Krishnasai Paleti
kxp5619

CSE 584 – Homework 1

Paper 1: Improving Generalization with Active Learning

Authors: DAVID COHN, LES ATLAS, RICHARD LADNER

Year: 1994 (possibly the first paper on Active Learning)

Motivation:

This paper addresses the limitations of the traditional learning methods from examples, specifically within neural networks. The main motivation is to improve generalization with a limited number of training examples by introducing a more efficient learning process called active learning.

In traditional learing by examples, the algorithm passively receives randomly selected examples from the input domain, leading to inefficiencies. This method often requires many examples to achieve good generalization performance

The main motivation of this paper is to explore a method where the learning algorithm has control over the input it receives, querying specific examples that provide the most useful information for learning. This approach, known as selective sampling, enables the model to focus on regions of uncertainty in the input space, thereby reducing the number of examples required to achieve high generalization accuracy.

Approach:

Instead of passively accepting random examples, the learning algorithm actively selects which parts of the input domain to query based on where it has the most uncertainty. By doing so, it focuses on the examples that provide the most information for improving its understanding of the problem space.

Selective Sampling is done by identifying the region of uncertainty, where the model is most likely to misclassify new examples. The learner focuses on querying data points from this region, effectively reducing the overall uncertainty in the model's predictions.

S-G network is proposed, The idea is to modify the standard neural network learning algorithm to include selective sampling. The network is trained in such a way that it only queries examples from the region of uncertainty, thereby refining its generalization ability more efficiently.

The paper also discusses a version space, here the model maintains a "most specific" and "most general" hypothesis about the target concept and reduces this version space by querying points that fall in the disagreement between these two hypotheses.

To reduce computational costs, the paper proposes batch sampling.

Novel Contributions:

- Selective Sampling as an Active Learning Method
- SG-Network for Selective Sampling
- Version Space Search in Neural Networks
- Approximation Techniques for Region of Uncertainty
- Batch-Based Sampling Strategy

Downsides:

- Implementing selective sampling requires recalculating the region of uncertainty, which can be computationally expensive. The cost of updating the region of uncertainty after every batch still adds overhead. This tradeoff between reducing generalization error and computational efficiency may not always be justified in practice, especially for larger datasets.
- In high-dimensional domains or complex concept classes, the region of uncertainty may extend over large portions of the input space, particularly in the early stages of learning. This could turn into random sampling.
- Selective sampling is more efficient when the underlying data distribution is known, which is not always realistic in many practical scenarios
- The experiments and domains used in the paper are controlled. Noisy data scenario is not discussed in detail.

Krishnasai Paleti
kxp5619

Paper 2: Active Learning by Learning

Authors: Wei-Ning Hsu, Hsuan-Tien Lin

Year: 2015

Motivation:

The motivation of the paper is the problem of inefficiency in **pool-based active learning**, where machines aim to reduce labeling efforts by strategically selecting which data points to query for labeling. Traditionally, most active learning strategies are based on human-designed philosophies, which are created with specific assumptions about what constitutes a "good" data point to label. While these strategies can work well in some cases, they often fail in others because no single human-designed approach can consistently handle the diverse characteristics of different datasets. *(This was a downside of paper 1, addressed here).*

The main problem the paper tries to solve is how to automatically and adaptively select and combine different active learning strategies in a way that improves overall performance across diverse datasets, without relying on fixed, human-designed strategies. The proposed solution is to let the machine "learn" which strategy works best on the fly, by applying techniques from the multi-armed bandit problem.

Approach:

Introduces a novel approach called Active Learning by Learning (ALBL), which adaptively selects and blends different active learning strategies based on their performance on the current dataset.

- The multi-armed bandit problem is used as the foundation. In this analogy, each active learning strategy (or algorithm) is treated as a "bandit machine" in a casino, where each machine gives a random reward. The challenge is to balance exploration (trying different strategies) and exploitation (using the strategy that has been most successful so far). In ALBL, each active learning algorithm is a "bandit machine," and the goal is to choose which strategy to "pull" (use) in each iteration to maximize learning performance.
- ALBL continuously monitors the performance of multiple existing active learning strategies and adjusts its selection dynamically. Instead of committing to one static

approach, ALBL can switch strategies based on how well each performs over time. The EXP4.P algorithm (a variant of multi-armed bandit solvers) is used to manage the selection of active learning strategies

- To evaluate the performance of each strategy in real-time, ALBL introduces a novel reward function based on importance-weighted accuracy (IW-ACC). This reward function estimates how much each strategy helps improve the learning model by re-weighting the accuracy on the sampled data.
- ALBL intelligently blends multiple strategies instead of choosing just one at any time. The probability of selecting each strategy is adjusted dynamically based on its estimated contribution to the learning performance. This probabilistic blending allows ALBL to adapt over time and switch between strategies as necessary.
- ALBL includes a random sampling strategy as a fallback option. This ensures that even when all human-designed strategies fail, the model can still resort to a basic random strategy, providing a robust solution across various scenarios.

Novel Contributions:

- ALBL framework for dynamically learning which active learning strategy to use.
- Novel connection between active learning and the multi-armed bandit problem.
- Use of EXP4.P for adaptive strategy selection and modification of its reward update.
- Introduction of Importance-Weighted Accuracy (IW-ACC) as a reward function.
- Inclusion of a random fallback strategy to handle extreme cases.

Downsides:

- The paper focuses on pool-based active learning, where the learner has access to both labeled and unlabeled pools of data. However, the method's scalability to very large datasets or streaming data scenarios is not discussed in detail.
- The paper does not discuss how the ALBL approach performs on imbalanced datasets, where the distribution of classes is highly skewed. In active learning, imbalanced data can be especially problematic because the querying strategy may focus too much on one class or fail to uncover rare but important examples.
- The paper's evaluation focuses on academic benchmarks (e.g., UCI datasets) and shows good performance on these datasets. However, there is limited discussion on the practical challenges of deploying ALBL in real-world scenarios, such as noisy labeling, evolving data distributions, or time-varying availability of data.

Krishnasai Paleti
kxp5619

Krishnasai Paleti
kxp5619

Paper 3: Compute-Efficient Active Learning

Authors: G ´abor N ´emeth, Tam ´as Matuszka

Year: 2023

Motivation:

The paper addresses the problems from the previous two papers (paper1 and paper 2) in this report of high computational costs in traditional active learning methods, especially when dealing with large-scale datasets. Active learning aims to reduce the labeling costs by selecting the most informative data samples for annotation, but this process itself can be computationally expensive due to the need to evaluate large amounts of unlabeled data.

The motivation behind the paper is to create a compute-efficient active learning framework that reduces the number of acquisition function evaluations (used to determine the importance of unlabeled samples) without sacrificing model performance. By leveraging historical values of the acquisition function to guide the selection of data points, the method aims to reduce the overall computational burden while maintaining or even improving the effectiveness of active learning.

Approach:

The paper proposes a compute-efficient active learning framework to solve the problem of high computational costs in traditional active learning.

The method Historical Acquisition Function Values leverages the historical values of the acquisition function (which ranks the importance of unlabeled data points). The assumption is that these historical values are good predictors of future importance, so not every data point needs to be evaluated in every iteration.

Instead of evaluating all the unlabeled data points in every iteration, the framework subsamples a smaller candidate pool based on the acquisition function's past evaluations. This smaller pool contains the most likely important samples, reducing the number of points that need to be processed.

The algorithm uses a softmax function to convert acquisition function values into probabilities, assigning higher probabilities to more informative samples. These are then selected into the candidate pool for further evaluation.

The acquisition function is only updated for the candidate pool rather than the entire unlabeled dataset. This reduces the computational load, as fewer points need to be re-evaluated. The framework selectively annotates the data points that have the most potential to improve model performance

Novel Contributions:

- A novel active learning framework designed to reduce computational costs by leveraging historical acquisition function values.
- The paper introduces the idea that historical acquisition function values are good predictors of future values.
- Subsampling strategy that prioritizes more informative samples for further evaluation and labeling, based on softmax probabilities derived from acquisition function values.
- The framework is designed to be method-agnostic, meaning it can work with various acquisition functions.
- The proposed method demonstrates that it can reduce computational demands (up to 25% less training time in experiments) while maintaining or even surpassing model performance compared to traditional active learning methods.

Downsides:

- The method assumes that historical acquisition function values are good predictors of future values. While this works well in many cases, it may not always hold, particularly in rapidly changing models where importance of data points evolves significantly as the model learns.
- Although the method is tested on MNIST and CIFAR-10, which are standard benchmark datasets, these datasets are relatively simple compared to more complex real-world datasets. The paper acknowledges this limitation and suggests further experiments on larger, more complex datasets are necessary to validate the method's generalizability
- The generalizability of the method across more diverse types of data, such as text or time series, has not been addressed.