第33卷第1期 管理科学 ISSN 1672-0334 Vol. 33 No. 1 30-45
2020年1月 *Journal of Management Science* January, 2020

基于深度神经网络的企业信息系统用户异常行为预测

尹 隽[1,2], 彭艳红[2], 陆 怡[3], 葛世伦[2], 刘 鹏[2]
1 江苏科技大学 江苏高校哲学社会科学重点研究基地, 江苏 镇江 212003
2 江苏科技大学 经济管理学院, 江苏 镇江 212003
3 中国工商银行 软件开发中心, 上海 200120

摘要: 随着企业信息化水平的不断提高, 企业核心业务越来越依赖于信息系统的可靠运行, 任何信息系统用户进行的异常操作都可能给企业带来不可估量的损失。企业更加重视用户异常行为可能对企业造成的负面影响, 如何有效预测企业信息系统的异常行为成为当前的研究问题。

设计企业信息系统用户异常行为的预测框架, 明确企业信息系统用户异常行为的界定标准, 基于用户日志数据, 在已有研究基础上加入业务维度构建特征模型, 采用深度神经网络方法进行用户异常行为预测。通过与经典统计方法和传统机器学习方法对比进行模型评估, 以某船舶企业为例进行实验分析, 初步验证该预测框架的有效性。

研究结果表明, 加入业务特征后的特征模型整体表现更好, 召回率、查准率和*AUC*分别提高3.52%、2.16%和3.36。基于数据驱动的深度神经网络模型可以层层抽取用户异常行为的抽象特征, 提高各个特征对异常行为预测的效率。与多重线性回归方法相比, 深度神经网络的召回率和查准率分别提高16.49%和7.46%; 与支持向量机算法相比, 召回率、查准率和*AUC*分别提高3.09%、5.09%和0.08。进一步比较3个部门的模型发现, 在与企业业务直接相关的业务部门和职能部门, 用户异常行为能被更好地识别出来, 而信息部门的分类效果欠佳。

研究结果为企业提供了一种可能适用于企业信息系统用户异常行为的预测框架, 有助于企业对用户异常行为进行预测, 从而及时采取措施以降低用户异常行为可能对企业造成的负面影响。

关键词: 企业信息系统; 深度神经网络; 用户异常行为; 特征工程; 预测
中图分类号: C931.6 **文献标识码**: A **doi**: 10.3969/j.issn.1672-0334.2020.01.003
文章编号: 1672-0334(2020)01-0030-16

收稿日期: 2019-07-04 **修返日期**: 2019-11-28
基金项目: 国家自然科学基金(71331003, 71972090, 71871108); 江苏省研究生科研创新计划项目(KYCX_19-1650)

作者简介: 尹隽, 管理学博士, 江苏科技大学江苏高校哲学社会科学重点研究基地和经济管理学院副教授, 研究方向为信息系统复杂性、信息系统使用和云计算等, 代表性学术成果为"信息系统'功能任务网络'中位置及关系特征对企业信息系统使用的影响研究", 发表在2018年第2期《系统工程理论与实践》, E-mail: bamhill@163.com
彭艳红, 江苏科技大学经济管理学院硕士研究生, 研究方向为信息系统使用等, 代表性学术成果为"Evolution analysis of newcomer-task network structure of enterprise information system: a case study of a shipbuilding enterprise", 发表在《International Society for Knowledge and Systems Sciences》(ISBN 978-981-15-1208-7), E-mail: 1206394813@qq.com
陆怡, 中国工商银行软件开发中心信息科技助理经理, 研究方向为信息系统运维和机器学习等, 代表性学术成果为"Analysis of dynamic complexity feature of information system data based on visualization", 发表在《2018 1st International Conference on Information Management and Management Science》(ISBN 978-1-4503-6486-7), E-mail: deerlet1993@163.com
葛世伦, 管理学博士, 江苏科技大学经济管理学院教授, 研究方向为管理信息系统等, 主持国家自然科学基金项目"基于云的管理信息系统再造研究"(71331003), E-mail: jzgsl@jzerp.com
刘鹏, 管理学博士, 江苏科技大学经济管理学院讲师, 研究方向为复杂社会网络演化分析等, 代表性学术成果为"Structure and evolution of co-authorship network in an interdisciplinary research field", 发表在2015年第1期《Scientometrics》, E-mail: liupeng19821017@126.com

引言

企业信息系统用户异常行为是系统正常用户行为模式之外、对企业信息系统正常运行造成影响的行为[1]。随着企业信息系统应用的深入,用户异常行为的威胁也日益严重,不仅影响用户的工作质量和效率,甚至给企业造成经济损失,威胁到企业的安全[2-3]。如近10年中国银行业的违规金额损失事故中,由内部员工系统使用异常行为引发的事故占比高达51%[4];2018年4月,韩国三星证券因用户违规操作,造成企业损失高达1.87亿美元[5]。对企业信息系统的用户异常行为进行预测成为业界和学界广泛关注的焦点问题。

系统使用日志忠实地记录了系统用户的行为数据,使捕捉和分析系统用户的异常行为成为可能[6],对系统日志进行分析逐渐成为识别用户异常行为的有效手段。关于系统日志的分析,学界主要形成基于模型[7]、基于规则[8]和数据驱动[9-10]3类方法,但是,随着信息系统复杂程度的不断加大,前两种方法逐渐难以满足数量呈指数级上升的系统日志的分析需求[11]。因此,结合系统日志提出有针对性的、数据驱动的用户异常行为分析方法成为学界广泛探索的开放性课题。

本研究采用数据驱动的系统日志分析方法,针对企业信息系统异常行为预测的问题,结合企业信息系统特征界定企业信息系统用户的异常行为,并构建一个加入业务维度特征的新特征工程方案,采用更能抽象复杂行为模式的深度神经网络方法进行预测,以船舶制造A企业日志数据为实验环境进行验证,初步实验结果表明,该预测框架在分析和预测企业信息系统用户异常行为问题上具有更好的效果。

1 相关研究评述

1.1 信息系统领域用户异常行为

ANDERSON[12]最早将信息系统用户异常行为定义为滥用对系统及其数据的授权访问权限。随后,DENNING[13]提出与软件系统常规行为完全不同的用户行为是异常行为。此外,由于异常行为预测与用户行为模式之间的密切关系,ZHANG et al.[1]称异常行为为系统正常用户行为模式之外的行为。

按照信息系统的架构层次,可以将信息系统用户异常行为分为网络层网络用户异常行为、数据层数据库用户异常行为、表达层用户鼠标异常行为和应用层社交网络用户异常行为等,详见表1。以网络用户异常行为研究最为活跃和深入,具体分为基于主机[14]、基于网络[15]和混合型[16-18]3类问题的研究,目前相关研究成果已运用到政府、能源、教育、电子商务、医疗和制造业等各个领域,多数已开发了对应的入侵检测系统。近些年,随着信息系统承载的数据越来越丰富且重要,有学者开始关注信息系统数据层的用户异常行为。李海斌等[19]提出一种无监督的检测数据库内部合法用户行为的方法;SALLAM et al.[20]研究基于query语句向量化特征的异常检测方法。此外,在信息系统越来越重视用户体验的同时,表达层的用户异常行为也开始受到关注。ZHENG et al.[22]根据用户标识,使用支持向量机分类器构建用户特征的鼠标移动模式;许洪军等[23]通过卷积神经网络分析用户鼠标轨迹,检测用户异常的鼠标行为。但相对而言,已有研究对应用层的关注并不够,仅有部分研究对社交网络用户异常行为进行探讨,针对恶意用户[24-26]、僵尸用户[27]、垃圾用户[28]和虚假用户[29]等的识别进行研究,但这仅仅是应用层的一个领域。实际上,企业信息系统用户异常行为的威胁不容忽视,因为企业内部用户能够通过系统驱动企业的核心业务,一旦异常行为造成损失,对企业而言可能就是致命的打击。

1.2 用户异常行为的特征模型

为保证预测方法的有效性,需结合具体情景选择有助于识别用户异常行为的特征。此外,还需要考虑数据获取和处理时的可行性和效率因素。目前在预测用户异常行为的研究中选取的特征主要分为两类,一类是用户的个体属性特征。李海斌等[19]在研究数据库用户异常行为时,选取用户角色和用户工作状态等用户属性特征;谈磊等[24]在分析社交网络恶意行为时选取用户资料为特征。另一类是用户的行为属性特征。李海斌等[19]选取数据库的单日内访问数据量、单日内访问不同表总个数作为特征;张艳梅等[30]在对新浪微博的异常用户行为进行分析时选取发文频率、发博文数和离线时间等行为属性;岳

表1 信息系统领域用户异常行为研究

Table 1 Research on User Abnormal Behavior in Information System Domain

信息系统架构层次	研究主题	来源
网络层	网络用户异常行为	杨宏宇等[14], BOUKHTOUTA et al.[15], ALJAWARNEH et al.[16], NAVARRO et al.[17], GEORGE et al.[18]
数据层	数据库用户异常行为	李海斌等[19], SALLAM et al.[20], KAMRA et al.[21]
表达层	用户鼠标异常行为	ZHENG et al.[22],许洪军等[23]
应用层	社交网络用户异常行为	谈磊等[24],岳虹等[25], ZHENG et al.[26], SARPIRI et al.[27], FIRE et al.[28], ZHANG et al.[29], 张艳梅等[30]

虹等[25]在对僵尸微博用户进行分析时选取转发比例和提及其他用户比例等属性。

1.3 预测用户异常行为的方法

根据已有研究,预测用户异常行为的方法可以分为统计方法和基于机器学习的方法。统计方法是指收集和分析用户行为数据并由数据得出结论的一系列方法,包括分析用户行为正常状态以及与正常行为不同的异常行为。然而,统计方法需要准确的统计分布,当统计特征值不明显或者变化较大时,误报率和漏报率高,而且随着数据量和特征维度的增长,还导致异常行为分析效率降低。例如,经典的多元线性回归方法(MLR)适用于线性相关情况的预测,且需事先筛选出对因变量影响较高的自变量。为此,很多研究通过机器学习方法取得了较好的预测效果,相应的方法有朴素贝叶斯方法(NB)、K近邻算法(KNN)、支持向量机算法(SVM)和神经网络算法等。朴素贝叶斯方法较适合小数据规模,且对于数据的表达形式较敏感,需要计算先验概率;K近邻模型的时间和空间复杂度都比较高,效率相对比较低;SVM能较好地解决高维问题并提高泛化能力,在预测大规模日志行为数据的异常行为研究中的应用越来越广泛[24,26],但当特征变量较多时,分类效率有所降低;而神经网络算法以神经元数学模型为基础,通过模拟人脑学习新事物的方式工作,通过获取主题的行为模式特征,利用神经网络的识别、分类和归纳能力,实现对用户行为模式的预测,其优势在于效率和准确率高、适应性强,目前越来越多地应用在专门的网络防御和预测社交网络用户行为异常等任务中[19]。

1.4 评述

综上所述,针对用户异常行为数据的多层次、大规模和数据不平衡等特点,已有研究从信息系统架构的多个层次开展了许多有益的工作。然而,作为直接面向用户的信息系统应用层,其识别模式有别于其他技术层次,具有高度的领域相关性和更直接的行为后果,已有研究还有待进一步丰富和深化。①从应用领域方面,已有研究大多关注社交网络用户的异常行为,这些仅反映一个代表性领域。②从方法方面,由于用户异常行为数据的特征,相关工作主要基于机器学习的方法进行预测,对于小规模的用户异常行为数据,贝叶斯的识别效率较高,但计算复杂;对于相对高维的数据,SVM具有较好的性能,因而在预测异常用户行为研究中被广泛使用[24,26],但当特征变量较多时,分类效果并不好;神经网络算法能克服上述两种方法的缺陷,对当前大规模、高维的用户异常行为数据,其预测的表现更好[19],但收敛速度慢,且特征的抽取只有一层。基于上述分析,本研究关注应用层的企业信息系统领域,该领域的用户异常行为不仅影响用户工作质量,更重要的是直接影响企业效益,甚至企业安全;在方法方面选用深度神经网络,建立现有特征模型到高层次语义特征之间的映射关系,以提高预测的准确率。

2 企业信息系统用户异常行为预测框架

信息系统中的用户异常行为带来的负面影响程度不一,但都不可小觑,如用户在不允许登录系统的时间段内[1]向未经授权的目的地发送敏感数据、企图入侵计算机或无意中非法操作了信息系统,可能造成数据泄露和丢失等情况。此外,用户由于自身原因出现的工作超时行为[31]和不当退出行为,使用户工作效率降低或数据损坏和丢失,甚至造成直接或间接经济损失[32]。因此,结合企业信息系统特征和已有研究,本研究将企业信息系统用户异常行为定义为:当企业信息用户偏离正常行为模式,或者用户的行为有可能具有非法操作系统的嫌疑,对企业业务造成威胁时,称为异常。

在前述分析的基础上,本研究提出建立企业信息系统用户异常行为预测框架,见图1,该模型主要包括用户异常行为分类和界定、特征工程、模型训练和模型评估4个模块。其基本思想为:首先,本研究关注应用层用户行为异常,这种异常往往与时间和地点相关,具体包括无意产生、基于规则和基于知识3类异常[33],本研究模型将结合已有研究和企业信息系统的管理特征进行异常行为的进一步分类和界定。其次,由于企业信息系统本质是对各企业具体业务的管理,不同的业务类型直接影响用户行为模式,因此本研究提出将业务维度纳入特征模型,以提升特征的识别度。此外,考虑到用户异常行为的复杂性,本研究采用深度神经网络作为用户异常行为的预测方法,该方法更能适应多维度和大规模的数据,有助于建立现有特征模型到高层次语义特征之间的映射关系,从而提高预测的准确率。

2.1 用户异常行为分类和界定

用户异常行为的分类和界定是进行分析预测的起点,目前用户异常行为研究涉及到多种异常行为的分类,如按复杂程度分为简单异常和复杂异常[19],按发生的层次分为运输层异常和网络层异常等[34]。本研究考虑可能对企业信息系统产生不良后果的用户异常行为,从用户认知特征的角度进行异常行为分类,即结合ZHAO et al.[33]的研究将异常分为无意产生、基于规则和基于知识。此外,为了能够更明确地界定用户异常行为,梳理相应的界定标准,目前应用层的异常行为主要基于时间和地点进行界定。

综合已有研究以及对企业情况的访谈调研,将企业信息系统异常行为的分类、界定方式以及可能的情景和后果进行梳理,结果见表2。

(1)基于时间的用户行为异常界定

主要考虑两种时间,用户登录时间 T_{in} 和退出时间 T_{out},且均以秒计算。假设企业规定每天 CT_{in} 时刻开始工作,CT_{out} 时刻结束工作,考虑到真实的企业情景,合理的登录和退出时间可能允许有偏差,因此用一个偏离值 ΔT 修正合理登录和退出时间,相应地,可以界定的3类基于时间的异常分别为无退出时间异常、非正常时间登录异常和超出合理操作时间异常。

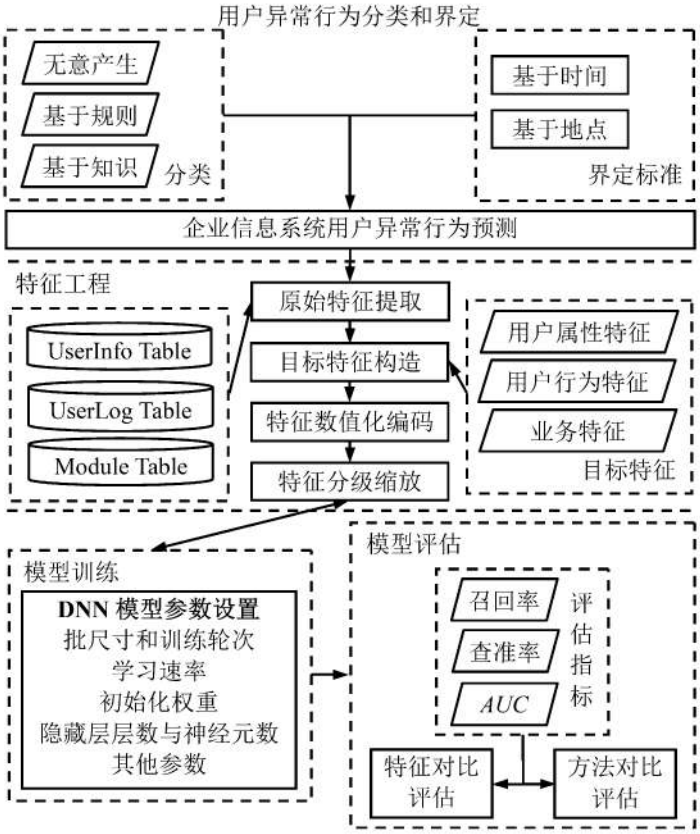


图 1 企业信息系统用户异常行为预测框架

Figure 1 Prediction Framework of User Abnormal Behavior in Enterprise Information System

表 2 企业信息系统用户异常行为分类和界定

Table 2 Classification and Dividing Line of User Abnormal Behavior of Enterprise Information System

分类	说明	企业信息系统异常界定	可能的情景和后果
无意产生	未按计划执行的动作[33]	基于时间:无退出时间	误操作或操作不规范,使业务未进行完就退出模块 可能后果:数据冗余、数据异常
基于规则	规则被错误地应用于熟悉的场景[33,35]	基于时间:非正常时间登录 基于地点:未在规定地点登录	(1)用户可能盗用账户信息,在非正常时间登录 (2)用户可能盗用账户信息,在非常规地点登录 可能后果:数据泄露或执行违规业务操作,如修改审批价格造成经济损失、违规审批采购单造成企业生产受阻
基于知识	思维模式错误或知识储备不足[33,35]	基于时间:超出合理操作时间	(1)用户登录后长时间不退出,占用系统资源 (2)用户登录后立刻退出,形成无效操作 可能后果:大量类似操作造成系统堵塞

①无退出时间异常

此类异常对应无退出时间的记录。当用户误操作、操作不规范等原因导致信息系统无响应或异常关闭时,用户正常的业务操作进程中断,则系统中的退出时间T_{out}丢失,此时的操作记录中$T_{out} \in \varnothing$。

② 非正常时间登录异常

该类异常对应非正常登录时间记录。根据企业实际情况,企业正常的工作时间为$CT_{in} - \Delta T \sim CT_{out} + \Delta T$。若操作记录中出现登录时间$T_{in} \notin (CT_{in} - \Delta T, CT_{out} + \Delta T)$,表明用户在不允许登录系统的时间段内非法登录系统,可能破坏信息系统的安全性,使企业机密信息被泄漏,严重时还导致其核心竞争力下降。因此,本研究将该行为对应的操作记录定义为非正常时间登录异常。

③ 超出合理操作时间异常

该类异常对应超时的记录。具体表现为登录时

长TL过长,此时$TL = T_{out} - T_{in}$,对于此类异常,本研究采用3σ原则(拉依达准则)标记异常行为。事先假设一组检测数据$\{x_1, x_2, \cdots, x_i\}$,$i \in N^*$,该组数据只含有随机误差数据,$i$为检测数据的个数,$N^*$为正自然数,对其进行计算处理,得到数据均值$\bar{x}$和标准偏差$\sigma$,以阈值$F$为条件确定一个区间,$F = \bar{x} \pm 3\sigma$,认为若$x_i \notin (\bar{x} - 3\sigma, \bar{x} + 3\sigma)$,则$x_i$就不属于随机误差数据,而是粗大误差数据,即明显偏离预期的数据。

(2)基于地点的用户行为异常界定

地点是指用户登录企业信息系统所处的位置,通常用IP地址表示主机所处的位置,用户通常在固定的地方使用信息系统完成企业业务。根据企业具体情况分为两种,一种是设定企业正常IP地址总集合IP_{set};另一种对用户行为控制要求高的企业,可以以用户为单位设置其能进行操作的正常IP地址集合$(UserID, IP_{set})$,相应地,可以界定超出正常IP地址集合登录的行为即为未在规定地点登录异常。

2.2 特征工程

特征工程主要实现从原始数据到可供算法直接使用的特征数据的转化,特征决定了机器学习的上限,而训练模型只是尽可能接近该上限,因此结合具体研究问题进行特征模型构建尤其重要。已有相关研究的特征模型主要分为用户属性特征和用户行为特征两类。由于企业信息系统承载了各企业的具体业务,应考虑将用户行为涉及的业务维度纳入特征模型。

基于上述分析,本研究提出构建用户属性特征、用户行为特征和业务特征3类特征,3类特征的选取思路如下。

(1)用户属性特征

已有研究发现,在信息系统操作过程中,性别是一个影响个体对信息系统认知和行为的重要因素[36-37],而年龄的差异使用户对系统的认知和处理方式等有所不同,从而影响用户使用信息系统的态度和行为[38-39]。此外,根据认知决策理论的阐述[40],用户在复杂环境中做出决策,受到个人经验和技术能力的影响,具体而言,工龄是用户在企业工作时间长短,主要体现用户的工作经验[41],职称级别主要用于区分用户的工作能力和技术水平[42]。因此,在用户基本属性方面,本研究选取性别、出生日期、进厂日期和职称级别4个特征。

(2)用户行为特征

用户的操作时间、时间间隔和操作技能成熟度等因素都对用户行为产生影响[43],本研究将这3个特征纳入用户行为特征的子集。

(3)业务特征

用户操作的功能与其业务职能直接挂钩,因此需要考虑用户操作功能的业务特性,具体包括用户操作系统的业务类型和业务成熟度。此外,在企业内部各部门之间,用户行政级别的高低体现了不同用户群体之间的业务责任差异,这类职责差异直接影响其行为模式[44]。因此,本研究选取业务层级、业务操作类型和业务成熟度3个特征。

特征模型的构建具体包含4个过程。①原始特征提取,就本框架而言,将从企业信息系统的用户基本信息数据、用户日志数据和业务数据中进行抽取;②目标特征构造;③特征数值化编码;④特征分级缩放。

2.3 模型训练

由于企业信息系统用户行为数据量大,模式复杂,为了能层层抽取用户异常行为的抽象特征,建立现有的特征模型到高层次语义特征之间的映射关系,结合对用户异常行为相关预测方法的分析,本框架选择深度神经网络(DNN)构建预测模型。目前深度神经网络被广泛用于图像处理、语音识别、搜索引擎等许多领域,它能够从大量数据中学习分类所需的高层和抽象的特征表示[45]。

DNN模型是一种前馈人工神经网络[45],根据节点在网络中的位置,可分为输入层、隐藏层和输出层。与浅层网络相比,DNN具有多个隐藏层,且每一层也可以有数量较多的神经单元,当前层的输出将作为下一层的输入。由此,可构造出层层叠加的网络结构,见图2。

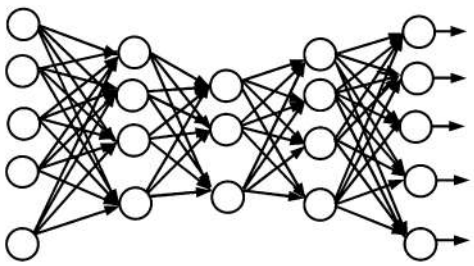


图2 DNN 结构概念图

Figure 2 DNN Structure Concept Diagram

DNN中各变量间都存在对应关系。假设存在$(N+1)$层的DNN,其中,输入层为第0层,隐藏层为第1层到第$(N-1)$层,输出层为第N层。存在$n \in (0, N]$,对任意的第n层,都有如下对应关系,即

$$net_k^n = \sum_{j=1}^{N^{n-1}} \omega_{j,k}^n \cdot z_k^{n-1} + b_k^l \tag{1}$$

$$z_k^n = f_l(net_k^n) \tag{2}$$

其中,net_k^n为第n层中第k个节点的输入值,N^{n-1}为第$(n-1)$层的节点个数,$\omega_{j,k}^n$为第n层中第j个节点与第n层的第k个节点间的权值,z_k^{n-1}为第$(n-1)$层中第k个节点的输出值,b_k^l为第n层中第k个节点的偏置,$f_l(\cdot)$为第n层的激活函数。

具体训练过程为:将原始特征输入深度神经网络的Sequential模型,进行多次的模拟训练,选择出最优的初始化参数设置,在最优参数的模型训练下获得最高层的特征表达,将其输入混淆矩阵分类模型[46]

表3 UserLog 表中原始数据的部分记录

Table 3 Partial Record of the Original Data in the UserLog Table

LoginName	LoginTime	LogoutTime	ModuleName
	……		
乔＊风	2013-12-26T13:26:04	2013-12-26T13:33:12	物料销售单
徐＊刚	2013-12-26T13:26:44	2013-12-26T15:24:03	领料申请单
缪＊中	2013-12-26T13:27:57	2013-12-26T13:28:16	出库单审核
	……		

表4 UserInfo 表中原始数据的部分记录

Table 4 Partial Record of the Original Data in the UserInfo Table

LoginName	Department	Gender	BirthDate	JoinDate	PRank	PositionRank
		……				
李＊松	机加工车间	男	1974-07-12	1992-07-01	经济师	科长
李＊	造船模块车间	男	1967-04-13	1989-11-01	技术员	安全监督员
李＊红	机电修理车间	女	1973-03-26	1993-07-01	助理工程师	科长
		……				

中进行模型的评估。

2.4 模型评估

为测量和验证本研究预测框架的有效性和准确性,本研究将进行两个层次的模型评估。第1层,考虑是否加入业务特征,比较模型的预测效果;第2层,与统计类经典方法(多元线性回归)和机器学习经典方法(支持向量机)进行比较,验证模型预测的准确性。

具体评估指标方面,采用召回率、查准率和*AUC*共3个常用指标,召回率和查准率反映预测方法针对信息系统异常行为的分类能力,*AUC*值主要用来评估二值分类器的好坏。

3 实验结果和分析

为验证预测框架的有效性,本研究选取A船舶制造企业为实验对象,因为:①该企业为行业内业绩领先的大型修造船企业,属于典型的大型单件小批制造企业,业务复杂,因而样本具有一定的代表性;②企业于2011年11月起开始正式启用ERP系统并应用至今,良好的应用基础为本研究提供了大量的实验数据,对该样本进行研究具有可行性;③企业在信息化应用过程中出现过多次由用户异常行为造成的损失,对用户异常行为的管理提出明确的需求,这为本研究提供了良好的案例环境。

3.1 数据准备

本研究选取A企业2011年10月至2017年9月共72个月的用户操作企业信息系统日志数据作为数据来源,采用覆盖用户范围较广的业务部门、职能部门和信息部门的信息系统作为研究对象。就本研究而言,需要用到系统中的日志信息表(UserLog)、用户信息表(UserInfo)和系统信息表(Module),基本数据情况见表3、表4和表5。字段含义分别为LoginName为用户名,LoginTime为登录时间,LogoutTime为退出时间,ModuleName为功能名,Department为所在部门,Gender为性别,BirthDate为出生日期,JoinDate为进厂日期,Prank为职称级别,PositionRank为行政级别,MoName为模块名,ModuleType为功能类型,SysName为系统名。

表5 Module 表中原始数据的部分记录

Table 5 Partial Record of the Original Data in the Module Table

ModuleName	MoName	SysName	ModuleType
	……		
CX9705 实际成本	成本分析	成本管理	DSS
保管入库单	库存管理	物资系统	TPS
保管入库单撤销	库存管理	物资系统	TPS
	……		

注:DSS为决策支持系统,TPS为事务处理系统。

经统计,UserLog表中源数据共1 611 288条,通过对表中空缺数据、噪音数据、不一致数据、重复数据以及不完整数据进行处理,共获取1 569 246条日志数据,结合UserInfo表得到研究样本,用户信息统计见表6。

表6 用户信息统计表

Table 6 User Information Statistics

变量	类别	人数	百分比/%
性别	男	402	76.86
	女	121	23.14
年龄	25岁及以下	55	10.52
	26岁~35岁	248	47.42
	36岁~45岁	135	25.81
	46岁及以上	85	16.25
职位	无职称	89	17.02
	助理职称	96	18.35
	初级职称	321	61.38
	中级职称	4	0.76
	高级职称	13	2.49
部门	业务部门	203	38.82
	职能部门	293	56.02
	信息部门	27	5.16

根据用户信息系统使用日志数据计算出登录时长(LoginTime-LogoutTime),单位为秒(s),统计结果见表7。

表7 登录时间数据描述

Table 7 Logintime Data Description

	最小值	最大值	均值	标准差
时长	0	55 421	1 467	23 415.98

3.2 用户异常行为数据

结合2.1,根据异常发生情况的不同,本研究将企业信息系统中可能的用户异常行为分为3类,即无退出时间异常、非正常时间登录异常和超出合理操作时间异常。该企业规定的工作时间范围为8:00-18:00,根据企业实际情况,设置偏离值为2小时,则正常登录时间范围为6:00-20:00,若操作记录中出现登录时间$T_{in} \notin (6{:}00, 20{:}00)$,则表明用户在不允许登录系统的时间段内异常登录系统,具体统计情况见表8。

表8 3类用户异常行为分类数据描述

Table 8 Three Types of User Abnormal Behavior Classification Data Description

	最小值	最大值	均值	标准差
无退出时间异常	4	3 022	1 164.99	611.30
非正常时间登录异常	0	301	37.39	52.84
超出合理操作时间异常	2	333	152.49	83.02

本研究以月份为单位对样本数据集进行初步统计分析,得到每月的用户异常操作次数A_x和操作总次数N_x,计算出异常操作发生率E_x,令$E_x = \frac{A_x}{N_x}$,x为月份。为分析各月异常操作发生率,绘制统计图,见图3。从图3可以看出,系统在投入使用后,异常率呈现先升后降的趋势,整体异常率较高,但在2016年10月异常率达到16.68%的高峰,通过对本月数据进行统计,发现异常行为数据主要集中在2016年10月8日至

图3 月异常率趋势

Figure 3 Monthly Abnormal Rate Trend

10月31日的工作日中,经调查得知A企业在国庆期间进行了全面的系统功能更新,此后的一个月内,用户可能处于对系统的适应期,异常行为普遍较多。

3.3 特征构建

本框架的特征模型包含4个过程。

(1)原始特征提取,从日志数据中提取所有9个原始特征。

在用户基本属性方面,选取用户名、性别、出生日期、进厂日期和职称级别5个原始特征,主要从用户信息表进行特征数据提取;在用户系统业务属性方面选取业务层级和操作功能类型两个原始特征,主要从系统信息表和日志信息表提取特征数据;在用户行为属性特征方面,选取登录时间和操作功能两个原始特征,主要从日志信息表提取特征数据。

(2)目标特征构建,通过特征提取得到9维特征子集后,进一步构建新特征。

在用户基本属性方面,对用户年龄和工龄进行目标特征构建,用户年龄 = 当前操作日期 - 出生日期,工龄 = 当前操作日期 - 进厂日期;在用户系统业务属性方面,构建业务成熟度特征,业务成熟度为功能投入使用至员工本次登录时间的间隔月数;在用户行为属性特征方面,增加技能成熟度、登录时间间隔和操作时间段3个目标特征,技能成熟度是指本次操作为止该用户操作的总次数,登录时间间隔为距上一次登录时间的间隔,操作时间段指用户登录的时间段。

与具体数据表的特征匹配情况见图4。

(3)特征数值化编码,见表9。

(4)特征分级缩放。由于用户原始特征对极端值不太敏感,故本研究使用分级缩放对操作技能成熟度和登录时间间隔两个特征进行数据的标准化,见表10。

3.4 DNN模型参数设置

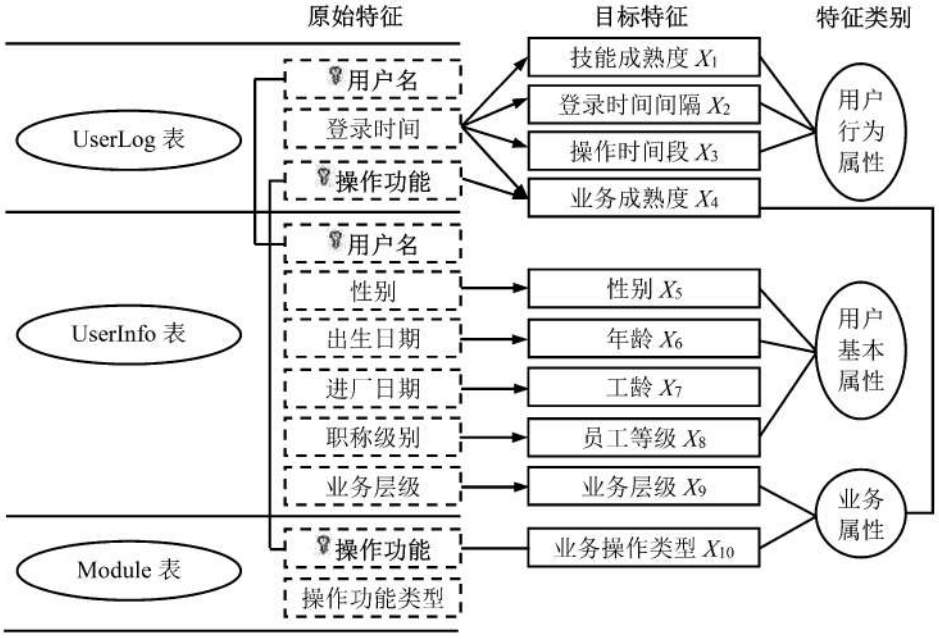
(1)在具体的DNN模型构建中,本研究设置适当的批尺寸($batch\_size = 128$)和训练轮次($epochs = 100$),使模型在内存不溢出的情况下达到最佳运算效率。为提高模型收敛效果,本研究采用可变的学习速率方案,令学习速率随着学习进展逐步减小。具体的动态学习率计算方法为

$$lrate = initial\_lrate \cdot \mathrm{math.pow}(drop, \mathrm{math.floor}\frac{1+epoch}{epochs\_drop}) \quad (3)$$

其中,$lrate$为学习率;$initial_lrate$为初始学习率,本研究模型中为0.10;$drop$为每个周期的衰减率,本研究模型中为0.50;$epoch$为当前训练轮次数量;$epochs_drop$为每个周期中包含的训练轮次数量,本研究模型中为4,即模型的学习率每经过4个训练轮次就会衰减50%。

(2)本研究在其他条件不变的情况下,改变模型中的隐藏层层数和层中神经元数量,采用业务信息系统的全特征子集进行训练和测试,得到的模型性能对比结果见图5和图6。其中,图例中每条线对应的数组表示输入层、隐藏层和输出层神经元的个数。例如,图5中紫色线对应的数组为[10,16,32,16,1],表示的神经网络配置为:包含10个神经元的输入层和1个神经元的输出层,隐藏层的数量为3个,3个隐藏层中包含的神经元个数分别为16、32、16。

综合分析图5和图6的结果可以发现,当网络配置为[10,64,128,256,128,64,1]时,模型性能处于相对最好、最稳定的状态。因此,本研究的DNN模型中采用该配置下的参数值。



注:🔑为主键标识符。

图4 企业信息系统用户异常行为特征匹配关系

Figure 4 Matching Relationship of User Abnormal Behavior Characteristics of Enterprise Information System

表 9 特征数值化编码

Table 9 Feature Numerical Coding

特征类别	变量	区间	编码
用户基本属性	性别	[1,2]	男性取值为 1,女性取值为 2
	年龄	[16,63]	无需编码
	工龄	[1,40]	无需编码
	职称等级	[1,5]	无职称为(1,0,0,0,0),助理职称为(0,1,0,0,0),初级职称为(0,0,1,0,0),中级职称为(0,0,0,1,0),高级职称为(0,0,0,0,1)
业务属性	业务层级	[1,2]	非管理层取值为 1,管理层取值为 2
	业务操作类型	DSS/TPS	DSS 取值为 1,TPS 取值为 2
	业务成熟度	[1,83]	无需编码
用户行为属性	技能成熟度	各部门取值区间存在差异,但连续	无需编码
	登录时间间隔	各部门取值区间存在差异,但连续	无需编码
	操作时间段	12:00 前后	6:00 – 12:00 取值为 1,12:01 – 20:00 取值为 2

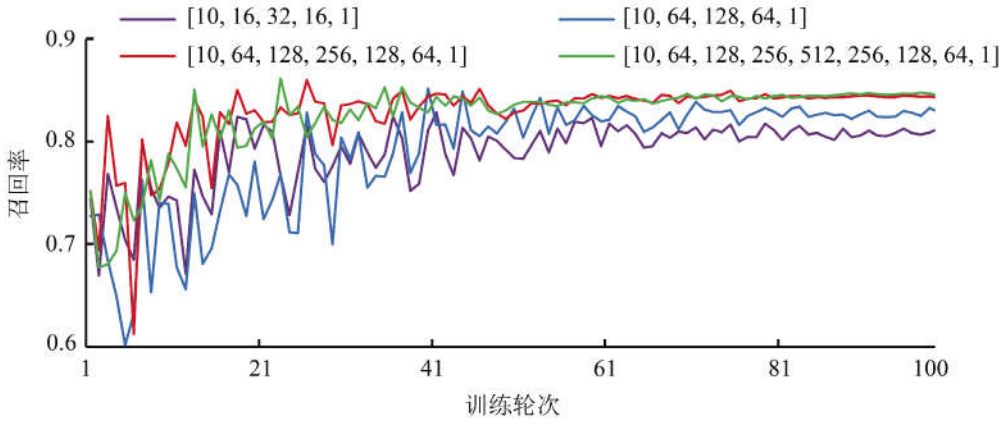


图 5 不同隐藏层层数与层中神经元数量配置下模型召回率对比

Figure 5 Comparison Diagram of Recall of Model Based on Different Number of Hidden Layers and Neurons in the Layer

表 10 技能成熟度和登录时间间隔分级映射对应关系

Table 10 Hierarchical Mapping Correspondence of Skill Maturity and Logintime Interval

部门	技能成熟度原区间	登录时间间隔原区间	统一映射区间
业务部门	[1,43 760]	[1,189 326 373]	
职能部门	[1,59 087]	[1,209 246 917]	[1,50]
信息部门	[1,26 119]	[1,185 419 565]	

(3)本研究在解决分类不平衡问题时采用设置惩罚系数的方法[47],通过降低负样本对模型的影响和提高正样本对模型的影响来保障正负样本对模型的影响大致相同,以提高模型训练的有效性。

(4)配置合理的激活函数[48],让模型具备非线性因素,提高模型的表达能力。此外,为避免过拟合现象[49],模型加入 Dropout 层,优化了网络层之间的连接结构。

3.5 模型分类效果评估

采用两个实验分别对本研究提出的预测框架中特征模型的有效性以及最终预测结果的有效性进行

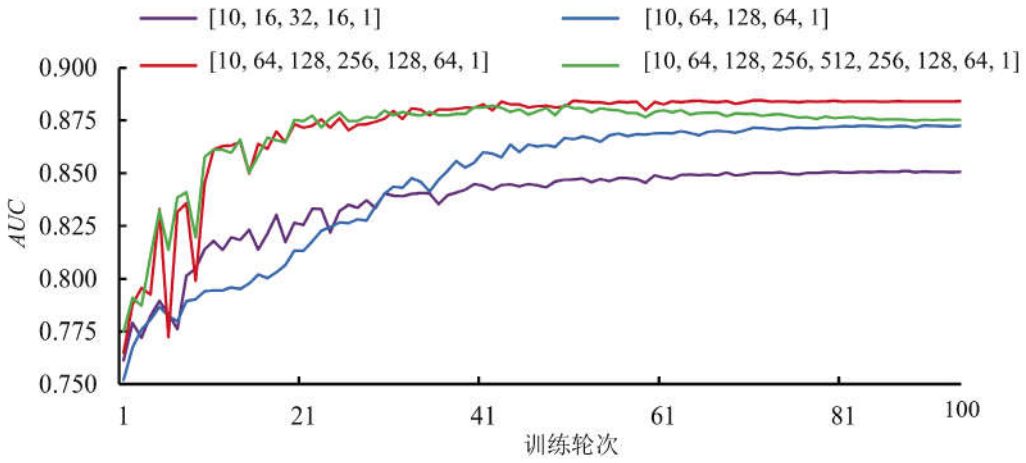


图 6 不同隐藏层层数与层中神经元数量配置下模型*AUC*对比

Figure 6 Comparison Diagram of *AUC* of Model Based on Different Number of Hidden Layers and Neurons in the Layer

分析。

3.5.1 实验1:特征模型对比

对比不考虑业务特征和加入业务特征的情况下,验证本预测框架的性能。具体步骤为:①选择所有用户的行为日志数据,基于已有研究,只采用包括用户基本属性和行为属性的7个经典特征进行预测;②加入代表业务特性的3个特征进行训练,比较不同特征数量下本研究方法的分类效果,验证本研究提出的特征模型有效性。

实验1的对比分析共输出100行模型性能数据,以训练轮次为横坐标,模型性能为纵坐标,绘制预测结果折线图,结果见图7。图7中的(a)、(b)、(c)分别给出召回率、查准率和*AUC*的变化趋势,可以看出,加入业务特性后的特征模型预测准确性有明显提高,召回率、查准率和*AUC*分别提高3.52%、2.16%和3.36,说明这些业务特征能够提高特征模型对用户异常行为的识别度。

这进一步说明就本案例而言,本预测模型的特征选取方式是合理的,符合企业信息系统的特点。但与传统预测方法相比是否具有优势,需要通过第2个实验进一步验证。

3.5.2 实验2:预测方法的对比

通过与MLR分类和SVM分类等其他预测方法的对比验证本研究模型的有效性。此外,考虑到A企业信息系统用户来自不同部门,不同部门用户产生的行为异常往往是不同的。因此,为了进一步考察预测框架的适用性,下面的预测实验也包含了针对不同部门的异常行为预测分析。

(1)MLR分类

使用Stata 15.0进行运算,结合MLR模型公式得到所有部门和3个部门的回归方程(4)式~(7)式;结合数据集获得MLR计算后的预估区间,结果见表11;得到 $\hat{Y}$(异常)映射到{0,1}后的用户异常行为的预测结果,见表12。

所有部门:

$$\hat{Y} = 0.1178X_1 + 0.0087X_2 - 0.8194X_3 + 0.1604X_4 + 0.4005X_5 + 0.1196X_6 - 0.2007X_7 - 0.3935X_8 + 0.6493X_9 + 0.1774X_{10} - 7.1149 \quad (4)$$

业务部门:

$$\hat{Y} = 0.1911X_1 + 0.0237X_2 - 0.0385X_3 - 0.0009X_4 + 0.0016X_5 + 0.0018X_6 + 0.0012X_7 + 0.0003X_8 - 0.0544X_9 - 0.0495X_{10} - 0.0157 \quad (5)$$

职能部门:

$$\hat{Y} = 0.4062X_1 - 0.0093X_2 - 0.0111X_3 + 0.0029X_4 - 0.0004X_5 + 0.001X_6 + 0.0013X_7 - 0.0012X_8 - 0.0447X_9 - 0.1096X_{10} - 0.0197 \quad (6)$$

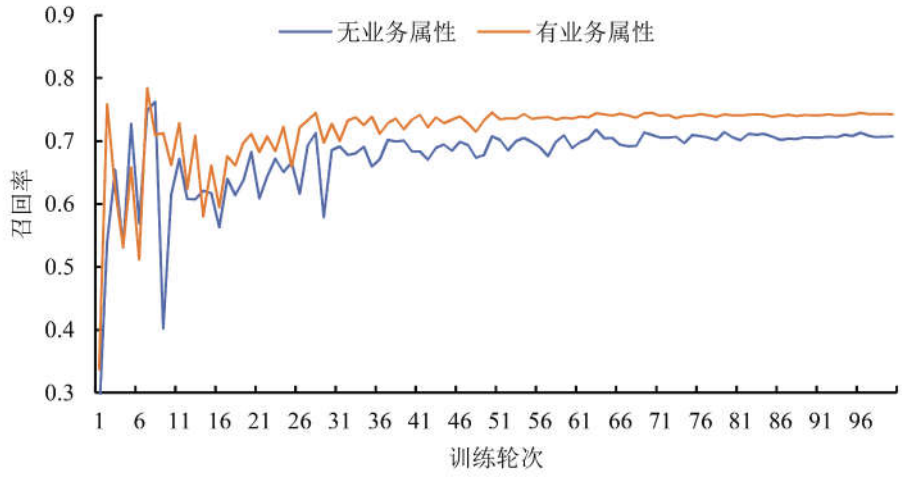
信息部门:

$$\hat{Y} = -0.0493X_1 - 0.0168X_2 - 0.0588X_3 - 0.0053X_4 + 0.0028X_5 + 0.0001X_6 + 0.0001X_7 + 0.0009X_8 - 0.0404X_9 - 0.0004X_{10} - 0.3809 \quad (7)$$

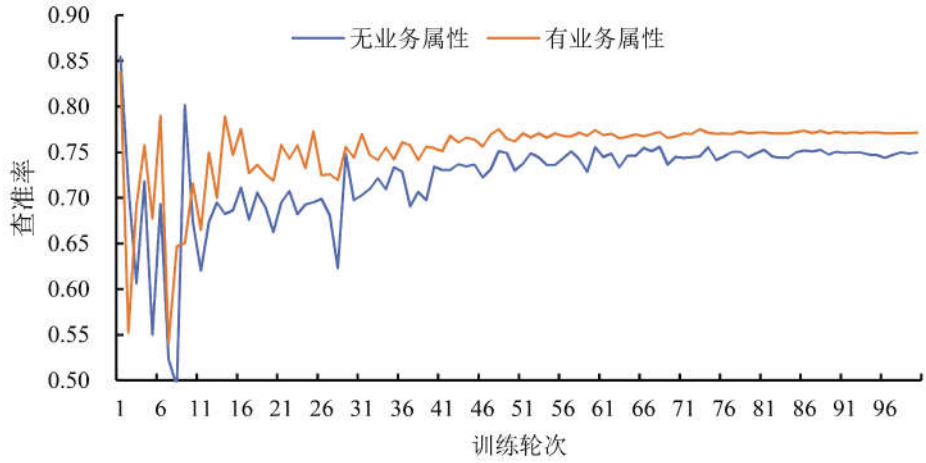
表 11 多重线性回归结果

Table 11 Results for Multiple Linear Regression

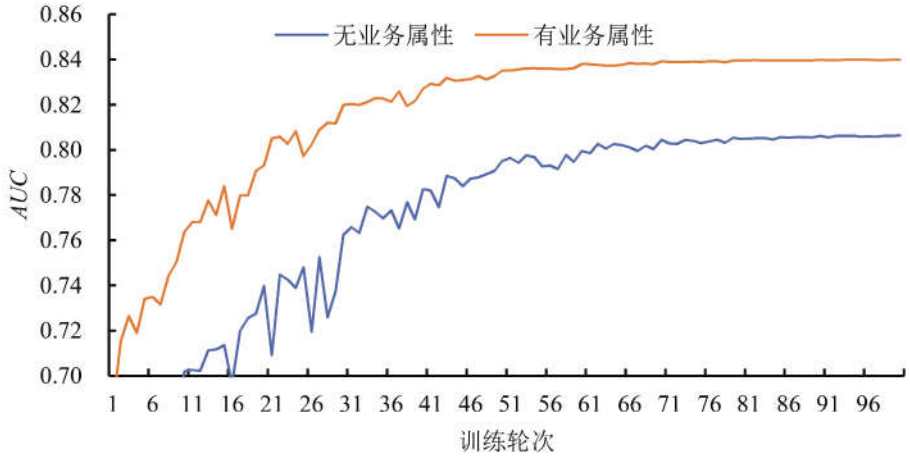
用户所在部门	$E(\hat{Y})$	F	$\hat{Y}\to 0$ 时的 $E(\hat{Y})$	$\hat{Y}\to 1$ 时的 $E(\hat{Y})$
所有部门	[-0.07,0.27]	0.10	[-0.07,0.10)	[0.10,0.27]
业务部门	[-0.10,0.45]	0.17	[-0.10,0.17]	[0.17,0.45]
职能部门	[-0.05,0.29]	0.12	[-0.05,0.12]	[0.12,0.29]
信息部门	[-0.03,0.12]	0.05	[-0.03,0.05]	[0.05,0.12]



(a)召回率对比



(b)查准率对比



(c) *AUC* 对比

图 7 不同特征数量下的比较结果

Figure 7 Comparison Results with Different Feature Quantities

表 12 MLR 模型的预测结果

Table 12 Prediction Results for the MLR Model

用户所在部门	召回率/%	查准率/%
所有部门	57.79	69.64
业务部门	27.89	88.94
职能部门	16.14	80.01
信息部门	15.43	88.36

根据表12,各部门的召回率都远低于50% 的自然分类结果,MLR的分类效果很差,表明用户行为模式呈现出线性不可分的状态。因此,不能采用MLR的方法进行用户异常行为的分类。

(2)SVM分类

MLR分类实验的研究结果表明,用户异常行为的预测问题是线性不可分的问题,因此属于非线性分类的SVM问题,需要引入内核扩展方法。本研究有10个输入维度,根据公式可知,如果映射到特征空间,会产生65个维度,故需要寻找合适的核函数,降低计算量,提高运算效率。测试发现,高斯核函数的性能相对较好,选该函数作为SVM模型的核函数,并对其γ值进行配置测试,发现$\gamma=20$时效果最佳;选择1 024作为批尺寸的大小,使模型在内存允许的情况下达到最大的运算速度;采用构建惩罚系数的计算方法,解决分类不平衡的问题。实验结果见表13。

表 13 SVM 模型的预测结果

Table 13 Prediction Results for the SVM Model

用户所在部门	召回率/%	查准率/%	*AUC*
所有部门	71.19	72.03	0.76
业务部门	74.86	63.03	0.82
职能部门	69.13	71.79	0.75
信息部门	62.29	68.54	0.70

根据表13,所有部门以及3个部门在SVM模型下的预测结果均高于自然分类的50%的阈值,明显优于MLR预测结果,但仍然没有达到理想状态,说明10个维度的特性仍然没有很好地抽取出来用于最后的训练。从分类效果看,SVM模型相当于单层神经网络的训练效果。因此,本研究的用户异常行为预测框架中采用深度神经网络模型是合理的,可以层层抽取各个特征的特性用于训练。

根据表11 ~ 表13,将3种预测模型或方法进行信息汇总,结果见表14。

统计方法中的MLR分类结果表明,3个部门的召回率都低于自然分类结果,表明用户异常行为呈现出线性不可分的数据状态,而采用非线性的SVM模型进行分类,分类效果得到显著提高。但由于非线性的SVM模型相当于单层的简单神经网络的特性,其抽取各个特征的特性的能力较弱,故在防止过度拟合的情况下,需要考虑增加模型的复杂度以提高模型的召回率。在最终采用的深度神经网络分类模型中,所有部门、业务部门和职能部门的用户异常行为预测的召回率分别为74.28%、77.40%和73.64%,查准率分别为77.12%、84.56%和74.68%,*AUC* 分别为0.84、0.88和0.82;但信息部门的召回率和查准率始终都低于70%,*AUC* 低于0.75,即该模型在信息部门的数据上表现较差。由此可以表明,DNN模型在用户异常行为分类问题的研究中,性能优于MLR和SVM预测模型。

此外,对3个部门的模型进一步比较可以发现,在与企业业务直接相关的业务部门和职能部门中,用户异常行为被较好地识别出来,而信息部门的分类效果不佳,这也恰好说明本研究选取的特征与企业的业务紧密相关,而信息部门用户的主要职责是辅助其他部门用户实施信息系统,其本身的操作不涉及企业的主要业务流程,因此用本研究的用户异常行为预测框架预测信息部门的用户异常行为效果欠佳。

综合以上实验结果可知,本研究提出的加入业务维度的特征模型能够有效提高模型的效果,与统计方法和机器学习方法相比,本研究模型表现得更好。

表 14 不同预测模型的预测结果对比信息汇总

Table 14 Comparison Information Summary for Prediction Results of Different Prediction Models

用户所在部门	召回率/%			查准率/%			*AUC*		
	DNN	MLR	SVM	DNN	MLR	SVM	DNN	MLR	SVM
所有部门	74.28	57.79	71.19	77.12	69.64	72.03	0.84		0.76
业务部门	77.40	27.89	74.86	84.56	88.94	63.03	0.88		0.82
职能部门	73.64	16.14	69.13	74.68	80.01	71.79	0.82		0.75
信息部门	64.15	15.43	62.29	69.28	88.36	68.54	0.74		0.70

4 结论

4.1 研究结果

针对企业信息系统用户异常行为的预测问题,为提高预测的准确性,本研究基于深度神经网络方法构建一种企业信息系统用户异常行为预测框架,并进行验证,得出研究结果如下。

(1)提出一套企业信息系统的非开放式用户异常行为预测框架,具体包括用户异常行为分类和界定、特征工程、模型训练和模型评估4个模块,并通过案例企业的实际数据初步验证了其有效性。

(2)加入业务特征后的新特征工程方案,在预测和分析企业信息系统异常行为方面有更好的表现,召回率、查准率和*AUC*分别提高3.52%、2.16%和3.36。

(3)通过与统计方法的MLR和机器学习的SVM比较,预测效果均有相应提升,与MLR相比,召回率和查准率分别提高16.49%和7.48%;与SVM相比,召回率、查准率和*AUC*分别提高3.09%、5.09%和0.08。

4.2 理论意义和实践意义

本研究的理论意义在于:①与已有研究主要考虑网络层、数据层和表达层的用户异常行为不同,本研究重点聚焦应用层的企业信息系统,提出基于深度神经网络的用户异常行为预测框架,补充和丰富了用户异常行为的研究成果,并通过一个典型企业的实验分析初步验证了该模型的有效性。②验证了深度神经网络方法对应用层用户异常行为预测研究的作用,一定程度上为深度学习在应用层用户异常行为的预测研究方面增加了新的证据。已有关于应用层用户异常行为的研究大部分集中在社交网络领域,对企业信息系统缺乏关注,且主要采用机器学习方法,如贝叶斯[30]和SVM[24,26]等,这些方法属于单层的特征学习,对复杂的行为模式缺乏多层次的特征抽取。本研究提出一个更为集成的特征方案,即在企业信息系统情景下考虑加入业务特征维度,并采用深度神经网络抽取并建立高层次语义特征的映射,丰富了用户异常行为的理论研究,也为后续其他复杂行为模式的用户异常行为预测研究提供了新的研究思路。

本研究的实践意义在于:①本研究使用实际的企业信息系统用户行为数据进行实验,提出的方法预测性能较好,可以将该方法推广到企业,辅助企业进行更有针对性的预防和管理决策,减少用户异常行为可能带来的损失;②除关注影响异常的用户特征和行为特征,业务特性也是影响异常的一个关键因素,企业需要充分关注用户操作所对应的属性,如业务类型和业务层级,这些可能是企业制定差异化安全策略的重要依据。

4.3 局限性和未来研究方向

本研究仍然存在一些局限性,需要在未来研究中进行扩展。①不同的用户异常行为可能带来不同的后果,本研究目前是将3类用户异常行为都视为一类进行研究,未来研究可细化分析不同用户异常行为的预测模型,使研究更具有针对性;②由于实验企业的实际情况,本预测框架只初步验证3类基于时间的用户异常,在未来研究中应逐步加入基于地点的用户异常情况,如结合企业的业务特征和政策环境,进一步收集其他类型企业的信息系统数据集进行分析;③本研究的相关结果主要通过构建预测框架和实验分析得到,在后续的研究中应尝试从理论层面寻找企业信息系统用户异常行为的影响机制,进而对用户异常行为的管理提出更有针对性的管控策略。

参考文献:

[1] ZHANG Y, MERATNIA N, HAVINGA P. Outlier detection techniques for wireless sensor networks: a survey. ***IEEE Communications Surveys & Tutorials***, 2010, 12(2): 159-170.

[2] BENLIAN A. IT feature use over time and its impact on individual task performance. ***Journal of the Association for Information Systems***, 2015, 16(3): 144-173.

[3] 王念新,李清香,倪丹,等. 信息系统使用对企业员工绩效影响的实证研究. **管理评论**, 2017, 29(6): 141-151.
WANG Nianxin, LI Qingxiang, NI Dan, et al. Effects of IT system usage on job performance: an empirical study. ***Management Review***, 2017, 29(6): 141-151.

[4] 袁粉侠,朱柏雅. 我国中小商业银行操作风险与对策探讨. **价值工程**, 2018, 37(32): 13-15.
YUAN Fenxia, ZHU Boya. Discussion on operational risks and countermeasures of small and medium-sized commercial banks in China. ***Value Engineering***, 2018, 37(32): 13-15.

[5] 券商中国. **韩国三星证券送乌龙大礼,竟向员工误发6 700亿**. (2018-05-28)[2019-05-18]. http://finance.sina.com.cn/stock/s/2018-05-28/doc-ihcffhsu4828535.shtml.
QUANSHANG ZHONGGUO. ***South Korea's Samsung Securities sent an oolong gift and actually sent 6 700 to employees***. (2018-05-28)[2019-05-18]. http://finance.sina.com.cn/stock/s/2018-05-28/doc-ihcffhsu4828535.shtml.

[6] LIM C, SINGH N, YAJNIK S. A log mining approach to failure analysis of enterprise telephony systems // KOOPMAN P, MADEIRA H. ***International Conference on Dependable Systems & Networks***. Alaska: Anchorage, 2008: 398-403.

[7] VENKATASUBRAMANIAN V, RENGASWAMY R, YIN K, et al. A review of process fault detection and diagnosis: Part I: quantitative model-based methods. ***Computers & Chemical Engineering***, 2003, 27(3): 293-311.

[8] KRAMER M A, PALOWITCH B L, Jr. A rule-based approach to fault diagnosis using the signed directed graph. ***AICHE Journal***, 1987, 33(7): 1067-1078.

[9] QIN S J. Survey on data-driven industrial process monitoring and diagnosis. ***Annual Reviews in Control***, 2012, 36(2): 220-234.

[10] YIN T Z X, WULFF S S, PIERRE J W, et al. A case study on the use of data mining for detecting and classifying abnormal power system modal behaviors. ***Quality Engineering***,

2019,31(2):314-333.

[11] BAO L, LI Q, LU P Y, et al. Execution anomaly detection in large-scale systems through console log analysis. ***Journal of Systems and Software***, 2018, 143: 172-186.

[12] ANDERSON J P. ***Computer security threat monitoring and surveillance***. Washington, Pennsylvania: James P Anderson Company, 1980.

[13] DENNING D E. An intrusion-detection model. ***IEEE Transactions on Software Engineering***, 1987, 13(2): 222-232.

[14] 杨宏宇,李博超. 基于逆向习得推理的网络异常行为检测模型. **计算机应用**, 2019, 39(7): 1967-1972.
YANG Hongyu, LI Bochao. Network abnormal behavior detection model based on adversarially′ learned inference. ***Journal of Computer Applications***, 2019, 39(7): 1967-1972.

[15] BOUKHTOUTA A, MOKHOV S A, LAKHDARI N E, et al. Network malware classification comparison using DPI and flow packet headers. ***Journal of Computer Virology and Hacking Techniques***, 2016, 12(2): 69-100.

[16] ALJAWARNEH S, ALDWAIRI M, YASSEIN M B. Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. ***Journal of Computational Science***, 2018, 25: 152-160.

[17] NAVARRO J, DERUYVER A, PARREND P. A systematic survey on multi-step attack detection. ***Computers & Security***, 2018, 76: 214-249.

[18] GEORGE A. Anomaly detection based on machine learning dimensionality reduction using PCA and classification using SVM. ***International Journal of Computer Applications***, 2012, 47(21): 5-8.

[19] 李海斌,李琦,汤汝鸣,等. 一种无监督的数据库用户行为异常检测方法. **小型微型计算机系统**, 2018, 39(11): 2464-2472.
LI Haibin, LI Qi, TANG Ruming, et al. User behavior anomaly detection for database based on unsupervised learning. ***Journal of Chinese Computer Systems***, 2018, 39(11): 2464-2472.

[20] SALLAM A, BERTINO E, HUSSAIN S R, et al. DBSAFE: an anomaly detection system to protect databases from exfiltration attempts. ***IEEE Systems Journal***, 2017, 11(2): 483-493.

[21] KAMRA A, TERZI E, BERTINO E. Detecting anomalous access patterns in relational database. ***The VLDB Journal***, 2008, 17(5): 1063-1077.

[22] ZHENG N, PALOSKI A, WANG H N. An efficient user verification system using angle-based mouse movement biometrics. ***ACM Transactions on Information and System Security***, 2016, 18(3): 1-27.

[23] 许洪军,张洪,贺维. 一种基于鼠标行为的云用户异常检测方法. **哈尔滨理工大学学报**, 2019, 24(4): 127-132.
XU Hongjun, ZHANG Hong, HE Wei. A cloud user anomaly detection method based on mouse behavior. ***Journal of Harbin University of Science and Technology***, 2019, 24(4): 127-132.

[24] 谈磊,连一峰,陈恺. 基于复合分类模型的社交网络恶意用户识别方法. **计算机应用与软件**, 2012, 29(12): 1-5, 17.
TAN Lei, LIAN Yifeng, CHEN Kai. Malicious users identification in social network based on composite classification model. ***Computer Applications and Software***, 2012, 29(12): 1-5, 17.

[25] 岳虹,张智,杨科. 基于磷虾群免疫神经网络的微博僵尸粉检测. **计算机应用与软件**, 2015, 32(12): 145-149.
YUE Hong, ZHANG Zhi, YANG Ke. Detecing microblogging zombie fans based on krill hred immune neural network. ***Computer Applications and Software***, 2015, 32(12): 145-149.

[26] ZHENG X H, ZENG Z P, CHEN Z Y, et al. Detecting spammeon on social networks. ***Neurocomputing***, 2015, 159: 27-34.

[27] SARPIRI M N, GANDOMANI T J, TEYMOURZADEH M, et al. A hybrid method for spammer detection in social networks by analyzing graph and user behavior. ***Journal of Computers***, 2018, 13(7): 823-829.

[28] FIRE M, KATZ G, ELOVOCI Y. Strangers intrusion detection detecting spammers and fake profiles in social networks based on topology anomalies. ***Human Journal***, 2012, 1(1): 26-39.

[29] ZHANG Y, LU J G. Discover millions of fake followers in Weibo. ***Social Network Analysis and Mining***, 2016, 6(16): 1-15.

[30] 张艳梅,黄莹莹,甘世杰,等. 基于贝叶斯模型的微博网络水军识别算法研究. **通信学报**, 2017, 38(1): 44-53.
ZHANG Yanmei, HUANG Yingying, GAN Shijie, et al. Weibo spammers′ identification algorithm based on Bayesian model. ***Journal on Communications***, 2017, 38(1): 44-53.

[31] 孟续铎,王欣. 企业员工超时工作成因与劳动时间特征. **经济与管理研究**, 2015, 36(12): 66-74.
MENG Xuduo, WANG Xin. Causes of enterprise staff working overtime and characteristics of labor time. ***Research on Economics and Management***, 2015, 36(12): 66-74.

[32] 李亚玲. 基于员工生产率非对称信息模型的超时工作原因探析. **企业经济**, 2011(10): 88-92.
LI Yaling. Analysis of the causes of overtime work based on the asymmetric information model of employee productivity. ***Enterprise Economy***, 2011(10): 88-92.

[33] ZHAO B, OLIVERA F. Error reporting in organizations. ***Academy of Management Review***, 2006, 31(4): 1012-1030.

[34] 马力,焦李成,董富强. 一种Internet的网络用户行为分析方法的研究. **微电子学与计算机**, 2005, 22(7): 124-126.
MA Li, JIAO Licheng, DONG Fuqiang. Research on a kind of categorised method and model of users′ behaviors based on Internet. ***Microelectronics & Computer***, 2005, 22(7): 124-126.

[35] REASON J. Generic error-modelling system (GEMS): a cognitive framework for locating common human error forms // RASMUSSEN J, DUNCAN K, LEPLAT J. ***New Technology and Human Error***. London: Wiley & Sons, 1987: 86.

[36] MERLUZZI J. Gender and negative network ties: exploring difficult work relationships within and across gender. ***Organization Science***, 2017, 28(4): 636-652.

[37] 沈校亮,厉洋军. 智能健康硬件用户间歇性中止行为

影响因素研究. **管理科学**, 2017, 30(1):31-42.
SHEN Xiaoliang, LI Yangjun. An empirical investigation of factors affecting smart health device users' intermittent discontinuance. ***Journal of Management Science***, 2017, 30(1):31-42.

[38] 周燕, 郭偲偲, 张麒麟. 内外双向因素与搭便车行为: 社会网络的调节作用. **管理科学**, 2015, 28(3):130-142.
ZHOU Yan, GUO Caicai, ZHANG Qilin. Free rider behavior and internal-external bidirectional factors: moderating effects of social network. ***Journal of Management Science***, 2015, 28(3):130-142.

[39] 卫旭华, 刘咏梅, 车小玲. 中国上市企业高管离职影响因素的跨层研究. **管理科学**, 2013, 26(6):71-82.
WEI Xuhua, LIU Yongmei, CHE Xiaoling. Multilevel study of the influencing factors of top managers' turnover in Chinese listed companies. ***Journal of Management Science***, 2013, 26(6):71-82.

[40] KOLSKI C, STRUGEON E L. A review of intelligent human-machine interfaces in the light of the ARCH model. ***International Journal of Human Computer Interaction***, 1998, 10(3):193-231.

[41] 孙华, 丁荣贵, 王楠楠. 研发团队共享领导力行为的产生和对创新绩效的作用: 基于垂直领导力的影响. **管理科学**, 2018, 31(3):17-28.
SUN Hua, DING Ronggui, WANG Nannan. Emergence of shared leadership behaviors and effect on innovation performance in R&D team: based on the influence of vertical leadership. ***Journal of Management Science***, 2018, 31(3):17-28.

[42] ODARCHENKO R, TKALICH O, KONAKHOVYCH G, et al. Evaluation of SDN network scalability with different management level structure // POPOVSKIY V V. **2016** ***Third International Scientific-Practical Conference Problems of Infocommunications Science and Technology (PIC S&T)***, 2016:128-131.

[43] JASPERSON J S, CARTER P E, ZMUD R W. A comprehensive conceptualization of post-adoptive behaviors associated with information technology enabled work systems. ***MIS Quarterly***, 2005, 29(3):525-557.

[44] 李怡文, 刘杰. 管理信息系统开发中的用户行为及系统开发策略. **计算机工程**, 2005, 31(16):61-63.
LI Yiwen, LIU Jie. User behavior and developing strategy in MIS exploitation. ***Computer Engineering***, 2005, 31(16):61-63.

[45] LECUN Y, BENGIO Y, HINTON G. Deep learning. ***Nature***, 2015, 521:436-444.

[46] OHSAKI M, WANG P, MATSUDA K, et al. Confusion-matrix-based kernel logistic regression for imbalanced data classification. ***IEEE Transactions on Knowledge and Data Engineering***, 2017, 29(9):1806-1819.

[47] BARANDELA R, SÁNCHEZ J S, GARCÍA A V, et al. Strategies for learning in class imbalance problems. ***Pattern Recognition***, 2003, 36(3):849-851.

[48] SCHMIDT-HIEBER J. ***Nonparametric regression using deep neural networks with ReLU activation function***. Enshede, Netherlands: University of Twente, 2017.

[49] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting. ***The Journal of Machine Learning Research***, 2014, 15(1):1929-1958.

Research on User Abnormal Behavior Prediction of Enterprise Information System Based on Deep Neural Networks

YIN Jun[1,2], PENG Yanhong[2], LU Yi[3], GE Shilun[2], LIU Peng[2]

1 Key Research Base of Philosophy and Social Sciences of Colleges of Jiangsu Provinces, Jiangsu University of Science and Technology, Zhenjiang 212003, China

2 School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang 212003, China

3 Software Development Center, Industrial and Commercial Bank of China, Shanghai 200120, China

Abstract: With the continuous improvement of enterprise informatization, its core business is increasingly dependent on the reliable operation of information system. For enterprises, any abnormal operation performed by information system users may bring inestimable losses to them. The enterprise pays more and more attention to the negative impact that the user's abnormal behavior may have on the enterprise. How to effectively predict the abnormal behavior of enterprise information system is our research question.

To address this need, the study designs a prediction framework for the abnormal behavior of enterprise information system users: firstly, the definition standard of abnormal behavior of enterprise information system users is defined; then, with the user log data, we add the business dimension to build the feature model based on previous research, and use the deep neural network method to predict the abnormal behavior of users; finally, the model is evaluated by comparing with classic statistical method and traditional machine learning method. Taking a shipbuilding enterprise as an example, the effectiveness of our prediction frame-

work is preliminarily verified.

The research results show that the overall performance of the feature model becomes better after adding business features. In addition, the data-driven deep neural network model can extract abstract features of users' abnormal behaviors layer by layer, improving the efficiency of each feature in predicting abnormal behavior. Compared with multiple linear regression, the recall rate and precision rate of deep neural networks increased by 16.49% and 7.46% respectively; compared with support vector machine, the recall rate, precision rate and *AUC* increased by 3.09%, 5.09% and 0.08, respectively. A further comparison of the models of the three departments found that in the business departments and functional departments directly related to the business of the enterprise, the abnormal behavior of users can be better identified, while the classification effect of the information department is not good.

The research results provide a prediction framework that may be applicable to the abnormal behavior of enterprise information system users. The feature model is built by integrating the extant research and business-oriented features of enterprise information system, and the deep neural network method is selected for model training, which improves the accuracy of prediction. This also helps enterprises to predict the abnormal behavior of users, so that timely measures can be taken to reduce the negative impact of abnormal user behavior on the enterprise.

Keywords: enterprise information system; deep neural networks; user abnormal behavior; feature engineering; prediction

Received Date: July 4th, 2019　　**Accepted Date**: November 28th, 2019

Funded Project: Supported by the National Natural Science Foundation of China(71331003,71972090,71871108) and the Jiangsu Graduate Research Innovation Program(KYCX_19-1650)

Biography: YIN Jun, doctor in management, is an associate professor in the Key Research Base of Philosophy and Social Sciences of Colleges of Jiangsu Provinces and the School of Economics and Management at Jiangsu University of Science and Technology. Her research interests cover information system complexity, information system use and cloud computing. Her representative paper titled "Study of effects of functional tasks network's position and relationship on enterprise information system usage" was published in the ***Systems Engineering – Theory & Practice*** (Issue 2, 2018). E-mail: bamhill@163.com

PENG Yanhong is a master degree candidate in the School of Economics and Management at Jiangsu University of Science and Technology. Her research interest focuses on information system use. Her representative paper titled "Evolution analysis of newcomer-task network structure of enterprise information system: a case study of a shipbuilding enterprise" was published in the ***International Society for Knowledge and Systems Sciences*** (ISBN 978-981-15-1208-7). E-mail: 1206394813@qq.com

LU Yi is an assistant manager in the Software Development Center at Industrial and Commercial Bank of China. Her research interests include information system operation and maintenance, and machine learning. Her representative paper titled "Analysis of dynamic complexity feature of information system data based on visualization" was published in the **2018 1*st International Conference on Information Management and Management Science*** (ISBN 978-1-4503-6486-7). E-mail: deerlet1993@163.com

GE Shilun, doctor in management, is a professor in the School of Economics and Management at Jiangsu University of Science and Technology. His research interest focuses on management information system. He is the principal investigator for the research project titled "Research on management information system reengineering based on cloud", supported by the National Natural Science Foundation of China(71331003). E-mail: jzgsl@jzerp.com

LIU Peng, doctor in management, is a lecturer in the School of Economics and Management at Jiangsu University of Science and Technology. His research interest focuses on evolution analysis of complex social networks. His representative paper titled "Structure and evolution of co-authorship network in an interdisciplinary research field" was published in the ***Scientometrics*** (Issue 1, 2015). E-mail: liupeng19821017@126.com □