

行政院國家科學委員會專題研究計畫 成果報告

以查詢為基礎之資料採擷技術

計畫類別：個別型計畫

計畫編號：NSC92-2416-H-002-051-

執行期間：92年08月01日至93年07月31日

執行單位：國立臺灣大學工程科學及海洋工程學系暨研究所

計畫主持人：張瑞益

共同主持人：陳彥良

計畫參與人員：賴良賓、羅嘉彥、徐嘉鴻、林岳進、吳怡靜、林宗岳、邱莉媛

報告類型：精簡報告

處理方式：本計畫可公開查詢

中 華 民 國 93 年 10 月 13 日

# 行政院國家科學委員會補助專題研究計畫成果報告

## 以查詢為基礎之資料採擷技術

計畫類別： 個別型計畫            整合型計畫

計畫編號：NSC 92-2416-H-002-051-

執行期間： 92 年 8 月 1 日至 93 年 7 月 31 日

計畫主持人：張瑞益      國立台灣大學工程科學系

共同主持人：陳彥良      國立中央大學資訊管理系

本成果報告包括以下應繳交之附件：

赴國外出差或研習心得報告一份

赴大陸地區出差或研習心得報告一份

出席國際學術會議心得報告及發表之論文各一份

國際合作研究計畫國外研究報告書一份

執行單位：國立台灣大學工程科學系

中 華 民 國      93 年      10 月      12 日

# 行政院國家科學委員會專題研究計畫成果報告

## 以查詢為基礎之資料採擷技術

### Data Mining by Query-Based Techniques

計畫編號：NSC 92-2416-H-002-051-

執行期限：91年8月1日至92年7月31日

主持人：張瑞益 國立台灣大學工程科學系

共同主持人：陳彥良 國立中央大學資訊管理系

計畫參與人員：賴良賓、羅嘉彥、徐嘉鴻、林岳進、

吳怡靜、林宗岳、邱莉媛（國立台灣大學）

#### 一、中英文摘要

資料採擷技術已被許多企業廣泛應用來從大量收集的資料中得到洞察先機的機會。可是針對處理大量資料所必須耗費許多訓練時間的問題，本研究針對此問題，應用以查詢為基礎的學習方法於資料採擷的分類與分群問題上，以解決傳統必須耗費許多時間的問題。初步實驗結果顯示，以查詢為基礎的方式比原方式佳。

**關鍵詞：**資料採擷，查詢學習，電子商務

#### Abstract

Data mining has been widely used in enterprises to analyze data. However, it is time-consuming for mining a large database. In this report, we focus on query-based learning and its applications on some data mining problems. Experiments show that the proposed algorithm is better than the previous one.

**Keywords:** data mining, query-based learning, electronic commerce

#### 二、計畫緣由與目的

傳統上，類神經網路的訓練樣本，是以樣本空間中所有的實例(Instance)做為訓練樣本，這種方法的主要缺點是在對大量的訓練資料進行學習的效率不佳。另外，由於在學習的過程中每個樣本所代表之資訊量皆一樣，使得困難或不易學習的部份，無法提供足夠的資訊量進行學習，一直重

複學習已經學過的知識，而學不到真正期望的知識。詢問式的學習方法與傳統使用訓練資料的差異在於，詢問式傳遞類神經網路能夠主動的選擇訓練的樣本，而不是被動的接受訓練資料。這種學習方法的目標是產生一個簡潔且高度富教育性的訓練資料集合。因此，詢問式的學習方法的主要精神在於針對學習者愈不清楚的地方，加強學習，也就是說「適當的學習時間，給學習者適當的資訊」。正如孔子所言：「因材施教」，所以往往能事半功倍。

以查詢為基礎的學習方法，其學習的程序，歸納為下列六個步驟：

- Step1：隨機產生所有權重的初始值。
- Step2：利用部分的訓練資料，訓練網路。
- Step3：執行反向演算法產生分類邊界點。
- Step4：對每個反向的分類邊界點，計算其梯度，得到共軛的輸入資料對。
- Step5：則將分錯的輸入資料對，加到訓練案例中。
- Step6：重新訓練網路。

圖一為 Query-Based Learning 的架構圖。

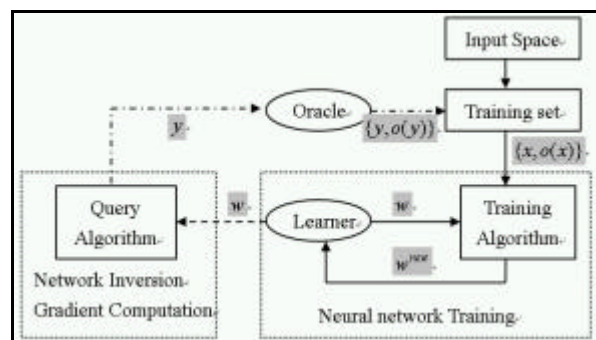


圖 1: Query-Based Learning 架構圖

研究與發展電腦輔助工具以協助醫師診斷工作一直是醫生或研究學者所致力於的方向。對於臨床醫師必須對患者的眾多數據資料進行分析時，很難用以醫師的記憶及個人大腦的判斷來診斷疾病，將繁雜的病患資訊一一分開判讀，而無法做到資訊整合分析的工作【1】。另外，在醫師進行診療判斷時，會因醫師當時的決策情境與個人精神、情緒等因素之影響而使得診斷之結果有所不同，而產生診斷錯誤的可能性是存在的。因此若能建立一套智慧型電腦輔助系統，將可幫助診斷及分析病患之檢查數據，協助醫師診斷，以減少人為判斷的失誤，提升病患更好的治療品質。

本研究的動機，是希望應用詢問式學習方法於疾病診斷資料挖掘的分類問題上，以解決處理大量資料時必須耗費許多時間的問題，並提高模式的正確性。計畫中我們考慮多項資料挖掘的應用問題，包括：心臟病、乳癌、糖尿病之疾病診斷資料。並以學習效率、預測正確率、預測可靠度，以包括假說判斷表(Contingency table)與接受者操作特徵曲線(Receiver Operating Characteristic Curve, ROC)等多項重要指標，來進行預測效率的比較分析。

### 三、結果與討論

實驗方式是應用以查詢為基礎的類神經網路(Query-Based Learning Neural Network, QBL)與傳統隨機選取訓練樣本的倒傳遞類神經網路(Back Propagation Neural Networks, BPN)，兩種方式所建構之疾病預測模式，進行評估比較。

在網路學習效率的比較上，兩種學習方式是以固定學習循環(Epoch)次數後，執行的時間做為比較基準，並以此訓練完成的類神經網路所建構的模式，此測試資料進行驗證，以評估模式預測的正確性。利用模式所預測出之假說判斷表進一步分析預測模式的可靠度。要建立某分類的假說判斷表，可將所有的驗證分類依其假說分類成「真」與「偽」兩類，凡其「目標」分類為疾病患者為真，否則為偽；另外再依其推論分類成「正」與「負」兩類，凡模式「推論」分類為疾病患者則為正，反之為負【3】。

表 1：假說判斷表(Contingency table)

目標 推論		假說	
		真	偽
判 斷	正	True Positive(TP)	False Positive(FP)
	負	False Negative(FN)	True Negative(TN)

根據表 1 可以得到四個主要的衡量指標【1】：這四個衡量指標，其值愈大表示此預測模式之效果愈佳。

- (1) 敏感度(Sensitivity)：衡量此模式能夠正確分類罹患疾病之病患的有效程度，以  $TP/(TP+FN)$  計算。
- (2) 鑑別度(Specificity)：衡量此模式能夠正確分類非罹患疾病之病患的有效程度，以  $TN/(FP+TN)$  計算。
- (3) 預測得病之概似比值(Positive Predictive Value; PPV)：若一病患被診斷為有病，則此病患有多少機率是真正的有病，以  $TP/(TP+FP)$  計算。
- (4) 預測健康之概似比值(Negative Predictive Value; NPV)：若一病患被診斷為健康，則此病患有多少機率是真正的健康，以  $TN/(FN+TN)$  計算。

ROC 分析主要應用於決策時的參考依據，藉由觀察 ROC 曲線上的操作點(Operating point)，可瞭解決策問題本身的損益平衡狀況，而應用在預測的問題時，可作為評估預測模式的局部性能【4】。評估方式是以曲線下方面積(Area Under the ROC Curve; AUC)作為 ROC 曲線評估的準則，此數值若是愈大則表示 ROC 曲線愈佳。本研究使用 MedCalc Software Inc.【5】發展的工具，繪製 ROC 曲線並計算 AUC 值。

#### a. 心臟病資料

根據衛生署統計數據心臟病已名列十大死因的第三位【6】。本研究利用克里夫蘭醫學中心的心臟病患者資料，總共有 270 筆資料，進行心臟病疾病預測模式之建構，其資料屬性包括年齡、性別、胸痛類型、血壓、膽固醇濃度、血糖濃度、心電圖結果、最大心跳速率、運動是否引發心絞痛和血管攝影結果等 13 類屬性資料，以

決定是否有心臟病。並隨機抽樣將資料分為 240 筆為訓練樣本以及 30 筆為測試樣本。

網路架構設計為多層式類神經網路，分為輸入層(13 個神經元)、隱藏層(26 個神經元)和輸出層(1 個神經元)。學習速率設定為 0.45，慣性項(Momentum Term) 設定為 0.9 和網路的收斂誤差值(Convergency error) 設定為 0.05，網路的學習循環設定基準為 1000 次。

BPN 從隨機選取的 240 組訓練樣本進行學習，而 QBL 初始學習是從隨機選取的 240 組訓練樣本中再隨機選取 60 組訓練樣本進行學習，在學習過程中查詢 Oracle 後，再取得 15 組訓練樣本加強學習。BPN 和 QBL 在學習 1000 次的學習循環後，圖 2 說明 QBP 與 BPN 網路學習的收斂過程，而表 2 和表 3 說明執行速度與預測的錯誤率，其中 QBL 僅需 14.76 秒即完成學習且對 30 筆測試資料驗證學習效果，QBL 的錯誤率也比 BPN 低。

預測模式的可靠度評比，如表 4 和表 5 所示，QBL 的敏感度、鑑別度和預測得病之概似比值皆比 BPN 佳，而在預測健康之概似比值方面，QBL 與 BPN 兩個方法的比值相同。另外 ROC 曲線的評估，如圖 5 所示，QBL 也比 BPN 結果好。

依據上述評估指標顯示，在建構心臟病預測模式的類神經網路方法中，QBL 明顯比 BPN 優異。

#### b. 乳癌資料

目前台灣的乳癌發生率為女性好發癌症的第二位，而死亡率則於八十五年首度超越子宮頸癌【6】。本研究利用 Wisconsin 醫學中心的乳癌患者資料，總共有 699 筆資料，進行乳癌疾病預測模式之建構，其資料屬性包括腫塊厚度、單一細胞的大小和形狀和細胞有絲分裂等 9 類屬性資料，判斷此腫瘤為良性或惡性，以決定是否患有乳癌。以隨機抽樣將資料分為 466 筆為訓練樣本以及 233 筆為測試樣本。

網路架構設計為多層式類神經網路，分為輸入層(9 個神經元)、隱藏層(18 個神經元)和輸出層(1 個神經元)。學習速率設定為 0.45，慣性項(Momentum Term) 設定為 0.9 和網路的收斂誤差值(Convergency error)

設定為 0.05，網路的學習循環設定基準為 1000 次。

BPN 從隨機選取的 466 組訓練樣本進行學習，而 QBL 初始學習是從隨機選取的 466 組訓練樣本中再隨機選取 46 組訓練樣本進行學習，在學習過程中查詢 Oracle 後，再取得 5 組訓練樣本加強學習。BPN 和 QBL 在學習 1000 次的學習循環後，圖 3 說明 QBP 與 BPN 網路學習的收斂過程，而表 2 和表 3 說明執行速度與預測的錯誤率，其中 QBL 僅需 5.88 秒即完成學習且對 233 筆測試資料驗證學習效果，QBL 的錯誤率也比 BPN 低。

預測模式的可靠度評比，如表 4 和表 5 所示，QBL 的敏感度、鑑別度和預測得病之概似比值皆比 BPN 佳，而在預測健康之概似比值方面，QBL 與 BPN 兩個方法的比值相同。另外 ROC 曲線的評估，如圖 6 所示，QBL 也比 BPN 結果好。

依據上述評估指標顯示，在建構乳癌預測模式的類神經網路方法中，QBL 明顯比 BPN 優異。

#### c. 糖尿病資料

台灣地區 40 歲以上成年人當中，更是每十人就有一人是糖尿病患者【6】。本研究利用 Pima Indians Diabetes DataBase，總共有 768 筆資料，進行糖尿病疾病預測模式之建構，其資料屬性包括空腹靜脈血血漿葡萄糖、心臟舒張血壓、三頭肌皮皺厚度、血清胰島素濃度、體質指標等資料，決定是否患有糖尿病。以隨機抽樣將資料分為 512 筆為訓練樣本以及 256 筆為測試樣本。

網路架構設計為多層式類神經網路，分為輸入層(8 個神經元)、隱藏層(24 個神經元)和輸出層(1 個神經元)。學習速率設定為 0.45，慣性項(Momentum Term) 設定為 0.9 和網路的收斂誤差值(Convergency error) 設定為 0.05，網路的學習循環設定基準為 10000 次。

BPN 從隨機選取的 512 組訓練樣本進行學習，而 QBL 初始學習是從隨機選取的 512 組訓練樣本中再隨機選取 46 組訓練樣本進行學習，在學習過程中查詢 Oracle 後，再取得 44 組訓練樣本加強學習。BPN 和 QBL 在學習 10000 次的學習循環後，圖 4 說明 QBP 與 BPN 網路學習的收斂過程，而表 2

和表 3 說明執行速度與預測的錯誤率，其中 QBL 僅需 106.44 秒即完成學習且對 256 筆測試資料驗證學習效果，QBL 的錯誤率也比 BPN 低。

預測模式的可靠度評比，如表 4 和表 5 所示，QBL 的敏感度和預測得病之概似比值比 BPN 佳，但是在鑑別度和預測健康之概似比值方面，BPN 比 QBL 評估指標好。另外 ROC 曲線的評估，如圖 7 所示，QBL 是比 BPN 結果好。

依據上述評估指標顯示，在建構糖尿病預測模式的類神經網路方法中，QBL 除了鑑別度和預測健康之概似比值兩個評估指標比 BPN 差，其他的指標仍比 BPN 優異。

表 2: BPN 與 QBL 執行速度(秒)比較表

	心臟病	乳癌	糖尿病
QBL	14.76	5.88	106.44
BPN	52.79	53.29	700.58

表 3: BPN 與 QBL 學習錯誤比較表

	心臟病	乳癌	糖尿病
QBL	4/30 =0.133	5/233 =0.021	66/256 =0.258
BPN	5/30 =0.667	13/233 =0.056	69/256 =0.270

表 4: BPN 與 QBL Contingency Table

	Heart Disease			Breast Disease			Diabetes Disease		
	推 論	假說		推 論	假說		推 論	假說	
		1	0		1	0		1	0
QBL	1	10	1	1	88	2	1	60	24
BPN	0	3	6	0	3	140	0	42	120
	Heart Disease			Breast Disease			Diabetes Disease		
	推 論	假說		推 論	假說		推 論	假說	
		1	0		1	0		1	0
BPN	1	9	2	1	80	10	1	62	22
QBL	0	3	6	0	3	140	0	47	115

圖 2 :Heart Disease

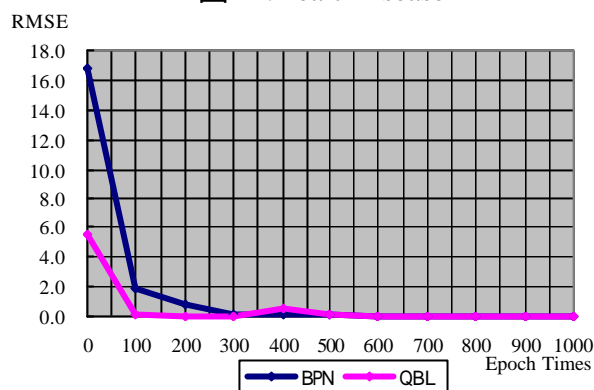


圖 3 :Breast Disease

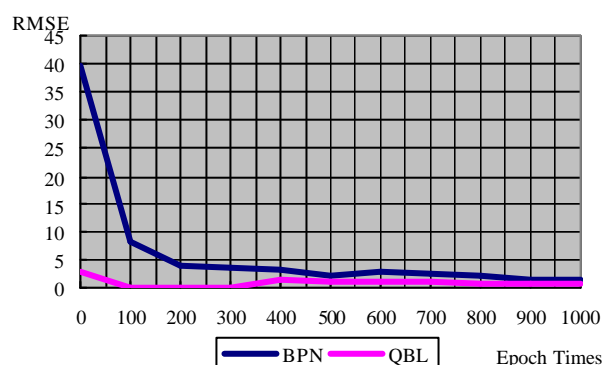
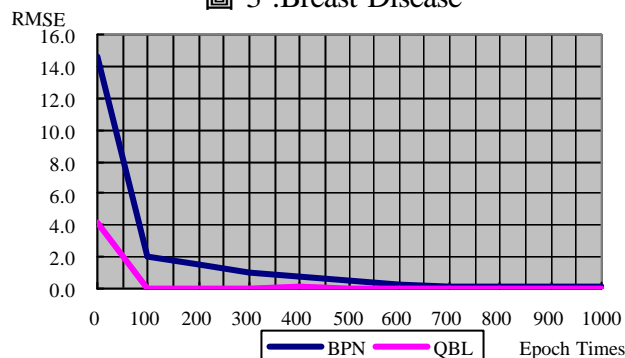


圖 4: Diabetes Disease

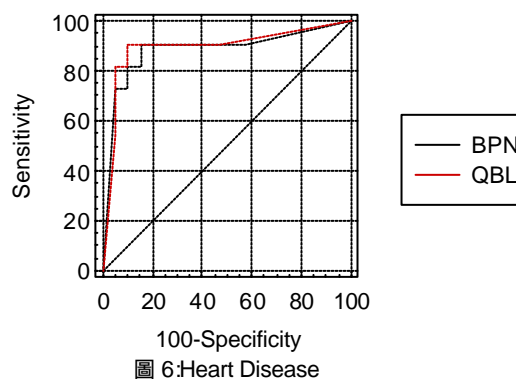


圖 6:Heart Disease

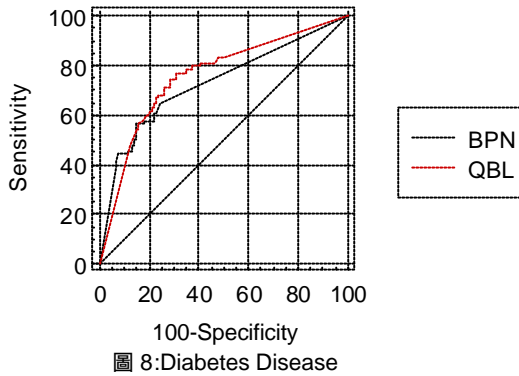
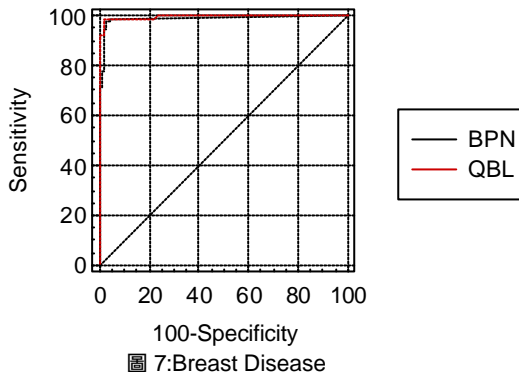


表 5: BPN 與 QBL 效益評估指標值

	Heart Disease		Breast Disease		Diabetes Disease	
Q B L	敏感度	0.77	敏感度	0.97	敏感度	0.59
	鑑別度	0.94	鑑別度	0.99	鑑別度	0.83
	PPV	0.91	PPV	0.98	PPV	0.71
	NPV	0.84	NPV	0.98	NPV	0.74
B P N	敏感度	0.75	敏感度	0.96	敏感度	0.47
	鑑別度	0.89	鑑別度	0.93	鑑別度	0.84
	PPV	0.82	PPV	0.89	PPV	0.74
	NPV	0.84	NPV	0.98	NPV	0.71

PPV : Post-test probability of a Positive test  
 NPV : Post-test probability of a Negative test

表 6: Area under the ROC Curve 比較表

	心臟病	乳癌	糖尿病
QBL	0.895	0.992	0.756
BPN	0.885	0.985	0.727

#### 四、計畫成果自評

本實驗結果證明所提出之詢問式倒傳遞類神經網路學習方法相較於先前之傳統類神經網路方法，能夠帶給使用者更為正確答案且更為迅速的分類。因此對於疾病

預估更能夠提供一個強而有力的工具。未來將多方運用本身的優勢，使資料挖掘的工具發展更為完備。希望能擴展此方法到其他資料挖掘技術上。

#### 五、參考文獻

1. 江宏志, 民 92, 運用基因演算法建構疾病預測模型之研究-以尿路結石疾病預測為例, 國立臺灣大學商學研究所博士論文.
2. Ray-I Chang and Pei-Yung Hsiao, "Unsupervised query-based learning of neural networks using selective-attention and self-regulation," *IEEE Trans. Neural Networks*, vol.8, no.2, March 1997, pp:205-217,...
3. 葉怡成, 民 90。類神經網路模式應用與實作, 儒林書局。
4. DeLeo, J. M. Rosenfeld, S. J. "Essential roles for receiver operating characteristic (ROC) methodology in classifier neural network applications", in Proc. Int. Joint Conf. Neural Networks, Vol.4, 2001, pp:2730-2731
5. MedCalc Software, <http://www.medcalc.be>, 2003 年 4 月 3 日, Accessed。
6. 行政院衛生署, <http://www.doh.gov.tw/statistic/index.htm>, 2003 年 4 月 3 日, Accessed。
7. J. N. Hwang, J. J. Choi, S. Oh, and R. J. Marks II, "Query-based learning applied to partially trained multilayer perceptrons," *IEEE Trans. Neural Networks*, Vol. 2, No. 1, pp. 131-136, January 1991.