相関ルールに対するクラスタ分析による ソフトウェアプロジェクトのリスク抽出

出張純也[†1] 菊野亨[†1] 水野修[†2]
菊地奈穂美[†3] 平山雅之[†3]

本稿では，ソフトウェア開発プロジェクトの好ましくない状況を抽出を目的として，ソフトウェア開発プロジェクトで収集されたデータに対して相関ルールマイニングを行った．さらに，抽出されたルールを集約するために，ルールに対するクラスタ分析を行った．ワークショップではその効果や妥当性について議論を行いたい．

Extracting Risk Factors of Software Projects by Clustering Association Rules

JUNYA DEBARI ,[†1] TOHRU KIKUNO ,[†1] OSAMU MIZUNO ,[†2]
NAHOMI KIKUCHI [†3] and MASAYUKI HIRAYAMA[†3]

In this paper, we applied association rule mining to the data collected from software development projects for the purpose of extracting the risks of software development projects. In addition, we applied the clustering analysis to the extracted data. We want to discuss the effect and the validity of clustering rules.

1. 議論したい課題

ワークショップでは，2，3 節に述べる現時点までの研究を通じて生じた以下の課題に関して議論を行いたい．

- ソフトウェア開発データに対して，そこからマイニングを実施することの意義について
- 相関ルールマイニングを実施する際のパラメータ設定について
- 相関ルールをクラスタ分析する際に，樹形図から情報を取り出す方法について
- 相関ルールをクラスタ分析する効果について

2. ソフトウェア開発データのマイニング

2.1 研究の背景

ソフトウェア開発の現場では，混乱状態の回避が急務である．そのため，現場で観察収集される様々なデータを有効に活用してプロジェクトを制御するための手法の確立が求められている．

従来研究 [4] では，ソフトウェア開発プロジェクトの早期に行う問題分析アンケートに対して相関ルールマイニングを適用することでリスク要因とプロジェクト混乱の関係の抽出を試み，リスク要因を複数抽出した．

本稿では，IPA/SEC の 2008 年版データ白書 [3] のデータに対して相関ルールマイニングを適用して，プロジェクトが失敗する条件を示す相関ルールを抽出し，得られたルールに対してクラスタ分析を適用することで，ソフトウェアプロジェクトのリスク要因を抽出する事を目指している．

2.2 相関ルールマイニング

相関ルールマイニングは，事象間の強い関係を相関ルールの形で発見する，データマイニング手法の一つである．A ⇒ B という形で表される相関ルールは，A という事象が発生した場合に B という事象が発生することを意味しており，A を前提，B を結論と呼ぶ．

この相関ルールを評価するパラメータとして，支持度 (support)，信頼度 (confidence)，リフト (lift) がある．

相関ルールマイニングの問題点として，これらのパラメータの設定によっては，大量のルールが抽出されたり類似したルールが多数抽出されたりするということが挙げられる．類似したルールを集約する方法とし

†1 大阪大学 大学院情報科学研究科
Graduation School of Information Science and Technology, Osaka University
†2 京都工芸繊維大学 大学院工芸科学研究科
Graduate School of Science and Technology, Kyoto Institute of Technology
†3 情報処理推進機構 ソフトウェア・エンジニアリング・センター
Information-Technology Promotion Agency Software Engineering Center

て，本稿では相関ルールをクラスタ分析するという方法を採用している．

2.3 クラスタ分析

クラスタ分析とは，個体間の距離を基に類似度の高い個体同士を集めてクラスタを作り，対象である個体を分類するデータ分析手法である [2]．相関ルールのクラスタ分析を行う手法としては Gupta らが距離を使ったクラスタリングを提案しており [1]，2 つの相関ルール r_1 と r_2 の距離 $d(r_1, r_2)$ を次のように定義している．

$$d(r_1, r_2) = \frac{|N(r_1) \cap N(r_2)|}{|N(r_1) \cup N(r_2)|}$$

ここで，$N(r_1)$ は相関ルール r_1 に含まれる要素の集合である．本稿では，この距離を相関ルール間の距離として利用する．

クラスタ間の距離の求め方には，最小距離法や最大距離法などがあるが，本研究では最も分類感度が高い方法の一つであるウォード法を用いている．ウォード法は，クラスタ C_i と C_j の距離 $D(C_i, C_j)$ を次のように設定する．

$$D(C_i, C_j) = E(C_i \cup C_j) - E(C_i) - E(C_j)$$

ただし，$E(C)$ はクラスタ C の全ての要素とクラスタ C のセントロイドの距離の二乗の総和である．最も距離の近い二つのクラスタを逐次的に併合し，全ての対象が一つのクラスタになるまで繰り返す．この結果得られる階層構造が樹形図として得られる．

3. 開発データへの適用

3.1 適用データ

本稿で利用するデータは，IPA/SEC が収集したものである [3]．このデータは国内の企業 20 社から収集されたもので，汎用計算機上で動作するアプリケーションソフトウェアの開発プロジェクトが 95%である．収集されたプロジェクトの形態は 9 割以上が受託開発である．それぞれのプロジェクトのデータには，規模・工数・工期・不具合数などの量的なプロジェクトデータ項目に加えて，プロジェクト特性を示す質的データが含まれている．収集されたメトリクスの詳細については文献 [3] に詳しいため，ここでは省略する．

相関ルールマイニングは連続値を含むデータを処理できないため，工数などの連続値をとるデータについては全て中央値で 2 分割を行った．また，順序尺度をとるメトリクスについては 2 群の差が少なくなるように 2 分割を行った．

3.2 適 用 結 果

前節で記述したデータに対して，結論を『発生不具合原因数$^{\star 1}$=多い』，最低支持度を 0.1，最低信頼度を 0.9 として相関ルールマイニングを行った．その結果，251 件の相関ルールを得た．得られた相関ルールの例を以下に示す．

- 設計支援ツールの利用 = 無し ∧ ユーザ担当者のシステム経験 = 経験が不十分 → 発生不具合原因数 = 多い
- 開発プロジェクトチーム内での役割分担・責任所在の明確さ = 不明確 ∧ ユーザ担当者のシステム経験 = 経験が不十分 → 発生不具合原因数　= 多い

次に，抽出された相関ルールに対してクラスタ分析を行った．その結果，相関ルールを葉とする樹形図を得た．得られた樹形図の概形を図 1 に示す．図 1 は得られた樹形図をある高さで切断したものである．樹形図の切断箇所より下部 (図 1 では C_1〜C_5) がクラスタとなる．上に示した 2 つの相関ルールは，いずれも図 1 におけるクラスタ C_1 に分類されるルールである．

現在は，得られた樹形図を，事前に取り決めた個数のクラスタに分割される高さで切断し，クラスタに含まれる相関ルールの前提となっている要因を調査することで相関ルールの傾向把握を行っている．

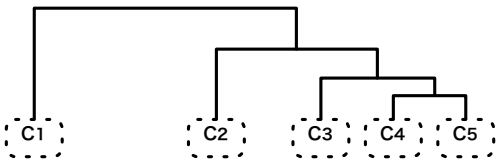

図 1　得られた樹形図

参 考 文 献

1) Gupta, G., Strehl, A. and Ghosh, J.: Distance based clustering of association rules, *Proc. of Intelligent Engineering Systems Through Artificial Neural Networks*, Vol.9, pp. 759–764 (1999).
2) 奥野忠一，芳賀敏郎，矢島敬二，奥野千恵子，橋本茂司，古河陽子：続多変量解析，日科技連出版社 (1976).
3) (独) 情報処理推進機構 ソフトウェア・エンジニアリング・センター：ソフトウェア開発データ白書 2008, 日経 BP 社 (2008).
4) 浜野康裕，天嵜聡介，水野修，菊野亨：相関ルールマイニングによるソフトウェア開発プロジェクト中のリスク要 因の分析，コンピュータソフトウェア， Vol.24, No.2, pp.79–87 (2007).

⋆1 リリース後 6ヶ月以内に報告された不具合の原因数