截断误差的光滑型支持向量顺序回归

何海江

(长沙学院计算机科学与技术系 长沙 410003)

【摘要】支持向量顺序回归算法已成功应用于解决顺序回归问题，但其易受训练样本中野点的干扰。为此，提出一种截断误差的光滑型支持向量顺序回归(TLS-SVOR)算法。学习顺序回归模型时，将错划样本形成的误差s限制在范围u内。TLS-SVOR首先用包含参数u的分段多项式近似s；再引入光滑型支持向量机分类算法的思路，将优化目标转变为二次连续可微的无约束问题，从而由牛顿法直接求得唯一的决策超平面。采用两阶段的均匀设计方法确定TLS-SVOR的最优参数。实验结果表明，相比其他顺序回归算法，TLS-SVOR在多个数据集能获得更高的精度。

关 键 词 顺序回归; 野点; 分段多项式; 支持向量机; 截断误差
中图分类号 TP391 文献标志码 A doi:10.3969/j.issn.1001-0548.2014.01.022

Truncated Loss Smooth Support Vector Ordinal Regression

HE Hai-jiang

(Department of Computer Science and Technology, Changsha University Changsha 410003)

Abstract Support vector ordinal regression (SVOR) has been proven to be the promising algorithm for solving ordinal regression problems. However, its performance tends to be strongly affected by outliers in the training datasets. To remedy this drawback, a truncated loss smooth SVOR (TLS-SVOR) is proposed. While learning ordinal regression models, the loss s of the misranked sample is bounded between 0 and the truncated coefficient u. First, a piecewise polynomial function with parameter u is approximated to s. Then, by applying the strategy of smooth support vector machine for classification, the optimization problem is replaced with an unconstrained function which is twice continuously differentiable. The algorithm employs Newton's method to obtain the unique discriminant hyperplane. The optimal parameter combination of TLS-SVOR is determined by a two-stage uniform designed model selection methodology. The experimental results on benchmark datasets show that TLS-SVOR has advantage in terms of accuracy over other ordinal regression approaches.

Key words ordinal regression; outlier; piecewise polynomial; support vector machine; truncated loss

顺序回归(ordinal regression，OR)算法是机器学习的重要工具，解决介于分类和数值回归之间的问题。给定训练样本集$\{(\boldsymbol{x}_j,y_j)|_{j=1,2,\cdots,n}\}$，特征组合向量$\boldsymbol{x}_j$属于输入空间$X$，标记值$y_j$属于输出空间$Y$，函数$f$实现$X$到$Y$的映射。在分类问题和OR问题中，$Y$是有限个离散元素的集合。不同的是，前者$Y$中元素没有顺序，后者$Y$中元素存在单调的顺序关系。而在数值回归问题中，$Y$是连续的实数。若OR模型的输出空间由$v$个序数$\{r_1,r_2,\cdots,r_v\}$组成，则有$r_v>_*\cdots>_*r_2>_*r_1$，其中$>_*$表示顺序关系，可解释为“比……优先”或“比……更受欢迎”等。有网站将浏览者对文章的评价划分为：鲜花、普通、鸡蛋，用$>_*$表示好评程度有：鲜花$>_*$普通$>_*$鸡蛋。

OR算法已成功应用于图像检索、生物识别[1]和信息检索[2]等领域。OR算法大致可分为3类：

1) 样本转换方法。将OR问题转变为分类问题，再用标准的分类算法求解。RankSVM[2]将序数相异的样本两两组合成一个新样本，OR问题转换为两类别分类问题，再以支持向量机作为分类算法。文献[3]则用代价敏感的支持向量机训练两类别分类模型。SVM-EBC要求错划代价确定且已知，而许多实际应用中，错划代价未知且模糊。文献[4]以决策超平面的间隔代理错划代价，由多目标优化算法求得OR模型。文献[5]也采用了类似的思路。

2) 多阈值方法。令$b_0=-\infty$，$b_v=+\infty$，以$v-1$个阈值$\{b_1,b_2,\cdots,b_{v-1}\}$构造$v$个区间$\{(b_i,b_{i+1})|_{i=0,1,\cdots,v-1}\}$，通过决策函数$\min_{1\leqslant i\leqslant v}\{r_i:f(\boldsymbol{x})<b\}$预测样本$\boldsymbol{x}$的序数。文献[6]从结构风险最小化原则出发，构造了$v-1$个平行的决策超平面作为OR模型，但没有考虑超平面阈值应该满足$b_1<b_2<\cdots<b_{v-1}$。SVOR-EXC和SVOR-IMC[7]

收稿日期：2012－09－24; 修回日期：2013－02－26
基金项目：国家自然科学基金(61100139)
作者简介：何海江(1970－)，男，副教授，主要从事机器学习、数据挖掘方面的研究.

将该限制纳入优化目标后，取得了很好的效果。ISVOR[8]则结合了两个文献的优化目标。NNRank[1]用神经网络求样本序数分别等于各序数的v个概率，与其他方法相比，加快了预测速度。相对于样本转换方法，多阈值方法更受到青睐，MINLIP[9]可归于此类。

3) 综合性方法。利用机器学习领域已有的研究成果，如聚类、判别分析等，结合顺序回归的特性，求解问题模型。BQSVR[10]由K-均值聚类寻找各序数的样本代表，再在代表性样本上学习OR模型，显著降低了训练时间复杂度。KDLOR[11]定义了相同序数样本的方差和相异序数样本的方差，组合核判别分析技术和结构风险最小化原理，不仅降低了优化问题规模，精度也不输于其他算法。

除了开发提高精度、降低时间复杂度的算法外，不少研究者还讨论了顺序回归的其他方面。文献[12]提出两种OR模型的误差上界。文献[13]设计了评估OR精度的新测度。值得注意的是，排序学习亦可构造OR模型[14]，只是它主要考虑两个样本之间的顺序关系，而非单一样本的序数。

和网站文章的评价一样，股票或债券评级、学生成绩、歌曲或电影打等级分类似应用中，样本的标记总是由人来确定的，自然难以避免误差。误差过大的样本，也就是野点的存在，使得顺序回归的决策函数偏离。因此，本文提出一种截断误差的光滑型支持向量顺序回归(TLS-SVOR)算法。首先，每个错划样本的误差s都限制在范围u内，用二次光滑的分段多项式来近似s；接着，引入用于解决分类问题的多项式光滑型支持向量机[15]的思路，将SVOR-IMC[7]的误差项由一次变更为二次，并转换优化目标为二次连续可微的无约束问题；最后，由牛顿法直接求无约束问题的唯一最优解。截断误差的思路已在分类问题文献[16]中出现，在OR问题中，则未见报道。本文和已有思路的另一个显著不同点是，u在优化过程中不固定，可作为OR模型的参数。

1 截断误差的光滑型支持向量顺序回归

若有v个序数，序数为r_i的样本有n_i个，$n=n_1+n_2+\cdots+n_v$是训练集的样本总数。令$\boldsymbol{x}_{k,j}$是序数为r_k的样本集中第j个样本，k=1,2,⋯,v，j=1,2,⋯,n_k；$\boldsymbol{b}=(b_1, b_2,\cdots, b_{v-1})$是构造$v$个区间的阈值向量。

1.1 通用核的OR问题

假设$\boldsymbol{A}\in \mathrm{R}^{n\times m}$，$\boldsymbol{B}\in \mathrm{R}^{m\times o}$，则通用核$K(\boldsymbol{A},\boldsymbol{B})$将$\mathrm{R}^{n\times m}\times\mathrm{R}^{m\times o}$映射为$\mathrm{R}^{n\times o}$[17]。依据通用核的定义，参考SVOR-IMC[7]，OR问题的优化目标可为问题1：

$$\min_{\boldsymbol{w},\boldsymbol{b},\xi,\varepsilon}\frac{1}{2C}(\boldsymbol{w}^2+\boldsymbol{b}^2)+\sum_{i=1}^{v-1}\Big(\sum_{k=1}^{i}\sum_{j=1}^{n_k}\xi_{i,k,j}^2+\sum_{k=i+1}^{v}\sum_{j=1}^{n_k}\varepsilon_{i,k,j}^2\Big) \quad (1)$$

使得：

$$\begin{gathered}K(\boldsymbol{x}_{k,j}^{\mathrm{T}},\boldsymbol{H})\times\boldsymbol{w}-b_i\leqslant -1+\xi_{i,k,j}\\ \xi_{i,k,j}\geqslant 0,\ k=1,2,\cdots,i,\ j=1,2,\cdots,n_k\\ K(\boldsymbol{x}_{k,j}^{\mathrm{T}},\boldsymbol{H})*\boldsymbol{w}-b_i\geqslant 1-\varepsilon_{i,k,j}\\ \varepsilon_{i,k,j}\geqslant 0,\ k=i+1,2,\cdots,v,\ j=1,2,\cdots,n_k;\ i=1,2,\cdots,v-1\end{gathered} \quad (2)$$

式(1)中，C是模型复杂度和经验风险间的平衡因子；$\boldsymbol{w}$是OR模型的权；$\boldsymbol{x}^{\mathrm{T}}$是$\boldsymbol{x}$的转置向量；$\xi$和$\varepsilon$是误差项，$\xi_{i,k,j}$和$\varepsilon_{i,k,j}$表示序数为$r_k$的第$j$个样本相对阈值$b_i$形成的误差。最坏的情况，每个样本有$v-1$个误差项，问题1共包含$n\times(v-1)$个相关的误差项。式(2)的核矩阵 $\boldsymbol{H}=(\boldsymbol{x}_{1,1},\cdots,\boldsymbol{x}_{1,n_1},\cdots,\boldsymbol{x}_{2,n_2},\cdots,\boldsymbol{x}_{v,n_v})\in \mathrm{R}^{m\times n}$。文章采用通用高斯核[18]完成所有实验，$\gamma$是高斯核宽度。

1.2 减弱野点的影响

在问题1中，野点降低了OR模型的泛化性能。为减小野点的影响，将误差s(ξ或ε)限制在范围u内(u>0)，也就是0⩽s⩽u，u被称为截断系数。有正号函数$(x)_+=\max(0,x)$，定义在截断系数u下的误差s的截断误差为$q(s,u)=(s)_+-(s-u)_+$。引入多项式光滑支持向量机[15]的思路，合并问题1的式(1)和式(2)，构造新的无约束问题2：

$$\begin{aligned}\min_{\boldsymbol{w},\boldsymbol{b}}\lambda_u(\boldsymbol{w},\boldsymbol{b})=\min_{\boldsymbol{w},\boldsymbol{b}}\Bigg\{&\frac{1}{2C}(\boldsymbol{w}^2+\boldsymbol{b}^2)+\\ &\sum_{i=1}^{v-1}\Bigg[\sum_{k=1}^{i}\sum_{j=1}^{n_k}q\Big[(1+K(\boldsymbol{x}_{k,j}^{\mathrm{T}},\boldsymbol{H})\times\boldsymbol{w}-b_i),u\Big]^2+\\ &\sum_{k=i+1}^{v}\sum_{j=1}^{n_k}q\Big[(1-K(\boldsymbol{x}_{k,j}^{\mathrm{T}},\boldsymbol{H})\times\boldsymbol{w}+b_i),u\Big]^2\Bigg]\Bigg\}\end{aligned} \quad (3)$$

因$\boldsymbol{w}$不可能为零向量，且式(3)各个子项皆为平方项，显然$\lambda_u(\boldsymbol{w},\boldsymbol{b})$是严格的凸函数，问题2有唯一最优解。$q(s,u)$不可微，为了直接用牛顿法求解问题2，借鉴多项式光滑函数[15]，定义分段多项式$p(s,u,\eta)$近似$q(s,u)$：

$$p(s,u,\eta)=\begin{cases}u & s\geqslant u+\eta\\ \dfrac{(s-u)^4}{16\eta^3}-\dfrac{3(s-u)^2}{8\eta}+\dfrac{s-u}{2}+u-\dfrac{3\eta}{16} & u-\eta<s<u+\eta\\ s & \eta\leqslant s\leqslant u-\eta\\ -\dfrac{1}{16\eta^3}(s+\eta)^3(s-3\eta) & -\eta<s<\eta\\ 0 & s\leqslant -\eta\end{cases} \quad (4)$$

式中，$\eta\in(0,1)$，$u>2\eta$。η越小，$p(s,u,\eta)$越逼近$q(s,u)$。

将式(4)代入式(3)，新OR算法TLS-SVOR的优化目标为问题3：

$$\min_{\boldsymbol{w},\boldsymbol{b}}\psi_{u,\eta}(\boldsymbol{w},\boldsymbol{b})=\min_{\boldsymbol{w},\boldsymbol{b}}\left\{\frac{1}{2C}(\boldsymbol{w}^2+\boldsymbol{b}^2)+\sum_{i=1}^{v-1}\left[\sum_{k=1}^{i}\sum_{j=1}^{n_k}p\left[(1+K(\boldsymbol{x}_{k,j}^{\mathrm{T}},\boldsymbol{H})\times\boldsymbol{w}-b_i),u,\eta\right]^2+\sum_{k=i+1}^{v}\sum_{j=1}^{n_k}p\left[(1-K(\boldsymbol{x}_{k,j}^{\mathrm{T}},\boldsymbol{H})\times\boldsymbol{w}+b_i),u,\eta\right]^2\right]\right\}\quad(5)$$

问题3同样有唯一最优解，且 $\psi_{u,\eta}(\boldsymbol{w},\boldsymbol{b})$二次可微，可用牛顿法直接求$(\boldsymbol{w},\boldsymbol{b})$。令$b_v=+\infty$，与式(5)对应的决策函数为：

$$f(\boldsymbol{x})=\min\nolimits_{1\leqslant i\leqslant v}\{r_i:K(\boldsymbol{x}^{\mathrm{T}},\boldsymbol{H})\times\boldsymbol{w}-b_i<0\}\quad(6)$$

1.3　分段多项式函数的性质

与逼近正号函数相似，逼近$q(s,u)$的任意偶数阶分段多项式函数或许都存在，选取形式简单的四阶分段多项式$p(s,u,\eta)$。

定理 1　$p(s,u,\eta)$具有二阶光滑性。

证明：由式(4)容易验证。

定理 2　$s\leqslant u-\eta$时，$p(s,u,\eta)\geqslant q(s,u)$；$s>u-\eta$时，$p(s,u,\eta)\leqslant q(s,u)$。

证明：从文献[15]可知，$s\leqslant u-\eta$时，$q(s,u)=(s)_+$，有$p(s,u,\eta)\geqslant q(s,u)$。

再证明第二个不等式，Δ为导数符号。

1) $s\in(u-\eta,u)$时，设$\mathrm{fa}(s)=p(s,u,\eta)-q(s,u)$。令$\mathrm{fb}(s)=\Delta\mathrm{fa}(s)=(s-u)^3/(4\eta^3)-3(s-u)/(4\eta)-1/2$，有$\Delta\mathrm{fb}(s)=3(s-u)^2/(4\eta^3)-3/(4\eta)<0$。故$\mathrm{fb}(s)$严格单调递减，$-1/2\leqslant\mathrm{fb}(s)<0$。因为$\Delta\mathrm{fa}(s)<0$，$\mathrm{fa}(s)$严格单调递减，$-3\eta/16\leqslant\mathrm{fa}(s)<0$，因此$p(s,u,\eta)<q(s,u)$。

2) $s\in[u,u+\eta]$ 时，类似步骤1)的方法可证明不等式仍成立。

3) $s\geqslant u+\eta$时，$p(s,u,\eta)=q(s,u)=u$。

综合步骤1)～步骤3)可知第二个不等式成立。

定理 3　$s\leqslant u-\eta$时，$p(s,u,\eta)^2-q(s,u)^2<0.052\,6\eta^2$；$s>u-\eta$时，$q(s,u)^2-p(s,u,\eta)^2<0.375u\eta$。

证明方法与定理2的相类似。

1.4　TLS-SVOR的收敛特性

接下来可以证明，随着η趋于零，问题3的最优解将收敛于问题2的最优解。

定理 4　假定两个问题最优解的欧几里得差为δ，则有$\delta<n(v-1)\times0.427\,6\,u\eta$。

证明：令$\boldsymbol{\theta}=(\boldsymbol{w},\boldsymbol{b})$，若$\boldsymbol{\theta}^*$是问题2的最优解，$\boldsymbol{\theta}^\#$是问题3的最优解，则由一阶最优性条件及$\lambda_u(\boldsymbol{w},\boldsymbol{b})$、$\psi_{u,\eta}(\boldsymbol{w},\boldsymbol{b})$的强凸性可知：

$$\lambda_u(\boldsymbol{\theta}^\#)-\lambda_u(\boldsymbol{\theta}^*)\geqslant\Delta\lambda_u(\boldsymbol{\theta}^*)(\boldsymbol{\theta}^\#-\boldsymbol{\theta}^*)+0.5(\boldsymbol{\theta}^\#-\boldsymbol{\theta}^*)^2=0.5(\boldsymbol{\theta}^\#-\boldsymbol{\theta}^*)^2\psi_{u,\eta}(\boldsymbol{\theta}^*)-\psi_{u,\eta}(\boldsymbol{\theta}^\#)\geqslant\Delta\psi_{u,\eta}(\boldsymbol{\theta}^\#)(\boldsymbol{\theta}^*-\boldsymbol{\theta}^\#)+0.5\times(\boldsymbol{\theta}^*-\boldsymbol{\theta}^\#)^2=0.5\times(\boldsymbol{\theta}^*-\boldsymbol{\theta}^\#)^2$$

两式相加，得$(\boldsymbol{\theta}^\#-\boldsymbol{\theta}^*)^2\leqslant\psi_{u,\eta}(\boldsymbol{\theta}^*)-\lambda_u(\boldsymbol{\theta}^*)+(\lambda_u(\boldsymbol{\theta}^\#)-\psi_{u,\eta}(\boldsymbol{\theta}^\#))$。最坏情况下：$\lambda_u(\boldsymbol{\theta}^*)$的误差项都小于等于$u-\eta$，依定理2和定理3，此时$0\leqslant\psi_{u,\eta}(\boldsymbol{\theta}^*)-\lambda_u(\boldsymbol{\theta}^*)\leqslant n(v-1)\times0.052\,6\eta^2$；$\lambda_u(\boldsymbol{\theta}^\#)$的误差项都大于$u-\eta$，同理$0\leqslant\lambda_u(\boldsymbol{\theta}^\#)-\psi_{u,\eta}(\boldsymbol{\theta}^\#)\leqslant n(v-1)\times0.375u\eta$。因此$(\boldsymbol{\theta}^*-\boldsymbol{\theta}^\#)^2\leqslant n(v-1)(0.052\,6\eta^2+0.375u\eta)<n(v-1)\times0.427\,6u\eta$。从而结论得证。

1.5　支持简约核的TLS-SVOR

样本个数n很大时，核矩阵计算非常耗时。为此，引入简约核[17]。从训练集随机采集$h<<n$个样本，组成简约矩阵$\overrightarrow{\boldsymbol{H}}\in\mathrm{R}^{m\times h}$，并保证新样本集的各个序数样本比例与原训练集相同。支持简约核的TLS-SVOR，优化目标为问题4，将式(5)的$\boldsymbol{H}$替换为简约矩阵即可；其决策函数与式(5)相似，同样将$\boldsymbol{H}$替换为简约矩阵即可。定义简约比例$\mathrm{Ratio}=n/h$，一般来说，Ratio越大，简约核TLS-SVOR的训练速度越快，但是OR精度越低。

问题3和问题4的优化目标都是二次连续可微的无约束最小化函数，可用牛顿法求解。

2　两阶段的最优参数选择方法

分类模型的参数选择方法得到较多的研究，大多数OR模型直接利用这些成果。SVOR-IMC[7]、SVOR-EXC[7]和SVOR-EBC[4]采用完全扫描方法，将参数均匀划分，每个参数组合都试一次。为了减少训练时间，采用两阶段的均匀设计实验方法[18]确定最优参数。算法1是具体的实现步骤。在问题3和问题4中，截断系数u既可固定，也可设定为待优化的参数。固定u时，算法1简称为两参数方法，选择C和γ；u可变时，算法1简称为三参数方法，选择u、C和γ。在实现算法时，先确定参数的范围，C从$C_{\min}$到$C_{\max}$，u从$u_{\min}$到$u_{\max}$，高斯核宽度γ的范围参考文献[18]。

算法 1　两阶段最优参数选择算法

输入数据：训练数据集、$C_{\min}$和$C_{\max}$、$u_{\min}$和$u_{\max}$、交叉验证折数nFold。

1) 计算$\gamma_{\min}$和$\gamma_{\max}$；从http://www.math.hkbu.edu.hk/UniformDesign下载均匀设计模式UD表；执行nFold折训练数据集拆分，每次都按相同样本数比例拆成学习集和验证集，并保证拆分后分属各序数的样本数比例与原训练集一致。对每一折的学习集与

验证集，执行步骤2)和步骤3)。

2) 依UD表及$C_{\min}$、$C_{\max}$、$u_{\min}$、$u_{\max}$、$\gamma_{\min}$和$\gamma_{\max}$，计算出13个参数组合，以每个参数组合在学习集上训练OR模型，获得验证集上的OR精度。最优参数记为(C^*,γ^*,u^*)，两参数方法则为(C^*,γ^*)。

3) 第二阶段，计算新的参数范围。新范围以第一阶段最优参数组为中心点，数值区间则减半。以C为例，令$C_{\text{span}}=\dfrac{0.5(C_{\max}-C_{\min})}{9-1}$，则$C_{\min}^{\text{new}}=C^*-4C_{\text{span}}$，$C_{\max}^{\text{new}}=C^*+4C_{\text{span}}$。

u和γ的新范围计算方法相同。依UD表及新参数范围计算出9个新参数组合。这一阶段得到的最优参数组记为$(C^{**},\gamma^{**},u^{**})$，两参数方法则为$(C^{**},\gamma^{**})$。

4) 获得nFold个第二阶段的最优参数组，分别计算它们在nFold个验证集上的平均OR精度，以最高精度对应的参数组确定为训练数据集上的最优参数。

值得注意的是，步骤3)的9个参数组中，有一个(中心点)在步骤2)已经试过，所以两阶段共需13+9−1=21次训练。

3 实验结果与分析

使用平均绝对误差MAE和平均错划误差MZE[7]评估OR精度。TLS-SVOR的优化算法中，最优解误差$\delta=10^{-3}$。使用到算法1的所有实验，$u_{\min}=1.5$，$u_{\max}=6$，$C_{\min}=10^{-3}$，$C_{\max}=10^{3}$；当用简约核时，$C_{\min}=10^{-2}$，$C_{\max}=10^{4}$。

3.1 人工数据集上测试野点的影响

先生成一个人工数据集，再在数据集上测试野点如何影响TLS-SVOR的精度。人工数据集由300个样本组成，样本$\boldsymbol{x}=(x_1,x_2,\cdots,x_5)$有5个属性，每个属性都是[0,1]间的随机数。按以下规则确定$\boldsymbol{x}$的标记y：

$$y=i \quad \Leftrightarrow \quad f(\boldsymbol{x})+\varepsilon\in[b_i,b_{i+1}) \tag{7}$$

$$f(\boldsymbol{x})=\mathrm{e}^{-\frac{1}{5}\sum_{i=1}^{5}(x_i-0.5)^2} \tag{8}$$

式(7)中，ε是期望值为0，方差为0.026的高斯型白噪声；$\boldsymbol{b}=(-\infty,0.905,0.925,0.945,+\infty)$是预先定义的超平面阈值向量。在300个样本中，序数为1、2、3、4的样本个数分别等于65、86、71、78。将人工数据集拆成3份$S1$、$S2$、$S3$，使得样本数相同，序数均匀分布。

为考察野点的影响，检验截断系数的作用，实现了一个和TLS-SVOR相似的算法np-SVOR，其优化目标与问题3相似，但不限制误差范围，即$u=+\infty$。使用算法1确定np-SVOR和TLS-SVOR的最优参数。TLS-SVOR的u固定为1.5。如图1和图2所示是np-SVOR和TLS-SVOR在人工数据集的表现，实验结果是三折的平均值。在第一折中，学习集$S1$、验证集$S2$、测试集$S3$；第二三折，$S2$和$S3$分别充当学习集。两图中，野点比例分别为0%、4%、8%、12%。以4%为例，将学习集的2个序数为1的样本错标为4，同时将2个序数为4的样本错标为1。

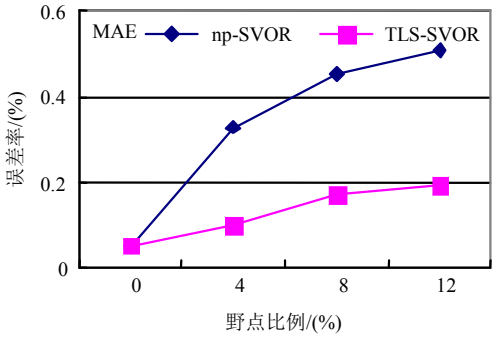

图1 人工数据集上比较MAE

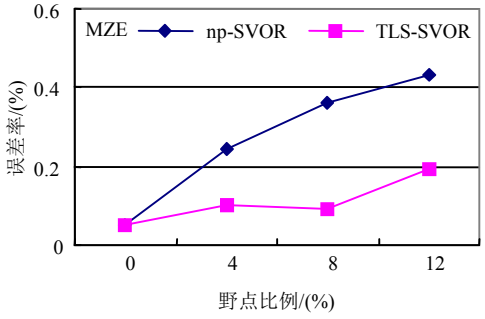

图2 人工数据集上比较MZE

从图中可看出，野点的出现，导致OR模型的精度明显下降。随着野点的增多，np-SVOR的精度急剧下降；相比而言，TLS-SVOR受到的影响小得多。另外，无论是MAE，还是MZE，TLS-SVOR都显著优于np-SVOR。由此可证明，截断误差在很大程度上减弱了野点的干扰。理论上，算法给野点赋以很大的误差，训练过程中，野点对应序数所属的决策超平面会尽力靠拢野点，这会引发连锁反应，其他样本的误差跟着变化；而TLS-SVOR在截断误差后，减弱了这种影响，自然更接近理想的学习器。

3.2 TLS-SVOR和其他算法的比较

文献[7]提出的8个数据集被其他研究者广泛采用，已经成为OR社区的基准数据集。它们的序数皆为10，每个数据集都被划分20次，每次按相同样本数比例划分成训练集和测试集。SVOR-EXC[7]、SVOR-IMC[7]、SVM-EBC[3]和KDLOR[11]都在8个基准数据集上测试算法的OR精度。本节将TLS-SVOR

和这4个算法对比。公平起见，实验结果同样取20次划分的平均值。当用算法1选择参数时，每次划分都使用五折交叉验证，每一折，学习集和验证集的样本比例同样为3:2，再求解优化目标。三参数方法选择TLS-SVOR的最优参数，截断系数的范围：$u_{\min}$=1.5，$u_{\max}$=6，该范围未经挑选。表1是5种算法的MAE比较，表2是4种算法的MZE比较。SVOR-EXC和SVOR-IMC的实验数据来自文献[7]；SVM-EBC的来自文献[3](缺MZE的数据)；KDLOR的来自文献[11]。几种算法在同一数据集表现最好的用粗体字标注，包括平均误差和20个误差的标准方差。

表1　5种算法比较MAE

数据集	Pyrimidines	MachineCPU	Boston	Abalone	Bank	Computer	California	Census
TLS-SVOR	**0.954±0.234**	**0.677±0.128**	0.742±0.087	**0.529±0.012**	1.401±0.011	0.613±0.012	0.997±0.010	1.248±0.009
KDLOR	1.100±0.100	0.690±0.015	**0.700±0.035**	1.400±0.050	1.450±0.020	0.601±0.025	**0.907±0.004**	1.213±0.003
SVOR-EXC	1.331±0.193	0.986±0.127	0.773±0.049	1.391±0.021	1.515±0.017	0.602±0.009	1.068±0.005	1.270±0.007
SVOR-IMC	1.294±0.204	0.990±0.115	0.747±0.049	1.361±0.013	**1.393±0.011**	0.596±0.008	1.008±0.005	1.205±0.007
SVM-EBC	1.304±0.040	0.842±0.022	0.732±0.013	1.383±0.004	1.404±0.002	**0.565±0.002**	0.940±0.001	**1.143±0.002**

表2　4种算法比较MZE

数据集	Pyrimidines	MachineCPU	Boston	Abalone	Bank	Computer	California	Census
TLS-SVOR	**0.671±0.088**	**0.408±0.071**	0.580±0.041	**0.439±0.010**	0.757±0.007	0.475±0.009	**0.634±0.003**	0.710±0.005
KDLOR	0.739±0.050	0.480±0.010	**0.560±0.020**	0.740±0.020	0.745±0.003	0.472±0.020	0.643±0.005	0.711±0.020
SVOR-EXC	0.752±0.063	0.661±0.056	0.569±0.025	0.736±0.011	**0.744±0.005**	**0.462±0.005**	0.640±0.003	**0.699±0.002**
SVOR-IMC	0.719±0.066	0.655±0.045	0.561±0.026	0.732±0.007	0.751±0.005	0.473±0.005	0.639±0.003	0.705±0.002

Abalone、Bank、Computer、California和Census这5个数据集的训练集样本数分别为1 000、3 000、4 000、5 000和6 000。考虑到训练时间的问题，测试这5个数据集时，使用简约核的TLS-SVOR，简约因子Ratio分别为2、6、8、10和12，也就是简约核样本数h都为500。从表1和表2可看出，无论以MAE还是MZE评估算法时，在Pyrimidines、MachineCPU和Abalone上，TLS-SVOR都比其他算法表现好。特别是在Abalone上，TLS-SVOR展现巨大的优势，MZE数据比其他算法下降40%以上，MAE数据甚至不到其他的一半。很可能Abalone中存在较大比例的野点，文献[11]也提到这一点，只是KDLOR无法应对这种情况。在California上，TLS-SVOR的MZE值最小，而其MAE值并不理想。总体来看，TLS-SVOR降低MZE的能力要胜过降低MAE的能力。TLS-SVOR在样本数少的数据集表现好，因为同一序数的样本少，一旦出现一个野点，就会严重干扰OR模型。TLS-SVOR在样本数多的数据集表现差，有两方面的原因：使用简约核，导致TLS-SVOR的OR精度下降；样本多，少数野点形成的误差被其他样本稀释。还有一点特别需要指出，在选择参数时，TLS-SVOR采用五折交叉验证，训练21次；SVOR-EXC和SVOR-IMC同样五折交叉验证，需要训练130次；SVM-EBC五折交叉验证，训练次数不会少于21次；而KDLOR采用十折交叉验证，但训练次数未知。

4　结 论

基于统计学习的顺序回归算法，训练时易受野点的干扰。为此，提出一种截断误差的光滑型支持向量顺序回归算法。在TLS-SVOR的训练过程中，限制错划样本的误差，使其不超过截断系数。在人工数据集上的实验证实，截断误差后，顺序回归模型的精度大为提高。本文还提出一种新思路，在OR模型加入截断系数，采用两阶段的均匀设计方法选择最优参数。在8个基准数据集的测试结果表明，与已有的一些算法相比，无论是比较MAE，还是比较MZE，新算法都有明显的优势。

本文的研究工作得到了长沙学院科研基金(CDJJ-11010208)的大力支持，在此表示感谢！

参 考 文 献

[1] CHENG Jian-lin, WANG Zheng, POLLASTRI G. A neural network approach to ordinal regression[C]//Proceedings of the International Joint Conference on Neural Networks. CA: IEEE Press, 2008: 1279-1284.

[2] HERBRICH R, GRAEPEL T, OBERMAYER T. Large margin rank boundaries for ordinal regression[C]//Advances in Large Margin Classifiers. Cambridge, MA: MIT Press, 2000: 115-132.

[3] LI Ling, LIN H T. Ordinal regression by extended binary classification[C]//Advances in NIPS 19. Cambridge, MA: MIT Press, 2007: 865-872.

[4] EMILIO C, BELEN M B. Maximizing upgrading and

downgrading margins for ordinal regression[J]. Mathematical Methods of Operations Research, 2011, 74(3): 381-407.

[5] DOBRSKA M, WANG Hui, BLACKBURN W. Ordinal regression with continuous pairwise preferences[J]. International Journal of Machine Learning and Cybernetics, 2012, 3(1): 59-70.

[6] SHASHUA A, LEVIN A. Ranking with large margin principle: Two approaches[C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2002: 937-944.

[7] CHU Wei, KEERTHI S S. Support vector ordinal regression[J]. Neural Computation, 2007, 19(3): 792-815.

[8] SUN Bin-yu, ZHANG Xiao-ming, LI Wen-bo. An improved ordinal regression approach with sum-of-margin principle[C]//Proceedings of Sixth ICNC. CA: IEEE Press, 2010: 853-857.

[9] BELLE V V, PELCKMANS K, SUYKENS J, et al. Learning transformation models for ranking and survival analysis[J]. Journal of Machine Learning Research, 2011, 12(3): 819-862.

[10] ZHAO Bin, WANG Fei, ZHANG Chang-shui. Block-quantized support vector ordinal regression[J]. IEEE Transactions on Neural Networks, 2009, 20(5): 882-890.

[11] SUN Bin-yu, LI Jiu-yong, WU D D, et al. Kernel discriminant learning for regression[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(6): 906-910.

[12] YANG Zhi-xia, TIAN Yin-gjie, DENG Nai-yang. Leave-one-out bounds for support vector ordinal regression machine[J]. Neural Computing & Applications, 2009,18(7): 731-748.

[13] CARDOSO J S, SOUSA R. Measuring the performance of ordinal classification[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2011, 25(8): 1173-1195.

[14] SAWADE C, BICKEL S, OERTZEN T von, et al. Active evaluation of ranking functions based on graded relevance[C]// Proceedings of the 2012 ECML/PKDD. Heidelberg: Springer- Verlag.

[15] 袁玉波, 严杰, 徐成贤. 多项式光滑的支撑向量机[J]. 计算机学报, 2005, 28(1): 9-17.
YUAN Yu-bo, YAN Jie, XU Cheng-xian. Polynomial smooth support vector machine(PSSVM)[J]. Chinese Journal of Computers, 2005, 28(1): 9-17.

[16] WU Yi-chao, LIU Yu-feng. Robust truncated hinge loss support vector machines[J]. Journal of the American Statistical Association, 2007,102(479): 974-983.

[17] LEE Y J, HUANG Su-yun. Reduced support vector machines: a statistical theory[J]. IEEE Transactions on Neural Networks, 2007, 18(1):1-13.

[18] HUANG Chien-ming, LEE Y J, LIN D K J, et al. Model selection for support vector machines via uniform design[J]. Computational Statistics and Data Analysis, 2007, 52(1): 335-346.

编 辑 张 俊