文章编号: 1001-0920 (2016) 10-1791-07 DOI: 10.13195/j.kzyjc.2015.1306

面向属性效应控制的等均值约束支持向量回归机

刘解放[1,2], 王士同[1], 王 骏[1], 邓赵红[1]

(1. 江南大学 数字媒体学院，江苏 无锡 214122；2. 湖北交通职业技术学院 交通信息学院，武汉 430079)

摘 要: 针对现有的属性效应控制方法无法有效控制非线性回归建模的属性效应问题, 基于间隔最大化和结构风险最小化原则, 通过向 SVR 目标学习准则中施加等均值条件约束, 提出等均值支持向量回归机 (EM-SVR). 所提出的方法具有较好的泛化能力, 同时继承了 EM-LS 的良好性能. 实验结果验证了所提出方法的有效性.

关键词: 支持向量回归机；属性效应控制；等均值约束；结构风险最小化

中图分类号: TP181 **文献标志码:** A

Equal mean based support vector regression for attribute effect control

LIU Jie-fang[1,2], *WANG Shi-tong*[1], *WANG Jun*[1], *DENG Zhao-hong*[1]

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. School of Traffic Information，Hubei Communications Polytechnic，Wuhan 430079，China．Correspondent：LIU Jie-fang，E-mail：ljf-it@163.com)

Abstract: In view of the problem that the existing attribute effect control method can not effectively control the attribute effect in the nonlinear regression model, an equal mean-support vector regression(EM-SVR) based on the principles of the margin maximization and the structural risk minimization is proposed by using the constraint condition of equal mean, which has the good generalization ability and the characteristic of nonlinear regression. At the same time, the good performance of equal mean-least square(EM-LS) is also inherited. Finally, the experiment results show the effectiveness of the proposed method.

Keywords: support vector regression；attribute effect control；equal mean；structural risk minimization

0 引 言

数据挖掘中, 数据的可靠性是最关键的因素. 然而, 由于科技水平制约、观测或选择误差、不同数据来源、性别或种族歧视等原因, 搜集的数据集 (尤其是历史数据) 大多存在偏差, 这在很大程度上影响了目标函数的建模和预测. 针对这类有偏数据影响数据建模的问题, Pedreschi 等[1]提出了防歧视概念, Calders 等[2]提出了属性效应控制概念, 二者略有不同. 前者的研究主要针对由于歧视或偏见引起的有偏数据的偏差控制, 相近的研究有文献 [1, 3-7]; 后者的研究主要针对敏感属性引起的有偏数据的偏差控制. 事实上, 歧视或偏见总是基于某个事物的属性特征, 所以本质上, 后者的研究包含前者, 相似的研究有文献 [2, 8-12]. 本文从推广性的角度出发, 针对敏感属性引起的有偏数据集, 以及在其上的非线性回归建模方法进行研究.

近年来, 属性效应引起了数据挖掘领域研究人员的广泛关注. 例如, 通过对带有属性效应现象的历史遗留数据进行合理利用, 可以生成适用于当前数据环境的新分类器模型. 然而, 针对有偏数据集的挖掘, 由于数据中存在偏差, 如 k-means、SVM 和 CART 等经典学习算法不但不能取得理想的效果, 反而会加剧属性效应 (数据偏差). 在早期研究中, 人们大多在训练分类器前对数据进行预处理来移除敏感属性, 从而达到移除数据之间依赖关系的目的. 这些方法的局限性在于, 只对数据进行必要的预处理, 没有针对属性效应问题对已有的分类学习算法进行实质性的改进[5-7]. 文献 [8] 指出, 由于多个相关属性的间接影响, 仅简单移除原始数据中的个别敏感属性并不能真正消除属性效应; 另一方面, 移除部分属性会丢失部分有价值的信息, 这不利于后续学习器的训练. 近年来, 人们多通过改造已有的学习器来解决面向属性效应

收稿日期: 2015-10-24；**修回日期:** 2016-02-29.

基金项目: 国家自然科学基金项目(61272210, 61572236)；江苏省杰出青年基金项目(BK20140001)；江苏省自然科学基金项目(BK20130155, BK20151299).

作者简介: 刘解放(1982−), 男, 讲师, 博士生, 从事人工智能与模式识别的研究；王士同(1964−), 男, 教授, 博士生导师, 从事模式识别、数据挖掘、模糊神经网络等研究.

控制的分类和回归问题. Calders 等[8]通过向贝叶斯模型中添加隐变量, 使用期望最大化学习准则来优化模型参数, 从而提出了 3 种不同的贝叶斯分类学习方法; Kamishima 等[9]提出了适用于任意概率判别模型的正则化分类器, 该方法通过向分类学习模型中引入正则化项来强制约束分类器, 使之独立于敏感属性, 并进一步使用该方法解决了 logistic 回归问题; Kamiran 等[10]提出了基于决策树分类器, 当选择非叶子节点特征时, 该方法不但考虑了关于目标的信息增益, 而且还考虑了关于敏感属性的信息增益. 这些方法较好地解决了属性效应控制的分类问题. 针对回归问题, 目前在该方面的研究成果还较为少见, Calders 等[2]提出的等均值约束最小二乘 (EM-LS) 方法是对属性效应引起的有偏数据集进行线性回归的典型代表. 它基于误差最小化原则, 通过对最小平方误差准则施加等均值条件约束, 实现了线性回归中属性效应控制. 然而, 由于它采用了经验风险最小化原则, 限制了它的泛化性和实用性. 在现实生活中, 诸如生物形态学和社会科学等各个领域, 非线性有偏数据随处可见. 如何面向这类复杂数据进行非线性回归建模尚是学术研究的一个空白.

另一方面, 基于支持向量回归学习理论的非线性回归学习模型得到了广泛的研究[13-15]. 该类方法通过将原特征空间中的数据映射到高维空间中, 从而使非线性数据线性可分, 并基于间隔最大化学习准则实现了非线性数据的回归学习. 但是, 支持向量学习技术均没有考虑属性效应现象对非线性回归学习性能的影响, 因此不能直接用来解决有偏数据的回归学习问题.

以此为出发点, 本文将探讨面向敏感属性引起的有偏数据的非线性回归建模问题. 通过向 SVR 目标学习准则中施加等均值条件约束, 提出一种新型的回归学习模型 EM-SVR, 以解决训练数据中的属性效应问题. 所提方法不但能保证具有 EM-LS 处理有偏数据的良好性能, 同时又适合于各种复杂的非线性数据集, 具有较好的泛化能力、灵活性和处理非线性回归的特点. 实验结果验证了所提出方法的有效性.

1 EM-LS 算法

EM-LS 方法是针对有偏数据的线性回归问题. 给定样本集 $D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N},\boldsymbol{x}_i\in R^d,y_i\in R$. 向量 $\boldsymbol{x}_i$ 代表输入属性, 标量 y_i 代表第 i 个样本的目标值. 其目的是学习线性回归模型 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$. 为了便于表示, 把 b 看作权向量 $\boldsymbol{w}$ 的第 1 个元素 w_0, $x_0=1$ 为输入向量 $\boldsymbol{x}_i$ 的第 1 个元素. 假设基于某个敏感属性 $\boldsymbol{x}_s$, 把 D 分成不同的组. 本文采用二值属性, 把样本集 D 分成 D_1 和 D_2 两组, 所有结果都能推广到多个组, N_1 和 N_2 分别表示 D_1 和 D_2 中的样本个数. 例如, $\boldsymbol{x}_s$ 是性别, 把 D 分成男性样本 D_1 和女性样本 D_2.

1.1 EM-LS 基本原理

EM-LS 不仅考虑所建模型的目标值与观测值之间的误差平方和最小, 而且考虑了由敏感属性 $\boldsymbol{x}_s$ 所划分的两组样本集 D_1 和 D_2 的目标平均值应该相等. 其思想可由下式表示:

$$\begin{aligned}&\min\ \mathrm{Risk}(f);\\&\text{s.t.}\ \frac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in D_1}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i}{N_1}=\frac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in D_2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i}{N_2}.\end{aligned}\tag{1}$$

其中: 目标函数 $\mathrm{Risk}(f)$ 表示整个样本集 D 的经验风险; 条件表示等均值约束. 因此, EM-LS 本质上是带等均值约束的最小二乘方法.

1.2 EM-LS 方法及推导

基于式 (1) 的思想, Calders 等[2]给出了 EM-LS 方法的目标函数形式如下:

$$\begin{aligned}&\min\ \sum_{(\boldsymbol{x}_i,y_i)\in D}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-y_i\right)^2;\\&\text{s.t.}\ \frac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in D_1}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i}{N_1}=\frac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in D_2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i}{N_2}.\end{aligned}\tag{2}$$

定义

$$\boldsymbol{q}=\frac{\sum\limits_{(x_i,y_i)\in D_1}\boldsymbol{x}_i}{N_1}-\frac{\sum\limits_{(x_i,y_i)\in D_2}\boldsymbol{x}_i}{N_2},$$

则 $\boldsymbol{q}$ 是样本集 D_1 和 D_2 的平均向量之差, 即等均值条件可以表示为 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{q}=0$.

通过拉格朗日乘子法, 上述问题可以表示为带约束的拉格朗日函数

$$L(\boldsymbol{w},\lambda)=\sum_{(\boldsymbol{x}_i,y_i)\in D}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-y_i\right)^2+2\lambda\boldsymbol{w}^{\mathrm{T}}\boldsymbol{q},\tag{3}$$

其中 λ 是拉格朗日乘子.

对 $L(\boldsymbol{w},\lambda)$ 求关于系数 w_j 的偏导数, 并赋值为 0, 即

$$\frac{\partial L(\boldsymbol{w},\lambda)}{\partial w_j}=\sum_{(\boldsymbol{x}_i,y_i)\in D}2(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-y_i)x_{ij}+2\lambda q_j=0,\tag{4}$$

可解得

$$\boldsymbol{w}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}-\lambda(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{q}.\tag{5}$$

然后, 由 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{q}=0$, 可解得

$$\lambda=\frac{\boldsymbol{q}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{q}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{q}}.\tag{6}$$

把式 (6) 代入 (5), 可得

$$\boldsymbol{w}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}-\frac{\boldsymbol{q}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}}{\boldsymbol{q}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{q}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{q}.\tag{7}$$

其中: $\boldsymbol{X}$ 为 $[x_{ij}]_{N\times d}$ 矩阵, $\boldsymbol{q}$ 为 $d\times 1$ 的列向量, q_j 为 $\boldsymbol{q}$ 的分量, 符号上标“T”表示转置, 上标“−1”表示矩阵求逆.

2　EM-SVR 算法

2.1　算法推导

给定样本集 $D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$, $\boldsymbol{x}_i\in R^d$, $y_i\in R$. EM-SVR 在核空间构造非线性函数 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x})$, 使之与样本集中各样本间的 ε- 不敏感损失函数

$$|\boldsymbol{y}-f(\boldsymbol{x})|_\varepsilon=\begin{cases}0, & |\boldsymbol{y}-f(\boldsymbol{x})|\leqslant\varepsilon;\\ |\boldsymbol{y}-f(\boldsymbol{x})|-\varepsilon, & \text{otherwise}\end{cases}$$

最小. 通过引入 L2 范式的惩罚项和结构风险项, 可以构造如下 EM-SVM 目标函数的优化问题:

$$\begin{aligned}&\min_{\boldsymbol{w}}\frac{1}{2}||\boldsymbol{w}||^2+C\sum_{i=1}^N({\xi_i}^2+{\xi_i^*}^2).\\&\text{s.t. }\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)-y_i\leqslant\varepsilon+\xi_i,\ i=1,2,\cdots,N;\\&\quad y_i-\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)\leqslant\varepsilon+\xi_i^*,\ i=1,2,\cdots,N;\\&\quad\boldsymbol{w}^{\mathrm{T}}\boldsymbol{q}=0.\end{aligned}\tag{8}$$

其中

$$\boldsymbol{q}=\frac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in D_1}\varphi(\boldsymbol{x}_i)}{N_1}-\frac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in D_2}\varphi(\boldsymbol{x}_i)}{N_2}$$

是 D_1 和 D_2 在核空间的平均向量之差. 因此, $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{q}=0$ 表示为等均值约束条件, 其中 $\xi_i,\xi_i^*\geqslant 0$ 自动满足.

定理 1　如下最优化问题是 EM-SVR 原始优化问题的对偶问题:

$$\begin{aligned}&\max_{\tilde{\boldsymbol{\alpha}}\in R^{2N}}\ -\frac{1}{2}\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{K}}\tilde{\boldsymbol{\alpha}}-\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\begin{bmatrix}\varepsilon\boldsymbol{1}+\boldsymbol{y}\\ \varepsilon\boldsymbol{1}-\boldsymbol{y}\end{bmatrix};\\&\text{s.t. }\tilde{\alpha}_i\geqslant 0,\ i=1,2,\cdots,2N.\end{aligned}\tag{9}$$

其中

$$\begin{cases}\tilde{\boldsymbol{\alpha}}=[\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}^{*\mathrm{T}}]^{\mathrm{T}},\\ \boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}},\\ \boldsymbol{\alpha}^*=[\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*]^{\mathrm{T}};\\ \boldsymbol{y}=[y_1,y_2,\cdots,y_N]^{\mathrm{T}};\\ \tilde{\boldsymbol{K}}=\\ \begin{bmatrix}(\boldsymbol{K}_1-\boldsymbol{K}_2)+\dfrac{1}{2C}\boldsymbol{I} & -(\boldsymbol{K}_1-\boldsymbol{K}_2)\\ -(\boldsymbol{K}_1-\boldsymbol{K}_2) & (\boldsymbol{K}_1-\boldsymbol{K}_2)+\dfrac{1}{2C}\boldsymbol{I}\end{bmatrix},\\ \boldsymbol{K}_1=[\bar{k}_{ij}]_{N\times N},\ \bar{k}_{ij}=\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j),\\ \boldsymbol{K}_2=\dfrac{1}{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}[\hat{k}_{ij}]_{N\times N},\ \hat{k}_{ij}=\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j),\\ \phi(\boldsymbol{x}_i)=\Big(\dfrac{\sum\limits_{t=1}^{N_1}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i)}{N_1}-\dfrac{\sum\limits_{t=1}^{N_2}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i)}{N_2}\Big).\end{cases}\tag{10}$$

证明　引入两组拉格朗日乘子 $\boldsymbol{\alpha}^{(*)}$ 和 γ, 对式 (8) 构造拉格朗日函数

$$\begin{aligned}&L(\boldsymbol{w},\xi^{(*)},\boldsymbol{\alpha}^{(*)},\gamma)=\\&\frac{1}{2}||\boldsymbol{w}||^2+C\sum_{i=1}^N({\xi_i}^2+{\xi_i^*}^2)+\\&\sum_{i=1}^N\alpha_i(\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)-y_i-\varepsilon-\xi_i)+\\&\sum_{i=1}^N\alpha_i^*(y_i-\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)-\varepsilon-\xi_i^*)+\gamma\boldsymbol{w}^{\mathrm{T}}\boldsymbol{q}.\end{aligned}\tag{11}$$

根据 KKT 条件, $L(\boldsymbol{w},\xi^{(*)},\boldsymbol{\alpha}^{(*)},\gamma)$ 取得极值时有

$$\partial L/\partial\boldsymbol{w}=0,\tag{12}$$

$$\partial L/\partial\xi_i^{(*)}=0,\tag{13}$$

$$\partial L/\partial\gamma=0.\tag{14}$$

因此, 由式 (12) ∼ (14) 可得

$$\boldsymbol{w}=\sum_{i=1}^N(\alpha_i^*-\alpha_i)\varphi(\boldsymbol{x}_i)-\gamma\boldsymbol{q},\tag{15}$$

$$\xi_i^{(*)}=\alpha_i^{(*)}/(2C),\tag{16}$$

$$\gamma=\frac{\sum\limits_{i=1}^N(\alpha_i^*-\alpha_i)\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{q}}{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}.\tag{17}$$

将式 (15) ∼ (17) 代入 (11) 中, 简化后可得

$$\begin{aligned}&L=\\&-\frac{1}{2}\sum_{j=1}^N\sum_{i=1}^N(\alpha_j^*-\alpha_j)(\alpha_i^*-\alpha_i)\varphi(\boldsymbol{x}_j)^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+\\&\frac{1}{2}\sum_{j=1}^N\sum_{i=1}^N\Big[(\alpha_j^*-\alpha_j)(\alpha_i^*-\alpha_i)\frac{1}{\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}}\times\\&\Big(\frac{\sum\limits_{t=1}^{N_1}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i)}{N_1}-\frac{\sum\limits_{t=1}^{N_2}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i)}{N_2}\Big)^{\mathrm{T}}\times\\&\Big(\frac{\sum\limits_{t=1}^{N_1}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)}{N_1}-\frac{\sum\limits_{t=1}^{N_2}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)}{N_2}\Big)\Big]-\\&\frac{1}{4C}\sum_{i=1}^N({\alpha_i}^2+{\alpha_i^*}^2)-\varepsilon\sum_{i=1}^N(\alpha_i^*+\alpha_i)+\\&\sum_{i=1}^N(\alpha_i^*-\alpha_i)y_i.\end{aligned}\tag{18}$$

通过定义式 (10), 式 (18) 可以转化为矩阵形式 (9), 由此定理 1 成立. □

定理 2　对偶问题 (9) 必有解 $\tilde{\boldsymbol{\alpha}}=[\alpha_1,\alpha_2,\cdots,\alpha_N,\alpha_1^*,\cdots,\alpha_N^*]^{\mathrm{T}}$.

证明　对偶问题 (9) 是欧氏空间 R^{2N} 上的最优问题, 其目标函数是连续函数, 而可行域是非空有界闭集, 所以一定有解. □

根据前面的推导, 得到等均值支撑向量回归算法EM-SVR如下.

算法 1 等均值支撑向量回归算法EM-SVR.

输入: 有偏数据集 D;

输出: 拉格朗日乘子 $\tilde{\boldsymbol{\alpha}} = [\alpha_1, \alpha_2, \cdots, \alpha_N, \alpha_1^*, \cdots, \alpha_N^*]^{\mathrm{T}}$.

Step 1: 读入有偏数据集 D;

Step 2: 选取适当的核函数及核宽, 以及适当的精度 $\varepsilon > 0$ 和惩罚参数 $C > 0$;

Step 3: 根据式 (10) 计算核矩阵 $\tilde{\boldsymbol{K}}$;

Step 4: 求解式 (9) 所示的二次规划 (QP) 问题, 解得拉格朗日乘子 $\tilde{\boldsymbol{\alpha}}$.

2.2 EM-SVR 算法相关讨论

1) 时间复杂度分析. 算法EM-SVR的时间复杂度主要包括两方面: 一方面是计算核矩阵 $\tilde{\boldsymbol{K}}$ 中每个元素所花的总时间 t_M, 另一方面是对式 (9) 进行QP计算 $\tilde{\boldsymbol{\alpha}}$ 的时间 $t_{\tilde{\boldsymbol{\alpha}}}$. 相对于 $t_{\tilde{\boldsymbol{\alpha}}}$, t_M 基本可以忽略. 因此主要关注 $t_{\tilde{\boldsymbol{\alpha}}}$ 部分. 根据式 (9), 对其进行QP计算的时间可达 $O((2N)^3)$.

2) 核函数及相关参数选择. 不敏感损失参数 ε、惩罚参数 C、核函数及其参数的优化选择对EM-SVR算法学习回归模型的精度和推广能力起着决定性的作用. 大量的文献表明, CV、LOO和PSO是非常经典的参数优化方法, 依据简单、高效的原则, 本文采用CV方法对 ε 和 C 两参数进行优化选择; 对于核函数的构造, 本文采用了文献 [16] 的方法, 选择了高斯核函数并自动计算核宽 β.

$$\beta = (1/N^2)\sum_{i,j=1}^{N} ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2.$$

其中: N 为样本个数, $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 为样本向量.

3 实 验

为了评价本文所提出算法的性能, 在真实数据集上进行了大量实验验证. 首先考察引入等均值约束条件对SVR在有偏数据集上进行回归建模的影响; 然后以经典的SVR方法为参照, 评估EM-SVR方法的回归精度.

3.1 实验方法和数据集

表1列出了实验中采用的所有算法及主要参数. 其中: ε 是不敏感损失参数, C 是惩罚参数.

表 1 实验所用各种算法及主要参数

算法	所用数学模型及求解方法	主要参数
EM-LS	基于式 (7) 进行矩阵计算	无
EM-SVR	基于式 (9) 求解QP	ε,C
ε-SVR	基于L2-SVR目标函数求解QP	ε,C

通过五折交叉验证的方法分别搜索网格 $\{2^{-6}, 2^{-5}, \cdots, 2^2\}$ 和 $\{10^{-1}, 10^{-2}, \cdots, 10^6\}$ 来获取 ε 和 C 的最优值. EM-SVR和L2-SVR分别采用

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\{-||\boldsymbol{x} - \boldsymbol{y}||^2/\beta\},$$
$$\beta = (1/N^2)\sum_{i,j=1}^{N} ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2$$

为核函数及核宽, 其他未列出参数均采用交叉验证获得最优值.

本文实验采用如下3种指标对不同算法所得回归结果进行比较.

1) RMSE (Root Mean Square Error) 指标[17]

$$\mathrm{RMSE} = \frac{1}{\max y_i}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(f(\boldsymbol{x}_i) - y_i)^2}. \tag{19}$$

其中: y_i 为第 i 个样本的真实值, N 为所有样本个数.

2) MD (Mean Difference) 指标[2]

$$\mathrm{MD} = \frac{\sum\limits_{(\boldsymbol{x}_i, y_i)\in D_1} f(\boldsymbol{x}_i)}{N_1} - \frac{\sum\limits_{(\boldsymbol{x}_i, y_i)\in D_2} f(\boldsymbol{x}_i)}{N_2}. \tag{20}$$

其中: D 为样本集, 根据二值敏感属性 $\boldsymbol{x}_s$ 的取值将其划分为两个子集 D_1 和 D_2; N_1 和 N_2 分别为 D_1 和 D_2 中样本的个数. 如果该指标等于0, 则表示不存在属性依赖效应.

3) AUC (Area Under the ROC Curve) 指标[2]

$$\mathrm{AUC} = \frac{\sum\limits_{(\boldsymbol{x}_i, y_i)\in D_1}\ \sum\limits_{(\boldsymbol{x}_j, y_j)\in D_2} I(f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j))}{N_1 \times N_2}, \tag{21}$$

其中 $I(\cdot)$ 是指标函数, 当它的参数为真时, 返回1, 否则为0. AUC的变化范围为[0,1], 当为0.5时, 表示随机预测, 即不存在属性效应.

表2列出了实验所使用的各种数据集及主要属性参数.

表 2 3个数据集的主要特征

	Communities and Crime	Wine Quality	Census Income
N	1 994	6 497	19 952
M	99	11	14
y	Crime Rate	Rating	Wage
$\boldsymbol{x}_s$	Race	Type	Sex
N_1, N_2	970, 1 024	4 898, 1 599	10 358, 9 594
MD	0.22	0.94	6.3
AUC	0.79	0.76	0.85

实验中, 所有算法都基于64位的Matlab开发实现. 实验环境为Intel Core i7 3.40 GHz CPU, 8 G RAM, Windows 7 X64, Matlab 2010a.

3.2 实验结果分析

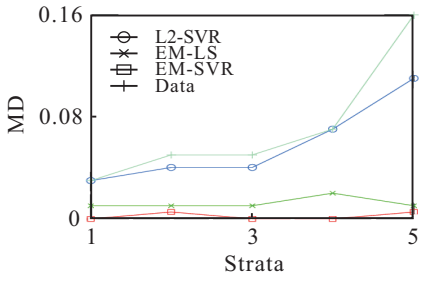
实验基于Communities and Crime[18]、Wine Quality[19]和Census Income (也即Adult data set)[18]数据集对本文所提算法进行评估, 这些数据集是数据挖掘领域公认的突显属性效应的3个典型有偏数据集.

Communities and Crime 数据集包含社区及社区犯罪率的社会经济信息, 含有 99 个属性. 该数据集被公认为关于种族 (Race ∈ {black, no-black}) 属性存在属性效应. 实验中, 对数据集进行了预处理, 删除了含有空值的属性, 并根据二值属性 Race 把数据集分为两组, 一组数据表示由全体黑人形成的社区, 另一组数据则表示由全体非黑人形成的社区. 对所有属性进行标准化, 在最终得到的数据集中, Communities and Crime 数据集总共包含 1 994 个实例, 其中黑人社区和非黑人社区分别包含 970 和 1 024 个样本. 对该数据集进行分析, 可以发现该数据集体现了目标 (犯罪率) 与敏感属性 Race 之间的强烈依赖关系. 黑人社区平均犯罪率为 0.35, 而非黑人社区平均犯罪率为 0.13 (MD = 0.22, AUC = 0.79). 数据集的相关信息如表 2 所示.

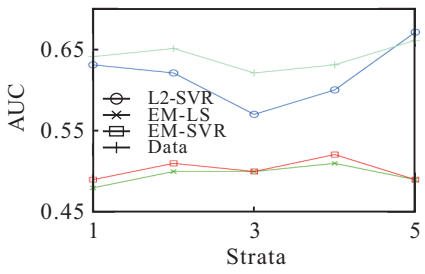
Wine Quality 数据集包含了对红酒和白酒评级的描述, 含有 11 个属性特征, 目标函数值描述了酿酒品质的评级, 取值范围为 [1,10]. 该数据集被公认为关于类型 (Type∈ {white, red}) 属性存在属性效应. 实验中, 对数据集进行归一化预处理. 原始数据集中, 两类酒的评级平均差较小. 为了构造出满足实验要求的有偏数据, 随机选取了 70 % 的白酒数据, 在它们的评级上加 1. 修改后红酒和白酒两类数据的 MD = 0.94, AUC = 0.76. 数据集的相关信息如表 2 所示.

Census Income 数据集抽取于人口普查数据. 该数据集被公认为关于性别 (Sex∈ {male, female}) 属性存在属性效应[2-3,10]. 总的说来, 女性的收入要低于男性. Census Income 数据集原本用于分类, 根据个人信息 (如职业、性别、学历等属性) 预测个人工资是否大于 5 万美金. 本文通过删除个别空值数据及属性值较少的字符属性, 并且离散化所需字符属性, 然后随机生成连续的目标工资. 修改后的数据集, 男性工资与女性工资平均差为 MD = 6.3, AUC = 0.85. 数据集的相关信息如表 2 所示.

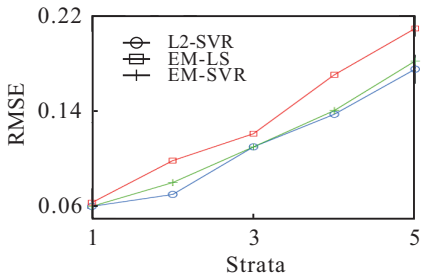
实验参考了文献 [2] 中的方法, 采用倾向评分分析 (PSA)[20] 对数据进行分层. 基于以上 3 个数据集, 分别运行 L2-SVR、EM-LS 和 EM-SVR 三个算法对分层后得到的每一层数据进行建模. 图 1 ∼ 图 3 给出了算法的运行结果, 这些结果均由五折交叉验证得到. 仿照文献 [2] 中的命名方法, 对各算法在分层数据上进行建模采用后缀 "-M" 进行标识. 为了便于比较, 在图 1 ∼ 图 3 的 (a) 和 (b) 中还给出了数据集真实的 MD



(a) 各层平均差

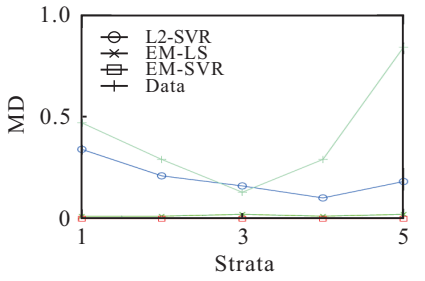


(b) 各层曲线下面积

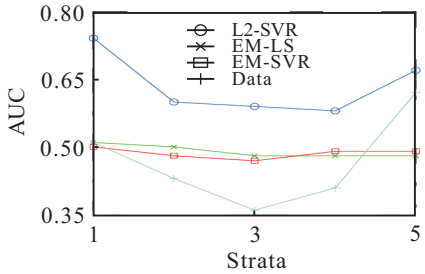


(c) 各层均方根误差

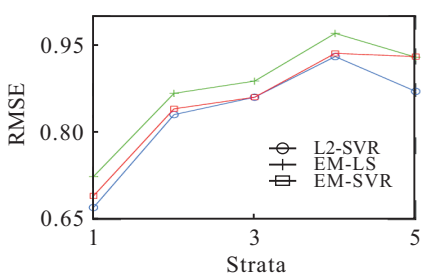
图 1　Communities and Crime 数据集实验结果



(a) 各层平均差

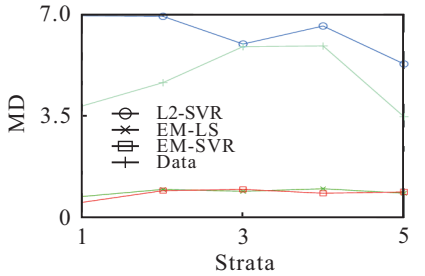


(b) 各层曲线下面积

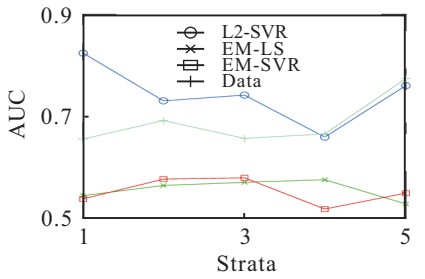


(c) 各层均方根误差

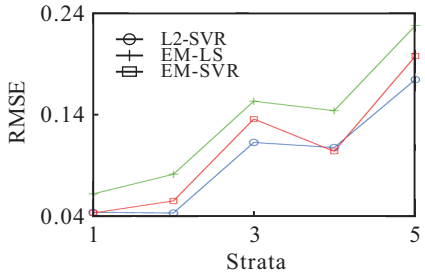
图 2　Wine Quality 数据集实验结果



(a) 各层平均差



(b) 各层曲线下面积



(c) 各层均方根误差

图 3 Census Income 数据集实验结果

值和 AUC 值.

图 1～图 3 分别给出了分层后每层平均值(MD)、曲线下面积(AUC)和均方根误差(RMSE). 如图 1～图 3 的 (a) 和 (b) 所示, 每层中, 犯罪率对种族的依赖、酒品评级对酒类型的依赖和工资对性别的依赖都显著降低.

从图 1～图 3 不难发现, 本文引入等均值约束是有效的, 它能消除每层的属性效应. 此外, 基于图 1～图 3 的 (a) 和 (b) 不难发现, EM-SVR 较 L2-SVR 和 EM-LS 提供了更好的属性效应控制效果. 基于图 1～图 3 的 (c) 来考察均方根误差, 由于采用了非线性回归模型, EM-SVR 的拟合效果明显优于 EM-LS 方法.

表 3 进一步比较了 3 种回归方法采用不同的模型得到的性能. 分全局模型 (相应的方法采用“-S”作为后缀进行标识, 如 L2-SVR-S, EM-LS-S, EM-SVR-S) 和分层模型 (如 L2-SVR-M, EM-LS-M, EM-SVR-M) 两种情况进行对比.

从表 3 不难发现: L2-SVR 没有考虑属性效应情况, 所以得到了较差的结果; 由于等均值约束的引入, EM-LS-S 和 EM-SVR-S 消除了整个数据集的所有属性效应, 但是 EM-LS 是线性回归模型, 所以得到的回归结果不令人满意, 而 EM-SVR 在施加等均值约束后, 均方根误差仍然能够得到与 L2-SVR 相接近的学习结果.

表 3 不同方法的实验结果比较

	Communities and Crime			Wine Quality			Census Income		
	MD	AUC	RMSE	MD	AUC	RMSE	MD	AUC	RMSE
Data	0.22	0.79	—	0.94	0.64	—	6.30	0.85	—
L2-SVR-S	0.22	0.82	0.13	0.92	0.89	0.84	7.68	0.92	0.17
L2-SVR-M	0.13	0.77	0.15	0.81	0.87	0.87	6.76	0.87	0.19
EM-LS-S	0.00	0.49	0.22	0.00	0.51	0.94	0.00	0.50	0.29
EM-LS-M	0.07	0.69	0.20	0.40	0.72	0.90	1.34	0.68	0.26
EM-SVR-S	0.00	0.48	0.15	0.00	0.51	0.89	0.00	0.51	0.19
EM-SVR-M	0.08	0.71	0.17	0.27	0.70	0.88	1.25	0.65	0.22

4 结 论

针对敏感属性引起的有偏数据进行建模是数据挖掘领域的一个重大挑战. 本文针对该问题对非线性回归的影响提出了新的算法 EM-SVR. 一方面, 基于核化的支撑向量回归机 SVR 学习框架, 使得所提算法具有处理非线性数据的能力; 另一方面, 吸收了 EM-LS 算法的优点, 通过等均值条件约束, 使得所提算法具有属性效应控制能力. 实验中的准确率及泛化能力验证了本文所提方法的有效性. 当然, 该算法时间复杂度相对较高, 如何降低时间复杂度, 使其能够处理大规模有偏数据集的属性效应是下一步的研究重点.

参考文献(References)

[1] Pedreshi D, Ruggieri S, Turini F. Discrimination-aware data mining[C]. Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. Las Vegas: ACM, 2008: 560-568.

[2] Calders T, Karim A, Kamiran F, et al. Controlling attribute effect in linear regression[C]. Proc of the IEEE 13th Int Conf on Data Mining(ICDM). Texas: IEEE, 2013: 71-80.

[3] Hajian S, Domingo-Ferrer J. A methodology for direct and indirect discrimination prevention in data mining[J]. IEEE Trans on Knowledge and Data Engineering, 2013, 25(7): 1445-1459.

[4] Ruggieri S, Pedreschi D. Data mining for discrimination discovery[J]. ACM Trans on Knowledge Discovery from Data, 2010, 4(2): 9.

[5] Kamiran F, Calders T. Classifying without discriminating[C]. Proc of the 2nd Int Conf on Computer, Control & Communication. Karachi: IEEE, 2009: 1-6.

[6] Kamiran F, Calders T. Classification with no discrimination by preferential sampling[C]. Proc of the 19th Annual Machine Learning Conf of Belgium and The Netherlands. Leuven: DTAI, 2010: 1-6.

[7] Pedreschi D, Ruggieri S, Turini F. Measuring discrimination in socially-sensitive decision records[C]. Proc of the SIAM Int Conf on Data Mining. Nevada: ASA, 2009: 581-592.

[8] Calders T, Verwer S. Three naive Bayes approaches for discrimination-free classification[J]. Data Mining and Knowledge Discovery, 2010, 21(2): 277-292.

[9] Kamishima T, Akaho S, Asoh H, et al. Fairness-aware classifier with prejudice remover regularizer[C]. Proc of the Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2012: 35-50.

[10] Kamiran F, Calders T, Pechenizkiy M. Discrimination aware decision tree learning[C]. Proc of the IEEE 10th Int Conf on Data Mining(ICDM). Sydney: IEEE, 2010: 869-874.

[11] Kamiran F, Karim A, Zhang X. Decision theory for discrimination-aware classification[C]. Proc of the IEEE 12th Int Conf on Data Mining(ICDM). Brussels: IEEE, 2012: 924-929.

[12] Kamiran F, Karim A, Verwer S, et al. Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset[C]. Proc of the IEEE 12th Int Conf on Data Mining Workshops(ICDMW). Brussels: IEEE, 2012: 370-377.

[13] 陈进东, 潘丰. 基于在线支持向量回归的非线性模型预测控制方法[J]. 控制与决策, 2014, 29(3): 460-464.
(Chen J D, Pan F. Online support vector regression-based nonlinear model predictive control[J]. Control and Decision, 2014, 29(3): 460-464.)

[14] 张军峰, 胡寿松. 基于无约束优化的非线性支持向量回归[J]. 控制与决策, 2009, 24(1): 125-128.
(Zhang J F, Hu S S. Nonlinear SVR based on unconstrained optimization[J]. Control and Decision, 2009, 24(1): 125-128.)

[15] 许敏, 王士同, 顾鑫, 等. 大样本领域自适应支撑向量回归机[J]. 软件学报, 2013, 24(10): 2312-2326.
(Xu M, Wang S T, Gu X, et al. Support vector regression for large domain adaptation[J]. J of Software, 2013, 24(10): 2312-2326.)

[16] Wang Shitong, Wang Jun, Chung Fu-lai. Kernel density estimation, kernel methods, and fast learning in large data sets[J]. IEEE Trans on Cybernetics, 2014, 44(1): 1-20.

[17] Deng Zhaohong, Jiang Yizhang, Choi Kup-Sze, et al. Knowledge-leverage-based TSK fuzzy system modeling[J]. IEEE Trans on Neural Networks and Learning Systems, 2013, 24(8): 1200-1212.

[18] Asuncion A, Newman D. UCI machine learning repository[DB/OL]. (1998-12-11)[2015-03-01]. http://archive.ics.uci.edu/ml/2007.

[19] Cortez P, Cerdeira A, Almeida F, et al. Modeling wine preferences by data mining from physicochemical properties[J]. Decision Support Systems, 2009, 47(4): 547-553.

[20] Rosenbaum P R, Rubin D B. Reducing bias in observational studies using subclassification on the propensity score[J]. J of the American Statistical Association, 1984, 79(387): 516-524.

(责任编辑：齐　霁)