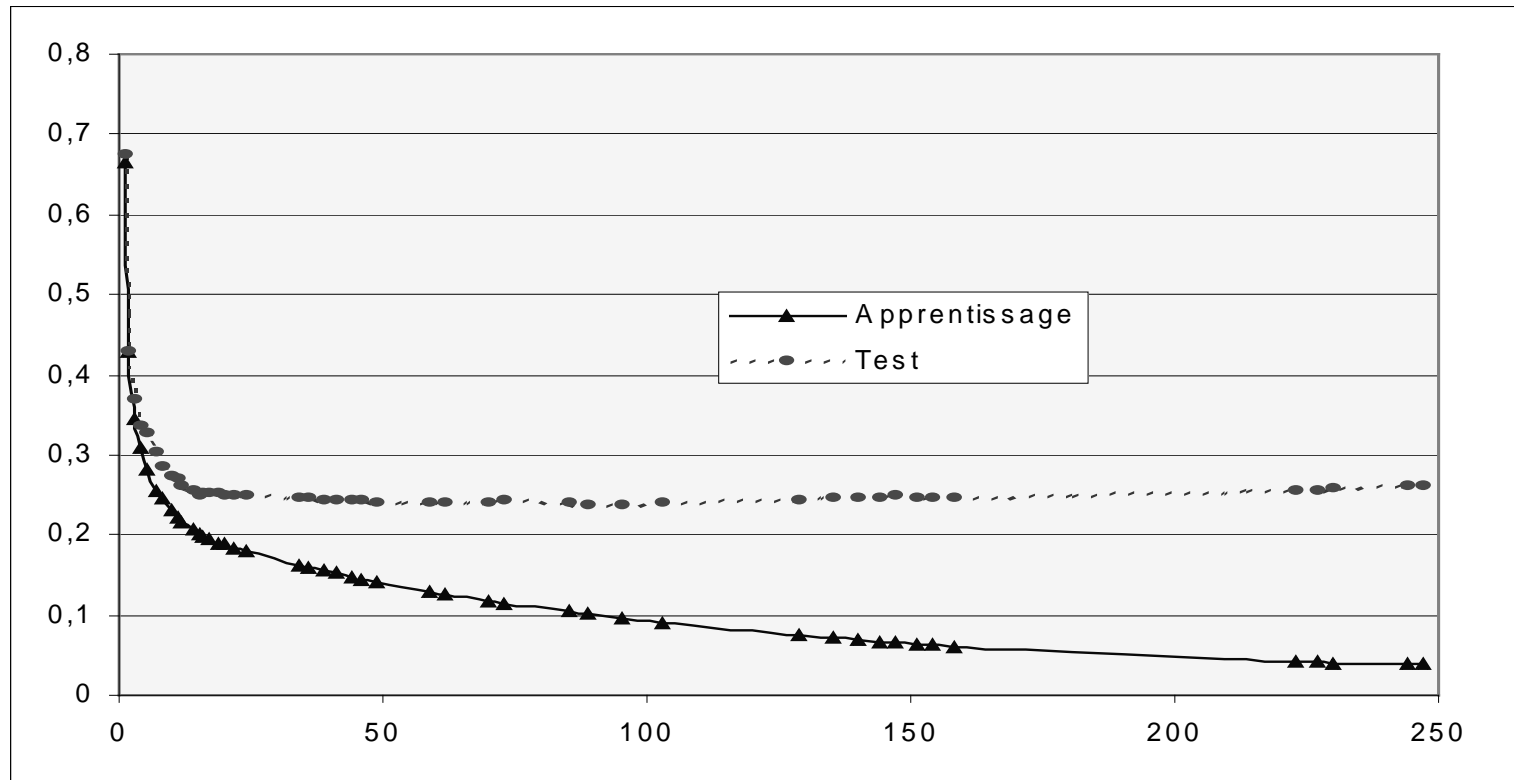
Fusion si (p-value > probabilité critique pour la fusion)

Détermination de la taille de l'arbre

Arbitrage biais - variance

Biais : (in)capacité à retraduire des fonctions / concepts « complexes »

Variance : dépendance au fichier d'apprentissage



Arbre sous-dimensionné Arbre « optimal » Arbre sur-dimensionné

Détermination de la taille de l'arbre

Pre-pruning

Critères empiriques

- Effectifs sur les nœuds et les feuilles : taille limite avant la segmentation et effectif d'admissibilité
- Pureté des feuilles : seuil de spécialisation
- Taille de l'arbre

Simple mais difficiles à déterminer (essais et tâtonnements, dépendant de la taille de la base et du domaine d'étude)

Critères statistiques -- CHAID

- Test d'indépendance du CHI-2

Difficile de déterminer un niveau de signification optimal (à fixer très bas à mesure que la taille de la base augmente)

Dans la pratique, ça marche quand même :

- la zone « optimale » est large
- rapidité en apprentissage (par rapport au post-pruning)
- à privilégier dans une phase exploratoire

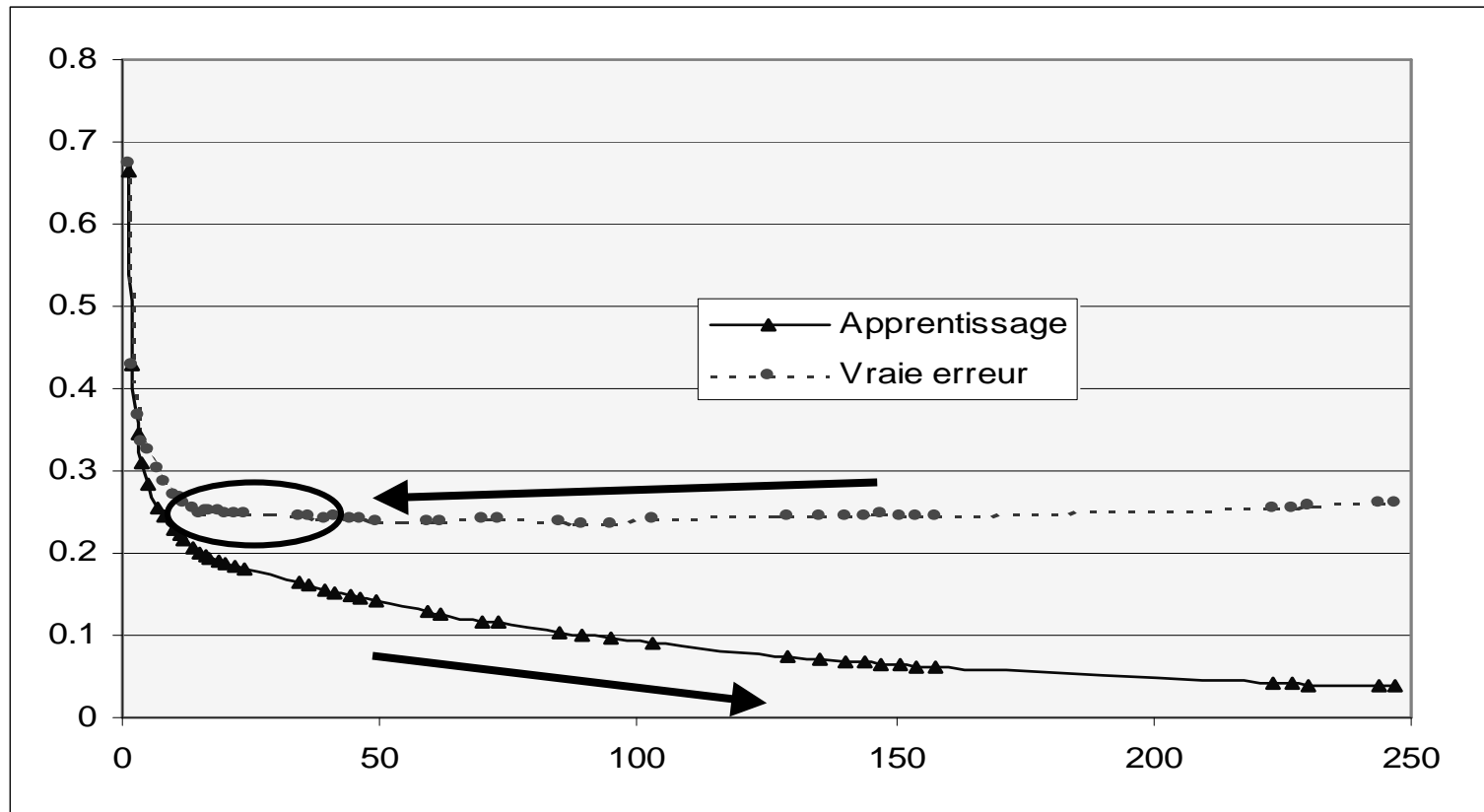


Détermination de la taille de l'arbre

Post-pruning

Apprentissage en deux phases

- (1) Expansion [growing] → maximiser la pureté
- (2) Élagage [pruning] → minimiser l'erreur de prédiction



Comment obtenir une estimation crédible de la « vraie » erreur



Détermination de la taille de l'arbre

Post-pruning avec un échantillon d'élagage -- CART

Subdivision de l'apprentissage en 2 parties

(1) Growing set (#67%)

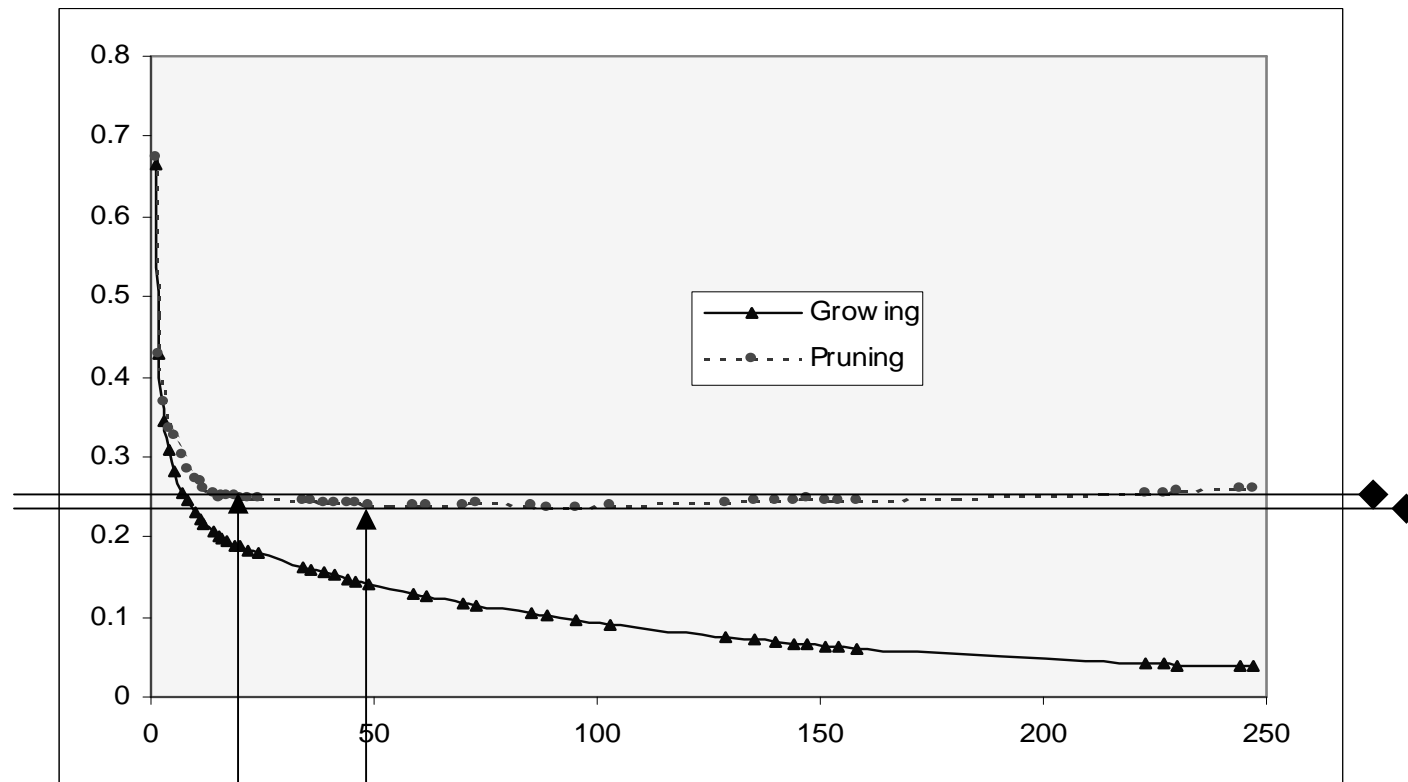
(2) Pruning set (#33%)

Estimation « honnête » de l'erreur

Séquences d'arbres de coût-complexité équivalents

$$E_{\alpha}(T) = E(T) + \alpha \times |T|$$

Éviter la trop grande dépendance à l'échantillon d'élagage

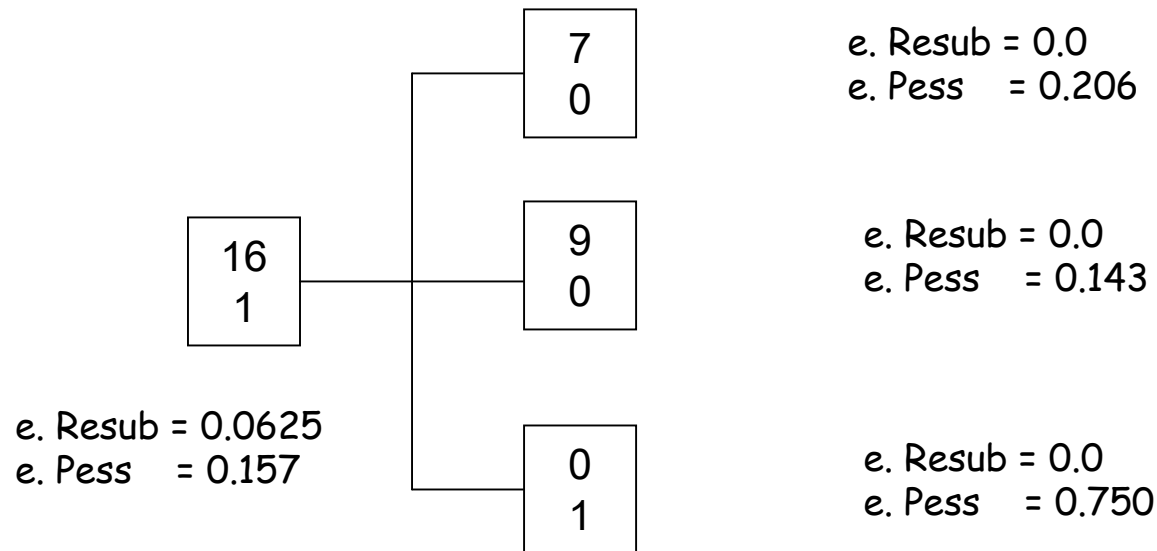


Arbre « 1-SE rule » Arbre « optimal »

Détermination de la taille de l'arbre

Post-pruning avec l'erreur pessimiste – C4.5

Erreur pessimiste = erreur pénalisée par les effectifs
= borne haute de l'intervalle de confiance du taux d'erreur



Stratégie :
Tester de proche en proche
chaque sommet précédant des
feuilles

$$\text{Élagage} : 0.157 < (7 \times 0.206 + 9 \times 0.143 + 1 \times 0.750)/17 = 0.2174$$

Récapitulatif

Caractéristiques des méthodes – CHAID, CART ou C4.5 ?

Carac. / Méthode	CHAID	CART	C4.5
Impact	T de Tschuprow	Indice de Gini	Gain informationnel (Gain Ratio)
Regroupement	M-aire Test d'équivalence distributionnelle	Binaire forcément	1 modalité = 1 branche
Détermination de la taille « optimale »	Effectif minimum pour segmenter Nombre de niveau de l'arbre Seuil de spécialisation Effectif d'admissibilité		
Détermination de la taille « optimale » (spécifique)	Pré-élagage avec test du CHI-2	Post élagage avec échantillon d'élagage	Post-élagage avec estimation pessimiste de l'erreur
Conseillé parce que / lorsque...	Phase exploratoire Grosses bases de données	Performances en classement Pas de paramétrage compliqué	Petits effectifs Incontournable chez les informaticiens (IA - ML) Peu sensible au paramétrage
Déconseillé parce que / lorsque...	Performances en classement Difficulté à trouver les « bons » paramètres	Petits effectifs Binarisation pas toujours appropriée	Post-élagage peu performant sur les grands effectifs Taille arbre fonction de la taille de la base

Aspect pratique

Prise en compte des coûts de mauvaise affectation -- CART

Dans les problèmes réels, les coûts de mauvaise affectation ne sont pas symétriques

Comment en tenir compte dans l'apprentissage ?

	Prédiction	
Observé \	Cancer	Non-Cancer
Cancer	0	5
Non-Cancer	1	0

Cancer	: 10 (33%)
Non-Cancer	: 20 (67%)

Ne pas tenir compte des coûts

$$\begin{aligned} E(\text{cancer}) &= 20/30 = 0.67 \\ E(\text{non-cancer}) &= 10/30 = 0.33 \end{aligned}$$

Décision = non-cancer \rightarrow E (Feuille) = 0.33

Tenir compte des coûts

$$\begin{aligned} C(\text{cancer}) &= 10/30 \times 0 + 20/30 \times 1 = 0.67 \\ C(\text{non-cancer}) &= 10/30 \times 5 + 20/30 \times 0 = 1.67 \end{aligned}$$

Décision = cancer \rightarrow C (Feuille) = 0.67

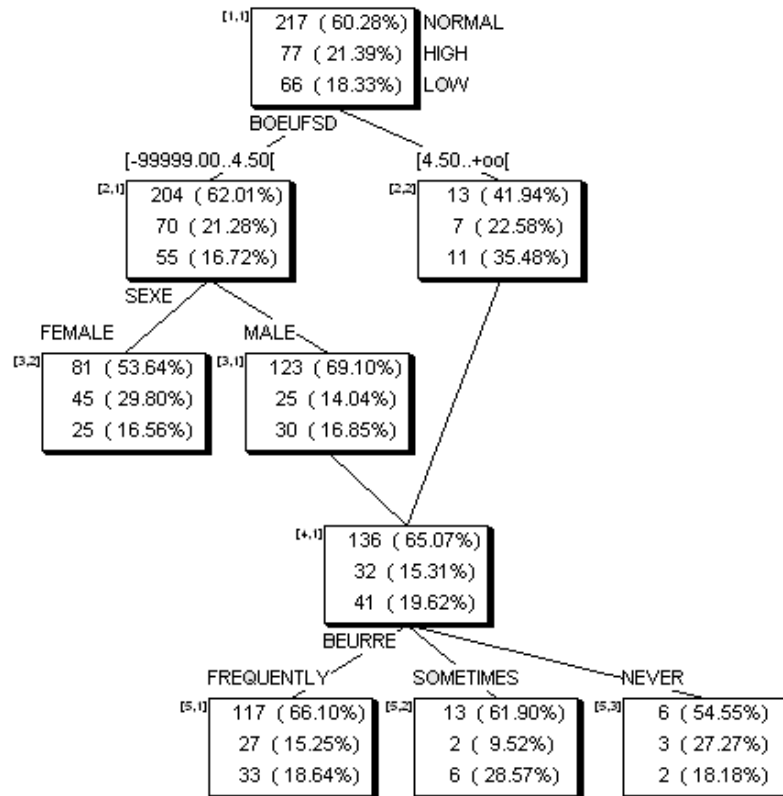
Stratégie de CART :

- (1) Définir séquences d'arbres de coût complexité équivalents
- (2) Choisir l'arbre qui minimise le coût de mauvaise affectation

$$C_{\alpha}(T) = C(T) + \alpha \times |T|$$

Autres subtilités

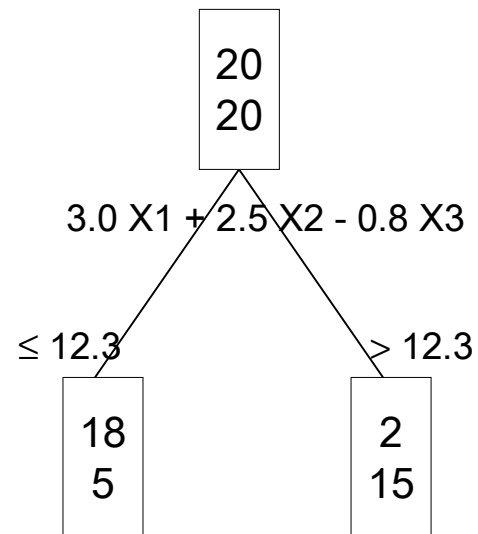
Les Graphes d'Induction – La méthode SIPINA (Zighed)



- Introduction de l'opérateur de « fusion »
- Amélioration du système de représentation
- Meilleure exploitation des petits effectifs
- Interprétation moins évidente des règles (ET / OU)
- Ne se démarque pas en termes de performances
- Graphes très « profonds » parfois

Autres subtilités

Les Arbres Obliques – OC1 (Murthy)



- Utilisation de combinaison linéaire de variables
- Amélioration du système de représentation
- Arbres plus concis

- Interprétation moins évidente des règles
- Complexité de calcul
- Pas tranchant face à des méthodes comme la LDA

Autres subtilités

Moralité de tout cela ?

- *Arbres flous*
- *Arbres à options*
- *Combinaisons logiques de variables*
- *Induction constructive*
- *Recherche en avant*

etc... cf. Rakotomalala (2005)

- (1) Les performances en classement sur données réelles sont peu probants
(2) Ces subtilités entraîne souvent une simplification de l'arbre (à performances égales)

