INFORMATION
SCIENCE
TECHNICAL
REPORT

NAIST-IS-TR2011001
ISSN 0919-9527

ソフトウェアエンジニアリング リポジトリを対象とした 例外ルール抽出

森崎 修司, 門田 暁人, 松本 健一

February 2011

NAIST

〒 630-0192
奈良県生駒市高山町 8916-5
奈良先端科学技術大学院大学
情報科学研究科

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan

ソフトウェアエンジニアリングリポジトリを対象とした例外ルール抽出

森崎 修司，門田 暁人，松本 健一
奈良先端科学技術大学院大学　情報科学研究科

1. はじめに

ソフトウェア開発管理が計算機上で行われるようになり，ソフトウェア開発の途中で様々なデータが電子データとして入手できるようになった．不具合情報，構成管理，開発計画や開発工数に関わるデータをはじめとして，これらの情報をソフトウェアエンジニアリングリポジトリと呼び，知見の検出や状況の分析のために利用することが一般的になりつつある．

ソフトウェアエンジニアリングリポジトリの分析には，様々な手法が用いられる．本稿では，例外ルール抽出と呼ぶ，一般的な傾向と対になっている例外的な傾向を計算機により網羅的に抽出する方法を提案する．例外ルールから人手によって有益な知見を得ることができる．

2. 例外ルール

2.1 概要

提案手法は，ソフトウェアエンジニアリングリポジトリに蓄積された複数のデータの間にある１つの比例尺度の項目 r の傾向を把握することを目的としている．傾向を把握することにより，開発プロセス，開発体制等の課題の抽出につながる．対象となるソフトウェアエンジニアリングリポジトリには，複数の名義尺度，順序尺度の項目(n_k)があり，１つだけ比例尺度の項目 r が含まれているものとする．文献[3]のような不具合管理データにおける，修正工数とその他の項目が該当する．

ソフトウェアエンジニアリングリポジトリに蓄積される多くのデータに共通する r と n_k の傾向を X → y のようなルールの形式で抽出するとともに，その傾向と類似していながら稀に存在する例外的な条件を X’ → y’の形式で抽出する．特に X と X’が似通っていながら，y と y’が大きく異なるものを比較することにより，問題の発見や現状の把握に役立つと期待される．ここで，X は n_k の複数の組合せから構成され，y は r の平均値と標準偏差から構成される．一般的な条件と例外的な条件を分析者が目視することにより，課題を推測する．

表１　対象とするソフトウェアエンジニアリングリポジトリ

ID	n_1	n_2	…	$n_{(k-1)}$	n_k	$n_{(k+1)}$	…	r
1			…				…	
2								
…	…	…	…	…	…	…	…	…

2.2 例外ルール抽出例

表１のようなソフトウェアエンジニアリングリポジトリを対象とする．１行が１件のデータに対応し，各列がそれぞれの情報である．１行のデータは，１件のプロジェクト，１件のソフトウェア，１件の構成管理ログ，１件の不具合をはじめとして様々なものが該当する．提案手法では，対象とするソフトウェアエンジニアリングリポジトリには作業工数，開発規模，修正時間をはじめとする比例尺度の項目 r を１つ持つことを想定している．その他は選択式の情報（名義尺度，順序尺度）であることを前提としている．比例尺度の項目が複数ある場合には，１つを残して順序尺度に変換する．

ツール等を用いて n_k が r に影響を与えていると推測される条件を全て列挙する．ソフトウェアエンジニアリングリポジトリに含まれる全ての傾向をルールとして抽出すると膨大な件数になるので，出現頻度などの外的基準を与え，有益である可能性の高いルールをなるべく残しながら，抽出ルール件数を削減する．

自由記述の項目は，ルールを抽出した後，分析者による目視や検討の材料として利用する．たとえば，ルールには “($n_k = v_1$) → r (平均: a, 標準偏差: s)”というようなものが含まれる．a や s の値が突出していれば，$n_k = v_1$ が表わしている状況からその背景や課題が推測できるか試みる．また，a, s の値が望ましくない場合，$n_k = v_1$ が表わす状況を制御したり，回避したりする．

2.3 手順

提案手法の実施手順を図１に示す．

1. ツールによるルール抽出

ソフトウェアエンジニアリングリポジトリに含まれる全ての条件と条件を満たす項目 r の平均値 a と標準偏差 v を X→ y の形式のルールとして求める．ルール抽出は文献[2]の方法を用いる．ルールは一般ルールと呼ぶ．

2. ツールによる例外ルール抽出

一般ルールと類似の形を持つ例外ルール(X’→y’)を抽出する．文献[1][4]等の方法を用いる

3. 目視による解釈可能ルールの選出

抽出した一般ルールと例外ルールから，目視により解釈可能なルールを選出する．ここ

での解釈可能なルールとは，開発における現状を表わしているものとし，対象ソフトウェアやプロジェクトに十分な知識のある分析者が判断するものとする．

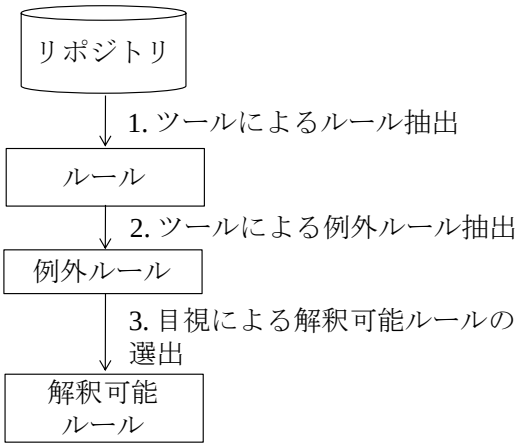

図 1 実施手順

2.4 ルール

2.4.1 一般ルール

ルールは X → y の形式で表わし，前提部 X と結論部 y から構成される．X は対象データの項目とその値を「∧」（かつ）で 0 回以上連結したものである．y は対象データのうち，X にあてはまる対象データの間の平均である．前提部を満たす対象データの件数の割合（データ中の出現頻度）を Pr(X → y)と表わす．Pr(X → y)が大きいほど，多くの対象データにあてはまるルールである．

2.4.2 例外ルール

例外ルールは一般ルールと類似のルールでありながら，結論部 y が異なるものであり，一般ルールを X → y としたときに，X ∧ Z → y'となるルールである．y'は y と同じデータ項目 r であり，a, s の値が異なるものを指す．

参考文献

[1] Balaji Padmanabhan and Alexander Tuzhilin, “Knowledge refinement based on the discovery of unexpected patterns in data mining.” Journal of Decision Support Systems, Vol. 33, No. 3, pp. 309-321 (2002)

[2] Shuji Morisaki, Akito Monden, Haruaki Tamada, Tomoko Matsumura and Ken-ichi Matsumoto, "Mining Quantitative Rules in a Software Project Data Set", 情報処理学会論文誌，Vol. 48, No. 8, pp. 2725-2734 (2007).

[3] Shuji Morisaki, Akito Monden, Tomoko Matsumura, Haruaki Tamada and Ken-ichi

Matsumoto, "Defect Data Analysis Based on Extended Association Rule Mining", In Proc. International Workshop on Mining Software Repository, pp. 17-24 (2007)
[4] 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, pp. 782-789 (2000)