5J-01

可逆圧縮を利用したプロダクト派生関係推定手法の実験的評価 - 圧縮アルゴリズム毎の誤りの傾向分析 -

索手　一平 †　早瀬　康裕 ‡　北川　博之 ‡

† 筑波大学情報学群情報科学類　　‡ 筑波大学システム情報系情報工学域

1　はじめに

既存のプロダクトに変更を加えることで，新しいプロダクト (派生先プロダクト) を開発することがある．派生先プロダクトはもととなったプロダクト (派生元プロダクト) と多くの部分を共有するため，それらの派生関係を記録することで保守を効率的に行うことができる．しかし，開発の過程でその記録が失われてしまうことがある．

この問題を解決するため，筆者らはプロダクトの派生関係グラフを推定する手法 EEGL[1] を提案した．EEGL はプロダクトを可逆圧縮し，データ量の増分が少ないプロダクト間を，データ量の小さいプロダクトを始点として接続することで派生関係グラフを推定する．

過去に行われた実験によって，EEGL が採用する圧縮アルゴリズムによって，推定結果の精度や誤りの傾向が変化することが分かっている．

そこで本研究では，新たな圧縮アルゴリズムを加え，出力される誤りの内容について比較分析を行った．その結果，プロダクトの派生時に加わる変更の単位や種類，規模と出力される誤りの間に相関があることがわかった．また，圧縮アルゴリズムによって，出力する誤りと相関のあるプロダクトの変化が異なることがわかった．

2　誤りの分析

2.1　実験の目的と方法

本研究では EEGL の出力する誤りの内容を分析することで，圧縮アルゴリズムごとの誤りの傾向の違いを明らかにする．圧縮アルゴリズムによって誤りの傾向が異なることから，誤りの内容も圧縮アルゴリズムによって異なることが期待される．

分析は EEGL の推定結果に含まれる正解の辺と誤りの辺について，それらの表すプロダクトの変化の違いを比較することで行う．プロダクトの変化とは，プロダクトに加わる変更のことであり，変更の単位と種類，その規模によって表される．例えば，変更は「行の削除量が 10 行である」というように表現される．このとき変更の単位は「行」であり，変更の種類は「削除量」，変更の規模は「10 行」となる．本分析で用いる，プロダクトの変更の単位と種類を表 1 に示す．

単位	
ディレクトリ	LOC
ファイル	拡張子数
行	言語数
サイズ	

(a) 変更の単位

種類	
増加量	増加率
追加量	追加率
削除量	削除率
追加・削除量	追加・削除率
共有量	共有率

(b) 変更の種類

表 1: プロダクトの変更の単位と種類

ディレクトリとファイルは，それぞれプロダクトに含まれるディレクトリとファイルの総数である．行とサイズは，それぞれプロダクトに含まれる全てのファイルの行数とサイズの総和である．LOC は，CLOC[2] によって計測される，プロダクトに含まれるファイルの LOC の総和である．拡張子数はプロダクトに含まれるファイルの拡張子の種類の総数である．言語数とは，プロダクトに含まれる全てのファイルのプログラミング言語を，CLOC を用いて推定した時の，言語の種類の総数である．変更の種類の計算方法は，派生元プロダクトと派生先プロダクトにおけるある単位 (例:ディレクトリ) の集合を N_B, N_D とすると，増加量は $|N_D|-|N_B|$，追加量は $|N_D \setminus N_B|$，削除量は $|N_B \setminus N_D|$，追加・削除量は $|N_D \triangle N_B|$($\triangle$ は 2 つの集合の対象差を表す)，共有量は $|N_D \cap N_B|$ で表される．増加率，追加率，削除率，変更率，共有率はそれぞれ対応する量を $|N_B|$ で

Evaluation of estimating method for product evolution graph using lossless compression
-error analysis on the basis of compression algorithm-

Ippei Nawate†(ippei.nawate@kde.cs.tsukuba.ac.jp),
Yasuhiro Hayase‡(hayase@cs.tsukuba.ac.jp),
Hiroyuki Kitagawa‡(kitagawa@cs.tsukuba.ac.jp)
†College of Information Sciences, University of Tsukuba
‡Faculty of Engineering, Information and Systems, University of Tsukuba

割った値である．

個々の圧縮アルゴリズムを分析する手順を以下に記す．この分析の結果を圧縮アルゴリズム間で比較することで，誤りの傾向の違いを明らかにする．

1. 以下の手順 (a)〜(c) を複数のプロダクト集合 (データセット) に対して行う．
   (a) EEGL を用いてデータセットの派生関係グラフを推定する．
   (b) 推定されたグラフの辺を，正解の辺の集合 E_c と誤りの辺の集合 E_e に分類する．
   (c) 表 1 に示した変更の単位と種類について，E_c と E_e の順位平均の有意差を信頼係数 95%で検定する
2. 表 1 の変更の単位と種類ごとに，有意差を示したデータセットを数える．このとき，変更の規模が大きいほど推定に成功するのか (positive な有意差)，失敗するのか (negative な有意差) によって別々に数える．

2.2 実験

gzip，bzip2，xz，ppmd，Re-Pair の 5 つのアルゴリズムを用いて EEGL による派生関係グラフの推定を行い，推定結果に含まれる誤りの傾向を比較分析した．推定を行うデータセットには，神田らが OSS をもとに作成・公開しているデータセット dataset1，2，5-9[3] の 7 つを使用した．

2.3 実験結果と考察

実験結果を図 1 に示す．左側のグラフは positive な有意差を数えたもの，右側のグラフは negative な有意差を数えたものである．縦軸は変更の単位と種類を，横軸は圧縮アルゴリズムを表す．各マスの色は有意差を示したデータセット数を表しており，数が多いほど黒に近くなる．

図 1 では多くのマスに色がついていることから，プロダクトの派生時に加わる変更と EEGL の出力する誤りとの間に相関があることがわかる．各列ごとに色のばらつき方が異なっていることから，圧縮アルゴリズムによって，出力する誤りと相関のある変更の単位と種類，規模が異なる事がわかる．いくつかの変更の単位と 種類では，positive と negative の両方で同じ圧縮アルゴリズムの列に色がついている．このことから，推定するデータセットの性質によって，誤りの傾向が変化していることがわかる．特に bzip2，xz の列ではその傾向が強く，データセットの性質によって推定結果が変化しやすいのだと考えられる．

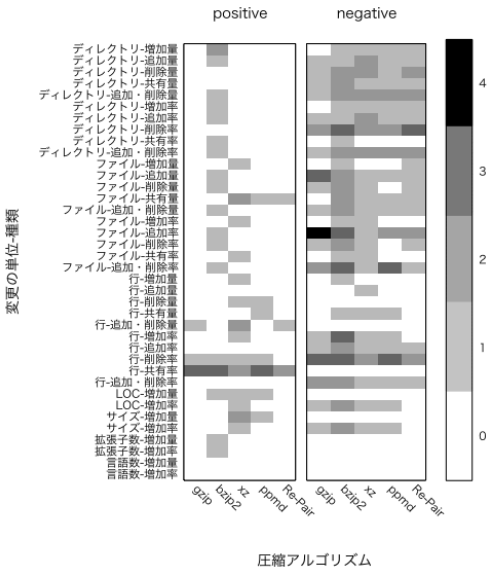

図 1: 実験結果

今回の実験は OSS をベースとしたデータセットを利用して行われたが，企業で開発される商用のプロダクトは OSS とは異なる性質の派生関係を持つ可能性がある．そのため，今回得られた結果が商用のプロダクト開発にも有用であるかは議論の余地があるといえる．

3 まとめ

本研究では，複数の圧縮アルゴリズムについて，EEGL の出力する誤りの内容を分析した．実験の結果，プロダクトの派生時に加わる変更の単位と種類，規模と出力される誤りとの間に相関があり，圧縮アルゴリズムごとにその相関は異なるということがわかった．今後の課題として，辺の表すプロダクトの変化から推定結果の誤理を予測することや，各圧縮アルゴリズムを組み合わせた推定手法を提案することが考えられる．

参考文献

[1] Yasuhiro Hayase, Tetsuya Kanda, and Takashi Ishio. Estimating product evolution graph using kolmogorov complexity. IWPSE 2015.

[2] http://cloc.sourceforge.net.

[3] Tetsuya Kanda, Takashi Ishio, and Katsuro Inoue. Extraction of product evolution tree from source code of product variants. SPLC '13.