

面向不平衡数据的逻辑回归偏标记学习算法

周 瑜, 顾 宏*

(大连理工大学 电子信息与电气工程学部, 辽宁 大连 116024)

摘要: 偏标记学习是近几年提出的新机器学习框架,已有的逻辑回归偏标记算法尚不能解决数据不平衡问题.建立了一种可以解决数据不平衡的逻辑回归模型偏标记学习算法.基本思想是在多元逻辑回归模型中定义新的似然函数以达到处理不平衡数据的目的.算法先根据训练集中各个类别样本所占比例定义了一个新的似然函数,之后通过逼近和求导等数学手段推导得到了能够求解的光滑的逻辑回归偏标记学习模型.在UCI数据集和真实数据集上的仿真实验表明,所提算法在数据存在不平衡问题时提高了样本的平均分类精度.

关键词: 偏标记学习;数据不平衡;逻辑回归;阻尼牛顿法

中图分类号: TP391

文献标识码: A

doi: 10.7511/dllgxb201702011

0 引言

偏标记学习是近几年提出的一种新的机器学习框架,国内外学者对它的研究已经有了一定的成果.最早的文献是Grandvalet对逻辑回归模型进行的拓展研究^[1],其提出了一种偏标记学习算法;随后Jin等^[2]将偏标记学习归结为一种新的机器学习框架.新的学习框架的提出促进了众多学者对偏标记学习的研究, k 近邻^[3]、最大间隔^[4]、线性支持向量机^[5-6]等方法均被用于偏标记学习算法研究.这些方法都是通过定义新的损失函数来改进传统分类模型,使其可以处理偏标记学习问题.但在很多的实际应用问题中,各个类别的样本数量之间是极度不平衡的,如在蛋白质亚细胞定位预测问题中^[7],数据集中两类数目差别近百倍.数据集的这种不平衡(也称数据不平衡)问题对学习算法性能具有很大的影响,通常会导致算法的分类面偏向少数类一侧,从而造成预测精度大幅下降,特别是对少数类样本的预测精度要远远低于多数类样本^[8].目前已有的偏标记学习算法都没有考虑数据的不平衡性.因此,考虑数据不平衡问题的偏标记学习算法也是将偏标记学习技术推向更加实用化所需要解决的关键问

题.本文建立一种逻辑回归偏标记学习算法,以期提高不平衡数据的平均分类精度.

1 逻辑回归偏标记学习模型

1.1 模型建立

偏标记学习的定义如下:

设 $\mathbf{X}$ 为样本的特征空间, $Y = \{1, 2, \dots, l\}$ 为类别标记集合.利用训练集 $\mathbf{D} = \{(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \dots, (\mathbf{x}_n, Y_n)\}$ (其中 $\mathbf{x}_i \in \mathbf{X}$ 是样本的特征向量; $Y_i = \{y_{i1}, y_{i2}, \dots, y_{in_i}\} \subset Y$,是含样本 $\mathbf{x}_i$ 真实标记的一个集合)确定一个函数 $f: \mathbf{X} \rightarrow Y$,使得 f 可以正确输出新(待预测)样本 $\mathbf{x}^* \in \mathbf{X}$ 的类别标记.

偏标记学习是一个多分类问题,与传统的多分类模型一样,都是在特征空间 $\mathbf{X}$ 中为每个类别标记 y_j ($j = 1, 2, \dots, Q$)定义一个潜变量函数 $f_{y_j}(\bar{\mathbf{x}}) = \bar{\mathbf{w}}_{y_j}^T \bar{\mathbf{x}} + b_{y_j}$,也可以写成 $f_{y_j}(\bar{\mathbf{x}}) = \mathbf{w}_{y_j}^T \mathbf{x}$ 的形式,其中 $\mathbf{w}_{y_j}^T = (\bar{\mathbf{w}}_{y_j}^T \quad b_{y_j})$, $\mathbf{x} = \begin{pmatrix} \bar{\mathbf{x}} \\ 1 \end{pmatrix}$, 样本 $\bar{\mathbf{x}} \in \mathbf{X}$ 的真实类别标记是 y_k ($k = 1, 2, \dots, Q$)的概率可以由这些潜变量函数 $\{f_{y_j} | j = 1, 2, \dots, Q\}$ 在 $\mathbf{x}$ 上的值来确定,即 $p(y_k | \mathbf{x}, \mathbf{W}) = \frac{e^{f_{y_k}(\mathbf{x})}}{\sum_{j=1}^Q e^{f_{y_j}(\mathbf{x})}}$, $\mathbf{W} = \{\mathbf{w}_{y_j} | j = 1, 2, \dots, Q\}$, 则 $f_{y_j}(\mathbf{x}) = \mathbf{w}_{y_j}^T \bar{\mathbf{x}} + b_{y_j} =$

$\mathbf{W}_{y_j}^T(\mathbf{y}_j \otimes \mathbf{x})$, 其中 $\mathbf{y}_j$ 是第 j 个元素为 1, 其他元素为 0 的 Q 维列向量. 由于均值似然函数本质是假定 Y_i 的所有元素都是 $\mathbf{x}_i$ 的真实标记, 为了体现偏标记学习“样本的候选标记中有且只有一个标记是样本的真实标记”的概念, 本文定义一个新的最大值似然函数:

$$p(Y_i | \mathbf{x}_i, \mathbf{W}) = \max_{y \in Y_i} p(y | \mathbf{x}_i, \mathbf{W}) = \max_{y \in Y_i} \left(e^{r_y f_y(\mathbf{x}_i)} / \sum_{j=1}^Q e^{r_{y_j} f_{y_j}(\mathbf{x}_i)} \right) = e^{\max_{y \in Y_i} r_y f_y(\mathbf{x}_i)} / \sum_{j=1}^Q e^{r_{y_j} f_{y_j}(\mathbf{x}_i)} \quad (1)$$

式中: 权重系数 $r_y = 1/(1 + \gamma e^{-n_y/n})$, 其中 $0 < r_y < 1$, 对平衡数据集 $\gamma = 0$, 对不平衡数据集 $\gamma > 0$; n_y 表示训练集中第 y 类样本个数. 由最大值似然函数的概率性质可知, $\sum_{i=1}^n (e^{r_y f_y(\mathbf{x}_i)} / \sum_{j=1}^Q e^{r_{y_j} f_{y_j}(\mathbf{x}_i)}) = 1, 0 \leq e^{r_y f_y(\mathbf{x}_i)} / \sum_{j=1}^Q e^{r_{y_j} f_{y_j}(\mathbf{x}_i)} \leq 1, p(Y_i | \mathbf{x}_i, \mathbf{W}) (i = 1, 2, \dots, n)$ 相互独立. 可以看出 n_y 越小, r_y 越大, $p(y | \mathbf{x}_i, \mathbf{W})$ 的值就越大, 因为增大了小类别样本 $\mathbf{x} \in \mathbf{X}$ 的真实类别标记是 $y_k (k = 1, 2, \dots, Q)$ 的概率, 从而使得模型能处理不平衡问题.

模型参数 $\mathbf{W}$ 可以通过最大化对数联合似然函数 $\ln L(\mathbf{W} | D) = \sum_{i=1}^n \ln p(Y_i | \mathbf{x}_i, \mathbf{W})$ 求得, 即

$$\mathbf{W} = \arg \max_{\mathbf{W}} \left\{ \ln L(\mathbf{W} | D) - \rho \sum_{j=1}^Q \|\mathbf{W}\| / 2 \right\} \quad (2)$$

其中 $\rho \sum_{j=1}^Q \|\mathbf{W}\| / 2$ 是为避免过拟合所加的正则化项.

由于 $\max(\cdot)$ 函数不可导, 用凝聚函数逼近最大值似然函数. 当 $p \rightarrow +\infty$ 时, 有

$$\frac{1}{p} \sum_{i=1}^n \ln \left(\sum_{y \in Y_i} e^{p f_y(\mathbf{x}_i)} \right) \approx \sum_{i=1}^n \max_{y \in Y_i} \{ f_y(\mathbf{x}_i) \}$$

则

$$\mathbf{W} = \arg \max_{\mathbf{W}} \left\{ \frac{1}{p} \sum_{i=1}^n \ln \left(\sum_{k \in Y_i} e^{p \mathbf{W}^T (\mathbf{r} \cdot \mathbf{y}_k \otimes \mathbf{x}_i)} \right) - \sum_{i=1}^n \ln \left(\sum_{j=1}^Q e^{p \mathbf{W}^T (\mathbf{r} \cdot \mathbf{y}_j \otimes \mathbf{x}_i)} \right) - \frac{\rho}{2} \mathbf{W}^T \bar{\mathbf{E}} \mathbf{W} \right\} \quad (3)$$

其中 $\mathbf{r} = (r_1 \ r_2 \ \dots \ r_Q)^T$, $\bar{\mathbf{E}}$ 为将与 $\mathbf{W}$ 同阶的单位矩阵与 $\{\mathbf{b}_{y_j} | j = 1, 2, \dots, Q\}$ 对应的对角元素置为 0 后所得的矩阵. 为表达方便, 设 $\mathbf{Z}(\mathbf{W}) = \frac{1}{p} \times$

$$\sum_{i=1}^n \ln \left(\sum_{k \in Y_i} e^{p \mathbf{W}^T (\mathbf{r} \cdot \mathbf{y}_k \otimes \mathbf{x}_i)} \right) - \sum_{i=1}^n \ln \left(\sum_{j=1}^Q e^{p \mathbf{W}^T (\mathbf{r} \cdot \mathbf{y}_j \otimes \mathbf{x}_i)} \right) -$$

$$\frac{\rho}{2} \mathbf{W}^T \bar{\mathbf{E}} \mathbf{W}.$$

下证 $\mathbf{Z}(\mathbf{W})$ 是光滑的凹函数. 设 $L = \frac{1}{p} \times$

$\sum_{i=1}^n \ln \left(\sum_{k \in Y_i} e^{p \mathbf{W}_k^T \mathbf{x}_i} \right) - \sum_{i=1}^n \ln \left(\sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} \right)$, w_{su} 是 $\mathbf{W}$ 第 s 行第 u 列的元素, 则有

$$\frac{\partial \ln L}{\partial w_{su}} = \begin{cases} \sum_{i=1}^n \left(e^{p \mathbf{W}_s^T \mathbf{x}_i} x_{iu} / \sum_{\xi=1}^n e^{p \mathbf{W}_{y_{i\xi}}^T \mathbf{x}_i} - e^{p \mathbf{W}_s^T \mathbf{x}_i} x_{iu} / \sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} \right); & s \in Y_i \\ \sum_{i=1}^n \left(-e^{p \mathbf{W}_s^T \mathbf{x}_i} x_{iu} / \sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} \right); & s \notin Y_i \end{cases}$$

当 $p \rightarrow \infty$ 时,

$$\frac{\partial \ln L}{\partial w_{su}} = \begin{cases} \sum_{i=1}^n \left(1 - e^{p \mathbf{W}_s^T \mathbf{x}_i} / \sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} \right) x_{iu}; & s = \arg \max_{c \in Y_i} \{ \mathbf{W}_c^T \mathbf{x}_i \} \\ \sum_{i=1}^n \left(-e^{p \mathbf{W}_s^T \mathbf{x}_i} / \sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} \right) x_{iu}; & s \neq \arg \max_{c \in Y_i} \{ \mathbf{W}_c^T \mathbf{x}_i \} \end{cases}$$

$$\frac{\partial^2 \ln L}{\partial w_{su} \partial w_{tv}} = \begin{cases} \sum_{i=1}^n \left(\frac{p e^{p \mathbf{W}_s^T \mathbf{x}_i} \sum_{\xi=1}^{n_i} e^{p \mathbf{W}_{y_{i\xi}}^T \mathbf{x}_i} - p e^{2p \mathbf{W}_s^T \mathbf{x}_i}}{\left(\sum_{\xi=1}^{n_i} e^{p \mathbf{W}_{y_{i\xi}}^T \mathbf{x}_i} \right)^2} - \frac{e^{p \mathbf{W}_s^T \mathbf{x}_i} \sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} - e^{2p \mathbf{W}_s^T \mathbf{x}_i}}{\left(\sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} \right)^2} \right) x_{iu} x_{iv}; & s, t \in Y_i, \text{ 且 } s = t \\ \sum_{i=1}^n \left(\frac{-p e^{p \mathbf{W}_s^T \mathbf{x}_i} e^{p \mathbf{W}_t^T \mathbf{x}_i}}{\left(\sum_{\xi=1}^{n_i} e^{p \mathbf{W}_{y_{i\xi}}^T \mathbf{x}_i} \right)^2} + \frac{e^{p \mathbf{W}_s^T \mathbf{x}_i} e^{p \mathbf{W}_t^T \mathbf{x}_i}}{\left(\sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} \right)^2} \right) \times x_{iu} x_{iv}; & s, t \in Y_i, \text{ 且 } s \neq t \\ \sum_{i=1}^n \left(\frac{e^{p \mathbf{W}_s^T \mathbf{x}_i} e^{p \mathbf{W}_t^T \mathbf{x}_i}}{\left(\sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} \right)^2} \right) x_{iu} x_{iv}; & s \in Y_i, t \notin Y_i \text{ 或 } t \in Y_i, s \notin Y_i \\ \sum_{i=1}^n \left(-\frac{e^{p \mathbf{W}_s^T \mathbf{x}_i} \sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} - e^{2p \mathbf{W}_s^T \mathbf{x}_i}}{\left(\sum_{j=1}^Q e^{p \mathbf{W}_j^T \mathbf{x}_i} \right)^2} \right) x_{iu} x_{iv}; & s, t \notin Y_i, \text{ 且 } s = t \end{cases}$$

当 $s \neq t, s \in Y_i$ 时,

$$\frac{pe^{\rho w_s^T x_i} e^{\rho w_t^T x_i}}{\left(\sum_{\xi=1}^{n_i} e^{\rho w_{y_{i\xi}}^T x_i}\right)^2} = \frac{pe^{\rho w_s^T x_i} e^{\rho w_t^T x_i}}{e^{2\rho \max_{c \in Y_i} \{w_c^T x_i\}}} = \frac{p}{e^{\rho \max_{c \in Y_i} \{ (w_c^T - w_s^T) x_i \}} e^{\rho \max_{c \in Y_i} \{ (w_c^T - w_t^T) x_i \}}} = 0$$

当 $s=t, s, t \in Y_i$ 时,

$$\frac{pe^{\rho w_s^T x_i} \left(\sum_{\xi=1}^{n_i} e^{\rho w_{y_{i\xi}}^T x_i} - e^{\rho w_s^T x_i}\right)}{\left(\sum_{\xi=1}^{n_i} e^{\rho w_{y_{i\xi}}^T x_i}\right)^2} = \frac{p \left(\frac{1}{\sum_{\xi=1}^{n_i} e^{\rho (w_{y_{i\xi}}^T - w_s^T) x_i}} - \frac{1}{\left(\sum_{\xi=1}^{n_i} e^{\rho (w_{y_{i\xi}}^T - w_s^T) x_i}\right)^2} \right)}{e^{\rho \max_{y_{i\xi} \in Y_i} \{ (w_{y_{i\xi}}^T - w_s^T) x_i \}}} - \frac{p}{\left(e^{\rho \max_{y_{i\xi} \in Y_i} \{ (w_{y_{i\xi}}^T - w_s^T) x_i \}}\right)^2} = 0$$

则当 $p \rightarrow \infty$ 时,

$$\frac{\partial^2 \ln L}{\partial w_{su} \partial w_{tv}} = \begin{cases} \sum_{i=1}^n \left(e^{(w_s^T + w_t^T) x_i} x_{iu} x_{iv} / \left(\sum_{j=1}^Q e^{w_j^T x_i} \right)^2 \right); s \neq t \\ \sum_{i=1}^n \left(e^{w_s^T x_i} \left(e^{w_s^T x_i} - \sum_{j=1}^Q e^{w_j^T x_i} \right) x_{iu} x_{iv} / \left(\sum_{j=1}^Q e^{w_j^T x_i} \right)^2 \right); s = t \end{cases}$$

则 $Z(W)$ 对 W 一阶和二阶导数可写成矩阵形式:

$$\begin{aligned} \nabla Z(W) &= \sum_{i=1}^n (\bar{Y}_i \otimes x_i - d_i \otimes x_i) - \rho \bar{E} W \\ \nabla \nabla Z(W) &= - \sum_{i=1}^n (\text{diag}\{d_i\} - d_i d_i^T) \otimes (x_i x_i^T) - \rho \bar{E} \end{aligned} \tag{4}$$

式中: $\otimes$ 表示克罗内克积 (Kronecker product)^[9], $\bar{Y}_i = \arg \max_{k \in Y_i} \{W^T(r \cdot y_k \otimes x_i)\}$, $d_i = (d_{i1} \ d_{i2} \ \cdots$

$d_{iQ})^T$, $d_{is} = e^{w^T(r \cdot y_s \otimes x_i)} / \sum_{j=1}^Q e^{w^T(r \cdot y_j \otimes x_i)}$ ($s = 1, 2, \cdots, Q$), $\text{diag}\{d_i\}$ 表示以 d_i 的元素为对角元的对角矩阵. 利用柯西不等式可以证明 $\text{diag}\{d_i\} - d_i d_i^T$ 是一个半正定矩阵, 则可知 $\nabla \nabla Z(W)$ 是一个半负定矩阵, 即 $Z(W)$ 是一个凹函数.

1.2 模型求解

本文应用阻尼牛顿法对模型进行求解, 阻尼牛顿法的迭代公式如下:

$$W^{k+1} = W^k - \lambda_k (\nabla \nabla Z(W^k))^{-1} \nabla Z(W^k)$$

其中步长 λ_k 通过一维优化问题 $\lambda_k = \arg \max_{\lambda} (Z(W^k - \lambda (\nabla \nabla Z(W^k))^{-1} \nabla Z(W^k)))$ 求得, 由于

$Z(W^k - \lambda (\nabla \nabla Z(W^k))^{-1} \nabla Z(W^k))$ 是一个凹函数, 可以通过利用二分法求解其导数的零点的方式来求解 λ_k , 二分法的结束条件是二分区间长度小于 10^{-5} , 算法流程图如图 1 所示.

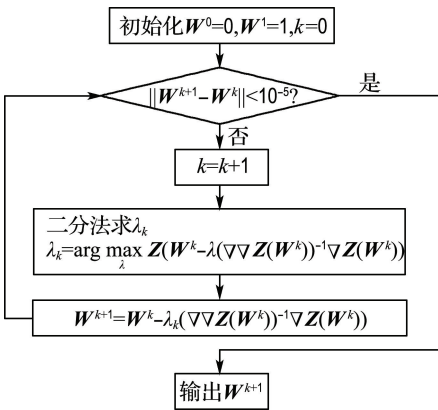


图 1 阻尼牛顿法求解 W

Fig. 1 Damped Newton method for solving W

2 数值实验

本文的实验数据包括 UCI^[10] 数据集中 5 个数据集和两个真实的候选标记集, 具体数据如表 1 所示. PLLOG^[11] 与本文算法 PLLOG-NB 比较结果如表 2~4 所示. 由于 UCI 的数据集不是候选标记集, 需要先生成候选标记集, 即先从原始的数据集中随机选取 pn 个样本, 然后随机地从样本的真实标记以外的类别标记中选取 r 个与真实标记一起构成该样本的候选标记. 其中 p 表示偏标记样本 (即 $|Y_i| > 1$) 在整个样本集中的比例, r 表示偏标记样本的除真实标记以外的候选标记个数, 即 $r = |Y_i| - 1$. 每个预测精度都是五次五折

表 1 算法验证所用的数据集

Tab. 1 The data sets for validating the algorithms

数据集	样本数	特征值数	类别数	最小类样本数	最大类样本数
Ecoli	336	7	8	2	143
CTG	2 126	21	10	53	384
Yeast	1 484	8	10	5	463
UCI	Image				
	Segmentation (Segment)	2 310	19	7	330
	Movement	360	90	15	24
真实	BirdSong	4 998	38	13	32
	MSRCv2	1 758	48	23	3

交叉实验的均值. A_c 是分类精度, $A_c = \sum_{j=1}^Q p_j / \sum_{j=1}^Q n_j$; $\overline{A_c}$ 是平均分类精度, $\overline{A_c} = \sum_{j=1}^Q \frac{p_j}{n_j} / Q$, 其中 p_j 是预测为第 j 类正确的数目. 所有结果都是在 CPU 主频 2.5 GHz, 内存 4 GB 的笔记本电脑上

运行得到的.

从表 1 可以看出, UCI 数据集前 3 个数据集和两个真实数据集都是不平衡数据集, 后两个是平衡数据集, 最大不平衡比例为 5/463. 由表 2~4 可以看出, 在平衡数据集上两个算法的预测精度

表 2 两个算法在 UCI 数据集上的预测精度

Tab. 2 The prediction accuracy of two algorithms on the UCI data sets

数据集	算法	$A_c / \%$								
		$r=1$			$r=2$			$r=3$		
		$p=0.15$	$p=0.45$	$p=0.75$	$p=0.15$	$p=0.45$	$p=0.75$	$p=0.15$	$p=0.45$	$p=0.75$
Ecoli	PLLOG-NB	83.93	83.93	83.04	83.93	83.04	82.74	84.23	82.74	82.44
	PLLOG	83.97	83.93	82.74	84.52	83.93	83.33	83.93	83.33	82.74
CTG	PLLOG-NB	67.59	69.19	70.08	68.77	68.72	69.71	69.00	69.85	69.80
	PLLOG	64.68	65.33	66.46	65.10	65.00	66.65	65.43	66.89	67.73
Yeast	PLLOG-NB	48.05	48.65	50.07	48.25	49.19	51.35	47.71	49.87	51.89
	PLLOG	49.06	50.67	52.36	48.38	50.74	50.20	47.04	48.65	50.94
Segment	PLLOG-NB	87.79	87.92	88.44	87.62	88.18	88.74	87.62	88.27	89.09
	PLLOG	87.84	87.92	88.44	87.71	88.10	88.70	87.66	88.18	89.09
Movement	PLLOG-NB	66.11	63.89	66.39	65.56	64.72	65.00	65.56	66.94	65.28
	PLLOG	66.11	64.17	66.39	65.83	64.72	65.28	65.28	66.94	65.83

表 3 两个算法在 UCI 数据集上的平均预测精度

Tab. 3 The mean prediction accuracy of two algorithms on the UCI data sets

数据集	算法	$\overline{A_c} / \%$								
		$r=1$			$r=2$			$r=3$		
		$p=0.15$	$p=0.45$	$p=0.75$	$p=0.15$	$p=0.45$	$p=0.75$	$p=0.15$	$p=0.45$	$p=0.75$
Ecoli	PLLOG	64.80	64.18	64.11	64.45	63.84	61.07	64.81	61.53	62.60
	PLLOG-NB	65.07	64.64	64.29	65.16	64.92	61.79	64.91	64.38	65.10
CTG	PLLOG	66.67	66.40	64.07	66.52	64.32	61.66	67.14	62.97	56.48
	PLLOG-NB	68.19	67.80	65.89	67.53	66.07	64.21	68.36	65.41	59.54
Yeast	PLLOG	53.20	51.68	52.83	52.75	51.09	49.90	50.67	52.19	46.41
	PLLOG-NB	53.59	52.42	53.83	52.75	51.64	49.82	50.82	52.48	49.67
Segment	PLLOG	87.79	87.92	88.44	87.62	88.18	88.74	87.62	88.27	89.09
	PLLOG-NB	87.84	87.92	88.44	87.71	88.10	88.70	87.66	88.18	89.09
Movement	PLLOG	66.11	63.89	66.36	65.56	64.72	65.00	65.56	66.94	65.28
	PLLOG-NB	66.11	64.17	66.39	65.83	64.72	65.28	65.28	66.94	65.83

表 4 两个算法在真实数据集上的预测精度

Tab. 4 The prediction accuracy of two algorithms on the real world data sets

数据集	算法	$A_c / \%$	$\overline{A_c} / \%$
BirdSong	PLLOG	58.30	40.97
	PLLOG-NB	58.20	41.37
MSRCv2	PLLOG	41.18	24.24
	PLLOG-NB	41.52	25.45

几乎相同, 这是因为模型建立的原理相同, 当 $\gamma=1$ 时, 模型计算结果几乎无差别. 本文算法在数据集 CTG 的 $\overline{A_c}$ 结果优势很明显, 在另外两个 UCI 的不平衡数据集上 $\overline{A_c}$ 也有一定的提高, 尤其当偏标记样本在整个样本集中的比例越大 ($p=0.75$), $\overline{A_c}$ 提高得越明显. 在几个不平衡数据集上的数据实验进一步证实了本文算法有效提高了样本的平均分类精度.

3 结 语

本文提出了可以处理数据不平衡问题的逻辑回归偏标记学习算法,在数据集上的实验结果验证了本文算法的有效性以及在处理不平衡问题方面的优势.下一步的工作是定义新的似然函数,应用更好的适合偏标记学习的机器学习算法,使其能够更好地处理数据不平衡偏标记学习问题.

参考文献:

- [1] GRANDVALET Y. Logistic regression for partial labels [C] // **Proceeding of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems**. Annecy: IPMU, 2002:1935-1941.
- [2] JIN R, GHAMRANI Z. Learning with multiple labels [C] // **Advances in Neural Information Processing Systems 15-Proceedings of the 2002 Conference, NIPS 2002**. Vancouver: Neural Information Processing Systems Foundation, 2003.
- [3] HÜELLERMEIER E, BERINGER J. Learning from ambiguously labeled examples [J]. **Intelligent Data Analysis**, 2006, **10**(5):419-439.
- [4] LUO J, ORABONA F. Learning from candidate labeling sets [C] // **Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010**. Red Hook: Curran Associates Inc., 2010: 1504-1512.

- [5] COUR T, SAPP B, TASKAR B. Learning from partial labels [J]. **Journal of Machine Learning Research**, 2011, **12**:1501-1536.
- [6] NGUYEN N, CARUANA R. Classification with partial labels [C] // **KDD 2008 - Proceedings of the 14th ACM KDD International Conference on Knowledge Discovery and Data Mining**. New York: Association for Computing Machinery, 2008: 551-559.
- [7] HE J, GU H, LIU W. Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites [J]. **PLoS One**, 2012, **7**(6):e37155.
- [8] LIU X Y, ZHOU Z H. **Imbalanced Learning: Foundations, Algorithms, and Applications** [M]. Hoboken: Wiley-IEEE Press, 2013:61-82.
- [9] HORN R, JOHNSON C. **Topics in Matrix Analysis** [M]. Cambridge: Cambridge University Press, 1991:239-297.
- [10] BACHE K, LICHMAN M. UCI machine learning repository [EB/OL]. (2013-04-04) [2016-08-12]. <http://archive.ics.uci.edu/ml>.
- [11] 周 瑜, 贺建军, 顾 宏, 等. 一种基于最大值损失函数的快速偏标记学习算法[J]. **计算机研究与发展**, 2016, **53**(5):1053-1062.
ZHOU Yu, HE Jianjun, GU Hong, *et al.* A fast partial label learning algorithm based on max-loss function [J]. **Journal of Computer Research and Development**, 2016, **53** (5): 1053-1062. (in Chinese)

Partial label learning algorithm for imbalanced data based on logistic regression

ZHOU Yu, GU Hong*

(Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Partial label learning is a new machine learning framework proposed in recent years, but existing partial label learning algorithms based on logistic regression have not solved the problem of data imbalance. A partial label learning algorithm for data imbalance is presented based on logistic regression model. The basic idea is to define a new likelihood function in the multiple logistic regression models to deal with imbalanced data. Firstly, a new likelihood function is defined according to the proportion of each class sample in the training set; then, the smooth and logistic regression-based partial label learning model is derived through derivation and approximation method. Simulation experiments on UCI data sets and real world data sets show that the proposed algorithm improves the average classification accuracy of sample for data imbalance problem.

Key words: partial label learning; data imbalance; logistic regression; damped Newton method