
fault-prone モジュール判別におけるアルゴリズム選択のための評価指標の提案

New criterion for algorithm selection on fault-prone module prediction

角田 雅照* 門田 暁人† 松本 健一‡

あらまし 本稿では fault-prone モジュール判別モデルの判別精度を高めるために、新たな評価指標を提案し、その効果を確かめた。

Summary. To enhance the accuracy of fault-prone predication models, we propose new evaluation Criterion. We analyzed the effect of the criteria.

1 背景

ソフトウェアの品質を高めるためには、fault-prone モジュールを判別することが非常に重要である。fault-prone モジュール判別とは、あるモジュールにバグが含まれているかどうかを予測（判別）することである。判別モデルは過去データに基づいて構築され、新規モジュールがバグを含んでいるかどうかを判別する。判別精度を高めるためには、アルゴリズムの選択が重要となる。そのため一般的な手順は以下の通りである。(1) 蓄積されているデータ全てを用いてモデル構築を行う、(2) 蓄積されているデータ全てを用いて判別を行う、(3) 評価指標を算出し、最も指標の値が優れていたアルゴリズムを選択する。判別のモデル構築時に用いられる評価指標のひとつとして、AUC (Area Under the Curve) がある。AUC は 0 から 1 の値をとり、1 に近いほど精度が高いことを示す。本稿ではモデルの判別精度を高めるため、新たな評価指標を提案する。

2 提案する評価指標

2.1 AUC の平均値, 標準偏差

アルゴリズムを選択するための指標として、(1) AUC の平均値、(2) AUC の標準偏差、(3) AUC のシャープレシオ (平均値と標準偏差に基づく。詳細は 2.4 節で後述) を用いることを提案する。AUC は直感的に説明すると正答率であるため、AUC の平均値、標準偏差は存在しない。そこで、 n -fold cross validation を用いて AUC を n 個算出し、それらの平均値と標準偏差を求める。 n -fold cross validation をごく簡単に説明すると (1) 蓄積されているデータの一部を用いてモデル構築を行う、(2) 蓄積されているデータのうち、(1) で用いなかったデータを用いて判別を行うことを n 回繰り返すことである。

2.2 ブートストラップ法の利用

2.2 節の方法では判別を n 回繰り返すため、時間を要する Analogy 法を用いて判別する場合、あまり適さない。そこでブートストラップ法に基づき以下の手順により AUC

*Masateru Tsunoda, 近畿大学理工学部

†Akito Monden, 奈良先端科学技術大学院大学情報科学研究科

‡Kenichi Matsumoto, 奈良先端科学技術大学院大学情報科学研究科

を複数個算出し、平均値と標準偏差を求めることを提案する。(1) 蓄積されているデータ全てを用いてモデル構築を行う、(2) 蓄積されているデータ全てを用いて判別を行う、(3) (2) の判別結果の一部をランダムに抽出する、(4) 評価指標を算出する、(5) (3)、(4) を繰り返す。これにより 1 度だけの判別で AUC を複数個算出できる。

2.3 シャープレシオ

AUC の平均値と標準偏差を同時に考慮するため、シャープレシオを用いることを提案する。シャープレシオはポートフォリオ（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、 $(\text{ポートフォリオの収益率} - \text{無リスク資産の収益率}) \div \text{ポートフォリオの収益率の標準偏差}$ で計算する。収益が高くてもリスクが高いポートフォリオは値が低くなる。分子で無リスク資産の収益率を減じているのは、無リスク資産よりもポートフォリオの収益率が低い場合、ポートフォリオを用いるべきではないためである。本稿ではシャープレシオを $(\text{AUC の平均値} - 0.5) \div \text{AUC の標準偏差}$ で計算する。分子で 0.5 を減じているのは、AUC が 0.5 を下回る場合、判別がランダムと同じとなるためである。

3 実験

実験では、評価指標の違いにより判別精度に違いが生じるのかを確かめた。ここでは変数選択をアルゴリズム選択の一種とみなした。実験に用いたデータは NASA IV & V Facility Metrics Data Program において公開されているデータの 1 つ (KC1) である。提案した指標それぞれに対し、以下の手順で実験を行った。(1) 説明変数の組合せが異なるパターンを複数用意する、(2) 交差妥当性を確認するため、データをモデル構築用とモデル評価用のサブセットに分割する、(3) モデル構築用のサブセットにおいて、説明変数の組み合わせごとに、提案した評価指標を算出する、(4) (3) で算出した評価指標に基づき、説明変数の組み合わせを選択する、(5) モデル評価用のサブセットにおいて、判別精度を評価するために AUC を算出する、(6) (2) から (5) を 4 回繰り返し、(5) で求めた AUC の平均値を求める。5,10-fold cross validation を用いて AUC の平均値と標準偏差を計算した。都合上、ブートストラップ法による評価指標の算出は省略した。

表 1 に各評価指標に基づくモデル選択結果を示す。10-fold により（モデル構築用のサブセットにおいて）AUC の標準偏差を求め、その値が最も小さいモデルを選択すると、最も見積もり精度が高くなった（モデル評価用サブセットでの AUC が最大になった）。

表 1 評価指標に基づくモデル選択結果

AUC (従来法)	5-fold			10-fold		
	AUC 平均値	AUC 標準偏差	AUC シャープレシオ	AUC 平均値	AUC 標準偏差	AUC シャープレシオ
0.63	0.63	0.67	0.68	0.61	0.73	0.62

4 おわりに

本稿では fault-prone モジュール判別モデルの判別精度を高めるために、新たな評価指標を提案した。実験では評価指標の違いにより判別精度に違いが生じることが示された。今後の課題はさらに実験を行い、最適な評価指標を明らかにすることである。

謝辞 本研究の一部は、文部科学省科学研究補助費（基盤 C：課題番号 25330090）による助成を受けた。