予測モデル構築・評価におけるリサンプリング法の活用

角田雅照[†1] 門田暁人[†1] 松本健一[†1]

工数見積もりモデルや fault-prone モジュール判別モデルを構築する際，フィットデータとテストデータ間で生じる差異を把握し，それを考慮することにより，予測精度の低下を防ぐことができると期待される．そのための方法として，本稿ではリサンプリング法に着目する．

Using the Resampling Method for Building and Evaluation of Prediction Models

Masateru Tsunoda,[†1] Akito Monden [†1]
and Ken-ichi Matsumoto[†1]

When building effort estimation model or fault-prone module discriminant model, grabbing and considering difference between fit dataset and test dataset is expected to make prediction accuracy high. To perform that, we focus on the resampling method.

1. はじめに

ソフトウェア開発，特に大規模ソフトウェア開発において，品質低下，納期遅延，コスト超過などのプロジェクトの失敗を避けるためには，定量的データに基づくプロジェクトマネジメントが必須となる．定量的データに基づくプロジェクトマネジメントでは，プロジェクトの計画を立案するために，数学的モデルによる予測が行われる．これまで，そのための数多くの工数見積モデルや fault-prone モジュール判別モデルが提案されてきた．予測モデルは過去データ（以降フィットデータと呼ぶ）に基づいて構築され，予測対象の（新規）データ（以降テストデータと呼ぶ）に適用されることにより，工数見積もりや fault-prone モジュール判別を行う．

フィットデータとテストデータが同一の組織から収集されている場合，データの性質が類似しているため，フィットデータにおいて高い精度で予測できたモデルは，テストデータにおいても，ある程度高い精度で予測できると期待される．ただし，フィットデータとテストデータには差異が存在するため，予測精度が完全に同一になることはなく，一般には予測精度が低下する．フィットデータとテストデータ間（標本間）で生じる差異を把握し，それを考慮してモデルを構築することができれば，予測精度の低下を防ぐ，すなわちより高い予測精度のモデルを構築することができると期待される．我々は，フィットデータとテストデータ間で生じる差異を把握する方法として，リサンプリング法に着目する．

†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. ブートストラップ法を用いたモデル構築

ブートストラップ法はリサンプリング法の 1 つであり，標本からランダムにデータを抽出することを繰り返すことにより，標本間のデータにおいて生じる差異を把握する．近年，ソフトウェア工学分野において，ブートストラップ法に基づく統計手法の適用例がいくつか見られるようになっている[1)–3)]．ブートストラップ法では，統計量の分布を把握することができ，例えば（同じ母集団から抽出された）標本によって，どれだけ標本平均に差異が生じうるかを推定することができる．

ブートストラップ法はモデル構築の様々な場面に適用可能であると考えられる．本稿ではブートストラップ法を用いた 2 つのモデル構築方法を提案する．

2.1 シャープレシオに基づくモデル選択

予測精度の高いモデルを構築するためには，フィットデータを用いて予測モデルを構築する際，（ロジスティック回帰分析や分類木などの）複数のモデル候補から最適なモデルを選択したり，説明変数の候補から

最適な変数を選択したりする必要がある．これらの選択を行うためには，相対誤差（MRE: Magnitude of Relative Error）や F1 値，相関係数などの指標が用いられる．ただし，これらの指標値はフィットデータとテストデータで一致せず，フィットデータでの指標値を基準にモデルを構築しても，テストデータでは最適なモデルとならないことが，予測精度が低下する原因の 1 つであると考えられる．

これまで，モデルや説明変数の選択は，指標値の大小関係のみに基づいて行われてきが，提案方法では，ブートストラップ法によりモデル構築に用いる指標値の分散を推定し，指標値の大小だけではなく分散の大きさも考慮してモデルを構築する．この分散の大きさの考慮に，シャープレシオを用いる．

シャープレシオとはポートフォリオ (金融商品の組み合わせ) の運用実績の優劣を示す指標であり，ポートフォリオの収益の高さだけではなく，リスクの大きさも考慮したものである．シャープレシオはポートフォリオの収益率/ポートフォリオの収益率の標準偏差 (正確には，分子から無リスク資産の収益率を減じる) で計算され，収益が高くてもリスクが高い (標準偏差が大きい) ポートフォリオは，値が低くなる．

提案方法では以下の手順でモデルを構築する．

(1) ブートストラップ法により，予測モデル構築のための（F1 値や相関係数などの）指標の大きさと分散を求める．
(2) 指標をポートフォリオの収益率と見なし，指標の大きさと分散に基づいてシャープレシオを求め，シャープレシオが大きくなるようにモデルを構築する．

これにより，テストデータとフィットデータの差異に影響されにくい，予測精度の高いモデルが構築できると考えられる．例えば fault-prone モジュール判別を行うために，ロジスティック回帰分析や分類木などから最適なモデルを選ぶ場合，それぞれの手法の F1 値の大きさと分散をブートストラップ法により求め，シャープレシオが最も大きくなるモデルを選択する．変数選択を行う場合，説明変数と目的変数の相関係数のシャープレシオを同様にして求め，値が大きくなる変数を選択する．

なお，相対誤差など，値が小さいほど予測精度が高いことを示す指標を用いる場合，値が大きいほど予測精度が高くなるように，何らかの変換をあらかじめしておく必要がある．また，分母については標準偏差の代わりに変動係数を用いることが考えられる（平均値が大きいほど標準偏差が大きくなりやすいため）．

2.2 偏回帰係数の差異を考慮したモデル構築

テストデータとフィットデータの差異に影響されにくい，予測精度の高い重回帰モデルを構築することを目的として，重回帰モデルの（標準化）偏回帰係数の差異を考慮したモデル構築法を提案する．提案方法は，線形判別モデルやロジスティック回帰モデルにも同様に適用可能である．以下に手順を示す．

(1) ブートストラップ法により重回帰モデルを複数作成し，他のモデルと偏回帰係数を比較する．この時，偏回帰係数が 1 つでも他のモデルと大きく異なるモデルを除外する．また，偏回帰係数の分散が大きい変数がある場合，説明変数から除外し，もう一度モデルを構築する．
(2) 残ったモデルの中で最も予測精度が高い（相対誤差や残差平方和が最も小さい）モデルを選択する．なお，1 つのモデルを選択する代わりに，残ったモデルを用いてアンサンブル予測を行ってもよい（アンサンブル予測とは，複数のモデルの予測値の平均値などを予測値として用いる方法である）．

3. おわりに

本稿ではフィットデータとテストデータ間で生じる差異を把握する方法として，リサンプリング法に着目し，シャープレシオに基づくモデル選択法と，偏回帰係数の差異を考慮したモデル構築法を提案した．ワークショップでは，予測モデル構築・評価に対するリサンプリング法の有効性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.: Evaluation of Ensemble Learning Methods for Fault-Prone Module Prediction, *Proc. International Workshop on Software Productivity Analysis and Cost Estimation (SPACE'07)* (2007).
2) Mittas, N., and Angelis, L.: Comparing cost prediction models by resampling techniques, *The Journal of Systems and Software*, Vol.81, No.5, pp.616–632 (2008).
3) Mittas, N., Athanasiades, M., and Angelis, L.: Improving analogy-based software cost estimation by a resampling method, *Information and Software Technology*, Vol.50, No.3, pp.221–230 (2008).