L2 损失大规模线性非平行支持向量顺序回归模型

石 勇 [1,2,3,4] 李佩佳 [2,4] 汪华东 [3,4]

摘 要 顺序回归是一种标签具有序信息的多分类问题, 广泛存在于信息检索、推荐系统、情感分析等领域. 随着互联网、移动通信等技术的发展, 面对大量具有大规模、高维、稀疏等特征的数据, 传统的顺序回归算法往往表现不足. 非平行支持向量顺序回归模型具有适应性强, 在性能上优于其他基于 SVM 的方法等优点, 该文在此模型基础上提出基于 L2 损失的大规模线性非平行支持向量顺序回归模型, 其中线性模型的设计可处理大规模数据, 基于 L2 的损失可使标签偏离较大的样本得到更大惩罚. 此外, 该文从模型的两种不同角度分别设计了信赖域牛顿算法和坐标下降算法求解该线性模型, 并比较了两种算法在性能上的差异. 为验证模型的有效性, 该文在大量数据集上对提出的模型及算法进行了分析, 结果表明, 该文提出的模型表现最优, 尤其采用坐标下降算法求解的该模型在数据集上获得了最好的测试性能.

关键词 顺序回归, 支持向量机, 信赖域牛顿算法, 对偶坐标下降算法

引用格式 石勇, 李佩佳, 汪华东. L2 损失大规模线性非平行支持向量顺序回归模型. 自动化学报, 2019, **45**(3): 505−517

DOI 10.16383/j.aas.2018.c170438

L2-loss Large-scale Linear Nonparallel Support Vector Ordinal Regression

SHI Yong[1,2,3,4] LI Pei-Jia[2,4] WANG Hua-Dong[3,4]

Abstract Ordinal regression, where the labels of the samples exhibit a natural ordering, is a kind of multi-classification problem. It has found wide applications in information retrieval, recommendation systems, and sentiment analysis. With the development of internet and mobile communication technology, traditional ordinal regression models often underperform when facing numerous large scale, high dimensional and sparse data. However, the nonparallel support vector ordinal regression model shows its advantages with strong adaptability and better performance compared with other SVM-based models. Based on this model, this paper presents a new L2-loss linear nonparallel support vector ordinal regression model, whose linear model could deal with large-scale problems and whose L2-loss could give a great punishment to the sample that deviates from the true label. Besides, two algorithms: trust region Newton method and the dual coordinate descent method (DCD) are developed in terms of different perspectives of the model and their performances are compared. To verify the effectiveness of the model, experiments are conducted on numerous datasets and the results show that the proposed model, especially the model with the DCD algorithm can achieve the state-of-art performance.

Key words Ordinal regression, SVM, trust region Newton method, dual coordinate descent method

Citation Shi Yong, Li Pei-Jia, Wang Hua-Dong. L2-loss large-scale linear nonparallel support vector ordinal regression. *Acta Automatica Sinica*, 2019, **45**(3): 505−517

顺序回归旨在对具有顺序标签结构的样本进行分类. 近些年随着数据挖掘和机器学习技术的发展, 顺序回归模型在情感分析、产品评论、信用评估、用户画像等领域得到了广泛应用[1−3]. 在这些领域中, 其样本标签包含顺序信息, 不同的错误样本代价往往不同. 如用户画像领域中对年龄的预估, 20 岁的青年用户被错分为 30 岁和 50 岁形成的用户画像有明显差异. 再如在信用评估领域, 一个信用值极低的公司被错分为一般低和较高所影响的决策大相径庭. 因此, 顺序回归问题受到越来越多的重视. 顺序回归在机器学习领域中介于分类问题和回归问题之间. 与分类问题不同, 顺序回归问题的标签集合具有顺序结构而不仅仅是一个多类别集合. 再者, 与回归问题不同, 顺序回归问题的标签不具有度量信息.

顺序回归领域拥有多种模型. Gutiérrez 等[4] 对顺序回归模型做了一个比较全面的综述性调研并且对比了当前 16 种流行的顺序回归模型的性能. 实

收稿日期 2017-08-01 录用日期 2017-10-30
Manuscript received August 1, 2017; accepted October 30, 2017
国家自然科学基金 (71110107026, 71331005, 91546201) 资助
Supported by National Natural Science Foundation of China (71110107026, 71331005, 91546201)
本文责任编委 何海波
Recommended by Associate Editor HE Hai-Bo
1. 中国科学院虚拟经济与数据科学研究中心 北京 100190 2. 中国科学院大学计算机与控制学院 北京 101408 3. 中国科学院大学数学科学学院 北京 100049 4. 中国科学院大数据挖掘与知识管理重点实验室 北京 100190
1. Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190 2. School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 101408 3. School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049 4. Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190

验表明, 基于 SVM 的方法[5], 如支持向量顺序回归机 (Support vector ordinal regression, SVOR)[6], 约简的 SVM (Reduction applied to support vector machines, RedSVM)[7] 和基于映射的顺序集成学习 (Ordinal projection-based ensemble learning, OPBE)[8] 在顺序回归数据上表现出良好的性能. 但是, 这些方法均采用非线性核函数把样本点映射到一个高维特征空间. 由于经非线性变换后的数据往往被表示成具有较高维度的向量, 那么在模型由于向量之间进行内积计算造成训练时间成本较高且内存消耗大. 所以基于 SVM 的非线性方法只适合于小规模及低纬度数据. 而在大量实际应用中, 数据往往表现出量多、特征维度多等特点, 如文本分类等. 在文本分类中, 文本通过特征提取和特征值计算的方法, 如词袋特征 (Bag-of-words, BOW), 词频 – 反文档频率 (Term frequency-inverse document frequency, TF-IDF) 等常被表示成高维向量. 所以非线性顺序回归模型无法满足大规模数据的需求. 此外, 研究表明[9] 在大规模数据中, 线性模型和非线性模型在性能上没有较大差异. 所以, 为了减少时间和内存消耗, 文本将提出基于 SVM 的线性顺序回归模型. 文献 [10] 提出了非线性的非平行支持向量回归机模型 (Nonparallel support vector ordinal regression, NPSVOR), 该模型采用 Hinge 损失函数建立. 实验表明在非线性情况下, NPSVOR 优于其他基于 SVM 的方法, 因此本文将研究可处理大规模数据的线性 NPSVOR 模型.

在解决顺序回归问题时一方面需要考虑顺序信息, 另一方面由于对不同分错样本的处理不同, 所以在构建模型时, 对损失项的特殊处理有助于使预测标签与实际标签尽可能接近, 提高模型的性能. 如此, 为了使与真实标签产生较大偏差的样本得到更大的惩罚, 我们在建模中采用 L2 损失 (均方 Hinge 损失) 作为模型损失函数, 旨在最小化真实值与估计值的距离的平方, 使得训练的模型能更好地处理与真实标签间的差异. 与此同时, L2 损失对离群点较敏感, 一个合适的训练模型算法显得如此重要. 针对求解算法, Hsieh 等研究了标准的线性支持向量分类 (Support vector classification, SVC)[11] 和支持向量回归 (Support vector regression, SVR)[12] 求解算法, 并提出了相应处理大规模数据的求解算法如对偶坐标下降算法 (DCD) 和信赖域算法. 虽然 Hsieh 等[11−12] 的模型和算法被广泛应用在文本挖掘中, 但这些模型和算法主要解决了标准的多分类和回归问题, 而在顺序回归模型中还没有得到应用. 所以该文提出能快速求解基于 L2 损失的非平行支持向量回归机的算法.

本文提出一种基于 L2 损失的线性非平行支持向量顺序回归模型. 从目前研究来看, 这是在顺序回归领域中第一个处理大规模问题的相关工作. 此外, 针对该模型, 该文设计了两种求解该模型的算法并比较了两种算法的性能表现.

本文组织结构如下: 第 1 节介绍基于 L1 损失的 NPSVOR 模型. 第 2 节介绍本文提出的基于 L2 损失的线性 NPSVOR (L2-NPSVOR) 模型, 并给出其对偶模型. 第 3 节研究求解 L2-NPSVOR 模型的优化算法, 从原问题和对偶问题两个角度分别给出了信赖域牛顿算法和对偶坐标下降算法. 第 4 节主要介绍数值实验, 将提出的 L2-NPSVOR 模型与其他相关模型进行分析比较, 验证模型的有效性. 最后, 对本文研究工作进行总结.

1 非平行支持向量回归机

在顺序回归问题中, 每个训练样本均由一个特征向量和一个有序标签组成. 假设顺序回归问题有 p 个不同的具有有序结构的类别, 为不失一般性, 我们用连续整数 $1,2,\cdots,p$ 表示其类别, 用 n 表示样本的数量. 则顺序回归样本集可以表示为:

$$\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1,\cdots,n}$$

其中 $\boldsymbol{x}_i\in\mathbf{R}^m$ 是输入向量, $y_i\in\{1,2,\cdots,p\}$ 是 $\boldsymbol{x}_i$ 的标签, 这里整数值的大小关系反应类别之间序关系.

文献 [10] 提出的非平行支持向量顺序回归模型 (NPSVOR), 其可以在原空间上学习多个非平行的超平面, 对数据分布具有更好的适应性. 并在性能上优于其他基于 SVM 的方法. 对于 p 类的顺序回归问题, NPSVOR 针对每个类别 $k\in\{1,\cdots,p\}$ 构建有序三元分解学习一个超平面, 即给定 $k\in\{1,2,\cdots,p\}$, 首先对每个索引 k 建立三个索引集: $\mathcal{L}_k=\{i|y_i<k\}$, $\mathcal{I}_k=\{i|y_i=k\}$ 和 $\mathcal{R}_k=\{i|y_i>k\}$. 其中 y_i 是 $\boldsymbol{x}_i$ 的标签. 然后, 学习一个映射 $f_k(\boldsymbol{x})=\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}+b_k$, 建立如下优化模型

$$\begin{aligned}\min_{\boldsymbol{w}_k}\quad & \tfrac{1}{2}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k+C_1\sum_{i\in\mathcal{I}_k}\max(0,|\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i+b_k|-\varepsilon)+\\ & C_2\sum_{i\notin\mathcal{I}_k}\max(0,1-\hat{y}_i^k(\boldsymbol{w}_k^{\mathrm{T}})\boldsymbol{x}_i+b_k)\end{aligned}\tag{1}$$

其中 $\hat{y}_i^k=\begin{cases}1, & y_i>k\\ -1, & y_i\le k\end{cases}$, $C_1,C_2>0$ 是惩罚系数. 由于模型中采用 L1 损失 (Hinge 损失) 作为损失函数, 这里记该模型 (1) 为 L1-NPSVOR.

在模型式 (1) 中, 第一项为正则项, 第二项和第三项为 L1 损失项, 其中第二项是要求学习的超平面尽可能考虑第 k 类样本 $\{\boldsymbol{x}_i:i\in\mathcal{I}_k\}$, 第三项要求其他类样本离该超平面尽可能远, 且其中标签大

于 k 的样本 $\{\boldsymbol{x}_i : i \in \mathcal{R}_k\}$ 和标签小于 k 的样本 $\{\boldsymbol{x}_i : i \in \mathcal{L}_k\}$ 分别位于该超平面两侧, 以更好地利用标签的有序信息. 值得强调的是, 学习的 p 个子优化模型 (1) 相互独立, 因而可并行学习.

图 1 是非平行支持向量回归机的几何解释. 图 1 中, "□", "◯" 和 "◇" 分别表示类别 1、类别 2 和类别 3. 该类的最优超平面应该离该类样本尽可能地近而离其他类的样本尽可能远. 同时, 第 1 类和第 3 类的样本点应位居该类最优超平面的两侧.

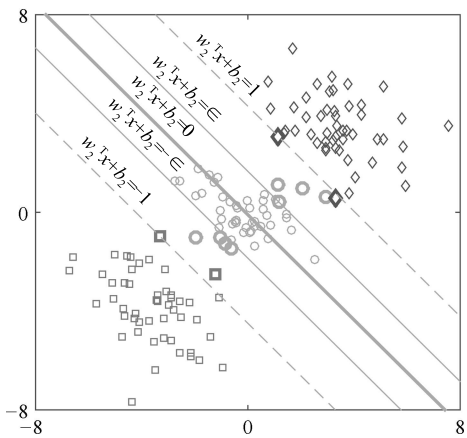

图 1　非平行支持向量顺序回归的几何解释 (以类别 2 超平面构建为例)

Fig. 1　Geometric interpretation of NPSVOR (It shows the construction of the 2-th proximal hyperplane)

若第 k 类对应的模型式 (1) 的解为 $\boldsymbol{w}_k{}^*, b_k^*$, 那么关于第 k 类的最优超平面即为 $f_k(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}_k{}^* + b_k^*$, 其中 $k = 1, 2, \cdots, p$. 预测准则被定义为:

$$r(\boldsymbol{x}) = 1 + \sum_{i=1}^{p-1} [\![f_k(\boldsymbol{x}) + f_{k+1}(\boldsymbol{x}) > 0]\!] \tag{2}$$

其中 $[\![\cdot]\!]$ 表示一个 0-1 示性函数, 即满足条件取 1, 否则取 0. 实际上, 预测函数通过学习的 p 个类别超平面 $f_k(\boldsymbol{x}), k = 1, \cdots, p$ 构建 $p-1$ 有序二元分类的决策函数, 即 $f_k(\boldsymbol{x}) + f_{k+1}(\boldsymbol{x}), k = 1, \cdots, p-1$, 然后以投票的方式进行决策.

2　基于 L2 损失的线性非平行支持向量顺序回归机

本节将在 NPSVOR 模型基础上描述本文提出的基于 L2 损失 (均方 Hinge 损失) 的线性 NPSVOR 模型. 考虑到模型 (1) 中 L1 损失对损失的惩罚是线性关系, 而顺序回归问题建模目标是使预测标签与真实标签尽可能接近, 这促使我们考虑采用 L2 损失函数, 其对于较大的损失给予更大惩罚, 使模型尽可能避免产生较大偏差预测.

据此, 我们考虑建立 L2 损失线性 NPSVOR 模型

$$\begin{aligned}\min_{\boldsymbol{w}_k} f(\boldsymbol{w}_k) = &\frac{1}{2}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k + \\ &C_1 \sum_{i\in\mathcal{I}_k} \max\left(0, |\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i| - \varepsilon\right)^2 + \\ &C_2 \sum_{i\notin\mathcal{I}_k} \max\left(0, 1 - \hat{y}_i^k(\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i)\right)^2\end{aligned} \tag{3}$$

其中 $\hat{y}_i^k = \begin{cases} 1, & y_i > k \\ -1, & y_i \le k \end{cases}$, $C_1, C_2 > 0$ 是惩罚系数. 模型 (3) 相对于模型 (1) 区别就是损失项采用 L2 损失, 为了区别两个模型, 这里记基于 L2 损失的线性 NPSVOR 模型 (3) 为 **L2-NPSVOR**. 引入松弛变量, 模型 (3) 等价于如下形式

$$\begin{aligned}\min_{\boldsymbol{w}_k, \boldsymbol{\xi}_k^+, \boldsymbol{\xi}_k^-, \boldsymbol{\xi}_k} &\frac{1}{2}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k + C_1 \sum_{i\in\mathcal{I}_k} \left((\xi_{ki}^+)^2 + (\xi_{ki}^-)^2\right) + \\ &C_2 \sum_{i\notin\mathcal{I}_k} (\xi_{ki})^2 \\ \text{s.t.}\quad & -\varepsilon - \xi_{ki}^- \le \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i \le \varepsilon + \xi_{ki}^+, i \in \mathcal{I}_k \\ & \hat{y}_i^k(\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x}_i) \ge 1 - \xi_{ki}, i \notin \mathcal{I}_k \\ & \xi_{ki}^+, \xi_{ki}^- \ge 0, i \in \mathcal{I}_k \\ & \xi_{ki} \ge 0, i \notin \mathcal{I}_k\end{aligned} \tag{4}$$

其中 $\boldsymbol{\xi}_k^+ = (\xi_{ki}^+)_{i\in\mathcal{I}_k}$, $\boldsymbol{\xi}_k^- = (\xi_{ki}^-)_{i\in\mathcal{I}_k}, \boldsymbol{\xi}_k = (\xi_{ki})_{i\notin\mathcal{I}_k}$. 在式 (4) 中对每个约束引入对偶因子, 根据 Lagrangian 函数, 利用 Karush-Kuhn-Tucker 定理, 得到原模型式 (3) 的对偶问题:

$$\begin{aligned}\min_{\boldsymbol{\alpha}_k} &\frac{1}{2}\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{w}_k + \sum_{i\in\mathcal{I}_k} \varepsilon(\alpha_{ki}^+ + \alpha_{ki}^-) - \sum_{i\notin\mathcal{I}_k} \alpha_{ki} + \\ &\left\{\frac{1}{4C_1}\sum_{i\in\mathcal{I}_k}\left[(\alpha_{ki}^+)^2 + (\alpha_{ki}^-)^2\right] + \frac{1}{4C_2}\sum_{i\notin\mathcal{I}_k}\alpha_{ki}^2\right\} \\ \text{s.t.}\ & \alpha_{ki} \ge 0, \forall i\end{aligned} \tag{5}$$

其中 $\boldsymbol{w}_k$ 与对偶变量 α_k 相关

$$\boldsymbol{w}_k = -\sum_{i\in\mathcal{I}_k}(\alpha_{ki}^+ - \alpha_{ki}^-)\boldsymbol{x}_i + \sum_{i\notin\mathcal{I}_k}\hat{y}_i^k\alpha_{ki}\boldsymbol{x}_i \tag{6}$$

对于具有 p 类的顺序回归问题, L2-NPSVOR 由 p 个子优化模型式 (3) (或对偶问题式 (5)) 组成.

3　训练算法

在本节中, 我们将针对 L2-NPSVOR 模型, 从原问题及其对偶问题两个角度, 分别设计了信赖域牛顿算法和对偶坐标下降算法求解该模型. 由于不

同 k 对应的原问题式 (3) 及其对偶问题式 (5) 具有相同形式, 为了方便讨论, 在不引起混淆的情况下, 我们将忽略模型式 (3) 和式 (5) 中的下标 k.

3.1 信赖域牛顿法

信赖域牛顿算法 (Trust region Newton method, TRON)[13] 是一种求解可微的无约束或有界约束问题的广义优化算法. Ho 和 Lin 等研究了 L2 损失 SVC 和 SVR 以及 Logistic 回归问题的 TRON 算法[12−14]. 这里将该算法应用于 L2-NPSVOR 模型的求解.

采用 TRON 算法求解原问题 (3), 优化过程包含两层迭代: 在第 t 步外迭代中, 给定 $\boldsymbol{w}_t$, TRON 算法构造 $f(\boldsymbol{w}^t+s)-f(\boldsymbol{w}^t)$ 在信赖域半径 Δ_t 下的二次优化问题, 即

$$\min_{\|\boldsymbol{s}\|\le\Delta_t} q_t(\boldsymbol{s}) \equiv \frac{1}{2}\boldsymbol{s}^{\mathrm{T}}\nabla^2 f(\boldsymbol{w}^t)\boldsymbol{s}+(\nabla f(\boldsymbol{w}^t))^{\mathrm{T}}\boldsymbol{s} \quad (7)$$

然后, 在内层迭代中, 求解该模型获得拟牛顿方向 $\boldsymbol{s}$. TRON 算法根据近似函数 $q_t(\boldsymbol{s})$ 调整优化半径 Δ_t, 具体调整方法参见文献 [14]. 在构造二次优化问题时, 需要计算梯度 $\nabla f(\boldsymbol{w}^t)$ 和 Hessian 矩阵 $\nabla f(\boldsymbol{w}^t)$. 由于 $f(\boldsymbol{w})$ 连续可微, 存在梯度

$$\begin{aligned}\nabla f(\boldsymbol{w}) = \boldsymbol{w}&+2C_1(X_{I_1,:})^{\mathrm{T}}(X_{I_1,:}\boldsymbol{w}-\varepsilon)+\\ &2C_1(X_{I_2,:})^{\mathrm{T}}(X_{I_2,:}\boldsymbol{w}+\varepsilon)+\\ &2C_2(X_{I_3,:})^{\mathrm{T}}(X_{I_3,:}\boldsymbol{w}-\hat{\boldsymbol{y}}_{I_3})\end{aligned}$$

其中

$$\begin{aligned}&I_1=\{i\in\mathcal{I}|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i>\varepsilon\}, I_2=\{i\in\mathcal{I}|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i<-\varepsilon\}\\ &I_3=\{i\notin\mathcal{I}|1-\hat{y}_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)>0\}\end{aligned}$$

这里 $X=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n]^{\mathrm{T}}$ 为样本矩阵, 每行代表一个样本向量, I_1, I_2 和 I_3 为指标集, $X_{I,:}$ 表示 X 中指标集 I 对应的行组成的样本矩阵. 这里 $f(\boldsymbol{w})$ 二阶不可微, 不存在 Hessian 矩阵, 因而不能直接采用牛顿方向对变量更新. 但 $f(\boldsymbol{w})$ 几乎处处二阶可微, 且一阶导数 Lipschitz 连续, 因而根据文献 [13] 可定义广义 Hessian 矩阵, 即

$$B(\boldsymbol{w}) = I + 2X^{\mathrm{T}}DX \quad (8)$$

其中 I 为 m 阶的单位矩阵, D 为 n 阶对角矩阵且

$$D_{ii}=\begin{cases}C_1, & i\in I_1\cup I_2\\ C_2, & i\in I_3\\ 0, & \text{其他}\end{cases}$$

然后, 更新 $\boldsymbol{w}^t$ 和 Δ_t. 先考察

$$\rho_t=\frac{f(\boldsymbol{w}^t+\boldsymbol{s}^t)}{q_t(\boldsymbol{s}^t)} \quad (9)$$

为原问题优化函数减少量与优化的二次函数 $q_t(\boldsymbol{s})$ 减少量的比值. 只要 ρ_t 足够大便更新 $\boldsymbol{w}^t$, 否则不更新. 设 $\eta_0>0$ 是给定的阈值, 则更新 $\boldsymbol{w}^t$ 为

$$\boldsymbol{w}^{t+1}=\begin{cases}\boldsymbol{w}^t+\boldsymbol{s}^t, & \rho_t>\eta_0\\ \boldsymbol{w}^t, & \rho_t\le\eta_0\end{cases} \quad (10)$$

关于更新信赖域半径 Δ_t: Lin 等[13] 给出了如下更新规则, 即根据给定的两常数 η_1 和 η_2 且满足 $\eta_1<\eta_2<1$, Δ_t 的更新依赖于正数 σ_1,σ_2 和 σ_3 满足 $\sigma_1<\sigma_2<1<\sigma_3$, 于是信赖域半径更新条件为

$$\begin{aligned}&\Delta_{t+1}\in[\sigma_1\min\{\|s^t\|,\Delta_t\},\sigma_1\Delta_t], && \rho_t\le\eta_1\\ &\Delta_{t+1}\in[\sigma_1\Delta_t,\sigma_3\Delta_t], && \rho_t\in(\eta_1,\eta_2)\\ &\Delta_{t+1}\in[\Delta_t,\sigma_3\Delta_t], && \rho_t\ge\eta_2\end{aligned} \quad (11)$$

式 (11) 仅给出了 Δ_t 更新满足的范围, 文献 [13] 给出了只要满足以上更新条件算法收敛的证明. 具体更新规则可以有多种方式, 这里我们采用与 LIBLINEAR[15] 中一致的 TRON 更新规则方式[14]:

$$\Delta_{t+1}=\begin{cases}\min(\max(\alpha_t,\sigma_1)\|\boldsymbol{s}^t\|,\sigma_2\Delta_t), & \rho_t<\eta_0\\ \max(\sigma_1\Delta_t,\min(\alpha_t\|\boldsymbol{s}^t\|,\sigma_2\Delta_t)), & \rho_t\in[\eta_0,\eta_1]\\ \max(\sigma_1\Delta_t,\min(\alpha_t\|\boldsymbol{s}^t\|,\sigma_3\Delta_t)), & \rho_t\in(\eta_1,\eta_2)\\ \max(\Delta_t,\min(\alpha_t\|\boldsymbol{s}^t\|,\sigma_3\Delta_t)), & \rho_t\ge\eta_2\end{cases} \quad (12)$$

在更新规则 (12) 中, $\alpha_t\|\boldsymbol{s}^t\|$ 被引入作为 Δ_t 的估计, 其中

$$\alpha_t=\frac{-(\nabla f(\boldsymbol{w}^t))^{\mathrm{T}}\boldsymbol{s}^t}{2(f(\boldsymbol{w}^t+\boldsymbol{s}^t)-f(\boldsymbol{w}^t)-(\nabla f(\boldsymbol{w}^t))^{\mathrm{T}}\boldsymbol{s}^t)} \quad (13)$$

对于算法终止条件, 我们考察算法第 k 次迭代步时目标函数的梯度 $\nabla f(\boldsymbol{w}^t)$ 相对初始梯度 $\nabla f(\boldsymbol{w}^0)$ 关系, 以及样本类别样本规模, 建立如下终止条件

$$\|\nabla f(\boldsymbol{w}^t)\|\le\varepsilon\cdot\frac{\min(|\mathcal{I}|,n-|\mathcal{I}|)}{n}\|\nabla f(\boldsymbol{w}^0)\| \quad (14)$$

其中 ε 为给定终止精度, $|\mathcal{I}|$ 表示指标集 $\mathcal{I}$ 元素个数, n 为训练样本个数.

算法 1 给出了 NPSVOR 的信赖域牛顿法算法主要步骤.

算法 1. TRON: 信赖域牛顿算法求解 L2-NPSVOR 的子模型式 (3)

1) 给定 $\boldsymbol{w}^0$, 子模型式 (3) 对应的类别 k.

2) 根据类别 k 定义 $\hat{y}_i=\begin{cases}-1, & y_i\le k\\ 1, & y_i>k\end{cases}, i=$

$1, \cdots, n$.

3) for $i = 1, 2, \cdots$ (外循环迭代):

共轭梯度算法近似求解信赖域子问题式 (7) 得到 $\boldsymbol{s}^t$.

根据式 (9) 更新 ρ^t.

根据式 (10) 更新 $\boldsymbol{w}^t$ 为 $\boldsymbol{w}^{t+1}$.

根据式 (12) 求解得到 Δ_{t+1}.

如果 $\boldsymbol{w}^t$ 满足终止条件式 (14), 终止迭代.

算法 1 的效率主要依赖于是否能快速求解子优化问题式 (7). 由于目标函数的广义 Hessian 矩阵式 (8) 是 $m \times m$ 阶, 对于高维问题直接计算该矩阵会使得内存难以存储. 同时, 由于样本矩阵 X 高度稀疏, 可采用共轭梯度算法进行求解, 因此只需要在算法优化过程中计算和存储 Hessian 矩阵式 (8) 与向量的乘积, 即

$$B(\boldsymbol{w})\boldsymbol{v} = \boldsymbol{v} + 2C_1((X_{I_1 \cup I_2,:})^{\mathrm{T}}(X_{I_1 \cup I_2,:}\boldsymbol{v})) + 2C_2((X_{I_3,:})^{\mathrm{T}}(X_{I_3,:}\boldsymbol{v})) \tag{15}$$

算法 2 给出了求解问题 (7) 的共轭梯度算法过程.

算法 2. 共轭梯度算法近似求解信赖域子问题 (7)

1) 给定 $\xi_t < 1, \Delta_t > 0$. 令 $\bar{\boldsymbol{s}}^0 = \boldsymbol{0}, \boldsymbol{r}^0 = -\nabla f(\boldsymbol{w}^t)$ 和 $\boldsymbol{d}^0 = \boldsymbol{r}^0$.

2) for $i = 1, 2, \cdots$ (内循环迭代):

如果 $||\boldsymbol{r}^i|| \le \xi_t ||\nabla f(\boldsymbol{w}^t)||$, 输出 $\boldsymbol{s}^t = \bar{\boldsymbol{s}}^i$, 并终止迭代.

$\alpha_i = ||\boldsymbol{r}^i||^2/((\boldsymbol{d}^i)^{\mathrm{T}}\nabla f(\boldsymbol{w}^t)\boldsymbol{d}^i)$

$\bar{\boldsymbol{s}}^{i+1} = \bar{\boldsymbol{s}}^i + \alpha_i \boldsymbol{d}^i$

如果 $\|\bar{\boldsymbol{s}}^{i+1}\| \ge \Delta_t$, 计算 τ 满足 $\|\bar{\boldsymbol{s}}^i + \tau \boldsymbol{d}^i\| = \Delta_t$, 那么输出 $\boldsymbol{s}^t = \bar{\boldsymbol{s}}^i + \alpha_i \boldsymbol{d}^i$, 终止算法.

$\boldsymbol{r}^{i+1} = \boldsymbol{r}^i - \alpha_i \nabla^2 f(\boldsymbol{w}^t)\boldsymbol{d}^i$

$\beta_i = ||\boldsymbol{r}^{i+1}||^2/||\boldsymbol{r}^i||^2$

$\boldsymbol{d}^{i+1} = \boldsymbol{r}^{i+1} + \beta_i \boldsymbol{d}^i$

3.2 对偶坐标下降算法

坐标下降算法 (Coordinate descent method, CD) 是一种无约束优化技术, 被用于求解大规模线性 SVM 模型. Chang 等[9] 利用 CD 算法求解 L2 损失的线性 SVM 模型的原始问题, 实验表明这种方法可以快速获取模型的解. Hsieh 等[11] 提出对偶坐标下降算法 (Dual coordinate descent method, DCD) 求解线性 SVM 模型, 即在 L1 和 L2 损失的线性 SVM 的对偶模型上利用 CD 算法, 并采用 Shrinking 和随机置换优化样本序列的加速技术. 当数据的规模和特征维度规模都比较大时, CD 算法比其他算法在求解线性 SVM 模型上能获得更好的效果[11, 16]. Yuan 等[17] 将 DCD 算法应用于求解 L1 正则化的优化问题. Tseng 和 Yun[18] 系统讨论了 L1 正则优化问题的分解算法, 给出分解算法的一般性框架[12]. 将 DCD 算法扩展到求解大规模 SVR 问题中, 但采用了与文献 [11] 中不同的 Shrinking 准则和算法终止策略, 研究表明这种策略在回归问题中可以快速获得优化模型的解. 本节将利用 DCD 算法求解基于 L2 损失的线性 NPSVOR, 实现大规模顺序回归问题的求解.

忽略原问题式 (3) 的下标 k, 其对偶问题式 (5) 可写为:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \sum_{i \in \mathcal{I}} \varepsilon(\alpha_i^+ + \alpha_i^-) - \sum_{i \notin \mathcal{I}} \alpha_i + \left(\frac{1}{4C_1}\sum_{i \in \mathcal{I}}((\alpha_i^+)^2 + (\alpha_i^-)^2) + \frac{1}{4C_2}\sum_{i \notin \mathcal{I}} \alpha_i^2\right) \tag{16}$$

$$\text{s.t. } \alpha_i \ge 0, \forall i \tag{17}$$

其中 $\boldsymbol{w}$ 与对偶变量 $\boldsymbol{\alpha}$ 相关

$$\boldsymbol{w} = -\sum_{i \in \mathcal{I}}(\alpha_i^+ - \alpha_i^-)\boldsymbol{x}_i + \sum_{i \notin \mathcal{I}} \hat{y}_i \alpha_i \boldsymbol{x}_i \tag{18}$$

根据 KKT 条件可知, α_i^+, α_i^- 不同时非零, 且有 $\alpha_i^+\alpha_i^- = 0$, 因此, 在最优解条件下有

$$\alpha_i^+ + \alpha_i^- = |\alpha_i^+ - \alpha_i^-|, (\alpha_i^+)^2 + (\alpha_i^-)^2 = (\alpha_i^+ - \alpha_i^-)^2 \tag{19}$$

性质成立. 令 $\alpha_i = \alpha_i^+ - \alpha_i^-$, 对偶问题转变为

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \sum_{i \in \mathcal{I}} \varepsilon|\alpha_i| - \sum_{i \notin \mathcal{I}} \alpha_i + \left(\frac{1}{4C_1}\sum_{i \in \mathcal{I}} \alpha_i^2 + \frac{1}{4C_2}\sum_{i \notin \mathcal{I}} \alpha_i^2\right) \tag{20}$$

$$\text{s.t. } \alpha_i \ge 0, i \notin \mathcal{I} \tag{21}$$

其中 $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^{\mathrm{T}}$, 同时式 (18) 变为

$$\boldsymbol{w} = -\sum_{i \in \mathcal{I}} \alpha_i \boldsymbol{x}_i + \sum_{i \notin \mathcal{I}} \hat{y}_i \alpha_i \boldsymbol{x}_i \tag{22}$$

对偶坐标下降算法, 每次仅更新一个变量, 同时固定其他变量. 由于目标函数变量中关于 $\alpha_i, i \in \mathcal{I}$ 存在不可微项, 这里对 $i \in \mathcal{I}$ 和 $i \notin \mathcal{I}$ 的变量分别讨论.

1) 当 $i \in \mathcal{I}$ 时, 令 $A_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i$, 更新变量 α_i 得到如下单变量优化子问题

$$\min_s g(s) = \frac{1}{2}\bar{A}_i s^2 - B_i s + \varepsilon|s| \tag{23}$$

其中 $\bar{A}_i = A_i + 1/(2C_1)$, $B_i = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + \alpha_i \bar{A}_i$ 为常数, 且 B_i 由最近一次更新变量得到. 尽管优化目标

函数 $g(s)$ 不可微, 但可以通过软阈值方法得到闭式解. 令 $g(s)$ 左导数和右导数分别为

$$g_p'(s)=\bar{A}_i s-B_i+\varepsilon,\ g_n'(s)=\bar{A}_i s-B_i-\varepsilon \quad (24)$$

根据软阈值的方法可知, 优化问题的解为

$$\alpha_i \leftarrow \alpha_i - d_i \quad (25)$$

其中

$$d_i=\begin{cases}\frac{g_p'(\alpha_i)}{\bar{A}}, & g_p'(\alpha_i)<\bar{A}_i\alpha_i\\ \frac{g_n'(\alpha_i)}{\bar{A}}, & g_n'(\alpha_i)>\bar{A}_i\alpha_i\\ 0, & \text{其他}\end{cases} \quad (26)$$

在算法迭代中, 需要判断优化变量是否达到最优性条件, 定义优化目标函数关于 α_i 的投影梯度为

$$v_i=\begin{cases}g_p'(\alpha_i), & \alpha_i>0\\ g_n'(\alpha_i), & \alpha_i<0\\ \max(0,g_n'(\alpha_i))-\min(0,g_p'(\alpha_i)), & \alpha_i=0\end{cases} \quad (27)$$

容易得到, α_i 为最优解当且仅当 $v_i=0$.

2) 当 $i\notin\mathcal{I}$ 时, 令 $A_i=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i$, 更新变量 α_i 得到关于步长 d_i 的优化子问题

$$\min_{d_i} h(d_i)=\frac{1}{2}\bar{A}_i d_i^2+G_i d_i,\ \text{s.t.}\ \alpha_i+d_i\geq 0 \quad (28)$$

其中 $\bar{A}_i=A_i+\frac{1}{2C_1}, G_i=\hat{y}_i\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}_i-1$. 令目标函数导数为 0, 即 $h'(d_i)$=0, 可得

$$d_i=\frac{G_i}{\bar{A}_i} \quad (29)$$

此时, 对 α_i 更新

$$\alpha_i\leftarrow\max(0,\alpha_i+d_i) \quad (30)$$

目标函数 $h(d_i)$ 关于 α_i 的投影梯度为

$$v_i=\begin{cases}\max(0,-G_i), & \alpha_i=0\\ G_i, & \alpha_i>0\end{cases} \quad (31)$$

同样, α_i 为最优点当且仅当 v_i=0. 根据以上分析, 我们可以给出优化问题式 (5) 的 DCD 算法过程, 如算法 3 所示, 记该算法为 DCD.

算法 3. DCD: 坐标下降法求解 L2-NPSVOR 的对偶问题 (5)

1) 给定 $\boldsymbol{\alpha}=\boldsymbol{0}$ 和 $\boldsymbol{w}=\boldsymbol{0}$.

2) 计算 $A_i=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i, i=1,\cdots,n$.

3) while $\boldsymbol{\alpha}$ 不满足最优性条件: do (4)

4) for $i=1,\cdots,n$

定义 $\hat{y}_i$: $\hat{y}_i=\begin{cases}-1, & y_i\leq k\\ 1, & y_i>k\end{cases}$

do 步骤 5), 6), 7).

5) if $y_i=k$:

$\bar{A}_i=A_i+\frac{1}{2C_1}$

$G_p=\hat{y}_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+\varepsilon$ 和 $G_n=\hat{y}_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-\varepsilon$,

根据式 (26) 更新 d_i, 根据式 (27) 更新 v_i.

if $v_i\neq 0$: $\bar{\alpha}_i\leftarrow\alpha_i$, $\alpha_i\leftarrow\alpha_i-d_i$.

6) else if $y_i\neq k$:

$\bar{A}_i=A_i+\frac{1}{2C_2}$

$G=\hat{y}_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-1$

根据式 (29) 更新 d_i, 根据式 (31) 更新 v_i.

if $v_i\neq 0$: $\bar{\alpha}_i\leftarrow\alpha_i$,$\alpha_i\leftarrow\max(0,\alpha_i+d_i)$

7) $\boldsymbol{w}\leftarrow\boldsymbol{w}+(\alpha_i-\bar{\alpha}_i)\hat{y}_i\boldsymbol{x}_i$,

关于终止条件, 我们可以采用文献 [17] 的终止条件, 即

$$\left\|\boldsymbol{v}^t\right\|_1<\varepsilon_s\left\|\boldsymbol{v}^0\right\|_1 \quad (32)$$

其中 $\boldsymbol{v}^0$ 和 $\boldsymbol{v}^t$ 分别是初始迭代的违反值 (Violation) 向量和第 t 步迭代的违反值向量. 注意到, $\boldsymbol{v}^t$ 的分量值是通过式 (27) 和式 (31) 在第 t 次迭代中得到.

Shrinking 策略[19], 是一种算法加速技术, 在算法迭代训练时删去一些值不变的变量, 通过减少优化问题的变量规模实现对算法的加速. 该策略常被用于 SVM 的分解算法, 只是不同的算法和模型在具体操作上有所不同. 在本文的 DCD 算法中, 也采用了该技术, 即考虑在算法的有序迭代中, 删除达到约束边界的最优变量 (即 $\alpha_i=0,\forall i\in\mathcal{I}$) 以及不可微点 ($\alpha_i=0,\forall i\notin\mathcal{I}$). 对于变量 α_i, Shrinking 条件为:

当 $i\notin\mathcal{I}$ 时,

$$\alpha_i=0\ \text{且}\ v_i>M \quad (33)$$

当 $i\in\mathcal{I}$ 时,

$$\alpha_i=0\ \text{且}\ g_n'(\alpha_i)<-M\ \text{且}\ g_p'(\alpha_i)>M \quad (34)$$

其中

$$M=\max_{\forall i}|v_i| \quad (35)$$

这里 v_i 是上一次迭代的违反值.

这里取梯度违反值绝对值的最大值式 (35) 作为 Shrinking 阈值条件, 并考察梯度违反值 $\boldsymbol{v}^t$ 与初始 $\boldsymbol{v}^0$ 值缩小比例式 (32) 进行算法终止. 在后面实验中, 我们将进一步对比在设计大规模线性 SVM 的 DCD 算法中提出的 Shrinking 技术和终止策略[11], 即对梯度违反值正负值分别维持阈值 M,m, 然后并将 $M-m<\epsilon$ 作为算法终止条件 (具体见文献 [11]), 以说明本文算法设计的合理性.

4 数值实验

为验证提出的 L2-NPSVOR 模型及算法的有效性, 本文在多个数据集上与其他基于 SVM 的顺序回归模型进行了性能比较. 其中比较的模型包括: L1-NPSVOR、SVM、SVR、RedSVM 等. 此外, 本文还比较了 TRON 和 DCD 在 L2 损失的 NPSVOR 模型的算法效率. 最后本文分析了 L2-NPSVOR 模型对参数的敏感性. TRON 和 DCD 算法均在 LIBLINEAR 框架基础上用 C++ 实现[1], 实验平台为 Intel Xeon 2.0 GHz CPU (E5504), 4 MB cache, 内存 4 GB, Linux 系统.

4.1 实验数据准备与评价标准

针对大规模高维稀疏的顺序回归问题, 目前还缺少相关的研究. 这里我们收集并整理了部分文本顺序回归数据集, 这些数据来自情感分析、电影评论、Amazon 商品评论等多个领域, 具体数据集如下:

1) TripAdvisor[2], 是一个酒店评论数据集, 最早被用于文本潜在语义分析[20]. 每条评论有一个 1 至 5 颗星的打分.

2) Treebank[3], 来自斯坦福大学构建的情感数据库 Treebank, 每条数据对应一个来自 {very negative, negative, neutral, positive, very positive} 的标签.

3) MovieReview[21], 电影评论数据集[4]. 顺序标签是从连续值 $[0,1]$ 离散化得到, 即 a) rating ≤ 0.3, b) $0.4 \leq$ rating ≤ 0.5, c) $0.6 \leq$ rating ≤ 0.7, d) $0.8 \leq$ rating. 该数据集常被用于情感分析.

4) LargeMovie[5], 是一个包含 8 种情感类别的电影评论数据.

5) Amazon 产品评论, 有 8 个数据集, 来自两个数据资源网站: 其中 4 个数据集来自文献 [20], 包括 AmazonMp3, VideoSurveillance, Mobilephone, Cameras[6]; 另外 4 个数据集 (Electronics, HealthCare, AppsAndroid, HomeKitchen)[22−23] 来自 Amazon 产品评论数据集[7], 所有数据均是文本评论, 且有 5 个类别. 由于实际数据集类别样本分布不均衡, 为了不影响模型测试表现, 这里对数据集进行降采样得到平衡数据集.

以上数据集均为文本数据, 故本文对这些数据进行下列预处理: 词干化、去停用词、删除词频小于 3 次的词, 以及在所有文本出现的频率大于 50 % 或出现少于 2 次的词. 此外, 我们采用 unigram, bigram 作为特征, 利用 TF-IDF 技术提取文本特征. 为了方便实验算法分析和方法比较, 将每个数据集随机划分为两部分, 即取 20 % 条数据作为测试集, 剩余 80 % 条数据作为训练集. 数据集的统计描述见表 1 所示, 其特征包括样本规模、特征维数、训练集非零元素个数等.

关于评价标准, 由于顺序回归问题与普通多分类问题不同, 预测标准是预测标签与真实标签尽可能接近, 因此, 这里采用平均绝对误差 (Mean absolute error, MAE) 和平均均方误差 (Mean square error, MSE) 作为评价准则, 即给定预测标签 $\{\hat{y}_1, \cdots, \hat{y}_n\}$ 和实际标签 $\{y_1, \cdots, y_n\}$, MAE 和 MSE 定义如下

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \quad \mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{36}$$

该指标被广泛用于刻画顺序回归模型预测与实际标签的接近程度[6−7, 24−25].

4.2 L2-NPSVOR 与其他模型比较

本节我们测试线性 L2-NPSVOR 的泛化效果, 并与其他 SVM 相关方法比较, 比较方法具体如下:

1) SVC[11], 将顺序回归看成普通多分类问题处理的朴素方法. 文献 [11] 给出了线性 SVM 模型的 DCD 算法, 在算法中采用了随机置换和 Shinking 加速技术. 该算法已经实现并集成在著名的 LIBLINEAR 软件包中, 采用 one-vs-all 的方式策略进行多分类预测.

2) SVR[12], 将顺序回归标签看成普通数值, 采用数值回归的方式进行处理, 同样属于一种朴素方法. SVR 模型预测值是连续的数值, 本文对预测后的连续数值按照相邻的整数离散成相应的类别标签. 文献 [12] 给出了线性 SVR 的 DCD 求解算法, 并在 LIBLINEAR 中实现, 这里仅对预测函数作了修改.

3) RedSVM[7], 对于 p 类顺序回归问题, 其学习一个线性映射将样本映射到一维实数轴上, 在该数轴上寻找 $p-1$ 个具有最大划分间隔的阈值点, 将直线分成 p 个连续的区间段进行预测. 文献 [7] 提出一种处理顺序回归的框架, 将顺序回归问题转化为一个二元分类问题, 对样本 $(\boldsymbol{x}, y)$ 通过扩展将其转化为二分类样本 $((\boldsymbol{x}_i^{\mathrm{T}}, \boldsymbol{e}_k^{\mathrm{T}}), \hat{y}_i^k)$ $(\forall i; k = 1, \cdots, p-1)$, 其中 $\hat{y}_i^k = -1$ 当 $y_i < k$, 否则 $\hat{y}_i^k = 1$.

4) L1-NPSVOR[10], 基于 L1 损失的 NPSVOR, 属于有序分解模型, 根据标签有序信息, 对每个类

[1] 算法代码已上传至 https://github.com/huadong2014/LinearNPSVOR/.

[2] 数据集取自 http://www.cs.virginia.edu/∼hw5x/dataset.html

[3] http://nlp.stanford.edu/sentiment/

[4] scale dataset v1.0: http://www.cs.cornell.edu/people/pabo/movie-review-data/

[5] http://ai.stanford.edu/∼amaas/data/ sentiment/

[6] http://sifaka.cs.uiuc.edu/∼wang296/Data/index.html

[7] Amazon product reviews datasets: http://jmcauley.ucsd.edu/ data/amazon/

表 1 数据集特征描述

Table 1 Data statistics

数据集	样本 (n)	特征 (m)	非零元素个数	类别	类别分布
AmazonMp3	10 391	65 487	1 004 435	5	≈ 2 078
VideoSurveillance	22 281	119 793	1 754 092	5	≈ 4 456
Tablets	35 166	201 061	3 095 663	5	≈ 7 033
Mobilephone	69 891	265 432	5 041 894	5	≈ 13 978
Cameras	138 011	654 268	14 308 676	5	≈ 27 602
TripAdvisor	65 326	404 778	8 687 561	5	≈ 13 065
Treebank	11 856	8 569	98 883	5	≈ 2 371
MovieReview	5 007	55 020	961 379	4	≈ 1 251
LargeMovie	50 000	309 362	6 588 192	8	≈ 6 250
Electronics	409 041	1 422 181	37 303 259	5	≈ 81 808
HealthCare	82 251	283 542	5 201 794	5	≈ 16 450
AppsAndroid	220 566	253 932	6 602 522	5	≈ 44 113
HomeKitchen	120 856	427 558	8 473 465	5	≈ 24 171

别均构造一个三划分并学习一个超平面, 从而构建了优化模型式 (1), 可直接将文献 [11] 的 DCD 算法直接扩展求解其对偶问题.

5) L2-NPSVOR (TRON), 即 L2 损失的 NPSVOR, 采用信赖域牛顿法 (见算法 1) 求解, 共轭梯度算法 (算法 2) 求解信赖域子问题式 (7).

6) L2-NPSVOR (DCD), 即 L2 损失的 NPSVOR, 采用对偶坐标下降法 (见算法 3) 求解, 终止条件为式 (32), Shrinking 策略为式 (33) 和式 (34).

该实验在训练集上进行 5- 折交叉验证进行参数选择, 参数选择范围设定在 $\{2^{-5}, 2^{-4}, \cdots, 2^{5}\}$. 以 MAE 作为交叉验证选参的标准, 通过参数选择后的最优参数作为模型训练的参数. 关于实验参数设置方面, 基于 DCD 算法求解的模型终止精度均设为 0.1, $\boldsymbol{\alpha}$ 和 $\boldsymbol{w}$ 均采用 0 向量作为初始化, TRON 的终止精度设定为 0.001. 另外, 为公平起见, 在 NPSVOR 算法中该实验固定参数 ϵ 值为 0.1, $C_1 = C_2$, 并与其他模型中的 C 采用同样的选参方式, 除 RedSVM 模型[8], 其他模型均采用有偏置项模型. 实验数据集如表 1 所示. 表 2 给出了各模型在不同数据集上 MAE、MSE 值和训练时间 (Time), 表中每行最好的结果均已经加粗显示. 表 2 的最后列出了各方法在所有数据集上关于 MAE、MSE 和训练时间的平均排序, 以方便比较各模型之间的性能.

从表 2 的结果中, 通过观察可以得到以下几点结论:

1) 根据各方法在所有数据集上平均排序可以看出, L2-NPSVOR 较其他方法在 MAE 和 MSE 上, 取得了最好的预测效果. 虽然 L2-NPSVOR 在 TRON 和 DCD 算法得到的预测效果接近, 但 DCD 在总体上得到了更好的 MAE 和 MSE 值, 算法的训练时间相对 TRON 优势明显.

2) RedSVM 模型在非线性情况下表现突出[4], 但在大规模数据集的表现略低于朴素方法线性 L1/L2-SVC.

3) 对比 L1-NPSVOR 和 L2-NPSVOR, 采用 L2 损失的模型在 MAE、MSE 优于 L1-NPSVOR 模型, 这与我们的预期一致, 即顺序回归问题预测要求预测标签与实际标签尽可能接近, L2 损失对于损失偏差较大的样本给予更大的惩罚, 可得到预测偏差更小的模型. 此外, 基于 L2 损失的 NPSVOR 在 DCD 算法下可以得到更快的优化速度.

4) 在算法的训练时间上, 基于 DCD 算法的 L2-NPSVOR 获得了除 SVR 外最快的训练速度. 尽管 SVR 在时间上具有优势, 但其在顺序回归上的预测结果相对较差, 这也说明将顺序回归问题等同于数值回归存在一定的缺陷.

4.3 TRON 与 DCD 算法比较

本文针对 NPSVOR 提出了信赖域牛顿算法和对偶坐标下降算法, 这里我们对算法效率进行比较. 假设原问题的优化目标函数为 $f(\boldsymbol{w})$, 通过观察算法训练过程中目标函数值 $f(\boldsymbol{w})$ 与最优值目标函数值 $f(\boldsymbol{w}^*)$ 的接近程度, 即 $f(\boldsymbol{w}) - f(\boldsymbol{w}^*)$ 来比较算法效率. 为说明本文算法设计的合理性, 考虑以下四种情形:

[8] 需要注意的是, 因为目前还没有针对顺序回归问题提出的大规模求解算法, RedSVM 模型只有非线性模型的求解算法, 故本文对线性 RedSVM 算法求解时采用文献 [7] 中 DCD 算法对 RedSVM 的线性版本进行实现.

表 2　方法在各数据集上测试结果, 包括 MAE、MSE 和最优参数下的训练时间 (s)

Table 2　Test results for each dataset and method, including MAE, MSE and training time (s)

数据集	指标	L1-SVC	L2-SVC	SVR	RedSVM	NPSVOR	L2-NPSVOR (TRON)	L2-NPSVOR (DCD)
AmazonMp3	MAE	0.564	0.557	0.534	0.535	0.488	0.481	**0.481**
	MSE	0.996	0.987	0.732	0.735	0.699	**0.670**	0.683
	TIME	0.209	0.660	**0.031**	0.186	0.165	0.830	0.144
VideoSurveillance	MAE	0.404	0.391	0.426	0.446	0.376	**0.371**	0.372
	MSE	0.709	0.668	0.578	0.592	0.511	0.493	**0.491**
	TIME	0.433	1.708	**0.087**	0.551	0.492	1.996	0.402
Tablets	MAE	0.306	0.299	0.334	0.346	0.280	**0.278**	0.278
	MSE	0.514	0.496	0.444	0.444	0.373	**0.362**	0.363
	TIME	0.821	3.400	**0.198**	1.029	0.948	2.958	0.674
Mobilephone	MAE	0.431	0.419	0.450	0.444	0.391	0.388	**0.385**
	MSE	0.736	0.705	0.604	0.587	0.536	0.524	**0.522**
	TIME	1.811	7.574	**0.353**	1.909	2.330	6.724	1.605
Cameras	MAE	0.246	0.240	0.273	0.301	0.227	0.232	**0.226**
	MSE	0.394	0.381	0.357	0.375	**0.296**	0.299	0.298
	TIME	9.552	34.480	**1.401**	6.016	6.341	30.132	5.388
TripAdvisor	MAE	0.398	0.388	0.433	0.429	0.365	**0.365**	0.366
	MSE	0.611	0.583	0.539	0.523	**0.445**	0.449	0.452
	TIME	2.331	12.778	**0.807**	2.110	2.857	9.238	3.505
Treebank	MAE	0.907	0.841	0.784	**0.752**	0.763	0.806	0.756
	MSE	1.652	1.455	1.116	**1.049**	1.126	1.229	1.068
	TIME	0.025	0.040	**0.004**	0.015	0.026	0.035	0.024
MovieReview	MAE	0.501	0.490	0.448	0.447	0.432	0.436	**0.431**
	MSE	0.615	0.582	0.486	0.485	0.476	0.476	**0.475**
	TIME	0.121	0.429	**0.029**	0.133	0.130	0.373	0.125
LargeMovie	MAE	1.205	1.176	1.182	1.093	**0.992**	1.008	1.002
	MSE	3.617	3.502	2.469	2.225	2.046	2.075	**2.020**
	TIME	3.311	10.416	**0.328**	1.965	2.569	7.523	2.493
Electronics	MAE	0.592	0.590	0.606	0.620	0.529	0.526	**0.520**
	MSE	1.069	1.050	0.840	0.848	0.747	0.736	**0.731**
	TIME	22.316	168.141	**4.878**	10.736	23.075	116.586	18.062
HealthCare	MAE	0.637	0.621	0.660	0.681	0.591	0.590	**0.589**
	MSE	1.338	1.282	1.004	1.062	0.945	**0.920**	0.929
	TIME	2.098	7.429	**0.439**	2.686	2.954	6.365	2.673
AppsAndroid	MAE	0.640	0.616	0.656	0.659	**0.584**	0.590	0.584
	MSE	1.179	1.106	0.922	0.920	**0.844**	0.872	0.859
	TIME	4.043	14.924	**0.634**	1.603	4.574	11.423	6.290
HomeKitchen	MAE	0.585	0.574	0.597	0.609	0.519	0.519	**0.510**
	MSE	1.050	1.015	0.829	0.842	0.745	0.723	**0.720**
	TIME	5.587	19.393	**0.896**	1.786	5.171	19.560	4.475
平均排序	MAE	5.64	4.57	5.64	5.86	2.50	2.21	**1.57**
	MSE	7.00	6.00	4.36	4.29	2.36	2.29	**1.64**
	TIME	3.57	6.79	**1.00**	3.29	4.07	6.21	3.07

TRON: 即本文给出的信赖域牛顿算法 1, 在算法中 NPSVOR 各子模型独立求解, 均初始化 $\boldsymbol{w}=\mathbf{0}$, 可分布并行求解各子模型. 为方便比较, 这里仅考虑串行求解方式.

TRON (WS): 在利用 TRON 算法求解 NPSVOR 各子模型时, 可采用暖启动策略 (Warm start, WS). 由于在 NPSVOR 模型中, 每个子模型的超平面是基于有序三元分解建立的, 其是根据标签的有序结构信息得到, 因而相邻类别对应的超平面相对比较接近, 对应的解具有相似的结构, 即 $\boldsymbol{w}_k \approx \boldsymbol{w}_{k+1}$. 如果依次求解 $k=1,\cdots,p$ 对应的模型, 在求解第 $k+1$ 个子模型 $\boldsymbol{w}_{k+1}$ 时, 可以利用第 k 个模型的解 $\boldsymbol{w}_k$ 作为其初始解, 这样可以利用顺序回归的特有性质来加速模型的求解.

DCD-M: 即本文给出的对偶坐标下降算法 3. 算法中采用了 Yuan 等[17] 给出的终止准则和 Shrinking 策略, 即采用最优梯度违反值的绝对值最大值式 (35) 作为判断最优性条件的阈值.

DCD-Mm: Hsieh 等[11] 在设计线性 SVM 的对偶坐标下降算法中, 根据梯度投影的最大幅度值作为终止条件的判断依据, 并根据上下振幅值作为 Shrinking 阈值, 将该策略应用到文本的 DCD 算法中, 即 Shrinking 条件为, 当 $i\in\mathcal{I}$ 时,

$$\alpha_i=0 \text{ 且 } g_n'(\alpha_i)<-m \text{ 且 } g_p'(\alpha_i)>M \tag{37}$$

其中 $m=\min_{\forall i} v_i$ 和 $M=\max_{\forall i} v_i$ 这里 v_i 是上一次迭代的违反值. 终止条件为:

$$M-m<\epsilon \tag{38}$$

该实验选择数据集中样本规模较大的 8 个数据集, 固定参数 $C=1$, 模型 NPSVOR 中的参数 ϵ 固定值为 0.1. TRON 算法和 DCD 算法的终止精度分别设为 0.01 和 0.001. TRON 的其他参数选择与文献 [14] 相同, 即

$$\eta_0=0,\eta_1=0.25,\eta_2=0.75$$
$$\sigma_1=0.25,\sigma_2=0.5,\sigma_3=4$$

在训练集上进行训练, 记录目标函数值变化情况. 由于 L2-NPSVOR 对于每个类别均需要求解一个子优化模型, 图 2 仅展示类别 3 对应的子优化模型 (即 $k=3$) 绝对目标函数差训练时间的变化.

从图 2 中, 我们可以观察到: 1) 采用暖启动的 TRON 算法 TRON (WS) 在其中 6 个数据集上有较为明显的加速, 但在短的训练时间内, 加速不明显. 2) 目标函数的 $f(\boldsymbol{w}^t)-f(\boldsymbol{w}^*)$ 在 DCD-M 算法优化下在给出的 8 个数据集中比 TRON 算法高效并且获得了更低的目标函数值, DCD-M 算法优势明显. 3) DCD-Mm 采用文献 [11] 的终止条件和 Shrinking 策略, 实验表明目标函数值过早趋于平稳, 并且

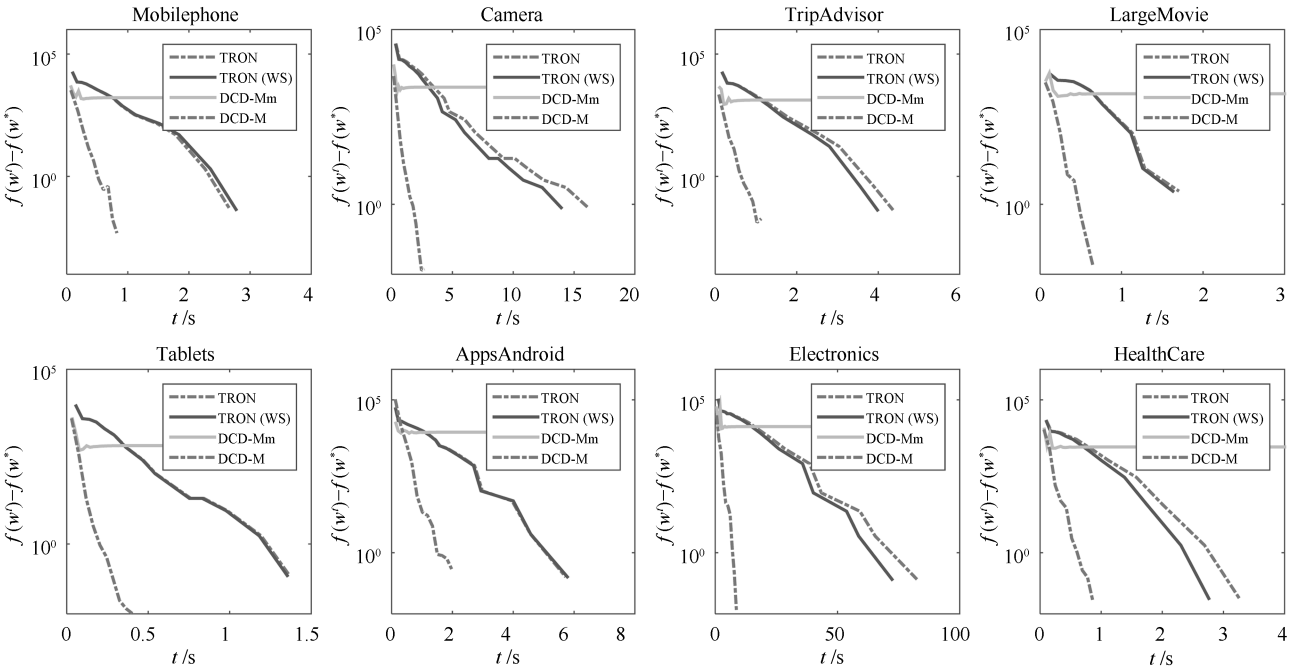

图 2 TRON, TRON (WS), DCD-M and DCD-Mm 在 8 个数据集上的比较 (这里展示了类别 3 对应的优化问题). 横坐标是时间 t, 纵坐标为 L2-NPSVOR 原问题目标函数的 $f(\boldsymbol{w}^t)-f(\boldsymbol{w}^*)$ 的值

Fig. 2 Comparison of TRON, TRON (WS), DCD-M and DCD-Mm on eight datasets (Show the optimization model for rank 3). The horizontal axis is training time in seconds and the vertical axis is the difference between $f(\boldsymbol{w}^t)$ and $f(\boldsymbol{w}^*)$

不能够及时有效终止算法.

4.4 L1-NPSVOR 和 L2-NPSVOR 参数敏感性

参数选择通常十分耗时, 故我们期望得到的模型对参数不敏感, 即参数值的变化不会有较大的测试结果变化. 本节实验将比较 L1-NPSVOR 和 L2-NPSVOR 关于 MAE 和 MSE 随参数值 C 改变的变化情况 (限定 $C_1 = C_2$ 并记为 C), 采用 DCD 算法求解. 实验选择与第 4.3 节相同的 8 个数据集进行实验, 参数 C 的变化范围设定为 $\{2^{-5}, 2^{-4}, \cdots, 2^5\}$, 采用 5 折交叉验证得到每个参数值下的测试 MAE/MSE 值. MAE 和 MSE 变化趋势分别如图 3 和图 4 所示.

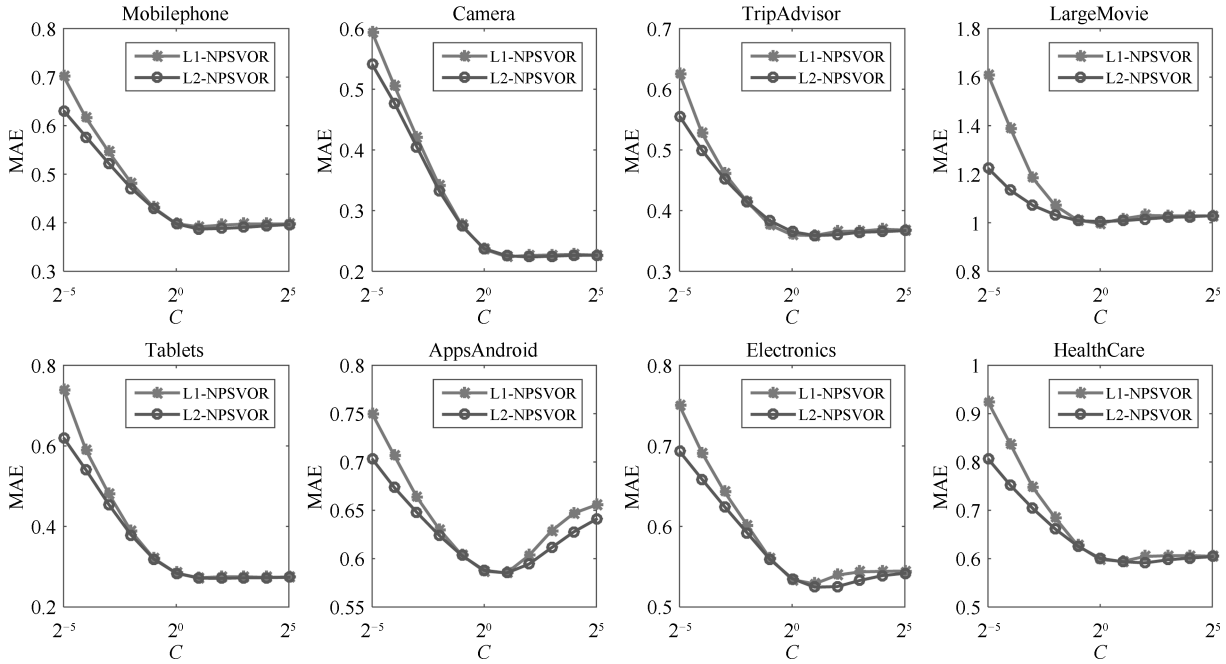

图 3　L1/L2-NPSVOR 的 MAE 分别随参数 C 变化

Fig. 3　Test MAE results of L1/L2-NPSVOR change with parameter C on eight datasets

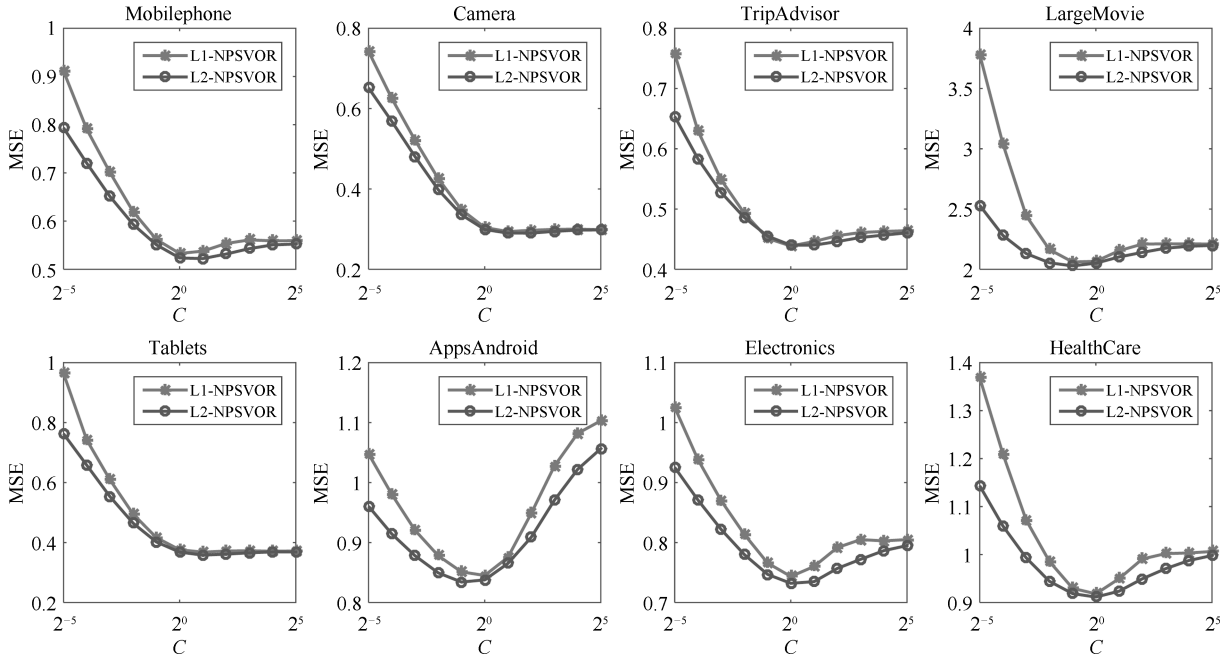

图 4　L1/L2-NPSVOR 的 MSE 分别随参数 C 变化

Fig. 4　Test MSE results of L1/L2-NPSVOR change with parameter C on eight datasets

从图 3 和图 4 中可以观察得到, 基于 L2 损失的 NPSVOR 在 MAE/MSE 上随参数 C 变化相对平稳, 变化幅度均低于采用 L1 损失的对应结果, 尤其是在参数值 C 较小的情况下, 在此情形下, L1-/L2-NPSVOR 均在训练数据集上出现欠拟合问题, 但是由于 L2 损失对损失的惩罚要高于 L1 损失, 故欠拟合问题严重性相对较弱一些. 另外, 从图 4 中最后 3 个数据集 (Apps Android、Electronics 和 Health Care) 对比看出, L2 损失在 MSE 上有显著地提高. 以上结果表明, L2 损失对参数 C 的敏感性低于 L1 损失, 即利用 L2 损失, 可以更容易选出较合适的参数.

5 结束语

本文针对大规模、高维、稀疏的顺序回归问题, 考虑到 L2 损失的引入对偏离较大的点给予更大的惩罚, 使得预测标签与真实标签更加接近, 而线性模型的提出能成功解决大规模数据面对的速度及内存消耗问题, 所以本文提出基于 L2 损失的线性非平行支持向量顺序回归模型 — L2-NPSVOR. 另外本文从原问题及其对偶问题两个角度, 分别设计了信赖域牛顿算法和对偶坐标下降算法求解该模型. 其中在 TRON 算法中, 考虑到顺序回归相邻的超平面具有相似的解, 在算法求解时提出暖启动的方法. 最后, 为验证模型及算法的有效性, 本文在收集的大量文本顺序回归数据上对提出的模型及算法进行了分析和比较. 结果表明, 相比其他基于 SVM 的顺序回归模型, L2-NPSVOR 在性能上表现最优; 关于求解算法, TRON 算法能够获得比 DCD 更加精确的解, 但当样本维数远远高于样本数时, DCD 算法比 TRON 更加高效.

References

1 Nakov P, Ritter A, Rosenthal S, Sebastiani F, Stoyanov V. SemEval-2016 task 4: sentiment analysis in Twitter. In: Proceedings of the 10th International Workshop on Semantic Evaluation. San Diego, CA, USA: ACL, 2016. 1−18

2 Dikkers H, Rothkrantz L. Support vector machines in ordinal classification: An application to corporate credit scoring. *Neural Network World*, 2005, **15**(6): 491

3 Chang K Y, Chen C S, Hung Y P. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA: IEEE, 2011. 585−592

4 Gutiérrez P A, Pérez-Ortiz M, Sánchez-Monedero J, Fernández-Navarro F, Hervás-Martínez C. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 2016, **28**(1): 127−146

5 Zhang Xue-Gong. Introduction to statistical learning theory and support vector machines. *Acta Automatica Sinica*, 2000, **26**(1): 32−42
(张学工. 关于统计学习理论与支持向量机. 自动化学报, 2000, **26**(1): 32−42)

6 Chu W, Keerthi S S. Support vector ordinal regression. *Neural Computation*, 2007, **19**(3): 792−815

7 Lin H T, Li L. Reduction from cost-sensitive ordinal ranking to weighted binary classification. *Neural Computation*, 2012, **24**(5): 1329−1367

8 Pérez-Ortiz M, Gutiérrez P A, Hervás-Martínez C. Projection-based ensemble learning for ordinal regression. *IEEE Transactions on Cybernetics*, 2014, **44**(5): 681−694

9 Chang K W, Hsieh C J, Lin C J. Coordinate descent method for large-scale L2-loss linear support vector machines. *The Journal of Machine Learning Research*, 2008, **9**: 1369−1398

10 Wang H D, Shi Y, Niu L F, Tian Y J. Nonparallel support vector ordinal regression. *IEEE Transactions on Cybernetics*, 2017, **47**(10): 3306−3317

11 Hsieh C J, Chang K W, Lin C J, Keerthi S S, Sundararajan S. A dual coordinate descent method for large-scale linear SVM. In: Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008. 408−415

12 Ho C H, Lin C J. Large-scale linear support vector regression. *The Journal of Machine Learning Research*, 2012, **13**(1): 3323−3348

13 Lin C J, Moré J J. Newton′s method for large bound-constrained optimization problems. *SIAM Journal on Optimization*, 1999, **9**(4): 1100−1127

14 Lin C J, Weng R C, Keerthi S S. Trust region newton method for logistic regression. *The Journal of Machine Learning Research*, 2008, **9**: 627−650

15 Hsia C Y, Zhu Y, Lin C J. A study on trust region update rules in newton methods for large-scale linear classification. In: Proceedings of the 9th Asian Conference on Machine Learning (ACML). Seoul, South Korea: ACML, 2017

16 Chiang W L, Lee M C, Lin C J. Parallel Dual coordinate descent method for large-scale linear classification in multi-core environments. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: ACM, 2016. 1485−1494

17 Yuan G X, Chang K W, Hsieh C J, Lin C J. A comparison of optimization methods and software for large-scale l1-regularized linear classification. *The Journal of Machine Learning Research*, 2010, **11**: 3183−3234

18 Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 2009, **117**(1−2): 387−423

19 Joachims T. Making Large-scale SVM Learning Practical, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, Germany, 1998

20 Wang H N, Lu Y, Zhai C X. Latent aspect rating analysis on review text data: a rating regression approach. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA: ACM, 2010. 783−792

21 Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. Ann Arbor, USA: ACL, 2005. 115−124

22 McAuley J, Targett C, Shi Q F, van den Hengel A. Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile: ACM, 2015. 43−52

23 McAuley J, Pandey R, Leskovec J. Inferring networks of substitutable and complementary products. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia: ACM, 2015. 785−794

24 Tang D Y, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: ACL, 2015. 1422−1432

25 Diao Q M, Qiu M H, Wu C Y, Smola A J, Jiang J, Wang C. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2014. 193−202

石 勇 中国科学院大学教授. 于 1991 在美国堪萨斯大学商学院获得管理科学和计算机系统专业博士学位. 主要研究方向为数据挖掘和多目标决策分析.
E-mail: yshi@ucas.ac.cn
(**SHI Yong** Professor at University of Chinese Academy of Sciences. He received his Ph. D. degree in management science and computer systems from University of Kansas, Lawrence, KS, USA, in 1991. His research interest covers data mining and multiple criteria decision making.)

李佩佳 中国科学院大学计算机与控制学院博士研究生. 目前在中国科学院虚拟经济与数据科学研究中心学习. 她于 2013 年获得河南师范大学工学学士学位. 主要研究方向为数据挖掘, 深度学习和自然语言处理. 本文通信作者.
E-mail: lipeijia13@mails.ucas.ac.cn
(**LI Pei-Jia** Ph. D. candidate at the School of Computer and Control Engineering, University of Chinese Academy of Sciences. She is also studying at the Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences. She received her bachelor degree from Henan Normal University in 2013. Her research interest covers data mining, deep learning and natural language processing. Corresponding author of this paper.)

汪华东 北京三星通信研究院助理研究员. 于 2017 年 7 月在中国科学院大学数学科学学院获博士学位. 于 2014 ∼ 2017 年在中国科学院虚拟经济与数据科学研究中心学习. 主要研究方向为支持向量机, 深度学习, 优化理论及应用和数据挖掘. E-mail:
wanghuadong14@mails.ucas.ac.cn
(**WANG Hua-Dong** Assistant research fellow at Samsung R&D Institute China – Beijing. He received the Ph. D. degree in July, 2017 from the School of Mathematica Science, University of Chinese Academy. He studied at the Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences from 2014 to 2017. His research interest covers support vector machines, machine learning, optimization theory and applications and data mining.)