doi:10.3969/j.issn.1001-4616.2025.04.010

基于局部邻接图的半监督稀疏回归算法

秦晓燕[1],郑晓晗[2],张莉[1,2]

(1.苏州高博职业学院信息与软件学院,江苏 苏州 215163)
(2.苏州大学计算机科学与技术学院,江苏 苏州 215006)

[**摘要**] 半监督回归算法可以利用少量有标签样本和大量无标签样本进行回归建模,进而在一定程度上解决了获取标签样本成本高的问题. 目前,已经提出了能利用邻接矩阵来挖掘数据潜在结构的图半监督回归算法. 然而,这些方法存在两个问题. 第一,现有半监督回归算法多使用全连接的图生成方式,很容易受到噪声或离群点的影响;第二,现有图半监督回归算法的稀疏性不足. 为了解决上述问题,本文提出了一种新的半监督回归学习算法——基于局部邻接图的半监督稀疏回归算法. 该算法构建了一种新的局部邻接图,在生成邻接矩阵时仅关注样本的局部信息,从而保留了数据的局部流形结构,缓解噪声样本对算法的影响. 另外,利用 1 范数正则能诱导稀疏性的特性,本文在优化问题中引入模型系数的 1 范数正则项,有效地提高了模型的稀疏性. 本文在 9 个真实数据集上对算法的半监督回归性能和稀疏性进行了实验验证. 实验结果表明,本文所提算法在不同实验设置下均能获得较好的回归性能.

[**关键词**] 半监督回归学习,图生成,邻接矩阵,稀疏性,1 范数正则
[**中图分类号**]TP181　[**文献标志码**]A　[**文章编号**]1001-4616(2025)04-0096-10

Semi-Supervised Sparse Regression Based on Local Adjacency Graph

Qin Xiaoyan[1],Zheng Xiaohan[2],Zhang Li[1,2]

(1.School of Information and Software,Suzhou Global Institute,Suzhou 215163,China)
(2.School of Computer Science and Technology,Soochow University,Suzhou 215006,China)

Abstract:Semi-supervised regression(SSR) algorithms leverage a small amount of labeled samples along with a large pool of unlabeled samples for modeling regression functions,which can alleviate the high costs associated with obtaining labeled data to some extent. At present,graph-based SSR algorithms have been proposed,which can employ adjacency matrices to dig the underlying data structure. However,existing graph-based SSR methods are confronted with two primary challenges. First,existing methods rely on the fully-connected graph on data that is susceptible to noise and outliers. Second,the sparsity of current SSR algorithms leaves room for enhancement. To address these issues,this paper proposes a novel semi-supervised sparse regression algorithm. This algorithm designs a new graph,the local adjacency graph,which focuses on the local connectivity of samples. This graph generation method preserves the local manifold structure of the data and mitigates the impact of noisy samples. Furthermore,our algorithm capitalizes on the sparsity-inducing property of the 1-norm regularization by incorporating a 1-norm regularization term for the model coefficients into the optimization problem,effectively enhancing the model's sparsity. Empirical validation is performed on nine real-world datasets to evaluate the SSR performance and the sparsity of the proposed algorithm. Results demonstrate that our method achieves superior performance across various experimental setups.

Key words:semi-supervised regression learning,adjacency matrix,graph generation,sparsity,1-norm regularization

半监督学习作为机器学习的重要分支,有效地融合了全监督与无监督学习的优势,通过结合少量标注数据和大量未标注数据的训练,旨在提升模型的泛化性能[1-4]. 该学习模式在打标签成本高或难以获取标签的场景下显得尤为重要. 对于回归任务而言,半监督学习通过挖掘未标注数据的内在结构信息,有助于模型更精准地预测连续变量,从而优化了模型的预测精度和泛化能力[5-6]. 这一策略为处理标签稀缺的回

收稿日期:2024-09-24.
基金项目:江苏省高校自然科学研究资助项目(19KJA550002)、江苏省六大人才高峰资助项目(XYDXX-054)、江苏省职业教育软件技术“双师型”名师工作室资助项目(苏教师函[2022]31 号).
通讯作者:张莉,教授,博士生导师,研究方向:机器学习. E-mail:zhangliml@suda.edu.cn

归问题提供了一种高效的解决路径,显著提升了模型在实际应用中的性能表现[7-9].

已有的半监督回归算法可以大致分为两类:基于协同训练的和基于图的方法. 协同训练采用两个或多个模型从不同视角学习,通过互相学习来提高各自的预测能力,其代表算法为协同训练回归器(co-training regressor,COREG)[10-11]. 基于图的半监督回归算法通过构建数据点之间的图表示,利用图的结构信息来指导模型学习,使得模型能够更好地捕捉数据的内在结构和分布,其常见的算法有拉普拉斯支持向量回归机(Laplacian support vector regression,LapSVR)[12]以及拉普拉斯正则化最小二乘回归机(Laplacian regularized least squares regression,LapRLSR)[13]. 这些半监督回归算法各有侧重,但都旨在利用未标记数据的潜在信息,以达到更优的学习效果. 本文主要关注的是基于图的半监督回归算法.

尽管上述算法在处理半监督回归任务时展现出了一定的优势,但是它们普遍缺乏稀疏性,这限制了其在特定场景下的应用效率和性能. 稀疏性是指在模型的表示中,大多数的模型系数为零或接近零,这一特性能带来诸多优势[14-17]. 具备稀疏性的算法能显著降低模型的复杂度,提高计算效率,同时减少内存消耗. 更重要的是,稀疏性有助于增强模型的泛化能力,通过消除不必要的数据干扰,使模型更专注于关键数据,从而提升算法的性能. 此外,稀疏模型更易于解释,有利于理解模型的决策过程和提升模型的透明度. 因此,在设计和选择算法时,考虑稀疏性对算法性能的提升具有重要的意义.

另一方面,邻接图的质量在很大程度上决定了算法性能的优劣[18-21]. 邻接图通过描绘数据点间的连接性、距离及其他相互作用,为算法提供了一种直观的数据关系表示. 这种表示的准确性直接关联到算法能否有效识别数据的内部结构,进而影响其在分类、聚类、回归任务中的准确性和效率. 不准确的邻接图可能会误导算法,忽视关键的结构信息,从而降低结果的准确度. 另外,常见的基于图的半监督回归算法多使用全连接图来表示训练数据的结构信息[13-14]. 在存在噪声或离群点时,采用全连接图会使得算法性能变差.

为了解决上述问题,本文提出了一种新的基于局部邻接图的半监督稀疏回归(semi-supervised sparse regression based on local adjacency graph,LAG-S3R)算法. 本文的主要贡献包括 2 个方面:

(1)本文提出了一种新的图生成方法:局部邻接图. 该图生成方法是对每个样本用其近邻的线性加权来表示,进而保留了数据的局部流形结构,减少噪声样本或离群点对算法性能的影响;

(2)在局部邻接图基础上,本文提出了 LAG-S3R 算法. 通过引入 1 范数正则来提高回归模型的稀疏性,并且在原空间而非对偶空间中求解 LAG-S3R 的优化问题,以保证模型的稀疏性不会在求解过程中丢失. 值得注意的是,LAG-S3R 可以同时实现回归预测和样本约减的任务.

1 相关工作

在此,首先给出本文相关的符号定义. 令训练样本矩阵为 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]^{\mathrm{T}}\in\mathbf{R}^{n\times m}$,这里 $\boldsymbol{x}_i$ 是训练样本,n 和 m 分别表示矩阵 $\boldsymbol{X}$ 的行和列,也就是样本个数和样本特征数. 对于半监督回归学习,我们定义矩阵 $\boldsymbol{X}$ 中前 l 行是有标签样本,其标签向量表示为 $\boldsymbol{y}\in\mathbf{R}^l$;剩余 u 行是无标签样本,且有 $l+u=n$. 对于任意矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$,令 a_{ij}、$\boldsymbol{a}_i=[a_{i1},a_{i2},\cdots,a_{in}]^{\mathrm{T}}$、$\boldsymbol{A}^{\mathrm{T}}$ 分别表示矩阵 $\boldsymbol{A}$ 的第 i 行第 j 列、$\boldsymbol{A}$ 的第 i 行元素组成的列向量、$\boldsymbol{A}$ 的转置矩阵. $\mathbf{0}$ 和 $\mathbf{1}$ 分别表示适当维度的全 0 或全 1 向量.

1.1 图的生成

图方法需要将样本点与图的顶点之间建立起一一对应关系,并以样本点之间的某种相似度作为图的边. 这样就根据样本点构建了一个与之对应的图. 如何构造合适的图是图方法的核心步骤. 不失一般性,用矩阵 $\boldsymbol{G}$ 来描述邻接矩阵.

图方法中采用得最多的是 Laplacian 矩阵. 令 Laplacian 矩阵为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{G}$,其中 $\boldsymbol{D}$ 是对角矩阵,其对角元素为 $D_{ii}=\sum_{j=1}^{l}g_{ij}$,$g_{ij}$表示矩阵 $\boldsymbol{G}$ 的第 i 行第 j 列. 目前,基于图的半监督回归方法通过构造全连接图来体现不同样本的相似关系,用样本之间的距离来表示相似度,计算方式如下:

$$\boldsymbol{G}^f=\exp\left\{-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{\sigma}\right\},\tag{1}$$

式中,σ 为预定义的热核参数. 可以看到,全连接图考虑了所有可能样本对之间的相似性信息. 如果数据集存在噪声或离群点,全连接图的生成质量可能会受到影响,进而降低学习算法的性能.

1.2　拉普拉斯支持向量回归机

LapSVR 是通过向支持向量回归机(support vector regression,SVR)[22]中添加了流形正则项而被提出的. 利用该流形正则项,LapSVR 可以实现半监督回归,其决策函数为:

$$f(\boldsymbol{x})=\sum_{i=1}^{n}\alpha_i\kappa(\boldsymbol{x}_i,\boldsymbol{x})+b, \tag{2}$$

式中,$\kappa(\cdot,\cdot)$表示核函数,$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_n]^{\mathrm{T}}$ 和 b 表示权重向量和阈值.

为了确定决策函数,LapSVR 构造了如下的优化问题:

$$\begin{aligned}&\min_{\boldsymbol{\alpha},\boldsymbol{\xi}^{(*)}}\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}+C_1\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{K}\boldsymbol{\alpha}+C_2\sum_{i=1}^{l}(\xi_i+\xi_i^*),\\&\text{s.t. }y_i-\sum_{j=1}^{n}\alpha_i\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)-b\leqslant\varepsilon+\xi_i,\\&\sum_{j=1}^{n}\alpha_i\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)+b-y_i\leqslant\varepsilon+\xi_i^*,\\&\xi_i,\xi_i^*>0,i=1,\cdots,l.\end{aligned} \tag{3}$$

式中,核 Gram 矩阵 $\boldsymbol{K}$ 的第 i 行可以表示 $\boldsymbol{\kappa}_i=[\kappa(\boldsymbol{x}_1,\boldsymbol{x}_i),\kappa(\boldsymbol{x}_2,\boldsymbol{x}_i),\cdots,\kappa(\boldsymbol{x}_n,\boldsymbol{x}_i)]^{\mathrm{T}}$,$\boldsymbol{\xi}^{(*)}$表示松弛向量,$C_1,C_2>0$ 为两个预定义的正则化参数,ε 为 ε 不敏感损失函数的参数.

优化问题中的第一项旨在降低模型复杂度;第二项为流形正则项,最小化这一项可以使所求流形空间尽可能地光滑,同时将有标签样本与无标签样本联系起来;第三项是 ε 不敏感损失函数. 然而,LapSVR 缺乏稀疏性且不具有特征约减的能力. 另外,该算法使用式(1)生成邻接矩阵,这会在一定程度上降低算法的鲁棒性.

2　本文算法

本节首先设计了一种新的图生成方法,生成的图名为局部邻接图. 接着,利用该图生成方法以及 1 范数正则化项,我们提出了一个新的半监督回归算法——基于局部邻接图的稀疏半监督回归算法. 本节将深入探讨该算法的优化问题及其求解策略,提供详细的算法设计和实现细节.

2.1　局部邻接图

定义 1(局部邻接图)　已知训练样本集合$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,$n$ 是样本个数,则局部邻接图 $\boldsymbol{G}$ 需要通过求解如下优化问题获得:

$$\begin{aligned}&\min_{\boldsymbol{G}}\Big(\sum_{i=1}^{n}\Big\|\boldsymbol{x}_i-\sum_{\boldsymbol{x}_{i_j}\in\pi_k(\boldsymbol{x}_i)}\boldsymbol{x}_{i_j}g_{ii_j}\Big\|^2\Big),\\&\text{s.t.}\sum_{\boldsymbol{x}_{i_j}\in\pi_k(\boldsymbol{x}_i)}g_{ii_j}=1,i=1,\cdots,n,\end{aligned} \tag{4}$$

式中,$\pi_k(\boldsymbol{x}_i)$表示 $\boldsymbol{x}_i$ 的 k 近邻集合,g_{ii_j}是样本点 $\boldsymbol{x}_{i_j}$重构 $\boldsymbol{x}_i$ 的加权系数,且若 $\boldsymbol{x}_{i_j}\in\pi_k(\boldsymbol{x}_i)$,则 $g_{ii_j}\neq0$;若 $\boldsymbol{x}_{i_j}\notin\pi_k(\boldsymbol{x}_i)$,则 $g_{ii_j}=0$.

优化问题的构造受局部线性嵌入(locally Linear embeeding,LLE)[23]算法的启发,其暗含的意义是让每个样本点可以通过其近邻的线性加权来表示. 这意味着只需要求解 k 个权值,不需要求 $l+u$ 个权值,进而可以提高邻接矩阵的稀疏程度.

2.2　优化问题

对于半监督回归任务,LAG-S3R 的学习目标是确定如下的决策函数:

$$f(\boldsymbol{x})=\phi(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{w}+b, \tag{5}$$

式中,$\boldsymbol{w}$ 和 b 分别表示假设函数的权重向量和阈值,$\phi(\boldsymbol{x})$表示对 $\boldsymbol{x}$ 的非线性映射,令 $\phi(\boldsymbol{x})=\kappa(\boldsymbol{x},\cdot)$. 此时特征空间维度和样本个数相同. 由于高斯核函数的泛化能力较强,所以被广泛使用,本文实验部分也采用了该核函数. 高斯核函数的表达式为:$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\rho\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2)$,其中 $\rho>0$ 是高斯核参数.

为了提高算法的稀疏性,我们向 LAG-S3R 中增加了 1 范数正则项. 稀疏正则是一种常见的诱导稀疏性的技术. 目前已证明 p 范数($0\leqslant p\leqslant1$)可以诱导稀疏性,其中 0 范数是实现稀疏解的最理想正则化手

段. 然而,求解含有 0 范数正则项的优化问题极为复杂. 相比之下,1 范数正则由于能够近似 0 范数并且具备分段可微性质,使得优化包含 1 范数的函数变得可行[14-16,24]. 因此,1 范数正则作为所有稀疏性正则中唯一具有凸性的正则而被广泛使用. 相关的算法有:1 范数支持向量机[14],基于 1 范数的最大化主成分分析[25],基于 1 范数的 Fisher 判别分析[26],基于 1 范数度量的聚类算法[27]等.

基于局部邻接图和 1 范数正则技术,LAG-S3R 构造了如下的优化问题:

$$\begin{aligned}\min_{\boldsymbol{w},b,\boldsymbol{\xi}^{(*)}}\quad & \frac{1}{l}\sum_{i=1}^{l}(\xi_i+\xi_i^*)+C_1(\|\boldsymbol{w}\|_1+b)+C_2\sum_{i,j=1}^{n}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2g_{ij},\\ \text{s.t.}\quad & f(\boldsymbol{x}_i)-y_i\leqslant\varepsilon+\xi_i,\\ & y_i-f(\boldsymbol{x}_i)\leqslant\varepsilon+\xi_i^*,\\ & \xi_i,\xi_i^*\geqslant 0,\ i=1,\cdots,l,\end{aligned}\tag{6}$$

式中,$\|\boldsymbol{w}\|_1$ 表示 $\boldsymbol{w}$ 的 1 范数. 优化问题(6)中共有 3 项:第一项是损失函数项,用来控制经验风险;第二项是模型稀疏的 1 范数正则项,控制了学习机的复杂度;第三项是图正则项,其中的局部邻接图 $\boldsymbol{G}$ 通过求解优化问题(4)计算得到.

2.3 求解算法

2.3.1 优化问题(4)的求解

虽然优化问题(4)是求一个邻接图,但是邻接矩阵 $\boldsymbol{G}$ 的每一行是独立的. 因此,可以分别求解其每一行. 以 $\boldsymbol{G}$ 中的第 i 行为例来考虑 g_i 的求解.

令 $J_1(\boldsymbol{g}_i)=\left\|\boldsymbol{x}_i-\sum_{\boldsymbol{x}_{i_j}\in\pi_k(\boldsymbol{x}_i)}\boldsymbol{x}_{i_j}g_{ii_j}\right\|^2$,于是有:

$$J_1(\boldsymbol{g}_i)=\left\|\sum_{\boldsymbol{x}_{i_j}\in\pi_k(\boldsymbol{x}_i)}g_{ii_j}\boldsymbol{x}_i-\sum_{\boldsymbol{x}_{i_j}\in\pi_k(\boldsymbol{x}_i)}\boldsymbol{x}_{i_j}g_{ii_j}\right\|^2=\left\|\sum_{\boldsymbol{x}_{i_j}\in\pi_k(\boldsymbol{x}_i)}g_{ii_j}(\boldsymbol{x}_i-\boldsymbol{x}_{i_j})\right\|^2=\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{s}_i,\tag{7}$$

式中,$\boldsymbol{q}_i=[(\boldsymbol{x}_i-\boldsymbol{x}_{i_1}),\cdots,(\boldsymbol{x}_i-\boldsymbol{x}_{i_k})]^{\mathrm{T}}$;$\boldsymbol{S}\in\mathbf{R}^{n\times k}$ 中第 i 行第 j 列($i=1,\cdots,n,j=1,\cdots,k$)的元素 s_{ij} 分别表示第 i 个样本与其第 j 个近邻样本的相似度,该矩阵可由邻接矩阵 $\boldsymbol{G}$ 删除其中的 0 元素得到;且有 $\boldsymbol{s}_i=[s_{i1},s_{i2},\cdots,s_{ik}]^{\mathrm{T}}$.

因此,优化问题(4)可以表示为:

$$\begin{aligned}&\min_{\boldsymbol{s}_i}\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{s}_i,\\ &\text{s.t.}\ \sum_{j=1}^{k}s_{ij}=1.\end{aligned}\tag{8}$$

利用拉格朗日乘子法求解上述优化问题,可得到其对应的拉格朗日函数:

$$L(\boldsymbol{s}_i)=\boldsymbol{s}_i^{\mathrm{T}}\boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{s}_i+\beta_i(\mathbf{1}^{\mathrm{T}}\boldsymbol{s}_i-1),\tag{9}$$

式中,β_i 为拉格朗日乘子. 对上式(9)求偏导,并令其偏导值为 0,得到:

$$\frac{\partial L(\boldsymbol{s}_i)}{\partial\boldsymbol{s}_i}=2\boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{s}_i+\beta_i\mathbf{1}=\mathbf{0},\tag{10}$$

进而可将 $\boldsymbol{s}_i$ 表示为:

$$\boldsymbol{s}_i=-\frac{\beta_i}{2}(\boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}})^{-1}\mathbf{1}.\tag{11}$$

利用 $\sum_{j=1}^{k}s_{ij}=1$ 对 $\boldsymbol{s}_i$ 归一化,可得最终的 $\boldsymbol{s}_i$ 计算公式:

$$\boldsymbol{s}_i=\frac{(\boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}})^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}(\boldsymbol{q}_i\boldsymbol{q}_i^{\mathrm{T}})^{-1}\mathbf{1}}.\tag{12}$$

利用上式求出矩阵 $\boldsymbol{S}$ 后,通过在该矩阵对应位置添加 0 元素即可获得我们所需的邻接矩阵 $\boldsymbol{G}$.

2.3.2 优化问题(6)的求解

为了方便对优化问题(6)的求解,引入两个非负的辅助向量($\boldsymbol{w}_+$及 $\boldsymbol{w}_-$)和两个非负的辅助常数变量(b_+及 b_-),进而可以将 $\boldsymbol{w}$ 和 b 分别表示为:

$$\begin{cases} \boldsymbol{w}=\boldsymbol{w}_+-\boldsymbol{w}_- & \boldsymbol{w}_+\geqslant\boldsymbol{0},\boldsymbol{w}_-\geqslant\boldsymbol{0} \\ b=b_+-b_- & b_+\geqslant 0,b_-\geqslant 0 \end{cases}, \tag{13}$$

将式(13)代入式(5),可以得到:

$$f(\boldsymbol{x})=\phi(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{w}_+-\boldsymbol{w}_-)+(b_+-b_-), \tag{14}$$

同时,优化问题(6)可以表示为:

$$\begin{aligned} \min_{\boldsymbol{w}_\pm,b_\pm,\boldsymbol{\xi}^{(*)}}\ & \frac{1}{l}\sum_{i=1}^{l}(\xi_i+\xi_i^*)+C_1((\boldsymbol{w}_+ +\boldsymbol{w}_-)+(b_++b_-))+C_2\sum_{i,j=1}^{n}(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2 g_{ij}, \\ \text{s.t.}\quad & f(\boldsymbol{x}_i)-y_i\leqslant\varepsilon+\xi_i, \\ & y_i-f(\boldsymbol{x}_i)\leqslant\varepsilon+\xi_i^*, \\ & \xi_i,\xi_i^*\geqslant 0,\ i=1,\cdots,l, \\ & w_{+_j},w_{-_j}\geqslant 0,\ j=1,\cdots,n, \\ & b_+,b_-\geqslant 0. \end{aligned} \tag{15}$$

为了避免模型的稀疏性在求解过程中丢失,我们在原空间直接求解优化问题(15). 这样才可能保证模型系数 $\boldsymbol{w}$ 具有稀疏性,也就是说该向量中含有为零的分量.

定义

$$\boldsymbol{u}=[(\boldsymbol{w}_+)^{\mathrm{T}},(\boldsymbol{w}_-)^{\mathrm{T}},b_+,b_-,\boldsymbol{\xi}^{\mathrm{T}},(\boldsymbol{\xi}^*)^{\mathrm{T}}]^{\mathrm{T}}, \tag{16}$$

$$\boldsymbol{p}=\left[C_1\boldsymbol{1}^{\mathrm{T}},C_1\boldsymbol{1}^{\mathrm{T}},C_1,C_1,\frac{\boldsymbol{1}^{\mathrm{T}}}{l},\frac{\boldsymbol{1}^{\mathrm{T}}}{l}\right]^{\mathrm{T}}, \tag{17}$$

$$\boldsymbol{A}_1=[\boldsymbol{K}_l,-\boldsymbol{K}_l,\boldsymbol{1},-\boldsymbol{1},-\boldsymbol{1},\boldsymbol{0}], \tag{18}$$

$$\boldsymbol{A}_2=[-\boldsymbol{K}_l,-\boldsymbol{K}_l,-\boldsymbol{1},\boldsymbol{1},\boldsymbol{0},-\boldsymbol{1}], \tag{19}$$

$$\boldsymbol{A}=[\boldsymbol{A}_1^{\mathrm{T}},\boldsymbol{A}_2^{\mathrm{T}}]^{\mathrm{T}}, \tag{20}$$

$$\boldsymbol{h}=[\varepsilon+\boldsymbol{y}^{\mathrm{T}},\varepsilon-\boldsymbol{y}^{\mathrm{T}}]^{\mathrm{T}}, \tag{21}$$

$$\boldsymbol{V}=\begin{bmatrix} 2C_2\boldsymbol{K}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{K} & -2C_2\boldsymbol{K}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{K} & \boldsymbol{0} \\ -2C_2\boldsymbol{K}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{K} & 2C_2\boldsymbol{K}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{K} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \tag{22}$$

式中,矩阵 $\boldsymbol{K}_l$ 中的第 i 行($i=1,\cdots,l$)为 $\boldsymbol{\kappa}_i=[\kappa(\boldsymbol{x}_1,\boldsymbol{x}_i),\kappa(\boldsymbol{x}_2,\boldsymbol{x}_i),\cdots,\kappa(\boldsymbol{x}_n,\boldsymbol{x}_i)]^{\mathrm{T}}$. 进而可将优化问题(15)表示为:

$$\begin{aligned} &\min_{\boldsymbol{u}}\frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{u}+\boldsymbol{p}^{\mathrm{T}}\boldsymbol{u}, \\ &\text{s.t. } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{u}\geqslant\boldsymbol{h},\boldsymbol{u}\geqslant\boldsymbol{0}. \end{aligned} \tag{23}$$

直接求解优化问题(23),便可获得 LAG-S3R 的决策函数. 对于一个不在 $\boldsymbol{X}$ 中的新样本 $\boldsymbol{x}$,便可通过该决策函数对 $\boldsymbol{x}$ 进行回归预测. LAG-S3R 的稀疏性体现在权重向量 $\boldsymbol{w}$ 上,也就是说,该算法得到的 $\boldsymbol{w}$ 中有较多的零元素. $\boldsymbol{w}$ 中元素值的大小表示对应样本对建模的重要程度. 通过剔除 $\boldsymbol{w}$ 中零元素对应的样本,即可完成样本约简. 算法 1 为 LAG-S3R 的具体步骤.

算法 1　LAG-S3R

输入:训练样本矩阵 $\boldsymbol{X}\in\mathbf{R}^{n\times m}$,其前 l 个样本对应的标签向量 $\boldsymbol{y}\in\mathbf{R}^l$,正则化参数 C_1 和 C_2,最近邻参数 k;

输出:$\boldsymbol{w}$ 和 b.

1. 计算式(12)得到矩阵 $\boldsymbol{S}$;
2. 通过向 $\boldsymbol{S}$ 中对应位置添加 0 元素,进而获得邻接矩阵 $\boldsymbol{G}$;
3. 通过 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{G}$ 计算得到拉普拉斯矩阵 $\boldsymbol{L}$;
4. 通过式(17)~式(22)得到 $\boldsymbol{p}$、$\boldsymbol{A}$、$\boldsymbol{h}$ 和 $\boldsymbol{V}$;
5. 求解优化问题(23),得到 $\boldsymbol{u}$;
6. 通过式(16)得到 $\boldsymbol{w}_+$、$\boldsymbol{w}_-$、b_+及 b_-的值;
7. 通过式(13)求出 $\boldsymbol{w}$ 和 b.

3 实验分析

为了验证 LAG-S3R 算法的回归预测和样本约简性能,我们在 9 个真实数据集上进行了实验. 实验在配备有 2.8 GHz Intel 核与 16GB 内存的 MacOS 系统的个人电脑上完成,并使用 Matlab R2016b 软件编程.

3.1 实验设置

3.1.1 评价指标

实验部分我们使用了两种评价指标:均方根误差(root mean squared error,RMSE)和稀疏度(degree of sparsity,SD). 前一个指标用来衡量回归算法的预测性能,指标的值越小表示模型性能越好;后一个指标用来计算模型的稀疏程度,稀疏度的值越大表示模型越稀疏. 下面分别对这两个指标进行介绍.

均方根误差:该指标可以表示真实值与预测结果之间的平均平方差,用公式表示为:

$$RMSE(y_i,f(\boldsymbol{x}_i))=\sqrt{\frac{1}{n'}\sum_{i=1}^{n'}(y_i-f(\boldsymbol{x}_i))^2}, \tag{24}$$

式中,n'是测试样本的个数.

稀疏度:该指标用来衡量模型的权重向量 $\boldsymbol{w}$ 中零元素的个数. 用公式表示为:

$$SD(\boldsymbol{w})=\frac{length(\boldsymbol{w})-\|\boldsymbol{w}\|_0}{length(\boldsymbol{w})}, \tag{25}$$

式中,$length(\boldsymbol{w})$表示向量 $\boldsymbol{w}$ 的长度,也就是其元素的个数,$\|\boldsymbol{w}\|_0$ 表示 $\boldsymbol{w}$ 的 0 范数.

3.1.2 对比方法

本文将所提算法与 6 种相关算法进行了比较,包括:

• $\text{COREG}_{\text{kNN}}$[11] 和 $\text{COREG}_{\text{SVR}}$[10]:这两种算法都是基于 COREG 的半监督回归算法. 它们的不同之处在于 $\text{COREG}_{\text{kNN}}$和 $\text{COREG}_{\text{SVR}}$分别使用了 kNN 和 SVR 进行回归预测.

• SS-SVR[28]:SS-SVR(semi-supervised SVR)是半监督支持向量回归机的简称. 该算法同 COREG 一样,都是基于协同训练的半监督回归算法.

• LapSVR[12] 和 LapRLSR[13]:这两种算法都是基于图的半监督回归算法. 两种算法的主要不同点:LapSVR 使用了 ε 不敏感损失函数,而 LapRLSR 使用了最小二乘损失.

• NILP[29]:NILP(Node influence-based label propagation)是基于节点影响标签传播算法的简称. 利用该标签传播算法,可以为训练集中的无标签样本提供标签值,然后使用 SVR 进行回归学习.

3.1.3 数据集

实验中我们共采用了来自于 4 个网站的 9 个公开数据集,包括来自 Kaggle 网站的 Boston housing、Garments worker productivity、Happiness Index 2018-2019、Marketing,来自 StatLib 网站的 Bodyfat,来自 Hastie 教授个人主页的 Los Angeles ozone,以及来自 UCI 数据集库的 Concrete slump test、Real estate valuation、Servo. 这些数据集的详细信息如表 1 所示.

表 1 9 个真实数据集的信息

Table 1 Description of nine benchmark datasets

数据集	数据集名称	样本数	特征数	数据集	数据集名称	样本数	特征数
Dataset 1	Boston housing[a]	506	13	Dataset 6	Servo[d]	167	4
Dataset 2	Bodyfat[b]	252	14	Dataset 7	Garments worker productivity[a]	691	13
Dataset 3	Los Angeles ozone[c]	330	9	Dataset 8	Happiness Index 2018-2019[a]	312	9
Dataset 4	Concrete slump test[d]	103	10	Dataset 9	Marketing[a]	171	3
Dataset 5	Real estate valuation[d]	414	6				

[a]下载自 https://www.kaggle.com;[b]下载自 https://lib.stat.cmu.edu/datasets/;[c]下载自 https://hastie.su.domains/ElemStatLearn/data.html;[d]下载自 https://archive.ics.uci.edu/ml/index.php.

3.1.4 参数设置

LAG-S3R 一共包含了 4 类参数:正则化参数、最近邻个数 k、ε 不敏感损失函数的参数 ε 和高斯核参数 ρ. 对于 LAG-S3R 和其他几种算法,我们皆在 $\{10^{-6},10^{-5},\cdots,10^{6}\}$ 的范围内确定其正则化参数的取值;LAG-S3R 的最近邻个数 k 在 $\{3,7,15,20\}$ 的范围内选取;对于所有使用 ε 不敏感损失函数的算法,我们都

令 $\varepsilon=0.01$;高斯核参数根据中值法[30]确定.

对比算法中涉及的其他参数均根据各自相应文献所推荐的方法进行设定. 对所有算法,本文均采用 5 倍交叉验证方法进行实验,取 5 次最优结果的平均值作为最终结果.

3.2　实验结果分析

3.2.1　半监督性能比较

为了验证算法在半监督回归任务上的性能,我们从训练集中随机挑选 10%的样本作为有标签样本,从剩余的训练样本中分别随机挑选 30%、60%、90%的样本为无标签样本.

图 1 展示了 LAG-S3R 及其他 6 种相关算法在数据集中含有 30%、60%、90%无标签样本情况下获得的 RMSE 变化曲线. 从图中可以得到如下结论:

(1)随着无标签比例的增加,算法的 RMSE 曲线在大部分情况下呈现出下降趋势,这也说明了标签信息对于回归预测算法性能的影响是正面的;

(2)NILP 算法的性能较差,这可能是因为该算法使用的邻接矩阵不能很好地表示数据的结构信息;

(3)LapSVR 和 LapRLSR 的表现相当,且优于 NILP. 在某些情况下,这两个算法能获得最优或次优的结果,但其总体性稍差于 COREG 和 LAG-S3R;

(4)两种 COREG 算法虽然在一些数据集上表现出了最优的回归预测性能,但总体来说仍稍逊于本文提出的 LAG-S3R 算法;

(5)LAG-S3R 算法在大部分情况下都获得了最优的回归预测结果,这也表明了局部邻接图和 1 范数正则项的引入,对算法的性能提升是有帮助的.

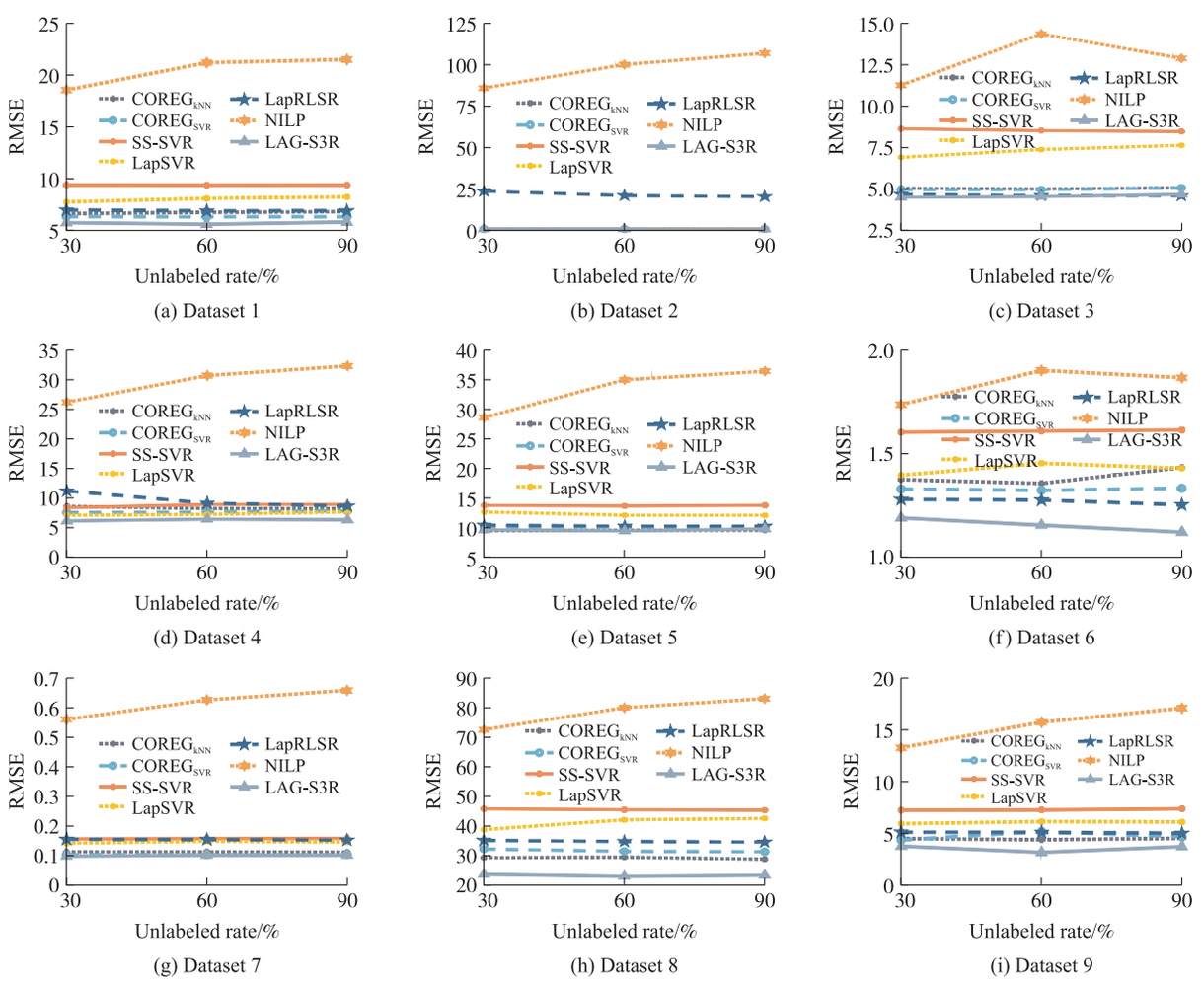


图 1　7 种方法在 9 个含有不同无标签比例的真实数据集上获得的 RMSE 值

Fig. 1　RMSE obtained by seven methods over nine benchmark datasets with different unlabeled rates

3.2.2 稀疏性能比较

为了验证回归算法的稀疏性,我们在训练样本中含有 90%无标签样本情况下,向训练样本中增加等同于有标签样本数量的 20%、40%、60%的样本作为噪声样本. 这些噪声样本服从均值为 0.5 方差为 0.25 的高斯分布. 对于这些噪声特征,其中一半数量的样本为有标签样本(其标签值服从$[y_{\min},y_{\max}]$的均匀分布,其中$y_{\min}$和$y_{\max}$分别表示原有有标签样本的最小和最大标签值),剩余一半数量的样本为无标签样本.

图 2 显示了几种算法在含有不同比例噪声样本情况下获得的 RMSE 结果. 从图 2 可以看出:(1)LapSVR 性能受噪声影响较大,在 Dataset 3 和 Dataset 5 上表现出了最差的回归性能. 这在一定程度上也暗示了一个稀疏的模型或更合理的图生成方法对模型的重要性;(2)NILP 的总体性能较差,其次是 LapSVR 和 LapRLSR;(3)LAG-S3R 的性能最优,两种 COREG 算法的性能次之.

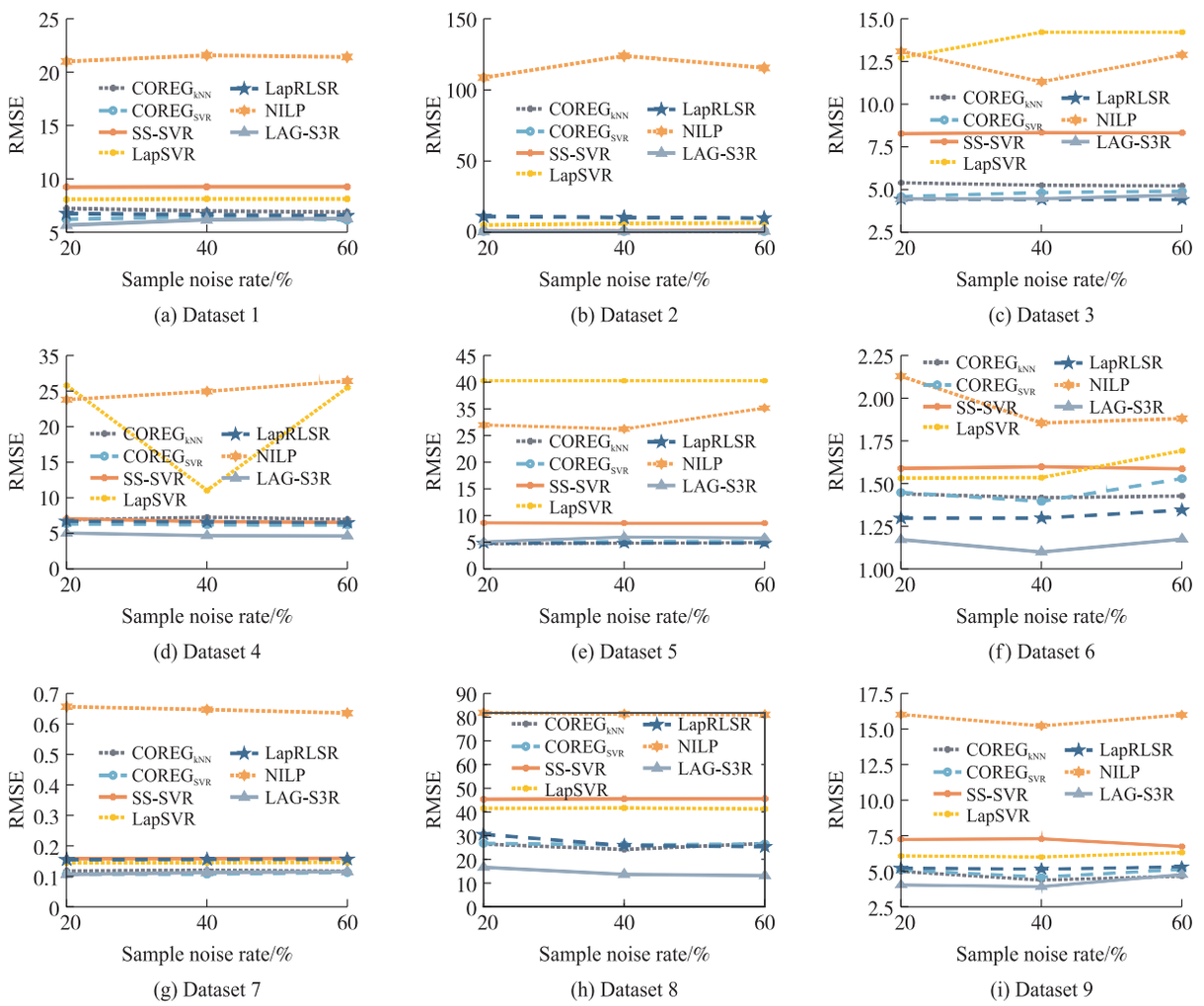


图 2　7 种方法在 9 个含有不同比例的样本噪声的真实数据集上获得的 RMSE 值

Fig. 2　RMSE obtained by seven methods over nine benchmark datasets with different sample noise rates

为了更直观地显示模型的稀疏程度,我们定义了如式(25)所示的稀疏度评价指标. 在实验中,如果模型的权重向量中某一元素的绝对值小于10^{-7},则将其值设置为 0. 表 2 显示了几种算法在数据集中含有 60%噪声样本的情况下表现出的稀疏度,表中加粗的数字显示了在每个数据集上的最优稀疏度. 从表中可以看到:(1)COREG$_{kNN}$在所有数据集上都没有表现出稀疏性,这是因为该方法使用 kNN 进行回归预测,而 kNN 算法不具有稀疏性;(2)COREG$_{SVR}$、SS-SVR 和 NILP 表现出了较弱的稀疏性,这些算法使用 SVR 算法进行回归预测,而 SVR 模型的建立和支持向量有关,具有一定的稀疏性;(3)LapSVR 和 LapRLSR 的稀疏性较差,而本文提出的 LAG-S3R 具有最优的稀疏性,这也验证了 1 范数正则的引入能有效提高模型的稀疏性.

表 2　7 种算法在 9 个含有 60%样本噪声的真实数据集上获得的稀疏度及其方差

Table 2　DS obtained by seven methods over nine benchmark datasets with 60% sample noise rate

数据集名称	$COREG_{kNN}$	$COREG_{SVR}$	SS-SVR	LapSVR	LapRLSR	NILP	LAG-S3R
Dataset 1	0.00±0.00	0.03±0.01	0.09±0.06	0.00±0.00	0.00±0.00	0.06±0.14	**0.35±0.48**
Dataset 2	0.00±0.00	0.03±0.04	0.15±0.08	0.00±0.00	0.00±0.00	0.00±0.00	**0.86±0.08**
Dataset 3	0.00±0.00	0.01±0.01	0.02±0.03	0.00±0.00	0.11±0.25	0.01±0.13	**0.29±0.15**
Dataset 4	0.00±0.00	0.06±00.04	0.00±0.00	0.00±0.00	0.00±0.00	0.01±0.01	**0.45±0.39**
Dataset 5	0.00±0.00	0.04±0.03	0.02±0.04	0.00±0.00	0.00±0.00	0.01±0.01	**0.19±0.16**
Dataset 6	0.00±0.00	0.03±0.04	0.12±0.08	0.20±0.45	0.00±0.00	0.01±0.01	**0.77±0.06**
Dataset 7	0.00±0.00	0.08±00.01	0.43±0.29	0.01±0.01	0.01±0.01	0.24±0.26	**0.91±0.03**
Dataset 8	0.00±0.00	0.02±0.02	**0.11±0.05**	0.00±0.00	0.03±0.05	0.00±0.00	**0.11±0.01**
Dataset 9	0.00±0.00	0.04±0.03	0.08±0.02	0.00±0.00	0.00±0.00	0.01±0.01	**0.57±0.23**

4　结论

本文设计了一种新的图生成方法——局部邻接图. 该方法在构造邻接矩阵时保留了样本的局部流形结构,可以在一定程度上减少噪声或离群点对算法性能的影响. 在该图生成方法的基础上,我们使用了 1 范数正则项,进而提出了一种新的半监督回归学习算法:基于局部邻接图的半监督稀疏回归算法. 该算法利用 1 范数能诱导稀疏的特性,使得学习得到的模型具有良好的稀疏性. 我们在 9 个真实数据集上将 LAG-S3R 与其他 6 种相关算法进行了比较. 为了验证算法的半监督回归学习性能,我们使得每个数据集中分别仅有 30%、60%、90%的无标签样本;为了验证算法的稀疏性,我们随机对每个数据集增加了 20%、40%、60%的噪声样本. 实验结果均验证了所提算法的有效性和良好的稀疏性.

[参考文献]

[1] VAN ENGELEN J E,HOOS H H. A survey on semi-supervised learning[J]. Machine learning,2020,109(2):373-440.

[2] 邓廷权,王强. 半监督类保持局部线性嵌入方法[J]. 智能系统学报,2021,16(1):98-107.

[3] 翟德明,沈斯娴,周雄,等. 基于最优传输理论的深度半监督学习伪标签生成算法[J]. 软件学报:1-14[2024-06-28]. https://doi.org/10.13328/j.cnki.jos.007054.

[4] 胡声丹,苗夺谦,姚一豫. 基于三支标签传播的半监督属性约简[J]. 计算机学报,2021,44(11):2332-2343.

[5] KOSTOPOULOS G,KARLOS S,KOTSIANTIS S,et al. Semi-supervised regression:a recent review[J]. Journal of intelligent & fuzzy systems,2018,35(2):1483-1500.

[6] 郭兰哲,李宇峰. 稳健选择伪标注的混合式半监督学习[J]. 中国科学:信息科学,2024,54(3):623-637.

[7] AZRIEL D,BROWN L D,SKLAR M,et al. Semi-supervised linear regression[J]. Journal of the American Statistical Association,2022,117(540):2238-2251.

[8] CHEN X,CAO W,GAN C,et al. Semi-supervised support vector regression based on data similarity and its application to rock-mechanics parameters estimation[J]. Engineering applications of artificial intelligence,2021,104:104317.

[9] 杨智勇,许倩倩,何源,等. 半监督 AUC 优化的 Boosting 算法及理论[J]. 计算机学报,2022,45(8):1598-1617.

[10] ZHOU Z,LI M. Semi-supervised regression with cotraining-style algorithms[J]. IEEE transactions on knowledge and data engineering,2007,19(11):1479-1493.

[11] ZHOU Z,LI M. Semi-supervised regression with co-training[C]//Proceedings of International Joint Conference on Artificial Intelligence. San Francisco:Margan Kaufmann,2005:908-913.

[12] YOO J,KIM H J. Semisupervised location awareness in wireless sensor networks using Laplacian support vector regression[J]. International journal of distributed sensor networks,2014,10(4):265801.

[13] YANG H,QIN F,LIANG Q L,et al. LapRLSR for NIR spectral modeling and its application to online monitoring of the column separation of Salvianolate[J]. Chinese chemical letters,2007,18(7):852-856.

[14] ZHANG L,ZHOU W. On the sparseness of 1-norm support vector machines[J]. Neural networks,2010,23(3):373-385.

[15] 郑晓晗. 稀疏与鲁棒的孪生支持向量机算法研究[D]. 苏州:苏州大学,2021.

[16] ZHENG X,ZHANG L,XU Z. L1-norm Laplacian support vector machine for data reduction in semi-supervised learning[J]. Neural computing and applications,2023,35(17):12343-12360.

[17] ZHENG X, ZHANG L, YAN L. Sparse discriminant twin support vector machine for binary classification[J]. Neural computing and applications,2022,34(19):16173-16198.
[18] WANG J, GUO S, NIE F, LI X. Local-global fuzzy clustering with anchor graph[J]. IEEE transactions on fuzzy systems, 2024,32(1):188-202.
[19] ZHU J, CHEN X, YANG H, et al. Unsupervised adaptive bipartite graph embedding[J]. IEEE transactions on knowledge and data engineering,2023,35(10):10514-10525.
[20] CHEN L, ZHONG Z. Adaptive and structured graph learning for semi-supervised clustering[J]. Information processing & management,2022,59(4):102949.
[21] 李静,陈秀宏. 基于竞争性协同表示的局部判别投影特征提取[J]. 智能系统学报,2019,14(5):974-981.
[22] SMOLA A J, SCHOLKOPF B. A tutorial on support vector regression[J]. Statistics and computing,2004,14:199-222.
[23] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science,2000,290(5500): 2323-2326.
[24] 陈聪. L1 正则化与 pinball 损失函数的极限学习机[J]. 信息技术与信息化,2023(3):37-40.
[25] KWAK N. Principal component analysis based on L1-norm maximization[J]. IEEE transactions on pattern analysis and machine intelligence,2008,30(9):1672-1680.
[26] WANG H, LU X, HU Z, ZHENG W. Fisher discriminant analysis with L1-norm[J]. IEEE transactions on cybernetics,2013, 44(6):828-842.
[27] 杨红鑫,杨绪兵,寇振宇,等. 基于 L1 范数的 k 平面聚类算法设计[J]. 南京航空航天大学学报,2019,51(5):681-686.
[28] WANG X, MA L, WANG X. Apply semi-supervised support vector regression for remote sensing water quality retrieving[C]// Proceedings of IEEE International Geoscience and Remote Sensing Symposium, Piscataway, NJ:IEEE,2010:2757-2760.
[29] HUA Z, YANG Y, QIU H. Node influence-based label propagation algorithm for semi-supervised learning[J]. Neural computing and applications,2020,33(7):2753-2768.
[30] ZHANG L, ZHOU W, CHANG P, et al. Kernel sparse representation-based classifier[J]. IEEE transactions on signal processing,2011,60(4):1684-1695.

[责任编辑:陆炳新]

(上接第 95 页)

[15] ZHAO Y, LV W, XU S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA:IEEE,2024:16965-16974.
[16] TAN M, PANG R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. ,Seattle, WA, USA:IEEE,2020:10781-10790.
[17] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canda:IEEE,2023:7464-7475.
[18] WANG A, CHEN H, LIU L, et al. YOLOv10:Real-Time End-to-End Object Detection[EB/OL]. (2024-5-23)[2024-5-18]. https://arxiv.org/abs/2405.14458.

[责任编辑:陆炳新]