## 研究方法和技术途径

### 研究方法概述

本项目研究的海量对抗样本生成算法，依照不同的场景，既希望算法体系具备生成电磁信号数据的能力。为了实现这个目的，生成对抗网络和扩散生成网络提供了解决方案。因此，设计可靠高效的生成对抗网络和扩散模型，并将两者有机的结合以形成海量对抗样本生成算法体系，是本项目希望突破的一大关键技术。

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：

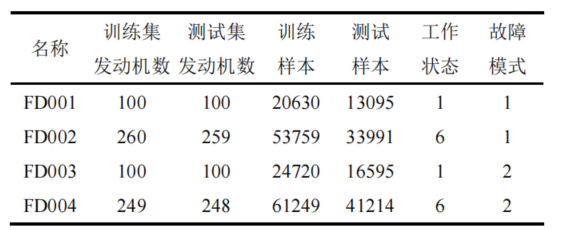
原来 输入5个标签--🡪数据--🡪计算得到值

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：

其中为高斯核函数。

#### 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，具体模型组成由数据生成器、数据鉴别器组成，如图xxx所示。

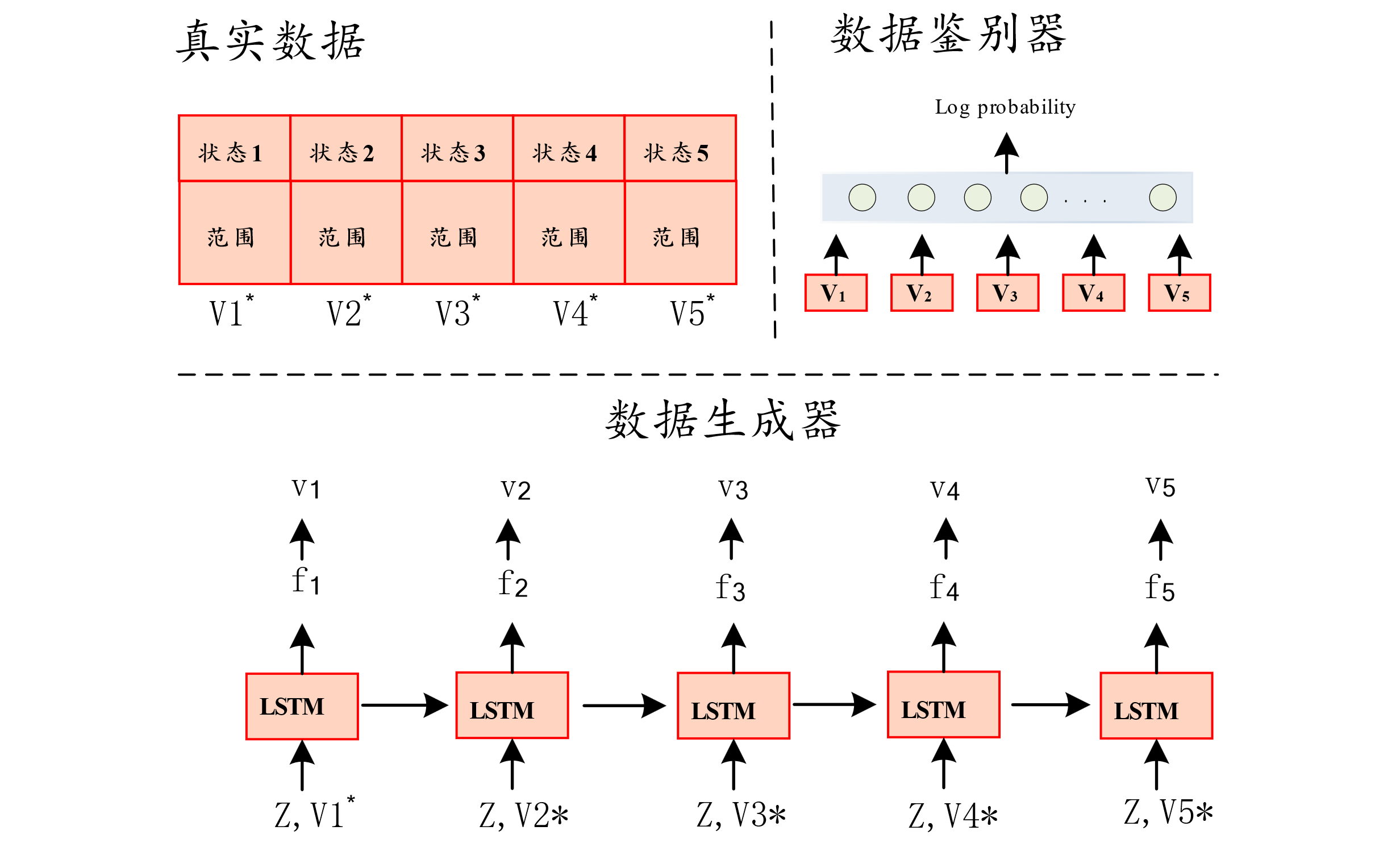


图 基于生成对抗的数据生成模型图

##### 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，具体公式如xx所示。

然而，仿真数据中包含连续数值变量和离散分类变量，为了处理这种多模态数据，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，如图xxx所示。

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。

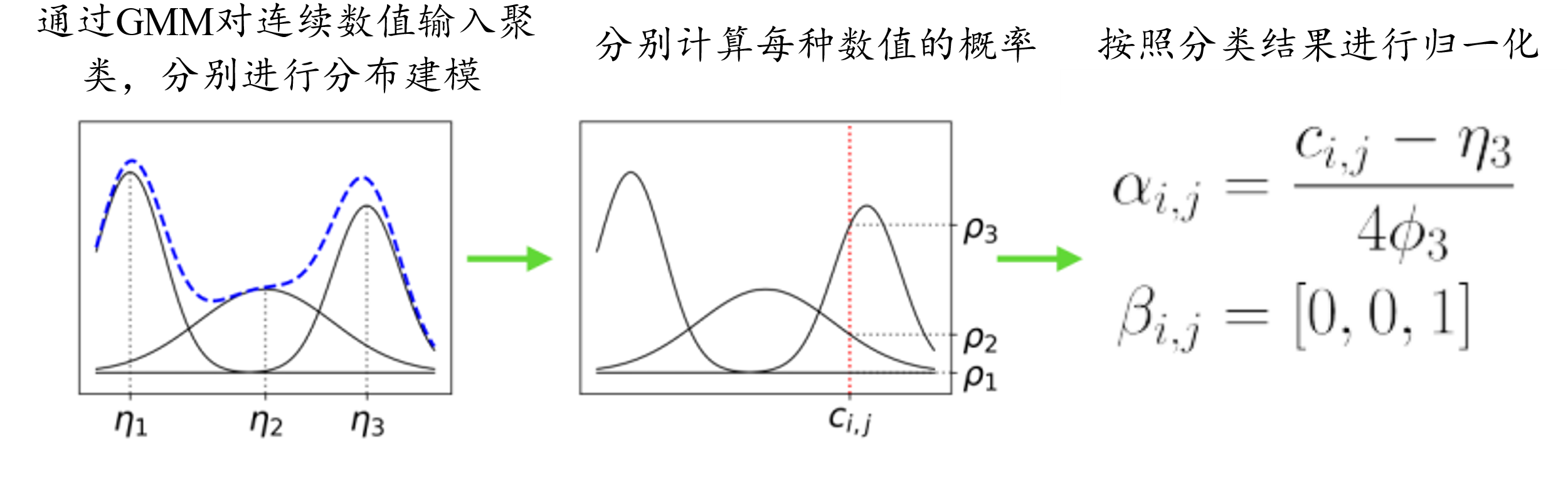


图 基于GMM的分类归一化

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：

海量数据生成器采用基于注意力机制的LSTM作为特征提取器，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v。在线性映射层中，使用带有LeakyReLU和BatchNorm的多层感知器 (MLP)。 损失函数为输入变量的 KL 散度项，具体公式如下所示。

其中P(x)和Q(x)为两个离散的概率分布，x表示分布的取值。X是所有可能的x的集合，log可以是以任意正数为底数的对数函数。且KL散度不具备对称性，即 KL(P||Q)与 KL(Q||P)的值可能不同。

##### 鉴别器：

海量数据鉴别器实现真实数据与生成数据之间的分类任务，结构较简单，由带有 LeakyReLU和BatchNorm的多层感知器 (MLP)实现，并采用数据生成器输出的级联向量v作为输入； KL 散度项作为损失函数。 详细模型结构如图xx所示

##### 实验结果分析 ：

本文通过半实物仿真数据集分别验证了所提生成对抗方法和扩散生成方法的有效性，如表x所示。

表1　 生成对抗方法对比实验结果

Table 1　 Comparative experimental results

|  |  |  |
| --- | --- | --- |
| Method | MMD | FID |
| cGAN | 0.9931 | 29.30 |
| BiCoGAN | 0.9916 | 25.60 |
| IcGAN | 0.9736 | 24.80 |
| MADGAN | 0.9527 | 21.70 |
| DropoutGAN | 0.9246 | 21.61 |
| EBGAN | 0.9212 | 19.42 |
| TripleGAN | 0.9031 | 18.78 |
| SGAN | 0.9002 | 18.34 |
| TGAN | 0.8901 | 18.23 |

表2　 扩散生成方法对比实验结果

Table 1　 Comparative experimental results

|  |  |  |
| --- | --- | --- |
| Method | MMD | FID |
| SDEs | 0.9747 | 23.46 |
| SGMs | 0.9624 | 23.05 |
| DDPM | 0.9432 | 21.12 |

由表x和表x的实验结果表明，在生成对抗方法中，TGAN通过对表格数据的特殊处理，实现了更好的数据生成效果；扩散生成方法中，DDPM方法对离散列和连续列分开考虑，符合数据特性且更具可解释性。

#### 方案二：TABDDPM

#### 方案三：GMMN