WEI-CHIH CHANG Q36111281
November 22, 2022

Abstract

分析不均勻資料集對不同分類模型的影響。

1. Related Work

本篇報告探討的偽資料主題是「Who is good teaching assistant」，而這次資料最特別的地方在於標籤結果的不平均，每次生成一筆的資料集，會有八成的資料是好助教(label is 1)，剩下兩成資料為不好的助教(label is 0)。如此的設計除了想表達現實中好助教占了大多數外，同時我也想探討在全部都預測特定一個標籤就會有八成的準確率的情況下，四種分類模型將會有什麼樣的分類結果。

2. Data Design

這次的偽資料有 15 種特徵，分別有 6 種數值型資料 (Age, Height, Weight, Sleeping, Research_time, Freq_Sport)以及 9 種類別型資料(ID, Gender, Birthplace, Attitude, Performance, Entertainment, Cost, Changed_Major, Health)，最後的標籤為 1 和 0，分別對應好助教與不好的助教。

這幾筆特徵中最有趣的特徵是 ID，我參照成功大學的學號設計方式，從碩士、博士的 122 個系所中，加入入學年以及流水號，合成偽造的成功大學 ID。我設計該特徵的用意是希望其他使用這能觀察該筆特徵，從中發現 ID 的前兩個字元能對應到不同的系所，進而將系所當成類時的特徵。

下列函式 f(x)是標籤生成的規則，分別使用到 ID, Age, Attitude, Research_time, Entertainment, Freq_Sport, Health, Changed_Major, Birthplace, Cost 等十種特徵，若經過計算後 f(x)大於等於 0，標籤為 1，反之 f(x)小於等於 0，標籤為 0。其中權重最重的特徵為 Age, Freq_Sport, Changed_Major, ID 以及 Birthplace 和 Cost，帶負號的特徵有 ID, Age, Attitude, Research_time, Health, Changed_Major。考慮各項特徵滿足的條件以及出現機率之緣故，因此在設計時我推斷若使用 Decision Tree 分類資料，Age 和 Attitude 將會是重要的節點。

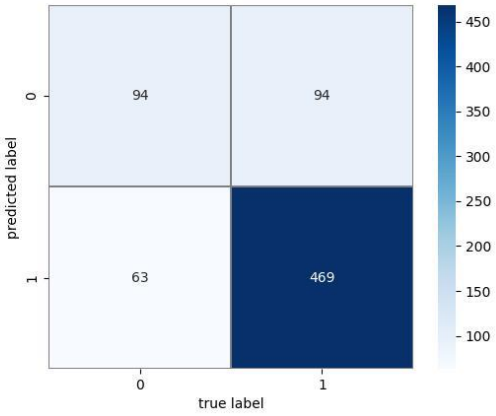
```
f(x) = 2*(Age%3-1) + (Attitude==positive) -
       (Attitude==negative) + 0.5*(Research_time-5)+
       (Entertainment==pokemon_go) +
       (Freq_Sport-2.5)^2 + (Health-2)/5 –
       2*(Changed_Major==1) –
       (Birthplace%7==1)/(Cost/200) + (ID==Q) –
       2*(ID==K2)
```

3. Classification Method

我選擇四種分類器做為這次探討的要素。透過 Decision Tree 我希望能從視覺化的樹中發現和真實規則相對應的結果；希望從 SVM 的預測解果，了解資料在高維空間中是否能適宜地分割；從 KNN 的預測結果分析資料間是否有特定類別的資料會聚集的情形；最後是類神經網路模型，藉此觀察其在不均勻的資料集中會有如何的表現。資料中 label 為 1 的機率為 0.7703，資料中 label 為 0 的機率為 0.2297。

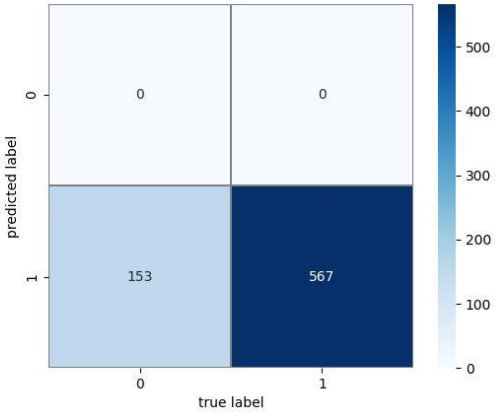
3.1 Decision Tree

建出來的決策樹最深深度為 18，不過令人意外的是，Decision Tree 的第一個節點是 Research_time，因為該特徵權重僅有 0.5，不過後續的幾個節點就如同預期為 Age 和 Attitude。Decision Tree 的準確率為 0.78，比用猜的好一點，預測是壞助教的有一半機率猜錯，但是在預測好助教上則有 0.88 的準確率。



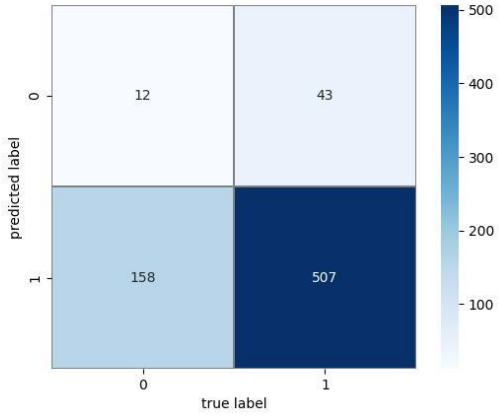
3.2 SVM (Support Vector Machine)

SVM 的預測結果也非常出乎預料，SVM 直接都預測標籤為 1，因此準確率和標籤機率差不多。



3.3 KNN(K Nearest Neighbor)

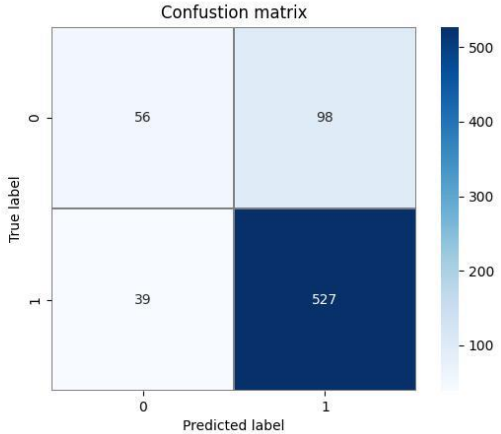
KNN 在預測壞助教上有八成都預測錯誤，在預測好助教上有 0.76 的準確率，該結果說明壞助教的資料在空間中分布並不集中，這才導致 KNN 傾向將標籤都分類為好助教。



3.3 NN(Neural Network)

類神經網路模型在預測壞助教的部分僅有 0.36 的準確率，但在預測好助教上則有 0.93 的準確率。該結果符合類神經網路模型需要大量資料訓練的特性，好助教因為占整體資料的八成，因此類神經網路模型能透過大量的資料去學習分辨是不是好助教，但壞助教的資料因為只佔整體資料的兩成，類神經網路模型還沒從資料中學習如何分辨壞助教時就已經沒有壞助教的資料了。

若是能生成好助教與壞助教數量平均的資料集，預期類神經網路模型能在分辨兩種標籤上都有不錯的表現。



4 Compare right rules with the rules generated by Decision Tree

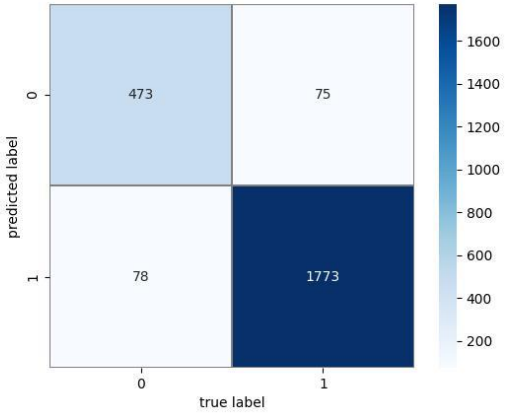
Decision Tree 的第一個節點是 Research_time，但是接續的幾個節點就符合設計時的預期，我們推測是因為雖然這兩個標籤能夠區分標籤，但效果並沒有十分明顯，因此 Decision Tree 才在第一個節點使用 Research_time 作為判斷依據，初步區分資料，該結果也有可能是因為分割訓練集和驗證集所導致。不過在多嘗試幾次後會發現第一個節點並不是固定的，但是接續幾個節點有極高機率是 Age 或是 Attitude。

5. Apply rules generated by Decision Tree in another data

我修改標籤的規則，將 Freq_Sport 那項拿掉，因為那項有平方的因次，因此在規則中大程度的決定了好助教的數量會遠多於壞助教，重新生成資料後好助教和壞助教的機率變為 0.64 和 0.35。

這裡我們使用第一次的資料集訓練 Decision Tree，接著使用該模型去預測修改規則後的資料集，從結果中我們可以發現準確率來到 0.93，其中預測好助教的準確率為 0.96，預測壞助教的準確率為 0.86。從這結果中我們可以發現透過舊資料集訓練出來的模型是可以順利對類似的新資料集進行預測。

可能 Decision Tree 在利用舊資料集訓練時已經大程度的找到除了 Freq_Sport 外如何判斷好壞助教，但因為 Freq_Sport 這項特徵的權重為太大，一直無法順利學習 Freq_Sport 在 label 中的影響。但在使用拿掉 Freq_Sport 後的規則後，就能順利預測標籤。



6. Conclusion

預測不均勻的資料對各種模型都是一件困難的事，除了某一標籤訓練資料不足外，若維持相同機率但是輸入進模型訓練的資料集是經過平衡後的資料，或許會有意想不到的事情發生。