现代计算机

XIANDAI JISUANJI

第29卷第14期（总第782期）

半月刊（1984年创刊）

2023年7月25日出版

主管单位 中山大学

主办单位 广州中山大学出版社有限公司

出版单位 广东现代计算机杂志社有限公司

发 行 广东省报刊发行局（全国公开发行）

印 刷 广州一龙印刷有限公司

社 长 黄少伟

主 编 石玉珍

编 委 邹岚萍 熊锡源 李 文 石玉珍 梁嘉璐

地 址 广州市海珠区新港西路135号
中山大学内（510275）

电 话 020-84112089（编辑部）

网 址 www.moderncomputer.cn

电子邮箱 tougao@moderncomputer.cn

ISSN 1007-1423
CN 44-1415/TP

邮发代码：46-121

定价：30.00元

邮局订刊二维码

现代计算机
官方网站二维码

ISSN 1007-1423
14
9 771007 142239

现代计算机 第29卷 第14期（总第782期） 2023年7月

2023年7月 第29卷

第14期 （总第782期）

ISSN 1007-1423
CN 44-1415/TP

现代计算机

MODERN COMPUTER

中山大学出版社 主办

中国期刊数据库CNKI全文收录期刊
中国学术期刊（光盘版）收录期刊
中文科技期刊数据库全文收录期刊
中国核心期刊（遴选）数据库收录期刊
中国学术期刊综合评价数据库收录期刊

- 研究与开发：计算机发展和软、硬件开发的理论研究
- 图形图像：重点为与图形图像相关的理论及实践研究
- 开发案例：基于某方面的计算机开发案例研究与分析
- 实践与经验：计算机应用的实例及心得

现代计算机
XIANDAI JISUANJI
第29卷 第14期（总第782期）2023年7月25日出版

目 次

研究与开发

图形图像

实践与经验

开发案例

期刊基本参数：CN44-1415/TP*1984*S*16*120*zh*P*¥30.00*1200*24*2023-07

Modern Computer
(Vol. 29, No. 14; Jul. 25, 2023)

CONTENTS

Research and Development

Graphic and Image

Practice and Experience

Development Solution

研究与开发

文章编号: 1007-1423(2023)14-0001-08 DOI: 10.3969/j.issn.1007-1423.2023.14.001

基于改进YOLOv5和DeepSort的交通流参数检测方法

单振宇[1*]，张 琳[1]，侯晓雯[2]

（1. 长沙理工大学交通运输工程学院，长沙 410114；2. 中国海洋大学信息与工程学部，青岛 266100）

摘要：为缓解基于视频的交通流参数检测效果差的问题，提出了基于改进YOLOv5s和DeepSort的实时交通流参数检测方法。首先，在YOLOv5s中引入MobilenetV3和SIoU Loss来轻量化检测模型，提高检测速度。其次，重构DeepSort外观特征提取网络，并基于车辆重识别数据集重训练，提高跟踪准确率。最后，基于虚拟检测线圈的方式检测双向交通流量和车速。实验结果表明：方法整体运行速度可达59帧/s，在平峰、高峰时段的检测准确率均在92%以上，在实际交通流参数检测任务中具有较强的鲁棒性。

关键词：交通流参数检测；YOLOv5s；DeepSort；实时

0 引言

交通流量、速度等交通流参数的检测是交通调查外业工作中的主要任务之一，采集的交通流信息可用于灵活的交通管理与交通安全评价等任务。目前交通调查中常用的交通流参数检测方法仍然以人工现场观测为主，但人工方法存在效率低、误检率高、成本大等问题；此外，高峰时期车流量较大，现场观测也存在一定的安全隐患。因此本研究基于外业采集的交通视频，利用深度学习技术在内业完成流量与速度的自动检测任务。

基于视频的交通流参数检测方法可分为传统机器学习方法和深度学习方法。对于传统机器学习方法，El Bouziady等[1]通过背景差分法检测车辆，由合加速稳健特征算法(speeded up robust features，SURF)对车辆显著外型信息特征进行匹配并生成稀疏深度图，经过像素坐标转换后由帧差法计算车速。Tourani等[2]把通过混合高斯背景差分法检测到的车辆使用形态学变换与过滤算法进行匹配，结合斑点跟踪算法跟踪车辆，由帧间行驶距离估计车速。蒋建国等[3]借助视觉前景提取算法(visual background extractor, ViBe)分离背景，随后基于像素以及视频帧相结合的算法更新背景，最后由虚拟线圈法统计道路交通流量。张冬梅等[4]针对无人机(unmanned aerial vehicle, UAV)视频，利用改进的多帧平均方法获取初始背景，由混合高斯背景算法更新背景，通过背景差分法检测车辆目标，最后由虚拟线圈法获取实时交通流量。上述传统机器学习方法检测过程相对简单，处理速度较快，但是对检测环境要求较高，检测精度不稳定。

对于深度学习方法，Arinaldi等[5]使用Faster R-CNN检测模型检测交通视频中的车辆目标，基于核相关滤波算法(kernel correlation filter，KCF)进行目标跟踪，根据被跟踪车辆行驶距离和时间检测车速。Ke等[6]利用组合Haar级联和卷积神经网络(convolutional neural networks，CNN)来检测UAV视频中的车辆目标，并通过KLT(Kanade Lucas Tomasi)跟踪算法跟踪车辆，进而获取道路上车辆速度以及交通流量等信息。

收稿日期：2023-03-24 **修稿日期：**2023-04-24

基金项目：长沙理工大学研究生科研创新项目(CX2021SS113)

作者简介：*通信作者:单振宇(1998—),男,山东滕州人,硕士,研究方向为交通图像处理与识别,E-mail:20101030026@stu.csust.edu.cn；张琳(1996—),女,重庆开州人,硕士,研究方向为智能交通；侯晓雯(1998—),女,山东枣庄人,硕士,研究方向为计算机应用技术

Liu等[7]使用迁移学习后的深度神经网络(deep neural network, DNN)检测道路中的车辆，然后基于卡尔曼滤波算法进行跟踪，根据跟踪结果估计了车流量与车速。Li等[8]提出一种基于Faster R-CNN和域自适应(DA)的车辆检测方法，通过空间金字塔Lucas Kanade光流法计算车辆在两帧之间的行驶距离，并结合检测道路区域面积和车辆数获取流量、密度、速度信息。Li等[9]在YOLOv3目标检测模型的基础上引入了金字塔网络结构，改进了小尺寸和遮挡场景下的车辆检测效果，随后使用多目标跟踪算法(confidence multi-object tracking，CMOT)跟踪车辆，构建了交通流三参数的检测框架。赖见辉等[10]使用YOLOv3目标检测算法检测路侧采集的交通视频，建立了透视投影变换加卡尔曼滤波预测加匈牙利分配的交通流量计数框架。

深度学习技术的发展有效降低了外界环境对目标特征信息提取的影响，提高了车辆检测与跟踪效果。然而，多数深度学习的方法采用的神经网络结构复杂，模型参数量与计算量较大，不适合小算力的移动端设备使用，且难以在实际应用中实时运行。意识到了这一点，刘磊等[11]提出YOLO检测和Meanshift快速跟踪的车流量检测方法。文奴等[12]通过使用可变性卷积和MobileNetV3轻量化的YOLOv4目标检测模型接合DeepSort算法实现了多车道的交通流量检测。上述工作的改进虽有一定成效，但准确率和实时性仍有一定的提升空间。

因此，本文提出一种基于改进YOLOv5s和DeepSort的端到端的实时交通流参数检测方法。为提高车辆检测模型运行速度、降低对使用设备的算力要求，引入MobileNetV3网络作为YOLOv5s的骨干网络；为提高车辆检测时的定位精度，使用SIoU Loss作为车辆边界框的损失函数。为提高车辆跟踪算法的跟踪精度与速度，轻量化原算法的外观特征提取网络并在车辆重识别数据集上重新训练。最后，融合改进的YOLOv5s目标检测模型和DeepSort多目标跟踪算法实施交通流参数检测任务。

1 交通流参数检测模型

本文提出的实时交通流参数检测方法主要分为车辆检测、车辆跟踪、交通流参数检测三部分，整体流程如图1所示。首先，使用改进的YOLOv5s模型提取视频中的车辆目标的位置与类型信息；其次，使用改进的DeepSort多目标跟踪算法结合车辆外观信息与运动信息跟踪车辆目标；最后，设置虚拟检测线并构建检测区域，通过检测区域中车辆的行驶方向和时间计算流量与速度信息。

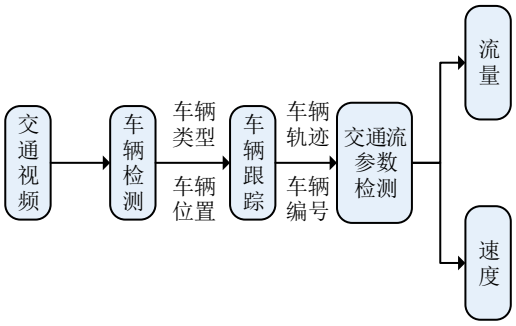


图1 交通流参数检测流程

1.1 车辆检测

1.1.1 YOLOv5s

YOLOv5是YOLO(you only look once)系列检测模型的最新版本，较前代版本(YOLOv3、YOLOv4等)在速度、精度上均有较大提升。YOLOv5按照模型大小，从小到大可分为YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x。鉴于实时性的考虑，本文选择YOLOv5s作为基础检测模型。

YOLOv5s模型的网络结构主要包括输入端、Backbone骨干网络、Neck融合网络和输出端四部分。其中，输入端对输入图像进行Mosaic数据增强、自适应锚框计算以及自适应图片缩放等数据预处理操作，丰富数据特征。Backbone骨干网络提取不同层次的图像特征，是占比整个模型参数量最大的部分，原模型使用Focus模块、CSP瓶颈层结构以及空间金字塔池化模块SPPF。Neck融合网络的设计采用了路径聚合结构PAN加特征金字塔结构FPN的结构，加强特征信息的融合能力，进一步扩大感受野，有效保留图像的上下文信息。输出端输出检测结果，包括目标位置和类别信息。

1.1.2　YOLOv5s改进

为进一步降低模型的参数量和计算量，使其更适合部署到移动端设备，本文使用MobileNetV3-Small[13]作为Backbone骨干网络。MobileNetV3通过使用逆向残差结构(inverted residual structure)、线性瓶颈模块(linear bottleneck)、深度可分离卷积(depthwise separable convolutions)降低模型参数大小与计算量，并引入具有SE(squeeze and excite)机制的轻量级注意力机制模块，通过神经结构搜索NAS来获取网络的最佳结构、参数配置，弥补网络轻量化带来的精度损失，既保持了良好的特征提取能力，又极大程度上压缩了模型。此外，本文还改进YOLOv5s的边界框损失函数，使用SIoU Loss[14]作为目标边界框损失函数，相比原损失函数增加了预测边界框和真实边界框中心距离和长宽比例的惩罚项，可以加快网络训练时预测框的收敛，提高回归定位精度。本文改进的检测模型记作M-YOLOv5s，网络结构如图2所示。

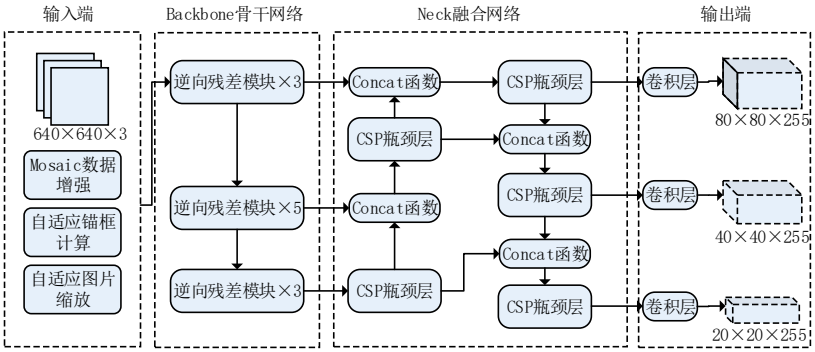


图2　改进的YOLOv5s网络结构

1.2　车辆跟踪

1.2.1　DeepSort

DeepSort算法是经典的多目标跟踪算法，由Sort算法改进而来，由于Sort算法仅考虑了目标运动信息，当目标被遮挡后容易出现跟踪丢失的问题。因此，DeepSort算法在运动信息的基础上添加了外观信息，即加入神经网络提取目标的外观特征信息，提高遮挡情况下的跟踪准确率。除此之外，整体依旧延续了卡尔曼滤波算法预测和匈牙利算法匹配的技术路线，如图3所示。首先，由卡尔曼滤波预测目标运动状态，得到描述目标位置和状态的信息$(x,y,r,h,\dot{x},\dot{y},\dot{r},\dot{h})$；其中：$(x,y)$表示检测框的中心坐标；$r,h$分别表示检测框的长宽比和宽度；$(\dot{x},\dot{y},\dot{r},\dot{h})$表示前四个参数对应的速度信息。其次，结合运动信息和外观信息进行目标关联，计算运动信息时使用马氏距离描述卡尔曼滤波算法预测结果和检测模型检测结果之间的误差，马氏距离表达式为

$$d^{(1)}(i,j)=(\boldsymbol{d}_j-y_i)^{\mathrm{T}}\boldsymbol{S}_i^{-1}(\boldsymbol{d}_j-\boldsymbol{y}_i) \tag{1}$$

其中：$\boldsymbol{d}_j$表示第j个检测结果的状态向量，$\boldsymbol{y}_i$表示第i个预测结果的状态向量。$\boldsymbol{S}_i$表示检测结果和跟踪结果之间的协方差矩阵。对于外观信息，则先计算出每一个检测框$\boldsymbol{d}_j$对应的外观特征描述符r_j。由于车辆和行人运动方式不同，车辆通常是直线运动，相邻两帧的外观相似度变化不大，而行人会经常更换行进位置导致相邻两帧的外观相似度变化较大。因此，对于车辆而言不需要再构建车辆外观特征合集，只计算检测结果和跟踪结果最近时间的相似度即可，一定程度上简化了算法复杂度。外观信息相似度由深度余弦距离衡量，其定义为

$$d^{(2)}(i,j)=1-r_j^{T}r_{j-1} \tag{2}$$

$$c_{i,j}=\lambda d^{(1)}(i,j)+(1-\lambda)d^{(2)}(i,j) \tag{3}$$

随后，使用线性加权的方式(见式(3))结合二者，λ为超参数，由匈牙利算法进行帧间目标匹配。

再次，对状态确定的目标通过前后两帧目标框之间的交并比(IoU)大小构建相似度矩阵，再次使用匈牙利算法匹配。最后，更新卡尔曼滤波算法。

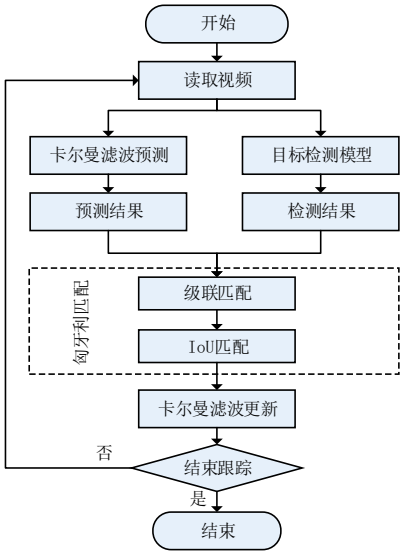


图 3 DeepSort跟踪算法流程

1.2.2 DeepSort改进

在提取外观信息时，原算法使用的是一个离线神经网络(wide residual network，WRN)，但由于WRN参数量和计算量较大，使得DeepSort算法相比Sort算法运行速度降低。鉴于实时性考虑，本文使用轻量型神经网络MobileNetV3-Small作为DeepSort算法的外观特征提取网络。此外，原算法的特征提取网络使用行人重识别数据集训练，并把输入图片限定为128 × 64大小来拟合行人宽高比。因此，为使其更适用于车辆特征提取，本文将输入图像大小调整到64 × 128，并在车辆重识别数据集VeRi[15]上重新进行训练。本文改进的跟踪算法记作M-DeepSort。

1.3 交通流参数检测

1.3.1 流量检测

考虑到DeepSort跟踪算法中卡尔曼滤波需要一定的时间确保预测的稳定性，本文设置了虚拟检测区域，如图4灰色区域所示。当一个全新的车辆ID通过检测区域时，自动生成数据字典保存该ID车辆在检测区域出现的帧次数，根据其出现的次数判断该次计数是否有效，可以减少车辆因出现身份切换IDs导致的流量统计错误，本文将其设置为三次。其次，根据同一ID车辆的在检测区域的运动轨迹的起终点坐标计算一个矢量方向来判断该车辆的运动方向。最后，根据行驶方向信息和流量计数区域的身份检测信息进行双向交通流量统计。

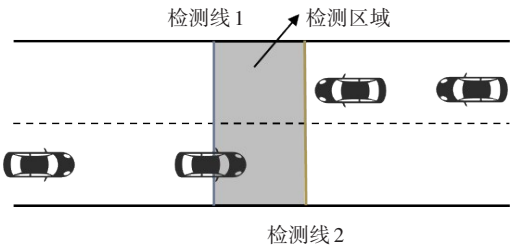


图 4 车辆检测线与检测区域示意图

1.3.2 速度检测

速度由车辆经过两虚拟检测线的时间除以检测线之间的实际距离获得。相比于根据像素距离推算车速的方法，本文方法优点如下：由于不需要对摄像机进行标定，该方法普适性更强；由于只需要计算车辆在两个检测线之间的行驶时间，方法更为简单。速度的计算公式如下：

$$V = 3.6\frac{S}{t}\ ,\ t = \frac{\left|f_2 - f_1\right|}{FPS} \tag{4}$$

其中：f_1为车辆ID检测框质点坐标经过检测线1时的帧数，f_2为经过检测线2时的帧数，FPS为检测视频的帧率，S为两条检测线之间的实际距离，由实际测量确定。

2 实验结果与分析

本文实验在Windows 10环境下使用Python语言，基于PyTorch深度学习框架进行，硬件配置为NVIDIA RTX3090 GPU，Intel(R) i9-10900K CPU。

2.1 模型训练

本节模型训练采用的评价指标包括检测精度AP、运行速度FPS、模型参数量/MB，计算量/G，IDs/次，其中，前两个指标的数值越高，表明模型效果越好，后三个指标的数值越低则表明效果越好。

2.1.1 车辆检测模型训练

车辆检测模型训练使用的数据集为UA-DETRAC数据集，其由北京市、天津市的24个不同卡口视频截取组成，包括近14万帧图像，

121万已标记的检测框，由于两帧图像之间相似度较高，为了减少训练时间，本文每隔10帧抽取一张图片，共获得13402张图像，按照7∶3的比例划分训练集与测试集。

为保证实验的客观性，本文将YOLOv5s和M-YOLOv5s在UA-DETRAC数据集上进行重新训练时超参数的设置保持一致，均使用YOLOv5s预训练模型进行权重初始化，加快网络拟合速度，缩短训练时间。训练图像大小为640×640，优化器为SGD，初始学习率为0.01，动量为0.935，衰减系数为0.0005，训练次数100个轮次，批大小为16，由于仅保留car类目标，故没有分类损失曲线。M-YOLOv5s训练过程的边界框损失和置信度损失随训练轮次变化如图5所示，网络已经基本收敛。

两个模型测试结果见表1，本文轻量化的YOLOv5s的模型参数量仅为3.54 MB，为原YOLOv5s的1/2；计算量为2.94 G，为原YOLOv5s的1/3，而准确率AP较原模型仅下降1.2%，取得了更好的速度与精度之间的均衡，更适合低算力设备部署使用。

表1　车辆检测实验结果

模型	AP@0.5/%	模型参数量/MB	计算量/G
YOLOv5s	97.5	7.23	8.25
M-YOLOv5s	96.3	3.54	2.94

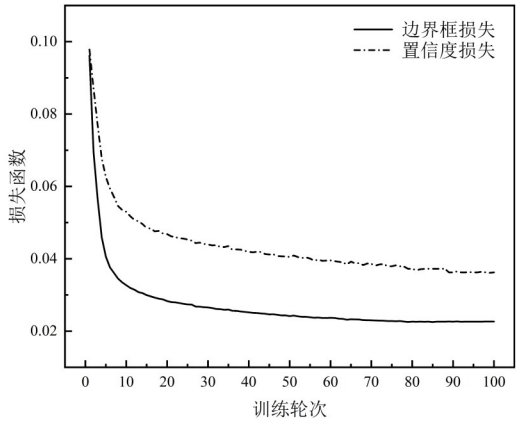


图5　M-YOLOv5s训练损失

2.1.2　车辆跟踪模型训练

将改进外观特征提取网络在VeRi车辆重识别训练集上重训练，VeRi[15]数据集来自20个不同位置的监控摄像头，拍摄场景包括交叉口和道路断面，包括776辆汽车以及从不同视角拍摄的近50000张图像。训练图像大小为64 × 128，训练200个轮次，其余超参数与检测模型训练时的配置相同，训练时的损失函数变化如图6所示，训练200个轮次后网络损失函数变化基本稳定。

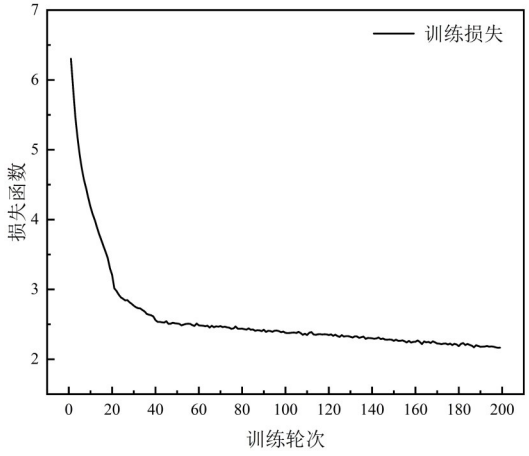


图6　外观特征提取网络训练损失

为验证本文改进跟踪算法的有效性，在UA-DETRAC测试集的MVI_123123、MVI_1234534上进行跟踪效果测试，测试结果见表2。结果表明本文改进算法在所列跟踪指标的表现中均优于原DeepSort，由于经过车辆重识别数据集的训练，本文算法较原算法IDs降低4次。此外，轻量型外观特征提取网络使得算法整体运行速度较原算法提高17%，改进后的跟踪算法更适用于车辆跟踪任务。

表2　车辆跟踪实验结果

模型	IDs/次	FPS/(帧·s^{-1})
M-YOLOv5s+DeepSort	28	52
M-YOLOv5s+M-DeepSort	24	61

2.2　交通流参数检测性能验证

验证使用的视频共两段，每段视频时长5 min，分辨率为1080 p，S=5 m。视频1、2分别对应平峰、高峰场景，每段视频拍摄角度、位置固定。使用准确率评价性能，定义为公式(5)，真实值

由人工采集，估计值由算法自动采集。性能对比时本文算法指M-YOLOv5s+M-DeepSort，原算法指YOLOv5s+DeepSort。

$$准确率 = \left(1 - \frac{|真实值 - 估计值|}{真实值}\right) \times 100 \quad (5)$$

2.2.1 流量检测验证

分别采用人工统计、本文算法、原算法统计每个视频片段上行、下行与总的交通流量，结果如图7所示。两段视频中真实的上行交通流量略小于下行交通流量，算法检测到的交通流量也同样符合此种情况。其次，本文算法较原算法的估计值更接近人工统计的真实值，表3展示了两种算法的统计准确率，本文算法在上行、下行和总和的准确率均高于原算法，表明了本文改进算法的有效性。此外，视频2中由于车流密度增加、车辆遮挡较为严重，本文算法的统计精度较视频1中的精度降低，但精度仍稳定在94%以上，整体表现效果依旧良好。

表 3 交通流量检测准确率对比

算法	视频1			视频2		
	上行/%	下行/%	总和/%	上行/%	下行/%	总和/%
原算法	93.9	95.6	94.8	92.7	94.6	93.7
本文算法	95.1	96.7	95.9	95.6	92.6	94.1

2.2.2 速度检测验证

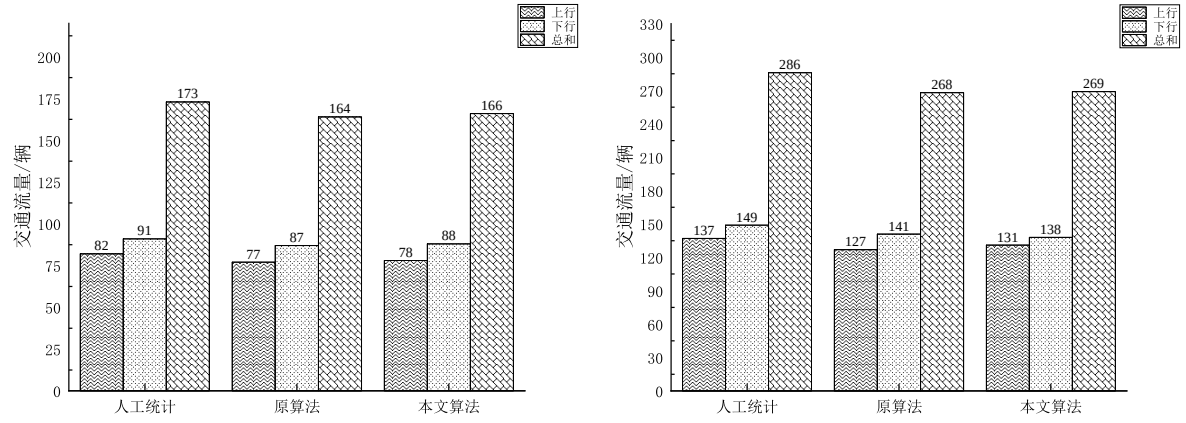
由于无法获取每个车辆真实的行驶速度，故以平均速度为基准进行比较。视频1的实际平均速度为41.56 km/h，视频2为39.87 km/h。由于车流密度增加，车辆之间的影响增大，车辆行驶速度会有所降低，因此视频1中车辆平均速度较视频2稍高，符合实际情况，且算法采集到数据样本均值也符合此情况。此外由表4对比结果可知，本文算法在视频1、2中采集到的速度数据的平均值的准确率比原算法都高，再次证明本文算法的有效性。

表 4 速度估计准确率对比

算法	视频1		视频2	
	准确率/%	平均速度/（km/h）	准确率/%	平均速度/（km/h）
人工检测	—	41.56	—	39.87
原算法	93.19	44.39	91.97	43.14
本文算法	94.08	44.02	92.17	42.99

2.2.3 检测实时性验证

当车流密度增加时算法的运行速度会相应降低，为检验本文算法在高峰时期的运行速度的情况，对算法主要部分耗时情况进行了统计，结果见表5。车辆检测任务耗时10.21 ms，占总消耗时间的61.25%，其次为车辆跟踪任务耗时4.18 ms，最后为视频读取任务2.28 ms。相对于日常视频25帧/s的速率，本文算法可达到59帧/s，可满足实时性的要求。图8展示了本文算法的实际应用效果，从图8可以看出本文算法具有较强



（a）视频1流量检测结果　　（b）视频2流量检测结果

图 7 不同视频车流量检测结果

的可靠性。

表5　算法实时性测试分析　　单位：ms

视频读取	车辆检测	车辆跟踪	总耗时
2.28	10.21	4.18	16.67

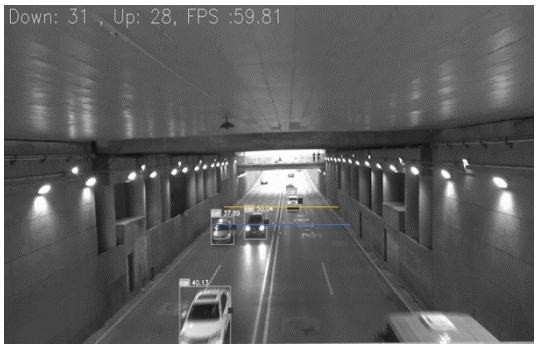


图8　车流量与车速检测效果

3　结语

本文提出了一种基于改进YOLOv5s和DeepSort的交通流参数检测方法，改进的YOLOv5s在精度较原YOLOv5s仅降低1.2%的情况下，参数量和计算量分别降低1/2和2/3；通过重构DeepSort外观特征提取网络，调整图像输入尺寸并进行车辆重识别训练，使得改进DeepSort算法更适合用于车辆的跟踪任务。实际试验结果表明，该方法的平均准确率在92%以上，运行速度可达59帧/s，满足实时使用的要求并具有较强的实用性。未来将进一步提高方法在雨、雪等恶劣天气、夜间或者道路极度拥挤等场景下的适用性。

参考文献：

[1] EL BOUZIADY A, THAMI R O H, GHOGHO M, et al. Vehicle speed estimation using extracted SURF features from stereo images[C]//Processing of the 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), Morocco Fez, 2018:1-6.

[2] TOURANI A, SHAHBAHRAMI A, AKOUSHIDEH A, et al. Motion-based vehicle speed measurement for intelligent transportation systems[J]. International Journal of Image, Graphics and Signal Processing, 2019, 11(4):42-54.

[3] 蒋建国，王涛，齐美彬，等. 基于ViBe的车流量统计算法[J]. 电子测量与仪器学报，2012，26(6)：558-563.

[4] 张冬梅，卢小平，张航，等. 一种基于无人机视频影像的车流量统计算法[J]. 遥感信息，2020，35(1)：142-146.

[5] ARINALDI A, PRADANA J A, GURUSINGA A A. Detection and classification of vehicles for traffic video analytics[J]. Procedia Computer Science, 2018, 144:259-268.

[6] KE R, LI Z, TANG J, et al. Real-time traffic flow parameter estimation from UAV video based on ensemble classifier and optical flow[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(1):54-64.

[7] LIU C, HUYNH D Q, SUN Y, et al. A vision-based pipeline for vehicle counting, speed estimation, and classification [J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(12):7547-7560.

[8] LI J, XU Z, FU L, et al. Domain adaptation from daytime to nighttime: a situation-sensitive vehicle detection and traffic flow parameter estimation framework [J]. Transportation Research Part C: Emerging Technologies, 2021, 124:102946.

[9] LI S, CHANG F, LIU C, et al. Vehicle counting and traffic flow parameter estimation for dense traffic scenes[J]. IET Intelligent Transport Systems, 2020, 14(12):1517-1523.

[10] 赖见辉，王扬，罗甜甜，等. 基于YOLOv3的侧视视频交通流量统计方法与验证[J]. 公路交通科技，2021，38(1)：135-142.

[11] 刘磊，赵栓峰，郭卫. 一种YOLO识别与Mean shift跟踪的车流量统计方法[J]. 制造业自动化，2020，42(2)：16-20.

[12] 文奴，郭仁忠，贺彪. 基于DCN-Mobile-YOLO模型的多车道车辆计数[J]. 深圳大学学报理工版，2021，38(6)：628-635.

[13] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Korea Seoul, 2019:1314-1324.

[14] GEVORGYAN Z. SIoU loss: more powerful learning for bounding box regression[EB/OL]. arXiv: 2205.12740, 2022.

[15] 刘鑫辰. 城市视频监控网络中车辆搜索关键技术研究[D]. 北京：北京邮电大学，2018.

（下转第13页）

文章编号: 1007-1423（2023）14-0008-06　　DOI: 10.3969/j.issn.1007-1423.2023.14.002

基于Charm算法挖掘基因表达保序子序列

廖旭红，江　华，廖　莎，李志杰*

（湖南理工学院信息科学与工程学院，岳阳　414006）

摘要：保序子序列（OPSS）是基因表达数据重要的定性测度双聚类方法，通常将基因实数表达值排序后替换成相应的列标签，OPSS的部分行在部分列下同升同降。提出一种基于Charm算法的保序子序列挖掘方法Charm_Seq，将Charm由频繁闭合项集挖掘改造为频繁闭合序列挖掘，充分利用了Charm高效的Itemset-Tidset前缀搜索树数据结构。在人工和实际基因表达数据集上进行实验，实验结果验证了该方法的高效性和有效性。

关键词：基因表达数据；双聚类；保序子序列；Charm；序列挖掘

0　引言

诞生于上世纪90年代的分子生物学微阵列实验技术，通过生物芯片同时测定成千上万基因在不同实验条件下的表达量，产生了海量的基因表达数据[1]。挖掘基因表达数据中基因活动模式信息，在生物医药等领域有广泛用途。聚类是一种重要的无监督机器学习和数据挖掘技术，基因表达数据传统聚类仅在基因或实验条件单一方向上聚类。

然而，一个生物基因不可能在所有的实验条件下展示共表达特性，也不可能在所有的实验条件下展示相同的水平，却常常参与多种遗传通路。这些特性意味着基因表达数据存在许多潜在的局部模式，只有对基因（行）和实验条件（列）两个方向同时聚类，才可能挖掘出大量有价值的局部模式。

基因表达数据双聚类主要有基于定量测度和基于定性测度的方法。Cheng等[2]引入元素残差与子矩阵均方残差（mean square residue，MSR）的概念，以MSR为评价函数贪婪求解约束优化问题，这种CC算法是典型的基于定量测度的双聚类方法。

多数双聚类方法通过不同基因表达样本相似性度量发现局部模式。Wang等[3]为了指导相似模式聚类，定义了一种新的最近邻测度方法。Liu等[4]以基因表达值排序的顺序而不是欧氏距离作为判断两个基因相似的标准，提出一种灵活有效的保序双聚类模型。保序子序列（order-preserving subsequence，OPSS）是部分行在部分列下具有相同的趋势，实质上是一种排序后的保序子序列挖掘问题。Ben-Dor等[5-6]证明OPSS是NP难题。

本文提出基于Charm[7]的基因表达数据保序子序列挖掘算法Charm_Seq。Charm是离线挖掘频繁闭合项集的最高效算法[8]。Charm_Seq将Charm由频繁闭合项集挖掘改造为频繁闭合序列挖掘，实验验证了算法的有效性。

收稿日期：2023-02-24　　**修稿日期：**2023-06-30

基金项目：湖南省自然科学基金（2019JJ40111）

作者简介：廖旭红（1997—），女，湖南醴陵人，硕士生，研究方向为计算生物学；江华（1996—），男，安徽合肥人，硕士生，研究方向为数据挖掘；廖莎（2000—），女，湖南涟源人，硕士生，研究方向为数据挖掘；*通信作者：李志杰（1964—），男，湖南永兴人，博士，副教授，研究方向为大数据在线学习，E-mail：lzj0019@163.com

1　相关工作

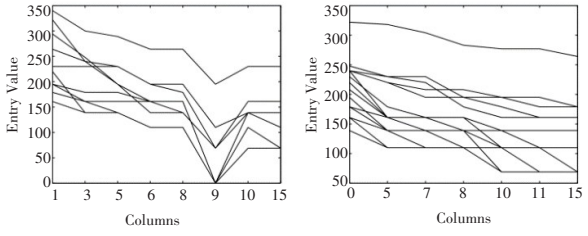
1.1　基因表达数据保序子序列

基因表达数据可表示为一个$n \times m$的数值矩阵$\boldsymbol{A}$，其中元素a_{ij}表示第i个基因(g_i)在第j个实验条件(t_j)下的表达实数值。$\boldsymbol{A}$可形式化表示为$\boldsymbol{A}=(G,C)$，其中，$G=\{g_1,g_2,\cdots,g_i,g_{i+1},\cdots,g_n\}$表示基因行集合，$C=\{c_1,c_2,\cdots,c_j,c_{j+1},\cdots,c_m\}$表示实验条件列集合。表1是一个基因表达数据序列示例。

表 1　基因表达数据序列示例

Gene	t_0	t_1	t_2	t_3	t_4	t_5
g_1	0.217	0.084	0.409	0.138	−0.159	0.129
g_2	0.375	0.115	−0.201	0.254	−0.094	−0.181
g_3	0.238	0.000	0.150	0.165	−0.191	0.132
g_4	−0.073	−0.146	0.442	−0.077	−0.341	0.063
g_5	0.394	0.909	0.443	0.818	1.070	0.227
g_6	0.385	0.822	0.426	0.768	1.013	0.226

在DNA微阵列分析中，密切相关的基因的表达值可能会随一组实验样本相应地同步上升和下降。尽管这些基因的强度表达水平可能不接近，但它们所呈现的模式却非常相似，这种模式即是双聚类局部模式。图1展示从GDS2267酵母菌数据集挖掘的两个局部模式示例，每个模式在条件列集上具有一致递减趋势。



（a）Cluster #1　　（b）Cluster #2

图 1　酵母菌两个双聚类模式示例

假设$I \subset G$，$J \subset C$，$\boldsymbol{A}_{IJ}=(I,J)$表示部分行$I$在部分列$J$下具有相似行为或趋势，$\boldsymbol{A}_{IJ}$也称之为保序子序列。OPSS是矩阵$\boldsymbol{A}$的一种双聚类局部模式，挖掘OPSS是要从给定的基因表达序列$\boldsymbol{A}$中发现具有相似行为或趋势的子序列$\boldsymbol{A}_{IJ}=(I,J)$的集合。

1.2　频繁项集与Charm算法

项集挖掘以事务型数据为挖掘对象，是数据挖掘领域很活跃的研究方向。Charm算法挖掘事务型数据的频繁闭合项集，是最有效的离线频繁项集挖掘算法。

定义1　事务型数据。事务型数据是由事务组成的集合，每个事务是项的集合，称为事务项集。设事务数据的属性集$A=\{a_1,a_2,\cdots,a_n\}$，项为属性的整型取值。每个属性在一个事务中最多一个项，因此，一个事务项集的长度不大于属性集长度。

定义2　频繁项集。一个项集X在事务型数据的所有事务中出现的次数称为项集的支持度$\sigma(X)$。假设事务数据集的最小支持度阈值为min_sup，如果$\sigma(X)\geqslant min_sup$，则称项集$X$为频繁项集。

定义3　频繁闭合项集。假设X是频繁项集，Y表示项集X的任一超项集。如果$\forall Y$，$\sigma(Y)<\sigma(X)$均成立，则称X为频繁闭合项集。

离线和在线频繁模式挖掘典型算法[9-10]有Apriori、Charm、IncMine、Moment等。其中Charm是频繁闭合项集离线挖掘最有效算法，其优越性能主要通过构建<项集×事务集>键值对搜索树，并且键值对表示采用Bitset编码技术。另外，算法采用差集技术减少中间计算节点的内存占用空间，使用基于hash的方法加速清除非闭合的项集等。实验显示[9]，使用Charm作为批处理挖掘器的IncMine算法，比Moment快几个数据级，且使用更少的内存。

Charm的数据结构是一种Itemset-Tidset(IT)前缀搜索树。树中每个节点为IT对，频繁闭合项集为ITSearchTree的叶子节点。该算法首先扫描事务数据库得到频繁项组成的集合I，然后对每个频繁项$X_i \in I$的节点P_i向下深度扩展。

2　基于Charm的频繁闭合序列挖掘

与Charm挖掘频繁闭合项集不同，保序子序列OPSS是挖掘频繁闭合序列，即保序子序列。挖掘频繁闭合项集与挖掘频繁闭合序列的区别如下：

(1) 频繁闭合项集首先搜索频繁项，而频繁闭合序列挖掘首先搜索的是长度为2频繁原

子序列；

（2）频繁闭合项集搜索树下层节点由当前节点与兄弟节点连接生成，而频繁闭合序列增长由当前序列与长度为2频繁原子序列连接实现；

（3）长度为2频繁序列是基本的原子序列，也是所有序列增长的连接对象。

然而，Charm有高效的Itemset-Tidset前缀搜索树数据结构，这是Apriori等没有的。Charm_Seq通过改造Charm算法实现基因表达数据频繁闭合序列挖掘。

基于Charm的保序子序列方法挖掘频繁闭合序列过程有如下三个步骤：

（1）每个基因的所有表达值按大小排序；

（2）各个基因表达值分别替换为相应列标签；

例如表1数据，经步骤(1)和(2)处理后将变成如表2所示的基因表达列序列。

表 2　基因表达列序列

Gene	基因表达值降序排序					
g_1	2	0	3	5	1	4
g_2	0	3	1	4	5	2
g_3	0	3	2	5	1	4
g_4	2	5	0	3	1	4
g_5	4	1	3	2	0	5
g_6	4	1	3	2	0	5

（3）挖掘列标签序列集的频繁闭合序列。

为了挖掘表2中g_1～g_6的频繁闭合序列，可以改造Charm算法为Charm_Seq算法，把挖掘目标由频繁闭合项集转变为频繁闭合序列。在Charm_Seq算法中，设$[P]$表示以P为父节点的所有子节点，$P_i \in [P]$，则P_i向下深度扩展即是$[P_i]$不断取代$[P]$的循环过程。Charm_Seq伪代码如算法1所示。

算法1　Charm_Seq ($\boldsymbol{A}$, min_sup, C=∅)

输入：基因表达数据矩阵$\boldsymbol{A}$，最小支持度阈值min_sup

输出：频繁闭合序列集合C

(1) G = Ordering($\boldsymbol{A}$, T)　　// T为列标签集

(2) I=$[P]$=$\{s_i s_{i+1} \times g(s_i s_{i+1}) : s_i s_{i+1} \subset G \wedge \sigma(s_i s_{i+1}) \geq min\_sup\}$

(3) for each $S_i \times g(S_i)$ in $[P]$

(4)　$[P_i]$ = ∅ and $S = S_i$

(5)　for each $S_j \times g(S_j)$ in I and $S_i \otimes S_j$　//可首尾连接

(6)　$S = S \cup S_j$, $G = g(S_i) \cap g(S_j)$

(7)　if ($\sigma(S) \geq min\_sup$) then

(8)　Add $S \times G$ to $[P_i]$;

(9) if ($[P_i] \neq \varnothing$) then $[P]$ = $[P_i]$, goto (3)

(10) delete $[P_i]$

(11) $C = C \cup S$

以表2中的$\{g_1, g_2, g_3, g_4, g_5, g_6\}$六个基因为例，图2说明Charm_Seq算法挖掘列标签频繁闭合序列的过程。

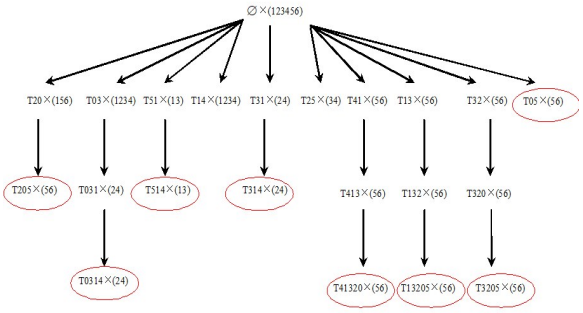


图 2　g_1 ~ g_6的列标签子序列×Gidset搜索树构建过程

3　实验结果与分析

本文使用GEO微阵列基因表达数据集、基于基因表达数据的肿瘤或非肿瘤分类数据集，以及人工数据集对算法的性能进行评价。比较算法包括Charm_Seq、OPSS、CC、Charm、Apriori等。算法用Java语言实现。实验在2.60 GHz、Intel(R) Core(TM) i7-6700HQ CPU、内存16 GB、操作系统Windows 10的计算机上进行。

3.1　数据集

GDS2267微阵列基因表达数据集来自GEO网站：http://www.ncbi.nlm.nih.gov/geo，是GEO公共资源网上关于酵母菌(Saccharomyces cerevisiae)微阵列基因表达数据，数据集名称是Metabolic cycle: time course。该数据集以12～25分钟的间隔对营养有限的连续培养细胞进行三个周期的分析。在这种条件下，生长的细胞以呼吸爆发的形式表现出强健的周期性。数据集对应实验的结果提供了对控制代谢振荡的分子机制的洞察。

四个基准数据集leukemia、colon-cancer、breast-cancer、unbalanced是基于基因表达数

据的肿瘤或非肿瘤分类数据集。其中，leukemia 和 colon-cancer 可从网站下载获得：http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/。breast-cancer 和 unbalanced 则是 Weka 数据分析工具的两个自带数据集。

T10I4D100K 和 T40I10D100K 是两个人工产生项集模式的事务数据集，使用 Zaki's IBM Datagen software 标准符号。该人工数据集句法规则为 TxIyDz[Pu][Cv]，其中 x 是平均事务长度，y 为项集大小(单位为 k)，z 表示所产生事务的数量(单位为 k)。

表 3　实验相关的七个数据集参数

数据集	基因个数	样本大小
GDS2267	9335	36
leukemia	7129	类标 ALL：47 类标 AML：25
colon-cancer	2000	类标 tumor：40 类标 normal：22
breast-cancer	10	类标 recurrence-events：85 no-recurrence-events：201
unbalanced	33	类标 Active：12 类标 Inactive：844
T10I4D100K	10	10000
T40I10D100K	40	100000

3.2　算法性能分析

3.2.1　Charm 挖掘频繁闭合项集

实验以人工事务数据集 T10I4D100K(T10)和 T40I10D100K(T40)为对象，说明 Charm 挖掘频繁闭合项集 FCIs 的有效性和高效性。

(1) 模式挖掘频繁闭合项集数与长度分布。

T10I4D100K 和 T40I10D100K 数据集的 Charm 算法模式挖掘结果如表 4 所示。

表 4　模式挖掘结果

database	#items	avg.length	std.dev.	#records	max.patttern	levels
T10	1000	10	3.7	100000	11	11
T40	1000	40	8.5	100000	25	19

(2) 时间性能。

图 3 和图 4 是 Charm 和 Apriori 算法时间性能对比，它们都是典型的离线挖掘频繁模式算法。实验充分显示了 Charm 算法的高效性。

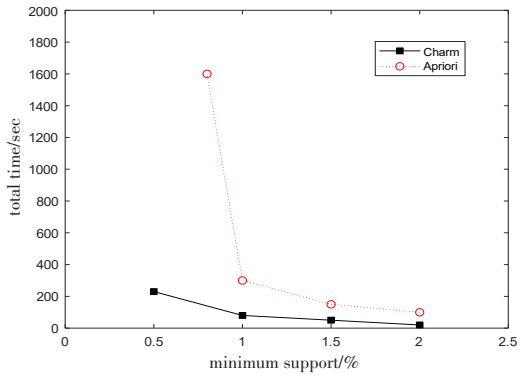


图 3　T40I10D100K 上时间性能对比

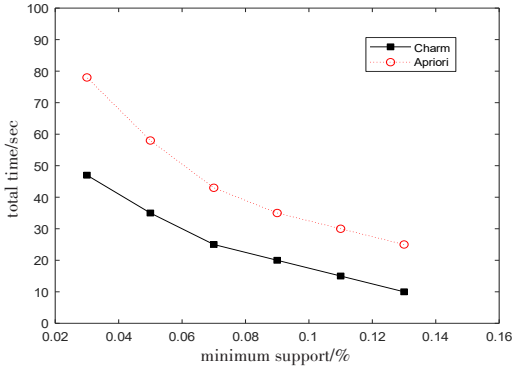


图 4　T10I4D100K 上时间性能对比

Charm_Seq 和 Charm 算法的数据结构相同，挖掘目标的改动并不会影响算法的时间性能。因此，Charm_Seq 和 Charm 算法一样具有高效特性。

3.2.2　Charm_Seq 算法性能分析

(1) 扩展性能。

Charm_Seq 算法行和列的扩展性能如图 5。

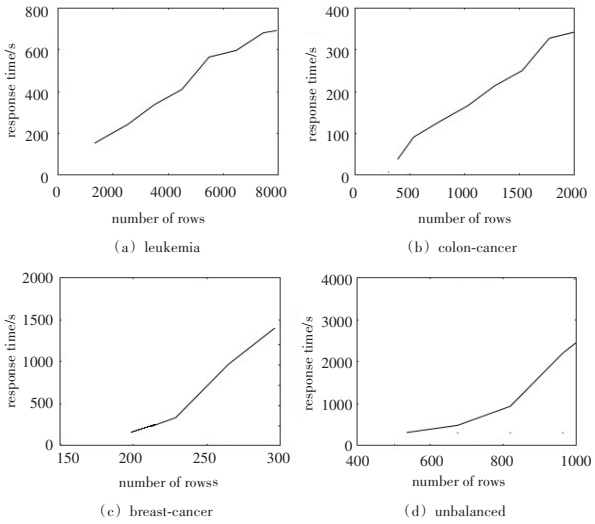


图 5　Charm_Seq 算法行和列的扩展性能

图5(a)、(b)、(c)、(d)分别是Charm_Seq算法在leukemia、colon-cancer、breast-cancer、unbalanced数据集上关于行和列的扩展性能图。从图5显示的趋势看，行扩展曲线与列扩展曲线有一定的对称性。

(2)保序子序列挖掘示例。

表5显示Charm_Seq算法从酵母GDS2267数据集挖掘的五个双聚类相关信息。

图6进一步比较Charm_Seq、CC、OPSS算法的GO功能类别富集程度，使用的数据集为GDS2267。

表5 算法挖掘的酵母五个聚类示例

编号	基因数	条件数	均方残差	行方差
1	79	17	205.44	711.08
8	101	16	221.12	685.33
12	621	11	200.11	1634.32
21	1156	10	221.42	1385.08
32	543	12	199.11	986.09

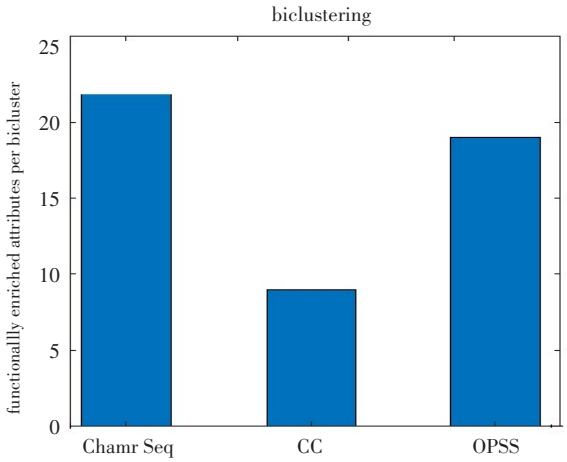


图6 在GO功能方面比较双聚类算法

从图6可以看出，在GDS2267数据集上，Charm_Seq双聚类算法的平均GO功能富集属性数与OPSS大致相当，比定量测度双聚类方法CC高，说明Charm_Seq所得双聚类有较好的生物学意义。

4 结语

与传统的相似测度基于欧氏距离或余弦距离不同，保序子序列基因相似标准是表达水平在相同条件下同升同降。针对NP-难的OPSS模型不适用于大规模基因表达数据分析，本文利用Charm的高效Itemset-Tidset前缀搜索树用于频繁闭合序列挖掘，为求解OPSS问题提供了一种新的尝试。

参考文献：

[1] 姜涛,李战怀.基因数据表达中的局部模式挖掘研究综述[J].计算机研究与发展,2018,55(11):2343-2360.

[2] CHENG Y,CHURCH G M. Biclustering of expression data[C]//Proc. of the 8th Int. Conf. on Intelligent Systems for Molecular(ISMB),Menlo Park,CA:AAAI,2000:93-103.

[3] WANG H,PEI J,YU P. Pattern-based similarity search for microarray data[C]//Proc. of the 11th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining,New York:ACM,2005:814-819.

[4] LIU J,WANG W. OP-clustering by tendency in high dimensional space[C]//Proc. of the 3rd IEEE Int. Conf. on Data Mining,Piscataway,NJ:IEEE,2003:187-194.

[5] BEN-DOR A,CHOR B,KARP R,et al. Discovering local structure in gene expression data[C]//Proc. of the 6th Annual Int. Conf. on Computational Biology,New York:ACM,2002:49-57.

[6] BEN-DOR A,CHOR B,KARP R,et al. Discovering local structure in gene expression data:the order-preserving submatrix problem[J]. Journal of Computational Biology,2003,10(3/4):373-384.

[7] ZAKI M J,HSIAO C-J. CHARM:an efficient algorithm for closed itemset mining[C]//Proc. of the Second SIAM Int. Conf. on Data Mining,Arlington,VA,USA,2002:457-473.

[8] BIFET A,GAVALDÀ R,HOLMES G,et al. 数据流机器学习:MOA实例[M]. 陈瑶,姚毓夏,译. 北京:机械工业出版社,2020.

[9] CHENG J,KE Y,NG W. Maintaining frequent closed itemsets over a sliding window[J]. Journal of Intelligent Information System,2008,31(3):191-215.

[10] 李志杰. 数据流挖掘与在线学习算法[M]. 北京:中国电力出版社,2022.

Mining gene expression order-preserving subsequence based on Charm algorithm

Liao Xuhong, Jiang Hua, Liao Sha, Li Zhijie*

（School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang 414006, China）

Abstract: Order-preserving subsequence（OPSS）is an important biclustering method for gene expression data based on qualitative measures. Usually, the expression value of genes is sorted and replaced with corresponding column labels, and then frequent sequence sets are mined. In this paper, we propose Charm_Seq which is a frequent subsequence mining method based on Charm algorithm. Charm_Seq transformed Charm from frequent closed itemset mining to frequent closed sequence mining, and made full use of Charm's efficient Itemset-Tidset prefix search tree data structure. Experiments are carried out on artificial and real gene expression data sets, and the experimental results verify the efficiency and effectiveness of this method.

Keywords: gene expression data; biclustering; order-preserving subsequence; Charm; sequence mining

（上接第7页）

Traffic flow parameter detection method based on improved YOLOv5 and DeepSort

Shan Zhenyu[1*], Zhang Lin[1], Hou Xiaowen[2]

（1. School of Traffic and Transportation Engineering, Changsha University of Science & Technology, Changsha 410114, China;
2. School of Information and Engineering, Ocean University of China, Qingdao 266100, China）

Abstract: Real-time traffic flow parameter detection method based on improved YOLOv5s and DeepSort is proposed to alleviate the problem of poor video traffic flow parameter detection. First, MobilenetV3 and SIoU Loss are introduced in YOLOv5s to lighten the detection model and improve the detection speed. Secondly, the DeepSort appearance feature extraction network is lightened and retrained on the vehicle re-identification dataset to improve the tracking accuracy. Finally, the two-way traffic flow and vehicle speed are detected based on the virtual detection coil. The experimental results show that the overall operation speed of the algorithm can reach 59 fps, and the accuracy in both flat and peak is above 92%, which has strong robustness in the actual traffic flow parameter detection tasks.

Keywords: traffic flow parameter detection; YOLOv5s; DeepSort; real-time

文章编号: 1007-1423（2023）14-0014-05 DOI: 10.3969/j.issn.1007-1423.2023.14.003

基于改进Faster-RCNN的小目标检测

张 杰*

（安徽理工大学计算机科学与工程学院，淮南 232001）

摘要：现阶段存在的Faster-RCNN基本可以满足普通目标的检测，但是对小目标的检测效果不佳。因此对传统的Faster-RCNN算法进行改进，用Resnet50残差网络替换之前的VGG16网络，让模型提取更多的小目标信息，并引入改进后的特征金字塔MC-FPN，使模型对于小目标信息的检测能力得到提升。实验结果表明，改进后的模型在HRRSD数据集达到86.2%的检测精度，较改进前的检测精度提升了4.7个百分点，证明改进后模型的有效性。

关键词：Faster-RCNN；小目标检测；MC-FPN；特征融合

0 引言

伴随着深度学习的快速发展，小目标检测技术在军事、遥感等领域得到了广泛的应用[1]。计算机视觉领域通常对小目标的定义有两种：一种是相对尺寸大小，在256×256像素图中目标面积小于80像素（即目标面积小于图像面积的0.12%）定义为小目标；二是绝对尺寸大小，以COCO数据集为例，尺寸小于32×32像素目标定义为小目标。小目标由于其分辨率低，在图像中占比低，特征信息得不到很好的利用，容易受到信息混淆等因素的影响，相对于常规目标检测任务来说，目前的主流模型对小目标的检测效果往往不佳。因此如何改善小目标的检测效果，一直是计算机视觉领域的重难点问题。

近些年来，随着深度学习和目标检测任务的结合，让目标检测在各个领域都获得了很好的发展前景，基于深度学习的目标检测算法主要分为两类：第一种是以SSD[2]、YOLO[3]系列为代表的单阶段检测算法，这类算法的检测速度通常比较快，但检测精度较低；第二种是以Faster-RCNN[4]、Mask-RCNN[5]等RCNN系列为代表的双阶段检测算法，检测精度得到了不错的提升，但网络参数比较大，导致检测速度劣于单阶段算法的检测速度。针对小目标容易被忽略，携带的信息有限等特点，提出了许多有效的改进模型。Qu等[6]把膨胀卷积和特征融合一起使用，增强深层特征的语义信息来加强对遥感小目标的检测效果。Li等[7]引入一种特征融合模块加入到特征金字塔中，改善对小目标的检测效果。亢洁等[8]提出了新的多尺度融合模块，通过通道注意力机制重新分配通道权重，增强浅层感受野来提高小目标的检测效果。在PANet当中，Liu等[9]对特征金字塔网络不同特征图进行二次融合，这样可以使特征金字塔的高层特征图同时包含低层的特征信息和高层的特征，从而达到提高小目标检测精度的目的；以上的方法通过不同的改进，直接或间接使小目标的检测效果得到了提高，但仍有一些不足。

本文基于Faster-RCNN算法，提出了一种改进的小目标检测算法：

（1）采用ResNet替换传统的VGG提取网络，减少VGG池化造成的语义特征丢失等问题，加强模型对小目标信息的提取能力。

（2）提出改进后的特征金字塔模型MC-FPN，通过扩充上下文特征信息和引入通道注意力机制来提高整个模型对小目标的检测精度。

收稿日期：2023-03-22 **修稿日期：**2023-04-06

作者简介：*通信作者:张杰(1999—),男,安徽六安人,硕士研究生,研究方向为目标检测,E-mail:2934049959@qq.com

1　基于改进Faster-RCNN的小目标检测算法

本文针对传统的Faster-RCNN算法对小目标检测精度不佳的问题，提出改进Faster-RCNN的小目标检测算法，使用ResNet50替换传统的VGG16作为新的特征提取网络，其次融合改进后的特征金字塔MC-FPN多尺度特征以提高整体模型对小目标的检测能力。

1.1　Faster-RCNN架构

Faster-RCNN是一种基于候选区域的端到端的两阶段目标检测算法，是以Fast-RCNN模型作为基础，加入新的区域建议网络(RPN)，通过滑动窗口在相应的特征图上生成所需要的候选区域，即锚框，紧接着通过锚框得到输出类别以及预测框，最后在使用非极大抑制算法对前面的预测结果进行分析，最终获得所需要的候选区域，总体框架如图1所示。

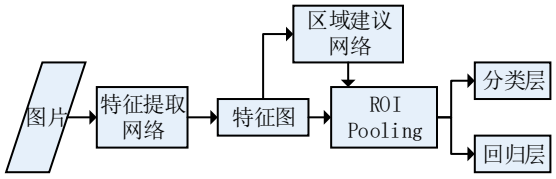


图1　Faster-RCNN总体框架

Faster-RCNN网络架构主要是由特征提取模块、区域建议网络(RPN)、ROI Pooling层和分类回归层四部分组成，首先对输入的图片进行特征提取，获得所需特征图，再将特征图传输到区域建议网络从而生成一系列预选框，同时再将生成的预选框与特征图一起传输到ROI Pooling层，目的是从一系列预选框当中选出最适合特征图的候选框，最后将选出的候选框传送到分类和回归层，这就是Faster-RCNN算法的流程。

1.2　ResNet50残差网络

自AlexNet发展以来，网络结构一直朝着深度进行研究，大家便认为随着网络深度的增加，特征提取网络的拟合能力也会不断变强，从而模型取得一个较好的结果。但逐渐发现，并不是网络层数越深，模型的检测效果越好，精度会一直提升，反而在训练甚至是测试的时候误差都开始变大；在这种情况下，ResNet于2015年被提出，ResNet网络可以在加深网络层数提取更多特征信息的同时解决之前深度增加所导致的训练精度下降的问题。ResNet50采用的是跳跃连接，其输入和输出端直接相连，如图2所示，其中两个1×1卷积的作用是降维和升维，3×3卷积的作用是提取特征信息。

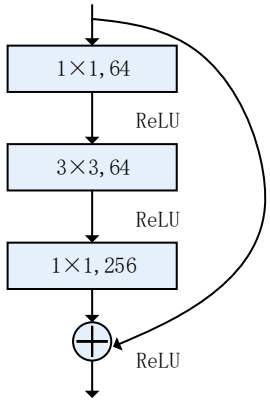


图2　ResNet50残差网络

1.3　MC-FPN

特征金字塔(feature pyramid network，FPN)通过引入自下而上，自顶向下的路径和横向连接的方式，将高层特征图的语义信息和低层特征图的位置信息相结合，从而提升小目标的检测能力，但P5仅通过C5得到，而特征金字塔的高层主要用于大中目标的检测，这就导致对小目标的检测效果较差。为解决上述问题，本文提出一种改进后的特征金字塔MC-FPN，如图3所示，对FPN的改进主要有以下两个部分：①在C5和P5直接添加一个多分支空洞卷积模块MCCM(multi-branch cavity convolution module)，如图4所示，此模块是由三个不同大小空洞率的空洞卷积并联而成，其作用是通过扩大感受野来捕获更多的小目标特征信息，然后注入到特征金字塔当中，使上下文信息得到更加充分的利用。②引入CBAM(convolutional block attention module)[10]注意力机制，如图5所示，注意力机制CBAM是结合了空间和通道的注意力模块，由通道注意力模块和空间注意力模块组成，通过融合这两

个模块，对特征图进行细化处理。CBAM注意力机制对最后输出的特征图进行处理，是因为深层特征图含有较多的小目标信息，而特征金字塔信息融合过程导致的信息混淆对深层的小目标信息不太友好，存在漏检等问题，所以在深层P2和P3处引入注意力机制，可以使模型更好地关注小目标信息，从而提高整个模型对小目标的检测效果。

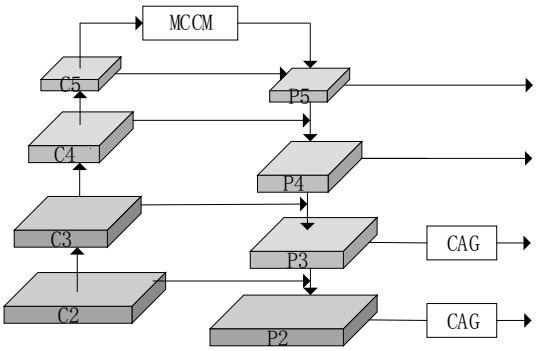


图3 MC-FPN结构

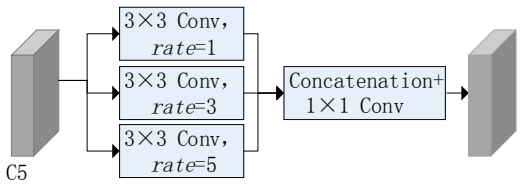


图4 多分支空洞卷积模块(MCCM)

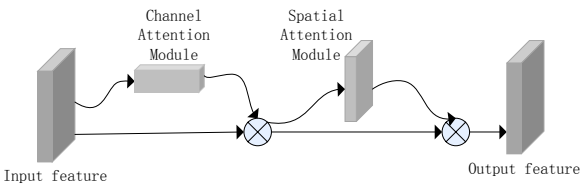


图5 CBAM注意力机制

2 实验与分析

2.1 实验平台与数据集

本文实验采用的是基于深度学习框架的PyTorch1.7.1，操作系统为Ubuntu18.04，CUDA为11.0，GPU型号为NVIDIA GeForce RTX 3090，Python的使用版本为3.7。模型训练时采用随机梯度下降(stochastic gradient descent, SGD)作为优化器。

本文选择中国科学院发布的高分辨率遥感图像公共数据集HRRSD，该数据集共含有图像21761张，13个类别，分别是飞机、棒球场、篮球场、桥梁、十字路口、田径场、港口、停车场、船、存储罐、丁字路口、网球场和汽车。各个类别之间样本的数量比较均匀，其中大部分类别是以密集排布的小目标形式进行分布，可以用来验证算法模型对小目标的检测效果。以数据集划分，选用1∶1∶2的方式选取图像进行训练，评估以及最后的测试，并对数据集样本进行简单地水平、垂直翻转，以提高模型的泛化能力。

2.2 评价指标

本文的评价指标是目标检测领域常用的平均精度(mAP)，表示所有类别AP的平均值。AP的定义为

$$AP = \int_0^1 p(r)\mathrm{d}r \tag{1}$$

mAP的定义为

$$mAP = (AP_1 + AP_2 + AP_3 + \cdots + AP_k)/k \tag{2}$$

2.3 实验结果分析

特征提取网络的优越性对于一个模型检测性能的好坏和分类结果有着直接的影响，为了证明改进后模型对于小目标的检测效果，在参数相同的情况下做了多组对比实验，首先验证了ResNet50特征提取网络与传统VGG16，实验结果见表1。

从表1可以看出，两种特征提取网络当中，ResNet50表现出来的性能优于VGG16，mAP达到82.7%，比VGG16高出1.2个百分点，其中各种类别的检测精度也有不同程度的提升，表明在Faster-RCNN模型当中，ResNet50特征提取网络对小目标的检测效果略优于VGG16特征提取网络。

单一地替换特征提取网络对小目标的检测效果提升不是很高，因此我们对FPN改进，提取一种改进后的特征金字塔MC-FPN，使之与特征提取网络相融合，这对于小目标的检测效果有

很好的提升，改进后的模型对比实验见表2。

表 1　特征提取网络的比较

Backbone	VGG16/%	ResNet50/%
Airplane	90.8	92.8
Baseball Diamond	86.9	87.2
Basketball Court	47.9	49.7
Bridge	85.5	85.8
Crossroad	88.6	88.7
Ground Track Field	90.6	90.7
Harbor	89.4	89.7
Parking lot	63.3	65.3
Ship	88.5	88.9
Storage Tank	88.7	89.9
T junction	75.1	75.8
Tennis Court	80.7	81.9
Vehicle	84.0	87.8
mAP/%	81.5	**82.7**

表 2　改进模型的比较

算法	Faster-RCNN/%	Faster-RCNN+ ResNet50/%	本文算法/%
Airplane	90.8	92.8	**97.8**
Baseball Diamond	86.9	87.2	**88.7**
Basketball Court	47.9	49.7	**65.7**
Bridge	85.5	85.8	**86.4**
Crossroad	88.6	88.7	87.6
Ground Track Field	90.6	90.7	**97.0**
Harbor	89.4	89.7	**92.9**
Parking lot	63.3	65.3	**65.7**
Ship	88.5	88.9	**89.6**
Storage Tank	88.7	89.9	**94.1**
T junction	75.1	75.8	68.6
Tennis Court	80.7	81.9	**91.7**
Vehicle	84.0	87.8	**95.1**
mAP/%	81.5	82.7	**86.2**

从表2可以看出，本文算法(ResNet50+MC-FPN)与传统的Faster-RCNN算法和使用ResNet50特征提取网络的Faster-RCNN算法在HRRSD数据集上检测结果的对比，mAP较改进之前提升了4.7个百分点，其中汽车、存储罐、飞机、网球场等小目标检测精度都有显著的提升，进一步验证本文模型的优越性。

图6展示了Faster-RCNN算法改进前后的检测效果图，可以明显看出改进后的算法对于小目标的检测明显比改进前更优。

（a）改进前　　（b）改进后

图 6　检测效果

3　结语

本文主要针对传统的Faster-RCNN算法对小目标检测效果不佳的问题做出改进，首先用ResNet50残差网络替换传统的VGG16网络，以便提取丰富的特征信息，一定程度上提升了模型对小目标的检测精度；紧接着提出MC-FPN模型，减少特征融合过程带来的信息混淆和冗余，从而增强模型对小目标的检测效果。通过前后对比实验证明，改进后的算法对于小目标的检测效果优于改进前的算法，证明改进算法的鲁棒性。

参考文献：

[1] TONG K, WU Y Q, ZHOU F. Recent advances in small object detection based on deep learning: a review [J]. Image and Vision Computing, 2020, 97: 103910.

[2] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector [C] //Proceedings of the European Conference on Computer Vision, Cham: Springer, 2016: 21-37.

[3] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C] //Proceedings of the IEEE International Conference on Computer Vision, Los Alamitos: IEEE Computer Society Press, 2016: 779-788.

[4] REN S, HE K, GIRSHICK R, et al. Faster R-CNN:

towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39 (6): 1137-1149.

[5] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 2017: 2961-2969.

[6] QU J S, SU C, ZHANG Z W, et al. Dilated convolution and feature fusion SSD network for small object detection in remote sensing images [J]. IEEE Access, 2020, 8: 82832-82843.

[7] LI Z X, ZHOU F Q. FSSD: feature fusion single shot multibox detector [EB/OL]. arXiv: 1712.00960, 2018.

[8] 亢洁, 刘港, 郭国法. 基于多尺度融合模块和特征增强的杂草检测方法 [J]. 农业机报, 2022, 53 (4): 254-260.

[9] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, New York: IEEE Press, 2018: 8759-8768.

[10] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu: IEEE, 2017: 936-944.

Small target detection based on improved Faster-RCNN

Zhang Jie*

(School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China)

Abstract: The Faster-RCNN that exists at this stage can basically satisfy the detection of ordinary targets, but is not effective in detecting small targets. Therefore, the traditional Faster-RCNN algorithm is improved by replacing the previous VGG16 network with the Resnet50 residual network, allowing the model to extract more information about small targets, and introducing the improved feature pyramid MC-FPN to enhance the detection capability of the model for small targets. The experimental results show that the improved model achieves 86.2% detection accuracy in the HRRSD dataset, which is 4.7 percentage point higher than the detection accuracy before the improvement, proving the effectiveness of the improved model.

Keywords: Faster-RCNN; small target detection; MC-FPN; feature fusion

文章编号: 1007-1423（2023）14-0019-04 DOI: 10.3969/j.issn.1007-1423.2023.14.004

基于信息融合的多列卷积神经网络异常驾驶研究

王富强*，龙 涛

（西安明德理工学院信息工程学院，西安 710100）

摘要：针对传统卷积神经网络（CNN）在异常驾驶检测方面识别率低、处理能力差的问题。对异常驾驶状态进行研究，提出了一种基于信息融合的多列卷积神经网络用来检测异常驾驶状态。首先建立三个卷积局部感受野核大小不同的CNN，然后将三列卷积神经网络卷积特征图进行融合，最后，融合的卷积特征图通过全连接层进行降维，输出不同的驾驶行为。实验结果表明，该方法相对于传统的卷积神经网络取得了更高的准确性、鲁棒性。在state farm distracted driver detection数据集上识别率达到了89.8%。

关键词：异常驾驶；信息融合；卷积神经网路；深度学习

0 引言

随着我国经济的快速增长，我国机动车拥有量已达到历史新高。据公安部统计，截至2023年1月11日，我国机动车保有量已达4.17亿辆，每年仍以10%左右的速度在快速增长[1]。由此引发越来越多的交通事故，有90%以上事故是由于驾驶员操作不当引起的，其中最突出的就是疲劳驾驶和分心驾驶引起的操作不当。驾驶员疲劳驾驶和分心驾驶导致的交通事故已占交通事故的30%～40%，尤其在高速上，高达40%以上，所以近几年来疲劳驾驶和分心驾驶已经成为轨道交通安全领域的研究热点[2]。

当前驾驶员在驾驶习惯中存在着玩手机、打电话、东张西望、喝水、吸烟、疲劳驾驶、和后排乘客聊天等不良驾驶习惯，这些行为都会给安全驾驶构成一定的威胁。在驾驶期间当驾驶员出现上述行为时如果能够提醒驾驶员以减少驾驶员分心，会减少交通事故发生，保护人民生命财产。

异常驾驶是一种注意力不集中的行为表现，美国汽车协会交通安全基金会（AAAFTS）将异常驾驶定义为驾驶员由于车内或车外发生的事件，导致驾驶员注意力从驾驶任务转移，对安全完成驾驶任务所需的信息识别较慢的反应。异常驾驶可以分为四种主要类型[3]：视觉干扰、听觉干扰、认知干扰和生物力学干扰。视觉干扰是指驾驶员在车内或车外观察其他事件、物体或人时视线的转移；认知干扰被定义为由于思考其他事情而从驾驶中转移注意力；听觉干扰的定义是由于使用手机、与其他乘客交流或使用其他音频设备而从驾驶中分心。

1 相关工作

为了减少交通事故和提高道路安全，人们提出了各种基于计算机视觉的方法。Kaggle发起了一项名为State farm distracted driver detection的竞赛，旨在通过一个仪表盘摄像头拍摄的图像，将注意力分散的驾驶行为与安全驾驶区分开来。在本文中主要利用图像识别技术检测驾驶员在驾车行驶过程中的不规范行为，以及时提醒驾驶员，减少交通事故的发生。

驾驶员异常检测基于传统的检测方法主要分为基于生理信号的检测、基于车辆行驶状态的检测和基于视觉的检测三类。基于脑电信号

收稿日期： 2023-05-16 **修稿日期：** 2023-06-24

作者简介： *通信作者：王富强（1994—），男，甘肃庆阳人，硕士研究生，助教，研究方向为图像处理、人工智能，E-mail：wangfuqiangabc@163.com；龙涛（1981—），男，陕西汉中人，硕士研究生，助教，研究方向为软件工程

的异常驾驶检测主要是通过传感器采集驾驶员生理信号来分析判断驾驶员是否处于异常驾驶状态。Li等[4]提出了通过小波变换分析心率变异性来检测驾驶员是否处于异常驾驶状态，此方法达到了95%的准确率，但其是一种侵入性检测方式，对正常驾驶有一定干扰，目前只应用于理论研究。基于车辆行驶状态的检测是通过判断车辆有无偏离车道线、方向盘偏转角度、车速等来判断驾驶员是否处于异常驾驶状态；屈肖蕾等[5]提出通过提取车辆转向操作特性和车辆状态特征，运用SVM算法判断驾驶员是否处于异常驾驶状态；Hu等[6]通过获取车辆实时速度运用局部设计的神经网络来判断驾驶员是否处于异常驾驶状态，其缺点是该方法受道路环境、驾驶员驾驶经验等因素影响。基于视觉的检测是通过摄像头实时采集驾驶员头部姿态，从采集的实时视频中提取帧图像来检测驾驶员是否存在喝水、东张西望、抽烟和玩手机等特征来判断驾驶员是否存在异常驾驶；Yan等[7]通过对驾驶员手部位置进行监测，来判断驾驶员是否处于异常驾驶；Ragab等[8]通过对6名受试者眼睛状态、手臂位置、面部表情和面部方向采用AdaBoost、隐马尔可夫模型、随机森林和神经网络进行异常检测；Eraqi等[9]提出了遗传加权的卷积神经网络进行异常驾驶检测，达到了90%准确率；Hu等[3]提出基于信息融合的多列卷积神经网络异常驾驶检测，但该方法存在网络中参数多、时间开销大且易过拟合等不足。

深度学习概念是由Hinton等[10]于2006年提出的，是机器学习中一种基于大量数据学习特征的学习方法，是机器学习的一个新的研究领域。受Hubel和Wiesel对猫视觉皮层电生理研究启发，提出卷积神经网络（convolutional neural network，CNN）。

本文主要通过对深度学习中经典的模型-卷积神经网络进行改进，来检测驾驶员在驾驶过程中出现的异常驾驶行为，从而达到发出精准警告信息的目的，进而有效地降低交通事故的发生。因此驾驶员违规行为识别研究就变得十分重要且有意义，本文正是基于此做的相关研究。

2 基于信息融合的多列卷积神经网络异常驾驶检测

考虑到普通卷积神经网络识别率低、鲁棒性差，本文提出一种基于信息融合的多列卷积神经网络模型，如图1所示，本模型由三列卷积神经网络构成，每列卷积神经网络结构相同，只是卷积核大小不一样，卷积核大小分别为3×3、5×5、7×7，每列卷积神经网络由VGG16结构改进而成，结构如图2所示。

每列卷积神经网络包含10个层、8个卷积层、一个全局平均池化层和一个全连接层，它以640×480的RGB图像作为输入，8个卷积层可以分为五个阶段来实现，全局平均池化层（global average pooling，GAP)将卷积后的每个卷积特征图均值，所有的卷积特征图经过全局平均池化层后输入全连接层（fully connected，FC），

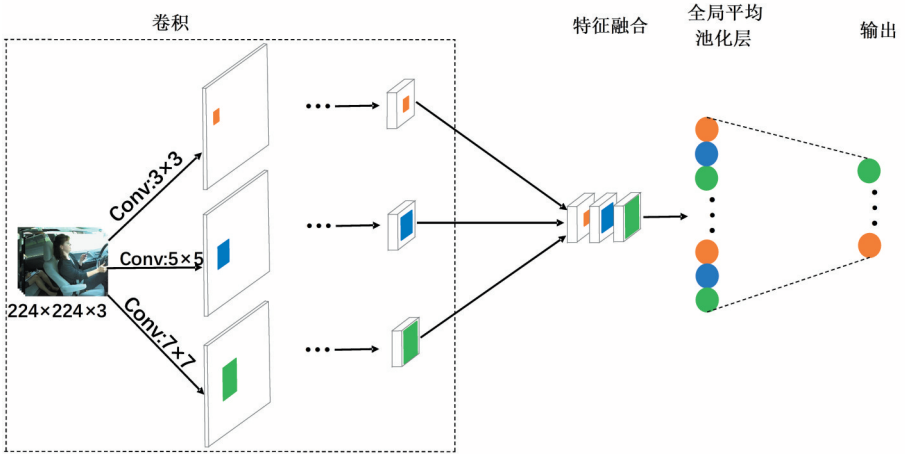


图1 信息融合多列卷积神经网络模型

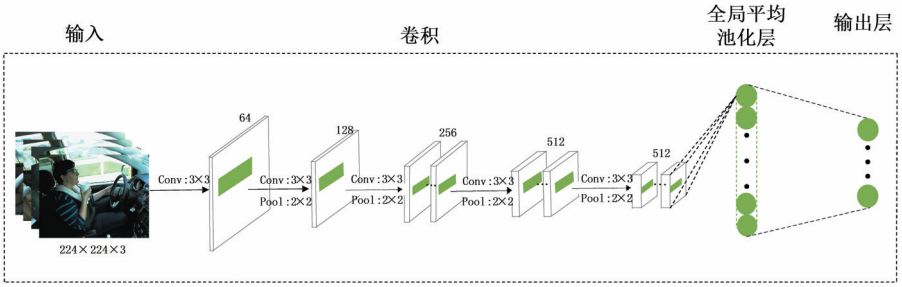


图 2　单列卷积神经网络结构

最后通过Softmax分类器输出不同驾驶行为的概率。每列卷积神经网络激活函数采用ReLU，步长设置为2，最大池化(MaxPool)尺寸选为2×2。

3　实验与结果分析

3.1　数据集介绍

实验数据集采用Kaggle竞赛官方提供的State farm distracted driver detection驾驶员行为标准数据集(https://www.kaggle.com/)，该数据集由102150张640×480的RGB彩色图片构成，包含有十种驾驶状态，其中，训练集提供了22424张图片，测试集提供了79726张图片，每种驾驶状态提供的数据样本数见表1，每种驾驶状态如图3所示。

图 3　State farm distracted driver detection数据集十种驾驶状态

表 1　数据集详情

司机行为类别	训练集图片/张	测试集图片/张
安全驾驶	2489	8726
右手发短信	2267	8353
右手打电话	2317	8293
左手发短信	2346	8264
左手打电话	2326	8344
操作收音机	2312	8435
喝东西	2325	8052
伸手到后面	2002	7261
整理发型和妆容	1911	6984
和乘客交谈	2129	7011

3.2　结果分析

在State farm distracted driver detection标准数据集上，将本文提出的多列卷积神经网络与其他算法进行了比较。在该数据集上将图片大小修改为224×224×3，学习率设置为0.0001。

大量研究者在State farm distracted driver detection数据集上做了相关研究，都取得了不错的研究成果，本文主要针对Alexnet、ResNet34和本文提出的融合算法进行了对比实验，主要从算法的识别准确率和精确率方面进行了对比分析，具体对比结果见表2和表3。

表 2　不同算法在State farm distracted driver detection数据集上的准确率/%

方法	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	总体
Alexnet	78.2	66.1	69.0	71.3	65.2	73.9	74.8	77.5	76.6	68.5	72.1
ResNet34	87.9	85.0	83.8	82.2	86.7	87.5	81.9	84.4	88.9	85.1	85.3
本文算法	87.9	89.6	91.0	89.9	90.3	90.6	89.8	91.6	89.2	88.3	89.8

表 3　不同算法在State farm distracted driver detection数据集上的精确率/%

方法	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	总体
Alexnet	73.1	73.4	72	73.1	69.5	70.7	68.7	70.6	73.8	72.6	71.8
ResNet34	82.3	84.1	81.6	82.5	84.7	83.8	82.1	81.9	84.2	81.4	82.9
本文算法	87.5	89.2	89.8	88.5	89.9	89.9	90.3	88.7	88.5	89.1	89.1

4 结语

本文提出了一种基于信息融合的多列卷积神经网络的异常驾驶行为识别方法。该方法首先利用卷积核大小不一样的卷积神经网络进行卷积，将每列卷积神经网络得到的卷积特征进行融合，然后通过全局平均池化层进行特征均值，全连接层将特征均值进行降维，最后利用多分类函数Softmax输出不同驾驶行为的概率。相对传统的卷积神经网络，本文方法有效减少了参数运算量，避免了全连接层带来的过拟合问题，提高了分类正确率。

参考文献：

[1] 我国新能源汽车保有量达1310万辆呈高速增长态势[N/OL]. https://www.gov.cn/xinwen/2023-01/11/content_5736281.htm.

[2] 王富强,刘德胜,刘云鹏. 基于面部特征的疲劳驾驶检测技术研究[J]. 现代计算机,2021,27(7):121-124.

[3] HU Y,LU M,LU X. Driving behaviour recognition from still images by using multi-stream fusion CNN [J]. Machine Vision and Applications, 2019, 30: 851-865.

[4] LI G ,CHUNG W Y . Detection of driver drowsiness using wavelet analysis of heart rate variability and a support vector machine classifier[J]. Sensors,2013,13(12):16494-16511.

[5] 屈肖蕾,成波,林庆峰,等. 基于驾驶员转向操作特性的疲劳驾驶检测[J]. 汽车工程,2013,35(9):803-807.

[6] HU J,XU L,HE X,et al. Abnormal driving detection based on normalized driving behavior[J]. IEEE Transactions on Vehicular Technology,2017,66(8):6645-6652.

[7] YAN C,COENEN F,ZHANG B. Driving posture recognition by convolutional neural networks [J]. IET Computer Vision,2016,10(2):103-114.

[8] RAGAB A, CRAYE C, KAMEL M S, et al. A visual-based driver distraction recognition and detection using random forest [C] //Proceedings of the 11th International Conference on Image Analysis and Recognition,Vilamoura,Portugal,2014:256-265.

[9] ERAQI H M,ABOUELNAGA Y,SAAD M H,et al. Driver distraction identification with an ensemble of convolutional neural networks[J]. Journal of Advanced Transportation,2019,2019:465-476.

[10] HINTON G E,OSINDERO S,TEH Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation,2006,18(7):1527-1554.

Research on abnormal driving in multi-column convolutional neural networks based on information fusion

Wang Fuqiang*, Long Tao

(Information Engineering College of Xi'an Mingde Institute of Technology, Xi'an 710100, China)

Abstract: The traditional convolutional neural network(CNN) has a low recognition rate and poor processing ability in abnormal driving detection methods. The abnormal driving state is studied, and a multi-column convolutional neural network based on information fusion is proposed to detect the abnormal driving state. First, three CNNs with different sizes of convolutional local Receptive field nuclei are established, and the three-column convolutional neural network convolution feature map is fused. Finally, the fused convolutional feature map is dimensionally reduced through the full connection layer, and different driving behaviors are output. The analysis of experimental results shows that this method achieves higher accuracy and robustness compared to traditional convolutional neural networks. The recognition rate on the state farm dispersed driver detection dataset reached 89.8%.

Keywords: abnormal driving; information fusion; convolutional neural network; deep learning

文章编号: 1007-1423（2023）14-0023-05 DOI: 10.3969/j.issn.1007-1423.2023.14.005

基于编辑距离的字符串相似度算法研究

张胜楠*

（湖北轻工职业技术学院信息工程学院，武汉 430000）

摘 要：基于编辑距离（LD）求解字符串相似度的算法非常经典，但其在普适性和精确性方面略有不足，基于最长公共子串（LCCS）和最长公共子序列（LCS）对其改进，使计算结果更有区分性、普适性和精确性。另外在计算相似度时，对LD和LCS的求解算法从数据结构的角度进行了优化，在数量级上降低了算法空间复杂度。对实验结果进行了对比分析，证明其可行性和正确性。

关键词：相似度计算；编辑距离；最长公共子序列；最长公共子串

0 引言

求字符串相似度的算法在自然语言处理、抄袭检测、网页搜索等领域应用广泛，如编辑距离算法、最长公共子串算法、贪心字符串匹配算法、Heckel算法等非常经典。当比较两字符串相似度时，编辑距离算法基于字面相似的方法计算，针对不同的需求对编辑距离算法改进的方法层出不穷，有作者考虑了非相邻字符间的交换操作，使编辑操作次数最小化；有作者从多公共字符串的熵出发来计算文本相似度等。这些方法往往只考虑他们的编辑次数而忽略了最长公共子序列（longest common subsequence，LCS）对字符串间相似度的影响，或者考虑了*LCS*但忽略了最长公共子串（longest common substirng，LCCS）的影响，使用传统的基于编辑距离求相似度的公式无法得到合理的结果。如情况一，字符串*S*=“abcd”与T_1=“dcba”和T_2=“cdab”比较相似性时，计算的编辑距离*LD*都是4，若不考虑*LCS*，就无法判断出T_2与*S*更相似；再比如情况二，字符串*S*=“abcdef”与T_1=“amcnf”和T_2=“abcmng”比较相似性时，计算的编辑距离（levenshtein distance，LD）和最长公共子序列*LCS*都是3，所以即使考虑了*LCS*，也无法判断出T_2与*S*更相似。

*LCS*是在字符相对位置不变的情况下并不要求必须连续，它反映了字符串的雷同性，常用于相似性检查。*LCCS*则必须是连续的最长的公共子串，对相似度计算具有一定的影响，本文提出一种基于*LD*综合考虑*LCS*和*LCCS*的字符串相似度算法，以改善准确性和适用性。

1 基于编辑距离的字符串相似度算法

1.1 基于编辑距离的字符串相似度计算

编辑距离由俄国科学家Vladimir Levenshtein于1965年提出，以字符为编辑单位，计算由一个字符串到另一个字符串的最少操作（删除、插入、替换）次数，常被用于字符串的相似度计算。设两个字符串$S=S_1S_2S_3\cdots S_m$和$T=T_1T_2T_3\cdots T_n$，构造矩阵$LD[m+1, n+1]$，使用动态规划的思想循环计算矩阵中每个单元格$LD(i,j)$的值，最右

收稿日期：2023-02-27 **修稿日期：**2023-06-02

作者简介：*通信作者：张胜楠（1991—），女，河南商丘人，硕士，助教，专任教师，研究方向为软件工程、系统集成项目管理工程师，E-mail：1653921195@qq.com

下角的$LD(m,n)$即为所求编辑距离大小，计算公式如下[1]：

$$LD(i,j)=\begin{cases}0,\ i=0,j=0\\ j,\ i=0,j>0\\ i,\ i>0,j=0\\ \mathrm{Min},\ i>0,j>0\end{cases}\tag{1}$$

$\mathrm{Min}=\min\{LD(i-1,j)+1,LD(i,j-1)+1,LD(i-1,j-1)+f(i,j)\}$，其中，当$S$的第$i$个词不等于$T$的第$j$个词时，$f(i,j)=1$；否则，$f(i,j)=0$。$LD$本身就能表示两个字符串的相似程度，直观上$LD$越大，相似性越小。表示相似度$Sim(S,T)$的计算公式如下：

$$Sim(S,T)=1-LD/\max(m,n)\tag{2}$$

但其并不具备好的普遍适用性，例如设字符串S: abcdef, T: mefngh 则$LD(S,T)=6$, $Sim(S,T)=1-6/6=0$，两字符串的相似度却是0，显然不符合实际情况。

1.2 结合LCS和LD的字符串相似度算法

求解LCS长度的方法很多[2]，如穷举法：对T串的所有子序列检查是否为S的子序列，从而确定是否为公共子序列，再选出最长的，但是对于两个长度分别为m、n的字符串，时间复杂度为指数级别的$O(2n*2m)$；还有递归法：编程简单、易于理解，但由于有大量重复递归调用，效率不高。经典的动态规划的方法：将复杂问题简单化，引入数组来存放所有子问题的解，从而省去重复求解相同子问题的麻烦，当求LCS时，用$L[i,j]$记录序列S_i和T_j的最长公共子序列的长度，最后，S和T的LCS的长度记录于$L[m,n]$中，建立递归关系如下：

$$L[i,j]=\begin{cases}0,i=0\text{或}j=0\\ L[i-1,j-1]+1,i,j>0,S_i==T_j\\ \max\left(L[i,j-1],L[i-1,j]\right),i,j>0,S_i!=T_j\end{cases}\tag{3}$$

当比较两个字符串S和T的相似度大小时，除了编辑距离LD的直观影响外，两字符串最长公共子序列LCS的大小也很重要，那么结合LD和LCS给出另一个相似度计算公式如下：

$$Sim(S,T)=LCS/(LCS+LD)\tag{4}$$

重新计算引言的情况一得：$S(S,T_1)=1/5$小于$S(S,T_2)=2/6$，能正确地表示合理的相似度。经实验证明此公式有良好的适用性且拓展了传统公式的普遍适用性，但引言中提到的情况二仍未得到解决，下面结合最长公共子串继续对相似度计算公式进行改进。

2 基于LD、LCS和LCCS的字符串相似度优化算法

2.1 基于列向量的编辑距离求解优化算法

传统编辑距离算法的空间、时间复杂度都是$O(m*n)$，考虑到每次计算$LD(i,j)$时只需用到$LD(i,j-1)$、$LD(i-1,j)$和$LD(i-1,j-1)$，即只需要当前列和前一列参加计算即可，因此本文利用两个列向量代替矩阵，每次只保存当前的状态和上次运算的状态，实验表明在基本不影响结果和时间复杂度的情况下，空间复杂度减少为$O(\min(m,n))$。

算法描述如下：输入：字符串S、T；输出：LD长度。

1. 设n是S的长度，m是T的长度(设$n<m$)，如果$m=0$，return n；如果$n=0$，return m；创建两个向量$\boldsymbol{v}_0[m+1]$和$\boldsymbol{v}1[m+1]$；

2. 初始化$\boldsymbol{v}_0[m+1]$的值为$0\cdots m$；

3. 检查S(i from 1 to n) 中的每个字符；

4. 检查T(j from 1 to m) 中的每个字符；

5. 设cost为编辑代价，if $s[i]==t[j]$，则$cost=0$；if $s[i]!=t[j]$，则$cost=1$；

6. 设置单元$\boldsymbol{v}1[j]$为下面的最小值之一：

7. 紧邻该单元上方+1：$\boldsymbol{v}_1[j-1]+1$

8. 紧邻该单元左侧+1：$\boldsymbol{v}_0[j]+1$

9. 该单元对角线左上角+$cost$：$\boldsymbol{v}_0[j-1]+cost$

10. 在完成迭代(3, 4, 5, 6)之后，$\boldsymbol{v}_1[m]$便是编辑距离的值。

2.2 基于降维的最长公共子序列算法优化

考虑到当计算$L[i,j]$时只和$L[i-1,j-1]$，$L[i,j-1]$，$L[i-1,j]$三个元素直接相关，即在计算$L[i]$这一行的子问题时，只需要知道$L[i-1]$一行的值和当前$L[i,j]$的前一个元素$L[i,j-1]$的值即可，此时，在时间复杂度不变的情况下，空间复杂度由$O(m*n)$降低为$O(n)$。定义一个长为$m+1$的数组$base[]$并初始化为0，一个递推前导$front=0$，一个当前计算元素pre(意义上的

$L[i,j]$），每次计算完$L[i,j]$后，把*front*的值更新到$base[j-1]$上，然后把*pre*更新到*front*上，然后继续向后扫描，注意最后一个元素的更新。这样迭代n次，最后$base[m]$的值即*LCS*的值。

算法描述如下：输入：字符串S、T，输出：*LCS*长度。

1. 输入两个字符串S，T；

2. 定义数组$base[\,]$大小为T.length()+1，定义递推前导*front*=0，当前元素*pre*=0；

3. 初始化$base[\,]$为0，$base[i]$=0；

4. 检查S(i from 1 to S.length())的每个字符；

5. 检查T(i from 1 to T.length())的每个字符；

6. 循环检查若$S[i-1]=T[j-1]$则$pre=base[j-1]+1$；否则pre=max(*front*, $base[j]$)；

7. 循环更新*base*和*front*，$base[j-1]$=*front*，*front*=*pre*；

8. 更新$base[T.\text{length}()]$=*front*，检查$base[j]$(j from 1 to T.length())并输出base[]值；

9. 循环(4,5,6,7,8)最后$base[T.\text{length}()]$就是*LCS*的值。

求解*LCCS*时，首先构造二维矩阵$L[m+1][n+1]$并初始化，若$S_i=T_j$，则$L[i][j]$=1；最后找到矩阵中最长的斜对角线上为1的字串就是*LCCS*，考虑到查找的麻烦，直接在矩阵需要填1时让他等于左上角值加1，这样矩阵中的最大元素就是*LCCS*的长度，*LCCS*的首字符位置就是第一个发生改变的单元格的坐标输出。

2.3　基于*LD*、*LCS*和*LCCS*的字符串相似度优化计算方法

当考虑连续的最长公共子串*LCCS*的影响力之后，引言中的情况二就有了解决办法，在*LD*和*LCS*都相同时，可以通过比较*LCCS*来判断相似度，计算得$LCCS(S,T_1)$=1小于$LCCS(S,T_2)$=3，可以判断出T_2和S的相似度更高，这也和人的主观判断方式相一致。设字符串S=“abcmg”，T_1=“abcnp”，T_2=“ebcmf”，现在发现这三个关键的影响因子大小都一样，但更接近人工主观判断的结果是S和T_1更相似，本文的解决方法是综合考虑*LCCS*的长度和出现位置，引入一个变量p表示目标串S中最长连续公共子串的首位置，当*LCS*长度和*LD*大小都一样时，*LCCS*值越大则相似度越大；当*LCCS*长度也一样时，考虑到前驱字符(当前处理字符前一个字符)比后继字符(当前处理字符后一个字符)对相似度的影响更大，则p越小说明*LCCS*首字符位置越靠前相似度越大，μ是一个平衡此因子的变量，可以根据对*LCCS*需求的严格程度人为设定。定义新的相似度计算公式如下：

$$Sim(S,T)=\frac{LCS}{LCS+LD+\frac{p}{LCCS}*\mu} \tag{5}$$

重新计算上面例子，设$\mu=1$，使用公式(5)计算$Sim(S,T_1)=0.6$，$Sim(S,T_2)=0.563$，得S和T_1的相似度大于S和$T2$的相似度，较之前结果更符合实际情况、区分性更好。

设字符串S和T长度分别是m和n，按照前文介绍的优化算法步骤求得编辑距离大小$LD(S,T)$和最长公共子序列长度$LCS(S,T)$，按照*LCCS*算法步骤求得最长公共子串长度$LCCS(S,T)$和子串首字符位置p，其中过滤掉了无意义的全符号比较，按照提出的相似度公式计算相似度，Java语言实现，算法流程如图1所示。

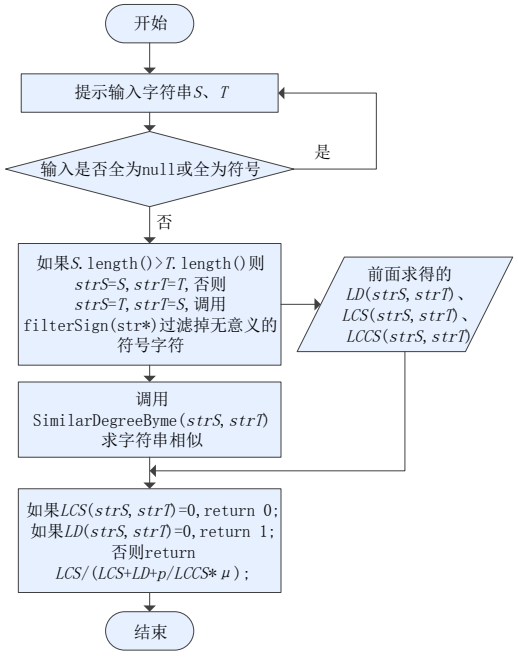


图1　优化算法的计算流程

3　实验分析及结论

实验用两组数据分析算法的准确度、适用性和区分性。数据一选取源串S=“expect”，目标串集合

$W=\{w_1,w_2,w_3,w_4,w_5,w_6,w_7\}=${“spectator”,“exercise”,“pecuniary”,“accept”,“excerpt”,“exempt”,“aspect”}。数据二源文本$S=$“abcde01234”，在英文词典中选出与word接近的一组单词构成目标串$T=\{t_0,t_1,t_2,t_3,t_4,t_5,t_6,t_7,t_8,t_9,t_{10},t_{11},t_{12},t_{13}\}=${“wqefghlm56234”，“e01abcd”，“fbcdm56789”，“fghlm51234”，“mgcde05678”，“w01234abc”，“mghln01234”，“abcde56789”，“fghde01234",“abcde01567”，“fgcde01234”，“abcde012fg”，“fbcde01234”,“abcde01235”}。

3.1 相似度结果分析

分别采用公式(2)、公式(4)和公式(5)三个相似度计算公式求解相似度，对两组数据的字符串相似度计算结果进行对比分析，另外为了更详细地描述实验结论，便于比较说明，给出第一组的对比数据如图2所示。

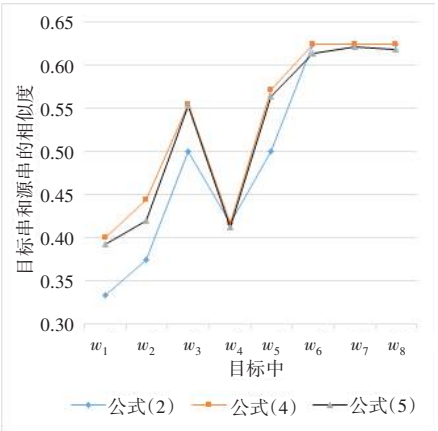


图2 数据集W的相似度比较

3.2 实验结论

可以看出，在对目标串集合w_1～w_8，t_1～t_{13}的考察中，公式(4)、公式(5)的结果较公式(2)更加合理精确，避免了由字符顺序变化导致的相似度极度不合理的情况，公式(5)求解的相似度和公式(4)走势大致一样，基本介于公式(2)和公式(4)的结果之间，说明了公式(5)的适用性和稳定性，当公式(2)、公式(4)两条线有重合持平时，即这两种方法无法区分相似度时，公式(5)的线条有上升或下降的变化，可以得到区分性更好的结果。

由表1的第3、5组数据可以看出，当LD相同时公式(4)、公式(5)较公式(2)有更合理、更具区分性的结果，由表第6、7、8组数据看出当LCS和LD都相同时，公式(5)较公式(2)和(4)得出了更合理且有区分性的相似度数据，数据因$LCCS$的长度和位置不同而变化，公式(5)相对于公式(2)、公式(4)有更适中的标准差和极差，使相似度结果更合理。

表1 数据对比

组	目标串	公式(2)	公式(4)	公式(5)
1	spectator	0.333	0.400	0.393
2	exercise	0.375	0.444	0.420
3	expceted	0.500	0.555	0.553
4	bespectacled	0.417	0.417	0.412
5	exempt	0.500	0.571	0.564
6	earpecte	0.625	0.625	0.614
7	expedite	0.625	0.625	0.622
8	expdctse	0.625	0.625	0.618
	极差	0.292	0.225	0.226
	标准差	0.117885	0.097	0.095

综上，可以看出公式(5)具有良好的普遍适应性，且比公式(2)和公式(4)结果更精确合理，结果具有更好的区分性。

4 结语

按照提出问题、分析问题和解决问题的思路对字符串相似度的计算进一步改善。考虑到LCS和$LCCS$对计算字符串相似度的重要影响，基于编辑距离的相似度算法进行改进，改进的字符串相似度算法提高了算法的普遍适用性和结果精确性，对于差异性很小的字符串也可以得到较好区分性的相似度。另外，对LD和LCS的传统计算过程在数据结构方面进行了优化，在时间复杂度基本不受影响的情况下进一步降低了算法的空间复杂度。

参考文献：

[1] 陈俊月,郝文宁,张紫萱,等.基于改进句子相似度算法的释义识别研究[J].计算机工程,2020,46(9):76-82.

[2] 藏润强,孙红光,杨凤芹,等.基于Levenshtein和TFRSF的文本相似度计算方法[J].计算机与现代化,2018(4):84-89.

(下转第32页)

图形图像

文章编号: 1007-1423(2023)14-0027-06 DOI: 10.3969/j.issn.1007-1423.2023.14.006

基于CHFM特征的约束图像拼接检测和定位

杨 海[1]，蔺 聪[2,3*]

（1. 广东财经大学信息学院，广州 510320；2. 广东财经大学统计与数学学院大数据与教育统计应用实验室，广州 510320；3. 中山大学广东省信息安全技术重点实验室，广州 510006）

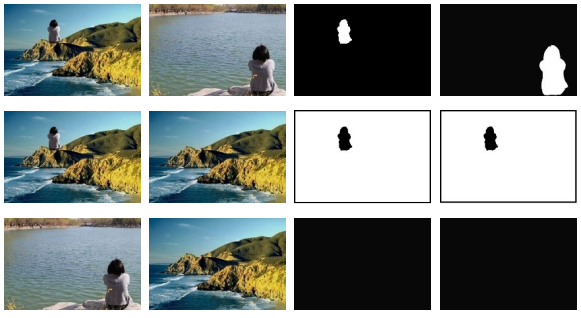
摘要： 提出了一种基于切比雪夫-傅里叶矩(CHFM)的约束图像拼接检测和定位方法。首先提取图像的CHFM特征，接着用PatchMatch算法对提取的特征进行匹配，最后进行后处理，得到掩模。在CASIA v2.0数据集上证明了本方法的有效性。

关键词： 数字取证；篡改检测；CISDL；PatchMatch；CHFM

0 引言

随着计算机技术的不断发展，新手只需经过简单的学习就可借助图像编辑软件对图像进行篡改。篡改后的图像很难用肉眼辨别其真伪，使图像的真实性和完整性受到威胁[1]。因此，迫切地需要可靠且全面的篡改检测技术来识别伪造图像。

图像伪造中，常用的篡改操作有：复制移动、修复和拼接。图像拼接伪造是复制一幅图像的部分区域，然后粘贴到另一幅图像中并合并成一幅新图像[2]。在2017年美国国家标准与技术研究所的Nimble挑战中[3]，图像拼接问题被重新表述为：给定一个查询(探针)图像Q和一个潜在供体图像P，其目标除了检测Q是否包含来自P中的区域，还要定位出Q的拼接区域和P提供给Q的区域。这将图像拼接检测和定位的研究约束在一对图像上，因此Wu等[4]将其定义为约束图像的拼接检测和定位(constrained image splicing detection and location，CISDL)。CISDL问题所用输入图像如图1所示，图中第三列Q_m中白色表示拼接区域，黑色为真实区域，第四列P_m中白色区域为提供给Q的区域。图中前两对图像的Q中包含了来自P的区域，为篡改图像，第三对图像为真实图像。



(a)查询图像Q (b)供体图像P (c)Q掩模图像Q_m (d)P掩模图像P_m

图1 CISDL问题所用图像

针对CISDL问题，本文提出一种基于切比雪夫-傅里叶矩(Chebyshev-Fourier Moments，

收稿日期： 2023-05-21 **修稿日期：** 2023-07-02

基金项目： 教育部产学合作协同育人项目(220605769202709)；广东省信息安全技术重点实验室开放基金(2020B1212060078)；广东省普通高校特色创新项目(自然科学)(2022KTSCX041)；广州市科技计划基础与应用基础研究项目(202102080316)；广州市海珠区科技计划项目(海科工商信计2022-45)；广东财经大学一流本科教学质量与教学改革工程项目(粤财大〔2022〕132号)：面向互联网+的《非结构化数据挖掘》混合式教学改革探索；广东财经大学大学生创新创业项目培育双百工程(2022XSZD103，2022XSYB380，2022XSYB382)：基于深度学习的图像复制粘贴篡改检测研究

作者简介： 杨海(2000—)，男，河南南阳人，在读硕士研究生，研究方向为图像取证；*通信作者：蔺聪(1979—)，男，河北邯郸人，博士，讲师，硕导，研究方向为信息安全、数字取证，E-mail:lotuslin2005@126.com

CHFM)特征的检测方法。本方法首先用CHFM矩[5]进行特征提取；其次用PatchMatch[6]算法进行特征匹配，得到各像素对应匹配像素的坐标索引；最后将上步得到的索引矩阵进行后处理，得到输入图像的掩模图像。

1 相关工作

关于CISDL问题，Wu等[4]提出了深度匹配和验证网络(deep matching and validation network，DMVN)，DMVN网络中有特征提取、深度密集匹配、掩模反卷积和视觉一致性验证器四个模块。Ye等[7]继承了DMVN中的深度密集匹配，提出了特征金字塔深度匹配和定位网络(feature pyramid deep matching and localization network，FPLN)。FPLN在深度密集匹配后添加了融合层，简化了DMVN网络结构，可以检测和定位到小的拼接区域，但空间信息丢失限制了模型的辨别能力和定位精度。Liu等[8]提出一种新的对抗性学习框架。该框架包含基于空洞卷积的深度匹配网络(deep matching network based on atrous convolution，DMAC)、一个检测网络和一个判别网络。Liu等[9]又提出一种新的注意感知编解码器的深度匹配网络，称为AttentionDM。Xu等[10]提出一种新的尺度自适应深度匹配网络(scale-adaptive deep matching network，SADM)，SADM的金字塔框架捕获了多尺度细节，提高了多尺度区域和边界检测与定位的精度。

2 提出的方法

本节分为三个部分，来详细介绍提出方法的各个步骤，其流程如图2所示。第一部分介绍特征提取的相关知识；第二部分介绍PatchMatch算法进行特征匹配的过程；第三部分是后处理，生成二值图像的过程。

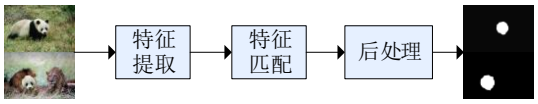


图2 提出方法的流程

2.1 特征提取

首先介绍一下图像矩。在数学上，图像矩一般定义为图像函数f和基函数V_{nm}的内积$\langle f, V_{nm}\rangle$[11]，公式为

$$\langle f, V_{nm}\rangle = \iint_D V_{nm}^*(x,y) f(x,y)\mathrm{d}x\mathrm{d}y \tag{1}$$

其中：$*$表示复共轭；$f(x,y)$是图像函数，其取值为$(x,y)\in R^2$。CHFM矩是圆形矩，需将图像矩用极坐标形式进行表示，因此式(1)可表示为

$$F_I(n,m) = \int_0^{2\pi}\int_0^{\infty} f(\rho,\theta) V_{nm}^*(\rho,\theta)\rho\mathrm{d}\rho\mathrm{d}\theta \tag{2}$$

式(2)中$F_I(n,m)$表示图像矩集$\langle f, V_{nm}\rangle$，$f(\rho,\theta)$是图像函数$f(x,y)$的极坐标表示，其中$\rho\in[0,\infty]$，$\theta\in[0,2\pi]$。式(2)中基函数$V_{nm}(\rho,\theta)$形式如下：

$$V_{nm}(\rho,\theta) = R_{nm}(\rho)\frac{1}{\sqrt{2\pi}}\mathrm{e}^{\mathrm{j}m\theta} \tag{3}$$

其中：$R_{nm}(\rho)$为径向基函数，可以是任意取值，j为虚数单位，取值为$\mathrm{j}=\sqrt{-1}$。

将式(3)代入式(2)中，式(2)可重新表示为

$$F_I(n,m)=\int_0^{\infty}\rho R_{nm}^*(\rho)\times\left[\frac{1}{\sqrt{2\pi}}\int_0^{2\pi} f(\rho,\theta)\,\mathrm{e}^{-\mathrm{j}m\theta}\mathrm{d}\theta\right]\mathrm{d}\rho \tag{4}$$

本方法中，选用的径向基函数为CHFM，其形式如下：

$$R_n^{(\mathrm{CHFM})}(r) = \frac{2}{\pi}\left(\frac{1-r}{r}\right)^{\frac{1}{4}}\sum_{k=0}^{\left\lfloor\frac{n}{2}\right\rfloor}\frac{(-1)^k(n-k)!(4r-2)^{n-2k}}{k!(n-2k)!} \tag{5}$$

以上是图像矩的定义及形式，接下来用这些理论定义描述图像特征。图像中各像素的特征$f(a)$是由其$F_{I(a)}(n,m)$值组成的集合，因此特征提取转变为求解各像素的$F_{I(a)}(n,m)$值。计算$F_{I(a)}(n,m)$值时，首先确定(n,m)对数量和块的大小。本方法选取16对(n,m)，即特征长度为16。块大小即滤波器大小为16×16。

其次，对式(2)进行近似，求解图像矩集。其近似形式为

$$F_I'(n,m) = \sum_{(x,y)\in W} f(x,y) V_{nm}^*\big(\rho(x,y),\theta(x,y)\big) \tag{6}$$

式(6)中的$\rho(x,y)=\sqrt{x^2+y^2}$，$\theta(x,y)=\pm\arctan(y/x)$。图像矩的近似可看成图像和滤波器的滤波。首先要求解$\rho(x,y)$和$\theta(x,y)$，然后再求解V_{nm}，不同的(n,m)能够得到不同的V_{nm}，所以最终的V_{nm}

是三维16×16×16大小的。随后将图像函数和求得的V_{nm}的复共轭进行滤波，最终得到整幅图像所有像素$F_{I(a)}(n,m)$值组成的集合，即图像的CHFM特征。

2.2　特征匹配

特征匹配前首先对特征进行归一化，其公式为

$$Y=\frac{X-MinValue}{MaxValue-MinValue} \tag{7}$$

式(7)中X是需归一化的像素对应的像素值，$MinValue$是需归一化的所有像素中像素值的最小值，$MaxValue$是需归一化的所有像素中像素值的最大值。

归一化后使用PatchMatch算法进行特征匹配。其匹配步骤如下：

2.2.1　初始化

首先介绍一下偏移量。偏移量值是二维的，其值记录的是特征向量距离最小的两个块中心像素坐标之间的横向和纵向差值。用公式表示为

$$\delta(a)=\arg\min_{\Phi:a+\Phi\in\Omega,\Phi\neq 0} D\big(f(a),f(a+\Phi)\big) \tag{8}$$

式中：Ω表示图像所有像素，$a+\Phi$是除像素a外的像素。$f(a)$表示以a为中心P像素邻域的块对应的特征向量，D表示两个块对应特征向量之间的距离。D取最小值时，两个像素之间的坐标差值，即为a到其最近邻的偏移量。a对应最近邻像素(即a对应匹配点)a'的坐标用公式表示为

$$a'=a+\delta(a) \tag{9}$$

初始化是为图像中每个像素随机赋予一个偏移量，使每个像素都能够在当前图像中找到一个像素与之匹配，其匹配点坐标由公式(9)得到。由于特征匹配是寻找与目标点相对较远的匹配项，所以本阶段提前排除小于阈值8的偏移量。

2.2.2　迭代

本步骤是将初始化阶段好的偏移量传播和更新到整幅图像，需扫描整幅图像，在扫描时，每扫描一个像素执行传播和随机搜索，当扫描完整幅图像即一次迭代。其迭代过程如下：

(1) 传播

$$\delta(a)=\arg\min_{\Phi\in\Delta^{P}(a)} D\big(f(a),f(a+\Phi)\big) \tag{10}$$

奇数次迭代时Φ取值为$\Delta^{P}(a)=\{\delta(a),\delta(a^{r}),\delta(a^{c})\}$，分别为当前像素的偏移量及其左相邻和上相邻像素的偏移量，扫描顺序如图3所示，从上至下，从左至右逐个进行扫描。首先扫描图像左上角第一个像素点，分别计算以该像素为中心的块与以该像素加上三个偏移量后得到的三个不同的像素为中心的块的特征向量之间的距离。取三个距离中的最小值，计算对应偏移量，接着执行随机搜索，更新当前像素的偏移量。然后往右扫描第二个像素，执行相同操作。当扫描完一行像素时，换行仍从左到右逐个像素进行扫描。当扫描到图像右下角最后一个像素时，扫描结束即完成一次迭代。在奇数次迭代时，好的偏移量能够传播到当前像素右面及下面各个像素。

偶数次迭代时Φ取值分别为当前像素的偏移量及其右相邻和下相邻像素的偏移量。偶数次迭代过程中，扫描是从下至上，从右至左进行扫描。在偶数次迭代时，好的偏移量能够传播到当前像素左面及上面各个像素。

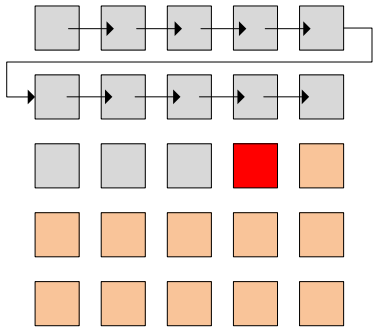


图3　奇数次迭代扫描顺序

(2) 随机搜索

传播得到的偏移量是在相邻像素中寻求的最优值，可能会出现在整个区域内不是最优的情况，所以需进行随机搜索，防止其陷入局部最优。

随机搜索时像素a偏移量取值公式为

$$\delta_i(a)=\delta(a)+R_i \tag{12}$$

其中：$\delta(a)$是传播后的偏移量，R_i是边长为2^{i-1}($i=1,2,\cdots,L$)的方形网格中，服从均匀分布的二维随机变量，公式(10)中Φ取值为

$$\Delta^{R}(a)=\{\delta(a),\delta_1(a),\cdots,\delta_L(a)\} \tag{13}$$

本步骤首先以当前像素的最近邻像素为中心，在初始搜索半径为图像最长边的区域中随机寻找一个像素，计算以当前像素为中心的块与以该像素最近邻像素为中心的块特征向量之间的距离，与以随机搜索到的像素为中心的块特征向量之间的距离，比较两距离值。如果与以最近邻像素为中心的块特征向量的距离最小，将搜索区域半径缩小为原来一半，继续搜索。相反，更新当前像素偏移量，并以随机搜索的像素为中心，在搜索半径为原来一半的区域中继续进行搜索。当搜索半径缩小为1时，当前像素的偏移量更新完成。

通过匹配，最终得到两个分别记录各像素匹配像素的x索引和y索引的矩阵。

2.3 后处理

理想情况下，匹配后的偏移量大多是混浊的，需后处理来规范和约束偏移量，以减少误检的概率。后处理步骤如下：

(1) 用半径为4的圆形域对偏移量进行中值滤波。首先生成半径为4的圆形域，随后用其对匹配后得到的索引矩阵进行顺序统计中值滤波，得到两个新的索引矩阵。

(2) 在半径为6的圆形域中计算密集线性拟合(dense linear fitting，DLF)误差。DLF是用线性仿射模型拟合出以$\boldsymbol{a}$为中心N像素邻域的偏移量，使匹配后的偏移量与经过仿射后的偏移量误差平方和最小化：

$$\varepsilon^2(\boldsymbol{a}) = (\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\delta}) - (\boldsymbol{\delta}^{\mathrm{T}}v_1)^2 - (\boldsymbol{\delta}^{\mathrm{T}}v_2)^2 - (\boldsymbol{\delta}^{\mathrm{T}}v_3)^2 \tag{14}$$

其中：$\boldsymbol{\delta} = \left[\delta(a_1), \delta(a_2), \cdots, \delta(a_N)\right]^{\mathrm{T}}$为特征匹配后$\boldsymbol{a}$的$N$像素邻域中各像素偏移量组成的向量，表$V_i$示长度为$N$的列向量。按照公式(14)的形式，用半径为6的圆形域对上步得到的两个矩阵进行滤波，计算各像素两个方向所对应的线性拟合误差平方和，并将其相加，与阈值300进行比较，最终得到一幅二值图像。

(3) 去除偏移对间距离小于50的像素。本操作需计算各像素与对应最近邻像素欧式距离的平方，与阈值50×50进行比较，会得到一幅二值图像，并与上步的二值图像取交集，最终得到去除偏移对间距离小于阈值的二值图像。

(4) 形态学处理。首先去除二值图像中面积小于600像素的区域；其次进行镜像检测，寻找二值图像中值为1的像素，将其对应像素标记为1，并再次剔除面积小于600像素的区域；随后对二值图像进行膨胀操作。特征生成后，去除了特征图边缘，需填充二值图像边缘，使其与输入图像大小相同。

3 实验

本实验选用数据集是CASIA v2.0[12]，图像大小从320×240像素到800×600像素不等，其中包含7200张真实图像和5123张篡改图像。篡改的5123张图像中，有3302张是复制移动问题的图像，有1821张是拼接检测问题的图像。本实验根据Wu等[4]上传到gitlab上的CASIA配对表格，随机选取100对标签为1的匹配对，将其对应图像拼接，并用Photoshop生成各对图像的掩模图像。随后，对合成的图像及掩模图像进行后处理，将大小都设置为384×512像素。

为了综合评估本方法的性能，使用评价指标F_1来度量，F_1公式如下：

$$F_1 = \frac{2TP}{2TP + FN + FP} \tag{15}$$

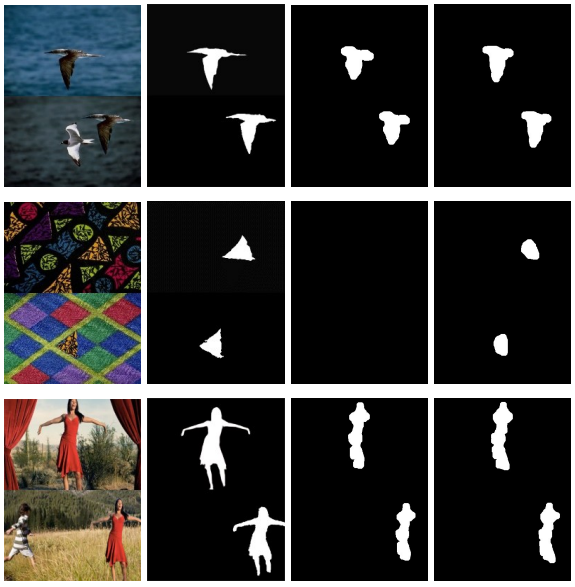
其中：图像级F_1中的TP(真正例)为检测到的伪造图像的数量、FN(假反例)为未检测到的伪造图像数量、FP(假正例)为错误检测到的真实图像的数量。像素级F_1中TP为检测到伪造像素的数量，FN为未检测到的伪造像素的数量，FP为误检测到真实像素的数量。在图像级上，F_1的值表示检测方法识别图像真伪的能力。在像素级上，F_1的值表示检测方法检测和定位图像伪造区域的精度。

为了检测本方法的性能，将本文方法与Cozzolino等[13]提出方法进行比较。图4展示了三组图像的掩模输出结果。从(c)和(d)中第一行和第三行可以看出两种方法检测到的篡改区域大小差不多，从第二行了解到本文方法检测到Cozzolino等[13]方法没有检测到的篡改区域。。

对比实验结果见表1，表1中报告了两种方法图像级和像素级的F_1值。通过表格了解到本方法图像级，特别是像素级的F_1优于Cozzolino等[13]方法。

表 1　对比实验结果/%

方法	图像级 F_1	像素级 F_1
Cozzolino[13]	87.64	70.88
本文方法	88.27	72.10



（a）伪造图像　（b）真实掩模　（c）Cozzolino　（d）本文方法

图 4　两种方法检测结果对比

4　结语

图像编辑软件不断发展，篡改后的图像效果逼真，很难辨别，给图像伪造检测带来极大的挑战。基于此，本文提出一种基于CHFM特征的方法解决CISDL问题，该方法首先提取CHFM特征，其次用PatchMatch算法进行特征匹配，最后进行后处理，输出二值图像，以此了解输入图像对中的重复区域。

实验结果表明，本文方法检测和定位精度与Cozzolino等[13]方法相比有了提高。但本文方法还有改造的空间，比如特征匹配方法和后处理阶段。在未来的工作中，将对这两个步骤的改进进行研究。

参考文献：

[1] 张怡暄，赵险峰，曹纭. 数字图像篡改盲检测综述[J]. 信息安全学报，2022，7(3)：56-90.

[2] BI X，WEI Y，XIAO B，et al. RRU-Net：the ringed residual U-Net for image splicing forgery detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPR)，Long Beach，CA，USA，2019：30-39.

[3] The National Institute of Standards and Technology (NIST). Nimble challenge 2017 evaluation[R/OL]. https://www. nist. gov/itl/iad/mig/nimble-challenge-2017-evaluation.

[4] WU Y，ABD-ALMAGEED W，NATARAJAN P. Deep matching and validation network：an end-to-end solution to constrained image splicing localization and detection[C]//Proceedings of the ACM International Conference on Multimedia，Mountain View，CA，USA，2017：1480-1502.

[5] PING Z L，WU R G，SHENG Y L. Image description with Chebyshev-Fourier moments[J]. Journal of the Optical Society of America A，2002，19(9)：1748-1754.

[6] BARNES C，SHECHTMAN E，FINKELSTEIN A，et al. PatchMatch：a randomized correspondence algorithm for structural image editing[J]. ACM Transactions on Graphics，2009，28(3)：24-35.

[7] YE K，DONG J，WANG W，et al. Feature pyramid deep matching and localization network for image forensics[C]//Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference(APSIPA ASC)，Honolulu，HI，USA，2018：1796-1802.

[8] LIU Y，ZHU X，ZHAO X，et al. Adversarial learning for constrained image splicing detection and localization based on atrous convolution[J]. IEEE Transactions on Information Forensics and Security，2019，14(10)：2551-2566.

[9] LIU Y，ZHAO X. Constrained image splicing detection and localization with attention-aware encoder-decoder and atrous convolution[J]. IEEE Access，2020，8：6729-6741.

[10] XU S，LV S，LIU Y，et al. Scale-adaptive deep matching network for constrained image splicing detection and localization[J]. Applied Sciences，2022，12(13)：6480.

[11] FLUSSER J，ZITOVA B，SUK T. Moments and moment invariants in pattern recognition[M]. United Kingdom：John Wiley & Sons，2009.

[12] WANG W，DONG J，TAN T. Image tampering detection based on stationary distribution of Markov chain[C]//Proceedings of the IEEE International Conference on Image Processing，Hong Kong，China，2010：2101-2104.

[13] COZZOLINO D，POGGI G，VERDOLIVA L. Efficient dense-field copy-move forgery detection[J]. IEEE Transactions on Information Forensics and Security，2015，10(11)：2284-2297.

Constrained image splicing detection and location based on CHFM features

Yang Hai[1], Lin Cong[2,3*]

(1.School of Information, Guangdong University of Finance and Economics, Guangzhou 510320, China;
2.Applied Laboratory of Dig Data and Education Statistics, School of Statistics and Mathematics, Guangdong University of Finance and Economics, Guangzhou 510320, China;
3. Guangdong Provincial Key Laboratory of Information Security Technology, Sun Yat-Sen University, Guangzhou 510006, China)

Abstract: In this paper, a constrained image splicing detection and localization method based on Chebyshev Fourier Moments (CHFM) is proposed. First, the CHFM features are extracted from the image, then the PatchMatch algorithm is adopted to match the CHFM features, finally the post-processing is performed to obtain the mask. The effectiveness of the proposed method was demonstrated on the CASIA v2.0 dataset.

Keywords: digital forensics; tamper detection; CISDL; PatchMatch; CHFM

(上接第26页)

Research on string similar degree algorithm based on Levenshtein distance

Zhang Shengnan[*]

(Information Engineering College of Hubei Light Industry Technology Institute, Wuhan 430000, China)

Abstract: The algorithm for solving string similarity based on Levenshtein Distance is very classic, but it has some shortcomings in terms of universality and accuracy. It is improved based on Longest Common Substring(LCCS) and Longest Common Subsequence(LCS) to make the calculation results more discriminative, universal, and accurate. In addition, when calculating similarity, the algorithm for solving LD and LCS is optimized from the perspective of data structure, reducing the spatial complexity of the algorithm in an order of magnitude. The experimental results were compared and analyzed to prove their feasibility and correctness.

Keywords: similarity calculation; levenshtein distance; LCS; LCCS

文章编号: 1007-1423（2023）14-0033-05 DOI: 10.3969/j.issn.1007-1423.2023.14.007

基于图像处理技术对植物叶片面积测量

郝晓峰*，于蓉蓉

（西京学院计算机学院，西安 710200）

摘要： 针对叶片面积测量，设计并实现一种基于图像处理对叶片面积进行测量的方法。利用CDD摄像机获取植物叶片图像，对采集的植物叶片图像进行图像分割和边缘检测，通过对比Sobel、Robert、Prewitt和Canny算子选取理想图片，再利用形态学处理对缝隙和噪音进行平滑处理，最后通过像素统计法获得测量面积，与真实面积进行相关性分析。该方法可以准确、便捷地测量植物叶片的面积。

关键词： 图像处理；边缘检测；面积；像素

0 引言

植物的光合作用影响着植物的作物产量，光合作用和植物叶片有关系，叶面积对植物的产物影响较大，叶面积与植物作物是成正比关系，因此测量叶面积对植物作物有非常重要的影响[1]。如何提高速度和准确无误地测量植物叶面积对农业发展具有重要的意义。在过去的几年里，已经提出了许多测量方法，如叶面积仪测定法、方格法、剪纸称质量法。韩殿元等[2]提出用颜色通道相似与自适应阈值分割计算植物叶面积和标准参照物的像素数目，然后通过两者对比计算植物叶面积。于东玉等[3]提出双边滤波和拉普拉斯算子首先对图形预处理，图像分割用分水岭算法，最后用比例法测量出植物叶片面积。

本研究采用相对参考物法[4]，通过CCD摄像机采集图像，用加权平均法进行灰度化，用最小差分法进行图像分割，再通过调节对比度使模糊不清的原始图像富含大量有用信息，边缘检测通过Sobel、Robert、Prewitt和Canny算子方法比较选用效果较好的Canny算子，通过形态学处理对空隙填充，再通过中值滤波处理去除噪音获取理想图片，最后通过经典算法像素统计法测量植物叶片的面积[5]。

1 实验方法

1.1 图像采集

利用CCD摄像机对目标进行图像采集，先将植物叶放在有参照矩阵和白色背景的纸板中；然后将CDD摄像机镜头与植物叶片垂直，并调节CDD摄像机的参数，保证拍摄的图像清晰，采集植物叶片的原始图像，如图1所示。

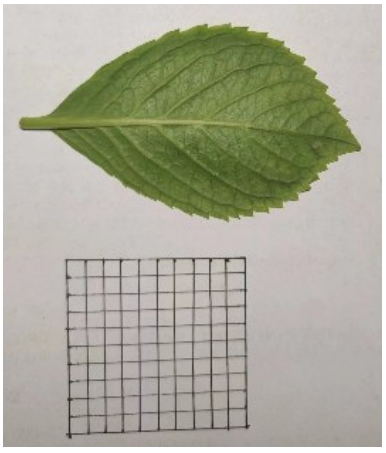


图 1 采集的原始图像

收稿日期： 2023-04-11 **修稿日期：** 2023-04-23

基金项目： 西京学院研究生横向课题项目（2023610002001334）：农业种植智能公共数据服务系统的设计与开发

作者简介： *通信作者：郝晓峰（1998—），男，山东菏泽人，在读硕士，研究方向为图像处理，E-mail：228974216@qq.com；于蓉蓉（1985—），女，河南南阳人，硕士，副教授，讲师，研究方向为非线性动力学与同步控制系统理论、混沌控制理论、复杂动力系统数值计算分析、神经元系统的动力学分析

图像采集完成后，需要对图像进行灰度变换，灰度变换可以提高图像的丰富度，突出植物叶片的主要特征，彩色图像转化灰度图像从而降低存储空间并提高图像的计算速度。该研究的整体流程如图2所示。

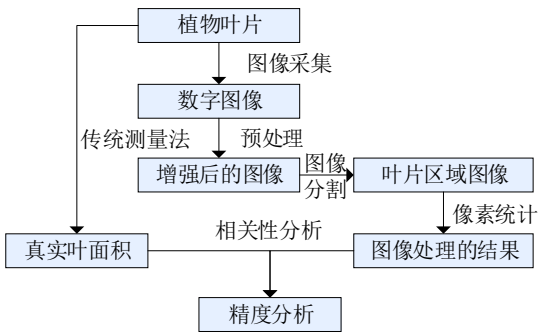


图2 整体流程

1.2 预处理

由于CDD摄像机拍摄的是彩色图像，彩色图像存储空间大，用图像处理机器计算速度较慢，为了进行下一步处理，需要先将图像转换为灰度图像。灰度变换是图像增强的一种手段，使图形动态范围加大，对比度在色素扩展，图形更清晰，特征更明显且灰色图像存储空间较小，同时可以提高图像数据处理速度。在色彩学理论中分为RGB颜色空间、HSV颜色空间、HSI颜色空间，其中RGB颜色空间通常用于显示器系统，利用物理学中的三原色彩叠加原理，可以产生各种颜色，因此本次研究选择RGB颜色空间。目前灰度化的常用方法是最大值法、平均值法、加权平均值法。本次研究选用的是加权平均法，由于人的视觉系统对RGB分量的敏感度不同，R、G、B的分量选取比例分别为0.30、0.59、0.11，该比例是较适合人眼视觉，加权平均法的公式为

$$f(i,j) = 0.30R(i,j) + 0.59G(i,j) + 0.11B(i,j) \tag{1}$$

其中：$f(i,j)$代表(i,j)处的灰度值，$R(i,j)$代表彩色图像的红色分量、$G(i,j)$代表彩色图像的绿色分量、$B(i,j)$代表彩色图像的蓝色分量。

1.3 图像分割

图像预处理后，对植物叶片图像背景与目标进行分割，图像分割适用于目标和背景占据不同灰度级范围的图像，既可以减少数据量又可以简化分析和处理[8]，为了突出目标和背景不同灰度级，本次研究使用白色背景板。对植物叶片进行图像分割时，为分割范围适用更广，尤其是室外破坏性质的采摘测量，目前图像分割常用的办法有极小点阈值法、最小均方误差法、双峰法、最大类间方差法。本研究采用最小均方误差法。具体步骤如下：令z表示灰度值，$p(z)$表示灰度值概率密度函数的估计值，则描述图像中整体灰度变换的混合密度函数是：

$$\begin{aligned} p(z) &= P_1 p_1(z) + P_2 p_2(z) \\ &= \frac{P_1}{\sqrt{2\pi}\,\sigma_1}\exp\left[-\frac{(z-\mu_1)^2}{2\sigma_1^{\,2}}\right] + \\ &\quad \frac{P_2}{\sqrt{2\pi}\,\sigma_2}\exp\left[-\frac{(z-\mu_2)^2}{2\sigma_2^{\,2}}\right] \end{aligned} \tag{2}$$

其中：μ_1和μ_2是背景和前景的平均灰度值；σ_1和σ_2分别是关于均值的均方差；$p_1(z)$与$p_2(z)$表示灰度值概率密度函数的估计值；P_1与P_2是前景和背景具有参数z的像素概率。如果$\mu_1 < \mu_2$，需要定义一个阈值T，小于T的灰度值的像素是背景，大于T的像素是目标。这样，把目标像素划错的概率是：

$$E_1(T) = \int_{-\infty}^{T} p_1(z)\mathrm{d}z \tag{3}$$

$$E_2(T) = \int_{-\infty}^{T} p_2(z)\mathrm{d}z \tag{4}$$

$$E(T) = p_2 E_1(T) + p_1 E_2(T) \tag{5}$$

其中：$E_1(T)$代表前景划错的概率；$E_2(T)$代表背景划错的概率；$E(T)$代表总的误差概率。

为求得误差最小的阈值可将$E(T)$对T求导并令导数为零得：

$$P_1 p_1(z) = P_2 p_2(z) \tag{6}$$

$$T = \frac{\mu_1 + \mu_2}{2} + \frac{\sigma^2}{\mu_1 - \mu_2}\ln\left(\frac{p_2}{p_1}\right) \tag{7}$$

若$p_1 = p_2 = 0.5$，最佳阈值是均值的平均数：

$$T = \frac{\mu_1 + \mu_2}{2} \tag{8}$$

调节对比度可以将灰度化模糊不清甚至无法分辨的原始图像处理成清晰的富含大量有用

信息的图像，通过调节对比度，可以有效去除图像中的噪音、增强图像中的边缘或其他需要的区域，从而更加容易对图像中需要的目标进行检测和测量。调节对比度前后效果分别如图3和如图4所示。

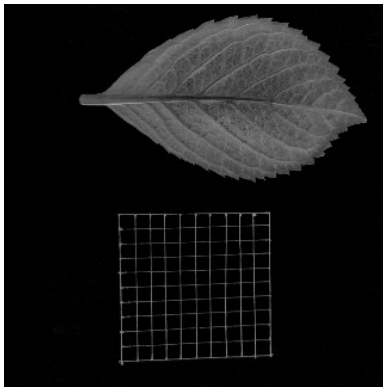


图 3　调节对比度之前的图片

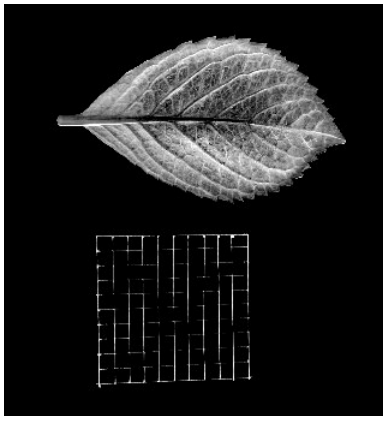


图 4　调节对比度之后的图片

1.4　边缘检测

边缘是目标区域的一部分，且承载了图像中大部分语义和含有大量的形状信息，边缘也是图像从彩色图像转变成灰色图像时发生突变的地方，同时也是图像两个灰度值相似像素点的集合，边缘检测是通过寻找边缘像素获取物体的轮廓，轮廓中保存图像中重要的信息，边缘检测的提取基本思想是检测不同像素点的灰度变化[8]，算子与图像进行卷积运算，对高频信号响应应该过零点。因为噪音在图像中是高频信号，所以边缘检测会对噪音产生呼应。不同算子对不同图像的噪音敏感度不同，去噪的实际效果不同。对于离散的数字图像进行图像处理时，存在一阶差分和二阶差分边缘检测算子。一阶边缘检测算子有Prewitt算子、Sobel算子、Roberts算子等，二阶边缘算子有拉普拉斯算子、高斯-拉普拉斯算子、高斯差分算子等，目前边缘检测常用的方法是Roberts算子、Laplace算子、Sobel算子、Prewitt算子、Canny算子，本次研究通过对比选用效果较好的算子。该五种算子得到的图像如图5所示。

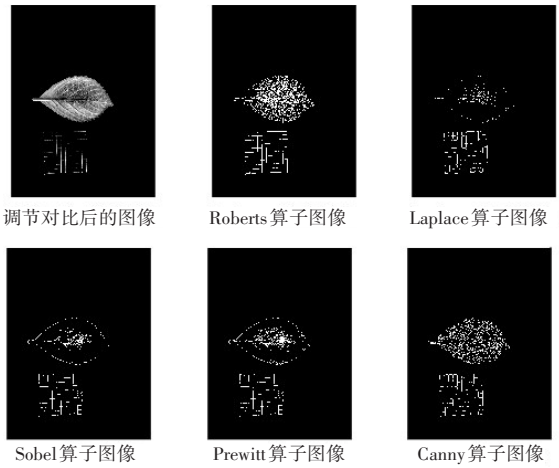


图 5　五种算子得到的图像

通过对比可以明显看出，Canny算子效果较好，符合实验要求，本研究选用Canny边缘检测[5]。叶片的边缘与内部存在许多的缝隙，需要对该结果进行进一步的优化处理，对于植物叶片内部缝隙采用形态学处理，对缝隙进行填充，由于拍摄时有外部因素存在，会在植物叶片图像中产生噪音，可以通过形态学处理来解决噪音这一问题，它不仅可以获得清晰的图像，也可以增强图像的频率。针对噪音通过中值滤波消除孤立的噪音点，从而获取到理想的植物叶片二值化图像[6]。中值滤波是非线性平滑滤波技术，它最关键的地方在于中值两字，通过排序选择的一种思想来消除独立的噪音，第一步是图像某一个邻域内的像素值用来排序，再用这一个领域排序获得的中值来代替需要处理的像素，该像素是这一邻域内最接近真实值的像素。第二步定义一个长度为L且长度L为奇数的窗口，因为L是奇数的情况下中值滤波器的识别效果较好；反之L为偶数时识别效果较差，所谓L一般等于$2N+1$，N为正整数，公式为

$$Y(i) = Med\left[X(i-N), \cdots, X(i), \cdots, X(i+N)\right], i \in Z \tag{9}$$

其中：Med表示中值，$Y(i)$代表用中值滤波器输出的像素中值，$X(i)$代表邻域中的不同像素值。经过形态学和去噪处理获得的最终图像如图6所示。

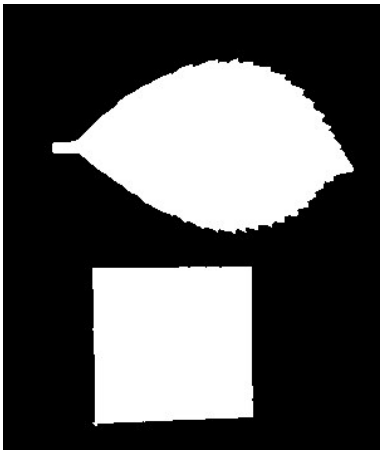


图6 形态学处理后的图像

1.5 标记与面积计算

通过像素统计计算叶面积[8]。设c_1和c_2分别表示叶面积和正方形面积。设p_1和p_2分别表示图像中植物叶子的像素数和图像中正方形内的像素数。因此计算叶面积的公式为

$$c_1 = c_2 \times \frac{p_1}{p_2} \tag{10}$$

其中：p_2是已知，p_1可以通过统计填充图像中灰度值为255的像素点的数量进行获取，因此可以获取c_1，再通过与传统测量法获取真实面积，与基于图像处理获得的叶片面积进行相关性分析，结果见表1。

表1 相关性分析

植物叶片	数据			
	植物像素数量	c_1/cm^2	实际面积/cm^2	测量误差
1	180609	33.5811	30.4377	3.1434
2	208680	38.8005	29.7374	3.0631
3	186830	34.7378	31.9537	2.7841
4	164485	30.5832	33.8613	3.2781
5	170219	31.6494	28.7852	2.8742

2 结语

本文以植物叶片作为研究对象，基于图像处理技术对植物叶片几何数据获得和面积测量进行研究，提出了一种利用CDD摄像机和矩阵标本作为参照物来测量植物叶片面积的方法[9]，该方法具有效率高、速度快、劳动量小等优点。整个过程从采集植物叶片开始，到植物叶片处理过程加权平均法的灰度化、Canny算子的边缘检测，调节对比度、叶片区域提取以及最后用像素统计法进行面积测量。与实际面积的误差很小，满足日常的需求。

参考文献：

[1] 吕书财,徐瑶,陈国兴,等.大豆冠层光合有效辐射、叶面积指数及产量对种植密度的响应[J].江苏农业科学,2018,46(18):68-72.

[2] 韩殿元,黄心渊,付慧,等.基于彩色通道相似性图像分割方法的植物叶面积计算[J].农业工程学报,2012,28(6):179-183.

[3] 于东玉,冯天祥,李奕昕,等.基于植物图像的活体叶片面积测量方法研究与实现[J].智能计算机与应用,2019,9(4):173-176.

[4] 徐贵力,毛罕平,胡永光.基于计算机视觉技术参考物法测量叶片面积[J].农业工程学报,2002,18(1):154-157.

[5] WANG Y,JIN G,SHI B,et al. Empirical models for measuring the leaf area and leaf mass across growing periods in broadleaf species with two life histories[J]. Ecological Indicators,2019,102:289-301.

[6] 韩少杰.基于图像综合特征与基于拉普拉斯高斯算子的零水印算法[D].天津:天津职业技术师范大学,2015.

[7] 蔡文宇.自然背景下的植物叶片多阈值分割方法研究[D].哈尔滨:东北林业大学,2018.

[8] 崔世钢,秦建华,张永立.基于图像处理技术的植物叶片面积和周长测量[J].江苏农业科学,2018,46(15):187-189.

[9] 李明,张长利,房俊龙.基于图像处理技术的小麦叶面积指数的提取[J].农业工程学报,2010(1):205-209.

（下转第44页）

实践与经验

文章编号: 1007-1423（2023）14-0037-08 DOI: 10.3969/j.issn.1007-1423.2023.14.008

基于招生数据利用不同的机器学习方法预测大一学生成绩

王 琛*

（温州肯恩大学信息技术中心，温州 325060）

摘要：利用机器学习方法从招生数据中挖掘影响高校学生大一GPA的因素并构建预测模型。首先，进行了数据清洗、筛选了招生相关数据和大一GPA成绩。随后，将学生的招生信息作为特征，训练了三种机器学习模型，分别为线性回归模型、逻辑回归模型和神经网络模型。最后，对三个模型的性能进行了评估，并对神经网络模型进行了优化。研究的成果可以为高校大一GPA预测建模提供借鉴，并有助于推进学业预警、学业成绩预测和评价的实践。研究结果表明，机器学习方法可以有效地挖掘影响高校学生大一GPA的因素并构建预测模型。通过训练和比较不同的机器学习模型，研究提供了一个可行的预测模型，并对神经网络模型进行了优化，提高了其预测精度。这些成果可以为高校学业预警、学业成绩预测和评价提供有用的参考信息，对于提高学生学习效果和改善教学质量具有积极的作用。

关键词：机器学习；预测建模；学业成绩；神经网络；招生数据

0 引言

随着高校信息化系统的不断完善，学生数据的信息量呈现出急剧增长的趋势。通过对数据进行适当的分析，可以获得有效的评估，从而预测学生的表现。近年来，机器学习技术在预测学生表现方面取得了显著进展[1-3]。与教育数据挖掘不同，机器学习技术可以结合任何学生属性对不同的学生群体进行有效的预测。本研究旨在借助不同的机器学习技术发现招生数据中的潜在规律或模式，继而为教学预警提供服务制定。通过招生数据进行建模，可以预测新生大一成绩，从而帮助高校教学管理者制定预警措施，如对学生进行预警警告或者尽早引导他们改进学习，以便学生在后续的学习中提高成绩和学术水平，这将是教学预警研究中的一个新亮点。

目前，对于招生数据和学业预测的相关研究多局限于对招生数据进行多维数据统计分析[4-5]。本研究将应用不同的机器学习模型对高校大一学生的平均学分绩点（grade point average, GPA）成绩进行预测，通过采集高校实际的招生数据和学业GPA成绩数据进行数据分析和模型预测，从而研究不同机器学习模型在大一学生GPA成绩预测方面的准确度。除了基本的招生数据处理和权重分析外，本研究还将添加大一各阶段性累计平均GPA，从而设计出一个具有预警功能的高校新生学业成绩评估模型，帮助教师及教学管理者为刚进入校门的学生提供初期成绩评估的预警信息和教学帮助参考，有利于高校对学生早期生涯的学业预警进行干预。

1 研究内容与方法

1.1 研究内容

本研究通过分析现有的招生基础数据和大一GPA成绩数据之间的关联，通过机器学习模型发现其中的规律，为高校学业成绩初期预警提供参考。为了达成这个目标，我们设计和采用了不同的机器学习模型。通过对这些模型进

收稿日期：2023-03-29 **修稿日期：**2023-04-07

作者简介：*通信作者：王琛（1988—），男，浙江温州人，硕士研究生，高级工程师，研究方向为教育信息化、人工智能、数字化改革、数字文化，E-mail：wangchen@wku.edu.cn

行训练和对比其在测试集(test set)上的预测结果，我们研究和分析了线性回归模型(linear regression)、多元逻辑回归模型(logistic regression)和神经网络(neural network)在大一GPA成绩预测准确率方面的表现。之后，我们还对进行了主成分分析(PCA)，期望能够从高校招生历史数据中挖掘重要的信息，消除招生新生历史数据之间的重叠数据和噪声数据。最后，将PCA和精度较高的预测模型相结合，首先采用PCA降低数据维度，提取对预测结果影响较大的主成分因子，然后使用预测准确率较高的模型对招生基础数据进行预测。图1为预测模型的流程。

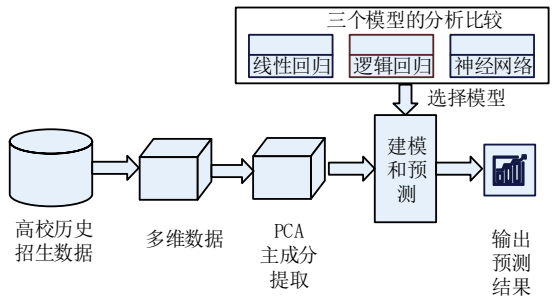


图1 预测模型流程

1.2 研究方法

本研究采用了模型研究和数据分析相结合的方法。在模型研究方面，使用PyTorch框架构建线性回归模型、多元逻辑回归模型和神经网络模型，并对它们进行训练。PyTorch是一个基于Torch的Python开源机器学习库，可用于自然语言处理等应用程序，支持动态神经网络，具有强大的灵活性和扩展性。在数据采集和分析方面，我们将收集高校的招生基础数据和学生大一的GPA成绩(四分制)数据，对其进行学习和训练。此外，我们还对数据进行处理，以便发现数据中的主要特征，消除噪声和冗余信息。

在选择机器学习模型方面，我们考虑了所采用的每个模型的优点和局限性，并根据实际情况进行了设计。在PCA步骤中，我们选择了最小化信息损失和最大化方差的方案，以确保提取到的主成分因子尽可能的准确和重要。在最后的优化步骤中，选择了最适合的模型来对GPA成绩数据进行预测。

1.2.1 数据和数据预处理

本研究采集的数据来自某大学2016年至2019年的招生基础数据和对应学生的GPA成绩(四分制)数据。使用的模型之一神经网络模型是一种基于大数据分析的人工智能模型，当数据量较大时，其预测精度较高。然而，当样本数据量较小时，模型容易陷入局部最优，出现过拟合现象，而降低预测精度。因此，在数据收集阶段，本研究将尽可能扩大数据样本的维度，收集新生的各种信息属性，以减少误差并提高准确率。

本研究还将采用数据重构方法，获取高校新生多维基础历史数据和学生成绩历史数据。采用数据重构的原因是，高校招生受到当年国家政策、社会需求、社会经济状态等因素的影响，导致招生数据变化存在非线性和复杂性。此外，不同省份给出的招生数据内容、格式和代码各不相同，成绩构成也各不相同，例如浙江省新高考改革后不分文理科。

表1展示了某省实际招生数据的变化。可以看出，该数据有着该省份特有的解析规则和信息，并且不同字段样本间存在数量级的差异。

在数据采集之后，本研究进行了数据清洗、数据集成、数据变换等数据预处理步骤。比如，对于采集到的招生数据中的缺失值本研究采用了插值法，通过计算平均值将其填充到缺失值的位置上。同时，由于招生数据中存在大量的类别值或离散值，为了提高模型的准确性，以及为了让模型的距离计算更合理，本研究使用Pandas中的get_dummies()函数对所有离散型特征进行了One-Hot编码。经过处理后，得到的招生数据集为一个5615行、50列的矩阵，如表2所示。

显然，上述数据具有高维度和高噪声等特性，并且不同省份或样本之间存在数量级的差异。当各招生数据指标间的水平相差很大时，如果直接用原始指标值进行分析，就会突出数值较高的指标在模型分析中的作用，相对削弱数值水平低的指标的作用。因此，在分析数据之前，我们还对数据集进行了标准化和归一化处理。通过归一化将招生数据按比例缩放至一

个小的特定区间，能有效降低数据集噪音，便于不同单位或数量级的指标能够进行比较和加权，这样便于后续数据处理和加快模型收敛[6]。本研究采用了Min-Max Scaling、Z-Score Normalization和L2 Normalization三种归一化算法进行尝试，基于训练样本上的总损失值比较结果，最终选择了Min-Max Scaling作为归一化算法。具体算法公式如下：

$$\frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (1)$$

最后，将数据集划分为训练集(train_Set)、评估集(valid_set)和测试集(test_set)，并进行了交叉验证(cross validation)。数据集的部分参数值详见表3。

1.2.2　基于数据统计的相关性分析

本研究首先对学生高考成绩和大一GPA成绩进行了皮尔逊相关系数的计算。皮尔逊(pearson)积矩相关系数是传统数据统计里常用的一种方法[7]，用于度量2个变量间相关程度，它是一个介于1和-1之间的值。其中：1表示变量完全正相关，0表示无关，-1表示完全负相关，计算结果见表4。

表1　某省实际招生数据

性别代码	科类代码	毕业类别	录取代号	高考成绩	特长	高考裸分成绩	专业代码	录取身份	外语代码	外语分数	批次
1	Z	00	11	582	?	582	012	10	1	89	1
1	Z	00	11	582	1	582	007	10	1	97	1
1	Z	00	11	586	1	586	006	10	1	102	1
1	Z	00	11	585	2	585	006	10	1	103	1
1	Z	00	11	579	?	579	005	10	1	106	1

表2　预处理后的招生数据集矩阵

序号	出生年份	出生月份	考生类型	获奖情况	英语分数	高考分数	一段分数	一段分差	录取均分	…
0	1999	9	0	0	106	473	511	-38	458.78	…
1	1998	9	0	0	114	483	511	-28	458.78	…
2	1999	3	2	0	112	485	511	-26	458.78	…
3	1999	8	0	1	110	446	511	-65	458.78	…
…	…	…	…	…	…	…	…	…	…	…
5680	1998	4	0	1	129	505	456	49	477.46	…
5681	1999	9	0	0	130	486	456	30	477.46	…
5614	1998	10	0	1	115	467	456	11	477.46	…

表3　归一化处理之后的训练集片段

0.329576	0.273933	0.000000	0.000000	0.000000	0.086317	0.109573	0.171208	…
0.284504	0.275883	0.000000	0.000000	0.000000	0.150155	0.129320	0.172427	…
0.341853	0.071034	0.000000	0.000000	0.000000	0.089533	0.131414	0.177586	…
0.313932	0.097849	0.358779	0.000000	0.000000	0.127068	0.061971	0.163081	…
0.291179	0.211767	0.000000	0.332776	0.332776	0.103993	0.068068	0.151262	…
…	…	…	…	…	…	…	…	…

表 4 不同特征与大一成绩间的相关系数

特征	Pearson 相关系数	Spearman 相关系数	Kendall 相关系数
高考外语	0.197846	0.217181	0.169518
高考数学	0.239372	0.255286	0.171152
高考语文	0.207756	0.223972	0.166089
高考总成绩	0.339547	0.293627	0.189425

从表4可以看出，高考总成绩和学生大一GPA成绩之间的相关系数值略高于0.3，表明二者具有一定程度的相关性，但不显著。而单门课程的高考成绩和大一GPA成绩的相关系数值都小于0.3，可以认为相关性不高。此外，从图2也可以明显看出高考成绩对学生大一GPA的影响没有呈线性递增。分析原因我们认为高校的教育以专业课程为主，更加细分化的专业知识减少高考成绩对学生大学学习的实际影响程度。因此，可以得出结论，虽然高考成绩与大一GPA成绩之间具有一定程度的相关性，但影响并不明显。此外，基于传统统计的相关性分析方法对于没有明显统计学规律的多元复杂数据的效果并不理想。

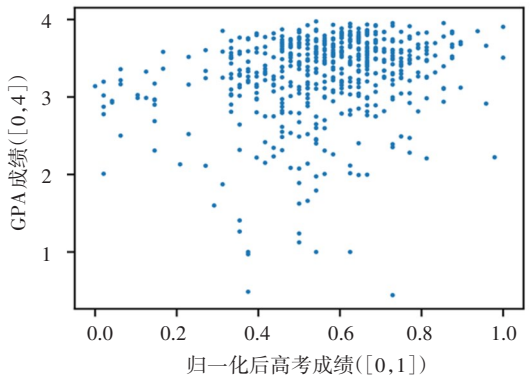


图 2 高考成绩和大一GPA关系分布

1.2.3 多元线性回归模型分析

线性回归是一种广泛应用的回归模型，其基于以下假设：自变量X和因变量y之间的关系是线性的，即y可以表示为X中元素的加权和[8]。我们假设招生数据和大一GPA之间存在线性关系，可以表示为特征(如英语成绩、总成绩、年龄等)的加权和。当使用线性代数来建模时，我们输入包含d个特征(在本研究中为50个)的高维招生数据集，将预测结果$\hat{y}$表示为

$$\hat{y} = w_1 x_1 + \cdots + w_d x_d + b \tag{2}$$

其中：每个x_i代表数据集中的一个特征；w_i代表该特征的权重；b表示偏差项。在给定招生训练数据特征X和对应的GPA成绩y，线性回归模型的目标是找到一组最优的权重向量$\boldsymbol{W}$和偏差项b，这组权重向量和偏差能够使得新样本预测标签的误差尽可能小。

为了寻找最好的模型参数W和b，我们需要定义损失函数和使用随机梯度下降方法来学习和训练模型参数。损失函数用来量化目标值和预测值之间的差距，在训练模型时我们的目标是找到一组参数(w^*, b^*)，通过计算训练集样本在损失函数$L(w,b)$上的损失均值，最小化在所有训练样本上的总损失：

$$w^*, b^* = \operatorname{argmin} L(w,b) \tag{3}$$

本研究使用PyTorch的MSELoss类来计算均方误差损失函数，也称为平方L2范数。梯度下降方法几乎可以优化所有机器学习模型。它通过在损失函数递减的方向上不断更新参数来降低误差。我们使用以下公式来表示这一更新迭代的过程(其中∂表示偏导数)：

$$w \leftarrow w - \frac{\eta}{|\beta|}\sum_{i\in\beta}\partial_w l^{(i)}(w,b) = w - \frac{\eta}{|\beta|}\sum_{i\in\beta} x^{(i)}\left(w^{\mathrm{T}}x^{(i)} + b - y^{(i)}\right) \tag{4}$$

经过多次迭代后，我们就能通过有限的数据来训练模型参数(w^*, b^*)。为了更好地评估训练效果，本研究会计算每次迭代周期后的损失，并使用更新后的参数计算评估集的准确率，以监控训练过程。记录的结果见表5。

表 5 线性回归模型每个迭代周期后的损失和评估集准确率

迭代次数	损失均值(loss)	准确率(acc)
epoch 1	0.891552	0.238997
epoch 2	0.461995	0.574387
epoch 3	0.261367	0.605596
epoch 4	0.254532	0.608116
epoch 5	0.252791	0.611502
epoch 6	0.245001	0.613754
epoch 7	0.214800	0.631189
epoch 8	0.226967	0.617542
epoch 9	0.228253	0.618833
epoch 10	0.226371	0.617951

通过这个已经训练好的线性回归模型，我们就可以使用招生数据测试集来预测未包含在训练数据和评估集中的学生大一GPA，并查看模型的准确度。

1.2.4　多元逻辑回归模型分析

多元逻辑回归本质上是多元线性回归，只是在特征到结果的映射中加入了一层函数映射[9-11]。该映射函数为Logistic函数（也称为Sigmoid函数），其形式如下：

$$s(z) = \frac{1}{1 + e^{-z}} \tag{5}$$

首先进行特征线性求和：

$$\theta^{T}x = \theta_0 + \sum_{i=1}^{n}\theta_i x_i \tag{6}$$

然后使用函数$s(z)$构造预测函数来进行预测，逻辑回归的最终表达式如下：

$$P_\theta(x) = s\left(\theta^{T}x\right) = \frac{1}{1 + e^{-\theta^{T}x}} \tag{7}$$

虽然逻辑回归是非常经典的分类算法模型，但主要用于解决二分类问题，因此需要对GPA数据进行预处理。在本次研究中，将学生大一GPA大于等于3.2的样本作为一个类别，小于3.2的样本作为另一个类别，通过训练模型得到属于每个类别的概率。本研究使用PyTorch的BCELoss类作为逻辑回归模型的损失函数。每次迭代周期后的损失以及使用更新后的参数计算的在评估集的准确率见表6。

表6　逻辑回归模型每个迭代周期后的损失和评估集准确率

迭代次数	损失均值(loss)	准确率(acc)
epoch 1	4.66181235315	0.81
epoch 2	2.87509441375	0.83
epoch 3	4.38195184967	0.82
epoch 4	7.71832617382	0.78
epoch 5	3.05447617865	0.83
epoch 6	1.58217272583	0.84
epoch 7	5.03531426950	0.81
epoch 8	5.23422103903	0.81
epoch 9	4.35069798423	0.82
epoch 10	1.72436655642	0.84

可以发现该模型在评估集上准确率很高，主要原因是该模型只将预测结果分为了两类。然而，这种分类方式颗粒度并不够精细，可能会导致欠拟合和精度不高的问题。

1.2.5　神经网络模型分析

前面两个模型都是基于单调性假设的线性回归模型，即任何特征值的增大或减小都会导致模型预测输出的增大或减小(对应权重为正或负)。例如，当我们试图预测学生大一GPA时，我们假设在其他条件不变的情况下，高考成绩高的学生可能比成绩较低的学生获得更高的GPA[12]。然而，即使高考成绩与大一GPA成绩之间存在单调性，它们之间的关系也不一定是线性的。高考成绩值在高分值段多5分的人可能比在低分值段多25分的人能获得更高很多的GPA。为了处理这类非线性问题，我们采用了神经网络模型。

神经网络模型是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。该模型通过调整内部大量节点之间相互连接的关系来实现信息处理，是目前应用最广泛的机器学习模型[13-15]。如图3所示，通过在网络中加入一个或多个隐藏层，可以克服线性模型的限制，使其能够处理更普遍的函数关系类型。

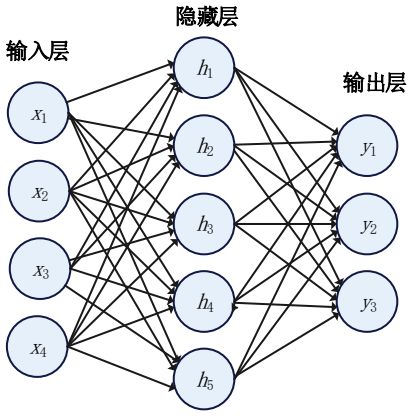


图3　一个单隐藏层的简单神经网络

本研究设计的神经网络模型为单隐藏层神经网络，该网络包含20个隐藏单元的单一隐藏层，使用CrossEntropyLoss交叉熵损失作为损失函数，并采用ReLU作为激活函数。ReLU提供了一种非常简单的非线性变换。对于给定的元

素 x，ReLU 函数被定义为该元素 x 与 0 的最大值：

$$ReLU(x) = \max(x, 0) \quad (8)$$

在本次研究中，我们采用传统的分段绩点法，将学生大一GPA成绩按照分数区间划分为11个等级类别，分别为A=4.0、A^-=3.7、B^+=3.3、B=3.0、B^-=2.7、C^+=2.3、C=2.0、C^-=1.7、D^+=1.3、D=1.0和F=0，因此设计的输出层包含11个输出单元。通过训练模型不断迭代来优化神经网络的权重和偏差参数，使神经网络可以学习到输入数据和输出之间的映射关系，从而能够准确地预测大一学生的GPA等级是多少。模型每次迭代周期后的损失和评估集的准确率如表7所示。

表 7　神经网络模型每个迭代周期后的损失和评估集准确率

迭代次数	损失均值(loss)	准确率(acc)
epoch 1	1.842503	0.605
epoch 2	0.611896	0.674
epoch 3	0.572317	0.713
epoch 4	0.557163	0.711
epoch 5	0.477176	0.702
epoch 6	0.429863	0.794
epoch 7	0.418220	0.775
epoch 8	0.404126	0.792
epoch 9	0.395612	0.783
epoch 10	0.375495	0.783

2　实验结果与优化

2.1　实验结果与分析

我们将训练好的三个模型在测试集上运行。三个不同的机器学习模型对测试集招生数据进行预测，结果显示多元线性回归模型构建的预测模型的准确率最低，为0.6019；其次是神经网络模型，准确率为0.7227；逻辑回归的准确率值最高，为0.8035。表8列出了不同算法的预测分类结果。

表 8　不同模型方案在测试集上的准确率

模型	多元线性回归模型	逻辑回归模型	神经网络模型
准确率	0.6019	0.8035	0.7227

可以看出，其中两种机器学习算法的准确率均在70%以上，预测效果良好。

为了获得更好的分类效果和更高的精准度，本研究尝试利用主成分分析(PCA)对招生数据进行预处理，提取其主成分后，在三个不同的模型上运行。最终测试集的准确率见表9。结果表明，对招生数据进行PCA主成分提取处理，对模型的精准率提高并不显著，而在某些情况下，可能会导致一定程度的信息丢失。

表 9　PCA主成分提取后模型方案在测试集的准确率

模型	多元线性回归模型	逻辑回归模型	神经网络模型
准确率	0.6594	0.8145	0.7382

2.2　模型优化

本次优化我们增加了学生大一不同阶段的累计GPA，包括大一上期中GPA、大一上期末GPA、大一下期中GPA数据。这样的优化考虑到了GPA的时序性，更能反映学生在大一不同时期的学习表现。这一优化有助于提高模型的预测精度，进而更准确地预测学生的大一期末GPA成绩。考虑到本研究在采用逻辑回归模型时分类颗粒度大，虽然准确度更高但在分类效果上存在一定的不足。相比之下，神经网络模型更适用于该任务。因此，本研究选择使用神经网络模型进行优化调整，以提高预测精度。优化后的模型如图4所示。

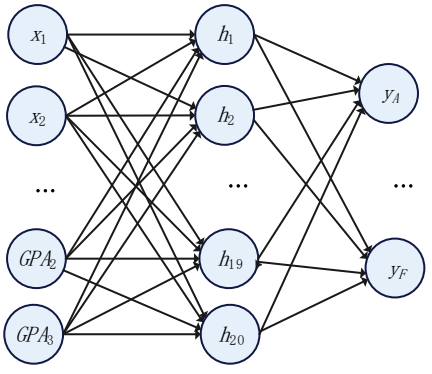


图 4　增加了不同阶段累计GPA的模型

使用训练集反复训练神经网络并进行权重和偏置的调整后，我们得到了一组优化后的神经网络模型。在测试集上的测试结果表明，这组优化后的模型预测准确率显著提高，其表现

见表10。

表 10　优化后模型方案在测试集的准确率

模型	未优化的神经网络模型	优化后的神经网络模型
准确率	0.7382	0.8474

3　结语

本研究的主要目的是通过分析招生数据的影响特征来预测学生的大一GPA，研究结果表明传统的统计模型在满足预测需求方面存在不足，因此本研究使用了多元线性回归模型、多元逻辑回归模型和神经网络模型构建预测模型，并通过比较它们的准确率来评估它们的性能。结果显示，逻辑回归模型表现最佳，神经网络模型也比传统多元线性回归方法具有更好的预测性能，尤其适用于分类预测。考虑到本研究在采用逻辑回归模型时分类颗粒度大，因此优化神经网络模型对大一GPA进行预测更具有应用前景和参考价值。

然而，本研究存在一些局限性。首先，所使用的数据量相对较少，因此在今后的研究中将进一步增加招生数据的训练集数据量以提高模型的拟合能力，使实验结果更具有普适性。其次，在选取的输出结果变量中，大一成绩分类颗粒度不够细。并且，神经网络模型也只使用了单层模型。虽然单隐层网络能够学习任何函数，但如果使用更深的网络可以更容易地逼近预测结果并获得更高的精准度。最后，本研究使用的招生数据集存在一定程度的样本缺失，虽使用了均值填补法对缺失值进行了填补，但仍可能对模型的准确性产生一定程度的影响。未来的研究可以进一步扩大训练集数据量、增加招生数据变量并构建更细粒度的模型、采用更深层的神经网络模型等来改善模型性能。

本研究有助于高校在入学阶段就预测学生未来的学业成绩，并提早规划出有针对性的学业辅助，从而提高学生整体成绩。本研究初步证明了机器学习模型通过招生数据对大一成绩的预测具有较好的准确性。未来可在此基础上进一步改良和优化机器学习模型，同时发掘更多与成绩有相关性的学生信息，随着学生信息大数据规模的不断扩展和增长，这项研究的应用潜力将得到进一步评估和优化。

参考文献：

[1] 吕品，于文兵，汪鑫，等. 基于机器学习的学生成绩预测及教学启示[J]. 计算机技术与发展，2019，29(4)：200-203.

[2] 徐铭希. 机器学习在学生成绩预测中的应用[J]. 电子制作，2019(2)：42-44.

[3] SUHAIMI N M，ABDUL-RAHMAN S，MUTALIB S，et al. Review on predicting students' graduation time using machine learning algorithms [J] International Journal of Modern Education and Computer Science，2019，11(7)：1-13.

[4] 杨曼，高飞，朱倩，等. 高考成绩与大学成绩相关性分析[J]. 云南民族大学学报(自然科学版)，2022，31(3)：360-365.

[5] 丁澍，缪柏其，叶大鹏. 高考成绩与大学成绩的相关性分析[J]. 中国大学教学，2008(11)：29-31.

[6] 班文静，姜强，赵蔚. 基于多算法融合的在线学习成绩精准预测研究[J]. 现代远距离教育，2022(3)：37-45.

[7] 刘辉，邵福波，宫响. 经典相关系数及统计功效对比研究[J]. 青岛科技大学学报(自然科学版)，2022，43(1)：111-119.

[8] 冯玉婷，黄雨欣，郭玉堂，等. 基于多元线性回归模型的网络安全漏洞预测研究[J]. 长春师范大学学报，2021，40(12)：16-23.

[9] 喻佳，白舒伊，吴丹新. 基于机器学习的在线教学学生成绩预测研究[J]. 电脑编程技巧与维护，2021(8)：118-119，154.

[10] 张俭鸽，李颖颖. 基于多元线性回归预测模型的sensor态势研究[J]. 计算机技术与发展，2011，21(9)：229-232.

[11] MOUSTRIS K P，NASTOS P T，LARISSI I K，et al. Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater athens area，greece [J]. Advances in Meteorology，2012，2012：978-988.

[12] 汪朝杰，谭常春，汪慧. 大学生在校成绩与高考成绩的统计分析[J]. 大学数学，2013，29(4)：79-86.

[13] 袁利平，陈川南. 人工智能视域下的宽度学习及在教育中的应用[J]. 远程教育杂志，2018，36(4)：49-56.

[14] STANKOVIC N L，BLAGOJEVIC M D，PAPIC M Z，et al. Artificial neural network model for prediction of students' success in learning programming[J]. J. Sci. Ind. Res.，2021，80.

[15] ZACHARIS N Z. Predicting student academic performance in blended learning using artificial neural networks [J]. Int. J. Artif. Intell. Appl.，2016，(5)：17-29.

Predicting freshman year college grades using different machine learning methods based on admission data

Wang Chen*

(Information Services Office, Wenzhou-Kean University, Wenzhou 325060, China)

Abstract: Using machine learning methods to mine factors influencing college students' first-year GPA from enrollment data and construct prediction models. Firstly, data cleaning was performed and relevant enrollment data and first-year GPA scores were filtered. Subsequently, students' enrollment information was used as features to train three different machine learning models: linear regression, logistic regression, and neural network models. Finally, the performance of the three models was evaluated, and the neural network model was optimized. The results of this study can provide a reference for modeling first-year GPA predictions in colleges and universities, and help promote the practice of academic early warning, academic performance prediction, and evaluation. The study shows that machine learning methods can effectively mine factors influencing college students' first-year GPA and construct prediction models. By training and comparing different machine learning models, the study provides a feasible prediction model and optimizes the neural network model, improving its prediction accuracy. These results can provide useful reference information for academic early warning, academic performance prediction, and evaluation in colleges and universities, and have a positive impact on improving student learning outcomes and teaching quality.

Keywords: machine learning; predictive modeling; academic achievement; neural network; enrollment data

(上接第36页)

Measurement of plant leaf area based on image processing technology

Hao Xiaofeng*, Yu Rongrong

(School of Computer Science, Xijing University, Xi'an 710200, China)

Abstract: Aiming at the measurement of leaf area, a method for measuring leaf area based on image processing is designed and implemented. Use the CDD camera to obtain plant leaf images, perform image segmentation and edge detection on the collected plant leaf images, select ideal pictures by comparing Sobel, Robert, Prewitt and Canny operators, and then use morphological processing to smooth the gaps and noises, and finally The measured area is obtained by the pixel statistics method, and the correlation analysis is carried out with the real area. This method can accurately and conveniently measure the area of plant leaves.

Keywords: image processing; edge detection; area; pixel

文章编号: 1007-1423（2023）14-0045-07 DOI: 10.3969/j.issn.1007-1423.2023.14.009

基于多维数据挖掘的学生学习画像构建

许惠惠*

（山西药科职业学院器械工程系，太原 030031）

摘要：以多维数据为基础，应用数据挖掘技术预测学生学习习惯并构建学习画像。通过收集273名学生的数据，涵盖人际关系、个性特征和健康状况等因素，经过数据预处理，成功建立了基于支持向量机、K近邻、多层感知器的多维数据挖掘模型，实现对学生学习习惯的预测。研究结果展现了全方位的学生画像，包括时间管理、学习方法等方面。该研究证实了数据驱动决策在教育领域的重要性，为提供个性化教育方案提供了科学依据。

关键词：数据挖掘；学生画像；支持向量机；K近邻；多层感知器

0 引言

随着信息技术和大数据时代的到来，数据驱动的决策在教育领域中受到极大的关注。学生学习画像可以提供有关学生学习习惯、人际关系、家庭背景等方面的信息，有助于教育工作者深入了解学生的需求，从而提供更为个性化的教育方案[1-3]。本研究旨在探讨多维数据在学生学习画像中的应用，利用多种数据挖掘技术，综合分析影响学生学习能力的因素，并尝试通过学生的多种学习属性构建学生的全方位学生学习画像。

数据挖掘是一种从大规模、复杂的数据集中提取有用信息、知识和规律的过程。这种技术结合统计学、机器学习、人工智能等多种方法，对原始数据进行处理、归纳和挖掘，为决策者提供有价值的参考依据[4]。利用这种技术，教育工作者可以更精确地评估学生的学习需求，建立相应的学生画像，从而为他们提供个性化的教育支持。

学生学习画像的构建具有重要的实际意义，它可以帮助教师发现学生的潜在问题，及时调整教学策略，提高教学质量。同时，学生学习画像还可以为学生提供有针对性的学习资源和建议，促进他们的自主学习[5]。然而，构建一个有效的学生学习画像并非易事，需要考虑诸多因素，如个性特征、家庭背景、社会经历等[6-7]。因此，本研究通过收集大量学生数据，分析这些因素与学生习惯的关系，为构建学生学习画像提供参考。

本研究共收集了273名学生的调查问卷数据，包括五个方面的属性信息：人际关系、个性特征、家庭背景、社会经历和健康状况，以及五个维度的学习习惯主题问题。通过对这些数据的分析，我们将提取与学生学习习惯相关的背景特征，并尝试利用数据挖掘方法探索学生各维度信息对学习习惯的影响，从而构建学生学习画像。

收稿日期：2023-04-14 **修稿日期：**2023-06-23

基金项目：2021年度山西省高等学校哲学社会科学研究项目（思想政治教育专项）（2021zsszsx207）：新时代高职大学生群体画像构建研究；山西省高职院校思想政治教育研究会2021年度思想政治教育研究项目课题（SYH2021-032）：高职院校学生心理健康测评与服务系统开发研究

作者简介：*通信作者：许惠惠（1983—），女，山西洪洞人，硕士，讲师，主要研究方向为思想政治教育、管理信息系统，E-mail：89442898@qq.com

1 方法

本研究旨在基于多维特征提取构建学生学习画像并预测学生学习习惯类型，最终构建不同类型的学习习惯类型学生画像。在此部分中，将详细阐述研究方法，包括数据收集、问题定义以及模型构建。

1.1 数据收集

首先，为了获取学生的多维背景信息数据，本研究设计了一份涉及学生五个背景领域的调查问卷，共计收集了273名学生的相关信息。调查问卷包含人际关系、个性特征、家庭背景、社会经历和健康状况五个方面的属性特征，这五个属性特征与学习习惯和学生学习画像之间的关系如图1所示。

其中每个属性含有三个具体的衡量指标，见表1。我们选择这五个属性是因为它们通常被认为是影响学生学习习惯的重要因素。例如，学生的人际关系和社会经历可能会影响他们的团队合作能力和社会适应性；个性特征可能会影响他们的学习方式和动机；家庭背景可能会影响他们的学习环境和资源；而健康状况则可能影响他们的学习效率和持久力。因此，这五个方面的信息为我们提供了学生学习画像的全面视角。

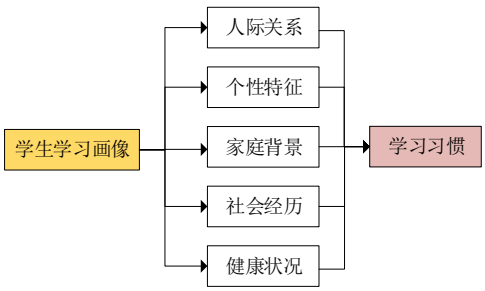


图1 学生学习画像属性图

同时为了对学生学习习惯做一个综合全面的描述和评分，我们还设计了一份用于统计学生学习习惯的调查问卷，见表2，该调查问卷主要从五个方面对学生的学习习惯进行刻画，以多个维度来评估学生的学习情况。

表1 调查问卷样例格式

指标/类别	人际关系	个性特征	家庭背景	社会经历	健康状况
1	朋友圈子数量（亲密朋友）	五大人格量表得分（1～5分）	家庭经济状况（高，中，低）	志愿者活动次数	每周锻炼次数
2	社交活动频率（每月）	学习动机评分（1～5分）	父母教育水平（普通，中级，高级）	兼职经历次数	睡眠质量评分（1～5分）
3	团队合作能力评分（1～5分）	时间管理能力评分（1～5分）	家庭对学习的支持程度（1～5分）	社团活动次数	心理健康状况评分（1～5分）

表2 学生学习习惯调查问卷

主题	指标	分数区间
学习时间管理	我经常制定学习计划并严格执行	（1～5）
学习方法	我通常会自我复习并总结学习内容	（1～5）
学习动机	我对我正在学习的主题感兴趣	（1～5）
学习环境	我可以在嘈杂的环境中集中精力学习	（1～5）
学习压力和应对策略	当我感到压力时，我有一套有效的应对策略	（1～5）

1.2 问题定义

本研究中，我们将学生画像中的学生学习习惯类型作为衡量学生学习画像的一项主要指标，将上述统计中学生五个方面的背景信息(人际关系、个性特征、家庭背景、社会经历和健康状况)作为影响学生学习习惯的因素。将学习习惯类型预测作为一项分类任务，目的是确定学生的哪些因素指标对学生学习能力的影响较高。在实验设计中，每个学生会有以上五维的表征向量，每个向量中包含三个衡量指标分数，标签即为学生的学习习惯类型，学习习惯类型

预测任务定义如下：

问题定义：给定学生S的五维表征向量，见公式(1)：

$$S=\{s_1,s_2,\cdots,s_n\},s_i=\{x_1,x_2,x_3\},n=5 \quad (1)$$

其中：s_i表示五个主要因素的指标得分，x_i代表每个因素下的具体指标，如人际关系中的朋友圈子数量、社交活动频率和团队合作能力评分等。

学生的学习习惯类型为$y\in\{1,2,3,4,5\}$，分别对应$\{A,B,C,D,E\}$不同的学生学习类型，这五种类型由表2中的学习习惯调查问卷得出，主题包括学习时间管理、学习方法、学习动机、学习环境、学习压力和应对策略等。

学生的学习习惯类型预测任务可以被描述为学习一个映射函数：

$$F:S\times \boldsymbol{A}\rightarrow y \quad (2)$$

其中：$\boldsymbol{A}$是映射矩阵，y是学生的学习习惯类型。预测问题为一项分类任务，目标是预测特定学生画像下的学习习惯类型。

本文目标是从各项背景指标中挖掘出与学生学习习惯具有较强关联的指标，这些指标将作为后续针对学生课程设计和学习习惯加强的关键依据。

1.3　模型构建

为了从多维数据中提取与学生学习习惯相关的特征，本研究采用了支持向量机(support vector machine，SVM)和K近邻(K-nearest neighbors，KNN)算法并行地提取处理调查问卷中的五个主要因素：人际关系、个性特征、家庭背景、社会经历、健康状况，将相关指标作为输入，提取出与学生学习习惯类型高度相关的特征表征。

例如，对于人际关系因素，本方法将朋友圈子数量、社交活动频率和团队合作能力评分这三个属性输入到SVM和KNN模型中，生成特征表征。同样的方法也被应用到其它四个因素的处理上。

这些特征表征被融合起来，作为学生的综合特征表示，这个向量代表了学生在各因素上的表现和属性的组合。随后将其输入到多层感知机(multilayer perceptron，MLP)模型中，MLP模型会根据这个综合特征向量预测学生的学习习惯类型。

这一研究设计旨在通过SVM和KNN对各因素的详细处理，以及MLP对各因素关系的深度理解，实现对学生学习习惯的准确预测和理解。以期通过这种方法提供一个准确和有深度的理解学生学习习惯类型的方式。本研究的模型如图2所示。

1.3.1　基于支持向量机的特征提取

支持向量机(SVM)是一种监督学习方法，主要用于分类和回归任务[8]。SVM算法的基本思想是找到一个最优超平面，将不同类别的样本尽可能地分开[9]。在本研究中，由于学生不同特征属性之间关系很难直观被发现，所以在我们的应用中，SVM主要用于处理调查问卷中关于学生个性特征和家庭背景的问题，比如学习动机评分、时间管理能力评分、家庭经济状况、父母教育水平等，这些特征在高维空间中的分布可能会影响学生的学习习惯类型。我们使用SVM对学生的原始数据进行特征丰富，进一步提升学生的特征维度，有益于后续模型从中学习到关于不同表征之间的关系。SVM通过构建最大间隔超平面，提取出对预测任务具有较高贡献的特征[10]。

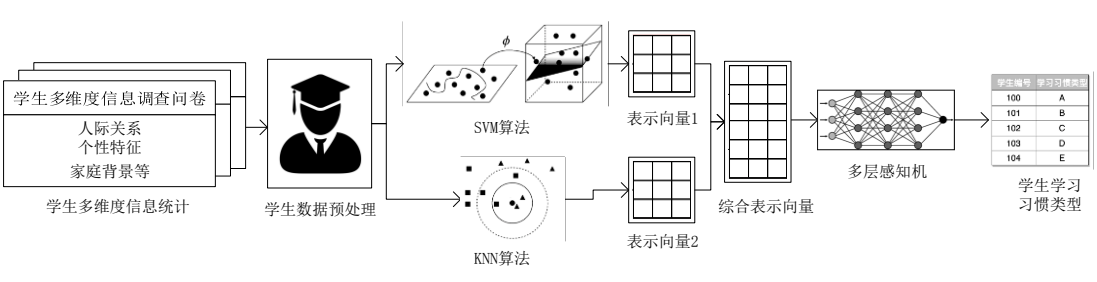


图2　模型结构

SVM模型的输出特征向量是从特征提取后的数据中得出的，使用多项式核函数的计算如公式(3)所示：

$$f(x) = sign\left(\sum_{i=1}^{n}\alpha_i y_i\left(\langle \boldsymbol{x}_i, x\rangle + c\right)^d + b\right) \quad (3)$$

其中：$\boldsymbol{x}_i$为训练样本的特征向量，y_i为样本的标签，α_i为对应样本的拉格朗日乘子，b为偏置项，c和d是多项式核函数的参数，最终得到样本同学的基于支持向量机的特征表征向量。

1.3.2 基于KNN的特征提取

K近邻(KNN)算法是一种基于实例的学习方法，同样可用于分类和回归任务[11]。KNN算法的核心思想是根据一个样本在特征空间中距离最近的K个邻居的类别来确定该样本的类别[12]。KNN能够处理调查问卷中关于学生人际关系的问题，比如朋友圈子数量、社交活动频率等，还可以处理调查问卷中关于学生社会经历的问题，比如志愿者活动次数、兼职经历次数等。这些特征在局部空间的相似性可能会影响学生的学习习惯类型。KNN通过计算样本间的距离并分析邻近样本的类别，找出具有较强预测能力的特征[13]。本研究中，我们使用欧几里得距离度量方法，对应的计算如公式(4)所示：

$$Distance = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \quad (4)$$

其中：x_i为学生i输入的样本特征，y_i为学生i对应的习惯类型。

1.3.3 多层感知机

多层感知机(MLP)是一种前馈神经网络，我们使用四层(输入层、两个隐藏层和输出层)网络的结构。MLP可以用于解决复杂的非线性问题，并广泛应用于分类和回归任务[14-16]。在预测阶段，我们将SVM和KNN提取的特征拼接，并输入到MLP模型中。MLP通过激活函数、权重更新和反向传播算法，在训练过程中学习到最优的权重参数，从而实现对学生学习习惯类型的预测。

我们首先初始化参数，包括隐藏层和输出层的权重和偏差，然后对于每个学生，将其输入向量馈送到网络中，计算输出，随后计算输出和真实标签之间的误差，使用误差来调整模型参数。其中每个隐藏层和输出层，计算加权和激活函数如公式(5)所示：

$$Z = X'\boldsymbol{W} + B, \boldsymbol{A} = f(Z) \quad (5)$$

其中：$\boldsymbol{W}$为权重矩阵，B表示偏置，f是ReLU激活函数。

进而计算输出层的误差，使用交叉熵损失函数，如公式(6)所示：

$$L = -\sum_{i=1}^{N} y_i \log(y_i') + (1 - y_i)\log(1 - y_i') \quad (6)$$

其中：N是样本数；y是真实标签；y_i'是预测值。

本研究通过对多维特征进行分析，构建了一个有效的学生学习习惯类型预测模型。该模型能够较好地预测学生的学习习惯类型，并为教育工作者提供有益的信息，以便更好地理解学生的学习状况和需求。

2 实验

2.1 实验参数和环境设置

在实验参数设置方面，我们为支持向量机(SVM)选择了多项式核函数，参数C设为1，参数γ设为0.1。对于K近邻(KNN)算法，我们设定邻居数量(K)为5，并采用欧氏距离作为距离度量。对于多层感知机(MLP)，我们将输入层节点数设置为与拼接后的特征数量相等，隐藏层节点数为128，输出层节点数与学习习惯类型类别数量相等。

本文使用Python 3.7.9作为主要的编程语言，并在Ubuntu 20.04 LTS系统上运行实验。使用的主要Python库包括NumPy 1.18.5、Pandas 1.0.5、Matplotlib 3.2.2、scikit-learn 0.23.1和PyTorch 1.13.0。实验运行在一台Intel(R) Core(TM) i7-9700K CPU的计算机上，配备了16 GB的内存和四块NVIDIA GeForce RTX 2070显卡。

2.2 性能评估指标

为了评估提出的方法在预测学生学习习惯类型方面的性能，我们使用准确率、精确率、召回率和$F1$分数等指标对分类结果进行评估，指标的数学描述如下。

准确度如公式(7)所示：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

精确度如公式(8)所示：

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

召回率如公式(9)所示：

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$F1$分数如公式(10)所示：

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

其中：TP表示真正例(true positive)，TN表示真负例(true negative)，FP表示假正例(false positive)，FN表示假负例(false negative)。

3　实验结果分析

3.1　模型结果分析

根据本研究的实验结果，我们得到了模型的准确率、精确率、召回率和$F1$分数。此外，本文还引入了三种经典的对比方法，分别为决策树(decision tree，DT)、随机森林(random forest，RF)和逻辑回归(logistic regression，LR)。表3是各模型的结果对比。

表3　模型结果对比/%

模型	准确率	精确率	召回率	$F1$分数
决策树(DT)	72.5	71.4	70.8	71.1
随机森林(RF)	76.3	75.6	75.0	75.3
逻辑回归(LR)	74.2	73.8	73.1	73.4
本研究方法	**79.5**	**78.2**	**77.5**	**77.8**

从表3可以看出，本研究所提出的模型在准确率、精确率、召回率和$F1$分数等指标上均优于其他三种对比方法。这说明我们的模型在预测学生学习画像中的学习习惯类型方面具有较高的性能，表现出较好的稳定性和泛化能力。同时，这也证实了将SVM、KNN和MLP结合使用的方法在学生学习习惯预测任务上具有一定的优势。

为了深入了解五个指标(人际关系、个性特征、家庭背景、社会经历、健康状况)对学生学习习惯的影响，我们对每个不同学习习惯进行了分析，其混淆矩阵如图3所示，可以发现本文模型对学习习惯最佳(标签为1)的学生群体预测准确率最高，对学生学习习惯中等(标签为2)的学生预测准确率最低。

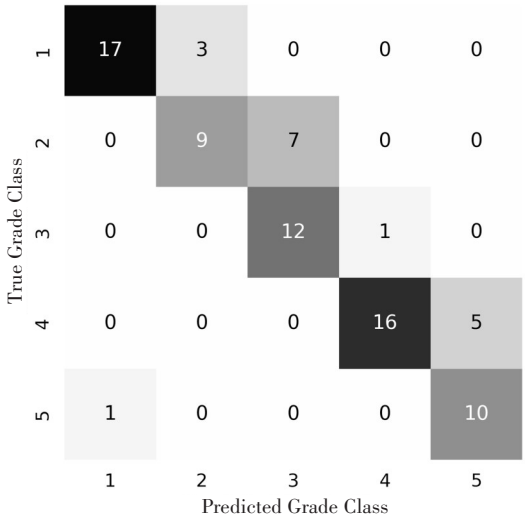


图3　结果混淆矩阵

我们进一步对模型中的权重进行了分析。权重分析能够帮助我们量化各个背景属性对学生学习习惯的影响程度。通过对本研究中学习习惯和背景属性的综合分析，我们发现以下规律：

(1)个性特征：在五个指标中，个性特征对学生学习习惯的影响最大。这可能是因为学习动机、时间管理能力等因素直接影响学生的学习效率和学习习惯。具有较高学习动机和良好时间管理能力的学生在学习过程中更有目标性，从而形成更好的学习习惯。

(2)家庭背景：家庭背景在五个指标中对学生学习习惯的影响排在第二位。家庭经济状况、父母的教育水平，以及家庭对学习的支持程度等因素可能间接影响学生的学习资源、学习环境和心理压力。在一个有利于学习的家庭环境中成长的学生往往能更好地专注于学业，从而形成较好的学习习惯。

3.2　学生画像分类

除了研究学生的背景信息对学生学习习惯的影响外，我们进一步根据不同的学生学习主题进行了学生画像的分类，以下是五种不同类型的学生画像及其所对应的不同的学习习惯类型和特点。

(1)时间管理型学生画像：这类学生的学习时间管理能力强，他们能有效地规划和利用

自己的时间，对于完成学习任务和准备考试有着出色的策略。他们通常对自己的时间表有严格的掌控，且往往在时间管理上展现出显著的自律性。

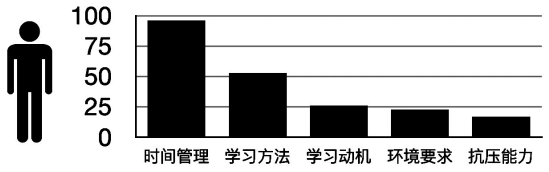


图 4　时间管理型学生画像

（2）探索式学习型学生画像：这类学生的学习方法倾向于探索和实验，他们善于寻找和试验新的学习方法，以提高学习效率和理解能力。他们乐于接受新的观念和思维方式，以及积极探索未知的领域。

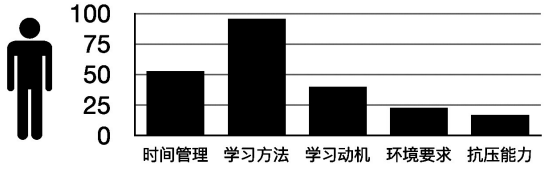


图 5　探索式学习型学生画像

（3）热情驱动型学生画像：这类学生具有强烈的学习动机，他们对学习充满热情和兴趣，能够主动并积极地进行学习。他们对学习的热情驱使他们在面对困难时坚持下去，激发他们不断进步的动力。

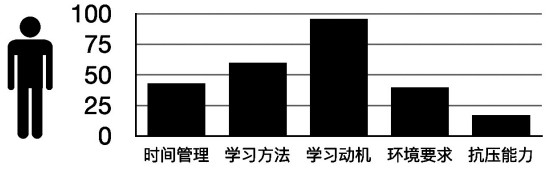


图 6　热情驱动型学生画像

（4）环境适应型学生画像：这类学生对学习环境有较高的适应能力，无论是在安静的图书馆还是在嘈杂的咖啡厅，他们都能保持良好的学习状态。他们能够利用各种环境资源，灵活调整自己的学习方式和习惯。

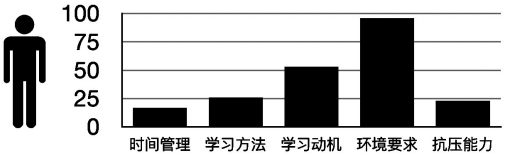


图 7　环境适应型学生画像

（5）压力应对型学生画像：这类学生对学习压力有出色的应对策略，他们能有效地管理和减轻学习压力，保持良好的学习心态。他们明白压力是学习过程的一部分，并已学会如何将其转化为推动自己前进的动力。

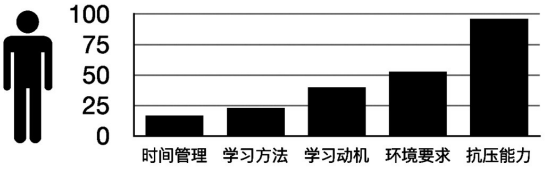


图 8　压力应对型学生画像

4　结语

本研究致力于揭示学生的五个核心维度——人际关系、个性特征、家庭背景、社会经历以及健康状况对学生学习习惯的影响，并通过多维度特征分析来预测学生的学习习惯类型。实验结果证明，本文模型在预测准确率、精确率、召回率以及 *F*1 分数等关键指标上，相较于决策树、随机森林和逻辑回归等常见方法，展现出更高的预测性能。此外，根据学生的学习习惯类型，我们成功绘制了五种不同的学生学习画像：时间管理型、探索式学习型、热情驱动型、环境适应型和压力应对型。

未来的研究应重点关注如何更有效地利用这些画像来实施个性化教育，特别是如何通过改变环境和教学方式来改善学生的学习习惯。此外，我们还需进一步探讨不同学生画像之间的动态变化，以及学生画像与其学术成绩的关联。

总体而言，本研究不仅提供了一种有效的方法来预测学生学习习惯，还为理解和改善学生学习过程提供了有益的理论支撑，开启了学生画像研究的新篇章。然而，我们仍需要在更

大规模和更多元化的样本上进一步验证和改进模型的有效性和稳定性，以便更深入地理解学生的学习习惯，并更好地指导教育实践。

参考文献：

[1] 邱文雅.基于大数据技术的高校学生信息管理平台研究[J].信息技术与信息化,2023(3):29-32.

[2] 张明丽,丁月华.基于大数据画像的个性化创新创业教育模式[J].高等工程教育研究,2023(2):183-189.

[3] 励凌凌,孙澄宇.基于大数据的职业院校学生画像系统构建[J].职教通讯,2023(2):79-86.

[4] 刘欢.基于可视化的教育数据挖掘综述[J].现代计算机,2023,29(4):60-63,74.

[5] 黄建国.职业能力导向的高职院校智慧学习模型构建及应用研究[D].长春:东北师范大学,2023.

[6] 王兵兵.基于学生画像的线上教学模式构建与动态优化策略探讨:以高职数学为例[J].科幻画报,2022(9):267-268.

[7] 黄炜,张治,胡爱花,等.基于"五育融合"的学生数字画像构建与实践分析[J].教育发展研究,2021,41(18):44-51.

[8] SCHULDT C,LAPTEV I,CAPUTO B. Recognizing human actions:a local SVM approach[C]//Proceedings of the 17th International Conference on Pattern Recognition,2004,3:32-36.

[9] CHERKASSKY V, MA Y. Practical selection of SVM parameters and noise estimation for SVM regression[J]. Neural Networks,2004,17(1):113-126.

[10] 赵卫,刘渊.基于深度学习与有向无环图SVM的局部调整年龄估计[J].计算机应用与软件,2023,40(1):189-195.

[11] GUO G, WANG H, BELL D, et al. KNN model-based approach in classification[C]//On the Move to Meaningful Internet Systems 2003:CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003,Catania,Sicily,Italy,2003:986-996.

[12] ZHANG H, BERG A C, MAIRE M, et al. SVM-KNN:discriminative nearest neighbor classification for visual category recognition[C]//Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06),2006,2:2126-2136.

[13] 高璇.应用自然邻居分类算法的大学生就业预测模型[D].重庆:重庆大学,2017.

[14] KARLIK B, OLGAC A V. Performance analysis of various activation functions in generalized MLP architectures of neural networks[J]. International Journal of Artificial Intelligence and Expert Systems, 2011, 1(4):111-122.

[15] TOLSTIKHIN I O,HOULSBY N,KOLESNIKOV A, et al. MLP-mixer:an all-MLP architecture for vision[C]//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021),2021,34:24261-24272.

[16] 申宇.头面部特征提取及基于MLP神经网络的号型预测[D].杭州:浙江理工大学,2022.

Construction of student learning portrait based on multi-dimensional data mining

Xu Huihui*

(Department of Instrument Engineering, Shanxi Pharmaceutical Vocational College, Taiyuan 030031, China)

Abstract: Based on multidimensional data, this study aims to apply data mining technology to predict students' learning habits and build learning profiles. By collecting data from 273 students, including interpersonal relationships, personality traits, and health status, and after data preprocessing, a multidimensional data mining model based on support vector machines, K-nearest neighbors, and multilayer perceptrons was successfully established, enabling prediction of students' learning habits. The research results present a comprehensive student profile, including aspects such as time management and learning methods. This study confirms the importance of data-driven decision-making in the field of education and provides a scientific basis for offering personalized education plans.

Keywords: data mining; student profile; support vector machine; K-nearest neighbors; multilayer perceptron

文章编号: 1007-1423(2023)14-0052-05 DOI: 10.3969/j.issn.1007-1423.2023.14.010

基于强化学习DQN算法的智能决策模型研究

韩中华*

（北方工业大学理学院，北京 100144）

摘要：针对强化学习DQN算法的三个优化因子(即Dueling、Double-Q以及Prioritized-replay)之间是否存在相互促进或抑制的关系，对三个优化因子之间进行随意组合作为交易策略进行研究，并将2020年9月2日至2022年9月2日期间雅虎金融网站上的HDFC银行股票的收盘价作为研究对象。研究结果发现，相较于基线模型，Dueling对股票短期收益预测最为贴合实际，并且对Double-Q与Prioritized-replay起到了促进作用；Prioritized-replay对Double-Q与Dueling起到了抑制作用，而Double-Q则对Prioritized-replay与Dueling未起到显著性改变。鉴于DQN算法在股票短期收益预测的随机性与预测精度的问题，其未来在金融预测领域将会有更好的应用前景。

关键词：DQN算法；深度学习；股票收益预测

0 引言

众所周知，作为毫无规律可循的时间序列数据，提升股票数据的预测精度一直备受金融领域的关注。然而，随着大数据与人工智能领域的不断发展，利用计算机实现自我学习以及大批量计算，对预测准度的提升不容小觑。其中，强化学习作为一种倍受瞩目的机器学习方法，它利用与环境交互过程中的奖励函数激励计算机学习，模拟人脑在学习过程当中的特点。自Alpha Go以4∶1击败著名围棋选手李世石后，强化学习瞬间掀起了人工智能领域的研究浪潮，这一技术特点被应用在包括金融在内的不同领域，进而凸显其强大的学习能力。因此，强化学习将在未来广泛应用于金融领域。

目前，强化学习在金融股票投资组合这一领域应用十分广泛。例如，为了验证DDPG算法的有效性，齐岳等[1]以我国股市为例，发现利用强化学习方法构建的投资组合，在实验期间的价值增幅远大于对照组，进而得以验证。焦禹铭[2]通过对比A2C、PPO与SAC三个模型，发现SAC模型无论是在获取收益还是在风险控制方面的优势最为显著。考虑到股票价格变化受到多种因素影响，陈诗乐等[3]将遗传算法(GA)与Transformer模型结合，提出GA-Transformer这一组合模型，该组合模型要优于传统单模型的效果。

本文将DQN(Deep Q Network)算法作为基线，将三个不同优化层面(Double-Q、Prioritized-replay以及Dueling)随机组合成七个优化模型对HDFC银行(HDB)历史交易数据进行预测，主要贡献在于：①找出三种优化方案分别对DQN算法预测的影响作用；②对比三个优化层面之间的联系以及影响程度；③对短期股票收益预测产生的随机性进行多次实验取平均值，与基线模型进行预期收益平均水平与波动情况的对比分析。

1 模型构建

1.1 DQN算法思想及其优势

相较于传统的Q-learning算法，DQN算法解决了其Q-table无法使用在状态与动作为连续的高维空间中。通过将Q-table更新替换为一个函数拟合的问题，以此来获得处于相近状态的输出动作。这样，将Deep Learning(DL)与Reinforcement Learning(RL)两者相结合，便可获得DQN。

收稿日期：2023-03-13 **修稿日期：**2023-03-28

作者简介：*通信作者：韩中华(1999—)，男，辽宁海城人，硕士，研究方向为基于强化学习算法的量化交易，E-mail：2021312060103@mail.ncut.edu.cn

在DQN算法中，存在下列问题：首先，由于DL中存在噪声与延迟性，导致许多状态与奖励的值普遍为0；其次，尽管DL当中每个样本之间相互独立，但是RL当前状态的返回值则依赖后续状态的返回值；最后，当使用非线性表达式表示值函数时，可能会出现不稳定的情况。

针对上述问题，DQN通过以下方面进行优化：首先，Q-learning中使用奖励来构造标签；其次，采用经验池(Prioritized-replay)来解决相关性以及非静态分布的问题；最后，采用主网络(Main Net)产生当前状态的Q值，使用另一个目标网络(Target Net)产生相应的目标Q值。

1.2　DQN神经网络的构建

由图1可得，在使用Tensorflow来实现DQN的过程中，一般选用两个神经网络进行搭建。其中target_net用于预测Q_target的值，但它并不会及时更新其中的参数；eval_net用于预测Q_eval，该神经网络拥有最新的神经网络参数。这两个神经网络的结构是完全一致，只是其中的参数并不一样。

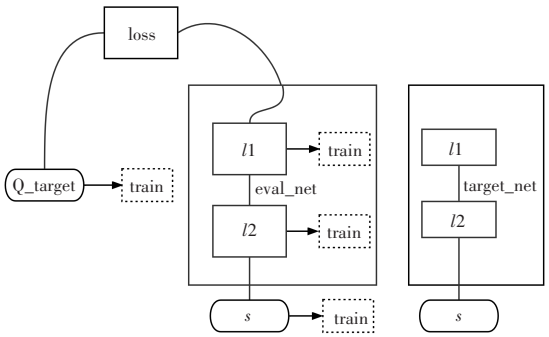


图1　DQN网络

除了传统的DQN方法之外，可以通过对其进行优化，得到下列三个DQN的优化算法：①Double DQN：解决了原本DQN算法中Q_{max}导致$Q_{reality}$当中的过估计问题；②Prioritized-replay DQN：在训练过程中对奖励较高或者较低的值加以重视；③Dueling DQN：将每个状态的Q值分解成每个状态的Value加上每个动作的advantage。

本文将DQN算法作为基准，与七个组合模型的交易策略进行比较。表1为DQN算法与三个优化因子之间的组合情况。

表1　三个优化因子与DQN算法的任意组合

模型名称	Double-Q	Prioritized-replay	dueling
Double DQN	√	×	×
Prioritized-replay DQN	×	√	×
Dueling DQN	×	×	√
Double Prioritized-replay DQN	√	√	×
Double Dueling DQN	√	×	√
Prioritized-replay Dueling DQN	×	√	√
Full DQN	√	√	√
DQN(baseline)	×	×	×

2　数据来源及样本选择

本文选取的股票数据取自雅虎金融(yahoo finance)网站，选取的对象为HDFC Bank(HDB)。本文设定的时间区间为2020年9月2日至2022年9月2日这两年期间所有交易日的历史价格与成交量数据，包含最高价(high)、最低价(low)、开盘价(open)、收盘价(close)、成交量(volume)以及调整后的收盘价(adj close)共计505条。图2是历史数据集的总体趋势图。

图2　HDB股票2020年9月2日到2022年9月2日历史收盘价轨迹

从图2中，大致可以看出HDB股票收盘价的历史变化趋势。从2020年10月到2021年4月保持上升趋势，但2021年4月到2022年7月之后整体呈下降趋势，2022年7月到2022年9月有所回暖。

3　实证研究

将原始数据分为以下两个部分。其中，训练集选取数据集中的前15～450条数据，从第450～505条数据则作为测试集来评估模型。

为了控制其他变量的影响，将固定移动预测步长(window_size)为15天，即预测下一天股票短期收益情况取决于前15天的历史数据；将总时间步数(*total_timesteps* = 150000)，即循环训练的总时间步长为150000次。

本文展开分析的角度共有两个。首先，通过七种组合方法与单模型DQN这一基线方法的对比，寻找出哪一个组合模型对短期股票预测效果较为贴切实际；其次，通过最终预测的短期股票收益情况，以此来确定这三种优化方面之间的相互促进或抑制作用。

3.1 七个组合模型以及基线模型的预测情况

在原始DQN算法的基础之上，可以优化的方面为以下三个，分别是：Double-Q，Prioritized-replay以及Dueling。通过这三个优化因子的任意随机组合，将产生七个组合模型以及一个基线模型。将Double DQN、Prioritized-replay DQN、Dueling DQN、Prioritized- replay Double DQN、Dueling Double DQN、Prioritized- replay Dueling DQN、Full DQN以及DQN(baseline)模型分别标记为1～8。考虑到每一次评估模型的预期涨跌情况为一随机数，选取20次实验的结果以避免偶然性发生。表2为这七个组合模型(序号1～7)以及基线模型(序号8)在测试集上最终持有资产相较于原始投入资产所占比重。

对表2中的七个组合模型与基线模型DQN算法的20次结果计算对应的平均资产变动情况(即平均值)与波动幅度(即标准差)。表3为这七个组合模型与DQN这一单模型的股票预期平均资产变动与波动情况。

表2 七个组合模型以及基线模型在测试集上最终持有资产占原始投入资产比重

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1（比重/%）	114.16	99.77	101.44	98.72	114.54	89.47	100.00	114.16	88.74	104.37	89.81	97.13	90.35	95.81	100.00	98.95	100.00	101.04	88.38	104.56
2（比重/%）	99.14	87.94	92.44	96.94	92.29	96.84	95.80	89.31	95.70	84.88	91.29	97.94	98.40	102.24	102.07	92.95	89.54	99.24	97.93	97.30
3（比重/%）	100.00	100.00	102.29	100.00	107.93	99.92	100.00	91.80	114.16	99.87	82.37	101.30	101.00	92.09	100.00	111.34	104.72	102.93	114.16	114.16
4（比重/%）	91.14	93.19	92.53	97.68	101.95	88.33	95.13	103.46	94.52	91.40	86.17	100.24	99.91	87.94	100.33	104.28	96.95	87.50	96.61	100.95
5（比重/%）	94.99	100.00	114.16	92.35	90.64	100.00	114.16	96.29	100.00	100.00	114.16	100.00	101.96	100.00	100.00	87.99	106.41	111.85	100.00	114.16
6（比重/%）	98.01	100.00	114.16	91.09	100.00	114.16	100.00	111.34	114.16	113.68	97.44	100.00	85.09	89.43	100.00	100.03	100.00	90.79	100.00	100.00
7（比重/%）	95.53	111.34	100.00	99.08	97.02	104.31	100.00	114.16	89.77	95.59	101.07	88.37	90.27	100.00	106.39	92.45	95.41	93.09	100.00	95.37
8（比重/%）	91.52	99.75	93.58	96.76	95.29	97.36	88.77	88.91	92.82	84.77	88.84	98.36	85.69	99.11	92.47	96.92	94.78	88.95	102.63	91.44

通过表3，可以发现在前三种加入单一优化方面组合算法当中，只有Dueling DQN这一组合算法的现有资产较原始投入资产上涨幅度最大，为2.00%。因此，说明Dueling DQN这一方法在预测股票短期收益方面较为贴合实际情况，其精度较高。其次则为Dueling Double DQN模型，其现持有资产相较于原始投入资产上涨1.96%，说明在原本的Dueling DQN模型中加入Double-Q后未对原始模型起到显著性的抑制作用，基本保持了原有的收益水平。最后则为Prioritized-replay Dueling DQN这一组合算法，相较于原始投入资产，其现有资产上涨了0.97%，说明在原

本的 Dueling DQN 模型中加入 Prioritized-replay 后，未对原本的Dueling DQN算法起到任何促进作用，仅仅保持了微弱的收益幅度。由此说明，在这三个优化因子中，Dueling这一优化因子的促进作用最强，其次是Double-Q，最后则为Prioritized-replay。

表 3　DQN算法及其组合模型的股票测试集上预期资产变动与波动情况

模型名称	现持有资产/原有资产/%	平均盈损率/%	标准差
Double DQN	99.57	−0.43	0.0811
Prioritized-replay DQN	95.01	−4.99	0.0474
Dueling DQN	102.00	+2.00	0.0792
Prioritized-replay Double DQN	95.51	−4.49	0.0560
Dueling Double DQN	101.96	+1.96	0.0808
Prioritized-replay Dueling DQN	100.97	+0.97	0.0861
Full DQN	98.46	−1.54	0.0678
DQN(baseline)	93.44	−6.56	0.0488

从最后一列的标准差，可以看出，相较于原本的DQN模型，只有加入Prioritized-replay这一优化因子后，标准差由原本的0.0488降低至0.0474，由此可见波动幅度有小幅度的降低。说明Prioritized-replay这一优化因子对DQN模型的稳定性具有促进作用。而Dueling DQN模型相较于原本的DQN模型，其标准差上涨了0.0304。而该上涨幅度小于Double DQN模型相较于原本的DQN模型的0.0323。由此可见，在这三个优化因子中，对DQN模型稳定性起到促进作用的为Prioritized-replay，对模型稳定性起到抑制作用的为Dueling以及Double-Q。

3.2　三个优化因子之间的关系

表4汇总了本文七个组合模型在测试集上的各项性能的表现，包括年华化收益率与夏普比率。

本文计算年化收益率的计算公式为：

$$\text{年化收益率}=\frac{\text{现持有资产/原有资产}}{\text{测试集天数}}\times 365\times 100\%$$

其中："现持有资产/原有资产"选取表3中七个组合模型的结果，测试集天数为55(即505−450=55)。最终结果如表4所示。

表 4　七个组合算法的回测结果(测试集)

模型名称	年化收益率/%	夏普比率
Double DQN	−2.85	0.7559
Prioritized-replay DQN	−33.12	0.3312
Dueling DQN	13.27	1.0808
Prioritized-replay Double DQN	−29.80	0.3696
Dueling Double DQN	13.01	1.0545
Prioritized-replay Dueling DQN	6.44	0.8746
Full DQN	−10.22	0.7404

对于夏普比率的计算，本文基于原本的公式进行了调整与改动。由于原始公式中包含无风险利率，本文将无风险利率替换为表3中DQN模型的现持有资产占原有资产的平均比重，为93.44%；而原始公式中的投资组合预期报酬率与其标准差也相应地替换为表3中七个组合模型的现持有资产占原有资产的平均比重以及标准差。本文计算夏普比率的公式为：

$$\text{夏普比率}=\frac{\text{组合模型现有资产/原有资产比重}}{\text{组合模型现有资产占原有资产比重的标准差}}-\frac{\textit{DQN}\text{现有资产/原有资产比重}}{\text{组合模型现有资产占原有资产比重的标准差}}$$

在年化收益率方面，基于Dueling DQN模型在测试集上达到了13.27%，基于Dueling Double DQN模型在测试集上达到了13.01%，基于Prioritized-replay Dueling DQN算法在测试集上达到了6.44%。这三个模型均为在回测中有所盈利的，剩下的模型则均有所亏损。从整体上看，在原本的Double DQN与Prioritized-replay DQN模型中分别加入Dueling这一优化因子后，其年化收益率均有显著性提升；在原本的Double DQN模型与Dueling DQN模型中加入Prioritized-replay这一优化因子后，年化收益率有明显的负增长；在原本的Prioritized-replay DQN模型与Dueling DQN模型中加入Double-Q这一优化因子后，年化收益率未起到了显著性改变。

在夏普比率方面，基于Dueling DQN算法在测试集上达到了1.0808，基于Dueling Double DQN算法达到了1.0545，剩下的组合模型均小于1。由此可见，Dueling与Double-Q这两个优化因子对DQN模型均起到了促进作用，并且将

二者同时加入模型中不会对模型起到显著性的抑制作用。

4 结论及建议

本文通过将DQN算法结合三个不同优化方面的七个相互组合模型，与单一模型DQN算法进行对比，得出以下结论：①这七个组合模型中，只有Dueling DQN算法的盈利率达到了最高，为2.00%，说明该模型对股票短期收益预测结果最为贴合实际；②这三个优化方面之间，Dueling对Double-Q与Prioritized-replay均起到了显著性促进作用，而Prioritized-replay对Double-Q与Dueling均起到了显著性抑制作用，Double-Q则对Prioritized-replay与Dueling未起到显著性改变；③将三个优化方面对DQN算法共同优化，相较于单一模型DQN算法而言，盈亏率有所上升，由此可以得出三个优化方面之间存在相互独立的作用。

本文在股票短期收益预测方面利用深度学习当中的DQN算法进行探索，针对不同三个优化方面进行对比验证，为后续将深度学习应用于金融领域奠定了基础。鉴于短期预测的不稳定性，在未来后续研究中仍有可以优化提升的方面。个人认为后续研究还可以在两大主题上进行展开：①引入较为稳定的预测方案及优化技巧的前沿方案，以此来抑制随机性带来的影响；②可将该DQN算法应用在金融其他领域，例如投资领域，以此来促进我国金融领域智能化大力发展。

参考文献：

[1] 齐岳,黄硕华. 基于深度强化学习DDPG算法的投资组合管理[J]. 计算机与现代化,2018(5):93-99.

[2] 焦禹铭. 基于深度强化学习的股票投资组合管理及实证研究[D]. 西安:西北大学,2021.

[3] 陈诗乐,王笑,周昌军. 基于GA-Transformer模型的多因子股票预测[J]. 广州大学学报(自然科学版),2021,20(1):44-55.

Research on intelligent decision making model based on reinforcement learning DQN

Han Zhonghua*

(North China University of Technology College of Science, Beijing 100144, China)

Abstract: Aiming at whether there is a relationship of mutual promotion or inhibition among the three optimization factors of the reinforcement learning DQN algorithm (namely Dueling, Double-Q, and Prioritized-replay), the random combination of the three optimization factors is studied as a trading strategy, and From September 2, 2020 to September 2, 2022, the closing price of HDFC (Housing Development Finance Corporation) bank (HDB for short) stock on the Yahoo Finance website was used as the research object. The results of the study found that, compared with the baseline model, Dueling is the most realistic for short-term stock return forecasts, and has promoted Double-Q and Prioritized-replay; Prioritized-replay has inhibited Double-Q and Dueling, while Double-Q did not significantly change Prioritized-replay and Dueling. In view of the randomness and prediction accuracy of the DQN algorithm in the short-term stock return prediction, it will have a better application prospect in the field of financial prediction in the future.

Keywords: DQN algorithm; deep learning; stock return forecast

文章编号: 1007-1423(2023)14-0057-05 DOI: 10.3969/j.issn.1007-1423.2023.14.011

基于Spark平台的恶意软件最大频繁子图挖掘方法

周显春*，肖 衡，焦萍萍，邹琴琴

（三亚学院信息与智能工程学院，三亚 572022）

摘要：为了解决传统子图挖掘算法时效性差的问题，设计了一种基于Spark平台的恶意软件最大频繁子图挖掘方法。该方法在保证挖掘信息完整的前提下，避免了挖掘所有频繁子图，采用了改进的FSMBUS方法来挖掘恶意软件的最大频繁子图，利用分布式架构Spark迭代计算优势，提高了挖掘效率。此外，改进算法还被应用于恶意软件同源性判定，改善了恶意软件检测效果。最后，通过对比实验结果，论证了该方法的高效性和可行性。

关键词：恶意软件；最大频繁子图；任务并行化；Spark

0 引言

子图挖掘(subgraph mining)是图数据挖掘的一个重要分支，它已经成为数据挖掘中的一个研究热点[1]，旨在从大规模复杂图数据中发现有意义的子图模式。随着社交网络、互联网、生物信息学等领域的快速发展，大量的图数据已经成为现代科学研究的重要来源，在社交网络中可以发现社交网络中的社区结构、关系网络等信息，在推进系统中可以发现用户间的相关关系、用户兴趣等信息，从而提高推荐系统的准确性，在网络安全中可以发现网络攻击行为、恶意软件等信息。但是，这些数据通常是非常复杂和庞大的，很难从中提取有意义的信息。

1 图数据挖掘算法相关研究

目前，图数据挖掘算法可以分为三类：针对图集合、针对单个大图[2]及分布式频繁子图挖掘。

在针对图集合的频繁子图挖掘中，可以使用邻接矩阵表示图数据，并使用基于Apriori算法的改进算法，如AGM[3]、FSG[4]和MARGIN[5]，或者基于DFS的改进算法gSpan[6]来挖掘频繁子图。其中，gSpan算法能够支持对边缘属性和标签等更丰富的信息进行挖掘，但在大规模数据集上性能可能受到影响。为此，UBW-gSpan算法[7]通过引入最小支持度阈值和最大标签数目阈值两个参数进行优化。

在针对单个大图的频繁子图挖掘中，SUBDUE[8]、SIGRAM[9]算法和GRAMI[10]算法是常用的算法。SUBDUE算法利用图匹配和图压缩进行子图挖掘，能够有效挖掘大规模图中的频繁子图，并且不需要事先指定子图的大小，但存在不能处理具有变化的图数据集以及可能生成重复子图的缺点。SIGRAM算法采用基于图同构的方法进行子图匹配，将每个图转换为一个特征向量来捕获其结构信息，并使用随机映射来降低计算复杂度。GRAMI[10]算法将图形模式挖掘问题转化为限制约束问题，从而实现高效的挖掘，但需要注意参数选择、候选生成复杂度和内存占用等缺点。

收稿日期：2023-03-28 **修稿日期：**2023-06-20

基金项目：海南省自然科学基金项目资助(620MS064)；三亚市院地科技合作项目资助(2019YD26)；三亚学院优势专业建设项目(SYJZUS202203)；三亚学院一流本科专业特色建设资助项目(SYJZZZ202212)

作者简介：*通信作者：周显春(1974—)，男，湖南常德人，副教授，工学硕士，研究方向为网络安全及数据挖掘，E-mail:354813786@qq.com；肖衡(1979—)，女，湖南衡阳人，硕士，副教授，研究方向为计算机网络通信、人工智能；焦萍萍(1983—)，女，江西景德镇人，硕士，讲师，研究方向为通信技术与嵌入式系统；邹琴琴(1982—)，女，海南三亚人，讲师，硕士，主要研究方向为大数据与学习分析

分布式频繁子图挖掘算法旨在Hadoop[11]/Spark[12-13]框架大规模分布式系统中挖掘频繁子图。该算法将数据集划分为多个分区，在每个分区上运行子图挖掘算法，最后将每个分区的结果进行合并，得到全局频繁子图。FSMBUS算法[14-15]通过分布式计算和分治策略来解决这些问题，减少了搜索空间，提高了挖掘效率，缺点是在处理大规模图数据库时会面临计算复杂度高的问题。

为了解决传统子图挖掘算法时效性差的问题，提出了一种基于Spark分布式平台的恶意软件最大频繁子图挖掘方法，利用改进FSMBUS算法的优点，能够利用子结构之间的相似性进行最大频繁子图挖掘，从而减少搜索空间并提高挖掘效率。该算法应用于恶意软件同源性判定，发现恶意软件中存在的共同子结构，以改善恶意软件检测的效果。通过实验验证了改进算法的效果。

2 最大频繁子图提取方法

2.1 提取方法的架构

最大频繁子图提取方法分为两个阶段：数据预处理和数据挖掘，流程如图1所示。数据预处理负责从恶意代码中提取程序依赖图；数据挖掘是核心部分，包括任务并行化、子图挖掘、子图剪枝、CSP模型判断[16]、是否存在超图等阶段。Master结点通过并行化把任务分解并分配给Worker结点。子图在每个Worker节点上对子图扩展进行迭代计算，通过频繁度比较完成边的扩展和剪枝。最后，在主结点通过判定频繁子图是否存在超图得到最大频繁子图(图2)。

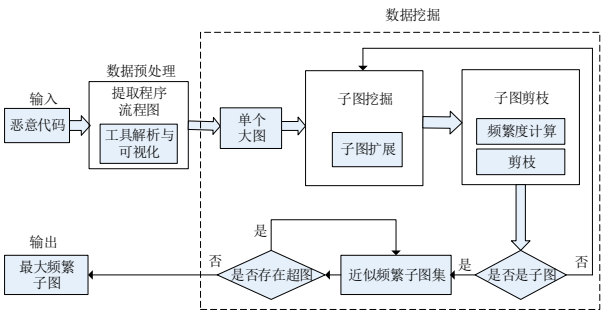


图1 最大频繁子图提取流程

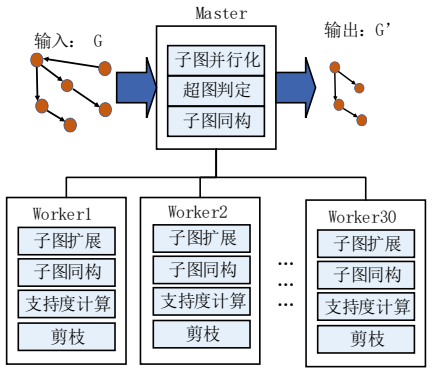


图2 基于Spark集群最大频繁子图提取架构

2.2 函数调用图提取

利用调试工具运行被混淆的代码，并在运行过程中监视其行为和状态，通过分析运行时数据和代码执行流程，提取函数调用图。

2.3 基于Spark的改进FSMB算法

2.3.1 FSMBUS算法[13]

它是一种基于状态机的流式数据挖掘算法，在候选频繁子串的每个位置中，查找所有长度为$k-1$的子串是否出现在之前已经确定为频繁子串的位置中，如果未出现的次数小于最小支持度阈值，则需要剪枝。剪枝可以有效地减少计算量，提高算法的效率。其伪代码如下：

输入：L：一个列表，包含所有长度为$k-1$的频繁子串；S：一个字符串列表，表示序列集合；P：一个候选频繁子串，用于判断是否需要剪枝；*threshold*：频繁子图的最小支持度阈值。

输出：*prune*：一个布尔值，表示是否需要对该候选频繁子串进行剪枝。

(1) 初始化一个计数器*count*为0

(2) 对于每个字符串$s \in S$，执行以下步骤：

1) 在s中使用后缀数组和后缀LCP数组找到所有长度大于等于$|P|$的子串，记为Si；

2) 在Si中使用位向量和比特位并行操作找到所有与P匹配的子串，记为Mi；

3) 将Mi中所有与P相等的子串对应的位置记录在一个集合中，记为Pi；

4) 对于Pi中的每个位置p，执行以下步骤：

(a) 将p转化为Pi的编号，记为*pid*

(b) 在L中查找长度为$k-1$的子串是否出现在*pid*中，如果出现，则将*count*加1

(3) 如果*count*小于*threshold*，则将*prune*设为True，

否则设为 False

(4) 返回 *prune* 的值

2.3.2　创建给定子图的所有不同超图的伪代码

输入：一个子图 $g(V', E')$，原图 $G(V, E)$

输出：子图 g 的所有不同超图的列表

(1) 初始化超图列表 H 为包含 V' 的初始超图

(2) 对于原图 G 中 $V'-|V'|+1$ 个节点的子集 S：

1) 如果 S 中至少包含 g 的所有节点，则创建一个新的超图 G'

2) 将 S 中的所有节点添加到 G' 的节点集合中

3) 对于 g 的每条边 $(e1, e2)$，如果 S 中同时包含 $e1$ 和 $e2$ 中的所有节点，则将它们添加到 G' 的超边集合中

4) 如果 G' 不是 H 中的任何超图的子集，则将 G′ 添加到 H 中

(3) 返回超图列表 H

2.4　基于 GraphFrames 的大规模单图上的子图匹配算法

GraphFrames 是一种基于 DataFrame 和 GraphX 的图处理库，提供了方便的 API 来实现图分析任务。包括加载大规模单图和查询子图、子图查询、子图匹配、结果输出。其伪代码如下：

(1) 读取原始图 G 和查询图 Q

(2) 对 G 和 Q 进行顶点和边的特征提取

(3) 在 G 中寻找与 Q 相同的顶点集合 V

(4) 对每个 V 中的顶点，寻找与 Q 相同的邻居结构，生成候选子图

(5) 将候选子图转换为 GraphFrames 格式，构建候选子图集合 CG

(6) 使用 GraphFrames 中的 find() 对 CG 进行子图同构匹配，得到匹配的子图集合 M

(7) 对 M 进行筛选和排序，输出最优的匹配子图

3　实验结果及分析

实验使用虚拟集群，虚拟机 PC 机 31 台，其中 1 台 master 主节点，30 台为 slave 从节点。每台 PC 机的配置相同：操作系统为 Centos7.9，CPU 为 Intel Core i7 3.8 GHz，16 GB 内存，使用 Scala 2.12.17 开发，集群环境建立在 Hadoop3.3.4、Spark3.3.1、jdk 1.8 上。

3.1　数据集

http://malware.lu 确实是一个公认的恶意代码样本库，一共有 4964137 个恶意软件样本，其中包含各种类型的恶意代码家族，例如病毒、蠕虫、木马、后门、根包和间谍软件等，大部分经过了加壳处理。每个实验样本，Mytob、Netsky、Allaple、Bagle、Mydoom、Agobot/Gaobot，均有 3000 个。

3.2　评价指标

采用最大频繁子图挖掘时间、恶意软件检测的准确率(PR)、误报率(FPR)、漏报率(FNR)四个指标对模型的检测效果进行评价。

3.3　实验结果及分析

3.3.1　支持度对运行时间的影响

由图 3 可知，MapReduce 方法挖掘最大频繁子图的运行时间较多，与改进 FSMBUS、传统 FSMBUS 算法相比，时间要多 2～4 倍。在支持度的值为 8.0 之后，三种分布式算法的运行时间基本稳定。改进 FSMBUS 算法运行时间稳定在 3.7 ms 左右，传统 FSMBUS 算法运行时间稳定在 2.7 ms 左右，说明了两种算法比较稳定。改进 FSMBUS 算法比传统 FSMBUS 算法多了 1 ms 左右，主要原因是传统 FSMBUS 算法是进行频繁子图挖掘，而改进 FSMBUS 算法是应用于挖掘最大频繁子图，与传统 FSMBUS 算法多了一个超图运算操作。

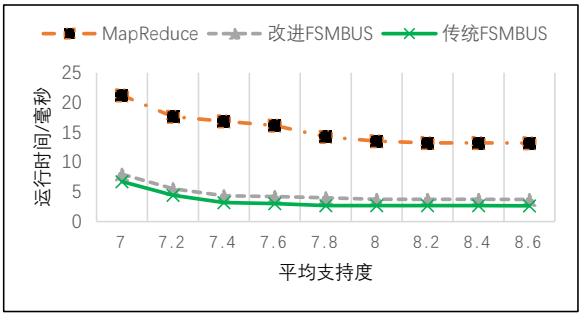


图 3　支持度与系统运行时间的关系

3.3.2　运行时间的分析

如表 1 所示，三种方法的最大频繁子图挖掘时间的对比，可以发现基于 Hadoopk 的方法频繁子图挖掘时间是基于 Spark 的方法的 3～4 倍；基于 Spark 的改进 FSMBUS 方法进行最大频繁子图挖掘时间要比 Top-Down 算法快 20～60 ms，平均

减少时间32 ms，说明了本文的方法对大图挖掘最大频繁子图的效果较好。

表1 最大频繁子图挖掘时间

恶意代码家族	挖掘时间/ms		
	Top-Down算法[17]	改进FSMBUS	MapReduce[11]
Allaple	157.3	132.4	421.3
Mytob	174.2	142.5	512.4
Mydoom	212.4	165.2	620.1
Bagle	225.3	168.4	654.2
Netsky	172.6	158.4	501.6
Agobot	145.2	125.7	432.5
平均值	181.17	148.77	523.68

3.3.3 恶意软件识别效果的分析

由表2可知，改进FSMBUS相较于传统FSMBUS方法，其平均准确率提高了1.3个百分点、平均误报率降低了3.6%个百分点；相对于MapReduce方法，改进FSMBUS的平均准确率提高了1.7个百分点、平均误报率降低了1.8个百分点。并且改进FSMBUS方法对大部分恶意软件的检测效果比较稳定，不仅说明最大频繁子图可以作为恶意软件的检测特征，而且对恶意软件的检测效果很好，平均准确率和平均误报率分别为95.6%和4.4%。

表2 传统FSMBUS、改进FSMBUS和MapReduce检测效果/%

恶意代码家族	传统FSMBUS			改进FSMBUS			MapReduce		
	准确率	误报率	漏报率	准确率	误报率	漏报率	准确率	误报率	漏报率
Allaple	96.2	3.8	0	97.8	2.2	0	95.8	4.2	0
Mytob	95.5	18.5	0	96.5	3.5	0	95.9	4.1	0
Mydoom	93.8	6.2	0	95.2	4.8	0	94.1	5.9	0
Bagle	94.8	5.2	0	96.2	3.8	0	94.6	5.4	0
Netsky	91.1	8.9	0	92.3	7.7	0	90.3	9.7	0
Agobot	94.5	5.5	0	95.6	4.4	0	92.7	7.7	0
平均值	94.3	8.0	0	95.6	4.4	0	93.9	6.2	0

4 结语

为了解决传统子图挖掘算法时效性低的问题，本文提出了一种基于Spark架构的恶意软件最大频繁子图挖掘方法。该方法采用次优树数据结构进行并行化处理，利用GRAMI算法计算支持度，并采用次优树连接和扩展边的方式获取最大频繁子图，从而提高挖掘效率。利用改进的FSMBUS方法在Spark分布式平台上进行实现，用于挖掘恶意软件中的最大频繁子图。实验结果表明，该方法相对于Top-Down算法平均挖掘时间减少了32 ms，与基于频繁子图的恶意软件检测相比，平均准确率提高了1.3个百分点、平均误报率降低了了3.6个百分点。子图相似匹配是最大频繁子图挖掘的关键，因此是下一步研究的重点和方向。

参考文献：

[1] HAN J, KAMBER M, PEI J. 数据挖掘：概念与技术[M]. 韩家炜，译. 北京：机械工业出版社，2007：79-165.

[2] KURAMOCHI M, KARYPIS G. Finding frequent patterns in a large sparse graph[J]. Data Mining and Knowledge Discovery, 2005, 11(3): 243-271.

[3] INOKUCHI A, WASHIO T, MOTODA H. An apriori-based algorithm for mining frequent substructures from graph data[C]//Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, Lyon, France, 2000: 13-23.

[4] KURAMOCHI M, KARYPIS G. Frequent subgraph discovery[C]//Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 2001: 313-320.

[5] THOMAS L T, VALLURI S R, KARLAPALEM K. Margin: maximal frequent subgraph mining[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2010, 4(3): 1-42.

[6] YAN X, HAN J. gSpan: graph-based substructure pattern mining[C]//Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, 2002: 721-724.

[7] KOURFOGIANNIS F, PAPPAS G J. Differential privacy for dynamical sensitive data[C]//Proceedings of the 2017 IEEE 56th Annual Conference on Decision and Control (CDC), Melbourne, VIC, Australia,

2017:1118-1125.
[8] HOLDER L B, COOK D J, DJOKO S. Substucture discovery in the SUBDUE system[C]//AAAIWS'94: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, Seattle WA, 1994:169-180.
[9] RANU S, SINGH A K. GraphSig: a scalable approach to mining significant subgraphs in large graph databases[C]//Proceedings of the IEEE International Conference on Data Engineering, Shanghai, China, 2009:844-855.
[10] ELSEIDY M, ABDELHAMID E, SKIADOPOULOS S, et al.GRAMI: Frequent subgraph and pattern mining in a single large graph [J]. Proceedings of the VLDB Endowment, 2014, 7(7):517-528.
[11] LU W, CHEN G, TUNG A K H, et al. Efficiently extracting frequent subgraphs using MapReduce[C]//Proceedings of the 2013 IEEE International Conference on Big Data, CA, USA, 2013:639-647.
[12] NIU X Z, NIU J J, SU D Z. FP-tree-based approach for frequent trajectory pattern mining[J]. Journal of the University of Electronic Science and Technology of China, 2016, 45(1):86-90.
[13] 严玉良，董一鸿，何贤芒，等. FSMBUS：一种基于Spark的大规模频繁子图挖掘算法[J]. 计算机研究与发展，2015，52(8)：1768-1783.
[14] 郭方方，王欣悦，王慧强，等. 基于最大频繁子图挖掘的动态污点分析方法[J]. 计算机研究与发展，2020，57(3)：631-638.
[15] BHATIA V, RANI R. Ap-FSM: a parallel algorithm for approximate frequent subgraph mining using Pregel[J]. Expert Systems with Applications, 2018, 106: 217-232.
[16] NIKOLOPOULOS S D, POLENAKIS I. A graph-based model for malicious code detection exploiting dependencies of system-call groups[C]//Proceedings of the 16th International Conference on Computer Systems and Technologies, Dublin, lreland, 2015: 228-235.
[17] 柴然，刘媛媛，郭彦颖. 最大频繁子图挖掘DMFS[J]. 中国管理信息化，2017(4)：153-154.

Maximum frequent subgraph mining method for malware based on Spark

Zhou Xianchun*, Xiao Heng , Jiao Pingping , Zou Qinqin

(School of Information and Intelligence Engineering, Sanya University, Sanya 572022, China)

Abstract: In order to solve the problem of poor timeliness in traditional subgraph mining algorithms, a malicious software maximum frequent subgraph mining method based on the Spark platform was designed. This method avoids mining all frequent subgraphs while ensuring the integrity of mining information. It adopts an improved FSMBUS method to mine the maximum frequent subgraphs of malicious software, utilizing the advantages of distributed architecture Spark iterative computation to improve mining efficiency. In addition, the improved algorithm has also been applied to malware homology determination, improving the effectiveness of malware detection. Finally, by comparing the experimental results, the efficiency and feasibility of this method were demonstrated.

Keywords: malware; maximum frequent subgraph; task parallelization; Spark

文章编号: 1007-1423（2023）14-0062-05 DOI: 10.3969/j.issn.1007-1423.2023.14.012

基于多维关系和用户聚类的智慧图书馆个性化图书推荐研究

闫俊辉*

（运城学院数学与信息技术学院，运城 044000）

摘要：如何为用户提供感兴趣的个性化推荐图书商品向来都是智慧图书馆最核心的难题之一。因此，利用优质的推荐算法构建推荐系统就变得尤为重要。基于多维关系和用户聚类两个方面构建推荐算法，在充分考虑用户之间相关关系和图书商品之间相关关系的基础上，不断更新“用户-图书商品”二维矩阵，使其数值更加合理和真实。随后使用*k*均值聚类方法聚拢高相关性用户。最后在类内选取目标用户实现个性化图书商品推荐。实验结果表明，该推荐算法能取得较高的评价指标*F*1值，即算法更加优质和有效。

关键词：智慧图书馆；多维关系；用户聚类；个性化；推荐系统

0 引言

“智慧图书馆”将人与人、物与物、人与物通过信息技术紧密结合在一起[1]，在智慧图书馆中，读者数据获取更便利、更全面、更广泛，可以更加准确地刻画读者的信息需求，并据此进行精准推荐。因此，怎样准确把握读者的阅读及信息需求并向其进行智慧型图书推荐成为需要考虑的关键问题。

推荐系统被证明是一种解决信息过载和长尾物品问题的有效工具，在日常阅读新闻资讯、网上购物时，都能看到各种各样的推荐。图书个性化推荐最早是亚马逊公司为了提升长尾图书的用户抵达率而提出的，据 VentureBeat 统计，图书个性化推荐为亚马逊贡献了 35% 的销售额[2]。目前的推荐算法主要包括协同过滤推荐算法、基于内容的推荐算法、基于用户-产品二部图关系的推荐算法及混合推荐算法。Mooney 等[3]开发了一种图书推荐系统，它利用简单的信息提取技术从 Web 收集到物品的半结构化信息，基于给定项目的信息和用户的配置文件来向用户推荐文档、项目和服务。多维度研究是从不同的角度以及不同的层次研究问题，不再是单向和单一维度思考问题，然而在个性化推荐系统领域中的多维度分析很少。其中，Ma 等[4]在考虑新浪微博的个性化推荐问题时，提出从用户维度和关键词语维度来研究微博推荐系统个性化以及准确性的问题。其他较少的学术论文只是提出多维研究的概念，尚没有明确的研究方法。

因此，本文提出了一种新型的基于多维分析和用户聚类研究的个性化图书推荐算法。一方面，通过用户在智慧图书馆借阅图书的数据构建起“用户-图书商品”二维矩阵，从用户维度考虑，以每本图书作为用户的属性，计算用户相似性更新“用户-图书商品”矩阵。同样，从商品的维度考虑，分析每个商品的介绍等内容，计算商品之间的相似度，将其作为权重再次更新“用户-图书商品”矩阵。另一方面，利用第一部分得到矩阵计算用户之间相似性，实现用户聚类。然后在目标用户类别内，研究目标用户与其他用户之间关系和选购商品上的差

收稿日期：2023-03-28 **修稿日期：**2023-04-23

作者简介：*通信作者：闫俊辉（1972—），男，山西运城人，本科，实验师，主要研究方向为计算机应用，E-mail：yjh9163@163.com

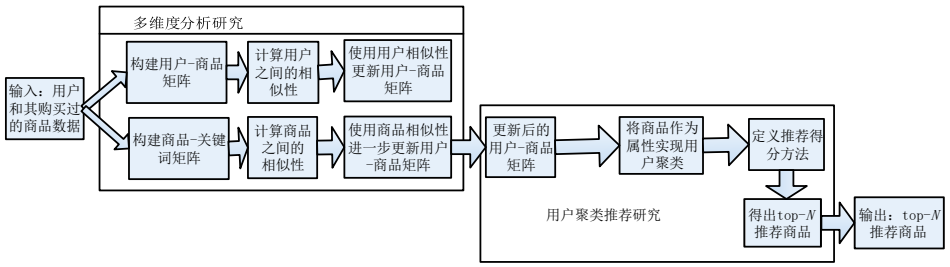


图1　个性化图书推荐系统框架

异，实现top-N商品推荐。图1概述了本文所提出的图书商品推荐系统的结构和内容。

1　理论基础与模型

1.1　相似度计算方法

本文研究的内容都是文本词语空间向量，由于余弦相似性对于词语空间向量以及稀疏向量的研究有较好的适用性，因此本文使用余弦相似性方法计算对象之间的相关系数。

$$\cos\theta = \frac{x_1x_2 + y_1y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}} \xrightarrow{\text{扩展到多维}} \cos\theta = \frac{\sum_1^n (A_i \times B_i)}{\sqrt{\sum_1^n A_i^2} \times \sqrt{\sum_1^n B_i^2}} \tag{1}$$

1.2　用户聚类算法

本文使用K均值作为用户聚类算法。K均值聚类算法是基于距离的典型聚类方法，其主要的思想是利用距离的概念判定节点之间的相似度，距离越大，相似度越小[5]。

$$Len = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2}$$

在n维欧氏空间中，每个点都是高维度数据，其中两个点$A = \{A_1, A_2, A_3, \cdots, A_n\}$，$B = \{B_1, B_2, B_3, \cdots, B_n\}$之间的距离$\rho(A, B)$可以由下式计算得出：

$$\rho(A,\ B) = \sqrt{\sum_1^n (A_i - B_i)^2} \tag{3}$$

再根据欧几里得距离计算K均值算法中用户之间相似度：

$$\mathrm{sim}(m,\ n) = \frac{1}{1 + \rho(A,\ B)} \tag{4}$$

1.3　图书个性化推荐模型

针对智慧图书馆的图书商品，本文提出了基于多维度关系和用户聚类的个性化推荐算法。爬取用户借阅过的图书商品的数据、智慧图书馆馆藏中图书商品的信息以及外部百科网站关于图书商品的介绍。引入用户相关与商品相关的因素到个性化推荐中，随后通过K均值聚类实现用户聚类，研究类内目标用户的图书个性化推荐。以下展示了本文的图书个性化推荐模型。

输入：用户和借阅过的图书商品数据以及图书商品的介绍信息

输出：针对目标用户的个性化推荐列表(top-N)

Step 1：数据预处理后构建用户-图书商品矩阵$\boldsymbol{M}_{cb}$

Step 2：依据原始用户-商品矩阵$\boldsymbol{M}_{cb}$计算用户之间的相关系数，获得用户相似方阵$\boldsymbol{M}_{c\times c}$，first $\boldsymbol{M}_{cb} = \boldsymbol{M}_{c\times c} \times \boldsymbol{M}_{cb}$，即考虑用户维度信息，第一次更新用户-图书商品矩阵，将first $\boldsymbol{M}_{cb}$矩阵标准化，得到$\boldsymbol{M}'_{cb}$

Step 3：根据百度百科和维基百科等资料网站得出图书商品的介绍信息，分词后构建图书-关键词矩阵，在此基础上计算图书商品之间的相似性，得到图书相关性方阵$\boldsymbol{M}_{b\times b}$，second $\boldsymbol{M}_{cb} = \boldsymbol{M}'_{c\times b} \times \boldsymbol{M}_{b\times b}$，即考虑图书维度信息，再次更新用户-商品矩阵，将second $\boldsymbol{M}_{cb}$矩阵标准化，得到$\boldsymbol{M}''_{cb}$

Step 4：根据更新后的用户-商品矩阵$\boldsymbol{M}''_{cb}$，使用K均值聚类方法，实现用户聚类

Step 5：每个类别内的用户联系紧密，亲密度更高。选取目标研究对象，研究类中其他用户购买过，而该目标用户没有买过的图书，图书集合为B

Step 6：根据其他用户借阅的信息，将图书集合B中的所有图书进行打分，依照评分前N位生成top-N推荐列表，将该列表作为个性化推荐结果推送给目标用户

该个性化推荐算法的创新点在于，将用户相关性和图书商品相关性考虑进“用户-图书商品”矩阵，以用户借阅图书商品的信息计算用

户相关性，以图书的外部资料测算图书商品相似性，实现两步更新“用户–图书商品”矩阵。在此矩阵基础上，完成用户K均值聚类，在每个类别内，用户之间相关性较高，根据图书商品的推荐得分，将目标用户没有借阅过，而类别内其他用户借阅过的商品形成top–N商品推荐，这种方式既考虑了推荐的准确性，也考虑了推荐的新颖性。

2 实证研究

本文选取了某图书馆的165条用户数据、借阅的35条图书商品的数据以及外部百科网站的35条图书商品介绍数据，从多维度和用户聚类来改善个性化推荐的研究。

2.1 用户相似性和图书商品相似性的计算

根据获取到的用户数据和其借阅过商品的数据，构建165×35的初始“用户–图书商品”二维矩阵 $\boldsymbol{M}_{165\times35}$，其中 $C_i(i=1,2,\cdots,165)$ 代表用户编号，$B_j(j=1,2,\cdots,35)$ 代表图书商品编号，矩阵里面的数值0代表没有买过，1代表买过。

根据图2所示的数据，以图书商品作为用户的属性，构建165条用户向量，使用余弦相似计算用户之间的相似度，构建165×165的用户相似矩阵 $\boldsymbol{M}_{165\times165}$。利用矩阵相乘 $\boldsymbol{M}_{165\times165}\times\boldsymbol{M}_{165\times35}=\text{first}\,\boldsymbol{M}_{165\times35}$，将新的矩阵标准化，得到基于用户关系更新后的“用户–图书商品”矩阵 $\boldsymbol{M}'_{165\times35}$，如图3所示。

2.2 基于用户聚类的个性化推荐

更新后的矩阵考虑了用户相关关系和图书商品相关关系，更加全面地体现了用户与图书商品之间的联系。根据更新后的矩阵的数据，将图书商品作为用户的属性特征，实现K均值用户聚类。聚类结果如图4所示。

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
C1	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
C2	1.00	1.00	1.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00
C3	1.00	1.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
C4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00
C5	1.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00
C6	0.00	1.00	0.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00
C7	1.00	0.00	1.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
C8	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
C9	0.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	0.00
C10	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00
C11	0.00	0.00	0.00	1.00	1.00	1.00	0.00	1.00	0.00	1.00
C12	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00
C13	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00
C14	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00
C15	0.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00
C16	0.00	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00

图2 初始“用户–图书商品”二维矩阵

	B1	B2	B3	B4	B5	B6	B7	B8
C1	0.3066	0.48098	0.51246	0.46089	0.45718	0.44407	0.42923	0.40142
C2	0.51321	0.46537	0.69276	0.49366	0.47407	0.63087	0.3107	0.40614
C3	0.39623	0.26829	0.39976	0.3351	0.36309	0.39262	0.19102	0.21015
C4	0.58679	0.49171	0.45907	0.44503	0.65621	0.62864	0.38205	0.50059
C5	0.73585	0.60683	0.83393	0.85201	0.55127	0.62304	0.60414	0.68831
C6	0.65	0.57756	0.45789	0.61839	0.34982	0.64989	0.65132	0.38961
C7	0.64434	0.3922	0.51127	0.40381	0.47768	0.42394	0.34522	0.43566
C8	0.53113	0.39415	0.51547	0.63636	0.60314	0.50112	0.71461	0.47107
C9	0.51226	0.43122	0.43535	0.28858	0.44632	0.29866	0.65017	0.23377
C10	0.44528	0.22732	0.37129	0.37421	0.25935	0.42841	0.42578	0.36128
C11	0.35755	0.35317	0.09371	0.14482	0.37153	0.28859	0.37975	0.28808
C12	0.63019	0.58829	0.55516	0.60148	0.48854	0.36801	0.46835	0.60449
C13	0.41604	0.52	0.36773	0.37526	0.38842	0.39485	0.37054	0.37662
C14	0.39587	0.45123	0.67452	0.54217	0.45231	0.42189	0.45282	0.60852
C15	0.41253	0.56374	0.38567	0.42583	0.56198	0.38756	0.64125	0.42798
C16	0.52315	0.38541	0.42158	0.53214	0.62184	0.52107	0.30297	0.36845

图3 更新后的“用户–图书商品”矩阵

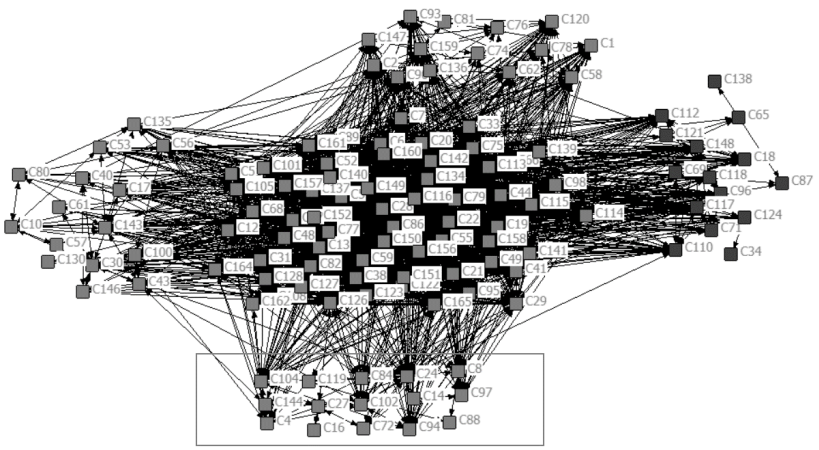


图4 K均值聚类结果示意图

同一类别中用户关联性较强，不同类别内用户关联性较弱。本文取图4中方框内的节点为代表类别进行研究，以C16为目标研究对象。由聚类结果可知，与C16在同一类别中的用户还有C104、C144、C4、C119、C27、C84、C102、C72、C24、C14、C94、C8、C97以及C88。在C16所在的类别中，除去C16已经购买过的图书商品，考虑类内其他用户借阅的商品推荐得分。以更新后的“用户-图书商品”矩阵为基础，将矩阵中的系数作为权重加到原始“用户-图书商品”矩阵中，通过对应相乘和商品列数值相加的方式计算出商品的总得分。从前一步全部图书商品数据中提取出C16所在类的图书商品数据，得出类内其他用户借阅（C16没有购买）的图书商品的得分。前5个得分最高的图书商品作为top-5图书推荐给用户C16，即针对目标用户完成图书商品的个性化推荐。top-5的推荐图书商品列表及其得分见表1。

表1　推荐商品得分

图书商品编号	商品加权得分
B6	5.307892594
B35	4.823407558
B15	4.423991637
B29	4.356848107
B8	4.284568748

2.3　实验结果与评估

准确度经常用来评价推荐系统中推荐算法的效果，本文使用准确度指标中准确率和召回率的加权平均*F*1指标来评估本文的推荐算法。

$$F1 = \frac{2*precision*recall}{precision + recall} \quad (5)$$

通过设置推荐商品数*N*的不同，得到推荐系统在一系列推荐实践下的评价指标准确率、召回率以及*F*1的值。以此来验证推荐系统的稳定性与持久性，排除偶然的因素。图5显示的是不同推荐算法下*F*1值的变化。

由图5可以看出，随着推荐商品个数的增加，本文的推荐算法评价指标*F*1值在0.4～0.7之间波动，呈上升趋势。同时，与基于原始矩阵推荐商品的方法比起来，本文的推荐算法*F*1值较大，说明本文算法更优，推荐效果更好。

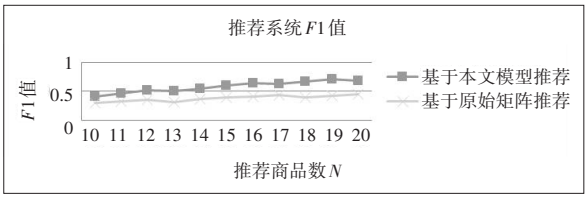


图5　不同推荐系统的评价指标*F*1值比较

3　结语

为提高图书推荐算法的准确性，解决推荐系统中冷启动的问题，本文从多维度加权和用户聚类两种方法实现图书商品个性化推荐。充分考虑用户之间的相关关系和图书商品之间的相关关系，将两者以数据加权的形式填补到“用户-图书商品”矩阵中，不断更新原始二维矩阵，使之更加合理准确。通过K均值聚类的方式给类别内目标用户推荐得分最高的top-5图书商品，本文的推荐算法获得较好的评价指标*F*1值。

优质的个性化推荐，能够精准地为用户提供其感兴趣的图书商品，不但提高用户的阅读体验，而且增强阅读者对于智慧图书馆的粘性。未来的研究可以使用外部信息对用户维度的数据再次进行优化，丰富“用户-图书商品”二维矩阵数据，进一步提高推荐系统的准确性。

参考文献：

[1] 严栋. 基于物联网的智慧图书馆[J]. 图书馆学刊，2010,7:8-10.

[2] 刘建国,周涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展,2009,19(1):1-15.

[3] MOONEY R J, BENNETT P N, ROY L. Book recommending using text categorization with extracted information[C]//Proceedings of the Recommender Systems Papers from 1998 Workshop, Technical Report WS-98-08,1998,1188.

[4] MA H,JIA M,ZHANG D,et al. Combining tag correlation and user social relation for microblog recommendation[J]. Information Sciences an International Journal,2017(385):325-337.

[5] 金建国. 聚类方法综述[J]. 计算机科学,2014,41(S2):288-293.

（下转第73页）

文章编号: 1007-1423(2023)14-0066-04 DOI: 10.3969/j.issn.1007-1423.2023.14.013

基于改进随机森林算法的防火墙日志异常检测并行化方法

刘 成，王佳斌*，洪继炜

（华侨大学工学院，泉州 362021）

摘要： 随机森林分类算法在产生决策树以及投票流程中各个决策树的分类准确度各不相同，由此带来的问题是少部分决策树会影响随机森林算法的整体分类性能。除此以外，数据集中的不平衡数据也能影响到决策树的分类精度。针对以上缺点，对Bootstrap抽样方法添加约束条件，以降低非平衡数据对生成决策树的影响；以及利用袋外数据(Out-of-Bagging)和非平衡系数对生成的决策树进行评估加权。试验结果表明，所提算法改善了随机森林对不平衡数据的分类精度。

关键词： Spark；随机森林算法；入侵检测；日志异常检测

0 引言

在现今的大数据时代，网络在人们生活中不可或缺。随着互联网的用户量和规模不断增长，网络流量也呈现出井喷式的增长，网络安全问题也变得越值得重视，如何保障网民的合法利益也变得越来越重要。因此，如何对大规模的网络流量进行异常检测并分类，是非常值得研究的课题。

近年来，许多国内外学者开始使用机器学习的方法来解决网络入侵流量异常检测和分类情景中所面临的问题[1]。Erman等[2]提出使用K-means算法对单向流信息进行分类。Moore等[3]改进了用于网络流分类的传统朴素贝叶斯算法，但该算法需要稳定的数据集，不适用于高速和不稳定的网络。雍凯[4]提出了决策树属性的权重评估，生成决策树时，通过优先选取权重较高的属性来提升单个决策树的分类性能。徐鹏等[5]提出利用训练数据的信息熵构建决策树分类方法，但该算法难以应用于高维样本。目前，国内外对网络流量异常检测与分类大多是在单机环境下进行研究，有限的资源难以胜任大数据时代下的大规模流量异常检测任务。

针对以上不足，本文提出了一种基于Spark平台的改进随机森林算法，对生成决策树的Bootstrap抽样方法添加约束条件，降低非平衡数据对生成决策树的影响；同时，利用不平衡度和袋外数据对决策树进行加权，提高随机森林算法整体的分类准确率。

1 相关技术及方法

1.1 随机森林及Bootstrap抽样介绍

随机森林是一种由Breiman[6]提出的集成学习分类算法，该算法利用若干个决策树来对样本进行学习和预测，随机森林算法的分类原理为：

步骤1：在样本集中用Bootstrap采样并有放回地抽取m个样本，产生一个新生成的子集，并在新生成的子集中选取所有特征中的s个特征，作为决策树的分离节点；

步骤2： 重复步骤1，直到得到n个决策树组成随机森林；

步骤3：将预测数据交给上一步骤中产生的

收稿日期： 2022-12-29 **修稿日期：** 2023-06-26

作者简介： 刘成(1999—)，男，安徽马鞍山人，硕士研究生，研究方向为分布式、异常检测；*通信作者：王佳斌(1974—)，男，福建泉州人，副教授，主要从事物联网、云计算、大数据、智能仪器的研究，E-mail:fatwang@hqu.edu.cn；洪继炜(1996—)，男，福建泉州人，硕士研究生，研究方向为分布式、推荐系统及其应用

随机森林中的每棵决策树进行预测，统计各个决策树的预测结果，最多决策树预测出的类别就是随机森林的分类结果。

1.2　抽样子集的不平衡度划分

随机森林算法使用的Bootstrap重抽样方法每次抽取总体的三分之二作为一个训练样本，不断地重复这一个抽取动作，以期望用一系列大小为原训练样本三分之二的训练样本搭建出一个空间，通过这个空间来无限接近总体。

对于抽样子集的不平衡度定义如下：①假设数据集中的样本数为$|D| = M$。$S = \{(x_i, y_i)\}$，$i = 1, 2, \cdots, m$，公式中x_i满足$x_i \in X$，X是维度为n的空间，$X = \{f_1, f_2, \cdots, f_n\}$，且$y_i \in Y$，$Y$是样本的特征值，$Y = \{1, \cdots, C\}$。②定义数据集的不平衡系数为$B$：

$$B = \left| \frac{S_{\max}}{S_{\min}} \right| \tag{1}$$

其中：$S_{\max}$和$S_{\min}$分别为数据集的多数类样本和少数类样本，满足$S_{\max} \cup S_{\min} = \{S\}$且$S_{\max} \cap S_{\min} = \{\varnothing\}$。

从这个角度可以把抽样得到的子集分为以下三种：①子集不平衡系数B'小于原数据集不平衡系数B。②子集不平衡系数B'大于原数据集不平衡系数B。③抽样子集中无少数类样本，即$S_{\min}$不存在，子集不平衡系数B'无法计算。

以上三种情况在随机森林随机抽样中都会出现，其中②和③情况所得的抽样子集只会加重样本的不平衡性，通过这些抽样子集训练得到的决策树会干扰最终的投票效果。

2　基于Spark的改进随机森林算法及其并行化

2.1　基于约束条件的Bootstrap重抽样

针对以上的决策树子集不平衡问题，设计了一种基于约束条件的重抽样Bootstrap算法。改进后的Bootstrap抽样会过滤掉不平衡系数B'较大的子集，进而使不平衡数据集对生成的决策树产生的影响降低。基于约束条件改进后的Bootstrap重抽样流程如下：

步骤1：利用Bootstrap抽样从数据集D中抽取三分之二的数据样本；

步骤2：计算所抽取的数据子集的数据非平衡度B，并添加约束条件为数据子集的非平衡度小于或等于原数据集的非平衡度：

$$B'(i) = \left| \frac{D'_{\max}}{D'_{\min}} \right| \leqslant B \tag{2}$$

其中：$D'_{\max}$为抽样得到的数据子集中的多数类数据，$D'_{\min}$为抽样得到的数据子集中的少数类数据；

步骤3：若Bootstrap抽样得出的数据子集满足约束条件，则可利用该数据子集来构造决策树。

2.2　改进随机森林算法的并行化建模

但在训练不平衡数据时，随机森林的精度和性能下降一直是该算法的应用局限性，本文提出了一种基于袋外数据和非平衡系数的加权随机森林算法。文献[7]研究了贝叶斯公式，得出了评估各个分类器性能的公式，将其中的$con(i)$用袋外数据的$F1$值代替，得出的加权公式为

$$W_{OOB}(i) = \frac{2}{N} - \frac{\dfrac{1}{F_1(i)}}{\displaystyle\sum_{j=1}^{t} \frac{1}{F_1(j)}} \tag{3}$$

其中：N为决策树分类器的数；$F_1(i)$为第i个决策树分类器的袋外数据的$F1$值；$W_{OOB}(i)$表示根据袋外数据所求得的第i个决策树分类器的权值。

同时，将得出的各个数据子集的非平衡度$B(i)$作为权值的另一个要素：

$$Weight(i) = W_{OOB}(i) \times B(i) = \left(\frac{2}{N} - \frac{\dfrac{1}{F_1(i)}}{\displaystyle\sum_{j=1}^{t} \frac{1}{F_1(j)}} \right) \times \left| \frac{D'_{\max}}{D'_{\min}} \right| \tag{4}$$

综上所述，改进随机森林算法主要分为以下步骤：

(1)随机Bootstrap抽样获取样本特征，计算不平衡度是否满足构建决策树的条件；

(2)利用Bagging抽样构建决策树；

(3)通过每棵决策树的袋外数据预测该决策树的$F1$值；

(4)利用分类器性能评价公式对各个决策树进行加权，并耦合成完整的并行化加权随机森林模型。

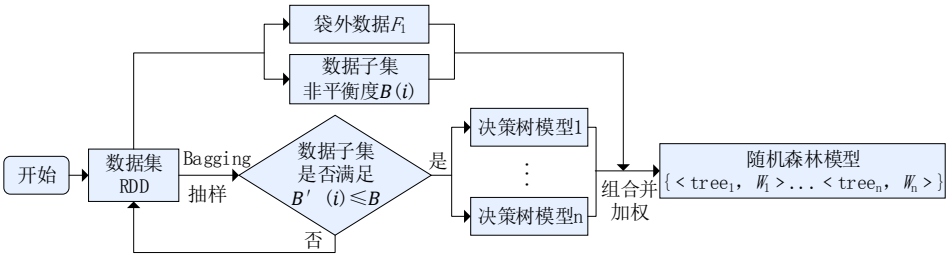


图 1 改进的随机森林算法建模

3 实验

3.1 实验环境及数据集

本文的实验环境由Windows11平台上Vmware Workstation安装的三台虚拟机组成，其中一台作为Master，两台作为Worker，系统为CentOS，Spark版本为2.4.4，Hadoop版本为2.7.1，使用的开发语言为Scala 2.13。

本文采用的数据集是加拿大通信安全机构和加拿大网络安全研究所发布的CIC-IDS-2018网络入侵检测数据集，数据集提供的流量模拟真实网络流量。CIC-IDS-2018数据集中包括多种不同的攻击场景，攻击包含Brute Force FTP，Brute Force SSH，DoS，Heartbleed(OpenSSL缺陷)，Web Attack，Infiltration(渗透)，Botnet(僵尸网络)和DDos。

3.2 评价标准

在分类任务中，一般使用精确率、召回率以及*F*1作为评价指标。为了便于介绍，用混淆矩阵来表示：*TP*表示实际与判定都为正类的样本；*FP*表示实际为负类，但被错误判定为正类的样本；*FN*表示实际为正类，但被预测为负类的样本；*TN*表示实际与判定结果都为负类的样本。

表 1 混淆矩阵

	真实属于正类的样本	真实不属于正类的样本
被判定属于正类的样本	*TP*	*FP*
被判定不属于正类的样本	*FN*	*TN*

精确率的数学公式为

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

召回率的数学公式为

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

*F*1综合考虑了召回率和精确率，公式为

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

3.3 实验结果

由表2可以看出，文中改进的随机森林算法在对CIC-IDS-2018的分类上要优于传统RF算法。综合来看，本文提出的改进随机森林算法利用了决策树的袋外数据和子数据集的不平衡系数，并通过加权来代替随机森林中相同权重的决策树投票，有效减少了随机森林中劣质树的干扰。

表2 文中算法与传统随机森林算法以及AUC值直接作为权重加权的随机森林算法比较

方法	准确率/%	召回率/%	*F*1/%
改进随机森林	97.28	97.13	97.21
传统随机森林	95.53	95.37	95.45

此外，得益于Spark分布式平台的特性，算法的运行时间大幅度缩短，这是因为Spark将文件读取进了内存，减少了对硬盘的频繁I/O操作。以下为算法在Spark分布式平台上与单机平台上的运行时间对比。

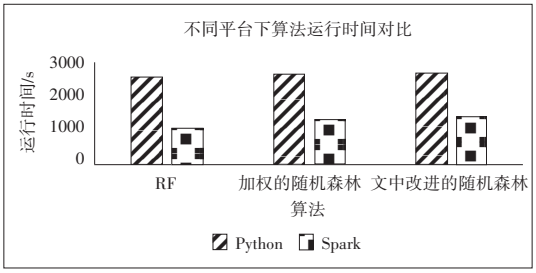


图 2 不同平台下各算法运行时间对比

4　结语

本文提出一种基于Spark的改进随机森林算法，先根据数据子集的不平衡度对决策树的生成过程施加约束，来改善随机森林分类算法对不平衡数据集条件下的适用性。接着采用加权投票的方式减少了随机森林中劣质树的干扰，提高随机森林算法的分类精度。实验结果表明，文中的改进随机森林算法在CIC-IDS-2018数据集上的分类精度比传统随机森林算法更胜一筹。在以后的研究中，考虑将文中算法与分层抽样相结合，来进一步改善数据子集与原数据集样本类别的一致性。

参考文献：

[1] NGUYEN T T T, ARMITAGE G. A survey of techniques for internet traffic classification using machine learning[J]. IEEE Communications Surveys & Tutorials, 2009, 10(4): 56-76.

[2] ERMAN J, MAHANTI A, ARLITT M, et al. Identifying and discriminating between web and peer-to-peer traffic in the network core[C]//Proceedings of the International Conference on World Wide Web, New York, NY, USA, 2007: 883-892.

[3] MOORE A W, ZUEV D. Internet traffic classification using Bayesian analysis techniques[J]. ACM Sigmetrics Performance Evaluation Review, 2005, 33(1): 50-60.

[4] 雍凯. 随机森林的特征选择和模型优化算法研究[D]. 哈尔滨：哈尔滨工业大学, 2008.

[5] 徐鹏, 林森. 基于C4.5决策树的流量分类方法[J]. 软件学报, 2009, 20(10): 2692-2704.

[6] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123-140.

[7] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

Parallel implementation method of firewall log anomaly detection based on improved random forest

Liu Cheng, Wang Jiabin*, Hong Jiwei

(College of Engineering, Huaqiao university, Quanzhou 362021, China)

Abstract: The classification accuracy of the random forest classification algorithm is different in the decision tree generation and voting process. The problem is that a small number of decision trees will affect the overall classification performance of the random forest algorithm. In addition, the unbalanced data in the dataset can also affect the classification accuracy of the decision tree. In view of the above shortcomings, add constraints to the Bootstrap sampling method to reduce the impact of unbalanced data on the generation of decision trees; And use out of bag data (Out of Bagging) and unbalanced coefficients to evaluate and weight the generated decision tree. The experimental results show that the proposed algorithm improves the classification accuracy of random forests for unbalanced data.

Keywords: spark; random forest; intrusion detection; log anomaly detection

文章编号: 1007-1423（2023）14-0070-04 DOI: 10.3969/j.issn.1007-1423.2023.14.014

基于Zeppelin+Hive的数据分析与可视化

张玉叶[1*]，孙延坤[2]

（1. 济南职业学院计算机学院，济南 250014; 2. 济南沃尔菲斯科技有限公司，济南 250014）

摘要：随着大数据时代的到来及大数据产业的迅速发展，快速有效地对海量数据进行分析处理及可视化，成为大数据产业从业人员的必备技能。通过对一组房屋销售数据的分析处理，介绍了如何利用Zeppelin和Hive来快速对海量数据进行分析及可视化，并给出了具体实现方法和代码。

关键词：Zeppelin；Hive；数据分析；数据可视化

0 引言

随着移动互联网和智能手机的迅速普及，电子商务、云计算、虚拟现实、人工智能等技术的不断发展，传统产业不断被重塑，大数据产业成为了新的产业革命核心[1]。各行各业正在以前所未有的速度每天产生海量数据，数据正在成为一种新的资源和新的生产要素[2]。大数据产业当前仍处于技术高速发展时期，用到的框架和工具有很多，其中用于做数据分析处理及可视化的方案也有很多，可根据实际使用场景来选择。本文选用的是Zeppelin和Hive，Zeppelin是一个基于Web的notebook，提供交互式数据分析和可视化功能。后台支持接入多种数据处理引擎，如Spark、Hive等。同时支持多种编程语言，如Scala、Python、Spark、Markdown、Shell等；Hive是一个基于Hadoop的数据仓库工具，可以将结构化的数据文件映射为一张数据库表，然后通过相应的SQL语句来对海量数据进行联机分析处理[3]。Zeppelin本身内置了柱状图、折线图、饼图、散点图、面积图等最常用的几种基本图形，只要将相应的数据查询出来，不需要写任何代码就可直接实现数据可视化，使用起来相对比较简单，因此利用Zeppelin和Hive做数据分析与可视化是目前业界应用较广也是比较容易掌握的一种方案。

本文利用Zeppelin和Hive来对一组房屋销售数据进行分析处理及可视化，通过此例来展示使用方法，从而快速上手数据分析与可视化。

1 数据集

数据来源于某地区的房屋销售记录，选取了某一年的销售记录作为原始数据集，数据集中的部分数据如图1所示。

sale_date	selling_price	bedrooms_num	housing_area	floor_num	housing_rating	year_built	year_repair
20140302	545000	3	1670	1	8	1974	0
20140211	785000	4	3300	2	10	1984	0
20140107	765000	3	3190	2	9	2007	0
20141103	720000	5	2900	2	9	1989	0
20140603	449500	5	2040	1	7	1969	0
20140506	248500	2	780	1	7	1958	0
20140305	675000	4	1770	1	8	1971	0
20140701	730000	2	2130	1	7	1941	0
20140807	311000	2	860	1	6	1903	0
20141204	660000	2	960	1	6	1942	0
20140227	435000	2	990	1	7	1947	0
20140904	350000	3	1240	1	7	1959	0
20140902	385000	3	1630	3	8	2008	0
20140413	235000	2	930	1	6	1930	0
20140930	350000	3	1300	1	6	1971	0
20140507	1350000	4	2000	1	9	1926	0

图1 数据集中部分数据

数据集中各字段含义依次为销售日期、销售价格、卧室个数、房屋面积、楼层数、房屋评分、房屋建造年代和房屋修缮年代。

2 环境准备

要使用Zeppelin和Hive，首先需要先搭建好相应的Hadoop，可根据实际情况选择搭建伪分

收稿日期：2023-03-22 **修稿日期：**2023-04-06

作者简介：*通信作者:张玉叶(1973—),女,山东青岛人,副教授,硕士,研究方向为大数据分析与应用,E-mail:yuyejn@163.com; 孙延坤(1969—),男,山东济南人,本科,研究方向工业控制与数据分析

布式或完全分布式，然后安装好相应的Hive[4]。确保Hadoop和Hive能够正常使用，然后安装部署Zeppelin。本文示例所使用的是Hadoop伪分布式，各组件相应的版本为Hadoop2.7.7、Hive2.3.4和Zeppelin0.10.0。

3　数据分析及可视化

数据分析是指用适当的分析方法对收集来的大量数据进行分析，提取有用信息和形成结论，对数据加以详细研究和概括总结的过程。数据分析包括狭义的数据分析和广义的数据分析。狭义的数据分析是指根据分析目的，采用对比分析、分组分析、交叉分析等分析方法，对收集的数据进行处理与分析，提取有价值的信息的过程；广义的数据分析是指依据一定的目标，通过统计分析、聚类、分类、回归等方法发现数据中隐含的信息的过程。即狭义的数据分析主要是从数据中提取显性的信息，而广义的数据分析则主要是从数据中挖掘隐含的信息，因此广义的数据分析实际上是狭义的数据分析加上数据挖掘[5]。

数据可视化就是以图形化方式表示数据。要想把数据可视化，必须知道它表达的是什么。数据是对现实世界的简化和抽象表达，当可视化数据时，其实是在将对现实世界的抽象表达可视化。数据可视化可帮助人们从一个个独立的数据点中解脱出来，换一个不同的角度去探索和理解它们。

通常，使用自然语言、数字等形式表达的概念是枯燥的、不易懂的，而可视化的技术可增加数据的生动性，图表的应用可使数据更具有可读性。“一图抵千言”，数据可视化能让决策者通过图形快速直观地看到数据分析结果，从而更容易理解业务变化趋势或发现新的业务模式[6]。

在人工智能和大数据时代，数据分析与可视化是人类理解和处理海量数据的关键技术，可以帮助人们快速从海量数据中发现和获取相应的信息或者帮助人们在错综复杂的数据中发现和验证不同维度和指标之间的关联。

不同的数据集有不同的分析目标和数据指标。通常情况下可将数据分为三类：用户数据、行为数据和商品数据。用户数据通常是记录用户基本信息的；行为数据则是记录用户做过什么，如对于电商数据则可能是用户的购买行为、浏览行为或是收藏行为等；而商品数据则是记录商品的基本信息。对于不同的数据有不同的数据分析指标，如对于用户数据比较关注的是新增用户数、用户活跃率和留存率等，对于行为数据则比较关注访问次数、访问人数、转换率等指标，而对于商品数据比较关注的则是成交量、成交额等。

上面列举的只是一些通用数据指标，对于具体的数据集除了一些通用数据指标外还可以根据具体情况来分析其它的一些相关数据指标。本数据集属于商品数据，因此对于此数据集，其分析目标主要是房屋的成交量、成交额，除此之外还可以分析房屋面积与价格之间的关系、不同评分或不同类型的房屋数量等。

3.1　统计每月的房屋销量

数据集中存放了2014年的房屋销售数据，通过统计每月的房屋销量，可以得出每个月的销量情况。相应的代码及查询结果如图2所示。

```
%hive
--统计每月的房屋销量
select date_format(from_unixtime(unix_timestamp(sale_date,'yyyyMMdd')),'yyyyMM') as month
    ,count(1) as total
from tb_house_2014
group by date_format(from_unixtime(unix_timestamp(sale_date,'yyyyMMdd')),'yyyyMM');
```

FINISHED

settings

month	total
201401	450
201402	580
201403	876
201404	1057
201405	1108
201406	974

图2　每月的房屋销量查询结果

Zeppelin中，默认情况下数据查询结果是以二维表格的形式显示的。可通过点击数据上方的相应按钮来快速实现数据的可视化。如单击折线图按钮，则数据以折线图显示。结果如图3所示。也可以直接通过点击柱状图按钮，将其以柱状图显示，结果如图4所示。柱状图适宜展示各分组数据之间的数量比较。

如果是想查看各月的销量占比情况，则可直接点击饼图按钮，将其以饼图显示，结果如图5所示。饼图适宜描述各分组数据在总数据中的占比情况。

图 3　每月房屋销量折线图

图 4　每月房屋销量柱状图

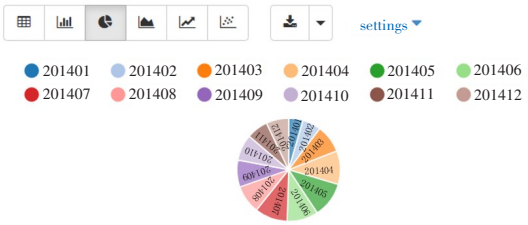


图 5　每月房屋销量占比饼图

同样的方法可统计分析每月的房屋销售额，还可以按季度来统计房屋的销量和销售额等。对于统计得到的数据可根据需要以相应的图形进行展示。

3.2　房屋面积与销售价格之间的关系

如果想要了解房屋面积与房价之间的关系，可通过绘制相应的散点图来查看。散点图适宜展示两个变量之间的相关性。代码及结果如图6所示。

图6　房屋面积与房价散点图

通过上述结果可以看出，房屋面积与房价之间具有一定的正相关性，即房屋面积越大，其房价越高。

3.3　不同评分的房屋数量

如果想要了解用户对房屋的评分情况，可通过分析不同评分的房屋数量来获取。代码及结果如图7所示。同样，如果是想查看数量，可用柱状图展示，如果想查看不同评分的房屋占比，则可改用饼图展示。类似的方法还可统计不同楼层数或不同卧室数的房屋数量等。

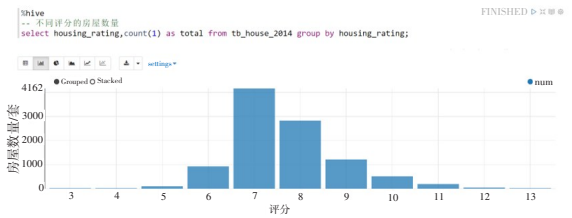


图 7　不同评分的房屋数量柱状图

4　结语

大数据时代如何快速有效处理海量数据，从中发现和提取人们所需要的有用信息已经成了各行各业的急需。而基于Zeppelin和Hive的数据分析及可视化可用相对简短的代码帮助人们快速有效地处理数据并获取到相关信息，掌握Zeppelin和Hive的使用已经成了各行业数据分析师所必备的基本技能。本文通过一个实际数据集展示了如何利用Zeppelin和Hive来进行数据分析及可视化，其实现简单方便，在目前大数据时代背景下有着广泛的应用领域和良好的使用前景。读者可参考其实现思路稍微改动一下即可实现更多类似的数据分析及可视化效果，从而快速提高数据分析及可视化能力。

参考文献：

[1] 肖芳，张良均．Spark大数据技术与应用[M].2版．北京：人民邮电出版社，2022.
[2] 谭旭，李程文．大数据存储[M]．北京：人民邮电出版社，2022.
[3] 肖睿．兰伟．廖春琼．Hadoop数据仓库实战[M]．北京：人民邮电出版社，2020.
[4] 张军，张良均．Hadoop大数据开发基础[M]．北京：人民邮电出版社，2021.
[5] 魏伟一，李晓红，高志玲．Python数据分析与可视化[M]．北京：清华大学出版社，2020.
[6] 周苏，王文．大数据及其可视化[M]．北京：中国铁道出版社，2016.

Data analysis and visualization based on Zeppelin+Hive

Zhang Yuye[1*], Sun Yankun[2]

（1.Computer Department of Jinan Vocational College, Jinan 250014, China;
2. Jinan Wolves Technology Limited Company, Jinan 250014, China）

Abstract: With the arrival of the Big data era and the rapid development of the Big data industry, data analysis and visualization based on Zeppelin+Hive has become a necessary skill for practitioners of the Big data industry to quickly and effectively analyze, process and visualize massive data. By analyzing and processing a set of housing sales data, this article introduces how to use zeppelin and hive to quickly analyze and visualize massive data, and provides specific implementation methods and codes.

Keywords: Zeppelin；Hive；data analysis; data visualization

（上接第65页）

Research on personalized book recommendation in intelligent library based on multidimensional relationship and user clustering

Yan Junhui[*]

（School of Mathematics and Information Technology, Yuncheng University, Yuncheng 044000, China）

Abstract: How to provide users with personalized recommended books has always been one of the most concerned problems of intelligent library. Therefore, it is very important to use high quality recommendation algorithm to build recommendation system. The recommendation algorithm is constructed from the two aspects of multidimensional relationship and user clustering. On the basis of fully considering the correlation between users and the relationship between books and commodities, the two-dimensional matrix of “user-book commodities” is constantly updated to make its value more reasonable and real. Then, the K-means clustering method was used to gather high correlation users. Finally, select target users within the class to realize personalized book product recommendation. The experimental results show that the recommendation algorithm can obtain a higher value of the evaluation index $F1$, and the algorithm is more high-quality and effective.

Keywords: intelligent library; multidimensional relationship; user clustering; personalization; recommendation system

开发案例

文章编号: 1007-1423(2023)14-0074-07 DOI: 10.3969/j.issn.1007-1423.2023.14.015

基于农业物联网的特早熟黄花菜智慧管理系统

盛晓雨[1,2]，蔡 鹏[1,2]，郑世玲[1,2]，张 霞[1,2*]

（1. 聊城大学物理科学与信息工程学院，聊城 252000；2. 山东省光通信科学与技术重点实验室，聊城 252000）

摘要：针对“特早熟”黄花菜种植管理过程中效率低、劳动成本高、管理不精准等问题，设计了一种基于农业物联网的“特早熟”黄花菜智慧管理系统。该系统包含集空气温湿度、土壤温湿度、光照强度等传感器模块为一体的环境监测系统以及由卷帘、照明、灌溉等设备组成的远程控制系统，为农业大棚提供了更智能的管理模式。在聊河农业黄花菜种植基地的应用结果表明，该系统能够显著减少人工投入、降低成本，推动黄花菜园管理升级，从而实现黄花菜产量和质量的提升，可以在农业生产管理领域普遍应用和推广。

关键词：农业物联网；环境监测；智慧管理系统

0 引言

近年来，随着物联网、5G通信等技术的飞速发展以及专家学者的深入研究，国内外涌现出许多与农业物联网相关的新技术，拓宽了物联网在现代农业中的应用[1]。比如美国应用“5S技术”、智能化农机技术等形成了农业精细化、规模化发展的智慧农业生产线系统[2]；我国应用传感器技术以及ZigBee、Wi-Fi、NB-IoT等网络通信技术，将环境监测设备从互联网单点设备逐步转变成多元化设备平台[3]。但当前国内农业智慧化覆盖程度不够广泛且相关硬件设备成本较高，导致在农业生产采集过程中，一些农作物种植管理还是人工操作。一方面，生产过程对温湿度等数据的把控要靠农民本身的经验或对相关数据进行逐一采集来判断作物所处环境的信息，缺乏准确性与权威性；另一方面，通风、灌溉等设施需要人力启动，耗费大量的人力物力，工作量大[4]。急需构建智慧农业管理系统，借助大数据平台进行农业数据的收集、分析以及环境的改善，以实现现代农业的提质增效增产[5]。其中，陈玉[6]搭建的智慧茶园云管理平台采用Spring Cloud架构完成，可实现对茶叶大棚的环境监测以及精准控制管理；郑春雨等[7]基于安卓平台和GIS技术，设计了一套智慧农业管理系统并在河湖长制遥感动态监测等工作中取得有效推广；刘飞飞等[8]设计的基于ZigBee的分布式农业环境监测系统，采用了ZigBee通信传输技术，进行数据采集处理与精准传输。

从上述研究背景和现状可知，农业智慧化管理从点的突破逐步转变成系统能力的提升，覆盖了农业生产中的多个环节。但是，智慧管理系统多用于种植水果蔬菜、茶叶等常见作物，应用类型不够广泛且系统的设计方案和核心技术还有待优化提高。针对上述问题，通过对聊河黄花菜市场需求的调研，设计了一种基于农

收稿日期：2023-03-28 **修稿日期：**2023-05-04

基金项目：山东省自然科学基金(ZR2018MA044，ZR2022MF284)；东昌府区人民政府与聊城大学深化校地合作项目：打造智慧农业应用基地——基于农业物联网的特早熟黄花菜智慧管理系统

作者简介：盛晓雨(1998—)，女，山东济宁人，硕士研究生，研究方向为农业物联网；蔡鹏(1998—)，男，山东济南人，硕士研究生，研究方向为农业物联网；郑世玲(1983—)，女，山东聊城人，讲师，博士，研究方向为相位恢复、数字图像处理等；*通信作者：张霞(1975—)，女，山东聊城人，教授，博士，研究方向为光纤通信系统及复用技术、农业物联网及智能光网络，E-mail：zhangxia2@lcu.edu.cn

业物联网的“特早熟”黄花菜智慧管理系统，应用于聊城市聊河农业科技有限公司的黄花菜种植大棚。系统利用数据采集硬件模块对黄花菜大棚中的空气温湿度、土壤温湿度、光照强度、二氧化碳等环境信息进行数据采集，依托OneNET物联网云平台搭建接收数据的界面，工作人员可通过手机客户端和网页客户端对采集数据进行监控并远程控制棚内的卷帘、风机、照明、灌溉等设备，让黄花菜始终处在适宜的生长环境中，从而有效提高大棚的管理效率。该系统通过后期测试和现场布设后，聊河基地用户反馈可实现黄花菜园区标准化生产的集中管控，提高劳动生产率和农作物采摘质量。

1　智慧管理系统总体架构设计

整个黄花菜智慧管理系统主要应用于现代化“特早熟”黄花菜园的综合管理。根据农业生产环境的实际需求进行分析，参考物联网体系结构模型，采用分层思想[9]对黄花菜智慧管理系统的总体架构进行设计。

智慧管理系统总体架构主要由三层结构组成，分别为感知执行层、传输层和应用层，如图1所示。

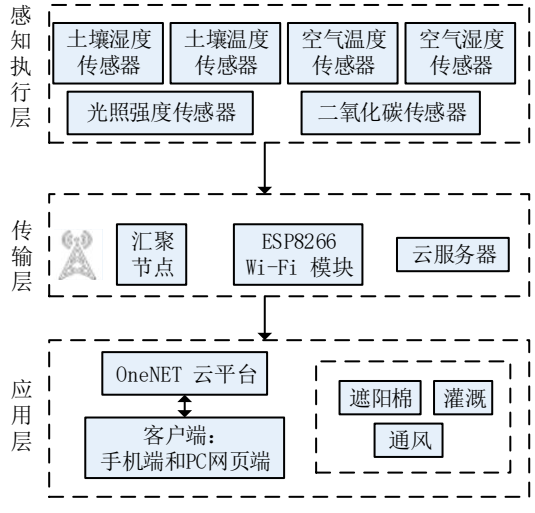


图1　智慧管理系统总体架构

感知执行层主要包括土壤温湿度、空气温湿度、光照强度等多种传感器，对黄花菜大棚中的温湿度、光照强度、二氧化碳等环境信息进行数据采集；传输层由汇聚节点、Wi-Fi模块等部分组成，主要负责数据传输与预处理；应用层是依托OneNET物联网云平台搭建而成的智慧黄花菜园管理平台，用户可以通过网页客户端或手机微信小程序实时查看农业大棚环境信息，以及对温室大棚内的遮阳、灌溉、通风等设备进行远程控制。

2　智慧管理系统硬件设计

该智慧管理系统从功能实现上又可以分为环境监测和远程控制。环境监测系统硬件结构由Arduino UNO PLUS主控模块、多种传感器数据采集模块、无线通信模块等构成。基于环境监测系统提供的数据，推进农业管理智能化，设计远程智能控制系统，该控制系统硬件结构主要由STM32F103C8T6主控芯片、无线通信模块、控制设备和相应的继电器、接触器控制电路组成，采用自动调节控制和远程控制相结合的设计，实现棚内卷帘、风机、照明、灌溉等设备的即时控制。系统硬件结构如图2所示。

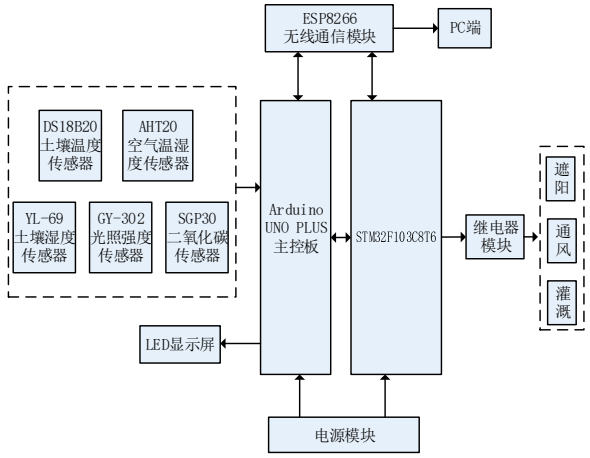


图2　系统硬件结构

2.1　数据采集系统硬件设计

在黄花菜园区，可以通过采集环境监测数据为大棚的智慧化管理提供数据指导，数据采集系统的硬件电路图如图3所示。选用Arduino UNO PLUS作为主控模块，配置了十四个数字I/O端口、六个模拟端口、I2C插头等引脚，支持UART、I2C、GPIO等，适用于多种物联网应用场合[10]。主控电路与DS18B20土壤温度检测模块、YL-69土壤湿度检测模块、AHT20空气温

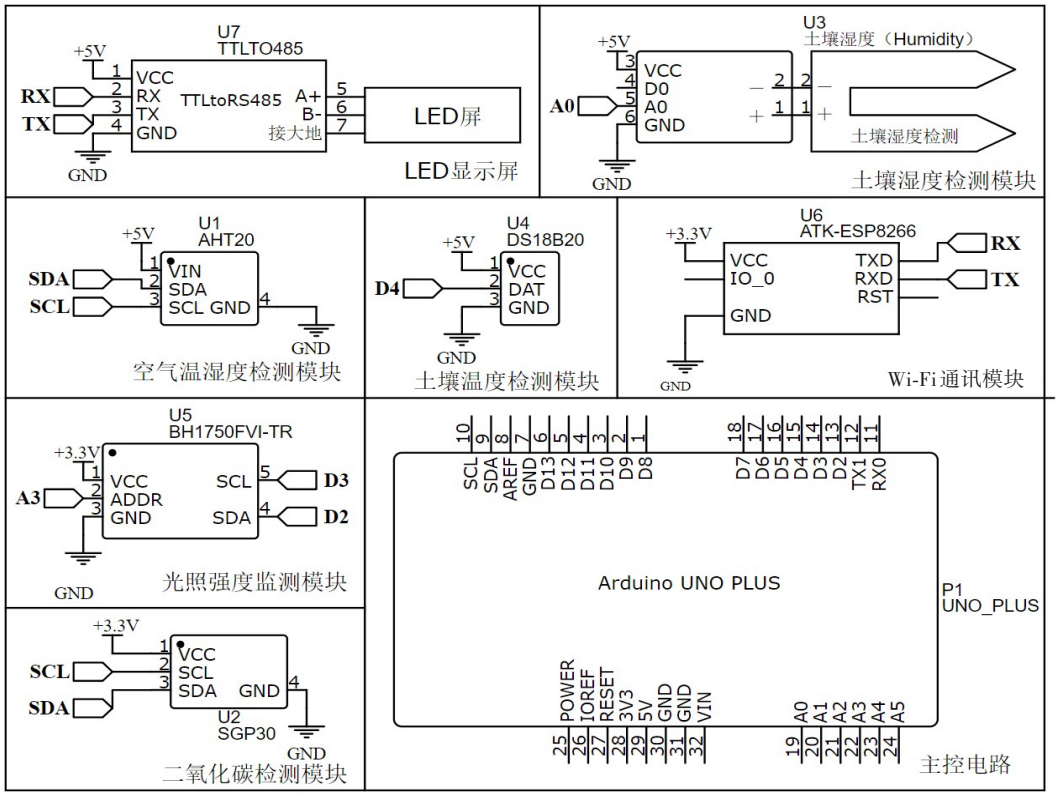


图 3　数据采集系统硬件电路图

湿度检测模块、GY-302光照强度监测模块和SGP30二氧化碳检测模块连接，进行多种环境数据的采集。采集到的传感器数据通过Wi-Fi模块上传到云平台，Wi-Fi模块采用串口与MCU通信，云平台存储和展示所收集到的环境信息。

除了PC端云平台展示，还安装了LED显示屏。主控模块本身不含RS485通讯接口，所以外接了TTL转RS485模块，通过RS485协议与LED显示屏进行通讯，通过LED显示屏可以更直观地观察到温室大棚内环境数据变化。

2.2　远程控制系统硬件设计

为了最大化地开发数据采集系统的黄花菜信息传输处理功能，发挥数据采集的作用，实现基于数据指导的远程控制照明、灌溉等设备的功能，选用性能更加稳定、操作更加便捷的远程控制系统，远程控制系统的硬件电路如图4所示。该系统选用STM32F103C8T6单片机作为主控模块，端口输出信号经过光耦隔离电路，使被隔离的低压主控电路和外部高压继电器控制电路之间没有电的直接连接，主要是防止因有电的连接而引起的干扰。

在继电器控制电路中，单片机通过发送高低电平信号来控制继电器线圈吸合，不同的继电器分别实现控制总供电和控制接触器切换功能。当控制总供电的继电器吸合时，负责切换的继电器吸合或断开会引起两个接触器交替吸合，以此来控制接触器控制电路中电机的正转反转；当控制总供电的继电器断开时，接触器都断开，控制的设备停止运作。此控制系统还加入了Wi-Fi通讯模块，将设备状态上传至云平台，平台可下发命令控制相应设备，适用于遮阳棉、喷水器、通风机等设备的远程控制。

3　系统软件设计

3.1　系统软件功能描述

基于Arduino、STM32嵌入式软件编程，数据采集部分软件实现主控板在主循环中完成对土壤温湿度、光照强度、二氧化碳等传感器数

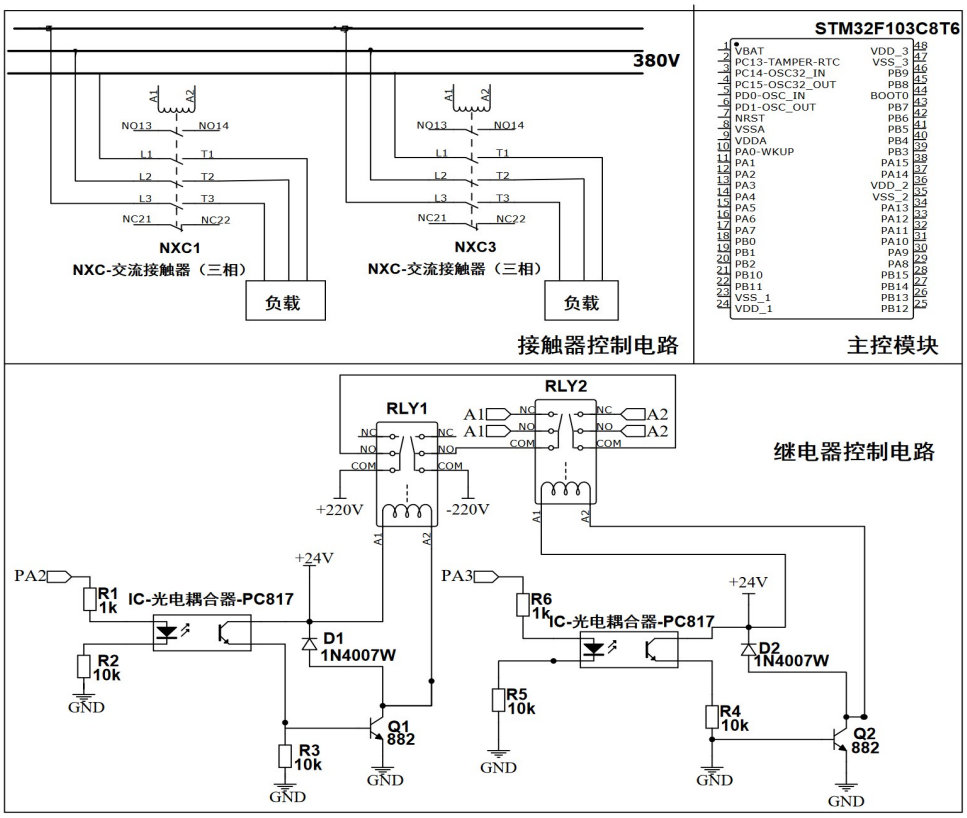


图 4　远程控制系统硬件电路图

据的采集，将数据汇集到终端节点，通过 Wi-Fi 模块将数据定时发送到 OneNET 云平台。还可以按照控制器规定的通讯协议与 LED 屏通讯，将传感器采集到的数据展示在大屏幕上。远程控制部分软件实现基于 STM32 控制设备与状态上传云平台，实时监听接收网关下达的控制命令。

3.2　数据采集部分软件设计

打开 Arduino 编程软件，写入硬件采样程序，系统对硬件、I2C 等通讯协议进行初始化设定（包括中断、延时、串口）。MCU 控制中心通过 Wi-Fi 模块发送 AT 指令连接云平台，平台得到响应显示设备在线。各种传感器通过 USART、单总线、I2C 通讯协议与 Arduino UNO PLUS 主控板进行通讯，单片机将采集到的环境监测数据上传到 OneNET 云平台，通过可视化界面及 LED 屏实时显示。软件流程如图 5 所示。

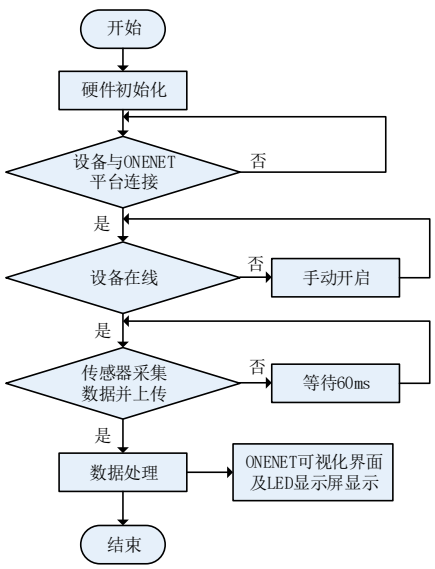


图 5　数据采集软件流程

3.3　LED 电子显示屏软件设计

该系统为了更直观地显示数据，在数据采集设备上外接了液晶拼接 LED 大屏，该 LED 屏

内置的控制卡支持二次开发，通过RS485通讯方式与主控模块进行通信。软件实现需要按照字库控制器的标准协议格式(帧头+包头数据+数据域+包校验+帧尾)生成数据帧，并将数据帧发送给内置的控制卡，控制卡会按照协议有所回复，同时会根据命令数据的不同进行相应的处理，包括将采集到的环境信息显示到LED屏上，通讯就可以正常进行了。

3.4 远程控制部分软件设计

首先，进行硬件初始化，配置所需的AT指令，通过串口把AT指令发送给Wi-Fi模块，即循环发送一个指针里面的所有数据。其次，通过EDP协议建立与平台的连接，调用平台封装好的接口函数，以EDP协议的Type3格式上传数据。最后，通过平台下发命令，提取到的命令会存储在数组中用字符匹配函数判断该数组是否有想要的命令，即可进行相应的设备控制操作。同时，为了防止设备掉线，每隔30 s发一次心跳包，以长时间保持跟平台的连接[11]。软件流程如图6所示。

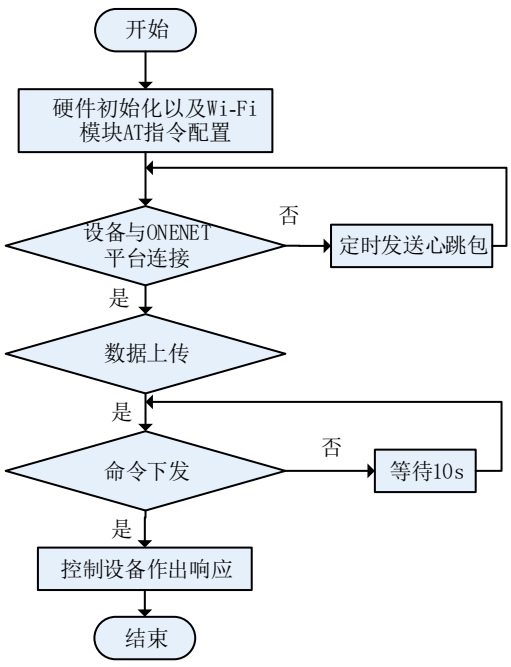


图6 远程控制软件流程

4 系统测试

为了检验该系统的性能，通过采集环境数据用于黄花菜大棚的控制。该管理系统在聊城市聊河农业科技有限公司农业大棚开启综合测试，现场部分设备布设展示如图7和图8所示。

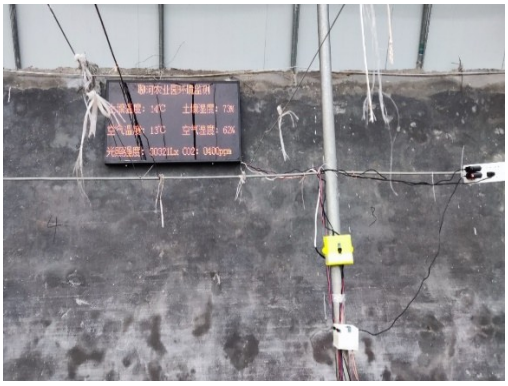


图7 传感设备分布情况

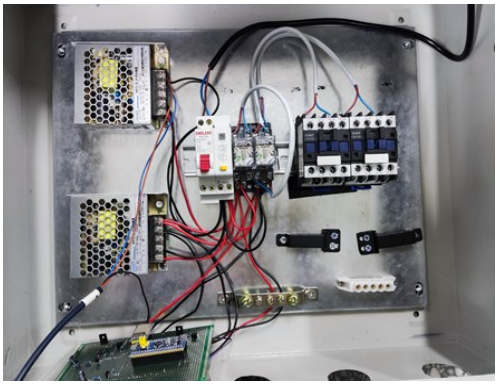


图8 电机控制实物图

其中，数据采集系统每隔4 s采集一次数据，技术测试人员可在编程系统Arduino IDE的串口监视器中查看到传感器采集的环境信息，与此同时各种数据会发送至云平台，产生发送数据成功响应，检查串口监视器与云平台数据流展示是否一致，如图9所示，显示一致后说明成功完成环境监测数据的上传。

环境监测数据成功上传后，如果土壤湿度低于预设阈值，工作人员可通过云平台远程控制系统打开喷淋、滴灌等设备，自动完成农田灌溉，补足作物所需水分。经过一系列测试，聊河基地用户反馈该系统操作简洁高效，极大

提升了数据采集的准确性，达到了远程控制的可靠性和及时性效果，用户能随时随地观察设备的运行状态，及时进行预警，明显减少了人工投入，对实现黄花菜的提质增效增产有显著帮助。

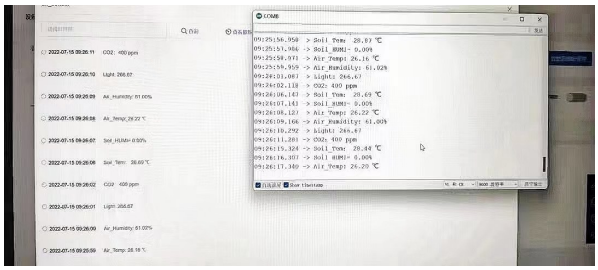


图 9　具体数据流和串口监视器显示

除了云平台显示和控制，用户还可以通过如图10所示的微信小程序界面，登录查看黄花菜生产过程中的环境监测数据，在系统中查看每一个采集点的环境实时数据和历史数据，实现遮阳棉、洒水器、风机等设备的远程控制。

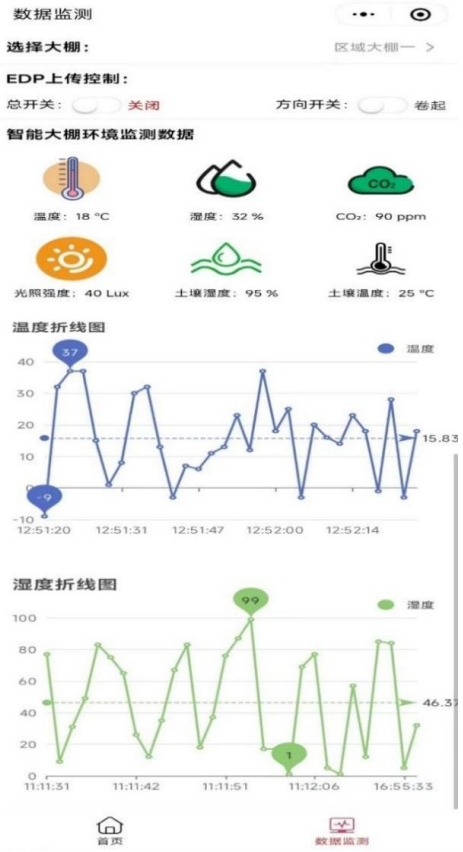


图 10　基于特早熟黄花菜智慧管理系统的微信小程序

聊河基地用户反映使用微信小程序可以提升日常农事作业记录的便利程度并有效反馈黄花菜生长过程中遇到的问题，及时采取相应措施保障黄花菜的健康生长状况。

5　结语

本文设计了一个基于农业物联网的特早熟黄花菜智慧管理系统，该系统采用传感器技术和Wi-Fi网络通信等技术，将采集数据传输至云平台进行处理和分析，实现环境信息的自动监测和实时报警，用来判断黄花菜的环境信息是否正常，如果不正常还可以及时采取措施修复环境，以尽量避免恶劣环境对黄花菜造成的减产甚至死亡问题。然后，将灌溉、遮阳等半自动化设施接入平台以根据环境信息来实现智能控制，大大降低人力成本。同时，用户还可以通过微信小程序进行数据提取及操控。另外，考虑到农业大棚网络覆盖范围不广和Wi-Fi模块传输距离较短等环境因素，可能会造成采集过程中数据的丢失，后期将数据处理与分析等方面作为研究重点，全面优化系统的数据传输性能，最大化地开发数据的功能，为现代农业的智慧化管理提供数据指导。

参考文献：

[1] 陶佳雯,徐楠,刘松霖,等.物联网技术助力农业产业化发展的简析与实施[J].物联网技术,2022,12(9):105-107.

[2] 赵春江,李瑾,冯献,等.“互联网+”现代农业国内外应用现状与发展趋势[J].中国工程科学,2018,20(2):50-56.

[3] 罗锡文,廖娟,臧英,等.我国农业生产的发展方向:从机械化到智慧化[J].中国工程科学,2022,24(1):46-54.

[4] 孙树莉.农业物联网构建与发展面临的主要挑战[J].现代农业科技,2022(15):142-151.

[5] 高嘉伟,夏磊.大数据在农业物联网中的应用研究[J].电子元器件与信息技术,2022,6(2):65-66,230.

[6] 陈玉.基于物联网技术的智慧茶园管理系统设计[D].济宁:曲阜师范大学,2020.

[7] 郑春雨,刘文静,孙阳.基于安卓平台的智慧农业管

理系统设计与实现[J]. 测绘与空间地理信息，2022，45(S1)：73-74，77.
[8] 刘飞飞，徐隆姬，马礼然. 基于ZigBee的分布式农业环境监测系统设计[J]. 传感器与微系统，2021，40(3)：90-92.
[9] 陈兰. 基于物联网的智慧茶园数据采集管理系统的设计与实现[D]. 杭州：浙江理工大学，2019.
[10] 张红艳，王萍，张振亚. 基于Arduino的温湿度数据采集系统实验报告[J]. 科学技术创新，2022(20)：62-65.
[11] SEKARAN K，MEQDAD M M，KUMAR P，et al. Smart agriculture management system using internet of things [J]. TELKOMNIKA (Telecommunication Computing Electronics and Control)，2020，18(3)：1276-1285.

Intelligent management system of super early-maturing day Lily based on agricultural Internet of Things

Sheng Xiaoyu[1,2], Cai Peng[1,2], Zheng Shiling[1,2], Zhang Xia[1,2*]

(1. School of Physics Science and Information Engineering, Liaocheng University, Liaocheng 252000, China;
2. Shandong Provincial key Laboratory of Optical Communications Science and Technology, Liaocheng 252000, China)

Abstract: Aiming at the problems of low efficiency, high labor cost and inaccurate management in the process of planting and managing super-early-maturing day Lily, an intelligent management system of "super-early-maturing" day Lily based on agricultural Internet of Things was designed in this work.The system includes an environmental monitoring system that integrates sensor modules such as air temperature and humidity, soil temperature and humidity, light intensity, and a remote control system for shutter, lighting, irrigation and other equipment, which provides a more friendly management mode for agricultural greenhouses. The application results in Liaohe Agricultural Cauliflower Planting Base showed that the system can significantly reduce the labor input and the cost, promote the management upgrade of day Lily garden, so as to improve the yield and quality of day Lily, and facilitate the universal application of agricultural production management and the popularization of controllable cost.

Keywords: agricultural Internet of Things; environmental monitoring; intelligent management system

文章编号: 1007-1423（2023）14-0081-05 DOI: 10.3969/j.issn.1007-1423.2023.14.016

基于深度学习的聊天机器人自动化平台设计

肖俊辉*，孙　丽，冉国翔，朱荣清，陈镜宇，张天乙

（东南大学成贤学院电子与计算机工程学院，南京 210000）

摘要：随着2023年的到来，新时代新科技飞速发展，同时人民日益增长的美好生活需要对科技提出的要求也愈加多元化、复杂化，单纯的程序执行已经不足以满足，搭载了人工智能（Artificial Intelligence）的机器人的出现使人们在柳暗花明中寻得一村，当今火热的ChatGPT标志着AI技术的又一里程碑式的进步，但是由于其造价高昂且获取、操作难度较大，不易为所有人所使用。因此本项目旨在创造一个亲民简单的个性化机器人搭建平台，即使是没有专业计算机知识技术的普通人都可以用其于生活中的各行各业。

关键词：AI；机器人；深度学习；自动化

0　引言

本项目的灵感来源于QQ群中的群聊机器人小冰，其功能包含入群欢迎、提醒打卡、简单的游戏交互等，那么我们能否创造一个功能更加完善、更加类似人类、更加贴合特定群聊特色的（个性化的）机器人加入群聊之中呢？一个对计算机编程了解甚少的人又如何能够在群聊中拥有自己所需要的机器人呢？于是我们便计划开发一个平台，使得更多的人能够通过低代码甚至无代码的简单方式获取和培养个性化的聊天机器人，旨在将本技术简单化、日常化，带入人们的日常生活。

经过多方渠道考察，从1950年开始，随着聊天机器人相关研究的不断发展，已有众多聊天机器人产品相继面世，目前的热点便是2022年11月30日由美国OpenAI公司发布的聊天机器人程序ChatGPT，其为人工智能技术驱动的自然语言处理工具，能够通过学习和理解人类的语言来进行对话，还能根据聊天的上下文进行互动，真正像人类一样来聊天交流。这一里程碑式的技术革命便是深度学习应用越来越广泛，技术越来越成熟的体现，其核心通过机器来模仿人类的对话内容和习惯，对聊天输入的内容做出决策和判断，给予相应的回应。

现如今国内外虽已出现众多的聊天机器人产品，但都在个性化和简便性方面有所不足，导致目前大部分聊天机器人还是需投入到客服环境中使用，因此如果实现了聊天机器人个性化和简便性的突破，便可走进普通人的日常生活，小到供人消遣娱乐、排忧解难，大到协助公司部门进行人事管理、甚至能够做到24小时不间断提供高质量人性化服务等，以其个性化程度进军各行各业，将拥有巨大的市场潜力。

本项目计划开发一个基于深度学习的聊天机器人自动化搭建平台的软件产品，试图在上述技术方面有所突破，弥补现阶段QQ小冰在群聊趣味性、个性化上的不足，以低代码甚至无

收稿日期：2023-03-30　**修稿日期：**2023-06-29

基金项目：东南大学成贤学院2022年大学生实践创新训练省级校企合作项目（SCX22043）

作者简介：*通信作者：肖俊辉（2002—），男，广东梅州人，本科，主要研究方向为深度学习，E-mail：threeax@foxmail.com；孙丽（1970—），女，江苏南京人，教师，主要研究方向为计算机科学与技术；冉国翔（2001—），男，江苏南京人，本科，主要研究方向为人工智能；朱荣清（2000—），男，江苏连云港人，本科，主要研究方向为软件工程；陈镜宇（2001—），男，江苏连云港人，本科，主要研究方向为网络安全；张天乙（2001—），男，江苏盐城人，本科，主要研究方向为自然语言处理

代码形式对机器人进行操作。本项目产品与QQ小冰的功能对比见表1。

表1 本项目与QQ小冰功能对比

	QQ小冰	项目产品
入群通知	√	√
定时通知	√	√
群聊互动	×	√
“斗图”	×	√
智能答复	×	√
个性化	×	√

1 总体设计

软件总共分为三个系统，如图1所示，包括聊天数据收集系统、聊天回复系统和自动化训练系统。

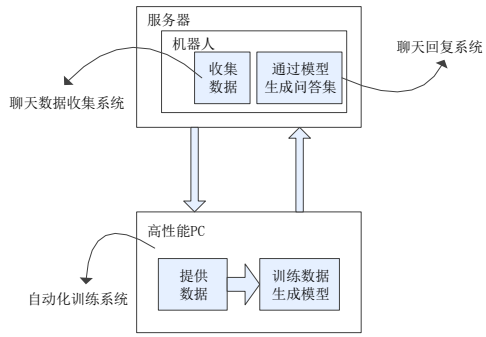


图1 总体架构

2 详细设计

2.1 软件设计

2.1.1 聊天数据收集系统

通过Mirai框架，运用Http技术[1]，使得QQ能与软件连接，可以自动将聊天记录生成一个“问—答”的词库，其中对数据集过滤方式包括：用户自定义策略、常用的无用语句、敏感隐晦字眼分析。最后根据词库链生成训练用语料集。

2.1.2 聊天回复系统

同样通过Mirai框架，让QQ群或者私聊作为一个聊天室的载体，可以收集数据的同时，使用训练的模型给出特定的回答，并回复在群聊中，系统功能架构如图2所示。

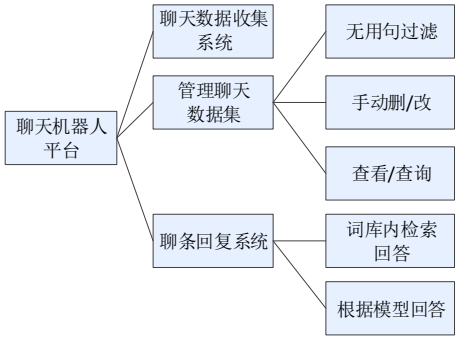


图2 收集/回复系统功能架构

2.2 模型训练设计

2.2.1 数据预处理

因为群聊内的聊天有时候会出现上文不接下文的情况，此时如果直接将聊天内容原封不动地生成对话场景模型，将会使得模型上下文逻辑混乱。我们设计词库链的初衷就是为了能更好地生成对话场景。

每个回答均是上一个“问题”的“答案”和下一个回答的“问题”，“问题”和“答案”均有一个“出现频率”的属性，据此可以较好地模拟聊天的对话场景，生成对话模型。

2.2.2 进行tokenize(标记化)

在文本分割的步骤上，我们从传统的词向量[2]转而使用了同为Transformer的BERT框架[3-4]的tokenize，能很好地应对一词多义的问题，从而提高模型对语言的理解能力，如图3所示。

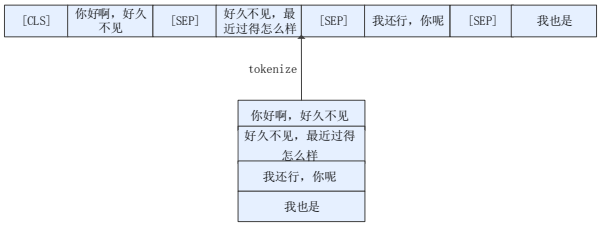


图3 tokenize示意图

2.3 模型训练

2.3.1 切分训练集和测试集

读取上一个步骤生成的预处理数据，将它

们按照一定比例划分为训练集和测试集。

2.3.2　读取预训练模型

群聊语料集对于训练一个模型来说还不够庞大，所以本文使用 GPT-2 预训练模型[5-6]来训练我们的模型。

2.3.3　自回归训练

在强大的 GPT-2 模型基础上，我们采用自回归训练方式，让模型输出能更加符合语料集的聊天场景，加强连续聊天能力，如图 4 所示。

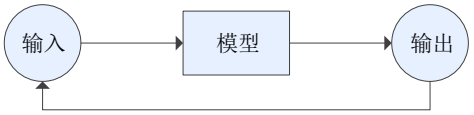


图 4　自回归概念图

2.4　训练结束指标

2.4.1　模型训练指标(loss)计算

在每一批次的训练中，通过前向传播计算出模型的预测输出和实际输出，使用反向传播算法计算出损失函数值(loss)以及对应训练模型参数的梯度，同时进行梯度裁剪[7]，防止发生梯度爆炸，进行一定次数的梯度积累后，根据梯度下降算法，更新模型的参数，完成一轮训练。

$$loss_{epoch} = \frac{\sum_{0}^{batch\_size} loss}{batch\_size} \tag{1}$$

2.4.2　生成困惑度最低模型(Perplexity)

困惑度可以被看作是一个语言模型中预测的不确定性大小的加权平均。在相同的测试数据集上，一般来说，困惑度越低，模型的性能就越好。

在一次训练中，通过对每个批次的 loss 值进行加权平均就可以得到一次训练的 loss 值，在测试集上使用同样的算法得出测试 loss 值后，与最佳测试 loss 值进行比较，低于最佳测试 loss 值的将保存，在每轮训练中不断更新与迭代困惑度最低模型，如图 5 所示。

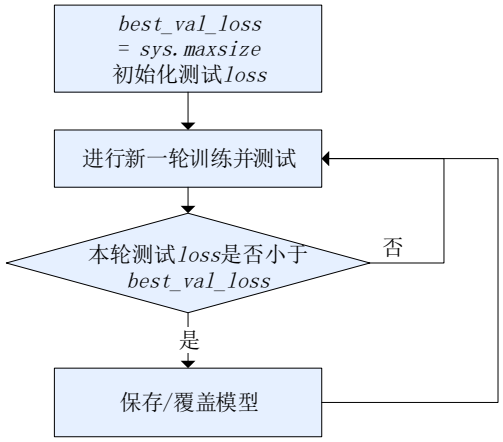


图 5　生成困惑度最低模型逻辑

在本文的多次测试中，有时候困惑度低，模型的生成效果不一定会越好，所以最后采用 loss 收敛来判断训练结束，loss 值稳定且不再下降则训练完成，如图 6 和图 7 所示。

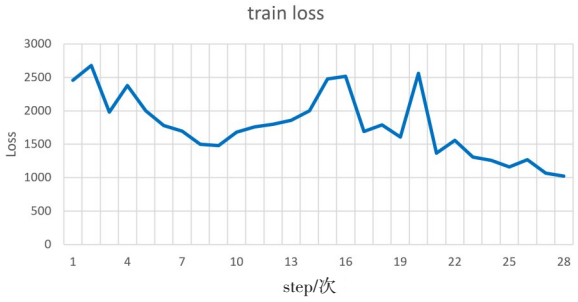


图 6　训练初期 loss 值变化

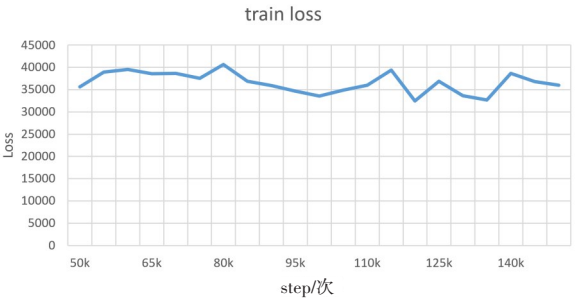


图 7　训练后期 loss 值变化

3　软件测试

3.1　软件操作说明

软件操作流程如图 8 所示。

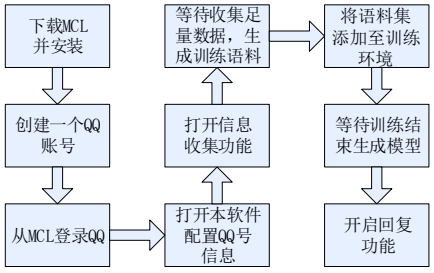


图 8　软件操作流程

3.2　软件流程测试

图 9　功能列表

首先进行收集系统测试，添加群聊“add learning”，开始记录“learning”，如图10所示，收集一段时间后得足量数据，进行数据预处理，如图11所示。

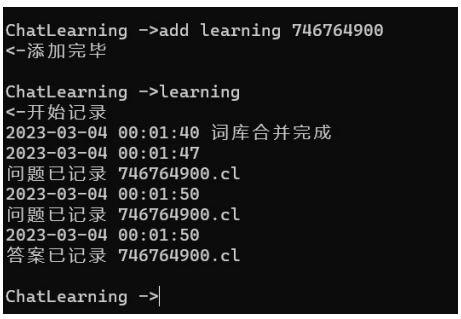


图 10　收集系统测试

图 11　数据预处理测试

将生成的语料集置于训练环境进行模型训练，如图12所示，注意关注loss值浮动幅度，等待训练结束，如图13所示。训练完成后将模型重新加载至机器人内部，如图14所示。

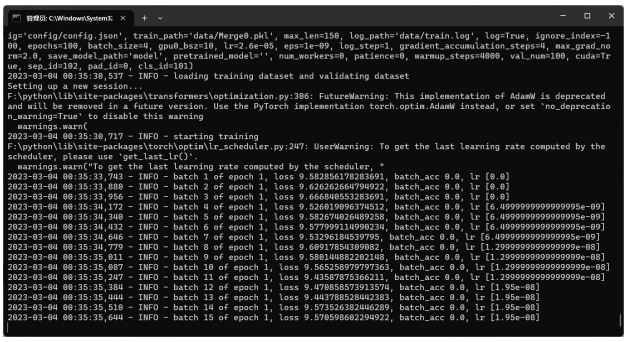


图 12　模型训练测试

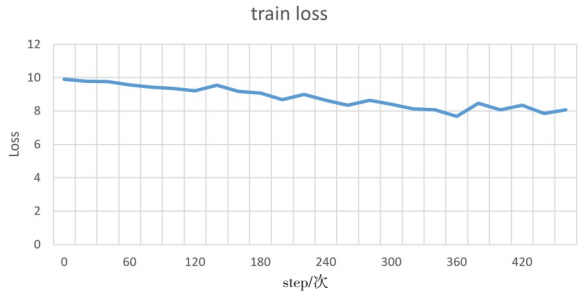


图 13　训练Loss可视化测试

图 14　训练模型加载测试

4　结语

本文详细介绍了一个基于深度学习的聊天机器人自动化平台，该平台旨在为普通用户提供一个亲民简单的方式来构建个性化的聊天机器人，从而将这项技术更广泛地应用于日常生活。所设计的平台分为“收集系统”“回复系统”“训练系统”三个模块，三个模块相互分离，方便用户根据需求灵活地选择和使用。

“收集系统”负责连接群聊，自动收集聊天记录并生成相应的问答词库。经过过滤无用信息和敏感内容后，将收集到的聊天数据用于训练语料集。而“回复系统”则负责在群聊中使

用训练好的模型或收集的词库进行智能回复。“训练系统”模块则包括数据预处理、模型训练和模型优化等环节。

通过这一设计，本文为普通用户提供了一个易于操作、个性化的聊天机器人搭建平台。这不仅有利于推动人工智能技术在日常生活中的应用，还能满足各种不同场景的需求。在未来的研究中，我们将不断改进和优化本平台的功能和性能，以满足用户不断增长的需求，推动聊天机器人领域的发展。

参考文献：

[1] 杨清玉,李金丽,陈吉兰,等. HTTP接口自动化测试方法研究[J]. 微型机与应用,2016,35(18):22-25.
[2] 席宁丽,朱丽佳,王录通,等. 一种Word2Vec构建词向量模型的实现方法[J]. 电脑与信息技术,2023,31(1):43-46.
[3] 陈旻杰,魏轩,郑堂,等. 基于BERT的智能问答系统[J]. 数字技术与应用,2022,40(1):161-163.
[4] 陈月月,李燕,帅亚琦,等. 基于BERT-CRF的中文分词模型设计[J]. 电脑知识与技术,2022,18(35):4-6.
[5] 张伟. 基于GPT-2模型的生成式对话系统应用研究[D]. 上海:华东师范大学,2022.
[6] 庄子源. 基于GPT-2的心理咨询聊天机器人[D]. 南京:南京邮电大学,2022.
[7] 吴德兵. 梯度裁剪下降算法的收敛性分析[D]. 杭州:浙江理工大学,2022.

Design of an automated platform for seep learning-based Chatbot

Xiao Junhui*, Sun Li, Ran Guoxiang, Zhu Rongqing, Chen Jingyu, Zhang Tianyi

(School of Electronics and Computer Engineering, Southeast University Chengxian College, Nanjing 210000, China)

Abstract: As 2023 arrives, new technologies are developing rapidly in this new era. At the same time, the increasingly diverse and complex needs of people's growing demand for a better life require more than just simple program execution. Therefore, the emergence of robots equipped with Artificial Intelligence (AI) has brought a glimmer of hope to people. The current popular ChatGPT represents another milestone in the advancement of AI technology. However, due to its high cost and difficulty in acquisition and operation, it is not easily accessible to everyone. Therefore, this project aims to create a user-friendly and simple personalized robot building platform that even ordinary people without professional computer knowledge and skills can use in various industries in their daily lives

Keywords: AI; Bot; deep learning; automated

文章编号: 1007-1423（2023）14-0086-05 DOI: 10.3969/j.issn.1007-1423.2023.14.017

智能巡更巡检管理系统开发与设计

王 东[1]，陈 阳[2*]

（1. 青海高等职业技术学院财经商贸系，海东 810006；2. 广州大学管理学院，广州 510006）

摘要：基于RFID近距离、高频的无线通信技术的优势，结合电子标签、自动控制技术、数据库、管理系统于一体设计一套智能巡更巡检管理系统。它由RFID电子标签、巡检器、PC端巡更巡检管理系统三部分组成。巡更巡检管理系统采用Visual Basic开发，数据库管理系统采用SQL Server，系统与数据库采用ODBC连接，以便有关部门提升和优化巡检管理工作能力和效率。

关键词：巡更巡检；管理系统；软件开发；工作效率

0 引言

目前许多行业仍然采用古老的人工巡检模式，浪费了人力物力，也使信息传递的及时性受到了阻碍[1]。如何采用先进的巡检系统来配合当今时代各行业的需求成为了一个备受人们关注的焦点。

基于RFID技术的智能巡更巡检管理系统，是一套融合无线电射频识别（radio frequency identification，RFID）、自动控制技术、数据库、管理系统于一体的智能化巡检系统。它由RFID电子标签、巡检器、PC端巡更巡检管理系统三部分组成。电子标签采用美国德州仪器半导体公司出品的RFID，是全球应用最广泛的电子标签产品。巡更巡检管理系统采用Visual Basic开发，数据库管理系统采用SQL Server，系统与数据库采用ODBC连接[2]。

1 需求分析

1.1 系统的设计思想

智能巡更巡检管理系统设计参照了相关的国家标准，其思路设计如下：

（1）先进性：采用行业领先和广泛使用的RFID技术和产品，这种近距离、高频的无线技术在相当长一段时间内，仍会是主流的行业解决方案。

（2）实用性：电子标签因近年来在电商领域应用极为广泛，反倒促进了其在各行各业的广泛普及，并且在使用的便利性、安全性方面得到了很大的发展，而且价格相当便宜，使得整体解决方案的总拥有成本（TCO）下降[3]。

（3）可靠性：采用成熟的开发工具和数据库系统，确保巡检管理系统的稳定性，并提供完善的数据库备份方案，保证数据的一致性和完整性。

1.2 系统的设计目标

针对系统的使用情况，其设计目标如下：

（1）操作界面简洁，容易上手，用户无需培训即可自行使用，体验良好。

（2）系统的响应速度要快。在低配硬件或旧系统上，仍能顺畅运行。

（3）要求原始数据只输入一次，通过设计的有效性检验机制，确保来自巡检机的数据的准确性和一致性。

（4）主界面所调用的子功能尽量都是独立的。

收稿日期：2023-06-12 **修稿日期：**2023-06-25

作者简介：王东（1984—），男，博士，副教授，研究方向为信息管理与信息系统；*通信作者：陈阳（1984—），男，硕士研究生，高级工程师，研究方向为计算机技术应用，E-mail：gdcc_chenyang@foxmail.com

1.3　技术原理

美国德州仪器TIRIS智能非接触式IC卡技术是RFID行业较为优质成熟的解决方案，本文在TIRIS的基础上，利用电子标签和巡检器实现自动化巡检[4]。在每个需要巡检的设备或位置安置一个电子标签，在执行巡检作业时，工作人员将巡检器在电子标签周围轻轻摇一摇，巡检器线圈发出的电磁波被电子标签的应答器线圈接收，在短时间内便聚集能量，能量足够时电子标签内部芯片执行指令，将该电子标签上全球唯一的序列号通过发射器发出，接着巡检器接收到电磁波后将其转译为电子标签的序列号，在巡检器的液晶屏幕就可以查询到该时刻扫描到的电子标签，巡检器还可以通过数据传输线将自身记录数据传到电脑[5]。如图1所示。

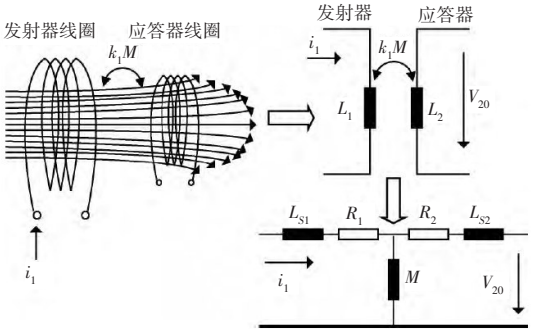


图1　电子标签感应电路示意图

RFID是一种近距离、高频的无线通信技术，它具有非常灵活的使用范围，125 kHz和134.3 kHz低频(LF)无源RFID标签的读取距离为10 cm内。在不利于人员操作的恶劣环境下可以使用13.56 MHz高频(HF)无源RFID标签，可在1～1.5 m距离读取电子标签。如果使用2.45 GHz超高频(SHF)有源RFID标签，甚至可以达到最高100 m的超远距离读取[6]。

1.4　系统的基本框架

系统的基本框架如图2所示[7]。

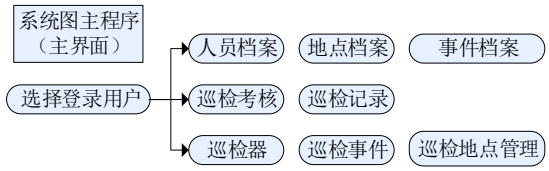


图2　主程序组成模块框架

2　系统设计

2.1　系统总体布局

整个巡检系统包括系统主模块和子模块两部分，子模块包含巡检管理、资源管理、巡检数据、系统维护四大模块[8]。其中巡检管理包括巡检器设置、下载档案、读取人员档案、读取地点档案、读取事件档案、巡检记录等内容。

由系统主模块支持各子模块功能的实现。如图3所示。

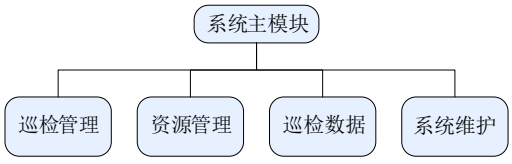


图3　巡检系统总体结构

2.2　巡检系统总体结构

巡检系统主模块的菜单如图4所示[9]。

系统管理▼	资源管理▼	巡检数据▼	系统维护▼	帮助▼
巡检器	巡检地点管理	巡检考核	参数设置	系统操作指南
操作员登录	巡检人员	读卡记录	网络设置	关于…
操作员注消	巡检器号码		数据备份	
退出	巡检事件		数据恢复	
	设置巡检线路		系统操作人员	
	制定巡检计划		修改密码	
			系统初始化	

注：▼表示下拉菜单

图4　菜单结构

2.2.1 系统主模块的流程图

主程序在运行时首先将所有变量加载到内存，业务逻辑问题则显示登录界面，主程序还检查用户是否输入信息并确定登录，然后用户进入子功能，系统主模块的流程如图5所示。

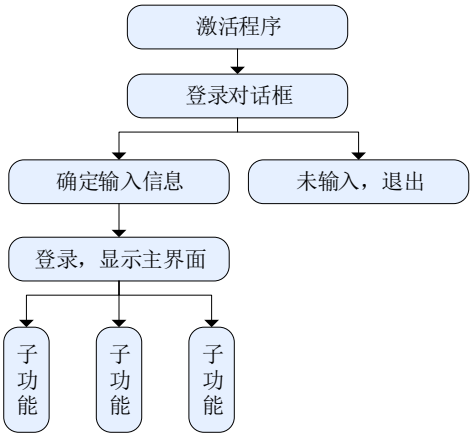


图5 登录系统流程

2.2.2 系统登录文件的制作与说明

为了确保数据不被外人看到，用户在登录以后所做的每个操作，都可以记录下来并保存为日志文件。本系统设计了用户登录环节，当主程序进入以后，要求用户输入密码，再决定是否让用户进入程序。

运行界面如图6所示。

图6 登录界面

如果输入了相关信息并点击确定，登录成功后会显示主界面，如图7所示。主界面上有一系列菜单，包括系统管理、资源管理、巡检数据、系统维护，还有快捷按钮，包括登录、注销、巡检器、计划设置、记录查询、巡检考核、参数设置。

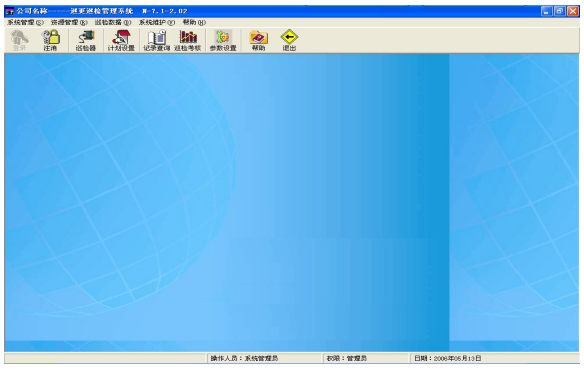


图7 主界面

2.3 数据库设计

在一个应用系统中，数据库能充分描述数据之间的内在联系，人们建立数据库的目的，就是想用数据来反映客观事物及其之间的联系。为了使数据库具有正确反映这种联系的能力，在建立数据库的初期确定能够表达客观事物的数据库模型十分重要。

2.3.1 基本数据表

基本数据表保存了巡检管理系统的主要数据，这些数据包括了记录计划、地点名称、到达时间等信息。表1是巡检记录表，其中只记录了巡检记录的资料，所以用一个数据库表就可完成，并且Visual Basic中所需要的数据是通过控件来输入的，数据库中的数据表项不需要在建立时填入数据，而是在Visual Basic中通过控件传递得到，因此只需要建立空表即可。

表1 巡检记录数据库表

编号	字段名称	类型	长度
001	计划	VARCHAR	50
002	地点名称	VARCHAR	50
003	到达时间	VARCHAR	50
004	人员	VARCHAR	50
005	时间	VARCHAR	50
006	方式	VARCHAR	50
007	机号	VARCHAR	50
008	数据	VARCHAR	50

2.3.2 巡检计划数据表

巡检计划数据表就是要存放制定巡检计划的各种数据，如表2所示。

表 2　巡检计划数据库表

编号	字段名称	类型	长度
001	序号	VARCHAR	10
002	计划时间	VARCHAR	50
003	线路名称	VARCHAR	50
004	最大限时	VARCHAR	50

2.3.3　巡检考核数据表

我们可以对从巡检计划的制定到巡检计划的执行进行考核，并把考核结果存入巡检考核数据表，如表3所示。

表 3　巡检考核数据表

编号	字段名称	类型	长度
001	线路	VARCHAR	50
002	地点	VARCHAR	50
003	计划时间	VARCHAR	50
004	到达时间	VARCHAR	50
005	人员	VARCHAR	50
006	事件	VARCHAR	50
007	考核结果	VARCHAR	50

2.3.4　线路设置数据表和线路中的巡检地点表

在设置巡检线路时，会要求设置巡检线路和线路中的巡检地点，这些数据对于巡检来说至关重要，当然也需要用数据库来保存，如表4和表5所示。

表 4　线路设置表

编号	字段名称	类型	长度
001	代号	VARCHAR	10
002	线路名称	VARCHAR	50
003	最大限时	VARCHAR	50
004	最小限时	VARCHAR	50

表 5　线路中的巡检地点表

编号	字段名称	类型	长度
001	序号	VARCHAR	50
002	巡检区域	VARCHAR	50
003	地点名称	VARCHAR	50
004	间隔时间	VARCHAR	50

2.3.5　巡检事件数据表

巡检事件是巡检计划制定的一部分，其具体数据项如表6所示。

表 6　巡检事件数据表

编号	字段名称	类型	长度
001	序号	VARCHAR	50
002	代码	VARCHAR	50
003	事件名称	VARCHAR	50

3　系统开发

3.1　系统主模块

系统主模块是系统的主要界面体现，它支持系统的各下属子模块功能的实现。在此设计的巡检系统中，包含有巡检管理、资源管理、巡检数据、巡检考核等主要子模块。系统主模块与各子模块通过主模块的支持和聚合作用有机地组成了一个整体[10]，实现了此巡检系统的各项功能。如图7所示。

3.2　巡检管理的设计

巡检器界面与主界面设计类似，在空白窗体中添加一个Sstab控件、一个Datagrid控件，考虑到要连接数据库，本设计采用ADO方式连接，故需要添加6个Adodc控件，并且添加相关控件，界面如图8所示。

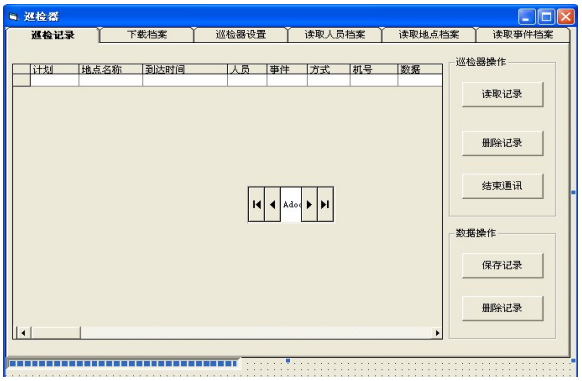


图 8　巡检器的功能标签

设计完成后界面如图9所示。

实现的功能如下：

（1）巡检记录。巡检器可采用RS-8232（或USB）接口与电脑连接，在电脑上可用厂商提供的测试工具来测试是否可以从巡检器上读取数据。

如果测试工具操作成功，再在巡检管理系统中调用Windows通讯API接口将巡检器中的数据库读取到内存，然后存入数据库中。在完成操作后按“结束通讯”，将巡检器置回功能选择状态。

（2）下载档案。巡检器中的档案包括了计划、计划点、人员、事件等，在将档案下载到巡检器时，需要做相应的数量限制，超过该数量则无法执行下载操作。例如，计划数为20个，计划中包含的地点为1024个，计划点总数为6000个，事件及人员数为250个。

（3）读取档案。从巡检器中读取地点、人员、事件等档案。

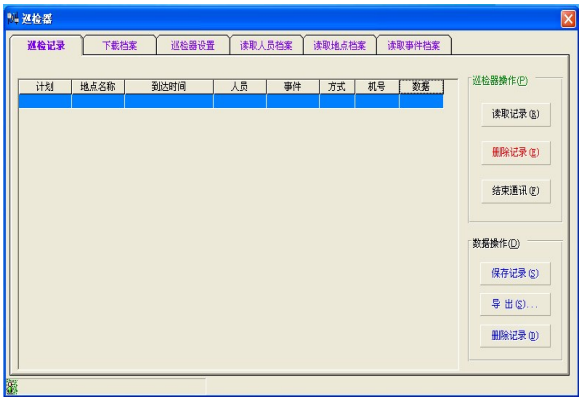


图9 巡检器菜单

4 结语

基于RFID的智能巡更巡检管理系统，可解决监管不完善、施工困难、布线复杂、信息不及时等问题，有效满足自动化巡检的需求，提高了巡检的效率，将巡检水平提高到一个新的高度[11]。笔者后续的工作将致力于将C/S架构管理后台迁移到GIS云平台，巡检器增加GPS定位功能和移动通信（GPRS、3G、4G或5G）技术，实现更加智能、完善的巡更系统。

参考文献：

[1] 李佑群，向道豪，丁德红. 基于二维码的电力设备巡检管理系统[J]. 现代计算机，2020(2)：103-108.
[2] 杨峰，王莹，程仁慧，等. 高速公路智能巡检养护管理系统研究与开发[J]. 智能与信息化，2021(4)：103-106.
[3] 刘莉，文勇军，唐立军. 高校消防设备移动巡检管理系统的设计与实现[J]. 现代计算机，2019(16)：91-96.
[4] 田芳，王珂. 基于5G+4K的无人机高速公路自动化巡检研究与应用[J]. 江苏科技信息，2021(6)：55-60.
[5] 苏婕，王忠. 基于NFC技术的巡更巡检管理系统的设计与实现[J]. 计算机工程与设计，2015，36(4)：1068-1073.
[6] 苏丰，陆志浩，毛颖科. 基于RFID技术的变电站运行巡检管理系统[J]. 电子设计工程，2020，28(3)：129-134.
[7] 张海军，赵雪松. 基于GPS的输电线路巡检管理系统的设计与实现[J]. 电网技术，2005，29(7)：78-82.
[8] 黄莉，李以策，郭江，等. 基于RFID技术的水电厂智能巡检管理系统[J]. 水电能源科学，2016，34(4)：166-169.
[9] 李温温. 基于低代码平台的多媒体教室巡检管理系统建设与应用研究[J]. 电脑知识与技术，2023，19(3)：116-119.
[10] 王志强，何建锋，温喜廉，等. 基于区块链的巡检管理系统的应用与实践[J]. 数字技术与应用，2021，39(10)：119-122.
[11] 邱俊豪，胡常伟，彭磊，等. 面向大型生产车间的设备点检与巡检管理系统设计[J]. 机电工程技术，2021，50(3)：35-39.

Development and design of intelligent patrol inspection management system

Wang Dong[1], Chen Yang[2*]

(1. Department of Finance and Commerce, Qinghai Higher Vocational and Technical Institute, Haidong 810006, China;
2. School of Management, Guangzhou University, Guangzhou 510006, China)

Abstract: Taking the advantages of RFID short range, high-frequency wireless communication technology, combining electronic tag, automation technology, database, management system to design the intelligent patrol inspection management system. This system adopts Visual Basic as development tool, SQL Server as database, ODBC for the connection of system and database, in order to help relevant departments improve and optimize the capability and efficiency of inspection management work.

Keywords: patrol inspection; management system; software development; work efficiency

文章编号: 1007-1423（2023）14-0091-05 DOI: 10.3969/j.issn.1007-1423.2023.14.018

基于中文分词算法和众包协同的高校课程思政资源共享与互助系统

张 露，童颖佳，马 华*

（湖南师范大学信息科学与工程学院，长沙 410081）

摘要：近年来，广大高校教师积极参与“课程思政”教改实践，但却经常面临创新思路单一、参考资源查找效率低、教学资源制作的技能和精力不足等困扰。针对现有研究的不足，设计并实现了一个基于中文分词算法和众包协同的高校课程思政资源共享与互助系统。该系统使用TF-IDF算法和众包机制来整合互联网上的课程思政资源，联结了多方平台和参与者，使用Spring Boot和Vue框架进行开发。该系统由协同互助、课程思政、讨论区、资源检索、个人中心等模块组成，可为高校教师提供课程思政教学的辅助支持，具有良好的应用价值。

关键词：课程思政教学；资源共享与互助；中文分词算法；众包协同

0 引言

自2014年“课程思政”这一理念提出并在上海的部分高校进行试验，取得了较好的成效。习近平总书记不断强调学校作为立德树人的地方，要牢牢抓住思想政治教育工作。2020年，教育部印发《高等学校课程思政建设指导纲要》[1]，高校思政课改革创新，各高校开始逐步探索课程思政模式。由此诞生的思政教育资源互联网平台，为教师提供了一些优质课程思政示范课及其配套资源，协助教师完成专业课和思政教育的有机融合。然而，目前已有平台仍存在如资源分类粒度大、数量有限、搜索困难等问题。

近年来，学者们对课程思政资源的共享和利用进行了研究，但尚缺乏面向高校课程的思政资源的共享与互助平台的系统性研究。例如，汤宇轩等[2]提出一种基于知识图谱的课程思政素材库的构建方案，对课程思政领域进行本体设计、命名实体识别、关系抽取等方法的研究，构建了面向大学生计算机基础课程的课程思政素材库。朱丽等[3]阐述了大数据平台商建立课程教学资源推荐系统的设计、框架及其应用，表示在大数据平台的基础上，课程教学资源推荐系统的建设得到推进。龙丹等[4]以CNKI数据库中刊载的课程思政的相关文献为数据来源，对我国高校课程思政研究的趋势、热点、问题进行调查分析。汪路金等[5]从继续教育思政教学出发，指出思政类网络课程资源的重要性，并提出建设策略。

针对此现状，我们引入TF-IDF(term frequency-inverse document frequency)中文分词算法和众包协同机制，挖掘和整合网络上现有的课程思政教育资源，联结多方平台和参与者以提供一个高校课程思政教育资源共享和互助的全新系统，构建了一个可动态扩充的智慧型课程思政资源库，即“智慧思政文库”。

收稿日期：2023-03-26 **修稿日期：**2023-04-23

基金项目：国家级大学生创新创业训练计划支持项目（202210542045）；湖南省教育科学规划课题研究成果（XJK18BXX001）

作者简介：张露（2002—），女，湖南邵东人，本科生，主要研究方向为个性化学习与推荐；童颖佳（2001—），女，江西宜春人，本科生，主要研究方向为推荐系统；*通信作者：马华（1979—），男，湖南株洲人，博士，教授，主要研究方向为推荐系统和个性化学习，E-mail：huama@hunnu.edu.cn

1 系统建设思路

1.1 爬取资源

国内已有一些课程思政教育资源平台，例如，新华思政(xhsz.news.cn)等。非专业教育学习平台上也散布有许多有借鉴意义和使用价值的课程思政资源，例如，知乎、微博等。面对目前互联网上众多的课程思政资源，我们使用Python开发数据爬虫，用Scrapy框架对不同的网站进行爬虫定制，对它们进行爬取，并分割保存为JSON格式数据。这些初始的资源数据被处理后存入MySQL数据库。

1.2 预处理数据

对JSON格式数据进行初步的处理加工。首先，以本校的院系专业划分以及课程开设为标准，建立基本的学科大类、专业、课程的树形结构分类。其后，利用TF-IDF算法从爬取的资源当中提取出关键词，根据关键词判断资源所属类别，从而对资源进行分类。最终，将处理过的数据存入本地数据库，供平台后续使用。

1.3 功能开发

使用Spring Boot框架和Vue进行平台的总体设计和开发。主要的模块包括协同互助模块、课程思政模块、讨论区模块、资源检索模块、登录注册模块、个人中心模块等。通过引入众包协同机制，平台允许用户自主上传资源和资源互助，以进一步扩充资源库。通过用户的自主上传，资源库得以不断扩充，而资源协同互助的创新功能，为教师和创作者的各取所需提供了交流和协作的渠道。

2 资源库建设及动态化构建

2.1 资源爬虫设计

本文从新华思政网站、CNKI、知乎、微博等爬取若干课程思政相关的各类资源，如课件、教学视频、教案等，作为资源库的一部分，并为之后的标签提取提供基础的数据支持，使用基于Python语言的Scrapy爬虫框架，针对不同网页开发定制个性化爬虫程序。对爬虫抓取的原始Web页面进行页面处理，通过去掉Web页面中的大量噪声数据，将Web页面转化为纯净统一的文本格式和元数据格式。

对初步处理得到的文本格式和元数据格式的数据进行进一步的加工，利用TF-IDF算法提取数据中的关键词，根据关键词判断资源所属类别，从而对资源进行分类，包括资源所包含的思政元素、所属课程、知识点等标签，将分类后的数据导入数据库中，作为资源库初始数据。

2.2 基于TF-IDF算法的中文关键词抽取

本文使用一种改进的TF-IDF算法[6]提取关键词。首先，计算类别间离散度，类别间离散度越高，词语的类别区分能力越强。计算公式如公式(1)所示。

$$D(t)=\begin{cases}\dfrac{\sqrt{\dfrac{1}{m-1}\sum_{i=1}^{m}\left(f_i(t)-\overline{f(t)}\right)^2}}{\overline{f(t)}}, & \overline{f(t)}\neq 0\\ 0, & \overline{f(t)}=0\end{cases} \tag{1}$$

其中：$f_i(t)$为类别C_i中包含术语t的文档数量；$\overline{f(t)}$为特征词在所有类别中出现频率的平均值，m为类别数。

之后，计算类别内信息熵，类别内信息熵越高，词汇的类别区分能力越强。使用公式(2)和(3)计算。

$$E(t,C_i)=-\sum_{j=1}^{n}e_j \tag{2}$$

$$e_j\begin{cases}\dfrac{Nd_j}{NC_i}\log_2\dfrac{Nd_j}{NC_i}, & NC_i\neq 0 \text{ and } Nd_j\neq 0\\ 0, & NC_i=0 \text{ and } Nd_j=0\end{cases} \tag{3}$$

类别间离散度或类别内信息熵越高，类别区分能力越强。因此，引入公式(4)表示特征项对类别C_i的区分能力，即类别判别：

$$CD(t,C_i)=D(t)*E(t,C_i) \tag{4}$$

最终得到TF-IDF的计算公式如公式(5)所示。

$$w_{ij}(t)=tf_{ij}*CD(t_j,C_i) \tag{5}$$

其中：w_{ij}为特征项t对类别C_i的权重；tf_{ij}为特征项t_j的词频；$CD(t_j,C_i)$为特征项t对类别C_i的类别判别，可由公式(1)～(4)计算。在改进的TF-IDF算法中，用类别识别$CD(t_j,C_i)$代替逆文

档频率IDF，从而弥补了传统TF-IDF算法没有考虑特征项的类别内和类别间分布的缺陷。

2.3　基于众包协同机制的资源库扩充与资源互助

一方面，利于设定好的爬虫程序，定时对目标平台进行资源爬取，根据内容发布时间，过滤掉已经爬取过的内容，处理存入资源库中。另一方面，设置“协同互助”模块，用户通过该模块所创作的个性化资源，经用户同意，也将录入资源库并公开展示。同时，用户也可以通过个人中心自主上传资源。以上两方面构成了资源库动态扩展的数据来源，使得资源库能够得到智能自主扩充。

平台所设置的“协同互助”模块显示已经发布的用户需求列表。教师可以发布相关需求，设定所需资源信息；创作者可以搜索感兴趣的需求申请承接。当需求有承接申请，教师可自主选定需求的最终承接者。由此产生的需求订单，平台将跟踪需求的完成状态，并以进度条的形式展示给相关用户。创作者可以是有闲暇时间且有相关创作技能的非教师用户、普通教师用户等，此机制将各类用户聚合在一起，共同完成资源的最终产出。另外，平台利用Java-WebSocket框架，实现了“私信交流”功能，使得教师与创作者能在系统中进行实时交流，教师可随时提出修改意见，协助成品的不断完善，便于需求的更好完成。

3　智慧思政文库系统的开发

3.1　基于Spring Boot和Vue的总体开发

3.1.1　基于Spring Boot框架的后端设计

（1）建立实体层，其中课程类包括id、名称、简介、一级学科、二级学科、目录、选课人数等属性；资源类包括id、类型、简介、上传时间、名称、链接、标签、来源、收藏量、下载量、所属章节等属性。

（2）建立数据访问层，封装对数据库的访问，CourseDao.java和ResourceDao.java用于对数据库中的课程信息表和资源信息表进行增删改查操作。

（3）建立Spring Boot配置文件，设置Spring Boot配置文件中的数据库连接参数和访问端口信息等。

（4）建立业务层，存放业务逻辑处理，提供控制器层调用的方法。

（5）建立控制器层，接收前端传过来的参数进行业务操作，并返回对应的数据。

3.1.2　基于前后端分离的系统框架设计

（1）使用Vue框架编写前端页面，使用Axios向后端发送HTTP请求。

（2）后端控制器层从HTTP请求中获取信息，提取参数，将其分发给业务层处理不同的服务。

（3）业务层调用数据访问层对数据库进行访问，并将其处理成JSON格式返回给控制器层。

（4）控制器层将数据返回给前端页面，前端根据不同的数据所在的位置进行页面的局部刷新。

3.2　系统组成及主要功能模块

基于上述设计，“智慧思政文库”系统包含了协同互助模块、课程思政模块、讨论区模块、资源检索模块、登录模块、注册模块和个人中心模块。系统的功能模块结构如图1所示。首页的界面设计如图2所示。课程资源的浏览界面设计如图3所示。

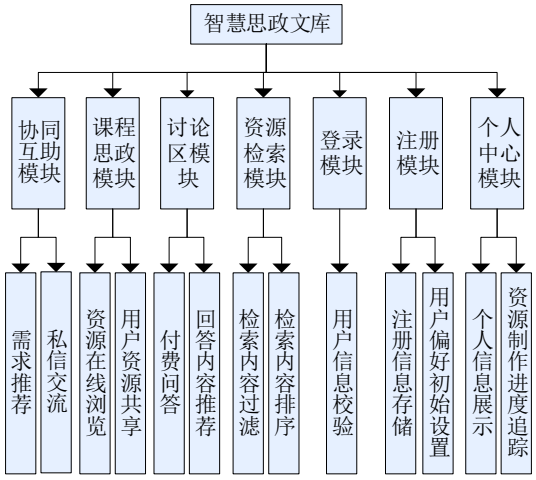


图1　系统功能模块

图 2 首页设计界面

图 3 课程资源浏览的设计界面

3.2.1 协同互助模块

在系统提供的第三方技术保障之下，教师基于自己或者其他用户上传的教学思路在此发布互助需求，创作者对于资源创作提供技术支持。教师可根据需求复杂度设置创作者人数及其分工。教师与创作者紧密合作，共同参与到资源的开发过程，教师提供资源设计理念和想法，创作者依据教师提出的需求制作出满足要求的资源。在整个过程中，创作者和教师在私聊区域进行资源内容开发的深入交流，在交流过程中逐步完善个性化资源的制作。制作完成后，教师对制作出的成品进行验收，结束整个互助流程。同时，平台将追踪资源互助完成进度，推动创作流程的顺利进行，促进生成更多优质资源。

3.2.2 课程思政模块

系统将资源所属课程根据学科大类进行分类，用户点击对应课程可查看相关的资源，资源类型分为课件、教案、思路、视频、论文。同时平台内的用户也可以进行资源上传操作，实现资源共享，用户上传的资源若评分较高，可以提升平台内的信誉值。

3.2.3 讨论区模块

该模块功能是发布问题以及回答问题，其中问题分为普通问题以及付费问题，付费问题用于激励平台内用户输出更为专业的解答，用户可以在个人中心修改付费问答的基本信息，以及选择最优回答给予酬金。

3.2.4 资源检索模块

检索的内容包括平台内的资源、需求、问题以及回答，根据用户在平台内资源的浏览记录、资源评分、创作记录等信息，分析用户潜在的资源浏览偏好，使用混合推荐算法对数据库中检索资源进行过滤及排序。

3.2.5 登录模块

用户提交登录信息后，向 Spring Boot 控制器层发送登录验证请求，访问数据库获取持久化对象，当对象密码校验正确时，用户成功登录系统，以便系统对用户提供资源协同互助的支持及个性化的资源推荐。

3.2.6 注册模块

用户提交注册信息后，向 Spring Boot 的控制器层发送注册请求，为用户生成持久化对象存储于数据库中。其中注册信息包括用户名、密码、所在专业、感兴趣的思政元素及课程等，以解决用户初始登录系统资源推荐冷启动问题。

3.2.7 个人中心模块

该模块用于公开展示用户的个人信息，包括用户名、所在院校、所在专业、自我介绍、兴趣爱好等，并提供用户的资源访问入口，包括互助情况、讨论情况、所发布的资源等。对用户本人开放资源制作进度追踪，确保双方能在指定时间完成资源制作。

4 结语

本文开发的基于中文分词算法和众包协同的高校思政教育资源共享和互助系统，使用了 Python 爬虫、Spring Boot 和 Vue 技术，基于 Scrapy 爬虫采集互联网上的课程思政资源，利用 TF-IDF 改进算法提取关键词并将其关联到知识点，通过用户的资源协同互助实现资源库的动态扩充，该系统借鉴了众包和协同思想，使用

界面简洁清晰，可为从事教学改革的高校教师和从事资源开发的专业技术人员搭建合作共赢的平台。可为广大教师提供课程思政教学的辅助支持，具有良好的应用价值。未来研究中，我们拟引入知识图谱技术对资源所包含的思政元素及知识点间的关联关系建模，进一步提高用户资源搜索的满意度和效率。

参考文献：

[1] 中华人民共和国教育部. 教育部关于印发《高等学校课程思政建设指导纲要》的通知[A/OL]. http://www.moe.gov.cn/srcsite/A08/s7056/202006/t20200603_462437.html.

[2] 汤宇轩,齐恒,申彦明,等. 基于知识图谱的课程思政素材库构建[J]. 软件导刊,2022,21(7):214-219.

[3] 朱丽,付海涛,冯宇轩,等. 基于大数据平台的课程教学资源推荐系统应用探究[J]. 计算机产品与流通,2020(6):209.

[4] 龙丹,卢一舟. 基于CiteSpace的高校课程思政研究的热点及趋势分析[J]. 当代教育理论与实践,2022,14(1):1-9.

[5] 汪路金,李庆玖,彭文兵,等. 继续教育思政类网络课程资源建设策略研究[J]. 成都中医药大学学报(教育科学版),2021,23(4):114-116.

[6] YI J, YANG G, WAN J. Category discrimination based feature selection algorithm in Chinese text classification[J]. Journal of Information Science & Engineering,2016,32(5):1145-1159.

A sharing and mutual assistance system of ideological and political education resources for university courses based on Chinese word segmentation algorithm and crowdsourcing collaboration

Zhang Lu, Tong Yingjia, Ma Hua*

(College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China)

Abstract: In recent years, more college teachers have participated in the teaching reform practice of "Curriculum Ideology and Politics". However, they often face the problems of single innovative ideas, low efficiency in finding reference resources, and insufficient skills and energy in creating teaching resources. To address the shortcomings of the existing research, a sharing and mutual assistance system of ideological and political education resources for university courses is designed and implemented based on Chinese word segmentation algorithm and crowdsourcing collaboration. The system uses the TF-IDF algorithm and crowdsourcing mechanism to integrate ideological and political resources of courses on the Internet, connects multiple platforms and participants, and is develops by using Spring Boot and Vue frameworks. The system consists of modules, including collaborative assistance, curriculum ideology and politics, discussion forum, resource retrieval, personal center, and on so, and can provide auxiliary support for college teachers in ideological and political education in course teaching and has good application value.

Keywords: ideological and political education in course teaching; resources sharing and mutual assistance; Chinese word segmentation algorithm; crowdsourcing collaboration

文章编号: 1007-1423（2023）14-0096-05 DOI: 10.3969/j.issn.1007-1423.2023.14.019

基于 Web3D 的中医嗅诊智能分析系统的研究

区锦锋，李振鹏，廖嘉镇，黄飞雄，周华英*，张 琦

（广东药科大学医药信息工程学院，广州 510006）

摘要：人体气味信息的复杂性和不稳定性，使得研究者们对中医嗅诊的研究远远滞后于其他面诊、脉诊等。基于仿生嗅觉技术和人工神经网络技术建立中医嗅诊智能分析模型，并将Web3D可视化技术应用于中医嗅诊智能化分析与研究，实现以直观生动的三维场景画面展示中医嗅诊的复杂分析过程。此研究将有助于促进中医四诊信息融合和中医远程辅助诊疗。

关键词：Web3D；中医嗅诊；三维可视化

0 引言

中医四诊“望、闻、问、切”是中医诊断的核心与精髓。人体因各种疾病会引起自身气味的改变，但由于人体气味信息的复杂性和不稳定性，导致中医嗅诊的研究远远滞后于其他舌诊、脉诊等[1-3]。中医嗅诊仍停留在凭医生的主观嗅气味诊法，缺乏客观标准。实现中医闻诊—嗅诊数字化和智能化分析是众多中医人多年努力的目标和期望。因此，为了促进中医四诊智能诊治融合发展，中医嗅诊智能化分析研究显得尤为迫切[4-5]。林雪娟等[6]、吴敏等[7]运用电子鼻采集口腔气味，研究得出可以根据口腔气味特征初步判断 2 型糖尿病的虚实病性。王忆勤[8]开展四诊信息融合辨证模型及中医四诊检测系统的研究，指出中医四诊信息融合在标准化、规范化方面尚处于起步阶段，在四诊信息融合研究及中医诊断智能平台研发方面尚需加强。

本文基于仿生嗅觉技术，对中医嗅诊展开数字化分析与研究。利用电子鼻获取口腔气体的气味信息，并对气味信息进行数字化分析和研究。同时，通过三维可视化技术实现中医嗅诊智能分析过程的三维可视化。本研究对中医嗅诊数字化、中医远程医疗等具有重要的理论意义和应用前景，有利于促进中医四诊数字化、中医四诊信息融合发展。

1 研究思路

本文基于 Web3D 的中医嗅诊智能化分析可视化，需要通过数据服务实现中医嗅诊数据的网络传输，然后在前端以网页的形式为用户呈现三维可视化的渲染结果。因此，本文从数据服务出发，将中医嗅诊三维数据全部存储于服务器中，并设计一个基于 Web 的中医嗅诊智能分析的三维可视化系统，在客户端的网页中进

收稿日期：2023-03-29 **修稿日期：**2023-06-30

基金项目：2022年全国大学生创新创业项目（202210573002）；2023年广东省科技创新战略专项资金（pdjh2023b0273）；2023年广东省中医药管理局项目（20231221）

作者简介：区锦锋（2002—），男，广东人，本科生，研究方向为中医药信息处理、计算机视觉；李振鹏（2003—），男，广东人，本科生，研究方向为Web3D前端应用及开发；廖嘉镇（2004—），男，广东人，本科生，研究方向为中医药信息处理；黄飞雄（2002—），男，广东人，本科生，研究方向位中医药信息处理；*通信作者：周华英（1974—），女，湖南邵阳人，博士，讲师，研究方向为中医药信息处理、模式识别等，E-mail：287059250@qq.com；张琦（1984—），女，吉林长春人，硕士，讲师，研究方向为医药虚拟仿真应用、计算机视觉

行三维可视化渲染处理，为用户呈现中医嗅诊智能分析的三维可视化效果，整体设计思路如图 1 所示。

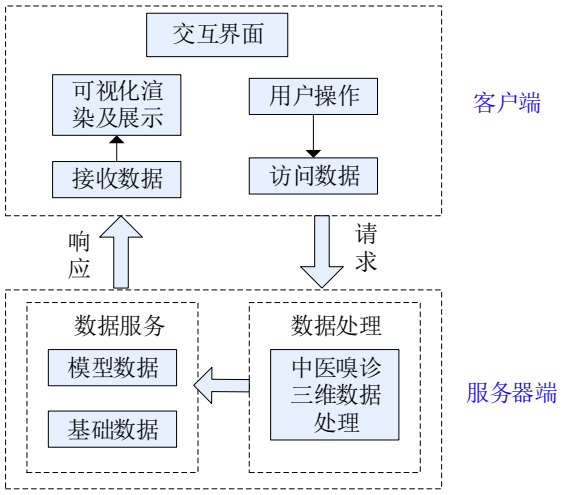


图 1　中医嗅诊智能分析可视化系统整体设计思路

本项目操作过程包括：客户端页面主要进行用户的交互操作处理以及三维图像的可视化渲染处理并展示；服务器端主要进行中医嗅诊智能分析三维数据的处理以及可视化数据的数据服务存储与发布，二者通过网络请求相互传输信息，其大致流程有以下三步：

（1）客户端页面根据用户的具体操作向服务器发起对应的数据请求；

（2）服务器端接收来自客户端的请求，选出符合条件的数据传输回客户端；

（3）客户端接收服务器响应，获取对应数据，再组织数据并在浏览器页面可视化渲染和显示。

2　研究方法和研究内容

针对本文的研究目标，设计的技术路线如图 2 所示。

根据图 2 的技术路线，我们将从以下四个方面重点开展研究。

2.1　气味信息采集与嗅诊数字化表征

项目组拟选择 100 例慢性肺炎住院病患者作为研究对象，100 例健康者作为对照组。研究对象均来自广州市某三甲医院呼吸与危重症医学科的慢性肺炎住院患者，对照组来自同一医院体检科的健康志愿者，所有数据采集均征求本人及家属同意。

根据口腔气味采集要求采集病患组和对照组的口腔气体，并利用电子鼻系统检测样品气味信息，检测完毕保存好检测数据，形成嗅诊初始数据库。根据电子鼻检测的口腔气体气味信息库，提取气味特征信息（例如曲线的起始水平、上升速率、变化速度、方差、均值、最终稳态值等）；建立一种数据结构，对气味特征进行描述，用以存储嗅诊参数信息（包含患者类别、检测平均值、最大值、方差、标准差等），实现中医嗅诊数字化表示。

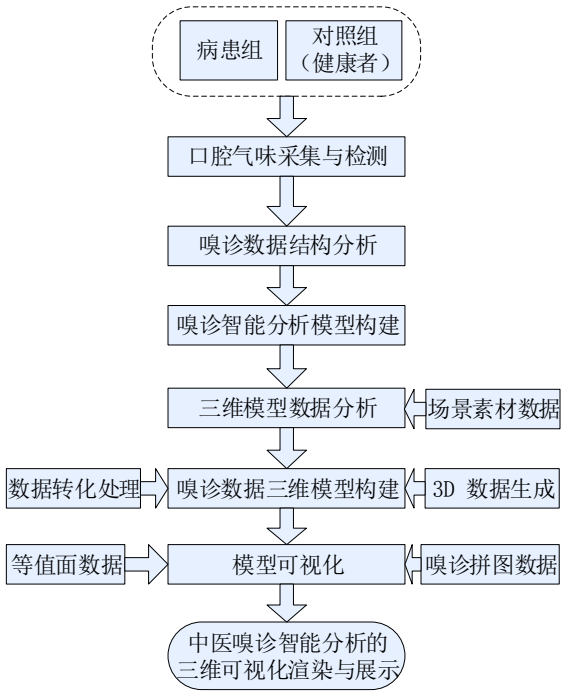


图 2　中医嗅诊智能分析可视化系统技术路线

2.2　中医嗅诊智能分析模型构建及实现

将针对 200 个实验样本，按照 7 ： 3 分成训练样本和测试样本。同时构建多层结构的人工神经网络：输入层、多个隐含层和输出层，并设置初始参数进行模型训练和学习，技术路线如图 3 所示。

课题组根据多层神经网络算法，研究激活函数：ReLU、Sigmoid、Tanh 等各函数特点，选

择适合气味特征的激活函数，不断训练和更新各层的权值和偏置值。同时，训练模型时根据训练误差条件选择学习率α和batchsize参数值大小，满足多层神经网络既充分训练又要避免训练时间过长而出现过拟合。训练结束，保存各参数，完成模型构建与实现。

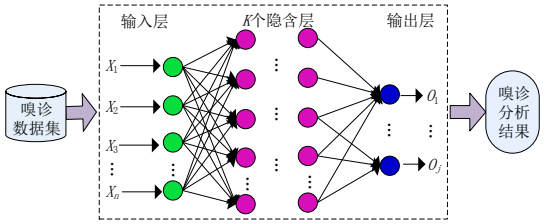


图3 中医嗅诊智能分析模型

2.3 中医嗅诊智能分析平台的三维数据分析及三维模型构建

根据中医嗅诊智能分析的场景设计，进行三维数据分析及三维模型设计。通过将源数据中的信息依次转换为几何结构、索引结构以及纹理结构三部分，完成数据转换处理[9-10]。然后根据三部分存储的信息以及3D数据的结构规范构建一个完整的3D模型数据，为后续的基于Web的三维可视化实现做准备。

2.4 中医嗅诊智能分析平台的三维可视化展示

服务器端将采用SSM架构分层设计中医嗅诊智能分析3D可视化平台，底层的数据内容除了常规参数数据外还包括三维数据；业务层主要对相关数据进行处理，即根据生成的中医嗅诊智能分析平台的三维数据的3D模型，通过对数据参数的逻辑处理将信息在视图层展示；视图层将基于Web GL技术实现3D可视化场景渲染与展示，完成Web视图页面的三维展示功能[9,11]。客户端无需安装其他APP应用软件，直接通过浏览器运行用3D引擎——Three.js组件，自动构建三维可视化场景并渲染展示[12-13]。

3 研究结果

3.1 获取口腔气味信息

本文利用电子鼻检测病患者和对照者的口腔气味，如图4为口腔气体气味信息检测过程。检测完毕保存好检测数据，形成嗅诊初始数据库。

图4 口腔气体气味信息检测

3.2 构建中医嗅诊智能分析模型

利用深度学习算法的平移、扭曲、缩放具有一定程度的不变性，研究深度学习算法参数优化方法，构建基于深度学习算法的中医嗅诊智能分析模型和系统，实现中医嗅诊数字化和智能化分析，系统界面和分析结果分别如图5和图6所示。

图5 中医嗅诊智能分析系统主界面

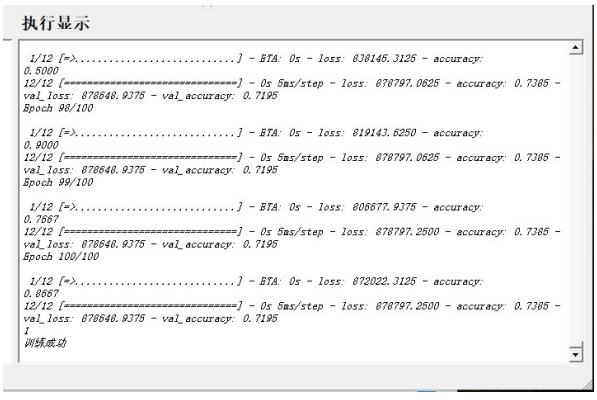


图6 中医嗅诊智能分析结果

3.3　三维数据模型构建

在三维数据可视化系统中，核心内容是营造三维虚拟场景，而虚拟环境的建立首先要进行三维数据建模，然后在建模的基础上再进行渲染显示。因此三维数据建模既是基础又是关键技术，通过将源数据中的信息完成数据转换处理，构建一个完整的 3D 模型数据是实现中医嗅诊智能分析平台三维可视化的核心内容。图7所示为三维可视化的一个截图。

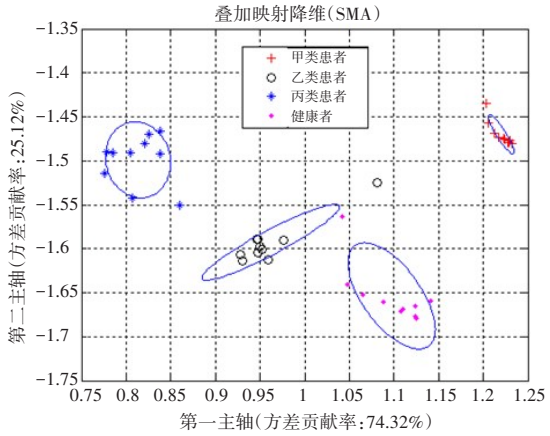


图 7　中医嗅诊智能分析三维可视化

4　结语

中医嗅诊信息客观化采集和数字化方法研究是中医四诊合参数字化、智能化发展的重要基础。本文通过将 Web3D 的可视化技术应用于中医嗅诊智能化分析与研究，通过Web3D技术完成中医嗅诊三维数据建模及建立中医嗅诊智能化分析的三维可视化平台，旨在依托互联网，通过直观生动的三维场景画面有效地展示中医嗅诊的复杂分析过程，有助于推动我国传统中医医术和中医文化在全世界范围内的广泛宣传和推广。本研究为中医嗅诊数字化研究、中医远程医疗等提供新方法，有利于促进中医四诊数字化、中医四诊信息融合发展。

参考文献：

[1] 林雪娟,梁丽丽,刘丽桑,等. 基于证素辨证的慢性胃炎常见病位间的气味图谱特征研究[J]. 中华中医药杂志,2016,31(10):3966-3969.

[2] 赵文,张佳,徐佳君,等. 四诊合参智能化发展现状及实现路径[J]. 中医杂志,2020,61(1):58-62,67.

[3] 宋雪阳,许朝霞,王寺晶,等. 中医闻诊客观化临床应用研究概述[J]. 中国中医药信息杂志,2019,26(3):146-149.

[4] 顾星,刘务勤,黄杨,等. 中医四诊数字化技术的应用前景[J]. 中国科技论文,2006,1(4):248-253.

[5] 张尚尚,杨学智,张健,等. 基于四诊合参辅助诊疗系统的抑郁症诊断研究[J]. 中国中医药信息杂志,2015(5):101-106.

[6] 林雪娟,李灿东,吴青海,等. 基于电子鼻技术的不同体质青年学生口腔呼气气味图谱研究[J]. 中华中医药杂志,2012,27(12):3184-3188.

[7] 吴敏,林雪娟,连梨梨,等. 基于电子鼻的热证患者口腔呼气的气味图谱特征分析[J]. 中华中医药杂志,2020,35(1):135-138.

[8] 王忆勤. 中医诊断技术发展及四诊信息融合研究[J]. 上海中医药大学学报,2019,33(1):1-7.

[9] 张永强,王波,申茂廷. 基于Cesium的3DWebGIS三维场景加载及开发[J]. 河南科技,2021,40(21):8-10.

[10] 刘璐,傅拥钢,高月娥. 基于WebGL3D技术的综合能源服务仿真环境及场景设备的构建研究[J]. 科技创新与应用,2021(5):152-154.

[11] 尹千慧,贺鹏飞,王玺联,等. 基于WebGL的3D立库可视化系统设计与实现[J]. 信息技术,2021(3):84-88.

[12] 张磊,王书旺,杨红兵,等. 基于WebGL的三维可视化展现系统及数据可视化方法:CN112256790A[P]. 2021.

[13] 徐鹏. 基于Unity3D开发平台的Web3D技术的数字科技馆展示系统设计制作[J]. 科技与创新,2021(3):51-52.

（下转第104页）

文章编号: 1007-1423（2023）14-0100-05　　DOI: 10.3969/j.issn.1007-1423.2023.14.020

基于云GIS的湖北省硒资源平台建设

杨　淑[1]，胡伟路[2,3*]

（1.湖北省地质科学研究院，武汉 430034；2. 中冶集团武汉勘察研究院有限公司，武汉 430080；
3. 中冶武勘智诚（武汉）工程技术有限公司，武汉 430073）

摘要：为提高湖北省硒资源数据信息化水平，实现硒资源大数据的应用，基于云GIS技术构建了湖北省富硒资源信息平台。以地图结合统计图表的云GIS形式建立湖北省富硒资源“一张图”，通过可视化大屏直观动态地展示湖北省硒资源调查的工作程度、采样点信息、调查区硒资源水平、富硒农作物等数据的统计分析情况，同时整个平台也作为湖北省富硒数据的管理应用平台。建设结果表明，开发的湖北省富硒资源信息平台实现了湖北省硒资源信息的动态可视化展示功能，为湖北省硒资源大数据的管理和应用提供了技术支撑。

关键词：硒资源；云GIS；一张图；动态可视化

0　引言

硒（Selenium）是一种非金属元素，化学符号是Se。硒是稀散非金属之一，可用于冶金等众多领域，同时也是人和动物体内不可或缺的微量元素，有提高免疫力、抗衰老、延年益寿的作用[1-2]。在世界上很多地方，土壤、农作物中均含硒元素，近年来我国随着土壤地球化学调查工作的开展，富硒土壤及作物得到了专家学者的关注[3-6]。自2014年以来，湖北省陆续完成了恩施州全域、洪湖市全域以及全省主要产粮区的土地质量调查工作[7]，调查工作查明了大量的硒相关信息，包括采样点的土壤硒含量、农作物硒含量、土壤富硒等级、富硒产业园区等数据。

随着信息技术的不断发展，地质工作与信息技术在不断地融合，中国地质调查局建成了地质云平台，提升了地质数据的信息服务能力和水平[8]，湖北省地质局也开发了湖北省地质局地质大数据平台作为地质云的重要节点[9]。地质大数据平台的建设能够推进地质工作的信息化和智能化，提高地质工作的服务水平。湖北省土地质量地球化学评价工作虽然获取了大量富硒资源调查数据、评价报告及评价图件，而且目前每年还有大量的新增数据，但这些数据和成果缺乏统一的在线数据管理和展示平台，造成信息查询、更新、分析、共享不方便，难以适应网络化传播环境和政府宏观智能管理的要求，巨大的信息需求与滞后的信息供给形成强烈的反差，亟需一种信息化管理工具将所有环节的信息规范化、标准化地管理起来，以便信息得到有效利用。

本文提出的基于云GIS构建的硒资源平台能够充分发挥硒资源信息的价值，将湖北省硒资源信息与湖北省时空地理信息相结合，实现硒资源的精细化管理和精准化服务。

1　硒资源大数据及其可视化

1.1　硒资源大数据概念

本文研究涉及到的硒资源数据主要指湖北省土地质量地球化学调查工作中得到的土壤硒元素含量、农作物硒元素含量、富硒产业相关

收稿日期：2023-05-25　　**修稿日期：**2023-06-12

作者简介：杨淑（1994—），女，山东菏泽人，硕士研究生，助理工程师，研究方向为地理信息技术、地质信息化；*通信作者：胡伟路（1993—），男，河北邢台人，硕士研究生，工程师，研究方向为地理信息软件开发技术，E-mail:825027247@qq.com

数据及其相关评价图件等数据。目前湖北省土壤硒资源采样面积达到45326平方千米，采样点数据达到上百万条，硒资源数据具有数据量大的特点；数据类型有采样表格数据、文字说明数据、地图数据、图片数据等，包含结构化、半结构化和非结构化的数据，数据类型呈现出多样性；硒资源数据的采样工作由专门的技术人员负责，样品分析由湖北省地质实验测试中心承担，并且由中国地质调查局区域化探样品分析质量检查监督组负责质量监控，数据质量较高。同时它还具备硒资源数据的多源性包括不同地域的土壤、各类农作物中的含量等，硒资源数据在不同地质背景的规律性反映体现了客观性以及硒资源数据之间、与其他元素之间的相互联系、相互作用的综合性特点。

1.2　硒资源大数据可视化方法

本文通过地图结合图表的大屏可视化形式做展示，主要通过饼状图、折线图和硒资源调查区分级设色评价图等方式，综合使用分类统计表法，分区统计表法等方法，采用柱状图、饼图等方式反映硒资源数据的变化情况，如图1所示。柱状图的横轴表示富硒产业基地的数量，纵轴表示富硒产业区类别；饼图中每一个板块代表富硒农作物的占比。平台通过数据集管理数据，将数据过滤、聚类、回归，实现同一份数据的多维度分析，同时采用不同种类的可视化方法去展现数据的统计分析结果，极大地提高了数据的直观性、易读性和空间对比性。

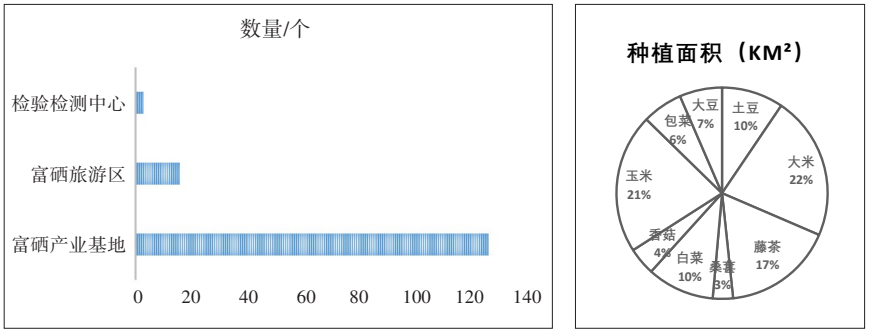


图1　数据统计图

2　平台总体架构

云GIS是在云计算环境中提供GIS服务的，它的软硬件支撑以及数据存储、计算等都在云服务器进行，用户只需拥有上网的终端即可[10]。基于云GIS的硒资源信息平台的建设需要紧密围绕硒资源的相关业务需求，按照统筹规划的设计原则，制定平台的总体架构，充分利用云GIS和大数据等技术，采用成熟先进的技术方法进行平台研发，实现硒资源的管理、查询、展示、分析和共享功能。平台总体架构使用传统的模型进行设计，如图2所示，包含基础设施层、数据层和应用层。

基础设施层是平台运行的重要支撑环境，主要包括服务器、存储、网络以及常用的基础软件。基础设施服务层通过将物理层硬件资源共享成资源池，并通过统一资源调度管理对外提供所需的物理资源。

数据层采用云端协同架构，为整个平台提供数据处理、存储、接入以及访问服务，在整个平台中起到了数据支撑的作用。数据层使用SQL语句完成业务逻辑的数据读写任务，利用完全开源免费、且易于做读写分离、负载均衡、稳定高效的PostgreSQL数据库[11]实现对业务数据的存储与管理。

应用层主要实现用户与平台之间的交互操作、可视化展示。应用层利用HTML结合CSS的前端开发技术实现平台的可视化交互页面。本研究以高德地图作为硒资源的可视化展示组件，

图 2　平台总体架构

以 Echarts 作为统计信息表达组件。高德地图可构建功能丰富、地图展示效果美观、交互性强的地图应用，具有稳定的服务可用性、快速的服务响应等优点；Echarts 是一个开源可视化库，涵盖各行业图表，能够满足各种需求[12]。本平台利用高德地图中的地图服务实现地图应用功能，结合 Echarts 的数据可视化框架，实现统计信息的可视化。

3　平台建设

基于平台的总体设计方案，基于云 GIS 的硒资源信息平台主要由五部分组成，分别为平台首页、一张图、产品中心、硒科普和相关政策页面，如图 3 所示。

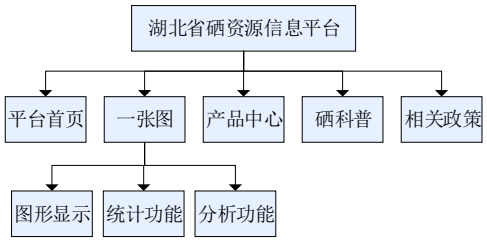


图 3　平台功能设计

3.1　平台首页

平台首页为整个平台客户端的门户，如图 4 所示，主要包含用户登录、图片轮播以及其他页面的索引内容。

图 4　平台首页

3.2　一张图

一张图是整个平台的核心模块，具有 GIS 常用的地图功能，包括地图放大、缩小、复位、全图显示、属性查询、面积量算、距离量算、图层透明度调节等功能，还可以将图层叠加显示。同时还具有统计分析和结果图表显示功能，提供选中数据范围内不同等级富硒土壤面积、富硒农作物组成结构、富硒产业基地数量、不同类型旅游资源点结构等的实时统计，并且以动态图表的形式直观展示，除此之外平台还开发了专门针对富硒农产品的种植选址分析功能模块，该功能结合硒元素含量、重金属元素、有机碳、土地利用数据分析某一区域的土壤富硒产品种植适宜性，为富硒产品基地选址提供技术支撑。

3.3 产品中心

产品中心分类展示富硒相关农产品信息，分为富硒茶叶、富硒大米、富硒禽蛋、富硒饮品四大类，提供每个产品的名称、图片、购买方式、产地富硒土壤等级、产品含硒量、特征描述等信息。产品中心生态的参与者主要有用户、开发者和经销商。其中用户可以通过我们的平台了解富硒产品、购买富硒产品；经销商可以将符合标准的富硒农产品放在我们的平台进行推广；开发者完成以上业务的运营和维护。此模块主要服务于富硒农产品推广，助力乡村振兴。

3.4 硒科普

硒科普模块主要展示硒元素以及硒与人体健康方面的科普知识，科普内容以文字、图片、视频、漫画等形式展示，做到浅显易懂、言简意赅、寓教于乐、生动诠释，实现知识与趣味相结合，让平台的硒科普模块成为公众了解硒资源的一个重要窗口，进一步提高公众对硒资源的认识，提升硒资源保护的认识，促进硒相关产业的科学持续发展。

3.5 相关政策

相关政策作为硒相关领域信息的聚合站，提供相关政策、行业资讯、精品文章、专题报道，依托语义分析、爬虫和并行计算技术，对相关的资讯全网采集、自动过滤、分类，为用户提供最新的相关政策和优质的资讯。

4 结语

本文平台以提高湖北省富硒资源利用为目标，基于B/S架构，综合运用云GIS、数据分析和数据动态可视化技术，研究湖北省富硒资源平台建设，实现富硒资源信息在线的管理、查询、展示、分析和共享功能。

富硒资源信息平台充分收集湖北省的硒相关数据，建成硒专题资源数据库，可存储、更新、显示、查询硒资源信息，减少了数据的冗余、节约了存储空间，减少了信息的滞后性，提高了数据的利用效率，满足了硒资源的智能管理。平台的建设过程，也将湖北省多年来在富硒产业研究工作中取得的数据、工作成效进一步发挥作用，落实了湖北省地质局加强数字地质云平台建设的意见，健全了地质大数据体系建设。硒资源信息平台推进了硒资源大数据汇聚、加工及产品服务，进一步提高富硒相关工作数字化程度和资料的利用程度。平台依托湖北省长江经济带耕地质量地球化学调查成果，在“富硒资源一张图”模块，将全省富硒产业基地信息与一张图相结合，实现富硒产业基地的可视化展示，提升湖北省富硒产业基地的知名度，“产品中心”模块，为富硒产品提供一个宣传平台，提升富硒特色产品的核心竞争力，从而增加农民收入，服务于农民脱贫致富，服务于乡村振兴战略。

参考文献：

[1] BIRRINGER M, PILAWA S, FLOHE L. Trends in selenium biochemistry[J]. Nat. Prod. Rep., 2002, 19(6): 693-718.

[2] 许凌凌，许月明. 微量元素硒与人体健康研究[J]. 农技服务，2016，33(6)：85-86.

[3] 闫加力，徐春燕，杨军，等. 湖北省不同农作物对土壤硒吸收富集规律的研究[J]. 资源环境与工程，2019，33(4)：481-485.

[4] 袁知洋，许克元，黄彬，等. 恩施富硒土壤区绿色富硒农作物筛选研究[J]. 资源环境与工程，2018，32(4)：569-575.

[5] 梁红霞，侯克斌，陈富荣，等. 安徽池州地区富硒土壤地球化学特征及成因分析[J]. 资源环境与工程，2022，36(2)：154-162.

[6] 秦王武，邵树勋，夏勇，等. 水城茶园硒的地球化学特征及富硒茶开发探讨[J/OL]. 地球与环境：1-10 [2023-06-19]. DOI：10.14050/j. cnki. 1672-9250. 2023.051.006.

[7] 杨军，项剑桥，徐春燕，等. 湖北省土地质量地球化学调查进展与展望[J]. 资源环境与工程，2021，35(6)：818-826.

[8] 谭永杰，文敏. 地质信息化建设研究进展与展望[J]. 中国地质调查，2023，10(2)：1-9.

[9] 黄家凯，樊旭东，秦丽娟. 省级地质大数据建设的总体框架研究[J]. 资源环境与工程，2019，33(1)：103-108.

[10] 郭建忠，谢耕，成毅，等. 网格GIS与云GIS辨析[J]. 测绘科学技术学报，2014，31(2)：111-114.

[11] 徐长庆，李雨晴，李德友，等. 基于PostgreSQL与PostGIS的空间数据一体化方案设计与实现[J]. 山东理工大学学报(自然科学版)，2023，37(3)：1-7.

[12] 王子毅，张春海. 基于ECharts的数据可视化分析组件设计实现[J]. 微型机与应用，2016，35(14)：46-48.

The construction of selenium resource platform in Hubei province based on cloud GIS

Yang Shu[1], Hu Weilu[2,3*]

(1. Hubei Institute of Geoscience, Hubei Selenium Industrial Research Institute, Wuhan 430034, China;
2. WSGRI Engineering & Surveying Incorporation Limited, Wuhan 430080, China;
3. WSGRI Smart City(Wuhan) Engineering Technology Co., Ltd., Wuhan 430073, China)

Abstract: In order to improve the information level of selenium resource data in Hubei province and realize the application of selenium resource big data, a selenium resource information platform in Hubei province was constructed based on cloud GIS technology. The "One map" of selenium resources in Hubei province is established in the form of cloud GIS combining maps and statistical charts, which visually and dynamically displays the work degree of selenium resource investigation in Hubei province, sampling point information, selenium resource level in the survey area, selenium rich crops and other data statistical analysis through a large visual screen. At the same time, the whole platform serves as a management and application platform for selenium rich data in Hubei province. The construction results show that the selenium rich resource information platform in Hubei province has realized the dynamic visual display function of selenium resource information in Hubei province, and provided technical support for the management and application of selenium resource big data in Hubei province.

Keywords: selenium resource; cloud GIS; one map; dynamic visualization

(上接第99页)

Research on Web3D based intelligent analysis system for traditional Chinese medicine olfactory diagnosis

Ou Jinfeng, Li Zhenpeng, Liao Jiazhen, Huang Feixiong, Zhou Huaying*, Zhang Qi

(College of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, China)

Abstract: The complexity and instability of human body odor information makes the research on traditional Chinese medicine olfactory diagnosis lag far behind other facial and pulse diagnoses. Based on bionic olfactory technology and artificial neural network technology, this work establishes an intelligent analytical model for traditional Chinese medicine olfactory diagnosis. At the same time, Web3D visualization technology is applied to the intelligent analysis and research of traditional Chinese medicine olfactory diagnosis, which realized the complex analysis process of traditional Chinese medicine olfactory diagnosis with intuitive and vivid three-dimensional scene images. This study will contribute to promoting information fusion of the four diagnostic methods of traditional Chinese medicine and remote assisted diagnosis and treatment of traditional Chinese medicine.

Keywords: Web3D; TCM olfactory diagnosis; 3D visualization

文章编号: 1007-1423（2023）14-0105-05 DOI: 10.3969/j.issn.1007-1423.2023.14.021

基于西门子PLC的智能储运粮仓设计

严佳佳，张志萍，吕宵宵*

（扬州大学广陵学院，扬州 225000）

摘要：基于西门子300PLC可编程控制器设计智能储运粮仓控制系统，采用G130变频器实现对电机速度的控制，光电传感器检测粮食进/出仓信号，指示灯显示粮仓实时存储量，编写程序实现启停、进出料动作、满/空仓报警等基本功能控制，利用WinCC组态软件仿真程序调试。所设计的智能储运粮仓成本低、运输效率高，实现存储管理系统精细化、现代化和智能化。

关键词：智能储运粮仓；西门子300PLC；智能控制；传感器

0 引言

我国是一个农业大国，加快建设仓储管理智慧化的粮食仓储体系有助于筑牢大国粮仓，推进乡村振兴。传统的粮食储备仓库大部分是以人工为主的粗放型管理，采用扦样取粮，依赖人力，劳动强度大，已经不适用于目前的社会形势，设计具备自动化、工业化特点的粮食存储以及运输仓迫在眉睫[1-3]。

本文基于西门子300PLC设计了一款全自动控制系统的粮食存储仓库，只需较少的人工，节省人力资源[4-5]。通过硬件和软件程序的设计，提升存储空间的利用率；通过程序和机械设备的结构设计，提高粮食运输的效率。

1 存储粮食仓库系统的控制方案简介

本次设计的智能储运粮仓系统，主要由输送装置、运转装置以及分拣装置三部分组成。

所设计的智能储运粮仓系统输送装置采用了带传动。因为粮食粒在传输过程中容易发生抖动，为了防止在输送过程中滑落到地面，采用防滑性传送带。而对于粮食的传输来说，为了节约输送时间、提升效率，采用光滑性传送带。

运转装置包括了设备的电机、牵引绳、改向装置、驱动装置等，整个传送带的运动以及运转装置的动力是由功率为5.5 kW的电机将电能转换成动能。牵引绳使用电机的力量为传送带提供牵引力，带动履带进行运转；绕过每一个滚筒形成一个闭合圈状运转带动履带，实现粮食运输工作。

分拣装置主要是对粮食须和粮食粒进行分拣区别。上部是类似筛子的一个装置，通过不断抖动可以将粮食粒和粮食须初步分离开来；中部是一个具有一定区域的空间，能够借用风力将粮食粒和粮食须再次分离，分离出的粮食粒进入下一步继续运输，而粮食须则进入处理区域进行下一步的利用。

2 粮食仓库系统的硬件设计

2.1 硬件控制方案

现如今城市工业通信网络化，使用较多的

收稿日期：2023-04-10 **修稿日期：**2023-04-25

基金项目：教育部高教学生司供需对接就业育人项目（20230111230）；江苏省高等学校大学生创新创业训练计划项目（202213987003Y）；江苏省高校哲学社会科学一般项目（2020SJA2373）

作者简介：严佳佳（2000—），女，江苏盐城人，学士，研究方向为机电一体化；张志萍（1982—），女，江苏扬州人，博士，副教授，研究方向为机械制造及其自动化；*通信作者：吕宵宵（1989—），女，江苏扬州人，硕士，讲师，研究方向为机械制造及其自动化，E-mail：060080@yzu.edu.cn

是西门子PLC，通信模块化和编程工具的多样化技术位于其他品牌前列，控制器之间的通讯连接能力强，两个型号相同的可编程控制器经过网络的连接，即可完成设计系统的整个通信，在拓展模块和通信协议方面，也有着先进的技术。这些既可以扩展PLC在各个行业的应用范围，又提高了系统运转的稳定性和可靠性。因此，选用PLC作为智能储运粮仓系统的核心控制器。

硬件部分主要包括PLC控制器、电源模块、输入输出模块、通信模块、拓展模块。

PLC控制器：接收输入信号、处理逻辑信号、输出控制信号，整个粮仓系统的运行由其控制。

电源模块：为PLC控制器以及其他模块提供相对稳定的电源。

输入输出模块：将传感器中信号转换为数字信号输入到PLC中，将PLC输出的数字信号转换为控制信号输出到执行器中。

通信模块：PLC控制器与其他设备之间的通信连接。

拓展模块：用于扩展PLC控制器的输入输出能力、存储能力等。

2.2 变频器的选型

将电气控制柜中电动机直接集成到机器设备中，通过对G130变频器的参数设定和通讯模块连接，采用TCP-IP通讯方式，实现对电机速度控制的功能。通过设置变频器输出的电压频率的大小、电机启停的加减速度大小、限制电机的最大输出电流、过载保护时间和过载保护电流值等参数变量的值，实现电机的低、中、高三档速度。

2.3 硬件接线图

智能储运粮仓系统的硬件部分主要包括PLC控制器、输入输出模块、电动机、光电传感器、指示灯。接线图如图1所示。借由这些硬件，可以实现对粮仓内粮食存储量的实时监控，并将收到的数据通过通讯模块(如以太网、无线网络等通讯方式)发送给上位计算机或云端服务器模块。

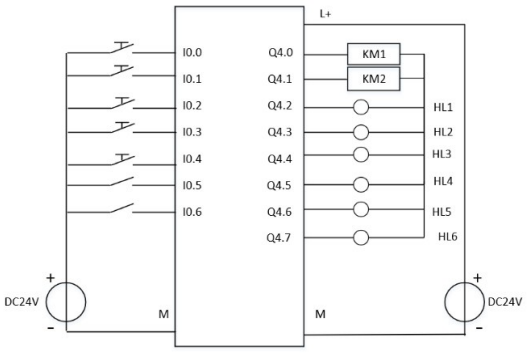


图1 智能储运粮仓系统的硬件接线图

2.4 智能储运粮仓系统的符号表

智能储运粮仓控制系统的控制使用四个按钮分别实现启动、停止、入仓、出仓动作，并通过指示灯显示工作状态。智能储运粮仓系统控制符号表如表1所示。具体描述功能实现原理：粮食经过传送带，对光源到传感器的光路

表1 智能储运粮仓系统控制符号

输入信号			输出信号		
名称	代号	输入编号	名称	代号	输入编号
启动按钮SB1	SB1	I0.0	电动机KM1	KM1	Q4.0
停止按钮SB2	SB2	I0.1	电动机KM2	KM2	Q4.1
启动按钮SB3	SB3	I0.2	粮仓空指示灯HL1	HL1	Q4.2
停止按钮SB4	SB4	I0.3	粮仓≥20%指示灯HL2	HL2	Q4.3
进料光电传感器PS1	PS1	I0.4	粮仓≥40%指示灯HL3	HL3	Q4.4
进料光电传感器PS2	PS2	I0.5	粮仓≥60%指示灯HL4	HL4	Q4.5
本地远程切换开关		I0.6	粮仓≥80%指示灯HL5	HL5	Q4.6
			粮仓满指示灯HL6	HL6	Q4.7

遮挡，形成脉冲信号。光电传感器PS1、PS2将检测到的信息转换成电信号输送到PLC的输入模块各输入端口中。检测到粮食进料时，则输出高电平信号，电机启动；检测到粮仓中粮食达到一定高度时，则输出低电平信号，电机停止；粮食存储量低于一定值时，会自动启动进料装置。PLC根据输入信号控制粮仓空指示灯：当PS1、PS2均未检测到粮食进料时，指示灯HL1亮起，粮仓为空；当PS1或PS2检测到粮食进料时，粮仓中的粮食存储量≥20%时，指示灯HL2亮起。HL3、HL4、HL5指示灯与其工作原理相同。粮仓中的粮食存储量满仓时，PS1、PS2同时输出低电平信号，PLC输出高电平信号，指示灯HL6亮。

2.5　输入输出地址表分配

通过对所设计智能储运粮仓系统进行分析，将提前设定的符号表用SIMATIC Manager软件进行符号表组态。智能粮仓系统的生产线符号表如图2所示。

S7 程序(1) (符号) -- liangcang\SIMATIC 300 站点\CPU314 C-2 DP(1)

	状态	符号	地址		数据类型		注释
1		Cycle Execution	OB	1	OB	1	
2		本地远程切换开关	I	0.6	BOOL		
3		电动机KM1	Q	4.0	BOOL		
4		电动机KM2	Q	4.1	BOOL		
5		进料光电传感器PS1	I	0.4	BOOL		
6		进料光电传感器PS2	I	0.5	BOOL		
7		粮仓≥20%指示灯HL2	Q	4.3	BOOL		
8		粮仓≥40%指示灯HL3	Q	4.4	BOOL		
9		粮仓≥60%指示灯HL4	Q	4.5	BOOL		
1		粮仓≥80%指示灯HL5	Q	4.6	BOOL		
1		粮仓空指示灯HL1	Q	4.2	BOOL		
1		粮仓满指示灯HL6	Q	4.7	BOOL		
1		启动按钮SB1	I	0.0	BOOL		
1		启动按钮SB3	I	0.2	BOOL		
1		停止按钮SB2	I	0.1	BOOL		
1		停止按钮SB4	I	0.3	BOOL		
1							

图2　智能粮仓系统的生产线符号表

3　粮食仓库系统PLC软件设计

3.1　粮仓系统的程序编写

粮食开始运入粮仓工作，传送带由按钮进行随时控制。按下SB1，电机M1接通自锁，传送带1动作，粮食运入粮仓；按下SB2，传送带1停止。系统自动检测粮仓的存储量，根据设定的指示灯显示出来。传送带的运行速度通过按钮进行控制，确保粮食的稳定运输。如图3所示，在粮仓系统中，入仓动作进行。同样，出仓工作原理类似，设计SB3、SB4两个按钮，分别控制传送带2开启和停止动作，实现将粮食运出粮仓。

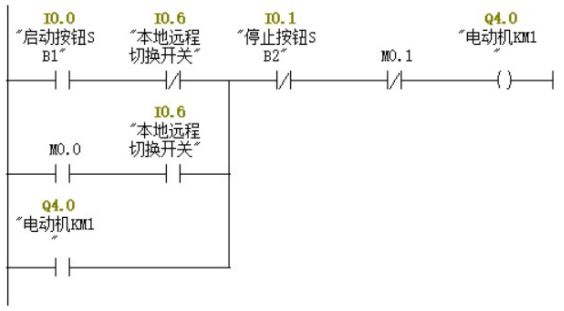


图3　智能储运粮仓系统控制入仓动作程序段

当启动粮食入仓动作时，安装在传送带两侧的进料光电传感器发出光束，通过检测光束是否被遮挡来判断粮食的数量，并且通过称重或测量体积的方式实时记录入仓粮食的数量，如图4所示。

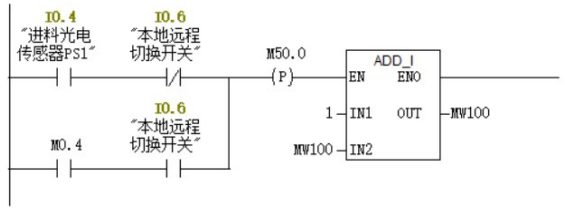


图4　智能储运粮仓系统光电传感器计数入仓数量程序段

粮仓系统将计数转换成百分数，通过指示灯来显示实时粮食存储量。粮仓总容量定为100%，粮食实时存储量转换为百分比。一旦粮仓内粮食存储量发生变化，系统会重新计算粮仓存储量百分比，以便指示灯动作。如图5所示程序段控制显示粮食储存量，并以百分比的形式显示。

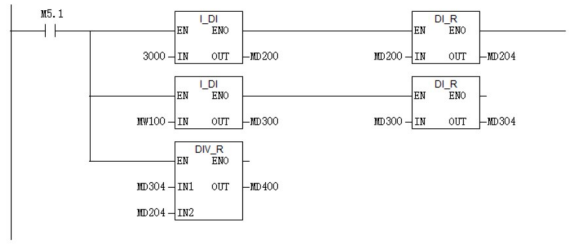


图5　智能储运粮仓系统显示粮食实时存储量程序段

3.2 基于西门子PLC粮食仓库系统的仿真调试

使用WICC软件，先创建系统项目，导入符号表，管理内部变量。打开图形编辑器，准备画图。最开始设置仓库内使用率为零，即仓库为空仓，并且仓库的第一个指示灯常亮为绿色，表示智能储运粮仓为空。如图6所示，粮仓系统光电感应器开始入仓动作。

图6 智能储运粮仓系统光电感应器开始入仓动作

3.3 仿真效果展示

粮食入仓工作状态，仿真效果如图7所示。按下启动按钮SB1，电机开始工作，传送带开始运行，安装在传送带两侧的进料光电传感器检测入仓数量，不同的指示灯亮起表示粮食入仓数量的变化情况。当粮食入仓达到粮仓的限制值后，“粮仓满”指示灯亮起，表示此时粮仓使用率为100 %，处于满仓状态，粮食入仓动作停止。

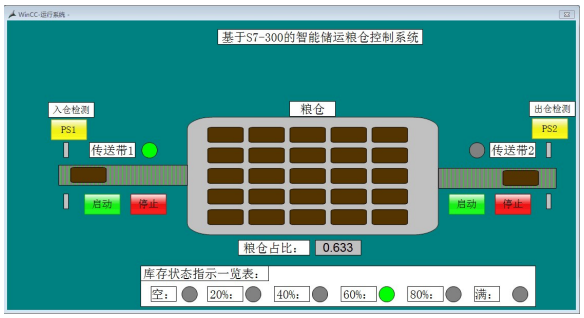


图7 智能储运粮仓系统工作状态仿真效果

按下SB3按钮，电机M2开始工作，出仓传送带开始运行，由指示灯指示粮仓内剩余的粮食数量。粮食出仓工作进行一段时间后，图7所示为粮仓内剩余的粮食数量为60 %的工作状态。当剩余粮食数量为空仓状态（即剩余数量为0时），“粮仓空”指示灯HL1再次亮起。光电传感器检测到粮仓空，出料完毕。按下停止按钮，电动机M2停止运转，粮仓停止出仓。

3.4 仿真模拟调试目的

通过WinCC软件仿真模拟真实的粮仓系统，验证控制程序和所需实现的功能是否符合要求，检测系统的稳定性能，对程序设计及时调试和排除故障，以便后续系统安全、稳定、高效地运行，满足储运粮仓的实际需求，为实际应用提供参考。

4 结语

本文阐述了基于西门子300PLC控制器设计的智能储运粮仓系统，通过硬件设计和软件设计，实现粮仓自动化和监测粮食储运过程。当粮食进出料时，光电传感器对粮食进行监测和数据采集，指示灯指示粮食存储量。西门子STEP 7软件编写PLC程序，WinCC软件仿真调试程序，实现智能储运粮仓自动控制和报警系统等功能。所设计的智能储运粮仓系统考虑粮食的储存容量以及粮食的运输等问题，实现粮食储运过程的智能化。

参考文献：

[1] 张立. 基于node-red的可视化粮仓温控系统设计[J]. 信息技术与信息化,2022(4):110-113.

[2] 刘东东,高绍和,梁明阳,等. 光纤测温传感器技术在粮仓中的应用[J]. 粮食科技与经济,2020,45(12):69-70.

[3] 王笑怡. 粮仓环境监测系统设计[J]. 微处理机,2020(3):39-44.

[4] 吴超. PLC技术在电气自动化控制中的应用[J]. 设备管理与维修,2021(20):119-121.

[5] 郭晓莹,何其文,谭雨祺. 基于S7-1200PLC的粮仓温湿度远程监控系统设计[J]. 工业仪表与自动化装置,2021(5):39-41.

（下转第120页）

文章编号: 1007-1423(2023)14-0109-04 DOI: 10.3969/j.issn.1007-1423.2023.14.022

NodeJs添加代码版权信息命令工具的设计与实现

张文豪*

（广东白云学院大数据与计算机学院，广州 510450）

摘要：为了给前端项目构建的代码文件添加版权信息，设计一个工具来自动完成，并可以提高工作效率。基于NodeJs包管理器NPM开发一个命令行工具，从package.json文件获取与作者相关的信息来生成版权信息，再扫描前端代码目标目录下的所有代码文件并自动将生成的版权信息添加到代码文件顶部。最后通过实际应用的结果表明，采用该方法设计的命令行工具简单易用，可以满足自动在代码文件顶部添加版权信息的需求。

关键词：NodeJs；NPM；NPX；GitHub；WorkFlow；版权

0 引言

随着Web技术的发展，软件项目中前端代码越来越多，前端技术在软件项目中的作用越来越重要，HTML、Javascript、Css等前端技术成为软件开发必备的技能[1]。2009年5月Ryan Dahl发布了NodeJs，是一个基于Chrome V8引擎的JavaScript运行环境，让JavaScript运行在服务端的开发平台，也使JavaScript成为了全球应用最广泛的开发语言[2]。2010年1月，NodeJs软件包管理工具NPM也正式上线，使得NodeJs在项目中的应用更加方便[3]。当前几乎所有前端项目都离不开NodeJs的相关技术，也形成了很多优秀的前端项目构建工具，如Vite、Webpack、Rollup等，并且也已经存在很多可以自动添加代码版权信息的插件，但是仍存在需要自己配置好版权信息才能完成的问题[4-6]。由于package.json文件已经包含了主要的版权信息，如软件名称、版本号、作者、开源协议等，每个NodeJs模块软件包又都必须包含一个package.json文件，可以通过package.json文件来直接获取版权信息[7]。

本文提出通过获取package.json文件里的相关版权信息来自动生成代码文件的版权信息，实现开箱即用的效果，无需任何配置就能自动添加代码的版权信息，同时考虑到前端项目构建工具众多，将该工具开发为一个NodeJs的命令行工具，能让该工具在所有前端项目中使用，可以有效提高该工具的通用性。

1 相关技术

NodeJs是Ryan Dahl于2009年发起的一个开源项目，是一个基于Chrome V8引擎，能够快速构建网络服务与应用的JavaScript执行平台，目前被广泛应用到各种Web应用的开发项目。

NPM的全称是NodeJs Package Manager，是一个NodeJs包管理和分发工具，已经成为非官方的发布NodeJs模块(包)的标准。

package.json是基于NPM标准的NodeJs模块(包)必须存在的描述文件，包括模块名称、版本号、作者、模块描述、关键字、软件许可证、运行脚本、软件依赖包等信息。

npx的全称是execute npm package binaries，是NPM5.2以后新增的命令工具，是一个NPM包的执行器，可以用于执行各种命令[8]。

GitHub于2008年正式上线，是一个面向开源及私有软件项目的托管平台[9]。

workflow文件是GitHub Actions的配置文件，

收稿日期：2023-04-03 **修稿日期：**2023-04-23

作者简介：*通信作者:张文豪(1983—),男,硕士,高工,研究方向为Web开发技术,E-mail:xqkeji@foxmail.com

存放在代码仓库的.github/workflows/目录下。比如，写一个npm_publish.yaml文件，存储的目录就是.github/workflows/publish.yaml。GitHub只要发现.github/workflows目录下里面有.yml文件，就会自动运行该文件[10]。

2 设计方案

首先根据package.json的信息自动生成代码版权信息，然后扫描项目构建的代码目录，并将版权信息添加到代码文件的顶部。

2.1 代码版权信息描述模板

```
/*!
 * ${pkg.name} v${pkg.version} (${pkg.homepage})
 * Author ${pkg.author}
 * LICENSE ${pkg.license}
 * Copyright ${year} ${pkg.author}
 */
```

pkg为解析package.json文件后的对象，name为模块名称/软件名称、version为版本号、homepage为官方主页、author为软件作者、license为软件许可证、year是根据当前时间获取的年份。

2.2 执行流程

通过NPM发布NodeJs模块通常都是指定dist目录(构建完成的代码目录)进行发布，可以默认扫描dist目录下所有代码文件，并添加代码版权信息就可以满足大部分的需求，同时也允许用户通过参数增加需要扫描的目录，来增加其通用性，如图1所示。

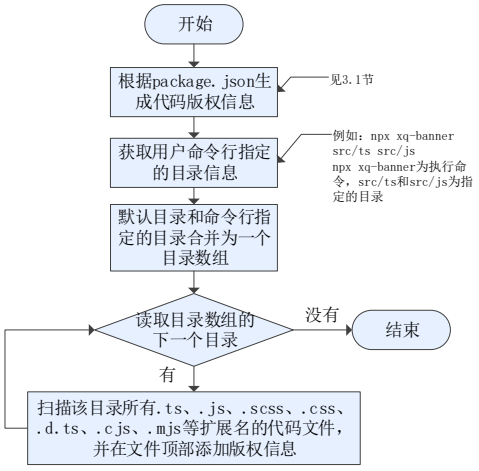


图1 具体流程

3 实现

3.1 初始化模块

打开命令行终端，使用mkdir新建模块目录xq-banner，使用cd命令进入模块目录，执行npm init初始化模块，按照提示完成相应信息的填写，系统自动完成package.json的创建，最后再根据模块实际情况调整后的package.json内容如下：

```
{
  "name": "xq-banner",
  "version": "1.0.0",
  "description": "NodeJs添加代码版权信息命令工具。",
  "type": "module",
  "files": ["bin"],
  "bin": {
      "xq-banner": "./bin/banner.js"
  },
  "repository": {
    "type": "git",
    "url": "https://github.com/xqkeji/xq-banner.git"
  },
  "keywords": ["HTML5","JavaScript","Nodejs","Banner"],
  "author": "xqkeji.cn",
  "license": "SSPL-1.0",
  "homepage": "https://xqkeji.cn/"
}
```

属性name要保证在npm模块包中是唯一的，才能发布成功。同时由于本文实现的是一个NodeJs的命令工具，必须指定bin属性的值，该属性由键值对组成，键表示命令，值表示对应的程序文件，例如{"xq-banner": "./bin/banner.js"}，表示该模块存在xq-banner命令，执行xq-banner命令时，就是执行'./bin/banner.js'的程序文件。

3.2 程序实现

(1) 确定package.json路径。用户安装本文实现的xq-banner模块时，会将程序安装在用户模块下的node_modules目录，同时xq-banner命令执行的是'./bin/banner.js'文件，banner.js文件在用户模块目录的路径为：'node_modules/

xq-banner/bin/banner.js’，因此用户模块的 package.json 路径相对于 banner.js 文件的路径为：‘../../../package.json’。

（2）生成代码版权信息。确定 package.json 路径后，使用 JSON 工具包解析 package.json 的文件，可以得到 package.json 的所有信息。同时使用日期工具，根据程序执行时间得到当前的年份。最后按照 2.1 节生成代码版权的信息。

（3）代码实现。代码实现可以到 github 平台查看：https://github.com/xqkeji/xq-banner。

3.3　发布

通常开发完成一个 NPM 模块后，可以使用 npm publish 命令来完成发布，但是一般的项目代码都是托管到 github 平台上，因此可以利用 github 的工作流来完成自动发布到 https://www.npmjs.com/服务器上。

（1）登录 npmjs 的账户，并创建一个 Access Token。

如图 2 所示，在 New Access Token 页面输入 Access Token 的 name 属性，然后选择用于 Publish，最后点击‘Generate Token’后就会生成一个 Access Token。

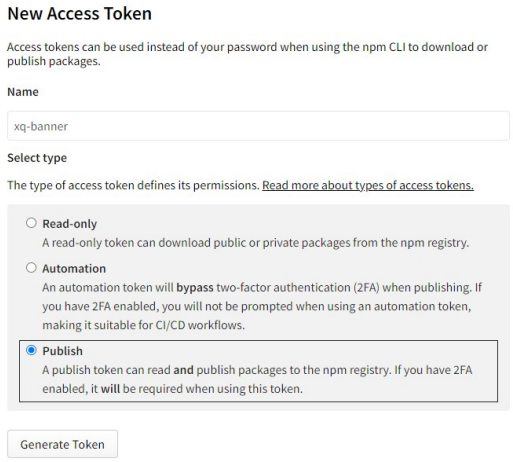


图 2　创建 Access Token

（2）在 xq-banner 的 github 代码库新建一个工作流程（workflow）。

如图 3 所示，模块发布工作流程的代码是相对固定的，只有${{secrets.npm_token}}是动态的，代表 npmjs 服务器上的 Access Token 值。

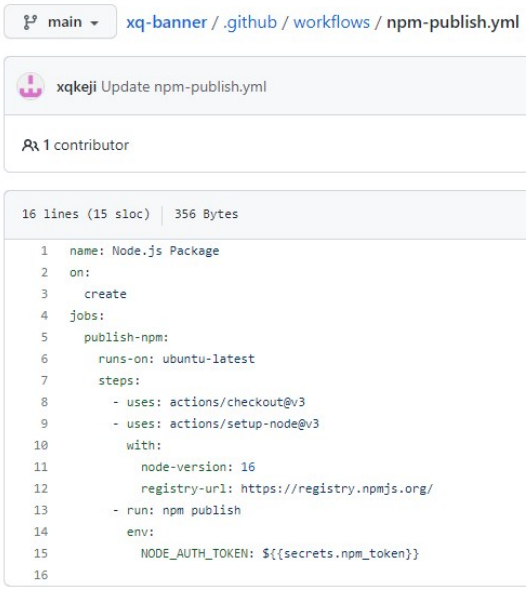


图 3　模块发布工作流程代码

（3）在 xq-banner 的 github 代码库新建密钥。

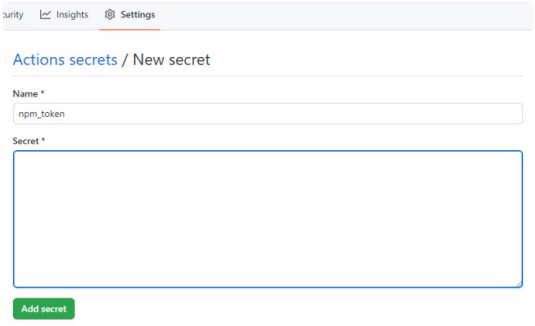


图 4　github 代码库新建密钥页面

Name 为密钥名称，与工作流程代码里的${{secrets.npm_token}}名称保持一致。Secret 填写 npmjs 服务器上生成的 Access Token 的值。

通过以上三个步骤后，只要在 github 代码库中新建一个版本标签，就会自动将 xq-banner 代码库的代码也同时发布到 npmjs 服务器上的 xq-banner 模块。

4　应用

（1）安装。

在需要自动添加代码版权信息的 npm 模块目录下，执行命令：npm i xq-banner，安装 xq-banner 模块。如图 5 所示。

```
PS F:\docker\npm\xq-confirm> npm i xq-banner

added 1 package, and audited 174 packages in 2s
```

图 5 安装xq-banner

（2）使用。

安装xq-banner模块后，可以直接执行npx xq-banner命令为dist目录下的所有代码添加版权信息，也可以将npx xq-banner命令串联到package.json的构建命令后面，构建完成后自动执行。package.json执行构建脚本示例：｛“build”：“npx unbuild && npm run convert && npm run min && npx xq-banner”｝。构建执行效果如图6所示。

```
PS F:\docker\npm\xq-confirm> npm run build

> xq-confirm@1.0.9 build
> npx unbuild && npm run convert && npm run min && npx xq-banner
```

图 6 执行package.json构建命令

执行后构建目录dist下的所有代码文件顶部都会自动添加如下的版权信息。

```
/*!
  * xq-confirm v1.0.9 (https://xqkeji. cn/demo/xq-confirm)
  * Author xqkeji.cn
  * LICENSE SSPL-1.0
  * Copyright 2023 xqkeji.cn
  */
```

5 结语

本文使用NPM模块的标准，开发了一个可以通过命令执行xq-banner模块，该模块默认实现了将dist目录下的所有代码文件顶部自动添加版权信息，也可以通过命令行参数指定其他目录来为代码自动添加版权信息。只要安装xq-banner模块，就能实现无需任何配置，开箱即用的效果，同时xq-banner是一个NodeJs命令行模块，适用于任何的前端项目，具有很强的通用性。

参考文献：

[1] 张文豪. Bootstrap模态对话框的开发与应用[J]. 福建电脑,2022,38(12):99-102.

[2] 周玉轩,杨絮,鲍富成,等. Cordova-NodeJS混合式物联网信息服务系统[J]. 计算机工程与应用,2019,55(6):209-217.

[3] 王伶俐,张传国. 基于NodeJS+Express框架的轻应用定制平台的设计与实现[J]. 计算机科学,2017,44(S2):596-599.

[4] 战仕霞. 基于React的信息管理系统前端自动化构建工具的设计与实现[D]. 北京:北京交通大学,2020.

[5] YOU E, CAPELETTO M, RAUCH G, et al. Getting started [EB/OL]. (2023-01-26) [2023-02-18]. https://vitejs.dev/guide/.

[6] TAEGERT ATKINSON L, BEDFORD G, KULAI A, et al. Introduction[EB/OL].(2023-01-21)[2023-04-01]. https://rollupjs.org/introduction/.

[7] THOMSON E, BORINS M, RIENSTRA M, et al. About npm [EB/OL]. (2022-08-28) [2023-02-18]. https://docs.npmjs.com/about-npm.

[8] MARCHÁN K, SCHLUETER I Z, CLARKE D, et al. Execute npm package binaries[EB/OL].(2021-03-10)[2023-02-18]. https://www.npmjs.com/package/npx.

[9] 汤佳杰,曹永忠,朱俊武,等. 基于混合神经网络的开源社区软件开发者人力资源价值预测[J]. 计算机应用与软件,2021,38(8):64-71,77.

[10] 张洋. 面向开源开发生态的群体贡献高效汇聚机理与方法研究[D]. 长沙:国防科技大学,2019.

Design and implementation of NodeJs add code copyright information command tool

Zhang Wenhao*

(Faculty of Big Data and Computing, Guangdong Baiyun University, Guangzhou 510450, China)

Abstract: In order to add copyright information to the code files built by the front-end project, designing a tool to complete it automatically and improve work efficiency. Developing a command line tool based on the NodeJs package manager NPM, obtain information related to the author from the package.json file to generate copyright information, and finally scan all code files in the target directory of the front-end code and automatically add the generated copyright information to the code file top. Finally, the results of practical application show that the command line tool designed by this method is simple and easy to use, and can meet the requirement of automatically adding copyright information at the top of the code file.

Keywords: NodeJs; NPM; NPX; GitHub; workflow; copyright

文章编号: 1007-1423（2023）14-0113-04 DOI: 10.3969/j.issn.1007-1423.2023.14.023

基于虚拟现实的油气井场安全培训系统设计

苑得鑫*，李嘉俊，李佳萍

（中国石油大学（华东）石油工业训练中心，青岛 266580）

摘要：针对传统安全培训存在的缺乏互动、形式枯燥、培训效果不佳等问题，将虚拟现实技术引入油气井场安全培训中，通过提炼油气井场典型的风险隐患，设计合理的交互情节，利用3dsMax建立模型，制作动态模型动画，以Unreal Engine 4作为开发引擎，构建交互体验场景，操作者可以通过HTC Vive设备与虚拟空间进行交互，开展油气井场事故体验与学习。实践表明，基于Unreal Engine和HTC Vive的油气井场安全培训系统对于提高学员的安全意识和安全技能具有良好的促进作用。

关键词：虚拟现实；Unreal Engine；油气井场；安全培训；3dsMax

0 引言

随着国民经济发展对油气资源需求的不断增长，油气企业进入快速发展阶段，然而，油气井场具有火灾、爆炸、高压物料刺漏、机械伤害、高处坠落、物体打击、触电等诸多风险，在生产运行和检维修过程中易出现安全事故[1]，因此，安全培训十分重要。传统安全培训的主要形式包括上课和观看视频，这种培训形式不仅缺少交流互动，枯燥乏味，更重要的是不易调动起学员的积极性，以致很难达到必要的培训深度，难以使学员对井场风险形成深刻认识。

随着虚拟现实（virtual reality，VR）技术的发展，利用其交互性、沉浸感和构想性三个最突出的特征，可以构建交互式三维动态视景和实体行为的系统仿真，为操作者提供视觉、听觉、触觉等感官体验，使操作者如同身临其境同三维空间内的事物进行交互[2-4]。目前，虚拟现实技术已越来越多地应用于矿山、医疗、电力、交通等领域的教育培训，对学员掌握操作技能、提升综合素质起到了巨大的辅助和促进作用[5]。

本文将虚拟现实技术与油气井场安全培训相结合，通过提炼油气井场典型的风险隐患，设计合理的体验场景和交互情节，利用3dsMax建立模型，并导入Unreal Engine 4引擎，利用HTC Vive作为输出设备，构建油气井场安全风险体验系统。实践表明，该系统对提高学员的安全意识和风险辨识能力，促进学员安全综合素质提升具有积极作用。

1 系统开发流程与功能

1.1 系统开发流程

油气井场安全培训系统既应符合一般软件工程项目的开发流程，又应注重油气井场安全工程背景，其整体开发流程如图1所示，主要分为三个阶段。第一阶段为总体策划，应结合油气井场风险辨识相关标准及典型事故案例，对交互场景和交互情节进行总体设计；第二阶段为模型库建立，根据设备CAD图纸及现场照片、视频等资料，利用3dsMax进行单体设备的三维建模，设计相关场景的平面图，将模型文件以

收稿日期：2023-03-29 **修稿日期：**2023-06-19

基金项目：中国石油大学（华东）大学生创新训练项目（202110032）

作者简介：*通信作者：苑得鑫（1990—），男，黑龙江鸡西人，硕士，讲师，研究方向为工程训练与创新创业教育，E-mail：540867565@qq.com；李嘉俊（2001—），男，广西柳州人，本科生，研究方向为计算机技术、虚拟现实技术；李佳萍（2002—），女，河北沧州人，本科生，研究方向为机电控制技术、虚拟现实技术

FBX格式导出；第三阶段是交互开发，将FBX格式的模型文件导入Unreal Engine 4中，进行贴图及灯光的处理，结合井场布局图进行场景构建和碰撞检测，通过蓝图可视化编程，根据设计好的交互情节，实现相应的交互功能，最终完成油气井场安全培训系统的开发。

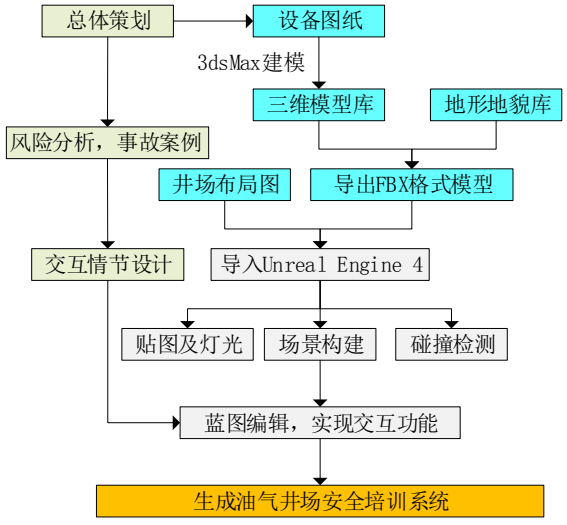


图1 系统开发流程

1.2 系统功能设计

油气井场安全培训系统旨在提高进入井场作业人员的安全意识、风险辨识能力以及事故预防与处置能力，重在直观体验。综合考虑事故发生发展规律及学习认知规律，设计了系统总体功能，如图2所示。

在充分辨识油气井场作业风险的基础上，结合GB 6441对于事故类型的划分，设定了皮带夹手、高处作业、阀门刺漏、人员触电、旋转挤入、光杆挤手、火灾及爆炸等八项体验内容，结合实际事故案例，设计合理的交互情节，引导操作者“犯下一个个小错误”，最终导致事故的发生。操作者通过HTC Vive设备第一视角直观体验事故带来的强烈感官刺激和严重后果，随后站在第三视角上观看事故回放，对事故发生发展的全过程进行复盘，对事故产生的原因形成直观认知。在事故体验的基础上，通过图文、语音、视频等形式，多方式提示事故预防措施及应急处置方法，强化知识吸收。此外，每个体验内容均包含了即时测试题目，学员考核合格后，方可解锁下一体验场景，否则需要重新体验。

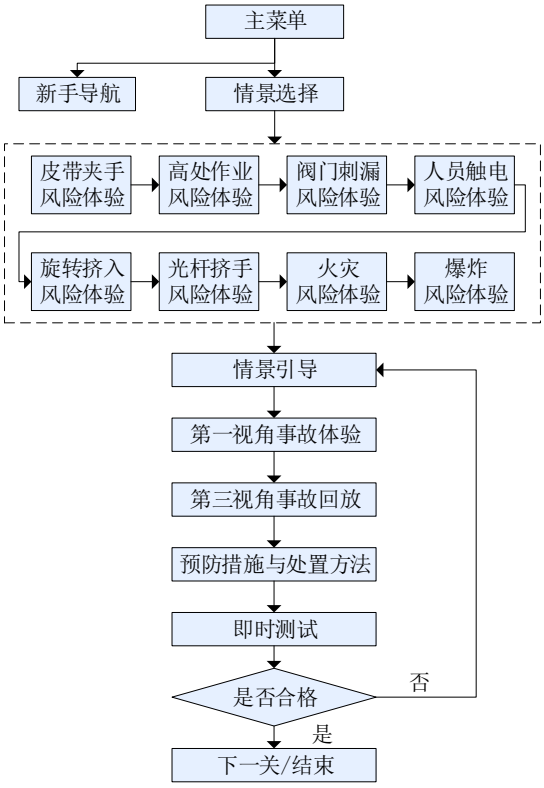


图2 系统功能框架

2 关键技术

2.1 基于3dsMax的抽油机模型动画建立

抽油机是油气井场中的关键设备，其零部件众多，根据CAD图纸，利用3dsMax作为建模工具进行建模。由于抽油机是动态运行的，为真实还原生产场景，应建立抽油机连续运动模型动画，并带入到Unreal Engine 4中。

通过分析抽油机运动特点，在皮带轮轴、曲柄销、中轴承等五处添加骨骼，其位置和作用见表1。为建立各骨骼之间的约束关系，按照曲柄、连杆、游梁、驴头的顺序，依次连环设置IK解算链。通过骨骼蒙皮技术建立骨骼与各运动模型间的物理和几何对应关系，并对各模型运动受到不同骨骼动作影响的权重进行精确设定[6]。

表 1　骨骼绑定位置及作用

序号	位置	作用
1	皮带轮轴	带动减速箱输入轴转动
2	减速箱输出轴	带动曲柄转动
3	曲柄销处	带动连杆运动
4	尾轴承	带动游梁上下摆动
5	中轴承	带动驴头上下摆动

根据抽油机运动特点，确定驴头上死点、驴头下死点、曲柄上死点、曲柄下死点、游梁水平位置为关键位置，即关键帧。在TimeLine上，分别对各关键帧下所有骨骼关键点的角度及位置进行调节，在曲线编辑器中，对各关键帧之间位置变化的加速度进行精细Bezier Curve校对，并进行逐帧检查，以保证抽油机动作连续平稳。根据实际抽油机的冲次范围，设定模型动画单次循环为120帧，若超出120帧范围，将模型动画的循环规律设定为“循环”模式，最终得到连续运动的抽油机模型动画。

2.2　基于UE4的交互场景构建

Unreal Engine 4(UE4)又称虚幻引擎，该引擎具有强大的蓝图可视化编程功能，可与C++编程结合互补使用，同时，其系统开源，利用编译版本可以十分方便地接入其他插件，十分便于开发；其光照渲染、物理引擎、材质编辑器等功能和效果基本处于业内顶尖水平，同时对VR手柄和VR控制器等虚拟现实设备提供了良好的支持。

将FBX格式的三维模型导入UE4后，在UE4引擎中完成环境创建及场景渲染，UE4材质系统能够创建出极为真实的材质，利用真实拍摄的抽油机照片贴图以及结合运用MindTex2软件生成的法线、高光等贴图，完成真实的、表面具有凹凸感的材质。导入UE4中的设备模型已经在3dsMax中完成了UV展开处理，能够实现为同一个物体的不同表面赋予不同的材质。灯光渲染能够让整体的场景更加真实，主要分为直接光照和全局光照。油气井场为野外环境，其直接光照主要为太阳光，在场景界面或蓝图界面均可设置太阳光的角度和位置；全局光照为环境光，利用Lightmass来存储全局光照，可使环境光更具真实感[7]。利用UE4渲染的油气井场效果如图3所示。

图 3　油气井场渲染效果展示

2.3　基于UE4的事故伤害特效设计

在安全体验系统中，事故伤害特效的展示尤为重要，直接决定了操作者受到感官震撼和冲击的程度。基于UE4强大的动态交互、实时渲染能力，通过与实际伤害现象进行对比，设计了符合各项事故体验内容的事故伤害特效，充分调动学员的视觉、听觉、触觉感知，使体验更加真实。

为展示事故特效，需进行多项参数的设定。以人物倒地效果为例，应进行如下设置：

(1) 物理倒下，即空间位置改变；

(2) 镜头旋转设置，模拟人物滚动效果；

(3) 尖叫声触发，模拟听觉感受；

(4) 屏幕材质贴图更换为红色血迹，模拟流血效果；

(5) 禁止手柄指令，保证触发后手柄无法继续操作，以便充分展示事故效果；

(6) 屏幕晃动，模拟人员站不稳的效果。

各项参数经过反复优化与测试，最终得到最具真实感的事故伤害体验效果。

3 结语

将虚拟现实技术的交互性、沉浸感和构想性应用于教育培训领域，可以提高学员学习的积极性，增加学习的趣味性，易于通过切身感受引发深入思考。本文以油气井场安全体验为主题，通过深挖工程内涵，利用3dsMax建立模型和动画，基于Unreal Engine 4引擎构建交互场景，实现交互功能，利用HTC Vive作为输出设备，建立了油气井场安全培训系统，能够使学员完全沉浸地、自主地、交互地进行安全体验和学习。实践表明，基于该系统开展的安全培训，更加有助于提高学员的安全意识和安全技能，培训效果优于传统培训方式。

参考文献：

[1] 马立权. 采油作业现场安全风险的预防与控制[J]. 化工设计通讯,2019,45(4):62.

[2] 黄仁东,吴同刚. 非煤矿山虚拟现实安全培训系统的研究与构建[J]. 中国安全生产科学技术,2017,13(8):36-41.

[3] 闫兴亚,王馨梅,魏梦婕. 基于虚拟现实的丝绸之路交互系统的设计与开发[J]. 计算机与数字工程,2020,48(4):838-842.

[4] 肖建良,张程,李阳. 基于Unity3D的室内漫游系统[J]. 电子设计工程,2016,24(19):54-56.

[5] 高俊涛,李林林. 基于虚拟现实的实验系统的设计及应用研究[J]. 计算机与数字工程,2018,46(8):1690-1696.

[6] 唐学军. 基于VR技术的多视觉动画角色3D模型设计与实现[J]. 现代电子技术,2020,43(16):142-145.

[7] 肖巍,冯时,王选遥. 基于虚幻4引擎的长白山虚拟现实场景地形制作[J]. 长春理工大学学报(自然科学版),2019,42(6):133-137.

Design of oil and gas well site safety training system based on virtual reality

Yuan Dexin*, Li Jiajun, Li Jiaping

(Petroleum Industry Training Center, China University of Petroleum, Qingdao 266580, China)

Abstract: Aiming at the problems of traditional safety training such as lack of interaction, boring format, and poor training effectiveness, virtual reality technology is introduced into oil and gas well site safety training, by refining typical risk hazards in oil and gas well sites, designing reasonable interaction scenarios, using 3dsMax to establish models, producing dynamic model animations, and using Unreal Engine 4 as a development engine to build interactive experience scenarios. Operators can interact with virtual spaces through HTC Vive devices to experience and learn oil and gas well site accidents. The practice has shown that the oil and gas well site safety training system based on Unreal Engine and HTC Vive promotes students' safety awareness and skills.

Keywords: virtual reality; Unreal Engine; oil and gas well site; safety training; 3dsMax

文章编号: 1007-1423（2023）14-0117-04 DOI: 10.3969/j.issn.1007-1423.2023.14.024

基于STM32单片机设计的睡眠检测装置

康楚阳[1*]，邵乙飞[1]，陈明蒂[2]，张景越[2]，冉志坚[2]

（1. 上海工程技术大学机械与汽车工程学院，上海 201620；2. 上海工程技术大学材料学院，上海 201620）

摘要：随着科技的进步与技术的发展，我们能够通过一些简单的装置和设备来监测身体状态，实时关注身体健康。针对睡眠时可能出现的打鼾情况提出了一种基于STM32单片机的智能止鼾装置。通过单片机驱动系统，对使用者的睡眠情况进行实时监测，通过振动模块的开闭对使用者进行一定程度的止鼾，来让使用者有一个更好的睡眠体验。

关键词：STM32；串口通信；睡眠检测

0 引言

随着科技的不断发展与社会的不断进步，健康已经成为人们生活中越来越重视的话题。比如近些年来科技厂商们都在其穿戴设备上添加了关于健康检测的模块。以 Apple Watch 为例，近年来 Apple Watch 的关注度越来越高，其过半的市占比就可以看出用户的肯定与追捧。第一代 Apple Watch 以“高端、奢华”为卖点，不过市场反应不积极、销量一般。其后的第二代、第三代产品逐渐转向以“健康、运动”为卖点，将睡眠检测、运动数据记录以及心率加入其互动系统中，而第四代产品更是将血氧检测加入其中，并为女性用户专门开发了经期检测系统，更加关注呵护用户的健康。可见围绕“科技与健康”这一主题，有相当大的空间进行实践与创新。

以此为起点，本项目组通过讨论并阅读相关文章，发现“睡眠问题”也是一个困扰很多人的问题，而打鼾是讨论最多的话题。通过前期制作相关问卷，统计总计200份的问卷调查报告，本项目组认为“打鼾”在某种程度上已经成为现在寝室的一个难题。基于网络上已经出现的相关文献，本项目组在现有算法的基础上进行改进，基于STM32单片机设计了一款睡眠检测装置。

1 设计方案

打鼾是指睡眠中因上呼吸道狭窄使悬雍垂发生振动而发出的鼾声，在日常生活中非常常见。打鼾的原因有很多，上了年纪的人几乎都有打鼾的症状，同时饮酒、吸烟、不恰当的睡眠姿势以及睡眠障碍也会成为打鼾的原因之一。打鼾带来的危害有很多，对于年轻人，可能会出现白天嗜睡、乏力、注意力不集中、头痛、工作能力下降等症状，对于老年人常伴有睡眠呼吸暂停综合征的出现，因此对个人健康有一定的不良影响。资料显示打鼾可以通过采取一

收稿日期：2023-03-22 **修稿日期：**2023-07-05

基金项目：2022—2023年度大学生创新项目（CX2224003）

作者简介：*通信作者:康楚阳(2001—),男,四川成都人,本科生,研究方向为汽车服务工程,E-mail:1263196639@qq.com;邵乙飞(2002—),男,河南新乡人,本科生,研究方向为智能制造工程;陈明蒂(2002—),男,广东湛江人,本科生,研究方向为材料成型及控制工程;张景越(2002—),男,黑龙江哈尔滨人,本科生,研究方向为材料成型及控制工程;冉志坚(2002—),男,贵州铜仁人,本科生,研究方向为材料成型及控制工程

些措施来缓解，比如在打鼾时由他人或外在装置轻推打鼾者，使其改变睡觉时的体位，可使打鼾得到缓解。

本文基于STM32F103C8T6这款单片机设计了可以监测穿戴者睡觉时所发出的声音分贝并做出一定振动输出的装置。其中STM32F103C8T6单片机负责将音频检测模块采集的数据处理后数字化，再用该数据与预先设置好的数字范围做对比，如果时间长达30秒及以上，则通过振动马达的输出振动穿戴者的手臂使其翻身，从而达到止鼾的目的。

2 止鼾装置模块介绍

止鼾装置模块如图1所示。

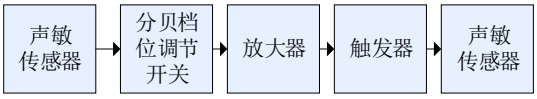


图1 装置模块

2.1 STM32单片机

STM32单片机是ST公司基于ARM Cortex-M内核开发的32位微控制器，STM32单片机常应用于嵌入式领域，如智能车、无人机、机器人、无线通信、物联网、工业控制等产品。STM32单片机主要包括1个ARM Cortex-M内核、储存器和外设。通过接口与不同的外设进行连接，实现和满足不同的需求与开发。

2.2 主控制器

考虑到成本以及装置的运行系统的稳定，项目组采用STM32F103C8T6控制芯片，一款增强型的单片机。内核为ARM Cortex-M3，最高工作频率为72 Mhz，1.25 MIPS/MHz。片上集成32～512 KB的Flash存储器。6～64 KB的SRAM存储器。2.0～3.6 V的电源供电和I/O接口的驱动电压。上电复位(POR)、掉电复位(PDR)和可编程的电压探测器(PVD)。拥有48个引脚，属于中容量产品。通过GPIO输入协议，将音频检测模块与STM32单片机进行连接。接受音频检测模块并将其数字化，判断其是否在实现所设定的范围内，判定结果后，通过GPIO输出协议，输出脉冲频率。主控制器所实现的原理即通过应用统一的智能化平台，实现各个独立子系统的有机连接，最终形成一个能及时进行信息交换和管控的网络。

2.3 音频检测模块

该功能模块指的是声音传感器，包括1个开关指示灯、电位器、信号输出、电源负极、电源正极(3.3～5 V)和一个电源LED指示灯。音频检测模块的作用是对环境声音强度敏感，作为检测周围声音强度的装置。模块在环境声音强度达不到设定阈值时，OUT输出高电平，当外界环境声音强度超过设定阈值时，模块OUT输出低电平。通过小板数字量输出OUT与STM32F103C8T6直接相连，通过单片机来检测高低电平，由此来检测环境声音。

2.4 输出模块

该功能模块指的是振动传感器，包括1个灵敏度调节电位器、开关信号输出、电源正极3.3～5 V、电源负极、3.1 mm可锁3.0螺丝。其中需注意的是开关信号指示灯亮时输出低电平、不亮则输出高电平、信号输出的电平接近电源压。振动模块在不振动时，振动开关呈断开状态，输出端输出高电平，绿色指示灯不亮；振动时，振动开关瞬间导通，输出端输出低电平，绿色指示灯亮。因为振动传感器输出端可以与单片机直接相连，通过单片机来检测高低电平，由此来检测环境是否有振动，起到报警作用。

2.5 OLED屏幕显示模块

该功能是由一块0.96寸OLED模块组成，OLED是性能优异的新型显示屏，具有功耗低、响应速度快、宽视角、轻薄柔韧等特点。选用0.96寸是考虑到成本因素，该显示屏简单易用、占用接口少。供电为3～5.5 V，通讯协议：I2C/SPI，分辨率：128×64。对于部分单片机编程而言，利用OLED屏幕可以使程序内部数据通过代码的编写来显示在OLED屏幕上，方便查找错误。同时在后续设备调试的时候使设备更加美观，使用更加方便。

3　软件设计

软件设计如图2所示。

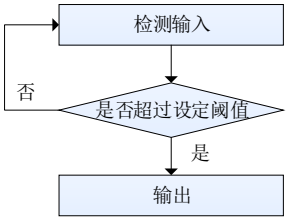


图2　设计逻辑

3.1　GPIO原理

GPIO也称为通用输入/输出，是控制器中最简单也是最重要的配置。一个程序接收到指令然后做出反应，接受指令就是一个输入的过程，做出反应就是一个输出的过程。GIPO就是这样一个输入和输出的总称。STM32F103C8T6的GPIO引脚电平为0～3 V，部分引脚可容忍5 V。输出模式下可控制端口输出高低电平，用以驱动LED、蜂鸣器、模拟通信协议输出时序等，输入模式下可读取端口的高低电平或电压，用于读取按键输入、外接模块电平信号输入、ADC电压采集、模拟通信协议接收数据等。GPIO配置8种输入输出模式。GPIO输入包括输入浮空、输入上拉、输入下拉和模拟输入四种输入模式。GPIO输出包括开漏输出、开漏复用功能、推挽式输出和推挽式复用功能四种输出模式。

3.2　中断原理

3.2.1　中断

处理器中的中断指的是在处理器中，中断是一个过程，即CPU在正常执行程序的过程中，遇到外部/内部的紧急事件需要处理，暂时中止当前程序的执行，转而去处理紧急的事件，待处理完毕后再返回被打断的程序处继续往下执行。中断在计算机多任务处理，尤其是即时系统中尤为重要。比如uCOS，FreeRTOS等。中断能提高CPU的效率，同时能对突发事件做出实时处理。实现程序的并行化，实现嵌入式系统进程之间的切换。

3.2.2　TIM定时器

TIM（Timer）定时器可以对输入的时钟进行计数，并在计数值达到设定值时触发中断。STM32的定时器拥有16位计数器、预分频器、自动重装寄存器的时基单元，在72 MHz计数时钟下可以实现最大59.65（$72/2^{16}/2^{16}$）的定时。定时器不仅具备基本的定时中断功能，而且还包含内外时钟源选择、输入捕获、输出比较、编码器接口、主从触发模式等多种功能。根据程序复杂和应用场景的不同定时器可分为高级定时器、通用定时器、基本定时器三种类型。

3.3　ADC

ADC全称为模拟-数字转换器。ADC可以将引脚上连续变化的模拟电压转换为内存中存储的数字变量，建立模拟电路到数字电路的桥梁。本次使用的STM32F103C8T6的ADC是12位逐次逼近的ADC，转换时间1 us。输入电压范围0～3.3 V，转换结果范围0～4095.18个输入通道，可测量16个外部和2个内部信号源，即温度传感器和内部参考电压。

本研究希望实现如下几种效果：首先是通过STM32单片机检测大概的声音强度并在OLED上显示出来。利用ADC原理对声音分贝的显示，通过ADC转化从而获得分贝数据并显示在OLED屏幕上方便观察。其次本研究不希望程序是听到声音立刻产生触发，考虑到鼾声是持续的、连续的。预期的程序是检测一段长达10秒及以上的声音之后才触发振动模块。通过代码的编写给予STM32单片机一个判断的数值范围，即分贝过小和分贝过大并不会触发单片机。该功能是防止误触发。比如分贝过小的连续性声音其实并不一定是鼾声，如果不划定一定的范围，那么就可能造成误触，降低使用者的使用体验。同样，分贝过大的声音足以令使用者醒来，因此也没有触发的必要。关于如何判断触发的因素是“一段长达10秒及以上的声音”，本研究采用定时器和中断的原理来完成这一设置。通过音频监测来让STM32单片机接收口电位以1秒的时间产生变化，定时器则可以记录这一电位变化，数据大于10则说明这一段声音是长达10秒的，即可以启动触发，开启振动模块。

4　结语

本研究针对寝室存在的打鼾问题设计了一款以STM32微处理器为核心，通过接收音频检

测器输入数字信号，再通过程序分析来达到一定的睡眠检测功能的装置，该装置具有一定的止鼾功能，控制系统也具有体积小、易于集成和调试的特点，方便后续的修改以及合成。

A sleep monitoring device based on STM32

Kang Chuyang[1*], Shao Yifei[1], Chen Mingdi[2], Zhang Jingyue[2], Ran Zhijian[2]

(1. School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;
2. School of Materials Science and Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: With the progress of technology and the development of techniques, we are able to detect our body status and monitor our health in real time through some simple devices and equipment. The article proposes an intelligent anti-snoring device based on STM32 for snoring during sleep. By driving the system with STM32, the user's sleep condition can be monitored in real time. The vibration module is turned on and off to stop snoring to some extent, providing users with a better sleep experience.

Keywords: STM32; serial communication; sleep monitoring

(上接第108页)

The design of intelligent storage and transportation granary based on Siemens PLC

Yan Jiajia, Zhang Zhiping, Lyu Xiaoxiao[*]

(Guangling College of Yangzhou University, Yangzhou 225000, China)

Abstract: The intelligent storage and transportation granary control system was designed based on Siemens 300PLC programmable controller. The control of motor speed for the control system by using a G130 frequency converter. The system detects the signal of grain entering and exiting warehouse through photoelectric sensors, and displays the real-time storage capacity of the warehouse through indicator lights. The basic functions of the warehousing system, such as start and stop, feeding and discharging action, full and empty warehouse alarm, was realized by writing control programs, which was debugged by WinCC configuration software simulation program. The designed intelligent storage and transportation granary realizes the refined, modern and intelligent storage management system, showing the advantages of low cost and high transportation efficiency.

Keywords: intelligent storage and transportation granary; Siemens 300PLC; intelligent control; sensor