データの経時的な性質変化を考慮した分析

角田雅照[†1] 門田暁人[†1] 松本健一[†1]
波多野亮介[†2] 福地豊[†2]

長期にわたりソフトウェア開発データを収集，蓄積していると，途中でデータの性質が変化する場合があり，それを考慮せずに分析すると，一般性の低い結果が得られる可能性がある．本稿では，その可能性を考慮し，変数間の関連が強く，かつその強さが経時的に変化しにくい関連を特定する方法を提案する．

Analysis Considering Change of Data Characteristic Over Time

MASATERU TSUNODA,[†1] AKITO MONDEN,[†1]
KEN-ICHI MATSUMOTO,[†1] RYOSUKE HATANO,[†2]
and YUTAKA FUKUCHI[†2]

When software development data is collected and stored for a long time, there is possibility that characteristic of the data is changed in the middle. Without consideration of that, analysis of the data may bring less general results. Considering the issue, we propose an analysis method which identifies relationships of variables which are strong and whose strength is stable over time.

1. はじめに

ソフトウェア開発マネジメントにおいて，定量的データ分析に基づいた知見を得て，それをマネジメントに活用することは非常に重要である．分析では，過去のプロジェクトにおいて収集，蓄積したデータを用い，目的変数（生産性など）とその他の説明変数（プログラミング言語など）との関係を分析したり，説明変数（ファンクションポイントなど）に基づいて目的変数（開発工数など）を予測するモデルを構築したりする．一般には，過去に蓄積したデータ全てを用いて分析，予測モデル構築を行う．

ただし，長期間にわたってデータを収集，蓄積している場合，その途中でプロセス改善などが行われ，組織の状態が変化する場合がある．組織の状態変化はデータの性質にも変化をもたらす，すなわち目的変数と説明変数との関係を変化させる可能性がある．蓄積されたデータの性質が途中で変化している（目的変数と説明変数との関係が異なるデータが混在している）にもかかわらず，全てのデータを用いて説明変数と目的変数の関係を分析したり，予測モデルの構築を行ったりすると，一般性の低い分析結果が得られたり，予測精度の低いモデルが構築される可能性がある．

近年，このようなデータの性質が経時的に変化する問題に対して，注目が高まりつつある．例えば，データの時間的変化を考慮し，直近に収集された n 件（n は自然数）のデータを用いて予測モデルを構築する方法などが提案されている [1),3)]．本稿では，目的変数と説明変数との関連の分析において，経時的な変化が少ない関連を特定することを試みる．

2. 経時的変化を考慮した分析

2.1 リスク項目とプロジェクトの結果

ある程度規模の大きなプロジェクトでは，コスト超過，納期遅延などにつながるリスクを持ちうる事項（例えば「プロジェクト計画書が作成され，レビューされているか? [2)]」など．以降リスク項目と呼ぶ）に関して，プロジェクトマネージャがその事項の状態をプロジェクト初期に評価し（計画書がレビューされていないならリスク高，レビューされているならリスク低など），リスクの高い事項を監視することが行われる．リスク項目はコスト超過などのプロジェクトの結果と関連を持つが，リスク項目によって関連の強さが異なり，これは必ずしも明らかではない．プロジェクトの結果と関連の強いリスク項目を，分析により明らかにすることにより，重点的に管理すべきリスク項目を特定することができる．

ただし，プロセス改善などが行われ，組織の状態が変化した場合，リスク項目とプロジェクトの結果との関連が変化する可能性がある．具体的には，プロセス改善が進

表 1 適合率，再現率の計算例

プロジェクトID	開始日	リスク項目1	リスク項目2	コスト
P001	06/07/03	高	高	超過
P017	07/03/11	高	*低*	超過
P025	08/04/24	低	低	非超過
P038	09/12/17	低	*高*	非超過
P046	10/08/10	*低*	高	超過
P054	11/01/09	*高*	低	非超過

[†1] 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
[†2] 株式会社日立製作所
Hitachi, Ltd.

表 2　適合率，再現率の計算例

適合率: 75%

プロジェクトID	リスク項目1	コスト超過
P001	1	1
P002	*1*	*0*
P003	1	1
P004	1	1

再現率: 75%

プロジェクトID	リスク項目1	コスト超過
P015	1	1
P016	*0*	*1*
P017	1	1
P018	1	1

んだ場合，ある種のリスク項目については，リスクの高低がプロジェクトの結果に影響しなくなる可能性がある．逆に，組織の状態がどのような場合でも，プロジェクトの結果に強く影響するリスク項目が存在する可能性もある．例えば，表 1 のリスク項目 1 と 2 は，全データではコスト超過に対する関連の強さは同じであるが，直近の 3 件では項目 1 のほうが関連が弱い．より一般性の高い分析結果を得るためには，後者のリスク項目を特定する必要がある．そこで，後者のリスク項目を特定する方法を提案する．

2.2 分析手順

以下の手順により，プロジェクトの結果と一定以上の関連を持ち，かつその関連の強さが経時的にあまり変化していない（組織の状態の変化に影響を受けにくい）リスク項目を特定する．

手順 1 各プロジェクトにおいて，複数のリスク項目それぞれについて評価値（3:リスク高，2: リスク中，1: リスク低など）が記録されており，かつプロジェクトの結果（予算と実績との差額など）が記録されているとする．このとき，リスク項目，プロジェクトの結果それぞれを二値に変換する（閾値を設定し，それ以上の場合は 1（リスク高，コスト超過など），それ以外は 0（リスク低，コスト非超過など）とする）．

手順 2 リスク項目とコスト超過の関連の強さを示す指標を，Moving-Window 法により求める．関連の強さを示す指標として，判別予測モデルの精度評価などに用いられる適合率と再現率を用いる．リスク項目の値を予測結果とみなし（例えば，ある項目のリスクが高く，かつコスト超過などか起きている場合，その項目は結果を正しく予測しているとみなす），表 2 のように適合率と再現率を計算する．関連の強さを示す指標として，相関係数やクラメールの V を用いることもできるが，適合率と再現率のほうが直感的にわかりやすいため（適合率はある項目のリスクが高い場合にコスト超過などが起きた割合，再現率はコスト超過などが起きた場合にある項目のリスクが高い割合を示す），指標として採用した．

Moving-Window 法は，経時的に変化するデータを分析する場合に用いられる方法であり，計測された順に並べた n 個（n は自然数）のプロジェクトを $p_1, p_2, \ldots, p_n$ とするとき，t 個（t は自然数．$t < n$）のデータ $p_1, \ldots, p_t$，$p_2, \ldots, p_{t+1}$ をそれぞれ用いて分析を行う方法である．Moving-Window 法を用いて，適合率と再現率の平均値と標準偏差を求める．

手順 3 手順 2 で求めた平均値と標準偏差を合成した指標に基づき，プロジェクトの結果との関連が強く，かつその強さが経時的にあまり変化していないリスク項目を特定する．本稿では以下の 4 つの指標を提案する．

1. 適合率の平均値 + (1 - 適合率の標準偏差)
2. 再現率の平均値 + (1 - 再現率の標準偏差)
3. 1 と 2 の合計
4. 1 と 2 の F1 値 （F1 値: 適合率と再現率の調和平均）

1 と 2 の計算時に，平均値と標準偏差の値域を，それぞれの最大値と最小値を用いて[0, 1]に変換する（指標に対する影響を等しくするため）．それぞれの指標は，値が大きいほど，プロジェクトの結果との関連が強く，かつその強さの経時的な変化が少ないことを示す．ここでは単純に平均値と標準偏差の和を用いているが，その他にシャープレシオ（ポートフォリオの評価に用いられる．平均値（収益率）と標準偏差の両方を考慮している）を応用すること [4]も考えられる．

3. おわりに

本稿では，プロジェクトの結果と一定以上の関連を持ち，かつその関連の強さが経時的にあまり変化していないリスク項目を特定する方法を提案した．ワークショップでは，データの経時的な性質変化を考慮することの必要性と，そのための分析方法について議論したい．

謝辞 本研究の一部は，「次世代 IT 基盤のための研究開発」の委託に基づいて行われた．また，本研究の一部は，文部科学省科学研究補助費（若手 B：課題番号 22700034）による助成を受けた．

参考文献

1) Amasaki, S.: Replicated Analyses of Windowing Approach with Single Company Datasets, *Proc. Product Focused Software Development and Process Improvement (Profes)*, Torre Canne, Italy (2011).
2) 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：IT プロジェクトの「見える化」 上流工程編，p.208，日経 BP 社 (2007).
3) Lokan, C. and Mendes, E.: Applying moving windows to software effort estimation, Proc. International Symposium on Empirical Software Engineering and Measurement (ESEM), pp.111-122, Washington, DC, USA (2009).
4) 角田雅照，門田暁人，松本健一：予測モデル構築・評価におけるリサンプリング法の活用，ウィンターワークショップ 2011・イン・修善寺，pp.105-106 (2011).