行政院國家科學委員會專題研究計畫　期中進度報告

演化式演算法應用於資料探勘之研究(1/2)

計畫類別：個別型計畫
計畫編號：NSC92-2213-E-009-073-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：國立交通大學工業工程與管理學系

計畫主持人：沙永傑

計畫參與人員：劉正祥

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 5 月 28 日

中文摘要

在知識發掘與資料探勘的領域中，分類技術被視為一個重要的研究議題。分類的目的在於定義每一個類別的特徵，透過訓練組的資料，建立一個判斷類別歸屬的模型，將未歸類的資料分門別類。

一般來說，資料類型可分成兩大類：數值型、類別型。在真實生活中，數值型資料之存在是相當普遍的，但是大多數的分類預測方法，只能處理類別型資料。針對數值型資料大多數研究會在資料前置處理（preprocessing）過程中將其離散成類別型資料，之後再利用資料探勘工具萃取分類規則。然而任何離散數值型資料的方法，都會造成原先隱藏於資料當中的資訊流失。因此本研究希望能發展一套分類預測方法，能在不離散數值型資料情況下進行分類規則的萃取。

本研究共分成兩個階段，第一階段將以基因演算法（GA）為主要研究工具，希望藉由基因演算法的高效率與彈性，設計出能同時處理數值型與類別型資料的分類預測方法。第二階段為針對螞蟻理論（ACO）設計出能同時處理數值型與類別型資料的分類預測方法，並比較兩種演化演算法在分類績效上的差異。

關鍵詞：資料探勘、離散化、基因演算法、螞蟻理論

ABSTRACT

Classification is one of the important issues in knowledge discovery and data mining. The goal of classification is to define the characteristics for each class in order to predict if a previously unknown object either belongs to the class or not.

In general, attribute data type can divide into two groups: numerical and categorical. Numerical attributes are very common in real-world application. There exist a large number of classification algorithms, which handle categorical attributes only. Therefore, the process of the descretization is an essential task for data preprocessing in knowledge discovery in databases (KDD). But during the discretization of numerical attribute, some information hidden in the data set can be lost, we will proposed classification algorithms can extract classified rules with numerical and categorical attributes simultaneously.

The first step of this project is based genetic algorithm to develop more effective classifier (GA-Classifier), which handle numerical and categorical attributes simultaneously than traditional classification algorithm—Decision Tree (C4.5). The second step of this project is based a novel evolutionary algorithm (Ant Colony Optimization, ACO) to develop more effective and efficient classifier than

GA-Classifier and Decision Tree.

Keywords: Data mining, Descretization, Genetic algorithm, Ant colony optimization

一、前言

為因應環境的變遷與商業競爭日益激烈，許多企業在面臨資訊科技發展的潮流下，均希望透過資訊科技的力量為企業帶來更多的競爭優勢。然而當企業引進資訊技術來有效率地收集資料時，卻發現無法有效率地從所收集數量龐大的資料中，發掘出有用的知識與規則。因此企業的焦點逐漸轉變到如何有效的利用所收集到的資料獲取有用的資訊。所以資料探勘的技術就逐漸被學術界與業界所重視。常見的資料探勘的定義有以下數種。

> Cabena(1997)定義資料探勘是將先前不知道，有效的資訊從龐大資料庫中萃取出的過程，並提供給決策者作為決策依據。
>
> Hall(1995)定義資料探勘乃是針對大量的資料，以全自動或半自動的方式進行分析，找出有意義的關係或規則。
>
> Berry(1997)定義資料探勘結合許多不同的技術，如資料視覺化（Data Visualization）、機器學習（Machine Learning）、統計（Statistics）以及資料庫（Databases）以便從龐大資料量中萃取以規則形式或其他模式所表達的知識。

資料探勘可以應用的領域幾乎涵蓋了各行各業，例如：生產製造、財務投資、信用卡交易、服務業、…、等等。基本上，資料探勘是知識發現（Knowledge Discovery in Databases, KDD）當中的一個步驟。知識發現大致可分成五大步驟[Bruha, 2000]：

1. 瞭解資料探勘所要應用的領域以及熟悉相關知識，並選擇所要使用的資料探勘技術。
2. 進行目標資料的收集。
3. 進行資料前置處理。針對目標資料當中不一致或是遺漏值進行必要的處理。由於所收集的目標資料當中，屬性的資料類型可分成兩大類：數值型、類別型。在資料前置處理過程中會將把數值型資料進行離散化，以轉換成類別型資料，以方便資料探勘工具的使用。
4. 將經過資料前置處理後的資料，以資料探勘工具進行知識萃取。
5. 資料後置處理。透過專家來驗證所萃取出的知識，並將有用的知識納入現有的決策系統。

資料探勘的相關作業應用可區分成五種[Collard, 2001]：資料特徵描述

（Description）、分類（Classification）、關連分析（Association Discovery）、順序樣式分析（Sequential Pattern Analysis）、迴歸（Regression）。

在本計劃中主要在探討資料探勘中的分類作業。在資料探勘的領域中，分類被視為一個重要的研究議題。分類的目的在於定義每一個類別的特徵，透過訓練組的資料，建立一個判斷類別歸屬的模型，將未歸類的資料分門別類。分類作業所萃取出的知識，其表達形式通常為 IF-THEN 規則，表達如下所示。

IF　<先決條件式>　THEN　<類別歸屬>

先決條件式由數個條件子（terms）所組成，每一個條件子利用邏輯運算符號（AND）連結。每一個條件子由三項資訊組成：屬性、比較運算符號、屬性值，例如：<性別=男性>。類別歸屬用來預測符合先決條件式的資料其應屬類別值。從使用者的觀點而言，IF-THEN 的知識表達形式比較簡單易懂，較能讓知識使用者所接受。

一般來說，進行分類作業時多數的分類預測方法，針對數值型資料會在資料前置處理（preprocessing）過程中將其離散成類別型資料，之後再利用資料探勘工具萃取分類規則。然而任何離散數值型資料的方法，都會造成原先隱藏於資料當中的資訊流失。因此本研究希望能發展一套分類預測方法，能在減少資訊流失的前提下離散數值型資料，進行分類規則的萃取。

二、研究目的

本計劃目的在於利用演化式演算法來發展分類預測方法，能同時處理數值型與類別型資料，在減少資訊流失的前提下離散數值型資料，並進行分類規則的萃取。本計劃共分成兩個階段，第一階段將以基因演算法（GA）為主要研究工具，希望藉由基因演算法的高效率與彈性，設計出分類績效較傳統分類工具—決策樹(Decision Tree, DT)好的分類器。第二階段將根據相同概念以螞蟻理論（ACO）設計出能同時處理數值型與類別型資料的分類器，並比較兩種演化演算法在分類績效上的差異。

三、文獻探討

目前已有許多機器學習(Machine Learning, ML)工具被應用於分類技術，如：決策樹、類神經網路、基因演算法。決策樹是目前最常被使用於分類作業的工具，其優點在於所產生的規則容易被人們所接受與解釋，缺點在於無法偵測與利用有交互作用的屬性[Clare, 2000]。同時決策樹針對數值型資料必須在知識發現（Knowledge Discovery in Databases, KDD）的資料前置處理（preprocessing）過程中使用離散工具(C4.5 Discretization)將其離散成類別型資料，之後再利用演算

法(C4.5)進行規則萃取。然而任何離散數值型資料的方法，都會造成原先隱藏於資料當中的資訊流失[Bruha, 2000]，所以應在不離散數值型資料情況下進行規則萃取。另外，類神經網路雖能同時處理數值型資料與類別型資料，但是其計算過程被視為黑箱(black box)且輸出的結果難以被人所解釋。至於基因演算法具有輸出結果容易被解釋以及有全域搜尋最佳解能力的優點，但缺點是計算時間較長。因此本計劃將基於演化演算法發展有效率的分類預測方法，能在不離散數值型資料情況下進行分類規則的萃取，具同時處理數值型與類別型資料的能力。

在過去有許多研究文獻利用基因演算法發展分類器[Congdon, 2000][Bandar, 1999][Lopes and Pozo, 2001][Fu and Mae, 2001][Pozo and Hasse, 2000][Shin and Lee, 2002][Bruha, 2000][Noda et al., 1999][Fidelis et al., 2000]，相關文獻證明利用基因演算法進行分類其分類正確率優於決策樹。表一彙整近年來使用基因演算法求解分類問題之研究，並說明各研究中針對數值型資料處理方式以及允許建構於規則當中的比較運算符號與邏輯運算符號。從表一可以了解利用基因演算法應用於分類問題時，早期對數值型資料均是加以離散化，同時允許建構於規則內的比較運算符號只限等於（＝），邏輯運算符號亦只限交集（AND）。近年來逐漸有學者發表可以同時處理數值型資料與類別型資料的基因演算法，同時所能建構於規則當中的比較運算符號也較以往來的多。至於邏輯運算符號還是只允許 AND 符號。

表一　基因演算法應用於分類技術研究之彙整

作者	年代	數值型資料處理方式	比較運算符號		邏輯運算符號	備註
Noda et al.	1999	無數值型資料	=		AND	--
Bruha et al.	2000	離散化	=		AND	--
Congdon	2000	離散化	＝		AND	--
Fidelis et al.	2000	不離散化	數值型資料	類別型資料	AND	--
			≧、＜	＝、≠		
Pozo and Hasse	2000	不離散化	數值型資料	類別型資料	AND	--
			≦、≧	＝		
Shin and Lee	2002	不離散化	數值型資料	類別型資料	AND	限制條件子數目為 5 個
			＜、≧	＜、≧		

經由表一可以了解使用基因演算法所能產生的分類知識表達形式(IF-THEN)其變化性較少，因此第一年計劃將針對基因演算法發展一套有效率、符號變化性較多的分類器，並能同時處理數值型與類別型資料。

螞蟻理論是由 M. Dorigo 在 1996 首次發表應用求解旅行者推銷問題(TSP)，此理論可以應用於其他 NP-Hard 問題，如：非對稱式旅行者推銷問題（ATSP)、二次規劃問題（QAP)、零工式工廠排程問題（JSP）以及車輛途程問題（VRP)、資料探勘(DM)等等。其中首次應用螞蟻理論於資料分類的是 Parpinelli 等人，不過 Parpinelli 等人針對數值型資料乃是利用離散工具(C4.5 Discretisation)將其離散成類別型資料，再進行規則萃取，允許使用的運算符號只限等於(＝)以及交集(AND)。因此在本計畫第二年度將根據之前相同的概念，針對螞蟻發展一套有效率、符號變化性較多的分類器，並能同時處理數值型與類別型資料。

四、研究方法

第一年計劃將以基因演算法作為主要研究工具，針對數值型資料將不執行離散化步驟，希望能保有原始資料內所隱藏的資訊；針對數值型資料將可擁有小於等於（≦）以及大於等於（≧）的比較運算符號，而類別型資料將可擁有小於等於（≦)、大於等於（≧)、等於（＝）以及不等於（≠）的比較運算符號；至於邏輯運算符號部分將可擁有交集（AND）與聯集（OR）符號的選擇。希望透過較多的比較運算符號與邏輯運算符號的選擇，讓多個規則能結合成單一規則，減少產生規則的數目，以方便決策者使用。最後將 GA-Classifier 分類績效與傳統分類工具決策樹進行比較。本計劃將利用爪哇語言(JAVA)設計一個分類績效較傳統分類工具決策樹準確的 GA-Classifier。

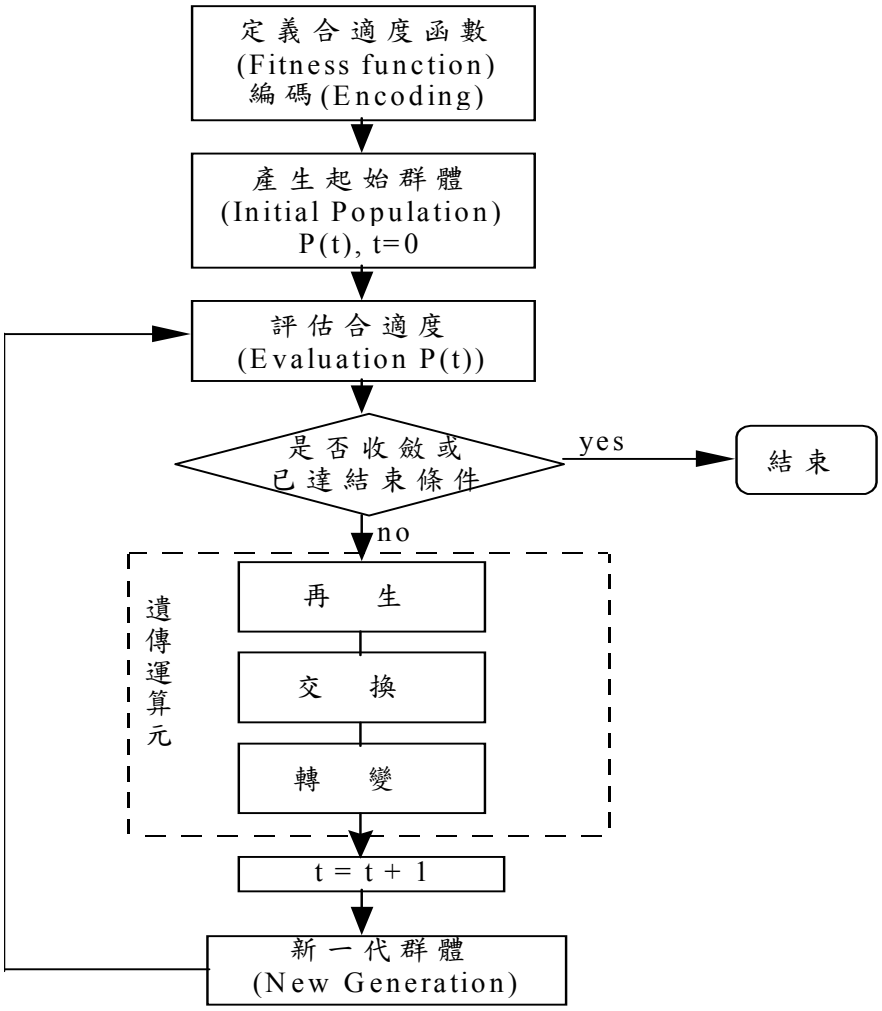
表二　GA-Classifier 功能說明

演算法	數值型資料處理方式	比較運算符號		邏輯運算符號
GA-Classifier	不離散化	數值型資料	類別型資料	AND、OR
		≦、≧	≦、≧、＝、≠	

基因演算法（genetic algorithm, GA）最早由 John Holland 等人在 1975 年首度發表，但直到 1980 年代後才逐漸有較多理論與應用之發展。GA 係將所有搜尋的參數轉換成另一種有限長度的表示式子（字串），再利用遺傳運算元產生新的下一代，配合評估的標準使子代具有比母代更好的表現。應用 GA 求解最佳化問題，須將問題目標轉化成對應的函數，稱為合適度函數（fitness function）；合適度函數代表系統對環境的適應能力，相當於系統的性能指標。GA 會將每一代中合適度高的解依據機率複製到下一代，再利用重組運算元（recombinational operator）去產生合適度更高的下一代，持續此反覆過程便可以逐步找到近似最佳解。這種演算的方式正與大自然的生物生存特性相似，產生更具適應能力的下一代以求在外部環境下能延續族群。

圖一為 GA 應用在搜尋最佳解問題的執行流程[王培珍, 1996]，以下針對其流程做說明：

(1) 定義合適度函數／編碼。
(2) 產生起始解。
(3) 評估目標值。
(4) 如果已達結束條件則停止，反之則執行步驟(5)。
(5) 遺傳運算元之運算。
(6) 產生新群體後，執行步驟(3)。



圖一　基因演算法流程圖[王培珍, 1996]

五、結果與討論

第一年計劃主要利用爪哇語言(JAVA)設計一個分類績效較傳統分類工具決策樹(Decision Tree)準確的 GA-Classifier。為了進一步驗證 GA-Classifier 分類的能力，本計劃首先採用四種常見的離散化技術：C4.5 discretization、Boolean reasoning algorithm、Entropy/MDL algorithm、Equal frequency binning，並使用三個標準資料庫 Cleveland、Australian、Iris 進行測試，以瞭解不同的離散化技術對

決策樹分類結果的影響性。表三為三個標準資料庫的屬性資料說明。本計畫針對每一個標準資料庫使用 Five-Folds Cross-Validation 方法測試四種離散化技術對決策樹分類結果之影響。表四列示各項離散化技術的平均分類錯誤率。從表四中，我們可以得知使用不同離散化技術進行連續型資料的離散，離散後資料經由決策樹分析後會得到不同分類結果。其中針對 Cleveland 資料庫，使用 C4.5 discretisation 技術的分類表現最佳；針對 Australian 資料庫，使用 Equal frequency binning 技術的分類表現最佳；針對 Iris 資料庫，使用 Boolean reasoning algorithm 技術的分類表現最佳。這也說明了使用不同的離散化技術會造成不同程度的資訊流失，因此有必要使用柔性演算法發展一能保有原有資料隱藏之資訊的智慧型分類器。

表三　資料庫說明

資料庫	類別型屬性數	連續型屬性數	資料筆數
Cleveland	7	6	303
Australian	8	6	690
Iris	0	4	150

表四　分類錯誤率

資料庫 / 離散化技術	Cleveland	Australian	Iris
C4.5 discretisation	***43.9%***	13.8%	6.0%
Equal frequency binning	47.2%	***12.2%***	5.3%
Entropy/MDL algorithm	44.2%	14.9%	6.7%
Boolean reasoning algorithm	44.6%	15.1%	***2.7%***

六、參考文獻

Bandar, Z, H. Al-Attar and D. McLean, "Genetic algorithm based multiple decision tree induction," *Proceedings of 6th International Conference on Neural Information Processing*, Piscataway, NJ, USA, pp.429-434 (1999).

Berry, M. J. A. and L. Gordon, *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley and Sons, New York, NY (1997).

Bruha, I., P. Kralik and P. Berka, "Genetic learner: Discretization and fuzzification of numerical attributes," *Intelligent Data Analysis*, vol:4, no:5, pp.445-460 (2000).

Cabena, P., P. O. Hadjinian, R. Stadler, J. Vehees and A. Zanasi, *Discoverying Data Mining from Concept to Implementation*, Prentice Hall, New York, NY (1997).

Collard, M, D. Francisci, "Evolutionary data mining: an overview of genetic-based algorithms," *Proceedings of 8th International Conference on Emerging Technologies and Factory Automation,* Piscataway, NJ, USA, pp.3-9 (2001).

Congdon, C. B., “Classification of epidemiological data: a comparison of genetic algorithm and decision tree approaches,” *Proceedings of the 2000 Congress on Evolutionary Computation*, Piscataway, NJ, USA, pp.442-449 (2000).

Dorigo, M., V. Maniezzo and A. Colorni, “Ant system: optimization by a colony of cooperating agents,” IEEE Transactions on Systems, Man, and Cybernetics—Part B:Cybernetics, vol:26, no:1, pp.29-41 (1996).

Fidelis, M. V., H. S. Lopes and A. A. Freitas, “Discovering comprehensible classification rules with a genetic algorithm,” *Proceedings of the 2000 Congress on Evolutionary Computation*, Piscataway, NJ, USA, pp.805-810 (2000).

Fu, Z. and F. Mae, “A computational study of using genetic algorithms to develop intelligent decision trees,” *Proceedings of the 2001 Congress on Evolutionary Computation,* Piscataway, NJ, USA, pp.1382-1387 (2001).

Hall, C., “The devil’s in the details: techniques, tools, and application for database mining and knowledge discovery part II,” *Intelligent Software Strategies*, vol:6, no:9, pp.1-16 (1995).

Lopes, F. M. and A. T. R. Pozo, “Genetic algorithm restricted by tabu lists in data mining,” *21st International Conference of the Chilean Computer Science Society,* Los Alamitos, CA, USA, pp.178-185 (2001).

Noda, E, A. A. Freitas and H. S. Lopes, “Discovering interesting prediction rules with a genetic algorithm,” *Proceedings of the 1999 Congress on Evolutionary Computation*, Piscataway, NJ, USA, pp.1322-1329 (1999).

Parpinelli, R. S., H. S. Lopes and A. A. Freitas, ”Data mining with an ant colony optimization algorithm,” *IEEE Transactions on Evolutionary Computation,* vol:6, no:4, pp.321-332 (2002).

Pozo, A. R. and M. Hasse, “A genetic classifier tool,” *Proceedings 20th International Conference of the Chilean Computer Science Society,* Los Alamitos, CA, USA, pp.14-23 (2000).

Shin, K. S. and Y. J. Lee, “A genetic algorithm application in bankruptcy prediction modeling,” *Expert Systems with Applications*, vol:23, no:3, pp.321-328 (2002).

王培珍，應用遺傳演算法與模擬在動態排程問題之探討，中原大學工業工程研究所碩士論文，1996.

七、計畫成果自評

本計畫第一年主要利用爪哇語言(JAVA)設計一個分類績效較傳統分類工具決策樹準確的 GA-Classifier，此 GA-Classifier 將能在不離散數值型資料狀況下，同時處理數值型與類別型資料。目前計畫執行進度已達預期目標，正在進行

GA-Classifier 程式驗證步驟。程式驗證完成時，即可與目前已執行完之決策樹分類結果進行比較與分析。本計畫相關執行人員，針對此計畫成果自評滿意，相信應能使用柔性演算法開發出高分類準確率之分類器。此一研究成果之學術價值非常適合在學術期刊發表。另外透過執行此項國科會計劃，相信也能使相關參與人員充分了解目前業界相當受重視的資料探勘技術，培植參與人員未來的就業的實力，同時參與人員的程式撰寫能力也將有顯著地提升。