



基于 5G 语音质差自适应算法研究及应用

赵宇翔, 纪雅欣, 余立, 周天一, 周航

(中国移动(浙江)创新研究院有限公司, 浙江 杭州 310030)

摘要: MOS 通常被业界用于评价语音质量,它能够客观公正地反映用户语音业务的感知。通过路测获取数据的方式难度大、成本高,通常采用训练好的监督学习模型预测 MOS。但运营商语音数据存在 MOS 低分数据占比低和时序变化的特性,这种数据特性影响了模型预测的精度和泛化性。在研究现有运营商数据采集系统和机器学习算法的基础上,提出了一种面向 5G 语音质差 MOS 评估的自适应算法。首先,基于全参评估的 POLQA 算法测试设备获取训练数据,保证了训练样本的准确性;其次,通过数据增强的方法,解决了质差样本获取难度大的问题;最后,基于自适应算法选型实现周期性动态地根据数据特征的时序变化选择最佳 MOS 预测模型,实现 5G 语音质量规模化、智能化的评估。

关键词: 5G 语音质量; MOS; 机器学习; 自适应

中图分类号: TN915

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2023249

Research and application of adaptive algorithm for 5G voice quality evaluation

ZHAO Yuxiang, JI Yaxin, YU Li, ZHOU Tianyi, ZHOU Hang

China Mobile (Zhejiang) Research & Innovation Institute Co., Ltd., Hangzhou 310030, China

Abstract: MOS (mean opinion score) is usually used to evaluate voice quality in the industry. It can objectively and fairly reflect the user's voice service perception. It is difficult and costly to obtain data by road test, so a trained supervised learning model is usually used to predict the MOS score. However, the operator voice data has the characteristics of low percentage of MOS low score data and time sequence change, which affects the accuracy and generalization of the model prediction. Based on the study of existing data acquisition systems and machine learning algorithms of operators, an adaptive algorithm for MOS evaluation of 5G speech quality was proposed. Firstly, POLQA algorithm test equipment based on full parameter evaluation obtained training data to ensure the accuracy of training samples. Secondly, by means of data enhancement, the difficulty of acquiring poor quality samples was solved. Finally, based on the adaptive algorithm selection, the optimal MOS prediction model could be selected periodically and dynamically according to the timing changes of data features, so as to achieve large-scale and intelligent evaluation of 5G voice quality.

Key words: 5G voice quality, MOS, machine learning, adaptive



0 引言

语音质量评估的通常做法是提取损伤信号和未损伤的原始语音信号的特征参数,再通过计算得到参数与平均主观意见分(mean opinion score, MOS)值的映射关系,使用映射关系对语音质量进行评估。2011年,国际电信联盟(ITU)颁布了语音质量评估的感知客观语音质量评估(perceptual objective listening quality analysis, POLQA)标准,它是语音质量感知评估(perceptual evaluation of speech quality, PESQ)算法的升级版,可用于互联网电话(VoIP)、3G/4G/5G网络语音等场景评价,是至今为止电话网络公认的双端语音质量的客观评价标准^[1-2]。文献[3]对比了POLQA算法(P.863算法)与PESQ算法(P.862算法),结果表明POLQA评分结果的离散程度更低。相比较而言,使用POLQA算法评估MOS的方式精度更好,可以作为模型训练的标签;但这种方式需要未损伤的原始语音信号作为参考进行对比,因此需要集成在特定的测试仪表或设备上,灵活性和普适性不足,无法做到大规模的MOS评估。

随着国内5G的大规模建设,5G语音业务量也迎来了快速增长,如何实现5G语音质量低成本、高效化、规模化、智能化的评估成为各运营商研究热点。前期研究^[4-8]主要集中在对语音通话本身的分析。文献[4]从质量、容量、覆盖分析了5G新空口承载语音(VoNR)的关键技术。文献[5]介绍了语音通话在无线层导致的分组丢失、时延问题进行分析、定位的基本思路和优化方法。文献[6]从电路交换语音回落(CS fallback, CSFB, 即CS语音回落)、呼叫建立、挂机返回3个阶段构建了语音业务质量评估体系。为满足模型训练需要,文献[7]建立了非结构化的语音数据集的质量评估标准,从广度覆盖性、选集区分性、领域深入性、数据完整性上进行了规范。文献[8]从覆盖、干扰、

参数、新功能算法改进等方面切入,定量分析各因素对语音质量的影响程度。此类语音质量研究聚焦在语音业务发生过程中网元设备生成的性能数据分析,分析方法的建立过程通常要结合相对固定的业务规则和专家经验,对不同业务场景质量评估的延展性较差。

随着人工智能的发展,越来越多的研究人员将语音质量提升的相关研究和机器学习模型结合。文献[9]提出了一种跨域多目标语音评估模型MOSA-Net,该模型采用卷积神经网络和双向长短期记忆架构,可以同时估计输入语音信号的语音质量、短时目标可懂度(short-time objective intelligibility, STOI)和线性相关系数(linear correlation coefficient, LCC)。文献[10]提出了一种基于双向长短时记忆(bidirectional long short-term memory, Bi-LSTM)的端到端非侵入式语音质量评估模型Quality-Net。上述两种模型都是将改造过的神经网络用于传统MOS语音评价。文献[11]针对标注缺失问题提出了半监督学习与优化准则联合优化方法。文献[12]基于华为端到端信令分析平台服务体验质量(service & experience quality, SEQ)分析平台对信令面数据进行相关性分析及聚类分析,建立了快速有效的端到端质量评估及优化体系。文献[13]引入机器学习对海量用户的测量报告(MR)进行建模,实现了对全网用户质差的有效甄别。文献[14]从大量外部数据表示法(external data representation, XDR)数据中通过梯度提升决策树(gradient boosting decision tree, GBDT)模型提取特征信息,从而进行质量评估。上述方法通过逻辑回归、特征工程建立及相关性分析等业界主流的建模方法搭建了智能语音评估体系,基于人工智能搭建的评估模型可针对不同业务场景选取不同特征、不同模型完成训练,提升了评估的灵活性和延展性。

围绕现网中语音质差数据占比低导致的可用训练样本不足、数据分布不均衡的问题,与数据

相关的研究和改进也在同步进行。文献[15]利用熵值法和效用函数法,构建指标间的模糊关系,提高了各指标间的相关关系。文献[16]基于对比学习的双分类器去除数据冗余信息,减少不同类别特征的混淆程度,提高了现有数据的可用性。文献[17]采用启发式滤波的密度峰值聚类实现不平衡数据集自适应加权过采样。文献[18]使用寻峰算法的神经网络进行拉曼光谱的样本扩充。利用样本扩充进行数据增强的方法具有很强的借鉴意义,数据增强提高了数据的质量,能在一定程度上解决数据问题。

部署在现网的模型长期不变,会导致训练数据和现网数据不一致的问题。在电网、路网匹配、信道接入等其他领域,这样的问题同样存在。文献[19-21]针对各自领域数据的特征,分别提出了适配的自适应算法来灵活应对影响因素与参数的多变性。文献[22]通过自动调整上行时隙重复次数,提高传输的覆盖性能。文献[23]为了对信道质量进行评估,对信道进行周期性检测,提高了信道的动态接入率。在语音领域,同样可以结合周期性检测,实现自动调整和自适应。

本文在对研究现状进行分析的基础上,结合运营商数据采集系统和语音包协议提出了一种基于自适应算法的 5G 语音质量评估方法,首先,基于全参评估的 POLQA 算法测试设备获取训练数据,保证了训练样本的准确性;其次,通过数据增强的方法,解决了质差样本获取难度大的问题;最后,基于自适应算法实现数据特征动态变化下的最佳 MOS 预测模型的选择。基于本文提出的方法,可以实现语音 MOS 质量准确评估,模型可在运营商生产系统上灵活部署,实现 5G 语音质量规模化、智能化评估。

1 语音质量评估自适应优化算法研究

1.1 算法优化模型

为了实现语音质量评估的自适应优化,本文采

用双层自适应模型建模。上层 F_n 为数据层,下层 $\tilde{F}_n$ 为模型层。针对质差样本获取难度大的问题,数据层 F_n 采用生成对抗网络 (generative adversarial network, GAN) 实现的数据增强算法 f_{GAN} 提升质差和欠质差数据的占比。模型层 $\tilde{F}_n$ 的目的是通过多模型联合优化方式提高模型精度,采用 f_{stacking} 模型融合算法。同时,周期性地获取训练数据,自适应地选择模型与参数。整体的任务量为:

$$F = F_n + \tilde{F}_n = f_{\text{GAN}} + f_{\text{stacking}} \quad (1)$$

算法流程描述如图 1 所示,具体描述如下。

步骤 1 从真实样本数据分布 $p_{\text{data}}(x)$ 采样 m 个小批量样本 $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ 。

步骤 2 通过 GAN 生成对抗过程生成一批构造数据 $\{z^{(1)}, z^{(2)}, \dots, z^{(m)}\}$, 真实数据与构造数据组成新的训练数据。

步骤 3 stacking 融合模型训练,基学习器提取高维特征,元学习器预测 MOS 值,遍历模型库更改 stacking 中的模型选择,保存精准率最高的模型选择为最佳模型。

步骤 4 使用最佳模型预测现网采集到的 XDR 数据得到 MOS 值。

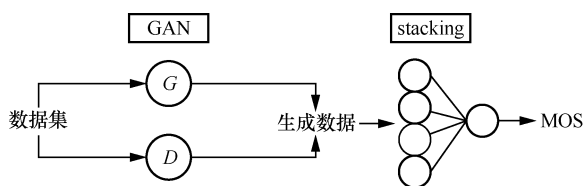


图 1 算法流程描述

(1) 数据层优化算法

最常用的 MOS 评分量表是 5 分绝对类别评分量表——优秀为 5、好为 4、一般为 3、差为 2、极差为 1,然而路测中的 MOS 值不限于整数,并且数值变化范围受业务类型的影响很大^[24]。数据质量划分标准见表 1。为了研究数据分布情况,按照表 1 中区间建议值将数据划分为良好、欠质差、质差 3 类。



表 1 数据质量划分标准

业务类型	建议值
MOS 良好区间	[3.8, 4.8]
MOS 欠质差区间	[3.0, 3.8)
MOS 质差区间	[1.08, 3.0)

基于大量训练数据统计,良好、欠质差、质差 3 类数据的数量分布呈现 4:1:1。将原始样本放入模型训练后,良好点的预测精准率可以达到 95%以上,而欠质差点和质差的训练效果一直很差,经过参数寻优和调整,也未超过 70%。这种情况严重影响了整体的预测效果。本质原因是质差点与欠质差点的数据量太少,训练不足;3 类数据样本不均衡,训练陷入局部最优。为了解决数据问题,有必要进行数据增强。噪声增强是一种经典的数据增强算法,通过给每条 XDR 切片的各个特征加入高斯噪声,从而构造出更多样本。但这种算法构造出的样本与原样本存在相关关系,容易造成过拟合。因此,引入 GAN 构造质差数据和欠质差数据。

GAN 生成对抗过程如图 2 所示^[25-28]。GAN 通过生成器 G 与判别器 D 相互对抗,通过最小化真实数据分布 p_x 和模型生成数据分布 p_z 之间的 Jensen-Shannon 散度来训练学习网络参数 θ 。通过生成对抗过程,生成一批构造数据,该数据与原始数据具有极高的相似性。

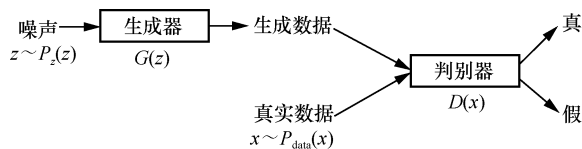


图 2 GAN 生成对抗过程

GAN 目标函数为 $V(D, G)$ 。 G 希望减少 V 的值让自己生成的分布无法识别,而 D 希望增大 V 的值让自己可以高效地判别数据的真假类别。因此 GAN 生成对抗过程如式 (2) 所示。

$$\min_G \max_D V(D, G) = E_{x \sim P_{\text{data}}(x)} [\lg D(x)] + E_{z \sim P_z(z)} [\lg(1 - D(G(z)))] \quad (2)$$

其中, G 表示生成器,选用全连接神经网络, D 表示判别器, E 表示数学期望。

训练过程中涉及的约束条件如下。

对抗损失:

$$L_D = \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m [\lg D(x^{(i)}) + \lg(1 - D(G(z^{(i)})))] < 0.8 \quad (3)$$

其中, θ_d 表示判别器的参数, m 表示样本的数量。

生成损失:

$$L_G = \nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^m \lg(1 - D(G(z^{(i)}))) < 0.65 \quad (4)$$

其中, θ_g 表示生成器的参数, m 表示样本的数量。

(2) 模型层优化算法

模型融合的经典方法有投票法、平均法、排序法等。平均法可以对多个回归模型的预测结果进行平均,优点在于平滑结果,从而减少过拟合,但是并不能获得多个模型的最佳效果。本文提出的模型层优化算法基于从模型库中选择的多种模型的融合。模型库涵盖了常见经典的回归模型^[29],如树模型、深度学习模型、时间序列模型等,举例如下。

- 树模型及其变种: XGBoost、LightGBM、CatBoost。
- 深度学习模型: CNN、DNN。
- 时间序列模型: LSTM、Bi-LSTM+Attention。

数据集被划分为训练集和测试集,标签采用路测 MOS 值。为了充分利用多个模型的训练效果,通过单层 stacking 方法进行模型融合,流程如图 3 所示;同时防止数据重复使用造成过拟合,采用了 K 折交叉验证, K 取 4 时,训练集被划分为 4 份。模型融合的过程包括以下 3 个步骤。

步骤 1 基学习器在模型库中随机选择 4 个,输入为特征,输出为提取到的新的高维特征。

步骤 2 元学习器从模型库剩余模型中选择

一个, 输入为提取到的新的高维特征, 输出为 MOS 值。

步骤 3 训练好的基学习器和元学习器共同组成 stacking 融合模型。

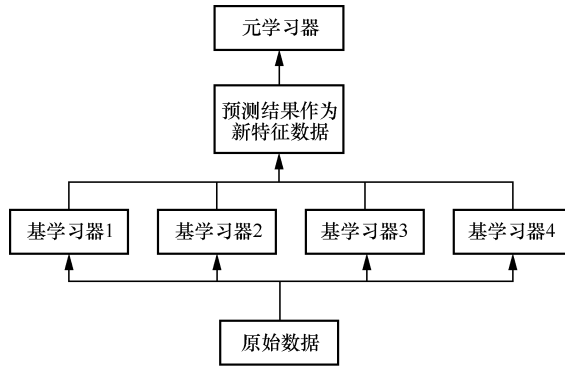


图3 单层 stacking 流程

基学习器 1 训练过程如图 4 所示, 以 CatBoost 模型为例阐述基学习器 1 的训练过程。样本 1/样本 2/样本 3 组成第一份数据集训练 CatBoost 模型, 样本 4 使用该 CatBoost 模型预测得到 MOS 值作为高维特征 $1'$, 同时测试集也使用该 CatBoost 模型预测得到 MOS 值作为 h_1 。同样, 样本 2/样本 3/样本 4、样本 1/样本 3/样本 4、样本 1/样本 2/样本 4 各组成第 2~4 份数据集重复上述过程。进行 4 次训练的同样都是 CatBoost 模型, 但在训练过程中模型参数不断优化。高维特征 $1'$ 、高维特征 $2'$ 、高维特征 $3'$ 、高维特征 $4'$ 共同组成新向量 m_1 , 而 h_1 、 h_2 、 h_3 、 h_4 的平均值作为新向量 n_1 。

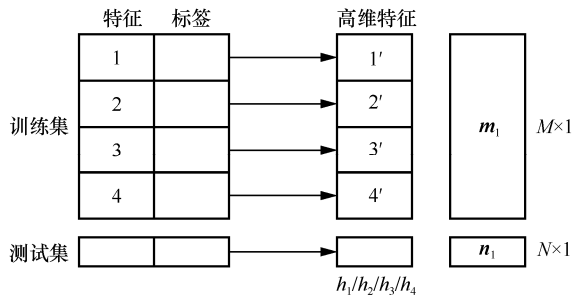


图4 基学习器 1 训练过程

其余 3 个基学习器同样在模型库中进行选择, 训练方法同上。4 个基学习器训练的结果共同组成新的矩阵, 如图 5 所示。

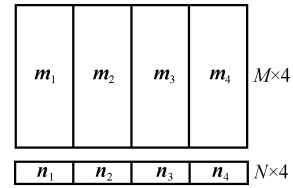


图5 4个基学习器训练的结果

元学习器使用高维特征作为输入, 标签同样采用路测 MOS 值。图 5 中的 $M \times 4$ 的矩阵作为元学习器的训练集训练模型, $N \times 4$ 的矩阵构成元学习器的测试集。元学习器的训练过程如图 6 所示。

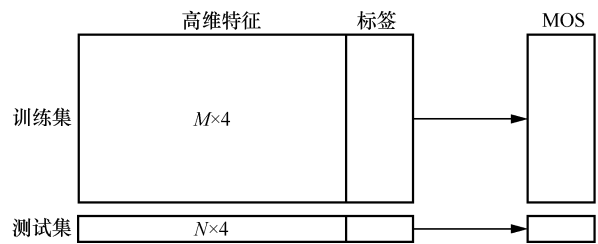


图6 元学习器训练过程

模型训练好后, 需要使用训练好的模型线上部署。 mo_1 、 mo_2 、 mo_3 、 mo_4 为 4 个基学习器, mo_5 为元学习器。一条 XDR 数据分别经过 4 个模型提取 4 个高维特征, 4 个高维特征预测得到 MOS 值。融合模型线上流程如图 7 所示。

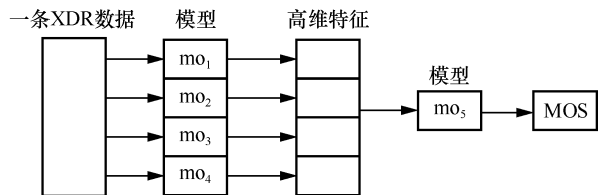


图7 融合模型线上流程

整体的目标函数为:

$$L = \max \text{PRE}(y, y') \quad (5)$$

其中, PRE 为精准度 (precision) 指标, y 为 MOS 真实值, 即路测 POLQA MOS, y' 为模型预测值。

设置提取终止迭代条件 early stopping, 当 $|L^i - L^{i-1}| \leq \varepsilon$ 时, 即相邻两次迭代的目标函数差值小于设定的阈值 ε 时, 模型训练迭代完成。

保存最佳模型对应的各个最佳参数, 最佳参



数表示为:

$$\theta_{i,\text{best}} = \max(F_j(\theta_i)) \quad (6)$$

其中, i 表示参数的个数, j 表示迭代的次数。

(3) 周期迭代

由于设备升级, 环境实时变化、配置更改等情况, XDR 数据和路测数据的映射关系会发生一定变化。为了减少训练数据与线上数据不一致导致的预测精准度下降的问题, 需要定期采集最新数据来更新数据集, 从而达到数据与模型相适应的效果。采集周期一般为 3 个月, 根据行业内发展变化速度, 周期可灵活调整。上一个周期保存的模型及模型参数将作为下一个周期的初始值, 以减少寻优时间。

1.2 算法应用

语音质量评估方法模块如图 8 所示, 整个评估方案包含数据预处理模块、数据划分模块、数据增强模块、模型训练共 4 个功能模块。

- 数据预处理模块: DPI 系统采集 XDR 数据, 道路拨测采集路测数据, 并将两部分数据通过时间进行定位关联, 路测 POLQA MOS 作为数据集的标签。
- 数据划分模块: 按照数据质量划分标准, 将数据划分为良好、欠质差、质差 3 类。

- 数据增强模块: GAN 构造欠质差和质差数据, 并将原始数据与构造数据合并, 形成新的数据集。
- 模型训练模块: 数据进入 stacking 模型融合训练, 模型选定后保存最佳模型。

线上获取当前 XDR 数据, 使用最佳模型预测得到 MOS 分。

周期性采集数据, 更新数据集, 重复上述过程。

在实际应用中, 会通过运营商的 DPI 系统获取现网用户实时语音业务的 RTP/RTCP 包相关性能数据, 完成数据处理后, 调用最佳模型用于线上实时语音 MOS 预测, 得到语音质量评估结果。

2 实验与结论

2.1 数据预处理及特征工程

路测数据和 XDR 数据关联后, 根据业务需要进一步生成衍生特征。原始特征和衍生特征共得到 105 个, 包括信令面特征、用户面特征和其他特征等, 部分业务特征及描述见表 2。

依据专家经验, 筛选出与 MOS 值强相关的特征包括时延、丢包、包间隔等。建立数据的特征工程, 一个特征越多地被用来在模

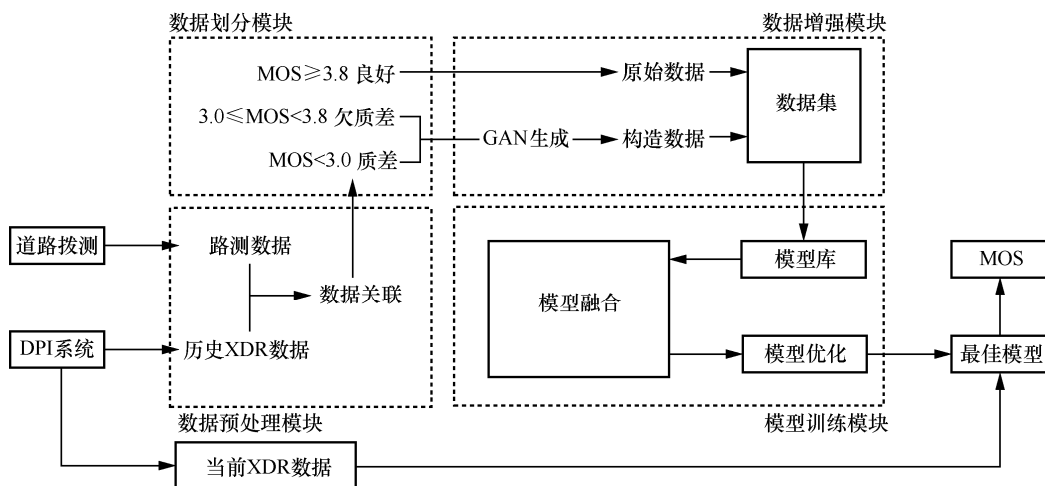


图 8 语音质量评估方法模块

表 2 业务特征及描述

业务特征	描述	
信令面特征	接入和承载建立特征	5G 承载建立标识
	5G 信令特征	信令面失败次数
用户面特征	业务行为特征	5G 业务时长
		5G 业务占比
	业务质量 KQI 特征	上/下行 RTP 丢包数
		上/下行 RTP 期望包数
		本周期收到的最后一个 RTP 包相对于开始时间的时延
其他特征	标识信息	切片周期内 RTP 连续丢包大于或等于 2 且小于 6 的次数
		MSISDN 码
		SESSION ID
		XDR ID
	Interface 字段	Gm、S1-U、N3 接口

型中构造决策树，它的重要性就越高。因此通过 XGBoost 模型的 plot_importan 函数得到各个特征的重要性得分并进行排序，默认参数为 weight，得到与 MOS 值相关度最强的前 10 个特征和得分，特征重要性排序如图 9 所示。

由特征重要性分析得到的与 MOS 值相关的前 10 个特征的特征包括平均和最大相对时延、包间隔、抖动等，并区分了静默帧和非静默帧。由特征重要性分析得到的与 MOS 值相关的前 10 个特征的详细含义见表 3。

表 3 与 MOS 值相关的前 10 个特征

序号	特征名称	含义
1	AvgInterDelay	平均相对时延
2	MaxIntervalNonsid	RTP 最大包间隔
3	MaxInterDelay	最大相对时延
4	MaxJitter	最大 RTP 抖动
5	MaxLossRate	RTP 最大连续丢包率
6	LossRateNonsid	RTP 丢包率（非静默帧）
7	RtcpPacketLossRateKeepTime	RTCP 单位时间的丢包数
8	MaxLoss	RTP 最大连续丢包数
9	MaxLossNonsid	RTP 最大连续丢包数(非静默帧)
10	CodeRateFeat	编码速率

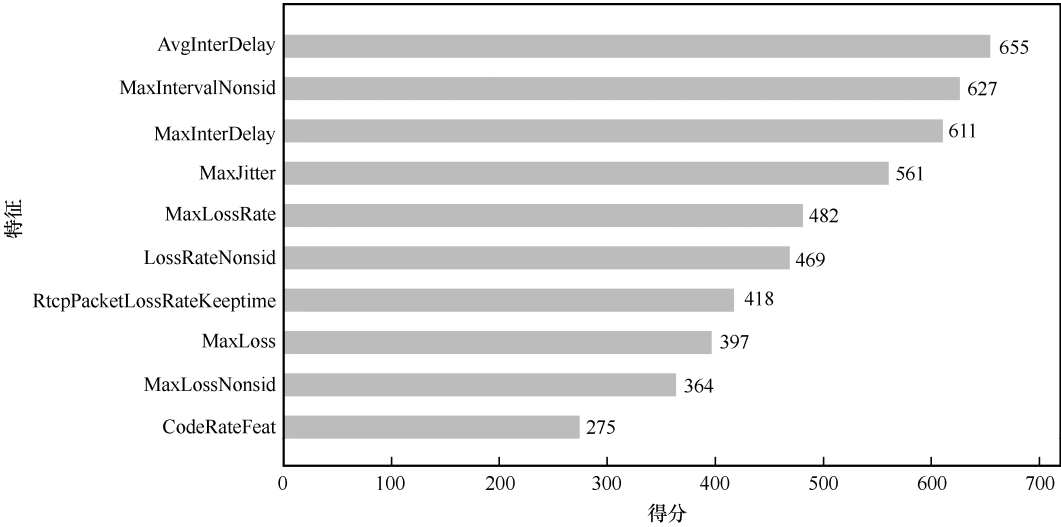


图 9 特征重要性排序



2.2 模型效果对比

数据预处理并选定特征后,通过本文提出的基于 5G 语音质差的自适应算法进行实验,分别设计了单个模型的数据增强对比实验、单模型与模型融合对比实验以及多批次数据对比实验 3 个实验。单个模型的数据增强对比实验是为了验证数据层的 GAN 数据增强效果,单模型与模型融合对比实验是为了对比模型层的模型融合效果,多批次数据对比实验是为了检验周期性采集到数据的不同表现。

(1) 单个模型的数据增强对比

使用 GAN 生成构造质差和欠质差数据,验证数据增强对模型的改进效果。原始数据 27 550 条,使用 GAN 生成 7 319 条,两部分合并后数据集共有 34 869 条数据。以 CatBoost 模型为例,使用原始数据与增强后的数据,对比 3 个评价指标——召回率 (recall)、精准度 (precision)、F1 值 (F1 score),CatBoost 模型数据增强前后对比见表 4。

表 4 CatBoost 模型数据增强前后对比

指标	数据	总体	质差	欠质差	良好
召回率	原始数据	0.582	0.518	0.488	0.739
	增强后数据	0.772	0.793	0.779	0.745
精准度	原始数据	0.731	0.876	0.376	0.824
	增强后数据	0.788	0.978	0.533	0.852
F1 值	原始数据	0.648	0.651	0.276	0.779
	增强后数据	0.768	0.876	0.472	0.795

实验结果表明,采用增强后的数据各项评价指标均有增益,其中质差点的评估指标增益明显,均已经达到较高的水平。欠质差数据的预测评估指标虽然得到一定程度的提高,但从数据表现来看准确率偏低。这是由于欠质差数据的 MOS 阈值处于 3.0~3.8,较小波动就会导致欠质差被划分为质差或者良好。在运营商日常网络运维运营中,更加关注质差通话的占比及影响。该实验表明数据增强可以明显改善质差数据的评估效果,对于模型训练有一定的帮助。

(2) 单模型与模型融合对比

相较于单一模型的方法,模型融合方法集成了多种模型,元学习器和基学习器均从模型库中选择,设计实验验证模型融合的效果。基学习器和元学习器的模型组合尽可能全面地选择各类模型,实验中选择的模型组合方式分别如下。

- 模型融合 stacking_1 , 其中基学习器选取 CatBoost、LightGBM、DNN、LSTM, 元学习器选取 XGBoost。
- 模型融合 stacking_2 , 其中基学习器选取 XGBoost、CatBoost、LightGBM、LSTM, 元学习器选取 DNN。
- 模型融合 stacking_3 , 其中基学习器选取 XGBoost、CatBoost、LightGBM、DNN, 元学习器选取 LSTM。

使用原始数据与增强后的数据构成的新数据集,分别得到单模型、平均法融合模型和本文提出的 stacking 融合模型的 precision 指标,单模型与模型融合对比见表 5。

表 5 单模型与模型融合对比

模型	总体	质差	欠质差	良好
XGBoost	0.782	0.968	0.528	0.850
LightGBM	0.705	0.952	0.313	0.843
CatBoost	0.788	0.978	0.533	0.852
CNN	0.759	0.934	0.374	0.963
DNN	0.759	0.902	0.374	0.999
LSTM	0.673	0.842	0.227	0.852
Bi-LSTM+Attention	0.692	0.823	0.212	0.868
平均法	0.737	0.914	0.366	0.889
stacking_1	0.804	0.970	0.623	0.818
stacking_2	0.793	0.971	0.602	0.806
stacking_3	0.821	0.964	0.676	0.823

实验结果显示,平均法中各个模型的结果更平滑,但没有提升整体效果;3 种 stacking 组合模型融合后的预测效果,在准确率方面均得到了提升。尤其是对于数据增强不能解决的欠质差类数据的数据精度问题, stacking 融合模型将欠质差数

据的 precision 指标提升到了 60% 以上。其中基学习器选取 XGBoost、CatBoost、LightGBM、DNN，元学习器选取 LSTM 时，融合模型效果最好。

(3) 多批次数据对比

为了验证不同数据分布表现下算法的优

劣，在现网共采集到 3 批次数据。3 个批次的的数据首先进行了 GAN 数据增强，原始数据与增强后的数据构成新的数据集进行单模型训练和融合模型训练，多批次数据精准度指标对比见表 6。

表 6 多批次数据精准度指标对比

模型	第一批次				第二批次				第三批次			
	总体	质差	欠质差	良好	总体	质差	欠质差	良好	总体	质差	欠质差	良好
XGBoost	0.782	0.968	0.528	0.850	0.810	0.956	0.602	0.872	0.809	0.954	0.613	0.861
LightGBM	0.705	0.952	0.313	0.843	0.767	0.923	0.535	0.844	0.779	0.924	0.559	0.856
CatBoost	0.788	0.978	0.533	0.852	0.808	0.927	0.638	0.859	0.811	0.922	0.656	0.854
CNN	0.759	0.934	0.374	0.963	0.729	0.851	0.456	0.880	0.725	0.863	0.467	0.845
DNN	0.759	0.902	0.374	0.999	0.702	0.905	0.618	0.584	0.718	0.884	0.471	0.781
LSTM	0.673	0.842	0.227	0.852	0.642	0.796	0.231	0.821	0.678	0.852	0.264	0.854
Bi-LSTM+Attention	0.692	0.823	0.212	0.868	0.654	0.827	0.225	0.836	0.693	0.868	0.273	0.868
stacking ₁	0.804	0.970	0.623	0.818	0.850	0.936	0.769	0.845	0.836	0.929	0.802	0.874
stacking ₂	0.793	0.971	0.602	0.806	0.868	0.955	0.772	0.877	0.879	0.955	0.834	0.847
stacking ₃	0.821	0.964	0.676	0.823	0.821	0.913	0.679	0.872	0.841	0.922	0.753	0.846

实验结果显示，面对不同数据的分布及表现，本文所提自适应算法均能显著提升欠质差点的准确率，从而达到整体数据的 MOS 评估效果提升。3 个批次的数据的总体效果，与单模型对比，第一批次数据效果最佳为 stacking₃，第二批次、第三批次数据效果最佳均为 stacking₂。

传统的语音质量评估的 Emodel 模型采用固定的人为制定的计算式和参数得到 MOS 评分结果，评估模型相对简单粗糙。本文通过以上 3 个实验验证了数据增强叠加模型融合的自适应算法，可以解决在数据不均衡及数据分布随机性变化带来的质差点与欠质差点识别准确率低的问题，与 Emodel 模型相比实现了整体数据的 MOS 评估效果提升。在运营商现网中的应用过程，需要重视数据表现差异，周期性采集数据进行迭代训练，根据不同数据分布情况采用自适应融合算法，灵活选择最佳模型，从而达到最佳评估效果。

3 结束语

本文提出的 5G 语音质差自适应算法，数据层通过数据增强提升质差数据的评估效果，模型层支持根据不同数据分布及特性自适应地进行模型融合提升欠质差数据的评估效果，从而在整体上降低数据不均衡的影响、提升了模型对于 5G 语音质差数据的整体评估精度。该算法可应用于电信运营商网管系统，完成实时、大规模 XDR 数据下语音的质差精准发现，指导运维人员发现故障并及时处理，保障正常用户通话效果，具有良好的规模化、智能化效果。

参考文献：

- [1] STREIJL R C, WINKLER S, HANDS D S. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives[J]. Multimedia Systems, 2016, 22(2): 213-227.
- [2] 玉荣娟. 语音质量评估及其优化策略[J]. 移动通信, 2009, 33(20): 79-81.
YU R J. Speech quality assessment and optimization strategies[J]. Mobile Communications, 2009, 33(20): 79-81.



- [3] 陈明. VoLTE 语音质量评估的 POLQA 算法研究[J]. 移动通信, 2016, 40(15): 20-25.
CHEN M. Research on the POLQA algorithm for VoLTE speech quality assessment[J]. Mobile Communications, 2016, 40 (15): 20-25.
- [4] 薛晓宇, 龙杰, 方义成, 等. 5G VoNR 语音关键技术分析[J]. 电信工程技术与标准化, 2022, 35(9): 81-85.
XUE X Y, LONG J, FANG Y C, et al. Analysis of key technologies for 5G VoNR speech[J]. Telecommunications Engineering Technology and Standardization, 2022, 35 (9): 81-85.
- [5] 吕军. 中国电信 FDD 4G VoLTE 语音感知 MOS 质量问题分析和优化提升的研究[J]. 数据通信, 2020(5): 1-9.
LYU J. Analysis and optimization of MOS quality problems in China Telecom FDD 4G VoLTE voice perception[J]. Data Communication, 2020 (5): 1-9.
- [6] 张磊, 张恒, 李一, 等. CSFB 语音业务质量评估与优化方法研究[J]. 邮电设计技术, 2016(2): 66-72.
ZHANG L, ZHANG H, LI Y, et al. Research on the quality evaluation and optimization method of CSFB voice service[J]. Posts and Telecommunications Design Technology, 2016 (2): 66-72.
- [7] 李荪, 曹峰, 刘姿杉. 面向算法模型的语音数据集质量评估方法研究[J]. 计算机科学, 2022, 49(S2): 519-524.
LI S, CAO F, LIU Z S. Research on the quality evaluation method of speech datasets for algorithm models[J]. Computer Science, 2022, 49 (S2): 519-524.
- [8] 辛建国, 刘洪波, 王浩. VoLTE 语音质量(MOS)提升研究[J]. 山东通信技术, 2021, 41(1): 40-45.
XIN J G, LIU H B, WANG H. Research on improving VoLTE voice quality (MOS)[J]. Shandong Communication Technology, 2021, 41 (1): 40-45.
- [9] ZEZARIO R E, FU S W, CHEN F, et al. Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features[J]. IEEE/ACM Transactions on Audio, 2023, 31(1): 54-70.
- [10] FU S W, TSAO Y, HWANG H T. Quality-net: an end to-end non-intrusive speech quality assessment model based on BLSTM[J]. Inter Speech, 2018(1): 1873-1877.
- [11] 李伦彬, 滕海坤, 王诗莹. 一种用于语音质量评价的半监督学习方法研究[J]. 长春工程学院学报(自然科学版), 2021, 22(3): 119-124.
LI L B, TENG H K, WANG S Y. Research on a semi-supervised learning method for speech quality evaluation[J]. Journal of Changchun Institute of Technology (Natural Science Edition), 2021, 22 (3): 119-124.
- [12] 陈森, 金岩华. VoLTE 端到端语音质量评估及优化研究[J]. 无线互联科技, 2018, 15(3): 10-11, 20.
CHEN S, JIN Y H. Research on VoLTE end-to-end voice quality evaluation and optimization[J]. Wireless Internet Technology, 2018, 15 (3): 10-11, 20.
- [13] 王希. 基于 MR 数据与机器学习的 LTE 用户感知评估方法[J]. 移动通信, 2018, 42(8): 21-26.
WANG X. LTE user perception evaluation method based on MR data and machine learning[J]. Mobile Communications, 2018, 42 (8): 21-26.
- [14] 黄鹂声, 冉金也, 罗静, 等. 基于 XDR 数据分析的 OTT 视频服务感知质量评估方法[J]. 计算机研究与发展, 2021, 58(2): 418-426.
HUANG L S, RAN J Y, LUO J, et al. A method for evaluating the perceived quality of OTT video services based on XDR data analysis[J]. Computer Research and Development, 2021, 58(2): 418-426.
- [15] 张瑜, 吴凯, 郭杰, 等. 基于数据质量评估的自适应序贯航迹关联算法[J]. 系统工程与电子技术, 2022, 44(11): 3477-3485.
ZHANG Y, WU K, GUO J, et al. Adaptive sequential track association algorithm based on data quality assessment[J]. Systems Engineering and Electronic Technology, 2022, 44(11): 3477-3485.
- [16] 孙艳丰, 陈亮, 胡永利. 基于对比学习的双分类器无监督域适配模型[J]. 北京工业大学学报, 2023, 49(2): 197-204.
SUN Y F, CHEN L, HU Y L. The unsupervised domain adaptation model of double classifiers based on contrastive learning[J]. Journal of Beijing University of Technology, 2023, 49(2): 197-204.
- [17] TAO X M, LI Q, GUO W J, et al. Adaptive weighted over-sampling for imbalanced datasets based on density peaks clustering with heuristic filtering[J]. Information Sciences, 2020(519): 43-73.
- [18] 张鸿, 黄保坤. 拉曼光谱神经网络分类与样本扩充方法[J]. 福建电脑, 2023, 39(4): 19-24.
ZHANG H, HUANG B K. Neural network classification and sample expansion in Raman spectrum[J]. Fujian Computer, 2023, 39(4): 19-24.
- [19] 李娜, 李琳琳, 王永远. 麦克风自适应算法在鲁棒语音合成中的应用研究[J]. 中国科技信息, 2013(11): 127, 130.
LI N, LI L L, WANG Y Y. Research on the application of microphone adaptive algorithm in robust speech synthesis[J]. China Science and Technology Information, 2013(11): 127, 130.
- [20] 朱永宽, 谷涓涓. 自适应遗传算法在聚类分析中的应用[J]. 黑龙江科技信息, 2010(25): 52-53.
ZHU Y K, GU J J. Application of adaptive genetic algorithm in cluster analysis[J]. Heilongjiang Science and Technology Information, 2010(25): 52-53.
- [21] 王超, 田波, 李子睿, 等. 基于自注意力子域自适应对抗网络的轴承故障诊断方法[J]. 中国舰船研究, 2023(6): 1-10.
WANG C, TIAN B, LI Z R, et al. A bearing fault diagnosis method based on self attention subdomain adaptive adversarial network[J]. China Shipbuilding Research, 2023(6): 1-10.
- [22] 谢顺文, 林高全, 邱宏洪, 等. 一种 5G 上行覆盖增强的方法[J]. 移动通信, 2023, 47(5): 90-96.
XIE S W, LIN G Q, QIU H H, et al. A method for enhancing 5G uplink coverage[J]. Mobile Communications, 2023, 47(5): 90-96.
- [23] 彭艺, 魏翔, 朱桢以, 等. 基于能量检测与信道质量评估的

- 动态信道接入算法[J]. 数据通信, 2023(1): 5-9.
- PENG Y, WEI X, ZHU Z Y, et al. Dynamic channel access algorithm based on energy detection and channel quality evaluation[J]. Data Communication, 2023(1): 5-9.
- [24] ITU. ITU-T Recommendation: P.863, Perceptual objective listening quality assessment (POLQA)[S]. 2011.
- [25] PAN Z Q, YU W J, YI X K, et al. Recent progress on generative adversarial network (GAN): a survey[J]. IEEE Access, 2019(7): 36322-36333.
- [26] KIM J Y, BU S J, CHO S B. Zero-day malware detection using transferred generative adversarial networks based on deep autoencoders[J]. Information Sciences, 2018(460/461): 83-102.
- [27] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. CoRR, 2015(1): 1-10.
- [28] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Mining of Massive Datasets, 2014(1):1-10.
- [29] 周文豪, 牛英滔, 施育鑫. 无线通信中的机器学习算法综述[J]. 无线电通信技术, 2022, 48(6): 1049-1057.
- ZHOU W H, NIU Y T, SHI Y X. Overview of machine learning algorithms in wireless communication[J]. Radio Communication Technology, 2022, 48(6): 1049-1057.

[作者简介]



赵宇翔(1993-), 男, 中国移动(浙江)创新研究院有限公司工程师, 主要研究方向为网络智能化技术。



纪雅欣(1997-), 女, 中国移动(浙江)创新研究院有限公司工程师, 主要研究方向为网络智能化技术。



余立(1981-), 男, 中国移动(浙江)创新研究院有限公司副院长, 主要研究方向为网络智能化技术。



周天一(1993-), 男, 中国移动(浙江)创新研究院有限公司工程师, 主要研究方向为网络智能化技术。



周航(1998-), 男, 中国移动(浙江)创新研究院有限公司工程师, 主要研究方向为网络智能化技术。