日本オペレーションズ・リサーチ学会
2023 年春季研究発表会

ソフトウェア開発工数予測における数量化 III 類の活用

非会員　東京都立大学　*田名網圭太　TANAAMI Keita
05000041　東京都立大学　肖霄　XIAO Xiao

1. はじめに

ソフトウェア開発工数予測はソフトウェア開発プロジェクトの初期段階において, 開発に必要な工数を見積もることである. 正確な見積もりはプロジェクトを成功に導くため, 予測精度の高い手法が必要とされている.

工数予測には過去のプロジェクトデータを用いることが一般的であり, 近年の傾向として, ファンクションポイントなどの量的変数よりも, 開発言語などの質的変数が多く記録されている. 質的変数を用いる従来の工数予測手法としてダミー変数化 [1] があげられるが, 線形回帰がベースであり, 説明変数の個数がサンプル数を上回ることが問題としてたびたび指摘される.

本稿では, 数量化 III 類 [2] を用いて従来手法のデメリットに対処し, さらに類似プロジェクト選出を取り入れたハイブリッド回帰を用いた予測手法を提案する. 特に, 類似度尺度としては量的変数間で線形従属関係がみられる場合にも予測精度の低減を阻止できるとされる重みつけ主成分分析類似度（Weighted Principal Components Analysis Analogy; WPCAA）[3] を用いる.

2. ダミー変数化による質的変数の利用

過去 n 件のプロジェクトデータに m 個の量的変数と l 個の質的変数がある場合を考える. ダミー変数化 [1] では, まず質的変数を 2 値データ (0 または 1) など量的変数に変換する. 例えば, 開発言語という質的変数に A, B, C のカテゴリーがある場合, ダミー変数 d_A, d_B, d_C を作成する. つまり, l 個の質的変数が s 個のダミー変数になる. そして, 線形回帰モデルを用いた目的変数と説明変数の関係は

$$y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \sum_{k=1}^{s} \alpha_k d_k + \varepsilon \qquad (1)$$

のように表される. ここで, y は開発工数, x_j (j=1,...,m) は j 番目の量的変数, d_k ($k = 1, ..., s$) は k 番目のダミー変数, β_0, β_j と α_k は回帰係数, ε は誤差項である.

3. 数量化 III 類による質的変数の利用

数量化 III 類 [2] とは, カテゴリーとサンプルの類似の該当パターンを集めることで, カテゴリーとサンプルの両方を同時に分類する成分分析的なカテゴリー解析法である. 数量化 III 類を質的変数に適用することで, 量的変数に対する主成分分析と同様に, 多くの変数で表現されているデータを少量の合成変数で表現することができる. これにより, データセットに含まれる質的変数は全て量的変数に変換され, 線形回帰モデルによる予測が可能となる. 加えて, 合成変数を 1 つでも用いれば全ての質的変数を考慮できることは, 数量化 III 類を活用する大きなメリットの 1 つである [2]. l 個の質的変数から作成した合成変数のうち u 個を用いる場合, 線形回帰モデルを用いた目的変数と説明変数の関係は

$$y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \sum_{k=1}^{u} \gamma_k c_k + \varepsilon \qquad (2)$$

のように表される. ここで, c_k ($k = 1, ..., u$) は k 番目の合成変数, γ_k は回帰係数である. なお, 本稿で用いる合成変数の個数を $u = 1$ とする.

ハイブリッド回帰では, 線形回帰に用いるデータをプロジェクトの類似性の観点から厳選する. 中でも, 予測プロジェクトと過去プロジェクトの量的変数をもとにユークリッド距離 (Euclidean Distance; ED) を計算し, 近いものからフィットデータとする方法がある. 本稿では, ED の代わりに WPCAA を用いることで, 量的変数間に線形従属関係がみられる場合にも, 相関重みづけにより偏ることなく類似プロジェクトを選出し, 予測精度の低減を防ぐ. 予測プロジェクト t と過去プロジェクト h の距離は式 (3) のように計算される.

$$Dist = \sqrt{\sum_{j=1}^{m} |r_j| \, |x_{tj} - x_{hj}|} \qquad (3)$$

ここで, r_j ($j = 1, ..., m$) は WPCAA による重みであり, x_{tj} と x_{hj} はそれぞれ予測対象プロジェクト t と過去プロジェクト h の j 番目の量的変数である. 類似プロジェクト数を n' とした場合, ハイブリッド回帰を用いた目的変数と説明変数の関係は式 (2) と同じように表されるが, 最小二乗法を用いて求めた回帰係数の推定値 $\hat{\beta}_0$, $\hat{\beta}_j$ と $\hat{\gamma}_k$ は式 (4) によって与えられる.

$$\arg\min_{\beta_0, \beta_j, \gamma_k} \sum_{i=1}^{n'} \left(y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij} - \sum_{k=1}^{u} \gamma_k c_{ik} \right)^2 \qquad (4)$$

ここで, x_{ij} と c_{ik} はそれぞれ i 番目のプロジェクトの j 番目の量的変数と k 番目の合成変数である. よって, 提案手法による予測プロジェクト t の開発工数の予測値 $\hat{y}_t$ は式 (5) によって与えられる.

$$\hat{y}_t = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_{tj} + \sum_{k=1}^{u} \hat{\gamma}_k c_{tk} \tag{5}$$

ここで, x_{tj} と c_{tk} はそれぞれ予測プロジェクト t の j 番目の量的変数と k 番目の合成変数である.

4. 数値実験

本章では, 提案手法と従来手法の予測精度を検証する. 実験にはプロジェクトデータの収集を行う非営利団体である ISBSG (The International Software Benchmarking Standards Group)[4] が収集しているデータセット ISBSG Release 2016 R1.1 を用いる. 工数予測が外部設計終了時に行われることを想定し, 目的変数は「実測工数」, 説明変数のうち量的変数に「未調整ファンクションポイント」, 質的変数に「業種」や「開発手法」など計 7 種類を用いる. 加えて, これらの変数が無欠損で, 信頼性のあるプロジェクトを厳選した結果, 実験で用いるプロジェクトデータは 510 件となった.

実験は leave-one-out cross validation 法に基づいて行う. 予測精度の評価指標として, SA (Standarized Accuracy)[5] を用いる. SA は値が大きいほど予測精度が高いことを示す.

$$SA = \left(1 - \frac{MAE}{MAE_{p0}}\right) \times 100 \tag{6}$$

ただし, $MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$, $MAE_{p0} = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{j<i} |y_i - y_j|$ である. ここで, MAE (Mean of Absolute Error) は i 番目のプロジェクトの実測工数 y_i とその予測値 $\hat{y}_i$ との絶対誤差平均で, MAE_{p0} は $\hat{y}_i$ の代わりに, $j\,(\neq i)$ 番目のプロジェクトの実測工数 y_j をランダムに割り当てた場合の絶対誤差平均である.

図 1 は提案手法と従来手法の SA を比較するグラフであり, 全体的に提案手法が従来手法よりも勝っていることがわかる. これは, 数量化 III 類によって回帰で用いるべき質的変数の要素のみを抽出したことが予測精度の向上をもたらしたと考えられる. また, 質的変数を用いた従来手法と提案手法のどちらにおいても WPCAA の類似度尺度は, 全体的に安定した予測精度をもたらすことに成功しており有効であると考察できる.

5. 結論

本稿ではソフトウェア開発工数予測手法として, WPCAA を用いたハイブリッド回帰のもと, 数量化 III 類を適用することで, ダミー変数化より予測精度が高くなることが示された.

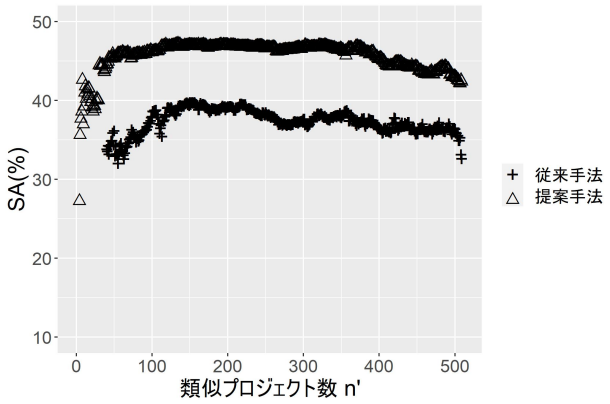

図 1: SA を用いた予測精度の比較

今後の課題としては, ダミー変数化以外の質的変数を用いた予測手法として例えば「データの層別」との比較検証を行う予定である. また, 本稿では類似プロジェクト数 n' を (回帰できる最小件数)$< n' <$ 509 (データセットの全件数 -1) で行ったが, 実際の予測では何件用いるべきかを事前に決めなければならないため, その決定方策について検討する予定である.

参考文献

[1] C.Lokan, E.Mendes, "Cross-company and single-company effort models using the ISBSG database: A further replicated study", Proceedings of 2006 ACM/IEEE International Symposium on Empirical Software Engineering (ISESE'06), pp.75-84, September, 2006.

[2] 菅 民郎, 例題と Excel 演習で学ぶ多変量解析: 因子分析・コレスポンデンス分析・クラスター分析編, オーム社, 東京都, 2017.

[3] Jianfeng Wen, Shixian Li, Linyan Tang, "Improve Analogy-Based Software Effort Estimation Using Principal Components Analysis and Correlation Weighting", Asia-Pacific Software Engineering Conference, 01-03 December, 2009.

[4] International Software Benchmarking Standards Group, URL:https://www.isbsg.org/, Access Date: 2022/01/06.

[5] M.Shepperd, S.Macdonell, "Evaluating prediction systems in software project estimation", Journal of Information and Software Technology, vol.54, pp.820-827, August 2012.