様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ５ 月 ３０ 日現在

機関番号：１５３０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００１０２

研究課題名（和文）ソフトウェア開発に関する機密データからの研究用データの生成

研究課題名（英文）Generation of mimic data sets from confidential software project data sets

研究代表者
門田　暁人（Monden, Akito）

岡山大学・自然科学研究科・教授

研究者番号：８０３１１７８６

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では，ソフトウェア開発実績データを対象として，データそのものではなく，データの特徴量（データの分布，データ間の関係など数値化したもの）のみをデータ収集企業から受け取り，特徴量の類似する研究用データを人工的に生成する方法を開発した．得られる生成データは，様々なデータ予測手法や分析手法に適用可能であり，特にデータ件数が多い場合には，元のデータと類似する予測性能や分析結果が得られることが分かった．

研究成果の学術的意義や社会的意義
本研究の成果により，データの機密性を保持したまま，その利活用が可能となることが実証された．提案方法により生成されたデータを広く公開することで，予測モデルの性能評価，ベンチマーキング，データ処理技術の評価等に活用でき，実証的ソフトウェア工学の研究分野の発展に寄与できると期待される．また，提案方法は，機密データを利用する様々な分野への応用が期待される．

研究成果の概要（英文）：This study proposes a method for artificially generating a mimic software project data set, whose characteristics (related to data distribution and data dependencies) are very similar to a given confidential data set. In this method, companies need to provide only several statistical values of their confidential data. Our experimental evaluation confirmed that the generated mimic data sets can be applied to various data prediction/analysis methods, and, in case that the original data contain enough data points, we could expect obtaining similar prediction/analysis results as the original data set.

研究分野：ソフトウェア工学

キーワード：データ機密保護　ソフトウェア開発実績データ　ソフトウェア開発工数予測　データ分析

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

実証的ソフトウェア工学の研究分野では，多数のソフトウェア開発プロジェクトの実績データが必要となる．ところが，一般公開されているデータは古いものが多く，研究の妥当性や信頼性を確保する上で大きな問題となっている．例えば，研究者向けに公開されている Maxwell, Desharnais，COCOMO'81，Kemerer，Albrecht などのソフトウェア開発プロジェクトデータセットは，1980～1990 年代のものであり，IEEE Trans. Soft. Eng.等のトップジャーナルにおいても，いまだにこれらの古いデータがベンチマーク用に使われ続けているのが現状である．これらのデータセットを用いた研究成果の妥当性や信頼性に疑問が残る．

一方，多くの企業では，最新のソフトウェア開発のデータを計測・蓄積しているが，昨今では，個人情報保護，及び，コンプライアンス重視のために，企業における機密保持がより厳格となり，大学の研究者が機密データを使った研究を行うことはますます困難となっている．

このような現状から，近年，データ・ミューテーションによりデータの匿名化を図るアプローチが研究されている．このアプローチでは，非公開データセット中の個々の値について，データセットの性質を大きく変えない範囲で値を増減させる．これにより，変換後のデータセットから，変換前のデータセットの各個体を特定することを難しくしている．ただし，データの匿名性を高めると，データから得られる分析結果やモデルに大きな影響を与えてしまう．逆に，匿名性が低いと，データ公開の理解が得られない．また，そもそもデータ匿名化は，必ずしもデータ公開に繋がるとは限らない．個々の値を変化させたとしても，データが外部に出ることには変わりがなく，利害関係者間の調整等，データ公開の敷居は依然として高いと考えられる．

２．研究の目的

本研究では，ソフトウェア開発実績データを対象として，データそのものではなく，データの特徴量（データの分布，データ間の関係など数値化したもの）のみをデータ収集企業から受け取り，特徴量の類似する研究用データを人工的に生成する方法を開発する．本アプローチでは，データ・ミューテーションと異なり，データの匿名性は原理的に確保されることになる．また，データの特徴量のみの提供であれば，企業から理解が得られやすいと期待される．一般に，ソフトウェア開発実績データセットは様々な尺度・分布を持った変数を含み，それらは外れ値や欠損値を含むことが考えられる．また，変数間には，様々な関係が存在することが考えられる．そこで，本研究では，様々な変数の分布や尺度，欠損値などを再現する方法と，変数間の関係を再現する方法を開発する．そして，多数のデータセット，データ分析方法を用いた評価を行う．

３．研究の方法

(1) データ生成方法の開発

申請者らの先行研究では，データセット中の各変数は，対数正規分布に従うことを仮定していたが，全ての変数が対数正規分布に近似できるわけではない．また，名義尺度，順序尺度といった，他の尺度についてもデータ生成方法が必要となる．そこで，様々な分布に対応したデータ生成方法を開発するとともに，名義尺度，順序尺度に対するデータ生成方法を開発した．

ソフトウェア開発に関するデータには，外れ値や欠損値が含まれることが多く，その取り扱いが重要な研究テーマとなっている．そこで，本研究では，外れ値，欠損値の再現方法を開発した．また，本研究では，多数の変数間の関係の定量化し，それを再現する方法についても開発した．

(2) データ生成方法の評価

本研究では，多数のソフトウェア開発実績データを対象としたデータ生成実験を行った．生成したデータの有用性を評価するために，まず，ソフトウェア開発現場でよく用いられている log-log 重回帰分析を用いたソフトウェア開発工数予測の性能評価を行った．また，ニューラルネット，ランダムフォレストを始めとする多様なモデリング手法を用いた評価を行った．さらには，欠損値の存在を前提としたデータマイニング手法であるアソシエーションルールマイニングについても適用実験を行った．

４．研究成果

(1) データ生成方法の開発

現実のソフトウェア開発データを分析した結果，多くの量的変数（比尺度）は対数正規分布に近い分布となっているが，歪みを含んでいることが明らかとなった．そこで，まず，ボックス＝ミュラー法により正規分布を生成し，対数変換を行うことで対数正規分布に従うデータを生成することとした．さらに，データの特徴量として，尖度と歪度を計測することとし，生成されたデータに対して sinh-arcsinh transformation による尖度と歪度の変換を施すことで，現実のデータにより近づける方法を提案した．実際のソフトウェア開発データに対して sinh-arcsinh transformation を用いた結果，与えられた尖度と歪度に一致するデータの生成が可能なことを確認した．図１に歪度と尖度の変換例を示す．また，外れ値を再現する方法として，元データの対数変換後のデータにおいて標準偏差σとした場合に，±2σ，および，±3σを超える値の含有率を特徴量として用い，生成データにおいて特徴量が一致するように外れ値を付与することとした．名義尺度と順序尺度については，取りうる値（カテゴリ）の割合が等しくなるように人工的な値を生成することとした．また，欠損値はカテゴリの一つとして与えることで再現した．

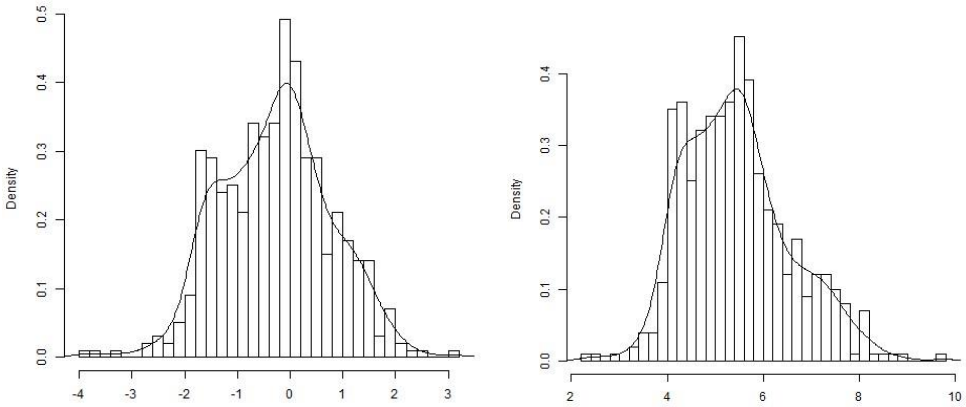

図 1. 歪度と尖度の変換前の分布（左）と変換後の分布（右）の例

Algorithm 1: Computation of correlation matrix based on Spearman's rank correlation coefficient.

Input: Values of project variables $\overrightarrow{V}_i, \forall i \in [1, T]$
Output: Correlation matrix χ

```
for i ← 1 : T do
    Set F to size of V_i
    Set S_i to ranked array of V_i
    for i ← 1 : F do
        for j ← 1 : F do
            χ(i,j) = cov(S_i,S_j) / (σ_Si σ_Sj)
```

Algorithm 2: Mimicking pairwise relations.

Input: Values of mimic variables $\overrightarrow{V'}_i \forall i \in [1, T]$
Correlation matrix of source data χ // See Alg. 1
Output: Reordered values of mimic variables $\overrightarrow{V'}_i, \forall i \in [1, T]$

```
Set S' to size of any array V'_i
Set ε_0 = ∞                          // Previous value of ε
do
    /* Reorder by swapping arbitrary values */
    Get a pair of arbitrary indices 1 < p, q < S', i ≠ j
    V'_i(p) ↔ V'_i(q)                                  // Swap
    Get χ' relating reordered mimic data      // See Alg. 1
    /* Similarity of correlation matrices ε is
       expressed in terms of sum of squared
       differences.  */
    Set ε = Σ_{i,j=1}^{T} (χ(i,j) − χ'(i,j))^2
    if ε < ε_0 then                  // There is improvement

        ε_0 = ε
    else                          // There is no improvement

        V'_i(p) ↔ V'_i(q)                          // Swap back
while ε converging
```

図 2. 変数間の関係を再現するアルゴリズム

変数間の関係の再現方法として, 3 変数以上の組み合わせを扱う方法を提案した. 具体的には, 相関の大きい 2 変数の組み合わせについて，一方を他方で除した導出尺度を設け，導出尺度を含むすべての 2 変数間の関係を再現することとした．提案アルゴリズムを図 2 に示す．まず，導出尺度を含む全ての変数に対して Algorithm 1 により順位相関係数行列を導出し，Algorithm 2 により元データと生成データの間で順位相関係数行列が類似するように生成データの入れ替えを行う．

(2) データ生成方法の評価

まず，歪度と尖度の変換がソフトウェア開発工数予測モデルに与える影響について，モデルの適合度と予測性能の両面から評価を行った．その結果，データセットによっては歪度と尖度の変

表 1. 評価実験で用いたソフトウェア開発実績データセット

Dataset	Number of categorical variables	Number of continuous variables	Number of projects
Albrecht	0	7	24
China	1	11	499
Coc81dem	14	4	63
Desharnais	4	5	77
Kemerer	2	5	15
Maxwell	22	4	62
Miyazaki94	0	8	48
Nasa93	24	2	93

表 2. ソフトウェア開発工数の予測結果

Dataset	Δm(AE)	Δm(MRE)	Δm(MER)	Δm(BRE)
Albrecht	0.63	0.02	0.30	0.01
China	0.02	0.01	0.03	0.01
Coc81-dem	0.10	0.05	0.17	0.08
Desharnais	0.02	0.05	0.02	0.02
Kemerer	0.33	0.53	0.44	0.18
Maxwell	0.04	0.03	0.08	0.02
miyazaki94	0.04	0.11	0.09	0.07
nasa93	0.04	0.02	0.30	0.08

換がモデルの適合度および予測性能の影響を与えることが分かり，変換の必要性を確認できた．次に，ソフトウェア工学分野でよく用いられている 8 つのソフトウェア開発実績データセット（表 1）を用いて，log-log 重回帰分析によるソフトウェア開発工数予測を対象とし，元データと生成データからそれぞれ得られる工数予測モデルの予測性能を比較する実験を行った．実験では，3-fold cross validation を 10 回繰り返した．表 2 に実験結果を示す．表中，AE は予測結果の絶対誤差, MRE は相対誤差, MER は予測値を分母とする相対誤差, BRE は balanced relative error を示している．Δ m は元データにおける予測結果と生成データにおける予測結果の相対差を示している．表 2 より，多くの場合において，元データと生成データは，似た性能を持つソフトウェア開発工数予測モデルを生成できていることが分かる．ただし，一部の結果（Kemerer と Albrecht の Δ m(AE)など）は高い値となっている．これは，表 1 に示されるように，Albrecht は 24 件，Kemerer は 15 件とデータ件数が極めて少ないことが原因であると考えらえる．ある程度の規模（例えば 50 件以上）のデータセットであれば，元データとそん色のない予測モデルが得られると期待される．

また，生成データに対し，ニューラルネット，ランダムフォレスト等の機械学習モデルを用いた評価を行った．その結果，log-log 重回帰モデルと同様，データ件数が多い場合には，元のデータをそん色のない予測性能を持つモデルが得られることを確認した．また，アソシエーションルールマイニングの適用実験を行った．その結果，やはりデータ件数が多い場合には，元のデータと類似するルールが得られることを確認した．

以上のことから，提案方法により生成されるデータは，人工的に作られたデータではあるものの，その特徴が元となる機密データと極めて似ていることから，様々な予測手法やデータ分析手法において，同じような性能のモデルや分析結果が得られることが分かった．本研究により，データの機密性を完全に保持したまま，その利活用が可能となることが実証され，ソフトウェア工学分野のみならず，機密データを利用する様々な分野の発展に寄与することが見込まれる．

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　7件／うち国際共著　2件／うちオープンアクセス　0件）

１．著者名 Maohua Gan, Zeynep Yucel, Akito Monden, Kentaro Sasaki	４．巻 -
２．論文標題 Empirical evaluation of mimic software project data sets for software effort estimation	５．発行年 2020年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Maohua Gan, Kentaro Sasaki, Akito Monden, Zeynep Yucel	４．巻 -
２．論文標題 Generation of mimic software project data sets for software engineering research	５．発行年 2018年
３．雑誌名 Proceedings of the 6th International Workshop on Quantitative Approaches to Software Quality	６．最初と最後の頁 38-43
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Seiji Fukui, Akito Monden, Zeynep Yucel	４．巻 -
２．論文標題 Kurtosis and skewness adjustment for software effort estimation	５．発行年 2018年
３．雑誌名 Proceedings of the 25th Asia-Pacific Software Engineering Conference	６．最初と最後の頁 504-511
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 福居誠二，門田暁人	４．巻 -
２．論文標題 尖度と歪度を考慮した予測モデルの検討	５．発行年 2018年
３．雑誌名 ウィンターワークショップ2018・イン・宮島 論文集	６．最初と最後の頁 32-33
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 齊藤 英和，門田 暁人	4．巻 -
2．論文標題 機密を保持したままソフトウェア開発データの分析を行う方法についての一考察	5．発行年 2017年
3．雑誌名 ソフトウェア工学の基礎XXIV	6．最初と最後の頁 111-116
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考