様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年　6 月　4 日現在

機関番号：14603
研究種目：若手研究（B）
研究期間：2010～2011
課題番号：22700034
研究課題名（和文）ソフトウェア開発プロジェクト予測フレームワーク

研究課題名（英文）Software Development Project Prediction Framework

研究代表者
角田　雅照（TSUNODA MASATERU）
奈良先端科学技術大学院大学・情報科学研究科・特任助教
研究者番号：60457140

研究成果の概要（和文）：ソフトウェア開発プロジェクトにおいて，精度の高い予測を可能とするための，予測モデル構築フレームワークを確立するための要素技術の研究を行った．フレームワークは(1)特異なデータ(外れ値)の除去，(2)データの層別，(3)予測に最適な変数の選定，(4)データに最適なモデルの選定の 4 つの要素技術から成り.順に実行される.研究期間において，それぞれの実現方法を考案するとともに，それらの効果を確認した．

研究成果の概要（英文）：To achieve high accurate prediction in a software development project, software development project prediction framework and its element technologies was studied. The framework consists of (1) peculiar data point (outlier) deletion, (2) stratification, (3) selecting appropriate variable for prediction, and (4) selecting appropriate prediction model for a dataset, and each element technology is applied in the numerical order. In the research period, each technology was invented and the effects were examined.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	2,100,000	630,000	2,730,000
2011 年度	900,000	270,000	1,170,000
年度			
年度			
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：総合領域
科研費の分科・細目：情報学・ソフトウエア
キーワード：ソフトウエア工学，ソフトウエア開発効率化・安定化，統計数学，モデル化

１．研究開始当初の背景

近年，ソフトウェアは社会のあらゆる場面で利用されており，それに伴い，ソフトウェアの複雑化，大規模化，短納期化がますます進んでいる．このため，ソフトウェア開発の失敗(納期遅れ，品質低下，コスト超過)の危険性が高まっているが，前述のように，ソフトウェアは社会のあらゆる場面で利用されており，開発の失敗は社会に与える影響も非常に大きい．このため，ソフトウェア開発が失敗する危険性を抑えることは，個別の企業にとって重要であるのみならず，社会的なニーズも非常に大きいといえる．

ソフトウェア開発の失敗の危険性を抑えるためには，定量的なデータに基づいて，プロジェクトを管理することが必要となる．定量的なデータに基づいてプロジェクトを管理する際には計画を立案する必要があり，そのために，開発工数の予測やバグが発生しやすいモジュールの予測などが行われる．しか

し，精度の高い予測モデルを構築することは容易ではないことから，現場でのモデルによる予測は必ずしも広く実施されていない．このことが定量的なデータに基づいてプロジェクトを管理する際の障害の１つとなっていた．

予測モデルの研究に取り組んできた過程で，精度の高い予測を行うためには，以下を考慮してモデルを構築する必要があることがわかってきた．

(1) 特異なデータ(外れ値)の除去
(2) データの層別
(3) 予測に最適な変数の選定
(4) データに最適なモデルの選定

そこで，上記の点を考慮した予測モデル構築フレームワークを確立することにより，予測精度を高めることができるとの着想に至った．

２．研究の目的

研究の目的は，予測精度の高いモデルを構築ためのフレームワークを確立することであった．具体的には，以下の４つの要素技術の開発を目指した．

(1) 特異なデータ(外れ値)の除去

モデル構築に用いるデータセットから，特異なデータ(外れ値)を除外する．外れ値とは，身長で例えると，170cm と入力するところを 170m と入力されたようなデータである．

(2) データの層別

モデル構築前に，データの層別を行う．データの層別とは，性別で例えると，男性のデータと女性のデータを分けてモデルを構築することである．

(3) 予測に最適な変数の選定

予測に最適な変数を選定する．ここでの選定とは，既にあるデータから変数を絞り込む(変数選択)のではなく，そもそもどのような変数が予測に一般的に有用であるか，ということを指す．

(4) データに最適なモデルの選定

予測に最適なモデルを選定する．2 つのアプローチがある．1 つ目のアプローチは，データセットの性質を考慮した予測モデルの選定である．データセットの性質とは，性別で例えると，男性のデータが 9 割，女性のデータが 1 割のデータのような，データに偏りなどを指す．もう 1 つは，データの性質を考慮せずに，複数の予測手法を適用し，その結果から，データに適した最適なモデルを選択するアプローチである．

３．研究の方法

(1) 特異なデータ(外れ値)の除去

(a) ソフトウェア工学の範囲に限定せず，従来提案されている外れ値除去法を調査するとともに，その方法を適用するためのツールや，具体的な手順を調査した．
(b) データセットから外れ値を除去する方法を新たに考案した．具体的には，生産性(成果物÷作業時間)に着目して外れ値を除去することを提案した．
(c) データセットに対し，複数の外れ値除去法を適用し，予測精度を確かめた．

(2) 予測に最適な変数の選定

プロジェクト管理のリスク項目を記録したデータ，システム保守におけるシステム構成や要員数を記録したデータ，システム運用におけるシステム構成や要員数を記録したデータなどを用いて分析を行った．

(a) 予測対象の変数(目的変数)を決定する．システム運用に必要な要員数を目的変数として分析した．システム運用に必要な要員数は，コストにかかわる重要な変数であるが，これまでシステム運用に関するデータはほとんど収集されておらず，要員数を予測することは困難であった．
(b) 相関係数などを用いて，目的変数と関連の強い説明変数を明らかにするとともに，多変量解析を行い，説明変数の予測に効果の高い変数を明らかにする．

(3) データセットの性質に適した予測モデルの選定

(a) データセットに人工的なデータを追加し，データの状態を段階的に変化させる．変化させるデータの状態として，以下のものを候補とした．
・成功プロジェクトと失敗プロジェクトの比率
(b) 従来提案されている予測モデルを調査するとともに，その方法を適用するためのツールや，具体的な手順を調査する．例えばマハラノビス・タグチ法などの，データの状態の変化にロバストであると考えられるものを候補とする．
(c) それぞれの状態において，複数の予測モデルを適用し，予測精度の違いを明らかにする．

(4) 複数の予測手法から最適なモデルを選択するための指標の考案

(a) 複数の指標を統合する方法について検討するとともに，類似の指標について調査を行った．指標として以下のものを候補とした．

・シャープレシオなどの，他の分野で用いられている複合的な指標

４．研究成果

(1) データセットの性質に適した予測モデル

コスト超過プロジェクトの発見には，線形判別分析などの判別予測法が用いられる．コスト超過プロジェクトが減少し，コスト非超過プロジェクトの割合が増加した場合における，各種判別予測方法の精度変化を分析した．その結果，コスト超過プロジェクトの割合にかかわらず，協調フィルタリングの予測精度が最も高く，汎用的に適用することができることがわかった．また，コスト超過プロジェクトとコスト非超過プロジェクトの割合に差が少ない場合は，線形判別分析を用いると比較的高い精度が得られ，コスト超過プロジェクトが少ない（35%以下）場合，マハラノビス・タグチ法を予測に用いると比較的高い精度が得られることがわかった（図 1）．

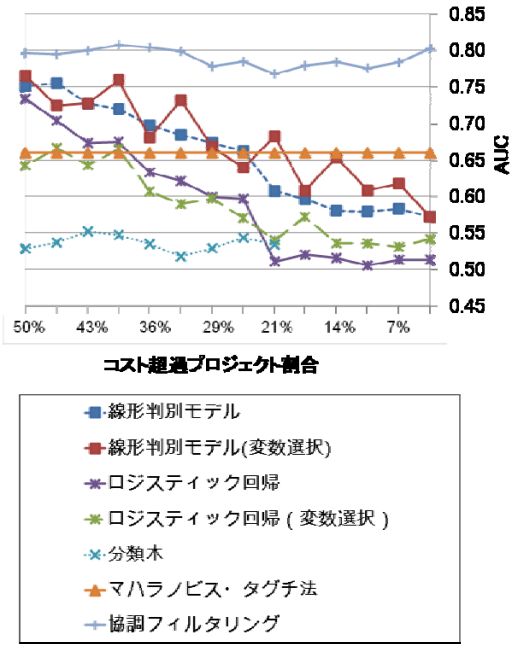

図 1 コスト超過プロジェクト割合と予測精度の関係

(2) 複数の予測手法から最適なモデルを選択するための指標

予測モデルを構築する際，（線形判別分析など）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．ブートストラップ法によりモデル・変数選択に用いる指標値（F1 値や相関係数）の分散を推定し，指標値の大小と分散の大小の両方を考慮してモデルを構築する方法を提案した．提案方法では以下の手順でモデルを構築する．

(I) ブートストラップ法により，予測モデル，説明変数選択のための（F1 値や相関係数などの）指標の大きさと分散を求める．

(II) 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．

類似性に基づく予測法(Analogy 法)による fault-prone モジュール判別を行う際，シャープレシオに基づく変数選択を適用し，効果を確かめた．まず，各指標に基づく説明変数と目的変数の関連の大きさを順位付けし，フィットデータとテストデータの順位の差を比較した．その結果，変動係数で補正したシャープレシオの差が最も小さく，相関係数の差が最も大きかった（表 1）．その後，モデルの精度を確かめたが，モデルの精度自体は向上しなかった．よって，説明変数をより適切に選択できる可能性が示されたといえるが，さらに実験を行い，信頼性の高い結果を得る必要がある．

表 1 フィットデータとテストデーにおける説明変数と目的変数の順位の差

	相関	シャープレシオ	シャープレシオ+変動係数
絶対誤差(AE)	4.3	4.16	3.78
相対誤差(MRE)	0.51	0.36	0.34
相対誤差(MER)	0.69	0.52	0.51

(3) モデル構築に用いるデータセットから，特異なデータ(外れ値)を除外する方法

類似性に基づくソフトウェア開発工数見積り方法（Analogy 法）における外れ値除去法の効果を，三つのプロジェクトデータセット（ISBSG データ，Kitchenham データ，Desharnais データ）を用いて実験的に比較した．比較対象の外れ値除去法は，これまで提案されている 4 種類の外れ値除去法(k-means 法を用いた除去法，LTS を用いた除去法，Cook の距離を用いた除去法，Mantel 相関を用いた除去法）と，本論文で新たに提案する除去法である．提案方法は，Analogy 法の特性を考慮して，類似プロジェクトにおいて工数（生産性）が極端に異なる類似プロジェクトを外れ値とみなし，見積り時の計算から除外する．実験の結果，提案方法は比較した外れ値除去法の中で平均的に最も高い見積り精度を示し，どのデータセットを用いた場合でも見積り精度が大きく低下することはなく，ABRE (Absolute Balanced Relative Error)平均値で最大 28.8%の改善が見られた（表 2）．

表 2 外れ値除去法を適用した場合の ABRE 平均値の変化

外れ値除去法	ISBSG データ	Kitchenham データ	Desharnais データ
提案方法	28.83%	7.61%	-1.54%
Mantel 相関	3.19%	1.30%	-1.11%
k-means 法	-17.86%	0.98%	-0.02%
Cook の距離	-5.99%	6.52%	-11.03%
LTS	11.25%	-0.42%	-4.72%

(4) モデル構築前に，データの層別を行う方法

回帰分析によるソフトウェア開発工数見積において，データに含まれるカテゴリ変数に対し，ダミー変数化を行った場合，層別を行った場合，階層線形モデルを適用した場合の見積精度を比較した．NASA のソフトウェア開発プロジェクトを用いて工数予測モデルを構築し，ABRE (Absolute Balanced Relative Error)に基づいて精度を比較した．その結果，ダミー変数化して重回帰分析した場合と，階層線形モデルを用いた場合では，一方の精度が高いとまではいえなかった．また，データを層別して重回帰分析した場合，その他の方法と比べ，精度が高くならなかった（表 3）．

表 3　各モデルの ABRE 平均値，中央値

	平均値	中央値
ダミー変数 変数選択あり	88.5%	31.9%
ダミー変数 変数選択なし	143.1%	53.0%
ダミー変数 開発規模のみ	82.1%	36.6%
層別 変数選択あり	742.5%	64.2%
層別 変数選択なし	12715.2%	98.9%
層別 開発規模のみ	111.3%	44.7%
HLM 切片にランダム効果	103.5%	40.8%
HLM 開発規模の係数にランダム効果	106.9%	39.5%
HLM 切片と開発規模の係数にランダム効果	107.5%	40.9%
HLM 開発規模のみ 切片にランダム効果	83.1%	39.7%
HLM 開発規模のみ 係数にランダム効果	77.6%	37.7%
HLM 開発規模のみ 切片と係数にランダム効果	77.7%	39.3%

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 10 件）

1. 角田 雅照, 門田 暁人, 渡邊 瑞穂, 柿元 健, 松本 健一, 類似性に基づくソフトウェア開発工数見積もりにおける外れ値除去法の比較, 電子情報通信学会論文誌 D, Vol.J95-D, No.4, pp.895-908, 2012, 査読有
2. Masateru Tsunoda, Akito Monden, Ken-ichi Matsumoto, and Tomoki Oshino, Analysis of Software Maintenance Efficiency Focused on Process Standardization, In Proc. of International Workshop on Empirical Software Engineering in Practice (IWESEP 2011), pp.3-8, November 2011, 査読有
3. Masateru Tsunoda, Akito Monden, and Ken-ichi Matsumoto, Sharpe Ratio Based Index for Building Fault Prediction Model, In Proc. of International Symposium on Software Reliability Engineering (ISSRE 2011) Fast Abstracts, Vol.1, No.6, pp.1-2, November 2011, 査読有
4. Masateru Tsunoda, Akito Monden, Takeshi Kakimoto, and Ken-ichi Matsumoto, An Empirical Evaluation of Outlier Deletion Methods for Analogy-Based Cost Estimation, In Proc. of International Conference on Predictive Models in Software Engineering (PROMISE 2011), No.17, pp.1-10, September 2011, 査読有
5. Masateru Tsunoda, Akito Monden, Jun-ichiro Shibata, and Ken-ichi Matsumoto, Empirical Evaluation of Cost Overrun Prediction with Imbalance Data, In Proc. of International Conference on Computer and Information Science (ICIS 2011), pp.415-420, May 2011, 査読有
6. Masateru Tsunoda, Akito Monden, Mizuho Watanabe, Takeshi Kakimoto, and Ken-ichi Matsumoto, Applying Outlier Deletion to Analosgy Based Cost Estimation, In Proc. of International Workshop on Empirical Software Engineering in Practice (IWESEP 2010), pp.13-18, December 2010, 査読有
7. Masateru Tsunoda, Akito Monden, Ken-ichi Matsumoto, Akihiko Takahashi, and Tomoki Oshino, An Empirical Analysis of Information Technology Operations Cost, In Proc. of International Workshop on Software Measurement (IWSM/Metrikon/Mensura 2010), pp.571-585, November 2010, 査読有
8. 角田 雅照, 天嵜 聡介, ソフトウェア開発工数見積もりにおけるカテゴリ変数の扱い, ウィンターワークショップ 2012・イン・琵琶湖, pp.57-58, January 2012, 査読有
9. 角田 雅照, 松本 健一, 大岩 佐和子, 押野 智樹, カスタムソフトの価格妥当性確認に向けた分析, ソフトウェア工学の基礎 XVIII, 日本ソフトウェア科学会

FOSE2011, pp.249-254, November 2011, 査読有
10. 角田 雅照, 門田 暁人, 松本 健一, 髙橋 昭彦, 押野 智樹,プロセス標準化に着目したソフトウェア保守ベンチマーク確立の試み, ソフトウェア工学の基礎XVII, 日本ソフトウェア科学会 FOSE2010, pp.113-118, November 2010, 査読有

〔学会発表〕(計1件)

1. 角田 雅照, 門田 暁人, 松本 健一, 波多野 亮介, 福地 豊, データの経時的な性質変化を考慮した分析, ソフトウェアエンジニアリングシンポジウム 2011 併設ワークショップ「開発マネジメントを取り巻く環境と課題」, 2011 年 9 月 12 日, 東京

６．研究組織

(1)研究代表者

角田　雅照 (TSUNODA　MASATERU)
奈良先端科学技術大学院大学・情報科学研究科・特任助教
研究者番号：60457140