

数据驱动测试报告

姓名：蔡挺，周泽龙

学号：2020214022，2020213990

课程：软件测试技术

日期：2020年12月5日

数据驱动测试报告

1 实验内容

1.1 要求

1.2 完成内容

2 现有分类器评估

2.1 AdaBoostClassifier

2.2 CBL0F（无监督）

2.3 DecisionTreeClassifier

2.4 GaussianNB

2.5 GaussianProcessClassifier

2.6 KNeighborsClassifier

2.7 MLPClassifier

2.8 QuadraticDiscriminantAnalysis

2.9 RandomForestClassifier

2.10 SVC

2.11 XGBOD

3 论文算法复现

3.1 ISDA

3.2 BaggingClassifierPU

3.3 JSFS

总结分析

参考文献

1 实验内容

1.1 要求

- 设计机器学习算法，利用NASA数据集和CK数据集进行软件缺陷预测；
- 利用10%，20%及30%的随机样本数据进行训练，使用剩余数据进行测试；
- 统计在两个数据集上的测试结果。

1.2 完成内容

- 调用库，评估现有分类器的效果
 - 如 MLPClassifier、DecisionTreeClassifier、RandomForestClassifier 等等
 - 包含无监督、半监督、监督算法
- 复现相关论文算法
 - ISDA [1]
 - BaggingClassifierPU [2]
 - JSFS [3]

2 现有分类器评估

部分子数据集实验结果较差，在报告中忽略，具体结果可查看 .\MultipleMethods\result 目录。

2.1 AdaBoostClassifier

AdaBoostClassifier (CK/ant1)			
	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.4838709677419355	0.5483870967741935	0.559322033898305
recall	0.3614457831325301	0.4594594594594595	0.5076923076923077
pf	0.13675213675213677	0.1346153846153846	0.14285714285714285
F-measure	0.41379310344827586	0.5	0.532258064516129
accuracy	0.7318611987381703	0.7588652482269503	0.7651821862348178
AUC	0.6123468231901967	0.6624220374220374	0.6824175824175824

AdaBoostClassifier (CK/jedit4)			
	10%	20%	30%
train samples	30	61	91
defective train samples	7	15	22
precision	0.4536082474226804	0.515625	0.4791666666666667
recall	0.6470588235294118	0.55	0.4339622641509434
pf	0.2548076923076923	0.16756756756756758	0.15432098765432098
F-measure	0.5333333333333333	0.5322580645161291	0.4554455445544555

	10%	20%	30%
accuracy	0.7210144927536232	0.763265306122449	0.7441860465116279
AUC	0.6961255656108598	0.6912162162162162	0.6398206382483113

AdaBoostClassifier (CK/lucene2)

	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.6451612903225806	0.6927710843373494	0.76
recall	0.546448087431694	0.7055214723926381	0.6643356643356644
pf	0.4435483870967742	0.4636363636363636	0.3125
F-measure	0.591715976331361	0.6990881458966565	0.7089552238805971
accuracy	0.5504885993485342	0.6373626373626373	0.6736401673640168
AUC	0.55144985016746	0.6209425543781372	0.6759178321678321

AdaBoostClassifier (CK/synapse1)

	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.5476190476190477	0.5492957746478874	0.5223880597014925
recall	0.2948717948717949	0.5652173913043478	0.5737704918032787
pf	0.12418300653594772	0.23529411764705882	0.2689075630252101
F-measure	0.3833333333333333	0.5571428571428572	0.5468749999999999
accuracy	0.6796536796536796	0.697560975609756	0.6777777777777778
AUC	0.5853443941679236	0.6649616368286444	0.6524314643890343

AdaBoostClassifier (CK/xalan2)

	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.5778894472361809	0.5813953488372093	0.6188925081433225

	10%	20%	30%
recall	0.6590257879656161	0.5645161290322581	0.7011070110701108
pf	0.448	0.3783783783783784	0.4006849315068493
F-measure	0.6157965194109772	0.5728314238952538	0.6574394463667821
accuracy	0.6035911602209945	0.5940902021772939	0.6483126110124334
AUC	0.605512893982808	0.5930688753269399	0.6502110397816308

AdaBoostClassifier (NASA/pc4)

	10%	20%	30%
train samples	128	257	386
defective train samples	17	35	53
precision	0.4125874125874126	0.475	0.59
recall	0.36875	0.4014084507042254	0.47580645161290325
pf	0.08408408408408409	0.07094594594594594	0.05276705276705277
F-measure	0.38943894389438943	0.4351145038167939	0.5267857142857142
accuracy	0.8403796376186368	0.8563106796116505	0.8823529411764706
AUC	0.642332957957958	0.6652312523791397	0.7115196994229253

AdaBoostClassifier (NASA/pc5)

	10%	20%	30%
train samples	171	342	513
defective train samples	47	94	141
precision	0.4470588235294118	0.5186567164179104	0.5358851674641149
recall	0.3584905660377358	0.3687002652519894	0.3393939393939394
pf	0.16845878136200718	0.13004032258064516	0.11175115207373272
F-measure	0.3979057591623037	0.43100775193798446	0.4155844155844156
accuracy	0.7012987012987013	0.7319211102994887	0.7370617696160268
AUC	0.5950158923378642	0.619329971335672	0.6138213936601032

2.2 CBLOF (无监督)

CBLOF (CK/ant1)

	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.5882352941176471	0.7021276595744681	0.5769230769230769
recall	0.3614457831325301	0.44594594594594594	0.23076923076923078
pf	0.08974358974358974	0.0673076923076923	0.06043956043956044
F-measure	0.4477611940298508	0.5454545454545454	0.3296703296703296
accuracy	0.7665615141955836	0.8049645390070922	0.7530364372469636
AUC	0.6358510966944703	0.6893191268191269	0.5851648351648352

CBLOF (CK/ivy2)

	10%	20%	30%
train samples	35	70	105
defective train samples	4	8	12
precision	0.33962264150943394	0.2926829268292683	0.3684210526315789
recall	0.5	0.375	0.5
pf	0.12455516014234876	0.116	0.1095890410958904
F-measure	0.4044943820224719	0.32876712328767116	0.4242424242424242
accuracy	0.832807570977918	0.8262411347517731	0.8461538461538461
AUC	0.6877224199288255	0.6295000000000001	0.6952054794520548

CBLOF (CK/jedit4)

	10%	20%	30%
train samples	30	61	91
defective train samples	7	15	22
precision	0.6296296296296297	0.6086956521739131	0.5263157894736842
recall	0.5	0.4666666666666667	0.18867924528301888
pf	0.09615384615384616	0.0972972972972973	0.05555555555555555
F-measure	0.5573770491803278	0.5283018867924527	0.27777777777777773

	10%	20%	30%
accuracy	0.8043478260869565	0.7959183673469388	0.7581395348837209
AUC	0.701923076923077	0.6846846846846847	0.5665618448637316

CBLOF (CK/lucene2)

	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.8243243243243243	0.84	0.8260869565217391
recall	0.3333333333333333	0.25766871165644173	0.13286713286713286
pf	0.10483870967741936	0.07272727272727272	0.041666666666666664
F-measure	0.47470817120622566	0.3943661971830986	0.22891566265060243
accuracy	0.5602605863192183	0.5274725274725275	0.46443514644351463
AUC	0.6142473118279569	0.5924707194645844	0.5456002331002332

CBLOF (CK/synapse1)

	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.5454545454545454	0.5789473684210527	0.6071428571428571
recall	0.6153846153846154	0.3188405797101449	0.2786885245901639
pf	0.26143790849673204	0.11764705882352941	0.09243697478991597
F-measure	0.5783132530120482	0.411214953271028	0.3820224719101123
accuracy	0.696969696969697	0.6926829268292682	0.6944444444444444
AUC	0.6769733534439417	0.6005967604433077	0.5931257749001241

CBLOF (NASA/mc2)

	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.4473684210526316	0.5769230769230769	0.5625

	10%	20%	30%
recall	0.85	0.4166666666666667	0.2903225806451613
pf	0.5753424657534246	0.16923076923076924	0.12280701754385964
F-measure	0.5862068965517242	0.48387096774193544	0.3829787234042554
accuracy	0.5752212389380531	0.6831683168316832	0.6704545454545454
AUC	0.6373287671232877	0.6237179487179487	0.5837577815506508

2.3 DecisionTreeClassifier

DecisionTreeClassifier (CK/ant1)			
	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.5	0.4657534246575342	0.4915254237288136
recall	0.3373493975903614	0.4594594594594595	0.4461538461538462
pf	0.11965811965811966	0.1875	0.16483516483516483
F-measure	0.4028776978417266	0.4625850340136054	0.46774193548387105
accuracy	0.7381703470031545	0.7198581560283688	0.7327935222672065
AUC	0.6088456389661209	0.6359797297297297	0.6406593406593407

DecisionTreeClassifier (CK/jedit4)			
	10%	20%	30%
train samples	30	61	91
defective train samples	7	15	22
precision	0.39344262295081966	0.4186046511627907	0.5357142857142857
recall	0.7058823529411765	0.6	0.5660377358490566
pf	0.3557692307692308	0.2702702702702703	0.16049382716049382
F-measure	0.5052631578947367	0.49315068493150693	0.5504587155963302
accuracy	0.6594202898550725	0.6979591836734694	0.772093023255814
AUC	0.6750565610859729	0.664864864864865	0.7027719543442813

DecisionTreeClassifier (CK/lucene2)

	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.624390243902439	0.7006369426751592	0.676056338028169
recall	0.6994535519125683	0.6748466257668712	0.6713286713286714
pf	0.6209677419354839	0.42727272727272725	0.4791666666666667
F-measure	0.6597938144329897	0.6875	0.6736842105263159
accuracy	0.5700325732899023	0.6336996336996337	0.6108786610878661
AUC	0.5392429049885422	0.623786949247072	0.5960810023310023

DecisionTreeClassifier (CK/synapse1)

	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.5263157894736842	0.4605263157894737	0.4418604651162791
recall	0.2564102564102564	0.5072463768115942	0.6229508196721312
pf	0.11764705882352941	0.3014705882352941	0.40336134453781514
F-measure	0.3448275862068965	0.48275862068965514	0.5170068027210885
accuracy	0.670995670995671	0.6341463414634146	0.6055555555555555
AUC	0.5693815987933634	0.60288789428815	0.609794737567158

DecisionTreeClassifier (CK/xalan2)

	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.56575682382134	0.5876923076923077	0.6212624584717608
recall	0.6532951289398281	0.6161290322580645	0.6900369003690037
pf	0.4666666666666667	0.4024024024024024	0.3904109589041096
F-measure	0.6063829787234042	0.6015748031496062	0.6538461538461537
accuracy	0.5911602209944752	0.6065318818040435	0.6483126110124334

	10%	20%	30%
AUC	0.5933142311365807	0.6068633149278311	0.649812970732447
DecisionTreeClassifier (NASA/mc2)			
	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.5	0.32	0.45454545454545453
recall	0.475	0.22222222222222222	0.3225806451612903
pf	0.2602739726027397	0.26153846153846155	0.21052631578947367
F-measure	0.48717948717948717	0.26229508196721313	0.3773584905660377
accuracy	0.6460176991150443	0.5544554455445545	0.625
AUC	0.6073630136986302	0.4803418803418804	0.5560271646859083
DecisionTreeClassifier (NASA/pc4)			
	10%	20%	30%
train samples	128	257	386
defective train samples	17	35	53
precision	0.452	0.41818181818181815	0.4921875
recall	0.70625	0.4859154929577465	0.5080645161290323
pf	0.13713713713713713	0.10810810810810811	0.08365508365508366
F-measure	0.551219512195122	0.4495114006514658	0.5
accuracy	0.8412424503882657	0.8359223300970874	0.8601553829078802
AUC	0.7845564314314314	0.6889036924248192	0.7122047162369742
DecisionTreeClassifier (NASA/pc5)			
	10%	20%	30%
train samples	171	342	513
defective train samples	47	94	141
precision	0.3974025974025974	0.4447058823529412	0.42771084337349397
recall	0.3608490566037736	0.5013262599469496	0.4303030303030303

	10%	20%	30%
pf	0.2078853046594982	0.23790322580645162	0.21889400921658986
F-measure	0.3782447466007416	0.4713216957605985	0.42900302114803623
accuracy	0.6733766233766234	0.6902848794740687	0.6844741235392321
AUC	0.5764818759721377	0.631711517070249	0.6057045105432202

2.4 GaussianNB

GaussianNB (CK/ant1)

	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.5128205128205128	0.5540540540540541	0.5909090909090909
recall	0.4819277108433735	0.5540540540540541	0.6
pf	0.1623931623931624	0.15865384615384615	0.14835164835164835
F-measure	0.4968944099378882	0.5540540540540541	0.5954198473282443
accuracy	0.7444794952681388	0.7659574468085106	0.7854251012145749
AUC	0.6597672742251055	0.697700103950104	0.7258241758241758

GaussianNB (CK/jedit4)

	10%	20%	30%
train samples	30	61	91
defective train samples	7	15	22
precision	0.22797927461139897	0.5423728813559322	0.4925373134328358
recall	0.6470588235294118	0.5333333333333333	0.6226415094339622
pf	0.7163461538461539	0.14594594594594595	0.20987654320987653
F-measure	0.3371647509578544	0.5378151260504201	0.55
accuracy	0.37318840579710144	0.7755102040816326	0.7488372093023256
AUC	0.46535633484162897	0.6936936936936936	0.7063824831120429

GaussianNB (CK/lucene2)

	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.6632124352331606	0.8	0.7674418604651163
recall	0.6994535519125683	0.4171779141104294	0.46153846153846156
pf	0.5241935483870968	0.15454545454545454	0.20833333333333334
F-measure	0.6808510638297872	0.5483870967741935	0.5764192139737991
accuracy	0.6091205211726385	0.5897435897435898	0.5941422594142259
AUC	0.5876300017627357	0.6313162297824875	0.6266025641025642

GaussianNB (CK/synapse1)			
	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.43902439024390244	0.5340909090909091	0.625
recall	0.23076923076923078	0.6811594202898551	0.6557377049180327
pf	0.1503267973856209	0.3014705882352941	0.20168067226890757
F-measure	0.3025210084033614	0.5987261146496815	0.64
accuracy	0.6406926406926406	0.6926829268292682	0.75
AUC	0.540221216691805	0.6898444160272804	0.7270285163245626

GaussianNB (NASA/mc2)			
	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.5925925925925926	0.8	0.5652173913043478
recall	0.4	0.2222222222222222	0.41935483870967744
pf	0.1506849315068493	0.03076923076923077	0.17543859649122806
F-measure	0.4776119402985075	0.3478260869565218	0.4814814814814815
accuracy	0.6902654867256637	0.7029702970297029	0.6818181818181818
AUC	0.6246575342465753	0.5957264957264957	0.6219581211092248

GaussianNB (NASA/pc4)			
	10%	20%	30%
train samples	128	257	386
defective train samples	17	35	53
precision	0.38926174496644295	0.5546218487394958	0.5777777777777777
recall	0.3625	0.4647887323943662	0.20967741935483872
pf	0.09109109109109109	0.059684684684684686	0.02445302445302445
F-measure	0.37540453074433655	0.5057471264367815	0.30769230769230765
accuracy	0.8334771354616048	0.874757281553398	0.8701442841287459
AUC	0.6357044544544544	0.7025520238548408	0.5926121974509072

2.5 GaussianProcessClassifier

GaussianProcessClassifier (CK/ant1)			
	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.5909090909090909	0.6461538461538462	0.6833333333333333
recall	0.3132530120481928	0.5675675675675675	0.6307692307692307
pf	0.07692307692307693	0.11057692307692307	0.1043956043956044
F-measure	0.40944881889763785	0.6043165467625901	0.6559999999999999
accuracy	0.7634069400630915	0.8049645390070922	0.8259109311740891
AUC	0.6181649675625579	0.7284953222453222	0.7631868131868133

GaussianProcessClassifier (CK/jedit4)			
	10%	20%	30%
train samples	30	61	91
defective train samples	7	15	22
precision	0.5909090909090909	0.5142857142857142	0.5757575757575758
recall	0.38235294117647056	0.3	0.3584905660377358
pf	0.08653846153846154	0.0918918918918919	0.08641975308641975

	10%	20%	30%
F-measure	0.46428571428571425	0.3789473684210526	0.441860465116279
accuracy	0.782608695652174	0.7591836734693878	0.7767441860465116
AUC	0.6479072398190044	0.6040540540540541	0.6360354064756581

GaussianProcessClassifier (CK/lucene2)

	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.5277777777777778	0.5585585585585585	0.7068965517241379
recall	0.3114754098360656	0.3803680981595092	0.5734265734265734
pf	0.4112903225806452	0.44545454545454544	0.3541666666666667
F-measure	0.3917525773195876	0.45255474452554745	0.6332046332046333
accuracy	0.4234527687296417	0.45054945054945056	0.602510460251046
AUC	0.4500925436277102	0.4674567763524819	0.6096299533799533

GaussianProcessClassifier (CK/synapse1)

	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.4857142857142857	0.5245901639344263	0.6086956521739131
recall	0.4358974358974359	0.463768115942029	0.45901639344262296
pf	0.23529411764705882	0.21323529411764705	0.15126050420168066
F-measure	0.45945945945945943	0.49230769230769234	0.5233644859813085
accuracy	0.6536796536796536	0.6780487804878049	0.7166666666666667
AUC	0.6003016591251885	0.625266410912191	0.6538779446204711

GaussianProcessClassifier (CK/xalan2)

	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116

	10%	20%	30%
precision	0.5833333333333334	0.5376344086021505	0.5833333333333334
recall	0.12034383954154727	0.16129032258064516	0.5166051660516605
pf	0.08	0.12912912912912913	0.3424657534246575
F-measure	0.1995249406175772	0.24813895781637718	0.5479452054794521
accuracy	0.5345303867403315	0.5287713841368584	0.5896980461811723
AUC	0.5201719197707737	0.516080596725758	0.5870697063135015

2.6 KNeighborsClassifier

KNeighborsClassifier (CK/ant1)			
	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.5263157894736842	0.6610169491525424	0.5769230769230769
recall	0.3614457831325301	0.527027027027027	0.46153846153846156
pf	0.11538461538461539	0.09615384615384616	0.12087912087912088
F-measure	0.42857142857142855	0.5864661654135338	0.5128205128205129
accuracy	0.7476340694006309	0.8049645390070922	0.7692307692307693
AUC	0.6230305838739574	0.7154365904365904	0.6703296703296703

KNeighborsClassifier (CK/jedit4)			
	10%	20%	30%
train samples	30	61	91
defective train samples	7	15	22
precision	0.546875	0.4166666666666667	0.5757575757575758
recall	0.5147058823529411	0.3333333333333333	0.3584905660377358
pf	0.13942307692307693	0.15135135135135136	0.08641975308641975
F-measure	0.5303030303030303	0.3703703703703704	0.441860465116279
accuracy	0.7753623188405797	0.7224489795918367	0.7767441860465116

	10%	20%	30%
AUC	0.6876414027149321	0.590990990990991	0.6360354064756581

KNeighborsClassifier (CK/lucene2)			
	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.6108108108108108	0.6613756613756614	0.6811594202898551
recall	0.6174863387978142	0.7668711656441718	0.6573426573426573
pf	0.5806451612903226	0.5818181818181818	0.4583333333333333
F-measure	0.6141304347826086	0.7102272727272727	0.6690391459074733
accuracy	0.5374592833876222	0.6263736263736264	0.6108786610878661
AUC	0.5184205887537459	0.592526491912995	0.599504662004662

KNeighborsClassifier (CK/synapse1)			
	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.4626865671641791	0.5263157894736842	0.6166666666666667
recall	0.3974358974358974	0.5797101449275363	0.6065573770491803
pf	0.23529411764705882	0.2647058823529412	0.19327731092436976
F-measure	0.42758620689655175	0.5517241379310345	0.6115702479338844
accuracy	0.6406926406926406	0.6829268292682927	0.7388888888888889
AUC	0.5810708898944191	0.6575021312872975	0.7066400330624054

KNeighborsClassifier (CK/xalan2)			
	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.6209150326797386	0.5955414012738853	0.59
recall	0.5444126074498568	0.603225806451613	0.6531365313653137

	10%	20%	30%
pf	0.30933333333333335	0.3813813813813814	0.4212328767123288
F-measure	0.5801526717557252	0.5993589743589743	0.6199649737302977
accuracy	0.6201657458563536	0.6111975116640747	0.6145648312611013
AUC	0.6175396370582618	0.6109222125351158	0.6159518273264925

KNeighborsClassifier (NASA/mc2)

	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.65	0.47058823529411764	0.6
recall	0.325	0.2222222222222222	0.3870967741935484
pf	0.0958904109589041	0.13846153846153847	0.14035087719298245
F-measure	0.43333333333333335	0.3018867924528302	0.47058823529411764
accuracy	0.6991150442477876	0.6336633663366337	0.6931818181818182
AUC	0.614554794520548	0.541880341880342	0.623372948500283

2.7 MLPClassifier

MLPClassifier (CK/ant1)

	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.5423728813559322	0.546875	0.6410256410256411
recall	0.3855421686746988	0.47297297297297297	0.38461538461538464
pf	0.11538461538461539	0.13942307692307693	0.07692307692307693
F-measure	0.4507042253521127	0.5072463768115941	0.4807692307692308
accuracy	0.7539432176656151	0.7588652482269503	0.7813765182186235
AUC	0.6350787766450416	0.6667749480249481	0.6538461538461539

MLPClassifier (CK/jedit4)			
	10%	20%	30%
train samples	30	61	91
defective train samples	7	15	22
precision	0.30120481927710846	0.5161290322580645	0.48717948717948717
recall	0.36764705882352944	0.5333333333333333	0.3584905660377358
pf	0.27884615384615385	0.16216216216216217	0.12345679012345678
F-measure	0.3311258278145696	0.5245901639344263	0.41304347826086957
accuracy	0.6340579710144928	0.763265306122449	0.7488372093023256
AUC	0.5444004524886878	0.6855855855855856	0.6175168879571394

MLPClassifier (CK/lucene2)			
	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.6048780487804878	0.6848484848484848	0.673469387755102
recall	0.6775956284153005	0.6932515337423313	0.6923076923076923
pf	0.6532258064516129	0.4727272727272727	0.5
F-measure	0.6391752577319587	0.6890243902439025	0.6827586206896552
accuracy	0.5439739413680782	0.6263736263736264	0.6150627615062761
AUC	0.5121849109818438	0.6102621305075293	0.5961538461538461

MLPClassifier (CK/synapse1)			
	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.6530612244897959	0.5303030303030303	0.55
recall	0.41025641025641024	0.5072463768115942	0.5409836065573771
pf	0.11111111111111111	0.22794117647058823	0.226890756302521
F-measure	0.5039370078740157	0.5185185185185185	0.5454545454545455
accuracy	0.7272727272727273	0.6829268292682927	0.6944444444444444

	10%	20%	30%
AUC	0.6495726495726495	0.6396526001705031	0.657046425127428

MLPClassifier (CK/xalan2)			
	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.5935828877005348	0.5975232198142415	0.6186046511627907
recall	0.6361031518624641	0.6225806451612903	0.4907749077490775
pf	0.4053333333333333	0.39039039039039036	0.2808219178082192
F-measure	0.6141078838174274	0.6097946287519748	0.5473251028806585
accuracy	0.6146408839779005	0.6158631415241057	0.6092362344582594
AUC	0.6153849092645655	0.6160951273854499	0.6049764949704292

MLPClassifier (NASA/mc2)			
	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.35398230088495575	0.3564356435643564	0.3522727272727273
recall	1.0	1.0	1.0
pf	1.0	1.0	1.0
F-measure	0.522875816993464	0.5255474452554745	0.5210084033613446
accuracy	0.35398230088495575	0.3564356435643564	0.3522727272727273
AUC	0.5	0.5	0.5

MLPClassifier (NASA/pc5)			
	10%	20%	30%
train samples	171	342	513
defective train samples	47	94	141
precision	0.14695945945945946	0.2788104089219331	0.28761061946902655
recall	0.20518867924528303	0.9946949602122016	0.9848484848484849

	10%	20%	30%
pf	0.4525089605734767	0.9778225806451613	0.9274193548387096
F-measure	0.17125984251968504	0.4355400696864112	0.4452054794520548
accuracy	0.45324675324675323	0.2899926953981008	0.32387312186978295
AUC	0.3763398593359032	0.5084361897835202	0.5287145650048877

2.8 QuadraticDiscriminantAnalysis

QuadraticDiscriminantAnalysis (CK/lucene2)			
	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.5609756097560976	0.6698113207547169	0.6739130434782609
recall	0.12568306010928962	0.8711656441717791	0.6503496503496503
pf	0.14516129032258066	0.6363636363636364	0.46875
F-measure	0.20535714285714288	0.7573333333333332	0.6619217081850534
accuracy	0.4201954397394137	0.6666666666666666	0.602510460251046
AUC	0.49026088489335445	0.6174010039040714	0.5907998251748252

QuadraticDiscriminantAnalysis (CK/xalan2)			
	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.56047197640118	0.601593625498008	0.5977653631284916
recall	0.5444126074498568	0.4870967741935484	0.3948339483394834
pf	0.3973333333333333	0.3003003003003003	0.2465753424657534
F-measure	0.5523255813953489	0.5383244206773619	0.4755555555555556
accuracy	0.574585635359116	0.5972006220839814	0.5808170515097691
AUC	0.5735396370582617	0.5933982369466241	0.5741293029368649

QuadraticDiscriminantAnalysis (NASA/pc5)

	10%	20%	30%
train samples	171	342	513
defective train samples	47	94	141
precision	0.3341584158415842	0.5095057034220533	0.6159420289855072
recall	0.6367924528301887	0.35543766578249336	0.25757575757575757
pf	0.482078853046595	0.13004032258064516	0.06105990783410138
F-measure	0.4383116883116883	0.41874999999999996	0.36324786324786323
accuracy	0.5506493506493506	0.7282688093498905	0.7512520868113522
AUC	0.5773567998917969	0.6126986716009241	0.5982579248708281

2.9 RandomForestClassifier

RandomForestClassifier (CK/ant1)			
	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.5510204081632653	0.6557377049180327	0.6862745098039216
recall	0.3253012048192771	0.5405405405405406	0.5384615384615384
pf	0.09401709401709402	0.10096153846153846	0.08791208791208792
F-measure	0.409090909090909	0.5925925925925926	0.6034482758620688
accuracy	0.7539432176656151	0.8049645390070922	0.8137651821862348
AUC	0.6156420554010915	0.7197895010395011	0.7252747252747253

RandomForestClassifier (CK/jedit4)			
	10%	20%	30%
train samples	30	61	91
defective train samples	7	15	22
precision	0.5348837209302325	0.5873015873015873	0.6923076923076923
recall	0.6764705882352942	0.6166666666666667	0.5094339622641509
pf	0.19230769230769232	0.14054054054054055	0.07407407407407407
F-measure	0.5974025974025974	0.6016260162601625	0.5869565217391305

	10%	20%	30%
accuracy	0.7753623188405797	0.8	0.8232558139534883
AUC	0.7420814479638009	0.738063063063063	0.7176799440950384

RandomForestClassifier (CK/lucene2)			
	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.6697674418604651	0.6813186813186813	0.7446808510638298
recall	0.7868852459016393	0.7607361963190185	0.7342657342657343
pf	0.5725806451612904	0.5272727272727272	0.375
F-measure	0.7236180904522613	0.7188405797101449	0.7394366197183099
accuracy	0.6416938110749185	0.6446886446886447	0.6903765690376569
AUC	0.6071523003701744	0.6167317345231456	0.6796328671328671

RandomForestClassifier (CK/synapse1)			
	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.5818181818181818	0.6567164179104478	0.6086956521739131
recall	0.41025641025641024	0.6376811594202898	0.6885245901639344
pf	0.1503267973856209	0.16911764705882354	0.226890756302521
F-measure	0.48120300751879697	0.6470588235294118	0.6461538461538463
accuracy	0.7012987012987013	0.7658536585365854	0.7444444444444445
AUC	0.6299648064353947	0.7342817561807331	0.7308169169307067

RandomForestClassifier (CK/xalan2)			
	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.5931758530183727	0.654275092936803	0.6486486486486487

	10%	20%	30%
recall	0.6475644699140402	0.567741935483871	0.7084870848708487
pf	0.4133333333333333	0.27927927927927926	0.3561643835616438
F-measure	0.6191780821917807	0.6079447322970639	0.6772486772486772
accuracy	0.6160220994475138	0.6469673405909798	0.6749555950266429
AUC	0.6171155682903534	0.6442313281022958	0.6761613506546024

RandomForestClassifier (NASA/mc2)			
	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.6190476190476191	0.375	0.5555555555555556
recall	0.325	0.25	0.3225806451612903
pf	0.1095890410958904	0.23076923076923078	0.14035087719298245
F-measure	0.42622950819672134	0.3	0.40816326530612246
accuracy	0.6902654867256637	0.5841584158415841	0.6704545454545454
AUC	0.6077054794520548	0.5096153846153846	0.5911148839841539

RandomForestClassifier (NASA/pc5)			
	10%	20%	30%
train samples	171	342	513
defective train samples	47	94	141
precision	0.4611872146118721	0.5955555555555555	0.5977011494252874
recall	0.23820754716981132	0.35543766578249336	0.3151515151515151
pf	0.1057347670250896	0.09173387096774194	0.08064516129032258
F-measure	0.31415241057542764	0.44518272425249167	0.41269841269841273
accuracy	0.7136363636363636	0.7560262965668371	0.7529215358931552
AUC	0.5662363900723608	0.6318518974073757	0.6172531769305962

2.10 SVC

SVC (CK/ant1)

	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.6046511627906976	0.6530612244897959	0.6724137931034483
recall	0.3132530120481928	0.43243243243243246	0.6
pf	0.07264957264957266	0.08173076923076923	0.1043956043956044
F-measure	0.41269841269841273	0.5203252032520326	0.6341463414634146
accuracy	0.7665615141955836	0.7907801418439716	0.8178137651821862
AUC	0.62030171969931	0.6753508316008315	0.7478021978021979

SVC (CK/synapse1)

	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.574468085106383	0.6037735849056604	0.6341463414634146
recall	0.34615384615384615	0.463768115942029	0.4262295081967213
pf	0.13071895424836602	0.15441176470588236	0.12605042016806722
F-measure	0.43200000000000005	0.5245901639344263	0.5098039215686274
accuracy	0.6926406926406926	0.7170731707317073	0.7222222222222222
AUC	0.6077174459527401	0.6546781756180733	0.6500895440143271

SVC (CK/xalan2)

	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.6703910614525139	0.6282722513089005	0.7121212121212122
recall	0.3438395415472779	0.3870967741935484	0.17343173431734318
pf	0.15733333333333333	0.2132132132132132	0.06506849315068493
F-measure	0.4545454545454545	0.4790419161676647	0.2789317507418398

	10%	20%	30%
accuracy	0.6022099447513812	0.5940902021772939	0.5683836589698046
AUC	0.5932531041069723	0.5869417804901677	0.5541816205833292

SVC (NASA/mc2)			
	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.8571428571428571	1.0	0.7333333333333333
recall	0.15	0.02777777777777776	0.3548387096774194
pf	0.0136986301369863	0.0	0.07017543859649122
F-measure	0.2553191489361702	0.05405405405405406	0.47826086956521735
accuracy	0.6902654867256637	0.6534653465346535	0.7272727272727273
AUC	0.5681506849315068	0.5138888888888888	0.6423316355404641

2.11 XGBOD

XGBOD (CK/ant1)			
	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.547945205479452	0.5686274509803921	0.5555555555555556
recall	0.4819277108433735	0.3918918918918919	0.5384615384615384
pf	0.14102564102564102	0.10576923076923077	0.15384615384615385
F-measure	0.5128205128205129	0.464	0.5468749999999999
accuracy	0.7602523659305994	0.7624113475177305	0.7651821862348178
AUC	0.6704510349088663	0.6430613305613305	0.6923076923076922

XGBOD (CK/jedit4)			
	10%	20%	30%
train samples	30	61	91

	10%	20%	30%
defective train samples	7	15	22
precision	0.37037037037037035	0.44285714285714284	0.6341463414634146
recall	0.14705882352941177	0.5166666666666667	0.49056603773584906
pf	0.08173076923076923	0.21081081081081082	0.09259259259259259
F-measure	0.21052631578947367	0.47692307692307695	0.5531914893617021
accuracy	0.7282608695652174	0.7224489795918367	0.8046511627906977
AUC	0.5326640271493213	0.6529279279279279	0.6989867225716283

XGBOD (CK/lucene2)			
	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.6363636363636364	0.7094972067039106	0.695364238410596
recall	0.6502732240437158	0.7791411042944786	0.7342657342657343
pf	0.5483870967741935	0.4727272727272727	0.4791666666666667
F-measure	0.6432432432432432	0.7426900584795323	0.7142857142857142
accuracy	0.5700325732899023	0.6776556776556777	0.6485355648535565
AUC	0.5509430636347612	0.653206915783603	0.6275495337995337

XGBOD (CK/synapse1)			
	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.4766355140186916	0.55	0.5757575757575758
recall	0.6538461538461539	0.4782608695652174	0.6229508196721312
pf	0.3660130718954248	0.19852941176470587	0.23529411764705882
F-measure	0.5513513513513513	0.5116279069767442	0.5984251968503937
accuracy	0.6406926406926406	0.6926829268292682	0.7166666666666667
AUC	0.6439165409753644	0.6398657289002557	0.693828351012536

XGBOD (CK/xalan2)			
	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.6184615384615385	0.597972972972973	0.6085626911314985
recall	0.5759312320916905	0.5709677419354838	0.7343173431734318
pf	0.3306666666666666	0.35735735735735735	0.4383561643835616
F-measure	0.5964391691394659	0.5841584158415841	0.6655518394648829
accuracy	0.6243093922651933	0.6080870917573873	0.6447602131438721
AUC	0.6226322827125119	0.6068051922890632	0.6479805893949351

XGBOD (NASA/mc2)			
	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.4166666666666667	0.5789473684210527	0.4666666666666667
recall	0.25	0.3055555555555556	0.6774193548387096
pf	0.1917808219178082	0.12307692307692308	0.42105263157894735
F-measure	0.3125	0.4000000000000001	0.5526315789473684
accuracy	0.6106194690265486	0.6732673267326733	0.6136363636363636
AUC	0.5291095890410958	0.5912393162393162	0.6281833616298811

XGBOD (NASA/pc4)			
	10%	20%	30%
train samples	128	257	386
defective train samples	17	35	53
precision	0.5740740740740741	0.44285714285714284	0.6753246753246753
recall	0.3875	0.21830985915492956	0.41935483870967744
pf	0.04604604604604605	0.04391891891891892	0.032175032175032175
F-measure	0.4626865671641791	0.29245283018867924	0.5174129353233831
accuracy	0.8757549611734253	0.8543689320388349	0.8923418423973363

	10%	20%	30%
AUC	0.6707269769769769	0.5871954701180052	0.6935899032673227

XGBOD (NASA/pc5)			
	10%	20%	30%
train samples	171	342	513
defective train samples	47	94	141
precision	0.4878048780487805	0.5741444866920152	0.59
recall	0.2830188679245283	0.4005305039787798	0.3575757575757576
pf	0.11290322580645161	0.11290322580645161	0.0944700460829493
F-measure	0.3582089552238806	0.47187500000000004	0.44528301886792454
accuracy	0.7207792207792207	0.7531044558071585	0.7545909849749582
AUC	0.5850578210590383	0.6438136390861641	0.6315528557464042

3 论文算法复现

3.1 ISDA

一种改进的子类判别分析算法（Improved Subclass Discriminant Analysis），主要用于处理类不平衡问题。同时，通过 SSTCA 进行特征迁移，可实现跨项目缺陷预测。[1]

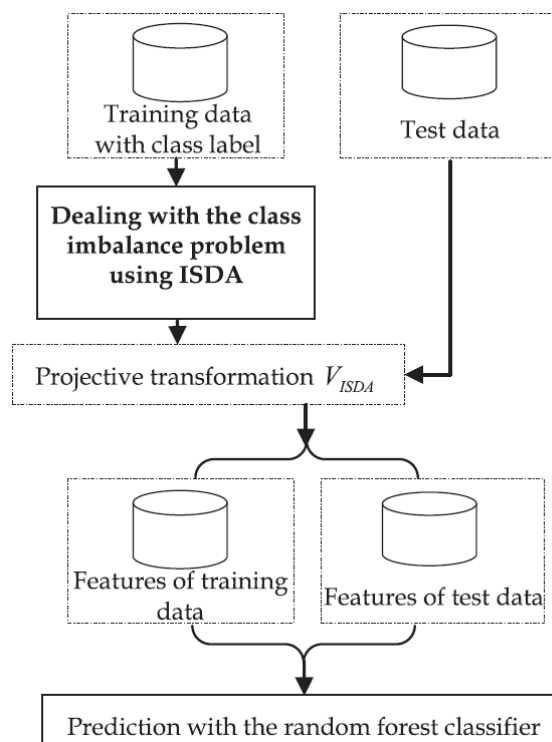


TABLE 3
ISDA-Based Within-Project Prediction

Input:	$X = [X_1, X_2]$ and y .
Output:	Class label of y .
Step 1.	Determine the optimal number of subclass H_{1opt} for the defective class and the optimal number H_{2opt} for the defect-free class by using the improved LOOT criterion in Formula (4).
Step 2.	Separately divide X_1 and X_2 into H_{1opt} and H_{2opt} subclasses by using the NNC algorithm [42].
Step 3.	For the obtained $H_{opt} = H_{1opt} + H_{2opt}$ subclasses, calculate Σ_B and Σ_X by using Formulae (2) and (3), respectively.
Step 4.	Obtain the projective transformation V by using Formula (1), which consists of eigen-vectors corresponding to the nonzero eigen-values of $\Sigma_X^{-1}\Sigma_B$.
Step 5.	Obtain the projected features of the training data and target instance by using $X^f = V^T X$ and $y^f = V^T y$.
Step 6.	Use the random forest classifier to classify y^f .

TABLE 4
SSTCA+ISDA for Cross-Project Prediction

Input:	X_S and X_T .
Output:	Class labels of instances in X_T .
Step 1.	Input X_S and X_T into the SSTCA method and achieve the transferred source and target project data X'_S and X'_T .
Step 2.	Regard X'_S as the training data and learn ISDA projective transformation V .
Step 3.	Obtain the projected features of X'_S and X'_T by using $X^f_S = V^T X'_S$ and $X^f_T = V^T X'_T$.
Step 4.	Use the random forest classifier to classify each instance in X^f_T .

ISDA (CK/ant1)			
	10%	20%	30%
train samples	34	69	104
defective train samples	9	18	27
precision	0.4666666666666667	0.47368421052631576	0.543859649122807
recall	0.42168674698795183	0.4864864864864865	0.47692307692307695
pf	0.17094017094017094	0.19230769230769232	0.14285714285714285
F-measure	0.4430379746835443	0.47999999999999999	0.5081967213114754
accuracy	0.722397476340694	0.723404255319149	0.757085020242915
AUC	0.6253732880238904	0.6470893970893971	0.6670329670329671

ISDA (CK/lucene2)			
	10%	20%	30%
train samples	33	67	101
defective train samples	20	40	60
precision	0.56	0.6162162162162163	0.6506849315068494
recall	0.5355191256830601	0.6993865030674846	0.6643356643356644
pf	0.6209677419354839	0.6454545454545455	0.53125
F-measure	0.547486033519553	0.6551724137931035	0.6574394463667821
accuracy	0.4723127035830619	0.5604395604395604	0.5857740585774058
AUC	0.4572756918737882	0.5269659788064696	0.5665428321678322

ISDA (CK/synapse1)			
	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.5164835164835165	0.46153846153846156	0.5
recall	0.6025641025641025	0.5217391304347826	0.5901639344262295
pf	0.2875816993464052	0.3088235294117647	0.3025210084033613
F-measure	0.5562130177514794	0.4897959183673469	0.5413533834586466
accuracy	0.6753246753246753	0.6341463414634146	0.6611111111111111
AUC	0.6574912016088487	0.606457800511509	0.6438214630114341

ISDA (CK/xalan2)			
	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.5866666666666667	0.5539568345323741	0.5475285171102662
recall	0.6303724928366762	0.4967741935483871	0.5313653136531366
pf	0.4133333333333333	0.37237237237237236	0.4075342465753425
F-measure	0.6077348066298343	0.5238095238095238	0.5393258426966292
accuracy	0.6077348066298343	0.5645412130637636	0.5630550621669627

	10%	20%	30%
AUC	0.6085195797516715	0.5622009105880074	0.561915533538897

ISDA (NASA/mc2)			
	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.5909090909090909	0.5333333333333333	0.4827586206896552
recall	0.325	0.4444444444444444	0.45161290322580644
pf	0.1232876712328767	0.2153846153846154	0.2631578947368421
F-measure	0.41935483870967744	0.4848484848484848	0.4666666666666667
accuracy	0.6814159292035398	0.6633663366336634	0.6363636363636364
AUC	0.6008561643835616	0.6145299145299146	0.5942275042444822

ISDA (NASA/pc5)			
	10%	20%	30%
train samples	171	342	513
defective train samples	47	94	141
precision	0.33729216152019004	0.4022346368715084	0.32558139534883723
recall	0.33490566037735847	0.3819628647214854	0.296969696969697
pf	0.25	0.2157258064516129	0.23387096774193547
F-measure	0.336094674556213	0.39183673469387753	0.31061806656101426
accuracy	0.6357142857142857	0.6734842951059167	0.6368948247078464
AUC	0.5424528301886792	0.5831185291349363	0.5315493646138807

3.2 BaggingClassifierPU

一种半监督分类器，利用正样本和无标签样本。[2]

BaggingClassifierPU (CK/ant1)			
	10%	20%	30%
train samples	34	69	104

	10%	20%	30%
defective train samples	9	18	27
precision	0.5333333333333333	0.4915254237288136	0.4576271186440678
recall	0.5783132530120482	0.7837837837837838	0.8307692307692308
pf	0.1794871794871795	0.28846153846153844	0.3516483516483517
F-measure	0.5549132947976878	0.6041666666666666	0.5901639344262295
accuracy	0.7570977917981072	0.7304964539007093	0.6963562753036437
AUC	0.6994130367624344	0.7476611226611226	0.7395604395604396

BaggingClassifierPU (CK/ivy2)

	10%	20%	30%
train samples	35	70	105
defective train samples	4	8	12
precision	0.2753623188405797	0.24210526315789474	0.30158730158730157
recall	0.5277777777777778	0.71875	0.6785714285714286
pf	0.17793594306049823	0.288	0.2009132420091324
F-measure	0.3619047619047619	0.36220472440944884	0.41758241758241754
accuracy	0.7886435331230284	0.7127659574468085	0.7854251012145749
AUC	0.6749209173586397	0.715375	0.7388290932811481

BaggingClassifierPU (CK/jedit4)

	10%	20%	30%
train samples	30	61	91
defective train samples	7	15	22
precision	0.3240223463687151	0.36363636363636365	0.4074074074074074
recall	0.8529411764705882	0.8	0.6226415094339622
pf	0.5817307692307693	0.4540540540540541	0.2962962962962963
F-measure	0.46963562753036436	0.5000000000000001	0.49253731343283585
accuracy	0.5253623188405797	0.6081632653061224	0.6837209302325581
AUC	0.6356052036199095	0.672972972972973	0.6631726065688329

BaggingClassifierPU (CK/synapse1)

	10%	20%	30%
train samples	25	51	76
defective train samples	8	17	25
precision	0.48863636363636365	0.4948453608247423	0.52
recall	0.5512820512820513	0.6956521739130435	0.8524590163934426
pf	0.29411764705882354	0.3602941176470588	0.40336134453781514
F-measure	0.5180722891566266	0.5783132530120482	0.6459627329192545
accuracy	0.6536796536796536	0.6585365853658537	0.6833333333333333
AUC	0.6285822021116139	0.6676790281329923	0.7245488359278137

BaggingClassifierPU (CK/xalan2)

	10%	20%	30%
train samples	79	160	240
defective train samples	38	77	116
precision	0.5528455284552846	0.5956873315363881	0.5988857938718662
recall	0.7793696275071633	0.7129032258064516	0.7933579335793358
pf	0.5866666666666667	0.45045045045045046	0.4931506849315068
F-measure	0.6468489892984542	0.6490455212922174	0.6825396825396826
accuracy	0.5897790055248618	0.6283048211508554	0.6447602131438721
AUC	0.5963514804202483	0.6312263876780005	0.6501036243239146

BaggingClassifierPU (NASA/mc2)

	10%	20%	30%
train samples	12	24	37
defective train samples	4	8	13
precision	0.4528301886792453	0.410958904109589	0.5454545454545454
recall	0.6	0.8333333333333334	0.7741935483870968
pf	0.3972602739726027	0.6615384615384615	0.3508771929824561
F-measure	0.5161290322580645	0.5504587155963302	0.64
accuracy	0.6017699115044248	0.5148514851485149	0.6931818181818182

	10%	20%	30%
AUC	0.6013698630136987	0.585897435897436	0.7116581777023203

BaggingClassifierPU (NASA/pc3)			
	10%	20%	30%
train samples	107	214	322
defective train samples	13	26	40
precision	0.2837370242214533	0.24456521739130435	0.25075528700906347
recall	0.6776859504132231	0.8333333333333334	0.8829787234042553
pf	0.24381625441696114	0.36821192052980134	0.3751891074130106
F-measure	0.4	0.3781512605042017	0.39058823529411774
accuracy	0.7463917525773196	0.657010428736964	0.6569536423841059
AUC	0.716934847998131	0.7325607064017661	0.7538948079956224

BaggingClassifierPU (NASA/pc4)			
	10%	20%	30%
train samples	128	257	386
defective train samples	17	35	53
precision	0.3347547974413646	0.35543766578249336	0.4166666666666667
recall	0.98125	0.9436619718309859	0.9274193548387096
pf	0.3123123123123123	0.27364864864864863	0.2072072072072072
F-measure	0.4992050874403816	0.5163776493256261	0.5750000000000001
accuracy	0.728213977566868	0.7563106796116504	0.8113207547169812
AUC	0.8344688438438438	0.8350066615911685	0.8601060738157512

BaggingClassifierPU (NASA/pc5)			
	10%	20%	30%
train samples	171	342	513
defective train samples	47	94	141
precision	0.4244604316546763	0.42199108469539376	0.4066115702479339
recall	0.6957547169811321	0.753315649867374	0.7454545454545455

	10%	20%	30%
pf	0.35842293906810035	0.39213709677419356	0.41359447004608296
F-measure	0.5272564789991063	0.540952380952381	0.5262032085561498
accuracy	0.6564935064935065	0.647918188458729	0.6302170283806344
AUC	0.6686658889565159	0.6805892765465902	0.6659300377042313

3.3 JSFS

一种联合贝叶斯半监督特征选择和分类算法（JSFS），该算法采用贝叶斯方法自动选择相关特征并同时学习分类器。 [3]

数据预处理：为解决数据量小，数据分布不均衡的问题, 使用SMOTE算法生成新的样本.

SMOTE算法流程：

- 1、采样KNN算法，计算出每个少数类样本的K个近邻;
- 2、从K个近邻中随机挑选N个样本进行随机线性插值;
- 3、构造新的少数类样本;
- 4、将新样本与原数据合成，产生新的训练集;

扩充后的数据集结果：

类别	比例	训练数据			测试数据			类别	比例	训练数据			测试数据		
		总数	正例	负例	总数	正例	负例			总数	正例	负例	总数	正例	负例
ant1	10%	50	25	25	317	83	234	cm1	10%	56	28	28	295	38	257
	20%	102	51	51	282	74	208		20%	114	57	57	262	34	228
	30%	154	77	77	247	65	182		30%	170	85	85	230	30	200
ivy2	10%	62	31	31	317	36	281	kc3	10%	30	15	15	176	33	143
	20%	124	62	62	282	32	250		20%	62	31	31	156	29	127
	30%	186	93	93	247	28	219		30%	94	47	47	137	26	111
jedit4	10%	46	23	23	276	68	208	mc2	10%	16	8	8	113	40	73
	20%	92	46	46	245	60	185		20%	32	16	16	101	36	65
	30%	138	69	69	215	53	162		30%	48	24	24	88	31	57
lucene2	10%	40	20	20	307	183	124	mw1	10%	44	22	22	229	25	204
	20%	80	40	40	273	163	110		20%	90	45	45	203	22	181
	30%	120	60	60	239	143	96		30%	134	67	67	178	19	159
synapse1	10%	34	17	17	231	78	153	pc1	10%	128	64	64	635	55	580
	20%	68	34	34	205	69	136		20%	256	128	128	565	49	516
	30%	102	51	51	180	61	119		30%	386	193	193	494	43	451
velocity1	10%	10	5	5	93	60	33	pc3	10%	188	94	94	970	121	849
	20%	30	15	15	83	50	33		20%	376	188	188	863	108	755
	30%	38	19	19	73	46	27		30%	564	282	282	755	94	661
xalan2	10%	82	41	41	724	349	375	pc4	10%	222	111	111	1159	160	999
	20%	166	83	83	643	310	333		20%	444	222	222	1030	142	888
	30%	248	124	124	563	271	292		30%	666	333	333	901	124	777
								pc5	10%	248	124	124	1540	424	1116
									20%	496	248	248	1369	377	992
									30%	744	372	372	1198	330	868

JSFS算法流程：

Algorithm 1 The proposed JSFS algorithm	
1: Input:	Training data $\mathbf{X} \in \mathbb{R}^{n \times d}$, parameters γ and μ .
2: Output:	The selected feature indexes and their corresponding weight vector \mathbf{w} for the linear classifier.
3:	Initialize w_i , λ_j , α_i , and c_j for $i = 1, \dots, d$ and $j = l + 1, \dots, n$.
4:	Construct the affinity matrix \mathbf{S} and graph Laplacian \mathbf{L} .
5:	Obtain the pseudo label vector $\tilde{\mathbf{y}}_u$ via label propagation.
6:	While $\max_i w_i^{\text{new}} - w_i^{\text{old}} > 10^{-3}$ do
7:	If $\ \mathbf{g}_w\ /d < 10^{-3}$ then
8:	Fix $\boldsymbol{\lambda}$, compute \mathbf{g}_w and \mathbf{H}_w by Eqs. (10) and (11), and update $\mathbf{w} \leftarrow \mathbf{w} - \mathbf{H}_w^{-1} \mathbf{g}_w$;
9:	end if
10:	Remove the i -th feature if $ w_i < 10^{-3}$;
11:	If $\ \mathbf{g}_\lambda\ /u < 10^{-3}$ then
12:	Fix \mathbf{w} , compute \mathbf{g}_λ and \mathbf{H}_λ by Eqs. (13) and (14), and update $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} - \mathbf{H}_\lambda^{-1} \mathbf{g}_\lambda$;
13:	end if
14:	Remove the j -th unlabeled sample if $ \lambda_j < 10^{-3}$;
15:	Update $\boldsymbol{\alpha}$ and \mathbf{c} using Eqs. (18) and (20);
16:	end while

JSFS与每个数据集的超参等配置见文件config.json

由于JSFS本身是半监督的，为了更好的效果，基本上保留了所有CK或NASA的数据标签

JSFS的情况与问题：

- 1、论文实验中部分超参缺少说明；
- 2、算法收敛条件固定，使得需要将x的特征值归一化到0-0.001等很小的值才能达到收敛条件；
- 3、算法要么不收敛，要么在前五次迭代即收敛，与数据集相关；
- 4、求出的y的值往往值域、量级或精确度差别较大，缺少二分类的可解释性说明，效果很受对y值二分类阈值划分的影响，而论文未说明；
- 5、算法过程中包含其他模型的使用，这些模型相关的选取与参数配置未说明，且没做消融实验等；

JSFS实验结果：

JSFS (CK/ant1)			
	10%	20%	30%
precision	0.5625	0.5434782608695652	0.4956521739130435
recall	0.43373493975903615	0.33783783783783783	0.8769230769230769
pf	0.11965811965811966	0.10096153846153846	0.31868131868131866
F-measure	0.4897959183673469	0.41666666666666663	0.6333333333333333
accuracy	0.7634069400630915	0.75177304964539	0.7327935222672065
AUC	0.6570384100504583	0.6184381496881496	0.7791208791208791

JSFS (CK/ivy2)			
	10%	20%	30%
precision	0.6666666666666666	0.25	0.21686746987951808
recall	0.1111111111111111	0.5625	0.6428571428571429

	10%	20%	30%
pf	0.0071174377224199285	0.216	0.2968036529680365
F-measure	0.1904761904761905	0.34615384615384615	0.32432432432432434
accuracy	0.8927444794952681	0.7588652482269503	0.6963562753036437
AUC	0.5519968366943456	0.67325	0.6730267449445533

JSFS (CK/jedit4)

	10%	20%	30%
precision	0.717948717948718	0.373015873015873	0.37815126050420167
recall	0.4117647058823529	0.7833333333333333	0.8490566037735849
pf	0.052884615384615384	0.42702702702702705	0.4567901234567901
F-measure	0.5233644859813085	0.5053763440860215	0.5232558139534883
accuracy	0.8152173913043478	0.6244897959183674	0.6186046511627907
AUC	0.6794400452488687	0.678153153153153	0.6961332401583974

JSFS (CK/lucene2)

	10%	20%	30%
precision	0.7687074829931972	0.7019867549668874	0.7007874015748031
recall	0.6174863387978142	0.6503067484662577	0.6223776223776224
pf	0.27419354838709675	0.4090909090909091	0.3958333333333333
F-measure	0.6848484848484848	0.6751592356687899	0.6592592592592593
accuracy	0.6612377850162866	0.6263736263736264	0.6150627615062761
AUC	0.6716463952053587	0.6206079196876743	0.6132721445221446

JSFS (CK/synapse1)

	10%	20%	30%
precision	0.4632352941176471	0.43283582089552236	0.4330708661417323
recall	0.8076923076923077	0.8405797101449275	0.9016393442622951
pf	0.477124183006536	0.5588235294117647	0.6050420168067226
F-measure	0.588785046728972	0.5714285714285715	0.5851063829787234
accuracy	0.6190476190476191	0.5756097560975609	0.5666666666666667

	10%	20%	30%
AUC	0.6652840623428858	0.6408780903665814	0.6482986637277862
JSFS (CK/velocity1)			
	10%	20%	30%
precision	0.7627118644067796	0.85	0.7058823529411765
recall	0.75	0.34	0.782608695652174
pf	0.42424242424242425	0.09090909090909091	0.5555555555555556
F-measure	0.7563025210084034	0.4857142857142858	0.7422680412371134
accuracy	0.6881720430107527	0.5662650602409639	0.6575342465753424
AUC	0.6628787878787877	0.6245454545454546	0.6135265700483092
JSFS (CK/xalan2)			
	10%	20%	30%
precision	0.5585585585585585	0.515625	0.48491879350348027
recall	0.3553008595988539	0.6387096774193548	0.7712177121771218
pf	0.2613333333333333	0.5585585585585585	0.7602739726027398
F-measure	0.4343257443082311	0.5706051873198847	0.5954415954415955
accuracy	0.5538674033149171	0.536547433903577	0.4955595026642984
AUC	0.5469837631327603	0.5400755594303982	0.5054718697871909
JSFS (NASA/cm1)			
	10%	20%	30%
precision	0.39361702127659576	0.6829268292682927	0.4520547945205479
recall	1.0	0.875	0.9705882352941176
pf	0.22093023255813954	0.05652173913043478	0.20408163265306123
F-measure	0.5648854961832062	0.767123287671233	0.616822429906542
accuracy	0.8067796610169492	0.9351145038167938	0.8217391304347826
AUC	0.8895348837209303	0.9092391304347827	0.8832533013205283
JSFS (NASA/kc3)			

	10%	20%	30%
precision	1.0	0.4146341463414634	0.5217391304347826
recall	0.29411764705882354	0.9444444444444444	1.0
pf	0.0	0.17391304347826086	0.088
F-measure	0.45454545454545453	0.576271186440678	0.6857142857142856
accuracy	0.9318181818181818	0.8397435897435898	0.9197080291970803
AUC	0.6470588235294118	0.8852657004830918	0.9560000000000001

JSFS (NASA/mc2)			
	10%	20%	30%
precision	0.6666666666666666	0.3783783783783784	0.3611111111111111
recall	0.4	0.7368421052631579	0.9285714285714286
pf	0.043010752688172046	0.2804878048780488	0.3108108108108108
F-measure	0.5	0.5	0.52
accuracy	0.8584070796460177	0.7227722772277227	0.7272727272727273
AUC	0.678494623655914	0.7281771501925546	0.8088803088803089

JSFS (NASA/mw1)			
	10%	20%	30%
precision	0.45121951219512196	0.49295774647887325	0.4262295081967213
recall	1.0	1.0	1.0
pf	0.234375	0.21428571428571427	0.23026315789473684
F-measure	0.6218487394957983	0.660377358490566	0.5977011494252873
accuracy	0.8034934497816594	0.8226600985221675	0.8033707865168539
AUC	0.8828125	0.8928571428571428	0.8848684210526316

JSFS (NASA/pc1)			
	10%	20%	30%
precision	0.2018348623853211	0.3978494623655914	0.7666666666666667
recall	1.0	1.0	0.7931034482758621
pf	0.29441624365482233	0.10606060606060606	0.015053763440860216

	10%	20%	30%
F-measure	0.33587786259541985	0.5692307692307692	0.7796610169491527
accuracy	0.7259842519685039	0.9008849557522124	0.9736842105263158
AUC	0.8527918781725888	0.9469696969696969	0.889024842417501

JSFS (NASA/pc3)

	10%	20%	30%
precision	0.6619718309859155	0.352112676056338	0.8222222222222222
recall	0.8703703703703703	0.9615384615384616	0.7872340425531915
pf	0.026200873362445413	0.11344019728729964	0.011299435028248588
F-measure	0.752	0.5154639175257731	0.8043478260869565
accuracy	0.9680412371134021	0.8910776361529548	0.976158940397351
AUC	0.9220847485039624	0.924049132125581	0.8879673037624715

JSFS (NASA/pc4)

	10%	20%	30%
precision	0.5925925925925926	0.25806451612903225	0.3142857142857143
recall	0.6956521739130435	0.8421052631578947	0.7333333333333333
pf	0.009683098591549295	0.04549950544015826	0.02708803611738149
F-measure	0.6399999999999999	0.3950617283950617	0.44
accuracy	0.9844693701466781	0.9524271844660194	0.9689234184239733
AUC	0.8429845376607471	0.8983028788588683	0.8531226486079759

JSFS (NASA/pc5)

	10%	20%	30%
precision	1.0	0.996742671009772	1.0
recall	0.6516257465162575	0.6866118175018698	0.5501285347043702
pf	0.0	0.09375	0.0
F-measure	0.7890719164323022	0.8131089459698848	0.7097844112769487
accuracy	0.6590909090909091	0.6917457998539079	0.5617696160267112
AUC	0.8258128732581287	0.7964309087509349	0.775064267352185

总结分析

对于现有分类器的评估，还是随机森林大法好，F值最高能达到0.7以上。其它方法各有优缺点，但统一的情况都会出现部分测试数据集的结果很不理想。对于论文算法的复现，ISDA 算法主要处理类不平衡问题，可能由于本次测试数据集规模较小，表现并不佳。BaggingClassifierPU 和 JSFS 在 CK 和 NASA 上的表现都不错，JSFS 甚至能达到 0.8 以上的F值。

参考文献

- [1] Jing, Xiao-Yuan, et al. "An improved SDA based defect prediction framework for both within-project and cross-project class-imbalance problems." *IEEE Transactions on Software Engineering* 43.4 (2016): 321-339.
- [2] Mordelet, Fantine, and J-P. Vert. "A bagging SVM to learn from positive and unlabeled examples." *Pattern Recognition Letters* 37 (2014): 201-209.
- [3] Jiang, Bingbing, et al. "Joint semi-supervised feature selection and classification through Bayesian approach." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019.