第五組資料探勘期末報告

吳國成 林鉉閎 許善捷 馮定安
B093040030 B093040032 B093040061 B093040014

一、摘要

現代人口老化及不健康生活型態影響，每年罹癌人數不斷上升。由於早期病徵不明顯，使得許多患者錯過最佳的治療時期，更是提高了癌症的死亡率，長期佔據台灣十大死因之首的位置。
這次實作我們利用RNA序列與癌症的關聯，以分類與分群的方法分析gene expression cancer RNA-Seq data set訓練模型，藉以分別各種癌細胞的特徵，並用於幫助療法的研發。
訓練過程中使用KNN演算法分類已知資料，用K-Means演算法處理未知資料，增進訓練模型的靈活度與效率。

二、流程

三、演算法

- KNN
  計算距離，取離traindata最近的K個testdata進行分類

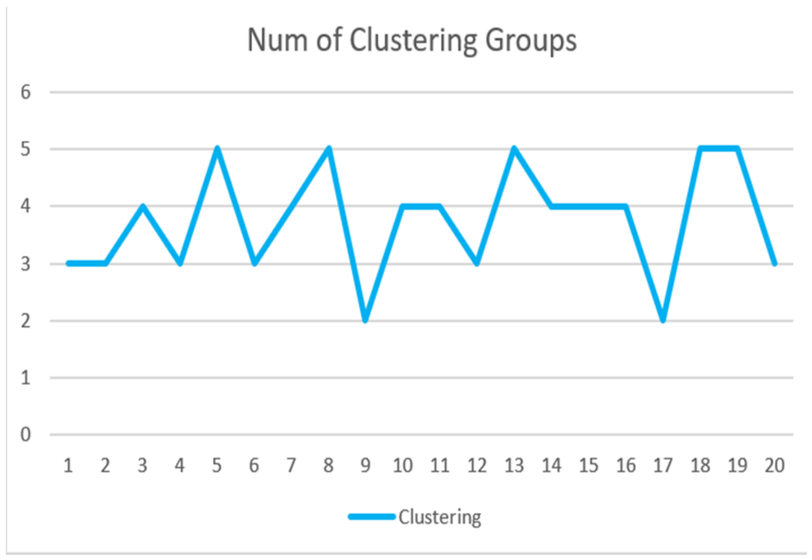
- K-Means Clustering
  K值用來決定要分成幾群。
  隨機取兩筆資料取K個初始形心，兩個形心相距太近或太遠時需重新選取。計算每個資料與形心的距離，劃分至距離最近的形心那群，取每一群所有資料的平均值為新的形心並再次分群直到形心不再變動或分群執行次數達到50次以上。

四、分析結果與總結

k	accuracy	分群次數
1	0.8945783132530121	3
2	0.9006024096385542	3
3	0.9126506024096386	4
4	0.9246987951807228	3
5	0.9186746987951807	5
6	0.9216867469879518	3
7	0.9216867469879518	4
8	0.9216867469879518	5
9	0.9216867469879518	2
10	0.9216867469879518	4
11	0.9216867469879518	4
12	0.9186746987951807	3
13	0.9186746987951807	5
14	0.9246987951807228	4
15	0.9186746987951807	4
16	0.9186746987951807	4
17	0.9186746987951807	2
18	0.9186746987951807	5
19	0.9156626506024096	5
20	0.9126506024096386	3
average	0.9173192771084338	3.75

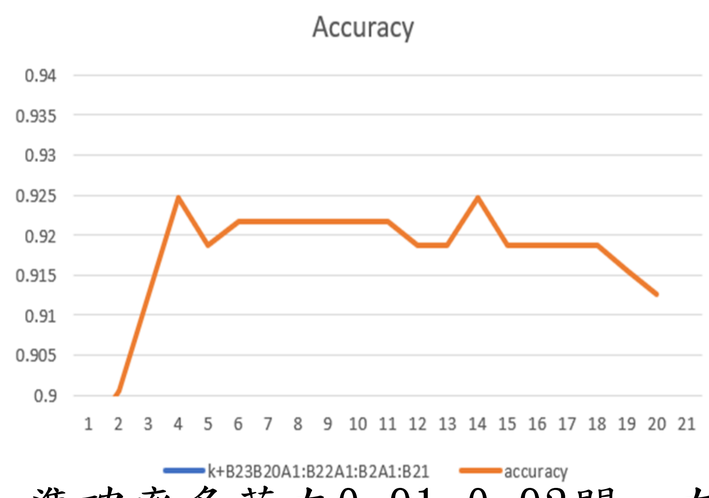
※註：k值為knn的取點個數

Num of Clustering Groups=3.75



分群次數介於2~5之間。

ACCURACY=0,9173192771084338



準確率多落在0.91~0.93間，在k=1及k=2時偏低，並在超過20時有下降的趨勢。

- 此程式使用Python實作knn及k-means演算法來預測罹患癌症腫瘤的類型，首先透過knn將資料分類並歸類出未知的資料，再透過k-means將未知資料分群。
- 可以透過分群的方法來更新已經訓練好的模型並預測新的未知類別資料。以預測罹病種類的例子為例，同時運用分類與分群訓練的模型就能更快速的因應病毒變異造成的病徵變化，並做適當的應對。
- 這次實作在測試knn參數變化時，耗費的時間動輒數小時，如何在保持準確度的同時減少預測所需的時間是未來需要努力的目標。