Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

协同训练(Co-training)

王海州　　硕士

2018年01月07日

内容提要

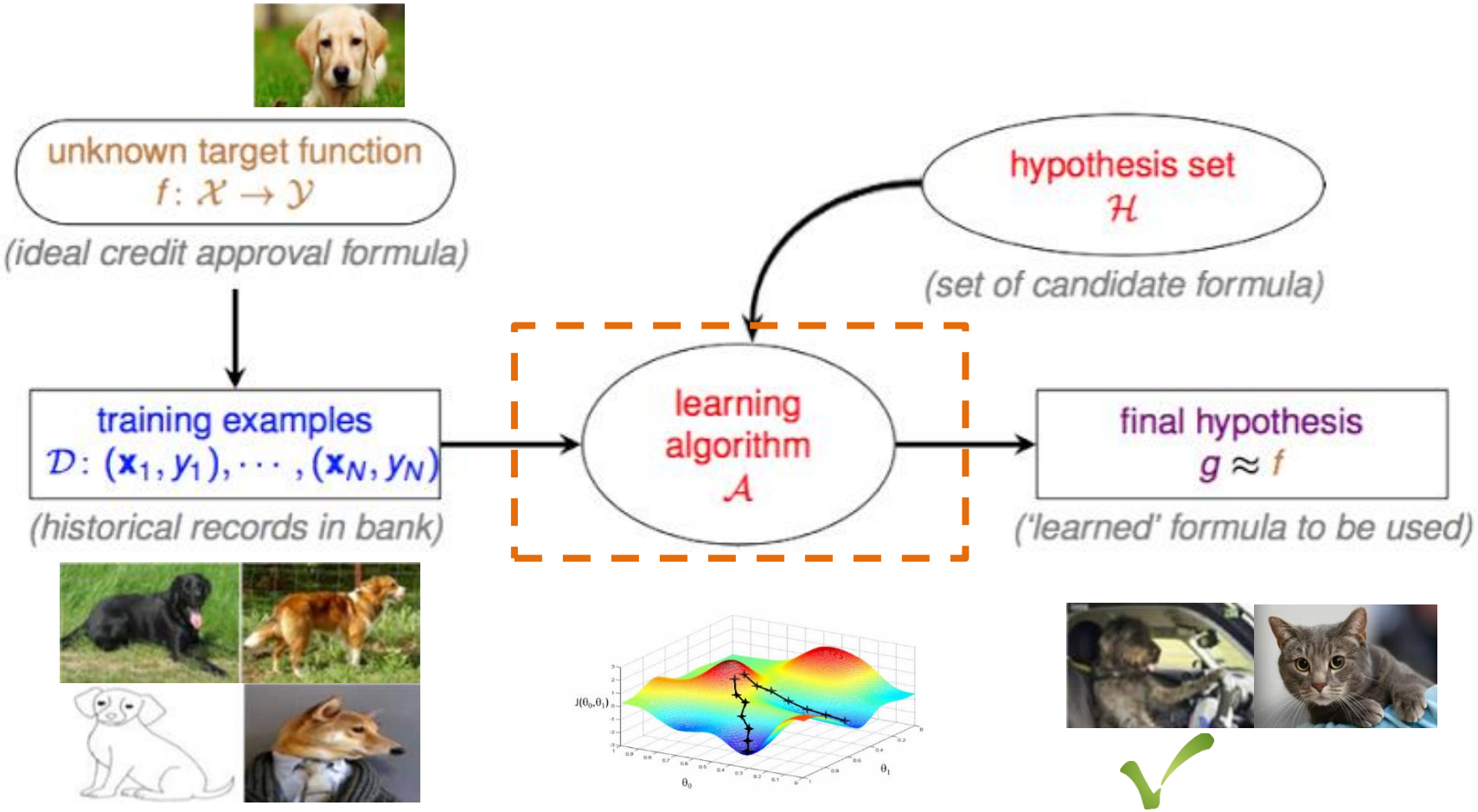
- 背景简介
- 基本概念
- 算法原理
- 扩展延伸
- 应用总结
- 参考文献

协同训练(Co-training)

背景简介

背景介绍

- 机器学习基础架构[1]



[1] 林轩田，机器学习基石

背景介绍

- 标签数据较难得到，无标签数据易获取
- 在仅有少量标签数据时，如何利用大量无标签数据提升学习性能？

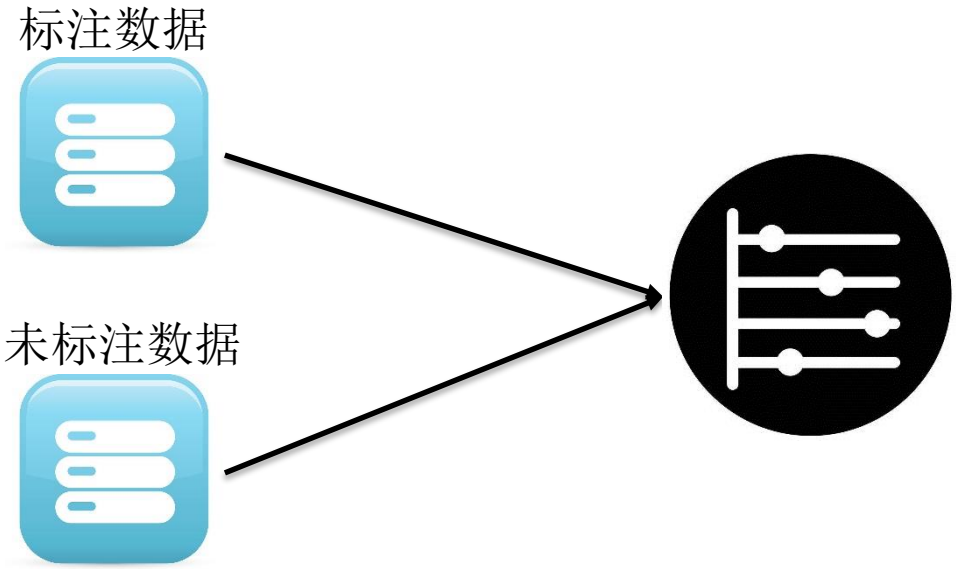
背景介绍

- 半监督学习最早由B. Shahshahani和D. Landgrebe在1994年提出
- 半监督学习的协同训练方法最早由A. Blum和T. Mitchell在1998年提出，在之后的十几年内迅速成为热点
- 2016年刘铁岩团队在nips发表了对偶学习的文章引起了争议，协同训练重新进入我们的视野

基本概念

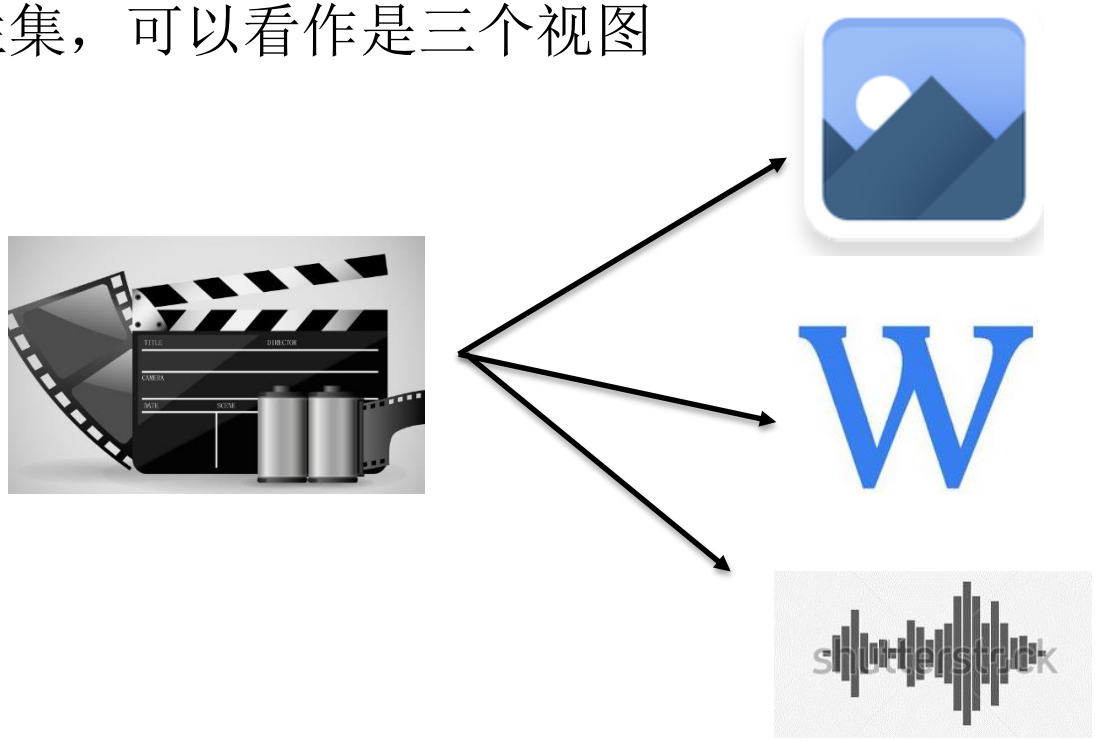
基本概念

- 半监督学习
  - 从含有标注数据和未标注数据的训练集中学习模型
  - 标注数据较难获得
  - 未标注数据较容易获得



基本概念

- 视图
  - 一个数据对象往往同时拥有多个“属性集”(attribute set)，每个属性集构成一个“视图”(view)
  - 例：电影对应数据对象，图像画面、声音、字幕信息对应三个属性集，可以看作是三个视图



基本概念

- 相容性
  - X^{v1}和X^{v2}分别表示X的两个视图，则一个样本可以表示为(x_1, x_2)，F表示X空间的目标函数，F_1、F_2分别表示在两个视图的目标函数，若满足$F_1(x_1) \neq F_2(x_2)$的样本(x_1, x_2)存在的概率为零，则称这两个视图具有相容性
  - 即F_1、F_2的判别的标记空间应存在交集
  - 例：通过画面判定的结果集合为{爱情片、动作片}，通过声音判定的结果集合为{爱情片、动作片}　　画面和声音相容
  - 若声音判定变为{爱情片、悬疑片}　　不相容

协同训练(Co-training)

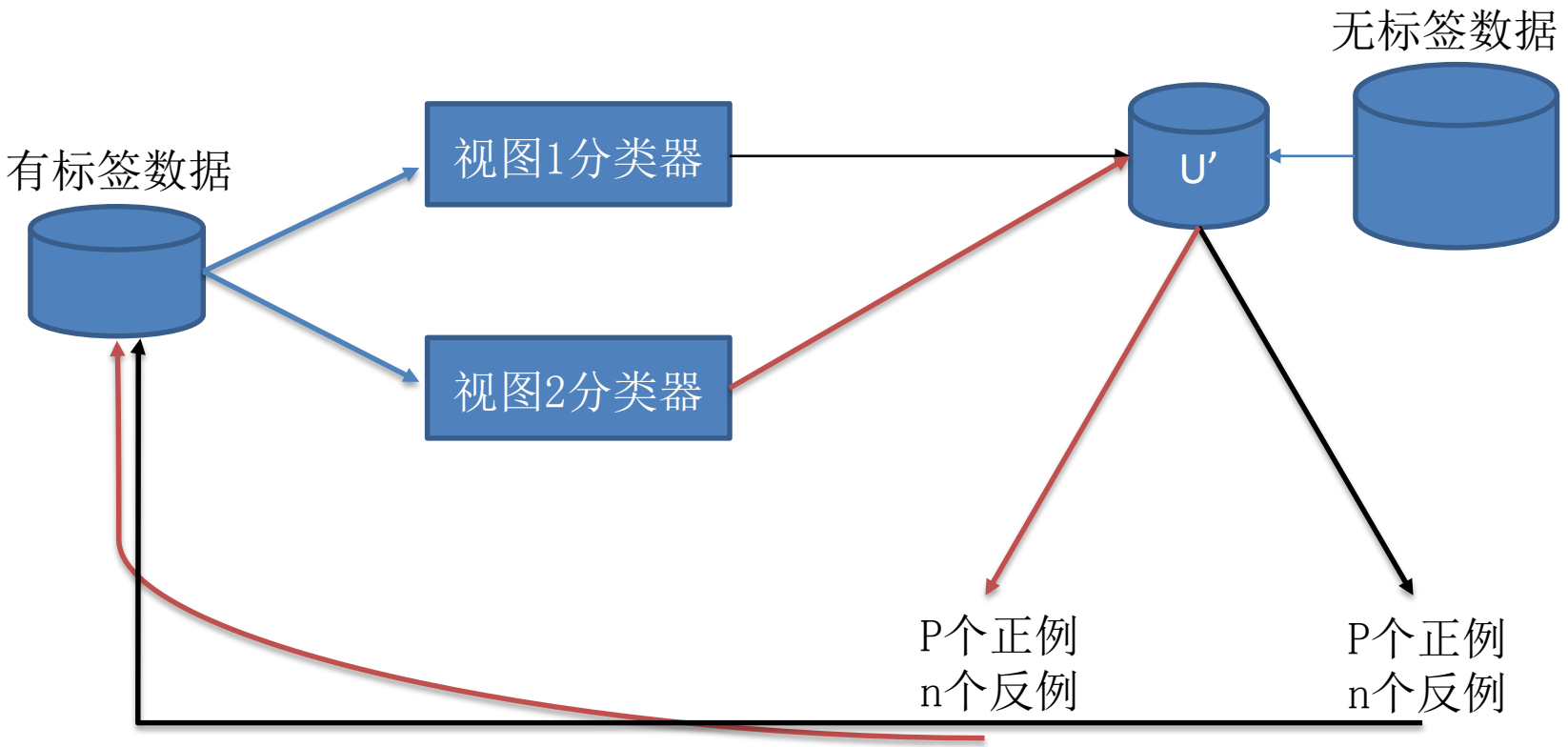
算法原理

算法原理

- 假设数据有两个充分且条件独立视图
  - 充分：包含足以产生最优学习器的信息
  - 条件独立：在给定的类别标记条件下，两个视图独立

算法原理

- 一次迭代



算法原理

- 伪代码

Given:

- a set L of labeled training examples
- a set U of unlabeled examples

Create a pool U' of examples by choosing u examples at random from U

Loop for k iterations:

Use L to train a classifier h_1 that considers only the x_1 portion of x

Use L to train a classifier h_2 that considers only the x_2 portion of x

Allow h_1 to label p positive and n negative examples from U'

Allow h_2 to label p positive and n negative examples from U'

Add these self-labeled examples to L

Randomly choose $2p + 2n$ examples from U to replenish U'

算法原理

- 约束条件
  - 视图之间需要具有相容性*
  - 每个视图都包含足够产生最优学习器的信息*
  - 在给定类别标签条件下两个视图独立

协同训练(Co-training)

扩展延伸

扩展延伸

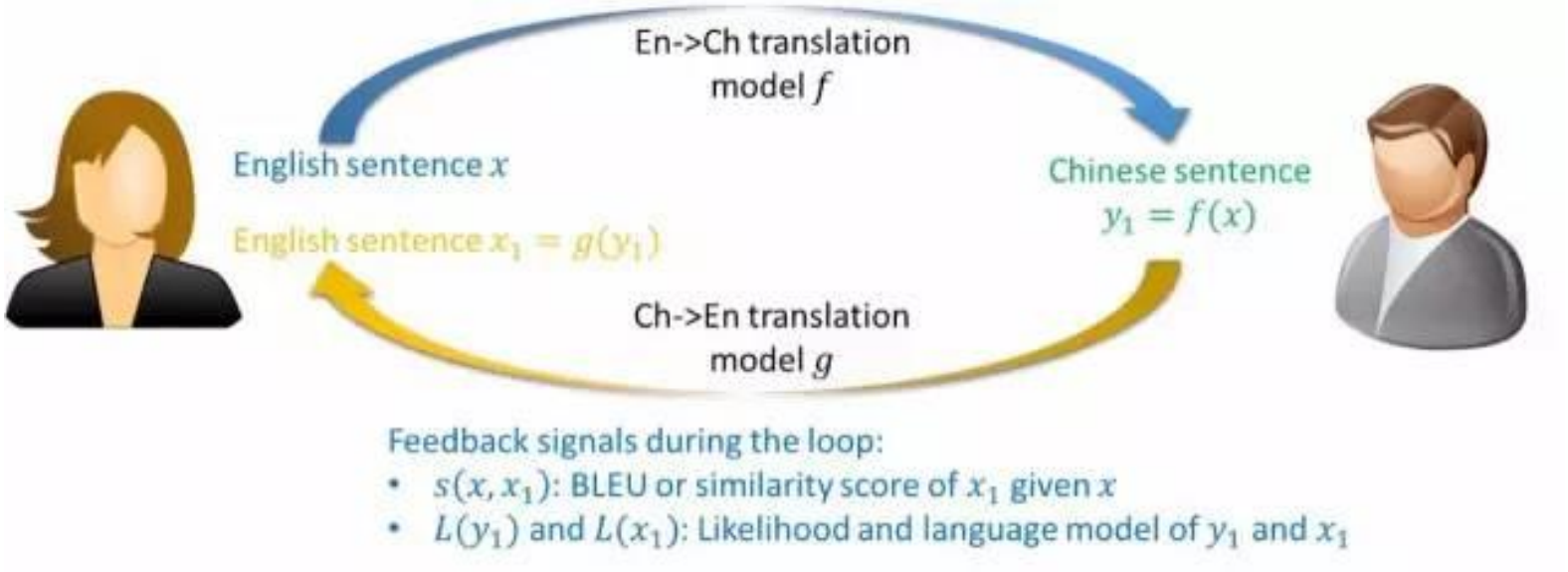
- Self-training
  - 两堆数据A和B，A为带标签数据，B是无标签数据
  - 算法流程
    - 从已标注A中训练一个分类模型M
    - 用模型M对B进行预测
    - 将预测结果置信度高的K个样本，连同预测结果加入A，并从B中删除
    - 回到第一步
- 最简单也最容易实现的半监督模型
- 若一个分类错误的样本加入了训练集，这个错误会越来越深

扩展延伸

- 对偶问题
  - 原始任务的输入空间为X，输出空间为Y
  - 对偶任务的输入空间为Y，输出空间为X
  - 例：中英文对译、图像分类与图像生成

扩展延伸

- 对偶学习
  - 例：小明和爱丽丝的中英文对译



扩展延伸

- 协同训练　　vs　　对偶学习
  - 相似性
    - 两个学习器相互影响
    - 致力于解决数据不够的情况
  - 差异性
    - 对偶学习的学习器任务不同
    - 约束条件不同

抄袭？

应用总结

应用总结

- 文本半监督分类

邓攀晓，罗涛，李剑峰. 基于不同文本表示协同训练的半监督文本分类算法[J]. 2017.

- 网页类型分类

Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Confer-ence on Computational Learning Theory. New York, USA: ACM, 1998. 92;100

- ……

参考文献

参考文献

[1] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]// Eleventh Conference on Computational Learning Theory. ACM, 1998:92-100.

[2] Xia Y, Qin T, Chen W, et al. Dual Supervised Learning[J]. 2017.

[3] Mcclosky D, Charniak E, Johnson M. Effective self-training for parsing[C]// Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics, 2006:152-159.

[4] 周志华. 基于分歧的半监督学习[J]. 自动化学报, 2013, 39(11):1871-1878.

道德经

道可道，非常道。名可名，非常名。无名天地之始。有名万物之母。故常无欲以观其妙。常有欲以观其徼。此两者同出而异名，同谓之玄。玄之又玄，众妙之门。

谢谢！