工数予測のための欠損値補完手法について

戸田 航史[1,a)]

概要：本稿では工数見積もりモデル構築に用いられるデータセットのための欠損値補完手法について検討する．既存手法としては類似性に基づく手法が提案されているが，近年，MissForest 法および多重代入法が新たに提案されている．本研究では MissForest 法および多重代入法の実装の 1 つである MICE(Multivariate inputation by chained equations) について検討する．

Discussion about Missing Value Imputation Methods for Effort Estimation

1. はじめに

ソフトウェア開発プロジェクトの初期段階において開発工数を予測すること，及び各工程の終了段階で開発工数の再予測を行うことは，プロジェクト完遂に必要な資源の確保や，スケジュール管理を行う上で重要である．そのために，過去のソフトウェア開発プロジェクトの実績データ (以下，プロジェクトデータ) を予測の根拠に用いる定量的予測手法が数多く提案され，用いられてきた [1][8]．定量的予測手法では，通常プロジェクトのメトリクス (開発規模，欠陥数など) を説明変数として用い，目的変数である開発工数との関係を数式等で表現する．

ただし多くの定量的予測手法は，その適用にあたって欠損値を含まないデータセットが必要となるが，一般に多数の部署・組織から収集されたプロジェクトデータには欠損値が含まれる [4][10]．欠損値が生じる原因としては，収集メトリクスが異なる複数の組織のデータをマージしたことや，時間的制約や不注意による記録漏れなどが挙げられる．このため，多くの場合定量的予測手法の適用にあたっては，何らかの手法で欠損を補完し，欠損の無いデータを作成することが必要となる．1 つの方法として，欠損値を含むメトリクスやプロジェクトを除去し，欠損値のないデータセットを作成する無欠損データ作成法がある [5]．ただし，無欠損データ作成法はデータセットのサイズが小さくなり，予測の根拠となる情報量を減らしてしまうため，モデルを構築したとしても十分な予測精度が得られない可能性がある．別の方法として，欠損値を何らかの値で補完 (欠損値補完法) の実施により，欠損値が無いままにデータセットのサイズを保つことである [3]．ただし，欠損値補完法によって適切な値を補完できない場合，それはデータセットにとってのノイズとなり，妥当なモデルが得られなくなる可能性がある．

本稿では近年になり提案された欠損値補完手法である MissForest 法 [9]，および多重代入法 [7] の実装アルゴリズムである MICE(Multivariate inputation by chained equations)[11] について紹介し，工数予測における欠損を含むデータに対するアプローチについてそれを踏まえた議論をしたい．

2. 欠損値補完手法

2.1 MissForest 法

MissForest 法は Random Forest 法 [2] を欠損値補完に適用した手法である．手順の概要を以下に示す．

(1) 欠損を含むメトリクスを 1 つ選ぶ
(2) 選んだメトリクスが欠損しているプロジェクトと欠損していないプロジェクトにデータセットを分割する
(3) 欠損していないプロジェクトを用いて，選んだメトリクスの予測モデルを Random Forest を用いて構築する
(4) 構築したモデルを用いて欠損値を予測し，補完する
(5) 1. から 4. までを全てのメトリクスについて実行する
(6) 所定の回数，もしくは欠損値が更新されなくなるまで 1. から 5. を繰り返す

[1] 福岡工業大学
Fukuoka Institute of Technology
[a)] toda@fit.ac.jp

初期状態では全ての欠損に対して平均値挿入法が実行され，上記の手順に従い，順次 RandomForest による予測値（補完値）により更新される．

2.2 多重代入法

多重代入法はマルコフ連鎖モンテカルロ法に基づく欠損値補完手法である．手順を以下に示す．

(1) 不完全な（欠損のある）データセットを n 個複製する
(2) 複製したデータセットに対し補定を実施し，n 個の補定済みデータセットを得る，
(3) 補定済みデータセットに対し分析を実施する
(4) 分析結果を統合し，各欠損値に対して点推定値（補完値）を得る

このように多重代入法は欠損値補完のためのおおまかな手順（枠組み）を提供してはいるものの，その具体的なアルゴリズムを規定していない．このため，多重代入法のアルゴリズムとして，大きく分けて 3 つの手法（Markov Chain Monte Carlo(MCMC), Fully Conditional Specification(FCS)，Expectation-maximization with Bootstrapping(EMB)）が提案されている．ここでは FCS の実装アルゴリズムである Multivariate inputation by chained equations (MICE) について，その概要を述べる．

(1) 不完全な（欠損のある）データセットを n 個複製する
(2) 各欠損値についてそのメトリクスの実績値からランダム取り出し，仮の補完値とする
(3) メトリクスを 1 つ選び，そのメトリクスが欠損していた（仮の補完値が充填された）プロジェクトと欠損していないプロジェクトに分割する
(4) 欠損していないプロジェクトを用いて何らかの予測手法を用いて各欠損値に対する補完値を算出する
(5) 2. から 4. までの手順を全てのメトリクスについて実行する
(6) 複製した n 個のデータセットに対して 3. から 5. までの手順を一定回数繰り返し，各欠損について n 個のデータの平均値を最終的な補完値とする

MICE を用いた実際の補完に当たっては，データセットの複製数，予測手法，繰り返しはハイパーパラメータとして与えられる．

3. 議論

多重代入法の枠組み自体の提案は 1987 年と新しいものではないが，上で例として示した MICE の提案は 2012 年とごく最近である．これは多重代入法の計算量の問題と考えられる．すなわち，ハイパーパラメータとして複製数を n，繰り返しを t として，データセット中の欠損を含むメトリクスが m 個であれば $n \times t \times m$ 回，予測手法を実施する必要がある．複製数については 100-1000 程度を推奨する文献 [6] もあり，例えば繰り返し回数，欠損を含むメトリクスが共に 5 であったとしても，予測手法は 2500-25000 回程度実行される事になり，予測手法の計算量が少なかったとしても，実行回数によって総計算量は非常に大きなものとなる．この計算量の問題が計算機の性能向上により解決され，多重代入法の新たなアルゴリズムが提案されるようになったものと考えられる．同様に MissForest 法についても，RandoForest 法自体が計算量が多い手法であり，これを複数のメトリクスに対し複数回行えば計算量が膨らむため，近年まで手法が提案されなかったと考えられる（RandomForest 法が 2001 年に対し MissForest 法が 2012 年）．

ワークショップでは新たに提案された欠損値補完手法の動向，および既存の類似度を用いた欠損値補完手法や欠損を含むプロジェクト，メトリクスを削除する方法を踏まえた上で工数予測における欠損値に対するアプローチと欠損値補完の必要性について議論できればと考えている．

参考文献

[1] Boehm, B.W.: Software engineering economics, Prentice Hall, New Jersey (1981).
[2] Breiman, L.: Random Forests, Machine Learning Vol.45 No.1(2001), pp.532.
[3] Jonsson, P. and Wohlin, C.: An evaluation of k-nearest neighbour imputation using likert data, Proc 10th IEEE International Softw. Metrics Symposium (Metrics'04), Chicago, Illinois (2004), pp.108-118.
[4] Kromrey, J. and Hines, C.: Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments, Educational and Psychological Measurement, Vol.54, No.3 (1994), pp.573-593.
[5] Myrtveit, I., Stensrud, E. and Olsson, U. H.: Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods, IEEE Trans. Softw. Eng., Vol.27, No.11 (2001), pp.999-1013.
[6] Royston, P and White, I.: Multiple Imputation by Chained Equations (MICE): Implementation in Stata, Journal of Statistical Software, Vol. 45, No. 1, pp.1-20 (2011).
[7] Rubin, D. B.: Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons (1987).
[8] Shepperd, M. and Schofield, C.: Estimating software project effort using analogies, IEEE Trans. Softw. Eng., Vol.23, No.12 (1997), pp.736-743.
[9] Stekhoven, D.J. and Buehlmann, P.: MissForest - nonparametric missing value imputation for mixed-type data, Bioinformatics, Vol.28, No.1 (2012), pp.112-118.
[10] 角田雅照, 大杉直樹, 門田暁人, 松本健一, 佐藤慎一: 協調フィルタリングを用いたソフトウェア開発工数予測方法, 情報処理学会論文誌, Vol.46, No.5 (2005), pp.1156-1164.
[11] Buuren. S.,: Flexible Imputation of Missing Data, London: Chapman & Hall/CRC (2012).