國立雲林科技大學資訊管理系

資料探勘-作業一

Department of Information Management

National Yunlin University of Science & Technology

Master Thesis

北京PM2.5資料集分析

Beijing PM2.5 Data Set Analysis

林冠州、葉家圻、張雅涵、劉柏陽

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

中華民國110年6月

June 2021

1. **緒論**

本章節將說明本研究的背景與動機和實驗的目的為何，以及對採用之資料集內容做簡略說明。

* 1. **動機**

中國自改革開放以來的發展過程面臨著眾多環境問題，至目前為止，其經濟增長主要依賴煤炭能源，雖然成功使國家經濟實力大幅度成長，但也導致環境迅速惡化。根據媒體報導，北京至上海這一沿線是中國空氣污染最為集中的地帶之一，其10％的土地面積造成了中國34%的PM2.5排放量。此外，2015年時中國記者柴靜的空氣污染紀錄片《穹頂之下》引起中國網友及周邊國家對污染問題的廣泛討論，其紀錄片由柴靜作為主講人對現狀進行介紹，插入採訪拍攝的片段，並採訪中國和洛杉磯的環境官員、工商管理人員及衛生官員，呼籲人們關注中國的空氣污染問題。除此之外，中國政府也自2011年起將造成灰霾天氣的PM2.5納入評價指標，由此可見，北京確實正努力試圖在全國範圍內降低空氣污染的濃度。

另外，為了多加嘗試使用決策樹(Decision tree)而不是機器學習(Machine learning)或神經網路(Neural network)等現今泛用技術進行預測及解決資料集所遇到之問題，本文還找了1994年之人口普查資料庫，預測任務即是確定一個人之年收入是否高於5萬美金。

* 1. **目的**

本文之第一項研究欲探討各氣象數據對於PM2.5之影響重要性，例如露點(Dew point)、溫度(Temp)、氣壓(Air Pressure)、風向(Wind direction)等因素，並使用UCI Machine Learning Respository所提供之「Beijing PM2.5 Data」資料集，內容涵蓋了北京美國大使館所監測之細懸浮微粒(PM2.5)數據與從Weather.nocrew.org獲得的北京首都國際機場(BCIA)的每小時氣象測量值作為資料集，藉由決策樹(Decision tree)做資料探勘，觀察各項因素與監測之細懸浮微粒(PM2.5)之關聯性，以期望日後能作為空氣品質預測與應對之參考，減輕災害程度。

除此之外，本文在第二項研究中採用成人資料集(Adult Data Set)所收集之人口普查收入數據，其中包含年齡(Age)、教育程度(Education level)、婚姻狀況(Marital Status)、職業(Occupation)、性別(Gender)等，及最主要的個人的年收入(Annual income)，用以預測何種個人背景的情況下年收入能超過5萬美金，提供各項屬性與收入的關聯性參考。

1. **方法**
   1. **實作說明**

本研究將「北京PM2.5資料集（Beijing PM2.5 Data Set）」、「成人資料集（Adult Data Set）」做數據的前置處理，其包括Missing Values的處理、Ono-hot encoding、資料聚合(Data Aggregation)、資料切割（即為將資料分成訓練資料(train data)以及測試資料(test data)），模型建構則利用決策樹並使用Entropy衡量各個樹節點（node）之不純度（Impurity），最後預測「何種環境數值會提升PM2.5並對人體造成危害」以及「何種成人的年收入會超過美金5萬元」，並輸出兩者的決策樹模型以及訓練資料與測試 資料的績效（accuracy），最終匯出分析結果表格。

* 1. **操作說明**

本研究執行環境皆為 Python3.6，並使用Visual Studio Code作為分析工具，利用 Pandas、Pandas\_Profiling以及Numpy來讀取資料以及做資料的前 處理，預測模型則利用 Sklearn 套件的 Tree 工具來建構，最後在使用 Graphviz、Matplotlib 套件將決策樹圖形輸出之。

1. **實驗**
2. **資料集**

本研究使用兩組資料集做預測分析，分別為2010至2014年的北京氣象數據與成人收入調查結果之相關數據，以下為該兩組資料集之資料名稱、資料筆數，以及資料表的欄位介紹。

### Beijing PM2.5 Data Set

名稱: Beijing PM2.5 Data Data Set

原始資料筆數: 43824

正規化後之訓練資料筆數: 29228

正規化後之測試資料筆數: 12527

表格一：Beijing PM2.5 Data Data Set資料集欄位介紹

|  |  |  |
| --- | --- | --- |
| 欄位 | 欄位名稱 | 內容 |
| 0 | No | row number |
| 1 | Year | year of data in this row |
| 2 | Month | month of data in this row |
| 3 | Day | day of data in this row |
| 4 | Hour | hour of data in this row |
| 5 | pm2.5 | PM2.5 concentration |
| 6 | DEWP | Dew Point |
| 7 | TEMP | Temperature |
| 8 | PRES | Pressure |
| 9 | cbwd | Combined wind direction |
| 10 | Iws | Cumulated wind speed |
| 11 | Is | Cumulated hours of snow |
| 12 | Ir | Cumulated hours of rain |

表格二：Beijing PM2.5 Data Data Set資料集欄位介紹

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 欄位 | 0 | 24 | 48 | 1772 | 28602 |
| 0 | 1 | 25 | 49 | 1773 | 28603 |
| 1 | 2010 | 2010 | 2010 | 2010 | 2013 |
| 2 | 1 | 1 | 1 | 3 | 4 |
| 3 | 1 | 2 | 3 | 15 | 6 |
| 4 | 0 | 0 | 0 | 20 | 18 |
| 5 | NaN | 129.0 | 90.0 | 30.0 | 11.0 |
| 6 | -21 | -16 | -7 | -11 | -12 |
| 7 | -11.0 | -4.0 | -6.0 | 1.0 | 12.0 |
| 8 | 1021.0 | 1020.0 | 1027.0 | 1026.0 | 1011.0 |
| 9 | NW | SE | SE | NW | SE |
| 10 | 1.79 | 1.79 | 58.56 | 132.31 | 9.39 |
| 11 | 0 | 0 | 4 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 |

### Adult Data Set

名稱：Adult Data Set

原始資料筆數：48842

正規化後之訓練資料筆數：22792

正規畫後之測試資料筆數：9769

表格 三 ：Adult資料集欄位介紹

|  |  |  |
| --- | --- | --- |
| 欄位 | 屬性 | 內容 |
| 0 | age | Continuous |
| 1 | workplace | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never worked |
| 2 | fnlwt | continuous |
| 3 | education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool |
| 4 | education-num | continuous |
| 5 | marital-status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| 6 | occupation | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces |
| 7 | relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| 8 | race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| 9 | sex | Female, Male |
| 10 | capital-gain | continuous |
| 11 | capital-loss | continuous |
| 12 | hours-per-week | continuous |
| 13 | native-country | United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands |
| 14 | salary | <=50K,>50K |

表格 四 ：顯示部分Adult資料集

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

1. **前置處理**

* Missing Values的處理：將資料中的缺失值以及過濾與預測不相關資訊。
* One-Hot encoding：對非數值的類別屬性進行特徵數字化。
* 資料聚合：將不同屬性的資料聚合成同一個屬性或是將同屬性中的相似值聚合成一個群組類以減少維度或是減少種類。
* 資料切割：將 70%當成訓練資料，其餘 30%為測試資料。

1. **前置處理-Beijing PM2.5 Data Set**

**One-Hot encoding：**

* 1. 對於cbwd屬性中的四個類別作One-Hot encoding

**資料聚合：**

* 1. 針對year、month、day、hour我們使用資料聚合將其合併成season屬性並以0、1、2、3代表四季。
  2. 將PM2.5依照中國的PM2.5指標聚合成0~6的PM(1)指標
  3. 將PM2.5依照是否對人體有害聚合成0、1兩類的PM(2)指標(此指標與上一個PM不同)
  4. 將TEMP減去DEWP得到新屬性DOTP(depression of the dew point)

表格一：Beijing PM2.5 Data Set資料集欄位介紹

|  |  |  |  |
| --- | --- | --- | --- |
| 欄位 |  | 欄位名稱 | 內容 |
| 13 |  | season | 0,1,2,3 |
| 14 |  | PM(1) | 0,1,2,3,4,5 |
| 15 |  | PM(2) | 0,1 |
| 16 |  | DOTP | TEMP - DEWP |
| 17 |  | cbwd\_NE | 表示東北風 |
| 18 |  | cbwd\_NW | 表示西北風 |
| 19 |  | cbwd\_SE | 表示東南風 |
| 20 |  | cbwd\_cv | 表示風向不明顯 |

表格一：Beijing PM2.5 Data Set前處理後部分資料

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 欄位 | 24 | 48 | 1772 | 28602 |
| 6 | -16 | -7 | -11 | -12 |
| 7 | -4.0 | -6.0 | 1.0 | 12.0 |
| 8 | 1020.0 | 1027.0 | 1026.0 | 1011.0 |
| 10 | 1.79 | 58.56 | 132.31 | 9.39 |
| 11 | 0 | 4 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 |
| 13 | 0 | 0 | 1 | 1 |
| 14 | 2 | 1 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 |
| 16 | 12 | 1 | 12 | 24 |
| 17 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 1 | 0 |
| 19 | 1 | 1 | 0 | 1 |
| 20 | 0 | 0 | 0 | 0 |

1. **前置處理-****Adult Data Set**

* 資料清洗：將資料中的缺失值及篩選與預測欄位不相關的資訊。
* 資料分群：給定尺度將資料劃分及群。
* One-Hot Encoding：對非數值的類別屬性資料轉換成一維向量資料。
* 資料切割：將70%作為訓練資料，其餘30%作為測試資料。

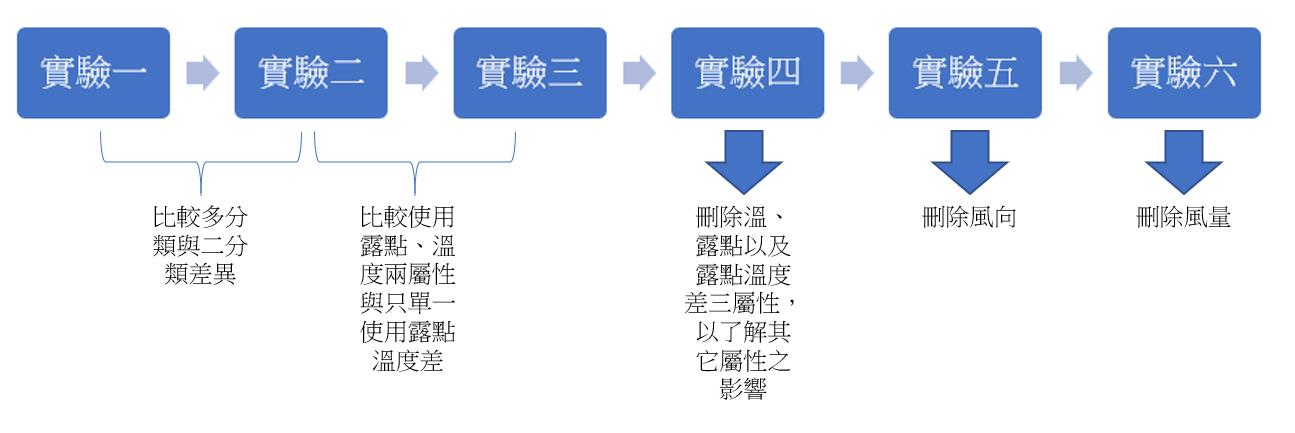
表格五 ： 成人資料及之資料前處理後部分資料

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| capital-gain | capital-loss | hours-per-week | encoding\_  income | age\_0-9 | age\_20-29 | age\_40-49 | age\_60-69 | … |
| 2174 | 0 | 40 | 0 | 0 | 1 | 0 | 0 | … |
| 0 | 0 | 13 | 0 | 0 | 0 | 1 | 0 | … |
| 0 | 0 | 40 | 0 | 0 | 1 | 0 | 0 | … |
| 0 | 0 | 40 | 0 | 0 | 0 | 1 | 0 | … |
| 0 | 0 | 40 | 0 | 0 | 1 | 0 | 0 | … |
| ... | ... | ... | ... | ... | ... | ... | … | … |

1. **實驗設計**

以下為兩個資料集之實驗設計

**3.3.1 Beijing PM2.5 Data Set實驗設計**



在Beijing PM2.5的部分主要分為六次實驗；首先為了瞭解結果的種類是否會影響分類效果，因此我們設計了兩種前處理c. d.並在第一次實驗中使用a.b.c.三個前處理做實驗，而後再第二次實驗中使用a.b.d. 三個前處理做實驗來比較效果。再對比完效果後我們在樹的結構中發現露點以及溫度對於分類的判斷十分重要(圖一)，而露點與溫度間的距離其實代表著空氣中的水氣是否飽和，基於此一觀點，我們設計了實驗三，將溫度減去露點，來獲得新屬性露點溫度差(欄位16)並實驗其影響，因此實驗三中使用a.b.d.e.四個前處理並將欄位6. 7.捨棄來做實驗對比效果。

為了排除其他屬性對決策樹的影響，延續了實驗三將欄位16也捨棄作為實驗四，希望了解剩下屬性中最重要的屬性；根據其結果可知，累積風量與風向是次重要之屬性因此分別設計兩個實驗五捨去欄位風向欄位(即不使用前處理a也不加入欄位9)，實驗六延續實驗五將累積風量欄位10也捨去，希望確定露點溫度差(欄位16)的重要性。



圖一：決策樹結構中的屬性重要性

**3.3.2 Adult Data Set實驗設計**

以下為Adult資料集模型建構之設計，依序的流程步驟：

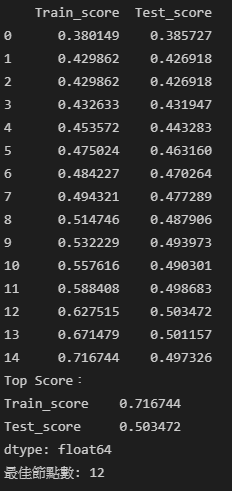
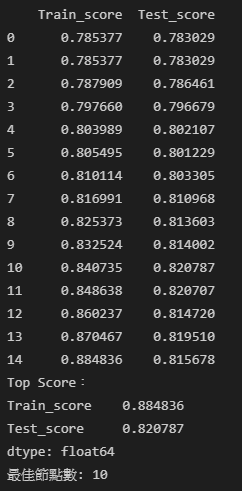
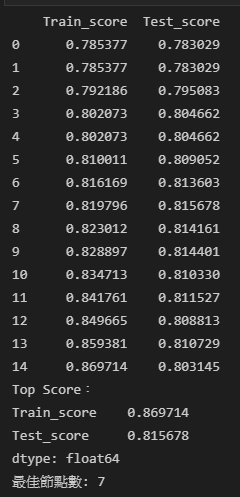
圖2：實驗流程圖

3.2.2薪資調查-部分國家成人資料

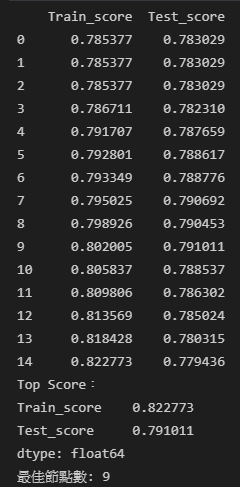
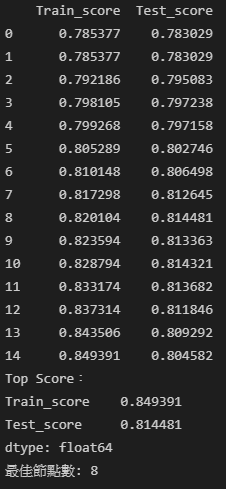
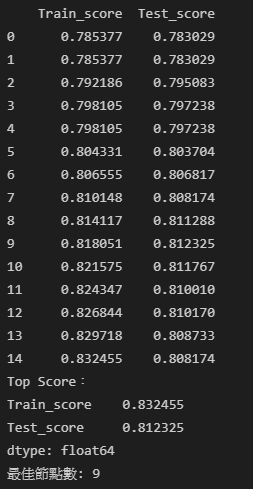
1. 建構數並設定樹身最大為11，分割方法為Entropy。
2. 將欲處理之資料匯入樹中做訓練。
3. 將樹結構視覺化並驗證樹之架構。
4. 產出訓練集及測試集之準確率。
5. 測試多個參數並使用最佳化之參數。
6. 匯出預測結果。
7. **實驗結果**

以下為Beijing PM2.5以及Adult資料集之預測圖表以及決策數圖：

3.4.1 **Beijing PM2.5 Data Set**

實驗一結果 實驗二結果 實驗三結果

實驗四結果 實驗五結果 實驗六結果

由結果可知實驗三雖然績效比最好的實驗二下降了0.005，但樹的結構簡化了非常多，也能由此了解，資料聚合前處理是比較有效的處理方法。

3.4.2 **Adult Data Set**

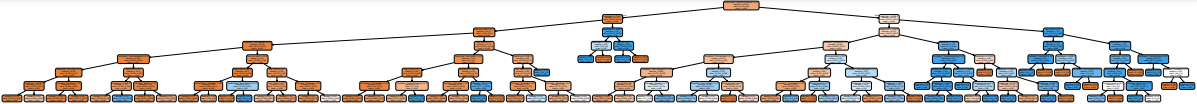


圖5：Adult資料集之決策數圖形

(完整圖表以傳至Github)

1. **結論**

空氣污染儼然是一項重大問題，空氣中細懸浮微粒主要來自人類活動排放造成的，而大氣中PM2.5的濃度主要受自然因素影響較大，其中露點溫度差尤為關鍵，露點溫度差與PM2.5濃度呈現負相關，露點溫度差代表當時空氣中的水氣飽和程度，露點溫度差越大，水氣越不飽和，則PM2.5濃度越低。有影響但影響程度較小的是風速，由觀測資料可知，除去溫度、露點、露點溫度差後累積風速成為最重要的影響因素，風速高，隨後PM2.5濃度略降；風速低，PM2.5濃度則略為升高。綜合上述，影響PM2.5濃度高低主要是露點溫度差。往後如有更深入探討PM2.5的來源與影響，建議可從這面向進行更深入的討論。