DJILLALI LIABES UNIVERSITY OF SIDI BEL ABBES
FACULTY OF EXACT SCIENCES
DEPARTMENT OF COMPUTER SCIENCES

Module : Apprentissage Automatique
1ST YEAR OF MASTER'S DEGEREE IN
NETWORKS,SYSTEMS & INFORMATION SECURITY(RSSI)
2021/2022

Construction d'arbre de Décisions Algorithmes ID3 et C4.5

Student:
HADJAZI Mohammed Hisham
Group: 01 / RSSI

Module Instructor:
Pr.ELBERRICHI Zakaria
TD Instructor:
Dr.FAHSI.Mahmoud

January 1, 2022

Contents

List of Figures

Chapter 1

Fiche TD-05 Solutions

1.1 Questions de cours.

1.1.1 1. Avantages et limites des arbres de décisions ?

Advantages

1. **Clear Visualization**: The algorithm is simple to understand, interpret and visualize as the idea is mostly used in our daily lives. Output of a Decision Tree can be easily interpreted by humans.
2. **Simple and easy to understand** : Decision Tree looks like simple if-else statements which are very easy to understand.
3. Decision Tree can be used for both classification and regression problems.
4. Decision Tree can handle both **continuous and categorical variables.**
5. **No feature scaling required**: No feature scaling (standardization and normalization) required in case of Decision Tree as it uses rule based approach instead of distance calculation.
6. **Handles non-linear parameters efficiently**: Non linear parameters don't affect the performance of a Decision Tree unlike curve based algorithms. So, if there is high non-linearity between the independent variables, Decision Trees may outperform as compared to other curve based algorithms.
7. Decision Tree can automatically **handle missing values**.
8. Decision Tree is usually **robust to outliers** and can handle them automatically.
9. **Less Training Period**: Training period is less as compared to Random Forest because it generates only one tree unlike forest of trees in the Random Forest.

Disadvantages

1. **Overfitting**: This is the main problem of the Decision Tree. It generally leads to overfitting of the data which ultimately leads to wrong predictions. In order to fit the data (even noisy data), it keeps generating new nodes and ultimately the tree becomes too complex to interpret. In this way, it loses its generalization capabilities. It performs very well on the trained data but starts making a lot of mistakes on the unseen data.
2. **High variance**: As mentioned in point 1, Decision Tree generally leads to the overfitting of data. Due to the overfitting, there are very high chances of high

variance in the output which leads to many errors in the final estimation and shows high inaccuracy in the results. In order to achieve zero bias (overfitting), it leads to high variance.

3. **Unstable**: Adding a new data point can lead to re-generation of the overall tree and all nodes need to be recalculated and recreated.

4. **Affected by noise**: Little bit of noise can make it unstable which leads to wrong predictions.

5. **Not suitable for large datasets**: If data size is large, then one single tree may grow complex and lead to overfitting. So in this case, we should use Random Forest instead of a single Decision Tree.

1.1.2 2. Faire un tableau comparatif des algorithmes de classification supervisée vus aux cours.

1.1.3 3. Domaines d'applications ?

1. **Business Management** In the past decades, many organizations had created their own databases to enhance their customer services. Decision trees are a possible way to extract useful information from databases and they have already been employed in many applications in the domain of business and management. In particular, decision tree modelling is widely used in customer relationship management and fraud detection, which are presented in subsections below.

2. **Customer Relationship Management** A frequently used approach to manage customers' relationships is to investigate how individuals access online services. Such an investigation is mainly performed by collecting and analyzing individuals' usage data and then providing recommendations based on the extracted information. Lee et al. (2007) apply decision trees to investigate the relationships between the customers' needs and preferences and the success

Algorithm	**Problem Type**
KNN	Either
Naive Bayes	Classification
Decision trees	Either
Average predictive accuracy	**Training speed**
Lower	Fast
Lower	Fast (excluding feature extraction)
Lower	Fast
Performs well with small number of observations?	**Handles lots of irrelevant features well (separates signal from noise)?**
No	No
Yes	Yes
No	No
Parametric?	**Features might need scaling?**
No	Yes
Yes	No
No	No
Results interpretable by you?	**Easy to explain algorithm to others?**
Yes	Yes
Somewhat	Somewhat
Somewhat	Somewhat
Prediction speed	**Amount of parameter tuning needed (excluding feature selection)**
Depends on n	Minimal
Fast	Some for feature extraction
Fast	Some
Automatically learns feature interactions?	**Gives calibrated probabilities of class membership?**
No	Yes
No	No
Yes	Possibly

of online shopping. In their study, the frequency of using online shopping is used as a label to classify users into two categories: (a) users who rarely used online shopping and (b) users who frequently used online shopping. In terms of the former, the model suggests that the time customers need to spend in a transaction and how urgent customers need to purchase a product are the most important factors which need to be considered. With respect to the latter, the created model indicates that price and the degree of human resources involved (e.g. the requirements of contacts with the employees of the company in having services) are the most important factors. The created decision trees also suggest that the success of an online shopping highly depends on the frequency of customers' purchases and the price of the products. Findings discovered by decision trees are useful for understanding their customers' needs and preferences.

3. **Fraudulent Statement Detection** Another widely used business application is the detection of Fraudulent Financial Statements (FFS). Such an application is particularly important because the existence of FFS may result in reducing the government's tax income (Spathis et al., 2003). A traditional way to identify FFS is to employ statistical methods. However, it is difficult to discover all hidden information due to the necessity of making a huge number of assumptions and predefining the relationships among the large number of variables in a financial statement.

4. **Engineering** The other important application domain that decision trees can support is engineering. In particular, decision trees are widely used in energy consumption and fault diagnosis, which are described in subsections below.

5. **Energy Consumption** Energy consumption concerns how much electricity has been used by individuals. The investigation of energy consumption becomes an important issue as it helps utility companies identify the amount of energy needed. Although many existing methods can be used for the investigation of energy consumption, decision trees appear to be preferred. This is due to the fact that a hierarchical structure provided by decision trees is useful to present the deep level of information and insight. For instance, Tso and Yau (2007) create a decision tree model to identify the relationships between a household and its electricity consumptions in Hong Kong. Findings from their tree model illustrate that the number of household members are the most determinant factor of energy consumption in summer, and both the number of air-conditioner and the size of a flat are the second most important factors. In addition to such findings, their tree model identifies that a household with four or more members with a flat size larger than 817ft2 is the highest electricity consumption group. On the other hand, households which have less than four family members and without air-conditioners are the smallest electricity consumption group. Such findings from decision trees not only provide a deeper insight of the electricity consumptions within an area but also give guidelines to electricity companies about the right time they need to generate more electricity.

6. **Fault Diagnosis** Another widely used application in the engineering domain is the detection of faults, especially in the identification of a faulty bearing in rotary machineries. This is probably because a bearing is one of the most important components that directly influences the operation of a rotary machine. To detect the existence of a faulty bearing, engineers tend to measure the vibration and acoustic emission (AE) signals emanated from the rotary machine.

However, the measurement involves a number of variables, some of which may be less relevant to the investigation. Decision trees are a possible tool to remove such irrelevant variables as they can be used for the purposes of feature selection. Sugumaran and Ramachandran (2007) create a decision tree model to identify the features that may significantly affect the investigation of a faulty bearing. Through feature selection, three attributes were chosen to discriminate the faulty conditions of a bearing, i.e., the minimum value of the vibration signal, the standard deviation of the vibration signal, and kurtosis. The chosen attributes, subsequently, were used for creating another decision tree model. Evaluations from this model show that more than 95

7. **Healthcare Management** As decision tree modelling can be used for making predictions, there are an increasing number of studies that investigate to use decision trees in health-care management. For instance, Chang (2007) has developed a decision tree model on the basis of 516 pieces of data to explore the hidden knowledge located within the medical history of developmentally-delayed children. The created model identifies that the majority of illnesses will result in delays in cognitive development, language development, and motor development, of which accuracies are 77.3

1.1.4 4. Quel est le proverbe arabe sur lequel s'appuie la construction des arbres de décisions?

A tree begins with a seed

1.2 Exercice 1 : Pour le tableau Data Weather « symbolique » vu au cours, construire l'arbre de décision, en utilisant le Rapport de Gain. Sans consulter les valeurs données au cours.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	0	no
sunny	hot	high	1	no
overcast	hot	high	0	yes
rainy	mild	high	0	yes
rainy	cool	normal	0	yes
rainy	cool	normal	1	no
overcast	cool	normal	1	yes
sunny	mild	high	0	no
sunny	cool	normal	0	yes
rainy	mild	normal	0	yes
sunny	mild	normal	1	yes
overcast	mild	high	1	yes
overcast	hot	normal	0	yes
rainy	mild	high	1	no

		play				total
		yes		no		
Outlook	Sunny	2	0.4	3	0.6	5
	Overcast	4	1	0	0	4
	Rainy	3	0.6	2	0.4	5
						14

		play				total
		yes		no		
Temperature	hot	2	0.5	2	0.5	4
	mild	4	0.666666666666667	2	0.333333333333333	6
	cool	3	0.75	1	0.25	4
						14

		play				total
		yes		no		
Humidity	high	3	0.428571428571429	4	0.571428571428571	7
	normal	6	0.857142857142857	1	0.142857142857143	7
						14

		play				total
		yes		no		
Windy	1	3	0.5	3	0.5	6
	0	6	0.75	2	0.25	8
						14

feature	Information Gain	Entropy
Outlook	0.246749819774439	1.58
Temperature	0.029222565658955	1.56
Humidity	0.151835501362342	1
Windy	0.04812703040827	0.99

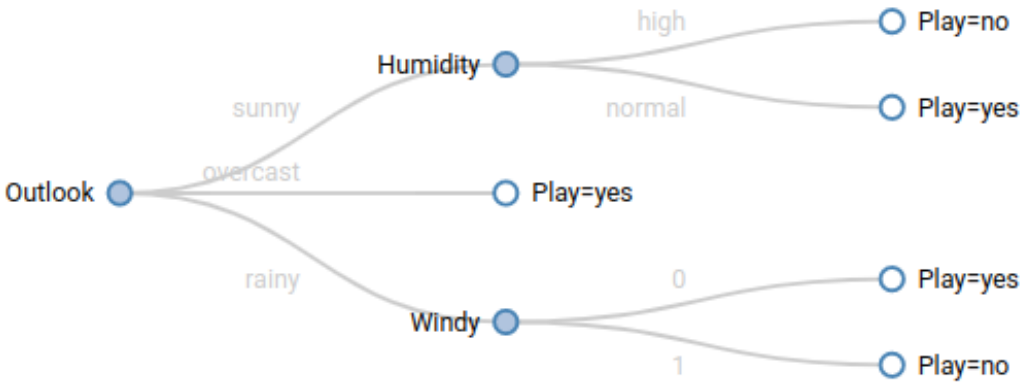

FIGURE 1.1: Decition Tree

1.3 Exercice 2 : En utilisant ID3, la mesure du gain, et la table d'apprentissage suivante (classe maladie), donner l'arbre de décision (détailler les étapes, et les calculs). Classer l'instance : oui Abdomen non ?

Fièvre	Douleur	Toux	Maladie
oui	Abdomen	non	Appendicite
non	Abdomen	oui	Appendicite
oui	gorge	non	rhume
oui	gorge	oui	rhume
non	gorge	oui	mal de gorge
oui	non	non	aucune
oui	non	oui	rhume
non	non	oui	refroidissement
non	non	non	aucune

		malade										total
		Appendicite		rhume		mal de gorge		aucune		refroidissement		
Fievre	oui	1	0.2	3	0.6	0	0	1	0.2	0	0	5
	non	1	0.25	0	0	1	0.25	1	0.25	1	0.25	4
												9

		malade										total
		Appendicite		rhume		mal de gorge		aucune		refroidissement		
Douleur	Abdomen	2	1	0	0	0	0	0	0	0	0	2
	gorge	0	0	2	0.67	0	0	0	0	1	0.33	3
	non	0	0	1	0.25	0	0	2	0.5	1	0.25	4
												9

		malade										total
		Appendicite		rhume		mal de gorge		aucune		refroidissement		
Toux	oui	1	0.2	2	0.4	1	0.2	0	0	1	0.2	5
	non	1	0.25	1	0.25	0	0	2	0.5	0	0	4
												9

feature	Informatiion Gain	Entropy
Fièvre	0.546631615393778	0.99
Douleur	1.22439444540599	1.51
Toux	0.462755226264504	0.99

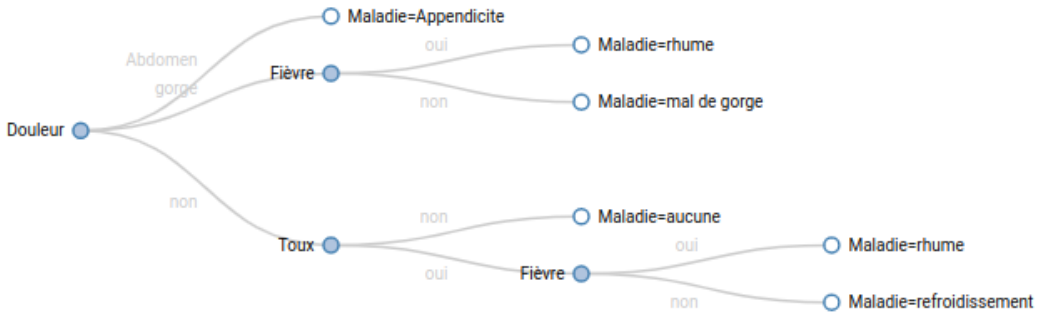

FIGURE 1.2: Decition Tree

1.4 Exercice 3 : En utilisant ID3, la mesure du gain, et la table d'apprentissage suivante (classe species),donner l'arbre de décision (détailler les étapes, et les calculs)

Touched	Hair	Breathes	Legs	Species
Touched	Hair	Breathes	Legs	Mammal
Touched	Hair	Breathes	Legs	Mammal
Touched	Not Hair	Breathes	Not Legs	Reptile
Not Touched	Hair	Breathes	Legs	Mammal
Touched	Hair	Breathes	Legs	Mammal
Touched	Hair	Breathes	Legs	Mammal
Touched	Not Hair	Not Breathes	Not Legs	Reptile
Touched	Not Hair	Breathes	Not Legs	Reptile
Touched	Not Hair	Breathes	Legs	Mammal
Not Touched	Not Hair	Breathes	Legs	Reptile

		Species				total
		Mammal		Reptile		
Touched	Touched	5	0.625	3	0.375	8
	Not Touched	1	0.5	1	0.5	2
						10

		Species				total
		Mammal		Reptile		
Hair	Hair	5	1	0	0	5
	Not Hair	1	0.2	4	0.8	5
						10

		Species				total
		Mammal		Reptile		
Breathes	Breathes	6	0.67	3	0.33	9
	Not Breathes	0	0	1	1	1
						10

		Species				total
		Mammal		Reptile		
Legs	Legs	6	0.857142857142857	1	0.142857142857143	7
	Not Legs	0	0	3	1	3
						10

feature	Information Gain	Entropy
Touched	0.007403392114697	0.72
Hair	0.609986547010987	1
Breathes	0.144484343805628	0.47
Legs	0.556779649447039	0.88

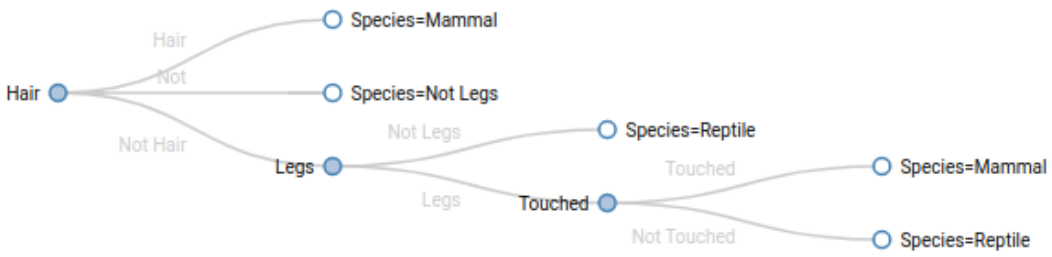

FIGURE 1.3: Decition Tree

1.5 Exercice 4 : En considérant seulement l'attribut Taille pour la classe Sexe. Trouver le meilleur point de split (S) pour l'algorithme C4.5 Pour le pseudo-attribut (Taille <S)

Taille	161	185	184	164	169	179	176	175	180
Sexe	F	F	M	F	M	F	F	M	M

1.6 Exercice 5 :

1.6.1 Expliquer le principe de C4.5.

The C4.5 algorithm is used in Data Mining as a Decision Tree Classifier which can be employed to generate a decision, based on a certain sample of data (univariate or multivariate predictors).

C4.5 is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals. C4.5 converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules. This accuracy of each rule is then evaluated to determine the order in which they should be applied. Pruning is done by removing a rule's precondition if the accuracy of the rule improves without it.

1.6.2 Qu'a-t-il apporté de plus qu'ID3 ?

- The algorithm inherently employs Single Pass Pruning Process to Mitigate overfitting.
- It can work with both Discrete and Continuous Data
- C4.5 can handle the issue of incomplete data very well

1.6.3 En utilisant l'algorithme C4.5 et la mesure rapport de gain, sur la table d'apprentissage suivantes, Donner le nœud racine de l'arbre Les calculs doivent être détaillés

A	B	C	Class
S	85	F	Y
S	90	T	N
O	86	F	Y
R	80	F	Y
R	70	T	Y
O	65	T	Y
S	95	F	N
S	70	F	N

1.7 Exercice 6 : En utilisant l'algorithme ID3 et la mesure du rapport du gain et la base d'apprentissage suivante, construire l'arbre de décision. Montrer toutes les étapes du calcul. Dessiner le ou les arbres possibles. Trouver la classe de l'instance : F M M

E	C	R	Class
M	E	E	Y
F	F	M	N
E	M	F	N
M	M	M	Y
F	M	E	Y
E	E	F	Y
F	F	F	N
M	M	E	Y

1.8 Exercice 7 : En utilisant l'algorithme ID3 et la mesure du gain, construire l'arbre du tableau d'apprentissage, puis classer l'instance : <=30 low no excellent

age	income	student	credit_rating	buys_computer
=30	high	no	fair	no
=30	high	no	excellent	no
31..40	high	no	fair	yes
40	medium	no	fair	yes
40	low	yes	fair	yes
40	low	yes	excellent	no
31..40	low	yes	excellent	yes
=30	medium	no	fair	no
=30	low	yes	fair	yes
40	medium	yes	fair	yes
=30	medium	yes	excellent	yes
31..40	medium	no	excellent	yes
31..40	high	yes	fair	yes
40	medium	no	excellent	no