计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20160519
54(9):1979-1991, 2017

面向大规模数据属性效应控制的核心向量回归机

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)
(ljf-it@163.com)

Core Vector Regression for Attribute Effect Control on Large Scale Dataset

Liu Jiefang[1,2], Wang Shitong[1], Wang Jun[1], and Deng Zhaohong[1]
[1](*School of Digital Media, Jiangnan University, Wuxi, Jiangsu* 214122)
[2](*School of Transportation and Information, Hubei Communications Technical College, Wuhan* 430079)

Abstract Attribute effect is a kind of phenomenon of data bias caused by sensitive attributes, which widely exists in real world. If not controlled, it will seriously affect the learning performance of regression model. In order to control the attribute effect in nonlinear regression model on large scale biased dataset, a novel fast equal mean-core vector regression (FEM-CVR) is proposed. First, a novel equal mean-support vector regression (EM-SVR) based on margin maximization criterion is proposed by using the constraint condition of equal mean. On this basis, the fact that the optimization problem of EM-SVR is equivalent to a center constrained-minimum enclosing ball (CC-MEB) problem is derived. Then a novel fast minimum enclosing ball based nonlinear regression learning algorithm for attribute effect control on large scale biased dataset, referred to as FEM-CVR, is further proposed by integrating the approximate minimum enclosing ball theory and reducing the original input dataset into the core set. In addition, some fundamental theoretical properties are deeply discussed. Finally, extensive experiments are conducted on synthetic and real datasets, and experimental results show that our FEM-CVR can effectively control attribute effect in nonlinear regression model on large scale biased dataset with good generalization ability, whose upper bound of the time complexity is independent of the size of the dataset, only related to the approximate parameter of the minimum enclosing ball $\bar{\varepsilon}$.

Key words regression learning; attribute effect control; center constrained-minimum enclosing ball (CC-MEB); equal mean constraint; large scale data

摘 要 属性效应在现实生活中广泛存在,如果不加以控制,将会严重影响回归学习的性能.针对大规模数据属性效应控制的非线性回归学习问题,提出了快速等均值核心向量回归机(fast equal mean-core vector regression, FEM-CVR).首先基于间隔最大化目标学习准则,通过施加等均值约束条件,提出了等均值支持向量回归机(equal mean-support vector regression, EM-SVR).在此基础上,证明了 EM-SVR 等价于一个中心约束最小包含球(center constrained-minimum enclosing ball, CC-MEB)问题,然后通过引入近似最小包含球理论,得到原始输入数据集的压缩集即核心集(core set),进一步提出了针对大规模数据属性效应控制的最小包含球快速非线性回归学习方法 FEM-CVR,并从理论上对相关性质

收稿日期:2016-07-13;**修回日期**:2016-12-09
基金项目:国家自然科学基金项目(61300151,61572236);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)
This work was supported by the National Natural Science Foundation of China (61300151, 61572236), the Distinguished Youth Foundation of Jiangsu Province (BK20140001), and the Natural Science Foundation of Jiangsu Province (BK20130155, BK20151299).

进行了深入分析.实验表明:该方法能够快速处理针对大规模数据属性效应控制的非线性回归学习问题,具有较好的泛化能力,并且其时间复杂度上限与数据集大小无关,仅与最小包含球近似参数 ε 有关.

关键词 回归学习;属性效应控制;中心约束最小包含球;等均值约束;大规模数据

中图法分类号 TP391

数据的可靠性是数据挖掘成败的关键因素之一.然而,由于科技水平制约、不同数据来源、系统误差、性别或种族歧视等原因,采集的数据(尤其是历史数据)往往存在对敏感属性的严重依赖[1-9].例如,早期的人口统计数据集(census income)[6-7]中,总的来说,女性工资远低于男性工资.类似该数据集中敏感属性(性别)所引起的数据严重偏差称为属性效应[8].它的存在严重影响学习器的训练和预测精度.因此,针对属性效应控制的问题引起了数据挖掘领域研究人员的广泛关注.

针对属性效应问题,研究人员从不同角度进行研究,提出了许多新的学习方法.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,它们只是对数据进行必要的预处理,而没有针对属性效应问题对已有的学习算法进行实质性的改进[2-5].文献[6]指出,由于多个相关属性的间接依赖,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,移除敏感属性会丢失部分有价值信息,这不利于后续学习器的训练.最近,研究人员大多通过改造已有的学习器来解决面向属性效应控制的分类和回归问题.例如,文献[6]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,提出了 3 种不同的贝叶斯分类学习方法.Kamishima 等人[7]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 等人[9]提出了基于决策树分类器,当选择非叶子节点特征时,该方法不但考虑关于目标的信息增益,而且考虑关于敏感属性的信息增益.这些方法较好地解决了针对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较为少见,Calders 等人[8]提出的等均值约束最小二乘(equal means-least square, EM-LS)方法是线性回归中属性效应控制的典型代表.它基于误差最小化原则,通过对最小平方误差和目标学习准则施加等均值约束条件而实现.然而,由于它使用了矩阵乘法和求逆运算,时间和空间复杂度都达到 $O(N^3)$,不但耗时且极易造成内存溢出,所以无法处理大规模数据的属性效应控制问题;另外,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.总之,这些方法虽然能够有针对性地解决属性效应在学习中的一些问题,但是仍然存在着局限性,主要表现在 2 个方面:1)算法复杂度较高,所以只适用于规模有限的数据集;2)大多面向属性效应控制的分类问题,对于非线性回归问题,却较少涉及.然而,在现实生活中诸如生物形态学和社会科学等各个领域,大规模非线性数据随处可见.如何面向复杂的大规模数据属性效应控制来进行非线性回归建模尚是学术研究的一个空白.

另一方面,基于最小包含球理论的大规模数据处理技术得到了深入的研究[10-12].该类方法通过求解近似最小包含球获得核心集,能够获得与原始输入数据集求解近似的结果且它的大小独立于原始输入数据集大小及样本维度,从而实现了大规模数据的压缩处理;此外,基于支持向量回归学习理论的非线性回归学习模型也得到了广泛的研究[13-14].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,并基于间隔最大化目标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对非线性回归学习性能的影响,因此不能直接用来解决针对属性效应控制的回归学习问题.

受上述思想的启发,本文将深入探讨面向大规模数据属性效应控制的非线性回归建模问题.首先,通过向支持向量回归机(support vector regression, SVR)目标学习准则中加入等均值约束条件提出了一种新型的非线性回归学习模型 EM-SVR(equal mean-support vector regression)以解决训练数据中的属性效应问题.进一步,针对大规模数据属性效应控制的学习问题,通过将其与中心约束最小包含球建立等价关系,提出基于最小包含球的快速非线性回归建模方法 FEM-CVR(fast equal mean-core vector regression),并从理论上深入探讨相关性质.最后通过实验验证了本文方法的有效性.

1 等均值支持向量回归机算法 EM-SVR

1.1 算法推导

给定 N 个输入-输出对$(\boldsymbol{x}_i,y_i)$,训练样本集 $D=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^N$,$\boldsymbol{x}_i\in\mathbb{R}^d$,$y_i\in\mathbb{R}$.选取合适的核空间变换 $\psi:\boldsymbol{x}\to\psi(\boldsymbol{x})$,则 EM-SVR 在核空间构造线性函数 $f(\boldsymbol{x})=\boldsymbol{\omega}^{\mathrm{T}}\psi(\boldsymbol{x})+b$,为了简化,该函数可以表示为 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x})$,其中 $\boldsymbol{w}=(\boldsymbol{\omega}^{\mathrm{T}},b)^{\mathrm{T}}$,$\varphi(\boldsymbol{x})=(\psi(\boldsymbol{x})^{\mathrm{T}},1)^{\mathrm{T}}$,使之与训练集中各样本间的 ε-不敏感损失距离

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

最小.通过引入 L2 范式的惩罚项和结构风险项,可构造并求解如下 EM-SVR 目标函数优化问题,

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

其中,$\boldsymbol{d}=\dfrac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in D_1}\varphi(\boldsymbol{x}_i)}{N_1}-\dfrac{\sum\limits_{(\boldsymbol{x}_i,y_i)\in D_2}\varphi(\boldsymbol{x}_i)}{N_2}$,$D_1$ 和 D_2 是由敏感属性 x_s' 划分数据集 D 所得到的 2 个子集;为便于描述,本文采用二值的敏感属性.N_1 和 N_2 分别表示 D_1 和 D_2 中样本个数,即 $\boldsymbol{d}$ 为 D_1 和 D_2 在核空间的平均向量之差.上述目标函数与 v-SVR[15] 类似,$\mu>0$,其作用与 v-SVR 中 v 相似,表示支持向量所占比重的下界;N 为样本个数,C 为惩罚因子,ε 为不敏感损失参数,$\xi_i,\xi_i^*\geqslant0$.

定理 1. 对于式(1)的优化问题,其对偶问题可描述为如下的凸二次规划形式,

$$\begin{aligned}\max_{\boldsymbol{\alpha}^{(*)}}\quad & (\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\alpha}^{*\mathrm{T}})\begin{pmatrix}\frac{2}{C}\boldsymbol{y}'\\ -\frac{2}{C}\boldsymbol{y}'\end{pmatrix}-(\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\alpha}^{*\mathrm{T}})\widetilde{\boldsymbol{K}}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\alpha}^*\end{pmatrix},\\ \text{s.t.}\quad & (\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\alpha}^{*\mathrm{T}})\mathbf{1}=1,\boldsymbol{\alpha},\boldsymbol{\alpha}^*\geqslant\mathbf{0},\end{aligned}\tag{2}$$

其中,$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_N)^{\mathrm{T}}$ 和 $\boldsymbol{\alpha}^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^{\mathrm{T}}$ 是拉格朗日乘子,$\boldsymbol{y}'=(y_1,y_2,\cdots,y_N)^{\mathrm{T}}$ 是已知输出;

$$\widetilde{\boldsymbol{K}}=\begin{pmatrix}(\boldsymbol{K}_1+\mathbf{1}\mathbf{1}^{\mathrm{T}}+\boldsymbol{K}_2)+\dfrac{\mu N}{C}\boldsymbol{I} & -(\boldsymbol{K}_1+\mathbf{1}\mathbf{1}^{\mathrm{T}}+\boldsymbol{K}_2)\\ -(\boldsymbol{K}_1+\mathbf{1}\mathbf{1}^{\mathrm{T}}+\boldsymbol{K}_2) & (\boldsymbol{K}_1+\mathbf{1}\mathbf{1}^{\mathrm{T}}+\boldsymbol{K}_2)+\dfrac{\mu N}{C}\boldsymbol{I}\end{pmatrix}\tag{3}$$

是核矩阵,其中,$\boldsymbol{K}_1=(\bar{k}_{ij})_{N\times N}$,$\bar{k}_{ij}=\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)$,$\boldsymbol{K}_2=\dfrac{1}{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d}}(\hat{k}_{ij})_{N\times N}$,$\hat{k}_{ij}=\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$,其中,

$$\varphi(\boldsymbol{x}_i)=\left(\frac{\sum\limits_{t=1}^{N_1}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i)}{N_1}-\frac{\sum\limits_{t=1}^{N_2}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i)}{N_2}\right),$$

$\boldsymbol{I}$ 为单位矩阵,$\mathbf{1}$ 是元素为 1 的向量,带上标(*)的参数表示带上标*的参数或不带上标*的参数.

证明.引入拉格朗日乘子 $\alpha_i^{(*)}$,β,构造式(1)的拉格朗日函数如下:

$$\begin{aligned}L(\boldsymbol{w},\xi_i^{(*)},\alpha_i^{(*)},\beta)=\ & \|\boldsymbol{w}\|^2+\frac{C}{\mu N}\sum_{i=1}^N(\xi_i^2+\xi_i^{*2})+\\ & 2C\varepsilon+\sum_{i=1}^N\alpha_i(y_i-\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)-\varepsilon-\xi_i)+\\ & \sum_{i=1}^N\alpha_i^*(\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)-y_i-\varepsilon-\xi_i^*)+\beta\boldsymbol{w}^{\mathrm{T}}\boldsymbol{d}.\end{aligned}\tag{4}$$

由 KKT(Karush-Kuhn-Tucker)条件可得,

$$\frac{\partial L}{\partial\boldsymbol{w}}=0\Rightarrow\boldsymbol{w}=C\Big(\sum_{i=1}^N(\alpha_i-\alpha_i^*)\varphi(\boldsymbol{x}_i)-\beta\boldsymbol{d}\Big),\tag{5}$$

$$\frac{\partial L}{\partial\xi_i^{(*)}}=0\Rightarrow\xi_i^{(*)}=\mu N\alpha_i^{(*)},\tag{6}$$

$$\frac{\partial L}{\partial\varepsilon}=0\Rightarrow\sum_{i=1}^N(\alpha_i+\alpha_i^*)=1.\tag{7}$$

由式(5)和式(1)的等式约束条件 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{d}=0$,可得:

$$\beta=\frac{\sum\limits_{i=1}^N(\alpha_i-\alpha_i^*)\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{d}}{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d}}.\tag{8}$$

将式(5)~(8)代入式(4),化简可得:

$$\begin{aligned}L=\ & -\sum_{i=1}^N\sum_{j=1}^N(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)-\\ & \sum_{i=1}^N\sum_{j=1}^N\Big[(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\frac{1}{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d}}\times\\ & \Big(\frac{\sum\limits_{t=1}^{N_1}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i)}{N_1}-\frac{\sum\limits_{t=1}^{N_2}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_i)}{N_2}\Big)^{\mathrm{T}}\times\\ & \Big(\frac{\sum\limits_{t=1}^{N_1}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)}{N_1}-\frac{\sum\limits_{t=1}^{N_2}\varphi(\boldsymbol{x}_t)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)}{N_2}\Big)\Big]-\\ & \sum_{i=1}^N\sum_{j=1}^N(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)-\frac{\mu N}{C}\sum_{i=1}^N(\alpha_i^*\alpha_i^*+\\ & \alpha_i\alpha_i)+\frac{2}{C}\sum_{i=1}^N(\alpha_i-\alpha_i^*)y_i.\end{aligned}\tag{9}$$

通过定义式(3),将式(9)写成对偶形式,也即得到式(2),因此,定理 1 成立. 证毕.

引理 1[16]. 设 X 是 $\mathbb{R}^d$ 上的一个紧集,若 $H(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 是 $X\times X$ 上的连续对称函数且关于任意 $\boldsymbol{x}_i\in X$ 的 Gram 矩阵半正定,则 $H(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 是 Mercer 核.

定理 2. 形如式(3)的 $\widetilde{\boldsymbol{K}}=(\widetilde{K}_{ij})$, $\widetilde{K}_{ij}=H(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathbb{R}$ 所描述的核函数是 Mercer 核.

证明. 首先定义 $\widetilde{\boldsymbol{K}}_1$ 和 $\widetilde{\boldsymbol{K}}_2$ 如下:

$$\widetilde{\boldsymbol{K}}_1=\begin{pmatrix}(\boldsymbol{K}_1+\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}})+\dfrac{\mu N}{C}\boldsymbol{I} & -(\boldsymbol{K}_1+\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}})\\ -(\boldsymbol{K}_1+\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}) & (\boldsymbol{K}_1+\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}})+\dfrac{\mu N}{C}\boldsymbol{I}\end{pmatrix},$$

$$\widetilde{\boldsymbol{K}}_2=\begin{pmatrix}\boldsymbol{K}_2 & -\boldsymbol{K}_2\\ -\boldsymbol{K}_2 & \boldsymbol{K}_2\end{pmatrix},$$

则式(3)可表示为 $\widetilde{\boldsymbol{K}}=\widetilde{\boldsymbol{K}}_1+\widetilde{\boldsymbol{K}}_2$,其中 $\widetilde{\boldsymbol{K}}_1$ 已由文献[17]证明它是 Mercer 核. 下面仅需证明 $\widetilde{\boldsymbol{K}}_2$ 是 Mercer 核,对于任意非零向量 $\boldsymbol{u}=(\boldsymbol{u}_1^{\mathrm{T}},\boldsymbol{u}_2^{\mathrm{T}})^{\mathrm{T}}\in X$,其中 $\boldsymbol{u}_1=(v_1,v_2,\cdots,v_N)^{\mathrm{T}}$, $\boldsymbol{u}_2=(v'_1,v'_2,\cdots,v'_N)^{\mathrm{T}}$,则有下列推导,

$$\boldsymbol{u}^{\mathrm{T}}\widetilde{\boldsymbol{K}}_2\boldsymbol{u}=(\boldsymbol{u}_1^{\mathrm{T}},\boldsymbol{u}_2^{\mathrm{T}})\begin{pmatrix}\boldsymbol{K}_2 & -\boldsymbol{K}_2\\ -\boldsymbol{K}_2 & \boldsymbol{K}_2\end{pmatrix}(\boldsymbol{u}_1^{\mathrm{T}},\boldsymbol{u}_2^{\mathrm{T}})^{\mathrm{T}}=$$

$$\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{K}_2\boldsymbol{u}_1-\boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{K}_2\boldsymbol{u}_1-\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{K}_2\boldsymbol{u}_2+\boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{K}_2\boldsymbol{u}_2=$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}v_i(\boldsymbol{K}_2)_{ij}v_j-\sum_{i=1}^{N}\sum_{j=1}^{N}v'_i(\boldsymbol{K}_2)_{ij}v_j-$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}v_i(\boldsymbol{K}_2)_{ij}v'_j+\sum_{i=1}^{N}\sum_{j=1}^{N}v'_i(\boldsymbol{K}_2)_{ij}v'_j=$$

$$\frac{1}{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d}}\Big[\sum_{i=1}^{N}\sum_{j=1}^{N}v_i\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)v_j-$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}v'_i\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)v_j-$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}v_i\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)v'_j+$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}v'_i\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)v'_j\Big]=$$

$$\frac{1}{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d}}\Big[\sum_{i=1}^{N}\sum_{j=1}^{N}v_i\sum_{k=1}^{N}\phi_k(\boldsymbol{x}_i)\phi_k(\boldsymbol{x}_j)v_j-$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}v'_i\sum_{k=1}^{N}\phi_k(\boldsymbol{x}_i)\phi_k(\boldsymbol{x}_j)v_j-$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}v_i\sum_{k=1}^{N}\phi_k(\boldsymbol{x}_i)\phi_k(\boldsymbol{x}_j)v'_j+$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N}v'_i\sum_{k=1}^{N}\phi_k(\boldsymbol{x}_i)\phi_k(\boldsymbol{x}_j)v'_j\Big]=$$

$$\frac{1}{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d}}\Big[\sum_{k=1}^{N}\Big(\sum_{i=1}^{N}v_i\phi_k(\boldsymbol{x}_i)\Big)^2-$$

$$\sum_{k=1}^{N}2\sum_{i=1}^{N}\sum_{j=1}^{N}v_i\phi_k(\boldsymbol{x}_i)\phi_k(\boldsymbol{x}_j)v'_j+$$

$$\sum_{k=1}^{N}\Big(\sum_{j=1}^{N}v'_j\phi_k(\boldsymbol{x}_j)\Big)^2\Big]=$$

$$\frac{1}{\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d}}\sum_{k=1}^{N}\Big(\sum_{i=1}^{N}v_i\phi_k(\boldsymbol{x}_i)-\sum_{j=1}^{N}v'_j\phi_k(\boldsymbol{x}_j)\Big)^2\geqslant 0.$$

另外,由定理 1 可知 $\boldsymbol{K}_2$ 为实数矩阵,且 $\boldsymbol{K}_2$ 的元素 $\hat{k}_{ij}$ 有 $\hat{k}_{ij}=\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)=\phi(\boldsymbol{x}_j)^{\mathrm{T}}\phi(\boldsymbol{x}_i)=\hat{k}_{ji}$,因此,$\boldsymbol{K}_2$ 是实对称矩阵,可推出 $\widetilde{\boldsymbol{K}}_2$ 也是实对称矩阵.

由上述推导可知,$\widetilde{\boldsymbol{K}}_2$ 是半正定矩阵且是 $X\times X$ 上连续对称函数,由引理1可知$\widetilde{\boldsymbol{K}}_2$是Mercer核矩阵.

又因 $\widetilde{\boldsymbol{K}}=\widetilde{\boldsymbol{K}}_1+\widetilde{\boldsymbol{K}}_2$,根据半正定的二次型定义,所以 $\widetilde{\boldsymbol{K}}$ 是 Mercer 核矩阵,也即定理 2 成立. 证毕.

根据定理 1 和定理 2 的推导,可得到等均值支持向量回归机算法 EM-SVR,其主要步骤如下:

算法 1. 等均值支持向量回归机算法 EM-SVR.

输入:数据集 D;

输出:拉格朗日乘子 $\boldsymbol{\alpha}^{(*)}$.

步骤 1. 读入数据集 D;

步骤 2. 根据式(3)计算核矩阵 $\widetilde{\boldsymbol{K}}$;

步骤 3. 求解式(2)所示的二次规划(quadratic programming, QP)问题,解得拉格朗日乘子 $\boldsymbol{\alpha}^{(*)}$;

步骤 4. 把解得的 $\boldsymbol{\alpha}^{(*)}$ 带入式(5)和式(8),即可求出相应回归模型 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x})$.

1.2 时间复杂度分析

算法 EM-SVR 的时间复杂度主要包括 2 方面:1)计算核矩阵 $\widetilde{\boldsymbol{K}}$ 中每个元素所花的总时间 t_{M};2)对式(2)进行 QP 计算 $\boldsymbol{\alpha}^{(*)}$ 的时间 $t_{\boldsymbol{\alpha}^{(*)}}$. 相对于 $t_{\boldsymbol{\alpha}^{(*)}}$, t_{M} 基本可以忽略. 所以,我们主要关注 $t_{\boldsymbol{\alpha}^{(*)}}$ 部分. 根据式(2),对其进行 QP 计算的时间可达 $O((2N)^3)$. 因此,从时间复杂度的角度考虑,EM-SVR 不适用于针对大规模数据属性效应控制的快速回归建模.

2 快速等均值核心向量回归机算法 FEM-CVR

由于等均值约束条件的引入,EM-SVR 可以很好地控制属性效应;核技巧的引入使之能够很好地解决非线性回归学习问题. 但是在求解 QP 问题的过程中,其时间复杂度可达 $O((2N)^3)$,因此面向大规模数据的属性效应控制,其处理效率低下. 本文将其与中心约束最小包含球问题建立等价关系,提出了基于中心约束最小包含球 CC-MEB 理论的快速等均值核心向量回归机算法 FEM-CVR.

2.1 最小包含球理论

文献[18]提出基于核心集的最小包含球$(1+\varepsilon)$近似算法. 在优化问题中,如果使用原始输入数据集的某个子集可获得与原始输入数据集求解近似的结果,那么就将这个子集称为核心集.

MEB (minimum enclosing ball)问题可描述为下列优化问题:

$$\begin{aligned}&\min_{R,\boldsymbol{c}} \quad R^2,\\ &\text{s.t.} \quad \|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2\leqslant R, i=1,2,\cdots,N.\end{aligned} \tag{10}$$

式(10)的对偶问题矩阵形式可以表示为

$$\begin{aligned}&\max_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}^{\mathrm{T}}\operatorname{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha},\\ &\text{s.t.} \quad \boldsymbol{\alpha}\geqslant\boldsymbol{0},\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1,\end{aligned} \tag{11}$$

其中,$\boldsymbol{K}=(k(\boldsymbol{x}_i,\boldsymbol{x}_j))_{N\times N}=(\varphi^{\mathrm{T}}(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j))_{N\times N}$为核矩阵,$\varphi$ 为核空间映射函数,$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_N)^{\mathrm{T}}$ 为拉格朗日乘子,$\boldsymbol{0}=(0,0,\cdots,0)^{\mathrm{T}}$,$\boldsymbol{1}=(1,1,\cdots,1)^{\mathrm{T}}$. 当核矩阵 $\boldsymbol{K}$ 对角线恒为常数 k 时,也即满足如式(12)时,

$$k(\boldsymbol{x}_i,\boldsymbol{x}_i)=k, \tag{12}$$

式(11)等价于式(13),

$$\begin{aligned}&\max_{\boldsymbol{\alpha}} \quad -\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha},\\ &\text{s.t.} \quad \boldsymbol{\alpha}\geqslant\boldsymbol{0},\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1.\end{aligned} \tag{13}$$

通过求解式(13),即可得到 MEB 的球心 $\boldsymbol{c}$ 和半径 R,

$$\begin{aligned}&\boldsymbol{c}=\sum_{i=1}^{N}\alpha_i\varphi(\boldsymbol{x}_i),\\ &R=\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\operatorname{diag}(\boldsymbol{K})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}}.\end{aligned} \tag{14}$$

Tsang 等人[18]指出,形如式(13)并满足式(12)的 QP 问题等价于最小包含球问题. 在此基础上,采用 MEB 理论的核心集方法开发了核心向量机(core vector machine, CVM)算法,研究表明,CVM 算法对处理大规模数据集表现出非凡的效率.

Tsang 等人[17]对核心向量机进行了扩展,提出广义核心向量机(generalized core vector machine, GCVM). 对形如式(11)的 QP 问题,即使所含的一次项不满足式(12),也可使用核心集技术进行快速求解,同时提出中心约束最小包含球 CC-MEB 来解决这一问题.

在 CC-MEB 中,给核空间任意样本点 $\varphi(\boldsymbol{x}_i)$ 增加一维新特征 $\delta_i\in\mathbb{R}$,形成新特征空间的新样本点 $\tilde{\varphi}(\boldsymbol{x}_i)=(\varphi(\boldsymbol{x}_i),\delta_i)^{\mathrm{T}}$,并约束 MEB 中增加的新特征维对应的球心定为圆点,即 CC-MEB 的中心是$(\boldsymbol{c},0)^{\mathrm{T}}$,这里 $\boldsymbol{c}$ 是原特征空间的 MEB 球心,然后求解新特征空间的 MEB 问题. CC-MEB 目标问题可描述为下列优化问题:

$$\begin{aligned}&\min_{R,\boldsymbol{c}} \quad R^2,\\ &\text{s.t.} \quad \|\varphi(\boldsymbol{x}_i)-\boldsymbol{c}\|^2+\delta_i^2\leqslant R, i=1,2,\cdots,N.\end{aligned} \tag{15}$$

式(15)的对偶问题矩阵形式可表示为,

$$\begin{aligned}&\max_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}^{\mathrm{T}}(\operatorname{diag}(\boldsymbol{K})+\boldsymbol{\Delta})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha},\\ &\text{s.t.} \quad \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1,\boldsymbol{\alpha}\geqslant\boldsymbol{0},\end{aligned} \tag{16}$$

其中,$\boldsymbol{\Delta}=(\delta_1^2,\delta_2^2,\cdots,\delta_N^2)^{\mathrm{T}}\geqslant\boldsymbol{0}$ 为用户定义的值,通过求解式(16),可得到 MEB 的球心 $\boldsymbol{c}$ 和半径 R,

$$\begin{aligned}&\boldsymbol{c}=\sum_{i=1}^{N}\alpha_i\varphi(\boldsymbol{x}_i),\\ &R=\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}(\operatorname{diag}(\boldsymbol{K})+\boldsymbol{\Delta})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}}.\end{aligned} \tag{17}$$

此外,任意点$(\varphi(\boldsymbol{x}_i),\delta_i)^{\mathrm{T}}$ 到球心$(\boldsymbol{c},0)^{\mathrm{T}}$ 的距离可表示为

$$\|\boldsymbol{c}-\varphi(\boldsymbol{x}_i)\|^2+\delta_i^2=\|\boldsymbol{c}\|^2-2(\boldsymbol{K}\boldsymbol{\alpha})_i+k_{ii}+\delta_i^2. \tag{18}$$

因为 $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1$,所以在式(16)的目标函数中增加任意一项 $-\eta\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}$,$\eta\in\mathbb{R}$,不会影响其最优解,于是,式(16)等价于式(19),

$$\begin{aligned}&\max_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}^{\mathrm{T}}(\operatorname{diag}(\boldsymbol{K})+\boldsymbol{\Delta}-\eta\boldsymbol{1})-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha},\\ &\text{s.t.} \quad \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{1}=1,\boldsymbol{\alpha}\geqslant\boldsymbol{0}.\end{aligned} \tag{19}$$

文献[17]指出,任意满足式(19)的 QP 问题都可认为是 CC-MEB 问题,可运用核心集技术进行快速求解.

2.2 FEM-CVR 算法

返回到式(2)的 QP 问题,首先定义 $\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}$ 和 $\boldsymbol{\Delta}$ 如下,

$$\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}=(\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\alpha}^{*\mathrm{T}}),\boldsymbol{\Delta}=-\operatorname{diag}(\tilde{\boldsymbol{K}})+\eta\boldsymbol{1}+\frac{2}{C}\begin{pmatrix}\boldsymbol{y}'\\-\boldsymbol{y}'\end{pmatrix}, \tag{20}$$

其中,η 为任意实数并保证 $\boldsymbol{\Delta}\geqslant\boldsymbol{0}$. 根据式(20),式(2)随即等价于式(21),

$$\begin{aligned}&\max_{\tilde{\boldsymbol{\alpha}}} \quad \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}(\operatorname{diag}(\tilde{\boldsymbol{K}})+\boldsymbol{\Delta}-\eta\boldsymbol{1})-\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{K}}\tilde{\boldsymbol{\alpha}}\\ &\text{s.t.} \quad \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{1}=1,\ \tilde{\boldsymbol{\alpha}}\geqslant\boldsymbol{0}.\end{aligned} \tag{21}$$

显然,式(21)满足式(19)的形式和约束条件,因此,它的 QP 问题可视为 CC-MEB 问题,也即 EM-SVR 可视为是 CC-MEB 问题,即可用核心集快速算法求解.

根据前面的推导,可得到快速等均值核心向量回归机算法 FEM-CVR,其主要步骤如下:

算法 2. 快速等均值核心向量回归机算法 FEM-CVR.

输入:大规模数据集 D、最小包含球逼近精度 $\bar{\varepsilon}$ 以及 η 和 $\boldsymbol{\Delta}$ 等参数;

输出:核心集 CS、拉格朗日乘子 $\tilde{\boldsymbol{\alpha}}$.

步骤 1. 设 t 为迭代计数器,且初值为 0,并初始化核心集 CS_0,最小包含球球心 $\boldsymbol{c}_0$,半径 R_0;

步骤 2. 若没有样本点 $\boldsymbol{x}$ 落在 $MEB(\boldsymbol{c}_t,(1+\bar{\varepsilon})R_t)$球外,则有 $CS=CS_t$,并转至步骤 6;

步骤 3. 根据式(18)查找离球心 $\boldsymbol{c}_t$ 最远的样本点 $\boldsymbol{x}$,并添加该点到核心集 $CS_{t+1}=CS_t\cup\{\boldsymbol{x}\}$;

步骤 4. 根据式(21)求解新的 CC-MEB,记为 MEB(CS_{t+1}),并且通过式(17)设定 $\boldsymbol{c}_{t+1}=\boldsymbol{c}_{MEB(CS_{t+1})}$, $R_{t+1}=R_{MEB(CS_{t+1})}$;

步骤 5. $t=t+1$,并返回步骤 2;

步骤 6. 终止训练,返回所需要的输出.

在实现 FEM-CVR 算法时有 2 点需要说明:

1)步骤 1 的初始化问题. 已有研究表明[17-19],合理选择数据点来初始化核心集可有效提高算法的性能. 本文中,我们采用如下方法:首先从原始输入数据集 D 中任取一点 $\boldsymbol{x}$,再选一点 $\boldsymbol{x}_a$,使其距离 $\boldsymbol{x}$ 最远;然后再找一点 $\boldsymbol{x}_b$,使其距离 $\boldsymbol{x}_a$ 最远. 最终初始化核心集为 $CS_0=\{\boldsymbol{x}_a,\boldsymbol{x}_b\}$,继而球心为 $\boldsymbol{c}_0=(\boldsymbol{x}_a+\boldsymbol{x}_b)/2$,半径为 $R_0=\|\boldsymbol{x}_a-\boldsymbol{x}_b\|/2$.

2) 步骤 2 和步骤 3 中涉及的距离计算问题. 对于每次迭代,就 N 个训练点来说,计算式(18)要花费时间为 $O(|CS_t|^2+(N-|CS_t|)|CS_t|)=O(N\times|CS_t|)$,当 N 非常大时,计算量巨大. 因此,可以使用概率加速方法[12,17],其思想在时间复杂度分析中有详细说明.

2.3 时间复杂度分析

算法 FEM-CVR 的运行时间主要集中在步骤 3~5. 当第 t 次迭代时,CS_t 的大小满足 $|CS_t|=t+2$. 在步骤 3 中需要利用式(18)计算原始输入数据集 D 减去核心集 CS_t 后剩余集合中每个样本点到 $\boldsymbol{c}_t$ 的距离,其运行时间为 $O(|CS_t|^2+(|D|-|CS_t|)\times|CS_t|)=O((t+2)N)\approx O(tN)$. 而步骤 4 中利用 QP 求解最小包含球的运行时间为 $O((t+2)^3)\approx O(t^3)$,所以,迭代过程中总的运行时间为 $O(tN+t^3)$. 一般情况下,$N\gg|CS_t|$,所以步骤 3 和步骤 4 的总运行时间近似为$O(tN)$,尽管它与训练样本个数呈线性关系,但对于大规模数据集而言运行时间也相当巨大. 因此,本文使用采样子集概率加速技术[12,17]完成步骤 3. 文献[12,17]指出当采样子集大小等于 59 时,$D-CS_t$ 中所有最远样本点的 5%在该子集中的概率不低于 95%,依照此方法,在确保性能的基础上,运行时间降为 $O(|CS_t|^2+59|CS_t|)\approx O(|CS_t|^2)\approx O(t^2)$,即迭代过程的总运行时间降为 $O((t+2)^3+t^2)\approx O(t^3)$,它仅与迭代次数 t 有关,而与训练集样本个数 N 无关. 而迭代终止条件 ε 控制着 t,一般 ε 越小,t 越大. 关于空间消耗,因为算法仅计算 CS 样本构成的核矩阵,所以其空间复杂度为 $O(|CS_t|^2)\approx O(t^2)$.

对于 1.1 节中提出的 EM-SVR,需要求解其对偶的 QP 问题,所以它的运行时间不小于 $O((2N)^{2.3})$,甚至达到 $O((2N)^3)$,而空间复杂度为 $O((2N)^2)$. 比较而言,FEM-CVR 在训练过程中的时间及空间复杂度都具有明显的优势.

FEM-CVR 算法是基于最小包含球近似算法的一个特例,因此在计算系统开销时,关于最小包含球核心集的结论同样适合 FEM-CVR 算法. 本文根据文献[11,17,19],给出如下性质:

性质 1. 给定最小包含球近似误差 ε,FEM-CVR 算法计算所得核心集大小的上界为 $O(1/\varepsilon)$,且该算法的迭代次数上界为 $O(1/\varepsilon)$.

性质 2. 给定最小包含球近似误差 ε,FEM-CVR 算法的时间复杂度上界为 $O(N/\varepsilon^2+1/\varepsilon^4)$.

性质 1 指出了 FEM-CVR 算法在最坏情况下的理论迭代次数;性质 2 指出了 FEM-CVR 算法在最坏情况下的理论运行时间,它与数据集大小 N 呈线性关系. 实际上,我们在实践中发现,在面向大规模数据属性效应控制时,算法的真实迭代次数和运行时间远低于理论最坏值,这也表明了 FEM-CVR 算法对大规模数据集处理的优势.

3 实验结果与分析

为评价本文所提算法的性能,我们分别在合成及真实数据集上进行了大量实验,主要从以下 2 个方面进行研究:1)面向属性效应环境,对比不同回归算法 EM-LS[8],SVR[20],FEM-CVR,查看它们控制属性效应的性能;2)面向大规模数据的属性效应环境,测试 FEM-CVR 的性能,并且对中心约束最小包含球学习理论中的重要参数 ε 作了深入研究.

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) 均方根误差(root mean square error, *RMSE*)指标[17]:

$$RMSE=\frac{1}{\max\limits_{1\leqslant i\leqslant N}y_i}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(f(\boldsymbol{x}_i)-y_i)^2}, \quad (22)$$

其中,y_i 为第 i 个样本的真实值,N 为所有样本个数.

2) 平均差(mean difference, *MD*)指标[8]:

$$MD(y,x'_s;D)=\frac{\sum\limits_{(\boldsymbol{x},y)\in D_1}f(\boldsymbol{x})}{N_1}-\frac{\sum\limits_{(\boldsymbol{x},y)\in D_2}f(\boldsymbol{x})}{N_2}, \quad (23)$$

其中,D 为训练样本集,根据二值敏感属性 x'_s 的取值将其划分为 2 个子集 D_1 和 D_2,N_1 和 N_2 分别表示 D_1 和 D_2 中样本个数,如果该指标等于 0,则表示数据中不存在属性效应.

3) 曲线下面积(area under the ROC curve, *AUC*)指标[8]:

$$AUC(y,x_s';D)=\frac{\sum_{(\boldsymbol{x}_i,y_i)\in D_1}\sum_{(\boldsymbol{x}_j,y_j)\in D_2}I(f(\boldsymbol{x}_i)>f(\boldsymbol{x}_j))}{N_1\times N_2},\quad(24)$$

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 $AUC=0.5$ 时,表示随机预测或不存在属性效应.

实验中,SVR 特指 L2-SVR,由 libSVM(Version 2.8)软件包实现. FEM-CVR 和 EM-LS 基于 MATLAB 环境实现,并根据文献[12]的建议,使用概率加速技术,每次采样子集的个数为 59,迭代终止条件 $\bar{\varepsilon}$ 设置为 10^{-6}. 所有实验采用高斯核 $k(\boldsymbol{x},\boldsymbol{y}')=\exp\{-\|\boldsymbol{x}-\boldsymbol{y}'\|^2/\beta\}$,核宽 β 由式 $\beta=(1/N^2)\sum_{i,j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2$ 自动计算. 其他未列出参数均采用交叉验证获得最优值. 实验环境为 Intel Core i7 3.40 GHz CPU, 8 GB RAM, Windows7 X64, MATLAB 2010a.

3.1 FEM-CVR, EM-LS, SVR 的比较

我们首先基于 Communities and Crime[8] 与 Wine Quality[21] 两个数据集对本文算法进行评估,它们是数据挖掘领域公认的突显属性效应的 2 个典型数据集.

Communities and Crime 数据集包含社区及社区犯罪率的社会经济信息. 本实验中,我们对数据集进行了预处理,删除了含有空值的属性,并根据二值敏感属性 Race 把数据集分为 2 组:1)表示由全体黑人形成的社区;2)表示由全体非黑人形成的社区. 并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 1994 个实例,其中黑人社区和非黑人社区分别包含 970 和 1024 个样本,该数据集共有 99 个属性. 对该数据集进行分析,我们发现该数据集体现了目标犯罪率 Crime Rate 和敏感属性 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($MD=0.22$, $AUC=0.79$). 表 1 给出了 Communities and Crime 数据集的相关信息.

Table 1 The Main Characters of Each Dataset

表 1 数据集的主要特征

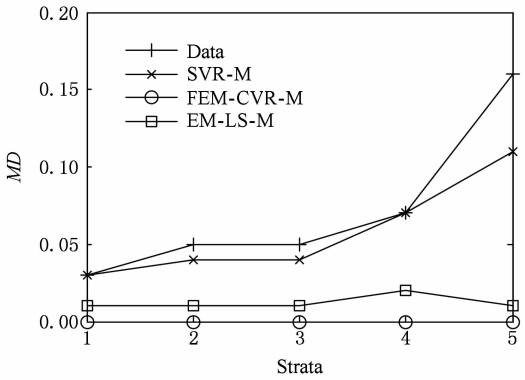
Datasets	N	M	y	x_s'	N_1	N_2	MD	AUC
Communities and Crime	1994	99	Crime Rate	$Race\in$ {black, non-black}	970	1024	0.22	0.79
Wine Quality	6497	11	Rating	$Type\in$ {white, red}	4898	1599	0.94	0.76
Census Income	199523	14	Wage	$Sex\in$ {male, female}	103582	95941	10.46	0.82
Friedman	100000	10	$f(\boldsymbol{x})$	$x_1'\in\{>0.5,\leqslant 0.5\}$	50023	49977	2.71	0.85
Census House	22784	121	Median Price	*black percentage* $\in\{>60\%,\leqslant 60\%\}$	8752	14032	3.65	0.77
ExtCrime	19940	99	Crime Rate	$Race\in$ {black, non-black}	9700	10240	1.52	0.72
ExtWine	64970	11	Rating	$Type\in$ {white, red}	48980	15990	2.68	0.84

Wine Quality 数据集包含了对红酒和白酒评级 Rating 的描述. 含有 11 个属性特征,函数输出描述了对酿酒品质的评级,取值范围为[1,10]. 实验中我们对数据进行归一化预处理. 原始数据集中,2 类酒的评级平均差较小. 为了方便观察试验结果,我们随机选取了 70%的白酒数据,在它们的评级上加 1. 修改后红酒和白酒 2 类数据的 $MD=0.94$, $AUC=0.76$. 数据集的相关信息如表 1 所示.

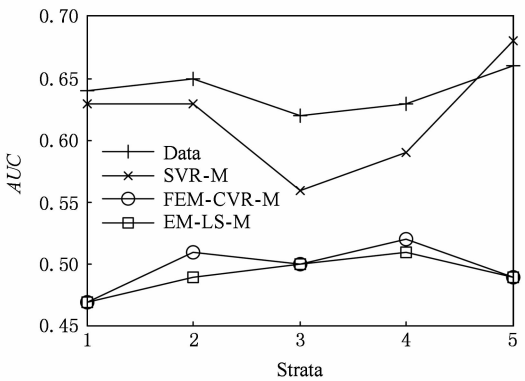
我们参考了文献[8]中的方法采用倾向评分分析(propensity score analysis, PSA)[22]对数据进行分层. 基于以上 2 个数据集,我们分别运行 EM-LS, SVR, FEM-CVR 三个算法对分层后得到的每一层数据进行建模. 图 1 和图 2 给出了算法的运行结果,它们均由十折交叉验证得到. 仿照文献[8]中的命名方法,我们对各算法在分层数据上进行建模采用后缀"-M"进行标识.

图 1 和图 2 分别给出了分层后每层 *MD*, *AUC* 和 *RMSE*. 为了便于比较,在图 1(a)(b)和图 2(a)(b)中我们还给出原始输入数据集的每层 *MD* 和 *AUC*. 如图 1(a)(b)和图 2(a)(b)所示,每层中犯罪率对敏感属性种族的依赖和酒品等级评定对酒类型的依赖都显著降低.

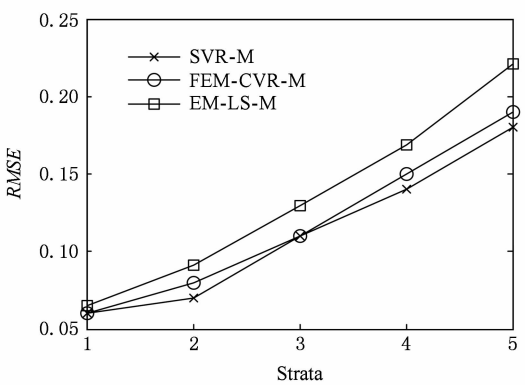
从图 1 和图 2 中不难发现,本文引入等均值约束是有效的,它能够使每层的 *MD* 值几乎为 0, *AUC* 值接近 0.5,也即几乎完全消除了属性效应. 而 SVR 没有考虑属性效应,所以它的 *MD* 值略大,而 *AUC* 值也趋向于 0 或 1,表明 SVR 不但不具有属性效应控制能力,甚至可能放大属性效应. 此外,基于图 1



(a) *MD* per layer



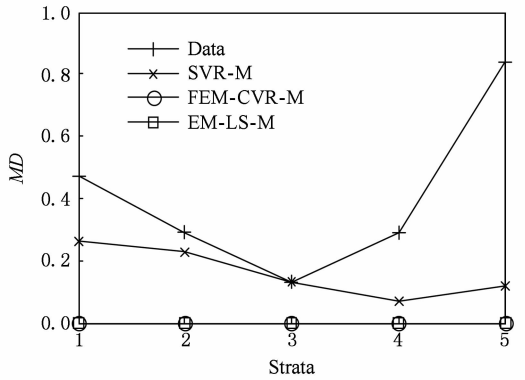
(b) *AUC* per layer



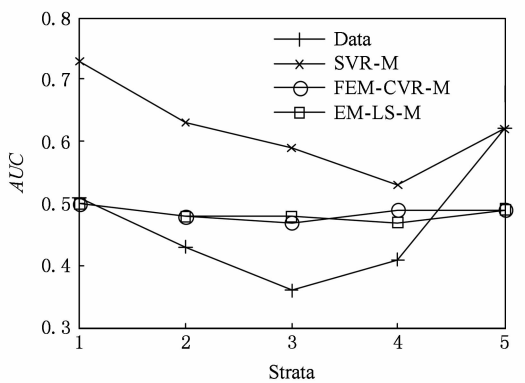
(c) *RMSE* per layer

Fig. 1 Experimental results on Communities and Crime dataset

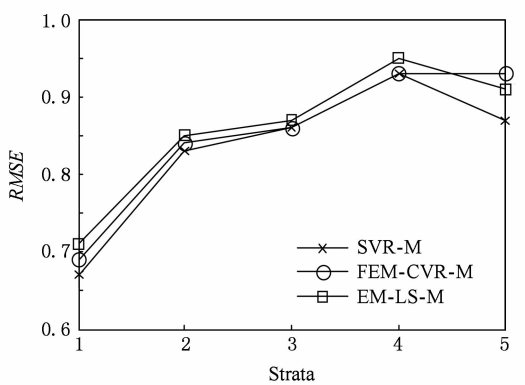
图 1 Communities and Crime 数据集实验结果



(a) *MD* per layer



(b) *AUC* per layer



(c) *RMSE* per layer

Fig. 2 Experimental results on Wine Quality dataset

图 2 Wine Quality 数据集实验结果

(c)和图 2(c)来考察均方根误差,本文的 FEM-CVR 由于采用了非线性回归模型,其拟合效果明显优于 EM-LS 方法. 因此,基于图 1 和图 2 我们不难发现 FEM-CVR 较 SVR 和 EM-LS 提供了更好的属性效应控制效果.

表 2 进一步比较了 3 种回归方法采用不同的模型得到的结果. 我们分全局模型(相应的方法采用"-S"进行标识,如 SVR-S,EM-LS-S,FEM-CVR-S)和分层模型(如 SVR-M,EM-LS-M,FEM-CVR-M)两种情况进行对比. 从表 2 中我们发现其结果类似图 1 和图 2. SVR 没有考虑属性效应,如 2 个数据集上 SVR-S 中 *AUC* 的值都大于原始数据集的 *AUC*,所以增大了数据偏差,获得了较差的结果. 由于等均值约束的引入,EM-LS 和 FEM-CVR 均能较好地消除数据集的属性效应,但是 EM-LS 由于是线性回归模型,所以得到的回归结果不令人满意. 而 FEM-CVR 在施加等均值约束后,仍然能够获得相对较小的均方根误差. 需要说明的是:为了消除数据属性效应(数据偏差),我们施加了等均值约束条件,此条件表示 2 组的预测结果应该相近;因此,其必定

导致误差加大.这也是 SVR 的均方根误差小于其他 2 个算法的原因,但其不具有属性效应控制能力.

Table 2 Comparison of Experimental Results for Different Methods

表 2 不同方法的实验结果比较

Method	Communities and Crime			Wine Quality		
	MD	*AUC*	*RMSE*	*MD*	*AUC*	*RMSE*
Data	0.22	0.79		0.94	0.64	
SVR-S	0.22	0.82	0.14	0.92	0.89	0.84
SVR-M	0.13	0.77	0.18	0.81	0.87	0.87
EM-LS-S	0.00	0.49	0.22	0.00	0.51	0.94
EM-LS-M	0.07	0.69	0.20	0.40	0.72	0.90
FEM-CVR-S	0.00	0.48	0.16	0.00	0.51	0.89
FEM-CVR-M	0.08	0.71	0.15	0.27	0.70	0.88

3.2 大规模数据环境实验

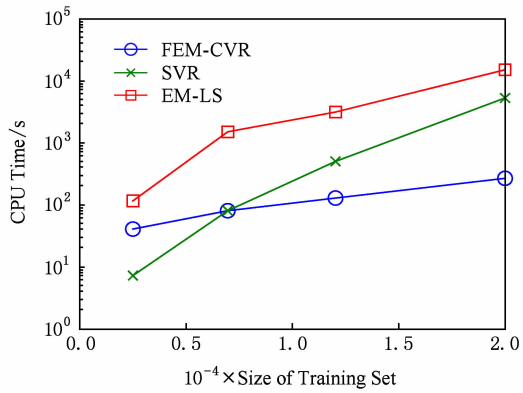
为了进一步验证大规模数据属性效应环境下 FEM-CVR 的性能,我们基于文献[10]的方法对 Communities and Crime 和 Wine Quality 数据集进行了扩充.扩充后的新数据集每个属性的随机偏移量服从正态分布 $\mathcal{N}(0,1)$,从而构造出大规模数据集,扩充后的 Communities and Crime 数据集记为 ExtCrime,样本数为 19 940,Wine Quality 数据集记为 ExtWine,样本数为 64 970.另外,新增加了 2 个 UCI 数据集 Census Income[6-7] 及 Census House[23] 和 1 个合成数据集 Friedman[24].表 1 显示了这些数据集的主要特征.

Census Income 数据集抽取于人口普查数据.该数据集被认为关于敏感属性性别 Sex 存在属性效应,总的说来,女性的工资远低于男性工资.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,并且离散化所有字符属性,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 $MD=10.46$,曲面下面积 $AUC=0.82$.

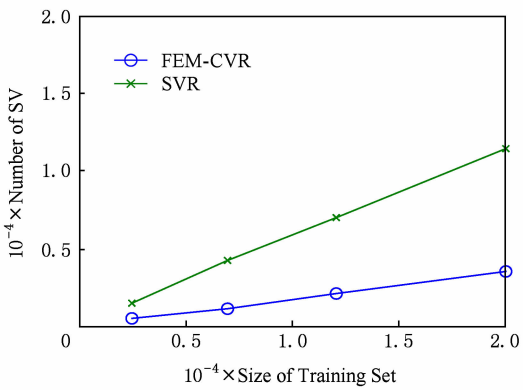
Census House 数据集是由美国统计局提供的房屋调查数据,它基于某地区的人口结构和房屋市场预测房子的平均价.Friedman 是 1 个合成数据集.Census House 和 Friedman 这 2 个数据集偏差并不明显.为了方便观察试验结果,通过采用 3.1 节处理 Wine Quality 数据集相同的方法放大它们的属性效应,处理后的数据集主要特征如表 1 所示.

为了验证 FEM-CVR 能够有效处理针对大规模数据属性效应控制的回归问题,我们首先从 Census House 数据集中分别随机抽取不同容量的子集,分别运行 EM-LS,SVR,FEM-CVR,并采用十折交叉验证,对比它们的 CPU 运行时间、支持向量个数(SV)和均方根误差.

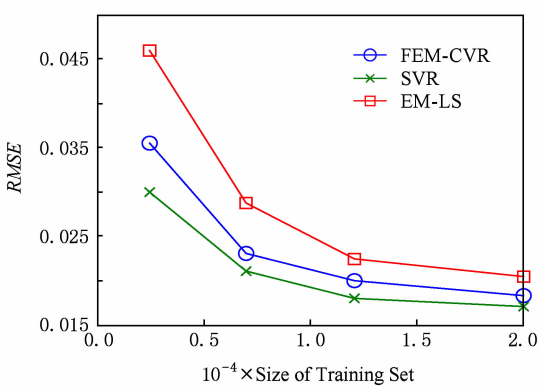
图 3(a)显示,训练样本个数较少时,FEM-CVR 在求解核心集过程中需要迭代外扩(多次求解 QP 问题),所以其速度优势表现不明显,甚至其运行速度低于 SVR;但是随着样本个数的增多,采用基于最小包含球的核心集进行优化求解的速度优势得到了充分的体现,其时间复杂度与训练样本个数基本呈线性关系,明显优于同样具有处理属性效应能力的 EM-LS 算法.



(a) CPU time on training set of different sizes



(b) Number of SV on training set of different sizes



(c) *RMSE* on training set of different sizes

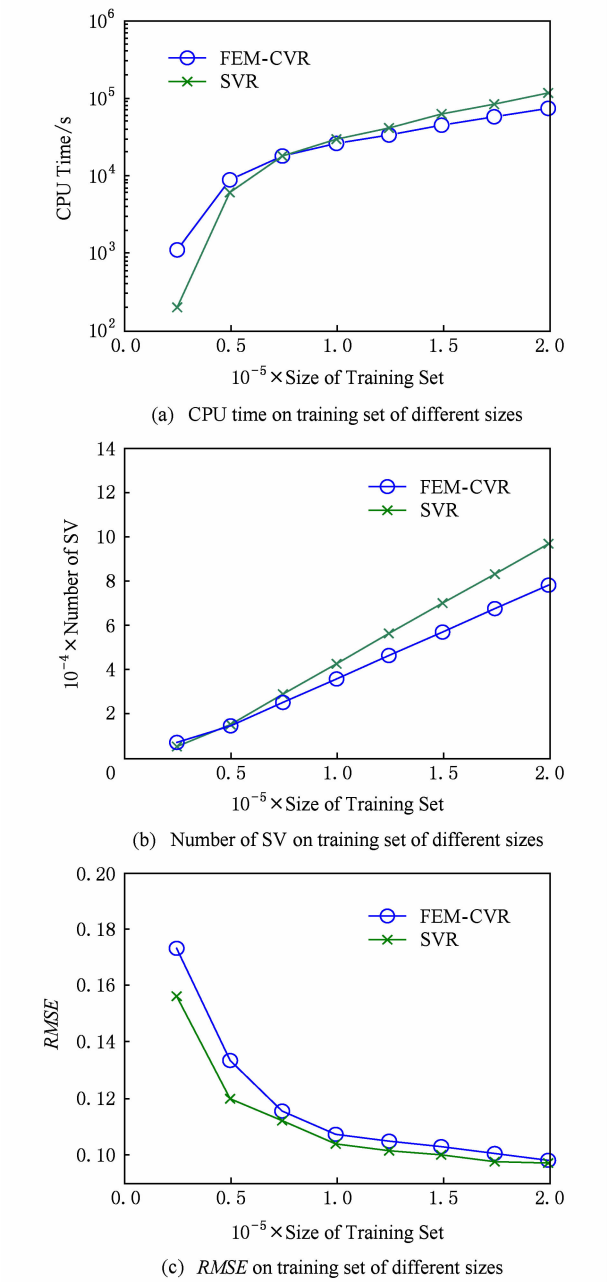
Fig. 3 Experimental results on Census House dataset

图 3 Census House 数据集实验结果

图 3(b)显示,采用不同大小的样本容量训练时,SVR 选择大约 60%样本作为的支持向量;而 FEM-CVR 的支持向量数目远低于 SVR. 较少的支持向量个数有助于减少运行时间.

图 3(c)显示,FEM-CVR 可以取得与 SVR 接近的均方根误差,其值明显小于同样具有处理属性效应能力的 EM-LS 算法.

采用相同的实验策略,我们在 Census Income, Friedman, ExtCrime 和 ExtWine 数据集上进行了与上文相同的实验,图 4～7 给出了相应的实验结果. 通过观察可以发现这些结果表现出类似上文的特征. 需要说明的是 EM-LS 算法在训练数据集大于 20 000 时,内存溢出,我们无法给出结果;小于 20 000 时,CPU 运行时间也明显高于 FEM-CVR,这充分说明了 EM-LS 算法在处理大规模数据回归问题方面的不足. 另外,表 3 给出了分别选取 Census Income 和 Census House 数据集中 10 000 个样本时,FEM-CVR 最大选择了不足 2 000 多个核心向量. 核心向量的减少,致使支持向量的减少,从而加快了运行速度.



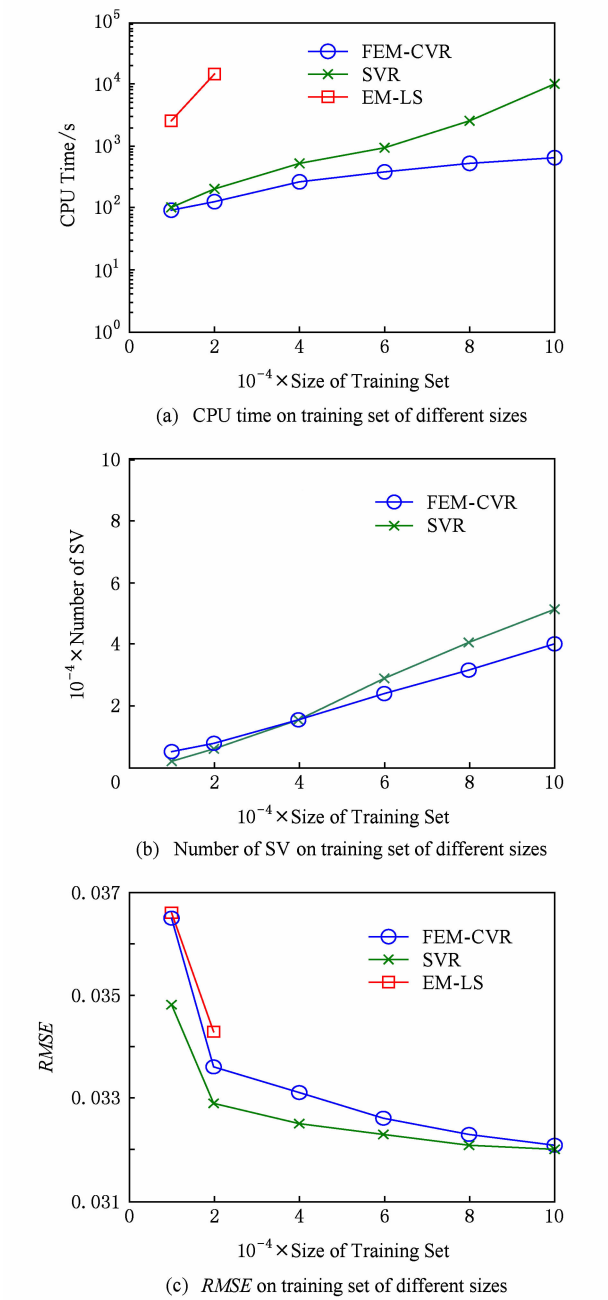
(a) CPU time on training set of different sizes

(b) Number of SV on training set of different sizes

(c) *RMSE* on training set of different sizes

Fig. 4 Experimental results on Census Income dataset

图 4 Census Income 数据集实验结果



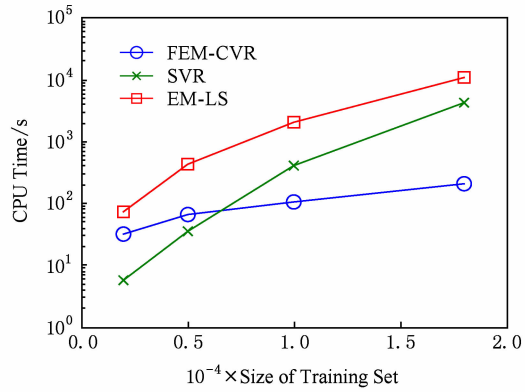
(a) CPU time on training set of different sizes

(b) Number of SV on training set of different sizes

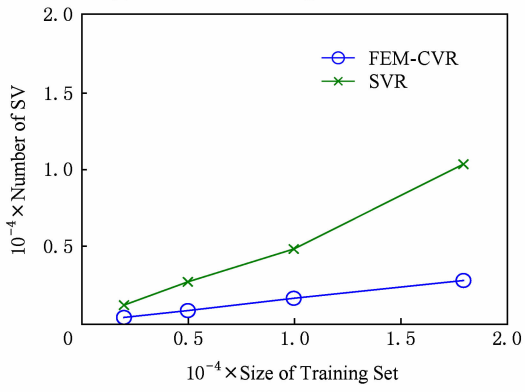
(c) *RMSE* on training set of different sizes

Fig. 5 Experimental results on Friedman dataset

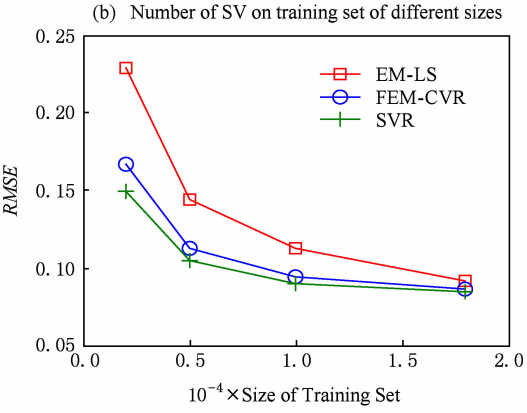
图 5 Friedman 数据集实验结果



(a) CPU time on training set of different sizes



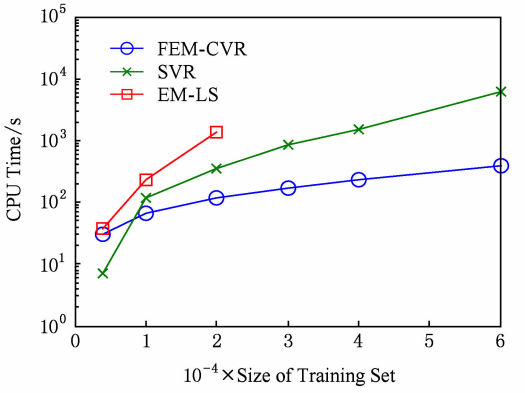
(b) Number of SV on training set of different sizes



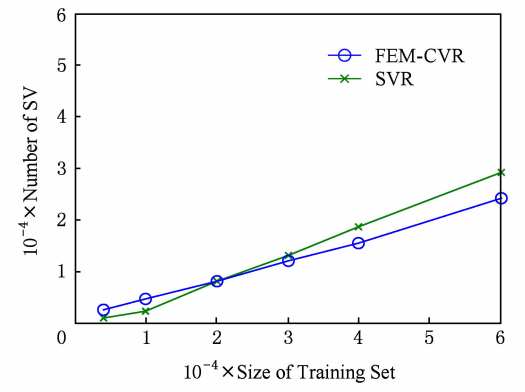
(c) *RMSE* on training set of different sizes

Fig. 6 Experimental results on ExtCrime dataset

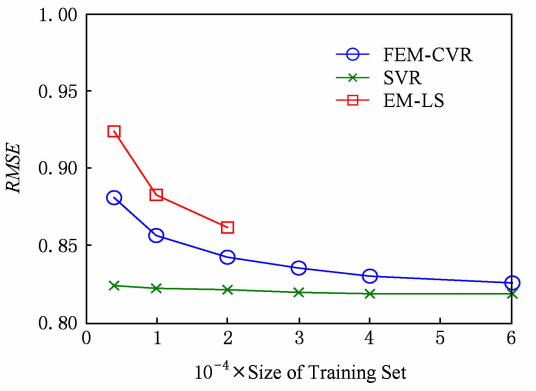
图 6 ExtCrime 数据集实验结果



(a) CPU time on training set of different sizes



(b) Number of SV on training set of different sizes



(c) *RMSE* on training set of different sizes

Fig. 7 Experimental results on ExtWine dataset

图 7 ExtWine 数据集实验结果

Table 3 The Comparison of Experimental Results for Different Parameters

表 3 不同参数的实验结果比较

$\bar{\varepsilon}$	Census Income				Census House			
	RMSE	Training Time/s	Testing Time/s	$\lvert CS\rvert$	*RMSE*	Training Time/s	Testing Time/s	$\lvert CS\rvert$
1E−8	0.132	103.690	20.648	1978	0.019	91.272	18.106	1822
1E−7	0.173	41.246	19.018	1719	0.022	34.012	16.128	1150
1E−6	0.178	29.952	18.944	1512	0.023	10.008	14.272	850
1E−5	0.178	10.828	16.312	951	0.025	3.128	13.096	664
1E−4	0.182	3.103	9.850	396	0.048	1.032	8.786	380
1E−3	0.187	1.376	6.572	144	0.050	0.744	6.304	121
1E−2	0.202	1.192	3.787	75	0.052	0.736	3.763	64
1E−1	0.494	0.752	1.250	31	0.073	0.728	1.248	40

总之,图 3～7 表明:样本容量较少时,FEM-CVR 由于采用了最小包含球$(1+\varepsilon)$的近似算法及采样子集概率加速方法,使得核心集和支持向量更少,因此,预测误差要稍大于 SVR,但是随着样本容量的增大,其误差逐渐接近 SVR.从而验证了 FEM-CVR 更适合大规模数据环境中的属性效应控制.与之不同的是,EM-LS 由于使用了矩阵相乘和求逆,时间复杂度和空间复杂度都超出$O(N^3)$,不但耗时且极易造成内存溢出,所以无法处理大规模数据的属性效应控制.

另外,表 3 给出了 FEM-CVR 算法在不同参数设置情况下得到的实验结果.我们不难发现,当$\eta=1000$时,随着$\bar{\varepsilon}$值的不断减小,所需的训练时间不断增加,但学习精度不断增加.根据我们大量的实验,并权衡精度和计算时间,$\bar{\varepsilon}=1E-6$是较为理想的设置.

4 结 论

面向属性效应控制的大规模数据进行建模是数据挖掘领域的一个重大挑战.本文针对该问题提出了新的算法 FEM-CVR.一方面,该算法基于核化的支持向量回归机 SVR 学习框架,这使得该算法具有处理非线性数据的能力;另一方面,算法 FEM-CVR 在求解 QP 问题时,基于最小包含球学习框架,通过核心集对数据进行压缩,从而快速得到了全局最优解.实验表明,在参数$\bar{\varepsilon}$固定的前提下,算法 FEM-CVR 的时间复杂度上限与数据集的大小呈线性关系,这为它高效地应用于大规模数据属性效应控制提供了坚实的保证.

参 考 文 献

[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. New York: ACM, 2008: 560-568

[2] Calders T, Kamiran F, Pechenizkiy M. Building classifiers with independency constraints [C] //Proc of the 9th IEEE Int Conf on Data Mining Workshops. Piscataway, NJ: IEEE, 2009: 13-18

[3] Kamiran F, Calders T. Classifying without discriminating [C] //Proc of the 2nd Int Conf on Computer, Control & Communication (IC4). Piscataway, NJ: IEEE, 2009: 1-6

[4] 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, Belgium: DTAI, 2010: 1-6

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

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

[7] 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

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

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

[10] Ying Wenhao, Xu Min, Wang Shitong. Fast adaptive clustering by synchronization on large scale datasets [J]. Journal of Computer Research and Development, 2014, 51(4): 707-720 (in Chinese)
(应文豪, 许敏, 王士同. 在大规模数据集上进行快速自适应同步聚类[J]. 计算机研究与发展, 2014, 51(4): 707-720)

[11] Xu Min, Wang Shitong, Gu Xin, et al. Support vector regression for large domain adaptation [J]. Journal of Software, 2013, 24(10): 2312-2326 (in Chinese)
(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)

[12] Wang Jun, Wang Shitong, Deng Zhaohong. Fast kernel density estimator based image thresholding algorithm for small target images [J]. Acta Automatica Sinica, 2012, 38(10): 1679-1689 (in Chinese)
(王骏, 王士同, 邓赵红. 面向小目标图像的快速核密度估计图像阈值分割算法[J]. 自动化学报, 2012, 38(10): 1679-1689)

[13] Ding Lizhong, Liao Shizhong. KMA-α: A kernel approximation algorithm for support vector machines [J]. Journal of Computer Research and Development, 2012, 49(4): 746-753 (in Chinese)
(丁立中, 廖士中. KMA-α:一个支持向量机核矩阵的近似计算算法[J]. 计算机研究与发展, 2012, 49(4): 746-753)

[14] Wang Zhen, Shao Yuanhai, Bai Lan, et al. Twin support vector machine for clustering [J]. IEEE Trans on Neural Networks and Learning Systems, 2015, 26(10): 2583-2588

[15] Schölkopf B, Bartlett P, Smola A, et al. Support Vector Regression with Automatic Accuracy Control [M]. Berlin: Springer, 1998: 111-116

[16] Zhang Jingxiang, Wang Shitong. Common-decision-vector based multiple source transfer learning classification and its fast learning method [J]. Acta Electronica Sinica, 2015, 43(7): 1349-1355 (in Chinese)
(张景祥, 王士同. 基于共同决策方向矢量的多源迁移及其快速学习方法[J]. 电子学报, 2015, 43(7): 1349-1355)

[17] Tsang I, Kwok J, Zurada J. Generalized core vector machines [J]. IEEE Trans on Neural Networks, 2006, 17(5): 1126-1139

[18] Tsang I, Kwok J, Cheung P. Core vector machines: Fast SVM training on very large data sets [J]. Journal of Machine Learning Research, 2005, 6(1): 363-392

[19] Deng Zhaohong, Chung Fulai, Wang Shitong. FRSDE: Fast reduced set density estimator using minimal enclosing ball approximation [J]. Pattern Recognition, 2008, 41(4): 1363-1372

[20] Chang C C, Lin C J. LIBSVM: A library for support vector machines [J]. ACM Trans on Intelligent Systems and Technology, 2011, 2(3): 27

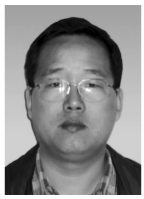
[21] Cortez P, Cerdeira A, Almeida F, et al. Modeling wine preferences by data mining from physicochemical properties [J]. Decision Support Systems, 2009, 47(4): 547-553

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

[23] Musicant D R, Feinberg A. Active set support vector regression [J]. IEEE Trans on Neural Networks, 2004, 15(2): 268-275

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

Liu Jiefang, born in 1982. PhD candidate at the School of Digital Media, Jiangnan University. Member of CCF. His main research interests include pattern recognition, intelligent computation.



Wang Shitong, born in 1964. Professor, PhD supervisor at the School of Digital Media, Jiangnan University. His main research interest include artificial intelligence, pattern recognition and bioinformatics.

Wang Jun, born in 1978. PhD, associate professor, master supervisor at the School of Digital Media, Jiangnan University. Senior member of CCF. His main research interests include pattern recognition, data mining, and digital image processing.

Deng Zhaohong, born in 1981. PhD, professor, master supervisor at the School of Digital Media, Jiangnan University. Senior member of CCF. His main research interests include fuzzy modeling and intelligent computation.