國立台灣師範大學
資訊工程研究所碩士論文

指導教授：　柯佳伶　博士

從 GPS 軌跡以遞迴類神經網絡預測個人活動意圖
Predicting Personal Activity Intention from GPS Trajectory with Recurrent Neural Networks

研究生：　朱修毅　撰

中華民國　一百零七　年　十一　月

摘要

從 GPS 軌跡資料以遞迴類神經網絡預測個人活動意圖

朱修毅

本論文研究活動意圖類型預測方法，以遞迴類神經網路架構為基礎，用建立群組模型的概念，比較四種建立模型的架構。第一種是全體資料模型，以所有 GPS 軌跡資料擷取特徵作為模型輸入，建立全體資料模型。第二種是群組模型，本論文提出兩種分群方法，分別為使用者為單位進行分群的使用者分群法，及以序列為單位進行分群的序列分群法，再以群組資料建立群組模型。第三種是遷移學習模型，以全體資料進行訓練，將全體資料訓練好的參數設置為初始參數，以群組資料作為訓練資料，只對模型中部份層的參數進行調整。第四種是合成模型，將全體資料模型和群組模型預測結果，學習調和參數將兩個預測結果進行比重相加。實驗評估顯示，遷移學習模型在 OSM 資料集的預測結果優於全體資料模型和群組模型，合成模型在大部分情況下可良好地結合兩模型，正確預測出使用者的活動意圖。在 Geolife 資料集中，合成模型在 Accracy@5 最高可達 89.31%的準確率，在 OSM 資料集中，則可達 74.12%的準確率。

關鍵字：GPS 軌跡、活動意圖預測、基於遞迴神經網路的學習網路

ABSTRACT

Predicting Personal Activity Intention from GPS Trajectory

with Recurrent Neural Networks

by

Hsiu-Yi Chu

In this paper, we study the problem of predicting activity intentions based on the recurrent neural network architecture. We constructed four learning models based on the various combinations of training data sets and recurrent neural network architectures. The first one is the global model, which uses all activity sequences as the training data set. The second one is the group model, in which two clustering methods: user-based clustering and sequence-based clustering methods are proposed to separate the data into groups. Accordingly, a prediction model is constructed respectively for each group of training data. The third one is the transfer-learning model, in which the parameters learned from all training data set are set as the initial parameters. Then the training data in each group is used to adjust the parameters from the middle layer of the RNN architecture to construct the predicting model for each group. The last one is the ensemble model, which concatenates the predicting results of the global model and the group model to learn the ensemble parameters to get a properly weighted sum of the two predicted results. The results of experiments show that the transfer-learning model on the OSM dataset has better performance than the global model and the group model. Furthermore, the ensemble model can combine the results of two models well in most cases and provide the highest accuracy. In the Geolife dataset, the accuracy@5 of the ensemble model achieves 89.31%, and gets 74.12% on the OSM dataset.

Keywords: GPS trajectory, activity intention prediction, learning network based on RNN

致謝

碩士班的兩年多中，能完成碩士學位與論文，首先感謝我的指導教授柯佳伶老師。感謝老師耐心且嚴格的指導，在我毫無頭緒時指引論文研究方向，讓我能順利完成研究。在處事上也學習面對問題的態度，並提醒自己能夠以實例來驗證自己的論述，論之有理。在論文撰寫階段，感謝老師耐心地逐字批改並指導我論文寫作技巧。由衷地感謝老師這些日子的包容與指導。

感謝沈錳坤教授和徐嘉連教授在百忙之中擔任我的口試委員，給予我許多寶貴的建議以及指導，讓此研究更加完善，在此致上最深的謝意。

感謝謹安、博文、仕翰與家儀，在論文研究上總是與我討論和指點，特別感謝博文和家儀，在程式設計和練習報告上給了我很多意見和幫助。感謝翊誠、盈翔、明潔以及碩 1 的學弟妹們，在口試時，幫忙餐點準備、布置教室、紀錄與攝影，謝謝你們。

感謝我的家人在我忙碌及壓力大的時候，給我無條件的支持與無虞生活環境，即使不在台灣，也時常關心我的健康和論文的進度。對於碩士這期間支持與陪伴我的人，在此獻上自己最大的感謝，因為有你們我才能順利走到這一步完成碩士學業，謝謝大家。

朱修毅 謹致

於國立台灣師範大學資訊工程研究所

2018 年 11 月

目錄

附圖目錄

附表目錄

第一章 緒論

1.1 研究動機與目的

隨著資訊科技日益發展，人們經常會利用行動裝置來記錄自己的位置。最常見的方式包括在社群媒體上留下記錄，例如在 Facebook 或是 Twitter 打卡，或是使用 GPS 定位系統留下軌跡記錄。觀察這些記錄可以得出許多有用的資訊，例如可以發現某個區域被拜訪的頻率較高，或是預測出人們的移動行為和模式，進而自動推薦使用者需要的資訊。

在許多研究中，用來預測位置或是推薦系統的研究，多以打卡資料作為輸入。打卡資料和 GPS 軌跡資料最大的不同之處，在於打卡資料會有明確的 POI(Point of Interest)和貼文訊息或標籤。POI 是指像 Google Map 等電子地圖上的地標或景點，包含名稱、類別、經緯度、海拔等資料，可標示出該地所代表的政府部門、商業機構、旅遊景點、古蹟名勝、交通設施等處所；而 GPS 軌跡則只有經緯度和時間等沒有語意的記錄點資訊。這意味著打卡資料可直接判別出地點類型，而 GPS 定位是原始經緯度座標，且由固定時間間隔抓取紀錄，無法明顯看出停留點。所以 GPS 資料需要結合其他處理方法，有效比對出使用者到達的 POI 位置及對應的地點類型，在資料處理上比較有挑戰性。然而人們通常不會在自己慣常的所在地打卡，所以無法完整反應出使用者的日常行為軌跡，且並不是所有人都習慣在社群媒體上留下打卡記錄，即使有也不見得每個行為都會留下紀錄。反而由 GPS 定位系統可以隨時記錄使用者蹤跡，有詳細的時間記錄，且不受限於需要自

已打卡。

近年來機器學習的技術被廣為使用，其中遞迴類神經網路在序列預測的效果很好，因此本論文嘗試將此技術應用在位置預測的研究。本論文認為，地點類型能夠表現出使用者的日常偏好及習慣，因此透過建立遞迴類神經網路模型，從使用者的 GPS 軌跡資料預測使用者接下來可能停留的目標地點類型，進而用於推斷出使用者接下來的活動類型或目的。訓練模型所使用的資料和預測效果有很大關係，當使用者的歷史軌跡不夠多，不足以建立個人的軌跡預測模型，但以全部使用者的行為序列進行預測模型，又可能使模型過於一般化。因此是否能透過協同式過濾的想法，對相似行為序列的資料分別建立群組預測模型，以提高預測準確率，為本論文的研究動機。

1.2 研究的範圍與限制

本論文的研究範圍是針對使用者 GPS 軌跡歷史資料，探討如何從這些歷史記錄建立預測下一個停留地點類型的模型。在 GPS 軌跡資料集中，每位使用者有多天的軌跡資料，每筆 GPS 軌跡資料記錄包含經緯度、海拔高度、日期及時間。本論文將探討如何從這些記錄資料，找出使用者的停留地點和類型，以及在該地點的時間和停留時間，形成使用者活動序列。本論文針對 GPS 軌跡資料中可以擷取出停留點的資料作為研究，無法擷取出停留點的資料不在本論文研究之內。使用者活動序列會被切割成長度 k 的序列，作為模型的訓練資料。此外，本論文將研

究如何對使用者活動序列進行分群以建立群組模型。

本論文的研究重點分成三部分：(1)如何將 GPS 原始經緯度座標軌跡轉換成停留地點類型序列。(2)如何對使用者活動序列進行分群，用來建立不同群組模型進行預測。(3)如何運用不同群組資料，組合建立遞迴類神經網路預測模型。

1.3 論文方法

(1) 將 GPS 原始經緯度座標軌跡轉換成停留地點類型序列：本論文將找出每段軌跡內停留時間較長的停留點，並使用騰訊位置服務 API 和 Foursquare Venue Search API 工具，判斷該經緯度座標對應或附近的 POI 名稱和類型。

(2) 將使用者活動序列進行分群：本論文提出兩種分群方法，分別針對使用者和序列進行分群。使用者分群是針對使用者，由其活動序列中擷取出不同活動間的移動模式(Transition Pattern)，根據移動模式計算使用者彼此的相似度，再用階層式分群方法對使用者(Hierarchical Clustering)進行分群。序列分群則是對所有長度為 k 的活動序列計算序列彼此的相似度，再用階層式分群方法對序列進行分群。

(3) 運用不同群組資料組合建立遞迴類神經網路預測模型：本論文首先建立全體資料模型和群組模型兩種模型，前者利用全部使用者的活動序列資料建立預測模型，後者則是對各分群的活動序列資料分別建立預測模型。此外，本研究提出兩種組合模型，包括合成模型和遷移學習模型。

為評估本論文所提出方法的成果，本研究將進行三部份的實驗：第一部份以全體資料模型與相關論文方法做比較，實驗結果以預測地點類型前 k 名有出現跟實際地點類型的比例(Accuracy@k)進行評估。第二部份則評估各群組模型的準確率，以及和全體資料模型的比較。第三部分探討合成模型和遷移學習模型的準確率，以及各模型在不同長度 k 活動序列的準確率表現。

1.4 論文架構

本論文以下章節內容簡介如下：第二章將說明相關文獻，第三章說明本論文之問題定義與系統架構，第四章將詳述資料的前處理和特徵擷取。接下來在第五章說明資料分群方法，第六章說明活動意圖預測方法。第七章將呈現本論文方法的實驗結評估果，最後在第八章提出總結並探討未來研究方向。

第二章　文獻探討

有鑑於行動裝置和社群媒體上的位置服務日益越新，近年來有愈來愈多研究探討位置服務相關的技術與應用。以下將依序介紹與本論文相關的研究，並將其分成軌跡分析處理技術和地點預測方法兩部分進行探討。此外，類神經網路的技術也逐漸被廣用在各個領域的研究當中，將人們移動行為視為序列，探討如何運用遞迴類神經網路的學習方法預測接下來的行為。本章節將分別探討上述相關文獻。

2.1 軌跡分析處理技術

近年來人們使用 GPS 導航記錄外出行為已越來越常見，這些位置記錄會形成 GPS 軌跡，不只記錄使用者在現實世界的位置走訪歷史，也能顯示使用者的習慣、興趣和偏好。以下為與軌跡分析處理技術有關的研究，分成軌跡樣式探勘和軌跡相似度評估。

2.1.1 軌跡樣式探勘

軌跡樣式探勘是指根據 GPS 的原始軌跡，找出人類常見的移動模式，也就是從一個地點到另一個地點，會經由那些地點。但以 GPS 軌跡紀錄當作原始資料時，因為單從 GPS 軌跡記錄不容易出現完全相同的地點，且因為 GPS 軌跡資料容易發生飄移，所以一個重要處理步驟是找出較可能的停留點，論文[10][17][3][4][7][19]皆提出或引用找出停留點的方法。停留點的定義因應用需求而定，廣泛的

定義為在距離門檻值以內的區域，停留超過某個時間門檻值的時間，這個區域即被稱作停留區域，而停留區域的中心點就被被稱作停留點。

透過 GPS 軌跡資料找出停留點以後，只提供一個經緯度座標，無法得知這個座標是什麼地點。因此論文[3]使用密度分群演算法，根據地理位置將停留點分群，對每個區域(群)中包含的多個停留點標記出地點類別，最後建立週期模型，計算每個地理類別區域被拜訪的週期性。和論文[3]不同，論文[4]則利用 Google Map 位置服務提供的 API 來查詢停留點的所屬類別，根據該經緯度座標的 POI 類型當作該停留點的類別。除了考慮地理空間外，論文[5]則比對是否有 GPS 軌跡段落在時間和空間上都匹配，將軌跡根據時間和空間相似性分群，擷取使用者的運動模式，預測該地理區域的地點類別。

論文[7]認為透過使用 Google Map API 反向地理編碼技術[4]查找出的 POI 標籤經常是一個郵政地址，例如 x 路、y 城市等等，但關於地理位置的含意(如家庭、工作場所等)無法從地理資訊資料庫中取得。於是論文[7]使用密度式分群和時間性分群兩個方法找出停留點的經緯度座標，再從每個經緯度座標擷取出走訪模式的時間特徵，例如停留時間和訪問次數等，根據時間特徵對這些位置指定預定義的類型，當作該地點的語意標籤。

除了 GPS 軌跡記錄的經緯度資料外，若透過社群媒體或手持裝置上的位置服務(打卡)取得地點資料，就可能包含該經緯度座標的其他資訊，例如使用者留下的訊息或標籤。論文[6]認為社群媒體上相關打卡和短信息中所提供位置的訊息中

可能含有地點語意資訊，卻沒有被有效利用。因此該論文透過數個包含短信息和位置的打卡資料，找出人們的移動模式。例如許多人會先到洋基棒球場，然後去時代廣場，形成一種移動模式。該論文首先假設存在一個 LDA 生成模型，持續訓練直到該生成模型產生的打卡資料和實際資料達到最大相似，找出人們的移動模式。但該論文方法適用的前提為，使用者必須留下打卡及文字資料。

根據上述論文，擷取停留點和標註地點類別，是分析 GPS 軌跡資料的重要處理步驟，停留點的擷取方法常用分群演算法或設定時間距離門檻值找出，本論文將使用設定時間距離門檻值的方法。此外，當距離門檻設定較高時，停留區域內可能有多個 POI，標註地點類別較為困難。而本論文定義停留點必須是使用者在該經緯度座標靜止不動，因此將採用較低的距離門檻值，進而較明確的標註出停留地點類型。

2.1.2 軌跡相似度評估

使用者的一段軌跡可以顯示多種資訊，就個人而言，如果相似的軌跡一直重複出現，可以根據地點屬性推斷使用者的習慣或是偏好；而如果兩個使用者的軌跡具有語意相似性的話，就顯示出使用者彼此間可能有相同的偏好或興趣。當新使用者出現的時候，也能夠透過尋找是否有和他相似的使用者之歷史記錄來進行參考。此外，把相似性高的使用者或序列軌跡資料分在同群組，可建立不同特性的群組模型，因此本論文將透過計算使用者或是序列間的相似度將軌跡資料進行分群。

論文[10]使用所有使用者的停留點資料，根據其地理區域進行階層分群，每層的群內包含數個停留點，再根據使用者的停留點序列資料建立群到群的有向圖。每個有向圖中，每個節點(群)代表使用者去過的地理區域，而邊則代表使用者的到訪順序。該論文不只考慮地理區域，也考量拜訪順序。兩個使用者的位置歷史中，若有愈長的到訪地理區域子序列發生重疊，表示這兩個使用者愈相似。

論文[15]認為[10]只考慮了地理區域交疊性，這樣子的分析並不全面，雖然人們所去的實際地點不同，但若兩個地點的類別是相同的，也可以推斷是有相似的興趣，所以作者提出以到訪地點的語意類別評估者用者的相似性。該論文從 GPS 歷史軌跡資料找出停留點，藉由 POI 地點類型的資料庫搜尋附近範圍內的 POI 類型，並用特徵向量表示每個停留點附近區域內 POI 類型的分佈，形成特徵向量將停留點分群。接下來把每個群代表一種抽象的語意區域，將使用者的停留點序列轉換成群組序列。最後再透過比對兩個使用者的群組序列之最大共同子序列，計算出使用者間的相似度。由於該方法不是考慮實際地理位置，因此即使使用者在不同的城市，仍然可以計算使用者間的相似度。

由上述兩個方法，論文[10]以地理位置來表示停留點的特徵，論文[15]則利用附近區域中 POI 類型分佈的特徵向量來表示。本論文探討的使用者意圖，應該要明確知道使用者的停留點座標，準確標註出停留點的地點類別，因此採用[15]的處理概念，但縮小搜尋範圍，並訂定出只決定一個地點類型的方法。

2.2 位置預測技術

近年來有許多研究將位置預測應用在行動位置服務 LBS(Location-Based Service)上，透過行動業者的無線電通訊網路或外部定位方式例如 GPS，來取得行動終端用戶的位置訊息（地理坐標）。此服務可辨認查詢一個人或物的位置，例如尋找最近的提款機或朋友同事的目前位置，也能透過客戶目前所在的位置提供地點推薦系統[2]，提供該地點使用者可能需要的資訊[1]等服務。而要提供此服務的前提，就是要有良好的位置預測技術。過往的研究在地點預測[11][12][14]通常採用傳統的機率模型，然而效果並不顯著，要準確預測出地點並不容易。

和準確的地點預測相比，許多研究認為能夠預測地點的類別才是重要的，因為地點類別能夠明確地顯示使用者的興趣、偏好、習慣，甚至是可能的行動，例如資料顯示一個使用者經常去西餐廳這個類別，雖然都是屬於餐廳，但和中餐廳或是路邊小吃的類別是有所區別的，因此可以從中知道使用者的喜好。

地點類別預測方法的研究，大多以社群媒體資料作為研究資料。論文[8]以打卡資料作為輸入，不過資料中只具有經緯度座標，因此該論文使用 FourSquare API 自動標註出地點座標所屬類別。該論文以建立條件機率模型的方式，將上一個地點類別與下一個視為一個組合，計算兩兩組合的條件機率。當給定一個地點類別後，帶入模型就可以預測下一個可能去的地點類別機率分佈。此模型只考慮前一地點，且機率模型較簡單，所以預測效果並不理想，因此論文[16]提出 HMM(Hidden Markov Model)機率模型來預測使用者下一個會去的地點類別。該論

文不只考慮走訪的前一個地點，而是考慮先前走訪過的地點類別而形成的序列，相較於論文[8]準確率有明顯的提升。

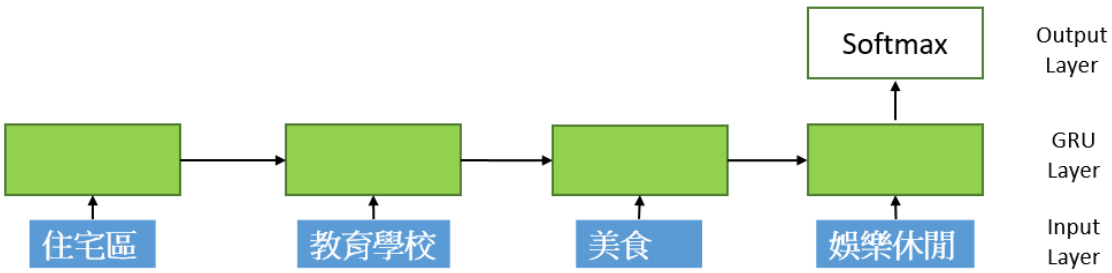

圖 2.1 RNN-based 地點類型預測模型架構[13]

除了採用機率模型，遞迴類神經網路經常被應用在處理序列的預測問題[9][13][17]。論文[13]將遞迴類神經網路用在地點類別預測，且驗證此方法優於傳統的預測方法。該論文以社群媒體地點資料，將地點類型依據時間前後排列形成序列，建立遞迴類神經網路 GRU 模型來預測使用者下一個會去的地點類型，如圖 2.1 所示。此外，該論文將資料以國籍和性別進行分群後，探討不同群組間的預測效果。

本論文採用的遞迴類神經網路基本架構和[13]相似，但因資料來源不同，無法以國籍和性別為單位進行分群，因此本論文提出新的分群方法，並進一步考慮整體資料模型和群組模型的組合方式。此外，模型的輸入資料也使用更多特徵，並增加隱藏層提升預測效果。

第三章　問題定義與系統架構

3.1 問題定義

使用者的 GPS 軌跡由一連串的 GPS 紀錄點組成，其中每個記錄點包含經度(Longitude)、緯度(Latitude)、日期(Date)以及時間(Time)，如以下定義 1。

[定義 1] 使用者 u 的 GPS 軌跡： $Tra_u = P_1 \rightarrow P_2 \rightarrow \ldots \rightarrow P_l$，

其中 $P_i = (Latitude、Longitude、Date、Time)(i = 0,1,\ldots,l)$。

本論文考慮透過 GPS 軌跡記錄，預測使用者下一個停留的地點類型。

[問題定義]透過 GPS 軌跡記錄，預測使用者 u 下一個停留的地點類型

對於一個使用者 $u \in U$ 的 GPS 軌跡資料 $Tra_i \in Tra_u$，本論文的目標是預測使用者在固定時間內下一個停留地點類型 c_i 的出現機率，其中 $c_i \in C$，而 C 為一個給定的地點類型集合。

3.2 系統架構與流程

論文方法的處理分為兩個階段：離線訓練以及線上預測。

3.2.1 離線訓練

離線訓練的處理主要分為三部分:(一)資料前處理及特徵擷取、(二)分群方法、及(三)遞迴類神經網路學習架構(RNN based Learning Network)，如圖 3.1 所示。

(一) 資料前處理及特徵擷取：包含停留點擷取、地點類型自動標註、及輸入資料產生方法。

(二) 分群方法：分成以使用者為單位的分群方法，及以序列為單位的分群方法

(三) 遞迴類神經網路學習架構(Rnn based Learning Network)：分成遷移學習模型(Transfer Learning Model)和合成模型(Ensemble Model)。

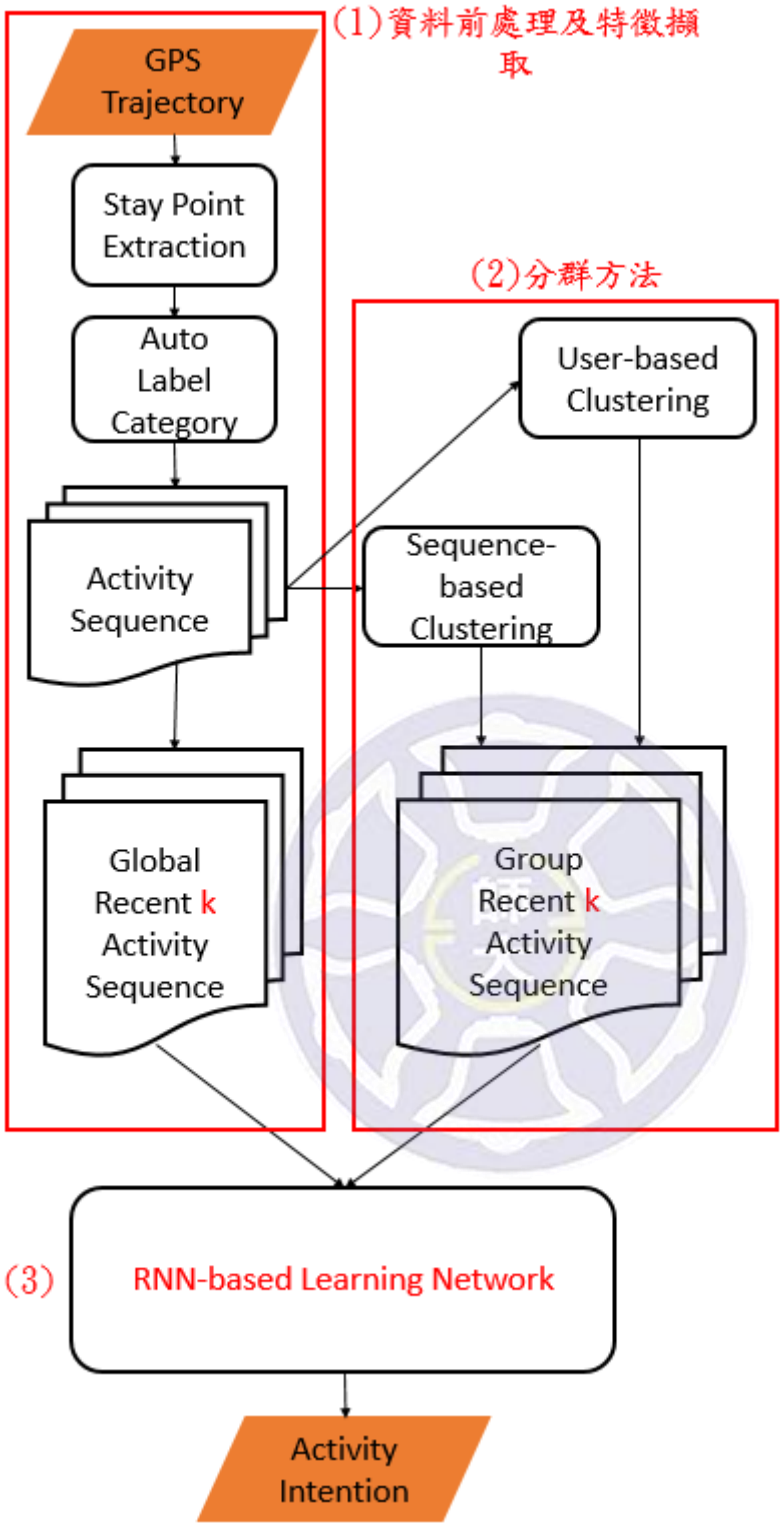

圖 3.1 使用者活動意圖預測系統離線訓練之架構

離線訓練的處理流程簡述如下：

(一) 資料前處理及特徵擷取

從 GPS 軌跡資料中計算出停留點並自動標註地點類型，形成使用者的活動序列。根據需求將每一個活動序列切分，形成長度為 k 的使用者活動序列作為模型訓練。

(二) 分群方法：

本論文提出兩種分群方法：

(1) 使用者分群法(User-based Clustering)

以使用者為單位進行分群，將每兩個連續的停留點子序列視為一組移動模式，將兩兩使用者出現的移動模式所成的集合進行相似度計算，根據相似度採用階層式分群演算法進行分群。

(2) 序列分群法(Sequence-based Clustering)

以序列為單位進行分群，將長度 k 的活動序列兩兩進行相似度計算，根據相似度採用階層式分群演算法進行分群。

(三) 遞迴類神經網路學習架構(RNN based Learning Network)

本論文提出三種遞迴類神經網路的學習架構，進行使用者活動意圖預測：

(1) 全體資料模型/群組模型(GRU Global/Group Model)

以未經分群過，全部長度 k 的活動序列所訓練出的模型稱為全體資料模型；以各分群資料所訓練出的模型統稱為群組模型。

(2) 遷移學習模型(GRU Transfer Learning Model)

遷移學習模型將全體長度 k 的活動序列作為輸入訓練好模型後，將參數記錄下來，並將該參數預設為未訓練模型的初始參數，然後對每個分群，採用該分群的資料輸入調整各群組模型，其中 GRU 層和隱藏層的參數將不再被訓練及改動，此模型稱為遷移學習模型。

(3) 合成模型(GRU Ensemble Model)

合成模型將第一種架構提出的全體資料模型和群組模型所預測的結果，透過調和參數(Ensemble Parameter)學習由前兩個模型預測結果的組合比重，並做 Softmax 處理。最後由原本全體資料模型和群組模型的預測結果各自乘上組合比重值後相加，得到各活動意圖的機率預測結果。

3.2.2 線上預測

線上預測的處理主要分為兩部分：(一)資料前處理及特徵擷取、及(二)模型選擇方法，如圖 3.2 所示。

(一) 資料前處理及特徵擷取：包含停留點擷取，自動標記地點類別及切割序列產生輸入資料方法。

(二) 模型選擇方法：使用者群組模型和序列群組分群模型有兩種不同的適用方法。

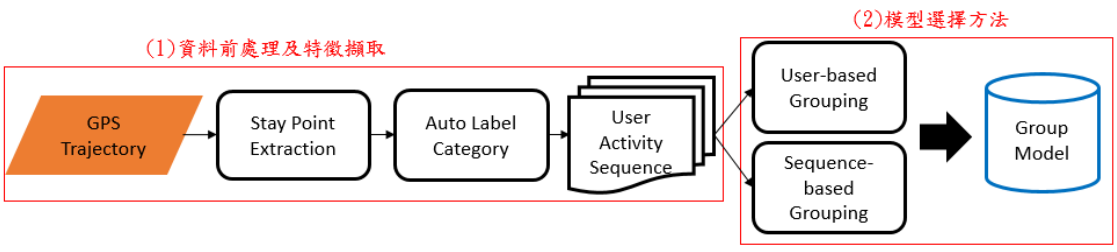

圖 3.2 使用者活動意圖預測系統線上預測之架構

線上預測的處理流程簡述如下：

(一) 資料前處理及特徵擷取

從 GPS 軌跡資料中計算出停留點並自動標註地點類型，形成使用者的活動序列，擷取最近長度為 k 的使用者活動序列作為預測模型輸入。若長度不及 k 則需要繼續蒐集資料。

(二) 模型選擇及預測

判斷測試資料應適用哪個群組模型。根據分群法的不同有以下兩種判斷方法：

(1) 使用者群組模型選擇方法

將使用者和各群的使用者兩兩進行相似度計算，取相似度最大值的使用者所屬分群，決定測試資料應採用的群組模型。

(2) 序列群組模型選擇方法

將序列和各群的序列兩兩進行相似度計算，取相似度的最大值的序列所屬分群，決定測試資料應使用的群組模型。

第四章　資料前處理和特徵擷取

4.1 軌跡資料格式與名詞定義

本論文使用研究資料為 GPS 軌跡，資料格式如表 4.1 所示。對於一個使用者，GPS 記錄器會在固定時間間隔記錄下使用者所在經緯度、日期以及時間。經過時間的累積，這些紀錄點就會形成使用者的 GPS 軌跡。

以下本論文將定義研究中使用的名詞。

[定義 2] 停留點：*在一個地點停留超過 t 分鐘的經緯度座標稱為停留點。s = (latitude, longitude, arrive time, staying time)。*在本研究中，t 設為 15 分鐘，latitude 和 longitude 分別表示緯度和精度，arrive time 表示使用者從該時間開始停留，staying time 表示使用者在該地點的停留時間。。

[定義 3] 位置歷史序列：*將一個使用者 u 的所有停留點自動標注類別以後，根據時間先後排序而成的序列稱為位置歷史序列。*$LH = S_1 \rightarrow S_2 \rightarrow \ldots \rightarrow S_k$，$S_i$ = *(Category, arrive time, staying time)。*其中 Category 表示該停留點的 POI 類別，arrive time 表示使用者從該時間開始停留，staying time 表示使用者在該地點的停留時間。

[定義 4] 使用者活動序列：*在使用者位置歷史序列中，若兩個連續停留點間的時間間隔超過門檻值時，將序列切開後形成的多個子序列，稱為使用者活動序列。*

[定義 5] 長度 k 的活動序列：*由使用者的活動序列，取出所有長度 k 的連續子序列稱為長度 k 活動序列。*

[定義 6] 長度 k 活動序列之活動意圖預測問題：

$$P_{next}(c_i) = P(S_{k+1}, Category = c_i | S_1 \rightarrow S_2 \rightarrow \ldots \rightarrow S_k)$$

以長度 k 活動序列 $S_1 \rightarrow S_2 \rightarrow \ldots \rightarrow S_k$ 為輸入，預測使用者下一個停留的地點類型為 c_i 的機率，其中 $c_i \in C$，C 為一個給定的地點類型集合。

Latitude	longitude	Date	Time
39.984702	116.31841	2008-10-23	02:53:04
39.984683	116.31845	2008-10-23	02:53:10
....			

表 4.1 GPS 軌跡資料格式

以下將介紹輸入給遞迴類神經網路前的資料處理方式：主要分成兩個步驟：(1)擷取停留點及(2)自動標註停留點類別，以下小節將分別說明這兩個步驟。

4.2 停留點擷取方法

由於 GPS 軌跡的特性，就算在定點位置上固定不動時，GPS 紀錄器所記錄的座標仍可能有所偏移，因此一個停留點會對應到一段連續的 GPS 軌跡 $Tra_u = Pm \rightarrow Pm+1 \rightarrow \ldots \rightarrow Pn$，其必須滿足以下三個條件：

（1）對於所有 $m < i < n$ 都必須滿足 $Distance(Pm, Pi) < \theta_d$，其中$\theta_d$為距離門檻值。

（2）$Distance(Pm, Pn+1) > \theta_d$其中$\theta_d$為距離門檻值。

（3）$Time(Pm, Pn) > \theta_t$，其中θ_t為時間門檻值。

則停留點 s 的屬性計算方式包括以下三個：

(1) 座標：(Pm + Pm+1 + … + Pn)/|P|，其中|P|表示走訪過的記錄點個數。

(2) 時間：Pm.t，以起始記錄點 Pm 的時間 Pm.t 代表停留點 s 的進入時間。

(3) 停留時間：Pn.t–Pm.t，表示以記錄點結束的時間 Pn.t 和記錄點起始的時間 Pm.t 差為停留時間。

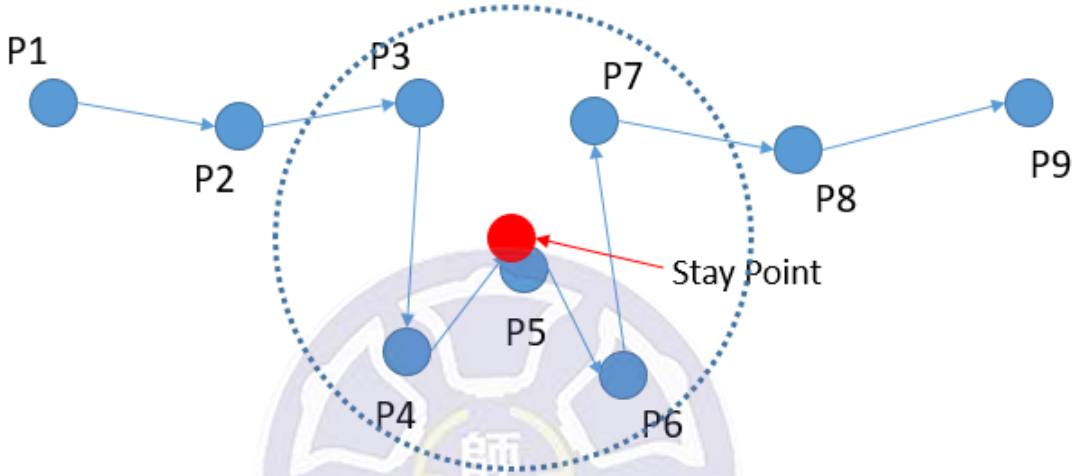

圖 4.1 GPS 軌跡和停留點示意圖

本論文設定容許的停留點誤差範圍為 20 公尺，也就是把θ_d設為為 20。θ_t則設為 15 分鐘，表示停留至少 15 分鐘才代表使用者在這個地點停留。

停留點擷取範例如圖 4.1 所示，有一個 GPS 軌跡中包含 P3→P4→P5→P6→P7，其中 P3 到 P4、P5、P6 或 P7 的距離都在設置距離門檻值θ_d以內，P3 到 P8 的距離則超過θ_d。若 P3 到 P7 的時間，間隔大於時間門檻值θ_t以上，就取出一個停留點，停留點經緯度座標由所經過之記錄點 P3 到 P8 的經緯度座標平均值決定。接下來再從下一個 GPS 記錄點 P8 開始往下計算，直到取出所有停留點為止。

4.3 自動標註停留點類別

本論文透過前一小節的方法，可以擷取出使用者所有的停留點，然而停留點只是一個經緯度座標，表示一個地理位置，並不知道使用者停留在什麼類型的地點，因此本論文透過騰訊位置服務 WebService API 中的逆地址解析系統，自動標註出停留點的 POI 和 POI 類型。

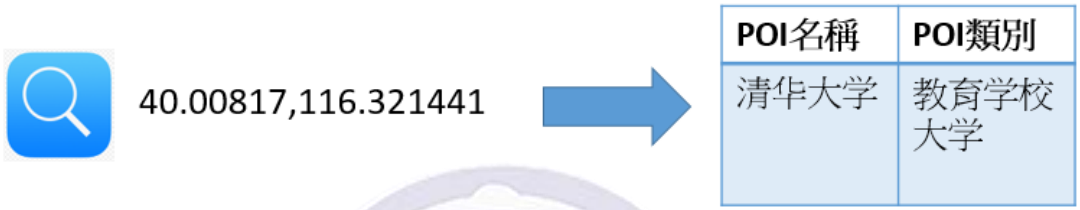

圖 4.2 騰訊位置服務 WebService API 逆地址解析示意圖

騰訊位置服務 WebService API 逆地址解析示意圖如圖 4.2 所示，當輸入一個經緯度座標，這個 API 系統會回傳距離該座標設定距離內的 POI 和 POI 類型，在此設為 20 公尺。由於回傳的 POI 可能不只一個，因此本論文使用以下四個規則依序判斷來自動標註出停留點所屬類別：

(1) 若回傳一個 POI，則以該 POI 類別標註該停留點。

(2) 若回傳多個 POI，取這些 POI 出現次數最多的 POI 類別來標記該停留點。

(3) 若回傳多個地點，且有多個類別出現的次數相同，則以距離停留點最近的 POI 類別來標記該停留點。

(4) 若無回傳，在範圍內找不到 POI，則將該停留點刪除。

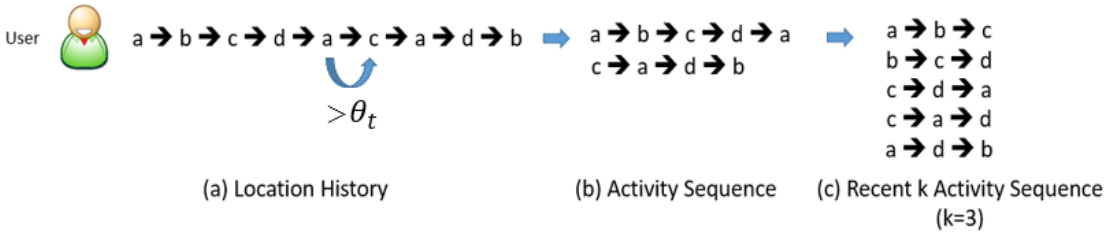

圖 4.3 輸入序列處理流程示意圖

自動標註停留點類別後就可得到每個使用者的位置歷史序列，如圖 4.3(a)所示。對一個使用者的位置歷史序列中，每兩個停留點之間的時間間隔若超過門檻值θ_t，則切開形成另一個序列，所形成的多段序列稱為使用者活動序列。一個使用者可能有多段活動序列，如圖 4.3(b)所示，本論文將θ_t設為 1 天。

最後根據模型訓練的設定，從使用者活動序列當中取出固定長度 k 的序列當作輸入資料，稱為使用者長度 k 的活動序列，如圖 4.3(c)為長度 k=3 的活動序列。

第五章　分群方法

本論文提出了兩種不同單位的訓練資料分群方法，當作建立各群組模型的依據，另外，要進行預測時，必須決定採用哪個群組模型。本章節將分別介紹兩種分群方法以及群組模型選擇方法。

5.1 使用者分群法(User-based Clustering)

使用者分群法是計算出使用者彼此的相似度對使用者進行分群，分到同一群使用者的所有活動序列，即形成同一個群組模型的訓練資料。使用者分群法 User-based Clustering 分成三個處理步驟：(一)找出使用者的轉移模式(Transition Pattern)、(二)計算兩兩使用者的雅卡爾相似度(Jaccard Similarity)、(三)階層式分群法(Hierarchical Clustering)。

(一) 找出使用者的轉移模式(Transition Pattern)

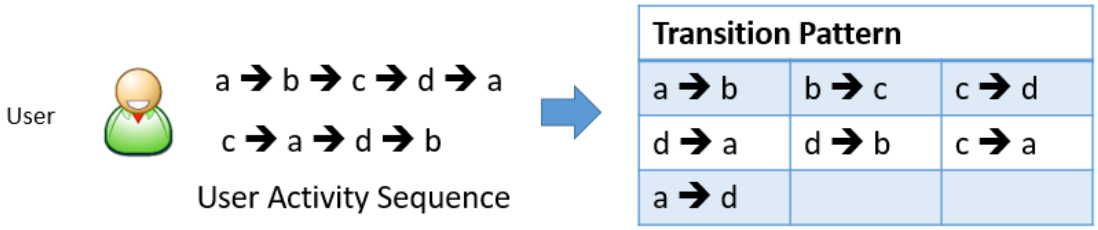

圖 5.1 找出使用者的轉移模式

在使用者的活動序列中，我們稱在固定時間內從一個地點移動到另一個地點的行為為一個轉移模式，轉移模式可以顯示出使用者去過哪些類型的地點及移動習慣。因此，本論文將使用者活動序列，以每兩個連續的停留點型成一個轉移模

式，如圖 5.1 所示，將這些轉移模式所成的集合 TP_u 用來代表一個使用者 u 的活動特徵。此集合表示可接受一個元素出現一次以上，為一個多集合(multi set)表示法。

(二) 計算兩兩使用者的雅卡爾相似度(Jaccard Similarity)

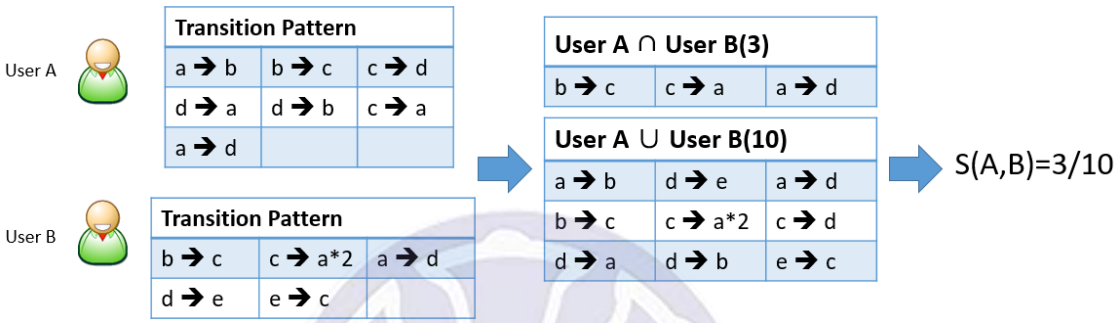

圖 5.2 雅卡爾相似度範例 1

當所有使用者都取出轉移模式之後，本論文計算兩兩使用者之活動特徵的雅卡爾相似度。

$$S(x,y) = \frac{TPx \cap TPy}{TPx \cup TPy}$$

(公式 1)

計算方式如公式 1 所示，將兩個使用者 x 及 y 的轉移模式 TPx 及 TPy 取交集和聯集，交集代表兩個使用者之間共同出現的移動模式，聯集則是兩個使用者所有出現的移動模式，如此一來便能夠計算出兩個使用者之間的相似程度，如圖 5.2 所示，舉例說明，使用者 A 的轉移模式有 7 個，使用者 B 有 6 個，交集的轉移模式有 3 個，聯集的交集模式有 10 個，則兩個使用者的相似度為 3/10。

(三) 階層式分群法(Hierarchical Clustering)

階層式分群法(Hierarchical Clustering)透過階層架構的方式，將資料層層反覆地進行分裂或聚合，產生一個的樹狀結構來組織資料的群組。本論文採用聚合式階層分群演算法(Agglomerative Hierarchical Clustering)中的完整連結聚合演算法(Complete-linkage Agglomerative Algorithm)，群聚間的相似度定義為兩個群聚中最不相似兩筆資料間的相似度，如此能夠保證在每一個群聚當中，兩兩資料間的相似度都大於此相似度。分群法會逐步聚合，直到下一步驟不再合併小群組而是把數量大的群組聚合為止，取當前步驟的群組分群結果。

5.2 序列分群法(Sequence-based Clustering)

序列分群法，則採用長度 k 的活動序列為單位進行分群。然而，長度 k 活動序列的數量非常多，若兩兩長度 k 活動序列都要計算相似度極為耗時，因此本論文進行兩階段分群：第一階段先採用抽樣的方式先對抽樣出來的長度 k 活動序列進行分群，第二階段再將其他資料指定到以找出的分群中。序列分群法主要分成以下四個處理步驟：(一)序列抽樣 (二)計算抽樣序列相似度 (三)第一階段分群 (四)第二階段分群。

(一) 序列抽樣

從所有長度 k 活動序列中抽樣 2000 筆序列進行第一階段分群處理處理，約佔全部資料的 15%，如此一來能夠大幅降低分群處理的時間。

(二) 計算序列相似度

使用長度 k 活動序列進行分群，若使用轉移模式，會因序列長度不夠長，造成序列間的相似度普遍很低，因此本論文在序列分群法另外採用以下兩種相似度的算法：(1) 地點類型集合的雅卡爾相似度(Jaccard Similarity)、(2) 序列的最長共同子字串相似度(LCS(Longest Common Sequence) Similarity)。

(1) 地點類型集合的雅卡爾相似度(Jaccard Similarity)

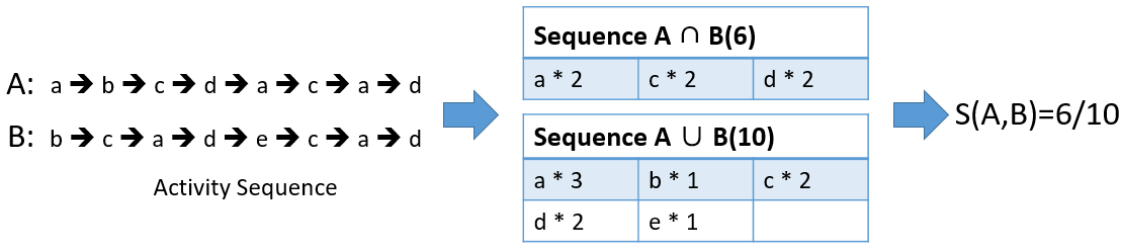

圖 5.3 雅卡爾相似度範例 2

將長度 k 活動序列的地點類型所成的集合作為該序列的特徵，計算兩兩序列特徵的雅卡爾相似度。舉例說明，序列 A：a→b→c→d→a→c→a→d 和序列 B：b→c→a→d→e→c→a→d 的交集有 6 個，聯集有 10 個，則兩個序列的相似度為 6/10，如圖 5.3 所示。

(2) LCS(Longest Common Sequence) Clustering

$$S(x,y) = \frac{|LCS(x,y)|}{k}$$ (公式 2)

第二種方法是找出兩個序列當中的最大共同子序列，如公式 2 所示，|LCS(x,y)| 代表共同子序列的長度，k 為活動序的長度。舉例說明，序列 A：a→b→c→d→a→c→a→d 和序列 B：b→c→a→d→e→c→a→d 比對之後，最大共同子序列為 b→c→d→c→a→d，長度為 6，序列 AB 為長度 8 個活動序列，因此兩個序列的相似度為 6/8，如圖 5.4 所示。

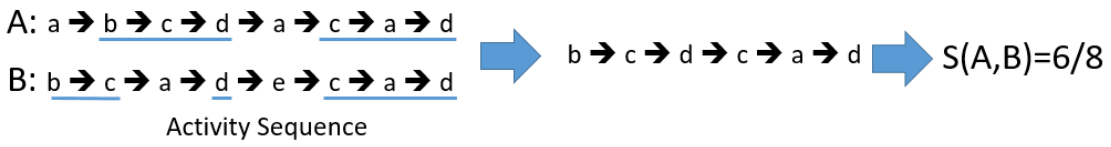

圖 5.4 LCS 相似度範例

在序列分群法當中，使用雅卡爾相似度並沒有考慮使用者走訪地點類型的前後關係，只考慮使用者去過了哪些地方，而 LCS 則會考慮走訪地點類型的前後順序。

(三) 第一階段分群

計算好抽樣序列兩兩的相似度之後，採用完整連結聚合演算法(Complete-linkage Agglomerative Algorithm)將抽樣序列進行分群。分群法會逐步聚合，直到下一步驟不再合併小群組而是把數量大的群組聚合為止，取當前步驟的群組分群結果。

(四)第二階段分群

第二階段分群是要將其餘未被抽樣出的長度 k 活動序列序列一一分到群組中。對每一筆上未被分群的序列 s，會和每一群組中的所有序列計算相似度，根據不同相似度計算方式分出來的群組分別適用該相似度計算方法，跟每一群組 Gi 中所有序列算出的相似度最最小值當作 s 跟 Gi 的相似度，將序列 s 分到跟他相似度最高的那一群組中。

另外，為了確保每個群內的序列之相似度達到一定基準，因此會設立最小相似度門檻值，若一個序列到每個群組的相似度皆沒有達到門檻值的話，則會自己獨立出來型成一個新的群組。下一筆序列會和第一階段分群所有的群和新群計算相似度決定其群組，以此方式處理，直到所有長度 k 活動序列都被分完群為止。

5.3 群組模型選擇方法

測試資料集中若有一個新的使用者或是最近長度 k 活動序列需要進行預測時，必須判斷應該適用哪個群組模型，本論文提出群組模型選擇方法如下：

(一) 使用者群組模型選擇方法

當有一個新的使用者 u 的資料要套用模型進行預測時，必須先進行前處理取出其活動序列。再取其轉移模式 TPu，然後和訓練資料中每一群組中的每個使用者進行雅卡爾相似度計算，取 Gi 中和 u 的最大相似度作為該群的適用分數，取和 u 有最大適用分數的群組模型作為適用模型。建立分群模型必須使用較嚴格的方式來對訓練資料進行分群，才能達到群組模型的效果。而當資料要進行預測時，只要取預測資料最相似的訓練資料為其參考基準，因此取使用者 u 在每一群算出的最大相似度值來做為該群對使用者 u 的適用分數。

(二) 序列群組模型選擇方法

當最近的長度 k 活動序列資料要套用模型預測時，則直接和每個序列群組做相似度計算，可分別採用再 5.2 提出的兩種相似度計算方法，再以與各群組算出的最大相似度最為該群組模型的適用分數，從中選出適用分數最高的模型套用。

第六章　活動意圖預測

關於使用者活動意圖預測模型建立方法，本論文採用遞迴類神經網路架構，。本論文實作採用深度學習系統 Tensorflow 的 Keras 工具建立各模型，本章節將針對在 3.2.1 提到的三個模型架構詳細說明。

(一) 全體資料模型和群組模型(GRU Global/Group Model)

5.1 將詳細說明全體資料模型和群組模型採用的網絡架構設計，並針對各個層(Layer)進行詳細說明：包括所採用之遞迴層、隱藏層和輸出層設計，並說明每層的參數設定。

(二) 遷移學習模型(GRU Transfer Learning Model)

5.2 將說明遷移學習模型採用的網絡架構設計，其目標為根據全體資料模型，在採用群組資料調整部分模型係數。

(三) 合成模型(GRU Ensemble Model)

5.3 將說明合成模型的網絡架構設計，並針對各層(Layer)進行說明。其目標為結合全體資料模型和群組模型中的預測結果，讓效果更好。

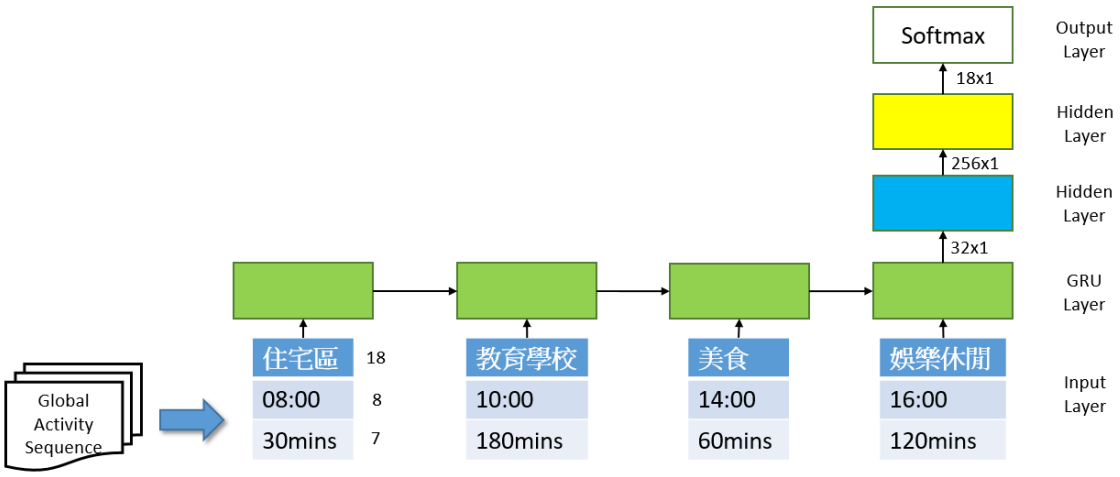

圖 6.1 全體資料模型

6.1 全體資料模型和群組模型

全體資料模型如圖 5.1 所示，本架構會採用全部使用者的序列資料作為模型訓練模型。

(一) 輸入層

輸入資料皆以 one-hot 表示：一筆活動序列由 k 個停留點組成，每個停留點包含的屬性有：地點類型、進入時段及停留時間。地點類型共分成 18 個類別，因此會有 18 個維度；進入時段將一天 24 小時以每 3 小時為單位分成 8 個時段，因此會有 8 個維度；停留時間則以每半小時為單位，從 30 分鐘以內至 180 分鐘以上分成 7 種停留時間，因此會有 7 個維度。因此，輸入序列的每個時間點輸入資料維度為 33 維，長度 k 的一筆活動序列維度為：k*33。

(二) GRU(Gate Recurrent Unit)層

GRU 是遞迴類神經網路的一種，和 LSTM 一樣，為了解決長期記憶和反向

傳播中的梯度下降等問題而被提出來。本論文選擇使用 GRU 是因為其效果和 LSTM 相似，且訓練參數較少，計算更有效率。

一個 GRU 層中的神經元單位如圖 5.2 所示，對應到本問題中，x_t 表示當前輸入的停留點輸入屬性質，h_{t-1} 為上一階段的輸出結果，h_t 則為當前的預測結果。GRU 透過控制閘門來調整訊息的去留，包含重設閥(reset gate) r_t 和更新閥(update gate) z_t，如公式 2 和公式 3 所示。

$$r_t = \sigma(W^r x_t + U^r h_{t-1}) \quad \text{(公式 2)}$$

$$z_t = \sigma(W^z x_t + U^z h_{t-1}) \quad \text{(公式 3)}$$

接下來候選隱藏層(candidate hidden layer)$\tilde{h}_t$ 的計算如公式 4，其中 r_t 用來控制需要保留先前多少記憶，若 r_t 設為 0，$\tilde{h}_t$ 就只包含當前的訊息。最後 z_t 控制從前一時間的隱藏層 h_{t-1} 中保留多少訊息，h_t 就是當前的輸出結果，如公式 5。

$$\tilde{h}_t = \sigma(W x_t + r_t U h_{t-1}) \quad \text{(公式 4)}$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t \quad \text{(公式 5)}$$

本論文將 GRU 層的輸出 h_t 設定為 32 維度的向量，輸入向量為 k*33，因此訓練的參數數量為(k * 33 + 32) * (32 * 3) + (32 * 3)，其中最後的 32*3 為偏差值(bias)。

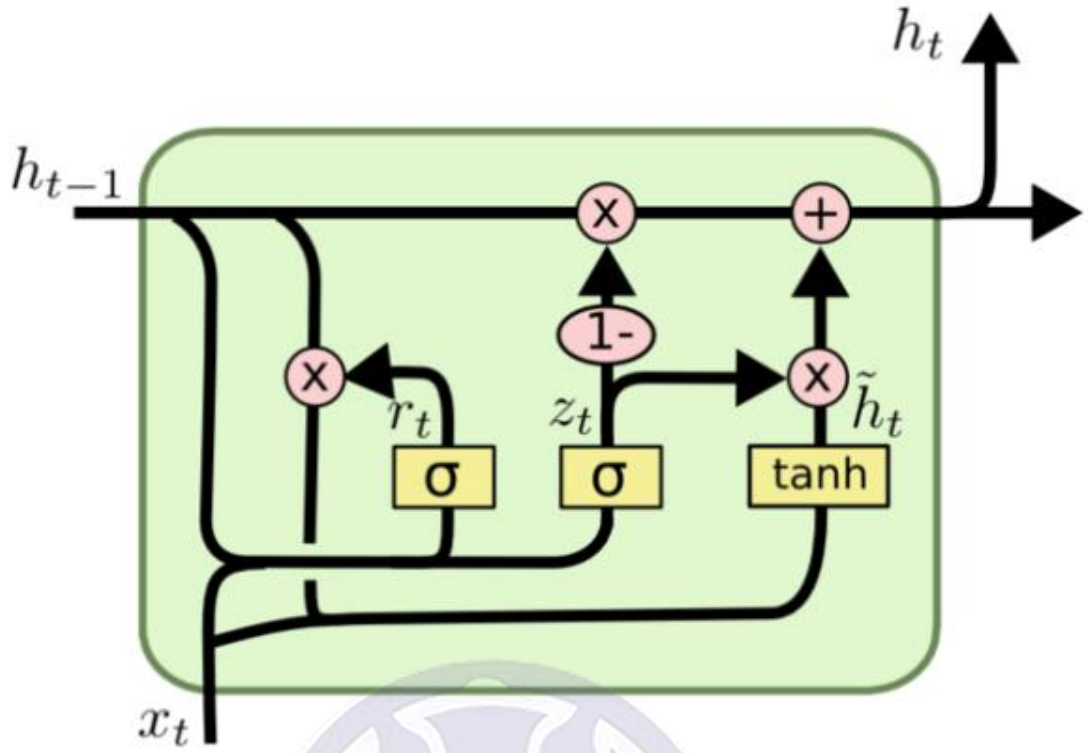

圖 6.2 GRU 架構流程

(三) 隱藏層及輸出層

本論文在輸出層前採用兩回合的隱藏層，第一回合隱藏層使用 256 個神經元，與 GRU 層輸出的 R^{32x1} 特徵向量接上，第二回合的隱藏層則採用 18 個神經元將與第一回隱藏層的 256 個神經元接上。第二回合的隱藏層輸出結果用來預測使用者活動意圖的 18 種類型，因此透過激活函數 Softmax 處理後，將 18 種活動意圖類型以機率分布呈現。預測結果時將機率值由大而小排序，取機率值最大的前 k 個活動意圖當作預測結果。

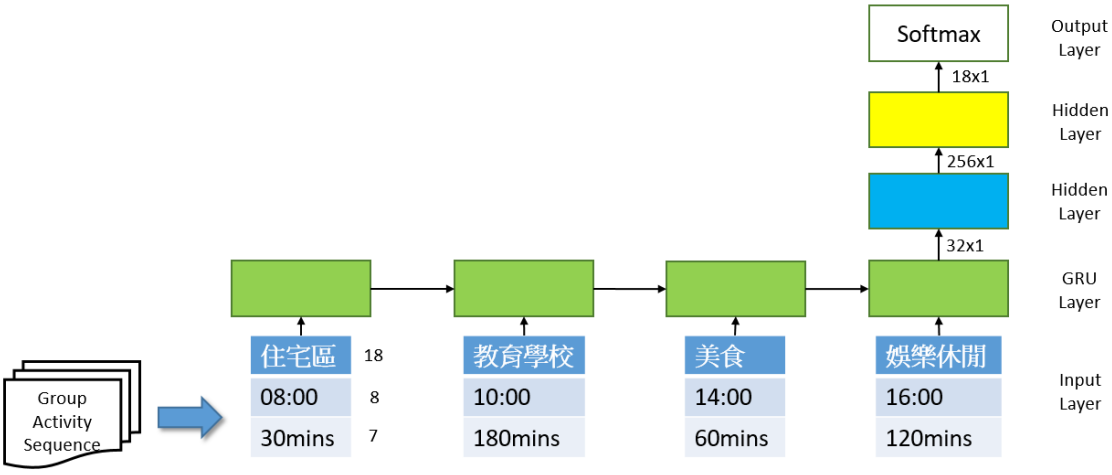

圖 6.3 群組模型

群組模型如圖 6.3 所示，群組模型和全體資料模型所使用的架構相同，差別只是輸入資料不同，透過分群後的各個群組資料訓練出不同的群組預測模型。

本論文訓練模型時，採用的損失函數(loss function)設定為 Keras 提供的多元交叉熵(categorical cross entropy)，如公式 6 所示。

$$\sum_i \sum_t (y_{i,t} * \log(\hat{y}_{i,t})) \text{(公式 6)}$$

其中$\hat{y}_{i,t}$表示預測結果，$y_{i,t}$表示真實的意圖類型，t 表示意圖類型數目，i 為樣本數。此外，為了避免發生過擬(overfitting)，本論文在 GRU 層和第一個隱藏層後採用 Dropout 模組，將每個神經元的輸出結果隨機設置為 0，本論文設定的機率值為 0.35。

6.2 遷移學習模型

同一筆資料透過全體資料模型和群組模型所產生的預測可能有差異，在全體資料模型預測正確但在群組模型預測錯誤，或相反的情況。因此本論文提出了兩種組合模型架構，希望綜合兩個模型的結果達到更好的預測效果，遷移學習模型為第一種架構。

全體資料模型和群組模型最大的不同就是全體資料資料量大，學到一般性的模型，群組資料量較小，但較可能學到符合群組特性的模型。因此，本論文先用全體資料進行訓練，將訓練好的模型參數記錄下來，如圖 6.4(a)所示。接下來將全體資料訓練好的參數設置為初始參數，在 GRU 層和第一個隱藏層將 trainable 變數設置為 false。因此，遷移學習模型採用群組資料對第二個隱藏層進行參數訓練，讓模型訓練更偏向群組資料的特性。最後一樣透過激活函數 Softmax 處理後，將 18 種活動意圖預測結果以機率分佈呈現。預測結果時將機率值由大而小排序，取機率值最大的前 k 個活動意圖當作預測結果。

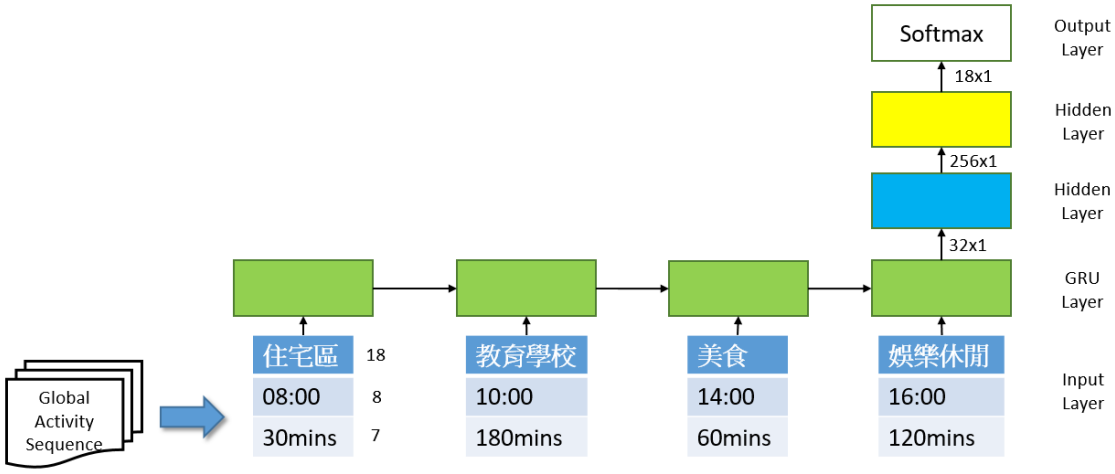

圖 6.4(a) 遷移學習模型第一階段

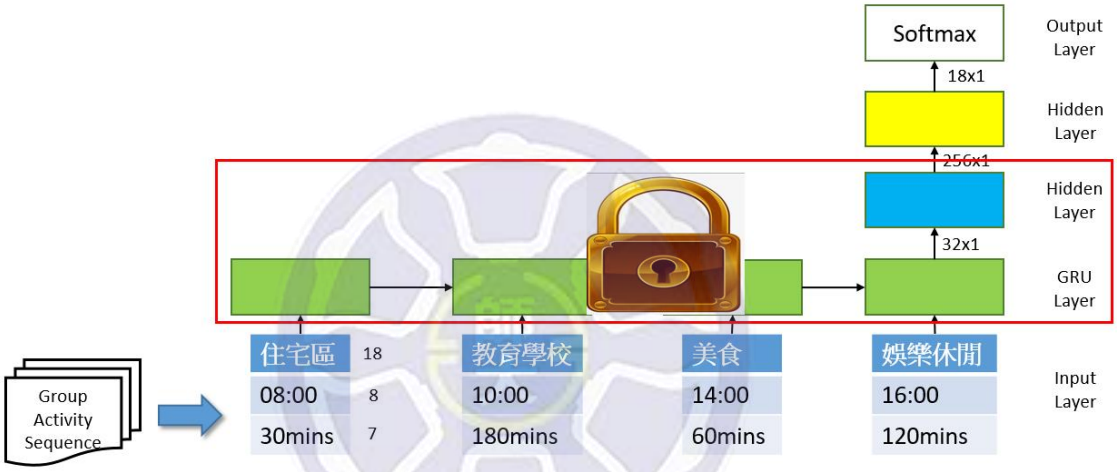

圖 6.4(b) 遷移學習模型第二階段

6.3 合成模型

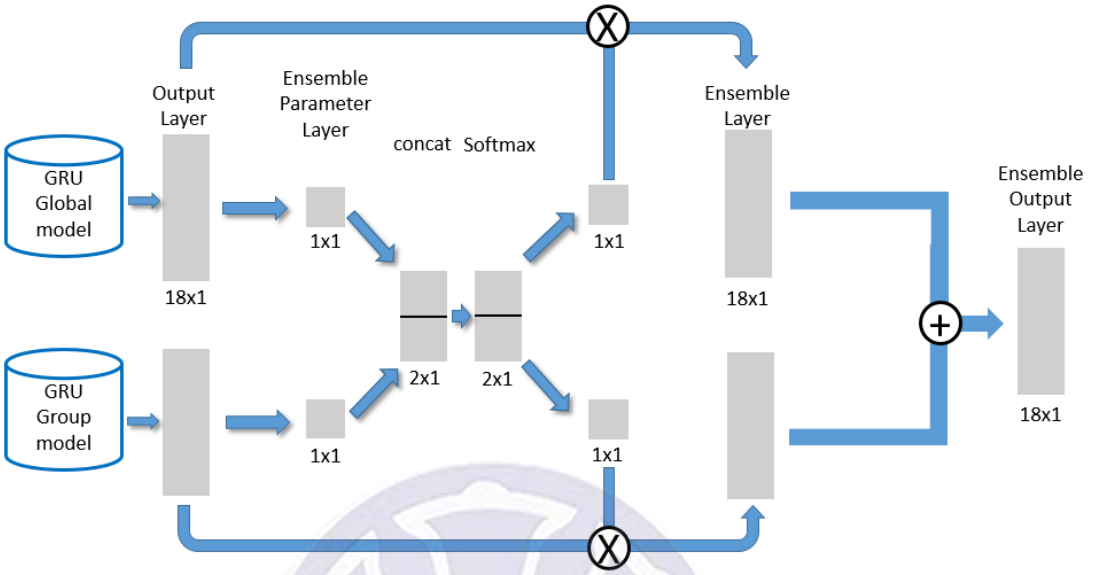

圖 6.5 合成模型

合成模型是本論文提出的第二種組合模型，將全體資料和群組資料，分別輸入各自的模型，如圖 6.1 和圖 6.3 所示，各自建立預測 18 種意圖類型的預測模型，結果以 Softmax 呈現，再將全體資料模型(GRU Global Model)和群組模型(GRU Group Model)的輸出結果當作合成模型的輸入資料。

合成模型的想法是學習一個調和參數，以一個隱藏層連接到一個神經元，如圖 6.5 所示，訓練投射參數並計算出組合比重值，用來將兩個模型的預測結果進行比重合成。當調和參數學習完成後，以激活函數 Softmax，以確保兩個組合比重值總和為 1。

因此在此模型中，會先套用全體資料模型和適用的群組模型，各自預測出 18

種活動意圖類型的機率分布值，以其訓練好的組合比重值相乘後加總，將 18 種活動意圖預測結果以機率分佈呈現。預測結果時將機率值由大而小排序，取機率值最大的前 k 個活動意圖當作預測結果。

第七章　實驗結果及探討

本論文依系統提出的不同模型建立方式，將實驗分為三部分進行評估：

(一) 全體資料模型(GRU Global Model)之效果評估

(1) 評估特徵及其組合之預測效果

(2) 模型參數設置實驗

(二) 群組模型(GRU Group Model)之效果評估與比較

(1) 使用者分群法(User-based Clustering)預測效果評估

(2) 序列分群法(Sequence-based Clustering)預測效果評估

(3) 群組模型選擇及預測效果評估

(三) 組合模型之效果評估與比較

(1) 遷移學習模型(GRU Transfer Learning Model)之效果評估與比較

(2) 合成模型(GRU Ensemble Model)之效果評估與比較

(3) 序列長度影響評估與比較

以下小節將詳細說明實驗資料、評估指標、以及上述三部分的實驗方法及結果。

7.1 資料來源與討論

資料集	GeoLife		OSM(Open Street Map)	
來源	China		World	
使用者數目	91		925	
	訓練資料	測試資料	訓練資料	測試資料
活動序列數目	18157	4549	49897	12480
長度3活動序列	4871	1196	14508	3512
長度5活動序列	4659	1153	12765	2843
長度7活動序列	4400	1116	11889	2974
長度9活動序列	4227	1084	10735	2801
長度k序列總數 (k=3,5,7or9)	22706		62377	

表 7.1 資料集資訊

本論文使用網路上公開的 GPS 軌跡記錄，共兩份資料集。Geolife 資料集是由 182 個使用者從 2007 四月到 2012 八月蒐集而來的 GPS 軌跡(https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/)，Open Street Map 資料集則是從世界各地蒐集使用者的 GPS 軌跡記錄形成的資料集(https://planet.openstreetmap.org/gps/)。

資料集基本統計資訊如表 7.1 所示，經過前處理後，形成長度 k 活動序列，本論文將 k 設定為 3,5,7 和 9 四種。本論文在 Geolife 資料集使用 91 個使用者取出共 22706 個活動序列，其中用於訓練的資料有 18157 筆，用於測試的資料有 4549 筆；OSM 資料集使用 925 個使用者取出共 62377 個活動序列，其中用於訓練的資料有 49897 筆，用於測試的資料有 12480 筆。

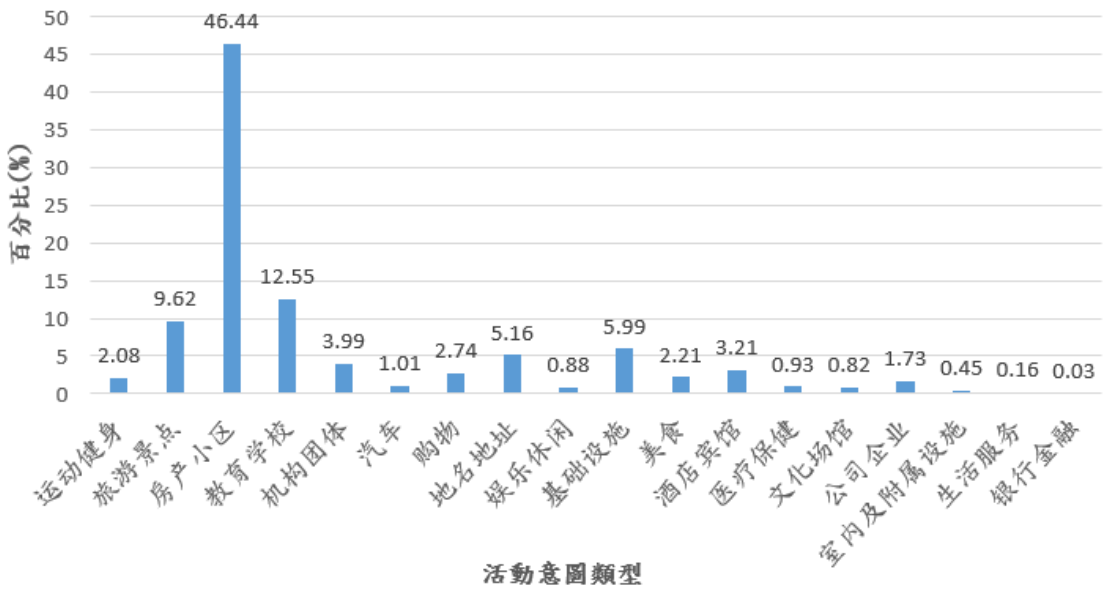

圖 7.1(a) Geolife 資料集地點類型統計分佈圖

圖 7.1(b) OSM 資料集地點類型統計分佈圖

Geolife 資料集中的 22706 筆活動序列，透過騰訊位置服務 API 自動標註序列中停留點的 18 種地點類型，其統計分佈圖如圖 7.1(a)所示，前三名最多的地點類型是(1) <房產小區,46.44%>，(2) <教育學校,12.55%>，以及(3) <旅遊景點,9.62%>。OSM 資料集中的 62377 筆活動序列，其停留點的地點類型統計分佈圖如圖 7.1(b)所示，前三名最多的地點類型是(1) <美食,20.62%>，(2) <購物,16.11%>，以及(3) <旅遊景點,9.48%>。

7.2 評估指標

本論文評估指標採用準確率@n，如公式 6 所示：

$$Accuracy@n = \frac{1}{|D|}|\{Y_i \mid Y_i \in Top_n(X_i) \wedge X_i \in D\}| \quad \text{(公式 6)}$$

D 表示測試資料集，$Top_n(X_i)$表示前 n 個活動意圖輸出結果所成的集合，X_i 表示 D 中某筆測試資料的長度 k 活動序列，Y_i 表示X_i 下個時間發生的真實活動意圖。

7.3 全體資料模型(GRU Global Model)之效果評估

7.3.1 評估特徵及其組合之預測效果

本實驗的目的是觀察全體資料模型輸入特徵數量是否影響預測效果，並透過統計方法 Max 做為比較基準。

本實驗使用所有使用者的長度 k 活動序列，使用全體資料模型進行預測，以論文[13]方法，輸入資料只包含地點類型一個特徵(實驗以 global_1f 表示)。本論文則擷取了更多特徵，包括地點類型、時間及停留時間(實驗以 global_3f 表示)。另外以統計的方式,將出現次數前 n 高的活動意圖類型作為預測結果(實驗以 Max 表示)，作為比較基準。

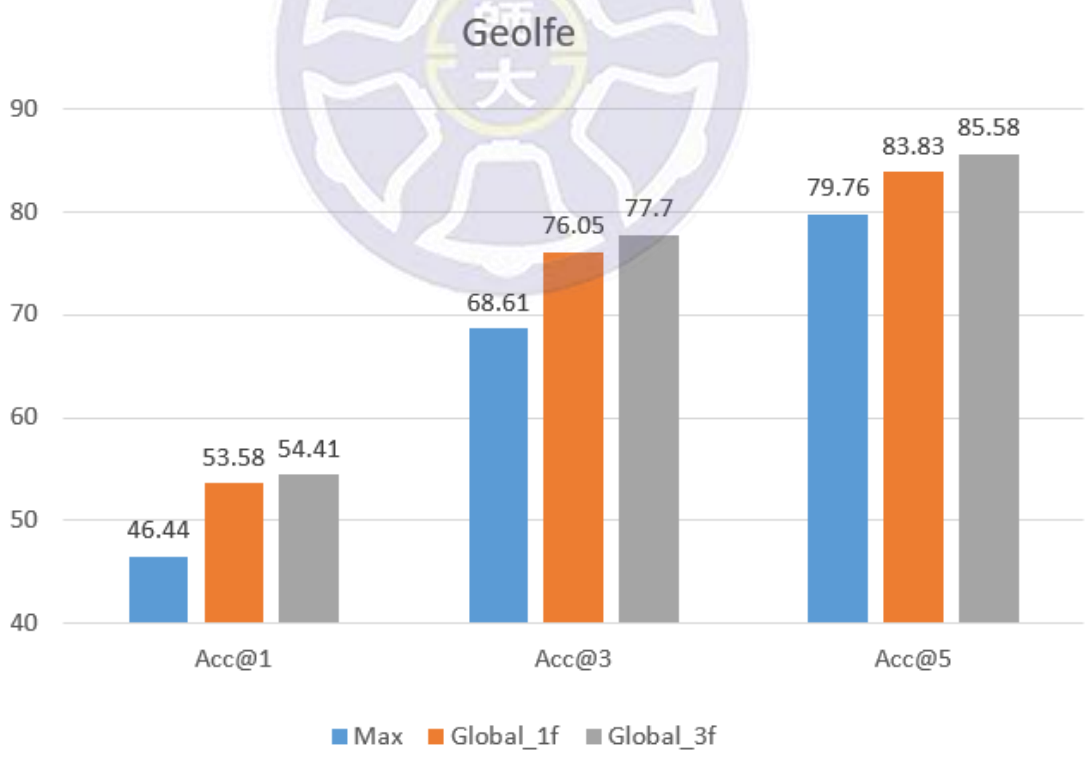

圖 7.2(a) Geolife 資料集全體資料模型預測效果

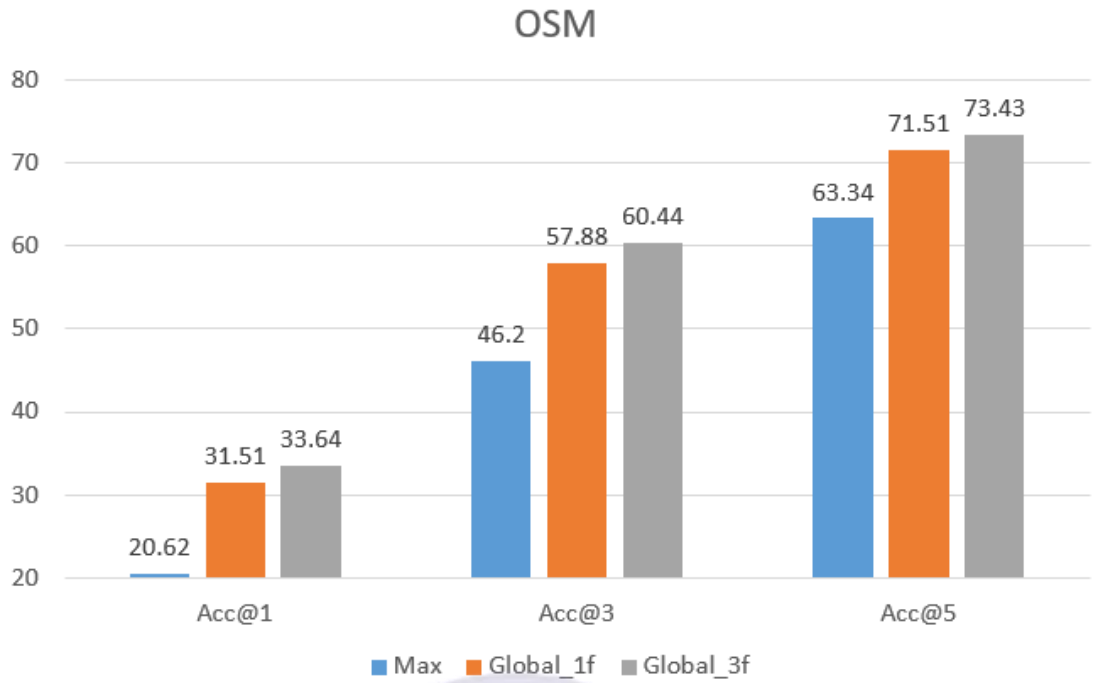

圖 7.2(b) OSM 資料及全體資料模型預測效果

從圖 7.2(a)所示，Geolife 資料集的活動意圖類型分佈較不均勻，因此 Max 方法在 Accuracy@1 時有 46.44%的準確率。論文[13]所提方法 Global_1f，在 Accuracy@1 有 53.58%的準確率，效果比 Max 好 7.14%，在 Accuracy@5 達到 83.83%的預測準確率。而本論文所提出的 Global_3f，輸入資料使用三個特徵，效果皆優於 Global_1f，在 Accuracy@5 有高達 85.58%的預測準確率，增加了 1.75%。從圖 7.2(b)所示，OSM 資料及的活動意圖類型分佈較為均勻，因此 Max 方法在 Accuracy@1 僅有 20.62%的準確率。論文[13]所提方法 Global_1f 在 Accuracy@1 有 31.51%的準確定，效果比 Max 好 10.89%，在 Accuracy@5 有 71.51%的準確率。本論文提出的 Global_3f，效果皆優於 Global_1f，在 Accuracy 有 73.43%的準確率，增加了 1.92%。

本論文所提方法不管在 OSM 資料集集 Geolife 資料集，與 Max 方法和論文

[13]所提方法 Global_1f 相比，全體資料模型採用三個特徵時效果皆較好，可顯示加入時間及停留時間特徵，可進一步提升預測效果。

7.3.2 模型參數設置實驗

本論文的模型架構為一層 GRU 層和兩回合隱藏層，最後以激活函數 Softmax 作為輸出結果。考慮每一層的輸出維度參數要設置多少可能會影響預測效果，因此本論文調整 GRU 層和第一回隱藏層的輸出維度觀察預測效果，第二回隱藏層輸出 18 維度為 18 種活動意圖，因此不變動維度。

從表 7.2(a)和表 7.2(b)所示，兩個資料集在 GRU 層的輸出維度改變和模型準確率無太大關係，輸出 32 維度和 64 維度的預測效果差異皆小於 0.2%，屬於誤差範圍。隱藏層維度設定愈高時，準確率有些微上升的趨勢。由於 GRU 層的輸出維度不太影響結果，因此後續實驗 GRU 層輸出設定為 32 維度，減少模型訓練時間，而隱藏層因為輸出維度愈高，預測效果有小幅增加的趨勢，因此將隱藏層輸出設定為 256 維度。

Geolife	GRU層輸出32維度	GRU層輸出 64維度
隱藏層輸出64維度	0.854015	0.853894
隱藏層輸出128維度	0.854237	0.854276
隱藏層輸出256維度	0.855849	0.855793

表 7.2(a) Geolife 資料集參數設定實驗(Accuracy@5)

OSM	GRU層輸出 32維度	GRU層輸出 64維度
隱藏層輸出64維度	0.732238	0.731167
隱藏層輸出128維度	0.732952	0.731524
隱藏層輸出256維度	0.734302	0.73438

表 7.2(b) OSM 資料集參數設定實驗(Accuracy@5)

7.4 群組模型(GRU Group Model)之效果評估與比較

此部分的實驗以群組資料訓練模型，和全體資料模型的預測效果進行比較，並比較各群組的準確率。以下將分別評估以使用者分群法及序列分群法所建立各群組模型的預測效果，以及模型選擇方法效果分析。

7.4.1 使用者分群法(User-based Clustering)預測較果評估

(一) 群組模型整體預測效果

以使用者為單位進行分群，本論文將群組資料分別使用群組模型進行預測，同時也以全體資料模型進行預測。本實驗將群組資料分成兩個方向討論：(1)將測試資料視為新的使用者，透過 5.3 小節的方法，決定以哪個群組模型進行預測(實驗以 Group_3f_new 表示)。(2)測試資料的使用者，以原使用者歷史資料所在群組，決定群組模型進行預測(實驗以 Group_3f_old 表示)。

如圖 7.3(a)及 7.3(b)所示，兩份資料集皆顯示，不論將使用者視為新使用者選擇群組模型，或是以使用者歷史資料所在群組模型直接預測，在 Accuracy@1 皆表現的比全體資料模型還要好，在 Accuracy@3 和 Accuracy@5 則略微降低。但 Group_3f_new 和 Group_3f_old 的預測效果差不多，各有較高準確率的部分，然而 Group_3f_old 會有冷啟動(cold-start)的問題，若有新的使用者卻沒有該使用者的歷史記錄時會無法預測。因此，本論文的實驗將以 Group_3f_new 的方式來進行評估，對新使用者可選定預測群組，且對擁有歷史資料使用者，也可考慮其習慣可能改變，而根據新的活動序到找到一個合適的群組模型。

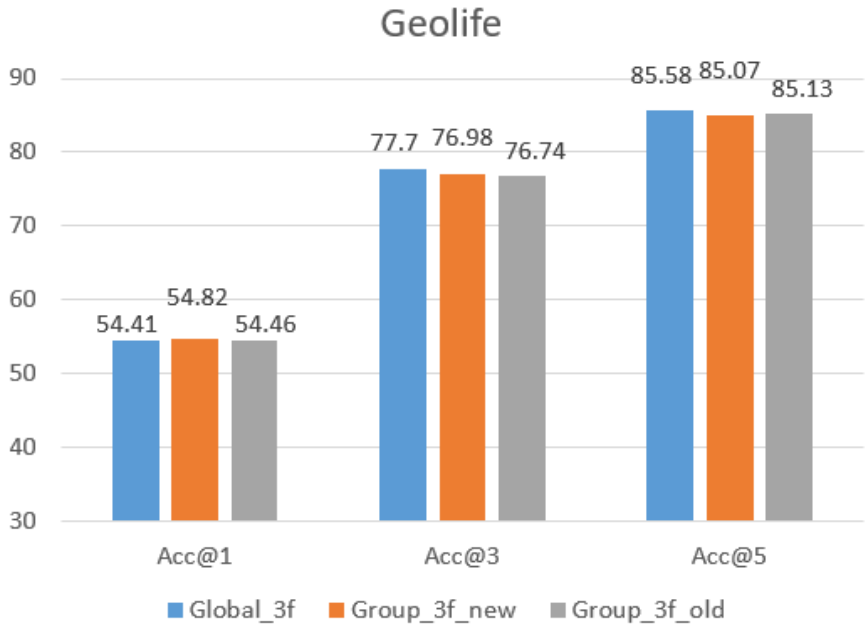

圖 7.3(a) Geolife 資料集群組模型預測效果

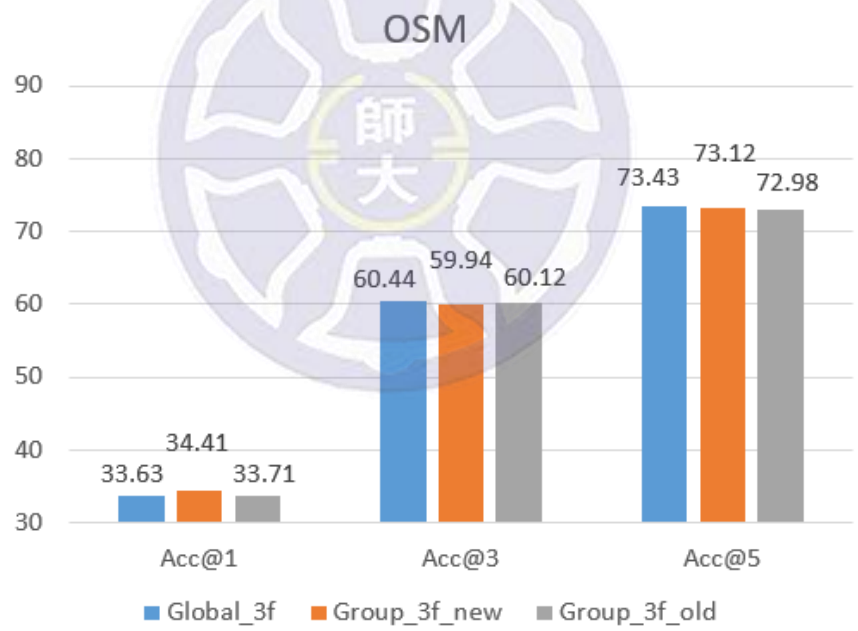

圖 7.3(b) OSM 資料集群組模型預測效果

(二) 各群組模型預測效果之比較

如圖 7.4(a)和圖 7.4(b)所示，將群組資料分別套用全體資料模型和所屬群組模型進行預測，以全體資料模型為比較基準，觀察使用群組模型的平均預測效果。Geolife 資料集在群組 1、2、3 的 Accuracy@1 有小幅提升，OSM 資料集在群組 1、3 的 Accuracy@1 也有提升。各群組的資料分佈如表 7.3(a)和表 7.4(b)所示。本論文認為，預測效果和訓練模型採用的資料數量有關係，可觀察到，Geolife 資料集在群組 4 的資料量是最少的，只有其他群組約 1/4 的數量，OSM 資料集則是在群組 2 的資料量最少，只有其他群組的 1/10 左右。反應在實驗結果裏，在資料量偏少的群組模型中，預測效果 Accuracy@1,3,5 皆筆採用全體資料模型有下降的趨勢。若在資料量充足的群組模型中，則預測效果在 Accuracy@1 會有所提升，Accuracy@3,5 則較不穩定。

本論文在使用者群組模型得到一個結論，若使用本論文分群方法，在群組資料量充足時，預測效果能夠在 Accuracy@1 有所提升。

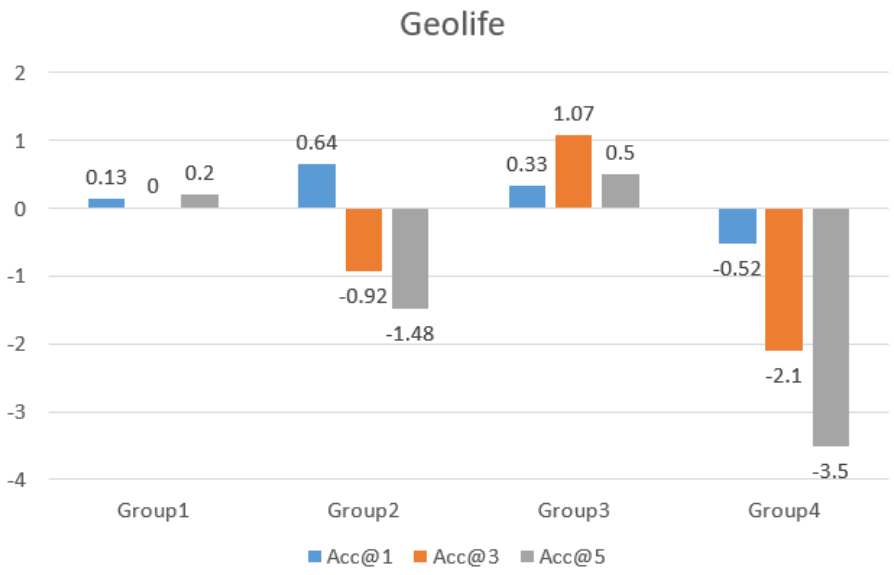

圖 7.4(a) Geolife 各群組模型以全體資料模型為基底比較效果

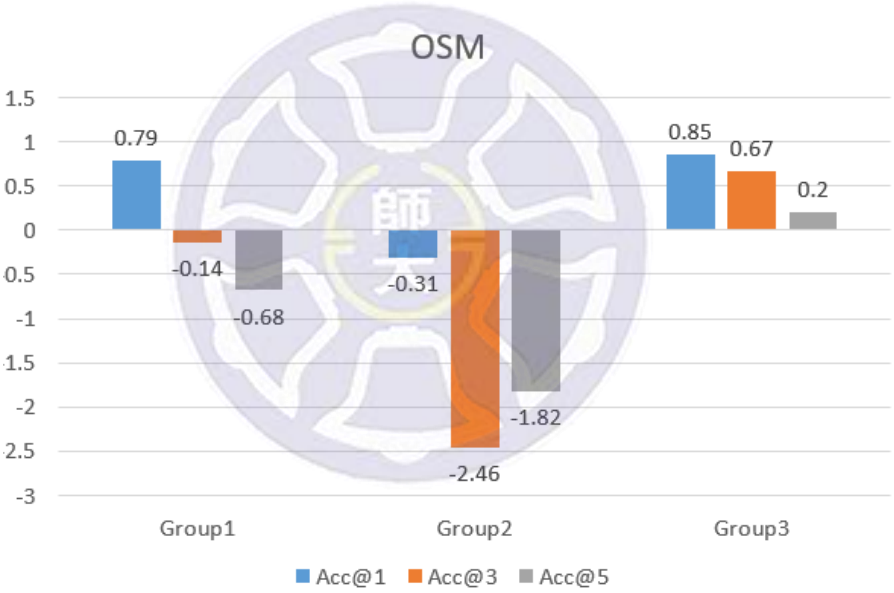

圖 7.4(b) OSM 各群組模型以全體資料模型為基底比較效果

	Training data	Testing data
Group1	6342	1588
Group2	5648	1414
Group3	4833	1211
Group4	1334	336

表 7.3(a) Grolife 資料集各群組資料分佈

	Training data	Testing data
Group1	23782	5947
Group2	2209	554
Group3	23906	5979

表 7.3(b) OSM 資料集各群組資料分佈

7.4.2 序列分群法(Sequence-based Clustering)預測效果評估

(一) 群組模型整體預測效果

此部分實驗以長度 k 活動序列為單位進行分群，將群組資料分別使用群組模型進行預測及全體資料模型進行預測。由於當序列過短時，相似度普遍偏低，會造成分群處理上的困難，因此本實驗將 k 設定為 9，使用長度 9 活動序列進行分群。以下將針對採用不同相似度計算進行分群，分別觀察其群組模型的預測效果。

(1) 分群處理步驟使用 LCS 相似度計算

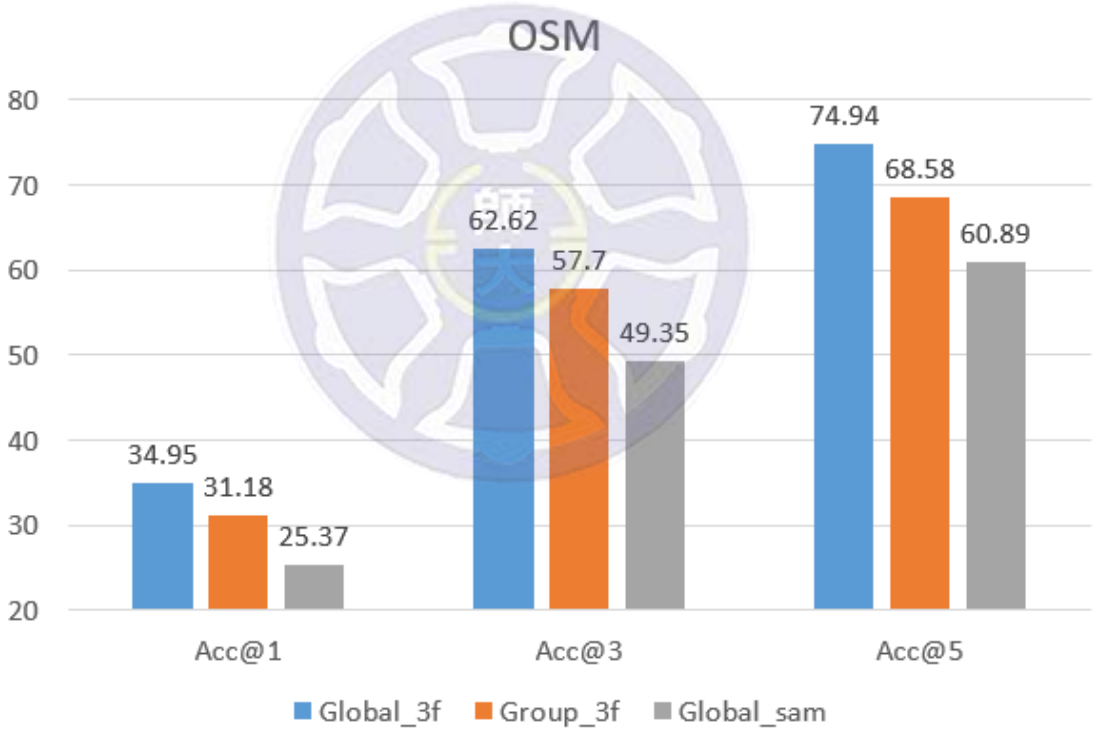

圖 7.5 OSM 序列群組模型預測效果評估(LCS 相似度)

如圖 7.5 所示，全體資料模型的預測結果以 Global_3f 表示，序列群組模型的預測結果以 Group_3f 表示。經觀察發現，序列群組模型預測效果比全體資料模型差，在 Accuracy@1 降低了 3.77%的準確率，在 Accuracy@5 更降低了 6.36%的準

確率。

本論文認為，以序列分群法找出的各群組資料量偏少，如表 7.4 所示，很可能是造成模型預測效果不好的原因，因此本論文從全體資料中，只抽樣 200 筆做為序列訓練模型，共抽樣 10 次，此模型的平均預測結果以 Global_sam 表示，並和 Group_3f 進行比較。

比較群組模型和抽樣全體資料模型，可顯示以接近 200 筆的訓練資料，群組模型的準確率明顯筆整體抽樣模型的準確率高，在 Accuracy@1 提升了 5.81%的準確率，在 Accuracy@5 提升 7.69%準確率，在接近的序列資料數量下，群組模型的確比混雜的全體資料取樣，對分群資料有更準確的預測結果。

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Group 11
Train	230	154	151	256	198	288	197	80	117	145	123
Test	173	108	115	193	134	189	142	75	92	109	153
	Group 12	Group 13	Group 14	Group 15	Group 16	Group 17	Group 18	Group 19	Group 20	Group 21	All
Train	168	212	128	208	152	137	112	97	99	127	3379
Test	174	176	101	164	158	92	87	121	98	147	2801

表 7.4 OSM 序列群組資料分佈(LCS 相似度)

(2) 分群處理步驟使用雅卡爾相似度計算

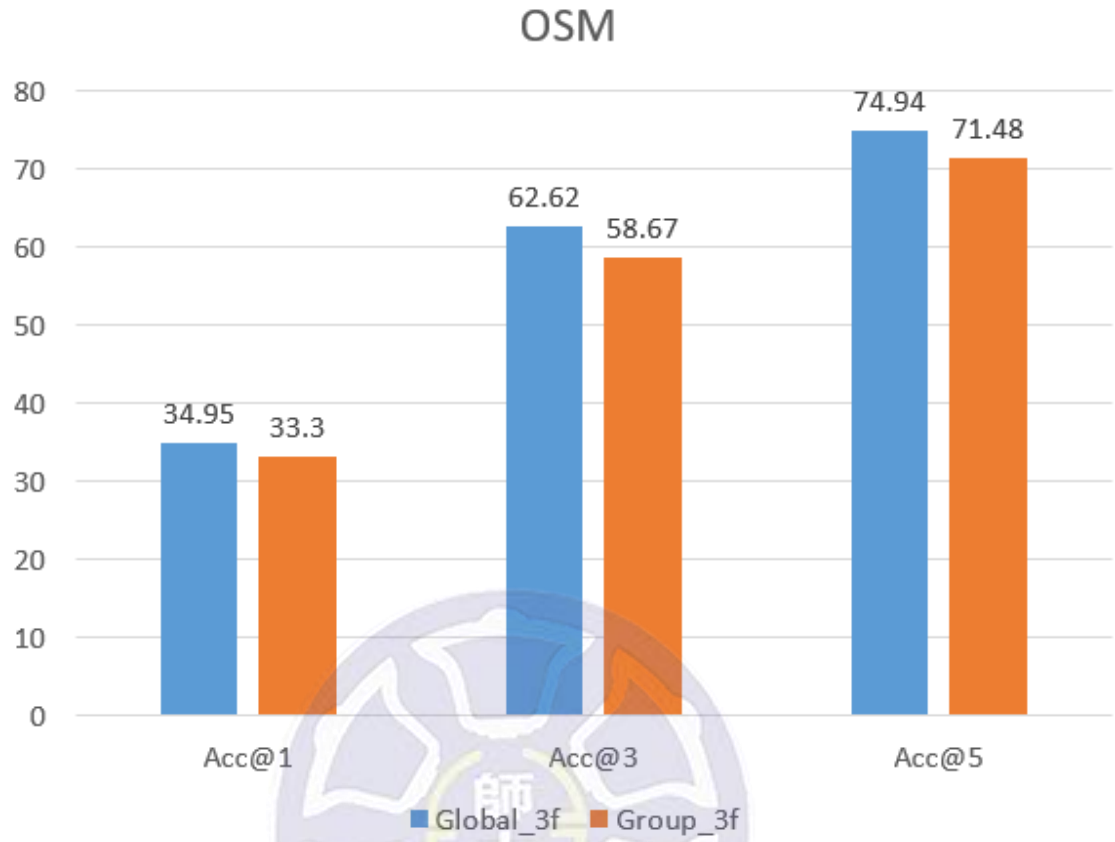

圖 7.6 OSM 序列群組模型預測效果評估(雅卡爾相似度)

	Training	Testing
Group1	1798	450
Group2	1814	454
Group3	1777	445
Group4	1738	459
Group5	1832	435
Group6	1819	454
Group7	416	104
Global	11202	2801

表 7.5 OSM 序列群組資料分佈(雅卡爾相似度)

如圖 7.6 所示，全體資料模型預測結果以 Global_3f 表示，序列群組模型預測結果以 Group_3f 表示。此結果顯示，序列群組模型預測效果比全體資料模型差，在 Accuracy@1 降低了 1.65%的準確率，降低幅度較使用 LCS 相似度少。本論文認為群組中資料的相似性和預測效果有關，使用者分群法以使用者的移動模式計算雅卡爾相似度，序列分群法以地點類別計算雅卡爾相似度時，沒有考慮地點出現前後順序，因此雖然每個群組包含的資料量比使用 LCS 相似度來的多，但因為分群中的資料相似性較低，即使資料充足，也無法訓練出較好的群組模型，因此接下來的序列群組模型實驗，在分群步驟皆用長度 k 活動序列以 LCS 計算相似度。

(二)、 各群組模型預測效果之比較

如圖 7.7(a)所示，將群組資料分別套用全體資料模型和所屬群組模型進行預測，以全體資料模型為比較基準，觀察使用群組模型的效果。實驗顯示，不論在哪個群組模型，預測效果都是下降的。但是和抽樣全體資料 200 筆所建立的模型比較時，不論哪個群組模型，預測效果都是提升的，如圖 7.6(b)所示。因此本論文得到一個結論，序列分群方法確實可以建立更準確的群組模型，但是訓練模型所使用的群組資料量必須要足夠。

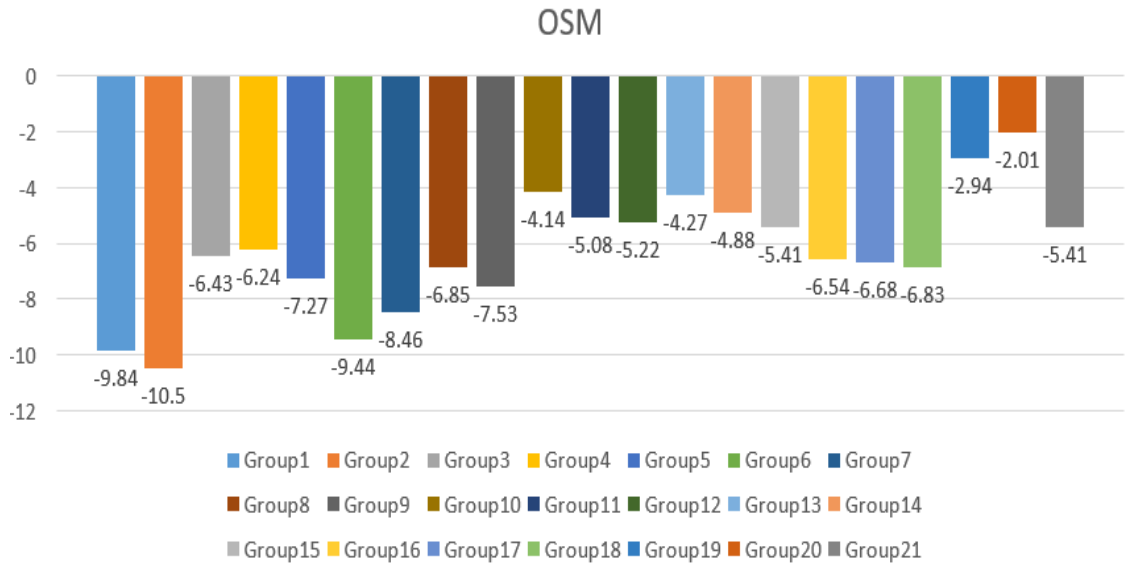

圖 7.7(a) OSM 各群組模型以全體資料模型為基底比較效果

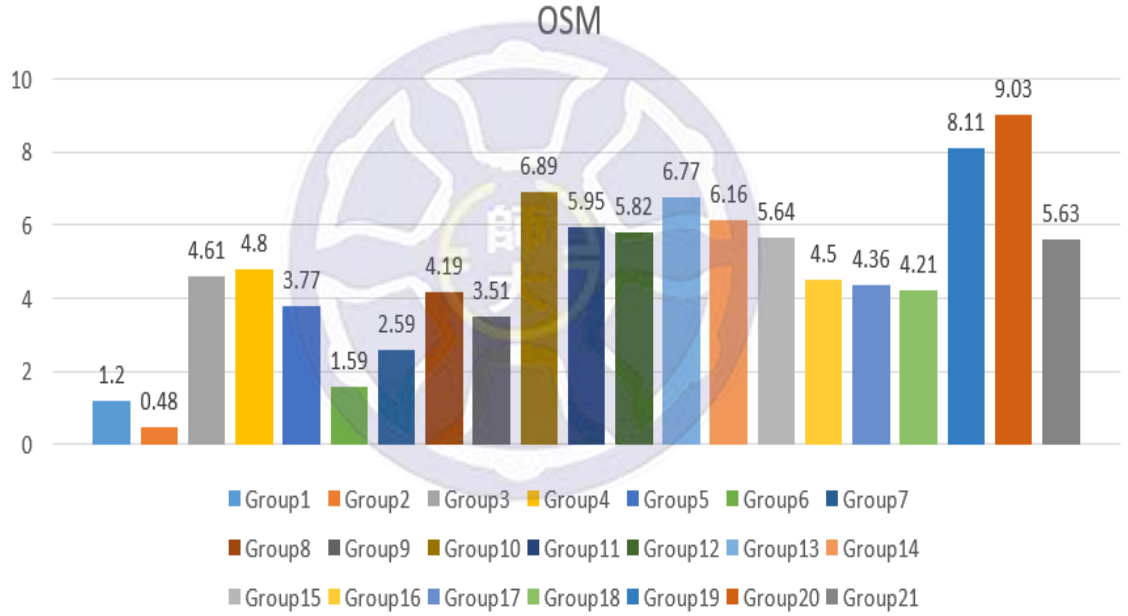

圖 7.7(b) OSM 各群組模型以抽樣全體資料模型為基底比較效果

7.4.3 群組模型選擇及預測效果評估

本實驗評估透過 5.3 對測試資料選擇適用模型的效果，因此，將各群組測試資料套用到其他未被挑選維適用群組的群組模型，相對比較其預測效果。表 7.6(a)和表 7.6(b)顯示，各群組測試資料，皆以所挑選的群組模型能達到最佳預測效果，其中 Geolife 資料集將群組 4 的測試資料套用其他的群組模型，下降幅度較少，而將其他群組的測試資料套用群組 4 模型時，下降幅度很高。這個實驗除了驗證本論文提的群組模型選擇方法適用之外，也能說明群組模型訓練好壞與否，和訓練的資料量有很大關係。OSM 資料集顯示的結果也和 Geolife 資料集相符，資料量少的群組資料套用其他群組模型時，雖然準確率有下降，但下降幅度不高，而其他其組資料套用訓練資料量少的群組模型時，下降幅度則非常高。

本小節呈現了本論文所提兩種分群方法，當分群中的資料量足夠，使用者分群法在特定條件之下有助於提升準確率，而序列分群法則因為分群中的資料量不足或是分群中序列的相似度計算過為寬鬆，比較無助於提升準確率。因此接下來的組合模型實驗，評估將全體資料模型和使用者群組模型做結合的預測結果。

Geolife (%)	Group1 Model	Group2 Model	Group3 Model	Group4 Model
Group1 Data		-3.36	-4.65	-6.12
Group2 Data	-2.63		-3.78	-5.66
Group3 Data	-4.51	-4.25		-5.96
Group4 Data	-1.53	-0.88	-2.15	

表 7.6(a) Geolife 群組模型選擇方法效果評估

OSM(%)	Group1 Model	Group2 Model	Group3 Model
Group1 Data		-7.63	-4.27
Group2 Data	-0.69		-1.24
Group3 Data	-5.19	-10.27	

表 7.6(b) OSM 群組模型選擇方法效果評估

7.5 組合模型之預測效果評估

本實驗將進行遷移學習模型和合成模型的預測效果評估，以全體資料模型和群組模型為比較基準，觀察準確率是否有提升。

7.5.1 遷移學習模型(Transfer Learning Model)之預測效果評估

圖 7.8(a)顯示，採用遷移學習模型在 Geolife 資料的預測效果，介於全體資料模型和群組資料模型之間。本論文認為，其原因是因為遷移學習的概念是有一個龐大的基底，透過重訓練部分模型的方式慢慢調整到適合預測群組資料的模型。當採用全體資料模型預測結果較好時，模型往準確率下降的群組模型學習，導致遷移學習模型相較於全體資料模型，預測效果會下降，相反的，若全體資料模型預測結果較差時，模型往準確率提高的群組模型學習，使遷移學習模型較全體資料模型的預測效果提升。如圖 7.8(b)所示，遷移學習模型在 OSM 資料的預測效果，則是在群組 2 和 3 有提升，分別提升了 0.93%和 0.43%，總平均也有提升。本論文認為以上結果和資料屬性有關係，Geolife 資料集屬性較為極端，在 18 種活動意圖類型中，其中一個活動意圖佔了將近一半，OSM 資料集則是活動意圖類型分佈較均勻，比較偏向常態資料，因此遷移學習模型在 OSM 資料集有助於提升準確率。

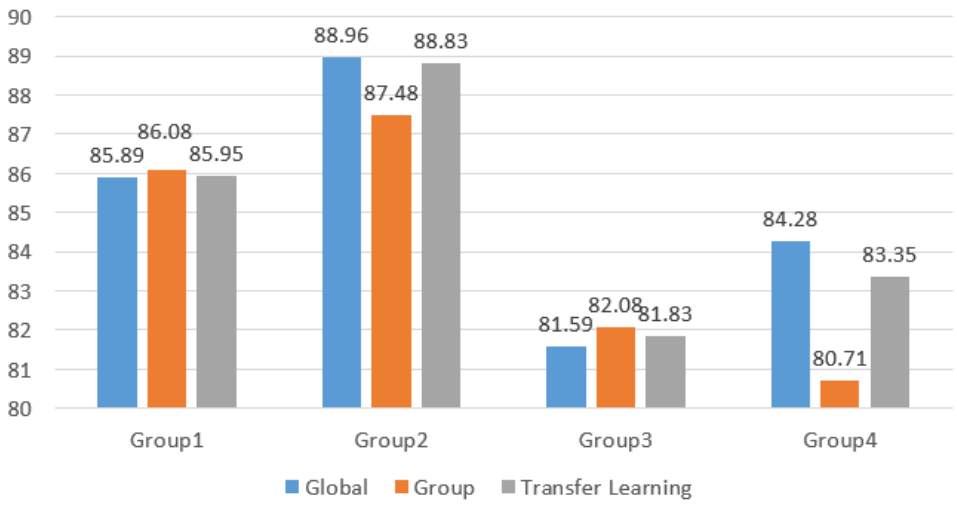

圖 7.8(a) Geolife 遷移學習模型預測效果與比較(Accuracy@5)

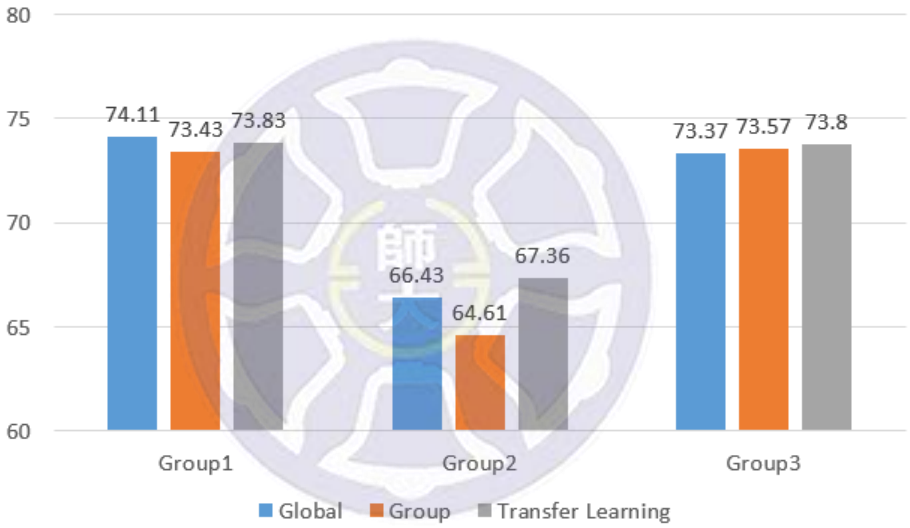

圖 7.8(b) OSM 遷移學習模型預測效果與比較(Accuracy@5)

7.5.2 合成模型(Ensemble Model)之效果評估與比較

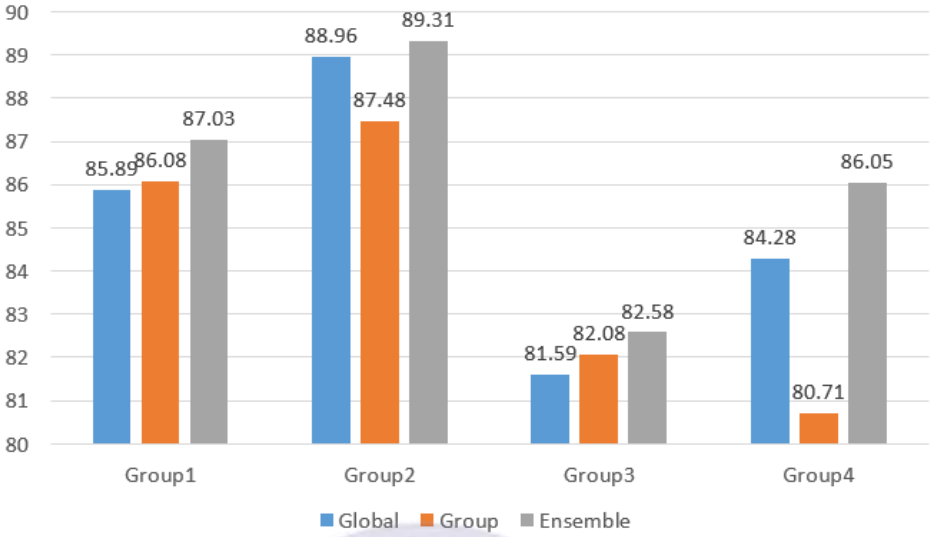

圖 7.9(a) Geolife 合成模型預測效果與比較(Accuracy@5)

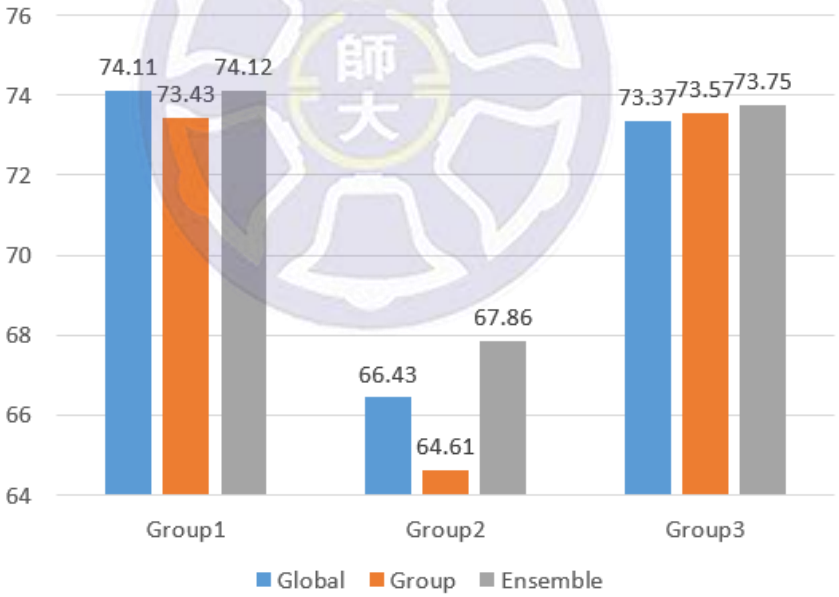

圖 7.9(b) OSM 合成模型預測效果與比較(Accuracy@5)

圖 7.9(a)和圖 7.9(b)分別顯示，兩份資料集以全體資料模型、群組模型，及合成模型的預測所果，在兩個資料及的各群組，皆顯示合成模型能有效地提升準確

率。在 Geolife 資料集中，合成模型在 Accracy@5 最高可達 89.31%的準確率，在 OSM 資料集中，也可達 74.12%的準確率。

7.5.3 序列長度影響評估

本實驗探討改變長度 k 序列觀察對模型預測效果的影響。本論文共提出四種模型架構：全體資料模型、群組模型、遷移學習模型及合成模型，以下將比較 k 的長度和各模型的預測效果，本實驗中，k 有 3,5,7,9 四種設定值。

對 Geolife 資料集，不同序列長度和各模型的預測效果如圖 7.10(a)所示。這份資料集的全體資料模型、群組模型和合成模型，以長度 5 和 7 的活動序列套用模型時有較好的準確率；當序列長度為 9 時，準確率反而相對較低。此外，在本實驗的交叉比對當中，合成模型仍有最好的預測效果。

對 OSM 資料集，不同序列長度和各模型的預測效果如圖 7.10(b)所示。這份資料集可顯示，當序列長度愈長時，預測效果愈好。從實驗的交叉比對當中，以長度 7 和 9 的活動序列套用遷移學習模型有較高的準確率，合成模型次之但很接近，長度 3 和 5 的活動序列則適用合成模型較好。

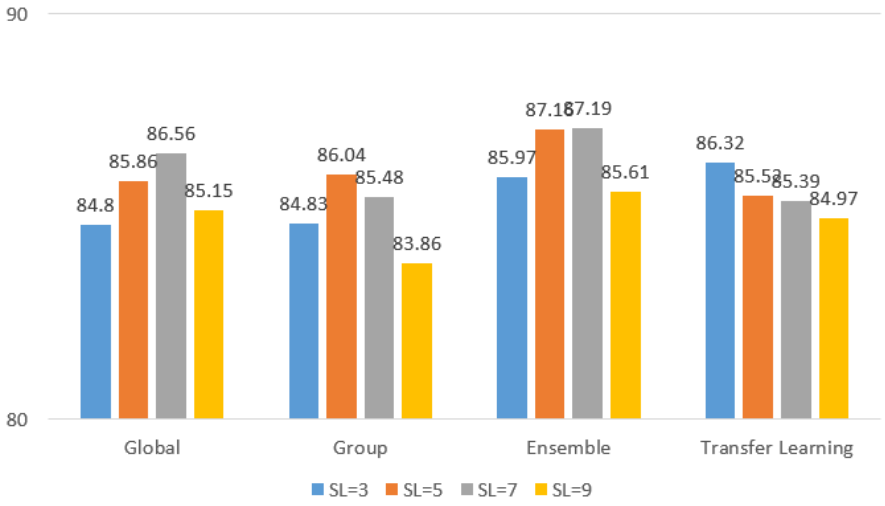

圖 7.10(a) Geolife 序列長度預測結果評估(Accuracy@5)

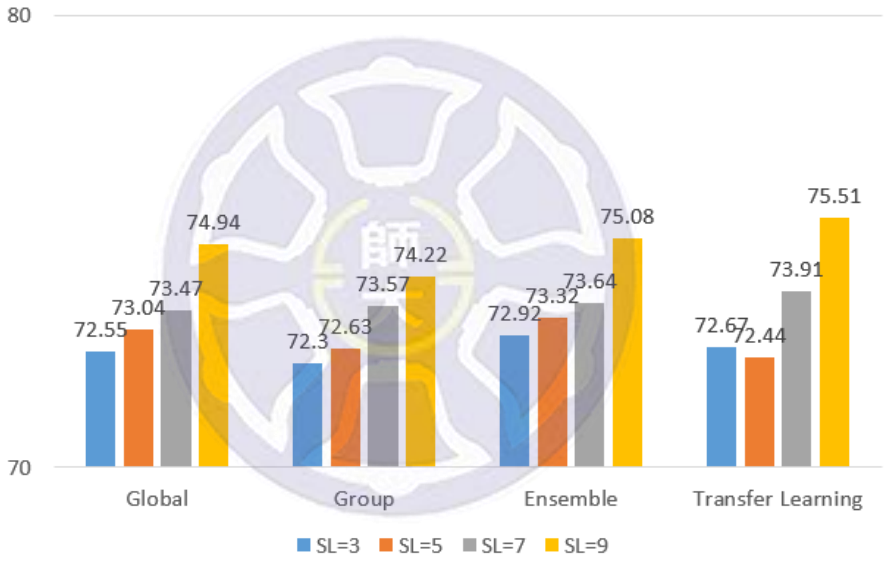

圖 7.10(b) OSM 序列長度預測結果評估(Accuracy@5)

7.5.4 加入時間條件影響評估

本實驗探討加入時間條件來預測特定時間點的活動意圖，觀察對模型預測效果的影響。架構如圖 7.11 所示，時間條件為一天當中的 8 個時段，以 one-hot 表示。之後將輸入維度 8 的時間條件和第一回隱藏層輸出結果串聯(concatenation)，並用第二回隱藏層輸出 18 個維度結果作為預測地點類別。本論文共提出四種模

型架構：全體資料模型、群組模型、遷移學習模型及合成模型，以下將比較模型加入時間條件後與先前所建立各模型的預測效果。

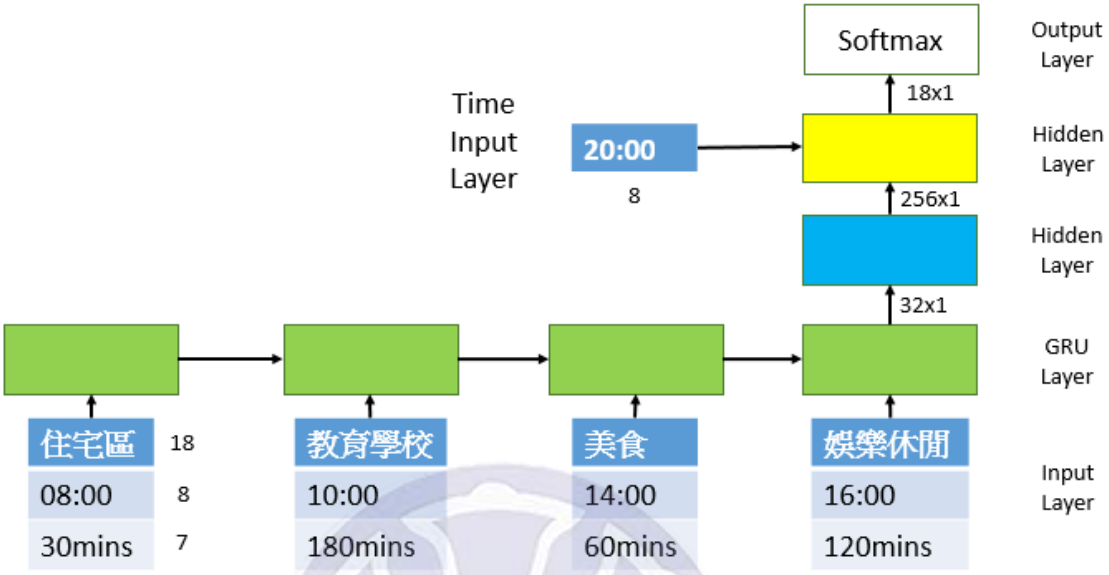

圖 7.11 加入時間條件之模型架構

如圖 7.12(a)和 7.12(b)所示，可觀察出，加入時間條件預測使用者的活動意圖，在全體資料模型的表現有提升。但加入時間條件後，資料分佈會變的較稀疏，加上群組中的資料不足，使得群組模型的效果皆略為降低，連帶影響到合成模型以及遷移學習模型。因此，本研究結果顯示，在群組資料不足的情況下，加入時間條件未能得到較好的預測結果。

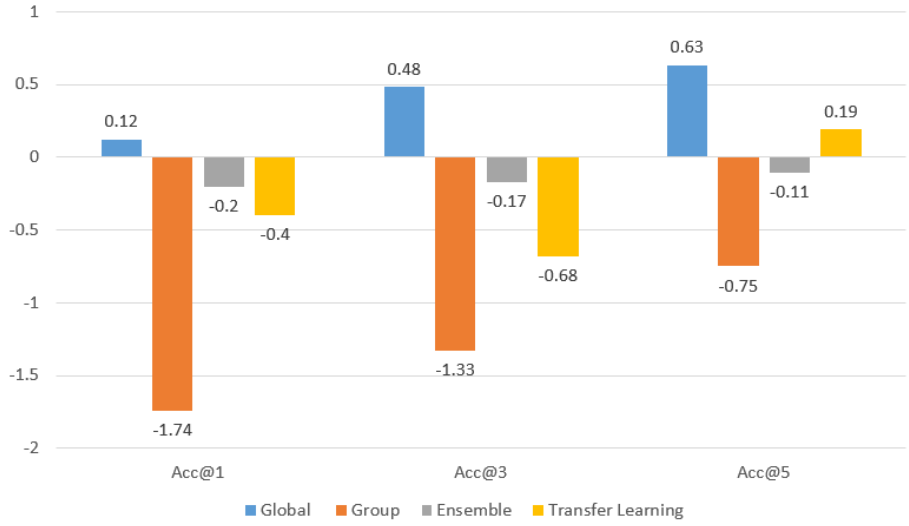

圖 7.12(a) Geolife 加入時間條件以未加入時間條件為基底比較結果

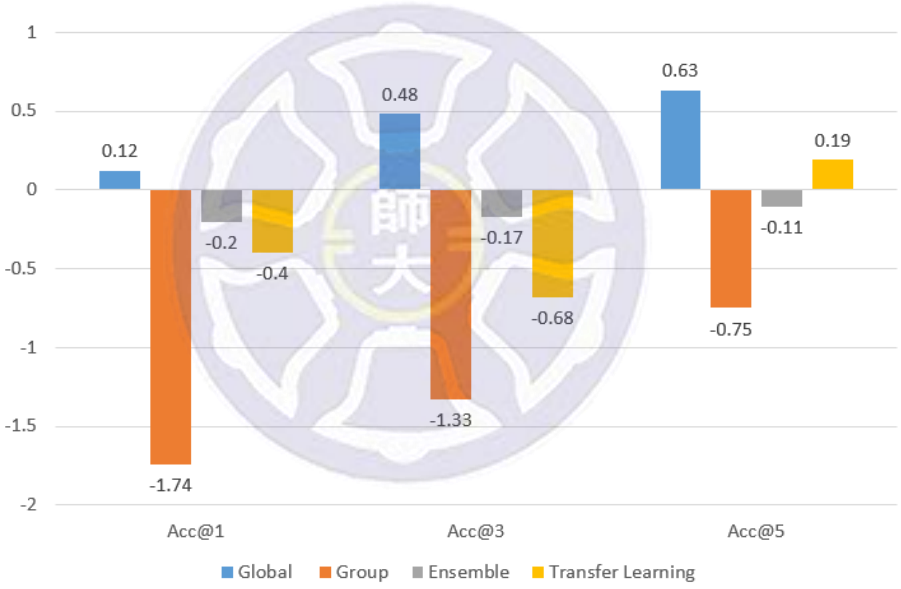

圖 7.12(b) OSM 加入時間條件以未加入時間條件為基底比較結果

第八章　結論與未來研究方向

本論文研究活動意圖類型預測方法，提出以相關研究[13]提出的遞迴類神經網路架構為基礎，以 GPS 軌跡資料擷取更多特徵作為模型輸入，並改良模型架構。本論文提出兩種分群方法，以使用者分群法建立的群組模型，實驗結果顯示在群組資料充足時，能提升預測結果的 Accuracy@1。以序列分群法建立的群組模型則因為資料不充足，導致效果不明顯。不過經由實驗觀察，透過抽樣與平均群組資料量差不多的全體資料進行模型訓練，可驗證出以序列分群法增進群組模型預測效果是有幫助的。此外，本論文提出兩個方式組合全體資料模型和群組模型。實驗評估顯示，遷移學習模型在 OSM 資料集的預測結果優於全體資料模型和群組模型，合成模型則是透過調和係數學習結合兩個模型輸出結果的比重，能有效幫助系統更正確預測出使用者活動意圖。本論文所提的組合模型中，合成模型比遷移學習模型更能達到本論文所期望之目標，綜合全體資料模型和群組模型結果，提升對各群組的預測準確率。

本研究未來可進一步根據活動意圖類型，搜索使用者附近符合的 POI，從這些 POI 中推薦地點或相關資料給使用者。

參考文獻

[1] J. R. Benetka, K. Balog and K.Norvag ,“Anticipating Information Needs Based on Check-in Activity,” in Proceedings of the 10th ACM International Conference on Web Search and Data Mining(WSDM), 2017.

[2] J. Bao, Y. Zheng and M. F. Mokbel ,“Location-based and Preference-Aware Recommendation Using Sparse Geo-Social Network Data,” in Proceedings of the 20th International Conference on Advances in Geographic Information Systems(SIGSPATIAL), 2012.

[3] X. Chen, D. Shi, B. Zhao and F. Liu ,“Periodic Pattern Mining Based on GPS Trajectories,” in Proceedings International Symposium on Advances in Electrical, Electronics and Computer Engineering (ISAEECE), 2016.

[4] S. Ghosh and S. K. Ghosh ,“Modeling of Human Movement Behavioral Knowledge from GPS Traces for Categorizing Mobile Users , “in Proceedings of the 26th ACM International Conference on World Wide Web Companion(WWW), 2017.

[5] S. Ghosh and S. K. Ghosh ,“THUMP: Semantic Analysis on Trajectory Traces to Explore Human Movement Patterns,” in Proceedings of the 25th International Conference Companion on World Wide Web(WWW), 2016.

[6] Y. Kim, J. Han and C. Yuan ,“TOPTRAC: Topical Trajectory Pattern Mining,” in

Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD), 2015.

[7] M. Lv, L. Chen, and G. Chen ,“Discovering Personally Semantic Places from GPS Trajectories,” In Proceedings of the 21st ACM international conference on Information and knowledge management(CIKM), 2012.

[8] A.Likhyani, D. Padmanabhan, S. Bedathur, and S. Mehta ,“Inferring and Exploiting Categories for Next Location Prediction,” in Proceedings of the 24th International Conference on World Wide Web(WWW), 2015.

[9] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma ,“Neural Attentive Session-based Recommendation,” in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management(CIKM), 2017.

[10] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu and W. Y. Ma ,“Mining User Similarity Based on Location History, ” in Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems(GIS), 2008.

[11] J. McInerney, A. Rogers, and N. R. Jennings ,“Improving Location Prediction Services for New Users with Probabilistic Latent Semantic Analysis,” in Proceedings of the 2012 ACM Conference on Ubiquitous Computing(UbiComp), 2012.

[12] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo ,“Mining User Mobility Features for Next Place Prediction in Location-based Services,” in Proceedings of the 2012 IEEE 12th International Conference on Data Mining(ICDM), 2012.

[13] E. Palumbo, G. Rizzo, R. Troncy, and E. Baralis ,“Predicting Your Next Stop-over from Location-based Social Network Data with Recurrent Neural Networks,” in RecTour, 2017.

[14] F. Wu and Z. Li ,“Where Did You Go: Personalized Annotation of Mobility Records,” in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management(CIKM), 2016.

[15] X. Xiao, Y. Zheng, Q. Luo, and X. Xie ,“Finding Similar Users using Category-Based Location History,” in Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS), 2010.

[16] J. Ye, Z. Zhu, and H. Cheng ,“What’s Your Next Move: User Activity Prediction in Location-based Social Networks,” in Proceedings of the 2013 SIAM International Conference on Data Mining(SIAM), 2013.

[17] M. Zhou, Z. Ding, J. Tang, and D. Yin ,“Micro Behaviors: A New Perspective in E-commerce Recommender Systems,” in Proceedings of the 16th ACM International Conference on Web Search and Data Mining(WSDM), 2018

[18] Y. Zheng, X. Xie and W. Y. Ma ,“GeoLife: A Collaborative Social Networking

Service among User, Location and Trajectory," in IEEE 2008.

[19] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang ,"Collaborative Location and Activity Recommendations with GPS History Data," in Proceedings of the 18th international conference on World wide web(WWW) 2010.

[20] Y. Zheng, L. Zhang and X. Xie, W. Y. Ma ,"Mining Correlation between Location Using Human Location History," in Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems(GIS), 2009