## 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：

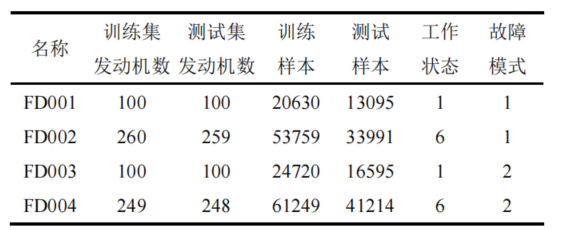
原来 输入5个标签--🡪数据--🡪计算得到值

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：

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。

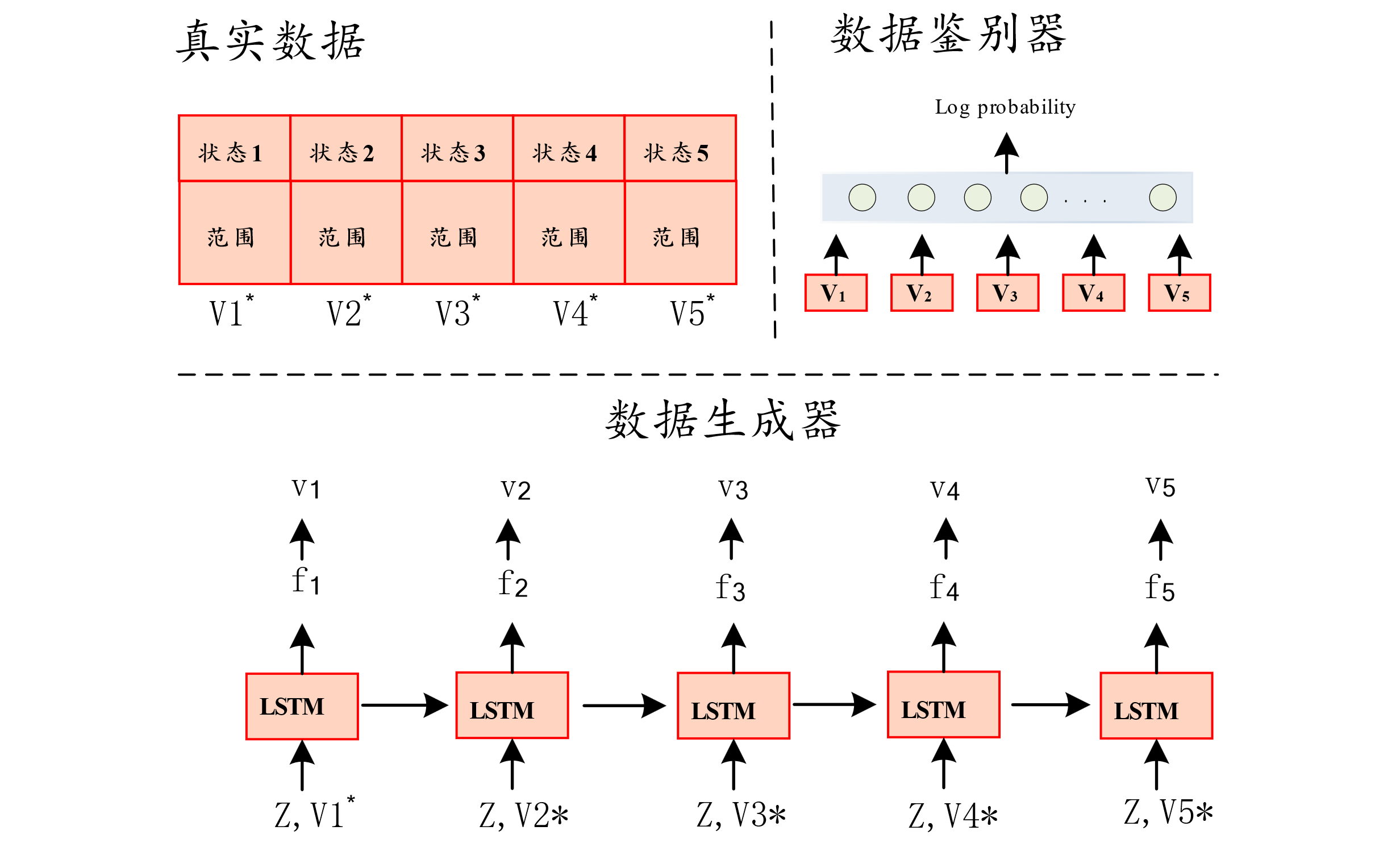


图 基于生成对抗的数据生成模型图

##### 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。

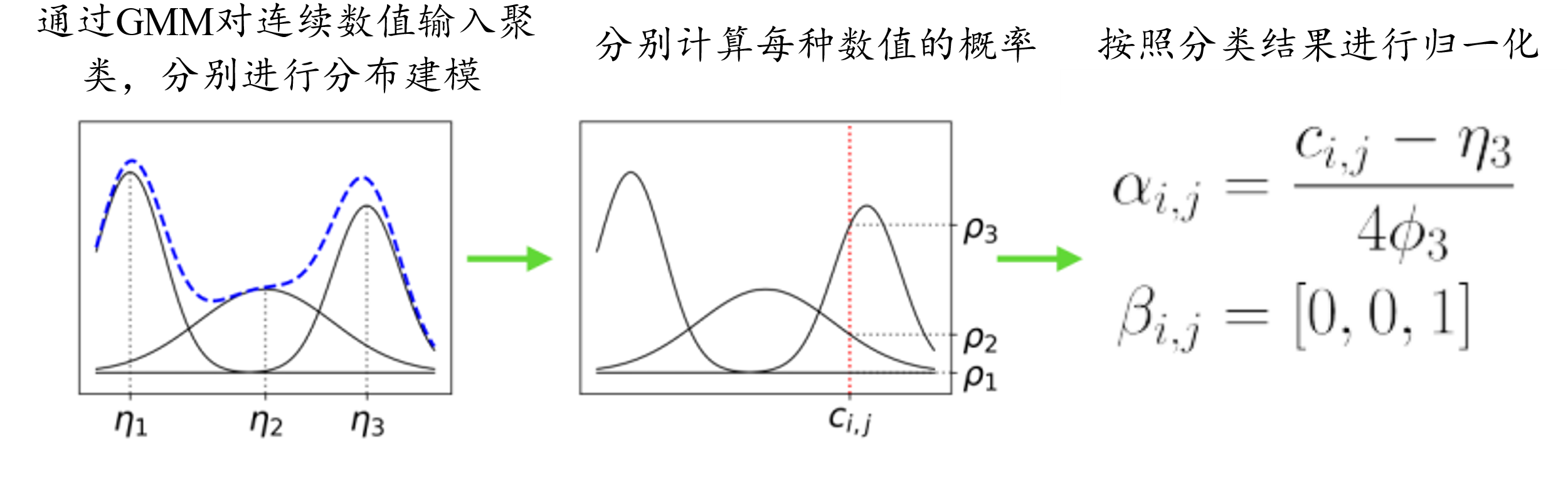


图 基于GMM的分类归一化

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 ：

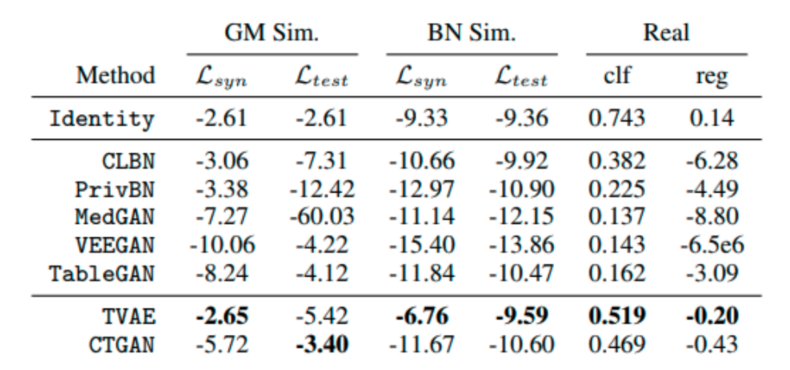
****

表1　 对比实验结果（RMSE）

Table 1　 Comparative experimental results（RMSE）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | Source  only | DANN | ADDA | CADA | WADA | TADA |
| FD001→FD002 | 43.25 | 48.62 | 19.87 | 19.52 | 18.60 | **15.15** |
| FD001→FD003 | 29.29 | 45.87 | 39.74 | 39.58 | **19.13** | 19.59 |
| FD001→FD004 | 47.27 | 43.82 | 31.78 | 31.23 | 21.93 | **16.79** |
| FD002→FD001 | 32.79 | 28.10 | 14.33 | 13.88 | 15.09 | **12.38** |
| FD002→FD003 | 34.50 | 37.46 | 32.60 | 33.53 | 19.13 | **17.14** |
| FD002→FD004 | 14.99 | 31.85 | 34.35 | 33.71 | 23.38 | **13.03** |
| FD003→FD001 | 11.66 | 31.74 | 19.97 | 19.54 | 17.64 | **11.59** |
| FD003→FD002 | 36.34 | 44.62 | 23.47 | 19.33 | **16.38** | 16.96 |
| FD003→FD004 | 37.31 | 47.94 | 26.33 | 20.61 | 19.49 | **17.47** |
| FD004→FD001 | 20.17 | 31.54 | 37.89 | 20.1 | 19.15 | **16.11** |
| FD004→FD002 | 14.74 | 24,93 | 28.77 | 18.5 | 20.30 | **11.49** |
| FD004→FD003 | 29.98 | 27.84 | **14.14** | 14.49 | 14.86 | 16.02 |
| Mean | 29.36 | 37.01 | 26.93 | 23.669 | 18.76 | **15.31** |

#### 方案二：TABDDPM

#### 方案三：GMMN