基于密度估计的逻辑回归模型

毛 毅 [1] 陈稳霖 [2] 郭宝龙 [1] 陈一昕 [2]

摘 要 介绍了一种基于密度的逻辑回归 (Density-based logistic regression, DLR) 分类模型以解决逻辑回归中非线性分类的问题. 其主要思想是根据 Nadarays-Watson 密度估计将训练数据映射到特定的特征空间, 然后组建优化模型优化特征权重以及 Nadarays-Watson 密度估计算法的宽度. 其主要优点在于: 它不仅优于标准的逻辑回归, 而且优于基于径向基函数 (Radial basis function, RBF) 内核的核逻辑回归 (Kernel logistic regression, KLR). 特别是与核逻辑回归分析和支持向量机 (Support vector machine, SVM) 相比, 该方法不仅达到更好的分类精度, 而且有更好的时间效率. 该方法的另一个显著优点是, 它可以很自然地扩展到数值类型和分类型混合的数据集中. 除此之外, 该方法和逻辑回归 (Logistic regression, LR) 一样, 有同样的模型可解释的优点, 这恰恰是其他如核逻辑回归分析和支持向量机所不具备的.

关键词 非线性分类, Nadarays-Watson 密度估计, 逻辑回归, 核函数

引用格式 毛毅, 陈稳霖, 郭宝龙, 陈一昕. 基于密度估计的逻辑回归模型. 自动化学报, 2014, **40**(1): 62−72

DOI 10.3724/SP.J.1004.2014.00062

A Novel Logistic Regression Model Based on Density Estimation

MAO Yi[1] CHEN Wen-Lin[2] GUO Bao-Long[1] CHEN Yi-Xin[2]

Abstract We propose a density-based logistic regression (DLR) model for classification to address the challenge of the nonlinear classification problem in this domain. Based on a Nadarays-Watson density estimator, the training data is mapped into a particular feature space. Then, an optimization model is set up to optimize the feature weights and the width in the Nadaraya-Watson density estimation algorithm. We show that it is superior to not only standard logistic regression but also kernel logistic regression (KLR) with radial basis function (RBF) kernels. The results show that DLR compares favorably against other nonlinear methods including KLR and support vector machine (SVM). The introduced approach achieves not only better classification accuracy but also better time efficiency. Another major advantage of our method is that it can be naturally extended to cope with hybrid data with both categorical features and numerical features. Moveover, our approach shares with logistic regression the same advantage of interpretability of the model, which is not obtained by kernel based methods such as KLR and SVM.

Key words Nonlinear classification, Nadarays-Watson density estimation, logistic regression (LR), kernel function

Citation Mao Yi, Chen Wen-Lin, Guo Bao-Long, Chen Yi-Xin. A novel logistic regression model based on density estimation. *Acta Automatica Sinica*, 2014, **40**(1): 62−72

作为数据挖掘、机器学习的关键部分, 分类在机器视觉[1−3]、药物开发[4]、语音识别、手写辨识等方面有着广泛的应用. 它是基于已知训练集识别一个新的实例属于哪个类别的有监督的学习问题. 在分类算法中, 有以下五个重要的衡量指标: 非线性分类的能力、数据类型的敏感度、多类分类的能力、效率、结果的可信任度. 几乎没有一种分类算法可以完美地处理所有这些问题. 目前存在很多分类模型, 如支持向量机 (Support vector machine, SVM)、逻辑回归 (Logistic regression, LR)、核逻辑回归分析 (Kernel logistic regression, KLR)、决策树 (Decision tree, DT)、朴素贝叶斯 (Naive Bayes, NB) 分类等, 但是它们大部分只适应于特定类型的数据或输出. 例如, SVM 是在处理数字和非线性可分数据方面非常有优势, 而 LR 则支持概率类型结果的输出.

支持向量机 (SVM)[5] 是一种经典的二值分类器, 通过解带约束条件的二次优化问题来建立数据集之间的最佳分界线. 与其他算法相比, 其重要优势在于: 通过使用不同的核函数, SVM 既可以用于线性分类, 也可以用于非线性分类. 但是由于其依赖于一对一模式, SVM 在多类分类上受到很大限制. 尽

收稿日期 2013-01-16 录用日期 2013-04-02
Manuscript received January 16, 2013; accepted April 2, 2013
国家自然科学基金 (61105066, 61201290, 61305041, 61305040) 资助
Supported by National Natural Science Foundation of China (61105066, 61201290, 61305041, 61305040)
本文责任编委 封举富
Recommended by Associate Editor FENG Ju-Fu
1. 西安电子科技大学机电工程学院智能控制与图像工程研究所 西安 710071, 中国 2. 圣路易华盛顿大学计算机工程学院 圣路易 63130, 美国
1. Institute of Intelligent Control and Image Engineering, School of Electromechanical Engineering, Xidian University, Xi′an 710071, China 2. Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis 63130, USA

管在将 SVM 扩展到多类分类上做了很多努力, 众所周知, 这些方法仍然有很多负面的影响[6−7]. 例如, 多类别分类中, SVM 一对多的决策方法就深受数据集类间不平衡的影响. 另一个重要的问题是它可能将同一实例分配给多个类. 虽然许多方法被提出来解决这些问题, 但是它们都有其他不利影响: 比如效率. SVM 的结果是纯粹二分的, 不支持概率输出. SVM 从一个任务的数值输出与另一个任务的数值输出不具有可比性. 此外, 与基于信任度的分类器相比, 这种没有限制的数值对于终端用户来讲很难解释其背后的意义.

逻辑回归 (LR) 是分类的重要方法之一. 标准逻辑回归使用输入变量的系数加权线性组合来分类. 相较于支持向量机, 自某一给定的类上, 标准逻辑回归能给出相应的类分布估计, 并且在模型训练时间上也占很大优势. 此外, 标准逻辑回归比支持向量机更容易扩展到多类分类. 美中不足的是, 标准逻辑回归是线性的分类器.

将核函数与标准逻辑回归相结合衍生出的核逻辑回归 (KLR)[8−9], 与 SVM 有着类似的损失函数. 通过使用总体精度矩阵, KLR 往往能实现与 SVM 类似, 甚至更好的分类性能, 并且支持概率类型结果输出. 但是 KLR 不能作为二次规划问题拟定, 需要解无约束的优化问题, 从而需要非线性的迭代优化算法, 如 Newton 和 Quasi-Newton 算法. 由于计算每次迭代的梯度代价高, 所以核逻辑回归的算法复杂度非常高.

SVM、LR、KLR 都可以处理数值类型的特征. 当处理数值和分类型混合类型数据时, 传统方法是将分类型数据转换为数值型数据. 对于一个有 m 个类型的分类数据, 一般需要 m 位的数值来表示[10]. m 位数值中只有一位为 1, 其余为 0. 但是, 这种预处理方式增加了特征的数量, 特别是当一个变量的类型多时, 很容易使算法陷入高维分类的尴尬境地.

朴素贝叶斯 (NB) 分类由先验概率 $p(x_i|y)$ 通过贝叶斯法则推出后验概率 $p(y|\boldsymbol{x})$. 当 x_i 为数值类型变量时, 朴素贝叶斯分类器被证明为线性分类器[11]. 此外, 朴素贝叶斯分类器是基于特征条件独立的假设. 一旦数据集不满足这一假设, 其分类准确度将会大受影响.

为了解决以上分类模型中的问题, 在本文中, 我们提出了一种基于密度的逻辑回归 (Density-based logistic regression, DLR) 模型. 接下来, 我们将详细论述所提出的新模型的理论合理性和实际应用的有效性. 在第 1 节中, 我们将介绍传统的 LR 算法. 在第 2 节中, 我们将讨论 DLR 在二元分类特征变量独立情况下中公式的推导表示. 然后, 在第 2.5 节中, 将模型与 Nadaraya-Watson 密度估计中参数 h 结合, 我们提出一种基于最大化训练集整体似然函数学习最优化参数的方法 DLRH.

1　逻辑回归 (LR) 算法

假设有数据集 $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}$, $i = 1, \cdots, N$, $\boldsymbol{x}_i \in \mathbf{R}^D$, $y_i \in \{0, 1\}$, 输入向量为 $\boldsymbol{x} = (x^{(1)}, \cdots, x^{(D)})$, 类标签 y 为二值函数: y 为 0 或 1. 逻辑回归 (LR) 为基于如下的概率模型:

$$p(y=1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^{\mathrm{T}}\phi) = \frac{1}{1+\exp(-\boldsymbol{w}^{\mathrm{T}}\phi)} = \frac{1}{1+\exp(-\sum\limits_{d=0}^{D} w_d\phi_d(\boldsymbol{x}))} \tag{1}$$

这里, $\boldsymbol{w}$ 是需要被学习的特征的权重值, 而 $\phi(\boldsymbol{x})$ 定义了特征向量的转换函数. $\phi_0 = 1$ 代表常数项. 本文提出了一种对于输入向量的新的变化方式 ϕ.

对于每一个 i, 定义 $b_i = p(y_i = 1|\boldsymbol{x}_i)$ 如式 (1). 逻辑回归最大化训练集的似然概率:

$$p(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{N} b_i^{y_i}(1-b_i)^{1-y_i} \tag{2}$$

$\boldsymbol{w} = (w_0, \cdots, w_D)^{\mathrm{T}}$ 是系数向量.

该问题等价于最小化负的似然概率, 或是以下形式的熵误差函数:

$$E(\boldsymbol{w}) = -\ln p(\boldsymbol{y}|\boldsymbol{w}) = -\sum_{i=1}^{N}\{y_i \ln b_i + (1-y_i)\ln(1-b_i)\} \tag{3}$$

求 $E(\boldsymbol{w})$ 关于 $\boldsymbol{w}$ 的导数, 得到:

$$\nabla E(\boldsymbol{w}) = \sum_{i=1}^{N}(b_i - y_i)\phi(\boldsymbol{x}_i) = \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{b} - \boldsymbol{y}) \tag{4}$$

建立 SVM 模型等价于最小化:

$$\frac{1}{N}\sum_{i=1}^{N}(1-y_i) \tag{5}$$

与 SVM 一样, 逻辑回归也可以与 “核” 结合, 产生核逻辑回归 (KLR)[8−9, 12]. 核 SVM 与 KLR 的时间复杂度均为 $\mathrm{O}(N^3)$. 核逻辑回归最小化下面的函数:

$$E(\boldsymbol{w}) = \sum_{i=1}^{N}\{y_i \ln b_i + (1-y_i)\ln(1-b_i)\} + \lambda\|\boldsymbol{w}\|^2 \tag{6}$$

2 基于密度的逻辑回归 (DLR) 模型

本节将描述针对逻辑回归 ϕ 的构造和参数自动调节的新方法. 这里, ϕ 是将 $\boldsymbol{x}$ 映射至特征空间的映射函数.

2.1 逻辑回归的局限性

假设输入向量是 $\boldsymbol{x} = (x^{(1)}, \cdots, x^{(D)})$, 对应的类标签 y 是二值的: $y \in \{0, 1\}$, 基于贝叶斯准则, 我们得到:

$$p(y=1|\boldsymbol{x}) = \frac{p(\boldsymbol{w}, y=1)}{p(\boldsymbol{w})} = \frac{p(\boldsymbol{x}, y=1)}{p(\boldsymbol{x}, y=1) + p(\boldsymbol{x}, y=0)} = \tag{7}$$

$$\frac{1}{1+\exp(-\tau(\boldsymbol{x}))} \tag{8}$$

其中,

$$\tau(\boldsymbol{x}) = \ln \frac{p(\boldsymbol{x}, y=1)}{p(\boldsymbol{x}, y=0)} \tag{9}$$

对比式 (1), (8) 和 (9), 能够看到在逻辑回归中的如下关系:

$$\sum_{d=0}^{D} w_d \phi_d(\boldsymbol{x})) = \tau(\boldsymbol{x}) \tag{10}$$

标准逻辑回归只简单地运用了 $\phi_d(\boldsymbol{x}) = x^{(d)}$, 从而得到:

1) 若 $w_d > 0$, 则 $p(y=1|\boldsymbol{x})$, 且随着 $x^{(d)}$ 的增长, $\tau(\boldsymbol{x})$ 增加;

2) 若 $w_d < 0$, 则 $p(y=1|\boldsymbol{x})$, 且随着 $x^{(d)}$ 的减少, $\tau(\boldsymbol{x})$ 增加.

因此, 采用简单映射函数 $\phi_d(\boldsymbol{x}) = x^{(d)}$ 的一个弊端是: 当 $p(y=1|\boldsymbol{x})$ 和 $x^{(d)}$ 为单调函数的时候, 逻辑回归的建模效果很好; 而当两者关系违背这一假设的时候, 所建模型的准确度会大打折扣.

2.2 DLR 模型和单变量映射

本节将详细描述我们提出的 DLR 模型. 为简化表示, 我们首先讨论当类标签 y 为二值的情况: $y \in \{0, 1\}$.

我们提出的 DLR 模型依然遵循 LR 在式 (1) 的参数模型. 但是, 它采用新的变化矩阵 ϕ, 我们的特征映射函数 ϕ_d 定义如下:

$$\phi_d(\boldsymbol{x}) = \ln \frac{p(y=1|x^{(d)})}{p(y=0|x^{(d)})} - \frac{D-1}{D} \ln \frac{p(y=1)}{p(y=0)} \tag{11}$$

右边第 1 项为考虑 $x^{(d)}$ 时 $y=1$ 的概率. 第 2 项 $\frac{D-1}{D} \ln \frac{p(y=1)}{p(y=0)}$ 为数据集类别不平衡的量度. 因此, 这里 $\phi_d(\boldsymbol{x})$ 可以认为是当 $x^{(d)}$ 时 y 为 1 的 "似然" 概率的评价量度. w_d 的大小评价了第 d 个特征变量对于 y 被标记为 1 的贡献度.

引理 1. 如图 1 所示, 对于给定 y, 如果所有变量 x 均条件独立时,

$$p(\boldsymbol{x}, y) = \prod_{d=1}^{D} \left(\frac{p(x^{(d)}, y)}{p(y)^{\frac{D-1}{D}}} \right) \tag{12}$$

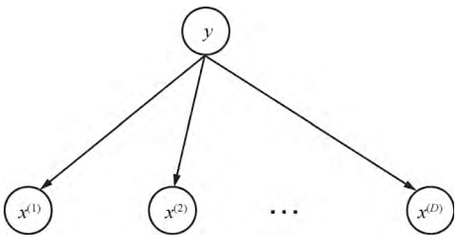

图 1 DLR 中的条件独立模型

Fig. 1 The independence model in DLR model

证明. 如果所有变量均条件独立时,

$$p(\boldsymbol{x}|y) = \prod_{d=1}^{D} p(x^{(d)}|y) \tag{13}$$

因此,

$$p(\boldsymbol{x}, y) = p(\boldsymbol{x}|y)p(y) = p(y) \prod_{d=1}^{D} p(x^{(d)}|y) = \frac{\prod_{d=1}^{D} p(x^{(d)}|y)p(y)}{p(y)^{D-1}} = \frac{\prod_{d=1}^{D} p(x^{(d)}, y)}{p(y)^{D-1}} = \prod_{d=1}^{D} \frac{p(x^{(d)}, y)}{p(y)^{\frac{D-1}{D}}} \tag{14}$$

□

定理 1. 对于给定 y, 如果所有变量 $\boldsymbol{x}$ 均条件独立时, 映射函数 ϕ (对于所有 d, $w_d = 1$) 将反映数据的真实分布.

证明. 根据式 (9) 和定理 1, 我们得到:

$$\tau(\boldsymbol{x}) = \ln \frac{p(\boldsymbol{x}, y=1)}{p(\boldsymbol{x}, y=0)} = \ln \frac{\prod_{d=1}^{D} \left(\frac{p(x^{(d)}, y=1)}{p(y=1)^{\frac{D-1}{D}}} \right)}{\prod_{d=1}^{D} \left(\frac{p(x^{(d)}, y=0)}{p(y=0)^{\frac{D-1}{D}}} \right)} =$$

$$\ln\frac{\prod\limits_{d=1}^{D}\left(\dfrac{p(y=1|x^{(d)})p(x^{(d)})}{p(y=1)^{\frac{D-1}{D}}}\right)}{\prod\limits_{d=1}^{D}\left(\dfrac{p(y=0|x^{(d)})p(x^{(d)})}{p(y=0)^{\frac{D-1}{D}}}\right)}=$$

$$\ln\frac{\prod\limits_{d=1}^{D}\left(\dfrac{p(y=1|x^{(d)})}{p(y=1)^{\frac{D-1}{D}}}\right)}{\prod\limits_{d=1}^{D}\left(\dfrac{p(y=0|x^{(d)})}{p(y=0)^{\frac{D-1}{D}}}\right)}=$$

$$\sum_{d=1}^{D}\left(\ln\frac{p(y=1|x^{(d)})}{p(y=0|x^{(d)})}-\frac{D-1}{D}\ln\frac{p(y=1)}{p(y=0)}\right)=$$

$$\sum_{d=1}^{D}\phi_d(\boldsymbol{x})$$

□

定理 1 表明, 在条件独立假设前提下, 当所有 $w_d = 1$ 时, DLR 可以完整地描述数据的真实分布. 同时, 我们也可以看到朴素贝叶斯和我们提出的 DLR 模型之间的联系. 朴素贝叶斯首先建立 $p(x^{(d)}|y)$ 模型, 然后根据贝叶斯准则得到 $p(y|\boldsymbol{x})$, 而我们提出的 DLR 直接建模得到 $p(y|x^{(d)})$, 通过式 (1) 即可得到 $p(y|\boldsymbol{x})$. 根据定理 1, 我们能够得到在条件独立假设前提下, 当所有 $w_d = 1$ 时, DLR 可以完整地描述数据的真实分布, 这与 NB 是吻合的. 虽然单变量的 DLR 模型有同样的条件独立的假设, 但是与 NB 不同的是, 这并不是 DLR 严格限制的. 当给定的数据集不能完全满足条件独立时, DLR 模型仍然能够通过最大化似然概率得到权重集 $\boldsymbol{w}$, 正如原始 LR 一样. 所不同的是, DLR 是非线性的, 而 LR 是线性的. DLR 包含了 LR 的处理不均衡的能力和 NB 衍生的能力, 而 NB 分类器可以看做是 DLR 的一个特例.

2.3 核密度估计函数 ϕ

在第 2.2 节中, 如式 (11) 所示, 我们提出了针对 DLR 的转换 ϕ. 下面将详细介绍如何估计式 (11) 中的概率. 给定训练集 $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}$, $i = 1, \cdots, N$, 我们需要对每一维 d 估计出 $\phi_d(\boldsymbol{x})$. 将 $\mathcal{D}$ 分为两个子集 $\mathcal{D}_1$ 和 $\mathcal{D}_0$, 分别包含 $y = 1$ 和 $y = 0$ 的数据点. $\phi_d(\boldsymbol{x})$ 的导数为 $p(x^{(d)}, y)$ 和 $p(y)$. $p(y)$ 可以由简单计算 y 在 $\mathcal{D}$ 中的比例得出.

$$p(y=k)=\frac{|\mathcal{D}_k|}{N},\quad k=0,1 \tag{15}$$

对于 $p(x^{(d)}, y)$, 分为离散、连续变量两种. 如果 $x^{(d)}$ 为分类型变量时, $\tilde{p}(y = k|x^{(d)})$ 仅仅为 $y = k$ 在第 d 个变量于 $x^{(d)}$ 训练集中的比值. 因此, ϕ 在分类型变量时可以通过简单计算得到.

定理 2. 当数据集所有变量 $x^{(d)}$ $(d = 1, 2, \cdots, D)$ 为分类型时, DLR $(w_d = 1)$ 与 NB 分类器有同样的决策边界.

证明. 给定实例 $\boldsymbol{x}$, NB 分类器采用如下准则给出类标签:

$$f(\boldsymbol{x})=\arg\max_y p(y)\prod_{d=1}^{D}p(x^{(d)}|y)=\arg\max_y p(\boldsymbol{x},y)$$

令 $\tau(\boldsymbol{x})=\ln\frac{p(\boldsymbol{x},y=1)}{p(\boldsymbol{x},y=0)}$, 我们能够得到: 如果 $\tau(\boldsymbol{x}) > 0$, 输出结果为 1, 反之为 0. 从定理 1, 我们得知当所有 $w_d = 1$ 时, DLR 模型对于 $\tau(\boldsymbol{x})$ 有同样的决策. 此外, 如定理 1 所示, NB 有同样的条件独立的假设. 对于分类型变量, DLR 模型和 NB 均通过计算每一维的发生频率来计算 $\tau(\boldsymbol{x})$. 因此, DLR $(w_d = 1)$ 与 NB 分类器有同样的决策边界. □

对于所有数值变量 $x^{(d)}$, 我们提出使用 Nadaraya-Watson 估计器来估计 $p(y = k|x^{(d)})$ $(k = 0, 1)$.

根据 Nadaraya-Watson 估计, 我们得到:

$$p(y=k|x^{(d)})=\frac{\sum\limits_{i\in\mathcal{D}_k}K\left(\dfrac{x^{(d)}-x_i^{(d)}}{h_d}\right)}{\sum\limits_{i=1}^{N}K\left(\dfrac{x^{(d)}-x_i^{(d)}}{h_d}\right)} \tag{16}$$

$K(x)$ 为核函数, 并且满足 $K(x) \geq 0$, $\int K(x)\mathrm{d}x = 1$. $h_d > 0$ 为和密度函数的宽度系数. 本文, 我们采用 Gaussian 核函数来构造 $K(x)$,

$$K(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) \tag{17}$$

将式 (16) 替换为式 (11), 得到第 k 维上的变换:

$$\phi_d(\boldsymbol{x})=\ln\frac{\sum\limits_{i\in\mathcal{D}_1}K\left(\dfrac{x^{(d)}-x_i^{(d)}}{h_d}\right)}{\sum\limits_{i\in\mathcal{D}_0}K\left(\dfrac{x^{(d)}-x_i^{(d)}}{h_d}\right)}-\frac{D-1}{D}\ln\frac{p(y=1)}{p(y=0)} \tag{18}$$

运用 Gaussian 核函数, 得到:

$$\phi_d(\boldsymbol{x}) = \ln \frac{\sum\limits_{i\in\mathcal{D}_1} \exp\left(-\dfrac{(x^{(d)}-x_i^{(d)})^2}{2h_d^2}\right)}{\sum\limits_{i\in\mathcal{D}_0} \exp\left(-\dfrac{(x^{(d)}-x_i^{(d)})^2}{2h_d^2}\right)} - \frac{D-1}{D}\ln\frac{p(y=1)}{p(y=0)} \tag{19}$$

$\phi_d(\boldsymbol{x})$ 为数据变量 $x^{(d)}$ 的转换. 本质上, 我们用式 (11) 给出了贝叶斯在 DLR 上的解释, 然后运用式 (11) 中基于核的方法来估计密度.

2.4 h_d 参数自调整

在 Nadaraya-Watson 估计中, h_d 作为自由参数, 仍然需要调节. 尽管有很多迭代算法来调节 h_d, 但是均与数据集本身无关. 本文提出一种基于最大化训练集整体似然函数学习最优化参数的方法, 其等价于优化式 (3) 中的 E. 为了解决这个优化问题, 我们需要计算 E 关于 h_d 的导数 $\frac{\partial E}{\partial h_d}$.

令

$$r_d = -\frac{1}{2h_d^2} \tag{20}$$

根据式 (19), 得到:

$$\phi_d(\boldsymbol{x}) = \ln \frac{\sum\limits_{i\in\mathcal{D}_1} \exp(r_d(x^{(d)}-x_i^{(d)})^2)}{\sum\limits_{i\in\mathcal{D}_0} \exp(r_k(x^{(d)}-x_i^{(d)})^2)} - \frac{D-1}{D}\ln\frac{p(y=1)}{p(y=0)} = g_1 - g_0 - S \tag{21}$$

这里

$$g_j = \ln \sum_{i\in\mathcal{D}_j} \exp(r_d(x^{(d)}-x_i^{(d)})^2) \tag{22}$$

$$S = \frac{D-1}{D}\ln\frac{p(y=1)}{p(y=0)} \tag{23}$$

因此, E 关于 h_d 的导数为

$$\frac{\partial \phi_d(\boldsymbol{x})}{\partial r_d} = \frac{\partial g_1}{\partial r_d} - \frac{\partial g_0}{\partial r_d} \tag{24}$$

对于 $j=0,1$,

$$\frac{\partial g_j}{\partial r_d} = \frac{\sum\limits_{i\in\mathcal{D}_j}[(x^{(d)}-x_i^{(d)})^2\cdot\exp(r_d(x^{(d)}-x_i^{(d)})^2)]}{\sum\limits_{i\in\mathcal{D}_j}\exp(r_d(x^{(d)}-x_i^{(d)})^2)} = \frac{\sum\limits_{i\in\mathcal{D}_j}[(x^{(d)}-x_i^{(d)})^2\cdot K(\frac{x^{(d)}-x_i^{(d)}}{h_d})]}{\sum\limits_{i\in\mathcal{D}_j}K(\frac{x^{(d)}-x_i^{(d)}}{h_d})} \tag{25}$$

下面计算 $\ln b_i$ 和 $\ln(1-b_i)$ 关于 r_d 的导数, 也就是 $\frac{\partial E}{\partial r_d}$ 的导数, 根据式 (1), 得到:

$$\frac{\partial \ln b_i}{\partial r_d} = \frac{\exp(-\sum\limits_{d=1}^{D} w_d\phi_d(\boldsymbol{x}_i))}{1+\exp(-\sum\limits_{d=1}^{D} w_d\phi_d(\boldsymbol{x}_i))} w_d\frac{\partial\phi_d(\boldsymbol{x}_i)}{\partial r_d} = (1-b_i)w_d\frac{\partial\phi_d(\boldsymbol{x}_i)}{\partial r_d} \tag{26}$$

以及,

$$\ln(1-b_i) = \ln\frac{\exp(-\sum_{d=1}^{D} w_d\phi_d(\boldsymbol{x}_i))}{1+\exp(-\sum_{d=1}^{D} w_d\phi_d(\boldsymbol{x}_i))} = -\sum_{d=1}^{D} w_d\phi_d(\boldsymbol{x}_i) + \ln b_i \tag{27}$$

$$\frac{\partial\ln(1-b_i)}{\partial r_d} = -w_d\frac{\partial\phi_d(\boldsymbol{x}_i)}{\partial r_d} + \frac{\partial\ln b_i}{\partial r_d} = -b_i w_d\frac{\partial\phi_d(\boldsymbol{x}_i)}{\partial r_d} \tag{28}$$

将式 (28) 和 (26) 代入式 (3), 得到如下表达式:

$$\frac{\partial E}{\partial r_d} = -\sum_{i=1}^{N}\left\{y_i\frac{\partial\ln b_i}{\partial r_d} + (1-y_i)\frac{\partial\ln(1-b_i)}{\partial r_d}\right\} = \sum_{i=1}^{N}(b_i-y_i)w_d\frac{\partial\phi_d(\boldsymbol{x}_i)}{\partial r_d} \tag{29}$$

从式 (20), 我们得到:

$$\frac{\partial E}{\partial h_d} = \frac{\partial E}{\partial r_d}\cdot\frac{\partial r_d}{\partial h_d} = \frac{1}{h_d^3}\sum_{i=1}^{N}(b_i-y_i)w_d\frac{\partial\phi_d(\boldsymbol{x}_i)}{\partial r_d} \tag{30}$$

最终, 将式 (24) 和 (25) 代入式 (30), 从而得到 $\frac{\partial E}{\partial h_d}$ 的完整表达式.

2.5 算法描述

令 $\boldsymbol{h} = (h_1, h_2, \cdots, h_D)$, 式 (30) 梯度函数 $\boldsymbol{h}$ 即可得到. 然而, $\boldsymbol{h}$ 的二阶偏导数却很难计算, 最终导致在 $\boldsymbol{w}$ 和 $\boldsymbol{h}$ 的联合空间中 Newton 算法很难实现. 这里, 我们提出一种学习包括 $\boldsymbol{w}$ 和 $\boldsymbol{h}$ 所有参数的分层优化框架, 如算法 1 所示. 整个优化框架包含两层. 与 SVM 超参数优化相似[13], 我们将整个数据

集分为三部分: 训练集、验证集和测试集. 首先, 在外层, 特征矩阵 ϕ 根据新的 $\boldsymbol{h}$ (第 4～8 行) 更新数据. 如果变量是分类型数据时, 相应的特征通过计算 $(x_i^{(d)}, y=1)$ 和 $(x_i^{(d)}, y=0)$ 频率得到. 对于数值型数据 $x^{(d)}$, 我们运用式 (19) 的核密度估计算法来计算相应的特征 $\phi_d(\boldsymbol{x})$ (第 9～11 行). 在内层循环中, 我们固定 $\boldsymbol{h}$, 根据训练集中的数据通过 Newton 算法来优化 $\boldsymbol{w}$. 最终, 根据式 (30) (第 12～16 行) 通过计算导数的下降方向来优化 $\boldsymbol{h}$.

算法 1. DLR 分层优化学习

1. Initialize $\boldsymbol{w}$
2. Initialize $\boldsymbol{h}$
3. **repeat** // Outer level
4. **for** $i=1$ to N **do** // Feature mapping
5. **for** $d=1$ to D **do**
6. $\boldsymbol{\Phi}_{i,d} \leftarrow \ln\frac{p(x_i^{(d)}, y=1)}{p(x_i^{(d)}, y=0)}$ // $\boldsymbol{\Phi}$ is feature matrix
7. **end for**
8. **end for**
9. **repeat** // Inner level. Fix $\boldsymbol{h}$ and optimize $\boldsymbol{w}$
10. $\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\nabla E}{\nabla^2 E}$ // Newton′s method
11. **until** $\boldsymbol{w}$ converges
12. **for** $d=1$ to D **do** // Fix $\boldsymbol{w}$ and tune each h_d
13. **if** $x^{(d)}$ is a numerical variable **then**
14. $h_d \leftarrow h_d - \gamma\frac{\partial E}{\partial h_d}$ // Gradient descent.
15. **end if**
16. **end for**
17. **until** $\boldsymbol{h}$ converges

通过算法 1, 我们能够推导出 DLR 中的两个重要变量. 我们将整个完整的算法称为 DLRH, 因为算法本身自动选择 $\boldsymbol{h}$. 如果我们在初始值时固定 $\boldsymbol{h}$, 去掉第 12～16 行的优化步骤, 则称此算法为 DLR. 因为它与传统 LR 算法一样, 只是在特征映射的时候才用了不同的算法. 在实验部分, 我们将对 DLR 与 DLRH 分别进行评估.

3　讨论: 一种新的概率鉴别核

本节我们将从核方法的角度来讨论 DLR 模型, 从而说明 DLR 不仅是一个正定的核算法, 更是一种鉴别的核算法. 我们将通过与现有的核比较来阐述这一点. 众所周知, LR 和 SVM 模型可以在一个统一的视角下研究. 它们都可以理解为优化下面这个函数:

$$\min_{\boldsymbol{w}} E = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i g(-y_i\cdot\boldsymbol{w}\boldsymbol{x}_i) \tag{31}$$

LR 模型采用 LOG 损失函数

$$g(\eta) = \log(1+\mathrm{e}^{\eta}) \tag{32}$$

SVM 模型利用 hinge 损失函数

$$g(\eta) = \max(1-\eta, 0) \tag{33}$$

由 Representer 定理, 两个模型都是利用核函数来获得非线性分离. 从这个角度看, 提出的 DLR 模型可以看作是采用 LR 模型中 LOG 损失函数, 但是 DLR 采用了如下的核模型:

$$k_{\mathrm{DLR}}(\boldsymbol{x},\boldsymbol{x}') = \phi(\boldsymbol{x})\phi(\boldsymbol{x}') = \sum_{d=1}^{D}\phi_d(\boldsymbol{x})\phi_d(\boldsymbol{x}') = \sum_{d=1}^{D}k_d(\boldsymbol{x},\boldsymbol{x}') \tag{34}$$

这里, ϕ_d 采用式 (11) 中的变换, 并且 $k_d(\boldsymbol{x},\boldsymbol{x}') = \phi_d(\boldsymbol{x})\phi_d(\boldsymbol{x}')$. 在某种程度上, 一个内核定义了两个数据实例之间的相似性. 例如, 高斯内核解释基于距离的相似性. 根据式 (11) 和第 2.2 节中的讨论, 每个 ϕ_d 代表当给定 $x^{(d)}$ 时, y 和标签 1 之间的似然概率. 特别是, $k_d(\boldsymbol{x},\boldsymbol{x}')$ 代表两个实例第 d 个变量之间的相似度, 而整体的相似度有每一维的相似度的整合来定义. 我们能够从下面的例子中得到上述事实.

1) 如果 $x^{(d)}$ 和 $x^{(d)\prime}$ 均有很高的概率被标记为 1, 那么 $\phi_d(\boldsymbol{x})$ 和 $\phi_d(\boldsymbol{x}')$ 均为正, 而且它们的乘积 (即 $k_d(\boldsymbol{x},\boldsymbol{x}')$) 应较大;

2) 如果 $x^{(d)}$ 和 $x^{(d)\prime}$ 均有很高的概率被标记为 0, 那么 $\phi_d(\boldsymbol{x})$ 和 $\phi_d(\boldsymbol{x}')$ 均为负, 而且它们的乘积应较大;

3) 如果 $x^{(d)}$ 和 $x^{(d)\prime}$ 有很高的概率被标记为不同的类别, 那么对应的 ϕ_d 应为一正一负, 而且它们的乘积应为负.

显然, DLR 在式 (34) 中的核是一个正定的核, 因为它在特征空间的内积核矩阵总是正定的. 如文献 [14] 所述, 对于一个良好的核函数, 我们需要 $k(\boldsymbol{x},\boldsymbol{x}')$ 满足条件当 $y=y'$ 时, $k(\boldsymbol{x},\boldsymbol{x}')>0$ 和当 $y\neq y'$ 时, $k(\boldsymbol{x},\boldsymbol{x}')<0$. 当一个核函数严格满足以上条件时, 数据集中的实例将有可能被完全分类正确. 通过上面的分析, 我们能够看到, 我们提出的 DLR 非常接近这个条件, 所以说我们的核函数不仅是一个正定的核, 而且是一个良好的核函数. 与现存的核函数, 比如 Polynomial 核、Gaussian 核相比, 我们提出的核函数通过鉴别性的条件概率考虑到了类别中隐藏的信息, 从而使得相似性的衡量更加准确. 而这一点在其他的核函数, 比如 Gaussian 函数中均没有考虑到.

3.1　时间复杂度分析

核 SVM 与 KLR 的时间复杂度均为 $\mathrm{O}(N^3)$. 其中 N 为数据集的大小. 逻辑回归的时间复杂度与其使用的具体优化算法有关, 一般来讲, 其时间复杂度为 $\mathrm{O}(M)$[15], 其中 M 为优化矩阵中非零元素的

个数. 在贝叶斯算法中, 模型训练的时间复杂度为 O(N). 本文提出的算法 DLRH 复杂度主要由三部分决定: 核密度转换的时间、逻辑回归的时间和训练 H 的时间. 核密度转换的时间与数据的特征量 D 成正比, 训练 H 的时间与算法的收敛度有关 (算法中每次都采用上一次优化得到的参数作为下一步的初始点, 大大加快了收敛的速度), 所以本文的时间复杂度为 O(M) + O(D). 而一般情况下 $D \ll M$, 所以忽略非主要项的时间复杂度为 O(M), 远远低于核 SVM 与 KLR 的时间复杂度.

4 相关工作

逻辑回归 (LR) 仍然是数据挖掘、数据分类中应用最常用的方法之一[15]. 在语义网服务匹配[16]、流行病研究[17]、预测地质灾害[18−19] 等方面都有着广泛的应用. Das 等[20] 提出将协变量与逻辑回归结合来解决逻辑回归在有缺失的数据集中预测偏差的问题. 多元逻辑回归 (Multinomial logistic regression, MLR) 模型作为一种推广的具有两个以上离散结果的逻辑回归模型[21], 在实际建模中也得到了广泛的应用. 文献 [22] 利用正规划逻辑回归对脑磁图 (Magnetoencephalography, MEG) 数据进行特征选择.

逻辑回归还可以从贝叶斯的角度来进行研究, 尽管确切的贝叶斯推理是不可行的, 或仍然需要一些近似, 如拉普拉斯近似[23]. KLR 中的对偶算法与 SVM 中的串行最小优化 (Sequential minimal optimization, SMO) 算法[12] 的思路基本相似. 在 KLR 损失函数中引入导入向量机 (Import vector machines, IVM), 使其具有了 SVM 的框架[9]. 导入向量 (IVM) 具有逻辑回归很多有利的特性, 包括估计隐藏概率和推广至多类问题的能力. 与其他复杂的算法相比, 如贝叶斯逻辑回归和 SVM, 我们的算法仅仅需要一个简单的预处理, 将输入向量变换为特征, 并且如逻辑回归一样只需要解一个无约束优化问题. 它仍然继承着逻辑回归相对 SVM 的优势.

Raina 等提出一种用于分类的混合生成/鉴别模型[24]. 同样地, 他们将每个变量 x_k 变换为特征 $\phi_k = \ln(\hat{p}(x, y = 1)) - \ln(\hat{p}(x, y = 0))$. 但是, 它只考虑到 x 取离散值的情况. 并且, $\hat{p}(x, y = 1)$ 和 $\hat{p}(x, y = 0)$ 概率是从贝叶斯模型学习得到, 需要大量采样. 这对于连续型 x 并不适用. 与此不同, 我们利用一个平滑的核密度估计算法, 从而使得对于连续 x 仅需少量的采样点. 与此同时, 我们的核宽度 h_k 可以自动优化生成.

在已有文献中, 还有其他相关的结合生成和鉴别的概率核. Jaakkola 等提出了在有识别力的分类模型中使用生成技术[25]. 他们提出使用 Fisher 分数从生成模型来生成一个内核, 从而可以用于区别模型, 例如 SVM 和 LR. 但是, 这种 “生成” 核仅仅适合于一些特定域的任务, 比如分类结构化数据, 并且对不同的任务要求不同的生成模型. 例如, 顺序数据如蛋白质和 DNA 的 Fisher 核通常由隐马尔可夫模型 (Hidden Markov model, HMM) 生成[26], 而文本分割的核通常来源于线性判别分析 (Linear discriminant analysis, LDA)[27]. 另外一种更普遍的概率核是文献 [28] 中描述的用来分类标记图像的边缘内核. 然而, 它仍然需要一个特定域的生成模型. 文献 [29] 引入鉴别核, 这里映射函数也是基于 $\log(p(y|\boldsymbol{x}, \theta))$. 并且, 它仍然依赖于一个特定的生成模型, 如 HMM, 因为 θ 取决于 HMM. 与其他基于模型的内核相比, 我们的 DLR 内核不需要对数据集假设任何底层模型, 而直接使用 Nadaraya-Watson 密度估计量在每一维中来近似 $p(y|\boldsymbol{x})$, 从而使得我们的算法不受数据集的限制, 有着更广泛的应用空间.

5 实验结果

本节我们将对本文提出的基于密度的逻辑回归 (DLR) 和自调节 “$\boldsymbol{h}$” 的 DLR (DLRH) 模型进行评估. 当比较本文算法与其他算法的有效性时, 在不同的数据集上, 我们采用其他 5 个主要的分类方法作为我们的比较对象. 实验结果主要集中在以下三个问题: 分类精度、执行速度和不同的基准数据集的鲁棒性. 用于比较的数据分类算法包括逻辑回归、支持向量机和内核逻辑回归. Polynomial 和径向基函数 (Radial basis function, RBF) 内核均被用于支持向量机和内核逻辑回归分类算法中来评估算法.

5.1 归一化

假设 $\phi(x)_{\min}$ 和 $\phi(x)_{\max}$ 分别为数据集中的最大值和最小值. 对于一个实例, 归一化算法如下:

$$\phi(x)_{\text{nor}} = \frac{\phi(x) - \phi(x)_{\min}}{\phi(x)_{\max} - \phi(x)_{\min}}$$

通过这种方式, 式 (11) 中的数据贡献不均衡将被取出, 所以对每一个实例做如下的归一化:

$$\phi(x)_{\text{nor}} = \frac{\phi(x)}{\|\phi(x)\|}$$

5.2 简单构造数据集上的实验结果

首先我们在如图 2 和图 3 所示的简单一维和二维数据上比较实验结果. 一维数据在 $[0, 1]$; $[10, 12]$; $[20, 21]$ 三个区间随机产生. 每个区间随机产生 100 个点, 其中第一个区间和第三个区间的数据有相同的类标签, 共计 300 个点, 两个大类. 二维数据从均

值为 [100], [−100], [0, 10] 和 [0, −10] 这 4 个多变量标准分布中产生, 共计 1 200 个点. 前两个分布的协方差为 $\sigma = [1, 0; 0, 100]$, 后两个分布的协方差为 $\sigma = [100, 0; 0, 1]$.

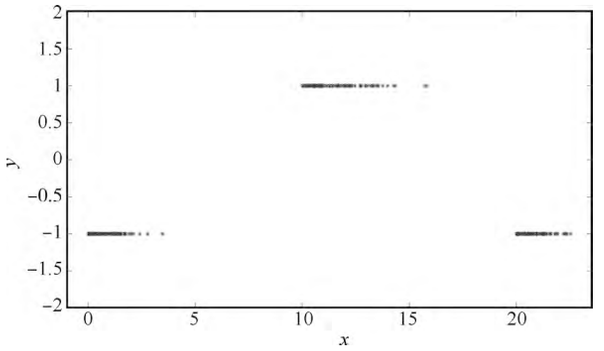

图 2　一维数据集分布

Fig. 2　1-D dataset

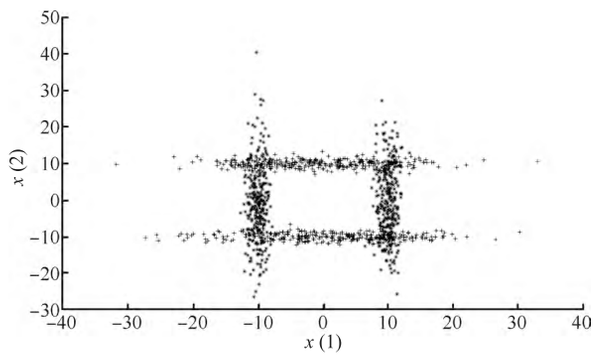

图 3　二维数据集分布

Fig. 3　2-D dataset

从表 1 的分类准确度比较结果中, 我们能够看到, 与其他算法相比, DLR 方法有相同甚至更好的分类精度. 特别当数据集为非线性分布时, DLR 与传统的逻辑回归算法相比, 分类精度大大提高. 相比 RBF 核的 KLR, SVM 分类算法有着更好的分类精度. 同时, DLRH 模型将精度从 0.865 进一步提升至 0.893, 从而证实了调节 $\boldsymbol{h}$ 的有效性. 除此之外, 从表 2 可以看出, 与 RBF 核的 KLR、SVM 分类算法相比, DLR 和 DLRH 的算法运行时间也大大缩减. 因此, 虽然我们的算法与线性逻辑回归有着相同的算法复杂度, 但是比很多先进的非线性分类算法有着更好的分类准确度和更短的分类时间.

5.3　UCI 数据集上的实验结果

为了进一步检验算法的鲁棒性, 我们将 DLR 和 DLRH 用于 UCI 数据集中[30]. 表 3 列出了这些数据集的主要特征, 包含了实例数、特征数量、正实例的个数. 第二个数据集有 167 个缺失数据, 可以用来进一步检验算法的鲁棒性. 缺失的数据用原始数据集中的特征期望值来替代. 表 4 和表 5 比较了 DLR、DLRH 与其他算法的分类精确度和 AUC. 表 6 为各个算法在 UCI 数据上的运行时间比较. SVM、KLR 非线性分类器往往比线性分类器, 如 LR、NB 分类效果更好. 并且, 我们能够看到与其他算法相比, DLR、DLRH 有着更好的分类精度, 同时和 LR 有着相近的执行时间. 如表 4 所示, DLR 对于不完整数据集有着更好的鲁棒性. KLR、SVM 对于核的选择有着相当的敏感性. 在大多数情况下, RBF 核比 Polynomial 核有着更好的分类精度, 但是对于 Hepatitis domain 数据集, Polynomial 核却有着更好的表现. 由于数据分布的差异性, 很难判断哪种核函数更合适, 除非将所有核函数一一试过. 而我们新提出的逻辑回归算法的优势在于: 我们对于数据集不假设核, 取而代之, 我们用核密度估计来发现数据集背后的分布.

5.4　分类型数据集上的实验结果

分类数据是一种统计的数据类型组成的分类变量, 用于观察到的数据, 它的值是固定数目的标称类别之一, 或已被转换成这种形式的数据, 例如, 分组数据. 当处理分类型 (Categorical data) 混合类型数据时, 传统方法是将分类型数据转换为数值型数据. 对于一个有 m 个类型的分类数据, 一般需要 m 位的数值来表示[10]. m 位数值中只有 1 位为 1, 其余为 0, 从而使得同一类之间的距离为 0, 不同类的距离为 1. 在我们提出的 DLR 算法中, 我们可以轻易跳过数据转换, 不用将特征空间扩展为 $m \times num$. 我们在 UCI 的分型数据上测试了我们的算法. 从表 7 和表 8 可以看到, 与其他基于核的分类器相比, 在保持高分类精度的基础上, DLR 可以大幅度降低训练时间.

6　结论

作为模式识别、数据挖掘的重要分支, 分类越来越广泛的应用领域, 使之成为医疗、航空电子侦察等系统的核心和关键技术. 本文提出了一种基于密度的逻辑回归模型. 该方法基于逻辑回归的二元基本模型, 构建了一种非线性的新型分类器. 实验结果表明其不仅在分类精度、分类鲁棒性上比传统分类方法有优势, 而且在训练时间上也明显优于其他算法. 除此之外, 该算法可以很自然地扩展到数值类型和分类型混合的数据集中. 并且其训练得到的模型比 SVM、贝叶斯等方法具有更强的解释性. 但是, 目前算法在算法优化时间上仍然有很大的优化空间, 在以后的工作中将尝试探寻压缩运行时间的方法. 和现有的工作[31−32] 相结合, 我们也将进一步研究新密度模型在流数据中的应用.

表 1 算法精度比较

Table 1 Accuracy of various methods

	Accuracy (%)							
Data set	LR	NB	KLR (Polynomial)	KLR (RBF)	SVM (Polynomial)	SVM (RBF)	DLR	DLRH
1-D data set	63.33	67.02	96.67	96.67	96.67	96.67	96.67	96.67
2-D data set	50.3	54.0	52.5	86.3	52.5	83.5	86.5	89.3

表 2 运行时间比较

Table 2 Running time of various methods

	Time (ms)							
Data set	LR	NB	KLR (Polynomial)	KLR (RBF)	SVM (Polynomial)	SVM (RBF)	DLR	DLRH
1-D data set	46.8	54.6	78.5	140.4	889.2	390.0	62.4	62.4
2-D data set	31.2	124.8	187.2	3 868.8	140.4	2 808.0	109.2	249.2

表 3 UCI 数据集

Table 3 UCI dataset

Data set	#Instances	#Attributes	#Positive
Breast cancer wisconsin	699	9	239
Hepatitis domain	155	19	32
Ionosphere	351	33	126
Cleveland heart disease	297	13	137
Pima indians diabetes	768	8	268

表 4 UCI 数据集上算法精度比较

Table 4 Accuracy of various methods on UCI dataset

	Accuracy (%)							
Data set	LR	NB	KLR (Polynomial)	KLR (RBF)	SVM (Polynomial)	SVM (RBF)	DLR	DLRH
1	93.9	94.3	93.0	96.5	93.9	96.2	96.5	**96.5**
2	85.2	82.4	80.4	84.3	84.6	84.3	86.2	**88.2**
3	85.1	77.8	82.8	93.1	84.6	94.3	93.1	**93.1**
4	84.5	81.6	79.7	83.8	75.0	77.7	85.1	**85.1**
5	74.7	74.4	74.0	75.0	70.8	71.1	75.5	**77.8**

表 5 UCI 数据集上算法 AUC 比较

Table 5 AUC of various methods on UCI dataset

	AUC							
Data set	LR	NB	KLR (Polynomial)	KLR (RBF)	SVM (Polynomial)	SVM (RBF)	DLR	DLRH
1	0.9923	0.9690	0.9534	0.9835	0.9890	0.9860	0.9960	**0.9960**
2	0.8004	0.8040	0.5000	0.5207	0.8049	0.8512	0.8732	**0.8781**
3	0.8711	0.7980	0.9264	0.9635	0.7994	0.9610	0.9695	**0.9695**
4	0.8060	0.8040	0.5000	0.6053	0.8217	0.8168	0.8128	**0.8134**
5	0.7602	0.7642	0.5000	0.6322	0.5949	0.7740	0.7665	**0.7873**

表 6 UCI 数据集上算法运行时间比较

Table 6 Running time of various methods on UCI dataset

	Time (ms)							
Data set	LR	NB	KLR (Polynomial)	KLR (RBF)	SVM (Polynomial)	SVM (RBF)	DLR	DLRH
1	46.8	249.6	1 700.4	312.2	2 620.8	1 716.1	296.4	702.0
2	62.4	156.0	842.4	124.8	93.6	93.6	78.5	249.5
3	62.4	140.4	1 154.4	1 201.2	421.2	3 822.0	124.8	546.0
4	31.2	124.8	1 060.8	1 154.8	499.2	93.6	124.8	604.2
5	62.4	140.4	3 042.0	3 556.8	22 245.7	5 772.0	452.4	764.4

表 7 分型数据集上算法精度比较

Table 7 Accuracy of various methods on categorical dataset

	Accuracy (%)							
Data set	LR	NB	KLR (Polynomial)	KLR (RBF)	SVM (Polynomial)	SVM (RBF)	DLR	DLRH
SPECFT heart data	57.8	65.4	65.6	74.5	63.3	76.7	77.8	**87.6**
Tic-tac	95.9	96.0	98.1	97.5	98.1	91.5	**98.1**	**98.1**
Monk problem	69.8	67.1	97.3	97.3	97.3	96.2	**97.3**	**97.3**

表 8 分型数据集上算法训练时间比较

Table 8 Training time of various methods on categorical dataset

	Time (ms)							
Data set	LR	NB	KLR (Polynomial)	KLR (RBF)	SVM (Polynomial)	SVM (RBF)	DLR	DLRH
SPECFT heart data	187.2	280.4	530.4	171.6	1 716.0	280.8	**46.8**	62.4
Tic-tac	218.4	256.8	8 907.6	1 872.0	11 548.7	764.4	46.8	**46.8**
Monk problem	187.2	124.8	1 388.4	795.6	12 230.4	390.0	**31.3**	**31.3**

References

1 Haralick R M, Shapiro L G. *Computer and Robot Vision*. Boston, MA: Addison-Wesley Longman Publishing Co., Inc. 1992

2 Jiang Li-Xing, Hou Jin. Image annotation using the ensemble learning. *Acta Automatica Sinica*, 2012, **38**(8): 1257−1262

3 Chen Rong, Cao Yong-Feng, Sun Hong. Multi-class image classification with active learning and semi-supervised learning. *Acta Automatica Sinica*, 2011, **37**(8): 954−962
(陈荣, 曹永锋, 孙洪. 基于主动学习和半监督学习的多类图像分类. 自动化学报, 2011, **37**(8): 954−962)

4 Ting N. *Dose Finding in Drug Development (Statistics for Biology and Health)*. New York: Springer, 2006

5 Wang Yao-Nan, Yuan Xiao-Fang. SVM approximate-based internal model control strategy. *Acta Automatica Sinica*, 2008, **34**(2): 172−179

6 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006

7 Zhao Zhi-Gang, Lv Hui-Xian, Li Yu-Jing, Li Jing. A multi-classification SVM based on clustering idea. *Journal of Qingdao Technological University*, 2011, **32**(1): 73−76
(赵志刚, 吕慧显, 李玉景, 李京. 一种基于聚类思想的 SVM 多类分类方法. 青岛理工大学学报, 2011, **32**(1): 73−76)

8 Green P J, Yandell B S. Semi-parametric generalized linear models. In: Proceedings on the 2nd International GLIM Conference. New York: Springer-Verlag, 1985. 44−55

9 Zhu J, Hastie T. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, Cambridge, MA: MIT Press, 2001. 1081−1088

10 Hsu C W, Chang C C, Lin C J. TA Practical Guide to Support Vector Classification, Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, China, 2003

11 Mitchell T M. *Machine Learning*. New York: McGraw-Hill Inc., 1997

12 Keerthi S S, Duan B K, Shevade S K, Poo A N. A fast dual algorithm for kernel logistic regression. *Machine Learning*, 2005, **61**(1−3): 151−165

13 Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing multiple parameters for support vector machines. *Machine Learning*, 2002, **46**(1−3): 131−159

14 Gärtner T. A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, 2003, **5**(1): 49−58

15 Maalouf M. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, 2011, **3**(3): 281−299

16 Wei Deng-Ping, Wang Ting, Wang Ji. A logistic regression model for semantic web service matchmaking. *Science China Information Sciences*, 2012, **55**(7): 1715−1720

17 Zhang Z, Liu A, Lyles R H, Mukherjee B. Logistic regression analysis of biomarker data subject to pooling and dichotomization. *Statistics in Medicine*, 2012, **31**(22): 2473−2484

18 Junek W N, Jones L W, Woods M T. Use of logistic regression for forecasting short-term volcanic activity. *Algorithms*, 2012, **5**(4): 330−363

19 Dong J J, Tung Y H, Chen C C, Liao J J, Pan Y W. Logistic regression model for predicting the failure probability of a landslide dam. *Engineering Geology*, 2011, **117**(1−2): 52−61

20 Das U, Maiti T, Pradhan V. Bias correction in logistic regression with missing categorical covariates. *Journal of Statistical Planning and Inference*, 2010, **140**(9): 2478−2485

21 Bham G H, Javvadi B S, Manepalli U R R. Multinomial logistic regression model for single-vehicle and multivehicle collisions on urban US highways in Arkansas. *Journal of Transportation Engineering*, 2012, **138**(6): 786−797

22 Santana R, Bielza C, Larrañaga P. Regularized logistic regression and multiobjective variable selection for classifying MEG data. *Biological Cybernetics*, 2012, **106**(6−7): 389−405

23 Jaakkola T S, Haussler D. Probabilistic kernel regression models. In: Proceedings of the 1999 Conference on AI and Statistics. Key West, FL: Morgan Kaufmann, 1999. 1−9

24 Raina R, Shen Y R, Ng A Y, McCallum A. Classification with hybrid generative/discriminative models. In: Proceedings of the 2003 Advances in Neural Information Processing Systems. MIT Press, 2003. 280−289

25 Jaakkola T, Haussler D. Exploiting generative models in discriminative classifiers. In: Proceedings of the 1998 Advances in Neural Information Processing Systems 11. MIT Press, 1998. 487−493

26 Jaakkola T, Diekhans M, Haussler D. Using the Fisher kernel method to detect remote protein homologies. In: Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology. Heidelberg: AAAI Press, 1999. 149−158

27 Sun Q, Li R X, Luo D S, Wu X H. Text segmentation with LDA-based Fisher kernel. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. 269−272

28 Kashima H, Tsuda K, Inokuchi A. Marginalized kernels between labeled graphs. In: Proceedings of the 20th International Conference on Machine Learning. Heidelberg: AAAI Press, 2003. 321−328

29 Tsuda K, Kawanabe M, Rätsch G, Sonnenburg S, Müller K R. A new discriminative kernel from probabilistic models. *Neural Computation*, 2002, **14**(10) 2397−2414

30 Frank A, Asuncion A. UCI Machine Learning Repository [Online], available: http://www.ics.uci.edu/∼mlearn/{MLR}epository.htm, August 28, 2012

31 Tu L, Chen Y X. Stream data clustering based on grid density and attraction. *ACM Transactions on Knowledge Discovery from Data*, 2009, **3**(3): 12:1−12:27

32 Chen Y X, Tu L. Density-based clustering for real-time stream data. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-07). New York, USA: ACM, 2007. 133−142

毛 毅 西安电子科技大学智能控制与图像工程研究所博士研究生. 2008 年获西安电子科技大学测控计量技术与仪器学士学位. 主要研究方向为数据挖掘与机器学习. 本文通信作者.
E-mail: olivia.maoy@gmail.com
(**MAO Yi** Ph. D. candidate at the Institute of Intelligent Control and Image Engineering, Xidian University. She received her bachelor degree from Xidian University in 2008. Her research interest covers machine learning and data mining. Corresponding author of this paper.)

陈稳霖 圣路易斯华盛顿大学计算机科学与工程系博士研究生. 2011 年获中国科学技术大学计算机系学士学位. 主要研究方向为数据挖掘, 机器学习, 优化.
E-mail: wenlinchen@wustl.edu
(**CHEN Wen-Lin** Ph. D. candidate at the Department of Computer Science and Engineering, Washington University in St. Louis, USA. He received his bachelor degree from University of Science and Technology of China, China in 2011. His research interest covers data mining, machine learning and optimization.)

郭宝龙 教授. 分别于 1988 年, 1995 获西安电子科技大学硕士、博士学位. 主要研究方向为模式识别与智能系统, 图像处理和图像通信.
E-mail: blguo1199@126.com
(**GUO Bao-Long** Professor. He received his master and Ph. D. degrees from Xidian University in 1988 and 1995, respectively. His research interest covers pattern recognition and intelligent systems, image processing, and image communication.)

陈一昕 圣路易斯华盛顿大学计算机科学与工程系副教授. 2005 年获伊利诺伊香槟分校计算机系博士学位. 主要研究方向为数据挖掘, 规划, 优化, 博弈论.
E-mail: chen@cse.wustl.edu
(**CHEN Yi-Xin** Associate processor in the Department of Computer Science and Engineering, Washington University in St. Louis, USA. He received his Ph. D. degree in computer science from the University of Illinois at Urbana-Champaign, USA in 2005. His research interest covers machine learning, planing, optimization, and game theory.)