Vol.2010-SE-167 No.36
2010/3/19

データ分布の偏りがコスト超過判別予測に及ぼす影響

柴田淳一郎[†1]　角田雅照[†1]
門田暁人[†1]　松本健一[†1]

プロジェクトのコスト超過を防ぐためには，コスト超過が発生しそうなプロジェクトを早期に発見し，対応策を実施する必要がある．コスト超過プロジェクトの発見には，線形判別分析，ロジスティック回帰，マハラノビス・タグチ法などの判別予測法が用いられる．ソフトウェアプロセスが改善されると，データセットに占める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が変化し，割合がアンバランスとなり，これらの判別予測法の精度に影響を与える可能性がある．本論文では，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する．実験を行った結果，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，線形判別分析が適しているが，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・タグチ法が予測に最も適しているといえた．

Influence of Imbalance Data toward Cost Overrun Project Prediction

JUN-ICHIRO SHIBATA,[†1] MASATERU TSUNODA,[†1]
AKITO MONDEN,[†1] and KEN-ICHI MATSUMOTO[†1]

To prevent cost overrun of software projects, it is necessary that project managers discriminate projects which has risk of cost overrun early, and they perform countermeasures. Linear discriminant analysis, logistic regression analysis, and Mahalanobis-Tagichi method are used to find cost overrun projects. When software process is improved, balance of cost overrun projects and cost non-overrun projects become unbalance on dataset, and it may affects accuracy of discriminant prediction. In this paper, we analysis accuracy of various discriminant prediction methods, changing ratio of cost overrun projects and cost non-overrun projects. As a result, when cost overrun projects and cost non-overrun projects are balanced, linear discriminant analysis showed highest accuracy. Also, Mahalanobis-Tagichi method showed highest accuracy when ratio of cost overrun projects is smaller than 35%.

1. はじめに

近年，ソフトウェアは社会のインフラストラクチャとして，様々な場面で利用されており，ますます大規模化している．それに伴い，ソフトウェア開発プロジェクトの管理に失敗した場合のコスト超過額も大きくなっており，ソフトウェア開発企業の業績に大きく影響する危険性が高まっている（ここでコスト超過額とは，プロジェクトの予算金額に対して実績金額が超過した分を指す）．このため，プロジェクトのコスト超過（失敗）を防ぐことの重要性が増している．

プロジェクトのコスト超過を防ぐためには，コスト超過が発生しそうなプロジェクトを早期に発見し，対応策を実施する必要がある．コスト超過プロジェクトの発見には，線形判別分析，ロジスティック回帰，マハラノビス・タグチ法などの判別予測法が用いられる．判別予測法は，過去のプロジェクトで収集されたデータに基づき，予測モデルを構築し，この予測モデルに現行のプロジェクトで収集されたデータを入力することにより，予測を行う．

一般的な判別予測モデルでは，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が大きく異なることは想定されていない．しかし，ソフトウェアプロセスが改善されると，データセットに占める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が変化し，割合がアンバランスとなる可能性がある．ソフトウェア開発企業では，プロジェクトが失敗する（コスト超過，納期超過，品質低下などが発生する）可能性を低下させるために，プロセス改善活動が実施される場合がある．CMMI[2)]などに基づくプロセス改善により，ソフトウェアプロジェクトが失敗する可能性が低下すると，データセットに占めるコスト超過プロジェクトの割合も低下する．

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が大きく異なる場合，判別予測モデルの精度が低下する可能性がある．コスト非超過プロジェクトの割合が高くなると，モデルに対するコスト非超過プロジェクトの影響が大きくなる．その結果，モデルがコスト非超過プロジェクトの予測に過適合してしまうと，コスト超過プロジェクトがうまく予測できなくなる．予測方法によってモデル構築の方法が異なるため，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が予測精度に与える影響も異なると考えられる．

そこで本論文では，プロセス改善により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する．これにより，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応じた最適な判別予測方法が選択可能となることが期待される．実験では，あるソフトウェア開発企業で収集されたリスク評価データを用いて，コスト超過プロジェク

[†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

トとコスト非超過プロジェクトの割合を実験的に変化させ，各種の判別予測方法を適用し，精度を確かめる．また，本論文では，判別予測方法の 1 つであるマハラノビス・タグチ法を判別予測方法として採用する．マハラノビス・タグチ法は，モデル構築時にコスト超過プロジェクト（異常データ）を用いないため，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の変化（コスト超過プロジェクトの減少）に影響を受けない．

以降，2 章では各種判別予測方法について解説する．3 章では評価実験の方法と手順について説明する．4 章ではケーススタディの結果と，結果に対する考察を行う．5 章では関連研究を説明し，最後に 6 章でまとめと今後の課題について述べる．

2. 判別予測方法

本論文では，コスト超過プロジェクトの予測にあたり，判別予測方法を用いる．判別予測方法は，複数の説明変数を用いて，二値で表される目的変数を予測する方法である．予測モデルは，過去に蓄積されたデータを用いて構築される．以降，各判別予測方法について説明する．マハラノビス・タグチ法以外の予測方法は，判別予測方法として広く用いられているものである．

2.1 線形判別分析

線形判別分析は，直線によりデータを 2 つのグループに分けることにより，判別予測を行う方法である．線形判別分析で作成されるモデルは，以下のようになる．

$$y = a_1 x_1 + \cdots a_n x_x + b \tag{1}$$

ここで y は目的変数，x_n は説明変数，a_n は係数，b は切片を表す．y の値によって，各ケースが 2 つのグループのどちらに属するかがわかる．

2.2 ロジスティック回帰分析

ロジスティック回帰分析は，ロジスティック式に基づいた判別予測モデルを構築する方法である．ロジスティック回帰分析で作成されるモデルは，以下のようになる．

$$y = \frac{1}{1 + e^{-(a_1 x_1 + \cdots a_n x_x + b)}} \tag{2}$$

ここで y は目的変数，x_n は説明変数，a_n は係数，b は切片を表す．グループの判別予測の結果は，確率として得られる．例えば，予測結果が 0.7 の場合，一方のグループに属する確率が 70%となる．

2.3 分類木

分類木は，説明変数の値によって，分岐を繰り返すことにより目的変数を判別予測するモデルである．作成されるモデルは木構造となる．分類木の作成方法には，CHAID(Chi-squared automatic interaction detection)，CART(classification and regression trees)，C4.5 など，様々なものが提案されているが [9]，本論文では CART に基づいたアルゴリズムを採用した．

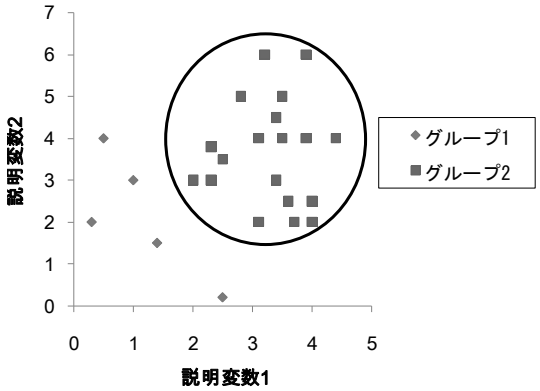

図 1 マハラノビス・タグチ法の例

2.4 マハラノビス・タグチ法

マハラノビス・タグチ法は田口玄一により考案された方法であり，主に製造業の品質管理のための一技法として用いられてきた．マハラノビス・タグチ法は正常なデータは類似しているが，異常なデータは個別性が高く，一定の傾向は見られないと考え，正常データのみを用いてモデルを構築する．マハラノビス・タグチ法では，正常なデータからのマハラノビス距離が大きいデータを，異常なデータと判定する．マハラノビス・タグチ法の例を図 1 に示す．図ではグループ 2 が正常データであり，正常データから離れたグループ 2 は異常データであると判別予測される．

3. 評価実験

3.1 実験の概要

実験では，コスト超過プロジェクトの割合が変化した場合の各予測手法の精度を比較し，コスト超過プロジェクトの割合に応じた最適な予測手法を明らかにする．実験では，あるソフトウェア開発企業で収集されたリスク評価データを用い，コスト超過プロジェクトの割合を実験的に変化させ，コスト超過プロジェクトの判別予測を行った．実験に用いた予測手法は，線形判別分析，ロジスティック回帰分析，分類木，マハラノビス・タグチ法である．これらの予測手法のうち，変数選択の手法が確立され

ている線形判別分析とロジスティック回帰分析については，変数選択を行った場合と行わない場合の両方で予測を行った．

3.2 実験に用いたデータセット

実験では，あるソフトウェア開発企業で収集されたリスク評価データを用いた．リスク評価データとは，プロジェクトの失敗（コスト超過，納期遅延，品質低下）につながるリスク（危険性）のある項目について，プロジェクトマネージャがそのリスクを評価したものである．リスク評価を行うことにより，リスクの高い項目，すなわち，重点的に管理すべき項目が明らかとなる．これらの高リスク項目を管理することにより，プロジェクトの失敗を未然に防ぐことを目的としたものである．リスク管理は，PMBOK[13]でも重要な技術の 1 つとして取り上げられている．

表 1 にリスク評価データの例を示す(リスク項目の例は，書籍 6)から引用している．実験に用いたデータセットに含まれるリスク項目ではない)．表の各行はリスク項目を示す．各リスクは，「リスク高」「リスク中」「リスク低」「対象外」の 4 段階により評価される．例えば，「S1:プロジェクト計画書が作成され，レビューされているか?」というリスク項目に対し,「プロジェクト計画書が作成されず，レビューもされていない」なら「リスク高」，「プロジェクト計画書が作成されたが，レビューされていない」なら「リスク中」，「プロジェクト計画書が作成され，かつレビューされた」なら「リスク低」となる．なお，例えば「外部委託先の管理は適切か」というリスク項目を評価する際に，そもそも開発の一部を外部委託していない場合，「対象外」となる．モデルを用いて予測をする際には，これらの評価を数値化する必要がある．評価は順序尺度であるため，実験では「リスク高」を 4，「リスク中」を 3，「リスク低」を 2，「対象外」を 1 とした．

リスク項目は，プロジェクトの上流（要求分析，設計）から下流（製造，試験）に関するものまで幅広く存在する [6) 7)]．下流に関するリスク項目は，下流の実施前に評価されることが多い．コスト超過プロジェクトの予測は，プロジェクトの初期（上流）に実施される．これは予測結果に基づき，早期に対応策を実施する必要があるためである．そこで実験では，プロジェクトの初期に不明である（下流に関する）リスク項目を除いた，プロエクトの初期（受注時点）で評価されるリスク項目のみを説明変数として用いた．

予測では，コスト超過を目的変数とした．コスト超過とは，コストの予定額と実績額を比較し，一定以上実績額が予定額を超過している場合を指す．これらの一定以上実績額が予定額を超過したプロジェクトを「コスト超過プロジェクト」，それ以外のプロジェクトを「コスト非超過プロジェクト」とし，コスト超過プロジェクトの判別予測を行った．

実験に用いたデータには，2000 年代に実施された 112 件のプロジェクトが含まれている．これらのプロジェクトのうち，29 件がコスト超過プロジェクト，83 件がコスト非超過プロジェクトであった．データにはリスク項目が 100 件以上記録されていたが，データ件数と比較して説明変数が多すぎる場合，予測モデルが適切に構築されないという問題がある．そこで，目的変数と相関が強いリスク項目のみを実験に用いた．具体的には，コスト非超過を 0，コスト超過を 1 として各リスク項目とコスト超過とのピアソンの積率相関係数を計算し，相関係数が 0.2 以上のリスク項目を説明変数の候補とした．

表 1　リスク評価データの例 [6)]

No.	リスク項目	評価	評価値
S1	プロジェクト計画書が作成され，レビューされているか?	リスク高	4
S2	プロジェクトの全体像を俯瞰できているか?	リスク中	3
S3	顧客のプロジェクト目的は明確か?	リスク低	2
S4	プロジェクト体制について，顧客も含めたステークホルダーの責任分担が明確になっており，体制上の不備不足や不安はないか?	対象外	1

さらに，これら相関係数が 0.2 以上のリスク項目から，欠損値が含まれないリスク項目を抽出した．欠損値とは，変数に値が記録されていないことを指す．欠損値が含まれる場合，何らかの方法（平均値挿入法などの欠損値除去法 [11)]）を用いて欠損値を除外する必要がある．実験では，欠損値の影響を除外するため，欠損値が含まれないように実験用データセットを作成した．相関係数が 0.2 以上，かつ欠損値が含まれないリスク項目を抽出した結果，説明変数となるリスク項目は 7 個となった．

3.3 実験手順

実験では，以下の手順によりデータセットに含まれる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を変化させ，予測を行った．実験結果の偏りをなくすために，下記手順を 10 回繰り返した（図 2）．

1. データセットからコスト非超過プロジェクト，超過プロジェクトをランダムにそれぞれ 14 件ずつ抽出し，ラーニングデータとする．
2. 同様にして作成したデータをテストデータとする．
3. ラーニングデータを用いて，判別予測モデルを構築する．
4. テストデータに構築したモデルを適用して判別予測を行い，予測精度を求める．
5. ラーニングデータから，コスト超過プロジェクトをランダムに 1 件除外する．

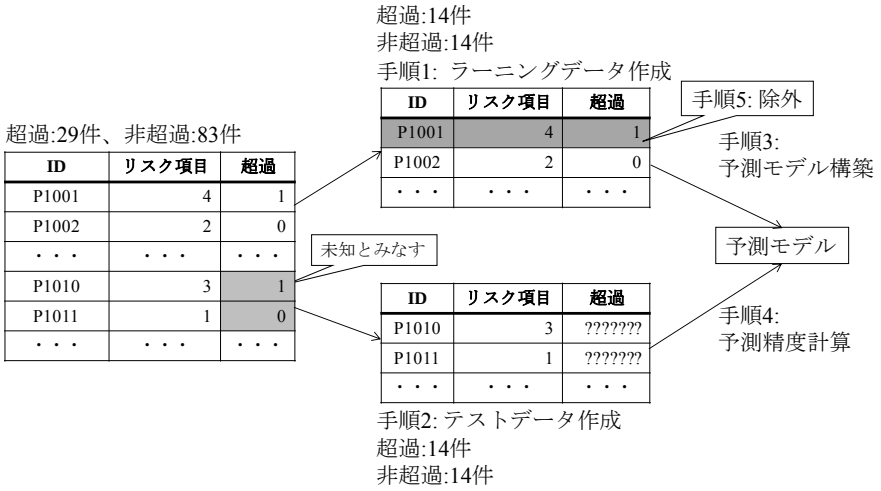

図 2 実験の手順

6. 手順 3〜5 をコスト超過プロジェクトがなくなるまで繰り返す.

手順 1 により，非超過プロジェクト，超過プロジェクトの割合を 50%にし，手順 5 により，コスト超過プロジェクトの割合を減少させる．テストデータについては，非超過プロジェクト，超過プロジェクトの割合は 50%とし，変化させない．なお，マハラノビス・タグチ法は，コスト非超過プロジェクトしかモデル構築に用いず，ラーニングデータのコスト超過プロジェクトの件数が減少しても，構築されるモデルは変化しないため，手順 3，手順 4 は 1 度だけ行う．

3.4 評価尺度

予測精度の評価尺度として，AUC(Area Under the Curve)[4]を用いた．AUC とは

表 2 *TP*，*FN*，*FP*，*TN* の定義

		実測結果	
		真	偽
予測結果	真	*TP*	*FP*
	偽	*FN*	*TN*

ROC(Receiver Operating Characteristic)曲線の下の面積であり，値域は[0, 1]である．AUC が 1 に近いほど予測精度が高いことを示す．ROC 曲線とは，閾値を変化させて陽性率と偽陽性率を計算し，線でつなげたものである．陽性率と偽陽性率は以下の式により計算する．*TP* (true positive)，*FN* (false negative)，*FP* (false positive)，*TN* (true negative) の定義を

表 2 に示す.

$$\text{陽性率} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{偽陽性率} = \frac{FP}{FP + TN} \tag{4}$$

4. 結果と考察

実験結果を図 3 に示す．図の横軸はコスト超過プロジェクト数，縦軸は AUC を示す．AUC は 10 回の実験結果から得られた値を平均したものである．9 件以上（64%以上）の場合，線形判別分析（変数選択あり，なし）の精度が最も高かった．8 件，7 件（50%，43%）の場合，線形判別分析（変数選択あり，なし）とマハラノビス・タグチ法の精度が同程度に高く，6 件（43%）では，線形判別分析（変数選択あり）とマハラノビス・タグチ法の精度が同程度に高かった．5 件以下（35%以下）の場合，マハラノビス・タグチ法の精度が最も高くなっていた．よって，データセットに含まれるコスト超過プロジェクトの割合が 35%以下の場合，マハラノビス・タグチ法が最も適した予測方法であり，コスト超過プロジェクトの割合が 35%よりも高い場合，線形判別分析（変数選択あり）が最も適した予測方法であるといえる．

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が同一の場合，線形判別分析（変数選択あり，なし）とロジスティック回帰分析（変数選択なし）の精度が高く，マハラノビス・タグチ法とロジスティック回帰分析（変数選択あり）はそれほど精度が高くなかった．また，分類木は最も精度が低くなっており，さらに，コスト超過プロジェクトが 5 件 (36%) の場合，予測モデルを構築することができなかった．分類木は,リスク評価データに基づくコスト超過予測には適していない可能性がある．

線形判別分析において，変数選択をした場合としない場合を比較すると，変数選択をした方が，コスト超過プロジェクトの割合によって精度の変化が大きいが，予測精度が高かった．ロジスティック回帰分析では，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が同じ場合，変数選択をした方が予測精度が高いが，コスト超

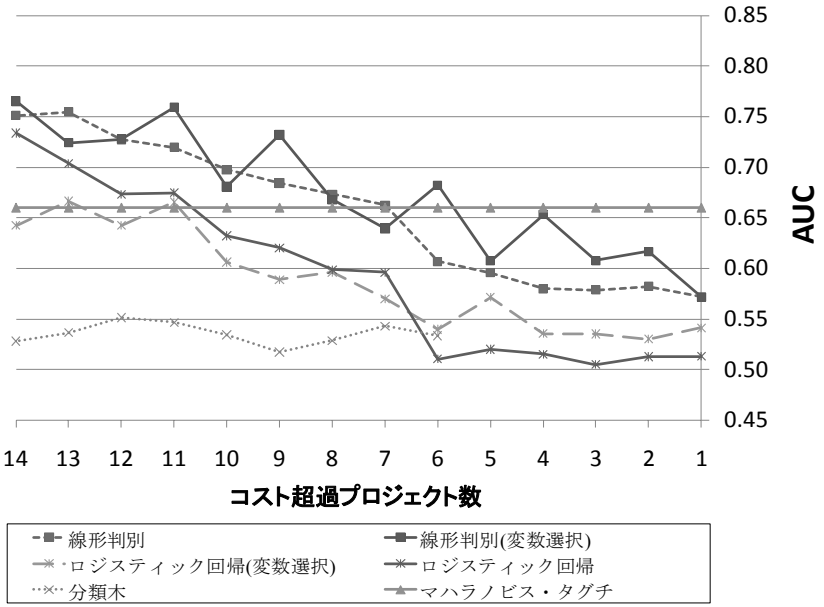

図 3 コスト超過プロジェクト数と予測精度の関係

過プロジェクトの割合が低くなると，変数選択をした方が予測精度が大きく低下していた．よって，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が変数選択に与える影響は，予測手法によって異なると考えられる．

マハラノビス・タグチ法が，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が同一の場合に，予測精度が他の手法に比べて低かった理由は，モデル構築時にコスト超過プロジェクトを使わない，すなわち他のモデルよりも情報量が少ない状態でモデルを構築しているためであると考えられる．逆に，コスト超過プロジェクトの割合が低い場合は，コスト超過プロジェクトが予測モデルに悪影響を与えるため，コスト超過プロジェクトを用いないマハラノビス・タグチ法の予測精度が相対的に高くなっていると考えられる．

5. 関連研究

これまで，リスク評価データなどの定性的なデータに基づき，コスト超過などのプロジェクトの結果を判別予測する研究が，いくつか行われている．本村ら [12]は，リスク評価データに協調フィルタリング法を適用し，コスト超過プロジェクトを判別予測する方法を提案している．Takagi ら [15]は，プロジェクトのリスク要因アンケートにロジスティック回帰分析を適用し，納期が遅延するプロジェクトを判別予測する方法を提案している．さらに工藤ら [10]は，リスク要因アンケートに対し，統計ソフト Weka [3]から利用できる 22 種類の判別予測法を適用し，手法間で精度の比較を行っている．ただし，これらの研究では，本論文のように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を変化させる実験は行っておらず，データに偏りがある場合の予測精度は明らかにしていない．

マハラノビス・タグチ法をソフトウェア工学分野で用いている研究がわずかながら存在する．Aman ら [1]は，マハラノビス・タグチ法を用いて，修正工数が掛かるモジュールの特定を行う方法を提案している．ただし，この研究でも，マハラノビス・タグチ法と他の手法間で，データ分布の偏りに対するロバスト性は比較していない．

マハラノビス・タグチ法とは逆に，コスト非超過プロジェクト（負例データ）を用いずに，コスト超過プロジェクトのみを用いる判別予測方法が存在する．畑ら [5]は，positive unlabeled learning アプローチの 1 つである Positive Naive Bayes 法（PNB）を用いてフォールトプローンモジュール（バグを含みやすいモジュール）を判別予測する方法を提案している．超過プロジェクトとコスト非超過プロジェクトの割合を変化させて PNB を適用し，予測精度を確かめることは今後の課題の一つである．

また，亀井ら [8]は偏りのあるデータに対し，オーバーサンプリング法を適用することを提案している．オーバーサンプリング法は，コスト超過プロジェクト，コスト非超過プロジェクトのどちらかデータ件数が少ないグループに対し，データの複製を行うことにより，グループのデータ件数をバランスさせる方法である．超過プロジェクトとコスト非超過プロジェクトの割合を変化させた後にオーバーサンプリング法を適用し，予測精度を比較することは，今後のもう一つの課題である．

6. おわりに

本論文では，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を実験的に変化させ，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よって，最適な判別予測方法が異なることを示した．ソフトウェア開発企業で収集されたリスク評価データを用いてコスト超過プロジェクトを判別予測する実験を行った結果，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，線形判別分析が適しているが，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・タグチ法が予測に最も適しているといえた．

今後の予定は，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偏り

がある場合にオーバーサンプリング法を適用し，マハラノビス・タグチ法とその他の判別予測方法の精度を比較することと，Positive Naive Bayes 法とその他の予測方法の精度を比較することである．

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．

参考文献

1) Aman, H., Mochiduki, N. and Yamada, H.: A Model for Detecting Cost-Prone Classes Based on Mahalanobis-Taguchi Method, IEICE transactions on information and systems, vol.E89-D, no.4, pp.1347-1358 (2006).
2) Chrissis, M., Konrad, M. and Shrum, S.: CMMI: Guidelines for Process Integration and Product Improvement, p.688, Addison-Wesley (2003).
3) Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.: The WEKA Data Mining Software: An Update, SIGKDD Explorations, vol.11, no.1 (2009).
4) Hanley, J. and McNeil, B.: The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology, no.143, pp.29-36 (1982).
5) 畑秀明，水野修，菊野亨：負例を用いない機械学習による fault-prone モジュール検出，ソフトウェアエンジニアリングシンポジウム 2009，pp.133-138 (2009).
6) 情報処理推進機構 ソフトウェア・エンジニアリング・センター：IT プロジェクトの「見える化」 上流工程編，p.208，日経 BP 社 (2007).
7) 情報処理推進機構 ソフトウェア・エンジニアリング・センター：IT プロジェクトの「見える化」 下流工程編，p.211，日経 BP 社 (2006).
8) 亀井靖高，柗本真佑，柿元健，門田暁人，松本健一：Fault-prone モジュール判別におけるサンプリング法適用の効果，情報処理学会論文誌，vol.48，no.8，pp.2651-2662 (2007).
9) 金明哲：R と樹木モデル(1)，ESTRELA, 5 月号, pp70-76, 統計情報研究開発センター (2005).
10) 工藤公太，水野修，菊野亨：複数の手法を用いたソフトウェア開発プロジェクトの混乱予測 ： 手法間での精度比較実験，電子情報通信学会技術報告，ソフトウェアサイエンス研究会，no.SS2004-36，pp.13-18 (2004).
11) Little, R. and Rubin, D.: Statistical Analysis with Missing Data, 2nd ed., p.408, John Wiley & Sons (2002).
12) 本村拓也，柿元健，角田雅照，大杉直樹，門田暁人，松本健一：協調フィルタリングを用いたプロジェクトコスト超過の予測，電子情報通信学会技術報告，ソフトウェアサイエンス研究会，no.SS2005-39，pp.35-40 (2005).
13) Project Management Institute：プロジェクトマネジメント 知識体系ガイド（PMBOK ガイド），p.484，Project Management Institute (2009).
14) R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, (2009). http://www.R-project.org.
15) Takagi, Y., Mizuno, O. and Kikuno, T.: An Empirical Approach to Characterizing Risky Software Projects Based on Logistic Regression Analysis, Empirical Software Engineering, vol.10, no.4 (2005).