Active Learning Benchmark

Towards Comparable Active Learning

February 31, 2021

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim

Motivation

The DAL result landscape is terrible

Every paper uses different datasets and use-cases

Every paper has a different classification model and training regime

No one-fits-all evaluation metric

Competitor Papers

Some Benchmark papers for DAL exist already:

- Focus on Image Classification
- Focus on best possible classifier performance
  - Data Augmentation
  - Type of Optimizer
  - Semi-Supervised Learning

Main Idea

We focus on the AL algorithms themselves:

- When we find the ”best” algorithm, classification performance will come naturally
- AL is lacking reproducible research, not strong classification models (random/uncertainty sampling does a good job already)
- We incorporate different domains instead of focusing on images
- We reduce the amount of true hyperparameters by design

Theoretical Problems

Some conflicting results from other papers:

- Data augmentations (seemingly) can replace diversity components from AL algorithms (under investigation currently)

Scope

The framework is planned to include

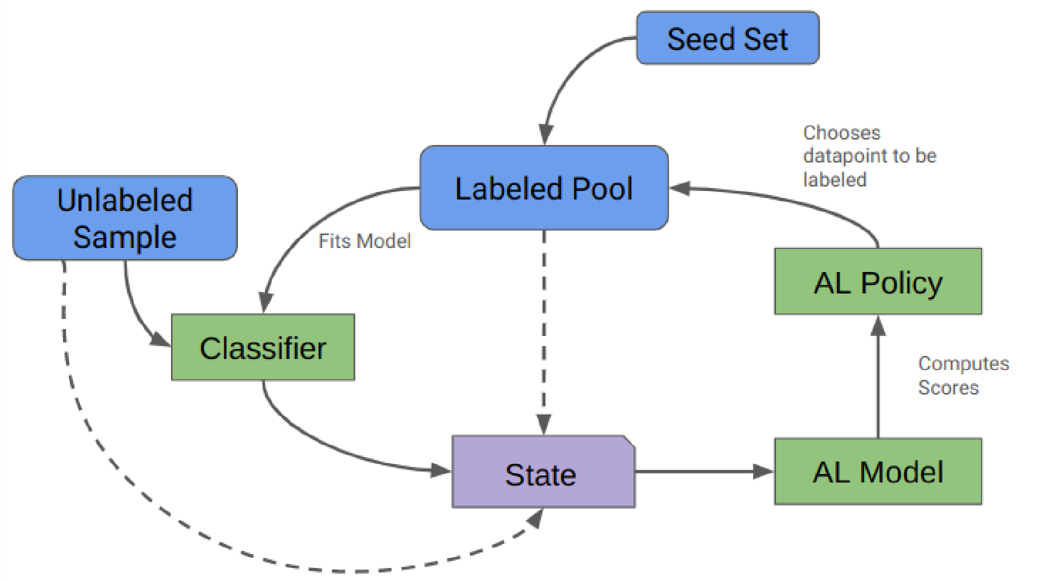
- Three different domains (Tabular, Image, Text)
- SOTA baseline methods
- Preprocessed datasets (train/test split and seed set)
- Tuned classifiers for each dataset
- Set logging and evaluation procedures
- (Ensemble Approaches)
- (Single Domain and Domain-Transfer use cases)

The framework will **not** include

- Batch Active Learning
- Elaborate Training (Data Augmentation, etc.)

Design Decisions

The AL Cycle



Seed Set
Labeled Pool
Chooses datapoint to be labeled
Unlabeled Sample
Fits Model
AL Policy
Classifier
Computes Scores
State
AL Model

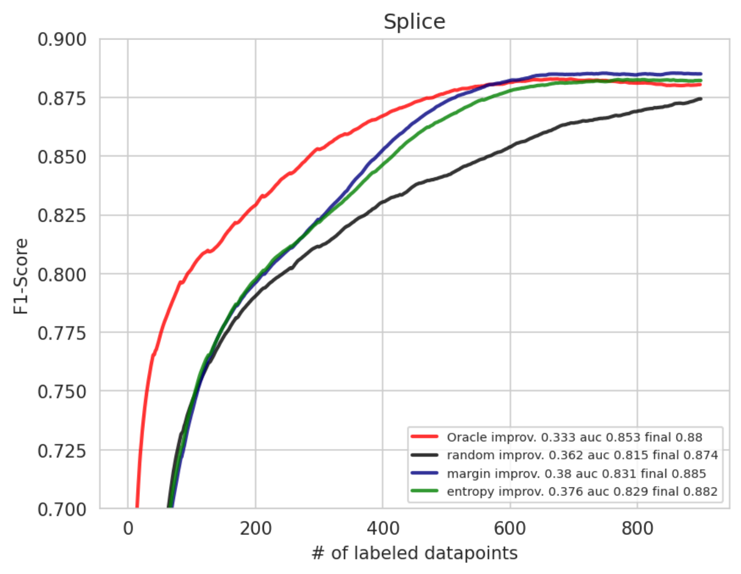
Datasets

Each dataset should be preprocessed - The features are normalized, the target are one-hot encoded and **the seed set should be fixed**

Datasets should be selected by their ”potential”

Each dataset needs an oracle curve that sets an upper bound performance

Dataset Potential



Classifier

The classification model should be governed by the dataset

There is no need to have the same model for every dataset, as long as the model in question is suited well for the data

Simpler models with less dynamic behaviour are better - SOTA performance is not so important

The model needs have tuned hyperparameters for each dataset (full dataset or subset?)

The expectation is that good AL algorithms will also work well on SOTA models

Evaluation Metrics

Multiple options are available

- Accuracy* / F1-Score
- AUC
- Advantage over random sampling*
- Regret from the oracle curve
- Relative performance (0% = random / 100% = oracle)

Cross Validation

AL can cross validate in two ways

- Changing the seed set
- Changing the presented sub-samples of unlabeled points
- Changing model initialization

Currently, I keep the seed set fixed

Is there value in fixing everything? (Model checkpoints, pre-sampling data*, etc.)

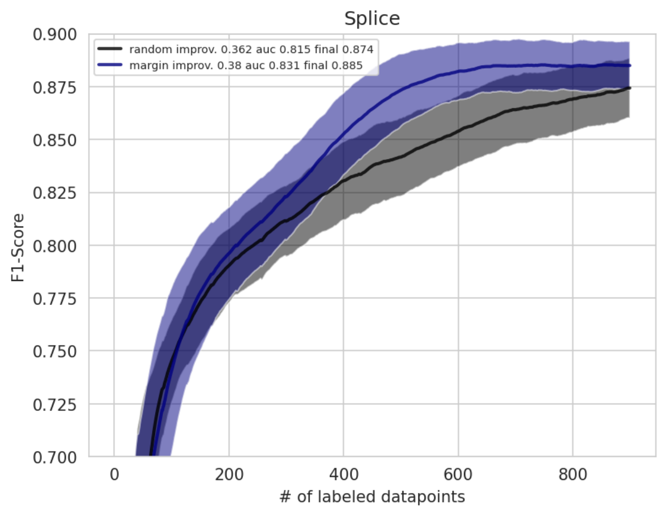
Cross Validation

Splice (Tabular):

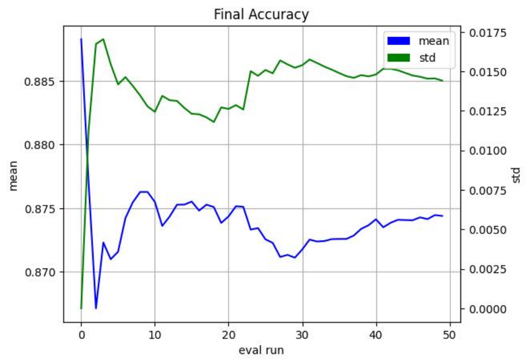
Random: 0.874 ± 0.01

Margin: 0.885 ± 0.015

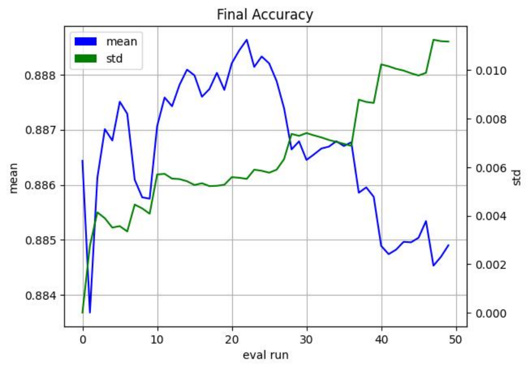
Difference: 0.011



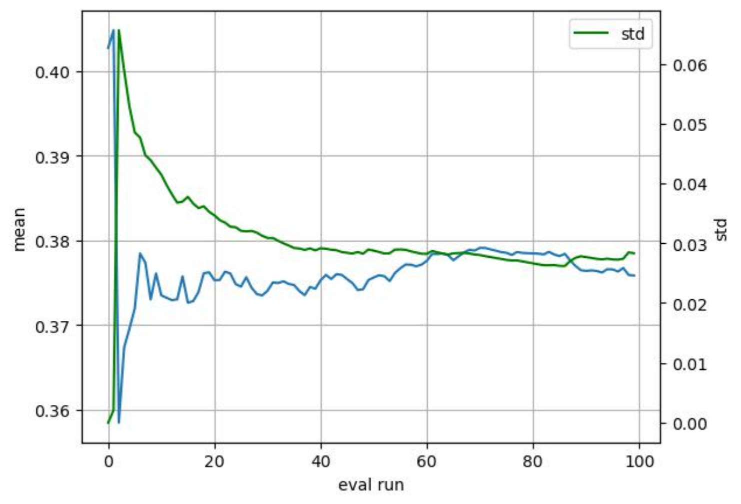
Random Sampling



Uncertainty Sampling



Cross Validation



Uncertainty Sampling

Hyperparameters: Dataset

Budget is set to allow convergence

Seed set is fixed to 1 point per class unless the cold start problem is to pronounced

Classifier batch size / LR / regularization is optimized for performance
Image datasets might require a trade-off for speed (Cifar10 currently needs 8 days)

Hyperparameters: Agent

Each AL agent defines its own state space with the information from the environment

Other hyperparameters for baselines need to be investigated

True Hyperparameters

This leaves very few ”true” hyperparameters:

- Unlabeled Sample Size
- # of cross validation trials

Ensemble Approaches

Uncertainty Sampling profits a lot from ensembles

The logical thing is to build the final classifier as an ensemble

How do you compare single model approaches to ensemble approaches?

Prefix a single evaluation model out of the ensemble?

Single Domain vs. Transfer

Single Domain: Fit AL agent on the same dataset that you evaluate on (not useful in practice)

Domain Transfer: Fit AL agent on domain with large resources and evaluate agent on domain with few resources

Often used in NLP (fit agent on English data and evaluate for Spanish / German)

The End

Cheers

Classifier Training

General Setup	Speed Optimizations
Cross-Entropy Loss	Fine tuning
No case weights	Aggressive Early Stopping
At least one epoch	
NAdam optimizer	

Results for Splice

Agent	Acc	AUC
Oracle	0.8806 ± 0.0141	0.8530 ± 0.0132
SAR	0.8839 ± 0.0142	0.8339 ± 0.0118
Margin	0.8849 ± 0.0112	0.8309 ± 0.0141
Entropy	0.8822 ± 0.0156	0.8285 ± 0.0154
Coreset	0.8788 ± 0.0105	0.8246 ± 0.0108
Random	0.8744 ± 0.0144	0.8152 ± 0.0129

Results for Splice