数理解析研究所講究録 1325 巻 2003 年 75-80

間接結合ルールによるデータマイニング
Data Mining via Indirect Association

大阪府立大学大学院理学系研究科 濱野 慎一*(Shinichi Hamano)
大阪府立大学総合科学部数理・情報科学科 佐藤 優子**(Masako Sato)
*Graduate Scool of Science, Osaka Prefecture University
**Department of Mathematics and Information Sciences
College of Integrated Arts and Sciences, Osaka Prefecture University

概要：本稿では同時に購入される割合が低い商品対 (a,b) がメディエータと呼ばれる商品集合を介在させることにより、間接的な従属性を高い割合でもつ、間接ルール $((a,b);M)$ を導入する。またルールの評価指標として P_A, P_D を適用し、実際のビジネスデータであるドラッグストアの POS データを解析する。ブランドの影響力や競合状況などを顧客行動レベルで観測し、間接結合ルールの有益性を示す。

1 結合ルールと指標

I を品目の有限集合とし、各品目を $a,b,a_1,a_2,\cdots$ 等で表す。トランザクションの集合を D とし、各トランザクションを T,T' 等で表す. I の部分集合を X,Y 等で表す。結合ルール (Association Rule) とは、$X \Rightarrow Y$ という形の関係であり、あるトランザクションが品目の集合 X を含むならば、それは品目集合 Y も含むということを表現する。ここで、X を条件部 (Assumption) 或いは本体 (Body)、Y を結論部 (Conclusion) 或いは頭部 (Head) と呼ぶ。結合ルールを最初に定式化した Agrawal 等 [1] は、結合ルールの重要性を評価する指標として、サポート (Support) 及びコンフィデンス (Confidence) と呼ばれる概念を導入した。一般に、品目集合 $Z \subseteq I$ のサポートとは、Z を含むトランザクションのデータベースでの割合、すなわち、出現頻度のことであり、$\sup(Z)$ で表す。結合ルール $X \Rightarrow Y$ のサポートは、条件部と結論部の双方の品目を含むトランザクションのデータベースでの割合、すなわち、$\sup(X \wedge Y)$ である。一方、コンフィデンスは、条件部を満たすトランザクションの内、結論部も満たすトランザクションの条件付の割合で定義され、$\mathrm{conf}(X \Rightarrow Y)$ とかく。すなわち、

$$\mathrm{conf}(X \Rightarrow Y) = \frac{\sup(X \wedge Y)}{\sup(X)}$$

である。ただし、$X \wedge Y$ は、X と Y 双方の品目集合を含む集合とする。Agrawal 等 [1] は、ユーザーが与えた2つの閾値 t_s, t_c(min-support, min-confidence) よりも大きなサポートとコンフィデンスをそれぞれもつ結合ルールを興味深い結合ルール (Interesting Association Rule) と考え、それらのルールを枚挙する APRIORI と呼ばれる演繹的アルゴリズムを提案した。APRIORI では、サポートの値が閾値 t_s よりも大きい品目集合 (多頻度品目集合、Freqent itemset) を先ず導出し、それらの中から、コンフィデンスに関する閾値 t_c より大きな conf の値をもつ結合ルールを検索する。

2 PA と PD

結合ルールの興味深さ (良さ) を計る指標 μ に関する要請として、Piatesky-Shapiro[5] は、次の3つの重要な性質を提唱した。

P_1 : X と Y が統計的に独立ならば、$\mu(X,Y) = 0$;

P_2 : $P(X) = P(Y)$ ならば、$\mu(X,Y)$ は $P(X \wedge Y)$ に関して単調の増加する;

P_3 : 固定された $P(X \wedge Y)$ と $P(Y)$ に対して、$\mu(X,Y)$ は $P(X)$ に関して単調増加する。
また、固定された $P(X \wedge Y)$ と $P(X)$ に対して、$\mu(X,Y)$ は $P(Y)$ に関して単調増加する。

性質 P_1 は、"正しい"結合ルールの指標として、従来から提唱されている要請である。そして、さらに、次の2つの要請がある。

P_4 : $P(X \wedge Y) = P(X)$ ならば、$\mu(X,Y) = 1$;

P_5 : $P(X \wedge Y) = 0$ ならば、$\mu(X,Y) = -1$

Zhang[13] は、品目集合 X と Y の間の結合関係の方向性に着目し、新しい指標を導入した。一般によく使用される指標である、χ^2 や相関係数 ϕ は、方向性がない。すなわち、対称である。彼は先ず、X と Y の間の、結合性と非結合性の違いを指摘した。条件付確率 $P(X|Y)$ が、$P(X|\neg Y)$ より大きければ、X の Y に対する関係は、結合的といえる。そうでなければ、その関係は、非結合的である。Zhang[13] は、結合的・非結合的の双方の場合に対して次の指標を導入した。

$$P_A(X \Rightarrow Y) = 1 - \frac{P(X|Y)}{P(X|\neg Y)}, \quad \text{if } P(X|Y) > P(X|\neg Y),$$

$$P_D(X \Rightarrow Y) = \frac{P(X|Y)}{P(X|\neg Y)} - 1, \quad \text{if } P(X|Y) \leq P(X|\neg Y).$$

上記の指標は、性質 P_1, P_4, P_5 を満たす指標として導入されたが、その他の性質 P_2, P_3 も満たすことが、容易に示される。

定理 2.1. P_A, P_D の性質
(1) $P_A(X \Rightarrow Y)$ と $P_D(X \Rightarrow Y)$ は、$\mathrm{conf}(X \Rightarrow Y)$ に関して単調増加する。
(2) $P_A(X \Rightarrow Y)$ と $P_D(X \Rightarrow Y)$ は、$P(Y)$ に関して単調減少する。

注) (1) では、$\mathrm{conf}(X \Rightarrow Y)$ 以外のパラメータ $P(X), P(Y)$ は、固定されているとする。(2) も同様である。

3 間接的結合ルールの定義

ここでは先ず、同時に購入されることが少ない2つの品目対の定式化から始める。品目の対の集合を $I^2 = \{(a,b) \mid a,b \in I\}$ とする。希少な品目対 (a,b) の候補者を設定するために、品目対サポート閾値 (Rare itempair threshold) $t_p (0 < t_p < 1)$ を導入し、

$$\mathrm{RP} = \{(a,b) \in I^2 \mid \sup(a,b) < t_p\}$$

とする。品目対 $(a,b) \in \mathrm{RP}$ を希少対と呼ぶ。また、対 (a,b) と $t_f (t_p < t_f < 1)$ に対して

$$M_{a,b} = \{c \in I \mid \sup(a,c) \geq t_f,\ \sup(b,c) \geq t_f\}$$

とする。品目 $c \in M_{a,b}$ は、a,b のいづれの品目とも同時に購入したトランザクションが多い品目である。ただし、$(a,b) \in \mathrm{RP}$ ならば、$\sup(a,b,c) \leq \sup(a,b) < t_p$ となるので、これらの3品目を同時に購入したトランザクションは少ない。t_f を品目対頻度閾値 (Frequent itempair threshold) という。$M_{a,b}$ の各品目は、a,b とそれぞれ、高いサポートを有する品目であるが、品目集合 $M_{a,b}$ 自身が同時に高い出現する割合をもつかの保障は必ずしもない。さらに、a,b との相関が正であるのか、負であるのかは、$M_{a,b}$ に含まれる個別品目に依存すると考えられる。

定義 3.1. $(a,b) \in \mathrm{RP}, 0 < t_A, t_D < 1$ とする。品目の集合 $M \subsetneq I$ は、次の条件を満たすとき、(a,b) のメディエータといい、$((a,b);M)$ を間接ルールという。ただし、$y \in \{a,b\}$ とする。

(i) $M \subseteq M_{a,b}, \quad P(M) \geq t_s,$
(ii) $P(M \wedge y) \geq P(M) \times P(y),$
(iii) $P(M \wedge y) < P(M) \times P(y)$

$y = a,b$ に対して (ii) の場合、$((a,b);M)$ を間接結合ルールといい、(iii) のとき間接非結合ルール、a と b で、(ii),(iii) となる場合、間接両結合ルールという。また、t_s をメディエータのサポート閾値、t_A, t_D を結合及び非結合閾値と呼ぶ。

メディエータ M の説明：
(i) メディエータのサポート、すなわち、同時に出現する確率は、メディエータサポート閾値 t_s 以上の高さが必要である。

(ii) 品目 y とメディエータ M が正の相関を持つならば、その結合の度合いを表す $P_A(M \Rightarrow y)$ が結合閾値 t_A 以上の高さが必要である。負の相関を持つ場合も同様である。

メディエータに含まれる品目間に関しての条件として次の概念を導入する。

定義 3.2. 間接結合ルール $((a,b);M)$ は、次の条件を満たすとき、admissible という。

任意の $M' \subsetneq M$ に対して、M' は、(a,b) の間接メディエータにはならない。

間接非結合ルールや、間接両結合ルールの admissibility についても同様に定義する。

定理 3.1. $((a,b);M_{a,b})$ が間接結合ルールならば、Admissible な間接結合ルール $((a,b);M)$ が存在する。間接非結合ルール及び間接両結合ルールの場合も同様である。

以下、$(a,b) \in \mathrm{RP}$ を固定し、$M_{a,b} \neq \phi$ とする。任意の $M \subseteq M_{a,b}$ に対して、定理 2.1 で示したように、$P_A(M \Rightarrow y), P_D(M \Rightarrow y)$ は、$\mathrm{conf}(M \Rightarrow y)$ に関して単調増加である。従って、次の等価性が成り立つ。

$$P_A(M \Rightarrow y) \geq t_A \iff P(M) \geq t_{A,y} \times P(M \wedge y),$$
$$P_D(M \Rightarrow y) \leq -t_D \iff P(M) \leq t_{D,y} \times P(M \wedge y).$$

ただし、

$$t_{A,y} = \frac{P(y)}{1 - t_A(1 - P(y))}, \quad t_{D,y} = \frac{P(y)(1 - t_D)}{1 - t_D P(y)}$$

とする。

定理 3.2. $M_1 \subseteq M_2 \subseteq M_{a,b}$ とする。このとき、

(1) $\max\{P(M_1), P(M_2)\} < t_{A,y} \times P(M_2 \wedge y)$ ならば、$M_1 \subseteq M \subseteq M_2$ を満たす任意の M に対して、$P_A(M \Rightarrow y) < t_A$ である。

(2) $\min\{P(M_1), P(M_2)\} < t_{D,y} \times P(M_1 \wedge y)$ ならば、$M_1 \subseteq M \subseteq M_2$ を満たす任意の M に対して、$P_D(M \Rightarrow y) > -t_D$ である。

系 3.1. $c \in M_{a,b}$ とする。$y = a$ または、$y = b$ に対して、

(1) $\max\{P(c), P(M_{a,b})\} < t_{A,y} \times P(M_{a,b} \wedge y)$ ならば、c を含む (a,b) の結合メディエータは存在しない。

(2) $\min\{P(c), P(M_{a,b})\} < t_{D,y} \times P(c \wedge y)$ ならば、c を含む (a,b) の非結合メディエータは存在しない。

4 アルゴリズム

この節では、希少品目対 (a,b) に対して、間接ルール $((a,b);M)$ を計算するアルゴリズムについて考察する。

メディエータ M を求めるアルゴリズムは、次の 2 つのステップからなる：

Step 1 : $\sup(M) \geq t_s$ を満たす $M \subseteq M_{a,b}$ を求める。

Step 2 : Step 1 で求めた各 M に対して、M と a,b との相関の符号を調べ、それに応じて、P_A, P_D の値がメディエータ閾値 t_A, t_B 以上であるかどうかを調べ、選別する。

Step 1 は、品目集合 $M_{a,b}$ に対する、Agrawal[2] の有名なアルゴリズム APRIORI を採用する。Tan 等 [9] では、品目集合全体 I に対して、このアルゴリズムを適用しているが、ここでは、希少品目対 (a,b) に依存して定まる品目集合 $M_{a,b}$ にこれを適用する点が異なる。

次に間接ルールとなるメディエータを計算するアルゴリズムを考える。以下、$((a,b);M_{a,b})$ は間接結合ルールとし、$\mathcal{M}$ を (a,b) の間接ルールを与えるメディエータの集合とする。更に、

$$\mathcal{M}_k = \{M \in \mathcal{M} \mid |M| = k\}$$

とおく。

Algorithm
入力：Item set I, Database D, Thresholds $(t_p, t_f, t_s, T_A, t_D)$;
出力: 間接ルールの集合
begin
1) $\mathrm{RP} = \{(a,b) \in I^2 \mid a \neq b, \sup(a,b) < t_p\}$;
$\mathrm{FP} = \{(a,b) \in I^2 \mid a \neq b, \sup(a,b) \geq t_f\}$;
2) **for each** itempair $(a,b) \in \mathrm{RP}$ **do**
begin

$M_{ab} = \{c \in I \mid (a,c),(b,c) \in \mathrm{FP}\}$;
$\mathcal{M}_1 = \{\{c\} \mid c \in M_{ab}, \sup(c) \geq t_s\}$;
for each item $c \in \mathcal{M}_1$ **do**
begin
$P\mathcal{M}_1 = \{\{c\} \mid c \in \mathcal{M}_1, \phi(c,a) \geq 0, \phi(c,b) \geq 0\}$;
$M_{A1} = \{\{c\} \mid c \in P\mathcal{M}_1, P_A(c,a) \geq t_A, P_A(c,b) \geq t_A\}$;
$N\mathcal{M}_1 = \{\{c\} \mid c \in \mathcal{M}_1, \phi(c,a) \leq 0, \phi(c,b) \leq 0\}$;
$M_{D1} = \{\{c\} \mid c \in N\mathcal{M}_1, P_D(c,a) \leq t_D, P_D(c,b) \leq t_D\}$;
$B\mathcal{M}_1 = \{\{c\} \mid c \in \mathcal{M}_1, \phi(c,a) \geq 0, \phi(c,b) \leq 0\}$;
$M_{B1} = \{\{c\} \mid c \in B\mathcal{M}_1, P_A(c,a) \geq t_A, P_D(c,b) \leq t_D\}$
end
for$(k = 2; \mathcal{M}_{k-1} \neq \emptyset; k++)$ **do**
begin
$\mathcal{M}_k = apriori - gen(\mathcal{M}_{k-1})$;
for each itemset $M \in \mathcal{M}_k$ **do**
begin
$P\mathcal{M}_k = \{M \mid M \in \mathcal{M}_k, \phi(M,a) \geq 0, \phi(M,b) \geq 0\}$;
$M_{Ak} = \{M \mid M \in P\mathcal{M}_k, P_A(M,a) \geq t_A, P_A(M,b) \geq t_A\}$;
$N\mathcal{M}_k = \{M \mid M \in \mathcal{M}_k \phi(M,a) \leq 0, \phi(M,b) \leq 0\}$;
$M_{Dk} = \{M \mid M \in N\mathcal{M}_k, P_D(M,a) \leq t_D, P_D(M,b) \leq t_D\}$;
$B\mathcal{M}_k = \{M \mid M \in \mathcal{M}_k \phi(M,a) \geq 0, \phi(M,b) \leq 0\}$;
$M_{Bk} = \{M \mid M \in B\mathcal{M}_k, P_A(M,a) \geq t_A, P_D(M,b) \leq t_D\}$
end
end
$\mathcal{M}_{ab} = \bigcup_k \{M_{Ak} \cup M_{Dk} \cup M_{Bk}\}$
end
$Answer = \bigcup_{(a,b) \in \mathrm{RP}} \{((a,b); M) \mid M \in \mathcal{M}_{ab}\}$
end

Apriori-gen:
$apriori - gen$ は $\mathcal{M}_{k-1}$ の入力に対して $\mathcal{M}_k$ を出力する。2 つの $\mathcal{M}_{k-1}$ 中の任意の集合同士から $\mathcal{M}_k$ を作り出す。ただし、$\mathcal{M}_{k-1}$ 中の $k-2$ 個の品目は同じであるが、最後の 1 つの品目は異なっている。

5 数値実験

	a	b	Mediator	P(M,a)	P(M,b)
1	アリエール	ニュービーズ	スコッティティッシュ スコッティロール ハイター	0.724	0.921
2	アリエール	ニュービーズ	エリエールティッシュ エリエールロール ハイター	0.905	0.565
3	エリエール ロール	ネピア ロール	エリエールティッシュ	0.793	-0.117
4	エリエール ロール	ネピア ロール	エリエールティッシュ エルモア	0.644	-0.633
5	エリエール ロール	ネピア ロール	アリエール ハイター	-0.272	0.809
6	エリエール ロール	ネピア ロール	アリエール ハミング	-0.295	0.766
7	エリエール ロール	ネピア ロール	ハイター ハミング	0.307	-0.128

データマイニングオリンピックで利用されたドラッグストアの POS データ (1999 年 4 月～2000 年 3 月) を利用して顧客行動を解析した。上記のルールは 32 店舗, トランザクション数 1422415, 品目数 5920 のデータに関する実験結果である。ただし本研究の目的は間接ルールの発見であるため、1 品目しか購入していない顧客のデータは削除してある。本実験での t_p, t_f, t_s はそれぞれ 0.00014, 0.0007($= t_p \times 5$), 0.0014($= t_p \times 10$)、また t_A, 及び t_D は 0.1, -0.1 である。

これらの閾値を用いて本実験で発見された間接結合ルール数は 774、間接非結合ルール数は 16、間接両結合ルール数は 200 であった。上記のルール 1,2 は発見された間接ルールの一部である。ルール 1 とルール 2 はブランドの影響力を認識することのできるルールとなっている。スコッティを購入する顧客はニュービーズを購入する傾向が強い。また、エリエールを購入する顧客はアリエールを購入する傾向が強い。すなわち、ティッシュメーカーと洗剤メーカーのブランド力は顧客行動に何らかの影響を与えていることが判断できる。ルール 3 以下の 5 つのルールは発見された間接両結合ルールの一部である。ルール 3 とルール 4 からは顧客は同一ブランドで購入する傾向が強いことが判る。エリエールティッシュを購入する顧客はネピアロールよりも同じメーカーによって製造されたエリエールロールを購入する傾向が非常に強い。よってエリエールは個々の製品を宣伝するよりもエリエールとしての全体のイメージを利用した宣伝が効果的であると考えられる。また、

ルール5, 6、及び7からは購入された個々の製品が顧客行動にどのように影響を与えているかが認識できる。アリエールを購入する顧客はエリエールを購入しない傾向にある。7ではアリエールを購入していないのでエリエールを購入する傾向が強くなっている。すなわち、アリエールがエリエールに与える影響はあまり良いものではなくエリエール側のマーケターにとってアリエールは悪い影響を与える潜在的な競合相手であることがわかる。同様の製品を製造しているわけではないが、その影響力が顧客行動レベルで存在するのでアリエール側のプロモーションに常に注意しておかなければならない、という仮説を導きだせる。

最後に本実験では特にティッシュメーカー間の競合度合いがわかるルールが多く発見された。エリエールはネピア、及び、スコッティに対してティッシュ部門、ロール部門どちらも競争優位にたっている。しかしホクシー、クリネックスにはどちらの部門でも顧客を奪われている。エリエールにとっての競合はティッシュメーカー4社ではなく、ホクシーとクリネックス2社であると推測される。

6 おわりに

本論文では同時に購入される割合が低い商品対にメディエータと呼ばれる商品集合を介在させた間接ルールの導入を行った。実際のビジネスデータであるドラッグストアのPOSデータを用いて興味深い顧客行動に関するルールを発見することができた。顧客行動に関する興味深い知見を得ることができ、ブランドの影響力や潜在的な競合相手を知ることができることを示した。

謝辞

本論文を執筆するにあたり、多岐にわたる御指導、御鞭撻を賜りました佐藤優子教授に厚く御礼申し上げます。また数値実験の実行にあたり、多大な御協力並びに貴重な御助言を下さった向内康人先生に厚く御礼申し上げます。最後に実験のために貴重なビジネスデータを提供してくださった宮野悟教授(東京大学)、及びデータに関する様々なアドバイスを下さった矢田勝俊助教授(関西大学)に厚く御礼申し上げます。

参考文献

[1] R. Agrawal, T. Imielinski and A. Swami: *Mining Association Rules between Sets of Items in Large Databases*, in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., USA, May 26–28, 1993, pp. 207–216, 1993.

[2] R. Agrawal and R. Srikant: *Fast Algorithm for Mining Association Rules*, in Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, 1994, pp. 487–499, 1994.

[3] R.J. Bayardo Jr. and R. Agrawal: *Mining the Most Interesting Rules*, in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 15–18, 1999, pp. 145–154, 1999.

[4] S. Brin, R. Motwani, and C.Silvertein: *Beyond market baakets: Generalizing association rules to correlations*, in Proceedings of 1997 ACM-SIGMOD International Conference on Management of Data, Tucson, AZ, 1997.

[5] G.Piatesky-Shapiro: *Discovery, analysis and presentation of strong rules*, In G. Piatesky-Shapiro and W. Frawley, editors, Knowledge Discovery in Data bases, pages 2299-248. MIT Press, Cambridge, MA, 1991.

[6] A. Savasere, E. Omiecinski and S. Navathe: *Mining for Strong Negative Associations in a Large Database of Customer Transactions*, in Proceedings of the Fourteenth International Conference on Data Engineering, Orlando,

Florida, USA, February 23–27, 1998, pp. 494–502, 1998.

[7] R. Srikant and R. Agrawal: *Mining Generalized Association Rules*, in Proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Switzerland, September 11–15, 1995, pp. 407–419, 1995.

[8] R. Srikant, and Q. Vu and R. Agrawal: *Mining Association Rules with Item Constraints*, in Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97), Newport Beach, California, USA, August 14–17, 1997, pp. 67–73, 1997.

[9] P.N. Tan and V. Kumar: *Interestingness Measures for Association Patterns: A perspective*, Technical Report # TR00-036, Department of Computer Science, University of Minnesota, 2000.

[10] P.N. Tan, V. Kumar and J. Srivastava: *Indirect Association: Mining Higher Order Dependencies in Data*, in Proceedings of the Fourth European Conference on Principles of Data Mining and Knowledge Discovery (PKDD-2000), Lyon, France, September 13–16, 2000, Lecture Notes in Computer Science **1910**, pp. 632–637, 2000.

[11] P.N. Tan, and V. Kumar and J. Srivastava: *Selecting the Right Interestingness Measure for Association Patterns*, in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 23–26, 2002, pp. 32–41, 2002.

[12] T. Washio, H. Matuura, and H. Motoda: *Mining Association Rules for Estimation and Prediction*, in Proceedings of the Second Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining (PAKDD-98), Melbourne, Australia, April 15–17, 1998, Lecture Notes in Computer Science **1394**, pp. 417–419, 1998.

[13] T. Zhang: *Association Rules*, in Proceedings of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications (PAKDD-2000), Kyoto, Japan, April 18–20, 2000, Lecture Notes in Computer Science **1805**, pp. 245–256, 2000.

[14] 中谷明弘, 森下真一: 分岐限定法を用いた並列グラフ探索による最適結合ルールの発見, 発見科学とデータマイニング, pages 149-158, 共立出版.