使用潛在資訊提升小樣本學習準確率

利德江[1]　張哲榮[2]　林良憲[3]
[1]國立成功大學工業與資訊管理學系 lidc@mail.ncku.edu.tw
[2,*]國立成功大學工業與資訊管理學系 r3795102@nckualumni.org.tw
[3]國立成功大學工業與資訊管理學系 r38991052@mail.ncku.edu.tw

摘要

在高度競爭的製造環境中，有效地控制製造系統是企業生存的必要條件。尤其，製造系統的早期階段更是管理的關鍵。在這個階段，我們通常僅能獲得少量的生產樣本，如何使用這有限的資料完成製程參數的決策，對工程師與管理者而言是一大挑戰。虛擬樣本產生技術是現行處理小樣本數據與不充足資訊的主要方法之一，但其無法有效地處理時間序列資料的缺點，限制了它的應用範圍。為此，本研究發展了潛在資訊函數來分析資料，並萃取出隱藏資訊，以協助小樣本資料的知識學習。經過虛擬控制圖時間序列資料集的實際測試，本研究提出的方法能有效地提升預測準確率，是一個處理小樣本資料的適當程序。

關鍵字：小樣本學習、預測、製造、時間序列。

使用潛在資訊提升小樣本學習準確率

1. 緒論

全球化促使製造環境產生劇烈的變化，不但產品的生命週期縮減，企業的競爭更因此而加劇。如何有效地控制製造系統，其重要性不言可喻。尤其製造系統的早期階段，更是管理的關鍵 (Li et al. 2009b) 。管理者必須要能及時地發現製造問題，並採取適當的處理方式，以維持企業的高效率與效能。因此，擁有適當的預測技術，除了能提升管理效益外，更能協助企業面對嚴峻競爭環境的挑戰。

在製造系統的早期階段，我們僅能獲得少量的觀察值，在這樣的特殊情況下，傳統的預測方式並無法直接地產生理想的結果 (Li et al. 2011) 。此時，若能利用有限樣本建立適當的預測模型，對管理者而言是極具價值的 (Li et al. 2010) 。本研究的主要目的是建立一個以小量樣本為基礎的預測模型，以降低工程師或管理者產生決策錯誤的風險。

小樣本學習工作的困難，在於樣本無法反映出完整母體的特性，為了解決這個問題，使學習過程更為穩定，過去的研究指出虛擬樣本產生法是處理小樣本問題的有效方式 (Abu-Mostafa 1990; Cho et al. 1997; Huang and Moraga 2004; Li et al. 2009a; Wen et al. 2008; Yang et al. 2011) 。然而，虛擬樣本擴展技術在使用上仍然有所侷限，其在人工樣本的產生過程中，並未考量屬性之間的相關性，使得屬性相依的資料並無法直接地適用。其中，時間序列資料就是這樣的一個確例，觀察值的趨勢發展通常與其發生順序有關，而不考慮資料與時間之間連結關係的虛擬樣本，並無法有效地改善模型的學習狀況。

為了解決少量時間序列資料的學習問題，本研究提出了潛在資訊 (latent information, LI) 函數來分析資料，並萃取出隱藏資訊，以協助小樣本資料的知識探索。這個方法能夠依照資料的特性獲得額外的資訊，並且能產生預測結果較佳的神經模型。為了證明方法的有效性，本研究使用虛擬控制圖時間序列資料集 (Synthetic Control Chart Time Series dataset, SCCTS) 進行驗證。實驗結果顯示，潛在資訊函數能夠顯著地改善預測的準確率，是一個處理小樣本資料的有效技巧。

本論文剩餘的部分組織如下：第二節將介紹潛在資訊函數的建構程序。第三節為本研究方法的驗證過程與實驗結果。最後，結論將在第四節呈現。

2. 潛在資訊函數的建構程序

在製造系統的早期階段，我們往往無法擁有足夠的樣本數量，這使得預測模型的學習並無法有效率地執行。在過去的研究中發現，適當地增加訊息量可以克服這個問題，並使我們能夠得到較為穩定的預測結果。此外，時間序列資料的收集可以視為一個連續且漸增的程序，進入的數據將逐步改善並更新我們所獲得信息，資料的進入順序對於模型的建構將會產生不相等的影響。本研究試圖在序列資料的特性下，探索出額外資訊以使知識獲取過程更為穩定。

為了分析潛在數據出現的可能性，本研究定義了 LI 函數。我們將資料的可能範圍進行適度地擴展以填補資料間距，並以目前所擁有的樣本數量決定值域的延展程度。簡單地說，當擁有大量的觀察值時，因資訊量較大使得資料輪廓較為明確，我們並不需要大幅度地擴展值域；反之，較大程度地擴展值域範圍是必須的。因此，我們以全距除以樣本數決定值域的擴展程度，而向左或向右延展的比例則使用偏度指標來確定，延伸後的界限稱為上界 (upper bound, UB) 與下界 (lower bound, LB) ，再結合中心趨勢 (central tendency, CT) 共同建構出 LI 函數。LI 值的範圍界於 0 與 1 之間，它代表資料出現的可能強度。LI 函數的完整建構程序敘述如下：

1. 假設我們擁有 n 期的時間序列資料 $X=\{x_1,x_2,\ldots,x_n\}$。令 $x_{\min}$ 為 X 中擁有最小值的元素，而 $x_{\max}$ 為 X 中擁有最大值的元素，則全距可經由公式 $\mathrm{R}=x_{\max}-x_{\min}$ 算出。
2. 使用加權平均的方式決定中心趨勢 CT。

$$\mathrm{CT}=\frac{\sum_{i=1}^{n} ix_i}{\sum_{i=1}^{n} i}\ ,\ i=1,2,\ldots,n \tag{1}$$

3. 找出現存資料的中心點 CL。

$$\mathrm{CL}=\frac{x_{\min}+x_{\max}}{2} \tag{2}$$

4. 計算資料集合中大於 CL 的元素個數 $\left|X^{+}\right|$ 與小於 CL 的元素個數 $\left|X^{-}\right|$。

5. 確定資料集合的遞增趨勢 (IT) 與遞減趨勢 (DT) 。

$$\begin{cases}\mathrm{IT}=\dfrac{\left|X^{+}\right|}{\left|X^{+}\right|+\left|X^{-}\right|}\\ \mathrm{DT}=\dfrac{\left|X^{-}\right|}{\left|X^{+}\right|+\left|X^{-}\right|}\end{cases} \tag{3}$$

6. 利用 IT 與 DT 非對稱地擴展值域範圍。擴展後的值域上界 (UB) 與下界 (LB) 可由下列的式子得到。

$$\begin{cases}\mathrm{UB}=x_{\max}+\mathrm{IT}\times\dfrac{R}{n}\\ \mathrm{LB}=x_{\min}-\mathrm{DT}\times\dfrac{R}{n}\end{cases} \tag{4}$$

7. 以 CT、UB 與 LB 建構如圖 1 的 LI 函數。我們設定 CT 的 LI 值為 1，UB 與 LB 的 LI 值為 0，則藉由相似三角形的特性，可以得到每一個資料點的 LI 值。

$$\text{LI值}=\begin{cases}\dfrac{x-\mathrm{LB}}{\mathrm{CT}-\mathrm{LB}} & \text{if } x<\mathrm{CT}\\ 1 & \text{if } x=\mathrm{CT}\\ \dfrac{\mathrm{UB}-x}{\mathrm{UB}-\mathrm{CT}} & \text{if } x>\mathrm{CT}\end{cases} \tag{5}$$

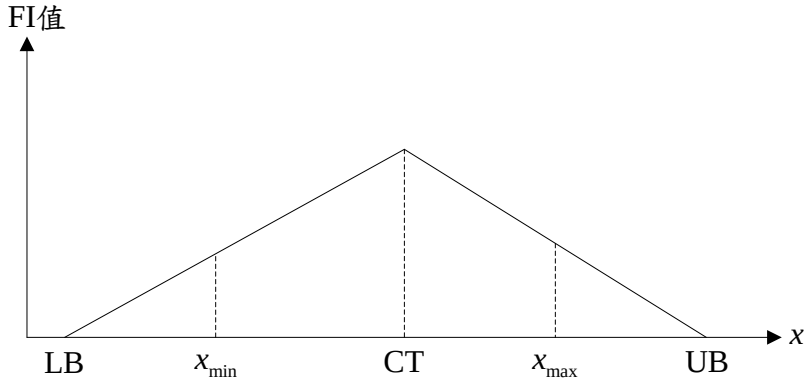

圖 1：潛在資訊函數的形狀

3. 實證分析

工業製程管制的個案將用來呈現潛在資訊函數的執行情況，以驗證 LI 函數於早期製造系統的適用性，實驗的詳細程序將於接下來的子節中描述。

3.1 製程控制資料

為了確認方法在早期製造階段的預測表現，我們由知識發現資料庫 (http://kdd.ics.uci.edu/databases/synthetic_control/synthetic_control.html) 選擇了虛擬控制圖時間序列資料集 (Synthetic Control Chart Time Series Dataset, SCCTS)。這個資料集合包含了 600 筆虛擬控制圖的觀察值，它分成六個不同的種類 (常態、週期、遞增、遞減、跳躍增加、跳躍減少) ，每一個種類有 100 個樣本，每一個樣本則有 60 個時期的資料。在研究中，我們選擇所有 600 筆樣本進行實證分析，部分資料如表 1 所示。

本研究的目的是在有限數據的條件下，萃取潛藏資訊以建立一個適用於小樣本預測的模型。在本文中，我們採用序列最前面五個時期的資料，建立預測模型，再利用這個模型對下一個輸出值進行預測。也就是說，$\{x_1, x_2, \ldots, x_5\}$ 已給定且用於產生預測 x_6 的模型。

3.2 使用原始資料建構 BPN 模型

倒傳遞神經網路 (backpropagation neural network, BPN) 是神經網路中最著名的一種類型 (Witten and Frank 2005) ，因為其使用便利而廣泛地被應用，本研究採用的學習軟體為 Pythia software。我們使用 4 個學習樣本來訓練 BPN，每對樣本包含一個輸入屬性與一個輸出屬性，即 $\{(x_1, x_2), (x_2, x_3), (x_3, x_4), (x_4, x_5)\}$。在有限的訓練資料限制下，我們使用 1-2-1 的 BPN 架構 (如圖 2) 來處理預測的工作，而 BPN 所訓練的模型將被用來推測對應值 $\hat{x}_6$。

我們以資料集的第一個個案為例，訓練集合有五筆觀測值 {28.7812, 34.4632,

31.3381, 31.2834, 28.9207}，將其重整後可以得到四個成對的學習樣本{(28.7812, 34.4632), (34.4632, 31.3381), (31.3381, 31.2834), (31.2834, 28.9207)}；其中，每對資料的第一個值與第二個值分別為訓練 BPN 的輸入值與輸出值。在網路訓練之後，我們輸入 $x_5 = 28.9207$ 以取得輸出值 $\hat{x}_6 = 31.4316$，個案資料如表 2 所示。

3.3 使用潛在資訊建立 BPN 模型

因為原始資料包含了較少的資訊以至於不容易產生穩健的神經網路學習，本研究試圖利用 LI 函數萃取資料的額外資訊以提升 BPN 的學習效能。這裡的學習樣本依舊是 4 個，每組樣本包含二個輸入屬性與一個輸出屬性，即{(x_1, LI_1, x_2), (x_2, LI_2, x_3), (x_3, LI_3, x_4), (x_4, LI_4, x_5)}。網路架構則採用 2-2-1 的 BPN 架構，如圖 3 所示。

在個案中，原始資料集合為{28.7812, 34.4632, 31.3381, 31.2834, 28.9207}，在使用 LI 函數擴充資訊後，可以得到每筆資料相對應的 LI 值，LI = {0.3144, 0.0579, 0.8538, 0.8677, 0.3626}。訓練集合加入 LI 值修正後可得{(28.7812, 0.3144, 34.4632), (34.4632, 0.0579, 31.3381), (31.3381, 0.8538, 31.2834), (31.2834, 0.3626, 28.9207)}，每組樣本的前兩個值為訓練網路的輸入值，而第三個值則為輸出值。在拓撲建立後，我們輸入 $(x_5, LI_5) = (28.9207, 0.3626)$ 以得到預測值 $\hat{x}_6 = 34.0560$，個案資料如表 3 所示。

表 1：虛擬控制圖時間序列資料

	時期						
No.	1	2	3	4	5	…	60
1	28.7812	34.4632	31.3381	31.2834	28.9207		25.8717
2	24.8923	25.7410	27.5532	32.8217	27.8789		26.6910
3	31.3987	30.6316	26.3983	24.2905	27.8613		29.3430
4	25.7740	30.5262	35.4209	25.6033	27.9700		25.3069
5	27.1798	29.2498	33.6928	25.6264	24.6555	…	31.0179
6	25.5067	29.7929	28.0765	34.4812	33.8000		35.4907
7	28.6989	29.2101	30.9291	34.6229	31.4138		26.4637
8	30.9493	34.3170	35.5674	34.8829	30.6691		34.5230
9	35.2538	34.6402	35.7584	28.5510	25.6518		32.3833
⋮				⋮		⋱	⋮
598	35.8990	26.6719	34.1911	35.8270	25.1009		17.4747
599	24.5383	24.2802	28.2814	27.1316	26.6623	…	17.4599
600	34.3354	30.9375	31.9529	31.1457	24.5189		10.1521

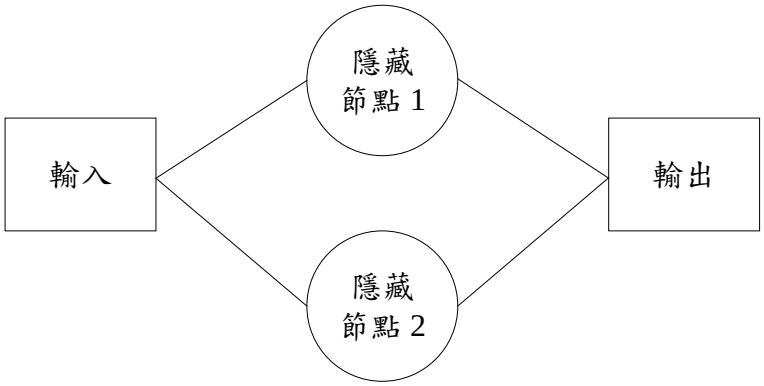

圖 2：1-2-1 的 BPN 架構

表 2：1-2-1 網路架構的學習個案

	時期	輸入	輸出
訓練樣本	1	28.7812(x_1)	34.4632(x_2)
	2	34.4632(x_2)	31.3381(x_3)
	3	31.3381(x_3)	31.2834(x_4)
	4	31.2834(x_4)	28.9207(x_5)
測試樣本	5	28.9207(x_5)	33.7596($\hat{x}_6$)

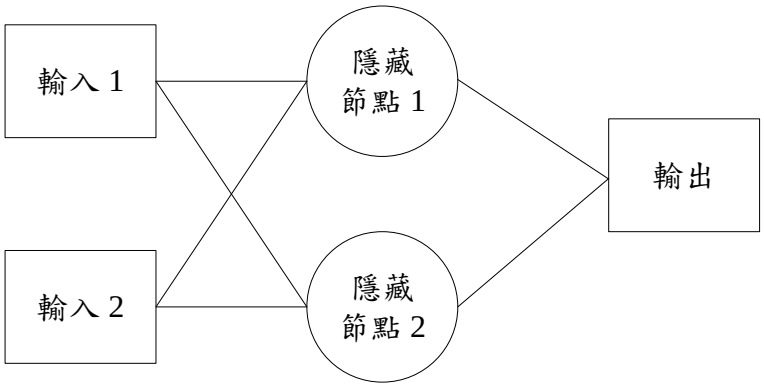

圖 3：2-2-1 的 BPN 架構

表 3：2-2-1 網路架構的學習個案

	時期	輸入 1	輸入 2	輸出
訓練樣本	1	28.7812(x_1)	0.3144(LI_1)	34.4632(x_2)
	2	34.4632(x_2)	0.0579(LI_2)	31.3381(x_3)
	3	31.3381(x_3)	0.8538(LI_3)	31.2834(x_4)
	4	31.2834(x_4)	0.8677(LI_4)	28.9207(x_5)
測試樣本	5	28.9207(x_5)	0.3626(LI_5)	33.7596($\hat{x}_6$)

3.4 結果比較

正確率是衡量預測方法最重要的指標 (Mahmoud et al. 1988; Yokum and Armstrong 1995)。本研究以 SCCTS 中不同類別的時間序列進行預測比較，為了檢測模型的效能，我們採用平均絕對百分誤差 (mean absolute percentage error, MAPE) 做為評估指標。

表 4 顯示本研究所提出的方法 (結合 LI 值的 BPN 架構) 擁有較佳的預測表現，其對 MAPE 的平均改善效果能達到 12%以上。顯示 LI 函數可以改善原先訊息量不充份的缺點，並透過反應資料特性而帶來較佳的預測結果。對不同資料型態而言，除了週期資料外，改善後的 BPN 架構之預測效能明顯優於原始的 BPN 架構。也就是說，本研究的方法適用於樣本數有限的常態與趨勢資料。因此，它是一個製造系統早期階段的適當預測工具。

4. 結論

企業為了要能有效地控制經營成本，並讓組織的整體規劃更有效率，擁有適當的預測技術是必須的。尤其，在製造系統的早期階段，因為成本與時間的考量，普遍僅能取得不充足的資料，如何藉由小量樣本取得精確的預測結果，是一個相當重要的議題。本研究企圖在這樣的限制下，找出一個可以改善預測準確率的方法。

我們在本研究中提出了一個新的資料分析程序，以克服發生在早期階段不完整資料與不充足資訊的問題。本研究所發展的 LI 函數，被用以萃取預測序列資料所需的資訊，其能夠考量資料的發生順序，並有系統地找出資料的額外訊息，以增進神經網路的學習效率。實驗結果清楚地顯示 LI 函數可以幫助提升預測表現，其訓練結構能在製造系統的早期階段學習知識，具有一定的實用價值的。在未來，LI 函數應可以使用到真實個案中，以強化其實務應用的價值。

表 4：LI 函數的改善表現

資料型態	原始架構的 BPN	結合 LI 值的 BPN	改善幅度 (%)
常　　態	12.1201	10.5176	13.22
週　　期	18.2104	17.5713	3.51
遞　　增	12.2547	10.5869	13.61
遞　　減	13.6885	11.2704	17.67
跳躍增加	13.8541	12.2666	11.46
跳躍減少	11.8284	9.5600	19.18
平　　均	13.6594	11.9621	12.43

參考文獻

1. Abu-Mostafa, Y.S. "Learning from hints in neural networks" Journal of Complexity (6:2) 1990, pp:192-198.
2. Cho, S.Z., Jang, M. and Chang, S.J. "Virtual sample generation using a population of networks" Neural Processing Letters (5:2) 1997, pp:83-89.
3. Huang, C.F. and Moraga, C. "A diffusion-neural-network for learning from small samples" International Journal of Approximate Reasoning (35:2) 2004, pp:137-161.
4. Li, D.C., Chang, C.J., Chen, C.C. and Chen, C.S. "Using non-equigap grey model for small data set forecasting - a color filter manufacturing example" Journal of Grey System (22:4) 2010, pp:375-382.
5. Li, D.C., Chang, C.J., Chen, W.C. and Chen, C.C. "An extended grey forecasting model for omnidirectional forecasting considering data gap difference" Applied Mathematical Modelling (35:10) 2011, pp:5051-5058.
6. Li, D.C., Fang, Y.H., Lai, Y.Y. and Hu, S.C. "Utilization of virtual samples to facilitate cancer identification for DNA microarray data in the early stages of an investigation" Information Sciences (179:16) 2009a, pp:2740-2753.
7. Li, D.C., Yeh, C.W. and Chang, C.J. "An improved grey-based approach for early manufacturing data forecasting" Computers & Industrial Engineering (57:4) 2009b, pp:1161-1167.
8. Mahmoud, E., Rice, G. and Malhotra, N. "Emerging issues in sales forecasting and decision support systems" Journal of the Academy of Marketing Science (16:3-4) 1988, pp:47-61.
9. Wen, Q.Y., Zhang, H.W., Yang, Q.H. and Zhang, P.X. "A virtual-sample technology based artificial-neural-network for a complex data analysis in a glass-ceramic system" Journal of Ceramic Processing Research (9:4) 2008, pp:393-397.
10. Witten, I.H. and Frank, E. "Data mining: practical machine learning tools and techniques" San Francisco, United States, Morgan Kaufmann Publishers, 2005.
11. Yang, J., Yu, X., Xie, Z.Q. and Zhang, J.P. "A novel virtual sample generation method based on Gaussian distribution" Knowledge-Based Systems (24:6) 2011, pp:740-748.
12. Yokum, J.T. and Armstrong, J.S. "Beyond accuracy: Comparison of criteria used to select forecasting methods" International Journal of Forecasting (11:4) 1995, pp:591-597.

Using latent information to improve learning accuracies of small samples

Der-Chiang Li[1] Che-Jung Chang[2] Liang-Sian Lin[3]
[1]Department of Industrial and Information Management, National Chen Kung University
lidc@mail.ncku.edu.tw
[2]Department of Industrial and Information Management, National Chen Kung University
r3795102@nckualumni.org.tw
[3]Department of Industrial and Information Management, National Chen Kung University
r38991052@mail.ncku.edu.tw

Abstract

In the current highly competitive manufacturing environment, controlling manufacturing systems effectively and efficiently is a necessary competence to maintain competitive advantages, especially in their early stages, when the sample sizes are usually very small. However, to determine proper manufacturing parameters using limited data is a challenge for engineers and managers. The virtual-sample-generating technique is an effective approach for dealing with uncertain and insufficient information, but it has a weakness as it usually can not process times series data. This research thus develops a Latent Information function to analyze data features and extract hidden information for knowledge learning with small data sets considering timing factors. The experimental results using the Synthetic Control Chart Time Series Datasets show that the proposed method can improve forecasting accuracies significantly, and thus is considered an appropriate procedure in general to forecast manufacturing outputs based on small samples.

Keywords: Small-data-set learning; Forecasting; Manufacturing; Time series