文章编号: 1001-0920 (2014) 06-1103-06 DOI: 10.13195/j.kzyjc.2013.0318

基于稀疏流形聚类嵌入模型和 L_1 范数正则化的标签错误检测

夏建明, 杨俊安

(合肥电子工程学院 a. 通信对抗系，b. 安徽省电子制约技术重点实验室，合肥 230037)

摘 要: 综合利用含错标签中的有用信息和数据结构中蕴含的鉴别信息, 提出一种基于稀疏流形聚类嵌入模型和 L_1 范数正则化的标签错误检测修正方法. 首先, 用稀疏流形聚类嵌入模型将数据投影到易分类的空间, 利用标注正确的极少量样本和最近邻分类器获得新标签; 然后, 构造标签错误检测模型, 获得仅含 0、1 元素的检测向量, 正确、错误的标签分别对应着 1、0 的位置; 最后, 给出了相应的优化算法及收敛证明, 并在相关实验上验证了算法的有效性.

关键词: 标签错误；稀疏流形聚类嵌入；L_1 范数正则化；凸松弛

中图分类号: TP181 **文献标志码:** A

Labeling errors detecting and correcting algorithm based on sparse manifold clustering and embedding and L_1 norm regularization

XIA Jian-ming, *YANG Jun-an*

(a. Department of Communication Countermeasure，b. Key Laboratory of Electronic Restriction，Electronic Engineering Institute，Hefei 230037，China．Correspondent：XIA Jian-ming，E-mail：jianmingeei@163.com)

Abstract: As to detect and correct the labeling errors, a labeling errors detecting and correcting algorithm based on sparse manifold clustering and embedding and L_1 norm regularization is proposed. The proposed algorithm is based on the useful information in the original labels and the natural discriminating information which is contained in the data structure. Firstly, the original data are projected to the new space by using the sparse manifold clustering and embedding model. Then, a nearest neighbor classifier and a very small amount samples which are labeled correctly are used to obtain new labels for the original data. Meanwhile, the constructing labeling error detection model is built and then the sparse label detection vector which consists of 0 and 1 is obtained to modify the detection errors. The inaccurate and accurate labels correspond to 0 and 1 in the label detection vector respectively. Finally, the convex optimization scheme is introduced to solve the optimization problem and the convergence proofs are given. The experiment results show the effectiveness of the proposed algorithm based on the artificial data of complex manifold structure and the typical low-dimensional, high-dimensional data.

Key words: labeling errors；sparse manifold clustering and embedding；L_1 norm regularization；convex relaxation

0 引 言

信息社会中, 生物、军事、经济等领域的数据爆炸性增长给相应的机器学习算法带来了极大的挑战. 监督型学习算法通过处理已标签的样本获得分类准则, 如果忽略学习策略的影响, 则分类准则的好坏将严重依赖于样本的质量. 能否获得高质量的训练数据已成为决定机器学习效果好坏的一个重要条件. 传统的算法往往假设样本标签是正确的, 但在实际问题中, 由于录入错误、缺乏有效信息等原因, 标签往往会发生错误, 而标签错误对分类准则的影响要更甚于属性中的噪声影响, 会显著恶化学习的效果[1-3].

传统的监督学习算法或简单地忽视了标签错误, 或者假设算法对标签错误具有一定的鲁棒性[4]. 在标签出错的条件下, 有几类获得分类准则的方法: 1) 数据预处理的方法, 它是最直接简单的方法, 在数据进入分类器前进行置信度的分配和过滤, 将标签错误数据移除或是重新进行标注[5], 但是这种方法有可能剔除有用信息, 尤其是在训练样本规模较小的情况下; 2) 变精度粗糙集方法, 通过引入一些附加的参数来增强算法对标签错误的鲁棒性[6-7]; 3) 多事例学习的框

收稿日期: 2013-03-24；**修回日期:** 2013-12-04.

基金项目: 国家自然科学基金项目(61272333)；安徽省自然科学基金项目(1208085MF94, 1308085QF99).

作者简介: 夏建明(1982−), 男, 博士, 从事数据挖掘、机器学习的研究；杨俊安(1965−), 男, 教授, 博士生导师, 从事信号处理、智能计算等研究.

架[8]. 这几类方法共有的主要问题是没有一个自适应的参数设定准则, 如: 预处理方法无法设定一个通用的置信度有效分配方法, 且选择保留高置信度的样本往往会导致分布偏差; 变精度粗糙集方法[7]的正确分类率只能主观设定; 多事例学习模型[8]无法自适应地确定数据包的规模, 每一个包的规模都是一样的, 不能适应数据结构多变的情况. 还有一类方法是通过对标签转化概率进行建模, 如 RMLR (Robust Multiclass Logistic Regression)[9]. 对标签转化概率和 logistic 回归分类器的权重向量进行联合优化, 标签检测效果较好, 但若标签错误不满足其转化概率的假设, 则检测性能明显下降.

不需要标签的非监督学习算法是另一类机器学习算法, 旨在挖掘蕴含在样本结构中的模式信息, 直接忽略标签的类别信息, 不受标签的影响. 但由于数据结构复杂, 这类方法不能很好地挖掘其中的鉴别信息, 如 Elhamifar 提出的稀疏流形聚类嵌入模型 SMCE (Sparse Manifold Clustering and Embedding)[10], 这种方法在目标点所附着的流形上寻找其重构点, 既不同于基于稀疏表示的机器学习方法在数据全局寻找稀疏重构点, 也不同于流形学习模型中保持全局或局部欧式结构的特点, 虽然可自适应设定邻域, 聚类效果较好, 但也无法做到完整地反映数据中的鉴别信息.

在没有先验信息的条件下, 从数据中所获得的知识只有样本数值和可能出错的标签, 此时仅依靠不确定的标签或仅通过非监督学习算法获得样本类别信息都是不全面的, 需要综合利用二类方法才能得到足够的鉴别信息, 提高检测并改正错误标签的成功率.

含错标签中可能存在一部分未出错的有用信息, 数据结构中蕴含着模糊的鉴别信息. 基于综合利用这两类信息的思路, 本文假设数据中每类都有一个预先标注好的样本. 首先, 利用稀疏流形聚类嵌入模型对数据进行降维, 将数据投影到易分类的空间; 然后, 根据重新构造的数据和已有的正确标注数据进行分类获得新标签; 最后, 依托线性回归分类器的框架, 以标签检测向量和回归分类器权重向量为优化变量, 通过添加的稀疏标签检测向量在原标签和新标签之间进行选择, 向量中 0、1 的位置分别对应原标签中的错误标签和正确标签, 从而获得较为正确的标签输出. 该方法无需主观设定任何参数, 且能够综合利用数据本身的信息和含错标签中的部分有用信息, 实验效果要好于以往算法.

1 基于稀疏流形聚类嵌入模型和 L_1 范数正则化的标签错误检测算法

1.1 稀疏流形聚类嵌入模型

在机器学习的很多领域, 人们研究发现, 数据往往附着或近似附着在一个拥有低本质维的流形上. 通过对数据进行降维, 获得一个把握其流形结构的紧致表达, 有助于后期的处理. 数据的降维处理需要建立邻域图, 邻域规模是构建邻域图的关键参数, 太小可能无法获取足够的流形结构信息, 太大又可能满足不了用来抓住流形信息的本质要求. 此外, 流形的曲率和数据点的密度在流形的不同区域都是不同的, 因此使用一个固定的邻域规模也是欠妥当的.

同以往的维数约简模型一样, SMCE 模型[10]对每个数据点选择一个小的邻域, 并赋予邻域内每个点合适的权值; 与以往 LLE、LEM 等模型不同, SMCE 模型自适应地选择邻域和赋予权值, 并尽量选择数据点周围附着在同一个流形上的数据作为邻域点, 对于不同的数据集以及同一数据集中的不同数据点, 邻域规模皆为自动选择, 这就使得模型能够较好地把握数据中不同的流形结构.

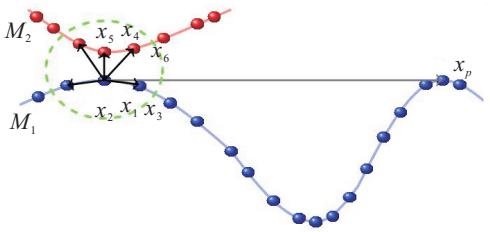

图 1 SMCE 模型邻域选择示意图

SMCE 模型算法核心思想如图 1 所示, M_1、M_2 为数据中的两个流形, 传统维数约简模型选中 x_1 的邻域为 $x_2 \sim x_6$, 而 SMCE 模型选中的邻域样本点仅有附着在流形上的 x_2、x_3.

设数据集 $\{x_i \in R^D\}_{i=1}^N$ 附着在 n 个不同的流形 $\{M_l\}_{l=1}^n$ 上, 流形各自的维度为 $\{d_l\}_{l=1}^n$. 其中一点 x_i 附着在 d_l 维流形 M_l 上, x_i 的重构向量为 $c_i \in R^{N\times 1}$ ($c_{ii}=0$), 那些附着在同一流形 M_l 上且包含在 x_i 邻域中的点展成了一个经过 x_i 的 d_l 维仿射子空间, 则重构向量 c_i 需满足

$$\begin{aligned} &||[x_1 - x_i \quad \cdots \quad x_N - x_i]c_i||_2 \leqslant \varepsilon, \\ &\text{s.t. } \mathbf{1}^{\mathrm{T}}\mathrm{c_i} = 1. \end{aligned} \tag{1}$$

其中 $\mathbf{1}$ 为 $[1 \quad \cdots \quad 1]_{N\times 1}$. 为了自动获取邻域规模, 文献 [10] 假设 c_i 是一个稀疏向量, 其中非零的元素便对应了 x_i 的邻域样本. 但是 c_i 的稀疏解并非唯一, 为了在获得 c_i 稀疏解的同时将结果局限在离点 x_i 较近的范围内 (如图 1, 仅选择点 x_2、x_3, 而舍去 x_p), 文献 [10] 对式 (1) 进行修改, 最终获得 SMCE 模型的核心目标函数表达式

$$\begin{aligned} &\min_{c_i} ||Q_i c_i||_1 + 0.5||X_i c_i||_2^2, \\ &\text{s.t. } \mathbf{1}^{\mathrm{T}} c_i = 1. \end{aligned} \tag{2}$$

其中 Q_i、X_i 分别设置为

$$Q_i = \frac{||x_j - x_i||_2}{\sum\limits_{t\neq i} ||x_t - x_i||_2} \in (0,1], \tag{3}$$

$$X_i \triangleq \left[\frac{x_1-x_i}{||x_1-x_i||_2} \quad \cdots \quad \frac{x_{i-1}-x_i}{||x_{i-1}-x_i||_2} \quad 0 \quad \frac{x_{i+1}-x_i}{||x_{i+1}-x_i||_2} \quad \cdots \quad \frac{x_N-x_i}{||x_N-x_i||_2}\right] \in R^{D\times N}. \tag{4}$$

相应地, 相似矩阵 $W=\{w_{ij}\}_{i,j=1}^{N}$ 由下式获得:

$$w'_{ii} \triangleq 0,$$
$$w'_{ij} \triangleq \frac{\dfrac{c_{ij}}{||x_j-x_i||_2}}{\displaystyle\sum_{t\neq i}\frac{c_{it}}{||x_t-x_i||_2}},\ j\neq i,$$
$$w_{ij}=|w'_{ij}|. \tag{5}$$

按照谱图的方法, D 为对角阵且对角线元素为

$$D_{ii}=\sum_j W_{ij}. \tag{6}$$

则最佳投影 X' 可由下式获得:

$$\max\ X'WX',$$
$$\text{s.t. } X'DX'=1. \tag{7}$$

通过广义特征值分解求解式 (7), 有

$$WX'=\lambda DX'. \tag{8}$$

通过稀疏流形聚类嵌入模型, 将原数据投影到易分类的空间, 获得聚类效果较好、类间距离较大的新数据 X'.

1.2　标签错误检测算法

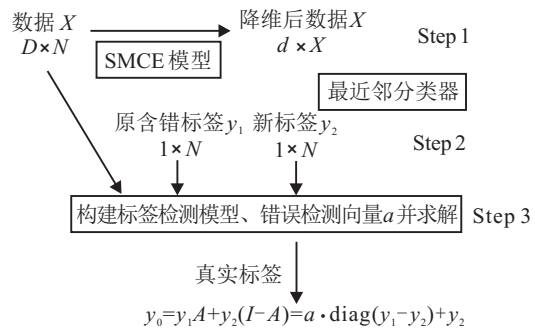

图 2　标签错误检测算法流程图

图 2 为本文提出的标签错误检测算法流程图. 设原数据集 $X\in R^{D\times N}$, 原标签 $y_1\in 1\times N$. 得到投影后的新数据 $X'\in R^{d\times N}$ 后, 基于已正确标注的样本, 可通过最近邻分类器获得新标签 $y_2\in 1\times N$. 设错误检测矩阵 $A\in N\times N$ 为对角阵, 错误检测向量为其对角线 $a\in 1\times N$, a 为仅含 0、1 元素的稀疏向量

$$A_{ij}=\begin{cases}1 \text{ or } 0,\ i=j;\\ 0,\ i\neq j.\end{cases} \tag{9}$$

其中 $a_i=A_{ii}$. 假设在两标签中选择可获得的最终标签为

$$y_0=y_1A+y_2(I-A). \tag{10}$$

设 $Y=\text{diag}(y_1-y_2)$, 即以 y_1-y_2 为对角线的对角阵, 则式 (10) 可转化为

$$y_0=(y_1-y_2)A+y_2=(aY+y_2). \tag{11}$$

其中: 稀疏向量 a 中为 1 的元素对应着含错标签中相对可靠的标签; 0 元素对应着通过稀疏流形聚类嵌入模型获得的新标签中的相对可靠标签. 综合利用含错标签和数据结构中有用的鉴别信息, 从而获得较为正确的标签.

在标签未受污染时, 一般的线性回归分类器通过求解 $y=w^{\mathrm{T}}X$ 对未知标签数据进行预测, 其目标函数一般为预测与实际值的误差最小化函数, 即

$$\min_w ||w^{\mathrm{T}}X-y||_2^2. \tag{12}$$

当标签受到污染后, 为了获得正确的标签, 将最终标签表达式 (11) 代入 (12), 进一步构建目标函数为

$$\min_{w,a} ||w^{\mathrm{T}}X-(aY+y_2)||_2^2+\lambda||a||_1;$$
$$\text{s.t. } a_i\in\{0,1\},\ i=1,2,\cdots,n. \tag{13}$$

即在最小化预测误差的同时得到 a 的稀疏解, 且内含元素约束为 0、1.

由于 L_1 范数为向量中元素的绝对值之和, 而起到选择标签作用的稀疏向量 a 内元素仅为 0 或 1, 则上式继续化简为

$$\min_{w,a} ||w^{\mathrm{T}}X-(aY+y_2)||_2^2+\lambda\sum_{i=1}^{N}a_i;$$
$$\text{s.t. } a_i\in\{0,1\}, i=1,2,\cdots,n. \tag{14}$$

目标函数需对 w、a 联合优化, 相当于一个同时求解 0、1 规划和连续变量优化的综合优化问题, 其整体并非一个凸函数. 通过交替固定一个变量, 同时优化另一个变量的方式来迭代优化: 当固定 a 求解 w 时, 优化问题是一个最小二乘问题, 可通过求导直接求解; 当固定 w 求解 a 时, 优化问题是一个 0、1 规划问题, 采取一种基于凸松弛类似于贪婪算法的优化方法.

1.3　交替迭代优化算法

1.3.1　固定 w 求解 a 的算法

1) 基于凸松弛的贪婪优化算法.

首先固定 w 求解 a, 将 a 中的元素松弛为连续值, 然后在优化的过程中, 获得元素仅为 0、1 的解. 通过对目标函数求导可以发现, 在获得的 n 维负梯度向量中, 仅有几个分量与负梯度向量夹角较小, 反映了主要下降方向. 若 a 在这几个分量上增大, 将使目标函数有最大的下降量. 设置初始值 a^0 为零向量 $[0,\cdots,0]$, 每次选取一个 a_i 置 1 获得新的 a, 使得 a 的变化方向 ∇a 与当前的负梯度方向夹角最小, 反复迭代操作, 最终可快速获得全局优化结果. 设 $b=w^{\mathrm{T}}X-y_2$, 则相应的拉格朗日函数为

$$L(a)=||b-aY||_2^2+\lambda\sum_{i=1}^{N}a_i=$$
$$(b-aY)(b-aY)^{\mathrm{T}}\cdots+\lambda a\mathbf{1}. \tag{15}$$

对 a 求导可得

$$\frac{\partial L(a)}{\partial a} = 2aYY^{\mathrm{T}} - 2bY^{\mathrm{T}} + \lambda \mathbf{1}^{\mathrm{T}} = \Big[\frac{\partial L(a)}{\partial a_1}, \cdots, \frac{\partial L(a)}{\partial a_i}, \cdots, \frac{\partial L(a)}{\partial a_N}\Big]. \tag{16}$$

设第 j 次迭代后获得的结果为 $a^j = \Big\{a_i^j \in \{0,1\} \Big| \sum_{i=1}^{N} a_i^j = j\Big\}$, 已选择为 1 的标签索引为 $S = \{s_1, s_2, \cdots, s_j\}$, 则第 $j+1$ 次迭代优化结果为

$$a^{j+1} = a^j + e_k. \tag{17}$$

其中 e_k 为第 k 个元素为 1、其余元素为 0 的单位向量. 则有

$$\nabla a^{j+1} = a^{j+1} - a^j = e_k. \tag{18}$$

为了使第 $j+1$ 次 a 的变化方向与负梯度的夹角最小, 此时目标函数可写为

$$\max_{k\notin S} \frac{(\nabla a^{j+1})^{\mathrm{T}}\Big(-\dfrac{\partial L(a^j)}{\partial a^j}\Big)}{||\nabla a^{j+1}||\Big\|\dfrac{\partial L(a^j)}{\partial a^j}\Big\|} = \max_{k\notin S} \frac{{e_k}^{\mathrm{T}}\Big(-\dfrac{\partial L(a^j)}{\partial a^j}\Big)}{||{e_k}^{\mathrm{T}}||\Big\|\dfrac{\partial L(a^j)}{\partial a^j}\Big\|} = \max_{k\notin S} \frac{-\dfrac{\partial L(a^j)}{\partial a_k^j}}{\Big\|\dfrac{\partial L(a^j)}{\partial a^j}\Big\|}. \tag{19}$$

式 (20) 表明: 在每次迭代时, 选择负梯度向量中最大分量对应的 a_i 为 1 即可, 具体算法流程如下.

输入: 当前的权重向量 w, 数据 X, 含错标签 y_1, 新标签 y_2, 系数 λ;

输出: 错误标签检测向量 a.

Step 1: 计算 $b = w^{\mathrm{T}}X - y_2$, $Y = \mathrm{diag}(y_1 - y_2)$.

Step 2: 初始化 S、S^0、a^0; $S = \varnothing$, S^0 为全体标签索引, $a^0 = [0, \cdots, 0]$.

Step 3: 第 j 次迭代求解步骤. 由式 (16) 求得梯度

$$\frac{\partial L(a^{j-1})}{\partial a^{j-1}} = \Big[\frac{\partial L(a^{j-1})}{\partial a_1^{j-1}}, \cdots, \frac{\partial L(a^{j-1})}{\partial a_i^{j-1}}, \cdots, \frac{\partial L(a^{j-1})}{\partial a_n^{j-1}}\Big],$$

$$\mathrm{s}_j = \arg\max_{i\notin S} \frac{-\dfrac{\partial L(a^{j-1})}{\partial a_i^{j-1}}}{\Big\|\dfrac{\partial L(a^{j-1})}{\partial a^{j-1}}\Big\|}.$$

若 $\dfrac{\partial L(a^{j-1})}{\partial a_{s_j}^{j-1}} < 0$, 且 $S \neq S^0$, 有

$$a^j = a^{j-1} + e_{s_j},\ S = S\bigcup s_j;$$

否则跳出迭代循环.

Step 4: 输出错误标签检测向量 a.

2) 基于凸松弛的贪婪算法的收敛证明.

定理 1 若目标函数为 $\min_a ||b - aY||_2 + \lambda\sum_{i=1}^{N} a_i$, 且有约束 $a_i \in \{0,1\}$, 则基于凸松弛的贪婪算法可以收敛到最优点.

证明 ①算法迭代步数是有限的. 因为 $a \in R^{1\times N}$, 每次优化前初始值为 $a^0 = [0, \cdots, 0]$, 优化过程中每迭代一次, 在 a 的相应位置将 0 置换为 1, 根据优化终止条件可知迭代步数最多只有 N 步, 算法迭代步数是有限的.

②目标函数在算法的优化过程中是递减的. 设 $a^{j+1} = a^j + e_k$, ε 为一极小正值, 则有

$$\frac{\partial L(a^j)}{\partial a_k^j} = \lim_{\varepsilon\to 0} \frac{L(a^j + \varepsilon e_k) - L(a^j)}{\varepsilon e_k}, \tag{20}$$

而选择的 e_k 与负梯度方向夹角最小, 则有

$$\frac{\partial L(a^j)}{\partial a_k^j} = \lim_{\varepsilon\to 0} \frac{L(a^j + \varepsilon e_k) - L(a^j)}{\varepsilon e_k} < 0, \tag{21}$$

即

$$L(a^j + \varepsilon e_k) < L(a^j), \tag{22}$$

进而可得 $L(a^j + e_k) < L(a^j)$, 即

$$L(a^{j+1}) < L(a^j). \tag{23}$$

因此, 目标函数在优化过程中是递减的. 由于算法迭代步数是有限的, 且目标函数值在本算法的优化过程中是递减的, 本文算法是收敛的.□

1.3.2 固定 a 求解 w 的算法及整体算法

当固定 a 求解 w 时, 设 $c = aY + y_2$, 目标函数可化简为

$$\min_w ||w^{\mathrm{T}}X - c||_2^2. \tag{24}$$

式 (24) 可通过对拉格朗日函数求导直接求解 w, 即

$$L(w) = (w^{\mathrm{T}}X - c)(w^{\mathrm{T}}X - c)^{\mathrm{T}}, \tag{25}$$

$$\frac{\partial L(w)}{\partial w} = 2(XX^{\mathrm{T}}w - xc^{\mathrm{T}}) = 0, \tag{26}$$

$$w = (XX^{\mathrm{T}})^{-1}Xc^{\mathrm{T}} = (XX^{\mathrm{T}})^{-1}X(aY + y_2)^{\mathrm{T}}. \tag{27}$$

当数据矩阵低秩时, XX^{T} 往往不可逆, 可依照文献 [11] 构建一个数值极小的单位矩阵 εI 与 XX^{T} 相加替代即可. 整体交替迭代优化算法流程如下.

输入: 数据 X, 含错标签 y_1, 新标签 y_2, 系数 λ, 极小正值 ε;

输出: 错误标签检测向量 a, 最终标签 y_0.

Step 1: 初始化 w=rand$(D,1)$, Y=diag$(y_1 - y_2)$.

Step 2: 计算 $b = w^{\mathrm{T}}X - y_2$.

Step 3: 根据前述基于凸松弛的贪婪优化算法计算获得错误标签检测向量 a, 由式 (27) 计算获得 w.

Step 4: 检验优化迭代终止条件. 若 $||w^{\mathrm{T}}X - (aY + y_2)||_2^2 > \varepsilon$, 则转到 Step 3 继续迭代优化; 否则, 输出错误标签检测向量 a、最终标签 y_0=$aY + y_2$.

基于稀疏流形聚类嵌入模型和 L_1 范数正则化的标签错误检测修正算法完整流程如下.

输入: 数据 X, 含错标签 y_1, 已正确标注样本, 系数 λ, 极小正值 ε;

输出: 错误标签检测向量 a, 最终标签 y_0.

Step 1: 通过 SMCE 模型将原数据投影到新的空间, 获得新数据 X', 并通过已正确标注样本和最近邻分类器获得新标签 y_2;

Step 2: 构建标签稀疏模型, 通过交替迭代优化算法获得错误标签检测向量 a、最终标签 y_0, 输出结果.

2　实验结果与分析

为了验证算法在复杂人工数据集和低维与高维数据的效果, 分别在人工数据集三叶草纽结、UCI 数据集 Iris、wine 以及人脸数据 YaleB 上对比验证算法效果. 将本文方法分别与对错误标签有一定的鲁棒性的 SVM 方法[9](核函数选择径向基核函数)、针对错误标签的 RMLR 方法进行对比. 数据描述如表 1 所示.

表 1　数据描述

数据集	样本数	属性数	类别数
三叶草纽结	200	3	2
Iris	250	4	3
Wine	345	6	2
YaleB	2 414	1 024	38

首先构造标签错误数据, SVM 方法以其在标签污染数据上建立的分类器在原数据上的分类精度为最终结果; 本文方法、RMLR 方法则先对标签错误数据进行处理获得标签修正数据, 再以各自的标签修正数据为训练集, 以原数据为测试集, 用 SVM 方法进行分类获得的分类精度为最终结果.

2.1　人工数据实验

在人工数据集上进行可视化实验, 两个三叶草纽结的流形互相近邻, 每个三叶草纽结为一类, 各含 100 个点, 原数据如图 3 所示. 按照 50% 的概率随机更改标注产生标签错误数据, 用 SVM 方法对数据进行处理获得分类器, 分别用 RMLR 方法、本文方法进行处理得到更新的标签及分类器, 重复 100 次获得其统计结果. 标签污染数据及各方法处理结果如图 4～图 7 和表 2 所示.

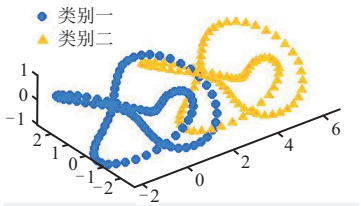

图 3　三叶草数据

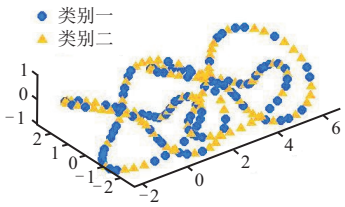

图 4　标签污染后数据

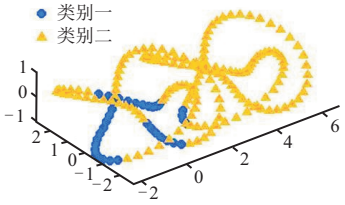

图 5　SVM 处理后标签

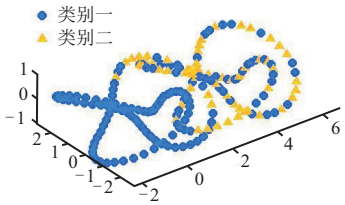

图 6　RMLR 方法处理后标签

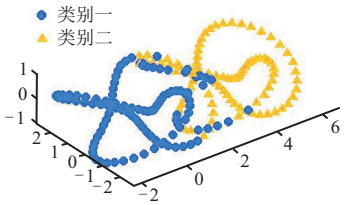

图 7　本文方法处理后标签

表 2　分类精度对比

数据集	SVM	RMLR	本文方法
三叶草纽结	59.23±6.54	66.46±7.43	77.55±6.75

图 4～图 7 是一次实验的结果. 图3、图 4 分别为原数据和标签错误数据, 如图 4 所示, 标签错误相对比较密集; 图 5 为 SVM 在标签错误数据上建立的分类模型对原数据的分类结果; 图 6、图 7 为 RMLR 方法、本文方法标签检测并修复后的数据. 由图 5～图 7 及表 2 可知, SVM 方法对标签错误具有一定的鲁棒性, RMLR 方法修复了一部分错误标签, 本文方法以较高的概率检测并修复了错误标签, 最终分类效果要好于其他算法.

2.2　UCI 数据实验

分别在 Iris、Wine 数据上进行对比实验, 按照从 10%～90% 的标签错误率随机更改标注产生标签错误数据, 再用 SVM 方法、RMLR 方法、本文方法进行处理, 分别重复 100 次获得统计结果. 图 8、图 9 分别为其分类精度对比.

由图 8、图 9 可知, SVM 方法对标签错误具有一定的鲁棒性, 但性能明显不如 RMLR 方法、本文方法;

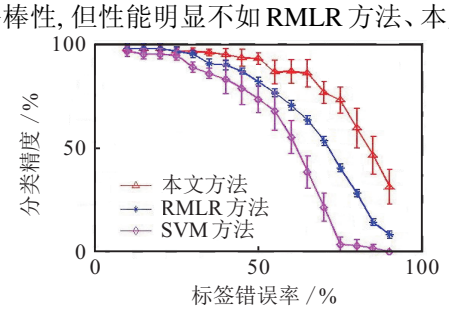

图 8　Iris 数据实验结果对比

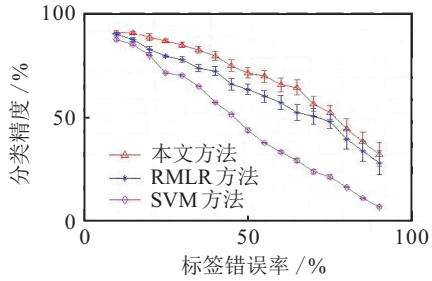

图 9 Wine 数据实验结果对比

RMLR 方法对标签错误有一定的处理能力, 但当标签错误率增大时, 性能下降较快, 而本文方法综合了数据结构中蕴含的自然鉴别信息和含错标签中的有用信息, 标签错误检测效果较好. 以 Iris 数据为例, 标签错误率小于 25% 时, SVM 方法与 RMLR 方法、本文方法的结果依然比较接近, 但当标签错误率增大时, 其性能迅速下降; 错误率小于 45% 时, RMLR 方法性能和本文方法相近, 但从 45% 后差距逐渐拉大; 本文方法的效果始终好于其他方法.

2.3 YaleB 数据实验

YaleB 人脸数据库中包含 38 个人在不同光照条件下的近似正面图像, 每人有约 64 张图像, 共 2 414 张, 像素为 32×32. 按照从 10%～90% 的标签错误率随机更改标注产生标签错误数据, 再用 SVM 方法、RMLR 法、本文方法进行处理, 分别重复 10 次获得统计结果. 图 10 为其分类精度对比.

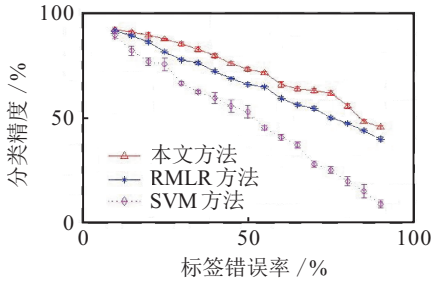

图 10 YaleB 数据实验结果对比

从图 10 可知, 本文方法在高维数据上的效果也明显好于 SVM 方法、RMLR 方法. 在标签错误率小于 20% 时, 本文方法和 RMLR 方法效果相近, 但在错误率继续增大的情况下, RMLR 方法与本文方法的差距始终保持在 5%～10%, 当标签错误达到 85% 以上时, 由于错误率太高, 本文方法与 RMLR 方法效果再次相近. 综合来看, 本文方法要优于 SVM、RMLR 方法.

3 结 论

标签错误在机器学习任务中经常出现, 它的存在严重影响到有监督机器学习算法的效果. 为了检测并修正标签错误, 区别于传统的有监督学习、无监督学习方法思路, 本文综合利用了含错标签中的有用信息和数据结构中蕴含的自然鉴别信息, 提出一种基于稀疏流形聚类嵌入模型和 L_1 范数正则化的标签错误数据检测修正方法, 并给出了相应的基于凸松弛的贪婪优化算法及其收敛证明. 在相关数据上的实验验证了本文算法的有效性.

参考文献(References)

[1] Li L, Darden T A, Weingberg C R, et al. Gene assessment and sample classification for gene expression data using a genetic algorithm/*k*-nearest neighbor method[J]. Combinatorial Chemistry and High Throughput Screening, 2001, 4(8) : 727-739.

[2] Yasui Y, Pepe M, Hsu L, et al. Partially supervised learning using an em-boosting algorithm[J]. Biometrics, 2004, 60(1): 199-206.

[3] Malossini A, Blanzieri E, Ng R T. Detecting potential labeling errors in microarrays by data perturbation[J]. Bioinformatics, 2006, 22(17): 2114-2121.

[4] Steven W Norton, Haym Hirsh. Classifier learning from noisy data as probabilistic evidence combination[C]. Proc of the 10 th National Conf on Artificial Intelligence. Menlo Park, 1992: 141-146.

[5] Jiang Y, Zhou Z H. Editing training data for knn classifiers with neural network ensemble[J]. Advances in Neural Networks, Lecture Notes in Computer Science, 2004, (7): 356-361.

[6] 纪霞, 李龙澍. 基于变精度动态容差关系的扩充粗糙集模型[J]. 系统仿真学报, 2009, 21(18): 5731-5734.
(Ji X, Li L S. Extended rough set model based on variable precision dynamic tolerance relation[J]. J of System Simulation, 2009, 21(18): 5731-5734.)

[7] Chen Dingjun. Promotion of variable precision covering rough set model[C] Int Conf on Intelligent Computation Technology and Automation. Changsha, 2010: 943-946.

[8] Thomas Leung, Yang Song, John Zhang. Handling label noise in video classification via multiple instance learning[C]. IEEE Int Conf on Computer Vision. Barcelona, 2011: 2056-2063.

[9] Jakramate Bootkrajang, Ata Kab an. Label-noise robust logistic regression and its applications[C]. The European Conf on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Bristol, 2012: 143-158.

[10] Ehsan Elhamifar, Rene Vidal. Sparse manifold clustering and embedding[C]. Proc of the 25th Annual Conf on Neural Information Processing Systems. Sierra Nevada, 2011: 55-63.

[11] Ehsan Elhamifar, Rene Vidal. Block-sparse recovery via convex optimization[J]. IEEE Trans on Signal Proc, 2012, 60(8): 4094-4107.

(责任编辑: 齐 霁)