

文章编号: 1000-8608(2010)05-0788-06

基于 Logistic 回归模型和凝聚函数的多示例学习算法

贺建军*, 王欣, 顾宏, 王哲龙

(大连理工大学 控制科学与工程学院, 辽宁 大连 116024)

摘要: 鉴于很多实际问题都可以转化到多示例框架下求解, 多示例学习越来越受到机器学习领域内学者们的关注。提出了一个基于 Logistic 回归模型的多示例学习算法。首先定义了一个新的似然函数来表示每个包的标签与其示例的隐含标签之间的关系, 然后利用凝聚函数把该似然函数转化为一个光滑的凹函数, 从而使问题可以用常用的无约束优化方法快速求解。在一些标准数据集和一个文本分类问题上的实验结果表明, 所提算法要优于其他常用多示例学习算法。

关键词: 多示例学习; Logistic 回归模型; 凝聚函数; 文本分类

中图分类号: TP181 **文献标志码:** A

0 引言

多示例学习(multi-instance learning, MIL)最早是由 Dietterich 等^[1] 在对药物活性预测(drug activity prediction)问题的研究中提出来的。随着该领域内研究的不断深入, 多示例学习也被更广泛地应用到物体检测^[2]、图像检索^[3] 和 Web 挖掘^[4,5] 等诸多方面。在多示例学习问题中, 样本——通常被称为包(bag), 由多个特征向量组成, 而每个特征向量称为一个示例; 训练集由具有标签(label)的包组成, 而包中的各个示例并没有标签。通常情况下, 假设包中每个示例都有一个隐含的标签, 如果一个包中至少有一个正示例, 则这个包的标签即为正, 反之, 则该包的标签为负。为方便问题描述, 称上面的假设为标准的多示例学习假设。

APRs 算法^[1] 是最早的多示例学习算法。随后, Maron 等^[6] 对 APRs 算法的思想做了进一步的改进提出了多样性密度(diverse density, DD)算法。然而, 由于多样性密度空间中存在多个局部极小点, 将每一个正包示例都作为初始点进行一次搜索, 会使该算法的训练时间开销相当大。Zhang 等^[7] 将 DD 算法和 EM 算法相结合, 提出了 EM-DD 算法。与 DD 算法相比, EM-DD 算法的计

算时间花费有所减少。在提出多示例学习的概念时, Dietterich 等就指出“当前一个非常值得研究的课题是如何对传统的机器学习算法进行修改, 使它们可以处理多示例学习任务”, 该问题对多示例学习的研究起到了推动作用。经过十多年的研 究, 常用的机器学习算法基本上都有了其多示例学习版本。例如, 多示例版本的支持向量机(SVM)算法有 Andrews 等^[8] 提出的 MI-SVM 和 mi-SVM 算法, Chen 等^[9] 提出的 MILES 算法及 Li 等^[3] 提出的 Ins-KI-SVM 和 Bag-KI-SVM 算法等; 多示例版本的神经网络算法有 BP-MIP 算法^[10] 和 RBF-MIP 算法^[11]; 多示例版本的核 miGraph-Kernel^[12] 和 Marginalized MI-Kernel^[13] 等; Cheung 等^[14] 提出了一个多示例版本的正则化框架(regularization framework); Zhou 等^[15] 则将集成学习(ensemble learning)技术引入到了多示例学习中。

作为一种常用的机器学习技术, Logistic 回归模型也被应用到了多示例学习中。Xu 等^[16] 建立了最早的基于 Logistic 回归模型的多示例算法。在该算法中, 每一个包的概率被定义为其示例的概率的均值, 而这与标准的多示例学习假设“每个包只要至少包含一个正示例则该包就是正包”

不相符.之后,Ray等^[17]提出用 softmax 函数来处理包的概率和其示例的概率之间的关系;Raykar 等^[18]提出用与 DD 算法一样的模型定义包的概率.后面的两个算法虽然满足标准的多示例学习假设,但是建立的目标函数都是非凸的,遇到了与 DD 算法一样的求解困难.

本文通过定义一个新的关系函数来建立包的概率和其示例的概率之间的联系,提出一种既能充分体现标准的多示例学习假设,又能使目标函数为凹函数的基于 Logistic 回归模型的多示例学习算法.在标准的多示例学习假设中,一个包是正包当且仅当该包至少包含一个正示例.这就意味着每个包都有一个关键的示例控制着该包的标签,因此可以定义每个包为正包的概率为其所含示例为正示例的概率的最大值.由于最大值函数不是一个光滑函数,本文提出用凝聚函数来逼近它.最后从理论上证明所建立的目标函数是凹函数.

1 模型建立

令 $\mathcal{X} = \mathbf{R}^d$ 表示示例空间, $\mathcal{D} = \{(\mathbf{B}_i, y_i) \mid 1 \leq i \leq n\}$ 表示训练集, 其中 $\mathbf{B}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{in_i}\}$ 是包含 n_i 个示例的包, $y_i \in \{+1, -1\}$ 是包 \mathbf{B}_i 的标签, $\mathbf{x}_{ij} \in \mathcal{X}$ 表示第 i 个包的第 j 个示例. 多示例学习的任务就是要利用训练集寻找一个能够给出未标记包的正确标签的函数.

因为在标准的多示例学习假设中,训练集中的每个示例都有隐含的标签,因此考虑用定义在示例空间 \mathcal{X} 上的线性判别函数 $f(\bar{\mathbf{x}}) = \mathbf{W} \begin{pmatrix} \bar{\mathbf{x}} \\ 1 \end{pmatrix}$ 来判断示例的标签,如果 $f(\bar{\mathbf{x}}) > 0$ 则示例 $\bar{\mathbf{x}}$ 的标签为正,反之则为负.为了公式书写方便,用齐次坐标表示法表示每个示例,从而可以把 $f(\bar{\mathbf{x}})$ 直接写作 $f(\mathbf{x}) = \mathbf{Wx}$. 利用 Logistic 函数 $\sigma(t) = 1/(1 + e^{-t})$,可以定义如下的概率:

$$P(y = 1 \mid \mathbf{x}, \mathbf{W}) = \sigma(f(\mathbf{x})) = 1/(1 + e^{-f(\mathbf{x})}) \quad (1)$$

另外,在标准的多示例学习假设中,一个包在当且仅当至少包含一个正示例时为正包.这就意味着每个包中都有一个关键的示例控制着该包的标签.于是可以用示例中正标签概率的最大值代表该包为正的概率:

$$P(y_i = 1 \mid \mathbf{B}_i, \mathbf{W}) = \max_{j \in \{1, 2, \dots, n_i\}} \sigma(f(\mathbf{x}_{ij})) = \sigma(\max_{j \in \{1, 2, \dots, n_i\}} f(\mathbf{x}_{ij})) \quad (2)$$

利用 $P(y_i = -1 \mid \mathbf{B}_i, \mathbf{W}) = 1 - P(y_i = 1 \mid \mathbf{B}_i, \mathbf{W})$ 可以推导出, \mathbf{B}_i 为负包的概率:

$$P(y_i = -1 \mid \mathbf{B}_i, \mathbf{W}) = \frac{1}{1 + e^{\max_{j \in \{1, 2, \dots, n_i\}} f(\mathbf{x}_{ij})}} = \sigma(-\max_{j \in \{1, 2, \dots, n_i\}} f(\mathbf{x}_{ij})) \quad (3)$$

综合式(2)、(3)可得

$$P(y_i \mid \mathbf{B}_i, \mathbf{W}) = \sigma(\max_{j \in \{1, 2, \dots, n_i\}} f(\mathbf{x}_{ij})) \quad (4)$$

下面用最大似然估计法求解参数 \mathbf{W} . 由于训练集中各个包是相互独立的,似然函数可以写为

$$L(\mathbf{W} \mid \mathcal{D}) = \prod_{i=1}^n P(y_i \mid \mathbf{B}_i, \mathbf{W}) = \prod_{i=1}^n \sigma(\max_{j \in \{1, 2, \dots, n_i\}} f(\mathbf{x}_{ij})) \quad (5)$$

因此,可以通过最大化对数似然函数 $\ln L$ 求得参数 \mathbf{W} ,即

$$\begin{aligned} \mathbf{W} = \arg \max_{\mathbf{W}} \ln L(\mathbf{W} \mid \mathcal{D}) = \\ \arg \max_{\mathbf{W}} \left\{ - \sum_{i=1}^n \ln (1 + e^{-y_i \max_{j \in \{1, 2, \dots, n_i\}} f(\mathbf{x}_{ij})}) \right\} = \\ \arg \max_{\mathbf{W}} \left\{ - \sum_{i=1}^n \ln (1 + e^{-y_i \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}}) \right\} \end{aligned} \quad (6)$$

2 对数似然函数逼近

由于 $\max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}$ 是非光滑函数,式(6)中的对数似然函数也是一个非光滑函数,不能用基于梯度的方法求解,而其他方法又需要花费大量的时间.一种直觉上的想法就是先用一个光滑函数逼近对数似然函数 $\ln L$,然后再用基于梯度的方法求解.

凝聚函数 (aggregate function) 又名指数惩罚函数 (exponential penalty function)

$$G_p(\mathbf{x}) = \frac{1}{p} \ln \left(\sum_{i=1}^m e^{pg_i(\mathbf{x})} \right) \quad (7)$$

是由 Li^[19] 最早在求解非线性规划问题时利用代理约束概念和最大熵原理推导出来的,这里 p 是一个正的控制参数. 该函数和对应的最大值函数 $G(\mathbf{x}) = \max_{i \in \{1, 2, \dots, m\}} g_i(\mathbf{x})$ 有如下的关系式:

$$G(\mathbf{x}) \leq G_p(\mathbf{x}) \leq G(\mathbf{x}) + (\ln m)/p \quad (8)$$

因此,对于有限正整数 m ,当 p 趋于正无穷时, $G_p(\mathbf{x})$ 能单调一致地逼近 $G(\mathbf{x})$.

令 $n_0 = \max_{i \in \{1, 2, \dots, n\}} n_i$, 即用 n_0 表示训练集中包

$$\text{的示例个数的最大值}, G_p^i(\mathbf{W}) = \frac{1}{p} \ln \left(\sum_{j=1}^{n_i} e^{p\mathbf{Wx}_{ij}} \right),$$

则利用式(8)可得

$$\begin{aligned} \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij} &\leq G_p^i(\mathbf{W}) \leq \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij} + \\ \frac{\ln n_i}{p} &\leq \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij} + \\ \frac{\ln n_0}{p}; \quad i = 1, 2, \dots, n \end{aligned} \quad (9)$$

由于 n_0 是有限的, 对任意的阈值 $\epsilon > 0$, 都存在 p 使得

$$|G_p^i(\mathbf{W}) - \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}| < \epsilon; \quad i = 1, 2, \dots, n \quad (10)$$

因此可以用 $G_p^i(\mathbf{W}) = \frac{1}{p} \ln \left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)$ 代替式(6) 中的 $\max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}$, $i = 1, 2, \dots, n$, 从而可得

$$\begin{aligned} \mathbf{W} = \arg \max_{\mathbf{W}} \ln L(\mathbf{W} | \mathbf{D}) = \\ \arg \max_{\mathbf{W}} \left\{ - \sum_{i=1}^n \ln \left(1 + e^{-y_i G_p^i(\mathbf{x})} \right) \right\} = \\ \arg \max_{\mathbf{W}} \left\{ - \sum_{i=1}^n \ln \left(1 + \left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)^{-y_i/p} \right) \right\} \end{aligned} \quad (11)$$

下面讨论式(11) 中的对数似然函数

$$\ln L(\mathbf{W} | \mathbf{D}) = - \sum_{i=1}^n \ln \left(1 + \left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)^{-y_i/p} \right) \quad (12)$$

的性质. 显然式(12) 是一个连续可微函数.

对式(12) 求导可得

$$\frac{\partial \ln L}{\partial w_k} = \sum_{i=1}^n \frac{y_i}{1 + \left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)^{y_i/p}} \sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} x_{ij}(k); \quad k = 1, 2, \dots, d+1 \quad (13)$$

$$\frac{\partial^2 \ln L}{\partial w_k \partial w_l} = - \sum_{i=1}^n (g_{il} + g_{li}); \quad k, l = 1, 2, \dots, d+1 \quad (14)$$

式中: w_k 表示 \mathbf{W} 的第 k 个元素, $x_{ij}(k)$ 表示 x_{ij} 的第 k 个元素;

$$\begin{aligned} g_{il} = & \frac{\left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)^{y_i/p}}{\left(1 + \left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)^{y_i/p} \right)^2} \times \\ & \sum_{j_0=1}^{n_i} \frac{x_{ij_0}(k)}{1 + \sum_{j \neq j_0} e^{\rho \mathbf{W}(x_{ij} - x_{ij_0})}} \times \\ & \sum_{j_0=1}^{n_i} \frac{x_{ij_0}(l)}{1 + \sum_{j \neq j_0} e^{\rho \mathbf{W}(x_{ij} - x_{ij_0})}} \end{aligned}$$

$$g_{il} = \frac{y_i}{1 + \left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)^{y_i/p}} \times \\ \sum_{j_0=1, j_0 \neq j_1}^{n_i} \frac{x_{ij_0}(k)(x_{ij_1}(l) - x_{ij_0}(l)) p e^{\rho \mathbf{W}(x_{ij_1} - x_{ij_0})}}{\left(1 + \sum_{j \neq j_0} e^{\rho \mathbf{W}(x_{ij} - x_{ij_0})} \right)^2}$$

观察 g_{il} 的表达式中各项, 当 p 趋近于正无穷大时, 有下式成立:

$$\frac{\left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)^{y_i/p}}{\left(1 + \left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)^{y_i/p} \right)^2} = \frac{e^{y_i \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}}}{(1 + e^{y_i \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}})^2} \quad (15)$$

$$\frac{x_{ij_0}(k)}{1 + \sum_{j \neq j_0} e^{\rho \mathbf{W}(x_{ij} - x_{ij_0})}} = \begin{cases} \frac{x_{ij_0}(k)}{|S|}; & j_0 \in S, S = \arg \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij} \\ 0; & \text{其他} \end{cases} \quad (16)$$

同样, 当 p 趋近于正无穷大时, g_{il} 的表达式中各项有如下性质:

$$\begin{aligned} \frac{y_i}{1 + \left(\sum_{j=1}^{n_i} e^{\rho \mathbf{Wx}_{ij}} \right)^{y_i/p}} &= \frac{y_i}{1 + e^{y_i \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}}} \quad (17) \\ \frac{x_{ij_0}(k)(x_{ij_1}(l) - x_{ij_0}(l)) p e^{\rho \mathbf{W}(x_{ij_1} - x_{ij_0})}}{\left(1 + \sum_{j \neq j_0} e^{\rho \mathbf{W}(x_{ij} - x_{ij_0})} \right)^2} &= \\ \begin{cases} \infty; \mathbf{W}(x_{ij_1} - x_{ij_0}) = 0, j_0 \in \arg \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij} \\ 0; \text{其他} \end{cases} \end{aligned} \quad (18)$$

综合式(14) ~ (18), 可以得到 $\frac{\partial^2 \ln L}{\partial w_k \partial w_l}$ 的如下性质:

假设对任意的 i , 集合 $S = \arg \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}$ 的势为 1, 即对任意的 i 有唯一的 $j_0 \in \{1, 2, \dots, n_i\}$ 使得 $\mathbf{Wx}_{ij_0} = \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}$, 则当 p 趋近于正无穷大时

$$\begin{aligned} \frac{\partial^2 \ln L}{\partial w_k \partial w_l} &= - \sum_{i=1}^{n_i} \frac{e^{y_i \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}}}{(1 + e^{y_i \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}})^2} \times \\ & x_{ij_0}(k) x_{ij_0}(l); \quad k, l = 1, 2, \dots, d+1 \end{aligned} \quad (19)$$

式(19) 可以写成矩阵形式

$$\nabla \nabla \ln L = - \sum_{i=1}^{n_i} \frac{e^{y_i \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}}}{(1 + e^{y_i \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}})^2} \mathbf{x}_{ij_0} \mathbf{x}_{ij_0}^T \quad (20)$$

根据式(20)可知, $\nabla \nabla \ln L$ 是一个半负定矩阵. 因此, 在假设条件“对任意的 i , 集合 $S = \arg \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}$ 的势为 1”下, 对数似然函数式(12)是凹函数. 由于其对立假设“存在 i , 使得集合 $S = \arg \max_{j \in \{1, 2, \dots, n_i\}} \mathbf{Wx}_{ij}$ 的势不为 1”是一个小概率事件, 在实际问题求解的过程中可以抛开这个假设不管, 直接把对数似然函数式(12)看做凹函数求解.

3 模型求解

前文已证明了本文建立的对数似然函数是一个凹函数, 因此可以用常用的优化方法求解问题(11). 本文采用拟牛顿法求解, 迭代公式如下:

$$\begin{aligned} \mathbf{W}^{(k+1)} &= \mathbf{W}^{(k)} - \lambda^{(k)} \mathbf{H}^{(k)} \nabla \psi(\mathbf{W}^{(k)}) ; \\ \mathbf{H}^{(k+1)} &= \mathbf{H}^{(k)} + \frac{(\mathbf{s} - \mathbf{H}^{(k)} \mathbf{y}) \mathbf{s}^T + \mathbf{s} (\mathbf{s} - \mathbf{H}^{(k)} \mathbf{y})^T}{(\mathbf{s}, \mathbf{y})} - \\ &\quad \frac{(\mathbf{s} - \mathbf{H}^{(k)} \mathbf{y}, \mathbf{y}) \mathbf{s} \mathbf{s}^T}{(\mathbf{s}, \mathbf{y})^2} ; \\ \mathbf{s} &= \mathbf{W}^{(k+1)} - \mathbf{W}^{(k)} ; \\ \mathbf{y} &= \nabla \psi(\mathbf{W}^{(k+1)}) - \nabla \psi(\mathbf{W}^{(k)}) \end{aligned} \quad (21)$$

式中

$$\psi(\mathbf{W}) = -\ln L(\mathbf{W} | \mathbf{D})$$

$$\lambda^{(k)} = \arg \min_{\lambda} \psi(\mathbf{W}^{(k)} - \lambda \mathbf{H}^{(k)} \nabla \psi(\mathbf{W}^{(k)}))$$

由于 $\lambda^{(k)} = \arg \min_{\lambda} \psi(\mathbf{W}^{(k)} - \lambda \mathbf{H}^{(k)} \nabla \psi(\mathbf{W}^{(k)}))$ 是一个一维的凸优化问题, 用单纯形法(Nelder-Mead Algorithm)^[20]对其求解. 下面给出了具体的算法流程.

训练

输入: 训练样本集 $\mathbf{D} = \{(\mathbf{B}_i, y_i) \mid 1 \leq i \leq n\}$, 控制参数 p ;

输出: \mathbf{W} ;

初始化: \mathbf{W}^0 为任意向量, \mathbf{H}^0 为单位矩阵;

(1) 计算最优步长 $\lambda^{(k)} = \arg \min_{\lambda} \psi(\mathbf{W}^{(k)} - \lambda \mathbf{H}^{(k)} \nabla \psi(\mathbf{W}^{(k)}))$;

(2) 利用式(21)计算 $\mathbf{W}^{(k+1)}$ 和 $\mathbf{H}^{(k+1)}$;

(3) 如果 $|\psi(\mathbf{W}^{(k+1)}) - \psi(\mathbf{W}^{(k)})| < \epsilon$, 则输出结果 $\mathbf{W}^{(k+1)}$, 否则, 转到(1).

测试

输入: 待测样本 $\mathbf{B}_* = \{\mathbf{x}_{*1}, \mathbf{x}_{*2}, \dots, \mathbf{x}_{*n_*}\}$;

输出: $P(y = 1 \mid \mathbf{B}_*, \mathbf{W}) = \max_j \sigma(\mathbf{Wx}_{*j})$.

4 实验

首先在多个标准数据集上测试了本文的算

法. 这些数据集包括文献[1]中提供的麝香分子数据集 Musk1 和 Musk2, 文献[8]中提供的 Elephant、Fox 和 Tiger 数据集. Musk1 数据集包含 47 个正样本和 45 个负样本; Musk2 数据集包含 39 个正样本和 63 个负样本. Elephant、Fox 和 Tiger 数据集均包含 100 个正样本和 100 个负样本, 正样本数据是由包含该种动物的图片生成的, 负样本则由不包含该种动物的图片数据任意组成. 表 1 给出了这些数据集的详细信息. 表 2 给出了本文算法与其他算法在标准数据集上的实验结果, 每个数据集上的最好实验结果由粗体显示. 所有算法的实验结果都是采用十倍交叉验证法计算所得, 其他算法的实验结果引自相关文献. 从表 2 可以看出, 本文的算法在对 Elephant 和 Fox 数据集的实验都得到了最高的准确率, 在其他数据集上的实验虽准确率没有达到最高, 但仍然与最好结果相近, 因此本文的算法在这些标准数据集上的整体表现要优于其他算法.

表 1 标准数据集的详细信息

Tab. 1 The details of benchmark data sets

数据集	示例维数	正包个数	正包例个数	负包个数	负包例个数
Musk1 ^[1]	166	47	207	45	269
Musk2 ^[1]	166	39	1 017	63	5 581
Elephant ^[8]	230	100	762	100	629
Fox ^[8]	230	100	647	100	673
Tiger ^[8]	230	100	544	100	676

表 2 在标准数据集上的实验结果

Tab. 2 Experimental results on benchmark data sets

算法	准确率/%				
	Musk1	Musk2	Elephant	Fox	Tiger
本文的算法	84.7	86.2	85.0	60.5	82.0
EM-DD ^[7]	84.8	84.9	78.3	56.1	72.1
mi-SVM ^[8]	87.4	83.6	82.0	58.2	78.9
MI-SVM ^[8]	77.9	84.3	81.4	59.4	84.2
MICCLLR_SVML ^[21]	86.0	82.2	80.7	52.1	79.1
MICCLLR_Vote ^[21]	91.6	86.1	81.7	56.2	78.5
MICA ^[22]	84.4	90.5	80.5	58.7	82.6

为了进一步验证本文算法的性能, 还针对一个文本分类问题进行了测试. 该文本问题包括 20 个文本数据集, 每个数据集由 50 个正样本和 50 个负样本组成, 关于这些文本数据集的详细信息, 可以参考文献[12]. 将本文算法与文献[12]中算

法(MI-Kernel 和 miGraph)的实验结果进行了比较,结果如表 3 所示,所有算法的实验结果都是采用 10 次十倍交叉验证法计算所得的平均精度,最好的实验结果由粗体显示。从表 3 可以看出,本文提出的算法分别在 13 个数据集上的准确率要高于其他两个算法,在 1 个数据集上与 miGraph 算法的精度一样但要优于 MI-Kernel 算法,另外在所有数据集上的结果均要优于 MI-Kernel 算法。因此本文的算法在文本分类问题中的整体表现要优于其他两个算法。

表 3 文本分类数据集上的实验结果

Tab. 3 Experimental results on text categorization

数据集	准确率/%		
	本文的算法	MI-Kernel	miGraph
alt. atheism	68.4	60.2	65.5
comp. graphics	65.3	47.0	77.8
comp. os. ms-windows. misc	65.0	51.0	63.1
comp. sys. ibm. pc. hardware	59.5	46.9	59.5
comp. sys. mac. hardware	63.7	44.5	61.7
comp. windows. x	72.3	50.8	69.8
misc. forsale	59.5	51.8	55.2
rec. autos	64.0	52.9	72.0
rec. motorcycles	79.4	50.6	64.0
rec. sport. baseball	71.0	51.7	64.7
rec. sport. hockey	76.0	51.3	85.0
sci. crypt	73.3	56.3	69.6
sci. electronics	56.6	50.6	87.1
sci. med	62.3	50.6	62.1
sci. space	69.3	54.7	75.7
soc. religion. christian	67.1	49.2	59.0
talk. politics. guns	64.9	47.7	58.5
talk. politics. mideast	76.0	55.9	73.6
talk. politics. misc	64.4	51.5	70.4
talk. religion. misc	66.4	55.4	63.6

5 结 论

本文在标准多示例学习的假设条件下,建立了一个基于 Logistic 回归模型的多示例学习算法。算法的创新之处是定义了一个新的似然函数,然后通过引入凝聚函数去逼近最大值函数,把对数似然函数转化为一个光滑的凹函数,从而使问题可以用常用的无约束优化方法快速求解,不存在局部极值的问题。在标准数据集和实际文本分类问题上的实验结果表明,本文的算法要优于其他算法。下一步将研究利用高斯过程推广该算法到非线性情形,从而使得该算法更具有竞争力。

参 考 文 献:

- [1] DIETTERICH T G, LATHROP R H, LOZANO-PÉREZ T. Solving the multiple-instance problem with axis-parallel rectangles [J]. *Artificial Intelligence*, 1997, **89**(1-2):31-71
- [2] VIOLA P, PLATT J, ZHANG C. Multiple instance boosting for object detection [C] // *Advances in Neural Information Processing Systems* **18**. Cambridge: MIT Press, 2006:1419-1426
- [3] LI Y F, KWOK J T, TSANG I W, et al. A convex method for locating regions of interest with multi-instance learning [C] // *Machine Learning and Knowledge Discovery in Databases-European Conference, ECML PKDD 2009, Proceedings*. Bled: Springer-Verlag, 2009:15-30
- [4] ZARRA A, VVENTURA S, ROMERO C, et al. Multi-instance genetic programming for web index recommendation [J]. *Expert System and Applications*, 2009, **36**(9):11470-11479
- [5] ZHOU Z H, JIANG K, LI M. Multi-instance learning based web mining [J]. *Applied Intelligence*, 2005, **22**(2):135-147
- [6] MARON O, LOZANO-PÉREZ T. A framework for multiple-instance learning [C] // *Advances in Neural Information Processing System* **10**. Cambridge: MIT Press, 1998:570-576
- [7] ZHANG Q, GOLDMAN S A. EM-DD:an improved multiple-instance learning technique [C] // *Advances in Neural Information Processing Systems* **14**. Cambridge: MIT Press, 2002:1073-1080
- [8] ANDREWS S, TSOCHANTARIDIS I, HOFMANN T. Support vector machines for multiple-instance learning [C] // *Advances in Neural Information Processing Systems* **15**. Cambridge: MIT Press, 2003: 577-584
- [9] CHEN Y, BI J, WANG J. MILES:Multiple-instance learning via embedded instance selection [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, **28**(12):1931-1947
- [10] ZHOU Z H, ZHANG M L. Neural networks for multi-instance learning [C] // *Proceedings of the International Conference on Intelligent Information Technology*. Beijing: ICIIT, 2002
- [11] ZHANG M L, ZHOU Z H. Adapting RBF neural networks to multi-instance learning [J]. *Neural Processing Letters*, 2006, **23**(1):1-26
- [12] ZHOU Z H, SUN Y Y, LI Y F. Multi-instance

- learning by treating instances as non-i. i. d. samples [C] // **Proceedings of the 26th International Conference on Machine Learning (ICML'09)**. Montreal:Omnipress, 2009:1249-1256
- [13] KWOK J T, CHEUNG P M. Marginalized multi-instance kernels [C] // **Proceedings of the 20th International Joint Conferences on Artificial Intelligence**. San Francisco:Morgan Kaufmann Publishers Inc., 2007:901-906
- [14] CHEUNG P M, KWOK J T. A regularization framework for multiple-instance learning [C] // **Proceedings of the 23rd International Conference on Machine Learning**. New York:ACM, 2006:193-200
- [15] ZHOU Z H, ZHANG M L. Ensembles of multi-instance learners [C] // **Proceedings of the 14th European Conference on Machine learning (LNAI 2837)**. Berlin:Springer-Verlag, 2003:492-502
- [16] XU X, FRANK E. Logistic regression and boosting for labeled bags of instances [C] // **Proceedings of the Pacific-Asia Conference on Machine Learning and Data Mining**. Berlin:Springer-Verlag, 2004:272-281
- [17] RAY S, CRAVEN M. Supervised versus multiple instance learning:an empirical comparison [C] // **Proceedings of the 22nd International Conference on Machine Learning**. Bonn:Association for Computing Machinery, 2005:697-704
- [18] RAYKAR V C, KKRISHNAERAM B, DUNDAR J, et al. Bayesian multiple instance learning: automatic feature selection and inductive transfer [C] // **Proceedings of the 25th International Conference on Machine Learning**. New York : Association for Computing Machinery, 2008:808-815
- [19] LI X S. An aggregate function method for nonlinear programming [J]. **Science in China Series A-Mathematics**, 1991, 34(12):1467-1473
- [20] WANG Z L, HE J J, SHANG H, et al. Forward kinematics analysis of a six-DOF Stewart platform using PCA and NM algorithm [J]. **Industrial Robot: An International Journal**, 2009, 36(5):448-460
- [21] EL-MANZALAWY Y, HONAVAR V. MICCLLR:Multiple-instance learning using class conditional log likelihood ratio. [C] // **Proceedings of the 12th International Conference on Discovery Science**. Heidelberg:Springer Berlin, 2009:80-91
- [22] MANGASARIAN O L, WILD E W. Multiple instance classification via successive linear programming [J]. **Journal of Optimization Theory and Applications**, 2008, 137(3):555-568

An algorithm for multi-instance learning based on Logistic regression model and aggregate function

HE Jian-jun*, WANG Xin, GU Hong, WANG Zhe-long

(School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Since many real-world problems could be transformed and solved in multi-instance learning (MIL) framework, the researchers in the field of machine learning have paid more and more attention to MIL. Based on the Logistic regression model, a new MIL algorithm is developed. Firstly, in this algorithm, a new likelihood function is defined to build the relationship between the bag's class label and its instances' hidden class labels. Then, the aggregate function is used to transform the likelihood function to a smooth concave function, thereby the problem can be solved by a general unconstrained optimization method. Experimental results of both the benchmark data sets and a text categorization problem show that the proposed algorithm can achieve superior performance to the published MIL algorithms.

Key words: multi-instance learning; Logistic regression model; aggregate function; text categorization