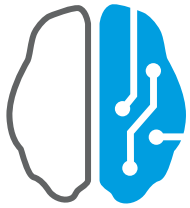


Intelligent
Embedded Systems

Evaluation in Active Learning

Tutorial

Interactive Adaptive Learning

September 18, 2017

DanielKottke

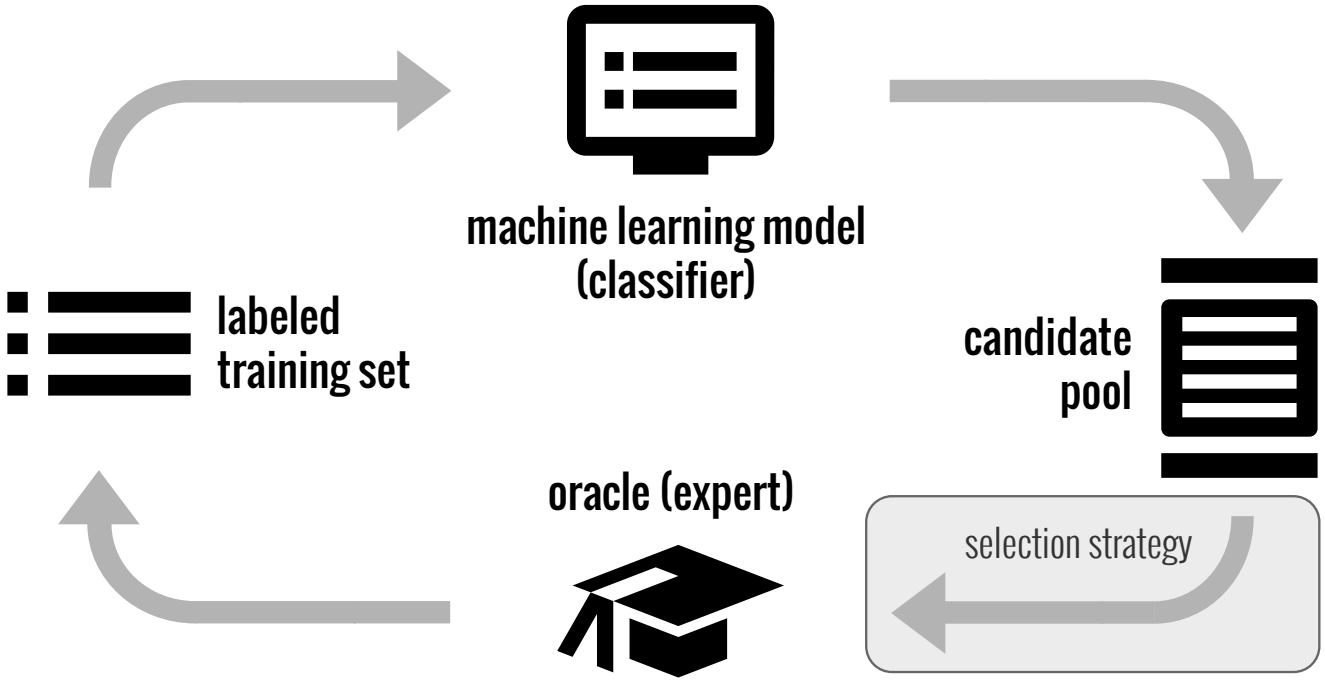
Motivation[1]

The evaluation methodology should be

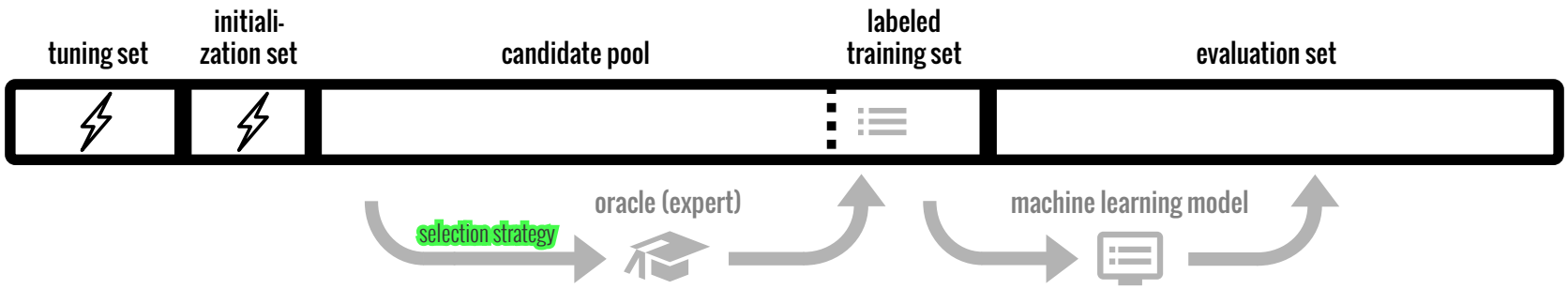
1. reliable
   - robust to varying seeds or shuffling data
   - reproducible (well-described, availability of data)
2. realistic
   - valid assumptions for real applications
3. comparable
   - development of a standardized active learning evaluation gold standard to compare algorithms without reimplementing

[1] This talk is based on "Challenges of Reliable, Realistic and Comparable Active Learning Evaluation" by Kottke et al., IAL@ECMLPKDD, 2017

Pool-based Active Learning Cycle [Set09]



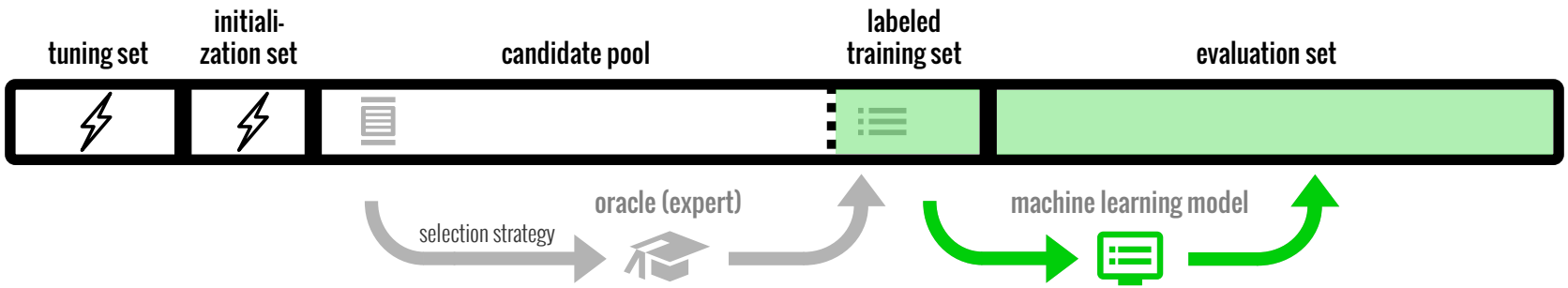
A Different View on the Active Learning Cycle



We want to evaluate the performance of the **selection strategy**.

Reliable evaluation

Evaluating the Model's Performance



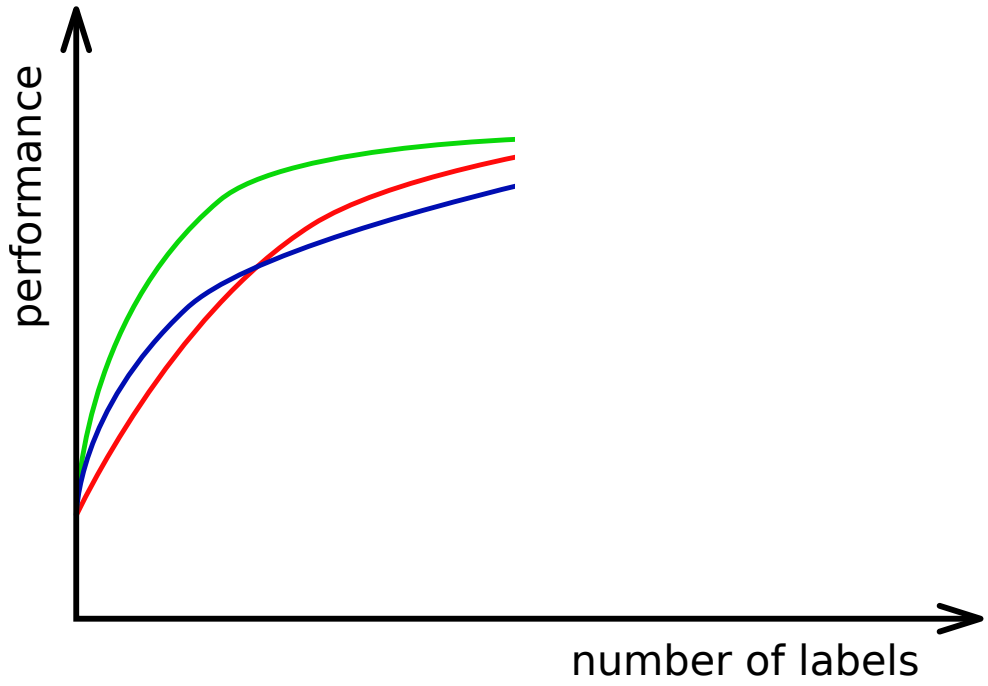
- training set is subsequently filled with selected candidates
- the learned model is evaluated on a hold-out evaluation set

Which performance measure should be used?

- depends on the application
  - balanced class priors (e.g., accuracy, error)
  - unbalanced class priors (e.g., f1-score, AUROC)
- complexity [Par11]:
  - point measures (e.g., accuracy, precision, recall)
  - integrated measures (e.g., AUROC, H-Measure)

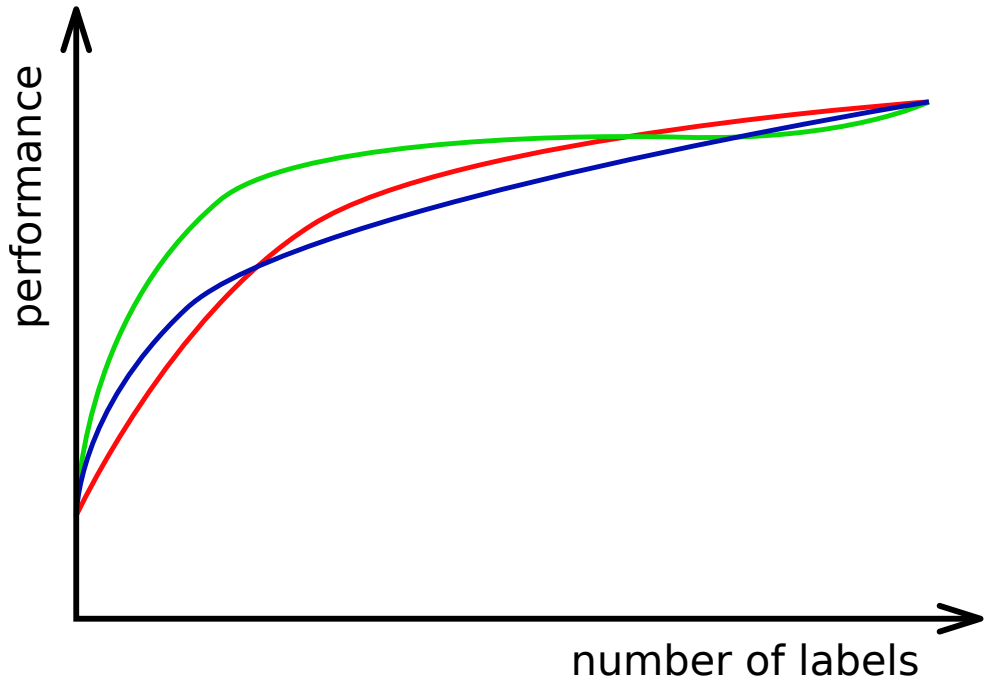
How to interpret the results of a learning curve?

- converging as fast as possible
- converging to the highest overall value



How to interpret the results of a learning curve?

- converging as fast as possible
- converging to the highest overall value



How to summarize results from a learning curve?

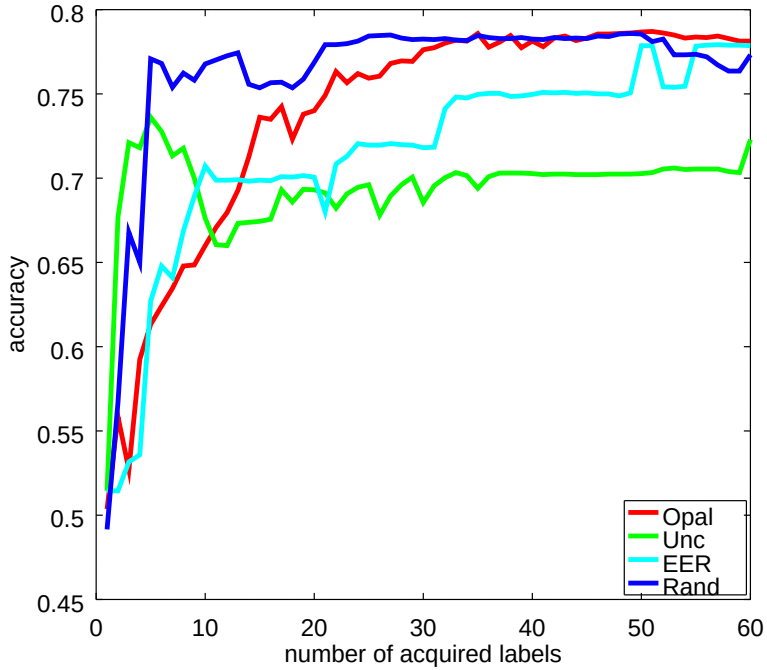
- Table at specific time points (early, mid, late)
- Area under the learning curve, mean (depends on stopping point)
- deficiency [YS15]
- data utilization rate [RS13]

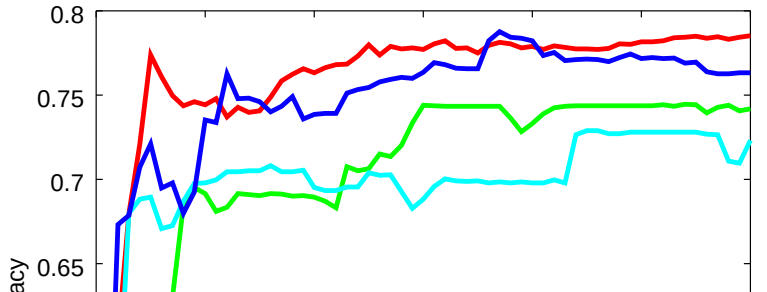
How to evaluate statistical significance?

- Which values to compare?
  - **not** across label acquisitions (highly correlated) but across multiple repetitions
  - at which point in time?
- Statistical tests
  - t-Test cmp. mean (assumes that mean is normal distributed)
  - Wilcoxon Signed Rank Test cmp. tendency (parameter-free test)
- always present results with **statistical significance** and **effect size**

How many repetitions are required?

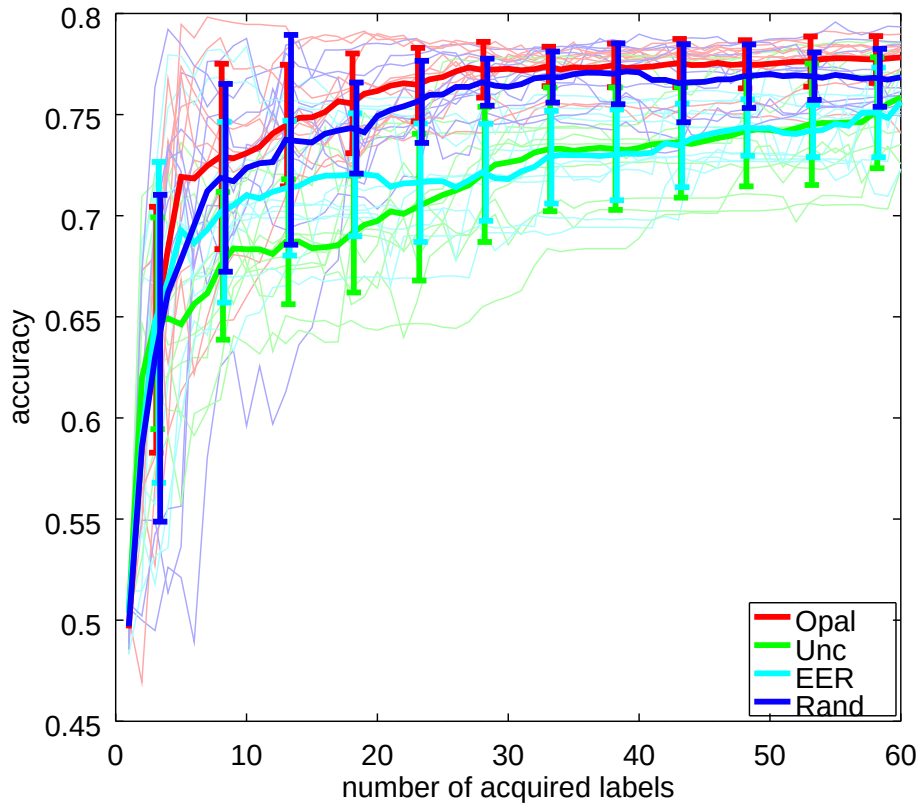
Comparison of algorithms using 5-fold cross validation



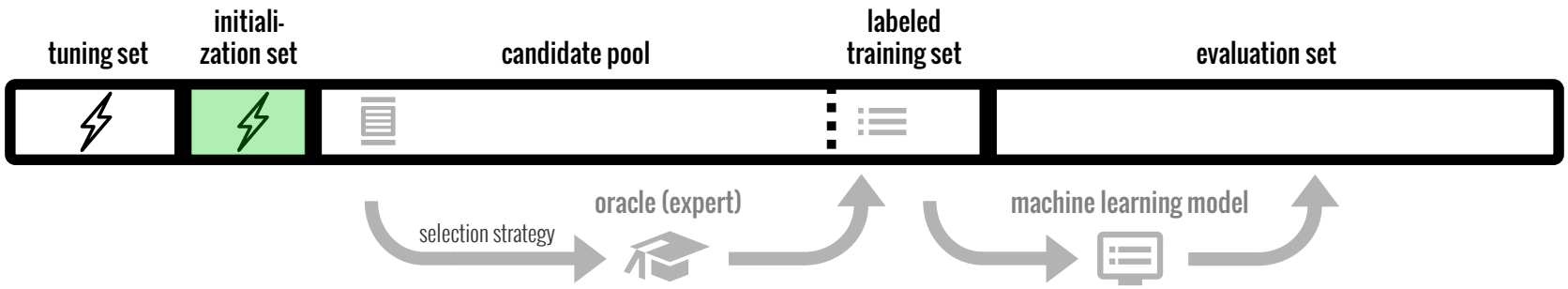


How many repetitions are required?

Comparison of algorithms using 5-fold cross validation



Initialization of Instance Selection



Objectives:

1. Compatibility issues
2. Improve the proposed selection method

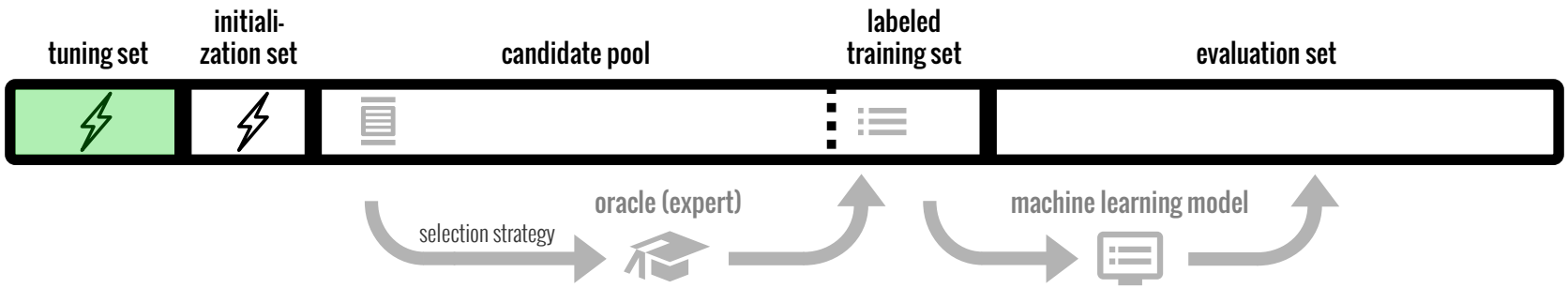
Initialization due to compatibility

- example: libSVM needs at least one instance per class
- selecting one instance from each class is not possible, as one does not know these in advance
- this is even more unlikely for unbalanced datasets
- include number of selected instances in analysis

Initialization to improve main selection method

- often random selection for an initial exploration phase
- number of labels is tuned (how much labels does the algorithm need)
- highly dataset dependent (how to tune?)
- include number of selected instances in analysis

Parameter Tuning



1. Determine hyperparameter and fix them across selection methods
2. How to tune without labels?

Parameter Tuning

- tuning instances should be considered in the number of acquisitions
- how many instances should be used for tuning? (many classifiers are sensitive to the number of instances)
- normally, no instances for supervised parameter tuning available
- tuning parallel to sampling may be complicated

Realistic evaluation

Real applications oft are more challenging

- Often highly specialized (hard to transfer approaches to related domains)
- Imperfect labelers (experts might be wrong)
- In real-world only one shot (mean results are not representative)
- Labels are not always available (in time and space)
- Performance guarantees (cmp. random sampling)
- Assess online performance of an actively trained classifier
- Different costs for different annotations or classes
- Ground truth might not be available

Comparable evaluation

Discussion on an Evaluation Gold Standard

- Use exactly the same robust classifier for every AL method when comparing and try to sync the parameters of these classifiers.
- Capture the effect of different AL methods on multiple datasets using at least 50 repetitions.
- Start with an initially unlabeled set. If you need initial training instances, sample randomly and explain when to stop.
- Use either a clear defined stopping criterion or enough label acquisitions (sample until convergence).
- Show learning curves (incl. quartiles) with reasonable performance measures.
- Present pairwise differences in terms of significance and effect size (Wilcoxon signed rank test).

References I

- Charles Parker, *An analysis of performance measures for binary classifiers*, Proceedings of the 11th IEEE International Conference on Data Mining (ICDM), IEEE, 2011, pp. 517–526.
- T Reitmaier and B Sick, *Let us know your decision: Pool-based active training of a generative classifier with the selection strategy 4DS*, Information Sciences - Informatics and Computer Science Intelligent Systems Applications, vol. 230, 2013, pp. 106–131.
- B. Settles, *Active learning literature survey*, Computer Sciences Technical Report 1648, University of Wisconsin, Department of Computer Science, 2009.

References II

Erelcan Yanik and Tevfik Metin Sezgin, *Active learning for sketch recognition*, Computers and Graphics (Pergamon) **52** (2015), 93–105.