软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(7):3206−3225 [doi: 10.13328/j.cnki.jos.006511]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于实例加权和双分类器的稳定学习算法*

杨 帅[1,2], 王 浩[1,2], 俞 奎[1,2], 曹付元[3]

[1](大数据知识工程教育部重点实验室 (合肥工业大学), 安徽 合肥 230601)

[2](合肥工业大学 计算机与信息学院, 安徽 合肥 230601)

[3](山西大学 计算机与信息技术学院, 山西 太原 030006)

通信作者: 俞奎, E-mail: yukui@hfut.edu.cn

摘 要: 稳定学习的目标是利用单一的训练数据构造一个鲁棒的预测模型, 使其可以对任意与训练数据具有相似分布的测试数据进行精准的分类. 为了在未知分布的测试数据上实现精准预测, 已有的稳定学习算法致力于去除特征与类标签之间的虚假相关关系. 然而, 这些算法只能削弱特征与类标签之间部分虚假相关关系并不能完全消除虚假相关关系; 此外, 这些算法在构建预测模型时可能导致过拟合问题. 为此, 提出一种基于实例加权和双分类器的稳定学习算法, 所提算法通过联合优化实例权重和双分类器来学习一个鲁棒的预测模型. 具体而言, 所提算法从全局角度平衡混杂因子对实例进行加权来去除特征与类标签之间的虚假相关关系, 从而更好地评估每个特征对分类的作用. 为了完全消除数据中部分不相关特征与类标签之间的虚假相关关系以及弱化不相关特征对实例加权过程的干扰, 所提算法在实例加权之前先进行特征选择筛除部分不相关特征. 为了进一步提高模型的泛化能力, 所提算法在训练预测模型时构建两个分类器, 通过最小化两个分类器的参数差异来学习一个较优的分类界面. 在合成数据集和真实数据集上的实验结果表明了所提方法的有效性.

关键词: 实例加权; 特征选择; 分布变化; 稳定学习

中图法分类号: TP18

中文引用格式: 杨帅, 王浩, 俞奎, 曹付元. 基于实例加权和双分类器的稳定学习算法. 软件学报, 2023, 34(7): 3206–3225. http://www.jos.org.cn/1000-9825/6511.htm

英文引用格式: Yang S, Wang H, Yu K, Cao FY. Stable Learning via Sample Reweighting and Dual Classifiers. Ruan Jian Xue Bao/Journal of Software, 2023, 34(7): 3206–3225 (in Chinese). http://www.jos.org.cn/1000-9825/6511.htm

Stable Learning via Sample Reweighting and Dual Classifiers

YANG Shuai[1,2], WANG Hao[1,2], YU Kui[1,2], CAO Fu-Yuan[3]

[1](Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Hefei 230601, China)

[2](School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China)

[3](School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract: Stable learning aims to leverage the knowledge obtained only from a single training data to learn a robust prediction model for accurately predicting label of the test data from a different but related distribution. To achieve promising performance on the test data with agnostic distributions, existing stable learning algorithms focus on eliminating the spurious correlations between the features and the class variable. However, these algorithms can only weaken part of the spurious correlations between the features and the class variable, but can not completely eliminate the spurious correlations. Furthermore, these algorithms may encounter the overfitting problem in learning the prediction model. To tackle these issues, this study proposes a sample reweighting and dual classifiers based stable learning algorithm,

* 基金项目: 国家重点研发计划 (2020AAA0106100); 国家自然科学基金 (61876206); 智能信息处理山西省重点实验室开放课题 (CICIP2020003)
收稿时间: 2021-07-10; 修改时间: 2021-08-21; 采用时间: 2021-10-14; jos 在线出版时间: 2022-12-16
CNKI 网络首发时间: 2022-12-19

which jointly optimizes the weights of samples and the parameters of dual classifiers to learn a robust prediction model. Specifically, to estimate the effects of all features on classification, the proposed algorithm balances the distribution of confunders by learning global sample weights to remove the spurious correlations between the features and the class variable. In order to eliminate the spurious correlations between some irrelevant features and the class variable and weaken the influence of irrelevant features on the weighting process of samples, the proposed algorithm selects and removes some irrelevant features before sample reweighting. To further improve the generalization ability of the model, the algorithm constructs two classifiers and learns a prediction model with an optimal hyperplane by minimizing the parameter difference between the two classifiers during learning the prediction model. Using synthetic and real-world datasets, the experiments have validated the effectiveness of the proposed algorithm.

Key words: sample reweighting; feature selection; distribution shift; stable learning

1 引 言

机器学习技术已经被广泛应用于语音识别[1–3]、图像分类[4,5]、目标检测[6–8]等诸多领域. 传统的机器学习算法大部分建立在训练数据和测试数据独立同分布的假设上. 然而, 在一些实际应用场景中, 例如图像分类、语音识别、推荐系统等, 这个假设经常被违背, 导致这些算法性能显著下降. 以语音识别为例, 说话语速的快慢、使用不同地方方言讲话都可能导致训练数据和测试数据的分布存在偏移, 致使基于一些用户语音信息构建的语音识别系统可能不能很好地识别新用户的语音信息.

为了解决上述问题, 迁移学习[9]放宽了训练数据和测试数据独立同分布的假设. 然而, 迁移学习算法需要依赖于未标记的测试数据来缩小领域差异. 迁移学习算法需要将训练数据和未标记的测试数据放在一起训练来学习预测模型. 但在一些实际应用场景中, 例如在线电影评论、医学应用、天气预测等, 数据往往是动态产生的, 测试数据很难获取甚至模型训练时测试数据是未知的, 这给已有的迁移学习算法带来了巨大的挑战. 例如, 在医学应用中, 不同患者的数据分布往往存在偏移[10], 预先收集每个新患者的数据不太可行. 为了缓解这个问题, 稳定学习[11,12]放宽了训练模型时未标记测试数据可用这个条件. 稳定学习的目标是仅利用单一的训练数据构建一个具有高泛化能力的预测模型, 使其在任意与训练数据具有相似分布的测试数据上都能实现精准的分类.

稳定学习的难点在于测试数据不提供的情况下如何缩小训练数据和测试数据的分布差异. 文献 [11–15] 表明训练数据和测试数据的分布发生偏移时, 去除特征与类标签之间的虚假的相关关系有利于缩小训练数据和测试数据的分布差异. 例如, 在识别“狗”的图像分类任务中, 如果训练数据中含有大量“狗在草地上”的图片, 则背景特征“草地”可能和类标签“狗”存在虚假相关关系; 一旦测试环境发生变化, 如测试数据中含有大量“狗在雪地里”的图片, 这种虚假的相关关系将会恶化预测模型的性能. 去除虚假的相关关系, 仅利用一些和类标签具有真实相关关系的特征 (如狗的鼻子、胡须、脚趾等) 构建预测模型, 即使在数据分布发生偏移时, 该模型也能适应于新数据. 因而稳定学习的核心在于如何去除特征与类标签之间的虚假相关关系. 已有的稳定学习算法主要通过特征选择、实例抽样、实例加权来实现这个目标.

基于特征选择的稳定学习算法[14]通过选取因果特征构建预测模型来去除特征与类标签之间的虚假相关关系. 选择因果特征时需要进行条件独立性测试, 而条件独立性的可靠性依赖于训练数据的实例数量和质量. 实例数量不足、噪音实例都会影响条件独立性测试的可靠性. 基于实例抽样的算法[16]通过从训练数据中选择一个实例子集使协变量相互平衡来减小噪音特征和类标签之间的虚假相关. 实例抽样时为了保证协变量相互平衡可能导致选择的实例子集仅是训练数据中极小的一部分, 未被选择的样本中可能包含很多重要信息, 而这些信息却没有被充分利用. 真实应用中数据往往是复杂的, 导致精准地进行因果特征选择和实例抽样是困难的. 而不精准的因果特征选择和实例抽样有可能导致部分特征与类标签之间仍存在虚假相关关系.

基于实例加权的稳定学习方法主要从两个角度来去除特征与类标签之间的虚假相关关系: 一是通过实例加权迫使特征相互独立来隔离每个特征对分类的影响, 例如 DWR[13]、DVD[17]和 SRDO[18]算法. 这 3 个算法主要针对线性回归模型, 不能很好地处理非线性数据和分类问题. 二是通过实例加权来平衡每个特征所对应的治疗组和对照组的数据分布, 从而评估每个特征对分类的真实作用, 例如 CRLR[11]和 DGBR[12]算法. 然而这两个算法存在以下缺陷: (1) 为了评估某个特征对分类的真实作用, CRLR 和 DGBR 算法需要平衡与该特征相关的混杂因子的数据分

布. 由于不知道哪些特征是混杂因子, CRLR 和 DGBR 算法将其余特征都作为混杂因子. 然而这些混杂因子中存在一些与类标签无关的特征, 这些不相关的特征作为混杂因子时会干扰实例加权的过程, 造成实例权重评估有所偏差, 导致部分不相关的特征和类标签仍然存在虚假的相关关系. (2) 在训练预测模型时只有训练数据可提供, 可能导致这些算法出现过拟合问题, 致使构建的预测模型不具有良好的泛化能力. 如图 1 所示, 在图 1(a) 和图 1(b) 中, 两个分类器都能精准地对训练数据进行分类. 但两个分类器在测试数据上取得较差的性能, 如图 1(c) 所示. 如何设计性能良好的稳定学习算法仍是一个极具挑战性的课题.

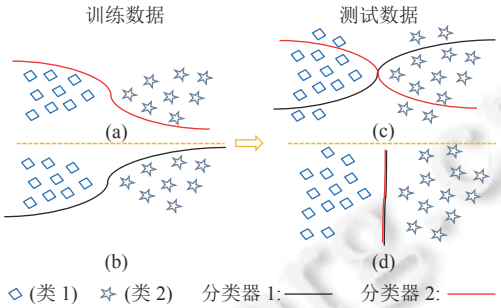


图 1 过拟合的一个示例

为了解决上述问题, 本文提出一种基于实例加权和双分类器的稳定学习算法. 该算法将实例加权和双分类器统一到一个模型中来学习一个鲁棒的分类模型. 该算法首先使用一种无监督学习方式从数据中选择信息最丰富的特征构建子数据集以减少不相关特征对实例加权过程的影响和消除部分不相关特征与类标签之间的虚假相关. 基于所构建的子数据集, 通过从全局角度平衡混杂因子对实例进行加权来去除特征与类标签之间虚假的相关关系从而更好地评估每个特征对分类的作用. 然后根据加权后的实例构建预测模型用于对测试数据进行分类. 为了提升预测模型的泛化能力, 在构建预测模型时, 该算法构建两个分类器, 并用不同的初始化值对这两个分类器进行初始化; 通过最小化两个分类器的参数差异来学习一个较优的分类界面, 如图 1(d) 所示. 合成以及真实数据集上的实验结果表明了所提算法的有效性.

本文第 2 节对稳定学习的相关工作进行总结; 第 3 节简单地介绍了自动编码器; 第 4 节详细介绍了本文提出的基于实例加权和双分类器的稳定学习算法; 第 5 节通过在合成和真实数据集上进行实验来验证所提算法的有效性; 最后对全文进行了总结.

2 相关工作

稳定学习的目标是利用单一的训练数据构建一个鲁棒的预测模型. 已有的稳定学习算法主要分为 3 类: 基于实例加权、基于实例抽样、基于特征选择的算法.

基于实例加权的算法主要通过实例加权来隔离每个特征对分类的影响, 如 CRLR[11]、DGBR[12]、DWR[13]、DVD[17]和 SRDO[18]算法. CRLR 和 DGBR 算法通过从全局角度平衡混杂因子对实例进行加权以评估每个特征对分类的真实作用, 从而去除虚假的相关关系构建稳定的预测模型. Zhang 等人[19]提出了一种 StableNet 方法, 该方法采用深度学习模型来学习特征表示, 并基于最后一层特征表示, 通过实例加权消除相关和不相关特征表示之间的统计相关性. SRDO 算法[18]通过进行实例加权来解决特征之间共线性问题, 并从理论上证明了实例加权后的线性模型在分布偏移情况下可以取得稳定的预测性能. DWR 算法[13]迫使特征之间相互独立来对实例进行加权. 为了促使特征之间相互独立, DWR 可能导致大量实例的加权权重为 0, 致使很多含有较多信息量的实例没有被充分利用. 为了缓解这个问题, DVD 算法[17]首先根据特征之间相关性对特征进行分组, 然后使不同组之间的特征相互独立, 同组之间的特征不进行特征去相关. SRDO、DWR 和 DVD 算法主要针对线性回归问题.

尽管基于实例加权的算法的性能较好, 但在训练实例数量庞大的场景下, 每个实例的权值的学习将变得困难. 为了缓解这个问题, CVS 算法[14]选择因果特征构建预测模型. CVS 算法首先根据先验知识选择一个因果特征, 然后将所选的因果的特征和剩余特征进行条件独立性测试来选择其他因果特征. 噪音实例以及实例量较小的情况下

可能导致不可靠的条件独立性测试, 从而影响 CVS 算法的性能. 在实际应用中, 收集到的数据中经常含有噪音实例和噪音特征, BSSP 算法[16]从实例抽样的角度出发, 基于部分析因设计理论从训练数据中选取实例子集使协变量相互平衡以减少噪音特征和类标签的虚假相关.

3 预备知识

自动编码器是一个由编码和解码两个阶段构成的 3 层神经网络, 包含一个输入层, 一个或多个隐藏层, 一个输出层. 自动编码器首先使用非线性函数对输入数据 X 进行编码来学习数据的低维非线性表示, 即隐藏层表示; 然后对隐藏层表示进行解码来获取输入数据的重构, 即输出层表示 X'. 通过最小化输入数据 X 和输出数据 X' 的重构误差来优化隐藏层表示. 自动编码器主要用于学习数据的低维表示. 具有 l 个隐藏层的自动编码器可以表示为:

$$\text{编码}: \xi^{(j)}=\sigma(\xi^{(j-1)}W_1^{(j)}+b_1^{(j)}),\ j=1,2,\ldots,l \tag{1}$$

$$\text{解码}: \psi^{(j)}=\sigma(\psi^{(j-1)}W_2^{(j)}+b_2^{(j)}),\ j=1,2,\ldots,l \tag{2}$$

其中, σ 是非线性激活函数 (如 Sigmoid 函数等), $\xi^{(0)}=X$, $\psi^{(0)}=\xi^{(l)}$. $\xi^{(l)}$ 为输入数据的低维表示. $W_1^{(j)}$ 和 $b_1^{(j)}$ 分别是第 j 个编码层的权重矩阵和偏差向量. $W_2^{(j)}$ 和 $b_2^{(j)}$ 分别是第 j 个解码层的权重矩阵和偏差向量. 自动编码器通过学习 $W_1^{(j)}$、$b_1^{(j)}$、$W_2^{(j)}$ 和 $b_2^{(j)}$ 来最小化输入数据 X 和输出数据 X' 的重构误差. 输入数据 X 的低维表示记为 $\xi(X)$. 自动编码器的目标损失函数可以表示为:

$$\frac{1}{n}\|X-X'\|^2+\lambda\sum_{j=1}^{l}\left(\left\|W_1^{(j)}\right\|^2+\left\|b_1^{(j)}\right\|^2+\left\|W_2^{(j)}\right\|^2+\left\|b_2^{(j)}\right\|^2\right) \tag{3}$$

其中, $\|X-X'\|^2$ 为数据重构损失; $\sum_{j=1}^{l}(\|W_1^{(j)}\|^2+\|b_1^{(j)}\|^2+\|W_2^{(j)}\|^2+\|b_2^{(j)}\|^2)$ 为模型参数正则化项, 用于约束模型复杂度; n 为输入数据的实例数量; λ 为平衡参数.

4 算 法

本节首先给出了稳定学习的问题定义, 其次简述了所提的基于实例加权和双分类器的稳定学习算法的整体框架, 最后介绍了所提算法的具体实施细节.

4.1 问题定义

令 X 表示特征变量空间, Y 表示标签变量. 基于 X 和 Y 的一个联合概率分布 P_{XY} 为一个环境. 令 π 为所有环境集合, $D^e=[X^e,Y^e]$ 为从环境 $e\in\pi$ 中收集到的数据集. 不同环境下, 特征变量和标签变量的联合概率分布不同, 即 $P_{XY}^{e_1}\neq P_{XY}^{e_2}$, 其中 e_1, $e_2\in\pi$. 给定训练数据 $D^e=[X^e,Y^e]$, 其中 $X^e\in\mathbb{R}^{n\times d}$ 为特征数据, $Y^e=(y_1,\ldots,y_n)\in\mathbb{R}^{n\times 1}$ 为标签数据, n 为实例数量, d 为特征数量, $y_i\in\{0,1\}$ 是第 i 个实例对应的标签. 稳定学习的目标是仅利用单一的训练数据 D^e 学习一个分类器使其可以精准地预测和训练数据存在数据分布差异的测试数据 $X^{e'}$, $e'\in\pi$.

假定 $X=\{S,V\}$. 令 S 表示稳定特征集合, $V=X/S$ 表示噪音特征集合.

假设 1. 存在一个概率函数 $P(y|s)$ 对于任意环境 $e\in\pi$, $\Pr(Y^e=y|S^e=s,V^e=v)=\Pr(Y^e=y|S^e=s)=P(y|s)$.

图 2 给出了 S、V 和 Y 满足假设 1 的 3 种关系, 包括 $S\perp V$、$S\to V$ 和 $V\to S$. 在图 2(b) 和图 2(c) 中, 噪音特征和类标签存在虚假的相关关系, 这种虚假的相关关系在跨领域间是不稳定的. 如果利用噪音特征构建分类器, 一旦测试数据的分布发生偏移, 性能就会显著下降. 根据假设 1, 可以利用稳定特征集合 S 中的特征构建分类器用于稳定预测.

4.2 基于实例加权和双分类器的稳定学习算法

4.2.1 算法框架

本文提出一个基于实例加权和双分类器的稳定学习算法. 该算法包含两个学习阶段, 如图 3 所示. 阶段 1 选择信息最丰富的特征以消除部分不相关特征与类标签之间的虚假相关关系以及弱化不相关特征对实例加权过程的干扰. 阶段 2 用于实例加权和构建预测模型. 为了构建稳定的预测模型, 所提算法通过实例加权来去除特征和类标

签之间的虚假关系从而更好地评估每个特征对分类的作用. 为了捕捉特征之间的非线性关系以及更好地平衡每个特征所对应的治疗组和对照组的数据分布, 所提算法采用一个自动编码器模型将每个特征所对应的治疗组和对照组数据都映射到一个低维非线性空间, 然后基于治疗组和对照组的低维非线性表示进行实例加权. 此外, 阶段 2 构建了两个分类器, 并通过最小化两个分类器的模型参数误差来学习一个稳定的预测模型.

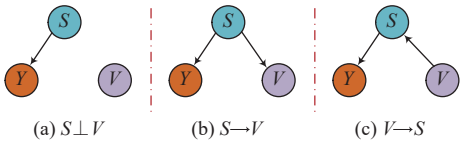


图 2 稳定特征 S, 噪声特征 V 和类标签 Y 的 3 种关系

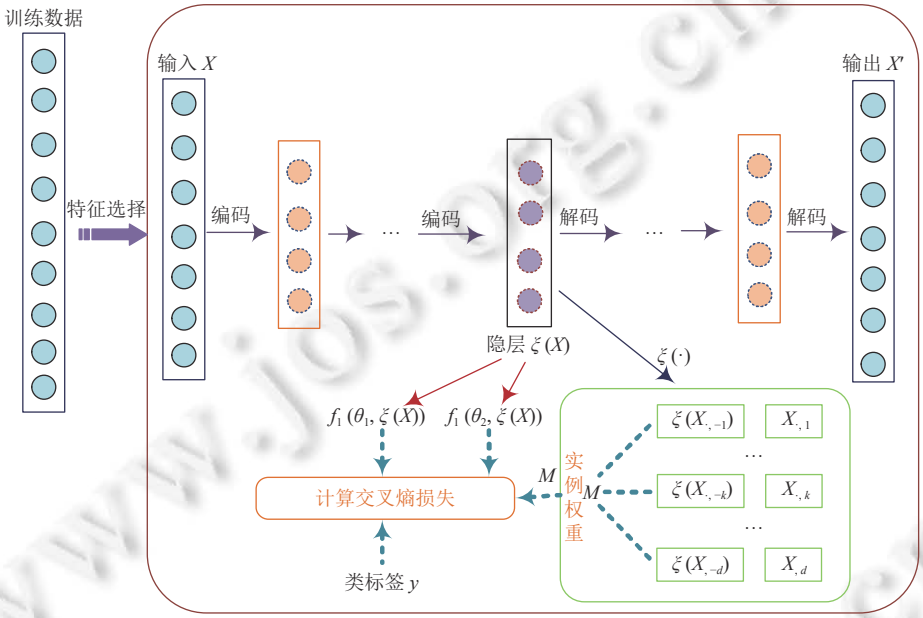


图 3 所提方法的框架图

4.2.2 特征选择

在实际应用中, 收集到的数据往往含有噪音特征. 如果直接利用所收集到的数据学习实例权重来去除特征和类标签之间的虚假相关关系, 噪音特征会干扰实例加权的过程, 从而恶化算法的性能; 此外, 虽然实例加权可以去除部分不相关特征与类标签之间的虚假相关关系, 但是不能完全消除它们之间的虚假相关关系. 一旦训练数据和测试数据的分布发生偏移, 不相关特征和类标签之间存在的微弱的虚假相关关系会使算法的性能下降. 特征选择, 即从原始特征空间中选择相关的、信息丰富的特征子集, 可以从某种程度缓解这个问题. 在特征选择时, 赋予每个实例一样的权重虽然可行但仍存在不足, 因为每个实例包含的信息量不同, 致使每个实例对分类的重要程度不同. 特征选择的过程可能会被信息含量较低的实例所误导, 因此信息较丰富的实例应赋予更高一点的权重. 在特征选择时, 对实例进行加权是有必要的.

同时进行特征选择和实例加权的核心是将它们统一到同一个框架中. 给定选定的特征, 加权后的实例要能够最好地逼近整个训练数据. 同时, 选择的特征要能够最好的保存原始数据的结构信息. 因此, 特征选择和实例加权方法从实例加权的角度被表述为数据重构问题, 从特征选择的角度被表述为结构保持问题. 总的来说, 进行特征选择之后的实例应该具有重建整个数据集的能力. 特征选择的目标函数如下所示:

$$\mathcal{L}=\|X-A^{\mathrm{T}}XW\|_{2,1}+\|A\|_{2,1}+\|W\|_{2,1}+g(A,W) \tag{4}$$

其中, A 和 W 分别为实例权重矩阵和特征权重矩阵. 公式 (4) 中的第 1 项 $\|X-A^{\mathrm{T}}XW\|_{2,1}$ 是数据的重构误差, 这里用 $l_{2,1}$ 范数来衡量数据的重构误差以减少噪音数据的影响. 公式 (4) 中的第 2 项 $\|A\|_{2,1}$ 和第 3 项 $\|W\|_{2,1}$ 分别用于实例

加权和选择特征子集. $l_{2,1}$ 范数结构稀疏正则化项用来约束实例加权矩阵 A, $l_{2,1}$ 范数可以促使行稀疏性. 由于矩阵 A 的每一行对应一个实例, 因此行稀疏性本质上有助于自适应实例重加权. $l_{2,1}$ 范数结构稀疏正则化项也用于约束特征权重矩阵 W. 近年来, 很多工作表明特征降维时保留数据局部几何结构是很有必要的, 公式 (4) 中第 4 项中采用图正则化项 $g(A,W)$ 来实现这个目标. 为了保存数据的局部几何结构信息, 首先利用训练数据中的所有实例构建一个对称矩阵 Z, Z 中元素值为实例之间的相似度, 计算公式如下:

$$Z_{ij}=\begin{cases}\exp\left(-\dfrac{\|x_i-x_j\|_2^2}{2\sigma^2}\right), x_i\in N_k(x_j)\ \text{ or }\ x_j\in N_k(x_i)\\ 0,\ \text{otherwise}\end{cases} \tag{5}$$

其中, x_i 是 X 第 i 个实例, $N_k(x_i)$ 为实例 x_i 的 k 个近邻实例, 本文中 k 的值为 10, 则图正则化项 $g(A,W)$ 可以表示为:

$$g(A,W)=\mathrm{tr}(W^{\mathrm{T}}X^{\mathrm{T}}ALA^{\mathrm{T}}XW) \tag{6}$$

其中, 拉普拉斯矩阵 $L=D-Z$, $D=\sum_j Z_{ij}$.

由于 L 是实对称的, 因此可以将拉普拉斯矩阵 L 分解为: $L=VOV^{\mathrm{T}}$; 则图正则化项可以表示为:

$$g(A,W)=\mathrm{tr}(W^{\mathrm{T}}X^{\mathrm{T}}AVOV^{\mathrm{T}}A^{\mathrm{T}}XW)=\|O^{\frac{1}{2}}V^{\mathrm{T}}A^{\mathrm{T}}XW\|_F^2=\|PW\|_F^2 \tag{7}$$

其中, $P=O^{\frac{1}{2}}V^{\mathrm{T}}A^{\mathrm{T}}X$. 由于 F (Frobenius) 范数对噪音数据敏感, 因此用 $l_{2,1}$ 范数代替 F 范数, 则最终特征选择的目标函数如公式 (8) 所示:

$$\mathcal{L}=\|X-A^{\mathrm{T}}XW\|_{2,1}+\|A\|_{2,1}+\|W\|_{2,1}+\|PW\|_{2,1} \tag{8}$$

通过求解公式 (8) 获取权重矩阵 A 和 W. 根据特征权重矩阵 W 可以选择特征.

求解: 采用迭代的方式求解权重矩阵 A 和 W: (1) 首先固定 A 求解 W; (2) 然后固定 W 求解 A; 交替求解 A 和 W 直到目标函数收敛或达到最大迭代次数.

固定 A 求解 W: 将公式 (8) 关于 W 的导数设为 0, 则有:

$$\frac{\partial\mathcal{L}}{\partial W}=-2X^{\mathrm{T}}AG_1X+2X^{\mathrm{T}}AG_1A^{\mathrm{T}}XW+2G_2W+2P^{\mathrm{T}}G_3PW=0 \tag{9}$$

$$[G_1]_{ii}=\frac{1}{2\|X_i-A^{\mathrm{T}}X_iW\|_F+\varepsilon},\ [G_2]_{ii}=\frac{1}{2\|W\|_F+\varepsilon},\ [G_3]_{ii}=\frac{1}{2\|PW\|_F+\varepsilon} \tag{10}$$

其中, ε 是一个非常小的正数. 如果 G_1、G_2 和 G_3 已知, W 可以由公式 (11) 计算可得:

$$W=(X^{\mathrm{T}}AG_1A^{\mathrm{T}}X+G_2+P^{\mathrm{T}}G_3P)^{-1}X^{\mathrm{T}}AG_1X \tag{11}$$

通过交替更新 G_1、G_2、G_3 和 W 直到收敛来或达到最大迭代次数求解 W. 同理固定 W 可以求解 A.

4.2.3 实例加权

进行特征选择可以减少噪音特征的影响. 然而, 精准地进行特征选择是困难的, 选择的特征子集可能仍含有噪音. 为了评估每个特征对分类的真实作用, 需要对实例进行加权来隔离每个特征对分类的作用.

传统的机器学习算法大都仅考虑特征和类标签之间的统计相关关系. 混杂因子的存在往往导致这些算法错误地评估特征与类标签之间的相关关系. 例如, 夏天冰激凌的售量会大增, 犯罪率也会上升. 冰激凌销售量和犯罪率存在统计相关关系, 但不存在因果关系. 它们之所以存在统计关系, 是因为它们有一个共因, 即天气炎热, 也称为混杂因子[20]. 为了评估某个特征对分类的真实作用, 需要平衡与该特征相关的混杂因子的数据分布. 由于不知道哪些特征是混杂因子, 可以将其余特征都作为混杂因子. 具体而言, 首先在数据集中去除该特征, 并构造一个由其余特征组成的新数据集. 然后, 根据实例在该特征上的取值将新数据集中的实例划分为两组: 治疗组和对照组. 如果实例在该特征上的取值为 1 则划分到治疗组; 否则划分到对照组. 为了精准估计该特征对分类的真实作用, 可以通过调整实例权重来平衡治疗组和对照组之间的数据分布, 如下所示:

$$\left\|\frac{\sum\limits_{i:\,E_i=1}M_i\cdot x_i}{\sum\limits_{i:\,E_i=1}M_i}-\frac{\sum\limits_{i:\,E_i=0}M_i\cdot x_i}{\sum\limits_{i:\,E_i=0}M_i}\right\|_F^2 \tag{12}$$

其中, M 为实例权重; M_i 为第 i 个实例的加权权重; x_i 为 X 第 i 个实例; E 是特征, $E_i \in \{0,1\}$ 表示实例 x_i 在特征 E 的取值; $\sum_{i:\,E_i=1} M_i \cdot x_i / \sum_{i:\,E_i=1} M_i$ 和 $\sum_{i:\,E_i=0} M_i \cdot x_i / \sum_{i:\,E_i=0} M_i$ 分别是治疗组和对照组的一阶矩. 这里使用一阶矩差异来衡量数据分布差异. 通过学习实例权重 M 来最小化治疗组和对照组的一阶矩差异来评估特征 E 对分类的作用.

为了评估每个特征对分类的真实作用, 需要平衡每个特征所对应的治疗组和对照组的数据分布. 为此, 所提算法从全局角度出发学习一组权重, 使其可以对齐每个特征对应的治疗组和对照组的数据分布, 如下所示:

$$\sum_{k=1}^{d}\left\|\frac{X_{\cdot,-k}^{\mathrm{T}}\cdot(M\odot X_{\cdot,k})}{M^{\mathrm{T}}\cdot X_{\cdot,k}}-\frac{X_{\cdot,-k}^{\mathrm{T}}\cdot(M\odot(1-X_{\cdot,k}))}{M^{\mathrm{T}}\cdot(1-X_{\cdot,k})}\right\|_F^2 \tag{13}$$

其中, $X_{\cdot,k}$ 为输入数据 X 第 k 个特征变量; $X_{\cdot,-k}=X\backslash\{X_{\cdot,k}\}$ 表示去除第 k 个特征变量后剩下所有特征变量构成的数据矩阵. 如果 $X_{\cdot,k}$ 在第 i 个实例上的取值为 1, 则 $X_{\cdot,-k}$ 中的第 i 个实例划分到治疗组; 如果 $X_{\cdot,k}$ 在第 i 个实例上的取值为 0, 则 $X_{\cdot,-k}$ 中的第 i 个实例划分到对照组. d 为原始特征空间特征的维度; $\odot$ 表示 Hadamard 乘积.

特征之间往往存在非线性关系, 而且数据中往往含有噪音, 其会干扰治疗组和对照组数据分布的平衡. 由于自动编码器在学习特征之间的非线性关系和压缩噪音方面具有优势, 所提算法采用自动编码器将治疗组和对照组数据映射到一个低维非线性空间, 然后再平衡它们之间的数据分布, 如下所示:

$$\sum_{k=1}^{d}\left\|\frac{\xi(X_{\cdot,-k})^{\mathrm{T}}\cdot(M\odot X_{\cdot,k})}{M^{\mathrm{T}}\cdot X_{\cdot,k}}-\frac{\xi(X_{\cdot,-k})^{\mathrm{T}}\cdot(M\odot(1-X_{\cdot,k}))}{M^{\mathrm{T}}\cdot(1-X_{\cdot,k})}\right\|_F^2 \tag{14}$$

在平衡第 k 个特征变量所对应的治疗组和对照组数据分布时, 首先使用自动编码器将 $X_{\cdot,-k}$ 映射到低维非线性空间上, 然后根据第 k 个特征变量 $X_{\cdot,k}$ 在实例上的取值将映射后的数据 $\xi(X_{\cdot,-k})$ 划分为两组. 如果 $X_{\cdot,k}$ 在第 i 个实例上的取值为 1, 则映射后的数据 $\xi(X_{\cdot,-k})$ 中的第 i 个实例划分到治疗组; 否则划分到对照组.

4.2.4 双分类器

加权后的实例仍有可能导致部分特征和类标签之间存在虚假的相关关系, 主要的原因有两个: (1) 由于从全局角度出发学习实例权重, 可能导致有的特征变量所对应的治疗组和对照组数据分布能够很好地被平衡, 而有的特征变量所对应的治疗组和对照组数据分布不能够很好地被平衡. (2) 由于把某一特征变量作为处理变量时, 剩下的特征变量都作为混杂因子, 而实际上并不是所有变量都是混杂因子, 导致特征变量对分类的真实作用计算有偏差从而构建的预测模型仍不是很稳定.

为了进一步提高所构建模型的泛化能力, 所提算法构建了两个分类器 $f_1(\theta_1,\cdot)$ 和 $f_2(\theta_2,\cdot)$, 并赋予它们不同的参数初始化值, θ_1 和 θ_2 分别是两个分类器对应的参数. 如果两个分类器参数经过学习之后尽可能一致, 则认为所学到的模型是稳定的. 因为一个分类器对应一个分类界面, 由于两个分类器参数初始化值设置不同, 虽然最终所学的两个分类器都可以对训练数据很好的分类, 但两个分类器所学的分类界面可能不同. 如果两个分类器所学到的参数保持一致或非常接近, 则说明所学到的分类界面是相对较优的分类界面, 分类器也具有较好的泛化能力, 也就有较好的稳定性. 交叉熵损失 $\ell(\cdot)$ 被用于衡量预测标签和真实标签的误差来学习分类器. 所提算法的目标损失函数如下所示:

$$\begin{aligned}
&\min\sum_{i=1}^{n}M_i\cdot(\ell(f_1(\theta_1,\xi(x_i)),y_i)+\ell(f_2(\theta_2,\xi(x_i)),y_i),\\
&\text{s.t.}\begin{cases}
\displaystyle\sum_{k=1}^{d}\left\|\frac{\xi(X_{\cdot,-k})\cdot(M\odot X_{\cdot,k})}{M^{\mathrm{T}}\cdot X_{\cdot,k}}-\frac{\xi(X_{\cdot,-k})\cdot(M\odot(1-X_{\cdot,k}))}{M^{\mathrm{T}}\cdot(1-X_{\cdot,k})}\right\|_F^2\leqslant\lambda_1\\
\|M\odot(X-X')\|_F^2\leqslant\lambda_2, M_i\geqslant 0, \|M\|_F^2\leqslant\lambda_3, \left(\displaystyle\sum_{j=1}^{n}M_i-1\right)^2\leqslant\lambda_4, \|\theta_1\|_F^2+\|\theta_2\|_F^2\leqslant\lambda_5,\\
\displaystyle\sum_{j=1}^{l}\left(\left\|W_1^{(j)}\right\|^2+\left\|b_1^{(j)}\right\|^2+\left\|W_2^{(j)}\right\|^2+\left\|b_2^{(j)}\right\|^2\right)\leqslant\lambda_6, \|\theta_1-\theta_2\|_F^2\leqslant\lambda_7
\end{cases}
\end{aligned} \tag{15}$$

其中, y_i 为第 i 个实例的标签; $\|M\odot(X-X')\|_F^2\leqslant\lambda_2$ 用于约束数据重构损失; $\|M\|_F^2\leqslant\lambda_3$ 用于约束全局实例权重的方差; $\left(\sum_{j=1}^{n}M_i-1\right)^2\leqslant\lambda_4$ 用于防止很多实例权重被设置为 0; $\|\theta_1\|_F^2+\|\theta_2\|_F^2\leqslant\lambda_5$ 用于防止过拟合; $\sum_{j=1}^{l}(\|W_1^{(j)}\|^2+\|b_1^{(j)}\|^2+\|W_2^{(j)}\|^2+\|b_2^{(j)}\|^2)\leqslant\lambda_6$ 用于约束自动编码器的模型复杂度; $\|\theta_1-\theta_2\|_F^2\leqslant\lambda_7$ 用于约束两个分类器的参数差异.

本文所提的基于实例加权和双分类器的稳定学习算法框架如算法 1 所示. 在算法 1 中, 特征选择的百分比 τ 为 80%, 最大迭代次数 MaxIter=4000. 所提算法使用双分类器, 由于最终两个分类器的参数一致或很接近, 所提算法仅使用分类器 1 去预测测试数据.

算法 1. 基于实例加权和双分类器的稳定学习算法.

输入: 训练数据 X, 标签 y, 特征选择的百分比 τ, 最大迭代次数 MaxIter;
输出: M, θ_1, θ_2, $W_1^{(j)}$, $b_1^{(j)}$, $W_2^{(j)}$, $b_2^{(j)}$, $j=1,2,\ldots,l$.

1. 求解公式 (4) 获取权重矩阵 W;
2. $\|W_i\|_F$ $(i=1,\ldots,d)$ 进行降序排序, 然后选取特征总数的 τ% 个特征;
3. 初始化: $t=0$, θ_1, θ_2, $W_1^{(j)}$, $b_1^{(j)}$, $W_2^{(j)}$, $b_2^{(j)}$, $j=1,2,\ldots,l$;
4. 重复执行:
5. $\quad t\leftarrow t+1$;
6. $\quad$ 固定 M, 更新 $W_1^{(j)}$, $b_1^{(j)}$, $W_2^{(j)}$, $b_2^{(j)}$, θ_1, θ_2;
7. $\quad$ 固定 $W_1^{(j)}$, $b_1^{(j)}$, $W_2^{(j)}$, $b_2^{(j)}$, θ_1, θ_2, 更新 M;
8. 直至收敛或达到最大迭代次数.

5 实验结果及分析

本节通过在合成和真实数据集上进行实验来评估所提算法的有效性.

5.1 对比算法

所提算法和以下 6 个对比算法进行了对比: LR (逻辑回归算法)、DLR[12]、CRLR[11]、DGBR[12]、CVS[14]和 DWR[13]. DWR 算法主要用于回归问题, 本文中对 DWR 算法进行了改动, 采用交叉熵损失来衡量预测误差以适应于分类问题.

5.2 合成数据实验

5.2.1 数据集

本文考虑了图 2 中的 3 种情况, 即 $S\perp V$ 、 $S\rightarrow V$ 和 $V\rightarrow S$, 根据这 3 种情况生成合成数据.

(1) $S\perp V$: 在这种情况下, 稳定特征集合 S 中的特征和噪音特征集合 V 中的特征相互独立. 首先生成独立高斯分布的特征变量 $\widetilde{S}_{\cdot,1},\ldots,\widetilde{S}_{\cdot,d_s},\widetilde{V}_{\cdot,1},\ldots,\widetilde{V}_{\cdot,d_v}\sim N(0,1)$, $d_s+d_v=d$. 为了使 $X=\{S_{\cdot,1},\ldots,S_{\cdot,d_s},V_{\cdot,1},\ldots,V_{\cdot,d_v}\}$ 为二进制数据, 本文采用文献 [9] 的方法: 如果 $\widetilde{X}_{\cdot,j}\geqslant 0$, 令 $X_{\cdot,j}=1$; 否则 $X_{\cdot,j}=0$.

(2) $S\rightarrow V$: 在这种情况下, 噪音特征集合 V 中的特征根据稳定特征集合 S 中的特征生成. 首先生成带有独立高斯分布的稳定特征 $\widetilde{S}$, 如果 $\widetilde{S}_{\cdot,j}\geqslant 0$, 令 $S_{\cdot,j}=1$; 否则 $S_{\cdot,j}=0$. 然后, 基于 $\widetilde{S}$ 生成噪音特征 $\widetilde{V}=\{\widetilde{V}_{\cdot,1},\ldots,\widetilde{V}_{\cdot,d_v}\}$, $\widetilde{V}_{\cdot,j}=\widetilde{S}_{\cdot,j}+\widetilde{S}_{\cdot,j+1}+N(0,2)$. 如果 $\widetilde{V}_{\cdot,j}\geqslant 0$, 令 $V_{\cdot,j}=1$; 否则 $V_{\cdot,j}=0$.

(3) $V\rightarrow S$: 在这种情况下, 稳定特征集合 S 中的特征根据噪音特征集合 V 中的特征生成. 首先生成带有独立高斯分布的稳定特征 $\widetilde{V}$, 如果 $\widetilde{V}_{\cdot,j}\geqslant 0$, 令 $V_{\cdot,j}=1$; 否则 $V_{\cdot,j}=0$. 然后, 基于 $\widetilde{V}$ 生成噪音特征 $\widetilde{S}=\{\widetilde{S}_{\cdot,1},\ldots,\widetilde{S}_{\cdot,d_v}\}$, $\widetilde{S}_{\cdot,j}=\widetilde{V}_{\cdot,j}+\widetilde{V}_{\cdot,j+1}+N(0,2)$. 令 $S_{\cdot,j}=1$ 如果 $\widetilde{S}_{\cdot,j}\geqslant 0$; 否则 $S_{\cdot,j}=0$.

将稳定特征进一步划分为两部分: 线性部分 S_1 和非线性部分 S_2, 即 $S=\{S_1,S_2\}$. 最后, 为上述 3 种情况生成

类标签 Y. 类标签 Y 的生成方式如下所示:

$$Y = 1/\left(1+\exp\left(-\sum_{X_{\cdot,i}\in S_1}\alpha_i\cdot X_{\cdot,i}-\sum_{X_{\cdot,j}\in S_2}\beta_j\cdot X_{\cdot,j}\cdot X_{\cdot,j+1}\right)\right)+N(0,0.2)$$

其中, $\alpha_i=(-1)^i\cdot(i\%3+1)\cdot d/3$, $\beta_j=d/2$. 为了使 Y 为二进制数据, 本文采用文献 [9] 的方法: 如果 $Y\geqslant 0.5$, 令 $Y=1$; 否则 $Y=0$.

为了测试所提算法的稳定性, 需要生成一个环境集合 π, 每个环境 $e\in\pi$ 都有一个不同的数据联合分布. 在假设 1 下, 由于 $P(Y|V)$ 或 $P(V|S)$ 在不同环境中不同, 所以会出现预测的不稳定性. 因此, 在实验中通过改变 $P(Y|V)$ 和 $P(V|S)$ 来生成不同的环境. 通过样本选择偏差来改变 $P(Y|V)$, 偏差率 $r\in(0,1)$. 如果样本的噪音特征的值等于类标签的值, 即 $V=Y$ 则以概率 r 选择该样本; 否则, 选择概率为 $1-r$ 选择该样本. $r>0.5$ 表示 Y 和 V 正相关.

评估指标. 本文采用均方根误差 *RMSE* 作为评估指标. *RMSE* 的计算过程如下所示:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(f_1(\theta_1,\xi(x_i))-y(x_i))^2}{n}},$$

其中, $f_1(\theta_1,\xi(x_i))$ 为实例 x_i 的预测标签, $y(x_i)$ 为实例 x_i 的真实标签, n 为测试实例数量.

5.2.2 实验结果

在 $S\perp V$, 实例数量 $n=\{500, 750, 1000\}$, 特征数量 $p=\{20, 40, 60, 80\}$, 偏差率 $r=\{0.15, 0.25, 0.75, 0.85\}$ 的情况下生成不同的数据集, 实验结果如图 4–图 7 所示. 从实验结果可以得到以下结论.

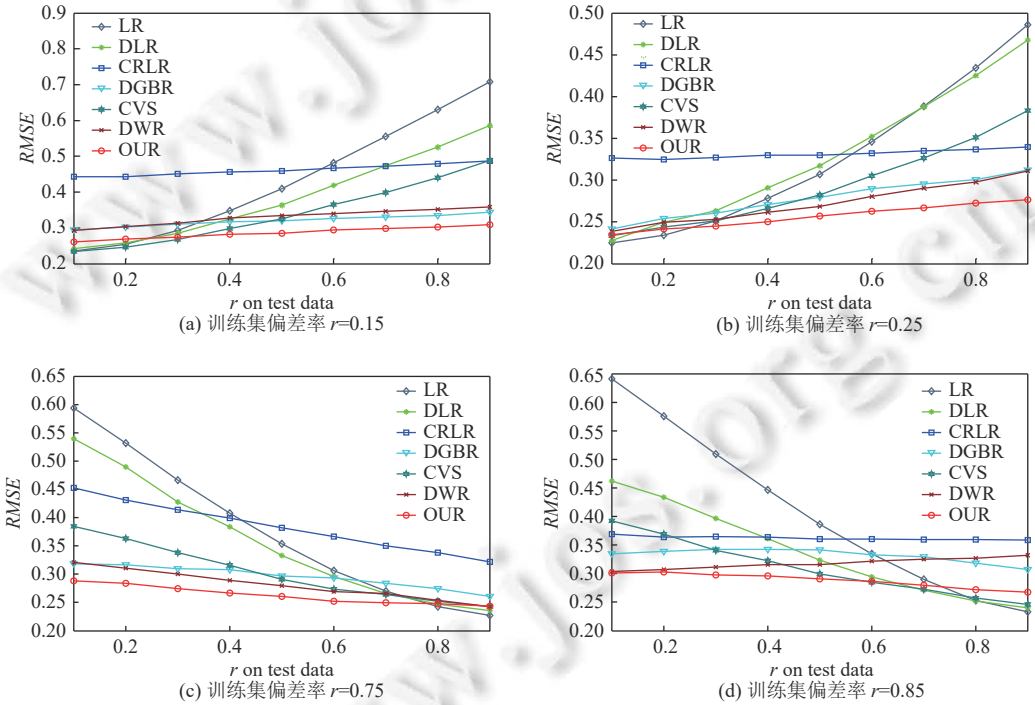
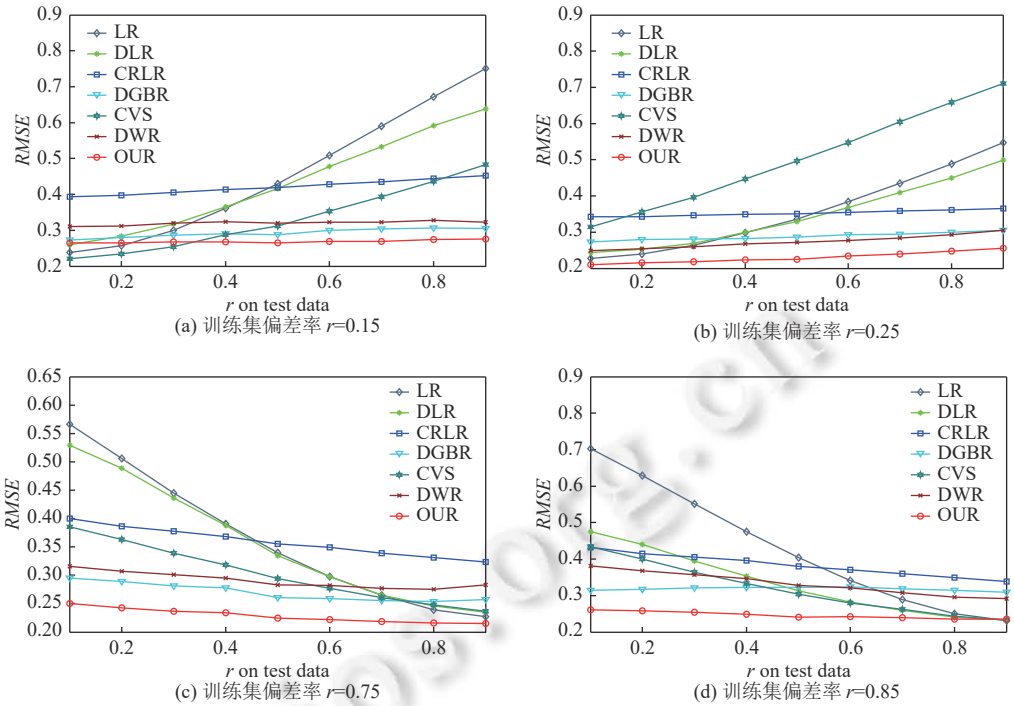


图 4 $S\perp V$, $n=500$, $p=20$ 设定: 不同偏差率 r 下各种测试数据集预测的均方根误差

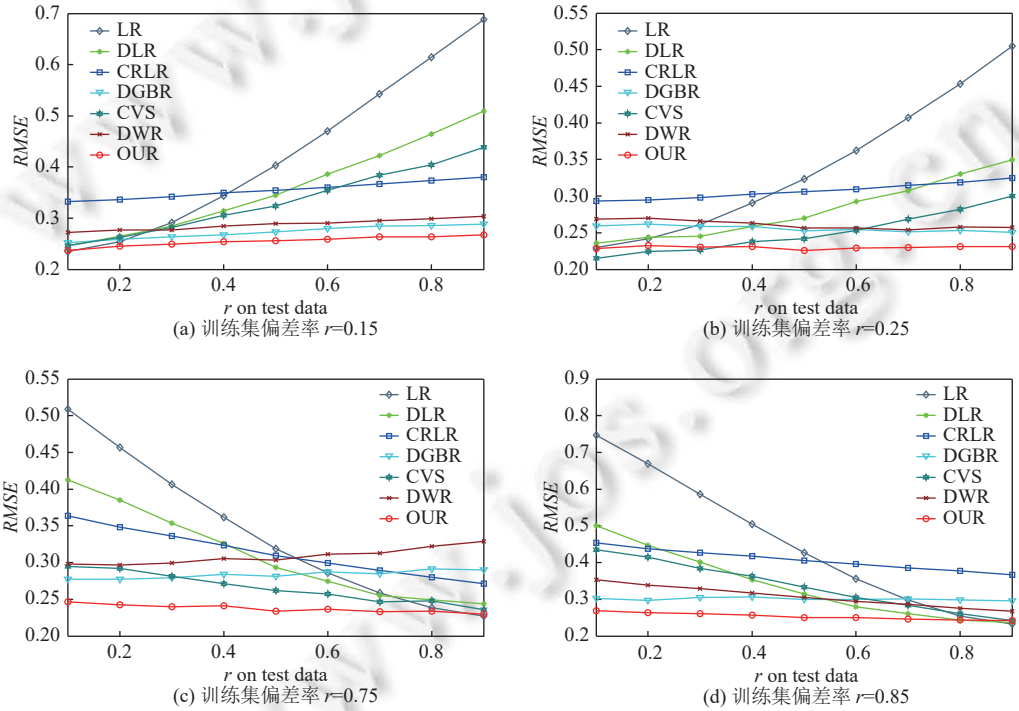
(1) 所提算法的性能优于 LR 和 DLR 算法. LR 和 DLR 算法在所有的数据集上性能都不佳, 表明了 LR 和 DLR 算法并不能解决稳定学习问题. 主要的原因如下: 由于样本选择偏差, 导致部分特征和类标签之间存在虚假的相关关系. 这些虚假的相关关系并不是稳定的, 一旦测试数据和训练数据的数据分布差异较大, 虚假的相关关系就会恶化性能. 而所提算法弱化了虚假的相关关系, 从而实现了更稳定的预测.



(a) 训练集偏差率 r=0.15　(b) 训练集偏差率 r=0.25

(c) 训练集偏差率 r=0.75　(d) 训练集偏差率 r=0.85

图 5　$S \perp V$, $n=750$, $p=20$ 设定: 不同偏差率 r 下各种测试数据集预测的均方根误差



(a) 训练集偏差率 r=0.15　(b) 训练集偏差率 r=0.25

(c) 训练集偏差率 r=0.75　(d) 训练集偏差率 r=0.85

图 6　$S \perp V$, $n=1000$, $p=20$ 设定: 不同偏差率 r 下各种测试数据集预测的均方根误差

(2) CRLR 算法也可实现稳定预测, 但其性能不如所提算法. CRLR 算法在原始特征空间平衡治疗组和对照组的数据分布. 在原始特征空间, 不相关的特征会影响治疗组和对照组的数据分布的平衡; 此外, 特征之间的非线性

关系无法被 CRLR 算法捕捉. 所提算法首先去除了部分不相关的特征, 从而弱化了不相关特征的影响; 其次使用自动编码器将治疗组和对照组的数据映射到一个低维非线性空间, 实现了治疗组和对照组数据分布更好的平衡, 由此表明了进行特征选择和采用自动编码器捕捉特征之间非线性关系的有效性.

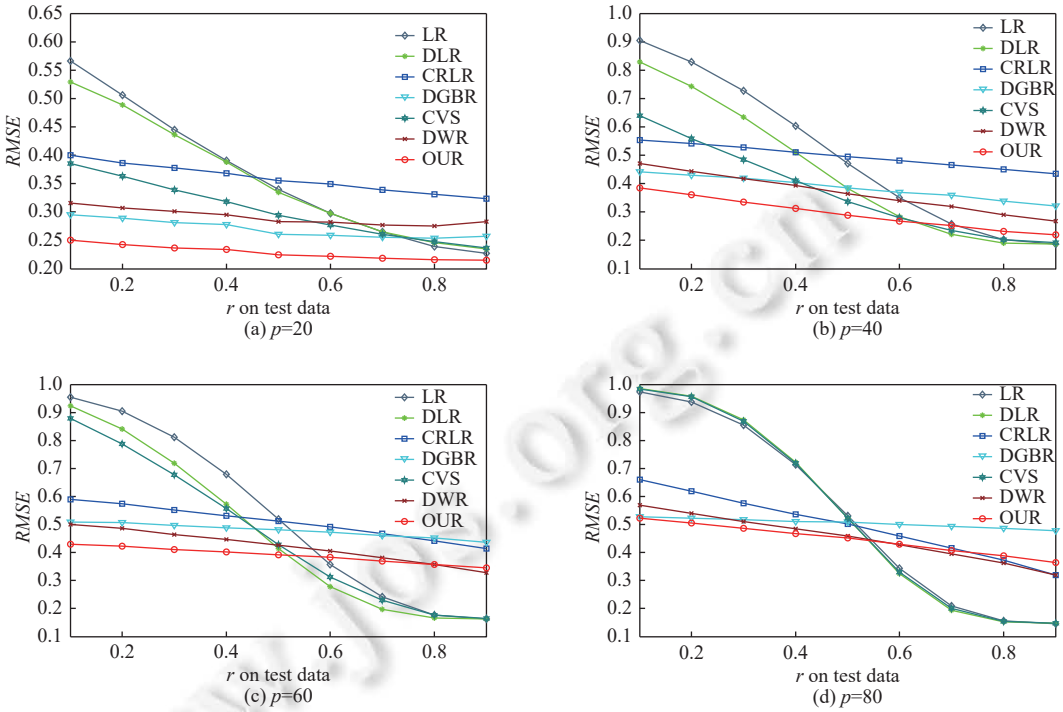


图 7 $S \perp V$, $n=750$, $r=0.75$ 设定: 不同特征数量下 p 各种测试数据集预测的均方根误差

(3) 所提算法优于 DGBR 算法. DGBR 可以实现稳定预测, 但性能略差于所提算法. 虽然 DGBR 算法首先将特征映射到低维非线性空间, 然后进行实例加权来平衡每个特征所对应的治疗组和对照组的数据分布. 但是数据中含有不相关的特征, 其会影响实例加权的过程. 此外, 构建分类器时, 分类器参数初始化的不同, 可能导致最后所学到的分类器的泛化性能不同, 而 DGBR 算法没有考虑到这一点. 而所提算法首先进行特征选择, 弱化了部分不相关的特征, 其次构建双分类器用于提高分类器的泛化能力. 由此表明特征选择和提高分类器的泛化能力有助于提高稳定学习的性能.

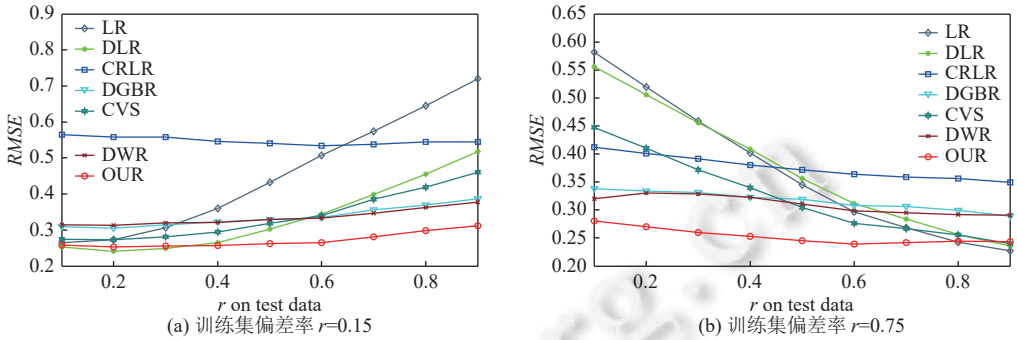
(4) 所提算法优于 CVS 算法. CVS 算法选择因果特征构建分类器. CVS 算法需要进行条件独立性测试来选择因果特征. 条件独立测试的可靠性依赖于样本量, 本文生成的数据集样本量并不是很大, 可能导致很多不可靠的条件独立性测试, 致使 CVS 的性能不佳. 虽然所提算法也进行特征选择, 但采用的特征选择方法并不依赖于条件独立性测试, 由此表明了所提算法的优势.

(5) 所提算法取得了比 DWR 算法更好的性能. DWR 算法通过实例加权来使特征之间相互独立. 为了促使特征两两独立, 很多实例的加权权重几乎为 0, 致使很多实例没有被充分利用. 而所提算法充分利用所有实例的信息.

(6) 所提算法在不同的选择偏差率 r 下都能取得较好的性能. 此外, 随着实例数量 n 的增加, 算法的性能越来越好, 原因如下: 所提算法去除了部分不相关的特征, 随着实例数量的增加, 所提算法可以更好评估每个特征对分类的作用, 从而更好地去除特征与类标签之间的虚假的相关关系. 随着特征数量的增加, 所有的算法的性能都在下降, 表明了特征维度的增加加大了学习的难度. 但所提算法在不同特征维度下, 都取得了最好的性能, 表明了所提算法的有效性.

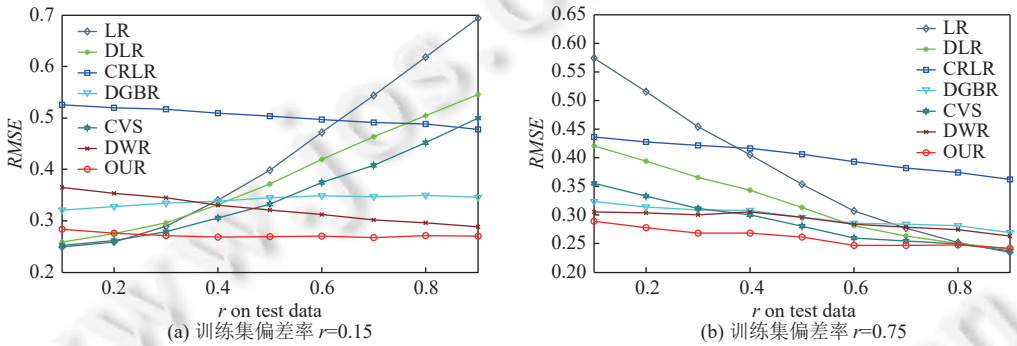
在 $S \to V$ 和 $V \to S$, 实例数量 n =750, 特征数目 p=20, 偏差率 r={0.15, 0.75}的情况下生成两个数据集, 实验结果如图 8 和图 9 所示. 从实验结果可以看出, 所提算法的性能优于 LR 和 DLR 算法. 由于样本选择偏差的原因, 导

致噪音特征集合 V 中的特征和类标签可能高度相关. 而 $P(Y|V)$ 在不同的测试环境下是不同的. LR 和 DLR 算法依赖于相关关系构建预测模型. 从图 8 和图 9 可以看出测试数据和训练数据具有相同或非常相似的数据分布时, LR 和 DLR 可以取得较好的性能, 但训练数据和测试数据分布差异变大时, 虚假的相关关系变得不可靠, 从而导致 LR 和 DLR 性能显著下降. 而所提算法可以评估每个特征对分类的真实作用, 因而构建的预测模型具有较强的鲁棒性.



(a) 训练集偏差率 r=0.15 (b) 训练集偏差率 r=0.75

图 8 $S \to V$, $n = 750, p = 20$ 设定: 不同偏差率 r 下各种测试数据集预测的均方根误差



(a) 训练集偏差率 r=0.15 (b) 训练集偏差率 r=0.75

图 9 $V \to S$, $n = 750, p = 20$ 设定: 不同偏差率 r 下各种测试数据集预测的均方根误差

所提算法的性能优于 CRLR 和 DGBR 算法. 从图 8 和图 9 可以看出, CRLR 和 DGBR 算法都可实现稳定预测. 但 CRLR 算法忽视了特征之间的非线性关系. DGBR 仅能弱化部分不相关特征与类标签之间的虚假相关关系. 而所提算法充分挖掘了特征之间的非线性关系并通过特征选择完全去除部分不相关特征和类标签之间的虚假相关; 此外, 所提算法通过构建双分类器进一步提高预测模型的泛化性能.

CVS 算法的性能略差于所提算法. CVS 算法选择因果特征构建模型, 实例数量不充分情况下可能导致 CVS 算法错选因果特征, 致使所选的部分特征和类标签之间仍存在虚假相关关系. 而所提算法即使在特征选择阶段没有去除所有无关特征, 通过实例加权也能很大程度弱化无关特征和类标签之间的虚假相关, 提高了预测性能.

从图 8 和图 9 可以看出 DWR 算法也可以实现稳定预测, 但性能略差于所提算法. DWR 算法通过实例加权迫使特征相互独立来评估每个特征对分类的作用. 通常数据集中小部分特征之间具有较强的相关性, DWR 算法通过实例加权使特征两两独立后可能会造成小部分信息丢失, 导致性能下降.

综上, 所提算法的性能优于所有对比算法, 表明了所提算法构建的预测模型具有较强的鲁棒性.

5.3 真实数据实验

5.3.1 数据集

Amazon 数据集: Amazon 数据集是跨领域情感分析的基准数据集, 其主要由 Book (B)、DVD (D)、Electronic (E) 和 Kitchen (K) 这 4 个领域的产品评论组成. 每个领域的产品评论包含 1 000 条正面评论和 1 000 条负面评论数据. 在本文中, 令一个领域产品评论数据为训练数据, 其他 3 个领域产品评论数据为 3 个测试数据. 在该数据集上

构造了 12 个任务, 如 B→D、D→E、E→K、K→B 等.

Office-Caltech10 数据集: Office-Caltech10 数据集是经典的领域适应图像数据集, 其包含从 4 个不同图像领域 Caltech256 (C)、Amazon (A)、Webcam (W) 和 DSLR (D) 收集的 2 533 个图片. 同样, 一个领域数据作为训练数据其他 3 个领域数据作为测试数据. 在该数据集上构造了 12 个任务: C→A, C→W, C→D, W→A 等.

评估指标. 本文使用的分类精度 *Accuracy* 作为评估指标. *Accuracy* 的计算过程如下所示:

$$Accuracy = \frac{\text{预测正确的实例个数}}{\text{测试实例总数}} \times 100\%.$$

5.3.2 实验结果

本节在 Amazon 和 Office-Caltech10 数据集上比较了所提算法与对比算法的分类性能, 实验结果如表 1 和表 2 所示. 在这两个数据集上, 所提算法的性能优于其他对比方法. 具体而言, 在 Amazon 数据集上, 所提算法的平均分类准确率为 75.70%. 与最优的基线算法 CRLR 相比, 分类性能提高了 0.88%. 在 Office-Caltech10 数据集上, 所提算法的平均分类准确率为 46.87%. 与最优基线算法 DGBR 相比, 分类性能提高了 1.15%, 表明了所提方法的有效性.

表 1 Amazon 数据集上的分类精度 (%)

Task	LR	DLR	CRLR	DGBR	CVS	DWR	OUR
B→D	71.09	75.64	77.94	76.31	66.83	72.64	**78.01**
B→E	70.67	73.08	74.52	73.27	66.47	71.97	**75.09**
B→K	70.38	74.73	75.08	75.79	65.18	74.09	**77.98**
D→B	70.00	66.25	**73.20**	66.60	65.80	70.05	67.87
D→E	70.62	70.94	72.17	71.02	66.61	71.77	**73.65**
D→K	72.09	74.84	76.49	74.06	65.38	74.43	**77.70**
E→B	66.35	69.57	70.55	71.17	62.60	66.70	**72.51**
E→D	65.88	69.15	73.49	72.36	65.53	68.38	**74.43**
E→K	75.33	80.02	82.54	79.97	69.43	78.10	**83.20**
K→B	65.95	70.26	69.60	68.30	67.20	68.15	**72.17**
K→D	67.53	71.23	71.19	72.46	70.87	70.54	**74.16**
K→E	72.52	79.66	81.08	80.95	70.93	78.39	**81.60**
Avg.	69.87	72.95	74.82	73.52	66.90	72.10	**75.70**

表 2 Office-Caltech10 数据集上的性能分类精度 (%)

Task	LR	DLR	CRLR	DGBR	CVS	DWR	OUR
C→A	39.25	49.66	48.75	51.30	28.39	39.98	**51.35**
C→W	26.10	38.99	36.44	38.50	31.19	28.81	**41.02**
C→D	36.31	42.23	39.81	41.50	34.39	29.94	**45.22**
A→C	36.77	**43.26**	40.25	42.09	30.63	35.08	43.00
A→W	28.14	33.53	35.93	**36.63**	26.44	32.88	35.59
A→D	**36.94**	35.63	37.58	35.19	24.84	26.11	**36.94**
W→C	26.71	33.26	30.77	36.19	19.67	26.27	**37.04**
W→A	28.28	34.34	33.87	37.66	22.80	27.04	**38.30**
W→D	66.87	80.96	74.52	**83.44**	49.68	62.74	82.16
D→C	24.39	30.39	27.28	32.60	17.72	31.30	**34.37**
D→A	27.35	31.25	27.84	32.51	16.70	31.83	**35.49**
D→W	63.72	75.59	60.29	81.00	34.24	72.88	**82.03**
Avg.	36.74	44.09	41.11	45.72	28.06	37.07	**46.87**

5.4 消融实验

1) 在本节中, 为了验证进行特征选择和使用双分类器的有效性, 本文提出了两个变体算法: 一个是不进行特征

选择也不使用双分类器, 记为 OURO; 另一个仅进行特征选择但不使用双分类器, 记为 OURS. 在合成数据集上对 3 个算法进行了比较, 实验结果如图 10 所示. 从图 10(a) 和图 10(b) 中可以看出: (1) OURS 在两个合成数据集上的性能优于 OURO 的性能, 验证了特征选择的有效性. 不相关特征会误导实例加权, 导致性能下降. 去除部分不相关特征有助于稳定学习. (2) 我们所提算法的性能优于 OURO 和 OURS, 表明进行特征选择和使用双分类器有益于稳定学习.

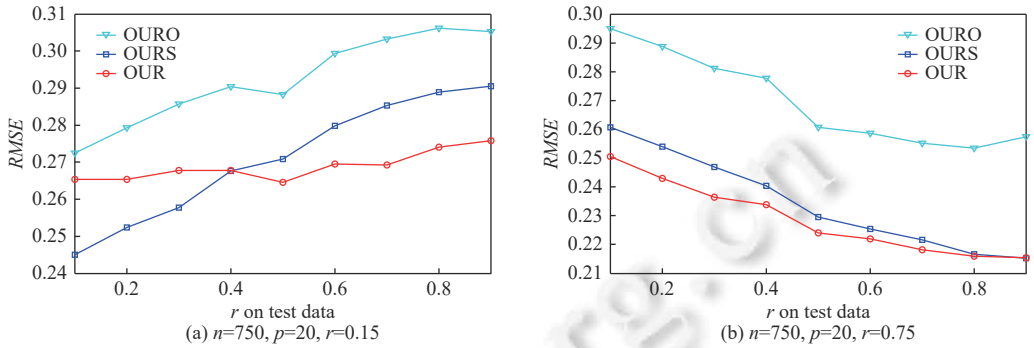


图 10 $S\perp V$ 设定: OURO、OURS 和 OUR 在两个合成数据集上的均方根误差

2) 为了进一步验证所提的特征选择策略的有效性, 本文提出 3 个变体算法, 分别使用 MRMR 算法[21]、FCBF 算法[22]、ℓ_1 - UFS 算法[23]替代所提的特征选择方法, 记为 MRMR-OUR、FCBF-OUR、UFS-OUR. 通过在合成数据集上对 4 个算法进行了比较, 实验结果如图 11 所示. 从图 11(a) 图 11(b) 中可以看出所提算法在两个合成数据集上的性能优于其他算法的性能, 验证了所提特征选择方法的有效性, 表明了特征选择时对含有较高信息量样本赋以更高的权重有利于稳定学习.

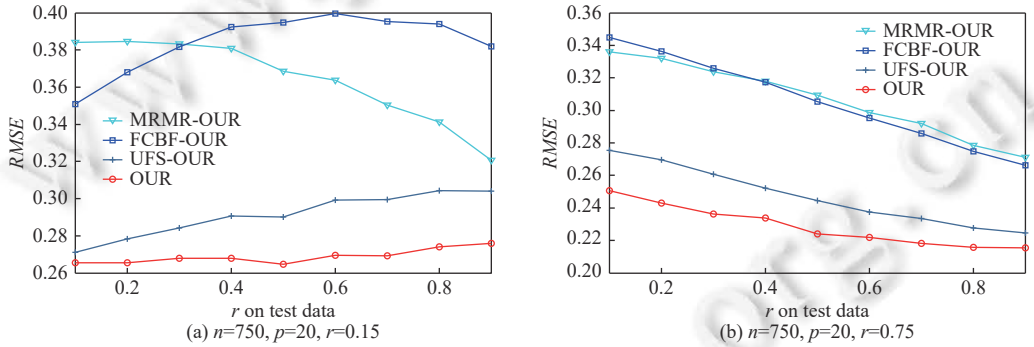
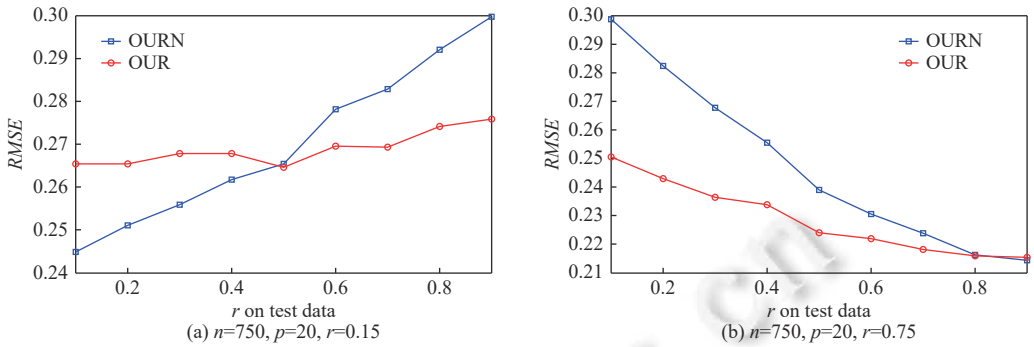


图 11 $S\perp V$ 设定: MRMR-OUR、FCBF-OUR、UFS-OUR 和 OUR 在两个合成数据集上的均方根误差

3) 在公式 (14) 中, 为了更好评估每个特征对分类的作用, 所提算法采用自动编码器将特征变量所对应的治疗组和对照组数据映射到一个低维非线性空间, 然后再平衡它们之间的数据分布. 本文提出一个变体算法, 在公式 (14) 中将特征变量 $X_{\cdot,k}$ 也进行了映射, 记为 OURN. 由于第 k 个特征变量 $X_{\cdot,k}\in\mathbb{R}^{n\times 1}$ 是个单变量, 无法直接进行低维映射, 因此将 $X_{\cdot,k}$ 的值重复 d 次构造一个新的数据矩阵 $X^*_{\cdot,k}\in\mathbb{R}^{n\times d}$, 其中 $X^*_{\cdot,k}$ 中的每一列的数据和 $X_{\cdot,k}$ 的值相同. 然后使用自动编码器将 $X^*_{\cdot,k}$ 映射到低维非线性空间获取映射后的数据 $\xi(X^*_{\cdot,k})$. 由于 $X^*_{\cdot,k}$ 中的数据每一列数据相同, 因此直接取 $\xi(X^*_{\cdot,k})$ 中的第 1 列替代公式 (14) 中的 $X_{\cdot,k}$. 在合成数据集上对所提算法和 OURN 算法进行了实验, 实验结果如图 12 所示. 从图 12(a) 图 12(b) 中可以看出, 所提算法在两个合成数据集上的性能优于 OURN, 表明将公式 (14) 中特征变量 $X_{\cdot,k}$ 进行映射后性能下降了, 其原因如下: 所提算法在评估特征变量 $X_{\cdot,k}$ 对分类的作用时, 需要将 $X_{\cdot,-k}$ 中实例划到治疗组和对照组, 如果 $X_{\cdot,k}$ 在第 i 个实例上的取值为 1, 则 $X_{\cdot,-k}$ 中的第 i 个实例划分到治疗组; 如果 $X_{\cdot,k}$ 在第 i 个实例上的取值为 0, 则 $X_{\cdot,-k}$ 中的第 i 个实例划分到对照组. 而将特征变量 $X_{\cdot,k}$ 映射后, 本文自动编码器采用 Sigmoid

非线性函数, 映射后的 $\xi(X^*_{\cdot,k}) \in (0,1)$, 不能精准的将实例划分到相应的治疗组和对照组, 从而影响实例加权的过程, 导致性能下降了.



(a) n=750, p=20, r=0.15　　(b) n=750, p=20, r=0.75

图 12　$S \perp V$ 设定: OURN 和 OUR 在两个合成数据集上的均方根误差

4) 所提算法仅使用一阶矩的差异来衡量治疗组和对照组的数据分布差异. 本文提出 2 个变体算法, 即使用二阶矩和四阶矩来度量数据分布差异, 分别记为 OUR-II 和 OUR-III. 二阶矩的计算过程如下所示:

$$\sum_{k=1}^{d}\left\|\frac{1}{M^{\mathrm{T}}\cdot X_{\cdot,k}}\sum_{i=1}^{n}\Delta_a^{\mathrm{T}}\Delta_a-\frac{1}{M^{\mathrm{T}}\cdot(1-X_{\cdot,k})}\sum_{i=1}^{n}\Delta_b^{\mathrm{T}}\Delta_b\right\|_F^2,$$

其中, Δ_a 和 Δ_b 的计算过程如下所示:

$$\Delta_a=X_{i,k}\cdot(M_i\cdot\xi(X_{i,-k})-u_a),\Delta_b=(1-X_{i,k})\cdot(M_i\cdot\xi(X_{i,-k})-u_b),\ \mu_a=\frac{\xi(X_{\cdot,-k})^{\mathrm{T}}\cdot(M\odot X_{\cdot,k})}{M^{\mathrm{T}}\cdot X_{\cdot,k}},\ \mu_b=\frac{\xi(X_{\cdot,-k})^{\mathrm{T}}\cdot(M\odot(1-X_{\cdot,k}))}{M^{\mathrm{T}}\cdot(1-X_{\cdot,k})},$$

其中, M_i 为第 i 个实例加权权重; $X_{i,-k}$ 为 $X_{\cdot,-k}$ 中第 i 个实例.

四阶矩的计算过程如下所示:

$$\sum_{k=1}^{d}\left\|\frac{1}{M^{\mathrm{T}}\cdot X_{\cdot,k}}\sum_{i=1}^{n}\Delta_a\otimes\Delta_a\otimes\Delta_a\otimes\Delta_a-\frac{1}{M^{\mathrm{T}}\cdot(1-X_{\cdot,k})}\sum_{i=1}^{n}\Delta_b\otimes\Delta_b\otimes\Delta_b\otimes\Delta_b\right\|_F^2,$$

其中, $\Delta_a\otimes\Delta_a=\Delta_a^{\mathrm{T}}\Delta_a$.

在两个合成数据集上评估 OUR-II、OUR-III 和所提算法的性能, 实验结果如图 13 所示. 从图 13(a) 和图 13(b) 中可以看出 OUR 的性能优于 OUR-II 和 OUR-III 的性能. 文献 [8] 和文献 [9] 表明数据集中特征变量为二进制变量时, 一阶矩也可以取得较好性能. 相对于一阶矩而言, 尽管理论上二阶矩和四阶矩能更好描述数据分布差异, 但二阶矩和四阶矩的计算过程较为复杂, 优化起来也更难, 导致从全局角度出发平衡每个特征变量所对应的治疗组和控制组的数据分布变得困难. 使用二阶矩和四阶矩衡量单个特征变量对应的治疗组和对照组的数据分布差异时有优势. 但所提算法需要平衡每个特征变量所对应的治疗组和对照组的数据分布. 二阶矩和四阶矩在平衡所有特征变量对应的治疗组和对照组的数据分布时, 为了保证实例加权后二阶矩值和四阶矩值较小, 可能导致实例加权权重差异不是很明显. 而一阶矩仅考虑均值差异, 能够对信息量大的实例赋予较大的权重, 使不同实例的加权权重有明显差异.

5.5　参数敏感性分析

本节研究选择特征的数量以及参数 λ_1、λ_2、λ_3、λ_4、λ_5、λ_6 和 λ_7 的取值对所提算法性能的影响. 在 $S \perp V$, $n=750$, $p=20$, $r=0.15$ 和 $r=0.75$ 两种设定情况下生成的数据集上进行实验. 当微调某个参数时, 其他参数值保持不变.

为了研究所提算法在阶段一选择的特征数量对实验结果的影响, 分别选取了总特征数量的 50% (记为 $0.5d$)、60%、70%、80%、90% 和 100% 的特征进行实验, 实验结果如图 14 所示. 从实验结果可以看出选择的特征数量占总特征数 70%–80% 时, 所提的算法性能比较好. 如果选取的特征数量太少, 一些和类标签具有相关关系的特征

可能被漏选; 如果选取的特征数量太多, 一些噪音特征可能会被选择, 噪音特征会干扰对照组和控制组的平衡过程, 从而导致实例权重评估不精准, 导致性能下降.

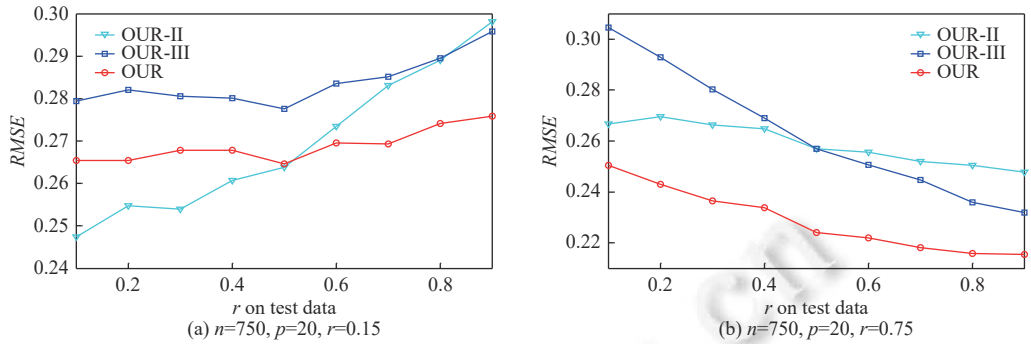


图 13 $S\perp V$ 设定: OUR-II、OUR-III 和 OUR 在两个合成数据集上的均方根误差

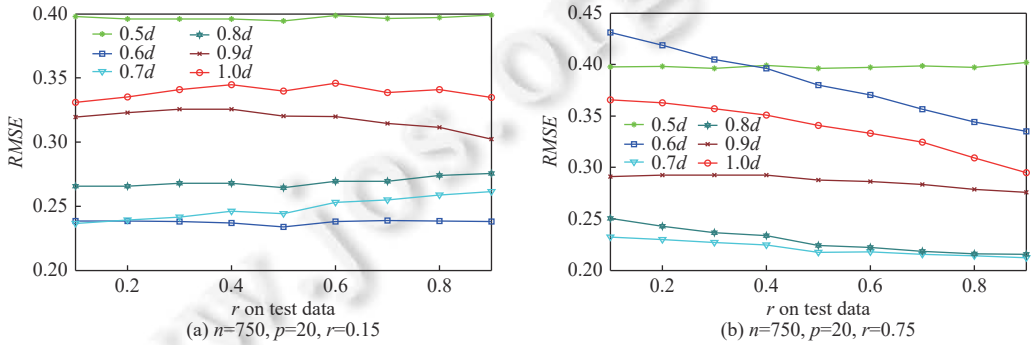


图 14 $S\perp V$ 设定: 选择的不同特征数量下各种测试数据集预测的均方根误差

参数 λ_1 主要用于约束全局平衡损失. λ_1 值的变化范围为{0.01, 0.1, 1, 5, 10, 20}. 图 15 展示了不同 λ_1 值的实验结果. 从实验结果可以看出所提算法的性能对 λ_1 的值敏感, $\lambda_1\in[0.1, 1]$ 时可以取得较好的性能. 当 λ_1 取值太小时, 对照组和控制组数据分布不能很好地被平衡, 导致实例权重计算不精准, 致使部分特征与类标签之间的虚假相关仍然存在; 当 λ_1 取值太大时, 即 $\lambda_1>1$ 时所提算法在优化参数时过多关注全局平衡损失项而较少关注预测损失等其他项, 导致性能下降.

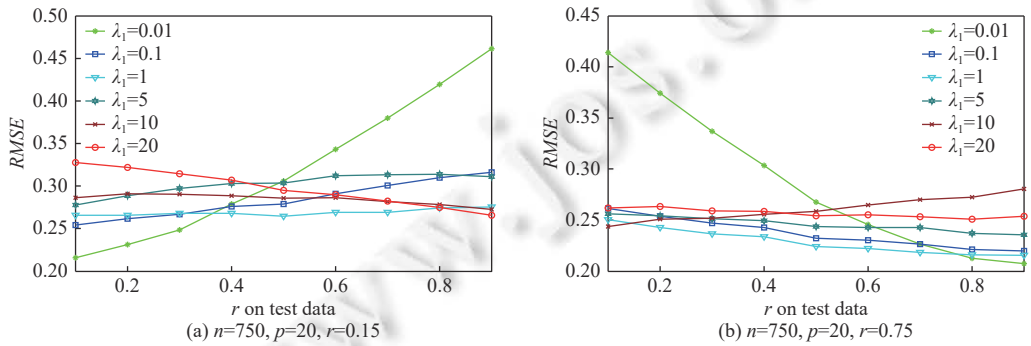
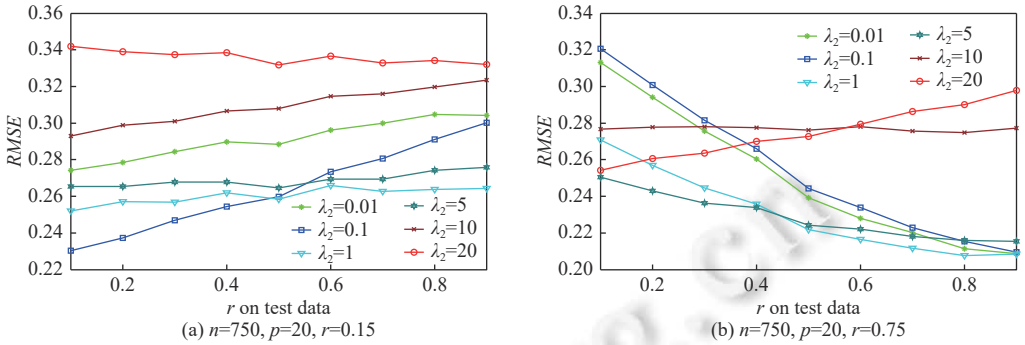


图 15 $S\perp V$ 设定: 参数 λ_1

参数 λ_2 主要用于约束自动编码器的重构损失. λ_2 值的选取范围为{0.01, 0.1, 1, 5, 10, 20}. 对不同的 λ_2 值分别进行实验, 实验结果如图 16 所示. 所提算法的性能对 λ_2 的值敏感, $\lambda_2\in[1, 5]$ 时所提算法的性能较优. 当 λ_2 取值很

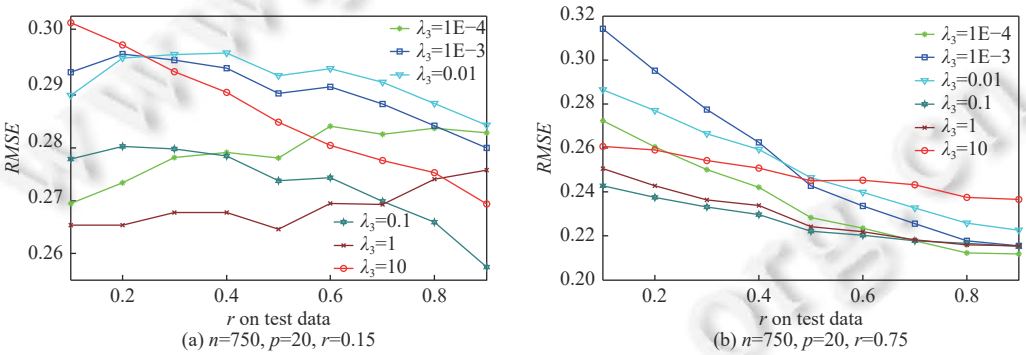
小时, 重构后的数据和原始数据可能差异较大, 自动编码器学到的隐藏层特征表示不能很好复现原始数据, 导致隐藏层特征表示损失部分原始数据信息; 当 λ_2 取值太大时使得所提算法在优化参数时过多关注重构损失项而较少关注全局平衡损失和预测损失等其他项, 导致实例权重评估不精准, 使得预测模型性能下降.



(a) n=750, p=20, r=0.15　(b) n=750, p=20, r=0.75

图 16　$S\perp V$ 设定: 参数 λ_2

参数 λ_3 主要用于约束全局实例权重的方差. λ_3 值的选取范围为{1E–4, 1E–3, 0.01, 0.1, 1, 10}. 对不同的 λ_3 值分别进行实验, 实验结果如图 17 所示. 所提算法的性能对 λ_3 的值敏感, $\lambda_3\in$[0.1, 1] 时所提算法可以取得良好性能. 当 λ_3 取值很小时, 全局实例权重方差较大, 导致部分实例权重被设置的很大而部分实例权重被设置的较小, 可能导致一小部分实例权重值很大而大部分实例权重值很小, 致使大部分权重被设置很小的实例的信息没有被充分使用; 当 λ_3 取值太大时使得所有实例的权重相差不大, 导致实例加权后不能很好平衡每个特征所对应的对照组和治疗组数据分布.

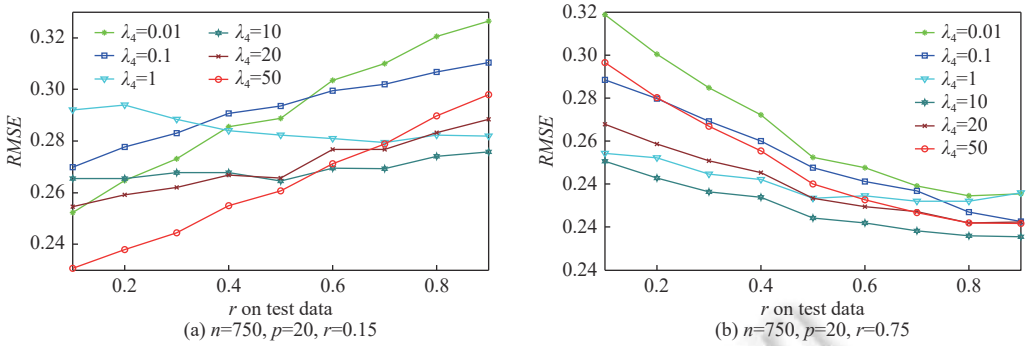


(a) n=750, p=20, r=0.15　(b) n=750, p=20, r=0.75

图 17　$S\perp V$ 设定: 参数 λ_3

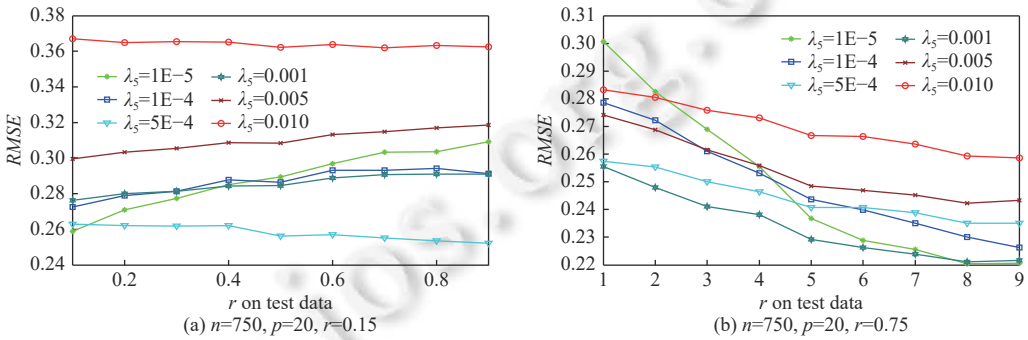
参数 λ_4 主要用于防止很多实例权重被设置为 0. λ_4 值的选取范围为{0.01, 0.1, 1, 10, 20, 50}. 对不同的 λ_4 值分别进行实验, 实验结果如图 18 所示. 所提算法的性能对 λ_4 的值敏感, $\lambda_4\in$[10, 20] 时所提算法可以取得较优性能. 当 λ_4 取值较小时, 可能导致一部分实例权重被设置为 0, 从而这部分实例信息没有被充分利用; 当 λ_4 取值太大时使得所有实例的权重都很接近, 导致信息量较高的实例作用没有被充分利用.

参数 λ_5 主要用于防止所学的预测模型过拟合. λ_5 值的选取范围为{1E–5, 1E–4, 5E–4, 0.001, 0.005, 0.01}. 对不同的 λ_5 值分别进行实验, 实验结果如图 19 所示. 所提算法的性能对 λ_5 的值敏感, $\lambda_5\in$[5E–4, 0.001] 时所提算法可以取得较优性能. 参数 λ_6 主要用于约束自动编码器的模型复杂度. λ_6 值的选取范围为{5E–5, 1E–4, 5E–4, 0.001, 0.005, 0.01}. 对不同的 λ_6 值分别进行实验, 实验结果如图 20 所示. 所提算法的性能对 λ_6 的值敏感, $\lambda_6\in$[5E–4, 0.001] 时所提算法可以取得较优性能. 参数 λ_7 主要用于约束两个分类器参数差异损失. λ_7 值的变化范围为{1E–5, 1E–4, 0.001, 0.01, 0.1, 1}, 实验结果如图 21 所示. 从实验结果可以看出 λ_7=0.001 时所提算法可以取得较佳性能.



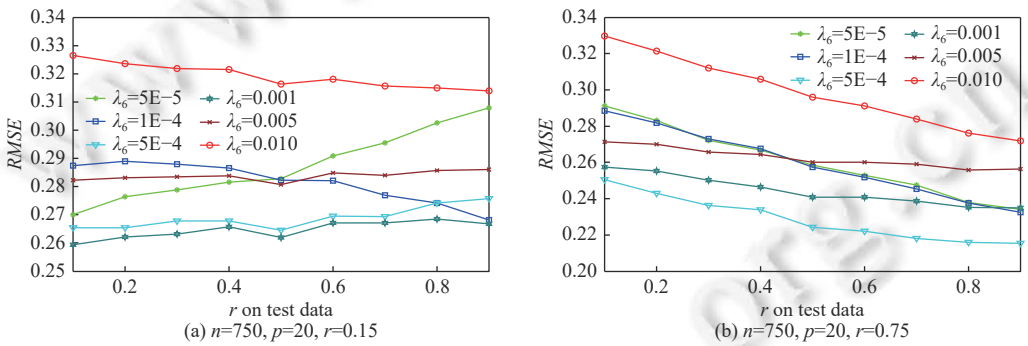
(a) n=750, p=20, r=0.15 (b) n=750, p=20, r=0.75

图 18 $S\perp V$ 设定: 参数 λ_4



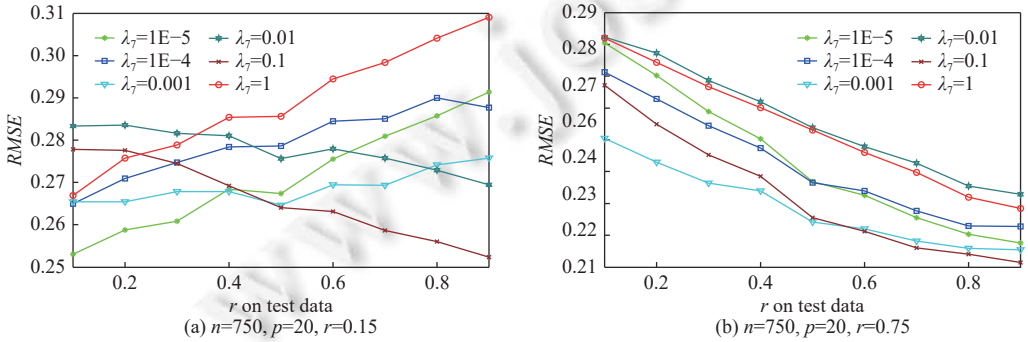
(a) n=750, p=20, r=0.15 (b) n=750, p=20, r=0.75

图 19 $S\perp V$ 设定: 参数 λ_5



(a) n=750, p=20, r=0.15 (b) n=750, p=20, r=0.75

图 20 $S\perp V$ 设定: 参数 λ_6



(a) n=750, p=20, r=0.15 (b) n=750, p=20, r=0.75

图 21 $S\perp V$ 设定: 参数 λ_7

6 结 语

本文提出了一种基于实例加权和双分类器的稳定学习算法. 该算法使用一种新的特征选择策略用于去除部分不相关的特征以完全消除部分不相关特征和类标签之间的虚假相关关系和减少不相关特征对实例加权过程的干扰, 并通过实例重新加权来评估每个特征对分类的作用从而实现稳定预测. 为了更好地进行特征选择, 所提的特征选择策略自适应对信息量丰富的实例分配更高的权重, 弱化了信息量较少的实例对特征选择的影响. 为了进一步提高预测模型的泛化能力, 该算法使用双分类器来学习一个较优的分类界面. 在合成数据和两个真实数据集上进行了实验, 实验结果验证了所提算法的有效性, 表明了去除不相关特征对实例加权过程的干扰以及使用双分类器来学习一个较优的分类界面有助于提高稳定学习的性能.

References:

[1] Fan CH, Yi JY, Tao JH, Tian ZK, Liu B, Wen ZQ. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2021, 29: 198–209. [doi: 10.1109/TASLP.2020.3039600]

[2] Kumar Y, Sahrawat D, Maheshwari S, Mahata D, Stent A, Yin YF, Shah RR, Zimmermann R. Harnessing gans for zero-shot learning of new classes in visual speech recognition. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2020. 2645–2652. [doi: 10.1609/aaai.v34i03.5649]

[3] Wang NY, Ye YX, Liu L, Feng LZ, Bao T, Peng T. Language models based on deep learning: A review. Ruan Jian Xue Bao/Journal of Software, 2021, 32(4): 1082−1115 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6169.htm [doi: 10.13328/j.cnki.jos.006169]

[4] Pei YT, Huang YP, Zou Q, Zhang XY, Wang S. Effects of image degradation and degradation removal to CNN-based image classification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2021, 43(4): 1239–1253. [doi: 10.1109/TPAMI.2019.2950923]

[5] Liu CY, Li J, He L, Plaza A, Li ST, Li B. Naive gabor networks for hyperspectral image classification. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(1): 376–390. [doi: 10.1109/TNNLS.2020.2978760]

[6] Cai Q, Pan YW, Wang Y, Liu JG, Yao T, Mei T. Learning a unified sample weighting network for object detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 14161−14170. [doi: 10.1109/CVPR42600.2020.01418]

[7] Wu Y, Chen YP, Yuan L, Liu ZC, Wang LJ, Li HZ, Fu Y. Rethinking classification and localization for object detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10183–10192. [doi: 10.1109/CVPR42600.2020.01020]

[8] Qi L, Yu PZ, Gao Y. Research on weak-supervised person re-identification. Ruan Jian Xue Bao/Journal of Software, 2020, 31(9): 2883−2902 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6083.htm [doi: 10.13328/j.cnki.jos.006083]

[9] Zhuang FZ, Qi ZY, Duan KY, Xi DB, Zhu YC, Zhu HS, Xiong H, He Q. A comprehensive survey on transfer learning. Proc. of the IEEE, 2021, 109(1): 43–76. [doi: 10.1109/JPROC.2020.3004555]

[10] Cai RC, Chen W, Zhang K, Hao ZF. A survey on non-temporal series observational data based causal discovery. Chinese Journal of Computers, 2017, 40(6): 1470–1490 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2017.01470]

[11] Shen ZY, Cui P, Kuang K, Li B, Chen PX. Causally regularized learning with agnostic data selection bias. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. ACM Press, 2018. 411–419. [doi: 10.1145/3240508.3240577]

[12] Kuang K, Cui P, Athey S, Xiong RX, Li B. Stable prediction across unknown environments. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM Press, 2018. 1617–1626. [doi: 10.1145/3219819.3220082]

[13] Kuang K, Xiong RX, Cui P, Athey S, Li B. Stable prediction with model misspecification and agnostic distribution shift. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2020. 4485–4492. [doi: 10.1609/aaai.v34i04.5876]

[14] Kuang K, Li B, Cui P, Liu Y, Tao JR, Zhuang YT, Wu F. Stable prediction via leveraging seed variable. arXiv:2006.05076, 2020.

[15] Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y. Toward causal representation learning. Proc. of the IEEE, 2021, 109(5): 612–634. [doi: 10.1109/JPROC.2021.3058954]

[16] Kuang K, Zhang HT, Wu RZ, Wu F, Zhuang YT, Zhang AJ. Balance-subsampled stable prediction across unknown test data. ACM Trans. on Knowledge Discovery from Data, 2022, 16(3): 45. [doi: 10.1145/3477052]

[17] Shen ZY, Cui P, Liu JS, Zhang T, Li B, Chen ZT. Stable learning via differentiated variable decorrelation. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. ACM Press, 2020. 2185–2193. [doi: 10.1145/3394486.3403269]

[18] Shen ZY, Cui P, Zhang T, Kuang K. Stable learning via sample reweighting. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2020. 5692–5699. [doi: 10.1609/aaai.v34i04.6024]

[19] Zhang XX, Cui P, Xu RZ, Zhou LJ, He Y, Shen ZY. Deep stable learning for out-of-distribution generalization. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5368–5378. [doi: 10.1109/CVPR46437.2021.00533]

[20] Spirtes P, Glymour C, Scheines R, Heckerman D. Causation, Prediction, and Search. 2nd ed., Cambridge: MIT Press, 2000.

[21] Peng HC, Long FH, Ding CHQ. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226–1238. [doi: 10.1109/TPAMI.2005.159]

[22] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research, 2004, 5: 1205–1224.

[23] Tang C, Zhu XZ, Chen JJ, Wang PC, Liu XW, Tian J. Robust graph regularized unsupervised feature selection. Expert Systems with Applications, 2018, 96: 64–76. [doi: 10.1016/j.eswa.2017.11.053]

附中文参考文献:

[3] 王乃钰, 叶育鑫, 刘露, 凤丽洲, 包铁, 彭涛. 基于深度学习的语言模型研究进展. 软件学报, 2021, 32(4): 1082–1115. http://www.jos.org.cn/1000-9825/6169.htm [doi: 10.13328/j.cnki.jos.006169]

[8] 祁磊, 于沛泽, 高阳. 弱监督场景下的行人重识别研究综述. 软件学报, 2020, 31(9): 2883−2902. http://www.jos.org.cn/1000-9825/6083.htm [doi: 10.13328/j.cnki.jos.006083]

[10] 蔡瑞初, 陈薇, 张坤, 郝志峰. 基于非时序观察数据的因果关系发现综述. 计算机学报, 2017, 40(6): 1470–1490. [doi: 10.11897/SP.J.1016.2017.01470]

杨帅(1995－), 男, 博士生, 主要研究领域为因果发现, 领域适应.

俞奎(1979－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 数据挖掘.

王浩(1962－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为人工智能, 数据挖掘.

曹付元(1974－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为机器学习, 数据挖掘.