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　６ 年　 ６ 月 １５ 日現在

機関番号：１５３０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ１１７４９

研究課題名（和文）ソフトウェアバグ予測を題材とする機械学習システムの評価技術の開発

研究課題名（英文）Development of evaluation techniques for machine learning systems for software bug prediction

研究代表者
門田　暁人（Monden, Akito）

岡山大学・環境生命自然科学学域・教授

研究者番号：８０３１１７８６

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：機械学習システムの評価においては，(1)学習データの品質の評価，および，(2)システム出力の性能評価が重要となる．(1)については，データ矛盾性の尺度Similar Case Inconsistency Level (SCIL)を定義した．評価実験によって，矛盾の少ないデータセットほど，得られる機械学習モデルの予測性能が高い傾向にあることを示した．(2)については，2クラス分類問題における性能評価指標の期待値を，データセットのneg/pos ratioに基づいて定義した．適用実験によって，従来の評価指標では予測性能を正しく評価できないケースがあることが分かり，提案尺度の有用性が示された

研究成果の学術的意義や社会的意義
本研究の成果によって，ソフトウェア開発データを対象とした機械学習システムにおいて，学習データを事前に評価すること，および，性能評価をより適切に行うことが可能となり，ソフトウェア工学分野のさらなる発展に寄与できると期待される．また，提案方法は，機械学習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）：In the evaluation of machine learning systems, it is important to (1) evaluate the quality of training data and (2) evaluate the performance of system output. For (1), we defined a data inconsistency measure, Similar Case Inconsistency Level (SCIL). Through evaluation experiments, we showed that the less inconsistent the dataset is, the better the prediction performance of the resulting machine learning model tends to be. For (2), we defined the expected values of performance measures for a two-class classification problem based on the neg/pos ratio of the dataset. Application experiments showed that there are cases in which conventional evaluation measures cannot correctly evaluate the prediction performance, indicating the usefulness of the proposed measures.

研究分野：ソフトウェア工学

キーワード：ソフトウェア開発データ　ソフトウェアバグ予測　ソフトウェアメトリクス　機械学習

１．研究開始当初の背景

今日，情報システムの開発は，「論理回路を組む」「プログラムを書く」といった従来型の方法に加えて，「データを例示する」という機械学習による新しい方法が用いられつつある．そのような機械学習システムの動作は，大量のデータから学習することで決定されるため，従来のソフトウェア開発方法論が必ずしも利用できず，システムの性能を担保するための新たな技術が必要となる．今日，AI を支える機械学習技術を内部に組み込んだ情報システム（以降，機械学習システム）が社会に普及しつつあり，その評価技術の確立が急務となっている．

機械学習システムの評価においては，機械学習の元となる(1)学習データの品質の評価，および，(2)多様な入力に対するシステム出力の性能評価が重要となる．(1)については，データ品質に影響する要素として，外れ値，ノイズ，完全性，矛盾性，冗長性などが知られているが，それらについてのコンセンサスのある定義はなく，機械学習の性能との関係も必ずしも明確でない．また，(2)については，機械学習システムの様々な（予測）性能指標が提案されてきたが，システムの適用対象となる入力データ群が違えば得られる性能値も異なることになる．例えば，特定のデータ群に対して高い予測精度が得られたとしても，別のデータ群に対しては低い予測精度しか得られないことも少なくない．そのため，入力データに依存しない評価方法が必要である．

２．研究の目的

本研究では，ソフトウェア開発に関するデータを題材として，上記(1)と(2)の解決を目指す．(1)については，学習データに偏りや矛盾が含まれる場合，そのデータに基づいて作成されたシステムは，期待通りの性能を発揮できなかったり，予期しない動作を行う可能性がある．本研究では，従来注目されてこなかった，学習データに含まれる「矛盾」に着目した学習データの評価方法の確立を目的とする．ここでいう矛盾とは，データセットを構成する複数の個体が，互いに矛盾する値を保持していることを指す．また，(2)については，ソフトウェアバグ予測においてよく使われる評価指標がデータの neg/pos ratio（ここでいう pos とはバグを含む個体の数であり，neg とはバグを含まない個体の数である）に影響される点に着目し， neg/pos ratio に影響されない評価方法の確立を目的とする．

３．研究の方法

(1)については，学習データ中の矛盾の定義，および，定義に基づく矛盾の量の定量化を行う．定量化にあたっては，データセットに含まれる（説明変数の）類似する個体のペアについて，目的変数の値の相違に基づいて矛盾の度合いを定量化する．

(2)については，機械学習による 2 クラス分類問題を対象とし，混同行列（Confusion Matrix）の値の期待値および分散を neg/pos ratio に基づいて算出し，それに基づいた neg/pos-normalized measures を定義する．

４．研究成果

まず，(1)については，データ矛盾性の尺度 Similar Case Inconsistency Level (SCIL)を定義した．具体的には，図 1 に示すように，データセットに関する定義をまず行い，その定義に基づいて，目的変数の類似性に関する定義（図 2），および，説明変数の類似性に関する定義（図 3）を行った．これら定義に基づいて図 4 の通り SCIL を定義した．図中の d_{NR}（normalized rank of relative similarity）の定義については，文献[1]を参照されたい．多数のデータセットを用いた評価実験の結果，矛盾の少ないデータセットほど，得られる機械学習モデルの予測性能が高い傾向にあることを示した．

次に，(2)については，2 クラス分類問題における代表的な性能評価指標である Precision (positive predictive value)，Recall (true positive rate)，NPV(negative predictive value)，Specificity (true negative rate)について，それらの期待値をデータセットの neg/pos ratio に基づいて定義した．多数のソフトウェアバグデータセットを対象とした適用実験の結果を図 5 に示す．図 5 では，X 軸に評価指標の期待値，Y 軸に従来の評価指標を示している．Y 軸の値が X 軸の値を下回る場合，その予測は成功とはいえない．特に，図 5(b)の recall については多数のデータセットにおいて成功とはいえないケースがあることが分かった．また，これら期待値に基づいて，neg/pos-normalized measures を提案し，それらと従来の評価指標との比較を行った．その結果，従来の評価指標と提案指標による結果は大きく異なっており，提案指標の必要性が明らかとなった．結果の詳細については文献[2]を参照されたい．

Let D be a data set of N software projects,

$$D = \{\mathbf{p}_i,\ 1 \le i \le N\}, \tag{1}$$

where $\mathbf{p}_i$ stands for the i^{th} project (or equivalently its feature vector).

Let the feature f_{m^*} represent the estimation target and consider that all other features, i.e. f_m such that $m \neq m^*$, are disposable for estimating its value. Henceforth, we refer to f_{m^*} as "estimation target variable" (or simply as "target variable") and f_m as "estimator variables".

Suppose that $\mathbf{P}_D$ stands for the set which contains all possible pairs of different projects belonging to the data set D,

$$\mathbf{P}_D = \{\mathbf{p}_{ij} \mid \mathbf{p}_i, \mathbf{p}_j \in D,\ i \le j\}, \tag{2}$$

where $\mathbf{p}_{ij}$ is simply an unordered pair $(\mathbf{p}_i, \mathbf{p}_j)$.

図 1. データセットに関する定義（[1]より抜粋）

The similarity of a project pair $\mathbf{p}_{ij}$ in terms of the target variable f_{m^*} is assessed based on their *relative distance*. Specifically, we denote the relative distance of a project pair $\mathbf{p}_{ij}$ in terms of the target variable f_{m^*} with $d_R(\mathbf{p}_{ij}, f_{m^*})$,

$$d_R(\mathbf{p}_{ij}, f_{m^*}) = \frac{\left|\mathbf{p}_i[f_{m^*}] - \mathbf{p}_j[f_{m^*}]\right|}{\frac{\mathbf{p}_i[f_{m^*}] + \mathbf{p}_j[f_{m^*}]}{2}}. \tag{7}$$

If $d_R(\mathbf{p}_{ij}, f_{m^*})$ is smaller than 1, i.e.

$$d_R(\mathbf{p}_{ij}, f_{m^*}) < 1, \tag{8}$$

then the project pair $\mathbf{p}_{ij}$ is considered to have *similar* target variables. Otherwise (i.e. $d_R(\mathbf{p}_{ij}, f_{m^*}) \ge 1$), it is regarded to have *dissimilar* target variables.

Let $\overset{\approx}{\mathbf{P}}_{D,m^*}$ denote the set of project pairs in D with similar target variable f_{m^*}:

$$\overset{\approx}{\mathbf{P}}_{D,m^*} = \{\mathbf{p}_{ij} \mid d_R(\mathbf{p}_{ij}, f_{m^*}) < 1,\ \mathbf{p}_{ij} \in \mathbf{P}_D\}. \tag{9}$$

Moreover, let $\overset{\not\approx}{\mathbf{P}}_{D,m^*}$ denote the set of project pairs with dissimilar target variables f_{m^*}:

$$\overset{\not\approx}{\mathbf{P}}_{D,m^*} = \{\mathbf{p}_{ij} \mid d_R(\mathbf{p}_{ij}, m^*) \ge 1,\ \mathbf{p}_{ij} \in \mathbf{P}_D\}. \tag{10}$$

図 2．目的変数の類似性に関する定義（[1]より抜粋）

The similarity of a project pair $\mathbf{p}_{ij}$ in terms of the estimator variables f_m is assessed based on their normalized rank of relative similarity d_{NR}.

Let $\overset{\approx}{\mathbf{P}}_{D,m}$ denote the set of project pairs with similar estimator variables f_m.

$$\overset{\approx}{\mathbf{P}}_{D,m} = \{\mathbf{p}_{ij} \mid d_{NR}(\mathbf{p}_{ij}, f_m) < \alpha,\ \mathbf{p}_{ij} \in \mathbf{P}_D\}, \tag{11}$$

where α is a threshold in the interval between 0 and 1. In addition, suppose that $\overset{\not\approx}{\mathbf{P}}_{D,m}$ is the set of project pairs with dissimilar estimator variables f_m.

$$\overset{\not\approx}{\mathbf{P}}_{D,m} = \{\mathbf{p}_{ij} \mid d_{NR}(\mathbf{p}_{ij}, f_m) \ge 1 - \alpha,\ \mathbf{p}_{ij} \in \mathbf{P}_D\}. \tag{12}$$

Note that when $\alpha \le d_{NR}(\mathbf{p}_{ij}, f_m) < 1 - \alpha$ we do not regard the project pair $\mathbf{p}_{ij}$ to be neither similar nor dissimilar in terms of the estimator variables.

Let $\mathbf{P}_{D,R1}$ denote the set of inconsistent project pairs of D, which satisfy (R1). Then, $\mathbf{P}_{D,R1}$ can be written as

$$\mathbf{P}_{D,R1} = \left\{\mathbf{p}_{ij} \middle| \ \mathbf{p}_{ij} \in \overset{\not\approx}{\mathbf{P}}_{D,m^*},\ \mathbf{p}_{ij} \in \overset{\approx}{\mathbf{P}}_{D,m}\right\}. \tag{13}$$

図 3. 説明変数の類似性に関する定義（[1]より抜粋）

$$\mathbf{SCIL}(D) = \frac{\#(\mathbf{P}_{D,R1})}{\#(\mathbf{P}_D)}. \tag{19}$$

図 4. SCIL の定義（[1]より抜粋）

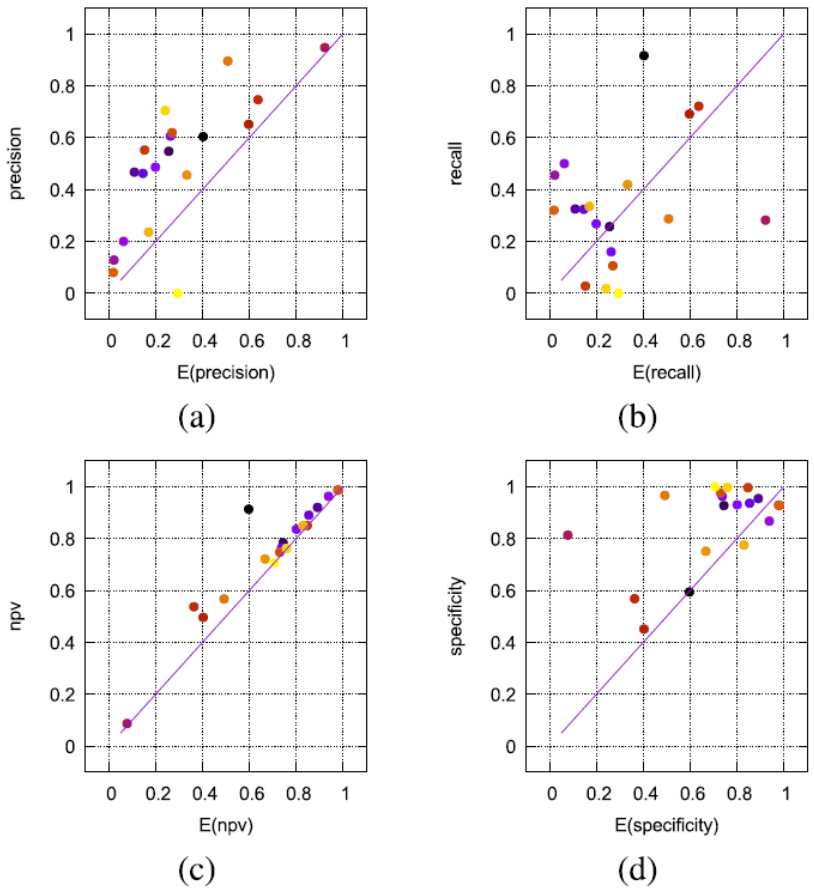

図 5. 性能評価指標の値とその期待値の比較（[2]より抜粋）

参考文献

[1] Maohua Gan, Zeynep Yücel, Akito Monden, "Improvement and evaluation of data consistency metric CIL for software engineering datasets," IEEE Access, Vol. 10, pp. 70053-70067, 2022.

[2] Maohua Gan, Zeynep Yücel, Akito Monden, "Neg/pos-normalized Accuracy Measures for Software Defect Prediction," IEEE Access, Vol. 10, pp. 134580-134591, 2022.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 西浦 生成、門田 暁人	４．巻 40
２．論文標題 Fault-proneモジュール予測における第三者データに基づいた外れ値除去	５．発行年 2023年
３．雑誌名 コンピュータ ソフトウェア	６．最初と最後の頁 4_22～4_28
掲載論文のDOI（デジタルオブジェクト識別子） 10.11309/jssst.40.4_22	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Maohua Gan, Zeynep Yucel, Akito Monden	４．巻 10
２．論文標題 Improvement and Evaluation of Data Consistency Metric CIL for Software Engineering Data Sets	５．発行年 2022年
３．雑誌名 IEEE Access	６．最初と最後の頁 70053-70067
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2022.3188246	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Maohua Gan, Zeynep Yucel, Akito Monden	４．巻 10
２．論文標題 Neg/pos-Normalized Accuracy Measures for Software Defect Prediction	５．発行年 2022年
３．雑誌名 IEEE Access	６．最初と最後の頁 134580-134591
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2022.3232144	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 田中 和也, 門田 暁人, Zeynep Yucel	４．巻 38
２．論文標題 ソフトウェア開発工数予測におけるauto-sklearnの適用	５．発行年 2021年
３．雑誌名 コンピュータソフトウェア	６．最初と最後の頁 4_46-4_52
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 Takumi Kanehira, Akito Monden, Zeynep Yucel	4．巻 1
2．論文標題 Association Metrics Between Two Continuous Variables for Software Project Data	5．発行年 2021年
3．雑誌名 Proc. 22nd IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing	6．最初と最後の頁 1-6
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 無
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 Teruki Hayakawa, Masateru Tsunoda, Koji Toda, Keitaro Nakasai, Amjed Tahir, Kwabena Ebo Bennin, Akito Monden, and Kenichi Matsumoto	4．巻 E104-D
2．論文標題 A Novel Approach to Address External Validity Issues in Fault Prediction Using Bandit Algorithms	5．発行年 2021年
3．雑誌名 IEICE Transactions on Information and Systems	6．最初と最後の頁 327-331
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6件（うち招待講演　0件／うち国際学会　2件）

1．発表者名 横山 大貴, 西浦 生成, 門田 暁人
2．発表標題 BERTによるセキュリティバグの判別の試み
3．学会等名 ソフトウェア工学の基礎ワークショップFOSE2023
4．発表年 2023年

1．発表者名 Kinari Nishiura, Takeki Kasagi, Akito Monden
2．発表標題 A Cost-Effectiveness Metric for Association Rule Mining in Software Defect Prediction
3．学会等名 2023 Congress in Computer Science, Computer Engineering & Applied Computing (CSCE)（国際学会）
4．発表年 2023年

１．発表者名
Hiroshi Demanou, Akito Monden, Masateru Tsunoda

２．発表標題
A Dynamic Model Selection Approach to Mitigate the Change of Balance Problem in Cross-Version Bug Prediction

３．学会等名
10th International Workshop on Quantitative Approaches to Software（国際学会）

４．発表年
2022年

１．発表者名
西脇将樹，門田暁人，笹倉万里子，西浦生成

２．発表標題
データ断片からのソフトウェア開発データ復元の実験評価

３．学会等名
電子情報通信学会ソフトウェアサイエンス研究会

４．発表年
2022年

１．発表者名
西脇 将樹，門田 暁人

２．発表標題
データ断片からのソフトウェア開発データの復元の試み

３．学会等名
第27回ソフトウェア工学の基礎ワークショップ

４．発表年
2020年

１．発表者名
伊永 健人，門田 暁人

２．発表標題
ソフトウェア開発工数予測におけるデータスムージングの実験的評価

３．学会等名
第27回ソフトウェア工学の基礎ワークショップ

４．発表年
2020年

〔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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関