

摘要

- WSL主要的三种弱监督: incomplete supervision, inexact supervision, inaccurate supervision

正文

incomplete supervision: 数据集中只有一小部分有标签

主动学习: 借助人工干预, 目标是最小化询问专家 (查询) 的数量

有两个指标可以衡量一个指标值得查询的程度:

- **informativeness:** 标签可以消除统计模型的不确定性的程度, 针对他的查询算法有:
 1. uncertainty sampling: 基于有限标签数据训练一个学习者, 查询在为标签数据中学习者预测置信度低的标签
 2. query by committee: 训练多个学习者, 对学习者的预测分歧最大的标签进行查询
- **representativeness:** 标签可以帮助表示输入的模式, 针对他的查询算法有:
 1. 聚类算法: 聚类数据, 然后在不同类中按数量比例查询标签

semi-supervised learning: 没有人工干预

$$h(\mathbf{x}) = \arg \max_{c \in \{Y, N\}} \sum_{j=1}^n P(y_i = c | g_i = j, \mathbf{x}_i) \\ \times P(g_i = j | \mathbf{x}_i),$$

上图证明了没有标签的数据确实可以提升模型准确度 (第二项)

在半监督学习中有两项假设: cluster assumption(数据拥有聚类结构, 并且在同一聚类的数据共享同一标签), manifold assumption(距离近的标签有同样的标签)

算法有: generative methods, graph-based methods, low-density separation methods, disagreement-based methods

inexact supervision: 粗粒度标注

一个被粗标注的个体称为bag, bag中装有实例, 这些实例共享该包所属的标签, 一个包为正当且仅当包中存在一个实例为正

多实例学习:

1. transfer focus from instance discrimination to bag discrimination
2. representation transformation

! instance 不能被假定为独立同分布

! 包中每个实例按不同的规则分类不能进行多实例学习

! 将其他任务迁移到多实例学习中可能需要一个bag generator来将bag分裂成若干个instances

inaccurate supervision: 标签有标错的

见 label noise learning tutorial