行政院國家科學委員會專題研究計畫　成果報告

連續資料粒子化程度於知識探勘效能之優劣分析

計畫類別：個別型計畫
計畫編號：NSC91-2416-H-168-002-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：崑山科技大學資訊管理系

計畫主持人：高淑珍

計畫參與人員：鄭琇馨,麥靜怡

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 2 月 10 日

行政院國家科學委員會專題研究計畫成果報告

連續資料粒子化程度於知識探勘效能之優劣分析

The impact of level of information granularity to knowledge discovery performance: an empirical study

計畫編號：NSC 91-2416-H-168 -002
執行期限：91 年 8 月 1 日至 92 年 7 月 31 日
主持人：高淑珍　　崑山科技大學資訊管理系

一、中文摘要

運用連續性資料之間斷化技術的主要目的是為了提高資料探勘技術的效能，包括精確度、效率、精簡度、以及實際效果。本研究專題係以三種不同的模糊隸屬函數來轉換連續性資料並求得其隸屬強度值，最後再比較其探勘結果精簡度的優劣。此三種模糊隸屬函數分別為採用有四分之一重疊的三角形函數、降低粒子化程度(Decreasing information granularity, DIG)、與提高粒子化程度(Increasing information granularity, IIG)，其中 DIG 將採正弦函數(Sine function)為轉換函數，而 IIG 則將以三角形為中心，對稱於 DIG 所構成的函數為基礎。本研究搜集了十八個實務性資料庫做為測試之用，探勘法係採 ID3。比較結果發現：DIG 有比較好的效能。
關鍵詞：資料探勘、間斷化、模糊隸屬函數

Abstract

Knowledge discovery has been successfully used in acquiring domain knowledge from large databases. It is a required process to transform continuous data into linguistic ones. The transformation function used in the previous research was triangle membership function. However, the impact of other type of information granularity remains unknown. It is believed that different level of information granularity will result in different knowledge discovery performance. In this research, two tasks were conducted. The first one contained four processes. They were collecting real-life datasets, granulizing continuous datasets via Increasing Information Granularity (IIG) and Decreasing Information Granularity (DIG), discovering decision rules via ID3, and documenting the mined results. The DIG was based on sine function while IIG the function that is centered on the triangle function and symmetrical to the DIG. The second task focused on the empirical evaluation where eighteen real-life datasets were utilized. The result indicated that DIG showed a better performance.
Keywords: Data mining, Granulation, membership function

二、緣由與目的

資料挖掘的主要目的是能在複雜繁瑣的資料堆中挖掘出可用的資訊，藉以輔助決策行動，並進一步增加企業知識，提昇企業的競爭優勢。根據 Chen 等人的研究 [1]，資料挖掘的參考與進行步驟如下：1.理解資料與進行的工作;2.獲取相關知識與技術（Acquisition）;3.融合與查核資料（Integration and checking）;4.去除錯誤或不一致的資料（Data cleaning）;5.發展模式與假設（Model and hypothesis development;6.實際資料挖掘工作;7.測試與檢核所挖掘的資料（Testing and verification）;8.解釋與使用資料（Interpretation and use）。從八個步驟來看，資料挖掘牽涉大量的規劃與準備，以系統的觀點來看，一階段的輸出即成為另一階段的輸入，於是每階段所使用的技術都將影響到最終的探勘結果。目前大多數的研究都著重在技術層面的探討與改進，例如引導式(supervised)與非引導式(unsupervised)粒子化的比較分析、挖掘機制之歸納法與統計法的比較分析，這些研究確實大大的提昇了資料挖掘的層次[2]。

以歸納法為基(induction-based)的探勘機制是目前應用較廣的技術，但有一個先決條件是連續性資料必須先經過粒子化才能以此法淬取出決策規則 [3, 4, 5]，，換言之，必須先以一個語意名稱來代表一段連續性資料。由於粒子化的內函是將某段資料壓縮在一組，然後用一個語意性資料來代表，例如人體體溫在 38.5 到 39.5 的所有連續性資料都以"高"來表示。非引導式是一種比較簡單方便而且完全依原始資料考量的分離方式，傳統已發展出等距法(EWI)及等頻法(EFI)兩種，等距法根據所蒐集到的資料，將每一屬性值之下的資料分成間隔相等的固定組數，然後編成一個轉換表，再依據此轉換表做資料的轉換。等頻法首先決定出組數，然後把

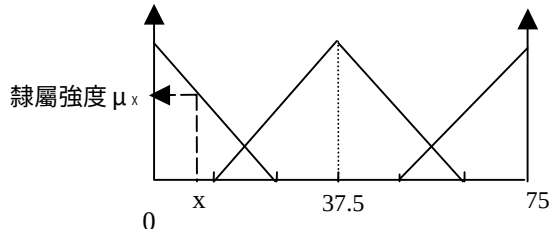
所有資料除以組數以得知每一組的資料筆數，最後將原資料依序排列，再依據每一組應有的資料筆數逐一做轉換。定義模糊函數做不同資料參與強度的轉換而提出了一個同屬非引導式的粒子化方法 EMFI，同時也已證明在淬取優質決策規則(即低條件數、高支持數)方面優於 EWI。它是透過定義模糊隸屬函數而使得不同的連續性資料對某一語意資料，其轉換後的隸屬強度介於 0 和 1 之間，因此而影響淬取後的結果。

採用模糊隸屬函數係透過提高粒子化程度的方法來做連續性資料的轉換[6]。此法也好可以提昇知識挖掘的效能，同時，我們相信，不同的粒子化程度會大大的影響最終的淬取結果，於是除了原四分之一重疊的三角形隸屬函數以外，本研究將試著比較 ” 降低 ” 以及 ” 增加 ” 粒子化程度對淬取效能的影響，以進一步瞭解選擇粒子化轉換函數的正確方向。本研究專題將以三角形隸屬函數為中心，利用可以降低粒子化程度(Decreasing Information Granularity, DIG)的 Sine 隸屬函數，以及可以提高粒子化程度(Increasing Information Granularity, IIG)且對稱於 Sine 所構成的隸屬函數為實驗主題，為求各類隸屬函數在相同的重疊基礎下做比較，DIG 與 IIG 都將設計為有四分之一重疊的轉換函數，而挖掘機制則採用 ID3 演算法，比較的方法則為精簡度。

三、研究方法

(一) 間斷化函數

1. 三角型函數



例如上圖的語意層次共有三層，且假設資料庫中含由一個名為「皮膚角化程度」的屬性，其最大值為 75，最小值為 0 ，則單位間距為 75/(2*3) = 12.5，我們並以普遍使用的三角模糊函數為轉換基準。其中，每個函數均由兩條線性方程式所決定，而每一線性方程式則由兩點座標決定，當間距決定之後，各線性方程式的點座標也都能決定：假設線性方程式為 $y = b_1x + b_0$ 且兩點座標分別為 (x1, y1) 與 (x2, y2)。

$$b_1 = \frac{\sum_{i=1}^{2} x_i y_i - \frac{\left(\sum_{i=1}^{2} x_i \sum_{i=1}^{2} y_i\right)}{2}}{\sum_{i=1}^{2} x_i^{\,2} - \frac{\left(\sum_{i=1}^{2} x_i\right)^2}{2}}$$

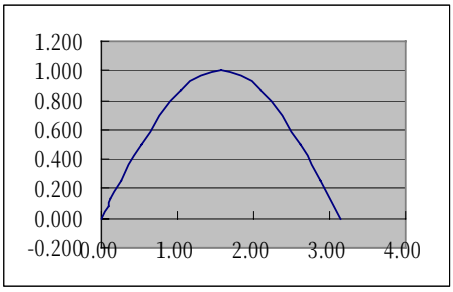
$$b_0 = \frac{1}{n}\sum_{i=1}^{2} y_i - b_1 * \frac{1}{n}\sum_{i=1}^{2} x_i$$

以第一個函數為例，兩點座標分別為(0.00, 1.00) 與 (25.00, 0.00)，則線性函數為 y = -0.04x + 1.00。當所有線性函數都決定出來之後，接下來便是決定它們所適用的 x 範圍，而由於模糊函數有重疊的部分，未來某一個觀察值可能會落在其中，於是我們採用模糊和(即取最大值，因為我們所要的是該筆資料所能提供的最大資訊量)，最後我們便可以得到如下的對照表：

	觀察值 x 的範圍	隸屬函數值	
n = 3	0.00<= x < 18.75	M1	y = -0.04x + 1.0
	18.75 <= x < 37.5	M2	y = 0.04x - 0.5
	37.50 <= x < 56.25	M2	y = -0.04x + 2.5
	56.25 <= x < 75.00	M3	y = 0.04x - 2.0

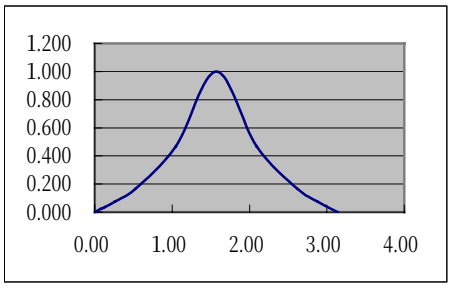
於是所有的數值性資料便可以依此表轉換成間斷性資料，例如當觀察值 x = 20 時，其轉換的結果為 M2(0.3)，即 x = 20 隸屬於 M2 函數，且隸屬程度為 0.3。

2. DIG 隸屬函數



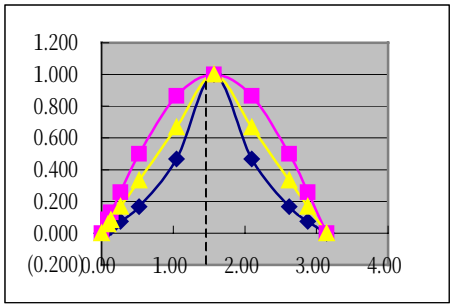
如上圖所示，Sine 在 Π/2 為 1.0，在 0 及 Π 為 0，Π = 3.1416。但由於實際資料並不一定會在該範圍之內，故在做轉換時，先對原資料做正規化(Normalization)，且須限制在第一和第三象限，使其值大於 0。最後的隸屬程度值則與上述相同。

2. IIG 隸屬函數



由於 IIG 隸屬函數(設其值為 y_2)以三角形隸屬函數為中心 ($y = b_1x + b_0$)，且對稱於 Sine 所構成的隸屬函數($y_1 = \sin(x)$)，故 $y_2 = 2*(b_1x + b_0) - \sin(x)$。

如果結合三角形隸屬函數、DIG、以及 IIG，可以得到如下的轉換函數：



(二) 探勘機制

ID3 是在 1986 年由 Quinlan 所提出的 [5]，它利用各屬性對類別所得到的資訊獲取量 (Gained information)來決定分解類別的能力。它是由以下的式子所決定的：

Equation (A)：I (nc_1,nc_2,…,nc_n) =

$$(-\frac{nc1}{N}\log_2\frac{nc1}{N})+(-\frac{nc2}{N}\log_2\frac{nc2}{N})+\ldots.$$

$$+(\frac{ncn}{N}\log_2\frac{ncn}{N})$$

nc_i：結論是屬於 c_i 的觀察紀錄數目，i=1,2…n

N ：總觀察紀錄數目。

Equation (B)：

$$E(A)=\sum_{i=1}^{m}(\frac{n_vi}{N})I(n_1v_1c_1,n_2v_2c_2,\ldots n_nv_nc_n)$$

E (A)：屬性 A 的訊息量(假設 A=outlook)

m ：屬性 A 的可能值的數目(m=3)

n_{vi} ：屬性 A 的值是 v_i 的觀察紀錄數目

$n_iv_ic_i$：屬性 A 的值是 v_i 且其結論是屬於 c_i 的觀察紀錄數目。

(三) 評估機制

我們用 18 個資料庫分別加以測試，使用三角型隸屬函數、DIG 及 IIG 三種不同的間斷法以及同一個探勘方式（ID3），並求得各演算法所產生的決策樹的精簡度，由此來討論其優劣，而比較的基準為決策樹的精簡度，其公式如下：

$$S_{DT}=\sum_{i=1}^{m}\left(\frac{1}{nc_i}\right)(ns_i)$$

S_{DT}：決策樹的精簡度；nc_i：第 i 個規則的條件數; ns_i：第 i 個規則的資料筆數；I=1，2，…，m.

四、實驗結果

本研究使用 18 個資料庫分別加以測試，並在處理器 Pentium 200 的電腦中對每個資料庫做處理。其結果如附表一所列。實驗結果顯示：DIG 有比較好的間斷效能。

五、結論

雖然許多研究對連續性資料轉換為間斷性資料對資料探勘的效能，有了很深入的研究，許多研究者也都努力的在探討如何有效又精確的淬取到有價值的決策規則，然而此些研究並未對隸屬函數的粒子化程度對探勘效能做更深入的分析探討，本研究專題在此動機下，比較 DIG、三角隸屬函數、以及 IIG 三種轉換函數的探勘效能，以對未來間斷化的隸屬函數之粒子化程度有所遵循。

我們深信我們週遭每天都會產生無數的資料，這些資料必定含有許多很有價值的知識，資料挖掘技術未來必能發揮淬取知識的功能，更多的研究也勢將提昇淬取知識的質與量。本研究除了提昇知識淬取技術的利用價值之外，也期望不同的資訊技術能在決策品質的提昇與知識的累積上發揮功能。

六、參考文獻

[1] Chen, M.S., Han, J. & Yu, P.S. (1996), “Data Mining: An Overview from a Database Perspective”, IEEE Trans. on Knowledge and Data Engineering, Vol. 8, pp. 866-883.

[2] Dougherty, J., Kohavi, R. and Sahami, M., “Supervised and Unsupervised Discretization of Continuous Features”, Proceedings of 1995 International Conference on Machine Learning, pp.194-202, 1995.

[3] Sestito, S. & Dillon, T., (1994), Automated Knowledge Acquisition, Prentice Hall, Englewood Cliffs, NJ.

[4] Quinlan, J.R. 1986, “Induction of decision tree”, machine Learning, 1, pp. 81-106.

[5] Imielinski, T & Mannila, H. 1996. A Database Perspective on Knowledge Discovery, Communication of ACM, 39(11), 58-64.

[6]Wu, C.H. & Kao, S.C., (2002), “An induction-based approach to rule generation using membership function”, International Journal of Computer Integrated Manufacturing, Vol. 15, No. 1, pp. 86-96.

附表二

探勘結果表

資料庫名稱	DIG	三角隸屬函數	IIG
B_c_data198	45.4146	44.3012	47.9607*
B_c_data569	96.0135	96.8499*	96.8416
B_c_data699	186.6905*	186.6905*	186.6905*
Bupa	50.3833*	50.3833*	50.3833*
Glass	38.8083*	24.5583	21.8119
Ionosphere	102.4644*	101.5349	99.3966
Iris	61.0000*	61.0000*	61.0000*
Letter	436.7277*	424.4113*	421.4073*
Pendigits3498	254.4496*	254.4496*	250.8310
Sattlite	344.1644	319.1459	311.0472
Segmentation	32.0000*	32.0000*	32.0000*
Shuttle	7.0000*	7.0000*	7.0000*
Sonar	51.8556*	50.7167	50.8722
Synthetic	89.7595*	89.6286	89.6286
Vehicle	111.2723	111.7307	110.8854*
Vowel	82.4167*	82.4167*	82.4167*
Waveform	1696.6667*	1696.6667*	1696.6667*
Wine	57.1667*	57.1667*	57.1667*