主动学习的问题和思考【持续补充】

一、假设、定义和目标

假设

- 数据层
    - 抽象假设
        - 不同的数据样本对于某项具体任务的作用和价值是不一样的
    - 具象假设
        - 数据密度说
            - 用少量的样本，把握大量数据上的分布
- 模型层
    - 抽象假设
        - 模型或者算法可以利用已经学到的一些知识，选择出对提升该模型效果有利的样本，从而减少训练数据量
    - 具象假设
        - 深度学习模型由于其复杂的结构，具有很强的学习容量

定义

- 通过机器学习的方法获取到那些比较**“难”**分类的样本数据，让人工再次确认和审核，然后将人工标注得到的数据再次使用有监督学习模型或者半监督学习模型进行训练，逐步提升模型的效果，将人工经验融入机器学习的模型中
    - 这里所谓的**难**，需要根据不同的情况去定义

目标

- 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

二、核心要素和结构

核心要素

- 实体要素
    - Labeled Dateset
    - Temp Model
    - Unlabeled Dataset
    - Selected Dataset
    - Annotator

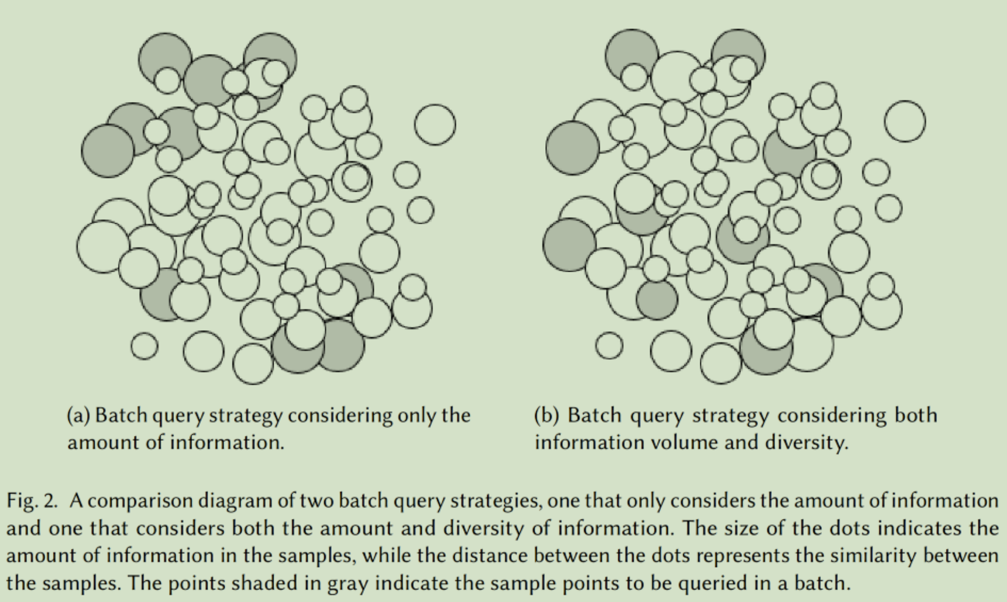
- 关系要素
  - Labeled Dateset To Temp Model
    - 无特殊关系，Task-specific
  - Temp Model To Unlabeled Dataset
    - 利用模型去抽取大数据集上的特征，一般是某种 Embedding
  - Unlabeled Dataset To Selected Dataset
    - Query Strategy，这个一般是 Task-specific
  - Selected Dataset
    - To Annotator
      - 无特殊情况，Task-specific
    - To Labeled Dateset
      - Query Strategy Specific
  - 停止闭环要素

结构

- 取决于目标
  - 通过某种手段，降低标注成本，提高模型效果
- 组织和方向
  - 单向组织

三、工具组件

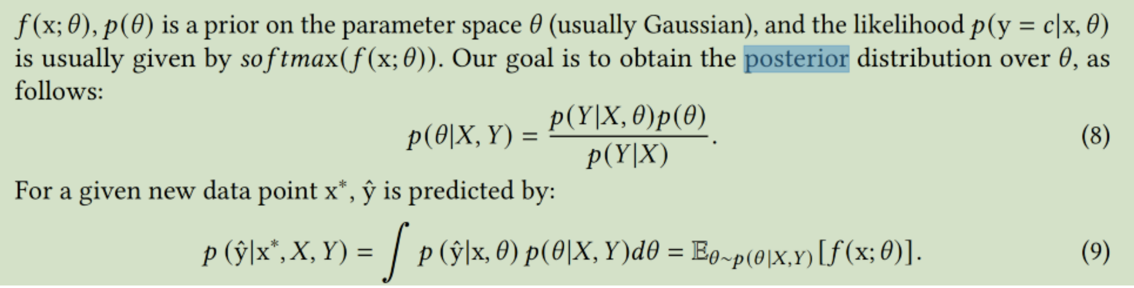
- Temp Model To Unlabeled Dataset
  - 蒸馏小模型来替代大模型进行特征映射
- Unlabeled Dataset To Selected Dataset
  - Batch Mode DeepAL (BMDAL)
    - 需要确认待标数据之间的互信息，这样可以考虑到分布情况
  - Uncertainty-based and Hybrid Query Strategies
    - Batch Mode DeepAL (BMDAL)
      - The batch-based query strategy is the foundation of DeepAL
      - The one-by-one sample query strategy in traditional AL is inefficient and not applicable to DeepAL, so it is replaced by batch-based query strategy
        - In traditional AL, most algorithms use a one-by-one query method, which leads to frequent training of the learning model but little change in the training data
      - **需要确认待标数据之间的互信息，这样可以考虑到分布情况**
    - Uncertainty-based and Hybrid Query Strategies
      - CEAL (Cost-Effective Active Learning) enriches the training set by assigning pseudo-labels to samples with high confidence in model prediction in addition to the labeled dataset sampled by the query strategy

    - 这个也是在贝叶斯中的做法，也是第一次应用主动学习到深度学习中的样例
    - **Cost-Effective Active Learning** 值得一看
  - Uncertainty-based query strategy refers to the model based on sample uncertainty ranking to select the sample to be queried. The greater the uncertainty of the sample, the easier it is to be selected
  - However, this is likely to ignore the relationship between samples. Therefore, the method that considers multiple sample attributes is called the hybrid query strategy
  - For example, in the margin sampling, margin M is defined as the difference between the predicted highest probability and the predicted second highest probability of an sample as follows: $M = P(y1 \mid x) - P(y2 \mid x)$, where $y1$ and $y2$ are the first and second most probable labels predicted for the sample x under the current model
  - A feasible strategy would thus be to use a hybrid query strategy in a batch query, taking into account both the information volume and diversity of samples in either an explicit or implicit manner
  - The performance of early Batch Mode Active Learning (BMAL) algorithms are often excessively reliant on the measurement of similarity between samples. In addition, these algorithms are often only good at exploitation (learners tend to focus only on samples near the current decision boundary, corresponding to high-information query strategies), meaning that the samples in the query batch sample set cannot represent the true data distribution of the feature space (due to the insufficient diversity of batch sample sets)
    - Exploration-P
      - Using a deep neural network to learn the feature representation of the samples, then explicitly calculates the similarity between the samples
    - DMBAL (Diverse Mini-Batch Active Learning)
      - Adding informativeness to the optimization goal of K-means by weight, and further presents an in-depth study of a hybrid query strategy that considers the sample information volume and diversity under the mini-batch sample query setting
      - 

        (a) Batch query strategy considering only the amount of information.

        (b) Batch query strategy considering both information volume and diversity.

        Fig. 2. A comparison diagram of two batch query strategies, one that only considers the amount of information and one that considers both the amount and diversity of information. The size of the dots indicates the amount of information in the samples, while the distance between the dots represents the similarity between the samples. The points shaded in gray indicate the sample points to be queried in a batch.
- Deep Bayesian Active Learning (DBAL)

- Traditional DL methods rarely represent such model uncertainty
- 

  $f(x;\theta)$, $p(\theta)$ is a prior on the parameter space θ (usually Gaussian), and the likelihood $p(y=c|x,\theta)$ is usually given by $softmax(f(x;\theta))$. Our goal is to obtain the posterior distribution over θ, as follows:

  $$p(\theta|X,Y)=\frac{p(Y|X,\theta)p(\theta)}{p(Y|X)}. \tag{8}$$

  For a given new data point x^*, $\hat{y}$ is predicted by:

  $$p\left(\hat{y}|x^*,X,Y\right)=\int p\left(\hat{y}|x,\theta\right)p(\theta|X,Y)d\theta=\mathbb{E}_{\theta\sim p(\theta|X,Y)}[f(x;\theta)]. \tag{9}$$

- Density-based Methods
  - The density-based method is a query strategy that attempts to find a core subset representing the distribution of the entire dataset from the perspective of the dataset to reduce the cost of annotation
  - **This idea is mainly inspired by the compression idea of the core set dataset and attempts to use the core set to represent the distribution of the feature space of the entire original dataset, thereby reducing the labeling cost of AL**
  - The Core-set approach attempts to solve this problem by constructing a core subset
  - Previous core-set-based methods often simply try to query data points as far as possible to cover all points of the data manifold without considering the density, which results in the queried data points overly representing sample points from manifold sparse areas

- 停止闭环
  - Most DeepALs often use the predefined stopping criterion, and when the criterion is satisfied, they stop querying labels from the oracle
    - Maximum number of iterations
    - Minimum threshold for changing classification accuracy
    - Minimum number of labeled samples
    - Expected accuracy value
  - Although these stopping criteria are simple, these predefined stopping criteria are likely to cause DeepAL to fail to achieve optimal performance

四、问题和回答

- Stream-based AL 和 Pool-based AL 的核心区别
  - 后者会考虑数据的总体分布，前者只考虑某条具体样本
  - 前者一般应用在移动小型设备上，主流的还是后者
-