Automated Parameter Optimization of Classification Techniques for Defect Prediction Models, ICSE’16

Andre Lustosa Motta
North Carolina State University
alustos@ncsu.edu

***Abstract*—Defect prediction models are classifier that are trained to identify defect-prone software modules. These classifiers have configurable parameters that control their behaviour. Recent studies show that these classifiers may underpeform due to the use of suboptimal default parameters. In this poster paper [1] is discussed where the performance of classifiers using Caret - an automated parameter optimization techinique - was studied. Going through a case study with 18 datasets from proprietary and open source systems it was concluded that parameter settings can have a large impact on the performance of defect prediction models.**

***Index Terms*—Software defect prediction, experimental design, classification techniques parameter optimization**

I. Introduction

The limited SQA resources of software organizations must focus on modules likely to be defective in the future, to identify those modules defect prediction models are trained. These models use classification techniques which have configurable parameters that control characteristics of the model. For example the number of decision trees in a random forest, or the k in *k*-nearest neighbours. Since optimal settings for these parameters are not known ahead of time, they are usually left on default values.

A literature analysis revealed that 26 of the 30 most common classification techniques require at least one parameter setting. And since such setting may impact the performance of the model, it should be carefully selected. It is however impractical to assess all of the possible settings for any given classification technique. For example there are at least 17,000 possible settings to explore when training *k*-nearest neighbours classifier.

In [1] they investigate the performance of defect prediction models with the use of Caret - an off-the-shelf automated parameter optimization technique. Caret evaluates candidate parameter settings and suggests the optimized setting that achieves highest performance. The results on a case study of 18 datasets were recorded with respect to two dimensions:

- **Performance improvement:** Caret improves the AUC preformance of the models by up to 40% and the performance improvement is non-negligible for 16 of the 26 studied classifiers.
- **Performance stability:** Caret-optimized classifiers are at least as stable as the ones trained with default settings and 9 out of the 26 are more stable.

The results lead to a conclusion that parameter settings can have a large impact on the performance of defect-prediction models and since the cost of including optimization techniques is manageable, they should be inluded in future defect prediction studies.

II. Research Questions

Even though prior work suggests that defect prediction models may underperform if they are trained using suboptimal parameter settings, parameters are left at their default values and little research has been applied to optimize the parameters of classification techniques for defect prediction models.

RQ1 How much does the performance of defect prediction models improve when automated parameter optimization is applied?

Like any form of classifier optimization, automated parameter optimization may increase the risk of overfitting, producing a classifier that is too specialized for the data from which it was trained.

RQ2 How stable is the performance of defect prediction models when automated parameter optimization is applied?

III. Case Study Approach

A. DataSets

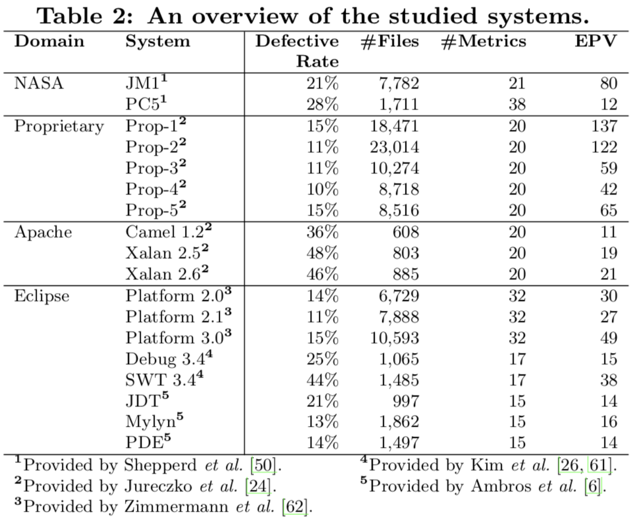


Table 2: An overview of the studied systems.

Domain	System	Defective Rate	#Files	#Metrics	EPV
NASA	JM1[1]	21%	7,782	21	80
	PC5[1]	28%	1,711	38	12
Proprietary	Prop-1[2]	15%	18,471	20	137
	Prop-2[2]	11%	23,014	20	122
	Prop-3[2]	11%	10,274	20	59
	Prop-4[2]	10%	8,718	20	42
	Prop-5[2]	15%	8,516	20	65
Apache	Camel 1.2[2]	36%	608	20	11
	Xalan 2.5[2]	48%	803	20	19
	Xalan 2.6[2]	46%	885	20	21
Eclipse	Platform 2.0[3]	14%	6,729	32	30
	Platform 2.1[3]	11%	7,888	32	27
	Platform 3.0[3]	15%	10,593	32	49
	Debug 3.4[4]	25%	1,065	17	15
	SWT 3.4[4]	44%	1,485	17	38
	JDT[5]	21%	997	15	14
	Mylyn[5]	13%	1,862	15	16
	PDE[5]	14%	1,497	15	14

[1] Provided by Shepperd *et al.* [50].
[2] Provided by Jureczko *et al.* [24].
[3] Provided by Zimmermann *et al.* [62].
[4] Provided by Kim *et al.* [26, 61].
[5] Provided by Ambros *et al.* [6].

B. Generate Bootstrap Sample

In order to ensure that the conclusions are robust the bootstrap validation technique was used. It consists in two steps:

- **(Step 1)** A boostrap sample of size N is randomly drawn with replacement from an original dataset which is also of size N
- **(Step 2)** A model is trained using the bootstrap sample and tested using rows that do not appear in the bootstrap sample.

C. Caret Parameter Optimization

Since it is impractical to assess all of the possible parameter settings of the suggested spaces, we use optimized parameter settings suggested by the **train** function of the **caret R** package. The process is made in three steps:

- **(Step 1)** Generate candidate parameter setting
- **(Step 2)** Evaluate candidate parameter settings
- **(Step 3)** Identify the Caret-optimized setting

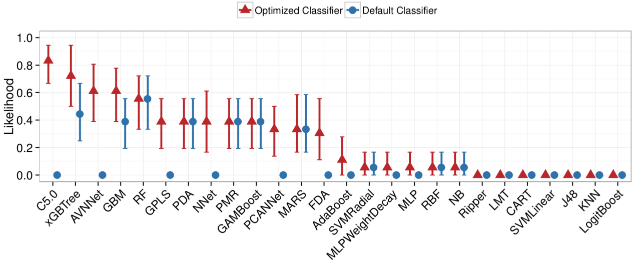
IV. Case Study Results

After the aforementioned analysis it was possible to get the following answers to the research questions:

RQ1 Caret improves the AUC performance of defect prediction models by up to 40 percentage points. Moreover, the performance improvement provided by Caret is non-negligible for 16 of the 26 studied classification techniques (62%)

RQ2 Caret-optimized classifiers are at least as stable as classifiers that are trained using the default settings. Moreover, the Caret-optimized classifiers of 9 of the 26 studied classification techniques (35%) are more stable than classifiers that are trained using the default values.

V. Revisiting the ranking of classification techniques for defect prediction models



As seen in the results. Automated parameter optimization increased the likelihood of appearing in the top Scott-Knott ESD rank by as much as 83%. And the computational cost involved was the addition of less than 30 minutes of additional computation time for 65% of the classifiers. Even those who had a big overhead, of up to 3 hours, such as AdaBoost, MLPWeightDecay and RBF could be computed overnight. Therefore this coupled with the fact that the models do not need to be built often in practice the cost is manageable.

VI. Threats to Validity

- **Construct Validity:** The results from RQ1 show that Caret improves the performance of defect prediction models. However, the performance improvement may increase the complexity of defect prediction models. Also the variation of metrics does not propose a threat to this study.
- **Internal Validity:** The performance of the classifiers was measured using AUC. Other performance measures might yield a different result. The refered paper [1] plans to expand the measures in future work. It states that the generalizability of the RLE (bootstrap based Ranking Likelihood Estimation) depends on how representative the sample is. That is why datasets of different sizes and domains were used. Even though a highly-controlled experiment was made to clean the datasets, the topic should be inspected in future work
- **External Validity:** The number of systems studied in this paper was limited. Therefore the results may not generalize to every software systems. However the point of this paper is not to generalize but to show that optimization may matter for certain datasets.

VII. Conclusions

Throughout the refered paper we were able to find these three big observations:

- Caret improves the AUC performance of defect prediction models by up to 40 percentage points. Moreover, the performance improvement provided by Caret is non-negligible for 16 of the 26 studied classication techniques (62%).
- Caret-optimized classiers are at least as stable as classiers that are trained using the default settings. Moreover, the Caret-optimized classiers of 9 of the 26 studied classication techniques (35%) are more stable than classiers that are trained using the default values.
- Caret increases the likelihood of producing a topperforming classier by as much as 83%, suggesting that automated parameter optimization can substantially shift the ranking of classication techniques.

These results lead to the conclusion that parameter settings can have a large impact on the performance of defect prediction models, suggesting that this should be experimented with to improve results of classification techniques. The paper finishes stating that given the availability of the automated parameter optimization in commonly-used research toolkits the recomendation of using this is rather simple and low-cost.

References

[1] C. Tantithamthavorn, S. McIntosh, A. Hassan, and K. Matsumoto, "Automated parameter optimization of classificationtechniques for defect prediction models."