RITSUMEIKAN

機械学習 第１２回 パターンマイニング Pattern mining

模式挖掘

立命館大学 情報理工学部

福森 隆寛

Beyond Borders

講義スケジュール

- 担当教員１：福森（第1回～第15回）

1	機械学習とは、機械学習の分類
2	機械学習の基本的な手順
3	識別（１）
4	識別（２）
5	識別（３）
6	回帰
7	サポートベクトルマシン
8	ニューラルネットワーク

9	深層学習
10	アンサンブル学習
11	モデル推定
12	**パターンマイニング**
13	系列データの識別
14	半教師あり学習
15	強化学習

- 担当教員２：叶昕辰先生（第16回の講義を担当）

今回の講義内容

- 取り扱う問題の定義
- パターンマイニング
  データ中に何度も出現するパターンを抽出したり（頻出項目抽出）、そのパターンに基づいた規則を発見する（連想規則抽出）手法
- Aprioriアルゴリズム 先验算法
  「a prioriな原理」の対偶を用いて、小さな項目集合から支持度の計算をはじめ、項目集合を大きくする際に、頻出でない項目集合を それ以上拡張させず、調べる項目集合を減らす方法
  - 頻出項目抽出（ひんしゅつ） 常用词提取
    パターンマイニングで扱うデータの単位
  - 連想規則抽出（れんそう） 关联规则抽取
- FP-Growthアルゴリズム
  Aprioriアルゴリズムを高速化する手法
  トランザクションデータをコンパクトな情報に変換して、そのコンパクトな情報に対してパターンマイニングする
- 推薦システムにおける学習
  - 協調フィルタリング（きょうちょう） 协同过滤Collaborative filtering
    新規ユーザがある商品を購入した時、購入パターンが似ているユーザを探し、「そのユーザが購入していて、かつ新規ユーザが購入していない商品」を推薦するというのが基本的な考え方
  - Matrix Factorization
- 演習問題

取り扱う問題の定義：教師なし・パターンマイニング

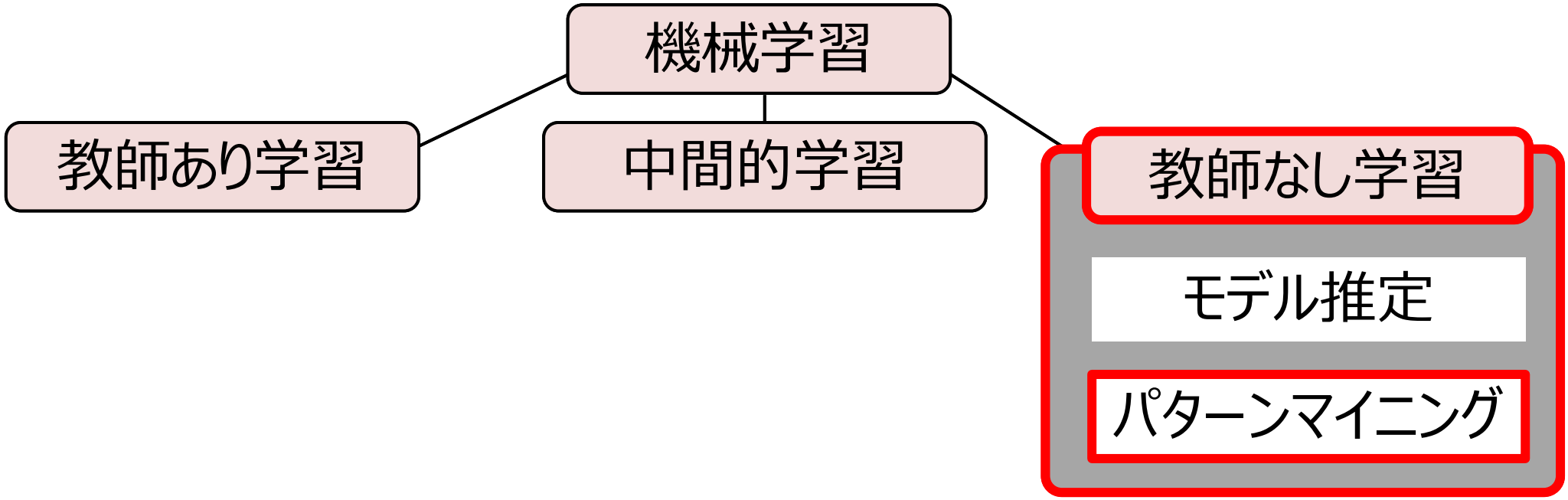
输入分类格式的特征向量，以查找隐藏在数据中的有用pattern

- **カテゴリ形式**の特徴ベクトルを入力して、
  そのデータに潜んでいる有用なパターンを見つける

※ 教師なし学習の問題での学習データは、以下で構成される

入力データの特徴ベクトル（カテゴリ形式） ← $\{\boldsymbol{x}_i\},\ i = 1,2,\ldots,N$ → 学習データの総数

※ 正解情報は与えられていない



パターンマイニング

- **パターンマイニング**
  - データ中に何度も出現するパターンを抽出したり（**頻出項目抽出**）、そのパターンに基づいた規則を発見する（**連想規則抽出**）手法

一种提取数据中多次出现的模式（提取频繁出现的项目）
并根据这些模式查找规则（提取关联规则）的方法。

- パターンマイニングの応用例
  - ネットショッピングサイトなどでのお勧め商品の提示（ていじ）
    - 商品Aと商品Bを購入している客が多くいるとき、商品Aのみを購入した客に対して商品Bを勧める
  - データからの連想規則の抽出による新たな知見の獲得
    - 血液の生化学検査（けっえきのせいかがくけんさ）項目の値から腫瘍（しゅよう）マーカの値の高低（こうてい）が推定できれば、効果な腫瘍マーカ検査の回数を減らせる

頻出項目抽出 常用词提取

- パターンマイニングで扱うデータの単位（たんい）
  - Transaction 交易
  - **トランザクション**：1個のデータ
    - 例：スーパーマーケットの売り上げ記録の場合
      - ある人の1回分の買い物（かいもの）で同時に買われた物の集合がトランザクション

- **支持度**（しじど） (support)
  - 全データに対して、ある項目集合が出現する割合

$$\mathrm{support}(\mathrm{items}) = \frac{T_{\mathrm{items}}}{T}$$

  - T：全トランザクション件数（けんすう）
  - T_{items}：項目集合itemsが出現するトランザクション件数

演習問題12-1（10分間）

- 以下のスーパーマーケットの売り上げ記録がある
  - 商品点数：{ミルク、パン、バター、雑誌(ざっし)} の4点
  - トランザクション：6件（6回分の購入記録）

1. このデータは「疎らなデータ」、「密なデータ」のどちらであるか考えよ
   疎らなデータ
2. 項目集合{ミルク、パン}の支持度を示す support({ミルク、パン})を求めなさい
3. 商品点数が1000種類であったとき、可能な項目集合の数はいくつか？
   ヒント：商品は「買った」「買っていない」の2種類に分類できると考える
   2^1000 -1通り 必须至少买一件东西，因此需要-1

No.	ミルク	パン	バター	雑誌
1	t	t		
2		t		
3				t
4		t	t	
5	t	t	t	
6	t	t		

スーパーマーケットの売り上げ記録
（tはその商品が買われたことを意味する）

頻出項目抽出：Aprioriアルゴリズム

Aprioriアルゴリズム

任何频繁项的子项一定是频繁的
任何不频繁的子项所对应的超项不频繁

- 「a prioriな原理」の対偶(たいぐう)を用いて、
  小さな項目集合から支持度の計算をはじめ、
  項目集合を大きくする際に、頻出でない項目集合を
  それ以上拡張(かくちょう)させず、調(しら)べる項目集合を減らす方法

- a prioriな原理 认为认识是先天的
  - “a priori” は、ラテン語で
    「経験的(けいけんてき)認識に先立(さきだ)つ先天的(せんてんてき)、自明的(じめいてき)な認識や概念」
  - 今回の場合、「学習データ（経験的認識）に関係なく、
    当たり前に成り立つこと」という意味

演習問題12-1（10分間）解答例

1. このデータは「疎らなデータ」、「密なデータ」のどちらであるか考えよ

スーパーマーケットの売り上げ記録

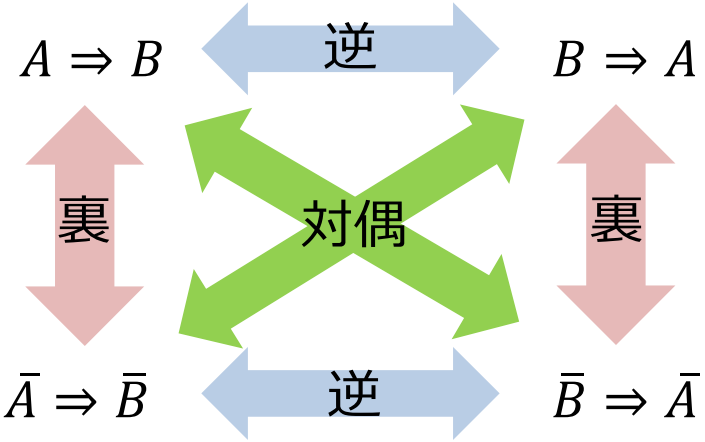
No.	ミルク	パン	バター	雑誌
1	t	t		
2		t		
3				t
4		t	t	
5	t	t	t	
6	t	t		

□ 解答例

- ■ 疎らなデータ
- ■ トランザクションの特徴ベクトルのほとんどの次元の値は空白(くうはく)で、少数(しょうすう)の次元のデータが埋(う)まっている
  - スーパーマーケットで揃えている商品の種類数：数千～数万
  - 1回の買い物で1人の客が買う商品の種類数：多くて数十

頻出項目抽出：Aprioriアルゴリズム

- 【補足】命題（めいだい）「AならばB」について
  - 逆：BならばA
  - 裏（うら）：Aでないならば**B**でない
  - **対偶**（逆の裏）：Bでないならば**A**でない
  - 「AならばB」が成立するとき
    - 逆や裏：必ずしも成り立つわけではない
    - 対偶：必ず成り立つ



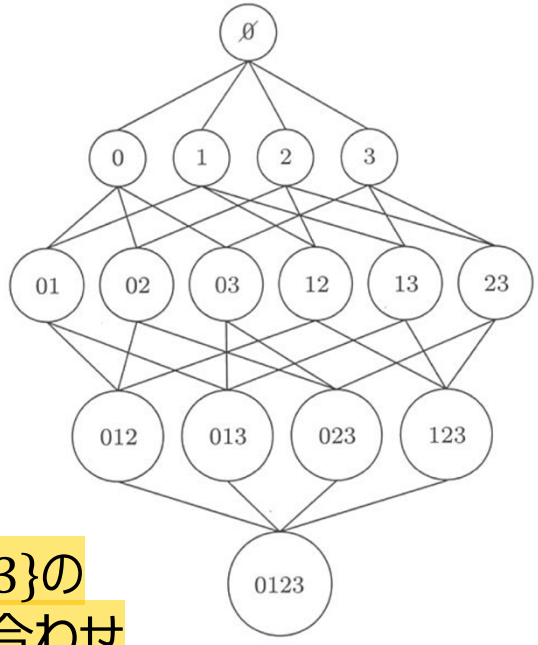
頻出項目抽出：Aprioriアルゴリズム

- 今回の「a prioriな原理」

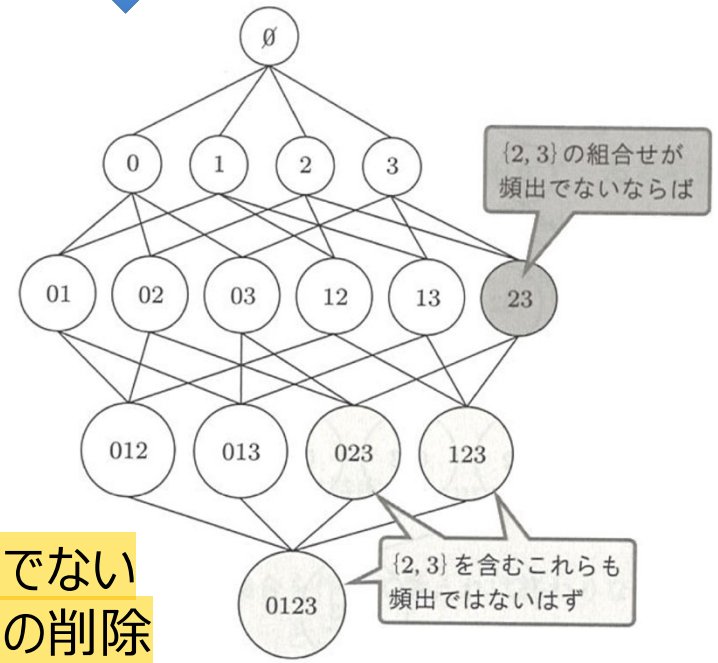
命題
ある項目集合が頻出ならば
その部分集合も頻出である

対偶
ある項目集合が頻出でないならば
その項目集合を含む集合も頻出ではない

これを図で表現すると



項目{0,1,2,3}の
全ての組み合わせ



頻出でない
項目の削除

頻出項目抽出

Apriorアルゴリズム（頻出項目抽出）

入力：正解なしデータ D
出力：頻出項目集合
$F_1 \leftarrow$ 要素数1の頻出項目集合
$k = 2$
while $F_1 \neq \emptyset$ **do** 频繁的items sets
　$C_k \leftarrow F_{k-1}$ の各要素の組み合わせ
　for all $x \in D$ **do** 扫描数据库中的每一条记录
　　for all $c \in C_k$ **do**
　　if $c \subset x$ **then**
　　$c.count \leftarrow c.count + 1$
　　end if
　　end for
　　$F_k \leftarrow \{c \in C_k | c.count >$ 閾値$\}$ Frequent item sets大于阈值，则选择
　end for
　$k \leftarrow k + 1$ 循环
end while
return $\cup_k F_k$

Frequent item sets
F_k：要素数kの頻出項目集合

C_k：F_{k-1}の要素を組み合わせて作られる頻出項目集合の候補
Candidate item sets
长度为k的候补频繁项目集合

※ 予め抽出する項目集合の支持度の閾値を決めておき、それを超えるものを頻出項目集合とする

連想規則抽出：Aprioriアルゴリズム

- 規則の学習では、どの特徴（またはその組み合わせ）が結論部になるかわからない
  - 「商品Aを購入したならば、商品Bを購入することが多い」、「商品Cを購入したならば、商品DとEを購入することが多い」のように得られた規則のそれぞれが、異なった条件部・結論部を持つことが多くなる
- 連想規則の作成手順
  1. Apriori アルゴリズムで頻出項目を抽出
  2. 頻出項目の要素を、条件部と結論部に分けて可能な規則集合を生成
  3. **規則の有用性**を評価し、役立ちそうなものを絞り込む

评估规则的有用性并缩小可能有用的范围

連想規則抽出：Aprioriアルゴリズム

連想規則の作成手順（つづき）

規則の有用性を評価する基準

- **確信度**（confidence）置信度
  - 規則の条件部が起こったときに結論部が起こる割合
  - 確信度が高いほど、この規則に当てはまる事例が多い

$$\mathrm{confidence}(A \Rightarrow B) = \frac{\mathrm{support}(A \cup B)}{\mathrm{support}(A)}$$

- **リフト値**（lift） 提升値
  - 規則の結論部だけが単独（たんどく）で起こる割合と条件部が起こったときに結論部で起こる割合との比
  - リフト値が高いほど、得られる情報の多い規則である

$$\mathrm{lift}(A \Rightarrow B) = \frac{\mathrm{condidence}(A \Rightarrow B)}{\mathrm{support}(B)}$$

置信度 / 支持度

演習問題12-2（5分間）

- 「ハム→卵（たまご）」という規則について、以下の情報が得られたとする
  - 項目集合 {ハム、卵} の支持度が0.1
  - 「ハム→卵」という規則の確信度が0.7、リフト値：5

- 上記の条件において、以下の項目を求めなさい
  1. ハムと卵を同時に購入している客の割合　支持度
  2. ハムを購入した客の中で卵も購入する客の割合　確信度
  3. 「ハムを既（すで）に買った客が卵を買う確率」は「任意の客が卵を買う確率」の何倍か？

     規則の結論部だけが単独で起こる

演習問題 解答例

- 「ハム→卵」という規則について、以下の情報が得られたとする
  - 項目集合 {ハム、卵} の支持度が0.1
  - 「ハム→卵」という規則の確信度が0.7、リフト値：5

- 解答例
  1. ハムと卵を同時に購入している客の割合：10％（支持度）
  2. ハムを購入した客の中で卵も購入する客の割合：70％（確信度）
  3. 「ハムを既に買った客が卵を買う確率」は「任意の客が卵を買う確率」の何倍か？：5倍（リフト値）

連想規則抽出：Aprioriアルゴリズム

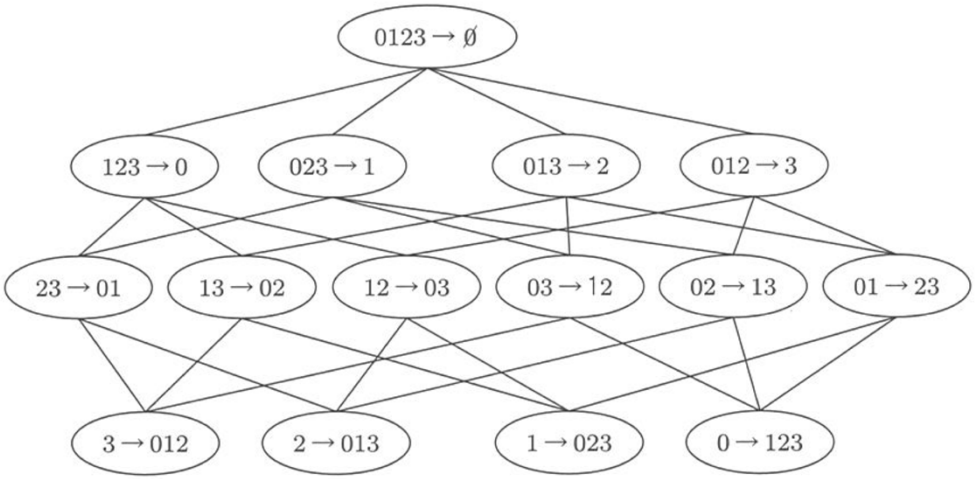
缩小评价值高的规则

□「a priori原理」から、評価値の高い規則を絞り込む

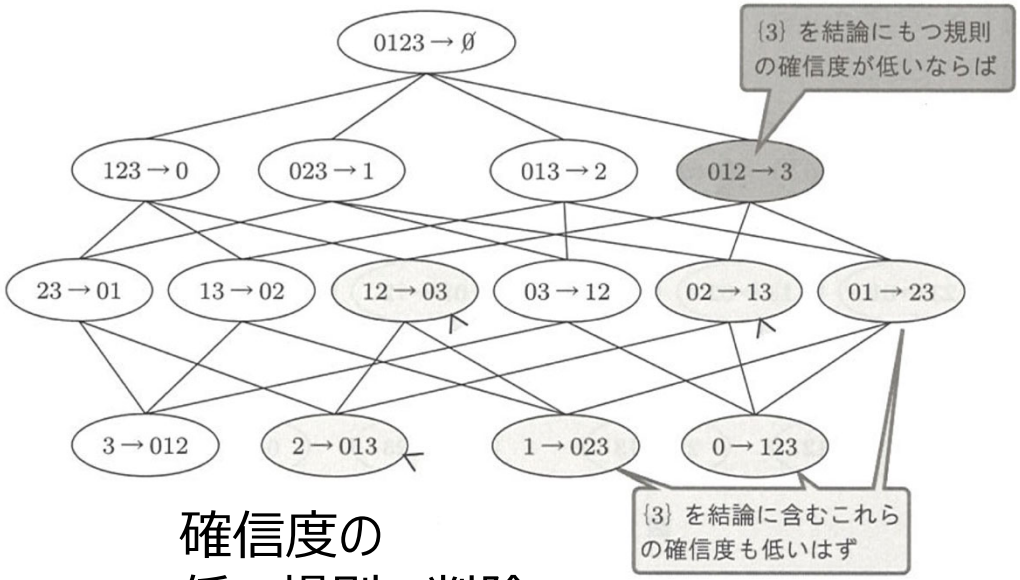
命題
ある項目集合を結論とする規則の確信度が高ければ、その部分集合を結論とする規則の確信度も高い

対偶
ある項目集合を結論とする規則の確信度が低ければ、その項目集合を含む項目集合を結論とする規則の確信度も低い

「商品Aを購入したならば、商品B・Cを購入する」という規則の確信度が高ければ
「商品Aを購入したならば、商品Bを購入する」という規則の確信度も高い…ということ



可能な連想規則集合
（頻出項目集合が {0,1,2,3} の場合）



確信度の
低い規則の削除

連想規則抽出

Aprioriアルゴリズム

F_k：要素数kの頻出項目集合
H_m：要素数mの結論部の集合

入力：頻出項目集合 F_k $(k \geq 2)$
出力：連想規則集合
for all $f_k \in F_k$ **do**
　$H_1 \leftarrow \emptyset$
　$A \leftarrow \{a_{k-1} | a_{k-1} \subset f_k\}$
　if confidence$(a_{k-1} \Rightarrow f_k - a_{k-1}) >$ 閾値 **then**
　　規則 $a_{k-1} \Rightarrow f_k - a_{k-1}$ を出力
　　H_1に $\{f_k - a_{k-1}\}$ を加える
　end if
　ap−genrules(f_k, H_1) /* 次スライドのアルゴリズムを参照（さんしょう） */
end for

※ 予め抽出する連想規則の確信度の閾値を決めておく

連想規則抽出

- Apriori アルゴリズム（つづき）
  - ap−genrules(f_k, H_m)

```
if k > m + 1 then
    H_m の要素を組み合わせて H_{m+1} を作成
    for all h_{m+1} ∈ H_{m+1} do
        if confidence(f_k − h_{m+1} ⇒ h_{m+1}) > 閾値 then
            規則 f_k − h_{m+1} ⇒ h_{m+1} を出力
        else
            H_{m+1} から h_{m+1} を削除
        end if
    end for
    ap−genrules(f_k, H_{m+1})
end if
```

FP-Growthアルゴリズム

FP-Growthアルゴリズム

将数据集存储在一个特定的称做FP树的结构之后发现频繁项集或者频繁项对，即常在一块出现的元素项的集合FP树

- Aprioriアルゴリズムを高速化する手法
- トランザクションデータをコンパクトな情報に変換して、そのコンパクトな情報に対してパターンマイニングする

  将交易数据转换为紧凑信息并对该紧凑信息进行模式挖掘
- トランザクションデータをコンパクトにする手順
  1. 特徴を頻出頻度順に並べ替える 按频率排序特征
  2. 頻度の高い順から順に、その情報をまとめると

     商品Aの購入が100件で、その内、商品Bの同時購入が40件、
     商品Cの同時購入が30件…

     のように多数のトランザクションの情報を手短(てみじか)に表現
  3. 2. の情報を木構造で表現

FP-Growthアルゴリズム

- FP-Growthアルゴリズムの例
  - 以下のトランザクション集合を学習データとして考える
  1. 学習データに対する特徴の出現頻度の計算とフィルタリング

     训练数据特征出现频率的计算和过滤

1	{r,z,h,j,p}
2	{z,y,x,w,v,u,t,s}
3	{z}
4	{r,x,n,o,s}
5	{y,r,x,z,q,t,p}
6	{y,z,x,e,q,s,t,m}

学習データ

出現する文字は特徴名を表す
（出現したときは、その特徴の値がtとなる）

・特徴を出現頻度順にソート
・出現頻度が低い特徴を
フィルタにかけて消去（しょうきょ）
（今回は2回以下しか
現れないものを消去）

1	{z,r}
2	{z,x,y,s,t}
3	{z}
4	{x,s,r}
5	{z,x,y,r,t}
6	{z,x,y,s,t}

ソーティングと
フィルタリング後の
学習データ

FP-Growthアルゴリズム

FP-Growthアルゴリズムの例（つづき）

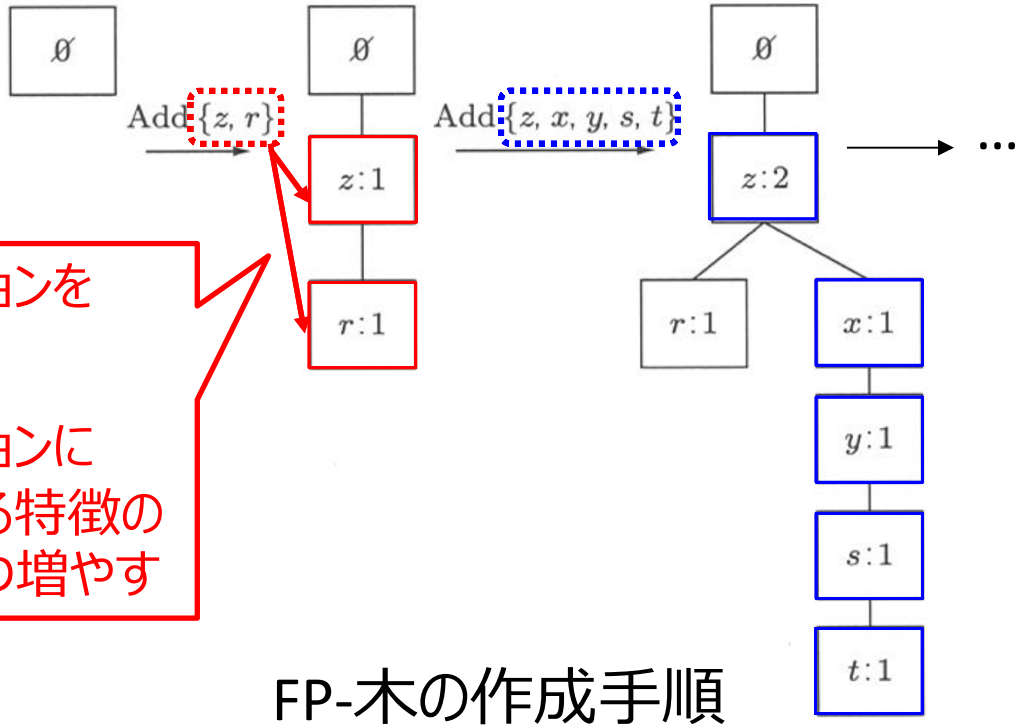
2. FP-木（Frequent Pattern Tree）の作成
   - 最初にnullというカテゴリを付けた根ノード（ルート）を用意し、トランザクションを順にその木に挿入する
     そうにゅう

1	{z,r}
2	{z,x,y,s,t}
3	{z}
4	{x,s,r}
5	{z,x,y,r,t}
6	{z,x,y,s,t}

ソーティングと
フィルタリング後の
学習データ

トランザクションを
木に挿入

トランザクションに
含まれている特徴の
出現数を1つ増やす



FP-木の作成手順

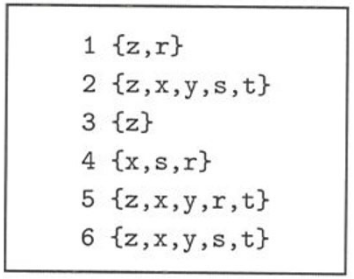
FP-Growthアルゴリズム

- FP-木挿入アルゴリズム
  - T：トランザクション、FP：FP-木

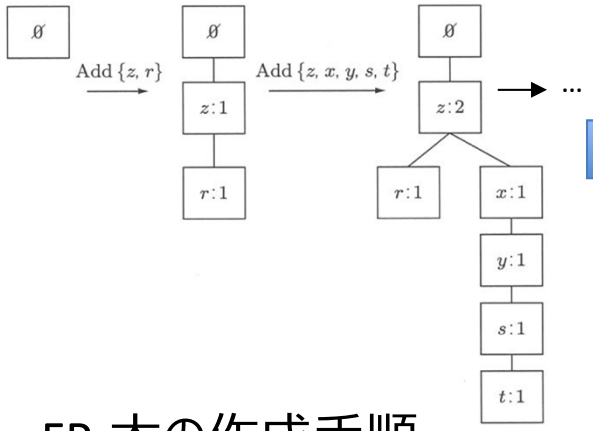
トランザクションTの先頭(せんとう)要素tを取り出す
if $t \subset FP$ **then**
　　tに対応するノードNのカウントを1増やす
else
　　ノードNを作成し、カウントを1としてFPにつなぐ
end if
if T に残りの要素がある **then**
　　FP−木挿入(Tの残りの要素, NをルートとするFP−木)
end if

FP-Growthアルゴリズム

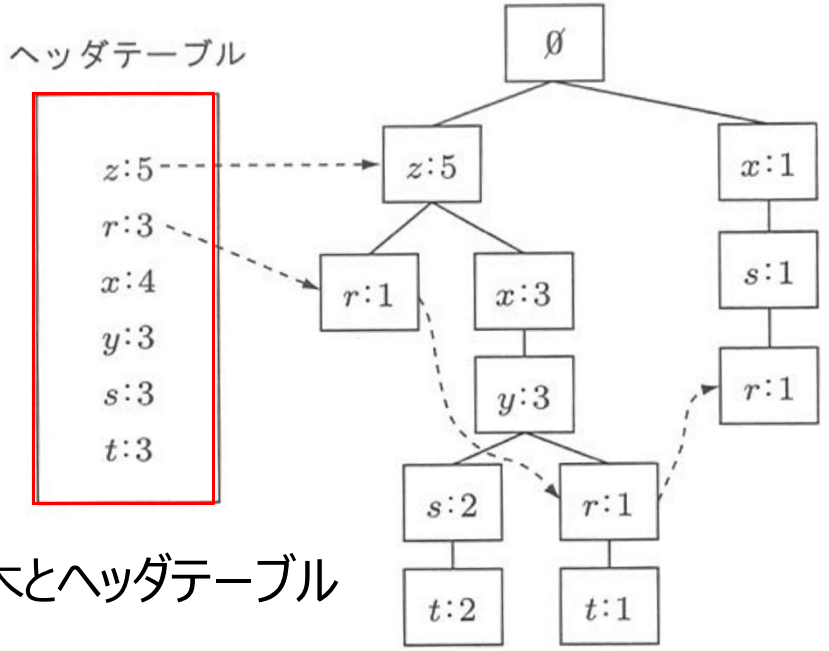
- FP-Growthアルゴリズムの例（つづき）
  3. 完成(かんせい)したFP-木に対して特徴を見出(みだ)しとするヘッダテーブルを作成し、その頻度を記録する
  4. FP-木に出現する同じ要素をリンクで結ぶ
     - リンクを辿って集めた出現数は、全体のトランザクション集合での出現数と一致



ソーティングとフィルタリング後の学習データ



FP-木の作成手順

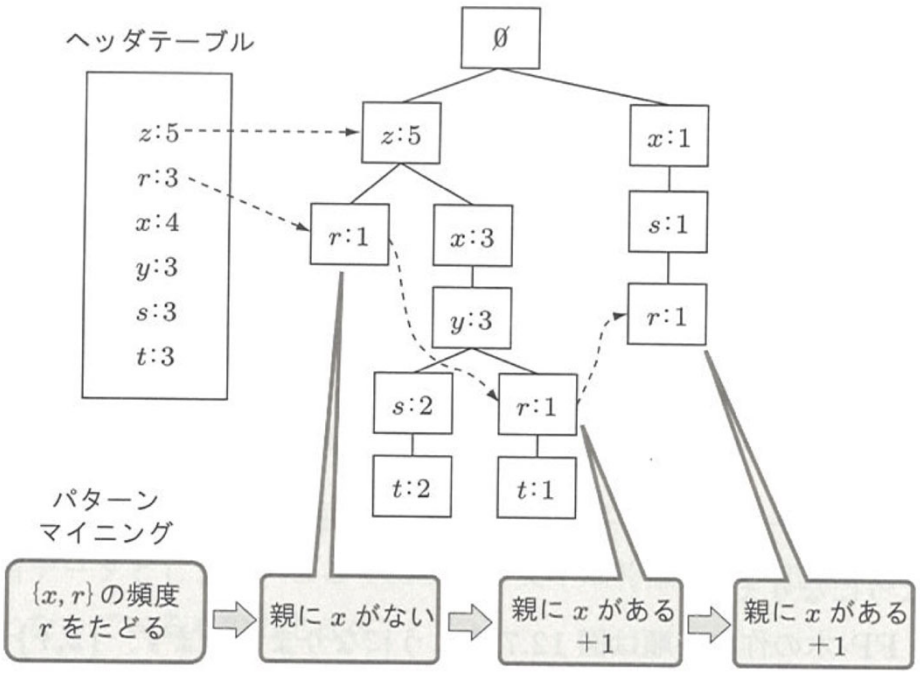


FP-木とヘッダテーブル

FP-Growthアルゴリズム

FP-Growthアルゴリズムの例（つづき）

5. FP-木に対して、パターンマイニングを行う



FP-木のマイニング

例えば

頻出項目抽出で{x,r}の頻度を求める

1. ヘッダテーブルから頻度の少ない方を選ぶ
   - rが3回、xが4回なので、rのリンクを辿りながら頻度を計算
2. 最初のrから親を辿る
   - xが無いので、{x,r}が共起していないと判断
3. 次のrのリンクを辿り、その親を辿る
   - xがあったので、rの頻度をカウントに加える
4. 親にxが出現するパスになるrの頻度を足し続けて、最終的な{x,r}の頻度を得る

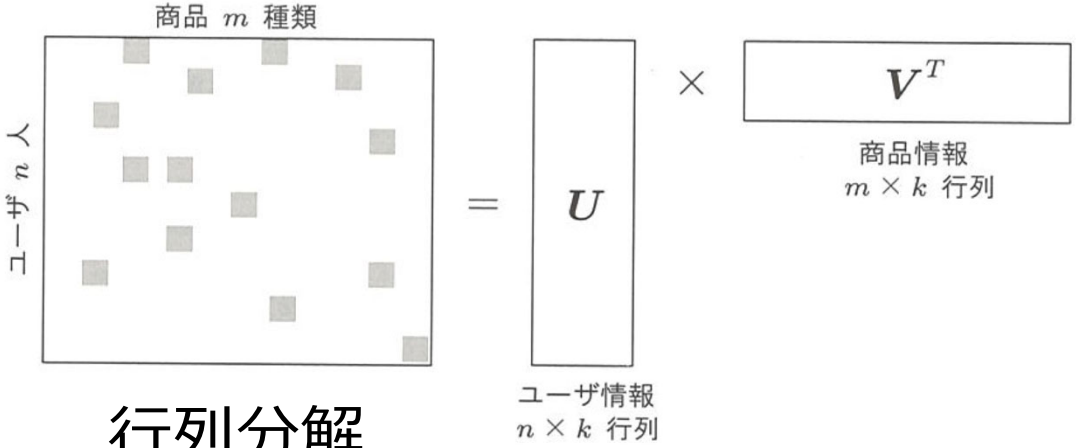
推薦システムにおける学習

构建个人推荐系统

- 個人(こじん)に対して推薦(すいせん)を行うシステムの構築
  - トランザクションデータ（1個のデータが、1件の売り上げに相当）を個人に対応付けてまとめると、どの個人がどの商品を購入しているのかがわかる
- **協調フィルタリング** 协同过滤
  - 推薦システムにおける学習手法の1つ
    - 新規(しんき)ユーザがある商品を購入した時、購入パターンが似ているユーザを探し、「そのユーザが購入していて、かつ新規ユーザが購入していない商品」を推薦するというのが基本的な考え方
  - 購入データを低次元の行列に分解し、ユーザ・商品の特徴を低次元のベクトルで抽出する

推薦システムにおける学習

- 協調フィルタリング（つづき）
  - 購入データを低次元の行列に分解し、
    ユーザ・商品の特徴を低次元のベクトルで抽出する
    将购买数据分解成低维列，用低维向量提取用户/产品特征
  - Matrix Factorization
    - 疎らなデータを低次元行列の積に分解 将稀疏数据分解为低维列的乘积
      - Alternating Least Squares手法：値のある要素のみを使って行列分解
      - Non-negative Matrix Factorization：分解した行列の要素が全て非負（ひふ）



行列分解

演習問題12-3（10分間）

- 実際に協調フィルタリングが用いられている事例を調べなさい