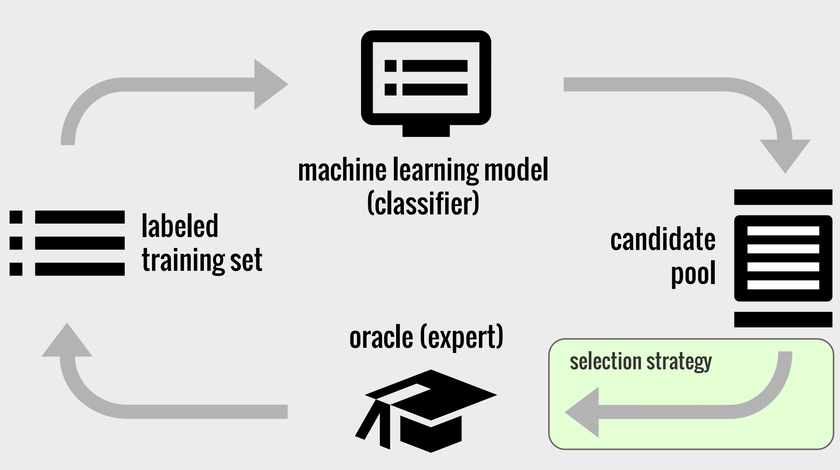
Intelligent Embedded Systems

KNOWLEDGE MANAGEMENT AND DISCOVERY

Challenges of Reliable, Realistic and Comparable Active Learning Evaluation

Daniel Kottke[1], Adrian Calma[1], Denis Huseljic[1], Georg Krempl[2], and Bernhard Sick[1]

Motivation



Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin, Department of Computer Science (2009)

Realistic Evaluation

Evaluating an AL algorithm in a lab setting (the lack of labels is just simulated) is not realistic. Often, implications for the real world do not hold. Hence, AL is not very common in industrial applications.

Challenges of real-world applications

- 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

Current evaluation methodologies vary a lot regarding its evaluation type, performance measure, number of repetitions, etc. Ideally, presented results are directly comparable with others.

Discussion of a 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 how to determine the number of samples.
- 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).

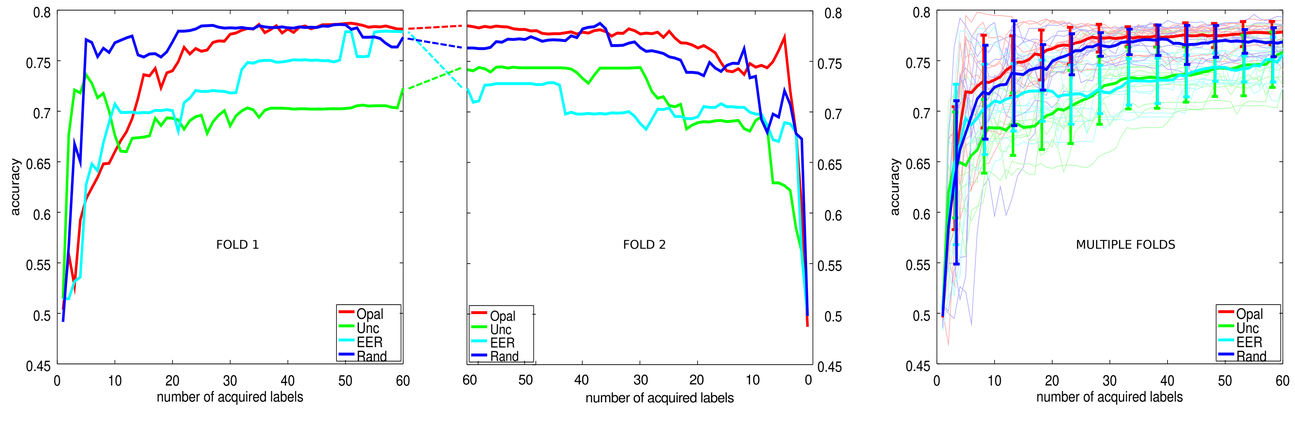
Reliable Evaluation

Reliable evaluation results require a robust and reproducible evaluation methodology. Hence, the methodology should be described in detail and should be robust to varying seeds or shuffled data.

Repetitions and Hold-Out Evaluation

- The performance is very sensitive to changes caused by the small number of training samples.
- It varies a lot depending on the concrete choice of instances.
- Lots of repetitions are needed to get a reliable trend of the performance.

Comparison between 5-fold cross valdations with and without repetitions

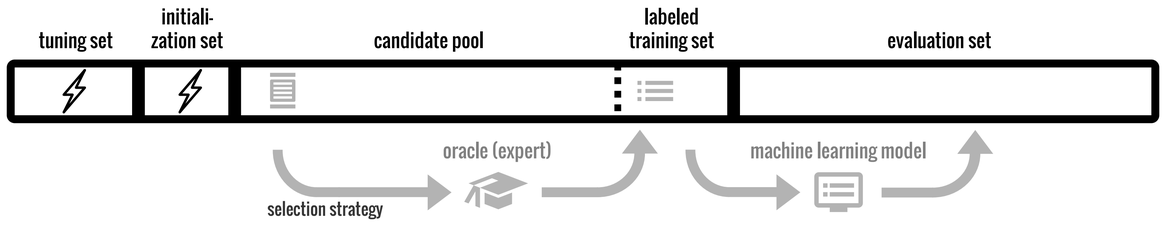


Performance Measures

- Two main objectives: 1) achieve a high performance level, 2) learn as fast as possible
- Best performance measure depends on the learning problem
- Results from multiple executions should be included in the evaluation by plotting standard deviations or ideally quartiles

Initialization of Active Learning

- Papers propose to initialize their AL cycle with some labels as an essential part of their algorithm. But there is not always an description on how the initialization works.
- Some authors added a fixed number of instances per class, although this is not possible in real applications.
- The initialization phase is a part of the active learning algorithm and should be somehow evaluated.



Parameter Tuning

- Tuning parameters for classifiers is very difficult with only a few labels available.
- Unfortunately tuning procedures are often not described in sufficient detail.
- We either use a pre-trained mediocre classifier because parameters are tuned for a specific labeling situation, or we re-calibrate the parameters during learning which means that classifiers become different across selection methods which also biases the results.

Affiliations: [1)] Intelligent Embedded Systems, University of Kassel; [2)] OvGU Magdeburg.

Published: International Workshop and Tutorial on Interactive Adaptive Learning, Skopje, Macedonia, 2017.

UNIKASSEL VERSITÄT

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF

Download presentation, paper and poster