臺東大學綠色科學學刊 2021，11(2)，pp.97-108

基於深度強化學習於混合式動態電影推薦演算法研究

張耀中[1]、陳世曄[1]、賴槿峰[2]、張淙垣[2]、賴盈勳[1]*

1. 國立臺東大學資訊工程學系 2. 國立成功大學工程科學所

摘要

隨著 5G 網路與行動裝置普及化影響，有越來越多新型態多媒體服務改變現有生活模式，包含線上音樂、串流影片、社交網路、電子商務等都出現不斷產生新型態娛樂互動模式。包含觀看時間長短、觀看類型喜好甚至是影片類型都隨著社群網路與行動模式而不斷出現改變，這樣娛樂模式的蓬勃發展也使得推薦系統能發揮的空間大增。本研究提出了一種新型態的推薦演算法模型，以深度強化學習做為基礎，定義其架構及相關參數，根據不同使用者點擊不同電影並產生的評分，進行對使用者的個性化推薦，搭配線上模擬環境進行預訓練，且針對在大量的電影系統下的問題進行改良，期望提升效能並做出良好的推薦。

實驗結果中，在大量動作下與傳統的 DQN 演算法比較顯示本研究方法能在早期收斂且提升 20 倍以上的速度，與 user based、item based 的 knn 以及 Bayesian Personalized Ranking 比較顯示，在多項指標中表現較佳且能動態更新模型，此方法即使在系統完全沒見過的使用者下也能給出不錯的推薦，在驗證和測試資料集中都沒有過度擬合的問題，證明此模型的通用性。

關鍵字: 推薦系統、深度強化學習、混合模型

張耀中，國立臺東大學資訊工程學系教授。E-mail: ycc@nttu.edu.tw
陳世曄，國立臺東大學資訊工程學系副教授。E-mail: sychen@nttu.edu.tw
賴槿峰，國立成功大學工程科學所教授。E-mail: cinfon@ieee.org
張淙垣，國立成功大學工程科學所碩士。E-mail: jungle0717@gmail.com
賴盈勳 (通訊作者)，國立臺東大學資訊工程學系副教授。E-mail: yhlai@nttu.edu.tw

Dynamic Movie Recommendation Algorithm Based on Deep Reinforcement Learning

Yao-Chung Chang, Shih-Yeh Chen, Chin-Feng Lai, Tsung-Yuan Chang, and Ying-Hsun Lai*

1. Department of Computer Science and Information Engineering, National Taitung University 2. Department of Engineering Science, National Cheng Kung University

Abstract

With the popularization of 5G networks and mobile devices, the current life mode have been changed by more and more new types of multimedia services. There are constantly emerging new interactive entertainment modes including online music, streaming videos, social networks, e-commerce, etc. This research proposes a new type of recommendation algorithm model and defines architecture and related parameters based on deep reinforcement learning. This personalized recommendation is based on the ratings generated by different users clicking on different movies. This research mainly improved the problems in a large number of movie systems and pre-trained with an online simulation environment to improve performance and make good recommendations.

In the experimental results, the comparison with the traditional DQN algorithm under a large number of actions shows that this research method can converge in the early stage and increase the speed by more than 20 times. Compared with user based, item based knn and Bayesian Personalized Ranking, it shows better performance in multiple indicators and can dynamically update the model. This proposed method can give good recommendations even for users who have never seen the system. There is no over-fitting problem in the validation and test data sets, which proves the generality of this model.

Keywords: Recommendation system, Deep Reinforcement Learning, hHbrid model

Yao-Chung Chang, Professor, Department of Computer Science and Information Engineering, National Taitung University, E-mail: ycc@nttu.edu.tw

Shih-Yeh Chen, Associate Professor, Department of Computer Science and Information Engineering, National Taitung University, E-mail: sychen@nttu.edu.tw

Chin-Feng Lai, Professor, Department of Engineering Science, National Cheng Kung University, E-mail: cinfon@ieee.org

Tsung-Yuan Chang, Master, Department of Engineering Science, National Cheng Kung University, E-mail: jungle0717@gmail.com

Ying-Hsun Lai (Corresponding author), Associate Professor, Department of Computer Science and Information Engineering, National Taitung University, E-mail: yhlai@nttu.edu.tw

壹、前言

近年來因為資訊科技與網路的迅速發展改變了人們的生活習慣,線上音樂、串流影片、社交網路、電子商務等蓬勃發展使得推薦系統能發揮的空間大增,生活中隨處可見其蹤影,而這些平台也是近年來研究者研究推薦系統的重要實驗場景，搭配人工智慧近年來的興起，將其結合在此應用的想法也開始出現，The ACM Recommender Systems conference (RecSys)會議也自2016年起開始舉辦定期的推薦系統深度學習研討會，旨在促進研究包括算法設計、系統實現和系統評估等，並且鼓勵基於深度學習的推薦系統的應用，而深度神經網路和推薦系統的結合將會是推薦系統未來的主要研究方向，在未來發展上隨著使用者評價與商品資料量大量增長，計算時間也勢必大幅拉長，推薦系統除了能良好的的推薦適合的商品外，如何降低運算級反應時間，也成為了推薦系統的重要議題。

近年來受惠於計算效能的提升，許多機器學習以及深度學習的演算法得以更加普及的實作，像是卷積神經網路(Convolutional Neural Network, CNN)可以用在各種影像辨識系統上(Krizhevsky et al., 2012)，長短期記憶模型(Long Short-Term Memory, LSTM)可以用在與時間序列有關的語音辨識、文章分析等(Hochreiter & Schmidhuber, 1997)，以及這幾年重新竄起的強化學習(Reinforcement Learning, RL)領域，其中強化學習能應用在不同領域且存在不少發展的可能性，因此相關研究快速興起，最著名的研究像是Alpha GO、語音助理(Voice assistant)等，其中與大眾生活息息相關的各式種類推薦系統(Recommender System)也獲得許多研究者的關注與發展。

然而推薦系統不是萬能的，以下為推薦系統常面臨的幾個問題，這些問題也成為了後續研究人員的改善重點，包含

- 數據稀疏問題：由於產品和使用者數量都很龐大，使用者評分產品佔產品總數的比例非常的少，有其他使用者共同評分的產品更是少之又少，若能解決此問題是提高推薦質量的關鍵。
- 冷啟動問題：如何在新系統沒有大量使用者歷史紀錄的情況下，設計個性化推薦系統並且讓使用者在早期就能對推薦結果滿意。
- 健壯性問題：推薦系統可能存在被惡意使用者的攻擊而影響推薦的執行結果，讓推薦清單經常或者很少出現某類商品，甚至破壞系統使其無法產生有效推薦。
- 反應速度問題:如何在資訊爆炸的時代高效且快速的處理大量的數據並實時的推薦商品。
- 可擴展性問題：通常在實際的推薦系統中不僅數據量大，而且新使用者新產品會不斷進入系統,使用者和商品間會不停的產生新的評分,使得系統中的數據處於動態變化中,既使能在小規模數據集上的離線測試表現良好的算法，在實際的大規模數據集上不見得奏效。

本研究目標在於設計出一種混合式推薦演算法，能根據系統現在所面臨的不同使用者與商品,學習出不同使用者可能的喜好關係並且找出潛在的其他興趣,進而預測出系統推薦良好的個性化商品,並預期可以提升大量電影下的系統效能。本研究採用近年來熱門的技術,使用強化學習並搭配深度學習的方法,將推薦系統的問題抽象化到強化學習領域,透過定義相關的動作(Action)、狀態(State)、價值函數(Value Function)、策略(Policy) 以及獎勵(Reward)等參數來實踐,且推薦系統也符合強化學習中不斷有新的使用者反饋的特性,在一次次錯誤中學習出更加適合的答案,達到良好的動態推薦效果。主要研究目標為以下幾點:

- 引進強化學習將其應用於推薦演算法：此部分主要著眼於未來推薦是否有能獲得良好的評分，可學習出每次每個使用者評分的隱藏特徵，透過過去的經驗來增加未來的準確率。
- 定義方法能在強化學習大量動作空間下收斂：此部分主要著眼於推薦系統通常伴隨著大量的使用者及電影，這對於強化學習會一個相當耗時且難以收斂的議題。
- 使用SVD模型作為線上模擬環境：在線下訓練中使用SVD模擬線上環境，解決數據稀疏(Data Sparsity)與冷啟動(Cold Start)問題，並期望能學習偕同過濾之特性。
- 推薦從未見過的使用者：利用協同過濾可以透過他人的經驗找出可能的興趣，但傳統model-based的限制，無法對不在原資料集中的使用者進行推薦。

貳、文獻探討

一、推薦系統介紹

自 1997 年 Resnick & Varian (1997) 首次提出推薦系統(Recommender System)一詞，並給出以下定義：「使用者提供推薦的項目作為系統的輸入，系統收集這些資訊並加以處理，再提供合適的項目給使用者。」從此之後推薦系統一詞被廣泛使用，推薦演算法也開始成為一個重要的研究領域，推薦系統也可以說是一種訊息過濾系統，其目的是從過去的歷史資料過濾出有用的特徵，用來預測使用者對未知商品的評分或偏好，一旦可以估計使用者對未知商品的評分，就可以找出分數最高的商品推薦給使用者，其應用的場景與我們的生活息息相關包括電影、音樂、新聞、書籍、學術論文以及各種產品，一些常用的平台像是社群媒體、搜尋引擎、影音及網路購物平台等，甚至是平台上投放的廣告內容都可以進行推薦，利用使用者產生的資訊創造價值，著名的案例像是 Netflix 自 2006 年 10 月開始舉辦電影推薦競賽獎並提供 100 萬美金的獎金，也帶動了推薦系統的研究風潮，是相當熱門的研究領域。

在過去 20 年中推薦演算法也發展出許多種不同的主流技術，根據推薦方式可分類為基於內容、協同過濾、基於關聯規則與基於知識推薦等。基於內容推薦演算法(Content-Based Recommendation)，是一種工業界應用比較廣的推薦演算法，不需要太多的使用者評分或者群體記錄，可以根據物品的特性和使用者的特殊偏好等特徵屬性進行較直觀的推薦(Pazzani & Billsus, 2007)。協同過濾推薦演算法(Collaborative Filtering Recommendation)，概念為利用某興趣相投且擁有共同經驗之群體的喜好，來推薦使用者感興趣的資訊，透過個人的合作機制給予資訊相當程度的反饋像是評分或給標籤，並記錄下來以達到過濾的目的進而幫助別人篩選資訊 (Sarwar et al., 2001; Herlocker et al., 2000)。協同過濾技術又分為兩種，以記憶為基礎(Memory-Based)及以模型為基礎(Model-Based)(Gong et al., 2009; Zahir et al., 2019)。以記憶為基礎這種方法需要計算表格中每一格的相似度，這種窮舉法會耗費大量計算資源，使用最近鄰搜尋(Nearest Neighbor Search)，根據資料的不同可以選擇不同的相似度演算法，像是餘弦相似性(Cosine-based Similarity)、Pearson Correlation Coefficient。以模型為基礎的協同過濾是先用歷史資料得到一個模型，再用此模型進行預測。可以使用機器學習與深度學習的演算法，直接找出機率最大的推薦商品，在使用者量和商品數很大的情況時，可以先進行離線訓練模型，達到後續推薦的實時性，可以改善以記憶為基礎的缺點，難以即時處理大資料量或在資料稀疏問題下影響(Jiang et al., 2015; Aggarwal, 2016)。另一種是利用矩陣分解(Matrix Factorization)產生模型，最有名的方法為奇異值分解(Singular Value Decomposition, SVD)，此方法在 Netflix 電影推薦的競賽中獲得了成功(Paterek, 2007)。

基於關聯規則推薦演算法(Association Rules-Based Recommendation)，從大量使用者行為數據中發現有強關聯的規則，找出那些同時被很多不同使用者購買的物品集合，從這些集合內的物品可以相互進行推薦，是一種無監督的機器學習方法，最常見的關聯法則演算法為 Apriori (Agrawal & Srikant, 1994) 和效能更佳的 FP-Growth (Han et al., 2000)及其延伸算法，最著名的例子為"啤酒與尿布"，已在零售業賣場中得到了成功的應用。基於知識的推薦演算法(Knowledge-Based Recommendation)，高度重視知識源，推薦的物品不是建立在使用者需要和喜好上，需要將領域專家的知識整理成為規範且可用的表達形式，可能需要主動的詢問使用者的需求(Burke, 2000)。例如購買相機時依使用者主要考慮重量、防水等需求所進行推薦，然後返回推薦結果。常應用的場景有汽車、電腦、房屋、理財產品等。這些場景通常很難在一個商品上獲取大量的使用者評分資訊。

二、強化學習

強化學習是一種即時性的機器學習的方法,和監督式學習最大的區別在於，它不需要額外對每個輸入標記答案，也不需要事先收集大量資料，而是在收集資料的過程中同時更新模型，主要概念是強調如何基於環境而進行動作，以獲得最大化的預期利益，透過跟外界的互動來不斷學習，又稱為近似動態規劃(Approximate Dynamic Programming, ADP)或增強式學習。

在強化學習的架構下存在環境(Environment)與代理人(Agent)兩個主要部分，強化學習是描述兩者互動的方法，環境是強化學習的使用場景，代理人是觀察環境並做出動作的智能體，只要有辦法定義適當的狀態、動作等相關參數，根據價值函數去幫不同的動作評估其價值，再依照所定義的策略選擇，最後從執行動作的結果給予獎勵並更新回代理人，不斷反覆訓練此流程直到收斂。

馬可夫決策過程(Markov Decision Processes, MDPs)是強化學習的基礎，用來描述環境與代理人間的狀態轉移互動，由五個元素(S，A，P，R，γ)組成，而我們的目標就是找到最佳的狀態節點，而它具有最高的獎勵。

- 狀態集(Set of States)是代理人從環境中獲得的資訊稱為狀態，所有可能獲得的狀態記作 S。
- 動作集(Set of Actions)是代理人根據狀態產生的對應行為稱為動作，所有可能的動作記作 A。
- 轉移機率(Transition Probability)表示狀態 S 進行動作 a 後轉移到狀態 S'的機率，此狀態轉移矩陣記作 P。
- 獎勵(Reward)表示再狀態 S 採取動作 A 後轉移到狀態 S'得到的獎懲分數記作 R。
- 折扣因子(Discount Factor)表示未來的獎勵在當前的價值，用來估計整體獎勵中使用的變數記作 γ，當 $\gamma=0$ 時，獎勵僅考慮立即回報，當 $\gamma=1$ 時，所有未來的回報都可以算在當前動作的結果，γ 越大代表越重視未來可得到的獎勵。

透過上述就可以估計在不同狀態下進行不同動作能獲得的期望獎勵，稱之為價值，在後面的決策過程中透過價值來判斷該如何決策，並在強化學習的學習過程中不斷更新價值，整個強化學習的學習過程就是在馬可夫決策過程中不斷的改變轉移。

強化學習的策略主要分為兩種，分別是 On-policy 與 Off-policy，On-policy 是基於當前的策略直接執行一次動作選擇，然後用這個樣本更新當前的策略，因此生成樣本的策略和學習時的策略是相同的，優點直觀且速度快，缺點是光利用目前已知的最優選擇，可能學不到最佳解，相關的方法有 Sarsa (Sutton, 1996)、Sarsa(λ)(Harutyunyan et al., 2016)、Policy Gradient (Silver et al., 2014)等。Off-policy 生成樣本的策略和學習時的策略是不同的，在計算下一狀態的預期收益時通常使用貪婪策略，直接選擇最優動作，而當前策略並

不一定能選擇到最優動作，優點是因為涵蓋行為更加全面所以較通用，缺點是收斂慢，相關的方法有 Q-learning、Deep Q Network(DQN)、Proximal Policy Optimization(PPO)、Trust Region Policy Optimization(TRPO)等(Roderick et al., 2017)。

在推薦系統結合深度神經網路的部分，因為深度學習也發展出很多不同特色的網路架構，所以也產生了不同的形態的推薦系統，像在監督學習中，其中較有名的方法像是 Steffen Rendle 等人(2012)提出了一種個性化排名的方法 BPR-OPT，方法是利用協同過濾的反饋矩陣為輸入，基於貝葉斯分析得到的最大後驗機率來對商品進行排序，搭配深度神經網路架構進行隨機梯度下降訓練學習推薦。Simon Stiebellehner 等人(2017)提出兩種混合濾波技術，從 Apple Store 和 Google Play 獲取數據，訓練深度神經網路 user2vec 及 context2vec 模型，混合為 doc2vec 模型並在廣告交易平台中提升預測性能。

另外在非監督式學習中的自動編碼器(AutoEncoder)也能作為推薦系統的架構，像是 Florian Strub 等人(2016)利用自動編碼器為基礎來訓練協同過濾中矩陣分解的方法，對於預測模型值的平均誤差有所改善。Vito Bellini 等人(2017)提出了一種新方法 SEM-AUTO 來提取和加權語義特徵，將電影與電影標簽之間的關係映射到自動編碼器，將使用者對電影的評分情況作為訓練資料，最後得到使用者對電影標簽的偏好。Wei Zhao 等人(2017)提出了一種基於生成對抗網絡的推薦系統 LSIC 模型，以循環神經網絡中的 LSTM 作為基礎架構，能動態調整歷史資料中長期偏好和短期會話的電影推薦模型。

另外在解決大量狀態與動作的部分，大量狀態與動作會造成推薦系統效能不佳，以致無法進行即時的推薦，因此近期少數研究也針對這個問題來解決，像是 Gabriel Dulac-Arnold 等人(2015)提出強化學習為基礎的 Wolpertinger 框架，使用近似最近鄰方法允許對數時間查找相對於動作數量的複雜性，任務最多可以多達到一百萬個動作。Xiangyu Zhao 等人(2017)根據來自真實電子商務網站的數據提出了深度強化學習架構 LIRD，利用 DDPG 算法訓練框架的參數，可應用於較大動態項目空間的場景。Sungwoon Choi 等人(2018)將推薦系統中的定義場景轉換至 gridworld 遊戲中，通過使用雙簇(Biclustering)技術可以減少狀態和動作空間。

參、研究方法

一、研究應用環境說明與問題描述

本文欲解決之問題為設計新型態的電影推薦演算法，其應用環境能夠處理大量電影數量與新進使用者，且不會受到使用者太少產生冷啟動問題的影響，在每次推薦都可以動態對系統進行更新，結合協同過濾能從他人的使用找出潛在的興趣推薦，和基於內容基特性輔助的混合式強化學習推薦系統。其中使用強化學習在推薦系統上有兩個優點，第一是他們能夠在互動過程中不斷更新策略，直到系統收斂到生成的最佳策略，並建議最適合使用者的喜好。第二是通過長期性最大化預期，來制定最佳使用者策略的累積獎勵，結合深度學習的方法使其可以靈活地實踐在擁有大量電影的推薦系統中。

方法整體的軟體架構如 Figure1 顯示，將環境中使用者評分轉換為電影標籤基因組狀態，送入代理人在大量離散動作問題下進行策略選擇動作，透過線上模擬環境對推薦清單中的電影進行評分與計算獎勵，最終更新模型並根據使用者的反饋決定是否繼續推薦。

一開始使用者先在電影推薦系統客戶端點擊電影並進行評分，在資料庫建立獨立金鑰的請求，將電影、使用者及評分資訊送入預先載入的強化學習模型，經演算法即時預測喜好電影並更新模型，回傳預測結果回資料庫中相同金鑰，客戶端得到推薦的電影並針對

該推薦進行反饋，此系統架構可以在多位使用者環境中進行線上推薦，且不會發生結果不同步問題，省下每次請求重新載入模型與連線之時間。系統有以下幾點假設：

1.假設電影種類數量為固定，且系統已知有多少種電影。

2.假設推薦的電影使用者如果給予高分則會繼續使用不會離開系統。

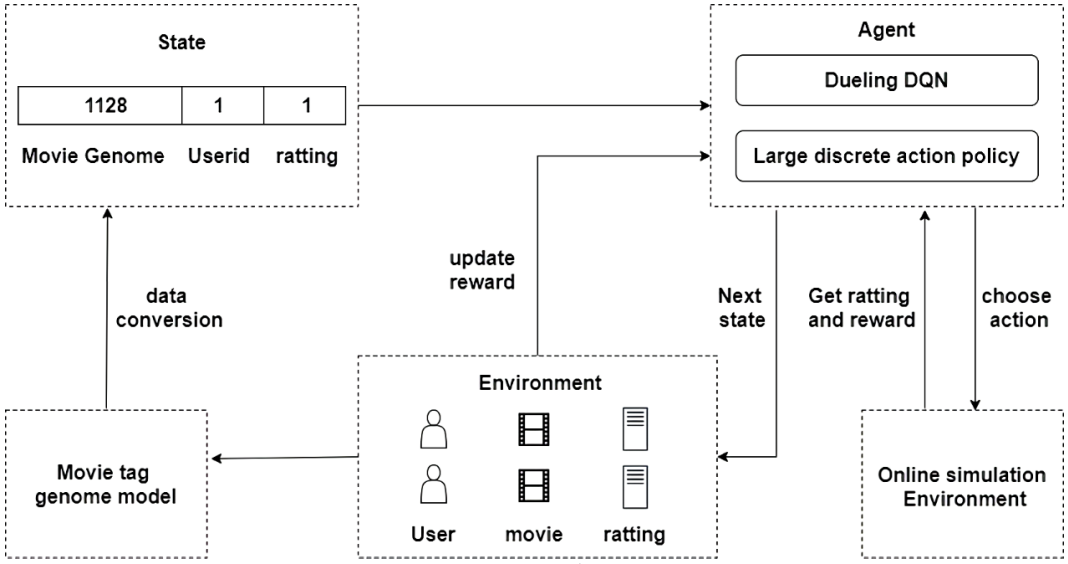

Figure 1 研究系統架構

二、混合式強化學習推薦模型

本研究以 Dueling DQN 方法架構進行實驗，在電影推薦的案例上，對於使用者當前看的電影與評分進行狀態評估，並對於所要選的推薦電影進行動作評分，若有些使用者沒有特別的潛在興趣或是一些較獨特的電影類別，只看動作價值可能無法達到最佳推薦，若能將當下觀看的電影狀態進去一起評估，才能根據現況選出更加適合的推薦電影提升推薦質量，為了推薦的即時性與效能，這裡大量簡化了神經網路的層數與參數，此處模型的輸出動作個數以 n 表示，隨著不同的情境所需要的動作個數會跟著改變，輸出的數值對應的是各動作之動作價值。

在強化學習動作價值函數與策略，透過貝爾曼方程式(Bellman Equation)來表示估計價值的函數，經由價值函數計算每個狀態或動作的價值，然後透過價值與策略函數選擇適當的動作，以貪婪策略為例，便能選出最佳的動作 $Q^*(s, a)$，又稱最佳化動作函數

$$Q^*(s, a) = \max_{\pi} Q^{\pi}(s, a) \quad \text{.....(1)}$$

將 Q 值分為兩個部分計算，分別為狀態價值函數 $V(s, \theta, \beta)$和優勢函數 $A(s, a, \theta, \alpha)$，他們是共用相同的輸入、相同的隱藏層神經網路參數來計算，加總後得出我們最後要的動作價值函數，搭配貪婪策略選出最佳的動作，但只利用(式 2)計算相加的 Q 值時可能會出現無法辨認(unidentifiable)的問題。

$$Q(s, a, \theta, \alpha, \beta) = V(s, \theta, \alpha) + A(s, a, \theta, \beta) \quad \text{...(2)}$$

因為當我們給定一個 Q 時，卻無法得到唯一的狀態價值函數和優勢函數，造成狀態價值函數不能反映真實 state 值，優勢函數不能反映真實 advantage 值，這裡改用(式 3)來解決，其中 θ 為公共部分(隱藏層)的網絡參數，α 為狀態值函數的網絡參數，β 為優勢函數的網絡參數，將優勢函數做中心化處理減去其平均值，不僅提高了模型的穩定性，效果也更佳。

$$Q(s,a,\theta,\alpha,\beta) = V(s,\theta,\alpha) + \left(A(s,a,\theta,\beta) - \frac{1}{|A|}\sum_{a'}(A(s,a',\theta,\beta) \right) \quad \text{.....(3)}$$

在強化學習中餵入模型的狀態為S_t，在我們定義的電影推薦情境中狀態將包含電影編號(movieId)、使用者編號(userId)與其評分(ratting)，由於對代理人來說單純的電影編號只是一個數字沒意義，必須將其轉換為有代表意義且相關性的數值，像是能代表電影的特徵集合或是它的相關資訊等，其概念就像生物的基因組序列DNA，而這裡是基於電影的標籤產生的電影基因組序列。

根據Jesse Vig (2012)所提出的方法，他們將使用者對於MovieLens資料集中收集的使用者對不同電影進行的評價標籤，總共有1128種不同類別的標籤，使用機器學習的方法，得到數據屬性(電影)和輸出變量(標籤相關性)之間的映射，其中tag-independent是所有標籤具有統一的模型，而tag-specific是每個標籤學習單獨的模型，將這些資料以上面兩種模型的特性結合訓練，不只可以捕獲特定標籤特徵值和標籤相關性之間的關係，同時避免了特定標籤的模型的過度擬合問題，以非線性回歸模型所訓練出來的電影基因組序列，在實作上可用R語言的glmer()函式完成訓練，經過轉換後可以得到每組相對應的1128維的基因組與13816部電影的標籤基因組模型，所有數值區間皆為[0,1]，將該電影以1128個不同的標籤所佔的比例來表示，例如電影編號1089的"Reservoir Dogs"包含在動作、暴力、黑暗、犯罪、有趣的標籤特徵值較高，代表其包含有上述的標籤特徵，最終框架輸入的狀態為進行基因轉換後，每個movieId所對應的1128維預訓練的電影特徵加上userId、ratting共1130維。

在強化學習中動作是指預測未來會轉移到哪一個狀態的這個過程,Dueling DQN模型在選擇該動作時，動作價值函數會根據當前狀態以及未來動作兩者進行評估對系統進行推薦，選取該系統在未來的狀態中所採取之動作，並調整其模型中的權重。獎勵函數是強化學習進行更新策略的基礎，學習的目標是想辦法最大化獎勵函數，而獎勵函數的設計會影響最終模型是否能收斂，根據使用者對於推薦清單的評分產生獎勵，這裡選出4種推薦評價指標作為獎勵。

- hit :用來對推薦清單中的電影進行加權分數。
- CTR(Click Through Rate) : 點擊率作為推薦清單命中結果的評價指標，用來評估所推薦的K個項目是否會被使用者點擊，這裡假設評分大於3分的商品會被使用者點擊。
- nDCG(Normalized Discounted Cumulative Gain) : 標準化折現累積收益作為推薦排序結果的評價指標，用來評估所推薦的項目在推薦清單中的表現，這樣設計使得排名越前面的電影權重越大，讓推薦結果會更專注在推薦好的電影在前面。
- Top-k Accuracy: Top-k準確率作為推薦清單相似類別結果的評價指標，用來評估所推薦的項目是否與原本的類別有相關，預測k種類別只要任何一部電影命中同類別都算對，全部都不是則為未命中。

獎勵R是由選擇前K個動作後，根據推薦清單結果個別計算上述4種動作評價指標，其所產生的值乘上各自加權權重相加而成(式4)，若推薦清單中有任何評分低於3分即不良推薦時給予懲罰，即將獎勵減半，否則維持原本的獎勵，最後回傳本次推薦產生的最終獎勵Reward。

$$R = hit_k + \omega_C * CTRk + \omega_n * nDCG_k + \omega_T * Topk_k \quad \text{.....(4)}$$

三、大量離散動作選擇問題

強化學習算法的時間複雜度是[O(|A| * |S|)]，其中S是狀態的總數，A表示代理人可以選擇的動作的總數。計算複雜度是[Q(|A| +M)]，其中A與前者相同表示動作總數，M是儲存起來過去經歷過的[狀態-動作]組合的數目，由此可見大量的動作數對於強化學習的計算與時間有著很大的影響，執行複雜度隨著|A|的數量線性增長，所以一般對於高維度的動

作空間與狀態空間問題的解決，常規的強化學習算法總是效能低落。在大部分的推薦系統中，所包含的動作量非常的多，且可能會隨系統上線時間再增加新產品，如果限制在小量的動作進行訓練無法符合現實的應用情況，但強化學習在大量動作的情況下，可能會面臨到一些問題，例如資料分布太稀疏導致模型無法往正確的方向收斂，且大量動作選擇下的計算成本昂貴，花費的時間會非常大，目前尚未有能夠完全解決此問題的架構，為了改善此問題，我們必須處理三件事情，首先必須先將動作集合上的所有動作泛化(Generalization)，使其能嵌入連續空間中，然後透過方法縮小動作的量級，來降低查找複雜度，最後進行最終動作的選擇策略，這裡我們提出一種演算法，運作流程如下。在動作泛化中的標籤基因組模型來達到此目的，這裡先將所有的原始動作(Proto-action)集合泛化到相對應的電影基因組，其分布的連續矩陣空間中。在縮小動作的量級中，這裡使用非監督學習中的k-means clustering，他的優勢在於在特徵學習的任務中，雖然較簡易但能展現出與其他複雜特徵學習方法如自動編碼器等相當的效果，計算成本較低，如Figure 2將泛化完的電影基因組對於其各項的特徵向量進行分群，將大量的原始動作降成維度較少的群，並輸出映射到離散的動作群集(Cluster-action)。

然而非監督中的分群最常見的問題就是要分幾群結果才是較好的，這邊利用Average silhouette Method來尋找適合的k值，在分不同的k群下， Sc值越大越好，代表不同群的點分得越開，同群的點更集中，不同動作集必須根據其分布的狀況來決定分群的方法，這邊通常需要一些經驗法則，若分群的效果不佳會產生雜訊點，影響整體模型的結果。

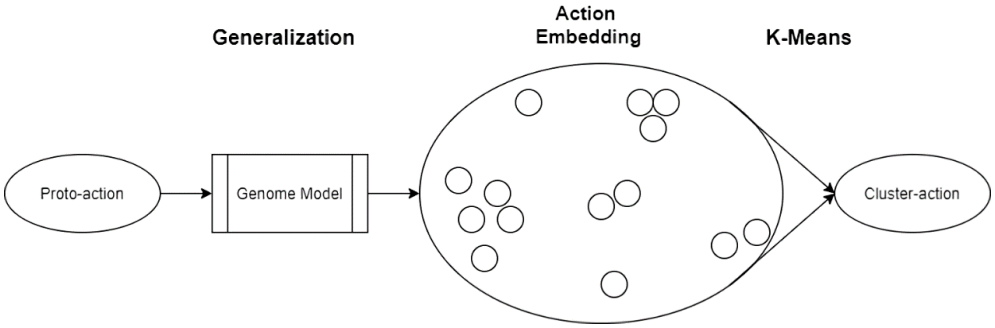

Figure 2.大量離散動作泛化並降維示意圖

肆、結果與討論

一、實驗資料集

實驗資料集採用由 GroupLens Research 實驗室於 1997 年創建的開源資料集 MovieLen，也是最常見且完整的電影推薦資料集之一，他們收集了各種不同數據大小的資料集與內容，以針對不同的應用場景，這裡選用 MovieLens 25M Dataset 來評估推薦系統演算法，此資料集未對於使用者的特徵或資訊(性別、年齡等)進行紀錄，也未給予電影的相關資訊(上映日期、導演、片長等)，主要是紀錄不同使用者對於不同電影的評分，下面為實驗會用到的資料如 ratings.csv 有 25000095 筆歷史資料，包含了使用者編號(userId)、電影編號(movieId)、評分(rating)，評分的範圍為 0 至 5 分包含小數，是用來訓練強化學習模型輸入的主要輸入資料集。movies.csv 包含了電影編號(movieId)、電影名稱(title)、電影類型(genres)共 20 類，一部電影可有一至多種類別，是用來計算推薦電影的 Top-k Accuracy 分類，在實驗資料採集過程中可以發現資料集類別分布很不平均，有超過 50%的電影都包含在類別六"Drama"及它項，這裡會根據資料集特性修改 Top-k Accuracy 公式，推薦的所有電影中需有大於等於 2 個相同類別才算為命中，避免即使代理人亂猜也很容易命中的情況發

生，導致此指標學習不佳。tags.csv 包含了使用者編號(userId)、電影編號(movieId)、評論標籤(tag)，是用來計算電影特徵的資料集也就是各電影的基因組，共以 13816 部電影進行建模。

二、演算法架構比較實驗

本實驗將比較 Natural DQN 與 Dueling DQN 兩種演算法在相同的線上模擬環境中之表現，用來證明所選擇之方法在此環境下其他的實驗表現皆較佳。本實驗將兩個演算法在相同環境中訓練固定的總回合數(episode)，使用資料集一進行初步少量資料集(100 部電影)的快速實驗，每 200 個訓練 episode 為單位比較各方法在驗證資料集中的 loss、Average Step 的分數表現。其結果如下 Figure3 與 Figure 4 所示。Dueling DQN 在早期便能縮小 loss 並穩定收斂，且 Dueling DQN 在各指標上表現也較佳 Dueling DQN 改善了不只針對評估各動作的分數，也會為當前看的狀態評估分數，對於推薦電影更有幫助。

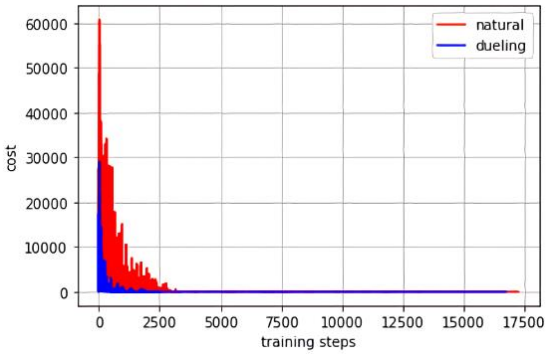

Figure 3. Natural DQN 與 Dueling DQN 於模擬環境中之 loss 的分數表現

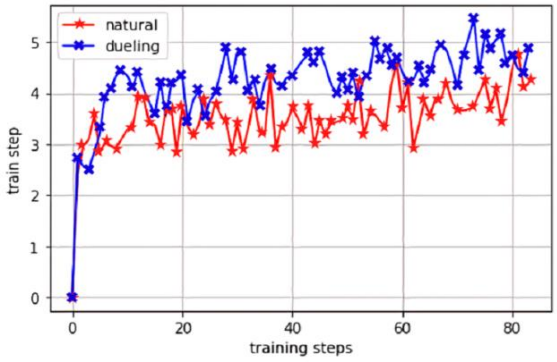

Figure 4. Natural DQN 與 Dueling DQN 於模擬環境中之 Average Step 的分數表現

三、冷啟動實驗

由於在一般的推薦系統中冷啟動問題是很常見的問題，我們必須驗證在只有少量使用者評分時，系統能否正常運作推薦，且保持良好的推薦效果。本實驗將使用大量動作 13816 部電影，即資料集中所有的電影作為動作，在環境中訓練固定的總回合數(episode)，使用資料集三只保留每個使用者的 10%評分紀錄進行大量資料及大量動作的實驗，每 500 個訓練 episode 為單位比較各方法在驗證資料集中的 loss、Average Step 的分數表現，如 Figure 5 與 Figure 6 所示。結果顯示即使在推薦系統中常見的冷啟動問題下，也能快速在各項評分指標中有良好的結果，且皆有穩定收斂的趨勢。整體實驗結果及訓練過程與沒有冷啟動情況下相差不遠，顯示本研究方法能有效改善冷啟動問題。

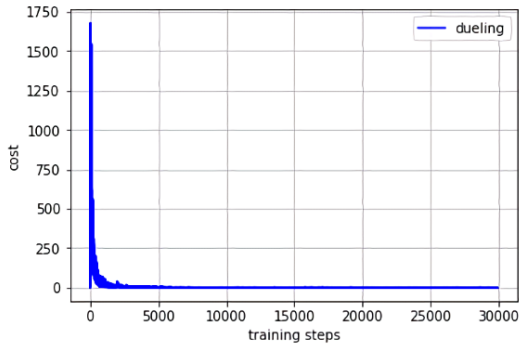

Figure 5. Dueling DQN 於模擬環境中之冷啟動 lost 的分數表現

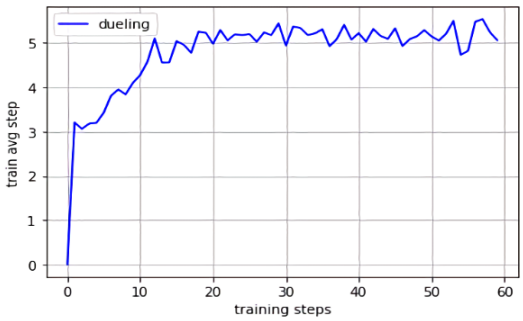

Figure 6. Dueling DQN 於模擬環境中之冷啟動 Average Step 的分數表現

伍、結論

本研究中提出了一種電影推薦的方法，期望可以利用多種不同的特性達到良好的推薦，依據實驗結果，本研究就可以整理出以下四點貢獻及特性:

1. 推薦演算法結合多種特性 : 從本研究的模型結果來看，此方法不僅有強化學習的動態學習優勢，結合了協同過濾的相同興趣使用者的推薦，與基於內容在少量資訊下也能針對類似電影推薦等這些個性化推薦方法之優點，組合產生的級聯混合式推薦系統，可以發現此演算法有一定程度的推薦準確率，並且持續的向更高的獎勵推薦評分收斂。
2. 避免冷啟動問題 : 在使用者冷啟動問題下，系統中各個使用者的數據量非常少的情況，可能造成無法良好的個性化推薦，在訓練過程中證明此方法還是能從少量使用者中，學習如何推薦並快速建立好推薦的模式。
3. 在相同的別迭代數下結果較佳 : 由實驗可以得出，使用 Dueling DQN 作為架構比一般的 DQN，在相同的迭代訓練下各項得分都是更佳的，此外在大量動作下問題中 Dueling DQN 也比一般的強化學習定義的方法，在相同迭代下各項得分也都是更佳的且更早收斂。
4. 模型通用性高 : 由實驗可以得出，模型在訓練時在驗證資料集的表現，與在訓練結束時在測試資料集的表現中，都沒有發生過度擬合的問題，且即使在完全沒有見過的使用者下，所推薦的電影在各項指標中也都有不錯的分數，代表模型的通用性是好的，對於沒見過的資料集也能有好的推薦。

本研究目前在此學習環境中有著不錯的實驗結果，但也是根據歷史資料集驗證的結果，礙於無法得到大量的實際使用者對於此系統的評估，未來若能將此電影推薦系統上線可以得到更多真實反饋證明。

陸、引用文獻

Aggarwal, C. C. (2016). Model-based collaborative filtering. In Recommender systems (pp. 71-138). Springer, Cham.

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

Bellini, V., Anelli, V. W., Di Noia, T., & Di Sciascio, E. (2017, August). Auto-encoding user ratings via knowledge graphs in recommendation scenarios. In Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems (pp. 60-66).

Burke, R. (2000). Knowledge-based recommender systems. Encyclopedia of library and information systems, 69(Supplement 32), 175-186.

Choi, S., Ha, H., Hwang, U., Kim, C., Ha, J. W., & Yoon, S. (2018). Reinforcement learning based recommender system using biclustering technique. arXiv preprint arXiv:1801.05532.

Dulac-Arnold, G., Evans, R., van Hasselt, H., Sunehag, P., Lillicrap, T., Hunt, J., ... & Coppin, B. (2015). Deep reinforcement learning in large discrete action spaces. arXiv preprint arXiv:1512.07679.

Gong, S., Ye, H., & Tan, H. (2009, May). Combining memory-based and model-based collaborative filtering in recommender system. In 2009 Pacific-Asia Conference on Circuits, Communications and Systems (pp. 690-693). IEEE.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. ACM sigmod record, 29(2), 1-12.

Harutyunyan, A., Bellemare, M. G., Stepleton, T., & Munos, R. (2016, October). Q ($$\lambda $$) with Off-Policy Corrections. In International Conference on Algorithmic Learning

Theory (pp. 305-320). Springer, Cham.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000, December). Explaining collaborative filtering recommendations. In Proceedings of the 2000 ACM conference on Computer supported cooperative work (pp. 241-250).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Jiang, S., Qian, X., Shen, J., Fu, Y., & Mei, T. (2015). Author topic model-based collaborative filtering for personalized POI recommendations. IEEE transactions on multimedia, 17(6), 907-918.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

Paterek, A. (2007, August). Improving regularized singular value decomposition for collaborative filtering. In Proceedings of KDD cup and workshop (Vol. 2007, pp. 5-8).

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In The adaptive web (pp. 325-341). Springer, Berlin, Heidelberg.

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2012). BPR: Bayesian personalized ranking from implicit feedback. arXiv preprint arXiv:1205.2618.

Resnick, P., & Varian, H. R. (1997). Recommender systems. Communications of the ACM, 40(3), 56-58.

Roderick, M., MacGlashan, J., & Tellex, S. (2017). Implementing the deep q-network. arXiv preprint arXiv:1711.07478.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (pp. 285-295).

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484-489.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014, January). Deterministic policy gradient algorithms. In International conference on machine learning (pp. 387-395).

Stiebellehner, S., Wang, J., & Yuan, S. (2017). Learning continuous user representations through hybrid filtering with doc2vec. arXiv preprint arXiv:1801.00215.

Strub, F., Gaudel, R., & Mary, J. (2016, September). Hybrid recommender system based on autoencoders. In Proceedings of the 1st workshop on deep learning for recommender systems (pp. 11-16).

Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. Advances in neural information processing systems, 1038-1044.

Vig, J., Sen, S., & Riedl, J. (2012). The tag genome: Encoding community

knowledge to support novel interaction. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3), 1-44.

Zahir, A., Yuan, Y., & Moniz, K. (2019). AgreeRelTrust—A Simple Implicit Trust Inference Model for Memory-Based Collaborative Filtering Recommendation Systems. Electronics, 8(4), 427.

Zhao, W., Chai, H., Wang, B., Ye, J., Yang, M., Zhao, Z., & Chen, X. (2017). Leveraging long and short-term information in content-aware movie recommendation. arXiv preprint arXiv:1712.09059.

Zhao, X., Zhang, L., Xia, L., Ding, Z., Yin, D., & Tang, J. (2017). Deep reinforcement learning for list-wise recommendations. arXiv preprint arXiv:1801.00209.