# Breast cancer classification

训练数据集来提高学习性能

主动学习被广泛应用于用少量训练样本获得最优预测结果

其次，将经过预测阶段的预测样本作为下一个需要训练的样本。通过迭代过程对预测样本进行校正，提高了癌症的准确性

# A survey of Deep Active Learning

主动学习是一种致力于研究如何通过标记尽可能少的样本来获得尽可能多的性能增益的方法。更具体地说，它旨在从未标记的数据集中选择最有用的样本，并将其交给人工进行标记，以尽可能地减少标记的成本，同时仍然保持性能

种类

Pool based AL cycle

Stream based AL cycle

基本过程

1. 从unlabeled pool中根据query strategy获取sample
2. 将sample交给oracle进行label 得到数据集L
3. 在L上训练模型

关键点

1. 设计的query strategy

方式

1. 随机的query strategy 导致样本抽样偏差，不同label的样本分布不合理。注重抽样多样性的strategy导致样本标记成本增加
2. Density Based 该思想主要是受到核心集数据集压缩思想的启发，试图用核心集来表示整个原始数据集的特征空间分布，从而降低人工智能的标记成本

# Active learning for human protein-protein interaction prediction

## 结论

主动学习算法能够使用更少的标签数据学习更准确的分类器，并被证明在手动注释数据非常困难的应用程序中非常有用。这里演示的主动学习技术也可以应用于其他蛋白质组学应用，如蛋白质结构预测和分类

数据

已知14600对蛋白质相互作用，随机生成400000个特征向量作为负样本。

方法

为了保持正配对和随机配对之间特征覆盖的平衡，创建了一个同质数据子集，使每个配对的特征覆盖超过80%;(通常来说，处理功能覆盖率很低的实例的挑战很重要，但这是一个单独的研究)。该子集用于算法的开发和评估。这个同质亚群共有55,950个蛋白质对。从这些蛋白质对中随机选择1万对进行训练，另外1万对进行测试。正负样本比例为1:4

Query stragegy:

A: random data

B: Density based

C: Uncertainty Based(random seed)

D: Uncertainty Based(density-based seed)