專家系統知識自動整合環境之研製

Developing an Automatic Knowledge Integration System for Multiple Knowledge Sources

計畫編號：NSC 87-2213-E-009-086
執行期限：86 年 8 月 1 日至 87 年 7 月 31 日
主持人：曾憲雄　　交通大學資訊科學系
E-mail: sstseng@cis.nctu.edu.tw

一、摘要

專家系統已成功地應用在各種領域並展現出優異之成效，但是當知識庫系統愈大時，其建構成本及困難度就愈高。本計畫中，我們完成一套專家系統的知識自動整合環境，包括知識擷取模組、機器學習模組、知識整合模組以及知識修正模組。此環境可以利用知識擷取與機器學習模組由各種不同的知識來源取得資料，加以整理後，再利用知識整合模組整合由知識擷取模組所產生的領域知識及由機器學習模組所歸納產生的規則，並產生專家系統的知識庫。知識修正模組則運用於知識庫推理中，若有不適用之規則存在，則將自動找出並加以修正，隨時維持知識庫的一致性及準確性。

除此之外，我們亦實際地將此環境應用在兩個真實領域，並建立腦瘤疾病診斷專家系統與甘蔗育種培育預測專家系統，由實驗的結果顯示，我們的方法與系統皆有相當好的效能。
關鍵詞：知識整合、基因演算法

Abstract

Expert systems have been successfully applied to many fields and have shown excellent performance. The cost of the effort is high and will become prohibitive as we attempt to build larger and larger systems. Reusing and integrating available knowledge from a variety of sources thus plays a crucial role in reducing the development cost. However, knowledge reuse and knowledge integration are still very difficult and full of challenge although they have many advantages.

In this project, we develop an automatic knowledge-integration environment which consists of four main modules: knowledge acquisition, machine learning, knowledge integration, and knowledge refinement. The knowledge acquisition module collects various knowledge acquisition tools to help domain experts input knowledge. The machine learning module stores various learning systems to induce domain knowledge from different training sets. In the knowledge integration module, an automatic knowledge-integration approach combines multiple knowledge inputs derived by knowledge acquisition tools or machine learning methods to construct the initial knowledge base. In the knowledge refinement module, a knowledge refinement scheme is proposed to modify the existing knowledge base during the process of inference.

Furthermore, we apply our approaches to two real-world application domains, Diagnosis of Brain Tumor and Prediction of Sugar-Cane Plant Breeding, for evaluation. The experimental results show that our approaches have good performance.
Keywords: Knowledge Integration, Genetic Algorithm

二、計畫緣由與目的

近年來，專家系統被廣泛的運用到許多應用範疇，並一再顯示出其優越特性。然而，要建構一個成功的專家系統，必須能有效地結合眾多專家的專業領域知識、文獻所記載的資料、以及真實的案例等資訊，才能開發出一個完整、一致且清晰的知識庫。相對地，在建構專家系統的過程中卻顯現了幾個瓶頸，最常見的就是專家知識擷取與專家知識整合的困難。近年來，由於知識擷取與機器學習技術的蓬勃發展，已克服了傳統知識擷取的瓶頸，然而知識的整合工作卻一直停留在需要專家介入處理的階段，造成因人為因素之關係而影響知識整合之效率。

針對傳統知識整合系統諸多問題，基因遺傳演算法的自我調整找尋技術是一個處理這些問題的好方法。因為如何建構一個完整且一致的知識庫，本身就是一個最佳化問題，而基因遺傳演算法的自我調整找尋技術正是處理最佳化問題的利器。此外，由於大部份的專家系統建構工具並未支援知識整合功能，使得使用者必須付出相當的代價，包括金錢及時間，來從事專家系統的發展。為了解決知識整合的困難，我們計畫建立一個知識自動整合系統，它不但可以協助知識工程師獲取專家們的知識，並且可以有效地將專家們的專業知識以及過去發生過的案例資訊自動整合成一個完整且一致的知識庫。

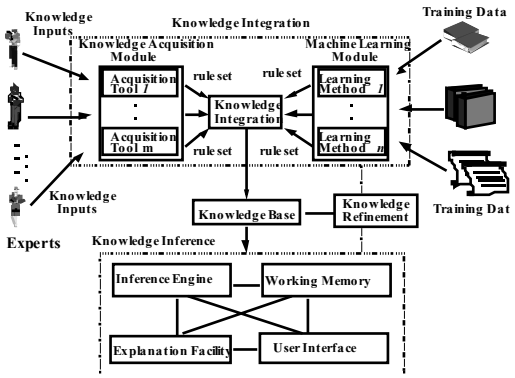
在實際應用方面，我們將以醫學上腦瘤診斷及農業上甘蔗育種的實際應用領域進行實作。在應用領域中將邀請專門醫師及學者們作為專家，並依實際病歷及歷年紀錄的資料進行測試，以驗証計畫之成果。

在腦瘤診斷的應用領域中，電腦斷層掃瞄器是必要的輔助工具，它可以協助醫生了解腦瘤生長部位及部份特性，但腦瘤診斷至今它還是一件困難的事，除了依靠設備之輔助幫忙外，尚需有經驗醫生的診斷知識。目前已有不少的研究試圖提昇腦瘤診斷的正確率，也有部份教學醫院已著手研究專家系統的開發以便技術的傳遞，但目前都還在雛型階段，為了突破知識取得的瓶頸及有效地結合各種知識資源，及專門醫師的知識或一些雛型系統所隱含知識；我們預計利用知識整合系統來整合不同的知識來源來提昇系統的準確性，並達到知識共享及知識重建的目的。

在甘蔗育種的應用領域中，我們面對的是歷年來紀錄的育種資料，其紀錄格式為每一株甘蔗的各項特徵及其親本資料，甘蔗的各項特徵包括莖長、莖寬等農藝特性、含糖量、花粉情況、及各種病蟲害特性等等，另外，親本資料則記錄了此株甘蔗是由哪兩株甘蔗交配培育而成；有了這些育種紀錄，育種專家可以依據歷年的紀錄及其專業知識，歸納出一些適合的規則以選擇新年度的交配組合，其最大的難題是交配組合數量很大，其中隱藏的規則太過於複雜，單純依據專家的歸納似乎不夠準確；我們預計利用知識整合系統學習出隱藏於歷年育種紀錄中的規則，並與專家歸納出來的結論作整合，以便完成甘蔗育種輔助系統，減少在傳統育種工作上可能出現的難題。

三、結果與討論

我們發展了一個知識自動整合環境 (如圖一)，整合由知識擷取工具所獲取的專家知識以及由機器學習方法所歸納而得的規則，並產生知識庫，以完成知識邏輯的推理。整個系統環境包含四個模組，分別敘述如下：



圖一 知識自動整合環境

1. 知識擷取模組

知識擷取模組可在存放一些知識擷取工具，專家們可依使用的習慣以及知識擷取工具之知識表示方式來選用適當的工具，它可以幫助領域專家輸入領域知識。另外，本模組具有擴充性，可加入不同方法的知識擷取工具。只要是適當的專家知識擷取工具皆可放入本模組中，待日後專家輸入知識時即可選用。

2. 機器學習模組

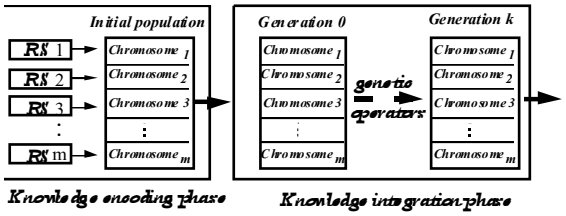
機器學習模組可存放一些著名的學習系統，使用者可依學習資料描述的形式以及學習系統之知識表示方式來選用適當的學習系統。它可以將過去發生的一些案例歸納出一些規則。目前，本模組收容了一些學習系統，如：ID3、PRISM、ASSISTANT、及 Version Space 等系統。另外，本模組亦具有擴充性，可放入不同方法的學習系統，只要是適當的學習系統皆可放入本模組中，以待日後選用。

3. 知識整合模組

本模組提供一個知識自動整合的方法，可將知識擷取模組所產生的領域知識或機器學習方法所歸納的規則加以整合，並產生知識庫。此整合方式不需要領域專家介入處理，同時，整合所需的時間遠比傳統方法所需要的少。

本計畫所提出的知識整合方法，是運用基因遺傳演算法之自我調整找尋技術，藉由一些基因遺傳演算子，將欲整合之知識或規則，彼此"交配"或"突變"，最後，產生一個知識庫。整個過程如圖二所示，共分為知識編碼及知識整合二階段，其中的 $RS_1, \ldots, RS_m$ 代表欲整合之知識或規則。為了方便將基因遺傳演算法技術運用在知識整合系統上，首先，必須將所有欲整合之知識或規則加以編碼，以

位元字串方式來描述每一個知識或規則。然後，再運用基因遺傳演算法之自我調整找尋技術加以整合，產生知識庫。



圖二 知識編碼及知識整合

3.1 知識編碼

在此，我們假設每個知識來源皆可用規則集合的型式來表示，主要是因為現行的一些專家知識擷取工具或機器學習系統所產生的知識，大多以規則的方式表示，或者能很容易地轉換成規則。在此我們運用 **Pittsburgh** 方法將每一個規則集合編碼成一串變動長度的 **0/1** 字串。以腦瘤診斷為例，假設有二種類型腦瘤如 {*Adenoma, Meningioma*}，可用三種特徵 {*Location, Calcification, Edema*} 來辨別。特徵 *Location* 有三種可能值 {*brain surface, sellar, brain stem*}，特徵 *Calcification* 有四種可能值 {*no, marginal, vascular-like, lumpy*}，特徵 *Edema* 有三種可能值 {*no, < 2 cm, < 0.5 hemisphere*}。假設有一個規則集合 RS_i，只含二條規則：

R1: **If** (*Location* = *sellar*) **and** (*Calcification* = *no*) **then** *Class* **is** *Adenoma*;
R2: **If** (*Location* = *brain surface*) **and** (*Edema* < *2 cm*) **then** *Class* **is** *Meningioma*.

我們先將這些規則以中間型式表示：

R_1': **If** (*Location* = *sellar*) **and** (*Calcification* = *no*) **and** *(Edema = no or Edema < 2 cm or Edema < 0.5 hemisphere)* **then** *Class* **is** *Adenoma*;

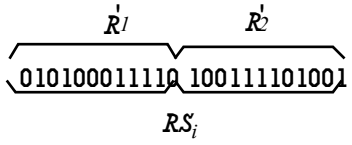
R_2': **If** (*Location* = *brain surface*) **and** *(Calcification = no or Calcification = marginal or Calcification = vascular like or Calcification = lumpy)* **and** (*Edema* < *2 cm*) **then** *Class* **is** *Meningioma*.

在此，*R1* 及 *R2* 與 R_1' 及 R_2' 邏輯相同。

每一條規則以固定字串方式來表示，例如：特徵 *Location* 之特徵值 {*brain surface, sellar, brain stem*}，則利用三個位元來代表這些特徵值。而字串 *101* 代表特徵 *Location* 之值為 *"brain surface"* 或 *"brain stem"*。所以，例子中的兩條規則編碼成如下：

	Location	*Calcification*	*Edema*	*Class*
R_1'	010	1000	111	10
R_2'	100	1111	010	01

最後，RS_i 編碼成如下：



3.2 知識整合

在此我們將基因遺傳演算法運用在知識整合上，其步驟如下：

1. 將欲整合之知識或規則表示成一群問題的可能解 (知識庫)；
2. 定義一個評估函數，評估每一個可能解 (知識庫) 的適用程度；
3. 從這一群問題的可能解 (知識庫) 中找出適用程度較高的一些可能解；
4. 定義一些基因遺傳演算子，對適用程度較高的可能解(知識庫)進行交配，並產生 新的可能解 (知識庫)；
5. 反覆進行步驟 2 至步驟 4 的動作，直到找到最佳解 (知識庫)；
6. 最佳解成為最後的知識庫。

由知識擷取模組所產生的領域知識或由機器學習方法所歸納產生的規則集合，皆可視為知識庫的"候選者"，而這些"候選者"可透過基因遺傳演算子彼此"交配"或"突變"，產生更優良的下一代。這些基因遺傳演算子功能如下：

1. **Dynamic Crossover**：將欲整合之知識或規則集合，彼此 "交配" 產生新的下一代。
2. **Mutation**：任意改變某一知識或規則集合內之某一條規則。
3. **Fusion**：解決知識整合過程中，知識或規則間的重覆 **(redundancy)** 和包含關係。

4. **Fission**：修正知識整合過程中，錯誤的知識或規則 (**misclassification**)，並解決知識或規則間之矛盾 (**contradiction**) 關係。

在整合過程中，我們考慮兩個重要因素：知識庫的準確度與知識庫的複雜度。我們希望能整合出一個準確性高以及複雜度低的知識庫，所以，在此我們準備一些例子來檢測知識庫 (RS) 的準確度，其中例子愈多，檢測效果愈準。其準確度的計算如下：

$$Accuracy(RS) = \frac{the\ total\ number\ of\ test\ instances\ correctly\ matched\ by\ RS}{the\ total\ number\ of\ test\ instances}$$

，而知識庫 (RS) 結構的複雜度的計算如下：

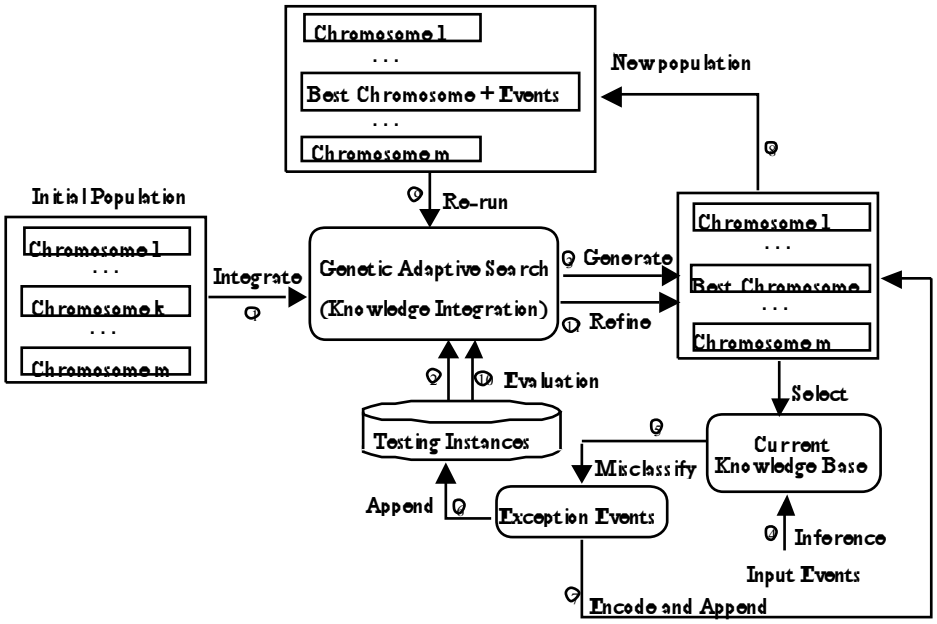
$$Complexity(RS) = \frac{Number\ of\ rules\ within\ the\ integrated\ rule\ set\ RS}{[\sum_{i=1}^{m}(Number\ of\ rules\ within\ initial\ RS_i)]/m}$$

其中 $RS_1, \ldots, RS_m$ 是欲整合之知識或規則集合。由上面兩個考慮因素，我們定義出一個知識庫的評估函數如下：

$$fitness(RS) = Accuracy(RS) * \frac{1}{Complexity(RS)}$$

4. 知識修正模組

當知識庫整合完成後，於推理過程中若有某些例子 (**events**) 無法正確地被目前之知識庫所辨斷，則將這些例子所隱含的資訊加入目前知識庫內，再運用知識整合程序自動地修正目前的知識庫，以便隨時維持目前知識庫的一致性及準確性。知識修正流程如圖三所示，其中符號 "○" 內的數字代表處理流程順序。



圖三 知識修正流程

四、計畫成果自評

在本計畫中，我們完成了下列幾項工作：

1. 完成知識自動整合環境，包括知識擷取模組、機器學習模組、知識自動整合模組、及知識修正模組。
2. 完成系統整體之模擬與測試
3. 完成腦瘤診斷專家系統並進行評估
4. 完成甘蔗育種輔助系統並進行評估

以上成果，與原先預期目標相符。完成的系統可以進一步的推廣至更實際的應用上，從理論研究的角度而言，我們完成了以基因演算法為基礎的知識整合模式的建立，可以作為日後此領域研究的參考。

五、參考文獻

[1] C. Baral, S. Kraus, and J. Minker, "Combining multiple knowledge bases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 3, no. 2, pp. 208-220, 1991.

[2] M. L. Shaw and B. R. Gaines, "KITTEN: Knowledge initiation and transfer tools for experts and novices," *International Journal of*

Man-Machine Studies, vol. 27, pp. 251-280, 1987.

[3] C. Y. Suen, Y. S. Huang and A. Bloch, "Multiple expert systems and multi-expert systems," *The Second World Congress on Expert Systems*, pp. 207-212, 1994.

[4] C. H. Wang, T. P. Hong, and S. S. Tseng, "Self-integrated knowledge-based brain tumor diagnostic system," Expert Systems With Applications, vol. 11, no. 3, pp.351-360, 1996.

[5] C. H. Wang, T. P. Hong, S. S. Tseng, and C. M. Liao, "Automatically integrating multiple rule sets in a distributed-knowledge environment," IEEE Transactions on Systems, Man, and Cybernetics :Part C," Vol. 28, No. 3, Aug. 1998, pp. 471-476.

[6] C. H. Wang, T. P. Hong, and S. S. Tseng, "Integration membership functions and fuzzy rule sets from multiple knowledge sources," accepted by Fuzzy Sets and Systems.

[7] C. H. Wang, T. P. Hong, and S. S. Tseng, "Knowledge integration by genetic algorithm," International Fuzzy Systems Association World Congress, vol. 2, pp. 404-408, 1997.

[8] C. H. Wang, T. P. Hong, and S. S. Tseng, "A hybrid genetic knowledge-integration strategy," accepted by IEEE International Conference on Evolutionary Computation, ICEC'98

[9] C. H. Wang, T. P. Hong, and S. S. Tseng, "Genetic-Fuzzy Knowledge-Integration Strategies" accepted by 10TH IEEE International Conference on Tools With Artificial Intelligence

[10] C. H. Wang, T. P. Hong, M. B. Chang, and S. S. Tseng, "Integrating Multiple Rule Sets By Genetic Algorithms" accepted by IEEE Conference on Systems, Man, and Cybernetics, 1998.