電影評論網站之資料分析

以 MovieLens 20M 資料集為例

廖信堯
國立政治大學資管所碩一

報 告 大 綱

電影評論網站之資料分析

主題介紹

命題出發點：

假設 MovieLens 是一家公司，其目標是**增加網站流量，進而增加收益**，要如何達成？

- **內容面**：網站內容對使用者來說，有用且有價值的，主要參考**電影評等**
- **功能面**：網站可依據使用者行為，**推薦**符合使用者偏好的電影

命題假設：

1. 電影受歡迎程度可從電影評等高低作為衡量
2. 使用者在 MovieLens 上的評分行為，可以代表使用者的觀影潛在偏好

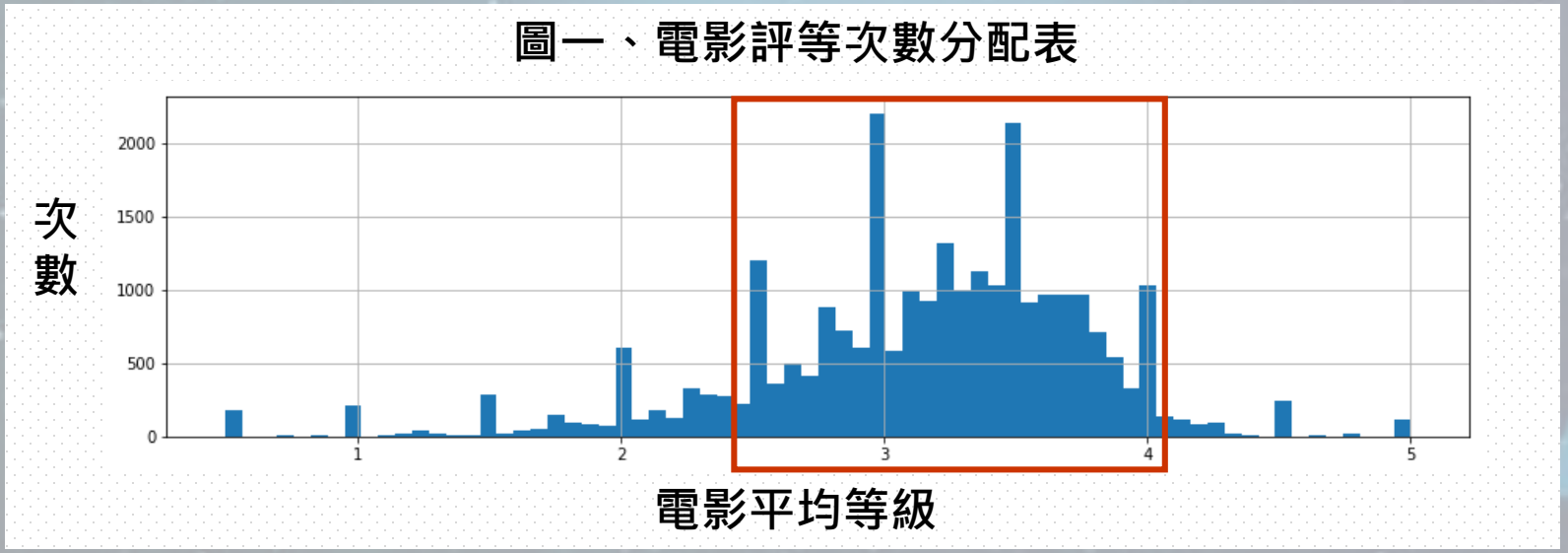
命題：

1. **新電影：如何知道新片受歡迎程度**
2. **舊電影：如何根據使用者行為推薦電影給使用者**

資料探索

資料集簡介：MovieLens 20M 裡，共有13萬使用者對2.7萬部電影的評等與標籤資料

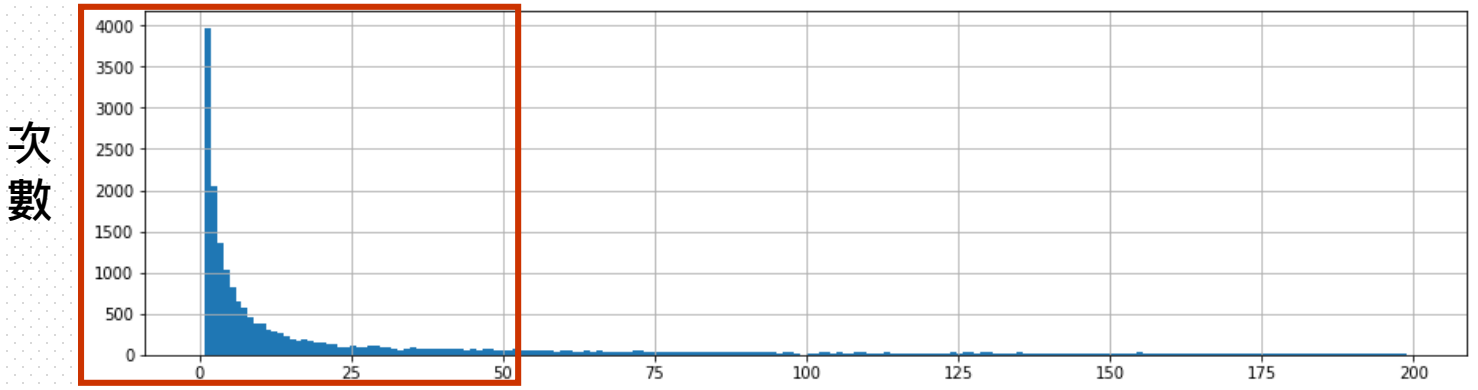
- 平均每部電影約30人給過評等，平均給分約3.5分，平均有約23個標籤
- 資料視覺化呈現如下：



- 大部分的電影評等落在2.5~4分之間，且特別集中於3~3.5分
- 表示電影評等與大眾口味接近，極好/差片較少，介於中間的片佔多數

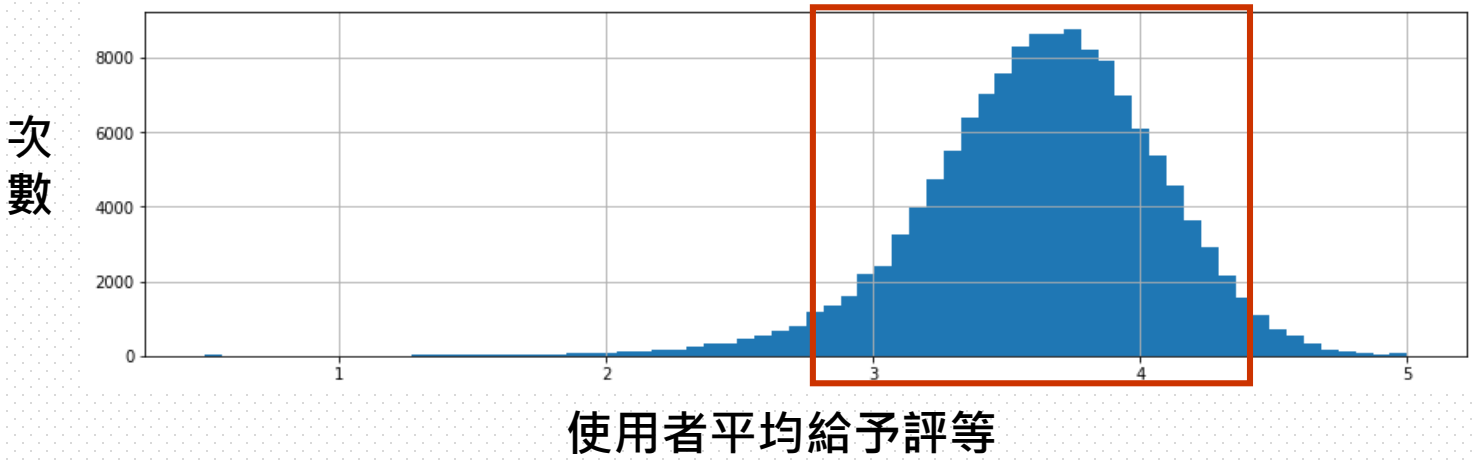
資 料 探 索 cont.

圖二、電影評等人數次數分配表



- 大部分的電影評等人數<50人
- 表示僅有少數片(如：全面啟動)會得到大眾關注
- 大部分電影關注度較低

圖三、使用者平均給予評等分配表



- 使用者平均給分約3.5分
- 整體分布接近鐘形常態分佈，介於2~5分間

資 料 探 索 cont.

表一、五大熱門電影

片名	使用者評等次數
Pulp Fiction (1994)	67310
Forrest Gump (1994)	66172
Shawshank Redemption, The (1994)	63366
Silence of the Lambs, The (1991)	63299
Jurassic Park (1993)	59715

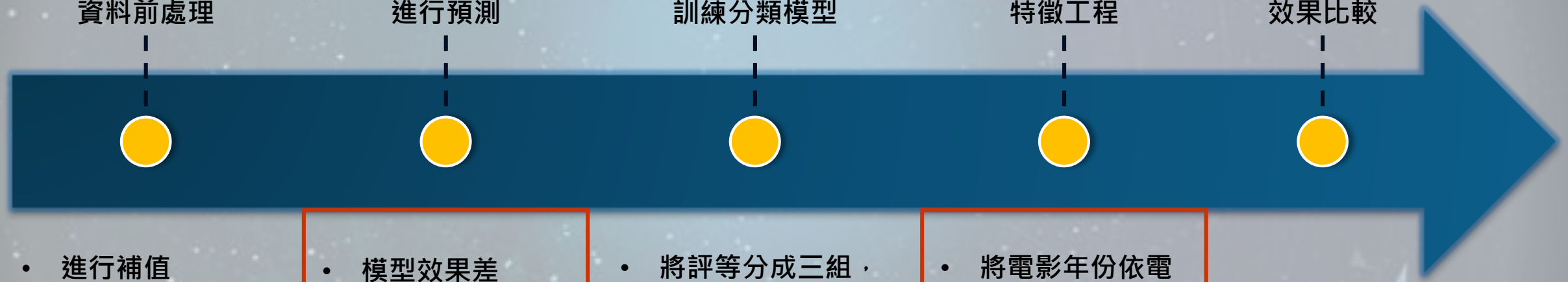
表二、五大熱門標籤

標籤名	使用者使用次數
sci-fi	3384
based on a book	3281
atmospheric	2917
comedy	2779
action	2657

資料分析 - 命題一

命題一：如何知道新片受歡迎程度 → 預測問題

每個人的1~5分不一樣
改為不推薦/推薦/極推薦



- 進行補值
- 萃取年份欄位
- 將電影類別做one-hot encoding處理

- 模型效果差 R-Square僅0.11
- 因為僅有2個特徵，年份與類別

效果不佳，修改方向

- 將評等分成三組，修改成分類問題
- 訓練多個模型，做集成式學習
- 隨機森林模型
- 梯度決策樹模型

- 將電影年份依電影發展史，分成三組
- 將電影類別做聚類，分成四組
- 依照電影類別，萃取Top5標籤資訊

增加特徵
期望提升分類效能

資料分析 - 命題一 cont.

圖四、進行特徵工程前 模型結果

```
The Accuracy of EnsembleLearning:  0.5971182634730539
The Accuracy of EnsembleLearning:  0.544578538684629
              precision    recall  f1-score   support

           0       0.16      0.33      0.21       818
           1       0.92      0.56      0.70      9817
           2       0.01      0.41      0.01        54

   micro avg       0.54      0.54      0.54     10689
   macro avg       0.36      0.43      0.31     10689
weighted avg       0.86      0.54      0.66     10689

------------------------------------------------------------
The Accuracy of GradientBoostingClassifier:  0.565993013972055
The Accuracy of GradientBoostingClassifier:  0.561324726354195
              precision    recall  f1-score   support

           0       0.07      0.48      0.12       251
           1       0.94      0.57      0.71      9882
           2       0.09      0.47      0.15       556

   micro avg       0.56      0.56      0.56     10689
   macro avg       0.37      0.51      0.33     10689
weighted avg       0.87      0.56      0.67     10689
```

圖五、進行特徵工程後 模型結果

```
The Accuracy of EnsembleLearning:  0.5799206206025618
The Accuracy of EnsembleLearning:  0.5511363636363636
              precision    recall  f1-score   support

           0       0.06      0.32      0.09       145
           1       0.90      0.56      0.69      6516
           2       0.15      0.51      0.23       731

   micro avg       0.55      0.55      0.55      7392
   macro avg       0.37      0.46      0.34      7392
weighted avg       0.81      0.55      0.63      7392

------------------------------------------------------------
The Accuracy of GradientBoostingClassifier:  0.5805520476276
The Accuracy of GradientBoostingClassifier:  0.5557359307359
              precision    recall  f1-score   support

           0       0.07      0.41      0.12       136
           1       0.88      0.57      0.69      6260
           2       0.20      0.51      0.29       996

   micro avg       0.56      0.56      0.56      7392
   macro avg       0.38      0.50      0.36      7392
weighted avg       0.77      0.56      0.62      7392
```

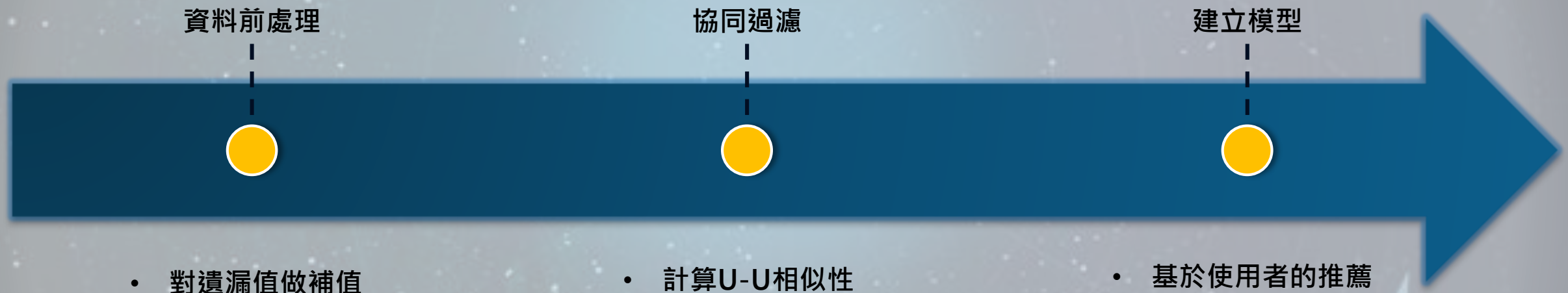
資料分析 - 命題一 cont.

進行特徵工程後，效果差異不大，可能原因：

1. **電影類型聚類方式可改進：**
   - 依據電影評等三類，計算各類別下的不推薦率，不推薦率接近的為一類，以不推薦率作為聚類依據，還可再改進
2. **電影類型 Top5 標籤與評等關係：**
   - 依照類別聚類結果，找出類別與標籤之關係，預測測試資料中的可能標籤，作為特徵
   - 可能標籤與評等的關係沒有想像中有關連
3. **樣本集中率問題：**
   - 訓練樣本大多集中在”推薦”類，導致訓練過程中比較難訓練到其他兩類的部分

資料分析 - 命題二

命題二：如何根據使用者行為推薦電影給使用者 → 推薦問題



資料分析 - 命題二 cont.

表三、基於物品相似性的推薦結果

排名	全面啟動 推薦Top 10
Top1	Watchmen (2009)
Top2	Super 8 (2011)
Top3	Strange Days (1995)
Top4	Contagion (2011)
Top5	Soylent Green (1973)
Top6	Donnie Darko (2001)
Top7	Forgotten, The (2004)
Top8	Jacket, The (2005)
Top9	One Point O (2004)
Top10	Prestige, The (2006)

圖六、基於U-U & V-V 協同過濾 預測之電影評等結果

```
Testing User-based CF RMSE: 31240.305881320917
Testing Item-based CF RMSE: 36390.790744219405
Training User-based CF RMSE: 19378.364232899054
Training Item-based CF RMSE: 153.8704496131768
```

- 訓練與驗證樣本RMSE落差大
- 可能有過度擬合
- 整體預測效果差

資料分析 - 命題二 cont.

陽春版推薦系統改良方向：

1. **推薦效果未進行驗證：**
   - 可用隨機化抽取使用者對電影評分結果進行遮蔽，將遮蔽部分當作答案，來驗證推薦系統的推薦覆蓋性與準確性
2. **輸入的特徵過少：**
   - 僅有評等的資訊，若將電影類型、Top5標籤等納入矩陣，或許可優化推薦效果
3. **改用類神經網絡進行Embedding或是降維：**
   - 由於使用者與電影數量眾多，造成相似性矩陣龐大，若能透過類神經網絡，進行Embedding動作，並用類神經網絡之多層架構進行預測，或許也可優化推薦效果

結論與討論

結論：

在這個資料集下，可使用的特徵較少，儘管已透過特徵工程產生潛在特徵，但在預測電影評等以及推薦使用者可能喜歡的電影上，效果仍有限

可能改進方向：

1. 時間足夠下，可從其他電影評論網站爬文，抓取電影的特徵，例如：導演、演員、電影評等、國家...等，當作預測時有用的特徵
2. 若使用深度學習網絡(DNN)，處理資料維度高的情形，可能可優化預測及推薦效果

討論：

1. **如何預測歌曲評等與推薦歌曲給使用者，來優化客戶體驗，想必是KKBOX的目標**
2. **從資料面來看，有限的使用者資訊、稀疏矩陣、推薦多樣性與精確性的兩難，仍需努力克服**

報告結束

謝謝您們的耐心閱讀

廖信堯
國立政治大學資管所碩一