PPML: Performance Profile for Machine Learning

1st HOUNDJI Vinasetan Ratheil
Institut de Formation et de Recherche en Informatique (of University of Abomey-Calavi)
Abomey-Calavi, Benin
0000-0002-5467-9448

***Abstract*—The classic way to compare the performance of Machine Learning models consists to use different well-known metrics. In general, these metrics only provide aggregate performances of each model separately. This paper introduces PPML (Performance Profile for Machine Learning), an approach to refine the analysis by comparing the performances of models on the same plot. With PPML, one can easily answer this kind of question: 1) the probability that a given model is the best among the other models? 2) the probability that a given model is x times worse than the other models? and 3) the probability that a given model is never x times worse than the other models? This paper shows how to use it in regression tasks with absolute error as the main metric. Moreover, it provides a use case in which PPML can be a very useful tool: multi-class classification when two wrong answers do not have the same impact.**

***Index Terms*—evaluation, performance profile, machine learning, regression, classification**

I. Introduction

There are many Machine Learning (ML) algorithms and it is relatively easy to implement and test them thanks to the available libraries. An important task is then the comparison and analysis of the implemented models on a given test set. For example, in regression task, to evaluate ML algorithms, one use metrics such as MAE, MSE, $RMSE$, R^2, etc. These metrics are good but only give the aggregate performances of each model separately. For example, one cannot say something like "Model 1 is twice worse than the best model on 25% of the instances of the data set".

To refine the analysis of the performances of the ML models, this paper proposes to use performance profiles, distribution functions for a performance metric [1]. For a given model, the performance profile provides its cumulative performance w.r.t. the best model on each instance. Performance profiles are usually used in Operations Research.

Thanks to this approach one can easily answer this kind of question:

- the probability (proportion of instances) that a given model is the best among the other models?
- the probability (proportion of instances) that a given model is x times worse than the other models?
- the probability (proportion of instances in the test set) that a given model is never x times worse than the other models?

This could also help to know if the dataset contains different "kinds" of instances. For example, if a model is very good on some instances and is very bad on other ones, further dataset analysis is necessary.

The remainder of this paper is organized as follows. Section II defines PPML for regression and illustrates it on a small example. Then Section II shows how PPML can be used for multi-class classification when two wrong answers have different consequences. To our knowledge, there is no metric for this kind of problem.

II. PPML in regression

In this section, the paper formally defines the performance profile for ML in regression tasks and shows a simple pedagogical example.

A. Definition

Performance profile offers the advantage that the performance is evaluated on each instance separately. This leads us to consider, for the illustration, the absolute error on each instance as the performance metric for regression.

Formally, consider a set of models $\mathcal{M}$ and a set of instances $\mathcal{I}$ (test set of the dataset). Let r_m^i be the performance ratio of the model $m \in \mathcal{M}$ w.r.t. to the best performance by any model in $\mathcal{M}$ on the instance i:

$$r_m^i = \frac{e_m^i}{\min\{e_x^i \mid x \in \mathcal{F}\}} \tag{1}$$

in which e_m^i is the absolute error on the prediction of the model m on the instance i. We have $e_m^i = |v_i - p_m^i|$ with v_i the true value to predict for the instance i and p_m^i the value returned by the model m for the instance i.

Then

$$\rho_p(\tau) = \frac{1}{|\mathcal{I}|} \cdot |\{i \in \mathcal{I} \mid r_m^i \leq \tau\}| \tag{2}$$

is the proportion of instances for which the model $m \in \mathcal{M}$ has a performance ratio r_m^i within a factor $\tau \in \mathbb{R}$ of the best possible ratio. In other words, for a point (x, y) on the performance profile, the value $(1 - y)$ gives the percentage of the instances where the given model is at least x times worse than the best model on each instance.

B. Example

Assume that we want to compare three models M_1, M_2 and M_3 on five instances of a test set of a regression problem. Table I provides: the true value to predict (column v with v_i is the true value of the instance i); the different values predicted by each model as well as the absolute error on each of the 5 instances (in $\alpha \| \beta$, α is the value predicted by the corresponding model and β is the absolute error $|\alpha - v_i|$).

Instance	v	M_1	M_2	M_3
1	2	14‖12	10‖8	**6**‖4
2	6	**11**‖5	**1**‖5	12‖6
3	3	21‖18	9‖6	**0**‖3
4	10	0‖10	**12**‖2	2‖8
5	23	2‖21	15‖8	**18**‖5

TABLE I
EXAMPLE: PERFORMANCE PROFILE - MEASURES

Instance	M_1	M_2	M_3
1	3	2	**1**
2	**1**	**1**	1.2
3	6	2	**1**
4	5	**1**	4
5	4.2	1.6	**1**

TABLE II
EXAMPLE: PERFORMANCE PROFILE - PERFORMANCE RATIOS r_p^i

One can deduce the different performance ratios by instance/model using the formula 1 (see Table II).

Based on the different ratios provided by Table II, Table III reports the different cumulative performances using the formula 2.

τ	M_1	M_2	M_3
1	0.2	0.4	0.6
1.2	0.2	0.4	0.8
1.6	0.2	0.6	0.8
2	0.2	1	0.8
3	0.4	1	0.8
4	0.4	1	1
4.2	0.6	1	1
5	0.8	1	1
6	1	1	1

TABLE III
EXAMPLE: PERFORMANCE PROFILE - CUMULATIVE PERFORMANCE $\rho_p(\tau)$

Figure 1 shows the corresponding performance profile. One can deduce the following conclusions from this figure:

- M_3 is the best model for 60% of the instances, M_2 is the best for 40% of the instances and M_1 is the best for 20% of the instances;
- M_1 is the worst model. M_1 is ≥ 3 times worse than the best model for 60% of the instances;
- M_2 is never > 1.6 times worse than the other models and M_3 is ≤ 1.2 times worse than the other models for 80% of the instances.

III. PPML IN MULTI-CLASS CLASSIFICATION

PPML can be a very useful tool for multi-class classification when two wrong answers do not have the same meaning. To illustrate this, let us consider the following example. One wants to set up an ML model to recognize traffic light colors for a driver-less taxi. Assume that the target is Red and two different models have failed by predicting respectively Green and Orange. The model that predicts Green is clearly worse and more dangerous than the second one that predicts orange.

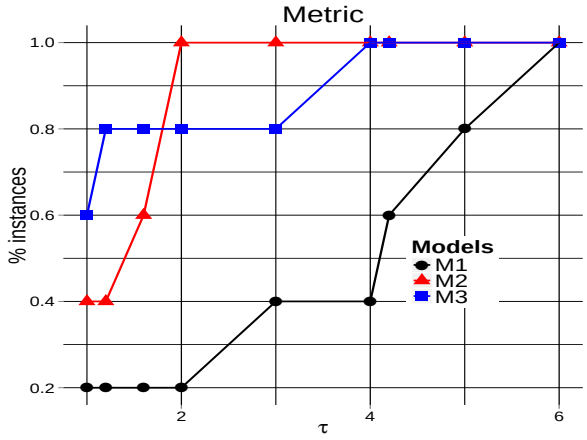

Fig. 1. Performance profile - Example

To the best of our knowledge, classic metrics such as precision, accuracy, F1-score, AUC, etc. cannot highlight this kind of performance when different "weights" are associated with wrong answers. In this case, performance profiles can easily be used. To adapt it to this kind of problem, it is sufficient to weigh the wrong answer. A model will get 1 if its prediction is correct and a weight otherwise (more the weigh is high, more the prediction is worse).

By considering our traffic light example, one can have Table IV to weigh the predictions of the models to evaluate. In this example, if the target is Green, a model that predicts Orange is twice times worse than the correct one and a model that predicts red is 4 times worse than the correct one.

	Green	Orange	Red
Target=Green	**1**	2	4
Target=Orange	4	**1**	2
Target=Red	10	4	**1**

TABLE IV
EXAMPLE OF WEIGHTS FOR ILLUSTRATION

From this step, the rest is straightforward to get the performance profiles.

IV. CONCLUSION

This paper has introduced Performance Profile for Machine Learning (PPML), an adaptation of the classic performance profiles for ML. For regression, PPML can be very useful to refine the analysis of the performances together with classic metrics. For multi-class classification when two wrong answers do not have the same impact, PPML offers a powerful tool for analysis that cannot be tackled by the classic state-of-the-art metrics.

REFERENCES

[1] E. D. Dolan, and J. J. Moré, "Benchmarking optimization software with performance profiles," Mathematical programming, 91, 2, Springer, pp. 201–213, 2002.