Active Learning for Probabilistic Hypothesis Using the Maximum Gibbs Error Criterion

Cuong, Nguyen Viet, et al. "Active learning for probabilistic hypotheses using the maximum Gibbs error criterion." *Advances in Neural Information Processing Systems* 26 (2013).

What problem does this paper try to solve?

This paper talks about a new objective function, called *the policy Gibbs error*, for Bayesian active learning with probabilistic hypotheses. The objective function is the expected error rate of a random classifier drawn from the prior distribution selected by the active learning policy. It can be formulated as a maximum coverage objective with a fixed budget. In other words, with a budget of k queries, we aim to select k examples such that the *policy Gibbs error* is maximal. This maximality implies the minimality of the posterior label entropy of the remaining unlabeled examples in the pool. Specifically, it aims to improve the selection of training data from unlabeled examples in a way that maximizes the generalization performance of a classifier while minimizing the number of labeled examples required.

How does it solve the problem?

Exact maximization of this function is hard, so it proposes a greedy strategy, called **maxGEC(maximum Gibbs Error Criterion)**, which selects examples that maximize the Gibbs error at each iteration to maximize this criterion and apply it to three active learning setting; non-adaptive, adaptive, and batch active learning.

- Non-adaptive: The set of examples is not labeled until all examples in the set have all been selected.
- Adaptive: The examples are labeled as soon as they are selected, and the new information is used to select the next example.

- Batch: Select a batch of examples, query their labels and proceed to select the next batch.

List of novelties/contributions

- It introduces a new objective function for Bayesian active learning called the policy Gibbs error.
- It proposes a greedy strategy called maxGEC to solve the problem.
- It provides mathematical proofs showing the near-optimality of the policy Gibbs error for each setting.
- It discusses how to compute maxGEC for some probabilistic models such as the Bayesian conditional exponential model, and the Bayesian transductive Naive Bayes.
- It shows some experimental results on NER(Named Entity Recognition) and text classification.

Downsides of the work

- Restrictions in size of batches: In non-adaptive and batch settings, since the proposed algorithms need to sum over all labels of the previously selected examples in a batch, it restricts the algorithms to small batches.
- For the case of noisy data, the formulation is simpler than other techniques such as EC^2 and the generalized binary search but the proposed algorithms are limited by computational concerns.