資工系實務專題研究計畫成果報告

(全民球探)

專題編號：108-CSIE-S013
執行期限：107 年第 1 學期至 108 年第 1 學期
指導教授：劉傳銘
專題參與人員： 105590026 黃彥穎
105590028 鄭宇翔
105590030 陳哲葦

一、摘要

本次專題實作主題為建構一個基於資料分析的球探系統，以 Vue.js 作為網頁前端框架，並使用 Node.js 系統實做後端服務，我們使用 Requests 將資料從網站[1]上爬取下來，再將資料利用 Pandas 做基礎資料前處理後儲存於 Firestore，使用 v-charts 將資料以視覺化方式呈現於網頁上，並 scikit-learn 進行隨機森林[2]機器學習，預測球隊排名及最有價值球員(MVP)。於機器學習訓練過程中，為提升 R^2決定系數[3]需不斷調整特徵及模型參數，以尋找出最佳模型，最後則是使用 K-fold 進行最後驗證，以確保模型的精準度並減少過度擬合的問題。

關鍵詞：
資料視覺化、機器學習、隨機森林

二、緣由與目的

對於剛開始接觸 NBA 的球迷，往往想要更進一步了解籃球的資訊時，會因為一堆文字及數據而退卻，因此我們希望能夠透過數據視覺化的方式，提供球迷更加友善的管道了解自己所喜愛的球隊及球員。近年機器學習蓬勃發展，於體育界也有許多案例，最著名的就是「魔球理論」，透過數據的預測及分析球員發展，因此我們也透過機器學習，來預測 MVP 及球隊排名，以提供有興趣的球迷當作參考。

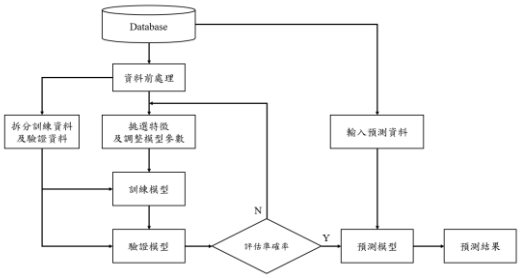

圖 1 機器學習流程圖

三、研究報告內容

(一) 資料視覺化

利用 v-charts 根據球員不同的數據來選擇不同類型的圖表，讓數據以最合適的方式呈現，能夠一目瞭然。

(二) 資料前處理

在機器學習前，需要將資料做清理，去除不需要的資料及干擾，再進行缺失資料處理。做完資料前處理後，會將資料分成訓練資料集及驗證資料集，以提供機器學習使用。

(三) 機器學習

本次專題我們選用隨機森林演算法，選擇原因有二，一是因為其較適合高維度的資料，二是因為此演算法選用訓練的特徵彼此相關性不能過高，而這也符合我們的需求。在訓練的過程中，我們需不斷的使用 GridSearchCV 調整參數及挑選特徵[4]，以尋找最佳模型(如圖 1)，最後使用 K-fold 分成 K 個賽季進行驗證，以驗證此模型的精準度(如圖 2)。

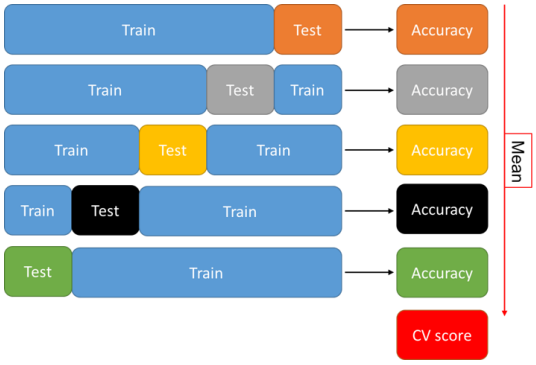

圖 2 K-fold 示意圖

(四) 系統架構

「NBS」是一個以資料分析為主軸的球探系統，以 Vue.js 做為網頁的前端，Node.js 為後端的網頁服務，並使用 Cloud Firestore 做為我們的資料庫，透過 Web crawler 來抓取我們所需的相關資料，再利用 python 進行資料前處理，最後將資料放入資料庫中，以提供網頁呈現視覺化資料及機器學習所需的訓練數據。(如圖 3)。

(五) 實驗結果

本次專題的實驗主題為資料視覺化及機器學習，資料視覺化方面，透過揣摩使用者的體驗來設計，我們認為與原本 NBA 網站，閱讀上已經改善很多，符合我們的初衷，設計給剛接觸 NBA 的人使用。在機器學習方面，我們運用隨機森林演算法來預測 MVP 排名及球隊排名，因兩者預測的方式都是以排名的方式進行預測，因此在準確率上雖不甚高，但結果仍有一定的參考價值，兩者的模型精確度都在 6 成左右(如表 1)，還有待改善的空間。

賽季	名次	預測排名	實際排名
2018-19	1	Giannis Antetokounmpo	Giannis Antetokounmpo
	2	James Harden	James Harden
	3	Trevon Duval	Paul George
2017-18	1	James Harden	James Harden
	2	LeBron James	LeBron James
	3	Russell Westbrook	Anthony Davis
2016-17	1	Russell Westbrook	Russell Westbrook
	2	James Harden	James Harden
	3	Kawhi Leonard	Kawhi Leonard
預測精準度(R^2 計算)：60.2%			
K-fold Cross validation：61.2% +/- 18.3%			

表 1 MVP 預測結果

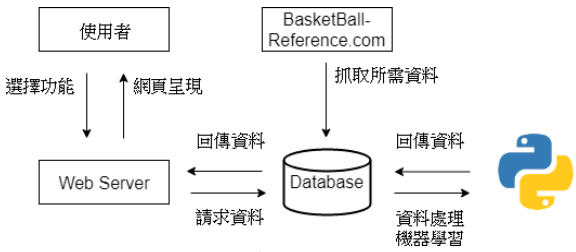

圖 3 系統架構圖

四、結論

於本次專題中，準確率未如預期的高，訓練期間我們發現除了調整挑選特徵及模型參數外，最重要的是特徵與目標之間的關係，除了挑選與目標關係度高的特徵外，也要挑選有意義的特徵，而不是一味挑選高相關度特徵即可。而另一問題則是訓練資料的選擇，在現今籃球球風與以前相差許多，若是選擇較早期的數據可能會影響模型，因此在預測模型中我們特地選擇近期的數據進行訓練。

雖然本次專題預測之數據擁有一定準確率，但仍有可進步的空間，在未來我們可透過蒐集更多資料、調整特徵及模型參數等方式，或是選擇其他演算法，來進行嘗試以提升預測的精準度。

五、參考文獻

[1] Basketball-Reference. Retrieved from https://www.basketball-reference.com/ (October 10, 2019)

[2] Random forest. Retrieved from https://en.wikipedia.org/wiki/Random_forest (October 10, 2019)

[3] 決定系數 from https://zh.wikipedia.org/wiki/%E5%86%B3%E5%AE%9A%E7%B3%BB%E6%95%B0 (October 13, 2019)

[4] Hyperparameter Tuning the Random Forest in Python. (Jan 10, 2018) Retrieved from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74