エンピリカルデータを対象とした相関ルール分析ツールNEEDLE

森崎　修司
EASEプロジェクト/奈良先端科学技術大学院大学

動機

- **ソフトウェア開発プロジェクトの特性をまとめたデータ（コスト、プロフィール、バグ）から規則性、傾向、例外を抽出したい。**
  - **改修プロジェクトのテスト工数比率は新規プロジェクトのテスト工数比率よりどのくらい大きい？**

プロジェクト特性データの例

ID	開発種別	・・・	アーキテクチャ	・・・	要件定義工数	結合試験工数	総合試験工数	・・・	不具合密度	・・・
001	新規	・・・	3階層CS	・・・	80	230	200	・・・	0.124	・・・
002	改修	・・・	スタンドアロン	・・・	120	200	360	・・・	0.086	・・・
003	拡張	・・・	3階層CS	・・・	60	260	400	・・・	0.158	・・・
・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・

相関ルール分析

- **対象データに含まれる「AならばB」という規則（相関ルール）を全て列挙する。**
- **列挙されたルールから解釈を与えることができるルールを人手により探し、役立てる。**
- **コンビニの購買履歴から得た相関ルールの例**
  休日に「レジャーシート」を買う顧客は「おにぎり」と「お茶」も同時に買っている。
  「（曜日＝土日） and おにぎり and お茶 ⇒ レジャーシート」
  → 休日には、レジャーシートの配置をおにぎりかお茶に近づけ、発見率、併せ買い率を上げる。

プロジェクト特性データから得る相関ルールの例

- **「(開発種別=拡張) and (アーキテクチャ=3階層CS)**
  ⇒テスト工数比率＝大」
  3階層アーキテクチャの機能拡張プロジェクトではテスト工数比率が高くなる。
  → 3階層アーキテクチャの機能拡張プロジェクトのテスト工数は他よりも大きく見積る。

ID	開発種別	・・・	アーキテクチャ	・・・	要件定義工数	結合試験工数	総合試験工数	・・・	不具合密度	・・・
001	新規	・・・	3階層CS	・・・	80	230	200	・・・	0.124	・・・
002	改修	・・・	スタンドアロン	・・・	120	200	360	・・・	0.086	・・・
003	拡張	・・・	3階層CS	・・・	60	260	400	・・・	0.158	・・・
・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・

相関ルール分析適用の問題点

- **項目の組合せによっては、利用価値の低いルールが多く含まれる。(開発種別とアーキテクチャ、OSとプログラミング言語など)**
- **数値データ(量的変数)を含むソフトウェア特性データにそのまま適用することはできない。**

ID	開発種別	・・・	アーキテクチャ	・・・	要件定義工数	結合試験工数	総合試験工数	・・・	不具合密度	・・・
001	新規	・・・	3階層CS	・・・	80	230	200	・・・	0.124	・・・
002	改修	・・・	スタンドアロン	・・・	120	200	360	・・・	0.086	・・・
003	拡張	・・・	3階層CS	・・・	60	260	400	・・・	0.158	・・・
・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・

プロジェクト特性データへの相関ルール分析適用

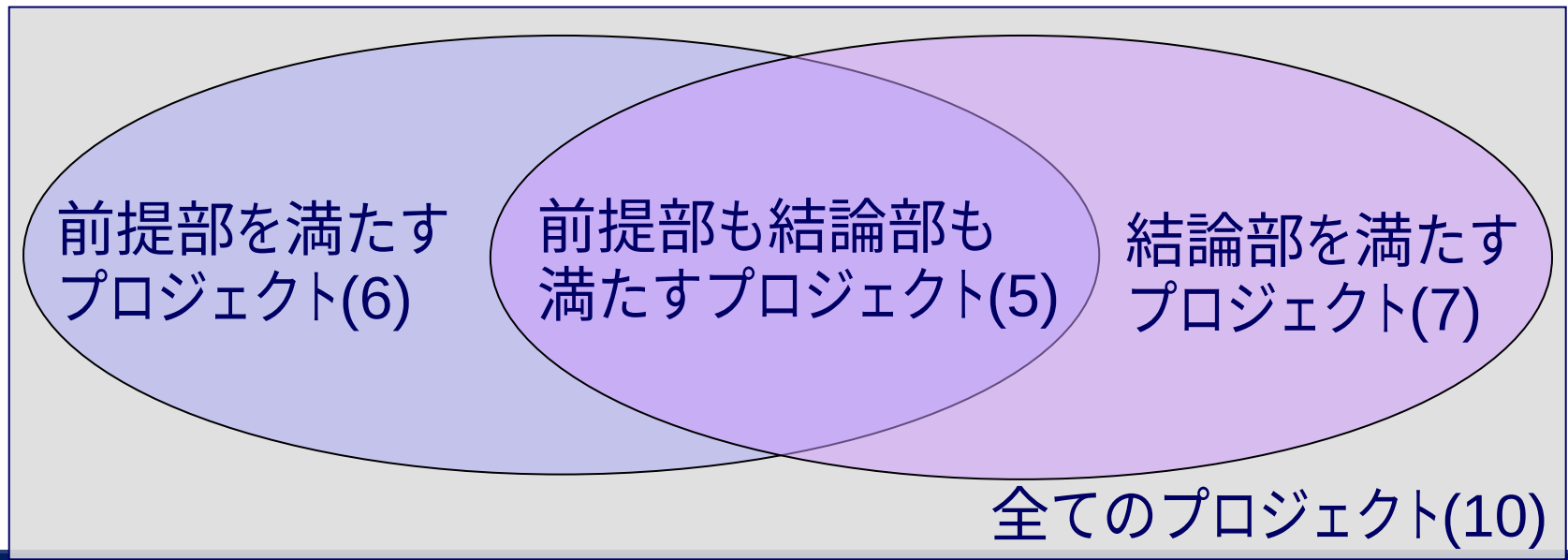
- **分析者が結論部を指定し、ルールを抽出する。**
  例）A かつ B ならば（不具合密度＝低い）
  指定する
- **数値データを扱えるようにする。**
  - **質的ルール（通常の相関ルール）**
    数値データを区間に離散化した後、ルール抽出する。
  - **量的ルール（相関ルールの拡張）**
    結論部以外は数値データを区間に離散化しておき、結論部は数値データの統計値（平均、標準偏差）とする。

質的ルール

- **表記：A ⇒ B，支持度、信頼度、リフト値**
  - **支持度：ルールの出現確率**
  - **信頼度：Aが起こっているときBも同時に起こっている確率**
  - **リフト値：Aがないとき（データセット全体）の信頼度とAがあるときの信頼度の比**
- **数値データの離散化**
  - **あらかじめ数値を区間に置き換え、ルール抽出する。**
    例）（1, 2, 4, 8, 9）→（小, 小, 中, 大, 大）
    小：1~3，**中**: 4~6，**大**:7~9

質的ルールの指標値例

- **（開発種別=機能拡張）** and **（アーキテクチャ=3階層CS）⇒テスト工数比率=大**
  支持度：0.50（5/10）、**信頼度**: 0.83（5/6）、**リフト値**: 1.2（0.83/0.7）



量的ルール

- **表記：A⇒B (平均、標準偏差) 、支持度、基準化平均、基準化標準偏差**
  - **基準化平均 (全体平均に対する倍率)**
    全プロジェクトの平均と前提Aを含むプロジェクトの平均の比
  - **基準化標準偏差 (全体標準偏差に対する倍率)**
    基準化平均と同様
- **結論部 (B) 以外の数値データは質的ルールと同様に区間に分割しておく。**

量的ルールの指標値 (基準化平均)

- **(顧客＝既存)** and **(アプリケーションサーバ＝** WebLogic) **⇒外部委託率 (平均**0.32 **標準偏差** 0.23) **、支持度**：0.38, **基準化平均**1.39, **基準化標準偏差**0.8

顧客	・・・	アプリケーションサーバ	・・・	外部委託率
既存	・・・	WebLogic	・・・	0.32
既存	・・・	自社プロダクト	・・・	0.13
新規	・・・	WebLogic	・・・	0.26
既存	・・・	自社プロダクト	・・・	0.12
既存	・・・	WebLogic	・・・	0.35
新規	・・・	自社プロダクト	・・・	0.17
新規	・・・	WebLogic	・・・	0.28
既存	・・・	WebSphere	・・・	0.24

前提部を含むプロジェクトの外部委託率の平均: 0.32

全てのプロジェクトの外部委託率の平均：0.23

基準化平均＝0.32/0.23=1.39

ルールの利用例（状況把握）

- **開発規模による外部委託率の違い**
  - **（顧客＝既存）** and **（開発規模＝小）⇒外部委託率（平均** 0.28、**標準偏差** 0.2）、**支持度**0.23、**基準化平均** 1.8、**基準化標準偏差** 0.8
    →開発規模が小さい既存顧客の案件は外部委託率が平均より1.8**倍高い。**
- **開発環境によるテスト工数比率の違い**
  - **（テスト環境=エミュレータあり）** and **（実機=既存）⇒プログラムテスト工数比率＝小、支持度** 0.16、**信頼度** 0.8、**リフト値** 1.4
    → 既存のハードウェアを利用した開発でソフトウェアによるエミュレーション環境がある場合、プログラムテスト工数比率が小さい

値と項目は架空のものです。

ルールの利用例（基準値）

- （業種=金融）and（委託形態=派遣）and（開発種別=改修）→ 工数比率（テスト）（平均 0.36，標準偏差 0.17）
- （業種=金融）and（委託形態=業務委託）and（開発種別=改修）→ 工数比率（テスト）（平均 0.38，標準偏差 0.082）

→金融業の改修では派遣も委託も同程度のテスト工数比率であるが、派遣のほうがブレが大きい

値と項目は架空のものです。

抽出ルール

質的ルールの拡張（仮説駆動型例外ルール*）

- **常識ルールと例外ルールのペアを発見する。**
  - **常識ルール：出現頻度の高いルール**
    例）
    抗生物質を服用⇒病気が治る
    支持度0.98、**信頼度**0.85、**リフト値**3.4
  - **例外ルール：常識ルールの前提部を含み、結論部が常識ルールと異なるルール**
    例）
    （ブドウ球菌を保有） and **（抗生物質を服用）**
    ⇒死亡する 支持度0.001、**信頼度**0.92、**リフト値**3.1

*: 鈴木英之進，共通データからの仮説駆動型例外ルール発見，人工知能学会誌vol. 15, no. 5（2000年9月）

例外ルールの利用方法

- **例外ルール例**
  - **常識ルール**

    (アーキテクチャ＝クライアント/サーバ) &
    (開発種別＝拡張) ⇒

    テスト工数比率 (小)　支持度0.32 **信頼度**0.85 **リフト値**2.3
  - **例外ルール**

    (アーキテクチャ＝クライアント/サーバ) &
    (開発種別＝拡張) &
    (協力会社＝新規) ⇒ 例外部分

    テスト工数比率 (大) 支持度0.04 **信頼度**0.95 **リフト値**1.6
- **利用方法**
  - **例外部分からチェックリスト作成**
  - **見積り精度の向上**

値と項目は架空のものです。

相関ルール分析ツール：NEEDLE

- **日本ユニシスMiningPro21®の開発部隊との共同開発**
- **入出力**
  - **入力：CSV形式（カンマ区切り）のエンピリカルデータ**
  - **出力：質的ルール、量的ルール、例外ルールに支持度等を付加したCSV形式**
- **Windows®で実行できるコマンドラインツール群**
  - **数値の離散化等の前処理プログラム**
  - **ルール抽出プログラム**
  - **ルール選択プログラム**

入出力例

- **入力** (40**プロジェクト**)

A	B	C	D	E	F	G	H
プロジェクト	開発種別	業種	アーキテクチャ	開発言語(	OS	要求仕様_	開発期間(, ビー
21	a: 新規開発	b: 通信	a: クライアントサーバ	b: C++	a: UNIX		9
22	a: 新規開発	b: 通信	a: クライアントサーバ	i: その他	a: UNIX	b:かなり明	9
23	a: 新規開発	b: 通信	b: スタンドアロン	e: JAVA	a: UNIX		3
24	a: 新規開発	b: 通信	b: スタンドアロン	i: その他	b: WINDOW	b:かなり明	11
25	a: 新規開発	b: 通信		d: VISUAL	f: WINDOWS 95		16

- **出力** (2000**ルール**)

ルール	支持度	平均	標準偏差	基準化平
(FP/工数＝mean0.628054;std0.608534←(開発種別＝a: 新規開発) & (業種＝f: 製造) & (FP計測手法＝a: IFPUG)	0.050633	0.628054	0.608534	3.743041
(FP/工数＝mean0.628054;std0.608534)←(プロジェクトID＝4.[31.000000,41.000000)) & (開発種別＝a: 新規開発) & (業種＝f: 製造) &	0.050633	0.628054	0.608534	3.743041
(FP/工数＝mean0.469859;std0.545643)←(プロジェクトID＝4.[31.000000,41.000000)) & (開発種別＝a: 新規開発) & (開発期間(月数)＝	0.050633	0.469859	0.545643	2.800238
(FP/工数＝mean0.463618;std0.435696)←(開発種別＝a: 新規開発) & (ピーク要員数＝3.[3.000000,5.000000)) & (FP計測手法＝a: IFPUG)	0.075949	0.463618	0.435696	2.763042

値と項目は架空のものです。

対象データ

- プロジェクト特性データ
  - プロジェクト毎にそのプロジェクトの特徴を表す値を記録したもの
  - 結論部：工数比率、規模、生産性、バグ密度
- 障害データ
  - バグ票
  - 結論部：修正工数、重要度
- その他CSV形式のもの
  (欠損値があってもよい)

適用中・適用検討中の企業様 (社名50音順)

- **適用中の企業**
  - **株式会社日立システムアンドサービス 様**
- **適用を検討いただいている企業**
  - **株式会社NTTデータ 様**
  - **株式会社デンソークリエイト 様**
  - **株式会社東芝 様**
  - **日本ユニシス株式会社 様**
  - **富士フイルム株式会社 様**
  - **他 4社**

EASE PROJECT

詳細利用法、NEEDLE貸与説明会のお知らせ

- **場所**
  本会場6F（**奈良先端大 リエゾンオフィス**）
- **日程候補**
  11/8（**水**）、9（**木**）、14（**火**）、15（**水**）
  いずれも15:00〜17:00
- **参加無料/事前登録制**
  needle-info@empirical.jp**へメールしていただく**
  と登録方法を自動返送いたします