文章编号：2095-2163(2021)03-0185-05 中图分类号：TE621 文献标志码：A

基于 XGBoost 和神经网络拟合预测模型的辛烷值损失的预测

朱怡欣
（上海工程技术大学 管理学院，上海 201620）

摘 要：汽油清洁化重点是降低汽油中的硫、烯烃含量，同时尽量保持其辛烷值。降低辛烷值（RON）损失是国内车用汽油质量升级的主要目标之一。本文针对某石化企业的催化裂化汽油精制脱硫装置运行收集的数据进行处理，探求数据样本中变量与变量本身、其他自变量及目标变量等之间的相关性，对特征变量进行多阶段降维，进而通过 XGBoost 和 LSTM 循环神经网络对汽油辛烷值损失进行建模，通过对预测结果的统计表明该方法在企业辛烷值损失预测中具有较好的表现，为国内车用汽油技术升级提供一定的指导作用。

关键词：辛烷值损失；XGBoost；LSTM

Prediction of octane loss based on XGBoost and neural network fitting prediction models

ZHU Yixin
（School of Management Studies, Shanghai University of Engineering Science, Shanghai 201620, China）

【Abstract】 The focus of gasoline cleaning is to reduce the sulfur and olefin content in gasoline while maintaining its octane number as much as possible. Reducing the loss of octane number (RON) is one of the main goals of China´s automotive gasoline quality upgrade. This paper deals with the data collected from the operation of the catalytic cracking gasoline refinery desulfurization unit of a petrochemical company, and explores the correlation between the variables in the data sample and the variables themselves, other independent variables and target variables, and performs multi-stage dimensionality reduction on the characteristic variables. Furthermore, the gasoline octane loss is modeled by XGBoost and LSTM recurrent neural network, and the statistics of the prediction results show that this method has a good performance in the prediction of enterprise octane loss, which provides a certain guidance for the upgrading of China´s automotive gasoline technology.

【Key words】 octane loss; XGBoost; LSTM

0 引 言

汽油作为小型汽车的主要燃料，其燃烧产生的尾气对环境造成了恶劣的影响。世界各国都制定非常严格的汽油标准。随着国内经济的迅速发展，汽车保有量在持续增长，汽油的需求量也在逐年加大。因此国内大力发展了以催化裂化为核心的重油轻质化工艺技术，充分利用了原油中的重油资源，将重油转化为汽油、柴油和低碳烯烃。超过 70%的汽油是由催化裂化生产得到，因此成品汽油中 95%以上的硫和烯烃来自催化裂化汽油。辛烷值是反映汽油燃烧性能的重要指标，辛烷值每降低 1 个单位，相当于损失约 150 元/吨。以一个 100 万吨/年催化裂化汽油精制装置为例，若能降低 RON 损失 0.5 个单位，其经济效益将达到 7 500 万元。

1 研究方法与基本假设

1.1 研究方法

首先，研究根据样本针对不良数据进行预处理，筛选出无关紧要的操作变量，辅助找出主要变量，其次，对预处理数据进行多阶段降维，利用了 Embedded Feature Selection 筛选出在建模过程中贡献度比较高的变量（特征），作为最终建模变量。经过多阶段特征降维后，遴选出了 30 个变量作为影响最终结果的自变量，选出了产品性质中的硫含量、辛烷值损失作为因变量，上述的 32 个变量用于对辛烷值损失的模型的建立和求解。在经过 325 个训练样本的训练后，根据 XGBoost 和神经网络两个拟合预测模型不同的预测能力，使用加权打分的形式进行组合，对最终的辛烷值损失进行预测。

1.2 基本假设

假设各训练样本之间相互独立，不存在强耦合的关系。

假设各样本内容虽然与真实环境存在一定误差，但不影响最终结果。

假设在预处理阶段剔除的变量，对最终结果的预测不会产生方向性错误。

作者简介：朱怡欣（1996-），女，硕士研究生，主要研究方向：企业战略管理。

收稿日期：2020-11-03

2 XGBoost 和神经网络拟合预测模型

2.1 多阶段特征降维

由于原始数据变量较多,工程技术应用中经常需要先降维,这有利于忽略次要因素,发现并分析影响模型的主要变量。所以,文中对预处理后的数据进行了多阶段降维,充分考虑到了多方面因素进行变量的选择。

首先,是业务逻辑降维。根据业务逻辑可以知道辛烷值(RON)损失是原料辛烷值与产品辛烷值的差值,所以在给定原料辛烷值的情况下,就不再将产品辛烷值作为建模特征,否则会出现信息泄露问题。随后,是标准化降维,利用样本数据预处理结果,已经删除了一部分变量。然后,是自变量间相关性降维,考虑到各变量之间的相关性,进行变量的两两比较,删除高度相关的变量,保留高度相关的其中一个变量即可,这样有利于减少变量维度并且降低变量之间的耦合性。其次,是目标变量与自变量间相关性降维,考虑到辛烷值 RON 损失作为因变量,其与剩余所有变量的相关性,故对辛烷值 RON 损失以及其余所有变量进行了两两相关性计算,有利于剔除与目标变量无关的变量,最大限度地保留对目标变量有意义的变量。接下来,是方差降维。考虑到变量自身的有效性,对变量进行了方差检验,剔除了方差小于 0.1 的变量,方差越小,表示该变量无法有效地去表征目标变量,在后续建立模型中会产生较大的影响。最后,再利用 Embedded Feature Selection 筛选出在建模过程中贡献度比较高的变量(特征),作为最终建模变量。最终保留变量如图 1 所示。

S-ZORB.DT_2001.DACA	S-ZORB.TC_5005.PV	S-reborn(S-再生吸附剂性质)
S-ZORB.PDT_1003.DACA	S-ZORB.PDT_2604.PV	S-ZORB.DT_2107.DACA
S-ZORB.FT_9001.PV	S-ZORB.SIS_TE_2802	S-ZORB.FT_1301.DACA
S-ZORB.FT_9301.PV	S-ZORB.TE_1001.PV	S-ZORB.LT_9101.DACA
S-ZORB.TE_5202.PV	S-ZORB.LC_5002.DACA	S-ZORB.PDT_2503.DACA
S-ZORB.FC_1203.PV	olefin(烯烃)	S-ZORB.FT_9403.PV
S-ZORB.TE_5201.DACA	S-ZORB.TE_1102.DACA	S-ZORB.FT_9302.PV
S-ZORB.3801.DACA	S-ZORB.FT_9002.DACA	S-ZORB.TE_5006.DACA
S-ZORB.PDT_2906.DACA	S-ZORB.AT_1001.DACA	S-target(硫含量)
RON-source(辛烷值 RON)	S-ZORB.FT_1003.PV	
S-ZORB.CAL.CANGLIANG.PV	S-ZORB.PT_2901.DACA	

图 1 最终保留变量

Fig. 1 Final retained variables

2.2 多模型融合下的辛烷值损失预测模型

在预测模型中,需要指出的是,目标变量是辛烷值损失值,而不是产品性质中的辛烷值。从上述的相关性分析中,可以得到产品性质中的辛烷值与原料性质中的辛烷值具有高度相关性,如果利用产品性质中的辛烷值作为目标变量,对于结果而言会存在一定的作弊行为。

经由多阶段特征降维处理后得到 30 个主要变量,本次研究将其认定影响最终结果的自变量,其中每个变量含有 325 个数据。进一步地,选取产品性质中的辛烷值损失作为因变量,同样含有 325 个数据。通过对这含有 31 个变量的 325 组数据构建模型,对处于不同操作条件(30 个主要变量的不同取值)下的辛烷值损失进行预测。

2.2.1 *RMSE* 和 *MAE* 指标介绍

假设存在一个训练集 $Train = \{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n),\cdots,(x_N,y_N)\}$,这里,$N$ 为训练样本总数,$n = 1,2,\cdots,N$;其测试集 $Test = \{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m),\cdots,(x_M,y_M)\}$,其中,$M$ 为训练样本总数,$m = 1,2,\cdots,M$。若经由训练模型 $f(x)$ 得到预测值 $\hat{y} = \{(\hat{y}_1),(\hat{y}_2),\cdots,(\hat{y}_m),\cdots,(\hat{y}_M)\}$,则均方根误差(*RMSE*)为真实值与预测值差平方的期望的平方根,其对应数学公式可写为:

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(y_m - \hat{y}_m)^2}, \tag{1}$$

RMSE 函数一般用来检测模型的预测值和真实值之间的偏差。*RMSE* 值越大,表明预测效果越差。

平均绝对误差(Mean Absolute Error, *MAE*),即误差绝对值的平均值,可以准确反映实际预测误差的大小,其对应数学公式可写为:

$$MAE = \frac{1}{M}\sum_{m=1}^{M} | y_m - \hat{y}_m | . \quad (2)$$

MAE 评估的是真实值和预测值的偏离程度，即预测误差的实际大小。*MAE* 值越小，说明模型质量越好，预测越准确。

2.2.2　辛烷值损失预测模型的建立-XGBoost

XGBoost（eXtreme Gradient Boosting）作为一种对多棵决策树进行集成学习的算法，其中的决策树之间具备一定的关联关系，这和随机森林有极大的不同。XGBoost 模型中，每棵决策树都是对前面所有决策树的预测结果之和与真实值的残差，其算法过程如下：

（1）假设原始训练集含有的样本数为 N[1-2]，随机且有放回地从原始训练集中抽取 n 个训练样本，并将其作为第一棵决策树的训练集。

（2）设定每个训练样本的特征数都为 M，随机从中抽取 m 个特征，并将其作为决策树选择最优划分特征的特征集合。

（3）利用这 n 个训练样本和 m 个特征构建第一棵决策树，得到第一棵树预测值。

（4）将第一棵决策树的预测值与真实值之间的残差作为第二棵树的输入值得到第二棵决策树的预测值[2]。

（5）重复地将第一棵树与第 $K-1$ 棵树之间的预测结果之和与真实值之间的残差作为第 K 棵树的输入值[2]，实验循环至达到项目停止的条件，最终得到 K 棵决策树，即 XGBoost。

（6）利用 XGBoost 对测试集进行预测得到最终预测结果，即 K 棵决策树的预测结果之和。

在 XGBoost 回归模型中，样本 D_i 的最终预测值为各棵决策树对该样本的预测结果之和[2]，如式（3）所示：

$$\hat{D}_i^{31} = \sum_{k=1}^{K} f_k(D_i^{'}) , \quad (3)$$

其中，K 为决策树的总数，f_k 为第 k 棵决策树。

XGBoost 模型的构建流程如图 2 所示，当决策树 1 加入到模型中时：

$$\hat{D}_i^{31(1)} = f_1(D_i^{'}) , \quad (4)$$

当决策树 2 加入到模型中时：

$$\hat{D}_i^{31(2)} = f_1(D_i^{'}) + f_2(D_i^{'}) = \hat{D}_i^{31(1)} + f_2(D_i^{'}) , \quad (5)$$

当决策树 $K-1$ 加入到模型中时：

$$\hat{D}_i^{31(K-1)} = \hat{D}_i^{31(K-2)} + f_{K-1}(D_i^{'}) , \quad (6)$$

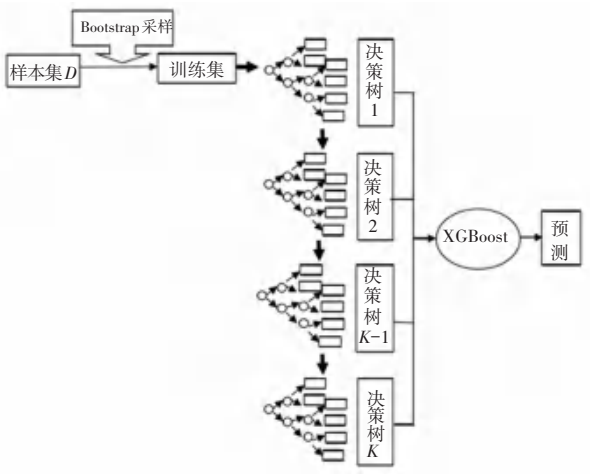

图 2　XGBoost 模型的构建流程图

Fig. 2　Construction flow chart of XGBoost model

当决策树 K 加入到模型中时：

$$\hat{D}_i^{31(K)} = \hat{D}_i^{31(K-1)} + f_K(D_i^{'}) , \quad (7)$$

XGBoost 回归模型的目标函数表示为：

$$\sum_{i=1}^{325} l(D_i^{31}, \hat{D}_i^{31}) + \sum_{k=1}^{K} \varphi(f_k) , \quad (8)$$

$$\varphi(f_k) = \gamma T + \frac{1}{2}\delta \sum_{t=1}^{T} w_t^2 . \quad (9)$$

其中，T 为第 K 棵决策树的叶子节点总数；w_t 为第 K 棵树的第 t 个叶子节点的预测值[2]；γ 和 δ 分别表示对这两部分的重视程度。

公式（8）表示在欠拟合和过拟合之间寻求平衡。其中，第一部分表示全部样本的真实值以及预测值的残差函数，该值越小，欠拟合的概率越低；第二部分表示正则化惩罚项，该值越大，过拟合可能性就越大，因此将该部分尽可能缩小化，可以使最终模型更加简单，具有更强的泛化能力。同时，XGBoost 中每个叶子节点的预测值是根据贪心策略，通过最优化目标函数求出。

2.2.3　辛烷值损失预测模型的建立-LSTM

循环神经网络 RNN 与传统神经网络不同的是，RNN 通过保存当前隐藏层的信息，并通过隐藏层之间的连接将信息传递到下一时刻的隐藏层[1]，赋予网络“记忆”属性，如图 3 所示。但 RNN 网络在反向传播的情况下，对模型的线性关系参数具有长期依赖性[1]，序列过长往往伴随着梯度消失，网络参数过大等条件将进一步导致梯度爆炸。

LSTM 模型是 RNN 模型的一种衍生，是为了避免 RNN 存在的长期依赖性问题，LSTM 网络利用时间进行反向传播训练，解决了梯度消失问题。LSTM 的具体结构如图 4 所示。图 4 中，h_{t-1} 是上一层的

输出，C_{t-1} 是上一个 LSTM 结构的数据信息，h_t 是该层的输出，C_t 是该 LSTM 结构的数据信息。

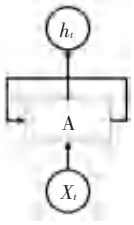

图 3 RNN 模型的简单架构图

Fig. 3 Simple architecture diagram of RNN model

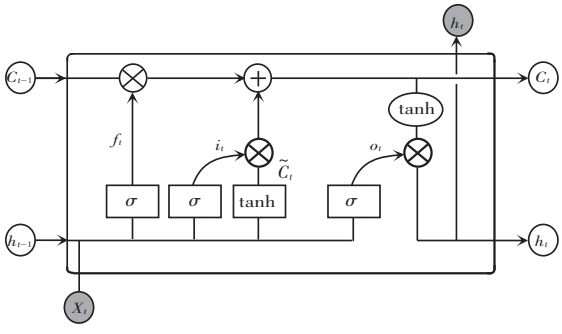

图 4 LSTM 的具体结构

Fig. 4 The specific structure of LSTM

LSTM 基于细胞状态和门控制对信息实现遗忘和更新，结构中包括输入门、输出门和遗忘门，其对应的方程式为：

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f), \tag{10}$$

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i), \tag{11}$$

$$\widetilde{C_t} = \tanh(W_c h_{t-1} + U_c x_t + b_c), \tag{12}$$

$$C_t = f_t^* C_{t-1} + i_t^* \widetilde{C_t}, \tag{13}$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o), \tag{14}$$

$$h_t = o_t^* \tanh(C_t). \tag{15}$$

其中，σ 为激活函数，U、W、b 分别为模型信息的相关参数和偏倚[1]。

之前隐藏层的“记忆”的保留和遗忘是由遗忘门决定的。式(10)通过激活函数 sigmoid，利用 h_{t-1} 和当前的输入 x_t 得到输出 f_t，输出数值在[0,1]之间表示上一个 LSTM 结构保留信息的概率[1]。式(11)、式(12)利用 sigmoid 和 tanh 两个激活函数实现了对新信息的选择保留。式(13)表示为对 LSTM 结构保留的信息进行更新，即由 f_t 与 C_{t-1} 取 *Hadamard* 积，表示部分保留旧信息；i_t 和 $\widetilde{C_t}$ 取 *Hadamard* 积，表示部分保留新信息，将两者相加来更新 LSTM 结构保留的信息 C_t。输出门将式(15)中的 tanh 激活函数应用于最新 LSTM 结构保留的信息，并利用式(14)得到的 o_t 取 *Hadamard* 积控制最终的输出 h_t。

在得到最终的训练样本 D 后，结合主要变量的长期时间序列的特点，建立了的 LSTM 循环神经网络结构如图 5 所示[1]。

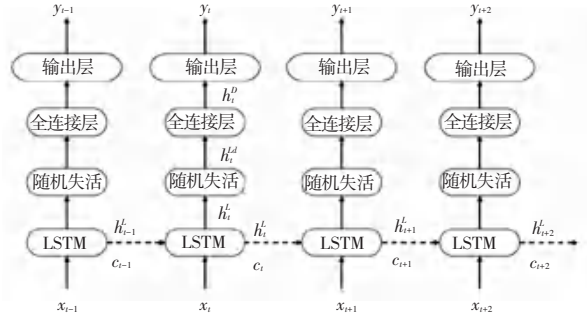

图 5 LSTM 循环神经网络结构图

Fig. 5 LSTM recurrent neural network structure diagram

图 5 中，除了当前时刻的变量数据，上一个 LSTM 结构的隐藏层输出和 LSTM 结构所包含的信息一起作为当前 LSTM 循环神经网络的输入[1]。该结构输出结果不仅传递给下一个 LSTM 结构，还利用随机失活模块进一步无差别舍弃部分隐藏层节点，以此预防过拟合现象的出现，同时也能避免 LSTM 循环神经网络因过度关注历史信息而导致新信息输入时一直出现不满意结果的现象。XGBoost 模型的参数设置见表 1。XGBoost 拟合对比如图 6 所示，LSTM 拟合对比如图 7 所示。

表 1 XGBoost 模型的参数设置

Tab. 1 XGBoost model parameter settings

参数	含义	实验设置
n_estimaotrs	决策树的数量	7 200
max_depth	决策树的最大深度，默认为 6	6
max_leaf_nodes	决策树的最大叶子节点数量，可以替代 *max_depth*	42
min_child_weight	叶子节点最小的样本权重和，默认为 1	1.5
gamma	节点划分所需的最小增益，默认为 0	0
subsample	行采样，随机采样的比例，默认为 1	0.2
colsample_bytree	采样，随机选取特征的比例，默认为 1	0.2
colsample_bylevel	每一级的每一次分裂，列采样的比例，默认为 1	1
reg_alpha	模型的 L_1 正则化参数	0.9
reg_lambda	模型的 L_2 正则化参数	0.6
leaming_rate	收缩参数，默认为 0.3	0.05

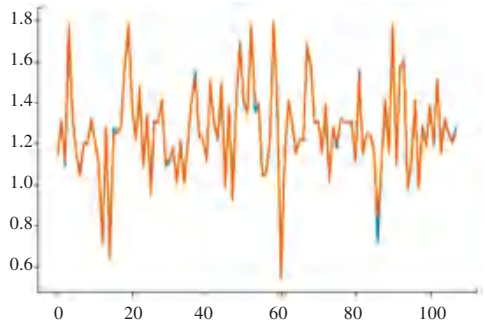

图 6　XGBoost 拟合对比图

Fig. 6　XGBoost fitting comparison chart

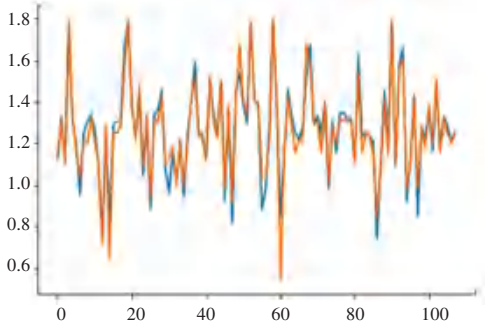

图 7　LSTM 拟合对比图

Fig. 7　LSTM fitting comparison chart

2.3　XGBoost 和 LSTM 的融合

由以上实验结果对比可得，XGBoost 的训练结果最好。但考虑到 LSTM 循环神经网络的构建是基于时间序列的，考虑到了时间因素，最终值的预测是对相比于 XGBoost 进行了更深的挖掘而得，且 LSTM 的训练结果也很好。因此，本文采用基于 XGBoost 和 LSTM 的融合模型对辛烷值损失进行预测，即对 XGBoost 的预测值和 LSTM 的预测值进行加权求和，进而得到最终的预测值 $\hat{D}^{31}$，如式(16)所示：

$$\hat{D}^{31} = \alpha \hat{D}^{31}_{XGB} + \beta \hat{D}^{31}_{LSTM}. \qquad (16)$$

其中，α、β 分别为 XGBoost 和 LSTM 的权重；$\hat{D}^{31}_{XGB}$ 为 XGBoost 的最终预测值；$\hat{D}^{31}_{LSTM}$ 为 LSTM 的最终预测值。

设定 $\alpha=\beta=0.5$，XGBoost 的参数设置同表 1，训练样本与上文相同，融合模型的拟合对比如图 8 所示。由图 8 可以看出，融合模型的拟合效果并未有 XGBoost 模型的效果好，但考虑到训练样本数据的数量并不多，可能存在过拟合问题，LSTM 相较于 XGBoost 多考虑了时间相关性因素，进行了更深层次的数据挖掘。因此，当前的融合模型虽然拟合效果不如 XGBoost，但具有更强的鲁棒性和适应性，如果拥有更多的数据量，融合模型的表现会更好。

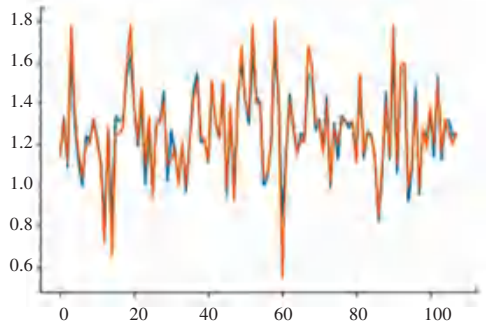

图 8　融合模型的拟合对比图

Fig. 8　Fitting comparison chart of fusion model

3　结束语

本文通过 2 种模型融合对石化企业的催化裂化汽油精制脱硫装置辛烷值损失程度进行了预测，结果表明该模型在预测精准度上有较好的表现，能够为有关部门对车用汽油质量升级关键技术上提供可靠参考。

参考文献

[1] 王炜，刘宏伟，陈永杰，等. 基于 LSTM 循环神经网络的风力发电预测[J]. 可再生能源，2020，38(9)：1187-1191.

[2] 邹玉莹. 基于机器学习的票据转贴现利率预测研究[D]. 南昌：江西财经大学，2020.

[3] 杨轶男，任晔，毛安国，等. 影响催化裂化装置汽油辛烷值变化的技术因素分析[J]. 炼油技术与工程，2019，49(6)：32-35.

[4] 马强，赵昌明. 降低 S-Zorb 装置汽油辛烷值损失的优化操作[J]. 当代化工研究，2020(15)：43-45.

[5] 刘宝，倪维起. S Zorb 装置汽油辛烷值损失影响因素分析[J]. 齐鲁石油化工，2019，47(2)：102-104，124.

[6] GERS F A，SCHMIDHUBER J，CUMMINS F. Learning to forget：Continual prediction with LSTM[J]. Neural Computation，2000，12(10)：2451-2471.

[7] ZHANG Dahai，QIAN Liyang，MAO Baijin，et al. A data-driven design for fault detection of wind turbines using Random Forests and XGboost[J]. IEEE Access，2018，6：21020-21031.

[8] 万黎，毛炳启. Spearman 秩相关系数的批量计算[J]. 环境保护科学，2008，34(5)：53-55，72.

(上接第 184 页)

[26] YIN Di，HUANG Shujian，DAI Xinyu，et al. Utilizing non-parallel text for style transfer by making partial comparisons[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19). Macao，China：IJCAI，2019：5379-5386.

[27] TAO Qian，ZHOU Yuchen，HUANG Jie，et al. A GAN-based transfer learning approach for sentiment analysis [C]// Proceedings of the 2019 International Conference on Artificial Intelligent and Computer Science. New York，USA：Association For Computing Machinery，2019：364-368.

[28] LARSEN A B L，SØNDERBY S K，LAROCHELLE H，et al. Autoencoding beyond pixels using a learned similarity metric[J]. arXiv preprint arXiv：1512.09300，2015.