各類機器學習方法在太陽黑子數目預測上之效能評估

胡文品[1]陳啟瑋[2]陳健勛[1]朱彥煒[2*]

[1]亞洲大學生物與醫學資訊學系（41354台 中市霧峰區柳豐路 500 號 ）

[2]國立中興大學基因體暨生物資訊學研究所（402 台中市南區國光路 250 號）

[*] ywchu@nchu.edu.tw

摘要

太陽黑子是一種經常出現在太陽上陰暗斑點之物理變化，也是太陽活動的基本標誌。由過去的研究中，推測太陽黑子的週期平均為 11 年；然而週期並非一直維持如此的規律性，其中可能存在許多因素的介入，導致太陽活動異常。因此，本研究選擇許多適合時間序列預測之機器學習方法，挖掘在太陽黑子資料庫中隱藏的訊息，進而建構太陽黑子數目之預測模型。實驗結果發現，迴歸方法的表現較類神經網路與決策樹為佳；再者，以日作為週期單位可獲得相較於年、月、雙週及單週更為準確的太陽黑子數目的預測，其預測準確度可達九成。

關鍵字：*太陽黑子*、*機器學習*、*時間序列預測*

1. 導論

太陽的組成結構大多數是幾個基本的氣體，如氫占 71.3%、 氦占 27%， 至於其它元素大概占 2%。而太陽上的氣層根據不同的性質也可分為三層，分別是光球層、色球層和日冕層。其中光球層之對流層上面的太陽大氣，就是我們平時所見的太陽圓盤，稱為太陽光球。光球是一層不透明的氣體薄層，厚度約 500 千米。它確定了太陽非常清晰的邊界，幾乎所有的可見光都是從這一層發射出來的。其中，光球層會有三個比較特別的特徵：(1) 光斑：在周圍背景有比較亮的光點散布在周圍較小的部分。(2) 臨邊昏暗：往邊緣方向有逐漸減小的日面亮度的現象。(3) 米粒組織：一種太陽表面的結構、小顆粒形狀，外型是多角型 [1]。事實上黑子在太陽表面中是一股炙熱的的巨大漩窩的氣體，大約攝氏 4500 度。也因此溫度相較於太陽的光球表面的光球層要來得較低，看上去是稍微深色的斑點，因為，局部磁場中溫度較低的地方所以也會比較陰暗。

另一方面，太陽黑子在兩極都會出現，時間從幾天到數月都不一定。比較大片的黑子在黃昏或凌晨時，就可以變成肉眼可以觀察到的黑子。早期太陽黑子在還沒被發現之前，有相當大的程度造成人類的危害及困擾，也基於天文科技的進步，人類對於太陽黑子越來越更加認識，但黑子有不固定約 11 年的週期，每次變化都是不同的樣貌，特別在 17 世紀後，有近五十年間，都沒有太陽黑子的出現，甚至相當稀少，而這種就稱做「蒙德極小期」。黑子經常成群出現，比較少單獨活動，活躍時會對地球的磁場產生影響，主要是使地球南北極和赤道的大氣環流作經向流動，從而造成惡劣天氣，使氣候轉冷，嚴重時會對各類電子產品和電器造成損害。也有研究指出除上述所提到會危害人類的情況外，太陽黑子也存在下列情況發生太陽黑子之於人類的影響，包括通訊、交通及經濟的動盪，例如：1. 太陽黑子數目增長後全球經濟將會進入停滯時期。2. 黑子數目上升也可能導致社會動盪不安，治安的變化。3. 太陽黑子增多的時候，地震頻繁，也有可能引發海嘯。基於以上幾點，我們若能預測太陽黑子出現的週期就顯得非常重要，可以提供各類專家學者做為決策或判斷的依據[2-3]。

過去較多的文獻都是從天文科學的角度去預測太陽黑子的發生，近代也有許多專家學者應用不同的資訊技術在這個議題上面，本篇論文我們利用 Java 程式去擷取 NOAA (國家海洋大氣管理局) 的資料庫當作我們的建構預測模型所需的學習資料。因為太陽黑子的發生性質是屬於時間序列，因此，在分析建構預測模型時，我們想要探討不同的時間週期是否會影響各種機器學習方法的效能，結果也顯示日的週期的確比其它時間週期的預測要來得佳。在預測方法的選擇上，除了使用許多迴歸類別的方法外，本篇論文亦加入人工智慧之類神經網路和規則類別之決策樹來比較；最後結果顯示結合規則與迴歸性質的方法 M5 Rules 常優於其它方法的表現。

2. 資料集的建立

本研究所使用的學習資料，係採用 NOAA (National Oceanic And Atmospheric Administration) [4] 國家海洋大氣管理局之 Space Weather 資料庫所組合而成，而這些數據是由 STP (Solar - Terrestrial Physics Division) 負責太陽能和空間環境數據和衍生產品由 NOAA 觀測系統收集，並通過世界太陽和地球物理學數據中心收購歸檔和訪問所整理而成的。其中國際太陽黑子數

(International sunspot number) 的資料是由 SIDC (Solar Influences Data Analysis Center) 在比利時皇家天文台觀測而來，資料的總數量是大約幾萬多筆，實際數目依太陽黑子每次的週期的總數量而有所不同[5]。

實驗的第一步將建構預測所需之資料集，因此，本研究將 NOAA 獲得之太陽黑子數目資料以年、月、雙週、單週及日區分做為週期單位，流程如圖 1 所示，獲得年 64 筆、月 23,680 筆、雙週 1,579 筆、單週 3,299 筆及日 23,680 筆。再以 13、15、17 和 21 作為週期刻度，例如 13 日；則以該日之前 13 日作為時間序列編碼、13 年：則以該年之前 13 年作為時間序列編碼，其數量如表 1 所示。

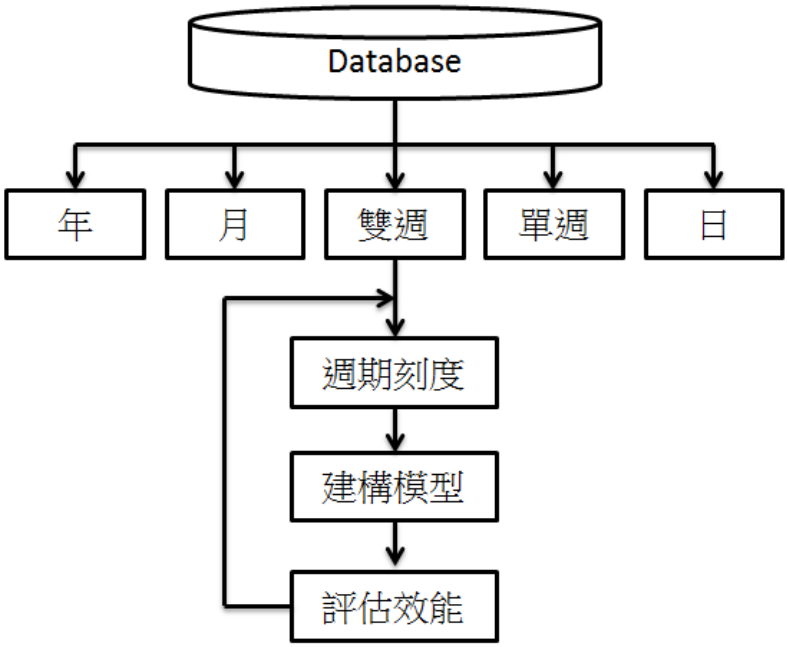

圖1 實驗流程

表1 太陽黑子訓練資料數量

序列長度 / 週期	13	15	17	21
年	51	49	47	44
月	775	753	751	747
雙週	1565	1563	1561	1557
單週	3286	3284	3282	3278
日	23668	23666	23664	23660

3. 實驗方法

3.1 測試週期區間

從 NOAA 資料收集太陽黑子資料並建立訓練所需之資料集後，將以機器學習方法建構預測模型。然而，預測太陽黑子乃時間序列之問題，若週期區間選擇適當較能呈現問題型態，則對於機器學習將有學習上的幫助。為此，本研究將選取 13、15、17 及 21 的週期間距，給予機器學習建構預測模型，並評估其週期區間之準確度。

3.2 機器學習方法

我們主要選擇迴歸、人工智慧、決策樹三種不同類別的方法來分析：

Linear Regression

線性迴歸是研究單一依變項 (dependent variable) 與一個或以上自變項 (independent variable) 之間的關係。本篇論文使用的模組是赤池信息量準則 (Akaike criterion)，並能處理權重實例。該方法主要有兩個用途：預測和因果分析。預測是用來觀察變數預測依變項；因果分析則是將自變分析成依變相發生原因。

Multilayer Perceptron

多層感知器神經網路是一個向前傳遞訊息，反向傳導學習的分類器。該方法多用於非線性之分類問題上，本篇論文採用的激勵函數為 sigmoid function。

SMO reg

此方法主要是實現支持向量機處理迴歸問題的能力。在最小風險函數之參數的尋找，本論文使用的演算法為 RegOptimizer。

K-Star

‘K’是一種基於實例的分類器，該種分類器係利用測試與訓練資料的相似度來達到分類預測。不同於其他基於實例的學習方法，該方法採用了基於熵的距離函數。

M5 Rules

利用 separate-and-conquer 方式來產生的決策列表以解決迴歸問題。在每一迭代，利用 M5 演算法建構模式樹，並選擇最好的分支作為規則。

REP Tree

快速決策樹學習法，使用信息增益/方差來建立決策/回歸樹，同時亦使用 reduced-error pruning 演算法加速樹的搜尋。如同 C4.5 演算法一樣，REP Tree 對於 missing value 是以相對應的實例做切片處理。

以上六個用來建構預測模型之機器學習方法，本研究以 Weka [6] 軟體實現，六個方法皆使用預設參數訓練模型。

3.3 評估模型

交叉驗證是一種用來確定模型有效性的一種做法，而交叉驗證的方法種類很多，其優缺也不同，本研究以常見之 K-fold Cross-validation [7] 做為實驗之驗證方法，將資料分為 K 的子集，每個子集均做一次測試驗證，其他 K-1 個子集做為訓練集，測試機器學習訓練之模型在沒學習該資料的情況下之正確率， 再將 K 次準確度之平均做為識別正確率的結果，而本研究將 K 設為 10 做 10 組交叉驗證，計算準確度的方式則以 Pearson's Correlation Coefficient 作為使用。

4. 實驗結果

以年、月、雙週、單週及日作為訓練資料，搭配不同性質之機器學習方法建構預測模型，測試並尋找適合之組合，其結果如表 2-6。由測試結果表 2 可見，利用 17 年作為編碼以 SMOreg 可得到 0. 9427 之準確度(Correlation Coefficient)。而表 3 月作為週期之編碼，其準確度最高為使用 LinearRegression 和 M5Rules 以 21 個月做編碼有較高的準確度 0.9442，從兩者結果顯示，年與月預測準確度差異並不大。表 4 為雙週週期作為編碼使用之預測結果，在 15 及 17 個雙週上以 LinearRegression 即 M5Rules 有相較於其他方法高的準確度 0. 9039，而單週結果如表 5，其最高準確度落於使用 SMOreg 之 15 個週的週期編碼，且準確度與雙週相差甚小，由此可見雙週與單週較無明顯之準確度差異。使用日做為週期結果如表 6 所示，可見 17 日週期編碼與 M5Rules 方法可獲得 0.9795 之準確度，高於年、月、雙週和單週之實驗結果。然而，使用年與月其準確度差異甚小，且雙週與單週除了兩者差異不大外，其準確度也低於年與月，而以日做為週期使用準確度卻可優於於其他年、月、雙週和單週，因此可能可作為太陽黑子研究上的一個參考訊息，待未來的研究做進一步的釐清。

表 2 年週期預測準確度

Method	**Window size**			
	13 年	15 年	17 年	21 年
LinearRegression	0.9067	0.8968	0.9136	0.8659
MultilayerPerceptron	0.8673	0.8394	0.9276	0.9286
SMOreg	0.9133	0.9115	**0.9427**	0.9348
Kstar	0.8847	0.9186	0.8883	0.8436
M5Rules	0.8920	0.8370	0.8608	0.8522
REPTree	0.8236	0.7741	0.6962	0.6939

表 3 月週期預測準確度

Method	**Window size**			
	13 個月	15 個月	17 個月	21 個月
LinearRegression	0.9433	0.9436	0.9440	**0.9442**
MultilayerPerceptron	0.9000	0.8661	0.8919	0.8836
SMOreg	0.9429	0.9439	0.9443	0.9434

Kstar	0.9166	0.9122	0.9183	0.9172
M5Rules	0.9433	0.9436	0.9440	**0.9442**
REPTree	0.9241	0.9235	0.9285	0.925

表 4 雙週週期預測準確度

Method	**Window size**			
	13 個雙週	15 個雙週	17 個雙週	21 個雙週
LinearRegression	0.9034	**0.9039**	0.9030	**0.9039**
MultilayerPerceptron	0.8639	0.8479	0.8351	0.7819
SMOreg	0.9024	0.9029	0.7917	0.9017
Kstar	0.8520	0.8405	0.8402	0.8382
M5Rules	0.9034	**0.9039**	0.9030	**0.9039**
REPTree	0.8748	0.8800	0.8739	0.8737

表 5 單週週期預測準確度

Method	**Window size**			
	13 個週	15 個週	17 個週	21 個週
LinearRegression	0.8923	0.8932	0.8928	0.8041
MultilayerPerceptron	0.8535	0.8322	0.8916	0.8041
SMOreg	0.8925	**0.8936**	0.8928	0.8930
Kstar	0.8359	0.8259	0.8232	0.8062
M5Rules	0.8923	0.8932	0.8921	0.8928
REPTree	0.8604	0.8575	0.8575	0.8585

表 6 日週期預測準確度

Method	**Window size**			
	13 日	15 日	17 日	21 日
LinearRegression	0.9791	0.9792	0.9136	0.8943
MultilayerPerceptron	0.9737	0.9729	0.9276	0.9319
SMOreg	0.9777	0.9791	0.9123	0.9427
Kstar	0.9653	0.9622	0.9603	0.8602
M5Rules	0.9793	0.9794	**0.9795**	0.9077
REPTree	0.9766	0.9766	0.9764	0.6321

5. 結論

本研究目的在於如何利用過去長時間所累積的太陽黑子數目資料，以機器學習方法預測未來太陽黑子可能的活動情形。由目前的實驗結果可推測，以日做為週期性預測較為準確，但由於時間序列其週期性組合較為複雜且費時費力，因此本研究測試 13 到 21 之週期以簡化計算之時間複雜度，其中可能並非最適當的週期性，未來可加入基因演算法 (Genetic Algorithms，GA) 及模擬退火法 (Simulated Annealing，SA) 等方法用以搜尋出可能之最佳週期。

此外，雖然預測模型具有太陽黑子數目的預測，但若有更多的數據輔佐，針對太陽黑子的發生情況做詳細的統計和分析可能將有助於進一步的發現。未來期望配合地球大氣變化和天文變化，針對太陽黑子對於人類的災害做更深入的預測及分析，已達到更完整的天文預測。

6. 參考文獻

[1] Preminger, D. G., & Walton, R. R. (2007). From Sunspot Area to Solar Variability: A Linear Transformation, *Solar Physics, 240(1),* 17-23.

[2] Li, K. J., Su, T. W., & Liang, H. F. (2004). The periodicity of sunspot activity based on modern sunspot observations, *Chinese Science Bulletin, 49,* 2511-2516.

[3] Li, K. J., Su. T. W., & Liang, H. F. (2008). The prediction of smoothed monthly mean sunspot number for the twenty-fourth solar cycle, *Science in China (Series G), 38,* 1097-1105.

[4] NOAA sunspot numbers database. *http://www.ngdc.noaa.gov/nndc/struts/results?t=102827&s=5&d=8,430,9*

[5] WU, Y. L., DONG, F. A., & LI, S. F. (2004). Predicting on the time series of sunspot number based on chaos theory, *Journal of Shanxi Normal University (Natural Science Edition), 32(6),* 81-83.

[6] Witten, I. H., & Frank, E. (2005). *Data Mining: In: Practical Machine Learning Tools and Techniques*, Elsevier, Amsterdam, Second edition.

[7] Kohavi, Ron. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 2 (12),* 1137–1143.

Evaluating various machine learning approaches for the sunspot number prediction

Wen-Pin Hu[1] Chi-Wei Chen [2] Jian-Shiun Chen [1] Yen-Wei Chu [2*]

[1] Biomedical Informatics, Asia University

[2] Institute of Genomics and Bioinformatics, National Chung Hsing University

[*] ywchu@nchu.edu.tw

ABSTRACT

Sunspots often appear on sun that is the basic signs of physical changes of dark spots and solar activity. Sunspot cycle is an average of 11 years by past studies speculated. However, the cycle was not always maintaining regularity which may be many factors involved so that there lead to abnormal solar activity. Thus, this study chose many suitable for time series prediction of machine learning methods. Mining sunspot database find the hidden information and then building sunspot number prediction models. Experimental results show that the performances of regression approaches are better than neural networks and decision trees. Moreover, the model using day as a cycle units is better than other units which had 90% accuracy.

Keywords: *Sunspot, Machine learning, Time series prediction*