ISS-P-039

機械学習を用いたソフトウェア開発工数の二段階予測方法の改善

角花　駿†　　榊原　一紀†　　中村　正樹†
† 富山県立大学　工学部　電子・情報工学科

1. はじめに

ソフトウェア開発においてプロジェクトを成功に導くためには、開発に必要な工数を精度よく見積もることが重要である。本稿では、機械学習を用いて二段階予測方法の改善を検討する。

2. 二段階予測方法

2.1 二段階予測方法の概要

ソフトウェア開発工数の二段階予測方法は、(1)予測の信頼度推定　(2)予測の実行という 2 つのステップから構成される。予測が大きく外れると想定される場合には予測を行わないという方法である。予測の信頼度推定では、予測対象のプロジェクトについて、開発工数を高い精度で予測できそうか否かを判断する。そして、予測の実行では、予測の信頼度が高い場合のみ、開発工数の予測を行う。開発工数予測モデルを構築するにあたって、開発工数への寄与率の高い変数を予測モデルの構築に使う。説明変数として採用しなかった変数を予測の信頼度の推定に用いる。

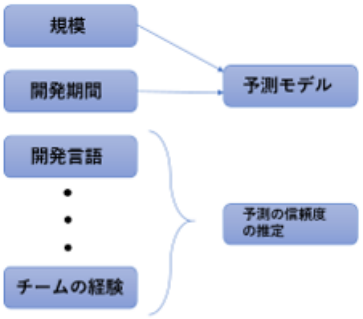

図1. ソフトウェア開発工数の二段階予測方法

Team Exp	Manager Exp	Length	Points Ajust	Effort
小	1 大	4	12	302	5152
小	0 小	0	4	315	5635
大	4 大	4	1	83	805
小	2 小	1	9	145	2569
小	1 小	2	13	214	3913
大	3 小	1	12	247	7854
大	3 大	4	4	103	2422

プロジェクト集合	残差分散
ManagerExp－小	2610577
ManagerExp－大	1520112
TeamExp－小	・・・
・・・	・・・

図 2. 予測の信頼度の低いプロジェクトの特定

2.2 予測の信頼度推定

予測の信頼度の推定方法では、予測モデル構築に用いない変数をカテゴリ変数への変換を行う。この図 2 の例では、各説明変数について、中央値より小さいか否かによって各値に「小」「大」のカテゴリ値を与えている。次に全ての変数についてカテゴリごとにプロジェクト集合を作成する。作成したプロジェクト集合について残差のばらつきに関する尺度を求める。そして、ばらつきが小さいカテゴリに属するプロジェクトについてのみ、予測を行う。

3. 評価実験

3..1 実験概要

本稿では複数の機械学習手法を用いて二段階予測方法をソフトウェア開発企業で収集された実績データを用いて実験的に評価する。データセットを用いてプロジェクトの開発工数を予測し、その精度と予測の信頼度の高い範囲を比較し評価を行う。

3.2 実験の題材

実験の題材として、ソフトウェア開発工数予測研究においてよく用いられている。Desharnais データセットを用いる。本データセットはカナダのあるソフトウェア企業の開発実績データを収集したものであり、無償で一般公開されているため追実験が可能である。

表 1 に、本データセットの各変数の内容を示す。本データセットは 81 プロジェクトのデータが含まれる。

本稿では本データセットにリッジ回帰、ラッソ回帰などを適用して予測精度を比較する。

表1. データセットの各変数の内容

Team Exp	開発チームの経験年数
Manager Exp	プロジェクトマネージャーの経験年数
Length	開発期間
Points Adjust	開発規模
Lang2	開発言語2を使用しているなら1そうでないなら0
Lang3	開発言語3を使用しているなら1そうでないなら0
Effort	実際にかかった工数

4.今後の課題

それぞれの手法を用いた結果を比較し、予測精度と予測可能な範囲がどれほど向上するのか比較検証する。

参考文献

[1] 木下直樹、門田暁人:ソフトウェア開発工数の二段階予測方法のフィージビリティスタディ