# データインテリジェンス特論

## 第12回
## アソシエーション分析 (1)

# 第12回 アソシエーション分析 (1)

「よく一緒に買われる商品」を見つけるためのデータ分析

- 相関ルールの基本概念・用語
- Apriori アルゴリズム
  - すべての頻出アイテムセットの抽出
  - アルゴリズムの詳細
- 相関ルールの生成
  - 頻出アイテムセットから強いルールを生成
- その他
  - FP-Tree アルゴリズムの概要

# アソシエーション分析

- マーケットバスケット分析とも呼ばれる
  - 「よく一緒に買われる（＝バスケットに入れる）商品」を見つけるために用いられるため
  - 多くの顧客がある商品を買ったら別の商品も一緒に買うという規則（相関ルール）を見つける
  - 例：購買POSデータの分析
    - 同時に購入される確率が高い商品を明らかにする
    - 店舗のレイアウト設計、棚割り、商品配置の最適化に活用する
    - セット商品を開発する
    - 利益率の高い商品と一緒に買われる商品をセールにする
    - お勧め商品を提案する（レコメンデーション）、クーポンを発行する
- 相関（アソシエーション）ルール
  - Aという事象が起きると、Bという事象が起きやすい

# アソシエーション分析の活用例

| 業種／業態 | 分析 | 効果 |
| --- | --- | --- |
| 小売業（スーパーマーケット、コンビニエンスストア等） | 同時に購入される商品の組み合わせ、組み合わせの店舗間での比較 | 商品陳列の最適化、購買動機の把握、店舗間での顧客嗜好の把握、新規出店時の地理的特徴の把握 |
| eコマース、デジタルコンテンツ産業 | 購買履歴、ページ閲覧履歴 | 顧客ごとにカスタマイズされた案内、広告、トップページによる購入促進 |
| 通信サービス、工業製品 | 製品本体とオプションの組み合わせ、オプション同士の組み合わせ | サービスプランやオプションの提案、オプションのセット販売によるインセンティブの提供 |
| 外食産業 | 食べ物、サイドメニュー、飲み物の組み合わせ | 追加オーダーのレコメンド、セットメニューの考案、顧客動向の調査 |
| 金融サービス | 金融商品の購入状況 | 投資対象銘柄の提案、顧客の嗜好・選択基準の把握 |

# 参考図書


- データマイニングと集合知 ―基礎からWeb,ソーシャルメディアまで―, 石川 博, 新美 礼彦, 白石 陽, 横山 昌平著, 共立出版, 2012.
- アクセンチュアのプロフェッショナルが教えるデータ・アナリティクス実践講座, 翔泳社, 2016.


# 相関ルールの基本概念

用語の説明

# 相関ルールの用語

- 全アイテムの集合　$I = \{I_1, I_2, \cdots, I_n\}$
  - $I_k$ は個々のアイテム

- トランザクション　$T \subseteq I$
  - $T$はアイテムの集合であり$I$に含まれる
  - 個々のトランザクションには識別子TIDが関連付けられ、TIDにより唯一に識別される

- マイニングの対象となるデータベース　$D$
  - $D$はトランザクションの集合

# トランザクションの例

データベース

各行がトランザクション

| TID | アイテム |
|---|---|
| T1 | I1, I2, I4, I5 |
| T2 | I2, I3, I5 |
| T3 | I1, I2, I4, I5 |
| T4 | I1, I2, I3, I5 |
| T5 | I1, I2, I3, I4, I5 |
| T6 | I2, I3, I4 |

個々のアイテム

| ItemID | アイテム名 |
|---|---|
| I1 | フランスワイン |
| I2 | イタリアワイン |
| I3 | スペインワイン |
| I4 | チリワイン |
| I5 | 国産ワイン |

# 相関ルール

- 相関ルール　$A \Rightarrow B$（$A$ ならば$B$ ）

$A$が買われるなら$B$も一緒に買われる

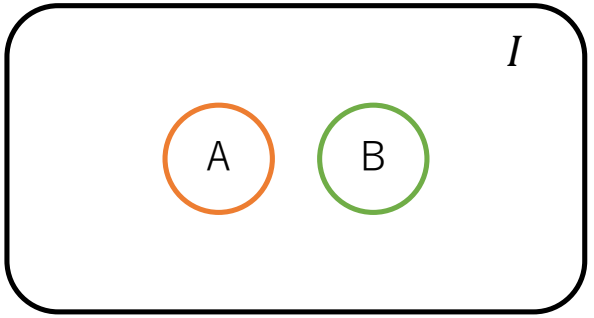
  - $A, B$はアイテムの集合
    - $A, B$は、それぞれ1つの商品でなくても良い
    - 複数でも良いので、一般化して集合とする

  - $A, B \subseteq I$かつ$A \cap B = \emptyset$
    - $A, B$は全アイテムの集合 $I$ の部分集合になっている
    - $A, B$の要素に重なりはない
      - 従って、積集合は空集合

∅: 空集合

# サポート

- 相関ルール
  $A \Rightarrow B$（$A$ ならば$B$ ）
  - $A, B$はアイテムの集合
  - $A, B \subseteq I$かつ$A \cap B = \emptyset$




- $A \Rightarrow B$ のサポート（支持度）

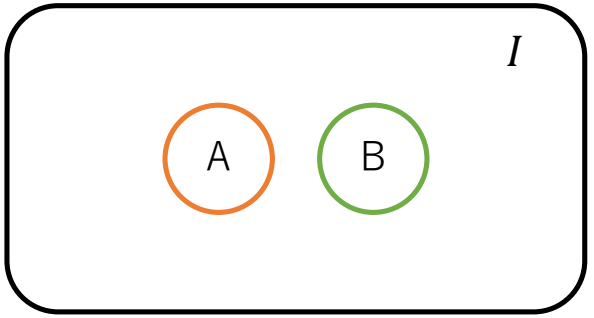
$$P(A \cup B)$$

$$\frac{A\text{と}B\text{を共に含むトランザクションの件数}}{\text{全トランザクションの件数}}$$

  - アイテムの集合$A, B$の両方 $A \cup B$ を含むトランザクションの割合
  - 相関ルールの重要性の尺度
    - 値が大きければ、 $A \cup B$ を含むトランザクションの割合が多く、相関ルールとして重要

# コンフィデンス

- 相関ルール
  $A \Rightarrow B$（$A$ ならば$B$ ）
  - $A, B$はアイテムの集合
  - $A, B \subseteq I$かつ$A \cap B = \emptyset$




- $A \Rightarrow B$ のコンフィデンス（確信度）

$$P(B|A) = \frac{A\text{と}B\text{を共に含むトランザクションの件数}}{A\text{を含むトランザクションの件数}}$$

  - $A$を含むトランザクションに占める、$B$を含むトランザクションの割合
  - 相関ルールの信頼性の尺度
    - 値が大きい　$A$を買うなら$B$を買うことが多い　相関ルールは信頼できる
    - 値が小さい　$A$を買っても$B$を買うことは少ない　相関ルールは信頼できない

# サポート、コンフィデンスの意味

- 相関ルール　パン、バター ⇒ミルク
  - サポート 4%

$$\frac{\text{パン、バター、ミルクを共に含むトランザクションの件数}}{\text{全トランザクションの件数}} = 4\%$$

    - サポートが大きいルールは、出現頻度が高い、影響が大きい重要なルール

  - コンフィデンス 90%

$$\frac{\text{パン、バター、ミルクを共に含むトランザクションの件数}}{\text{パン、バターを含むトランザクションの件数}} = 90\%$$

    - パンとバターを購入する顧客は90%の確率でミルクも購入する
    - コンフィデンスが大きいルールは、信頼できるルール

# 頻出アイテムセット

- アイテムセット
  - アイテムの集合
- サポートカウント (support_count)
  - あるアイテムセットを含むトランザクションの件数
  - アイテムセット$A$の出現頻度 $\mathrm{support\_count}(A)$
- 最小サポートを満足する
  - ある最小サポートが与えられた時
    $\mathrm{support\_count}(A) \geq$ 最小サポート$\times |D|$
- 頻出アイテムセット
  - 最小サポートを満足するアイテムセット

最小サポートカウント

$|D|$ データベースに含まれる全トランザクションの件数

# サポート、コンフィデンスの再定義

- $A \Rightarrow B$ のサポート（支持度）

$$P(A \cup B) = \frac{\text{support\_count}(A \cup B)}{|D|}$$

$$= \frac{A\text{と}B\text{を共に含むトランザクションの件数}}{\text{全トランザクションの件数}}$$

- $A \Rightarrow B$ のコンフィデンス（確信度）

$$P(B|A) = \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)}$$

$$= \frac{A\text{と}B\text{を共に含むトランザクションの件数}}{A\text{を含むトランザクションの件数}}$$

# 強いルール

- 意味のないルールを排除するため、サポートとコンフィデンスに、ある閾値を設定
  - 最小サポート (min_sup)　　…重要性
  - 最小コンフィデンス(min_conf)　　…信頼性

- 強いルールとは
  ・最小サポート以上のサポート（重要性）
  ・最小コンフィデンス以上のコンフィデンス（信頼性）
  を持つルール

# 相関ルールのマイニング

- 相関ルールのマイニングは以下の2つのステップからなる
  1. すべての頻出アイテムセットを抽出
     - トランザクションの集合であるデータベースをスキャンし、アイテムセットのサポートカウントを計算
     - サポートカウントが　最小サポート×$|D|$　以上ならば頻出アイテムセット
  2. 頻出アイテムセットから強いルールを生成

- ステップ1では繰返しデータベースをスキャンすることになるため、この部分の効率化が必要

→Apriori アルゴリズム、FP-Tree アルゴリズム

# Aprioriアルゴリズム

相関ルールのマイニング手順（１）

- すべての頻出アイテムセットの抽出

# Aprioriアルゴリズム

$A$が頻出アイテムセット
$\mathrm{support\_count}(A) \geq$ 最小サポート$\times |D|$

## 頻出アイテムセット抽出の効率化が目的

- 基本的な考え方
  - Aが頻出アイテムセットならば、
    Aの部分集合のBも頻出アイテムセットになる
  - Bが頻出アイテムセットでなければ、
    Bを含む集合Aも頻出アイテムセットではない

頻出かどうかのチェック
を無駄に行わないように
することによる効率化

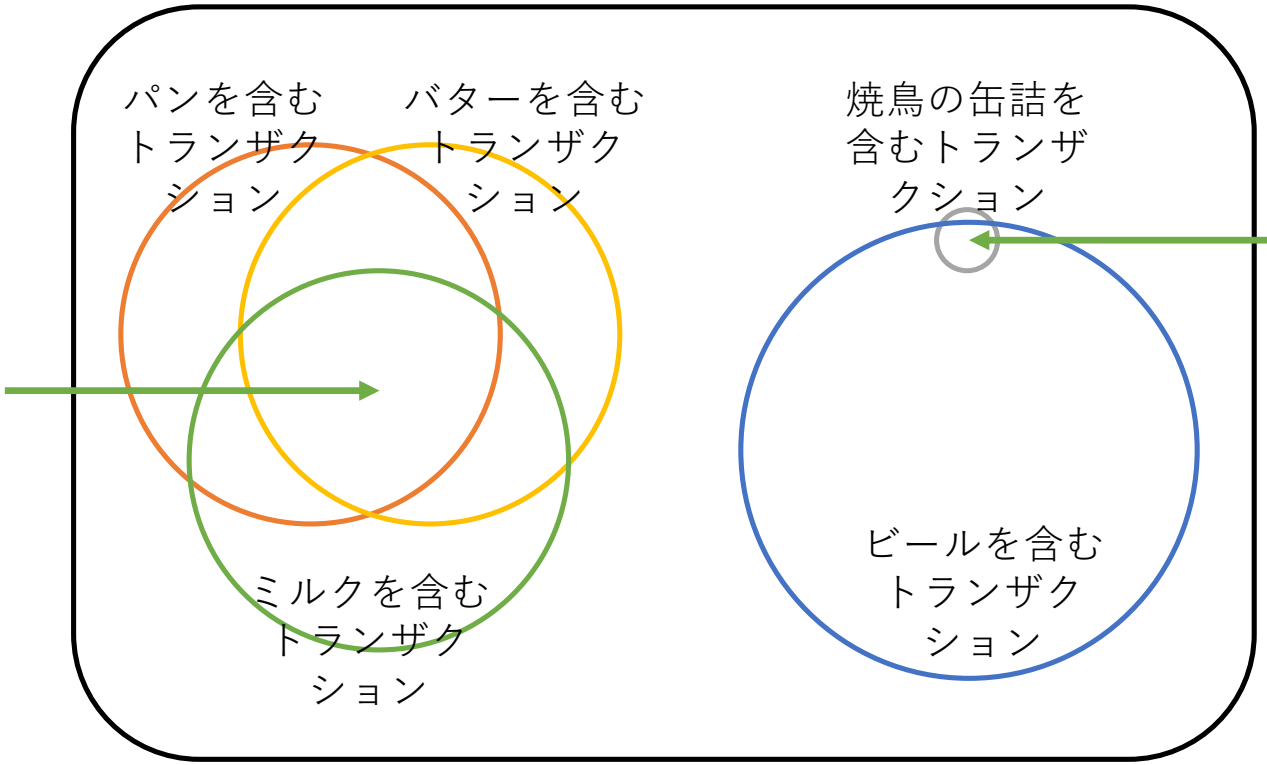
$$\frac{\text{パン、バター、ミルクを共に含むトランザクションの件数}}{\text{全トランザクションの件数}} = 10\%$$

- ならば、パン、バターを共に含むトランザクション件数は少なくとも同数以上

$$\frac{\text{焼鳥の缶詰を含むトランザクションの件数}}{\text{全トランザクションの件数}} = 0.1\%$$

- ならば、焼鳥の缶詰と他のアイテムを含むトランザクション件数は同数以下

# Aprioriアルゴリズム：考え方




パン、バターを含むトランザクションの件数は、パン、バター、ミルクを含むトランザクションの件数よりも多い

焼鳥の缶詰を含むトランザクションの件数がそもそもすくなければ、焼き鳥の缶詰とビールを含むトランザクションの件数が多くなることはありえない

パン、バター、ミルクが頻出アイテムセットならば、パン、バターも頻出アイテムセット

焼鳥の缶詰が頻出アイテムセットでなければ、焼鳥の缶詰とビールは頻出アイテムセットにはならない

# Aprioriアルゴリズム：定式化と手順化

- 単調減少性

  $X, Y \subseteq I, X \subseteq Y \Rightarrow \mathrm{support\_count}(X) \geq \mathrm{support\_count}(Y)$

  - $X, Y$はアイテムの集合であり、 $X$は$Y$の部分集合ならば、 $Y$を含むトランザクションの個数（サポートカウント）は、$X$を含むトランザクションの個数（サポートカウント）以下になる

アルゴリズムは
この性質（アプリオリ）を利用する

- 要素数の少ないアイテムセットからサポートカウントを計算
- 頻出アイテムセットとならないアイテムセットが見つかったら、そのアイテムセットを包含するアイテムセットは候補から除外（枝刈り）

# Apriori アルゴリズム：手順の実行例

- 手順
  - 要素数の少ないアイテムセットからサポートカウントを計算
  - 頻出アイテムセットとならないアイテムセットが見つかったら、そのアイテムセットを包含するアイテムセットは候補から除外

1. 要素が1つのアイテムセットのサポートカウントを計算
   - サポートカウントが最小サポートカウントより小さいと頻出ではない
   - 頻出ではない（要素が１つの）アイテムセットを含む、より要素の多いアイテムセットは、頻出ではないので、候補から除外
2. 要素が２つのアイテムセットのサポートカウントを計算
   - 要素を２つにすることで、重なりが少ないアイテムセットのサポートカウントは小さくなり、頻出ではなくなるかもしれない
   - 頻出ではなくなったアイテムセットがあれば、それを含むアイテムセットも頻出ではないので、候補から除外

# 相関ルールの生成

相関ルールのマイニング手順（２）

- 頻出アイテムセットから強いルールを生成

# 相関ルールのマイニング手順

- 相関ルールのマイニングは以下の2つのステップからなる

1. すべての頻出アイテムセットを抽出　　Aprioriアルゴリズム
2. 頻出アイテムセットから強いルールを生成　　相関ルールの生成

# 強いルールの生成とは

- 強いルール
  - 最小サポート以上のサポート
    - 最小サポート以上のサポートを持つ頻出アイテムセットは抽出済み
  - 最小コンフィデンス以上のコンフィデンス

- コンフィデンス $(A \Rightarrow B) \equiv \dfrac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)}$

$$= \frac{A\text{と}B\text{を共に含むトランザクションの件数}}{A\text{を含むトランザクションの件数}}$$

# 相関ルールの生成：手順

$l$を除く部分集合
この場合は空集合も除く

1. 各頻出アイテムセット $l$ について、空集合以外の真部分集合 $s$ を求める
   - $k$個のアイテムを含むアイテムセットの場合、 $s$ は $2^k - 2$ 個できる
2. ルール $s \Rightarrow (l - s)$ を作成し、そのコンフィデンスを計算する

$$\text{コンフィデンス}\big(s \Rightarrow (l-s)\big) \equiv \frac{\text{support\_count}(s \cup (l-s))}{\text{support\_count}(s)} = \frac{\text{support\_count}(l)}{\text{support\_count}(s)}$$

3. コンフィデンスが最小コンフィデンス以上であれば採用する

- 採用されたルールは、頻出アイテムセットから作られたため、最小サポートを満足し、かつ最小コンフィデンスも満足するため、強いルールとなる

# 相関ルールの生成：例

頻出アイテムセット $l = \{\mathrm{I2}, \mathrm{I3}, \mathrm{I5}\}$ からルールを生成

1. 空集合以外の真部分集合 $s$ を $2^3 - 2 = 6$ 個求める
   - $\{\mathrm{I2}\}, \{\mathrm{I3}\}, \{\mathrm{I5}\}, \{\mathrm{I2}, \mathrm{I3}\}, \{\mathrm{I2}, \mathrm{I5}\}, \{\mathrm{I3}, \mathrm{I5}\}$
2. ルール $s \Rightarrow (l - s)$ を作成し、そのコンフィデンスを計算

$$\text{コンフィデンス}\big(s \Rightarrow (l - s)\big) \equiv \frac{\text{support\_count}(s \cup (l - s))}{\text{support\_count}(s)} = \frac{\text{support\_count}(l)}{\text{support\_count}(s)}$$

- $\{\mathrm{I2}\}$ $3/6 = 50\%$
- $\{\mathrm{I3}\}$ $3/4 = 75\%$
- $\{\mathrm{I5}\}$ $3/5 = 60\%$
- $\{\mathrm{I2}, \mathrm{I3}\}$ $3/4 = 75\%$
- $\{\mathrm{I2}, \mathrm{I5}\}$ $3/5 = 60\%$
- $\{\mathrm{I3}, \mathrm{I5}\}$ $3/3 = 100\%$

サポートカウント

| ItemID | サポートカウント |
|---|---|
| I2 | 6 |
| I3 | 4 |
| I5 | 5 |

| ItemID | サポートカウント |
|---|---|
| I2, I3 | 4 |
| I2, I5 | 5 |
| I3, I5 | 3 |

| ItemID | サポートカウント |
|---|---|
| I2, I3, I5 | 3 |

# 相関ルールの生成：例（続き）

頻出アイテムセット $l = \{\mathrm{I2}, \mathrm{I3}, \mathrm{I5}\}$ からルールを生成

3. 例えば最小コンフィデンス 70% 以上のルール $s \Rightarrow (l - s)$ を採用すると
   - $\{\mathrm{I3}\}$ $3/4=75\%$ $\{\mathrm{I3}\} \Rightarrow \{\mathrm{I2}, \mathrm{I5}\}$
   - $\{\mathrm{I2}, \mathrm{I3}\}$ $3/4=75\%$ $\{\mathrm{I2}, \mathrm{I3}\} \Rightarrow \{\mathrm{I5}\}$
   - $\{\mathrm{I3}, \mathrm{I5}\}$ $3/3=100\%$ $\{\mathrm{I3}, \mathrm{I5}\} \Rightarrow \{\mathrm{I2}\}$

採用されたルール

# リフト値

$$\frac{P(A \cup B)}{P(A)P(B)} = \frac{\text{コンフィデンス}(A \Rightarrow B)}{\text{サポート}(B)}$$

- アイテム $A, B$ の相関を示す指標
  - 1に等しければ $A, B$ に相関はない
  - 1よりも大きければ $A, B$ は一緒に起きやすい（正の相関）
  - 1よりも小さければ $A, B$ は一緒に起きにくい（負の相関）

# リフト値：例

| | チリワイン | $\overline{\text{チリワイン}}$ | **小計** |
|---|---|---|---|
| イタリアワイン | 3500 | 3000 | 6500 |
| $\overline{\text{イタリアワイン}}$ | 1500 | 500 | 2000 |
| **小計** | 5000 | 3500 | 8500 |

サポート・コンフィデンスともに高いが
リフト値からはやや負の相関が出ている
このルールを用いる時には注意が必要

- ルール：チリワイン ⇒ イタリアワイン
  - サポート $\dfrac{\text{チリワインとイタリアワインを共に含むトランザクションの件数}}{\text{全トランザクションの件数}} = \dfrac{3500}{8500} = 41\%$
  - コンフィデンス $\dfrac{\text{チリワインとイタリアワインを共に含むトランザクションの件数}}{\text{チリワインを含むトランザクションの件数}} = \dfrac{3500}{5000} = 70\%$

サポート・コンフィデンスともに高い

- リフト値 $\dfrac{\dfrac{\text{チリワインとイタリアワインを共に含むトランザクションの件数}}{\text{全トランザクションの件数}}}{\dfrac{\text{チリワインを含むトランザクションの件数}}{\text{全トランザクションの件数}} \cdot \dfrac{\text{イタリアワインを含むトランザクションの件数}}{\text{全トランザクションの件数}}} = \dfrac{\frac{3500}{8500}}{\frac{5000}{8500} \cdot \frac{6500}{8500}} = 0.92$

リフト値はわずかに負の相関

チリワインを買うとイタリアワインを買う確率はやや下がる

# JupyterLab

Pythonでのアソシエーション分析