

Associating Code Clones with Association Rules for Change Impact Analysis

Manishankar Mondal, Banani Roy, Chanchal K. Roy,
Kevin A. Schneider

*2020 IEEE 27th International Conference on
Software Analysis, Evolution and
Reengineering (SANER), 2020, 93-103*

影響箇所予測

- 変更時に影響する箇所を予測する研究が盛んに行われている
- コードクローンは同時に変更されることが頻繁にある[36]
- コードクローンの情報を用いて影響箇所予測を行った研究は存在していない

目的

- コードクローン情報を用いると影響箇所予測の精度向上につながるかを明らかにする
- 精度向上に役立つコードクロンの情報を明らかにする

Evolutionary Coupling(EC)

- 同時に変更される 2 組以上のコード片
- 影響分析の際に必要な概念

使用する影響箇所予測手法

- 使用する手法はTarmaq[45]
 - システムの過去のECを取得
 - 過去のECをもとに関連規則を生成
 - 生成した関連規則を用いて影響箇所を予測

関連規則

- 発生したある事象によって発生する結果が決まる規則
- $X \Rightarrow Y$ X :先例, Y :結果
- 同時変更はコードエンティティを $E1$, $E2$ とすると
$$E1 \Rightarrow E2, \quad E2 \Rightarrow E1$$

コードクローン情報の使用

- 関連規則の生成の際に過去のECとコードクローン集合の両方を用いる
- 後述する分析のために通常のクローンとミクロクローンを分ける
 - 通常のクローン:5行以上のクローン
 - ミクロクローン[33]:5行未満のクローン

リサーチクエストション

1. ECとクローンを組み合わせた場合、予測手法の精度は向上するか
2. 通常のクローンとミクロクロローンのどちらが精度向上に寄与するか
3. 精度向上に寄与するミクロクロローンのサイズは
4. 精度向上に寄与する通常クローンのサイズは
5. ECとクローンを組み合わせた場合、影響箇所はより頻繁に現れるか

データセット

TABLE II
THE SUBJECT SYSTEMS THAT WE INVESTIGATE

Systems	Lang.	Domains	LLR	SRev	LRev
Ctags	C	Code Definition Generator	33,270	1	774
Camellia	C	Multimedia	85,015	1	207
BRL-CAD	C	Solid modeling system	39,309	1	735
Carol	Java	Game	25,091	1	1700
Freecol	Java	Game	91,626	1000	1950
DNSJava	Java	DNS Protocol	23,373	1	1635
Greenshot	C#	Multimedia	37,628	1	999
MonoOSC	C#	Formats and protocols	14,883	1	355

LLR = LOC in the Last Revision

SRev = Starting Revision LRev = Last Revision

RQ1: ECとクローンを組み合わせた場合、予測手法の精度は向上するか

- 精度の計測

- $precision = \frac{|A \cap S| \times 100}{|S|}$

- $recall = \frac{|A \cap S| \times 100}{|A|}$

- A :実際に影響したコード, S :予測手法で検出した影響箇所

RQ1: ECとクローンを組み合わせた場合、予測手法の精度は向上するか

- 結果

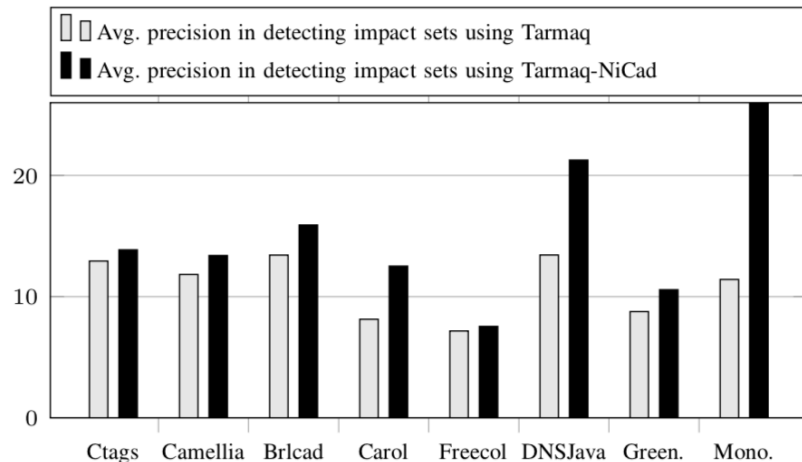


Fig. 3. Comparing the precisions of the techniques Tarmaq and Tarmaq-NiCad for detecting impact sets

precision
 $p < 0.01$ cohen's $d = 0.63$

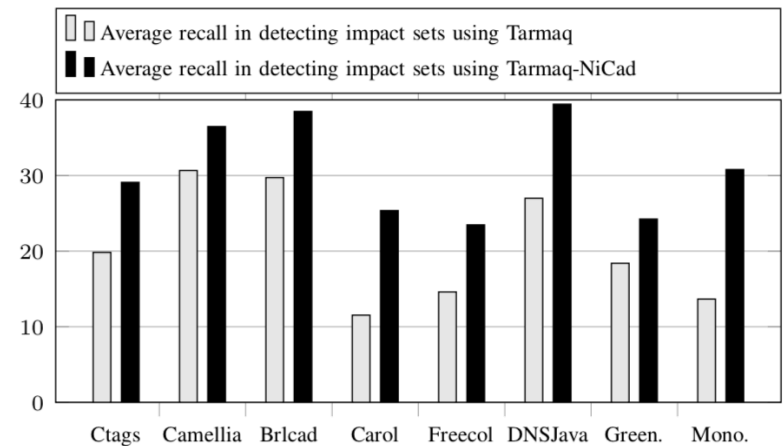


Fig. 2. Comparing the recalls of the techniques Tarmaq and Tarmaq-NiCad for detecting impact sets

recall
 $p < 0.01$ cohen's $d = 0.63$

RQ2:通常のクローンとミクロクローンのどちらが精度向上に寄与するか

- 結果

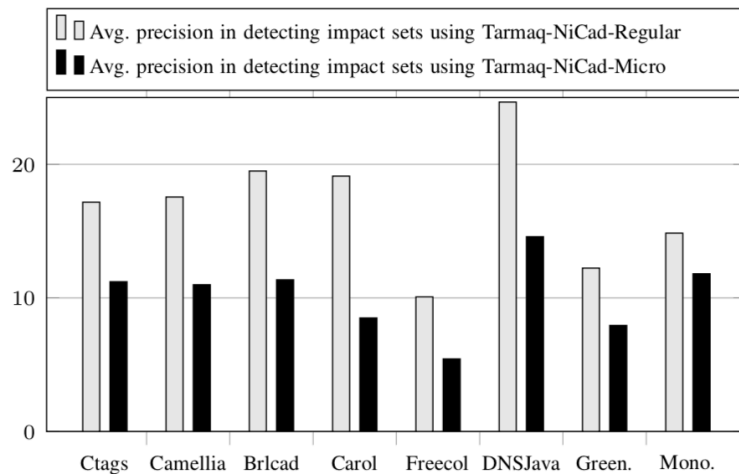


Fig. 5. Comparing the precisions of the techniques Tarmaq-NiCad-Regular and Tarmaq-NiCad-Micro for detecting impact sets

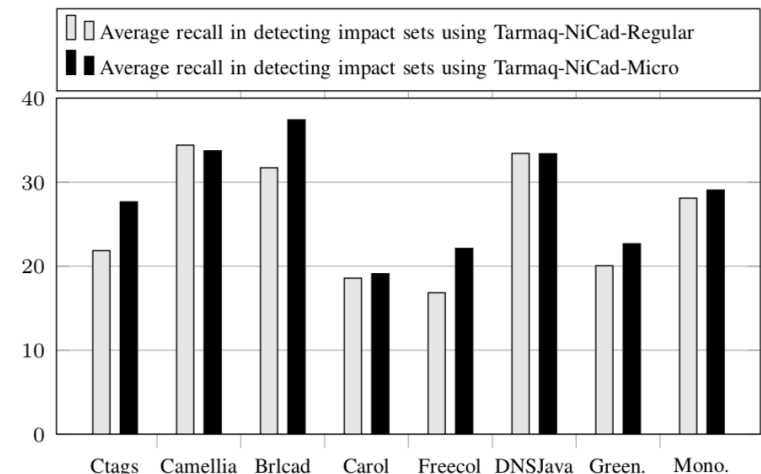


Fig. 6. Comparing the recalls of the techniques Tarmaq-NiCad-Regular and Tarmaq-NiCad-Micro for detecting impact sets

recallはミクロクローン, precisionは通常クローンの方が大きい
->ミクロクローンは数が多くrecallを押し上げる役目を果たしている

RQ3:精度向上に寄与するミクロクロンのサイズは

- 結果

TABLE IV
NUMBER OF TRUE POSITIVES OBTAINED BY ANALYZING DIFFERENT SIZES
OF MICRO-CLONES

Systems	Size = 1LOC	Size = 2LOC	Size = 3LOC	Size = 4LOC
Ctags	6	292	318	74
Camellia	2	0	138	108
Brlcad	4	2	138	966
Carol	0	30	1014	454
Freecol	94	4	2328	558
DNSJava	38	100	1338	1610
Greenshot	12	94	936	248
MonoOSC	0	8	224	272

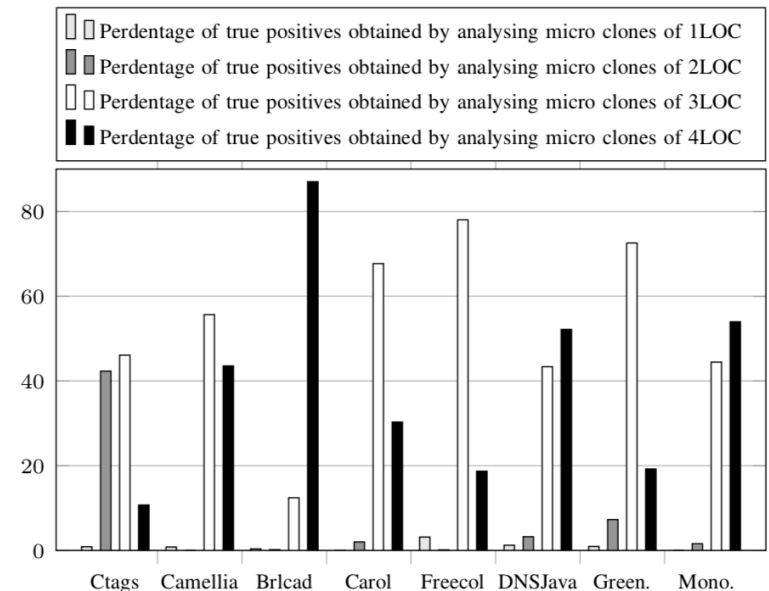


Fig. 7. Percentage of true positives obtained from different sizes of micro-clones

3, 4行の割合が多いため, 予測に用いるミクロクロンのサイズは3, 4行に絞ると精度が向上する

RQ4:精度向上に寄与する通常クローンのサイズは

- 結果

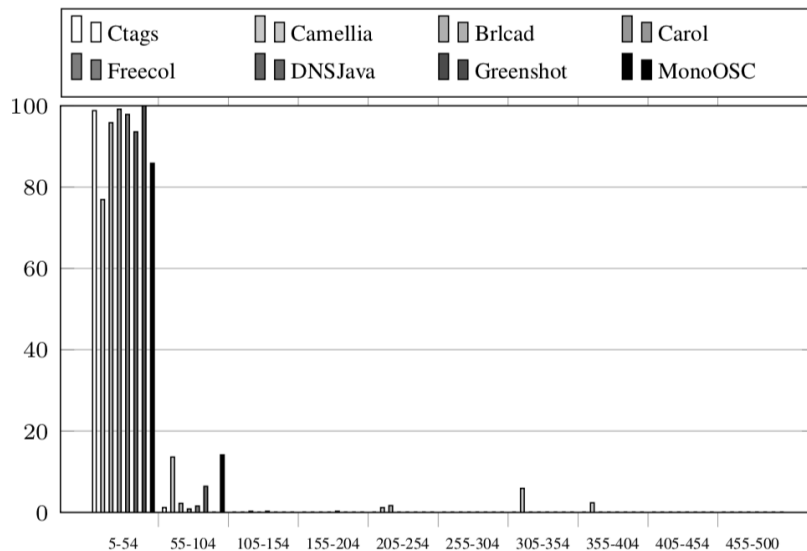


Fig. 8. Possibilities of getting true positive clone fragments at different ranges of clone sizes

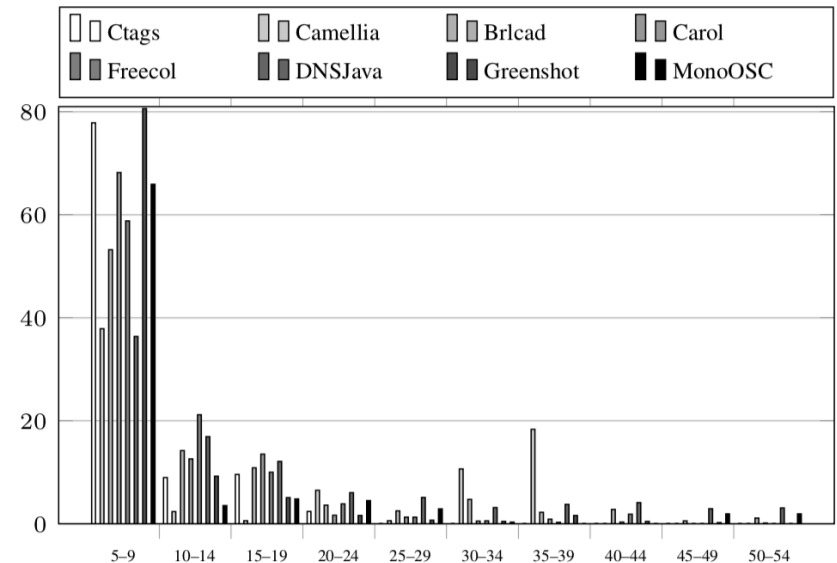


Fig. 9. Possibilities of getting true positive clone fragments at different intervals of clone sizes in the range 5 to 55

検出するクローンの行数は5-20行に絞ってもあまり精度は変わらないことがわかった

RQ5: ECとクローンを組み合わせた場合、影響箇所はより頻繁に現れるか

- 結果

TABLE V

STATISTICS REGARDING THE NUMBER OF COMMITS WHERE TARMAQ OR TARMAQ-NICAD PROVIDED IMPACT SETS WITH TRUE POSITIVES

Measure	Ctags	Camellia	Brlcad	Carol	Freecol	DNSJava	Greenshot	Mono.
No. of commits where more than one method got changed	152	88	261	243	485	613	336	150
No. of commits where Tarmaq provided true positives	71	68	167	96	267	382	212	109
No. of commits where Tarmaq-NiCad provided true positives	89	74	192	178	336	485	247	120

- 2つ以上のメソッドに変更があったコミット数
- 過去のECを使ってtrue positiveだったコミット数
- 過去のECとクローンを組み合わせてture positiveだったコミット数

過去のECとクローンを組み合わせた方が頻繁に影響箇所が現れることがわかった

まとめ

- 過去のECとクローンを組み合わせた方が影響箇所予測の精度が上がる
- ミクロクローンはrecallを底上げする
- 通常クローンは5-20行に絞っても良い

所管

- クローンタイプ別で比べるとどうなるか気になった
- 他のクローンの性質を考慮することでより精度を向上できるかもしれない
- 分析結果を用いて予測してみると精度向上や計算コスト削減ができているか気になる