利用遮罩法提升自組映射圖訓練效率

Improve Training Efficiency of Self-Organizing Maps by Canopy Technique

陳慶隆　　蘇育葳　　許中川

國立雲林科技大學資管所　國立雲林科技大學資管所　國立雲林科技大學資管所

gmi111@mis4k.mis.yuntech.edu.tw gmi002@mis4k.mis.yuntech.edu.tw hsucc@mis.yuntech.edu.tw

摘要

在資料探勘的演算法中，由 Kohonen 所提出的非監督式類神經訓練演算法：自組映射圖，可以將高維度的資料投射到低維度的空間上，同時保留資料在高維空間的相對關係，因此在降低資料複雜度方面有許多應用。特別是投射到二維平面空間時，可以成為一種良好的視覺化互動式分群技術。然而，自組映射圖在處理大量資料時，卻有著沉重的運算效率問題。本研究提出應用遮罩法技術的自組映射圖，透過限制訓練階段尋找最適配神經元的範圍，在不降低分群精確度前提下，提升自組映射圖的訓練速度。本研究分別針對不同遮罩啟動門檻、不同維度、資料筆數、神經元個數及不同資料集進行提升速度之實驗，其結果證實速度提升率與資料維度呈線性正相關、與資料筆數、神經元個數無顯著差異，且能在不影響分群效果的條件下提升效率。

關鍵詞：資料探勘、自組映射圖、遮罩法技術。

Abstract

The self-organizing map is a well-known algorithm in data mining purposed by Kohonen. It provides a good visualizing mechanism to project multidimensional data to low-dimensional space. However, there is an efficiency problem on training the SOM. In this paper, we propose a method which integrates a canopy technique to the training algorithm of self-organizing maps for improving the training efficiency without decreasing clustering performance. We conducted several experiments to test the proposed method, including different setups of the canopy threshold、data dimensionality、training data quantity and the size of SOM neurons. The result confirms that the performance of training efficiency can be improved with positive correlation of data dimensionality but there is no correlation with data quantity and the size of neurons.

Keyword: data mining、Self-Organizing Maps、canopy technique

壹、緒論

資料探勘(Data Mining)**是指從大量資料中萃取隱藏、未知與潛在且具有實用性的資訊處理過程**[Fayyad, 1996]**。它被應用在許多的領域，如：市場行銷**[Piatetsk-Shapiro, 1996]**、生物科技蛋白質二級結構預測**[Rost, 1997]**、股票市場分析**[Gavrilov, 2000]**等。透過資料探勘技術可有效的萃取潛藏在大量資料中的資訊或知識，而這是一般的資料分析技術所不能達到的。對於現今資料量爆炸的時代，資料探勘是非常重要且不可或缺的資料分析技術。**

在資料探勘的分群(clustering)**技術中，自組映射圖**(self-organizing maps, SOM)**是一項重要的工具。它是由** Kohonen **所發表的一種非監督式類神經學習演算法**[Kohonen, 1982]**。運用屬性圖**(feature map)**及投射技術，自組映射圖可以將高維度資料以低維度的方式呈現，並且保有原始資料的拓樸特性。由於自組映射圖可以低維度方式呈現資料特徵，因此提供了一個很好的視覺化機制，這也讓它成為目前重要且被廣泛應用的技術，例如：網路入侵偵測**[施東河與黃于爵, 民 92]**、**DNA **晶片設計**[Douzono, 2001]**、採購管理**[Davis, 2001]**、財務分析**[Deboeck, 1998]**與化學應用**[Tokutaka, 1998]**等。**

然而，自組映射圖卻有訓練時間效率的問題，其主要的問題在於每一個訓練回合的每一筆訓練資料尋找最適神經元(best matching unit)**的過程中，需要和自組映射圖上的所有神經元進行距離(相似度)比對，其訓練時間複雜**

度為 O(t*n*m*f)，其中 t 為訓練回合數，n 為訓練資料筆數，m 為自組映射圖上神經元個數、f 為訓練資料的維度。因此，若過多的資料量或神經元個數，將會造成自組映射圖在訓練計算上沉重的負擔，進而影響自組映射圖於巨量資料的應用。

基於上述的訓練效率問題，本研究提出結合遮罩法技術(canopy technique)[McCallum, 2000]與自組映射圖訓練演算法，透過限制訓練時期尋找最適配神經元的搜尋範圍，在不降低分群準確度前提下，提升自組映射圖的訓練速度。經一系列實驗證明，本論文所提出的方法，確能提升自組映射圖的訓練速度。

本篇論文共包含五個章節，第一章為緒論，說明研究背景及研究動機與目的。第二章為文獻探討，介紹自組映射圖基本理論、改善自組映射圖效率的相關研究與遮罩法技術。第三章為研究方法。第四章為實驗，實際開發一雛型系統以驗證本研究架構。最後一章為結論。

貳、文獻探討

本章將針對與本研究相關的議題進行描述，包含自組映射圖、改善自組映射圖速度研究與遮罩法技術等相關研究。

2-1 自組映射圖

自組映射圖是 Kohonen 所提出的非監督式類神經網路架構，是一個被廣泛應用的分群演算法[Kohonen, 1982]。自組映射圖能將多維的資料投射(project)至二維的平面，且仍可保有原始資料特徵。它依據計算資料的相似程度，在二維平面上形成數個大小不一的群聚。在學習的過程中，神經元會與所有的輸入資料向量進行相似度計算，最接近輸入向量的神經元稱為最適配神經元(best matching unit)，而此神經元會調整至更接近輸入向量，同時也會對最適配神經元所有鄰近的點進行調整，使得分群中的相似資料更加接近，如此即可在二維神經元平面形成數個具有高度相似之分群。

因此，自組映射圖可分為二個步驟，第一步為計算最適配神經元，如式(1)，其中 c 為二維向量中的每個神經元、m_c 為神經元的模式向量；第二步為調整最適配神經元鄰近的神經元，如式(2)，其中 t 表任一時間、$a(t)$ 為一介於 0 與 1 之間遞減的學習率(learning rate)，式(3)為調整範圍函式，r_b 與 r_c 是指神經元 b 與 c 在自組映射圖神經元方格的位置、$s(t)$ 是以最適配神經元為中心的半徑，隨著時間的增加而遞減。而自組映射圖最主要處理效率的問題，就在於尋找最適配神經元的過程。

$$\|x - m_b\| = \min_c \{\|x - m_c\|\} \quad (1)$$

$$m_c(t+1) = m_c(t) + a(t)h_{bc}(t)[x(t) - m_c(t)] \quad (2)$$

$$h_{bc}(t) = \exp\left(-\frac{\|r_b - r_c\|^2}{2s^2(t)}\right) \quad (3)$$

2-2 改善自組映射圖速度研究

結合基因演算法於訓練自組映射圖：這是由 Huang 與 Hung 提出的方法，主要的精神是利用基因演算法決定自組映射圖神經元的初始值，透過這樣的方式可以避免神經元初始值，因為太過隨機，而造成沉重的神經元微調計算。經過該研究的實驗證明，利用基因演算法決定神經元初始值，雖然需要額外的時間，不過，整體的運算時間確是顯著下降的，大約提升 31.3%[Huang and Hung, 1995]。

結合 K-means 之 SOM：Su 與 Chang 提出透過 K-means 演算法來決定自組映射圖神經元初始值與個數[Su and Chang, 2000]。共分成三階段，第一階段為透過 K-means 來決定 N^2 個中心點，N^2 為自組映射圖神經元個數；第二階段將 K-means 所決定的 N^2 個中心，適當的分配至自組映射圖上，以形成神經元初始值；最後，視神經元的狀況，判斷是否要進行微調動作。運用此方法，將大符減少自組映射圖神經元調整的動作。並比傳統的降低(N+1)/2 倍時間。

調整資料順序之 SOM：是由 Miyoshi、Kawai 與 Masuyama 所提出的方法[Miyoshi et al., 2002]。其主要的精神為，先將資料依照其對映類別作距離計算，排出資料順序，然後將排序好的資料依序輸入 SOM。透過這樣的方式，可以避免神經元的值大幅度跳動，以節省訓練時間。經過實驗證明較傳統自組映射圖提升 9%的速度。

針對上述改善自組映射圖訓練時間問題的研究，我們可以歸納其研究方法之原理：前兩種方法都是透過神經元的初始值設定；避免神經元的初始值太過隨機，而造成後續沉重的神經元微調處理。第三種方法是從另一個角度：妥善安排訓練資料的輸入順序。本篇論文，我們提出另一種思考方向，改從縮減最適神經元範圍著手。我們這些方法的運用並不會互相排斥，反而可以結合使用，形成互補，進一步提升自組映射圖的訓練效率。

2-3 遮罩法技術

遮罩式技術(canopy technique)是一種在資料具有大量的資料筆數、資料中包含大量的屬性，以及資料隱藏大量分群的特性時，仍可有效率進行分群技術的方法[McCallum, 2000]。如圖一所示，它是一種兩段式的分群方法，首先使用較不費時(cheap)及粗糙(rough)的快速量測方式，將距離小於門檻值(T1)的資料放在同一集合，叫作遮罩(canopy)；接下來，以每一個遮罩為單元，以精準的量測方式，如K-means[MacQueen, 1967]或最大期望值(expectation-maximization)[Lauritzen, 1995]等技術，以另一門檻值(T2)進行同一遮罩的分群運算。透過遮罩法的技術，將可避免對全部資料進行運算，如此，即可解決一般分群技術的時間複雜度問題，並且經過實驗證明確實可行[McCallum, 2000]。

遮罩分群法是一個可以大幅減少時間複雜度的方法。以 k-means 或最大期望值為例，在未使用遮罩法之前，建立分群的流程需 O(nk) 的複雜度，n 為資料筆數，k 為群數；而在加入遮罩法後，即降低為 O(fn/c)，f 為每一個遮罩平均所涵蓋的資料筆數，c 為遮罩數，而在一般情況下 c>f [McCallum, 2000]。

本研究將應用遮罩法的精神於自組映射圖，其目的在於降低尋找最適配神經元所需的大量時間複雜度。此外，遮罩法技術也可以應用於資料雜訊或離群值的偵測。

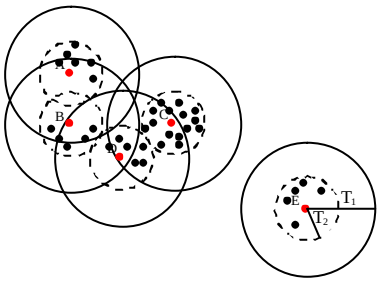


圖 1 遮罩法技術

參、研究方法

3.1 結合遮罩法之自組映射圖

本研究在改善自組映射圖效率的方法上，係將遮罩法理念應用於自組映射圖搜尋最適神經元來完成。如文獻探討所述，在同一個遮罩內的資料是指具有一定的相似程度的資料集合，因此，只要對遮罩內的資料進行運算，即可獲得分群結果。本研究以相同的精神來提升自組映射圖速度，透過只搜尋遮罩區域的神經元來尋找最適配神經元，如此即可避免比對所有神經元進行相似度運算。

本研究採用遮罩法技術以提升運算時間的主要原理為，類神經網路學習過程可分成兩階段，分別是剛開始進行學習的粗調階段，以「排列階段」(ordering phase)表示；以及後段的細部調整神經元，以「收斂階段」(converging phase)表示。因此，我們利用「收斂階段」的特性，在學習後段，將資料尋找最適配神經元的範圍，侷限於訓練資料上回合之最適配神經元的有效半徑內，以縮減運算範圍，如圖二所示。

關於「排列階段」與「收斂階段」的切換判斷條件，我們提出採用平均誤差。所謂平均誤差是指在每一訓練回合(training epoch)，每筆訓練資料與其最適配神經元的平均距離，計算公式如(4)：

$$E_{avg} = \left(\sum_{i=1}^{X}\sum_{j=1}^{DIM}\left\|x_{i,j} - mb_{i,j}(t)\right\|\right) \Big/ \left((X \times DIM)/2\right) \quad (4)$$

式中，X 為訓練資料筆數，DIM 為訓練資料維度。$x_{i,j}$ 為訓練資料 i 的第 j 個維度，$mb_{i,j}$ 為訓練資料 i 的最適神經元的第 j 個元素。式子右邊分子部分為所有訓練資料與其最適神經元距離總合，除以分母部分為計算平均值及進行誤差正規化。在每個訓練資料維度的屬性值及自組映射圖神經元元素值都正規化到零與一之間的情況下，所有訓練資料和其最適神經元之距離總合最大值為(X*DIM)。此極端的最大值發生在當最適神經元的每個元素值和訓練資料的每個維度差值為一。式 (4)右邊分母部分為此極大值的二分之一。由於自組映射圖神經元的起始值為隨機設定，因此 E_{avg} 通常會介於零與一之間。

在自組映射圖的訓練過程中，平均誤差會隨著訓練回合的增加而遞減，這意謂著神經元的訓練已愈來愈成熟，且資料的分佈及群組特性已在自組映射圖的神經元中成形。因此，若 E_{avg} 的值小於遮罩啟動門檻值 s，$0 \le s \le 1$，則訓練資料尋找最適配神經元的範圍，就只須侷限在遮罩內搜尋，而不再搜尋整個自組映射圖。此遮罩範圍的中心點為該筆資料上次的最適神經元，因此此回合的搜尋範圍為與訓練資料上回合最適配神經元的座標距離，小於遮罩半徑(CR(t))即可，如圖二所示。

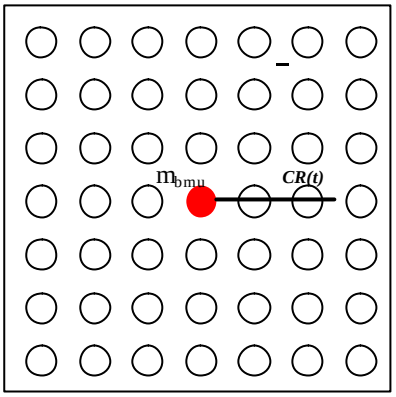


圖 2 當平均誤差小於門檻值後，於遮罩範圍內尋找訓練資料的最適配神經元

搜尋範圍及最適神經元如式(5)、式(6) 所示，

$$Search(m_{bmu}(t), CR(t)) = \begin{cases} \{\|r(m_{bmu}(t)) - r(m_c)\| < CR(t),\ c = 1,2,...n\}, & E_{avg} < s \\ \{m_c, c = 1,2,...,\ n\}, & \text{else} \end{cases} \quad (5)$$

$$m_{bmu}(t+1) = \min_c \{\|x - m_c\| \mid c \in Search(m_{bmu}(t), CR(t))\} \quad (6)$$

$$CR(t) = NR(t) \quad (7)$$

t 為時間以訓練回合表示，n 為自組映射圖神經元個數，CR(t)表示遮罩半徑。式(5)定義搜尋範圍為當平均誤差小於門檻時，搜尋範圍限於以第 t 回合的最適神經元為中心的遮罩範圍內，否則為整個映射圖。式(6)定義第 t+1 回合訓練資料 x 的最適神經元為在搜尋範圍內，與訓練資料 x 距離最近的神經元 m_c。

式(7)遮罩半徑 CR(t)是一個隨時間增加而遞減的函數，使得搜尋最適神經元的搜尋範圍隨時間而縮減。本研究定義此搜尋範圍半徑為自組映射圖的鄰近範圍半徑(neighborhood radius, NR(t))，鄰近範圍為調整神經元的有效範圍。我們如此設定的主要理由是，在此鄰近範圍內的神經元因獲得調整，而其與訓練資料的相似程度，應比此範圍外的神經元的相似程度還要高。因此，下一次最適配神經元出現在此鄰近距離的機率也較高。

肆、實驗

在實驗方面，本研究共採用三個資料庫，分別為 UCI Machine Learning 所提供的 Animal 資料集，其中包含 15 個布林型態、1 個數值型態與 1 個類別欄位，共 17 個欄位，101 筆資料，7 類動物種類。此資料集是 Ritter 與 Kohonen 在描述自組映射圖處理高維度資料時，所採用的資料集[Ritter and Kohonen, 1989]。另外兩個資料庫為人工資料集，分別具有 16 個數值型及一個類別欄位，資料同樣為 101 筆，透過類別欄位分為四種類別。第一個人工資料內容產生方式為常態分配(synthetic_normal)，分別為四種類別資料的各個欄位設定平均數及標準差，然後依常態分配產生屬性值。第二個人工資料產生方式為一般亂數分配(synthetic_random)，分別為四種類別資料的各個欄位設定屬性值範圍，然後亂數產生屬性值。

本研究分別對這三個資料集，進行不同遮罩啟動門檻值 s、不同資料維度、不同資料筆數以及自組映射圖不同神經元個數等進行實驗，觀察對提升自組映射圖訓練效率之影響。

4.1 CPU 優先權影響

本實驗的目的在於觀察不同的 CPU 優先權對處理資料的速度的影響，以排除電腦系統內常駐程式對實驗的影響，讓後續的實驗更為精確。我們針對相同資料在 CPU 不同優先權下進行速度提升的實驗，實驗資料採用人工資料集，共 170 筆，在實驗參數設定上，神經元為 64、學習率初始值為 0.9、鄰近距離初始值為 1，鄰近遞減函數 $NR(t) = NR(0) \times 1/2t$，學習率遞減函數 $a = a_0 \times e^{-t}$，在實驗過程中，我們關閉系統以外不必要的軟體，並拔除網路線，然後才調整 CPU 優先權，以求最精確的數據，其實驗結果如圖 3。

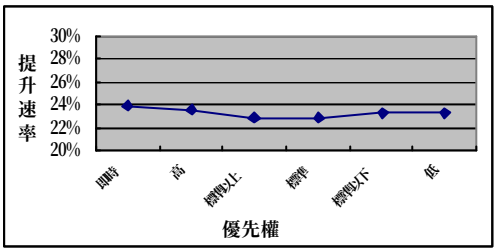


圖 3 CPU 優先權與速度關係圖

如圖 3 所示，CPU 在優先權「即時」的時候具有較佳的提升速率，且經由多次實驗證實在「即時」的優先權下具有較穩定的執行效能，且平均執行的誤差時間可控制在約 0.5 秒之內，故本研究後續實驗將設定 CPU 優先權為「即時」，以求較精確之數值。

4.2 啟動門檻值實驗

本實驗的目的在於，在不同啟動門檻值 s 的參數下，對分群效果與提升效率的影響性。s 參數值的設定由零至一，每次漸增 0.1。s 參數值為零的時候，永遠不會啟動遮罩，相當於搜尋整個映射圖的傳統訓練演算法。其他實驗參數設定上，神經元為 400 個、學習率初始值為 0.9、鄰近距離初始值為 1，鄰近遞減函數 $NR(t) = NR(0) \times 1/2t$，學習率遞減函數 $a = a_0 \times e^{-t}$，訓練停止條件為學習率低於 0.001 與鄰近函數低於 0.01 為止。

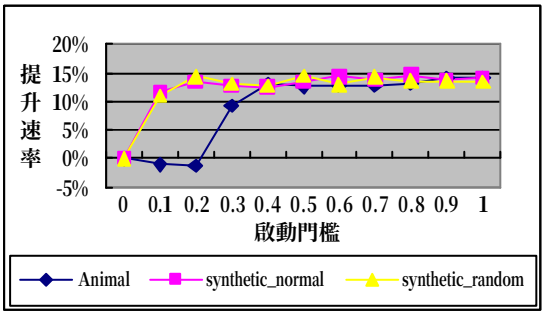
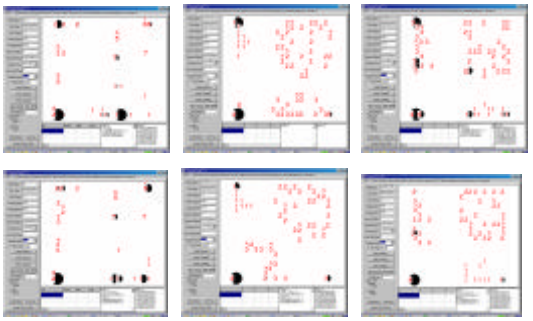


圖 4 不同啟動遮罩門檻與提升映射圖訓練速度之關係

表 1 實驗資料每一訓練回合的平均誤差

回合	Animal	Synthetic normal	Synthetic random
一	0.618613	0.496432	0.489857
二	0.363363	0.127905	0.108113
三	0.27034	0.029544	0.030049
四	0.224219	0.024746	0.026965
五	0.205939	0.023572	0.025846
六	0.200016	0.023164	0.025469
七	0.197885	0.023016	0.025330



(a) animal data set (b) synthetic_normal data set (c) synthetic_random data set

圖 5 上半部及下半部分別為未使用及使用遮罩，對自組映射圖訓練的結果

實驗結果如圖 4 所示，基本上啟動門檻 s 參數值設定愈高，則提升的速度就愈多。因為參數值越大，就越早啟動遮罩，縮減搜尋範圍。但圖 4 顯示，提升到達一定的程度後就停止，其主要原因可由表 1 各資料集在每一訓練回合的平均誤差得知，由於遮罩是在平均誤差小於啟動門檻值時才會啟動，所以只要設定的啟動門檻值 s 大於第一回合的平均誤差時，則都會在第二回合啟動遮罩，因此其提升的速度都會相同。例如 Animal 最大的誤差為 0.6186，因此啟動門檻大於 0.7 以上(含)的 s 均有相同的提升效率，且速度提升率和遮罩啟動回合有直接關係，如 Animal 資料集門檻值 0.2~0.3 的速度提升約 10%，但 0.6~0.7 提升率卻只有 1%~2%，其主要原因為門檻值在 0.2 時只有第七回合才滿足啟動條件，但門檻值在 0.3 時卻有第三到第七回合都滿足啟動條件，而門檻值 0.6~0.7 只增加第一個回合滿足啟動條件，所以速度提升率才會如圖 4 所示。另外人工資料因為是規則產生的，所以平均誤差才會遞減那麼快，第二回合就降到 0.2 以下，造成速度提升在門檻值 0.2 之後就沒什麼差別。

本實驗在遮罩啟動之後，自組映射圖所呈現的分群結果如圖 5 所示，上半部分別為 Animal、normal、random 三個資料集未啟動遮罩時的自組映射圖，下半部為完全啟動遮罩，即遮罩門檻值 s 為 1 時的自組映射圖，也就是

第二回合開始就使用遮罩縮減搜尋範圍，由圖中我們可看出遮罩啟動前後雖然分群的位置有些許不同，但整體的群組類別還是沒有太大改變，由此可證明本研究確實能在不影響分群效果之下，提升自組映射圖的效率。

4.3 資料維度實驗

本實驗的目的在於，測試資料維度與速度提升的關係，我們分別對三個資料集進行維度 10、13、16、19、22、25、28 及 32 的速度提升測試。由於原始資料有 16 個維度，所以我們利用刪除欄位與複製欄位的方式來產生不同維度個數的資料，複製、刪除的欄位為隨機抽取。除資料集維度不同外，實驗參數設定均與上一個遮罩門檻值實驗的參數設定相同。我們採用未啟動遮罩與完全啟動遮罩(第二回合即使用)，所需的執行時間來計算速度提升率。

實驗結果如圖 6 所示，由圖中我們得知本研究所採用的遮罩加速方式，在不同的資料集皆具有相同的特性，結果顯示速度提升率與資料維度成正相關的線性成長，資料維度愈大，速度率越大，且自組映射圖所呈現的分佈狀態也無太大改變，如圖 7 所示，由於版面關係，我們只放置 Animal、normal 及 random 未啟動遮罩及完全啟動遮罩時，所產生的自組映射圖。

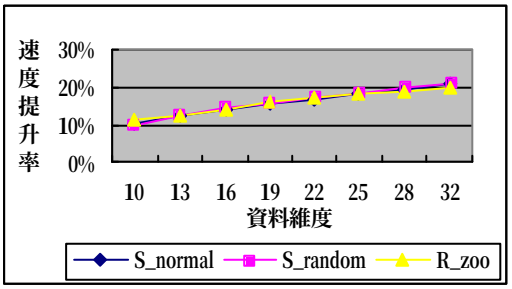
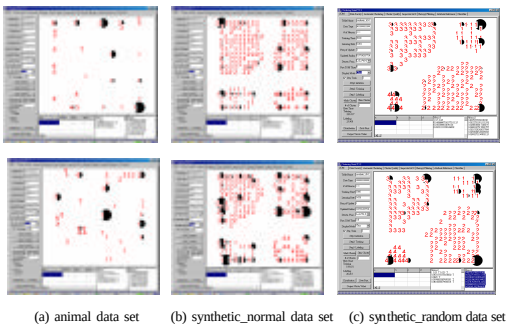


圖 6 資料維度與速度提升率關係圖



(a) animal data set (b) synthetic_normal data set (c) synthetic_random data set

圖 7 變動維度實驗中，上半部及下半部分別為未使用及使用遮罩，對自組映射圖訓練的結果

4.4 資料筆數實驗

本實驗的目的在於，測量資料筆數與速度提升率的關係，我們同樣對三個資料集進行不同資料筆數的實驗，資料筆數採用複製的方式，分別依倍數產生 101、505、1010、1515、2020、2525、3030 筆資料，再隨機打亂資料順序進行實驗。測量參數除資料庫筆數不同外，其餘皆與上述實驗相同。

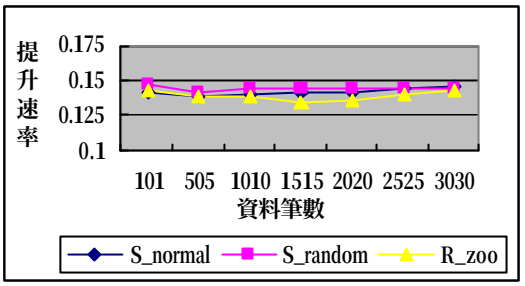


圖 8 資料筆數與速度關係圖

實驗結果如圖 8 所示，採用遮罩法提升自組映射圖的速率，並不會受資料筆數所影響，在不同的筆數中速度的提升率並無顯著不同，不同資料集亦是如此，且自組映射圖所呈現的分群結果也無太大差異，但因版面關係我們不再放置分佈圖。

4.5 神經元個數實驗

本實驗的目的在於，測量不同神經元個數與訓練速度提升的關係，我們分別針對神經元個數 25、100、625、2500、5625、10000 進行實驗，測量參數除神經元個數不同外，其餘皆與上述實驗相同。

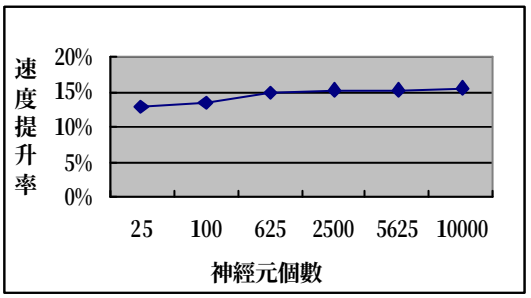


圖 9 神經元個數與訓練速度提升率之關係

實驗結果如圖 9 所示，自組映射圖所提升的速度，在神經元個數較少時會有所差異，但是大於 625 個神經元之後就不再提升，其主要原因是在神經元個數很少時，自組映射圖無法有效將資料分群，但神經元到達一定數量之後即可完整分群，因此神經元個數對速度提升是不影響的。

伍、結論

非監督式類神經網路自組映射圖將高維度資料投射至低維度空間，同時保留資料在高維度時的相對關係，因此應用極為廣泛，其中一個應用為視覺化交互式資料分群。傳統自組映射圖訓練演算法，在搜尋每筆訓練資料的最適神經元時，搜尋範圍為整個自組映射圖。本論文中，我們提出應用遮罩技術，縮減最適神經元的搜尋範圍，以降低自組映射圖的訓練時間。經本研究的實驗證明，確實能在不影響分群品質的條件下，提升訓練自組映射圖的速度。實驗結果顯示，使用遮罩技術時，訓練資料筆數和速度提升率沒有關係。然而，訓練資料維度與速度提升率呈線性正相關的成長，資料維度愈大，速度提升率愈大。以本研究的實驗為例，資料維度從 10 增加到 32 時，速度提升率從 10%提升到 21%。另外，當自組映射圖神經元個數增加時，訓練速度提升率會隨著略為增加，當到達相當的數量之後，速度提升率即趨為穩定，不再隨之提升。

陸、誌謝

能完成本篇論文首先要感謝許中川老師辛勤的指導，育威學長的細心傳授技巧，勝玄學弟的幫忙跑實驗，及實驗室夥伴和家人的支持，讓我能順利完成此論文。

柒、參考文獻

[1] 施東河、黃于爵,"網站入侵偵測系統之分析與研究", 中華民國資訊管理學報,第九卷．第二期：183~214 頁, 2003.

[2] B. Rost and S. O' Donohue "Sisyphus and protein structure prediction," BioInformatics(13), pp:345-356, 1997.

[3] Davis, R.G. and J. Si "Knowledge discovery from supplier change control data for purchasing management," Info-tech and Info-net, Proceedings. ICII 2001 Beijing. 2001 International Conferences on(3),pp:67-72, 2001.

[4] Deboeck, G. and T. Kohonen "Visual Explorations in Finance using self-Organizing Maps," Springer, London, 1998

[5] Douzono, H., S. Hara and Y. Noguchi "A design method of DNA chips for SNP analysis using self-organizing maps," Neural Networks, Proceedings. IJCNN '01. International Joint Conference on(4), pp: 2467-2471, 2001.

[6] Fayyad , U. and R. Uthurusammy "Data mining and knowledge discovery in databases," Communications of the ACM(1:39), pp:24-26, 1996.

[7] G. Piatetsky-Shapiro, R. Brachman, T. Khabaza, W. Kloesgen, and E. Simoudis "An overview of issues in developing industrial data mining and knowledge discovery application," in Proceedings, Second International Conference on Knowledge Discovery and Data Mining, AAAI Press:Menlo Park, CA, 1996

[8] Gavrilov, M., D. Anguelov, P. Indyk and R. Motwani "Mining the Stock Market:Which Measure Is Best?," Proceedings of International Conference on Knowledge Discovery and Data Mining, pp:487-496, 2000.

[9] Huang, S. J. and C. C. Hung "Genetic algorithms enhanced Kohonen's neural network," In Proc. IEEE Int. Conf. Neural Networks, pp:708-712, 1995.

[10] Kohonen, T. "Self-organized formation of topologically correct feature maps," Biological Cybernetics(43), pp:59-69, 1982.

[11] Kohonen, T. "Self-Organizing Maps," Springer-Verlag: Berlin, 1997.

[12] Lauritzen, S. L. "The EM algorithm for graphical association models with missing data," Computational Statistics and Data Analysis(19), pp:191-201, 1995.

[13] MacQueen, J. "Some methods for classification and analysis of multivariate observation," Proc. 5th Berkeley Symp. Math. Statist, Prob.(1), pp281-297, 1967

[14] McCallum, A., K. Nigam, and L. Ungar. "Efficient clustering of high-dimensional data sets with application to reference matching." In KDD, pp:169-178, 2000

[15] Miyoshi, T., H. Kawai and H. Masuyama "Efficient SOM Learning by Data Order Adjustment," Neural Networks, IJCNN '02. Proceedings of the 2002 International Joint Conference(1), pp:784-787, 2002.

[16] Ritter, H.J. and T. Kohonen "Self-Organizing semantic maps," Biolog. Cybern.(61), pp:241-254, 1989.

[17] Su, Mu-Chun and Hsiao-Te Chang "Fast Self-Organizing Feature Map Algorithm," IEEE Transaction on Neural Network(11:3), 2000.

[18] Tokutaka, H, K. Yoshihara, K. Fujimura, K. Iwamoto, T. Watanabe, and S. Kishida "Applications of self-organizing maps to the composition determination of chemical products," Neural Networks Proceedings. IEEE World Congress on Computational Intelligence.(1) pp:4-8, 1998