

M 现代计算机

XIANDAI JISUANJI

第30卷第7期（总第799期）

半月刊（1984年创刊）

2024年4月10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 广州市友盛彩印有限公司
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



M

现代计算机

第30卷 第7期（总第799期）

2024年4月

2024年4月 第30卷

第7期（总第799期）

M

ISSN 1007-1423
CN 44-1415/TP

MODERN COMPUTER

现代计算机



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

版权声明

1. 本刊版权属于杂志社所有，其他报刊或网站如需转载，须经本刊同意，注明转载自本刊并付作者稿酬。
2. 本刊来稿恕不退还，请自留底稿。请勿一稿多投。来稿文责自负，严禁抄袭。对侵犯他人版权或其他权利的稿件，本刊概不承担连带责任。
3. 对所投稿件，本刊编辑有权根据刊物的需要进行删改或调整。
4. 凡是刊登在本刊的稿件，即表示作者同意稿件在《现代计算机》网站、中国期刊数据库CNKI、中国学术期刊（光盘版）、中文科技期刊数据库、中国科技期刊（遴选）数据库、中国学术期刊综合评价数据库等媒体发布。

目次

研究与开发

- 基于改进YOLOv5的水下目标检测模型 张征鑫, 张笃振 (1)
- ECPANet: 一种基于注意力的深度卷积神经网络通道剪枝方法
..... 余显冰, 杨礼友, 李 健 (9)
- 一种针对低质量指纹证据图像的增强算法设计 卢超超 (17)
- 基于遗传算法和SMOTE的网络入侵检测模型优化研究 戴周浩 (24)
- 基于神经网络的民航客运量的预测研究 唐甜甜, 张佳明, 姜 为, 王 海 (31)
- 改进DeepLabV3+的复杂场景倒塌建筑物分割算法 钟联升, 陶青川 (38)
- 基于注意力机制深度学习的轻量级无线网络数据分选方法 何靓华, 赵 英 (44)
- 基于改进的CNN-BiLSTM和三支决策的网络入侵检测方法 王 宇 (49)
- 基于NEDI插值的图像分辨率提升算法 刘文字, 郝慧艳, 郭 楠, 张志伟 (54)
- 基于深度学习的小目标检测综述 王聪敏, 李秀丽 (59)

实践与经验

- 基于边缘缓存的蜂窝网络的资源分配 潘 雨, 华 伟 (64)
- 融合LDA-LSTM算法的微博档案关注度和情感分析 孙思怡, 王家强, 罗子江 (70)
- 软件设计模式下装饰模型的研究与应用 向晓婷 (76)
- 计算模型在化合物毒性预测方面的应用及展望 许志旺, 何王秋, 孔 韧 (81)
- 基于粒子群优化随机森林算法的河南省葡萄价格预测研究
..... 王 哲, 张文慧, 付金鹏, 杨进进 (86)

开发案例

- 教师职教能力评估管理系统的设计与应用 李进豪 (90)
- 智能终端APP适老化设计研究 崔晓旭, 徐文燕 (96)
- 非遗印记APP的UI分析与设计 吴 冰, 张晶晶, 程 妮 (101)
- 基于FreeSWITCH的电话约车系统方案与应用 安文政, 张 文 (105)
- 基于“互联网+”的智能车位管理系统的研究与设计
..... 石亮翻, 朱贵平, 张凯淇, 毛 湘, 张 燕 (109)
- 基于在线句子分析的系统设计与实现 汪倩倩 (113)
- 基于SpringBoot的网上商城管理系统设计与实现 王培培 (117)

Modern Computer

(Vol. 30, No. 7; Apr. 10, 2024)

CONTENTS

Research and Development

Underwater object detection model based on improved YOLOv5	(1)
ECPANet: An attention-based method for channel pruning of deep convolutional neural networks	(9)
Design of an enhancement algorithm for low quality fingerprint evidence images	(17)
A study on improving the network intrusion detection model using genetic algorithm and SMOTE	(24)
Research on predicting civil aviation passenger transport volume based on neural networks	(31)
Complex scene segmentation of collapsed buildings algorithm based on improved DeepLabV3+	(38)
Data sorting method of lightweight wireless network based on deep learning of attention mechanism	(44)
Network intrusion detection based on improved CNN-BiLSTM and three-way decision	(49)
Image resolution enhancement algorithm based on NEDI interpolation	(54)
A review of deep learning based on small target detection	(59)

Practice and Experience

Resource allocation in cellular networks based on edge caching	(64)
Analysis of emotion and attention of Weibo archives based on LDA-LSTM algorithm	(70)
Research and application of decorative models in software design patterns	(76)
Application and prospects of computational models in predicting compound toxicity	(81)
Research on grape price forecasting of Henan Province based on particle swarm optimization random forest algorithm	(86)

Development Solution

Design and application of teacher vocational education ability evaluation management system	(90)
Research on age-friendly design of intelligent terminal APP	(96)
UI analysis and design of APP on “Impression of Intangible Cultural Heritage”	(101)
Telephone taxi booking system based on FreeSWITCH and its application	(105)
Research and design of intelligent parking space management system based on “Internet +”	(109)
Design and implementation of a system based on online sentence analysis	(113)
Design and implementation of an online mall management system based on SpringBoot	(117)

研究与开发

文章编号: 1007-1423(2024)07-0001-09

DOI: 10.3969/j.issn.1007-1423.2024.07.001

基于改进 YOLOv5 的水下目标检测模型

张征鑫, 张笃振*

(江苏师范大学计算机科学与技术学院, 徐州 221116)

摘要: 在复杂的深海环境中, 精准地检测和识别目标是海洋探索和保护的关键。此研究对 YOLOv5 目标检测算法进行了优化, 以提升水下目标的检测效果和效率。为了增强模型的特征提取能力, 对 YOLOv5 的关键卷积结构——C3 模块进行了替换。此外, 引入 SEAttention 模块, 这种模块利用空间注意力机制, 使模型更专注于重要目标的识别。评估时, 采用 EIoU 作为标准以准确地衡量目标检测的表现。实验主要在 DeepTrash 数据集上进行, 与原始 YOLOv5 相比, 优化后的模型在精确度上提高了 3.1 个百分点, 召回率提高了 1.8 个百分点, mAP@0.5 提高了 2.4 个百分点。而且, 优化后的模型参数量也减少了 17.4 个百分点, 从而提升了计算效率。即使与最新的 YOLOv8 相比, 优化后的模型也展现出了优越的性能。

关键词: 水下目标; 目标检测; 注意力机制; YOLOv5; 深度学习

0 引言

海洋是人类无尽的宝藏, 但随着工业化和城市化的加速, 海洋垃圾污染日益严重。化学废物、金属和塑料等污染物的排放破坏了生态平衡, 对渔业、旅游业和全球气候造成了不良影响。为了解决这一问题, 水下垃圾检测技术应运而生。

在目标检测领域有诸多经典算法, 例如, R-CNN^[1]采用选择性搜索生成候选区域, 再利用卷积神经网络提取特征。但其速度较慢。Fast R-CNN^[2]和 Faster R-CNN^[3]分别进一步优化了 R-CNN, 提高了速度和准确性。SSD^[4]和 YOLO 系列则是基于单阶段的目标检测算法, 均能有效地检测多尺度目标。

近些年, 随着海洋保护意识的增强, 许多研究者开始运用 YOLO 算法进行水下目标检测。我们选择了基于 YOLOv5^[5]进行研究而不是最新的 YOLO 模型, 主要是因为 YOLOv5 提供了更快的推理速度、更小的模型大小和更高的准确性, 特别是在特定应用场景中表现更为突出。例如,

通过改进的 YOLOv5 模型, 实现了自主水下车 (AUV) 的水下垃圾检测与收集功能^[6]; Wang 等^[7]考虑到水下图像采集的缺陷, 如弱光、低分辨率和许多干扰, 提出了一种基于注意力改进的 YOLOv5 的水下目标检测算法; 而 Deng 等^[8]则基于 Mask R-CNN 设计了改进算法以适应海洋环境。另有许多研究者也在这一领域做出了贡献^[9-13]。

本文在此背景下进行了一系列改进。首先, 考虑到水下环境的复杂性, 引入了新模块以增强特征提取能力。其次, 为适应各种大小和形状的垃圾, 加入了 SEAttention 模块。最后, 使用 EIoU 评估指标取代了原有的 CIoU, 引入 Focal Loss 以解决样本不平衡问题, 从而提高模型的检测精度。

为验证效果, 在公开的深海垃圾数据集上进行了实验。结果显示, 与原始 YOLOv5 相比, 改进后的模型在精确度、召回率和 mAP@0.5 上均有明显提升, 同时参数量减少了 17.4%。这证明了提出的方法在水下垃圾检测中的有效性和优越性。

收稿日期: 2023-10-31 修稿日期: 2023-11-28

基金项目: 江苏省高等学校自然科学研究面上项目(19KJB520032); 江苏师范大学博士学位教师科研支持项目(20XSRS018)

作者简介: 张征鑫(1999—), 男, 江苏徐州人, 硕士, 研究方向为深度学习、计算机视觉; *通信作者: 张笃振(1967—), 男, 江苏徐州人, 博士, 副教授, 研究方向为图像处理、模式识别、机器学习等, E-mail: zhduzhen@jsnu.edu.cn

1 YOLOv5 模型

1.1 YOLOv5 整体架构介绍

YOLOv5 是由 Glenn Jocher 开发的, 代表了深度学习目标检测领域的先进水平, 实现了实时推理速度与高精度的完美平衡。它兼容多种部署环境, 已成为当前广泛采用的实时对象检测算法之一。

其前身版本 YOLOv1 至 YOLOv4 各有卓越之处。YOLOv1^[14]创新地将目标检测转化为回归问题, 但是在小目标和定位精度上存在局限。YOLOv2^[15]引进了多尺度预测和锚框, 显著改进了小目标检测和定位性能。YOLOv3^[16]利用更多特征层和不同尺度的锚框, 进一步提高了分类和检测的精度。YOLOv4^[17]融入了众多优化技术, 如 CIoU 损失和跨阶段连接, 实现了在精度和速度方面的双重飞跃。YOLOv5 经过模型结构和训练策略的全面优化, 具备了更小的模型参数和更快的速度, 适应性强。

从架构上看, YOLOv5 包括三个主要部分: backbone、neck 和 head。backbone 负责提取图像特征, neck 进行特征融合, head 通过 Detect 层精确输出检测结果。其高效率和易用性使其在自动驾驶、智能监控和医疗诊断等领域表现卓越, 展现了强大的性能和广泛的应用前景。这一算法不仅性能优异, 且易于使用, 在目标检测领域中占据了领先地位。

1.2 YOLOv5 主要模块

C3 模块是 backbone 的一部分, 用于深入提取图像特征。它通过卷积层和 Bottleneck 层对输入特征图进行处理, 实现了信息的压缩和特征的提取, 并利用残差结构避免了梯度消失, 确保了网络的稳定训练。

SPPF 模块是对 SPP 模块的改进, 位于 backbone 的最后一层, 执行空间金字塔池化操作, 通过多尺度池化捕捉不同尺度的特征信息, 增强模型对目标的感知能力, 提升目标检测的性能和精确度。

head 模块采用特征金字塔的结构, 通过结合浅层的图像特征和深层的语义特征, 形成不同分辨率的图像层次。这种设计使模型能够在不同尺度上学习图像特征, 提升预测的准确性。

这三个模块共同构成了 YOLOv5 的关键部

分, 优化了模型对多尺度目标的感知能力并提高了目标检测的精确度。通过集成这些高效的特征提取和融合机制, 模型能够更好地适应不同的检测场景, 展现出卓越的性能。这些优化确保了 YOLOv5 在复杂和挑战性的任务中的竞争力, 使其成为水下目标检测等应用的理想选择。

2 YOLOv5 模型的改进与优化

2.1 模型的改进方法

在水下环境中进行目标检测面临诸多挑战, 其中包括较暗的光线条件、海洋生物和地质结构的干扰, 以及垃圾类别间数量差异极大的问题。这些因素共同导致获取的图像质量不佳, 垃圾检测变得更加困难。

为了克服这些难题, 我们对 YOLOv5 模型的主干网络进行了深入的改进。引入了 MLP_Block, 一个将局部卷积和多层感知机相结合的灵活且高效的模块, 用以替换原有的 C3 模块。这种设计使得模型能够在光照不足和视觉模糊的水下环境中, 更好地提取深海垃圾的特征。实验证明, MLP_Block 在处理复杂任务和大数据集时, 相较于传统的 C3 模块, 展现了更为卓越的性能。

为了进一步提升模型的性能, 我们在模型的 head 层引入了 SEAttention 模块, 专注于关键特征的提取, 从而解决了背景干扰和数据不均衡的问题。SEAttention 通过学习不同通道间的相关性和分配权重, 增强了模型对重要特征的关注度, 提升了其判别能力和泛化性能。通过这种方式, 使得本文模型在检测水下垃圾时, 能够更准确地区分目标和背景, 减少了误检和漏检的情况。

与此同时, 我们将 CIoU 替换为 EIou, 以更精确地衡量预测和真实边界框之间的相似度, 从而提升水下垃圾检测的定位准确性。EIou 在考虑重叠面积、中心点距离和宽高比的基础上, 引入了额外的度量标准, 如边界框的形状和大小, 确保在水下复杂多变的环境中实现稳健的性能, 特别是对形状不规则或尺寸较小的垃圾, 能提供更准确的边界框定位。这一改进, 结合 MLP_Block 和 SEAttention 模块的引入, 显著提升了模型在水下垃圾检测任务上的准确性和鲁棒性。改进后的模型为 YOLOv5-MS (YOLOv5-C3_MLP-SEAttention), 如图 1 所示。

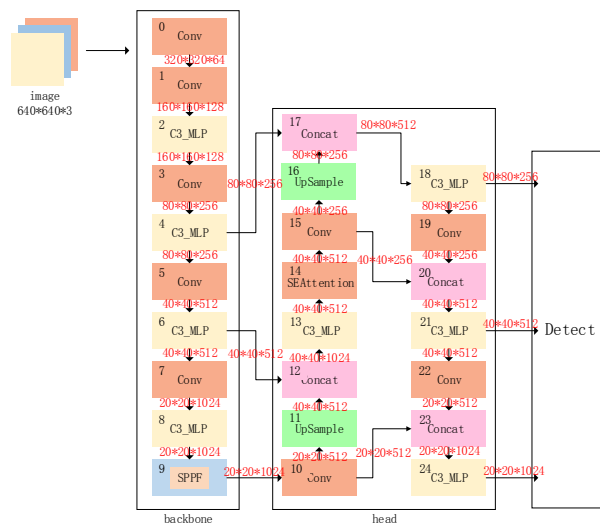


图1 YOLOv5-MS模型图

2.2 YOLOv5-MS网络结构

YOLOv5目标检测算法是一个优秀的检测算法，具有很好的实用性，但仍有可以改进的空间。根据前面的理论分析与研究，可以得到YOLOv5-MS，其模型结构见表1。

表1 YOLOv5-MS网络整体结构

序号	from	params	module	arguments
0	-1	3520	Conv	[3, 32, 6, 2, 2]
1	-1	18560	Conv	[32, 64, 3, 2]
2	-1	13248	C3_MLP	[64, 64, 1]
3	-1	73984	Conv	[64, 128, 3, 2]
4	-1	71168	C3_MLP	[128, 128, 2]
5	-1	295424	Conv	[128, 256, 3, 2]
6	-1	357888	C3_MLP	[256, 256, 3]
7	-1	1180672	Conv	[256, 512, 3, 2]
8	-1	826368	C3_MLP	[512, 512, 1]
9	-1	656896	SPPF	[512, 512, 5]
10	-1	131584	Conv	[512, 256, 1, 1]
11	-1	0	Upsample	[None, 2, 'nearest']
12	[-1, 6]	0	Concat	[1]
13	-1	272896	C3_MLP	[512, 256, 1, False]
14	-1	8192	SEAttention	[256]
15	-1	33024	Conv	[256, 128, 1, 1]
16	-1	0	Upsample	[None, 2, 'nearest']
17	[-1, 4]	0	Concat	[1]
18	-1	68608	C3_MLP	[256, 128, 1, False]
19	-1	147712	Conv	[128, 128, 3, 2]
20	[-1, 15]	0	Concat	[1]
21	-1	207360	C3_MLP	[256, 256, 1, False]
22	-1	590336	Conv	[256, 256, 3, 2]
23	[-1, 10]	0	Concat	[1]
24	-1	826368	C3_MLP	[512, 512, 1, False]

表中的from列表示输入来自哪一层，列中-1表示来自上一层的输出，params列表示参数的大小，module为子模块，arguments表示模块参数的信息，包括输入通道数、输出通道数、卷积核大小、步长信息等。

2.3 C3_MLP部分

Chen等^[18]提出了PConv卷积，并在此基础上提出了MLP_Block和FasterNet，它在许多设备上取得了比其他神经网络更高的运行速度，而不会影响各种视觉任务的准确性。本文使用该模型中的MLP_Block来对YOLOv5的主干网络进行改进，将其C3模块更换成C3_MLP模块。在整体上降低了参数量，使得训练和预测更加快速。

C3_MLP模块包括了两个部分，Partial_conv3卷积和MLP_Block模块。

Partial_conv3卷积核心思想是将输入的特征图在通道维度上分为两个部分，然后仅对其中的一个部分进行卷积，以实现局部卷积的效果。当向前传播方式设为“slicing”时，通过复制输入特征图为“x”，仅对一个部分进行卷积，最后重新合并，适用于推理阶段，减轻计算负担，保留原始输入信息，同时实现了局部卷积。当向前传播的方式设为“split_cat”时，使用torch.split()分割输入特征图为两个部分，然后只对一个部分进行卷积操作，再将两个部分通过cat()操作进行合并，适用于训练和推理。这两个方式总体上所做的事情相同，其不同点在于“slicing”方式不需要额外的内存，直接在输入特征图上进行操作，同时避免了分割和拼接的开销。而“split_cat”方式生成了新的向量，需要额外的内存，并且包含了分割和拼接操作，执行效率更低。

Partial_conv3卷积的优点在于提供了局部特征的卷积处理，保留了原始输入信息，避免了信息丢失，同时有效地提升了计算效率，适用于多种场景。具体结构如图2所示。

MLP_Block模块用于在神经网络中的某个层级上进行特征变换和增强。首先，对输入数据实行了Partial_conv3卷积操作，这有助于从原始特征中提取关键信息；接下来，引入了两个卷积层构成的多层感知机(MLP)，这些层可以

将特征进一步映射和转换,来获得更有用的信息;然后,经过了DropPath操作,通过随机丢弃MLP输出的一部分内容,有助于避免模型过拟合;最后,将MLP的输出与输入数据进行残差连接,以确保在特征转换后保留原始信息,构成了最终的输出。具体结构如图3所示。

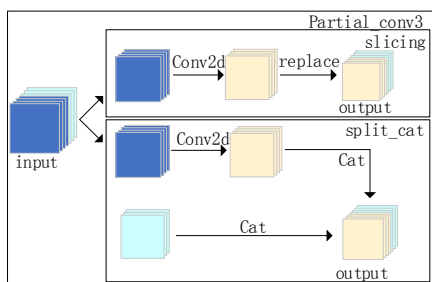


图2 Partial_conv3 结构图

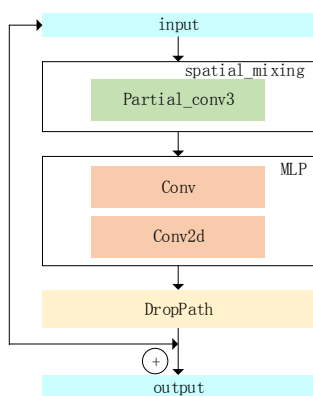


图3 MLP_Block 结构图

MLP_Block 模块提供了一个功能丰富、灵活可配置、具有正则化和鲁棒性的模块化设计。相比于原始的C3模块来说,增加了正则化操作,有助于减少过拟合的风险。MLP_Block 模块采用了部分卷积的操作更加注重原始输入信息的提取。同时,MLP_Block 模块中的MLP操作有助于提取输入更丰富的特征,提高模型的性能和泛化能力,总体而言,相比于C3模块更适合复杂的计算机视觉和图像处理任务。

改进后的模型中的C3_MLP模块继承了YOLOv5原模型中的C3模块,通过参数的修正确定了C3_MLP模块中的通道数缩放后的通道数,最后调用了上文提到的MLP_Block模块来增强特征表示能力。

2.4 SEAttention 部分

为了进一步提高YOLOv5目标检测的算法性能,我们引入了SEAttention模块,基于注意力机制思想,增强模型对关键目标特征的学习能力。这一改进为了提高在水下环境下的检测精确度,并通过实验证明了其有效性。

SEAttention模块最初由Hu等^[19]在2018年提出,其灵感来源于人类快速识别图像全局信息,以突出需要关注的候选区域,从而获取更多目标细节^[20]。其核心思想是通过学习通道之间的关系,自动为每个通道分配适当的权重,从而使模型能够更好地关注和利用重要的特征信息,提高模型的表示能力和鲁棒性。具体结构如图4所示。

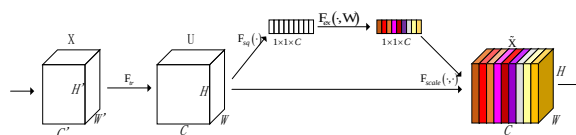


图4 SEAttention 模型图

在SEAttention模型中包含三个操作,挤压(Squeeze)操作、激励(Excitation)操作、缩放(Scale)操作。

挤压操作通过全局平均池化,将特征图的 $H \times W$ 维度降至 1×1 。激励操作引入小型多层感知机(MLP)作为注意力模型,逐步减少通道数并引入非线性,最终产生C维的注意力权重。缩放操作通过注意力权重重新加权原始特征图,强调重要性特征。

总体而言,SEAttention模型通过Squeeze-and-Excitation块实现通道注意力机制,自适应学习通道权重,并将这些权重应用于特征图,以提高重要特征的表示能力。这种注意力机制的引入对于提升CNN性能和提高特征提取能力具有重要意义。

2.5 EIou 部分

在目标检测任务中,生成各种评估指标的过程很大程度上取决于候选框与真实框之间的交并比(IoU)^[21]。IoU用于度量两个边界框之间的相似性和重叠程度,它是目标检测和物体跟踪任务中最常用的评价指标之一。提高IoU值也

是提高目标检测性能的一个关键方法。通过计算边界框与真实边界框的IoU，我们可以确定检测结果的准确性，并进一步评估目标检测算法的性能。其计算公式如下：

$$IoU = \left(\frac{inter}{union + \varepsilon} \right)^\alpha \quad (1)$$

其中： $inter$ 表示两个边界框的交集区域的大小， $union$ 表示两个边界框的并集区域的大小。

然而，标准IoU仅考虑了边界框的位置和形状，忽略了中心点距离以及凸包信息，为了更全面地度量相似性，YOLOv5算法引入了完整交并比(CIoU)^[22]的评价指标，其公式如下：

$$CIoU = IoU - \left(\frac{\rho^2}{c^2} + v \times \alpha \right) \quad (2)$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w2}{h2} - \arctan \frac{w1}{h1} \right)^2 \quad (3)$$

$$\alpha = \frac{v}{v - IoU + 1 + \varepsilon} \quad (4)$$

其中： ρ^2 表示两个边界框中心点的欧几里得距离的平方， c^2 表示两个边界框的凸包对角线的长度的平方，参数 v 与边界框宽度和高度的比例变化相关，参数 α 用于调整评估的特性，较大的值更强调更大的重叠，较小的值更加平衡。相比于原始的IoU作为相似性评估指标，CIoU考虑了边界框之间的位置和尺寸信息，提供了更准确的相似性度量，尤其是在边界框尺寸相差比较大的情况下，提高了目标检测的精度。

本文采用了扩展交并比(EIoU)^[23]作为相似性评估指标，EIoU是对CIoU的进一步改进，引入了更多的因素，更全面地考虑了边界框之间的相似性，其公式如下：

$$EIoU = IoU - \left(\frac{\rho^2}{c^2 + \varepsilon} + \frac{\rho_w^2}{cw^{2\alpha} + \varepsilon} + \frac{\rho_h^2}{ch^{2\alpha} + \varepsilon} \right) \quad (5)$$

其中： ρ^2 和 c^2 所表示的含义和CIoU相同， ρ_w^2 和 ρ_h^2 分别表示两个边界框宽度和高度之差的平方， cw 和 ch 分别表示两个边界框的凸包宽度和高度。

由于EIoU相比于CIoU不仅考虑了中心点距离和凸包信息，还加入了高度和宽度的变量，可以更好地捕捉不同尺寸和形状的物体。EIoU的复杂度更高，因此可能在某些情况下能够更准确地识别重叠的目标，特别是对于高度不均匀的边界框。由于水下场景中通常具有更复杂

的环境和目标，因此EIoU可以为水下目标检测提供更高的准确性和可靠性。

3 实验

3.1 实验环境

实验是在Kaggle在线平台上进行的，该平台上的GPU型号为Tesla P100-PCIE，显存大小为16 GB，内存大小为13 GB。YOLOv5模型基于PyTorch深度学习框架，编程语言为Python3.10.10，使用CUDA11.8对GPU进行加速。

3.2 数据集

实验所用的数据集DeepTrash^[24]是通过加州(南太浩湖、博德加湾、旧金山湾)海洋塑料的视频整理的。这些视频在质量、深度和能见度方面差异很大，能更好地反映海洋环境的严酷程度。该数据集标注了深海图像中的垃圾物体，为目标检测算法提供了定位和识别目标的关键信息，数据集中包括了各种垃圾的类别，包括塑料袋、塑料瓶、易拉罐、包装纸等，这些垃圾都被分到目标检测的同一类别。

该数据集包含了1900幅训练图像、637幅测试图像、637幅验证图像，以6:2:2的比例划分，其中包含各类垃圾和各种环境下的图像，部分图像如图5所示。



图5 DeepTrash数据集中图像示例

3.3 不同模型实验结果对比

表2给出了不同模型在数据集DeepTrash上的检测结果。可以看出改进后的YOLOv5-MS模型在参数量、精确度、召回率和mAP各个指标上都有明显提升。本文没有采用更高版本的YOLO算法，是因为更高版本的算法拥有更多的参数量，相比于YOLOv5需要更长的训练时间，但从表2也可以看出它们之间各种指标的差距不大，因此本文采用YOLOv5模型进行改进。

表2 不同模型实验结果对比

模型	参数量	Precision/%	Recall/%	mAP@0.5/%
SSD	23612246	82.5	64.0	70.6
EfficientDet	3828002	78.2	61.2	69.7
YOLOv3	8666692	82.6	66.4	72.4
YOLOv4	9103334	84.4	67.3	75.1
YOLOv5	7012822	84.8	68.0	76.7
YOLOv6	18751698	85.3	70.9	78.1
YOLOv7	36481772	87.5	70.0	77.9
YOLOv8	3005843	81.9	69.4	75.7
YOLOv5-MS	5790486	87.9	69.8	79.1

3.4 消融实验

为了更全面地验证本文提出的算法在目标检测性能方面的效果,以及深入探究各项改进方法的有效性,我们在YOLOv5s的基础上设计了七组消融实验。每组实验采用相同的超参数和训练技巧,以确保实验的可比性和可信度。实验结果见表3。

表3 消融实验

组别	C3_MLP	SEAttention	ElIoU	参数量	Precision/%	Recall/%	mAP@0.5/%
1	×	×	×	7012822	84.8	68.0	76.7
2	√	×	×	5782294	85.6	67.7	76.0
3	×	√	×	7021014	82.6	70.5	77.6
4	√	√	×	5790486	82.1	65.2	75.3
5	×	√	√	7021014	85.5	69.7	77.9
6	√	×	√	5782294	86.6	71.5	79.6
7	√	√	√	5790486	87.9	69.8	79.1

通过观察表3的数据,我们可以发现,在将原始的YOLOv5模型中的C3模块替换为C3_MLP模块后,模型的参数量减少了约17.4%。这不仅大大降低了计算负担,还在一定程度上提高了召回率和平均精度均值,然而精确度稍微有所下降。同时,引入SEAttention注意力机制和采用ElIoU来评估目标框的方法,虽然增加了很少的参数,却对精确度有了显著的提升,相较于未做任何改动的模型,精确度提高了2~3个百分点。

此外,需要注意SEAttention注意力机制的位置也可能影响实验结果。因此,我们进行了三组实验来探究不同位置下SEAttention注意

力机制的效果。第一组实验在YOLOv5的head模块的每一层C3_MLP后都添加了SEAttention注意力机制。第二组实验在YOLOv5的head模块中的第三层C3_MLP后引入SEAttention注意力机制。第三组实验在head模块的第一层C3_MLP后引入SEAttention注意力机制。实验结果见表4。

表4 不同位置的SEAttention模块的实验结果

组别	参数量	Precision/%	Recall/%	mAP@0.5/%
1	5833494	84.5	70.8	78.3
2	5800726	82.2	70.1	77.2
3	5790486	87.9	69.8	79.1

综合考虑,从表现最佳的角度出发,将C3_MLP模块与SEAttention注意力机制相结合,并通过ElIoU方法来评估检测框的质量,能够取得最优的实验结果。

在这两组实验中,我们通过在不同的位置引入SEAttention注意力机制,旨在探索最佳的注意力机制应用策略。通过对比实验结果,我们可以评估不同位置的SEAttention注意力机制对检测性能的影响。最终确定在head模块中第一层C3_MLP后引入SEAttention注意力机制能够取得最佳的检测结果。

3.5 实验结果对比

图6展示了YOLOv5算法和改进后的YOLOv5-MS算法在DeepTrash数据集上进行的目标检测结果对比。

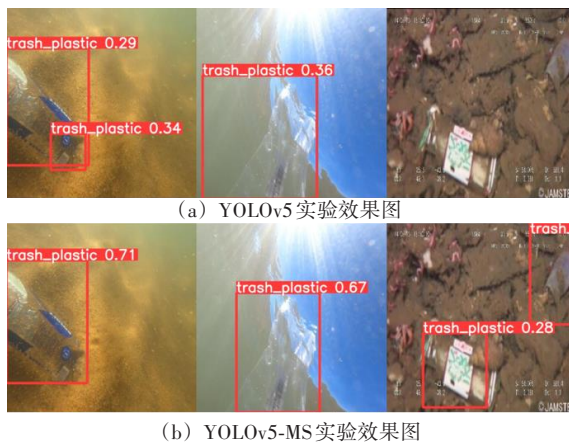


图6 YOLOv5和YOLOv5-MS效果对比

从图6可以看到, YOLOv5算法有可能会将一个目标检测成多个目标, 在光照强的图片里检测目标的置信度偏低, 还会出现一些漏检的情况, 经过改进后的YOLOv5-MS可以避免这些问题。结果表明, 与原版YOLOv5相比, 改进算法YOLOv5-MS在该数据集上的目标检测精度更高, 检测结果更佳。这验证了改进算法在目标检测任务上的有效性。

3.6 其他数据集验证

鉴于水下目标检测中水下垃圾检测的普遍性, 本文实验主要采用了DeepTrash数据集来进行水下检测的训练, 为了确保YOLOv5-MS模型的有效性, 我们在TrashCan数据集上^[25-26]进行了实验。TrashCan数据集包含了7212幅标注图像, 其中训练集包含了6065幅图像, 验证集包含了1147幅图像, 记录了水下垃圾、遥控潜水器(ROVs)以及各种海底植物和动物的观察结果。TrashCan数据集的图像来自于由日本海洋科学技术机构(JAMSTEC)策划的J-EDI(JAMSTEC深海图像电子图书馆)数据集。该数据集包含自1982年以来JAMSTEC操作的遥控潜水器在日本海等地采集的图片。通过发布该数据集, 希望能够推动海洋机器人在应对紧迫的自主垃圾检测和清理方面发挥更大作用^[27]。

表5展示了YOLOv5-MS模型在TrashCan数据集上进行训练测试的结果。从表5可以看出, 相比于原始网络YOLOv5, 改进后的YOLOv5-MS在精确度上提高了1.9个百分点, 在mAP指标上提高了0.3个百分点。同时在验证时, 两个模型的图像预处理和模型推理时间相同, 但NMS(非极大值抑制)所花费的时间YOLOv5模型花费0.8 ms, YOLOv5-MS模型花费0.7 ms。由此可见YOLOv5-MS的模型的复杂程度相对较低, 训练时长也相对较低, 更利于训练和预测。

表5 TrashCan数据集结果对比(%)

模型	Precision	Recall	mAP@0.5
YOLOv5	76.3	65.0	67.7
YOLOv5-MS	78.2	63.0	68.0

4 结语

本文针对传统的YOLOv5算法进行了改进, 通过替换C3模块为C3_MLP、使用EIoU来评估检测框的质量以及在YOLOv5的head模块中引入了SEAttention模块, 取得了显著的性能提升。相较于原始网络结构, 本文的改进算法在DeepTrash水下垃圾数据集中的精确度上提高了3.1个百分点, 召回率提高了1.8个百分点, mAP@0.5提高了2.4个百分点, 同时参数量下降了17.4个百分点。同时在TrashCan数据集上的实验结果也有所提升。验证了所提出的改进策略的有效性。替换C3模块为C3_MLP增强了特征提取能力, EIoU的使用优化了检测框的匹配度, 而SEAttention模块则有助于提升对重要目标的关注度。尽管在精确度上取得了显著的提升, 但仍存在一些改进的空间。未来的研究可以进一步优化改进算法的设计, 并在更广泛的数据集和应用场景中进行验证。

参考文献:

- [1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, 2014: 580-587.
- [2] GIRSHICK R. Fast R-CNN [C] // Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440-1448.
- [3] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal net-works [C] // Advances in Neural Information Processing Systems, 2015, 91-99.
- [4] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [C] // Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, Springer International Publishing, 2016: 21-37.
- [5] ULTRALYTICS. Ultralytics. yolov5 [EB/OL]. [2022-06-01]. <https://github.com/ultralytics/YOLOv5>.
- [6] TENG X, FEI Y, HE K, et al. The object detection of underwater garbage with an improved YOLOv5 algorithm [C] // Proceedings of the 2022 International

- Conference on Pattern Recognition and Intelligent Systems, Hammamet, Tunisia, 2022:55-60.
- [7] WANG J, YU N. UTD-YOLOv5: a real-time underwater targets detection method based on attention improved YOLOv5[EB/OL]. arXiv:2207.00837, 2022.
- [8] DENG H, ERGU D, LIU F, et al. An embeddable algorithm for automatic garbage detection based on complex marine environment [J]. Sensors, 2021, 21 (19):6391.
- [9] MACE T H. At-sea detection of marine debris: overview of technologies, processes, issues, and options [J]. Marine Pollution Bulletin, 2012, 65 (1/2/3) : 23-27.
- [10] VALDENEGRO-TORO M. Submerged marine debris detection with autonomous underwater vehicles [C]//Proceedings of the 2016 International Conference on Robotics and Automation for Humanitarian Applications(RAHA), Kerala, India, 2016:1-7.
- [11] FULTON M, HONG J, Islam M J, et al. Robotic detection of marine litter using deep visual detection models [C]//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, Canada, 2019:5752-5758.
- [12] POLITIKOS D V, FAKIRIS E, DAVVETAS A, et al. Automatic detection of seafloor marine litter using towed camera images and deep learning [J]. Marine Pollution Bulletin, 2021, 164:111974.
- [13] KYLILI K, HADJISTASSOU C, ARTUSI A. An intelligent way for discerning plastics at the shorelines and the seas [J]. Environmental Science and Pollution Research, 2020, 27(34):42631-42643.
- [14] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016:779-788.
- [15] REDMON J, FARHADI A. YOLO9000: better, faster, stronger [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, 2017:7263-7271.
- [16] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. arXiv:1804.02767, 2018.
- [17] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. arXiv:2004.10934, 2020.
- [18] CHEN J, KAO S, HE H, et al. Run, don't walk: chasing higher FLOPS for faster neural net-works [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023:12021-12031.
- [19] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018:7132-7141.
- [20] TANG X I, ZHONG B, PENG J P, et al. Multi-scale channel importance sorting and spatial attention mechanism for retinal vessels segmentation [J]. Applied Soft Computing Journal, 2020, 93 (5) : 321-329.
- [21] RAHMAN M A, WANG Y. Optimizing intersection-over-union in deep neural networks for image segmentation [C]//Proceedings of the International Symposium on Visual Computing. Springer, Cham, Las Vegas, USA, 2016:234-244.
- [22] ZHENG Z, WANG P, REN D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation [J]. IEEE Transactions on Cybernetics, 2021, 52 (8) : 8574-8586.
- [23] ZHANG Y F, REN W, ZHANG Z, et al. Focal and efficient IOU loss for accurate bounding box regression [J]. Neurocomputing, 2022, 506:146-157.
- [24] ZHENG Z, WANG P, REN D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation [J]. IEEE Transactions on Cybernetics, 2021, 52(8):8574-8586.
- [25] TATA G, ROYER S J, POIRION O, et al. A robotic approach towards quantifying epipelagic bound plastic using deep visual models [EB/OL]. arXiv: 2105.01882, 2021.
- [26] HONG J, FULTON M, SATTAR J. Trashcan: a semantically-segmented dataset towards visual detection of marine debris [EB/OL]. arXiv: 2007.08097, 2020.
- [27] KAKO S, MORITA S, TANEDA T. Estimation of plastic marine debris volumes on beaches using unmanned aerial vehicles and image processing based on deep learning [J]. Marine Pollution Bulletin, 2020, 155:111127.

(下转第63页)

文章编号: 1007-1423(2024)07-0009-08

DOI: 10.3969/j.issn.1007-1423.2024.07.002

ECPANet: 一种基于注意力的深度卷积神经网络通道剪枝方法

余显冰*, 杨礼友, 李 健

(四川大学电子信息学院, 成都 610065)

摘要: 在深度学习领域中, 卷积神经网络的快速发展导致了先进模型需要大量的计算和存储资源。然而, 将这些模型部署到计算和存储资源受限且高实时性的嵌入式设备上变得越来越具有挑战性。为了解决这个问题, 通道剪枝已成为网络压缩的主要方法之一。传统的通道剪枝方法存在着精度下降和难以确定通道重要性的问题。针对这些问题, 提出了一种高效的通道注意力剪枝方法。通过将ECPANet模块嵌入到深度卷积神经网络中以增强其表征能力, 评估每个通道在特征映射中的重要性, 并根据通道重要性因子剪枝掉不重要的通道以减小模型的大小和计算量。实验结果表明, 与传统的通道剪枝方法相比, 基于注意力的通道剪枝方法能够更准确地确定通道重要性, 从而提高剪枝效果和模型性能。

关键词: 深度卷积神经网络; 通道剪枝; 注意力机制

0 引言

随着深度学习的飞速发展, 深度卷积神经网络(deep convolutional neural networks)已成为计算机视觉^[1]和自然语言处理^[2]等领域的主流技术。但是, 这些网络复杂性越来越高, 它们的训练和部署成本也随之增加。在资源有限的设备上部署这些复杂的神经网络需要考虑模型规模、内存和计算操作数量的影响。

神经网络模型压缩的方法包括网络剪枝、网络量化、矩阵分解和知识蒸馏等技术。其中, 网络剪枝是压缩神经网络模型最有效的方法之一, 目的是去除神经网络中的冗余参数。通道剪枝是一种流行的结构化剪枝方法, 它不需要使用专门的软件和硬件, 基本思想是在性能损失^[3]可以忽略的情况下, 减少卷积层的输入/输出通道数。剪枝过程的关键步骤是度量每个通道的重要性, 以确定通道是否可以被移除。然而, 早期的研究只考虑了每个通道对重构损失的独立贡献, 而忽略了相邻通道之间的相关性对推断性能的影响。He等^[3]提出剪枝方法, 通

过考虑每个通道对特征图的重要性来决定保留哪些通道。该方法与其他通道剪枝方法相比, 压缩率更高但性能下降幅度也较大。Zhong等^[4]提出一种新的卷积核剪枝方法, 在神经网络训练过程中, 对不同的卷积核随机初始化并分配不同的学习任务。在训练结束后, 只有部分卷积核在网络中真正发挥了作用, 通过删除未发挥作用的卷积核进行剪枝。Wang等^[5]提出了使用SENet模型评估通道重要性, 以识别和删除最不重要的冗余通道。然而, SENet本身的限制导致其生成的比例值不能完全反映通道的重要性, 从而无法有效提高剪枝性能。Liu等^[6]通过引入空间注意力和通道注意力来指导通道剪枝, 提出了一种新的注意力模型, 并基于注意力统计量评估通道的重要性。然而, 这种方法忽略了不同剪枝层之间的关联。

为了解决上述问题, 本文提出了一种基于ECPANet (efficient channel pruning attention network) 注意力机制的通道剪枝方法。与传统的通道剪枝方法不同, 本方法考虑相邻通道之间的相关性, 使用注意力机制来动态地选择通道。

收稿日期: 2023-10-31 修稿日期: 2024-03-13

作者简介: *通信作者: 余显冰(1998—), 男, 四川攀枝花人, 硕士研究生, 研究方向为神经网络模型压缩、边缘计算, E-mail: yuxianbing@stu.scu.edu.cn; 杨礼友(1991—), 男, 贵州开阳人, 硕士研究生, 研究方向为神经网络模型压缩、边缘计算; 李健(1969—), 女, 四川成都人, 博士, 副教授, 研究方向为物联网与边缘计算、智能感知与认知、目标识别

通过使用注意力机制，我们可以更准确地衡量每个通道对模型的贡献，并且能够在准确率和剪枝率之间取得更好的平衡。

1 方法

1.1 注意力网络模块

注意力模块能够从输入特征中选择有用的信息，并将其集中在更有意义的特征上，从而提高模型的性能。本文提出轻量级的注意力机制模块ECPANet来让网络自主学习通道的重要性。ECPANet增强了DCNN的特征聚合，通过学习每个通道之间的关联性来提高网络的特征表现能力。ECPANet模块结构如图1所示，包括池化层、卷积层和激活函数。输入特征图 $X \in \mathbf{R}^{H \times W \times C}$ ，通过注意模块后得到通道重要性特征图 $X_s \in \mathbf{R}^{H \times W \times C}$ ，然后注意力特征图乘以标准化后的特征图，得到最终的通道注意力特征 $\tilde{X} \in \mathbf{R}^{H \times W \times C}$ ，如公式(1)：

$$\tilde{X} = X \otimes X_s(X) \quad (1)$$

在通道注意力计算阶段，对输入的特征图分别进行全局平均池化和全局最大池化得到 $1 \times 1 \times C$ 的特征图，通过一个 1×1 的动态卷积核来对压缩后的通道特征进行重要性学习。将全局平均池化后卷积运算和全局最大池化后卷积运算的结果进行聚合，可以更好地表现通道的重要性和相邻通道之间的相关性。将聚合后的特征通道通过一个 1×1 的2D深度可分离卷积计算每个通道的重要性评估值，然后通过sigmoid函数将其映射到0~1之间的范围，表示每个通道被保留的重要性程度。最后，将特征聚合后得到的

权重乘以标准化后的特征，得到最终的通道注意力特征。ECPANet模块引入了ECANet^[7]的自适应动态调整K值方法，如公式(2)：

$$K = \psi(C) = \left\lceil \frac{\log_2^{(C)}}{\gamma} + \frac{b}{r} \right\rceil_{\text{odd}} \quad (2)$$

其中： $\lceil \cdot \rceil_{\text{odd}}$ 表示最近奇数。在本文中，将 γ 和 b 设定为2和1，以确保在所有实验中使用相同的参数设置。

1.2 ECPANet注意力机制的通道剪枝方法

本文将ECPANet模块插入到每一层网络中，对其进行通道重要性评分，然后根据预定义的通道剪枝比例，将排序后的通道重要性因子列表的长度乘以预定的剪枝比例，计算出通道剪枝重要性评判阈值。如公式(3)：

$$\text{pruning_threshold} = \text{channel_important_factors_index} \times \text{pruning_ratio} \quad (3)$$

将每一层的通道重要性因子与剪枝阈值进行比较，大于阈值的通道保留，小于阈值的通道移除。如公式(4)：

$$\text{select_remain_channel} = \begin{cases} 1, & \text{channel_important_factors} > \text{pruning_threshold} \\ 0, & \text{channel_important_factors} < \text{pruning_threshold} \end{cases} \quad (4)$$

并将这些通道对应的滤波器和所有插入的ECPANet模块移除。最后，对剪枝后的网络进行微调，以恢复剪枝后网络的精度损失。图2展示了本文的通道剪枝方法的整体流程。

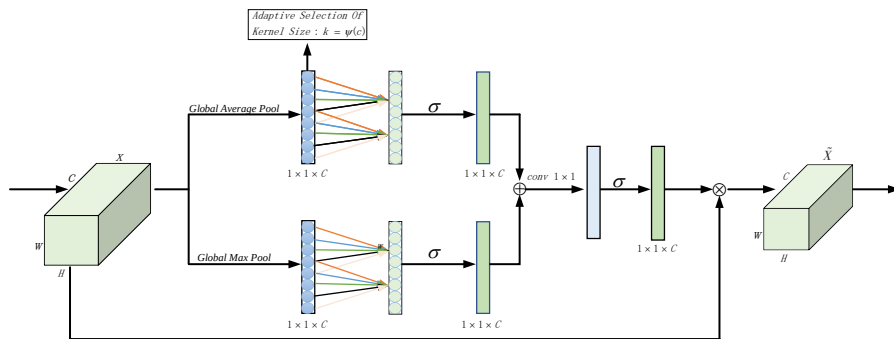


图1 ECPANet模块结构图

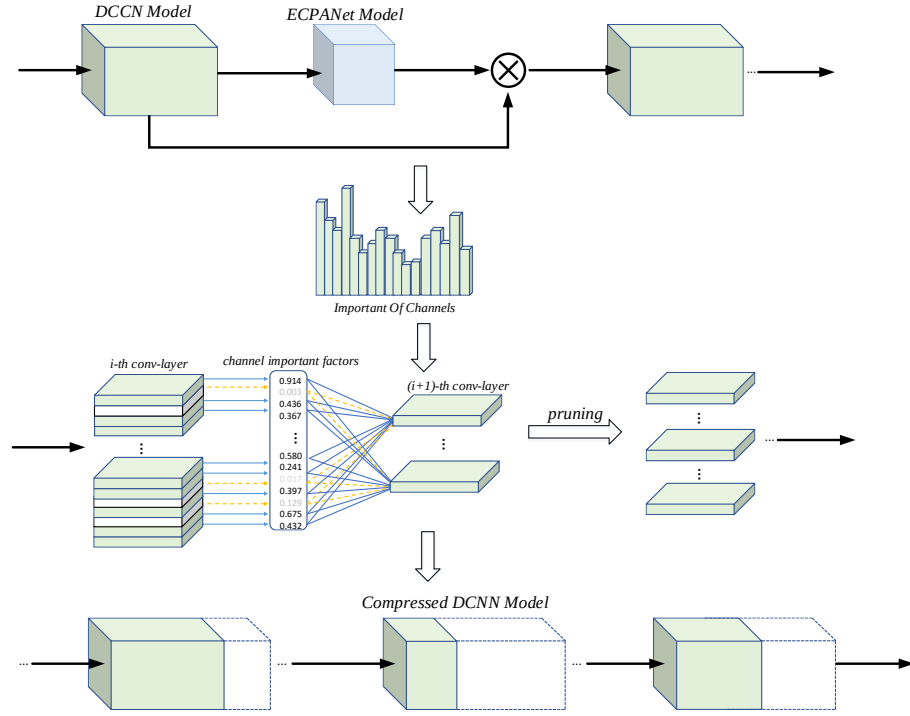


图2 ECPANet注意力机制的通道剪枝方法

在剪枝过程中,当某个通道被移除时,其对应的卷积核也会被移除。这意味着下一层中处理这个通道的卷积核也将不存在。因此,本文通过计算每一层的参数量和浮点运算量来评估剪枝后网络的计算复杂度,并对网络进行进一步优化。参数计算量如公式(5):

$$W_i = C_{in}^i \times (\Omega^i)^2 \times C_{out}^i \quad (5)$$

其中: W 是参数量, i 为卷积层数索引, C_{in} 为输入通道数, C_{out} 为输出通道数, Ω 是卷积核的大小。

浮点运算量如公式(6):

$$P_i = C_{in}^i \times (\Omega^i)^2 \times C_{out}^i \times S_{out}^i \quad (6)$$

其中: P 是浮点运算量, i 为卷积层数索引, C_{in} 为输入通道数, C_{out} 为输出通道数, Ω 是卷积核的大小, S_{out} 为输出层的特征尺寸。

根据图2中的剪枝方法,当对第 i 层进行通道剪枝,假定剪枝比例为 α ,将移除通道数为 $C_{out}^i \times \alpha$,与其对应的卷积核也将被移除。对应的参数量如公式(7):

$$W_i = C_{in}^i \times (\Omega^i)^2 \times C_{out}^i \times (1 - \alpha) \quad (7)$$

而浮点运算量如公式(8):

$$P_i = C_{in}^i \times (\Omega^i)^2 \times C_{out}^i \times (1 - \alpha) \times S_{out}^i \quad (8)$$

类似地,当第 i 层完成通道剪枝后,对第 $i + 1$ 层进行通道剪枝时,假定剪枝比例为 β ,将移除通道数为 $C_{out}^i \times \beta$,与其对应的卷积核也将被移除。对应的参数量如公式(9):

$$W_i = C_{in}^i \times (1 - \alpha) \times (\Omega^i)^2 \times C_{out}^i \times (1 - \beta) \quad (9)$$

而浮点运算量如公式(10):

$$P_i = C_{in}^i \times (1 - \alpha) \times (\Omega^i)^2 \times C_{out}^i \times (1 - \beta) \times S_{out}^i \quad (10)$$

最终,当 i 层剪枝完成后, i 层的参数量和浮点运算量都将减少为原来的 $(1 - \alpha) \times (1 - \beta)$ 。

1.3 剪枝步骤

本文的剪枝步骤如图3所示。具体步骤如下:

(1) 添加 ECPANet 注意力模块: 在需要剪枝的网络层上添加注意力模块,并对模型进行重新训练。

(2) 通道重要性评估: 利用注意力模块中的2D的 1×1 深度可分离卷积来对当前层的通道进行重要性评估。对新模型中的通道重要性因子进行读取,并进行排序。

(3) 选择剪枝通道: 根据预先定义的剪枝比例,选择需要剪枝的通道。

(4) 卷积核剪枝: 移除步骤(3)中所选择剪枝通道对应的卷积核, 并移除步骤(1)中添加的注意力模块。

(5) 再训练: 对剪枝后的网络进行再训练, 以恢复剪枝带来的精度损失。

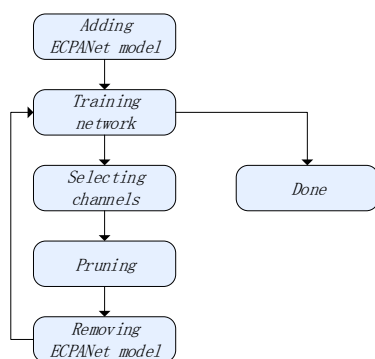


图3 剪枝步骤

传统卷积神经网络(如VGGNet^[8])可在每个卷积层后添加注意力模块进行剪枝。但对于带残差模块的模型(如ResNet^[9]、DenseNet^[10]), 由于短路连接的存在, 短路连接两端的通道数和特征图大小必须相同。若对这些层进行裁剪, 则会影响连接另一端层的通道重要性关系, 这会破坏整个网络结构, 进而影响网络性能。因此, 对于有残差结构的神经网络, 本文仅剪枝残差模块内除最后一个卷积层外的卷积层, 即在需要剪枝的卷积层之后添加注意力模块进行剪枝。

2 实验

2.1 数据集和网络模型

本文在实验中使用了CIFAR-10^[8]和

CIFAR-100^[8]数据集作为基准数据集。针对三种不同网络架构, 即VGGNet、ResNet和DenseNet, 对本文的剪枝方法进行了评估。对于VGGNet网络进行了修改, 取消了全连接层, 并采用了VGGNet-16、VGGNet-19进行实验。对于ResNet网络采用了通用的ResNet-56和具有瓶颈结构的ResNet-164进行实验。对于DenseNet网络使用了DenseNet-40进行实验。

2.2 实验设置

本文使用Python 3.7和PyTorch 1.4开发框架, 64位Ubuntu 20.04操作系统, CPU处理器为Intel Xeon E5-1650 v4 @ 3.60GHz, GPU处理器为NVIDIA GeForce GTX 1080 Ti, 内存为12 GB。使用动量为0.9的随机梯度下降算法(SGD)从头开始训练所有的网络。在CIFAR数据集上, 将批量大小设置为64, 进行160轮训练。初始的学习率设置为0.1, 并在总训练轮数的50%和75%时将学习率衰减10倍。使用权重衰减为 10^{-4} 。本文通过与几种先进的方法进行对比, 证实了本文提出的剪枝方法的有效性。

2.3 通道剪枝

本文基于CIFAR-10和CIFAR-100数据集上选择通道剪枝最常用的测试网络VGGNet-19、ResNet-164和DenseNet-40作为测试网络, 它们具有大多数深度神经网络模型所拥有的结构, 这样可以更好地评估本文方法的性能。实验结果见表1、表2, 本文的剪枝方法在不同比例的通道剪枝下取得了较好的效果。

表1 基于CIFAR-10数据集的实验结果

Model	Channel pruned/%	Accuracy/%	Parameters/ $\times 10^6$	FLOPs/ $\times 10^8$	Memory/MB
Original VGGNet-16	0	93.04	14.73	3.14	56.24
	30	93.62	6.89	1.81	26.32
	50	93.12	3.31	1.10	12.67
	70	92.71	1.12	0.64	4.31
	80	91.72	0.48	0.39	1.87
Original VGGNet-19	0	93.59	20.04	3.99	76.52
	30	93.85	9.41	2.29	35.95
	50	93.53	4.58	1.43	17.51
	70	92.86	1.56	0.72	5.99
	80	91.92	0.67	0.46	2.57

续表1

Model	Channel pruned/%	Accuracy/%	Parameters/ $\times 10^6$	FLOPs/ $\times 10^8$	Memory/MB
Original ResNet-56	0	93.63	0.59	0.90	2.41
Pruned	30	93.52	0.40	0.64	1.62
	50	93.22	0.29	0.49	1.20
	70	91.82	0.19	0.35	0.82
	80	89.74	0.14	0.26	0.61
Original ResNet-164	0	95.02	1.71	2.54	6.95
Pruned	30	94.83	1.09	1.80	4.44
	50	94.27	0.75	1.36	3.13
	70	93.40	0.45	0.94	1.99
	80	91.72	0.31	0.70	1.45
Original DenseNet-40	0	93.84	1.07	2.90	4.31
Pruned	30	94.32	0.75	2.19	3.06
	50	93.74	0.55	1.73	2.26
	70	92.07	0.34	1.21	1.46
	80	90.39	0.24	0.93	1.07

表2 基于CIFAR-100数据集的实验结果

Model	Channel pruned/%	Accuracy/%	Parameters/ $\times 10^6$	FLOPs/ $\times 10^8$	Memory/MB
Original VGGNet-16	0	72.06	14.77	3.14	56.39
Pruned	30	72.97	6.84	1.89	26.11
	50	71.97	3.32	1.16	12.69
	70	69.98	1.14	0.63	4.37
	80	66.77	0.49	0.39	1.89
Original VGGNet-19	0	72.71	20.08	3.99	76.71
Pruned	30	72.33	9.42	2.30	35.99
	50	71.41	4.64	1.43	17.73
	70	68.78	1.57	0.76	6.02
	80	65.66	0.68	0.49	2.62
Original ResNet-56	0	73.75	0.62	0.90	2.49
Pruned	30	72.21	0.39	0.64	1.59
	50	70.56	0.29	0.50	1.18
	70	65.00	0.18	0.34	0.78
	80	58.22	0.13	0.26	0.59
Original ResNet-164	0	76.95	1.73	2.54	7.04
Pruned	30	75.31	1.13	1.84	4.59
	50	73.15	0.78	1.39	3.25
	70	70.61	0.47	0.93	2.05
	80	65.52	0.32	0.69	1.49
Original DenseNet-40	0	72.93	1.10	2.90	4.67
Pruned	30	73.95	0.79	2.22	3.16
	50	72.42	0.57	1.76	2.34
	70	68.88	0.35	1.23	1.51
	80	63.52	0.24	0.92	1.08

2.4 与先进注意力模块的对比实验

本文针对流行的VGGNet-16和ResNet-56网络结构进行了图像分类实验，以评估本文的ECPANet注意力模块。

表 3 CIFAR-10数据集上不同注意力模块在VGGNet-16、ResNet-56上的对比实验

Model	Variant	Accuracy/%	与基线差值/个百分点
VGGNet-16	baseline	93.04	—
	+SENet	93.96	+0.92
	+CBAM	93.73	+0.69
	+ECANet	94.03	+0.99
	+ECPANet	94.11	+1.07
ResNet-56	Baseline	93.63	—
	+SENet	94.05	+0.42
	+CBAM	93.72	+0.09
	+ECANet	93.90	+0.19
	+ECPANet	94.25	+0.62

在VGGNet-16网络中，本文在每个卷积块上放置了ECPANet模块，而在ResNet-56网络中，ECPANet模块被放置在每个残差块的卷积输出上。本文在CIFAR-10数据集上对ECPANet与SENet^[11]、CBAM^[12]、ECANet等最先进的注意力模块进行了比较，实验结果表明，ECPANet注意力模块具有先进性。

2.5 与先进剪枝方法的对比实验

本文基于ECPANet注意力的通道剪枝方法与一些代表性的剪枝方案进行了比较，包括CHIP^[13]、TMI-GKP^[4]、SENet^[5]、CPSCA^[6]、Slimming^[14]、DACP^[15]、GFP^[16]、FPVG^[17]、GAL^[18]、HRank^[19]、Random^[20]。本文在CIFAR-10数据集上比较了不同剪枝方案在相似剪枝比例下的效果，结果如表4所示。

2.6 实验结果分析

实验结果表明，本文的方法具有先进性。

表 4 CIFAR-10数据集上使用VGGNet-16、ResNet-56进行不同剪枝方法的对比

Network	Method	CIFAR-10			
		Params Drop/%	Channel Pruned/%	FLOPs Reduction/%	Accuracy Variation/个百分点
VGGNet-16	Slimming	79.96	—	77.02	-0.26
	DACP	93.50	—	65.00	-0.58
	GAL	77.60	—	39.60	-0.19
	HRank	82.90	—	53.50	-0.53
	CHIP	81.60	—	58.10	-0.10
	SENet	75.50	—	74.97	-0.27
	CPSCA	79.96	—	77.02	-0.08
	ECPANet	77.51	50	64.81	+0.08
		85.86	60	72.31	-0.32
		92.38	70	79.58	-0.33
ResNet-56		96.71	80	87.36	-1.32
	Slimming	47.67	—	32.76	-0.72
	DACP	70.40	—	47.10	-0.01
	GFP	53.30	—	60.10	-0.33
	FPVG	40.00	—	52.60	+0.10
	TMI-GKP	43.49	—	43.23	+0.22
	Random	44.92	—	48.97	+0.22
	SENet	50.59	—	49.80	-1.02
	ECPANet	42.16	40	38.21	-0.01
		50.93	50	54.93	-0.41
		59.19	60	52.47	-0.97
		67.79	70	61.10	-1.81
		76.73	80	70.55	-3.89

在VGGNet-16上,当通道剪枝比例达到50%时,参数量减少了77.51%,浮点运算量降低了64.81%,本文仍然可以提高0.08%的精度;当通道剪枝比例达到80%时,参数压缩率达到96.71%,浮点运算量降低了87.26%,而精度只降低了0.22%。在ResNet-56上,传统的网络剪枝实验采用ResNet V1网络结构,这种网络结构参数量和计算量相对较大,网络中的冗余参数更多,通道剪枝粒度大。本文采用的是ResNet V2网络结构,该结构下的ResNet-56网络参数量和计算量分别是ResNet V1的0.69倍和0.36倍,网络中的冗余参数和计算量更少,更深层次的网络采用该结构可减少硬件资源并提高运行速度。在同样的通道剪枝粒度下,虽然本文的方法损失了些许精度,但本文的参数量和计算量压缩倍数分别是传统方法的1.6倍和3倍以上,在保证一定精度要求下,实现了更大的参数量和计算量压缩。

3 结语

本文改进了一种新型的注意力模块,并提出基于该模块的通道剪枝方法。在需要剪枝的卷积层后引入注意力模型,对当前层的特征通道进行增强或抑制,并进行通道重要性评分。根据评分结果对需要剪枝的卷积层进行合理的剪枝。实验结果表明,在同样的剪枝粒度下,本文的方法取得了更好的效果。在一定精度损失范围内,本文的方法还可以进一步压缩网络模型。此外,本文的方法灵活简便,不依赖任何专用的库和硬件平台。还可以与其他模型压缩策略结合,进一步压缩网络模型。

参考文献:

- [1] KENDALL A, GAL Y. What uncertainties do we need in Bayesian deep learning for computer vision? [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.
- [2] CHOWDHARY K. Natural language processing[J]. Fundamentals of Artificial Intelligence, 2020, 17(6): 603-649.
- [3] He Y H, ZHANG X Y, SUN J. Channel pruning for accelerating very deep neural networks[C]//Proceedings of the IEEE International Conference on Computer Vision, Hawaii, 2017: 1389-1397.
- [4] ZHONG S C, ZHANG G Q, HUANG N J, et al. Revisit kernel pruning with lottery regulated grouped convolutions [C]//Proceedings of the International Conference on Learning Representations, 2022.
- [5] WANG X J, YAO W B, FU H Y. A convolutional neural network pruning method based on attention mechanism [C]//Proceedings of the International Conference on Software Engineering and Knowledge Engineering, Pittsburgh, 2019: 343-452.
- [6] LIU M R, FANG W W, MA X D, et al. Channel pruning guided by spatial and channel attention for DNNs in intelligent edge computing[J]. Applied Soft Computing, 2021, 110: 107636.
- [7] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Haikou, 2020: 11534-11542.
- [8] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images [J]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4): 1-10.
- [9] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016: 770-778.
- [10] HUANG G, LIU Z, VAN D M L, et al. Densely connected convolutional networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2017: 4700-4708.
- [11] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Utah, 2018: 7132-7141.
- [12] WOO S, PARK J, LEE J, et al. CBAM: convolutional block attention module [C]//Proceedings of the European Conference on Computer Vision (ECCV), Munich, 2018: 3-19.
- [13] SUI Y, YIN M, XIE Y, et al. CHIP: channel independence-based pruning for compact neural networks [J]. Neural Information Processing Systems, 2021, 34: 24604-24616.
- [14] LIU Z, LI J D, SHEN Z Q, et al. Learning efficient convolutional networks through network slimming [C]//Proceedings of the IEEE International Conference on Computer Vision, Hawaii, 2017: 2736-2744.
- [15] ZHUANG Z W, TAN M K, ZHUANG B H, et al.

- Discrimination-aware channel pruning for deep neural networks [C] // Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018.
- [16] YOU Z H, YAN K, YE J M, et al. Gate decorator: global filter pruning method for accelerating deep convolutional neural networks [C] // Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 2133–2144.
- [17] HE Y, LIU P, WANG Z W, et al. Filter pruning via geometric median for deep convolutional neural networks acceleration [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019: 4340–4349.
- [18] LIN S H, JI R R, YAN C Q, et al. Towards optimal structured CNN pruning via generative adversarial learning [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019: 2790–2799.
- [19] LIN M B, JI R R, WANG Y, et al. HRank: filter pruning using high-rank feature map [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2020: 1529–1538.
- [20] LI Y W, ADAMCZEWSKI K, LI W, et al. Revisiting random channel pruning for neural network compression [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, 2022: 191–201.

ECPANet: An attention-based method for channel pruning of deep convolutional neural networks

Yu Xianbing*, Yang Liyou, Li Jian

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: In the field of deep learning, the rapid development of Convolutional Neural Networks (CNNs) has led to advanced models that require substantial computational and storage resources. However, deploying these models on resource-constrained and highly real-time embedded devices has become increasingly challenging. To address this issue, channel pruning has become one of the primary methods for network compression. Traditional channel pruning methods suffer from accuracy degradation and difficulties in determining channel importance. To tackle these issues, an efficient channel attention pruning method has been proposed. This method involves embedding ECPANet modules into deep convolutional neural networks to enhance their representational capacity, evaluate the importance of each channel in feature maps, and prune unimportant channels based on their importance factors to reduce the model's size and computational load. Experimental results show that, compared to traditional channel pruning methods, attention-based channel pruning methods can more accurately determine channel importance, thereby improving pruning effectiveness and model performance.

Keywords: deep convolutional neural network; channel pruning; attention mechanism

文章编号: 1007-1423(2024)07-0017-07

DOI: 10.3969/j.issn.1007-1423.2024.07.003

一种针对低质量指纹证据图像的增强算法设计

卢超超*

(山西警察学院侦查系, 太原 030401)

摘要: 指纹图像采集过程中存在旋转、平移以及图像采集不全等内在差异, 导致每次提取到的细节点数目是不固定的, 这种特征形式后期匹配计算复杂度较高。当前基于模板匹配的图像增强算法针对低质量指纹证据图像匹配计算复杂度较高的情况, 以固定阈值完成原始图像滤波处理, 未考虑单像素细化方向场特征丢失问题。提出基于卡尔曼滤波的低质量指纹证据图像增强方法。根据图像高斯噪声分布函数, 选用 VisuShrink 阈值降噪, 构建卡尔曼滤波器, 完成指纹图像预处理。将原始图像转化为单像素细化图, 提取滤波后指纹图像的奇异点与特征点。划分指纹图像, 根据奇异点与方向向量确定指纹的方向场, 进行图像增强滤波处理, 完成基于卡尔曼滤波的低质量指纹证据图像增强方法设计。构建实验环节, 验证该方法应用效果, 实验方法表明: 该方法使用后, 低质量指纹图像变得更加清晰, 并且平均错误率最高仅为 6.21%, 评估分数均值达到了 0.85, 该方法有效提高了图像的增强效果。

关键词: 卡尔曼滤波; 指纹增强; 图像增强; 方向场; 特征点提取; 图像预处理

0 引言

生物识别技术是融合计算机、智能识别以及传感器等的综合性技术。在科技高速发展的今天, 为人们的日常生产与生活提供了大量的便利^[1-2]。通过生物识别技术, 确定人体的关键信息, 有助于确定人物身份。当前, 在大量的案件侦破与人体信息识别过程中, 指纹识别是其中较为重要的项目之一。由于指纹采集设备的不同以及留下指纹物体材质的差异, 会产生大量的低质量指纹。所谓的低质量指纹, 是指存在异常或是缺失的指纹片段, 其对指纹识别技术的应用造成了恶劣的影响。只有解决低质量指纹所造成的问题, 才能实现生物识别技术的识别率整体提升。因此, 低质量指纹处理与增强问题是一个亟需解决的问题。

低质量指纹大多具有信息可信度不足以及有效信息较少的特征, 针对此问题, 文献[3]针对指纹图像的特点, 以阈值函数为基础, 结合小波变换实现指纹图像增强。但是该方法存在关键信息点错误率较高的问题。文献[4]提出基

于滤波和深度玻尔兹曼机重构的指纹增强方法, 对指纹图像进行滤波处理, 根据处理结果, 通过深度玻尔兹曼机重构方法对处理后指纹图像进行增强。但是该方法存在评估分数较低的问题。文献[5]提出基于各向异性扩散和冲击滤波的指纹图像增强方法, 结合冲击滤波算法对采集的指纹图像进行滤波处理, 以处理结果为基础, 通过各向异性扩散进行指纹增强。但是该方法也存在图像增强平均错误率较高的问题。

以上方法存在问题的主要原因是: 当前基于模板匹配的图像增强算法针对低质量指纹证据图像匹配计算复杂度较高的情况, 以固定阈值完成原始图像滤波处理, 未考虑单像素细化方向场特征丢失问题。针对上述情况, 本文在对比多种图像处理方法后, 选择卡尔曼滤波技术对传统方法进行优化, 设计出一种基于卡尔曼滤波的低质量指纹证据图像增强方法。针对图像增强方法中的图像处理环节进行优化, 以期在低质量指纹图像的处理根源优化图像质量, 提升图像增强效果, 获取高质量指纹图像。

收稿日期: 2023-10-23 修稿日期: 2024-02-01

作者简介: *通信作者: 卢超超(1986—), 女, 山西祁县人, 硕士, 讲师, 主要研究方向为侦查学, E-mail: a13393433201@

163.com

1 卡尔曼滤波的指纹证据图像增强方法

对当前低质量指纹证据图像增强方法各环节进行分析后,将本次研究提出的指纹图像基础环节设定如图1所示。

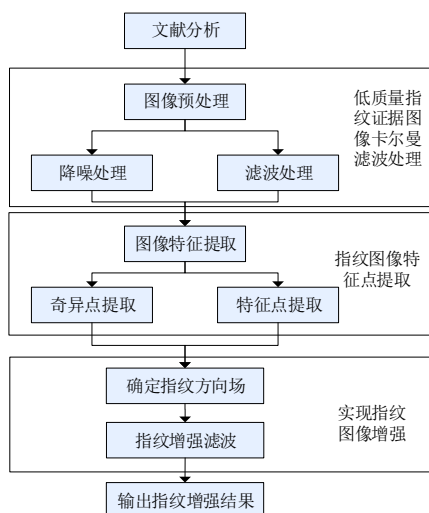


图1 基于卡尔曼滤波的低质量指纹证据图像增强过程

如图1所示,本次研究将指纹证据图像增强过程整体划分为三个环节。

(1) 低质量指纹证据图像卡尔曼滤波处理。采用 VisuShrink 阈值进行降噪处理,然后使用卡尔曼滤波技术对降噪后的图像进行去噪处理,降低噪声对图像的影响,从而提高图像质量。

(2) 指纹图像特征点提取。将处理后的指纹图像转化为单像素细化图,提取滤波后的指纹图像的特征点。

(3) 实现指纹图像增强。以提取的特征点为基础,通过变形计算得到指纹图像的方向场,平滑处理获取到方向场数据,得到指纹方向信息,并对方向信息进行图像增强滤波处理。

按照图1所示内容完成基于卡尔曼滤波的低质量指纹证据图像增强方法设计。

1.1 低质量指纹证据图像卡尔曼滤波处理

通过分析大量的低质量指纹图像增强案例可以发现,在图像的预处理过程中,并没有较好地完成图像滤波过程,导致图像增强过程从根源处得不到重视。故而,本研究采集到低质量指纹证据图像后,首先对其进行降噪处理,为后续的指纹图像处理提供基础。在对比多种方法后,选择 VisuShrink 阈值^[6-7]完成降噪过程。已知

在图像处理的过程中会产生大量的高斯噪声。那么,指纹图像的高斯噪声密度函数可表示为

$$g(b) = \frac{1}{\sqrt{2\pi}\alpha} r^{-\frac{(b-\beta)^2}{2\alpha^2}} \quad (1)$$

式(1)中: b 表示原始图像灰度值; β 表示灰度值期望值; α 表示灰度值的标准差, r 表示灰度系数。根据公式(1)得到高斯噪声密度函数后,针对多维变量,构建 VisuShrink 阈值函数,则有:

$$T = \alpha \sqrt{2 \ln N} \quad (2)$$

式(2)中: N 表示信号长度。通过上述公式完成降噪处理后,使用卡尔曼滤波技术对其进行去噪处理,彻底去除噪声对图像的影响。构建卡尔曼滤波器,将图像中未知的噪声变量 v 的条件概率分布设定为

$$p(v|y) = \frac{p(y|v)p(v)}{p(y)} \quad (3)$$

式(3)中: $p(v|y)$ 表示此噪声条件的后验概率分布函数; $p(v)$ 表示此噪声条件的先验概率分布函数; $p(y)$ 为图像的先验条件。根据公式(3),结合公式(1)和公式(2)得到卡尔曼滤波器基础模型^[8],即获得卡尔曼滤波处理后的指纹证据图像,公式为

$$Y(v_i, y_i) = \left\{ \frac{p(v|y) f_i(v_{i-1}, d_i)}{\emptyset g(b)}, \frac{p(v|y) u_i(y_{i-1}, e_i)}{g(b)} \right\} \quad (4)$$

$$g(b) < T$$

式(4)中: v_i, y_i 表示图像的状态向量; f_i 表示图像状态模型; u_i 表示图像测量模型; d_i 表示图像处理过程噪声; e_i 表示图像测量噪声,此两种噪声均属于高斯白噪声; $\emptyset$ 表示滤波处理转换系数。当公式(4)中两部分函数均为线性函数时,此滤波器的使用效果最为理想。使用此卡尔曼滤波器基础模型对降噪处理后的图像进行滤波去噪处理,完成图像整体的预处理过程,如图2所示。

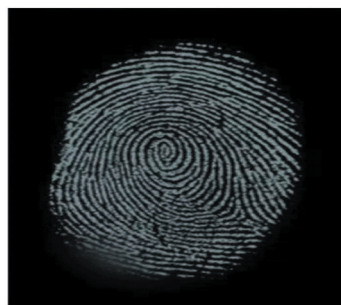


图2 滤波去噪处理后的指纹图像

1.2 指纹图像特征点提取

对采集到的指纹图像进行卡尔曼滤波处理后，将原始低质量指纹图像转化为单像素细化图^[9-10]，以该图像为基础，提取滤波后的指纹图像特征点。由于指纹图像存在方向场，在特征点提取的过程中需要对图像中的奇异点进行分折，以保证指纹特征点的获取精度。指纹奇异点可以表示为

$$\wp(v_i, y_i) = \frac{\int_0^{2\pi} KY(v_i + x \cos \delta, y_i + x \cos \delta) d\delta}{2\pi} \quad (5)$$

式(5)中： δ 表示指纹图像角度； d 表示指纹图像中的距离； K 表示计算系数。在图像中如存在距离较近的奇异点，可采用相邻奇异点的计算均值代替。

确定奇异点计算公式后，将指纹图像划分为多个图像区块，分别计算图像区块的信息熵^[11]以及方向场密度，得到每个图像区块中的奇异点个数，从而得到整幅图像的奇异点个数。此时，图像区块信息熵计算公式可表示如下：

$$G_i = \frac{N_i(\wp(v_i, y_i))}{P*O} \quad (6)$$

式(6)中： N 表示指纹方向场落在各个图像区块的段落； $P*O$ 表示图像区块的尺寸。根据公式(6)提取指纹图像特征点，公式为

$$W_i = -\sum G_i \lg G_i \quad (7)$$

通过公式(7)确定不同区块图像的代表信息变化程度，根据此程度判断奇异点的分布情况，确定奇异点的主要分布区域，将此区域特征作为指纹图像特征点。提取的指纹图像特征点如图3所示。

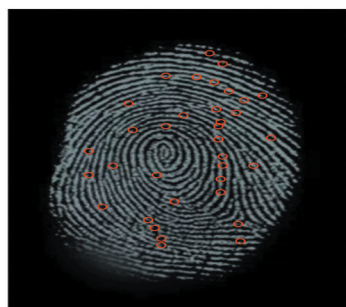


图3 指纹图像特征点提取结果

1.3 实现指纹图像增强

上文中完成的卡尔曼滤波处理低质量指纹

证据图像和指纹图像特征点提取均为指纹图像增强准备环节，在此基础上，优化指纹图像增强环节，实现指纹图像增强效果的整体提升。在本研究中，首先补偿指纹图像的方向场，设定图像的方向能量为 $B(Z)$ ，使用其对图像中的任意一个区块的方向强度进行衡量，则有：

$$B(\varpi) = [\Im_n^\varpi]^2 + [\Re_n^\varpi]^2 \quad (8)$$

式(8)中： $\Im_n^\varpi$ 表示高斯函数的相关导数； $\Re_n^\varpi$ 表示Hilbert变换系数^[12-13]； ϖ 表示图像方向能量因子。将上述函数转化为傅立叶技术的三角函数形式，计算得到图像中指纹的主要方向。确定主要方向后，通过变形计算得到指纹图像的方向场：

$$Z = \text{mod} \left(\frac{\arg(B(\varpi))}{2} + \pi, \pi \right) - \frac{\pi}{2} \quad (9)$$

式(9)中： mod 表示取余运算； $\arg$ 表示复数求值计算中的辐角。而后，对获取到的方向场数据进行平滑处理，得到指纹方向信息，并对此方向信息进行滤波处理。经过平滑处理后指纹方向信息所含噪声较少，且纹理结构清晰，在此图像上进行增强处理可得到可信度较高的信息。在前面的图像处理过程完成后，进行本环节的核心内容增强滤波处理^[14-15]，得到低质量指纹图像增强处理后的图像，公式为

$$P = \varepsilon F \left(\cos^2(Z) W_{ixx} + 2 \sin(Z) W_{ixy} + \sin^2(Z) W_{iyy} \right) \quad (10)$$

式(10)中： W_{ixx} 表示指纹横向特征； W_{iyy} 表示指纹纵向特征； W_{ixy} 表示指纹曲线特征； F 表示频率； ε 表示增强处理因子。为了更好地完成增强处理过程，对图像的滤波频率展开设定，则有：

$$F = \frac{1}{\pi f'} \quad (11)$$

式(11)中： f' 表示初始频率。

使用公式(11)对图像的增强滤波频率加以控制，在保证增强效果的同时，提升指纹图像的处理速度。在此环节设定完成后，将其与上文内容1.1和1.2节进行有序连接，并将其与当前方法中的计算部分融合处理，整合为闭环算法。至此，基于卡尔曼滤波的低质量指纹证据图像增强方法设计完成。增强后的指纹图像如图4所示。

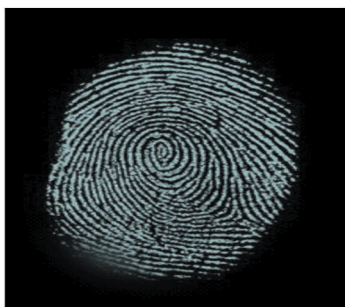


图4 增强后的指纹图像



图5 实验主观评估指纹图像

2 实验论证与结果分析

本研究提出了一种基于卡尔曼滤波的低质量指纹证据图像增强方法,为进一步将其与现有的指纹增强方法进行比较,以验证文中所提方法对指纹图像增强的有效性,构建实验环节,对文献[4]方法、文献[5]方法与本文方法的应用效果进行对比分析。

2.1 实验数据

在本文实验中,为进一步验证文中方法的应用效果,实验中使用两个FVC数据库作为实验数据样本来源。FVC数据库1使用光学传感器完成指纹采集工作,FVC数据库2使用电容传感器进行指纹采集。在指纹采集结束后,保证每个数据库中含有500个400 DPI的指纹图像。将上述两个指纹库中的指纹图像整合到实验数据库中,对原始指纹进行筛选,其中可作为实验数据的指纹图像共计800个,将其划分为10个子集:ZY1,ZY2,ZY3,⋯,ZY10。其中,ZY1,ZY2,ZY3,⋯,ZY9均为相关设备采集得到的指纹,ZY10为假指纹图像。为降低实验难度,将指纹图像均整合为1080*1080分辨率。为提升实验的有效性,在此数据库中选择一幅指纹图像作为指纹增强方法主观评估图像,具体图像选择结果如图5所示。

首先对此图像进行增强处理,并对增强结果进行主观分析,对文中方法应用效果进行初步分析。其余数据库中的实验图像作为本次实验中的样本集,为后续的应用效果客观评估实验过程提供数据来源。

2.2 实验环境

为测试所设计方法的性能效果,将对本文方法进行测试。为了验证本文方法有效性,整个实验都需要在性能强大的服务器上完成,服务器的硬件环境为:Inter(R)Core(TM)i7-7700HQ、CPU@3.8 GHz的服务器、5TG的内存、NVIDIA RTX 2080Ti 12 GB的显卡。服务器的软件环境为:Python 4.8的代码编写软件平台、Windows10操作系统。

2.3 实验方案

在对大量的文献以及实验进行分析后,应用预设的指纹图像数据库,构建本次实验的实验方案。本次实验过程中,将指纹证据图像增强方法应用效果的验证过程大致分为两个环节,具体设定如下。

2.3.1 主观评估过程

使用文中方法与文献[4]提出的基于滤波和深度玻尔兹曼机重构的指纹增强方法和文献[5]提出的基于各向异性扩散和冲击滤波的指纹图像增强方法,预先选定的当前指纹增强方法对图像进行增强处理,在获取图像增强结果后,对图像进行肉眼评析,初步确定不同方法的使用效果。将该评析作为增强方法效果的评定内容之一。

2.3.2 客观评估过程

为对文中方法的应用效果进行更加全面的分析,使用文中方法与文献[4]和文献[5]的方法对数据库中的图像进行整体处理,并对处理结果的整体增强效果与局部增强效果进行对比

分析。在此部分实验过程中，将实验图像在实验中成功匹配的指纹点数量作为 a ，匹配失败的指纹点数量作为 c ，匹配错误的指纹点数量作为 s ，得到图像增强的评估分数：

$$Q = \frac{\sum_{i=1}^h k_i [a_i - c_i - s_i]}{\sum_{i=1}^h k_i t_i} \quad (12)$$

式(12)中： h 表示图像分割块数； k_i 表示图像原始质量分析结果。使用此公式对不同方法的局部增强效果进行分析。与此同时，将图像整体增强效果通过平均错误率指标进行研究，具体指标计算过程设定如下：



图6 指纹图像增强主观评估结果

对图6进行分析可知，本文方法与文献[4]方法和文献[5]方法在图像增强后所得结果存在细节上的差别。文献中的两种方法存在严重的块效应，对于图像中的脊线断裂问题修复效果较差，在增强的过程中容易引入虚假的指纹特征。其中，文献[5]方法增强效果不够明显，对于脊线的修复过程存在错误，导致真正的特征点丢失。本文方法可对图像进行明显的改进，弥补指纹中的断裂情况，提升指纹图像质量。通过肉眼分析可初步确定，本文方法的指纹增强效果较为理想。

2.4.2 客观评估结果分析

在主观评估不同方法的指纹图像增强效果后，客观评估图像增强效果，以公式(12)和公式(13)获取的数据为评估指标，实验结果如图7所示。

$$\bar{f} = \frac{l_1 + l_2 + l_3}{n} * 100\% \quad (13)$$

式(13)中： l_1 表示增加指纹特征点错误率； l_2 表示遗失指纹特征点错误率； l_3 表示错误指纹特征点错误率； n 表示计算项目个数。使用上述两项指标对不同方法的指纹增强结果进行计算，并比对不同方法的指标计算结果确定其应用效果。

2.4 实验结果分析

2.4.1 主观评估结果分析

应用三种方法增强处理指纹图像，获取增强后的图像，主观评估图像质量，实验结果如图6所示。

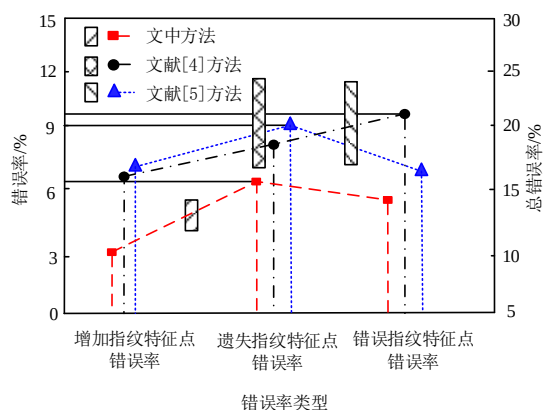


图7 实验方法图像增强平均错误率计算结果

图7中的柱形图为总错误率。对图7中的数据展开分析可以看出，文献[4]和文献[5]的指纹图像增强方法在整体实验图像库处理的过程中出现了错误率异常增加的问题，这两种方法的

总错误率均值分别达到了24.89%和24.00%，此实验结果从侧面证实了两种方法的指纹识别与处理性能相对较低。而本文方法的总错误率仅为14.88%，三种方法相比可知，本文方法的总错误率与另外两种方法相比，降低了9个百分点以上，并且该方法的单项最高平均错误率仅为6.21%。其对此两类方法进行系统分析可以发现，文献方法在使用过程中出现指纹关键点特征丢失或虚假的问题，而本文方法在指纹图像增强处理的过程中，多次滤波图像，在一定程度上可以提升图像的处理效果，从而降低了错误率。

在上述实验的基础上，分析不同方法的图像增强评估分数，对比三种方法的分数，结果如图8所示。

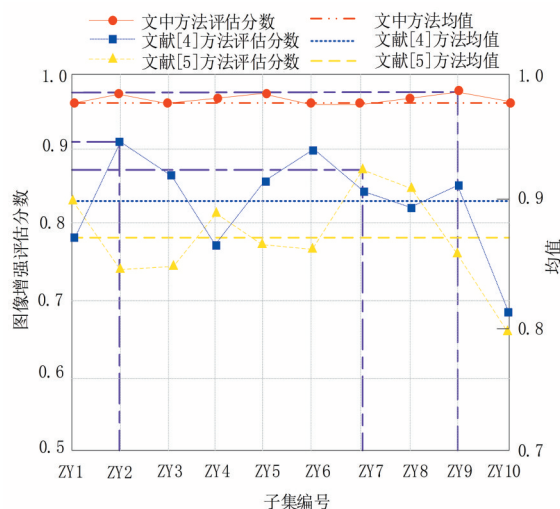


图8 实验方法图像增强评估分数对比

图像增强评估分数的取值结果越高，说明图像增强效果越好，根据该原理，结合图8中的数据可知，本文方法的图像增强评估分数的均值达到了0.85，而文献[4]和文献[5]方法的评估分数均值分别为0.63和0.77。三种方法相比，本文方法的图像增强评估分数最高，其比文献方法最高高出0.08，说明该方法增强处理后的图像的增强效果优于文献方法，并且本文方法的评估分数最高达到了0.95，而文献方法的分数最高仅为0.84，本文方法比文献方法的最高值高了0.11，该实验结果也说明本文方法具备更好的指纹图像增强效果。

在实验方法图像增强平均错误率实验结果与实验方法图像增强评估分数实验结果分析完成后，整合上述分析结果可知，本文方法无论在图像增强平均错误率，还是在图像增强评估分数方面，均取得了较好的结果，表明该方法在增强处理低质量指纹图像时，提高了图像增强效果。与此同时，将此分析结果与指纹图像增强主观评估结果一起分析可知，进一步确定了本文方法应用效果的优异性。

3 结语

针对当前低质量指纹证据图像增强过程中存在的图像增强评估分数低、增强效果差等问题，应用卡尔曼滤波技术，研发出一种新的图像增强方法。该方法以卡尔曼滤波为基础，滤波处理了低质量指纹图像，提高图像质量，结合了信息熵，提取指纹图像特征点，通过方向场和图像的滤波频率，实现低质量指纹图像增强处理。在方法设计完成后，对其进行应用测试确定其使用效果，经实验证实该方法的低质量指纹图像的增强效果优于文献方法，并且图像增强总错误率仅为14.88%，图像增强评估分数均值达到了0.85，均高于文献方法，说明该方法具有较高的图像增强效果。

参考文献：

- [1] 徐武,杨昊东,汤弘毅,等. 基于改进加权引导滤波的红外图像增强算法[J]. 激光杂志, 2022, 43(1): 101-106.
- [2] 潘晓英,魏苗,王昊,等. 多尺度融合残差编解码器的低照度图像增强方法[J]. 计算机辅助设计与图形学报, 2022, 34(1): 104-112.
- [3] 张兆茹,邓彩霞,岳欣华. 改进阈值函数的指纹图像去噪方法[J]. 哈尔滨理工大学学报, 2022, 27(1): 55-60.
- [4] 卞维新,丁世飞,张楠,等. 结合滤波和深度玻尔兹曼机重构的指纹增强[J]. 软件学报, 2019, 30(6): 1886-1900.
- [5] 王丹,张小波. 基于各向异性扩散和冲击滤波的指纹图像增强算法[J]. 智能计算机与应用, 2019, 9(5): 320-322, 325.
- [6] 丛爽,张坤. 基于卡尔曼滤波的在线量子态估计优化算法[J]. 系统工程与电子技术, 2021, 43(6): 1636-1643.

- [7] 刘明珠, 苏桃宇, 王依宁. 压扩多尺度Retinex图像增强算法研究[J]. 哈尔滨理工大学学报, 2020, 25(5): 93-99.
- [8] 常路宾. 基于SE(3)的扩展卡尔曼滤波姿态估计算法(英文)[J]. 中国惯性技术学报, 2020, 28(4): 499-504, 550.
- [9] 陈忠辉, 唐晓颖, 林志权, 等. 畸变图像自适应校正及质量增强算法研究与实现[J]. 计算机应用, 2020, 40(S1): 180-184.
- [10] 蒋啸虎, 佟金, 马云海, 等. 基于卡尔曼滤波融合算法的深松耕深检测装置研究[J]. 农业机械学报, 2020, 51(9): 53-60.
- [11] 贾镕, 王峰, 袁宏武, 等. 潜指纹紫外偏振图像模糊自适应融合算法研究[J]. 计算机工程, 2020, 46(11): 267-272, 278.
- [12] 田会娟, 蔡敏鹏, 关涛, 等. 基于YCbCr颜色空间的Retinex低照度图像增强方法研究[J]. 光子学报, 2020, 49(2): 173-184.
- [13] 张韬, 宗和厚, 陈华, 等. 高聚物黏结炸药CT图像像素级裂纹识别方法研究[J]. 含能材料, 2020, 28(5): 442-448.
- [14] 吴蔚, 谭献海, 钱晓群, 等. 基于卡尔曼滤波的室内定位可信指纹库研究与实现[J]. 计算机应用研究, 2020, 37(6): 1835-1838.
- [15] 陈文燕, 范文博, 杨钧宇. 基于密集连接卷积网络的小面积指纹识别方法[J]. 计算机工程与应用, 2020, 56(10): 134-140.

Design of an enhancement algorithm for low quality fingerprint evidence images

Lu Chaochao*

(Department of Investigation, Shanxi Police College, Taiyuan 030401, China)

Abstract: During the process of fingerprint image acquisition, there are inherent differences such as rotation, translation, and incomplete image acquisition. The number of fine nodes extracted each time is not fixed, and this feature form has a high computational complexity for later matching. The current image enhancement algorithms based on template matching aim to address the high computational complexity of low-quality fingerprint evidence image matching, and complete the original image filtering process with a fixed threshold without considering the issue of single pixel thinning direction field feature loss. Propose a low quality fingerprint evidence image enhancement method based on Kalman filtering. Based on the Gaussian noise distribution function of the image, VisuShrink threshold denoising is selected to construct a Kalman filter and complete fingerprint image preprocessing. Convert the original image into a single pixel refined image, and extract the singular points and feature points of the filtered fingerprint image. Divide fingerprint images, determine the direction field of the fingerprint based on singular points and direction vectors, perform image enhancement filtering processing, and complete the design of a low-quality fingerprint evidence image enhancement method based on Kalman filtering. Construct an experimental section to verify the application effect of this method. The experimental method shows that after using this method, low-quality fingerprint images become clearer, and the average error rate is only 6.21%, with an average evaluation score of 0.85. This method effectively improves the image enhancement effect.

Keywords: Kalman filter; fingerprint enhancement; image enhancement; direction field; feature point extraction; image preprocessing

文章编号: 1007-1423(2024)07-0024-07

DOI: 10.3969/j.issn.1007-1423.2024.07.004

基于遗传算法和 SMOTE 的网络入侵检测模型优化研究

戴周浩*

(贵州财经大学信息学院, 贵阳 550000)

摘要: 针对网络入侵检测存在数据不平衡和特征冗余的问题, 提出一种新的检测模型。该模型结合了遗传算法和 SMOTE 算法, 通过对数据进行采样和特征选择, 提高了网络入侵检测的准确性。首先, 为了解决数据不平衡的状况, 采用了 SMOTE 算法。这个算法通过在入侵类样本中嵌入随机样本, 有效地提高了入侵类样本的数量, 使得数据达到平衡。其次, 为了缓解特征冗余, 引入了基于遗传算法和随机森林方法的包装式特征选择技术, 选择有用特征, 减少冗余信息, 从而提升最终的入侵检测性能。最后, 采用随机森林算法对经过上述处理的数据集进行分类, 实现对网络入侵样本的有效检测。在 NSL-KDD 数据集上的实验表明, 基于遗传算法和 SMOTE 的网络入侵检测模型从整体上提高了入侵检测的识别率。

关键词: 特征选择; SMOTE 过采样; 随机森林; 网络入侵检测; 遗传算法

0 引言

随着互联网的普及和人们对其依赖程度的不断提高, 网络安全问题日益凸显, 其中网络入侵行为的威胁成为亟待解决的焦点。网络入侵检测在保障网络安全方面起着关键作用, 近年来机器学习方法在网络入侵检测领域得到了广泛应用。然而, 由于网络流量数据的高维度和不平衡分布等挑战, 传统的机器学习算法在网络入侵检测中面临着困境。

在这一背景下, 为了提高网络入侵检测的效果, 研究者们纷纷探索新的方法。在已有的研究中, 朱琨等^[1]系统介绍了传统机器学习算法在网络入侵检测中的应用, 对各分类算法的优缺点进行了比较。为了应对训练集不平衡的问题, 研究者们采用了数据采样的方法^[2]。文献[3]详细介绍了利用合成少数过采样技术(SMOTE)来提高不平衡类型的样本数量, 以此来降低测试模型对不平衡类型假阳性的比率。另外, 有些研究人员, 如封化民等^[4]提出了一个新的入侵检测模型, 它结合 SMOTE 与梯度升

级决策树(GBDT)的技术。但是, 单纯的过度采样可能会导致数据的规模扩大。因此, 为了减小网络侵入测试数据的规模, 也有研究人员提出了使用数据降维的策略^[5]。

通过融合特征选择方法和机器学习的策略, 能够有效地应对特征冗余的问题, 减少网络侵入检测的误判, 并提升系统的执行效率^[6]。而方彬皓等^[7]则推出了一种结合了多元特征选择和混合采样的 CatBoost 网络入侵检测模型, 优化了特征选择, 缓解了特征冗余及类别不均衡的问题。另外, Song 等^[8]提出了采用信息论模型进行侵入检测特征抽取的策略, 有效地提高了检测的准确性。Ren 等^[9]利用 K 近邻模型找出并剔除异常数据, 显著地提升了网络入侵检测的准确度。Wang 等^[10]结合人工神经网络和遗传算法的网络入侵检测方法, 通过降低特征维度实现了较好的检测性能。以上研究已经在 NSL-KDD 数据系列上, 利用搜索算法、随机森林以及极限学习机等多种机器学习方法, 进行了实证测试。

鉴于上述分析, 本文采用了随机森林(random forest, RF)的方法, 并与多目标遗传算法

收稿日期: 2023-12-19 修稿日期: 2024-01-23

作者简介: *通信作者: 戴周浩(1998—), 男, 浙江温州人, 硕士研究生, 研究方向为最优化理论, E-mail: z0815564@

163.com

NSGA-II融合，构建出一种包装型特征选择算法。然后，运用SMOTE的采样方法来处理经过降维的特征子集，以实现数据集的平衡。本算法充分发挥了NSGA-II在多目标优化上的优秀性能，并结合SMOTE的采样技巧，成功地解决了特征冗余和类别不平衡的问题，从而获得了出色的分类效果。在网络入侵检测研究领域，该算法将为提高系统准确率和鲁棒性提供新的思路。

1 融合封装式特征选择和SMOTE采样的入侵检测模型

本模型综合运用了封装式特征选择算法、SMOTE采样技术和随机森林算法，以应对机器学习在网络入侵检测中常见的问题，如特征冗余和类别不平衡。入侵检测模型的设计包括三个主要组成部分，即封装式特征选择模块、SMOTE采样模块以及入侵检测模块。

首先，通过将随机森林(RF)与多目标遗传算法NSGA-II相融合，创造性地构建了一种封装特征选择算法。该算法旨在挑选出对网络入侵检测最具信息量的特征，减少特征冗余，提高模型的效率和准确性。其次，SMOTE采样模块利用合成少数过采样技术，在数据集中引入新的合成样本，以平衡不同类别之间的样本分布，解决类别不平衡问题，提高模型对入侵类样本的识别能力。最后，入侵检测模块采用随机森林算法，通过构建多个决策树的集成模型，实现对网络入侵的有效检测。三个模块协同工作，构成了一个综合的入侵检测模型，旨在克服传统机器学习方法在处理网络入侵检测问题时所面临的挑战，为提高准确性和性能提供了全面而有效的解决方案。

1.1 封装特征选择算法

图1是封装式特征选择算法的流程。这一算法是通过融合多目标遗传算法NSGA-II和随机森林，使得封装式特征选择的效率得以提高。

NSGA-II算法的运作和两个核心参数有着紧密的关联，即 x_{rank} 和 x_{dist} ，它们的主要用途是对解 x 的品质进行评估。 x_{rank} 是表示解 x 在帕累托前沿排名的参数，其计算方式是使用快速非支配排序法。 x_{dist} 则是反映了解 x 的密集程度，它

是通过计算拥挤度来对解在帕累托前沿上的分散程度作出判断。拥挤度指标的引入有助于确保帕累托前沿上的解呈现均匀分布，以防止过度集中在某一特定区域。解 x 的拥挤度计算公式如下：

$$x_{\text{dist}} = \sum_{k=1}^m \frac{t_k(x+1) - t_k(x-1)}{t_k^{\text{max}} - t_k^{\text{min}}} \quad (1)$$

式中：目标函数的数量用 m 来表示，在本文特征选择问题中， m 的数值被设为2。 $t_k(x)$ 表示解 i 在第 k 个目标任务中的确切数值。 t_k^{max} 和 t_k^{min} 则代表了第 k 个目标函数的极大值和极小值。

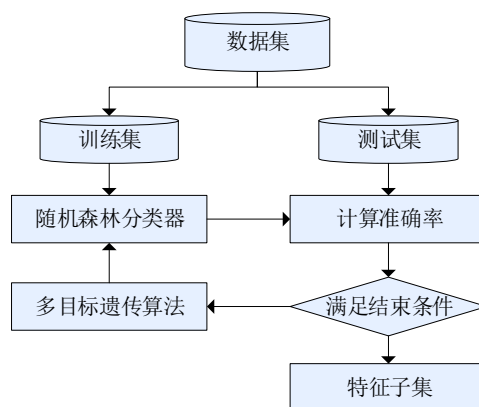


图1 基于多目标遗传算法和随机森林的特征选择算法

选择算子在NSGA-II中是依据参数 x_{rank} 和 x_{dist} 来进行实现的，其标准是更倾向于挑选在前沿中更高的解。但若两个解答都是同一前沿，那么就会优先选择拥挤度相对较低的那个解。这一选择策略确保了在不同前沿中均匀地选择解，并通过保留拥有最优帕累托前沿的解来维持算法的多样性。在NSGA-II算法的第 t 次迭代过程中，通过遗传算子处理种群 P_t 形成后代种群 Q_t ，接着把 P_t 和 Q_t 融合为一个 $2N$ 大小的种群 R_t 。用非支配排序策略识别出 R_t 中的各个前沿，并选用精英机制把最优的帕累托前沿解留到接下来的迭代。这一过程旨在确保算法在演化的过程中保持多样性，同时保留最具竞争力的解。

特征选择的目标是通过NSGA-II来寻找多目标优化问题的帕累托前沿，并采用RF算法来评估特征的优良程度，以实现以最少的特征获取最高的分类准确率。

算法1 多目标遗传算法 NSGA-II 和随机森林组合封装式特征选择算法

输入：种群个体数量 N ，特征数量 n ，交叉率 pc ，变异算子采用概率 pm ，最大迭代次数 G

输出：特征子集。

步骤1：随机初始化 N 个染色体的种群 P_t ，每个染色体表示一个潜在的特征子集，假设总特征数量为 n ，特征子集为 $X = \{x_1, x_2, \dots, x_n\}$ ，其中 $x_i \in \{0, 1\}$ ， $1 \leq i \leq n$ 。 $x_i = 1$ 表示特征 i 被选择，否则被取消。

步骤2：通过 RF 评估初始化种群的染色体，将准确率和特征数量作为两个目标函数。将 RF 评估过的每个子集保存于列表 L ，避免重复评估相同的特征子集。

步骤3：采用非支配排序方法将初始化种群排序，经过遗传算子（选择算子、交叉算子及变异算子）处理产生后代种群 Q_t 。

步骤4：根据非支配排列程序的排序结果直接生成第1个后代种群，选择算子选择拥挤度低的染色体，交叉算子采用概率为 pc 的单点交叉，变异算子采用概率为 pm 的随机变异。

步骤5：将父种群 P_t 和后代种群 Q_t 合并获得 R_t ，对种群 R_t 进行非支配排序处理和拥挤度排序处理，选出排名前一半的染色体建立下一代种群。

步骤6：获得特征子集

1.2 SMOTE 采样算法

SMOTE 算法是一种被广泛使用的数据扩充技术，是 Chawla 为解决不平衡数据问题提出的。该算法是在少数类样本之间进行随机的线性插值，产生大量的合成样本。通过这种方式，有效地增加了少数类别样本的数量，从而实现了数据集的平衡。SMOTE 算法的具体步骤如下：

Step1 对于每个少数类别样本 x_i ($i = 1, 2, \dots, n$)，计算其与其他样本的距离，得到样本 x_i 的 k 个近邻。

Step2 选用 m 个样本（从 k 个近邻样本中随选）并用 x_{ij} ($j = 1, 2, \dots, n$) 来标识。

Step3 通过在 x_i 与 x_{ij} 之间进行随机的线性插值构造新的样本。插值公式为

$$x_{\text{new}} = x_i + \text{rand}(0, 1)(x_{ij} - x_i) \quad (2)$$

式中： x_{new} 表示人为构造的样本， $\text{rand}[0, 1]$ 是区间 $[0, 1]$ 上满足均匀分布的随机数。

1.3 基于随机森林的检测分类算法

作为集成学习的一种方法，随机森林能够

通过创建和结合多棵决策树的结果，从而提升整体模型的性能和健壮性。基于随机森林的入侵检测建立需要三个步骤。

在本文中，通过有放回地提取 N 样本，进而创建了一个新的数据集 D_i ，其中 N 代表了初始数据集的大小。将这一抽样过程重复执行 B 次，得到 B 个包含 N 个样本的 bootstrap 样本集合，表示为 $\{D_1, D_2, \dots, D_B\}$ 。其次，本文通过运用分类回归树 (CART) 算法来生成 n 个决策树。在一棵决策树的每一个节点上，从所有特征中随机挑选 m 个特征加以分析，其中 m 是一个比总特征数要小的数字。最后，通过投票机制将生成的决策树组合成随机森林。

$$F(x) = \frac{1}{B} \sum_{i=1}^B f_i(x) \quad (3)$$

式中： $F(x)$ 是随机森林的最终预测， $f_i(x)$ 是第 i 个决策树的预测， B 是随机森林中决策树的数量。

1.4 融合封装式特征选择和 SMOTE 采样入侵检测模型

在整个入侵检测模型处理过程中，首先对 NSL-KDD 数据集进行分类，从而形成了训练集以及测试集。对这两个集合进行了数值归一化和字符型特征数值化操作，得到了训练集 M_1 和测试集 M_2 。随后，通过封装式特征选择对 M_1 和 M_2 进行了处理。采用多目标遗传算法和基于随机森林的特征评价方法，循环执行特征选择过程，最终获得了经过优化的特征子集 F 。基于得到的特征子集 F ，对原始训练集和测试集进行了进一步的处理，得到了优化后的训练集 FM_1 和测试集 FM_2 。

为确保实验的有效性，本文仅对训练集 FM_1 执行 SMOTE 过采样操作，生成了进一步优化的训练集 SFM 。这一系列处理步骤旨在提升入侵检测模型的性能和鲁棒性，使其能够更好地适应实际数据集的特点。最后，使用 SFM 进行随机森林分类器的训练，并利用测试集 FM_2 验证分类效果，评估分类准确率、召回率、精确率和 $F1_score$ 值等指标。这一系列步骤构成了本文提出的综合模型的处理流程，从数据预处理、特征选择、过采样到最终的分类器训练和性能评估，形成了一个完整而有效的入侵检测系统。

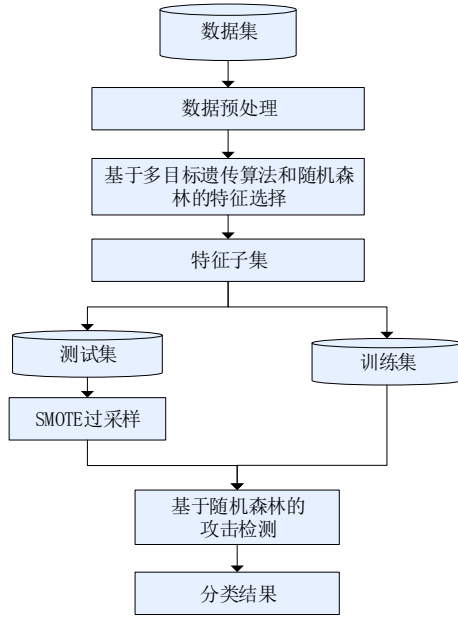


图2 网络入侵检测模型流程图

2 实验

2.1 数据集

作为KDD-99的扩展数据集，NSL-KDD含有125973个训练样例和22543个测试样例。这个数据集覆盖了41个特质和5个类别，其中一类代表的是网络的正常流动(Normal)，其余四类就象征着各种不同形式的攻击流量，分别是DoS、Probe、R2L和U2R。图3详细呈现了NLS-KDD训练组(KDDTrain+)和测试组(KDDTest+)中每一类样例的数量统计，数据样例的分布情况非常不平衡，尤其在训练组里，R2L和U2R这两类的样例数量仅占总数的0.790%和0.041%。

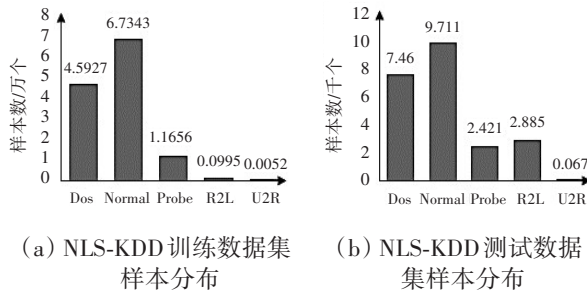


图3 NLS-KDD样本分布

2.2 数据预处理

NSL-KDD数据集包含有41个特征，其中有3个非数值型特征，分别是protocol Type、service、flag。由于机器学习算法无法直接处理非数值型数据，因此本文采用了独热(one-hot)编码的方法，将这些非数值型特征转换为数值表示。此外，部分数据的数值范围存在差异，这可能会影响后面模型训练的效率。为了避免这种情况，本文引入了标准化数据的策略。对于每个特征 x ，使用以下公式进行归一化处理：

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (4)$$

式中： x 代表转化后的数据，而 x' 则代表经过归一化处理的数据。 $x_{\max}$ 和 $x_{\min}$ 则代表着特征值的最大值和最小值。

2.3 评价指标

通过分类的准确率来衡量入侵检测的准确性，同时使用精度、召回率、F1_score评分来评估入侵检测算法的整体性能。下面将介绍所需的分类性能指标。

精度的计算式为

$$Precision = \frac{\sum_{i=1}^l TP_i}{\sum_{i=1}^l (TP_i + FP_i)} \quad (5)$$

式中： TP_i 为真正阳性样本， FP_i 为误报阳性样本。

召回率的计算式为

$$Recall = \frac{\sum_{i=1}^l TP_i}{\sum_{i=1}^l (TP_i + FN_i)} \quad (6)$$

式中： FN_i 为假阴性样本。

F1_score的计算式为

$$F1_score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (7)$$

2.4 实验方法设计

本文针对传统机器学习算法在算法优化中未充分考虑特征冗余和类别不平衡问题的缺陷，提出了一种综合解决方案，结合了封装式特征选择算法和SMOTE过采样算法。在本节中，首先展示了封装式特征选择算法的执行结果，并详细探讨了其对分类器性能的影响。然后，对

SMOTE过采样算法执行前后样本分布情况进行了对比分析,同时研究了SMOTE过采样算法对分类器性能的影响,对实验成果进行了深入分析。最后,将本研究提出的模型与国内外一些相应模型进行了比较。实验环境采用了一台PC机,配备Intel Core i5-7200U处理器,运行Windows 10操作系统,使用Python编程语言实现了入侵检测算法。

2.4.1 封装式特征选择的结果分析

NSL-KDD数据集里的每个样本都包含了41种特征,然而其中部分特征对最终分类的效果并不明显,有时甚至可能带来不良影响。因而,需要将特征选择作为必须进行的步骤,挑选出那些在分类结果上起着关键作用的特征,并且剔除那些对结果几乎没有影响的特征。表1展示了封装式特征选择算法在使用前后对随机森林算法结果的影响。

表1 数据集的特征选择结果

特征选择阶段	特征数量	Accuracy/%	Precision/%	Recall/%	F1_score/%
选择前	41	76.15	83.42	76.62	79.88
选择后	23	81.48	83.78	82.03	82.89

图4展示了未进行特征选择的RF算法与执行封装式特征选择的RF算法的对比结果。根据图4,特征选择前后随机森林除精确率(Precision)指标外,其他指标表现均存在显著差异。这个情况表明数据集中存在的特征重复和无效特征等问题,这些因素都对最终结果产生了负面效应。封装式特征选择算法在应对这些问题方面发挥了积极作用,成功提高了检测结果的准确性。

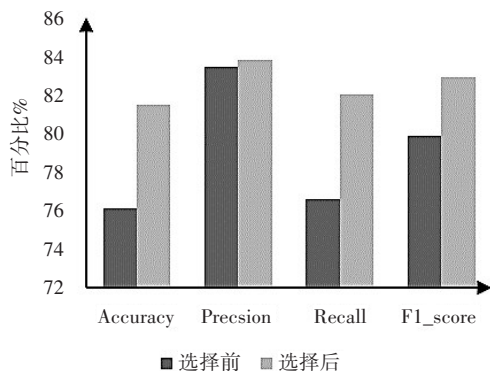


图4 执行封装式特征选择前后的结果对比

2.4.2 SMOTE过采样算法结果分析

从图3观察的结果来看,在采样之前,Probe类样本、R2L类样本,以及U2R类样本分别约为总样本的9.00%、0.82%和0.004%。这组数据明显显示,不同类型的训练样本在训练集中的数目有着相当大的差距。为了实验结果的真实性和有效性,本文对数据集运用了SMOTE过采样算法进行处理。SMOTE过采样算法仅在训练集KDDTrain+_20Percent.txt上进行采样。最终的采样结果如图5所示,展示了数据经过采样处理后的情况。

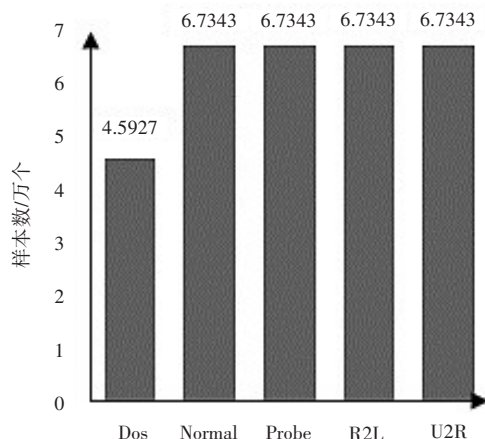


图5 训练集过采样后的样本分布

图6对比了没有应用SMOTE过采样技术的模型和应用了该技术的模型的结果。可以看出,除精确度(Precision)以外,随机森林(RF)在其他各项指标上的表现都有显著的差异。这种差异的主要原因是没有采用过采样技术的模型,在处理R2L和U2R类别样本时未能充分考虑其较少的情况,导致其在预测这两类数据时性能下降,从而对整体效果产生了负面影响。相反,模型整体性能在经过采样的训练集中表现更为出色,因为每个类别的样本数都得到了相对平衡。如果训练数据集的类别分布不均匀,就会使得学习模型难以精确识别数量较少的类别,即使建立了分类模型,也可能产生过度拟合的问题。这将直接影响模型在新数据上的表现,尤其对入侵类的分类结果产生负面影响。

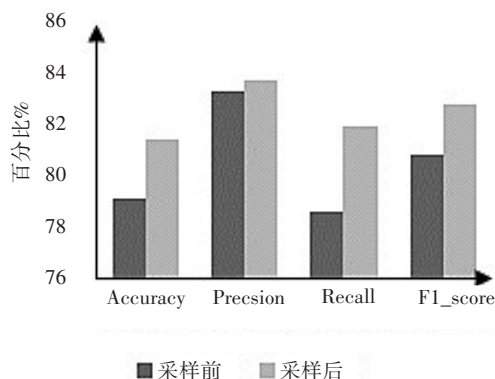


图6 执行SMOTE前后结果对比

2.4.3 模型整体指标分析

(1) 流量检测准确率。Yin等^[11]利用递归神经网络和SoftMax方法开发出一套入侵检测系统，GA-SVM^[12]使用遗传算法对流量数据集执行降维操作，同时采用过滤性特征选择技术剔除冗余特征，之后应用支持向量机方法对流量数据样本进行种类划分。GA-ANN^[13]采用遗传算法对流量数据集进行过滤式特征选择，通过降维技巧剔除冗余特征，最后使用人工神经网络和支持向量机的复合模型对流量数据样本执行分类操作。本文提出的模型简称为SMGA-RF。对Accuracy的对比结果详见表2。

表2 多种模型检查Accuracy的结果(%)

序号	模型	Accuracy
1	Recurrent Neural Network ^[11]	81.29
2	SVM ^[11]	79.00
3	Decision Tree ^[11]	80.43
4	GA-SVM ^[12]	79.27
5	GA-ANN ^[13]	84.38
6	SMGA-RF	86.83

从Accuracy的角度看，本文提出的SMGA-RF算法在NSL-KDD数据集上表现出明显的优势。这一优势主要来自NSGA-II具备高效多目标优化特性，加上随机森林模型对不平衡数据分类表现了出色的性能。因此，本文挑选的特征子集在处理不平衡数据分类上极其有效。此外，本文的模型不仅解决了数据集存在的特征冗余的问题，还通过结合SMOTE采样方法平衡了数据

集。因此，相较于表2中的其他五个模型，本文提出的模型在入侵检测效果上取得了更为显著的成果。

(2) 入侵检测性能。在上述实验中，对算法框架的每个阶段进行了性能测试，并在本节中对入侵检测算法的整体性能进行了实验验证。图7对NSL-KDD数据集的统计效果做了展示。从NSL-KDD数据集的分析结果看，SMGA-RF在统计分析上已经取得了明显的优势。本文充分利用了NSGA-II较强大的多目标优化能力，结合SMOTE对数据集进行平衡处理，最终实现了较出色的检测性能。

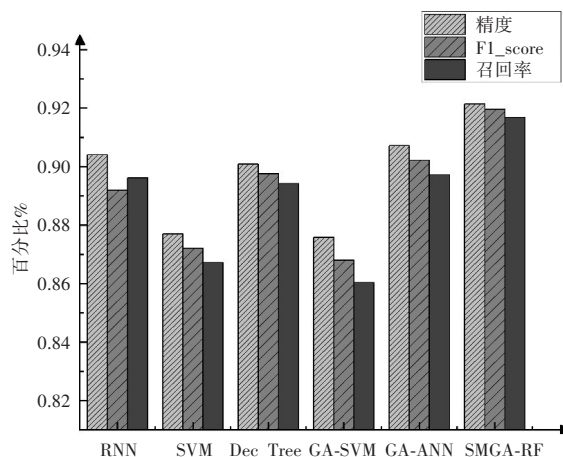


图7 NSL-KDD数据集的统计性能指标

3 结语

本文提出了一种综合利用封装式特征选择算法、SMOTE过采样算法以及RF算法的入侵检测模型，以应对当前机器学习入侵检测中普遍存在的特征冗余和类别不平衡问题。通过采用多目标遗传算法，该算法有效地选择了具有较强判别能力的特征子集，利用随机森林模型对不平衡数据的分类能力，提高了模型的性能。同时，SMOTE算法的使用增加了样本类别的数量，从而平衡了数据集，进而提升了入侵检测模型的效率。实验结果显示，本研究提出的方法在入侵检测数据集中表现出出色的识别能力，明显超过了同类的方法。这一研究为解决机器学习入侵检测中的挑战性问题的提供了一种有效的方法，为进一步提高网络安全水平提供了有益的思路。

参考文献:

- [1] 朱琨,张琪. 机器学习在网络入侵检测中的应用[J]. 数据采集与处理,2017,32(3):479-488.
- [2] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE Trans on Knowledge and Data Engineering,2009,21(9):1263-1284.
- [3] ZHANG X Y, WANG H Z. Intrusion detection of ICS based on improved border SMOTE for unbalance data[J]. Netinfo Security,2020,20(7):70-76.
- [4] 封化民,李明伟,侯晓莲,等. 基于SMOTE和GBDT的网络入侵检测方法研究[J]. 计算机应用研究,2017,34(12):3745-3748.
- [5] WANG J Z. Geometric structure of high-dimensional data and dimensionality reduction[M]. New York: Springer Heidelberg Dordrecht London,2011.
- [6] KASSEM AK, ELSAYED M, DAYA B, et al. A new feature representation method for intrusion detection system[C]//Proceedings of the International Conference on Computational Mathematics & Engineering Sciences, Antalya, Turkey,2019:1-7.
- [7] 方彬皓,黄洪,周子云. 多重特征选择和混合采样的入侵检测方法研究[J]. 四川轻化工大学学报(自然科学版),2022,35(4):85-93.
- [8] SONG Y, CAI Z P. An intrusion detection feature extraction method based on information theory model[J]. Journal of University of Electronic Science and Technology of China,2018,47(2):267-271.
- [9] REN J D, LIU X Q, WANG Q, et al. An multilevel intrusion detection method based on KNN outlier detection and random forests[J]. Journal of Computer Research and Development,2019,56(3):116-125.
- [10] WANG L L, LIU J H, FU X M. An intrusion detection method based on extreme learning machine and modified K-means[J]. Computer Engineering and Science,2018,284(8):66-72.
- [11] YIN C, ZHU Y, FEI J, et al. A deep learning approach for intrusion detection using recurrent neural networks[J]. IEEE Access,2017,5:21954-21961.
- [12] KHREICH W, KHOSRAVIFAR B, HAMOULHADI A, et al. An anomaly detection system based on variable N-gram features and one-class SVM[J]. Information and Software Technology,2017,91(11):186-197.
- [13] OMRANI T, DALLALI A, RHAIMI BC, et al. Fusion of ANN and SVM classifiers for network attack detection[C]//Proceedings of the International Conference on Sciences and Techniques of Automatic Control and Computer Engineering, Monastir, Tunisia,2017:1-13.

A study on improving the network intrusion detection model using genetic algorithm and SMOTE

Dai Zhouhao*

(School of Information Science and Technology, Guizhou University of Finance and Economics, Guiyang 550000, China)

Abstract: A new detection model is proposed to address the issues of data imbalance and feature redundancy in network intrusion detection. The model integrates genetic algorithms and the Synthetic Minority Over-sampling Technique (SMOTE) to enhance the accuracy of network intrusion detection. Firstly, to address data imbalance, the SMOTE algorithm is employed, inserting random samples between minority class instances to effectively increase their quantity and achieve inter-class balance. Moreover, a wrapper feature selection process, based on genetic algorithms and random forests, has been established to reduce feature redundancy. This process not only picks useful features and lessens unnecessary data, but also enhances the overall efficiency of intrusion detection. Finally, the processed dataset undergoes classification using the random forest algorithm, enabling effective detection of network intrusion instances. Experiments on the NSL-KDD dataset show that the network intrusion detection model based on genetic algorithm and SMOTE improves the overall recognition rate of intrusion detection.

Keywords: feature selection; SMOTE oversampling; random forest; network attack detection; genetic algorithm

文章编号: 1007-1423(2024)07-0031-07

DOI: 10.3969/j.issn.1007-1423.2024.07.005

基于神经网络的民航客运量的预测研究

唐甜甜*, 张佳明, 姜 为, 王 海

(中国民用航空飞行学院理学院, 广汉 618300)

摘要: 为准确预测民航客运量, 解决传统模型无法精准捕捉不稳定客运量的波动问题, 选取更可靠的 BP 模型, 以此挖掘数据的非线性以及非平稳特征和规律。针对近 17 年的时间序列民航客运量进行预测研究, 构建 BP 神经网络预测模型, 并与传统模型 ARIMA 对比分析。进行对比预测曲线图, 直观地反映出 BP 模型比 ARIMA 模型在波动点的预测上更加稳定, 表现更好。结果表明, BP 模型相比于 ARIMA 模型有更高的 R^2 和更低 MSE, 能够更有效提高民航客运量的预测精度和预测稳定性, 为制定航空运输生产计划和发展航空运输业提供了重要参考。

关键词: 神经网络; BP 模型; ARIMA 模型; 民航客运量; 预测分析

0 引言

随着社会经济的发展和人民生活水平的提高, 民航客运量在不断增加。如何准确地预测民航客运量, 对于航空公司、机场管理部门以及政府相关部门都具有非常重要的意义。在研究中, 有学者^[1]将注意力机制引入 Bi-LSTM 模型, 进而在不同时刻关注输入特征的信息, 考虑权重计算重要性, 得出重要特征, 以此提高预测的精准度。也有学者^[2]使用求和自回归移动平均 (ARIMA) 模型捕捉时间序列数据中的线性时序特征并以均方根误差 (RMSE) 等指标检验预测精确度。研究者^[3-7]构建 BP 模型, 部分选取 T-RH 模型继续对其参数进行优化的方法, 从而找到最佳参数进行一季稻发育期预测分析。传统的基于统计方法的预测模型中, 如 ARIMA 模型预测存在一些问题, 无法有效处理序列数据和容易出现过拟合等问题。而有学者提出人工神经网络模型^[8-10], 结合具有强大的输入输出拟合能力的机器学习算法, 能够有效处理非线性复杂过程, 在序列预测等领域得到了广泛应用。

比较了 ARIMA 模型^[11]、BP 模型和基于 Bi-LSTM 神经网络模型的民航客运量预测方法^[12-13], 在实验中验证了 Bi-LSTM 神经网络模型具有更好的预测效果, 其中, Bi-LSTM 模型是具有双向性的序列模型, 既可以处理历史数据还可以评估未来因素的影响, 可以有效快速地处理与序列相关的数据。ARIMA 模型是时间序列分析中常用的模型之一, 通过对历史数据进行拟合, 寻找出时间序列中的趋势、周期性等规律性。BP 模型则是一种基于神经网络的模型, 可以通过不断的迭代来优化参数, 并进行预测。

本文基于近 18 年的民航客运量数据, 构建 ARIMA 模型, 对时间序列进行处理, 结果表明 ARIMA 对时间序列分析的处理具有较高的可靠性和实用性。构建 BP 神经网络模型对数据进行误差分析模拟, 进而提高民航客运量的预测精确度。最后, 通过实验对比不同模型在时间序列预测上的表现, 分析其优缺点, 在实际应用中, 该方法可以为航空公司、机场管理部门以及政府相关部门提供有力的决策依据。

收稿日期: 2023-10-26 修稿日期: 2023-11-28

基金项目: 校级青年基金项目 (XJ2022002101)

作者简介: *通信作者: 唐甜甜 (1997—), 女, 甘肃陇南人, 助教, 研究方向为深度学习与民航预测, E-mail: 1064198457@qq.com; 张佳明 (1999—), 男, 重庆秀山人, 在读本科生, 研究方向为深度学习; 姜为 (2002—), 男, 黑龙江哈尔滨人, 在读本科生, 研究方向为机器学习; 王海 (2000—), 男, 贵州毕节人, 在读本科生, 研究方向为大数据分析

1 模型原理

1.1 ARIMA模型概述

ARIMA (autoregressive integrated moving average) 称为差分自回归移动平均模型, 一般记为 $ARIMA(p, d, q)$ 模型, 由自回归过程 ($AR(p)$)、单整 ($I(d)$) 以及移动平均过程 ($MA(q)$) 三部分构成。其中, AR 为时间序列的当前值, 为滞后 p 期观测值的线性组合; ($I(d)$) 具体指将一个非平稳时间序列经过 d 次差分后转化为平稳序列所需的差分处理次数; MA 即模型值, 可以表示为 q 阶残差项的线性函数。该模型的表达式如下:

$$x_t = \Phi_0 + \Phi_1 x_{t-1} + \dots + \Phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q} \quad (1)$$

ARIMA 模型的 p 阶自回归模型 $AR(p)$ 公式如下:

$$X_t = \sum_{i=1}^p \Phi_i X_{t-i} + \omega_t \quad (2)$$

ARIMA 模型的 q 阶自回归模型 $MA(q)$ 公式如下:

$$X_t = \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \omega_t \quad (3)$$

将公式(2)和公式(3)结合, 得到自回归移动平均模型 $ARMA(p, q)$ 公式如下:

$$X_t = \omega_t + \sum_{i=1}^p \Phi_i X_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \quad (4)$$

式中: X_t 表示任意一个 t 时刻的序列值; ω_t 代表 t 时刻的随机误差; $\sum_{i=1}^p \Phi_i X_{t-i}$ 表示 t 时刻的前 p 个序列值的线性组合; $\sum_{j=1}^q \theta_j \varepsilon_{t-j}$ 表示 t 时刻的前 q 个序列的白噪声累加和的线性组合。ARMA(p, q) 模型只能适用于预测平稳性的时间序列数据, 对于非平稳时间序列, 则需要进行 d 阶差分处理, 由 ARIMA(p, d, q) 模型进行预测。

1.2 BP神经网络建模原理

BP 神经网络, 即为误差反向传播 (back propagation) 的多层前馈网络。BP 网络框架是由输入层、输出层和若干隐含层组合成的, 如图 1 所示。在同一层, 不同神经元的不连通, 相邻层的神经元以权重的形式连接传递特征信息。在预测过程中, 首先输入变量输入到输入层,

逐层向前通过权重传递特征的重要信息, 最后预测结果由输出层输出, 在每一层网络中, 特征信息都会经过相应函数计算变换。BP 网络的预测过程分为两个部分完成, 一是通过向前计算, 而是由误差向后传播。整个预测过程中, 误差会实时在传播中进行不断的调整和修正, 进而提升模型的预测精度, 使得网络达到平衡, 最后达到收敛的状态。

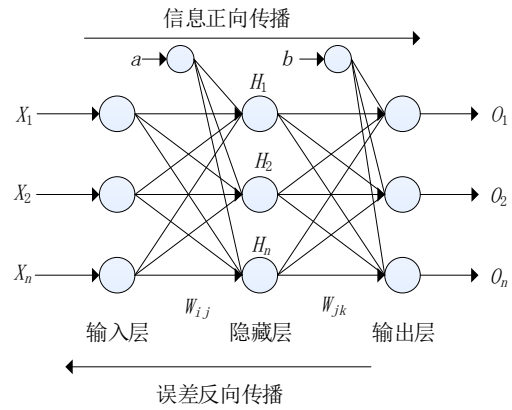


图 1 BP神经网络结构图

图中, $X_1, X_2, \dots, X_n$ 为输入层接收的数据, $H_1, H_2, \dots, H_n$ 为隐含层的节点, 输入层到隐含层的权重 W_{ij} , $O_1, O_2, \dots, O_n$ 为输出层的最终输出, 隐含层到输出层的权重为 W_{jk} , 输入层到隐含层的偏置为 a_j , 隐含层到输出层的偏置为 b_k , 设置激励函数为 $g(x)$, 选取 Sigmoid 函数, 公式如下:

$$g(x) = \frac{1}{1 + e^{-1}} \quad (5)$$

公式(6)中, X_i 表示输入变量, W_{ij} 、 a_j 表示输入层和隐含层间的连接权值及隐含层阈值, H 是隐含层节点数, g 为隐含层激活函数。隐含层 H 的输出为 H_j , 公式如下:

$$H_j = g\left(\sum_{i=1}^n W_{ij} X_i + a_j\right), j = 1, 2, 3, \dots, l \quad (6)$$

公式(7)中, W_{jk} 和 b_k 分别是连接权值和阈值。输出层 O 的输出为 O_k , 公式如下所:

$$O_k = \sum_{j=1}^l H_j W_{jk} + b_k, k = 1, 2, 3, \dots, m \quad (7)$$

公式(8)中 Y_k 是期望输出值, O_k 是实际输出值。误差的计算公式如下:

$$e_k = Y_k - O_k \quad (8)$$

权值通过预测误差值 e 对网络连接的权值完

成更新, 其中, η 为学习率, 如公式(9)~(12)所示:

$$W_{ij} = W_{ij} + \eta H_j (1 - H_j) x_i \sum_{k=1}^m W_{ij} e_k \quad (9)$$

$$i = 1, 2, 3, \dots, n; j = 1, 2, 3, \dots, l$$

$$W_{ij} = W_{ij} + \eta H_j e_k \quad (10)$$

$$k = 1, 2, 3, \dots, m; j = 1, 2, 3, \dots, l$$

$$a_j = a_j + \eta H_j (1 - H_j) \sum_{k=1}^m W_{jk} e_k \quad (11)$$

$$j = 1, 2, 3, \dots, l$$

$$b_k = b_k + \eta e_k \quad (12)$$

$$k = 1, 2, 3, \dots, m$$

2 构建模型与预测分析

2.1 数据来源

选取了中国民航总局官网中2005年1月至2023年1月的民航客运量数据作为研究对象, 部分数据见表1。

2.2 构建ARIMA模型

2.2.1 建模步骤

下面给出ARIMA模型的详细建模步骤:

1) 对时间序列数据进行平稳性检验, 若该序列平稳, 将进行白噪声检验;

2) 如时间序列为不平稳序列, 可通过 d 阶差分处理使非平稳序列转化为平稳序列;

3) 求解观测序列的样本自相关系数(ACF)

和偏自相关系数(PACF)的值;

4) 根据样本自相关系数和偏自相关系数表现出来的截尾性和拖尾性质, 确定 p 和 q 的值, 对ARIMA(p, d, q)模型进行拟合;

5) 估计模型中的未知参数的值;

6) 通过残差序列和数据的正态分布情况的白噪声检验, 说明模型的有效性。

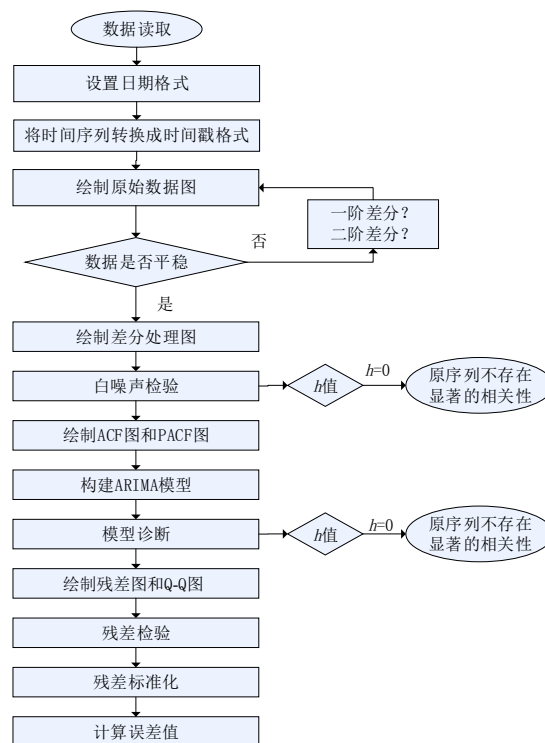


图2 构建ARIMA模型的流程图

表1 国内民航客运量2008—2022年月度数据

单位: 万人次

月份	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
1	1500	1748	1941	2267	2567	2567	3058	3246	3736	4393	4647	4647	4401	3004	2942
2	1502	1702	2025	2163	2349	2787	3109	3493	3898	4279	4843	5341	729	2386	3121
3	1581	1800	2173	2282	2503	2878	3025	3670	3894	4431	5140	5341	1664	4768	1527
4	1659	1883	2157	2437	2605	2849	3142	3579	4000	4402	5074	5341	2573	5091	1197
5	1536	2264	2600	2758	3071	3476	3730	4165	4642	5046	5657	5930	4598	2227	2189
6	1417	1771	2185	2325	2530	2865	3061	3404	3800	4374	4938	5341	3060	4116	2189
7	1718	2103	2547	2660	2995	3266	3580	3915	4353	4860	5378	5930	3894	4899	3385
8	1596	2264	2660	2758	3071	3476	3730	4165	4642	5046	5657	5930	4598	2227	3213
9	1641	1931	2260	2501	2744	3049	3315	3672	4169	4655	5029	5930	4775	3599	1989
10	1834	2171	2448	2638	2829	3165	3488	3855	4374	4883	5408	5930	5014	3875	1572
11	1700	1958	2078	2371	2584	2851	3240	3498	3971	4646	5006	5930	4425	2142	1234
12	1567	1875	2103	2300	2563	2793	3193	3500	4041	4666	5018	5930	4216	2698	1841

2.2.2 平稳化处理

由原始数据时序图3知,原始数据时序图存在不平稳性和周期性,差分法可以剔除周期性因素,将对归一化后的数据进行差分处理,图4是一阶差分之后的数据图,观察后发现时序还存在不平稳性,从而需要做二阶差分,如图5为二阶差分的数据周期性展示,说明数据已经具有平稳性。

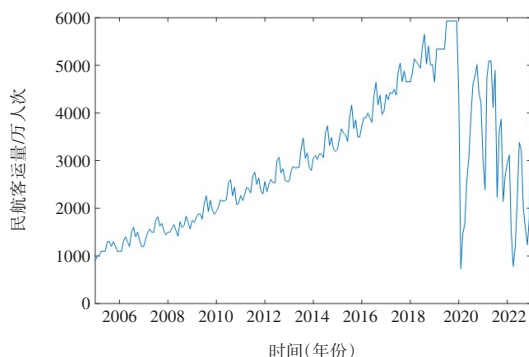


图3 民航客运量2005—2023年原始数据时序图



图4 一阶差分之后的数据图

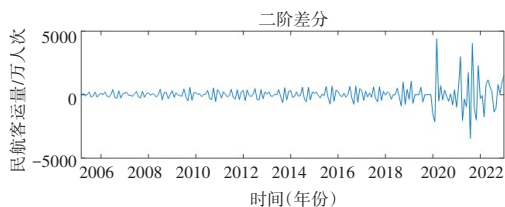


图5 二阶差分之后的数据图

2.2.3 参数 p 和 q 的确定

接下来将对数据进行求解,观测序列的样本自相关系数(ACF)和偏自相关系数(PACF),结果分别如图6、图7所示。

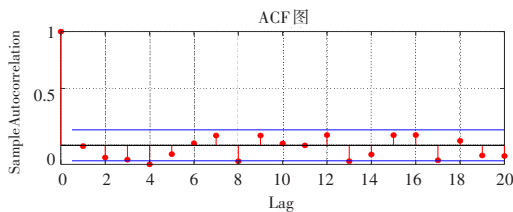


图6 自相关图

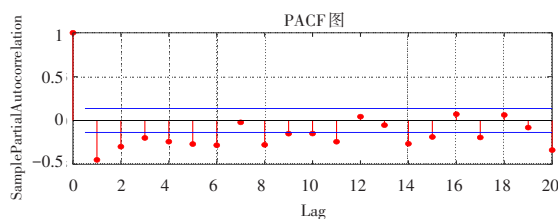


图7 偏自相关图

观察图6、图7,根据样本自相关系数和偏自相关系数的截尾性和拖尾性质,构建多个ARIMA模型,通过对多个ARIMA模型的检验,最终确定取值 $p=12$ 和 $q=7$ 最优,确定构建ARIMA(12,2,7)模型。

2.2.4 平稳性检验

构建模型的基础之上,进行Ljung-Box Q检验,检验时间序列的自相关性,模型进行残差序列白噪声检验,结果如图8所示,通过原假设表明残差序列存在显著的自相关性,并通过输入样本得到正态分布的QQ图如图9所示,由此可知数据明显呈现正态分布趋势。

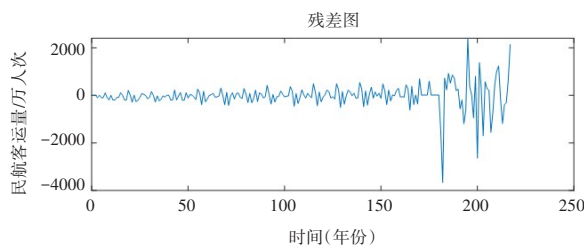


图8 残差序列白噪声检验

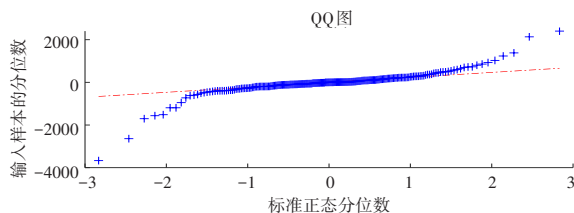


图9 正态分布-QQ图

通过对残差数据进行检验并输出误差指标,进而说明模型的有效性,可进行模型拟合,预测该时间序列未来三年的民航客运量,采用80%数据作为训练集,剩余的作为测试集,通过测试数据和原始数据的对比,检验模型的准确性。

2.2.5 模型的预测

选取 ARIMA(12,2,7)模型对预处理的民航客运量数据进行部分月份的预测分析。从图 10 可以看出,个别点民航客运量是增长还是降低存在误判的情况,但是总体的趋势预判较为合理,在后期民航客运量突然出现大幅度变化时,ARIMA 模型也能较好地预测销售额。其中,蓝色曲线表示差分处理后的原始数据曲线,红色曲线表示 ARIMA 模型预测的曲线。从红色的预测曲线拟合蓝色的真实曲线的情况来看,ARIMA 模型适用于短期的时间序列预测。

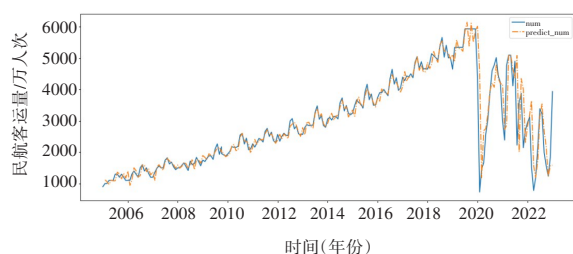


图 10 ARIMA 模型预测图

2.3 构建 BP 神经网络模型

2.3.1 建模步骤

BP 神经网络运算基本流程共有如下几步：

1) 根据所确定的风险指标体系,收集分析样本数据,对数据进行归一化。根据训练集与测试集的样本数量分配要求,确定训练样本数据与测试样本数据的样本量,将原始数据的数据格式根据 BP 神经网络训练与测试时的样本格式要求进行相应的修改;

2) 网络训练,通过输入层输入训练样本数据进行网络权值初始化,创建神经网络并对训练样本进行深度训练,按照设置的训练参数不断训练,直到输出结果的误差达到了预期设计的误差要求;

3) 网络测试,在 BP 神经网络对训练集的样本数据训练完毕后,将待测试的样本数据输入到训练好的神经网络之中进行测试,并得出测试样本的预测值,计算出训练集和测试集的误差指标 RMSE、MAE、 R^2 ,最后对预测值与预期结果进行对比分析;

4) 计算隐含层和输出层的输入输出结果,

绘制实际数据拟合图以及预测数据图,再将实际曲线和预测曲线进行对比。

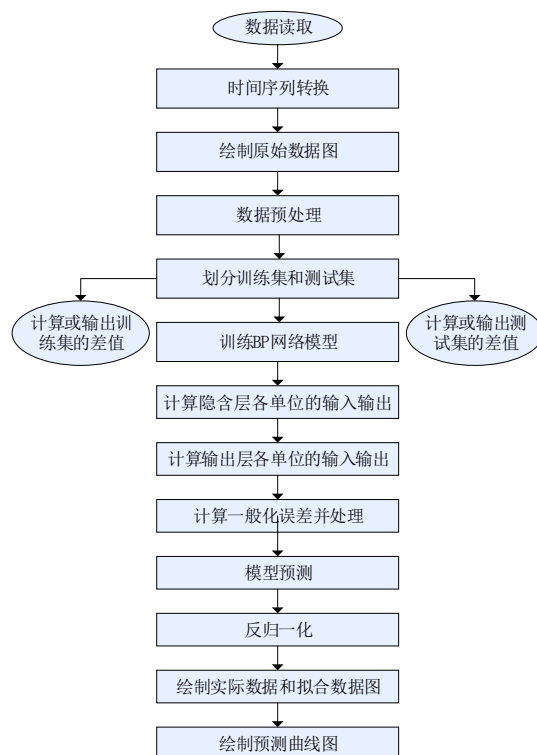


图 11 BP 模型构建流程图

2.3.2 数据预处理

通过对原始数据时序图分析发现数据不平稳,并且受到季节性等因素的影响,在对神经网络模型训练之前进行数据预处理,在前期数据预处理的基础上,继而采用归一化的方式将数据按照最大值和最小值缩放到特定区间 $[0,1]$,去除量纲化带来的影响,转化为无单位的数值,在后续的时间序列预测中,更便于进行加权组合等方式比较,从而加快训练网络的收敛性,归一化公式如下所示:

$$x^* = \frac{x - \min}{\max - \min} \quad (13)$$

在神经网络训练结果可视化中,采用反归一化中 scaler 函数进行数据转换,在可视化阶段,将其转换为初始数据形式。

2.3.3 模型的训练

为了确定模型的最优参数,需要通过训练集反复训练此网络,确定输入节点和输出节点

以及数据维度。选用三层的网络架构。为了得到较为理想的预测模型,将神经元个数、学习率、激励函数作为控制变量进行模型训练,并试预测后得到相对较优的网络参数值,将最大迭代次数设为100次,学习率设为0.001,中间层的传递函数选用ReLU函数,以此使网络的学习周期大幅度缩短。在梯度下降的思想中,学习率决定着算法每一步迭代中的更新步长,若学习率较小,则需要较高的迭代次数才能达到收敛;若学习率较大,则容易产生振荡,甚至会出现无法收敛的情况。神经网络中常用的激活函数为Logistic、Tanh、ReLU等函数。

2.3.4 模型的预测

为更清晰直观地展示预测效果,局部预测结果如图12所示,根据指标计算得出测试集的误差值。从图中可以看出,实际数据与预测数据拟合效果较为完美,预测的民航客运量数据贴合真实的数据值,预测值的波动性较小,由此可知模型测试成功。综上所述,说明BP神经网络模型在对非线性问题进行预测时具有较好的逼近能力。

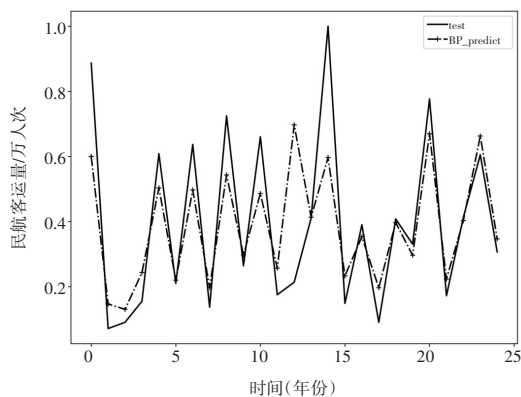


图12 BP模型预测图

2.4 预测结果对比分析

为了验证模型的有效性、民航客运量预测的准确性,将ARIMA预测得到的数据与BP网络的输出值进行对比分析。在民航客运量的预测研究方法上,为了更加清楚直观地对比,将结果进行量化分析,计算传统的经典模型ARIMA以及捕捉非线性特征的预测网络BP在同一时间序列预测数据集上的误差。将选取均方

根误差RMSE、平均绝对误差MAE和回归直线拟合优度 R^2 作为误差评判指标。ARIMA以及BP模型的误差对比见图13。评判指标的计算公式如下:

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_i - \hat{y}_i| \quad (14)$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)(y_i - \hat{y}_i)} \quad (15)$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2} \quad (16)$$

其中: y_i 是真实值, $\hat{y}_i$ 是预测值, $\bar{y}$ 是样本平均值, $\sum_i (\hat{y}_i - y_i)(\hat{y}_i - y_i)$ 为预测产生的误差, $\sum_i (\bar{y} - y_i)(\bar{y} - y_i)$ 为均值产生的误差。

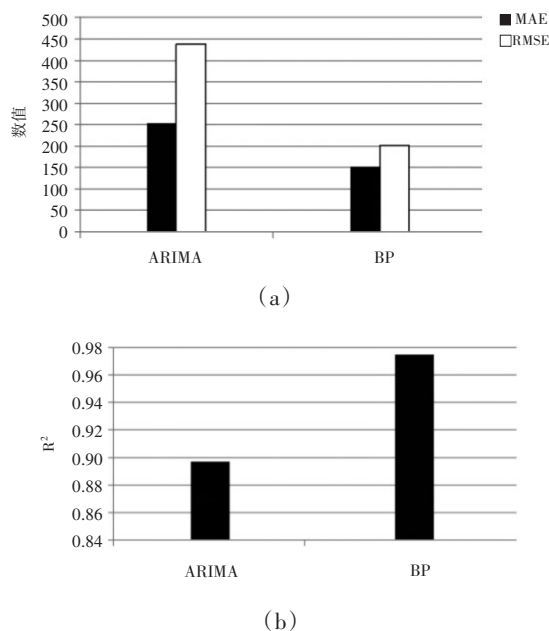


图13 各模型预测结果误差

对两种模型的预测结果,从图13中不难看出预测结果均比较理想,但是BP模型的各项指标均优于ARIMA模型。ARIMA预测模型在时序民航客运量数据的预测表现上也较为客观,但误差指标以及拟合效果不如BP模型表现出色。

主要选取经典预测时序数据的ARIMA模型来描述民航客运量数据的线性关系,选取BP网络模型捕捉民航客运量数据的非线性规律,以此解决不确定性以及ARIMA无法提取的非线性特征等问题。在时序民航数据的预测中,主要

分析预测了客运量,有助于了解民航业的发展形势,也为制定航空运输生产计划和发展航空运输业提供了重要参考。

3 结语

深度学习技术在时间序列预测方面取得了较大的发展与成功应用。选取民航客运量时间序列数据,对其进行分析和处理后,构建ARIMA和BP的民航客运量的预测模型,采用时间序列预测方法对民航航线的客运量进行预测。搭建民航客运量的预测模型,将传统模型与神经网络模型进行对比分析,试图找出更符合实际情况的模型。通过对基于时间序列民航客运量的预测,对比两类方法,得出结论:相较于传统时间序列预测的ARIMA模型,神经网络BP模型表现出更加优秀的预测精度,BP模型在学习与预测序列能力方面更为优秀,实验数据的预测效果更好。

参考文献:

- [1] 赵恒辉,黄德启,曾蓉,等.基于时空注意力Bi-LSTM模型的短时交通流预测[J].计算机仿真,2022,39(9):177-181,455.
- [2] 梁乃兴,闫杰,杨文臣,等.基于ARIMA-LSTM的高速公路交通安全组合预测模型研究[J].重庆交通大学学报(自然科学版),2023,42(4):131-138.
- [3] 樊闯,赵子皓,张雪松,等.基于BP神经网络的一季稻发育期预测模型[J].浙江农业报,2023,35(2):434-444.
- [4] 窦益华,张佳强,李国亮,等.基于优化BP神经网络的连续管疲劳寿命预测[J].石油机械,2023,51(10):144-149.
- [5] 李其操,董自健.基于GA-BP神经网络的温室温度预测研究[J].智能计算机与应用,2023,13(9):168-171.
- [6] 李娟娟,薛元,徐志武,等.基于BP神经网络的全色域彩色纱颜色预测模型构建[J].纺织高校基础科学学报,2023,36(4):37-46.
- [7] 李轩,周新苗,吴晓峰.基于HW-EEMD-SVM模型的民航客运量预测[J].数量经济研究,2023,14(2):189-204.
- [8] 陈聪聪,李程.基于PSO-ELM算法的民航客运量预测[J].控制工程,2022,29(11):2042-2047.
- [9] 张启凡,王永忠,王圣堂,等.基于NARX动态神经网络的民航客运量预测研究[J].计算机与数字工程,2022,50(7):1485-1488,1493.
- [10] 甘国育,游进国,张婷.基于双向LSTM的民航客运量预测[J].现代电子技术,2022,45(14):175-180.
- [11] 王承杰,孟彦菊.基于季节性ARIMA模型的民航客运量预测[J].统计与管理,2022,37(5):88-93.
- [12] 甘国育,游进国,段培娟.融合1D-CNN和LSTM的民航客运量预测模型[J].信息技术,2022(3):6-11.
- [13] 刘铭基,田雅楠,张亮,等.基于Prophet-ARIMA模型的民航周转量预测研究[J].计算机技术与发展,2022,32(2):148-153,160.

Research on predicting civil aviation passenger transport volume based on neural networks

Tang Tiantian*, Zhang Jiaming, Jiang Wei, Wang Hai

(College of Science, Civil Aviation Flight University of China, Guanghan 618300, China)

Abstract: In order to accurately predict civil aviation passenger traffic and solve the problem that traditional models cannot accurately capture the fluctuation of unstable passenger traffic, a more reliable BP model is selected as a way to mine the nonlinear and nonsmooth features and laws of the data. Aiming at the time series of civil aviation passenger traffic in the past 17 years, we carry out forecasting research, construct the BP neural network forecasting model, and compare and analyze it with the traditional model ARIMA. Comparison of prediction graphs is carried out, which intuitively reflects that the BP model is more stable and performs better than the ARIMA model in the prediction of fluctuation points. The results show that the BP model has higher R^2 and lower MSE than the ARIMA model, which can more effectively improve the prediction accuracy and prediction stability of civil aviation passenger traffic, and provides an important reference for the development of air transportation production plan and the development of air transportation industry.

Keywords: neural network; BP model; ARIMA model; civil aviation passenger traffic; predictive analysis

文章编号: 1007-1423(2024)07-0038-06

DOI: 10.3969/j.issn.1007-1423.2024.07.006

改进 DeepLabV3+ 的复杂场景倒塌建筑物分割算法

钟联升, 陶青川*

(四川大学电子信息学院, 成都 610065)

摘要: 受逆光、热噪声和云雾遮挡因素的影响, 现有分割算法在震后倒塌建筑物提取上精度和效率表现不佳。针对上述问题, 基于 DeepLabV3+ 提出了应对复杂场景的倒塌建筑物分割算法。首先, 选择 MobileNetV2 骨干网络进行替换, 降低网络的计算参数, 提高分割速度。然后, 在特征融合阶段引入 AS 模块, 选择性增强与倒塌建筑物相关的特征通道权重系数, 解决分割精度低的问题。此外, 在骨干网络中引入了 FPN 架构, 使得多级特征在解码阶段得到整合, 均衡捕捉不同倒塌范围的建筑物信息。实验结果表明, 改进后的网络在一定程度上解决了复杂场景下倒塌建筑物分割效率与精度的难题。

关键词: 深度学习; 图像分割; 倒塌建筑物; 复杂场景

0 引言

世界上许多地区都面临着高密度建筑和居住人口的挑战, 而其中一个主要的自然灾害威胁就是地震^[1]。地震发生后, 随时可能出现严重伤亡的情况。救援组织通常需要了解地震的规模、分布以及建筑物的受损情况, 以便采取有效的行动^[2]。现场损毁排查工作虽然可以提供更为详细的信息, 但费时费力。此外, 余震与倒塌的建筑材料易对人员造成二次伤害。若通过遥感技术以最小延时获取建筑物损坏程度图像, 救援团队即可不在现场观测的情况下做出决策, 因此快速获取倒塌建筑物位置与区域的方法对灾后管理评估至关重要。

图像分割技术可快速获取倒塌建筑物的损毁信息, 随后还可进行损毁分析和修缮跟踪等后续管理^[3]。国内外许多学者利用不同方法来检测倒塌建筑物区域, 也取得了一些实质性的成果^[4-6]。叶昕等^[7]利用受损房屋光谱分布杂乱的特点, 设计了梯度空间统计的方法来检测损毁建筑物; 葛小三等^[8]结合生成对抗网络和变

化检测网络, 使模型在有限数据集下拥有更好的检测性能; Li 等^[9]提出在跳跃连接中添加一个压缩步骤, 以减少网络权重; Cui 等^[10]基于 Swin Transform 提出了具有编码器-解码器结构的模型, 可同时融合多级特征使网络更注重非局部信息, 以应对倒塌建筑物样本的多样性。尽管前述的方法在倒塌建筑物检测方向已经有广泛研究, 对建筑物尺寸和几何形状的变化也表现出较强的鲁棒性, 但大部分分割工作环境都在高分辨图像下完成的, 并且遥感图像不可避免地易受到噪声干扰、曝光过度、云雾遮蔽复杂因素的挑战。这些场景下的研究仍然相当有限。

为了应对上述难题, 提出一种改进 DeepLabV3+ 的算法, 在面对复杂场景下可有效改善倒塌建筑物分割精度低的问题, 且平衡了精度和速度。

1 复杂场景分割模型设计

1.1 DeepLabV3+ 算法

编解码结构的奠基者 U-Net 在语义分割任务

收稿日期: 2023-10-28 修稿日期: 2023-11-27

作者简介: 钟联升(1998—), 男, 四川南充人, 在读硕士研究生, 研究方向为智能信息系统; *通信作者: 陶青川(1972—), 男, 四川南充人, 博士, 副教授, 研究方向为模式识别与智能系统, E-mail: taoqingchuan@scu.edu.cn

中提供了卓越的解决方案,但其偏向于小目标分割任务,而且需大量的计算资源来训练和推理。在实验中发现 DeepLabV3+相对于 U-Net 方法综合表现更出色^[11-12]。DeepLabV3+算法具备强大的能力来捕捉视觉数据不同部分之间的局部和全局关系,而这出色的能力可追溯到当中两个关键点。首先,采用空间金字塔结构来更有效地编码多尺度上下文信息。其次,编码器中使用融合技术拼接骨干网络的低级特征,使模型能够更准确地捕捉细节。

1.2 改进的 DeepLabV3+算法

从以下三个方面对 DeepLabV3+进行了改进：①骨干网络采用 MobileNetV2 替换原 Xception 特征提取网络，减小网络参数与计算负担；②对骨干网络使用 FPN(feature pyramid network) 架构，在解码模块中实现低、高级特征语义信息的整合；③提出了一种新颖的特征提取模块 AS-block，以提高模型在复杂场景下的分割准确率。改进后的整体网络如图 1 所示。

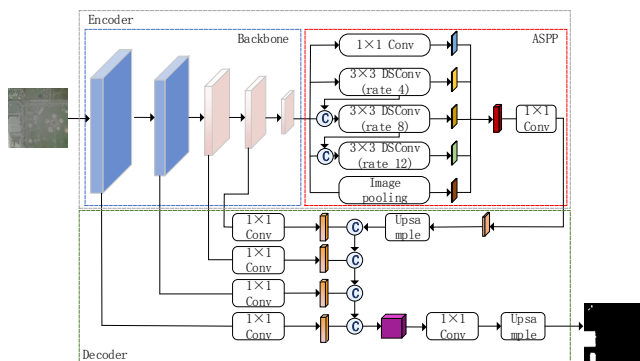


图 1 改进后的整体网络

1.2.1 骨干网络的改进

Xception 骨干网络在运算时可能面临计算资源不足与效率低下的问题^[13]，而 MobileNetV2 采用深度可分离卷积可在保持高性能水平的同时减少计算复杂性，与标准卷积不同之处如图 2 所示。为了改善这一问题，将骨干网络替换为 MobileNetV2。深度可分离卷积包括两个关键步骤：首先，进行深度卷积操作，逐通道与卷积核进行独立卷积；接着，通过逐点卷积操作，将特征通道的数量还原为原始值。

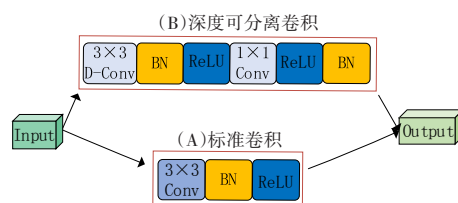


图 2 特征提取过程

下面分析了两种卷积的计算量。假设卷积核尺寸为 $F_A \times F_A$, 输入特征图尺寸为 $F_B \times F_B \times M$, 输出特征图尺寸为 $F_B \times F_B \times N$ 。当中的 F_A 与 F_B 为特征图的长度与宽度, M 与 N 为通道数。标准卷积与深度可分离卷积计算量如式(1)和式(2)。

$$Conv_{std} = F_A \times F_A \times M \times F_B \times F_B \times N \quad (1)$$

$$Conv_{ds} = F_A \times F_A \times M \times F_B + F_B \times F_B \times N \times M \quad (2)$$

根据式(3),可知使用深度可分离卷积,计算复杂性将降低约90%。

$$Ratio = \frac{Conv_{std}}{Conv_{disc}} = \frac{1}{N} + \frac{1}{A^2} \quad (3)$$

此外，在解码阶段低级特征图拥有较少的语义信息但倒塌位置准确，而高级特征图拥有丰富的语义信息但倒塌位置更为粗糙。为了在不增加原始模型的计算复杂性的情况下，解决对象的多尺度问题并提高模型对不同倒塌范围的分割能力，本文在骨干网络中引入特征金字塔架构^[14]，整体架构如图3所示。具体的实现过程如下：首先，利用 1×1 卷积核对骨干网络提取的特征层进行特征维度调整；随后，将这些特征与深层特征逐层融合，生成最终的预测结果。

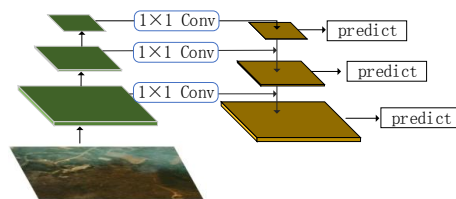


图 3 FPN 架构

1.2.2 AS 模块

在复杂场景倒塌建筑物识别过程中，很可能会面临噪声、云雾遮挡、曝光过度等问题。随着网络的深度增加，在特征提取阶段易使模型提取到与倒塌建筑物无关的高阶特征，如受损建筑周边不相关的环境特征，而忽视倒塌建

建筑物本身的特征。

为解决这一挑战,本文提出了AS-block特征提取模块,以进一步提升模型在复杂场景下的分割准确率。AS-block是通过组合ASPP-block和SE-block获得的,如图4所示^[15-16]。该模块旨在建立特征通道之间的相互关联性,允许模型使用全局信息来有选择性地增强与建筑物倒塌相关的特征通道,并抑制那些不相关的通道,使网络能够提取更丰富的高级特征。

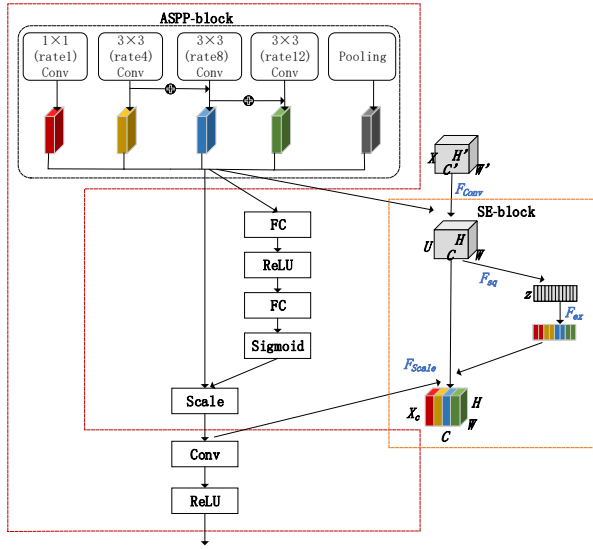


图4 AS模块

首先, F_{Conv} 是将特征 X 经过四个不同倍率的空洞卷积, 中间的三个分支采用逐层堆叠的方式实现像素级的密集采样, 获得待处理特征 U , 操作细节见图5。然后在特征 U 上分别执行 F_{sq} 与 F_{Scale} 。 F_{sq} 是对特征 U 进行压缩, 即对每个通道 C 进行全局平均池化操作, 以生成一个 $1 \times 1 \times C$ 的特征序列 z , 操作表达式如式(4)。

$$z_c = F_{sq}(U_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (4)$$

其中: H 和 W 分别代表特征 U 的第 c 个通道的高度和宽度。

紧接着为SE-block最关键的 F_{ex} 操作, 计算不同通道特征图的权重系数。以 z 作为输入, 通过权重为 W_1 与 W_2 的两个全连接网络计算, $F_{ex}(*, W)$ 表示全连接层的计算操作, 结果输出为 $1 \times 1 \times C$ 特征图, 表达式如式(5)。

$$s = F_{ex}(z, W) = \sigma(W_2 \delta(W_1 z)) \quad (5)$$

式中: δ 表示 $ReLU$ 函数, σ 表示 $Sigmoid$ 函数, W_i 代表全连接层权重。

在获得权重系数后, 完成最后的 F_{Scale} 操作, 即重新标定特征 U 每个通道的权重值。将每个通道的权重值 s_c 与相应通道的特征图 u_c 进行乘积, 得到不同通道特征图重要性程度的 X_c , 其中 $X_c = [x_1, x_2, \dots, x_c]$, 表达式如式(6)。实验结果表明(见2.3小节), 采用AS模块可以显著提高倒塌建筑物识别的准确率。

$$X_c = F_{Scale}(u_c, s_c) = u_c \cdot s_c \quad (6)$$

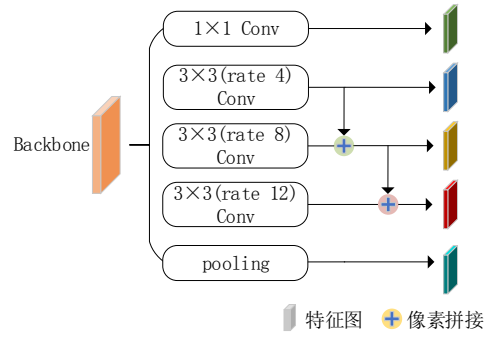


图5 ASPP处理细节

2 训练实验与结果分析

2.1 数据集制作

当前尚未有公开可用的复杂场景下倒塌建筑物数据集, 为了完成研究, 以高质量的灾害数据集xBD为基础, 从中挑选出震后倒塌明显的建筑物相关图像, 然后通过云雾合成、曝光增强与图像加噪技术自制复杂场景的数据集, 数据集中的图像如图6所示^[17]。



图6 复杂场景倒塌建筑物数据集

实验数据集共计2330张, 按8:1:1的比

例分成了训练、测试和验证集，然后对每个集合分别执行相同的增强策略，以确保测试和验证的稳定性。

2.2 实验平台与训练细节

本文所有的实验均在表1所示的配置环境下完成。

表1 训练平台配置

模块	参数
操作系统	Ubuntu 20.04
CPU	Intel(R) Core(TM) i5-6500 CPU@3.20GHz
GPU	GTX 1080Ti 11GB
内存	16 GB
Torch	1.10.0
Python	3.7

网络使用 MobileNetV2 权重来完成对复杂场景下倒塌建筑物的迁移学习。同时，为了确保网络充分学习数据集，输入网络的图像尺寸设置为 512×512 像素。模型的参数通过使用 RMSProp 优化器进行更新，初始学习率为 10^{-4} ，进行 300 轮训练迭代。为了动态调整学习率，采用了 Cosine 策略，同时设置了权重衰减为 $\beta = 0.001$ 。

2.3 网络评价指标

为了评估模型在复杂场景下倒塌建筑物分割中的表现，选择了四个评估指标，包含 mAP（平均精度）、mIoU（平均交并比）、FPS（每秒帧数）以及 Parameters（模型参数）。其中，mIoU 和 mAP 用于度量图像分割的准确性，而 FPS 和 Parameters 则用于评估模型的效率和复杂性。准确性相关指标计算公式如下：

$$mAP = \frac{1}{m+1} \sum_{i=0}^m \frac{p_{ii}}{\sum_{j=0}^m p_{ij}} \quad (7)$$

$$mIoU = \frac{1}{m+1} \sum_{i=0}^m \frac{p_{ii}}{\sum_{j=0}^m p_{ij} + \sum_{j=0}^m p_{ji} - p_{ii}} \quad (8)$$

式中： $m+1$ 为总类别数量， i 表示真实标定数据， j 表示预测分割数据， p_{ij} 表示将 i 预测为 j ，其余同理。

2.4 网络对比实验

为了评估不同骨干网络的性能，将骨干网络替换成了常用的 ResNet50 与 MobileNetV2，并将比较结果列于表2。最终选择的轻量化 MobileNetV2 相较于原骨干网络 Xception 在牺牲较少精度的情况下，FPS 提升近两倍，模型大小减小至 6.82 M，实现了更出色的性能。

表2 不同主干网络验证结果

模型名称	mAP/%	mIoU/%	Params/M	FPS/(f·s ⁻¹)
Xception	89.10	80.20	54.71	5.65
ResNet50	81.11	74.13	39.65	7.89
MobileNetV2	87.07	79.82	6.82	15.81

为了更进一步验证本文提出网络模型在复杂场景中的有效性，选择当前主流的语义分割算法在自制的复杂场景数据集上进行比较，实验数据见表3，分割效果见图7。

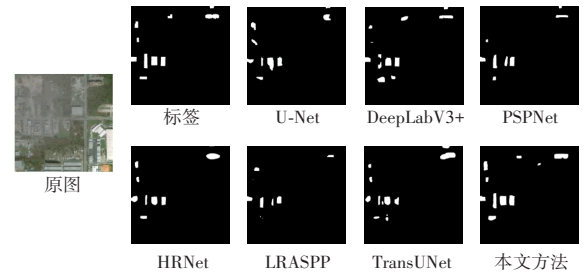


图7 不同模型分割效果图

表3 不同网络模型验证结果

模型名称	mAP/%	mIoU/%	Params/M	FPS/(f·s ⁻¹)
U-Net	82.20	74.50	24.89	7.32
DeepLabV3+	80.37	72.51	54.71	6.55
PSPNet	79.85	69.18	46.26	6.99
HRNet	80.10	71.27	29.54	7.93
LRASPP	71.12	60.36	3.29	21.58
TransUNet	92.13	85.21	93.20	1.38
本文模型	87.07	79.82	6.82	15.81

从表3可以看出，尽管 LRASPP 的网络模型参数较小，分割速度领先其他算法，但分割的精度相对较低，与实际需求不符。另一方面，TransUNet 在 mIoU 方面表现最佳，但其参数计

算量较大, FPS 仅有 1.38 帧/秒。然而, 本文提出的改进算法相较于其他算法具有更快的分割速度和更小的模型参数, 同时在精度和速度之间取得了平衡, 在复杂场景的倒塌建筑物分割任务中更具实际意义。

2.5 网络消融实验

由于本文提出的算法是基于 DeepLabV3+ 网络进行改进, 为了验证网络改进的有效性, 本节将改进模块逐步添加到基准网络中, 以评估其性能。

表 4 网络消融实验结果(%)

网络模型			普通场景数据集		复杂场景数据集	
BASE	FPN	AS-block	mAP	mIoU	mAP	mIoU
√			84.02	77.33	80.02	72.51
√	√		86.44	79.29	81.63	73.16
√		√	87.11	81.06	83.11	75.18
√	√	√	90.62	84.19	87.07	79.82

表 4 显示 AS-block 的特征融合能力显著提高了整个网络的预测准确性。这表明, 在处理复杂场景下倒塌建筑物图像时, 不同特征的互补和融合是至关重要的, 尤其在寻找重要目标时。值得注意的是, 对主干网络在使用 FPN 架构后对网络性能的改善效果不太明显, 但当与 AS-block 协同工作时, 改进效果显著。因此, 可以得出结论, 改进的每个部分不仅在个体层面有助于提高网络性能, 而且在整体集成时表现更为卓越。

3 结语

针对当前大部分算法在复杂场景中倒塌建筑物分割精度低的问题, 本文综合了 ASPP-block 与 SE-block 的优势, 设计了多级特征融合模块 AS-block, 并将骨干网络调整为轻量级 MobileNetV2。随后, 将骨干网络中各特征提取分支与解码模块进行级联整合, 形成一个高效而精确的分割模型。实验结果表明: 该模型在构建的复杂场景数据集下 mIoU 可达 79.82%, 模型大小 6.82 M, 并且 FPS 达到 15.81 帧/秒。总的来说, 所提出的算法在面对复杂场景的挑战时, 无论是在分割精度还是在计算效率方面, 均表现出色。本文的研究不仅为倒塌建筑物分割问题提

供了一种高效的解决方案, 还为更多类型复杂场景的分割问题奠定了基础。

参考文献:

- [1] AIXIA D O U, ZONGJIN M A, SHUSONG H, et al. Building damage extraction from post-earthquake airborne LiDAR data[J]. Acta Geologica Sinica(English Edition), 2016, 90(4): 1481-1489.
- [2] ZARYABI E H, KALANTAR B, MORADI L, et al. MSBDA-Net: multi-scale siamese building damage assessment network [C] // Proceedings of the 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Queensland, Australia, 2022: 1-6.
- [3] 杨玉荣, 张仕廉, 铁中用, 等. 震后建筑物受灾程度遥感监测方法研究[J]. 地震研究, 2018, 41(4): 630-636, 658.
- [4] CHENG X M, SHEN Z F, XING T Y. Damaged building extraction and rapid assessment for earthquake disasters based on high-resolution remote sensing images[J]. Journal of Natural Disasters, 2016, 25(3): 22-31.
- [5] TU J, LI D, FENG W, et al. Detecting damaged building regions based on semantic scene change from multi-temporal high-resolution remote sensing images [J]. ISPRS International Journal of Geo-Information, 2017, 6(5): 131.
- [6] TOBY T, GOPALAKRISHNAN U, RAO S N. A deeper CNN approach for detection of collapsed buildings in drone images [C] // Proceedings of the 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonipat, India, 2022: 404-410.
- [7] 叶昕, 秦其明, 王俊, 等. 利用高分辨率光学遥感图像检测震害损毁建筑物[J]. 武汉大学学报(信息科学版), 2019, 44(1): 125-131.
- [8] 葛小三, 陈曦, 赵文智, 等. 基于生成对抗网络的建筑物损毁检测[J]. 测绘学报, 2022, 51(2): 238-247.
- [9] LI L, LIANG J, WENG M, et al. A multiple-feature reuse network to extract buildings from remote sensing imagery[J]. Remote Sensing, 2018, 10(9): 1350.
- [10] CUI L, JING X, WANG Y, et al. Improved swin transformer-based semantic segmentation of postearthquake dense buildings in urban areas using remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,

- 2022,16:369-385.
- [11] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation [C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany. Springer International Publishing, 2015:234-241.
- [12] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C]//Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 2018:801-818.
- [13] LIU X, XU B, GU W, et al. Plant leaf veins coupling feature representation and measurement method based on DeepLabV3+ [J]. Frontiers in Plant Science, 2022,13:1043884.
- [14] 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, Honolulu, America, 2017: 2117-2125.
- [15] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation [EB/OL]. arXiv:1706.05587, 2017.
- [16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, America, 2018:7132-7141.
- [17] GUPTA R, HOSFELT R, SAJEEV S, et al.xBD: a dataset for assessing building damage from satellite imagery [EB/OL]. arXiv:1911.09296, 2019.

Complex scene segmentation of collapsed buildings algorithm based on improved DeepLabV3+

Zhong Liansheng, Tao Qingchuan*

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Existing segmentation methods for post-earthquake collapsed buildings suffer from low accuracy and real-time processing challenges due to factors like backlighting, thermal noise, and rain and fog interference. In response to these issues, a method is proposed for segmenting collapsed buildings in complex scenarios based on DeepLabV3+. Firstly, we select the MobileNetV2 backbone network as the feature extractor for DeepLabV3+ to reduce network computational parameters and address the slow segmentation speed. Subsequently, an AS block is introduced in the feature fusion stage to selectively enhance feature channel weight coefficients related to collapsed buildings, thereby improving the model's segmentation accuracy. Additionally, within the backbone network, FPN is introduced to integrate multi-level features in the decoding stage, effectively capturing information from various collapsed building ranges. The experimental findings indicate that this network has, to a certain extent, effectively tackled the challenges related to segmentation efficiency and accuracy in intricate scenarios involving collapsed structures.

Keywords: deep learning; image segmentation; collapsed buildings; complex scenes

文章编号: 1007-1423(2024)07-0044-06

DOI: 10.3969/j.issn.1007-1423.2024.07.007

基于注意力机制深度学习的轻量级无线网络数据分选方法

何靓华*, 赵 英

(南昌应用技术师范学院电子与信息工程学院, 南昌 330000)

摘要: 轻量级无线网络通常用于资源受限的设备, 使得数据传输和处理能力有限, 且数据量庞大, 特征识别较为困难, 导致数据分选精度较低。为此, 提出基于注意力机制深度学习的轻量级无线网络数据分选方法。计算数据的特征值, 对基础数据量化预处理, 采用领域粗糙集算法测算出数据分选的冗余限值, 多阶提取数据特征, 构建注意力机制深度学习网络数据分选流程, 采用最优筛选的方式来实现数据分选处理。最终的测试结果表明: 针对选定的 6 个数据分选测试周期, 注意力机制深度学习网络数据分选方法最终得出的数据分选平均 F-Score 均可以达到 85% 以上, 说明在注意力机制深度学习技术的辅助下, 当前所设计的数据分选方法的针对性更强、效率更高, 具有实际的应用价值。

关键词: 注意力机制; 深度学习; 轻量级无线网络; 数据分选; 分选方法

0 引言

物联网和智能设备的普及, 使得越来越多的设备被连接到网络中, 并产生大量的数据。这些数据需要进行准确和高效的处理和分析, 以便支持智能决策和应用程序的执行。轻量级无线网络通常由资源有限的设备和基础设施组成, 如移动设备、低功耗传感器、无线路由器等, 导致处理能力和存储容量有限, 无法承载大量的数据流量。但由于轻量级无线网络中数据具有高维度、动态性和异构性, 因此, 如何以高效和可靠的方式进行数据处理和分析, 在保证数据分选效果的同时, 减少对设备资源的占用, 提高网络的可扩展性和效率, 成为亟需解决的问题。

数据分选实际上是一种十分复杂且繁琐的工作, 尤其是轻量级无线网络数据的数据分选处理, 更是涉及多个关联性的环节, 较难把控, 导致整体的数据分选效果未达到预期的标准。针对轻量级无线网络的数据分选方式多设定为单向结构, 如文献[1]传统谱聚类算法下网络数据分

选方法与文献[2]传统高光谱网络数据分选方法, 这一类数据分选方式虽然能够实现预期的分选处理任务及目标, 但是缺乏针对性与稳定性, 在不同的背景环境下, 并不能依据数据的特征进行分类汇总与筛选剔除, 致使数据分选出现混乱、效率低、误差等问题^[3]。

针对以上研究存在的问题, 轻量级无线网络在日常的运行过程中形成的数据量较为庞大, 增加了数据分选工作日常的压力, 造成数据分选在回归处理的过程中难以控制, 成为影响最终分选结果精准性、可靠性的重要原因之一^[4]。为此提出基于注意力机制深度学习的轻量级无线网络数据分选方法。注意力机制下的深度学习主要指结合深度学习技术, 对可见性的信息进行部分识别筛选的一种方法, 将其与轻量级无线网络数据分选方法相融合, 一定程度上可以进一步扩大实际的数据分选处理范围, 综合性提升数据处理的综合性效率^[5]。此外, 在注意力机制深度学习的辅助与支持下, 当前设计了更加灵活、多变的数据分选结构, 更好地解

收稿日期: 2023-10-11 修稿日期: 2024-02-01

基金项目: 南昌应用技术师范学院校级课题(NYSJG2209)

作者简介: *通信作者: 何靓华(1996—), 女, 江西庐山人, 硕士研究生, 主要研究方向为深度学习、教学设计, E-mail: a15279270810@163.com; 赵英(1995—), 女, 贵州遵义人, 硕士研究生, 主要研究方向为强化学习

决了数据回归处理的问题，将数据分选存在的问题更为明确地表示出来，尽量加强对数据分选过程中归纳偏差的控制，提升分选精度和稳定性，推动相关网络与技术迈入一个新的发展台阶。

1 轻量级无线网络数据分选方法设计

1.1 数据量化预处理

一般情况下，轻量级无线网络的日常数据传输、共享量是较为庞大的，对应的数据整合处理压力也在不断增加^[6]。为解决这一问题，提升整体的数据分选效率，可以在设计分选方法之前，对基础数据量化预处理^[7]。数据量化预处理可以通过提取数据的特征值，即将数据转化为能够代表数据重要属性的数值，以减少数据维度，简化数据集，更方便进行后续的数据分选。

首先针对轻量级无线网络的复杂数据进行定向整合处理，设定多个数据处理周期，并实时采集对应周期阶段的数据、信息，汇总之后存储并备份在预期的位置之上^[8]。将采集的数据筛选、剔除、分类，建立一个对应的数据集，同时明确数据集内部汇总数据的对应特征^[9]。但由于轻量级无线网络中的数据实际属性以及单位均是不同的，对于后续的数据分选形成了一定的阻碍^[10]。为此，结合实际的量化处理需求和标准，划分出数据处理的特征区间，由此计算出数据的特征值，实现数据量化预处理，计算式如下所示：

$$D = (p \cdot I - \varepsilon \cdot u) \cdot \frac{1}{\varepsilon \cdot x} \tag{1}$$

式中： D 表示数据的特征值， p 表示数据分选定值， I 表示数据转换比， ε 表示分选剔除占比， u 表示权重值， x 表示数据分选转换差值。

将获取的数据特征值作为基础，结合特征值的变化，二次划分当前数据的特征区间。同时，依据当前的数据分类处理结果，为了确保数据在实际量化之后得到统一管理控制，利用数据特征值对数据按照隶属度进行排序，在当前的数据处理环境下，尽量减小数值转换后形成的误差，对数据进行归一量化处理，以便于后续的分析 and 分选处理。

1.2 数据特征多阶提取

由于轻量级无线网络中数据存在多样化的特性，使得数据分选涉及大量的特征，而单一采用数据特征值进行数据量化预处理，无法充分描述数据的信息，忽略了数据中更重要、更有区分度的特征。而传统的数据分选特征提取方式多为目标式提取，虽然可以实现预期的提取任务，但是缺乏稳定性与可靠性，且数据量庞大也无法逐一进行特征识别，导致整体的数据处理效果难以达到预期的水平。因此，进行数据多阶特征提取，有助于提高数据表达的丰富性和多样性，从而获取最优数据分选结果。

在进行特征提取之前，需要先去除数据中的不相关特征属性。冗余限值是用来衡量特征之间的相关性和重要性的指标。通过应用领域粗糙集算法，可以识别出在数据分选过程中不必要或冗余的特征，从而优化数据分选效果，提高筛选结果的准确性和可靠性。因此，根据数据量化预处理的特征值，采用领域粗糙集算法测算出数据分选的冗余限值 N ，如公式(2)所示：

$$N = D(m + n)^2 \times (n - \nu)y \tag{2}$$

式中： m 表示数据采集总范围， n 表示堆叠区域， y 表示去除冗余差值， ν 表示重复冗余处理领域。

完成对数据分选冗余限值的计算后，依据冗余限值的变动，划定出对应的冗余数据特征提取范围。为了避免数据丢失问题的出现，提升数据分选的可靠与精度，利用领域粗糙集算法，结合注意力机制深度学习原理，根据数据的不同特征阶段，设置数据分选的基础指标及参数，见表1。

表 1 数据分选指标参数设置

数据分选注意力机制深度学习下指标名称	第一特征阶段	第二特征阶段
特征权重值	16.35	18.57
数据分选查准率/%	80.35	92.11
最优解	11.35	17.24
数据分选残差值	4.2	3.1
冗余差值	1.09	0.97

结合表1，实现对数据分选注意力机制深度

学习下指标参数的设置。根据计算得到的每个特征的冗余限值,筛选出低于冗余限值的冗余特征,将其剔除。依据表1设置的数据分选指标参数,基于已筛选的特征集和设置的数据分选指标参数,进行数据特征的多阶提取。结合深度学习技术,调整特征分类的框架,利用多个层次、多个特征提取方法得到特征,并将其融合,生成最终的特征表示,实现数据特征多阶提取。

1.3 深度学习网络数据分选流程设计

通过对数据特征进行分选和分析,可以深入挖掘数据的内在规律和特征之间的关系,选择出对于目标变量最有影响的特征,从而实现特征选择和降维。对轻量级无线网络数据进行分选,可以剔除无效、冗余或异常的数据,减少网络传输负荷和资源占用,从而优化网络性能。因此,基于经过对数据多阶特征提取后,构建适当的数据分选流程,以实现数据的准确分类和分选。

结合注意力机制深度学习原理,设计轻量级无线网络数据分选流程。基于当前的数据分选需求,明确深度学习范围下制定的注意力机制覆盖范围,具体如下所示:

- (1) 进行轻量级无线网络数据的基础性采集,汇总整合以待后续使用。
- (2) 注意数分选的敏锐度。
- (3) 采集实时数据局部处理,合理分类存储。
- (4) 结合深度学习技术,在数据分选过程中对不稳定数据、信息进行剔除,保证实际分选效率。
- (5) 设置多头注意力处理结构和结构化注意力处理结构,对数据分选工作进行双向辅助处理。

结合数据多阶特征提取与上述制定的数据分选注意力机制覆盖范围,依据深度学习技术,设计对应的轻量级无线网络数据分选流程,如图1所示。

结合图1,完成对注意力机制深度学习网络数据分选流程的设计。按照上述的顺序和环节,对采集的数据进行单向处理,建立一个深度学习结合分解式的复合数据分选矩阵,设定矩阵的辅助控制指标参数,见表2。

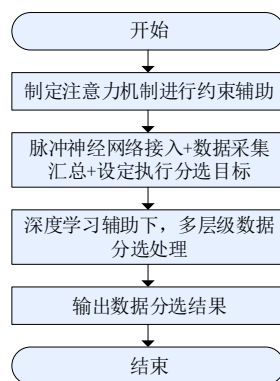


图1 注意力机制深度学习网络数据分选流程图

表2 数据分选流程辅助控制指标参数调整表

数据分选流程辅助控制指标	基础参数标准值	实际参数标准值
数据筛选匹配次数/次	6	12
实时特征值	21.06	25.48
数据最优逼近值	3.02	4.08
隶属度	8.5	9.7
目标函数	11.03	12.54

结合表2,实现对数据分选模型辅助控制指标参数的设置,以及上述的调整基本完成数据的分选处理,并将同类型的数据分类、存储在预设的位置之上。以此为基础,结合深度学习技术进行后续数据分选多维筛选处理层级,具体如图2所示。

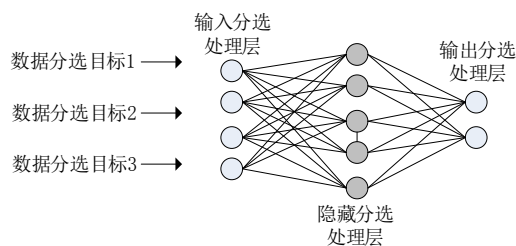


图2 深度学习下数据分选多维筛选处理层级图示

结合图2,实现对深度学习下数据分选多维筛选处理层级的设计,通过多个深度处理层级,过程中遵循注意力机制,对可控范围之内的数据进行同类型分类处理,同时设定与之匹配的分选结构,促使分选流程的内置分选框架逐步得到优化完善,进一步强化分选流程的数据分选综合能力。

1.4 数据分选最优处理

所谓最优分选处理，主要指在经过分选流程处理之后得出的数据分选结果，通过择优分类、定向筛选处理的方式进行分选层级划分，最终获取最佳分选结果，具体的分选结构如图3所示。

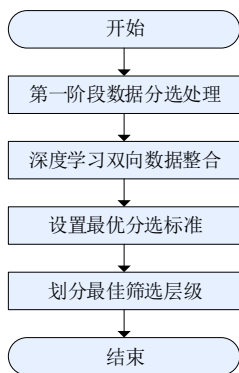


图3 最优分选结构图

结合图3，完成对最优筛选结构的设计。依据该流程进行扩展性数据分选处理。可以设计一个自动化的数据分选程序，完成数据分选之后，通过程序直接进行传输、存储，一定程度上进一步提升了数据分选处理的整体效率，强化最终的分选质量，使得数据分选具有真实性与可靠性。

2 方法测试

为了验证基于注意力机制深度学习的轻量级无线网络数据分选方法的实际应用效果，进行了分析与验证研究，考虑到最终测试结果的真实性与可靠性，采用对比的方式展开分析，选择F区域的轻量级无线网络作为测试的主要目标对象，将文献[1]传统谱聚类算法下网络数据分选方法与文献[2]传统高光谱网络数据分选方法，以及此次所设计的注意力机制深度学习网络数据作为分选测试组。结合注意力机制深度学习原理，对最终获取的测试结果比照研究，接下来，进行初始测试环境的关联搭建。

2.1 测试准备

结合注意力机制深度学习原理，对当前所选择的F区域的轻量级无线网络数据分选方法的实际测试环境进行基础搭建。将OPNET作为实验的软件环境，其是一种广泛应用于通信网络设计、性能评估的集成开发环境，具有可视化

工具和多个通信网络模型。OPNET提供了多种无线网络模型和数据分选算法，可以进行轻量级无线网络数据分选的仿真研究。选取KKD数据集作为本次实验的主要形式，同时构建与之匹配的数据集，从F区域轻量级无线网络之中，设定六个对应的测试周期，每一个周期的数据采集时间为48 h，获取的数据按照用途、属性或者类型进行分类存储，待后续使用。

在轻量级无线网络之中接入一个辅助性的数据处理平台，并在网络之中部署一定数量的监测节点，以便于实时数据、信息的及时获取。结合注意力机制深度学习原理，设计三个阶段的数据分选辅助测定环节，具体的测试控制指标见表3。

表3 数据分选辅助测定阶段环节指标参数设定表

数据分选辅助测定阶段环节指标	基础参数设定	可控边缘参数设定
权重值	3.2	2.6
数据分选阶段耗时/s	0.21	0.13
目标函数	+10.25	+14.25
最优值	19.6	27.5
数据分选限值	5.2	41
隐层节点数量/个	12	
调优差值	1.0261	0.92

结合表3，完成对数据分选辅助测定阶段环节指标参数的设定。结合注意力机制深度学习原理，设计一个多目标的数据分选处理结构，并将设计的数据分选环节导入当前的结构之中，与布设节点搭接，形成循环性的分选测试环境，以待后续的测试和实践性验证。

2.2 测试过程及结果分析

在上述搭建的测试环境之中，结合注意力机制深度学习原理，对选定的F区域轻量级无线网络数据进行分选处理测试。在深度学习的辅助下，设计多阶的数据处理层级，并利用学习目标作为引导，获取基础性的数据预处理结果，对六个周期的数据汇总整合，分类存储在初始接入的数据处理平台之中，以待后续的分选。结合随机森林算法和注意力机制，建立一个辅助性的数据分选矩阵，以数据集的形式对同类型、同属性的数据进行归类，并测算出数据分选的单向数据集训练时间 C ，如公式(3)所示：

$$C = \varpi \cdot W \frac{V + o}{Q \cdot V} \quad (3)$$

式中： ϖ 表示数据分选次数， W 表示数据分选单向数量， Q 表示分选平均值， V 表示重复分选区域， o 表示堆叠分选区域。在数据分选的单向数据集训练时间范围之内，在制定的注意力机制之中，通过深度学习技术对分类选定的数据作二次调整，如图4所示。

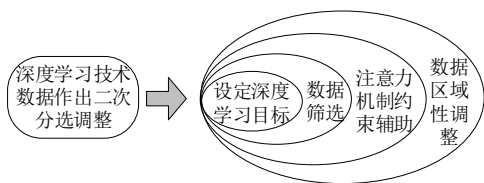


图4 深度学习数据作出二次调整流程图

完成对图4深度学习数据作出二次调整流程的设计之后，按照此顺序，对各个周期所采集的数据进行精准化的分选处理。

在轻量级无线网络数据分选中，数据集的类别分布可能存在不均衡的情况，即正例和负例的数量差异较大。F-Score相比其他评价指标对于不均衡数据集具有更好的鲁棒性，在评估数据分选的性能时更具可靠性。F-Score值越高对于数据分选的性能和可靠性更有意义。计算出最终数据分选的平均F-Score，如公式(4)所示：

$$L = G \times \zeta \times a \quad (4)$$

式中： L 表示数据分选的平均F-Score， G 表示数据分选覆盖范围， ζ 表示数据转换比， a 表示分选频次，结合当前测试，实现对测试结果数据的分析，见表4。

表4 测试结构数据比对分析表(%)

数据采集周期	文献[1]方法 的数据分选平 均 F-Score	文献[2]方法 的数据分选平 均 F-Score	所提方法的数 据分选平均 F-Score
周期1	75.2	71.9	89.3
周期2	52.3	80.3	90.2
周期3	76.1	81.6	87.6
周期4	86.9	70.2	92.4
周期5	80.1	79.8	88.8
周期6	75.6	84.9	91.6

结合表4，得出以下测试结论：针对选定的六个数据分选测试周期，对比于文献[1]方法与

文献[2]方法，所提方法最终得出的数据分选平均F-Score均可以达到85%以上，说明在注意力机制深度学习技术的辅助下，当前所设计的数据分选方法的针对性更强、效率更高，具有实际的应用价值。

3 结语

结合上述分析与研究，对基于注意力机制深度学习的轻量级无线网络数据分选方法的设计与验证分析。与初始的轻量级无线网络数据分选结构相比对，此次结合注意力机制深度学习原理所设计的数据分选方法更为灵活、多变，自身具有较强的针对性和稳定性。面对存在的数据分选问题，可以通过多层级、多目标、多结构的方式展开来，结合制定的数据分选规则，调整对应的数据分选阶段，为后续的数据处理工作的执行提供参考依据和理论借鉴。

参考文献：

- [1] 张震,胡贵恒,盖昊宇,等.基于谱聚类算法的高速网络数据流快速分类方法研究[J].齐齐哈尔大学学报(自然科学版),2023,39(5):24-30.
- [2] 崔利,秦浩然,姜海山.高光谱技术的复杂场景遥感测绘数据分类方法[J].激光杂志,2022,43(9):159-163.
- [3] 张宁.基于特征提取与聚类的医院档案数据分类方法[J].微型电脑应用,2022,38(9):45-47.
- [4] 陆骏.基于人工智能的网络复杂数据分类方法研究[J].信息记录材料,2022,23(8):195-197.
- [5] 尧雪莉.基于数据挖掘的应用型高校创新创业数据分类方法[J].数字通信世界,2022(6):194-196.
- [6] 韩磊,黄瑞龙,范文静,等.基于Weka平台和代价敏感特征选择的基因表达数据分类研究[J].智慧健康,2022,8(17):1-4.
- [7] 孟雅蕾,周千明,师红宇,等.基于改进ID3算法的数据分类方法[J].计算机仿真,2022,39(5):329-332,417.
- [8] 彭丽艳.虚拟存储产生容灾问题的数据备份关键技术分析[J].信息通信,2017(10):117-118.
- [9] 周兰兰,仇洪冰,周隼,等.基于数据场聚类与时差的雷达信号分选方法[J].桂林电子科技大学学报,2021,41(2):92-98.
- [10] 陈晶.基于数据驱动的网络安全数据分类方法[J].国外电子测量技术,2021,40(4):42-46.

(下转第75页)

文章编号: 1007-1423(2024)07-0049-05

DOI: 10.3969/j.issn.1007-1423.2024.07.008

基于改进的 CNN-BiLSTM 和三支决策的网络入侵检测方法

王 宇*

(安徽理工大学计算机科学与工程学院, 淮南 232001)

摘要: 针对网络入侵检测模型特征提取不足以及相关数据集不平衡问题, 提出一种基于卷积神经网络与双向长短期记忆网络的网络入侵检测方法。首先结合 SMOTE 过采样算法对 UNSW-NB15 数据集进行预处理; 其次建立基于 CNN-BiLSTM 的入侵检测模型, 提取数据集的局部特征和长距离依赖特征, 通过注意力机制进一步加强特征的重要性; 最后通过基于三支决策的分类器获得分类结果。实验结果表明, 所提方法在各个评价指标上均有所提升, 能够有效提高检测性能。

关键词: 入侵检测; 卷积神经网络; 双向长短期记忆网络; 三支决策

0 引言

随着互联网的不断发展和普及, 网络安全问题变得越来越突出。因此, 如何保障网络安全成为一个至关重要的任务^[1]。入侵检测系统是保护网络安全的重要手段之一, 它可以监控和识别网络流量中的异常和威胁行为, 并采取相应的措施来阻止攻击和恢复网络功能。目前入侵检测系统的研究和应用, 是网络安全保护的一个重要方向。

近年来, 深度学习技术在处理当前网络流量中的复杂数据具有突出优势^[2], 逐渐被应用到入侵检测算法中。关生等^[3]提出了一种卷积神经网络的入侵检测方法, 针对卷积神经网络待参数量过大、多层网络导致特征和梯度消失等问题, 增加网络深度, 减少参数数量来解决问题。武洋名等^[4]基于深度信念网络的自动提取特征的能力, 将极限学习结合提出了一种检测算法模型, 解决了数据不均衡以及收敛速度的问题。Tang 等^[5]使用 DCNN 构建检测系统, 同时利用 LSTM 提取信息。蔡宇航等^[6]提出了降噪自编码模型的概念, 以解决重要数据特征遗漏的问题, 进一步提高了算法的检测精度。谭

继安^[7]提出基于决策树的增量学习算法来识别数据的细粒度标签, 通过混合神经网络模型来提取数据特征, 解决传统深度学习在大规模网络入侵时存在的过拟合问题。吴启睿等^[8]提出了一种卷积神经网络和三支决策结合的入侵检测方法, 通过 CNN 的特征提取能力以及基于三支决策的分类方法达到特征提取充分以及分类精确的目的。李国燕等^[9]提出了一种居于并行多尺度融合特征的检测模型, 将残差卷积神经网络和 BiLSTM 相结合, 解决了网络流量中时空特征检测率低的问题。

目前入侵检测存在如下问题: ①数据集不平衡对模型训练的影响; ②在特征提取方面还存在一些不足, 因此在保证性能的前提下如何构建深层次的网络模型来获取更多特征, 还有待进一步研究。

针对上述问题, 本文将卷积神经网络和双向长短期记忆网络相结合, 建立入侵检测模型; 其次使用注意力机制对特征序列进一步处理, 引入过采样算法; 最后将提取到的特征通过基于三支决策理论的分类器进行分类。本文所提出的入侵检测方法在 UNSW-NB15 数据集上进行了实验对比, 验证了算法的综合性能。

收稿日期: 2023-10-17 修稿日期: 2023-11-29

作者简介: *通信作者: 王宇(1999—), 男, 安徽宣城人, 硕士研究生, 研究方向为入侵检测, E-mail: 2415319927@qq.com

1 相关工作

1.1 卷积神经网络

卷积神经网络(convolutional neural network, CNN)具有将入侵流量数据特征表示为高级特征的能力,同时还具备突出的特征学习能力。卷积神经网络有效解决了神经网络参数爆炸的问题,也保证了分类的准确性。在卷积神经网络中,局部感知是通过卷积层的卷积计算来实现的,卷积层由卷积核在输入数据上实现。

1.2 双向长短期记忆神经网络

双向长短期记忆神经网络(bidirectional long shortterm memory, BiLSTM)是在LSTM的基础上进一步发展的。它更全面地学习输入序列的特征。BiLSTM有两个LSTM层,分别按时间正序和时间逆序处理输入序列,然后将它们的输出拼接起来作为最终的输出。这种双向处理的方式使得模型能够同时考虑过去和未来的输入信息,来捕捉时间序列中的长期依赖关系,可以有效地保证检测结果的准确性。BiLSTM在求解时需要分别计算两个新的隐向量 $\vec{h}_t$ 和 $\overleftarrow{h}_t$ 的结果,同时将两个输出相结合就可以得到最终结果,计算公式如下:

$$\vec{h}_t = LSTM(x_t, \vec{h}_{t-1}) \quad (1)$$

$$\overleftarrow{h}_t = LSTM(x_t, \overleftarrow{h}_{t-1}) \quad (2)$$

$$y_t = \tanh(W_y [\vec{h}_t, \overleftarrow{h}_t] + b_y) \quad (3)$$

式中: W_y 代表每一层连接到前一个隐层状态的权重矩阵, $\vec{h}_t$ 和 $\overleftarrow{h}_t$ 代表正反两项LSTM在 t 时刻的输出, b_y 代表偏置值, y_t 是最终BiLSTM的输出结果。BiLSTM结构如图1所示。

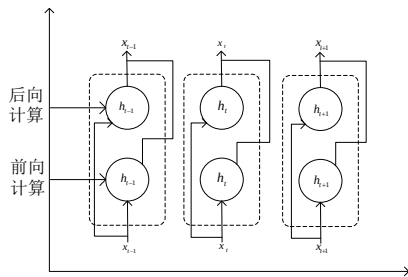


图1 BiLSTM模型结构图

1.3 三支决策

三支决策来源于粗糙集理论^[10],是一种决策方法,用于处理数据分类问题。在传统的二支决策的基础上,引入了延时决策的概念。当现有的特征无法充分判断数据的分类时,延迟决策可以避免将正常的数据误判为攻击数据,从而提高分类的准确性和可靠性。在三支决策中,数据可以被归为三类:正向数据、边界数据和负向数据。正向数据是指具有充分的特征,可以明确地归为某一类的数据;负向数据是指特征不足以确定数据归属的数据;边界数据则处于正向数据和负向数据的边界上,分类不确定。三支决策的决策集可以表示为 $D = \{D_P, D_B, D_N\}$,分别表示正向决策、边界决策以及负向决策。三支决策的代价函数见表1。

表1 三支决策的代价函数

	$X(P)$	$\neg X(N)$
P	λ_{PP}	λ_{PN}
B	λ_{BP}	λ_{BN}
N	λ_{NP}	λ_{NN}

由表1可以得到如下两个相关阈值的计算公式:

$$\alpha = \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PN})} \quad (4)$$

$$\beta = \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})} \quad (5)$$

2 基于CNN-BiLSTM和三支决策的网络入侵检测方法

2.1 系统的理论模型描述

为了实现更快的入侵检测研究,本文提出了一种基于CNN与BiLSTM融合神经网络的入侵检测模型。首先是数据预处理,将数据集中的原始数据进行数值化以及归一化,使其能够放入融合模型进行特征提取;其次将处理好的数据输入到融合模型中进行特征提取;最后将提取到的特征用三支决策的理论完成分类过程。本文提出的模型原理框图如图2所示。

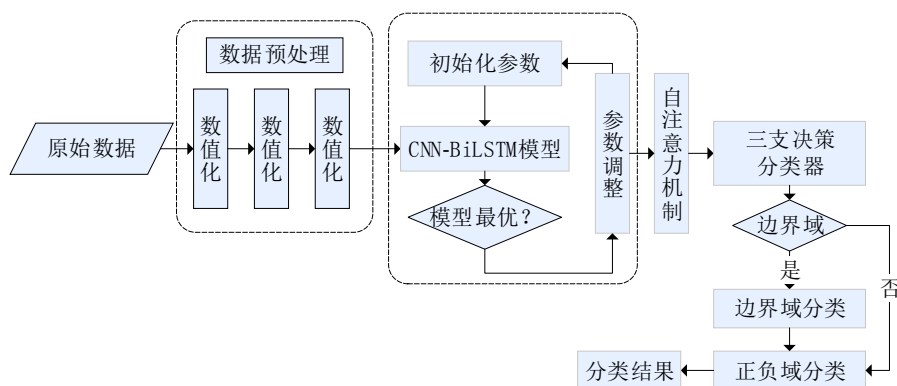


图2 入侵检测系统原理框图

2.2 数据预处理模块

为验证本文的入侵检测算法的性能，选用的数据集是 UNSW_NB15 数据集，包含有标签的真实正常流量和恶意攻击流量。本文做的是二分类，正常流量用0，攻击数据用1表示。为了更好地训练深度学习模型，需要对数据进行预处理，包括符号型特征转化为数值型特征、归一化等。预处理过程如下：

第一步，在数据的预处理阶段，首先需要将数据集中的符号型特征转化为数值型特征，以便模型训练的顺利进行。这里采用了 one-hot 编码的方法进行转换。

第二步，归一化处理。归一化可以统一各特征的取值范围，消除因取值尺度不同导致的模型训练时的偏差。将标准化后的数值归一化到[0,1]区间，计算公式如下：

$$x'_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (6)$$

式中： x_i 是标准化后的特征值， x'_i 是归一化后的特征值。

2.3 样本过采样

当训练集中存在类别不平衡问题时，使用 SMOTE 技术可以生成一些人工样本来增加该类别的样本数量，从而使得训练集更加均衡，避免训练模型时偏向于样本数量多的类别。具体步骤是：首先寻找少数类样本中的 k 个最近邻，从 k 个近邻中随机选择一个样本，再随机线性差值生成合成样本，重复此过程直至样本平衡。计算公式如下：

$$x' = x + r * (x_i - x) \quad (7)$$

式中： x' 是新样本， x 是已知的少量样本， x_i 是 k 个近邻样本中随机选的， r 是0~1之间的随机数。

2.4 基于改进的CNN-BiLSTM模型

融合神经网络模块主要由卷积神经网络、双向长短期记忆网络以及基于三支决策的分类器构成。CNN可以捕捉到输入数据中的局部模式和特征；同时，BiLSTM可以对序列数据进行建模，能捕捉到序列中的长距离依赖关系，通过在前向和后向两个方向上进行信息传递，能够更好地处理时序数据。自注意力机制可以加强模型对于重要特征的关注程度，通过自适应地调整特征的权重，从而提升模型的性能。基于三支决策的分类器可以有效地避免误分类的问题，通过设定不同的阈值，可以灵活地根据置信度进行样本的分类，从而降低误报率和漏报率，提高系统的检测性能。

3 实验设计与分析

3.1 实验环境

本实验采用了 AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz、NVIDIA GeForce 3060 和 16 GB RAM 作为硬件支持，操作系统为 Windows 10，采用 Python 3.7 实现本文算法。

3.2 评价指标

为了评估模型的检测性能，我们采用准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall) 和 $F1$ 值作为实验效果的评价标准。其中 TP 代表攻击数据的样本数量， FN 代表记录为攻击数据的正常数据的样本数量， FP 代表记录为正常

数据的攻击数据的样本数量， TN 代表正常数据的样本数量。计算式如下：

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (11)$$

3.3 实验结果分析

实验 1：有无自注意力机制的影响。从表 2 可以得知，本文模型在引入自注意力机制的情况下的性能要优于没有注意力机制的，可以说明自注意力机制可以提升重要特征的关注程度，达到提高性能的效果。

表 2 在 UNSW_NB15 数据集上不同研究的性能比较(%)

注意力机制	Accuracy	Precision	Recall	F1
有	95.82	96.53	95.75	96.29
无	92.71	94.74	92.35	93.52

实验 2：为了验证本文提出算法的检测效果，选择了较为传统的神经网络模型来进行比较，分别选择了 DNN、CNN、LSTM、BiLSTM 对本文数据集进行分类。结果见表 3。

表 3 在 UNSW_NB15 数据集上传统网络研究的性能比较(%)

模型	Accuracy	Precision	Recall	F1
本文	95.82	96.53	95.75	96.29
DNN	87.31	82.94	84.85	83.88
CNN	85.36	89.83	82.74	86.10
LSTM	91.42	91.94	90.11	91.02
BiLSTM	91.88	92.16	91.94	90.43

从实验结果可以看出，本文提出的分类算法在综合性能上优于传统的神经网络对比模型。可以说明，相较于传统的神经网络，本文提出的模型在特征提取以及分类方面要高于传统的神经网络算法。

实验 3：该实验主要是将提出的方法与其他研究者所提出的算法进行对比，实验选择的对比模型包括 BCNN 模型、GRU-MLP 模型、

SSL-3WD 模型、SEMI-GRU 模型。具体的对比实验结果见表 4。

表 4 在 UNSW_NB15 数据集上与其他研究模型的对比(%)

模型	Accuracy	Precision	Recall	F1
本文	95.82	96.53	95.75	96.29
BCNN	90.25	91.00	90.00	90.45
GRU-MLP	93.98	96.73	94.84	95.78
SSL-3WD	94.71	94.94	96.31	95.62
SEMI-GRU	88.11	96.29	81.56	88.13

实验结果表明，本文提出的模型在准确率方面达到了 95.82%、精确率方面达到了 96.53%、召回率方面达到了 95.75%、F1 值达到了 96.29%。综合四项指标来看，与其他模型相比，本文模型在 UNSW-NB15 数据集上取得了良好的检测效果。

综合上述实验结果，本文实验模型能够取得较好的性能，是利用了卷积神经网络可以有效解决局部特征丢失的问题，可以捕捉到输入数据中的局部平行特征，利用双向长短期记忆网络提取数据中的长距离依赖特征，较好地解决了单一神经网络特征学习不全面的问题。考虑到数据集前后平衡对当前实验的影响，最后使用基于三支决策的分类器有效地避免误分类的问题，通过设定不同的阈值，降低误报率和漏报率，提高系统的检测性能。

4 结语

本文基于卷积神经网络和双向长短期记忆网络构建了一个入侵检测方法，并使用 UNSW-NB15 数据集来进行验证。同时为了处理数据集中的不平衡问题，采用了过采样技术来提高分类效果。经过实验结果表明，实验模型在准确率、精确率上都有提升，达到了提高检测效果的目的。未来的工作可以包括在不同的新型网络入侵数据集上实现深度学习模型的多分类性能，以进一步验证所提出的模型的适用性和泛化能力。同时，也可以考虑将该模型应用到实际的网络安全场景中，并结合其他安全措施，构建更全面的网络安全防御系统。

参考文献：

- [1] GAO X, LIU L, ZHU X. Research on the main threat and prevention technology of computer network security[C]//IOP Conference Series:Earth and Environmental Science. IOP Publishing, 2021, 632(5):052065.
- [2] 肖建平,龙春,赵静,等. 基于深度学习的网络入侵检测研究综述[J]. 数据与计算发展前沿, 2021, 3(3):59-74.
- [3] 关生,周延森. 拆分降尺度卷积神经网络入侵检测方法[J]. 科学技术与工程, 2022, 22(36): 16108-16115.
- [4] 武洋名,宗学军,何戡. 基于数据增强的DBN-ELM入侵检测方法[J]. 科学技术与工程, 2022, 22(34): 15195-15202.
- [5] TANG K, ZENG C, FU Y, et al. Security analysis of social network topic mining using big data and optimized deep convolutional neural network[J]. Computational Intelligence and Neuroscience, 2022, 2022: 8045968.
- [6] 蔡宇航,廖光忠. 基于改进降噪自编码模型的网络入侵检测[J]. 计算机技术与发展, 2023, 33(2): 119-124.
- [7] 谭继安. 基于决策树和混合神经网络的大数据攻击增量检测研究[J]. 计算机应用与软件, 2022, 39(7):329-335, 349.
- [8] 吴启睿,黄树成. 结合卷积神经网络和三支决策的入侵检测算法[J]. 计算机工程与应用, 2022, 58(13):119-127.
- [9] 李国燕,周相茹,刘毅,等. 一种并行多尺度特征融合的入侵检测模型[J]. 火力与指挥控制, 2022, 47(11):29-35.
- [10] 鞠恒荣,李华雄,周献中,等. 基于Local约简的序贯三支分类器[J]. 计算机科学, 2017, 44(9): 34-39, 57.

Network intrusion detection based on improved CNN-BiLSTM and three-way decision

Wang Yu*

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

Abstract: Aiming at the problems of insufficient feature extraction of network intrusion detection model and imbalance of related dataset, a network intrusion detection method based on convolutional neural network and bidirectional long short-term memory network is proposed. Firstly, the UNSW-NB15 dataset is preprocessed by combining the SMOTE oversampling algorithm. Secondly, the intrusion detection model based on CNN-BiLSTM is established to extract the local features and long-distance-dependent features of the dataset, and the importance of the features is further strengthened by the attention mechanism. Finally, the classification results are obtained by the classifier based on the three-branch decision. The experimental results show that the proposed method is improved in each evaluation index and can effectively improve the detection performance.

Keywords: intrusion detection; convolutional neural networks; bidirectional long short-term memory networks; three-way decision

文章编号: 1007-1423(2024)07-0054-05

DOI: 10.3969/j.issn.1007-1423.2024.07.009

基于 NEDI 插值的图像分辨率提升算法

刘文字*, 郝慧艳, 郭楠, 张志伟

(中北大学信息与通信工程学院, 太原 030051)

摘要: 医学图像、遥感图像等图像受到成像设备等因素的限制, 一般分辨率较低。为得到高质量的高分辨率图像, 提出了对低分辨率图像采用 NEDI 插值方法提高图像的分辨率, 利用高斯滤波器进一步对高分辨率图像去噪的算法。将该算法应用于灰度图像的分辨率提升后, 不仅获得了边缘较为清晰的图像, 同时有效地抑制了边缘噪声, 图像视觉质量较好。

关键词: 图像插值; 图像去噪; NEDI 插值

0 引言

视觉作为人类获取外界信息的主要来源之一^[1], 图像质量对大多数基于视觉的应用性能至关重要。高分辨率图像中包含更多图像细节和信息, 因此在医学、遥感等实际应用中具有重要价值^[2-3]。提高图像分辨率主要有两种方法: 一种是通过增强成像设备性能来提高图像分辨率, 但这种方法会导致制造成本大量增加, 另一种方法是通过图像处理算法来提高图像分辨率, 这种方法成本低且不受硬件设备的制约。因此, 从算法方面来改进, 获得高质量的高分辨率图像是一个很有意义的研究方向。

提高图像分辨率的方法主要有插值法^[4-5]、基于学习的算法^[5-6]等。基于学习的算法能获得质量极高的高分辨率图像, 但运算时间较长, 实时性较差。因此本文使用图像插值方法提高图像分辨率, 但图像插值算法在提高图像分辨率的同时, 不可避免地在高分辨率图像中引入噪声, 使得生成的高分辨率图像质量较差, 所以需对生成的高分辨率图像进行进一步处理。本文针对低分辨率图像利用插值技术提高分辨率

和去除插值图像噪声两个方面进行了研究, 提出的方法能够得到较高质量的高分辨率图像。

1 图像插值方法

图像插值是通过提高低分辨率图像的分辨率和尺寸来生成高分辨率图像的一种方法^[4]。一些传统的线性插值方法, 如最邻近插值、双线性插值^[7], 双立方插值^[8-9]等, 这些方法算法简单、容易实现, 但由于插值像素的值是通过取相邻像素的平均值或卷积和来计算的, 在使用这些方法进行图像插值时, 通常会在生成的图像中出现模糊的边缘或楼梯伪影, 使得图像的主观质量大大下降。为克服这些问题, 国内外专家提出了基于图像边缘的插值方法^[1,10-12], 这些方法针对图像边缘特性采用了自适应插值, 能够更好地保护图像的边缘信息, 有效提高了插值图像的图像质量。基于边缘信息的边缘导向插值算法(new edge directive interpolation, NEDI)^[1]是其中一种较为流行的算法, 这种算法能提供主观质量很好的高分辨率图像。综上所述, 本文采用 NEDI 插值算法对低分辨图像进行插值。

收稿日期: 2023-11-01 修稿日期: 2023-12-02

基金项目: 山西省留学回国人员科技活动择优资助项目(20210039); 山西省留学回国人员科研资助项目(2022-146); 生物医学成像与影像大数据山西省重点实验室开放基金项目

作者简介: *通信作者: 刘文字(1996—), 女, 山西晋中人, 硕士研究生, 研究方向为图像处理, E-mail: 3352456559@qq.com; 郝慧艳(1980—), 女, 山西长治人, 博士, 副教授, 研究方向为光学探测、成像及信息处理; 郭楠(1998—), 女, 山西吕梁人, 硕士研究生, 研究方向为近红外漫射光血流成像技术; 张志伟(1964—), 男, 山西河津人, 教授, 博士, 研究方向为光电探测理论与技术

NEDI算法的基本思想是先计算低分辨率图像各像素点的局部协方差系数，然后再利用低分辨率图像与高分辨率图像协方差之间存在的几何对偶性来估计高分辨率图像的局部协方差系数^[1,13]。

设低分辨率图像为 $X_{i,j}$ ，低分辨图像放大两倍后得到的相应的高分辨率图像为 $Y_{i,j}$ ，则 $Y_{2i,2j} = X_{i,j}$ 。将NEDI插值分为两步，在第一步插值中利用原始的四个像素点 $Y_{2i,2j}$ 、 $Y_{2i,2j+2}$ 、 $Y_{2i+2,2j}$ 、 $Y_{2i+2,2j+2}$ 进行加权平均，得到待插值像素点 $Y_{2i+1,2j+1}$ 的值，公式如式(1)所示：

$$\hat{Y}_{2i+1,2j+1} = \sum_{k=0}^1 \sum_{l=0}^1 \alpha_{2k+1} Y_{2(i+k),2(j+l)} \quad (1)$$

第一步插值如图1所示。图中白点为原始像素点，黑点为待插值的像素点。

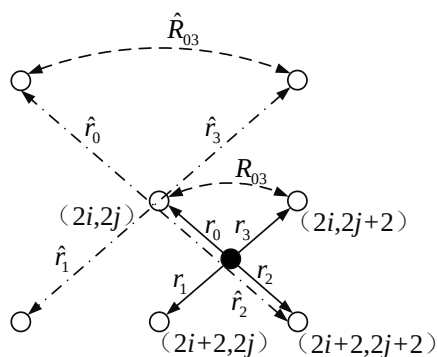


图1 第一步插值

将图像中的像素看作局部平稳的高斯过程，由经典的 Wiener 滤波理论^[14]可知，最优的最小均方误差 (minimum mean squared error, MMSE) 线性插值系数 α 可表示为

$$\alpha = R^{-1}r \quad (2)$$

其中： $R = [R_{kl}]$ ($0 \leq k, l \leq 3$), $r = [r_k]$ ($0 \leq k \leq 3$) 表示高分辨率图像中的局部协方差。高分辨率图像的相关信息是未知的，所以直接求解 R_{kl} 与 r_k 是行不通的，但可以利用低分辨率图像中 $\hat{R}_{kl}$, $\hat{r}_k$ 与高分辨率图像中 R_{kl} , r_k 存在的几何对偶性，将求解 R_{kl} , r_k 转化为求解 $\hat{R}_{kl}$, $\hat{r}_k$ 。

$\hat{R}_{kl}$, $\hat{r}_k$ 可采用标准方法在一个局部窗 (大小为 $M \times M$)^[1] 中计算得到：

$$\begin{cases} \hat{R} = \frac{1}{M^2} CC^T \\ \hat{r} = \frac{1}{M} Cy \end{cases} \quad (3)$$

其中： $y = [y_1 \cdots y_k \cdots y_{M^2}]^T$ 为一个局部窗中的 $M \times M$ 个点像素的数值向量； C 为 $4 \times M^2$ 数值矩阵，其中第 k 列向量为 y_k 对角线方向的最近邻 4 点像素，由公式(2)和公式(3)，可得到公式(4)：

$$\alpha = (CC^T)^{-1}(Cy) \quad (4)$$

公式(4)给出了最优的插值权值 α ，将 α 代入公式(1)中，即可得到待插值像素点 $Y_{2i+1,2j+1}$ 的值。

在第二步插值中，对 $Y_{2i+1,2j}$ 进行插值，依旧利用点 $Y_{2i+1,2j}$ 邻近的 4 个像素点进行插值，即

$$Y_{2i+1,2j} = \alpha_0 Y_{2i,2j} + \alpha_1 Y_{2i+1,2j-1} + \alpha_2 Y_{2i+2,2j} + \alpha_3 Y_{2i+1,2j+1} \quad (5)$$

通过观察公式(1)与公式(5)可以发现 $Y_{2i+1,2j}$ 的计算与 $Y_{2i+1,2j+1}$ 相似，只需按照第一步插值的方法就能对 $Y_{2i+1,2j}$ 进行求解。

第二步插值方法如图2所示，图中白色为原始像素点，灰色为第一步插值生成的像素点，黑色为待插值的像素点，通过观察图2可发现，第二步的插值中用到了第一步插值生成的像素点，这样的迭代计算会导致误差不断累计，最终导致插值后的图像边缘噪声比较大。

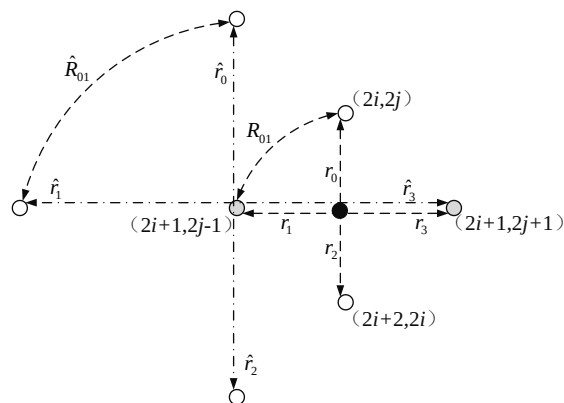


图2 第二步插值

2 图像去噪算法研究

使用NEDI插值算法对低分辨图像进行插值得到的高分辨图像边缘出现振铃效应，图像噪声较大，需要对图像进行去噪处理以进一步提高图像质量，图像去噪技术可分为线性滤波器和非线性滤波器^[15]，线性滤波主要包括均值滤

波、高斯滤波等，非线性滤波包括中值滤波、BM3D滤波等^[16-17]。

2.1 均值滤波

均值滤波是选用一个 $N \times N$ 的局部窗对图像进行卷积处理^[18]，即将待处理像素点作为中心像素点，以中心像素点邻域内各像素的平均值来代替该中心像素点原来的值。其定义公式可表示为公式(6)：

$$h[i, j] = \frac{1}{M} \sum_{(k, l) \in N} f[k, l] \quad (6)$$

其中： $h[i, j]$ 是滤波处理后图像的像素值， $f[k, l]$ 为原始图像的像素值， M 是邻域窗口 N 内的像素点总数。

均值滤波的滤波效果和图像的模糊程度都与邻域窗口 N 呈正相关，所以经均值滤波器处理后，图像中的噪点被去除的同时也会导致图像中大量细节信息丢失，造成图像边缘过度模糊。

2.2 高斯滤波

高斯滤波器是一种服从高斯分布的平滑线性滤波器，它的原理与均值滤波器类似，不同点在于待测处理像素点的值不用邻域内像素的平均值代替，而是用像素的加权平均值来代替。即在均值滤波器中邻域内所有像素的权重值是相同的，都为1，而在高斯滤波中，邻域内的像素权重值随着离中心像素点距离的增大而减小，这使高斯滤波在滤除噪声的同时，还能使图像边缘有较好的保真性，其定义公式可表示为公式(7)：

$$h(i, j) = \frac{\sum_{(k, l) \in N} f(k, l) w(i, j, k, l)}{\sum_{(k, l) \in N} w(i, j, k, l)} \quad (7)$$

其中： $w(i, j, k, l)$ 表示以 (i, j) 为中心像素点的邻域像素 (k, l) 的权重值，观察公式(7)，高斯滤波去噪效果的好坏关键在于选取合适的权重值。权重值可表示为公式(8)：

$$w(i, j) = \exp - \frac{(i - k)^2 + (j - l)^2}{2\sigma_d^2} \quad (8)$$

其中： σ_d 为高斯核大小。使用二维高斯函数作为滤波的权重值，使得高斯滤波器具有旋转对称性，即对图像每个方向上的平滑效果一样。 σ_d 决定了光滑的水平， σ_d 越大，表明滤波器的频带越宽，平滑结果越佳。

2.3 中值滤波

中值滤波是一种非线性的平滑滤波器，其基本原理为对中心像素点邻域内各像素进行排序，用其中值来代替中心像素点的值，它对非线性噪声的去除效果较好，在去除噪声的同时能有效保护图像边缘信息。其定义公式可表示为公式(9)：

$$h(i, j) = \text{Median}\{f(i - k, j - l), (k, l) \in N\} \quad (9)$$

2.4 三维块匹配算法滤波器

三维块匹配算法(block matching 3D, BM3D)是将图像分成一定大小的像素块，将像素块与相邻图块进行相似度匹配后得到一个三维的组合，然后对其进行滤波处理得到降噪后的图像块，并将降噪后的图像块聚合到原图像位置，最终生成去噪后的图像的方法。BM3D算法由有两个步骤组成。

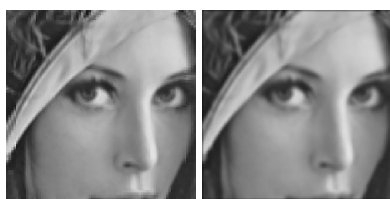
第一步：基础估计，首先将图片分成尺寸相同的图像块，在图片中搜索与目标块相似的图像块，搜索过程中可以使用欧氏距离来计算图像块相似性，将筛选出的相似图像块堆叠到一起形成3D图像区块；其次，将3D线性变换应用于生成的每一个3D图像区块，使用硬阈值进行系数收缩来衰减噪声，再通过逆变换得到初步处理后的图像^[17]；最后，按照上述步骤对所有图像块进行去噪处理，根据得到的块的质量，赋予不同的权值加权平均，得到去噪图像的基础估计。

第二步：最终估计，步骤和基础估计的方法类似，不同之处在于：①使用的是基础估计得到的图像，而不是原始图像；②不再使用硬阈值方法，而是使用维纳滤波处理生成的3D图像区块。

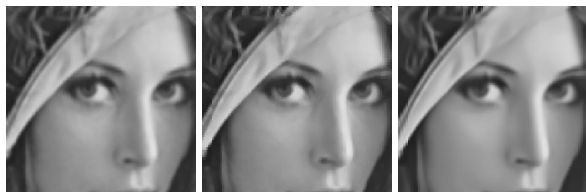
对低分辨率图像使用NEDI算法进行插值后，分别使用四种滤波器对NEDI算法插值后的图像进行去噪，去噪结果对比如图3、图4所示，表1为输出图像的信噪比。

表1 峰值信噪比对比

	均值滤波	高斯滤波	中值滤波	BM3D
测试图片1	27.3507	28.3110	28.4009	28.4814
测试图片2	25.5883	26.9319	27.1472	27.3553
平均值	26.4695	27.6215	27.7740	27.9184

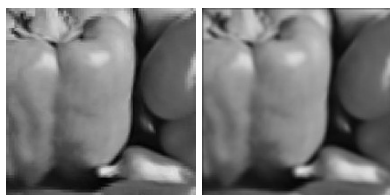


(a) NEDI插值图像 (b) 均值滤波



(c) 高斯滤波 (d) 中值滤波 (e) BM3D

图3 测试图片1经NEDI算法插值后的图像去噪结果



(a) NEDI插值图像 (b) 均值滤波



(c) 高斯滤波 (d) 中值滤波 (e) BM3D

图4 测试图片2经NEDI算法插值后的图像去噪结果

从图3、图4可以看出，对插值后的图像使用均值滤波后，原图边缘的振铃效应大部分被消除，但是图像整体比较模糊。而对图像使用BM3D滤波及中值滤波操作后，相比于均值滤波的效果，可以观察到生成的图像更加清晰，但原图中的振铃效应没有完全消除。在使用高斯滤波对图像进行滤波后，其图像清晰度接近BM3D滤波及中值滤波，同时图像边缘的振铃效应基本消除。如表1所示，通过使用四种滤波器对插值后的图像进行去噪，对比发现，均值滤波的峰值信噪比最小，平均值为26.4695，与其他三种滤波器的峰值信噪比相差较大；高斯滤波、中值滤波和BM3D的峰值信噪比分别为27.6215、27.7740、27.9184。三者相差不大，

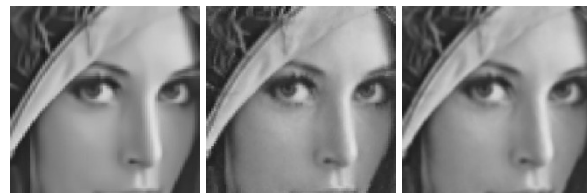
综合图像主观效果，这四种方法中高斯滤波对NEDI插值后产生的噪声更为有效。

3 实验

对两副测试图像进行比例为1:2的下采样，将下采样后的图像作为低分辨率图像分别使用最邻近、双线性、NEDI算法进行插值，使用高斯滤波对NEDI算法插值后的图像进行滤波处理，结果如图5、图6所示。



(a) 测试图像1 (b) 最邻近插值



(c) 双线性插值 (d) NEDI插值 (e) NEDI+高斯

图5 测试图像1结果对比



(a) 测试图像2 (b) 最邻近插值



(c) 双线性插值 (d) NEDI插值 (e) NEDI+高斯

图6 测试图像2结果对比

从图5、图6可以看出，使用最邻近插值算法进行插值后，插值后的图像出现了明显的“马赛克”效应，双线性插值方法插值后图像的细节受到损失，导致插值图像较为模糊，NEDI算法采用自适应边缘方向的插值方法，插值后

的图像有很好的边缘保留效果,但在边缘的交界噪声较大,出现振铃效应。而使用本文提出的NEDI与高斯滤波相结合的方法所生成的图像在保持边缘较为清晰的同时,降低了图像边缘的噪声,图像质量得到了进一步提升。

4 结语

本文在分析了常见的图像插值及图像去噪方法后,提出了一种NEDI插值与高斯滤波相结合的提高图像分辨率的方法,相较于传统算法,本文算法能够在保持图像边缘信息的基础上减少图像噪声,得到的高分辨率图像具有良好的视觉效果。

参考文献:

- [1] LI X, ORCHARD M T. New edge-directed interpolation[J]. IEEE Transaction on Image Processing, 2001, 10(10):1521-1527.
- [2] 曹春萍,李哲. 基于ResNet101多特征融合的新型冠状病毒感染图像分类方法[J/OL]. 小型微型计算机系统:1-7[2023-10-28]. <http://kns.cnki.net/kcms/detail/21.1106.tp.20230809.1358.012.html>.
- [3] YAN J H, YU L Q, XIA C X, et al. Super-resolution inversion and reconstruction of remote sensing image of unknown infrared band of interest [J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2023, 40(4):472-486.
- [4] 钟宝江,陆志芳,季家欢. 图像插值技术综述[J]. 数据采集与处理, 2016, 31(6):1083-1096.
- [5] 陈晓,荆茹韵. 单图像超分辨率方法综述[J]. 电子测量技术, 2022, 45(9):104-112.
- [6] 钟梦圆,姜麟. 超分辨率图像重建算法综述[J]. 计算机科学与探索, 2022, 16(5):972-990.
- [7] BLU T, THEVENAZ P, UNSER M. Linear interpolation revitalized [J]. IEEE Transactions on Image Processing, 2004, 13(5):710-719.
- [8] 郑心草,孙忠贵. 图像超分中双三次插值的非局部拓展[J]. 南京师范大学学报(工程技术版), 2023, 23(3):53-59.
- [9] KEYS R G. Cubic convolution interpolation for digital image processing[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1981, 29(6):1153-1160.
- [10] WANG Q, KREIDIEH R. A new orientation-adaptive interpolation method[J]. IEEE Transactions on Image Process, 2007, 16:889-900.
- [11] HWANG J W, LEE H S. Adaptive image interpolation based on local gradient features[J]. IEEE Signal Processing Letters, 2004, 11(3):359-362.
- [12] BATTIATO S, GIUFFRIDA E U, RUNDO F. A cellular neural network for zooming digital colour images[C]//Int Conf on Consumer Electronics(ICCE) Tech. Dig. New Jersey:IEEE, 2008.
- [13] 刘楠,毕笃彦,林嘉豪,等. 改进型快速NEDI图像插值实现算法[J]. 计算机应用, 2012, 32(7):1864-1867.
- [14] JAYANT N, NOLL P. Digital coding of waveforms: principles and applications to speech and video[M]. Englewood Cliffs, Prentice-Hall, 1984.
- [15] 常文斌,牟明任,贾海鹏,等. 基于OpenGL ES的图像滤波算法实现及优化研究[J/OL]. 计算机工程:1-15[2023-10-28]. <http://kns.cnki.net/kcms/detail/31.1289.TP.20230907.1346.003.html>.
- [16] 张娜娜,张媛媛,丁维奇. 经典图像去噪方法研究综述[J]. 化工自动化及仪表, 2021, 48(5):409-412, 423.
- [17] 刘利平,乔乐乐,蒋柳成. 图像去噪方法概述[J]. 计算机科学与探索, 2021, 15(8):1418-1431.
- [18] MAKOTO S, KUMAR P R, YOSHIKI K, et al. Improved iterative reconstruction method for compton imaging using median filter[J]. PLoS ONE, 2020, 15(3):e0229366.

Image resolution enhancement algorithm based on NEDI interpolation

Liu Wenyu*, Hao Huiyan, Guo Nan, Zhang Zhiwei

(School of Information and Communication Engineering, North University of China, Taiyuan 030051, China)

Abstract: Restricted by imaging equipment, human interference and other factors, most images have low resolution. In order to obtain high-quality high-resolution images, an algorithm is given to improve the resolution of images by using NEDI interpolation method for low-resolution images, and further denoising the high-resolution images by using Gaussian filter. After applying the algorithm to the resolution improvement of gray scale images, not only the images with clearer edges are obtained, but also the edge noise is effectively suppressed, and the visual quality of the images is better.

Keywords: image interpolation; image denoising; NEDI interpolation

文章编号: 1007-1423(2024)07-0059-05

DOI: 10.3969/j.issn.1007-1423.2024.07.010

基于深度学习的小目标检测综述

王聪敏^{1*}, 李秀丽²

(1. 华北水利水电大学信息工程学院, 郑州 450000; 2. 华北水利水电大学工程训练中心, 郑州 450000)

摘要: 该研究综述了基于深度学习的小目标检测技术, 首先, 梳理介绍了关键技术, 对不同方法进行了比较分析, 评估了它们的优劣势。其次, 讨论了相关数据集和评估指标, 为算法性能评估提供了坚实基础。此外, 探讨了小目标检测在智能交通和安全监控等领域的各种应用场景。最后, 提出了小目标检测的未来发展方向, 强调了提高检测精度和速度的必要性。

关键词: 小目标检测; 深度学习; 目标检测

0 引言

在计算机视觉领域, 目标检测主要涉及在图像或视频中特定目标的精确定位和识别。随着深度学习技术的发展, 目标检测方法主要分为 Anchor based、Anchor free。Anchor based 方法又分为单阶段和两阶段。从 2014 年到 2022 年目标检测的发展路线如图 1 所示。目标检测算法近年来从 RCNN 等两阶段模型发展到 YOLO 和 SSD 等单阶段方法。单阶段模型提高了效率和速度, 但精度问题仍然存在^[1]。为了解决这一问题,

新算法如 YOLO9000、YOLOv3 和 RetinaNet^[2] 已被引入, 改进了泛化能力和精度。虽然在检测大型和中型目标任务中取得了很大进步, 但小型目标体积小、特征不完整, 检测仍然困难。

本文旨在概述基于深度学习的小目标检测技术, 分析其关键技术和方法, 并介绍常见的小目标检测数据集, 分析各自特点和面临的挑战。与此同时, 还讨论小目标检测在自动驾驶、视频监控等多个领域的应用, 最后讨论了基于深度学习的小目标检测的未来发展方向。

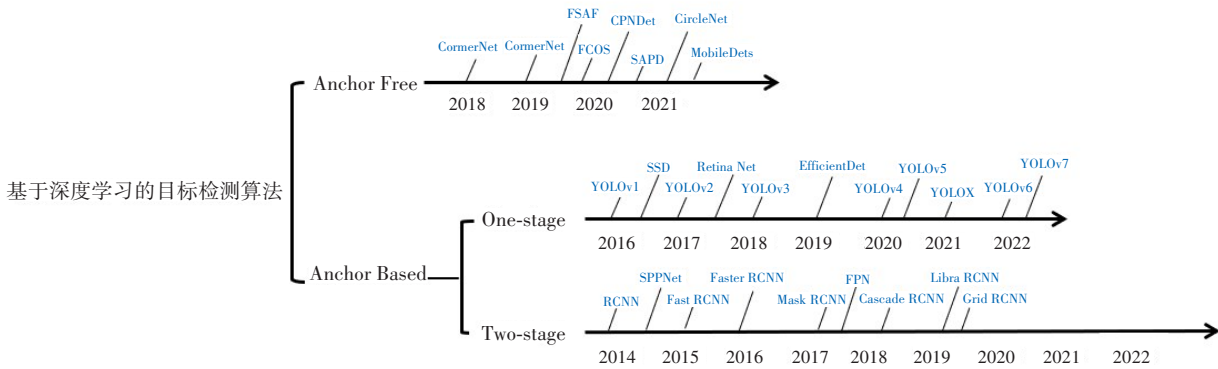


图 1 目标检测发展路线图

收稿日期: 2023-10-26 修稿日期: 2023-11-23

作者简介: *通信作者: 王聪敏(1998—), 女, 河南周口人, 硕士, 主要研究方向为图像处理、目标检测, E-mail: 3375925854@qq.com; 李秀丽(1967—), 女, 河南郑州人, 教授, 研究方向为人工智能、深度学习

1 小目标的定义和挑战

小目标主要根据两个标准分类。第一，绝对尺度定义，以目标的像素尺寸为基准。如在MS COCO数据集中^[3]，小目标是指尺寸小于 32×32 像素的目标；第二，相对尺度定义，小目标在图像中所占比例的相对大小。根据国际光学工程学会的标准，当一个目标在 256×256 像素的图像中所占的比例小于0.12%时，称为相对小目标^[4]。

正是因为小目标本身的特征，所以小目标检测面临多重挑战和困难。首先，由于数据集不足，获取高质量的小目标数据非常困难，这使得算法的训练受到限制。其次，传统特征提取方法通常不适用于小目标，深度学习方法可能会在处理小目标时失去关键特征信息。此外，小目标聚类 and 正负样本不平衡问题使算法难以应对目标重叠、遮挡和背景噪声。最后，传统损失函数难以准确计算模型误差，导致边界框的准确性下降。为克服这些问题，需要采用专门针对小目标检测的方法和技术，以提高算法性能，解决数据不足、特征限制和样本不平衡等问题。

2 关键技术和方法

2.1 数据增强

数据增强通过多种图像转换技巧来扩充和增强训练数据，以帮助模型更好地理解小目标的特征和变化。数据增强分为单样本增强和多样本增强。单样本增强操作包括翻转、裁剪、缩放、添加噪点和颜色变化等，以改变图像属性，增加数据多样性，提高模型泛化能力^[5]。如图2所示为单样本数据增强。



图2 单样本数据增强

多样本增强方法采用创新方式将多个图像合并成新的样本，从而扩展数据集。例如，Mosaic数据增强通过组合多个图像来增加背景信息，从而提高小目标检测性能^[6]。如图3所示为多样本数据增强。

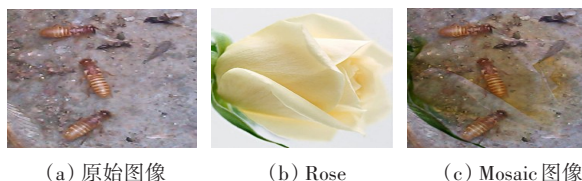


图3 多样本数据增强

2.2 多尺度检测

与传统目标相比，小型目标的像素较少，这增加了特征提取的难度。深度神经网络可能会导致小目标失去特征和位置信息，从而降低了精确检测的能力。为了解决这些问题，需要结合深层语义信息和浅层表征信息，多尺度检测通过整合不同尺度的特征数据，可以更好地捕捉小目标的细节和背景信息，从而提高检测性能。SSD算法通过对多尺度特征图进行预测，提高了小目标的检测性能，但仍有改进的空间，尤其是在缺少锚点和浅层特征中的语义信息方面。一些研究已经改进了SSD的训练方法，例如使用ResNet代替VGG模型^[7]，这有助于减少漏检和提高小目标检测性能。

2.3 上下文学习

小目标检测容易受到背景环境信息的干扰，因此添加上下文信息有助于增强特征表示和提高检测精度。研究采用不同方法实现上下文增强，如Tang等^[8]的上下文锚帧方法、Lim等^[9]的高级特征和上下文感知信息引入，以及Fu等^[10]的上下文推理方法。这些方法有助于模型更好地理解小目标与周围环境的关系，显著提高小目标检测的准确性和鲁棒性。

2.4 注意力机制

引入注意力机制是提高小目标检测性能的有效途径。SE模块允许网络自适应地调整特征通道权重，提高感知和检测性能。CBAM结合通道和空间注意模块，以自适应方式增强特征图权重，有助于提高检测性能和特征提取。自注意力机制通过捕捉长距离依赖性，增强感知和特征提取能力，广泛用于图像处理和目标识别。通道注意机制可增强不同通道的特征，基于注意力的候选区域生成方法则有助于提高小目标检测的准确性。这些方法使模型更智能地分配

资源，更有效地捕捉重要信息，显著提高小目标检测性能^[11]。

2.5 无锚机制

传统目标检测算法使用预定义锚点框，但难以适应小目标的大小和形状差异。无锚点框算法解决了这一问题，不依赖锚点框，直接预测目标边界框，使模型能自由预测大小和形状，适用于各种小目标^[12]。常见的无锚点算法如 CenterNet 和 CornerNet，CenterNet 通过预测中心位置、边界框和类别来定位目标，无需锚点框；CornerNet 通过角点识别实现目标定位，避免了复杂的锚点框架设计。这些方法提高了小目标检测的精确性和适应性。

从数据增强、多尺度检测、上下文学习、注意力机制、无锚机制对比小目标检测算法，结果见表1。

表1 小目标检测算法对比

关键技术	具体内容	优点	缺点
数据增强	通过随机裁剪、缩放、翻转扩大数据量	增加小目标的样本量	计算成本增加
多尺度检测	将浅层位置信息与深层语义信息相结合	融合不同尺度特征图的信息	难以避免干扰噪声的影响
上下文学习	特征图上下文信息融入训练	提升小目标的质量	速度较慢
注意力机制	对目标区域特征进行加权以抑制无关信息	增强对小目标特征的表达	计算复杂度大
无锚机制	根据关键点定位目标	几乎没有超参数	边框回归困难重重

3 小目标检测的数据集和评价指标

3.1 常用的小目标检测数据集

当前的小目标数据集有 MS COCO、Wider Face、TT100k、DOTA、VisDrone、WiderPerson、iSAID、TinyPerson、AI-TOD、SODA，这些数据集的注释如图4所示，基于名称、年份、数量、类别的小目标检测主流数据集见表2。

表2 小目标检测的主流数据集

数据集	年份	图像数量	类别
MS COCO	2014	800	10
Wider Face	2016	32203	60
TT100K	2016	100000	128
DOTA	2018	2806	15
VisDrone	2019	10209	10
WiderPerson	2019	13382	5
iSAID	2019	2806	15
TinyPerson	2020	1610	5
AI-TOD	2021	28036	8
SODA	2022	27214	18

3.2 评价指标

(1) IoU (Intersection over Union)，一个交集与并集比率。它用于评估预测结果与真实结果之间的相似性。在目标检测任务中，IoU 是预测边界框和真实边界框之间的交集面积除以它们的合并面积。



图4 数据集的注释

计算的公式如下:

$$IoU = \frac{A \cap B}{A \cup B} \quad (1)$$

通常,检测框的确定有一个阈值,称为IoU的阈值。当IoU值大于0.5时,就可以认为检测到目标。

(2) 准确率(Accuracy),表示符合标签的预测样本占样本总数的比例;精确率(Precision),表示正确预测正样本占实际预测为正样本的比例;召回率(Recall),表示正确预测正样本占正样本的比例;F1分数是精确率和召回率之和的调和平均。计算公式如下:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1-Score = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (5)$$

式(2)~(5)中:TP表示正样本被正确预测为正样本,FP表示负样本被错误地预测为正样本,TN表示负样本被正确预测为负样本,FN表示正样本被错误预测为负样本。

(3) mAP(Mean Average Precision),描述模型在不同类别中的平均性能。它的取值范围在0~1之间,取值越高意味着模型的性能越好。

(4) 每秒帧数(FPS),即算法每秒平均检测的图像数,用于显示算法的实时性能。

4 小目标检测的应用前景

4.1 自动驾驶

自动驾驶系统中,准确识别道路上的交通标志、行人、自行车和其他车辆是确保驾驶安全和做出正确决定的重要组成部分。小目标检测技术的实时性和精确性使自动驾驶系统能够更好地识别和适应复杂的交通环境。

4.2 视频监控和安全

在视频监控和安全领域,小目标检测技术可以用于重要设施、公共场所和城市交通监控系统,以快速识别和报警异常事件和行为,例如盗窃、遗失物品和交通违章,这将提高安全系统的效率和准确性。

4.3 无人机应用

无人机广泛用于环境监测、农业、救援等领域,其中就包括小目标检测和跟踪任务。例如,通过小目标检测技术,无人机可以快速、准确地定位和跟踪植物病虫害,提高农业生产效率,还可以在自然灾害中寻找受困人员,提供救援支持。

4.4 医学图像分析

在医学图像处理中,小目标检测有助于识别和定位医学图像中的小病灶或异常结构,如肿瘤、微血管和病变。这对于疾病的早期诊断和精确治疗至关重要,并可提高医学影像的诊断准确性和效率。

5 结语

本文简要概述了深度学习中小目标检测算法的重要性以及相关挑战。介绍了解决小目标检测问题的多种方法,如数据增强、多尺度检测、上下文学习和注意力机制等,这些方法可有效提升小目标检测性能和应用效果。虽然小目标检测取得了显著进展,但仍需解决一些问题,如特征融合、模型的泛化能力以及时间与准确性之间的平衡。鼓励未来研究在这些领域继续努力,以进一步提升小目标检测技术的实际应用性。

参考文献:

- [1] 张阳婷,黄德启,王东伟,等. 基于深度学习的目标检测算法研究与应用综述[J]. 计算机工程与应用, 2023,59(18):1-13.
- [2] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C] // Proceedings of the IEEE International Conference on Computer Vision, 2017:2980-2988.
- [3] LIN T, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context [C] // Proceedings of the 13th European Conference on Computer Vision, Zurich, Switzerland, Springer International Publishing, 2014:740-755.
- [4] XIE W L, ZHU D, TONG X X. Small target detection method based on visual attention [J]. Computer Engineering and Applications, 2013, 49(12):125-128.
- [5] INOUE H. Data augmentation by pairing samples for images classification [EB/OL]. arXiv: 1801.02929, 2018.
- [6] 董子平,陈世国,廖国清. 基于YOLOv5s的密集多

- 人脸检测算法[J]. 计算机工程与科学, 2023, 45(10):1838-1846.
- [7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. arXiv:1409.1556, 2014.
- [8] TANG X, DU D K, HE Z, et al. Pyramidbox: a context-assisted single shot face detector [C] //Proceedings of the European Conference on Computer Vision (ECCV). Cham:Springer, 2018:797-813.
- [9] LIM J S, ASTRID M, YOON H J, et al. Small object detection using context and attention [C] //Proceedings of the 2021 international Conference on Artificial intelligence in information and Communication (ICAIC), 2021:181-186.
- [10] FU K, LI J, MA L, et al. Intrinsic relationship reasoning for small object detection [EB/OL]. arXiv:2009.00833, 2020.
- [11] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018:7794-7803.
- [12] 潘晓英, 贾凝心, 穆元震, 等. 小目标检测研究综述 [J]. 中国图象图形学报, 2023, 28(9):2587-2615.

A review of deep learning based on small target detection

Wang Congmin^{1*}, Li Xiuli²

(1. School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450000, China;
2. Engineering Training Center, North China University of Water Resources and Electric Power, Zhengzhou 450000, China)

Abstract: This study reviews the deep learning-based small target detection techniques. Firstly, combing and introducing the key techniques, comparing and analyzing the different methods, and evaluating their advantages and disadvantages. Secondly, relevant datasets and evaluation metrics are discussed in the article, providing a solid foundation for algorithm performance evaluation. In addition, the article discusses various application scenarios of small target detection in fields such as intelligent transportation and security monitoring. Finally, the article proposes the future development direction of small target detection and emphasizes the necessity of improving detection accuracy and speed.

Keywords: small target detection; deep learning; target detection

(上接第8页)

Underwater object detection model based on improved YOLOv5

Zhang Zhengxin, Zhang Duzhen^{*}

(School of Computer Science and Technology, Jiangsu Normal University, Xuzhou 221116, China)

Abstract: In complex deep-sea environments, precise detection and identification of targets are crucial for marine exploration and conservation. This study optimizes the YOLOv5 target detection algorithm to enhance the detection effects and efficiency of underwater targets. To augment the feature extraction capability of the model, the critical convolutional structure of YOLOv5—the C3 module—has been replaced. Moreover, the SEAttention module, which utilizes a spatial attention mechanism, is introduced, enabling the model to focus more on recognizing important targets. For evaluation, EIoU is adopted as the standard to accurately measure the performance of target detection. Experiments are primarily conducted on the DeepTrash dataset. Compared to the original YOLOv5, the optimized model has achieved a 3.1% improvement in precision, a 1.8% increase in recall rate, and a 2.4% rise in mAP@0.5. Furthermore, the parameter count of the model has been reduced by 17.4% after optimization, thereby enhancing computational efficiency. Even when compared to the latest YOLOv8, the optimized model has demonstrated superior performance.

Keywords: underwater object; object detection; attention mechanism; YOLOv5; deep learning

实践与经验

文章编号: 1007-1423(2024)07-0064-07

DOI: 10.3969/j.issn.1007-1423.2024.07.011

基于边缘缓存的蜂窝网络的资源分配

潘雨, 华伟*

(四川大学电子信息学院, 成都 610065)

摘要: 物联网(IoT)的广泛应用和通信技术的飞速发展引起移动数据流量爆炸式增长, 给网络运营商造成巨大的能源消耗。因此, 提高能效是 5G 和 6G 时代面临的重要挑战。在支持缓存的蜂窝网络中做了资源分配, 通过将问题建模为混合整数问题, 可获得最优解。目标是在满足服务质量(QoS)要求的前提下, 联合优化基站发射功率和任务传输调度策略, 并探讨了缓存命中率对总能耗的影响。仿真结果表明, 文章方案在能耗方面表现良好, 并且提高缓存命中率有降低总能耗的作用。

关键词: 边缘缓存; 资源分配; 边缘云计算

0 引言

在物联网和 5G 愿景的推动下, 设备数量呈指数级增长。同时, 许多新的应用场景不断涌现, 使得传统云计算无法满足低延迟和高速度的新要求。在移动边缘云计算(mobile edge computing, MEC)中, 可以通过将网络中设备所需要的计算任务卸载到边缘服务器中进行计算, 这样能扩展用户设备的计算和存储能力, 并且节约移动设备的电量^[1]。其中 MEC 是物联网和 5G 架构的重要组成部分。虽然 MEC 架构解决了用户体验问题, 但物联网设备的大量增加导致移动数据流量快速增长, 从而大大增加了能耗。同时, 由于温室效应严重, 物联网网络的能耗优化问题亟待解决。文献[2]的研究表明, 提高基站的能效是降低蜂窝网络能耗的重要途径之一。近年来, 联网设备的数量一直在迅速增长, 预计到 2022 年底将达到 146 亿台^[3]。因此, 优化 MEC 架构下基站的能耗具有重要意义。

MEC 可通过提供本地缓存大大减少移动设备的延迟和能耗^[4]。因此, 为了减少内容传输延迟并提高服务质量, 移动运营商将越来越多的内容缓存到 MEC 服务器上^[5]。这就涉及到一个热点问题——资源分配。资源分配是提高物联网网络能效的重要方法之一。然而, 当考虑

到缓存时, 资源分配就变得更加复杂和具有挑战性。因此, 如何在高速缓存辅助网络中更有效地利用资源分配和高速缓存引起了很多人的研究兴趣。此外, 许多物联网设备都是移动的, 而现有的大多数研究在进行资源分配时都没有考虑用户的移动性。实际上, 如果能正确利用移动信息, 用户移动背景可能会带来有益的节能效果。因此, 我们在支持缓存的边缘蜂窝网络中引入了移动性感知。

利用移动背景的一个潜在方法是利用预测的未来信息做出资源分配决策。先前的研究表明, 预测用户的未来环境很容易, 尤其是未来位置^[6]。根据用户在不久的将来的位置, 可以更容易地预测他们的信道状态, 而边缘服务器可以利用这些预测来确保它们拥有适量的资源。

1 相关工作

MEC 可实现创新服务场景, 确保增强个人体验和优化网络运行, 可为用户提供更好的体验质量(QoE), 并通过将计算密集型任务迁移到 MEC 服务器来扩展无线网络边缘的计算能力^[7]。在支持 MEC 的网络中, 资源分配主要从两个方面进行研究, 即缓存和卸载。近年来, 研究人员专注于 MEC 网络中的缓存和资源分配或卸载

收稿日期: 2023-12-08 修稿日期: 2024-01-16

作者简介: 潘雨(1999—), 男, 四川绵阳人, 在读硕士研究生, 研究方向为嵌入式系统设计、MEC 资源分配; *通信作者: 华伟(1967—), 男, 山东聊城人, 副教授, 研究方向为电子与通信工程、MEC 资源分配, E-mail: hua23557@163.com

和资源分配^[8]。

有很多关于卸载和资源分配的研究^[9-14]，而关于缓存和资源分配的研究相对较少^[15-17]。对于那些考虑缓存和资源分配的研究中，有一些是为了减少延迟。例如，文献[15]提出了一种新模式，该模式可以大大减少请求互联网内容时的延迟。此外，这种新模式可通过对现有蜂窝系统稍加修改来实现。文献[16]的研究重点是车载网络，并提出了一种缓存辅助的更新和交付方案。文献[17]提出了一种动态资源分配策略，用于云边缘环境中延迟关键和计算密集型的应用。文献[18]和文献[19]都考虑了计算卸载、内容缓存和资源分配，但文献[18]的工作目标是最大化计算任务的总延迟消耗，而另一个关注在多用户缓存辅助的MEC系统中最小化能量消耗。此外，文献[20]、文献[21]将深度学习方法添加到缓存辅助MEC网络中，以减少延迟。文献[22]考虑了高速缓存问题，其重点是软件定义移动网络中的高效资源分配。然而，上述论文并未考虑移动感知。

还有一些文章考虑了移动感知。针对缓存内容，文献[23]提出了一种协调优化延迟和能耗的MEC视频缓存策略。该文还分析了用户请求数据，并利用神经网络模型预测用户后续时间片请求的内容，提前缓存用户请求。文献[24]基于移动性预测，标准的MEC应用移动性服务被扩展为支持预先缓存，该机制允许他们在移交前将多媒体内容预缓存在MEC服务器中。

在本文中，我们并不预测用户未来的位置。已经有大量的工作讨论如何预测行人和车辆未来的运动轨迹。换句话说，用户在未来的轨迹或位置可以被基站认为是可预测的。根据预测的移动用户的未来位置，我们可以得到用户与每个基站之间的距离，以及一段时间内的信道状态信息(CSI)。在此基础上，我们在支持缓存的蜂窝网络中做了资源分配，联合优化了基站发射功率和任务传输调度，并探讨了缓存命中率对总能耗的影响。

2 系统模型

2.1 网络概览

考虑一个含有 K 个活跃用户和一个覆盖范围为500 m的基站的时分单播蜂窝无线网络，将

第 k 个用户记为 U_k ，其中 $k = 1, 2, \dots, K$ 。这些用户在基站的覆盖范围内移动并向基站请求文件。基站通过核心网络与云端连接，在基站旁架设边缘服务器，该边缘服务器可以根据用户历史喜好、当前内容流行度等信息提前缓存一些文件，并且可以获取基站内活跃用户接下来 T 个时隙的移动轨迹，然后根据这些用户信息制定资源分配策略，实时控制基站的文件发送。为了方便模拟，我们假设每个数据包都有相同的长度，每个文件都由不同数量的数据包组成，故每个用户请求不同数量的数据包表征出其请求的文件的不同大小。例如当我们通过TCP请求一个视频文件时，原始文件将被分割成多个数据包，每个数据包的长度为1500字节，这是因为TCP的最大传输单元通常是1500字节。我们将第 k 个用户请求的数据包的数量记为 L_k ，其中 $k = 1, 2, \dots, K$ 。

在系统中，我们假设 K 个活跃用户同时向基站请求文件，基站收到请求后会将请求转入边缘服务器内处理，边缘服务器根据用户移动轨迹计算出对应位置的信道情况，然后以能耗最优原则给不同的用户任务分配时隙，完成资源分配策略。我们定义基站在一个时隙只能给一个用户传输一个数据包，故基站会根据边缘服务器做的时隙分配策略完成用户所请求的任务。值得注意的是，用户请求文件后，边缘服务器会首先检索本地服务器是否缓存该文件，若已缓存，那么可以直接发送给用户；若未缓存，那么需要从云端下载到本地后再发送给用户。系统模型图如图1所示。

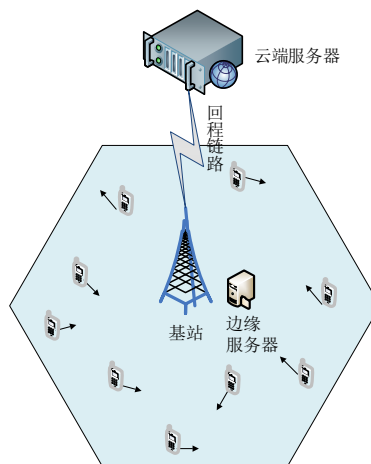


图1 系统模型

在我们的系统中, 假设每个文件都由多个大小相同的数据包组成, 并且每个时隙都可以传输一个数据包。TCP 是应用最广泛的传输协议, 它将每个长文件分割成多个长度为 1500 字节的数据包, 以便通过以太网进行传输。因此, 我们假设每个时隙需要传输 1500 个字节。在 GPRS 网络中, 数据传输速率通常从 56 kbps 到 114 kbps, 平均约为 72 kbps。协议 CDMA2000 1xRTT 的最大传输速率为 144 kbps, 平均传输速率也约为 72 kbps。因此, 我们选择一个时隙的持续时间为以 72 kbps 的速率传输 1500 字节所需的时间, 即 $1500 \times 8 / 72000 = 0.167$ s。我们知道, 与 4G 或 5G 的峰值速率相比, 这个速率相对较低, 但大多数应用程序不使用最大传输速率。并且随着物联网技术的发展, 越来越多的移动终端开始使用无线传输协议和 LoRa 协议, 其速率甚至可以低于 72 kbps。

在工作流程中, 一旦用户的未来发展轨迹可以被预测出来, 我们就可以开始做资源分配了。工作流程如下: 首先, 收集用户的请求数据, 并决定接下来 T 个时隙的资源分配。然后, 将结果进一步通知给基站以进行执行。在 T 个时隙后, 当前一批任务将完成。基站对下一段 T 个时隙重复资源分配决策过程。因此, 系统每 T 个时隙更新一次, T 的值取决于应用场景。如果对延迟敏感, 则应选择较小的 T 值, 否则 T 值可以较大。

2.2 信道模型与能量消耗

在无线通信系统中, 假设用户 U_k 接收信号时的信号与干扰加噪声比 (signal to interference plus noise ratio, SINR) 大于或等于一个特定阈值 γ , 那么代表该用户可以成功接收到基站发送的数据包, 否则就会产生丢包现象。我们定义用户 U_k 在第 t 个时隙接收基站信号的 SINR 为 S_{kt} , 并且表示为

$$S_{kt} = \frac{G_{kt} P_t}{n_0} \quad (1)$$

式中: G_{kt} 是用户 U_k 在第 t 个时隙的信道增益, P_t 是基站在第 t 个时隙时的发射功率。为计算方便, 我们定义一个常数 n_0 来表示传输过程中的噪声与干扰总和。设基站的最高发射功率为 $P_{\max}$, 则基站在每个时隙的功率不应大于 $P_{\max}$,

即满足 $P_t \leq P_{\max}$, 其中 $t = 1, 2, \dots, T$ 。

定义一个 $1 \times T$ 的矩阵 $\mathbf{P} = \{P_t\}$, 它表示基站在 T 个时隙的发射功率。为了方便计算, 本文将一个时隙的持续时间看为单位时间, 并且基站在一个时隙的功率视为不变, 那么基站在一个时隙的功率乘以该时隙的持续时间就可以得到基站在该时隙所消耗的能量, 即功率 P_t 又可以代表基站在 t 时刻所消耗的能量。因此, 基站在 T 个时隙内的发送文件所消耗的能量计算公式为

$$E = \sum_{t=1}^T P_t \quad (2)$$

如前所述, 若用户请求的文件已缓存, 那么直接通过本地服务器发送给用户, 这个过程只需要计算基站的发射功率, 不需要占用回程链路, 而当用户请求的文件未缓存在边缘服务器上时, 基站需要额外的能量去请求文件。在本文中, 未缓存的单个数据包的额外能耗我们设置为 P_t^c , 并且在做时隙分配策略时候, 未缓存的文件从云端取回的时间成本将会考虑在内, 即在基站从云端下载好内容之前, 这些用户是无法被分配时隙的。我们用二元变量 α_t 表示文件是否缓存。

$$\alpha_t = \begin{cases} 0, & \text{第 } t \text{ 个时隙传输的数据包未缓存} \\ 1, & \text{第 } t \text{ 个时隙传输的数据包已缓存} \end{cases} \quad (3)$$

加入未缓存文件所需额外能量后, 基站完成所有文件传输所需要的总能耗的计算公式为

$$E_{\text{总}} = \sum_{t=1}^T (P_t + (1 - \alpha_t) P_t^c) \quad (4)$$

如上一节所述, 基站需要向用户 U_k 发送 L_k 个数据包, 并且基站在一个时隙只能向一个用户发送一个数据包, 那么基站需要给用户 U_k 分配 L_k 个时隙。并且总时隙数必须大于总包数, 否则基站不能完成所有用户的任务。定义一个大小为 $K \times T$ 的矩阵 $\mathbf{X} = \{X_{kt}\}$ 表示一个用户是否是目标用户。 $X_{kt} = 1$ 则表明在第 t 个时隙, 用户 U_k 是基站的目标用户; 否则就不是目标用户。

我们称一个基站为了覆盖某一个特定用户所需要的最小功率为最小所需功率, 记为 W_{kt} 。定义一个大小为 $K \times T$ 的矩阵 $\mathbf{W} = \{W_{kt}\}$ 表示所有的最小所需功率。而当用户接收信号的 SINR

刚好等于 γ 时所需的基站发射功率最小。因此可以通过式(1)计算得到,如式(5)所示:

$$W_{kt} = \frac{\gamma n_0}{G_{kt}} \quad (5)$$

2.3 目标函数构建

针对上述问题我们构建如下目标函数:

$$\text{Minimize } \sum_{t=1}^T (P_t + (1 - \alpha_t) P_t^e) \quad (6)$$

Subject to

$$\text{C1 } P_t \geq W_{kt} X_{kt}, \text{ for } t \in [1, T], k \in [1, K];$$

$$\text{C2 } \sum_{t=1}^T X_{kt} \geq L_k, \text{ for } k \in [1, K];$$

$$\text{C3 } X_{kt} = 0, \text{ if } W_{kt} > P_{\max}, \text{ for } k \in [1, K], \\ \text{for } t \in [1, T];$$

$$\text{C4 } \sum_{t=1}^T X_{kt} \leq 1, \text{ for } t \in [1, T];$$

$$\text{C5 } \alpha_t \in \{0, 1\}, X_{kt} \in \{0, 1\}, \text{ for } t \in [1, T], \\ k \in [1, K];$$

文献[25]已经证明过无缓存的单播能量最小化问题是一个NP-hard问题,而无缓存的优化问题可以看作本文所有 $\alpha_t = 0$ 的特殊情况。因此,本文的优化问题可以归约为文献[25]中的单基站无缓存优化问题,从而证明本文的最小能耗问题也是一个NP-hard问题。

由此一来,此优化问题为一个混合整数规划问题,可以用一些标准的数学方法求解,如割平面法和分支定界法等。本文将该方法称为Mobility-Awareness and Cached with Mixed Integer Programming(MCMIP)。

3 实验分析

通过仿真平台对本文提出的方法进行了仿真,并与无移动性预测的传统调度算法(traditional algorithm, TA)进行了能耗比较,最后讨论了边缘服务器缓存内容的有效程度对能耗的影响。

在能耗性能比较方面,除了传输过程中的能源消耗外,我们认为该过程中涉及的其他成本主要包括基站从远程服务器检索数据包的成本。值得注意的是,无论是否使用资源分配,基站都需要从服务器上下载文件,从而产生成本,而当用户请求的文件未缓存在边缘服务器上时,基站需要额外的能量去请求文件。此外,

即使下载相同数量的文件,由于使用了缓存机制,我们也在调度基站以较低的成本下载文件时有更大的灵活性,例如在低负载或拥塞期间。因此,我们关注于用户请求开始后基站的能耗,不包括基站提前缓存文件所耗费的能量。

首先令 $T = 100$,测试用户数量 K 从20增大到100,结果如图2所示。从图2可以看到,当 T 保持不变,在 K 从20增大到100时,MCMIP能量消耗始终远低于传统方案TA。当 $K = 50$ 时,相较于传统算法TA,MCMIP能节省能量70.8%。当 $K = 100$ 时,相较于常规的TA,MCMIP能节省能量68.1%。这是由于当用户数量增多时,边缘服务器需要分配更多的时隙完成任务,但能耗较小的时隙已经被分配了,所以当任务数增多时,服务器需要选择能耗较大的时隙来完成增加的任务,这样就造成了平均能耗的上升。

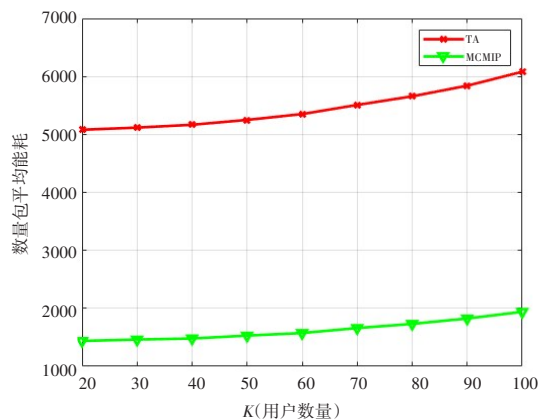


图2 用户数量对能耗的影响

第二组实验我们探讨时隙数对能耗的影响,令 $K = 80$,测试时隙数 T 从60增大到200,结果如图3所示。从图3可以看到,当 K 保持不变,在 T 从60增大到200时,MCMIP能量消耗依旧远低于传统方案TA。当 $T = 60$ 时,相较于常规的TA,MCMIP能节省能量63.1%。当 $T = 200$ 时,相较于常规的TA,MCMIP能节省能量75.1%。由于MCMIP是考虑整体情况的算法,当时隙数增多时,边缘服务器在分配时隙时就有更多能耗较低的时隙可以选择,所以当时隙数增多时,对MCMIP带来的增益也越来越大。对TA算法而言,因为TA是按照时间顺序分配时隙,后面增加的时隙基本不会影响算法的决

策,故TA算法的能耗几乎不受时隙数的影响。

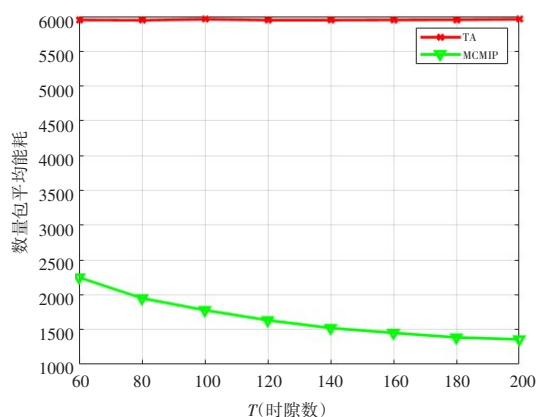


图3 时隙数对能耗的影响

前面的随机性实验充分测试了在不同状况下的资源优化结果,接下来我们讨论的是缓存命中率对算法的影响。缓存未命中会导致基站花费额外的开销利用回程链路去云端下载文件并传输,故缓存命中率越高,基站花费的额外开销会越低。因此,这里会测试在其它条件相同时,缓存命中率不同对能耗的影响。实验结果如图4所示。

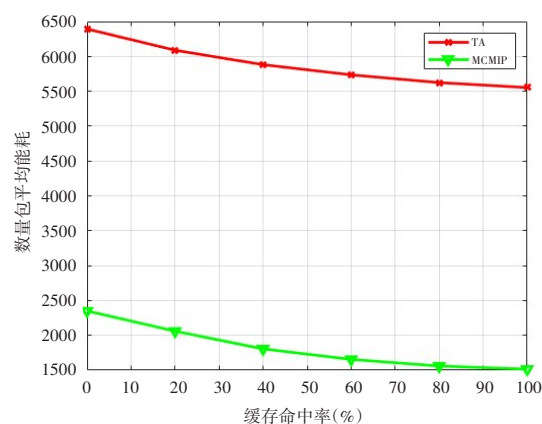


图4 缓存命中率对能耗的影响

从结果可以看到, MCMIP和TA的能耗都随着缓存命中概率的下降而上升,当缓存命中率为100%时,相当于所有用户请求的文件均被提前缓存至本地边缘服务器,在传输文件时基站会将本地的缓存文件直接发送给用户,此时对回程链路是不会产生负担的。而当缓存命中率为0%时,这代表所有的文件均未被缓存至本地

服务器,基站在传输文件时将占用回程链路下载文件,造成额外开销。可以看到,针对所提算法MCMIP,当缓存命中率为0%时,消耗的平均能耗比命中率为100%时多了55.1%。

4 结语

本文在支持缓存的蜂窝网络中做了资源分配,通过将问题建模为混合整数问题来获得最优解。目标是在满足服务质量(QoS)要求的前提下,联合优化基站发射功率和任务传输调度策略,并探讨了缓存命中率对总能耗的影响。根据仿真结果,所提算法MCMIP比传统的无任务传输调度的算法TA节省能耗最高达到75.1%,表明了引入传输调度机制显著降低能耗;而MCMIP的收益会随着时隙数的增加而增加,会随着用户数的增加而略微降低。另外本文还证明了缓存命中率的提高会降低所提算法MCMIP的能耗。

参考文献:

- [1] 黄星,殷锋,袁平. MEC系统中的资源分配算法研究综述[J]. 现代计算机,2020,(11):21-23,29.
- [2] FEHSKE A, FETTWEIS G, MALMODIN J, et al. The global footprint of mobile communications: The ecological and economic perspective[J]. IEEE Communications Magazine, 2011, 49(8):55-62.
- [3] ZHANG Z, XIAO Y, MA Z, et al. 6G wireless networks: vision, requirements, architecture, and key technologies[J]. IEEE Vehicular Technology Magazine, 2019, 14(3):28-41.
- [4] AI Y, PENG M, ZHANG K. Edge computing technologies for Internet of Things: a primer[J]. Digital Communications and Networks, 2018, 4(2):77-86.
- [5] VU T X, CHATZINOTAS S, OTTERSTEN B, et al. Energy minimization for cache-assisted content delivery networks with wireless backhaul[J]. IEEE Wireless Communications Letters, 2017, 7(3):332-335.
- [6] NADEMBEGA A, HAFID A, TALEB T. A destination and mobility path prediction scheme for mobile networks[J]. IEEE Transactions on Vehicular Technology, 2014, 64(6):2577-2590.
- [7] LI J, GAO H, LV T, et al. Deep reinforcement learn-

- ing based computation offloading and resource allocation for MEC [C] //Proceedings of the 2018 IEEE Wireless Communications and Networking Conference(WCNC), 2018:1-6.
- [8] LI S, ZHANG N, LIN S, et al. Joint admission control and resource allocation in edge computing for internet of things [J]. IEEE Network, 2018, 32 (1) : 72-79.
- [9] LIU J, MAO Y, ZHANG J, et al. Delay-optimal computation task scheduling for mobile-edge computing systems[C]//Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), 2016:1451-1455.
- [10] FANG F, WANG K, DING Z, et al. Energy-efficient resource allocation for NOMA-MEC networks with imperfect CSI[J]. IEEE Transactions on Communications, 2021, 69(5):3436-3449.
- [11] HU H, SONG W, WANG Q, et al. Mobility-aware offloading and resource allocation in MEC-enabled IoT networks[C]//Proceedings of the 2020 16th International Conference on Mobility, Sensing and Networking(MSN), 2020:554-560.
- [12] SALEEM U, LIU Y, JANGSHER S, et al. Mobility-aware joint task scheduling and resource allocation for cooperative mobile edge computing [J]. IEEE Transactions on Wireless Communications, 2020, 20(1):360-374.
- [13] HU H, WANG Q, HU R Q, et al. Mobility-aware offloading and resource allocation in a MEC-enabled IoT network with energy harvesting[J]. IEEE Internet of Things Journal, 2021, 8(24):17541-17556.
- [14] LI C, WANG H, SONG R. Mobility-aware offloading and resource allocation in NOMA-MEC systems via DC [J]. IEEE Communications Letters, 2022, 26 (5):1091-1095.
- [15] HSU H, CHEN K C. A resource allocation perspective on caching to achieve low latency [J]. IEEE Communications Letters, 2015, 20(1):145-148.
- [16] ZHANG S, LI J, LUO H, et al. Towards fresh and low-latency content delivery in vehicular networks: An edge caching aspect[C]//Proceedings of the 2018 10th International Conference on Wireless Communications and Signal Processing(WCSP), 2018:1-6.
- [17] TANG H, LI C, BAI J, et al. Dynamic resource allocation strategy for latency-critical and computation-intensive applications in cloud-edge environment [J]. Computer Communications, 2019, 134:70-82.
- [18] ZHANG J, HU X, NING Z, et al. Joint resource allocation for latency-sensitive services over mobile edge computing networks with caching[J]. IEEE Internet of Things Journal, 2018, 6(3):4283-4294.
- [19] CUI Y, HE W, NI C, et al. Energy-efficient resource allocation for cache-assisted mobile edge computing [C]//Proceedings of the 2017 IEEE 42nd Conference on Local Computer Networks (LCN), 2017:640-648.
- [20] YANG Z, LIU Y, CHEN Y, et al. Deep reinforcement learning in cache-aided MEC networks [C] // Proceedings of the 2019 IEEE International Conference on Communications(ICC), 2019:1-6.
- [21] NATH S, WU J. Deep reinforcement learning for dynamic computation offloading and resource allocation in cache-assisted mobile edge computing systems [J]. Intelligent and Converged Networks, 2020, 1 (2) : 181-198.
- [22] LIANG C, HE Y, YU F R, et al. Energy-efficient resource allocation in software-defined mobile networks with mobile edge computing and caching [C]//Proceedings of the 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2017:121-126.
- [23] LI C, ZHANG Y, SUN Q, et al. Collaborative caching strategy based on optimization of latency and energy consumption in MEC [J]. Knowledge-Based Systems, 2021, 233:107523.
- [24] LENTISCO C M, BELLIDO L, GONZÁLEZ-SÁNCHEZ D, et al. Extending the MEC mobility service to support cache prefetching [C] //Proceedings of the 2022 18th International Conference on the Design of Reliable Communication Networks (DRCN), 2022:1-4.
- [25] HUANG L, SUNG C W, CHEN C S. Context-aware wireless broadcast for next generation cellular networks [C]//Proceedings of the 2014 IEEE International Conference on Communication Systems, 2014:349-353.

(下转第95页)

文章编号: 1007-1423(2024)07-0070-06

DOI: 10.3969/j.issn.1007-1423.2024.07.012

融合 LDA-LSTM 算法的微博档案关注度和情感分析

孙思怡¹, 王家强², 罗子江^{3*}(1. 贵阳人文科技学院经济与管理学院, 贵阳 550025; 2. 贵州财经大学信息学院, 贵阳 550025;
3. 顺德职业技术学院智能制造学院, 顺德 528333)

摘要: 为了解析《档案法》实施三年后社会公众对档案的关注度与情感态度, 利用 Python 工具和 LDA 模型对数据进行提取和主题聚类, 获得不同时间下档案热点主题; 采用 LSTM 模型得到各档案主题的情感倾向, 分析用户产生不同情感倾向的原因。根据各主题间的联系, 得到档案项目、影视娱乐档案、学生-学校档案以及专项档案四类主题。各类档案主题具有较高的积极倾向, 说明公众对档案事业的发展比较支持和理解, 对于消极倾向较高的学生-学校档案类别, 有关部门应加强档案宣传教育, 从根本上保障公众利用档案的权利, 增强全社会的档案意识。

关键词: 档案关注度; LDA-LSTM 算法; 情感分析; 新浪微博

0 引言

档案是人类文明传承的载体, 同时也是人们精神文化的产物^[1]。随着数据信息化观念和新《档案法》的提出, 国家鼓励社会公众参与到档案事业的建设过程中, 使得公众不仅作为档案的“消费者”, 同时还是档案的“生产者”和“维护者”。在“互联网+”时代, 社交媒体由于具有开放性、互动性、多元性等特性, 为传播档案文化提供了新途径。微博作为具有代表性的社交媒体, 其信息传播高效、方式多样, 信息效应明显和观点表达真实, 深受青年人的欢迎^[2]。微博作为当今主流的社交媒体, 能够实时将信息通过裂变式进行传播, 实现信息即时分享、获取和交互。相比于其他社交媒体平台, 在微博上更容易形成意见、观点和情感的聚集, 并且信息传播路径具有一定的可视性, 因此本文选取微博作为数据源。

1 相关工作

很多学者依据相关理论和模型, 从不同角度对档案领域的关注度和情感分析进行探索。

易黎等^[3]为有效解决档案数据在构建知识图谱时出现多源异构的现象, 利用深度学习技术对数据进行结构化转变, 实现全域档案数据间的知识融合。Lian 等^[4]将扎根理论应用到数字档案中, 通过半结构化访谈方式发现档案管理制度、档案认知度等因素会影响数字档案的整合。周娟娟等^[5]利用文献连续体理论, 分析影响档案数据长期保存的因素, 并构建了能使档案数据长期保存的模型。王杰^[6]通过分析档案信息传播过程中面临的困境, 提出将新媒体技术与档案信息传播融合应用的策略及其实现路径。就目前的研究来看, 多以档案的利用、保护和传播为主, 鲜少以微博文本为数据源, 从用户视角去探究公众对档案领域的关注情况以及情感态度。同时, 通过对公众参与档案馆志愿服务的调查发现, 公众大多只了解传统档案, 对新式档案的认知水平和重视程度偏低、档案意识薄弱, 档案素养存在较大的提升空间^[7-8]。国外主要通过学科教育来提升档案素养, 而我国对档案素养的研究起步较晚且比较分散^[9]。

因此, 本文将主题模型和情感分析有效结

收稿日期: 2024-01-26 修稿日期: 2024-03-14

作者简介: 孙思怡(1993—), 女, 贵州贵阳人, 硕士, 助教, 研究方向为经济学、管理学、市场营销; 王家强(1998—), 男, 安徽阜南人, 硕士, 研究方向为信息服务; *通信作者: 罗子江(1980—), 男, 贵州遵义人, 教授, 硕士生导师, 研究方向为信息服务, E-mail: lah52000@sina.com

合,以微博文本为数据源,借助 LDA 模型识别档案主题,根据关注度算法计算各档案主题的关注度,得到各主题档案的变化趋势,最后利用 LSTM 模型分析各档案主题的正负情感,发现档案工作中存在的不足,进而提高公众在使用档案过程中的满意度和服务体验,同时为档案学相关领域专家学者的后续研究提供理论依据和研究方向。

2 研究方法

本文从用户视角下,以微博文本为数据源,利用 LDA-LSTM 模型算法深入挖掘微博文本中的潜在信息,得到微博档案主题关注度及其情感倾向。该方法共包括数据采集与预处理、LDA 主题挖掘-关注度分析和微博文本情感分析三个部分。

(1) 数据采集与预处理。利用 Python 和 XPath 语言编写爬虫,爬取新浪微博上关于“档案”的微博文本。剔除重复文本、无效文本数据,采用 jieba 工具对微博文本进行分词,通过构建自定义分词表,对档案领域专有名词精确分词。然后,利用哈工大停用词表并结合档案领域的特点,过滤停用词。

(2) LDA 主题挖掘-关注度分析。首先通过困惑度确定最优主题数,保证主题的拟合性和聚类效果明显,其次利用 LDA 模型对微博文本进行档案主题和关键词抽取。再对各档案主题的关注度进行计算,得到各档案主题关注度随时间变化趋势。

(3) 微博文本情感分析。运用 LSTM 模型对分类后的微博文本进行情感倾向分析,并借助可视化的形式展现各档案主题下情感分布情况。对各档案主题结果分析,探究档案主题关注度变化的原因,为档案建设和管理提供建议。

3 实证研究

3.1 数据准备

新《档案法》自 2021 年 1 月份正式实施以来,在社会上引起了广泛关注,推进依法治档、管档工作的同时,也增强了公众档案法治意识和认同感。为了更好了解公众对档案领域的认知水平和关注度,本文以微博用户关于“档案”的文本内容为实验样本。时间跨度从

2023 年 3 月至 8 月,共收集了 205396 条有关档案的微博文本。首先,在剔除无关文本(如表情文本、无意义文本等)和重复文本后,共得到 156871 条标准实验样本;然后,构建档案自定义分词表进行专有名词识别(如“蔚蓝档案”“时尚档案”等),精确分词,保证研究的真实性;最后,基于哈工大停用词表并结合档案领域的特点,过滤停用词(如“时候”“大卫”等)及标点符号。

3.2 微博文本内容的档案主题挖掘和关注度分析

3.2.1 主题提取和特征词选取

本文采用 Gibbs Sampling 来确定 LDA 模型的超参数,并利用 gensim 工具从原始的非结构化文本中实现微博文本潜在主题的挖掘和向量表达。根据式(1)计算不同主题数所对应的困惑度^[10],并借助 Python 工具将计算结果进行可视化,结果如图 1 所示,LDA 模型的困惑度随着主题数的增加逐渐减少,当主题数在 10 以上时,困惑度逐渐趋于平稳,此时主题拟合性和聚类效果明显。因此,本文 LDA 最优主题数选为 11。

$$perplexity(D) = \exp \left\{ - \frac{\sum_{d=1}^M \log(p(w)_d)}{\sum_{d=1}^M N_d} \right\} \quad (1)$$

式中: D 代表爬取微博文本的数量, $(w)_d$ 为词汇 d 的序列, N_d 代表 d 的词量。一般来说,困惑度越小,聚类效果和主题拟合性就越好。

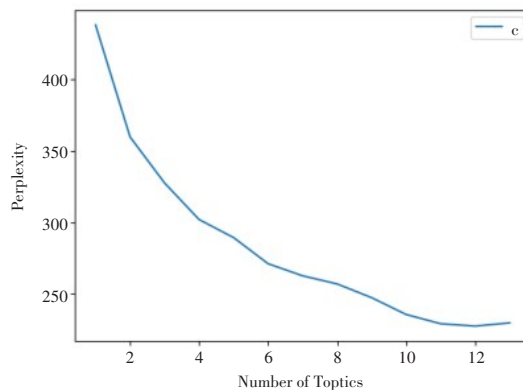


图 1 主题-困惑度变化趋势

利用 LDA 主题模型对微博文本进行分析得到相关数据,包括主题以及各主题对应的主题

词的概率向量,并结合TF-IDF获取各主题下概率最大的前11个描述词(见表1)。从表1可知,微博用户对档案的关注主题呈现出关注范围广、种类多样的特点,不仅涵盖了学术档案、政府档案等传统档案,还包括时尚档案、娱乐档案等新式档案。

由于档案的功能和用途受到服务人群、保存期限和类型等因素影响,我们将上述档案主题归纳为四大类:档案项目、影视娱乐档案、学生-学校档案以及专项档案。新《档案法》明确各类群体在保护、收集、利用档案的权利,鼓励社会公众参与到档案事业中。同时,档案

作为信息内容存储的载体,与国家治理活动、历史文化、经济科技发展和社会历史面貌等方面密不可分。

3.2.2 档案主题关注度计算

为获知微博用户对档案领域中各主题的关注度,将关注度的主题词与式(2)有效结合^[11],计算各档案主题的关注度,得到各档案主题随时间的变化趋势,如图2所示。

$$\overline{\theta}_{k,t} = \frac{1}{M_t} \sum_d \theta_{d,k,t} \quad (2)$$

式中: $\theta_{d,k,t}$ 为 t 时刻文本 d 在主题 k 的关注度; $\overline{\theta}_{k,t}$ 为 t 时间下主题 k 的关注度, M_t 为时间 t 下的

表1 主题-关键词表

类别	主题	关键词	数目
档案项目	Topic 1	工作 服务 社区 管理 检查 建设 企业 培训 健康 档案 监督 市场 信息化	14759
学生-学校档案	Topic 2	档案 毕业生 毕业 孩子 教育局 校长 文章 高校 材料 解析 公章 报到证	9005
	Topic 8	档案 学校 档案袋 老师 女儿 学生 毕业 大学 整理 母亲 考研 朋友 影响	39611
专项档案	Topic10	档案 公司 工作 档案管理 资料 专业 管理 信息 时间 领导 单位 系统 电子	14751
	Topic 5	档案 部队 侵华日军 档案资料 罪证 作者 犯罪 历史 乡村 社会 身份 出版社	5888
	Topic 6	档案 照片 民警 状元 女士 父母 人生 女子 女生 派出所 遗失 创作 笔记	19854
	Topic 7	档案 档案馆 中国 文化 国际 历史 主题 建设 世界 博物馆 数字 发展 时代	17842
影视娱乐档案	Topic 9	朱正廷 时尚档案 画师 事件 奇迹 认证 劳动争议 帅气 好帅 品牌 档案	12193
	Topic 0	链接 网页 视频 微博 档案 小说 档案室 公告 档案馆 刑事侦缉档案 全世界	15294
	Topic 3	蔚蓝档案 档案 记录 游戏 评论 角色 剧情 案件 图片 网友 抽奖 手机	12804
	Topic 4	肖战 品牌 艺术 典藏 代言人 档案 故事 寰宇 校园 序幕 时光 起源 实验室	4798

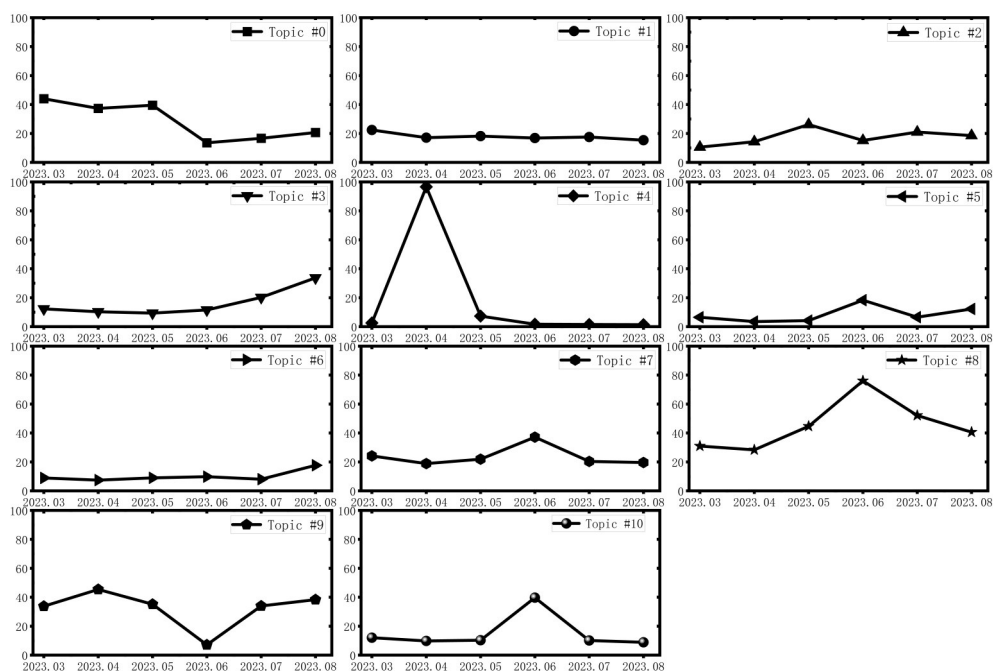


图2 档案主题关注度时间演化

主题集合总数。

由表1可知微博文本以学生档案、明星时尚档案为主，这是微博用户的主要关注点。由图2可知，在2023年3月至8月期间，相比于其他主题，Topic 0、Topic 4和Topic 8变化趋势较为明显。Topic 0呈现前期平稳后期下降的趋势，说明用户对档案监督工作和信息化建设有所关注，但关注程度随时间逐渐降低；Topic 4出现短期的高关注度，这很可能与“明星突发热点事件”有关，引起了粉丝的大量转发和传播；Topic 8、Topic 10前期平稳，到中后期先增加后下降，这是由于此阶段正处于学生升学或就业准备期间，导致该主题出现短期的高关注度；Topic 1、Topic 2、Topic 6关注度相对稳定，波动幅度不大；Topic 3呈现逐渐升高趋势，原因可能在于某游戏从早期宣传到后期公测，引起玩家的持续关注；Topic 9整体呈现平稳趋势，但在6月出现短期的低关注度，此时正处于学生升学或者就业的毕业季，微博用户更关注学生档案等方面，导致对明星档案关注度降低。综上，从事档案服务的人员和机构，应当把握热点档案领域，针对不同用户提供个性化服务，在践行新《档案法》的同时不断完善档案工作过程中存在的缺陷，提高档案在整个社会的关注度，让“档案事业”在整个社会中发挥更重要的作用。

3.3 档案主题情感倾向分析

为防止LSTM模型过拟合，同时保证实验结果的真实性和有效性，将微博文本按照6:2:2比例随机分为训练集、测试集和验证集。本模型将Dropout参数^[12]设置为0.2，epochs执行周期设为30，每句话处理长度设为50，词表长度设为10000，classes类别设为3(negative/active/neutral)，为防止某次异常实验的干扰，准确度选用30次情感分类的平均值(如图3所示)。最后模型在训练集上达到92.48%的精度，分类结果共得到积极倾向文本85152条，中性倾向文本27034条，消极倾向文本44685条。

为获悉微博用户对于各档案主题的真实情感态度，本文在充分考虑档案特点的基础上，基于LSTM算法对各分类档案的微博文本进行情

感倾向分析，结果如图4所示。

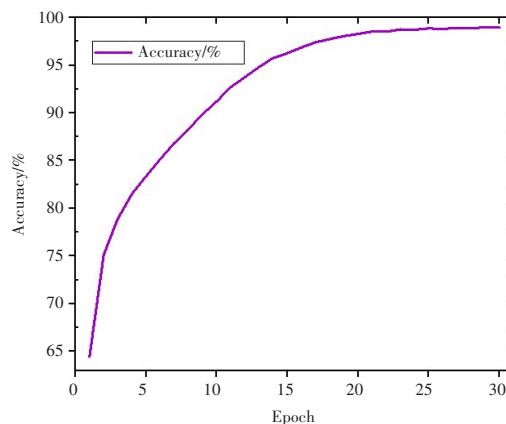


图3 准确度随执行周期演化

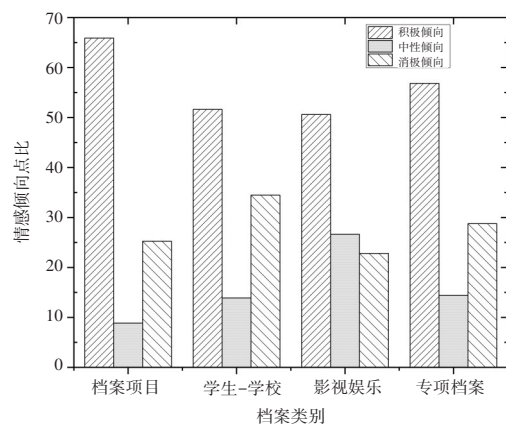


图4 分类档案主题的情感分析

图4中，各分类档案主题的积极倾向均在50%以上，表明用户对档案领域整体上认可度较高，其中档案项目类别的积极倾向占比最高，说明新《档案法》关于“档案的信息化建设”和“档案管理和监督”的内容得到公众较大认同和支持，同时国家也鼓励公众参与到档案事业的发展中，充分发挥档案在社会生产过程中的重要性。其次，在学生-学校档案中，将近35%微博用户呈现消极倾向，查阅微博原文了解到，消极情绪占比较高的原因除了3月至8月正处于毕业季，学生面临调档问题而产生疑惑、厌烦情绪以外，还可能与一些突发新闻事件有关，例如家长私拆学生档案，致使档案失效，引起部分用户对某些家长的负面评价；部分学生犯了重大错误，部分用户义愤填膺，希望将此记录到学生档案中，因而产生了大量的消极

倾向。影视娱乐档案类别的中性倾向最高,说明电影、游戏和明星等相关档案在受追捧的同时,可能由于部分用户对此类档案不置可否,在网上表达出比较中性的内容。在专项档案类别中,有45%的用户具有非积极倾向,可能由于部分微博用户对政府、企业等档案工作流程存在不了解和不满意的情绪,对此,新《档案法》实行统一领导、分级管理的原则,从根本上保障公民利用档案的权利。在落实档案管理和信息化工作机制的同时,加强档案知识宣传教育,创新档案服务形式和功能,进一步增强全社会档案意识。

为了更直观地看出用户情感倾向程度的变化趋势,对2023年3月至8月的微博文本倾向程度进行赋值,分别对每日的积极倾向和消极倾向进行加权计算,得到情感倾向程度随时间变化的趋势图,如图5所示。

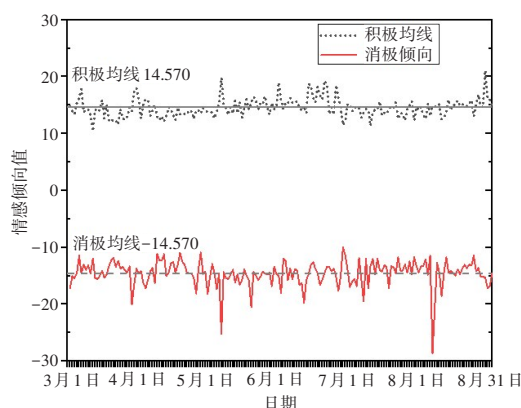


图5 档案主题情感倾向程度时间演化

如图5所示,在研究时间段内,微博用户对档案话题的积极倾向波动幅度不大,整体保持着稳定,而消极倾向在5月上旬以及8月上旬变化较大,形成了明显的峰值,查阅微博原文可知,5月上旬正处于学生升学、就业的阶段,学生面临调档、查档的问题,产生了较多的不解以及烦躁情绪;8月上旬,在黑龙江省牡丹江市举行的侵华日军细菌战与毒气战学术研讨会上,侵华日军第731部队罪证陈列馆向社会首次公布《日军哈尔滨第一陆军病院原簿》。通过这份档案公众全面认识到日军陆军病院参与活体解剖犯罪,对此产生了大量激愤情绪,并在微博上表达对日军此类行为的不满情绪。以上两点表明,公众对档案话题的情感倾向容易受到某个

特定时期或一些热点新闻事件的影响。

4 结语

本文将LDA-LSTM模型应用到档案学领域,从用户视角去探究公众对档案话题关注程度,该模型不仅能准确识别微博文本的档案主题,而且能够了解到用户对各档案的正负情感倾向,发现档案领域的关注热点和工作中存在的不足,从而提出合理的应对措施,进一步提高公众对档案关注度和满意度,增强全社会档案意识。相关部门应完善档案利用规则,创新档案服务形式和功能,为档案的高效利用提供条件,同时加强相关档案知识和档案工作机制的宣传教育,提高档案服务质量和效率,从根本上保障公民利用档案的权利,增强全社会对档案的重视度和认同感。新《档案法》中强调的档案信息化、多方参与、便民利民等规定得到了一定的落实,同时也得到了公众的关注与支持,这说明新《档案法》紧跟时代的步伐,既符合人民群众的利益,也满足了人民群众的基本要求。通过本文对新浪微博档案话题关注度与情感倾向的研究,为档案从业者与档案相关部门机构如何提升服务质量、提高档案工作的关注度、增强社会的档案意识和认可度提供了一定的参考。

参考文献:

- [1] 徐拥军. 档案在推进文化自信自强中的价值及实现路径[J]. 档案学通讯, 2023(1): 20-27.
- [2] 赵建平. 新媒体环境下档案馆公共服务能力提升研究[J]. 兰台世界, 2022(2): 40-44.
- [3] 易黎, 林扬. 基于深度学习的档案知识图谱构建研究[J]. 中国档案, 2022(11): 33-35.
- [4] LIAN Z Y. Factors influencing the integration of digital archival resources: a constructivist grounded theory approach[J]. Archives and Manuscripts, 2016, 44(2): 88-102.
- [5] 周娟娟, 李泽锋. 档案数据长期保存模型构建研究[J]. 北京档案, 2022(10): 10-14.
- [6] 王杰. 新媒体与档案信息传播的融合应用研究[J]. 新闻爱好者, 2022(7): 101-103.
- [7] 李宗富. 共建共享理念下公众参与档案馆志愿服务活动现状调查研究[J]. 档案学研究, 2019(2): 88-98.
- [8] 连志英. 公众参与数字档案资源社会化开发项目动

- 机研究[J]. 档案学研究, 2022(4):83-90.
- [9] 崔月, 高淑珍. 新《档案法》背景下公众档案素养培育现状及路径探究[J]. 兰台世界, 2023(8):67-69.
- [10] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003(3):993-1022.
- [11] 曾莉, 杨添宝, 周慧. 基于 LDA 与注意力机制 BiLSTM 的微博舆情分析模型[J]. 南京理工大学学报, 2022, 46(6):742-748.
- [12] 张涛, 马海群. 一种基于 LDA 主题模型的政策文本聚类方法研究[J]. 数据分析与知识发现, 2018, 2(9):59-65.

Analysis of emotion and attention of Weibo archives based on LDA-LSTM algorithm

Sun Siyi¹, Wang Jiaqiang², Luo Zijiang^{3*}

(1. School of Economics and Management, Guiyang College of Humanities and Technology, Guiyang 550025, China;

2. School of Information, Guizhou University of Finance and Economics, Guiyang 550025, China;

3. Institute of Intelligent manufacturing, Shunde Polytechnic, Shunde 528300, China)

Abstract: In order to analyze the public's attention and emotional attitude towards the archives two years after the implementation of the Archives Law, the authors used Python tools and LDA model to extract and cluster the data, hot topics of archives at different times were obtained in this paper; the LSTM model was utilized to obtain the emotional tendency of each file theme and analyzed the reasons for the different emotional tendencies of users. According to the connection between themes, there were four categories of archives projects, such as film and television entertainment files, student-school archives and special archives; all kinds of archives categories had high positive tendency, which indicating the public support and understand the development of archives. For the students with high negative tendency-school archives category, the relevant departments should strengthen the publicity and education of archives, protect the public's right to use archives fundamentally, and enhance the archives awareness of the whole society.

Keywords: attention of archives; LDA-LSTM algorithm; emotion analysis; Sina Weibo

(上接第 48 页)

Data sorting method of lightweight wireless network based on deep learning of attention mechanism

He Lianghua*, Zhao Ying

(School of Electronics and Engineering, Nanchang Normal College of Applied Technology, Nanchang 330000, China)

Abstract: Lightweight wireless network is usually used for resource-limited equipment, which makes the data transmission and processing capacity is limited, and the amount of data is huge, and the feature identification is difficult, resulting in low data sorting accuracy. Therefore, a lightweight wireless network data sorting method based on deep learning of attention mechanism is proposed. Calculate the characteristic value of the data, quantitatively preprocess the basic data, use the rough set algorithm of the field to calculate the redundant limit of data sorting, the data features are extracted in multiple stages, construct the data sorting process of attention mechanism deep learning network, and adopt the optimal screening method to realize the data sorting processing. The final test results show that for selected six data sorting test cycle, attention mechanism deep learning network data sorting method final data sorting average F-Score can reach more than 85%, in the attention mechanism of deep learning technology, the current design data sorting method more pertinence, higher efficiency, has the practical application value.

Keywords: attention mechanism; deep learning; lightweight wireless network; data sorting; sorting method

文章编号: 1007-1423(2024)07-0076-05

DOI: 10.3969/j.issn.1007-1423.2024.07.013

软件设计模式下装饰模型的研究与应用

向晓婷*

(贵阳信息科技学院信息工程系, 贵阳 550025)

摘要: 软件设计模式有很多种类, 当需要通过一些基本功能的排列组合形成大量的功能时, 简单地使用继承关系难以实现, 针对该问题对软件设计模型进行研究。首先, 介绍装饰模式和 UML 统一建模语言; 然后, 针对装饰模型设计对句的案例; 最后, 使用面向对象的 Java 语言进行实践, 丰富装饰模型的应用案例。

关键词: 设计模式; 装饰模型; 面向对象; Java; UML

0 引言

设计模式有很多, 专业术语晦涩难懂, 如何以简单易懂的方式阐述清楚设计模式中的装饰模型和应用, 是个值得研究的问题。金强山等^[1]针对“软件设计与体系结构”课程中设计模式的抽象性与复杂性, 为了使学能够更好地理解“观察者模式”的设计思路, 设计了温室蔬菜大棚中温度、湿度、CO₂、光照的数据采集处理问题实例。樊建平等^[2]在 QT 开发环境中, 基于组合设计模式与观察者设计模式, 提出一种开发模拟器终端软件的编程方法。张宇^[3]在一项提高软件质量的研究中发现, 将软件设计模式组合应用到某类软件是提高其质量的可行方法之一。这说明模式应用到软件是提升软件质量的一种手段, 模式的应用性研究也是极其必要的。王巍等^[4]介绍了 Decorator 模式在咖啡店计费场景下的应用。肖卓宇等^[5]针对设计模式识别精确率不高的问题, 提出一种基于特征分类表示的设计模式识别方法, 以结构型、行为型及创建型三类归纳特征信息进行设计模式识别实验。

综上所述, 目前的软件设计模式的研究主要关注于软件需求的多样性, 以及类和对象之间紧密耦合。近年来, 新工科、一流课程和课

程思政是热点^[6]。然而, 刚迈进大学的计算机专业新生, 往往对思政教育和计算机专业中核心内容之一的软件设计认识不足。针对此问题, 在众多的软件设计模式中, 如何以一种简单、通俗易懂的例子为刚进入大学的计算机专业本科生阐述清楚软件设计模式下装饰模型的应用, 是个值得探讨的问题。

1 设计模式

在软件开发过程中, 软件设计模式用于解决普遍遇到的软件设计问题。针对不同的业务内容, 引导软件技术人员在缺乏编程经验的同时也能够根据软件设计模式做出更好的项目决策。软件设计模式的分类有不同的方法。根据模式应用目的的不同可将设计模式分为创建型模式、结构型模式、行为型模式三种^[7-8]。本文围绕结构型装饰器模式进行研究和实践。

装饰器模式, 也称为装饰模式, 其定义是在不改变接口的情况下, 动态地给对象添加新的功能。为了更好地理解装饰器模型, 举个例子来说明。如: 所有人都要吃饭, 但是张三吃完饭之后想要锻炼身体, 准备跑五公里, 那么就需要针对张三这类群体增加特殊的装饰(使用装饰器 A), 从而让张三吃完饭之后还可以跑步这一动作, 如果吃完饭还有其他活动的安排,

收稿日期: 2023-12-25 修稿日期: 2024-01-16

作者简介: *通信作者: 向晓婷(1994—), 女, 贵州贵阳人, 硕士, 助教, 主要研究方向为面向对象编程技术、Linux 虚拟化技术, E-mail: 18275382376@163.com

还可以再添加多个装饰器(如装饰器B、装饰器C等)。装饰器相比于继承来说更灵活,在不改变原有对象的前提下,可以动态地为一个对象扩展功能。装饰器模式不仅完全遵守软件开发的开闭原则,还可以通过使用(或不用)装饰类及这些装饰类的排列和组合,从而实现不同的效果。

2 UML类图

统一建模语言(unified modeling language, UML)是一种为面向对象系统的产品进行说明、可视化和编制文档的一种标准语言,是非专利的第三代建模和规约语言^[9]。针对项目进行系统设计时,采用UML建模和编写文档,往往可以达到事半功倍的效果。UML具有很多不同类型的图表,如静态图和动态图。静态图有用例图、类图、包图,动态图有活动图、状态图、顺序图、协作图。

本文主要用到类图,因此只介绍类图。类图是UML类图中的主要元素,类图的组成包括三行,第一行是类的名称,如果是抽象类,一般采用斜体;第二行是类的属性,也就是类的各种成员变量;第三行是类的方法,即类的各种成员操作。对于属性和方法前面的符号有如下说明,公有类型public用“+”表示,私有类型private用“-”表示,受保护类型protect用“#”表示。

3 装饰模型应用设计

历史上留下了很多经典的对句,比如清代李文甫,年少时跟随老师出行,老师指着远处的雪峰吟出上句“青山原不老,为雪白头。”要求学生对出下句。李文甫低着头沉思,看见身边的一口池塘,树枝倒映在水中,微风吹过,池面泛起涟漪,于是悟出其对句“绿水本无忧,因风皱面。”

这个场景蕴含软件设计的一个重要思想:不改变原始类的代码,动态地为一个对象扩展其功能,这种思想被总结为设计模式中的装饰模式。这个场景中,给出上句的人不需要知道下句是什么内容。通过上句,给出下句的人属于动态扩展角色,需要知道上句具体的内容^[10]。

换句话说,任何动态扩展角色都需要知道上句角色,才能完成扩展任务。从面向对象的角度来说,要完成扩展任务,任何动态扩展角色都必须知道上句角色才行。在该场景需要注意的是,经过动态扩展角色最终形成的是一个完成的“对句”,这里的下句刚好是动态角色的特色工作之一。为实现对句功能,UML类图如图1所示,随后采用面向对象Java语言实现。

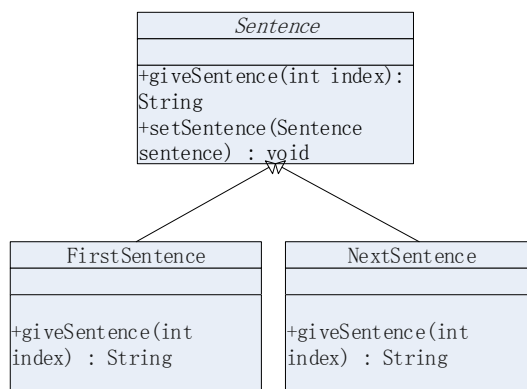


图1 对句UML类图

4 装饰模型应用实现

针对上述场景需要设置三个类,类名分别是 Sentence 类、FirstSentence 类和 NextSentence 类,其中 Sentence 类需要设置成抽象类而非接口,其原因是可以共享代码的实现。FirstSentence 类和 NextSentence 类不仅仅共享方法的签名,同时还需要共享相同的代码来实现。为了避免代码的冗余,抽象类是更好的选择。类的扩展性,即在不破坏已有实现的类的前提下,为类提供新的方法或者是属性,在更新迭代的代码版本中抽象类比接口更佳,其原因是通过在抽象类中添加新的方法或属性就可以保证现有实现的类不受到重写的影响。访问非公共成员,相比于接口只能定义没有访问权限公共的方法,抽象类可以定义受保护的成员变量和方法,并能够保证这些成员可以在子类中之直接访问,不受访问权限的限制。需要提供默认实现,接口只能包含方法的签名没有默认的实现,而抽象类可以包含具体方法的实现。针对该场景,并不要求 FirstSentence 类和 NextSentence 类都实现相同的代码,因此将 Sentence 类设置成抽象

类更适合。

```
public abstract class Sentence {  
    public abstract String giveSentence(int index);  
    public void setSentence(Sentence sentence){  
        }  
}
```

抽象类 Sentence 是父类，该类中有两个方法，方法名分别是 giveSentence 和 setSentence，用于实现对句的方法，这两个方法的访问权限都是 public，体现在 UML 类图中是用“+”来表示。FirstSentence 类和 NextSentence 类是子类，从 Sentence 类中继承过来，FirstSentence 类用于给出对句的上句，NextSentence 类是针对上句给出对句的下句。在抽象类 Sentence 中，giveSentence() 是抽象方法。抽象方法，是在方法声明时在方法的返回类型前加上 abstract 关键字，抽象方法没有方法体内容，具体内容对应的子类来重写实现。之所以将 giveSentence() 设置成抽象方法，是因为部分对象具有多态性，不同的对象对相同的行为可以用不同的处理方式。比如小猫和小狗都是动物，都可以吃东西，但是吃的东西各不相同，小猫吃鱼，而小狗吃骨头，针对吃这个动作如果都定义在父类中就会有冲突，因此在父类中将其定义成抽象方法，而具体吃的这个行为由具体的子类来实现。

输入：输入数组 sentence[] 内容，用于存放对句的上句内容

输出：获取对句的上句内容

```
public class FirstSentence extends Sentence {  
    String [] sentence;//存放上句  
    int index = 0;  
    FirstSentence() {  
        sentence = new String[3];  
        sentence[0] = “青山原不老,为雪白头。”;  
        sentence[1] = “水车车水,水随车,车停水止。”;  
        sentence[2] = “天作棋盘星作子,日月争光。”;  
    }  
    public String giveSentence(int index) {  
        return sentence[index];  
    }  
}
```

FirstSentence 类是实现抽象类 Sentence 的子类，在 Java 中用 extends 表示继承，FirstSentence

类用于编写对句的上句内容。在该类中，不再采用 Java 默认的构造方法，而是自定义构造方法，并在构造方法中先给出对句的上句内容，并通过 giveSentence() 方法把对句的上句内容进行输出，从而让其他对象根据对句的上句内容给出下句。

输入：sentence 对象作为参数，输入索引的内容

输出：对句的下句内容

```
public class NextSentence extends Sentence {  
    Sentence sentence;  
    public void setSentence(Sentence sentence) {  
        this.sentence = sentence;  
    }  
    public String giveSentence(int index) {  
        String str = sentence.giveSentence(index);  
        if(index == 0) {  
            str = str+ “\n绿水本无忧，因风皱面。”;  
        }  
        else if(index == 1) {  
            str = str+ “\n风扇扇风，风出扇，扇动风生。”;  
        }  
        else if(index == 2) {  
            str = str+ “\n雷为战鼓电为旗，风云际合。”;  
        }  
        return str;  
    }  
}
```

NextSentence 类是实现 Sentence 抽象类的子类，在该类中用抽象类先生成一个对象 sentence，在方法 setSentence() 传递刚创建的对象作为方法的参数，并在方法体中针对 FirstSentence 类给出的上句内容给出下句内容。

MainClass 是该应用场景的主类，程序从该类开始逐行运行。用抽象类生成两个对象，对象名分别是 firstSentence 和 nextSentence，并用默认的构造方法为其分配内存空间，这一过程是实例化的过程。firstSentence 对象去调用 FirstSentence 类中的方法读取对句的具体内容后，调用 nextSentence 类中的 giveSentence 方法，并将其对句的结果存放在成员变量 result 后进行输出。程序的运行先进入 MainClass 类，创建 firstSentence 对象，调用 FirstSentence 类中自定义的 FirstSentence() 构造方法，为对象分配内存空

间，读取成员变量 index 的值。读取 FirstSentence 类一维数组名为 sentence、数组长度为3的数组内容，从而获取对句的上句内容。创建 nextSentence 对象，调用 NextSentence 类中的 setSentence (firstSentence) 方法和 giveSentence (int index) 方法后，调用 FirstSentence 类中的 giveSentence(int index) 方法，并将 sentence[index] 值返回。判断 NextSentence 类中 giveSentence() 方法体内 index 的值是否超出长度为3的数组，如果超过数组长度则程序结束，否则判断 index 的值是否为 index == 0。如果值为零，让 str 的值连接上“绿水本无忧，因风皱面。”将保存在 result 变量中的“青山原不老，为雪白头。绿水本无

忧，因风皱面。”进行输出显示。如果值不为零，则判断 index 的值是否为 index == 1，如果判断的结果为真，让 str 的值连接上“风扇扇风，风出扇，扇动风生。”将保存在 result 变量中的“水车车水，水随车，车停水止。风扇扇风，风出扇，扇动风生。”进行输出显示。如果判断的结果为假，则继续判断 index 的值是否为 index == 2，如果判断的结果为真，让 str 的值连接上“雷为战鼓电为旗，风云际合。”并将保存在 result 变量中的“天作棋盘星作子，日月争光。雷为战鼓电为旗，风云际合。”进行输出显示，否则跳出 if-else 循环，结束运行。程序流程图如图2所示。

运行结果如图3所示。

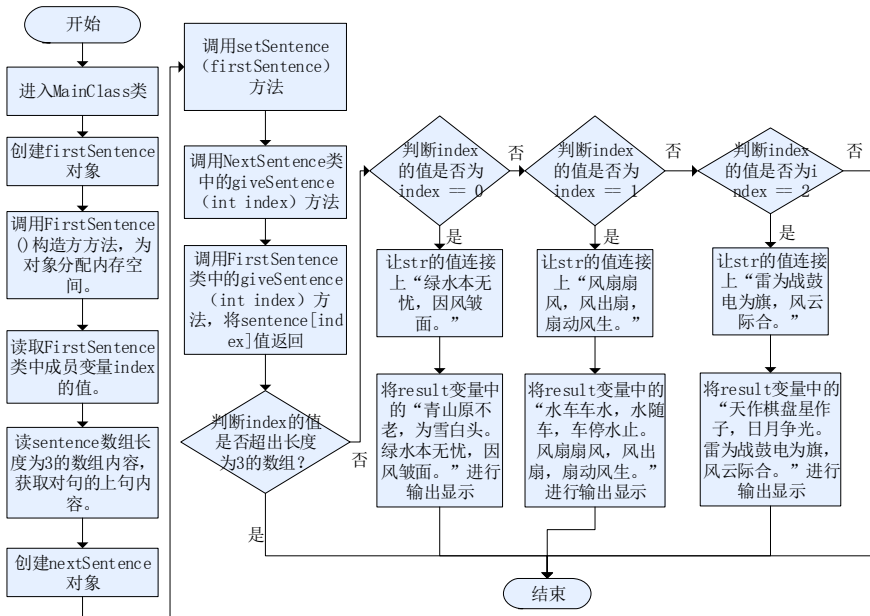


图 2 程序执行流程图

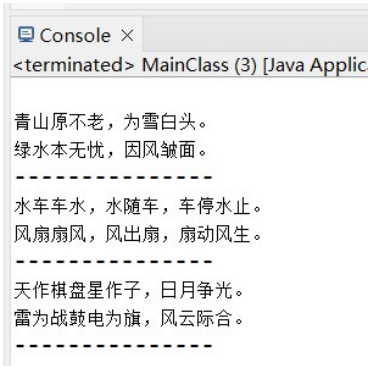


图 3 运行结果

5 结语

在软件设计过程中，动态增强子类对象的功能，使用对象之间关联的关系来替代类之间的继承关系，这种用于替代继承的技术就是装饰模式。在该模式中，通过引入新的装饰类，不仅可以调用原来类中的方法，还可以添加新的方法从而扩展原有类的功能。使用装饰模式不仅可以降低系统的高耦合度，还可以动态添加或删除原有对象的职责，从而让具体装饰类可以独立变化，让系统的可扩展性和可维护性

增强,符合软件设计开闭原则。如果想要创造出多种不同的新行为,通过对具体装饰类进行排列组合就可以完成。当希望动态透明地给对象添加功能而不是通过继承方式扩展类功能时,装饰模式比其他软件模式更适用。

参考文献:

- [1] 金强山,冯光,孙传国.“观察者模式”软件工程综合性实验设计[J].计算机时代,2022(5):105-108.
- [2] 樊建平,宋众艳,朱胤,等.设计模式在开发模拟器终端软件中的应用[J].火力与指挥控制,2023,48(7):99-103.
- [3] 张宇.设计模式在外观管理系统中的应用研究[D].太原:中北大学,2021.
- [4] 王巍,辛国栋.装饰模式应用研究[J].计算机工程与设计,2008(8):1884-1885,1889.
- [5] 肖卓宇,何锴,徐运标,等.设计模式识别的特征信息分类研究[J].计算机应用与软件,2022,39(6):7-11,20.
- [6] 陈晓丹,李思明,李立,等.基于OBE的智慧教室环境下新工科课程思政建设:以计算机科学导论课程为例[J].电脑知识与技术,2023,19(30):132-134.
- [7] 钟茂生,王明文.软件设计模式及其使用[J].计算机应用,2002(8):32-35.
- [8] 王竹荣.软件设计模式基础及应用[M].北京:电子工业出版社,2021.
- [9] 刘亚倩,赵大平,方洁,等.基于统一建模语言UML流程图的测试方法,设备及介质:CN202211498923.5[P].2022-11-28.
- [10] 耿祥义,张跃平.Java 2实用教程[M].6版.北京:清华大学出版社,2021.

Research and application of decorative models in software design patterns

Xiang Xiaoting*

(Department of Information Engineering, Guiyang Institute of Information Science and Technology, Guiyang 550025, China)

Abstract: There are many types of software design patterns. When it is necessary to form a large number of functions through the arrangement and combination of some basic functions, it is not practical to simply use inheritance relationships to achieve this. To address this issue, research software design models. Firstly, introduce the decoration pattern and UML unified modeling language. Then, design a case study for the decorative model with corresponding sentences. Finally, practice using object-oriented Java language to enrich the application cases of decorative models.

Keywords: design patterns; decorative models; object oriented; Java; UML

文章编号: 1007-1423(2024)07-0081-05

DOI: 10.3969/j.issn.1007-1423.2024.07.014

计算模型在化合物毒性预测方面的应用及展望

许志旺, 何王秋, 孔 韧*

(江苏理工学院电气信息工程学院生物信息与医药工程研究所, 常州 213001)

摘要: 毒性研究是一项重要且复杂的科学工作, 涉及到评估化学物质及其他物质对人类健康的潜在危害, 也对环境保护和可持续发展起着重要作用。传统的基于分子、细胞与动物模型的毒理学实验, 除了成本高昂且耗时的问题, 还存在对动物不友好的局限性。随着计算机技术的迅猛发展, 出现了计算毒理学模型的开发研究。相比于传统的毒理学实验, 计算毒理学模型具有高效率、低成本、方便快捷的特点, 逐渐成为毒理学研究的热点方向之一。介绍了包括机器学习在内的计算模型预测化合物毒性的研究方法及其在环境保护、毒性风险评估和药物设计等领域的应用, 并进一步讨论了计算模型在毒性预测中存在的不足之处及可能的发展方向。

关键词: 计算毒理学; 机器学习; 毒性预测

0 引言

随着化工行业的飞速发展, 化合物的数量呈现出快速增长的趋势。这些化学品在整个生命周期中, 在生产、储存、运输、应用和废弃物处理过程中难以避免其释放到人们生活的环境中, 造成生态和健康危害^[1]。这既为人们的生活提供了便利, 同时也增加了对环境和人类健康的潜在风险, 因此需要加强毒性测试和管理措施。化合物的毒性评估在现代科学研究和工业生产中占据核心地位, 直接关系到人类、环境的健康和新化合物及药物的研发。传统的评估方法, 如动物实验、细胞实验, 由于时间、资源消耗大和伦理问题, 其准确性和全面性受到限制^[2]。定量构效关系 (quantitative structure-activity relationship, QSAR) 是化学和药物学中一项非常重要的研究方法, 根据化合物结构定量/定性地预测化合物的活性^[3]。

早在 20 世纪 60 年代, 科学家开始引入计算模型和神经网络等先进技术, 提供更多确定且可靠的定量构效模型。2007 年, 美国环境保护

署(EPA)启动了ToxCast或“毒性预测器”计划, 该计划旨在使用计算机技术, 通过对化合物的毒性进行评估, 以预测它们对人体健康和环境的潜在毒性^[4]。化合物毒性评估在多个领域中具有广泛应用。这包括药物开发, 其中新的候选药物必须经过毒性评估, 以确保它们是安全的^[5]。随着计算机技术和数据科学的进步, 计算机预测已成为毒性测试的新方向。这种方法能快速、准确地评估化合物毒性, 减轻实验的负担和解决伦理问题。计算机技术在生物和化学领域应用广泛, 特别是在基因序列分析、分子建模等方面, 不仅提升了毒性测试的准确性, 也加快了新药或其他功能化学品的研发速度^[6]。

1 研究方法

1.1 建模流程

计算毒理学经典研究方法之一为 QSAR 模型, 是一种建立化学物质结构与毒性之间定量关系的方法。其建模一般分为数据获取、数据处理、计算分子描述符、建立模型、模型评估, 如图 1 所示。

收稿日期: 2023-10-30 修稿日期: 2023-11-23

作者简介: 许志旺(1994—), 男, 江苏泰州人, 硕士研究生, 研究方向为深度学习、环境毒理学; 何王秋(1999—), 男, 江苏盐城人, 硕士研究生, 研究方向为分子动力学模拟; *通信作者: 孔韧(1981—), 女, 湖南岳阳人, 博士, 研究员, 硕士生导师, 从事领域为深度学习、高性能计算和分子识别、人工智能辅助药物设计等, E-mail: rkong@jsut.edu.cn

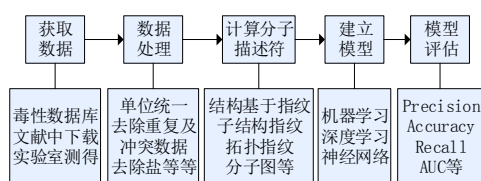


图1 QSAR建模流程

1.2 数据收集

数据的可靠性是建立模型的关键，模型的预测能力和准度很大程度上取决于数据是否可靠。如果数据有错误或者数据不完整，那么建立出来的模型很可能出现过拟合、欠拟合等问题，从而导致预测结果不准甚至无法使用^[7]。因此，公开毒性数据库的建立解决了数据获取的来源问题，为研究人员提供了便利。本文列举了一些比较有特点的化合物毒性数据库，见表1。

除了公开的毒性数据库外，还有各大学术数据库，例如知网、Google scholar、Pubmed等学术数据库，这些数据库中的文献包含了先前研究的数据和结果，提供了丰富的化合物结构、性质和毒性数据。

1.3 数据处理

在QSAR建模过程中，数据预处理是非常重要的一步，不仅能够提高模型的精度和训练效率，而且使得模型更加可靠^[8]。需要删除一些

不完整或者容易混淆的化学分子结构以及不能被大部分软件计算分子结构描述符的数据，如有机金属化合物、混合物、无机物以及离子等。由于下载到的数据会有重复，并且重复的数据其活性也不尽相同，这是由于数据来源于不同的实验室，测得的数据的实验条件并不一样，因此需要对数据进行分析，一般是去除极端数据后，将平均活性作为最后的活性值。但如果多条数据是同一种化合物并且其活性差异较大时，便去除这种化合物的数据。最后，还需要对SMILES标准化。

1.4 计算分子指纹

分子指纹是分子描述符的一种，它是用固定长度的二进制向量表示化学分子的结构特征，从分子的结构图中提取一定范围内的特征，例如化学键的种类、长度、偏轨和环的个数等，然后根据这些特征进行编码，得到一个长度固定、数值离散的指纹向量。分子指纹通常被用于快速比较和查找化学分子之间的相似性，是化学信息检索和高通量筛选中常用的特征提取方法^[9]。2018年，Moriwaki等^[10]提出了一种分子描述符计算工具Mordred，它的速度是Padel-Descriptor软件的两倍，并且可以计算大分子的描述符。此外，还列举了其他计算分子指纹的工具，见表2。

表1 毒性数据库

数据库	简介	网站	国家	年份
Cal/EPA	加州环保局毒性数据库，可以获得所需物质的风险评估信息	https://oehha.ca.gov	美国	1991
ACToR	超过50万种环境化学品的数据	https://actor.epa.gov/actor/home.xhtml	美国	2010
PubChem	115 M化合物，306 M物质，304 M生物活性	https://pubchem.ncbi.nlm.nih.gov	美国	2004
Drugs@FDA	自1939年以来批准的大多数药品	http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm	美国	1939
ECOTOX	1万多种化合物毒性数据，1万多物种毒性数据	http://cfpub.epa.gov/ecotox	美国	1995
DrugBank	主要的生物信息学和生物医学数据库。超过50万种药物和药品	https://go.drugbank.com	加拿大	2006
ChEMBL	2.4 M种化合物，14 k种药物	https://www.ebi.ac.uk/chembl/	欧洲	2011
ChemIDplus	有超过490万种化合物数据以及大量的毒性数据	https://www.nlm.nih.gov/pubs/techbull/ja22/ja22_pubchem.html	美国	2002
ToxCast	包含了对8000多个化学物质的超过1000种生物活性端点的数据	https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data	美国	2007
Lincs	涵盖了超过10万种化合物，包括药物、小分子化合物、天然产物等	https://lincs.hms.harvard.edu/	美国	2007

表 2 分子指纹计算工具

工具	说明
RDKit	是Python化学工具包,支持分子的表示、转换和计算分子描述符等多种功能
Open Babel	一个开源的化学工具包,支持许多流行的分子描述符,如化学键信息、元素等
Padel-Descriptor	基于Java语言,提供了计算超过1000个分子描述符的能力
Mordred	它是一个开源工具,可以通过编程接口或命令行界面来访问和使用。提供了广泛的分子描述符,包括物理化学性质、构象特征、电荷分布、极性、溶解度,等等
ChemDes	一个基于Python的化学工具箱,能够计算多种分子描述符,并提供了一些机器学习算法的实现
Dragon	一个商业化学软件,提供了多种计算分子描述符的算法和工具,并支持进行QSAR建模和分析等任务
MOE	一个商业化学软件,拥有丰富的分子描述符计算功能,并可以进行药物设计、机器学习建模等应用
ChemAxon	一个商业化学软件包,它具有广泛的分子描述符计算功能,并且可以扩展到大规模化学信息管理和分析

1.5 模型建立

不同的机器学习算法在不同的数据集上会表现出不同的预测性能,通常会出现欠拟合或者过拟合的问题发生。因此,在进行QSAR建模前要为特定的数据集选择最佳的算法。常用作构建QSAR模型的机器学习算法有:k最近邻算法、支持向量机、决策树、随机森林、线性回归、逻辑回归、人工神经网络、朴素贝叶斯等^[11]。通常需要对算法模型的参数进行调整,提高预测准确性和泛化能力。

1.6 模型评估

模型评估用于衡量模型的性能和可靠性,一般使用验证集来评估模型的性能。在化合物毒性预测中,选择适当的评估指标取决于任务的性质(二元分类、多类分类或回归)、数据的分布以及领域特定的需求。通常,多个指标结合使用可以提供更全面的模型评估。常用的评估指标包括准确性(Accuracy, ACC)、召回率(Recall, RE)、精确度(Precision, PR)、F1分

数等,计算公式如下:

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \quad (1)$$

$$RE = \frac{TP}{TP + FN} \quad (2)$$

$$PR = \frac{TP}{TP + FP} \quad (3)$$

$$F1 = \frac{2*PR*RE}{PR + RE} \quad (4)$$

2 毒性预测模型的应用

传统的毒性测试通常涉及大量的实验动物和耗费大量的时间和资源。毒性预测模型可以先于实验阶段,通过计算机建立模型进行预测,从而节省大量时间和金钱。计算机预测方法在化学、环境科学和药物开发等领域的应用已经非常广泛^[12]。特别是包括深度学习在内的机器学习方法具有在非常复杂的数据集中识别其特征的潜力,化合物毒性预测的准确性得到了极大的提高。在过去的十年里,机器学习方法在毒理学领域的研究有了急剧的增长。

2.1 环境化学方向

2019年,Liu等^[13]预测化学物质对甲壳海洋生物的毒性影响,从而判断出是否对海洋生态造成破坏。他们使用RF、SVM、kNN、ANN、朴素贝叶斯(NB)以及决策树(CT)这六种机器学习算法来预测化合物的毒性,结果较为理想的模型的受试者特征曲线下的面积(AUC)都能在0.8左右。2020年,Wang等^[14]将无向图(UG)和注意卷积神经网络(ACNN)相结合,建立了一种新的深度图注意卷积神经网络(GACNN)深度学习模型用于蜜蜂化学中毒预测,该模型取得了较为准确且稳定的结果。2021年,Kumar等^[15]利用深度神经网络(DNN)、支持向量机(SVM)、k近邻(k-nearest neighbor)和随机森林(random forest)建立了诱变预测模型。所建立的分类器是基于3039种化合物,对1014种化合物进行了验证。结果表明基于DNN的预测模型对训练数据和测试数据的预测准确率最高,分别为0.93和0.84。2023年,Sun等^[16]采用偏最小二乘(PLS)、多元线性回归(MLR)和遗传函数近似(GFA)三种回归方法预测了多环芳烃的沸点、辛醇-水分配系数和保留时间指数,通过相关系数(R²)和均方根误

差(RMSE)评估表明模型可用于环境化学物质的理化性质估计和毒性预测。2022年, Xu等^[17]利用随机森林(RF)、支持向量机(SVM)、k最近邻(kNN)、人工神经网络(ANN)、深度学习模型GCN五种机器学习方法结合九种分子指纹以及分子图预测化合物对四种鱼类的毒性, 每种鱼类的模型准确度都能达到0.83以上。此外, 该团队还建立了四种鱼类的整体模型来对鱼类的化合物毒性进行预测, 模型对于其他鱼类的预测效果能达到0.77。

2.2 药物化学方向

2017年, Zhang等^[18]基于朴素贝叶斯分类器开发了一种新的突变预测模型, 对已上市的446种药物的预测准确率达到0.91。2018年, 张兆津^[19]提出了一个基于深度神经网络的回归预测模型用于预测药物的毒性, 他将深度神经网络和单层神经网络进行对比, 在多组数据集实验中, 结果显示深度神经网络的学习性能和预测结果优于单层神经网络。2018年, Lysenko等^[20]提出了一种原创的机器学习方法TargeTox, 利用药物靶点和非靶点的识别, 从基因本体注释计算的功能影响评分和生物网络数据来预测药物毒, TargeTox可以区分具有潜在特异质毒性的药物和安全药物, 也适用于不同靶点集的推测性评估, 以支持最佳低毒性组合的设计。2023年, Ylipaa等^[21]利用20多万化合物进行训练, 使用RF、SVM、极限梯度提升(XGBoost)、深度神经网络(DNN), 以及基于门控递归单元的DNN(GRU-DNN)和图神经网络(GNN)六种机器学习模型对hERG(human ether-a-go-go-related gene)的毒性进行预测, 经过8万多的测试集验证, 模型的AUC都能达到0.94以上。

3 挑战与展望

虽然计算机预测方法在化合物毒性评估方面展现出巨大的潜力和价值, 但在实际应用中仍面临一系列挑战和问题。这些问题涵盖数据、技术等多个方面。

首先, 数据质量是影响计算机预测方法准确性的关键因素, 不准确、不完整或不一致的数据会导致模型的预测误差增大。数据的一致性问题, 如不同研究和实验的数据标准和格式

不一, 也会影响数据的整合和分析。研究人员可以标明实验条件, 并在分析中考虑这些因素的影响。这有助于降低数据的偏差, 并提高科学研究的可信度。当然, 由于商业、法规或伦理原因, 一些重要的数据无法获取的问题也困扰着很多科研人员。数据共享问题也影响着计算机预测方法的发展, 而开放和共享的数据资源更能促进方法的改进和创新。

另外, 虽然计算机预测方法在一些案例中表现出色, 但其准确性和可靠性仍有待提高。模型的复杂性、训练数据的质量和数量、算法的选择和优化等都会影响模型的性能。计算资源是限制计算机预测方法发展的另一个重要因素。复杂的模型和大量的数据需要大量的计算资源和时间。优化算法和提高计算效率是解决这一问题的关键。此外, 更新计算机硬件也可以提高计算机的计算效率。

总之, 计算机预测方法在化合物毒性评估领域已经取得了显著的进展。通过集成先进的人工智能技术, 研究人员能够更准确、快速地预测化合物的毒性, 这对于化学、环境科学和药物开发等领域具有重要意义。未来, 通过国际合作引入新技术和方法并制定标准化规范, 将进一步提升计算机预测模型的性能和应用范围。

参考文献:

- [1] TANG W, CHEN J, WANG Z, et al. Deep learning for predicting toxicity of chemicals: a mini review[J]. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev, 2018, 36(4): 252-271.
- [2] 沈超, 孙成亮, 郭沛霖, 等. 计算毒理学在农药毒性预测、活性筛查及风险评估中的应用[J]. 植物保护, 2023, 49(5): 279-287.
- [3] 谭壮生, 齐丽娟, 敬海明, 等. 基于QSAR预测模型的有机磷类农药急性毒性的健康风险评估[J]. 毒理学杂志, 2023, 37(3): 189-193.
- [4] KAVLOCK R, CHANDLER K, HOUCK K, et al. Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management[J]. Chem Res Toxicol, 2012, 25(7): 1287-302.
- [5] 徐雯丽, 黄迎, 胡静, 等. 化合物毒性预测及其应用[C]//2017(第三届)毒性测试替代方法与转化毒理学(国际)学术研讨会论文集, 2017.
- [6] 孙也之. QSAR模型的构建、应用与研究[D]. 沈

- 阳:中国医科大学,2009.
- [7] 包彤,李潍,于项毅,等. 面向化学品风险管理的数据库比较与分析[J]. 环境化学,2023,42(10):3381-3395.
- [8] 唐伟豪. 化学品线粒体毒性的机器学习分类模型[D]. 大连:大连理工大学,2021.
- [9] FENG H, ZHANG L, LI S, et al. Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints[J]. Toxicol Lett,2021,340:4-14.
- [10] MORIWAKI H, TIAN Y S, KAWASHITA N, et al. Mordred: a molecular descriptor calculator[J]. J Cheminform,2018,10(1):4.
- [11] WU Z, ZHU M, KANG Y, et al. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets[J]. Brief Bioinform,2021,22(4):1-17.
- [12] 钱跃言,吴永刚. 快速有效评估毒性:当今,以计算机为基础的软件预测化合物的毒性,节省时间和费用,减少动物试验[J]. 浙江化工,2000(4):53-55.
- [13] LIU L, YANG H, CAI Y, et al. In silico prediction of chemical aquatic toxicity for marine crustaceans via machine learning[J]. Toxicol Res (Camb),2019,8(3):341-352.
- [14] WANG F, YANG J F, WANG M Y, et al. Graph attention convolutional neural network model for chemical poisoning of honey bees' prediction[J]. Sci Bull (Beijing),2020,65(14):1184-1191.
- [15] KUMAR R, KHAN F U, SHARMA A, et al. A deep neural network-based approach for prediction of mutagenicity of compounds[J]. Environ Sci Pollut Res Int,2021,28(34):47641-47650.
- [16] SUN L, ZHANG M, XIE L, et al. In silico prediction of boiling point, octanol-water partition coefficient, and retention time index of polycyclic aromatic hydrocarbons through machine learning[J]. Chem Biol Drug Des,2023,101(1):52-68.
- [17] XU M, YANG H, LIU G, et al. In silico prediction of chemical aquatic toxicity by multiple machine learning and deep learning approaches[J]. J Appl Toxicol,2022,42(11):1766-1776.
- [18] ZHANG H, KANG Y-L, ZHU Y Y, et al. Novel naïve Bayes classification models for predicting the chemical Ames mutagenicity[J]. Toxicology in Vitro,2017,41:56-63.
- [19] 张兆津. 基于深度学习框架的药物特性预测[D]. 兰州:兰州大学,2018.
- [20] LYSENKO A, SHARMA A, BOROEVIKH K A, et al. An integrative machine learning approach for prediction of toxicity-related drug safety[J]. Life Science Alliance,2018,1(6):e201800098.
- [21] YLIPAA E, CHAVAN S, BANKESTAD M, et al. hERG-toxicity prediction using traditional machine learning and advanced deep learning techniques[J]. Curr Res Toxicol,2023,5:100121.

Application and prospects of computational models in predicting compound toxicity

Xu Zhiwang, He Wangqiu, Kong Ren*

(Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou 213001, China)

Abstract: Toxicity research is an important and complex scientific task that involves assessing the potential hazards of chemicals and other substances to human health, and plays an important role in environmental protection and sustainable development. Traditional toxicological experiments based on molecular, cellular, and animal models not only have high cost and time-consuming issues, but also have limitations in being unfriendly to animals. With the rapid development of computer technology, the research works of computational toxicology models have been emerged. Compared to traditional toxicology experiments, computational toxicology models have the characteristics of high efficiency, low cost, and convenience, gradually becoming one of the hot topics in toxicology research. This article introduces the research methods for predicting compound toxicity using computational models, including machine learning, and their applications in environmental protection, toxicity risk assessment, and drug design. It further discusses the shortcomings and potential development directions of computational models in toxicity prediction.

Keywords: computational toxicology; machine learning; toxicity prediction

文章编号: 1007-1423(2024)07-0086-04

DOI: 10.3969/j.issn.1007-1423.2024.07.015

基于粒子群优化随机森林算法的河南省葡萄价格预测研究

王 哲*, 张文慧, 付金鹏, 杨进进

(华北水利水电大学信息工程学院, 郑州 450000)

摘要: 河南省是中国重要的葡萄种植基地之一, 葡萄种植面积广泛且产量丰富。因此预测河南省葡萄价格至关重要。为了提高葡萄价格预测准确性和稳定性, 提出一种 PSO-RF 预测模型, 该模型用粒子群(PSO)优化随机森林中决策树深度, 选择最优预测器个数, 进一步得到最优预测器组合。实验结果表明, 相比于单一的随机森林预测模型, PSO-RF 模型具有更高的预测精度, MAE 仅为 0.0095, R^2 达到 0.968。因此 PSO-RF 预测模型可以对河南省葡萄价格进行更准确的预测。

关键词: 葡萄价格预测; 决策树; 随机森林; 粒子群算法

0 引言

葡萄作为中国最受欢迎的水果之一, 被誉为“水果之王”。我国是世界最大的鲜食葡萄生产国和消费国, 近 20 年来, 河南省葡萄的种植面积增长较快且葡萄产业已成为很多地方农业生产的支柱产业, 在乡村振兴、精准扶贫工作中发挥着重要作用。随着葡萄种植面积和产量持续增长, 河南省葡萄产业面临的问题也日益突出, 例如鲜食葡萄集中上市导致阶段性供过于求、市场价格波动大。因此, 利用价格预测方法准确把握葡萄价格波动趋势, 为农户提供即时的信息, 对促进葡萄市场发展具有现实意义。

近年来, 农产品价格预测已经成为了众多研究者关注的焦点, 尤其在食品安全和全球经济不稳定的背景下。不同的研究者使用了各种不同的方法来预测农产品的价格, 如机器学习、数据挖掘技术等。

为了寻找更为准确的价格预测手段, 何文靓等^[1]根据生猪价格波动规律, 利用随机森林模型进行预测, 但是没考虑到随机森林模型容易过拟合的问题。王一飞^[2]提出 PSO-Prophet 模型预测农产品价格, 但粒子群优化更适合连续

性数值。于浩^[3]利用 PSF-RF 模型进行股票预测, 但重点研究的是处理特征, 并未对随机森林本身进行优化。张禄^[4]通过深入分析影响玉米价格的多种因素, 采用相应的统计和模型方法对玉米价格进行准确预测。刘莲莲^[5]研究大蒜价格变动的规律性, 为农产品价格波动的预测提供新的视角。杜茜亚^[6]利用随机森林算法开发了一个海兰褐蛋鸡产蛋率预测系统, 证明随机森林在农业预测性能良好, 但在处理复杂的非线性关系和大量数据特征时的有效性有待提升。唐玉辰^[7]利用随机森林算法探讨了影响中国绿色债券发行价格的因素并预测发行价格, 发现预测精度不高。曹洁^[8]通过随机森林模型对二手车的价值进行评估, 发现随机森林参数影响预测精度。张兵等^[9]采用 SARIMA 模型对河北省葡萄价格进行预测分析, 该研究通过时间序列分析方法, 发现传统预测方法预测葡萄价格的预测效果差, 模型训练时间长。

通过阅读大量文献发现, 随机森林回归已在生猪、股票价格预测、农产品预测等方面被部分学者采用, 并获得良好效果。考虑到生猪、股票价格预测与水果市场的价格预测在某些方

收稿日期: 2023-11-20 修稿日期: 2023-12-13

作者简介: *通信作者: 王哲(2000—), 女, 河南商丘人, 硕士研究生, 研究方向为农业工程与信息技术, E-mail: 1183837332@qq.com; 张文慧(1999—), 女, 河南濮阳人, 硕士研究生, 研究方向为大数据与云计算; 付金鹏(1998—), 男, 河南漯河人, 硕士研究生, 研究方向为农业工程与信息技术; 杨进进(1994—), 女, 河南商丘人, 硕士研究生, 研究方向为图像处理与模式识别

面有较强的相似之处，且众多研究均表明随机森林在预测能力上表现出色，泛化能力强，基于此，本文采用随机森林方法来研究葡萄市场的价格预测。

上述方法虽然一定程度上能够获取预测价格的准确率，但是模型单一，随机森林容易有过拟合问题。粒子群算法是一种基于群体智能的优化算法，它通过模拟鸟群的社会行为来搜索最优解。粒子群算法可以优化随机森林的决策树深度和预测器个数，它具有较强的全局寻优能力，可以帮助随机森林预测模型解决容易过拟合的问题，提高模型预测精度。基于此，本文提出一种改进的河南省葡萄价格预测模型 PSO-RF。

1 预测模型

1.1 数据来源及影响因素确定

依据《河南省统计年鉴》《中国农村统计年鉴》、国家统计局相关网站、农产品价格信息网，选取河南省 2017—2023 年月度葡萄市场价格。

通过深入分析现有的研究文献，对影响河南省葡萄价格的因素进行了归纳和整合^[10]。综合考虑，最终确定了种植面积、产量、出口价格指数和食用农产品价格指数^[11]这四个关键变量作为本次研究的主要影响因素。

1.2 PSO-RF 河南省葡萄价格预测模型的建立

1.2.1 RF 葡萄价格预测模型建立

随机森林包含决策树 (Decision Tree) 算法，该算法是一种树形结构的机器学习算法，用于处理回归和分类问题。一般决策树由根结点、内部结点和若干叶子结点构成。其中根节点和内部节点应用于同一个属性测试，各个样本根据属性测试结果，被划分其余子结点中，重复不断递归此过程，一直到结点上的样本无法划分，此时的叶子结点相对应一个决策的结果。

此节探讨随机森林中决策树的深度，并深入分析其对模型整体性能的影响。目标是确保在模型训练过程中达到最小的偏差。将葡萄种植面积、产量、出口价格指数、食用农产品价格指数四个影响因素作为输入变量，河南省葡萄价格作为输出变量，通过试错法进行测试，经过反复大量实验训练模型，发现设置随机森林模型决策树最大深度为 3，预测器个数为 9 时

偏差最小。随机森林预测模型如图 1 所示。

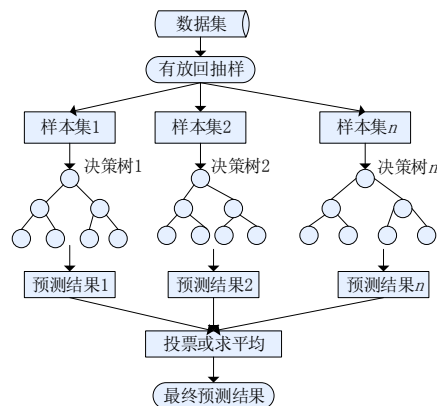


图 1 随机森林预测模型流程图

1.2.2 PSO-RF 算法设计

粒子群 (particle swarm optimization, PSO) 是一种演化算法，模拟了鸟群或鱼群等生物群体中个体之间的社会行为，通过协作和信息共享来搜索最优解。在粒子群算法中，每个粒子根据自己的当前位置和速度，以及历史最佳位置和全局最佳位置的信息，通过迭代更新这些值来找到最优解，最终达到全局寻优的目的。粒子群算法的全局寻优能力，可以用来优化随机森林 (RF) 预测模型，寻找个人最佳值 (pbest) 和全局最佳值 (gbest)，从而弥补随机森林预测模型容易过拟合的缺点。

PSO-RF 预测模型具体实现步骤如下：

Step1 设置相关参数。设置粒子数量为 100、最大迭代次数为 10。设置个体速度的上界为 1，下界为 -1，搜索范围是 2 到 10。设置学习因子 $c_1=0.5$ ， $c_2=0.3$ 和惯性权重 $w=0.9$ 。

Step2 初始化粒子群。每个粒子的位置表示一个可能的随机森林模型参数组合。初始速度可以设置为 0 或小的随机值。

Step3 设置适应度函数。适应度函数基于交叉验证的 R^2 分数来评估随机森林的预测性能。

$$Fitness(p) = -MSE(p) \quad (1)$$

Step4 更新粒子速度和位置。

使用以下公式更新每个粒子 i 的速度：

$$V_i(t+1) = W * V_i(t) + c_1 * rand_1() * (pbest_i - x_i(t)) + c_2 * rand_2() * (gbest_i - x_i(t)) \quad (2)$$

式中： $pbest_i$ 是粒子*i*到目前为止的最佳位置， $gbest$ 是整个群体的最佳位置， $rand_1()$ 和 $rand_2()$ 是随机数。

使用以下公式更新每个粒子*i*的位置：

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (3)$$

Step5 评估新的位置。对于每个粒子位置，使用适应度函数评估其性能。

Step6 更新个体和全局最佳值。如果一个粒子在新位置找到了更好的适应度，更新其个体最佳位置 $pbest$ 。如果在新位置找到了更好的群体适应度，更新全局最佳位置 $gbest$ 。

Step7 检查终止条件。如果达到最大迭代次数或满足其他停止准则，停止算法。否则，返回Step4。

Step8 通过粒子群算法优化后，将 $gbest$ 输入随机森林预测模型进行训练模型，得到最优解。

2 仿真实验结果与分析

收集河南省2017—2023年月度葡萄市场价格数据，对应的月度影响因素作为输入变量，其中75%的数据作为训练集用于模型的训练，25%的数据作为测试集进行预测模型评估。

2.1 评价指标确定

为评估 PSO-RF 预测模型的有效性和准确性，考虑均方误差 (MSE)、平均绝对误差 (MAE)、均方误差根 ($RMSE$) 和决定系数 (R^2)。决定系数是用于评估回归模型拟合程度的一种常用指标，它表示模型能够解释因变量的方差的比例。

四个指标作为模型评价指标，具体公式如下。式中 y_{pre} 、 y_{test} 分别为预测值、真实值。 MSE 计算了所有样本预测值与真实值之差的平方的平均值。 MSE 的取值范围从 0 到正无穷，越接近 0 表示模型的预测精度越高。式(7)中 SSR 代表回归平方和，表示回归模型拟合后残差的平方和。 SST 代表总平方和，表示因变量与其均值的差别的平方和。 R^2 的计算公式可以解释为拟合模型能够解释的因变量的方差所占据的比例。 R^2 的取值范围从 0 到 1，越接近 1 表示模型对数据的拟合越好。

$$MSE = \frac{1}{n} \sum_{i=1}^m w_i (y_{pre} - \hat{y}_{test})^2 \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{pre_i} - y_{test_i}| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n |y_{pre_i} - y_{test_i}|^2} \quad (6)$$

$$R^2 = 1 - \left(\frac{SSR}{SST} \right) \quad (7)$$

2.2 仿真实验与分析

2.2.1 模型预测结果分析

图2描绘了运用随机森林(RF)与经粒子群优化(PSO)增强的随机森林(PSO-RF)两种方法对葡萄价格进行预测的结果。通过比较，发现 PSO-RF 模型相对于 RF 模型在预测精度方面有显著提高。

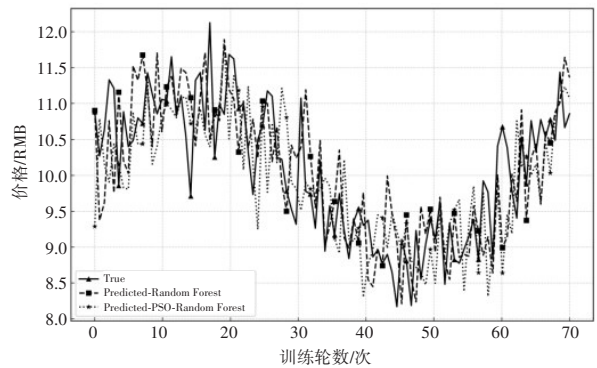


图2 预测值与真实值对比

2.2.2 模型误差分析

由于 MSE 、 $RMSE$ 、 MAE 和 R^2 四种评价指标可以全面地反映预测结果的准确性和稳健性，输出了两种预测模型在这四个指标上的对比图，如图3所示。

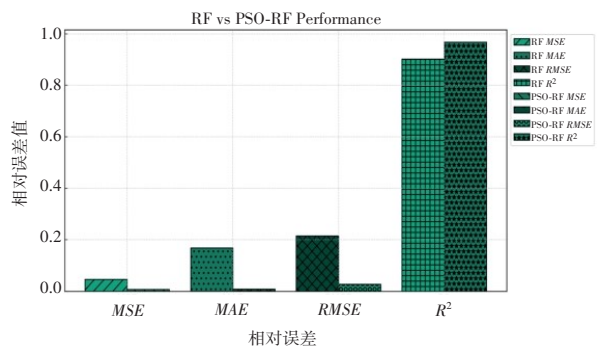


图3 各项误差对比图

从图3可以直观地看出,优化后的预测模型在各项指标上都有所改进,特别是在 MSE 和 $RMSE$ 上的降低以及 R^2 的提高,均显示了其更出色的预测能力。

从表1也明显观察到,优化后的河南省葡萄价格预测模型在各项评估指标上都表现得更为优越,这意味着其预测能力更为精准和稳定。

表1 模型评价指标

模型	MSE	MAE	$RMSE$	R^2
RF	0.0467	0.1688	0.2162	0.902
PSO-RF	0.0078	0.0095	0.0279	0.968

3 结语

以四个关键影响因素作为河南省葡萄价格随机森林中决策树的输入变量,构建一个随机森林预测模型,并采用粒子群算法进行模型优化,确保随机森林中的决策树深度和预测器数量达到最佳状态。完成了模型的训练、验证和误差分析过程。结果显示,经过粒子群算法优化随机森林预测模型的表现更为出色,预测准确度得到显著提升。葡萄种植户今后应准确把握葡萄市场价格的变动规律,在生产中着重提高产品品质,在现代消费者追求食品质量与安全的前提下,实现葡萄生产的绿色、高质、可持续发展。在产品进入上市期时期,可以利用电商平台新销售模式拓展销售渠道,以减缓葡

萄市场价格波动,降低收入风险。今后应该着眼于市场的需求变化,提高葡萄品质,加强对外省及国际市场的输出,提高收益。

参考文献:

- [1] 何文靓,付莲莲,廖静萍.基于随机森林模型的生猪价格预测及调控机制研究[J].价格月刊,2023(1):21-27.
- [2] 王一飞.基于深度学习的农产品价格预测研究[D].郑州:河南农业大学,2023.
- [3] 于浩.基于随机森林和LSTM网络的股指预测模型优化研究[D].哈尔滨:哈尔滨商业大学,2023.
- [4] 张禄.玉米价格影响因素分析及预测模型研究[D].泰安:山东农业大学,2023.
- [5] 刘莲莲.大蒜价格波动特征分析及价格预测研究[D].泰安:山东农业大学,2023.
- [6] 杜茜亚.基于随机森林算法的海兰褐蛋鸡产蛋率预测系统研究[D].秦皇岛:河北科技师范学院,2023.
- [7] 唐玉辰.基于随机森林算法的中国绿色债券发行价格影响因素研究[D].上海:华东师范大学,2022.
- [8] 曹洁.基于随机森林模型的二手车价值评估研究[D].石家庄:河北经贸大学,2020.
- [9] 张兵,白思远,蒋冰执,等.基于SARIMA模型的河北省葡萄价格预测分析[J].安徽农学通报,2019,25(17):159-161.
- [10] 王鹏.苹果品质评估与价格预测模型研究[D].泰安:山东农业大学,2020.
- [11] 李俊青.水果市场价格预测与波动预警研究[D].武汉:华中农业大学,2016.

Research on grape price forecasting of Henan Province based on particle swarm optimization random forest algorithm

Wang Zhe*, Zhang Wenhui, Fu Jinpeng, Yang Jinjin

(College of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450000, China)

Abstract: Henan Province is one of the important grape planting bases in China, with extensive grape planting area and abundant yield. Therefore, it is very important to predict the price of grapes in Henan Province. In order to improve the accuracy and stability of grape price prediction, a PSO-RF prediction model is proposed in this paper. In this model, particle swarm optimization (PSO) is used to optimize the depth of decision tree in random forest, select the optimal number of predictors, and further obtain the optimal combination of predictors. The experimental results show that compared with the single random forest prediction model, the PSO-RF model has higher prediction accuracy, MAE is only 0.0095, R^2 is 0.968. Therefore, the PSO-RF prediction model can make a more accurate prediction of grape price in Henan Province.

Keywords: grape price forecast; decision tree; random forest; particle swarm optimization

开发案例

文章编号: 1007-1423(2024)07-0090-06

DOI: 10.3969/j.issn.1007-1423.2024.07.016

教师职教能力评估管理系统的设计与应用

李进豪*

(广东生态工程职业学院信息工程学院, 广州 510520)

摘要: 以OKR为教师职教能力评估管理系统的设计理念, 旨在基于数据分析为整体竞争力提升提供决策支持, 基于AES打造核心竞争力。基于实际工作项目的需求, 重点阐述教师职教能力评估管理系统的设计与实现方法, 并以实践案例佐证其应用效果。

关键词: 工作室; 教师; 能力评估; 数据分析; VBA

0 引言

职业教育作为国民教育体系的重要组成部分, 担负着传承技术技能、促进就业创业、乡村振兴等重要职责, 是实施人才强国战略的重要保障。积极推进“三教”改革, 精准聚焦, 培养“肯做事+能做事+做成事”的“经师+人师”统一者, 是推进专业高质量发展的需要, 是打造专业品牌的必由之路。其中, 如何基于数据分析为教师职教能力整体提升提供决策支持, 如何基于AES打造属于教师的核心竞争力, 值得我们研究与探索。以名师工作室为依托, 本文重点阐述教师职教能力评估管理系统的设计与应用成效, 旨在展现以OKR为核心的管理系统设计理念和管理系统在教师专业成长中的助推作用。

1 问题的提出

1.1 数据分析

落实教育立德树人的根本任务, 基于培养优秀的职业教育学科(课程)带头人并形成有核心竞争力的教学团队, 是广东省职业教育“双

师型”名教师工作室的初心使命。调研与实践表明, 教师成长的内因固然重要, 但外因能起催化促进作用, 充分发挥管理系统的数据分析功能, 将有利于实现精准式帮扶与指导, 是提升名教师工作室工作效率的有效方法。

1.2 AES

AES(经验积累与分享)是Accumulated Experience Sharing的缩写, 特色化(教师专业成长)的AES实际上是一个不断扩充的教育教学资源库, 有利于借鉴和传承, 有利于提升教师的教科研能力和专业能力等^[1]。调研与实践表明, 在可推广、可借鉴、可触摸的典型案例分析支持下, 有利于教师实现聚焦式成长, 是提升名教师工作室培养质量的有效支点。

2 教师职教能力评估管理系统的设计

2.1 需求分析

(1) 开发背景。本研究基于以下三个课题, 2022年广东生态工程职业学院服务乡村振兴重点课题: 基于名师工作室的职业教育师资培养帮扶路径研究; 2023年度广东省教育科学规划

收稿日期: 2023-10-27 修稿日期: 2023-11-23

基金项目: 广东生态工程职业学院服务乡村振兴重点课题(2022-06); 2023年度广东省教育科学规划课题(2023GXJK804); 广东生态工程职业学院教育教学改革研究与实践项目(2022zlgc-xj-jxgg16)

作者简介: *通信作者: 李进豪(1969—), 男, 广东汕头人, 硕士, 正高级讲师, 高级技师, 研究方向为教学团队建设、软件工程、智慧校园等, E-mail: 56014274@qq.com

课题：基于名师工作室的教师专业成长的研究与实践；2022年广东生态工程职业学院教育教学改革研究与实践项目：基于名师工作室提升课程思政育人有效性的研究与实践。

(2) “三师型”教师。一是指具备教育教学能力的教师身份：具有教师资格证，学生信赖，得到社会认可；二是指具备专业技术能力的工程师（技师）身份：行业认可的技术型人才，能服务企业；三是具备社会服务能力的培训师身份：具有社会服务能力，能助推行业（同行）前行。

(3) 需求分析。名师工作室是在主持人的引导下，以优秀骨干教师为培养对象，以师带徒为主要培养形式，指导提升学员的专业能力和水平，共同开展基于线上和线下的教育教学改革的实体与网络相结合的新型工作室^[1]。

站在成果导向的角度，主体程序（职教能力评估管理系统）主要功能包括：基于职教能力评估指标（师德师风、科研与社会服务、教学改革-质量工程、教学改革-竞赛、公共服务）相关信息的录入，以“成员”为管理对象系统自动形成“工作表”形式的职教能力评估资料记录。以个人预设的业绩目标为基准，系统自动统计各项指标的完成情况。

站在能力提升与成果管理的角度，辅助程序（个人备忘录管理系统）主要功能包括：基于“部门年度任务（含工作室）+教学团队年度任务+个人任务+归档资料”提高预设工作目标完成度的可预见性和可控制性，基于分类协助用户提升工作效率和完成质量（确保任务100%完成）。以用户（教师）为管理单元，根据登录用户自动匹配个人备忘录工作表，提供个人任务等相关信息的智能化管理，主要包括：个人备忘录登改与智能导入、智能筛选、智能归档、部门信息导入、教学团队信息导入、信息导出（共享）等。

2.2 基本框架

(1) OKR。OKR（目标与关键成果法），用于明确教学团队的“目标”以及明确每个目标达成的可衡量的“关键结果”，旨在确保教学团队共同成长，并集中精力做出可衡量的成果。于教师个人也是如此，即明确自己服务于学校（或教学团队）的目标（充分沟通后的共识）及每个目标达成的“显性”业绩。

(2) 基于OKR的设计理念。以OKR为教师职教能力评估管理系统的设计理念，旨在基于数据分析为整体竞争力提升提供决策支持，基于AES打造核心竞争力。在服务教师专业成长的前提下，确保教学团队达成预期的“显性”业绩。

(3) 整体框架

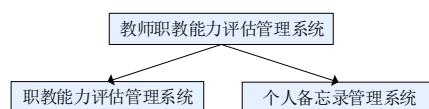


图1 教师职教能力评估管理系统

(4) 主体程序（职教能力评估管理系统）

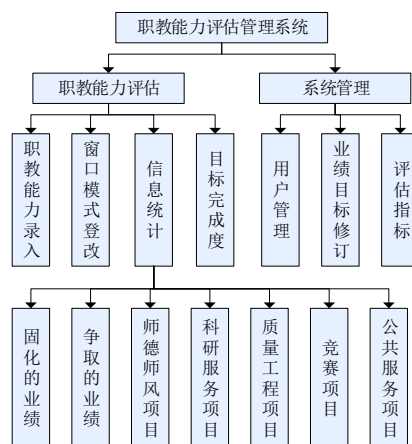


图2 职教能力评估管理系统

(5) 辅助程序（个人备忘录管理系统）

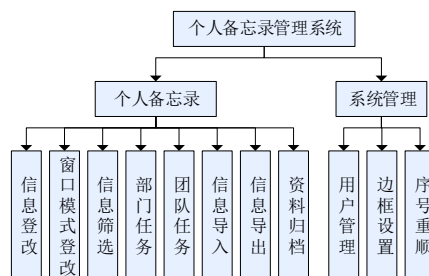


图3 个人备忘录管理系统

(6) AES资源

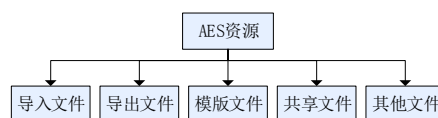


图4 AES资源

3 实现方法

3.1 开发平台与语言

开发平台：Office 2016(Excel)；开发语言：VBA；优势：运行环境简易；安全性：基于登录实现了数据的全方位保护。

3.2 核心技术

(1) 信息加密

```
mypassword = "ABC123"  
passnum = 9  
mypasswordnew = ""  
For j = 1 To Len(mypassword)  
mys = Mid(mypassword, j, 1)  
mypasswordnew = mypasswordnew + Chr(Asc(mys) +  
j - passnum)  
Next
```

■说明：mypassword为原始信息，mypasswordnew为加密后信息。作为一种常用的加密技术，加密参数j和passnum是可以改变的，以提高信息加密的安全性^[2]。

(2) 账号管理

```
myoldname = ComboBoxname.Text  
myi = ComboBoxname.ListIndex  
Sheet2.Cells(myi + 3, 2) = TextBoxname.Text  
For myk = 1 To Sheets.Count  
If Sheets(myk).Name = myoldname Then  
Sheets(myk).Name = TextBoxname.Text  
End If  
Next
```

■说明：myoldname为原账号名称，TextBoxname.Text为新账号名称，当账号名称改变时，对应的工作表名称也随之改变，实现智能化管理。

(3) 成员登录自动识别

```
xm = ThisWorkbook.mynamestr  
For myi = 1 To Sheets.Count  
If Sheets(myi).Name = xm Then  
Label88.Caption = myi  
End If  
Next
```

■说明：根据登录用户名与工作表名称的匹配实现成员登录的自动识别（myi为工作表的索引号），实现数据读写的一致性和安全性保护。

(4) 工作表隐藏

```
For myi = 1 To Sheets.Count
```

```
If Sheets(myi).Name = xm Or Len(Sheets(myi).Name)  
>= 12 Then  
Sheets(myi).Visible = True  
Else  
Sheets(myi).Visible = False  
End If  
Next
```

■说明：只有与用户名相同的工作表和公共工作表（名称长度大于或等于12）才为可见，其他工作表均为隐藏（私有工作表）。

(5) 自动获取当前日期

```
datelong = Format(Date, "yyyy/mm/dd")  
Me.ComboBox2.Text = Mid(datelong, 1, 4)  
Me.ComboBox3.Text = Mid(datelong, 6, 2)  
Me.ComboBox4.Text = Mid(datelong, 9, 2)
```

■说明：利用Date与Mid函数自动获取当前日期，以提高工作效率和录入的智能性。

(6) 自适应添加工作任务

```
For i = 2 To 50000  
If Sheet3.Cells(i, 1) <> "" Then  
If Sheet3.Cells(i, 6) = xm And Sheet3.Cells(i, 8)  
<> "是" Then  
Sheets(mysavenum).Cells(mynum, 1) =  
mynum - 1  
Sheets(mysavenum).Cells(mynum, 5) = Trim  
(Sheet3.Cells(i, 2))  
Sheets(mysavenum).Cells(mynum, 6) = Trim  
(Sheet3.Cells(i, 3))  
Sheets(mysavenum).Cells(mynum, 9) = "否"  
Sheets(mysavenum).Cells(mynum, 10) = "团队"  
Sheet3.Cells(i, 8) = "是"  
mynum = mynum + 1  
End If  
Else  
i = 60000  
End If  
Next
```

■说明：将团队教学工作任务中的个人工作任务自适应添加到个人工作表中，实现教学团队任务与个人工作任务智能对接。

(7) 序号重顺

```
For i = 2 To 50000  
If Sheet39.Cells(i, 1) <> "" Then  
Sheet39.Cells(i, 1) = i - 1  
Else
```

```
i = 60000
```

```
End If
```

```
Next
```

■说明：将指定工作表的序号进行重顺，以保证序号的唯一性。

(8) 单元格边框设置

```
wz = "A1:I" + Trim(Str(mynum))
```

```
Set rng = Range(wz)
```

```
rng.Borders.LineStyle = xlContinuous
```

■说明：mynum为工作表中有信息的单元格最后行号，该范围内(wz)的所有单元格边框均为实线，以提升报表的美观度。

(9) 工作表(信息)导入

```
Set sourceWorkbook = Workbooks.Open(This
```

```
Workbook.Path & "\input\" & "教学团队" & ".xlsx")
```

```
Set sourceWorksheet = sourceWorkbook.Worksheets
```

```
(1)
```

```
Set targetWorksheet = ThisWorkbook.Sheets(4)
```

```
Set sourceRange = sourceWorksheet.Range("A2:H1000")
```

```
Set targetRange = targetWorksheet.Range("A3")
```

```
sourceRange.Copy targetRange
```

```
ThisWorkbook.Sheets(4).Cells(1,1) = "临时导入  
的教学团队年度任务"
```

■说明：将指定工作簿的第一个工作表内容复制到系统工作簿的目标工作表。

(10) 工作表(信息)导出

```
Sheet39.Activate
```

```
ActiveSheet.Copy
```

```
ActiveWorkbook.Close SaveChanges:=True,
```

```
Filename:=ThisWorkbook.Path & "\xlsx\" & "教学  
团队" & datelong & ".xlsx"
```

■说明：将当前工作表导出为一个新工作簿(相关内容保存在第一个工作表)。

(11) PDF 报表生成

```
Sheet39.PageSetup.CenterHeader = "教学团队备忘录  
(&D & T)第&P 页，共&N 页"
```

```
Sheet39.ExportAsFixedFormat xlTypePDF, This
```

```
Workbook.Path & "\pdf\" & "教学团队" &
```

```
datelong & ".pdf"
```

■说明：将当前工作表以pdf报表格式导出，有效范围为设置了实线边框的单元格(动态设置)。

3.3 常用小技巧

(1) 输入式对话框

```
mysuper = InputBox("请输入维护密码:", "个人备  
忘录管理系统")
```

■说明：弹出一个输入式对话框，供用户输入相关信息。

(2) 选择式对话框

```
myqr = MsgBox("真的要修改吗?", vbYesNo,  
"警告")
```

■说明：弹出一个拥有两个按钮(是、否)的对话框，供用户进行选择。

(3) 居中显示

```
Sheet1.mygo.Left = (ActiveWindow.UsableWidth-
```

```
Sheet1.mygo.Width) / 2
```

■说明：控件(mygo)居中显示，提升界面设计的美观度。

(4) 保存

```
ThisWorkbook.Save
```

■说明：保存当前工作簿。

(5) 关闭

```
ThisWorkbook.Close
```

■说明：关闭当前工作簿。

(6) 当前窗口

```
Sheets(mysavenum).Activate
```

```
Range("a3").Select
```

■说明：激活指定的工作表，并将工作表切换为当前窗口。

(7) 动态文件名

```
datelong = Format(Date, "yyyy-mm-dd")
```

```
myfilename="教学团队" & datelong & ".pdf"
```

■说明：基于当前日期对报表文件进行动态命名。

(8) 文件检测

```
myfilename = ThisWorkbook.Path & "\input\"  
& "教学团队" & ".xlsx"
```

```
If Dir(myfilename) <> "" Then
```

```
'文件存在
```

```
Else
```

```
'文件不存在
```

```
End If
```

■说明：利用 ThisWorkbook.Path 动态获取当前路径，利用 Dir 检测文件是否存在，以提升容错性。

4 运行界面

4.1 职教能力评估管理系统

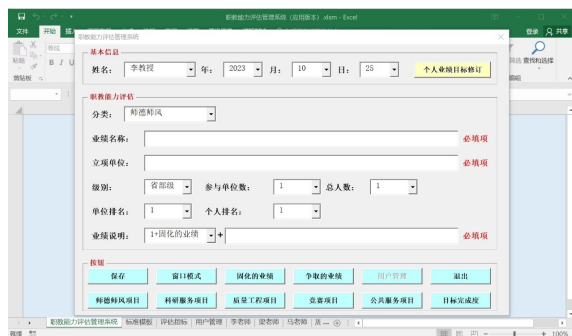


图5 职教能力评估管理系统运行界面

4.2 个人备忘录管理系统



图6 个人备忘录管理系统运行界面

5 研究成果与应用效果

5.1 研究成果

(1) 计算机软件著作权。以OKR为设计理念的教师职教能力评估管理系统在实践过程中,成功申请了计算机软件著作权1项(职教能力评估管理系统,登记号:2023SR0349261)。基于推广应用申请了相关计算机软件著作权3项: BH蔬菜销售管理系统、红色文化融入爱国主义教育管理系统、乡村振兴项目管理系统。

(2) 主题征文。以培养“肯做事+能做事+做成事”的“经师+人师”统一者为宗旨,论文《基于立德树人的教师教学胜任力评价体系的思考》荣获广东生态工程职业学院2023年教育评价改革主题征文一等奖。

5.2 应用效果

(1) 案例1: 名师工作室学员成长

近两年,共主持(含参与)课题27项,获得成果20项,荣誉33项,指导学生获奖20项,开设公开课5次,专题讲座7次等。

(2) 案例2: 名师工作室成员成长

近两年,共主持(含参与)课题16项,各类成果23项,专题讲座(含分享)12次等。

5.3 典型案例

(1) 帮扶路径。在管理系统支持下,通过“共享文件+多级联动”不断提升教师的专业能力和核心竞争力,以核心成员(中山市名师工作室主持人、四会市名师工作室主持人、博罗县名师工作室主持人等)为重要链接点,点、线、面相结合,助力更多的职业教育教师专业成长。

(2) 课题研究。2023年,在管理系统支持下,工作室学员申报的课题《特教中职残疾人融合教育的探索研究——以肢体残疾学生的心理融合为例》经过校内竞争性推荐,最后在与高校同台竞争且立项率仅为34%的课题评审中获得2023年度广东省残疾人事业理论与实践研究课题立项。目前,该课题已获得计算机软件著作权1项(心理融合辅助管理系统),开展省级主题分享1次,撰写相关论文1篇等。

(3) 专题讲座。2023年,在管理系统支持下,工作室主持人先后在多所兄弟院校等开设专题讲座《做“经师”和“人师”的统一者》《做“德技双馨”的职教老师》。

6 结语

基于数据分析提供决策支持,基于AES打造核心竞争力,依托“名师工作室+管理系统”,推行“站位高+反思深+业绩强”的多元能力重构提升,以职业能力为导向、以产教融合为抓手、以综合能力为核心、以真实任务为载体、以“学习圈+成长圈”为平台的融合贯通培养,有利于教师快速成长,有利于打造高水平教学团队。文中提及的代码全部在Window 10+Office 2016+VBA环境下调试通过,并在实际工作的应用效果良好。

参考文献:

- [1] 李进豪. 职业教育的探索与实践[M]. 广州:羊城晚报出版社,2020.
- [2] 李进豪. 基于VBA的异构数据源自适应读写技术研究与实践[J]. 软件导刊,2017(4):152-155.

Design and application of teacher vocational education ability evaluation management system

Li Jinhao*

(College of Information Engineering, Guangdong Eco-Engineering Polytechnic, Guangzhou 510520, China)

Abstract: The design concept of OKR as teacher vocational education ability evaluation management system aims to provide decision support for overall competitiveness improvement based on data analysis and build core competitiveness based on AES. Based on the needs of actual work projects, this paper focuses on the design and implementation of teacher vocational education ability assessment management system, and demonstrates its application effect with practical cases.

Keywords: studio; teachers; ability evaluation; data analysis; VBA

(上接第69页)

Resource allocation in cellular networks based on edge caching

Pan Yu, Hua Wei*

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: The widespread application of Internet of Things (IoT) and the rapid development of communication technologies have led to the explosive growth of mobile data traffic, which has brought huge energy consumption to network operators. Therefore, improving energy efficiency is an important challenge in the 5G and 6G era. In this paper, we make resource allocation in cache-enabled cellular networks and the optimal solution can be obtained by modeling the problem as a mixed integer programming problem. The objective is to jointly optimize the base station transmit power and task transfer scheduling while satisfying the quality of service (QoS) requirements, and explore the impact of cache hit rate on the total energy consumption. Simulation results show that our proposed scheme performs well in terms of energy consumption and that increasing the cache hit rate has the effect of reducing the total energy consumption.

Keywords: edge caching; resource allocation; edge cloud computing

文章编号: 1007-1423(2024)07-0096-05

DOI: 10.3969/j.issn.1007-1423.2024.07.017

智能终端 APP 适老化设计研究

崔晓旭¹, 徐文燕^{2*}

(1. 天津市电子信息技师学院信息技术系, 天津 300350; 2. 天津市电子信息技师学院商务服务系, 天津 300350)

摘要:随着我国通信技术与产业的巨大发展, 智能终端产品已在我国迅速普及, 但由于部分老年人随着年龄的增长, 视觉、听觉、运动、语言、感知等能力不同程度地下降, 在使用智能终端 APP 过程中面临着各种各样的困难和问题, 无法充分享受移动终端智能化服务带来的便利, “数字鸿沟”问题日益凸显, 因此, APP 适老化设计研发是时代发展的必然趋势。通过分析调研, 从老年群体的生理特征和心理需求出发, 探究适老化 APP 的设计策略, 并完成了设计实践, 助力我国智能终端适老化 APP 的设计改造研发。

关键词: 智能终端; 适老化; APP; 设计实践

0 引言

据国家统计局 2023 年数据显示, 我国 60 岁以上人口已超 2.8 亿, 随着老龄化的不断加剧, 预计到 2050 年, 中国 65 岁及以上的老年人将达到 3.8 亿。我国老龄化日趋严峻, 而随着互联网、大数据、人工智能等信息技术的快速发展, 智能终端产品得到广泛普及和应用, 在这一过程中, 部分老年人无法享受数字技术带来的便利, “数字鸿沟”问题日益凸显^[1]。

2021 年 1 月, 工信部在全国范围内组织开展了为期一年的互联网应用适老化及无障碍改造专项行动, 并在同年 4 月发布了《互联网网站适老化通用设计规范》《移动互联网应用适老化设计规范》等文件。经过一年多的实践, 多个互联网平台推出了适老化措施, 助力老年人融入互联网, 各大手机厂商也推出了老人模式、长辈模式等功能。然而, 单纯的放大字体图标、加入语音交互功能、一键直连人工客服、设置关怀模式等只是缓解了界面操作的不便性, 并没有从根本上解决老年人使用互联网及移动终端时遇到的困难^[2]。本文尝试从老年群体的具体需求出发, 提出真正符合老年人操作, 界面友好、功能实用的适老化 APP 设计策略, 助力

智能终端适老化应用发展^[3]。

1 用户调研与 APP 设计分析

1.1 用户群体分析

部分老年人随着年龄的增长, 视觉、听觉、运动、语言、智力等会有不同程度的下降, 在使用智能终端过程中面临着各种各样的困难和问题, 无法充分享受移动终端智能化服务带来的便利^[4]。

老年人使用互联网平台及移动终端的困难主要体现在以下几个方面:

——视力减退, 无法看清智能终端屏幕信息内容;

——听力减退, 无法听清通话内容和音频相关信息;

——运动能力减退, 无法进行点击操作或触控方面的复杂操作;

——认知能力减退, 无法进行复杂应用的操作;

——不会使用, 对于智能手机或复杂的移动应用不会使用或不能熟练使用;

——紧急状态, 由于身体原因, 遇到紧急状态时无法快速发送紧急救助信息或帮助信息^[2]。

收稿日期: 2023-10-26 修稿日期: 2023-11-28

作者简介: 崔晓旭(1987—), 女, 吉林松原人, 硕士, 讲师, 研究方向为网络营销策略与选择、电商运营模式研究与分析;

*通信作者: 徐文燕(1987—), 女, 新疆昌吉人, 硕士, 讲师, 研究方向为网络营销策略与选择、电商运营模式研究与分析, E-mail: 25434106@qq.com

1.2 改造偏好需求分析

对于智能终端APP适老化的改造和设计,老年消费群体的需求从高到低依次是操作的便捷性、功能的实用性、有效保障财产安全、界面风格简单清晰、保护隐私信息等。其中,操作便捷、功能实用更是老年群体迫切期待的改造优化方向。具体到页面的设计、风格的设定、安全的控制等方面,通过调研,总结出老年群体的具体期待有:界面简约,有良好的引导和层级区分;字体、图标放大,有明显的辨识度;操作友好,有辅助的读屏、语音交互等功能^[5]。

1.3 功能模块需求分析

APP功能的筛选是适老化评价的重要指标,根据调研,我们从社交通讯、餐饮购物、导航出行、运动健康、实用工具、教育新闻、影音娱乐、休闲游戏、金融理财、便捷生活十大类,基本覆盖了老年消费者关于衣、食、住、行、娱、教、医等领域进行分析,得出以下结论,餐饮购物、便捷生活、社交通讯、教育新闻类APP在整体设计时应更多地考虑到了老年人的需要,能够较多兼顾老年人的身心特点。

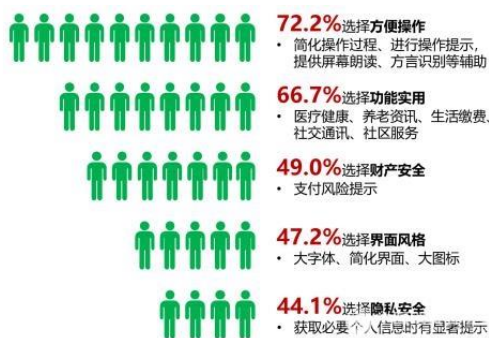


图1 老年消费者APP改造偏好

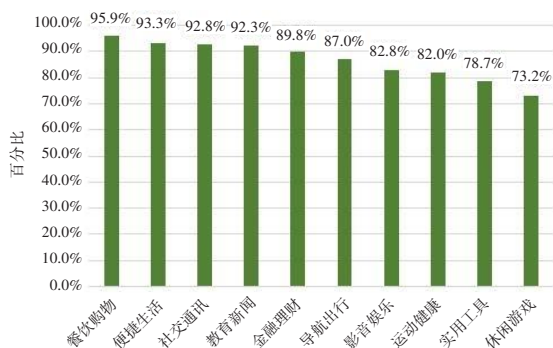


图2 老年消费者APP功能需求

2 APP适老化设计策略

老年群体由于身体素质、认知水平、行为能力受限,在使用智能终端APP时经常会遇到技术层面的操作困境,同时也受到弹窗广告、恶意软件的干扰,因此,适老化APP的设计要充分考虑老年群体的生理特征、使用习惯和心理需求,从产品定位、技术指标优化、特色功能、交互体验等方面进行分析研究。

2.1 明确产品定位

适老化APP应根据当前市场行情、互联网思维趋势,致力于为老年人提供全面、易于使用的集新闻资讯、娱乐服务、交通出行、金融服务、医疗健康、紧急处理等服务一体化的全平台应用软件,打破原有各种生活服务APP间壁垒,服务老年人多种生活场景,让老人用的安心,用的省心。

2.2 优化技术指标

适老化APP的研发应参考电信终端产业协会发布的重要文件,对APP从屏幕显示、读屏技术、触控交互等方面进行技术优化。将APP文字显示效果设置为不小于印刷文字4号字;文字视觉呈现为高对比度;图标采用方形图标,直线边长为12毫米;屏幕放大程度可达200%。页面采用的色彩基调偏暖色为主,方块按键以醒目的颜色为辅,图形的设计加上中间镂空,增强设计感,简单大方。

2.3 开通特色功能

除了助力老年人拥抱生活中常见数字化场景外,适老化APP还应增设“远程协助”及“紧急报警”等功能。远程协助可以帮助老年人和家人建立远程互助,实现远程视频通话、屏幕共享、同屏批注等。紧急报警功能可以实现老年人在锁屏状态下,向紧急联系人发出报警信息、发送实时定位,实现地理围栏功能。这大大提高了老年人在使用智能终端设备时的安全性及便捷性,也从心理上对老年人起到了慰藉作用。

2.4 交互体验良好

适老化APP应通过触屏、读屏等移动端适老化技术的综合应用,让老人与手机的交互变得更为人性化,方便老人更好地使用手机功能,视觉、听觉和触觉的感受更为直观。优质的内

容能增加用户的使用频率和活跃度，氛围感元素也提升了老人的使用感受和情感体验。

3 APP适老化设计实践

3.1 基本框架搭建

为了更好地使老年群体对适老化APP从感官刺激和情感认知上认可，我们将此次APP的名称设计为“老友季”，取自美国著名电视情景喜剧《老友记》的谐音，“老友季”APP基于前期对产品目标用户进行问卷调查，选取目标用户最为关心的服务模块进行开发设计，首期推出八大模块，分别是新闻资讯、社交通讯、生活购物、支付服务、交通出行、医疗健康、远程协助和紧急报警，产品针对性较强，适老化优势明显。



图3 “老友季”首页展示



图4 首页下滑展示

(1) 新闻资讯：用户可以观看最新的新闻事件、热点话题、人物动态、产品资讯并浏览短视频、电影和其它媒体。

(2) 社交通讯：用户可以扩大和维护社交网络，与居民之间建立联系和互动，分享信息、发布社区活动、组织线下聚会、参与社区建设等。

(3) 生活购物：用户可以网上点餐、买药、买菜、日常购物等，并通过优化的界面和功能体验一站式购买退换服务体验。

(4) 支付服务：用户可以完成线上生活交费及网络支付，线下零售支付等常见场景的支付操作。

(5) 交通出行：用户可以使用“一键叫车”“语音叫车”“实时公交”功能，解决短途出行问题，也可体验老龄群体的旅行项目，如语音导览、旅游咨询热线等。

(6) 医疗健康：用户可以根据年龄和身体情况，制定实时的健康管理计划，也可以进行网上就医、预约挂号等操作。

(7) 远程协助：用户可以通过设置和家人建立远程互助，实现远程视频通话、屏幕共享、同屏批注等，辅助完成其他软件相关操作。

(8) 紧急报警：用户遇到紧急情况时，可以在锁屏状态下通过点击屏幕或实体按键，向紧急联系人发出报警信息、发送实时定位。

3.2 技术指标设定

参考电信终端产业协会发布的《移动终端适老化技术要求》和《移动终端适老化测试方法》两份文件，对“老友季”APP从屏幕显示、读屏技术、触控交互等方面进行优化。

(1) 屏幕显示：适老化屏幕显示技术包括字体放大、文本对比调整、对比度调整、屏幕放大等。按照电信终端产业协会 T/TAF090—2021《移动终端适老化技术要求》标准，“老友季”APP主要信息的文字大小显示效果不小于印刷文字4号字；文字的视觉呈现为高对比度；图标采用方形图标，直线边长为12毫米；屏幕放大程度可达200%，其显示内容不缺失，被放大后的屏幕内容应可执行相关的操作，操作后的屏幕应仍处于被放大状态直至放大功能被用户关闭^[6]。

(2) 读屏技术：结合Android系统自带的TALKBACK、iOS操作系统自带的VoiceOver软件，设置读屏功能，用户可以通过对屏幕的两次点击操作完成对屏幕内容的读取和指令的操作，使得用户在听到相应的语音提示信息后，能够做出准确的操作，这一技术指标的设定，对低视力的老年群体特别友好，尤其是视障人群，可以有效实现智能终端的无障碍操作。

(3) 触控交互：老年人由于手的运动机能和视力机能降低，对屏幕的操作往往出现触控时间过短或过长、重复性点击等情况，造成智能终端程序混乱。针对这一情况，“老友季”APP对相关触控参数进行了调整，以降低老年人对于屏幕操作的难度。具体设置为：长按交互操作定义为通过按压屏幕，以显示内容预览、操

作或关联菜单的操作;对屏幕的两次点击操作完成对屏幕内容的读取和语音信息提示;支持在设定的时间内忽略重复点击的功能,而仅识别第一次点击操作。

3.3 APP界面优化

“老友季”作为一款面向老年群体的APP,充分考虑老年人思维和能力的局限性,简化视觉设计,避免有效信息被掩盖,从主页到二级

页面、三级页面都有有效的引导和结束提示。色彩应用方面,基调以偏暖色为主,方块按钮以醒目的颜色为辅,增强视觉辨识度,便于色彩的层级区分,界面常用的生活模块图标,加上中间镂空,增强设计感,简单大方^[7]。在页面布局方面,图标主要采用4行2列的排列形式,使老年人能够直观清楚地了解常用功能的分布情况,清晰明了、美观整洁,从而进行良好的界面交互,提高操作成功率^[8]。

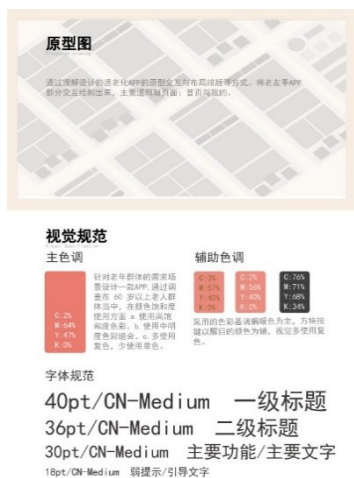


图5 “老友季”视觉规范



图6 “老友季”色彩优化



图7 “老友季”界面优化



图8 生活常用部分展示



图9 定位守护设计展示



图10 防诈骗指南设计宣传



图11 交通出行设计宣传

4 结语

近年来,老年人如何拥抱数字化问题持续引发社会关注,他们经历了国内科技发展跨度最大的年代,为当今的便捷社会、智慧社会的发展做出了巨大贡献,踏入新的时代,老年人一样可以享受到数字技术带来的便捷生活。适老化 APP 的设计研发是对老年群体生理和心理各层级需求考量的结果,设计者应做好调研分析,充分了解老年群体的生理特征和心理需求,设计出适合老年人操作,满足老年人生活并能守护老年人健康安全的产品,帮助老年人快速融入智能生活,成为数字智能技术的受益者。

参考文献:

- [1] 中国政府网. 关于印发《“十四五”积极应对人口老龄化工程和托育建设实施方案》的通知[A/OL]. (2021-06-17) [2023-11-27]. https://www.gov.cn/zhengce/Zhengceku/20-21-06/25/content_5620868.htm.
- [2] 工业和信息化部,中国残联. 关于推进信息无障碍的指导意见[A/OL]. (2020-09-11) [2023-11-27]. https://www.miit.gov.cn/jgsj/xgj/scgl/art/202/art_eca74f5b0c4a493e9e466195bd71ec74.html.
- [3] 电信终端产业协会. 智能终端适老化研究报告[R/OL]. (2021-06-17) [2023-11-27]. <http://www.taf.org.cn/upload/notice/2022-0919-135816-8656067.pdf>.
- [4] 国务院办公厅. 印发关于《切实解决老年人运用智能技术困难实施方案》的通知.(2020-11-24)[2023-11-27]. https://www.gov.cn/zhengce/zhengceku/2020-11/24/content_5563804.htm.
- [5] 王悦. 基于老年认知特性的界面交互设计研究[D]. 南京:南京理工大学,2018.
- [6] 万昱君. 基于用户体验的老年大学 APP 设计与研究[D]. 武汉:武汉理工大学,2020.
- [7] 刘丽,边卓. 色彩在 APP 界面设计中的应用研究[J]. 工业设计,2021(1):97-98.
- [8] 李佳玉,司峥鸣. 智慧产品 APP 的适老化设计策略[J]. 设计,2021,34(17):8-9.

Research on age-friendly design of intelligent terminal APP

Cui Xiaoxu¹, Xu Wenyan^{2*}

(1. Department of Information Technology, Tianjin Electronic Information Technician College, Tianjin 300350, China;

2. Department of Business Service, Tianjin Electronic Information Technician College, Tianjin 300350, China)

Abstract: With the great development of China's communication technology and communication industry, intelligent terminal products have been rapidly popularized in China, but because some elderly people with the growth of age, vision, hearing, movement, language, perception and other abilities decline to varying degrees, in the use of intelligent terminal APP process faced with a variety of difficulties and problems. Unable to fully enjoy the convenience brought by intelligent services of mobile terminals, the problem of "digital divide" has become increasingly prominent, therefore, APP aging design and development is an inevitable trend of the development of The Times. Through analysis and research, starting from the physiological characteristics and psychological needs of the elderly group, this paper explores the design strategy of the age-appropriate APP, and completes the design practice to help the design and transformation of the age-appropriate APP in our country.

Keywords: intelligent terminal; age-friendly; APP; design practice

文章编号: 1007-1423(2024)07-0101-05

DOI: 10.3969/j.issn.1007-1423.2024.07.018

非遗印记 APP 的 UI 分析与设计

吴 冰, 张晶晶, 程 妮*

(运城学院数学与信息技术学院, 运城 044000)

摘要: 为了更好地弘扬非物质文化遗产, 在分析用户需求的基础上, 使用 Photoshop、Illustrator 和 Visio 等软件开发了一款名为“非遗印记”的 APP, 并对 APP 的 UI 界面进行了设计和研究。非遗印记 APP 作为一种方便、高效、互动性强的移动程序, 为用户提供了一个便捷、全面了解非物质文化遗产的信息平台, 便于用户了解和体验传统文化, 从而加深对非遗文化的认识和理解, 促进非遗的保护、传承和推广。

关键词: 非物质文化遗产; APP; UI 设计; 传统文化; 传承

0 引言

非遗是人类社会在长期发展过程中逐渐形成的具有历史、文化、艺术、风俗等多种内涵的非物质文化遗产。作为一种独特的文化, 非遗的文化价值不可替代, 历史意义重大。因受现代化、城镇化等的冲击, 非物质文化遗产面临失传、衰退、遗忘等风险, 这给非遗保护工作带来了严峻的挑战。为了更好地保护和弘扬非遗文化, 笔者设计、开发了一款名为“非遗印记”的 APP, 并对用户界面(user interface, UI)进行了详细设计, 很好地呈现和传播了非遗文化, 旨在提高用户的非遗保护意识, 传承和发展非遗文化。

1 概述

作为中华优秀传统文化的重要组成部分, 非遗的传承和弘扬是延续历史文脉、推动文明交流的体现, 也是坚定文化自信的重要途径。伴随着新媒体技术和数字技术的发展, APP 成为人们日常了解外界的常用载体, 同时也是有效传播非遗文化的载体之一。因此, 借助 APP 推动非遗

的保护与传承, 从而促进文化交流与互鉴, 可为建设社会主义文化强国做出重要贡献。

1.1 非遗类移动应用的 UI 设计现状

UI 在非物质文化遗产应用程序中扮演着重要的角色, 它应该直观、易于使用, 并且能够有效地传达文化遗产的价值和信息。现阶段市面上虽存在一些非遗文化类 APP 应用(如四川省非遗 APP “智慧非遗”、湖北省非遗 APP “武汉非遗”等), 但很少有介绍山西非遗文化的 APP, 同时也鲜有研究者结合非物质文化遗产和 UI 界面设计来开发 Android 移动应用程序^[1]。在为数不多的以“非物质文化遗产”为主题的同类型应用中, 部分应用程序存在一些问题, 如布局不美观、导航混乱、ICON 图标与实际功能不匹配等, 这些都给用户造成了较差的使用体验。可见, 做好 APP 的界面设计至关重要, 同时开发一款方便实用、功能强大、内容丰富、交互性强的山西省非遗 APP 很有必要。

1.2 以用户思维为导向的 UI 界面设计

通过对 APP 的用户进行调研、访谈和测试,

收稿日期: 2023-10-25 修稿日期: 2023-11-24

基金项目: 2023 年山西省高等学校大学生创新创业训练计划项目(20230960); 2022 年山西省高等学校哲学社会科学项目(2022W133)

作者简介: 吴冰(2001—), 女, 山西运城人, 学生, 专业为数字媒体技术; 张晶晶(2003—), 女, 山西吕梁人, 学生, 专业为应用统计学; *通信作者: 程妮(1986—), 女, 山西运城人, 硕士研究生, 讲师, 研究方向为教育技术、计算机应用, E-mail: chengni_521@163.com

获得用户的反馈信息,开发者掌握了用户需求、偏好和行为模式,从而提供出色的用户体验。同时,在进行UI界面设计时,要保持界面元素的一致性,即颜色、图标和交互模式的统一,提升用户初次使用APP的舒适感和满意度^[2]。另外,APP导航界面的设计也要考虑用户的需求,清晰、直观的导航菜单和搜索功能,有助于用户快速浏览并准确定位需要的非遗内容^[3]。

2 相关理论基础

2.1 UI

UI即用户界面或人机界面,是用户与程序或设备互动过程的界面/接口。在计算机软件、应用程序或网站中,UI通常是指用户与软件或网站交互的各种元素和设计,例如按钮、菜单、图标、文本框、下拉列表、滑动条,等等。出色的UI也需要对用户进行需求分析,了解用户的特征和操作习惯等,从而为用户提供直观、易用和高效的界面^[4],增强用户的交互体验感。

2.2 用户体验

用户体验是UI设计的核心,指用户在使用产品或系统时的行为、情绪和态度等主观感受。UI设计时要遵循一些设计原则,如易用性、可访问性、反馈机制、导航结构等,还要考虑APP的社交功能,从而促进用户之间的交流与互动,提高用户对APP的使用体验和满意度。

2.3 APP界面视觉分析

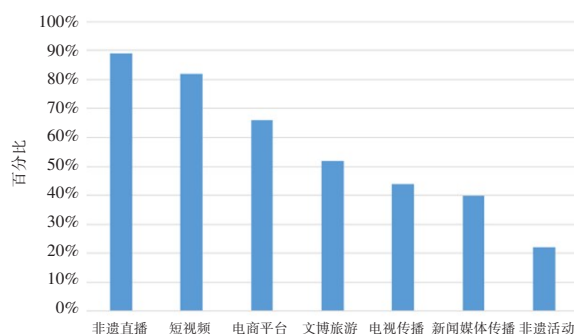
在非遗印记APP中使用了深色背景和明亮的文本颜色,这种颜色组合使得APP的界面更加醒目。此外,清晰的图标和直观的导航栏也会使UI界面简洁明了,同时使用一些线条、边框和背景颜色等在视觉上分隔不同的功能区域,这些设计都有助于用户分辨APP的功能,便于用户快捷找到需要的信息。

3 UI分析

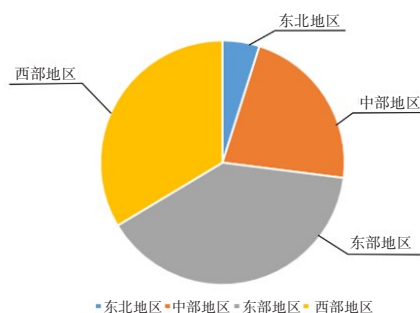
3.1 市场调研

在“互联网+”时代,随着新兴媒体的快速普及和广泛使用,非物质文化遗产呈现出依赖于新媒体进行多元化传播的趋势。APP的市场

调研数据来源于“易观数据”^[5],主要从非遗各渠道传播率和各地区非遗影响力两个方面进行分析(如图1所示)。市场调研结果表明,在非遗传播渠道方面,短视频媒体与直播媒体的兴起极大地促进了非遗在不同人群中的传播,深入到生活的方方面面。数字媒体为非遗文化的推广提供了新的途径,有助于年轻一代的参与、保护和传承。在非遗影响力方面,东部地区依托着非遗资源丰富和经济发达的优势位列非遗影响力的第一,同时中部和西部地区的非遗影响力也在日益增强^[5]。



(a) 各平台传播率



(b) 地区非遗影响力

图1 各渠道传播率及地区非遗影响力

3.2 交互场景分析

非遗印记APP的目标用户主要分为以下四类。

(1) 学生: APP可以为学生提供关于非物质文化遗产的学习资源,例如历史、文化和艺术等方面的知识和背景资料等。

(2) 文化爱好者: APP可以吸引对非遗文化和艺术感兴趣的爱好者,他们可以使用这个APP了解和探索不同的非物质文化遗产。

(3) 旅游者：当人们旅游时，可能会使用这个APP来了解当地的非物质文化遗产，例如传统节日、手工艺品和舞蹈等。

(4) 文化保护人士：这个APP可以吸引那些关注保护文化遗产的人，他们可以使用这个APP来了解不同的非物质文化遗产，并采取实际行动来保护这些遗产。

3.3 竞品分析

APP成为人们日常了解外界的常用载体，同时也是有效传播非遗文化的载体之一。现阶段市面上存在的非遗文化类APP尚存在一些不足，如用户体验不佳、登录界面使用繁琐、关于非遗的项目内容只进行了简单划分、项目分类单一化、UI设计未融入当地的民间艺术元素、ICON图标视觉上缺乏亮点、用户使用过程中代入感较弱等^[6]。非遗印记APP很好地弥补了以上不足，如设计了微信一键登录，避免了繁琐的登录步骤；同时对APP的功能模块进行了细致的划分；对ICON图标进行了细致的色彩划分；UI设计中融入了非遗特色等。

3.4 功能分析

通过对四类目标用户分析，目前非遗印记APP确定了如下几个功能。

(1) 非遗好物：对地方特色的非遗产品提供网上售卖平台，并且和匠人故事相结合，带动非遗地区的经济发展。

(2) 匠人故事：介绍非遗传承人和非物质文化遗产之间的故事，拓宽非遗产品的背景故事，保留非遗背后的故事。

(3) 我的收藏：将感兴趣的帖子以及非遗好物收藏起来，便于下次查找。

(4) 社区消息：与其他用户共同组成社区，便于用户之间的交流和分享。

(5) 非遗美食：将传统非遗文化与现代美食文化相结合，促进食品类非遗项目保护。

(6) 个人中心：查看自己的发帖记录，修改个人资料和系统设置等。

4 UI设计与实现

UI设计即用户界面的设计过程，目的是将

用户需求转化为具体的界面设计方案。在UI设计阶段，需要综合考虑用户体验、交互逻辑、界面美观等因素，以提高产品价值和用户体验。

4.1 功能结构图

通过市场调研和同类竞品分析，确定了非遗印记APP的基本功能：资源展示区、交互体验区、社交平台区、教育学习区、文化推广区和资讯推广区，而这六大主要功能又划分为14个子模块^[7]（如图2所示）。

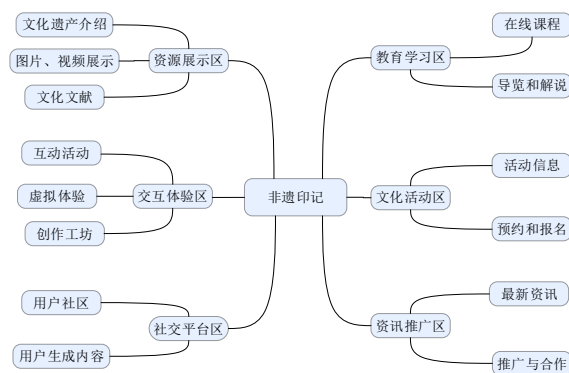


图2 APP功能结构图

4.2 APP界面设计

界面设计是开发应用程序的视觉和交互方面的设计，非遗印记APP的设计中保持了简单的设计风格，避免过于复杂的布局，将重要功能置于易于找到的位置；同时，选择清晰易读的字体，确保了可读性和易于理解^[8]。社区作为非遗印记APP的主要功能之一，为不同地区用户提供了很好的信息交流渠道，有利于应用积累沉淀用户，增加用户黏性。当然，对产品的反馈也可通过社区来进行，便于开发者及时了解用户的需求。APP的部分界面如图3所示。

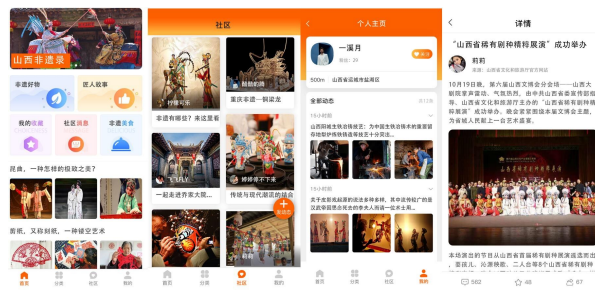


图3 非遗印记APP的部分界面

4.3 LOGO设计

作为APP的启动图标, LOGO要清晰地表达应用程序的目的和功能。非遗印记APP旨在保护和弘扬非物质文化遗产, 因此其LOGO的设计要与APP的主题保持一致。非遗印记APP的图标以中国非遗建筑为主体, 色彩选择了褐色, 字体采用了鸿雷板书简体, 背景设为纯白色, 整体突显了以非遗为核心的特色(如图4所示)。APP图标的布局参考了苹果应用图标的设计规范, 确保了应用图标的统一性和风格的一致性, 从而避免过多的细节和混乱的元素, 以使用户在应用商店中快速识别和区分该APP^[9]。总之, 简洁明了的应用图标、协调美观的手机桌面保证了在小尺寸的显示屏的可识别性, 极大地提升了用户的使用体验。



图4 非遗印记图标

4.4 ICON图标设计

ICON图标主要通过 Adobe Illustrator 进行设计, 采用了大量圆形图案和简洁且具有识别性的图形符号来代表功能、操作或内容的视觉元素。非遗印记APP的ICON图标简单生动, 界面协调美观, 给用户直观的导航和操作指引, 便于用户在界面上快速定位和浏览所需的内容, 使用户感到舒适和愉悦^[10]。非遗印记APP的部分ICON图标如图5所示。



图5 ICON图标效果

5 结语

本文设计了一款以保护和弘扬非物质文化遗产为主题的非遗印象APP, 将非物质文化遗产和移动应用相结合, 将非遗以数字的形式传播, 便于用户了解非遗文化, 加深对非遗文化的理解, 从而提高了非遗文化的保护意识, 促进了非遗的传承和推广。但在APP的UI设计过程中仍存在一些不足, 如APP功能不够完善、界面的过渡动画不够自然、社区功能比较单一等, 这些都是以后需要积极改进的地方。同时, 还应充分考虑非遗文化和UI之间的协调性, 促进非遗与数字的有机结合, 并以数字形式发扬非物质文化遗产, 使其焕发出新的生命力, 创作出更加美观协调的UI界面, 更好地保护和传承非物质文化遗产。

参考文献:

- [1] 马知遥, 刘智英, 刘鑫瑶. 中国非物质文化遗产保护理念的几个关键性问题[J]. 民俗研究, 2019, 6: 40-46.
- [2] 汪鸿, 陈锦渝. 以用户思维为设计导向的生鲜电商UI界面设计研究: 以食来优APP为例[J]. 湖北经济学院学报(人文社会科学版), 2021, 18(6): 66-68.
- [3] 周正, 司占军, 旷玲惠. 校园学付宝APP界面UI设计与应用研究[J]. 电脑知识与技术, 2020, 16(4): 54-55.
- [4] 李亭. 基于用户体验的智能手机APP界面设计研究[D]. 太原: 太原理工大学, 2015.
- [5] 文旅产业指数实验室. 2022非物质文化遗产消费创新报告[R]. 2022-10-27.
- [6] 曹小劲. 数字信息技术下非遗数字资源库建设研究[J]. 办公自动化, 2023, 11: 51-53.
- [7] 程妮, 刘宇腾. 基于Android的背单词APP的设计与开发[J]. 电脑与信息技术, 2022, 30(6): 30-33.
- [8] 程妮, 董晓雪, 张紫舒. 中国大学生计算机设计大赛的参赛创作过程与启示[J]. 电脑与信息技术, 2020, 28(4): 76-78.
- [9] 李倩. 基于用户体验的UI界面中图标设计研究[D]. 沈阳: 沈阳师范大学, 2018.
- [10] 郭威林. 宠物类APP平台UI设计对用户交互体验的影响[J]. 科技与创新, 2021, 3(15): 98-100.

(下转第108页)

文章编号: 1007-1423(2024)07-0105-04

DOI: 10.3969/j.issn.1007-1423.2024.07.019

基于 FreeSWITCH 的电话约车系统方案与应用

安文政^{1*}, 张 文²

(1. 中国电信股份有限公司日照分公司研发中心, 日照 276800; 2. 日照市大数据发展局, 日照 276800)

摘要: 随着移动互联网技术的发展, 智能交通出行系统日益普及, 老年人等不方便使用智能手机的人群面临约车难的问题。对这一难题进行了研究, 提出一种基于 FreeSWITCH 的电话约车方案。该方案通过对接 IMS/PSTN 实现全自动约车功能, 并由约车平台进行车辆调度和结果反馈。该方案操作简便、无需智能手机, 提高了出行效率和便利性。实际应用案例验证了系统的可行性和实用性, 为不方便使用智能手机的人群提供了便捷的出行选择。

关键词: 软交换; 电话约车; 无人值守; 订单处理

0 引言

近年来, 随着移动互联网技术的发展, 智能交通出行系统的建设逐渐得到推广。在这种背景下, 各种约车服务平台应运而生。这些平台通过手机应用程序提供服务, 为用户提供便捷的出行方式。然而, 对于部分不熟悉或无法操作智能手机的老年人等人群, 智能手机服务平台约车存在一定困难。

为了满足老年人等不方便使用智能手机的人群约车需求, 交通部从 2022 年开始在各地推广“95128”出租汽车约车服务电话。这项服务的推出旨在提供一种简单、便捷的约车方式, 让老年人和那些没有智能手机或无法熟练操作智能手机的人们也能够方便地约到出租车。

本文介绍的电话约车方案采用基于 FreeSWITCH(以下简称 FS)开源电话软交换平台研发的中间件(处理底层呼叫), 通过对接软交换/PSTN 电话, 实现全自动约车功能, 并将约车结果通过电话语音回复乘客。当然, 根据各地交通局需求, 也可通过软电话进行人工坐席接听派单。

1 系统架构

本文的电话约车方案基于软交换技术, 主要由以下几个部分构成。

(1) 电话及语音处理模块: 此模块基于 FS 接口开发, 用于自动接听呼入的电话, 处理(录制、压缩)和合成语音, 提供语音交互功能, 并通过自动语音识别(automatic speech recognition, ASR)对语音进行接口识别。

(2) 约车信息处理模块: 用于处理和解析识别出用户语音(文字)、关联地址信息, 获取相对准确的地理位置(经纬度)。

(3) 订单处理模块: 用于将用户号码、语音、解析出来的经纬度生成订单, 发送给周边司机。并将抢单车辆车牌号、司机电话转成语音回复乘客。

(4) 服务监控模块: 用于监控系统各模块是否正常运行, 对宕掉的服务进行重启及通知。

由于订单处理本身是个复杂的系统, 包括订单生成、车辆位置采集、订单就近推送、司机抢单、订单评价等多个过程, 鉴于篇幅, 本文不做详细介绍, 只对电话约车部分进行着重介绍。

收稿日期: 2023-11-07 修稿日期: 2023-11-23

作者简介: *通信作者: 安文政(1980—), 男, 山东日照人, 本科, 高级工程师, 系统分析师, 研究方向为基于通信能力的软件研发及应用, E-mail: anwenzheng.sd@chinatelecom.cn; 张文(1988—), 男, 山东日照人, 硕士研究生, 高级工程师, 研究方向为大数据应用

极大地降低了存储和带宽占用。

(7) 上传语音文件并获取语音 URL：将压缩后的语音文件上传至文件服务器，并生成一个唯一的语音 URL，以便司机听单时进行播放。建议使用对象存储(object storage service, OSS)，如阿里云、天翼云，对于业务量大的叫车系统，有良好的高并发兼容性。如果业务量不大，也可使用自建 FTP 服务器存储。

(8) 订单提交：语音上传成功后，可将乘客号码、经纬度信息(如有)、语音文件 URL 提交到订单处理模块，订单处理模块负责存储并发送订单。可使用消息队列遥测传输(message queuing telemetry transport, MQTT)方式^[4-5]，将订单消息推送到乘客周边的司机手机上。该过程涉及到司机实时位置采集、乘客所在地点与周边司机匹配等多个复杂环节。在实际使用过程中，“语义识别”过程对于方言可能做不到很好的识别，或产生地点无法匹配的语音，导致无法匹配经纬度就近发单，这类订单只能全网发送。

(9) 获取接单车号及司机号码：提交订单后，FS 侧等待订单处理模块返回司机抢单信息，包含抢单车号、司机电话等信息。一般设置 30 s 的抢单等待时间。如超时，则提示乘客：暂无车辆抢单；如有接单车辆，会有司机与您联系。

(10) 给乘客回复接单车号及司机号码：最后，系统会根据订单处理模块的返回结果，向乘客回复接单车号和司机号码等信息。由于 FS 本身不具有从文本到语音(text to speech, TTS)功能，需要将文字信息转换为语音进行播放。可以调用第三方平台的 TTS 功能直接整句转换，也可以对公用语句进行抽取，作为通用语音进行本地拼凑播放。比如，有司机抢单，给乘客回复：“您的订单已被车号：鲁 LT9999 抢单，司机电话 19900006666”，这句回复中，除了车号、电话号码外，其他信息是通用的，可以单独对“您的订单已被车号”“抢单，司机电话”这两部分录音，对车牌号的各省简称，如“鲁”“京”“苏”，以及车号里面的字母和数字都单独录音，再用算法进行合成即可在本地完美实现播报。

3 方案特点

特点 1：并发性强，适合大规模部署。采用多线程技术，能同时处理多部电话接入，并发

处理多个订单。经压力测试，单台 4 核 8 G 的 FS 服务器可实现 1000 路并发呼入。

特点 2：使用简易，适用范围广。只要会拨打电话就能叫车，固话、手机均可叫车。

特点 3：系统可全自动运行，采用预设的程序，完成语音提示-订单发送-接单回复工作。支持人工客服参与接听派单。还可对接第三方语音识别系统进行自动化 AI 地址识别精准发单。

4 方案实例

在交通部文件出台之前，山东日照多个区县(市区、五莲、岚山)在 2018 年已开始使用电话自动约车方案，当地出租车公司在电信的技术支持下，通过市场化运作，已在全国十几个地市推广，每日订单 6000~10000 余单，服务超过 60 余万乘客，订单突破 1000 余万单，其中叫车电话使用 600 余万次。做到了全程零人工值守，系统自动化运行。

本文介绍的语音文件预处理部分算法，已申请国家发明专利。

5 结语

本文通过对电话约车过程的详细研究，深入探讨了 FreeSWITCH 及软交换技术在其中的应用和实现方法。该系统的成功应用，可以提高约车服务的效率和质量，同时为乘客和司机提供更好的使用体验。随着移动互联网技术的不断发展，约车方案的功能和性能也在不断升级和完善。未来，我们将继续研究和改进该系统，以适应市场和用户的需求，推动智能交通系统的发展。

参考文献：

- [1] 杜金房,张令考. FreeSWITCH 权威指南[M]. 北京:机械工业出版社,2014:227-259.
- [2] 强磊,陈卉. IMS 核心原理与应用[M]. 北京:人民邮电出版社,2008:20-49.
- [3] 石凤贵. 基于 jieba 中文分词的中文文本语料预处理模块实现[J]. 电脑知识与技术,2020,16(14):248-251,257.
- [4] 付强. 物联网系统开发:从 0 到 1 构建 IoT 平台[M]. 北京:机械工业出版社,2020:62-74.
- [5] 姬广龙,孙丹,王珍珍,等. 关于 MQTT 通信协议的分析与研究[J]. 物联网技术,2023,13(1):63-66.

Telephone taxi booking system based on FreeSWITCH and its application

An Wenzheng^{1*}, Zhang Wen²

(1. China Telecom Rizhao Branch, Rizhao 276800, China;

2. Rizhao Big Data Development Bureau, Rizhao 276800, China)

Abstract: With the development of mobile internet technology, intelligent transportation systems are becoming more common, but there is still a problem of difficulty in booking taxis for people who are not proficient in using smartphones. This article studies this problem and proposes a telephone taxi booking system based on FreeSWITCH. This system can achieve automatic taxi calling by connecting IMS/PSTN, and the taxi platform can perform vehicle scheduling and result feedback. This system is simple to operate and does not require a smartphone, improving travel efficiency and convenience. Practical application cases have verified the feasibility and practicality of the system, providing a convenient travel option for people who are not proficient in using smartphones.

Keywords: soft switch; telephone taxi booking; unattended; order processing

(上接第 104 页)

UI analysis and design of APP on “Impression of Intangible Cultural Heritage”

Wu Bing, Zhang Jingjing, Cheng Ni^{*}

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

Abstract: In order to better carry forward the intangible cultural heritage, on the basis of analyzing the needs of users, the author developed an APP named “Impression of Intangible Cultural Heritage” by using several software such as Photoshop, Illustrator and Visio, and designed the UI interface of the APP. As a convenient, efficient and highly interactive mobile program, the “Impression of Intangible Cultural Heritage” APP provides users with a fast and comprehensive information platform for understanding intangible cultural heritage, facilitating users to understand and experience traditional culture, thereby deepening their understanding of intangible cultural heritage, and then advancing the protection, inheritance and promotion of intangible cultural heritage.

Keywords: intangible cultural heritage; APP; user interface design; traditional culture; inheritance

文章编号: 1007-1423(2024)07-0109-04

DOI: 10.3969/j.issn.1007-1423.2024.07.020

基于“互联网+”的智能车位管理系统的研究与设计

石亮翻¹, 朱贵平¹, 张凯淇¹, 毛 湘², 张 燕^{1*}

(1. 长沙医学院信息工程学院, 长沙 410219; 2. 湖南财政经济学院法学与公共管理学院, 长沙 410205)

摘要: 目前, 我国城市发展迅速, 城市建筑物复杂多样, 汽车数量越来越多, 城市用地资源紧张, 例如大多数城市大型商场未全方面规划泊车、人们出行找不到车位等诸多问题得不到改善。而相比传统的人工管理模式, 借助互联网+和线下人工管理结合的模式在此方面优势更加突出, 针对上述问题提供了一种解决方案, 并进行展开。

关键词: 车位系统; 车位管理; 预约车位; 互联网

0 引言

随着科技发展, 建筑越来越复杂多变, 交通越来越便利, 汽车数量超快增长, 市民日常停车车位的需求增大, 车位管理模式已从管理员手动记录车辆的进出时间和停放位置转变成通过安装传感器、摄像头等设备, 实现对车位的自动监测和管理。随着车辆数量日益增长, 大型商场地下停车场、大型收费停车场安装大量传感器成本过高, 后期设备的定期维护和更换需要大量成本, 此外用户出行不能明确知道车位的使用情况等, 这些直接影响用户出行体验, 因此本文设计“基于互联网+的智能车位管理系统”, 通过线上轻松预约车位便利用户出行。

1 技术背景

通过日常生活中遇到的问题得知需求, 通过需求得知应该应用哪方面的技术栈, 通过技术栈的实现来真正解决实际问题。项目的成功离不开技术的支持, 首先要做的就是对技术进行检索。实现智能车位管理系统在技术检索过程中, 本文鉴借文献[1-3]提出的基于Java语言搭建的前后端系统模块, 吸取了文献[4-5]使用

开源计算机视觉库 (OpenCV) 对图像视频的识别处理。对于系统的高效、完整, 需使用高效的数据结构和算法来存储和处理数据, 对于系统的安全需使用支付 API 和网络安全技术, 对于作用在服务器端的系统还需要实时操作系统、服务器技术和网络架构。对此本文重点介绍基于Java作为整体后端框架, OpenCV 识别车位自动统计模块。

2 系统架构

本系统由底层车位自动统计模块、中间层逻辑处理模块、上层功能表现模块构成。车位自动统计模块是实现系统的核心, 主要作用是信息处理, 在服务器响应时进行调用, 调用停车场摄像头获取实时视频, 通过获得的实时数据对车位进行识别处理并存储车位的具体信息。逻辑处理模块是连接车位自动统计模块、数据库、功能表现模块, 接受来自服务器端的功能表现模块的用户服务请求, 根据服务请求处理数据库中的车位信息或者管理用户个人信息, 将服务请求后的数据通过服务器响应传达给功能表现模块。功能表现模块为用户提供功能选

收稿日期: 2023-10-26 修稿日期: 2023-11-26

基金项目: 长沙医学院大学生创新创业训练计划项目(长医教[2022]41号-198)

作者简介: 石亮翻(2001—), 男, 湖南娄底人, 在读本科, 研究方向为 Web 应用程序开发、客户端开发; 朱贵平(2001—), 男, 湖南邵阳人, 在读本科, 研究方向为机器视觉和医学图像处理; 张凯淇(2003—), 女, 湖南岳阳人, 在读本科, 研究方向为数据库与大数据应用、算法设计与分析、算法歧视、生信分析; 毛湘(2002—), 女, 湖南娄底人, 在读本科, 研究方向为知识产权法、诉讼法、经济法; *通信作者: 张燕(1981—), 女, 湖南长沙人, 博士研究生, 副教授, 研究方向为生物信息学、疾病基因预测, E-mail: 2417860392@qq.com

择、数据接受窗口、数据渲染等，并调用 HTTP Web 服务传送数据给逻辑处理模块。

3 系统总体分析

3.1 实现方式

系统将部署在云端服务器上，用户通过功能表现层发送请求传达给服务器，服务器调用逻辑处理模块对数据库进行增、删、查、改操作，数据库将操作完成的结果反馈给逻辑处理模块进行逻辑处理，处理完成后将渲染展示在功能表现模块，完成用户的请求服务。车位自动统计模块是在服务器启动时自动开启，将得到车位数据信息传递给数据库，完成数据的实时数据收集与分析。具体实现方式如图 1 所示。

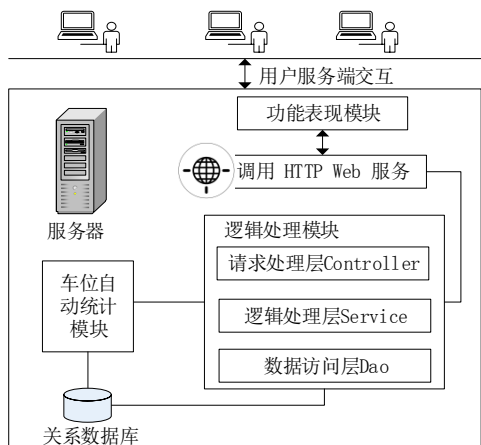


图 1 具体实现方式

3.2 技术设计

针对本车位管理系统设计，Web 前端开发使用的 HTML 超文本标记语言、CSS 层叠样式表、JavaScript 开发 Web 页面的脚本语言以及基于数据双向绑定的 MVVM 思想 Vue.js 构建的用户界面的渐进式 JavaScript 框架，简化 Web 页面开发。小程序端采用 .wxml、.wxss 文件负责小程序端的系统架构和渲染。后端采用 JavaSpring-Boot 主流框架结合 javaSE 面向对象、封装、面向接口编程、集合提供 RESTful API 接口，简化应用的开发和维护信息安全，连接数据方面采用 MyBatis 技术可以简化数据库操作，数据访问层使用 Java 数据库连接池连接数据库，完成逻辑处理模块与数据库的连接。车位自动统计模块采用开源发行的 OpenCV 对实时视频读取、图

像过滤、灰度处理、Canny 边缘检测、直线的检测、车位的划分等进行处理。基于 OpenCV 对车位识别的车位检测效果图如图 2 所示。



图 2 车位检测效果图

图像处理结束后建立训练神经网络模型切割停车位置图片、训练匹配模型并实时将数据处理后的结果传入数据库进行存储。如下为使用框架部分的代码展示。

```
public class ParkingDetection {
    public void ImageDel() {
        System.loadLibrary(Core.NATIVE_LIBRARY_NAME);
        //加载车辆特征分类器
        CascadeClassifier carCascade = new CascadeClassifier("haarcascade_car.xml");
        //haarcascade_car.xml 本地分类器文件路径
        //连接数据库
        try (Connection conn = DriverManager.getConnection("jdbc:mysql://localhost:3306/your_database", "username", "password")) {
            //创建数据库表
            String createTableQuery = "CREATE TABLE IF NOT EXISTS parking_tb (parking_id tinyint auto_increment primary key, parking_station tinyint not null, parking_createtime varchar(20) default Now(), parking_updatetime varchar(20) default Now(), x double not null, y double not null)";
            try (PreparedStatement createTableStmt = conn.prepareStatement(createTableQuery)) {
                createTableStmt.executeUpdate();
            }
            //调用摄像头
            VideoCapture capture = new VideoCapture(0);
            if (!capture.isOpened()) {
                System.out.println("未检测到实时视频!");
                return;
            }
            //实时处理视频帧
            Mat frame = new Mat();
            Mat grayFrame = new Mat();
```

```
while (true) {
    capture.read(frame);
    //将帧转为灰度图像
    Imgproc.cvtColor(frame, grayFrame,
        Imgproc.COLOR_BGR2GRAY);
    Imgproc.equalizeHist(grayFrame, gray-
        Frame);
    // 执行车辆识别
    MatOfRect cars = new MatOfRect();
    carCascade.detectMultiScale(grayFrame,
        cars);
    Rect[] carsArray = cars.toArray();
    // 将识别结果插入数据库,进行预编译
    处理,保证数据安全性
    String insertQuery = "INSERT INTO parking_
        results (x, y, width, height) VALUES
        (?, ?)";
    try (PreparedStatement insertStmt = conn.
        prepareStatement(insertQuery)) {
        for (Rect car : carsArray) {
            insertStmt.setInt(1, car.x);
            insertStmt.setInt(2, car.y);
            insertStmt.executeUpdate();
        }
    }
    //在图像上绘制识别到的车辆矩形框,
    并显示图像
    for (Rect car : carsArray) {
        Imgproc.rectangle(frame, car.tl(), car.br(),
            new Scalar(0, 255, 0), 2);
    }
    // 按下ESC 键退出循环
    if (capture.wait(1) == 27) {
        break;
    }
}
// 释放资源
capture.release();
System.out.println("数据传入数据库成功!");
} catch (SQLException e) {
    e.printStackTrace();
}
}
```

3.3 功能设计

智能车位管理系统主要有预定车位功能、用户定位功能、搜索附近的车库信息、订单交

充值功能、用户反馈功能、功能介绍等功能。功能展示图如图3所示。

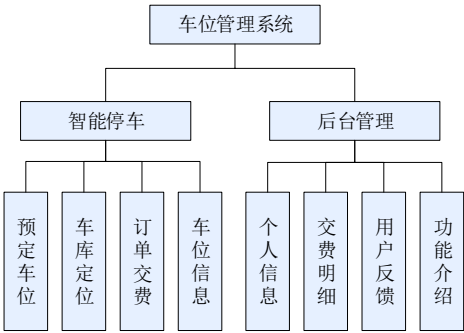


图3 功能展示图

3.4 数据库关系设计

根据业务需求，数据库的设计按照事务管理的原则，保证数据的一致性、数据的可靠性以及并发控制，避免数据冲突和死锁等问题。在数据库中，建立四张表来存储车位和用户信息，分别为“用户信息表”“车位信息表”“预订信息表”和“订单信息表”。其中，表1所示为“用户信息表”，表2所示为“车位信息表”，二者皆为基础表，而在“预定信息表”和“订单信息表”中分别设置预定车位外键、用户预定外键和订单车位外键、用户订单外键来连接“用户信息表”和“车位信息表”，同时也设置主键索引等提高数据的查询效率，防止数据量过多引起磁盘访问过慢的问题。

表1 用户信息表

属性	数据类型	约束	描述
user_id	tinyint	非空且唯一	用户编号
user_name	varchar	非空约束	用户姓名
user_gender	tinyint	非空约束	用户性别，0为女性，1为男性
user_telephone	varchar	非空约束	用户电话
user_createtime	varchar	非空约束	用户注册时间格式为：YYYY-MM-DD:TT: mm: ss
user_updatetime	varchar	非空约束	用户信息修改时间格式为：YYYY-MM-DD:TT: mm: ss

表 2 车位信息表

属性	数据类型	约束	描述
parking_id	tinyint	非空且唯一	车位编号
parking_station	tinyint	非空约束	车位状态, 0表示禁用, 1表示启用
parking_x	double	非空约束	车位长度
parking_y	double	非空约束	车位宽度
parking_createtime	varchar	非空约束	车位开始使用时间格式为: YYYY-MM-DD: TT: mm: ss
parking_updatetime	varchar	非空约束	车位使用结束时间格式为: YYYY-MM-DD: TT: mm: ss

4 结语

本文以 OpenCV 视觉库检测车位作为技术切入点的智能化车位管理系统, 通过系统实现找到车库、预约车位、停车交费等功能, 突破了传统的车位管理的模式, 高效管理车位资源等。本次研究与设计为智能车位管理系统的实现提

供了一种可行性的解决方案, 相信随着人工智能不断发展与优化, 智能车位管理系统将在实际应用中发挥更大的作用。因相关计算机技术不够成熟且经费不足, 本项目目前的研究与设计仍处于起步阶段, 项目开发过程中存在设计不完善、潜在安全风险以及技术风险。团队后期将努力检索最新文献、借鉴 AI 技术、钻研计算机技术, 后期对各个功能模块进行单元测试、集成测试、回归测试等多重测试, 深入了解用户需求, 进一步完善项目的功能。

参考文献:

[1] 宋俊芳, 麻文睿. 基于 Java 的校园停车智能管理系统设计[J]. 科技风, 2018(30):95.

[2] 苟坤炎, 韩利凯. 基于 Web 停车场车辆管理系统的设计与实现[J]. 电子世界, 2019(18):193-194.

[3] 谭明亮, 代玲林. 基于微信小程序的医院医疗设备管理系统设计与实现[J]. 价值工程, 2023, 42(3): 137-140.

[4] 付召斌, 王革. 基于 OpenCV 的校园车辆管理系统的设计与实现[J]. 通讯世界, 2020, 27(2):3-4.

[5] 沈哲. 基于视觉的校园停车位占用检测算法研究[D]. 吉林: 吉林大学, 2020.

Research and design of intelligent parking space management system based on “Internet +”

Shi Liangfan¹, Zhu Guiping¹, Zhang Kaiqi¹, Mao Xiang², Zhang Yan^{1*}

(1. College of Information Engineering, Changsha Medical University, Changsha 410219, China;

2. College of Law and Public Administration, Hunan University of Finance and Economics, Changsha 410205, China)

Abstract: At present, China’s cities are developing rapidly, urban buildings are complex and diverse, the number of cars are increasing, and the number of urban land resources are tight. For example, many problems such as large shopping malls in most cities have not planned parking in all aspects, and people can’t find parking spaces when travelling have not been improved. Compared with the traditional manual management model, the combination of Internet + and offline manual management model has more prominent advantages in this regard. This article will strive to develop a solution for the above problems.

Keywords: parking space system; parking space management; parking space reservation; Internet

文章编号: 1007-1423(2024)07-0113-04

DOI: 10.3969/j.issn.1007-1423.2024.07.021

基于在线句子分析的系统设计与实现

汪倩倩*

(华北水利水电大学信息工程学院, 郑州 450018)

摘要: 在分析句子结构的过程中, 为明确各短语的依赖关系和最终的分析树, 基于带有范畴转换的组合范畴语法 (CCG- C^2), 设计了一个在线句子分析的系统, 实现对句子的分词, 确定单词范畴, 使用范畴规则进行传递, 对产生的新短语进行歧义消解, 最后形成分析树并存储树库。该系统克服了句子分析的时间限制, 通过人机交互的方式可以实现句子分析过程中的难点即歧义消解, 大大提高了人工分析的效率。

关键词: 句子分析; 系统设计; 人机交互

0 引言

随着互联网大数据的迅速发展, 对计算机处理自然语言的需求日益增长。自然语言处理 (natural language processing, NLP) 包括自然语言的输入、表示、处理和输出等环节。内部表示可采用 ASCII 码、ANSI 字符集、Unicode 标准、GB2312 等编码方式。处理过程中, 可使用文档编辑软件 (如 Word、WPS 等) 进行加工处理。近年来, 随着人工智能的进步以及对机器快速处理大量自然语言的需求增加, 自然语言处理在问题回答、自动文摘以及机器翻译等方面面临着更高的要求。

自然语言处理是计算机和人工智能领域的一个重要研究领域, 致力于探索理论和方法, 使计算机能够像人类一样理解和处理自然语言。自然语言处理是一门跨学科研究领域, 涵盖语言学、计算机科学、数学、心理认知学以及神经科学等多个领域。该领域致力于研究人类语言, 通过形式化处理自然语言, 并利用各种算法实现对其进行处理。近年来, 自然语言处理的相关应用领域越来越广泛, 并取得了长足发展。这些应用包括机器翻译、舆情监测、自动摘要、观点提取、文本分类、问题回答、文本语义对比^[1]、语音识别以及中文光学字符识别

等。自然语言处理在这些领域的应用逐渐增多, 取得了显著的成果。NLP 在不同领域的应用, 不但带来了发展机遇, 也带来了巨大挑战, 因为人们对人工智能的需求不断增长。自然语言的语义分析不仅仅关注词法和句法层面的语法分析, 还与单词、词组、句子和段落所承载的意义相关联^[2]。该分析旨在获取句子的语义表示, 进而得到表达言语的语义表示结构^[3]。为了理解自然语言, 首先需要用计算机支持的语言来表达自然语言的语义, 特别是句子的语义表达, 因为句子是完整的语言语义表达单元。

句法分析在自然语言处理中扮演了关键角色, 它是支撑和影响后续处理任务的重要环节。基本目标是确定句子的内部结构。相比于欧美语言的屈折语言, 汉语属于分析语言系统, 所以不能直接套用欧美国家的语法理论。汉语语法研究的发展程度较低, 朱德熙先生在他的著作《语法答问》中曾探讨汉语句法分析的重要性和困难之处: ①汉语缺乏固定的形态; ②语序较为灵活; ③句法成分和词类之间缺乏一对一的对应关系; ④汉语句子和词组的构造原则基本相同; ⑤汉语语法形式化的研究滞后。

该研究开发了一个涵盖范畴表示、规则表示、句法分析、输入输出等模块的分析软件分析树库的机器构建系统, 旨在验证和指导中文

收稿日期: 2023-11-12 修稿日期: 2023-12-18

作者简介: *通信作者: 汪倩倩 (1997—), 女, 河南周口人, 硕士, 研究方向为自然语言处理, E-mail: 2374269533@qq.com

组合范畴语法的设计。同时,该系统能够评估分析器^[4]的性能,并成为未来构建分析树库的重要基础。本文设计的汉语句子分析器支持歧义消解^[5],能够支持自然语言处理中的问题-回答、自动文摘^[6]等多项任务,并对其性能进行改进。

1 Haskell 程序设计语言

为了验证组合范畴语法在汉语句法分析中的合理性和有效性,需要增加支持转类套叠的规则。为了让计算机完成句法分析,需要选择一种程序语言来实现这种新语法的分析器。范畴结合在组合范畴语法中实际上是函数运算,在实现该语法的分析器时使用函数式程序语言会更自然。Haskell 是一种这样的语言。

Haskell 是一种常见的、普遍适用的纯函数式编程语言,得名于美国逻辑学家 Haskell Brooks Curry。他在数理逻辑领域的贡献为函数式编程语言的发展提供了广泛基础。作为一门函数式编程语言,Haskell 的主要控制结构是函数。函数式编程具有以下基本特点:允许将函数作为值进行传递,并在定义函数时从外部获取值并放入函数对象中(通过 lambda 表达式实现)。对于函数对象,只能调用它,而不能直接访问其内部数据。Haskell 语言具有“证明即程序、命题即类型”的特征。Haskell 的语言特性包括纯函数式、不可变值、支持高阶函数和类型推断。

软件采用编程语言 Haskell,它简洁明了,主要有以下几个优点:

(1) 在组合范畴语法的句法分析中,我们成功实现了 74 种不同的句法类型转换。为了最大程度地降低结构歧义,我们通过引入范畴转换的方式,在句法分析过程中根据开关(OnOff)的值灵活选择所需的范畴转换。

(2) 实现了对树库和歧义片段库的统计和搜索功能。

(3) 完成了基于人机交互的命令层次结构的实现。在人机交互的过程中进行传递剪枝,模拟了人脑按需分析句子结构的过程,相较于机器分析保留所有句法歧义片段并生成很大的分析树森林,具备更好的可计算性。

2 句子分析系统的体系结构

为了研究基于组合范畴语法的中文歧义消解,我们设计了一个中文句法分析器。从“平衡语料库”中获取了词类标注样本,并为句中的词语标注了相应的典型句法成分(包括句法类型和语义)。通过支持范畴转换的组合范畴语法分析范畴序列,构建了分析树,并提取了句法结构基因,这些基因将被更新或添加到结果基因库中。通过深度学习等技术,我们在分析树库和结构基因库上对分析器进行了改进。创建了一个实验平台,如图 1 所示。其中,“范畴转换的组合范畴语法”模块使用了带有范畴转换的组合范畴语法 CCG-C^{2[7]},而“分析器”则用于实现人机交互的具体分析过程。

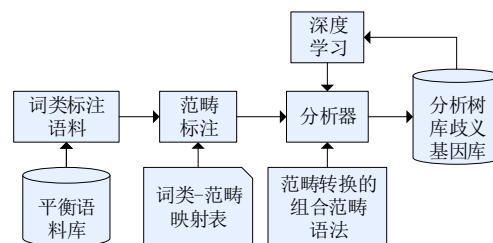


图 1 实验平台

分析软件的总体结构由范畴(Category)、规则(Rule)、短语(Phrase)、语料(Corpus)、分析(Parse)、输出(Output)、工具(Utills)、句子分析(SentParse)、统计模块(Statistics)、主入口模块(Main) 十大模块结构组成,模块间依赖关系如图 2 所示。

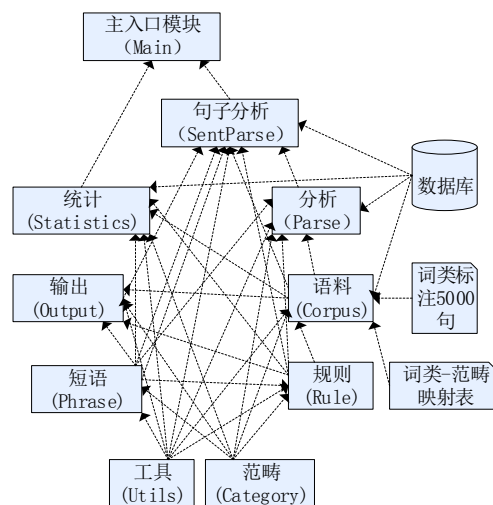


图 2 模块间依赖关系图

这些模块的主要功能如下：

范畴(Category)：定义范畴类型，以及作用于范畴的一些函数。主要是适合汉语的 CCG-C² 类型，及针对汉语范畴转换的情况给出其转换规则，使之更好地用范畴表示汉语类型。

工具(Utils)：定义一些有关元组、列表、字符串的工具函数。

短语(Phrase)：定义短语类型和一些应用于短语的函数。包括短语和传递的定义，以及句子生成森林的算法。

规则(Rule)：定义规则类型和一些应用于规则的函数。

语料(Corpus)：定义结构，一些有关基因类型、连接数据库、初始化再生语料的函数、词法规则列表、句法规则列表、分析树列表与相应范畴进行互相转换的一些函数。此模块叙述相关语料库的构建工作，并在此语料库上进行相应的分析、统计等工作。

分析(Parse)：定义范畴转换规则，以及范畴结合传递、剪枝相关的一些函数。

输出(Output)：定义一些标准输入输出函数和测试函数，并把数据类型值包括句法树输出到标准输出设备上。

句子分析(SentParse)：软件的中心任务是定义从 MySQL 数据库读取有词类标注句子、小句分隔、用户输入从某个小句开始分析、人机交互下完成范畴结合的句子所有传递闭包计算、小句分析结果存入到树库中、显示某个句子的分析树和分析森林等函数。把之前所叙述的范畴转换规则、传递闭包算法、森林算法、数据库构建等结合在一起之后对语料进行系统的分析，从而得出具体的分析树，进而形成一个树库。

主入口模块(Main)：实现命令层次，如分析哪一个语料库、生语料生成带范畴的语料、分析句子等命令。

统计模块(Statistics)：实现数据库表的数据统计和搜索。

统计分析有两类功能统计和搜索。对分析树的统计包括，统计使用的范畴转换频次以及使用的句法规则、统计句子个数、小句长度等；对分析树的搜索包括使用 CCG-C² 标记，以及对分

析树库 Tree 和结构基因库 Stru-gene 的搜索功能。

3 系统使用方法

所谓命令式交互层次是系统提供给用户的所有命令的分层次组织，从顶级到第二级再到第三级……每一级有一个命令列表，命令 A 完成统计分析。第一级菜单如下：“1”显示命令列表；“2”获得给定序号的词类标记句子；“3”复制该词类标记句子到 raw-sent2 字段；“4”获得该修订后的词类标记句子；“5”从修订后的词类标记产生 CCG 标记句子获得给定序号的 CCG 标记句子；“6”复制给定序号的 CCG 标记句子到 cate sent2 字段；“7”获得给定序号的修订后的 CCG 标记的句子；“8”分析给定序号的句子；“9”显示给定序号的句子的分析树；“A”对数据库中的 corpus 表和 Stru-gene 表进行统计分析；“0”退出。

假设选择“9”，那么就会出现给定句子的分析树，如图3所示。

```

Clause No.: 1
--Desig --Desig --Desig --Desig --Desig --Desig
((s\.np)/#(s\.np))/*np (s\.np)/.np np/.np np s\.np np np\*np
:在' (中' (开展' ((文明' 班级') 评比' 活动' :中'
:
np
:文明' 班级'
-----A/n-Hn/v->
np
:(文明' 班级') 评比'
-----A/n->
np
:((文明' 班级') 评比') 活动'
----->
s\.np
:开展' (((文明' 班级') 评比') 活动')
-----Hn/v-<
np
:中' (开展' (((文明' 班级') 评比') 活动'))
----->
(s\.np)/#(s\.np)
:在' (中' (开展' (((文明' 班级') 评比') 活动'))

Clause No.: 2
--Desig --Desig --Desig --Desig --Desig --Desig
np/.np np np/.np (s\.np)/.np np np s\.np np
:许多' 学校' :紧密' :配合' :中学生' :守则' :教育'
:
np
(s\.np)/.np
:许多' 学校' :紧密' 配合'
-----D/a->B
np
:中学生' 守则'
-----A/n-Hn/v->
np
:((中学生' 守则') 教育'
----->
s\.np
:(紧密' 配合') ((中学生' 守则') 教育')
-----<
s
:((紧密' 配合') ((中学生' 守则') 教育')) (许多' 学校')

```

图3 软件一级命令“9”结果图

根据图3句子分析结果可以看出，软件支持中文句法结构的转类套叠。对于是否能够解释汉语中的所有转类套叠还需要进行进一步验证。接下来，我们将改进分析树库的构建软件，现已将教育部语言文字应用研究所发布的平衡语料库(<http://corpus.zhonghuayuwen.org/>)的词类标注语料与 CCG-C² 进行分析，进而可以形成成分

析树库。通过使用这个分析树库,可以指导改进组合范畴语法,以便解释中文的普遍表达方式,同时还可以评估分析器的性能,指导改进分析器。另一方面,进一步深入研究汉语语法,从汉语和印欧语言的区别入手,探索改进组合范畴语法的方法,发现 CCG-C² 在汉语机器处理领域的应用价值。

4 结语

本文使用 Haskell 语言对 CCG-C² 进行了实现。这个软件由十个模块组成,包括范畴、规则、短语、语料、分析、输出、工具、句子分析、统计和主入口。它可以通过输入范畴序列来生成分析树森林。该软件可作为基础,进一步开发构建组合范畴语法分析树库的软件。该软件能够对输入的范畴序列进行处理,输出包含符合中文句法的分析树森林。使用该软件,用户不仅可以分析简单的小句,还可以对复杂句子进行人工交互的歧义消解。以该软件为基

础,可以进一步设计脱离人工消歧的机器分析软件。

参考文献:

- [1] 李强. 汉语句子语义成分分析系统[J]. 河南财经学院学报, 1990(2): 78-81.
- [2] 左静妮. 系统功能语法视角下语篇分析在英语阅读中的应用研究[J]. 牡丹江教育学院学报, 2023(2): 80-84.
- [3] 张继培. 句子分析在阅读理解中的影响[J]. 济南职业学院学报, 2010(5): 85-87.
- [4] 吴安迪. 左角句子分析器与中心语驱动句子分析器[J]. 国外语言学, 1993(2): 7-15.
- [5] 张觉. 从歧义分析看句子分析法[J]. 贵阳师专学报(社会科学版), 1995(3): 65-70.
- [6] 高东辉, 路伟. 基于自动文摘的地震灾情信息分析处理系统研究与实现[J]. 化工中间体, 2015, 11(12): 76-78.
- [7] 王庆江, 陈淑娴. 一个适合汉语的带有范畴转换的组合范畴语法[C]//第二十一届中国计算语言学大会(CCL2022), 南昌, 2022.

Design and implementation of a system based on online sentence analysis

Wang Qianqian*

(Department of Information Engineering, North China University of Water Resources and Electric Power,
Zhengzhou 450018, China)

Abstract: In the process of analyzing sentence structure, in order to clarify the dependency relationship of each phrase and the final analysis tree, an online sentence analysis system was designed based on the combined category grammar with category transformation (CCG-C²). It achieved sentence segmentation, determined word categories, used category rules for transmission, resolved ambiguity in the generated new phrases, and finally formed an analysis tree and stored it in a tree library. This system overcomes the time limit of sentence analysis and can achieve ambiguity resolution, a difficult point in the process of sentence analysis, through human-computer interaction. This system greatly improves the efficiency of manual analysis.

Keywords: sentence analysis; system design; human-computer interaction

文章编号: 1007-1423(2024)07-0117-04

DOI: 10.3969/j.issn.1007-1423.2024.07.022

基于SpringBoot的网上商城管理系统设计与实现

王培培*

(河南工业职业技术学院电子信息工程学院, 南阳 473000)

摘要: 根据网上商城管理系统的实际需求, 利用 SpringBoot、MyBatis、MySQL、BootStrap 等开发技术, 设计并开发出网上商城管理系统。系统可以有效地管理商品、订单生成、商品地址等相关流程, 提高商家运营的效率并节省了运营成本。

关键词: 网上商城; SpringBoot; MySQL

0 引言

在传统的商业模式中, 商家需要投入大量的人力、物力和财力来维护商品库存、处理订单、管理用户数据等, 然而, 随着电子商务的兴起, 商家面临着更为复杂的问题和挑战, 比如快速更新商品信息、处理大量订单、分析用户行为等, 这些问题单靠传统的方式已经难以解决, 在此情况下, 网上商城管理系统应运而生。线上购物逐渐赶上甚至超过了线下购物, 成为国内消费者的主要消费方式^[1]。网上商城管理系统不仅可以提高商家的管理效率, 降低运营成本, 减少人力资源的浪费。同时, 也可以提升用户体验、增加销售机会, 为网上商城带来更长久的发展和成功。

1 相关概述

本系统采用了 Linux+Tomcat+MySQL+Spring-Boot 的组合, 使用 BootStrap 来设计前端页面, 并实现了页面交互式设计。这种组合不仅提高了系统的稳定性和安全性, 还简化了开发流程, 提高了开发效率。同时, BootStrap 的运用也使得页面更加美观、友好, 用户体验更加流畅。

1.1 SpringBoot

SpringBoot 是当前流行的 Java 后端开发框架, 可以有效整合开发过程中使用的技术, 提供默认配置信息, 简化其他框架的配置过程^[2]。Spring Boot 是基于 Spring 的框架, 能够实现对应用软件的自动化, 无需手动设置。Spring Boot 集成了几乎所有的框架, 能快速搭建 Spring 项目。

1.2 MySQL 数据库

MySQL 是一种关系型数据库管理系统 (RDBMS)^[3], 它是许多网站和应用程序的基础组成部分。MySQL 是一个快速、多线程、多用户的 SQL 数据库服务器, 其重要性在于它提供了一种有效的方式来存储、管理和检索数据, 以满足网站和应用程序的需求。

SpringBoot+MySQL 的组合具有开发效率高、数据库安全性高、可靠性高、高度可定制性、支持多语言、维护成本低等优点, 是一种非常优秀的开发框架组合。在系统启动后, SpringBoot 自动与数据库相连接, 减少手动操作失误^[4]。

2 系统总体设计

经过前期的调研和分析, 本系统实现了商

收稿日期: 2023-10-19 修稿日期: 2023-11-22

作者简介: *通信作者: 王培培 (1995—), 男, 河南南阳人, 助教, 本科, 从事领域为软件工程和网络安全, E-mail: 980684377@qq.com

品管理、订单管理、会员管理、常规管理等四部分的功能。商品管理模块包括添加分类、分类列表、添加商品、管理商品和商品搜索功能子模块，方便商家管理库存。订单管理模块则实现了订单管理和评价管理两个功能子模块，便于商家处理订单和了解客户反馈。会员管理模块支持会员管理和会员查询等功能子模块，帮助商家维护用户数据。常规设置模块则包括留言管理和管理员密码两个功能子模块，确保商城的稳定运行。这些功能的实现，可以提高商家的管理效率，降低运营成本，提升用户体验，增加销售机会。网上商城管理系统如图 1 所示。

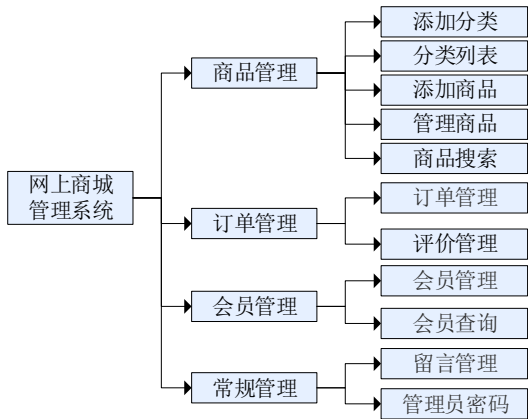


图 1 网上商城管理系统

3 页面设计

网上商城系统分为商品管理、订单管理、会员管理和常规管理四部分，方便管理员对商品、订单和会员进行统一管理。其中，index.jhtml 是系统的首页，管理员可以通过该页面查看商品页面、订单页面和会员页面，并进行相应的操作。本系统共包含 15 个类似的页面，每个页面及其对应的功能见表 1。

管理员成功登录系统后，可以进入系统主页面，方便浏览商品、订单、会员信息以及留言信息。同时，管理员具有增删改查等权限，可以进行各项操作，有效地管理商城。系统的功能子模块和功能权限灵活配置，让管理员可以根据实际需要进行自定义。严谨的权限管理

功能，确保了数据的安全性。

表 1 页面及相应功能

页面名称	功能
index.jhtml	主页
comdy.jhtml	商品管理页面
comdy_sort_add.jhtml	添加分类页面
comdy_sort_list.jhtml	分类列表页面
comdy_add.jhtml	添加商品页面
comdy_search.jhtml	商品搜索页面
order.jhtml	订单管理页面
evaluation.jhtml	评价管理页面
member.jhtml	会员管理页面
member_search.jhtml	会员查询页面
routine.jhtml	常规管理页面
message.jhtml	留言管理页面
ad_passwd.jhtml	管理员密码设置页面
login.jhtml	登录页面
register.jhtml	注册页面

4 数据库设计

在系统开发过程中，数据库的主要功能有存储和管理数据，在做数据库相关表的设计时，要考虑到它长期使用的情况，尽量做到“以不变应万变”，即当用户需求发生变化时，在保证数据库整体设计不变的情况下，也可以适应用户的功能需求^[5]。

根据前期和用户的对接以及实际的调研分析，设计出以下的数据库表来存储网上商城管理系统中所使用的数据。

(1) 商品表(goods)：存储商品的基本信息，包括商品 ID、内容、分类、价格等。

(2) 订单表(orders)：存储订单的基本信息，包括订单 ID、商品 ID、购买价格、订单号等。

(3) 会员表(members)：存储会员的基本信息，包括会员 ID、用户名、密码等。

(4) 留言表(comments)：存储留言的基本信息，包括留言 ID、会员 ID、留言内容等。



图4 订单管理界面

5.4 常规管理功能

常规设置模块包含留言管理和管理员密码两个功能子模块。密码界面采用可视化操作,管理员可以随时修改密码,确保账号的安全。留言管理支持商家快速回复买家的留言信息,让买家感受到更优质的服务。这些功能子模块可以帮助商家更好地管理商品库存、处理订单和掌握用户需求等,进而提高店铺的管理效率,增加商品的销售机会。留言管理界面如图5所示。

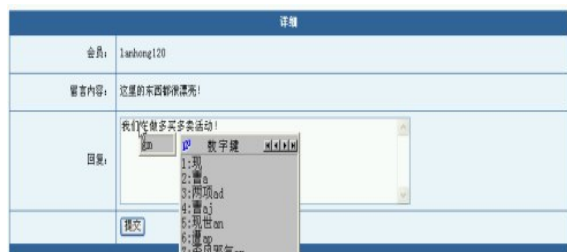


图5 留言管理界面

6 展望

当前系统的功能,例如商品、订单以及留言等,都需要进一步完善和改进,以保证能够满足不同用户的需求。同时,界面设计也需要进行改进,以提供更友好、美观、便捷的用户体验。系统的可操作性和易用性也需要进一步提高,以方便商家快速上手并高效地进行商品

管理和订单处理。此外,系统的稳定性和安全性也需要更加重视,以确保数据的安全和系统的稳定性。总之,未来还需要不断地改进和完善系统功能和界面设计,以适应电子商务的不断发展和变化。

7 结语

本系统从用户的实际需求出发,既实现了对商品信息的增删改查,又实现了订单信息的便捷查询和删除。此外,系统还支持商品搜索功能,能够帮助商家更好地管理商品信息。同时,管理员还可以通过系统快速查找到用户订单信息,快速帮助用户解决问题。系统的可操作性和易用性也得到了极大的提高,方便商家高效地进行商品信息、订单信息、留言信息的管理。总之,本系统在开发过程中不断与用户进行沟通和交流,从而实现了系统后台管理员高效地处理商品、订单、会员以及留言等信息,为商家提供了更加完善的后台管理解决方案。

参考文献:

- [1] 郝磊. 网络商城微信小程序的设计与实现[D]. 兰州:西北民族大学,2020.
- [2] 王亚东,李晓霞,陈强强,等. 基于SpringBoot的需求发布平台设计[J]. 信息与电脑(理论版),2023,35(1):105-107.
- [3] 田娟. 基于PHP+MySQL员工信息管理系统后台设计与实现[J]. 电脑知识与技术,2023,19(23):47-49.
- [4] 赵晓琴,李振江,赖炎峰,等. 基于Spring Boot的分布式微服务审计管理系统设计[J]. 自动化技术与应用,2023,42(1):182-186.
- [5] 桑冉航,李晓明. 基于Spring Boot的健身房管理系统的设计与实现[J]. 电脑知识与技术,2023,19(22):54-56.

Design and implementation of an online mall management system based on SpringBoot

Wang Peipei*

(School of Electronic Information Engineering, Henan Vocational and Technical College of Industry, Nanyang 473000, China)

Abstract: Based on the actual needs of the online shopping mall management system, this article designs and develops an online shopping mall management system using development technologies such as SpringBoot, MyBatis, MuSQL, and Bootstrap. The system can effectively manage processes related to products, order generation, and product addresses, improving the efficiency of merchant operations and saving operational costs.

Keywords: online shopping mall; SpringBoot; MySQL

