FDGAN—基于生成对抗网络的保险欺诈检测器

队名：向日葵

队长：王晓光

队员：相洪振

单位：西安交通大学

**1. 项目简介**

**1.1 项目背景**

根据银保监会数据显示，2021年全年我国保险公司各类保险赔款与给付支出为1.56万亿元，全年保险欺诈规模可高达1600-3000亿元。由于保费占比高、金额大，**车险和健康险**这两大险种成为保险欺诈的高发区。因此，准确高效智能地检测保险欺诈意义巨大，具有巨大的研究价值。

在保险行业中，欺诈样本数量相对于正常样本显著偏少，但其带来的经济损失却非常严重。由于数据集的类不平衡，传统的深度学习方法容易出现过拟合，即倾向于将欺诈实例错误地分类为正常实例。此外，保险的样本数据为表格数据，准确学习表格数据的分布也是一件极具挑战性的事情。为了解决这个问题，本项目引入了针对表格数据特点的开发的生成对抗网络（FDGAN），以便合成更多的欺诈样本，从而实现数据的类别平衡，提升模型的识别真实欺诈样本和正常样本的能力。此外，我们选择集成学习算法 XGBoost作为欺诈检测器。XGBoost 是一种成熟的集成学习算法，它采用多棵树进行决策，服从“少数夫从多数”的原则，从而具有更加鲁棒和准确的欺诈检测表现。我们的方法命名为 FDGAN，是一种基于生成对抗网络的保险欺诈检测器。

**1.1 项目目标**

本项目的目标是通过基于针对表格数据的生成对抗网络（FDGAN）的合成样本生成方法，为保险反欺诈提供一种新的解决方案。通过生成真实和欺诈样本的合成数据集，我们能够缓解数据不平衡问题，并提升分类模型对欺诈行为的检测能力。我们选择集成学习算法 XGBoost作为欺诈检测器，与FDGAN结合一起使用，进一步提升欺诈检测准确率。

**2. 技术实现**

**2.1 数据生成模块 （FDGAN）**

我们采用了FDGAN（Conditional Tabular GAN）模型生成合成数据。FDGAN擅长处理类别特征和连续特征的数据，适合用于解决保险数据中的不平衡问题。FDGAN模型根据真实数据的分布生成逼真的欺诈和正常样本，确保在样本分布上与真实数据保持一致。生成的数据与真实数据以1:1的比例混合，形成一个平衡的二分类数据集，从而有效解决了类不平衡问题，提高模型的鲁棒性。

FDGAN是生成对抗网络的一种变体，专门用于生成高质量的表格数据，特别适合解决类别不平衡问题。FDGAN的结构和原理可以通过以下几点来说明。

* + 1. **FDGAN的结构**

在FDGAN中，有两个关键组件：生成器（Generator, G）和判别器（Discriminator, D）。如图1所示。

* 生成器 G：接收噪声输入（通常是随机正态分布噪声），并生成合成样本。在FDGAN中，生成器会基于噪声和条件变量（通常是类别标签）生成模拟真实数据分布的合成样本。
* 判别器 D：接受真实样本和生成样本，并判断其来源，即判断样本是真实数据还是生成数据。判别器的目标是最大化对真实样本的识别准确率，并最小化对生成样本的错误判断。

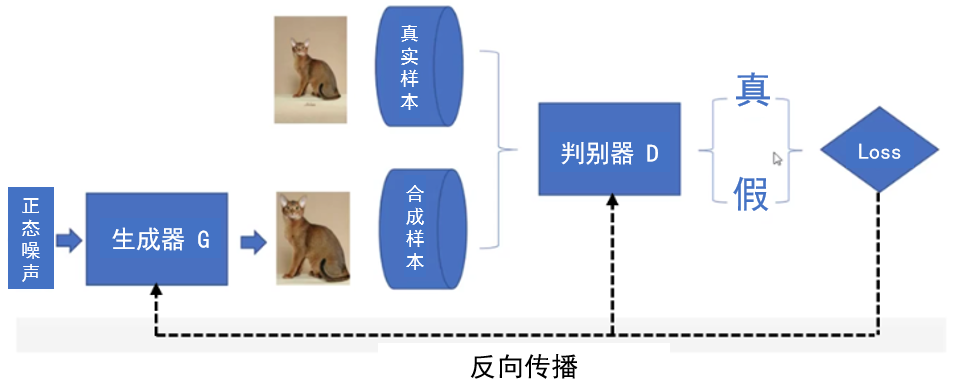


图2.生成对抗网络基本原理。

**2.1.2 FDGAN的工作原理**

图1展示了FDGAN的基本原理和训练流程：

* 噪声输入：生成器 G 从噪声分布中采样（例如标准正态分布，并利用该噪声生成合成样本。生成器尝试学习真实样本的分布，使得生成的样本在数据特征上与真实样本接近。
* 生成样本与真实样本的对比：生成的样本和真实的样本一起输入判别器 D。判别器对输入数据进行判断，输出“真”或“假”的判断。
* 损失反馈与参数更新：通过对判别器的判断结果计算损失，将损失反馈给生成器和判别器，更新它们的参数。生成器试图“欺骗”判别器，让生成样本被判别器判断为“真”；判别器则不断优化，以提高区分真实样本和合成样本的能力。
* 循环迭代：生成器和判别器交替优化，直到生成的样本无法被判别器区分为假样本为止，从而达到数据分布上的一致性。
  + 1. **特征处理与条件变量**

图2展示了生成合成样本的流程和FDGAN对特征的处理方式：

* 条件变量的使用：FDGAN支持在生成过程中利用条件变量，使生成的样本在不同类别上保持一致性。通过引入条件变量，生成器能够更好地模拟不同类别的特征分布。
* 高斯混合模型（VGM）：对于连续特征，FDGAN利用高斯混合模型（VGM）来描述数据的分布，以更准确地生成符合实际特征的合成数据。图中展示了对连续特征的处理方式，包括对特征进行标准化、混合建模等。
* 模式特定归一化：在生成过程中，FDGAN采用模式特定归一化的方法，以确保生成数据与真实数据在模式上相匹配。

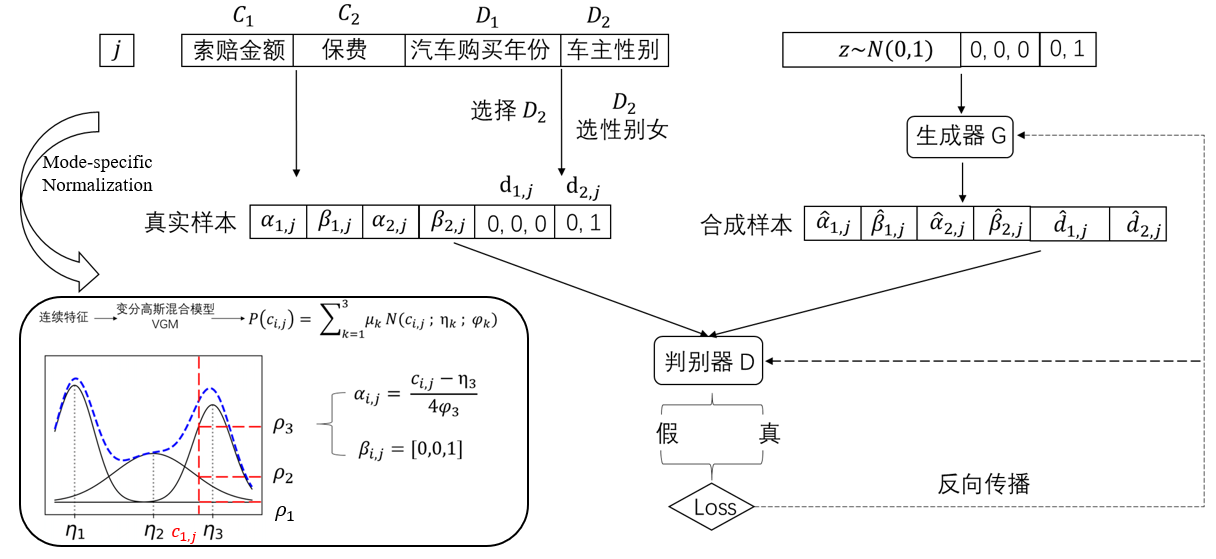


图2. 利用FDGAN生成合成样本。

* 1. **反欺诈分类器设计**
     1. **模型选择**

我们选择了XGBoost作为主要的分类器，并通过五折交叉验证进行评估。XGBoost在处理不平衡数据方面表现优异，适合本项目的反欺诈任务。XGBoost的原理如下。

XGBoost基于梯度提升决策树（Gradient Boosting Decision Tree, GBDT）算法。GBDT是一种集成学习方法，通过逐步构建和叠加决策树模型来提升预测性能。其核心思想是每一棵新树都在前一棵树的基础上，针对前一棵树的残差（即预测误差）进行拟合，从而逐步减小整体模型的误差。

XGBoost通过定义一个目标函数来衡量模型的性能，该目标函数通常包含两部分：

* **训练损失（Loss Function）**：衡量模型预测值与真实值之间的差异，常用的损失函数包括均方误差（MSE）、对数损失等。
* **正则化项（Regularization Term）**：用于控制模型的复杂度，防止过拟合。XGBoost在目标函数中引入了对树的结构进行正则化的项，如叶子节点数和权重的平方和。

XGBoost不仅利用了一阶梯度信息，还利用了二阶梯度（即二阶导数）信息来更精确地优化目标函数。这样可以更快地收敛，并提高模型的预测性能。

XGBoost采用贪心算法构建树，通过在每个节点选择最佳的特征和分裂点来最大化增益（即减少的损失）。为了防止过拟合，XGBoost对树的深度和叶子节点的数目进行了限制，并通过剪枝策略进一步优化树的结构。

* + 1. **XGBoost 的独特性**

（1）**正则化。**与传统的GBDT相比，XGBoost在目标函数中引入了正则化项，这有助于控制模型的复杂度，防止过拟合，提高模型的泛化能力。

（2）**列抽样（Column Subsampling）**。XGBoost支持对特征进行随机抽样，即在构建每棵树时，仅使用部分特征进行分裂。这类似于随机森林中的特征子采样，能够增强模型的多样性，进一步提高泛化能力。

（3）**并行化。**XGBoost通过并行化算法提升了训练速度。具体来说，在树的每一层中，特征的分裂点可以并行计算，从而充分利用多核处理器的计算能力，显著缩短训练时间。

（4）缺失值处理。XGBoost内置了对缺失值的处理机制。它通过在树的分裂过程中自动学习缺失值的最佳分裂方向，无需预先填补缺失值，从而简化了数据预处理流程。

（5）树的分裂方式。XGBoost使用近似算法（如直方图算法）来高效地找到最佳分裂点。这种方法能够在处理大规模数据时显著减少计算量，同时保持高精度。

（6）支持多种目标函数。XGBoost灵活地支持多种目标函数，包括回归、分类、排序等任务。同时，用户也可以自定义目标函数，以满足特定需求。

**2.3 实验设计**

**2.3.1 模型训练与优化**

数据集划分和实验流程如图3所示。

* 数据集分为80%的训练集和20% 的测试集，并进行五折交叉验证。
* 使用生成数据和真实数据共同训练XGBoost模型，优化模型参数以提升分类性能。

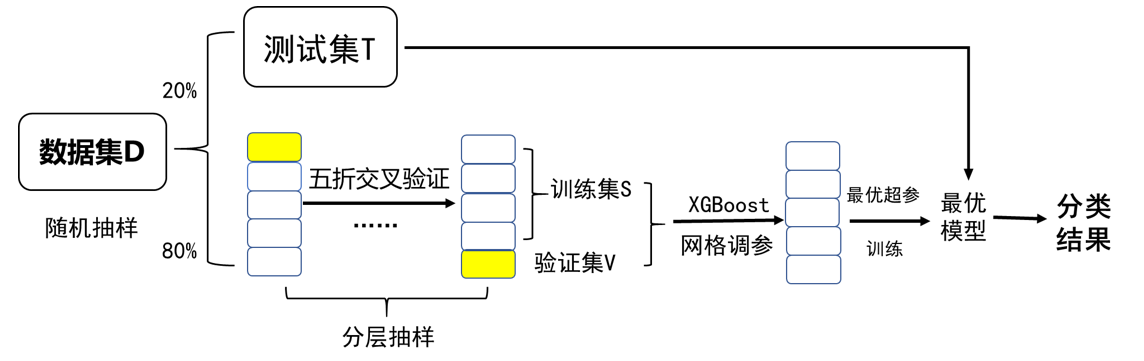


图3.实验流程。

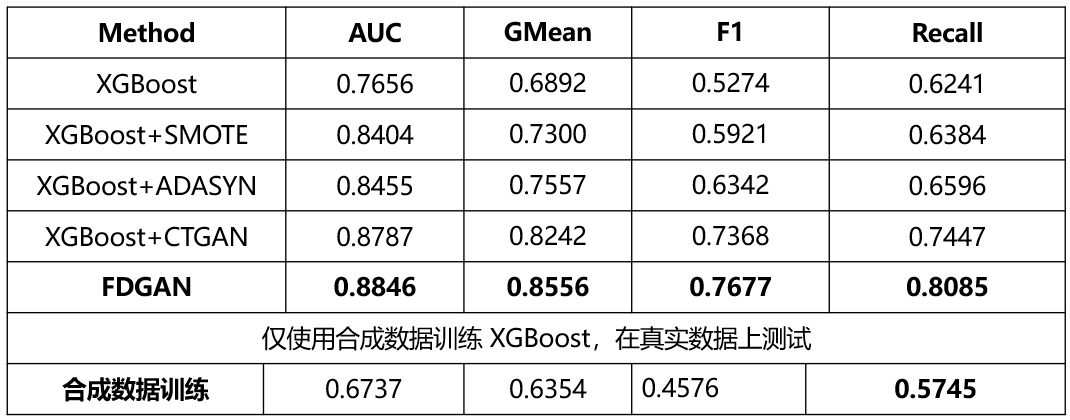
* + 1. **基准对比实验**

我们与传统的过采样方法进行了对比实验：SMOTE和ADASYN，并对不进行过采样的实验也作为基线方法，记为 None。结果表明，使用FDGAN生成的样本进行训练显著提升了模型的召回率（Recall）和F1分数，优于传统的过采样技术。

**2.3.3 实验结果与分析**

实验结果（如表1所示）展示了在不同方法下的模型性能力。使用FDGAN生成的数据进行训练，模型在AUC、GMean、F1和Recall指标上均取得了更高的分数，特别是在召回率 (Recall) 上提升了12.90%。

表1. 实验结果。



**2.4 进一步实验**

为了进一步验证我们提出的方法的有效性。我们仅仅使用合成数据训练 XGBoost，并且使用训练好的 XGBoost 去对相同的真实的测试集样本进行分类，结果如表1所示。仍然获得高达57.45% 的 Recall，验证了生成数据的有效性。进一步证明了本项目方案的优势。

**3.总结与未来展望**

本项目通过生成对抗网络生成欺诈样本，有效解决了数据不平衡问题，并在保险反欺诈任务中取得了显著的效果。未来，我们计划：

（1）探索更多生成模型，如扩散模型，以提升数据生成质量。

（2）结合更复杂的特征工程与特征选择方法，进一步优化模型性能。

（3）对于**团伙化保险欺诈**的研究不够深入，在后续的工作中，尝试**投保人-保险公司**、**投保人-投保人**关系网图，充分利用图的结构信息；