2018 NASAC报告交流

How far have we progressed in the journey?

An examination of cross-project defect prediction

周毓明

南京大学

2018.11.23

提 纲

① 跨项目缺陷预测

② 简单的预测模型

③ 实验结果的对比

④ 实验结论与建议

提 纲

① 跨项目缺陷预测

② 简单的预测模型

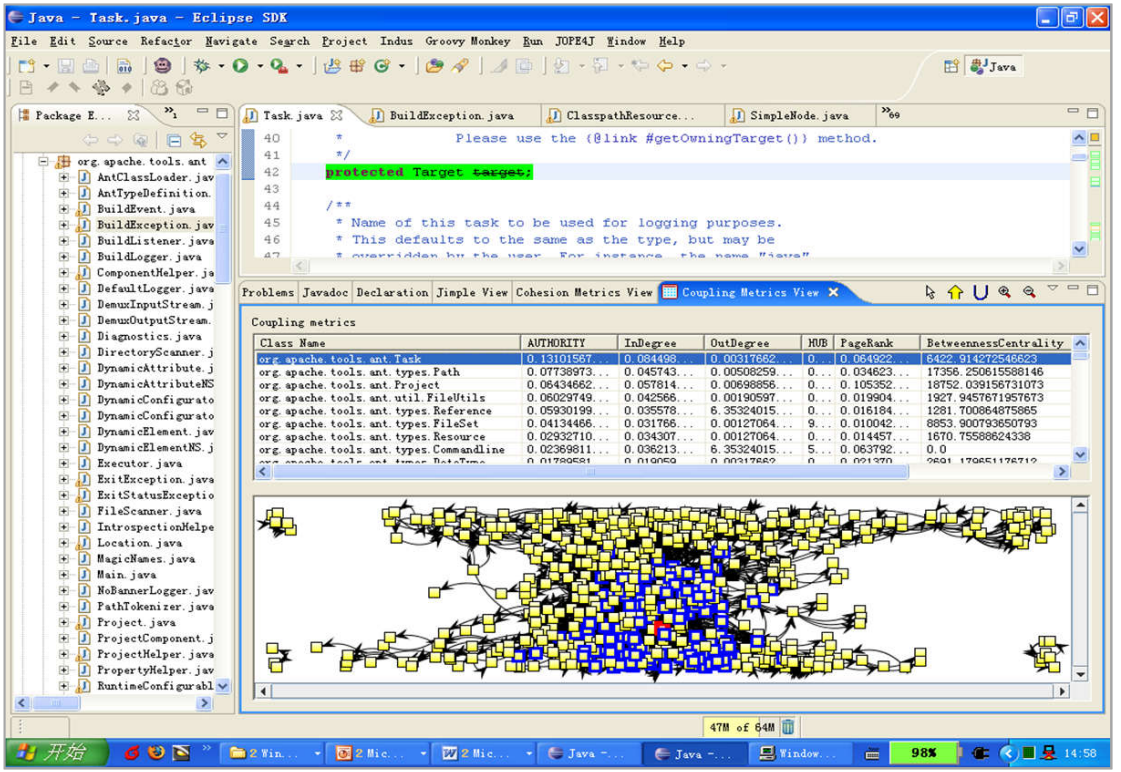
③ 实验结果的对比

④ 实验结论与建议

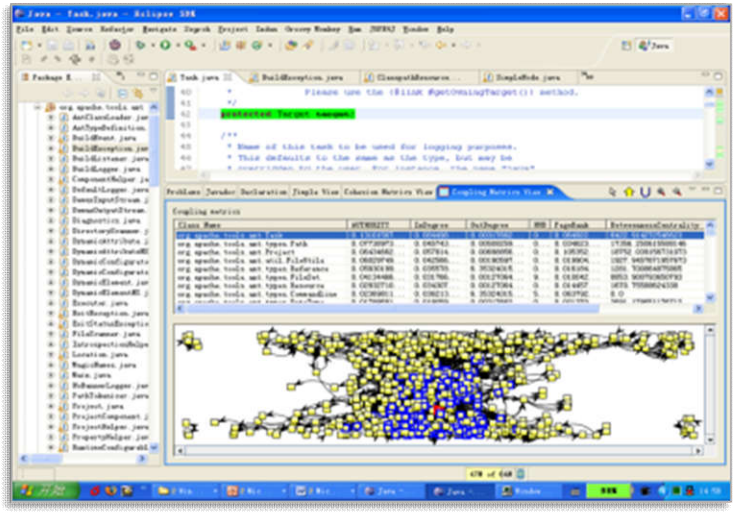
缺陷预测：基本问题

哪些模块有可能包含bug?

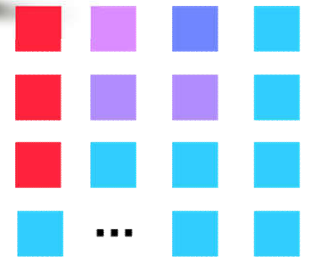
一个软件系统



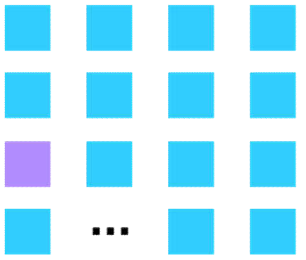
缺陷预测：分类场景



分类

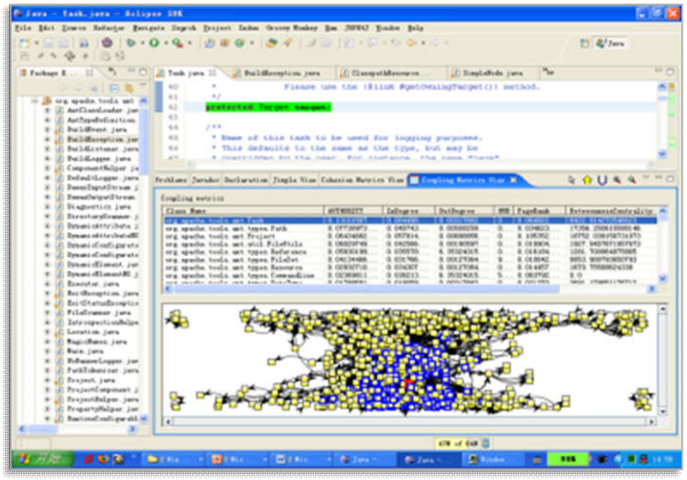


可能
有bug

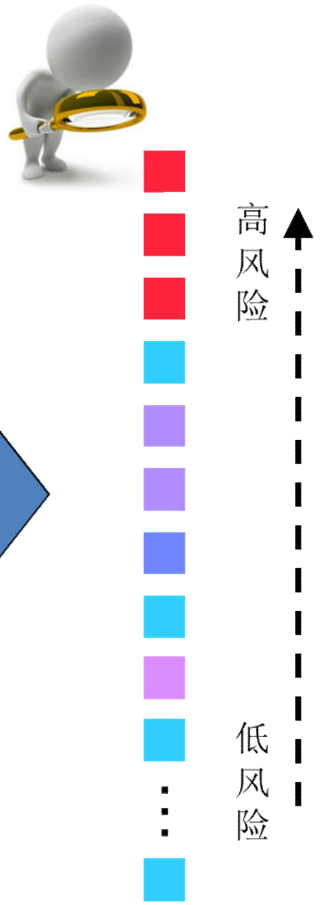


可能
无bug

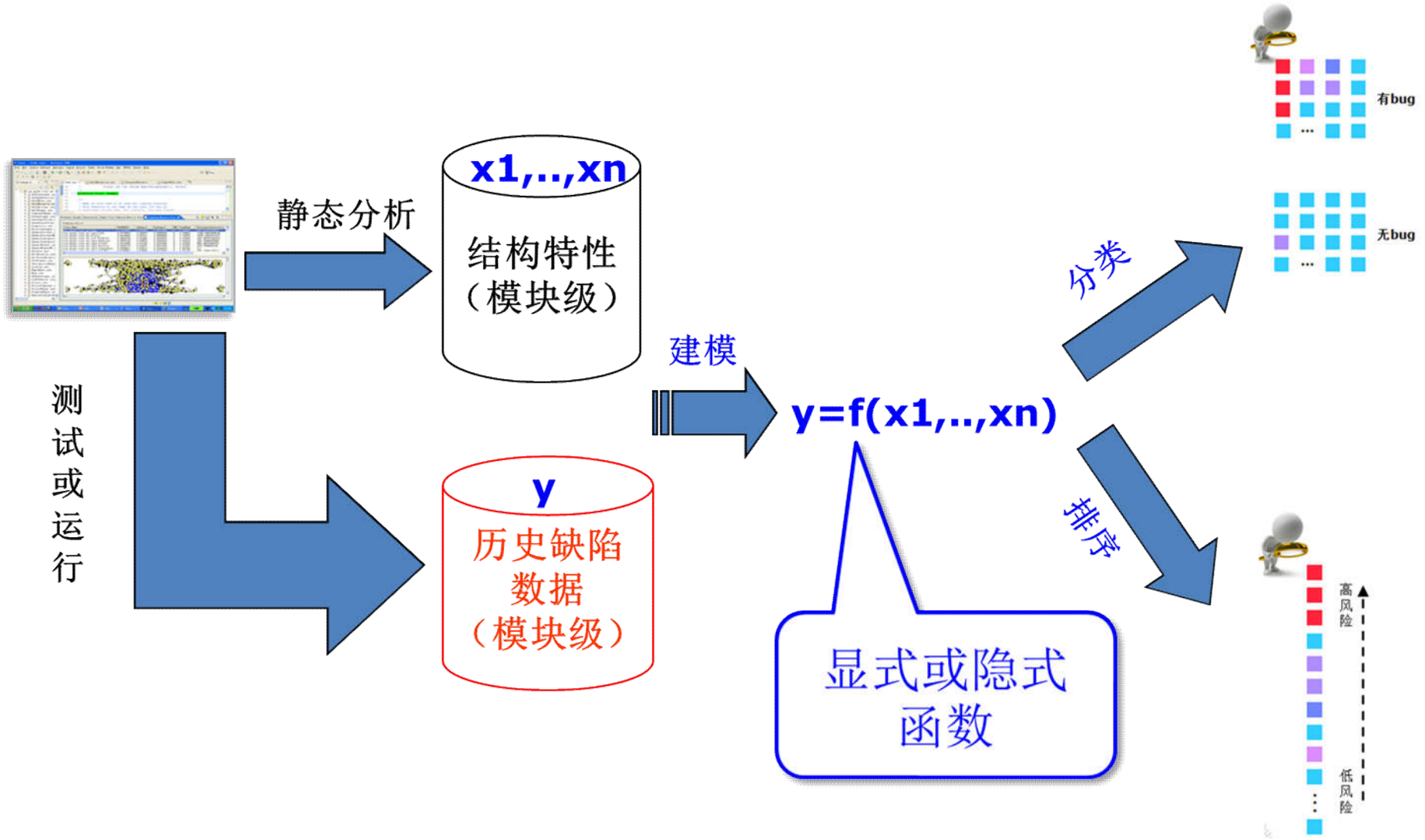
缺陷预测：排序场景



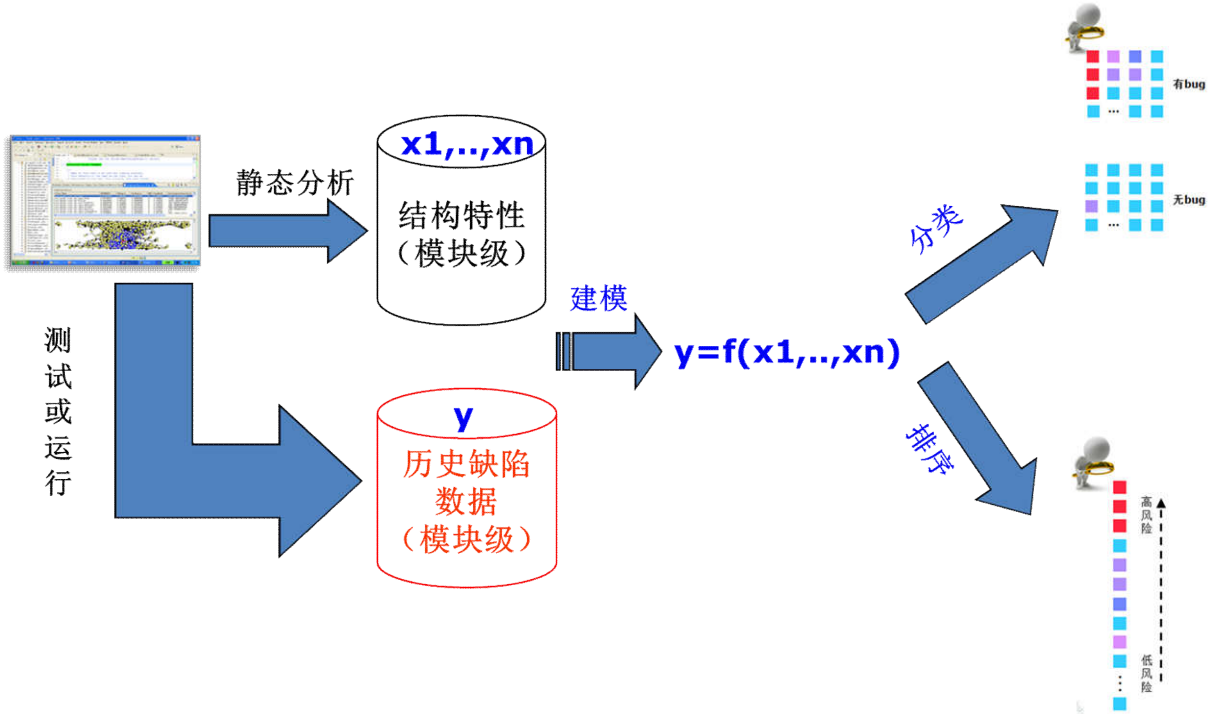
排序



缺陷预测：一般流程


静态分析
x1,..,xn
结构特性
（模块级）
测试或运行
y
历史缺陷
数据
（模块级）
建模
y=f(x1,..,xn)
显式或隐式
函数
分类
排序
有bug
无bug
高风险
低风险

缺陷预测：一般流程



如果训练数据**足够大**，
缺陷预测一般具有较好性能

跨项目缺陷预测：基本概念

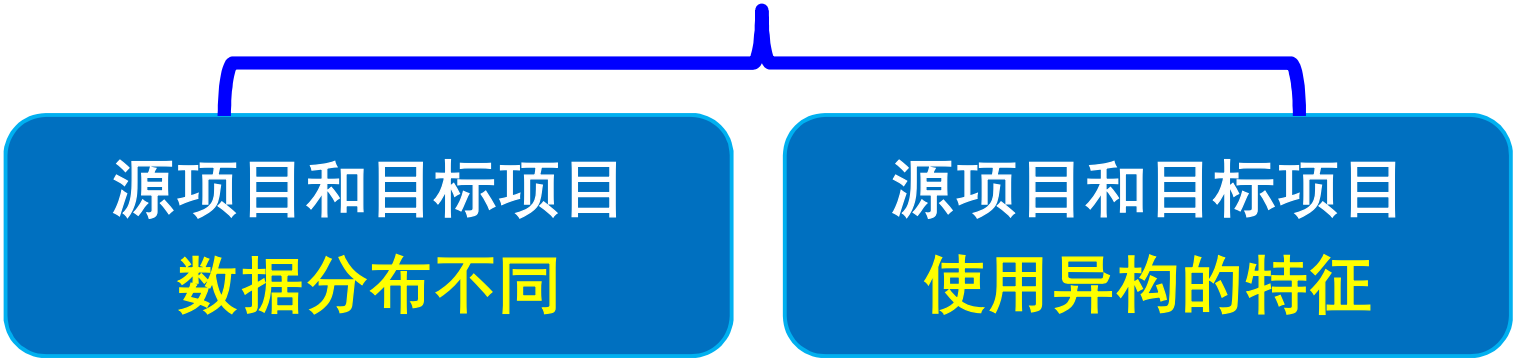
如果训练数据足够大，
缺陷预测一般具有较好性能

能否使用其他项目（源项目）的数据训练预测模型，在目标项目上使用（CPDP）？

跨项目缺陷预测：主要挑战

如果训练数据足够大，
缺陷预测一般具有较好性能

能否使用其他项目（源项目）的数据训练预测模型，在目标项目上使用（CPDP）？



跨项目缺陷预测：主要挑战

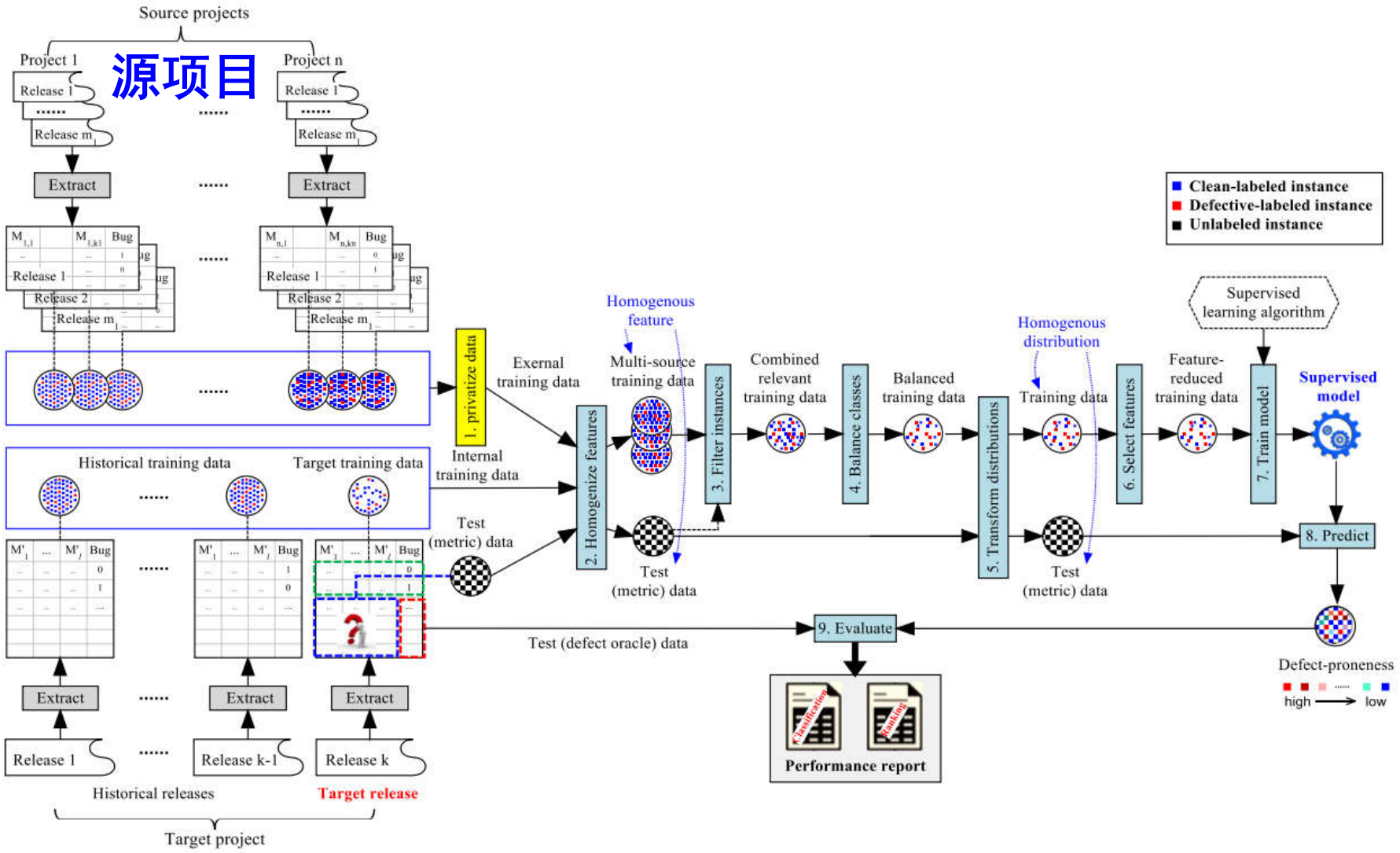
源项目和目标项目
数据分布不同

挑战1：常规的建模技术要求训练集和测试集具有相似的数据分布特性？

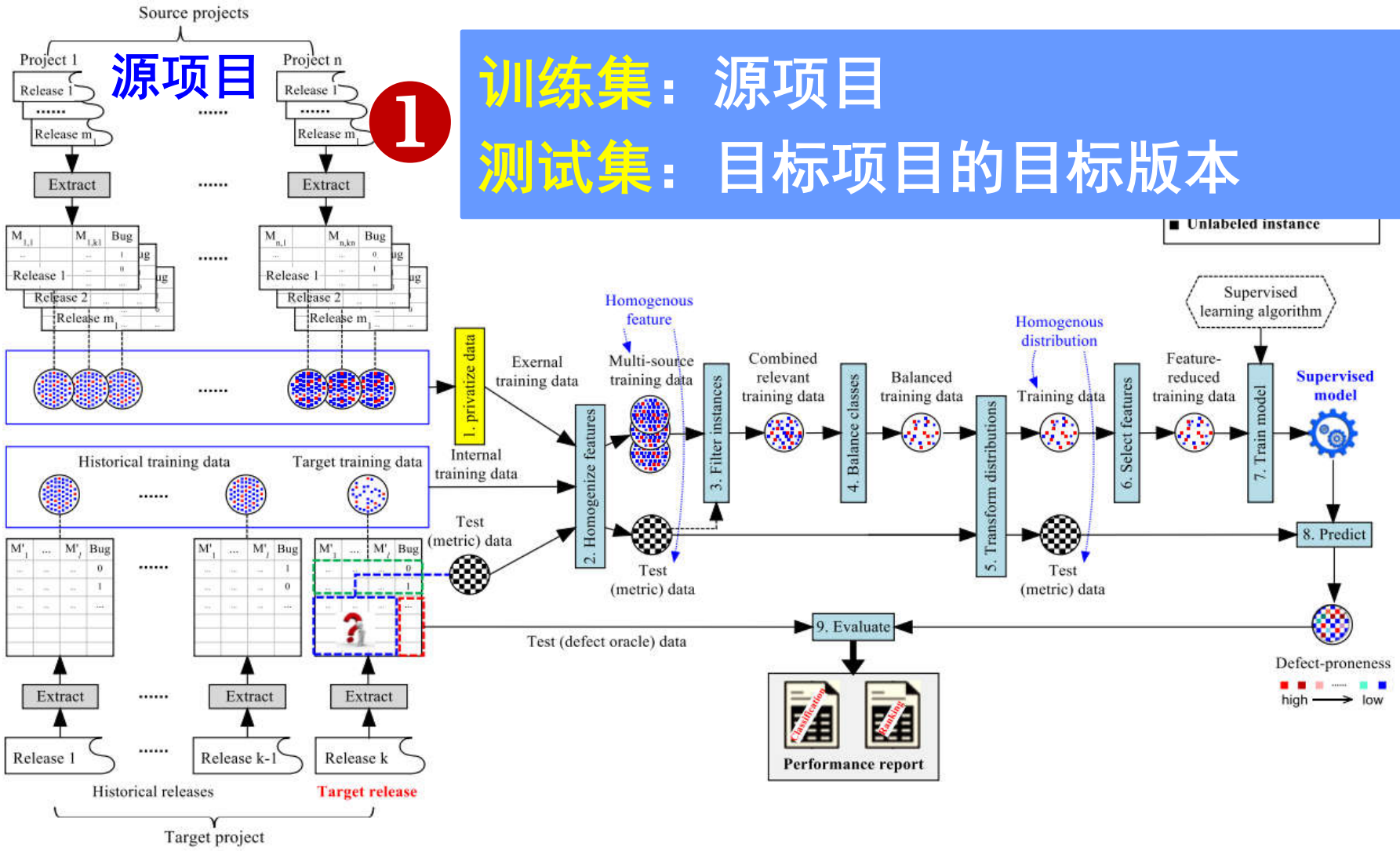
源项目和目标项目
使用异构的特征

挑战2：常规的建模技术要求训练集和测试集使用相同的特征集？

跨项目缺陷预测：通用框架



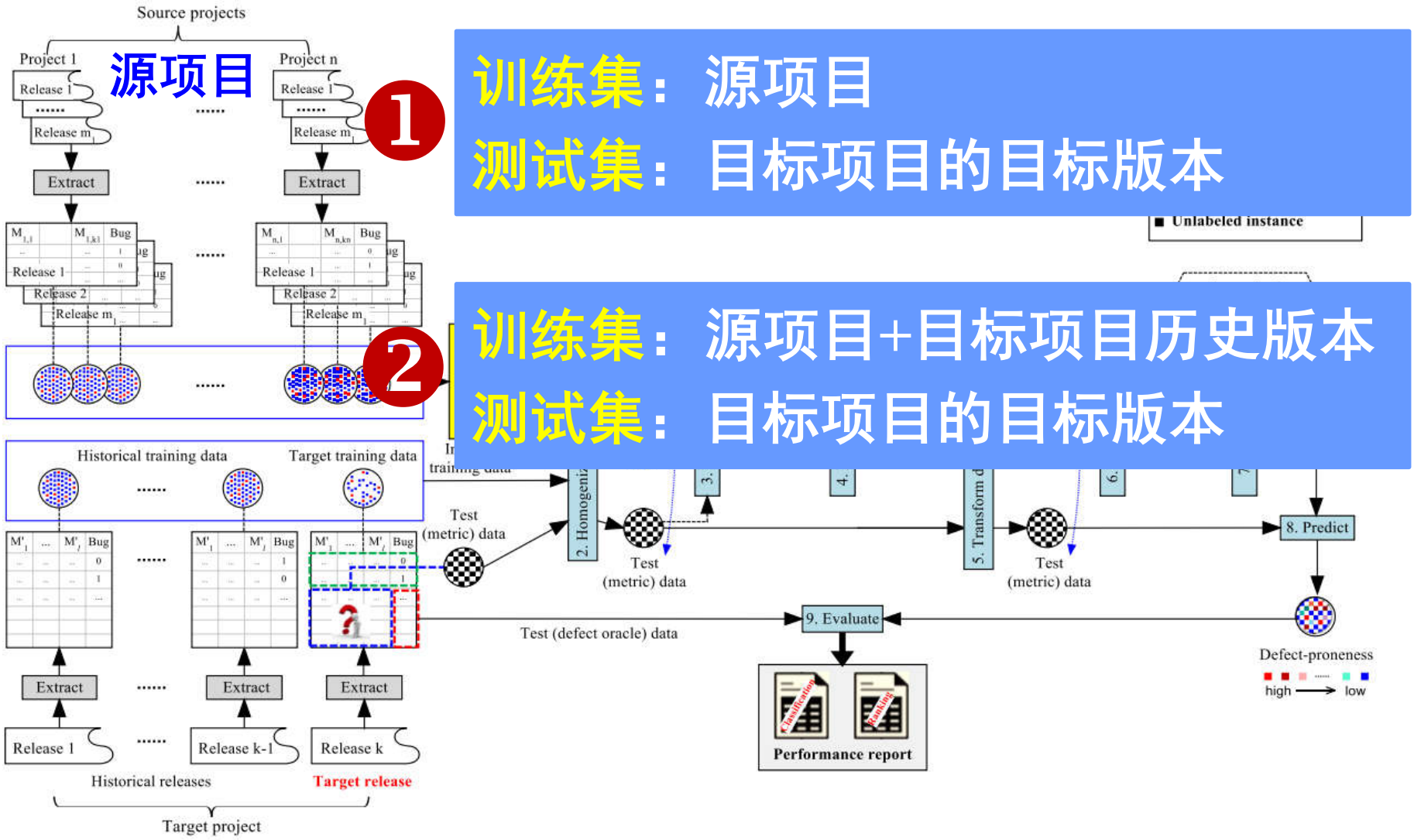
跨项目缺陷预测：通用框架



训练集：源项目

测试集：目标项目的目标版本

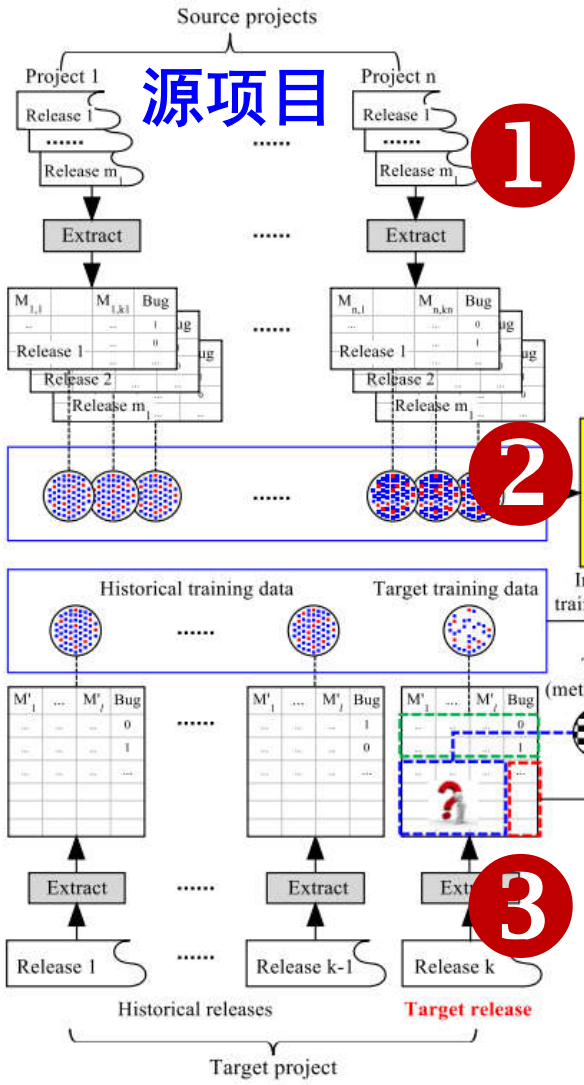
跨项目缺陷预测：通用框架



训练集： 源项目
测试集： 目标项目的目标版本

训练集： 源项目+目标项目历史版本
测试集： 目标项目的目标版本

跨项目缺陷预测：通用框架

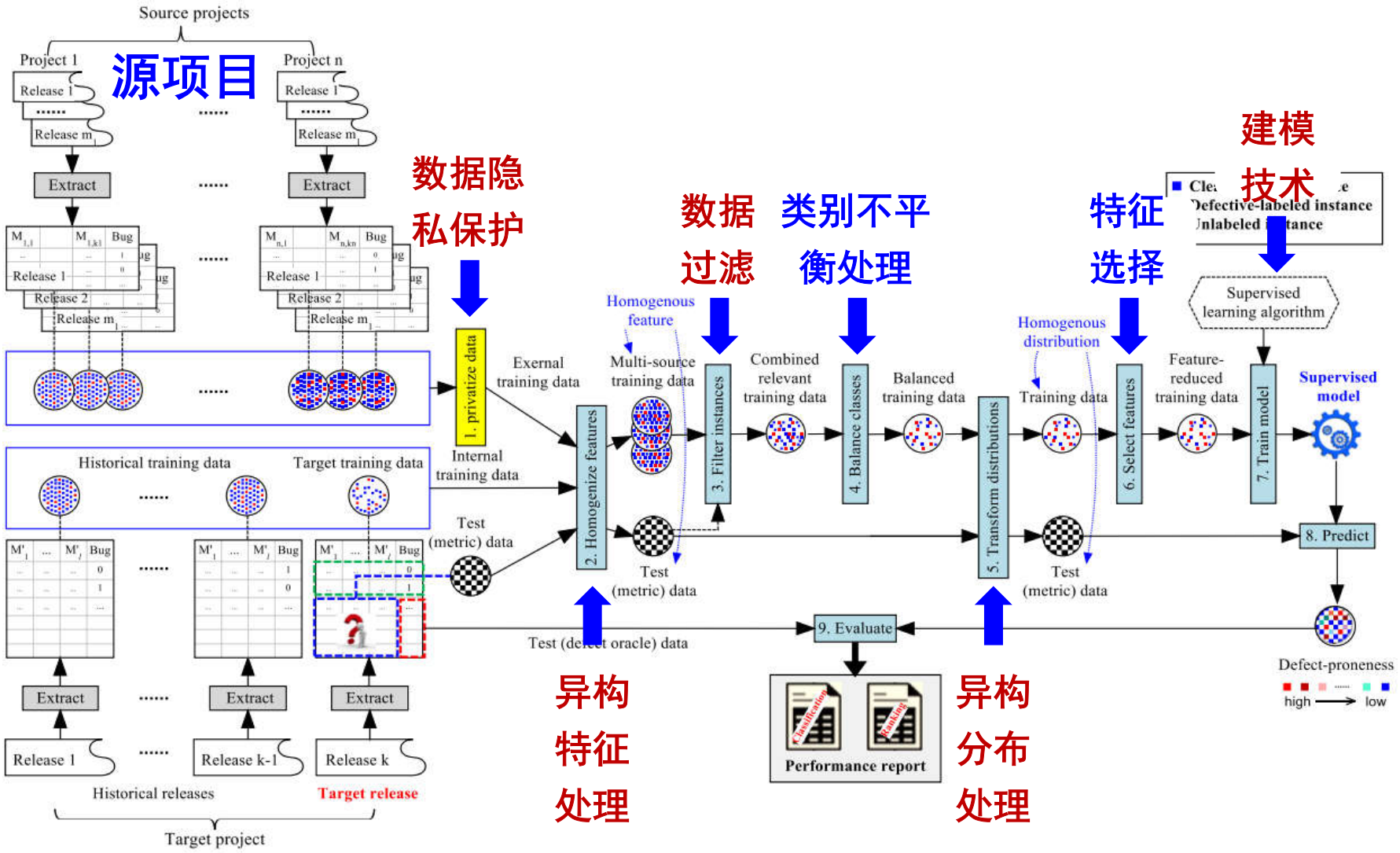


训练集： 源项目
测试集： 目标项目的目标版本

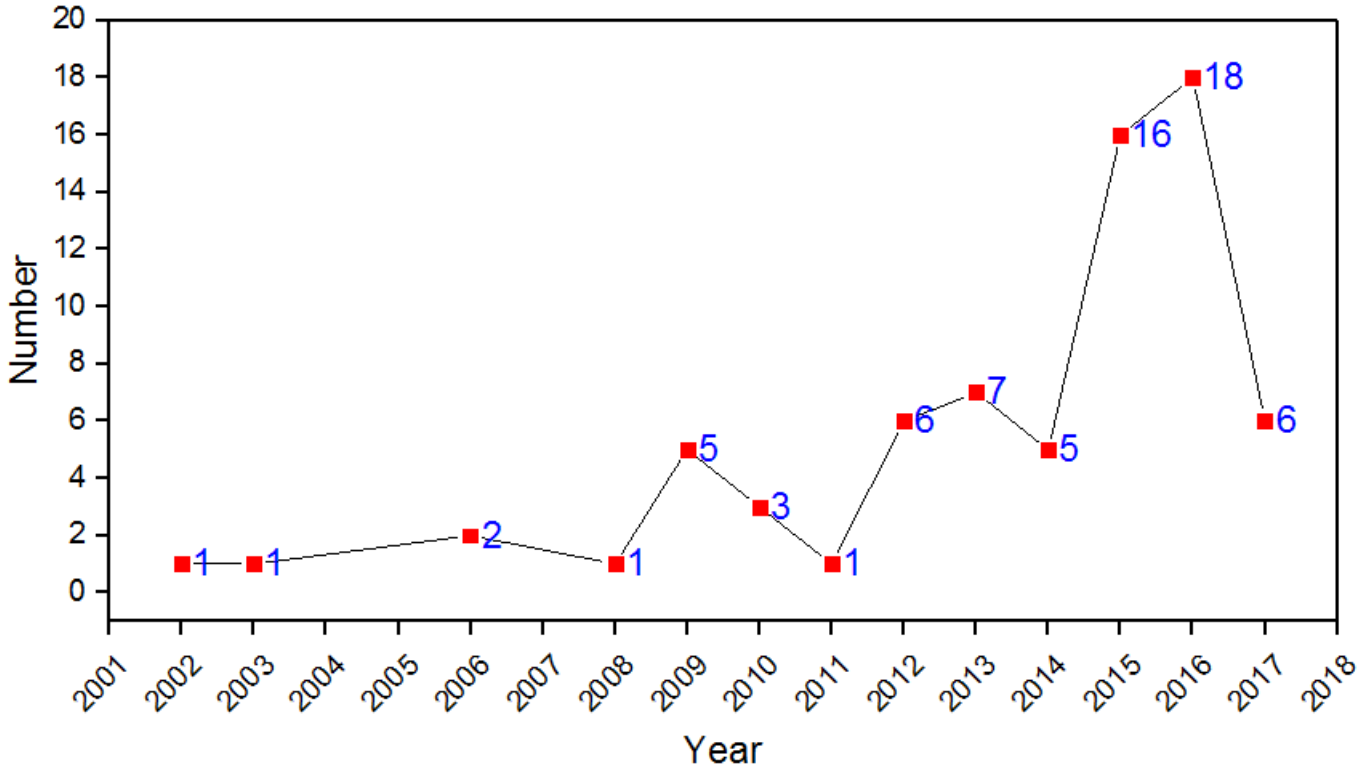
训练集： 源项目+目标项目历史版本
测试集： 目标项目的目标版本

训练集： 源项目+目标项目历史版本
+目标版本的小部分数据
测试集： 目标版本的其余数据

跨项目缺陷预测：通用框架



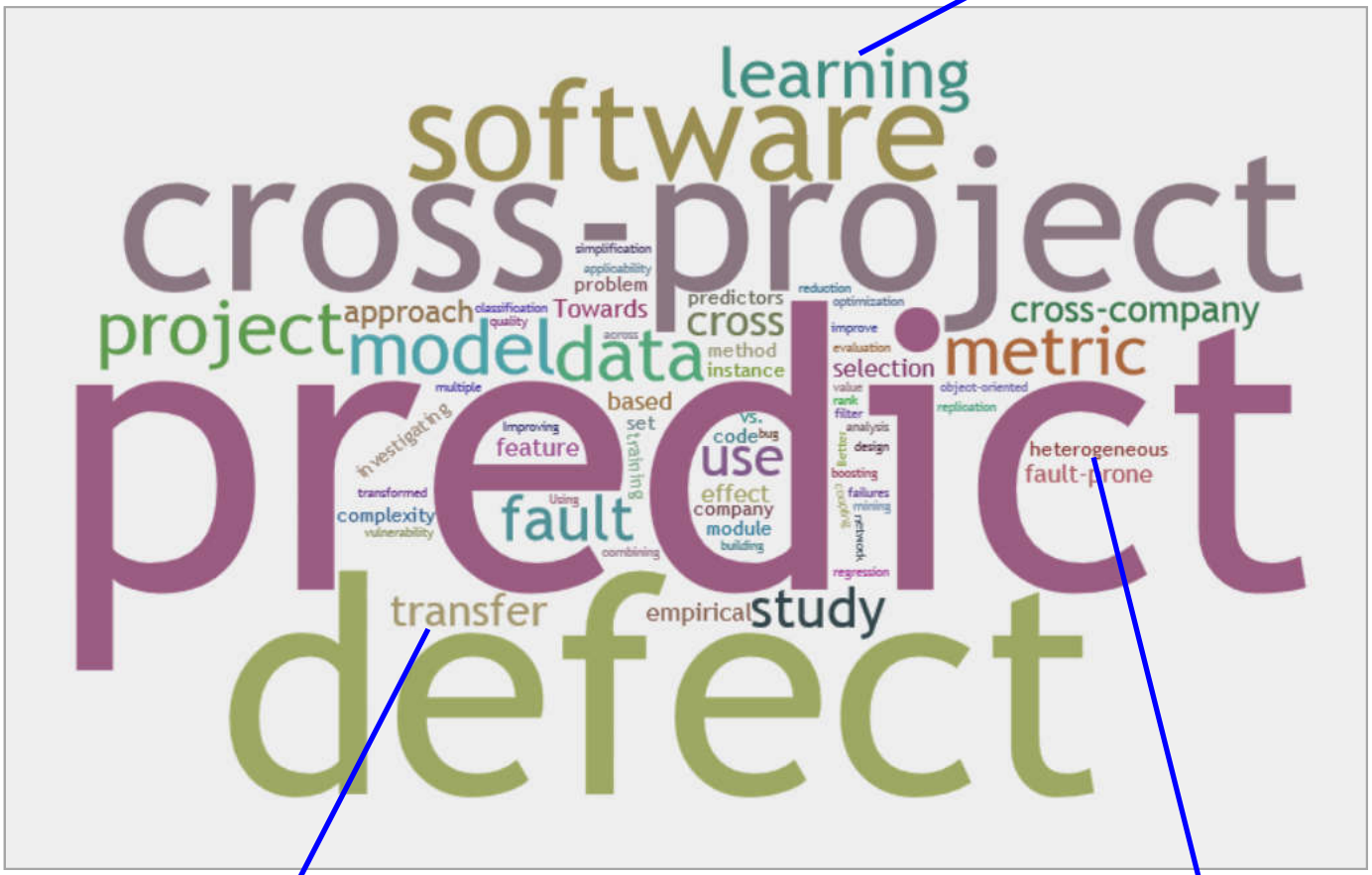
跨项目缺陷预测：当前文献



到2017年3月止，发表72篇

（英文、全文可得、期刊/会议版只算期刊、分类/排序）

跨项目缺陷预测：当前文献



learning

transfer

heterogeneous

提 纲

① 跨项目缺陷预测

② 简单的预测模型

③ 实验结果的对比

④ 实验结论与建议

简单模型：重要性

Does Bug Prediction Support Human Developers?
Findings from a Google Case Study

Chris Lewis[1], Zhongpeng Lin[1], Caitlin Sadowski[2], Xiaoyan Zhu[3], Rong Ou[2], E. James Whitehead Jr.[1]
[1]University of California, Santa Cruz, USA [2]Google Inc., USA [3]Xi'an Jiaotong University, China
{cflewis,linzhp,ejw}@soe.ucsc.edu, {supertri,rongou}@google.com, xyxyzh@gmail.com

C. Scaling

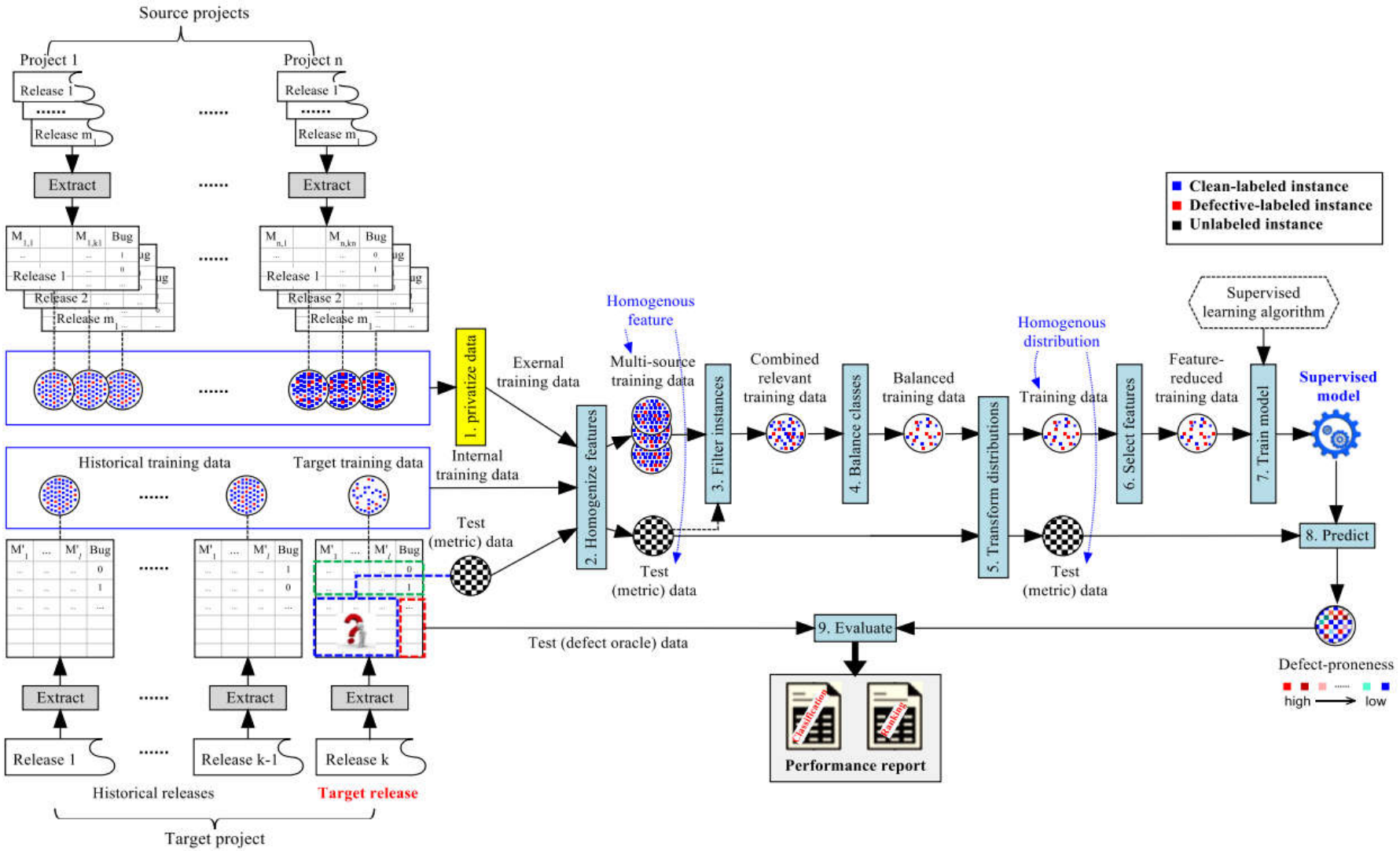
As a pre-requisite for deploying a bug prediction algorithm at Google, we had to ensure that the bug prediction algorithm would scale to Google-level infrastructure. We ran into two

简单 ✚ 有效

简单模型：重要性

Less, but better

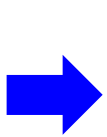
简单的预测模型


Source projects
Project 1
Project n
Release 1
Release m1
Extract
Release 2
Historical training data
Target training data
Historical releases
Target release
Target project
Release k-1
Release k
1. privatize data
Exernal training data
Internal training data
Test (metric) data
2. Homogenize features
Homogenous feature
Multi-source training data
Test (metric) data
3. Filter instances
Combined relevant training data
4. Balance classes
Balanced training data
5. Transform distributions
Homogenous distribution
Training data
Test (metric) data
6. Select features
Feature-reduced training data
7. Train model
Supervised learning algorithm
Supervised model
8. Predict
Defect-proneness
high
low
Clean-labeled instance
Defective-labeled instance
Unlabeled instance
9. Evaluate
Test (defect oracle) data
Performance report

简单的预测模型

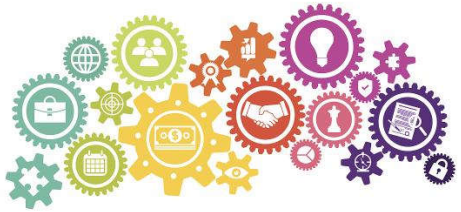
训练集

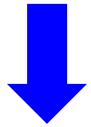
No	WMC	LCOM	…	SLOC	Fault
1	35	96	…	1255	2
2	73	100		2828	0
…	…	…	…	…	…



训练模型

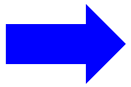
复杂模型



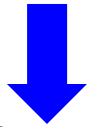


测试集

No	WMC	LCOM	…	SLOC	Fault
1	15	9	…	135	?
2	45	10		282	?
…	…				

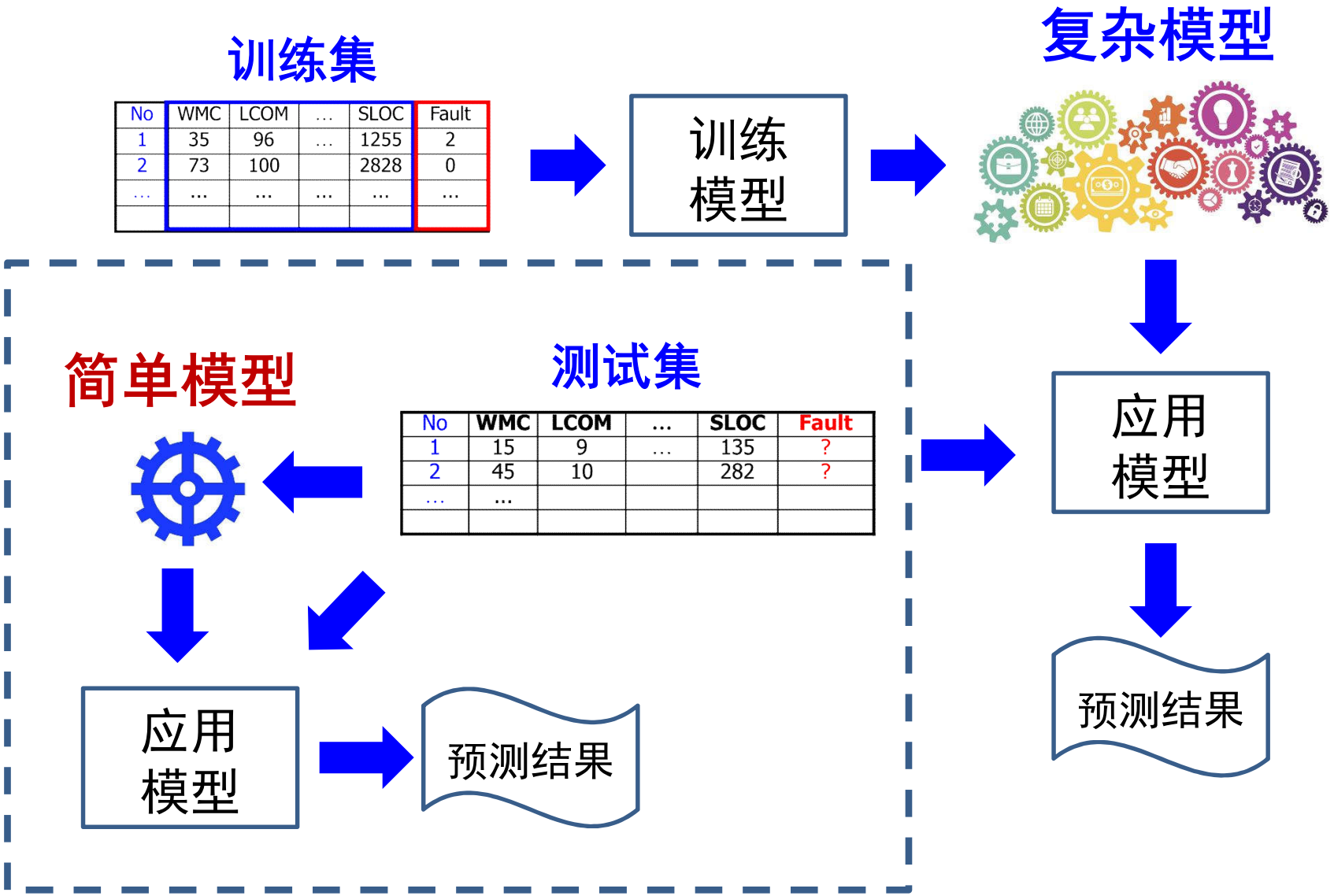


应用模型

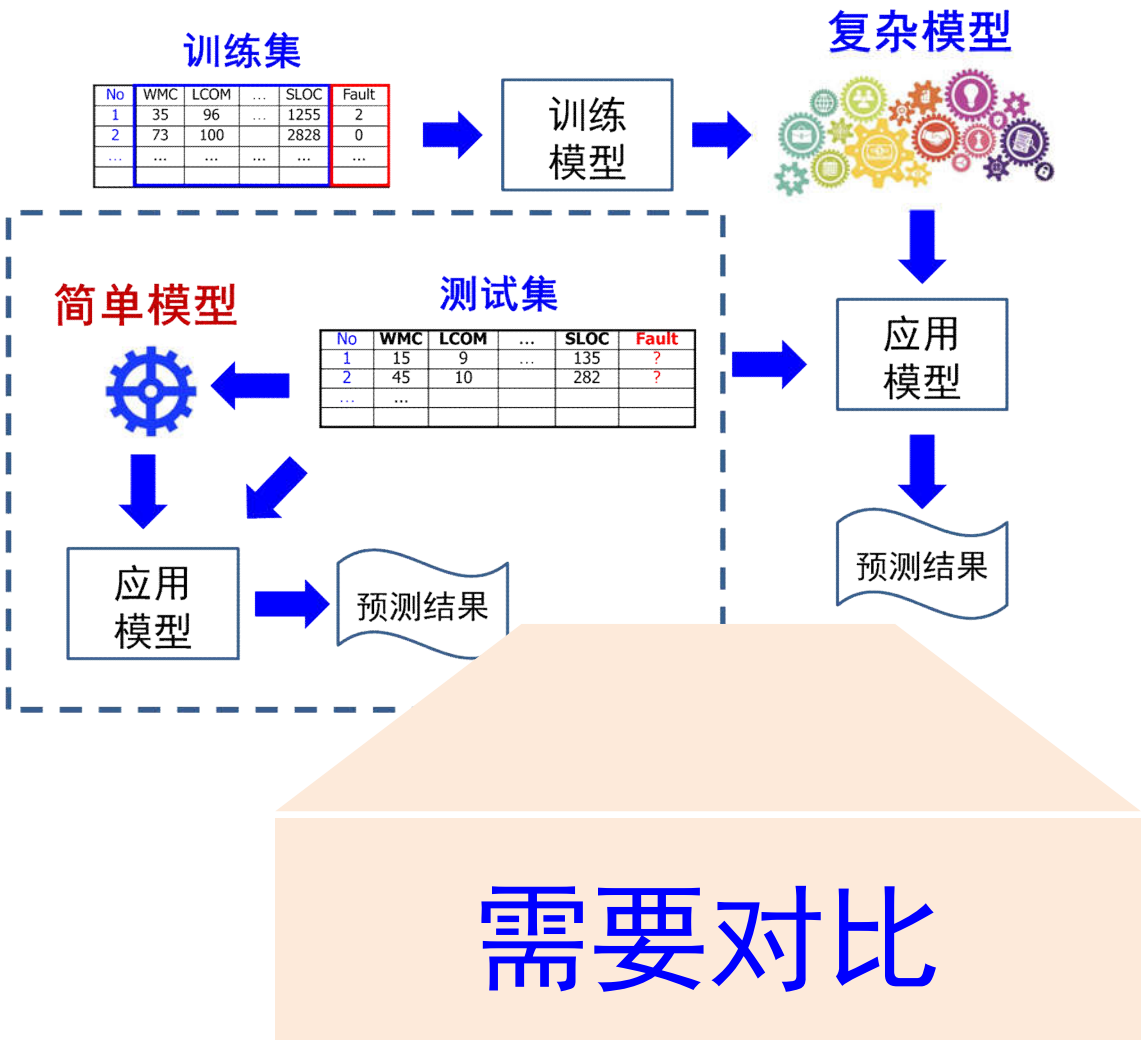


预测结果

简单的预测模型


训练集
No WMC LCOM … SLOC Fault
1 35 96 … 1255 2
2 73 100 2828 0
… … … … … …
训练
模型
复杂模型
简单模型
测试集
No WMC LCOM … SLOC Fault
1 15 9 … 135 ?
2 45 10 282 ?
… …
应用
模型
预测结果
应用
模型
预测结果

简单的预测模型


训练集
No WMC LCOM … SLOC Fault
1 35 96 … 1255 2
2 73 100 2828 0
… … … … … …
训练
模型
复杂模型
简单模型
测试集
No WMC LCOM … SLOC Fault
1 15 9 … 135 ?
2 45 10 282 ?
… …
应用
模型
预测结果
应用
模型
预测结果
需要对比

简单的预测模型

测试集

No	WMC	LCOM	…	SLOC	Fault
1	15	9	…	135	?
2	45	10		282	?
…	…				

分类场景 (评价指标：F1和AUC等)

ManualDown: (规模越大，缺陷数目可能越多)

（1）模块按源代码行SLOC从大到小排序

（2）前50%的模块预测为“有缺陷”，其他为“无缺陷”

（3）审查或者测试“有缺陷”的模块

许多文献表明：ManualDown有较好的分类性能

简单的预测模型

测试集

No	WMC	LCOM	…	SLOC	Fault
1	15	9	…	135	?
2	45	10		282	?
…	…				

排序场景 (评价指标：PofB20和Popt等)

ManualUp: (规模越小，缺陷密度可能越大)

（1）模块按源代码行SLOC从小到大排序

（2）从前往后依次审查或者测试

许多文献表明：ManualUp有较好的排序性能

简单的预测模型

对一个给定的跨项目缺陷预测模型

分类场景

跨项目预测模型应与测试集上的**ManualDown**比较

排序场景

跨项目预测模型应与测试集上的**ManualUp**比较

提 纲

① 跨项目缺陷预测

② 简单的预测模型

③ 实验结果的对比

④ 实验结论与建议

模型对比原则

现有CPDP模型 vs. 简单模型

原则1： 用原始论文使用的测试数据

原则2： 用原始论文使用的性能指标

原则3： 用原始论文中报告的性能值

原则4： 检查差别在统计上是否显著

(Wilcoxon signed-rank test)

检查差别在实用上是否重要

(cliff's delta)

预测结果对比：分类场景

Study	The CPDP model	N	Indicator	Model performance		Difference			Statistical test		Effect size	
				CPDP	ManualDown	mean	median	sd	p-value*	power	cliff's δ	Size
Zhang et al. [122]	MT+	18	F_1	0.448(0.170)	0.484(0.178)	-0.036	-0.021	0.146	0.728	0.237	-0.136	Trivial
			AUC	0.722(0.092)	0.749(0.065)	-0.027	-0.024	0.080	0.525	0.395	-0.148	Small
Stuckman et al. [101]	Metric+CFA	6	F_1	0.294(0.245)	0.269(0.258)	0.025	-0.001	0.052	1.000	0.249	<-0.001	Trivial
Aarti et al. [1]	ANN	19	F_1	0.266(0.130)	0.382(0.134)	-0.116	-0.086	0.177	0.045	0.920	-0.457	Moderate
Herbold et al. [37]	EM	79	F_1	0.419(0.176)	0.464(0.201)	-0.045	-0.022	0.136	0.025	0.947	-0.141	Trivial
			AUC	0.643(0.093)	0.734(0.111)	-0.090	-0.093	0.099	<0.001	1.000	-0.507	Large
Jing et al. [42]	SSTCA+ISDA	8	F2	0.661(0.034)	0.582(0.093)	0.080	0.068	0.068	0.030	0.959	0.563	Large
			AUC	0.796(0.022)	0.760(0.062)	0.036	0.049	0.065	0.525	0.417	0.438	Moderate
Krishna et al. [52]	Bellwether	19	ED	0.386(0.073)	0.343(0.073)	0.043	0.024	0.093	0.228	0.660	0.277	Small
Xia et al. [116]	HYDRA	29	F_1	0.544(0.223)	0.515(0.164)	0.029	0.008	0.114	0.728	0.382	0.062	Trivial
Zhang et al. [123]	Universal	5	AUC	0.741(0.060)	0.725(0.056)	0.016	-0.013	0.051	1.000	0.112	0.120	Trivial
Wang et al. [109]	DBN	22	F_1	0.568(0.151)	0.534(0.131)	0.033	-0.032	0.168	1.000	0.199	0.103	Trivial
Ryu et al. [93]	VCB-SVM	9	AUC	0.768(0.128)	0.788(0.074)	-0.020	0.005	0.066	1.000	0.179	0	Trivial
Ryu et al. [92]	MONBNN	13	Balance	0.621(0.073)	0.629(0.088)	-0.008	-0.014	0.039	0.620	0.133	-0.189	Small
			AUC	0.686(0.105)	0.701(0.133)	-0.015	-0.016	0.062	0.526	0.179	-0.172	Small
Hosseini et al. [39]	GIS	13	F_1	0.484(0.284)	0.486(0.216)	-0.002	-0.024	0.099	1.000	0.051	0.018	Trivial
			G_2	0.543(0.241)	0.526(0.184)	0.017	-0.014	0.085	0.938	0.137	0.065	Trivial
Cheng et al. [17]	CCT-SVM	32	F_1	0.556(0.156)	0.400(0.135)	0.156	0.113	0.197	<0.001	1.000	0.541	Large
			AUC	0.741(0.062)	0.763(0.068)	-0.022	-0.019	0.076	0.369	0.506	-0.179	Small
Kaur et al. [45]	POP	16	AUC	0.762(0.087)	0.747(0.086)	0.015	0.022	0.034	0.341	0.549	0.102	Trivial
Catal et al. [14]	Threshold	6	AUC	0.685(0.055)	0.755(0.064)	-0.070	-0.068	0.037	0.094	0.998	-0.611	Large
He et al. [29]	TDSelector	15	AUC	0.720(0.080)	0.721(0.110)	-0.002	-0.003	0.059	0.938	0.053	-0.102	Trivial
Ryu et al. [94]	HISNN	6	Balance	0.669(0.041)	0.646(0.039)	0.023	0.036	0.035	0.369	0.427	0.389	Moderate
Peters et al. [85]	LACE2	10	G_1	0.583(0.051)	0.662(0.048)	-0.078	-0.066	0.039	0.010	1.000	-0.760	Large
Jing et al. [41]	CCA+	74	F_1	0.592(0.170)	0.418(0.156)	0.174	0.141	0.188	<0.001	1.000	0.553	Large
Nam et al. [79]	HDP	28	AUC	0.711(0.108)	0.713(0.114)	-0.003	-0.001	0.026	0.648	0.102	-0.036	Trivial
Singh et al. [99]	NB	42	AUC	0.701(0.099)	0.758(0.083)	-0.052	-0.030	0.095	0.002	0.954	-0.286	Small
Zhang et al. [126]	Max	10	F_1	0.413(0.115)	0.399(0.136)	0.014	0.038	0.120	0.588	0.069	0.100	Trivial
Cao et al. [13]	TCANN	26	F_1	0.451(0.186)	0.450(0.178)	0.001	<0.001	0.035	1.000	0.054	-0.009	Trivial
Panichella et al. [82]	CODEP(BN)	10	AUC	0.861(0.072)	0.730(0.111)	0.131	0.088	0.105	0.010	0.994	0.820	Large
Mizuno et al. [66]	Text	6	F_1	0.463(0.223)	0.503(0.255)	-0.040	0.002	0.120	1.000	0.145	-0.167	Small
He et al. [27]	IFS(tca)	19	F_1	0.475(0.104)	0.469(0.172)	0.007	0.004	0.164	1.000	0.055	0.058	Trivial
Peters et al. [84]	LACE1(m40)	10	G_1	0.598(0.101)	0.604(0.104)	-0.006	-0.010	0.064	1.000	0.062	-0.060	Trivial
Peters et al. [86]	Peter-filter	21	G_1	0.694(0.265)	0.667(0.172)	0.028	0.061	0.342	0.622	0.072	0.263	Small
Turhan et al. [107]	Mixed	32	Balance	0.623(0.105)	0.638(0.063)	-0.015	0.010	0.110	1.000	0.157	-0.021	Trivial
Nam et al. [80]	TCA+	26	F_1	0.454(0.151)	0.450(0.178)	0.004	-0.013	0.061	1.000	0.066	0.030	Trivial
Ma et al. [61]	TNB	9	F_1	0.376(0.172)	0.370(0.152)	0.006	0.003	0.031	0.938	0.102	0.012	Trivial
			AUC	0.699(0.076)	0.788(0.074)	-0.089	-0.076	0.031	0.018	1.000	-0.630	Large
Uchigaki et al. [108]	Ensemble	20	AUC	0.701(0.058)	0.724(0.064)	-0.024	-0.033	0.019	0.004	1.000	-0.430	Moderate
He et al. [32]	DT	34	F_1	0.627(0.156)	0.490(0.172)	0.137	0.121	0.067	<0.001	1.000	0.420	Moderate
Liu et al. [59]	GP(V-V)	18	NECM	0.741(0.153)	0.730(0.102)	0.011	-0.005	0.069	1.000	0.093	-0.022	Trivial
Khoshgoftaar et al.[48]	MLMD	18	NECM	0.862(0.214)	0.730(0.102)	0.133	0.139	0.167	0.031	0.978	0.380	Moderate

*: Benjamini-Hochberg-corrected p-value.
BN: Bayes Network.

Win/Tie/Loss in the Classification Prediction Scenario

	Statistical significance	Practical importance
Number of comparisons	5/29/8	5/30/7
Number of studies	5/23/7	5/23/7

预测结果对比：分类场景

Study	The CPDP model	N	Indicator	Model performance		Difference			Statistical test		Effect size	
				CPDP	ManualDown	mean	median	sd	p-value*	power	cliff's δ	Size
Panichella et al. [82]	CODEP(BN)	10	AUC	0.861(0.072)	0.730(0.111)	0.131	0.088	0.105	0.004	0.994	0.820	Large
Jing et al. [42]	CODEP	8	F2	0.415(0.071)	0.582(0.093)	-0.167	-0.197	0.068	0.018	1.000	-0.875	Large
			AUC	0.635(0.035)	0.760(0.062)	-0.125	-0.136	0.067	0.012	1.000	-1.000	Large
Xia et al. [116]	CODEP(LR)	29	F_1	0.417(0.116)	0.515(0.164)	-0.098	-0.055	0.129	<0.001	0.998	-0.400	Moderate
Zhang et al. [126]	CODEP(LR)	10	F_1	0.301(0.131)	0.399(0.136)	-0.098	-0.032	0.212	0.261	0.385	-0.340	Moderate
Herbold et al. [35]	CODEP(BN)	86	F_1	0.360(0.169)	0.475(0.204)	-0.115	-0.089	0.174	<0.001	1.000	-0.339	Moderate
			G_1	0.467(0.197)	0.646(0.085)	-0.178	-0.152	0.192	<0.001	1.000	-0.573	Large
			MCC	0.238(0.151)	0.255(0.169)	-0.017	-0.013	0.130	0.268	0.324	-0.082	Trivial
			AUC	0.617(0.080)	0.732(0.113)	-0.115	-0.111	0.093	<0.001	1.000	-0.626	Large

*: Benjamini-Hochberg-corrected p-value.

Study	N	Indicator	Model performance		Difference			Statistical test		Effect size	
			CPDP	ManualDown	mean	median	sd	p-value*	power	cliff's δ	Size
Jing et al. [41]	74	F_1	0.592(0.170)	0.418(0.156)	0.174	0.141	0.188	<0.001	1.000	0.553	Large
Cheng et al. [17]	32	F_1	0.535(0.161)	0.400(0.135)	0.135	0.102	0.201	<0.001	0.996	0.480	Large
		AUC	0.719(0.062)	0.763(0.068)	-0.044	-0.037	0.078	0.009	0.969	-0.346	Moderate
Li et al. [57]	30	AUC	0.641(0.062)	0.749(0.072)	-0.108	-0.109	0.075	<0.001	1.000	-0.739	Large
Li et al. [58]	28	F2	0.482(0.112)	0.517(0.144)	-0.035	-0.058	0.097	0.074	0.622	-0.154	Small
		G_1	0.635(0.051)	0.642(0.079)	-0.007	-0.021	0.073	0.399	0.099	-0.230	Small
		AUC	0.669(0.058)	0.713(0.114)	-0.043	-0.059	0.094	0.009	0.830	-0.421	Moderate

*: Benjamini-Hochberg-corrected p-value.

Study	N	Indicator	Model performance		Difference			Statistical test		Effect size	
			CPDP	ManualDown	mean	median	sd	p-value*	power	cliff's δ	Size
Jing et al. [42]	74	F2	0.661(0.034)	0.582(0.093)	0.080	0.068	0.068	0.008	0.959	0.563	Large
		AUC	0.796(0.022)	0.760(0.062)	0.036	0.049	0.065	0.250	0.417	0.438	Moderate
Li et al. [57]	30	AUC	0.659(0.064)	0.749(0.072)	-0.089	-0.086	0.032	<0.001	1.000	-0.633	Large

*: Benjamini-Hochberg-corrected p-value.

分类场景

ManualDown不比大多数现有的CPDP模型差

预测结果对比：排序场景

Study	the CPDP model	N	Indicator	Model performance		Difference			Statistical test		Effect size	
				CPDP	ManualUp	mean	median	sd	p-value*	power	cliff's δ	Size
Xia et al. [116]	HYDRA	29	PofB20	0.330(0.146)	0.481(0.192)	-0.150	-0.109	0.157	<0.001	1.000	-0.467	Moderate
You et al. [118]	ROCPDP	39	Prec@10	0.382(0.179)	0.621(0.256)	-0.238	-0.200	0.263	<0.001	1.000	-0.563	Large
Canfora et al. [12]	MODEP(LR)	10	AUCEC	0.836(0.129)	0.759(0.158)	0.077	0.085	0.044	0.023	1.000	0.280	Small
Zhang et al. [126]	Bagging(J48)	10	NofB20	40.6(33.964)	37.7(31.457)	2.900	-2.000	15.481	1.000	0.104	-0.030	Trivial
Panichella et al. [82]	CODEP(LR)	10	AUCEC	0.541(0.079)	0.718(0.158)	-0.177	-0.111	0.224	0.046	0.814	-0.740	Large

*: Benjamini-Hochberg-corrected p-value.

	Statistical significance	Practical importance
Number of comparisons	1/1/3	1/1/3
Number of studies	1/1/3	1/1/3

Study	the CPDP model	N	Indicator	Model performance		Difference			Statistical test		Effect size	
				CPDP	ManualUp	mean	median	sd	p-value*	power	cliff's δ	Size
Canfora et al. [12]	MODEP(LR)	10	AUCEC	0.836(0.129)	0.759(0.158)	0.077	0.085	0.044	0.028	1.000	0.280	Small
Xia et al. [116]	MODEP(LR)	29	PofB20	0.191(0.109)	0.481(0.192)	-0.290	-0.295	0.243	<0.001	1.000	-0.772	Large
Chen et al. [16]	MULTI-M	30	ACC	0.730(0.078)	0.747(0.073)	-0.017	-0.007	0.119	0.487	0.148	-0.211	Small
			Popt	0.889(0.045)	0.899(0.043)	-0.009	-0.007	0.070	0.487	0.143	-0.289	Small

*: Benjamini-Hochberg-corrected p-value.

排序场景

ManualUp不比现有的CPDP模型差

Herbold等的CPDP Benchmark

亮点1：实现了2008-2015期间24个CPDP模型

亮点2：实现了4个项目内的baseline模型

亮点3：在86个公共数据集上对比24+4个模型

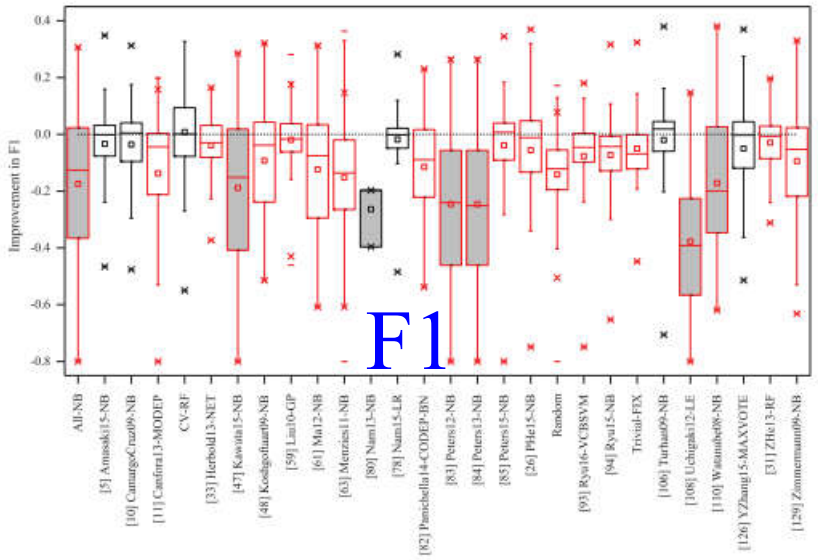
亮点4：采用F1、G1、AUC和MCC性能评价

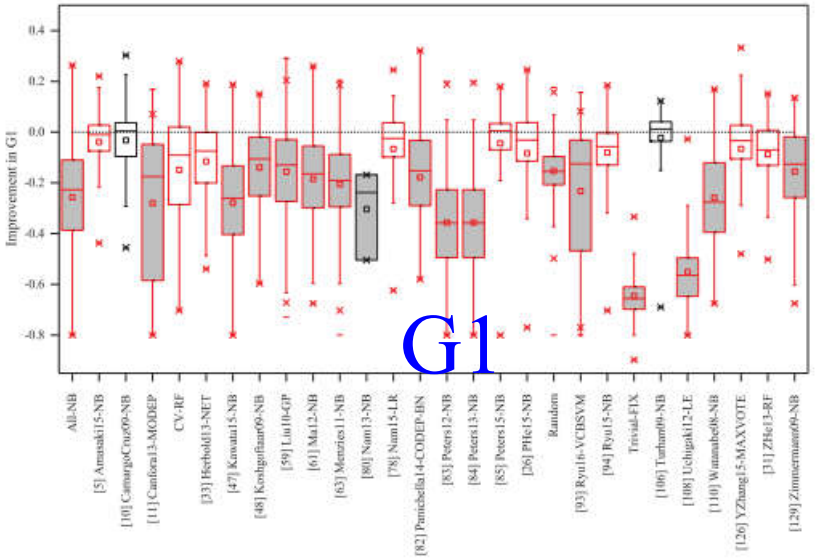
亮点5：源代码和分析结果公开

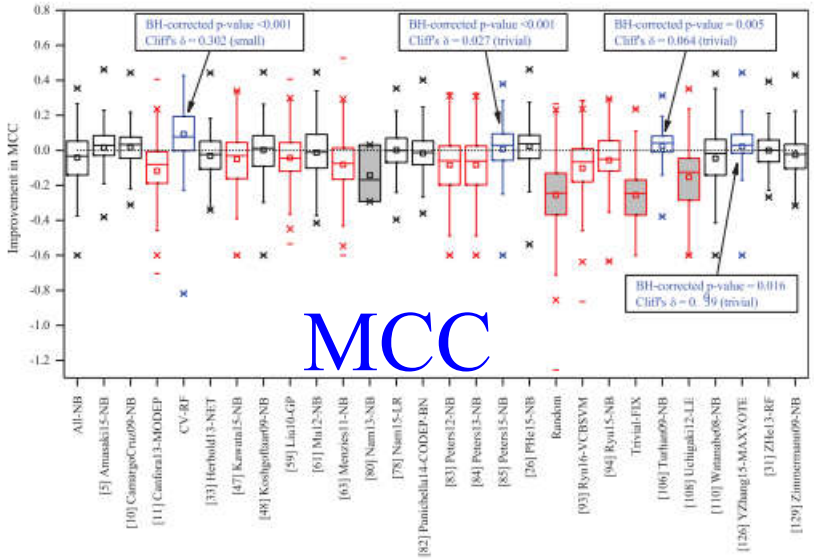
分类场景

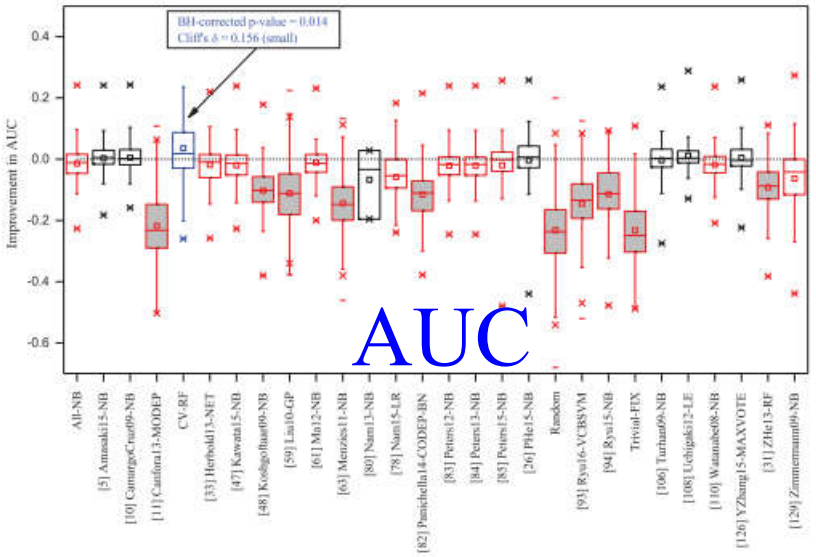
ManualDown vs. 24个CPDP模型

与Herbold等的Benchmark对比

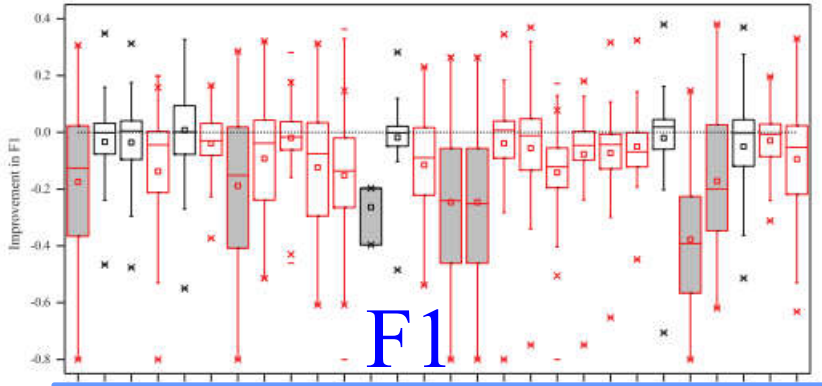


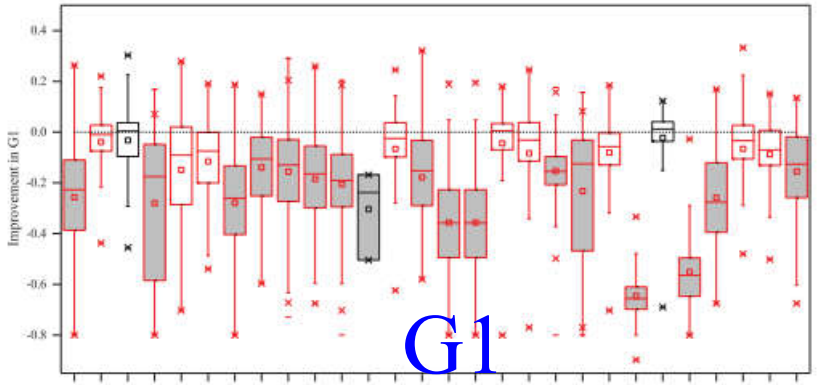




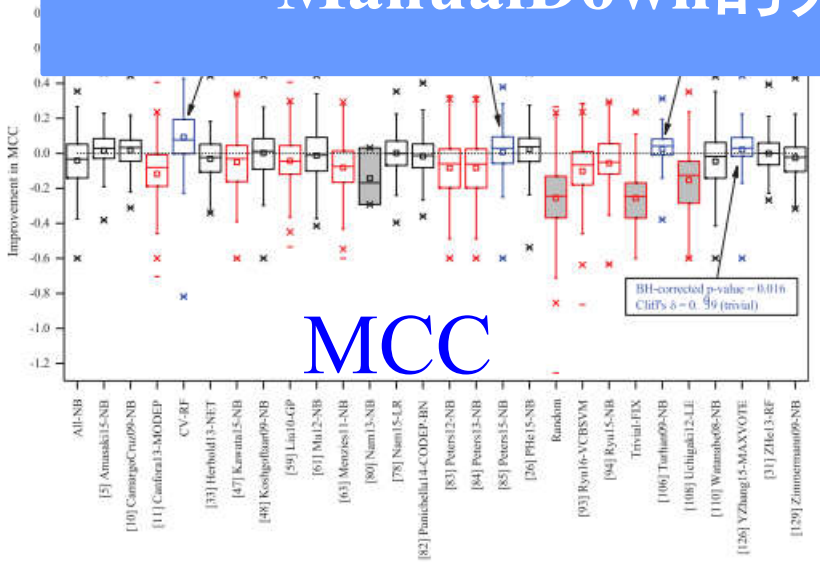


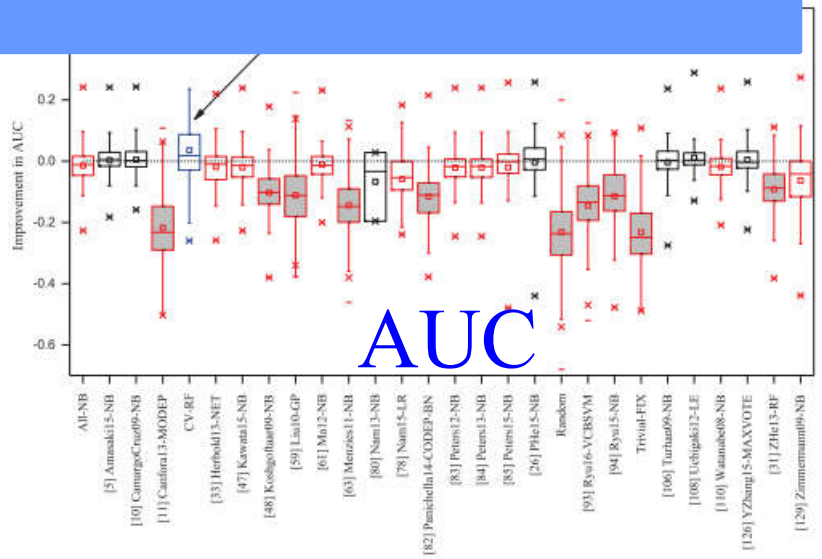
与Herbold等的Benchmark对比





与24个CPDP模型相比，
ManualDown的分类效果相似或者更好





提 纲

① 跨项目缺陷预测

② 简单的预测模型

③ 实验结果的对比

④ 实验结论与建议

实验结论与建议

1. 应该将简单模型作为baseline模型对比
2. 应该全面考虑CPDP的挑战性问题
3. 使用大量数据集和多个性能指标全面评价

谢谢！