doi:10.20396/revpibic262018369

Handling classification errors by learning to reject classifications

Edson Duarte*, Alexandre Ferreira, Jacques Wainer

Abstract
This work investigates how to improve classification metrics by learning when a classifier has higher chances of misclassification. The rejection technique increases the classification reliability and reduces the costs associated with misclassifications on cost-sensitive scenarios. A scenario of bug triaging classification shows the approach effectiveness where classification accuracy is increased from 67% up to 76%.

Key words:
Machine Learning, Classification Rejection, Bug Triaging

Introduction

When dealing with classification errors on unbalanced datasets that have different or unknown costs, the higher costs must be properly handled on cost-sensitive scenarios. One such scenario is automatic bug triaging systems where some classes have few samples and error costs are very different on each class.
Usually, the first approach to these problems is to balance out the dataset, be it by forcing classes with similar sizes by sampling the bigger classes or by data augmentation techiniques, or by weighting samples giving more relevance to smaller classes.
Alternative approaches, instead, focus on leveraging the results from classifiers that output confidence intervals, using these results to learn when a classifier has lower chances of a correct classification so that it can be handled properly.

Results and Discussion

Working with issue tickets from a private project management system, we developed a natural language processing model to help triaging teams to forward these tickets to their responsible tech teams. Considering the tech teams as classes, we are able to reach state-of-the-art performance.
For comparison with the literature [1] [2], we used our model on 42,000 tickets from Mozilla's Bugzilla instance [3], with the 150 most common assigned developers as classes, achieving accuracy of 66.8%.
Considering our working context, this performance, although good, is still not desirable as allocating teams and resources are expensive, and to deal with this situation we investigated options on rejecting classifications [4].
Our proposed method consists in training another classifier, the critic, on the confidence intervals given by the first classifier, the critic learns when the classifier has higher chances of giving a wrong classification.
Figure 1 shows how accuracy increases as we use the critic to remove samples with higher chances of misclassification, and compares the critic to using a metric, such as entropy, directly on the confidence intervals.
Considering this approach, we are able to achieve 76.2% of accuracy by rejecting classification on 20% of the samples. This value can be adjusted depending on the defined range of interest, which limits how many samples we remove from classifications. Rejecting too much is not desirable, we still want to classify as many as possible.
The selection of how many samples will be removed is then a trade-off between increasing accuracy and confidence in the model, and allocating people and resources (of both tech and triaging teams).

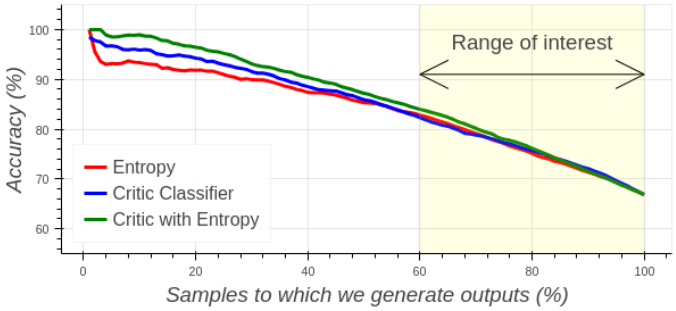

Figure 1. Rejection plot, highlighted is the range of interest where we increase accuracy by rejecting to classify a fraction of samples.

Conclusions

In this work, we managed to increase accuracy by training a critic classifier to learn when we have higher chances of properly classifying a ticket. This way, wrong classifications do not end up sending tickets to wrong tech teams, saving on time and resources.
As next steps, we plan on using additional and more sophisticated metrics and improving both classifiers.

Acknowledgment

This research was supported by Motorola Mobility LLC – Brazil, in partnership with Instituto de Pesquisas Eldorado, under grant 5078.

[1] Xuan J, Jiang H, Hu Y, Ren Z, Zou W, Luo Z, Wu X (2014). "Towards effective bug triage with software data reduction techniques". *IEEE transactions on knowledge and data engineering*.

[2] Xia X, Lo D, Ding Y, Al-Kofahi JM, Nguyen TN, Wang X (2017). "Improving automated bug triaging with specialized topic model". *IEEE Transactions on Software Engineering*. 43(3): 272-297.

[3] Mozilla's Bugzilla triaging system (2018). https://bugzilla.mozilla.org

[4] Guan H, Zhang Y, Cheng HD, Tang X (2018). "Abstaining Classification When Error Costs are Unequal and Unknown". *arXiv preprint.* arXiv:1806.03445.