B-007

ソフトウェア開発コスト見積における類似性に基づく欠損値処理の改良

Improvement of Similarity-Based Missing Data Techniques for Software Development Cost Estimation

渡辺　竜† Ryo Watanabe
柿元　健† Takeshi Kakimoto

1．諸言

ソフトウェア開発プロジェクトでは，開発初期段階においてコスト見積が行われる．開発初期段階においてコスト見積を正確に行うことが，プロジェクトを成功させるためには重要である．

開発プロジェクトのコスト見積を行う方法の一つとして，過去の開発プロジェクトデータに重回帰分析を適用する方法がある．しかし，過去の開発実績データには通常欠損値が含まれており，欠損値が含まれていると重回帰分析を適用することはできない．そこで，欠損値を含むデータに重回帰分析を適用するために，欠損値処理が行われている．

欠損値処理とは，欠損値の削除，あるいは，別の値による補完により，欠損値のないデータを作成する処理の総称である．欠損値処理の種類によって作成されるデータは異なる．従来の欠損値処理は削除もしくは補完のどちらかを行う手法であり，それぞれ，情報量の低下，誤差の含有を著しく招く場合がある．

そこで，本稿では，類似度に基づく補完法を改良し，欠損値の削除と補完を併用する手法であるハイブリッド手法を提案する．ハイブリッド手法は，欠損値が多く含まれるプロジェクトを削除した後に類似度に基づく補完法を適用することで，欠損値が多く含まれると適切な補完が難しい類似度に基づく補完法の欠点を改良した手法である．また，提案手法の性能と削除するプロジェクトの基準を確認するために行った評価実験についても報告する．

2．提案手法

2.1　類似度に基づく補完法

類似度に基づく補完法は，プロジェクト間の類似度にコサイン類似度を用いて，補完対象と類似しているプロジェクトの値を用いて補完値を算出する方法である [3]．

コサイン類似度はベクトルの内積を利用したものであり，以下の式で求めることができる．

$$\cos(\vec{P},\vec{Q}) = \frac{\vec{P}\cdot\vec{Q}}{|\vec{P}|\cdot|\vec{Q}|} = \frac{\sum_{i=0}^{n}\vec{p_i}\cdot\vec{q_i}}{\sqrt{\sum_{i=0}^{n}p_i^2}\cdot\sqrt{\sum_{i=0}^{n}q_i^2}} \qquad (1)$$

式(1)において，$\vec{P}$および$\vec{Q}$は，欠損値を含むプロジェクトと，そのプロジェクトに類似しているかどうか比較対象とするプロジェクトである．また，p_iおよびq_iはプロジェクト P と Q に含まれる欠損値でないメトリクスを正規化した値である．

次に，式(1)で求めた類似度から類似プロジェクトを選択し，補完値を類似度を重みとした重み付き平均値によって算出する．重み付き平均値は以下の式で求めることができる．

$$\widehat{X_n} = \frac{w_1x_1 + w_2x_2 + \cdots + w_ix_i}{w_1 + w_2 + \cdots w_i} \qquad (2)$$

式(2)において，$\widehat{X_n}$は n 番目のプロジェクトの欠損しているメトリクスに補完する重み付き平均値，i は類似プロジェクト数，$w_1, w_2, \cdots, w_i$はプロジェクトごとの類似度，$x_1, x_2, \cdots, x_i$は類似プロジェクトごとのメトリクスの値を表している．

2.2　ハイブリッド法

類似度に基づく手法には，欠損率が 60%以上のデータでは精度が低下してしまうという問題点がある [2]．そこで，類似度に基づく手法の問題点を改善するために，まず欠損値を多く含むプロジェクトを削除し適用データの欠損率を低くする．そして，削除後のデータに対して類似度に基づく補完法を適用することで，類似度に基づく補完法の精度を低下させないようにする．

したがって，ハイブリッド法の手順は以下のようになる．

① プロジェクトに含まれるメトリクス数のうち削除基準以上のメトリクスが欠損値であるプロジェクトを削除する．
② 削除後のデータに含まれる欠損値について，コサイン類似度を用いて類似度を算出し，類似プロジェクトを選択する．
③ 選択した類似プロジェクトより，重み付き平均値を算出し欠損値部分の補完値とする．
④ 全欠損値に対して，②~③を行う．

3．評価実験

提案手法であるハイブリッド法では，プロジェクトを削除する基準を変化させることで類似度に基づく補完法を適用するデータが変化し，コスト見積精度に影響を与える．したがって，本稿では，削除過程を加えることによる類似度に基づく補完法の見積精度の改善，および，削除基準の違いによる見積精度の変化を確認するために評価実験を行った．

評価実験に用いるデータセットには，見積手法の評価に広く利用されている ISBSG が収集したデータ [1]を基に作成したデータセットを用いた．データセットの作成において，ISBSG データセットから，目的変数である Summary Work Effort が欠損であるプロジェクト，ファンクションポイントのカウント手法が LOC と記録されているプロジェクト，設計工程以前には得ることができないメトリクス，および，欠損率が極端に多いメトリクスを削除した．その結果，使用データセットに含まれるプロジェクトは 1857 件，メト

† 国立高等専門学校機構 香川高等専門学校，National Institute of Technology, Kagawa College

表 1　使用データセットに含まれるメトリクス

	メトリクス名	欠損率
説明変数	Adjusted Function Point	0.5%
	Project Elapsed Time	10.6%
	Effort Plan	83.3%
	Effort Specify	70.0%
	Effort Design	84.3%
	Average Team Size	86.4%
	Input count	56.9%
	Output count	60.5%
	Enquiry count	61.3%
	File count	60.5%
	Interface count	61.3%
目的変数	Summary Work Effort	0%

リクスは表 1 に示す 12 個となった．このデータセットに対してハイブリッド法を適用した．なお，類似プロジェクト数は削除後のプロジェクト数の 10%とした．

評価実験では精度評価に多く用いられる Leave-one-out 法を用いた．以下に評価実験の手順を示す．

① データに含まれる全プロジェクト n 件から，見積対象とする 1 件のプロジェクトを抽出する．
② 抽出した 1 件のプロジェクトを，開発を行う現行プロジェクトと見立て，残った n-1 件のプロジェクトから線形重回帰分析によるコスト見積を行う．
③ コストの見積値と実測値から誤差を求める．
④ 抽出するプロジェクトを変更し，①~③の手順を全プロジェクト n 件に対して行う．
⑤ 求めた誤差から，ハイブリッド法の削除基準の違いによる見積精度の比較を行う．

評価実験における評価指標には MMRE を用いた． MRE は一般的な相対誤差で実測値に対する見積値の誤差である．そして，MMRE は MRE の平均値である．MRE は以下の式で求めることができる．

$$\mathrm{MRE} = \frac{|X_i - \hat{X}_i|}{X_i} \qquad (3)$$

式(3)において，X_iは実測値，$\hat{X}_i$は見積値，$i = 0,1,2,\cdots n$である． MMRE は値が小さいほど見積精度が高いことを示す．

4. 実験結果と考察

図 1 は，各削除基準の見積精度の値をプロットしたグラフである．横軸は各削除基準で作成されたデータの欠損率としている．各削除基準においてプロジェクトを削除した後のプロジェクト数と欠損率は表 2 の様になった．

欠損率 36.2%（削除基準 60%）で見積精度は最も小さく，削除しない場合と比較して 0.56 向上している．また，欠損率 12.2%（削除基準 20%），および，欠損率 16.8%（削除基準 30%）では見積精度が著しく低下している．削除基準 20%，30%で見積精度が著しく低下した原因として，プロジェクトを大量に削除したために類似しているプロジェクトが少なくなり，補完値に大きな誤差が生じたこと考えられる．

図 1 から見て取れるように，欠損基準が高すぎても低すぎて見積精度は低くなった．欠損率とプロジェクト数はトレードオフの関係であり，適切な欠損基準を選択すること

表 2　プロジェクト削除後のデータ

削除基準	プロジェクト数	欠損率[%]
20%	144	12.2
30%	244	16.8
40%	701	28.0
50%	757	29.1
60%	1016	36.2
70%	1241	41.7
80%	1808	52.2
削除なし	1857	53.0

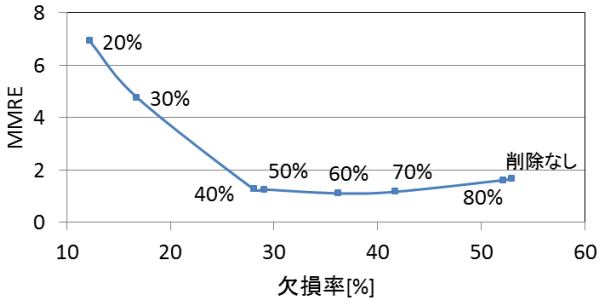

図 1　各削除欠損率の見積精度

で高い見積精度が得られると言える．本稿の実験では 1 つのデータセットしか扱っていないため，他のデータセットでも評価実験を行うことで，削除基準を決定する方法について検討していく必要がある．

5. まとめ

本稿では，類似度に基づく補完法を改良し，欠損値処理の削除と補完を併用する手法であるハイブリッド法を提案した．評価実験の結果，類似度に基づく補完法をそのまま適用した結果よりも，削除基準を 40%以上とした提案手法を適用した結果の方が高い見積精度が得られた．しかし，過度のデータ削除は，見積において十分なプロジェクト数とならず，見積精度が低下する原因となるため，類似度に基づく補完法の改良としては削除基準の調整が重要である．

また，今回評価実験に使用したデータセットでは見積精度の改善がみられたが，ハイブリッド法の汎用性を確認するために，他のデータセットでの評価実験も行う必要がある．その他にも，重み付き平均値以外の補完値算出方法や類似プロジェクトの数の違いによる見積精度の影響を考慮することも必要である．

参考文献

1) ISBSG Estimating, “Benchmarking and Research Suite Release 11 : International Software Benchmarking Standards Group,” http://www.isbsg.org/, 2009.
2) 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に基づく工数見積もり手法のデータの欠損に対するロバスト性の評価，電子情報通信学会論文誌 D, Vol.J89-D，No.12，pp.2602-2611，2006.
3) 田村晃一，柿元健，戸田航史，角田雅照，門田暁人，松本健一，大杉直樹：工数予測における類似性に基づく欠損値補完法の実験的評価，コンピュータソフトウェア，Vol.26，No.3，pp.44-55，2007.