Final Project

R02921044 童宇凡

**R02921074 吳家輝**

# 摘要

　　共合作關係預測在社群網路上是非常著名的問題，至今已經有許多人研究這一領域。本專案參考了許多文章所使用的特徵，並加入一些新的想法來做實驗。

# 介紹

　　共合作關係網路是指研究員們共同合作產出書、期刊、研究報告，彼此之間有合作便是有連結。共合作關係網路成長改變非常地快，而共合作關係預測是指給定時間Ｔ以前的共合作社群網路，藉以預測時間Ｔ＋１的共合作社群網路的連結的改變。

　　而在本研究中，重點在於時間Ｔ以前兩個作者u與v曾經合作，預測時間Ｔ＋１時，這兩個作者會不會合作。

　　共合作關係網路有兩種不同的型態，分別為：同質性網路與異質性網路，同質性網路的意思是網路上每一個節點代表的是相同類型的事物，舉例來說每個節點都代表一位作者。而異質性網路的意思是網路上每一個節點可以代表不同類型的事物，舉例來說節點可以是作者、會議、期刊、報告等，本研究是採用將共合作關係網路視為同質性網路的作法。

Network

Feature Vector

Prediction

　　本研究的每一筆輸入資料代表兩位不同作者在指定年、指定期刊、指定研究報告的一次合作，並不包含其他資料：像是作者的產量與研究報告的關鍵詞。然後產出在指定年兩位作者之間存不存在合作關係。

　　在這研究之中，我們利用監督式學習來學習這些資訊並做出預測，為此我們需要找出一些有用的特徵，並建構監督式學習的框架，利用這些特徵來提高預測的準確度。

# 研究方法

用於研究共同作者預測的主流方法，可分為三種(2013Minh)，列舉如下。

1. 相似度方法(Similarity Based Strategies)

進行相關度預測時，會先將每對欲預測的點進行評估，並計算每一對點的相似度。常使用的相似度指標，包含最短路徑長、共同相鄰點數、Kats叢集指標、Jaccard指標、Adamic/Adar指標等等。另外也有許多較為複雜的計算方法，包括PageRank、AuthorRank、PropFlow等等。多種指標常常交互混用，以達到更好的效果。整體來說，由於架構相對簡單，執行預測時的時間複雜度較低，因此適用於大型的網路。

1. 最大似然估計(Maximum Likelihood Estimation)

對於共同作者的預測，可以透過計算每對點相鄰的機率以求解。在實用上，可以使用最大似然估計的方法求解。進行這種方法時，通常會假設原始圖形具有某種性質，再利用這種性質以及原有的圖形，計算出未相連的點相連的似然度。常見的方法包含Stochastic Block Model及Hierarchical Structure Model等。然而使用這種方法，通常需要對原始的圖形有相當的認知。由於共同作者網路相當的稀疏，所以使用這種方法並不十分適合。

1. 監督式學習(Supervised Learning)

對於預測問題，另一常用的方法即為監督式學習。透過將已有的資訊轉換為訓練資料，可以對未知的資訊進行預測。此種方法的特點在於，需要擷取出每筆資料的特徵(feature)，並對這些特徵進行機器學習，將得到的參數及結果套用至欲預測的資料上。常用的方法包含Naïve Bayes、Decision Tree、Multilayer Perceptron、Support Vector Machine等等。數個先前的研究指出，使用監督式學習的效果，較相似度方法為優秀(2010Lü, 2006Hasan)。

這些主流的方法，均有其可取之處。對於我們的研究，主要採用了Support Vector Machine的方法，並以libSVM做為我們的工具。以下介紹我們的實作。

由於所給定的已知資料為2008至2012年，而欲預測的資料為2013年，因此我們作了如下的分割。所使用的訓練資料，是從2008至2011年的資料中擷取特徵，並將2012年合作與否視為分類的依據。藉由這些訓練的結果，再以2009至2012所擷取出的特徵，進而預測2013年是否合作。如此安排的考量在於，都是使用連續4年的資料作為特徵，並將下一年的資料作為標籤。而為了調整所需要的參數以及研究所需要的特徵，我們先使用了另一套框架來測試。在以下的測試中，訓練資料使用了2008至2010年的資料擷取特徵，並以2011年作為標籤。測試資料則是2009至2011年的資料作為特徵，並預測2012年的資料，比對準確率。

在Support Vector Machine的實作中，必須要對每對點擷取出其特徵(feature)，作為訓練及預測的基準。參考了數篇先前的研究，我們使用了如下的基準。在這些特徵中，多數特徵在引用出處均有嚴謹定義，可以參考引用來源。未有明確定義的，在以下附上說明。

1. Sum of Papers (2006Hasan)
2. Sum of Common Neighbors (2006Hasan)
3. Weighted Sum of Common Neighbors (2011Sá)
4. Sum of Conferences  
   兩位作者所合計參與過的Conference總數。
5. Coauthor Counts

兩位作者直接合作過的次數。

1. Inner Product of Conferences

將兩位作者在各個conference上發表過的paper記錄為兩個向量。將這兩個向量歸一化後，再算出兩個向量的內積。

1. Sum of log(Second Neighbors Count) (2006Hasan)
2. Clustering Index (2006Hasan)
3. Weighted Clustering Index (2011Sá)
4. Salton Index (2010Lü)
5. Jaccard Index (2010Lü)
6. Weighted Jaccard Index (2011Sá)
7. Hub Promoted Index (2010Lü)
8. Hub Depressed Index (2010Lü)
9. Leicht-Holme-Newman Index (2010Lü)
10. Preferential Attachment Index (2010Lü)
11. Weighted Preferential Attachment Index (2011Sá)
12. Adamic-Adar Index (2010Lü)
13. Weighted Adamic-Adar Index (2011Sá)
14. Resource Allocation Index (2010Lü)
15. Weighted Resource Allocation Index (2011Sá)
16. Maximal/Minimal PageRank

PageRank原本為估計網頁重要性的指標。在我們的情景中，若兩位作者合作過，就類比為兩個作者互相有一個指向彼此的連結。而依這種圖可以算出每位作者的PageRank值。分別取出兩個作者中較大及較小的PageRank即成為兩個特徵。

1. Maximal/Minimal Weighted PageRank

類似於PageRank的概念，然而不同點在於，較多的合作次數會增加連結的數量，而多重數量適用於加權PageRank。

另外由於給定測試資料的特性，一些其他的features，如Katz及Shortest Path等等，考慮到可能成效不佳，而不在我們的實作範圍中。

為了評估些特徵的有效性，我們選用了常見的F-Score做為Feature Selection的工具(2006Chen)。F-Score主要用於評估單獨一個特徵對於分類是否有效。F-Score僅需要有正負樣本的統計數值即可計算，為一個簡便的指標。然而其仍有部份缺點。舉例而言，某些資料可能透過兩個特徵的線性組合可被完全的線性分割，那這兩種特徵就相當具有使用的價值，但是這種特徵可能無法透過F-Score所檢驗出。即使如此，F-Score仍被廣泛認定為合用的標準。以下為各個特徵的F-Score值，其中較高的F-Score值對於SVM的幫助一般被認定為較大的。

透過這些F-Score，我們進行了特徵選取(Feature Selection)。透過選出較適合的特徵，期望可以提升整體的效能。對於給定的特徵數*n*，我們透過使用前*n*個具有較高F-Score的特徵供SVM。對於不同的*n*，所獲得的測試結果如下。由實驗的成效，我們選定了前24個特徵，作為後續使用的標準。換言之，我們捨棄了Minimum Weighted PageRank這項指標。

另外，由於已知的資料中，另外包含了年份的資料。我們認為年份可能對合作有重要影響，因此在我們的架構中，我們另外考慮，將每個已知連結乘上一個比重。以參數*y*調整這個比重。對於不同年份的連結，若兩者相差*d*年，則較新的連結對於整張圖的影響，就為舊連結的*yd*倍。由前面的結果，我們以24個特徵作為我們使用的標準，對不同的*y*，得到了如下的準確度。由這個實驗結果，我們選定了*y*=1.8。

除此之外，對於SVM的使用，參數的選擇也相當的重要。礙於計算能力，此處僅討論不同的cost對於SVM的影響。對於不同的參數*c*，得到的準確度如下。依據這些數據，雖然我們合理相信更大的*c*值會有更好的成效，但基於運算時間限制，我們選取*c*=128做為我們的參數。

# 遇到的困難

　　在做實驗的時候其實有遇到一些困難，像是記憶體問題，輸入的資料量很大，作者數目很多，要儲存將每個作者視為一個節點所建成的圖的話，用相鄰矩陣方式會有記憶體不夠的問題，不過幸好是稀疏矩陣，所以改用相鄰串列的方式便可以改善記憶體不夠的問題，雖然說執行速度會因此受到影響。

　　再來我們發現其實有些研究的特徵在這次的研究不適用，因為問題的定義有些不相同，像是有篇有用到最短路徑作為特徵，但是本研究中因為預測的兩點之間都有直接相連，所以最短路徑其實都是定值。

# 結論

　　在本研究中，我們介紹了共合作關係網路，並如何預測共合作關係網路的連結。然後把重點放在要用那些特徵才能有較好的表現，並加入了一些時間對作者之間影響力的衰減。在實驗中其實我們發現其實只用少數特徵便可以有不錯的效果，後面再加上一些特徵對準確度的改進其實不太明顯。

# 參考資料

* David Liben-Nowell and Jon Kleinberg. The Link Prediction Problem for Social Networks, *2004.*
* Le Nhat Minh. Link Prediction in Co-authorship Network, *2013.*
* Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link Prediction using Supervised Learning, *2006*.
* Hially Rodrigues de Sá and Ricardo B. C. Prudêncio. Supervised Link Predictin in Weighted Networks, *2011*.
* Linyuan Lü and Tao Zhou. Link Prediction in Complex Networks: A Survey, *2010*.
* Yi-Wei Chen and Chih-Jen Lin. Combining SVMs with Various Feature Selection Strategies, *2006*.