資料探勘專題作業一

決策樹預測

指導教授:
許中川 教授

成員:
M11112075 徐紹鈞
M11121028 陳韋諠
日期:
2022 年 10 月 27 號

摘要

本次研究主要是利用 Adult Dataset、Abalone Dataset，Adult Dataset 主要是在預測薪水有無高於 50000，而 Abalone Dataset 則是預測鮑魚的年齡，會先透過資料前處理並且將資料集切割為 80%為訓練資料，20%則為測試資料，本研究主要會利用決策樹進行分類預測，分類預測後會進行決策樹修剪，讓績效更加提升。

一、緒論

1.1 動機

1-1-1 Abalone Dataset

現實生活中，人類如果想得知鮑魚的年齡，必須是用物理的方式來測量，那便是通過將殼切開、染色並通過顯微鏡計算環的數量來確定，然而，這是一項無聊且耗時的任務。本研究希望透過機器學習來預測鮑魚的年齡，來節省人工作業耗時的問題。

1-1-2 Adult Dataset

為了查詢現代人類於薪資方面落差中所具備特質及專長，便在人口普查資料及當中進行資料探勘實作並利用決策樹相對應的特質，來深入探討人口中收入值變化，以供研究者研究。

1.2 目的

由於人工實際測量鮑魚的年齡非常耗時，為了解決這個問題，本研究選取 UCI Datasets，並使用決策樹進行分類，投入鮑魚的性別、長度、直徑、高度、整體重量、去殼重量、內臟重量、殼重量等 9 個屬性資料。

二、方法

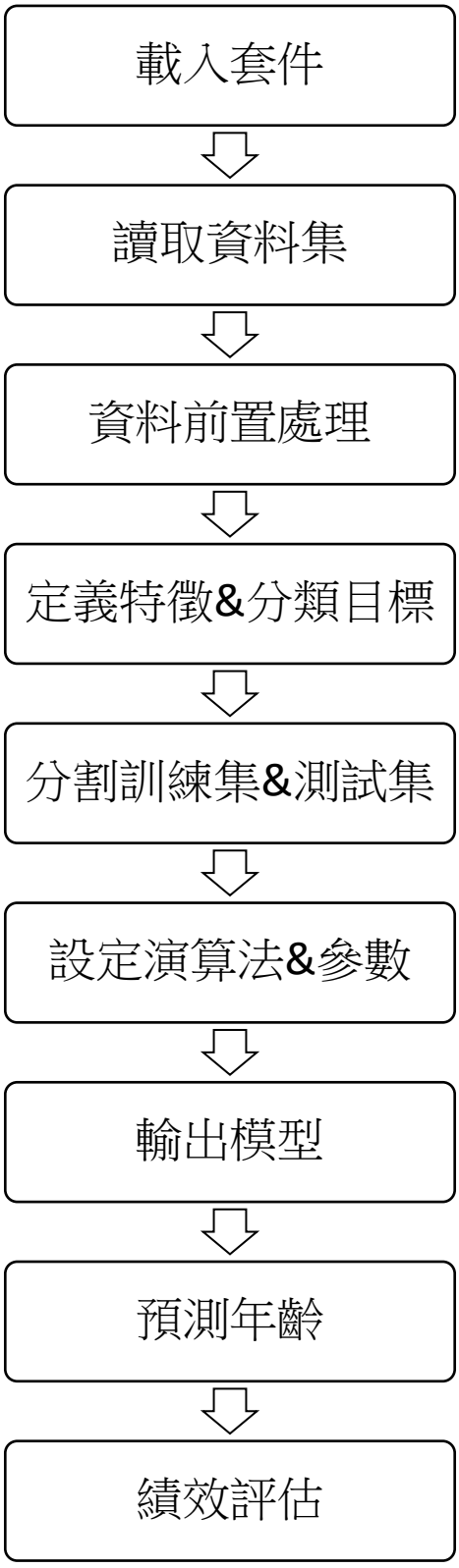

載入套件
讀取資料集
資料前置處理
定義特徵&分類目標
分割訓練集&測試集
設定演算法&參數
輸出模型
預測年齡
績效評估

圖 1 程式架構

2-1 程式執行方法

2-1-1 決策樹(Decision Tree)

決策樹（Decision tree）由一個決策圖和可能的結果（包括資源成本和風險）組成，用來創建到達目標的規劃。決策樹是一個利用像樹一樣的圖形或決策模型的決策支持工具，包括隨機事件結果，資源代價和實用性。它是一個算法顯示的方法。決策樹經常在運籌學中使用，特別是在決策分析中，它幫助研究者確認一個能最可能達到目標的策略。

2-1-2 隨機森林(Random Forest)

在機器學習中，隨機森林是一個包含多個決策樹的分類器，並且其輸出的類別是由個別樹輸出的類別的眾數而定。隨機森林的優點在於:1.對於很多種資料，它可以產生高準確度的分類器、2.它可以處理大量的輸入變數、也可以在決定類別時，評估變數的重要性、在建造森林時，它可以在內部對於一般化後的誤差產生不偏差的估計 3.它包含一個好方法可以估計遺失的資料 4.如果有很大一部分的資料遺失，仍可以維持準確度 5.學習過程是很快速的。

2-1-3KDD(Knowledge Discovery in Database)

KDD 是從資料集中識別出有效的、新穎的、潛在有用的，以及最終可理解的模式的非平凡過程。知識發現將信息變為知識，從資料礦山中找到蘊藏的知識金塊，將為知識創新和知識經濟的發展作出貢獻。

三、實驗

3-1-1 Abalone 資料集

此資料集建立於 1995 年 12 月 1 日共 4177 筆，10 個欄位

表一 Abalone Dataset 欄位資料說明彙總表

Sex	性別
Length	長度
Diameter	直徑
Height	高度
Whole weight	整體重量
Shucked weight	去殼重量
Viscera weight	內臟重量
Shell weight	外殼重量
Rings	環數量

3-1-2 Adult 資料集

此資料集建立於 1996 年 5 月 1 日共 48842 筆，14 個欄位

表二 Adult Dataset 欄位資料說明彙總表

age	歲數
workclass	工人階級
fnlwgt	NA
education	教育程度
education-num	教育程度
marital-status	婚姻狀況
occupation:	職業
relationship	關係
race	種族
sex	性別
capital-gain	資本收益
capital-loss	資本損失
hours-per-week	每週工時
native-country	國度

3.1.3 Abalone 資料集實驗數據

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550	7
...	...	...	...	...	...	...	...	...	...
4172	F	0.565	0.450	0.165	0.8870	0.3700	0.2390	0.2490	11
4173	M	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605	10
4174	M	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080	9
4175	F	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960	10
4176	M	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950	12

圖 2 Abalone Dataset

3.1.4 Adult 資料集實驗數據

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week	workclass_?	workclass_Federal-gov	workclass_Local-gov	workclass_Never-worked	...	native-country_Scotland	native-country_South
0	39	77516	13	2174	0	40	0	0	0	0	...	0	0
1	50	83311	13	0	0	13	0	0	0	0	...	0	0
2	38	215646	9	0	0	40	0	0	0	0	...	0	0
3	53	234721	7	0	0	40	0	0	0	0	...	0	0
4	28	338409	13	0	0	40	0	0	0	0	...	0	0

5 rows × 94 columns

圖 3 Adult Dataset

3-2 前置處理

在各資料集中，由原本的 name 檔案轉換成 csv 檔以利觀察資料，我們可以發現在資料集中有許多的空值，故本研究將會利用人工篩選的方式去除空值。

3-3 實驗設計

1.載入:由 Scikit-Learn 演算法載入決策樹的分類 (Classifier) 套件、決策樹建模的分離(train_test_split)套件、決策樹度量的分類報表(classification_report)，最後載入 python 的繪圖語言(pydot)與建立記憶體內的 str 輸入輸出空間(StringIO)。

2.森林繪製:由 StringIO 的輸入 pydot 的指令繪製 pydot 語言敘述的圖形，最後透過特徵值與分類值，繪製出隨機森林。

3-3-1 抓取 adult.data 網址引入 jupyter

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data'

import pandas as pd
```

3-3-2 引入資料呈現下圖，無 attribute 之說明只有代碼

```
df=pd.read_csv(url,header=None)
df.head()
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

3-3-3 將 attribute 類別依序填入程式中

```
df.columns = ['age', 'workclass', 'fnlwgt', 'education', 'education-num', 'marital-status', 'occupation',
              'relationship', 'race', 'sex', 'capital-gain', 'capital-loss', 'hours-per-week', 'native
               'income']
df.head()
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40

3-3-4 因電腦無法辨識字串故利用 pd.get_dummies()將資料修改，變成 attribute 新增子項目讓資料呈現 0 與 1。

```
df = pd.get_dummies(df)
df.head()
```

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week	workclass_?	workclass_Federal-gov	workclass_Local-gov	workclass_Never-worked	...	native-country_Scotland	native-country_South
0	39	77516	13	2174	0	40	0	0	0	0	...	0	0
1	50	83311	13	0	0	13	0	0	0	0	...	0	0
2	38	215646	9	0	0	40	0	0	0	0	...	0	0
3	53	234721	7	0	0	40	0	0	0	0	...	0	0
4	28	338409	13	0	0	40	0	0	0	0	...	0	0

5 rows × 94 columns

3-3-5 Class 方面已 Income 做為分類代表

```
# Show value counts of last column
df.iloc[:, -1].value_counts()
```

```
0    24720
1     7841
Name: income_ >50K, dtype: int64
```

```
# Split data into X and y
X = df.iloc[:, :-1]
y = df.iloc[:, -1]
```

3-3-6 帶入 sklearn.tree import DecisionTreeClassifier 模板即可預測衡量訓練資料及測試資料的分類正確率。

```
[11]  from sklearn.tree import DecisionTreeClassifier

[12]  # Build Decision Tree and score on test set
      model = DecisionTreeClassifier()
      model.fit(X_train, y_train)
      model.score(X_test, y_test)

      1.0

[13]  decisionTreeModel = DecisionTreeClassifier()
      decisionTreeModel.fit(X_train, y_train)

      print('train set accurancy: ',decisionTreeModel.score(X_train, y_train))
      print('test set accurancy: ',decisionTreeModel.score(X_test, y_test))

      train set accurancy:  1.0
      test set accurancy:  1.0
```

score: Any

3.4 實驗結果

下表二為鮑魚年齡預測結果表

表二 Abalone 預測結果

資料編號	預測	真實
0	8	15
1	5	7
2	10	9
3	10	10
4	8	7
5	19	8
6	5	20
7	9	16
8	15	9
9	13	19
10	9	14
11	6	10
12	7	11
13	9	10

下表三為成人資料集分類結果表

表三 成人資料集分類結果表

adult 分類結果

預測	真實		
0	0		Income<=50k
0	0		Income>50k
1	0		
0	0		
0	0		
0	0		
0	0		
0	1		
0	1		
0	1		
0	1		
0	1		
1	0		
0	0		
1	1		
0	0		
1	0		
0	0		
0	0		
1	1		
1	1		
0	0		
0	0		

四、結論

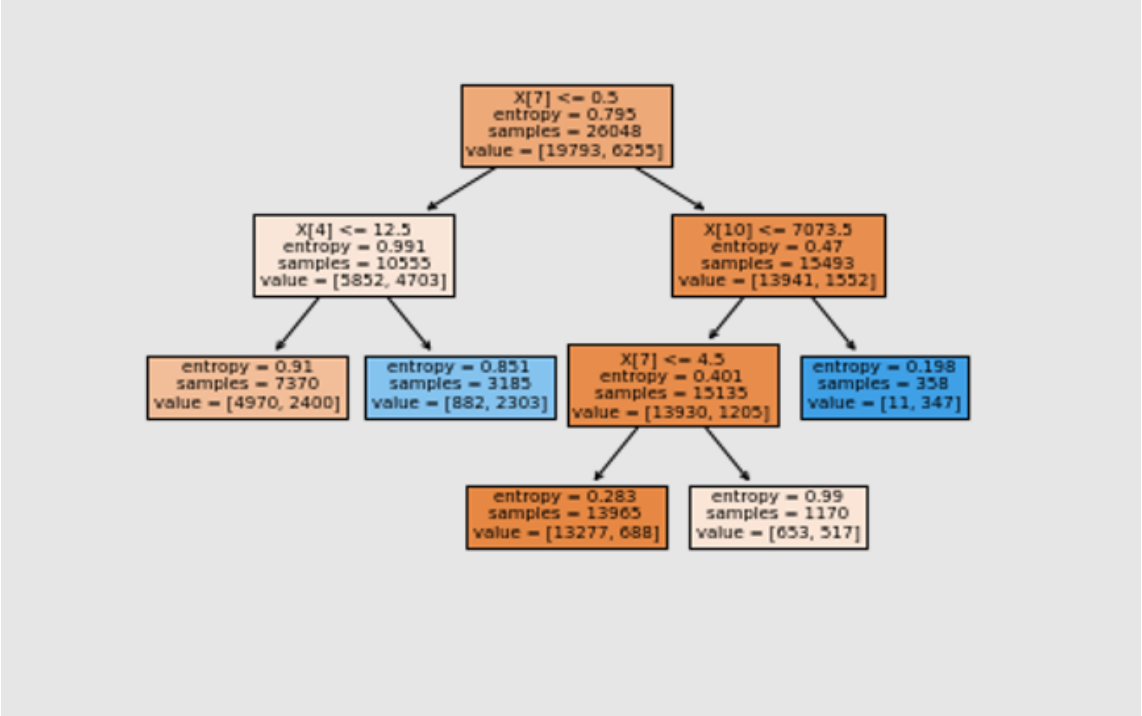
4.1 Abalone Dataset

本研究將 Abalone Dataset 使用分為 80%的訓練資料與 20%的測試資料，並使用決策數與隨機森林對 rings 類別做分類。

4-1-1 修剪前後績效

修剪前:0.81

修剪後:0.82



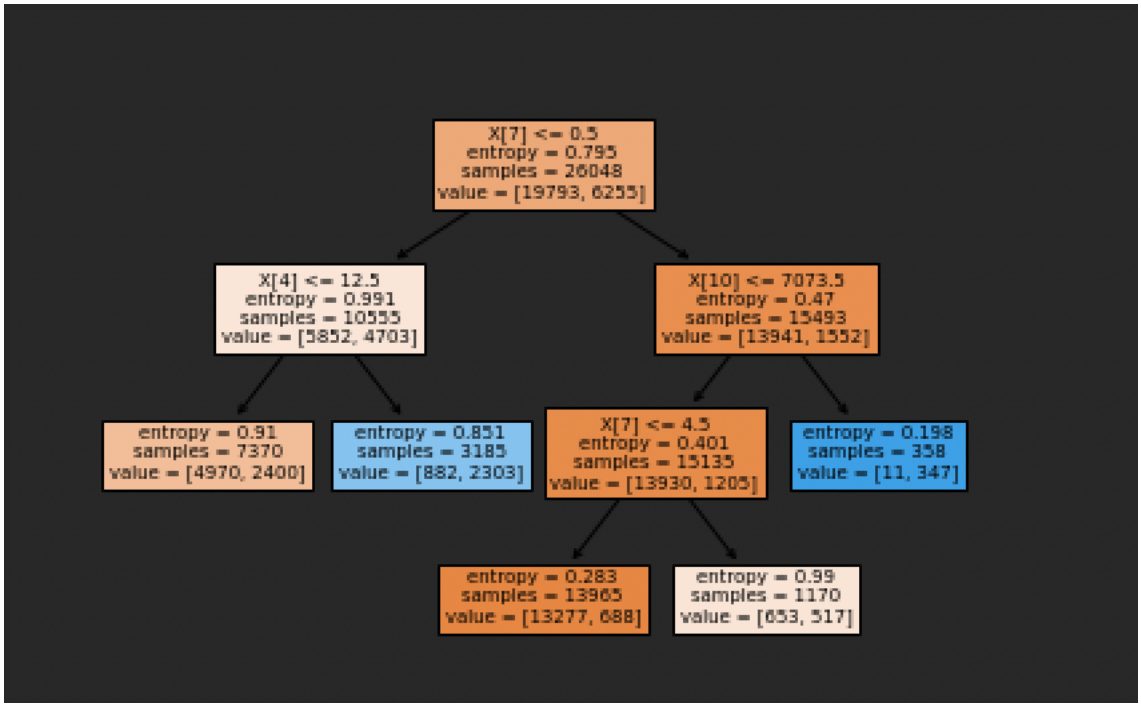
4.2 Adult Dataset

本研究將 Abalone Dataset 使用分為 80%的訓練資料與 20%的測試資料，並使用決策數與隨機森林對 rings 類別做分類。

4-2-1 修剪前後績效

修剪前:0.21

修剪後:0.27



五、參考文獻

1.Decision Tree Intro Variance Bias 資料參考：
https://www.youtube.com/watch?v=3apz65oBS-Y&t=416s
2.決策樹 (Decision tree)資料參考：
https://ithelp.ithome.com.tw/articles/10271143?sc=hot

3.鮑魚資料集參考資料：
https://archive.ics.uci.edu/ml/datasets/Abalone
4.Adult 資料集參考資料：https://archive.ics.uci.edu/ml/datasets/Adult
5.網站程式碼參考資料：
https://colab.research.google.com/drive/1IZGeRZZwCN5xk3cJPZqnhC5EEe5rxfzC?usp=sharing#scrollTo=SvcEtK0itKZu
6.網站程式碼參考資料：
https://colab.research.google.com/drive/1IZGeRZZwCN5xk3cJPZqnhC5EEe5rxfzC?usp=sharing