基于多任务深度神经网络的企业纳税行为甄别研究*

李国锋　李祚娟　王哲吉

内容提要：随着数字经济时代的到来，丰富的数据资源有利于全面精准地刻画企业纳税情况，但数据来源广、类别不平衡以及噪音多等问题，也给企业纳税行为的甄别工作带来挑战。本文融合企业报表以及证监会、海关和税务等部门的多来源涉税数据，基于K-S检验和随机森林算法，构建了企业纳税行为甄别指标体系；将不同行业企业纳税行为甄别工作视为不同任务，提出基于多任务深度神经网络的企业纳税行为甄别模型，充分利用了不同行业任务间的相关性和差异性信息；针对样本数据集不平衡问题，引入焦点损失函数进一步改进了甄别模型。研究发现，相对于传统Logistic、支持向量机和神经网络等单任务模型，本文多任务模型的企业纳税行为甄别能力、泛化能力和稳健性更强。当模型预测某企业纳税不遵从的概率超出阈值时，即可判定该企业为重点稽查对象，以辅助税务部门提升稽查效率。本研究为政府智慧税务治理工作提供了新的思路。

关键词：多源数据；多任务深度神经网络；企业纳税行为甄别

DOI: 10.19343/j.cnki.11−1302/c.2022.07.011

中图分类号：C81；F812　　**文献标识码：**A　　**文章编号：**1002−4565(2022)07−0137−13

Research on Corporate Tax Paying Behavior Identification Based on Multi-Task Learning in Deep Neural Networks

Li Guofeng　Li Zuojuan　Wang Zheji

Abstract: With the advent of the digital economy era, rich data resources are conducive to comprehensively and accurately depicting the tax payment situation of enterprises. However, multiple sources, imbalances, and noise in the data also pose challenges to the identification of corporate tax paying behavior. This paper integrates tax-related data from multiple sources such as corporate statements, China Securities Regulatory Commission, customs and tax authorities. Based on the K-S test and the random forest algorithm for feature filtering, we construct a set of indicators for screening corporate tax paying behavior. This paper proposes a multi-task deep neural network-based corporate tax paying behavior identification model with tax behavior identification in different industries as different sub-tasks, which takes into account the correlation and heterogeneity between the different industry tasks. Aiming at the imbalance of the sample data set, the focus loss function is introduced to improve the model. The study results show that compared with the traditional single-task models such as logistic, support vector machine and neural network, the multi-task model constructed in this paper has better performance in tax behavior identification, generalization, and robustness. When the model predicts that the probability of tax non-compliance of an enterprise exceeds the threshold, the enterprise can be identified as a key audit target

*基金项目：国家社会科学基金一般项目“多源数据融合下企业纳税行为甄别智能学习方法研究”（19BTJ023）。

to improve the efficiency of the tax authorities' audit. This study provides a new idea for the government's smart taxation governance.

Key words: Multi-Source Data; Multi-Task Learning in Deep Neural Networks; Corporate Tax Paying Behavior Identification

一、引言

企业纳税行为是指企业是否依法纳税或如何纳税的各种行为表现，一般分为纳税遵从和纳税不遵从。纳税不遵从行为典型表现为偷税、欠税、骗税或抗税等形式，且在当今数据来源丰富、信息网络发达的数字经济时代，日益呈现出多样化、复杂化和隐蔽化趋势。税收流失依旧是政府面临的难题之一，税务稽查仍是税务部门的重点工作。据报道，自2018年8月以来，国家税务总局联合多部门持续开展打击企业虚开骗税违法犯罪行动，至2020年10月底，共检查出口企业4092户，挽回税款流失达270.88亿元[①]。

传统税务稽查方法主要包括人工选案和计算机辅助选案，典型的有判别函数系统、峰值分析法、财务报表分析和Tobit模型等（李选举，2000；González和Velásquez，2013）。人工选案工作量大且繁杂，主要依赖于选案人员的专业水准和主观判断；基于指标的计算机辅助选案，部分指标体系的建立和指标权重的匹配也是根据人工经验而得（娄元英和楼文高，2014）。当选案人员技术水平或经验不足时，会降低选案结果的准确性。随着现代信息技术的发展，2015年国家税务总局印发《"互联网+税务"行动计划》[②]，明确提出"智慧税务"的发展理念，持续鼓励运用互联网思维，引入云计算和人工智能等技术，深挖细掘税收大数据这座"金山银库"。

近年来，诸如决策树、Logistic回归、支持向量机和神经网络等各种机器学习方法被应用到税务稽查选案工作中（Dhiman等，2013；佘镜怀和李亚民，2015；Didimo等，2018），较大地提升了企业纳税行为甄别工作的效率和准确率。其中，鉴于模糊神经网络、BP（Back Propagation）神经网络和深度神经网络等方法在关于对噪声数据的鲁棒性和容错性以及有效识别处理涉税指标间的非线性关系等方面存在优势，因而其在税务稽查选案研究中受到关注（刘尚希和孙静，2016；Kleanthous和Chatzis，2020）。Kleanthous和Chatzis（2020）针对增值税审计选案问题，设计了一个新颖的变分自动编码器深度神经网络模型，与塞浦路斯税务局合作开发并部署该方案，获得了76%的样本外预测准确性。

虽然神经网络等方法被广泛应用于企业纳税行为的甄别研究中，但仍存在局限性。一方面，数据资源利用范围有限，多数研究主要采用企业财务指标数据（娄元英和楼文高，2014）。另一方面，不同行业企业纳税行为甄别工作间的关系利用较少，多数研究将不同行业数据直接合并作为一个整体或只针对某单一行业数据展开研究。然而，各行业发展程度不一，经营管理活动和适用税种税率不同，会计核算和财务管理制度也存在差异，这些差异可能会对不同行业企业纳税行为产生不同的影响。针对这类数据场景，如果按传统处理做法只针对某单一行业企业数据集独立建模，可能会忽略不同行业企业纳税行为甄别工作间的相关性信息；如果将所有行业企业数据混合为整体数据集建模，则又可能会忽略其差异性信息。加之，在税务稽查的实际工作中，存在纳税不遵从行为的上市

①参见中华人民共和国中央人民政府网：打击虚开骗税专项行动延长至2021年6月底。http://www.gov.cn/xinwen/2020-12/28/content_5573903.htm。

②参见国家税务总局：关于印发《"互联网+税务"行动计划》的通知。http://www.chinatax.gov.cn/n810341/n810755/c1843071/content.html。

企业样本数相对较少，所收集到的数据集呈现出较强的类别不平衡性，而传统分类器为了保证模型的整体预测精度，会优先考虑多数类样本的准确率（Lin等，2017），这将导致对纳税不遵从企业少数类样本的误分率较高。

本研究正是针对上述问题展开的。第一，基于文献调查与理论分析，收集了企业报表以及中国证券监督管理委员会（以下简称证监会）、海关和税务部门等多来源涉税数据，整合不同来源的特征指标，形成初始的可选指标集；利用K-S检验（Kolmogorov-Smirnov Test）和随机森林算法，构建企业纳税行为甄别指标体系。第二，借鉴在自然语言处理、图像识别、疾病预测等领域中表现出优良性能的多任务深度神经网络算法（Chen等，2014；El-Sappagh等，2020），将不同行业企业纳税行为甄别视为不同任务联合建模，设计了含有共享隐藏层和特定任务层的多任务深度神经网络模型，以充分利用不同行业任务间的相关性和差异性信息①。第三，引入焦点损失函数，进一步改进甄别模型，使之更加关注少数类纳税不遵从企业样本，以处理数据集的不平衡性问题。第四，以制造业上市企业样本数据为例，应用本文所构建的企业纳税行为甄别模型，进行实验比较分析和样本外预测，以验证模型的企业纳税行为甄别能力、泛化能力和稳健性。

二、数据来源与指标体系构建

（一）基于不同数据来源的可选指标

本文的指标选取主要基于以下考虑。首先，文献研究发现，企业纳税行为甄别相关指标的选取基本以企业财务报表为基础，使用频率较高的指标有企业资产负债率、销售费用率和存货周转率等财务指标。现实中，企业常常会采取隐藏收入、虚增成本费用、虚构原始凭证等手段以减少税源或推迟纳税，能够揭示该类行为的指标有销售收入成本率、销售毛利率、流动比率和资产周转率等。如果企业盈利能力指标明显异常或“常亏不倒”，此类企业也可能存在一定的税收风险。另外，不同股权性质和结构对上市企业税收筹划的非税成本具有不同影响，可能导致控股股东以及高层管理者产生不同的税收行为（郑红霞和韩梅芳，2008），因此企业股东性质也是本文考虑的指标。

其次，来自证监会、海关部门的数据，主要包括企业是否因信息披露违法违规、涉嫌内幕交易和假报进出口等行为而受到通报或处罚等事项，旨在揭示企业经营活动与其纳税行为之间的关系。来自税务部门的数据，主要包括企业在某段时间内因偷税、欠税和骗税等行为受到通报或处罚等事项，旨在考察企业是否存在纳税不遵从行为。以上指标数据均是确定企业纳税行为甄别模型目标变量的基础。

最后，考虑到数据获取的可行性及行业代表性，本文以制造业上市企业的纳税行为甄别为研究对象，依据制造业各子行业数量占比、存在纳税不遵从行为企业数量占比等原则，选取计算机、通信和其他电子设备制造业，汽车制造业，医药制造业，化学原料和化学制品制造业，电气机械和器材制造业和造纸业等6个制造业子行业的企业为研究样本。财务指标和股权指标主要来自于锐思金融研究数据库②2011—2018年的上市企业报表数据；证监会、海关及税务部门对上市企业的处罚公告数

①不同文献关于多任务间的相关性和差异性的理解不同。一是基于领域知识和共享特征方面的解释，如Caruana（1997）认为，如果不同任务共享某些相同特征进行决策，则表示任务之间存在一定的相关性，而不同任务通过各自特定的特征进行决策，就体现出了它们的差异性。二是基于分类边界和数据生成机制方面的解释，如Xue等（2007）提出，如果两个任务的分类边界（权重向量）接近，则这两个任务是相似的。Maurer等（2016）提出，如果两个任务中的样本数据分布产生自同一类概率分布变换，则两个任务是相关的。本文借鉴Caruana（1997）的观点来解释不同行业任务间的相关性和差异性。

②锐思金融研究数据库是由来自清华大学、北京大学等单位的著名专家参与，参照国际通用数据库设计标准，结合中国金融市场实际情况设计而成的，涵盖了财政、经济、金融等多个领域的数据信息，可为实证研究、学科与实验室建设等提供强力支持。

据，包括事件涉及的企业名称及代码、发生的时间和具体违规事项内容等，主要来自于锐思金融数据库2011—2018年公司重大事项违规处罚模块，而部分处罚公告数据则来自于相关部门官方网站和巨潮资讯网等第三方网络平台，如国家税务总局网站中的重大税收违法失信案件信息公布栏。具体数据来源与可选指标描述如表1所示。

表1　　基于不同数据来源可获取的指标

数据来源	指标描述	
上市企业报表	财务指标	盈利能力指标：资产报酬率、销售成本率和营业利润率等 偿债能力指标：流动比率和速动比率等 成长能力指标：营业收入增长率、净利润增长率和净资产收益率等 营运能力指标：总资产周转率、流动资产周转率和应收账款周转率等 现金流量指标：现金流动负债比、总资产现金回收率和净利润现金含量等 股票情况指标：市盈率、市净率和市销率等
	股权指标	股东性质：国有股东标记为1，其他股东等标记为0 股东规模：股东户数和户均持股数等
证监会	企业是否存在信息披露违法违规、涉嫌内幕交易和定期报告存在虚假记载等事项，用二元虚拟变量表示（是标记为1，否标记为0）	
海关部门	企业是否存在进出口货品或价格申报不实、虚假单证和假报进出口等事项，用二元虚拟变量表示（是标记为1，否标记为0）	
税务部门	企业在某段时间内是否因偷税、欠税和骗税等受到税务局的通报或处罚等事项，用二元虚拟变量表示（是标记为1，否标记为0）	

根据表1采集到的多源数据无法直接用于模型构建与分析，需要将不同来源的数据进行融合处理（Boström等，2007）。首先，将来自于上市企业报表以及证监会、海关和税务等部门的样本数据，按照企业代码进行匹配；然后，根据行业类型将同一子行业的企业样本融合在一起。融合后的数据集有部分缺失值，横向看，若某样本企业信息缺失超过20%，则直接删除该条样本记录；纵向看，若某个特征指标所对应的数据缺失超过20%，则直接删除该指标。针对信息缺失不多的企业样本，若是数值变量，采取相邻年份均值插补法进行填补；若是类别变量，则找出与该缺失值相匹配的非缺失变量样本数据，用其观察值进行填补。为消除样本数据量纲差异的影响，本文选用Z-Score方法对数据进行标准化处理。最终获得制造业6个子行业的上市企业样本数据，样本总数为3522条，指标数为112个。处理后的数据集呈现出较强的类别不平衡性，少数类样本与多数类样本比例约为1:34，不平衡程度最高的数据集所属行业为化学原料和化学制品制造业，最低的为造纸业。

（二）分行业企业纳税相关指标的K-S检验

本文将不同行业的企业纳税行为甄别工作视为不同任务，引入多任务机器学习方法开展建模。建模前，首先利用K-S检验考察不同行业指标是否具有判别企业纳税行为遵从与否的能力。

K-S检验是比较两组数据的经验分布是否有显著差异的一种非参数检验方法。根据企业纳税行为遵从与否将企业样本分为两类：一类为纳税遵从企业样本，另一类为纳税不遵从企业样本。若某个指标在两类样本观测数据上的经验分布差异越大，相应指标的K-S检验显著性越高，则表明该指标对企业纳税遵从与否的区分能力越强。

以计算机、通信和其他电子设备制造业数据为例，该数据集包含896个企业样本、112个指标。图1绘出了营业收入增长率和总资产现金回收率两个指标各自在两类企业样本中经验分布比较的结果。其中，横坐标轴为指标观测值，纵坐标轴为其经验分布估计值，营业收入增长率和总资产现金回收率的K-S检验统计量值分别为1.689和0.860，P值分别为0.007和0.450。结果表明，营业收入增长率在两类样本企业中的分布差异显著大于总资产现金回收率在两类样本企业中的分布差异，营业收入增长率指标具有更强的区分能力。

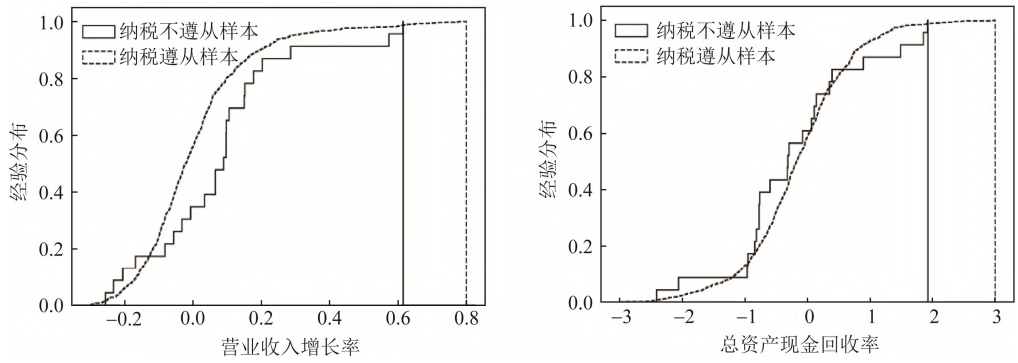

图1　计算机、通信和其他电子设备制造业中两个指标的K-S检验示例

继而分别对制造业6个子行业的各指标进行K-S检验，在α=10% 的显著性水平下，共有87个指标检验结果显著。表2列出的是至少在3个子行业中均显著的部分指标检验结果，从中可以看出，同一行业中有的指标K-S检验是显著的，有的指标是不显著的；同一指标在不同行业中的K-S检验统计量的值也不相同，表明这些指标在不同行业中的区分能力存在差异。

表2　　制造业6个子行业部分相关指标的K-S检验结果

行业 / 指标	电气机械和器材制造业	计算机、通信和其他电子设备制造业	汽车制造业	化学原料和化学制品制造业	医药制造业	造纸业
管理费用率	1.518**	1.689***	1.736***	0.885	2.242***	1.193
销售费用率	2.099***	1.047	0.972	1.486**	1.220*	1.826***
长期资产适合率	1.259*	0.713	1.264*	0.858	1.700***	1.826***
固定资产周转率	1.487**	0.784	0.896	0.669	2.354***	1.267*
经营现金净流量	0.589	0.816	1.396**	1.377**	0.978	1.453**
净利润	1.571**	0.565	1.321*	0.936	0.822	1.752***
销售毛利率	1.807***	0.854	0.764	1.043	1.710***	1.752***
自由现金流量	0.794	1.823***	2.010***	1.309*	0.766	0.820
每股营业总收入	0.828	0.750	1.481**	0.890	2.195***	1.304*
市净率	1.339*	0.762	1.510**	0.594	1.491**	0.559
流动资产周转率	1.380**	0.638	0.859	0.673	1.816***	1.603**
销售期间费用率	1.566**	0.510	1.698***	1.041	1.596**	0.783
市盈率	1.615**	0.906	1.481**	0.625	0.616	1.304*
销售成本率	1.807***	0.854	0.764	1.043	1.710***	1.752***
每股资本公积金	2.106***	0.727	1.227*	0.829	1.694***	0.969
营业收入增长率	0.563	1.689***	0.547	0.821	1.527**	1.342*
营业周期	1.585**	0.635	0.878	0.853	1.444**	1.342*
每股留存收益	1.382**	0.831	1.500**	1.267*	1.008	1.043
营运资金	0.928	1.292*	2.029***	0.863	0.807	2.273***

注：*、**、***依次表示统计量在10%、5%和1%显著性水平下显著，临界值分别为1.22、1.36和1.63（Young，1977）。

（三）企业纳税行为甄别指标体系构建

本文的研究目标是甄别企业纳税行为遵从与否，属于二分类问题。为了提高模型的运行效率和预测精度，采用随机森林等方法进一步精炼特征指标集，构建企业纳税行为甄别指标体系。

首先，考虑到不同行业含有区分能力显著的指标不完全相同，分别对6个子行业数据集利用随机森林算法进行指标筛选。具体操作为：分别计算各行业指标的重要性评分并递减排序，各保留前$\sqrt{F}$个指标（Geurts等，2006），F为数据集中的指标总数。

假设所构建的随机森林模型共含有T棵分类树，指标X_j在第t棵树上的重要性评分，定义为使用指标X_j进行空间切分时导致的基尼不纯度变化量，而在随机森林上的重要性评分（VIM_j），则为所有分类树基尼不纯度变化量的平均值（Jiang等，2009）：

$$VIM_j = \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M_t} IG_{mt} \times I\left(split\left(m\right)=X_j\right),\ j=1,2,\cdots,F \tag{1}$$

其中，M_t为第t棵分类树含有的内部节点数，$I\left(split\left(m\right)=X_j\right)$是示性函数，表示分类树的内部节点$m$依据指标$X_j$决策是否进行下一步切分，$IG_{mt}$为第$t$棵树中节点$m$分枝前后的基尼不纯度变化量：

$$IG_{mt} = GI_m - f_l GI_l - f_r GI_r \tag{2}$$

其中，GI_l和GI_r表示节点m分枝后的两个新节点l和r的基尼不纯度，f_l和f_r分别是节点m中属于节点l和r的企业样本比例。对二分类问题，节点m的基尼不纯度表示如下：

$$GI_m = 2P_m\left(1-P_m\right) \tag{3}$$

其中，P_m为样本企业在节点m中属于某一类的概率值（李航，2012）。利用该方法即可对每个行业指标的重要性评分展开测度。

以电气机械和器材制造业为例，利用随机森林算法计算其指标重要性评分，最终保留了每股现金及现金等价物、市盈率、销售成本率等11个指标，排序结果如图2所示。其他5个子行业的数据集特征指标选择按照同样流程进行，各自筛选出排名前11的指标。

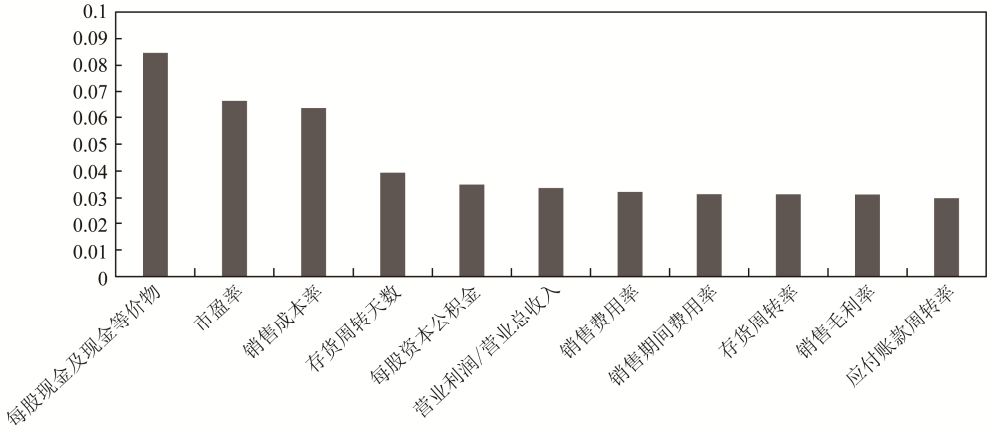

图2　电气机械和器材制造业部分指标重要性评分排序

其次，如果指标之间的相关性较强，说明可能存在冗余信息，不利于模型学习训练，进一步对保留下来的指标做相关性分析，删去高度相关指标中重要性评分较低的指标。

最后，将6个子行业筛选后剩余的指标取合集[①]，删除在所有子行业中K-S检验均不显著的指标，构建了包含49个指标的企业纳税行为甄别指标体系，如表3所示。

①之所以取合集，是由模型算法设计要求的。多任务模型要求不同任务输入的数据集应具有相同的特征指标，并将其添加到共享网络结构中，通过参数共享来学习不同任务特征的共享表示（shared representation），有利于提升模型的泛化能力。

表3　　企业纳税行为甄别指标体系

指标维度	指标名称
盈利能力	销售成本率、销售费用率、营业利润/营业总收入、销售毛利率、财务费用率、管理费用率、每股收益、利润总额/息税前利润、净资产收益率、销售期间费用率、净利润、营业总成本/营业总收入、营业收入增长率。
偿债能力	流动比率、速动比率、资产负债率、权益乘数、产权比率、股东权益/负债合计、流动资产/总资产、固定资产/总资产、经营现金净流量、流动负债/负债合计、有形净值债务率、每股资本公积金。
营运能力	存货周转率、应付账款周转率、流动资产周转率、固定资产周转率、存货周转天数、股东权益周转率、营运资金、应付账款周转天数、应收账款周转率、应收账款周转天数。
成长能力	利润总额增长率、净资产收益率、营业收入3年复合增长率。
现金流量	每股经营活动现金流量、每股现金及现金等价物、自由现金流量。
股票市场	市盈率、市净率、市销率。
股权结构	股东性质、股东户数、户均持股数。
证监会监管	证监会处罚。
海关监管	海关处罚。

三、相关模型设计

（一）模型构建与算法

目前，有关企业纳税行为甄别问题研究中，多数是基于单任务模式将不同行业的企业数据集作为整体，或只针对某单一行业的企业数据集展开研究。但现实场景是，不同行业的企业纳税行为甄别工作之间既有相关性也有差异性。为此，本文引入多任务深度神经网络方法（Multi-Task Learning in Deep Neural Networks，MTL-DNN），视不同行业企业纳税行为甄别工作为不同任务，通过构建共享隐藏层网络，使不同任务共享相同的参数来学习任务间的通用特征，同时设计拥有独自参数的特定任务层网络，来学习每个任务特定的特征，可有效降低模型过拟合的风险（Caruana，1997）。

本文构建的企业纳税行为甄别模型如图3所示，设共有 K 个任务。全部数据集 $\boldsymbol{X}$ 是由 K 个大小为 $n_k \times F\left(1 \leqslant k \leqslant K\right)$ 的样本数据矩阵组合构成的三维张量[①]，n_k 和 F 分别为第 k 个任务的样本量和特征指标数，每个任务的指标数相同；$\boldsymbol{y}$ 为目标变量，存在纳税不遵从行为的企业样本标签值记为1，否则为0。模型网络结构包括：数据输入层、两个共享隐藏层、两个特定任务层及特定输出层。鉴于

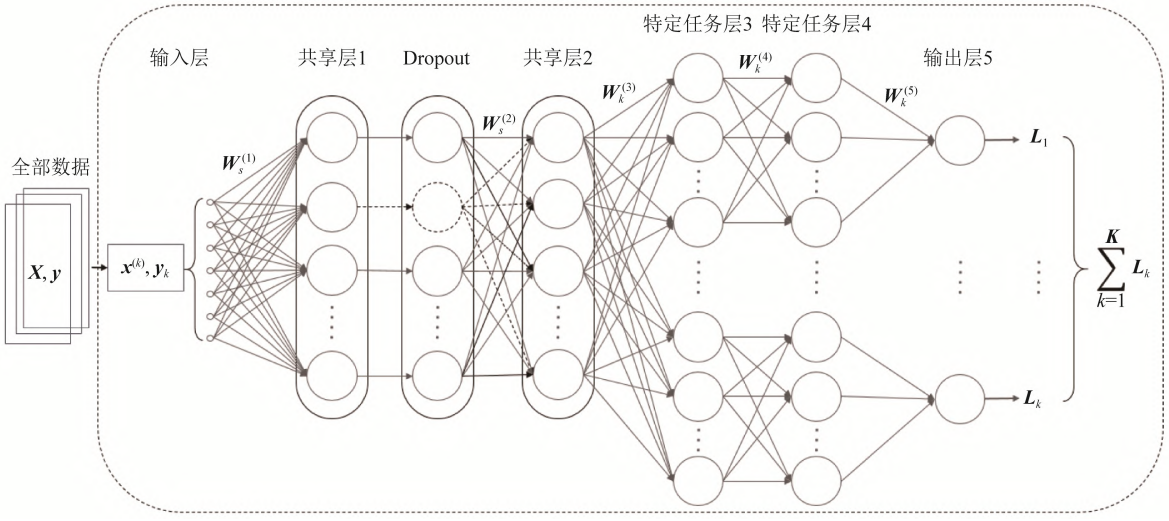

图3　基于MTL-DNN的企业纳税行为甄别模型

①张量（Tensor）是TensorFlow框架中最基础的数据结构，用来存储和处理多维数组。直观来看，向量可视为一维张量，矩阵可视为二维张量，三维张量可视为多个二维张量在深度方向上的组合。

深度神经网络的结构层数与节点存在个数多、学习速度慢和易出现过拟合等问题，因此对共享层1使用Dropout方法进行精炼。即在前向传播的过程中，随机舍弃一定比率 δ 的神经元节点，如图3中虚线标出的圆所示，设置其输出值为零。

本文采用Mini-batch方法训练网络，算法步骤如下。

步骤1：划分数据集，初始化网络参数。将数据集划分为训练集、验证集和测试集；设置迭代次数I、每批次数据中样本数等参数；随机初始化共享层权重矩阵 $\boldsymbol{W}_s^{(l)}$ 和偏置矩阵 $\boldsymbol{b}_s^{(l)}(l=1,2)$ 、特定层权重矩阵 $\boldsymbol{W}_k^{(l)}$ 和偏置矩阵 $\boldsymbol{b}_k^{(l)}(l=3,4,5)$ 。

步骤2：训练集上进行 K 个任务的交替学习，获取模型优化参数。

（1）假设轮到第 k 个任务训练，首先通过网络前向传播，输出 $\hat{\boldsymbol{y}}_k$ 和损失函数值 L_k：

$$\hat{\boldsymbol{y}}_k=\sigma(\boldsymbol{W}_k^{(5)}(\sigma(\boldsymbol{W}_k^{(4)}(\sigma(\boldsymbol{W}_k^{(3)}(\sigma(\boldsymbol{W}_s^{(2)}(\mathrm{Relu}(\boldsymbol{W}_s^{(1)}\boldsymbol{x}^{(k)}+\boldsymbol{b}_s^{(1)}))+\boldsymbol{b}_s^{(2)}))+\boldsymbol{b}_k^{(3)}))+\boldsymbol{b}_k^{(4)}))+\boldsymbol{b}_k^{(5)}) \tag{4}$$

$$L_k=f(\hat{\boldsymbol{y}}_k,\boldsymbol{y}_k) \tag{5}$$

其中，$\boldsymbol{x}^{(k)}=(\boldsymbol{x}_1^{(k)},\boldsymbol{x}_2^{(k)},\cdots,\boldsymbol{x}_F^{(k)})^{\mathrm{T}}$ 和 $\boldsymbol{y}_k$ 为从第 k 个任务数据集中选取的小批量样本数据，是共享层1的输入。假设第 l 层的神经元节点数为 M_l，$\boldsymbol{W}_s^{(1)}\in R^{M_1\times F}$ 和 $\boldsymbol{W}_s^{(2)}\in R^{M_2\times M_1}$ 分别为输入层到共享层1、共享层1到共享层2的权重矩阵，$\boldsymbol{b}_s^{(1)}\in R^{M_1\times 1}$ 和 $\boldsymbol{b}_s^{(2)}\in R^{M_2\times 1}$ 分别为两个共享层的偏置矩阵。$\boldsymbol{W}_k^{(l)}\in R^{M_l\times M_{l-1}}$ 和 $\boldsymbol{b}_k^{(l)}\in R^{M_l\times 1}(l=3,4,5)$ 分别为特定层的权重矩阵和偏置矩阵。共享层1的激活函数设为Relu函数，其他各层的激活函数设为Sigmoid函数①。$f(\cdot)$ 为损失函数②，用来衡量模型对第 k 个任务的输出与观测结果之间的差异。

其次通过误差反向传播，采用Adam算法（Kingma和Ba，2015）更新模型共享层的参数 $\boldsymbol{W}_s^{(l)}$、$\boldsymbol{b}_s^{(l)}$ 和特定层的参数 $\boldsymbol{W}_k^{(l)}$、$\boldsymbol{b}_k^{(l)}$。

（2）第 k 个任务训练完成后，继续进行第 $k+1$ 个任务的训练，再次进行前向传播，输出 $\hat{\boldsymbol{y}}_{k+1}$ 和损失函数值 L_{k+1}。此时，第 $k+1$ 个任务共享了第 k 个任务更新后的参数 $\boldsymbol{W}_s^{(l)}$ 和 $\boldsymbol{b}_s^{(l)}$，在保证其他任务特定层参数不变的前提下，通过误差反向传播，继续更新共享层的参数 $\boldsymbol{W}_s^{(l)}$、$\boldsymbol{b}_s^{(l)}$ 和特定层的参数 $\boldsymbol{W}_{k+1}^{(l)}$、$\boldsymbol{b}_{k+1}^{(l)}$。如此多个任务不断地进行交替训练，直至实现每个任务的目标优化。这样既归纳了隐含在相关任务训练信号中的有用信息，学习到了适用于多个任务的共享网络结构，又强化了每个任务特定信息的学习效率。

步骤 3：验证集上计算不同任务的损失函数 L_k，选择使整体损失函数式（6）最小化的参数方案。

$$\min_{1\leqslant i\leqslant I} L^i=\sum_{k=1}^{K}L_k \tag{6}$$

步骤 4：测试集上评估模型性能，输出模型的准确率和召回率等相关评价指标。

（二）模型损失函数设计

多任务深度神经网络模型分类问题中，常采用交叉熵损失函数，其形式为：

$$L_{CE}=\begin{cases}-\log\hat{y}_i, & y_i=1\\ -\log(1-\hat{y}_i), & y_i=0\end{cases} \tag{7}$$

其中，第 i 个样本的真实类别为 $y_i\in\{0,1\}$，$\hat{y}_i$ 表示预测第 i 个样本类别为 1 的概率，预测第 i 个样本类别为 0 的概率则为 $1-\hat{y}_i$。

鉴于本文使用的数据集中纳税不遵从企业数量远小于纳税遵从企业数量，属于不平衡数据。为此，引入 Lin 等（2017）针对不平衡数据分类问题提出的焦点损失函数（Focal Loss）改进模型，使

①Relu函数：$\mathrm{Relu}(x)=\max(0,x)$，Sigmoid函数：$\sigma(x)=1/(1+\mathrm{e}^{-x})$。

②分类问题中，具体选择怎样的损失函数将视问题场景和数据特点而定。本文模型损失函数的设计详见本章第（二）小节。

其更加倾向于关注易分类错误的企业和存在纳税不遵从行为的少数类企业。

首先，引入聚焦系数$\gamma(\gamma>0)$调整易分类企业和难分类企业的损失权重。模型对某样本企业的预测概率越高，表明对该企业的识别能力越强。以预测概率为基础，引入权重系数$(1-\hat{y}_i)^{\gamma}$，γ越大，难易分类企业的权重差别越大。其次，引入系数α，定义$1-\alpha$为某行业中少数类企业数目占全部行业少数类企业总数目的比例，以调整多数类企业和少数类企业在损失函数中的权重大小，增加少数类企业的损失权重。将二者结合起来，本文最终采用的焦点损失函数为：

$$L_{FL}=\begin{cases}-\alpha(1-\hat{y}_i)^{\gamma}\log\hat{y}_i, & y_i=1\\ -(1-\alpha)\hat{y}_i^{\gamma}\log(1-\hat{y}_i), & y_i=0\end{cases} \tag{8}$$

（三）模型性能评价指标

鉴于本文采用的数据集为不平衡数据，为全面评估模型分类效果，选择了准确率、召回率、*F1-Score*、*G-mean*和*AUC*作为模型性能的评价指标。准确率是分类正确的企业数量占全部样本企业数量的比例；召回率是被准确甄别出存在纳税不遵从行为的企业数量占全部纳税不遵从样本企业数量的比例；*F1-Score*是精确率和召回率的加权调和平均值；*G-mean*表示对纳税不遵从样本企业的分类精度和纳税遵从样本企业分类精度的几何平均值，可有效衡量不平衡数据的整体分类精度；*AUC*值（Area Under the Curve）是通过ROC曲线下的面积来比较不同甄别模型的分类效果，ROC曲线反映了模型在不同阈值下召回率和假正率的关系曲线。一般情况下，这些评价指标的值越高越好（Raschka，2019）。

四、企业纳税行为甄别实验分析

（一）模型效果比较

为了验证本文构建的MTL-DNN模型效果，实验中采用了以往研究中使用频率较高的Logistic回归、支持向量机和深度神经网络等单任务学习方法进行对比。将单任务学习方法针对6个子行业整体数据集建模的同时，还将6个子行业作为6个独立数据集分别建模，以便于比较分析。

针对不平衡数据集，将少数类样本采用SMOTE过采样方法进行处理（Chawla等，2002），过采样比例设置为3:1。整体数据集按70%和30%的比例分别随机划分为训练集和测试集，将训练集中20%的数据作为验证集，用于选择整体最优模型。测试集用于评估模型性能和稳健性，并采用10轮验证结果的平均值作为实验结果。

本文构建的MTL-DNN模型算法程序，是在Python语言环境中，基于TensorFlow框架实现的（Gonçalves等，2016）。利用计算图构建了如图3所示的从输入到输出的模型，包括损失函数设计和优化算法的选择。模型中，输入层为6个子行业的样本企业涉税数据集，每个行业的指标数均为49个。经实验，设置共享层和特定任务层的神经元数均为128个，Dropout函数失活概率δ=0.4；Adam优化算法学习率η=0.001，指数衰减率β_1=0.9，β_2=0.9999；焦点损失函数参数设定为γ=2。

模型训练效果如表4所示，为各模型算法分别针对整体数据集和6个子行业单一数据集训练建模得到的相应指标输出结果的均值。不难看出，单任务方法针对单一行业数据集的独立建模性能略优于针对整体数据集建模的结果，基于单一数据集构建的DNN模型的*Accuracy*、*G-mean*等指标结果较优。此外，本文构建的MTL-DNN模型，基于6个子行业数据集联合学习，所得到的召回率和*AUC*等评价指标均优于单任务模型，且其整体性能更优。

表4　　单任务学习和多任务学习方法模型训练效果对比

算法 / 指标	Logistic		SVM		DNN		MTL-DNN	
	整体数据集	单一数据集	整体数据集	单一数据集	整体数据集	单一数据集	交叉熵损失函数	焦点损失函数
Accuracy	0.7262	0.8179	0.8943	0.9172	0.8788	0.9225	0.9211	0.9394
Recall	0.7292	0.7840	0.6888	0.7304	0.6915	0.7388	0.8263	0.8741
AUC	0.7210	0.8133	0.8595	0.8625	0.7976	0.8602	0.8900	0.9190
F1-Score	0.7315	0.8266	0.7980	0.8083	0.7097	0.8148	0.8407	0.8840
G-mean	0.7323	0.8273	0.8149	0.8276	0.8533	0.9056	0.8992	0.9258

进一步考察表4最后两列，分别为使用交叉熵损失函数和焦点损失函数得到的MTL-DNN模型评估结果。首先，比较几种单任务模型和使用交叉熵损失函数的多任务模型的各项评价指标，以召回率为例，不难看出，使用交叉熵损失函数构建的多任务模型的召回率较单任务模型中的最好结果提升了4.23个百分点，更为精准全面地识别到了存在纳税不遵从行为的企业。其次，考察引入焦点损失函数改进的MTL-DNN模型的各项评价指标，分析发现，其召回率比使用交叉熵损失函数的MTL-DNN模型结果又提升了4.78个百分点，表明本文构建的基于焦点损失函数的MTL-DNN模型，更加关注存在纳税不遵从行为的企业样本，其*F1-Score*、*G-mean*等模型整体的性能评价指标也得到进一步提升。

综上所述，本文引入焦点损失函数构建的多任务深度神经网络企业纳税行为甄别模型，融合多来源数据，关注不同行业任务间的相关性和差异性信息，考虑样本数据集的不平衡问题，较大地提升模型的运行效率和整体预测性能。

（二）稳健性与可扩展性

为进一步检验模型的稳健性，分别将过采样的比例设置为5:1以及10:1等不同的数据不平衡比，对比使用交叉熵损失函数和焦点损失函数的MTL-DNN模型的多种评价指标，结果如表5所示，包括了6个子行业的各种指标输出结果以及其均值。可以发现，在不同的过采样比例下，针对不平衡数据使用焦点损失函数改进的MTL-DNN模型的性能更优；且随着不平衡比例的上升，使用焦点损失函数改进的多任务模型的召回率和*AUC*等指标优势进一步扩大。

以召回率为例，在10:1的过采样比例下，使用焦点损失函数较使用交叉熵损失函数的模型召回率提升了8.35个百分点，而在3:1的过采样比例下仅提升了4.78个百分点。实验结果表明，针对企业涉税样本数据不平衡的特点，使用焦点损失函数改进的多任务模型更具稳健性，能够有效地将注意力集中于存在纳税不遵从行为的企业上，并对这些企业能更精准全面地进行识别。

需要说明的是，本文所提出的基于焦点损失函数改进的MTL-DNN模型，虽然实验主要集中于制造业的部分行业数据集中，但该模型可以扩展到包含其他更多制造业子行业的税务稽查工作中。图4展示的是，在分别将子行业数随机组合从2个增加到6个的过程中，多任务模型的整体性能评价指标*F1-Score*和*G-mean*的对比结果。不难看出，随着不同子行业任务数目的增加，模型的整体性能渐优，表明了模型良好的任务扩展性能。同样，该模型具有良好的普适性，也可以便利地扩展到其他行业，如服务业的税务稽查工作[①]。

①运用本文提出的建模思路和算法程序对服务业中部分行业企业样本做了进一步的验证，鉴于篇幅及研究属于重复验证等原因，该部分测算结果未在正文中展示，感兴趣的读者可向作者索取。

表5　不同数据非平衡比例下两种损失函数的MTL-DNN模型效果对比

评价指标	子行业	3:1		5:1		10:1	
		交叉熵损失函数	焦点损失函数	交叉熵损失函数	焦点损失函数	交叉熵损失函数	焦点损失函数
Accuracy	电器	0.9221	0.9286	0.9421	0.9368	0.9560	0.9811
	化学	0.8963	0.9259	0.9096	0.9398	0.9528	0.9575
	计算机	0.9314	0.9078	0.9310	0.9368	0.9595	0.9625
	汽车	0.8993	0.9549	0.9667	0.9278	0.9604	0.9736
	医药	0.9097	0.9271	0.9021	0.9536	0.9692	0.9297
	造纸	0.9680	0.9920	0.9886	0.9886	0.9783	0.9785
	均值	0.9211	0.9394	0.9400	0.9472	0.9627	0.9638
Recall	电器	0.8511	0.8542	0.7037	0.7632	0.7875	0.8571
	化学	0.7586	0.8049	0.7419	0.8501	0.7077	0.7391
	计算机	0.7745	0.8846	0.6543	0.7500	0.6154	0.7544
	汽车	0.7638	0.8583	0.7519	0.7597	0.6800	0.8261
	医药	0.8205	0.8761	0.7615	0.8750	0.7027	0.8002
	造纸	0.9893	0.9667	0.9286	0.9474	0.8571	0.8750
	均值	0.8263	0.8741	0.7569	0.8242	0.7251	0.8086
AUC	电器	0.9022	0.9082	0.8426	0.8717	0.8368	0.9251
	化学	0.8463	0.8918	0.8450	0.9243	0.8904	0.8616
	计算机	0.8779	0.9000	0.8181	0.8655	0.8020	0.8682
	汽车	0.8529	0.9560	0.9194	0.8414	0.8478	0.9082
	医药	0.8817	0.8745	0.7159	0.9221	0.8115	0.8940
	造纸	0.9787	0.9833	0.9643	0.9737	0.9091	0.9315
	均值	0.8900	0.9190	0.8509	0.8998	0.8496	0.8981
F1-Score	电器	0.8696	0.8817	0.7755	0.8286	0.7786	0.8889
	化学	0.7586	0.8684	0.7541	0.8438	0.7577	0.7907
	计算机	0.8449	0.8251	0.7465	0.8197	0.7033	0.7748
	汽车	0.8010	0.9139	0.7346	0.7719	0.7879	0.8636
	医药	0.8312	0.8320	0.7581	0.8615	0.7805	0.8312
	造纸	0.9394	0.9831	0.9130	0.9730	0.8995	0.8750
	均值	0.8407	0.8840	0.7803	0.8497	0.7846	0.8374
G-mean	电器	0.9120	0.9234	0.9069	0.9245	0.9040	0.9542
	化学	0.8917	0.9314	0.8495	0.8809	0.8866	0.9074
	计算机	0.8807	0.8617	0.9034	0.9231	0.8913	0.9519
	汽车	0.8828	0.9278	0.8967	0.8926	0.9531	0.9419
	医药	0.8868	0.9155	0.8563	0.9096	0.9346	0.9210
	造纸	0.9411	0.9948	0.9933	0.9928	0.9579	0.9298
	均值	0.8992	0.9258	0.9010	0.9206	0.9212	0.9344

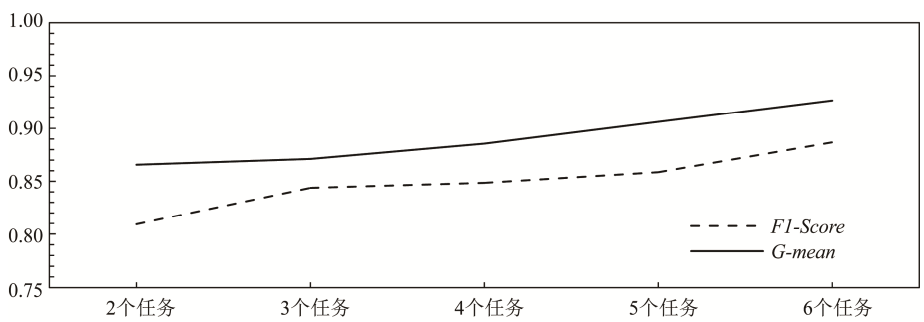

图4　不同任务数目下MTL-DNN模型效果对比

（三）样本外预测

为了检验所构建模型的实际应用能力，本文引入2019年的部分制造业企业数据进行样本外预测。引入的样本外企业包括被监管部门通报存在纳税不遵从行为的企业，标记为1；也包括至今未被通报有纳税不遵从行为的企业，标记为0。将上述企业的相关指标数据代入本文构建的已训练好的MTL-DNN模型中，估计各企业发生纳税不遵从行为的概率，以判定其是否可能存在纳税不遵从行为，具体甄别结果如附表1[①]所示。

由附表1可看出，本文构建的模型在实际应用中也表现出了较好的效果，预测概率的阈值设定为0.5，将实际预测概率大于50%的企业标记为存在纳税不遵从行为的企业，模型预测判定情况与实际情况相符，已被监管部门通报存在纳税不遵从行为的企业均被识别出来。基于来自国家税务局官网的样本企业的纳税信用评级情况（附表1最后一列），可以发现，每个行业中预测纳税不遵从概率较小的企业中大部分纳税信用评级为A，进一步验证了本文多任务模型的预测效能。

五、小结

本文融合了上市企业报表以及证监会、海关和税务等部门的多源涉税数据信息，形成初步的可选指标集；针对多源融合数据集中的高维指标，利用K-S检验和随机森林算法进一步筛选，确立了企业纳税行为甄别指标体系；利用焦点损失函数改进多任务深度神经网络算法，构建了分行业视角下的企业纳税行为甄别模型，以提升对存在纳税不遵从行为目标企业的甄别能力。

研究结果表明，一是相比于Logistic、SVM和DNN等单任务模型，本文将不同行业企业纳税行为甄别工作视为不同任务，构建的多任务深度神经网络企业纳税行为甄别模型的准确率、召回率和*AUC*等评价指标值更高，能够更为精准地甄别出具有纳税不遵从行为的目标企业。二是针对实际中存在的数据不平衡问题，引入焦点损失函数改进的多任务模型，较使用传统交叉熵损失函数的多任务模型，其召回率提高了4.78个百分点，进一步提升了对存在纳税不遵从行为少数类目标企业的甄别能力；另外，不同数据不平衡比例下的实验结果表明，引入焦点损失函数的MTL-DNN模型具有更强的稳健性。三是基于不同任务数和样本外数据集的实验结果表明，MTL-DNN模型具有良好的任务扩展性能和实际应用能力，当模型预测某行业企业纳税不遵从的概率超出阈值时，即可认定该企业为重点稽查对象，有利于辅助税务部门实时智能化地打击违反税法的企业。

实际研究中，本文还将所构建的多任务模型运用于服务业企业纳税行为甄别进行验证，取得了较好的效果，表明本文所构建的多任务模型具有良好的拓展能力和普适性，可以便利地扩展到除制造业外其他不同行业的企业纳税行为甄别工作中。本文融合多来源税收大数据，以及将不同行业企业纳税行为甄别工作视为不同任务提出的多任务建模思想，也可以应用到企业纳税信用评级、企业其他行为甄别等相关方面的分类预测问题研究中。

参考文献

[1] 李航. 统计学习方法[M]. 清华大学出版社, 2012.

[2] 李选举. Tobit 模型与税收稽查[J]. 统计研究, 2000(1): 46–50.

[3] 刘尚希, 孙静. 大数据思维: 在税收风险管理中的应用[J]. 经济研究参考, 2016(9): 19–26.

[4] 娄元英, 楼文高. 企业纳税评估指标和模型的研究综述[J]. 中国集体经济, 2014(30): 66–67.

[5] 佘镜怀, 李亚民. 税务稽查选案方法研究改进支持向量机[J]. 税务研究, 2015(11): 55–60.

[6] 郑红霞, 韩梅芳. 基于不同股权结构的上市企业税收筹划行为研究: 来自中国国有上市企业和民营上市企业的经验证据[J]. 中国软科学,

①因篇幅所限，基于样本外企业数据预测效果以附表1展示，详见《统计研究》网站所列附件。

2008(9): 122–131.
[7] Boström H, Andler S F, Brohede M, et al. On the Definition of Information Fusion as a Field of Research[J]. Neoplasia, 2007, 13(2): 98–107.
[8] Caruana R. Multitask Learning[D]. School of Computer Science Carnegie Mellon University, 1997.
[9] Chawla N V, Bowyer K W, Hall L O, et al. Smote: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357.
[10] Chen D, Mak B, Leung C, et al. Joint Acoustic Modeling of Triphones and Trigraphemes by Multi-task Learning Deep Neural Networks for Low-resource Speech Recognition[A]. IEEE International Conference on Acoustics, Speech and Signal Processing. 2014: 5592–5596.
[11] Dhiman R, Vashisht S, Sharma K. A Cluster Analysis and Decision Tree Hybrid Approach in Data Mining to Describing Tax Audit[J]. International Journal of Computers & Technology, 2013, 4(1): 114–119.
[12] Didimo W, Giamminonni L, Liotta G, et al. A Visual Analytics System to Support Tax Evasion Discovery[J]. Decision Support Systems, 2018, 110(6): 71–83.
[13] El-Sappagh S, Abuhmed T, Islam S, et al. Multimodal Multitask Deep Learning Model for Alzheimer's Disease Progression Detection Based on Time Series Data[J]. Neurocomputing, 2020, 412(10): 197–215.
[14] Geurts P, Ernst D, Wehenkel L. Extremely Randomized Trees[J]. Machine Learning, 2006, 63(1): 3–42.
[15] Gonçalves A R, Zuben F J V, Banerjee A. Multi-Task Sparse Structure Learning with Gaussian Copula Models[J]. Journal of Machine Learning Research, 2016, 17: 1–30.
[16] González P C, Velásquez J D. Characterization and Detection of Taxpayers with False Invoices Using Datamining Techniques[J]. Expert Systems with Applications, 2013, 40(5): 1427–1436.
[17] Jiang R, Tang W W, Wu X B, et al. A Random Forest Approach to the Detection of Epistatic Interactions in Case-Control Studies[J]. BMC Bioinformatics, 2009, 10: 1–12.
[18] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[A]. International Conference on Learning Representations (ICLR), 2015: 1–15.
[19] Kleanthous C, Chatzis S. Gated Mixture Variational Autoencoders for Value Added Tax Audit Case Selection[J]. Knowledge-Based Systems, 2020, 188: 1–9.
[20] Lin T Y, Goyal P, Girshick R, et a1. Focal Loss for Dense Object Detection[A]. IEEE International Conference on Computer Vision, 2017: 2999–3007.
[21] Maurer A, Pontil M, Romera-Paredes B. The Benefit of Multitask Representation Learning[J]. Journal of Machine Learning Research, 2016, 17(1): 2853–2884.
[22] Raschka S, Mirjalili V. Python Machine Learning[M]. Birmingham: Packt Publishing Ltd, 2019.
[23] Xue Y, Liao X, Carin L, et al. Multi-task Learning for Classification with Dirichlet Process Priors[J]. Journal of Machine Learning Research, 2007, 8: 35–63.
[24] Young I T. Proof without Prejudice: Use of the Kolmogorov-Smirnov Test for the Analysis of Histograms from Flow Systems and other Sources[J]. Journal of Histochemistry and Cytochemistry, 1977, 25(7): 935–941.

作者简介

李国锋，山东财经大学统计与数学学院教授。研究方向为应用统计学、数据科学与大数据分析、商务智能及经济行为分析等。

李祚娟（通讯作者），山东财经大学统计与数学学院博士研究生。研究方向为经济统计学、计量经济学、机器学习。电子邮箱：lzj901231@163.com。

王哲吉，山东财经大学统计与数学学院硕士研究生。研究方向为大数据分析与机器学习。

(责任编辑：张　亮)