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

内田眞司[†1], 藤原寛仁[†2], 内垣聖史[†3]

本稿では，ソフトウェア開発工数予測においてオーバーフィッティングの軽減を目的とした単回帰モデルのアンサンブル学習による予測モデルを用いた工数予測について述べる.

An Ensemble Approach of Simple Regression Models to Software Effort Estimation

Shinji UCHIDA[†1], Hiroto FUJIWARA[†2], Satoshi UCHIGAKI[†3]

In this paper, we explain software effort estimation by using an ensemble approach of simple regression models to reduce the overfitting problem.

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，過去のプロジェクトにおいて収集，蓄積されたデータ(以降，フィットデータという)を用いて工数予測モデルを構築し，予測対象プロジェクトから収集されたデータ(以降，テストデータという)を構築され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，重回帰分析が広く用いられている.

しかし，ソフトウェア開発は独自性が高いため，多種多様なプロジェクトを精度良く予測できるモデルの構築が難しい. プロジェクトの多様性を表現できるようにモデルの説明変数を増やすと過学習(オーバーフィッティング)の問題が起きる. 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る.

我々は，ソフトウェア工数予測におけるオーバーフィッティング問題の改善のためにアンサンブル学習に着目する. 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あわせ最終的な出力値とする方法である. 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]. 筆者らは，単回帰モデルをアンサンブル学習する手法を fault-prone モジュール判別問題に対して適用した[1].

本稿では，筆者らが提案したアンサンブル学習に着目したモデルを用いてソフトウェア開発工数予測を試みる.

†1 奈良工業高等専門学校
Nara National College of Technology
†2 JR 西日本株式会社
West Japan Railway Company
†3 奈良先端科学技術大学院大学
Nara Institute Science and Technology

2. 予測モデルと適用実験

2.1. モデル概要

筆者らは，複数の単回帰モデルをアンサンブル学習することにより予測モデルを構築する手法を提案した. 具体的には，各単回帰モデルを，その当てはまり具合を示す寄与率により加重平均する. 式(1)にモデル式を示す. ここで，y は目的変数，x は説明変数，n は説明変数の数，f(x)は単回帰モデル式，w は f(x)の寄与率とする. なお本研究では，単回帰モデル式 f(x)に，対数変換を行った線形単回帰モデル[3]を用いた.

$$y = \frac{\sum_{i=1}^{n} w_i f(x_i)}{\sum_{i=1}^{n} w_i} \quad (1)$$

2.2. 実験概要

実験では International Software Benchmarking Standard Group（ISBSG）が収集したデータセットを用いた．このデータセットには 232 件のプロジェクトが含まれており，変数は 39 個であった．データセットをランダムに二等分し，一方をフィットデータとして予測モデルを構築し，もう一方をテストデータとして開発工数を予測する．この過程を 10 回試行しその平均値により，従来手法(線形対数重回帰モデル)とアンサンブル法のオーバーフィッティングに対する効果の比較と予測精度の比較を行った．オーバーフィッティングに対する効果の比較では，フィットデータとテストデータが同じデータセットを用いて予測を行う(自己予測)．一方，予測精度の比較では，フィットデータとテストデータが異なるデータセットを用いて予測を行う(評価予測)．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．なお，両手法ともに変数選択を行っている．

3. 実験結果

3.1. 予測精度

表 1 に各手法の予測精度を示す．従来手法と比較して MMAE,MMRE 共にアンサンブル手法の予測精度が向上した．また，標準偏差に着目すると，アンサンブル手法の予測精度のばらつきがおさえられている．

表 1 予測精度

	MMAE		MMRE	
	従来手法	アンサンブル法	従来手法	アンサンブル法
平均	2834.8	2555.5	0.54	0.53
標準偏差	456.3	325.0	0.09	0.04
最大値	3662.9	3143.0	0.73	0.58
最小値	2203.8	2133.1	0.43	0.47
中央値	2808.2	2548.2	0.53	0.54

3.2. オーバーフィッティングに対する効果

従来手法とアンサンブル手法において，自己予測と評価予測の MMAE を箱ひげ図で図 1 に示す．従来手法では，自己予測と評価予測の予測精度の差，ばらつきの差が大きい．予測モデルがフィットデータに過度に適合している状態となり，テストデータに対する予測精度が低下している．従来手法で構築されたモデルは，プロジェクトの多様性を十分に表現できているとは言えず，オーバーフィッティングの状態にある．一方，アンサンブル手法は，自己予測と評価予測の予測精度の差，ばらつきの差が小さい．予測モデルがフィットデータ，テストデータそれぞれに適合しおり，オーバーフィッティングによる予測精度の低下を軽減させている．なお，MMRE についても同様の傾向が確認できた．

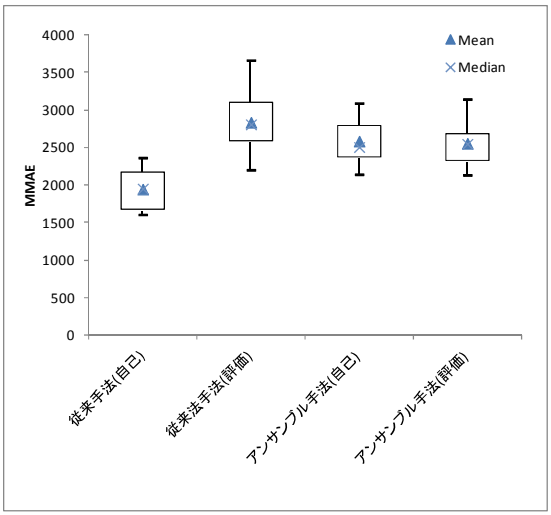

図 1 MMAE の箱ひげ図

4. おわりに

本稿では，ソフトウェア開発工数予測において単回帰モデルのアンサンブル学習による予測モデルを用いた工数予測を試みた．ワークショップでは，アンサンブル手法モデルの有用性と性能向上のための改善点について議論したい．

参考文献

[1] S. Uchigaki, S. Uchida, K. Toda and A.Monden,"An Ensemble Approach of Simple Regression Models to Cross-Project Fault Prediction,"In International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2012), pp.476-481, 2012.

[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] 門田，小林, "線形重回帰モデルを用いたソフトウェア開発工数予測における対数変換の効果", コンピュータソフトウェア, Vol.27, No.4, pp.234-239, 2010.