應用機器學習平台於衛星產品的開發

章鶴群 [1]　陳冠儒 [1]　張育承 [1]

中央氣象局第四組 [1]

林賢宏 [2]　羅時宜 [2]

樺鼎商業資訊股份有限公司 [2]

藍秀仁 [3]　蔡宜真 [3]　楊惠婷 [4]

DataRobot

Taiwan[3]　Singapore[4]

摘　要

傳統的衛星產品的開發過程中，需花費心力，收集各式研究文獻，並對資料特性有所了解，才能夠利用資料的物理特性開發出準確且實用的產品。若產品的特性並不十分明確，或者在定義上非常主觀，則十分難以找出資料和產品之間的關連性，這時機器學習則是一個好用的工具，能夠分析大量數據搭配人工智慧演算法，找出資料與目標產品的最佳關聯性。DataRobot 是一款雲端的機器學習平台，可以減少機器學習演算法開發的時間，將收集到的大量衛星觀測數據及相對應的驗證資料上傳到雲端平台，經過演算法績效排序，即可得到自動推薦的一組最佳演算法。DataRobot 提供統計及機器學習功能計算資料特性及視覺化界面，能達到快速偵測的精度與準度，得知產品的可信度及可用性，讓產品開發的人員可以省下大量撰寫程式碼的時間。

霧和雲量這兩個天氣現象，為人工觀測項目，目前難以自動觀測儀器取代，期望能夠使用機器學習找出觀測數據與人工觀測的關係。若能夠透過機器學習找出一個適合的演算法，則可以應用衛星資料，在沒有人工觀測的地點，使用機器學習的結果取代。本研究嘗試以 DataRobot 應用在霧的產品和雲量的產品，利用其演算法自動選擇的特色，快速得到最佳的訓練結果，並且使用其具有人性化和可解釋性的視覺化介面，檢視機器學習的統計結果。

關鍵字：機器學習、衛星、霧、雲量

一、　前言

近來為配合地面氣象站的轉型，對於使用電子儀器觀測的氣象參數，都改用電子式的儀器記錄，一是能夠減少人工判讀的誤差，增加觀測的次數，二來是能夠將觀測資料電子化並匯入電腦的資料庫中，以利於日後的分析及應用。但是仍有一些需要人工觀測的項目無法直接用觀測儀器得到，本研究所要探討的霧及雲量的觀測即是其中的項目，而衛星資料的衍生產品或許可以解決無法使用電子儀器直接觀測的困難。但是，衛星的觀測是高空往地面觀測，而氣象觀測員則是於地面往天上觀看，這兩者因視角的不同對於所觀測到的現象必定不同，如何在衛星的觀測和觀測員的觀測之間找到一個關聯性？機器學習或許能夠給出一個合理的解答。

市面上有眾多機器學習平台提供各式各樣演算法，如要一一研究、編寫程序並交叉比對所有的演算法，其工作量過於龐大，氣象局內部研發能量並不足以完成上述工作，故期待有一個開發平台，能夠引用新的機器學習演算法，快速建構預測模型，增加產品開發速度並同時提升準確度。

DataRobot(https://www.DataRobot.com/)是個能夠滿足上述條件的一個雲端自動化機器學習平台，具有自動化機器學習平台必備的所有功能，於上傳預測建模資料之後，可自動執行資料預處理、特徵工程、預測模型建立、準確度排序、參數調整、預測模型視覺化呈現等機器學習建模所需之各項功能。因此本研究採用 DataRobot 的各項功能來進行霧和雲量產品的開發。

二、 研究方法

（一）、 資料來源

本研究的資料來源為向日葵 8 號衛星的可見光及紅外線頻道觀測資料及其雲相關的衍生產品。衍生產品是由 CLAVRx 所產生的，CLAVRx 全名為 The Clouds from AVHRR Extended processing System，是由 NOAA/NESEDIS 和 UW/CIMSS 所發展（Heidinger 2012），原為了從 AVHRR 資料中取得雲的定量估計和定性分析產品，後來經過多次改版和擴充，除了能夠應用在繞極衛星的儀器 AVHRR、MODIS、VIIRS，也能應用在傳統的地球同步衛星 MTSAT、COMS，和新一代的地球同步衛星 HIMAWARI 和 GOSE-R。CLAVRx 的雲產品可分三類，一類是定性的分類，有雲遮（Cloud Mask）、雲類（Cloud Type）、和雲相（Cloud Phase），對於分辨雲的屬性十分有幫助；一類是針對雲頂特性的定量估計，包含雲頂高度（Cloud Top Height）、雲頂氣壓（Cloud Top Pressure）和雲頂溫度（Cloud Top Temperature），可做為判斷對流系統發展情況，以做為天氣預報的參考。第三類是雲的物理和光學特性，雲放射率（Cloud Emissivity）、雲光學厚度（Cloud Optical Depth）和雲滴粒子有效半徑（Cloud Effective Radius），對於雲物理的研究有幫助（章 2015）。

（二）、 霧

低雲與霧反演式理論可參考 GOES-R ABI（Calvert 2010）文件，修改部份反演方法及閾值即可應用於向日葵 8 號的 AHI 的觀測資料。衛星反演霧與低雲產品受限於衛星資料特性，例如：難以偵測小範圍的天氣現象、高雲覆蓋時無法偵測地表資訊等。氣象局現有使用衛星資料進行偵測霧與低雲產品，受限於上述條件的限制，僅能針對觀測範圍內搜尋霧的輻射特性，如霧頂亮溫、反照率及其均勻度等等，無法代表地表真實能見度狀況，因此，地面觀測員依據溼度，能見度等數據所判定之霧、靄等天氣現象，若改用氣象衛星的觀測，很可能沒有足夠的訊息來判斷是雲、霧或雲底下有霧。

使用衛星觀測的亮度溫度、反照率和 CLAVRx 產品共同做為機器學習的特徵（feature），氣象站觀測員的觀測數據為預測目標（target）進行建模。觀測員判斷霧的方式則是以目視的能見度為依據，能見度小於 1 公里為霧，大於 1 公里且小於 10 公里則可再依相對溼度分為霾（<75%）和輕霧（>75%）。

本研究資訊使用的時間為 2019 年 1 月到 6 月，範圍為東亞地區，資料內容則包含做為預測目標的氣象站觀測資訊（有霧及沒霧），及做為 Features 的氣象站所在位置之地理資訊、衛星觀測資料和 CLAVRx 提供的 Level2 產品。本研究設計了三組不同的 Features，讓 DataRobot 進行訓練：Run0 包含所有衛星頻道和 CLAVRx 所能提供的 Features（共 160 項）；Run1 則是從 Run0 中去除幾何資訊，包含方位角、天頂角、高度等（共 155 項），此設計原因是主觀上認為與衛星和地理相關的幾何資訊不會因為有霧或沒霧而有顯著的差異；Run2 則是從 Run0 中去除幾何資訊及 CLAVRx 提供 Level 2 產品（共 83 項），目的是想測試在只有衛星頻道資訊的情況下是否也能夠有好的預測結果。

（三）、 雲量

齊（2008）利用日本 MTSAT 衛星紅外線數據資料推估氣象站總雲量，以 103 年全年 27 個有觀測員觀測記錄的氣象站觀測進行比對，發現使用衛星資料估計的雲量與氣象站的觀測相關性達 0.68，若日間及夜間分開統計，日間的相關高達 0.75，夜間亦有 0.65（齊和謝，2014），表示衛星估計的雲量可信度高。前述之方法使用向日葵 8 號衛星進行雲量估計的測試，結果顯示衛星的估計雲量仍然十分可靠，並沒有因為不同世代的衛星而有明顯差異（章，2016）。

欲利用機器學習找出衛星觀測的資料和雲量之間的關係，要用於學習的特徵主要有：（1）測站處衛星觀測的亮溫/反照率。（2）以測站為中心半徑 2 到 16 公里亮溫/反照率之平均、標準差、亮溫/反照率高於某閾值的網格數。（3）經緯度、太陽天頂角等。台灣人工氣象站資料（26 站）觀測員記錄的雲量為十分量，分別以數字 1 到 10 代表十分之一到十分之十，是訓練的預測目標。訓練資料的時間為 2018 年 1 月到 2019 年 9 月，每日測站的觀測及衛星資料，共有 162100 筆。

三、 分析與結果

（一）、 霧

經 DataRobot 以 run0,run1,run2 三種不同的 Features 組合進行訓練，表 1 列出訓練後排名前面的三名的 Model，以 RMSE（Root Mean Square Error）作為排序的依據。Run0 包含的 Features 最多，訓練後的 RMSE 最低，排名前三名的演算法均能夠達到 0.2014，Run1 和 Run2 的 RMSE 則隨著 Features 減少而稍為增加，但是 RMSE 均在 0.202 以下，跟 Run0 的分別很小。AUC 則是在所有的實驗中都達到 0.9 以上，表示有命中率高且誤報率低，是十分漂亮的數字。

取一組實際的衛星資料進行測試來察看不同的演算法的差異，圖 1 為選取個案的真實色影像，除了雲帶的分布外，仍可看到一些霧的現像，例如東南亞及韓國的海岸。圖 2 由左至右三張圖分別是 run0 的排名前三名的演算法得到的結果。霧發生地區主要三處特別明顯（東南亞、韓國西邊海岸及中國內陸部份），其中東南亞和韓國海岸的霧，所有 DataRobot 的 Model 都能夠掌握到，霧的預測機率都可以超過 0.8，至於中國內陸的霧區，可能因為十分零星分散且範圍小，所以 DataRobot 沒有掌握得很好。在北海道東北部的群島，大部份的演算法都判為大範圍的霧（應是誤判），只有 Light Gradient Boosting on ElasticNet Predictions 這方法，只判斷霧的範圍侷限在群島上，並且島上的氣象站確實觀測有霧，為較合理的結果。

（二）、 雲量

表 2 為雲量的實驗結果，採用不同的 Feature 組合所得到的 RMSE。衛星中心目前提供的雲量資料的 RMSE 為 2.363，若只統計白天的雲量資料則是 2.144，夜間為 2.762，DataRobot 的結果均優於目前正在作業中的產品，不論白天晚上或全日的統計均是如此，其中最佳的 Feature 組合為 DataRobot 建議名單（使用 13 種頻道），能夠得到最低的 RMSE。

圖 3 為一個實際觀測資料的分析結果，黃色數字為每個衛星網所得到的雲量，而紅色數字則是氣象觀測員在氣象站所觀察到的結果，仔細對照可以發現差異並不大，表示機器學習的預測接近觀測。

四、 結論

本研究首次將機器學習平台 DataRobot 應用於衛星產品研發上，並評估機器學習應用在氣象衛星產品上的可行性。結果顯示機器學習方法具備一定的精確度，並不遜色於有物理法的偵測結果。觀測資料相對不足的位置，如海面及山區等地區，機器學習方法仍舊保有偵測能力。

而 DataRobot 這套系統最大的優勢在於能夠大大的減少程式開發的時間，並且有詳細的視覺化介面能夠洞察機器學習的結果，以衛星資料確實能夠經過大量的資料學習，而學會判斷霧發生的可能性，但是當實際進行預測時，仍然面臨諸如難於結合天氣分析的自動化作業與短時間內必須完成龐大的預測需求的相關問題。若要實際應用於天氣分析與預報的作業需另行測試與評估考量。

參考文獻

齊祿祥，2008：利用衛星多頻道資料估計氣象站觀測之總雲量研究，97 年度中央氣象局研究發展專題，CWB97-1A-08，p43。

齊祿祥和謝瑩薰，2014，利用衛星資料推估氣象站雲量觀測之可行性分析，103 年天氣分析與預報研討會。

章鶴群，2015：地球同步衛星之可見光及紅外線頻道衍生產品，104 年天氣分析與預報研討會。

章鶴群，2016：日本向日葵 8 號衛星資料衍生產品之應用—氣象站雲量估計，2016 年海峽兩岸害性天氣分析與預報研討會。

Calvert,C. and M. Pavolonis, 2010: GOES-R Advanced Baseline Imager (ABI) Algorithm Theoretical Basis Document for Low Cloud and Fog. NOAA/NESDIS STAR.

Heidinger, A.K., 2012: The Clouds from AVHRR Extended (CLAVR) User's Guide. NOAA/NESDIS STAR.

圖表

Feature List name		Run0			Run1			Run2		
Features		160			155			83		
Best Model		Light Gradient Boosting on ElasticNet Predictions			Light Gradient Boosted Trees Classifier with Early Stopping			eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning)		
		Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout
	RMSE	0.2011	0.2012	0.2017	0.2034	0.2038	0.2042	0.2149	0.2150	0.2158
	AUC	0.9513	0.9513	0.9514	0.948	0.948	0.9484	0.9304	0.9309	0.9301
Second Best Model		Light Gradient Boosted Trees Classifier with Early Stopping			Light Gradient Boosting on ElasticNet Predictions			Light Gradient Boosted Trees Classifier with Early Stopping		
		Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout
	RMSE	0.2013	0.2012	0.2019	0.2037	0.2042	0.2045	0.2149	0.2149	0.2159
	AUC	0.9518	0.9511	0.9520	0.9446	0.9469	0.9472	0.9289	0.9295	0.9291
Third Best Model		eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features			eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning)			eXtreme Gradient Boosted Trees Classifier with Early Stopping (Fast Feature Binning) and Unsupervised Learning Features		
		Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout	Validation	Cross Validation	Holdout
	RMSE	0.2014	0.2016	0.2022	0.2038	0.2036	0.2048	0.2150	0.2151	0.2157
	AUC	0.9501	0.9512	0.9503	0.9478	0.9483	0.9481	0.9305	0.9304	0.9306

表 1、DataRobot 的實驗以 Features 的內容分為 Run0,Run1,Run2 三個，表列出經過訓練之後排名前面的三名的 Model（以 RMSE 作為排序的依據）。

Fetures的組合	Features個數	驗證期間之RMSE		
		全日	白天	晚上
衛星中心目前提供的雲量資料		2.363	2.144	2.762
所有788個數據輸入	887	1.611	1.379	2.007
只有16個基本頻道	16	1.858	1.597	2.307
16基本頻道加上半徑4公里內的額外資訊	208	1.685	1.437	2.109
16頻道+半徑2公里資訊，無大值數量	48	1.85	1.553	2.351
4種基本頻道加上半徑4公里內的額外資訊	52	1.819	1.557	2.269
4基本頻道+2～12公里額外資訊	148	1.656	1.432	2.044
4基本頻道+2～14公里額外資訊	172	1.65	1.429	2.034
4基本頻道+2～16公里額外資訊	196	1.648	1.427	2.03
4基本頻道+2～16公里額外資訊+太陽天頂角	167	1.665	1.427	2.073
4基本頻道+10～16公里額外資訊	100	1.669	1.437	2.068
B07、B13、B15之 2～16公里平均值資訊	27	1.88	1.696	2.213
參考DataRobot建議名單，刪減部份特徵	96	1.621	1.383	2.026
參考DataRobot建議名單，刪減低貢獻特徵	88	1.634	1.387	2.056
參考DataRobot建議名單， 僅保留5種頻道（B03,B10,B11,B12,B15）	61	1.801	1.529	2.262
參考DataRobot建議名單， 僅保留6種頻道（B03,B07,B10,B11,B12,B15）。	69	1.712	1.493	2.094
DataRobot建議名單（使用13種頻道）	101	1.613	1.375	2.019

表 2、進行雲量產品訓練時所設計的各種 Features 組合的列表，及其結果的 RMSE。

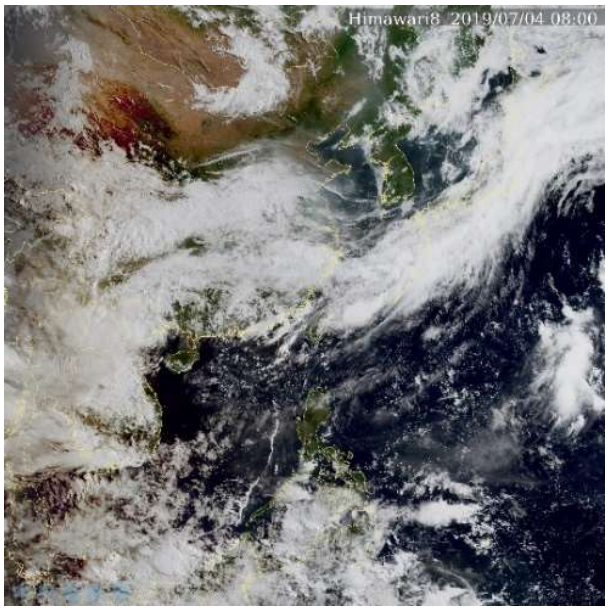

圖 1、2019 年 7 月 4 日 00Z 的東亞地區真實色影像。

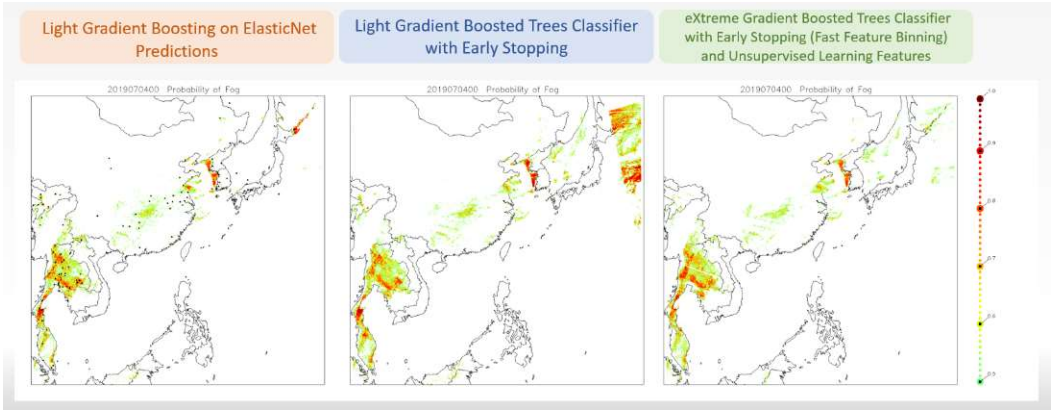

圖 2、2019 年 7 月 4 日 00Z 的衛星資料經過 RataRobot 計算後的結果，左至右為 Feature 組合為 Run0 時間，最佳排名最佳的三個演算法。

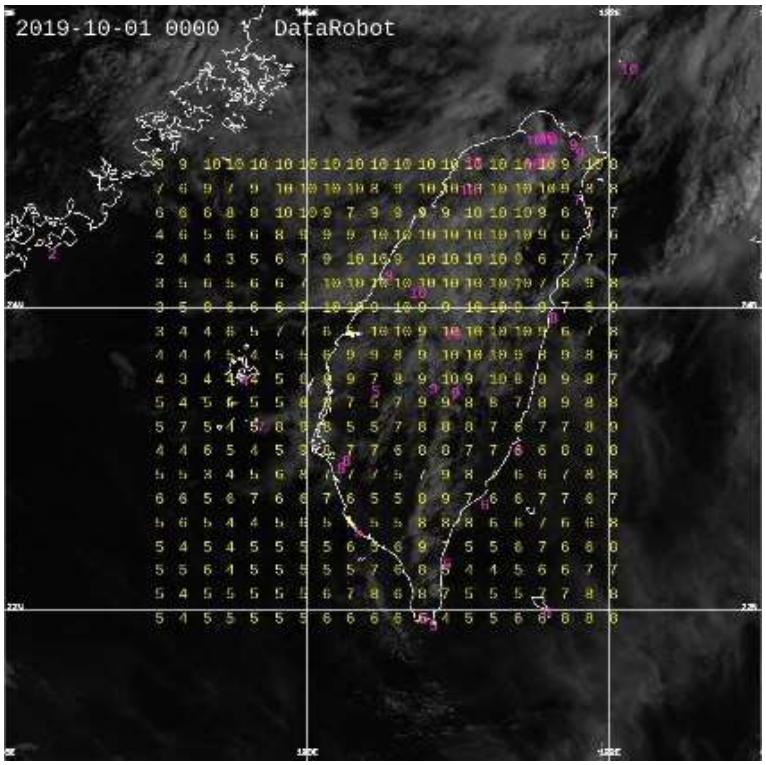

圖 3、可見光雲圖疊加網格上的雲量數值，黃色數字為 DataRobot 的結果，紅色數字為氣象站所觀測的雲量。