

基于集成学习与多头注意力机制的海关欺诈检测算法

于志文¹, 施水才^{1,2}, 王洪俊^{1,2}

(1. 北京信息科技大学 计算机学院, 北京 100192; 2. 拓尔思信息技术股份有限公司, 北京 100196)

摘要: 针对海关的报关欺诈行为, 提出一种深度自注意集成模型, 通过结合树结构信息、深度神经网络与注意力机制的优势, 从而更准确地识别和应对报关欺诈行为。首先, 利用集成学习 XGBoost 模型将海关数据转化为树形结构的叶子节点, 并使用实体嵌入向量表示每个节点; 其次, 利用深度学习模型计算节点之间以及节点内欺诈风险之间的注意力权重, 并得到融合向量; 最后, 利用线性层和激活函数对融合向量进行变换与分类。在真实海关报关数据集上进行实验, 发现与基线模型相比, 所提模型在准确性和鲁棒性方面展现出明显优势, ACC 和 AUC 指标分别达到 89.81% 和 0.814 2。该模型为海关部门发现潜在的欺诈行为并采取相应措施提供了高效可靠的工具, 适用于分析大规模海关数据集, 有助于保护国家利益和国门安全。

关键词: 海关欺诈检测; 注意力机制; 深度学习; 集成学习

DOI: 10.11907/jjdk.241833

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2026)001-0103-07

扫描二维码阅读全文:



Customs Fraud Detection Algorithm Based on Ensemble Learning and Multi-head Attention Mechanism

YU Zhiwen¹, SHI Shuicai^{1,2}, WANG Hongjun^{1,2}

(1. Computer School, Beijing Information Science and Technology University, Beijing 100192, China;

2. TRS Information Technology Co Ltd, Beijing 100196, China)

Abstract: A deep self attention ensemble model is proposed to address customs declaration fraud. By combining tree structure information, deep neural networks, and attention mechanisms, it can more accurately identify and respond to customs declaration fraud. The model first utilizes the ensemble learning XGBoost model to transform customs data into tree structured leaf nodes, and uses entity embedding vectors to represent each node; Secondly, using deep learning models to calculate attention weights between nodes and between fraud risks within nodes, and obtaining fusion vectors; Finally, the fusion vector is transformed and classified using linear layers and activation functions. Experiments were conducted on a real customs declaration dataset, and it was found that the proposed model exhibited significant advantages in accuracy and robustness compared to the baseline model, with ACC and AUC indicators reaching 89.81% and 0.814 2, respectively. This model provides an efficient and reliable tool for customs departments to detect potential fraudulent activities and take corresponding measures, suitable for analyzing large-scale customs datasets, and helps protect national interests and national security.

Key Words: customs fraud detection; attention mechanism; deep learning; ensemble learning

0 引言

在全球化背景下, 国际贸易快速发展, 但随之而来的

海关报关欺诈问题也日益严峻。海关欺诈通常指通过虚报商品信息、瞒报实际进口金额或数量、错误申报税率等手段, 意图逃避或减少关税的行为^[1]。因此, 报关数据中的欺诈样本往往表现为与正常申报不符的异常数据。这

收稿日期: 2024-10-28

基金项目: 国家重点研发计划项目 (2021YFC3340500)

作者简介: 于志文 (2001-), 男, 北京信息科技大学计算机学院硕士研究生, 研究方向为深度学习、自然语言处理; 施水才 (1966-), 男, 硕士, 北京信息科技大学计算机学院教授、硕士生导师, 研究方向为信息检索、中文信息处理、大数据和人工智能; 王洪俊 (1975-), 男, 博士, 北京信息科技大学计算机学院教授级高级工程师、硕士生导师, 研究方向为自然语言处理、文本挖掘、信息检索。

些报关数据结构包括但不限于报关单、箱单、入货通知单、发票、合同等,主要包含的特征包括商品名称、HS 编码、申报数量、原厂地、出发港口、商品类型、单价、是否征免、税率等。例如,申报品名为智利车厘子,进口时间为7月,中转港口在美国。然而,智利车厘子成熟期为10月至次年2月,美国车厘子成熟期为6-8月,且美国车厘子进口关税为10%,智利水果进口关税为0%。因此,极可能是美国车厘子虚报为智利车厘子,以逃避关税。再如,申报品名为樱桃干,但使用了生鲜水果的HS编码0809290000,而实际樱桃干的HS编码应为0803900000,关税为30%。这类虚假申报行为常用于偷税漏税,给国家财政和合法贸易秩序造成了严重威胁。

传统的欺诈检测方法主要依赖于基于规则的系统或监督学习模型,虽然这些方法具备一定的可解释性,且在处理HS编码不匹配等显而易见的错误时表现良好,但面对现代跨境交易中海量、复杂且高维度的数据时表现出较大的局限性,因为其难以应对非线性特征多、数据不平衡以及欺诈模式的动态变化等问题。在实际场景中,许多欺诈行为远比单纯的编码错误复杂,涉及多字段交互关联、跨时间段的异常行为,甚至涉及对申报要素等信息的细微隐性操控。这些复杂且多维度的欺诈模式往往不能被简单的规则系统捕捉,传统的欺诈分类算法并不适用,需要更高级的网络模型来识别。

1 相关研究

海关数据具有高度的敏感性和保密性。早期研究主要依赖于基于规则的方法进行欺诈检测。例如,任尔伟等^[2]通过分析报关单数据,并结合海关人员的经验知识生成风险规则,从而识别潜在的欺诈行为。Singh等^[3]依据海关官员提出的规则来检测印度海关数据中的欺诈行为。尽管基于规则的方法具有较强的可解释性,但效率较低,且规则库需要频繁更新,以适应不断变化的欺诈手段。为克服这些局限性,研究者引入无监督学习方法。例如,Rad等^[4]采用数据聚类和马氏距离分类器检测欺诈行为,但在处理高基数和非线性特征时仍存在不足。

随着欺诈模式复杂性的增加,研究者们逐渐转向监督学习和集成学习方法。例如,Li等^[5]将支持向量机应用于海关商品的自动化分类任务。Kunickaite等^[6]应用模糊最小-最大分类器和逻辑回归方法来进行海关商品欺诈检测研究,取得了比传统监督学习方法更高的分类精度,但在应对数据不平衡和高基数问题时仍存在不足。Li等^[7]采用XGBoost和AdaBoost等集成学习算法进行海关商品潜在欺诈检测研究,但在一些非线性特征提取方面仍面临挑战,影响了模型准确性。González等^[8]将XGBoost与随机森林(Random Forest, RF)结合起来,提高了西班牙海关欺诈检测精度。虽然这些基于监督学习的方法在某些场景

中表现良好,但在处理高基数数据和非线性特征时仍面临挑战。为此,研究者们尝试将深度学习与集成学习方法相结合。例如,Singh等^[9]采用深度神经网络进行海关欺诈检测研究,展现出较强的非线性建模和泛化能力,但容易受到样本偏差的影响,可能导致评估指标倾斜。Vanhoeveveld^[10]采用集成学习方法证实了细粒度的行为特征(如特定实体如收件人、发货人和申报人,以及所涉商品)在欺诈检测中的高度预测性,这些行为特征能够揭示欺诈模式,帮助海关将有限的资源集中在最可能的欺诈案件上。Dangsawang等^[11]将社交网络分析引入海关欺诈行为检测研究中,揭示了欺诈行为的群体联系,但由于数据非结构化和高基数的限制,模型的可信度受到影响。Hooi等^[12]对马来西亚皇家海关总署提供的数据进行分析,引入实体嵌入(Entity Embedding)缓解了高基数问题,通过对高基数分类变量进行特征编码,可在较低维度中更好地表示唯一值,但在实际应用中仍面临样本偏差的挑战。

综上所述,现有研究虽然在海关欺诈检测领域取得了一定进展,但在处理高基数数据和数据不平衡问题时仍有不足。为此,本文提出深度自注意集成学习(Deep Self-Attention Ensemble Model, DAE)模型,首先使用CTGAN(Conditional Table GAN)技术对海关数据进行增强处理;其次通过集成XGBoost的树结构嵌入特征、注意力机制和深度神经网络来捕捉复杂特征之间的关联,有效处理高基数数据和稀疏数据,进而提升欺诈检测的精度与鲁棒性。相较于传统的基于规则或单一监督学习的模型,DAE模型不仅检测精度更高,而且能够通过合成数据提升模型的泛化能力和数据利用效率,从而在数据不全或隐私保护要求较高的场景下依旧保持较高的检测精度。

2 理论基础

2.1 CTGAN

在现有研究领域,数据增强技术已成为解决数据稀缺问题的关键手段,尤其在生成高质量合成数据以扩充数据集、提升模型性能方面展现出显著优势与广泛应用前景。CTGAN(Conditional Table GAN)在处理具有统计属性的海关申报表单数据时具有一定优势,能够在保护隐私的前提下增强海关数据的多样性^[13]。本文创新性地将CTGAN应用于海关数据增强,生成高质量的合成数据,并通过深入分析数据特征关系提升合成数据的质量,提高了模型的训练效果和泛化能力。

2.2 注意力机制

注意力机制凭借其卓越的性能,已在多个领域的特征提取任务中取得了显著成功。该机制的核心优势在于能够自适应地为不同特征分配权重,从而精准捕捉关键信息。基于这一特性,注意力机制已被广泛集成至众多先进的Transformer模型之中,成为提升模型性能的关键组

件^[14]。然而,当前针对报关单数据的研究通常没有考虑到细粒度特征对检测结果的影响。为克服这一局限,本文通过引入自定义的融合注意力机制(Fusion Attention Mechanism)来动态调整模型对重要特征的关注度,尤其是对 HS 编码、申报人、申报要素具体字段等信息进行重点关注,精确捕捉数据中的关键信息^[15]。

2.3 集成学习

XGBoost 和 LightGBM 等算法在复杂模式和非线性关系建模中表现出色,但现有研究多偏向于单一算法,忽视了与深度学习算法结合的潜力^[16-17]。本文将 XGBoost 生成的叶子节点结构与实体嵌入相结合,以提供更丰富的结构化输入。随后进一步与深度神经网络相结合,使模型在检测性能和鲁棒性方面实现较大提升。这一设计区别于传统的集成方法,在处理高维度、非线性特征时展示出更优性能。

针对报关单中的高基数特征问题,现有研究多采用实体嵌入来解决^[18]。然而,单一嵌入的方式难以充分挖掘多

维特征之间的交互关系。本文在 DAE 模型中创新性地引入交叉特征嵌入(Cross Feature Embedding),通过嵌入重要特征组合得到交叉特征嵌入,并与实体嵌入融合,为深度神经网络提供更加细粒度的输入^[19]。该设计能更有效地捕捉特征之间的复杂关系,使模型在欺诈检测精度上获得显著提升。

综上所述,本文针对报关欺诈检测的具体需求对模型进行了创新优化,弥补了现有研究在数据多样性、特征细粒度处理和高维复杂特征识别中存在的不足。

3 DAE 算法构建

DAE 算法包括 3 个关键阶段:①采用基于树的分类器进行预训练来生成交叉特征;②采用深度神经网络学习交叉特征与高风险特征之间的相互作用;③采用注意力机制优化分类模型完成欺诈检测任务。算法总体架构如图 1 所示。

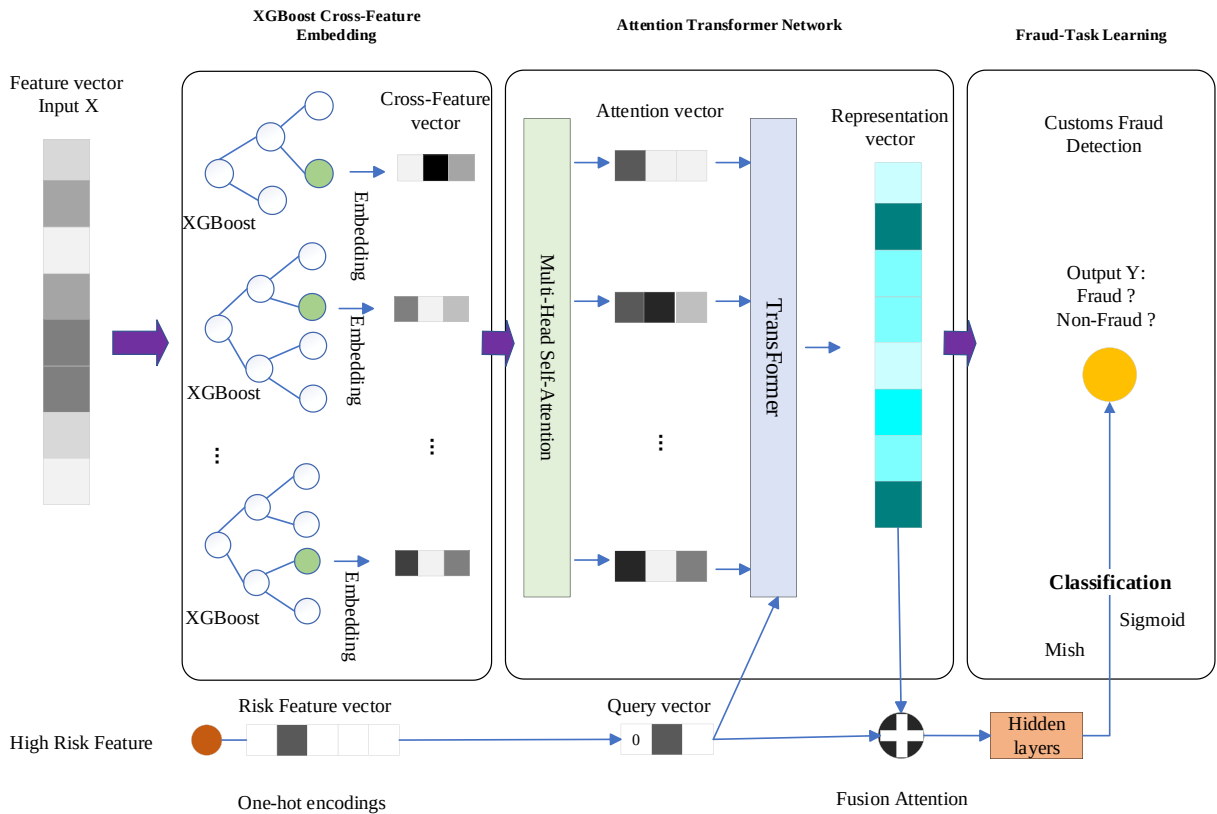


Fig. 1 Overall architecture of DAE algorithm

图1 DAE算法总体架构

在第一阶段,为应对数据的高基数问题并捕捉细粒度特征之间的关系,采用基于树的分类器进行预训练。树结构能够捕捉高风险特征之间的关联性,从而更准确地捕捉相关特征之间的相互作用,并且可以从叶节点路径中提取有价值的交叉特征。树模型的输出能够转化为稀疏的高维嵌入向量,为下游的深度神经网络模型提供结构化输入,从而作为 DAE 算法一个高效稳定的起点。首先,通过

XGBoost 生成叶节点嵌入向量,采用 GBDT(Gradient Boosting Decision Tree)方法构建多个 CART(Classification and Regression Tree)树基学习器。每棵树中的每个叶子节点表示一个特定的海关交易项,每个节点生成唯一的索引。从树的根节点到叶子节点的路径代表一个决策路径,如“PACK_NO > 2”、“NET_WT > 500”和“TRADE_COUNTRY == SHWKNE(已脱敏加密)”。这些路径经过多个中

间节点形成多个特征组合在一起的交叉特征。其次,通过实体嵌入技术将每个海关交易项表示为一个嵌入向量,将交叉特征表示为交叉特征向量,这样不仅捕捉了特征之间的关联性,而且生成了高稀疏的嵌入向量,为后续的深度神经网络模型提供结构化输入,从而提高模型的鲁棒性和检测准确性。

在第二阶段,为捕捉树节点之间的关联,以及高风险特征(如申报人、HS10 编码等)与节点之间的相关性,同时增强对关键特征的关注意,构建一个深度自注意力网络模型。首先,进行嵌入层的映射,将特征中的叶子节点映射为向量表示 Leaf_vector,将节点的交叉特征向量表示为 Cross_feature_vector;其次,将 Leaf_vector 与 Cross_fea-

ture_vector 拼接,作为 key_vector 和 value_vector。使用多头自注意力机制计算叶子节点的权重,捕捉叶子节点之间的相关性。通过自定义的注意力块学习特征与叶节点之间的注意力权重,更新嵌入向量;再次,生成查询向量,将申报人、物品和物品归类的标识经过嵌入层处理后,得到相应的嵌入向量 AGENT_vector、HScode_vector 和 CODE_TS_vector,将它们相乘得到查询向量 query_vector,表示高风险特征之间的关联;最后,使用融合注意力模块将 query_vector 与更新之后的 key_vector 输入到融合注意力模块中,对注意力权重向量进行加权求和,生成融合后的特征向量,动态调整对重要特征的关注度。神经网络部分的结构如图 2 所示。

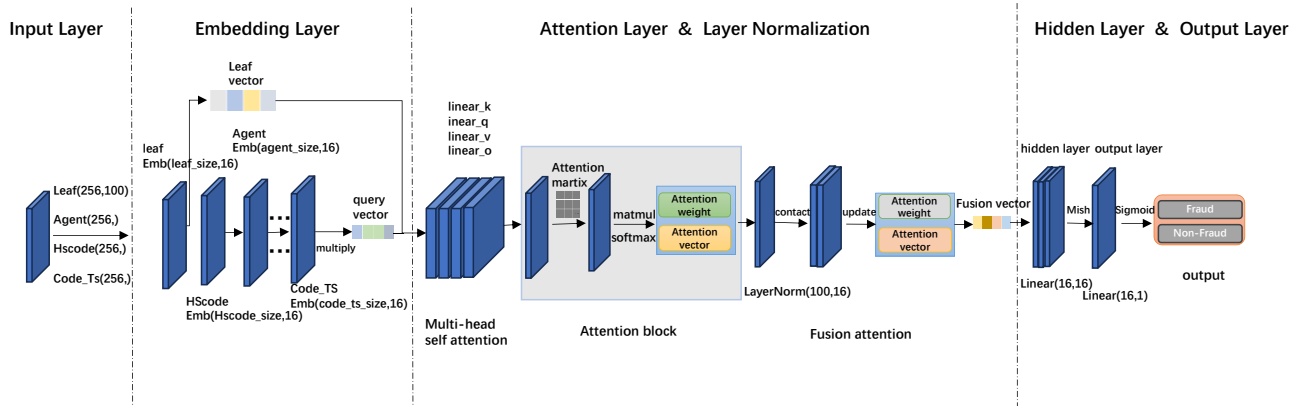


Fig. 2 Neural network structure

图 2 神经网络结构

在第三阶段,将更新后的融合向量作为输入传递到隐藏层。首先,通过融合向量进行层归一化(Layer Normalization),并传递到线性层进行线性变换;其次,通过非线性激活函数对线性层输出进行非线性变换与处理,得到隐藏层输出,该输出包含从原始特征和交叉特征中提取的高阶特征;再次,将隐藏层输出送入分类输出层(Classifier_layer),使用二元交叉熵损失函数(BCELoss)计算分类损失,生成每个交易项的最终预测结果。同时通过反向传播算法来最小化损失函数,使用 Ranger 优化器更新模型参数,确保模型在训练中收敛;最后,向模型中输入新的向量进行预测,将隐藏层的输出映射为欺诈检测分类结果,检测是否存在欺诈行为。

DAE算法的伪代码为:

输入:训练集 D , 隐藏层维度 dim , epoch 数 $epochs$, 学习率 lr , 正则化系数 $weight_decay$, 注意力头数 $head_num$, 损失函数权重 cls_loss_func , 融合注意力策略 $fusion$, 激活函数 act

输出:报关数据预测结果 OUTPUT

1 数据预处理

$D_after \leftarrow preprocess(D)$

2 初始化 XGBoost 模型

$xgb_clf \leftarrow XGBClassifier(n_estimators=100, max_depth=30, n_jobs=-1, learning_rate=0.3, \dots)$

3 提取 XGBoost 的叶节点特征

$D_leaves \leftarrow xgb_clf.apply(D_after)$

$transformed_leaves, leaf_num, new_leaf_index \leftarrow process_leaf_idx(D_leaves)$

4 叶节点嵌入表示

$entity_embedding + cross_feature_embedding \leftarrow nn.Embedding(transformed_leaves, embedding_dim)$

5 初始化模型和优化器

$init_model \leftarrow DAE(D, dim, head_num, \dots)$

$init_optimizer \leftarrow Ranger(model.parameters, lr, weight_decay)$

6 模型训练

FOR i IN range($epochs$):

FOR j IN batch_data(batch_feature, batch_cross_feature, batch_cls):

$model.train()$

Forward:

$classification_output, \dots, hidden_vector \leftarrow model(batch_feature, batch_cross_feature)$

Compute loss:

$cls_loss \leftarrow cls_loss_func(classification_output, batch_cls)$

Backward:

$optimizer.zero_grad()$

$cls_loss.backward()$

$optimizer.step()$

7 模型评估

$model.eval()$

```

WITH torch.no_grad():
    test_feature, test_cross_feature, test_cls ← D_test_preprocessed # 预处理之后的数据
    test_output, _ ← model(test_feature, test_cross_feature)
    OUTPUT ← torch.sigmoid(test_output)
    # 计算测试集上的性能指标(如准确率、AUC等):
    performance_metrics ← evaluate_model(OUTPUT, test_cls)
8 输出最终的预测结果和性能指标
RETURN OUTPUT, performance_metrics

```

4 实验方法与结果分析

4.1 实验设置

在 DAE 模型训练阶段,首先使用 XGBoost 模型将海关数据转化为树形结构,生成交叉特征。XGBoost 模型参数设定为:树的数量为 100,深度为 4。这些树的叶节点表示特定的海关交易项,每个节点生成唯一索引,从根节点到叶节点的路径代表一个决策路径,如“PACK_NO > 2”、“NET_WT > 500”等。这些路径形成多个交叉特征,通过实体嵌入技术将这些特征表示为高维稀疏嵌入向量,为后续深度神经网络提供输入。

在实体嵌入过程中,为每个分类变量的类别设置嵌入尺寸,嵌入尺寸大小根据类别数量设定。如果某分类变量的类别数量 n 小于 2 000 个,则嵌入尺寸设为 n 的一半,即 $n/2$;否则,嵌入尺寸设为 1 000。该阈值基于模型复杂度与计算资源之间的平衡来设置,以确保在捕捉足够信息的同时不引入过多计算负担。

在深度自注意力网络构建中,首先将叶子节点和交叉特征向量通过嵌入层映射为向量表示,通过多头自注意力机制计算叶子节点之间的权重,设置 head_num 为 8,用于增强模型的表达能力,可以同时关注海关交易项的不同部分,捕捉它们的关联性。同时自定义一个 Mish 激活函数。该函数是近年来提出的一种新型激活函数,计算公式见式(1)^[20]。其中,softplus(x)函数表示软正值函数,计算公式见式(2)^[21]。Mish 激活函数在 ReLU、LeakyReLU 等激活函数的基础上进行了改进,以更好地应对梯度消失和梯度爆炸等问题^[22-23]。本文算法创建了一个 Mish 类,并重写了 forward 函数,实现了 Mish 激活函数的计算。具体来说,forward 函数接受一个张量 x 作为输入,使用 tanh 和 F.softplus 函数分别计算 $\tanh(\text{softplus}(x))$ 和 $\text{softplus}(x)$,然后将它们相乘,得到 Mish(x)的输出结果。

$$\text{Mish}(x) = x * \tanh(\text{softplus}(x)) \quad (1)$$

$$\text{softplus}(x) = \log(1 + \exp(x)) \quad (2)$$

融合向量 fusion_vector 经过隐藏层后得到输出 hid_output,然后通过线性分类层的 sigmoid 函数得到最终的 $y_{\text{pred_prob}}$ 。计算公式见式(3)。根据得到的预测值和原始标签 y_{true} 计算损失值 cls_loss,计算公式见式(4)。根据 loss 值进行反向传播,不断优化模型。

$$y_{\text{pred_prob}} = \frac{1}{1 + \exp(-\text{hid_output})} \quad (3)$$

$$\text{cls_loss}(y_{\text{pred_prob}}, y_{\text{true}}) =$$

$$-\frac{1}{N} \sum_{i=1}^N [y_{\text{true},i} \cdot \log(y_{\text{pred_prob},i}) + (1 - y_{\text{true},i}) \cdot \log(1 - y_{\text{pred_prob},i})] \quad (4)$$

4.2 数据集

本文数据集来自某海关部门提供的真实报关单数据,经过脱敏处理后得到 15 000 条记录,包括录入编号、海关编号、商品编号 HS10 编码、申报单位、备案号、进出境关别、申报日期、运输方式、原产国、最终目的国、启运国、进口类型、进口用途、币制、税率、数量及单位、价格等信息。为扩充样本量,使用 CTGAN 算法生成 50 000 条合成数据,确保合成数据与原始数据分布一致。报关单数据包含多达 70 个必填字段,由于部分字段为空值,需要对数据进行清洗与特征选择,仅保留高质量数据进行模型训练。对剩余字段进行重命名以确保一致性,同时创建额外的数值和分类属性,如单价、重量单价、税金比率等,并将商品 HS10_Code 的 10 位编码细分为 5 个子类属性。在数据集中,真实的欺诈标签来自于海关历史记录和人工核验结果,标记为“欺诈”或“非欺诈”。具体来说,每条记录有一个二分类标签,1 代表欺诈样本,0 代表正常样本。这些标签基于海关部门的实际审核结果进行标注,从而确保模型能够识别实际业务中的欺诈行为。

为进一步提升模型的鲁棒性,进行数据增强操作。将高风险分类属性(如进口关别、进口商、HS 编码、原产国、商品类别、出口享惠情况等等)进行组合,生成新的属性列,并合并到原始数据中。通过风险分析识别并标记高风险实体,计算每个特征的欺诈货物数量,将排名前 20% 的实体标记为高风险(标签为 1),其余标记为低风险(标签为 0)。

4.3 实验结果与分析

为全面评估 DAE 算法的性能,将其与 LR、DT、RF、AdaBoost、XGBoost、LightGBM、CatBoost 等算法进行比较,使用 AUC (Area Under Curve) 值、准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall)、F1-score 等指标进行评估。不同模型的准确率、AUC 值比较见表 1。不同模型在查验率分别为 1%、5% 和 20% 时的精确率和召回率见表 2。其中,DAE1 为 LightGBM 版本的 DAE,DAE2 为 XGBoost 版本的 DAE。可以看出,DAE 算法在多个性能指标上明显优于其他分类算法,在低查验率下仍然能够准确识别欺诈案例,不会受到数据维度和特征复杂性的影响,说明其成功捕捉了复杂数据集中隐藏的欺诈行为,具有较高的预测精度。

不同算法的 ROC 曲线和 Precision-Recall (P-R) 曲线见图 3 和图 4。可以看出,DAE2、DAE1 算法的 ROC 曲线更靠近左上角,这意味着它们在较低的假正率下便能获得较高的真正率。DAE2 的 AP 值最高,为 0.20,在所有对比模型中表现最好,DAE1 紧随其后,说明二者在整体上能够更好地平衡精确率与召回率。

Table 1 Comparison of accuracy and AUC values of different algorithms

表 1 不同算法的准确率、AUC 值比较

算法	Accuracy	AUC
LR	0.576 2	0.596 8
DT	0.676 2	0.631 4
RF	0.676 6	0.724 9
AdaBoost	0.660 8	0.670 5
XGBoost	0.820 5	0.713 1
LightGBM	0.787 5	0.692 7
CatBoost	0.695 1	0.707 4
XGBoost + LR	0.829 7	0.733 2
LightGBM + LR	0.805 1	0.674 5
DAE1	0.889 5	0.803 3
DAE2	0.898 1	0.813 9

Table 2 Comparison of accuracy and recall rates of different algorithms under different inspection rates

表 2 不同查验率下不同算法的精确率和召回率比较

算法	Precision			Recall		
	1%	5%	20%	1%	5%	20%
LR	0.152 2	0.159 5	0.126 5	0.020 0	0.105 2	0.334 0
DT	0.106 1	0.117 5	0.130 2	0.013 9	0.077 5	0.343 9
RF	0.209 9	0.206 5	0.174 9	0.027 6	0.136 3	0.461 8
AdaBoost	0.646 7	0.540 9	0.528 7	0.035 7	0.149 7	0.449 8
XGBoost	0.610 3	0.574 8	0.548 1	0.036 5	0.151 2	0.468 7
LightGBM	0.632 0	0.575 7	0.540 5	0.033 4	0.146 8	0.452 6
CatBoost	0.253 7	0.230 9	0.182 6	0.033 4	0.152 4	0.482 2
XGBoost + LR	0.679 6	0.565 3	0.545 1	0.038 3	0.151 5	0.482 8
LightGBM + LR	0.510 8	0.535 0	0.526 4	0.036 5	0.148 4	0.478 1
DAE1	0.565 2	0.560 6	0.535 9	0.039 7	0.148 5	0.546 6
DAE2	0.673 9	0.593 0	0.569 2	0.040 3	0.149 5	0.554 7

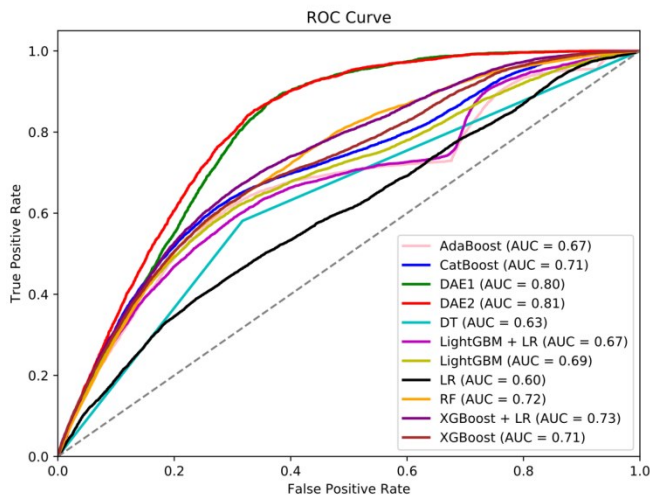


Fig. 3 ROC curve

图 3 ROC 曲线

4.4 消融实验

为验证 DAE 算法中各个模块的贡献,进行以下消融实验,重点考察数据增强技术 CTGAN 的影响。各算法不使用和使用 CTGAN 时的精确率、召回率和 F1-score 见表 3。可以看出,使用 CTGAN 后,多数算法的精确率、召回率和 F1-score 指标均有不同程度的提升。例如,DAE2 算法使

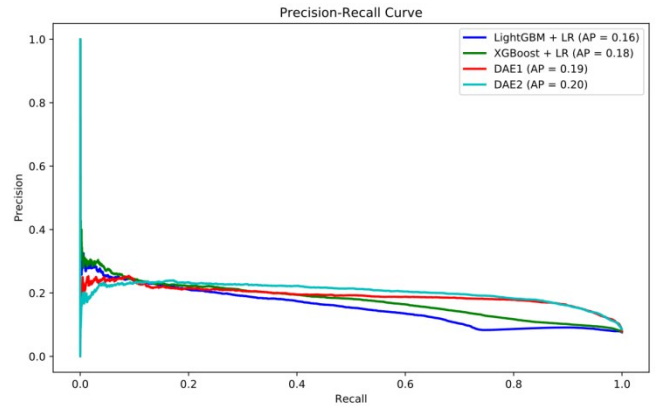


Fig. 4 P-R curve

图 4 P-R 曲线

用 CTGAN 后,精确率由 0.377 3 增加至 0.642 3, F1-score 由 0.534 9 上升至 0.939 3。这是由于 CTGAN 能够有效生成多样化的样本,对模型的决策边界有积极影响。消融实验结果表明,CTGAN 在提升模型性能方面具有显著作用,尤其是在处理复杂数据特征和样本不均衡问题时展现出很好的泛化能力和鲁棒性,证实了 DAE 算法设计的有效性和合理性。

Table 3 Result of ablation experiment

表 3 消融实验结果

算法	无 CTGAN			有 CTGAN		
	Precision	Recall	F1-score	Precision	Recall	F1-score
LightGBM	0.410 8	0.683 0	0.513 0	0.560 2	0.505 9	0.627 6
LightGBM+LR	0.408 7	0.676 8	0.509 7	0.554 2	0.500 8	0.620 5
XGBoost	0.423 2	0.695 7	0.526 2	0.589 0	0.511 3	0.694 5
XGBoost+LR	0.426 8	0.754 7	0.545 2	0.596 7	0.498 4	0.743 5
DAE1	0.349 4	0.926 0	0.507 4	0.650 2	0.510 9	0.893 6
DAE2	0.377 3	0.918 9	0.534 9	0.642 3	0.562 8	0.939 3

5 结语

在全球化、数字化贸易快速发展的背景下,海关报关欺诈检测变得尤为重要。为保护国家利益、维护公平竞争环境和捍卫消费者权益,本文提出 DAE 算法,通过三阶段的处理流程提高了欺诈检测的准确性。该算法结合树结构信息和深度学习模型的强大能力,通过融合注意力机制和集成学习算法,同时借鉴实体嵌入和交叉特征嵌入策略,有效缓解了数据不平衡和高基数等问题,并取得了较优的性能表现。该算法为海关部门发现并处置潜在的欺诈行为提供了一种高效、可靠的工具。

然而,DAE 算法仍存在一定的局限性。一方面,模型复杂度较高,对计算资源需求较大,影响了其在大规模应用中的效率;另一方面,深度神经网络的可解释性不足,难以为决策者清晰呈现预测依据。此外,在涉及国家安全、商业机密和个人隐私的海关数据处理上,仍需进一步强化安全性与合规性。未来研究可从以下几个方面展开:①提升深度模型的可解释性,为政策制定和执法提供更透明的

技术支持;②优化模型的训练与推理过程,提升其效率与可扩展性;③增强模型的抗攻击能力与数据安全保障,以适应更加复杂的应用环境。同时,还可以将DAE算法推广至金融、医疗保险等其他领域,验证其跨场景的适用性与鲁棒性。

参考文献:

- [1] WANG Z R. Research on tax collection and administration system in the era of digital economy[J]. Open Journal of Legal Science, 2023, 11(4): 3188-3198.
王珍容. 数字经济时代下税收征管制度研究[J]. 法学, 2023, 11(4): 3188-3198.
- [2] REN E W, MU Q J, SUN X W. Application of data mining technology in customs inspection and price fraud auxiliary decision-making[J]. Journal of Shanghai Customs University, 2002(3): 58-61.
任尔伟, 牟青杰, 孙学文. 数据挖掘技术在海关查验和价格瞒骗辅助决策中的应用[J]. 上海海关学院学报, 2002(3): 58-61.
- [3] SINGH A K, SAHU R, UJJWAL K. Decision support system in customs assessment to detect valuation frauds [C]//IEMC'03 Proceedings: Managing Technologically Driven Organizations; The Human Side of Innovation and Change, 2003: 546-550.
- [4] RAD H A, ARASH S, RAHBAR F, et al. A novel unsupervised classification method for customs fraud detection[J]. Indian Journal of Science and Technology, 2015, 8: 35.
- [5] LI X, ZHANG Y. A supervised ensemble learning approach for customs product classification and grouping [J]. International Journal of Customs and Trade Logistics, 2021, 15(3): 123-138.
- [6] KUNICKAITE R, BRAZINSKAITE A, SALTIS I, et al. Machine learning approaches for customs fraud detection [C]//26th Conference on Information Society and University Studies, 2021: 54-63.
- [7] LI G, LI N. Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network [J]. Electronic Commerce Research, 2019, 19(4): 779-800.
- [8] GONZÁLEZ G I, MATEOS C A. A comparison between Bayesian dialysis and machine learning to detect tax fraud and its causes: the case of vat, corporate tax and customs duties in Spain [J]. SN Computer Science, 2022, 4(1): 80.
- [9] SINGH K, TSAI Y C, LI C T, et al. GraphFC: customs fraud detection with label scarcity [C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023: 4829-4835.
- [10] VANHOEYVELD J, MARTENS D, PEETERS B. Customs fraud detection: assessing the value of behavioral and high-cardinality data under the imbalanced learning issue [J]. Pattern Analysis and Applications, 2020, 23(3): 1457-1477.
- [11] DANGSAWANG B, NUCHITPRASITTHAI S. A machine learning approach for detecting customs fraud through unstructured data analysis in social media [J]. Decision Analytics Journal, 2024, 10: 100408.
- [12] HOOI E K J, ZAINAL A, KASSIM M N, et al. Feature encoding for high cardinality categorical variables using entity embeddings: a case study in customs fraud detection [C]//2022 International Conference on Cyber Resilience, 2022: 1-5.
- [13] XIANG X Y, WANG J H, WANG Z R, et al. Medical simulation data generation method based on generative adversarial network technology [J]. Journal on Communications, 2022, 43(3): 211-224.
向夏雨, 王佳慧, 王子睿, 等. 基于生成对抗网络技术的医疗仿真数据生成方法[J]. 通信学报, 2022, 43(3): 211-224.
- [14] DAI Y, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 3560-3569.
- [15] VASWANI A. Attention is all you need [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2017: 5998-6008.
- [16] SUN D, SHI W L, RAO L X, et al. Credit card fraud detection method based on improved hybrid sampling and XGBoost algorithm [J]. Computer and Modernization, 2022(9): 111-118.
孙丹, 施炜利, 饶兰香, 等. 基于改进混合采样和 XGBoost 算法的信用卡欺诈检测方法[J]. 计算机与现代化, 2022(9): 111-118.
- [17] LIU S R. Research on credit card fraud recognition based on feature engineering [D]. Lanzhou: Lanzhou University, 2021.
刘思茹. 基于特征工程的信用卡欺诈识别研究 [D]. 兰州: 兰州大学, 2021.
- [18] GUO C, BERKHAHN F. Entity embeddings of categorical variables [DB/OL]. <https://arxiv.org/abs/1604.06737>.
- [19] YU R, YE D, WANG Z, et al. CFFNN: cross feature fusion neural network for collaborative filtering [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 34(10): 4650-4662.
- [20] MISRA D. Mish: a self regularized non-monotonic activation function [DB/OL]. <https://arxiv.org/abs/1908.08681>.
- [21] MURORUNKWERE B F, TUYISHIMIRE O, HAUGHTON D, et al. Fraud detection using neural networks: a case study of income tax [J]. Future Internet, 2022, 14(6): 168.
- [22] SCHLÖR D, RING M, KRAUSE A, et al. Financial fraud detection with improved neural arithmetic logic units [C]//Mining Data for Financial Applications: 5th ECML PKDD Workshop, 2021: 40-54.
- [23] XU J, LI Z, DU B, et al. Reluplex made more practical: Leaky ReLU [C]//2020 IEEE Symposium on Computers and communications, 2020: 1-7.

(责任编辑:尹晨茹)