複雜資料分析的統計降維法

國立交通大學統計所　盧鴻興

前　言

由于高通量(high throughput)科技的發展，產生的資料型態愈來愈複雜，資料分析的任務具備更多的挑戰。如何從這些資料中找出眞實的資訊，並有效的進行分析是一個重要的議題。通常這些資料的變量維度非常高，像是大量的複雜文件、高解析度影像，又或者是基因類型的資料變量等等。所以研究者經常將資料做降維(dimension reduction)處理，再進行分析，有效地萃取有用的資訊。本文介紹常用的統計降維法，研究者能運用這些方法進行複雜資料之分析。

複雜資料的變量維度高，分析之進行必須應用自動化機器(例如電腦…等等)，執行機器學習(machine leaning)。透過大數據(big data)的資料探勘(data mining)，可以發掘知識(knowledge discovery)，並產生人工智慧(artificial intelligence)。以下介紹統計降維法如何與兩個主要類型的機器學習方法結合，進行數據科學(data science)的有效分析及其應用。

非監督式學習(Unsupervised learning)的統計降維法

非監督式學習是運用所有特徵變數的數值變化來學習資料所隱含的分佈結構，將資料分群(clustering)。常用的非監督式學習之統計降維法包括主成分分析(principal component analysis, PCA)以及相關的方法。主成份分析的基本想法是將資料正交投影到具有最大投影量變化的低維度線性空間上，做法是將各個變數的數值乘上不同的權數，也就是投影到一個由權數向量所組成的方向上，使得在這個方向上的投影量之變化最大化。如此就可以將高維度的資料投影到一維上，構成第一主成份。接著在第一主成份的垂直方向，尋找第二主成份的投影方向，使得變數在第二主成份的投影方向的數字變化最大化。持續這樣的步驟，就可以把第三主成份以及後面的主成份之投影方向找出來。這樣的主成份分析方法是可以對資料的共變量矩陣做特徵值分解(eigenvalue decomposition)來找出投影的方向。

另一類常用的非監督式學習之統計降維法是多維標度法(multidimensional scaling, MDS)以及相關的方法。多維標度法基本想法是將資料投影到保有資料相對距離的低維度空間上，先以任意兩個資料點間的距離來構成整體資料的距離矩陣，接著對這一個距離矩陣做雙中央化(double centering)以轉換成內積矩陣，然後做奇異值分解(singular value decomposition)，找出對應的投影方向。這樣的統計降維法經常可以用來製造數據地圖，描寫數據之間的關聯性。當資料的分佈屬於歐式空間時，可以證明主成份分析找出來的投影方向和多維標度法找出來的投影方向是一樣的。

可是這樣的特徵值分解或者奇異值分解的計算複雜度是 $O(n^3)$，n 代表數據量的大小，因此很難使用應用到大型的資料所構成的矩陣上面。所以我們運用類似分治演算法(divide-and-conquer)的想法，提出了分割合併多維標度法(split-and-combine multidimensional scaling, SCMDS)的方法[1]。這樣的做法可以將一個大地圖的製造分解成數個小地圖的製造，再透過小地圖之間重疊的部分，將所有的小地圖組成一個完整的大地圖。因爲每個小地圖所運用的奇異值分解的複雜度比較小，重疊地圖的計算運用到小資料矩陣的 QR 分解(QR decomposition)，複雜度也不大，所以完成整個大地圖的複雜度就可以降成 $O(p^2n)$，p 代表降成的維度。所以這個分割合併多維標度法就可以有效的處理大型的矩陣，分析複雜的資料。這樣的分割合併法也能與主成分分析和奇異值分解結合[2]，分析現代的大型數據所構成的矩陣。

監督式學習(Supervised learning)的統計降維法

監督式學習是運用所有特徵變數的數值變化及其關聯的標記資訊(label information)來學習。如果要學習的標記資訊是離散的類別，監督式學習會將資料分類(classification)，常見的監督式學習之統計降維法包括線性識別分析(linear discriminant analysis, LDA)以及相關的方法。這是由 R. A. Fisher 提出，要找一個投影方向，使得投影後的組間的差異最大，投影後的組內的差異最小。Fisher 提出的準則(Fisher criterion)，定義爲投影後的組間差距除以投影後的組內差距。這個準則是對應於廣義雷利商數(generalized Rayleigh quotient)，商數最大化的解法可以解一個對應的廣義特徵值問題(generalized eigenvalue problem)來獲得。

如果要學習的標記資訊不是離散的類別而是連續的數值，常見的監督式學習之統計降維法包括迴歸分析(regression analysis)以及相關的方法。李克昭創新地提出分片逆迴歸法(sliced inverse regression, SIR) [3]，可以進行有效的降維(effective dimension reduction)。這個方法將連續的變數切成不同的分片，進行逆迴歸，推估有效降維空間(effective dimension reduction space)。當學習的標記資訊是離散的類別時，分片逆迴歸法也可以根據離散的類別來分片，進行逆迴歸，找出來的有效降維空間和線性識別分析找出來的有效降維空間是一樣的。因此，分片逆迴歸法可以學習離散和連續的標記資訊，執行監督式學習之統計降維。

這樣的想法可以進一步推廣成更多的方法和各種的應用。例如，我們發展動態分片逆迴歸法(dynamic sliced inverse regression, DSIR) [4]，可以分析動態的資料與影像。我們也發展遞迴分片逆迴歸法(iterative sliced inverse regression, ISIR) [5]，可以遞迴分析資料與影像，進行影像分割(image segmentation)與圖形識別(pattern recognition)。我們已進一步應用於神經科學的研究，瞭解神經元的激發模式，與神經元控制的行爲之關連[6]。如果有其他的關聯性和輔助型資訊也可以進一步結合，發展充分維度降維法(sufficient dimension reduction, SDR)及相關的推廣方法[7]，詳細的文獻回顧可以參考我們的著作[8]。

結　語

總結而言，不管是非監督式或監督式的學習，都可以與合適的統計降維方法結合。如此一來，就可以降低模型與計算的複雜度並且從中提取出重要的特徵，較能完整分析複雜資料，並能充分運用大數據與機器學習結合，創造資料的價值。於是，我們就可以建構特徵的相關地圖或者是學習到關聯模型。進一步我們可以從資料中學習辨別模式，挖掘關聯，做出有效的預測，發現新的科學特性及知識。

參考文獻

[1] J. Tzeng, H.H.S. Lu and W.H. Li, BMC Bioinformatics, 9:179 (2008).

[2] J. Tzeng, Journal of Applied Mathematics, Article ID 683053 (2013)

[3] K.C. Li, Journal of the American Statistical Association, 86, 316-327 (1991)

[4] H.M. Wu and H.H.S. Lu, Statistica Sinica, 14, 413-430 (2004).

[5] H.M. Wu and H.H.S. Lu, Pattern Recognition, 40, 12, 3492-3502 (2007).

[6] S.H. Yang, Y.Y. Chen, S.H. Lin, L.D. Liao, H.H.S. Lu, C.F. Wang, P.C. Chen, Y.C. Lo, T.D. Phan, H.Y. Chao, H.C. Lin, H.Y. Lai and W.C. Huang, Frontiers in Neuroscience (2016).

[7] H. Hung, C.Y. Liu and H.H.S. Lu, Biostatistics, 17, 3, 405-421 (2016).

[8] H. Hung and H.H.S. Lu, Wiley Interdisciplinary Reviews: Computational Statistics, 9:e1401 (2017).