



政府統計與機器學習的距離 — 從國際應用實例談起

隨資訊技術進步與硬體設備處理效能快速提升，加上豐沛的大數據，使此波人工智慧蓬勃發展，先進國家也積極研究如何運用人工智慧，強化政府統計效能。本文介紹近年國際應用機器學習於政府統計之實例，期藉他山之石，作為我國政府借鏡之參考。

李文德（行政院主計總處綜合統計處專門委員）

壹、前言

在數位時代，經由資訊應用產生大數據，再從這些數據萃取出有用資訊以供決策運用作業模式，已成顯學。尤其隨著資訊技術及設備處理效能提升，及源源不斷的大數據，提供人工智慧（Artificial Intelligence, AI）極佳的發展空間。所謂 AI 係使電腦模仿人類如感官的聽覺及視覺辨識、大腦的學習及決策等能力，以提高處理效能。近年 AI 之

發展又以機器學習（Machine Learning, ML）最為蓬勃，應用實例遽增。

先進國家為肆應外界對大數據運用的期待，亦積極評估以 ML 萃取大數據資訊，並套用於政府統計編製流程中。因此本文將簡單說明 ML 之概念，以及近年國際上應用 ML 於政府統計之案例。

貳、何謂機器學習 (ML)

近年 ML 技術之應用已相當常見，如手機人臉辨識，亞

馬遜無人商店及特斯拉自動駕駛等。ML 為 AI 之一環，強調使電腦藉由資料分析，從模型中推導學習，以改善效能，最終解決問題。常見之 ML 可分為以下 3 個類型：

一、監督式學習 (Supervised Learning)

提供標記資料，即標註特徵值 X 及其目標值 Y，使電腦從中學習，藉由模型（如 $Y=F(X)$ ）推導，進而對未知資料進行預測。惟為使預測確度

提高，常需大量標記資料 X 及 Y，且多為人工標記。此類型常用於分類或迴歸，應用於分類者（目標值 Y 是類別變數），如分辨貓或狗照片等，而應用於迴歸者（目標值 Y 是連續數值變數），如預測股市走勢等。

二、非監督式學習

(Unsupervised Learning)

亦是使電腦從模型（如 $Y=F(X)$ ）中學習，最大差異為不標註目標值 Y（即不提供正確答案），使電腦自行從給定特徵資料 X，進行分類或分群，如提供貓及狗多張身體部位特徵的照片，但不告知照片是貓或狗。此類型適用無較多人力標記 Y，或無背景知識判斷 Y 之集群問題等。

三、強化學習 (Reinforcement Learning)

為使電腦與外在環境互動，透過獎勵或懲罰方式而修正下一次動作，以尋找最大獎勵為策略，進而完成目標，屬互動式學習。如運用在自動駕駛之自助停車，目標為停在正

確停車格，透過與環境（感測器資訊）互動，當每次動作（煞車、踩油門或轉方向盤等）正確時獲得加分，錯誤則減分，為求最高分數而不斷調整動作，最終停在正確車格；及如應用於 AlphaGo，使所落每子均依棋局環境，以追求最高勝率而下等。

另近期熱門之深度學習 (Deep Learning, DL)，亦是一種 ML，相較 ML（須標註特徵），深度學習僅從資料 X，經由 3 層（含）以上人工神經網路（模仿生物神經網路的結構和功能的計算模型），透過模型學習並歸納推導，自行提取足以代表資料 X 特性之重要特徵，故亦稱特徵學習。如訓練電腦辨識貓或狗照片，僅須提供貓或狗大量照片，毋須標註特徵，照片內容是貓或狗，

則不一定要告知，電腦即可透過大量資料，自行提取特徵，將照片分類。由於深度學習的目標值 Y 不一定要標註，故其模型可以是監督式或非監督式，另可結合強化學習，即是深度強化學習，如應用於自動駕駛，透過深度學習辨識路上行人或建物之特徵，再透過強化學習，以行駛至某個定點為目標，經由與外界環境互動，不斷地調整下一個動作（煞車、踩油門或轉方向盤等），以安全行駛為策略，最後抵達定點。AI、ML 及 DL 之關係如圖 1 所示。

參、政府統計之國際應用實況

一、國際應用情形

部分先進國家統計局近期

圖 1 人工智慧、機器學習與深度學習的隸屬關係



資料來源：作者參考「機器學習探究」自行繪製。

論述》統計 · 調查



研究在聯合國「政府統計基本準則」(Fundamental Principles of Official Statistics) 下，是否可將 ML 技術導入政府統計產製流程之中。德國於 2018 年調查 37 個國家 (27 個歐盟成員國、4 個歐洲自由貿易協會 (EFTA) 成員國、6 個非歐國家) 及 2 個國際組織 (Eurostat 及 OECD) 有關應用 ML 於政府統計編製之情形，計 27 個國家或組織 (含 OECD) 回復已成立相關應用的專案，共計 136 個。依序以加拿大 (36 個)、美國 (20 個)、荷蘭 (16 個) 及紐西蘭 (9 個) 專案較多，合占近 6 成。依專案成熟階段來看，已屬產品者占 15.4%，仍在開發者占 20.6%，尚在實驗及點子階段則占 64.0%；就應用領域分，以跨統計領域 (Cross-Statistical)、勞動力市場 (Labour Market)、企業統計 (Business Statistics) 及家戶統計 (Household Statistics) 較多，合占逾 5 成，另物價統計占 3.7%；依專案應用類型分 (此問項複選)，應用於分類

(Classification) 占 48.4%、及插補 (Imputation) 占 13.7% 較多；另以使用演算法分 (此問項複選)，以隨機森林 (Random Forests)、人工神經網路 (Neural Network)、支援向量機 (Support Vector Machine) 及決策樹 (Decision Tree) 較多。

另聯合國 UNECE 與 Eurostat 等國際組織製定通用統計業務處理流程模型 (Generic Statistical Business Process Model, 簡稱 GSBPM)，提供一個標準框架及統一術語，協助各國

政府因應現今環境變化以編製政府統計，共分 8 個階段，每個階段均有數個子流程。前揭 136 個專案僅涉及 GSBPM 之 10 個子流程 (圖 2)，顯見應用尚未全面。若從應用類型來看，屬分類及插補者較多，仍較偏向處理 (Process) 階段，可見 ML 對資料處理之流程改善方面，發展較速；另從產品階段觀察，多數專案尚處於萌芽階段。

二、聯合國成立專案實作

依前揭應用情形，部分

圖 2 2018 年調查國際政府統計應用 ML 之專案涉及 GSBPM 流程情形

Specify needs	Design	Build	Collect	Process	Analysis	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collect	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination Products	
1.5 Check data Availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

說明：紅框表示該子流程應用 ML。
資料來源：Machine Learning in Official Statistics, Martin Beck, et al.。

國家統計局雖正在研究如何在官方統計使用 ML，然官方統計必須維持嚴格的統計品質標準，各國實務應用因而面臨許多問題，這些問題可以透過官方統計界的廣泛討論，取得共識而獲得解決。爰此，聯合國 UNECE 於 2019 年成立 ML 專案，致力於驗證 ML 的附加價值（如增加相關性、提高品質及減少成本等）、增進國家統計局運用能力，及確認將其納入組織的常見問題，並討論在標準流程下，評估應用 ML 之品質框架，預計於 2020 年底公布結果。為利專案研究，聯合國於 GSBPM 流程中挑選處理階段之 5.2 分類及編碼（Classify and Code）與 5.4 編輯及插補（Edit and Impute）子流程，以及影像資料處理部分，分為三個主題，邀請會員國，進行主題之實例研究：

（一）編碼與分類主題

此主題係研究將統計調查之文字項目，依填復內容進行歸類及自動編碼。包括加拿大與冰島研究行、職業別自動編碼、墨西哥與塞

爾維亞研究經濟活動自動編碼，以及美國研究職業別、職業傷害及疾病等自動編碼。

（二）編輯與插補主題

此主題係研究將公務登記或統計調查資料之數據，若判斷為不正確或屬遺失者，則分別更正或增補。包括德國及義大利研究公務登記的教育程度（Attained Level of Education）資料修補、澳大利亞研究人口普查的居住使用（Dwelling Occupancy）資料修補，以及英國研究 3 個家戶面調查資料修補等。

（三）影像主題

此主題係研究運用非由統計目的所蒐集衛星影像圖或航空照片，再從圖像萃取大數據之資訊，以補充公務登記資料之不足，或者取代（或部分取代）調查資料。包括墨西哥研究用衛星影像圖推估非普查年人口數、荷蘭研究用航空照片圖與衛星影像圖推估家戶面統計資料（包括家戶數、居住人口數

及居住場所數等），以及加拿大研究用衛星影像圖預測農作物產量等。

三、實例分享－美國勞工統計局 (BLS) 之自動編碼

（一）問題

BLS 按年辦理職業傷害及疾病調查（Survey of Occupational Injuries and Illnesses），需對職業別（Standard Occupational Classification, SOC）、傷害或疾病性質（nature）、身體受影響的部位（part）、事件（event）、致員工傷病的主要來源（source）及次要來源（secondary source）之 6 個調查項目進行編碼，因每年均耗用大量人力，且即使進用專業人士編碼，編碼相同程度僅達七成。

（二）ML 應用方法及效益

1. 資料集：將 2011~2015 年所調查且可用之 130 萬筆資料，分為訓練、驗證及測試資料集，前 2 項資料集供模型訓練及驗證之用，最後一項資料集用於

論述》統計 · 調查

測試模型之有效性。

2. 模型處理方法：BLS 係採監督式深度學習模型，進行分類及編碼。在提供標準答案 Y 下，首先將調查填復項目如職業別內容（資料 X）例：Registered Nurse（護理師），各個單字分別轉為電腦可認得之數字向量（如 R 轉為 (0,...,0,1,0,...,0)，第 18 個位置為 1，其他為 0，e 則轉為 (0,0,0,0,1,0,...,0) 等）；惟因所轉換的資訊較稀疏，須經大量學習，再篩選單字特性之重要資訊，使資訊轉為較密集之特徵向量（使電腦辨識並篩選單字之特徵）；再將各單字特徵向量依序組回填復內容，並模仿人類僅將較重要特徵保留在長期記憶，並經上下文多個單字比較，獲得該填復內容之重要特徵矩陣（包含重要單字特徵及單字間之關係特徵）；最後將矩陣轉為目標編碼之可能機率向量，並訂定門檻值，若編

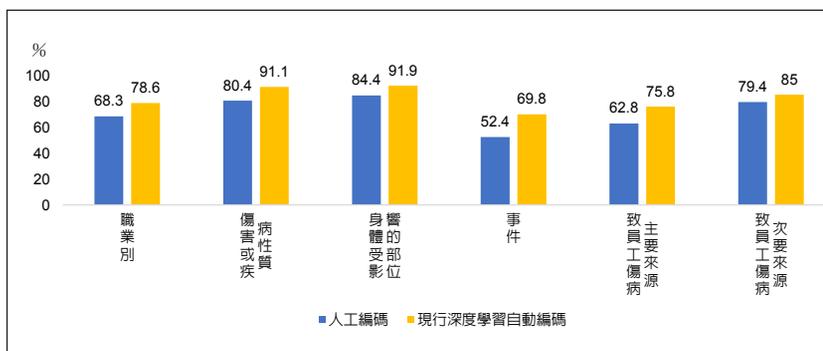
碼之可能機率值大於或等於門檻值則採用。

3. 模型有效性：透過訓練及驗證資料集，在損失函數最小下，獲致模型之適當參數；為確認模型對測試資料集是否有效，以準確（accuracy）率比較。依圖 3 所示，監督式深度學習模型之各項編碼準確度均優於人工編碼。
4. 模型效益：目前各項目編碼之自動程度平均已達 85%，此模型已大量減輕編碼的人力，使人力轉至審查自動編碼之品質，並增進調查回復資料之內容確認等工作。

肆、值得我國借鏡之地方

由前述聯合國分類及編碼之實例可知：將統計調查之人工編碼，改為電腦自動編碼，除減少人工編碼成本外，更可增進編碼效率及品質。我國所辦理政府統計調查如工業及服務業普查之製造業表，亦需對經營方式、生產品項及原材物料項，依填復內容各別檢核行業代碼、編撰產品及原材物料碼，也耗用大量人力，若能引用美國 BLS 之模型，除可提升處理編碼之效能及確度外，亦能將人力轉移至調查內容品質之精進。惟 BLS 模型為英文之

圖 3 美國勞工局有關職業傷害及疾病調查－編碼方法之準確度比較



資料來源：Deep neural networks for worker injury autocoding, Alexander Measure, U.S. Bureau of Labor Statistics。

編碼，英文單字間係以空白間隔，單字區隔較易，若要套用該模型，則須對中文分詞，如產品項之中文描述為「其他電腦週邊設備」，分詞後為「其他 / 電腦 / 週邊 / 設備」，分詞雖可引用中研院所開發之中文斷詞系統，但中文斷詞可能使模型之矩陣維度加大，參數調整之難度亦提高，將是中文分類及自動編碼須克服的部分。

另外二個主題之研究，未來可用於處理統計調查之遺失值（Missing Value）及運用如從衛星影像圖萃取資訊，取代（或勾稽）部分統計調查項目。惟此三個主題之實例多採監督式學習的模型，需大量人工標註答案，且模型是否適用不同的訓練資料集，其適用性仍須加以驗證。

伍、結語

觀諸目前國際政府統計對 ML 技術之應用情形，多屬 ML 之文字及影像辨識之應用，於 GSBPM 處理階段之部分流程，確實有改善效能及提升確度之效益，對現行編製統計做法可

扮演輔助角色，惟各國政府統計係依據「政府統計基本準則」編製，為確保統計品質，須符合準則之公正、標準及透明等特性，而 ML 所萃取之統計資訊，是否符合前揭準則，或者可導入 GSBPM 之其他子流程，仍維持統計高品質，皆是評估應用於產製政府統計流程所必須面對的課題與挑戰，未來仍須廣泛討論以取得共識。隨著 ML 應用之品質框架訂定，以及政府統計基本準則之適用性調整，將可望使 ML 與政府統計之合作距離，越來越近，值得持續關注。

參考文獻

1. 科技部（民 106），人工智慧（AI）推動策略。
2. 曾婉菁（民 107），機器學習探究，印刷科技，34 卷 2 期。
3. High-Level Group for the Modernisation of Official Statistics（2019）. Introduction to GSBPM, UNECE.
4. Martin Beck, Florian Dumpert, Joerg Feuerhake（2018）. Machine Learning in Official Statistics.
5. Wesley Yung（Canada），Jukka Karkimaa（Finland），Monica Scannapieco（Italy），Giulio Barcarolli（Italy），Diego Zardetto（Italy），Jose Alejandro Ruiz Sanchez（Mexico），Barteld Braaksma（Netherlands），Bart Buelens（Netherlands），Joep Burger（Netherlands）（2018）. The use of machine learning in official statistics, UNECE Machine Learning Team.
6. Yoon Kim, Yachine Jernite, David Sontag, Alexander M. Rush（2016）. Character-Aware Neural Language Models. ❖