

正则化框架下半监督本体算法

朱林立¹, 戴国洪², 高 炜³

(1 江苏理工学院 信息中心, 江苏 常州 213001; 2 常熟理工学院 机械学院, 江苏 常熟 215500;

3 云南师范大学 信息学院, 云南 昆明 650500)

摘 要: 提出正则化框架下的半监督本体算法, 将未标记数据的相关信息充分融入到计算模型中. 同时, 通过将数据嵌入到特征空间, 利用特征表示从理论上得到特征空间上的正则化半监督本体学习算法. 通过两个实验表明新算法对特定的应用领域具有较高的效率.

关键词: 本体; 相似度; 本体映射; 再生核希尔伯特空间; 收缩系数

中图分类号: TP393.092

文献标识码: A

文章编号: 1000-7180(2014)03-0126-04

Semi-supervised Ontology Algorithm in Regularization Setting

ZHU Lin-li¹, DAI Guo-hong², GAO Wei³

(1 Information and Education Technology Center, Jiangsu University of Technology, Changzhou 213001, china;

2 School of Mechanical Engineering, Changshu Institute of Technology, Changshu 215500, China;

3 School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China)

Abstract: In this paper, we pose semi-supervised ontology algorithm under the framework of the regularization such that the information of unlabeled data is fully integrated into the calculation model. At the same time, the data is embedded in the feature space, and feature space regularization semi-supervised ontology learning algorithm is given via the feature representation theory. Two experiments show that the new algorithm for specific applications with high efficiency.

Key words: ontology; similarity; ontology mapping; RKHS; shrinkage factor

1 引言

本体模型作为一种结构化表示和存储数据的工具, 已广泛用于计算机科学的各个领域.

近年来, 学习算法被广泛融入到本体相似度计算和本体映射中. 文献[1-2]中提出了基于正则化模型的本体相似度计算和本体映射算法, 其核心思想是将谱图理论与机器学习相融合, 通过正则化模型将问题转化为计算带权本体图拉普拉斯矩阵次小特征值对应的特征向量. 通过该特征向量, 将本体图中的每个顶点映射成实数(该特征向量的第 i 个分量代表顶点 v_i 对应的实数). 由此, 两顶点间的相似

程度通过它们对应实数的差值来判定. 其算法模型为

$$\arg \min_f \Omega(f) = \sum_{uv \in E} w(uv) \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2,$$

或其变形:

$$\arg \min_f \Omega(f) = \sum_{uv \in E} \pi(u) p(u, v) \left(\frac{f(u)}{\sqrt{\pi(u)}} - \frac{f(v)}{\sqrt{\pi(v)}} \right)^2.$$

这里, $p(u, v)$ 是顶点对上的传递概率函数, π 是本体图顶点的静态分布函数. 文献[3]中对此类算法的收敛性进行了相关的理论分析.

但是, 文献[1-2]的算法框架只适用于监督学习的情况, 无法处理半监督的场景. 本文在文献

收稿日期: 2013-05-29; **修回日期:** 2013-07-03

基金项目: 国家自然科学基金项目(60903131); 教育部科学技术研究重点项目(210210); 江苏省高校自然科学研究项目资助(10KJD52002)

[1-2] 基于正则化模型本体算法的基础上,提出基于正则化模型的半监督本体算法. 另一方面,通过空间变换,从理论上给出特征空间上对应的基于正则化模型的半监督本体算法.

2 半监督学习算法

设本体图 $G = (V, E)$ 对应本体 O , 图中顶点对应本体中的概念. 为了将顶点集 V 看成数据集合, 将每个顶点所有相关信息(包括名称、属性、实例、结构)用一个多维向量来表示. 设 $f: V \rightarrow \mathbb{R}$ 为预测函数, y_i 是顶点(向量) v_i 对应的标记, l 为亏损函数, 它根据 $f(v_i)$ 和 y_i 之间的偏差来给出相应的惩罚. 设 $(v_1, y_1), \dots, (v_n, y_n)$ 为服从分布 D 的训练样本. 函数 f 的选取范围称为假设空间. 假设空间的设置对学习模型至关重要, 如果其过大(例如: 假设空间为实值函数集合), 则会使得算法从理论上丧失稳定的估计. 因此, 我们通常会对特征空间加以限定. 本文中, 假设函数 f 有如下核方法特征:

$$f(\alpha, v) = \sum_{i=1}^n \alpha_i k(v_i, v) \quad (1)$$

式中, $\alpha = [\alpha_i]_{i=1, \dots, n}$ 是由数据得到的参数向量, k 是定义在本体图顶点集上的对称正定核函数. 从而 $n \times n$ Gram 矩阵 $K = [k(v_i, v_j)]_{i, j=1, \dots, n}$ 是半正定的. 通过该假设, 寻找抽象函数 f 的问题就转化为寻找参数向量 α .

设 $H_0 = \{ \sum_{i=1}^m \alpha_i k(v_i, v), m \in \mathbb{N}, \alpha_i \in \mathbb{R} \}$. 在 H_0 上定义如下范数:

$$\| \sum_i \alpha_i k(v_i, \cdot) \| = \left(\sum_{i, j} \alpha_i \alpha_j k(v_i, v_j) \right)^{\frac{1}{2}}.$$

则 H_0 在此泛数下成为内积空间. 设 H 是 H_0 在上述泛数下的闭包. 则 H 为希尔伯特空间, 也称为核函数 k 的再生核希尔伯特空间(reproducing kernel Hilbert space, RKHS). H 上对应的泛数记为 $\| \cdot \|_H$, 内积记为 $\langle \cdot, \cdot \rangle_H$. RKHS 的再生性是指: 对任意 $f \in H$, 有

$$f(v) = \langle f, k(v, \cdot) \rangle_H.$$

设 $\hat{f} \in H$ 为有噪框架下通过正则经验模型得到的估计函数, 即

$$\hat{f}(\cdot) = \arg \inf_{f \in H} \left\{ \frac{1}{n} \sum_{i=1}^n l(f(v_i), y_i) + \lambda \| f \|_H^2 \right\} \quad (2)$$

其中 $\lambda > 0$ 为正则化参数, 它用来平衡函数 f 对样本数据的精度和自身的光滑性. 根据表示定理^[4], 式

(2) 的解可表示为 $\hat{f}(v) = \sum_{i=1}^n \hat{\alpha}_i k(v_i, v)$, 因此, 式

(2) 等价于

$$\begin{aligned} \hat{f}(\cdot) &= \sum_{i=1}^n \hat{\alpha}_i k(v_i, \cdot), \\ [\hat{\alpha}_i] &= \arg \inf_{[\alpha_i] \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n l \left(\sum_{j=1}^n \alpha_j k(v_i, v_j), y_i \right) + \lambda \sum_{i, j=1}^n \alpha_i \alpha_j k(v_i, v_j) \right\}. \end{aligned} \quad (3)$$

将式(2)转换成式(3)的优点在于将无穷维问题转换为求解有限维向量 α .

设 $|V| = m$, 在半监督框架下, 设 $\{(v_i, y_i)\}_{i=1, \dots, n}$ 为标记数据, 其 $v_j (j = n+1, \dots, m)$ 余均为未标记数据. 利用正则化模型得到本体相似度的本质是得到 m 维向量, 该向量的第 i 个分量的值即为顶点 v_i 对应的实数. 进而, 我们将函数 f 和它的估计 $\hat{f}$ 看成向量, 即 $\hat{f}_j = \hat{f}(v_j)$, $j = 1, \dots, m$. 对应地, 式(2)变成如下形式:

$$\hat{f} = \arg \inf_{f \in \mathbb{R}^m} \left\{ \frac{1}{n} \sum_{i=1}^n l(f_i, y_i) + \lambda f^T K^{-1} f \right\} \quad (4)$$

如何在模型中充分利用未标记数据, 是半监督学习所关心的问题. 在基于核方法的半监督学习中, 一种有效的技术是利用未标记数据的相关信息, 将核函数 k 替换成 $\bar{k}$, 同时 K 变换为 $\bar{K}$.

文献[5]提出了如下方法, 在式(4)中用图拉普拉斯矩阵 L 来取代 K^{-1} . 进而有

$$\begin{aligned} f^T L f &= \beta f^T f + \frac{1}{2} \sum_{i, j=1}^m k(v_i, v_j) \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2 \\ &= f^T [(1 + \beta) I - D^{-1/2} K D^{-1/2}] f. \end{aligned}$$

这里, $d_i = \sum_{j=1}^m k(v_i, v_j)$ ($i = 1, \dots, m$), 对角矩阵 D

对角线元素为 d_i , 参数 $\beta > 0$ 用来保证拉普拉斯核 $\bar{K} = ((1 + \beta) I - D^{-1/2} K D^{-1/2})^{-1}$ 是正定的. 可见, 该方法通过变换对应核矩阵, 模型中增加了未标记顶点 $v_j (j = n+1, \dots, m)$ 的信息.

从这一思想中得到启发, 我们给出本文的第一个算法如下:

算法 1: 基于正则化模型的半监督学习算法

输入: 标记数据 $\{(v_i, y_i)\}_{i=1, \dots, n}$ 和未标记数据 $v_j (j = n+1, \dots, m)$; 核函数 k ; 收缩因子 $s_j \geq 0 (j = 1, \dots, m)$.

过程: 由核矩阵 $K = [k(v_i, v_j)]_{i, j=1, \dots, m}$ 计算核特征分解 $K = m \sum_{j=1}^m \mu_j \gamma_j \gamma_j^T$, 其中 (μ_j, γ_j) 是 K 的特征

对且满足 $\gamma_j^T \gamma_j = 1$.

修改核矩阵为

$$\bar{\mathbf{K}} = m \sum_{j=1}^m s_j \mu_j \gamma_j \gamma_j^T.$$

计算

$$\hat{f} = \arg \inf_{f \in \mathbb{R}^m} \left\{ \frac{1}{n} \sum_{i=1}^n l(f_i, y_i) + \lambda f^T \bar{\mathbf{K}}^{-1} f \right\}.$$

输出: 在顶点 $v_j (j = 1, \dots, m)$ 上的预测值 $\hat{f}_j$.

算法 1 的核心思想是通过核矩阵的特征分解, 得到包含所有顶点信息的 m 个特征对, 并由此来修改原有核矩阵. 在算法 1 中, 收缩系数的作用是调节各个顶点对 $\bar{\mathbf{K}}$ 的贡献, 从而来控制出错率. 但在实际应用中, 收缩系数的选择很难确定. 目前在工程上还没有普遍接收的方法来确定 s_j 的值. 为此, 用收缩函数来统一取代收缩系数, 得到算法 1 的如下改进:

算法 2 (算法 1 的变种):

输入: 标记数据 $\{(v_i, y_i)\}_{i=1, \dots, n}$ 和未标记数据 $v_j (j = n+1, \dots, m)$; 核函数 k ; 收缩函数 $g(\cdot)$.

过程: 由核矩阵 $\mathbf{K} = [k(v_i, v_j)]_{i,j=1, \dots, m}$ 计算

$$\hat{f} = \arg \inf_{f \in \mathbb{R}^m} \left\{ \frac{1}{n} \sum_{i=1}^n l(f_i, y_i) + \lambda f^T \mathbf{K}^{-1} g(\mathbf{K}/m)^{-1} f \right\}.$$

输出: 在顶点 $v_j (j = 1, \dots, m)$ 上的预测值 $\hat{f}_j$.

在算法 2 中, $s_j = g(\mu_j)$, 其中 g 是有理函数, 从而 $\bar{\mathbf{K}} = g(\mathbf{K}/m) \mathbf{K}$. 算法 2 相对算法 1 的本质改进在于: 用收缩函数取代收缩系数后, 无需计算 $\mathbf{K}$ 的特征分解.

3 算法在特征空间上的扩展

在经典基于核方法的分类算法中, 可将数据嵌入到高维空间 F (称为特征空间), 此时核分类器等价于线性分类器^[6]. 在此经典框架下, 特征空间 F 中的特征可看成列向量, 特征表示 $\psi(v) \in F$ 满足 $k(v_i, v_j) = \psi(v_i)^T \psi(v_j)$. 此时, 任意 $f \in H$ 都可表示为 F 上的线性分类器和权值向量 ω 的组合. 即 $f(v) = \omega^T \psi(v)$ 且 $\|f\|_H^2 = \omega^T \omega$.

基于上述思想, 我们给出算法 1 和算法 2 的特征空间上的变种. 设 H 的特征空间 F 是向量表示满足: 任意 $f \in H$ 对应 $\omega_f \in F$ 且 $\|f\|_H^2 = \omega_f^T \omega_f$. 对本体图中的每个顶点 v , 定义它的特征表示为 $\psi(v) = \omega_{k(v, \cdot)}$. 基于此特征空间表示, 式(2)等价于

$$\begin{aligned} \hat{f}(v) &= \hat{\omega}^T \psi(v), \\ \hat{\omega} &= \arg \inf_{\omega \in F} \left\{ \frac{1}{n} \sum_{i=1}^n l(\omega^T \psi(v_i), y_i) + \lambda \omega^T \omega \right\} \quad (5) \end{aligned}$$

在半监督框架下, 核的变换就对应成为特征空间的变换. 设 $\mathbf{S}$ 是 F 上的半正定矩阵, 可将特征向量从 $\psi(v)$ 变换为 $\psi'(v) = \mathbf{S}^{\frac{1}{2}} \psi(v)$, 从而核函数从 $k(v_i, v_j) = \psi(v_i)^T \psi(v_j)$ 转换为 $\psi(v_i)^T \psi'(v_j) = \psi(v_i)^T \mathbf{S} \psi(v_j)$.

4 算法的整体描述

基于正则化模型的半监督本体算法的整体描述如下:

算法 3: 基于正则化模型的半监督本体相似度计算算法

步骤 1: 预处理. 将本体图中每个顶点的信息用一个向量表示.

步骤 2: 选取样本集.

步骤 3: 通过算法 1 或算法 2 将每个顶点映射成实数.

步骤 4: 给出 G 中每对顶点对应实数的差值, 由此判定它们的相似程度.

算法 4: 基于正则化模型的半监督本体映射算法

步骤 1: 预处理. 设图 $G_1, G_2, \dots, G_m$ 分别对应本体 $O_1, O_2, \dots, O_m$, 令 $G = G_1 + G_2 + \dots + G_m$. 将 G 中每个顶点的信息用一个向量来表示.

步骤 2: 选取样本集.

步骤 3: 通过算法 1 或算法 2 得到每个顶点对应的实数.

步骤 4: 给出 G 中来自不同分支的顶点对所对应的实数差值, 由此判定它们的相似程度. 同时选择适当的本体映射策略.

5 实验

在以下两个实验中, 亏损函数统一使用平方亏损. 对于正则化参数, 从理论上来说, 对不同的函数空间有不同的选择方法, 相关内容可参考 cross validation 技术^[7]. 本文为了简化模型的计算, 统一取 $\lambda = 10^{-1}$. 此外, 虽然算法 2 不需要特征分解过程, 但有理函数 g 的具体表达方式在工程上难以事先确定. 因此, 两个实验均运用算法 1, 并且考虑到收缩系数的难确定性, 我们统一使用 $s_j = 1 (j = 1, \dots, m)$ 来简化实验过程.

本体相似度实验, 首先我们验证算法 5 对特定本体相似度计算的效率, 数据集采用生物 GO 本体^[8] O_1 . 选取一半顶点作为标记数据, 其余作为未标记数据. 采用 P@N^[9] 平均准确率来评价实验结

果.对比可知,本文所提出的算法5对于GO本体的效率明显高于快速排序算法和标准排序算法.

通过本体映射实验证明,本文所提出的算法6在“计算机软件”本体 O_2 和 O_3 间建立本体映射的效率明显高于快速排序算法和标准排序算法.

6 结束语

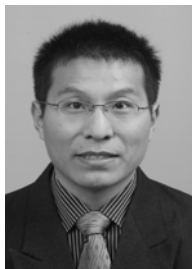
本文提出基于正则化模型的半监督本体相似度计算和本体映射算法.算法设计的核心思想是充分利用未标记数据的信息,将其融入到正则化模型中.此外,运用特征空间变换的技术,从理论的角度提出对应的半监督算法.通过两个具体的实验,验证了新算法具有较高的效率.

参考文献:

- [1] 高伟,梁立.基于超图正则化模型的本体概念相似度计算[J].微电子学与计算机,2011,28(5):15-17.
- [2] 高伟,朱林立,梁立.基于图正则化模型的本体映射算法[J].西南大学学报:自然科学版,2012,34(3):118-121.
- [3] 高伟,周定轩.与一般相似度函数相关的谱聚类的收敛性[J].中国科学:数学,2012,42(10):985-994.
- [4] Felipe C, Zhou D. Learning theory: an approximation theory viewpoint[M]. England: Cambridge University Press, 2007.
- [5] Zhou D, Bousquet O, Lal T, et al. Learning with local and global consistency[C]//NIPS. [s.l.]: IEEE, 2003: 321-328.

- [6] Zhu X, Kandola J, Ghahramani Z, et al. Nonparametric transforms of graph kernels for semi-supervised learning[J]. Advances in Neural Information Processing Systems, 2005(17): 1641-1648.
- [7] Caponnetto A, Yao Y. Cross validation based adaptation for regularization operators in learning theory[J]. Anal Appl, 2010(8): 161-183.
- [8] Gene Ontology website. Gene Ontology Database [EB/OL]. [2013-01-02]. <http://www.geneontology.org>.
- [9] Craswell N, Hawking D. Overview of the TREC 2003 web track[C]//Proceedings of the Twelfth Text Retrieval Conference. Gaithersburg, Maryland: NIST Special Publication, 2003: 78-92.

作者简介:



朱林立 男,(1975—),硕士,高级工程师.研究方向为计算机网络、人工智能及云计算.

戴国洪 男,(1967—),博士,教授.研究方向为CAD/CAPP/CAM、虚拟装配.

高伟 男,(1981—),博士,讲师.研究方向为统计学理论、图论.