

硕士学位论文

数据驱动的互联网卡用户离网预测及干预
算法研究

□□□□

Central South University

一 级 学 科	计算机科学与技术
---------	----------

二 级 学 科	计算机应用技术
---------	---------

作 者 姓 名	钱凯
---------	----

指 导 教 师	吕丰 教授
---------	-------

2022 年 4 月

中图分类号 TP391

学校代码 10533

UDC 004.9

学位类别 学术学位

硕士学位论文

数据驱动的互联网卡用户离网预测及干预 算法研究

□□□□

Central South University

作 者 姓 名	钱凯
一 级 学 科	计算机科学与技术
二 级 学 科	计算机应用技术
研 究 方 向	深度学习
二级培养单位	计算机学院
指 导 教 师	吕丰 教授
副 指 导 教 师	

论文答辩日期_____ 答辩委员会主席_____

中 南 大 学

2022 年 4 月

学位论文原创性声明

本人郑重声明，所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了论文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中南大学或其他教育机构的学位或证书而使用过的材料。与我共同工作的同志对本研究所作的贡献均已在论文中作了明确的说明。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

学位论文作者签名：_____ 日期：_____ 年 ____ 月 ____ 日

学位论文版权使用授权书

本学位论文作者和指导教师完全了解中南大学有关保留、使用学位论文的规定：即学校有权保留并向国家有关部门或机构送交学位论文的复印件和电子版，允许本学位论文被查阅和借阅。本人授权中南大学可以将本学位论文的全部或部分内容编入有关数据库进行检索和公开传播，可以采用复印、缩印或其它手段保存和汇编学位论文。本人同意按《中国优秀博硕士学位论文全文数据库出版章程》规定享受相关权益。本人保证：毕业后以学位论文内容发表的论文作者单位注明中南大学；学位论文电子文档的内容和纸质学位论文的内容相一致。

延缓公开论文延缓到期后适用本授权书，涉密论文在解密后适用本授权书。

本学位论文属于：(请在以下相应方框内打“√”)

☐ 公开

☐ 延缓公开，延缓期限（____ 年 ____ 月 ____ 日至 ____ 年 ____ 月 ____ 日）

学位论文作者签名：_____

指导教师签名：_____

日期：_____ 年 ____ 月 ____ 日

日期：_____ 年 ____ 月 ____ 日

(填写阿拉伯数字)

数据驱动的互联网卡用户离网预测及干预算法研究

摘要： LaTeX 利用设置好的模板，可以编译为格式统一的 pdf。目前国内大多出版社与高校仍在使用 word，word 由于其强大的功能与灵活性，在新手面对形式固定的论文时，排版、编号、参考文献等简单事务反而会带来很多困难与麻烦，对于一些需要通篇修改的问题，要想达到 LaTeX 的效率，对 word 使用者来说需要具有较高的技能水平。

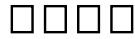
为了能把主要精力放在论文撰写上，许多国际期刊和高校都支持 LaTeX 的撰写与提交，新手不需要关心格式问题，只需要按部就班的使用少数符号标签，即可得到符合要求的文档。且在需要全篇格式修改时，更换或修改模板文件，即可直接重新编译为新的样式文档，这对于 word 新手使用 word 的感受来说是不可思议的。

本项目的目的是为了创建一个符合中南大学研究生学位论文（博士）撰写规范的 TeX 模板，解决学位论文撰写时格式调整的痛点。

图 43 幅，表 2 个，参考文献 8 篇

关键词： 中南大学；学位论文；LaTeX 模板

分类号： TP391



Central South University

Abstract: LaTeX can be compiled into a pdf of uniform format using the set template. At present, most domestic publishers and universities still use word. Because of its powerful function and flexibility, when faced with fixed-form papers by novices, simple matters such as typesetting, numbering, and reference documents will bring many difficulties and troubles. For some problems that need to be modified throughout, to achieve the efficiency of LaTeX, it requires a high level of skill for word users.

In order to focus on the writing of papers, many international journals and universities support the writing and submission of LaTeX. Novices don't need to care about formatting issues. They only need to use a few symbolic labels step by step to get the documents that meet the requirements. And when you need to modify the entire format, you can directly recompile the template file by replacing or modifying the template file. This is incredible for the word novice to use the word.

The purpose of this project is to create a TeX template that meets the specifications of the graduate degree thesis (PhD) of Central South University, and to address the pain points of format adjustment during the dissertation writing.

Keywords: CSU; LaTeX; Template

Classification: TP391

目 录

摘要	I
ABSTRACT	II
目录	III
插图索引	VII
表格索引	IX
符号说明	X
第 1 章 绪论	1
1.1 研究背景与意义	1
1.2 国内外研究现状	1
1.3 主要研究内容	1
1.4 论文组织结构	1
第 2 章 相关理论概述	1
2.1 用户画像	2
2.2 深度学习算法	2
2.2.1 卷积神经网络	2
2.2.2 循环神经网络	2
2.2.3 基于注意力机制的神经网络	2
2.3 强化学习算法	2
2.3.1 基于随机过程的多臂老虎机	2
2.3.2 基于上下文的多臂老虎机	2
2.4 本章小结	2
第 3 章 系统描述与问题建模	2
3.1 系统描述	2
3.1.1 离网用户预测模型	2
3.1.2 预离网用户偏好生成模型	2
3.1.3 预离网用户干预模型	2

3.2	问题建模	2
3.3	问题挑战	2
3.4	本章小结	2
第 4 章	用户数据分析和特征工程	2
4.1	平台描述	2
4.2	数据描述	4
4.3	数据分析	5
4.4	特征工程	9
4.4.1	静态特征工程	9
4.4.2	时序特征工程	13
4.5	本章小结	14
第 5 章	基于自注意力机制的互联网卡用户离网预测模型设计	15
5.1	系统描述与问题建模	15
5.2	基于自注意力机制的编码算法	16
5.3	基于主成分分析算法的特征降维算法	17
5.4	基于多层感知机的分类器设计	17
5.5	本章小结	18
第 6 章	关于离网预测模型的实验评估与结果分析	19
6.1	实验设置	19
6.1.1	基准模型	20
6.1.2	评估指标	20
6.2	用户离网预测模型性能评估	21
6.2.1	系统总体性能	21
6.2.2	Top-U 用户的性能	22
6.3	参数影响	24
6.3.1	性别参数的影响	24
6.3.2	年龄参数的影响	25
6.3.3	APP 参数的影响	26
6.3.4	套餐参数的影响	26
6.4	消融实验	27
6.5	本章小结	28

第 7 章 预离网用户偏好生成算法设计	29
7.1 系统描述	29
7.2 离网原因与偏好的相关性分析	29
7.3 离网偏好排名归一化	30
7.4 不可信用户过滤机制	30
7.5 本章小结	30
第 8 章 基于汤普森采样的预离网用户干预算法设计	31
8.1 系统描述与问题建模	31
8.2 奖励生成模型设计	31
8.3 基于汤普森采样的用户-干预措施匹配算法设计	31
8.3.1 动作空间	32
8.3.2 奖励机制设计	32
8.4 基于模拟干预结果机制的训练	32
8.5 本章小结	32
第 9 章 实验评估与结果分析	33
9.1 实验设置	33
9.1.1 对比方案	33
9.1.2 评估指标	33
9.2 预离网用户干预框架性能评估	33
9.2.1 总体性能	33
9.2.2 健壮性测试	35
9.3 参数影响	35
9.3.1 城市	35
9.3.2 年龄	35
9.3.3 离网风险	35
9.4 本章小结	35
第 10 章 总结与展望	36
10.1 工作总结	36
10.2 未来工作展望	36
第 11 章 绪论	37
11.1 研究背景与意义	37

11.2 主要研究工作.....	39
11.3 论文组织结构.....	39
第 12 章 图像布局	40
12.1 单图布局	40
12.2 横排布局	40
12.3 竖排布局	40
12.3.1 竖排多图横排布局	40
12.3.2 横排多图竖排布局	41
12.4 本章小结	41
第 13 章 表格插入示例	42
第 14 章 算法示例	43
第 15 章 公式、定理、证明插入示例	44
第 16 章 参考文献插入示例	46
第 17 章 总结与展望	47
17.1 工作展望	47
参考文献	48
附录 A （附录名称）（三号黑体，加粗）（必要时）	49
攻读学位期间主要的研究成果	50
致 谢	51

插图索引

图 1-1	论文组织架构图	1
图 4-1	大数据平台系统架构	3
图 4-2	用户侧数据描述	4
图 4-3	物品侧数据描述	5
图 4-4	互联网卡发展趋势	6
图 4-5	互联网卡离网用户-日期热力图	7
图 4-6	互联网卡用户 Top10 离网原因	8
图 4-7	互联网卡用户账户余额对比	9
图 4-8	月不同阶段的互联网卡用户平均消耗流量值对比	10
图 4-9	互联网卡用户月 APP 使用频次对比	10
图 4-10	互联网卡用户流量活跃熵对比	11
图 4-11	互联网卡用户目标编码值对比	12
图 4-12	互联网卡用户流量记录条数序列对比	13
图 4-13	互联网卡用户流量异常天数对比	14
图 5-1	互联网卡用户离网预测模型架构	15
图 6-1	基于滑动窗口的离网预测实验设置	19
图 6-2	离网预测性能对比	21
图 6-3	Top-U 用户的召回率对比	22
图 6-4	Top-U 用户的精准率对比	23
图 6-5	Top-U 用户的 F1 分数对比	23
图 6-6	性别参数的影响	24
图 6-7	年龄参数的影响	25
图 6-8	APP 参数的影响	26
图 6-9	套餐参数的影响	27
图 7-1	预离网用户偏好生成模块图	29
图 7-2	互联网卡用户所有离网原因示意图	29
图 7-3	互联网卡用户可建模的离网原因示意图	30
图 8-1	干预策略匹配模块图	31
图 8-2	奖励模型内部示意图	31
图 9-1	匹配算法性能对比图	33
图 9-2	干预框架总体性能对比图	34
图 9-3	奖励总和对比图	34

图 9-4	平均奖励对比图	35
图 9-5	收入总和对比图	35
图 9-6	城市参数对于奖励总和影响的对比图	36
图 9-7	城市参数对于平均奖励影响的对比图	36
图 9-8	年龄参数对于奖励总和影响的对比图	37
图 9-9	年龄参数对于平均奖励影响的对比图	37
图 12-1	单图布局示例	40
图 12-2	横排布局示例	40
图 12-3	竖排布局示例	41
图 12-4	竖排多图横排布局	41
图 12-5	横排多图竖排布局，斜体 <i>emph A</i> ， <i>A</i> ，斜体 <i>text A</i>	41

表格索引

表 6-1 消融实验.....	27
表 13-1 表格为三线表斜体 <i>emph A</i> , <i>A</i> , 斜体 <i>text A</i>	42

符号说明

符号	意义	单位（量纲）
频率	赫 [兹]	Hz

第1章 绪论

- 1.1 研究背景与意义
- 1.2 国内外研究现状
- 1.3 主要研究内容
- 1.4 论文组织结构

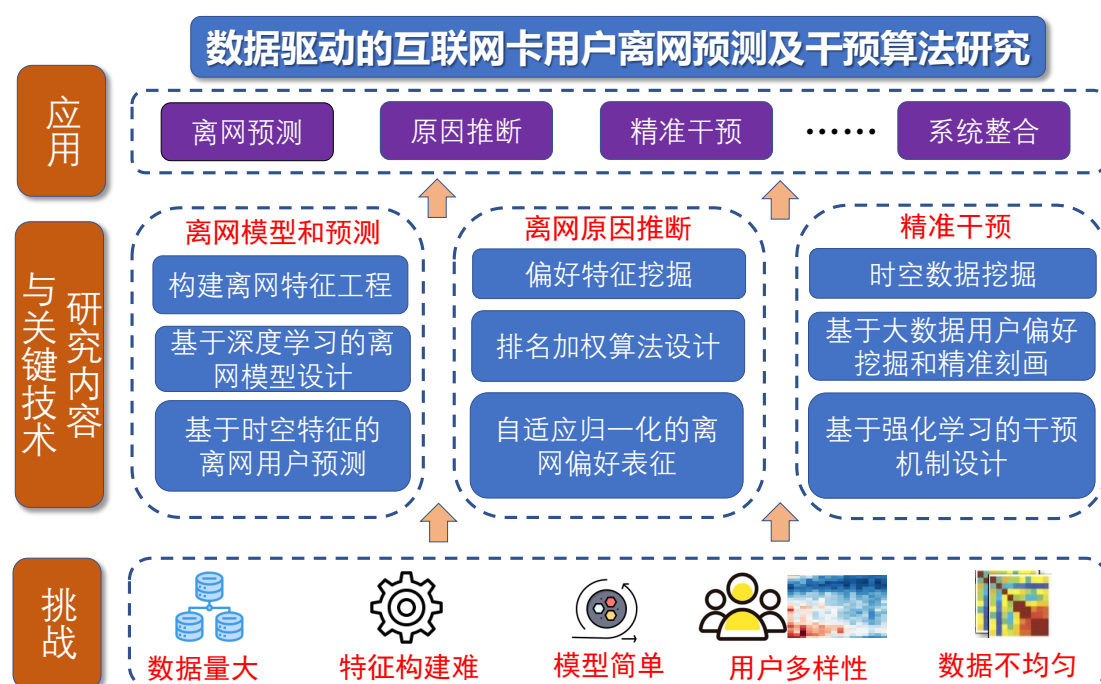


图 1-1 论文组织架构图

第2章 相关理论概述

2.1 用户画像

2.2 深度学习算法

2.2.1 卷积神经网络

2.2.2 循环神经网络

2.2.3 基于注意力机制的神经网络

2.3 强化学习算法

2.3.1 基于随机过程的多臂老虎机

2.3.2 基于上下文的多臂老虎机

2.4 本章小结

第3章 系统描述与问题建模

3.1 系统描述

3.1.1 离网用户预测模型

3.1.2 预离网用户偏好生成模型

3.1.3 预离网用户干预模型

3.2 问题建模

3.3 问题挑战

3.4 本章小结

第4章 用户数据分析和特征工程

4.1 平台描述

在本章中，本文会首先介绍平台架构，然后描述数据格式、规模等信息，接着进行了三个方面的数据分析，最后进行了相应的特征工程。

运营商们每天都会生产和存储巨量的数据，其中分为业务支持系统（BSS）和运营支持系统（OSS），这两者也构建了大数据平台的底层，从而用来提升业务和运营表现。具体来说，图4-1展示了流量运营商的大数据平台架构，其中包

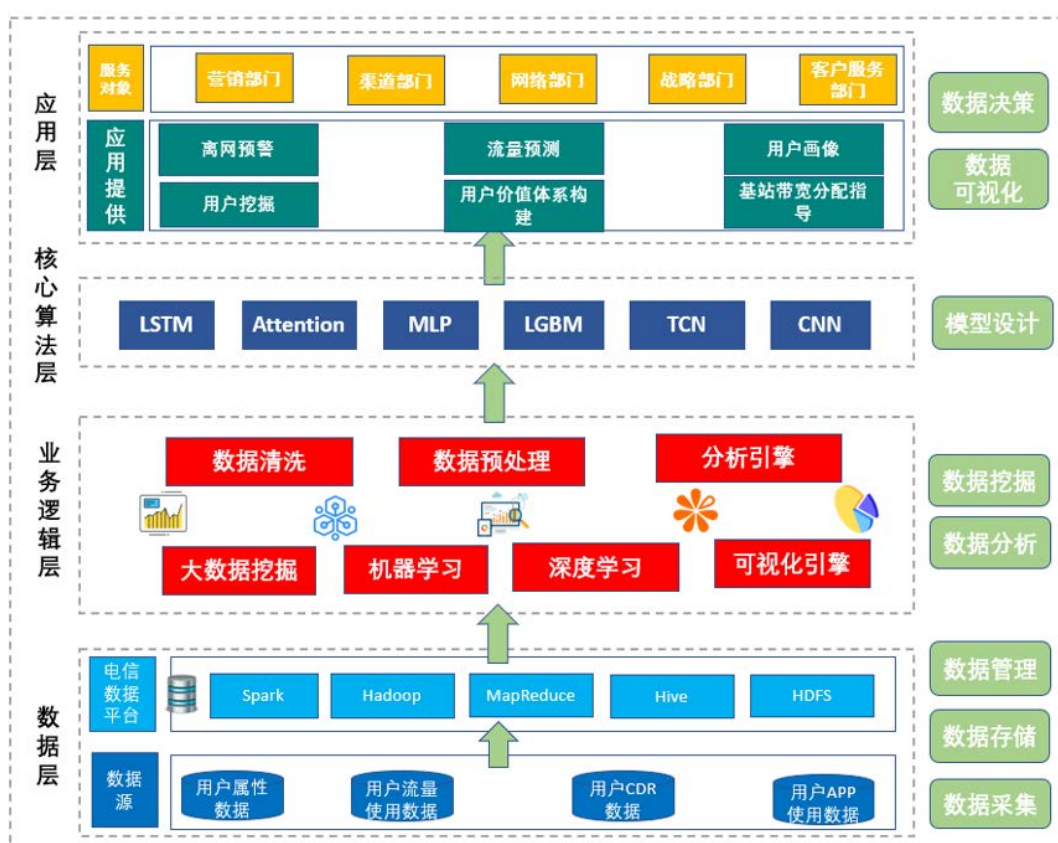


图 4-1 大数据平台系统架构

括数据层、业务逻辑层、核心算法层和应用层。

首先是数据层，数据层主要是承担数据采集、数据存储和数据管理的功能。首先从数据源中保存来自业务支持系统（BSS）和运营支持系统（OSS）的多维度数据，包括用户属性数据、用户流量使用数据、用户 CDR 数据、用户 APP 数据等。然后通过数据操作工具来做这些数据做周期性的更新、修改和删改工作从而提供给上层做其他工作。其中，Hadoop 分布式文件系统（HDFS）通过分布式硬件平台来提供基础的数据存储功能。Hive/Spark SQL 则是提供数据查询、清洗、过滤等功能，尤其是的提高了针对海量数据的开发和处理效率。而 MapReduce 则是提供大数据的并行计算范式从而缩短数据处理时间。

然后是业务逻辑层，次层主要为不同的业务部门提供数据分析、挖掘和建模功能。举例来说，其中包括数据预处理、统计分析、特征工程、机器学习等模块。

接着是核心算法层，主要是实现一些机器学习和深度学习模型的核心算法，其中包括针对表格型数据的轻度梯度提升机（LGBM）和多层感知机（MLP）等模型，针对序列数据的长短期记忆网络（LSTM）、时域卷积网络（TCN）、基于自注意力机制的深度学习神经网络（Transformer）等模型，针对图像的卷积神经网络（CNN）、自注意力视觉神经网络（ViT）等模型等模型。

最后是应用层，主要是提供经过开发人员开发和封装好的全流程自动的应用程序，为下游的营销部门、渠道部门等提供用户画像、离网预警等功能。

4.2 数据描述

CDR 表							
主叫号码 F7D97A	被叫号码 BDA794	开始时间 20200421123236	结束时间 20200421123305	时长 29	基站 ID 218090	城市 ID 11
流量表							
加密号码 F7D97A	开始时间 20200406041136	结束时间 20200406051136	下行流量 3700375	上行流量 1371049	时长 3600	基站 ID 218090
用户属性表							
加密号码 F7D97A	年龄 18	性别 male	余额 304.75	出账金额 19.00	套餐金额 19	城市 ID 11	流量 21.47
月份 ID 202004	在网时长 16	经度 112.59	纬度 28.19	单停次数 3	双停次数 1	套餐 ID 1186881
APP 表							
加密号码 F7D97A	一级标签 社交软件	二级标签 通信	三级标签 微信	使用流量 323	使用次数 7812	天数 10
停机表							
加密号码 F7D97A	停机时间 2020.11.10	停机类型 单停	是否复机 否
2020.04 至 2021.02 的数据统计							
互联网卡用户数量:400 万				停机表记录条数:1000 万条			
CDR 记录条数: 35 亿				流量表记录条数: 400 亿			
属性表记录条数: 400 万				APP 表记录条数: 4 亿			

图 4-2 用户侧数据描述

首先，本文来描述一下用户侧数据，如图4-2所示。

时间范围. 本文拥有 2020 年 4 到 6 月，11 月到 12 月和 2021 年 1 到 2 月的 7 个月的数据。

用户类型. 本文过滤掉了政企用户、家庭用户和其他用户，只留下互联网卡个人用户。

数据规模. 在这 7 个月的数据中，一共有 400 万的互联网卡个人用户，400 万条以月为粒度的属性表记录，35 亿条以次为粒度的 CDR(通话细节记录) 表记录，400 亿条以次为粒度的流量表记录，4 亿条以月为粒度的 APP(应用程度) 表记录，1000 万条以次为粒度的停机表记录。其中属性表记录的为用户属性数据，而其他四个表记录的为用户行为数据，尤其是流量表和 CDR 表的数据尤为珍贵，能够刻画用户的序列行为。但是从另一方面来说，如此海量的数据也给数据分析和模型训练推理带来了极大的硬件资源、方法性能、时间压力。

具体字段.
数据用途.

表 2
数据描述

干预表							
加密号码 F7D97A	营销结果 挽留失败	离网原因 套餐不合适	营销时间 201123 13:58	通话时长 136	联系电话 13826011809	城市 岳阳
套餐信息表							
套餐名称 B 站权益卡	主套餐 ID 9013357	主套餐名称 20200406051136	套餐内容 19 元	可选包 ID 9015724	可选包名 流量优惠包	可选项 1GB 流量
2021.11 至 2021.02 的数据统计							
互联网卡用户数量:5 万				套餐信息表记录条数: 80			
干预表记录条数: 5 万							

图 4-3 物品侧数据描述

接着，本文来描述物品侧数据，如图4-3所示。

时间范围. 本文拥有 2020 年 11 月 12 月和 2021 年 1 2 月的 4 个月的数据。

用户类型. 本文也同样过滤掉了政企用户、家庭用户和其他用户，只留下互联网卡个人用户。

数据规模. 在这 4 个月的数据中，一共有 20 万的离网的互联网卡个人用户，20 万条以次为粒度的干预表记录，80 条以个为粒度的套餐信息表记录。其中干预表记录的为运营商人工客服对离网客户的干预数据，而套餐表记录的则为互联网卡套餐的相关数据。

具体字段.

4.3 数据分析

在本章中，本文会首先针对互联网卡正常用户和离网用户做全面的数据分析，接着探索在哪些属性和行为数据上互联网卡离网用户和正常用户表现出较大差异，然后针对这些数据做相应的特征计算，提取用以区分互联网卡正常用户和离网用户的关键特征。

本文首先定义了互联网卡离网用户，然后分析了互联网卡的离网趋势以及互联网卡用户的离网原因分布。

离网用户定义. 离网用户也被称为流失用户，往往是用户在使用过程中因服

务不满意、资费贵、改用其他竞品等原因而选择不再使用互联网卡。其中又分为两种，分别是主动性离网和被动性离网。其中主动性离网是指用户主动到运营商营业厅提出销卡的请求，还包括退还余额等行为。而被动性离网则指用户保持了至少两周 14 天的双停状态，没有通过充值话费等行为使得相应手机号复机。则运营商会主动将这类号码销户，之后再销售给其他用户。本文的研究主要是针对被动性离网，因为主动性离网行为只占互联网卡所有离网行为的不到 10%，尚且不受运营商们重视。

互联网卡趋势。 为了了解互联网卡这个业务的发展趋势，本文绘制了从

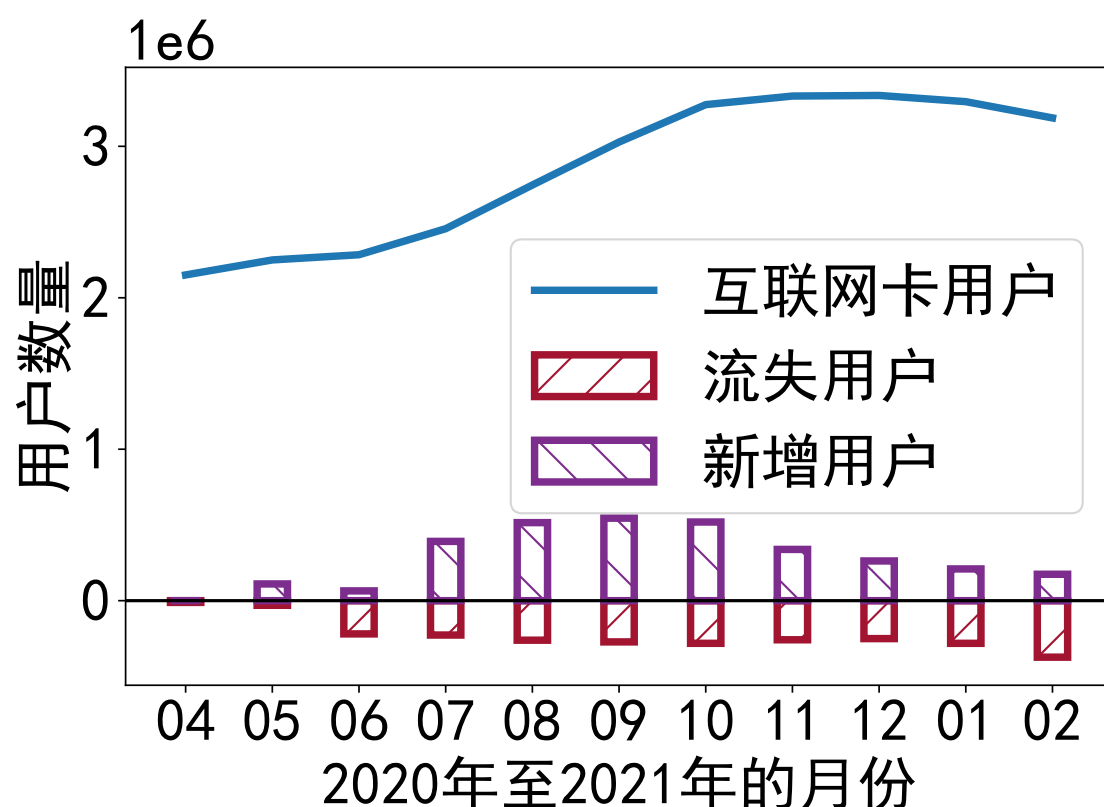


图 4-4 互联网卡发展趋势

2020 年 4 月到 2021 年 2 月的发展趋势图，如图 4-4 所示。从图中，本文可以观察到互联网卡在初期增长得十分迅速，但是后期增长趋势放缓，从 2020 年 4 月到 2021 年 2 月一共增长了 50% 的用户，数量约为 100 万。究其背后的原因是因为尽管在一开始互联网卡新增用户占大多数，但是随着时间流逝，离网用户的数量迅速增长，甚至超过了月新增用户。这使得离网问题变得日益严重起来，对运营商维系互联网卡业务的稳定性提出了挑战。

离网时间分析。 为了进一步地理解互联网卡用户的离网行为，本文分析了互联网卡用户 7 个月的停机数据并分析了每个用户的离网时间。图 4-4 展示了 7 个月份内不同日期的离网用户数量，本文可以观察到在每个月的 9 号至 19 号拥

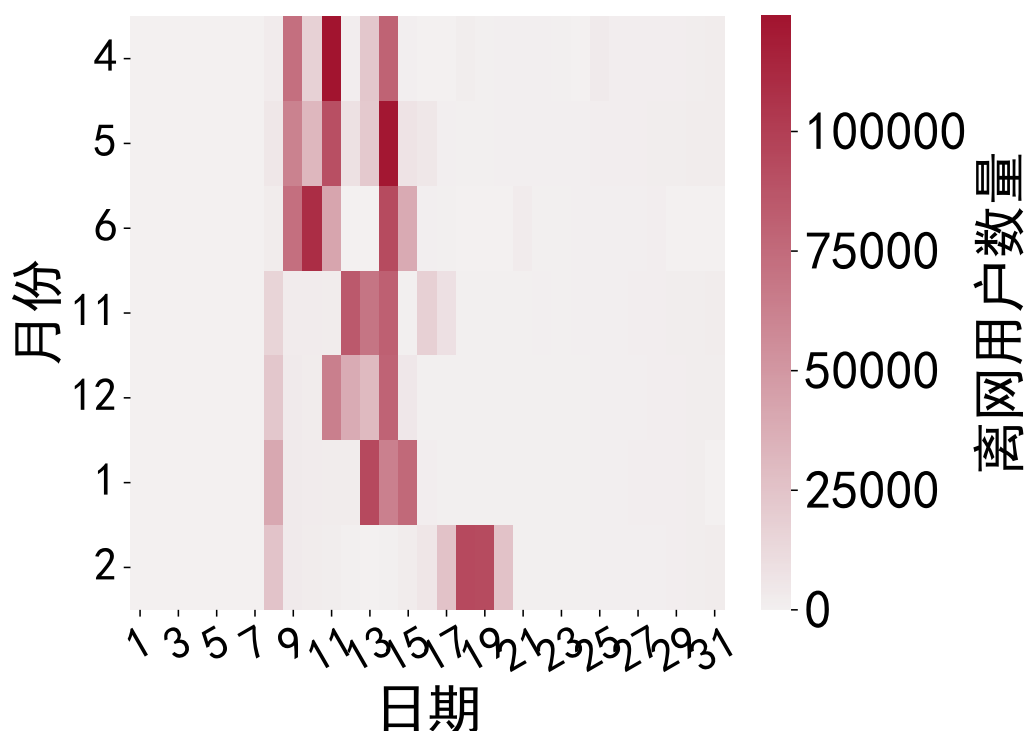


图 4-5 互联网卡离网用户-日期热力图

有绝大多数的离网用户（超过了 95%）。本文还可以发现不同月份的离网用户数量和离网时间分布具有一定差异。举例来说，2021 年 1 月 14 日的离网用户数量可以达到 122029 个，然而其他月份只有更少的离网用户。此外，本文还观察到在每个月的月初和月末都基本没有互联网卡用户离网。这可能是与运营商的工作特点有关联，运营商的每个月初和月末常常需要做些清点工作。上述发现带给本文两个启示，首先，由于用户的离网行为在每月都有相同和不同之处，不是非常稳定，这带给整个系统的建模带来了一定挑战。其次，这也启示本文互联网卡用户偏好月中离网，因此本文应当在上月末或者当月初就完成当月互联网卡的离网用户预测，才对运营商有实际意义。又因为互联网卡庞大的用户体量，运营商只能在每月初才能完成对所有数据的采集和基本处理工作，因此本文的工作最终也是在每月月初进行推理和产出的。

离网原因分析。 为了理解为什么一部分互联网卡用户倾向于俩王，本文基于运营商客服收集的离网反馈分类了用户离网原因。图4-6展示了互联网卡用户离网的数量最大的前 10 个原因。本文可以看到互联网卡用户最多是因为“号卡太多”这个原因离网的。因为现在不同的运营商都在计划占据整个互联网卡市场。他们通过向新用户频繁宣传和给予大量折扣的优惠来吸引他们使用自家的

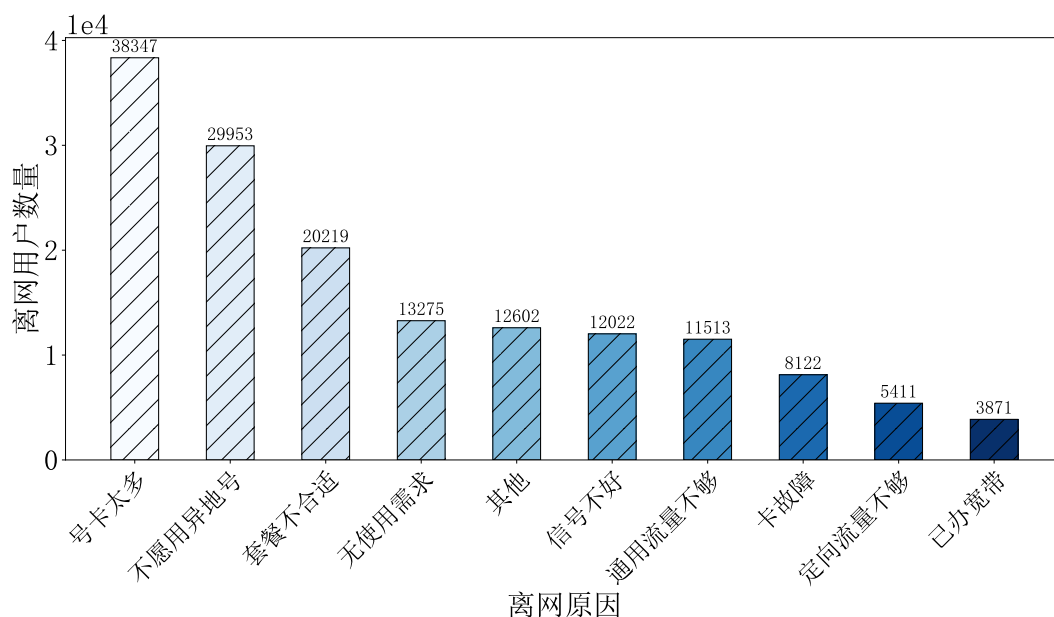


图 4-6 互联网卡用户 Top10 离网原因

互联网卡。因此，部分互联网卡用户可能会被其他运营商的优惠政策所吸引，从而离网转向使用他运营商的互联网卡。第二大离网原因是互联网卡用户不愿使用异地号。因为运营商的不同省份都是各自独立的，都会向全国各地兜售互联网卡。因此部分互联网卡用户使用的互联网卡的负责公司所出的省份与自己所处的省份并不同，在接听和拨打电话时常常会引起误会，而这个现象则会导致部分用户不想使用异地的互联网卡。并且由于异地的关系，手机信号和网络可能变得不稳定，从而影响用户的使用体验。此外，套餐不合适和无使用需求也占据了相当大的比例，这些是能被运营商优化的。从长期来说，互联网卡用户和原因的分布随着时间流逝和新用户的加入可能会变得不同。因此，本文只关注用户离网原因的类别。然后，本文可以为不同离网原因类别设计相应的干预策略从而挽留住那些已经离网或者将要离网的互联网卡用户。总的来说，用户离网原因分布的改变并不会影响系统的总体性能。

因为运营商通过向互联网卡用户收取基本的套餐费用以及额外的服务费用来赚取利润的，所以互联网卡这个业务市场高度依赖于互联网卡用户的数量。因为用户离网问题对运营商来说至关重要，所以本文设计和实现的系统不得不理解互联网卡离网用户的潜在行为并且提前预测用户的离网行为从而发起早期的干预。因此，本文迫切需要了解互联网卡用户画像，在此基础上制定更有效的业务策略，防止他们流失，这也是本工作的动力之一。

4.4 特征工程

在本小节中，本文会展示基于数据分析的一些重要特征，主要分成两类，一类是静态画像特征，另一类是时间序列特征。

4.4.1 静态特征工程

账户余额. 当互联网卡用户即将离网的时候，他们通常倾向于花光账户里

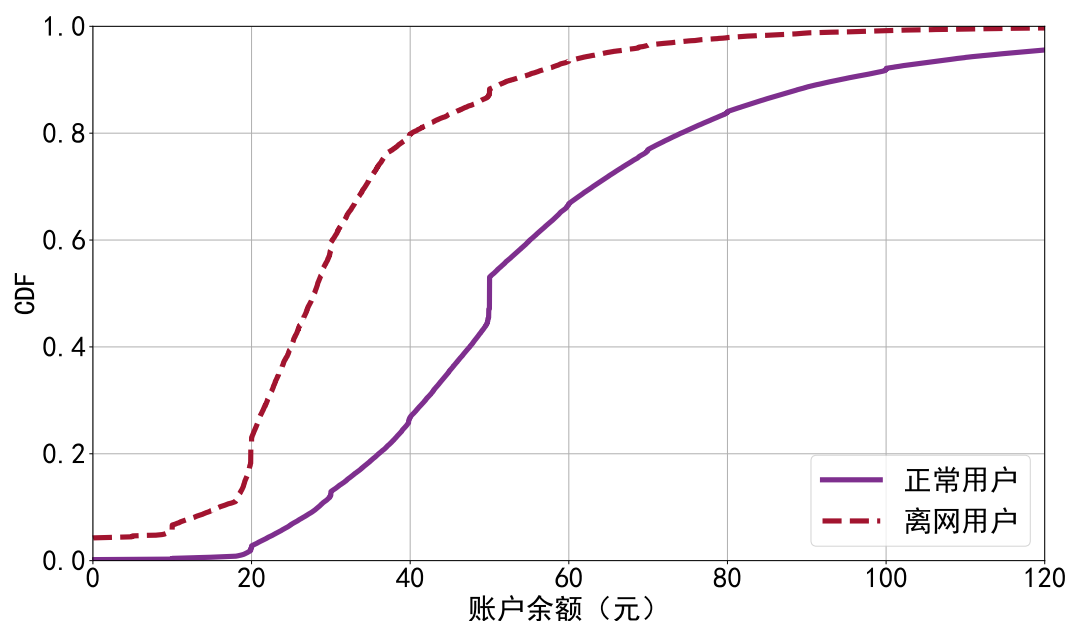


图 4-7 互联网卡用户账户余额对比

的余额，因此账户余额越低，用户越容易离开运营商。图4-7展示了互联网卡正常用户和离网用户的累积分布函数（CDF）图，可以看出这两条曲线有着非常大的差距。详细来说，对于 80% 的用户来说，离网用户的账户余额都小于 40 元，而正常用户的账户余额都小于 75 元，这意味着账户余额在用户发生离网行为前是一条关键的线索。

平均流量消耗阶段. 对于即将离网的用户来说，他们的网络行为往往会发生变化。为了捕捉这个特征，本文把每个月份平均分成三个等长的阶段，分别是上旬，中旬和下旬。图4-8显示了正常用户和离网用户在月份不同阶段的流量消耗变化。从中可以观察到对于正常用户来说，平均流量消耗并没有什么区别。但是在离网用户的三个不同阶段，流量消耗曲线有很明显的差距。举个例子，在上旬，中旬和下旬，平均消耗流量为 0 的用户在所有离网用户占比分别达到了 30%,80% 和 96%。

APP 使用频次. 为了捕捉互联网卡用户的 APP 使用习惯，本文基于采集的

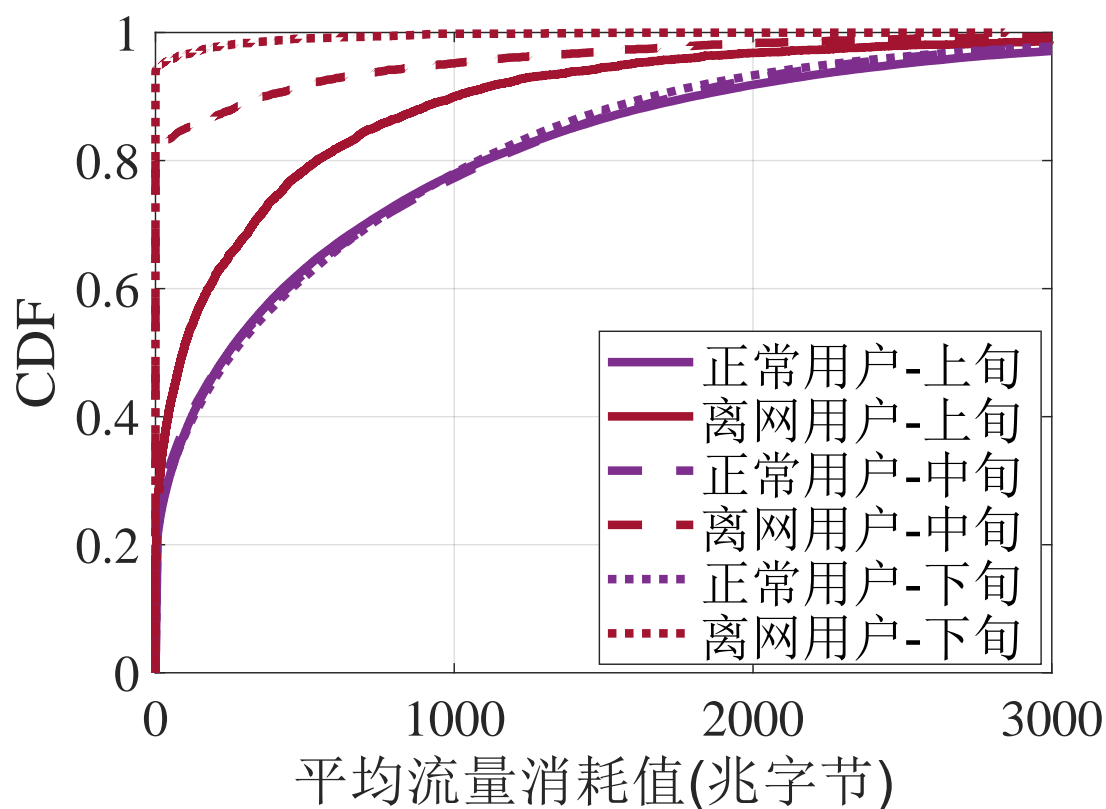


图 4-8 月不同阶段的互联网卡用户平均消耗流量值对比

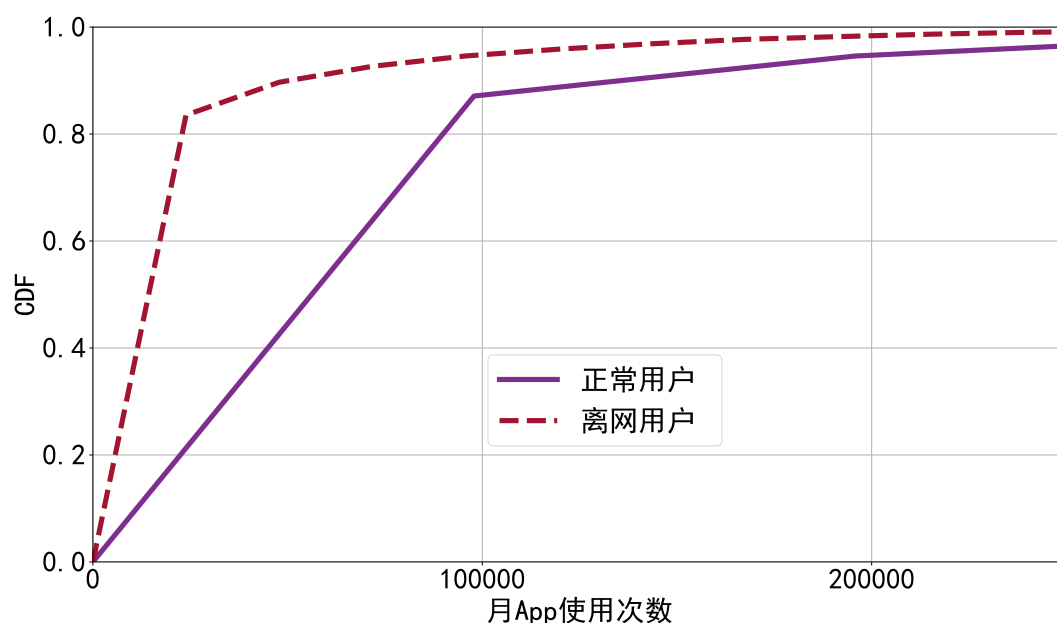


图 4-9 互联网卡用户月 APP 使用频次对比

APP 表计算了每个互联网卡用户在一个月内的所有 APP 使用频次。在图4-9中，本文描绘了正常用户和离网用户关于 APP 使用频次的对比 CDF 图，并且其中有非常大的不同。具体来说，正常用户的 APP 使用频次的中位数是 36075 次，而

离网用户的对应中位则是 18442 次。这意味着 APP 使用频次对于互联网卡用户来说是一个非常有价值的特征用以区别正常用户和离网用户。

活跃熵.

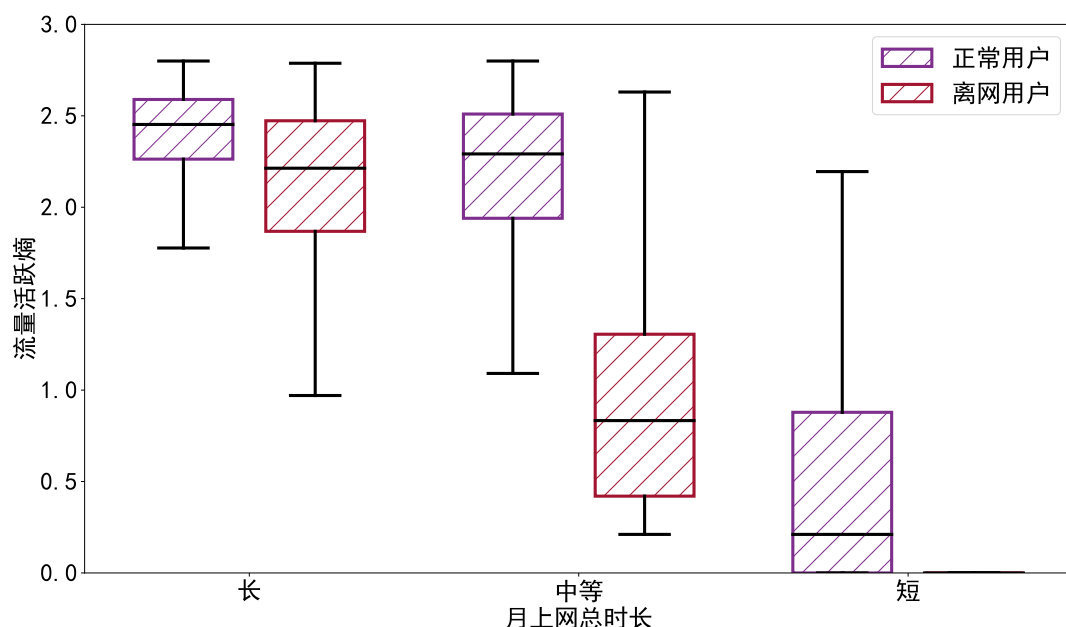


图 4-10 互联网卡用户流量活跃熵对比

和上述思路保持一致, 本文基于香农信息熵探索了在一个月內日指标观察值 (包括, 使用流量大小, 流量记录条数, 上网时长等) 的不确定性。对于某个互联网卡用户 u 来说, 他关于上网时长序列的活跃熵 $H(D_u)$ 可以被如下公式(4-1) 计算。

$$H(D_u) = \sum_{k=0}^{\min(\maxbin, \text{len}(D_u))} p_k \log \frac{1}{p_k} \quad (4-1)$$

其中 D_u 是用户 u 的上网时长序列, p_k 表示上网时长序列中的数值落在第 k 个箱子的概率。此外, \maxbin 是分箱的数量, $\text{len}(D_u)$ 是 D_u 的长度。如果上网时序列的活跃熵比较大, 这意味着上网时长序列中的数值在区间 $[\min(D), \max(D)]$ 中更为分散和混乱。否则, 如果上网时序列的活跃熵比较小, 这意味着上网时长序列中的数值都集中在某个较小的确定区间内。在图4-10中, 本文同时绘制了正常用户和离网用户关于日上网时长活跃熵的箱线图。并且用户们被分成了三组, 分别是上网时长较长, 上网时长中等和上网时长较短。本文可以观察到两种类型用户的不同行为模式, 其中离网用户有着更小的熵值, 这显示了他们有这更简单的网络行为模式, 从而使得他们能够同正常用户区分开来。

目标编码. 为了构建目标编码, 比如将分类特征替换为相应目标的后验概

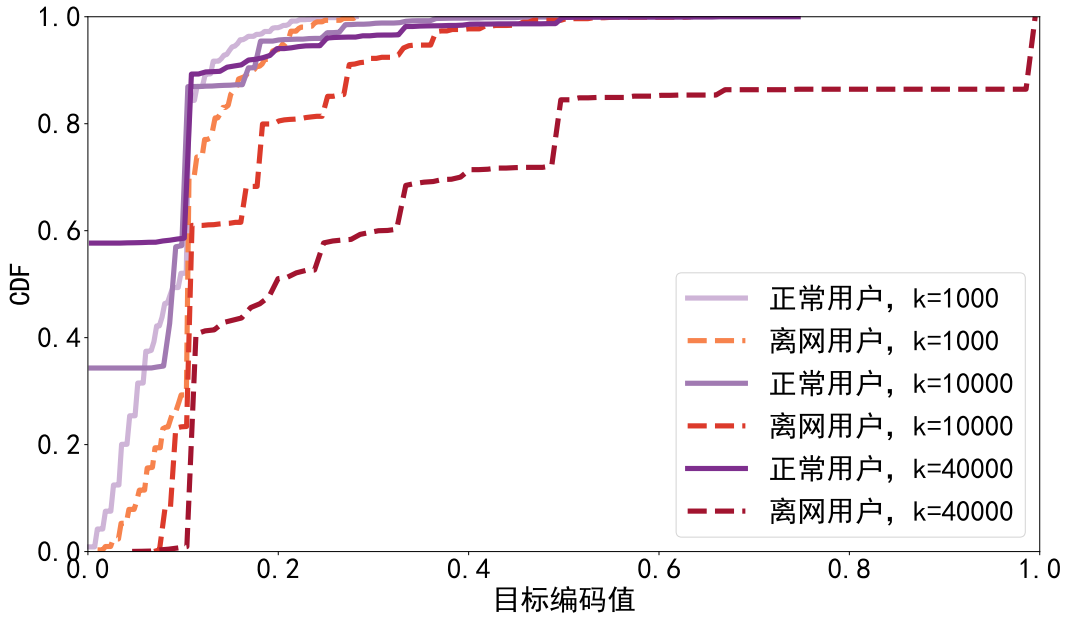


图 4-11 互联网卡用户目标编码值对比

率，本文首先根据用户的流量消耗值将互联网卡用户分组。特别地，本文将流量数据等宽地平均分到了 k 个箱子中，每个箱子的宽度 (w) 等于

$$w = (Max - Min)/k \quad (4-2)$$

其中 Max 和 Min 分别代表了用户在一个半月内的总流量使用值的最小值和最大值，并且每个箱子的边界值分别是 $\{Min + w, Min + 2w, \dots, Min + (k - 1)w\}$ 。因此，针对流量分箱的目标编码值，比如以符号 \vec{R} 表示，可以被如下公式(4-3)计算

$$\vec{R} = Concat\left(\frac{\sum_{j=1}^{n_i} u_{ij} \cdot y_{ij}}{n_i}\right), i = 1, 2, \dots, k, \quad (4-3)$$

其中 n_i 表示在第 i 个箱子中的用户数量， u_{ij} 表示在第 i 个箱子中的第 j 个用户， $y_{ij} \in \{0, 1\}$ 表示目标值，比如， $y_{ij} = 1$ 表示这个用户是离网用户，否则，这个用户就是正常用户、此外， $Concat$ 函数用于拼接 k 个值到向量 \vec{R} 中。图4-11描绘了在分箱数量不同时正常用户和离网用户的目标编码值的 CDF 对比图。本文可以观察到，当 $k=1000$ 时，离网用户和正常用户之间的差距还不是特别明显。但是，当分箱数量增加的时候，比如 $k=40000$ 时，一个良好的性能差距浮现出来，这也意味着正常用户和离网用户被分发到不同的箱子后计算的目标编码值可以有效地区分两者。

值得注意的是，除了上述提到的重要特征，其他基础画像特征，比如年龄、性别、开卡日期和终端类型等也被提取和注入到模型当中了。

4.4.2 时序特征工程

除了静态特征，时序特征对于学习模型来说也是非常重要的因为离网行为通常是一个渐进进程而不是一个突发事件。

流量序列。 图4-12展示了在一个月内的互联网卡用户的日产生流量记录条

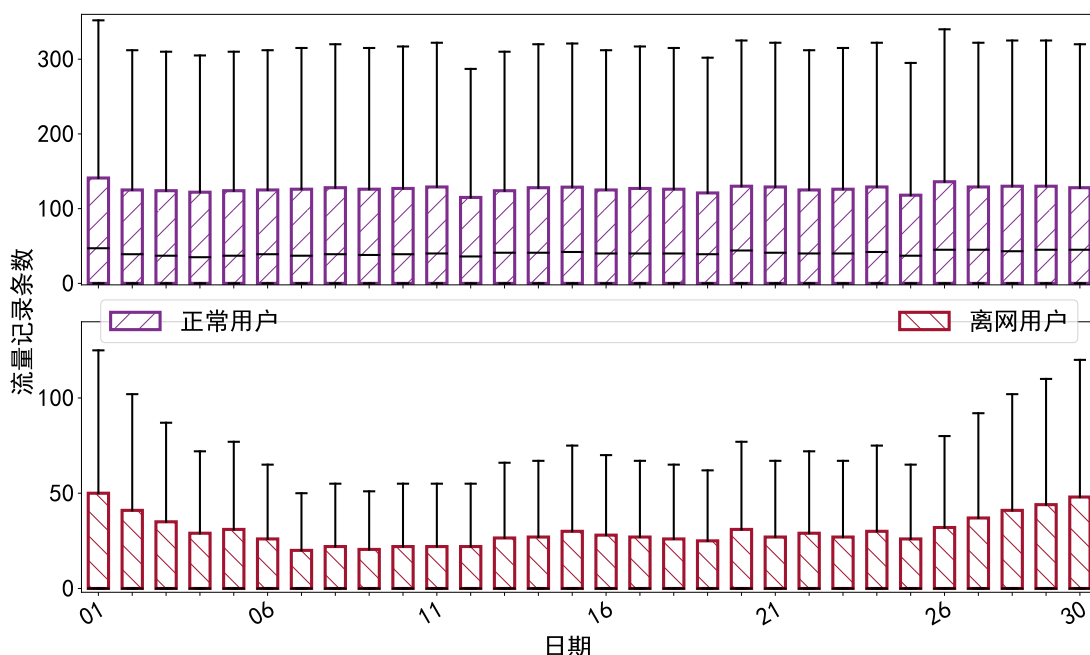


图 4-12 互联网卡用户流量记录条数序列对比

数的箱线图。本文可以从中观察到正常用户每日通常比离网用户都产生了更多数量的流量记录条数。更加重要的是，对于正常用户来说，时序行为的差异十分小，但是对于离网用户来说他们的流量使用行为显得更不稳定。这也意味着这种时间相关性能被加以利用用来区别这两种类型的用户。需要指出的是，除了每日流量记录，其他日粒度特征还包括上行流量值，下行流量值，上网时长，通话次数等，也都被提取成时序特征并且喂给了后续的学习模型。

流量异常天数。 除了分析互联网卡用户的流量统计特征，本文还检测了互联网卡用户在一个月内的哪些日期出现了流量异常行为，因为异常值往往表征着此用户表现同以往不同的行为，很有可能会趋向离网。明确来说，基于用户的日序列行为，本文对所有用户累加了使用特征值，包括上行流量值，下行流量值，上网时长和流量记录条数。因此，对每个序列特征来说，它都能被表征为 $X = [B_1, B_2, \dots, B_i, \dots, B_n]$ ，其中 B_i 表示这个月内第 i 天的统计特征。为了获得更平稳的序列特征，本文对上述得到的序列特征计算了它们的一阶前向差分形式，

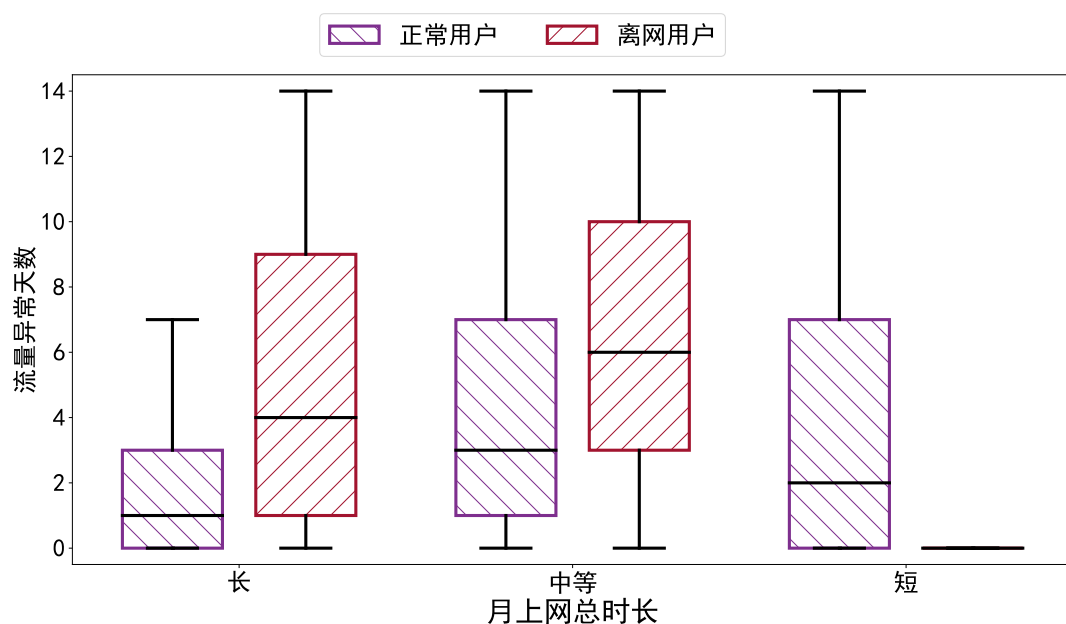


图 4-13 互联网卡用户流量异常天数对比

用来 $X' = [F_1, F_2, \dots, F_i, \dots, F_{n-1}]$ 表示，可以用以下公式(4-4)计算。

$$F_i = B_{i+1} - b_i \quad (4-4)$$

然后，本文定义了异常值 $E(X')$ ，可以通过使用四分位距（IQR）加上以下公式(4-5)来判定。

$$E(X') > Q_u + \gamma * IQR || E(X') < Q_l * IQR \quad (4-5)$$

其中 $IQR = Q_u - Q_l$ ， $\gamma = 1.5$ 。 Q_u 是上四分位值，显示只有 1/4 的观察值比它大，然后 Q_l 是下四分位值，这就意味着只有 1/4 的观察值比它小。因此，每天的流量行为是否是异常值可以通过计算用户序列特征前向差分值的异常值来判断。在图4-13中，本文为正常用户和离网用户都描绘了月异常天数的箱线图。类似地，其中用户也被分成了三组，依据为在一个月内的总上网时长从高到低依次排序。具体来说，其中较高和中等的离网用户组明显拥有更多的异常天数，能够被利用来识别离网事件。但是对于使用较少的离网用户组来说，离网用户比正常用户有着更少的异常天数。这是合理的，因为他们有着稀疏的上网行为，这也导致了没有异常值产生。

4.5 本章小结

第5章 基于自注意力机制的互联网卡用户离网预测模型设计

5.1 系统描述与问题建模

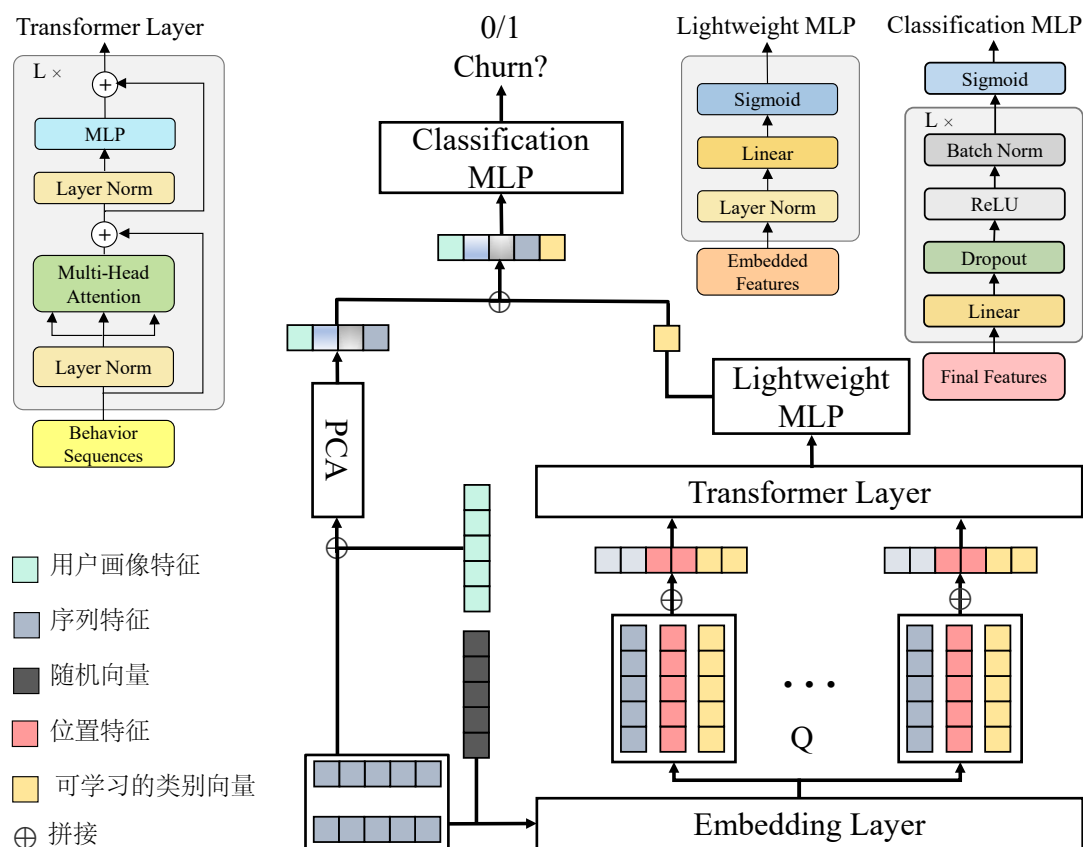


图 5-1 互联网卡用户离网预测模型架构

为了预测潜在的互联网卡离网用户，本文设计了一个可学习的模型架构：互联网卡离网预测 Inter Card Churn Prediction (ICCP)，其中包括基于主成分分析算法的特征降维算法，基于自注意力机制的编码算法，轻量级的多层感知机和分类的多层感知机的，如图5-1所示。

特征输入. 本文基于采集的数据构建了特征工程来提取用户属性和行为特征。对于这两种类型的特征，本文定义符号 $\mathcal{P} \in \mathbb{R}^{1 \times M}$ 来表示画像特征矩阵，其中 M 表示所有静态特征的数量，定义符号 $\mathcal{T} \in \mathbb{R}^{Q \times D}$ 来表示时序特征矩阵，其中 Q 是所有日粒度特征的数量， D 是序列特征的长度。在为所有用户计算完这两类特征值后，它们将和离网标签一起被输入到可学习的模型来做监督学习。

5.2 基于自注意力机制的编码算法

算法 5-1 针对多重序列特征的嵌入变换算法

Input: 时序特征 $T^{(N*L)}$, 嵌入向量数量 Q , 块长度 V , 块宽度 W , 嵌入向量大小 D , 变换块数量 L'

Output: 用户时序特征的编码向量 $E^{(Q*D)}$

- 1: 把时序特征 $T^{(N*L)}$ 变形成块 $B^{Q*(V*W)}$
- 2: 训练一个全连接神经网络来编码块 $B^{Q*(V*W)}$ 到块 B^{Q*D}
- 3: 随机初始化一个可学习的服从高斯分布的向量 X^D
- 4: 随机初始化一个可训练的服从高斯分布的位置嵌入矩阵 $M^{(Q+1)*D}$
- 5: 拼接块 B^{Q*D} , 向量 X^D , 矩阵 $M^{(Q+1)*D}$ 成 Z_0
- 6: 初始化 $Q_i = W_i^Q * Z_0, K_i = W_i^K * Z_0, V_i = W_i^V * Z_0, i = 1, \dots, 3$
- 7: 计算 $Head_i = softmax(\frac{Q_i * K_i^T}{\sqrt{d_{k_i}}})V_i, i = 1, \dots, 3$
- 8: 计算 $MSA(Z_0) = Concatenate(Head_1, \dots, Head_3) * W^O$
- 9: 计算 $Z'_l = MSA(LN(Z_{l-1})) + Z_{l-1}, l = 1 \dots L'$
- 10: 计算 $Z_l = MLP(LN(Z'_l)) + Z'_l, l = 1 \dots L'$
- 11: 计算 $E^{(Q*D)} = LN(Z_{L'})$
- 12: 返回 $E^{(Q*D)}$

为了捕捉潜在的时间关联性, 矩阵 \mathcal{T} 和一个随机向量会首先被喂入到一个嵌入层 (Embedding layer), 它会输出一个固定维度大小的低维空间向量 Q , 在图5-1可以看出, 这个向量 Q 包含序列, 位置和类别信息, 并用不同的颜色标注了出来。然后这个向量 Q 被输入到了变形层 (Transformer Layer), 其中包括 L 个块, 每个块中包含层归一化 (LN), 多头自注意力 (MSA) 和多层感知机 (MLP), 这相对应的函数可以被如下公式计算。

$$z_0 = [s_{class}; \mathcal{T}^1; \mathcal{T}^2; \dots; \mathcal{T}^{|D|}; \mathcal{T}_{pos}^1; \mathcal{T}_{pos}^2; \dots; \mathcal{T}_{pos}^D], \quad (5-1-a)$$

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1}, l = 1 \dots L \quad (5-1-b)$$

$$z_l = MLP(LN(z'_l)) + z'_l, \quad (5-1-c)$$

$$y = LN(z_L^0), \quad (5-1-d)$$

其中 z_0 是嵌入层的输出, \mathcal{T}^i 表示第 i 个时序特征的信息, 而 \mathcal{T}_{pos}^i 表示 \mathcal{T}^i 的位置嵌入信息。嵌入层和变形层的具体工作流程可见算法8-1。在这之后, 变形层的输出将被输入到轻量级的多层感知机 (Lightweight MLP), 其中包括层归一化, 全连接层和 Sigmoid 层。然后轻量级的多层感知机将会输出一个离网概率值, 后者将会被视为一个新特征注入到之后的分类多层感知机中。

5.3 基于主成分分析算法的特征降维算法

为了捕捉用户的画像信息，本文首先将用户画像特征和时序特征拼接成一个一维向量，比如说， $V_i = \{x_1, x_2, \dots, x_{M+Q \times D}\}$ 。然后这个向量将被输入到主成分分析算法这个组件。它会输出一个压缩向量 $W_i^* = \{w_1, w_2, \dots, w_{d'}\}$ ，其中 $w_{d'}$ 是新维度的大小。接着这个新向量和轻量级多层感知机的输出会被拼接，然后输入到最终的分类多层感知机中。需要特别指出的是，针对画像特征和时序特征的主成分分析算法能够在保持大多数特征成分的同时，比如最大限度地保留原始信息，这也能够减少用于分类的特征的复杂度，加速模型训练和收敛。具体的工作流可见算法5-2。

算法 5-2 针对画像特征和时序特征的主成分分析

Input: 用户画像特征 P , 时序特征 T , 信息阈值 I

Output: 被信息阈值截断的关于用户特征的主成分 P_c

- 1: 初始化时序矩阵 t 为空
 - 2: 初始化信息权重 w 为 0 和主成分数量 n 为 0
 - 3: **for** $i \leftarrow 1 \cdots Q$ **do**
 - 4: $t \leftarrow$ 拼接 t 和 T 的第 i_{th} 个时序特征
 - 5: **end for**
 - 6: $P \leftarrow$ 拼接 t 和 P
 - 7: 并行对所有特征去中心化 $P_i = P_i - \frac{1}{M+Q \times D} \sum_{j=1}^{M+Q \times D} P_j$
 - 8: 计算特征的协方差矩阵 $C = \frac{1}{M+Q \times D} P P^T$
 - 9: 计算特征值 λ 和协方差矩阵 C 对应的特征向量 \vec{v}
 - 10: 根据 λ 降序排列 \vec{v} 来得到矩阵 C'
 - 11: **repeat**
 - 12: $w = \frac{\sum_{i=1}^n \lambda_i}{\sum_{i=1}^{M+Q \times D} \lambda_i}$
 - 13: $n \leftarrow n + 1$
 - 14: **until** $w < I$
 - 15: 选择 C' 的前 n 行来得到变换矩阵 A
 - 16: $P_c = AP$
 - 17: 返回 P_c
-

5.4 基于多层感知机的分类器设计

模型训练. 为了分类一个互联网卡用户是否是离网用户，本文将其建模成一个二分类的问题。本文通过使用分类多层感知机，其中包括 L 个块，每个块包含小批量归一化 (Batch Norm)，ReLU，Dropout 和全连接层。为了训练这个模型，本文采用交叉熵作为损失函数，可以被以下公式(5-2)所计算。

$$\mathcal{L} = \frac{1}{N} \sum_i -[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \quad (5-2)$$

其中 p_i 表示用户 i 离网概率值, 而 $y_i \in \{0, 1\}$ 是表示用户是否离网的标签, 具体来说, $y_i = 1$ 代表此用户是一个离网用户, 否则, $y_i = 0$ 此用户是一个正常用户。 N 表示所有训练样本的数量。

5.5 本章小结

第 6 章 关于离网预测模型的实验评估与结果分析

6.1 实验设置

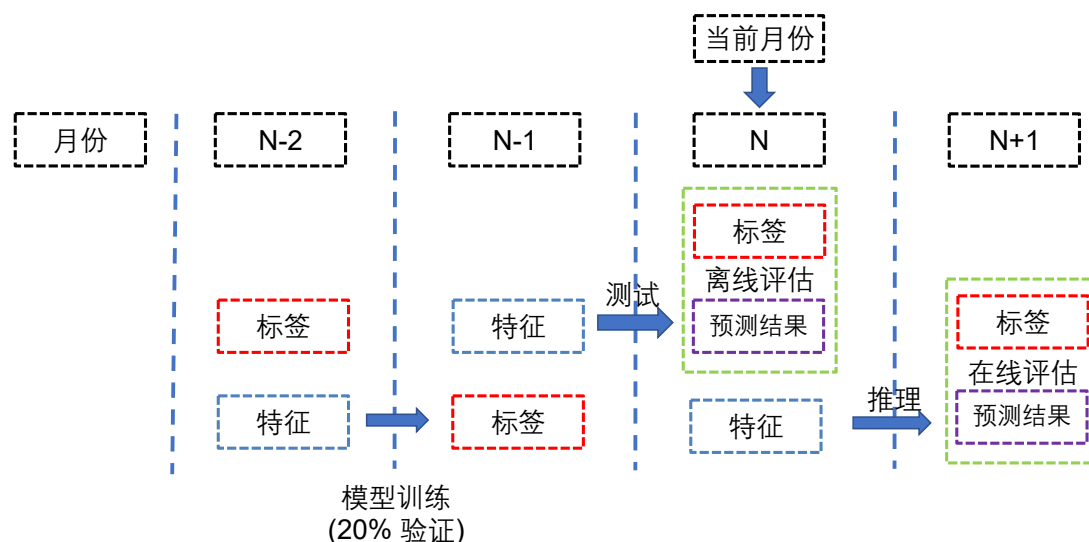


图 6-1 基于滑动窗口的离网预测实验设置

为了评估 ICCP 的性能，本文首先使用了某个运营商的互联网卡用户离网信息来标记离网用户，比如，如果一个互联网卡用户双停超过两个礼拜并且没有复机行为，那该用户就会被标记为离网用户。图5-2展示了在不同月份基于滑动窗口的训练和预测实验设置。需要特别指出的是，假设当前月份是 N 月，本文首先会使用 $N-1$ 月的用户离网信息来标记拥有 $N-2$ 月特征的用户，然后用相同的方法使用 N 月的用户离网信息来标记拥有 $N-1$ 月特征的用户。然后，我们使用拥有 $N-2$ 月特征的用户和 $N-1$ 月的标签来训练 ICCP，其中 20% 的数据样本用于验证。接着，本文使用拥有 $N-1$ 月特征的用户和 N 月的标签来测试 ICCP 的离线性能。最后，依赖于上述机制，本文把 N 月无标签的用户特征数据输入到 ICCP 中，就可以得到下一个月的互联网卡离网用户名单，并且可以在下月评估 ICCP 的在线真实性能。综上所述，本文利用滑动窗口机制，通过将窗口从上个月移动到当前月实现用户的特征捕捉，然后使用 ICCP 模型来预测下月的潜在离网用户。

对于特征工程和性能评估，本文使用了横跨五个月采集的数据，分别是 2020 年的 5 月、6 月，11 月、12 月和 2021 年的 1 月。其中 2020 年的 6 月、12 月和 2021 年的 1 月的数据分别用来标注 2020 年 5 月、11 月和 12 月互联网卡用户（是离网用户还是正常用户）。然后，本文使用相应的数据样本来进行离网预测模型的训练和测试。由于系统资源的限制，本文随机挑选了 50 万的互联网卡用户作为一

个月的数据样本，其中在 2020 年 5 月有 493251 个正常用户和 6749 个离网用户。此外，对于 2020 年 11 月和 12 月的互联网卡用户而言，则有 871049 个正常用户和 73289 个离网用户。因此，本文在 2020 年 5 月、6 月、11 月和 12 月的 500000 个个 944338 个数据样本上构建了实验，其中 80% 的样本用于训练，剩下 20% 的样本用于测试。

6.1.1 基准模型

为了显示本文提出的 ICCP 模型的优越性，本文设计和实现了一下可以比较的分类基准模型，包括机器学习和深度学习模型。值得注意的是，虽然这些算法在科研文献中被广泛采纳用来解决分类问题，在其中却没有解决互联网卡离网分类和预测的。另外，为了保证比较的公平性，所有的基准模型的特征输入都与 ICCP 保持一致。

- **随机森林 (RF)**：是一种被广泛采用用来分类问题的基于决策树的传统机器学习算法。
- **轻量梯度提升机 (LGBM)**：是一种基于决策树算法的分布式梯度提升机器学习框架。这种机器学习提供了高效率并行训练，低内存消耗，更高准确率和大量数据的快速处理。
- **长短期记忆网络 (LSTM)**：是深度学习领域内的一种循环神经网络架构 (RNN)。它能很好地捕捉序列数据中的短期时间相关性。
- **多层感知机 (MLP)**：是人工神经网络中的一种基础模型。它利用反向传播机制来进行模型训练。多层感知机架构是由包含全连接层，批量归一化 (BN) 层，ReLU 层，Dropout 层的块堆积而成的。
- **残差前馈神经网络 (ResMLP)**：是一种基于残差连接的神经网络架构。其中包括一个隐层前馈神经网络和一个线性的残差连接。和多层感知机相比，残差前馈神经网络是一种拥有三个块层的更简单的残差网络架构。在本文离网预测场景当中，我们在每三个块层中添加了一个残差连接用来解决反向传播过程当中的梯度消失问题。

6.1.2 评估指标

对于离网预测问题来说，基于以下这四个基础测试结果，分别是真阳性样本 (TP)，假阳性样本 (FP)，真阴性样本 (TN)，假阴性样本 (FN)，我们采用了一下 5 个评价指标来评估相应性能。

- **精准率 (Precision)**：指的是一个被预测为离网的用户被正确预测的概率，具体公式为， $\frac{TP}{TP+FP}$ 。

- **召回率 (Recall)**: 指的是在真实标签中所有离网用户被正确预测的概率, 具体公式为, $\frac{TP}{TP+FN}$ 。
- **F1 分数 (F1-Score)**: 被定义为精准率和召回率的调和平均数, 具体公式为, $2 \times \frac{Precision \times Recall}{Precision + Recall}$ 。
- **接收者操作特征曲线下面积 (AUC)**: 被定义为随机挑选一个正样本和负样本, 正样本能排在负样本之前的概率。该评价指标用来判断模型是否擅长分类。具体公式为, $\frac{\sum_{i \in positiveClass} rank_i - \frac{M(M+1)}{2}}{M \times N}$, 其中 M 和 N 分别表示正样本和负样本的数量。
- **精准率-召回率曲线下面积 (PR-AUC)**: 被定义在同一张图中根据不同阈值同时画出的精准率-召回率曲线下的面积。这项评价指标等同于根据不同阈值计算的召回率和精准率的平均乘积累加和。具体公式为, $\sum_{k=1}^N P(k) \Delta R(k)$ 。

6.2 用户离网预测模型性能评估

6.2.1 系统总体性能

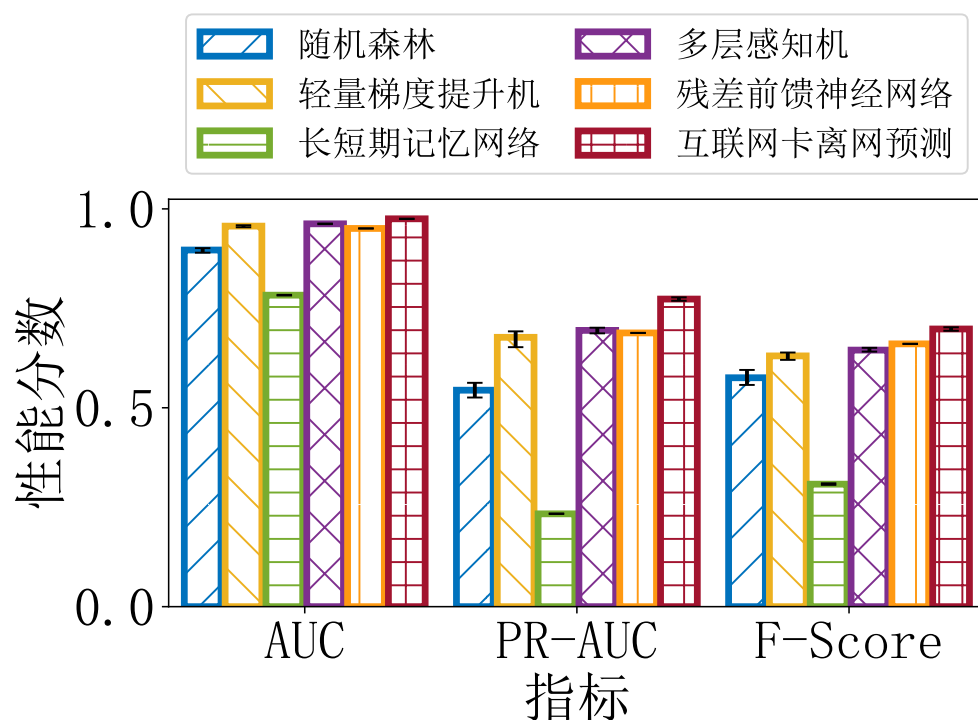


图 6-2 离网预测性能对比

本文首先通过与基线模型的对比来检验本文提出的模型 ICCP 的系统总体性能。在把所有数据平均分成 5 组 (比如, k 折交叉验证中方法 k=5) 来计算性能的

平均值和上下界，图6-2展示了不同分类模型的性能表现。本文可以做出以下推断：首先就 AUC，PR-AUC 和 F1 分数而言，本文提出的 ICCP 模型可以显著超越基线模型。具体来说，随机森林，轻量梯度提升机，长短期记忆网络，多层感知机和残差前馈神经网络的平均 PR-AUC 分数分别为 0.54，0.67，0.23，0.69 和 0.68，在另一方面，ICCP 的平均 PR-AUC 分数为 0.77。其次，与其他基准预测模型相对比而言，长短期记忆网络表现出了最差的性能，这是因为长短期记忆网络只能捕捉时序特征而忽略了静态画像特征。第三，通过比较随机森林、轻量梯度提升机和长短期记忆网络、多层感知机和残差前馈神经网络，本文可以发现深度学习模型展示了提高互联网卡用户离网预测准确率的潜力。这里面的主要原因为深度学习模型可以通过利用多维度特征学习隐含信息表示。需要指出的是，本文也测试了其他机器学习模型，包括决策树，支持向量机和梯度提升树等，最终本文选取了表现性能最好的随机森林和轻量梯度提升机作为机器学习模型的代表。

6.2.2 Top-U 用户的性能.

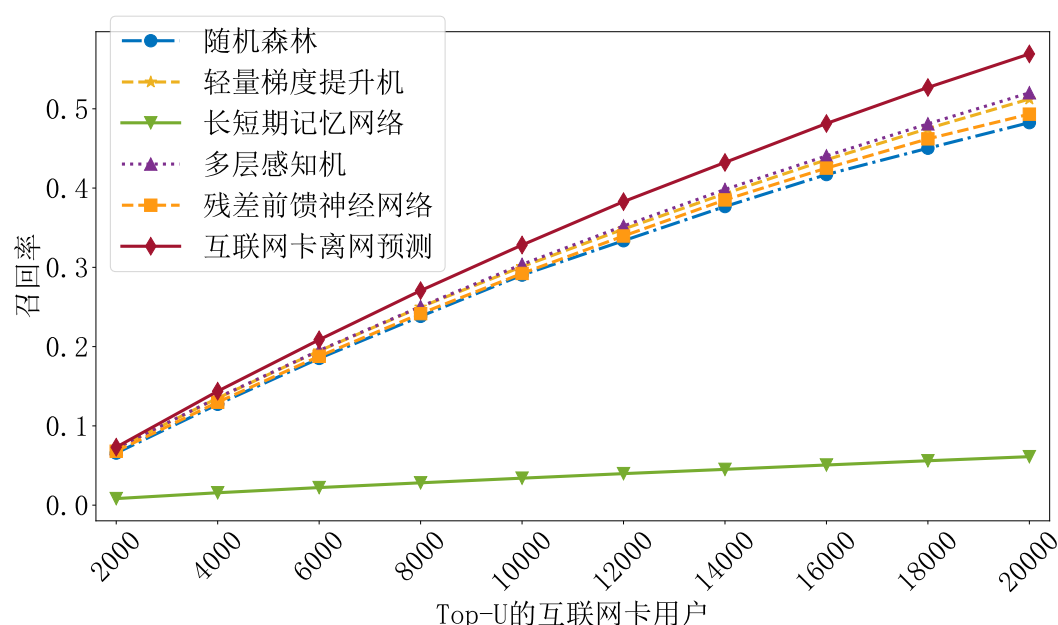


图 6-3 Top-U 用户的召回率对比

为了进一步评估离网预测模型的性能，我们测试了 Top-U 个互联网卡用户的模型表现。需要特别指出的是，整个预测系统将会产生 Top-U 个互联网卡用户的名单，其中是最有可能在下个月离网的用户，接着我们评估了他们的预测性能。这个 Top-U 用户性能的评价指标可以作为运营商做决策的重要参考。图6-4，图6-3和图6-5分别展示了 Top-U 用户的精准率（Precision@U）、召回率（Recall@U）和 F1 分数（F1-Score@U）。其中用户范围是从前 2000 个一直到前 20000 个。本

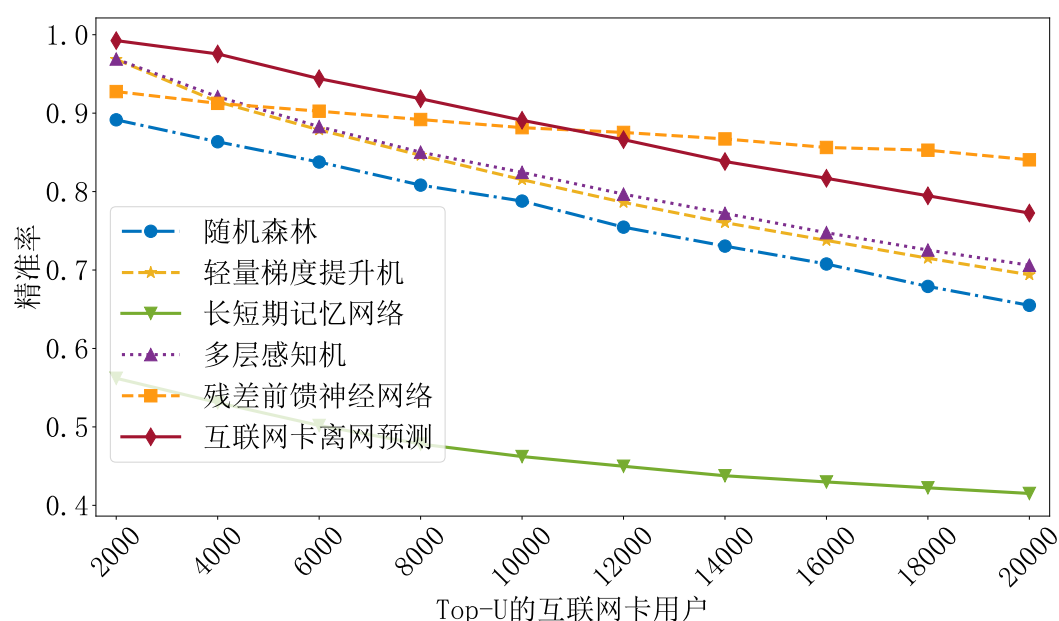


图 6-4 Top-U 用户的精准率对比

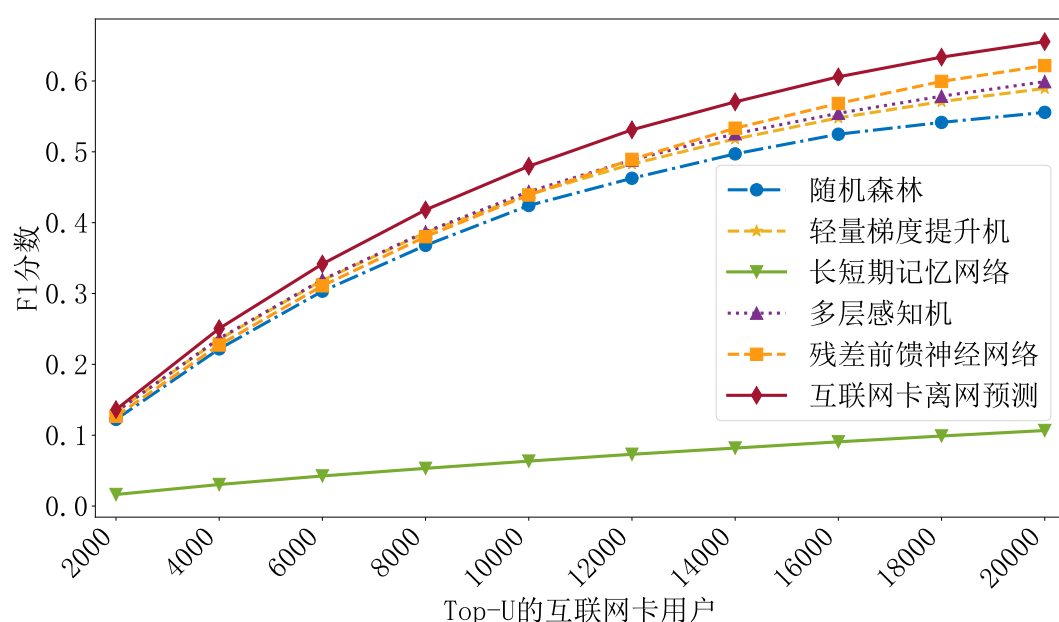


图 6-5 Top-U 用户的 F1 分数对比

文做出了以下的主要观察。

首先，在 Top-U 用户的设置下，我们提出的 ICCP 模型能在 Recall@U 和 F1-Score@U 这两项指标上明显超过其他基准模型。举例来说，ICCP 的 Precision@20000, Recall@20000, F1-Score@20000 分数为 {0.77, 0.60, 0.65}，与此同时残差前馈神经网络，多层感知机，轻量梯度提升机，长短期记忆网络和随机森林的对应分数分别为 {0.84, 0.49, 0.62}，{0.70, 0.52, 0.53}，{0.69, 0.51, 0.59}，{0.65, 0.48, 0.55}

和 $\{0.41, 0.06, 0.11\}$ 。第二，对于 Precision@U 和 Recall@U 这两个评价指标而言，随着用户 U 数量的增加，精准率是逐渐下降的，与此同时，召回率却是逐渐上升的。这是合理的，因为在用户 U 很少的情况下，这些被识别出来的预离网用户往往拥有更高的可能性离网运营商，这导致了精确率较高，但在另一方面，由于用户基数较少，也导致召回率较低。第三，随着用户 U 的增加， Precision@U 的下降趋势和 Recall@U 的增长逐渐放缓的趋势共同说明了随着用户 U 的增加，从中识别出离网用户变得越来越困难。此外，当 Top 用户数量大于 12000 时，残差前馈神经网络的精确率要高于其他方法是因为被添加到残差前馈神经网络中的残差连接结构。这个结构大大改善了在反向传播中的梯度消失问题，同时也使得残差前馈神经网络在预测 Top- U 用户时更加平衡，下降趋势也更加平缓。但是，本文提出的 ICCP 模型在 F1 分数这个综合指标上能超过残差前馈神经网络这个基线模型，是因为后者在召回率分数上增长缓慢。

6.3 参数影响

在保证了预测性能后，我们进一步探讨了用户性别、年龄、App 使用情况、套餐选择等用户属性参数对预测性能的影响。

6.3.1 性别参数的影响

我们将用户样本分为男性和女性两个子集，采用 $k=5$ 的 k 折交叉验证方法分别评估模型性能。图6-6绘制了平均性能度量分数，其中男性和女性用户都有

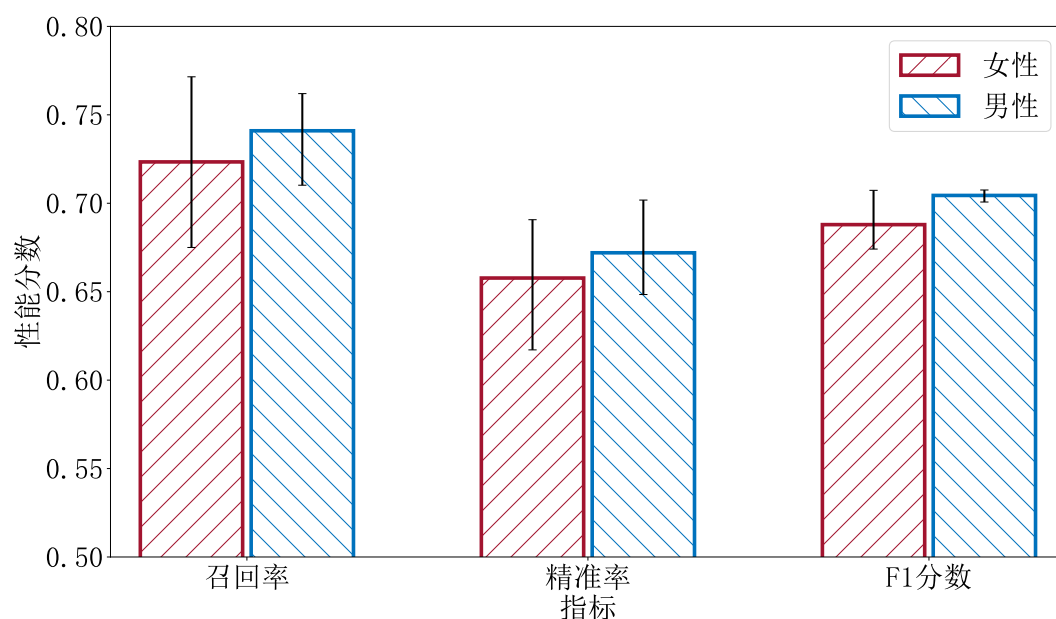


图 6-6 性别参数的影响

错误条用来，显示模型性能的上下界。本文可以得到以下结论：ICCP 对男性和女性均是有效的，他们的指标得分非常相似，并且与系统的整体性能相当。如，对于召回率、精准率和 F1 分数的指标，女性用户可以分别获得 0.72、0.66 和 0.69 的平均分，而男性用户可以分别获得 0.74、0.67 和 0.71 的平均分。另一方面，我们可以发现男性的指标得分略大于女性用户，这与男性用户比女性用户有更多的流量消耗行为是一致的。可以解释为，流量消费行为较多的男性用户离开运营商，更有利于学习模型捕捉潜在的离网行为线索。

6.3.2 年龄参数的影响

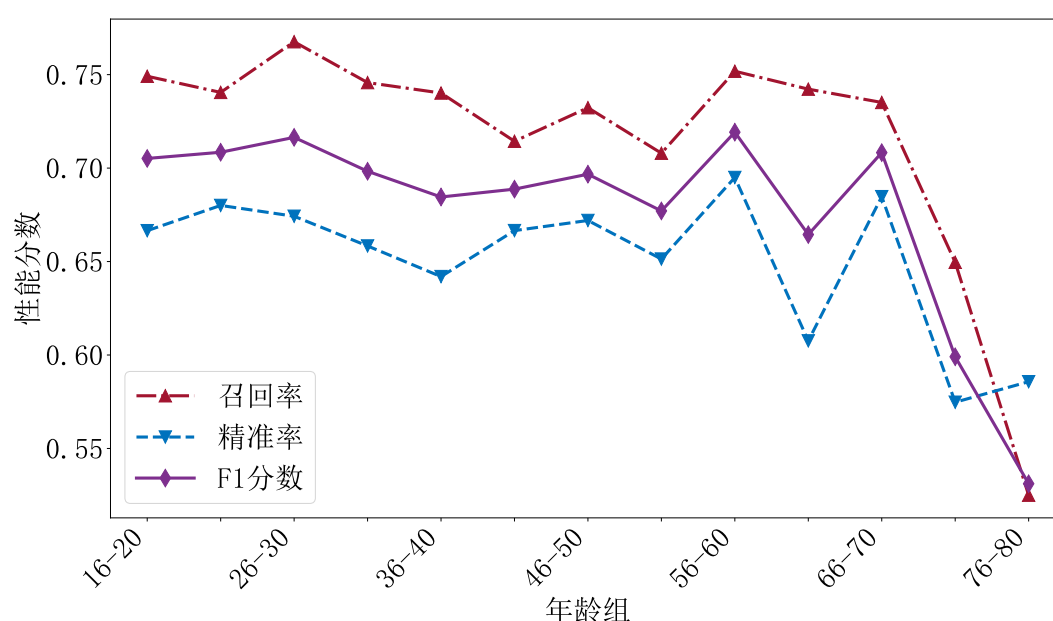


图 6-7 年龄参数的影响

为了研究用户年龄的影响，我们将 IC 用户分为 13 个年龄组，年龄从 16 岁到 80 岁，每个年龄组的范围为 5。图6-7显示了 ICCP 在不同年龄组中获得的召回率、精准率和 F1 分数的平均指标得分。虽然 16 岁至 60 岁用户的分数略有变化，但几乎所有的分数都在 70% 以上，这意味着 ICCP 可以在所有用户上稳健运行，而不受用户年龄影响流量消耗。另外，对于 16 岁至 35 岁的用户，他们可以获得最高的精度分数，这是因为年轻用户在互联网服务中更活跃，他们在离开系统前的异常行为更容易被识别出来。另一方面，对于 61 岁至 80 岁的用户，我们可以看到指标得分有严重的抖动，这是因为老年用户在互联网服务中不太活跃，很难捕捉他们的网络使用行为。

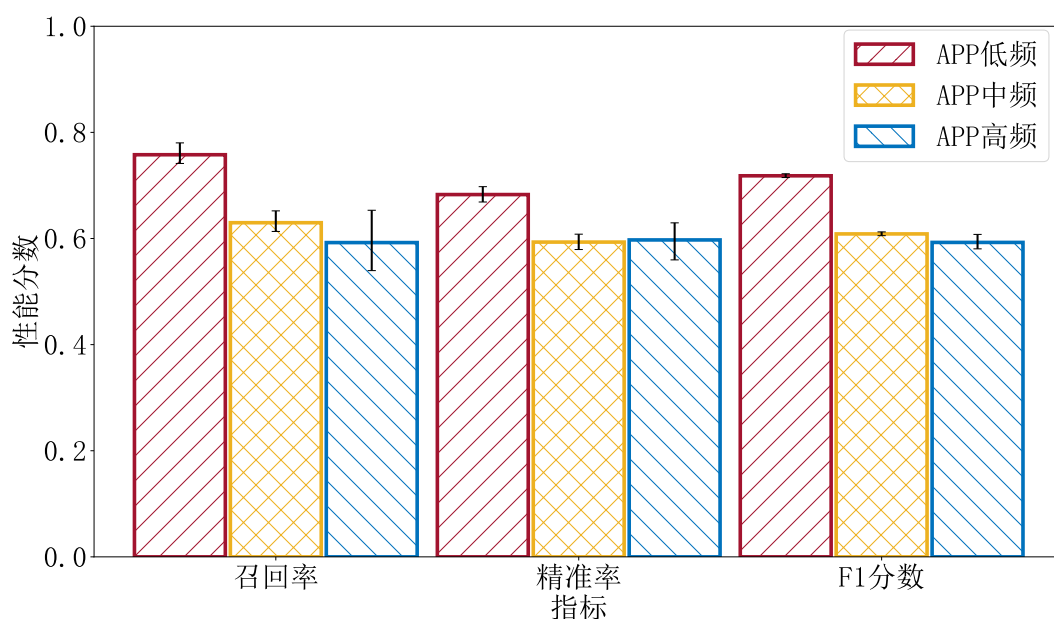


图 6-8 APP 参数的影响

6.3.3 APP 参数的影响

我们通过对用户应用程序使用频率的排序，将用户样本分为三个级别，即：高、中、低。图6-8显示了采用 $k=5$ 的 k 折交叉验证方法对三组 APP 使用水平不同的互联网卡用户的平均度量得分和误差条。我们可以看到，ICCP 在 APP 使用频率中等和较高的用户中都能很好地工作，这两个用户的指标得分非常相似。但是，我们可以观察到，APP 使用频率较低用户的指标得分比其他 App 使用级别的用户要大，这与我们之前的分析是一致的，低 App 使用频率的特征可以有效区分出离网用户。例如，对于召回率、精确度和 F1 分数的指标，APP 使用频率较低用户组可以分别获得 0.76、0.68 和 0.72 的平均分，APP 使用频率中等用户组可以分别获得 0.63、0.59 和 0.61 的平均分，而 APP 使用频率较高用户组用户可以分别获得 0.59、0.58 和 0.59 的平均分。

6.3.4 套餐参数的影响

为了检验用户套餐选择的影响，我们采用 k 折交叉验证方法，用误差条图绘制前 2 个受欢迎套餐（例如，19 元和 39 元）的平均度量分数，如图6-9所示。我们可以观察到，ICCP 对于这两种套餐的用户都可以稳健地工作，其中他们的精准率度量得分非常相似，并且与系统的总体性能相当。此外，我们可以看到 39 元套餐的指标得分（例如，召回率和 F1 分数）比 19 元套餐的用户略大，这是因为根据我们之前的分析，19 元套餐比 39 元套餐有更多的离网用户，从而带来了更

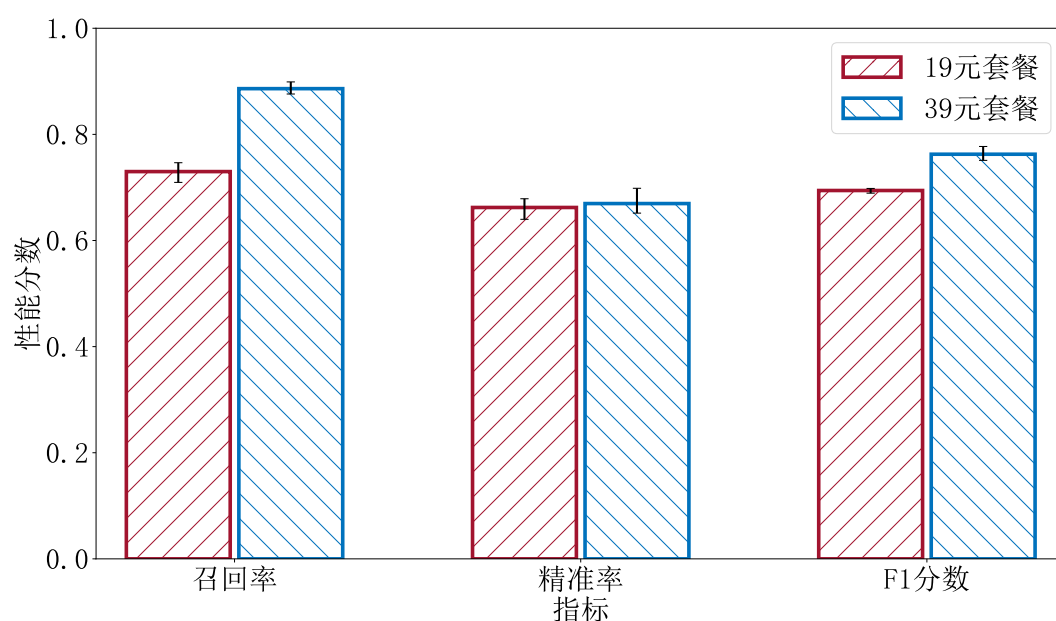


图 6-9 套餐参数的影响

多的预测挑战。

6.4 消融实验

表 6-1 消融实验

	AUC	PR-AUC	F1 分数	召回率	精准率
所有特征	0.9754	0.7749	0.7022	0.7464	0.6630
-目标编码	0.8656	0.4646	0.4548	0.4783	0.4335
-账户余额	0.9441	0.6420	0.5750	0.5795	0.5706
-CDR 序列	0.9727	0.7380	0.6781	0.7378	0.6274
-异常天数	0.9752	0.7742	0.7014	0.7513	0.6577
-开卡月份	0.9745	0.7699	0.6970	0.7392	0.6593
-流量序列	0.9735	0.7624	0.6912	0.7336	0.6535
-年龄	0.9735	0.7624	0.6912	0.7336	0.6535
-活跃熵	0.9439	0.6411	0.5716	0.6355	0.5195
-APP 使用频次	0.9712	0.7321	0.6754	0.7668	0.6035

在本小节中，我们将进行消融实验来验证提取特征的有效性。消融实验结果如表6-1所示。我们可以观察到，对于我们提取的主要特征，它们可以显著地影响预测性能，特别是目标编码和帐户余额这两个特征。特别是，当去除目标编码、帐户余额和活跃熵特征时，精度从 0.6630 下降到 0.4335、0.5706 和 0.5195，性

能分别下降 34.6%、13.9% 和 21.6%。同样地，对于 F1 分数和召回率，分数可以分别从 0.7022 降低到 0.4548、0.5750 和 0.5716，从 0.7464 降低到 0.4783、0.5795 和 0.6355。

6.5 本章小结

第7章 预离网用户偏好生成算法设计

7.1 系统描述

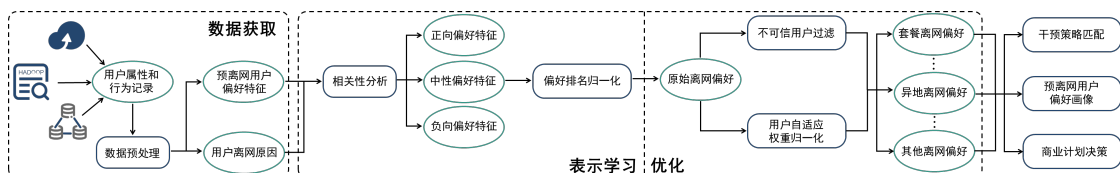


图 7-1 预离网用户偏好生成模块图

7.2 离网原因与偏好的相关性分析

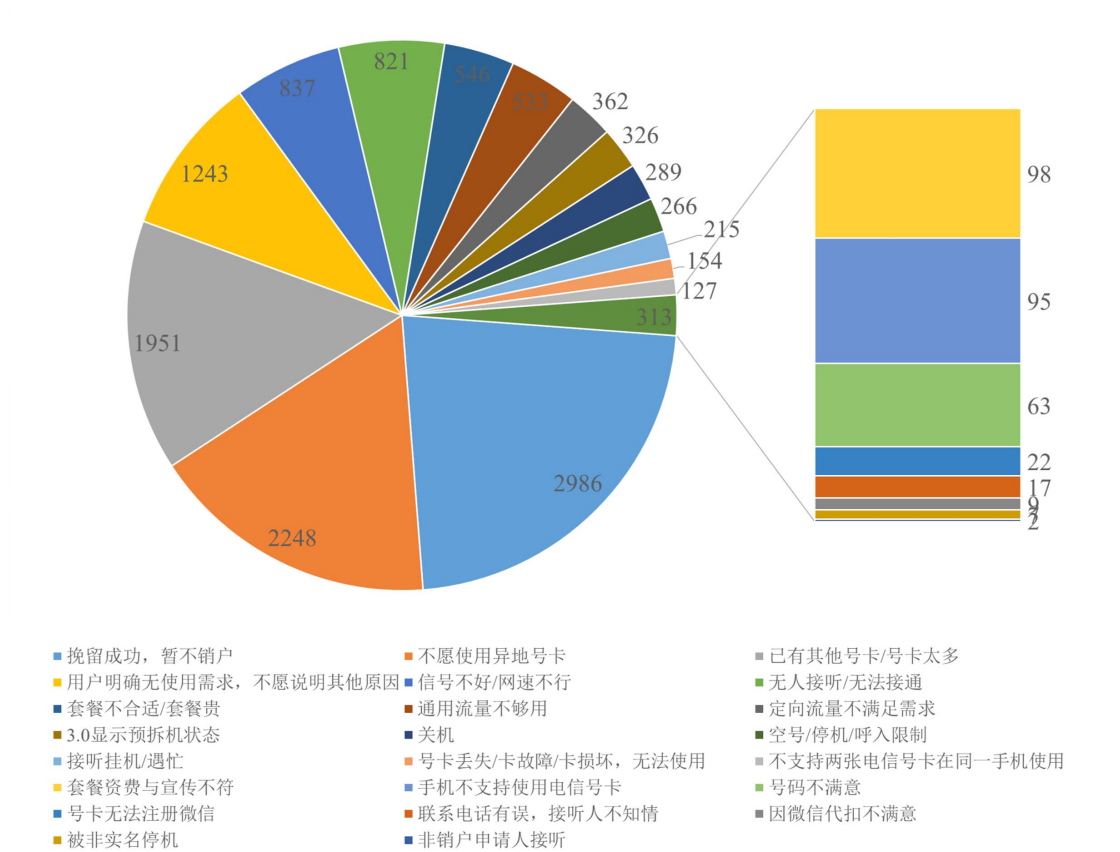


图 7-2 互联网卡用户所有离网原因示意图

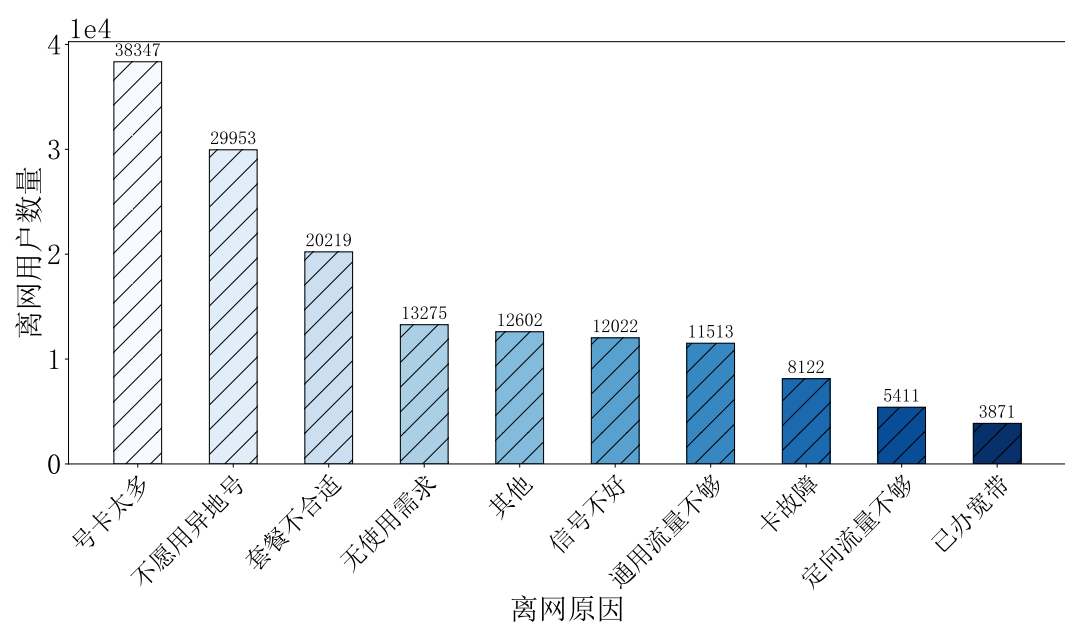


图 7-3 互联网卡用户可建模的离网原因示意图

	套餐不合适	信号不好	异地号卡	通用流量不够	定向流量不够
套餐不合适	1	0.04	0.63	0.71	0.06
信号不好	-	1	0	0	0
异地号卡	-	-	1	0.75	0.05
通用流量不够	-	-	-	1	0
定向流量不够	-	-	-	-	1

7.3 离网偏好排名归一化

离网原因	偏好特征 1	相关性 1	偏好特征 2	相关性 2
套餐不合适	出账金额	正向	账户余额	负向
信号不好	最大网速	负向	平均网速	负向
异地号卡	异地流量记录条数	正向	异地流量消耗值	正向
通用流量不够	通用流量记录条数	正向	通用流量消耗值	正向
定向流量不够	定向流量记录条数	正向	定向流量消耗值	正向

7.4 不可信用户过滤机制

7.5 本章小结

第 8 章 基于汤普森采样的预离网用户干预算法设计

8.1 系统描述与问题建模

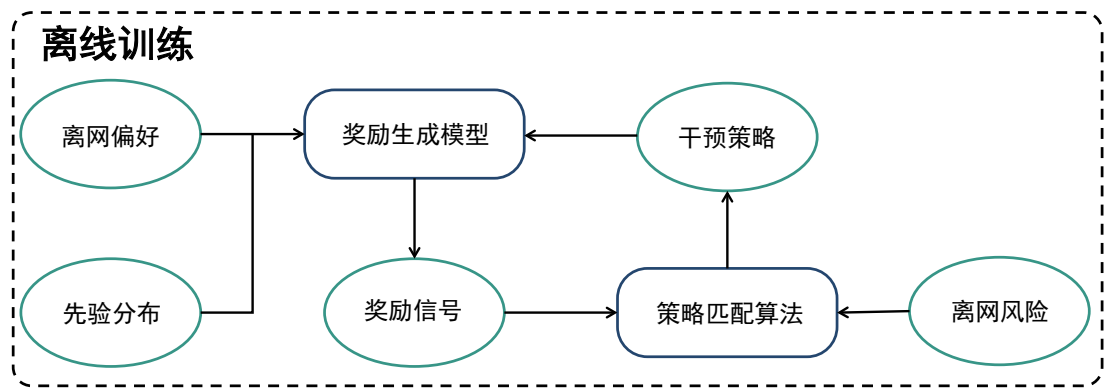


图 8-1 干预策略匹配模块图

8.2 奖励生成模型设计

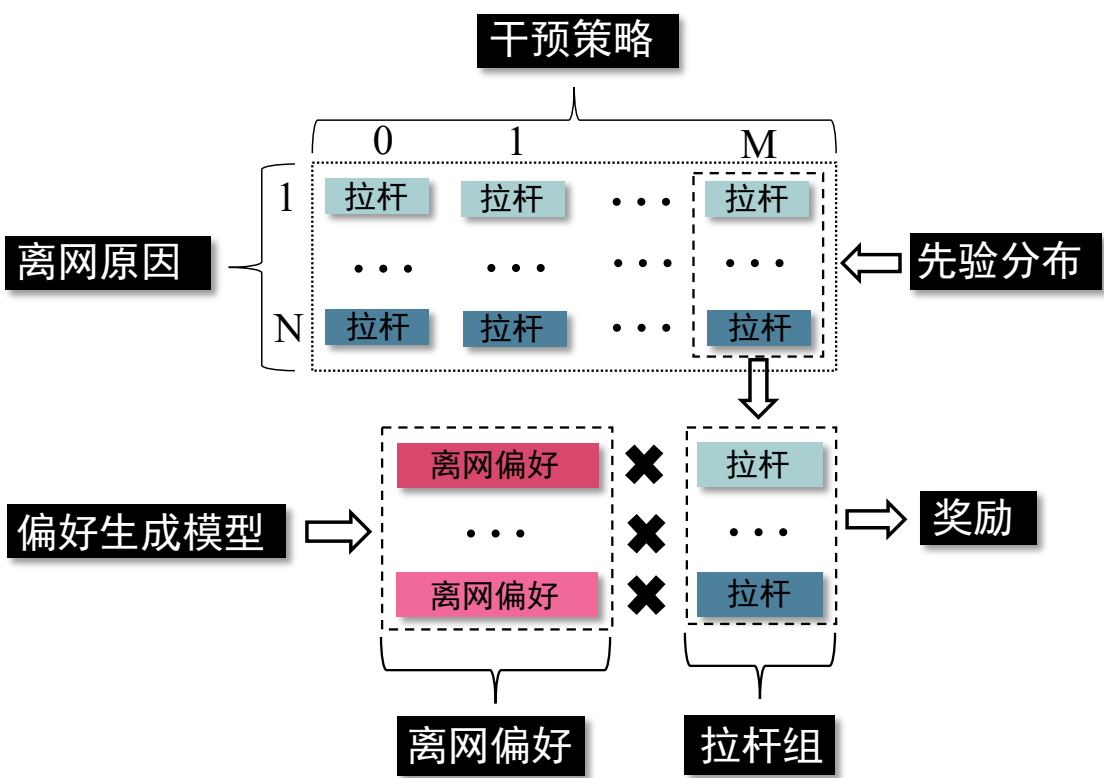


图 8-2 奖励模型内部示意图

8.3 基于汤普森采样的用户-干预措施匹配算法设计

算法 8-1 ICSM: 在资源有限上下文中的投资回报率优先的汤普森采样算法

Input: 预离网用户集 \mathcal{C} , 干预策略集 \mathcal{S} , 成本向量 \vec{E} , 预算 \mathcal{B} , 离网偏好矩阵 Cp , 离网风险向量 \vec{Cr} , 离网指示向量 \vec{Ci} , 生命周期价值向量 LV , 奖励生成模型 RGM

Output: ICSM, 预算 \mathcal{B} , 奖励列表 RL

```

1: 初始化  $S(1) = 0, F(1) = 0$ 
2: for  $c = 1, 2, \dots, \mathcal{C}$  do
3:   if 预算  $\mathcal{B} <$  干预策略最小的成本  $\min(\vec{E})$  then
4:     break
5:   else
6:     初始化最大概率  $MaxProb = -1$ 
7:     初始化选择拉杆号  $SelectedArm = -1$ 
8:     for  $m = 0, 1, \dots, \mathcal{M}$  do
9:       从  $Beta(S_m(t) + 1, F_m(t) + 1)$  分布中采样奖励期望  $r_m$ 
10:      if  $\mathcal{B} \geq \vec{E}_m$  and  $LV_c \geq \vec{E}_m$  and  $r_m / \vec{E}_m > MaxProb$  then
11:        最大概率  $MaxProb = r_m / \vec{E}_m$ 
12:        选择拉杆号  $SelectedArm = m$ 
13:      end if
14:    end for
15:    奖励  $r_c = RGM(Cp_c, SelectedArm)$ 
16:     $RL$  在尾部追加  $(r_c * Ci_c)$ 
17:     $S_{SelectedArm}(t + 1) = S_{SelectedArm}(t) + r_c * Cr_c$ 
18:     $F_{SelectedArm}(t + 1) = F_{SelectedArm}(t) + 1 - r_c * Cr_c$ 
19:    预算  $\mathcal{B} - = \vec{E}_{SelectedArm}$ 
20:  end if
21: end for

```

8.3.1 动作空间

8.3.2 奖励机制设计

8.4 基于模拟干预结果机制的训练

8.5 本章小结

第9章 实验评估与结果分析

9.1 实验设置

9.1.1 对比方案

9.1.2 评估指标

9.2 预离网用户干预框架性能评估

9.2.1 总体性能

匹配算法性能.

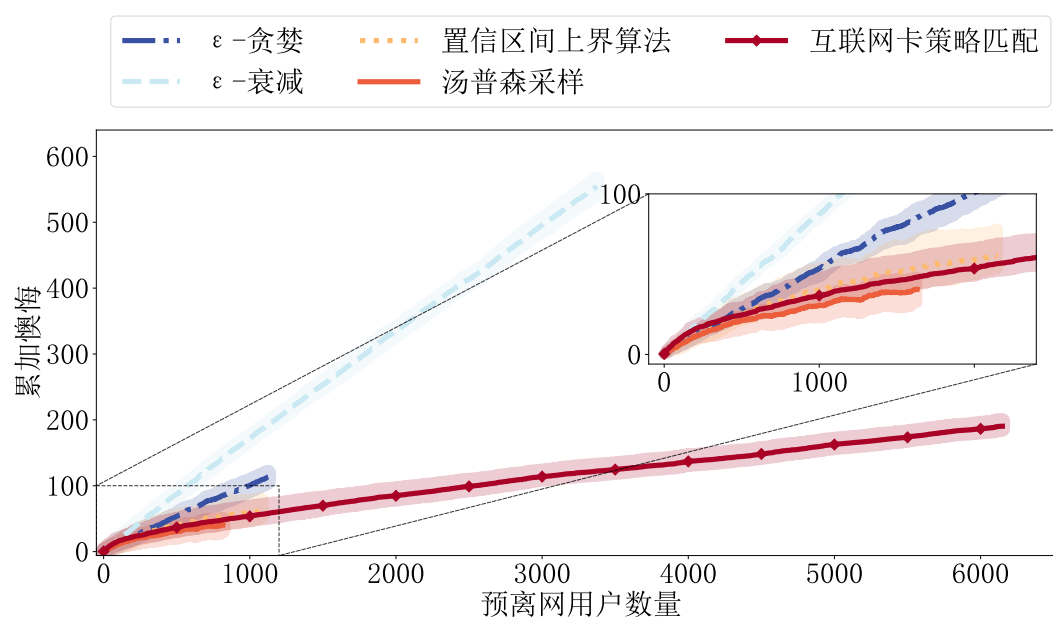


图 9-1 匹配算法性能对比图

干预框架总体性能.

商业指标.

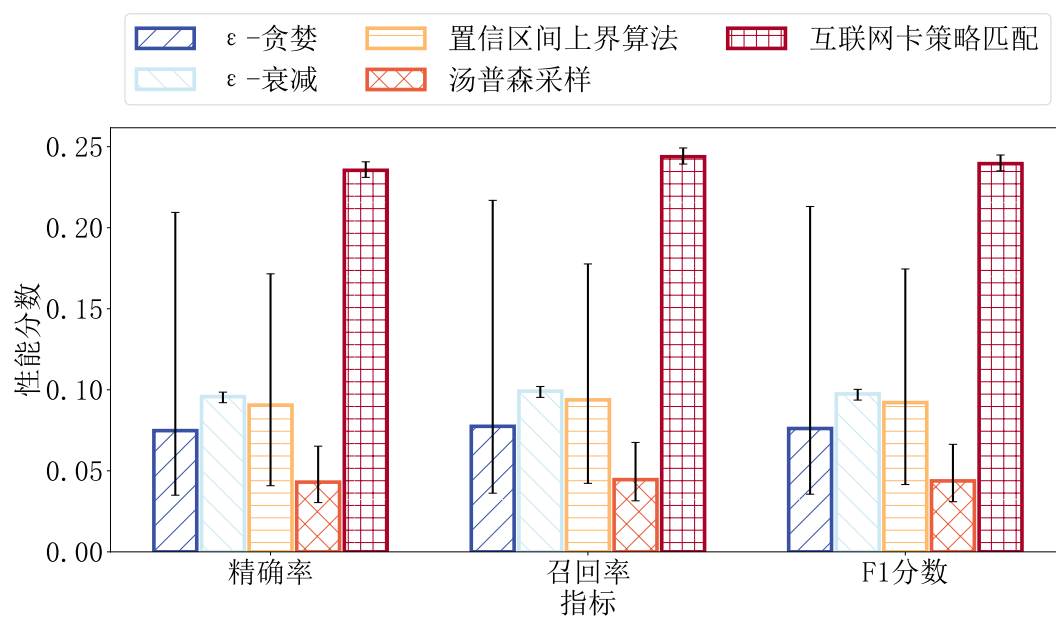


图 9-2 干预框架总体性能对比图

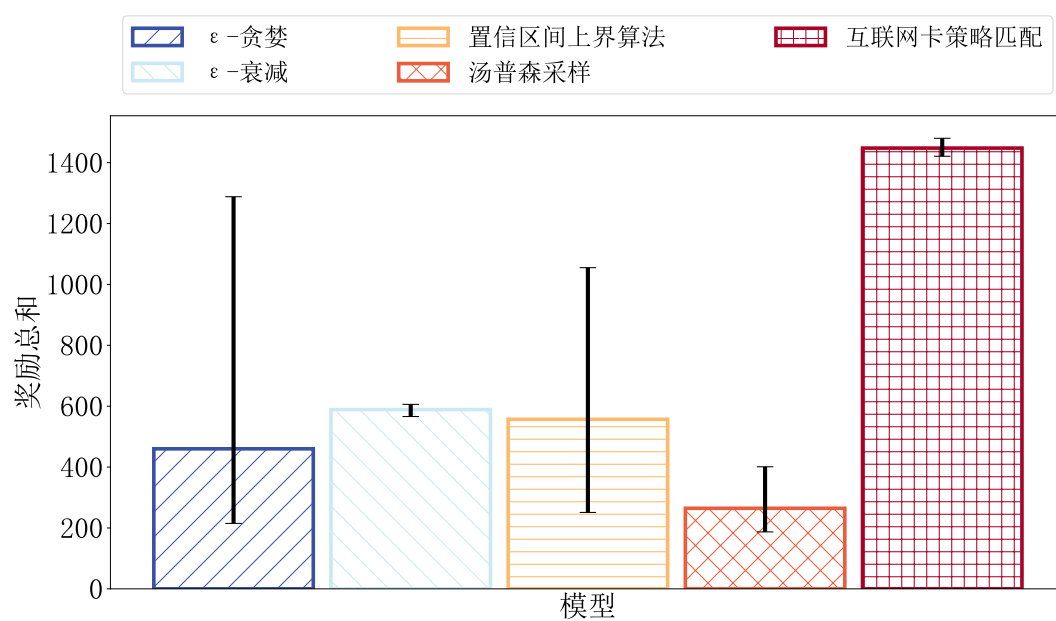


图 9-3 奖励总和对比图

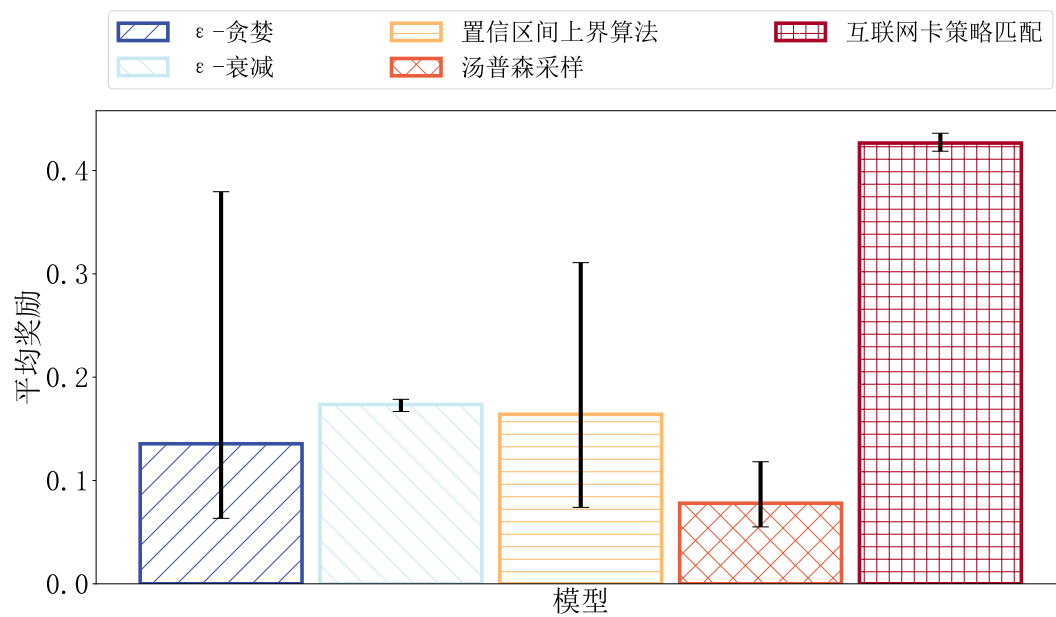


图 9-4 平均奖励对比图

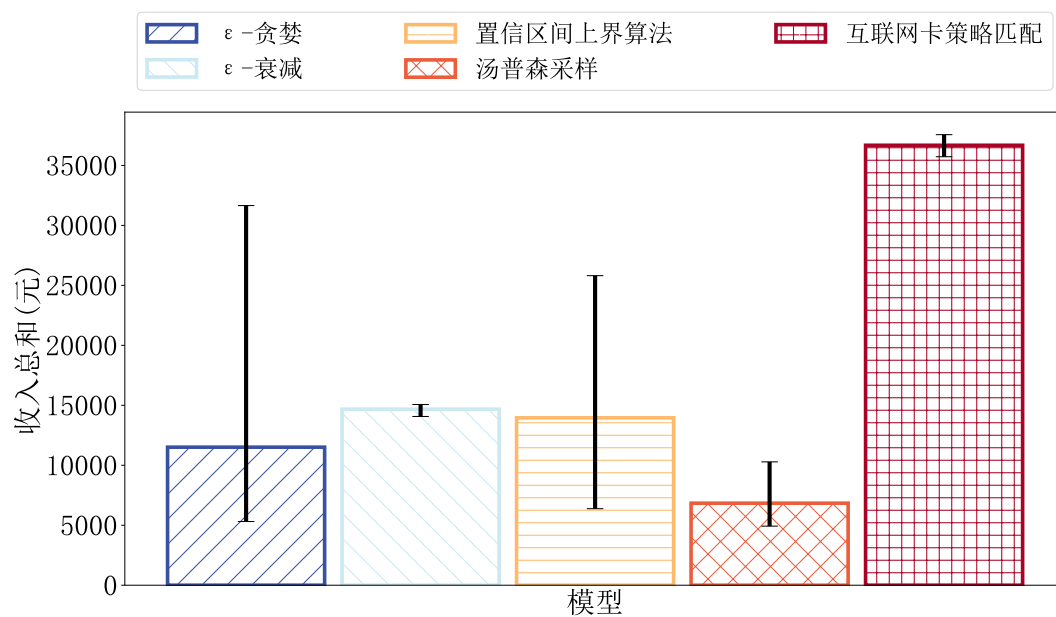


图 9-5 收入总和对对比图

9.2.2 健壮性测试

9.3 参数影响

9.3.1 城市

9.3.2 年龄

9.3.3 离网风险

9.4 本章小结

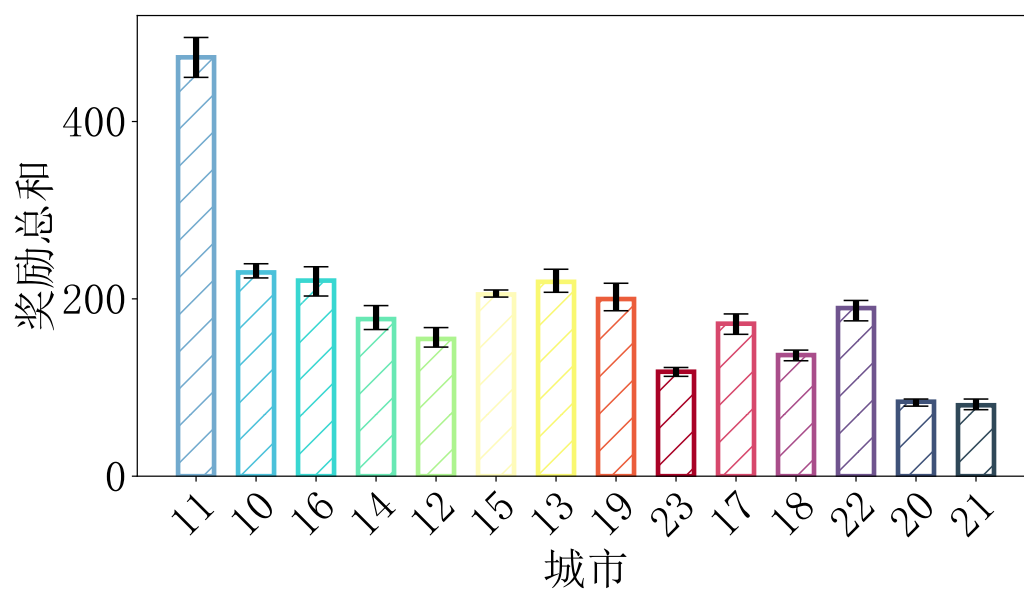


图 9-6 城市参数对于奖励总和影响的对比图

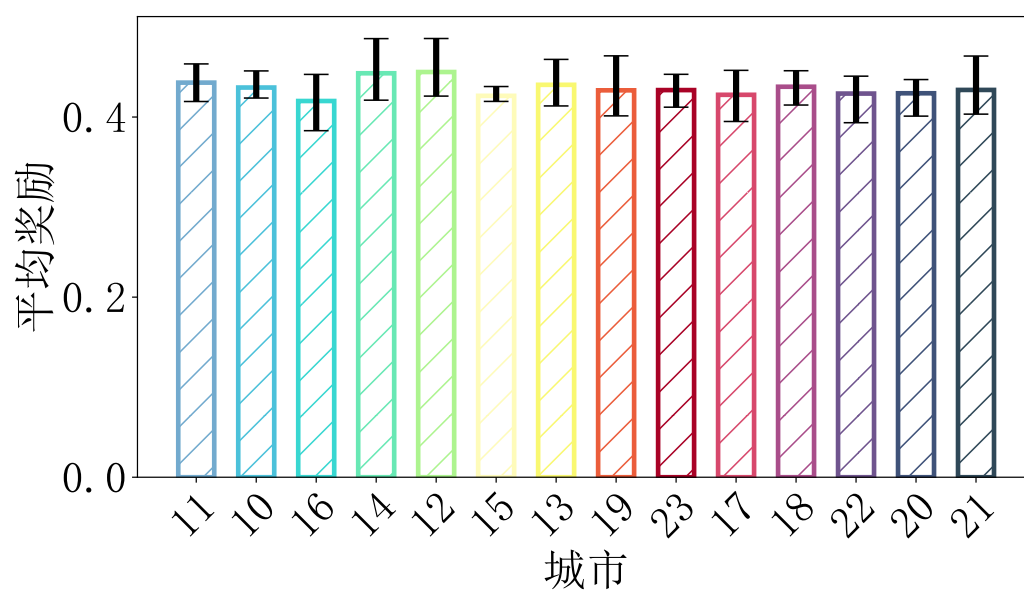


图 9-7 城市参数对于平均奖励影响的对比图

第 10 章 总结与展望

10.1 工作总结

10.2 未来工作展望

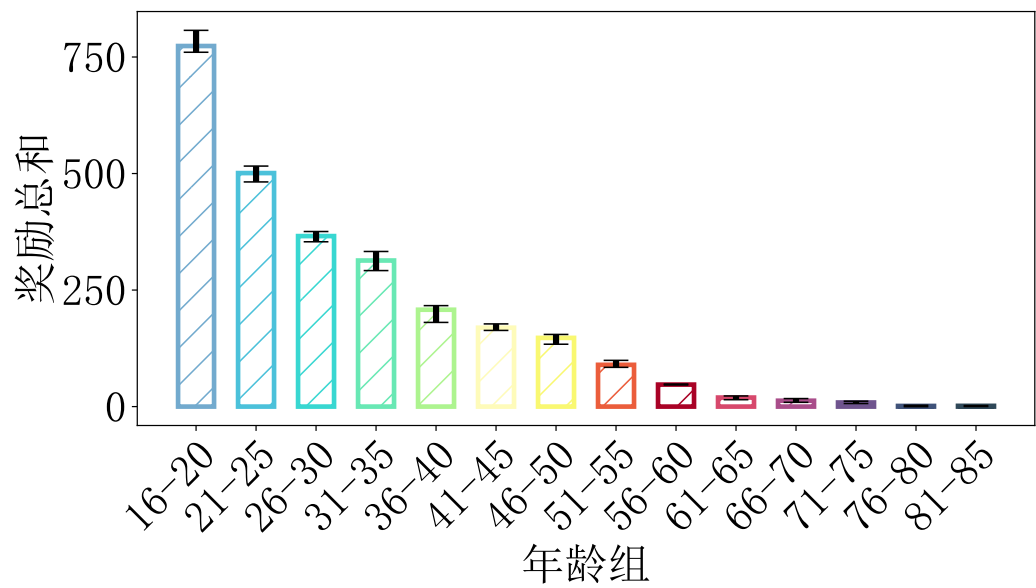


图 9-8 年龄参数对于奖励总和影响的对比图

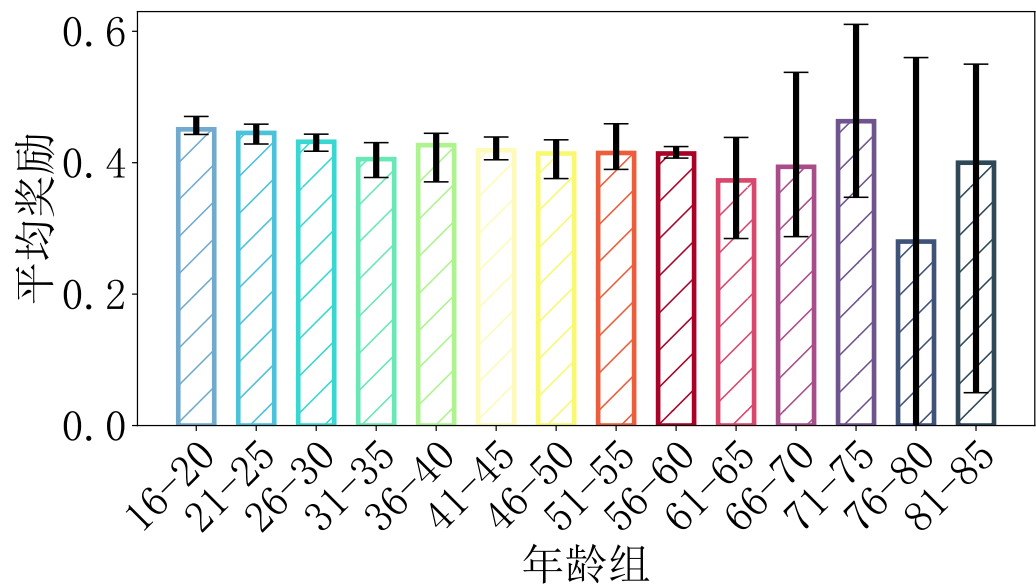


图 9-9 年龄参数对于平均奖励影响的对比图

第 11 章 绪论

11.1 研究背景与意义

目的是创建一个符合中南大学研究生学位论文（博士）撰写规范的 LaTeX 模板，解决学位论文撰写时格式调整的痛点。

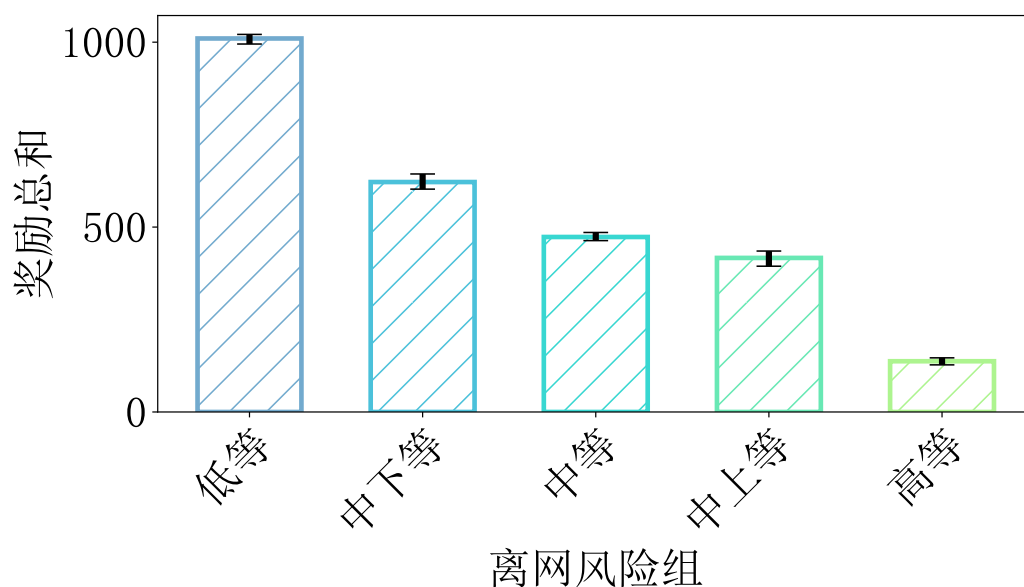


图 9-10 离网风险参数对于奖励总和影响的对比图

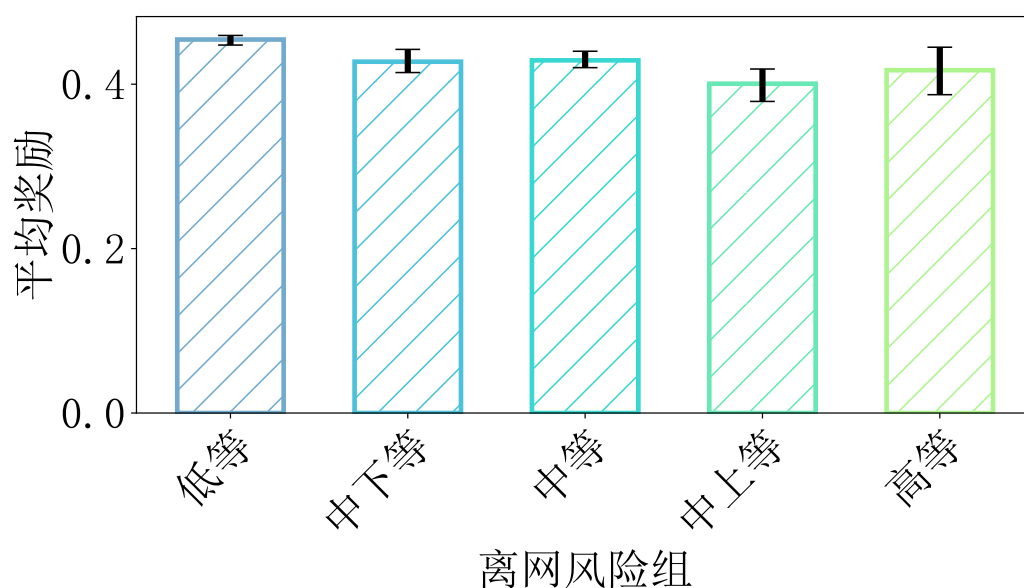


图 9-11 离网风险参数对于平均奖励影响的对比图

已有珠玉在前，本文之所以还要重新造轮子，主要依据 2022 年 4 月 18 号学校下发的 [《中南大学研究生学位论文撰写规范》中大研字【2022】8 号] (<http://oa.its.csu.edu.cn/Home/ReleaseMainText/9CFE8926B13143009D5EB424333AAD6C>), 重新修改了页面布局、字体类型和大小、标题内容，以期做到与 Word 模板尽可能的相似。学校要求：2022 年起，申请博士、硕士学位的学位论文必须按新文件执行。主要修改如下：

- 按要求修订段落与各级标题间距；
- 按要求修订中英文段落间距，章节间距，附录标题段落间距，研究成果及致谢标题间距，参考文献间距等；
- 增加博士和硕士论文模板选项，只需要 `info.tex` 选择即可，方便使用；
- 按新版撰写规范修改主要格式如下：修订目录章节标题间距；修订中英文段落间距；修订图片与表格标题的段落间距；
- 按要求更新“学位论文版权使用授权书”；
- 依据 2022 最新撰写规范修订封面和扉页-“封面”及“扉页”关于学科专业的表头更新为：一级学科/专业学位类别，二级学科/专业领域；
- 依据专家意见修订定理和证明等环境，如“定理”使用小四黑体，编号随章节变化重新编号（如定理 4-1），定理内容使用小四宋体，且内容行距与正文一致；“证明”无需编号，且以黑色小方块结尾；
- 修订算法在每个章节重新编号问题；
- 增加符号说明页和附录页（如果不需要，请在 `.cls` 文件对应处注释掉即可）；
- 增加参考文献按国标 `gbt7714-2015` 要求，只核对了常用的图书、中英文期刊，会议格式，其余未常使用的未进行核对（如有问题请改回 `gbt7714-2005`）；
- 修订多个子图 `Caption` 居中问题；
- 依据专家意见调整成果与致谢部分间距，并增加目录中的点密度；
- 按照图书馆最新要求（2020 年 12 月份），去除目录中红色边框；
- 增加页眉信息：中南大学博士论文与右侧的章节名保持一致，以及无需号章节名保持一致；
- 增加中英文摘要至目录，并保持与章节名对其；
- 参考文献完全依照国标 `gbt7714-2005`，修正了部分 Bug，提供了新的引用命令；
- 按照最新版本要求，在声明扉页前后各增加一页空白页，保证装订单独成页；
- 章节标题居中，并改成‘第 1 章’样式；
- 目录中，将原章节标题换成‘第几章’样式，字体按要求加粗；
- 中文摘要到目录结束用罗马数字编写页码，小五号 Times New Roman，居中；
- 增加插图索引和表格索引；
- 所有的章节题目和中英文摘要均按要求修改字体和间距；

11.2 主要研究工作

博士和硕士模板选择说明：

- 当前模板默认是博士，学术型。

- 如选择硕士模板, 只需要将对应的 `content/info.tex` 文件中, 选择 `\Doctor false` % 硕士学位论文, 注释掉对应的博士模板就行。
- 学术型和专业型, 盲审和正常版本, 公开和涉密版本, 均是同样操作;
- 其它模板, 可以根据自己需要修改 `CSUthesis.cls` 文件。

- (1) 提供图片插入示例。
- (2) 提供表格插入示例。
- (3) 提供公式插入示例。
- (4) 提供参考文献插入示例。

11.3 论文组织结构

全文内容共六章, 具体内容组织如下:

第一章为绪论。

第二章为图片插入示例。

第三章为表格插入示例。

第四章为公式插入示例。

第五章为参考文献插入示例。

第六章总结与展望, 总结了本文的主要工作, 展望了下一阶段的研究方向。

第 12 章 图像布局

12.1 单图布局

单图布局如图12-1所示。



图 12-1 单图布局示例

12.2 横排布局

横排布局如图12-2所示。

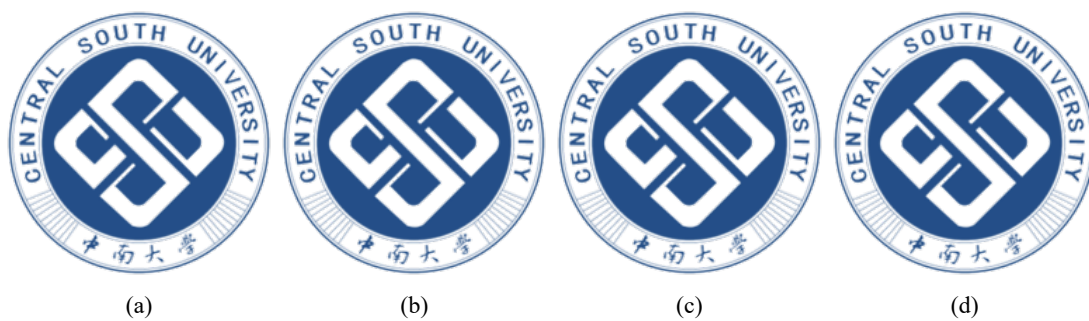


图 12-2 横排布局示例

12.3 竖排布局

竖排布局如图12-3所示。

12.3.1 竖排多图横排布局

竖排多图横排布局如图12-4所示。注意看 (a)、(b) 编号与图关系。



(a)



(b)

图 12-3 竖排布局示例



(a)



(b)

图 12-4 竖排多图横排布局

12.3.2 横排多图竖排布局



(a)



(b)

图 12-5 横排多图竖排布局，斜体 *emph A*, *A*, 斜体 *textit A*

横排多图竖排布局如图12-5所示。注意看 (a)、(b) 编号与图关系。

12.4 本章小结

本章示例图片布局。

第 13 章 表格插入示例

表 13-1 表格为三线表斜体 *emph A*, *A*, 斜体 *text A*

	<i>AA A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
1	212	414	4	23	fgw
2	212	414	v	23	fgw
3	212	414	vfwe	23	嗯
4	212	414	4fwe	23	嗯
5	af2	4vx	4	23	fgw
6	af2	4vx	4	23	fgw
7	212	414	4	23	fgw

表格如表13-1所示，**latex** 表格技巧很多，这里不再详细介绍。

第 14 章 算法示例

算法 14-1 Fourier-Mellin Based KCF

Input: Image I

preprocessed kernelized template T_κ

Output: scale σ , angle θ relation between I and T

- 1: fourier transform: $F = \mathcal{F}(I)$
 - 2: high pass filter: $F_h = \mathcal{H}(F)$
 $\mathcal{H}(x, y) = (1.0 - \cos(\pi x)\cos(\pi y))(2.0 - \cos(\pi x)\cos(\pi y))$
 - 3: log-polar transform: $F_{lp} = \mathcal{L}(F_h)$
 - 4: apply kernel function: $F_\kappa = \mathcal{K}(F_{lp})$
 - 5: phase correlation: $(\Delta x, \Delta y) = \mathcal{C}(F_\kappa, T_\kappa)$
 - 6: resolove scale and rotation:
 $\theta = \alpha \Delta x, \sigma = \log(\Delta y)$
 where α is translation factor of pixel translation on fourier domain and polar angle on origin image
-

算法 14-2 算法示例

Input: 相关输入。。。。

Output: 相关输出。。。

- 1: 算法描述
 - 2: **for** $i \leftarrow 1 \cdots N$ **do**
 - 3: 算法描述
 - 4: **for each** $j \leftarrow 1 \cdots K$ **do**
 - 5: 算法描述
 - 6: **end for**
 - 7: **end for**
 - 8: **repeat**
 - 9: **repeat**
 - 10: 令 $\tau \leftarrow \tau + 1$
 - 11: **until** 内循环迭代终止条件
 - 12: 。。。
 - 13: **until** 外循环迭代终止条件
-

如算法14-1所示，latex 算法技巧很多。按需调整，这里不再详细介绍。

第 15 章 公式、定理、证明插入示例

$$\text{P1: } \min_{\eta, R_u > 0, R_d > 0} \{T_{\text{latency}}(\eta, R_u, R_d)\} \quad (15-1)$$

$$\text{s.t. } 0 \leq \eta \leq 1 \quad (15-2)$$

公式插入示例如公式 (15-3) 所示。

$$\gamma_x = \begin{cases} 0, & \text{if } |x| \leq \delta \\ x, & \text{otherwise} \end{cases} \quad (15-3)$$

$$\text{P1: } \max_{\substack{P_{m,i}, P_{n,i} \\ q_{m,i}, q_{n,i} \\ \forall n, m, i}} [R_{\text{sum}}(P_{m,i}, P_{n,i}, q_{m,i}, q_{n,i}, \forall n, m, i)], \quad (15-4)$$

$$\text{s.t. } q_{m,i} \in (0, 1), q_{n,i} \in (0, 1), \forall n, m, i, \quad (15-5)$$

$$0 \leq \sum_{i=1}^R q_{n,i} P_{n,i} \leq P_n^{\text{sum}}, \forall n, \quad (15-6)$$

$$0 \leq \sum_{i=1}^R q_{m,i} P_{m,i} \leq P_m^{\text{sum}}, \forall m, \quad (15-7)$$

$$\sum_{i=1}^R q_{n,i} \leq 1, \sum_{i=1}^R q_{m,i} \leq 1, \forall m, n, \quad (15-8)$$

$$\sum_{n=1}^N q_{n,i} \leq 1, \sum_{m=1}^M q_{m,i} \leq 1, \forall i, \quad (15-9)$$

$$C_{m,BS,i}(P_{m,i}, q_{m,i}) \geq \varepsilon_{m,i}, \forall m, \quad (15-10)$$

公式子编号示例：

$$\varphi_{n,t} \in \{0, 1\}, \forall n \in \mathcal{N}, t \in \mathcal{N}_t, \quad (15-11-a)$$

$$\varphi_{n,t} \in \{0, 1\}, \forall n \in \mathcal{N}, t \in \mathcal{N}_t, \quad (15-11-b)$$

$$\varphi_{n,t} \in \{0, 1\}, \forall n \in \mathcal{N}, t \in \mathcal{N}_t, \quad (15-11-c)$$

其中，公式15-11-a表示。


$$H_j = \text{Concat}(\text{GAP}(F_j), \text{GMP}(F_j)), \quad (15-12)$$

$$\tilde{H}_{j-1,j} = \text{Concat}(H_{j-1}, H_j), j = 5, \quad (15-13)$$

$$p_c^{(i)} = \text{Softmax}(\mathbf{P}_\theta(\tilde{H}_{j-1,j})), \quad (15-14)$$

定理和证明环境说明：如”定理”使用小四黑体，编号随章节变化重新编号（如定理 4-1），定理内容使用小四宋体，且内容行距与正文一致；”证明”无需编号，且以黑色小方块结尾。

定理 15-1 开始定理。。。

证明 开始证明。。。 

第 16 章 参考文献插入示例

LaTeX[1] 插入参考文献最方便的方式是使用 bibliography[2]，大多数出版商的论文页面都会有导出 bib 格式参考文献的链接，建议使用 Jabref 管理参考文献，把每个文献的 bib 放入 “thesis-references”，然后用 bibkey 即可插入参考文献。

中文文献 [3]，注意手动编辑 bibkey 为英文的即可。

可以将文献标注为右上角^[4]，只需要在现有的 cite 后加 “ss” 即可。

英文会议 [5], [6].

英文期刊 [7], [8].

特别强调：从 Google 下载的 bib 也不一定全是对的，如发现有信息缺失，请下载原文核对。比如已发表的期刊，要包保证年、卷、标。

注意：如发现替换后的参考文献没有更新，请删除主文件夹下 xxx.bbl 文件，重新编译即可。

第 17 章 总结与展望

纯数字编号

1. XXXXXXXXXXXX
2. XXXXXXXXXXXX
3. XXXXXXXXXXXX

罗马编号

- (i) XXXXXXXXXXXX
- (ii) XXXXXXXXXXXX
- (iii) XXXXXXXXXXXX

括号编号

- (1) XXXXXXXXXXXX
- (2) XXXXXXXXXXXX
- (3) XXXXXXXXXXXX

半括号编号

- 1) XXXXXXXXXXXX
- 2) XXXXXXXXXXXX
- 3) XXXXXXXXXXXX

小字母编号

- a) XXXXXXXXXXXX
- b) XXXXXXXXXXXX
- c) XXXXXXXXXXXX

引用测试, 正如1、(i)、(1)、1)、a)所示

17.1 工作展望

手动编号

本课题针对 XX, 鉴于 XXX, 对 XX 进行了提高, 但是 XXX, 所以有如下 XX:

- (1) 目前 XX 虽然 XX, 但是 XX 仍然 XX, 所以 XX 仍然是一个值得 XX 的问题。
- (2) 随着 XX, XX 具有 XX 的问题, 仍值得进一步 XX。
- (3) 本课题在 XX 有了 XX, 但是 XX 的 XX 还存在 XX, 所以 XX。

参考文献

- [1] Lamport L. Latex: a document preparation system: user's guide and reference manual [M]. Addison-wesley, 1994.
- [2] Pritchard A, et al. Statistical bibliography or bibliometrics [J]. Journal of documentation, 1969, 25(4): 348-349.
- [3] 施巍松, 刘芳, 孙辉, 等. 边缘计算 [M]. 北京: 科学出版社, 2018.
- [4] 施巍松, 张星洲, 王一帆, 等. 边缘计算: 现状与展望 [J]. 计算机研究与发展, 2019, 56(01): 69-89.
- [5] Krauß V, Boden A, Oppermann L, et al. Current Practices, Challenges, and Design Implications for Collaborative AR/VR Application Development [C]//Proc. ACM CHI, Virtual Event/Yokohama, Japan. ACM, 2021: 454:1-454:15.
- [6] Wu F, Yang W, Lu J, et al. RLSS: A Reinforcement Learning Scheme for HD Map Data Source Selection in Vehicular NDN [J]. IEEE IoT J., 2021: 1-14.
- [7] Luo C, Zeng J, Yuan M, et al. Telco User Activity Level Prediction with Massive Mobile Broadband Data [J]. ACM Trans. Intell. Syst. Technol., 2016, 7(4): 63:1-63:30.
- [8] Wu F, Ren J, Lyu F, et al. Boosting Internet Card Cellular Business via User Portraits: A Case of Churn Prediction [C]//Proc. IEEE INFOCOM. 2022: 1-10.

附录 A （附录名称）（三号黑体，加粗）（必要时）

附录正文……（格式参考正文）。换行示例。

攻读学位期间主要的研究成果

一、学术论文

[1] **Daxia Mou**, Director, Someone. CSU Latex Template[J]. CSU player: 1(1):1-10. **(SCI 检索, JCR 1 区)**

[2] Director, **Daxia Mou**, Someone, Someother. XXXXXX[J]. Transactions on Image Processing. **(SCI Under Review, JCR 1 区)**

[3] Director, **Daxia Mou**, Someone, Someother. XXXXXX[J]. Transactions on Circuits and Systems for Video Technology. **(SCI Under Review, JCR 1 区)**

二、发明专利

[1] 某大侠, XXX, XXX. 一种用 Latex 写中南大学学位论文的方法. 申请号: CN20190415xxxx, 公开号: CNXXXXXXXXXXA

三、主持和参与的科研项目

[1] 国家自然科学基金面上项目《XXXXXXXXXXXXXXXXX》, 项目编号: XXXXXXXXX, 参与.

四、个人获奖情况

[1] XX 金奖

[2] XX 奖学金

致 谢

作者对给予指导、各类资助和协完成研究工以及提供种论文有作者对给予指导、各类资助和协完成研究工以及提供种论文有利条件的单位及个人表示感谢。

致谢应实事求是，切忌浮夸与庸俗之词。