主旨：

Undersampling

說明：

在二元分類問題中，如果二個類別比例差異過大，我們稱為資料不平衡，例如：某疾病正常與患病的人數、產線產品正常與異常等。在機器學習建模中，如果資料不平衡將導致演算法失效(演算法直接預測成多數類)，為了避免此問題，在資料前處理中我們會採用 oversampling 以及 undersampling 的技巧，讓二類別的資料平衡。最簡單的 undersampling 為隨機挑選多數類，讓多數類的樣本數減少，但隨機挑選有可能導致重要的資訊遺失，且無法保證模型的穩定性，因此有人提出先利用資料分群後取樣的方式，以降低資訊遺失的可能性。

聚合式階層分群法（agglomerative hierarchical clustering, HAC）為常見的資料分群法之一，聚合式階層分群法剛開始由底層開始架構，每個樣本當成一個 cluster，接著找出最相近的 cluster 進行合併，合併完之後產生新的 cluster，接著再合併最相近的 cluster，一直到全部合併為止，最後可以表示成一個 Tree(如圖 1)。在 cluster 合併的過程中，我們利用 2 個 cluster 之間的距離當作高(height)(如圖 1 中 8 跟 9 的距離為 0.5679)。在 HAC 中，可利用設定不同的高，將樹分成想要的群數，如圖 2 中，將原本的樹分成 3 群，接著在每群中找最小的名稱當作 undersampling 後的樣本，即可達到減少多數類的效果，例如：圖 2 中原本樣本數為 10，取樣後為 3 (取 1, 4, 8)。

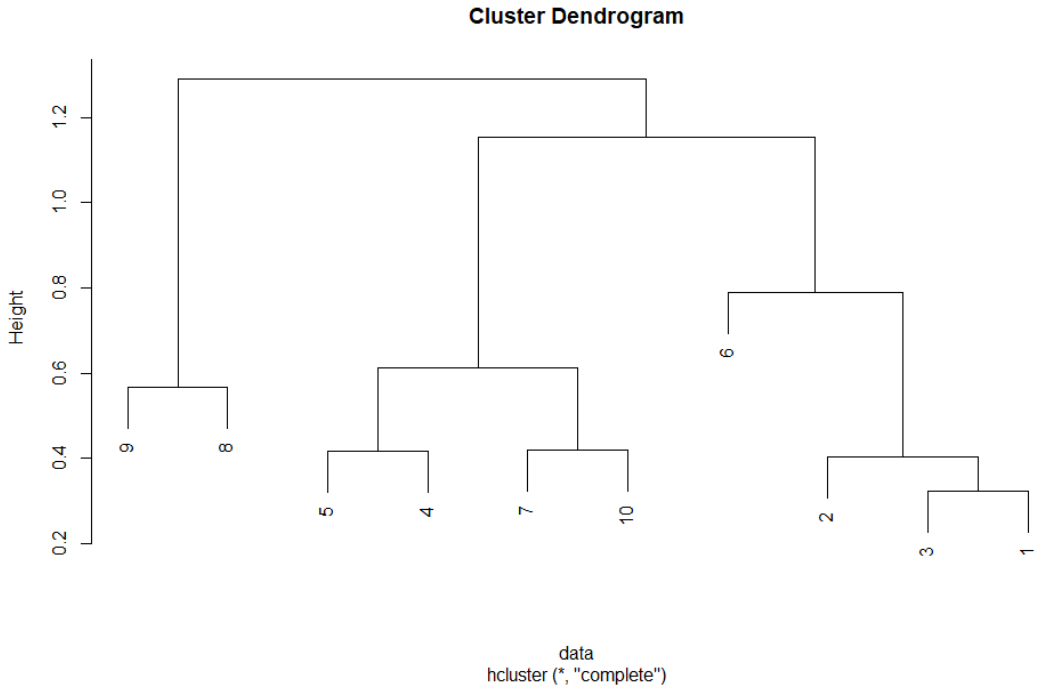

圖 1 HAC 範例圖

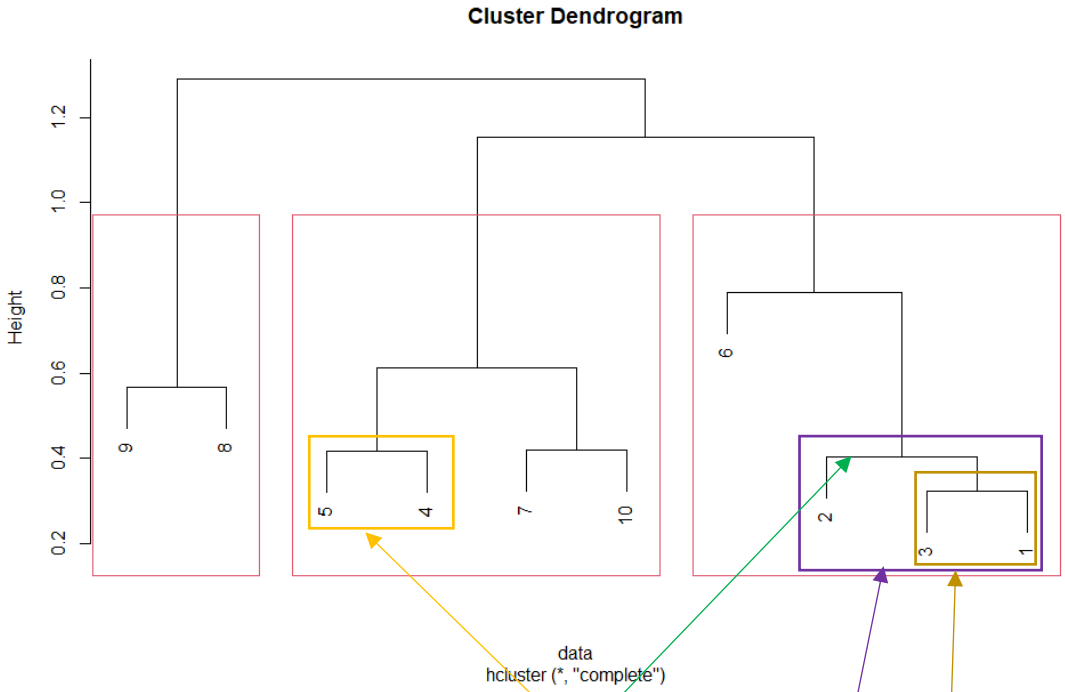

圖 2 原始 tree 分 3 群

我們可以利用三個檔案來表示 HAC，包含 merge、height、label，merge 的內容為 (n-1) x 2 的矩陣，記錄合併資訊，第 i 個 row 代表第 i 次合併的 2 個 cluster 資訊，若數值(假設為 j)<0, 代表合併第-j 個樣本，若數值(假設為 j)>0，代表與第 j 列的結果合併，例如：第 1 列為-8 跟-10，代表要合併第 8 個樣本(編號: 3)與第 10 個樣本(編號:1)(樣本編號參考 labels)，高度為: 0.3220248437620 92(高度參考 height)；第 2 列為-4 跟 1，代表要合併第 4 個樣本(編號 2)與第 1 次合併的結果，高度為: 0.403236903073119；第 3 列為-1 跟-6，代表要合併第 1 個樣本(編號: 5)與第 6 個樣本(編號:4)，高度為: 0.416052881254294，依此類推。

Input (merge.txt):

```
-8,-10
-4,1
-1,-6
-5,-9
-2,-3
3,4
-7,2
6,7
5,8
```

Input (labels.txt)

```
"5"
"9"
"8"
"2"
"7"
"4"
"6"
"3"
"10"
"1"
```

Input (height.txt)

```
0.322024843762092
0.403236903073119
0.416052881254294
0.418449519058154
0.567978872846517
0.613351449007826
0.788542960148653
1.15360305131358
1.28988371568913
```

Output:

```
1 4 8
```

注意事項:

1. 每筆測資有 4 個參數，merge 檔案位置、labels 檔案位置、height 檔案位置
2. 樣本名稱與編號不一致，例如：樣本 1 在編號 10 的位置、樣本 2 在編號 4 的位置
3. 每群選擇樣本名稱最小(數字排序)，樣本名稱均為數字
4. Output 樣本名稱中間用空白分隔
5. 每筆測資分群數均設定為 10 群

截止時間：

2021.12.29 23:59

繳交方式：

批改系統、Portal

作業系統：

Ubuntu 16.04

程式語言：

C or C++ (gcc version 9.4.0)

Command：

./hw3.exe 1_merge.txt 1_labels.txt 1_height.txt

Provide data：

mailto: tinin@saturn.yzu.edu.tw

title:[DSHW3] SID 測資提供

參考資料:

https://medium.com/ai-academy-taiwan/clustering-method-4-ed927a5b4377

https://mropengate.blogspot.com/2015/06/ai-ch17-6-clustering-hierarchical.html

注意事項：

1. 傳值方式

```
int main(int argc,char* argv[])
{
   ifstream fin1, fin2, fin3;
   fin1.open(argv[1]);
   fin2.open(argv[2]);
   fin3.open(argv[3]);
}
```

2. 不要有 system(“pause”);
3. 遲交一天扣 10 分