基于自适应低秩表示的多任务 AUC 优化算法

孙宇辰[1),2)] 许倩倩[1)] 王子泰[3),4)] 杨智勇[2)] 黄庆明[1),2),5),6)]

[1)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

[2)](中国科学院大学计算机科学与技术学院 北京 101408)

[3)](中国科学院信息工程研究所信息安全国家重点实验室 北京 100093)

[4)](中国科学院大学网络空间安全学院 北京 100049)

[5)](中国科学院大学大数据挖掘与知识管理重点实验室 北京 101408)

[6)](鹏城实验室 广东 深圳 518055)

摘 要 多任务学习是一种基于相似任务之间的关联性进行学习迁移,使得模型在数据不足场景下仍能表现出良好泛化性能的学习方法.在该领域内,大多数现有以准确率作为基准评价标准的方法只适用于平衡分布场景.然而,诸多实际应用如疾病检测、垃圾邮件检测等,均涉及样本分布不平衡问题.针对多任务学习面向任务相关性的高要求,即当模型学习和共享不相关知识时,负迁移可能会影响模型朝着错误方向训练.因此,大多数现有方法在此类场景中无法得到有效应用.为解决该实际问题,设计一种能适用于样本不平衡场景的多任务学习算法变得尤为重要.本文提出了一种基于自适应低秩表示的多任务 AUC 优化算法,首先引入了对标签分布不敏感的 ROC 曲线下面积(AUC)作为该学习任务的评价指标,并建立了一种用于 AUC 优化的多任务学习算法,以提高模型在样本不平衡场景下的性能表现.同时,为进一步有效优化模型,本文将原始成对优化问题重构为逐样本极大极小优化问题,使得每一轮迭代复杂度由 $O(Ln_{i,+}n_{i,-})$降低至 $O(L(n_{i,+}+n_{i,-}))$.针对多任务学习中存在的负迁移现象,本文引入了一种自适应低秩正则项,以消除模型冗余信息,同时提高模型的泛化性能.最后,通过与多个对比方法在四个仿真数据集和三个真实数据集 Landmine、MHC-I 和 USPS 上的比较,所有实验结果一致证明了本文所提出算法的有效性.

关键词 多任务学习;AUC 优化;低秩表示

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2024.02678

An Adaptive Low-Rank Algorithm for Multi-Task AUC Learning

SUN Yu-Chen[1),2)] XU Qian-Qian[1)] WANG Zi-Tai[3),4)] YANG Zhi-Yong[2)] HUANG Qing-Ming[1),2),5),6)]

[1)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing* 101408)

[3)](*State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093)

[4)](*School of Cyber Security, University of Chinese Academy of Sciences, Beijing* 100049)

[5)](*Key Laboratory of Big Data Mining and Knowledge Management, University of Chinese Academy of Sciences, Beijing* 101408)

[6)](*Peng Cheng Laboratory, Shenzhen, Guangdong* 518055)

Abstract In recent years, benefiting from the excellent performance and work efficiency of deep neural networks (DNNs), machine learning technology has achieved great success in various fields, such as natural language processing, computer vision, medical named entity recognition,

收稿日期:2023-06-05;在线发布日期:2024-07-04.本课题研究由科技创新 2030-"新一代人工智能"重大项目(2018AAA0102000)、国家自然科学基金项目(62236008,U21B2038,U23B2051,61931008,62122075,61976202)、中央高校基本科研业务费专项基金、中国科学院青年促进会会员项目、中国科学院战略性先导科技专项(XDB0680000)、中国科学院计算技术研究所创新基金(E000000)资助.**孙宇辰**,硕士研究生,主要研究方向为机器学习与计算机视觉.E-mail:sunyuchen22s@ict.ac.cn.**许倩倩**,博士,研究员,中国计算机学会(CCF)高级会员,主要研究方向为统计机器学习及其在多媒体和计算机视觉领域的应用.**王子泰**,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为机器学习与数据挖掘.**杨智勇**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为机器学习及其理论.**黄庆明**,博士,讲席教授,中国计算机学会(CCF)会士,主要研究领域为多媒体计算、图像处理、计算机视觉、模式识别.

and medical image analysis. In this field, multi-task learning (MTL) is based on the correlation between similar tasks for learning transfer, enabling the model to still exhibit good generalization performance in scenarios with insufficient data. In the past decade, most existing methods are proposed for the balanced category distribution, and use accuracy-based metrics as the benchmark evaluations. However, many practical applications, such as disease detection and spam, suffer from imbalanced sample distributions, which causes the performance degradation of DNNs. Furthermore, multi-task learning has high requirements for task relevance and is apt to the negative transfer phenomenon. Specifically, when models learn to share knowledge among tasks, the irrelevant knowledge may mislead the model training in the wrong direction. This process would result in an unexpected dilemma while most existing methods cannot be effectively applied in such scenarios. Hence, to address this learning problem, designing a multi-task learning algorithm that can learn in imbalanced sample scenarios with low-correlation tasks is of paramount importance to practical applications, as well as represents a critical machine learning challenge. This paper proposes a multi-task AUC optimization method based on an adaptive low-rank Factor Nuclear Norm minus Frobenuis Norm (FNNFN) regularizer to achieve robustness on imbalanced and irrelevant data, doomed MTAUC-FNNFN. Firstly, the area under the ROC curve (AUC), which is usually adopted as the measure for imbalanced distribution, is introduced for directly reflecting the model performance among tasks. Considering the discontinuity and non-differentiability of the loss function for AUC, this work establishes a novel multi-task learning algorithm for AUC optimization, which greatly improves the AUC value in imbalanced scenarios. Meanwhile, in order to effectively optimize, this method reconstructs the original pairwise AUC formulation into an instance-wise minimax optimization problem, reducing the complexity of per-iteration from $O(Ln_{i,+}n_{i,-})$ to $O(L(n_{i,+}+n_{i,-}))$. On top of this well-formed optimization objective, the factor parameters could be easily updated with the gradient descent ascent method. For resisting the negative effect of irrelevant tasks, this paper further introduces an adaptive low-rank regularization term FNNFN to eliminate negative transfer phenomena in multi-task learning and improve the generalization performance of the model. Specifically, penalizing the small singular values empirically equates to dropping the trivial data. In this case, this low-rank structure remains the relevant information within the matrix parameters for sharing knowledge. For the purpose of achieving a comprehensive assessment, we make a comparison between the proposed method and other methods including multi-task learning methods, AUC optimization methods, and low-rank representation methods. For fairness, we uniformly apply them to multi-task datasets and estimate their performance with the AUC metric. A series of experimental results of our method on four simulated datasets and three actual datasets, Landmine, MHC-I, and USPS, consistently demonstrate the effectiveness of our proposed algorithm.

Keywords　multi-task learning; AUC optimization; low-rank representation

1　引　言

近年来，随着深度神经网络(Deep Neural Networks，DNN)的快速发展，机器学习在众多识别和分类任务上展现出的高效率和高准确率使其取得了巨大成功并受到广泛应用. 然而，众所周知的是，机器学习的成功在很大程度上依赖于大量的训练样本. 在某些难以获取样本的场景下，研究人员可能无法为模型训练收集足够多的数据，从而导致模型欠拟合. 为了解决这个问题，多任务学习(MTL)[1-2]引起了许多研究人员的关注. 具体而言，多任务学习旨在通过学习多个相似任务之间相关知识，从而综合提升模型在各个任务上的性能表现. 与传统单任务

(STL)机器学习算法相比[1],多任务学习可基于相关任务之间的内在关联性产生正向知识迁移,从而减少模型对大规模数据集的依赖,取得更加良好的泛化性能.因此,在过去的十年里,多任务学习已被应用于各种领域,如自然语言处理[3]、计算机视觉[4]、医学命名实体识别[5]、医学图像分析[6-7]等.

尽管多任务学习可以在知识共享方面表现出良好性能,但模型基于不相关任务训练时并不总尽如人意.多任务模型在共享非相关知识时往往会产生严重的性能下降.而这种臭名昭著的"负迁移"现象[8]在近些年的研究中受到了广泛关注.具体而言,当模型进行非相关知识共享时,负迁移将会误导模型沿着错误的方向训练,从而导致模型无法有效完成分类与识别任务[9].鉴于此,对多任务模型参数施以合适的正则化约束,促进任务间知识正向迁移,成为目前研究人员攻克该难题的主要方法[10-12].其中,低秩表示作为一种颇为有效的约束模型参数的方法,在近期的一些多任务学习研究中取得了初步成功[13-15].该类方法的关键假设为,矩阵参数中零奇异值的数量相当于不相关知识的数量,而非零奇异值的个数决定了参数矩阵的秩.因此,为了有效抑制负迁移现象,一些研究选择将小奇异值替换为零,并保留若干主要奇异值.已有结果表明[16-17],低秩结构能够有效促进多任务模型进行相关知识共享,并减少由非相关知识带来的负面影响.

目前,随着对基于低秩表示的多任务学习的进一步研究,其应用场景也逐渐增加.如在罕见病检测[18]、风险内容识别[19]等现实应用中,多任务模型通常只能利用不平衡数据训练.然而,现有的低秩多任务方法大多基于平衡数据的学习,难以在数据不平衡场景发挥作用.为解决此问题,一些研究人员将ROC曲线下面积(AUC)指标引入多任务学习,并在最小化AUC损失方面上取得了一系列研究进展[20-21].众所周知,AUC对标签分布的不敏感性决定了AUC指标较准确率更适用于不平衡场景[22].迄今为止,关于低秩多任务AUC学习算法的研究仍然有限,其主要原因主要为,当AUC直接被应用于一般的多任务学习算法时,深度神经模型仍然会产生负迁移现象.具体而言,优化成对AUC较为困难,其复杂度达到了$O(n^2)$.因此,优化困难直接导致多任务学习产生负迁移现象.同时,在实际优化时,考虑到巨大的计算复杂度,优化成对AUC通常需要先对数据进行采样,而这导致模型优化可利用的学习样本减少,进而降低了模型性能.因此,探究如何构建一种高效的低秩多任务AUC学习方法以满足数据不足和不平衡下的任务需求是一个仍待解决的问题.

直接使用原始AUC损失直接替换多任务学习的逐点损失可能会导致难以优化、效率低下等问题.具体而言,替换后的成对损失的每轮迭代复杂性可以达到$O(Ln_{i,+}n_{i,-})$,其中L表示任务数,$n_{i,+}$,$n_{i,-}$表示第i个任务的正负样本数量.为了解决该问题,该文章进一步引入了一种新的AUC损失函数,旨在利用逐点损失优化形式替换原始AUC的成对损失函数.该方法每轮迭代复杂度可以降低至$O(L(n_{i,+}+n_{i,-}))$,从而实现了一种基于自适应低秩表示方法的高效多任务AUC学习算法.综上所述,本文贡献主要如下:

(1)本文提出了一种高效的用于AUC优化的新型低秩多任务学习方法,该方法通过低秩化模型结构去除参数无关信息,进而提升多任务AUC优化的性能表现.与已有方法相比,该研究融合了逐点AUC优化和低秩表示方法,弥补了此前面向低秩表示的多任务AUC优化的研究空缺与不足.

(2)针对所提出的目标,本文建立了一种效率更高的学习算法,使优化迭代复杂度由原先$O(Ln_{i,+}n_{i,-})$降低至$O(L(n_{i,+}+n_{i,-}))$.

(3)该方法在三个真实数据集和四个仿真数据集上与多个对比方法进行了一系列对比实验,实验结果证明了所提出算法的有效性.

2 相关工作

本节将对现有多任务学习和AUC优化方法进行总结和介绍.

2.1 多任务学习

多任务学习是一种通过在相关任务之间共享知识从而提高单任务模型泛化性能的学习方式.近年来,关于多任务学习的研究方向主要有三种,包括基于特征的方法、基于样本的方法和基于参数的方法[23-24].此外,基于优化的多任务学习方法[25-26]正逐渐受到关注.其中,基于特征的多任务学习,如特征变换法[10,27]和特征选择法[12,28],是一种旨在从相关任务之间的特征中学习模型结构,并将其作为共享知识的方法.在基于样本的多任务学习方法中,Bickel等人[29]提出的一个代表性工作则是通过不同权重的数据估计样本与任务之间关联概率,从而使每个样本及其标签均可以属于其自己的任务.而

基于主导参数的一类方法则直接对模型参数进行操作，如参数分解[13]、低秩化[11,30-31]和稀疏化[32]等，从而约束模型沿预期的方向进行学习. 具体而言，矩阵分解和参数生成作为两种典型的基于低秩表示的多任务学习方法，被 Jeong 等人[14]的工作所采用，并取得了良好的结果. 受他们工作的启发，本文考虑将这两种方法与一种新的自适应分解正则项[33]相结合，以高效更新多任务模型分解后的两部分参数矩阵.

2.2　AUC 优化

AUC 优化是一种基于 AUC 损失更新模型参数的方法. 具体而言，原始成对 AUC 最大化方法的迭代复杂度可以达到 $O(n^2)$，这使得 AUC 优化难以得到有效发展. 最近，一系列新型 AUC 优化方法[34-38]被提出来试图解决此问题. 其中，Ying 等人[36]首次提出了一种新的 AUC 鞍点优化形式以求解模型最优参数，并将传统成对 AUC 函数转化为基于单样本的 AUC 优化形式，从而将 AUC 优化的迭代复杂度降低至 $O(n)$. 此外，该方法是一种随机在线迭代方法，与其他的一般离线算法相比，具有更高的迭代效率. 随后，受到该工作的启发，Liu 等人[38]将其引入至深度神经网络，并验证了他们所提出的两种算法的良好性能. 然而，上述所有关于 AUC 优化的研究均为单任务学习方法. 尽管 Yang 等人[20]提出了一种将 AUC 引入到多任务学习中的方法，但在多任务学习和 AUC 优化之间仍然存在一些研究缺陷. 具体而言，Yang 等人的工作使用的 AUC 损失为成对优化形式，计算复杂度较大. 鉴于此，本文旨在为多任务学习寻求一种更合适且高效的 AUC 优化算法，以建立二者之间的联系.

3　前备知识

在本节中，文章将主要介绍一般多任务学习范式和 AUC 优化方法的概念、定义及具体含义，并对各个部分的物理意义进行相应建模.

3.1　多任务学习范式

给定 L 个任务，$\mathcal{S}=\{\boldsymbol{Z}^{(1)},\cdots,\boldsymbol{Z}^{(L)}\}$ 表示一个多任务数据集，其中 $\boldsymbol{Z}^{(i)}=(\boldsymbol{X}^{(i)},\boldsymbol{Y}^{(i)})$，$i\in N_+$，$1\leq i\leq L$. 对于 $\mathcal{S}$ 中的每个任务，$\boldsymbol{X}^{(i)}\in R^{d\times n_i}$ 表示第 i 个任务的特征矩阵，其中 d 和 n_i 分别表示第 i 个任务的特征维度和样本，$\boldsymbol{X}^{(i)}$ 的每一列则对应第 i 个任务的特征向量. 相应地，$\boldsymbol{Y}^{(i)}\in R^{1\times n_i}$ 表示第 i 个任务的输出变量，$\boldsymbol{Y}^{(i)}$ 的每一元素则为 $\boldsymbol{X}^{(i)}$ 对应样本的标签. 鉴于此，特征矩阵 $\boldsymbol{W}\in R^{d\times L}$ 作为任务的联合参数矩阵，其具体形式为 $\boldsymbol{W}=[\boldsymbol{W}^{(1)},\cdots,\boldsymbol{W}^{(L)}]$. 在本文中，多任务模型被统一设置为线性模型，即面向每个任务的线性预测模型被表示为 $g^{(i)}(\boldsymbol{X})=\boldsymbol{W}^{(i)\mathrm{T}}\boldsymbol{X}$. 鉴于此，多任务学习优化问题可以表示为

$$\min_{\boldsymbol{W}}\ell(\boldsymbol{W})+\lambda\varphi(\boldsymbol{W}) \tag{1}$$

其中 $\ell(\boldsymbol{W})=\sum_i\ell_i$，$\ell_i$ 表示第 i 个任务的经验风险，$\varphi(\boldsymbol{W})$ 为正则项，λ 为超参数.

3.2　ROC 曲线下面积

ROC 曲线下面积是一种被广泛应用的经典指标. 该指标衡量了正样本排名高于负样本的概率，从而间接表现模型的性能. 这种评价方式使得其对于样本分布高度不敏感，在数据不平衡分类场景下具有良好的表现. 为了详细解释 AUC 优化，首先给出一系列符号. 令 $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$ 表示样本空间，样本 $\boldsymbol{z}=(\boldsymbol{x},y)\in\mathcal{Z}$ 和 $\boldsymbol{z}'=(\boldsymbol{x}',y')\in\mathcal{Z}$ 是一对独立且随机采样的样本，其中 $\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}$ 分别为两个样本的特征向量，$y,y'=\{+1,-1\}\in\mathcal{Y}$ 表示两个样本的标签. 根据 AUC 的定义，即衡量对正负样本进行正确排序的概率，对于任意得分函数 $f:\mathcal{X}\to R$，其被定义为

$$\mathrm{AUC}(f)=\Pr(f(\boldsymbol{x})\geq f(\boldsymbol{x}')\,|\,y=+1,y'=-1) \tag{2}$$

其中，该优化问题等价于以下表达式：

$$\min_f\mathbb{E}[\mathbb{I}_{f(\boldsymbol{x}')-f(\boldsymbol{x})<0}\,|\,y=+1,y'=-1] \tag{3}$$

其中，$\mathbb{I}_{[\cdot]}$ 为指示函数，当 $[\cdot]$ 成立时取为 1，否则为 0. 在一般情况下，很难通过这种不光滑且不可微的目标函数(1)来优化模型的参数. 为了解决该问题，一种常见的策略是利用可微替代损失替换原始问题，例如平方损失函数[39]：

$$\min_f\mathbb{E}_{z,z'}[(1-f(w;\boldsymbol{x})+f(w;\boldsymbol{x}'))^2|y=+1,y'=-1] \tag{4}$$

然而，由于分布不可知，无法直接计算其期望. 因此，通常将该目标转向优化其无偏估计，即

$$\min_f\frac{1}{n_+n_-}\sum_{i=1}^{n_+}\sum_{i=1}^{n_-}(1-f(\boldsymbol{w};\boldsymbol{x})+f(\boldsymbol{w};\boldsymbol{x}'))^2\mathbb{I}_{[y_i=+1,y_i'=-1]} \tag{5}$$

其中，n_+，n_- 分别表示正负样本的个数. 鉴于该优化目标，AUC 优化算法的优化则更加容易实现.

4　模型建立

本节将展示本文所提出的多任务 AUC 优化算法，并针对该方法的建模过程进行详细说明.

4.1　低秩表示的多任务学习

对于多任务学习而言，强关联任务所蕴含的信

息往往是相似的.如果两个任务之间存在强关联,那么训练后的多任务模型在对应参数上则更可能会拥有相似的结构.因此,矩阵参数线性相关的模型通常比一般模型具有更好的泛化性能,此类模型将更倾向于识别相关任务并进行相关知识共享.鉴于此,具有低秩结构的多任务模型往往具备抑制模型学习负迁移的能力,且可以利用相关知识提高多任务模型的性能.

对于参数矩阵 $\boldsymbol{W}$,令其秩 $rank$ 满足下列关系:$0<rank\ll\min(d,L)$,则使该矩阵变为低秩矩阵.为了获得此类结构,可以借助正则项以约束高维模型的秩.给定正则项 $\varphi(\boldsymbol{W})$,则可以构建以下参数分解形式的优化问题:

$$\min_{\boldsymbol{M},\boldsymbol{N}}\ell(\boldsymbol{M},\boldsymbol{N})+\lambda\varphi(\boldsymbol{M},\boldsymbol{N})\ \text{s.t.}\ \boldsymbol{W}=\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}}\tag{6}$$

其中,$\boldsymbol{M}\in R^{d\times r}$为"特征"潜在矩阵,$\boldsymbol{N}\in R^{L\times r}$是"任务"潜在矩阵,$r\in\mathbb{N}_+$为潜在变量.$\boldsymbol{M}$ 和 $\boldsymbol{N}$ 分别是从包含"特征"维度和"任务"维度分解后的两个独立部分.基于上述形式,适用于多任务学习的线性模型则可以表示为 $g^{(i)}(\boldsymbol{X})=(\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}})^{(i)\mathrm{T}}\boldsymbol{X}$,且潜在变量 r 需满足条件 $0<rank\leq r\ll\min(d,L)$.由于分解后的两个部分参数的秩均远小于 $\min(d,L)$,这两部分参数在进行优化时总会保证低秩结构.

因此,得益于该低秩结构,式(3)的 AUC 优化可以被转化为优化分解参数的 AUC 优化方法,从而保证模型不受负迁移的影响.通过形式化 $\ell(\boldsymbol{M},\boldsymbol{N})$,可以得到以下基于低秩表示的多任务 AUC 优化算法:

$$l(\boldsymbol{M},\boldsymbol{N})=\frac{1}{L}\sum_{i=1}^{L}\frac{1}{n_{i,+}n_{i,-}}\sum_{\boldsymbol{x}_j\in\mathcal{S}_+^{(i)}}\sum_{\boldsymbol{x}_k\in\mathcal{S}_-^{(i)}}F(\boldsymbol{M},\boldsymbol{N},\boldsymbol{x}_j,\boldsymbol{x}_k)\tag{7}$$

$F(\boldsymbol{M},\boldsymbol{N},\boldsymbol{x}_j,\boldsymbol{x}_k)$被表示为

$$F(\boldsymbol{M},\boldsymbol{N},\boldsymbol{x}_j,\boldsymbol{x}_k)=(1-f((\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}})^{(i)};\boldsymbol{x}_j)+f((\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}})^{(i)};\boldsymbol{x}_k))^2\tag{8}$$

其中,$f((\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}})^{(i)};\boldsymbol{x})=(\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}})^{(i)\mathrm{T}}\boldsymbol{x}$;$\mathcal{S}_+^{(i)}$ 和 $\mathcal{S}_-^{(i)}$ 分别表示正负样本数据集;$n_{i,+}$,$n_{i,-}$ 分别表示第 i 个任务中的正负样本个数.显而易见,这种成对 AUC 形式仍会产生巨大的迭代复杂度.因此,下一小节将集中处理此问题,并对 AUC 优化函数进行重构.

4.2 AUC 优化重构

从成对 AUC 损失函数(5)可以得知,AUC 优化过程中每次的迭代复杂度为 $O(n_+n_-)$,该问题为 AUC 优化带来了一定挑战.为获得一种更加简单的目标函数,现有一些研究探索了关于重构 AUC 最大化函数的可行方法.其中,一种被称为随机在线 AUC 最大化的方法[36,40]在更新模型方面颇为高效,而且其优化目标被证明等价于原始优化目标(3).鉴于此,本文引入这种等价优化形式以更好地拟合多任务模型.具体而言,这个等价问题被表示为以下极大极小化问题:

$$\min_{\boldsymbol{w},a,b}\max_{\alpha\in R}h(\boldsymbol{w},a,b,\alpha)=\mathbb{E}_{\boldsymbol{z}}[F(\boldsymbol{w},a,b,\alpha;\boldsymbol{z})]\tag{9}$$

其中,$\boldsymbol{w}\in\mathbb{R}^d$是线性模型 $f(\boldsymbol{w};\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ 的权重向量;$h(\boldsymbol{w},a,b,\alpha)$是关于 $\boldsymbol{w},a,b$ 的凸函数和关于 α 的凹函数;a,b,α 是三个辅助优化变量.令 $p=\Pr(y=1)$表示正样本的比例,则有

$$\begin{aligned}F(\boldsymbol{w},a,b,\alpha;\boldsymbol{z})=&(1-p)(f(\boldsymbol{w};\boldsymbol{x})-a)^2\mathbb{I}_{[y=1]}+\\&p(f(\boldsymbol{w};\boldsymbol{x})-b)^2\mathbb{I}_{[y=-1]}+\\&2(1+\alpha)(pf(\boldsymbol{w};\boldsymbol{x})\mathbb{I}_{[y=-1]}-\\&(1-p)f(\boldsymbol{w};\boldsymbol{x})\mathbb{I}_{[y=1]})-p(1-p)\alpha^2\end{aligned}\tag{10}$$

该优化问题是一个有低迭代复杂度的鞍点问题,其通过极大极小优化得到最优模型 $f(\boldsymbol{w}^*;\boldsymbol{x})$.与式(4)相比,该目标利用单样本优化形式取代了原始成对优化形式,极大地降低了 AUC 优化难度.

4.3 基于分解 NNFN 正则的多任务 AUC 优化

在上述 AUC 最大最小化方法的基础上,这种基于低秩表示的多任务 AUC 经验风险最小化问题(ERM)可被建立为

$$\begin{aligned}&\min_{a,b,\boldsymbol{M},\boldsymbol{N}}\max_{\boldsymbol{\alpha}}F(\boldsymbol{M},\boldsymbol{N},\boldsymbol{a},\boldsymbol{b},\boldsymbol{\alpha};\mathcal{S})=\\&\min_{a,b,\boldsymbol{M},\boldsymbol{N}}\max_{\boldsymbol{\alpha}}\ell(\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}},\boldsymbol{a},\boldsymbol{b},\boldsymbol{\alpha};\mathcal{S})+\varphi(\boldsymbol{M},\boldsymbol{N})\end{aligned}\tag{11}$$

$$\begin{aligned}&\ell(\boldsymbol{W},\boldsymbol{a},\boldsymbol{b},\boldsymbol{\alpha};\mathcal{S})=\\&\frac{1}{L}\sum_{i=1}^{L}\left(\frac{1-p}{n_{i,+}}\sum_{\boldsymbol{x}_j\in\mathcal{S}_+^{(i)}}(f(\boldsymbol{W}^{(i)};\boldsymbol{x}_j)-\boldsymbol{a}_i)^2\right)+\\&\frac{1}{L}\sum_{i=1}^{L}\left(\frac{p}{n_{i,-}}\sum_{\boldsymbol{x}_k\in\mathcal{S}_-^{(i)}}(f(\boldsymbol{W}^{(i)};\boldsymbol{x}_k)-\boldsymbol{b}_i)^2\right)+\\&\frac{1}{L}\sum_{i=1}^{L}\left(\frac{2(1+\boldsymbol{\alpha}_i)p}{n_{i,-}}\sum_{\boldsymbol{x}_k\in\mathcal{S}_-^{(i)}}f(\boldsymbol{W}^{(i)};\boldsymbol{x}_k)\right)-\\&\frac{1}{L}\sum_{i=1}^{L}\left(\frac{2(1+\boldsymbol{\alpha}_i)(1-p)}{n_{i,+}}\sum_{\boldsymbol{x}_j\in\mathcal{S}_+^{(i)}}f(\boldsymbol{W}^{(i)};\boldsymbol{x}_j)\right)-\\&p(1-p)\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha}\end{aligned}\tag{12}$$

其中,$\boldsymbol{a},\boldsymbol{b},\boldsymbol{\alpha}\in\mathbb{R}^{L\times1}$是三个辅助优化向量,$\boldsymbol{a}_i,\boldsymbol{b}_i,\boldsymbol{\alpha}_i\in\mathbb{R}$ 分别是 $\boldsymbol{a},\boldsymbol{b},\boldsymbol{\alpha}$ 的第 i 个分量.设 $\boldsymbol{W}=\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}}$,其中 $\boldsymbol{W}$ 表示多任务模型的参数矩阵,$\boldsymbol{W}$ 被分解为 $\boldsymbol{M}\in\mathbb{R}^{d\times r}$和 $\boldsymbol{N}\in\mathbb{R}^{L\times r}$.受 Wang 等人[33]提出的低秩矩阵学习的启发,本文将该工作所提出的一种可扩展的、自适应的非凸正则项引入至式(7),从而辅助约束参数优化,并通过一般的梯度下降方法来解决.具体来说,正则项 $\varphi(\boldsymbol{M},\boldsymbol{N})$被表示为

$$\varphi(\boldsymbol{M},\boldsymbol{N})=\frac{\lambda}{2}(\|\boldsymbol{M}\|_F^2+\|\boldsymbol{N}\|_F^2)-\lambda\|\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}}\|_F \quad (13)$$

其中，λ 为超参数，$\|a\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}^2}$ 表示 Frobenius 范数. 因此，上述凸优化问题可以借助梯度下降-上升方法得到有效解决.

多任务 AUC 方法在 Yang 等人[20]工作中被证明在不平衡场景下有效. 本文基于该结论，引入逐点优化的 AUC 方法，提升优化效率. 此外，由于本文采用线性模型，所提方法满足 Liu 等人[38]工作中定理 1 假设条件，即优化 $F(\boldsymbol{M},\boldsymbol{N},\boldsymbol{a},\boldsymbol{b},\boldsymbol{\alpha};\mathcal{S})$ 时，参数能够以 $O(\epsilon^{-6})$ 的迭代复杂度收敛到 ϵ^2 以内. 同时，关于非对称低秩分解问题工作[41]证明，尽管非对称低秩矩阵分解问题的低秩优化是非凸且非光滑的，但随机初始化梯度下降能够以多项式收敛速度收敛到全局极小值. 因此，本文利用上述优势，使用随机初始化变量和梯度更新优化分解参数，在多任务 AUC 方法上进一步提升优化速度和性能.

4.4　优化框架

受 Chen 等人[42]工作的启发，问题(11)可以使用近端梯度方法进行优化. 具体而言，采用近端梯度下降-上升方法作为优化器来更新参数，同时利用梯度下降法来更新 $\boldsymbol{M}$ 和 $\boldsymbol{N}$. 具体算法如算法 1 所示.

算法 1.　MTAUC-FNNFN.

输入：多任务数据集 $\mathcal{S}$

参数初始化：潜在变量 r，迭代次数 K，更新步长 $\eta_{\boldsymbol{M}}$，$\eta_{\boldsymbol{N}}$，$\eta_{\boldsymbol{v}}$，$\eta_{\boldsymbol{\alpha}}$，动量 μ；随机初始化 $\boldsymbol{M}^1\in R^{d\times r}$，$\boldsymbol{N}^1\in R^{L\times r}$，$\boldsymbol{v}^1=(\boldsymbol{a}^1,\boldsymbol{b}^1)=0$，$\boldsymbol{\alpha}^1=0$

输出：$\boldsymbol{M}^K\boldsymbol{N}^{K\mathrm{T}}$

1. 令 $k=1$
2. WHILE $k\leq K$ DO
3. $\quad\widehat{g}_{\boldsymbol{M}}=\nabla_{\boldsymbol{M}}F(\boldsymbol{M}^k,\boldsymbol{N}^k,\boldsymbol{v}^k,\boldsymbol{\alpha}^k;\mathcal{S})$
4. $\quad\widehat{g}_{\boldsymbol{N}}=\nabla_{\boldsymbol{N}}F(\boldsymbol{M}^k,\boldsymbol{N}^k,\boldsymbol{v}^k,\boldsymbol{\alpha}^k;\mathcal{S})$
5. $\quad u_{\boldsymbol{M}}^{k+1}=\mu u_{\boldsymbol{M}}^k-\eta_{\boldsymbol{M}}\widehat{g}_{\boldsymbol{M}}$，$u_{\boldsymbol{N}}^{k+1}=\mu u_{\boldsymbol{N}}^k-\eta_{\boldsymbol{N}}\widehat{g}_{\boldsymbol{N}}$
6. $\quad\boldsymbol{M}^{k+1}=\boldsymbol{M}^k+u_{\boldsymbol{M}}^{k+1}$，$\boldsymbol{N}^{k+1}=\boldsymbol{N}^k+u_{\boldsymbol{N}}^{k+1}$
7. $\quad\widehat{g}_{\boldsymbol{v}}=\nabla_{\boldsymbol{v}}F(\boldsymbol{M}^k,\boldsymbol{N}^k,\boldsymbol{v}^k,\boldsymbol{\alpha}^k;\mathcal{S})$
8. $\quad\widehat{g}_{\boldsymbol{\alpha}}=\nabla_{\boldsymbol{\alpha}}F(\boldsymbol{M}^k,\boldsymbol{N}^k,\boldsymbol{v}^k,\boldsymbol{\alpha}^k;\mathcal{S})$
9. $\quad\boldsymbol{v}^{k+1}=prox_{\eta_{\boldsymbol{v}}}(\boldsymbol{v}^k-\eta_{\boldsymbol{v}}u_{\boldsymbol{v}}^{k+1})$
10. $\quad\boldsymbol{\alpha}^{k+1}=prox_{\eta_{\boldsymbol{\alpha}}}(\boldsymbol{\alpha}^k+\eta_{\boldsymbol{\alpha}}u_{\boldsymbol{\alpha}}^{k+1})$
11. $\quad k=k+1$
12. END WHILE
13. RETURN $\boldsymbol{M}^K,\boldsymbol{N}^K,\boldsymbol{v}^K,\boldsymbol{\alpha}^K$

(1) 首先，算法输入多任务训练数据，并随机初始化模型参数；

(2) 在步 3～6 中，算法计算 $\boldsymbol{M}$ 和 $\boldsymbol{N}$ 的梯度来更新参数；

(3) 在步 7～10 中，通过近端梯度下降-上升法来更新变量 $\boldsymbol{v}$ 和 $\boldsymbol{\alpha}$，其中 $\boldsymbol{v}=(\boldsymbol{a},\boldsymbol{b})$；

(4) 最后，经过 K 轮迭代，算法输出最优参数 $\boldsymbol{M}^K$ 和 $\boldsymbol{N}^K$ 作为多任务模型的最终参数.

5　实　验

本节将对上一节提出的算法进行一系列的实验验证. 具体而言，本节首先给出了实验的基本设置，并介绍了一些典型对比方法与本文方法进行性能比较. 为了充分验证多任务模型在不平衡数据集上的有效性，本文方法在几个典型的真实数据集和四个大型仿真数据集上进行了实验验证，并针对实验结果进行了系统分析.

5.1　数据集介绍

5.1.1　真实数据集

首先介绍三个真实多任务数据集 Landmine、MHC-I 和 USPS，具体信息如表 1 和下文所示：

表 1　三个真实数据集基本信息(其中 L、f、P、N、T、r 分别为任务数、特征维度、正样本数、负样本数、总样本数和不平衡率(其中，不平衡率 $r=\frac{N}{P}$))

数据集	L	f	P	N	T	r
Landmine	29	9	904	13916	14820	15.4
MHC-I	8	180	10189	8390	18579	1.2
USPS	5	256	4876	4422	9298	1.1

(1) Landmine 数据集[43]. Landmine 数据集由若干 9 维数据组成. 其中，这些数据是从 29 个地雷探测任务的雷达图像中提取的特征，这些任务的目标是将所有 14820 幅图像分为两类：地雷图像或非地雷图像(干扰图像).

(2) MHC-I 数据集[44]. MHC-I 数据集是生物信息学领域的二元多任务数据集，包含不同肽与不同 MHC-I 分子的结合亲和力数据，即短氨基酸序列. 每项任务的目标是预测肽是否通过氨基酸序列与 MHC-I 分子结合. 该数据集包含 8 种不同分子的 18579 个数据. 本文只选择了序列长度为 9 的数据组成数据集. 其中，序列中的每个字符均被视为长度为 20 的特征(序列中总共只有 20 种字符，字符在向量中的对应位置表示为 1，其他位置标记为 0)，因此每个氨基酸都被表示为一个 180 维数据.

(3) USPS 数据集[45]. USPS 数据集是一个手写的数字图像数据集，其中包含从 0 到 9 的数字的图像. 每个图像的大小是 16×16，处理后的数据被表

示为 256 维数据. 本文将所有数据划分为五个部分，每个部分用于一项任务，即数字 0 和数字 1、数字 2 和数字 3、数字 4 和数字 5、数字 6 和数字 7 以及数字 8 和数字 9 的分类. 同时，将每个任务中的偶数数据设置为正例，将所有奇数数据设置为负例. 因此，该多任务数据集包含 5 个二分类相关任务，共有 9298 个具有 256 个特征的数据.

5.1.2 仿真数据集

除了真实数据集外，本文还生成了四个仿真数据集对模型进行实验验证. 具体而言，本文设计的每个仿真数据集共有 100 个任务，其中每个任务 $\boldsymbol{X}^{(i)} \in \mathbb{R}^{1000\times 80}$. 每个样本均提取自分布 $\mathcal{N}(0, \boldsymbol{I}_{80})$，其中 $\boldsymbol{I}_{80}$ 表示维度为 80 的 1 向量. 除上述生成数据，本文设置线性模型 $s^{(i)} = \boldsymbol{W}^{(i)\mathrm{T}}\boldsymbol{X}^{(i)} + \epsilon^{(i)}$ 来分割仿真数据，其中 $\boldsymbol{W} \sim \mathcal{U}(0,1)$ 且其秩不大于 5，$\epsilon^{(i)} \in \mathbb{R}^{1000\times 1}$ 且 $\epsilon^{(i)} \sim \mathcal{N}(0, 0.01^2 I_{1000})$. 为了获得具有不同不平衡率的四个仿真数据集，每个任务中得分前 80 个、50 个、40 个和 10 个样本被标记为正样本，其余所有样本均被标记为负样本. 仿真数据集的具体信息如表 2 所示. 为了评估真实参数 $\boldsymbol{W}^*$ 和训练后模型参数 $\boldsymbol{W}$ 之间的距离，本文引入了一个评价标准 E.W.，其具体表达形式如下式所示：

$$\text{E.W.} = \frac{\|\boldsymbol{W}^* - \boldsymbol{W}\|_F^2}{L}.$$

表 2 四个仿真数据集的基本信息

数据集	L	f	P	N	T	r
Simu1	100	80	8 k	92 k	100 k	11.5
Simu2	100	80	5 k	95 k	100 k	19.0
Simu3	100	80	4 k	96 k	100 k	24.0
Simu4	100	80	1 k	99 k	100 k	99.0

5.2 实验设置

对于每个数据集，其中 80% 的数据用于模型训练，剩余的 20% 用于模型测试. 为了寻找最合适的模型参数设定，本文在不同的参数设置下对模型采用了 5 折交叉验证，根据其平均结果选择有助于模型在验证集上获得最佳结果的参数进行测试. 在测试阶段，本文在每个方法下均完成了 10 次随机重复实验，并计算其平均 AUC 结果和标准差呈现在表 2 中. 为了获取尽可能好的结果，所有实验均在集合 $\{1, 1\mathrm{e}{-1}, 1\mathrm{e}{-2}\}$ 上进行网格搜索以确定超参数 λ 的值，并将动量 μ 统一设置为 0.9，并令潜在变量 r 满足条件 $r \ll \min(d, T)$ 使模型满足低秩结构. 最后，所有的实验结果均为 AUC 指标下的测量结果.

5.3 对比方法

在本节中，本文设计并选择了多任务学习领域和 AUC 优化领域内的一系列对比方法，对本文提出方法进行实验比较. 除此之外，本文还根据均方误差(MSE)损失和 AUC 损失两种常见损失函数提出了一系列对比方法，并设置 ℓ_2 范数和 FNNFN 范数作为正则项. 首先，本文所提出三个基于 MSE 的对比方法如下所示：

单任务 ℓ_2 均方差(STL-MSE-L2)：

$$\arg\min_{\boldsymbol{w}_i} \sum_{j=1}^{n_i} \ell^{(i,j)}(\boldsymbol{w}_i) + r_{\ell_2\text{-norm}}(\boldsymbol{w}_i).$$

多任务 ℓ_2 均方差(MTL-MSE-L2)：

$$\arg\min_{\boldsymbol{W}} \frac{1}{Ln_i} \sum_{i=1}^{L}\sum_{j=1}^{n_i} \ell^{(i,j)}(\boldsymbol{W}) + r_{\ell_2\text{-norm}}(\boldsymbol{W}).$$

多任务低秩均方差(MTL-MSE-LowRank)：

$$\arg\min_{\boldsymbol{M},\boldsymbol{N}} \frac{1}{Ln_i} \sum_{i=1}^{L}\sum_{j=1}^{n_i} \ell^{(i,j)}(\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}}) + r_{\mathrm{FNNFN}}(\boldsymbol{M},\boldsymbol{N}).$$

对于上述基于最小均方差的对比方法，定义单任务最小均方差损失函数为 $\ell^{(i,j)}(\boldsymbol{w}_i) = (\boldsymbol{Y}^{(i,j)} - \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{X}^{(i,j)})^2$，多任务最小均方差损失函数为 $\ell^{(i,j)}(\boldsymbol{w}_i) = (\boldsymbol{Y}^{(i,j)} - \boldsymbol{W}^{(i)\mathrm{T}}\boldsymbol{X}^{(i,j)})^2$. 其中，单任务模型 $\boldsymbol{w}_i \in \mathbb{R}^d$，多任务模型 $\boldsymbol{W} \in \mathbb{R}^{d\times L}$；$\boldsymbol{X}^{(i,j)}$ 是第 i 个任务的第 j 个样本，$\boldsymbol{Y}^{(i,j)}$ 表示相应样本的标签；$r_{\ell_2\text{-norm}}$ 表示 ℓ_2 正则项，r_{FNNFN} 表示分解 NNFN 惩罚项. 三个基于 AUC 损失优化的对比方法如下所示：

单任务 ℓ_2 AUC 损失(STL-AUC-L2)：

$$\arg\min_{\boldsymbol{w}_i} \frac{1}{n_{i,+}n_{i,-}} \sum_{\boldsymbol{x}_j\in\mathcal{S}_+^{(i)}} \sum_{\boldsymbol{x}_k\in\mathcal{S}_-^{(i)}} (1 - f(\boldsymbol{w}_i;\boldsymbol{x}_j) + f(\boldsymbol{w}_i;\boldsymbol{x}_k))^2 + r_{\ell_2\text{-norm}}(\boldsymbol{w}_i).$$

多任务 ℓ_2 AUC 损失(MTL-AUC-L2)：

$$\arg\min_{\boldsymbol{W}} \frac{1}{Ln_{i,+}n_{i,-}} \sum_{i=1}^{L} \sum_{\boldsymbol{x}_j\in\mathcal{S}_+^{(i)}} \sum_{\boldsymbol{x}_k\in\mathcal{S}_-^{(i)}} (1 - f(\boldsymbol{W}^{(i)};\boldsymbol{x}_j) + f(\boldsymbol{W}^{(i)};\boldsymbol{x}_k))^2 + r_{\ell_2\text{-norm}}(\boldsymbol{W}^{(i)}).$$

多任务低秩 AUC 损失(MTL-AUC-LowRank)：

$$\arg\min_{\boldsymbol{W}} r_{\mathrm{FNNFN}}(\boldsymbol{M},\boldsymbol{N}) + \frac{1}{Ln_{i,+}n_{i,-}} \sum_{i=1}^{L} \sum_{\boldsymbol{x}_j\in\mathcal{S}_+^{(i)}} \sum_{\boldsymbol{x}_k\in\mathcal{S}_-^{(i)}} (1 - f((\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}})^{(i)};\boldsymbol{x}_j) + f((\boldsymbol{M}\boldsymbol{N}^{\mathrm{T}})^{(i)};\boldsymbol{x}_k))^2$$

其中，$f(\boldsymbol{w}_i;\boldsymbol{x}) = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}$，$f(\boldsymbol{W}^{(i)};\boldsymbol{x}) = \boldsymbol{W}^{(i)\mathrm{T}}\boldsymbol{x}$.

对于其他对比方法，本文将在以下内容中进行简要介绍.

(1) RMTL[13] 在学习相关任务的同时，还旨在通过两个分离的结构来判别不相关的异常任务，这

两个结构分别被命名为低秩模块、具有迹范数和 ℓ_2 范数的组稀疏模块；

(2) VSTG-MTL[14] 利用两个稀疏矩阵和 k 范数来同时训练参数和筛选相关任务组；

(3) NC-CMTL[46] 是校准多任务学习的扩展，它可以训练非凸低秩模型来直接解决多任务回归问题；

(4) LPAP-MTAUC[20] 是一种用于个性化属性多级学习的方法，其将参数分解为三部分，并利用 AUC 损失优化实现精确分类；

(5) KMSV、KMSV-new[15] 是两种基于重加权方法的相似算法，用于解决使用紧松弛方法时而产生的 NP 难秩最小化问题.

5.4　实验结果

(1) 性能比较. 表 3 列出了本文方法和对比方法在不同数据集上的 10 次平均结果. 实验结果表明通过本文算法训练得到的线性多任务模型在大多数数据集中都优于其他算法. 这说明本文提出的方法能够生成具有更加优异泛化性能的模型. 在训练阶段，本文算法使得模型在不平衡数据集上的训练更加稳定. 具体而言，Landmine 数据集的不平衡率高达 15.4，其远高于 MHC-I(1.2)和 USPS(1.1)的不平衡率. 而如表 3 的左半部分显示，本文算法在 Landmine 数据集上的表现要优于其他两个数据集上的表现. 表 3 的右半部分表明，随着四个仿真数据集的不平衡率变高，本文算法优化得到模型取得的 AUC 值也逐渐升高，而通过其他方法得到模型的 AUC 值却逐渐降低. 因此通过上述观察可知，本文方法在采用适当的 AUC 优化方法后显著提高了模型在不平衡数据的分类效果，并始终优于大多数其他对比方法. 综上，本文提出方法在不平衡多任务场景下拥有更加优异的性能表现.

表 3　不同数据集上 MTAUC-FNNFN 与其他对比方法的 AUC 性能比较

算法	AUC						
	真实数据集			仿真数据集			
	Landmine	MHC-I	USPS	Simu1	Simu2	Simu3	Simu4
STL-MSE-L2	68.2±0.00	77.4±0.00	97.9±0.00	93.4±0.00	93.7±0.00	90.3±0.00	88.1±0.01
MTL-MSE-L2	65.4±2.09	86.2±0.59	94.0±1.00	95.8±0.16	95.5±0.16	94.7±0.12	87.9±0.23
MTL-MSE-LowRank	67.8±0.91	91.8±0.14	96.3±0.71	99.1±0.00	99.1±0.01	99.0±0.05	91.3±0.17
STL-AUC-L2	70.3±0.51	76.7±0.17	84.2±4.32	97.9±0.10	98.2±0.06	98.6±0.04	98.1±0.25
MTL-AUC-L2	69.3±2.68	85.7±2.36	66.8±4.13	95.7±0.01	96.1±0.03	95.4±0.09	89.3±0.10
MTL-AUC-LowRank	70.7±4.50	87.7±2.36	81.2±6.44	99.3±0.02	99.4±0.01	99.5±0.01	99.5±0.04
RMTL	68.6±0.27	88.7±0.00	77.8±0.46	94.9±0.00	94.1±0.00	94.2±0.00	87.5±0.02
VSTG-MTL($k=5$)	70.9±1.67	91.3±0.87	99.8±0.09	99.1±0.06	99.2±0.06	99.2±0.06	97.4±0.38
NC-CMTL	72.3±1.93	**93.2±0.30**	99.7±0.06	95.6±0.13	94.9±0.21	94.6±0.29	88.9±0.86
KMSV	71.5±1.54	90.1±0.42	99.8±0.10	96.0±0.12	95.2±0.20	95.0±0.30	89.6±1.05
KMSV-new	70.9±2.31	87.3±0.68	99.3±0.26	94.8±0.13	93.1±0.24	92.3±0.45	79.3±1.74
LPAP-MTAUC	69.3±1.47	82.3±0.53	97.3±0.29	97.5±0.12	97.1±0.15	96.7±0.18	91.3±0.63
MTAUC-FNNFN(Ours)	**73.0±0.61**	90.4±0.27	**99.9±0.10**	**99.3±0.04**	**99.5±0.02**	**99.6±0.04**	**99.7±0.03**

(2) 效率对比. 除上述 AUC 性能比较外，表 4 还展示了使用 AUC 优化的每个算法每一轮迭代花费的平均运行时间. 具体而言，与其他三种原始成对 AUC 优化算法相比，本文算法在迭代更新上所消耗的时间更少. 可以观察到，本文算法的迭代速率甚至优于单任务 AUC 优化算法. 对于 Ori1 的多任务学习版本 Ori2 和 Ori3 而言，多任务学习使得模型的学习成本骤然增加. 而本文算法基于重构 AUC 和近端梯度下降-上升方法有效地避免了此问题. 综上所述，表 4 中的结果有力地证明了本文方法在多任务学习中的高效率.

表 4　各 AUC 优化算法的运行时间对比

算法	单次平均迭代时间/s			
	Landmine	MHC-I	USPS	Simu
Ori1	2.2969	0.6406	0.4160	7.8594
Ori2	3.7507	1.7677	1.2679	15.1294
Ori3	4.7215	2.1543	1.3438	19.9120
Ours	**0.2423**	**0.1557**	**0.0875**	**1.9657**

注：Ori1、Ori2 和 Ori3 分别表示本文提出的三种基于原始 AUC 优化的算法，即 STL-AUC-L2、MTL-AUC-L2 和 MTL-AUC-LowRank. Simu 表示任意不平衡率下的仿真数据集(不平衡率改变不影响算法复杂度)

(3) 模型参数可视化. 除分析模型性能和效率之外，本文还研究了提出算法训练得到的模型结构. 由图 1 展示的算法迭代过程可以得知，本文的算法在各数据集上进行一系列更新迭代后均表现出了较好的收敛性. 不同优化算法还原模型参数矩阵的情况如图 2 所示. 当 E.W. 值越小时，说明还原矩阵与原始矩阵之间的距离越小，表明还原效果更佳. 通过 E.W. 指标可知，随着不平衡率的增加，本文方法恢

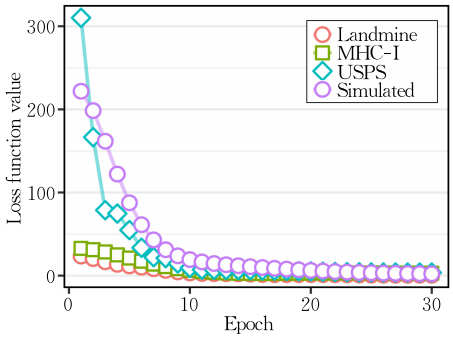

图 1 本文算法在真实数据集和仿真数据集上的收敛曲线

复出的矩阵参数在大多数情况下要优于其他算法，尤其当不平衡率增加到 99 时，本文算法的性能依旧优于所有其他算法. 虽然 VSTG-MTL 方法在低不平衡率下还原效果更优，但是其对于不平衡数据仍较为敏感.

为了更清晰地显示模型参数的恢复，本文选择了两个对比模型与本文模型进行比较，并将对比结果呈现在图 3 和图 4 中. 由于本文方法引入了低秩表示方法，本文模型参数矩阵的对角线上只存在几个主要的奇异值，图 3 清晰地展示了本文模型参数具有明显的低秩结构. 同时，图 4 表明训练得到的模型的对角矩阵中只有不超过 5 个大于 0.01 的奇异值，这种低秩结构使得模型训练得到的参数与生成仿真数据的模型参数更加相近. 这意味着本文所提出的算法在低秩表示方面具有更大的优势，以更好地实施模型在数据不平衡场景下的训练与学习.

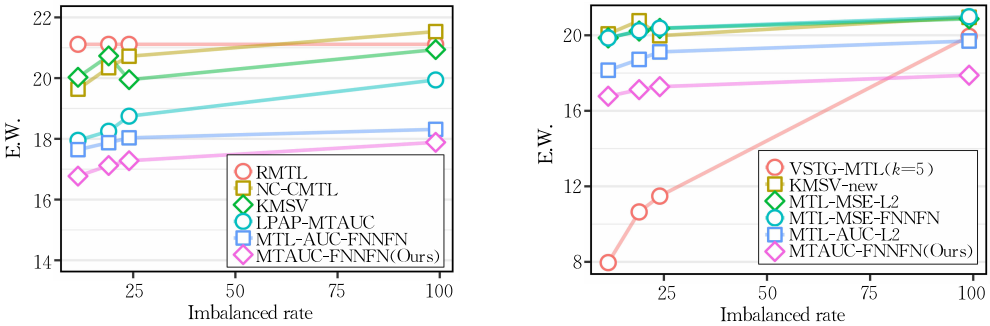

图 2 所有算法训练得到模型的 E.W. 值对比

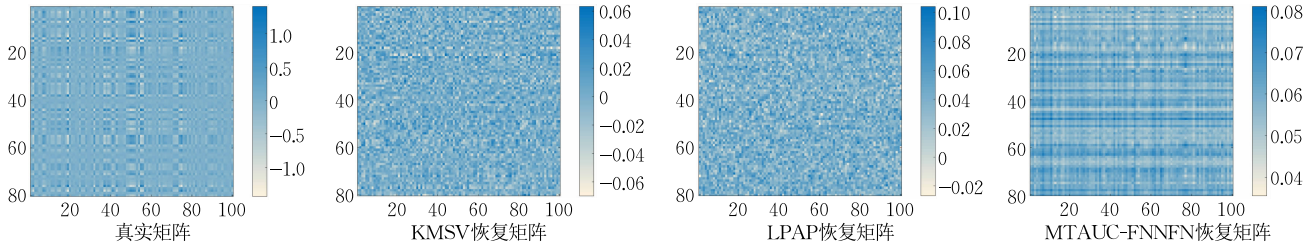

图 3 在 Simu1 数据集上通过 3 种代表算法训练得到的模型的矩阵恢复情况

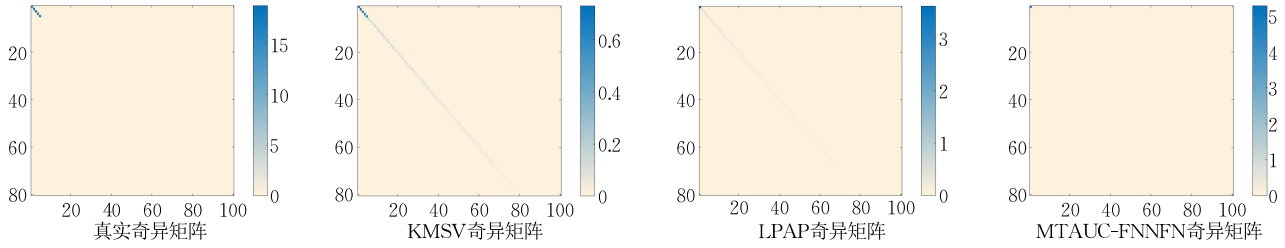

图 4 在 Simu1 数据集上通过 3 种代表算法训练得到的模型参数矩阵的奇异值矩阵

6 结 论

本文提出了一种新的基于低秩表示的多任务 AUC 优化算法. 具体而言，该算法采用了一种自适应非凸正则项用于低秩多任务学习. 为降低算法迭代复杂度，该算法采用了一个重构单样本 AUC 损失函数，以提高模型在不平衡数据集上表现的同时，保证了模型参数更新的可行性. 最后，该方法在多个多任务二分类数据集上进行了一系列性能表现和算法效率的实验验证. 实验结果一致证明了本文所提出算法的性能优越性.

参 考 文 献

[1] Caruana R. Multitask learning. Machine Learning, 1997, 28(1): 41-75

[2] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning//Proceedings of the International Conference on Machine Learning. Helsinki, Finland, 2008: 160-167

[3] Lim K T, Lee J Y, Carbonell J, et al. Semi-supervised learning on meta structure: Multi-task tagging and parsing in low-resource scenarios//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 8344-8351

[4] Liu S, Johns E, Davison A J. End-to-end multi-task learning with attention//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1871-1880

[5] Zhou B, Cai X, Zhang Y, et al. MTAAL: Multi-task adversarial active learning for medical named entity recognition and normalization//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 14586-14593

[6] Niu J, Yang Y, Zhang S, et al. Multi-task character-level attentional networks for medical concept normalization. Neural Processing Letters, 2019, 49(3): 1239-1256

[7] Moeskops P, Wolterink J M, van der Velden B H M, et al. Deep learning for multi-task medical image segmentation in multiple modalities//Proceedings of the Medical Image Computing and Computer-Assisted Intervention. Athens, Greece, 2016: 478-486

[8] Liu S, Liang Y, Gitter A. Loss-balanced task weighting to reduce negative transfer in multi-task learning//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 9977-9978

[9] Kollias D. ABAW: Learning from synthetic data & multi-task learning challenges//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 157-172

[10] Maurer A, Pontil M, Romera-Paredes B. Sparse coding for multitask and transfer learning//Proceedings of the International Conference on Machine Learning. Atlanta, USA, 2013: 343-351

[11] Barzilai A, Crammer K. Convex multi-task learning by clustering//Proceedings of the Artificial Intelligence and Statistics. San Diego, USA, 2015: 65-73

[12] Wang J, Ye J. Safe screening for multi-task feature learning with multiple data matrices//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 1747-1756

[13] Chen J, Zhou J, Ye J. Integrating low-rank and group-sparse structures for robust multi-task learning//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 42-50

[14] Jeong J Y, Jun C H. Variable selection and task grouping for multi-task learning//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 1589-1598

[15] Chang W, Nie F, Wang R, et al. New tight relaxations of rank minimization for multi-task learning//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. Queensland, Australia, 2021: 2910-2914

[16] Jain P, Netrapalli P, Sanghavi S. Low-rank matrix completion using alternating minimization//Proceedings of the Symposium on Theory of Computing Conference. Palo Alto, USA, 2013: 665-674

[17] Collins L, Hassani H, Mokhtari A, et al. Exploiting shared representations for personalized federated learning//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 2089-2099

[18] Hao H, Fu H, Xu Y, et al. Open-narrow-synechiae anterior chamber angle classification in AS-OCT sequences. arXiv preprint arXiv:2006.05367, 2020

[19] Wu H, Hu Z, Jia J, et al. Mining unfollow behavior in large-scale online social networks via spatial-temporal interaction//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 254-261

[20] Yang Z, Xu Q, Cao X, et al. Learning personalized attribute preference via multi-task AUC optimization//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 5660-5667

[21] Hu Q, Zhong Y, Yang T. Multi-block min-max bilevel optimization with applications in multi-task deep AUC maximization//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022

[22] Cortes C, Mohri M. AUC optimization vs. error rate minimization//Proceedings of the Advances in Neural Information Processing Systems. Vancouver and Whistler, Canada, 2003: 313-320

[23] Zhang Y, Yang Q. A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering, 2021, 34(12): 5586-5609

[24] Vandenhende S, Georgoulis S, Van Gansbeke W, et al. Multi-task learning for dense prediction tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(7): 3614-3633

[25] Yang E, Pan J, Wang X, et al. AdaTask: A task-aware adaptive learning rate approach to multi-task learning//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023: 10745-10753

[26] Navon A, Shamsian A, Achituve I, et al. Multi-task learning as a bargaining game//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 16428-16446

[27] Xu L, Huang A, Chen J, et al. Exploiting task-feature co-clusters in multi-task learning//Proceedings of the AAAI Conference on Artificial Intelligence. Austin, USA, 2015:

1931-1937

[28] Han Y, Zhang J, Xu Z, et al. Discriminative multi-task feature selection//Proceedings of the Late-Breaking Developments in the Field of Artificial Intelligence. Bellevue, USA, 2013: 41-43

[29] Bickel S, Bogojeska J, Lengauer T, et al. Multi-task learning for HIV therapy screening//Proceedings of the International Conference on Machine Learning. Helsinki, Finland, 2008: 56-63

[30] Han L, Zhang Y. Multi-stage multi-task learning with reduced rank//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, Arizona, 2016: 1638-1644

[31] McDonald A M, Pontil M, Stamos D. Spectral k-support norm regularization//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 3644-3652

[32] Zhang Y, Yang Q. Learning sparse task relations in multi-task learning//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 2914-2920

[33] Wang Y, Yao Q, Kwok J. A scalable, adaptive and sound nonconvex regularizer for low-rank matrix learning//Proceedings of the Web Conference. Ljubljana, Slovenia, 2021: 1798-1808

[34] Gao W, Jin R, Zhu S, et al. One-pass AUC optimization//Proceedings of the International Conference on Machine Learning, Atlanta, USA, 2013: 906-914

[35] Ding Y, Zhao P, Hoi S, et al. An adaptive gradient method for online AUC maximization//Proceedings of the AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2568-2574

[36] Ying Y, Wen L, Lyu S. Stochastic online AUC maximization //Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 451-459

[37] Huo J, Gao Y, Shi Y, et al. Cross-modal metric learning for AUC optimization. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(10): 4844-4856

[38] Liu M, Yuan Z, Ying Y, et al. Stochastic AUC maximization with deep neural networks//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2019

[39] Gao W, Zhou Z H. On the consistency of AUC pairwise optimization//Proceedings of the International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 939-945

[40] Liu M, Zhang X, Chen Z, et al. Fast stochastic AUC maximization with $O(1/n)$-convergence rate//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 3195-3203

[41] Ye T, Du S S. Global convergence of gradient descent for asymmetric low-rank matrix factorization//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021: 1429-1439

[42] Chen Z, Zhou Y, Xu T, et al. Proximal gradient descent-ascent: Variable convergence under KL geometry//Proceedings of the International Conference on Learning Representations. Virtual Event, Austria, 2021

[43] Jawanpuria P K, Lapin M, Hein M, et al. Efficient output kernel learning for multiple tasks//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 1189-1197

[44] Peters B, Bui H H, Frankild S, et al. A community resource benchmarking predictions of peptide binding to MHC-I molecules. PLoS Computational Biology, 2006, 2(6): e65

[45] Hull J J. A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(5): 550-554

[46] Nie F, Hu Z, Li X. Calibrated multi-task learning//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 2012-2021

附录 1. 其他实验结果，如图 5～图 10 所示.

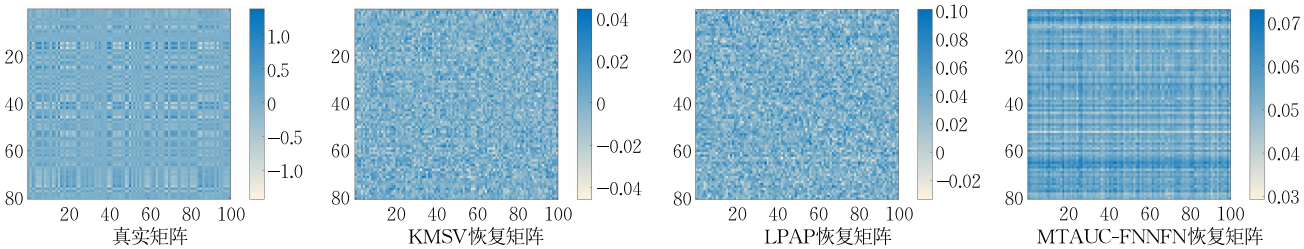

图 5 在 Simu2 数据集上通过 3 种代表算法训练得到的模型的矩阵恢复情况

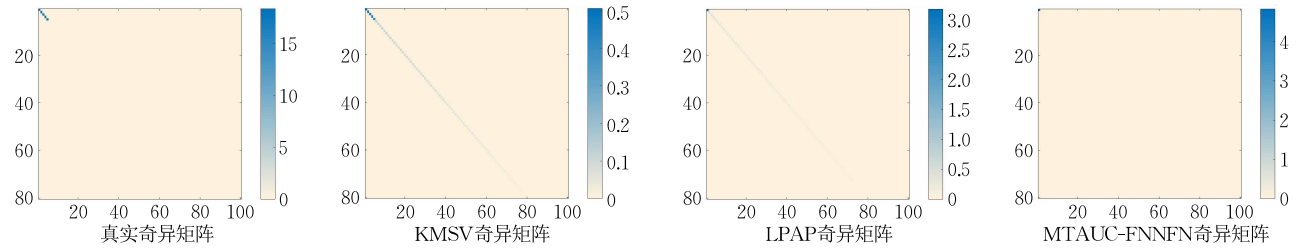

图 6 在 Simu2 数据集上通过 3 种代表算法训练得到的模型参数矩阵的奇异值矩阵

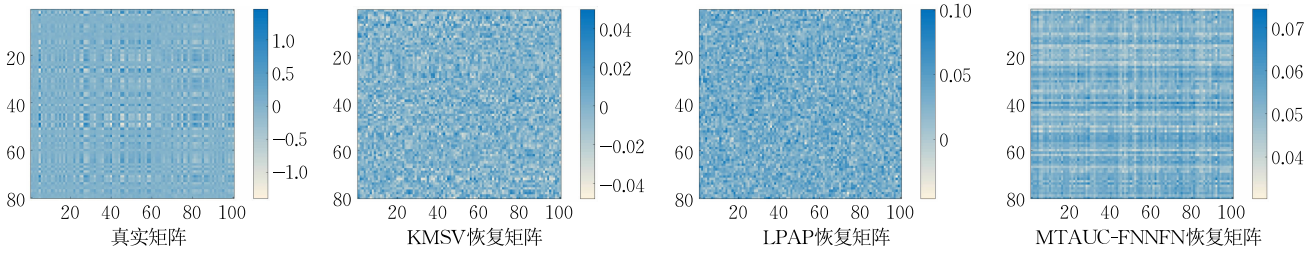

图 7　在 Simu3 数据集上通过 3 种代表算法训练得到的模型的矩阵恢复情况

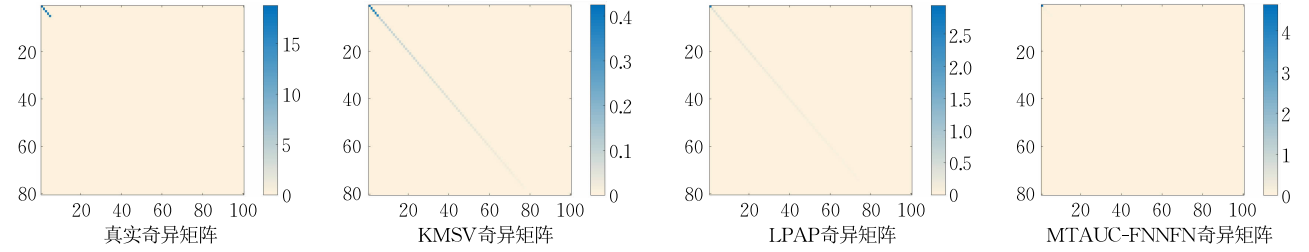

图 8　在 Simu3 数据集上通过 3 种代表算法训练得到的模型参数矩阵的奇异值矩阵

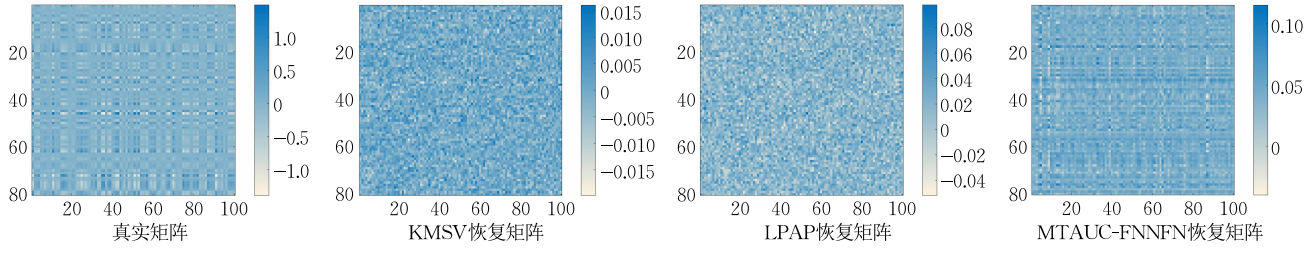

图 9　在 Simu4 数据集上通过 3 种代表算法训练得到的模型的矩阵恢复情况

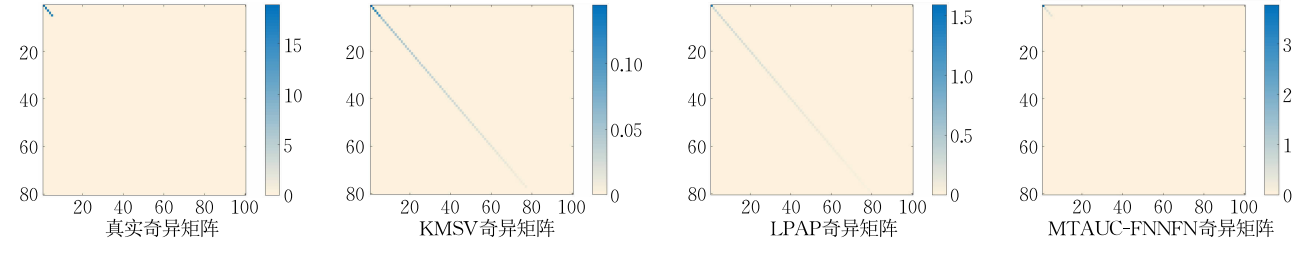

图 10　在 Simu4 数据集上通过 3 种代表算法训练得到的模型参数矩阵的奇异值矩阵

SUN Yu-Chen, M. S. candidate. His research interests include adversarial learning and multi-label learning.

XU Qian-Qian, Ph. D., professor. Her research interests include statistical machine learning, with applications in multimedia and computer vision.

WANG Zi-Tai, Ph. D., candidate. His research interests include machine learning and data mining, with special focus on the metrics of ML tasks.

YANG Zhi-Yong, Ph. D., associate professor. His research interests lie in machine learning and learning theory, with special focus on AUC optimization, meta-learning/multi-task learning, and learning theory for recommender systems.

HUANG Qing-Ming, Ph. D., professor. His research areas include multimedia computing, image processing, computer vision and pattern recognition.

Background

In recent years, machine learning has developed rapidly and achieved brilliant results. However, it is well-known that the success of machine learning relies on a large number of data to a considerable extent. In some scenarios, it might be hard to collect enough data for model training. To address this problem, multi-task learning (MTL) has attracted the attention of many researchers. To be specific, MTL methods aim to share knowledge among multiple tasks, thereby improving the model performance on each task. Compared with traditional algorithms, MTL is capable of utilizing the intrinsic relationships among related tasks and thus relies less on large-scale datasets. What's more, it has been proved that the MTL classifier has a more impressive generalization performance than being trained via the general method named single-task learning (STL). Hence, in the past decade, MTL has been applied to various kinds of areas such as natural language processing, computer vision, medical named entity recognition, medical image analysis and so forth.

Although MTL can present outstanding performance in sharing knowledge, it is not always true among unrelated tasks. When the multi-task model shares irrelevant knowledge, it could produce a serious performance degradation with a high probability. This notorious phenomenon, known as "negative transfer", has received widespread attention. Specifically, when the model studies and shares irrelevant knowledge, the negative transfer probably misleads the trained model in the wrong direction. In view of this, the regularization operation of the parameters of the multi-task model becomes the major idea of researchers. Among these approaches, the low-rank approach, as an effective kind of rank minimization method, has achieved numerous successes in recent MTL studies. One key assumption behind low-rank MTL algorithms is that the number of singular values of zero in parameters is almost equivalent to the number of irrelevant knowledge. Meanwhile, the number of singular values of non-zero determines the rank of the parameter matrix. Hence, to inhibit the negative transfer, it becomes natural to penalize the small singular values to zero and keep several main singular values. And empirical results also demonstrate that this low-rank structure encourages sharing related knowledge and ignoring unrelated tasks.

As low-rank MTL develops further, the standard for it has gradually grown higher. In some real-world applications, like rare disease detection and risk content identification, it is a common case that only imbalanced training data are available. While existing low-rank MTL methods generally focus on classifying balanced data. The rise of recognition requirements for imbalanced data makes prior arts in this direction hard to perform well. To solve this problem, some researchers introduce the Area Under the ROC curve (AUC) metric to MTL and have gained a few achievements in minimizing the AUC loss. It is well-known that AUC is insensitive to label distribution, and this appealing property determines that AUC can be a more appropriate performance metric under imbalanced scenarios. To the best of our knowledge, there still exist few studies that study low-rank algorithms for multi-task AUC learning. When we add the AUC to general MTL simply, the model still possibly tends to produce a few negative transfer problems so that making the generalization performance of models slump. Thus, it gives us a reason to seek a low-rank multi-task AUC learning method to satisfy the classification tasks with both insufficient and imbalanced data.

However, directly replacing the original pointwise loss with the AUC loss might induce an unfavorable efficiency issue. Specifically, the per-iteration complexity can reach $O(Ln_{i,+}n_{i,-})$, where L is the number of tasks and $n_{i,+}$, $n_{i,-}$ are the number of positive/negative samples in the i-th task, respectively. To handle this problem, we further introduce a novel reformulation of the AUC loss, where the pairwise terms are replaced with a pointwise optimization problem. On top of this, the overall complexity decreases to $O(L(n_{i,+}+n_{i,-}))$. In the end, we achieve an efficient multi-task AUC learning algorithm based on an adaptive low-rank approach.