行政院國家科學委員會專題研究計畫　成果報告

決策樹形式知識整合之研究

計畫類別：個別型計畫
計畫編號：NSC93-2416-H-004-018-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：國立政治大學資訊管理學系

計畫主持人：林我聰

計畫參與人員：馬芳資、李亞暉、董惟鳳

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 94 年 10 月 26 日

行政院國家科學委員會專題研究計畫成果報告

決策樹形式知識整合之研究

The Research on Decision-Tree-Based Knowledge Integration

計畫編號：NSC 93-2416-H-004-018
執行期限：93 年 8 月 1 日至 94 年 7 月 31 日
主持人：林我聰　政治大學資訊管理學系
計畫參與人員：馬芳資、李亞暉、董惟鳳

一、摘要

隨著知識經濟時代的來臨，掌握知識可幫助組織提昇其競爭力，因此對於知識的產生、儲存、應用和整合，已成為熱烈討論的議題，本研究針對知識整合議題進行探討；而在知識呈現方式中，決策樹（Decision Tree）形式知識為樹狀結構，可以用圖形化的方式來呈現，它的結構簡單且易於瞭解，本研究針對決策樹形式知識來探討其知識整合的課題。

本研究提出決策樹合併修剪方法 DTBMPA (Decision-Tree-Based Merging-Pruning Approach)方法，此方法包括三個主要程序：決策樹合併、合併樹修剪和決策樹驗證。其做法是先將兩棵原始樹經由合併程序結合成一棵合併樹，再透過修剪程序產生修剪樹，最後由驗證程序來評估修剪樹的準確度。本研究提出的 DTBMPA 方法藉由合併程序來擴大樹的知識，再利用修剪程序來取得更精簡的合併樹。

本研究利用實際信用卡客戶的信用資料來進行驗證。在 DTBMPA 方法的實驗中，合併樹的準確度優於原始一棵樹的比率有 90%，而修剪樹的準確度大於或等於合併樹的比率有 80%。在統計檢定中，合併樹和修剪樹的準確度優於一棵樹的準確度達顯著差異。且修剪樹的節點數較合併樹的節點數平均減少約 15%。

關鍵詞：知識整合、決策樹、決策樹合併、決策樹修剪

Abstract

In the knowledge economy era, mastering knowledge can improve organization competitive abilities. Therefore, knowledge creation, retention, application, and integration are becoming the hot themes for discussion nowadays.

Our research focuses on the discussion of knowledge integration and related subjects. Decision trees are one of the most common methods of knowledge representation. They show knowledge structure in a tree-shaped graph. Decision trees are simple and easily understood; thus we focus on decision-tree-based knowledge in connection with the theme of knowledge integration.

First, this research proposes a method called DTBMPA (Decision-Tree-Based Merging-Pruning Approach) to solve this problem. There are three steps in this approach. In the merging step, the first step, two primitive decision trees are merged as a merged tree to enlarge the knowledge of primitive trees. In the pruning step, the second step, the merged tree from the first step is pruned as a pruned tree to

cut off the bias branches of the merged tree. In the validating step, the last step, the performance of the pruned tree from the second step is validated.

We took real credit card user data as our sample data. In the DTBMPA simulation experiments, the percentage accuracy for the merged tree will have 90% of chance that is greater than or equal to the accuracy for those primitive trees, and the percentage accuracy for the pruned tree will have 80% of chance that is greater than or equal to the accuracy for merged tree. And we also find that the average number of nodes of the pruned tree will have 15% less than that of the merged tree.

Keywords: Knowledge Integration, Decision Tree, Decision Tree Merging, Decision Tree Pruning.

二、計畫緣由與目的

在機器學習範疇內的歸納學習法是從實際發生的案例資料中，以自動化的方式來進行知識的擷取與學習。歸納學習法在適當應用領域中確實可以有效地替代知識工程師，來完成知識擷取的工作。近年來決策樹歸納法已有相當多的學者投入研究，並且將演算法進行改良且提昇其效果。甚至已發展出許多商業軟體，其具有學習決策樹的功能，而且它們成功的應用在解決實際的問題中。對於經由學習演算法的複雜運算所獲得的決策樹形式知識之後續知識整合是值得進一步研究的課題；再者決策樹可用圖形化的方式來呈現其知識結構，且層級式結構十分簡單而易於瞭解，因此本研究針對決策樹形式知識整合課題來進行探討。

首先有關決策樹形式知識的合併方面，有些學者利用投票表決、加權投票、或其他演算法來綜合多棵樹的知識，而這些處理方法並沒有實際整合決策樹形式的知識，僅是外加一個結合策略來綜合多棵樹的預測結果值。另外有些學者是把決策樹的樹形結構轉換成多條法則，再利用法則的形式來整合多棵決策樹，但是這樣的處理方法會失掉原有樹形知識的結構，且轉換為法則後，亦會引發其間法則順序的安排、法則修剪、法則重複與法則衝突等問題。

而 Quinlan 以兩兩合併的方式把三棵決策樹合併成一棵決策樹，而他所提出的合併演算法是把第二棵樹取代第一棵樹的葉節點，如此增長了決策樹的路徑，而且使得合併樹的節點數大幅成長。為了減少合併樹的節點數，他採用簡化策略，刪除沒有例子落入的葉節點，然而卻使得合併樹的平均準確度降低，甚至低於一棵樹的平均準確度。有鑑於此，本研究提出新的合併方法，在結合多棵決策樹形式知識的同時，不會大幅成長合併樹的節點數，且期能提昇其預測準確度。

其次在修剪決策樹方面，傳統決策樹修剪的目的是為了避免演算法過於配合訓練例子集來進行分群，或是因為資料內含雜訊而形成過度分支的現象，因而造成產出的決策樹過於龐大且複雜，因此將過度分支進行修剪處理。本研究針對已合併後的決策樹進行修剪處理，期能藉由修剪方法，讓合併樹在保有原有預測能力下，能減少其樹的節點數，換句話說，修剪的目的在於取得一棵較精簡合併樹。

綜合上述，本研究的目的在於探

討如何把決策樹形式知識進行整合，使多棵決策樹形式知識能整合包含於一棵決策樹中，以達成知識累積的目的；同時藉由適當的修剪方法，使合併後的決策樹在保有其良好的預測能力下, 能減少其樹的節點數／複雜度。

三、結果與討論

3-1 決策樹形式知識管理架構

本研究提出一決策樹形式知識管理架構（請參考圖一），此架構包括五個主要元件，『決策樹建立』、『決策樹前置處理』、『決策樹知識整合』、『決策樹儲存』，及『決策樹應用』；其中『決策樹知識整合』元件，為本研究的研究範圍，包含了三個處理程序，即『決策樹合併』、『合併樹修剪』和『決策樹驗證』。

決策樹形式知識管理架構其運作流程說明如下：首先依據研究問題搜集其相關的原始資料，經由前置資料處理，去除雜訊並處理空缺值，然後進行隨機抽樣，把案例資料分成訓練例子集和測試例子集。接著把訓練例子集放入『決策樹建立』來產生出決策樹。再者為了進行決策樹的知識整合，我們建立兩棵原始決策樹，然後將此兩棵決策樹放入『決策樹前置處理』，進行修剪過度分支、處理代表性不足的葉節點、及重複子樹等問題，以及決策樹資料的轉檔等清理動作。

接著進行決策樹知識整合的處理（請參考圖二），即將處理過的兩棵決策樹經由『決策樹合併』程序，合併成一棵決策樹。然後再將合併後的決策樹，利用『合併樹修剪』程序來進行修剪，最後將修剪過後的決策樹放入『決策樹驗證』，利用額外的測試例子集來進行評估，以瞭解合併修剪後決策樹的準確度，即評估其預測未知案例的能力。經檢視並分析測試結果後，倘若準確度符合期望值，則將此合併修剪樹利用『決策樹儲存』單元，儲存在資料庫內，以供『決策樹應用』將來使用。未來有新進例子載入時，我們只需針對這些新進例子集建立決策樹，再將此新建的決策樹合併於我們原有儲存的合併修剪樹之中，即可達成知識整合與累積的目的。

綜合上述，圖一所示架構如同一個知識管理工廠，裡面包括知識產生、知識清理、知識整合、知識重整、知識評估、知識儲存和知識應用等知識管理的功能。

3-2 『決策樹知識整合』元件說明

1. 決策樹合併

決策樹合併的目的在於把兩棵原始決策樹合併成一棵決策樹。而它的做法是由上而下比對兩棵決策樹的相對應節點，當發現兩個節點型態不同，或型態相同但選取不同分徑屬性時，則在既有決策樹新增一個連結節點，將兩個比對節點及其下所有分支子節點，接枝到此連結節點下，而成為其子節點。換句話說，即利用接枝（Grafting）的技術，把新的決策樹中不同於既有決策樹的分支，利用一個連結節點，接枝到既有決策樹，以擴增既有決策樹的知識。

2. 合併樹修剪

合併樹修剪的目的在於取得一棵較精簡的樹，同時也避免合併樹經由合併程序而產生龐大的合併樹。由下而上計算每一子樹之估計錯誤率及把它還原成葉節點之錯誤率，倘若子樹

的錯誤率大於還原成葉節點的錯誤率時，就把此子樹還原成一個葉節點。以下為計算公式：

- 葉節點之估計錯誤率

$$e_t = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}$$

f = E/N 為節點的誤判率，而其中 E 是錯誤分類的例子數，N 是此落入此葉節點的總例子數；當信賴水準為 25%，則 z 值為 0.69。

- 子樹之估計錯誤率

$$e_T = \sum_{i=1}^{k} \frac{N_i}{N} e_i$$

其中 N 是子樹 T 的根節點的總例子數，N_i 是 T 的第 i 子節點的例子數，而 T 有 k 個子節點。

3. 決策樹驗證

目的在於評估決策樹在合併修剪前後之績效的差異。本研究中，決策樹的績效評估指標包括複雜度和準確度；複雜度衡量決策樹的大小，即決策樹的節點數、法則數（葉節點個數）、及樹的層級等；準確度衡量決策樹分類測試例子集的正確命中比率。驗證的進行方式採用保留法(Holdout Method)，即在建立決策樹之前，先將例子集分為兩群，其中一群用來建立決策樹，另一群則保留來驗證此決策樹。這個程序主要是利用測試例子來評估合併修剪樹的預測未知例子的能力及計算其節點數目。

3-3 實驗設計與結果分析

本研究採用發卡銀行之信用卡客戶歷史信用資料來進行實驗，經過資料整理，刪除無效樣本及空缺值，共計可用樣本為103,653筆。在目標值選取上，本研究以正常流通卡為信用良好客戶，而以強制停卡者為信用不良客戶。

實驗樣本內所搜集到的屬性包括年齡、學歷、年收入、有無甲存、有無不動產、行業別、公司等級、職等、一般卡張數及金卡張數，以及帳款餘額和六個月的繳款記錄等，共計有十七個；其中年齡、年收入及帳款餘額為數值性線索，其餘為非數值性線索。期能讓系統充分學習信用良好與信用不良客戶之特徵，所以將此十七個屬性完全納入學習之範圍。

實驗結果與發現，列示如下：

1. 準確度的比較分析：合併樹的準確度比原始一棵樹的準確度有顯著提昇。就七十次的準確度比較之中，合併樹優於一棵樹之比率為90%。就統計檢定分析而言，在顯著水準 α 值為0.05且自由度是69的情況下，合併樹的準確度優於一棵樹的準確度達顯著差異。由此可見，藉由樹形知識合併，可以擴大決策樹的知識涵蓋面，提昇其預測未知例子的準確度。
2. 合併樹的節點數成長問題：合併樹的平均節點數是原始樹的平均節點數的1.8倍，其中在最差的情形下是合併樹的節點數是原始兩棵樹的節點數加總；相對於Quinlan的合併方法，其在合併三棵樹的規模下，合併樹的平均節點數約為一棵樹的平均節點數的73.09倍；由此可見，本研究所提出的方法所產生的合併樹之節點數較不會大幅成長。而比較合併樹與修剪樹的節點數，得知修剪樹的平均節點數約為合併樹的平

均節點數的百分之八十五；換言之，修剪方法使得合併樹的節點數減少約15%。

3. 結合的順序：本研究採用兩兩合併的方式進行多棵決策樹的知識合併，對於決策樹的合併順序是否影響合併後決策樹的準確度，本研究經由統計檢定得知合併順序並不影響合併樹的準確度。
4. 判定最終預測值策略：當有多條法則在預測某一案例時，其預測值互相衝突或預測之葉節點內的各類別例子數之比例相當時，則以此策略來綜合出最佳之預測值。在此我們利用強態葉節點來加強預測之準確度。經由實驗結果得知，藉由設定強態法則可以提升合併樹的準確度（優於原始樹的準確度約26%）。
5. 知識累積：本實驗案例中，持卡人的消費資料是與日俱增，發卡銀行不需就所有資料重新去產生一棵完整決策樹，而可僅就新增資料部分產生一新增決策樹，再利用本研究所提出之DTBMPA方法將此新增決策樹加入原有決策樹以擴充原有決策樹的知識。此外，發卡銀行亦可將公司內部之授信法則或專家知識法則，轉換成樹形知識，然後再利用本研究所提出之方法來整合至合併決策樹內，以提昇決策樹內部知識的廣度，以達知識累積的目的。

四、計畫成果自評

以往學者對於決策樹知識整合的作法法大抵採用投票表決、轉換成法則集的方法以綜合多棵樹的知識，而這些處理方法或者沒有實際整合決策樹形式的知識，或者失去了原有一棵樹形結構的知識體系。而Quinlan提出的合併方法雖可以將多棵樹合併成一棵樹，然而產生的合併樹的節點數呈乘數的巨幅成長；經修剪後，其準確度低於一棵原始樹的準確度。有鑑於此，本研究針對這個議題，提出一個新的方法，進行決策樹形式知識的整合，且讓合併後的決策樹之節點數呈加數的成長；同時加入修剪方法，以取得一棵較為精簡的合併樹。實驗結果顯示合併樹的準確度優於一棵樹的準確度；而在保有合併樹的準確度下，修剪方法可以減少合併樹的節點數／降低樹的複雜度。

非常感謝國科會對本研究計畫執行經費的補助，本研究經過一年的努力，已達成當初研究計畫提出時的預期成果。本研究兼顧理論創新與實務應用，相關成果已著手整理成論文，投稿至學術會議與期刊上，希冀能對學術界相關研究及業界實際應用皆能有所助益。

五、重要參考文獻

[1] Frank, E., Pruning Decision Trees and Lists, Department of Computer Science, University of Waikato, Hamilton, New Zealand. 2000.

[2] Mingers, J., An empirical comparison of pruning methods for decision tree induction, Machine Learning, Volume 4, 1989, pp.227-443.

[3] Quinlan, J. R., MiniBoosting Decision Trees, Journal of Artificial Intelligence Research, 1998.

[4] Quinlan, J.R., C4.5: Programs for Machine Learning, San Mateo: Morgan Kaufmann, 1992.

[5] Quinlan, J.R., Simplifying decision trees. International Journal of Man-Machine Studies, 1987, 27(3), pp.221-234.

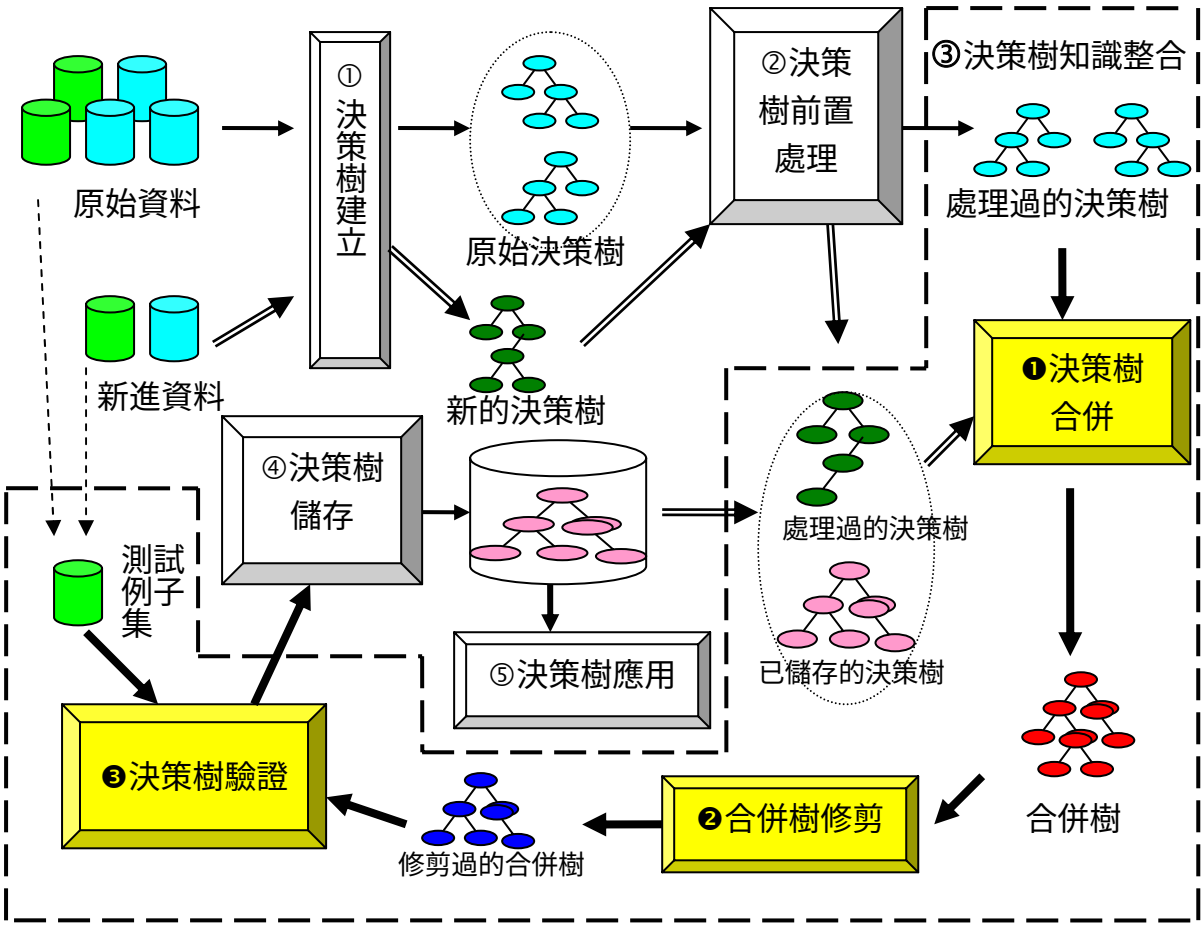
[6] Williams, G., Induction and Combining Multiple Decision Trees, Ph.D. Dissertation, Australian National University, Canberra, Australia, 1990.

[7] Windeatt T. & Ardeshir G., An empirical comparison of pruning methods for ensemble classifiers, Proc. of Int. Conf Intelligent Data Analysis, Sept 13-15, Lisbon, Portugal, Lecture notes in computer science, Springer-Verlag, 2001, pp.208-217.

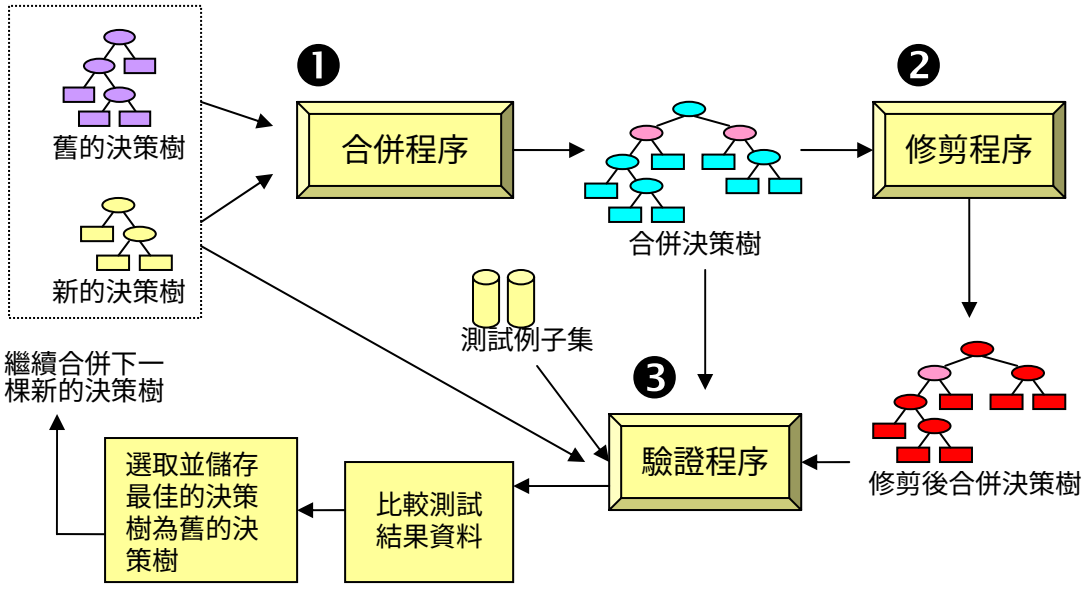
[8] Witten, I. H. & Frank, E, Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations, Morgan Kaufmann, 2000.

[9] 馬芳資，信用卡信用風險預警範例學習系統之研究，第十屆全國技職及職業教育研討會，技職研討會，商業類 I，1995 年，pp.427-436。

[10]陳重銘，結合直線最適法於決策樹修剪之影響研究，國立中山大學資訊管理研究所碩士論文，1995 年。



圖一：決策樹形式知識管理架構



圖二：DTBMPA 方法之運作流程