1M-1

重回帰モデルをアンサンブル学習する工数予測モデル

松下 貴徳† 山口 佳子‡ 小西 文章* 内田 眞司‡

奈良工業高等専門学校 専攻科 電子情報工学専攻† 奈良工業高等専門学校 情報工学科‡

奈良先端科学技術大学院大学 情報科学研究科*

1. はじめに

ソフトウェア開発工数の予測は，対象プロジェクトの品質低下，納期遅延，コスト超過などの失敗を避けるために，開発に必要な資源の確保やスケジュール管理を行う上で必要である．その手段として，数学的モデルによる工数予測が行われている．この場合まず，過去のプロジェクトにおいて収集，蓄積されたプロジェクトの特性データを説明変数とし，目的変数である工数との関係をモデル化する．次に予測対象プロジェクトから収集された特性データを構築したモデルに入力することで予測値を得る．予測モデルとして，様々なモデルが提案されているが，線形重回帰分析が広く用いられている．

ところで，ソフトウェア開発プロジェクトはその多様性から，様々なプロジェクトを精度良く予測できるモデルの構築は難しい．構築するモデルの説明変数を増加させることでプロジェクトの多様性を表現するモデルを構築できる．しかし，説明変数を増加させると，たとえ変数選択を行ったとしても過学習（オーバーフィッティング）の問題が生じ予測精度が低下する．また，説明変数を少なくすると，プロジェクトの多様性を十分に表現できなくなり，やはり予測精度が低下する．

筆者らはこれまでに単回帰モデルをアンサンブル学習するソフトウェア開発工数予測モデル(WASRA モデル）を提案した[1]．アンサンブル学習は，それほど精度の高くない学習機械を複数足し合わせ最終的な出力値とすることで，より学習精度の高い学習機械を形成する手法である．アンサンブル学習は，学習精度の高い学習機械を見出す必要がないことであると同時に過学習が起きにくいことが指摘されている[2]．従来手法である対数重回帰分析とアンサンブル手法をテストサンプル法にて予測精度を比較した結果，アンサンブル手法の予測精度の向上とばらつきが抑えられることを確認した．

本稿では，アンサンブル学習するモデルを単回帰モデルから，説明変数を 2 つに増やした重回帰モデルに変更する．一般的に，回帰モデルの説明変数を適切に選択して増やすことでモデルの性能は向上することから提案手法の予測精度向上が期待できる．

2. 予測モデル

予測精度を向上させるために，本稿ではアンサンブル学習するサブモデルの説明変数を 2 つに増やす．言い換えれば重回帰モデルをアンサンブル学習する．モデル式を(1)に示す．

$$y = \frac{\sum_i \sum_j w_{ij} f(x_i, x_j)}{\sum_i \sum_j w_{ij}} \quad where \ i < j \ \cdots (1)$$

式中，y は目的変数，x は説明変数を表している．$f(x_i, x_j)$ は説明変数が 2 つのサブモデル，w_{ij}はサブモデル$f(x_i, x_j)$ の寄与率を表している．

3. 評価実験

3.1. 概要

評価実験の目的は，2 つの従来手法(従来手法 1: 対数重回帰，従来手法 2: WASRA モデル)と式(1)の予測精度とオーバーフィッティングに対する効果を比較し，説明変数を増やした効果を検証することである．評価実験に用いたデータセットは，ISBSG[3]が収集した，20 カ国のソフトウェア開発企業の実績データである．データセットに含まれる全データセットのうち，欠損を含まないプロジェクト 234 件を評価実験に使用した[4]．データセットに含まれる特性は 8 種類あった．

予測精度の比較ではデータセットをランダムに 2 等分し，一方をモデル構築用データセット(fit データセット)，もう一方を評価用データセット（test データセット）とした（未知データの予測）．予測精度の評価基準として，絶対誤差平均値（MMAE）と相対誤差平均値（MMRE）を用いた．実験結果の信頼性を確保するために，この試行を 10 回繰り返し，その平均値を比較した．

An Ensemble Approach of Multiple Regression Model to Software Development Effort Estimation
† Takanori MATSUSHITA
Advanced Electronic and Information Engineering Course, Nara National College of Technology
‡Yoshiko YAMAGUCHI and Shinji UCHIDA
Information Engineering, Nara National College of Technology
* Fumiaki KONISHI
Graduate School of Information Science, Nara Institute of Science and Technology

オーバーフィッティングに対する効果の比較では，まず fit データセットと test データセットが同じデータセットを用いて予測を行う（自己予測）．その後，自己予測と未知データの予測の差を比較する．この差が小さいほど，オーバーフィッティングを軽減できていると言える．

3.2. 実験結果

表 1 に未知データの予測精度を示す．従来手法 2 と比較すると提案手法の MMAE, MMRE はともに向上していた．また，提案手法の予測精度のばらつきはおさえられていた．従来手法 1 と比較すると提案手法の MMAE は向上し，MMRE については同程度であった．予測精度のばらつきはおさえられていた．

表 2 に各手法の自己予測精度と未知データの予測との差を示す．未知データの予測精度との差はそれぞれの平均値の差である．MMAE, MMRE ともに提案手法の差が最も小さかった．以上の結果から，従来手法と比較して提案手法は，未知データに対する予測精度を向上させるとともに，オーバーフィッティングを軽減させることが確認できた．

4. まとめ

本稿では，重回帰モデルをアンサンブル学習するソフトウェア開発工数予測モデルを提案した．評価実験の結果，単回帰モデルをアンサンブル学習するモデルと比較して予測制度が向上することを確認した．また，オーバーフィッティングを軽減できることも確認した．説明変数を増やしても，オーバーフィッティングを軽減しつつ予測精度の向上が可能である．

今後の課題としては，他のデータセットに対する提案手法の評価実験，予測精度をさらに向上させる最適な説明変数の個数を明らかにすることがある．

5. 参考文献

[1] 内田眞司, 藤原寛仁, 内垣聖史, “単回帰モデルのアンサンブル学習によるソフトウェア開発工数予測の試み”, ウィンターワークショップ 2013・イン・那須, pp.15-16, 2013.

[2] R. E. Schapire, Y. Freund, P. Bartlett and W. S. Lee, “Boosting the margin: A new explanation for the effectiveness of voting methods. “, The Annals of Statistics, 26(5):1651-1686, 1998.

[3] International Software Benchmarking Standards Group, “ISBSG Estimating Benchmarking and research suite release 9”, 2004.

[4] 小西文章, 内田眞司, 戸田航史, 門田暁人, “説明変数のランダムサンプリングを用いた対数重回帰モデルによるソフトウェア開発工数予測の試み”, 情報処理学会第 175 回ソフトウェア工学研究会, pp.1 - 8, 2012.

表 1 未知データの予測精度

	MMAE			MMRE		
	従来手法 1	従来手法 2	提案手法	従来手法 1	従来手法 2	提案手法
平均	2834.8	2790.9	**2680.5**	0.54	0.61	**0.56**
標準偏差	456.3	475.8	**254.8**	0.09	0.06	**0.05**
最大値	3662.9	3598.0	**3094.1**	0.73	0.73	**0.64**
最小値	2203.8	2185.0	**2326.7**	0.43	0.54	**0.49**
中央値	2808.2	2877.2	**2661.0**	0.53	0.60	**0.56**

表 2 自己予測精度と未知データの予測との差

	MMAE			MMRE		
	従来手法 1	従来手法 2	提案手法	従来手法 1	従来手法 2	提案手法
平均	1905.8	2585.4	**2674.9**	0.76	0.51	**0.54**
標準偏差	140.0	296.2	**258.9**	0.16	0.02	**0.03**
最大値	2146.0	3091.3	**3016.1**	1.02	0.55	**0.59**
最小値	1739.6	2143.7	**2213.9**	0.55	0.46	**0.49**
中央値	1856.2	2506.0	**2700.1**	0.75	0.52	**0.54**
未知データの予測との差	929.0	205.5	**5.6**	0.22	0.10	**0.02**