行政院國家科學委員會專題研究計畫成果報告

診斷資料粒子化之相容程度 SQL 於資料挖掘之應用

Data mining: a SQL-based approach to tuple consistency recognition for granulized datasets

計畫編號：NSC 90-2416-H-168 -007
執行期限：90 年 8 月 1 日至 91 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

許多探討連續性資料之間斷化技術的研究，其主要目的是為了提高資料探勘技術的效能。然而，在間斷化的過程中，不相容資料的比率卻並未被加以探討，導致所得到的結論並不是非常可靠。本研究專題以此為動機，探討連續性資料間斷化後之資料不相容程度，藉此鼓勵相關研究在研究過程中能考慮這個問題，以提高研究結論的可靠度。本研究搜集了十個實務性資料庫做為測試之用。間斷法係採等距法，每個區間以三十筆為基礎，最多不超過七個區間。實驗結果發現：有 50%的資料庫含有不相容的資料；其中有 30%的資料庫，其不相容程度超過 15%。
關鍵詞：資料探勘、間斷化、不相容資料

Abstract

In this study, the problem of inconsistent tuples that exist in the granulized datasets was considered. It is believed that too high degree of tuple inconsistency will definitely decrease the reliability of conclusion in the data mining related research. An empirical investigation where 10 real life continuous datasets were used was conducted to confirm the study concern. The technique of equal width interval was employed to transform the continuous data. The remarkable results showed that 50% of the used datasets contained inconsistent tuples; 30% contained more than 15% inconsistent tuples.
Keywords: Data mining, Granulization, inconsistency

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]。而由於資料庫有各種不同的欄位數、資料筆數、以及結論數，使得粒子化的結果很有可能存在太多的不相容資料，若超出某個比例，則此資料庫所探勘出來的結果就不可靠了，尤其是在做技術提昇的比較時，更會使結論被誤導，所以連續性資料庫粒子化結果的診斷工作就變的益加重要了。目前的文獻雖然在資料挖掘的各階段都有很深入的研究，但如果能指出其對連續性資料庫粒子化相容度之診斷結果，進而使用粒子化相容度較高的資料庫，必能使結論更具可信度。

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，對於數量龐大的資料庫而言，往往會產生許多不相容資料錄(tuple)，不相容資料錄表示有共同條件值(即屬性值)，但會產生不同結論。這些不相容資料對挖掘機制而言是沒有意義的，雖然在日常生活中，我們常常會遇到相同資訊卻得到不同結論，但對歸納法之資料挖掘機制而言，我們必須也只能接受相同資料只能得到共同的結論的假設。雖

然如此，但這個觀念是合理的，例如以下的敘述是沒有意義的：”你順著這條路往下走，在第一個紅綠燈右轉，你可能「會」也可能「不會」找到你的小狗”，此類「相同的條件卻有截然不同的結論」的資料對我們而言是沒有任何幫助的，所以應用挖掘機制之前必須檢測每個資料庫之資料錄的相容程度，如此方能確保所得到的探勘結果是可靠的。

然而，由於每個資料庫所包括的屬性數、資料筆數，結論數、以及各個屬性所可能反應出來的資料值都不一樣，使得資料不相容性的檢驗工作更加的複雜。為了能有效處理這個問題，Han 等人[5]、Meo 等人[6]、以及 Imielinski 等人[7]分別成功的應用結構化查詢語言(Structured Query Language, SQL)於 DM 的工作上，SQL 係為一資料庫查詢語言，它可以根據使用者的需求，由大型資料庫中快速且準確的取得相關的資料。本研究計畫也將深入探討 SQL 用於分離不相容資料的應用效能。

三、研究方法

(一) 間斷化技術

1. EWI（Equal Width Interval，等距法）

等距法係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。首先在所有的資料內找出最大值（x_{max}）及最小值（x_{min}），依其筆數（N）分成 3、5 或 7 組(k)，它可以由以下的公式做轉換：

$\delta=(x_{max}-x_{min})/k$

δ：間距

x：連續性資料值

k：組數（3、5 或 7 組，總筆數＜150 筆者，組數為 3；150＜總筆數＜210 者，組數為 5；總筆數＞210 者，組數為 7）

例：$x_{max}=108$，$x_{min}=24$，$k=7$

則 $\delta=(x_{max}-x_{min})/k=(108-24)/7=84/7=12$

決定 k 的方式：

k=Min (7 , int(N/30))

N：資料總筆數

例如：當資料總筆數為 150 筆時，Min (7 , int(150/30)) = 5，

又若資料總筆數為 10000 筆時，Min (7 , int(10000/30)) = 7。

（二）診斷機制

先探討 SQL 的結構與應用優點，將一個有不相容的資料庫分解成如下的部分集合，然後再探討檢測模式的建構內容：

S_o = {T_{Oi} | T_{Oi} is a tuple that belongs to the original dataset, i = 1, 2, 3, …, N; N 是間段資料的筆數}.

S_s = {T_{SCj} | T_{SCj} is a tuple of which the combination of attribute values is unique, j = 1, 2, 3, …, n; n<=N}.

S_m = {T_{MCk} | T_{MCk} is the k^{th} subset, its tuples have identical combinations of attribute values and the number of tuple in T_{MCk} is equal to or greater than 2, k = 1, 2, 3, …, m; m < N}.

S_{Same} = {T_{MCSp} | T_{MCSp} is the p^{th} subset in S_m, its tuples have the same conclusion, p = 1, 2, 3, …, q; q <= m}.

於是，所得到的不相容資料集便可由下列式子得到：

$$S_D = (S_o - S_s) - S_{Same}$$

然而，上述式子並非只是一個簡單的代數問題而已，就如前面提到的，資料庫所包括的屬性數、資料筆數，結論數、以及各個屬性所可能反應出來的資料值都會使問題複雜化，何況本研究並非只限定在某個資料庫的檢測，而是要能適用於所有的資料庫，也就是所謂的一般化。所以我們初步將透過三個 SQL 程序將不相容資料分離出來，包括群組功能(Grouping)、分離功能(Separating)、以及合併功能(Joining)。此三個程序如附件一所列。

（三）實驗結果

本研究使用 10 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 20 次的處理，取其平均值做為處理時間。其結果如附表二所列。實驗結果如下：有 50%的資料庫含有不相容的資料；其中有 30%的資料庫，其不相容程度超過 15%。本研究確認有很大比例不相容資料確實存在於間斷化後的資料中。

四、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，然而此些研究並為對間斷後的資料相容程度有所分析，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。本研究專題除了分析不相容資料對資料探勘效能的影響，同時也發展一個不相容資料檢測模式，以提昇資料探勘的應用價值。雖然本研究對間斷化相容程度提供了一個檢驗的方式，然而確也存在許多的限制，例如其他間斷法的節果如何？決定 k 的方式有有何影響，這些都有待做進一步的探討。

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。

五、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] Han, J., Fu, Y., Koperski, K. & Zaiane, O. 1996. DMQL: A Data Mning Query Language for relational Databases, DMKD-96 (SIGMOD-96 Workshop on KDD), Montreal Canada.

[6] Meo, R., Psaila, G. & Ceri, S. 1998. An Extension to SQL for Mining Association Rules, Data Mining & Knowledge Discovery, 2(2), 195-224.

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

附表一：SQL 程序

程序 #	SQL 作業	SQL 內容 (pseudo format)	描述
1	作業一	SELECT *, COUNT(dataset.a); FROM S_O; GROUP BY all_attribute + class; ORDER BY all_attribute + class; INTO TABLE S_A	Let S_O be the original granulized dataset; Create a temporary dataset S_A that lists the number of tuples of which the conditions and conclusion are the same.
	作業二	SELECT *, COUNT(dataset.b); FROM S_A; GROUP BY all_attribute; ORDER BY all_attribute; INTO TABLE S_B	Create a temporary dataset S_B from S_A that lists all kinds of combinations of conditions and their number of occurrences.
2	作業三	DELETE FROM S_B ; WHERE cnt_b<2	Create a temporary dataset S_C from S_B that lists all combinations of conditions for inconsistent tuples.
3	輸出結果	SELECT * ; FROM S_A , S_C ; INTO TABLE $S_{DifClass}$; WHERE S_A.all_attribute = S_C.all_attribute	Return the final subset $S_{DifClass}$ that contains all inconsistent tuples.

附表二

表二：測試資料庫與檢查不相容資料錄的結果

資料庫編號	資料庫名稱	資料筆數	屬性數目	結論的種類	不相容筆數	不相容比例	處理時間(秒)
1	Bupa	345	6	2	136	39.4203	0.2641
2	BC198	195	33	2	0	0	0.3612
3	Glass	214	9	7	74	34.5794	0.1578
4	Iris	150	4	3	23	15.3333	0.1350
5	Pendigit	3498	16	10	0	0	1.1564
6	Sattlite	2000	36	6	0	0	2.2547
7	Symthetic	600	61	6	0	0	0.9851
8	Vehicle	846	18	4	6	0.7092	0.3684
9	Waveform	5000	21	3	0	0	1.7942
10	Wine	178	13	3	0	0	0.2301