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

图半监督学习

Graph-based Semi-supervised Learning

尹继泽　硕士研究生

2018年09月02日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 1.理解图半监督学习的算法原理
  - 2.熟悉图半监督学习算法的应用场景
  - 3.了解图半监督学习在命名实体识别任务中的应用

背景简介

- 算法提出的原因
  - 现实中的有标记样本数量远远少于未标记样本数量（成本）
  - 有标记样本数量有限，造成学得模型的泛化能力不佳
- 提出问题
  - 为了解决未标记样本难以利用、有限的有标记样本训练出的模型泛化能力不佳的问题，在本次演讲中介绍图半监督学习算法及其在命名实体识别任务中的应用

基本概念

- 基本概念
  - 有标记样本：存在训练样本集$D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$，这$l$个样本的类别标记已知
  - 未标记样本：存在$D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}, l \ll u,$ 这u 个样本的类别标记未知
  - 学习器：学习算法在给定数据和参数空间上的实例化
  - 半监督学习：让学习器不依赖外界交互、自动地利用未标记样本来提升学习性能

算法原理

T	基于少量有标记样本和大量未标记样本，训练模型，使模型的泛化能力高于监督学习得到的模型
I	有标记样本集；未标记样本集；构图参数；折中参数

P	For {
	1. 图构建
	2. 标记传播
	3. 模型训练
	}

O	高性能分类器模型

P	1.现实中的有标记样本数量远远少于未标记样本数量 2.有标记样本数量有限，造成学得模型的泛化能力不佳
C	具备大量未标记样本，和少量有标记样本
D	如何利用未标记样本来训练模型
L	CCF C类会议

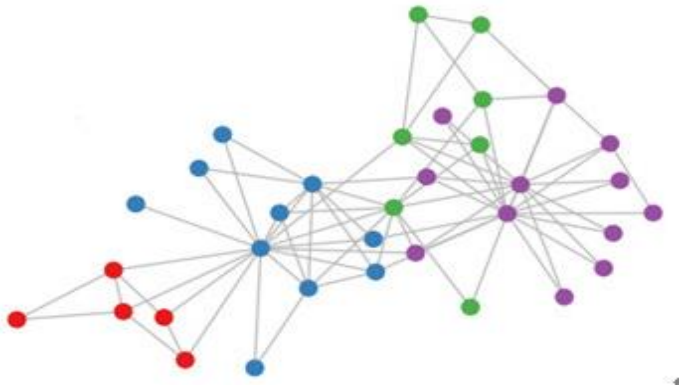
算法原理

- 必要的假设
  - 目的：将未标记样本所揭示的数据分布信息与类别标记相联系
  - 聚类假设：假设数据存在簇结构，同一个簇的样本属于同一个类别
  - 流形假设：假设数据分布在一个流形结构上，邻近的样本拥有相似的输出值

- 本质
  - 相似的样本拥有相似的输出

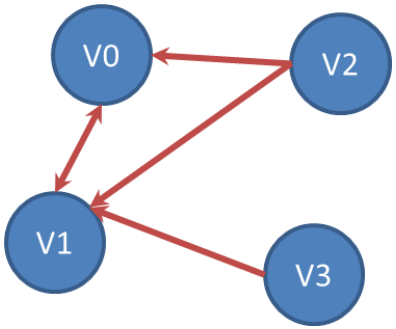
算法原理

- 从数据集到图
  - 数据集中每个样本对应于图中一个结点
  - 两个样本之间相似度很高（或相关性很强），则对应的结点之间存在一条边，边的强度正比于样本之间的相似度（或相关性）
  - 对有标记样本对应的结点染色
  - 对未标记样本对应的结点不染色

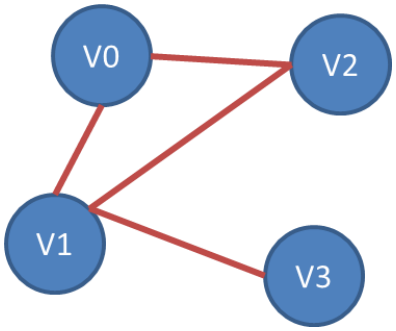


- 目的
  - 将“颜色”在图上扩散或传播

算法原理



	V0	V1	V2	V3
V0	0	1	0	0
V1	1	0	0	0
V2	1	1	0	1
V3	0	1	0	0



	V0	V1	V2	V3
V0	0	1	1	0
V1	1	0	1	0
V2	1	1	0	1
V3	0	0	1	0

算法原理

- 亲和矩阵
  - $(\boldsymbol{W})_{ij} = \begin{cases} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), \text{if } i \neq j; \\ 0, \quad \text{otherwise} \end{cases}$

- 对角矩阵
  - $\boldsymbol{D} = \text{diag}(d_1, d_2, \ldots, d_{l+u})$, $d_i = \sum_{j=1}^{l+u} (\boldsymbol{W})_{ij}$

算法原理

- 非负标记矩阵$[(l+u)\times c]$
  - $\boldsymbol{F}=\left(\boldsymbol{F}_1^{\mathrm{T}},\boldsymbol{F}_2^{\mathrm{T}},\ldots,\boldsymbol{F}_{l+u}^{\mathrm{T}}\right)^{\mathrm{T}}$
  - $\boldsymbol{F}_i=\left((\boldsymbol{F})_{i1},(\boldsymbol{F})_{i2},\ldots,(\boldsymbol{F})_{ic}\right)$为示例$\boldsymbol{x}_i$的标记向量
  - 分类规则：$y_i=\arg\max_{1\le j\le c}(\boldsymbol{F})_{ij}$
  - 对$i=1,2,\ldots,l+u,j=1,2,\ldots,c$，$\boldsymbol{F}$初始化为

$$\left(\boldsymbol{F}(0)\right)_{ij}=(\boldsymbol{Y})_{ij}=\begin{cases}1,\text{if }(1\le i\le l)\wedge(y_i=j);\\0,\text{otherwise}\end{cases}$$

  - 例如：$(2+1)\times 2$的矩阵

$$\boldsymbol{F}(0)=\boldsymbol{Y}=\begin{bmatrix}0&1\\1&0\\0&0\end{bmatrix}$$

算法原理

- 迭代式标记传播算法

输入： 有标记样本集$D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\}$；

未标记样本集$D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \dots, \boldsymbol{x}_{l+u}\}$；

构图参数σ；

折中参数α；

$$(\boldsymbol{W})_{ij} = \begin{cases} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), & \text{if } i \neq j; \\ 0, & \text{otherwise} \end{cases}$$

过程：

1: 基于亲和矩阵和参数σ得到$\boldsymbol{W}$；

2: 基于$\boldsymbol{W}$构造标记传播矩阵$\boldsymbol{S} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$;

3: 初始化$\boldsymbol{F}(0)$；

4: $t = 0$；

$$\boldsymbol{D} = \text{diag}(d_1, d_2, \dots, d_{l+u}), \quad d_i = \sum_{j=1}^{l+u}(\boldsymbol{W})_{ij}$$

算法原理

5: **repeat**

6: $\boldsymbol{F}(t+1) = \alpha \boldsymbol{S}\boldsymbol{F}(t) + (1-\alpha)\boldsymbol{Y}$;

7: $t = t+1$

8: **until**迭代收敛至$\boldsymbol{F}^*$

9: **for** $i = l+1, l+2, \ldots, l+u$ **do**

10: $y_i = \arg\max_{1 \le j \le c} (\boldsymbol{F}^*)_{ij}$

11: **end for**

输出：未标记样本的预测结果：$\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

算法原理

- 迭代式标记传播算法对应于以下正则化框架：

$$\min_{\boldsymbol{F}} \frac{1}{2}\left(\sum_{i,j=1}^{l+u}(\boldsymbol{W})_{ij}\left\|\frac{1}{\sqrt{d_i}}\boldsymbol{F}_i-\frac{1}{\sqrt{d_j}}\boldsymbol{F}_j\right\|^2\right)+\mu\sum_{i=1}^{l}\|\boldsymbol{F}_i-\boldsymbol{Y}_i\|^2$$

- 正则化参数$\mu=\frac{1-\alpha}{\alpha}$时，最优解$\boldsymbol{F}$为迭代收敛解$\boldsymbol{F}^*$
- 第一项：迫使相近样本具有相似的标记
- 第二项：迫使学得结果在有标记样本上的预测与真实标记尽可能相同

优劣分析

- 优势
  - 在概念上相当清晰
  - 易于通过对所涉矩阵运算的分析来探索算法本质
- 劣势
  - 存储开销大，样本数为$O(m)$，矩阵规模为$O(m^2)$，难以直接处理大规模数据
  - 构图过程仅能考虑训练样本集，难以判知新样本在图中的位置

应用总结

- 算法的应用领域
  - 情感分类
  - QA任务
  - 中文分词和词性标注

应用总结

- 中文命名实体识别
  - n元语法
    - 对于$p(w_i|w_1 \dots w_{i-1})$, w_i由$w_1 \dots w_{i-1}$决定，实际中只选用历史最近的$n-1$个词
  - 一元文法unigram
    - 当$n=1$时，出现在第i位上的词w_i独立于历史
  - 二元文法bigram
    - 当$n=2$时，出现在第i位上的词w_i仅与它前面的一个历史词w_{i-1}有关

应用总结

- 中文命名实体识别
  - Graph construction
    - 使用字符三元组
    - 对称k-NN图
    - 基于共现统计值，用于计算两个结点相似度的特征集

Feature	Meaning
$U_n, n \in (-4,2)$	Unigram, from previous 4th to following 2nd character
$B_{n,n+1}, n \in (-2,1)$	Bigram, 4 pairs of features, from previous 2nd to following 2nd character

  - Label propagation
  - CRF learning

参考文献

周志华.机器学习[M].北京: 清华大学出版社, 2016 (2016.10重印)

宗成庆.统计自然语言处理[M].北京：清华大学出版社，2013（2016.9重印）

Han et al.2015. Chinese named entity recognition with graph-based semi-supervised learning model. In proceedings of the eighth SIGHAN workshop on Chinese language processing (SIGHAN-8), pages 15-20.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。