

神经网络模型报告文档 4.10

许敬然

网络基本情况说明

网络模型的目的是替代传统优化手段，更快速地完成参数反演。在已知血药浓度的时间离散点时，利用该模型可以快速得到一个机器学习意义上最可能与之相对应的三个参数 ($DSC <$ 角质层中的有效扩散系数 $>$ 、 $PFO <$ 毛囊的渗透系数 $>$ 、 $u1 <$ 由脱屑而向皮肤表面转移的速度 $>$)。从接收输入到输出结果，网络模型的工作时间大概为 $\alpha\alpha\alpha\alpha$ 待填补 $\alpha\alpha\alpha\alpha$ 。

网络的输入是形状为 $N \times 68$ 的二维数组，其中“ N ”代表输入数据的条目数，“68”代表 68 个时间采样点。输入数据代表的是抽血采样得到的血药浓度。网络的输出是形状为 $N \times 3$ 的二维数组，其中“ N ”代表输出数据的条目数，“3”代表 3 个反演参数。网络的架构是由全连接层和 LayerNormalization 层构成的共 30 层的残差网络。

在网络训练中，数据集的标签 (label) 是经过各种方式在参数空间的合理范围内取得的三参数组，数据集的特征 (feature) 是将标签中的三参数代入至 BPTK 模型计算得到的血药浓度曲线在 68 个时间点的取样。(取样规则：从 0.5 时至 20 时结束，每 0.5 小时取一次样；从 20.5 时至 74.5 时结束，每 2 小时取一次样。总共得到 68 个时间取样点。) 损失函数设置为标签三参数和模型输出的三参数的 MSE。

评价指标说明

为了更好地评价输出结果的拟合程度，除了损失函数外，还使用了能度量“原始标签参数代入至 BPTK 模型计算得到的血药浓度曲线”和“模型输出参数代入至 BPTK 模型计算得到的血药浓度曲线”之间差异的评价指标 (简称两个曲线分别为原始曲线与模型曲线)。

1. 标签参数与网络输出参数的 MSE

简称为 **loss**。该指标沿用了模型训练中的损失函数，即单组标签三参数和单组模型输出三参数的 MSE (若有多组，取平均)。

2. 标签参数与网络输出参数的相对误差数组的平均

简称为**标签 MRE**。该指标是一个含有三个元素的一维数组，是标签参数逐元素与网络输出参数计算相对误差的结果 (若有多组，取平均)。

3. 原始曲线与模型曲线之间的绝对误差的平均值

简称为 **MAE**。单组的原始曲线与模型曲线本质上是两个含有 15000 个元素的一维数组 (共 15000 个时间戳)，该指标即为两个数组之间绝对误差的平均 (若有多组，取平均)。

4. 原始曲线与模型曲线之间的相对误差的平均值

简称为 **MRE**。类似上一条，该指标即为两个数组之间相对误差的平均值 (若有多组，取平均)。但该指标不稳定，在其他指标描述的效果很好的情况下，该指标可能会很高。如图1，从可视化角度看，两条曲线是极为接近的，但是 MRE 却有 49.775%。原因可能是在计算相对误差时，分母的数值十分小，导致了相对误差指标描述能力的失真，尤其是在曲线的末端，血药浓度接近于 0，失真效应就更强了。

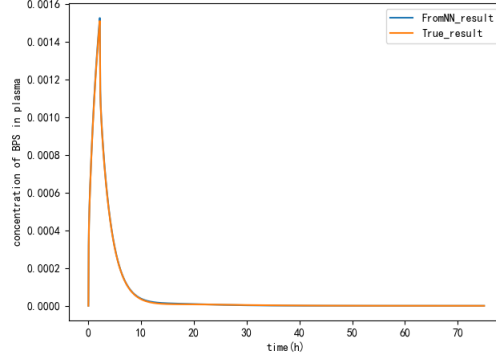


图 1: 某组数据集外的三参数生成的原始曲线与模型曲线

5. 决定系数 R^2

简称为 R^2 。该指标原用于评价线性回归模型的拟合程度，但用于非线性的拟合模型时也可提供一个拟合程度的大致度量。 $R^2 \in (0, 1]$ ，越接近 1 说明拟合程度越好 (若有多组，取平均)。它的公式是 $R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS}$ ，其中 $ESS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 、 $RSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 。

6. 原始曲线与模型曲线之间修正后的绝对误差的平均值

简称为**修正 MAE**。不同的参数计算出的浓度曲线的绝对数值是不同的，很难从绝对误差的大小上得到确切的拟合效果评价，故对绝对误差做一次“标准化”，除以一个 $scaling$ 。设置 $scaling = \frac{\sum_{j=1}^N \sum_{i=1}^{15000} C_{t_i}^{TrueLine(j)}}{15000 \times N}$ ，其中 15000 代表的是 15000 个时间戳 ($75 \div 0.005$)，N 代表的是 N 组不同的原始曲线， $scaling$ 代表的是 N 条原始曲线的平均线的平均值。

7. 平均原始曲线与平均模型曲线之间的相对误差的平均值

简称为**平均线 MRE**。由于在评价模型时输入不止一组数据集外的标签数据，考虑将不同组的三参数得到的原始曲线合成为一条平均线，得到的模型曲线也合成为一条平均线 (相加后除以数量)，分别称为平均原始曲线与平均模型曲线。计算平均原始曲线和平均模型曲线之间相对误差的平均即为该指标。

8. 修正平均原始曲线与修正平均模型曲线之间的相对误差的平均值

简称为**修正平均线 MRE**。简单地将曲线叠加并平均再一齐计算误差仍不妥，因为不同曲线之间的绝对数值之间有差异。为了减小由绝对数值的差异带来的误差，现设置修正平均曲线。 $\forall i = 1, 2, \dots, N$ ，将第 i 组的原始曲线与模型曲线除以一个 $scaling^i$ ， $i = 1, 2, \dots, N$ ，其中 $scaling^i = \frac{\sum_{j=1}^{15000} C_{t_j}^{TrueLine(i)}}{15000}$ 。 $scaling^i$ 代表的是原始曲线的平均值。之后再除以 $scaling^i$ 后的各组原始曲线叠加平均，各组模型曲

线也叠加平均，分别得到修正平均原始曲线与修正平均模型曲线，计算二者之间的 MRE，即可得到该指标。

数据集说明

数据集的构建需要在三维参数空间中的合理范围内取样，得到若干组三参数，即为数据集的标签。将标签代入至 BPTK 模型求解，得到若干组血药浓度曲线，对曲线再取样，即得到数据集的特征。参数空间内的取样是重要的一步。合适的取样在模型训练过程中能起到事半功倍的效果，不合适的取样可能导致网络模型仅在某个参数范围内拟合效果好，离开这个范围后的效果可能很差。以下是参数空间取样时使用过的几个方法与具体取样情况。

1. 等距取样

分别对三个参数独立地取样，再利用笛卡尔乘积得到三参数组合。根据参考资料，为每个参数设置一个合理的一维取样区间。将三个参数的取值范围界定为： $DSC_0 \in (10, 40)$, $PFO_0 \in (0, 15)$, $u_1^0 \in (0, 15)$, ($DSC_0 \triangleq DSC \times 10^{-9}$, $PFO_0 \triangleq PFO \times 10^{-5}$, $u_1^0 \triangleq u_1 \times 10^{-5}$)。又考虑到三个参数在这个取值范围内仍有一个更合理的取值范围，设置参数在更合理的区间内频次更高地等距取点。 DSC_0 被设置了三个取样区间，在 $(15, 20)$ 中 (严格地) 每隔 0.5 取一次点，在 $(10, 15) \cup (20, 30)$ 中每隔 1.2 取一次点，在 $(30, 40)$ 中每隔 4 取一次点，在后两个区间的取点间隔不一定严格遵守，而是根据实际情况进行了更合适的调整。 PFO_0 被设置了两个取样区间，在 $(2, 8)$ 中 (严格地) 每隔 0.4 取一次点，在 $(0, 2) \cup (8, 15)$ 中每隔 1.8 取一次点。 u_1^0 被设置了两个取样区间，在 $(3, 8)$ 中 (严格地) 每隔 0.4 取一次点，在 $(0, 3) \cup (8, 15)$ 中每隔 $\frac{4}{3}$ 取一次点。 DSC_0 共取样 25 个， PFO_0 与 u_1^0 各取样 20 个，取三个一维数组的笛卡尔积后便可以得到 10000 组不同的三参数组合。靠等距取样的方式得到了 10000 组数据。

2. 截断正态分布取样

分别对三个参数独立地取样，再利用笛卡尔乘积得到三参数组合。在参考资料中， DSC_0 的平均值为 17.28， PFO_0 的平均值为 6.39， u_1^0 的平均值为 5.7。它们的变异系数 CV(Coefficient of Variation) 统一被设置为了 30%，相应地，标准偏差由给出 $\sigma = CV \times \mu$ 。资料内的截断正态分布设定的左截断点为 $l = \mu - 1.96 \times \sigma$ ，右截断点为 $u = \mu + 1.96 \times \sigma$ 。表格1描述了三个参数的分布情况。

分布情况	平均值	标准差	左截断	右截断
三参数				
DSC_0	17.28	5.184	7.12	27.44
PFO_0	6.39	1.917	2.63	10.14
u_1^0	5.7	1.71	2.35	9.05

表 1: 三参数的截断正态分布情况

我对原分布做了一点改动，使得后两个参数的标准差和取样范围更大，改动后的分布情况如表格2所示。

DSC_0 共取样 32 个， PFO_0 与 u_1^0 各取样 28 个，取三个一维数组的笛卡尔积后便可以得到 25088 组不同的三参数组合。靠等距取样的方式得到了 25088 组数据。

分布情况	平均值	标准差	左截断	右截断
三参数				
DSC_0	17.28	4	9.28	27.68
PFO_0	6.39	4	2.39	11.99
$u1_0$	5.7	4	1.7	11.3

表 2: 改动后的三参数的截断正态分布情况

3.SparseGrids 取样

SparseGrids(稀疏网络) 是一种高效的数值离散化技术, 它在多变量问题的求解中尤其有用。这种技术通过特殊的张量积构造和后续的截断来创建多维多层次基, 从而允许在保持一定精度的同时显著减少所需的离散点数量。将 SparseGrids 用于参数空间取样, 可以在减少采样点的同时增加数据集的均匀性, 因为稀疏网络的结构要比各维度等距取样后乘积更优秀。

设置取样空间为 $[10.28, 24.28] \times [1.39, 11.39] \times [1.2, 10.2]$, 复杂度等级为 9, 意为在每个维度的区间上等距设置 $2^9 + 1$ 个格点 (包括区间端点), 之后按照 SparseGrids 格点的生成规则, 在这个取样空间内共生成 18943 组参数。靠稀疏网络取样的方式得到了 18943 组数据。

使用不同数据集训练出的较优模型的效果展示

在数据集之外, 使用截断正态分布取样的方式取得 DSC_0 共 6 个, PFO_0 与 u_1^0 各 5 个, 总共 150 个三参数组合。分布的截断范围比上述数据集都要大, 其分布信息如表格3

分布情况	平均值	标准差	左截断	右截断
三参数				
DSC_0	17.28	5	7.28	28.78
PFO_0	6.39	5	1.39	12.89
$u1_0$	5.7	5	0.7	11.7

表 3: 用于测试的三参数的截断正态分布情况