

基于分离 Bregman 技术的本体稀疏向量学习算法

高 炜

(云南师范大学 信息学院, 云南 昆明 650500)

摘要:为适应大数据应用背景下本体数据的计算和处理,越来越多的稀疏学习算法被应用于本体相似度计算和本体映射.在稀疏学习框架下,本体函数的学习归结于本体稀疏向量的学习.因此,利用分离 Bregman 方法得到本体稀疏向量计算策略,通过原始优化问题和对偶优化问题的交替迭代策略得到鞍点,进而得到最优本体稀疏向量,最后通过实验验证算法的有效性.

关键词:本体;稀疏向;分离 Bregman;扩展拉格朗日函数;鞍点;软阈值

中图分类号:TP393.092 **文献标识码:**A **文章编号:**1674-5639(2015)06-0112-04

DOI:10.14091/j.cnki.kmxyxb.2015.06.028

Ontology Sparse Vector Learning Algorithm Based on Split Bregman Technology

GAO Wei

(College of Information, Yunnan Normal University, Yunnan Kunming 650500, China)

Abstract: In order to adapt the computing and processing of ontology data in the background of big data applications, more and more sparse learning algorithms are applied to the ontology similarity calculation and the ontology mapping. Under the setting of sparse learning, the learning of ontology function attributes to the learning of sparse vector. So we present an ontology sparse vector computing strategy by virtue of split Bregman methods. The saddle point is obtained in terms of iterative algorithm alternating between the primal and the dual optimization to get the optimal solution of ontology sparse vector and last, the effectiveness of the algorithm is verified by experiments.

Key words: ontology; sparse vector; split Bregman; augmented Lagrangian function; saddle point; soft thresholding

本体是一种数据存储和表示的结构化语义模型,用本体图 $G=(V, E)$ 来表示本体 O . 本体在工程应用中的核心是本体概念的相似度计算,因而本体学习算法的目标是获得最优本体函数 $f: V \rightarrow \mathbb{R}$, 即将每个顶点通过本体函数映射成实数. 进而, 设顶点 v 和 v' 之间的相似度可以通过 $|f(v) - f(v')|$ 的值来衡量.

文献[1]首次将排序学习方法应用于本体映射的获得;文献[2]将图学习的方法应用于本体相似度计算,该方法的重点在于图拉普拉斯算子的应用;文献[3]和文献[4]将图正则化模型应用于单个本体概念相似度计算和多个本体之间建立本体映射,并注意到多数本体均是自上而下的层次结构,且由于本体概念分化的原因导致其结构多为树形或近似树形;文献[5]充分利用本体图的结构特征得到 k -

部排序半监督学习算法;更进一步,文献[6]得到基于正则化瑞利系数的半监督 k -部排序学习算法,并将其应用于本体相似度计算;文献[7]给出基于迭代拉普拉斯计算方法的本体半监督学习算法.

随着大数据时代的到来,本体应用面临很大的挑战,其原因在于所要处理的数据量越来越大,并且随着本体中概念数的增加,其本体图的结构越来越复杂,这导致本体图的路径和邻域结构越来越复杂.面对本体应用的这一挑战,稀疏算法被引入到本体相似度计算和本体映射中来.

本文将分离 Bregman 方法用于本体相似度计算和本体映射的构建.通过鞍点的计算得到最优本体稀疏向量,再通过该稀疏向量获得本体实值函数,而本体图中的每个顶点则通过该函数映射为对应的实数.

收稿日期:2015-10-14

基金项目:国家自然科学基金资助项目(11401519).

作者简介:高炜(1981—),男,浙江绍兴人,副教授,博士,主要从事机器学习和图论研究.

本体顶点对应概念之间的相似度则通过它们对应实数在实数轴上的相对距离来衡量. 最后, 将该算法应用于两个具体的本体应用领域来验证算法的有效性.

1 新算法描述

对于本体中的每个概念, 对应本体图中的一个顶点, 并且用一个 p 维向量来表示概念对应的所有信息. 为了方便表示, 用 v 来同时表示顶点以及对应的向量, 即用 $v = \{v_1, \dots, v_p\}$ 表示顶点 v 对应的向量. 本体优化模型的目的是通过本体样本集的学习得到本体实值函数 $f: V \rightarrow \mathbb{R}$, 并由此将每个本体图上的顶点找到对应的实数, 进而将本体相似度计算问题转化为实数轴上 1 维距离计算的问题. 此类算法的本质是学习降维算子 f , 将原本 p 维向量表示的本体信息降为 1 维实数, 即本体函数 f 是可以看成降维映射 $f: \mathbb{R}^p \rightarrow \mathbb{R}$.

在大数据应用背景下, 向量的维度 p 会非常的大. 比如在生物基因本体中, 该向量可能包含所有基因信息. 此外, 在概念数量庞大的本体图中, 其邻域结构信息和路径信息也会非常庞大, 比如地理信息系统本体. 由于表示向量的维度很大, 因而导致计算的复杂度大大提高. 但针对某个具体的应用, 只有 p 维向量中的少数分量才对相似度计算起作用. 比如在遗传疾病相关的本体应用中, 导致某种遗传病的是少数基因, 而其他绝大部分基因无任何联系; 在 GIS 的应用中, 假如某个地点有刑事案件发生, 那么我们需要第一时间寻找事发地点附件医院和派出所的位置, 而附近的商场、学校、娱乐场所和公园与此无关. 因此, 稀疏向量学习技术被应用于本体相似度计算和本体映射算法中.

具体地说, 通过本体稀疏向量的学习, 本体函数可作如下表示:

$$f_{\beta}(v) = \sum_{i=1}^p v_i \beta_i + \delta, \quad (1)$$

(1) 式中的 $\beta = (\beta_1, \dots, \beta_p)$ 即为本体稀疏向量, 其中稀疏的含义是指 β 的大部分分量为 0 或者小到可以忽略不计; δ 表示干扰项, 一般服从某个具体的分布. 由 (1) 式可知, 本体函数的学习就是本体稀疏向量的学习. β 的学习模型可表示为:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} l(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|V\beta\|_1, \quad (2)$$

其中 $l(\beta)$ 为亏损项, 用于表示 $V\beta$ 和 y 的近似程度. 这里 $V \in \mathbb{R}^{n \times p}$ 是本体数据矩阵或者本体图邻接矩

阵, $y \in \mathbb{R}^n$ 是目标向量. $l(\beta)$ 的一类常见取法为 $l(\beta) = \|V\beta - y\|_2^2$. 而 $\lambda_1 \|\beta\|_1$ 项用来控制向量 β 的稀疏度, 其中 λ_1 为平衡参数用来控制稀疏度在模型中占的比重. $\lambda_2 \|V\beta\|_1$ 项用来控制结构化稀疏程度, 其中 λ_2 为权值参数.

例 在特殊的框架下, (2) 式可以表示为:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^n (y_i - \sum_{j=1}^p v_{ij} \beta_j)^2 + \lambda_1 \sum_{i=1}^p |\beta_i| + \lambda_2 \sum_{i=2}^p |\beta_i - \beta_{i-1}|.$$

记 $a = \beta, b = V\beta$, 本体模型 (2) 式可以表示为:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} l(\beta) + \lambda_1 \|a\|_1 + \lambda_2 \|b\|_1. \quad (3)$$

本文将使用分离 Bregman 方法对优化模型 (3) 式进行求解, 该方法已广泛应用于机器学习, 见文献 [8~9]. 记

$$L = l(\beta) + \lambda_1 \|a\|_1 + \lambda_2 \|b\|_1,$$

则, L 的拉格朗日函数为:

$$\bar{L}(\beta, a, b, u, x) = l(\beta) + \lambda_1 \|a\|_1 + \lambda_2 \|b\|_1 + \langle u, \beta - a \rangle + \langle x, V\beta - b \rangle, \quad (4)$$

其中 $u \in \mathbb{R}^p$ 是对应线性约束 $\beta = a$ 的对偶变量; $x \in \mathbb{R}^n$ 是对应线性约束 $V\beta = b$ 的对偶变量; $\langle \cdot, \cdot \rangle$ 表示内积. 扩展拉格朗日函数可表示为:

$$\begin{aligned} L(\beta, a, b, u, x) = & l(\beta) + \lambda_1 \|a\|_1 + \\ & \lambda_2 \|b\|_1 + \langle u, \beta - a \rangle + \langle x, V\beta - b \rangle + \\ & \frac{\mu_1}{2} \|\beta - a\|^2 + \frac{\mu_2}{2} \|V\beta - b\|^2, \end{aligned} \quad (5)$$

其中 $\frac{\mu_1}{2} \|\beta - a\|^2$ 项和 $\frac{\mu_2}{2} \|V\beta - b\|^2$ 项分别用于表示对线性限制 $\beta = a$ 和 $V\beta = b$ 破坏的情况下的惩罚, μ_1 和 μ_2 为各自的惩罚参数.

考虑寻找扩展拉格朗日函数 $L(\beta, a, b, u, x)$ 的鞍点 (saddle point) $(\beta^*, a^*, b^*, u^*, x^*)$ 满足

$$\begin{aligned} L(\beta^*, a^*, b^*, u, x) \leq & L(\beta^*, a^*, b^*, u^*, x^*) \\ \leq & L(\beta, a, b, u^*, x^*) \end{aligned} \quad (6)$$

对所有 β, a, b, u 和 x 都成立. 可知 β^* 是 (3) 式的最优解的充要条件是对某些 a^*, b^*, u^* 和 x^* , $(\beta^*, a^*, b^*, u^*, x^*)$ 是上述鞍点问题的解.

下面我们通过迭代算法来求解鞍点问题, 其基本思路是在原始优化和对偶优化之间来回替换:

$$\begin{aligned} \text{Primal: } & (\beta^{k+1}, a^{k+1}, b^{k+1}) \\ & = \underset{\beta, a, b}{\operatorname{argmin}} L(\beta, a, b, u^k, x^k), \\ \text{Dual: } & u^{k+1} = u^k + \delta_1 (\beta^{k+1} - a^{k+1}), \\ & x^{k+1} = x^k + \delta_2 (V\beta^{k+1} - b^{k+1}), \end{aligned} \quad (7)$$

其中第 1 步是基于当前的估计 u^k 和 x^k 来更新原始变量;第 2 步是根据当前原始变量的估计值来更新对偶变量. 由于扩展拉格朗日函数对 u 和 x 是线性的,因此只需要使用步长为 δ_1 和 δ_2 的梯度下降算法即可计算对偶变量. 而对于原始问题,则通过交替最小化 β, a 和 b 得到:

$$\begin{cases} \beta^{k+1} = \operatorname{argmin}_{\beta} l(\beta) + \langle u^k, \beta - a^k \rangle + \langle x^k, \\ V\beta - b^k \rangle + \frac{\mu_1}{2} \|\beta - a^k\|_2^2 + \frac{\mu_2}{2} \|V\beta - b^k\|_2^2, \\ a^{k+1} = \operatorname{argmin}_a \lambda_1 \|a\|_1 + \langle u^k, \beta^{k+1} - a \rangle + \\ \frac{\mu_1}{2} \|\beta^{k+1} - a\|_2^2, \\ b^{k+1} = \operatorname{argmin}_b \lambda_2 \|b\|_1 + \langle x^k, V\beta^{k+1} - b \rangle + \\ \frac{\mu_2}{2} \|V\beta^{k+1} - b\|_2^2. \end{cases} \quad (8)$$

由于目标函数是二次的,并且不可导部分可以被完全分离,因此在(8)式中关于 a, b 的最小化可通过软阈值(soft thresholding)方法得到. 设 Γ_{λ} 是定义在向量空间上的软阈值算子,并满足

$$\Gamma_{\lambda}(w) = [t_{\lambda}(w_1), t_{\lambda}(w_2), \dots]^T, \quad (9)$$

其中 $t_{\lambda}(w_i) = \operatorname{sgn}(w_i) \max\{0, |w_i| - \lambda\}$, $\operatorname{sgn}(\cdot)$ 是符号函数. 利用该软阈值算子, (8)式中 a, b 的解可表示为:

$$\begin{aligned} a^{k+1} &= \Gamma_{\mu_1^{-1}\lambda_1}(\beta^{k+1} + \mu_1^{-1}u^k), \\ b^{k+1} &= \Gamma_{\mu_2^{-1}\lambda_2}(V\beta^{k+1} + \mu_2^{-1}x^k). \end{aligned} \quad (10)$$

由此,最后得到本体稀疏向量优化模型(3)式的最优解.

2 实验

下面是设计两个具体的实验来验证算法的有效性,其使用的数据分别是 GO 生物基因本体和物理教育学本体. 为了方便计算,在稀疏向量后,直接通过内积的形式得到顶点对应的实数,即忽略噪声项 δ ,通过 $f_{\beta}(v) = \sum_{i=1}^p v_i \beta_i$ 得到本体函数. 然后根据 $|f_{\beta}(v_1) - f_{\beta}(v_2)|$ 的值来判定 v_1 和 v_2 对应的概念之间的相似度. 在具体操作过程中,我们取 $l(\beta) = \|V\beta - y\|_2^2$.

2.1 在 GO 本体上的本体相似度计算实验

第 1 个实验采用 <http://www.geneontology.org> 网站构建的生物基因 GO 本体 O_1 (下图 1 为 GO 本体的基本结构)对本文本体算法在特定应用领域

对于相似度计算的效率来进行验证. 使用 $P@N$ [10] 平均准确率作为衡量实验数据的标准. 为了说明本文算法的效率高于已有的算法,将基于回归的本体学习算法 [11]、基于快速排序的本体算法 [12] 和基于标准排序方法的本体算法 [1] 应用于生物基因 GO 本体. 当 $N = 3, 5, 10$ 时的数据结果比较如下表 1 所示.

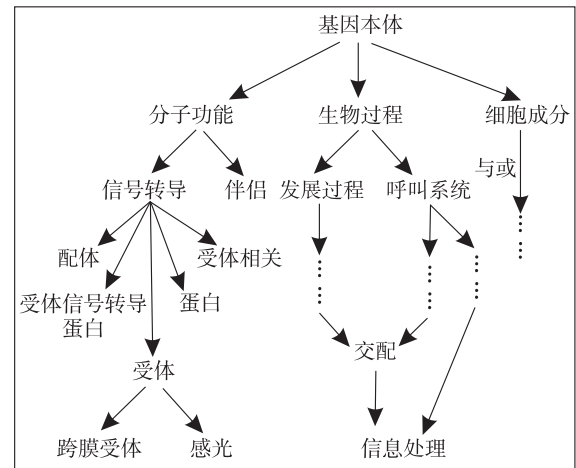


图1 GO本体 O_1

表 1 实验 1 部分数据

算法名称	$P@3$ 平均 准确率/%	$P@5$ 平均 准确率/%	$P@10$ 平均 准确率/%
本文算法	57.32	66.61	84.10
本体回归算法	56.44	63.48	78.41
快速排序算法	47.73	55.52	69.93
标准排序算法	52.37	60.62	72.96

通过表 1 中 $P@N$ 准确率对比可知,本文提出的新稀疏向量计算策略对于在生物基因 GO 本体上进行相似度计算的效率要明显优于另外 3 类算法.

2.2 本体映射实验

关于本文本体稀疏向量学习算法对构建本体映射的效率用下面两个“物理教育”本体 O_2 (见下图 2) 和 O_3 (见下图 3) 来验证. 同样,为了比较算法的计算准确率,将基于回归的本体学习算法、基于快速排序的本体算法和基于标准排序方法的本体算法分别作用于“物理教育”本体,并采用 $P@N$ 准确率进行比较. 当 $N = 1, 3, 5$ 时的数据如下表 2 所示.

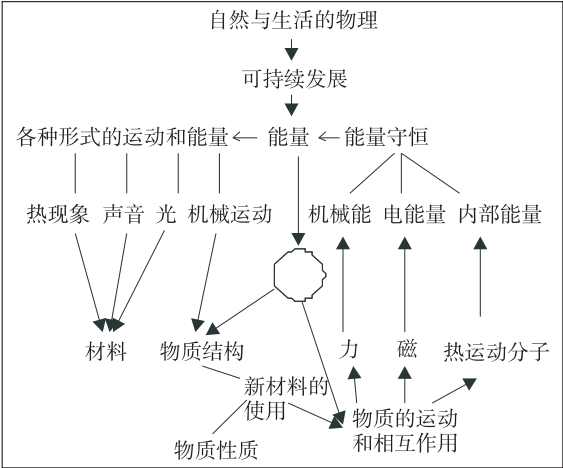


图2 “物理教育” 本体O₂

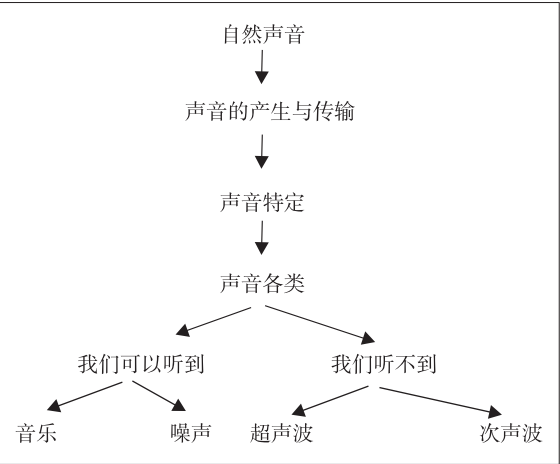


图3 “物理教育” 本体O₃

表 2 实验 2 部分数据

算法名称	P@1 平均	P@3 平均	P@5 平均
	准确率/%	准确率/%	准确率/%
本文算法	54.84	65.59	90.32
本体回归算法	48.39	52.69	65.81
快速排序算法	41.94	49.46	59.35
标准排序算法	45.16	56.99	64.52

由以上准确率对比可知,本文稀疏向量学习算法对于在“物理教育”两本体之间计算了相似度,并在此基础上建立本体映射的效率要高于基于回归的本体学习算法、基于快速排序的本体算法和基于标准排序方法的本体算法。

3 结语

分离 Bregman 方法是通过扩展拉格朗日函数

鞍点的计算来得到原优化模型最优解的方法,其鞍点又通过原始优化问题和对偶优化问题的交替迭代得到. 该方法被广泛应用于图像处理、信号系统、人工智能和各种机器学习领域. 本文通过梯度下降策略得到对偶优化问题的解,并通过软阈值算子的应用得到原始优化问题的解,最终得到最优本体稀疏向量. 实验结果表明该算法有广泛的本体应用前景.

[参考文献]

[1]高炜,兰美辉. 基于排序学习方法的本体映射算法[J]. 微电子学与计算机,2011,28(9):59-61.

[2]高炜,梁立,张云港. 基于图学习的本体概念相似度计算[J]. 西南师范大学学报:自然科学版,2011,36(4):64-67.

[3]高炜,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.

[4]高炜,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.

[5]高炜,梁立,徐天伟,等. 半监督 k-部排序算法及在本体中的应用[J]. 中北大学学报:自然科学版,2013,34(2):140-146.

[6]高炜,梁立,徐天伟. 基于正则化瑞利系数的半监督 k-部排序学习算法及应用[J]. 西南师范大学学报:自然科学版,2014,39(4):124-128.

[7]彭波,徐天伟,李臻,等. 迭代拉普拉斯半监督学习本体算法[J]. 计算机工程与科学,2014,36(11):2164-2168.

[8]GOLDSTEIN T,OSHER S. The split Bregman method for L1-regularized problems[J]. Imaging Sci,2009,2(2):323-343.

[9]CAI J F,OSHER S,SHEN Z. Split bregman methods and frame based image restoration[J]. Multiscale Model Simul,2009,8(2):337-369.

[10]CRASWELL N,HAWKING D. Overview of the TREC 2003 web track[C]//Proceedings of the Twelfth Text Retrieval Conference. Gaithersburg: NIST Special Publication,2003:78-92.

[11]GAO Y,GAO W. Ontology similarity measure and ontology mapping via learning optimization similarity function[J]. International Journal of Machine Learning and Computing,2012,2(2):107-112.

[12]HUANG X,XU T,GAO W,et al. Ontology similarity measure and ontology mapping via fast ranking method[J]. International Journal of Applied Physics and Mathematics,2011,1(1):54-59.