## 研究方法和技术途径

### 研究方法概述

原生成方法+扩散方法移到此处

### 样本生成技术途径

简介

图1‑1 信号发生器图结构

经过一系列调研分析与论证，最终选择了三个适用于项目的具体样本生成方案如下：

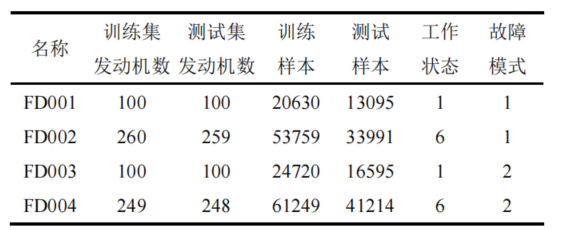
原来 输入5个标签--🡪数据--🡪计算得到值

现在 输入5个标签--🡪深度学习 /得到值

#### 数据集介绍

模拟仿真数据集介绍：

行数、列数、列名、含义



#### 评价指标介绍

为了保证生成数据的质量,将三种度量损失：FID，MMD等加入到生成对抗网络模型中，作为评价指标。

（2）Fréchet Inception Distance

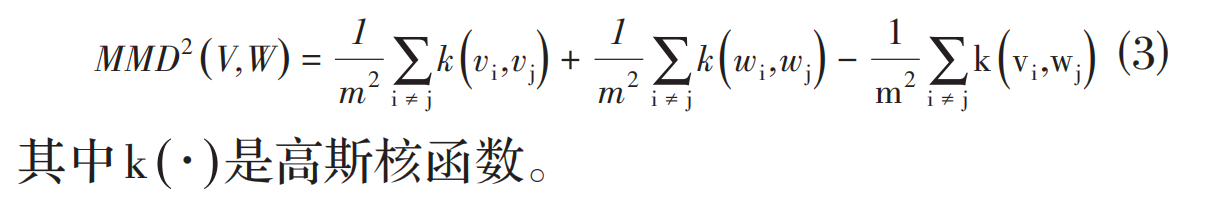
Fréchet Inception Distance(FID)的基本思想是用Inception网络的卷积特征层作为一个特征函数φ，并且用特征函数将真实数据分布Pr和生成数据分布Pg建模为两个多元高斯随机变量。从而计算多元高斯分布的均值μx，μg和方差∑x，∑g。因此，生成信息的质量可以通过下式由两个高斯分布之间的Fréchet距离来计算：

FID度量方式的思想和人类判断是一致的，该评价指标值越小，表示生成的图像越接近真实图像,生成的图片质量越好。FID和生成图像的质量之间有很强的负相关性。

(2) Maximum Mean Discrepancy (MMD)

MMD是一种基于最大均方差的统计检验来优化两类样本的分布，常用于评估生成图像的质量。此处，我们使用MMD衡量生成的结构化数据，定义如下：

给定两类结构化数据集，和，以下为MMD计算公式：



#### 方案一：基于生成对抗网络的海量对抗样本生成方法

由于实际中，数据的获取成本高、数据受限以及数据涉密，获取足够多的真实数据非常困难；真实数据无法涵盖所有可能的可能性的情况和变化；某些类别或情况出现的频率较低，导致模型在训练过程中对这些类别或情况的学习不足。为了克服这种问题，本项目设计一种基于表格数据生成对抗网络的样本生成方法，有针对性地生成更多的样本，以平衡不同类别或情况之间的数据分布，从而提高模型在这些类别或情况下的预测能力，具体模型组成由生成器、鉴别器组成，如图xxx所示。

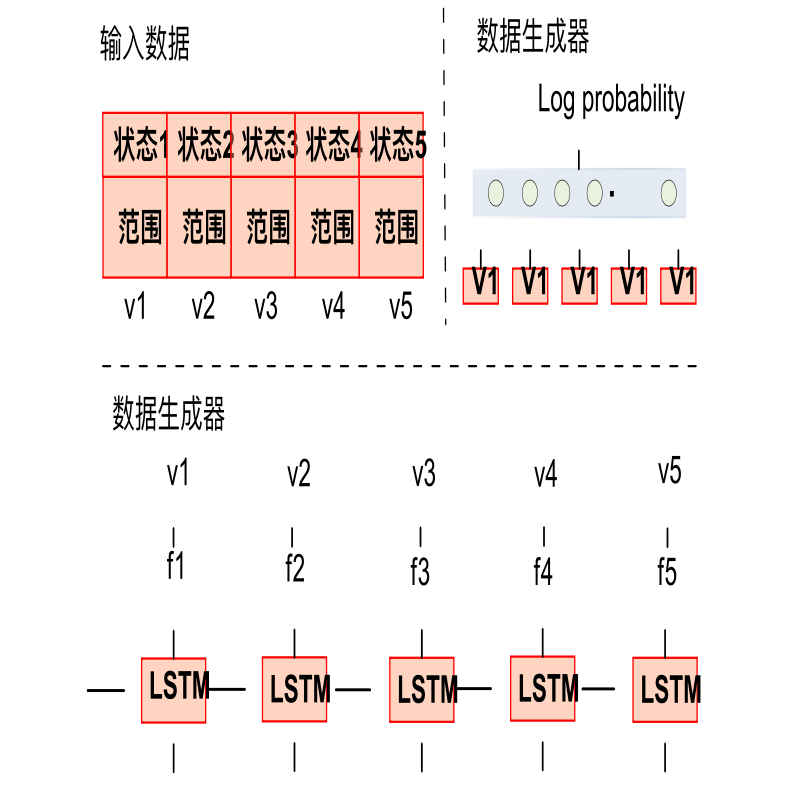


图 基于生成对抗的数据生成模型图

##### 数据预处理：

数据预处理是神经网络应用中至关重要的一步，它直接影响模型的性能和泛化能力。在这一部分，我们将详细介绍对表格数据的预处理过程，包括对连续数值变量和离散分类变量的处理。

首先，针对数值变量，我们希望神经网络能够有效地生成以 (-1, 1) 为中心分布的值。为了达到这个目标，我们采用了 tanh 激活函数，这有助于保持生成数据的平均值接近零。

然而，研究表明神经网络在处理多模态数据时可能存在一些挑战。为了应对这个问题，我们引入了高斯混合模型 (GMM) 来对数值变量进行聚类。具体而言，我们将每个数值变量划分为 C 个聚类，每个聚类有 m (m=5) 个分量。这样，每个数值变量就被表示为 C 维向量，其中每一维对应于一个聚类分量。利用训练好的 GMM 对每个聚类进行归一化，得到最终的 V 向量。这样的处理方式有助于网络更好地利用数值变量的信息，并生成更具多模态特征的数据。

对于分类变量，由于其基数通常较低，我们可以直接使用 softmax 函数生成概率分布。然而，为了更好地处理这些分类变量，我们需要将它们转换为带有噪声的单热编码表示形式，将其表示为二进制变量。这有助于引入一定的随机性，使得生成的数据更具多样性。

最后，将处理后的数值变量 V、聚类概率 U、以及分类变量 D 合并为一个综合的向量，该向量将作为生成器的输出和判别器的输入。需要注意的是，GAN 无法直接访问 GMM 的参数，这样能够增加模型的鲁棒性和泛化性能。

总体而言，通过综合处理连续数值变量和离散分类变量，我们可以更好地为神经网络提供输入，使其更有效地学习数据的特征，提高生成数据的质量和多样性。

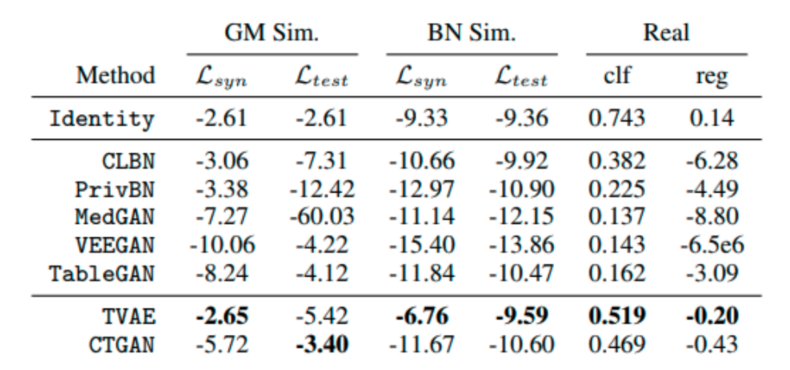
##### 生成器：

生成器 他们分两步生成一个数值变量。 首先，生成值标量 V，然后生成聚类向量 U，最终应用 tanh 使用 softmax 生成所有可能标签上的概率分布分类特征。 为了生成所需的行，使用了带有注意力机制的 LSTM [3]。 每个步骤中 LSTM 的输入是随机变量 z、带有先前隐藏向量和嵌入向量的加权上下文向量。 使用带有 LeakyReLU [10] 和 BatchNorm [5] 的判别器多层感知器 (MLP)。 第一层使用级联向量（V、U、D）以及来自 LSTM 的特征向量的小批量多样性。 损失函数是输入变量的 KL 散度项与序数对数损失函数之和。

##### 鉴别器：

判别器：使用带有 LeakyReLU [10] 和 BatchNorm [5] 的多层感知器（MLP）。 第一层使用级联向量（V、U、D）以及来自 LSTM 的特征向量的小批量多样性。 损失函数是输入变量的 KL 散度项与序数对数损失函数之和。 详细模型结构如图4所示

##### **实验结果分析 ：**

****

#### 方案二：TABDDPM

#### 方案三：GMMN