Active Learning Literature Survey

Abstract

- The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the data from which is learns
- 在主动学习（Active Learning）领域，其关键在于如何选择出合适的标注候选集给人工进行标注，而选择的方法就是所谓的查询策略（Query Strategy）。查询策略基本上可以基于单个机器学习模型，也可以基于多个机器学习模型，在实际使用的时候可以根据情况来决定。整体来看，主动学习都是为了降低标注成本，迅速提升模型效果而存在的

Introduction

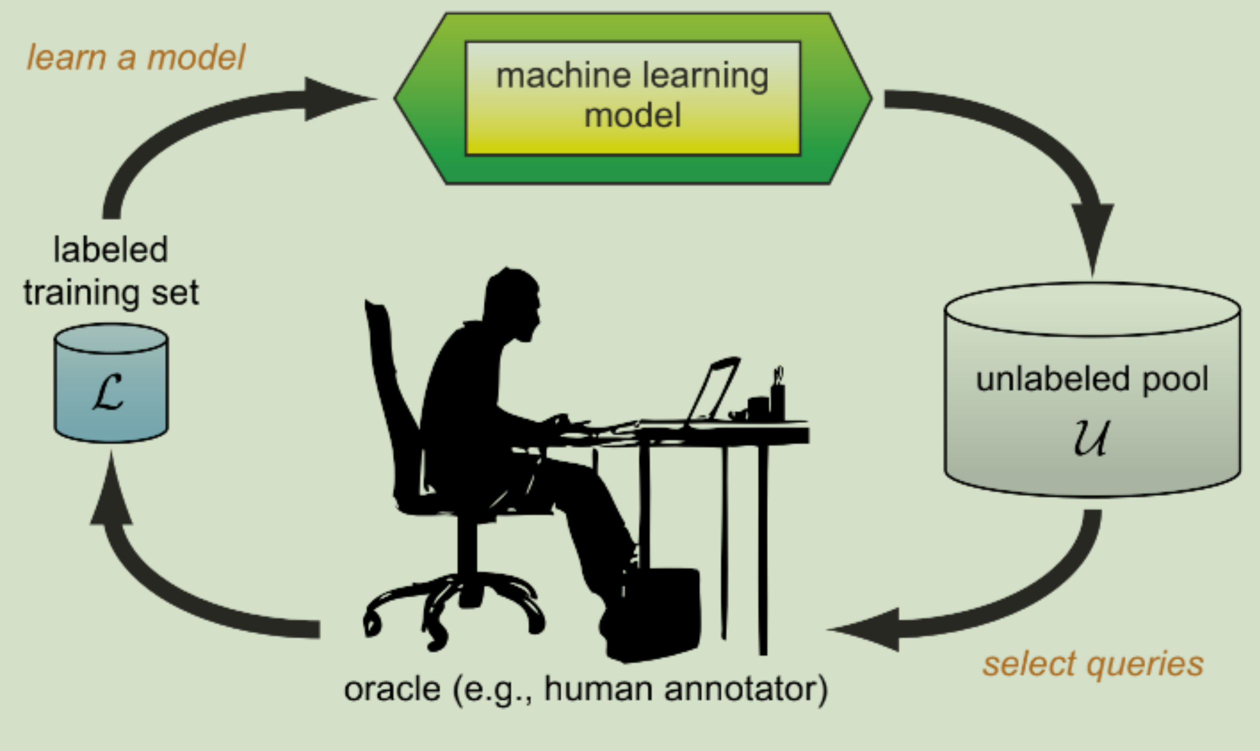
- What is Active Learning
  - Active learning (also called “**query learning**,” or sometimes “optimal experimental design” in the statistics literature)
  - **The key hypothesis is that, if the learning algorithm is allowed to choose the data from which it learns, it will perform better with less training**
- Active Learning Examples
  - 

    Figure 1: The pool-based active learning cycle.
  - A learner may begin with a small number of instances in the labeled training set L, request labels for one or more carefully selected instances, learn from the query results, and **then leverage its new knowledge to choose which instances to query next**

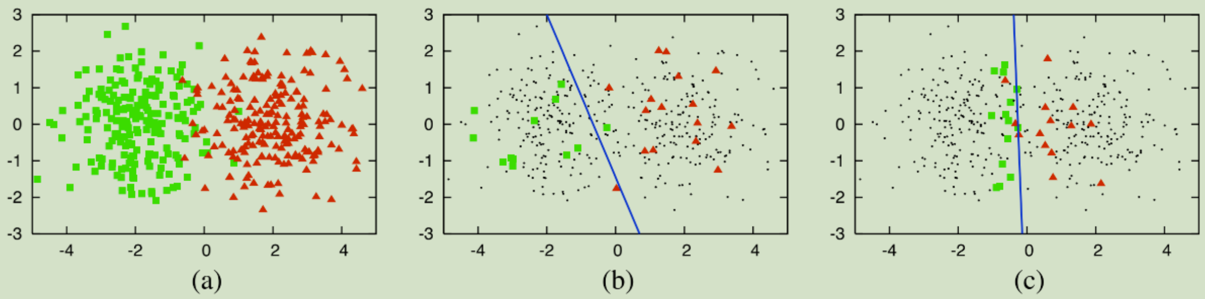
- 

  Figure 2: An illustrative example of pool-based active learning. (a) A toy data set of 400 instances, evenly sampled from two class Gaussians. The instances are represented as points in a 2D feature space. (b) A logistic regression model trained with 30 labeled instances randomly drawn from the problem domain. The line represents the decision boundary of the classifier (accuracy = 0.7). (c) A logistic regression model trained with 30 actively queried instances using uncertainty sampling (accuracy = 0.9).
- **用少量的样本，把握大量数据上体现的分布，这是最核心的思想**

Query Scenarios

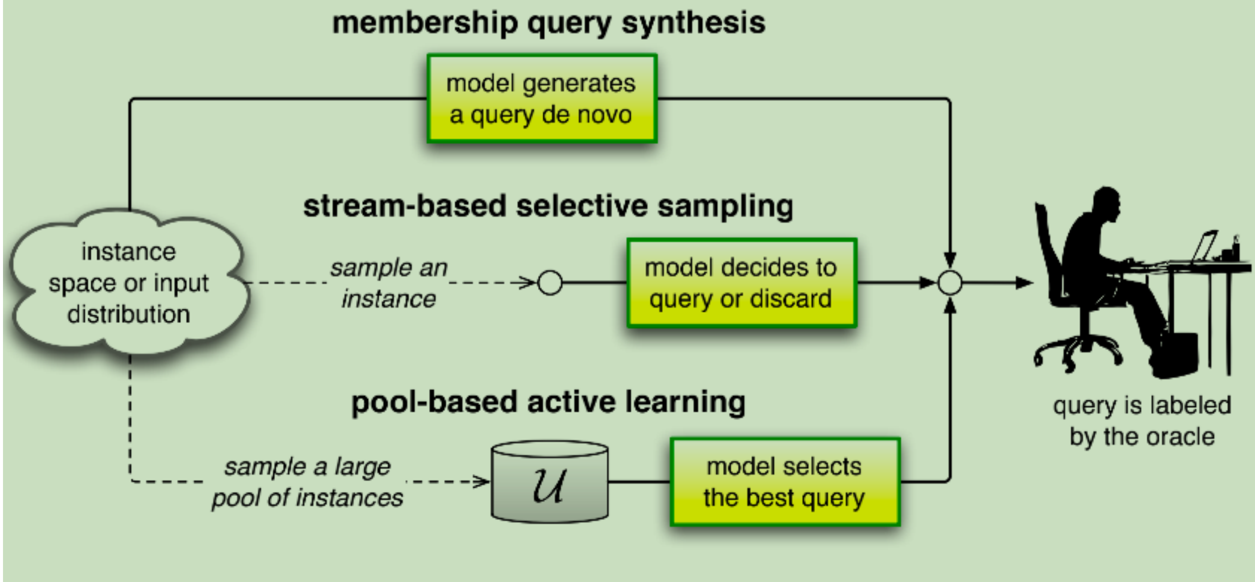
- 三种主动学习方式
  - Membership query synthesis
  - Stream-based selective sampling
  - Pool-based active learning
  - 
membership query synthesis
model generates a query de novo
stream-based selective sampling
instance space or input distribution
sample an instance
model decides to query or discard
pool-based active learning
sample a large pool of instances
$\mathcal{U}$
model selects the best query
query is labeled by the oracle

    Figure 4: Diagram illustrating the three main active learning scenarios.
- Membership query synthesis
  - In this setting, the learner may request labels for any unlabeled instance in the input space, including (and typically assuming) **queries that the learner generates de novo, rather than those sampled from some underlying natural distribution**

    - 查询或者说问题是学习者【学习器】生成的，而不是从某种查询中采样得到的
- Stream-based selective sampling
  - This approach is sometimes called stream-based or sequential active learning, as each unlabeled instance is typically
    drawn one at a time from the data source, and the learner must decide whether to query or discard it
  - If the input distribution is uniform, selective sampling may well behave like membership query learning
  - **However, if the distribution is nonuniform and (more importantly) unknown, we are guaranteed that queries will still be sensible, since they come from a real underlying distribution**
- Pool-based active learning
  - Which assumes that there is a small set of labeled data *L* and a large pool of unlabeled data *U* available

Query Strategy Frameworks

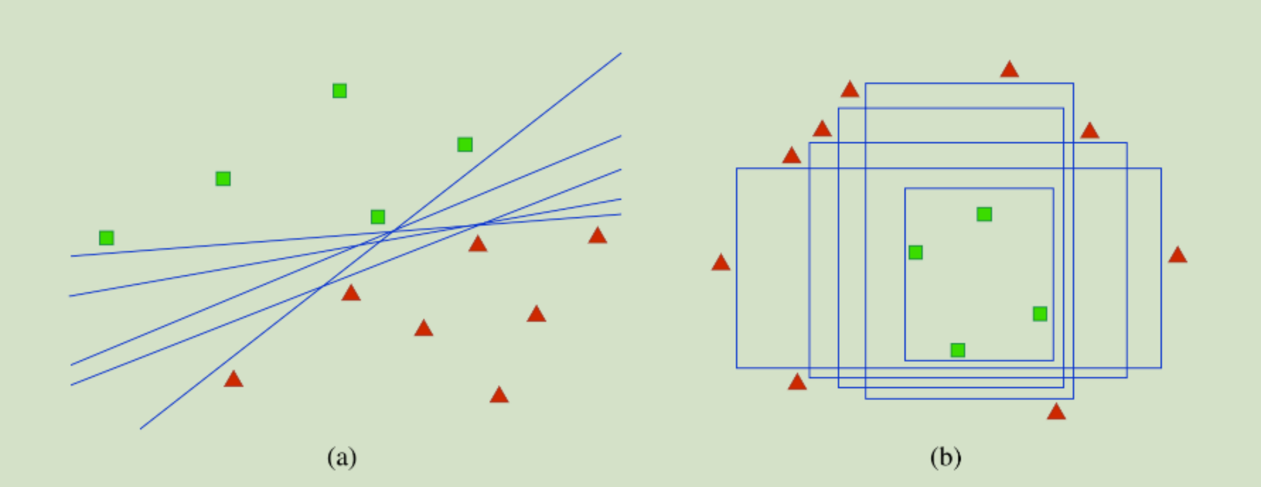
- Uncertainty Sampling
  - In this framework, an active learner queries the instances about which it is least certain how to label
  - Least Confident
  - Margin Sampling
  - Entropy
- Query-By-Committee
  - The QBC approach involves maintaining a committee C = {θ(1), . . . , θ(C)} of models which are all trained onthe current labeled set L, but represent competing hypotheses
  - 

    Figure 5: Version space examples for (a) linear and (b) axis-parallel box classifiers. All hypotheses are consistent with the labeled training data in $\mathcal{L}$ (as indicated by shaded polygons), but each represents a different model in the version space.
- Expected Model Change
  - Another general active learning framework is to query the instance that would impart the greatest

change to the current model if we knew its label. An example query strategy in this framework is the “expected gradient length” (EGL) approach
for discriminative probabilistic model classes

- Variance Reduction and Fisher Information Ratio
- Estimated Error Reduction
- Density-Weighted Methods

Analysis of Active Learning

- Empirical Analysis
    - 大多数情况是有用的
    - 模型更新改变【大改】则会失效
- Theoretical Analysis
    - 这个比较复杂，需要继续看一些资料

Problem Setting Variants

- Active Learning for Structured Outputs
    - Active learning for classification tasks has been widely studied. However, many important learning problems involve predicting structured outputs on instances, **such as sequences and trees**
    - 对于信息抽取任务，主动学习有一定的困难，需要再看看
- Batch-Mode Active Learning

Related Research Areas

- Research in active learning is driven by two key ideas
    - The learner should be allowed to ask questions
    - Unlabeled data are often readily available or easily obtained
- Semi-Supervised Learning
- Reinforcement Learning