CSE 584: Machine Learning - Homework 1

Sinjoy Saha

September 15, 2024

1 Introduction

The following papers are discussed here:

1. Ebert et. al., RALF: A Reinforced Active Learning Formulation for Object Class Recognition CVPR 2012 [1]
2. Konyushkova et. al., Learning Active Learning from Data. NeurIPS 2017. [2]
3. Muldrew et. al., Active preference learning for large language models. ICML 2024. [3]

2 Paper 1

Ebert et. al., RALF: A Reinforced Active Learning Formulation for Object Class Recognition CVPR 2012 [1]

2.1 What problem does this paper try to solve, i.e., its motivation

This paper addressed the challenge of balancing the selection of informative samples in pool-based active learning, where the entire dataset is treated as a pool, and samples are selected based on given criteria, for labeling by an oracle. The paper highlights the "exploitation-exploration dilemma" which often arises when only a single criterion is used for selecting examples from the data pool. The main issues with using either pure exploitative or pure explorative criteria can be summarized as follows:

- **Pure exploitative criteria** - These are often implemented using **uncertainty-based metrics**. These metrics tend to prioritize examples that the model is uncertain of i.e. those that are difficult to learn are ranked much higher. Thus, a pure exploitative strategy often leads to **sampling bias** and does not provide enough coverage of the entire data space since the model only focuses on outliers. This becomes more pronounced in multi-class scenarios where the dataset can be imbalanced and some classes are requested more while other classes are ignored. This problem also becomes prominent on more challenging datasets when examples in the same class are spatially separated into dense regions in the latent space.
- **Pure explorative criteria** - These are often implemented using **density-based metrics** leveraging the underlying unlabeled data distribution, using clustering or linear reconstructions, to find the most representative samples. This approach, although samples from the entire data space, ignores feedback during the labeling process, and thus needs many iterations i.e. too many labeling requests before a good decision boundary is found.

Prior methods have attempted to solve this problem by combining different criteria. The main problems with these approaches pointed out by the authors are as follows:

- These typically combine only two criteria and are difficult to balance effectively in practice.
- The combination of these criteria is usually fixed, instead of time-varying. Thus, they lack the flexibility to adapt to time-varying trade-offs.
- These methods do not generalize well to multi-class scenarios.
- Prior methods mostly overlook the incorporation of feedback from the classifier to learn from previous AL rounds to improve subsequent sample selections and label requests.
- They are limited in flexibility in accommodating more criteria and thus, the range of possible strategies that can be achieved.
- Many of these methods face computational challenges.
- The authors show that no single, pre-determined scheme works well across all datasets with different properties.

2.2 How does it solve the problems?

The paper addresses the issues listed above by proposing a reinforced active learning formulation (RALF) that considers the entire active learning as a meta-learning process that is optimized by learning a strategy from feedback "on the fly".

In the first part, the authors motivate the problem by discussing two exploitation criteria, namely **Entropy** and **Margin**, and three exploration criteria, namely **Node potential**, **Kernel farthest first** and **Graph density**. Graph density is introduced by the authors as a novel sampling criterion that uses a k-nearest neighbour graph to identify highly connected nodes. This is further discussed in the contributions Sec. 2.3.

Next, the paper describes the datasets and the classifiers used in the experiments. According to the increasing number of object classes, the datasets used are ETH-80 (ETH) with 8 classes, Cropped PASCAL (C-PASCAL) with 20 classes and Caltech 101 with 102 classes. The paper uses two classifiers, one semi-supervised and the other supervised, namely **Label Propagation (LP)** with k-NN graph and **Support Vector Machine (SVM)** with RBF kernel.

Multiple experiments are performed to analyze each single criterion and classifier pair, with random sampling from uniform distribution as the baseline. The paper also highlights that some datasets might need more exploration at the beginning and more exploitation at the end and vice versa, while others might need a constant trade-off. Thus, numerous experiments are also performed to analyze the fixed and time-varying combinations of exploration and exploitation criteria, with Eq. 10 in the paper being the AL framework. Essentially, different $\beta(t)$ such as $0.5, 0, 1, log(t), t$, etc. denote fixed, pure exploration, pure exploitation, and time-varying strategies respectively. These experiments lead to the following major observations:

- LP is always better than SVM, so subsequent experiments are done only on LP.
- Exploitation criteria work better than exploration criteria due to local feedback after each iteration.
- The novel criteria, Graph density, works best for all datasets in combination with Entropy or Margin criteria.
- No single, pre-determined scheme, whether single criteria or time-varying combinations, works well across all datasets with different properties.

Finally, motivated by these experimental results, the paper aims at modeling the progress of the learnt classifier and uses it to control trade-offs between criteria. This is the main difference from earlier works which overlooked the feedback from the classifier. It formulates the active learning process as a feedback-driven **Markov decision process** (MDP) which uses **Q-learning** to learn

a transition table where the states are the different combinations of two criteria with n actions representing n fixed trade-offs. The authors claim it is still time-varying since it can switch between different actions. The authors have chosen Q-learning since it is model-free and computationally fast. The initialization problem, where the method starts with an empty Q table, is solved using a guided initialization phase which uses a prediction probability-weighted entropy aggregated over each label j for each sample x_i. This is used to select the next action.

2.3 A list of novelties/contributions

The following are the main contributions of the paper:

- The paper proposes a novel exploration-based sampling criteria called **Graph density** leveraging a k-NN graph with Manhattan distance (and $k = 10$) which is weighted with a Gaussian kernel and normalized by number of edges, and used to rank the data points corresponding to the representatives. The main intuition is that each class representative is highly embedded in the graph and, thus is well-connected having many edges. The normalization and weight reduction of direct neighbors avoid over-sampling dense regions or outliers. It is more robust than the Node potential criteria due the underlying k-NN graph structure.
- The paper empirically shows that no single criterion works best across all datasets.
- The paper integrates the probability of switching between exploration and exploitation into the already parameterized time-varying AL framework so that there is always a mix of two criteria.
- Since previous reward rescaling functions do not generalize well to new datasets, the paper proposes a new rescaling function to map all previously observed rewards in the interval $[-1, 1]$.
- The authors also propose a new reward function using the difference in the overall entropies between the two time steps. This is closer to the learning progress of the classifier instead of just a change in the prediction.
- The main AL process is formulated as a Markov decision process which is learnt using the Q-learning method. This is a model-free and computationally efficient RL algorithm. There are only two parameters for Q-learning and no dataset-specific tuning is needed.
- To tackle the challenge of initialization with empty Q table and to avoid the computationally method of visiting each state-action-pair, the authors propose a guided initialization phase.

2.4 What do you think are the downsides of the work?

The authors run several experiments to show that their novel sampling criterion works best compared to other explorative criteria and that time-varying strategies are better than fixed ones. Their proposed MDP-based method, RALF, outperforms the compared methods and works well in the multi-class scenario for three different datasets. However, there are a few downsides to this work.

- Kernel and classifier - The authors choose a k-NN-based label propagation (LP) and RBF-based SVM classifier and show that LP consistently does better than SVM and all subsequent experiments are thus done only on LP. It is not clear from the experiments if the choice of kernel affects the classification performance and the AL process. An additional RBF-based LP could have been included to solve this ambiguity. Also, the results of subsequent experiments might differ with a different classifier.

- Task and dataset - The authors choose three different multi-class datasets with varying difficulty. However, all experiments show the out-performance of the method only on a single task of multi-class object classification. It is not obvious if the proposed method would generalize to other tasks and domains, either in computer vision, NLP or time series data.
- Limited sampling criteria - The parameterization of the AL framework to balance between explorative and exploitative sampling limits the strategies to only a mixture of two criteria, one from each category.
- Fixed trade-offs - The authors also use n fixed trade-offs in $[0, 1]$ and argue that it is still time-varying since it can switch between different actions. However, the range of possible time-varying strategies still becomes limited.
- RL algorithm - The authors choose Q-Learning for learning the strategies. As the authors point out, Q-learning is computationally efficient only when the number of states and actions are kept small to speed up initialization, thus defeating the flexibility with a much lesser range of strategies.

Minor point - U is overloaded, i.e. in Sec. 3.1, U denotes the set of all unlabeled examples and in Sec. 6, U denotes the set of all uncertainty-based criteria.

3 Paper 2

Konyushkova et. al., Learning Active Learning from Data. NeurIPS 2017. [2]

3.1 What problem does this paper try to solve, i.e., its motivation

The main problems this paper tries to solve can be summarized as follows:

- Prior active learning methods use a combination of hand-crafted heuristics and are restricted to these existing techniques.
- The assessment of the classification performance of previous AL methods is unreliable as it highly depends on the scarce annotated data.
- The paper also highlights the problem of uncertainty sampling (US), the most popular sampling criterion, in unbalanced datasets. The more imbalanced the classes are in a dataset, the further from the optimum choice made by US. Although query selection procedures can account for statistical properties of datasets and classifiers, for complex scenarios with many factors such as label noise, outliers and shape of distribution, there is no easy way to take into account all possible factors.

3.2 How does it solve the problem?

This paper attempts to solve the above problems using two features.

- They look at a continuum of AL strategies instead of combinations of pre-specified heuristics.
- They avoid the need for performance evaluation of the classification quality for application-specific data as their approach can learn from previous tasks and easily transfer strategies to new domains.

To achieve this, the authors propose a data-driven approach and formulate Learning Active Learning (LAL) as a regression problem. Specifically, given a trained classifier and its output for an unlabeled sample, they predict the reduction in generalization error if that label is added to that data point. They show that this regression function can be trained on synthetic data by using simple features.

These features may be classifier output variance or predicted probability distribution over the labels for a data point. Since these features are not domain-specific, it means that a regressor trained on synthetic data can be directly applied to other classification tasks.

The authors use uncertainty sampling (US) as it is the most popular and widely applicable sampling criterion. To motivate the need for data-driven approaches to improve AL strategies and to deal with scenarios where US fails, the authors present two toy examples with balanced and unbalanced classes. For balanced datasets, US is the best greedy approach. However, they show that US by design becomes sub-optimal for the imbalanced case since the data point that corresponds to the largest expected error reduction is different from 0.5 for the unbalanced class with much more data than the other.

The paper formulates AL as a data-driven Monte Carlo technique and proposes two approaches to constructing the datasets for training the regressor. The proposed method is quite fast during the online AL steps.

- **Independent LAL** strategy incorporates unused labels at random to retrain the regressor to correlate the change in test performance with classifier and new data point properties. Specifically, the classifier is characterized by K parameters and any new randomly selected data point is characterized by R parameters. These two sets of parameters form the characterization vector ξ of the learning state for the classifier-data point pair. The difference between the losses ($\delta_x = l_\tau - l_x$) for the classifiers f_τ and f_x, trained without and with the new randomly selected data point is recorded for each learning state. This is repeated for Q different initializations, for T different labeled subset sizes each with M different data points. Thus, a dataset $\Xi \in \mathbb{R}^{(QMT)\times(K+R)}$ is created and a regressor can be trained to learn the mapping from learning state $\xi \in \mathbb{R}^{K+R}$ to expected error reduction δ. Thus, the LALIndependent method greedily looks for samples that have the highest potential to reduce the classifier error at each time step of the AL process.

- **Iterative LAL** accounts for selection bias in AL by simulating the AL procedure and data point selection considers the strategy learnt on previously collected data. Thus, in each iteration selection of the most promising sample depends on the samples and strategies in the previous iteration. Hence, the final strategy in AL learns the sampling bias represented in the data.

Experiments are conducted for LALIndependent and LALIterative with (a) cold start with one sample from each class, and (b) warm start with a larger dataset. The synthetic datasets used are (a) 2D Gaussian point clouds and (b) XOR-like data. Once the regressors are trained on these synthetic datasets, they are tested on real-world data. The authors use a Gaussian Process classifier (GPC) and Random Forest (RF). The two LAL methods with cold start on 2D synthetic outperform baseline methods of random (RS) and uncertainty sampling (US) and previous AL methods, Kapoor (citation) and ALBE (citation) in both speed and accuracy. Subsequent experiments are done only on RF due to computational cost. The proposed methods trained on synthetic datasets do remarkably well on real data like the Striatum, MRI and Credit card datasets and better than RS, US and ALBE. The LALIndependent with warm start is applied on Splice and Higgs datasets and outperforms RS, US and ALBE. The paper reports a variety of metrics showing the robustness of the method to choice of loss function for measuring error reduction.

3.3 A list of novelties/contributions

The following are the main contributions of the paper:

- The paper empirically shows that uncertainty sampling (US), which is the most popular and widely used exploitative sampling method, selects sub-optimal samples for imbalanced datasets. The more the imbalance in the classes, the further from the optimum the choice is made by US.

- The paper formulates the learning AL as a regression problem and shows that the error reduction for the addition of a new labeled data point can be predicted by a regressor, given the properties of the classifier and that data point.
- The classifier can be characterized by simple features such as kernel parameters for kernel-based, average depths of trees for tree-based or prediction variability for ensemble classifier. The data points can also be characterized by simple features like predicted probability for a class, distance to the closest point in dataset, distance to the closest labeled point, etc.
- The authors show that the generalization capability AL regressor from synthetic to real data. It can be trained using 2D synthetic datasets and this can be applied to a warm start and further tuning on real-world datasets for both binary classification and segmentation with remarkable performance.

3.4 What do you think are the downsides of the work?

Although the paper highlights some key challenges in previous AL strategies and sampling methods and formulates a data-driven regression problem to tackle these issues, there are a few drawbacks of the proposed method as follows.

- Task and dataset - The experiments are performed only on binary classification and binary segmentation datasets. It is unclear if the proposed method would generalize to multi-class datasets. Also, the chosen datasets are mostly image or numerical data and further experiments may be needed to show efficacy on text, video or time-series datasets.
- Hand-designed features for classifier and data point - The hand-crafted features used to characterize the classifier and the data point under consideration need careful design decisions and are time-consuming to come up with. In Sec. 5, the paper lists six features which are used. These features constitute the learning state and can heavily influence the training of the regressor. This defeats the prior claims of avoiding hand-crafted heuristics.
- RF Classifier - The paper mostly reports results for the RF classifier for task model with RF regressor for learning AL with only the first experiment performed on GPC and RF, stating computational reasons. This creates an ambiguity in whether the choice of the classifier influences the performance of the regressor. It is also not obvious how this classifier can be generalized to more complex image segmentation or text models.

4 Paper 3

Muldrew et. al., Active preference learning for large language models. ICML 2024. [3]

4.1 What problem does this paper try to solve, i.e., its motivation

The fine-tuning techniques for aligning large language models (LLMs) require careful consideration and effective use of human resources where reinforcement learning is used (RLHF). Prior works have also used feedback from AI models (RLAIF) to align smaller language models. However, these methods are quite complex and unstable. Direct Preference Optimization (DPO) is a much simpler and more stable technique for aligning LLMs. Methods have also been developed using active learning to improve fine-tuning LLMs. The main problems with these previous methods of fine-tuning and aligning LLMs using active learning are as follows:

- Relying only on human feedback becomes increasingly unviable in today's era where language models have become extremely large and fine-tuning requires a lot of data.

- Methods relying on an AI model as an oracle for generating responses tend to consult it for every data point which becomes computationally inefficient.
- Prior methods have explored RLHF/RLAIF for active learning which are complex and unstable.
- Many AL sampling criteria are not straight-forward and might require modifications to the model architecture and fine-tuning process itself.

4.2 How does it solve the problem?

This paper aims to solve the above problems in the following manner:

- The paper leverages the DPO technique within the active learning process to make the fine-tuning process much simpler and more stable. The need for a separate reward model and multi-stage process to adapt to the autoregressive LLM is eliminated in DPO.
- The authors propose a few different acquisition functions. First, predictive entropy (PE) is a widely used measure of uncertainty in LLMs and it has previously been shown to be well calibrated for LLMs. Second, the preference model certainty under the Bradley-Terry model captures the oracle's preferences better than PE. The proposed function is maximised when the difference between the implicit rewards for the two generations is large and vice versa. Lastly, these two approaches are complementary and can be combined into a hybrid approach. The authors propose selecting a relatively large batch of prompts ranked by entropy. Then, only the top subset is used for generating prompt/completion pairs. Finally, the pairs are scored and ranked according to preference certainty. This filtering step minimizes the number of oracle consultations required for generation.
- The paper shows that GPT-4 is much more self-consistent than GPT-3.5 and thus selects GPT-4 as the oracle despite the high cost and latency.
- The main LLMs fine-tuned using active learning are GPT-2 and Pythia (GPT-3-like) models. The pre-trained versions of these models were obtained from Hugging Face.
- In the fine-tuning step, the authors use a straight-forward implementation of re-initialization, uniform sampling from the set of all previously sampled data and fine-tuning to convergence. The authors show that the proposed acquisition functions perform much better than the baseline of random sampling.
- The analysis of the histogram of the results shows that there is a much better differentiation for preference certainty and hybrid approach with the random sampling baseline.

4.3 A list of novelties/contributions

The main contributions of the paper are as follows:

- The paper introduces the combination of the DPO technique and active learning to simplify and stabilize the complex and data-hungry steps of fine-tuning of autoregressive LLMs. This eliminates the need for a separate reward model as in the case of RLHF/RLAIF base techniques.
- The authors propose a hybrid acquisition function combining predictive entropy and preference certainty. The filtering step using entropy reduces the number of requests to the oracle and the preference certainty captures the oracle preference much better.

4.4 What do you think are the downsides of the work?

Although the paper highlights the pressing challenge of fine-tuning LLMs with less data and proposes an AL method leveraging DPO to tackle this challenge, there are a few drawbacks of the proposed method as follows.

- Dataset - The paper only focuses on text generation tasks on generic text datasets. The first dataset is a text completion task for movie reviews and the second is a summarization of Reddit posts. Further analysis is required to show that the method works for domain-specific datasets and tasks.
- Model - The paper only runs experiments on GPT-2 and Pythia (GPT-3-like) models. These are mainly decoder-based models. For the sake of completeness, it would be interesting to experiment with encoder-decoder models too.
- Closed model as oracle - Since the paper mainly uses GPT-4 as the oracle for generating answers, the effects of the oracle cannot be studied very well. Also, closed models come at a much higher cost which might be economical for running multiple iterations of active learning. Also, the paper compares the choice of oracle by measuring the average self-consistency of only GPT-3.5 and GPT-4. However, a comparison including open models would have been better.
- Prompts - The authors only provide two prompts for the two datasets and thus the effect of prompt engineering is not well-studied.
- Comparative Analysis - The paper mainly compares the proposed acquisition functions predictive entropy, preference certainty and hybrid (PE + PC) with the baseline of random sampling. No comparison with other widely used sampling strategies is shown. The paper points out that the Reward rAnked Fine-Tuning (RaFT) technique consults the oracle on every data point before filtering. However, no comparative study has been done with previous AL methods for fine-tuning LLMs.

Despite some drawbacks, the paper proposes a novel method of leveraging Direct-Preference Optimization (DPO) for active learning of the fine-tuning of LLMs.

References

[1] Sandra Ebert, Mario Fritz, and Bernt Schiele. Ralf: A reinforced active learning formulation for object class recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633. IEEE, 2012.

[2] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[3] William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In *Forty-first International Conference on Machine Learning*, 2024.