巢狀超矩形學習模式應用於逕流量推估之研究

金紹興
經濟部水資源局副局長

馮德榮
經濟部水資源局科長

陳莉
中華工學院副教授

楊人傑
中華工學院研究生

馬家驎
經濟部水資源局助理工程司

摘要

本研究利用巢狀超矩形學習模式之理論架構，來預估全球氣候變遷對台灣地區逕流量變化之影響。此模式是以人工智慧領域中機器學習方式來達到預測之目的，藉由許多歷史資料點的自變數形成幾何上的超矩形結構，可節省大量的記憶體空間外，並在訓練階段加入新資料時不斷修正參數值，使預測的準確性愈來愈高。

首先以一理論之數學函數為測試範例，結果顯示當訓練資料個數增加時，模式預測值與真值之線性相關係數逐漸趨近於 1，表示其學習能力優良。

實例應用於推估逕流量，以台灣之北部地區為研究對象，視降雨量和蒸發量為輸入因子（自變數），經由巢狀超矩形學習模式來預測逕流量（應變數）與傳統之線性複迴歸比較，結果顯示其推估之準確度較佳。

一、前言

本研究希望探討氣候變遷對台灣地區水文環境之影響，故旨在建立逕流推估方法，其中最直接者即由降雨和蒸發資料為輸入因子，而逕流為輸出變數，形成一水文黑盒模式，而應用巢狀超矩形學習模式來達成預測，以北部地區為研究對象進行分析。

研究方法採用一種模擬人類智慧對事物學習、經驗累積的『巢狀超矩形學習模式』(Nested Hyper -rectangles Learning Model，NHRL) 來作推估工作。此方法由於無預設函數的型態，所以不須推估各種水文或地文因子，模式內所含的參數完全隨著所學習的觀測資料而動態的調整，所以若遇到不合理或異於常態的擾動(noise)點也會隨著學習對象的增加而逐漸將其淘汰。也由於其基本架構的簡易，所以適用於任何參數個數，不必因參數個數的改變而重新推導公式，所以為一種彈性極高的模式。

二、超矩形學習模式之理論特質

『依範例學習』之理論簡稱為 EACH（ Exemplar-Aided Constructor of Hyper-rectangles），其基本假設為以相同領域中先前發生過之例子（examples）為基礎，作為預測或分類之依據，經由過去的例子持續的增加，最初係以點（point）儲存在歐氏 n 維空間（E^n）中，n 代表在一個例子中的變數（ variables ）或特徵（ features ）數目，NGE（ Nested Generalized Exemplar Theory ）理論又將上述各點形成超矩形結構（ Medin and Schaffer, 1978 ），當例子的個數增多時，則一些例外（exceptions）可能在超矩形中產生，即所謂的「洞」（holes），這些洞裡面又可能有其它的洞，結果形成了「巢狀」（nested）的超矩形結構。

EACH 主要的概念為將新的樣本與先前發生過的例子做比對，至於如何決定那一個例子和新的樣本最為近似，係利用『類似度計量』（Similarity Metric），也可稱為『距離計量』（Distance Metric），因為它量測了各例子與新樣本之間的距離。我們使用 “Exemplar” 表示在記憶中已儲存的 「範例」，而以 “Example” 代表新加入系統中的「樣本」。每一個範例（ Exemplar ）都存著輸出變數的值，並用於預測或分類。簡而言之，EACH 將新樣本的預測或分類變數值與最接近的範例中存在的值設為相同，而所謂學習，是發生在系統經由預測而獲得一些回饋（ feedback ）之後，使系統能依最新的信息，作必要的修正或適度的改善，使其更能與實際狀況相吻合，詳細的步驟將在後幾節中說明。

三、巢狀推廣範例學習之演算法

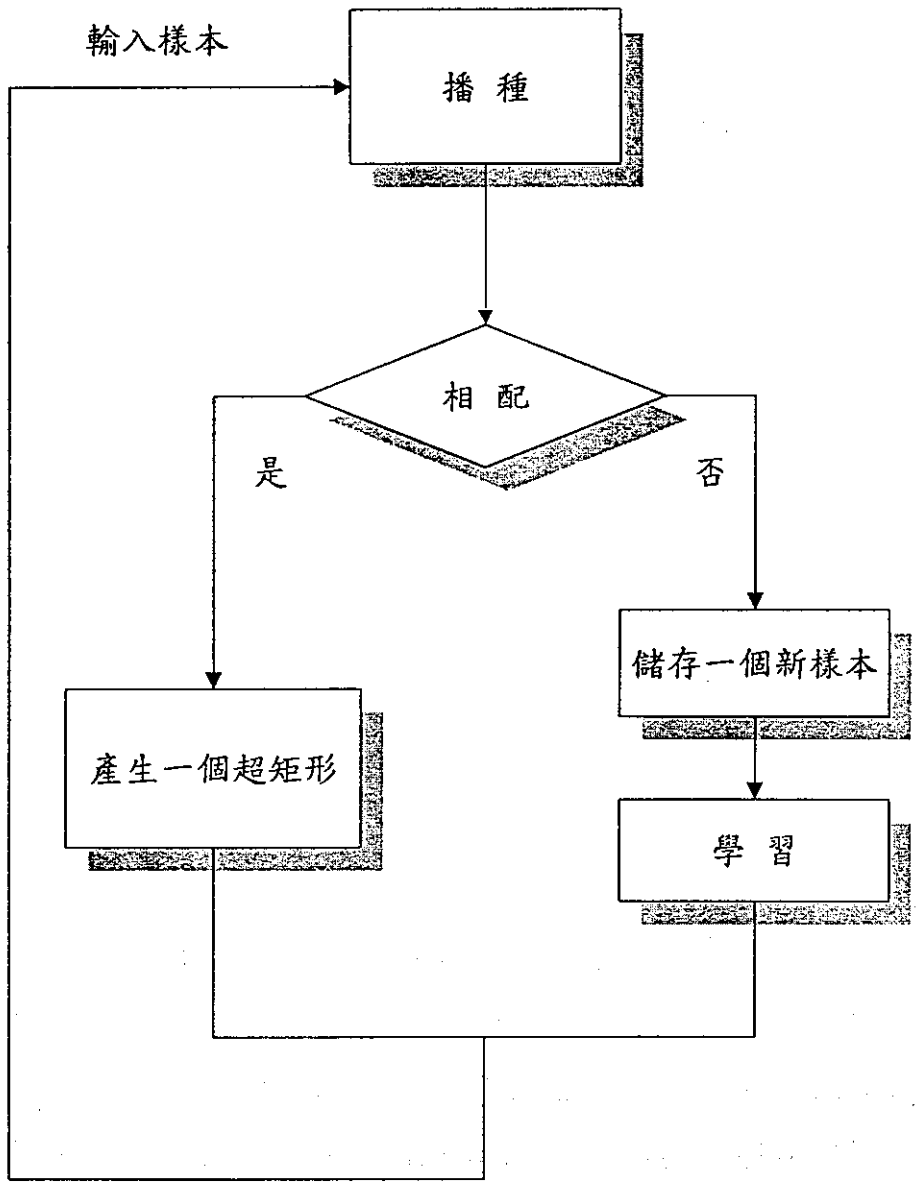


圖 3.1 EACH 演算法之流程圖

圖 3.1 為 EACH 演算法之流程圖，其詳細內容如下：

1.播種（Seeding）：EACH 必須具有一個以上之歷史性的範例，藉以為預測的基礎。

2.相配（Matching）：使用『距離計量』。系統可預測這個新加入的樣本 E 將落入原來範例中那一最接近的群，並計算是否在誤差容許範圍之內？若是則建立一個超矩形，若非則單獨記憶此一新樣本點。

3.回饋（Feedback）：如果系統做了正確的預測，則以 H 和 E 為對角線形成一個新的超矩形。如果系統做了不正確的預測，則在記憶中將這個新的樣本 E 視為一獨立的點而加以貯存。

4.學習（Learning）：如果發現系統預測錯誤時，EACH 可自動調整每一 f_i 的權重 W_i：

(1)如果某項 E_{fi} 太接近 H_{fi}，則設定 $W_i = W_i(1 + \Delta f)$

(2)如果某項 E_{fi} 不接近 H_{fiv}，則設定 $W_i = W_i(1 - \Delta f)$

四、模式驗證

為驗證巢狀超矩形學習模式在連續變數的推估能力，本研究利用蒙地卡羅（Monte Carlo）方法模擬函數值作為驗證，繁衍（generate）出兩組具均勻分佈特性的X_1, X_2資料，再依下式計算得一系列相對的Y值。

$$Y = \text{Sin}\, X_1 + 10\, \text{Cos}\, X_2$$

其中 $X_1 = 0 \sim 2\pi$

$X_2 = -\pi \sim \pi$

為評斷模式的優劣，以真正Y值和推估$\hat{Y}$值之線性相關係數r做比較。設資料個數為n，則：

$$r = \frac{n\sum_{i=1}^{n} y_i \hat{y}_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} \hat{y}_i)}{\sqrt{[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2][n\sum_{i=1}^{n} \hat{y}_i^2 - (\sum_{i=1}^{n} \hat{y}_i)^2]}}$$

r代表是真正Y值和推估$\hat{Y}$值之間線性關係的衡量值，也就是Y值和$\hat{Y}$值之相關程度，通常r介於 +1 與 -1 之間，當r等於±1時，樣本中的Y值和$\hat{Y}$值即呈現出完全的線性相關，若趨近於±1時，則代表良兩者具有高度的線性相關，當然當r接近零時，則Y值和$\hat{Y}$值之間的線性關係很微弱或不存在。另外，r^2（稱為樣本決定係數(Sample Coefficient of Determination)）代表$\hat{Y}$的總變異中可以被線性相關中Y值解釋的比例。即當r＝0.7時，就代表$\hat{Y}$的總變異中有49%可以用線性相關中的Y值解釋。

將實驗分為訓練程序與預測程序，先進行訓練程序，隨機選取n組（X_1，X_2，Y）依序加入巢狀超矩形模式中，不同資料點個數n為10、20、30、40、50、60、70、80、90、100，因其特徵變數（自

變數）只有 X_1 與 X_2 兩個，所以『距離計量』公式中的 i＝2，而『誤差寬容』設為 0.1$\bar{Y}$，經整個演算過程後(如圖 3.1 所示)，此 n 個數逐漸形成多個不同的超矩形或單獨點，即完成訓練程序。預測程序為推估相同的 100 個點。將此 100 組未知 Y 值之（X_1，X_2）一一加入已訓練完成的巢狀超矩形模式中，此時不對模式的任何參數做調整，故可驗證模式預測 Y 值之正確度。

本研究在訓練程序中分別針對上述 10 種情況（資料點個數為 10 至 100）重複 10 次實驗，再將不同訓練個數對應之 10 次平均線性相關係數 r 繪於圖 4-1，可看出當訓練個數增加時使 r 值逐漸遞增，表示預測結果愈來愈好。當訓練個數為 100 時，巢狀超矩形模式之線性相關係數已高達平均 0.96，相當趨近於 1，顯現極佳的預測能力。

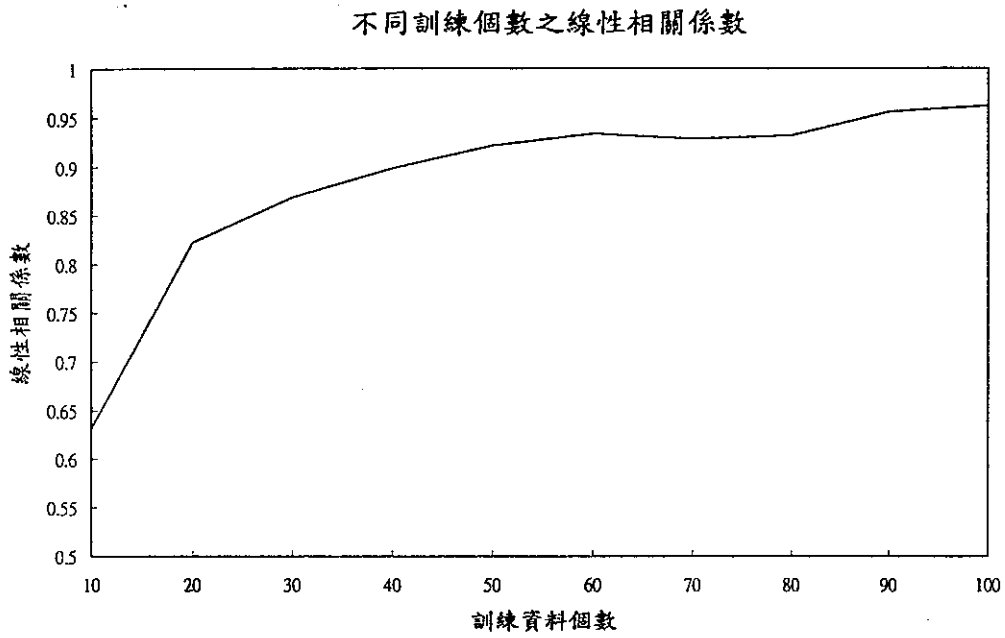


圖 4-1 巢狀超矩形學習模式於數學函數之預測結果

五、應用實例

本研究流域之選取考慮使模式的結果能較有代表性，在台灣北部區域選定一流域，並挑選一水文測站，而測站上游集水區即為研究對象。經選定為北部淡水河支流大漢溪玉峰站上游流域作為本模式演算模擬區域。在選定水文站集水區內擇選數個雨量站及蒸發量測站；至於資料採計年限，則以同一流域內各水文氣象測站同時有記錄者為準。選定之流量站、雨量站及蒸發量站如附表 6-1 所示。

流域	面積	流量站	雨量站（權重）			蒸發量站	年份	缺少	檢定年份	驗證年份
大漢溪	335.29	玉峰	白石 (0.43)	鎮西堡 (0.5)	玉峰 (0.07)	玉峰	1967-1990	1986，1987	1967-1981	1982-1990

表 5-1. 模式之輸入資料

首先將月逕流量與本月之蒸發量、降雨量做相關分析，再與前一月之逕流量、蒸發量及降雨量進行相關分析，如表 5-2 所示，找出線性相關係數最高的兩個因子：本月之降雨量（r＝0.59）與前一月之蒸發量（r＝0.61）為模式輸入之自變數。

	逕流量	蒸發量	降雨量
本月	1.0	0.25	0.59
前一月	0.26	0.61	0.23

表 5-2 逕流量與各因子之線性相關係數 r

以巢狀超矩形學習模式進行訓練程序，採用之檢定年份為

1967-1981 年，將此 15 年（180 個月）之資料依序輸入模式，其中誤差寬容 $e=\left(\frac{1}{3}\right)\bar{Y}$，$\bar{Y}$為逕流量之平均值，結果逕流預測值與其真值之線性相關係數為 0.9971（圖 5-1），本研究與線性複迴歸模式比較，以同樣資料的建立之迴歸方程式為

$$Y=-0.250198+0.304526X_1+0.592293X_2$$

式中 Y＝月流量，

X_1＝前月蒸發量，

X_2＝本月降雨量，

而線性相關係數為 0.855，故在訓練階段時巢狀超矩形模式之結果較理想。

接著以驗證年份 1982-1990（缺 1986，1987）為輸入，共計 7 年（84 個月）之資料，此時巢狀超矩形學習模式中之各項參數已固定不調整，結果逕流預測值與真值之線性相關係數為 0.9323（圖 5-2），而以上述之線性複迴歸模式進行預測所得之線性相關係數為 0.8972，顯示巢狀超矩形學習模式之預測能力亦較佳。

茲將兩種模式之檢定與驗證結果（線性相關係數）整理如表 5-3 所示。

表 5-3 兩種模式所得線性相關係數之比較

	訓練程序	預測程序
巢狀超矩形學習模式	0.9971	0.9323
線性複迴歸模式	0.855	0.8972

圖6-1　訓練（檢定）結果

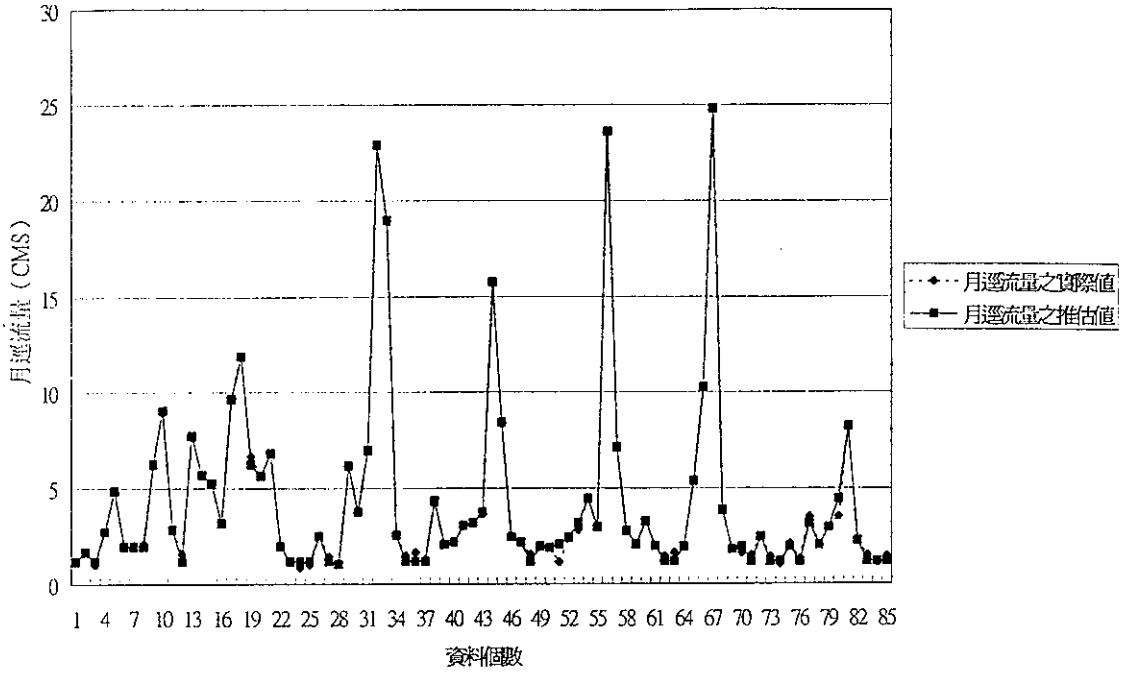
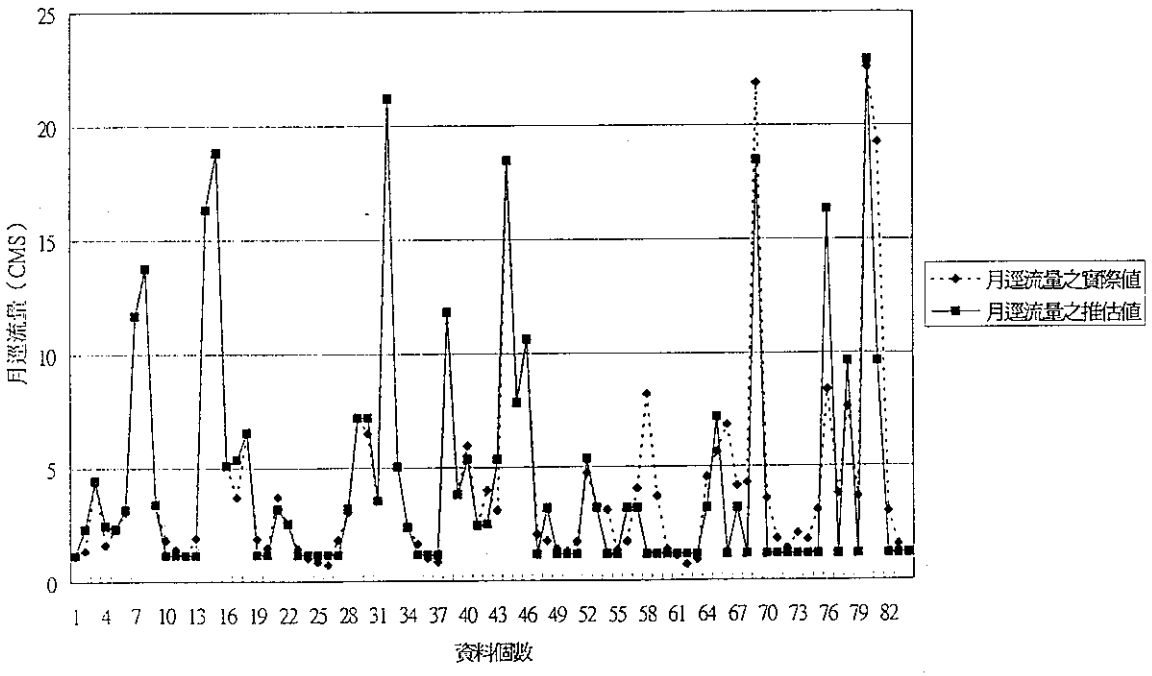


圖6-2　預測（驗證）結果



六、結論與建議

1. 巢狀超矩形學習模式是以歷史性範例為基礎的學習方式，適合應用於資料分類或預測，其基本假設為事件具有相似性，即過去的事件將來有重複發生的可能性。而水文領域中許多資料都符合此種特性，故本研究希望能達到推估逕流之目的。
2. 本模式主要的優點在於利用極少的資源，每個超矩形只需記憶對角線上的兩個點，即可達成令人滿意的預測效果，將不必要儲存的資料都從記憶體中刪除，使資料結構精簡而具代表性，並且隨著訓練樣本個數的增加而使預測結果更準確。
3. 模式中的距離計量公式可隨新資料的加入而動態調整各項參數值，使其預測能力愈來愈佳，故與一般的分群理論不同。但如何更精確的逐步修正參數需要靠經驗法則，若能結合最佳化模式來進行參數調整，例如遺傳演算法，將可獲致更完美的結果。
4. 本研究提出一個理論上的數學函數為測試範例，並實際應用於建立逕流推估模式。前者因完全符合巢狀超矩形學習模式之基本假設，所以可獲得相當理想的預測結果。後者以北部地區為實例，由降雨量及蒸發量來推估逕流量，結果比傳統的線性複迴歸分析方法所得的準確度高。
5. 逕流推估模式利用水文歷史資料架構完成之後，可應用於預測未來因全球氣候變遷所造成的溫度持續上升，進而引起降雨量逐年減少，而蒸發量逐年增加，不可避免的將使逕流量趨於更少。為事先明瞭其後果的嚴重程度，可藉本模式加以估測以便及早擬定因應措施。

參考文獻

1. Medin, D. and Schaffer, M. "Context Theory of Classification Learning" , Psychological Review, 85(3), 207-238, 1978.
2. Michalski, R., Carbonell, J., and Mitchell, T., "Machine Learning" , Tioga Publishing Co., 1983.
3. Salzberg, Steven, "Exemplar-Based Learning: Theory and Implementation.", Technical Report TR-10-88, Center for Research in Computing Technology, Harvard University , 1988.
4. Salzberg, Steven, " Nested Hyper-Rectangles for Exemplar -Based Learning.", Edited by J. Siekmann, Lecture Notes in Artificial Intelligence, 184-201, 1989.
5. Salzberg, Steven, " A Nearest Hyper-rectangle Learning Method." , Machine Learning, 6, 251-275, 1991.
6. 陳莉、張斐章:“巢狀超矩形學習模式於水資源系統之研究”，中國農業工程學報，第三十八卷第三期，1992。
7. 林惠芬:“巢狀超矩形學習模式於河川流量之研究”，台大農業工程學研究所碩士論文，1994 。
8. 林錦全:“衛星影像資訊於集水區地表覆蓋分類之研究”，台大農業工程學研究所碩士論文，1995。
9. 氣候變遷對臺灣水文環境影響之研究，經濟部水資源統一規劃委員會，1995。