國立政治大學資訊科學系
Department of Computer Science
National Chengchi University

碩士論文
Master's Thesis

具概念飄移的動態社群網絡之類別預測
Label Prediction on Dynamic Social Networks with Concept Drifting

研 究 生：游詳閔
指導教授：沈錳坤

中華民國九十九年十月
October 2010

DOI:10.6814/NCCU201901233

具概念飄移的動態社群網絡之類別預測

Label Prediction on Dynamic Social Networks with Concept Drifting

研究生：游詳閔　　Student: HsiangMin Yu

指導教授：沈錳坤　　Advisor: Man-Kwan Shan

國立政治大學

資訊科學系

碩士論文

A Thesis

submitted to Department of Computer Science

National Chengchi University

in partial fulfillment of the Requirements

for the degree of

Master

in

Computer Science

中華民國九十九年十月

October 2010

DOI:10.6814/NCCU201901233

具概念飄移的動態社群網絡之類別預測

摘要

社會網絡在電腦科學的研究範疇中扮演一個日漸重要的角色，類別預測正是其中一項熱門的議題。類別預測的研究目標，是利用網絡中部分已知類別的節點，預測出其他未知類別節點之類別。

以往類別預測之研究，皆以靜態社會網絡為主；然而，社會網絡往往是隨著時間動態演進的。在動態網絡中，網絡中的節點、連結、類別，皆可能隨著時間演進而更動。連帶的，節點之間相互影響的關係也會隨著時間改變。此變動可以視為一種概念飄移 (Concept Drift)。

不同於過往的研究，我們指出了動態網絡中類別分類的問題，並利用靜態網絡中類別分類的技術，結合概念飄移的方法，提出能夠在動態網絡中預測類別的解法。

實驗所採用的資料是 IMDb (Internet Movie Database) 的社會網路，我們用以預測演員的類別，根據實驗結果顯示，將動態社會網絡的演化過程，加入作為類別預測的參考指標，能夠提高動態網絡中類別分類的準確性。

DOI:10.6814/NCCU201901233

Label Prediction on Dynamic Social Networks with Concept Drifting

Abstract

Label prediction is one of the central questions of social network research. The core of label prediction is the use of labeled nodes to predict labels of un-labeled nodes in a social network. The definition of a labeled social network is a social network of partial or complete labeled nodes. The nodes in the same social network have a mutual impact on each other's labels.

Previous research on label prediction have been focused on static social networks. However, social networks are more dynamic in reality. In a dynamic social network, the links of nodes, even the labels of nodes, can be changed with time. The mutual influence of nodes can also be changed. The changing is called "Concept Drift."

This thesis predicts the labels on a dynamic labeled social work. We address the problems of classification for a dynamic social network. The technique of label prediction on static social networks and algorithms used to tackle concept drift are combined to solve the label prediction problem on dynamic social networks.

Experiments were performed on a labeled social network constructed from the Internet Movie Database. The results show that we can use the evolution of dynamic social networks to generate a more precise prediction of labels.

DOI:10.6814/NCCU201901233

致謝

這是一篇用了九年書寫，關於兩年的致謝。這幾年發生太多太多事，很開心終能從 DMLab 畢業，這段日子裡最感謝的人是我的指導教授：沈錳坤老師。第一次走進老師研究室的那個夏天傍晚，記憶仍然鮮明。謝謝老師在研究所兩年以及往後的人生中，給了我許多的鼓勵、教誨與幫助，謝謝老師陪我們熬了無數個漫長的夜、在就業及人生轉折時的支持，對我來說老師不只是學業上的教授，更是生命裡的導師，真的非常感謝您。

接著，要感謝 DMLab 的學長姐與學弟們：政吉、孟芝、棋安、伯丞、家奇、宏哲、紘維、柯柯柯，謝謝你們讓實驗室充滿了歡樂與活力。感謝華富學長，謝謝您不只當我的口試委員，更在最後鼓勵我，希望您在的世界平靜美好。謝謝阿德學長在寫論文時給我很多鼓勵，你一直是我們最好的榜樣。謝謝 UFO 在程式上與幽默感上幫助我精進；謝謝李法賢，你是我在實驗室最好的戰友，也是往後人生裡重要的朋友。謝謝容瑜，妳的認真總是我們學習的對象。謝謝戴張、斯越、柏堯、世宏、建成，跟你們一起說笑聊天絕對是研究所生活最快樂的時光之一。謝謝王裕炫、陳力瑋、阿三跟阿順，每次上樓前的閒聊總能讓人活力充滿。感謝韓助教以及譚助教，謝謝妳們在學校的幫忙。特別是韓助教，衷心謝謝妳在畢業期間給了我莫大的幫助。能在政大唸書、加入 DMLab，遇到你們每一個人，真是太好了。

此外，我要感謝我的媽媽，爸爸，爺爺，奶奶。謝謝你們把我帶來這個世界上。特別感謝我的媽媽，謝謝妳把我們養成現在的樣子，如果我們有什麼地方特別出色，那都是因為妳。最重要的，我要感謝我的妹妹聖賀，謝謝妳在我快樂的時候分享我的喜悅，在我痛苦的時候撐著我，在我懷疑自己的時候仍然信任，直視我告訴我可以。妳是我生命裡面的網，謝謝妳的陪伴。謝謝妳和媽媽，才有今天的游詳閔。

還要感謝這些年來陪著我的伴侶，謝謝在不同人生階段出現的妳，謝謝妳曾

DOI:10.6814/NCCU201901233

愛過我，燙平我的不安與彆扭。謝謝現在的女友，謝謝妳給了我很多的愛與包容，往後的人生也請多擔待了。感謝這些年的朋友們：大學同學、HOOP, MovieSpotNYC, SpoonRocket, HOMESEEN, WhiteWall. 謝謝你們豐富我的人生。

謝謝每一個人出現在我的生命裡的你/妳，無論相遇的時間多長，慶幸我們都曾是彼此生命裡的好人。謝謝未來也將一直在我身邊的你/妳，以後的日子也請多多指教。謝謝老天的眷顧，讓我當一個如此幸運的人。

最後，想把論文送給我最愛的爺爺，希望你會很驕傲拿給朋友看。希望你跟笨狗在天上都過得很好、很好。

游詳閔 謹誌于

政治大學資訊科學所 2019.9.4

DOI:10.6814/NCCU201901233

目錄

DOI:10.6814/NCCU201901233

DOI:10.6814/NCCU201901233

圖目錄

DOI:10.6814/NCCU201901233

表目錄

DOI:10.6814/NCCU201901233

第一章

前言

近年來，諸如 FaceBook、無名小站等 Web 2.0 社群網站如雨後春筍般地的崛起。隨著使用者人數越來越多，這些注重使用者彼此關係的社群網站，使得原本是社會科學的議題，漸漸在電腦科學的領域中發酵。

與社會網絡有關的研究範疇越來越多，光是社會網絡的資料來源就有各式各樣，簡單舉幾個常見的資料庫，像是描述學術論文與其參考文獻之間的網絡、描述研究學者彼此有無參考對方論文關係的網絡、還有描述電影演員合作關係的網絡...等等；其中還有一個常見資料來源，就是前面提到的社群網站。

我們可以從資料來源發現，社會網絡並非受限於人與人之間的網絡關係；事實上，可以把社會網絡這個抽象的名詞想像成一張圖，組成圖的兩個元素就是點(Node)與邊(Edge)；所謂的點，就是一筆資料，而資料可以是有生命的，如 FaceBook 中的使用者；當然也可以是無生命的，如學術論文資料庫 DBLP 中的論文；都可以當成社會網絡中的點；另一個組成社會網絡的基本元素是邊，邊描述著點與點之間的關係，例如 FaceBook 中的朋友關係、DBLP 中論文彼此的參考關係。電腦科學在研究社會網絡的議題時，就是用上述的方法來表示社會網絡。

有關社會網絡的議題非常的多也相當的廣，以社群網站舉例來說；有從社會社會科學衍生而來的，如探討每個人在所屬的社會網絡中扮演的角色是甚麼，即為所謂的角色中心性；也有新起的應用，例如結合社會網絡與廣告行銷，像是如果處在一個預算有限的狀況下，要挑選社會網絡中的哪　幾個人做行銷，可以達到最大的效益；還有針對社會網絡整體演化的研究，像是鏈結(Link)的預測，如應用在 FaceBook 中，鏈結預測即為

DOI:10.6814/NCCU201901233

推薦好友給使用者的功能。除此之外還有更多夠多的研究範疇，上述舉的例子只是關於社會網絡研究的冰山一角，相關的議題可說是不勝枚舉。

除了上述所舉的研究方向之外，還有一項近年來被討論的議題，稱為類別(Label)預測。舉一個實際的例子來說 [3]。，如果今天給定一個行動電話的網絡，每一個號碼視為社會網絡中的一個點，而點與點之間的連線表示其有無通聯紀錄，若有則兩點有連線，反之則無。在這個行動電話的網絡中，可能會出現兩種類別的號碼，分別是「詐騙」與「合法使用」;那麼在這個詐騙電話如此猖獗的現實生活中，如果可以把詐騙的號碼，與合法使用的號碼分別標記出來，那麼一定可以減少更多的受騙案例。但是，真實的情況下，已經確定是詐騙電話的號碼，與確定合法的電話號碼，在整個行動電話的網絡中往往佔非常少的比例。因此，我們便要想辦法利用，這些已知類別的號碼與未知類別的號碼出現在網絡中的關係，設法將整個行動電話網絡中的號碼，都給予適當的類別。這樣的例子，就是所謂的類別預測。

在上例中，類別是二元的，只有詐騙號碼與合法號碼兩種。然而，現實生活中，類別絕對可以是多於兩種的。所謂的類別，是為了讓網絡中的點有所區分而出現的，類別可以是一種行為或是一種狀態；把行為當作類別的話，像是：對政治的立場、購物的喜好、感興趣的議題、或是對某一件事的反應是正面或是負面，都可以是一種類別。若以狀態當作類別的話，像是婚姻的情況、服兵役的情況等等都可以。另外像是生活中的職業如：老師、學生、行政人員，這都算是類別的一種。

然而，現階段關於類別預測的研究，都是針對某個時間點中的社會網絡作預測，也就是在一個靜止的社會網絡中將未知類別的資料，作類別的分類；但這種情況與真實世界並不相符，真實世界裡的社會網絡通常是動態的，意即會隨著時間變化的，不管是網絡的結構或是其中的類別都可能隨著時間而改變。

舉一個動態社會網絡的實際例子，以 FaceBook 來說，假設網絡中有一個類別是每個人喜愛的牛仔褲品牌；隨著時間的變化越來越多人加入 FaceBook，社會網絡因此逐

DOI:10.6814/NCCU201901233

漸擴張，而每個時間階段都會有新的人加進來，如果這些新加入的人已經對於牛仔褲品牌有所喜好，可能影響他們的朋友，因而造成他們朋友喜愛的牛仔褲品牌有所改變。而使用者喜愛的牛仔褲品牌有所改變，也就是所謂的類別會隨著社會網絡而動態改變。

以往對於社會網絡中類別的探討，比起鏈結而言偏少許多；這幾年鏈結預測的研究如雨後春筍般冒出，而隨著鏈結預測日漸成熟，有關類別預測的研究也漸漸受到矚目。通常一個有類別的社會網絡，裡面已標記好的類別是相當稀疏的，也就是在實際的狀況下，一個社會網絡裡通常會有很多點的類別是未知的。而類別預測就是設法將這些未知類別的點，分配到一個最恰當的類別。

本研究即探討在動態社會網絡下的類別預測。在這樣的前題下，會有幾個衍生的問題，第一，由於本研究是針對動態的社會網絡作類別的預測，與以往針對靜止的社會網絡做類別預測的研究，不同的地方在於，所謂動態社會網絡指的是，社會網絡中的點與邊可能會隨著時間增加或減少，此前提與以往靜止的社會網絡中點與邊的數量都是固定的，並不考慮時間之於社會網絡的關係，是兩個不同的出發點；為簡化問題，本研究假設社會網絡會隨著時間擴張，也就是每個時間階段會有新的點加入網絡，網絡中的點以及邊的個數都隨著時間增加。

社會網絡中類別分類的問題，有別於一般傳統的分類問題。最大的不同在於社會網絡中每一筆資料之間是相關的，不像傳統的分類問題中，每筆資料都是獨立的。因此，既然社會網絡中的每一筆資料，在這邊的資料也就是指社會網絡中的點，其之間是彼此息息相關的；那麼，在社會網絡的演化過程中，點與點之間的相互的變化，以及每個點隨著時間之於整個社會網絡意義的變化，事實上蘊藏了許多潛在的類別資訊。

因此，我們要探討的是，在一個未來的時間點中，如何利用網絡的演變過程，以及在未來的時間點中的部分資料，來預測其社會網絡中點所屬的類別，並提高預測的準確率。

以下是本篇論文的結構，在第二章將介紹類別預測的相關研究。第三章則是我們的

DOI:10.6814/NCCU201901233

研究方法流程，包括 3.1 問題說明與定義、3.2 研究架構，包括：分類器的設定與特徵選取方法。第四章為實驗，4.1 為實驗設計、4.2 實驗評估方法與設計。第五章，描述本研究的結論，針對實驗結果作討論，並且說明未來可能的研究方向。

研究的方向將從兩個部分進行，分別是先了解傳統在靜止的社會網絡中的類別預測，並找出可以應用在動態社會網絡中的部分；以及針對解決與時間相關，找出趨勢的概念飄移做探討。

DOI:10.6814/NCCU201901233

第 二 章

相關研究

現階段針對社會網絡中類別預測(Label Prediction)的研究逐漸熱門，Gallagher, B. [3] 將類別預測的方法大致分成兩大類，分別是 Collective Classification 與 Graph based Semi-supervised Learning；大部分的做法，無論分類的著眼點為何，大多是基於這兩大類衍生的；除此之外，還有一個是結合以上兩大類的做法，由 Gallagher, B.等學者 [3] 所提出，稱為 Ghost Edge。事實上，不管是上述的哪一種方法，皆是將社會網絡中的類別預測視為一種網絡中分類的問題 (within-network classification)；

一般的分類(classification)問題中，資料與資料之間彼此是獨立的，分類問題的重點在於，利用已知類別資料的屬性學習出一個類別的分類規則，接著根據此分類規則，利用未知類別資料的屬性，將其分類到所對應的類別中。

然而，社會網絡中的資料，是彼此相關的；換句話說，社會網絡中的資料皆以點表示，而點與點之間有邊的存在，意即資料與資料(點與點)之間是彼此相關的。因此，社會網絡中類別分類的問題核心是，如何利用資料與資料之間連結的關係，設法讓未知類別的資料，分類到所對應的類別中。

2.1 Collective Classification

第一類研究此種問題的方法稱做 Collective Classification [11]，P. Sen 等學者認為社會網絡中每個點都代表一個資料(可能是網頁、人...等)，而每個點都有其屬性與類別；點的屬性都是已知的，而類別則是有些已知有些未知；如圖 2.1 所示，這個例子中每個點代表一個網頁，用橢圓形表示；而每個網頁的屬性以圓形表示，至於類別則是橢圓中的文

DOI:10.6814/NCCU201901233

字，可以看到這個例子中有兩筆已知類別的資料：其類別分別是 SH 與 CH；同時也有兩筆未知類別的資料，用空白的橢圓表示。因此，若給定一個上述的社會網絡，與一個未知類別的資料 (以下用 u 代替)，可以知道以下三種資訊：(1) *u* 的類別與 *u* 本身屬性的關聯、(2) *u* 的類別與 *u* 已知類別的鄰居的關聯、(3) *u* 的類別與 *u* 未知類別的鄰居之間的關聯。

Collective Classification 即是綜合利用上述的三種資訊，來預測未知類別的資料會是甚麼類別，也就是利用這三種結構上的資訊對社會網絡中的資料分類。有別於以往在機器學習(Machine Learning)與資料探勘(Data Mining)中，都是將資料獨立分類，而資料分類的順序並不會影響最後的結果；但是這種設定對於社會網絡中的類別分類並不適用，因為社會網絡中的類別分類的依據是資料在社會網絡中的結構，也就是點與點之間的連結關係，而每個未知的點所被分到的類別結果，則會影響其連結的未知的點；因此在社會網絡中的類別預測也必須考慮將資料分類的順序。

Collective Classification 將做法分為兩大類：第一種是以區域性的分類器為分類基礎的演算法，第二種是將分類問題表示成一個方程式，並用演算法求得最佳解。

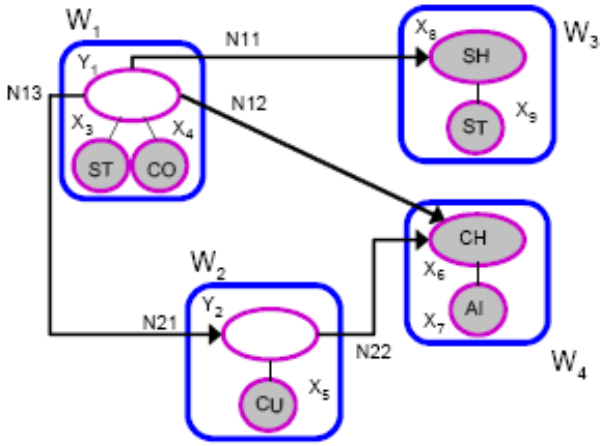

圖 2.1 Collective Classification 示意圖 [11]

DOI:10.6814/NCCU201901233

Collective Classification 以區域性的分類器為分類基礎的演算法，最常見的是ICA(Iterative Classification Algorithm)與 Gibbs Samplings。

ICA 與 Gibbs Sampling 的精神相當類似，皆是在每個回合先暫時決定一個類別給未知的資料，重複幾個回合後，最後指定最佳的類別給某個未知類別的資料。

不同的是，ICA 會不斷的重覆直到每個未知類別資料的類別機率分佈穩定為止，並未規定共要做多少回合才停止；而 Gibbs Samplings 則是一開始便決定總共要做幾回合即停止(稱為 burn-in)，在每個回合裡，同樣也是指定當下最適合的類別給每個資料，並且每次留下記錄，直到做滿一開始定下的次數停止之後，最後選定過程中指定最多次的類別為分類的結果。特別一題的是，由於 Gibbs Sampling 要判斷結果是否已經到達穩定的狀況是一件很複雜的事，因此如何選擇一個適當的次數變成了一個議題。

除了以區域性的分類器為分類基礎的演算法外，Collective Classification 還有另一種解法，是將此種分類問題表示成一個方程式，並用演算法求得最佳解。這一類型的做法，是將問題轉化成配對馬可夫隨機場(pairwise Markov random field)來解釋，常見的做法有 Loopy belief propagation 與 mean-field relaxation labeling 兩種。

Collective Classification 主要的精神在於利用未知類別資料的已知類別鄰居來預測其類別，因此如果社會網絡中已知類別資料數量稀少時，效果會變得非常不好。另外，根據 L. McDowell [11] 所提出的方法，他們認為與其把所有的鄰居全部列入考慮，不如只考慮前 K 個重要性最高的鄰居，會獲得比較好的結果，此種作法稱為 Cautious Collective Classification [11]。那麼，所謂鄰居的重要性之定義是甚麼？

前面有提到，在 ICA 的作法中，一開始對未知類別的點會先利用少數已知類別的鄰居當作分類的依據，給予此未知類別的點一個暫時的類別，而事實上這個暫時類別，是由多個可能的類別中取出機率值最大的一個類別，來當作此點暫時的類別；有鑑於此，L. McDowell [16] 等人便認為，這個暫時類別的機率值，另一層面也代表著在這個點上作類別分類的可信度；也就是說，既然每個暫時類別已經是多個可能的類別中，機率值

DOI:10.6814/NCCU201901233

最大者，那麼如果此暫時類別的機率值越小的話，就表示此次分類的結果不準確的可能性越大；因此，Cautious Collective Classification 有別於原本的 ICA，對每個未知類別的點作類別分類的時候，利用了所有的鄰居：在原始的 ICA 中，無論是已知類別的、或是原本為未知類別後來得到一個暫時類別的鄰居，全都拿來當作分類的依據。Cautious Collective Classification 對未知類別的點進行類別分類的時候，以較嚴格的方式在過濾鄰居；除了已知類別的鄰居一定要挑以外，對於原本為未知類別的鄰居，依據其暫時類別的機率值排序，只挑選前 K 個加入分類的依據，其他未知類別的鄰居，即便已經有了一個暫時的類別，但因可信度太低則不考慮。這就是 Cautious Collective Classification 的作法。

2.2 Local Structure Similarity

上述的做法皆是考慮已知類別鄰居其類別的分佈，做為將未知類別分類的依據。除了考慮鄰居的類別分布之外，C. Desrosiers [3] 提出另一種觀點是考慮未知類別的點，在社會網絡中的結構與哪個已知類別的點較為相像，便認為它們較有可能是相同的類別。

此作法是精神是，若某兩個點在社會網絡中扮演的角色非常相似，那麼此兩點應具有相同的類別。那要如何判斷某兩個點，在社會網絡中扮演的角色是否類似？

此篇研究使用的方法是，計算某個未知類別的點 u 與某個已知類別的點 v，其分別在社會網絡中做 random walk，並記錄每一個經過的點之類別，這麼一來，便可分別得到 u 與 v 在社會網絡中作 random walk 經過的類別序列；接著，再算出 u 與 v 會出現一模一樣的類別序列之機率為何，若 u 與 v 產生出相同的類別序列機率越高；則表示，從 u 和 v 出發越有可能會經過一樣類別的點，因此 u 與 v 這兩點在社會網絡中越有可能是扮演同樣的角色；因此未知類別的點 u 越有可能與已知類別的點 v，有一樣的類別。此方法是屬於 Collective Classification 中的 Relaxation labeling。

DOI:10.6814/NCCU201901233

2.3 Graph-based Semi-supervised Learning

類別預測兩大類方法，除了 Collective Classification 之外，還有另外一種做法是 Graph-based Semi-Supervised Learning，此種做法可以解決 Collective Classification 在已知類別數量偏少的社會網絡中效果不彰的缺點；Graph-based Semi-supervised learning 除了利用已知類別的資料以外，還有利用未知類別的資料一起當作已知的資訊[24]。

這一類方法的基本前提是，在社會網絡所表示的圖中，分布在附近區域的點，其類別皆會相同。因此 Graph-based Semi-supervised learning 必須滿足兩個限制：(1) 類別的一致性(Local Consistency) (2) 類別分布在社會網絡中的平滑性。所謂類別的一致性指的就是散佈在附近區域的點其類別皆會相同；而平滑性指的是，對整個社會網絡的圖而言，類別的分佈必須是平滑的，也就是必須避免在圖上有一區一區不同類別，出現非常明顯的界限的情況。所有使用 Graph-based Semi-supervised learning 的做法皆是在這兩個限制下取得平衡 [3] [23]。

由於此種作法是建立在附近區域的點其類別都相同的假設下，因此對於某些不符合此種假設的社會網絡，效果便會大打折扣。例如：描述師生關係的社會網絡，這一類型的社會網絡，類別為學生的點直接與類別為老師的點相連；這樣一來，某個點所直接聯結的鄰居，其之間類別事實上是相反的。在此種狀況下 Graph-based Semi-supervised learning 的解法便完全不適用，反之 Collective Classification 因為是倚賴已知類別的資料當作訓練資料，因此無論區域類別的一致性高或是低，只要已知類別的資料量夠多都可以做出理想的分類。下表 2.1 列出以上兩種方法的比較。

DOI:10.6814/NCCU201901233

表 2.1 Collective Classification 與 Graph-based Semi-Supervised Learning 比較表

Performance	Classification	Semi-supervised Learning
very few labeled nodes in the networks	largely degraded due to the lack of sufficient neighbors	Outperformance
Different dependency structures	powerful ability to learn various kinds of dependency structures	affected in negative auto-correlation

由上表 2.1 可以看出，在已知類別資料數量較少的情況下，Graph-based Semi-supervised learning 會得到較佳的分類結果；而在不知道類別的區域一致性是高或低，特別是區域一致性低的情況下，使用 Collective Classification 會有較佳的結果。

2.4 Ghost Edges

除了 Collective Classification 與 Graph-based Semi-supervised learning 之外，Gallagher, B. 等學者 [3] 也提出了一個可以解決以上兩者最大缺點的方法，稱為 Ghost Edge。其主要的精神是利用圖中所有已知類別的點來當作分類的資訊，在這篇研究中，主要針對只有二元類別的社會網絡作探討。

此做法是將未知類別的點，與圖上每一個已知類別的點都以 Ghost edge 相連；此種做法就是參考 Collective Classification 用已知類別當作分類基礎的精神。將未知類別的

DOI:10.6814/NCCU201901233

點與圖上所有已知類別的點以 Ghost edge 相連之後，分別計算每一個 Ghost edge 的權重，再根據每一條 Ghost edge 的權重，算出某個已知類別的點，對此未知類別的點的貢獻是多少。若權重越大，則表示此未知類別的點，越可能受到連結於此 Ghost edge 的已知類別的點影響；因此越有可能與其屬於相同的類別。

DOI:10.6814/NCCU201901233

第三章

研究方法

3.1 問題定義

[定義一]

一個動態社會網絡 $G = < G_1, G_{2,...}G_T, G_{T+1}>$，是由多個不同時間點的社會網絡所組成的序列，其中 $\forall t, 1 \leq t \leq (T+1)$, $G_t = (V_t, E_t, L)$，是一個 Unweighted graph.

V_t 為 G_t 的節點集合，E_t 為 G_t 的連結集合。

$|V|_t$ 表示 G_t 中所有節點的個數，$|E|_t$ 為 G_t 中所有連結的個數，|L|為 G 中所有類別的種類個數。

對於一個動態演化的社會網絡而言，$|V|_t$、$|E|_t$ 可能隨著時間 t 而改變。在本研究中，|L|為一常數，類別的種類是固定的。且動態網絡逐年增加規模，因此 $V_t \subseteq V_{t+1}$ 且 $E_t \subseteq E_{t+1}$

[定義二]

給定 $G = < G_1, G_{2...}G_T, G_{T+1}>, \forall t, 1 \leq t \leq (T+1)$

(1) $\forall G_t, 1 \leq t \leq T$，所有節點的類別皆為已知。

(2) G_{T+1} 只有部分節點為已知類別，其中 X_{T+1} 為已知類別的節點集合，Y_{T+1} 為未類別的節點集合。 $X_{T+1} \cup Y_{T+1} = V_{T+1}$

則動態社會網絡的類別預測問題是利用 $G_t = (V_t, E_t, L)$，$t \in [1,T]$，以及 X_{T+1}；將所有 $y_{i\,(T+1)} \in Y_{T+1}$ ，分類至對應的類別 l_i，$l_i \in L$。

圖 3.1 為動態社群網絡之類別預測示意圖，其中黑色點為已知類別的節點，白色點為未知類別的節點；針對動態的社群網絡，目標是利用所有已知類別的資料，針對未知類別的節點預測其類別。

DOI:10.6814/NCCU201901233

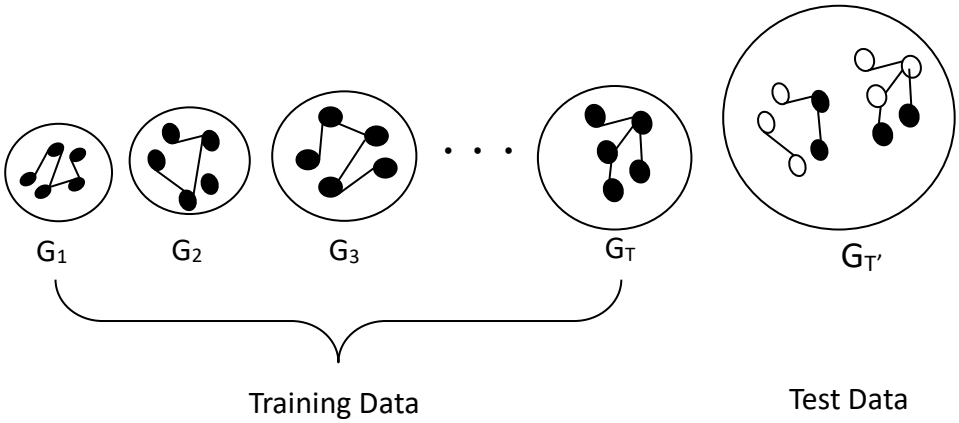

圖 3.1 問題定義示意圖 (黑色點為給定的資料)

3.2 研究架構

本研究的架構主要分三個部分：Base Classifier Learning、Ensemble Box Learning、Labeling，如下圖 3.2 所示。

在 Base Classifier Learning 中，主要的重點在於使用動態社會網絡過去的演化資料，對每一時間點的的社會網絡產生一個分類器，使得過去的動態演化過程，可以當成判斷未來時間點的社會網絡類別預測的線索。

Ensemble Box Learning 則針對 Base Classifier Learning 中所產生的眾多分類器，我們利用一個 Ensemble Box 組成了 Ensemble Classifier，當成最後作類別分類時的分類器。

在最後一個階段 Labeling 中，我們運用了 Collective Classification 中的方法 ICA (Iterative Classification Algorithm) 將上一階段中產生的 Ensemble Box，用於未知類別的節點來作類別的預測，最後產生每一個點都有類別的 G'_{t+1} 。

DOI:10.6814/NCCU201901233

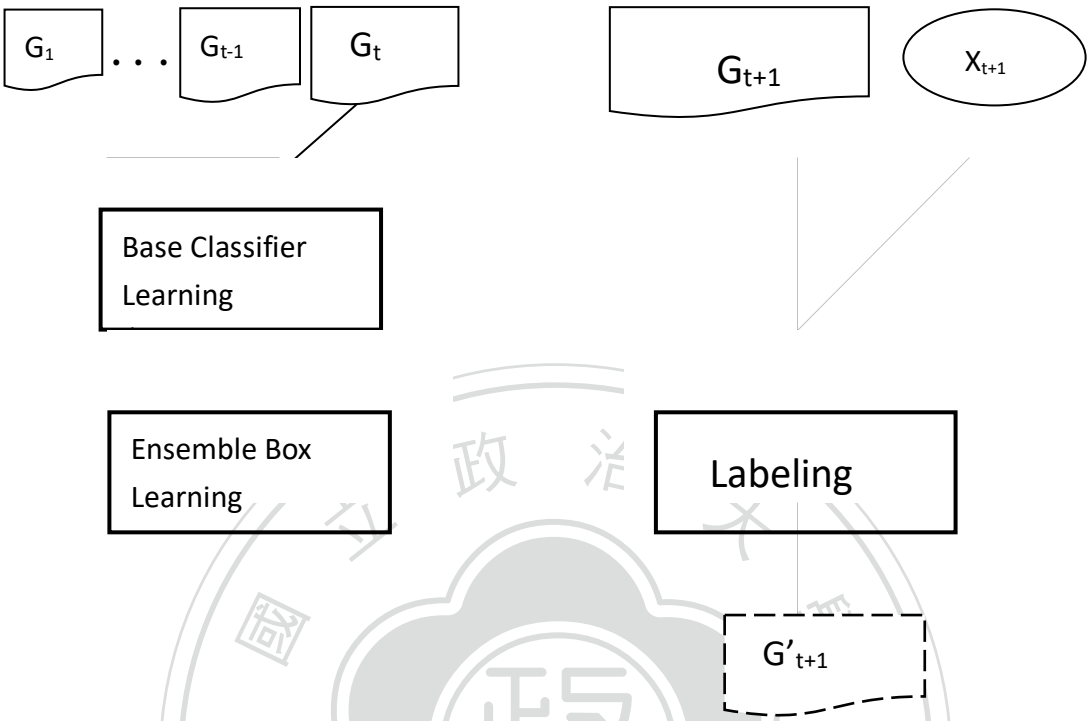

圖 3.2 研究架構流程圖

3.3 Base Classifier Learning

Base Classifier Learning 這個步驟的重點使用社會網絡演化的資料；訓練資料為：$G = < G_1, G_2, \ldots, G_T >$。 針對每個 $G_t = (V_t, E_t, L)$，$1 \leq t \leq T$，都使用 Base Classifier 學習出一個分類器出來，此分類器就代表了社會網絡在 t 時間點的類別規則。

社會網絡的類別預測，最直覺的方法就是觀察大部分直接相鄰的節點，多數為甚麼類別，則指定其最多數之類別給此未知類別的節點。

若將此方法對應到真實的情況，當社會網絡中點與點彼此聯結數量較少時，每個人的確受到身邊的人影響的程度非常大，尤其是直接相連的第一層鄰居；在點與點之間，

DOI:10.6814/NCCU201901233

以及社會網絡中整體的連結數量都不多時，如果利用此直觀的方法，假設鄰近的點彼此類別的同質性較高，將新加入的點作類別的分類或許是符合邏輯的。

然而，當社會網絡隨著時間逐漸增長的時候，網絡中每個點的聯結越來越多了，這個時候，每個點受到第一層鄰居影響的程度，也許也會受到改變，有可能除了第一層的鄰居以外，還會受到第二、三層、甚至更多層鄰居的影響。

如下圖 3.3 為例：點 *i* 為未知類別的節點，點 *i* 與節點 N_1、N_2、N_3、N_4 直接相連，若是以最直觀的方法，直接觀察其直接相鄰的節點；可能會將點 *i* 的類別分類成 A；但是在社會網絡中，雖然直接與點 *i* 相連的點只有四個，但是 N_1、N_2、N_3、N_4 與社會網絡中其他點是有連結存在的，由於網絡中的節點彼此會相互影響，所以不論是點 *i* 的直接鄰居，包括點 *i* 的間接鄰居，都可能會影響著點 *i* 的類別；因此在這種情況下，只看直接相鄰的鄰居，當做分類的特徵也許就不足夠當成分類的依據。

下一章節，將說明本研究中針對每一筆資料，所使用的特徵、以及使用的 Base Classifier。

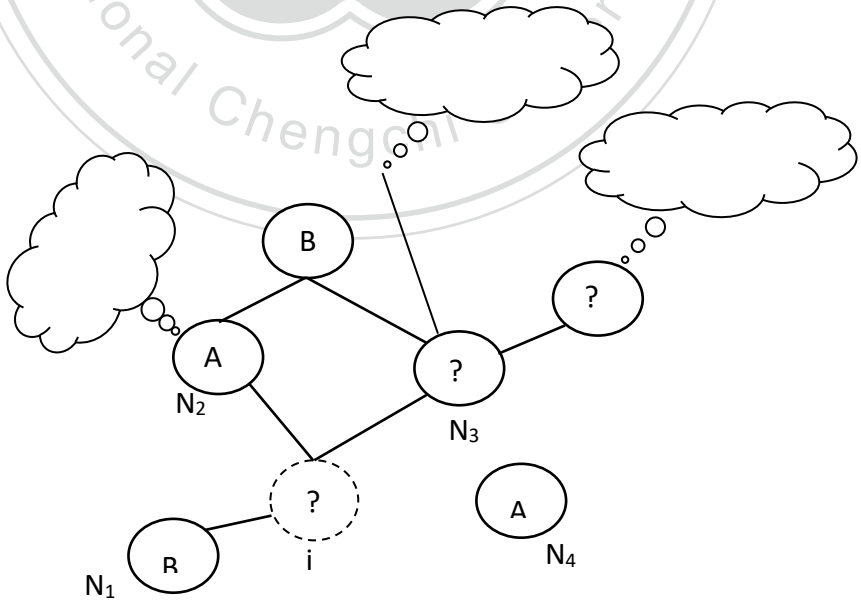

圖 3.3 類別預測示意圖

DOI:10.6814/NCCU201901233

3.3.1 Ghost Edge

對一個已知類別的點稀疏分佈的社會網絡，如下圖 3.4 所示，為了盡可能利用已知類別的點來當作類別預測的根據，Ghost Edges [4] 將每一個未知類別的點，以 Ghost Edges 與網絡中所有已知類別的點相連。如此一來，即便原本不直接相鄰的已知類別，都可以透過網絡中的結構，成為對未知類別的點做分類時的依據。

若以 *u* 表示某個未知類別的點，*v* 表示某個已知類別的點，若 *v* 與 *u* 並未直接相連，則而 Ghost edge 將 *u* 與 *v* 相連，藉著 Ghost edges 利用網絡的結構算出 *u* 與 *v* 之間的 proximity；此 proximity 即代表 *u* 與 *v* 在社會網絡中的遠近關係。

某兩點 *u* 與 *v* 之間 Ghost edge 的 proximity 值，是來自在社會網絡的圖中，從 *u* 出發做 Random Walk with Restart (RWR)，會到達 *v* 的機率。若從 *u* 出發做 RWR 到達 *v* 的機率越高，則此 *u*，*v* 相連的 Ghost edge 其 proximity 值越高；表示 *u* 越可能受到 *v* 的影響。

但若以 RWR 到達某個已知類別點 *v* 的機率，來判斷某個未知類別的點 u，是否受 *v* 影響，會出現以下的問題。

若 *u* 與 *v* 直接相連，那麼 *v* 對 *u* 的 proximity 必然較高，意即有較大的權重。然而，這麼一來又回到 Graph-based Semi-supervised learning 的缺點，也就是若是在區域一致性 (Local Consistency)較低的網絡中，如果網絡中直接相鄰的點剛好都是相反的類別時，其分類的效果將會變得極差。因此為了克服此問題，[4] 便將 RWR 變形。

其改變 RWR 原本一次走一步的定義，改成一次走兩步。計算 proximity 的方式變成：從某個未知類別點 *u* 出發，做 even-step RWR 到某個以 Ghost edge 相連的已知類別點 v 的機率。

這麼一來便可以克服 Graph-based Semi-supervised learning 無法在區域類別一致性較低的社會網絡中做分類的缺點。因為若在區域類別一致性極低的社會網絡中，鄰居互為相反的類別，一次走兩步剛好會避開相反的類別。而若在區域類別一致性高的社會網

DOI:10.6814/NCCU201901233

絡中，既然附近鄰居都會是相同的類別，那麼走兩步依然會是相同的類別。proximity 是利用[19]的 Random Walk with Restart 提出的方法得之，以下將詳細說明。

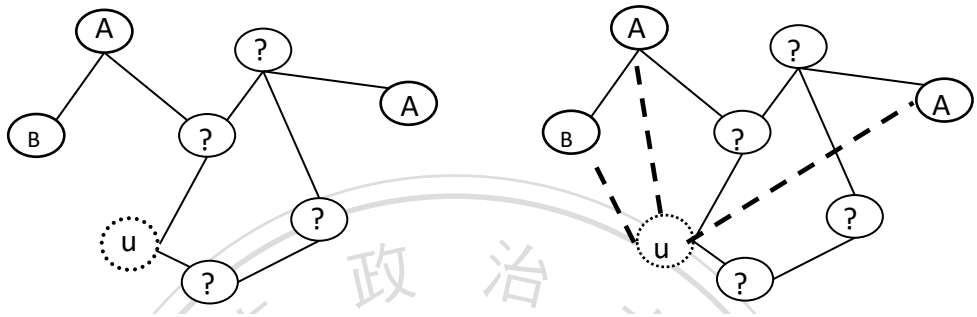

圖 3.4 Ghost Edges 示意圖

3.3.2 Random Walk with Restart

本研究中，兩點(*i, j*)之間的 RWR 值，代表 *i* 與 *j* 在網絡中結構上的遠近，[13]定義如下

$$\vec{r_j} = c\widetilde{W}\vec{r_j} + (1-c)\vec{e_j} \qquad (1)$$

說明：

$\vec{r_j}$是排序後的向量，是一個 n× 1 的矩陣，r_{ij} 為點 i 到點 j 之間相關的分數。

C 是從任一點 i 出發，回到原點 *i* 的機率；0≤c≤1

$\widetilde{W}$ 是一個正規劃之後的 weighted graph，$1 \le i, j \le n$ (n 為所有點的個數)

$\vec{e_j}$ 是一個起始向量，大小為 N-1：記錄從哪一個點開始，以 *i* 點出發，第 *i* 個元素為 1，其他皆為 0

RWR 的做法為：從某一個點 *i* 出發，隨機造訪在圖上的點，每一次有 1-c 的機率回

DOI:10.6814/NCCU201901233

到原點 i；本研究中，c = 0.9，且每一條邊的 weighted 都為 1。

若圖中任兩點(i，j)之間有越多權重值大的邊，或是存在最短路徑，那麼(i , j)之間的 proximity 會越高。

3.3.3 Base Classifier

本研究中，我們使用的 Base Classifier 為 Random Forest [1]。以下詳細說明 Random Forest。

Random Forest 是一種以決策樹 (Decision Tree) 為基礎分類器的 Ensemble Classification 方法。其分割訓練資料的方式是從資料的特徵為著眼點下手，假設某個基礎分類器的設定是 K 個特徵，那麼 Random Forest 就會在建立此分類器的決策樹時，在每個 split node 先隨機選出 K 個特徵，再從這 K 個特徵中找出最好的特徵，當作節點(split node)，以此方式建立起決策樹。

然而，既然 Random Forest 是一種 Ensemble Classification，其基礎分類器的個數當然不只一個，也就是說會有多個決策樹，其中每一個決策樹所選定的特徵數可以不同，這表示每一個決策樹的高度可能不一樣，而最後的準確率也會關係著基礎分類器的歧異度，如果歧異度越大則表示每個基礎分類器之間越獨立，那麼就越有可能分的越準確。因此，在指定各個決策樹的特徵數量時，要特別注意，通常是以$\log_2 d+1$ 為主，d 為原本訓練資料中的總特徵數。

3.4 特徵選取

在 Base Classifier Learning 當中，最主要的目的是為了利用社會網絡演化的資料，來做為分類的依據之一。對演化過程中每一個時間點 t 下的社會網絡 $G_t = (V_t, E_t, L)$，$t \in [1,T]$；我們針對每個 $x_i \in X_t$ 皆產生一組特徵，最後使用 X_t 當作 Base Classifier 的訓練資料，以學習出社會網絡在時間 t 時的類別規則。以下將說明特徵產生的方法 [4]。

DOI:10.6814/NCCU201901233

首先，利用前一章節所題及的計算 proximity 的方式，算出每一個點 $x_i \in X_t$，對網絡上其他所有點的 proximity 值。

接著，再將每一個點按照其proximity值分成六個等級；第一個等級為proximity值在前3%的點，接著第二個等級是proximity介於前3%-6%的點、第三個等級為前6%-12%的點、第四個等級為12%-25%的點、第五個等級為25%-40%的點、第六個等級為40%-80%的點。

如此一來，對點V而言，依照proximity值整個社會網絡中的點就可以分成六個等級，再加上直接相鄰的鄰居，一共為七個等級。假設社會網絡一共有兩種類別，分別為A以及B。

接著，再看點V在第i階層，$1 \leq i \leq 7$的鄰居中，屬於類別A的點有幾個，以及類別為B的點其個數為何，就以此當做點V的特徵。我們也是利用一樣的方法，得到社會網絡中每個點的特徵。

如此一來，在每一個時間t，$t \in [1,T]$中，我們有了當時的社會網絡$G_t = (V_t, E_t, L)$當作訓練資料，也利用B. Gallagher[4]的方法得到了特徵，便可以使用Base Classifier對每一個時間點 t的社會網絡學習出一個分類器，此分類器即代表此社會網絡在時間點t時，其類別分類的規則。

3.5 Ensemble Box Learning

再次以上圖 3.3，類別示意圖為例：在一般只考慮靜態社會網絡的類別預測中，若要對點 i 作類別的預測，無論單純考慮第一層鄰居 N_1、N_2、N_3、N_4，或是連帶考慮間接的鄰居，都是在此靜態情況下所做出的預測。但是一個社會網絡的演化過程中，其類別分類的規則是有可能隨著社會網絡的變化有一些趨勢產生；例如：受到第一層鄰居影響的程度遞減、受到不同層鄰居影響程度的變化...等等；這些趨勢，事實上是緊緊關係著網

DOI:10.6814/NCCU201901233

絡上所有的點，像是點 N_2 與點 N_3 所連結的雲朵，代表著在社會網絡上類別分類的趨勢，而這個趨勢都是有可能隨著社會網絡的變化改變的。我們希望可以藉著動態社會網絡的演化狀態，來得知此分類規則的趨勢。

此類與時間有關的演變資訊，機器學習領域中將此稱為概念飄移(Concept Drift)；因此我們的解法中，必須利用 Concept Drift 的解法，來找出分類規則改變的趨勢。以下將簡短說明 Concept Drift、並詳細說明本研究中 Ensemble Box Learning 的方法。

3.5.1 Concept Drift

概念飄移(Concept Drift)指的是資料在一連串的變化中出現的趨勢，而所謂的 Concept 可以是任何想要預測的概念；例如在氣象預報中，溫度、濕度等等都可以是一種 Concept；又例如在網路行銷中，商家可能想藉由過去的資料，了解消費者在不同季節或是不同時節的消費狀況，來預測消費者下一個可能的動作機制，就可以正中下懷釋出折扣或是大量廣告，進而達到最高的效益；在此例中，消費者的消費模式，也是一個 Concept。

前面提及本研究利用了社會網絡隨著時間增長的演進過程，設法找出位於下一個時間之社會網絡中的點，及其正確類別的對應脈絡。事實上，我們在本研究中，把這個可能會隨著社會網絡演變的類別分類規則，視為我們問題的 Concept。

Indre Zliobaite[25]等學者指出 Concept Drift 變化的方式可以分成四種，以下圖 3.5 為例，假設此圖中的 concept 表示產生資料的機率模型，此例中每一個圓柱型都是一個機率模型也就是一個 concept

(1)Sudden drift：

在 Sudden drift 的情況下，concept 會在某一個時間點忽然的改變，也就是說，資料原本全部出自下方的機率模型，卻在某一個時間點，變成了全部出自上方的機率模型，

DOI:10.6814/NCCU201901233

可以看到在接近中心的時間點，concept 出現了全然不同的改變。

(2)Gradual drift：

在 Gradual drift 的情況下，資料屬於兩個不同 concept 出現的機率則是逐漸改變，隨著時間，資料漸漸的從屬於下方的 concept 變成屬於上方的 concept 這種狀況稱為 Gradual drift；

(3)Incremental drift：

在 incremental drift 的情況下，除了原本上下兩個截然不同的 concept 之外，還會有許多介於這兩者之間的 concept，例如在此例中，隨著時間資料的出處就會漸漸的從下方機率模型，轉換到一個介於下方與上方之間的機率模型，以此類推，漸漸的出自於上方的機率模型。

incremental drift 與 gradual drift 最大的不同在於，gradual drift 只有兩種 concept，而資料屬於這兩種 concept 的機率會逐漸改變；incremental drift 還多了許多介於兩個極端之間的 concept，資料是靠著這些介於兩者之間的 concept 逐漸轉換的

(4)Reoccurring context：

最後一種 concept drift，稱為 reoccurring context，指的就是 concept 會隨著時間反覆有周期性的轉換，例如現實生活中，有時候會流行復古風，這樣的例子即為 reoccurring context。

Concept Drift的解法有許多，概括而言主要分成兩種：(1) Trigger based Learner與(2) Evolving based Learner [21] [25] [15]。Trigger based Learner 主要是用在concept會突然改變的情況下，例如sudden drift的情形。然而，本研究關注的對象是，類別分類的規則受到社會網絡演化影響而產生的變化，而這個類別分類規則的變化過程，不會倏忽出現大

DOI:10.6814/NCCU201901233

幅度的改變，而是潛移默化的。

若由 [23] 提出得四類Concept Drift分類，本研究隸屬於incremental drift的情況，因此並不適用trigger based learner的解法，需要的是evolving based learner。

相對於Trigger based Learner會偵測concept的變化，立即做更動；Evolving based Learner則是不主動偵測 concept，取而代之的是，不斷的調整 learner 本身，使其不斷的適應新的訓練資料，使其變的越來越準確。Evolving based learner這樣的特性，正是我們需要的。

DOI:10.6814/NCCU201901233

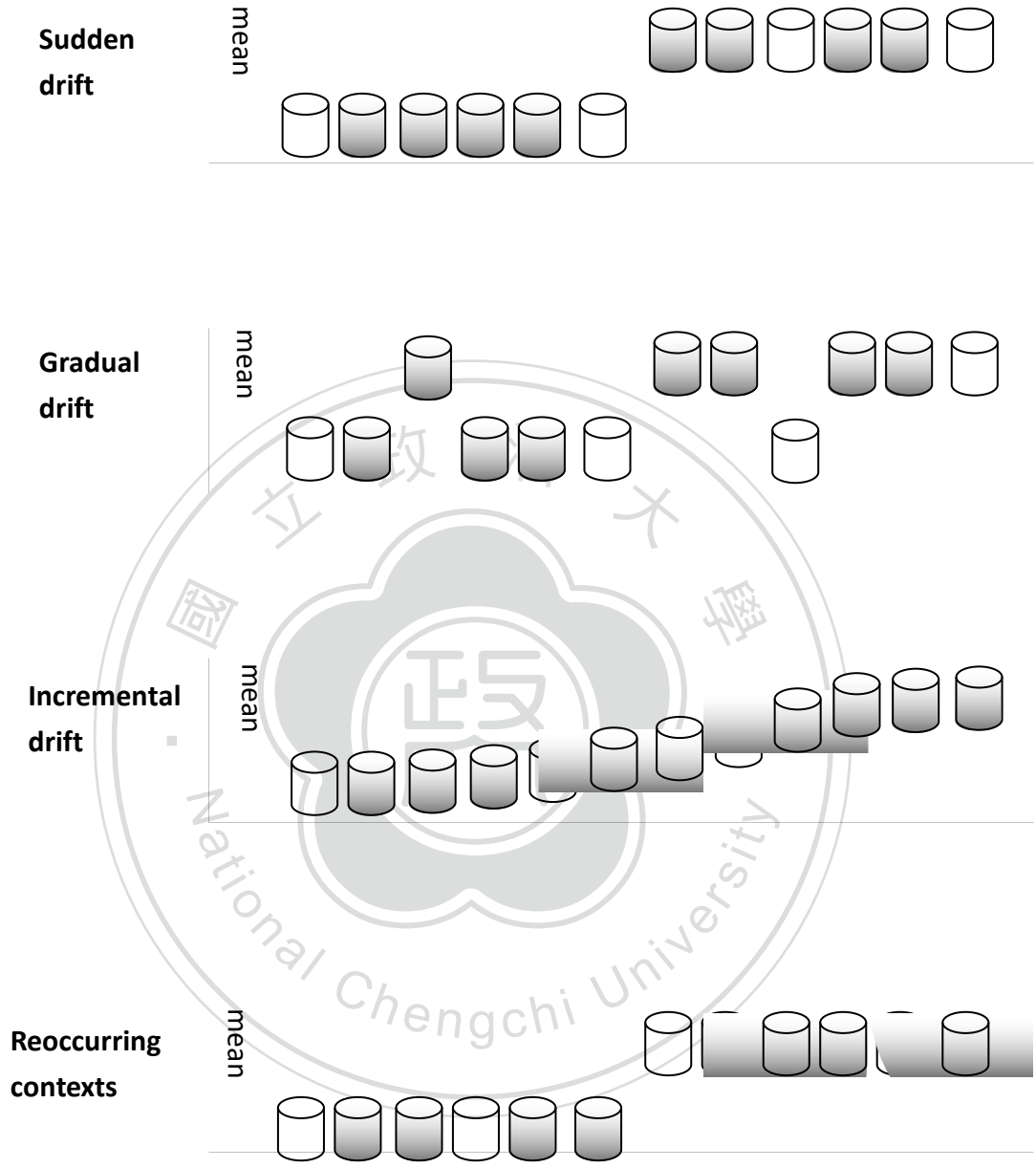

圖 3.5 概念飄移種類示意圖 [15]

Ensemble Classification是Evolving based learner常見的一種方法，此為資料探勘領域的分類問題中，為了提高分類的準確率因應而生的方法。Ensemble Classification的精神在於，若是單一分類器很難分的準確，那麼不如利用多個基礎分類器(based classifiers)來分類，

DOI:10.6814/NCCU201901233

而最後的分類結果則採用多數決的方式，利用此方法來提升分類的準確率。

此外，只要每個基礎分類器是彼此獨立的，並且每個基礎分類器的準確率比隨機猜來的好，那麼準確率就會比使用單獨一個分類器來的高[17]。

Ensemble Classification的關鍵是必須將原本的訓練資料分割成數個訓練資料；原因是Ensemble Classification是利用多個基礎分類器來做分類，那麼假設某個Ensemble Classifier是以N個基礎分類器所組成，那就必須要將原本的訓練資料，分割成N筆訓練資料，以產生N個分類器。最後再結合這n個分類器，綜合分類的結果，對未知的資料作分類。

3.5.2 Ensemble Box

- Streaming Ensemble Algorithm (SEA) [18]

SEA是針對data streaming發展的演算法；由於data streaming的資料量非常的大，若要將所有的data拿來當成訓練資料，記憶體與時間的代價都會很高。

因此SEA的做法是，設定一個固定大小的window size與一個固定基礎分類器數量的Ensemble classifier。新資料進來時，便將此window中的資料，當成訓練資料產生一個分類器；並且將此分類器用於下一次產生的資料中，若效果超過門檻值且Ensemble Classifier中的基礎分類器數量未達上限，則將此分類器加入Ensemble Classifier當中。最後，以此Ensemble Classifier當成最終的分類器，對新的資料做分類。SEA演算法如圖3.6所示：

DOI:10.6814/NCCU201901233

Streaming Ensemble Algorithm (SEA)

while more data points **are available**
  read d points, creating training set D
  build classifier ***Ci*** using D
  evaluate classifier *Ci-1* on D
  evaluate all classifiers in ensemble E on D
    If E not full
      insert *Ci-1*
    else if Quality(Ci-1) > Quality(Ej) **for some** j
      replace Ej with *Ci-1*
  end

圖 3.6 SEA 演算法 [11]

圖 3.6 SEA 演算法 [11]

給定已知的動態社會網絡 $G = < G_1, G_2, ..., G_T >$，經過了上一個步驟 Base Classifier Learning 之後，我們會得到一個 Classifier 的序列 $CF = < CF_1, CF_2, ..., CF_T >$ ，一共 T 個 base classifiers；$\forall CF_t, 1 \leq t \leq T$，都是由對應的 G_T 當作訓練資料，經過 Base Classifier Learning 而產生。

我們利用 SEA[18]的概念，設定一個 Ensemble Box，此 Ensemble Box 包含了 K 個 base classifiers，$1 \leq K \leq T$。針對每一年 G_T 當測試資料，我們拿以 G_{T-1} 當訓練資料所產生的分類器 CF_{t-1} 對 G_T 做測試；若分類準確率高於門檻值δ，且 Ensemble Box 中的 base classifiers 尚未到達上限，則將 CF_{t-1} 加入 Ensemble Box ;

如果分類準確率高於δ，而 Ensemble Box 中的 base classifiers 已達上限；則先比較若以 CF_{t-1} 替換掉 CF_j ，$1 \leq j \leq t$，是否使得 Ensemble Box 的準確率提升，若有，則以 CF_{t-1} 取代 CF_j 加入 Ensemble Box 中。

我們採用此種作法產生 Ensemble Box 之目的在於，由於動態社會網絡中類別分類的規則是漸漸改變的，那麼我們利用此種方式便可以留住針對下一個時間階段也依然準

DOI:10.6814/NCCU201901233

確的類別規則。

3.6 Labeling

經過了上一個步驟 Ensemble Box Learning 之後，我們便得到一個 Ensemble Box 當作我們最終的分類器；我們將 Ensemble Box 中的 base classifiers，稱為 Remained classifiers，這些 Remained classifiers 即代表動態社會網絡演化的過程中，所留下來的類別分類規則，由於我們在 base classifier learning 的步驟，使用的特徵為各個階層的鄰居中，屬於各個類別的鄰居數。因此，這些 Remained classifier 事實上潛在隱含著，對於某個未知類別的節點，其受到各個階層的影響之權重。

針對 G_{T+1} 中每一個未知類別的節點，使用此 Ensemble Box 當作最終的分類器，以多數決的方式，來預測 G_{T+1} 中每一個未知類別節點的類別。

在 Labeling 的步驟中，我們使用了 Collective Classification 的做法，加入了 ICA 在我們的 Labeling 步驟中。下一章節將說明 ICA 的作法。

3.6.1 Iterative Classification Algorithm

屬於 Collective Classification 中 Based on Local Classifier 的 ICA，其基本概念是：若給定一個未知類別的資料 u、以及所有與 u 直接連接的資料、與一個分類器 f；假設 u 的所有直接連接的鄰居其類別都是已知的，則 f 即可利用所有鄰居的類別，產生出 u 屬於各個類別的機率是多少，這樣一來就可以選擇機率最高的當作 u 的類別。

但是，事實上對一個未知類別的點而言，並非所有鄰居的類別都是已知的；也就是說，一個未知類別的資料也有可能是另一個未知類別的鄰居；那麼究竟要如何開始？

如圖 3.7 所示 ICA 的設定為，對每個未知類別的點，一開始只利用少數已知類別的鄰居，來做為類別分類的依據；這樣一來，這個未知類別的點便有了暫時的類別，而與

DOI:10.6814/NCCU201901233

其相鄰且類別未知的點，便可以利用這個暫時的類別，作類別的分類，暫時選擇每個階段得到最有可能的類別，再繼續做完網絡中其他未知的資料；接著，不斷的重複這樣的回合，直到每個未知資料所屬的類別都穩定為止。ICA 的演算法如下圖。

Iterative Classification Algorithm(ICA)

For each node $Y_i \in Y$
  do// bootstrapping
    //compute label using only observed nodes in N_i
    Compute attributes a_i only $X \cap N_i$
    $y_i \leftarrow f(a_i)$
end for
repeat// iterative classification
  generate ordering O over nodes in Y
  for each node $Y_i \in O$ **do**
    // compute new estimate of y_i
    Compute a_i using current assignments to N_i
    $y_i \leftarrow f(a_i)$
  end for
until all class labels have stabilized or a threshold number of iterations have elapsed

圖 3.7 ICA 演算法 [11]

在 Labeling 的步驟，有無 ICA 的差別在於，未知類別的節點之特徵是否為固定的。

假設，在社會網絡中的節點共有 N 個，其中已知類別的節點共有 K 個，那麼，若在沒有 ICA 的情況下，網絡中其他 N-K 個未知類別的節點，其特徵皆來自於此 K 個已知類別的節點，特徵值不會改變。然而，若在有 ICA 的情況下，在回合開始時，第一個未知類別的點被暫時分類至某類別後，此時其它未知類別的節點，便會把這個已經有暫時類別的節點，加入特徵中；因此，在 ICA 的過程裡，最後一個被分類的未知類別節點，其特徵是來自於前面的 N-1 個節點，包括原本類別為已知的節點，以及原本類別為未知的節點，因此 ICA 的回合中，每個未知類別的節點，事實上其特徵值是一直改變的。

DOI:10.6814/NCCU201901233

本研究在 Labeling 的步驟，加入了 ICA 的方式，以回合制的方式，逐漸預測所有未知類別的節點其類別為何。

DOI:10.6814/NCCU201901233

第 四 章

實驗

4.1 資料庫

本研究使用的資料庫來源為 IMDb (Internet Moies Database)。其中，以每一個演員當成節點，以合演過同部電影當作決定連結的條件，若某兩位演員有共同合演過一部電影，則此兩節點之間有連結。我們收集 1980 年至 1995 一共 15 年的電影，以及其演員，當作實驗中的動態的社會網絡。實驗目的為，利用動態網絡的資料來預測演員的類別。

表 4-1 資料庫中每個年份的演員數量

年份	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
人數	193	220	270	324	393	456	513	591	708	812	963	1552	2265	2803

4.1.1 資料庫特性

■ **動態網絡**

原始的資料為 1982-1995 年的電影資料，一共 14 年份。我們利用 1982-1993 年當作訓練資料。因此，我們實驗中的動態網絡 $G = <G_1, G_2, \ldots, G_T>$，每一個 G_T 都是當年的社會網絡，$\forall G_{t-1} \subseteq G_t$ ，$2 \leq t \leq 12$。

■ **電影類別**

IMDb 一共將電影種類分成 24 類。分別為：Drama、Thriller、Horror、Action、Comedy、Adventure、Romance、Fantasy、Sci-Fi、Music、Mystery、Crime、Biography、History、

DOI:10.6814/NCCU201901233

Family、Animation、War、Sport、Musical。

其中每一部電影包含一個以上的電影類型，為了簡化問題，我們根據視訊分段領域將電影歸類的方法[2] ，一共歸納成常見的四類，分別為：Drama、Action、Comedy、Horror。其中，由於 Thriller 與 Horror 在電影類型上十分相近，我們將 Thriller 一併列入 Horror 類。其它不屬於這四類的電影類型，我們將之分成第五類 Others。

■ **演員類別**

在訓練資料的產生過程中，演員的類別，是根據其演過的電影大多數為何種類型所決定。為了更能夠確切得看出演員的類別，我們以三年當作一個 sliding window，演員的類別則由此三年中演過的電影決定。

例如：在 1984 年的社會網絡中，此時社會網絡中所有演員的類別，是取決自 1982-1984 年間，演過的電影中最多數的電影類型。假設某演員 A 在此三年內一共演出了兩部電影，其電影類型分別為{ Drama, Romance, Comedy}與{ Drama }。首先我們依照電影種類，加總此演員在這三年內演過的電影數量，可以看出演員 A 在此三年的類別分佈為：Drama*2、Comedy*1、Romance*1。那麼我們將此演員的類別視為 Drama。

然而，若某演員於此三年演出的電影，其最多數的類型不只一種時，我們則以隨機的方式，在這些最多數的類別中，隨機選出一個當作此演員在此三年的類別。

■ **訓練資料的特徵**

由於目標是預測每個演員在下一個時間階段的類別，針對每個在 G_{T+1} 中未知類別的演員 *i*，我們以 [4] 的作法，利用 proximity 值將網絡中已知類別的的點 *u* 對 *i* 分成以下六個等級。X_{t+1}，為 G_{T+1} 中已知類別的點集合。

第一個等級：proximity(*u*,*i*)為前 3%，$u \in X_{t+1}$

第二個等級：proximity(*u*,*i*)為前 3%-6%， $u \in X_{t+1}$

第三個等級：proximity(*u*,*i*)為前 6%-12%，$u \in X_{t+1}$

DOI:10.6814/NCCU201901233

第四個等級：proximity(*u*,*i*)為前 12%-25%，u ∈ X_{t+1}

第五個等級：proximity(*u*,*i*)為前 25%-40%，u ∈ X_{t+1}

第六個等級：proximity(*u*,*i*)為前 40%-80%， u ∈ X_{t+1}

第七個等級：直接與 i 相鄰的點 u，u ∈ X_{t+1}

本實驗中，我們將電影類型一共分成五類，代表網絡中類別的種類個數為 5。GhostEdges [4] 中，以每個等級中屬於各個不同類別的相連點個數當成特徵值。因此對每個演員 a 而言，一共有 35 個特徵。

若以某特徵值為例，在所有與 A 的 proximity 值屬於第一個等級的已知類別節點中，總共有多少個節點其類別是 Drama，此節點的個數即為特徵值。以此類推，一共有 7 個等級，5 種類別；

因此每個演員會有 35 個特徵值。代表第 n 個等級的鄰居中，屬於 m 類型鄰居個數共有幾人。其中 $1 \leq n \leq 7$ 且 $m \in \{Drama, Action, Comedy, Horror, Others\}$

4.2 實驗設計

在實驗中我們設計了三種實驗：

[實驗一] Ensemble Box 與 wvRN 以及 GhostEdgesL 的比較

第一個實驗，我們設定了兩組分類器來做比較。分別為wvRN [9] 與GhostEdgesL [4]。

針對wvRN以及GhostEdgeL，我們給定X_{T+1}，為時間點T+1時的社會網絡中的已知類別點集合，當作其訓練資料。X_{T+1} 是以隨機的方式，產生$|V_{T+1}| \cdot K\%$個已知類別的點，其中K={10, 20, 30, 40, 50, 60, 70, 80, 90}。

而針對Ensemble Box，其訓練資料為$G = <G_1, G_2, \ldots, G_T>$，以及$X_{T+1}$；其中$X_{T+1}$與wvRN、GhostEdgeL的訓練資料相同。

DOI:10.6814/NCCU201901233

由於X_{T+1} 是以隨機的方式產生，因此實驗數據為隨機產生20組X_{T+1} 所得的準確率平均。

- wvRN (weighted-voted relational neighbor classifier) [9]。針對某個未知類別的點*i*，給定與點*i*直接相連的點集合N，wvRN利用方程式(2)，計算點*i*屬於每個class *c*的機率，最後指定機率值最大的class給點*i*。本實驗中$w(n, n_j)$皆等於1

$$P(C|\, n) = \frac{1}{Z}\sum_{\{n_j \in N\,|\text{label}\,(n_j)=c\}} w(n, n_j) \qquad (2)$$

- **GhostEdgesL**

GhostEdgeL [4]，是一種以logForest當作分類器的分類方法。logForest是與Random Forest一樣概念和做法的Ensemble classifier。不同的是，logForest以Logistic Regression取代了原本在Random Forest中的decision trees成為基礎分類器。在GhostEdgeL中，使用的Logistic Regression個數為500個，參考的特徵為$\log_2 35+1$ ，一共7個特徵值。

- Ensemble Box

Ensemble Box為考慮動態網絡演變過程的classifier，其訓練資料為$G = <G_1, G_2, \ldots, G_T>$加上$X_{T+1}$，在這邊所選用的Base Classifier為Random Forest，參考的特徵為$\log_2 35+1$，一共7個特徵值，且Ensemble Box的大小為3。

DOI:10.6814/NCCU201901233

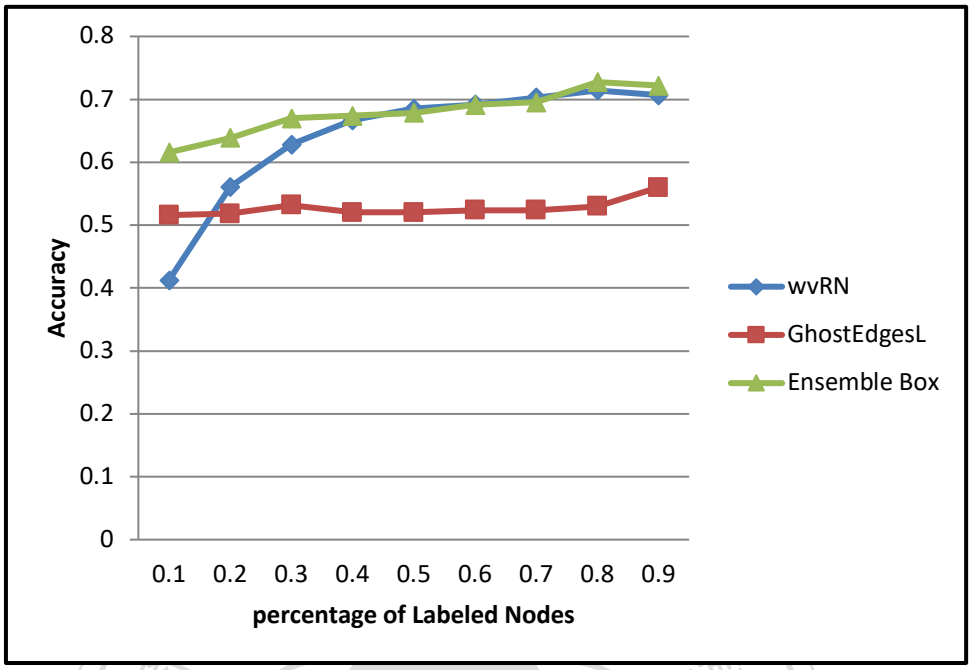

圖 4.1 ConceptClassifier、GhostEdgeL、WVRN、之比較

實驗結果如圖 4.1 所示，在已知類別點占整個社會網絡的 10%、20%時，Ensemble Box 的準確率比起 wvRN 與 GhostEdgeL 都來的好。然而，當已知類別的點所占的比例逐漸提高時，Ensemble box 的準確率與 wvRN 趨近相同。此實驗結果證明了，針對動態網絡中的類別預測，的確是可以由過去的資料中找出類別的規則，進而在下一個時間階段做分類的預測。

[實驗二] 訓練資料多寡對於準確率的影響

實驗二中，我們改變訓練資料的多寡，觀察是否訓練資料的個數與準確率為正相關。原本的實驗中，訓練資料為$G = <G_1, G_2, \ldots, G_T>$，$1 \leq t \leq 12$。在本實驗中，我們調整訓練資料的年份從原本$G = <G_1, G_2, \ldots, G_T>$，$1 \leq t \leq 12$，逐漸減少t的數目；我們調整訓練資料的數量如下：

$G = <G_1, G_2, \ldots, G_{12}>$ 一共12年

DOI:10.6814/NCCU201901233

$G = <G_2, G_{2,}\ldots, G_{12}>$ 一共11年

$G = <G_3, G_{2,}\ldots, G_{12}>$ 一共10年、

•

•

$G = <G_9, G_{2,}\ldots, G_{12}>$ 一共四年

藉此比較當訓練資料變少時，對準確率的影響。

圖4.2可以看出來，由於我們所減少的訓練資料數目是從最前代開始，使得越靠近測試年份G_{T+1}的Concept都有留下來，足以對訓練資料提供正確的分類線索，因此從圖中看出減少前代的Concept，對準確率的影響並不大。

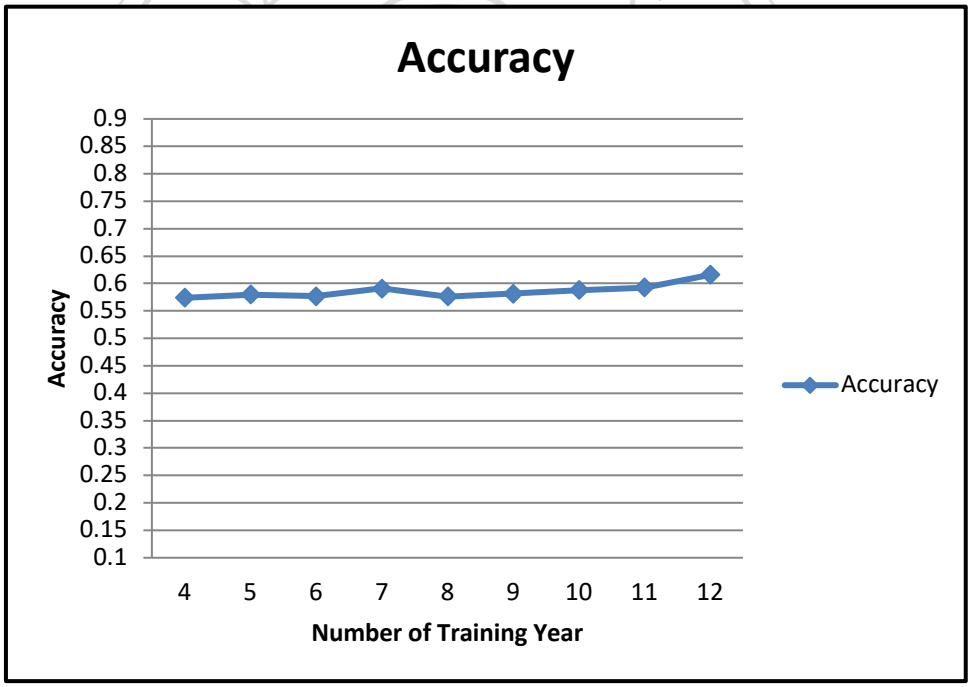

圖 4.2 訓練資料數量與準確率關係圖

[實驗三]

實驗三我們藉由改變 Ensemble Box 中，Classifier 的數量，藉以瞭解 Ensemble Box 的 size 與準確率之間的關係。我們原本的方法中，Ensemble Box 的大小設為 3，

DOI:10.6814/NCCU201901233

本實驗中，調整 Ensemble Box 的 size 為：1,3,5,7，實驗結果如圖 4.3 所示。

可以看得出來，準確率並不會隨著 Ensemble Box 的大小而增加，也就是說，並非考慮越多的過去資料，就會使得分類越準確。Ensemble Box 增大的情況下，可能將不具代表性的分類規則也加入預測的考量中，因此造成準確率些微的下降。Ensemble Box 的大小，與其挑的多，不如挑的巧。

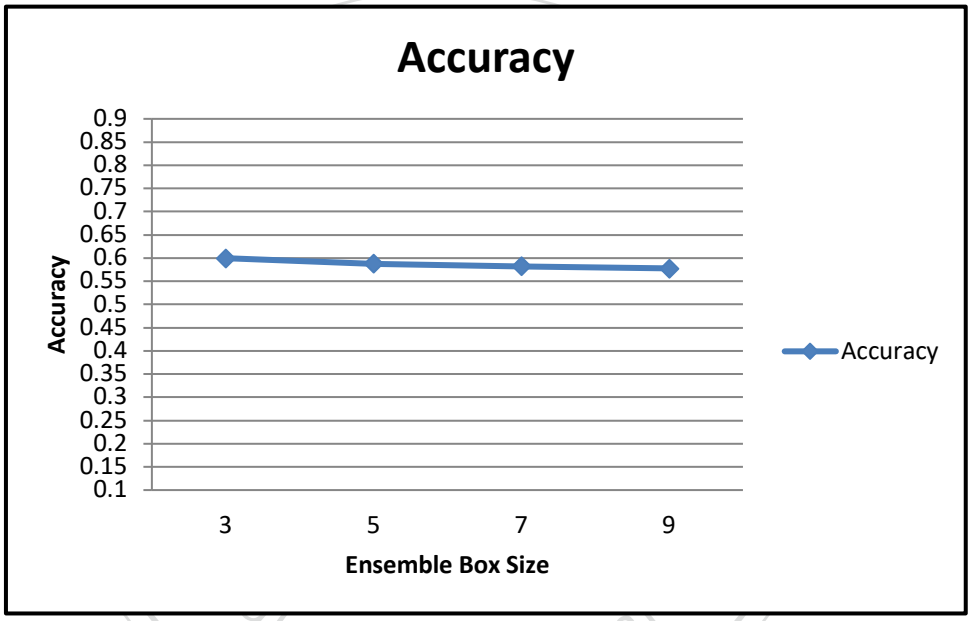

圖 4.3 Ensemble Box 與準確率關係圖

DOI:10.6814/NCCU201901233

第 五 章

結論與未來研究方向

5.1 結論

本研究是針對動態的社會網絡作類別的預測；社會網絡裡所有的點之間，無論在網絡結構分布的全域關聯性，或是實際連結的區域結構關聯性上，都對彼此產生了影響，而這些相互影響便決定了類別。

本研究提出了動態社會網絡中類別預測的問題。我們不只單單利用當下的社會網絡狀態當作分類的線索，更利用了社會網絡隨著時間增長的演進過程，設法找出位於下一個時間之社會網絡中的點及其正確的類別對應脈絡，並且透過實驗，證明此脈絡是會隨著時間改變，並且影響位於下一個階段中節點之類別。

5.2 未來研究方向

本研究提出了動態網絡中類別預測的問題；在偵測概念飄移的部分，現在是使用SEA[18]，而偵測概念飄移的方式還有許多方法，如何發展一種更適合於動態社會網絡類別預測的問題的解法，是一個值得探討的方向。另外，本研究討論的是單一類別的問題，現實中往往很多時候的分類問題都是多重類別的，因此針對多重類別的動態社會網絡的類別預測，也是一個值得深入之處。除此之外，不同的資料庫會造成不同的概念飄移，若可以應用在不同性質的資料庫中，針對不同型態的概念飄移找出合適的解法，同樣為一個需要深思的範疇。

DOI:10.6814/NCCU201901233

參考文獻

[1] L. Breiman. “Random Forests,” *Machine Learning*, Vol. 15, No. 1, pp. 5-12, 2001.

[2] D. Brezeale and D. J. Cook. “Automatic Video Classification: A Survey of the Literature,” *Systems, Man, and Cybernetics*, *Part C: Applications and Reviews*, *IEEE Transactions* on, Vol. 38, pp. 416-430, 2008.

[3] C. Desrosiers and G. Karypis, “Within-Network Classification Using Local Structure Similarity,” *Proc. of the European Conference on Machine Learning (ECML)* and *the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 260-275, 2009.

[4] B. Gallagher, H. Tong, T. Eliassi-Rad, and C. Faloutsos, “Using Ghost Edges for Classification in Sparsely Labeled Networks,” *Proc. of the 14th ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) at International Conference on Knowledge Discovery and Data Mining*, pp. 256-264, 2008.

[5] J. He, J. Carbonell, and Y. Liu, “Graph-Based Semi-Supervised Learning as a Generative Model,” *Proc. of the International Joint Conference on Artificial Intelligence*, pp. 2492-2497, 2007.

[6] J. He, M. J. Li, H. J. Zhang, H. H. Tong, and C. S. Zhang, “Manifold-Ranking based Image Retrieval,” *Proc. of the 12th annual ACM International Conference on Multimedia*, pp. 9-16, 2004.

[7] J. Z. Kolter and M. A. Maloof. “Dynamic Weighted Majority: A New Ensemble Method for Tracking Concept Drift,” *Proc. of the 3rd International IEEE Conference on Data Mining*, pp. 123-130, 2003.

[8] F. Lin and W. W. Cohen, “Semi-Supervised Classification of Network Data Using Very

DOI:10.6814/NCCU201901233

Few Labels," *Proc. of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, pp. 192-199, 2010.

[9] S. A. Macskassy and F. Provost. "A Simple Relational Classifier," *Proc. of the Second Workshop on Multi-Relational Data Mining (MRDM-2003) at International Conference on Knowledge Discovery and Data Mining*, pp. 64-76, 2003.

[10] S. A. Macskassy and F. Provost, "Classification in Networked Data: A Toolkit and a Univariate Case Study," *The Journal of Machine Learning Research*, Vol. 8, pp. 935-983, 2007.

[11] L. McDowell, K. M. Gupta, and D. W. Aha, "Cautious Inference in Collective Classification," *Journal of Machine Learning Research*, Vol. 10, pp. 596-601, 2007.

[12] L. McDowell, K. M. Gupta, and D. W. Aha, "Meta-Prediction for Collective Classification," *Proc. 23th International FLAIRS Conference*, 2010.

[13] J. Y. Pan, H. J. Yang, C. Faloutsos, and P. Duygulu, "Automatic Multimedia Cross-Modal Correlation Discovery," *Proc. of the 10th ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) at International conference on Knowledge discovery and data mining,* pp. 653-658, 2004.

[14] P. Sen, G. M. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective Classification in Network Data," *AI magazine,* vol. 29, No.3, pp. 93-106, 2008.

[15] W. Street and Y. Kim, "A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification," *Proc. of the 7th ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) at International conference on Knowledge discovery and data mining,* pp. 377-382, 2001.

[16] A. Sultan and A.Hegami, "Classical and Incremental Classification in Data Mining Process," *International Journal of Computer Science and Network Security*, Vol. 7, No.

DOI:10.6814/NCCU201901233

12, pp. 179-187, 2007.

[17] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison-Wesley Longman Publishing Co., 2005.

[18] H. Tong and C. Faloutsos, "Center-Piece Subgraphs: Problem Definition and Fast Solutions," *Proc. of the 12th ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) at International Conference on Knowledge Discovery and Data Mining*, pp. 404-413, 2006.

[19] H. Tong, C. Faloutsos, and J. Y. Pan, "Fast Random Walk with Restart and Its Applications," *Proc. of the 6th International IEEE Conference on Data Mining*, pp. 613-622, 2006.

[20] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," *International Journal of Data Warehousing and Mining*, Vol. 3, No. 3, pp. 1-13, 2007.

[21] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen, "Dynamic Integration of Classifiers for Handling Concept Drift," *Information Fusion,* Vol. 9, pp. 56-68, 2008.

[22] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with Local and Global Consistency," *Advances in Neural Information Processing Systems, Vol. 16*, pp. 321-328, 2004.

[23] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," *Proc. of the 20th International Conference on Machine Learning,* pp. 912-919, 2003.

[24] X. Zhu. *Semi-supervised Learning Literature Survey,* University of Wisconsin-Madison Department of Computer Sciences, 2005.

[25] I. Zliobaite, "Learning under Concept Drift: an Overview," *Technical Report*, Vilnius University, 2010.

DOI:10.6814/NCCU201901233