扫雷反诈

基于机器学习实现涉诈网址自动分类识别

项目建模思路

目录

1. 问题定义

1.1 痛点问题分析

1.1.1 社会问题分析

随着互联网的快速发展，信息爆炸式传递，网络诈骗凭借其迅速扩散、隐蔽性强的特点，已经成为了一个全球性的问题。识别诈骗网址的重要性在当今互联网时代尤为重要。识别恶意网址的必要性体现在：

- 防止信息泄露：识别并避免访问这些网址可以帮助用户保护自己的隐私和敏感信息，防止其被不法分子盗取或滥用；
- 防范网络攻击：恶意网址还可能包含病毒、木马、间谍软件等恶意程序，会对用户设备和系统造成严重威胁。通过识别这些网址，可以避免下到恶意软件和受到其他网络攻击的威胁；
- 提高安全性：通过识别恶意网址，用户可以提高使用互联网的安全性，减少受到网络陷阱和骗局的风险。这有助于促进社会的网络文明和健康发展。

1.1.2 技术问题分析

通过识别恶意网址保障用户上网安全是企业尤其是几大通信公司的当务之急，然而，目前传统的恶意网址系统存在以下问题：

1. 传统的机器学习技术，采用 SVM、线性回归、决策树等方式对恶意网址进行分类，模型的训练效果与训练加入的特征有很大关系，而**特征提取和实验需要耗费大量时间**。

2. 传统的网址识别是基于网址的黑白名单特征，无法有效处理新生成 URL。通常词向量采用 one-hot 编码，每个单词的编码结果是独立的，不同单词间不存在相关性，导致泛化能力不强，尤其对**单词拼接类**恶意域名检测效果不佳，**缺少有效处理高质量涉诈 URL 的能力**。

3. 机器学习的最终效果依赖于高质量数据集的获取，然而实际训练场景存在样本不平衡及标签不可靠的问题，**缺少对不可靠数据集的应对方案**。

1.1.3 数据问题分析

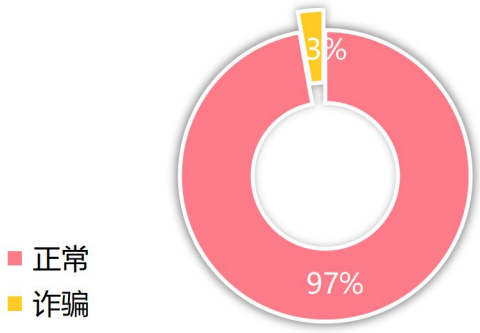

图 1-1 正常网址与诈骗网址数据饼状图

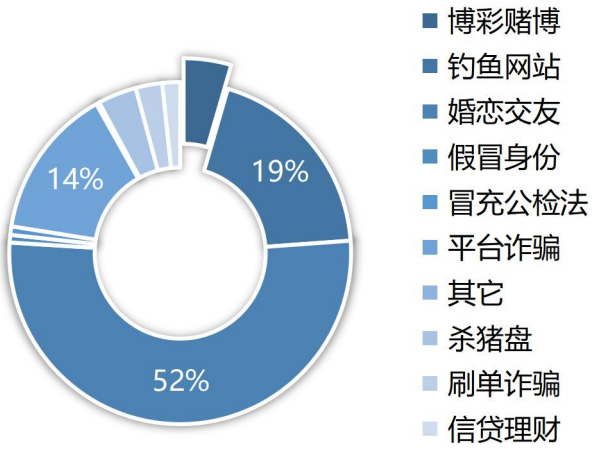

图 1-2 诈骗网址数据饼状图

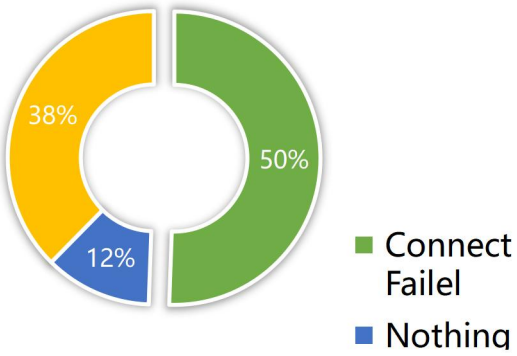

图 1-2 有效网址与无效网址饼状图

训练集中正常网址与诈骗网址比例大概约为 32.3:1，且**不同诈骗类别间分布极其不均匀**，比例差距最大为 1793：1。

将能够爬取到内容的网址称为有效网址，训练集中仅有**38%**左右的网址能够访问且爬取到内容。所以**仅仅基于内容特征**对涉诈网址进行分类时间成本高且有效网址数量少。

1.2 问题解决方案

1.2.1 针对特征提取问题：

我们同时应用 ngram 字符提取和 CNN 神经网络对原始数据预处理，**从学习低层特征不断进化为高层特征，不需要人工提取复杂特征**。在提取词法特征时，我们将字符序列通过编码形成字符 ID，同时提取所有特殊字符和单词生成 URL_BOW 词袋模型，最后将字符和单词分别映射进 word embedding 词嵌入矩阵；在提取文本内容特征时，我们采用 CNN 提取方式，进行文本分词以及去除停用词，最后转化为词向量。

运用级联框架多层递进对 URL 进行分析，在我们的框架下，我们同时采用两套算法+双分支处理，首先运用 minesweeping-CNN 对诈骗与非诈骗网址进行分类，再经过 CNN 对诈骗类

别间网址进行分类。

1.2.2 针对高质量诈骗网址的问题处理

我们同时结合**词法特征和文本内容特征**来分析网址，充分挖掘网址的特征，更加结构化地解析网址。在特征工程上，我们使用了 word embedding 词嵌入方式深度学习 URL 表征。在词法特征提取上，我们同时采用了**字符级+单词级词嵌入**，考虑了 URL 的长组件信息、字符边界及特殊字符处理，有效对抗了恶意网址对良性网址的模仿。

1.2.3 针对数据不平衡的问题处理

对于诈骗与良性网址之间的数据不平衡问题，我们主要采用 ADASYN 过采样技术来对少数类样本赋权，对少数类诈骗网址样本进行生成，从而达到诈骗与非诈骗样本之间的数据平衡；对于诈骗类网址不同类别间数据不平衡的问题，我们运用 seqGAN 序列对抗网络，参考最大化激励函数，进行少数类样本的文本生成，对少数类样本进行文本扩充与增强，从而达到数据平衡的目的。

1.3 小结

针对目前社会的需求以及待解决的问题，我们希望通过机器学习的方式训练出能够精准预测并对恶意网址进行分类的系统，并且支持模型的动态更新与云计算部署，能够对多种数据进行处理，高效识别涉诈网址，提升人工智能识别的精度和速度，以期进一步保障互联网的安全稳定运行 。

2. 建模过程解析

2.1 系统结构说明

针对建模过程，我们分为六层结构：数据获取层、数据预处理层、特征工程层、模型训练层结果输出层、展示层分层处理：

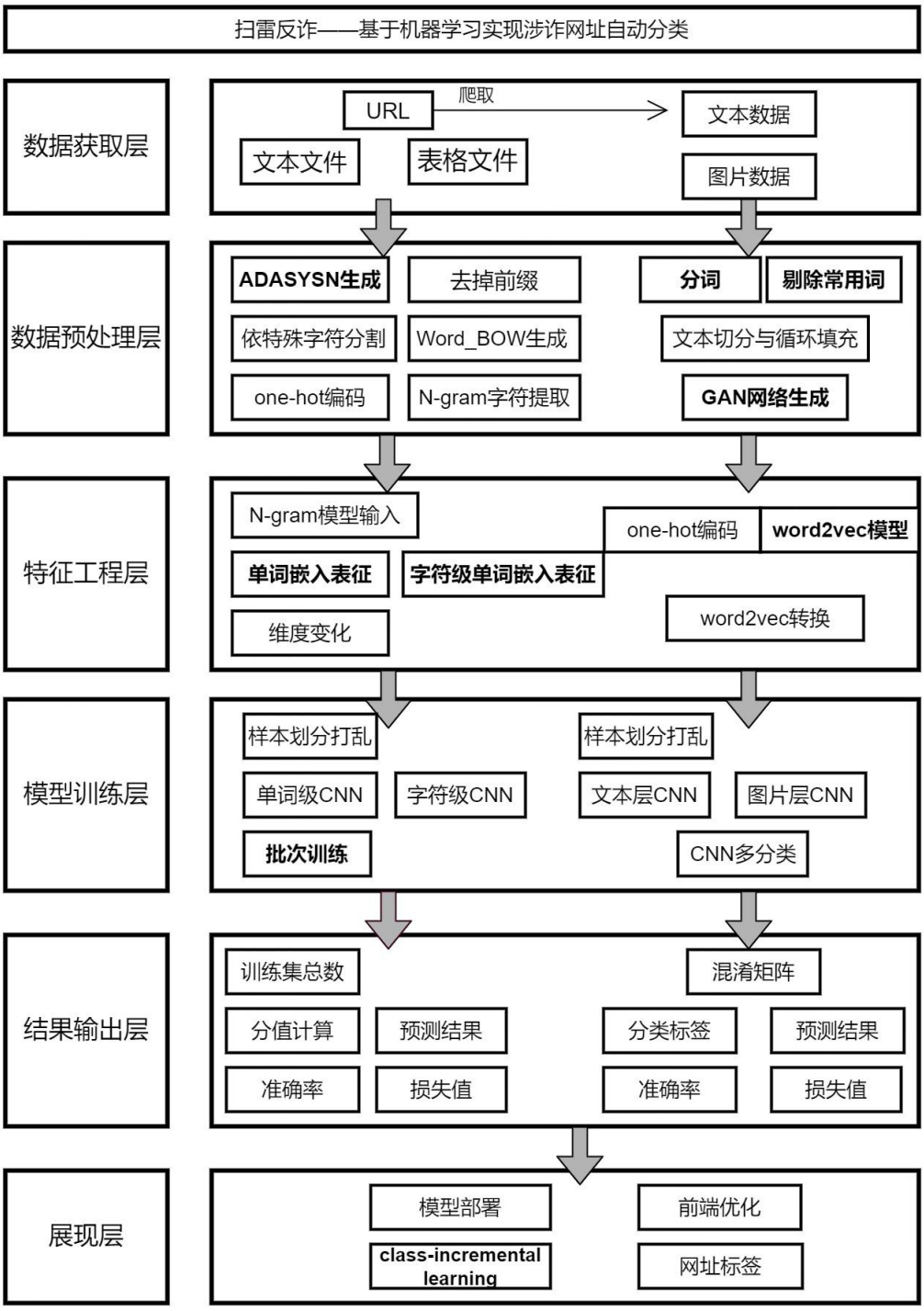

图 2-1 系统层次架构图

2.2 数据获取层

“扫雷反诈"系统具有强大且灵活的数据获取能力，支持多种类型数据的读入和输出。可以支持网址、文本文件、表格文件的输入，并解析为网址作为输出；也可以支持网站文本数据、图片数据的输入，并直接作为输出，也可以支持网址输入，通过爬取技术获得该网页快照的文本数据与图片数据，作为数据获取层的输出。

2.3 数据预处理层

对于 URL 的预处理，我们从词法特征和文本内容特征解析：

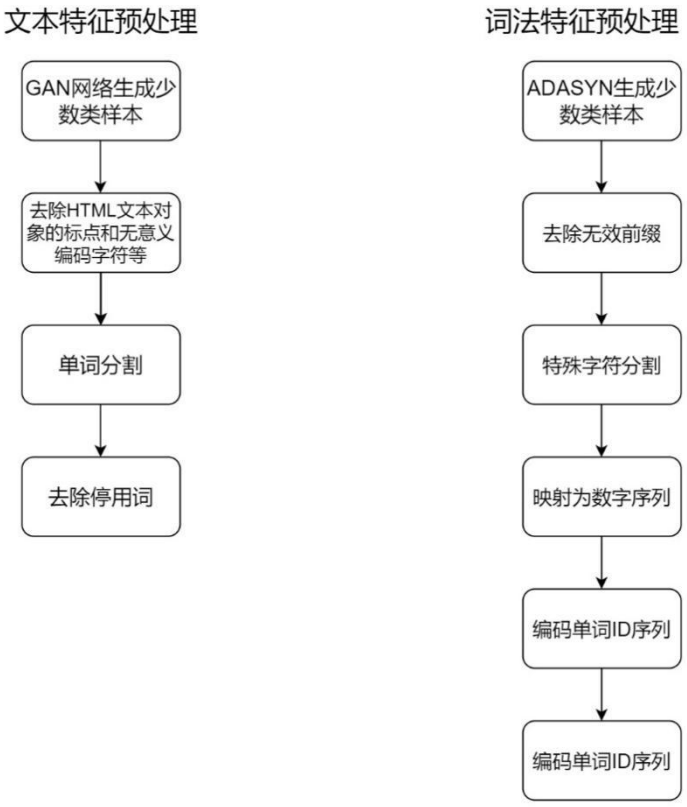

图 2-2 数据预处理流程图

2.3.1 基于网址词法特征预处理

首先读取文件网址，进行两类网址的数量统计。分析两类网址的数量不平衡情况。对于少数类诈骗网址的样本，我们采用 ADASYN 为少数类样本赋予权重，对少数类样本进行生成。再统计处理后的样本数据，达到样本平衡为止。

对于处理后的 URL：首先去除网址无效前缀"http"、"https"、"www"等字样，再对网址域名进行分割，包括依"\"分割，依"?"分割，依"."分割，最终返回网址的主域名、参数、路径、文件名、文件拓展名等参数。然后调用 tensorflow 生成单词到数字的映射表，将单词序列再通过独热编码方式转换为以 N-gram 为模型的单词 ID 序列。字符序列则不需要复杂的预处理，直接转换为对应的字符 ID 序列。

2.3.2 基于网页快照的文本内容预处理

根据上级分类提供的网址，首先进行网址内容的爬取。对于网址内容的文本数据进行长度判断，不达到 600 词的进行长度扩充，长度过长的进行截断。首先进行分词——采用结巴中文分词工具库完成分词，然后去除停用词——对一些中性词和常用词进行过滤，目前使用 Spacy 工具库内置中文停用词列表。

2.4 特征工程层

对于 URL 的特征处理，我们同样从词法特征和文本内容特征解析：

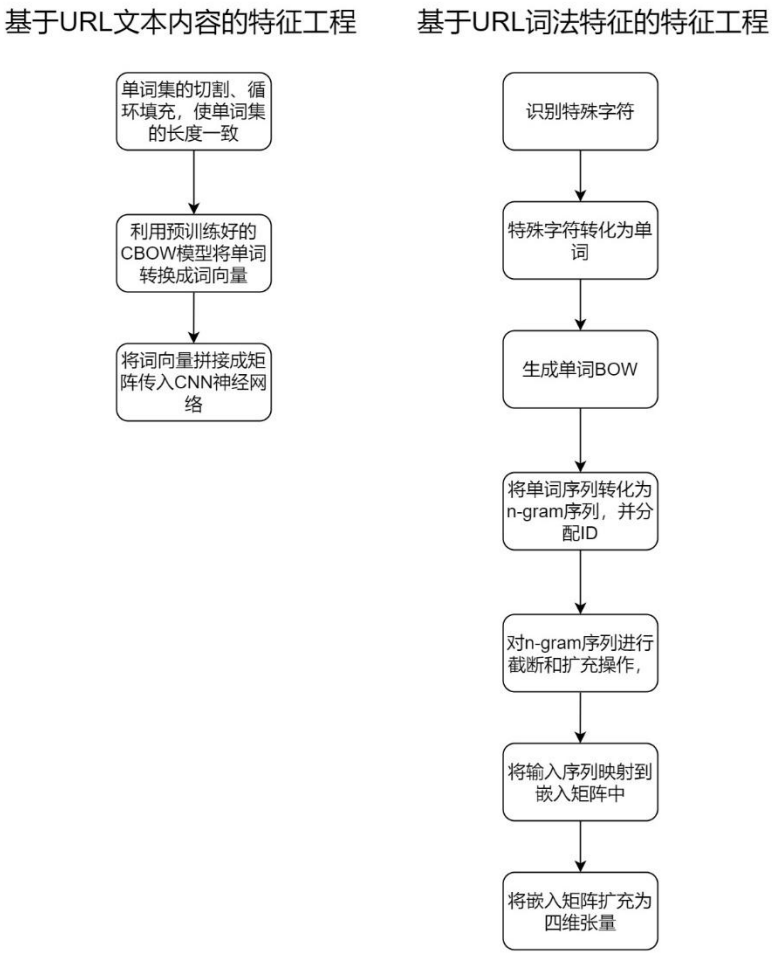

图 2-3 特征工程建立流程图

2.4.1 基于词法特征的特征工程

首先读取预处理后的 URL，识别 URL 中的特殊字符，将特殊字符视为单词。对于单词序列，我们将会生成一个 BOW 词袋模型，收录所有 URL 中的特征单词。再读入预处理后的单词 ID 编码和字符 ID 编码，生成字符＋单词 n-gramID 序列，对 n-gram 进行截断和扩充操作，再映射到嵌入矩阵中，最后将扩充为四维张量。生成单词级+字符级嵌入矩阵。

2.4.2 基于文本内容特征的特征工程

特征工程的最后一步采用一张预训练好的 word2vec 神经网络，将单词转译为独热编码的形式输入，取神经网络隐藏层的权值矩阵作为单词的向量描述，在现行方案中，我们使用一个收录 20000+词向量的 word2vec 模型，这个词向量的长度为 300。将词向量编码后，作为下层 CNN 的输入。

2.5 模型训练

对于我们的双分支级联框架模型，我们均采用 CNN 神经网络进行训练，同样分为两个分支。

2.5.1 基于网址词法特征的模型训练

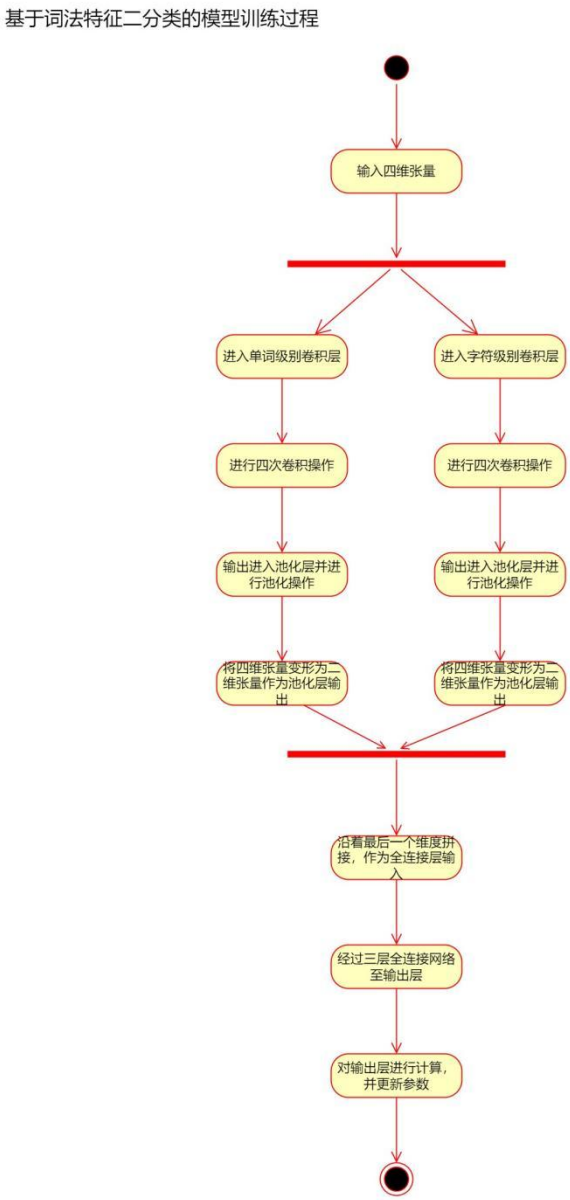

图 2-4 模型训练流程图

首先对输入样本进行随机打乱，再对样本训练集和测试集进行划分，在实际训练中我们采用了 8：2 的比例划分。再对整个训练集以命令行参数进行特定 size 的 Batch 与 epoch 划分为模型批训练准备。

然后应用 minesweeping-CNN，卷积层分为单词级卷积和字符级卷积，通过命令行定义滤波器个数和卷积层大小。经过四次不同大小卷积后输出至池化层，池化层将进行最大操作，经池化 dropout 化后将维度变化为二维 FC。将两级 CNN 池化层输出的 FC 按最后一个维度凭借成一个新 FC，对其进行三层全连接层处理，每层应用 Relu-activation 激活函数处理。

2.5.2 基于文本内容特征的模型训练

在接收到文本和图像内容(一一对应)之后，将文本和图像分别出入进文本处理的 CNN 模型和图像处理的 CNN 模型。

文本将会经历三层卷积与池化其中文本还会经历一次 l2 正则化以防止其过度拟合，图像将会经历两层卷积与池化，然后它们会各自经历 Flatten 层和有 128 个 uints 的全连接层，最后在 Concatenate()函数的帮助下按最后一个维度串联成一个一维向量作为输出层的输入。

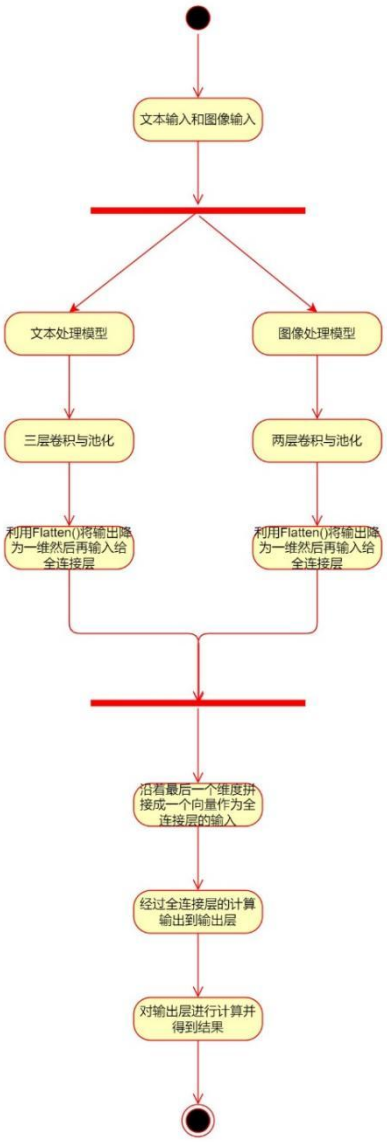

图 2-5 模型训练流程图

2.6 结果输出层

在第一个分支下，最后输出一个二维向量，经过 softmax 分类器输出层会根据分值得到最接近的预测值，从而得到二分类的结果。最后会统计输出训练集大小，并根据统计数据进行模型效果评估，包括 AUC、FPR、TPR 及损失值计算。

在第二个分支下，分类器根据 CNN 操作得到十二个标签值，并打上标签、输出混淆矩阵。最后根据统计数据进行模型效果评估，包括 AUC、FPR、TPR 及损失值计算。

2.7 结果展示层

我们对两种模型进行了云端部署，并设计了用户友好的 UI 界面，在用户输入对应 URL 或其他数据时，可以即时显示出经过我们模型的分类结果。

我们在云服务器后端部署了基于 CIL (Class-Incremental-Learning)增量学习算法，接受新输入的 URL 数据，并将其加入模型训练中进行在线学习，实现模型的动态更新，解决了新老数据的平衡问题，同时使模型最优化。

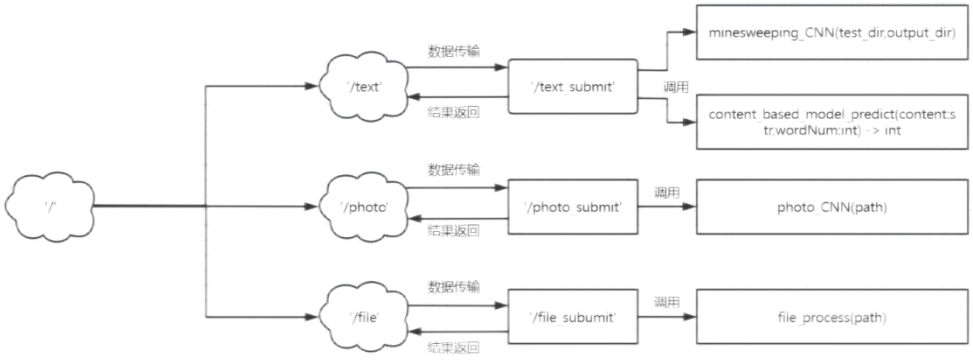

图 2-6 模型部署流程图

在模型部署上，我们使用 python 的 flask 库来实现后端，分别使用"/text"、"/photo"、"/file"三个路由作为网址，图片和文件的三种提交界面。当通过"/text"路由提交时，会进入到"/text_submit"路由，获取用户 post 的数据，进行进程通信，交由另一个二分类文件进行 url 分类，获得结果后若为恶意网址则调用函数进入多分类，再将结果返回给用户端，使用 Ajax 动态更新页面，显示内容；同样的，在文件和图片路由下提交会分别进入"/photo_submit"和"/file_submit"路由调用对应的处理函数进行分类，然后返回用户端动态显示内容。

3. 模型优化

3.1 优化模型架构

①增加了卷积层和池化层的数量，以使网络更深且更具表达能力。

①采用 ReLU 激活函数、L2 正则化和 Dropout，使模型更具泛化能力，提高模型性能。

3.2 数据增强

①采用 ADASYN 过采样+GAN 生成对抗网络生成少数类样本，提高分类器对于少数类的识别能力。

②通过数据增强网页文本切割方法，生成大量不同长度的文本片段，从而扩充训练集，增加数据的多样性。

③通过对图像进行旋转、切割、翻转等处理，增强数据集，减轻过拟合问题，提高模型

的泛化能力。

3.3 超参数调整

使用 sklearn 中的 GridSearchCV 模块来进行网格搜索，对超参数组合进行遍历来选择最佳超参数。通过网格搜索得到 minesweeping-CNN 最佳参数组合为（emb_dim: 32, filter_sizes: '3,4,5,6', nb_epochs: 5, batch_size: 1048）

3.4 优化损失函数

使用交叉熵损失函数（Cross-Entropy Loss Function）来计算模型输出和真实标签之间的差异，指导模型进行参数更新。

在交叉熵损失函数中为信贷理财和刷单诈骗设置较大的权值，以便在训练过程中更关注这两个标签，并提高其准确率。

4.模型效果

4.1 minesweeping-CNN 卷积神经网络

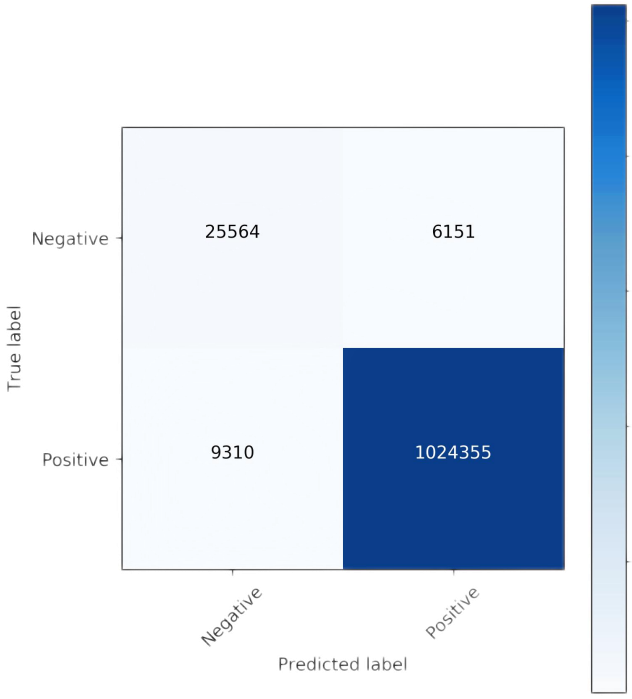

图 5-1 混淆矩阵热力图

	Positive	Negative
Precision	0.994	0.900

recall	0.991	0.820
F1 指标	0.9925	0.857
Score	1.00	0.965

表 5-1 模型评估指标数据

模型常用评估指标如下：

TP（True Positive）：预测结果为正例且实际为正例的样本数。

FN（False Negative）：预测结果为负例且实际为正例的样本数。

FP（False Positive）：预测结果为正例且实际为负例的样本数。

TN（True Negative）：预测结果为负例且实际为负例的样本数。

Precision（精确率）：指的是预测为正例的样本中，实际为正例的比例。其计算公式为：Precision = TP / (TP + FP)

Recall（召回率）：指的是实际为正例的样本中，被正确地预测为正例的比例。其计算公式为：Recall = TP / (TP + FN)

F1 Score：综合考虑了模型的精确率和召回率，是一个比较全面的评价指标。其计算公式为：F1 = 2 * (Precision * Recall) / (Precision + Recall)

AUC（Area Under the Curve）：是 ROC 曲线下面积，常用于衡量二分类模型的表现。AUC 值越大，说明模型的性能越好。

Score = 0.5 * $P_{R=0.7}$+0.3 * $P_{R=0.8}$ +0.2 * $P_{R=0.9}$ 反映了模型在不同条件下的全面表现，可以用来评估模型的鲁棒性和稳定性，分数越高，性能越好。

当决策边界设置为 0 时，TP 值为 19213，即真正例的数目为 19213 个，而 FN 值为 185.0，即误负例数目为 185 个，此外，FP 值为 108，即假正例数目为 108 个，而 TN 值为 494.0，即真负例数目为 494 个。可以看出，模型的准确度非常高，正确率为 0.9944，召回率为 0.9905，并且 F1 值也很高，达到了 0.9924。

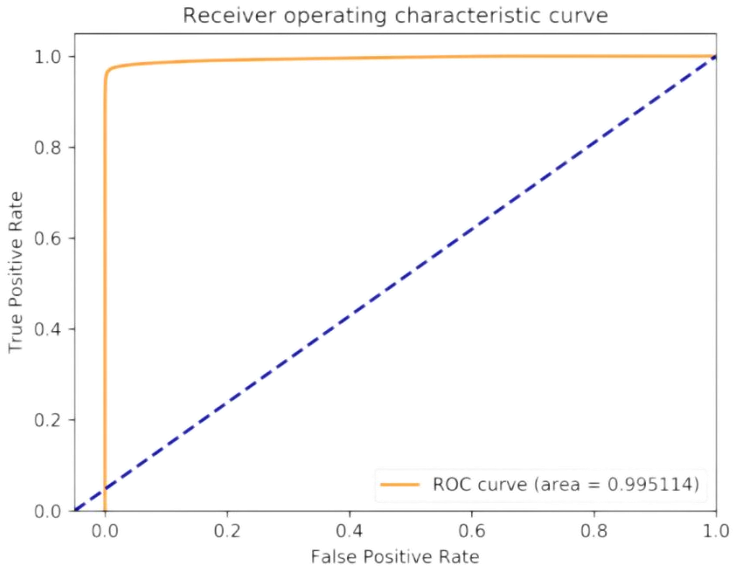

图 5-2 ROC 曲线图

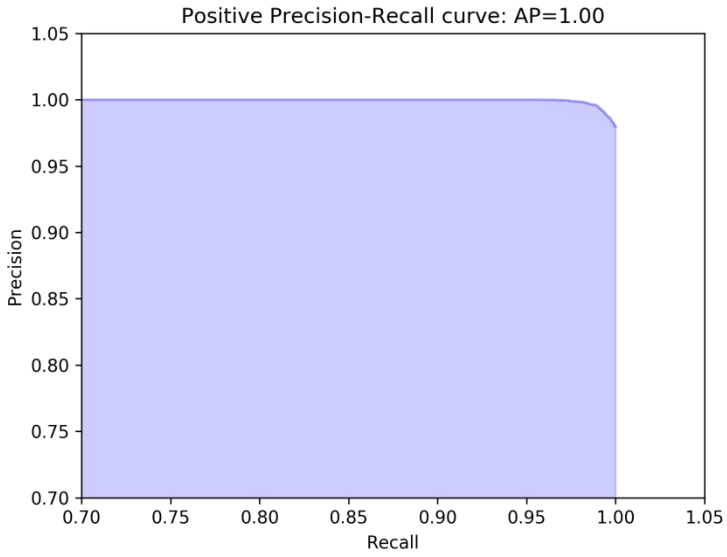

图 5-3 正样本 PR 曲线图

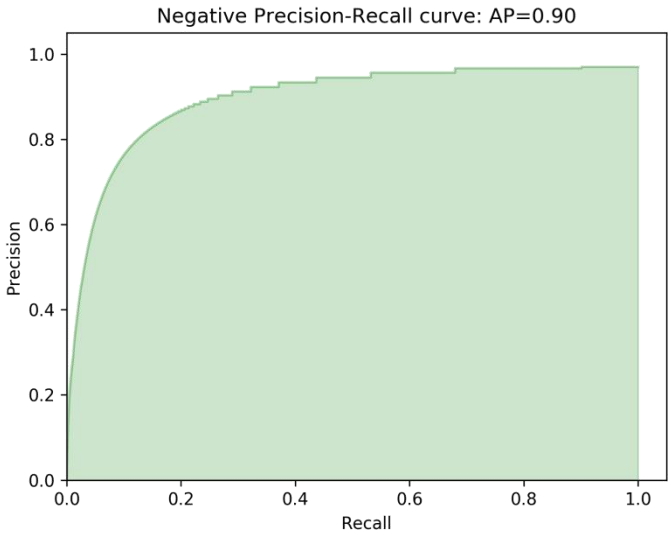

图 5-4 负样本 PR 曲线图

由图 5-2 可知，AUC 值非常接近于 1，因此可以判断该分类器具有非常优秀的性能，能够准确地对不同的正负样本进行分类，并且具有很好的鲁棒性和泛化性能。

由图 5-3 和图 5-4，该模型在正样本上具有很好的分类精度和查全率，而在负样本上的表现稍逊于正样本，但仍然具有较高的准确率和总体得分，说明该模型仍然具有很高的综合性能，适用于很多实际应用场景。

上述数据反映，基于机器学习实现的涉诈网址分类识别系统 minesweeping-CNN 卷积神经网络的效果非常优秀，能够很好地对涉诈网址进行分类，具有很高的准确性和可靠性。

4.2 FraudDetection-Classifier 诈骗网址分类器

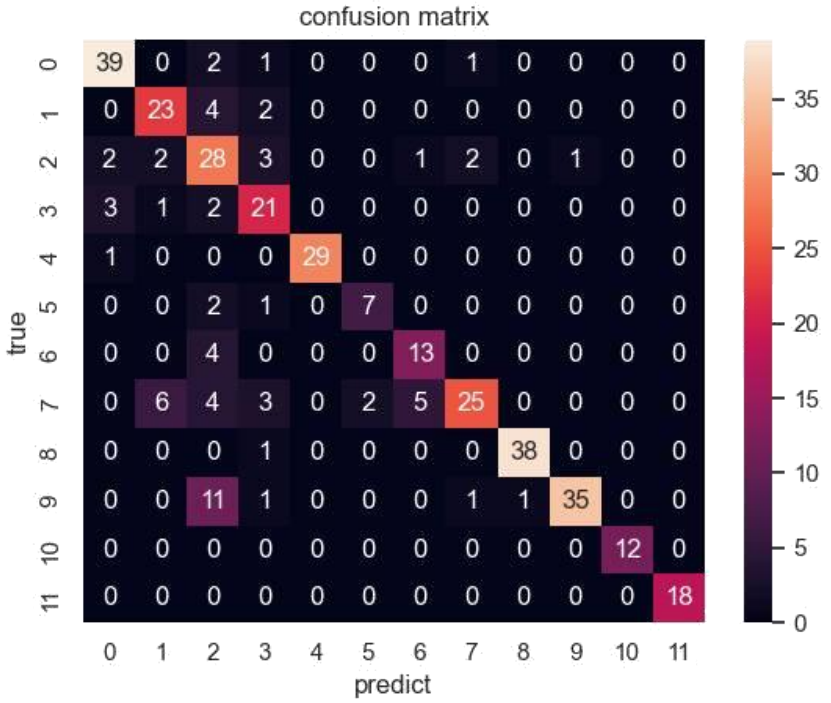

图 5-5 混淆矩阵热力图

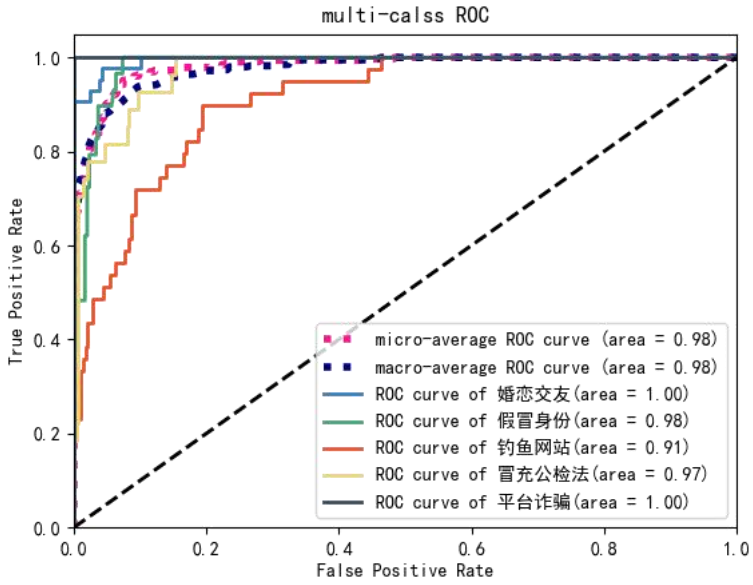

图 5-6 ROC 曲线图 1

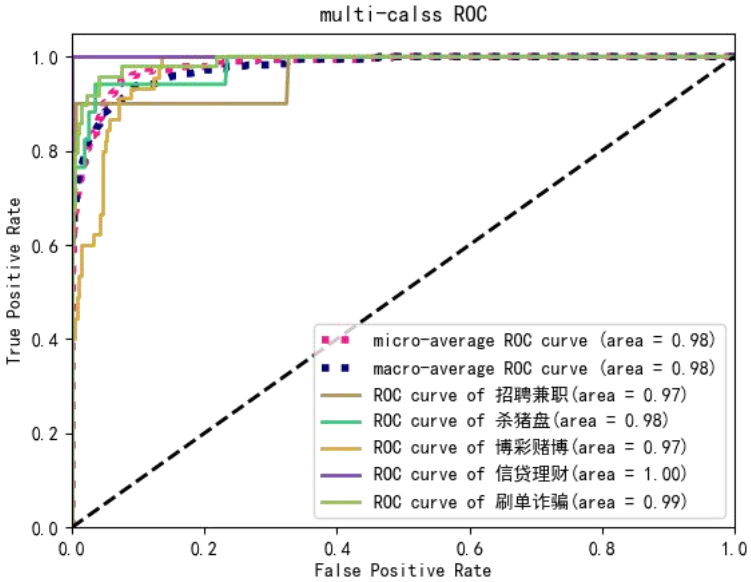

图 5-7 ROC 曲线图 2

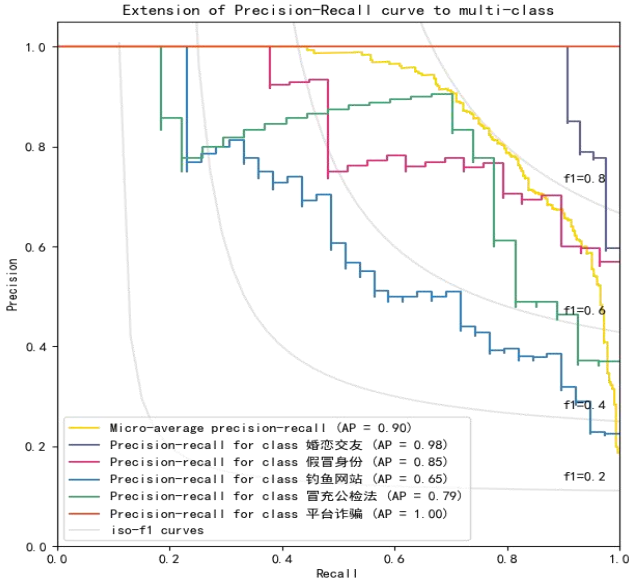

图 5-8 PR 曲线图 1

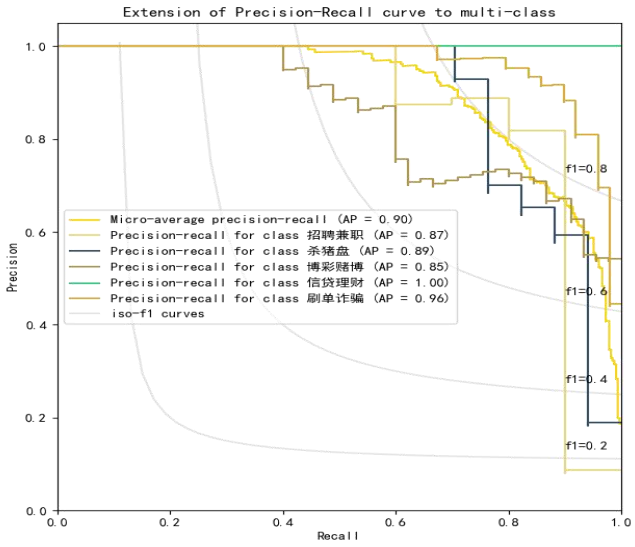

图 5-9 PR 曲线图 2

图 5-10 诈骗网址不同标签精确率和召回率折线图

图 5-11 诈骗网址不同标签得分雷达图

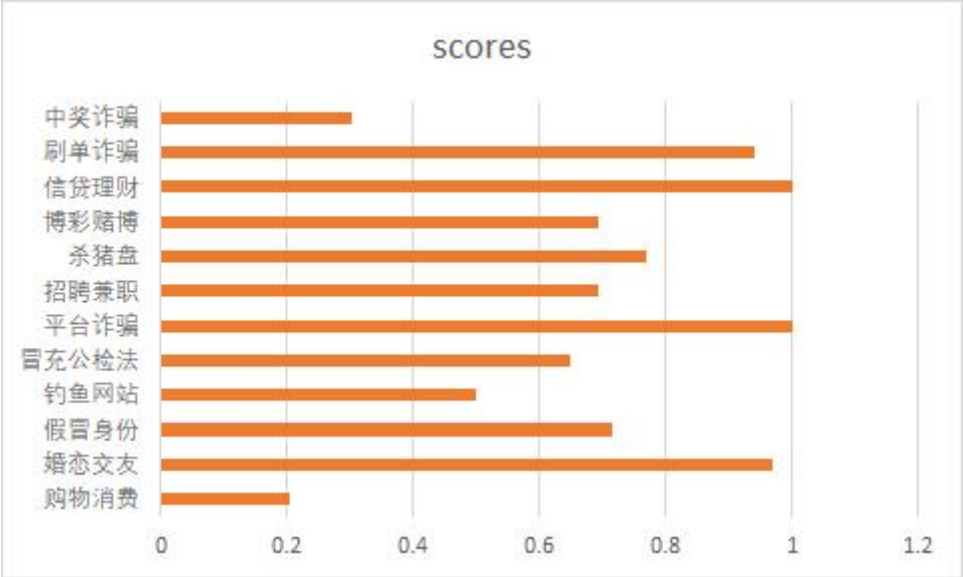

图 5-12 诈骗网址不同标签得分柱状图

根据上述数据分析得出下述结论：

在平台诈骗这个类别下，该神经网络模型的精确率和召回率都达到了很高的水平（分别是 1 和 0.967），该模型能够高度准确地将平台诈骗的样本识别出来，并且没有将其他类型的样本误判成平台诈骗类型。

我们还可以看到企业特殊关注的信贷理财、刷单诈骗这两个类别的精确率和召回率都非常高（分别是 0.974 和 0.784 以及 0.972 和 0.784），该模型对于这两种类型的样本进行分类时非常准确，不仅正确预测了绝大多数该类别的样本，而且并未将其他类别的样本误判为该类别。

此外，其他类别的精确率和召回率也非常不错，尤其是在婚恋交友和假冒身份这两个类别中，该神经网络模型的精确率和召回率都达到了 0.8 以上。总体来看，该神经网络模型的

性能优良，能够准确地对不同类别的样本进行分类，并且在每个类别中表现都非常出色。

5. 模型消融验证

5.1 对于双分类级联分类器的消融验证

在项目开始初期，我们使用单分类 CNN 来对 13 种网址基于网页快照文本特征进行分类，发现网页失效比例很大，且爬取时间非常长，从结果混淆矩阵来看，正常类标签对异常类标签影响较大。

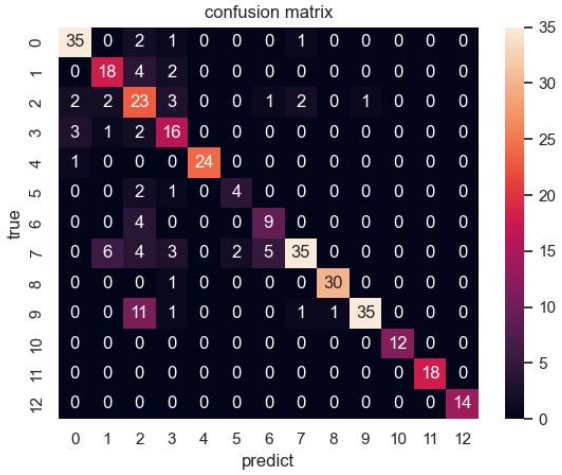

图 5-1 单分类 CNN 混淆矩阵图

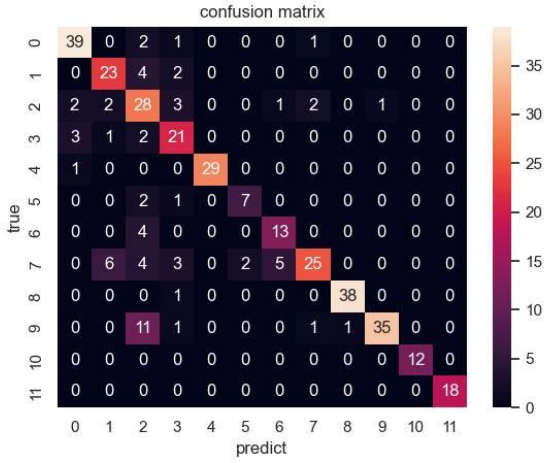

图 5-2 级联分类器混淆矩阵图

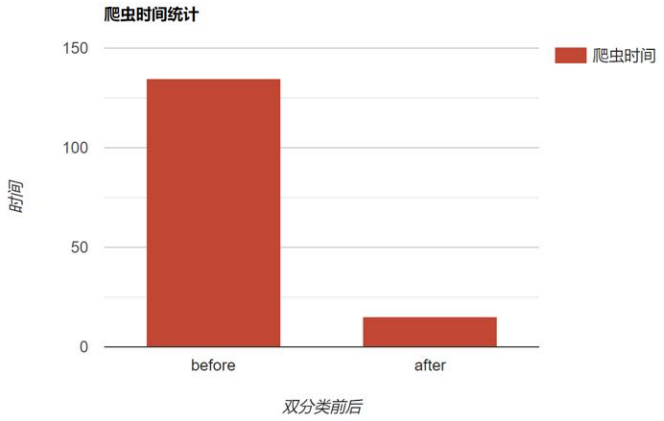

图 5-3 使用级联分类器前后效率对比图

在我们引入双分支级联分类框架后，采用先二分类再十二分类的方式，结合网址词法特征和文本内容特征分析，极大减少了网页爬取时间，从结混淆矩阵来看，正确率大大提升，且消除了正常类标签对异常类标签的分类影响。

5.2 对于增加单词级词嵌入的消融验证

在字符级嵌入 CNN 的基础上，我们增加了单词级嵌入 CNN，有效提高了模型效果。

以下是 minesweeping-CNN 的效果图：

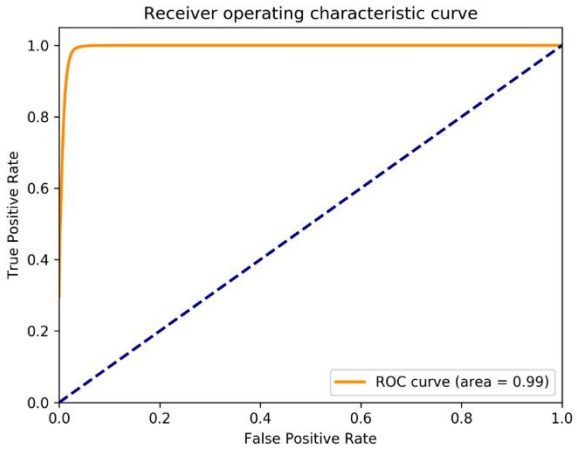

图 5-4 minesweeping-CNN AUC 曲线图

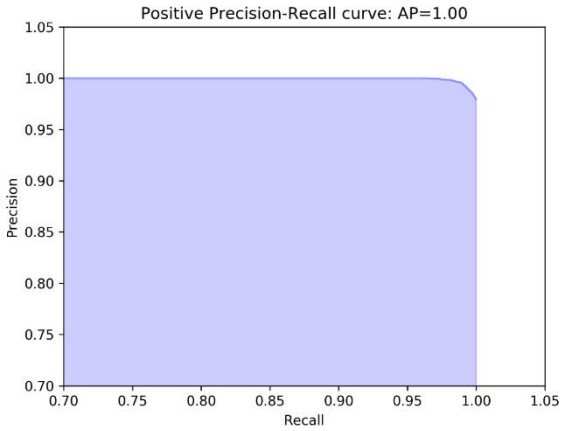

图 5-5 minesweeping-CNN PR 曲线图

以下是对比字符级或单词级单基线嵌入数据统计：

	Training Size = 100,000				Training Size = 1000,000			
		TPR@FPR Level				TPR@FPR Level		
	AUC	0.001	0.01	0.1	AUC	0.001	0.01	0.1
Char-level	0.9114	0.5931	0.8172	0.9211	0.9544	0.6734	0.7012	0.7990
Word_level	0.9557	0.7501	0.8506	0.9413	0.9712	0.8344	0.8425	0.9114
Combined	0.9621	0.7239	0.8233	0.9487	0.9889	0.8709	0.9212	0.9327

表 4-1 针对字符级和单词级特征表征技术消融验证

可以看出：结合字符级和单词级来对 URL 词法特征进行分析，可以有效对抗恶性网址对良性网址的模仿，同时在大样本训练集下，表现更见优越。

5.3 对基于内容特征的卷积神经网络消融实验

在这项研究中，我们研究了在卷积神经网络（CNN）模型中加入批量规范化（BN）层，数据增强以及 Dropout 层对网站分类的影响。这项消减研究的目的是评估 BN 层，数据增强和 Dropout 层对模型性能的影响，以分类精度来衡量。

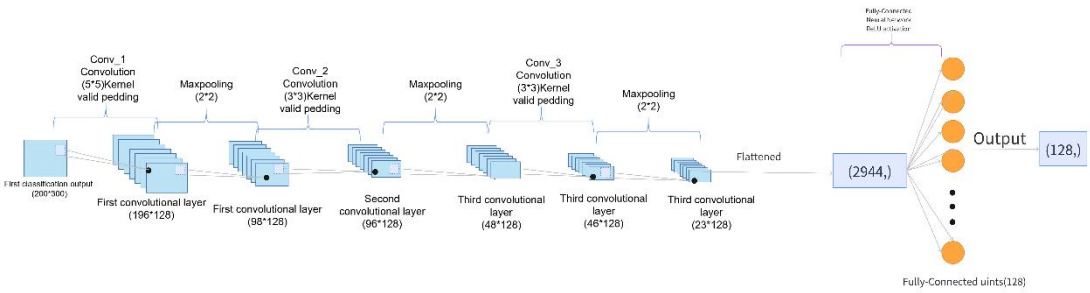

图 5-6 FraudDetection-Classifier 模型结构图

本研究使用的基本模型是一个具有三个 Conv1D 层的 CNN 模型，每个层后面有一个 MaxPooling1D 层，一个 Flatten 层和两个 Dense 层。该模型对卷积层和密集层使用 relu 激

活函数，对输出层使用 softmax 激活函数。该模型用 adam 优化器和 sparse_categorical_crossentropy 损失函数训练。

为了评估 BN 层和数据增强对模型性能的影响，我们设计了以下实验：

①基线模型： 没有 BN 层和数据增强的 CNN 模型

②有 BN 层的模型： 在 Dense 层之后、激活函数之前加入 BN 层的 CNN 模型

③有 Drop 层的模型：在第一个卷积层之后和第一个池化层之前加入 Dropout 层的 CNN 模型

④有数据增强的模型：在模型训练阶段采取数据增强的 CNN 模型

⑤既有 BN 层也有数据增强的模型：在特征提取阶段采取数据增强，在 Dense 层之后、激活函数之前加入 BN 层的 CNN 模型

我们对每个模型进行了 6 个历时的训练，并在网站数据的测试集中评估了它们的性能。结果如下：

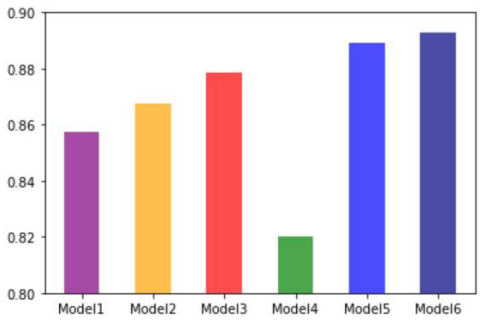

图 5-7 模型消融验证效果对比图

图中的 Model1-6 分别对应以上六种模型，结果显示，在 CNN 模型中加入 BN 层后，其分类精度提高了 2.14%。在 CNN 模型中加入数据增强后，分类精度提高了 1.02%。在 CNN 模型加入 Dropout 层之后，分类精度下降了 3.70%。既有 BN 层也有数据增强的，分类精度提高了 3.22%这表明，BN 层和数据增强可以成为改善 CNN 模型在网站分类中的性能的有用技术。所有技术都有的模型分类精度提高了 3.58%，故而 Dopout 层对于整个模型的作用还有待商榷。

在这项消融研究中，我们评估了在网站分类的 CNN 模型中添加 BN 层和数据增强的影响。结果显示，添加 BN 层后，模型的分类精度提高了 2.14%，加入数据增强后，分类精度提高了 1.02%。这表明，BN 层和数据增强都可以作为一种有用的技术来提高 CNN 模型的网站分类性能，然而再加入 Dropout 层之后，精度却下降了，但在整体模型中却使分类精度上升，这些结果表明，仔细考虑模型的设计和用于实现最佳性能的组件非常重要。

6. 方案亮点与创新点

6.1 系统运行速度快

采用级联框架，引入了多层检测器机制，降低不必要的网页爬取，训练模型快，测试数据快，能够实时检测，分钟级出结果。涉诈网址的数量非常庞大，需要对每一个网址进行分类识别，如果模型运行速度较慢，会给用户带来不良的使用体验和可能因为时间问题造成用

户不必要的损失。级联分类器可以在不降低模型准确率的前提下，提高模型的运行速度，让用户能够快速得到识别结果，增强用户体验。

6.2 系统性能好

采用过采样和网格搜索调整参数，改进模型结构，模型准确率高，诈骗网址的召回率高，误判率低，模型鲁棒性和解释性强。

采用了 ADASYN 过采样和 GAN 生成对抗网络来增加少数类样本的数量，使得模型在处理极度不平衡的数据集时更具有鲁棒性。

采用了 Class-Incremental Learning 类增量算法，可以在不重新训练整个模型的情况下，只针对新的数据进行增量更新，从而大大减少了重新训练的时间和计算资源，提高了系统的效率和可靠性。

为了提高系统的可扩展性和易用性，采用了云部署技术，将模型部署在云端，用户可以通过浏览器等终端设备直接使用系统，无需在本地安装和配置复杂的软件，提高了系统的便捷性和灵活性。

6.3 系统方法新

级联分类器通过逐步分离正例和负例来增强分类器的性能，与传统的单一分类器相比具有更高的准确度和灵敏度。

采用最新的 Class-Incremental Learning 增量学习算法和特征工程，不断更新和优化模型，泛化能力和适应性强。

将多种技术和算法结合起来应用，通过不同的角度对问题进行分析解决，从而提高了算法准确率、识别精度和鲁棒性。同时，这些算法也具有较强的拓展性和适应性，可以应用于多个场景和应用领域，从而成为互联网领域中新颖而实用的算法和方法。

采用 ADASYN 过采样和 GAN 生成对抗网络在涉诈网址识别领域应用是一个新的尝试，相比于传统的算法更加有效。

6.4 功能性特色与非功能性特色

序号	功能性特色	非功能性特色
1	网址多层次分类	高性能和可靠性
2	高准确率预测	数据隐私保护
3	样本不平衡处理	用户友好界面
4	动态更新	可维护性
5	预测速度快	可扩展性

7. 总结

本文档以机器学习对涉诈网址分类识别项目为背景，提出了基于级联分类器、ADASYN 过采样和 GAN 生成对抗网络、minesweeping-CNN 卷积神经网络和 FraudDetection-Classifier 诈骗网址分类器等多种算法和方法。其中详细介绍了模型建立的流程，包括系统结构说明、

数据获取层、数据预处理层、特征工程层、模型训练、结果输出层和结果展示层。此外，还介绍了模型优化和模型效果等相关内容，并通过消融验证方式进行了实验，验证了模型的效果。最后，总结了方案的亮点与创新点，包括系统运行速度快、系统性能好、方法新、功能性特色与非功能性特色等。