Resumé sur les arbres de decision

- **Algorithme**: Générer arbre de décision.
- *Générer un arbre de décision à partir des tuples d'apprentissage de données *
- partitionner D.
- **entrée:**
- Partition de données: D, un ensemble de tuples d'apprentissage avec leurs étiquettes de classe associés;
- Liste: l'ensemble des attributs;
- Attribut méthode de sélection, une procédure pour déterminer le critère de découpage
- les données tuples en classes individuelles. Ce critère constitue un attribut de fendage
- et, éventuellement, soit un point de partage ou sous-division.

Principe de construction d'un arbre de décision

LES ALGORITHMES de construction d'arbres :
ID3, C4.5 (Quinlan 84 et 86), CART(Breiman84) et CHAID(G. KASS 80).

Partant de:
un échantillon **S**,
un ensemble de classes {1,…, c}
un arbre de décision **t**,

Principe de construction d'un arbre de décision

- on opère le processus suivant :
- Si tous les exemples d'une branche sont de la **même classe**, alors on crée **une feuille** et on lui attribut **la classe correspondante**.
- Si on est dans le cas contraire, on refait le processus en enlevant l'attribut qui a déja été sélectionné.

Les algorithmes d'apprentissage pour la construction d'arbre de décision procèdent en 3 étapes.

Algorithme d'apprentissage générique

- **Entrée :** langage de description ; échantillon S ;
- **Début**
- Initialiser l'arbre vide ; la racine est le noeud courant
- **Répéter**
- Décider si le noeud courant est terminal ;
- **Si** le noeud est terminal
- **Alors**
- Affecter une classe ;
- **Sinon**
- Sélectionner un test et créer le sous-arbre ;
- **FinSi**
- Passer au noeud suivant non exploré s'il en existe ;
- **Jusqu'à** obtenir un arbre de décision
- **Fin.**

C4.5

- **Gain Ratio:** L'algorithme d'apprentissage C4.5 utilise la notion complémentaire au gain informationnel appelée **Gain Ratio**

- $SplitInfo(X) = -\sum_{i=1}^{n} \frac{|Di|}{|D|} * log_2\left(\frac{|Di|}{|D|}\right)$

- $GainRatio(X) = \frac{Gain(X)}{SplitInfo(X)}$

Gain informationnel

- $Gain(X) = Info(D) - \sum_{i=1}^{nbTest} \frac{|Di|}{|D|} * Info(Di)$
- |D| représente le nombre d'exemples à évaluer, nbTest est le nombre de valeurs pour l'attribut testé, |Di| est le nombre d'exemples qui correspondent à la valeur i de l'attribut testé.

- *Info(D)* est **la fonction entropie**, dont l'équation est la suivante :

- $Info(D) = -\sum_{j=1}^{k} \frac{freq(Cj,D)}{card(D)} * log_2\left(\frac{freq(Cj,D)}{card(D}\right))$

CIEL	TEMPERATURE	HUMIDITE	VENT	JOUER
Ensoleillé	Elevé	Forte	Faux	Non
Ensoleillé	Elevé	Forte	Vrai	Non
Couvert	Elevé	Forte	Faux	Oui
Pluvieux	Moyenne	Forte	Faux	Oui
Pluvieux	Basse	Normale	Faux	Oui
Pluvieux	Basse	Normale	Vrai	Non
Couvert	Basse	Normale	Vrai	Oui
Ensoleillé	Moyenne	Forte	Faux	Non
Ensoleillé	Basse	Normale	Faux	Oui
Pluvieux	Moyenne	Normale	Faux	Oui
Ensoleillé	Moyenne	Normale	Vrai	Oui
Couvert	Moyenne	Forte	Vrai	Oui
Couvert	Elevé	Normale	Faux	Oui
Pluvieux	Moyenne	Forte	Vrai	Non

Pour cet ensemble :

$$\text{Info}(D) = -9/14 * \log_2(9/14) - 5/14 * \log_2(5/14) = 0{,}940$$

- Par la suite, on évalue le Gain, SplitInfo, et Ratio pour chaque attribut.
- Pour l'attribut Ciel le Gain est :
- **Gain (Ciel)** = 0,940- 5/14 * (-3/5 * $\log_2(3/5)$ –2/5 * $\log_2(2/5)$) – 4/14 *(- 4/4 * $\log_2(4/4)$ - 0/4 * $\log_2(0/4)$) - 5/14 * (-3/5 * $\log_2(3/5)$ - 2/5 * $\log_2(2/5)$) = 0,247

- **SplitInfo(Ciel)=** -(5/14) * $\log_2(5/14)$-(4/14) * $\log_2(4/14)$- (5/14) * $\log_2(5/14)$ = 1,577
- GainRatio(Ciel) = 0,247 / 1.577 = 0.157

Attribut	SplitInfo	Gain	Ratio
CIE_L	1.577	0,247	0.157
TEMPERATURE	1.557	0.029	0.0186
HUMIDITE	1	0.152	0.152
VENT	0.985	0.048	0.0487

Pour la branche Ensoleillé on sélectionne les attributs qui n'ont pas été sélectionné avant, en recalculant pour chacun de ces derniers les différentes mesures.

Tableau : Calcul du gain informationnel pour l'ensemble d'apprentissage avec Ciel = 'Ensoleillé'.

Attribut	SplitInfo	Gain	Ratio
Température	1.522	0.571	0.375
Humidité	0.971	0.971	1
Vent	0.981	0.020	0.002

C4.5

- Pour VENT le ratio est le meilleur
- Donc:

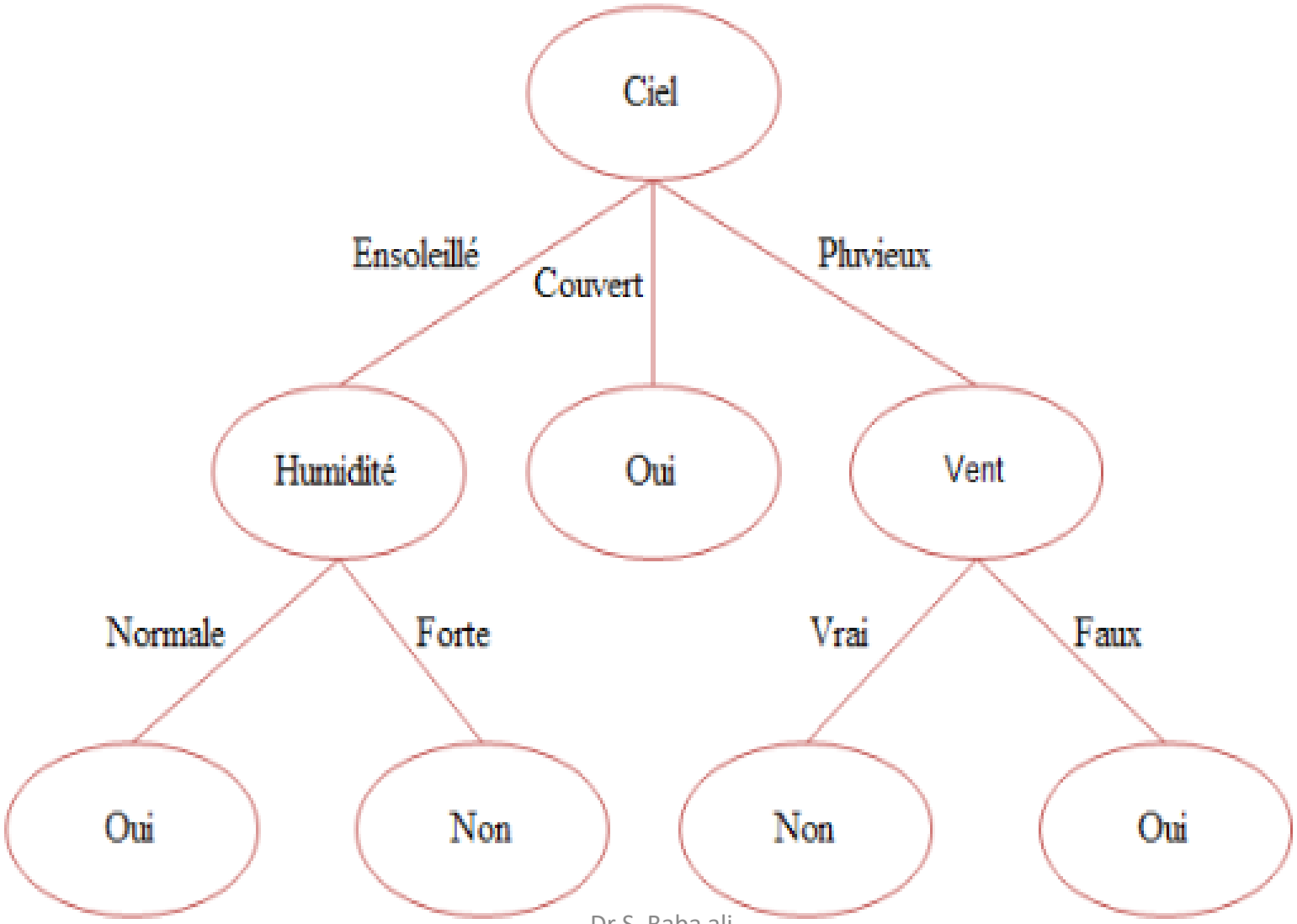
Si ciel = pluvieux et Vent =non alors Jouer =OUI

Dans cet ensemble, les exemples donnent:

si Ciel= Pluvieux et Vent= Vrai alors la classe Jouer =Non

on peut créer un nœud terminal.

L'ARBRE obtenu par C4.5 pour WEATHER



Donc on trouve cet il faut retrouver cet arbre

1- Rederouler cet exemple dans tous ces details pour retrouver l’ arbre

2- derouler c4.5 SUR L’EXEMPLE DE La Base d’ apprentissage « Bys computer »