Experiment with removing frequent patterns

Aibolit team

June 4, 2020

Experiment

According to previous experiments patterns 'Non final class', 'Non final attribute', 'Null check' and 'Var in the middle' are most frequent and important.

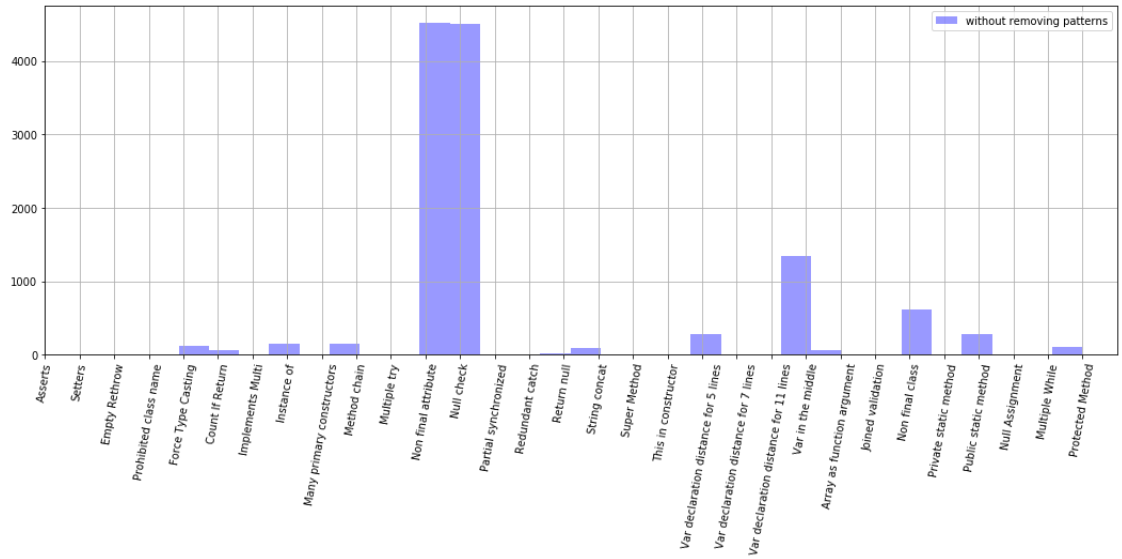

Figure 1: without removing patterns

In the experiment we trained model without this patterns and got best distribution of patterns' importances. Our model predict pattern which has the greatest impact on the complexity. On the graphics there is a distribution of most important patterns.

Results of the experiment are on the table.
p1 - remove pattern 'Non final class'
p2 - remove pattern 'Non final attribute'
p3 - remove pattern 'Null check'
p4 - remove pattern 'Var in the middle'

Patterns	mse	mae	r2
no removing	0.0182	0.0922	0.4792
p1	0.0181	0.0922	0.4811
p2	0.0195	0.0993	0.4395
p3	0.0209	0.1	0.4017
p4	0.0183	0.0932	0.4763
p1, p2	0.0196	0.0993	0.4366
p1, p3	0.0207	0.1	0.4052
p1, p4	0.0185	0.0935	0.4698
p2, p3	0.0228	0.1085	0.3453
p2, p4	0.02	0.1009	0.4259
p3, p4	0.0213	0.1016	0.3892
p1, p2, p3	0.0228	0.1087	0.3467
p1, p2, p4	0.02	0.1013	0.4264
p1, p3, p4	0.0214	0.1018	0.3874
p2, p3, p4	0.0235	0.1114	0.327
p1, p2, p3, p4	0.0237	0.1125	0.32

Table 1: Results of the experiment

Conclusion

According to the results only removing one pattern 'Non final class' could improve quality of prediction complexity.
More balanced distribution of patterns' importances is got in two cases:

- removing all 4 patterns (2)
- removing patterns 'Non final attribute', 'Null check' and 'Var in the middle' (3)

But in this 2 cases quality of prediction reduced.
In sum, better to remove all 4 patterns because balanced distribution is more important than quality of prediction complexity.

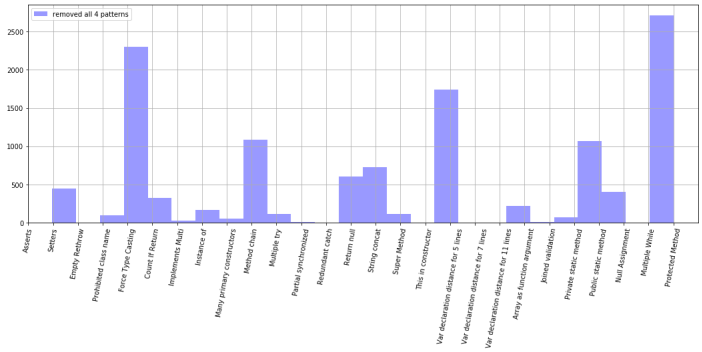

Figure 2: Removing all 4 patterns

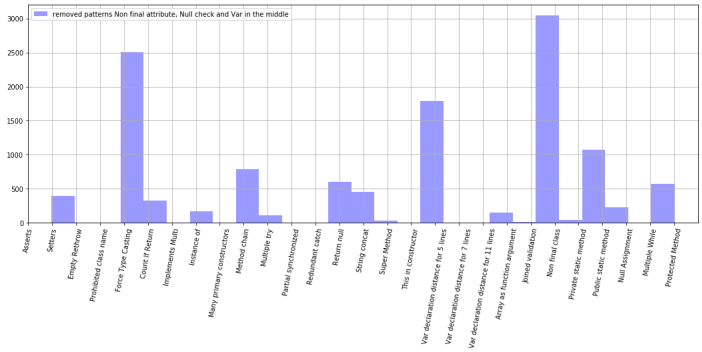

Figure 3: removing 'Non final attribute', 'Null check' and 'Var in the middle'