**Scaling Up CrowdSourcing to Very Large Datasets:**

**A Case for Active Learning**

论文中提出两种主动学习方案：

1. upfront

首先利用标准训练集对模型进行训练，然后根据预算，选择m大小的数据集进行众包，收集好数据之后对模型重新训练，将剩余未标注的数据进行标注。（只进行一次众包的行为）

1. iterative

通过迭代的方法进行标注。

迭代停止的条件：

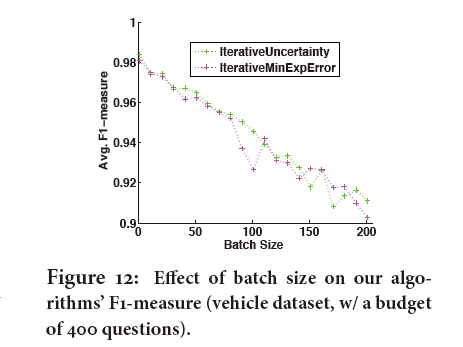
1. 到达预算的上限
2. **逐步**增加训练集的数量直到达到需要的准确率。

有标准数据：利用训练好的模型对标准数据集进行训练，计算模型的准确率。

无标准数据：利用k折交叉验证的方法对模型的准确率进行估算。（论文中利用实验验证了k折交叉验证的估计值与实际值比较接近，可以用于估算）

每次迭代时加入标记的个数（batch size），记为a：

a的值越小对提升模型的效果越好，a的最优值为1，但同时会提升整个学习过程的时间。考虑到二者的平衡，找到一个合适的a的值，可以从1开始，逐步增大a的值，直到时间或者准确率符合实验者的要求。



F1值浮动范围在8-10 随着每次迭代个数增加，运行时间会降低，迭代个数从1到200 训练模型时间也随之降低，降低大约两个数量级

**Support Vector Machine Active Learning with Applications to Text Classification**

论文中提出在SVM上引入主动学习对文本进行分类：

比较三种主动学习挑选query的方法与随机挑选的方法的预测准确率：

（这里的query指的是从未标记样本中挑选给众包标注的样本）

实验一：

逐步加入query观察模型准确率的变化直到和被动学习得到的准确率基本持平。

（被动学习：不加入主动学习单纯用SVM训练模型）

结论：

1） 利用主动学习的方法扩展到100条训练集时，能够达到原本1000条训练出来的模型性能。成本大大减少。

2） 3种主动学习方法的曲线比较接近，但都明显优于随机挑选

实验二：

比较不同topic下，不同挑选策略下挑选同样个数后，模型准确率提升情况。以及达到相同效果情况下，初始训练集所需大小的比较。

结论：

1）总体的提升效果不明显

2）被动学习需要相当于主动学习6倍的query个数

迭代停止的条件：

1 准确率达到要求（有标准答案）