利用分類技術發掘具有購買次序之產品項目最適性的消費者

陳垂呈　　董志源　　吳閔慧
南台科技大學資訊管理研究所
E-mail: ccchen@mail.stut.edu.tw, {n9290012, m9390229}@webmail.stut.edu.tw

摘要

隨著資訊技術的發展，企業可以更輕易地記錄、儲存消費者的交易資料、及分析消費者的購物傾向，使得企業有機會由傳統單向的大量行銷方式，轉變成客製化、適性化與互動式的行銷方式，這對改善企業與顧客之間的關係、提升顧客的忠誠度與滿意度、及擴展市場利基都有顯著的影響。在本篇論文中，我們以消費者之交易資料為探勘的資料來源，每一筆交易資料除了記錄有消費者曾經購買的產品項目，也記錄著其購買的次序性，並以某 d 個產品項目為探勘的目標，$d \geq 1$，利用分類(classification)技術來發掘具有購買次序之此 d 個產品項目最適性的消費者。在探勘的過程中，我們對包含有此 d 個產品項目的交易資料，將其在購買此 d 個產品項目之前的產品項目依序地作分解，並視分解後的項目組為欲分類的屬性，利用 ID3 演算法來對交易資料進行分類分析，以建構出一棵決策樹。根據決策樹所顯示出的特徵，我們可以找出那些項目屬性會影響購買此 d 個產品的意願為高，藉此做為發掘具有購買次序之此 d 個產品項目最適性的消費者特徵的依據。我們根據所提出的方法，設計與建置一個探勘系統，以發掘具有購買次序之產品項目最適性的消費者。此探勘結果，對企業在擬訂具有購買次序之產品的行銷策略時，鎖定產品最適性之消費者將可以提供非常有用的參考資訊。

關鍵詞：資料探勘、分類分析、購買次序。

1. 簡介

資訊技術的進步、資料儲存媒體容量的增加及價格快速地滑落，導致企業儲存消費者曾經購物的交易資料，已成為一件既輕易、快速又便宜的事[2]。這些交易資料可能來自於消費者的信用卡交易記錄、超級市場的收銀機、消費者填寫的特徵資料、或是網頁的瀏覽記錄等。因此，如何利用這些大量的交易資料，深入分析消費者的交易行為，以改善與顧客之間的關係，並提供最貼切的產品服務、提昇顧客的滿意度與忠誠度，是企業經營者必須思考的問題之一。

資料探勘(data mining)是從大量資料中找出有用的資訊與知識，目前已廣泛的應用在各領域中[2]，並已被證明可以有效地應用在產品行銷、銷售及顧客服務上，是企業提昇經營優勢與競爭力的重要工具之一[3, 4]。

在本篇論文中，我們以消費者之交易資料為探勘的資料來源，每一筆交易資料記錄有消費者曾經購買的產品項目及其購買的次序，並以某 d 個產品為探勘的目標，$d \geq 1$，利用分類(classification)技術來發掘具有購買次序之此 d 個產品項目最適性的消費者。在探勘的過程中，我們將曾經購買過此 d 個產品之消費者的交易資料，設定其購買此 d 個產品項目的意願度為「高」，否則設定其意願度為「低」。然後，我們對包含有此 d 個產品項目的交易資料，將其在購買此 d 個產品項目之前的產品項目依序地作分解成各項目組，並視分解後的項目組為欲分類的屬性，然後對消費者之交易資料進行分類分析。我們依據 ID3 演算法所建立的決策樹，可以找出那些項目屬性會影響購買此 d 個產品項目的意願為高，我們即定義這些項目屬性為「影響因子」。在消費者的交易資料中，我們依據其是否包含有「影響因子」，做為發掘具有購買次序之此 d 個產品項目最適性的消費者特徵的依據。

我們根據所提出的方法，設計與建置一個探勘系統，以發掘具有購買次序之產品項目最適性的消費者。本篇論文的探勘結果，對於擬訂具有購買次序之產品的行銷策略，可以在鎖定產品最適性之消費者的問題上，提供企業非常有用的參考資訊。

本篇論文的架構如下：下一節中，我們介紹資

料探勘技術、及其具有產品次序之應用的相關研究；第3節中，我們以某些產品為探勘的目標，說明利用分類技術來發掘具有購買次序之這些產品最適性的消費者的探勘過程，並以一個實例做說明；第4節中，我們根據所提出的方法，設計與建置一個探勘系統，以發掘具有購買次序之產品最適性的消費者；最後，我們在第5節中做一結論。

2. 相關研究

資料探勘(data mining)是從大量資料中挖掘出潛在有用的資訊與知識，發現專家尚且未知的新關係，以提供給企業專業人員參考。資料探勘可完成以下任務或是更多：關聯規則(association rules)、分群(clustering)、分類(classification)、次序相關分析(sequential pattern analysis)及預測等[5]，利用資料探勘技術於企業從事行銷決策及市場預測等活動時，可以提供非常有價值的參考資訊。在過去的研究中，次序相關是最主要被使用來分析消費者購物之次序性的探勘技術之一，其將每一消費者的交易資料，視為一群有次序性的購物行為，次序相關分析的目的就是在擷取具有次序性之最常出現的項目組，且其包含有最大的產品項目個數，其相關研究可參考[6, 7]。

分類分析是從已知的物件群中，根據所訂立的屬性條件來進行分類，決策樹(decision trees)與決策法則(decision rules)是分類分析最常用的兩種表示法。例如，我們可對消費者曾經購買過的產品項目來進行分類，把消費者對某一產品的購買意願分為「高」與「低」兩種類別，再將消費者之交易資料中其他產品項目視為影響屬性來進行分類分析，便可得知影響購買意願之高低的關鍵屬性。

在資料進行分類分析時，一般可以產生出許多的分類模式，但其期望得到的分類模式是越精簡越好。以決策樹為例，若決策樹的高度愈小，則表示可用愈少的屬性便能分類出所有物件。因此，一個好的分類技術，應該具有精簡與預測能力佳的特性，目前常被利用的分類技術有 ID3[8]、CN2[9]、倒傳遞類神經網路(back- propagation)[10]等。在本篇論文中，我們以消費者之包含有購買次序的交易資料為探勘的資料來源，並以某 d 個產品項目為探勘的目標，$d \geq 1$，利用分類技術來發掘具有購買次序之此 d 個產品項目最適性的消費者。

3. 發掘具有購買次序之產品項目最適性的消費者

從消費者有購買次序性的交易資料中，可顯示出消費者是依序本身的需求及產品的特性來有次序性地購買產品項目。在此一章節中，我們以消費者之具有購買次序的交易資料為探勘的資料來源，並某 d 個產品項目為探勘的目標，$d \geq 1$，利用分類技術做為發掘具有購買次序之此 d 個產品項目最適性的消費者的方法依據。此章節共分為兩小節如下：第 3.1 節中，我們說明發掘具有購買次序之產品項目最適性的消費者的探勘方法；第 3.2 節中，我們以一實例做說明。

3.1 探勘方法

在分類技術的方法中，ID3 演算法是最常被使用來建構決策樹的分類方法之一，其目的是選擇最佳的屬性來當作節點，以建構出的決策樹為一最簡單狀態、或接近最簡單狀態。最佳節點是依據其節點所產生的熵值(entropy)所決定，其計算方式如下：

若某一物件集合 C，其物件分屬於 j 個不同類別，則此物件集合之熵值 $E(C)$ 為：

$$E(C) = -\sum_i P_i \log_2 P_i \quad .....(1)$$

其中 C 為物件集合；I 為類別數；P_i=(屬於類別 i 的物件總數)/(C 的物件總數)。

接下來選擇某一屬性 X_j 為決策樹節點，在此節點下建立 m 個子節點，並將原本屬於節點的所有物件，分配至具有適當的子節點下。而分配至相同子節點的物件，其屬性 X_i 值必為相等。故以 X_j 為節點所產生的子決策樹熵值 $E(X_j)$ 為：

$$E(X_j) = \sum_k (n_k/n) \times E(C_k) \quad .....(2)$$

其中 C_k 為物件集合 C 中其 X_j 屬性相同的物件子集合 k；$E(C_k)$ 為物件 C_k 的熵值；n 為物件集合 C 的總物件數；n_k 為物件子集合 C_k 的物件數。

資訊收益(information gain)是原來物件集合的熵值與 X_j 為決策樹子節點的熵值間得差距，其公式如下：

$$G(X_j)=E(C)-E(X_j)\text{…….(3)}$$

根據熵值和資訊收益，我們將 ID3 演算法的計算過程說明如下[1, 8]：

(1) 設立決策樹的根節點為 *C*，此時所有物件都屬於 *C* 的物件集合。
(2) 若 *C* 中所有的物件都屬於同一類別，則定義 *C* 節點為此類別並停止，否則繼續執行步驟(3)。
(3) 對屬於 *C* 的所有物件，分別計算其熵值 $E(C)$。
(4) 從根節點至目前節點中，若有尚未當過節點的屬性 X_j，則以 X_j 對 *C* 物件集合進行分割，並分別計算部分決策樹的熵值 $E(X_j)$及資訊獲利 $G(X_j)$。
(5) 選擇具有最大資訊獲利的候選屬性，並當做 *C* 節點的分類屬性。
(6) 在 *C* 節點下建立子節點分別為 C_1、C_2、…、C_m(假設選擇了 *m* 個屬性值做為分類屬性)，並將 *C* 中的所有物件集合，分派至適合的子節點中。
(7) 對每個子節點 C_i 當做節點 *C*，$1\le i\le m$，並由(2)重覆執行。

在本篇論文中，我們以 ID3 演算法做為建構決策樹的方法依據，並以某 *d* 個產品項目為探勘的目標，假設此 *d* 個產品項目為 *X*，*X* 為包含有 *d* 個產品項目且為有次序性的項目組，$d\ge 1$。在探勘的過程中，我們將曾經購買過 *X*-產品之消費者的交易資料，設定其購買意願為「高」，否則設定其購買意願為「低」，然後對包含有 *X*-產品之交易資料中產品項目進行以下的處理：

假設 $B_1B_2B_3…B_i$ 依序分別為 *X*-產品之前所購買的產品項目，$i\ge 1$，其會影響是否購買 *X*-產品的前序產品，我們依序分解這些產品項目成為以下項目組：B_i, $B_{i-1}B_i$, …, $B_2B_3…B_i$, $B_1B_2B_3…B_i$。在分解的過程中忽略 *X*-產品之後所購買的產品項目。

經由對所有交易資料進行前述的處理，並將分解後的項目組視為影響屬性，若交易資料中有出現影響屬性的項目組，則設定為曾經購買此項目組，否則設定為未曾購買此項目組。

接下來，我們對消費者之交易資料進行分類分析，依據 ID3 演算法所建立的決策樹，可找出那些的影響屬性會影響 *X*-產品的購買意願為高，我們即稱這些影響屬性為「影響因子」，其顯示若消費者曾經購買過「影響因子」之產品項目，則具有購買 *X*-產品意願高的傾向。因此，在消費者的交易資料中，若包含有「影響因子」，則稱之為具有購買次序之 *X*-產品最適性的消費者。

3.2 實例說明

我們以一實例來說明發掘具有購買次序之產品項目最適性的消費者的探勘過程。假設{A, B, C, D, E, F, G, H}為全部的產品，有一包含 16 筆交易資料的資料庫，分別記錄有消費者曾經購買過之產品與其次序，如表 1。假設目前欲探勘之產品為 A，則發掘具有購買次序之 A-產品最適性的消費者的探勘過程說明如下。

表 1 交易資料庫

編號	產品項目的購買順序
1	EDBAE
2	DEGB
3	GDCAF
4	DEC
5	FDCB
6	FDCAE
7	FDCAB
8	GDEC
9	ECB
10	DCABE
11	DHB
12	EHGB
13	DCABG
14	CBADH
15	BDCAF
16	GEFB

若交易資料中包含有 A-產品，則分解 A-產品以前所購買的產品項目成為各有次序的項目組，並

不考量 A-產品之後所購買的產品項目。經由對所有交易資料進行前述的分解過程，並將分解後的項目組視為影響屬性，若交易資料中有出現影響屬性的項目組，則表示曾經購買此項目組，並以“x”作標示；否則設定為未曾購買此項目組並以“〇”作標示。交易資料經由上述的處理資之後，可轉換成如表 2 之交易資料庫。

表 2 交易資料庫

項目組 / 編號	EDB	DB	B	GDC	DC	C	FDC	CB	BDC	購買意願
1	〇	〇	〇	×	×	×	×	×	×	高
2	×	×	〇	×	×	×	×	×	×	低
3	×	×	×	〇	〇	〇	×	×	×	高
4	×	×	×	×	×	〇	×	×	×	低
5	×	×	〇	×	×	×	×	〇	×	低
6	×	×	×	×	〇	〇	〇	×	×	高
7	×	×	×	×	〇	〇	〇	×	×	高
8	×	×	×	×	×	〇	×	×	×	低
9	×	×	〇	×	×	×	×	〇	×	低
10	×	×	×	×	〇	〇	×	×	×	高
11	×	×	〇	×	×	×	×	×	×	低
12	×	×	〇	×	×	×	×	×	×	低
13	×	×	×	×	〇	〇	×	×	×	高
14	×	×	〇	×	×	×	×	〇	×	高
15	×	×	×	×	〇	〇	×	×	〇	高
16	×	×	〇	×	×	×	×	×	×	低

〇：曾經購買過　　×：未曾購買過

接下來，我們說明發掘具有購買次序之 A-產品最適性的消費者的探勘過程。首先，我們利用公式(1)，計算物件集合的熵值，在表 2 中共有 8 位對購買 A-產品意願高和 8 位對購買 A-產品意願低的消費者，計算所有消費者的熵值，其計算如下：

E(所有消費者)
$= -(8/16)\log_2(8/16)-(8/16)\log_2(8/16)=1$

接著計算出各個屬性下其子節點的熵值，以購買產品 EDB 為例，其屬性值的熵值計算如下：

E(EDB-曾購買)
$= -(1/1)\log_2(1/1)-(0/1)\log_2(0/1)= 0$

E(EDB -未曾購買)
$= -(7/15)\log_2(7/15)-(8/15)\log_2(8/15)$
$= 0.996792$
E(EDB)$= (1/16)\times E$(EDB-曾購買)$+(15/16)$
$\times E$(EDB -未曾購買)$= 0.934492$

最後，再利用公式(3)來計算 EDB 項目屬性的資訊收益：

G(EDB)$= 1- E$(EDB) $= 0.065508$

依此類推，可求出所有產品統計項目屬性的資訊收益如下：

G(DB)$= 0.065508$

G(B)= 0.188722

G(GDC)= 0.065508

G(DC)= 0.548795

G(C)= 0.188722

G(FDC)= 0.137925

G(CB)= 0.018791

G(BDC)= 0.065508

由於 DC 項目屬性的資訊收益最大,所以把 DC 做為該決策樹的根節點，其餘項目屬性依照此方法一直做到所有節點下的消費者都屬於同一類別為止。將所有分類屬性依照先後順序連接起來，就形成如圖 1 的決策樹。

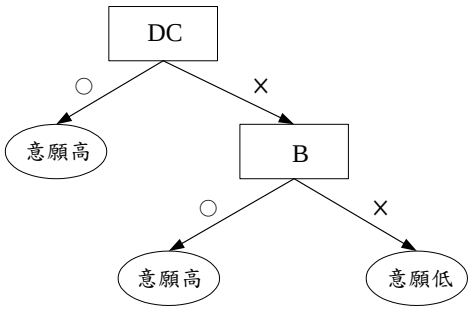

圖 1 分類決策樹

從圖 1 所建立的決策樹中，可發掘具有購買次序之 A-產品最適性的消費者的「影響因子」為：DC 及 B；其顯示若消費者曾經購買過 DC 或 B，則具有購買 A-產品意願高的傾向，因此，若消費者之交易資料中包含有 DC 或 B，則稱之為具有購買次序之 A-產品最適性的消費者。

4. 探勘系統之設計與實作

我們將前一章節所描述的探勘方法，應用到發掘具有購買次序之產品項目最適性的消費者的探勘系統實作上。我們以Delphi 6.0為撰寫的程式語言，在不失一般性的條件下，假設產品項目全部共有26項，分別以A, B, C, …, Z來表示之，並以亂數隨機產生每一筆交易資料，每一筆交易資料包含有最多10個有次序性購買的產品項目，共產生100筆交易資料，以下為此一系統探勘的執行過程。

圖2為此一探勘系統的交易資料庫，包含「交易編號」及「產品購買順序」等欄位資料。

圖 2 交易資料庫

圖3表示探勘畫面中包含有的欄位：

(1) 交易資料欄位：顯示出欲探勘之產品在購買之前的各產品項目，若交易資料未包含有欲探勘之產品，則顯示出全部曾經購買的產品項目。

(2) 項目次序型樣欄位：顯示出全部交易資料經由分解處理之後的項目組。

(3) 探勘產品欄位：為輸入欲探勘的產品項目。

(4) 探勘資料欄位：表示交易資料經由處理之後，準備分類分析的各項目屬性的資料。

(5) 探勘結果：顯示出建構之決策樹的決策路徑。

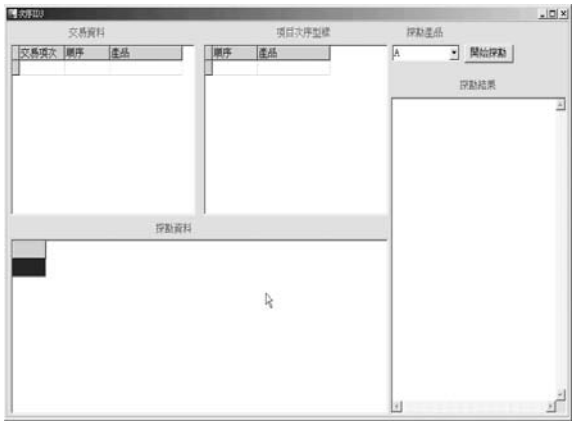

圖 3 探勘系統畫面

假設目前欲探勘之產品為A，並點選「開始探勘」的功能按鈕，經由第3章節演算法的探勘過程，可在「探勘結果」欄位中顯示出探勘的結果，如圖4。

圖 4 探勘結果的執行畫面

5. 結論

在目前企業的經營環境中，資訊技術已扮演著非常重要的角色，其功能除了協助支援企業平常的交易處理，企業的決策上也提供了非常有用的資訊與知識。在消費者曾經購買產品的記錄中，可從其購買的時間性，反映出消費者對產品有其次序性的需要特徵，若能從這些資料中找出那些的消費者對於某 d 個產品項目有其購買次序性的需求，$d \geq 1$，對企業經營者鎖定此 d 個產品項目最適性的消費族群，必可提供相當有用的資訊。在本篇論文中，我們以某 d 個產品項目為探勘的對象，利用分類技術來發掘具有購買次序之此 d 個產品項目最適性的消費者。從資料的蒐集、分析、方法的設計、及結果推導出的消費者特徵，顯示出我們所提出之探勘方法具有實務應用及方法創新的學術價值。此探勘結果，對企業經營者在擬訂具有購買次序之產品的行銷策略時，發掘產品最適性之消費者將可以提供非常有用的參考資訊。

參考文獻

[1] 魏志平、董和昇。電子商務理論與實務(2 版)。華泰書局，頁 167-205，2002。

[2] K. C. Laudon and J. P. Laudon, Management Information Systems, Sixth Edition, Upper Saddle River, NJ: Prentice Hall, 2000.

[3] M. J. A. Berry and G. Linoff, Data Mining Techniques for Marketing, Sales, and Customer Support, New York: John Wiley, 1997.

[4] S. C. Hui and G. Jha, "Data Mining for Customer Service Support," Information and Management, vol. 38, pp. 1-13, 2000.

[5] M. S. Chen, J. Han, and P. S. Yu, "Data Mining: An Overview from a Database Perspective," IEEE Trans. on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, 1996.

[6] R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proceedings of the International Conference on Data Engineering (ICDE), 1995.

[7] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements," Proceedings of the Fifth International Conference on Extending Database Technology (EDBT), 1996.

[8] C. L. Sabharwal, K. R. Hacke and D. C. St. Clair, "Formation of Clusters and Resolution of Ordinal Attributes in ID3 Classification Trees," Proc. of ACM/SIGAPP Symposium on Applied Computing: Technological Challenges of the 1990's, pp. 590-597, 1992.

[9] P. Clark and T. Niblett, "The CN2 Induction Algorithm," Machine Learning, vol. 3, pp. 261-283, 1989.

[10] E. Rich and K. Knight, Learning in Neural Network, 2nd Ed., McGraw-Hill, New York, 1991.