區間離散化對約略集合的影響於分類上之研究

黃美玲
國立勤益科技大學
工業工程與管理系
e-mail：
huangml@ncut.edu.tw

汪亭妏
國立勤益科技大學
工業工程與管理系
e-mail：
cutedabby@hotmail.com

陳智凱
國立勤益科技大學
工業工程與管理系
e-mail：
ckchen123@gmail.com

摘要

約略集合理論近幾年在資料探勘領域中受到許多學者的注意，主要原因是因為約略集合論具有移除遺漏值、屬性化簡、產生分類規則等等的優點。約略集合在面對連續型數值時必須給予離散化的處理，而離散化對最後分類效果具有相當大的影響。為了驗證區間離散化對約略集合於分類上的影響，本研究採用兩種不同區間離散化的方法對約略集合理論作資料的前處理，之後再進行屬性篩選而產生規則。

關鍵詞：區間離散化、約略集合、資料探勘、基因演算法

Abstract

Rough set theory was widely used in the field of data mining recently. The main advantages of rough set include deleting missing values, variable reduction, and generating classification rules. Discretization should be applied on continuous data before classification , and different discretization method affect the classification results. This study uses two different discretization methods for data preprocessing for variable reduction and generating classification rules.

Keyword：discretization、rough set、data mining、genetic algorithm

1. 前言

資料探勘技術近年來逐漸受到重視，不僅是因為企業或研究社會科學的機構單位將它用在描述數據型態與結構理論上，更重要的是此技術將管理技巧與管理行為實際運用於各領域中，例如醫療診斷、製造業的客訴案件等[1]。運用資料探勘技術尋找具有價值的規則、樣式，並從其中解讀出有用的訊息，提供決策者做為參考，在資料探勘領域中最常接觸的領域為分類問題，其工具如類神經網路、支援向量機、約略集合等。

在上述這些工具裡面，約略集合論是資料探勘領域中一種新興資料分析技術，這幾年受到許多學者的注意，而應用在許多不同的領域中，根據 Palwlak（1997）[16]所整理其主要原因是：

1. 能夠提供有效率的演算法來尋找資料中隱藏的資料型態。
2. 找到資料的最小集合。
3. 評估資料的顯著性。
4. 從資料中產生決策規則的最小集合。
5. 容易了解。
6. 能夠直接解釋分析的結果。
7. 能夠同時使用數量與非數量資料進行分析。
8. 應用約略集合進行分析時，不需服從任何假設。
9. 獲得一般統計方法所無法得到的關係。
10. 大部分與約略集合結合的演算法都能夠並行來處理資料。

此外，約略集合和類神經網路、支援向量機最大差別在於約略集合可以產生分類規則數，讓使用者能清楚了解系統分類的過程。

約略集合論具備這些優點，但在實際的生活中，所收集到的資料，其資料型態可能會包含一些名目型和數值型，因此必須對其資料作區間離散化，如此一來，約略集合才可以產生規則做分類，所以區間離散化對於約略集合而言有當大的影響。本研究採用兩階段法（第一階段為布林搜尋法；第二階段為等頻率法）和熵法這兩種區間離散化方法對約略集合作資料前處理，再結合基因演算法進行屬性篩選，再依篩選後的屬性以約略集合產生分類規則。研究發現應用熵法離散化有助於約略集合在分類準確性的提升。

2. 文獻回顧

2.1 約略集合論

約略集合論是由 Pawlak 於 1985 年提出，約略集合論又稱作粗糙集，該理論採用等價類別的概念來分割訓練例子，作為處理資料分類問題的一種數學工具[2]。執行約略集合的首先要建立資訊系統(Information funtion)，又稱為資訊表(Information table, IT)資訊系統可表示如下：

IT=(U，A) (1)

以下介紹各符號代表意義：

U：全域(Universe)，是一個非空集合，表示資料表中的所有物件的集合。

A：屬性(Attribute)的集合。存在一個屬性 $a \in A$，則可定義一資訊函數(Information function) $f_a: U \to V_a$，其中 Va 為 a 的屬性值集合，稱作 a 的屬性域(Domain of attribute a)。

τ：全域 U 及屬性 A 的交集。

以表一資訊表為例，若以數學式呈現結果如下所示：

U={n_1, n_2…n8}；

A={a_1, a_2, a_d}；

V={V_{a1}, V_{a2}, V_d}={{0, 1}, {0, 1, 2}{0, 1}}；

τ(n_1, a_1)={1}。

表一：資訊表

Objects	Attributes(a_i)		Decision(a_d)
(n_i)	a_1	a_2	d
n_1	1	0	0
n_2	1	1	1
n_3	1	2	1
n_4	0	0	0
n_5	0	1	0
n_6	0	2	1
n_7	0	1	1
n_8	0	2	0

在表一中，事件 n_5 及 n_7 在屬性 P={a_1, a_2} 中具有不可辨關係，所謂不可分關係即當個體之間為不可區分時，則稱具有不可分辨關係(Indiscerniblility relation)[14]。可以表示如下：

ρ(n_5, a_1, a_2)= ρ(n_7, a_1, a_2)={1, 1}。

若採用屬性集合{a_1, a_2}是無法辨別所有事件，尤其當 P = A, 在資訊表中 A-elementary 所產生的集合，以屬性 a_1 可以產生以下的集合：

{a_1}-elementary sets

E_1={n_1, n_2, n_3} forρ(x, a_1)={1}。

E_2={n_4, n_5, n_6, n_7, n_8} forρ(x, a_2)={0}。

將決策屬性集合(Decision-elementary sets)同樣以集合的概念(Concepts)表示如下：

C_1={n_1, n_4, n_5, n_8} => Class = 0 (d=0)。

C_1={n_2, n_3, n_6, n_7} => Class = 1 (d=1)。

更進一步可以發現，表一中事件 n_5 及 n_7 屬性值 a_1 及 a_2 相同，但所產生的決策屬性 d 卻不相同，而同樣的衝突同樣發生在事件 n_6 及 n_8。意思是指當相同事件卻無法經由屬性 a_1 及 a_2 精確判斷。

決策值為不一致的情況發生時，可藉由下限近似(Lower approximation)、上限近似(Upper approximation)與近似空間(Approximation space)來定義資料的不確定性。

事件 Y 在全域 U 所有確定目標族群的描述值；R 為一集合，滿足下限近似條件，以公式 2 表示下限近似：

$$\underline{R}(C) = \cup\{Y \in U / R : Y \subseteq C\} \quad (2)$$

事件 Y 在全域 U 可能屬於目標族群的描述值；R 為一集合，滿足上限近似條件，以公式 3 表示上限近似：

$$\overline{R}(C) = \cup\{Y \in U / R : Y \cap C \neq \phi\} \quad (3)$$

由上限近似及下限近似相減所得的差，稱為邊界區域(Boundary region)或稱為懷疑區域(Doubtful area)，如公式 4 所示：

$$BN_R(C) = \overline{R}(C) - \underline{R}(C) \quad (4)$$

以表一為例，首先求得概似集合 C_1 在決策屬性 d=0 的上、下限近似及邊界區域：

下限近似 => $\underline{R}$ (C_1)={n_1, n_4}；

上限近似 => $\overline{R}$ (C_1)={n_1, n_4, n_5, n_6, n_7, n_8}；

邊界區域 => $BN_R(C_1)$={n_5, n_6, n_7, n_8}= $BN_R(0)$。

同理；概似集合 C_2 (d=1)，結果表示如下：

下限近似 => $\underline{R}$ (C_2)={n_2, n_3}；

上限近似 => $\overline{R}$ (C_2)={n_2, n_3, n_5, n_6, n_7, n_8}；

邊界區域 => BN_R (C_2)={n_5, n_6, n_7, n_8}= $BN_R(1)$。

若 $BN_R = \varphi$，代表上限近似 $\overline{R}(C)$ 與下限近 $\underline{R}(C)$ 完全相同時，即可完整定義集合 C，但若有模糊的情況發生，則又另外對不精確的情況有以下四種之不同的定義，如表二所示：

表二：約略集合定義之類型

上、下限近似集合關係	定義名稱
$\overline{R}(C)\neq\psi$ 且 相等於 $\underline{R}(C)\neq\psi$	Definable
$\underline{R}(C)\neq\psi$ & $\overline{R}(C)\neq U$	Roughly-definable
$\underline{R}(C)=\psi$ & $\overline{R}(C)\neq U$	Intermally-indefinable
$\underline{R}(C)\neq\psi$ & $\overline{R}(C)=U$	Externally-definable
$\underline{R}(C)=\psi$ & $\overline{R}(C)=U$	Totally-indefinable

2.2 區間離散化

區間離散化的概念是將屬性值域劃分為區間，資料離散化技術可以用來減少給定連續屬性值的個數。區間的標記可以替代實際的資料值。用少數區間標記替換連續屬性的數值，從而減少和簡化了 原來的資料。這導致挖掘結果的簡潔、易於使用的、知識層面的表示。離散化技術可以根據如何進行離散化加以分類；可以分成全域和區域、靜態和動態、監督式和非監督式、參數式和非參數式、嚴格和模糊這幾類。

簡單來說區間離散化指的是將連續型資料轉換成離散型資料的一個過程。而離散化的好處根據唐文政（2004）[5]所整理如下：

1. 可以很清楚的呈現資料。
2. 使得各種資料探勘方法對資料和結果的 解釋能力更好。
3. 讓更多演算法能讀取資料。

常見的的離散化方法有等頻率法（Equal-frequency binning）、布林邏輯法（Boolean reasoning algorithm）、熵法（Entropy method）……等。

2.3 熵理論

熵(Entropy)的提出是由一位德國物理學家 Rudolph Clausius 於 1854 年提出[10]。熵是一種廣延量，其量值由一定熱力學狀態物質的量決定。孤立系統的熵只會增加不會減少，此一現象有時也被說成是熱力學第二定律。所以 Entropy 也是一個熱力學的專有名詞。熵理論可以從物理和資訊這兩種角度做解釋，從物理理論角度解釋，可以說明熱力學系統的演化方向、熱平衡的達成與否亦或是代表系統的混亂程度等。從資訊理論角度解釋，資訊熵則代表量測資訊系統的可信度或者是忽略度[6]。

「資訊熵」與「物理熵」是息息相關的。舉例而言，當我們將 1 莫耳（mol）理想氣體的體積在恆溫下壓縮一半時，「物理熵（ΔS）」減少了，但由於分子分布的空間縮小，其位置的確定性增高，因此「資訊熵（ΔI）」增加了，如果我們認定「資訊熵」的增加與「物理熵」的減少是相同的（$\Delta I = -\Delta S$），則取得一位元的資訊相當於每莫耳理想氣體溫度上升 1K 所需能量（J／mol・K）[10]。

熵理論近年來受到許多學者應用在許多領域，Fayyad & Irani[12]於1993年提出得到最佳分割點可用熵法，利用class-entropy 的概念找出所有屬性之最佳分割。本研究利用此概念來對連續型資料做分割。

2.4 基因演算法

基因演算法(Genetic Algorithms, GA)這個技術，是由學者 John Holland[13]於 1975 年提出的，它的原理主要參考了達爾文進化論—「物競天擇，適者生存」，其演算的流程是經由模仿生物演化特性，讓母代隨著每一代的變動來進行演化，其中主要包括「選擇(Selection)」、「交配(Crossover)」及「突變(Mutation)」這三種運算。經由保留較佳之染色體，淘汰較差之染色體的方法，產生「適者生存」的效果，使較好的染色體能擁有較高的機率將其基因遺傳至子代中[13]。因為 GA 具有簡單運算與平行處理的特性，使其成為解決多項式困難(NP-Hard)最佳化問題的最佳方法之一。雖然 GA 在搜尋的過程不易陷入局部最佳解，卻無法保證找出真正的全域最佳解，而是找出近似最佳解，因此比較適合解決多項式困難的複雜問題[5]。

GA 的主要有選擇(Selection)、交配(Crossover)、突變(Mutation)等三種運算。而在執行運算之前，除了要決定染色體的編碼方式以外，還有適應函數以及族群大小、代數等參數。

2.5 約略集合之相關運用

本研究所使用的資料庫是針對醫療診斷方面之研究，因此另外整理出將約略集合理論應用於醫學領域之文獻表，應用於心血管疾病、癌症、燒燙傷等相當多不同的醫療領域都曾使用此理論，將文獻整理於表三中。

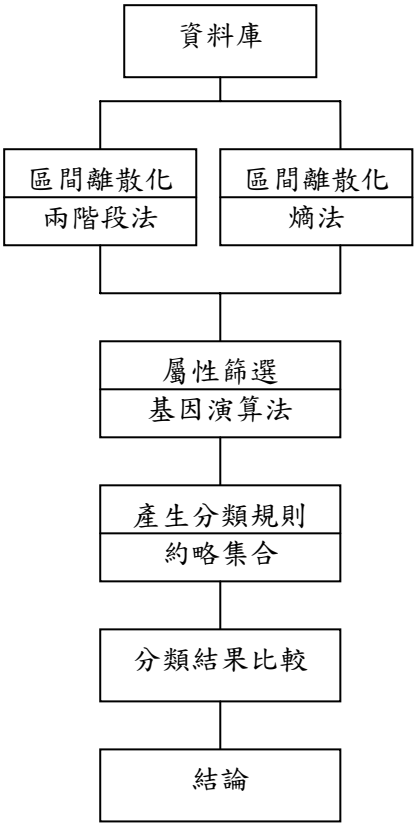
表三：約略集合理論應用於學之文獻表

學者	論文題目	應用領域
吳建興 [4]	以約略集合與決策樹萃取危險因子–以逆流性食道炎為例	逆流性食道炎
李昆鴻 [3]	醫學影像之診斷–應用粗集合理論與類神經網路	極座標靶心圖
Tsumoto[17]	Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model	頭痛、腦膜炎、腦血管疾病
雷賀君 [8]	前十字韌帶傷害快速診斷系統–以粗略集合、基因演算法與倒傳遞網路為工具	前十字韌帶
賴威利 [9]	利用約略及合理論預測燒燙傷患者死亡率	燒燙傷分級

3. 研究方法

3.1 研究流程

本研究分別採用兩種不同離散化方法先給予資料前處理，之後再結合基因演算法進行屬性篩選，再依篩選後的屬性以約略集合產生分類規則，其研究流程如下圖一所示：



圖一： 研究流程圖

3.2 兩階段法

3.2.1 第一階段－布林搜尋法（Boolean reasoning algorithm）

學者 Nguyen 及 Skowron[15]於 1995 提出布林搜尋法與約略集合相結合的一種離散方法，其步驟及例子如下所示，表四為連續型態的資訊表，共有 7 筆事件、兩個條件屬性及一個決策屬性所構成[9]：

表四：連續型態資訊表

Objects (n_i)	Attributes		Decision
	a	*b*	*d*
n_1	0.8	2	1
n_2	1	0.5	0
n_3	1.3	3	0
n_4	1.4	1	1
n_5	1.4	2	0
n_6	1.6	3	1
n_7	1.3	1	1

● 步驟一：首先先以集合表示所有屬性值 a(U)={0.8, 1, 1.3, 1.4, 1.6}；b(U)={0.5, 1, 2, 3}。

● 步驟二：以符號C_j^i表示屬性 j 的第 i 個間隔

C_1^a=[0.8, 1]； C_2^a=[1, 1.3]； C_3^a=[1.3, 1.4]；

C_4^a=[1.4, 1.6]； C_1^b=[0.5, 1]； C_2^b=[1, 2]；

C_3^b=[2, 3]。

● 步驟三：對每兩條決策不同的紀錄 i, j 用上述符號的並聯式(Disjunctive, ∨)φ(i, j)進行表示：

φ(2, 1)= $C_1^a \vee C_1^b \vee C_2^b$

φ(2, 4)= $C_1^b \vee C_3^a\ C_1^b$

φ(2, 6)= $C_2^a \vee C_3^a \vee C_4^a \vee C_1^b \vee C_2^b \vee C_3^b$

φ(2, 7)= $C_2^b \vee C_1^b$

φ(3, 1)= $C_2^b \vee C_1^a\ C_3^b$

φ(3, 4)= $C_3^a \vee C_2^b \vee C_3^b$

φ(3, 6)= $C_3^a \vee C_4^a$

φ(3, 7)= $C_2^b \vee C_3^b$

φ(5, 1)= $C_1^a \vee C_2^a \vee C_3^a$

φ(5, 4)= C_2^b

φ(5, 6)= $C_4^a \vee C_3^b$

φ(5, 7)= $C_3^a \vee C_2^b$

● 步驟四：將所有的並聯式用串連一般式 CNF(Conjunctive normal form)進行表示：

$\varphi^s=(C_1^a \vee C_1^b \vee C_2^b) \wedge (C_1^b \vee C_3^a\ C_1^b) \wedge (C_2^a \vee C_3^a \vee C_4^a \vee C_1^b \vee C_2^b \vee C_3^b) \wedge (C_2^b \vee C_1^b)$

$(C_2^b \vee C_1^a\ C_3^b) \wedge (C_3^a \vee C_2^b \vee C_3^b) \wedge (C_3^a \vee C_4^a) \wedge (C_2^b \vee C_3^b) \wedge (C_1^a \vee C_2^a \vee C_3^a) \wedge C_2^b \wedge (C_4^a \vee C_3^b) \wedge (C_3^a \vee C_2^b)$

● 步驟五：將串聯一般式化為並聯一般式 DNF(Disjunctive normal form)：

$\varphi^s=(C_2^a \wedge C_4^a \wedge C_2^b) \vee (C_2^a \wedge C_3^a \wedge C_3^b \wedge C_2^b) \vee (C_3^a \wedge C_1^b \wedge C_2^b \wedge C_3^b) \vee (C_1^a \wedge C_4^a \wedge C_1^b \wedge C_2^b)$

● 步驟六：在並聯一般式中的串聯式(Conjuctive, ∨)即為離散化結果：

以$(C_2^a \wedge C_4^a \wedge C_2^b)$來表示其中之一組的分割集合，將表一離散化後得到表五之結果。

表五：離散型態資訊表(Boolean reasoning algorithm)

Objects (n_i)	Attributes		Decision
	a	*b*	*d*
n_1	0	1	1
n_2	0	0	0
n_3	1	1	0
n_4	1	0	1
n_5	1	1	0
n_6	2	1	1
n_7	1	0	1

布林搜尋法的缺點是時間複雜度和空間複雜度高，以及資料量大的情況下，此法就不適用[11]。

3.2.2 第二階段－等頻率法（Equal frequency binning）

以表四的條件屬性 a 為例依序介紹 Equal frequency binning 之步驟：

● 步驟一：首先先定義欲產生之區間 k，找出屬性之最大值($v_{\max}^a$)及最小值($v_{\min}^a$)：

假設區間個數 k=3；最大值$v_{\max}^a$ 為 1.6；最小值$v_{\min}^a$=0.8。

● 步驟二：求區間寬度及分割點，這些分割點距離相等：

區間寬度：$\omega = \frac{v_{\max}^a - v_{\min}^a}{k}$ (5)

分割點：$p_i^a = v_{\min}^a + i * \omega$，i=1, 2, …k-1 (6)

屬性 a 區間寬度

$\omega_a = \frac{v_{\max}^a - v_{\min}^a}{k} = \frac{1.6-0.8}{3} = 0.27$

分割點 p_1^a=0.8+1*0.27=1.07

分割點 p_2^a=0.8+2*0.27=1.34

產生了三個區間：{(-∞, 1.07), (1.07, 1.34), (1.34, ∞)}={0, 1, 2}

屬性 b 區間寬度

$\omega_b = \frac{v_{\max}^b - v_{\min}^b}{k} = \frac{3-0.5}{3} = 0.83$

分割點 p_1^b=0.5+1*0.83=1.33

分割點 p_2^b=0.5+2*0.83=2.16

產生了三個區間：{(-∞, 1.33), (1.33, 2.16),(2.16, ∞)}={0, 1, 2}

- 步驟三：根據步驟二將表一連續型態的資訊表轉成離散型態的資訊表。

步驟二分別求出每個條件屬性的區間分段值，將屬性值找出位於哪個區間並轉換，即可得到離散型態的屬性值，如表六所示：

表六：離散型態資訊表(Equal frequency binning)

Objects (n_i)	Attributes		Decision
	a	*b*	*d*
n_1	0	1	1
n_2	0	0	0
n_3	1	2	0
n_4	2	0	1
n_5	2	1	0
n_6	2	2	1
n_7	1	0	1

3.3 熵法（Entropy）

假設某屬性*a*(U)={ 109, 105, 113, 107,107, 105 }為例依序介紹 Entropy 離散化之步驟[18]：

- 步驟一：首先將樣本集合內的樣本依照屬性 *a* 值依小排到大：

a(U)={105, 107, 107, 109, 113, 115}

- 步驟二：對於每一個候選分割點都可以依據公式（7）算出分割後的類別資訊熵：

$$E(a,c_i;U)=\frac{|U_1|}{|U|}Ent(U_1)+\frac{|U_2|}{|U|}Ent(U_2) \quad (7)$$

U_1, U_2 分別為 c 點左邊及右邊的集合。

其中：$Ent(U)=-\sum_{i=1}^{k}p(c_i,U)\log_2 p(c_i,U)$ 代表 *U* 的類別資訊熵。

$p(c_i,u)=\frac{|\{e\in U | e.u=c_i\}|}{|U|}$ 代表類別 *Ci* 的樣本在 U 子集合中的比例；$|U|$ 代表樣本集合總和。

下列為分割點 107 和 113 之類別資訊熵為：

$$E(a,107;U)=\frac{1}{6}(-1\cdot\lg 1)+\frac{5}{6}(-\frac{3}{5}\cdot\lg\frac{3}{5}-\frac{2}{5}\cdot\lg\frac{2}{5})=0.809$$

$$E(a,113;U)=\frac{4}{6}(-\frac{1}{4}\cdot\lg\frac{1}{4}-\frac{3}{4}\cdot\lg\frac{3}{4})+\frac{2}{6}(-1\cdot\lg 1)=0.541$$

同理可以算出其他分割點的類別資訊熵。

- 步驟三： 選出資訊熵值最小者為最佳分割點。

3.4 屬性篩選

原本資料庫內包含 13 個條件屬性，本研究結合基因演算法找出重要之屬性。在基因演算法之參數設定上，針對交配率(Crossover rate)及突變率(Mutation rate)採試誤法的方式找出能使分類結果較好之組合。交配率選擇了 0.3 及 0.5；突變率共設定四種情況，分別為 0.5、0.1、0.15 及 0.2，此設定會產生八種不同的組合。以下為固定參數，如選擇單點交配、族群大小(Population size)為 100、在選擇機制(Selection)使用菁英法(Elitism)，強迫遺傳演算法在每一代中保留一些最好的染色體直接複製到下一代基因。

根據交配率及突變率之選擇上，選擇能使整體正確率為最高之組合，其結果交配率選擇 0.3 及突變率為 0.1。

3.5 評估

本研究使用「UCI Machine Learnong -Repoitory」資料庫，資料庫中保括不同領域的資料，選用一組應用於醫學方面的資料庫進行分析，此資料庫為 Heart-disease（心臟病）中的 Cleveland（克里蘭夫）資料庫。

在此資料庫共有 13 個條件項目，其決策項目共有 2 種分類結果 ，其原本受測者有 303 位，但因為有些資料有遺漏的情況發生所以予以刪除之後剩下 297 位受測者。之後再將資料分成訓練組和測試組，如表七所示，其中訓練組佔資料的百分之七十，測試組佔資料的百分之三十。

表七：測試與訓練分組表

	罹患心臟病	未罹患心臟病	總計
訓練組	97	111	208
測試組	40	49	89
總計	137	160	297

同一個資料庫不同方法的離散後再經過屬性篩選之後，哪一種離散化方法對於約略集合在分類準確度有所提升，透過最後分類的正確率可以作為評估的準則。

4. 實證分析

本研究所使用的醫學資料庫中有 5 個是連續型資料，在此利用兩階段法和熵法做離散化，其離散化的結果如表八和表九所示：

表八：兩階段屬性離散化表

屬性	原始形態	第一階段 Boolean reasoning algorithm	第二階段 Equal frequency binning
Y1	連續型	[*, 57), [57,*) =(0, 1)	[*, 57), [57,*) =(0, 1)
Y2	離散型（0，1）	離散型（0，1）	離散型（0，1）
Y3	離散型 (1，2，3，4)	離散型 (1，2，3，4)	離散型 (1，2，3，4)
Y4	連續型	連續型	[*,123),[123,139),[139,*) =(0, 1, 2)
Y5	連續型	[*, 235),[235, *) =(0, 1)	[*,235),[235,) =(0, 1)
Y6	離散型 (0，1)	離散型（0，1）	離散型（0，1）
Y7	離散型 (0，1，2)	離散型 (0，1，2)	離散型 (0，1，2)
Y8	連續型	[*, 151), [151,*) =(0, 1)	[*,151), [151,*)=(0, 1)
Y9	離散型 (0，1)	離散型（0，1）	離散型（0，1）
Y10	連續型	連續型	[*,0.2), [0.2,1.4), [1.4, *) =(0, 1, 2)
Y11	離散型 (1，2，3)	離散型 (1，2，3)	離散型 (1，2，3)
Y12	離散型 (0，1，2，3)	離散型 (0，1，2，3)	離散型 (0，1，2，3)
Y13	離散型 (1，2，3)	離散型 (1，2，3)	離散型 (1，2，3)

表九：熵法屬性離散化表

屬性	原始形態	Entropy algorithm
Y1	連續型	[*,71), [71,77), [77,*) =(0, 1, 2)
Y2	離散型（0，1）	離散型（0，1）
Y3	離散型 (1，2，3，4)	離散型 (1，2，3，4)
Y4	連續型	[*, 186), [186, *)=(0, 1)
Y5	連續型	[*,276),[276,277), [277,280), [280,295), [295,299), [299,301), [301,310), [310,312), [312,319), [319,320), [320,322), [322,324), [324,326), [326,338), [338,341), [341,342), [342,348), [348,354), [354,401), [401,413), [413,*)=(0, 1, 2, 3, 4, 5, 6, 7, 8,9,10,11,12,13,14, 15, 16, 17, 18,19,20)
Y6	離散型 (0，1)	離散型 (0，1)
Y7	離散型 (0，1，2)	離散型 (0，1，2)
Y8	連續型	[*,148), [148,151), [151,162), [162,170), [170,172), [172,175), [175,176), [176,178), [178,183),[183,.195), [195,199),[199,*)= (0,1,2,3,4,5,6,7, 8,9,10,11)
Y9	離散型 (0，1)	離散型 (0，1)
Y10	連續型	[*, 3), [3, 5),[5,*) = (0, 1,2)
Y11	離散型 (1，2，3)	離散型 (1，2，3)
Y12	離散型 (0，1，2，3)	離散型 (0，1，2，3)
Y13	離散型 (1，2，3)	離散型 (1，2，3)

離散化之後再進行屬性篩選，本研究使用基因演算法找出重要屬性，其屬性篩選的結果如表十及表十一所示：

表十：兩階段離散化資料經基因演算法之屬性篩選表

	全部屬性	經過基因演算法
條件屬性之項目個數	13	9
條件屬性項目	y1～y13	y1、y2、y3、y4、y5、y7、y11、y12、y13

表十一：熵法離散化資料經基因演算法之屬性篩選表

	全部屬性	經過基因演算法
條件屬性之項目個數	13	11
條件屬性項目	y1～y13	y2、y3、y5、y6、y7、y8、y9、y10、y11、y12、y13

兩階段法離散化後經過屬性篩選剩下 9 個重要屬性，熵法離散化後經過屬性篩選剩下 11 個重要屬性，將這些重要屬性透過基因演算法群找最小集合，以最小集合產生分類規則數，在依據這些分類規則進行分類，而分類結果如表十二和表十三所示。

表十二：兩階段離散化測試組之分類正確率

	罹患心臟病	未罹患心臟病	正確率(%)
罹患心臟病	39	10	79.59
未罹患心臟病	5	35	87.50
正確率(%)	88.64	77.78	83.15

表十三：兩階段離散化訓練組之分類正確率

	罹患心臟病	未罹患心臟病	正確率(%)
罹患心臟病	111	0	100
未罹患心臟病	0	97	100
正確率(%)	100	100	100

表十四：熵法離散化測試組之分類正確率

	罹患心臟病	未罹患心臟病	正確率(%)
罹患心臟病	49	0	100
未罹患心臟病	0	40	100
正確率(%)	100	100	100

表十五：熵法離散化訓練組之分類正確率

	罹患心臟病	未罹患心臟病	正確率(%)
罹患心臟病	109	2	98.2
未罹患心臟病	1	96	98.9
正確率(%)	99	97.96	98.56

表十六： 正確率之比較

離散化 / 組別	兩階段法之正確率(%)	熵法之正確率(%)
訓練組	100	98.56
測試組	83.15	100

由表十六分類正確率的比較可以看出利用熵法進行離散化有提高分類準確度之效用。

5. 結論

利用約略集合作為資料探勘工具時，面對連續型的資料型態，首先必須給予資料作離散化，才能產生規則，而離散化對於最後分類準確度有很大影響。因此必須慎選離散化的方法。本研究採用兩種離散化方法做比較，結果發現，熵法離散化之分類結果比兩階段離散化的分類準確度更高。

參考文獻

[1] 李金鳳，2001，“資料探勘面面觀”，資訊與教育雜誌，頁 10-19，8 月。

[2] 李昆鴻，2003，“醫學影像之診斷–應用粗集合理論與類神經網路”，東海大學，工業工程學系，碩士論文

[3] 林君娥，2003，“利用約略集合論學習跨階層的確定性和可能性規則”，義守大學，資訊管理系，碩士論文。

[4] 吳建興，2002，“以約略集合與決策樹萃取危險因子–以逆流性食道炎為例”，華梵大學，資訊管理學系碩士班，碩士論文。

[5] 唐文政，2004，“基因演算法應用於約略集合理論之屬性化簡及屬性值離散化”，華梵大學，資訊管理學系，碩士論文。

[6] 曾致遠，“淺談最大熵原理和統計物理學”

[7] 陳亮宇，2004，“一個針對分群問題的關係基因演算法之原理與應用”，國立中央大學，資訊管理研究所，碩士論文。

[8] 雷賀君，2004，“前十字韌帶傷害快速診斷系統-以粗略集合、基因演算法與倒傳遞網路為工具”，大葉大學，工業工程學系碩士班，碩士論文。
[9] 賴威利，2005，“利用約略及合理論預測燒燙傷患者死亡率”，南台科技大學，國際企業系，碩士論文。
[10] 《科學發展》2004 年 5 月，377 期，34-37 頁
[11] Aleksander,ψhrn., 1999, “Discernibility and Rough Sets in Medicine: Tools and Applications”, NTNU report 1999:133, IDI report 1999:14, December.
[12] Fayyad, U.M., Irani, K.B.,“On the Handling of Continuous-Valued Attributes in Decision Tree Generatiuon.” Machine Leaning, Vol.8, pp.87-102,1992.
[13] Holland, J. H., 1975 “Adaptation in Natural and Artificial Systems,” Ann Arbor: The University of Michigan Press.
[14] Komorowski, j., Pawlak, Z., Polkowski, L., and Skowron, A.,“Rough Sets: A Tutorial”, http://www.idi.ntnu.no/~janko/rstutorial.pdf, Poland.
[15] Nguyen, H.S., and Skowron A., 1995, “Quantization of Real Values Attributes ”, In Proc, 2nd International Joint Conference on Information Sciences, edited by Wang P. P., Durham: Duke University, pp.34-37.
[16] Pawlak, Z ., 1997,“Rough Sets: Theoretical Aspects of Reasoning about Data about Data ”, Kluwer Academic Publishes.
[17] Tsumoto, T., 2004,“Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model”, Information Sciences: an Inter- national Journal, Vol: 162, pp.65-80.
[18] http://pc33.mis.hfu.edu.tw/DM/ppt/ch3.ppt