給定模型下即時分類遺失部份資訊的資料

(Real-time classification with missing data given models)

一、摘要

資訊網路擴展導致資訊量爆增，同時也使得機器學習 (machine learning) 這塊領域的研究更加蓬勃發展，但是直接將機器學習理論應用到現實世界中的時候卻難免會遇到諸多的問題，一項常常被大家提出來討論的議題就是「資料遺失 (data loss)」，不論是因為偵測裝置的異常或是儲存設備的遺漏，一旦資料出現遺失就很難再被復原成原本的樣貌，如此一來就導致很多機器學習演算法在實務應用上的困難。

資訊發生遺失的機會多到令人難以置信；舉例來說，一個放置在網路上的演講錄音檔案就有以下幾個可能導致資訊遺失：麥克風的接收端因為收訊不佳導致接收到過多雜音、錄音設備因為機器老舊而在轉錄音訊時有所遺漏，除此之外，上傳該檔案到網路上時還可能因為音訊壓縮而讓資料出現失真的情形；只要資訊出現遺失，模型的建立與預測便會出現誤差，因此如何兼顧表現以及效率，並在某種程度上填補、預測遺漏的資訊，同時還要做好分類，這就是我們急需解決的問題。

此次的研究將會專注在如何「即時」做遺失資訊資料的分類，更明確地來說，就是假設待補的資料跟資料之間都不存在著相依關係，而且對於每筆資料的詢問都必須立即回答其所屬的分類。

二、研究動機與研究問題

(一) 研究動機：

在現實世界中，有很多事情的結果都需要利用過去獲得的經驗、紀錄加以預測；透過機器學習用一套良好的方法將過去的資料歸納出結論，讓我們可以有效地預測未來發生類似事件的結果。但是很不幸的，搜集獲得的事件資訊可能是不完整的，這時就沒有辦法直接用曾經發生的類似事件來做比擬。

資訊遺漏的情形可能發生在兩個地方，第一個是在過去搜集而來的資料上，那些資料通常都已經被附加標記，我們將會利用那些資料來建構可供預測的模型，第二個則是出現在正在觀察的資料上，對此我們將會利用已經建立好的模型來預測該筆資料的標記。

前者已經有諸多前輩研究並提出有效的解決方法，卻鮮很少有人研究後者，但是在實務上兩者其實是同等重要的；如果只有一個良好的模型，卻沒有良好的分類演算法來處理未標記資料的資訊遺失問題，最後得到的結果也會同樣無法令人滿意。

資訊遺失的情形非常常見，發生的原因也很多種，因此如何找到一個有效率、同時又擁有高精準度的一般性 (general) 演算法來處理這類型的問題可以說是當務之急。

(二) 研究問題：

本次研究的目的是要找出一個兼顧高效率與高精準度的演算法來處理資料的資訊遺失問題，在此我將會只專注在預測端的資訊遺失，即假設模型已經給定，而忽略建構模型時可能會遇到的問題。

- Input: 又被稱作訓練資料 (training data)，給定過去搜集而來的資料 $D = \{d_1, d_2, ...\}$ ，每筆資料 d_i 都有一個特徵向量 (feature vector) v_i，和一個標記 s_i，在簡化題目的狀況下，假設每筆資料的向量都為固定長度 L，而且所有的元素值都是已知確定的，不會有遺失資訊的狀況發生。
- Query: 又被稱作測試資料 (testing data)，每次系統都會接收到一筆測試用的詢問資料，該筆詢問 q_i 會有一個特徵向量 u_i，該向量亦為固定長度 L，但是元素值可能會有少部分是不確定的。
- Output: 對於每筆詢問都要輸出一個最有可能的標記 t_i，作為我們系統在訓練之後的預測值。

一些之後可能的延伸研究方向包括：不定長度的特徵向量、特徵元素跟元素之間出現相依關係等。

三、文獻回顧與探討

關於建立模型時的資訊遺失已經有多篇論文提及不同方法，這裡將只會簡單列出重點。

1. **部分刪除法 (Partial deletion)**：
   大致上包含交互刪除 (pairwise deletion) 以及條列刪除 (listwise deletion) 兩種；在實作上，前者會枚舉所有特徵值出現的情形，模型的整體複雜度將會隨著特徵向量的長度做指數遞增，明確來說，模型的複雜度為 $O(2^L)$；後者則同字面上的意思，將所有遺失部份元素值的資料刪除，此作法的缺點是會犧牲掉許多有意義的資訊。

2. **部分插補法 (Partial imputation)**：
   顧名思義就是將遺漏的元素填入數值，遺漏的元素值可以被填入該特徵向量的中位數、均值、眾數等，甚至也可以選擇使用內插法 (interpolation) 來填值，缺點是此作法的表現通常較不理想。

3. **貝氏推估法 (Bayesian inference)**：
   Bayesian Inference in Statistical Analysis(George E. P. Box & George C. Tiao, 1973) 書中提出的方法 [1]，如果實際用枚舉的方式替模型找一組最佳化的參數解，在經過一連串的計算後一定能獲得針對訓練資料的最佳模型，但是這個方法有兩大缺點，第一個缺點是我們獲得的模型可能會過度貼近訓練資料的特性，而產生過度合適 (overfitting) 的情形，這將會使我們在測試資料上的預測失準，第二個更大的問題是此種演算法會花費極大量的時間，更明確地來說，模型的複雜度為 $O(L^{|D|})$。(這裡 $|D|$ 代表訓練資料的數量)

4. **EM 演算法 (Expectation-maximization algorithm, EM)**：
   Maximum likelihood from incomplete data via the EM algorithm (Arthur Dempster & Nan Laird & Donald Rubin, 1977) 論文中提出的作法 [2]，透過交替計算「固定係數下的期望值」(E-step)、「固定期望值的最佳係數」(M-step) 可以使結果趨向局部最大值 (local maximum)，透過吉布斯不等式 (Gibbs inequality) 可以證明每一次的交替運算都會使得結果呈現非嚴格遞增 (non-strict increasing)。

5. **多重插補法 (Multiple imputation, MI)**：
   Multiple Imputation for Nonresponse in Surveys(Joseph L. Schafer & Maren K. Olsen, 1987) 書中提出的想法 [3]，主要有兩種插補方法：

第一種稱作傾向分數方法 (propensity score method)，該方法利用邏輯迴歸分析模型 (logistic regression model) 來計算各個觀測值 (observed value)，並將觀測值當作遺失資訊的機率，最後再使用 (Bayesian bootstraping) 來做資料插補。

第二種稱作蒙地卡羅 - 馬可夫鏈法 (Monte Carlo Marcov Chain, MCMC)，是一套類似 EM 的演算法，透過交替使用「利用已知期望值算出的條件機率來插補遺漏的元素值」(I-step) 與「計算插補後元素的期望值」(P-step) 可以使得結果收斂在某個局部最大值。

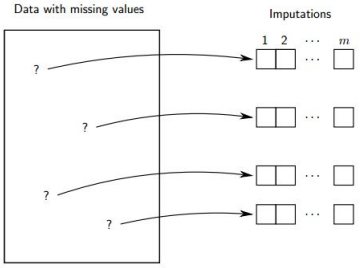

Figure 1: 利用多重插補法來處理遺失資訊的多元資料矩陣

6. **連續插補法 (Sequential imputation, SI)**：
   Sequential imputations and Bayesian missing data problems(Augustine Konga & Jun S. Liub & Wing Hung Wonga, 1994) 論文中使用的方法 [5]，該方法強調在順序地放入訓練資料的同時會替每筆遺失的資料調整出一個適當的權重。它的優勢在於不需要做交替優化，在實務上所需花的時間將遠小於前面的作法，缺點為該作法強調資料間的順序關係。

有關測試時出現資料的資訊遺失研究就相對少很多，2007 年發表論文的 Saar-Tsechansky, Maytal & Provost, Foster 也有特別提到這一點；已知的作法大致可以被歸類為下列幾項：

1. **部分刪除法 (Partial deletion)**：
   方法類似上述討論，可以選擇拋棄特徵值也可以直接選擇放棄整筆資料。

2. **部分插值法 (Value Imputation)**：
   方法類似上述討論，效果有限。

3. **分類樹 (Classification tree)**：

Handling Missing Values when Applying Classification Models (Saar-Tsechansky, Maytal & Provost, Foster, 2007) 論文提出的作法 [6]，可以依實作方式分成兩類：

第一類是數值插值 (predictive value imputation)，利用其他特徵值的數值替缺少的元素填值，一個叫做替代拆分 (surrogate splits) 的作法還會依填補的值來決定存放在子樹上的位置。

第二類是分布插值 (distribution-based imputation)，為演算法 C4.5[9] 的基礎策略，一旦出現資訊遺失的資料它就會嘗試分裂出多個可能的完整資料和權重，只有當抵達樹上的葉子 (leaf) 時才去計算它的機率，這個作法與前者的最大不同在於：前者只考慮如何求出最有可能的填補值，但是後者會考慮到特徵向量中各元素值的分布情形。

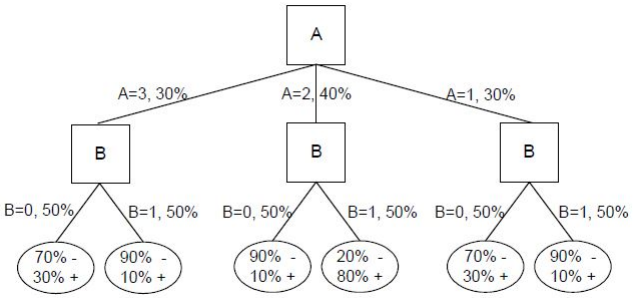

Figure 2: 分類樹的使用例子，A 和 B 都是一個隨機變數

4. **限縮特徵模型 (Reduced-feature model)**：

此方法同樣出自於 Handling Missing Values when Applying Classification Models [5]，針對不同的資訊遺失狀況限縮該分類樹，缺點是依據遺失的情形不同會保留多種模型。

四、研究方法及步驟

此次研究目的為解決如何處理測試資料的資訊遺失情形，所以會把重心放在設計分類演算法上，但是同樣的演算法配合不同的模型可能會得到不同的結果，所以我們在實驗中還是會測試不同的模型組合；實驗依照流程可以大致分為：資料搜集、資料切割、資料處理、模型設計、模型訓練、模型測試等步驟。

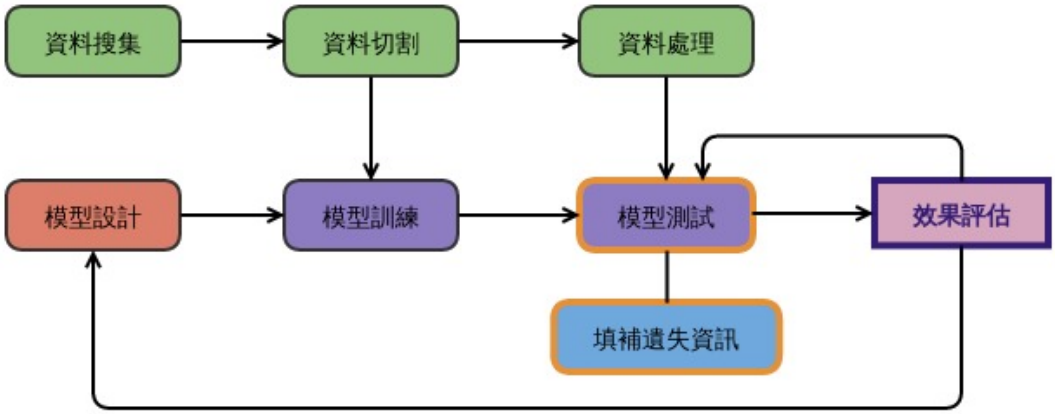

Figure 3: 實驗架構圖

1. **資料搜集 (Data collection)**：
   為了要產生出一個比較一般性的演算法，在資料集 (data set) 的選用上的條件限制比較寬鬆，決定好資料集後將會由步驟 2. 及步驟 3. 來產生出可以使用的訓練資料與測試資料。

   目前已經從國際氣候資料中心 (National Climate Data Center, NOAA) 中找到一個紀錄全世界氣候的網站：全球歷史氣候網路 (Global Historical Climate Network, GHCN)，其中包含著自 1763 年至 2014 年世界各地每日的天氣資料，從 181 個國家、74 州，一共 90905 個觀測站蒐集而來的資料，對於每個區域每天至少包含：降雨量 (精準至 0.1 mm)、下雪量 (精準至 1 mm)、下雪深度 (精準至 1 mm)、當日最高溫 (精準至 0.1 °C)、當日最低溫 (精準至 0.1 °C) 等五項觀測值，其餘可能的資訊還包含平均雲量、平均風速、最大風速、最大陣風等，甚至在特殊氣候下還會額外紀錄連續的天氣變化情形；整體而言可以說是記載得非常詳細、完整。

   有關此資料庫的其他細節格式與資料請查閱參考文獻。[10]

2. **資料切割 (Data partition)**：
   將搜集來的資料切割成訓練部分 (trainning set) 以及測試部分 (testing set)，要特別注意切割的選擇上是否足夠隨機 (random)，以維護訓練資料與測試資料的數據分布的一致性；之後再讓模型去學習訓練資料中可能隱含 (latent) 的模式 (pattern)。

3. **資料處理 (Data preprocessing)**：
處理測試資料的時候有三種可能的做法，分別代表資料背後的隱含意義，不同種的做法將會影響到最後測試出來的結果：

- **完全隨機遺失 (missing completely at random, MCAR)**：
代表遺失的特徵變數 (feature variable) x_i 與遺失的值 (value) v_i 是完全無關的。
- **隨機遺失 (missing at random, MAR)**：
代表遺失的特徵變數之間可能存在著某種相依關係，但是遺失的方式是隨機的。
- **非隨機遺失 (missing not at random, MNAR)**：
代表遺失的方式是非隨機的。

在現實世界之中觀察到的資料可能是三者中的任何一種，但在手動構造的實驗中，我們無法區別第一、二種的不同，也無法構造出第三種測試資料，因此在之後的實驗裡都統一使用「隨機遺失 (MAR)」這個方法。

4. **模型設計 (Model design)**：
決定訓練與測試時使用的模型，除了將前輩的分類樹 [6] 實作出來當作基線算法 (baseline algorithm) 外，大致上有兩個想要嘗試的方向：

第一種是改進傳統的圖形模型 (graphical model)，這類型的模型通常都能夠得到不錯的結果，缺點是也需要付出很高的計算代價。具體上的作法可以再分為兩類型：

- **貝氏網路 (Bayesian network)**：
用來解決有向 (directed) 的圖形模型的機器學習問題，依據不同的變數相依關係可以將原本獲得的式子轉變成比較容易計算的形式。
舉例來說：假設有 n 個事件，分別為 $X_1, ..., X_n$，該集合被稱作 S，每一事件都為一隨機變數 (random variable)，且都只跟前一項有關，則我們可以將原本一般化的式子 $P[X_1, ..., X_3] = \prod_{i=1}^{n} P[X_i|S \backslash X_i]$ 簡化成 $P[X_1, ..., X_3] = P[X_1] \times P[X_2|X_1] \times ... \times P[X_n|X_{n-1}]$。
因此，如果模型可以依照資料性質調整至一個較好的貝氏網路架構，就會提昇不少運算上的效率。
- **馬可夫隨機場 (Markov random field)**：
類似貝氏網路，主要是用來解決無向 (undirected) 的圖形問題。

第二種是嘗試減少在測試時插補所需的計算量，以改進 EM 演算法為例，使得它能夠在少量的迭代 (iteration) 下收斂至一組定值，但是這麼一來勢必會犧牲掉一些精準度，如何在速度與精準度兩者中做出適當的取捨也是這次的實驗目的之一。

5. **模型訓練 (Model training)**：
   在選定好要使用的模型之後，就用切分出來的訓練資料來決定模型的參數 (parameter)，這部分會依照不同模型的建構方式而有不同的訓練過程，最終目的是希望找到一組針對訓練資料的最佳解參數。

6. **模型測試 (Model testing)**：
   利用建構出來的模型來預測整體特徵向量值的分布，同時將測試資料分類；假設訓練與測試使用的資料分布類似，應該可以得到不錯的結果。

   精準度高低的評估根據資料性質不同而有不同的選擇，一些常用的損失函數 (loss function) 包括 0-1 損失函數 (0-1 loss function)、平方損失函數 (quadratic loss function) 等。

五、預期結果

1. **設計高效率 (efficient) 的方式替測試資料標記**：
   這是此次研究的主要目標，提出的作法必須符合即時標記這個要求，即時的定義在這裡有些模糊，但是至少要能跟得上測試資料的產生速度。

2. **設計高準確度 (high accuracy) 的演算法替測試資料標記**：
   同樣也是此次研究的重要目標，在兼顧速度的條件下盡可能的提高演算法的準確度，希望最後能夠將此作法使用在一般化的實務問題上。

3. **實際填補遺失的資訊**：
   除了替測試資料附加標記外，也要設法填補遺失的資訊，如此一來可能可以進一步觀察遺失資訊的原因並嘗試避免。

4. **將一般化的問題應用在其他領域**：
   在得到良好的一般化演算法後，就要嘗試是否能夠應用在其他地方，幾個急需解決的現實問題包括在線的語音辨識、影像辨識等，這些問題常常都會因為資訊的遺漏而大幅降低了辨識的精準度以及效率。

六、參考文獻

[1]: George E. P. Box & George C. Tiao(1973) Bayesian Inference in Statistical Analysis, Wiley Classics Library Edition Pulished 1992

[2]: Arthur Dempster & Nan Laird & Donald Rubin(1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1 (1977), pp. 1-38

[3]: Donald B. Rubin(1987) Multiple Imputation for Nonresponse in Surveys, Department of Statistics Harvard University

[4]: Joseph L. Schafer & Maren K. Olsen(1998) Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective, Multivariate Behavioral Research, 33:4, 545-571, DOI: 10.1207/s15327906mbr3304_5

[5]: Augustine Kong & Jun S. Liu & Wing Hung Wong(1994). Sequential Imputation for missing values, Computational Biology and Chemistry 31 (2007) 320 - 327

[6]: Saar-Tsechansky, Maytal & Provost, Foster(2007). Handling Missing Values when Applying Classification Models, Journal of Machine Learning Research. 7/1/2007, Vol. 8 Issue 7, p1625-1657. 33p. 1 Diagram, 8 Charts, 13 Graphs

[7]: Steffen L. Lauritzen(1995) The EM algorithm for graphical association models with missing data, Computational Statistics and Data Analysis, Volume 19, Issue 2, February 1995, Pages 191 - 201

[8]: Martin A. Tanner & Wing Hung Wong(1987) The Calculation of Posterior Distributions by Data Augmentation, Journal of the American Statistical Association, Computational Statistics and Data Analysis, Volume 82, Issue 398, 1987

[9]: Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[10]: Daily weather throughout the world provided by GHCN(Global Historical Climate Network). Database available checked on February 11, 2014, from *ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/*

[11]: David C. Howell. Treatment of Missing Data. Retrieved on February 11, 2014, from *http://www.uvm.edu/dhowell/StatPages/More_Stuff/Missing_Data/Missing.html* & *http://www.uvm.edu/dhowell/StatPages/More_Stuff/Missing_Data/Missing-Part-Two.html*

七、需要指導教授指導的內容

我第一次接觸到機器學習這塊領域就是修習林守德教授開授的暑假課程，迄今也差不多才半年的時間，因此仍有許多艱澀難懂的名詞、演算法尚不熟悉，很慶幸能夠得到教授的親自指點。

接下來的一學期（三年級下學期），林守德教授還會率領台灣大學的隊伍參加 ACM SIGKDD 會議舉辦的資料探勘比賽 KDD Cup(台灣大學已經連續多年得到冠軍，我也很榮幸今年可以成為其中的一員)，比賽中將很有可能會遇到與此研究相關的議題以及應用，屆時也會需要教授從旁協助。

林守德教授不僅在機器學習領域學有專精，過去也已經多次擔任國科會大專生計畫的指導教授，相信在他的指導下我能夠從中學到很多寶貴的研究經驗，並將此次的研究計畫完成。