資料探勘專題作業二

訓練模型預測數值

指導教授:

許中川 教授

成員:

M11123047 劉穎謙

M11123055 蕭旭朝

日期:

2022年11月22號

摘要

1-1

本研究使用adult資料集，從 1994 年人口普查數據庫中提取數據，來預測一個人的每週工作時數。先使用三種演算法(KNN、SVR、Random Forests)進行績效評估，再使用scikit-learn的feature-importance找出特徵屬性後，透過計算特徵重要性，對欄位特徵進行篩選，最後比較特徵欄位刪減前後績效之差異。

1-2

本研究使用Metro Interstate Traffic Volume資料集，為了探討MN DoT ATR 站 301 的每小時 94 號州際公路西行交通量，及每小時天氣特徵和假期包括對交通量的影響。先使用三種演算法(KNN、SVR、Random Forests)進行績效評估，再使用scikit-learn的feature-importance找出特徵屬性後，透過計算特徵重要性，對欄位特徵進行篩選，最後比較特徵欄位刪減前後績效之差異。

一、緒論

1.1 動機

1.1.1 adult資料集

為了查詢現代人類於工作時數方面落差中所具備特質及專長，從1994 年人口普查數據庫中提取數據，本研究使用機器學習來預測出每週工作時數。

1.1.2 Metro Interstate Traffic Volume資料集

為了探討MN DoT ATR 站 301 的每小時 94 號州際公路西行交通量，及每小時天氣特徵和假期包括對交通量的影響。

1.2目的

1.2.1 adult資料集

本研究針對adult資料集的訓練集與測試集分別使用KNN、SVR、RandomForest三種回歸模型計算MAE、RMSE、MAPE三種績效指標，刪減重要性最高之特徵後，再次進行KNN、SVR、RandomForest的三種績效指標的計算，最後比較刪減前後績效指標的區別，來了解刪減重要欄位後對績效指標的影響。

1.2.2 Metro Interstate Traffic Volume資料集

本研究針對adult資料集的訓練集與測試集分別使用KNN、SVR、RandomForest三種回歸模型計算MAE、RMSE、MAPE三種績效指標，刪減重要性最高之特徵後，再次進行KNN、SVR、RandomForest的三種績效指標的計算，最後比較刪減前後績效指標的區別，來了解刪減重要欄位後對績效指標的影響。

二、方法

2.1 adult資料集

2.2 Metro Interstate Traffic Volume資料集

三、實驗

3.1資料集

3.1.1 adult資料集

|  |  |
| --- | --- |
| Age | 歲數 |
| Workclass | 工人階級 |
| Fnlwgt | 比例 |
| 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 | 國籍 |
| income | 年收入 |

3.1.2 Metro Interstate Traffic Volume資料集

|  |  |
| --- | --- |
| holiday | 節日 |
| temp | 溫度 |
| rain\_1h | 每小時降雨量 |
| snow\_1h | 每小時降雪量 |
| clouds\_all | 雲量 |
| weather\_main | 天氣 |
| weather\_description | 天氣描述 |
| date\_time | 日期與時間 |
| traffic\_volume | 交通流量統計 |

3.2前置處理

3.2.1 adult資料集

因 Workclass 中有遺漏值，而資料中 Workclass 為遺漏值時 Occupation 也會產生遺漏值，所以先將資料集中 Workclass 為遺漏值的資料刪除，再接著將 workclass、marital-status、occupation、relationship、race、gender、native-country 等…類別型的特徵屬性轉為數值型資料。

3.2.2 Metro Interstate Traffic Volume資料集

因本資料集是用完整一年度的時間劃分每小時的交通流量，因此界定好固定的時間級距後，刪除了日期與時間的欄位，只用每小時降雨量、降雪量、雲量來做預測。

3.3實驗設計

3.3.1 adult資料集

針對adult資料集的訓練集與測試集分別使用KNN、SVR、RandomForest三種回歸模型計算MAE、RMSE、MAPE三種績效指標，刪減重要性最高之特徵後，再次進行KNN、SVR、RandomForest的三種績效指標的計算，最後比較刪減前後績效指標的區別，來了解刪減重要欄位後對績效指標的影響。

3.3.2 Metro Interstate Traffic Volume資料集

本研究針對adult資料集的訓練集與測試集分別使用KNN、SVR、RandomForest三種回歸模型計算MAE、RMSE、MAPE三種績效指標，刪減重要性最高之特徵後，再次進行KNN、SVR、Random Forest的三種績效指標的計算，最後比較刪減前後績效指標的區別，來了解刪減重要欄位後對績效指標的影響。

3.4實驗結果

3.4.1 adult資料集

|  |  |  |
| --- | --- | --- |
| KNN | 調整前 | 調整後 |
| MAE | 198.842 | 199.396 |
| RMSE | 14.101 | 13.834 |
| MAPE | 0.284 | 0.281 |

|  |  |  |
| --- | --- | --- |
| SVR | 調整前 | 調整後 |
| MAE | 123.930 | 124.477 |
| RMSE | 11.132 | 11.112 |
| MAPE | 0.272 | 0.271 |

|  |  |  |
| --- | --- | --- |
| Random Forest | 調整前 | 調整後 |
| MAE | 116.947 | 127.191 |
| RMSE | 10.814 | 11.277 |
| MAPE | 0.272 | 0.280 |

3.4.2 Metro Interstate Traffic Volume資料集

|  |  |  |
| --- | --- | --- |
| KNN | 調整前 | 調整後 |
| MAE | 5726472.165 | 5840386.783 |
| RMSE | 2393.004 | 2416.689 |
| MAPE | 2.105 | 2.993 |

|  |  |  |
| --- | --- | --- |
| SVR | 調整前 | 調整後 |
| MAE | 3922843.025 | 3908589.284 |
| RMSE | 1980.616 | 1977.015 |
| MAPE | 2.907 | 2.914 |

|  |  |  |
| --- | --- | --- |
| Random Forest | 調整前 | 調整後 |
| MAE | 4045009.815 | 3767996.171 |
| RMSE | 2011.220 | 1941.132 |
| MAPE | 2.288 | 2.586 |

四、結論

4.1

經過對比刪除特徵屬性後的績效，可以得知三種績效指標再刪除特徵屬性後的值普遍會變大。

4.2

經過對比刪除特徵屬性後的績效，可以得知三種績效指標再刪除特徵屬性後的值普遍會變大。

五、參考文獻

Huang Liz.(Apr 10, 2019). Python學習筆記#14：機器學習之KNN實作篇。

<https://medium.com/@search.psop/python%E5%AD%B8%E7%BF%92%E7%AD%86%E8%A8%98-14-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E4%B9%8Bknn%E5%AF%A6%E4%BD%9C%E7%AF%87-64071fbd0ac8>

桔子菌(2020年11月26日)。Python错误集锦：numpy数组下标方式访问时提示：IndexError: index 5 is out of bounds for axis 0 with size 5。

<http://www.juzicode.com/python-error-numpy-indexerror-index-5-is-out-of-bounds/>

plusone.(Dec 27,2017). Pandas資料的取得與篩選！

<https://ithelp.ithome.com.tw/articles/10194003>