データマイニング　第4回
相関ルールの抽出（1）

総合政策学部　古谷知之

相関ルールの概要

バスケット分析とも呼ばれ、ある事象とある事象がよく同時に起こるということを発見する手法。

例えば、ある電機小売店では“マウス”と“CD-R”がよく買われると分かったり、有名な例では、アメリカのスーパーマーケットでは、「ビールと紙おむつ」がよく買われるという伝説的な例もある。

このように、「X ならば、Yである」「If X=x, then Y=y」という“ルール”を発見する方法である。ルールは下記のようなフォーマットで表現される。

結論 <= 前提条件1&前提条件2&・・・前提条件N

例: マウスを買う <= CD-Rを買う

アソシエーション ルールの欠点: データの対象が多い場合（商品数が数万点あるような場合）、処理に多大な時間がかかってしまう場合がある。

アソシエーション ルールは“支持度”と“信頼度”という集計値から得られるルールで決して何らかの因果関係を示しているものではない!

相関ルールの概要

- 「ある項目が単独で起こる場合の度数」と、「組み合わせで同時に起こる場合の度数」に基づく
- ルールは、「ある項目Aが発生すると、項目BもXパーセントの割合で同時に生じる」と表現される

相関ルールの例

- 顧客が靴を購入するとき、その10%は同時に靴下も購入する
- 食料品チェーン店が発見したところでは、買い物客がトルティーヤチップスを購入するとき、その80%は同時に瓶詰めのサルサソースも購入する。
- 日曜大工をする人がラテックスペンキを購入するとき、その85%は同時にローラーも購入する。
- 株式インデックスファンドを所有する投資家の40%は、そのポートフォリオに成長型投資信託ファンドも含めている

「X ならば、Yである」
「If X=x, then Y=y」

- 顧客が靴を購入するとき、その10%は同時に靴下も購入する
  - X（X＝x）： 靴
  - Y（Y＝y）： 靴下
- Xにも、Yにも、複数の項目を含むことが可能

評価基準

■支持度（Support）

- データベース全体で、その組み合わせが起こる割合

■信頼度（Confidence）

- 相関の強さ。Xに続けてYが起こる割合のパーセント表示

相関ルール発見のための データレイアウトの例

顧客　(id)	製品　(target)
Skip Hops	Natural Choice Tomato Soup
Skip Hops	Washington Carver Honey Ale
Mary Lamb	Natural Choice Black Bean Soup
Mary Lamb	Otto's Oatmeal

相関ルール分析

相関ルール分析は、アソシエーション分析やマーケットバスケット分析とも呼ばれ、「商品Aを購入する人は商品Bも買う傾向がある」というように2つ以上のアイテムの関連性ルールを探るものである。分析結果から、どの商品が一緒に買われるか、どの商品を販売促進すべきかといった洞察を得ることができる。

＜入力データ＞

target　id

商品名	顧客ID	購入した順番
ソーダ	0	0
ペプシ	0	1
ソーセージ	0	2
たばこ	0	3
シリアル	0	4
サルサ	0	5
アボカド	1	0
ホットドッグ	1	1
コーヒー	1	2
ワインクーラー	1	3
鶏肉	1	4
牛乳	1	5
クラッカー	2	0
ハイネケン	2	1
ポテトフライ	2	2
サラミ	2	3
ピーナッツ	2	4
キューリ	2	5

＜出力データ＞

Relations	Lift	Support	Confidence	Transaction Count	Rule
3	7.59	7.6	58.46	76	鶏肉 ==> 野菜ジュース & 胡椒
3	7.59	7.6	98.7	76	野菜ジュース & 胡椒 ==> 鶏肉
3	7.73	7.6	97.44	76	野菜ジュース & 鶏肉 ==> 胡椒
3	7.42	7.6	95	76	胡椒 & 鶏肉 ==> 野菜ジュース
3	7.61	7.6	59.38	76	野菜ジュース ==> 胡椒 & バドワイザー
3	7.64	7.6	60.32	76	胡椒 ==> 野菜ジュース & バドワイザー
3	5.19	7.6	40	76	バドワイザー ==> 野菜ジュース & 胡椒
3	5.19	7.6	98.7	76	野菜ジュース & 胡椒 ==> バドワイザー
3	7.64	7.6	96.2	76	野菜ジュース & バドワイザー ==> 胡椒
3	7.61	7.6	97.44	76	胡椒 & バドワイザー ==> 野菜ジュース
3	7.33	7.6	59.38	76	野菜ジュース ==> 鶏肉 & バドワイザー
3	7.4	7.6	58.46	76	鶏肉 ==> 野菜ジュース & バドワイザー
3	5.13	7.6	40	76	バドワイザー ==> 野菜ジュース & 鶏肉
3	5.13	7.6	97.44	76	野菜ジュース & 鶏肉 ==> バドワイザー
3	7.4	7.6	96.2	76	野菜ジュース & バドワイザー ==> 鶏肉
3	7.33	7.6	93.83	76	鶏肉 & バドワイザー ==> 野菜ジュース

結果の読み方

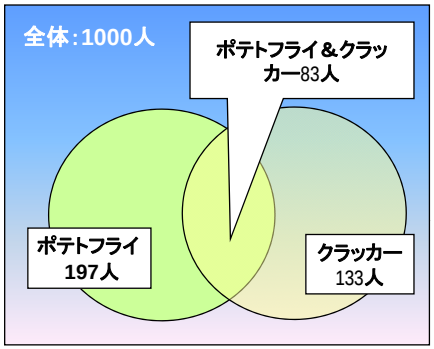
ルール:
顧客がポテトフライを購入するとき、
その約42%は同時にクラッカーも購入する。

- **支持度 (Support)**

$$\frac{\text{同時購入した人数}}{\text{全体の人数}} = 8.3\%$$

- **信頼度(Confidence)**

$$\frac{\text{同時購入した人数}}{\text{ポテトフライを買った人数}} = 42.13\%$$



	Relations	Lift	Support(%)	Confidence(%)	Rule
1	2	3.17	8.30	62.41	クラッカー ==> ポテトフライ
2	2	3.17	8.30	42.13	ポテトフライ ==> クラッカー
3	2	3.08	8.20	60.74	サラミ ==> ポテトフライ
4	2	3.08	8.20	41.62	ポテトフライ ==> サラミ

相関ルールの見つけ方

- 相関ルールを見つけるためには、頻出する項目の組を見つけなければならない。
- 「紙おむつ」と「ビール」を同時に買っている人が多くないと、そのようなルールは意味がない。
- 頻出する項目の組を「頻出アイテム集合」と呼ぶ。
- 相関ルールマイニングアルゴリズムは、「頻出アイテム集合」を高速に求めることが出来る。

なぜ支持度、信頼度か

- アイテム集合の頻出度を測らなければならない。
- ルール $A \Rightarrow C$（おしめ⇒ビール）に対して:
  - 支持度 =支持度($\{A, C\}$) = $P(AC)$
  - 信頼度 = $P(C|A)$ = 支持度($\{A, C\}$)/支持度($\{A\}$)
- 支持度、信頼度の両者で、頻出度を測定できる。
- 支持度は、たとえば購買者全体の中で、「紙おむつ」と「ビール」を同時に買っている人の割合を示す。
  - 支持度はルールの一般性の尺度となっている. ごく少数の事例にしか関係しないパターンは, 如何に特徴的であっても, 一般的な傾向とはみなされない.
  - 例外的な相関ルールを求めたいときには、支持度を低くする。
- 信頼度は、「紙おむつ」を買った人の中で、さらに「ビール」を買った人の割合を示す。
  - 信頼度はルールの確からしさの尺度を与える. 信頼度が高い場合には、ルールの左辺が成り立つと, 右辺が成り立つ確率が大きいことを意味する.

条件付き確率の定義

定義:事象A, Bに関して, 事象Bが起こったという条件のもとで事象Aのおこるという事象をA|B と書くとき, **条件付き確率P(A|B)**は、以下の式で与えられる。

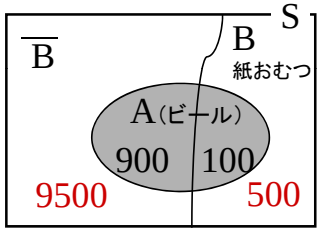
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)}$$

ここで $P(B) > 0$ とする。

Venn図による条件の表現

例 10000人の顧客のうち、500人が紙おむつを購入しとする。また、全体で1000人がビールを購入し、そのうち同時に紙おむつを購入した人が100人であった。

- 事象A:ビールを購入した顧客
- 事象B:紙おむつを購入した顧客



支持度: $P(AB) = \frac{100}{10000}$

信頼度: $P(A|B) = \frac{100}{500} = \frac{100/10000}{500/10000} = \frac{P(AB)}{P(B)}$

相関ルールマイニングの例

ID	Items
2000	A,B,C
1000	A,C
4000	A,D
5000	A,B,C,E,F

最小支持度 50%
最小信頼度 50%

頻出アイテム集合	支持度
{A}	100%
{B}	50%
{C}	75%
{A,B}	50%
{A,C}	75%
{B,C}	50%
{A,B,C}	50%

- 基本アルゴリズム
  - 支持度が最小支持度を超える(等しくてもよい)すべての頻出アイテム集合を求める。
  - 頻出アイテム集合の任意の部分集合を左辺とし、残りを右辺として、ルールを作り、その中で、信頼度が最小信頼度を上回る(等しくてもよい)ルールを集める。

問題

- 如何にして、支持度が最小支持度を超える(等しくてもよい)すべての頻出アイテム集合を求めるのか。
  - 生成検査法:すべてのアイテム集合に対して、支持度を計算する。データベースをすべてのアイテム集合の個数分、繰り返し調べる必要がある。
  - 各レコード(トランザクション)に対して、該当するアイテム集合に積算する。該当するアイテム集合を探すのに手間がかかる。
- もっとうまい方法はないのか。

相関ルールマイニングの例

ID	Items
2000	A,B,C
1000	A,C
4000	A,D
5000	A,B,C,E,F

最小支持度 50%
最小信頼度 50%

頻出アイテム集合	支持度
{A}	100%
{B}	50%
{C}	75%
{A,B}	50%
{A,C}	75%
{B,C}	50%
{A,B,C}	50%

ルール $A \Rightarrow B$ に対して:

支持度 =支持度({A, B}) = 50%

信頼度 = 支持度({A, B})/支持度({A}) = 50%

Apriorの計算原理:

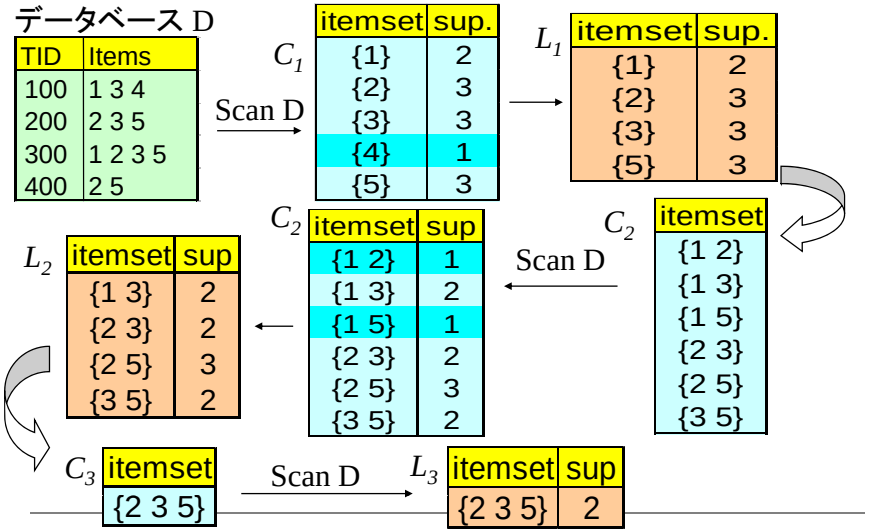
頻出アイテム集合(最小支持度を越えるアイテム集合)の任意の部分集合はふたたび頻出アイテム集合である

Transaction ID	Items Bought
2000	A
2000	B
2000	C
1000	A
1000	C
4000	A
4000	D
5000	A
5000	B
5000	C
5000	E
5000	F

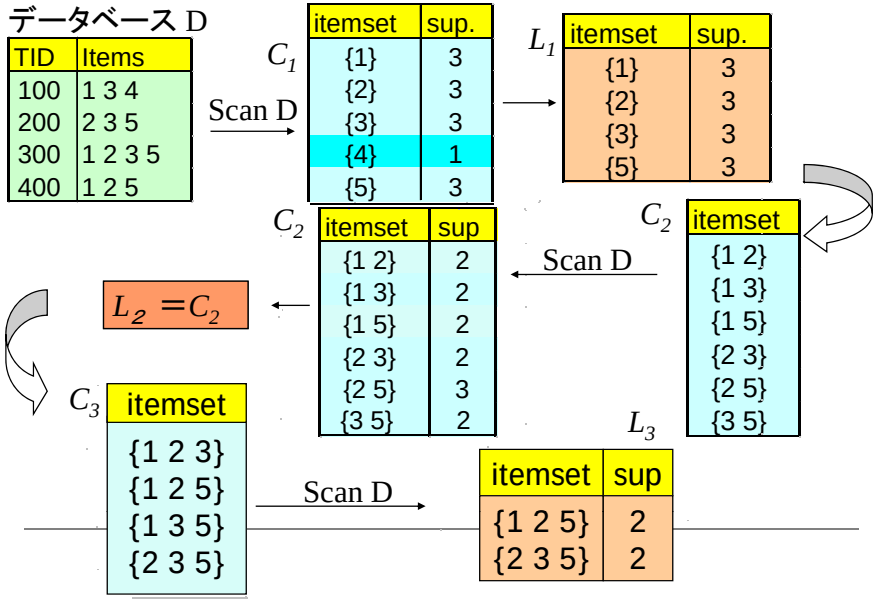
頻出アイテム集合のマイニング

- *頻出アイテム集合*(最小支持度を越えるアイテム集合) を見つける
  - 頻出アイテム集合の任意の部分集合はふたたび頻出アイテム集合でなければならない
    - 例　もし{*AB*}が頻出アイテム集合なら、{*A*}、{*B*}はともに頻出アイテム集合でなければならない。
  - 頻出アイテム集合を集合の大きさの順に1から*k* (*k*-itemset)まで、順繰りに求める。
  - 頻出アイテム集合を用いて相関ルールを求める。

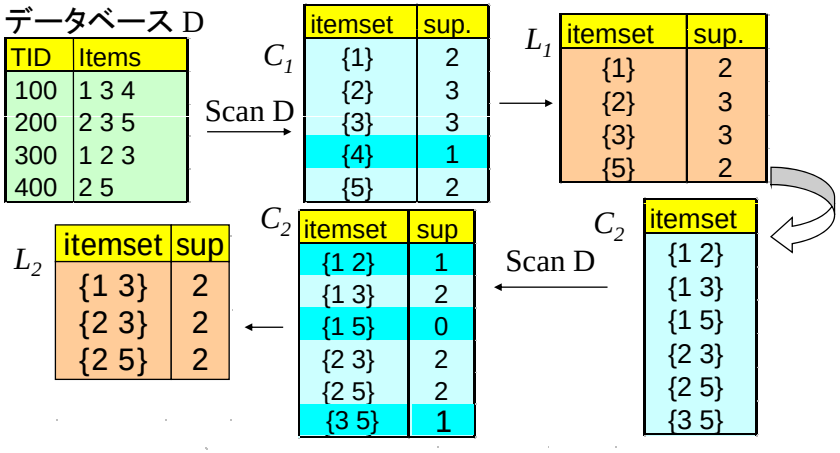
Apriorアルゴリズム — 例1



Apriorアルゴリズム — 例2



Apriorアルゴリズム — 例3



相関ルールの作成

- ステップ1：各頻出アイテム集合 l に対して，それをすべての方法で2つの空でない部分集合 $s, l-s$ に分割する。
  - {2,3,5} ⇒{2,3}と{5}，{2,5}と{3}，{3,5}と{2}，
    {5}と{2,3}，{3}と{2,5}，{2}と{3,5}
- ステップ2： ルール候補 $s \Rightarrow (l-s)$ に対する信頼度 $\frac{\text{support\_count}(l)}{\text{support\_count}(l\text{-}s)}$ を求める。もしこの値が最少信頼度よりも大きいか等しければ，それを相関ルールとする。

信頼度の計算

support_count({2})= support_count({3})=
support_count({5})=3

support_count({2,3})= support_count({3,5})=2

support_count({2,5})=3　　support_count({2,3,5})=2

{2,3}⇒{5}: $\frac{\text{support\_count}(\{2,3,5\})}{\text{support\_count}(\{2,3\})}$ = 1.0

{2,5}⇒{3}: $\frac{\text{support\_count}(\{2,3,5\})}{\text{support\_count}(\{2,5\})}$ = 0.667

{3,5}⇒{2}: $\frac{\text{support\_count}(\{2,3,5\})}{\text{support\_count}(\{3,5\})}$ = 1.0

{5}⇒{2,3}: $\frac{\text{support\_count}(\{2,3,5\})}{\text{support\_count}(\{5\})}$ = 0.667

{3}⇒{2,5}: $\frac{\text{support\_count}(\{2,3,5\})}{\text{support\_count}(\{3\})}$ = 0.667

{2}⇒{3,5}: $\frac{\text{support\_count}(\{2,3,5\})}{\text{support\_count}(\{2\})}$ = 0.667

アソシエーションルールの概要

- ルールは因果関係ではなく、複数の項目間の相関（association）と解釈するべき
- 項目の「有／無」のみに着目。量的評価はしない。
  - （例）ある顧客が項目Aを1単位購入するか、複数単位購入するかは問題にしない。マーケットバスケットに項目Aがあるかどうかだけに着目

逐次関係の発見とデータレイアウト

ルール「A ==> B」(AならばB)は事象Bは事象Aの後に起こるということを意味する･

(例)

- 現在、ポートフォリオに株式インデックスファンドを含めている顧客のうち、15%が次の年に国際ファンドの口座を開きます。
- 新たにコンピュータを購入した顧客のうち、25%が翌月にレーザープリンタを購入します

顧客	来店	製品
Skip Hops	1	Natural Choice Tomato Soup
Skip Hops	1	Brown Cow 2% Milk
Skip Hops	1	Washington Carver Honey Ale
Skip Hops	2	Lucky Dog Chow
Mary Lamb	1	Natural Choice Black Bean Soup
Mary Lamb	1	Brown Cow 2% Milk
Mary Lamb	2	Washington Carver Honey Ale
Mary Lamb	2	Otto's Oatmeal

結果の解釈:アソシエーション

結果 - アソシエーション

ルール | 度数 | コード | ログ | ノート

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	1.25	36.56	61.00	366.00	heineken ==> cracker
2	2	1.25	36.56	75.00	366.00	cracker ==> heineken
3	2	1.11	26.07	43.50	261.00	heineken ==> baguette
4	2	1.11	26.07	66.58	261.00	baguette ==> heineken
5	2	1.35	25.67	80.82	257.00	soda ==> heineken
6	2	1.35	25.67	42.83	257.00	heineken ==> soda
7	2	1.11	25.57	54.12	256.00	olives ==> hering
8	2	1.11	25.57	52.67	256.00	hering ==> olives
9	2	1.38	25.17	42.00	252.00	heineken ==> artichok
10	2	1.38	25.17	82.62	252.00	artichok ==> heineken
11	2	1.62	25.07	78.93	251.00	soda ==> cracker
12	2	1.62	25.07	51.43	251.00	cracker ==> soda
13	2	1.31	24.88	51.23	249.00	hering ==> baguette
14	2	1.31	24.88	63.52	249.00	baguette ==> hering
15	2	1.14	24.88	41.50	249.00	heineken ==> avocado
16	2	1.14	24.88	68.60	249.00	avocado ==> heineken
17	2	1.29	24.48	51.80	245.00	olives ==> bourbon
18	2	1.29	24.48	60.79	245.00	bourbon ==> olives

結果の解釈:逐次

結果 - シーケンス

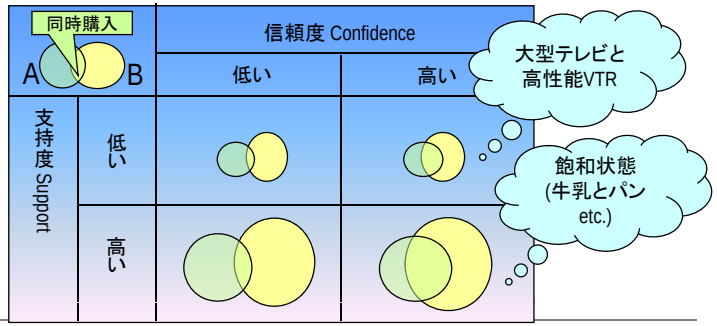
ルール | 度数 | コード | ログ | ノート

	Chain Length	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	33.67	69.06	337	cracker ==> heineken
2	2	23.48	48.35	235	hering ==> heineken
3	2	23.28	49.26	233	olives ==> bourbon
4	2	22.88	47.12	229	hering ==> corned_b
5	2	22.58	46.50	226	hering ==> olives
6	2	22.48	57.40	225	baguette ==> heineken
7	2	21.98	69.18	220	soda ==> cracker
8	2	21.98	56.12	220	baguette ==> hering
9	2	21.98	46.51	220	olives ==> turkey
10	2	21.78	68.55	218	soda ==> heineken
11	2	21.68	73.31	217	coke ==> ice_crea
12	2	21.28	52.85	213	bourbon ==> cracker
13	2	20.98	53.71	210	corned_b ==> olives
14	2	20.88	53.32	209	baguette ==> avocado
15	2	20.78	57.30	208	avocado ==> heineken
16	2	20.68	57.02	207	avocado ==> artichok
17	2	19.78	64.92	198	artichok ==> heineken
18	2	18.28	30.50	183	heineken ==> chicken
19	2	16.68	27.83	167	heineken ==> ice_crea

結果の読み方

信頼度 confidence	高い	当たり前の組合せ
	低い	未発見の組合せ
支持度 support	高い	全顧客に対するインパクトが大きい
	低い	全顧客に対するインパクトが小さい
Lift値	高い	組合せで購入されることが多い
	低い	単品で購入されることが多い

組み合わせのベン図

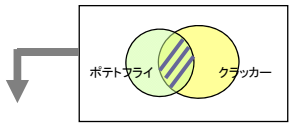


相関ルールマイニングの問題点

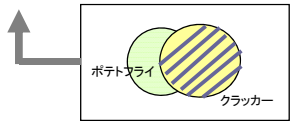
- 相関ルールマイニングでは、しばしば大量の相関ルールが出力される。
- その中には、当たり前のルールが数多く含まれている。
  - たとえば、「パン⇒牛乳」などのルールである。
- ルールとして欲しいのは、思っても見なかった組み合わせのルールである。そのような予想外の組み合わせを発見するためには、その組み合わせの意外性を測る尺度が必要である。
- そのような尺度の1つがリフト値である。

結果の読み方

Lift値について



$$\text{Lift値} = \frac{\text{ポテトフライを買った顧客の中でクラッカーも同時に購入する顧客の割合(信頼度)}}{\text{全顧客のうちクラッカーを購入した顧客の割合(期待信頼度)}}$$



Lift値が1に近ければ、クラッカーの売れ行きが、ポテトフライを同時に買った場合と、それを考慮に入れなかった場合があまり変わらないことを示している。すなわち、そこには組み合わせの意外性が認められないことになる。

順序を考慮した分析

購入順序の分析と分析後の対策

商品Aが買われた後に商品Bが購入される割合が高いかどうか？

もし割合が高ければ、商品Aが買われた後に、商品Bの購入を促すアプローチが考えられる。

アプローチとしては、割引券の配布、製品の発売時期、商品の配置レイアウトなどが考えられる。

評価基準