統合ソフトウェア開発データにおけるプロジェクト欠損率を基にしたコスト予測方法の検討

柿元 健[†1] 渡辺 竜[†2]

複数のソフトウェア開発データを統合しビッグデータとした場合，予測等には役立たないとされている．本稿では，プロジェクト欠損率に着目し，統合ソフトウェア開発データにおけるコスト予測方法について検討する．

A Study of the Cost Estimation Method based on Missing Rate of a Project in Consolidated Software Development Data

Takeshi Kakimoto[†1] and Ryo Watanabe[†2]

In this paper, we discuss about the cost estimation method in consolidated software development data focused on missing rate of a project.

1. はじめに

ソフトウェア開発プロジェクトの管理技法のひとつとして，過去に実施された実績データであるソフトウェア開発データに基づいた定量的な管理技法は広く研究され，開発現場にも広まりつつある．しかし，単独組織において実施されるプロジェクト数はそれほど多くはないため，定量的管理技法を有用に適用可能な量のソフトウェア開発データを蓄積するには長い年月が必要となっている．

ソフトウェア開発データの量を補う方法として，パブリックデータを含めた他組織のデータを自組織のデータと統合することが行われる．しかし，複数組織データを統合したデータ（cross company data）は，定量的管理技法を用いた予測において高い精度は得られないとされている[1)3)]．

統合ソフトウェア開発データにおいて高い予測精度が得られない原因として，組織ごとに収集メトリクスが異なり，統合した際に大量の欠損値が含まれることが挙げられる．また，計測方法や計測対象，単位が異なる場合もあり，さらに組織によって値域も大きく異なる場合がある．そのため，複数組織のソフトウェア開発データを統合しても量は増えるが質が低下してしまい，予測精度の向上にあまり寄与せず，場合によっては精度低下を招いてしまう．

本稿では，統合ソフトウェア開発データのプロジェクト欠損率に注目し，統合ソフトウェア開発データにおいてどのように予測を行うことで精度向上が見込めるかについて検討する．

†1 国立高専機構 香川高等専門学校 電気情報工学科
Department of Electrical and Computer Engineering, National Institute of Technology, Kagawa Campas
†2 国立高専機構 香川高等専門学校 専攻科 創造工学専攻
Advanced Course in Industrial and Systems Engineering, National Institute of Technology, Kagawa Campas

2. プロジェクト欠損率を基準とした予測方法

2.1 概　　要

統合ソフトウェア開発データにおいて，各プロジェクトの欠損率を基準としてどのような予測方法を用いることで精度向上が見込めるかを分析する．予測手法には，コサイン類似度に基づく補完法により欠損値補完を行い，重回帰分析によりコストを予測する方法を用いた．

2.2 欠損率による層別

欠損率の違いにより同一組織のデータを抽出することを意図した方法で，各プロジェクトの欠損率によって層別を行い，層別した欠損率ごとのデータセットにおいてコスト予測を行う．ただし，今回用いたデータセットにおいては同じ欠損率でも複数の欠損パターンが含まれているため，完全に同一組織のデータとはなっていないと考えられる．

2.3 欠損率によるプロジェクト選択

欠損率が高いプロジェクトは予測精度に悪影響を与えるという仮定の基に，閾値以上の欠損率のプロジェ

クトを削除，すなわち，閾値以下の欠損率のプロジェクトを用いてコスト予測を行う方法である[2)]．つまり，閾値以下の層別したデータを統合して用いることになる．

2.4 そ の 他

参考情報として，層別やプロジェクト選択を行わず元のデータ全体においてコスト予測を行った結果，および，その結果から欠損率ごとにプロジェクトを抽出し，それらのプロジェクトにおける予測精度の平均値を示す．

3. 評 価 実 験

それぞれの方法におけるコスト予測精度を得るために評価実験を行った．実験方法としては一般的に用いられている leave-one-out cross validation を用いた．

統合ソフトウェア開発データには，ISBSG データ（Release.11）から抽出したプロジェクト 1857 件，メトリクス 12 個，欠損率 53%のデータを用いた．抽出方法は，予測対象メトリクスとする Summary Work Effort が欠損してないプロジェクト，ファンクションポイントのカウント手法が LOC ではないプロジェクト，設計工程以前に得られる（予測できる）メトリクス，および，欠損率が極端に多くないメトリクスを抽出したものである．

また，精度基準としては相対誤差の平均値である MMRE を用いた．

4. 実 験 結 果

表 1 に欠損率ごとのプロジェクト数を，図 1 に評価実験の結果を示す．なお，プロジェクト選択の基準欠損率の値は他と合わせるために，閾値ではなく含まれる最大の欠損率でプロットしている．また，閾値の刻み幅が大きいため，他の結果よりもプロット数が少なくなっている．

欠損率ごとに層別して予測した結果を見ると，欠損率ごとに予測精度は大きく異なっており，欠損率やプロジェクト数とも強い関係性は見られない．データ全体を用いた予測結果から各欠損率の結果を抽出した平均値と比べると，層別して予測すると高い精度が得られる場合とデータ全体を用いて予測すると高い精度が得られる場合が混在している．各欠損率の層別が単独組織のデータの抽出と仮定すると，前者はデータの統合が悪影響を及ぼしており，後者はデータの統合が良い効果を生んでいると言える．しかし，どちらが良いかの判断基準は今回の結果から推測することは難しい．

欠損率によるプロジェクト選択の結果を見ると，閾値を厳しくしすぎるとプロジェクト数の不足からか高い精度が得られていない．しかし，欠損が非常に多いプロジェクトのみを除外し，ある程度のプロジェクト数が確保されている場合には高い精度が得られている．

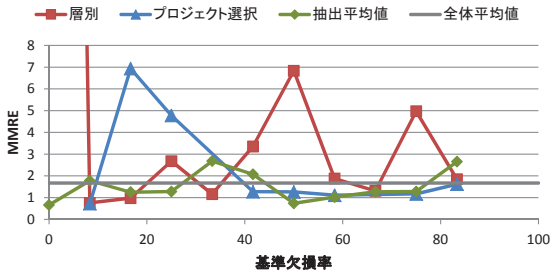

図 1 評価実験の結果

表 1 欠損率ごとのプロジェクト数

欠損率 (%)	0	8.3	16.7	25	33.3	41.7
プロジェクト数	5	67	72	80	477	56
欠損率 (%)	50	58.3	66.7	75	83.3	91.7
プロジェクト数	39	220	225	567	49	0

今後は，欠損パターンごとに同様のことを行う等の欠損率以外の指標についても調査が必要であると考えられる．

5. お わ り に

本稿では，統合ソフトウェア開発データの欠損率に注目し，欠損率で層別した場合，閾値以下の欠損率のプロジェクトを用いた場合について評価実験を行い，コスト予測の精度向上のための予測方法について検討した．

ワークショップでは，統合ソフトウェア開発データの質と量を向上させるための方法について議論したい．

参 考 文 献

1) Kitchenham, B.A., Mendes, E. and Travassos, G.H.: Cross versus within-company cost estimation studies: A systematic review, IEEE Trans. Softw. Eng., Vol.33, No.5, pp.316–329 (2007).
2) 渡辺 竜，柿元 健：ソフトウェア開発コスト見積における類似性に基づく欠損値処理の改良，情報科学技術フォーラム講演論文集，Vol.13, No.1, pp.123–124 (2014).
3) Zimmermann, T., Nagappan, N., Gall, H., Giger, E. and Murphy, B.: Cross-project defect prediction: a large scale experiment on data vs. domain vs. process, In Proc. of the the 7th joint meeting of ESEC and FSE, pp.91–100 (2009).