Data Mining

Project2 Classification

孫啟慧

P78063033

一、 實驗設計

本次作業要求生成一組絕對規則（absolutely right rules）並使用其生成 positive data 和 negative data。因為對於生成資料並沒有什麼頭緒，又想到 kaggle 上面有很多開源的資料集，所以我按照興趣選擇了一份關於動物收容所的記錄，他有記錄貓和狗兩種動物的詳細資訊以及最後的去向。通過使用我想出來的絕對規則對其進行資料進行篩選和改造，相應的部分實驗資料 in.csv 如下。

	A	B	C	D	E	F	G
1	AnimalTyp	SexuponOutcome	AgeinDays	SimpleBreed	SimpleColc	OutcomeType	
2	Cat	Intact Male	21	Domestic Shorthair	Blue	Transfer	
3	Cat	Intact Male	21	Domestic Shorthair	Blue	Transfer	
4	Cat	Spayed Female	21	Domestic Shorthair	Brown	Transfer	
5	Dog	Neutered Male	300	Pit Bull	Brown	Adoption	
6	Dog	Neutered Male	2730	Yorkshire Terrier	Black	Return_to_owner	
7	Cat	Neutered Male	90	Domestic Shorthair	Orange	Adoption	
8	Cat	Intact Male	21	Domestic Shorthair	Brown	Transfer	
9	Dog	Intact Female	14	Pit Bull	Brown	Transfer	
10	Cat	Spayed Female	12	Domestic Shorthair	White	Transfer	
11	Dog	Neutered Male	300	German Shepherd	Brown	Adoption	
12	Dog	Neutered Male	1660	German Shepherd	Black	Return_to_owner	
13	Cat	Intact Female	30	Domestic Shorthair	Orange	Transfer	
14	Dog	Neutered Male	1120	Pit Bull	Brown	Return_to_owner	
15	Cat	Intact Male	60	Domestic Shorthair	Blue	Transfer	
16	Cat	Neutered Male	920	Domestic Shorthair	Black	Transfer	
17	Cat	Intact Male	21	Domestic Shorthair	Brown	Transfer	
18	Cat	Spayed Female	2555	Domestic Shorthair	Blue	Adoption	
19	Cat	Intact Male	60	Domestic Shorthair	Orange	Transfer	
20	Cat	Intact Female	60	Domestic Shorthair	Brown	Transfer	
21	Cat	Neutered Male	300	Domestic Shorthair	Black	Adoption	
22	Cat	Spayed Female	3285	Pit Bull	Brown	Return_to_owner	
23	Cat	Intact Female	21	Domestic Shorthair	Black	Transfer	
24	Dog	Spayed Female	2555	German Shepherd	Black	Adoption	
25	Cat	Spayed Female	2555	Domestic Shorthair	Brown	Adoption	
26	Cat	Intact Female	21	Domestic Shorthair	Blue	Transfer	
27	Cat	Spayed Female	2555	Domestic Shorthair	Orange	Adoption	
28	Cat	Intact Male	7	Domestic Shorthair	Black	Transfer	
29	Cat	Intact Male	365	Domestic Shorthair	Orange	Transfer	
30	Dog	Spayed Female	4015	German Shepherd	Brown	Euthanasia	
31	Cat	Intact Female	30	Domestic Shorthair	Black	Transfer	
32	Cat	Spayed Female	14	Domestic Shorthair	Black	Transfer	
33	Cat	Spayed Female	60	Domestic Shorthair	Brown	Transfer	
34	Dog	Neutered Male	730	Pit Bull	Brown	Transfer	
35	Dog	Spayed Female	1095	German Shepherd	White	Adoption	
36	Cat	Spayed Female	2555	Domestic Longhair	Brown	Adoption	
37	Cat	Spayed Female	130	Domestic Shorthair	Black	Transfer	
38	Cat	Spayed Female	5014	German Shepherd	Blue	Euthanasia	
39	Cat	Intact Male	730	Domestic Shorthair	Blue	Transfer	

In.csv 中的資料是按照絕對規則生成資料，主要包括五個

features，也就是 K 為 5，相應的欄位分別為：Animal Types（cat 和 dog 兩種動物）、SexuponOutcome（Intact Male，Intact Female，Spayed Female 和 Neutered Male 四種狀態，分別表示完整雄性，完整雌性，切除卵巢的雌性和被閹割的雄性）、AgeinDays（表示該動物的年齡，以天數來計算）、SimpleBreed（表示該動物的品種，值得注意的是，貓跟狗的品種不同，也就是說當已知品種的時候就可以表示被分析物件是貓還是狗）、SimpleColor（表示該動物的毛色，因為原有資料包含混色，所以我有簡單對其進行處理，選取主要毛色作為其毛色）、OutcomeType（則是表示該動物的結果，包含轉移到其他收容所，被人收養，年紀過大被安樂死以及送還給其原本的主人，這四種）。最後生成的 data（不包括表頭）一共 356 項，也就是 M 為 356。

各項 feature 的值如下：

```
Console D:/NCKU/修课/Data mining/project/data mining project2/
> unique(ssinput$AnimalType)
[1] "Cat" "Dog"
> unique(ssinput$SexuponOutcome)
[1] "Intact Male"   "Spayed Female" "Neutered Male" "Intact Female"
> unique(ssinput$SimpleBreed)
[1] "Domestic Shorthair"   "Pit Bull"             "Yorkshire Terrier"    "German Shepherd"
[5] "Domestic Longhair"    "Domestic Medium Hair" "Siberian Husky"       "Chihuahua Longhair"
[9] "Siamese"
> unique(ssinput$SimpleColor)
[1] "Blue"   "Brown"  "Black"  "Orange" "White"
> unique(ssinput$OutcomeType)
[1] "Transfer"        "Adoption"        "Return_to_owner" "Euthanasia"
> |
```

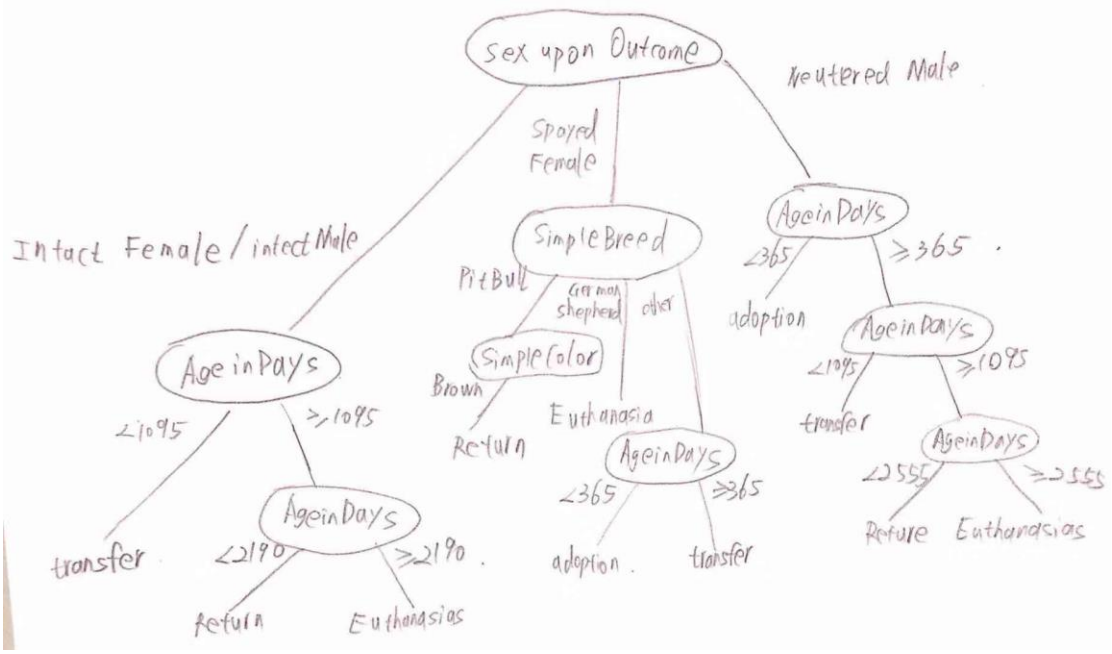
動物的種類一共有九種，而相應的 color 有五種，其他屬性的值上面已經有介紹。因為 AgeinDays 屬性有太多的可能，我就不一一列舉在這邊，具體可以參照 in.csv。

絕對規則主要如下：

1.若該動物為 Intact Male/Female，當其 AgeinDays 小於 1095 天的時候該動物會被 transfer 到其他動物收容所,若在 1095 和 2190 之間的時候該動物會被 return 給原主人，若大於 2190 則會因為年紀過大被安樂死。

2.若該動物為 Spayed Female，則當動物的品種為 pit bull，顏色為 Brown 的時候會被返還給原主人，其品種為 German Shepherd 的時候會被安樂死，對於其它品種的貓和狗，若其小於 365 天則會被 adoption,而大於 365 天的則會被 transfer 去其他的動物收容所。

3.若該動物為 Neutered Male 則當其年紀小於 365 天的時候會被 adoption，結餘 365 和 1095 之間的時候會被 transfer 去其他的收容所，介於 1095 和 255 之間的時候則會被 return 給原主人，當大於 2555 的時候會被安樂死。



二、 對比試驗

為了比較絕對規則和現有的 decision tree 的演算法，我將絕對規則生成的 in.csv 作為 input，使用 R 語言撰寫的決策樹演算法對資料進行建模，主要採用的是 C5.0 和 rpart。相應的代碼如 decision_tree.r 中所示。(因為 R 有撰寫過相應的演算法了，故直接 install 和 library 相應演算法所在 package 即可)

I. C5.0 演算法

C5.0 相較於 C4.5 主要進行了如下的改進：增加了對 Boosting 的支援，它同時也用更少地記憶體。它與 C4.5 演算法相比，它構建了更小地規則集，因此它更加準確。

使用該演算法對絕對規則生成的 data 進行 tree model 建立的詳細結果如下

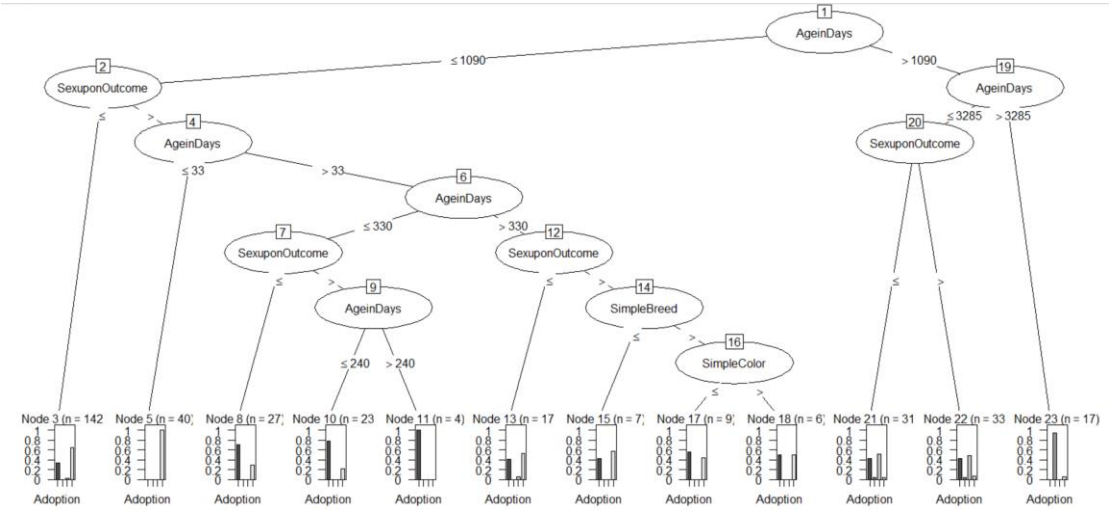
```
Call:
C5.0.formula(formula = OutcomeType ~ ., data = ssinput)


C5.0 [Release 2.07 GPL Edition]        Mon Nov 16 17:05:53 2018
-------------------------------

Class specified by attribute `outcome'

Read 356 cases (6 attributes) from undefined.data

Decision tree:

AgeinDays > 1090:
:...AgeinDays > 3285: Euthanasia (17/1)
:   AgeinDays <= 3285:
:   :...SexuponOutcome in {Intact Male,Neutered Male,
:       :                  Intact Female}: Return_to_owner (36/5)
```

```
:       SexuponOutcome = Spayed Female: Adoption (28/1)
AgeinDays <= 1090:
:...SexuponOutcome in {Intact Male,Intact Female}: Transfer (96)
    SexuponOutcome in {Spayed Female,Neutered Male}:
    :...AgeinDays <= 33: Transfer (27)
        AgeinDays > 33:
        :...AgeinDays <= 330:
            :...SexuponOutcome = Neutered Male: Adoption (77)
            :   SexuponOutcome = Spayed Female:
            :   :...AgeinDays <= 240: Transfer (14/2)
            :       AgeinDays > 240: Adoption (2)
            AgeinDays > 330:
            :...SexuponOutcome = Neutered Male: Transfer (30/1)
                SexuponOutcome = Spayed Female:
                :...SimpleBreed in {Domestic Shorthair,Yorkshire Terrier,
                    :               German Shepherd,Domestic Longhair,
                    :               Domestic Medium Hair,Siberian Husky,
                    :               Chihuahua Longhair,Siamese}: Adoption (21)
                    SimpleBreed = Pit Bull:
                    :...SimpleColor in {Blue,Black,Orange,White}: Adoption (4)
                        SimpleColor = Brown: Return_to_owner (4)


Evaluation on training data (356 cases):

             Decision Tree
           ----------------
           Size      Errors

             12   10( 2.8%)   <<


            (a)   (b)   (c)   (d)    <-classified as
           ----  ----  ----  ----
            131                3    (a): class Adoption
                  16     2          (b): class Euthanasia
              1          35         (c): class Return_to_owner
                    1     3   164    (d): class Transfer


        Attribute usage:

        100.00%    AgeinDays
         95.22%    SexuponOutcome
```

8.15% SimpleBreed
2.25% SimpleColor

Time: 0.0 secs

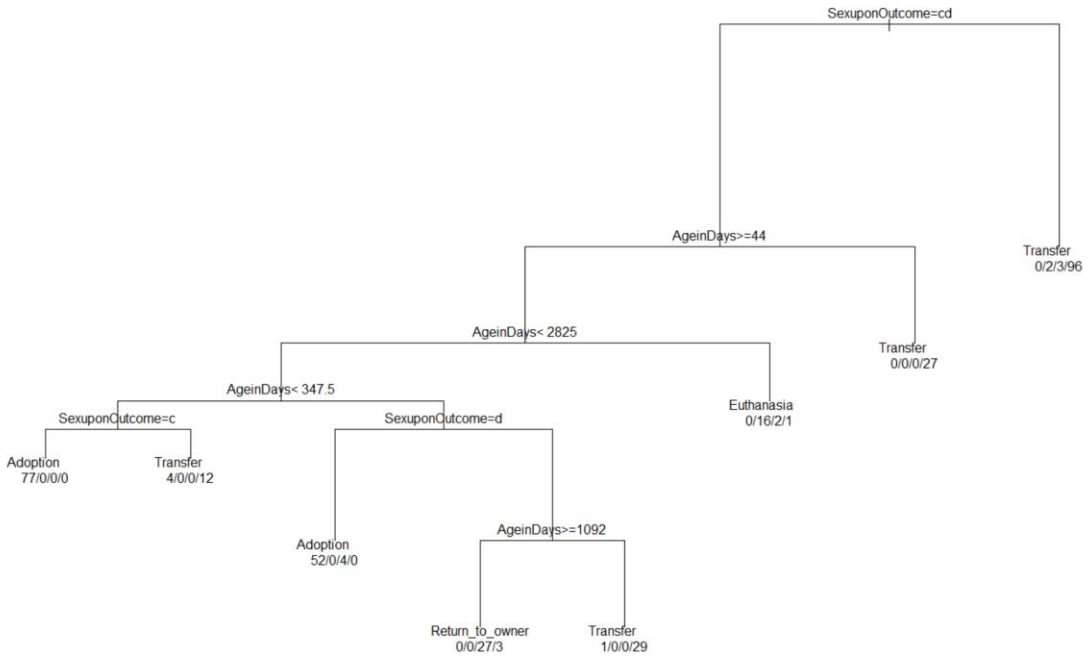
可以看出整個 decision tree 首先用 AgeinDays 來做屬性區分，之後針對於不同的區間再使用 SexuponOutcome，SimpleBreed，SimpleColor 三個屬性來生成最後 decision tree，跟我用來生成 data 的絕對規則相比有所不同，這一點從 AgeinDays 的分割區間就可以看出來了。而且對於毛色還有品種這兩個屬性來看，C5.0 生成的 model 和絕對規則也不一樣。相比較而言既包含了絕對規則，又因為某些屬性的不同分類準則，從而多了很多其他的分類準則。以圖的形式表示則如下，每各葉子節點的圖都表示預測為相應的 outcome 類別的可能性，值越高則可能性越大，因為結果總共有四種所以會有四條表示可能性的柱子，最高的一條表示最後分類的結果（由於圖的 size 的關係，所以部分文字有所缺失）。每一個內部結點都表示相應的分割依據。



CART 分類演算法

CART (Classification and Regression Trees)與 C4.5 演算法是非常相似的，但是 CART 支持預測連續的值（回歸）。CART 用訓練集和交叉驗證集不斷地評估決策樹的性能來修剪決策樹，從而使訓練誤差和測試誤差達到一個很好地平衡點。

下圖是採用 CART 演算法對於絕對規則生成的資料進行訓練從而得到的二叉樹 model，從結果可以看出主要採用了 SexuponOutcome 和 AgeinDays 兩個屬性來進行分類，關於毛色和種類的屬性已經被忽略了，結果的屬性主要是按照可能性來分配，最大可能性的一類就被當做最後的結果。在這裡跟我們的絕對規則相比較之後就會發現，分類的規則相差很多。



分類的詳細資訊如下圖所示，n 表示有多少個結果，而 node), split, n, loss, yval, (yprob)則是其最終的結果的形式，最大的可能性所在的結果類會被判斷成最後的結果。

```
> fit
n= 356

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 356 188 Transfer (0.37640449 0.05056180 0.10112360 0.47191011)
   2) SexuponOutcome=Neutered Male,Spayed Female 255 121 Adoption (0.52549020 0.06274510 0.12941176 0.28235294)
     4) AgeinDays>=44 228  94 Adoption (0.58771930 0.07017544 0.14473684 0.19736842)
       8) AgeinDays< 2825 209  75 Adoption (0.64114833 0.00000000 0.14832536 0.21052632)
        16) AgeinDays< 347.5 93  12 Adoption (0.87096774 0.00000000 0.00000000 0.12903226)
          32) SexuponOutcome=Neutered Male 77   0 Adoption (1.00000000 0.00000000 0.00000000 0.00000000) *
          33) SexuponOutcome=Spayed Female 16   4 Transfer (0.25000000 0.00000000 0.00000000 0.75000000) *
        17) AgeinDays>=347.5 116  63 Adoption (0.45689655 0.00000000 0.26724138 0.27586207)
          34) SexuponOutcome=Spayed Female 56   4 Adoption (0.92857143 0.00000000 0.07142857 0.00000000) *
          35) SexuponOutcome=Neutered Male 60  28 Transfer (0.01666667 0.00000000 0.45000000 0.53333333)
            70) AgeinDays>=1092.5 30   3 Return_to_owner (0.00000000 0.00000000 0.90000000 0.10000000) *
            71) AgeinDays< 1092.5 30   1 Transfer (0.03333333 0.00000000 0.00000000 0.96666667) *
       9) AgeinDays>=2825 19   3 Euthanasia (0.00000000 0.84210526 0.10526316 0.05263158) *
     5) AgeinDays< 44 27   0 Transfer (0.00000000 0.00000000 0.00000000 1.00000000) *
   3) SexuponOutcome=Intact Female,Intact Male 101   5 Transfer (0.00000000 0.01980198 0.02970297 0.95049505) *
```

三、 總結

1.通過比較兩種 decision tree（C5.0 和 CART）對於絕對數據生成的 tree model 可以發現兩種方式得到的分類準則與我生成資料時候所用的絕對準則不一致。

出現上述主要問題的原因在於，我在制定絕對規則的時候只是先按照 SexuponOutcome 的種類開始制定不同屬性情況下的收養結果的影響，之後再觀察其他 feature 的結果對於最後收養結果的影響來設計規則並篩選資料。而眾人熟知的 ID3 是使用 Information Gain 來計算，C4.5 是使用 gain ratio 來計算，C5.0 和 C4.5 類似，只是加入 boosting，而 CART 則是使用 gini index 來劃分。不同的計算方法導致不同的特徵的權重不一樣，根據各個演算法的規定，相應的特徵排序也就不同，而後決策樹演算法得到的分類規則和我

的絕對規則並不一樣讓我知道特徵排序的不同可能造就整個 tree model 完全不一致。

2.由於 C5.0（C5.0 是商業用的，所以其原始程式碼並沒有開源，只能直接進行使用，所以只能知道其官方文檔中公佈的針對 C4.5 的性能比較）和 CART 演算法均是 R 語言的 package 有寫好，所以根據其提供的結果可以看到他們均有被優化過，速度很快，跑我自己準備的 in.csv 只需要大約 0 秒。