

Heterogeneous Defect Prediction

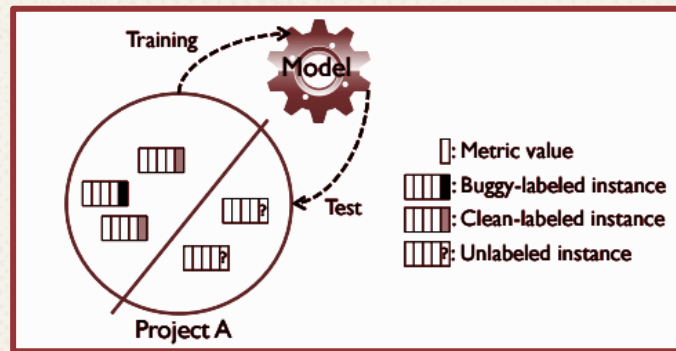
Jaechang Nam and Sunghn Kim

背景

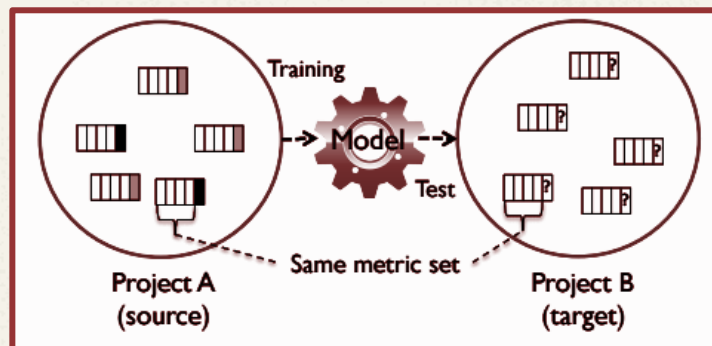
- 近年、予測研究が盛んに行われている
- 新しいプロジェクトの予測を行うのが困難
 - 過去の情報が少ない

WDPとCPDP

- WDP —Within-Project Defect Prediction
機械学習を行って、同プロジェクトの未来を予測する.



- CPDP —Cross-Project Defect Prediction
プロジェクトAで機械学習を行って、プロジェクトBの未来を予測する.



本研究の目的

- CDPDの課題
 - 寄与率の高いメトリクスが抜けてしまう可能性
→ 予測精度の低下
- 提案手法
 - HDP —Heterogeneous Defect Prediction
 - Promise Datasetなどを機械学習モデル構築に使えるように



Heterogeneous Defect Prediction



Source: Project A

X_1	X_2	X_3	X_4	Label
1	1	3	10	Buggy
8	0	1	0	Clean
⋮	⋮	⋮	⋮	⋮
9	0	1	1	Clean

Target: Project B

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Label
3	1	1	0	2	1	9	?
1	1	9	0	2	3	8	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	1	1	1	2	1	1	?

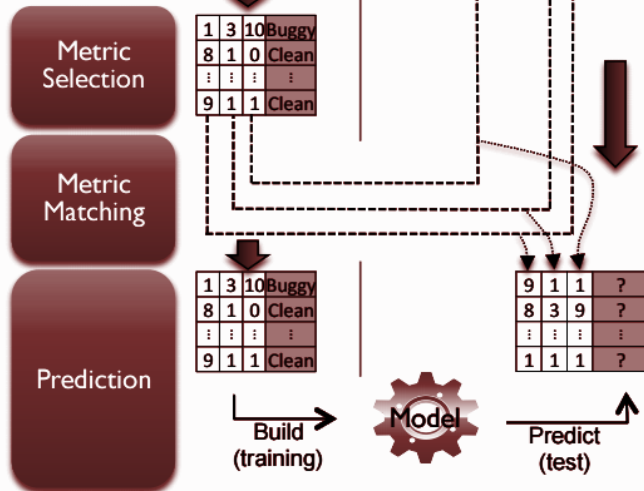
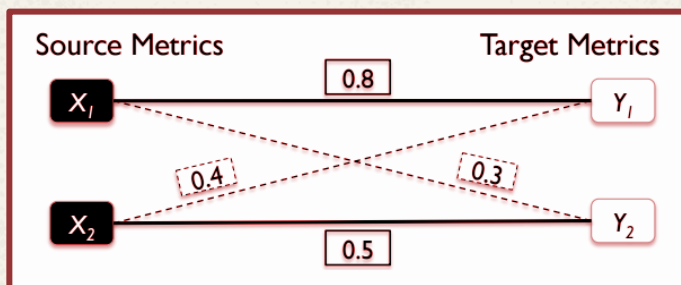


Figure 2: Heterogeneous defect prediction

- Sourceのデータセットから価値のあるメトリクスを抽出する
- メトリクス同士の類似度を測って、違うメトリクスも使う.
- 類似度測定
 - Percentile BM
 - Kolomogorov-Smirnov BM
 - Spearman's correlation BM



AUCの測定

Source	WPDP (Baseline1)	CPDP-CM (Baseline2)	CPDP-IFS (Baseline3)	HDP KS,0.05	Target Coverage of HDP
AEEEM	0.654	0.736	0.528	0.739*	48%
ReLink	0.654	0.665	0.500	0.702*	88%
MORPH	0.657	0.667	0.590	0.736*	100%
NASA	0.654	0.527	0.500	0.734*	52%
SOFTLAB	0.695	0.612	0.554	0.708*	100%

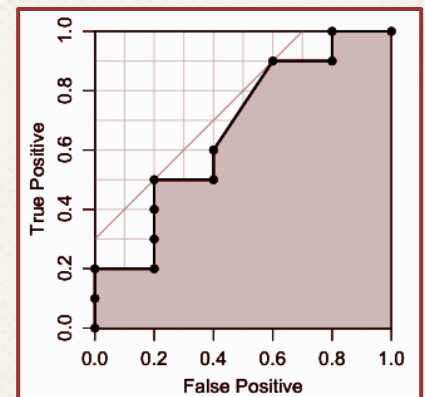
5銘柄でそれぞれのAUCを測定

- AUCのメジアンはHDPが一番良い結果だった。

- AUC

ROC曲線の下面積。

ランダムに1つ選んだとき, Trueになる確率。



各アルゴリズムの 詳細な比較

Target	Against								
	WPDP (Baseline1)			CPDP-CM (Baseline2)			CPDP-IFS (Baseline3)		
	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss
EQ	4	0	0	2	2	0	4	0	0
JDT	0	0	5	3	0	2	5	0	0
LC	6	0	1	3	3	1	3	1	3
ML	0	0	6	4	2	0	6	0	0
PDE	3	0	2	2	0	3	5	0	0
Apache	6	0	5	8	1	2	9	0	2
Safe	14	0	3	12	0	5	15	0	2
Zxing	8	0	0	6	0	2	7	0	1
ant-1.3	6	0	1	6	0	1	5	0	2
arc	3	1	0	3	0	1	4	0	0
camel-1.0	3	0	2	3	0	2	4	0	1
poi-1.5	2	0	2	3	0	1	2	0	2
redaktor	0	0	4	2	0	2	3	0	1
skarbonka	11	0	0	4	0	7	9	0	2
tomcat	2	0	0	1	1	0	2	0	0
velocity-1.4	0	0	3	0	0	3	0	0	3
xalan-2.4	0	0	1	1	0	0	1	0	0
xerces-1.2	0	0	3	3	0	0	1	0	2
cm1	7	1	2	8	0	2	9	0	1
mw1	5	0	1	4	0	2	4	0	2
pc1	1	0	5	5	0	1	6	0	0
pc3	0	0	7	7	0	0	7	0	0
pc4	0	0	7	2	0	5	7	0	0
ar1	14	0	1	14	0	1	11	0	4
ar3	15	0	0	5	0	10	10	2	3
ar4	16	0	0	14	1	1	15	0	1
ar5	14	0	4	14	0	4	16	0	2
ar6	7	1	7	8	4	3	12	0	3
Total	147 66.2%	3 1.4%	72 32.4%	147 66.2%	14 6.3%	61 27.5%	182 82.0%	3 1.3%	37 16.7%

AEEM, ReLink, MORPH,
NASA, SOFTLABの5銘柄, 28
プロジェクトでAUCの値の高さを
競わせた.

- 多くの場合HDPが優秀だと
わかった.

アルゴリズムを変えて比較

HDP Learners	Against					
	WPDP		CPDP-CM		CPDP-IFS	
	Win	Tie	Win	Tie	Win	Tie
Logistic	66.2%	1.4%	66.2%	6.3%	82.0%	2.7%
RandomForest	10.4%	2.3%	42.3%	1.4%	65.8%	2.2%
BayesNet	34.7%	4.1%	45.9%	2.7%	66.2%	2.7%
SVM	24.3%	23.0%	27.5%	0.0%	32.9%	14.9%
J48	30.2%	11.7%	32.4%	1.4%	41.9%	12.1%
SimpleLogistic	45.5%	2.7%	69.4%	6.8%	84.2%	3.2%
LMT	42.8%	3.2%	64.4%	6.3%	79.7%	3.2%

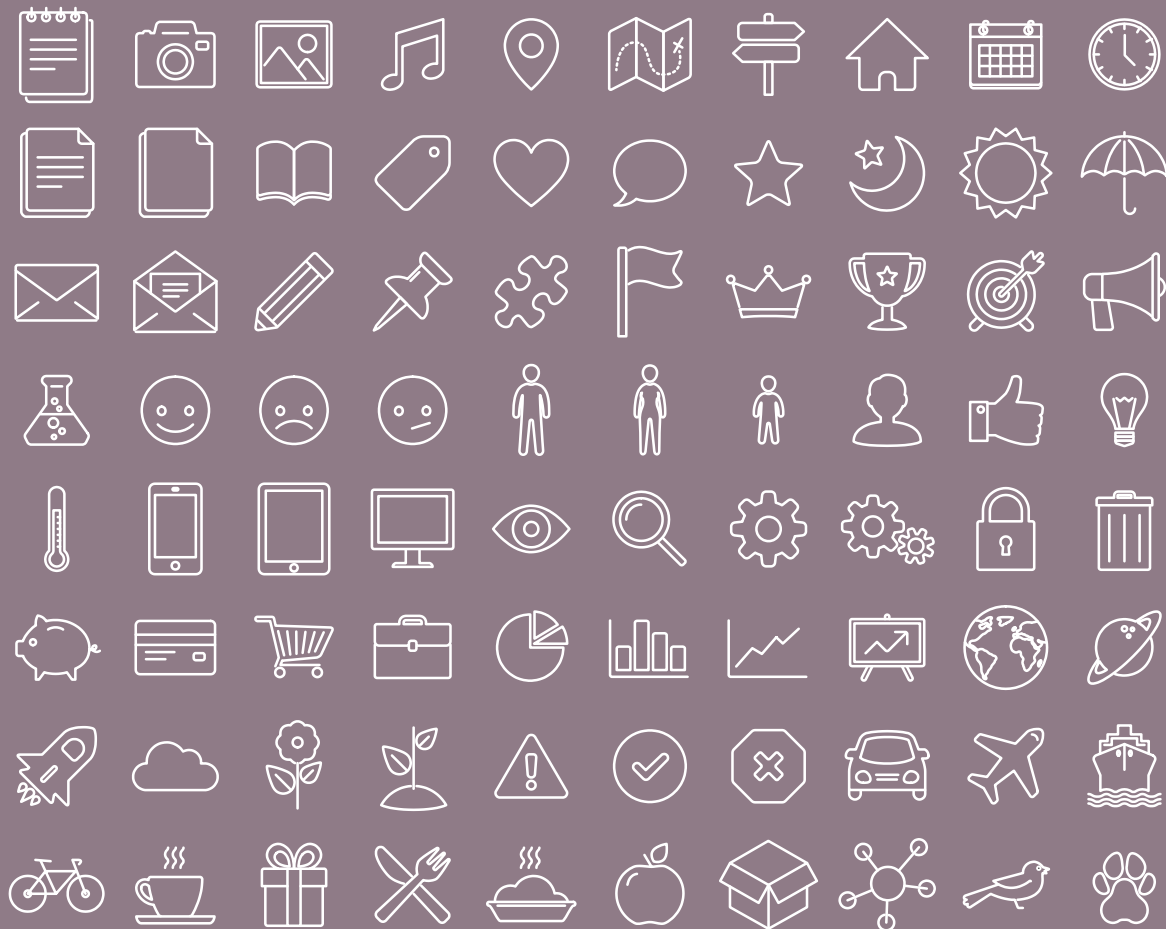
- 多くの場合, TieかWin (少なくともHDPの方が良いということ)
- 中でもLogisticによる予測は, 既存の手法を上回った

まとめ

- あるプロジェクトで構築した機械学習モデルで, 違うプロジェクトに予測を適応する手法, **HDP**を提案した.
 - **HDP**...メトリクスの類似度を使った手法
- 多くの場合は, 既存の手法よりも良い結果が得られた.
 - 特に**Logistic**を使った場合は顕著
- **Promise Datasets**を使った研究に役立ててほしい

所感

- 予測研究の新しい手法
- 類似度の数式は難解



SlidesCarnival icons are editable shapes.

This means that you can:

- Resize them without losing quality.
- Change line color, width and style.

Isn't that nice? :)

Examples:

