相関ルールマイニングを利用したソフトウェアプロジェクト混乱要因の関連性に関する調査

出張 純也 †1　尾形 憲一 †1　菊野 亨 †1　水野 修 †2　菊地 奈穂美 †3
平山 雅之 †3

†1 大阪大学 大学院情報科学研究科　†2 京都工芸繊維大学 大学院工芸科学研究科
†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

1 まえがき

1.1 研究の背景

ソフトウェア開発の現場では，混乱状態の早期検出と回避が重要である．しかし，プロジェクトを混乱させるリスク要因の特定や，特定した後のリスク回避方法の検討は経験に基づいて行われることが多く，データによる裏付けは必ずしも多くない．そのため，現場で観察収集される様々なデータを活用して，プロジェクトの混乱を早期に予測し，回避するための手法を確立することは有効である．

我々の研究グループでは，ソフトウェア開発の現場で収集される問題分析アンケートを分析する事で，ソフトウェアプロジェクトの混乱に影響を与えるリスク要因の抽出を試みてきた．従来研究 [2] では，この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し，プロジェクト混乱と特に関連が強いと思われるアンケート項目を複数抽出した．

1.2 研究の目的

プロジェクトの混乱回避のためには，プロジェクト早期のほうが有効な対策を実行しやすいため，出来る限り早い段階で対策を取ることが好ましい．しかし，プロジェクトの早期に得られる情報は限られているため，プロジェクトの混乱を回避するという観点で重点的に着目するべきメトリクスを特定する必要がある．

本研究では，IPA/SECの2008年版データ白書 [1] のデータに対して相関ルールマイニングを適用し，得られたルールの分析を通して，ソフトウェア稼働後に発生する不具合数に影響していると思われるメトリクスの特定を目指す．

Extracting Relationships between Risk Factors of Software Projects with Association Rule Mining
†1 Junya Debari　†1 Kenichi Ogata　†1 Tohru Kikuno
†2 Osamu Miusuno　†3 Nahomi Kikuchi　†3 Masayuki Hirayama
†1 Guraduate School of Information Science and Technology, Osaka University
†2 Graduate School of Science and Technology, Kyoto Institute of Technology
†3 Information-Technology Promotion Agency Software Engineering Center

2 研究のアプローチ

2.1 データの加工

本研究で利用するデータは，IPA/SECが収集したものである [1]．このデータは国内の企業20社から収集されたものである．収集されたメトリクスの詳細については文献 [1] に詳しいため，ここでは省略する．

本研究では，ソフトウェア稼働後の不具合数に関する分析を行うため，[1] に収録されているプロジェクトデータのうち，『発生不具合現象数』のメトリクスに回答値のある816件のプロジェクトデータを利用する．分析には，『発生不具合現象数』『工数』『工期』などの量的なメトリクスだけでなく，『開発プロジェクトの種別』などのプロジェクトの特性を示すメトリクスも含めた76個のメトリクスを利用する事とした．

分析に利用する手法であ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．

さらに，空白値の影響を少なくするため，記入率が低い (空白値が多い) メトリクスの削除を行う．分析に利用するために用意した76個のメトリクスについてそれぞれ記入率を計算し，記入率が25 percentile未満であるメトリクスを削除した．削除後のメトリクス数は57個である．次に，記入率が低いプロジェクトの削除を行った．816件のプロジェクトのうち，記入率が50 percentile未満であるプロジェクトを削除した．

この加工の結果，分析に利用するデータはメトリクス数が57件，プロジェクト数が425件となった．425件のプロジェクトデータ中，不具合が0件であるようなプロジェクトは205件存在した．

2.2 データの分析

2.1節の結果得られたデータに対して，相関ルールマイニングを適用する．相関ルールの結論部は，『発生不具合現象数=0』に設定する．

これは，『発生不具合現象数』のデータを持つプロジェクトの7割のプロジェクトで『発生不具合現象数』が

5件以下となっており，このような偏ったデータ分布では『不具合数が多い』プロジェクトに共通して発生している事象のマイニングは難しいと考えたためである．そこで，不具合が多いプロジェクトでは不具合が少ないプロジェクトと逆の事が起こっていると仮定し，不具合が0件であったプロジェクトで発生している事象を相関ルールマイニングによって明らかにする．得られたルール(不具合0ルールと呼ぶ)の前提に含まれているメトリクスについて，不具合が多いプロジェクト[1]での値を調査する．不具合0ルールに出現しているメトリクスの値と，不具合が多いプロジェクトでの値の一致率を計算し，一致率が低いプロジェクトほど不具合と関係があると考えた．

相関ルールマイニングを行う際のパラメータは，プレ実験の結果，最低信頼度を0.9，最低支持度を0.1と設定した．信頼度とは，相関ルールの前提部が成立する場合に結論部が成立する確率のことである．最低信頼度0.9とは，前提部が成立するプロジェクトの9割以上で結論部が成立しているルールのみを抽出することを意味している．支持度とは，相関ルールの前提部と結論部が同時に成立している確率のことである．最低支持度0.1とは，全体の1割以上，つまり425件のプロジェクトのうち43件以上のプロジェクトで前提部と結論部が成立する存在するようなルールを抽出するということを意味している．この結果，582件の相関ルールが抽出された．抽出されたルールのうち，最も支持度が高かったものが次のルールである．

実績の評価(品質) = a,b ∧ 要員スキル_開発プラットフォーム使用経験 = a,b ∧ SLOC_実績値_SLOC = 中央値未満
→ 発生不具合数現象数 = 0

注　上のルールにおいて，実績の評価(品質) = a,bは，稼働後不具合数が計画値より20%以上少ない(a)，あるいは計画値以下(b)を意味する．要員スキル_開発プラットフォーム使用経験 = a,bは，プロジェクトメンバの開発プラットフォームの使用経験状況が全員が十分な経験をしている(a)，あるいは半数が十分な経験，残り半分はいくらかの経験をしている(b)を意味する．なお，表1中のa,bも基本的に同じであるが，ここでは省略する．

3　分析結果と考察

3.1　分析結果

表1は，得られた不具合0ルールの前提に含まれているメトリクスの値と，不具合が多いプロジェクトでの値の一致率を計算し，一致率が低い順に並び替えたものである．ここでは，一致率が60%以下であったもののみ掲載している．

[1]不具合数が多いプロジェクトの上位5%とする

表1: 不具合が多いプロジェクトとの一致率

不具合0ルールに出現したメトリクス	一致率
実績開発工数 = 中央値未満	2.6 %
SLOC_実績値_SLOC = 中央値未満	2.6 %
実績月数_プロジェクト全体 = 中央値未満	4.9 %
月あたりのSLOC = 中央値未満	7.3 %
要員スキル_分析・設計経験 = a,b	33.3 %
新技術の利用 = b:なし	54.2 %
要員スキル_開発プラットフォーム使用経験 = a,b	57.1 %
要員スキル_業務分野経験 = a,b	58.8 %

3.2　分析結果の考察

表1では，実績開発工数，SLOC，実績月数などプロジェクトの規模を示すメトリクスの一致率が極めて低い．このことから，プロジェクトの規模が小さいプロジェクトでは不具合数が少ないということと，不具合数が多いプロジェクトではプロジェクトの規模が大きいということがわかる．月あたりのSLOCから，開発の速度に余裕があるプロジェクトでは不具合数が少ないということと，開発の速度が速いプロジェクトでは不具合数が多いということがわかる．また，要員スキルが高いプロジェクトでは不具合数が少ないということ，不具合が多いプロジェクトでは要員スキルが低い傾向があるということがわかる．このことから，稼働後の不具合を未然に防ぐという観点では，表1に挙げられるメトリクスを重点的に観察することが重要であると考えられる．

これらは，いずれも開発の現場では経験的に知られている事である．客観的に集められたデータを分析することによって，経験的に知られている事実に対して裏付けが為されたと考える事ができる．

4　まとめ

本研究では，IPA/SECの収集したソフトウェア開発データ[1]に対して，記入率の低いデータを削除した後に相関ルールマイニングを適用した．さらに，得られたルールを分析することで，ソフトウェア稼働後の不具合数に関連があると思われるメトリクスを複数抽出した．

謝辞　この研究の一部は，日本学術振興会科学技術研究費補助金特別研究員奨励費(課題番号：21・3963)，及び日本学術振興会科学技術研究費補助金基盤研究(C)(課題番号：21500035)の助成を受けている．

参考文献

[1] (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).

[2] 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).