

Hyperparameter Tuning for the Comparative Evaluation of Unsupervised Anomaly Detectors



Jonas Soenen*, Elia Van Wolputte*, Lorenzo Perini, Vincent Vercruyssen, Wannes Meert, Jesse Davis and Hendrik Blockeel ✉ firstname.lastname@kuleuven.be

Tune hyperparameters using a small labeled validation set for a reproduceable, fair and realistic comparison between competitors.


Problem statement

When comparing several anomaly detection algorithms over a set of datasets, how can we select the hyperparameters of each algorithm to ensure a fair and realistic comparison?

Ways to select hyperparameters

Out-of-the-box performance

Use default hyperparameters or rules of thumb




Assumes practitioners will do **no effort** to find good hyperparameters for the task at hand

Underestimates the potential of algorithms
Ambiguous: how to select defaults?
Unfair: are all defaults selected the same way?

Tuned performance

Use a small labelled validation set to tune the hyperparameters




Assumes practitioners will do **an honest effort** to select good task-specific hyperparameters

Aims at a **realistic** performance estimate
Reproduceable, fair and sound

Peak performance

Use optimal hyperparameters



Assumes practitioners can select **optimal hyperparameters** for a task

Overestimates the potential of algorithms
Unsound: tuning on the test set

Experiments

Exp 1.

The final ranking when comparing six popular AD algorithms based on out-of-the-box (on the left) and peak (on the right) performance over 16 standard anomaly detection datasets.

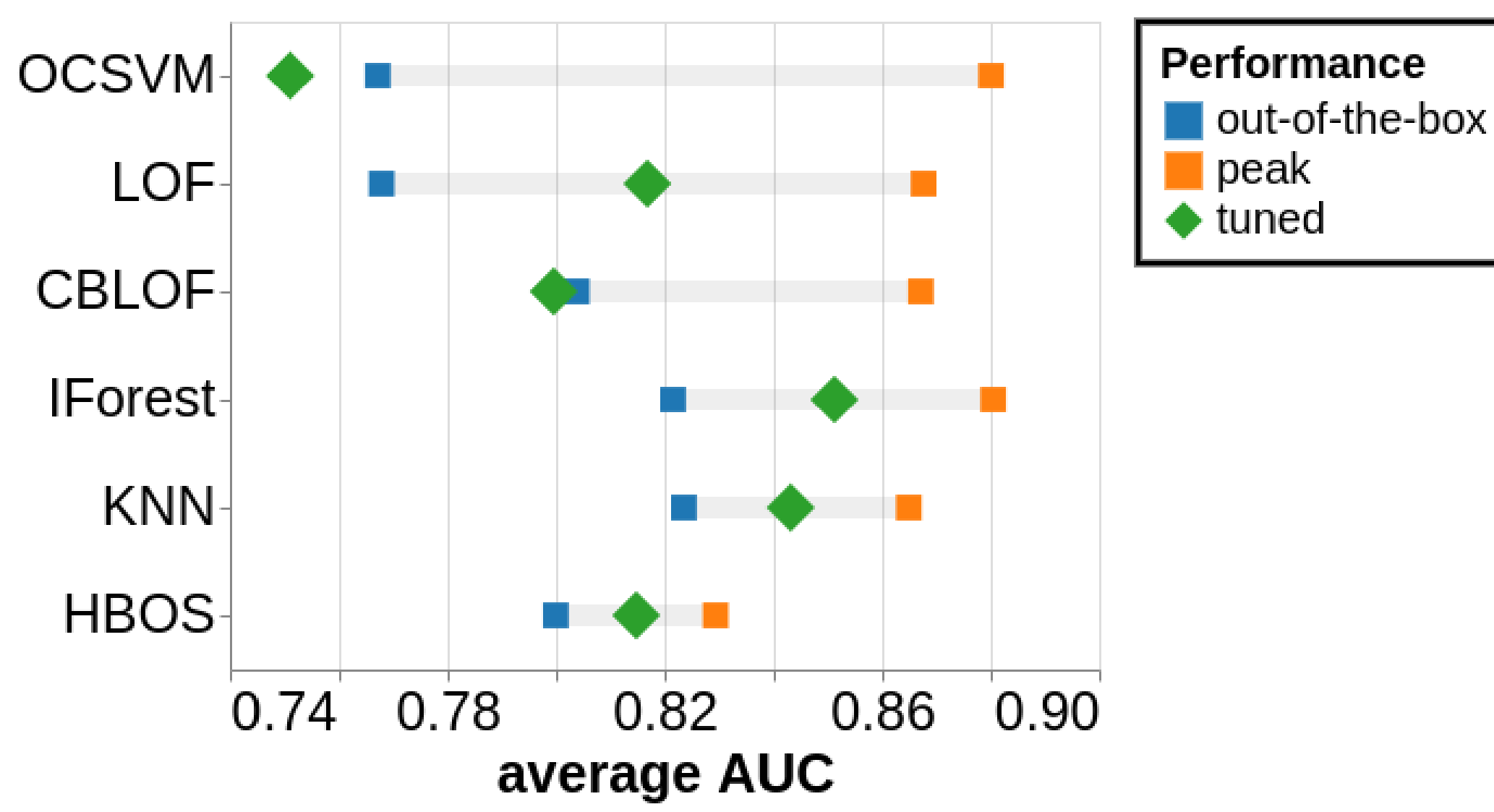
Out-of-the-box (Eq. 1)		
algorithm	avg AUC	rank
IForest	0.82	2.66
KNN	0.82	2.91
CBLOF	0.8	3.22
HBOS	0.8	3.81
LOF	0.77	4.12
OCSVM	0.77	4.28

Peak (Eq. 2)		
algorithm	avg AUC	rank
IForest	0.88	2.72
CBLOF	0.87	3.03
OCSVM	0.88	3.28
LOF	0.87	3.5
KNN	0.86	3.88
HBOS	0.83	4.59

The final ranking depends on the hyperparameter selection methodology

Exp 2.

The average AUC of six popular AD algorithms over 16 standard anomaly detection datasets using different hyperparameter selection methodologies



Generally, tuning results in better-than-default hyperparameters but doesn't reach peak performance
However, for OCSVM and CBLOF, tuning is counterproductive
For these algorithms, good performance on the validation set does not translate to the test set

Conclusions

1. Out-of-the-box performance is overly pessimistic, whereas peak performance is overly optimistic.
 2. The proposed tuned performance yields *realistic* performance estimates because it considers the difficulty of tuning a particular algorithm.
 3. Our methodology is practically feasible, as a small validation set with few anomalies is sufficient to tune the hyperparameters.
- For more details, notes about the validation set size and more experiments see [1].

[1] Soenen, Jonas, et al. "The Effect of Hyperparameter Tuning on the Comparative Evaluation of Unsupervised Anomaly Detection Methods." *Proceedings of the KDD'21 Workshop on Outlier Detection and Description*, 2021. <https://oddworkshop.github.io/>

* These two authors contributed equally to the paper.