Une approche probabiliste pour le classement d'objets incomplètement connus dans un arbre de décision

Lamis HAWARAH

Laboratoire TIMC-IMAG
Grenoble

22 octobre 2008

Les valeurs manquantes

Les valeurs manquantes

Problème

- Analyse statistique des données
- Aide à la décision à partir de données

Les valeurs manquantes

Problème

- Analyse statistique des données
- Aide à la décision à partir de données

Fouille de données (Data Mining)

Construction et exploitation du modèle

- Réseaux bayésiens
- Règles d'association
- Réseaux de neurones
- **Arbres de Décision**
- ....

Les valeurs manquantes

Pourquoi ?

- Oubli de l'enregistrement de la valeur
- Coût d'acquisition élevé
- Examen médical non effectué
- Refus de répondre à certaines questions

Les valeurs manquantes

Pourquoi ?

- Oubli de l'enregistrement de la valeur
- Coût d'acquisition élevé
- Examen médical non effectué
- Refus de répondre à certaines questions

Conséquence

- Donnée manquante : donnée complexe
- Décision non représentative, voire dangereuse (médecine)

Réponse (1)

Ignorer la valeur manquante : suppression des objets

Réponse (1)

Ignorer la valeur manquante : suppression des objets

Conséquences

- Réduction du volume des données
- Base non représentative (perte d'information)
- Information extraite non significative

Réponse (1)

Ignorer la valeur manquante : suppression des objets

Conséquences

- Réduction du volume des données
- Base non représentative (perte d'information)
- Information extraite non significative

Réponse (2)

Imputation : valeur la plus commune, moyenne...

Réponse (1)

Ignorer la valeur manquante : suppression des objets

Conséquences

- Réduction du volume des données
- Base non représentative (perte d'information)
- Information extraite non significative

Réponse (2)

Imputation : valeur la plus commune, moyenne...

Conséquences

- Relations entre les attributs modifiées
- Probabilité élevée de se tromper

Réponse (1)

Ignorer la valeur manquante : suppression des objets

Conséquences

- Réduction du volume des données
- Base non représentative (perte d'information)
- Information extraite non significative

Réponse (2)

Imputation : valeur la plus commune, moyenne...

Conséquences

- Relations entre les attributs modifiées
- Probabilité élevée de se tromper

Réponse (3)

Traitement de données manquantes internes à l'algorithme
Ex : C4.5, CART...

Objectif

Aide à la décision en présence de données manquantes de *manière probabiliste*

Objectif

Aide à la décision en présence de données manquantes
de *manière probabiliste*

Exemple

En médecine, pour un patient qui pourrait avoir trois maladies m_1, m_2, m_3, il serait préférable d'estimer la probabilité relatives de m_1, m_2, m_3 plutôt que de lui affecter une seule maladie

Objectif

Aide à la décision en présence de données manquantes
de *manière probabiliste*

Exemple

En médecine, pour un patient qui pourrait avoir trois maladies m_1, m_2, m_3, il serait préférable d'estimer la probabilité relatives de m_1, m_2, m_3 plutôt que de lui affecter une seule maladie

Arbre de décision

- Prédictif et descriptif
- Facile à comprendre et simple à construire et à utiliser
- Proche des structures de connaissances manipulées naturellement par l'esprit humain

Plan

Arbre de Décision

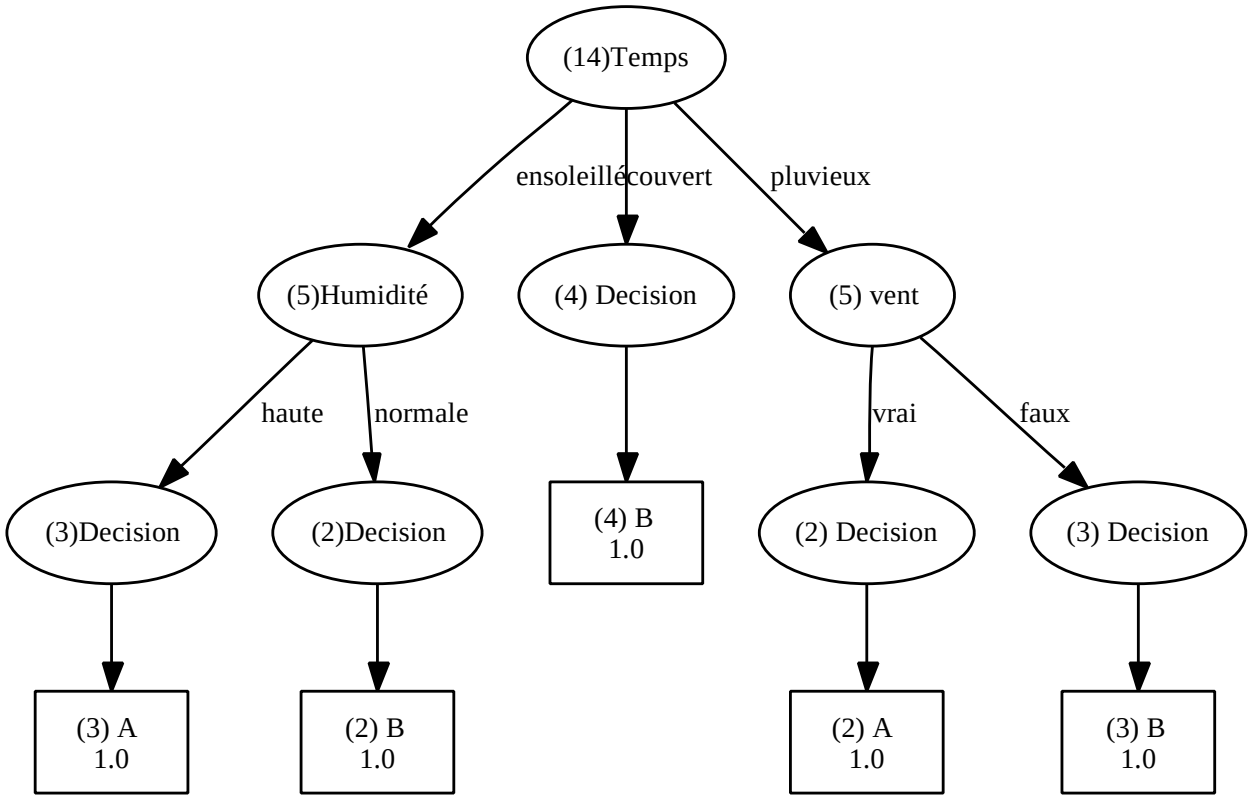
Apprentissage
inductif, supervisé

Extrait de la base météo : Ensemble d'apprentissage

id	Temps	Température	Humidité	Vent	Classe
1	Ensoleillé	Elevée	Haute	Faux	A
2	Ensoleillé	Elevée	Haute	Vrai	A
3	Couvert	Elevée	Haute	Faux	B
7	Couvert	Basse	Normale	Vrai	B
12	couvert	Moyenne	Haute	Vrai	B
13	Couvert	Elevée	Normale	Faux	B

L'attribut à prédire (classe) A : Je sors, B : Je reste à la maison

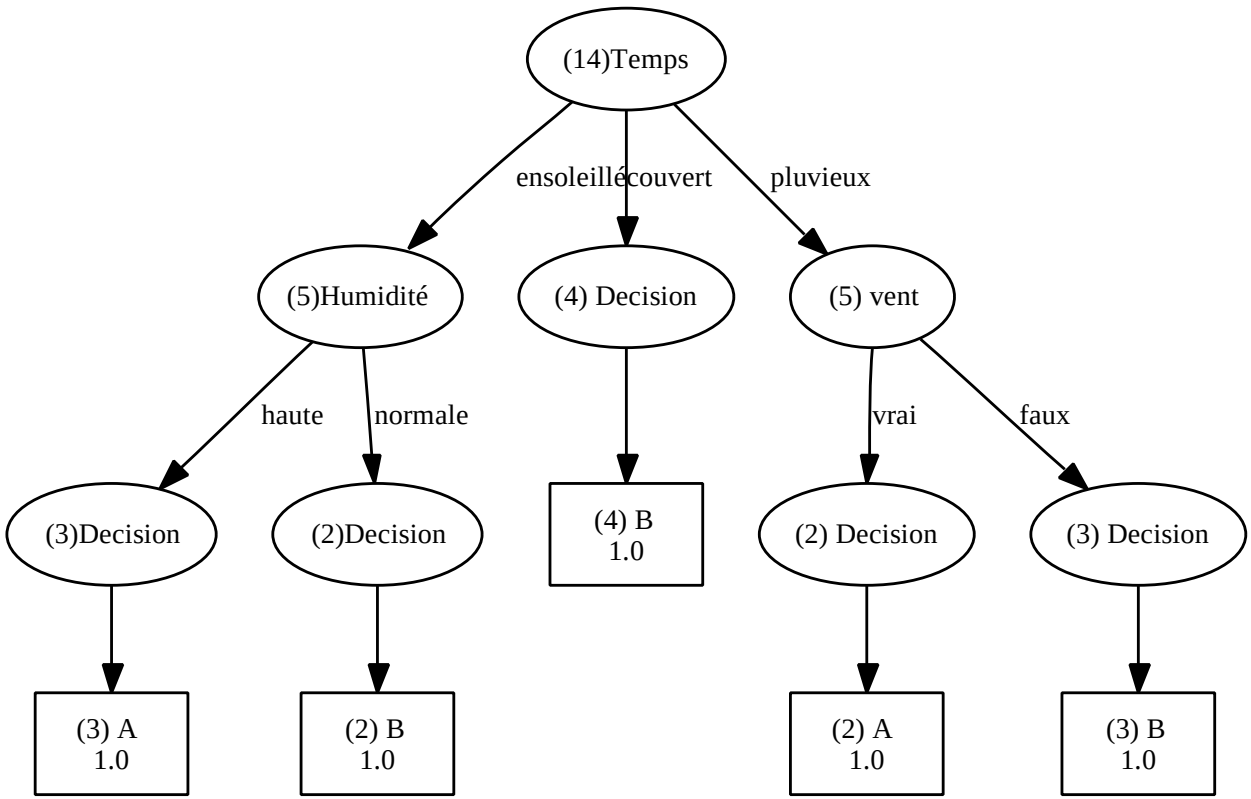
Arbre de Décision : Phase de construction



Arbre de Décision : Phase de classement

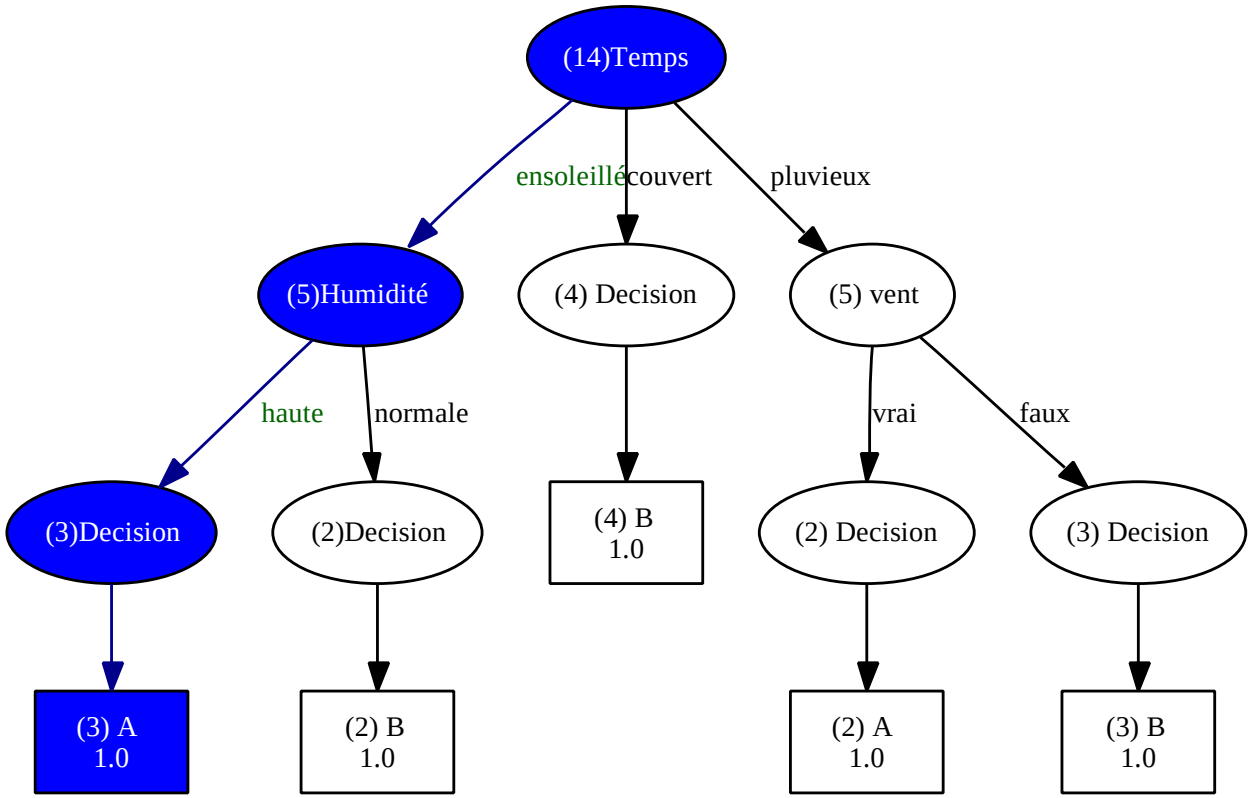
Détermination de la classe d'un nouvel objet **complet**

Temps	Température	Humidité	Vent
Ensoleillé	moyenne	haute	Faux



Arbre de Décision : Phase de classement

Temps	Température	Humidité	Vent
Ensoleillé	moyenne	haute	Faux



- Parcours l'arbre depuis sa racine jusqu'à une feuille
- Classe associée à cette feuille : classe de l'objet : A

Arbre de Décision : Phase de classement

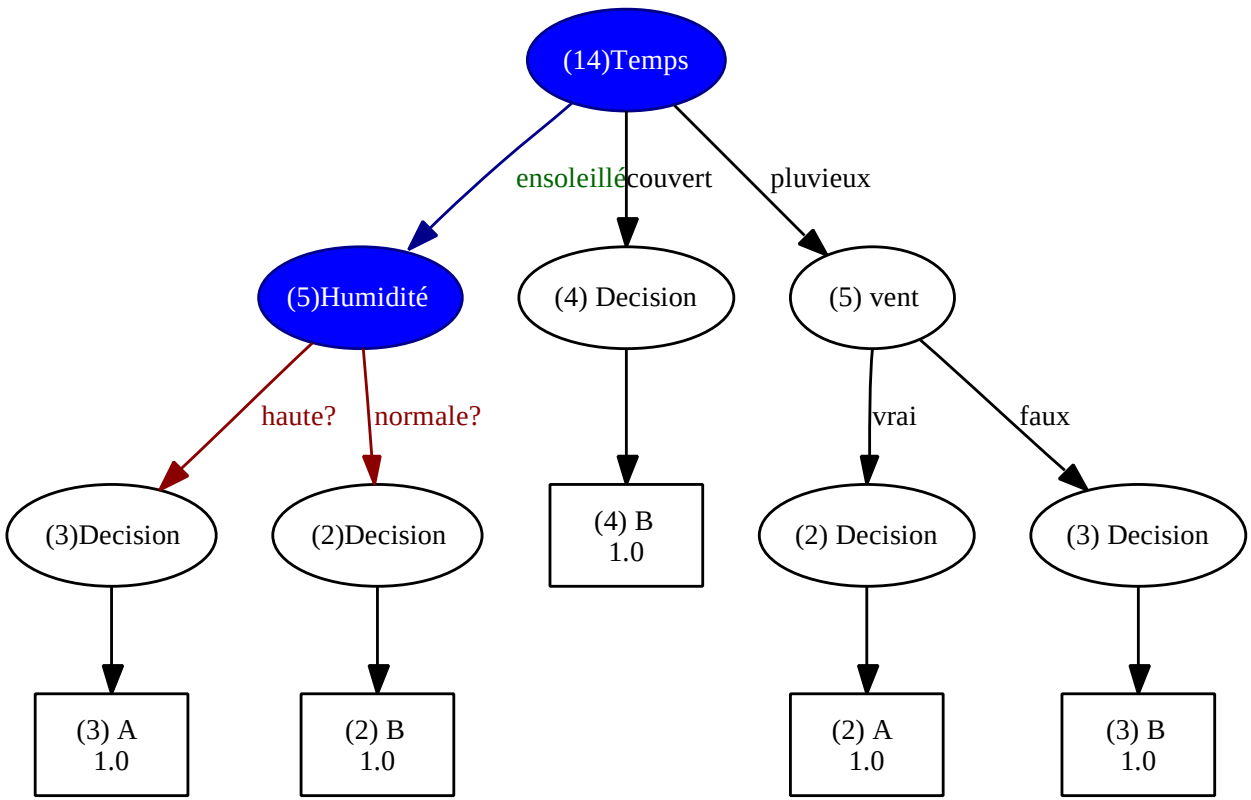
Détermination de la classe d'un nouvel objet **incomplet**

Temps	Température	Humidité	Vent
Ensoleillé	moyenne	?	Faux

Arbre de Décision : Phase de classement

Détermination de la classe d'un nouvel objet **incomplet**

Temps	Température	Humidité	Vent
Ensoleillé	moyenne	?	Faux



Plan

Algorithme de construction d'un arbre de décision

- Partition récursive de l'ensemble d'apprentissage en sous-ensembles plus homogènes
- Choix des attributs à tester
  - Attributs permettant le mieux de classer les exemples
  - Mesure formelle : choix des attributs pertinents et élimination des attributs inutiles
- Information Mutuelle [Shannon, 1949] :
  - Mesure de la force de la relation entre deux attributs
  - Réduction de l'incertitude sur C lorsque A est connu

$$gain(A, C) = IM(A, C) = H(C) - H(C|A)$$

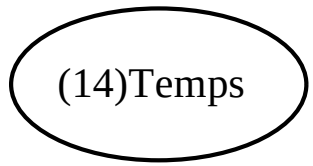
$$H(C) = \sum_{i=1}^{i=k} P_i log_2 P_i$$

$$H(C|A) = \sum_{i=1}^{i=n} P(A = a_i) H(C|a_i)$$

Algorithme de construction d'un arbre de décision

IM(Classe, Temps) = 0.246
IM(Classe, Température) =0.030
IM(Classe, Humidité) = 0.154
IM(Classe, Vent)= 0.049

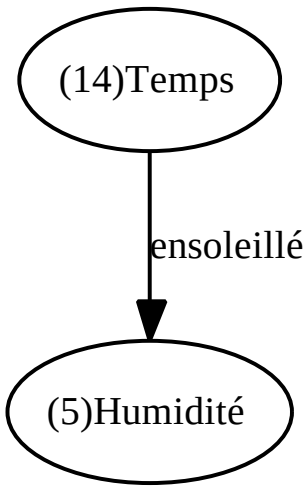
- Sélection de l'attribut pertinent qui maximise le gain d'information



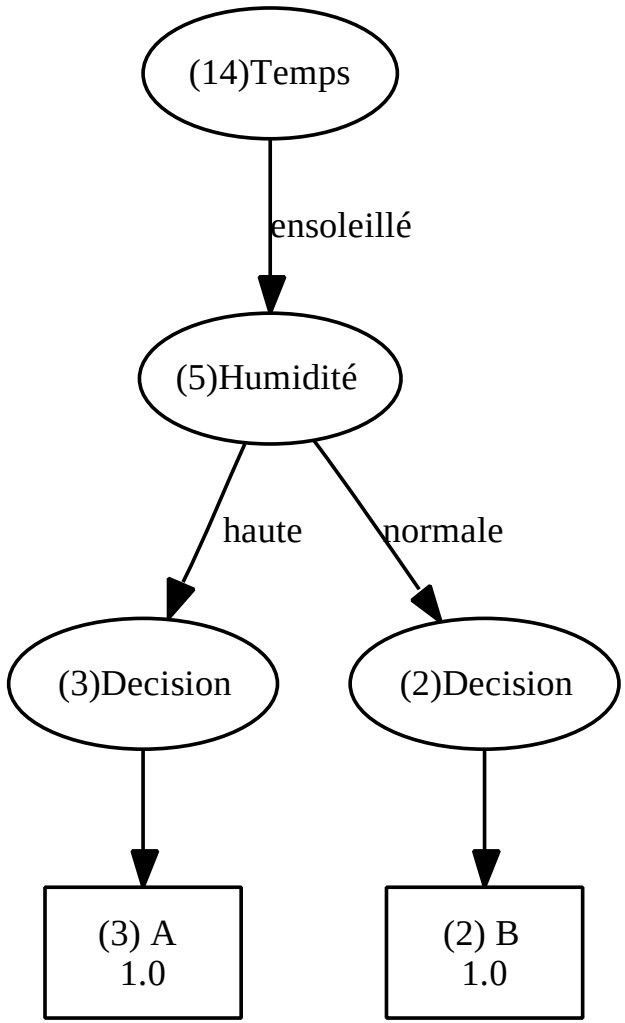
- Partition de l'ensemble d'apprentissage en 3 sous-ensembles selon les valeurs de l'attribut *Temps*

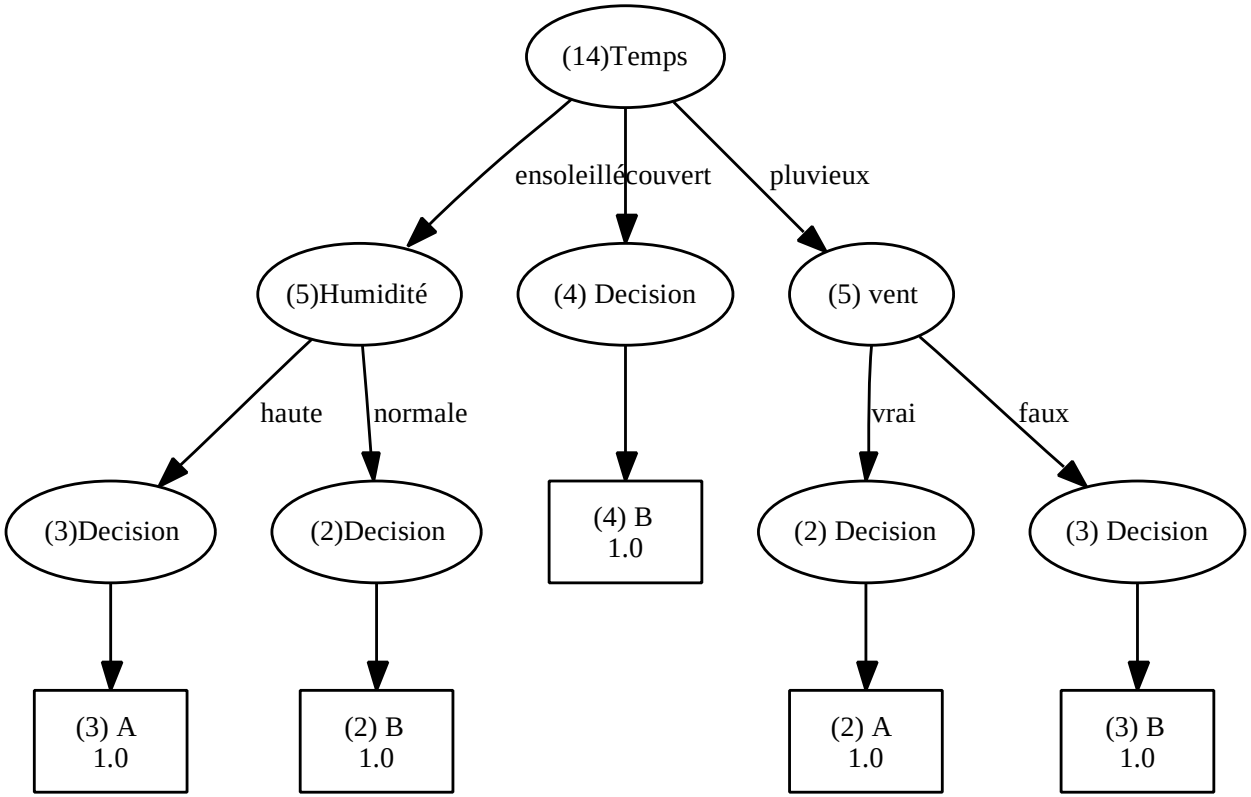
id	Temps	Température	Humidité	Vent	Classe
1	Ensoleillé	Elevée	Haute	Faux	A
2	Ensoleillé	Elevée	Haute	Vrai	A
8	Ensoleillé	Moyenne	Haute	Faux	A
9	Ensoleillé	Basse	Normale	Faux	B
11	Ensoleillé	Moyenne	Normale	Vrai	B

- Répétition du processus pour chaque sous-ensemble



- Création d'une feuille avec la valeur de la classe la plus probable si un critère d'arrêt est vérifié




(14)Temps
ensoleillé
couvert
pluvieux
(5)Humidité
(4) Decision
(5) vent
haute
normale
vrai
faux
(3)Decision
(2)Decision
(4) B
1.0
(2) Decision
(3) Decision
(3) A
1.0
(2) B
1.0
(2) A
1.0
(3) B
1.0

Plan

Valeurs manquantes

Pendant la phase de **construction**

- Calcul du gain d'information pour choisir l'attribut test
- Partition de l'ensemble d'apprentissage selon l'attribut test choisi

Valeurs manquantes

Pendant la phase de **construction**

- Calcul du gain d'information pour choisir l'attribut test
- Partition de l'ensemble d'apprentissage selon l'attribut test choisi

Pendant la phase de **classement**

Si la valeur d'un attribut test est manquante, il est impossible de décider quelle branche on doit choisir pour classer l'objet

Plan

Arbres d'Attributs Ordonnés (AAO) : Lobo et Numao 2000

Phase de construction : principe

- Ordonner les attributs par ordre croissant en fonction de leur Information Mutuelle avec la classe
- Construire un arbre de décision (appelé arbre d'attribut) pour chaque attribut

Arbres d'Attributs Ordonnés (AAO) : Lobo et Numao 2000

Phase de construction : principe

- Ordonner les attributs par ordre croissant en fonction de leur Information Mutuelle avec la classe
- Construire un arbre de décision (appelé arbre d'attribut) pour chaque attribut
  - Sous-ensemble : instances (valeurs connues pour cet attribut)
  - Attributs utilisés déjà traités

Arbres des Attributs Ordonnés (AAO)

Extrait de la base météo

id	Temps	Température	Humidité	Vent	Classe
11	Ensoleillé	Elevée	Haute	Faux	A
2	Ensoleillé	Elevée	Haute	Vrai	A
3	Couvert	Elevée	Haute	Faux	B
4	Pluvieux	?	Haute	Faux	B
5	Pluvieux	Basse	Normale	Faux	B
6	Pluvieux	Basse	Normale	Vrai	A
7	Couvert	Basse	Normale	Vrai	B
8	Ensoleillé	Moyenne	Haute	Faux	A
9	Ensoleillé	Basse	Normale	Faux	B
10	Pluvieux	Moyenne	Normale	Faux	B
11	Ensoleillé	Moyenne	Normale	Vrai	B
...	...	...	...	...	...

Arbres des Attributs Ordonnés (AAO)

IM(Classe, Température) =0.030
IM(Classe, Vent)= 0.049
IM(Classe, Humidité) = 0.154
IM(Classe, Temps) = 0.246
Température, *Vent*, *Humidité*, *Temps*

Arbres des Attributs Ordonnés (AAO)

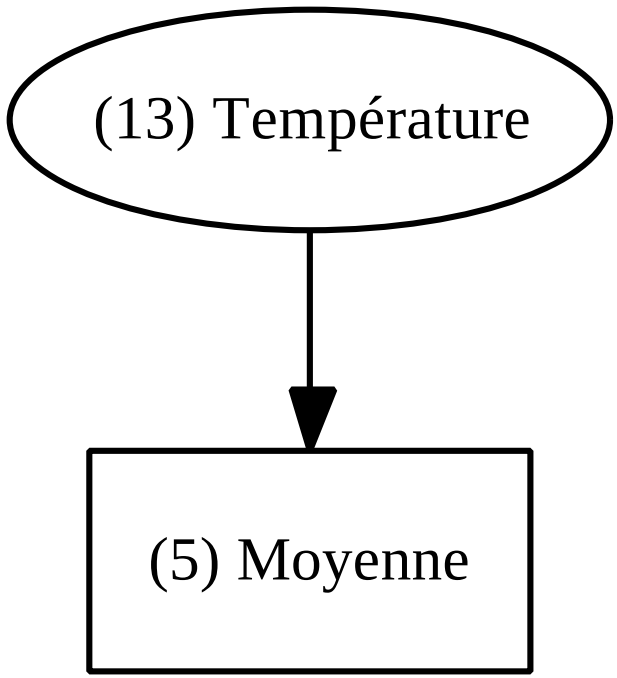
IM(Classe, Température) =0.030
IM(Classe, Vent)= 0.049
IM(Classe, Humidité) = 0.154
IM(Classe, Temps) = 0.246
Température, *Vent*, *Humidité*, *Temps*

id	Température
1	Elevée
2	Elevée
3	Elevée
5	Basse
6	Basse
8	Moyenne
10	Moyenne
11	Moyenne
12	Moyenne
..	..

Arbres des Attributs Ordonnés (AAO)

IM(Classe, Température) =0.030
IM(Classe, Vent)= 0.049
IM(Classe, Humidité) = 0.154
IM(Classe, Temps) = 0.246
Température, *Vent*, *Humidité*, *Temps*

id	Température
1	Elevée
2	Elevée
3	Elevée
5	Basse
6	Basse
8	Moyenne
10	Moyenne
11	Moyenne
12	Moyenne
..	..



Arbres d'Attributs Ordonnés (AAO)

Température, *Vent*, *Humidité*, *Temps*

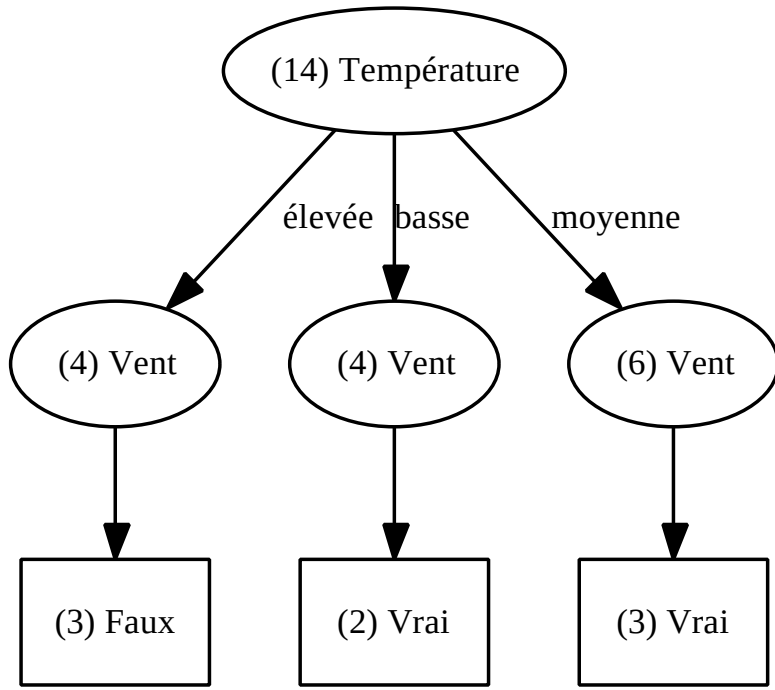
id	Température	Vent
1	Elevée	Faux
2	Elevée	Vrai
3	Elevée	Faux
4	**Moyenne**	Faux
5	Basse	Faux
6	Basse	Vrai
7	Basse	Vrai
..	..	..

Arbres d'Attributs Ordonnés (AAO)

Température, *Vent*, *Humidité*, *Temps*

id	Température	Vent
1	Elevée	Faux
2	Elevée	Vrai
3	Elevée	Faux
4	**Moyenne**	Faux
5	Basse	Faux
6	Basse	Vrai
7	Basse	Vrai
..	..	..

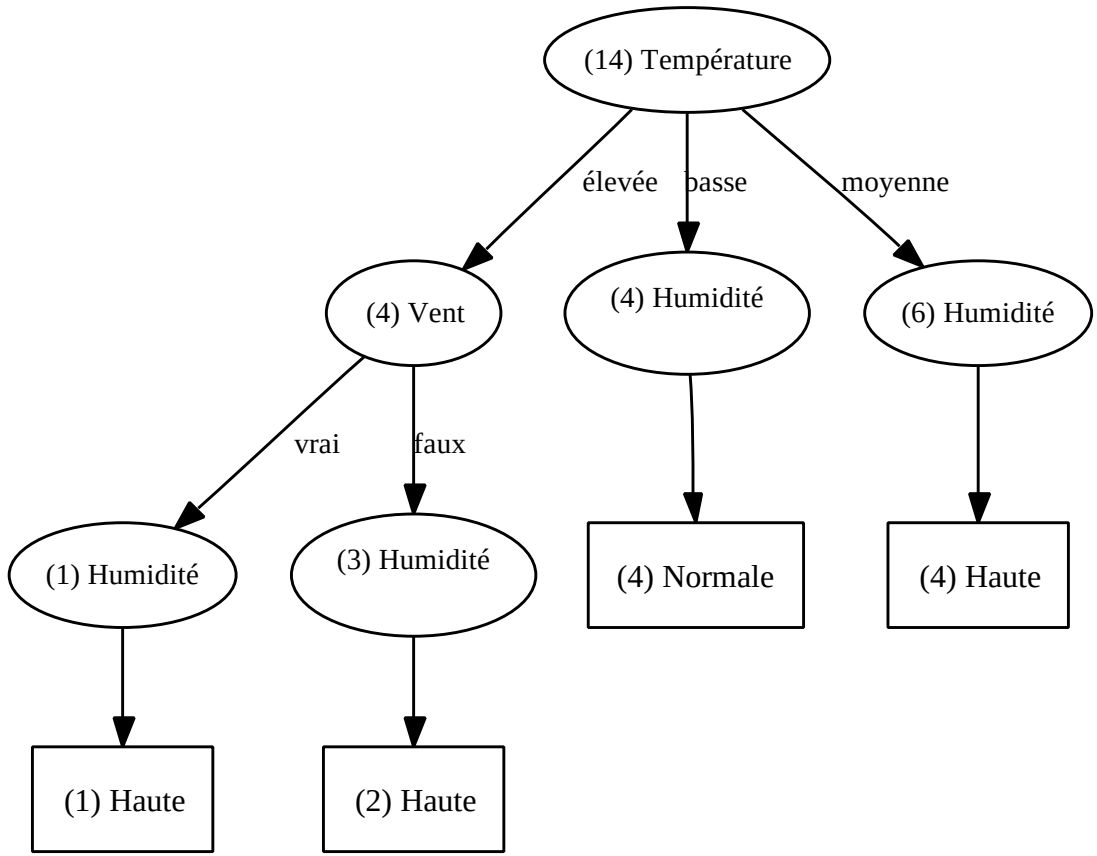
L'AAO de Vent



Arbres d'Attributs Ordonnés (AAO)

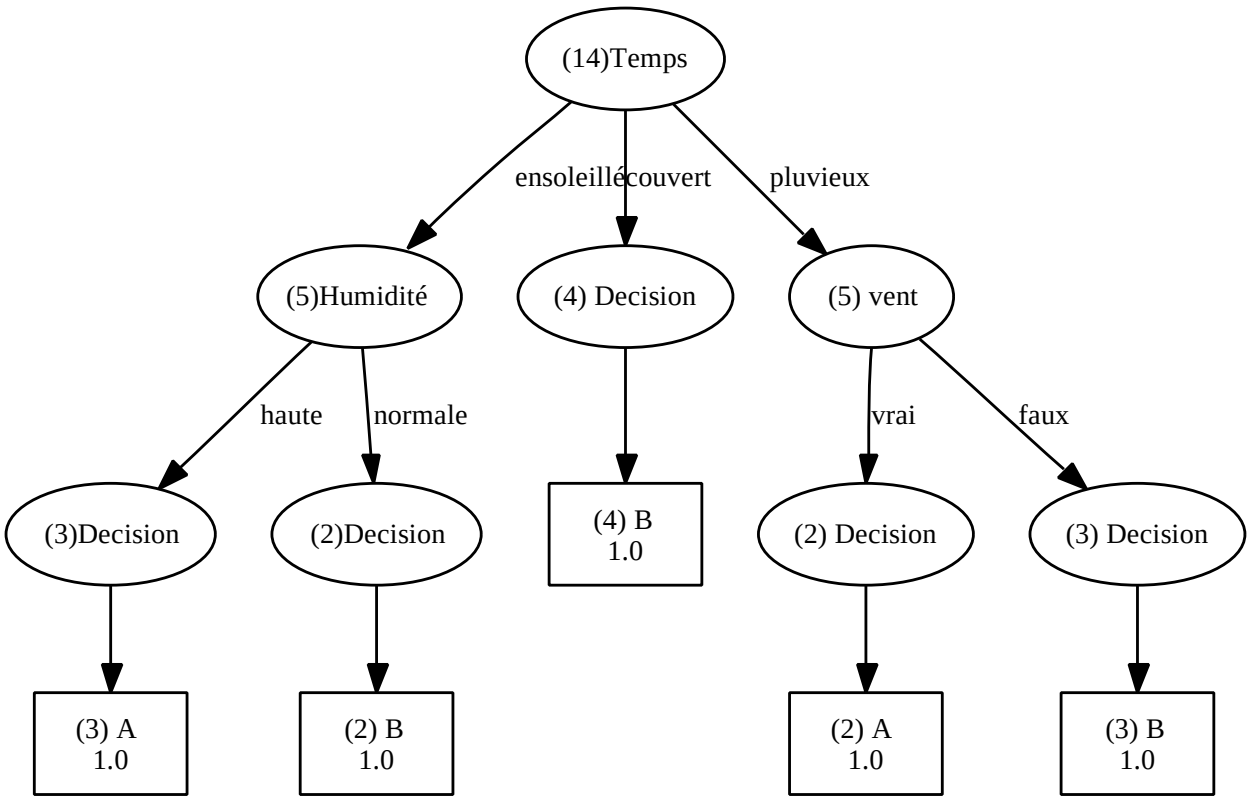
Température, *Vent*, *Humidité*, *Temps*

Construit en utilisant *Vent* et *Température*



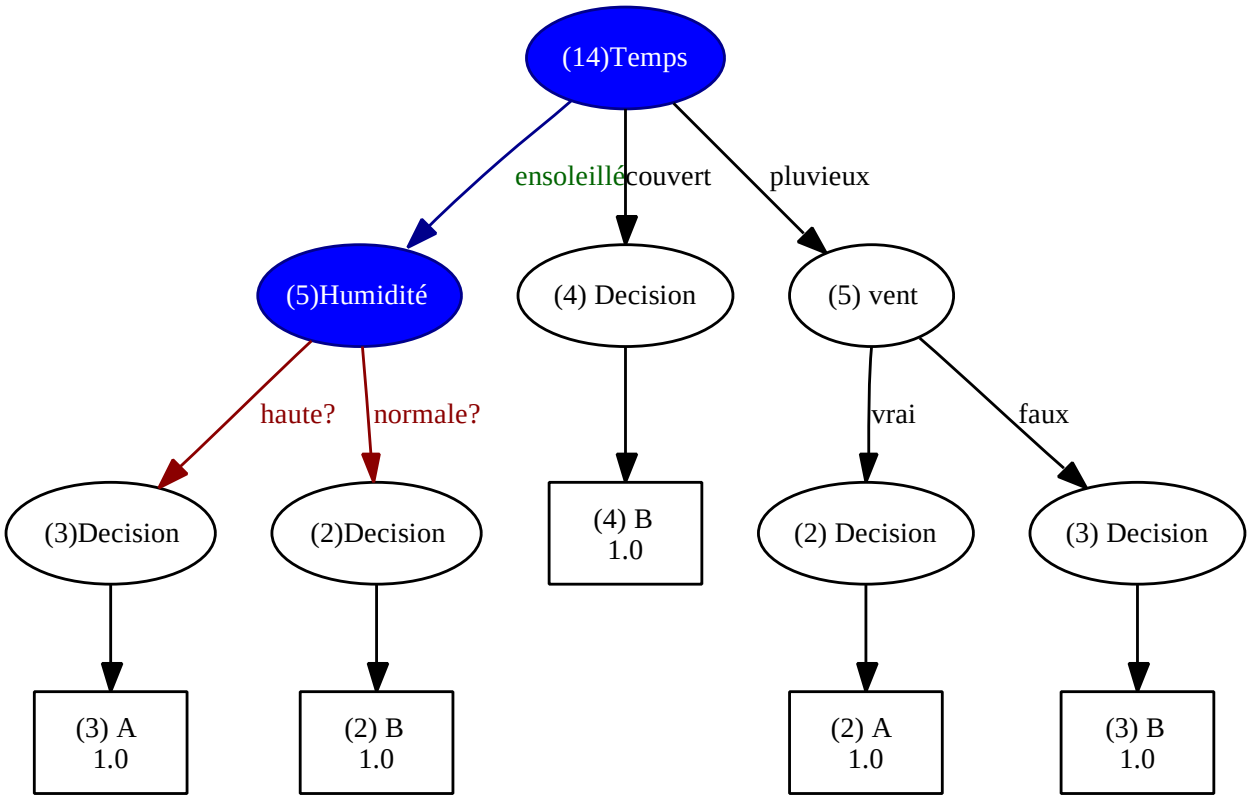
AAO : Exemple de classement

Temps	Température	Humidité	Vent	Classe
Ensoleillé	basse	?	Faux	?



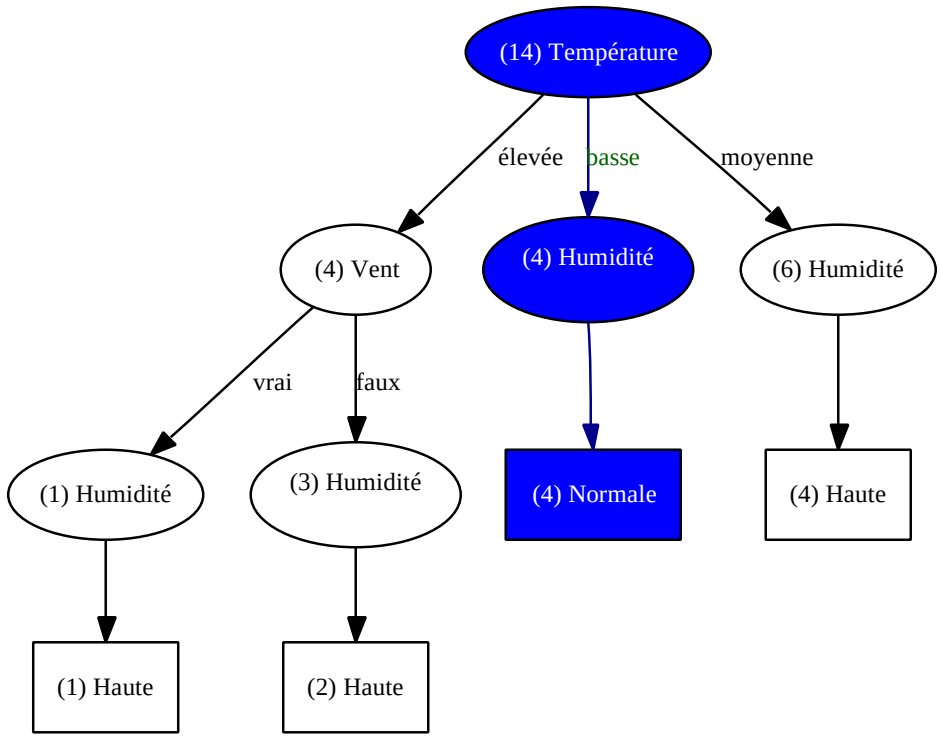
AAO : Exemple de classement

Temps	Température	Humidité	Vent	Classe
Ensoleillé	basse	?	Faux	?



AAO : Exemple de classement

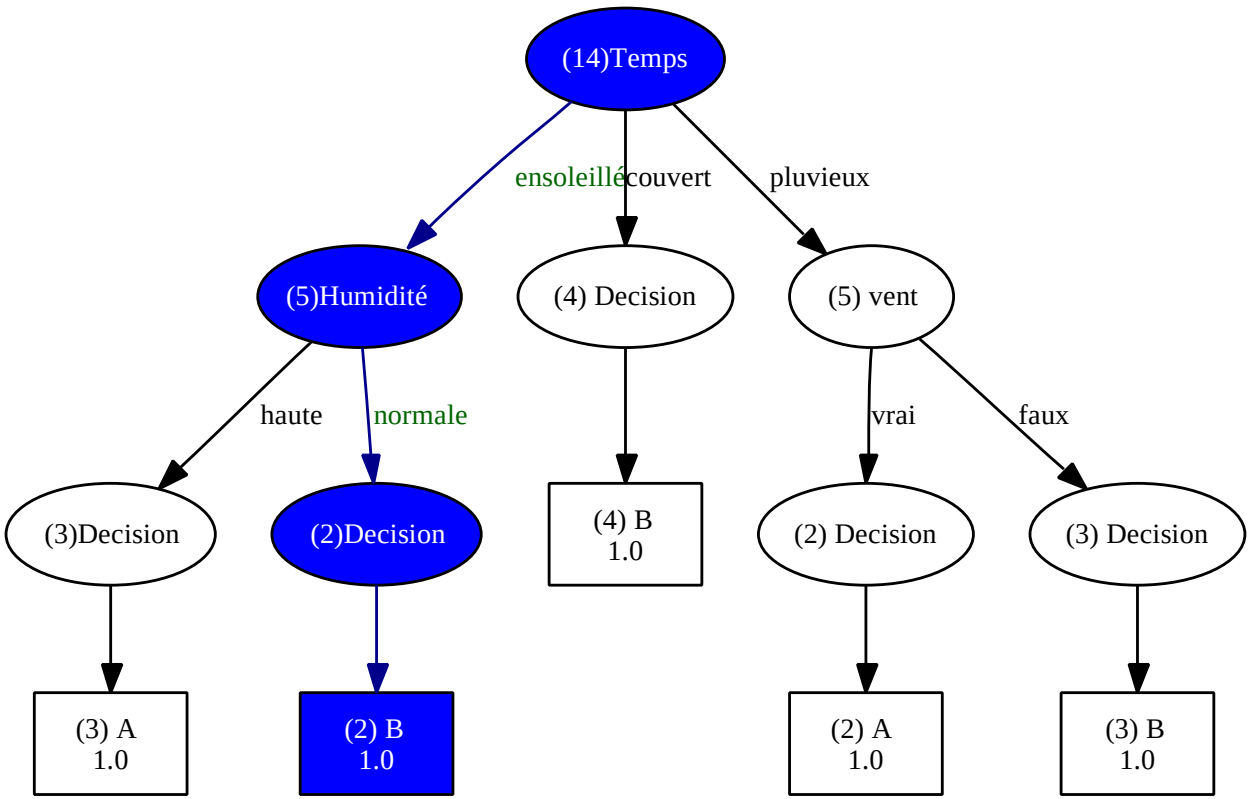
Temps	Température	Humidité	Vent	Classe
Ensoleillé	basse	?	Faux	?



Humidité est Normale

AAO : Exemple de classement

Temps	Température	Humidité	Vent	Classe
Ensoleillé	basse	normale	Faux	?



L'objet appartient à la classe B

AAO : Avantages et Inconvénients

Avantages

AAO : Avantages et Inconvénients

Avantages

- Utilise les arbres de décision pour déterminer la valeur manquante d'un attribut

AAO : Avantages et Inconvénients

Avantages

- Utilise les arbres de décision pour déterminer la valeur manquante d'un attribut
- Commence par l'attribut le moins dépendant de la classe

AAO : Avantages et Inconvénients

Avantages

- Utilise les arbres de décision pour déterminer la valeur manquante d'un attribut
- Commence par l'attribut le moins dépendant de la classe

Inconvénients

AAO : Avantages et Inconvénients

Avantages

- Utilise les arbres de décision pour déterminer la valeur manquante d'un attribut
- Commence par l'attribut le moins dépendant de la classe

Inconvénients

- Attribut manquant remplacé par une seule valeur

AAO : Avantages et Inconvénients

Avantages

- Utilise les arbres de décision pour déterminer la valeur manquante d'un attribut
- Commence par l'attribut le moins dépendant de la classe

Inconvénients

- Attribut manquant remplacé par une seule valeur
- Objet incomplet affecté à une seule classe

AAO : Avantages et Inconvénients

Avantages

- Utilise les arbres de décision pour déterminer la valeur manquante d'un attribut
- Commence par l'attribut le moins dépendant de la classe

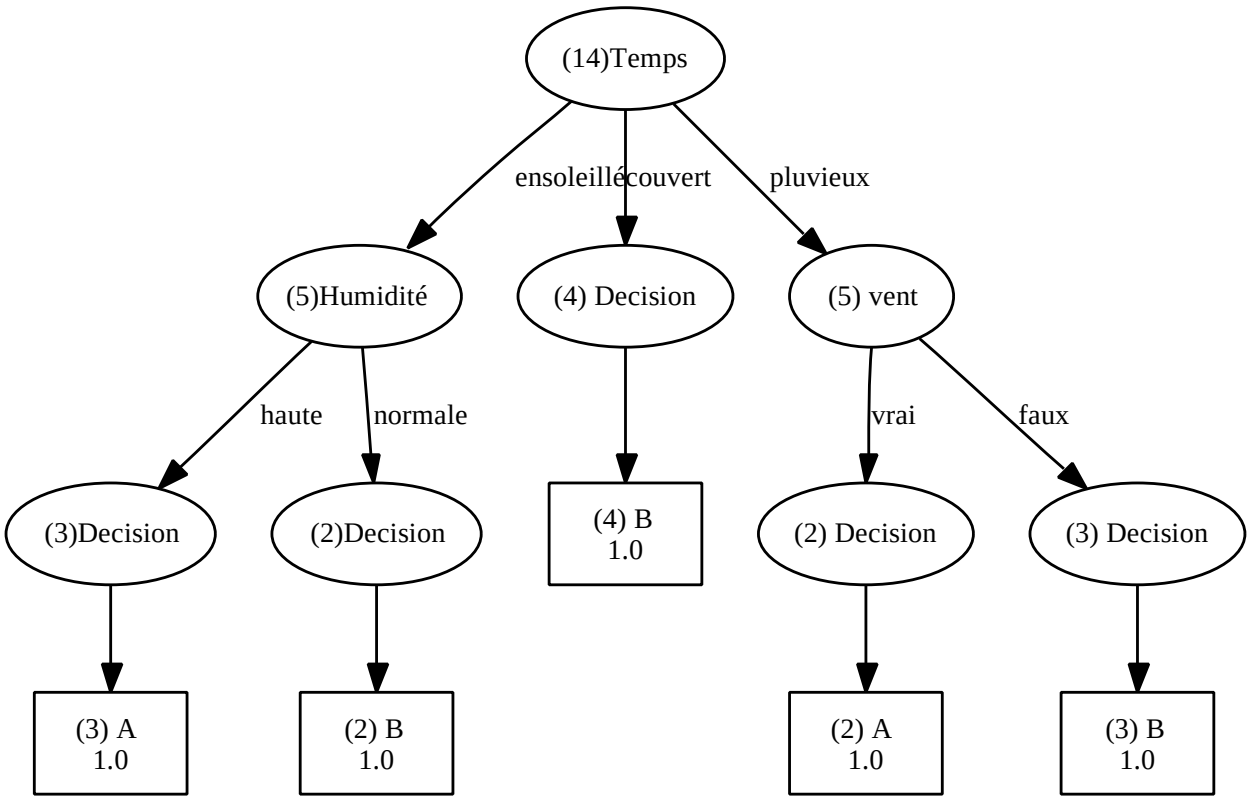
Inconvénients

- Attribut manquant remplacé par une seule valeur
- Objet incomplet affecté à une seule classe
- Ne prend pas en compte les dépendances entre les attributs

Plan

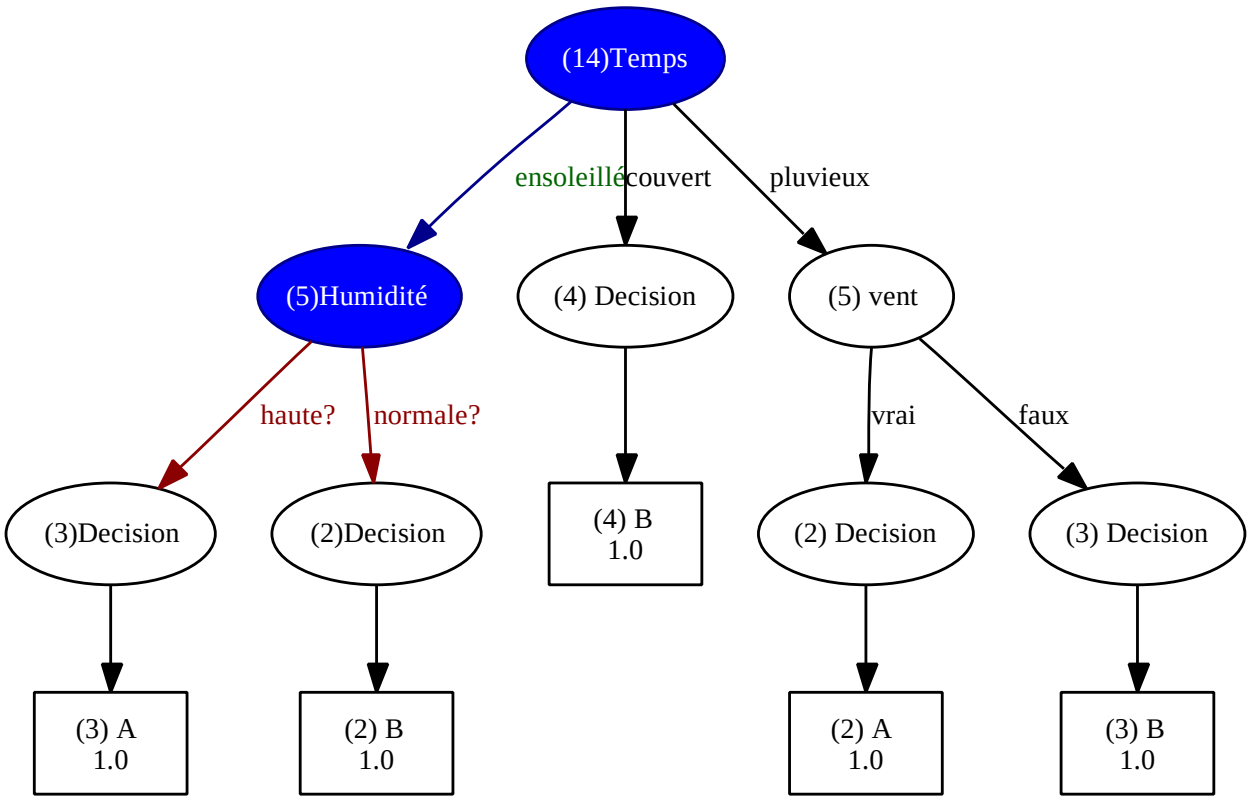
C4.5 : Phase de classement

Temps	Température	Humidité	Vent	Classe
Ensoleillé	basse	?	Faux	?



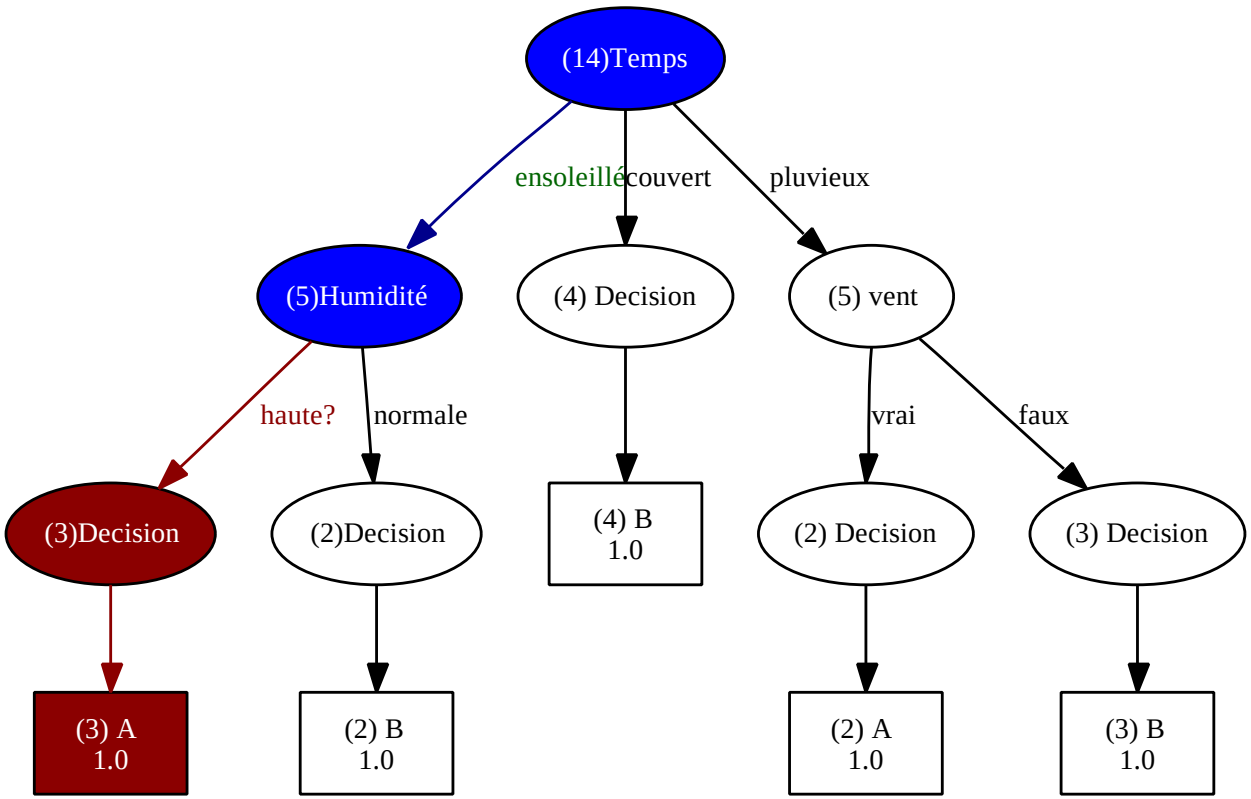
C4.5 : Phase de classement

Temps	Température	Humidité	Vent	Classe
Ensoleillé	basse	?	Faux	?



C4.5 : Phase de classement

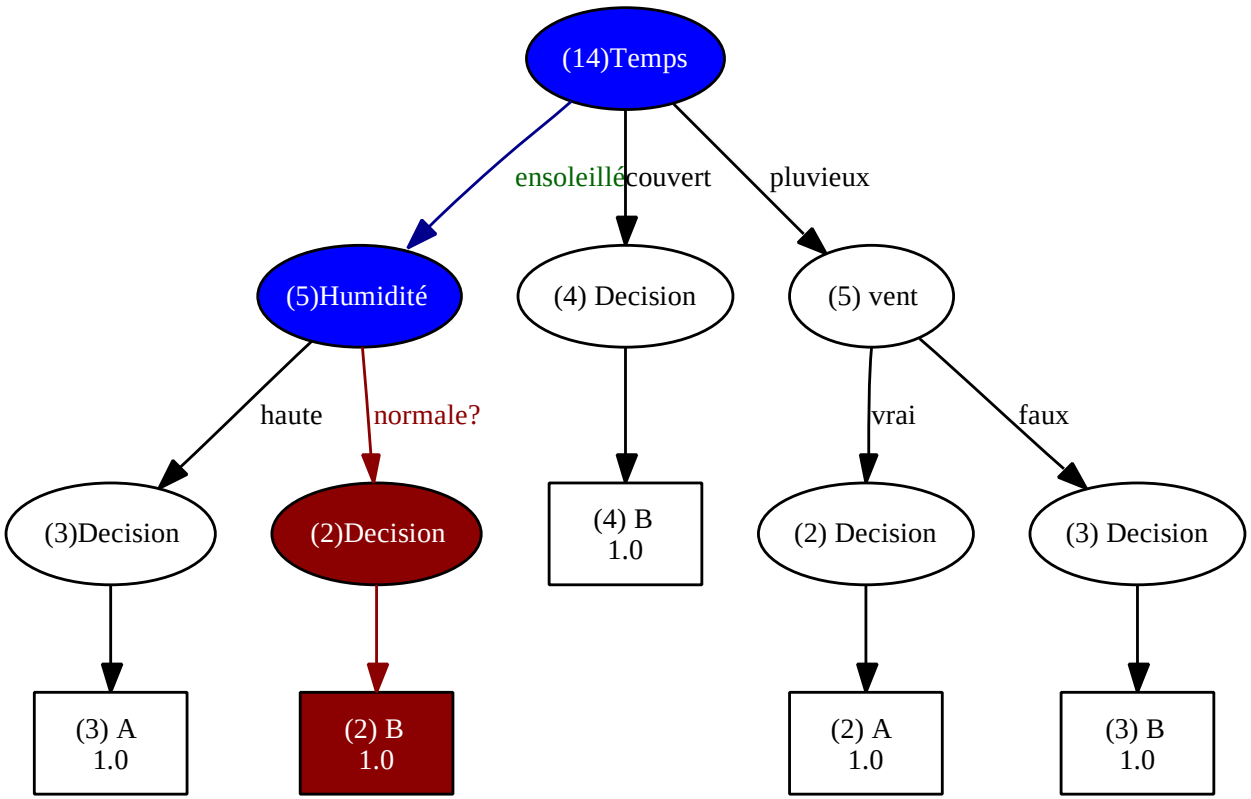
Temps	Température	Humidité	Vent	Classe
Ensoleillé	basse	?	Faux	?



$$P(A) = P(A|\text{haute}) \times P(\text{haute}) = 1 \times \left(\frac{3}{5}\right) = 0.6$$

C4.5 : Phase de classement

Temps	Température	Humidité	Vent	Classe
Ensoleillé	basse	?	Faux	?



$$P(B) = P(B|\text{normale}) \times P(\text{normale}) = 1 \times \left(\frac{2}{5}\right) = 0.4$$

C4.5 : Avantages et Inconvénients

C4.5 : Avantages et Inconvénients

Avantages

- Simple à utiliser
- Distribution de probabilité de classe

C4.5 : Avantages et Inconvénients

Avantages

- Simple à utiliser
- Distribution de probabilité de classe

Inconvénient

- Ne prend pas en compte les dépendances éventuelles entre les attributs

Plan

Approche proposée : Objectif

Classement probabiliste d'objets incomplètement connus dans un arbre de décision

Approche proposée : Objectif

Classement probabiliste d'objets incomplètement connus dans un arbre de décision

- Valeur manquante d'un attribut est prédite sous forme d'une **distribution de probabilité**

Approche proposée : Objectif

Classement probabiliste d'objets incomplètement connus dans un arbre de décision

- Valeur manquante d'un attribut est prédite sous forme d'une **distribution de probabilité**
- Utilisation des **dépendances** entre les attributs

Approche proposée : Objectif

Classement probabiliste d'objets incomplètement connus dans un arbre de décision

- Valeur manquante d'un attribut est prédite sous forme d'une **distribution de probabilité**
- Utilisation des **dépendances** entre les attributs
- Résultat du classement est **probabiliste**

Approche proposée : Objectif

Classement probabiliste d'objets incomplètement connus dans un arbre de décision

- Valeur manquante d'un attribut est prédite sous forme d'une **distribution de probabilité**
- Utilisation des **dépendances** entre les attributs
- Résultat du classement est **probabiliste**

Arbres de Décision Probabiliste

Utilisation d'un arbre de décision probabiliste au lieu d'un arbre de décision classique en gardant sur chaque feuille la distribution de probabilités de classe au lieu de la classe la plus probable

Plan

AAOP : Arbres d'Attributs Ordonnés Probabilistes

AAOP

- **Extension** de la méthode des *AAOs* (Lobo et Numao)
- *Attributs par ordre croissant selon l'Information Mutuelle relativement à la classe*

AAOP : Arbres d'Attributs Ordonnés Probabilistes

AAOP

- **Extension** de la méthode des *AAOs* (Lobo et Numao)
- *Attributs par ordre croissant selon l'Information Mutuelle relativement à la classe*
- Pour chaque attribut, construire un *AAO***P** au lieu d'un *AAO*

AAOP : Arbres d'Attributs Ordonnés Probabilistes

AAOP

- **Extension** de la méthode des *AAOs* (Lobo et Numao)
- *Attributs par ordre croissant selon l'Information Mutuelle relativement à la classe*
- Pour chaque attribut, construire un *AAO***P** au lieu d'un *AAO*
- Utilisation des attributs *déjà traités* et **dépendants** de l'attribut courant

AAOP : Arbres d'Attributs Ordonnés Probabilistes

AAOP

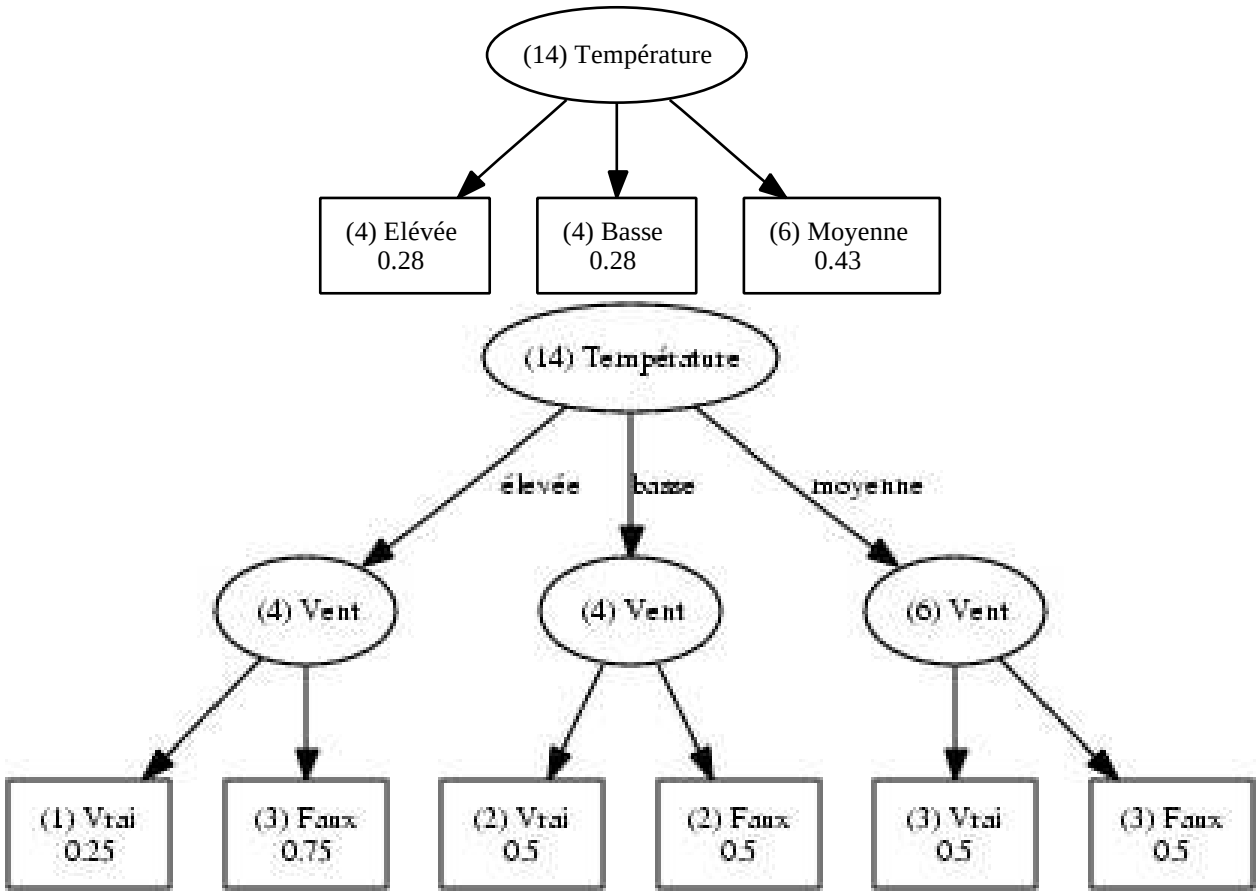
- **Extension** de la méthode des *AAOs* (Lobo et Numao)
- *Attributs par ordre croissant selon l'Information Mutuelle relativement à la classe*
- Pour chaque attribut, construire un *AAO***P** au lieu d'un *AAO*
- Utilisation des attributs *déjà traités* et **dépendants** de l'attribut courant

Différence par rapport à AAO (Lobo et Numao)

+ Feuilles probabilistes
+ Suppression des attributs déjà traités et **indépendants** de l'attribut courant

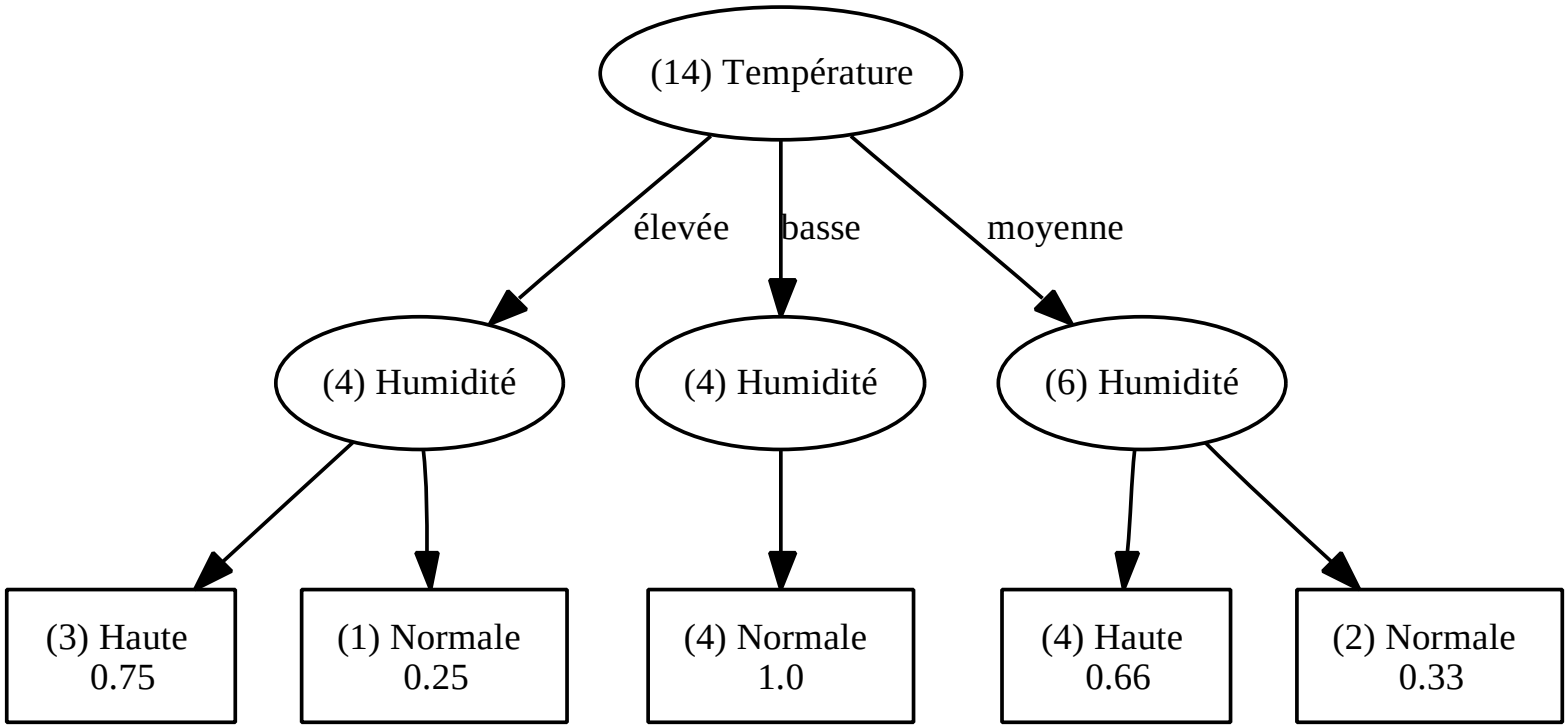
Exemple : AAOP

L'AAOP de Température et l'AAOP de Vent



Exemple : AAOP

L'AAOP d'Humidité



N'utilise pas l'attribut *Vent*

Plan

AAP : Arbres d'Attributs Probabilistes

La première proposition (AAOP) ne prend pas en compte toutes les dépendances

AAP : Arbres d'Attributs Probabilistes

La première proposition (AAOP) ne prend pas en compte toutes les dépendances

AAP

- Calcul pour chaque attribut de ses attributs dépendants en utilisant l'Information Mutuelle :

$$Dep(A_i) = \{A_j \mid IM_N(A_i, A_j) > Seuil\ \}$$

Seuil $\geq$ l'Information Mutuelle Normalisée moyenne [Lobo et Numao 2000]

AAP : Arbres d'Attributs Probabilistes

La première proposition (AAOP) ne prend pas en compte toutes les dépendances

AAP

- Calcul pour chaque attribut de ses attributs dépendants en utilisant l'Information Mutuelle :

$$Dep(A_i) = \{A_j \mid IM_N(A_i, A_j) > Seuil\ \}$$

Seuil $\geq$ l'Information Mutuelle Normalisée moyenne [Lobo et Numao 2000]

- Construction d'un AAP pour chaque attribut en utilisant les attributs dont il dépend

AAP : Arbres d'Attributs Probabilistes

La première proposition (AAOP) ne prend pas en compte toutes les dépendances

AAP

- Calcul pour chaque attribut de ses attributs dépendants en utilisant l'Information Mutuelle :

$$Dep(A_i) = \{A_j \mid IM_N(A_i, A_j) > Seuil\ \}$$

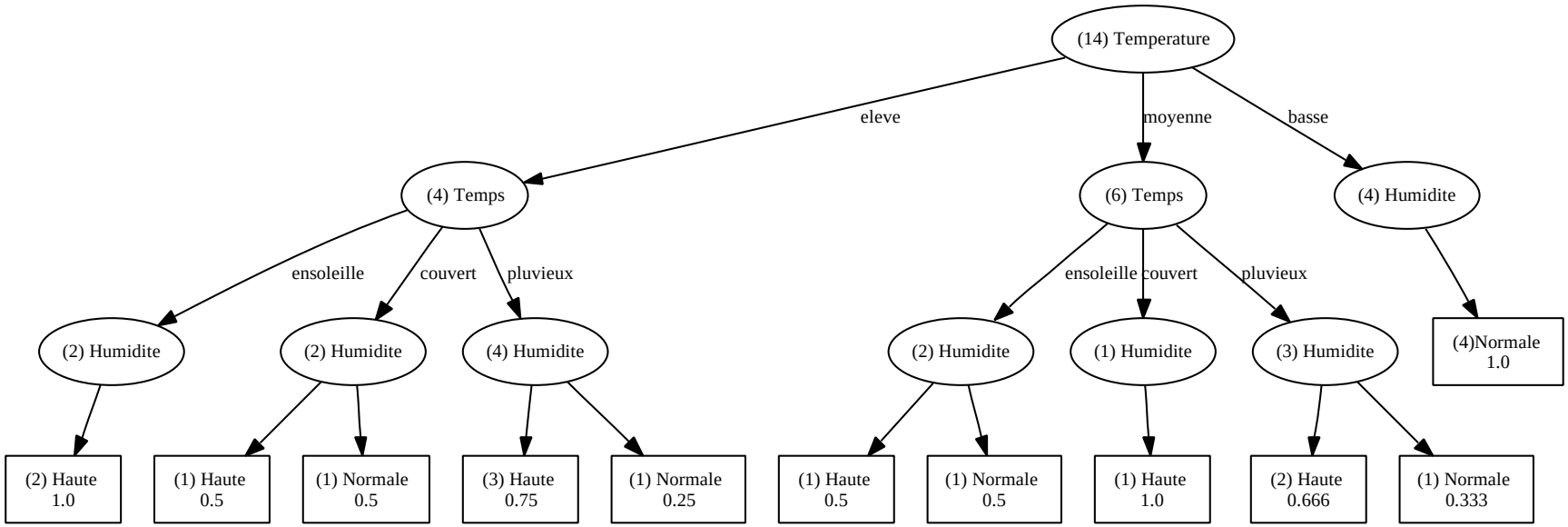
Seuil $\geq$ l'Information Mutuelle Normalisée moyenne [Lobo et Numao 2000]

- Construction d'un *AAP* pour chaque attribut en utilisant les attributs dont il dépend

Différence par rapport à AAOP

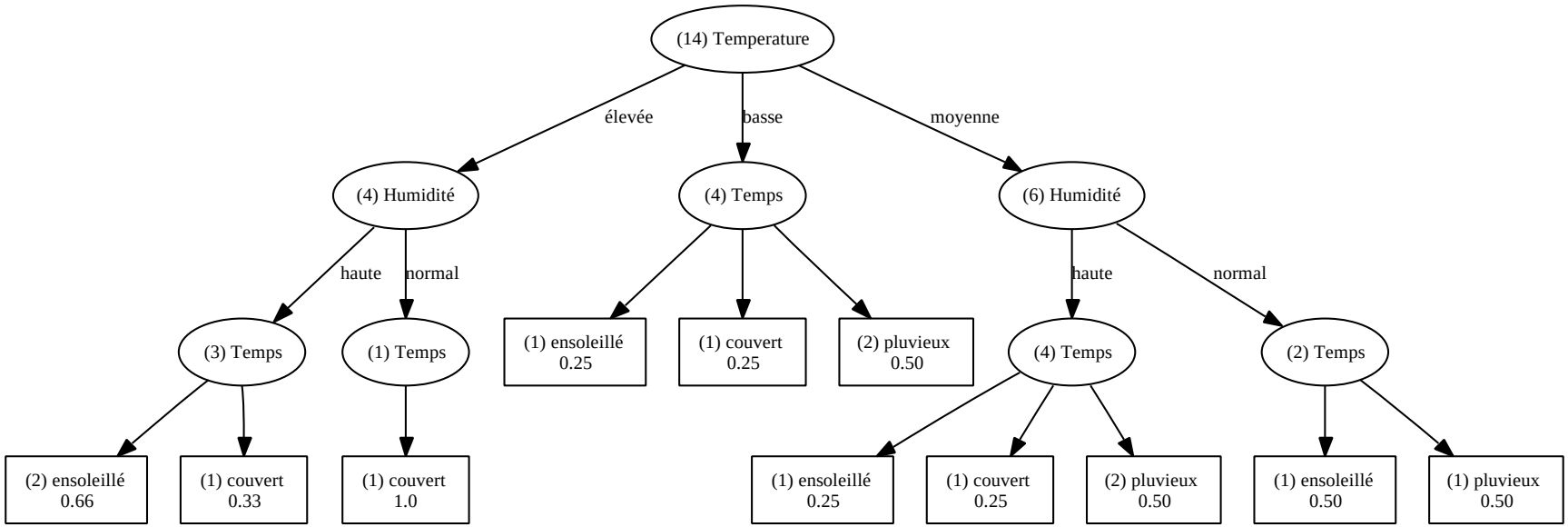
+ Pas d'ordre de construction
+ Arbre d'attribut construit en utilisant ses attributs dépendants

Exemple : AAP



L'arbre d'Humidité construit avec Temps et Température

Exemple : AAP



L'arbre de Temps construit avec Température et Humidité

Problème avec AAP

Cycle

Deux attributs dépendants et manquants en même temps

Problème avec AAP

Cycle

Deux attributs dépendants et manquants en même temps

Solution

- Attribut le moins dépendant de la classe : appel de son arbre d'attribut construit selon *AAOP*
- Autre attribut : appel de son arbre d'attribut construit selon *AAP*

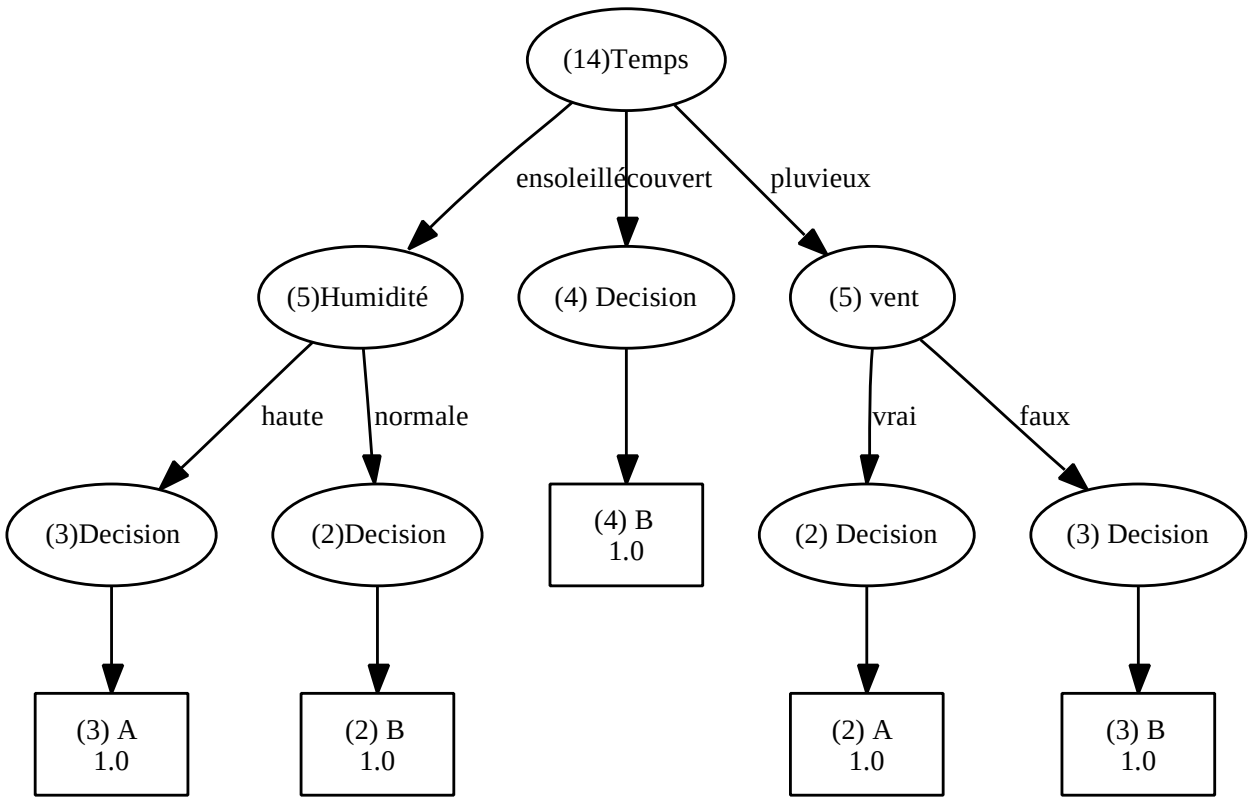
Plan

Exemple de classement

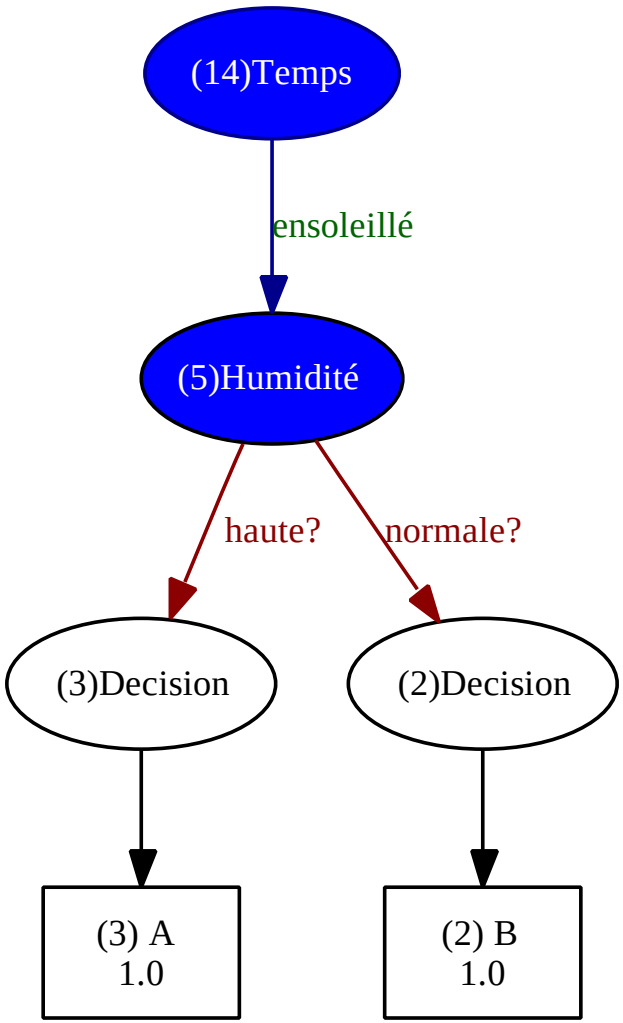
Temps	Température	Humidité	Vent	Classe
Ensoleillé	?	?	Faux	?

Exemple de classement

Temps	Température	Humidité	Vent	Classe
Ensoleillé	?	?	Faux	?



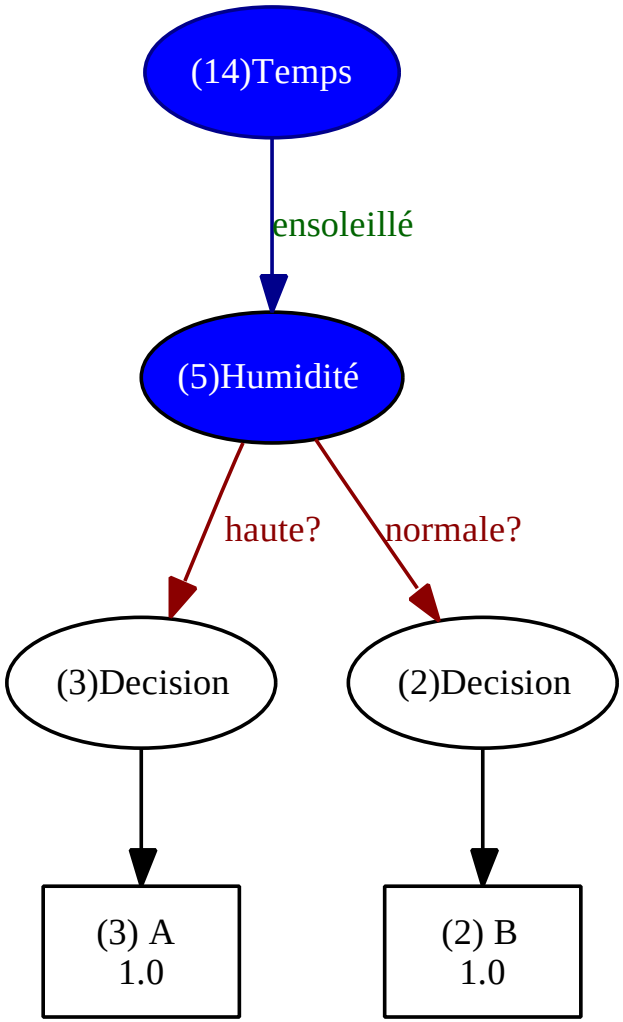
Exemple de classement

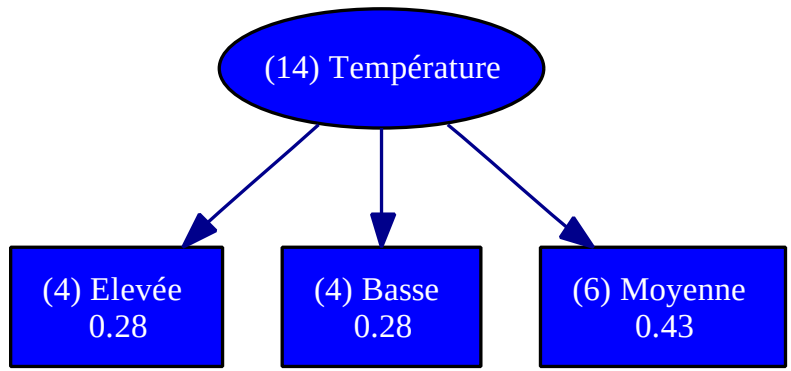


$P(A) = P(A|\text{haute})\ P(\text{haute})$
$P(B) = P(B|\text{normale})\ P(\text{normale})$
$P(A|\text{haute}) = 1 \Longrightarrow$
$P(A) = P(\text{haute})$
$P(B|\text{normale}) = 1 \Longrightarrow$
$P(B) = P(\text{normale})$

Humidité et Température :
dépendants
AAP d'Humidité
AAOP de Température

Exemple de classement

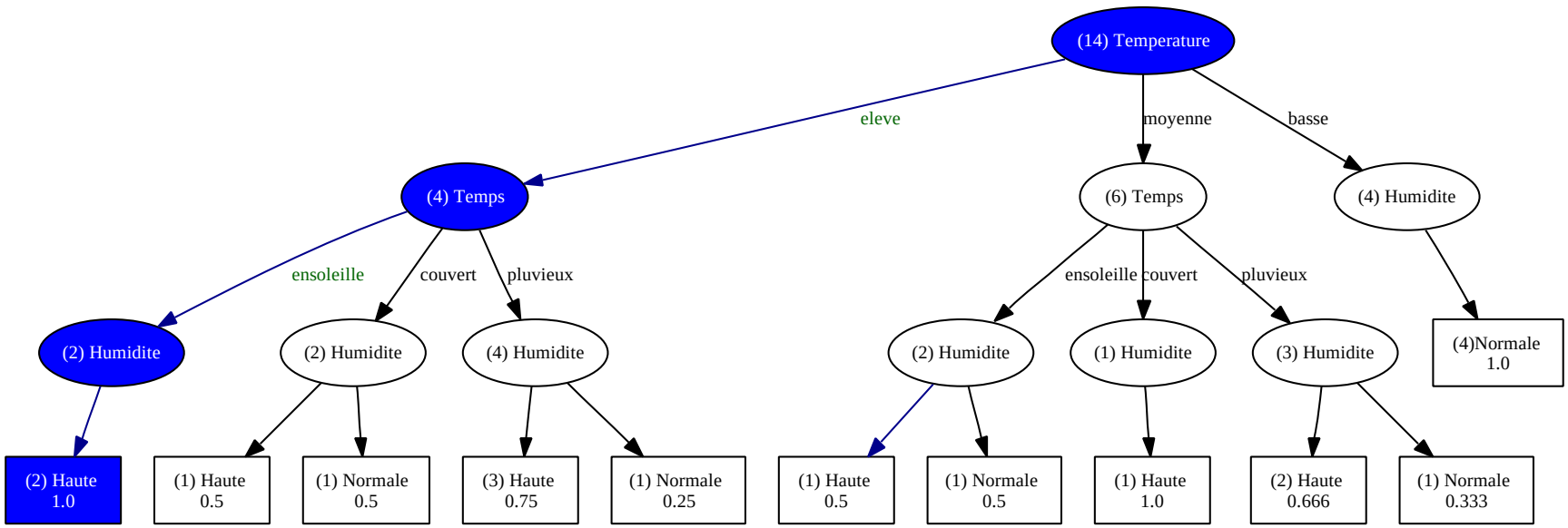




P(élevée) =0.28
P(moyenne)=0.43
P(basse)=0.28

Exemple de classement

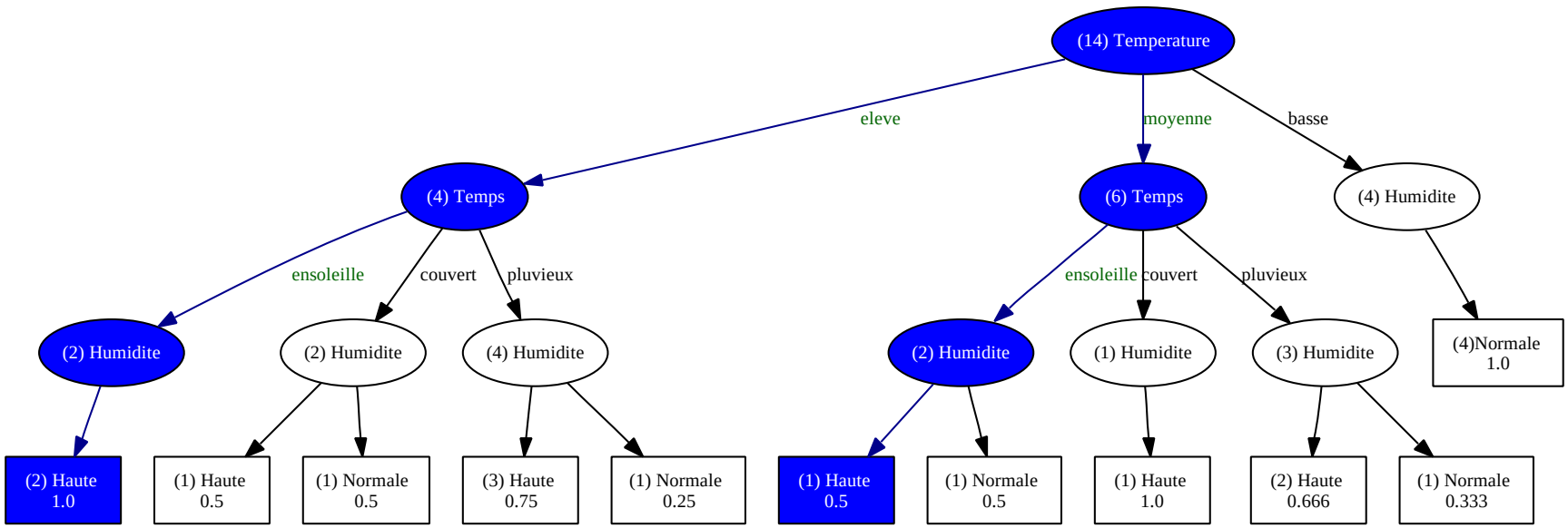
Temps	Température	Humidité	Vent	Classe
Ensoleillé	?	?	Faux	?



P(haute)= P(haute| ensoleillé, élevée) $\times$P(ensoleillé, élevée)

Exemple de classement

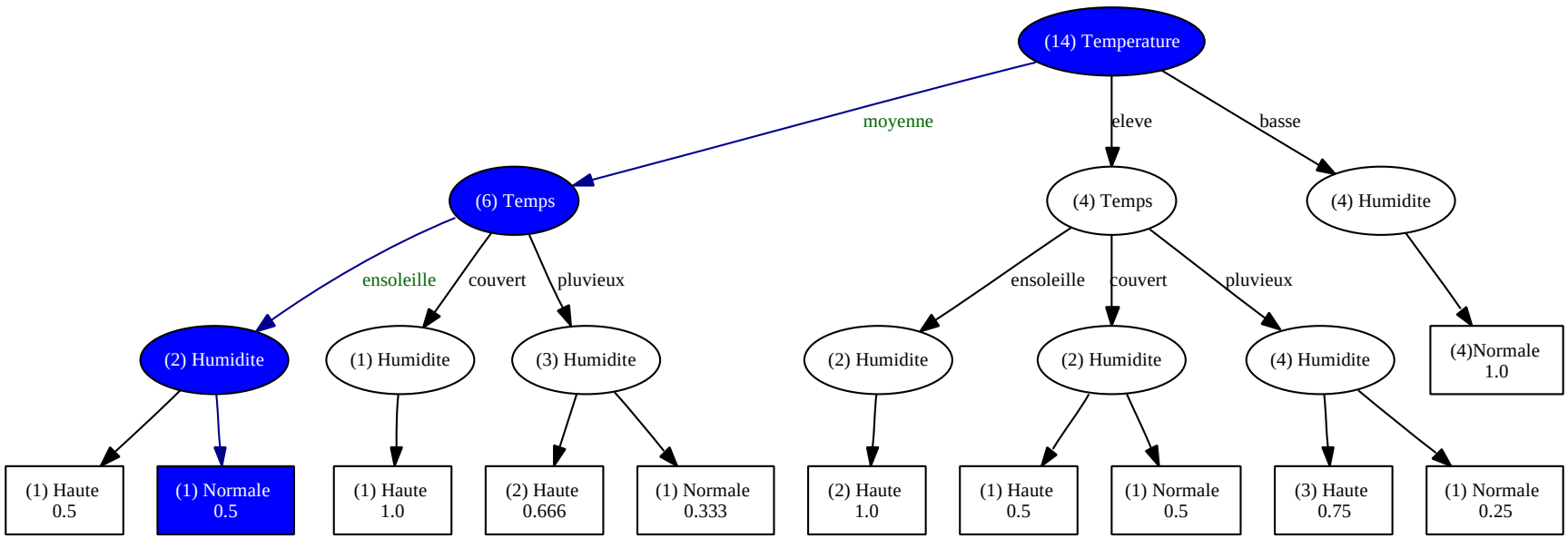
Temps	Température	Humidité	Vent	Classe
Ensoleillé	?	?	Faux	?



$$P(\mathit{haute}) = \texttt{P(haute| ensoleillé, élevée)} \times \texttt{P(ensoleillé, élevée)}$$
$$+ \texttt{P(haute| ensoleillé, moyenne)} \times \texttt{P(ensoleillé, moyenne)}$$
$$= 1 \times 0.28 + 0.5 \times 0.43 = 0.495$$

Exemple de classement

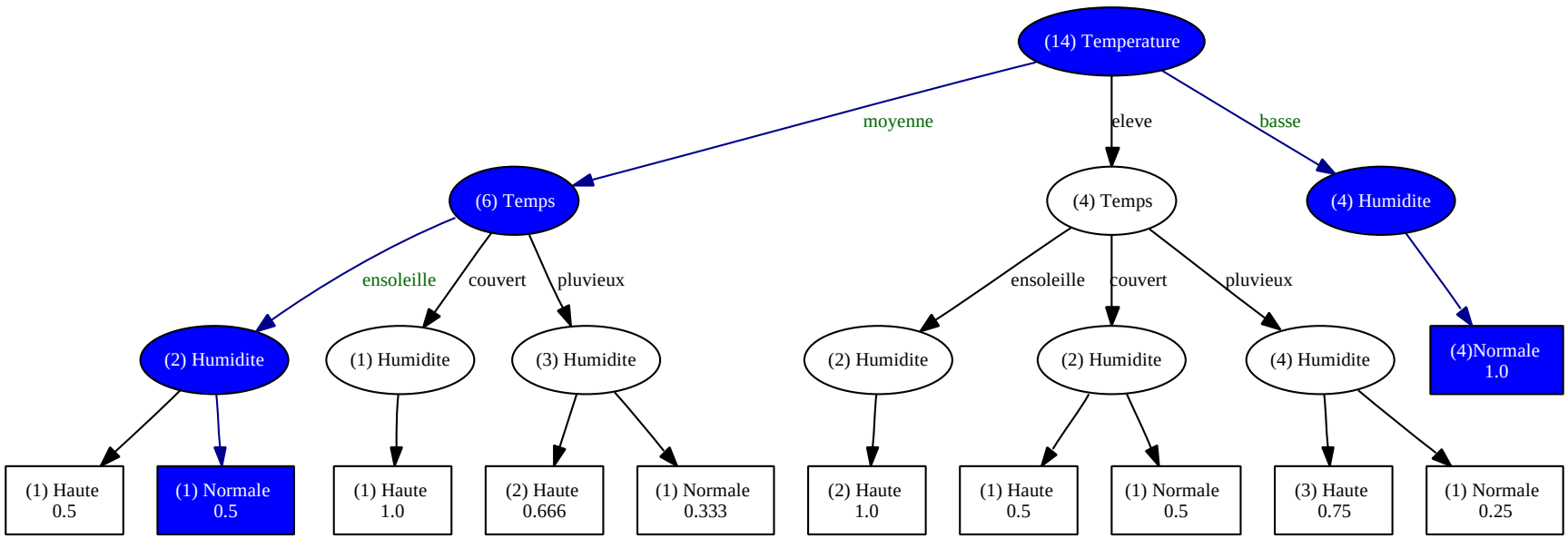
Temps	Température	Humidité	Vent	Classe
Ensoleillé	?	?	Faux	?



$$P(\mathit{normale}) = P(\mathit{normale}|\texttt{ensoleillé}, \mathit{moyenne}) \times P(\texttt{ensoleillé}, \mathit{moyenne})$$

Exemple de classement

Temps	Température	Humidité	Vent	Classe
Ensoleillé	?	?	Faux	?



$$
\begin{aligned}
P(\mathit{normale}) &= P(\mathit{normale}|\texttt{ensoleillé}, \mathit{moyenne}) \times \texttt{P(ensoleillé}, \mathit{moyenne}) \\
&\quad + P(\mathit{normale}|\mathit{basse}) \times P(\mathit{basse}) \\
&= 0.5 \times 0.43 + 1 \times 0.28 = 0.495
\end{aligned}
$$

Avantages et Inconvénients

Avantages

- Par rapport à C4.5 : prendre en compte d'autres attributs qui ne sont pas forcément dans l'arbre
  Température lors du calcul de la probabilité de *Humidité*

Avantages et Inconvénients

Avantages

- Par rapport à C4.5 : prendre en compte d'autres attributs qui ne sont pas forcément dans l'arbre
  Température lors du calcul de la probabilité de *Humidité*
- Par rapport à *AAO* : utilise un arbre d'attribut construit en fonction de ses attributs dépendants
  Temps lors du calcul de la probabilité de *Humidité*

Avantages et Inconvénients

Avantages

- Par rapport à C4.5 : prendre en compte d'autres attributs qui ne sont pas forcément dans l'arbre
  Température lors du calcul de la probabilité de *Humidité*
- Par rapport à *AAO* : utilise un arbre d'attribut construit en fonction de ses attributs dépendants
  Temps lors du calcul de la probabilité de *Humidité*

Inconvénient

Complexité de classement exponentielle en fonction du nombre d'attributs manquants dans l'objet à classer

$$Complexity(Trees) = O(m \times \bar{v}^m log(\bar{L}_T))$$

Plan

Bases de test

Test de notre approche, C4.5 et *AAO* sur 11 bases

- *Attributs dépendants* : Vote, Breast-cancer, Zoo, Lymphography, Mushroom, Dermatology, Splice, Iris, Breast-w
- *Attributs indépendants* : Nursery, Car

Bases de test

Test de notre approche, C4.5 et *AAO* sur 11 bases

- *Attributs dépendants* : Vote, Breast-cancer, Zoo, Lymphography, Mushroom, Dermatology, Splice, Iris, Breast-w
- *Attributs indépendants* : Nursery, Car

- Base de test construite à partir de la base d'apprentissage
- Attributs pertinents inconnus
- Tests sur plusieurs seuils de dépendance
- Comparaison avec les résultats de classement donnés par C4.5 et *AAO* pour la même base
- Calcul de la matrice de confusion de chaque méthode

Bases de test

La base *mushroom*

- Champignons sont comestibles ou toxiques
- Taille base d'apprentissage : 5644 instances, 22 attributs discrets
- Taille base de test : 73 instances
- La racine : *odor*, 75.34% de valeurs manquantes

Mushroom	Seuil	bien classés	mal classés	50%
AAP	**0.1** 0.2	**80.82%** 75.34%	**19.17 %** 24.65 %	
C4.5		**58.90%**	**41.09%**	
AAO		67.12%	32.87%	

Plan

Taux d'erreurs

Root Mean Squared Error

Pour une instance x le *RMES* est donné par l'équation suivante :

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{j=n}(t(j|x) - P(j|x))^2}$$

x est l'instance, j est la valeur de classe
$t(j|x)$ est la vraie probabilité de la classe j pour x
$P(j|x)$ est la probabilité estimée par la méthode pour l'instance x et la classe j

Taux d'erreurs

DataBase	AAP	C4.5	AAO
Zoo	**0.133817**	0.44812	0.245
Mushroom	**0.412543**	0.643147	0.535865
Dermatology	**0.25183339**	0.4108025	0.3068860
Vote	**0.310443**	0.52039	0.315079
Breast-cancer	**0.525338**	0.632477	0.6238318
Lymphography	**0.260477**	0.477835	0.420603
Splice	0.35292	0.38507	**0.22929**
Iris	0.290535	0.534436	**0.226616**
Breast-w	**0.29610**	0.55673	0.32818
Nursery	0.4456728	0.44999	0.436149
Car	0.3292286	0.32743	0.35666

Plan

Comparaison de performances entre C4.5 et *AAP*

Test McNemar

Soient deux hypothèses de classement h_1 (AAP) et h_2 (C4.5) :

$$M = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{10} + n_{01}}$$

Ce test suit une loi χ^2 à 1 degré de liberté. L'hypothèse que h_1 et h_2 aient le même taux d'erreur peut être rejetée avec une probabilité supérieure à 95% si $|M| > 3.84$

où

- n11 : le nombre d'instances bien classées par $h_2 1$ et h_2 ;
- n01 : le nombre d'instances mal classées par h_1 et non par h_2 ;
- n10 : le nombre d'instances mal classées par h_2 et non par h_1 ;
- n00 : le nombre d'instances mal classées par h_2 et h_1.

Les résultats du test McNemar sur les bases de données

n_{01} : nombre d'instances mal classées par *AAP* et non par C4.5
n_{10} : nombre d'instances mal classées par C4.5 et non par *AAP*

Base	n_{01}	n_{10}	McNemar	Différence
Vote	12	32	8.205	significative
Breast-cancer	5	14	3,368	non significative
Lymphography	2	11	4.923	significative
Zoo	1	23	18.375	**très significative**
Mushroom	0	12	10.083	**très significative**
Dermatology	1	8	4.000	significative
Splice	2	5	0.571	non significative
Iris	1	19	14.450	**très significative**
Breast-w	3	19	10.227	**très significative**
Nursery	2	0	3.448	non significative
Car	1	5	1.500	non significative

Plan

Analyse des résultats de classement

Objectif

- Analyse du résultat de classement de chaque objet test
- Comparaison des distributions de probabilités données par *AAP* et C4.5
- Choix de la distribution la plus représentative par rapport à la base d'apprentissage

Algorithme Analyser-Instance

Pour un objet test o

- Calcul de la distance entre l'objet o et tous les objets de la base d'apprentissage
- Si (Distance $<$ Near), deux objets Plus Proches Voisins (PPV)
- Fréquence de ses Plus Proches Voisins de chaque classe

RELIEF

Relief : Kira et Rendell 1992

- Mesure d'impureté pour évaluer la qualité d'un attribut
- Basé entièrement sur l'analyse statistique
- Dépendances entre les attributs
- Valeurs manquantes
- Calcul de la distance entre les objets

RELIEF

Fonction de Distance

$$Distance(I_1, I_2) = \sum_{j=1}^{j=n} diff(A_j, I_1, I_2)$$

$$diff(A, I_1, I_2) = \begin{cases} 0 \text{ if } & V^{(A,I_1)} = V^{(A,I_2)} \\ 1 \text{ if } & V^{(A,I_1)} \neq V^{(A,I_2)} \\ 1 - P(V^{(A,I_2)}/Class_{I_1}) \text{ if } & \texttt{A is unknown in } I_1 \end{cases}$$

- $Class_{I_1}$ est la valeur de la classe dans l'instance I_1
- $1 - P(V^{(A,I_2)}/Class_{I_1})$ est la probabilité que deux instances I_1 et I_2 prennent des valeurs différentes pour l'attribut A dans le cas où une de ces instances (I_1 ici) possède une valeur inconnue pour A

Exemple de calcul de distance

Instance test

id	Temps	Température	Humidité	Vent	Classe
5	Ensoleillé	Moyenne	?	Faux	A

Base d'apprentissage

id	Temps	Température	Humidité	Vent	Classe
1	Ensoleillé	Moyenne	Haute	Faux	A
2	Pluvieux	Moyenne	Normale	Faux	B
3	Pluvieux	Moyenne	Haute	Vrai	A
4	Ensoleillé	Moyenne	Normale	faux	A

Exemple de calcul de distance

Instance test

id	Temps	Température	Humidité	Vent	Classe
5	Ensoleillé	Moyenne	?	Faux	A

Base d'apprentissage

id	Temps	Température	Humidité	Vent	Classe
1	Ensoleillé	Moyenne	Haute	Faux	A
2	Pluvieux	Moyenne	Normale	Faux	B
3	Pluvieux	Moyenne	Haute	Vrai	A
4	Ensoleillé	Moyenne	Normale	faux	A

$$Distance(5,1) = 0 + 0 + (1 - P(Haute|A)) + 0$$

$$= (1 - \frac{2}{3}) = 0.333$$

Exemple de calcul de distance

Instance test

id	Temps	Température	Humidité	Vent	Classe
5	Ensoleillé	Moyenne	?	Faux	A

Base d'apprentissage

id	Temps	Température	Humidité	Vent	Classe
1	Ensoleillé	Moyenne	Haute	Faux	A
2	Pluvieux	Moyenne	Normale	Faux	B
3	Pluvieux	Moyenne	Haute	Vrai	A
4	Ensoleillé	Moyenne	Normale	faux	A

Distance(5, 1)=0.33
Si Near= 3 : Distance(5, 1) $<$ 3, les deux objets sont PPV de même classe

Exemple de calcul de distance

Instance test

id	Temps	Température	Humidité	Vent	Classe
5	Ensoleillé	Moyenne	?	Faux	A

Base d'apprentissage

id	Temps	Température	Humidité	Vent	Classe
1	Ensoleillé	Moyenne	Haute	Faux	A
2	Pluvieux	Moyenne	Normale	Faux	B
3	Pluvieux	Moyenne	Haute	Vrai	A
4	Ensoleillé	Moyenne	Normale	faux	A

$$Distance(5,2) = 1 + 0 + (1 - P(normale|A)) + 0$$

$$= 1 + (1 - \frac{1}{3}) = 1.66666$$

Exemple de calcul de distance

Instance test

id	Temps	Température	Humidité	Vent	Classe
5	Ensoleillé	Moyenne	?	Faux	A

Base d'apprentissage

id	Temps	Température	Humidité	Vent	Classe
1	Ensoleillé	Moyenne	Haute	Faux	A
2	Pluvieux	Moyenne	Normale	Faux	B
3	Pluvieux	Moyenne	Haute	Vrai	A
4	Ensoleillé	Moyenne	Normale	faux	A

Distance(5, 2)=1.666

Si Near= 3 : Distance(5, 2) $<$ 3 et les deux objets sont PPV des classes différentes

Exemple : Analyser-Instance

Zoo

- Base d'apprentissage :101 objets et 17 attributs
- Classe : 7 valeurs
- Base de test : 71 objets

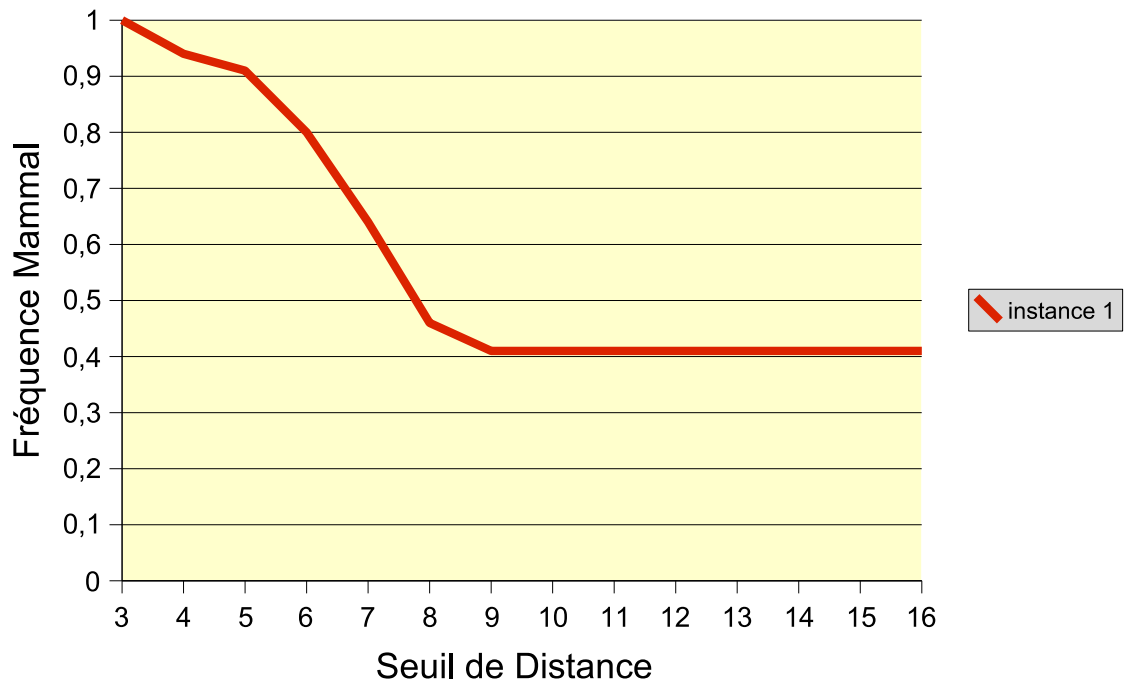
Attributs	Instance1
hair	true
feathers	false
eggs	true
milk	?
airborne	false
aquatic	true
predator	?
toothed	false
backbone	true
breathes	true
venomous	false
fins	false
legs	4
tail	?
domestic	false
catsize	true
classe	mammal

Exemple : Analyser-Instance

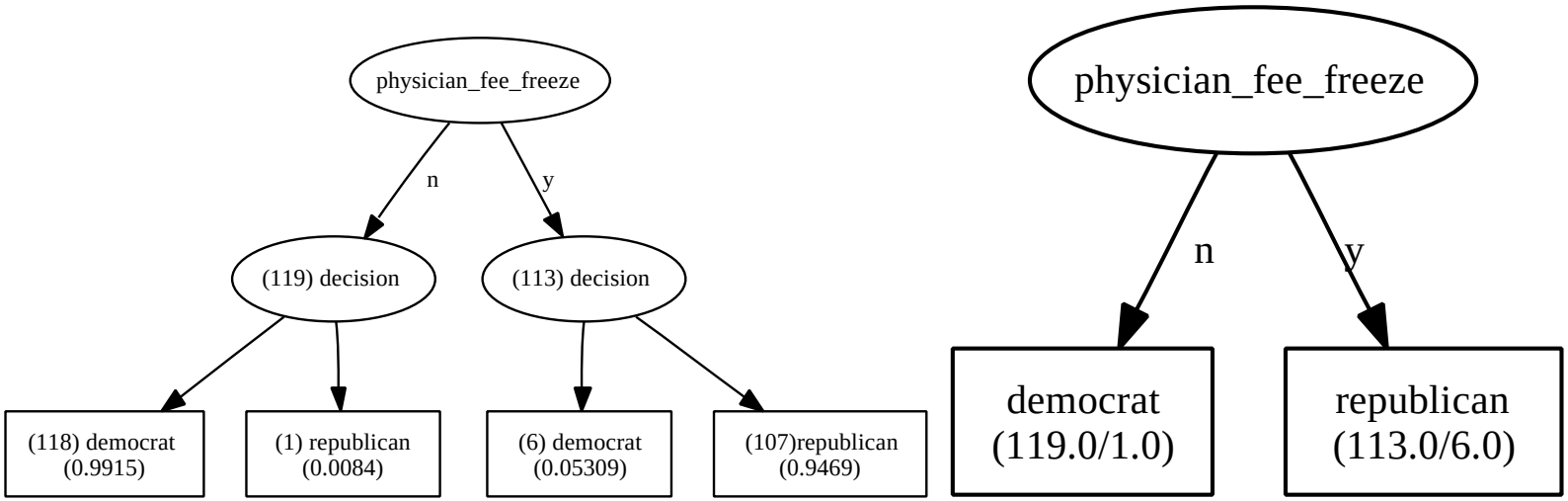
Valeurs de classe :
mammal, bird, reptile, fish, amphibian, insect, invertebrate

Near	Résultat	Instances proches
3	(**100%**, 0%, 0%, 0%, 0%, 0%, 0%)	2
4	(**94%**, 0%, 5%, 0%, 0%, 0%, 0%)	19
5	(**91%**, **3%**, **3%**, 0%, **3%**, 0%, 0%)	34
7	(**64%**, 17%, 5%, 5%, 6%, 0%, 3%)	64
9	(41%, 20%, 5%, 13%, 4%, 7%, 10%)	100
AAP	(**100%**, 0%, 0%, 0%, 0%, 0%, 0%)	
AAO	(0%, 0%, **100%**, 0%, 0%, 0%, 0%)	
C4.5	(**51%**, 0%, **27%**, 0%, **22%**, 0%, 0%)	

Fréquence de la classe *mammifère*



La base Vote



L'AAP de la base Vote pour un seuil 0.5 et l'arbre de C4.5

La base Vote

Vote

- Base d'apprentissage : 232 objets, 16 attributs
- Base de test : 240 objets
- Classe : Democrat et Republican

attributes	instance 1	instance 2	instance 3
physician-fee	?	?	?
el-salvador	y	?	y
education	y	n	n
crime	y	n	n
near=8	**(16%, 83%)**	**(91%, 08%)**	**(92%, 07%)**
near=10	(29%, 70%)	(84%, 15%)	(75%, 24%)
near=12	(38%, 61%)	(70%, 29%)	(57%, 42%)
AAP	(11%,89%)	(99%,01%)	(85%,15%)
AAO	(0%,100%)	(100%,0%)	(100%,0%)
C4.5	(53%,47%)	(53%,47%)	(53%,47%)

Plan

Conclusion 1

- Traitement du problème des valeurs manquantes
- Deux propositions :
  - Arbres d'Attributs Ordonnés Probabilistes (AAOP)
  - Arbres d'Attributs Probabilistes (AAP) : *Dépendances*
- Tests sur plusieurs bases de données
- Comparaison avec les résultats de classement donnés par C4.5 et *AAO*

Conclusion 1

- Traitement du problème des valeurs manquantes
- Deux propositions :
  - Arbres d'Attributs Ordonnés Probabilistes (AAOP)
  - Arbres d'Attributs Probabilistes (AAP) : *Dépendances*
- Tests sur plusieurs bases de données
- Comparaison avec les résultats de classement donnés par C4.5 et *AAO*

- Résultat de classement : une distribution de probabilité
- Performance meilleure quand les attributs sont dépendants
- Complexité *exponentielle* en fonction du nombre d'attributs manquants
  - Délimite l'usage : peu de valeurs manquantes $\Rightarrow$ amélioration par rapport à C4.5 et AAO est importante

Conclusion 2

- Proposition d'un algorithme, *Analyser-Instance*, issu de la méthode des k plus proches voisins
- Calcul pour chaque objet à classer de la fréquence de ses objets les plus proches de chaque classe
- Utilisation d'une fonction de distance qui prend en compte les valeurs manquantes dans l'objet à classer

Conclusion 2

- Proposition d'un algorithme, *Analyser-Instance*, issu de la méthode des k plus proches voisins
- Calcul pour chaque objet à classer de la fréquence de ses objets les plus proches de chaque classe
- Utilisation d'une fonction de distance qui prend en compte les valeurs manquantes dans l'objet à classer

- Résultats de l'algorithme *Analyser-Instance* proches des résultats données par notre approche
- Mauvaise performance de C4.5 lorsque l'attribut manquant est la racine de l'arbre

Perspectives

- Comparaison des distribution de probabilités avec celles obtenues avec les Réseaux Bayésiens et les Règles d'Association
- Arbres de décision probabilistes

Perspectives

- Comparaison des distribution de probabilités avec celles obtenues avec les Réseaux Bayésiens et les Règles d'Association
- Arbres de décision probabilistes

Simplification de notre approche pour diminuer la complexité de classement

- Construction d'une seule famille d'arbres *AAP* qui prend en compte les dépendances
- Pour résoudre le problème de Cycle : une solution hybride qui combine :
  - Arbres d'Attributs Probabilistes (AAP)
  - Analyser-Instance (utiliser pour l'attribut le moins dépendant de la classe)

Merci pour votre attention