

# 团 体 标 准

T/CES XXX-XXXX

## 电力人工智能边端侧模型技术规范

Technical specification for electric power artificial intelligence side and end  
measurement model  
(征求意见稿)

XXXX-XX-XX 发布

XXXX-XX-XX 实施

中国电工技术学会发布

# 目 录

前言 .....	III
1 范围 .....	1
2 规范性引用文件 .....	1
3 术语和定义 .....	1
4 通用要求 .....	5
5 通用算子要求 .....	6
6 模型封装要求 .....	6
7 模型评价指标及方法 .....	9
附录 A (资料性附录) 常见框架算子对照表 .....	17
附录 B (资料性附录) 模型评价指标计算 .....	19

## 前 言

本标准按照 GB/T1.1—2009《标准化工作导则 第1部分：标准的结构与编写》给出的规则起草。  
请注意本文件的某些内容可能涉及专利，本文件的发布机构不承担识别这些专利的责任。

本标准由国网信息通信产业集团有限公司提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本标准起草单位：国网信息通信产业集团有限公司、安徽继远软件有限公司、福建亿榕信息技术有限公司、北京国网信通埃森哲信息技术有限公司、国网重庆市电力公司电力科学研究院、四川大学、四川中电启明星信息技术有限公司、国网重庆市电力公司、中国电力科学研究院有限公司。

本标准主要起草人：李强、赵峰、邱镇、刘迪、庄莉、李炳森、廖逍、黄晓光、向辉、王秋琳、余江斌、许中平、谭洪恩、苏少春、杨迎春、周孔均、王晓东、钟加勇、彭舰、田鹏、吕小红、厉仄平、黄飞虎、王金策、梁翀、郭庆、张琳瑜、崔迎宝、刘璟、宫晓辉、尹玉、周伟、梁翀、李温静、李道兴。

本标准为首次发布。

# 电力人工智能边端侧模型技术规范

## 1 范围

本标准规定了电力人工智能边端侧模型技术应用相关的通用要求、安全性要求、效率要求、算子要求、模型应用封装要求和模型评价指标及方法。本文件适用于电力输变配和安检等数据采集终端智能分析应用场景，提升电力边端侧数据汇集分析智能化水平，构建低延时、低带宽依赖、经济智能、数据可控和高可靠性的边端人工智能算力赋能体系。

## 2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.28 信息技术 词汇 第28部分：人工智能基本概念与专家系统

GB/T 34952 多媒体数据语义描述要求

## 3 术语和定义

GB/T 5271.28界定的以及下列术语和定义适用于本文件。

### 3.1 电力边端侧模型 Edge model of electric power

面向电力生产数据端侧（数据采集装置）和边侧数据（变电站、作业现场等小范围设备集控场景）筛选归集与数据分析任务，基于边缘硬件架构优化与深度学习技术构建的可集成在边侧或端侧具备基础算力设备上的模型，能够满足低时延、低功耗、低带宽的基本业务要求。

### 3.2 开放神经网络交换 ONNX

一种针对机器学习所设计的开放式的文件格式，用于存储训练好的电力人工智能边端模型。它使得不同的人工智能框架可以采用相同格式存储电力人工智能边端模型数据并交互。

### 3.3 应用容器引擎 Docker

开源的应用容器引擎，打包电力人工智能应用以及依赖包到一个可移植的镜像中，然后发布到任何流行的 Linux 或 Windows 操作系统的机器上，也可以实现虚拟化。容器是完全使用沙箱机制，相互之间不会有任何接口。

### 3.4 算子 Operator

是一个函数空间到函数空间上的映射  $O: X \rightarrow X$ 。广义上的算子可以推广到任何空间，如内积空间等。

### 3.5 功率分析仪 Power analyzer

用来测量电机、变频器、变压器等功率转换装置的功率、效率等参量。

### 3.6 人工智能 artificial intelligence

利用数字计算机或者由数字计算机控制的机器，模拟、延伸和扩展人类的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术和应用系统。

### 3.7 电力人工智能模型 artificial intelligence model

通过学习海量电力样本数据中的内在规律和表现层次，实现包括分类、回归、聚类、目标检测、光学字符识别、自然语言处理、语音识别等任务。

#### 3.8 分类模型 classification model

判断一个新的观察电力样本数据所属的电力类别的模型。

#### 3.9 回归模型 regression model

预测一个新的观察电力样本数据的连续型目标值的模型。

#### 3.10 聚类模型 clustering model

划分无标记的电力样本数据集为多个电力类别的模型。

#### 3.11 目标检测模型 object detection model

确定电力图像上目标所属电力类别和位置属性信息的模型。

#### 3.12 光学字符识别模型 optical character recognition model

将电力影像上的文字内容转换为直接可编辑文字的模型。

#### 3.13 自然语言处理模型 natural language processing model

人与计算机之间用自然语言进行有效通信的模型，用于人类与机器的交互，能够挖掘电力领域自然语言文本蕴含的电力相关信息和知识。

#### 3.14 语音识别模型 speech recognition model

将电力语音数据转换为机器可读数据（文本内容、二进制编码、字符序列等）的模型。

#### 3.15 声音检测类 sound detection and recognition

主要以检测电力设备运行声音为主要目的任务。

#### 3.16 文字输入类 text input

以文字录入为主要目的，要求把电力领域的语音转化为电力领域相关文字的任务。

#### 3.17 训练数据集 training set

电力人工智能边端模型构建过程中使用的电力数据集合。

#### 3.18 测试数据集 testing set

评估电力人工智能边端模型构建质量的电力数据集合。

#### 3.19 真正例 true positives

电力人工智能边端模型正确判定的正类电力样本。

#### 3.20 假正例 false positives

电力人工智能被边端模型判定为正类的负类电力样本。

### 3.21 真正例率 true positive rate

电力人工智能边端模型正确判定的正类电力样本占所有正类电力样本的比例。

### 3.22 假正例率 false positive rate

电力人工智能边端模型判定为正类的负类电力样本占所有负类电力样本的比例。

### 3.23 正确率 accuracy

电力人工智能边端模型判断正确的电力样本占所有电力样本的比例。

### 3.24 精确率 precision

电力人工智能边端模型正确判定的正类电力样本占所有边端模型判定的正类电力样本的比例。

### 3.25 召回率 recall

电力人工智能边端模型正确判定的正类电力样本占所有正类电力样本的比例。

### 3.26 F1 值 F1-score

电力人工智能边端模型的精确率和召回率的调和平均。

### 3.27 对数损失 logloss

电力人工智能边端模型决策所包含的信息量。

### 3.28 P-R 曲线 precision recall curve

以电力人工智能边端模型精确率为纵轴、召回率为横轴作图得到的曲线。

### 3.29 平均精确率 mean precision

电力人工智能边端模型判定所有电力领域类别的精确率的均值。

### 3.30 平均召回率 mean recall

电力人工智能边端模型判定的所有电力领域类别的召回率的均值。

### 3.31 平均精度 average precision

在 P-R 曲线下，电力人工智能边端模型的召回率从 0 到 1 各个点的精确率的均值，即 P-R 曲线下的面积。

### 3.32 平均精度均值 mean average precision

电力人工智能边端模型判定的所有电力领域类别的平均精度的均值。

### 3.33 平均绝对误差 mean absolute error

电力人工智能边端模型预测结果与目标值的差的绝对值的平均值。

### 3.34 均方误差 mean square error

电力人工智能边端模型预测结果与目标值的差的平方的平均值。

**3.35 均方根误差 root mean square error**

电力人工智能边端模型预测结果与目标值的差的平方的平均值的根。

**3.36 决定系数 r-squared**

描述回归方程与真实电力样本数据输出之间的相似程度。

**3.37 校正决定系数 adjusted r-squared**

描述回归方程与真实电力样本数据输出之间的相似程度，基于决定系数的调整。

**3.38 兰德系数 rand index**

电力人工智能边端模型划分正确的电力样本对数占所有电力样本对数的比例。

**3.39 调整兰德系数 adjusted rand index**

兰德系数的去均值归一化。

**3.40 互信息 mutual information**

描述两个变量之间重叠的信息量。

**3.41 调整互信息 adjusted mutual information**

一种基于互信息的聚类效果评价方法。

**3.42 轮廓系数 silhouette coefficient**

结合簇内凝聚度和分离度的一种聚类效果评价方式。

**3.43 平均编辑距离 average edit distance**

电力人工智能边端模型识别的字符串变换到标准字符串进行的插入、删除、替换操作次数的均值。

**3.44 字符识别准确率 character recognition accuracy**

电力人工智能边端模型正确识别字符数占所有识别字符数的比例。

**3.45 字符识别召回率 character recognition recall**

电力人工智能边端模型正确识别字符数占所有标准字符数的比例。

**3.46 文本行定位准确率 text line positioning accuracy**

电力人工智能边端模型正确定位的文本行数占所有文本行数的比例。

**3.47 文本行定位召回率 text line positioning recall**

电力人工智能边端模型正确定位的文本行数占所有标准文本行数的比例。

**3.48 词错误率 word error rate**

电力人工智能边端模型输出词序列与标准词序列的标准编辑距离占标准词序列中所有词语的比例。

**3.49 字错误率 character error rate**

电力人工智能边端模型输出字序列与标准字序列的标准编辑距离占标准字序列中所有字的比例。

**3.50 句错误率 sentence error rate**

电力人工智能边端模型输出错误句子的个数占所有句子的比例。

**3.51 双语评估替换 bilingual evaluation understudy**

用于评估电力领域自然语言处理生成类文本的质量，简称 **BLEU**。

**3.52 鲁棒性 robustness**

描述扰动、异常和危险情况下电力人工智能模型的工作能力。

**3.53 完备性 completeness**

电力人工智能边端模型具有算法框架、开发语言、模型版本、运行环境等完整信息以及电力模型文件及附属源信息齐全等。

**3.54 黑盒攻击 black box attack**

攻击者未知攻击电力人工智能边端模型的内部结构，训练参数，防御方法，通过一定规则构造攻击电力样本以完成攻击。

**3.55 白盒攻击 white box attack**

攻击者已知攻击电力人工智能边端模型的内部结构，训练参数，防御方法，构造特定的攻击电力样本以完成攻击。

**4 通用要求**

电力人工智能边端侧模型主要应用电力业务边端数据采集分析场景，需能够快速适配主流边缘计算硬件架构，支持基于不同运算框架之间的模型超参转换，边端推理效率和安全性方面符合基本要求。

**4.1 应用场景**

电力人工智能边端侧模型主要通过集成到输变配和安监的源端数据采集装置或边侧数据集控终端进行快速数据分析，提升电力数据分析处理效率。

**4.2 硬件架构适配**

电力人工智能边端侧模型根据实际应用场景和边端硬件架构不同需支持差异化的适配，需支持的硬件架构包括但不限于FPGA和ARM内核等。

**4.3 模型超参转换**

电力人工智能边端侧模型为适应不同运算框架下的超参转换需求，需满足ONNX或TensorRT转换标准，采用相同格式存储模型数据并交互。

**4.4 边端推理效率**

边端侧模型应考虑资源占用情况,运行算力要求不宜高于1TFLOPS,边端推理功耗不超过5W,单张图像的响应处理时间应小于0.5s。

#### 4.5 应用安全

安全性方面应符合Q/GDW 1597-2015标准的要求。

### 5 通用算子要求

边端侧应当支持常见的深度学习底层框架,边端侧模型算子根据模型实际应用场景,进行差异化组合应用,但边端侧算子必须满足底层框架相关API使用规则。各类常见深度学习框架如下表所示,具体的框架对应的常规算子详见附录1

**算子融合要求:**对于按顺序出现的基础算子如卷积、激活函数以及池化等,需支持按照算子出现的顺序进行单级融合或多级融合。

**自定义算子要求:**对于常用算子外的由大量基本操作构成的运算过程,需支持提供相应接口以完成自定义计算节点的搭建及调用。

**高性能计算要求:**对于常用算子及自定义算子不同的计算硬件需求,需支持不同硬件间的并行处理,并支持依据算子相关性进行数据传输和共享。

表 5-1 框架支持列表

深度学习框架	指令参数项	模型文件类型
TensorFlow	tf	.pb, .ckpt
Caffe	caffe	.onnx
ONNX	onnx	.protxt, .caffemodel
Pytorch	torch	.pt, .pth

### 6 模型封装要求

边端侧模型封装应考虑模型管理与数据贯通要求,实现与边缘智能服务平台或人工智能平台等管理系统的数据贯通,实现模型初始化消息队列服务和数据管理服务,并采用Docker容器化方式进行封装。

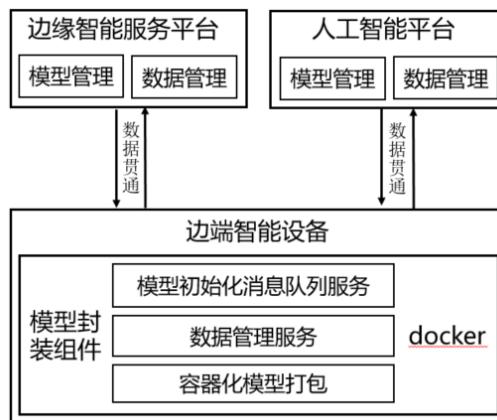


图 6-1 模型封装体系

#### 6.1 模型初始化服务

用于与管理系统进行交互，实现两级贯通与初始化算法需要输出的消息队列。

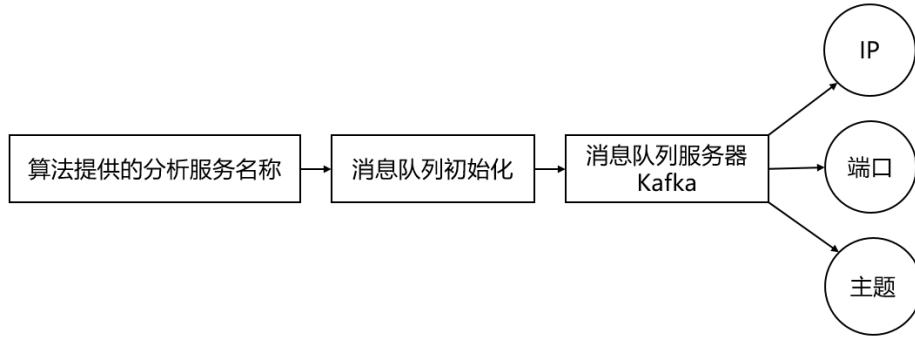


图 6-2 模型初始化服务流程

表 6-1 模型初始化服务接口

edge_computing_msg_queue_output_init	初始化需要输出的算法消息队列
edge_computing_msg_queue_output_deinit	反初始化需要输出的算法消息队列

## 6.2 数据管理服务

数据管理服务依托边缘侧模型推理服务为结构化数据或非结构化数据提供具备一定安全性的数据管理服务。

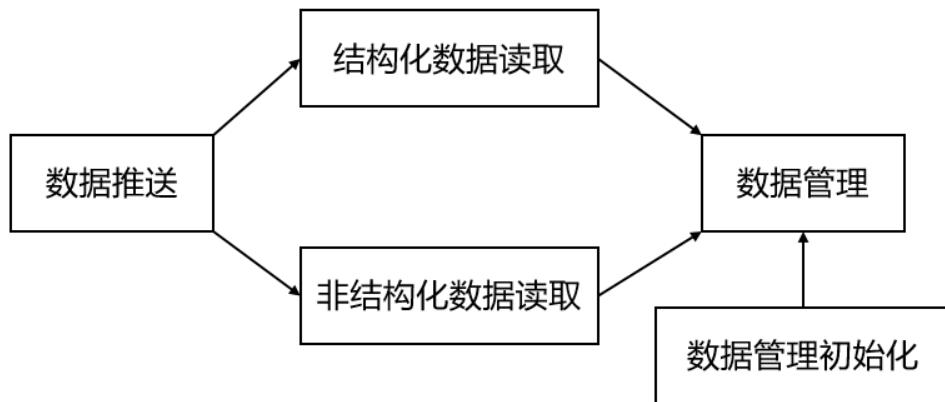


图 6-3 数据管理服务流程

表 6-2 数据管理服务接口

edge_computing_data_push	数据推送
edge_computing_structured_data_read	结构化数据读取
edge_computing_unstructured_data_re ad	非结构化数据读取

### 6.2.1 数据推送

推送结构化数据或非结构化数据进行数据管理。数据管理服务接口的type字段用于区分结构化数据和非结构化数据；其余字段按照接口说明传入参数，从而调用edge\_computing\_data\_push接口来实现对托管数据的推送。

数据推送接口如下：

### 6.2.2 结构化数据读取

结构化数据读取，主要根据其key字段，获取对应的value值。

结构化数据读取接口如下：

```
/* 结构化数据读取接口
 * @name 结构化数据读取
 * @param key: 结构化数据的 key 字段
 * @param address: 存储结构化数据 value 值的空间地址
 * @param capacity: 存储结构化数据 value 值的空间大小
 * @return 1:成功; 0:失败
 */
int edge_computing_structured_data_read(const char *key, void *address, size_t
capacity);
```

### 6.2.3 非结构化数据读取

非结构化数据读取，主要根据其存储地址，找到并读取其内容。

非结构化数据读取接口如下：

```
/* 非结构化数据读取接口
 * @name 非结构化数据读取
 * @param file: 非结构化数据的文件名称
 * @param path: 非结构化数据的存储地址
 * @return 1:成功; 0:失败
 */
int edge_computing_unstructured_data_read(const char *file, const char *path);
```

## 6.3 容器化模型打包

容器化模型打包通过Docker方式，定制化模型打包配制文件Dockerfile，通过执行docker build命令来构建镜像：

```
docker build -t <repo>:<tag> .
```

其中，-t为镜像的名字及标签，格式为<repo>:<tag>，Dockerfile的内容及docker build中的相关选项可参考Docker官方文档。

在完成制作镜像后，可以通过docker save 将指定镜像保存成 tar 归档文件：

```
docker save -o <repo_target>.tar <repo_source>:<tag>
```

其中，-o为指定输出tar文件，输出为<repo\_target>.tar即导出的镜像tar名称，<repo\_source>:<tag>为原始镜像的名称及标签。

在管理系统管理发布与初始化边缘侧模型过程中，除了构建完成的docker镜像包文件，还包含其它信息如镜像名称、镜像版本（用户自定义）以及Docker镜像启动参数（如 `--rm -i -t -d -P`）等。

## 7 模型评价指标及方法

边端测模型评价通用指标包括边端模型推理速度和边端模型推理功耗，针对不同的边端算法任务分为分类、回归、聚类、目标检测、光学字符识别、语音识别和自然语言处理模型评价指标。

### 7.1 模型评价体系

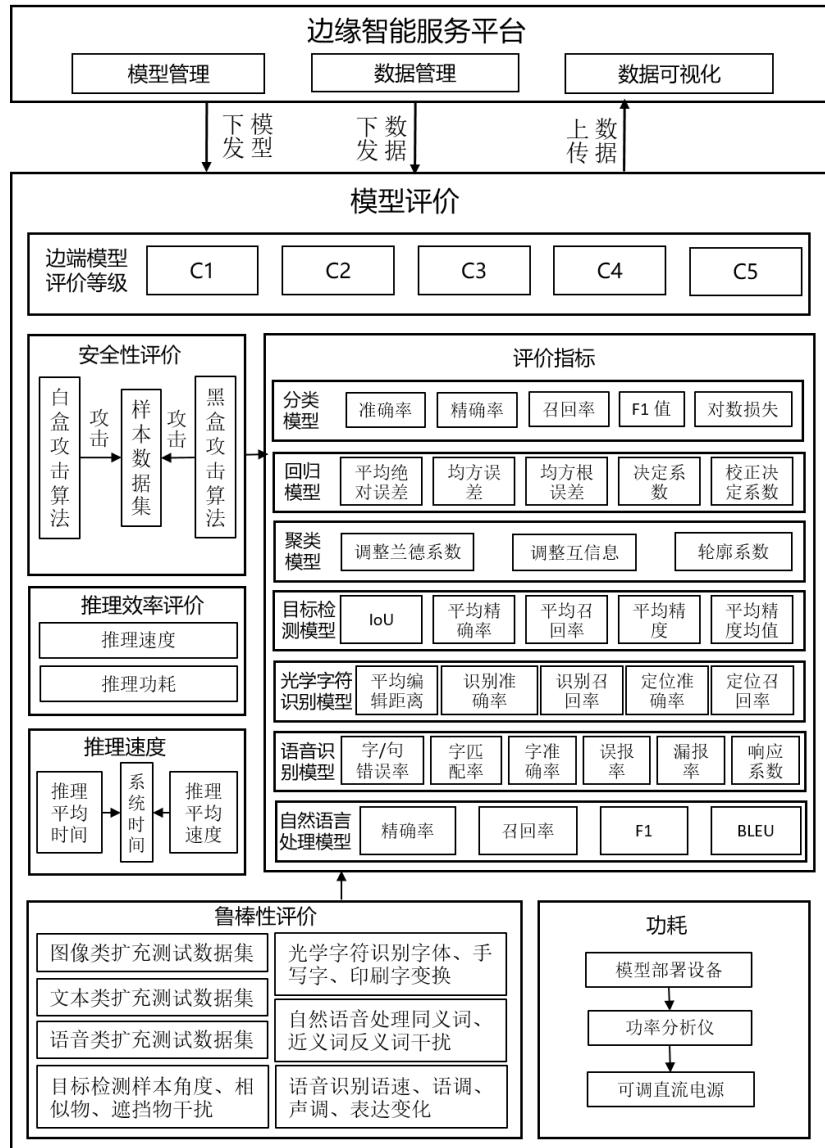


图 7-1 边端模型评价体系

边端模型评价主要包括边端模型评价等级、评价方法和评价指标三个大方面。边端模型评价的模型和测试数据从边缘智能服务平台下发至边缘设备中，通过测试集和扩充测试集数据计算各类边端模型的评价指标，获得模型的安全性、效率性、鲁棒性等指标，最终得到各边端模型的等级，并上传至边缘智能服务平台。

### 7.2 安全性评价

用于评价电力人工智能边端模型的安全程度，宜包括但不限于下列内容：

- a) 通过黑盒攻击算法构建攻击样本数据集，依据7.5计算边端模型攻击样本数据集上的功能性指标。
- b) 通过白盒攻击算法构建攻击样本数据集，依据7.5计算边端模型攻击样本数据集上的功能性指标。

### 7.3 鲁棒性评价

用于评价电力人工智能边端模型是否在样本变动时仍能维持性能，应包括但不限于下列内容：

- a) 图像类样本应经过旋转、裁剪、平移、模糊、加噪、缩放构建新的测试数据集，计算边端模型在新测试数据集上的功能性指标。
- b) 文本类样本应扩展、插值、交换、删除、分隔、词汇替换构建新的测试数据集，计算边端模型在新测试数据集上的功能性指标。
- c) 语音类样本应经过扩展、加噪构建新的测试数据集，计算边端模型在新测试数据集上的功能性指标。
- d) 目标检测类模型的鲁棒性评价样本应包含光照或角度变化、相似物与遮挡物干扰。
- e) 光学字符识别模型宜包含字体变换，手写与印刷字符变换等，可根据具体业务应用。
- f) 自然语言处理模型鲁棒性评价样本应包含同义词、近义词和反义词干扰，停用词干扰。
- g) 语音识别类模型的鲁棒性评价样本应包含语速、语调、声调、口音、表达方式变化。

### 7.4 模型推理效率评价

用于评价电力人工智能边端模型运行效率是否满足要求，宜包括但不限于下列内容：

- a) 边端模型推理速度：计算边端模型完成测试数据集推理的平均速度。
- b) 边端模型推理功耗：计算边端模型推理需要的功耗。

边端模型推理速度、边端模型推理功耗指标应作为边端模型推理效率性的重要参考指标，具体指标需根据业务实际需求确定。

## 7.5 评价指标

### 7.5.1 分类模型评价指标

用于评价电力人工智能分类模型实现的功能是否满足要求，宜包括准确率、精确率、召回率、F1值和对数损失等内容，具体计算过程见附录2.1。

### 7.5.2 回归模型评价指标

用于评价电力人工智能回归模型实现的功能是否满足要求，宜包括平均绝对误差、均方误差、均方根误差、决定系数和校正决定系数等内容，具体计算过程详见附录2.2。

### 7.5.3 聚类模型评价指标

用于评价电力人工智能聚类模型实现的功能是否满足要求，宜包括调整兰德系数、调整互信息、和轮廓系数等内容，具体计算过程见附录2.3。

### 7.5.4 目标检测模型评价指标

用于评价目标检测模型实现的功能是否满足要求，宜包括IoU大于或等于预设值、平均精确率、平均召回率、平均精度和平均精度均值等内容，具体计算过程见附录2.4。

### 7.5.5 光学字符识别模型评价指标

用于评价光学字符识别模型实现的功能是否满足要求，宜包括平均编辑距离、字符识别准确率、字符识别召回率、文本行定位准确率和文本行定位召回率等内容，具体计算过程见附录2.5。

### 7.5.6 语音识别模型评价指标

用于评价语音识别模型实现的功能是否满足要求，宜包括字错误率、句错误率、字匹配率、字准确率、误报率、漏报率、检出率和语音识别模型响应系数等内容，具体计算过程见附录2.6。

### 7.5.7 自然语言处理模型评价指标

用于评价自然语言处理模型实现的功能是否满足要求，宜包括精确率、召回率、F1和BLEU等内容，具体计算过程见附录2.7。

### 7.5.8 边端模型推理效率评价指标

#### (1) 边端模型推理速度评价指标

用于评价电力人工智能边端模型推理速度是否满足要求，宜包括但不限于内容：

- a) 计算边端模型完成测试数据集推理的平均时间。
- b) 计算边端模型完成测试数据集推理平均速度。

#### (2) 边端模型功耗评价指标

用于评价电力人工智能边端模型功耗是否满足要求，宜包括但不限于内容：

- a) 计算边端模型完成测试数据集推理的平均功耗。

## 7.6 评价方法

人工智能边端模型评价测试流程包括模型完备性评价、评价测试集选取、选择模型评价指标、模型封装测试和模型推理效率测试等五个步骤，如下图所示。

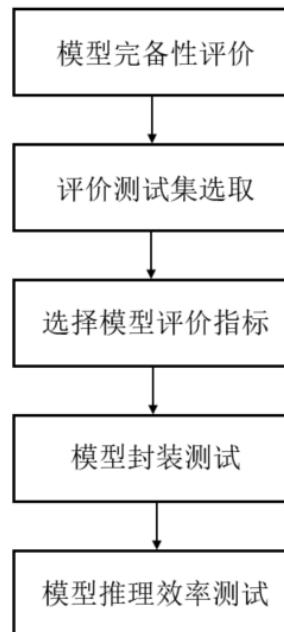


图 7-2 边端模型评价测试流程

### 7.6.1 模型完备性评价

具备完备性的人工智能边端模型应具备以下条件:

- a) 边端模型应有对应的模型描述文件,具体应包含模型名称、模型用途、运行模式、模型类型、模型运行环境、开发语言、开发框架、模型版本、模型提供单位、模型训练数据集规模等基本描述信息;
- b) 边端模型宜提供模型源文件和模型相关附属源文件等。

### 7.6.2 评价测试数据集选取

- a) 测试数据集应与训练数据集具有互斥性,即测试数据集与训练数据集不重合;
- b) 测试数据类型为图像文件时,图片宜为RGB三通道彩色图像,图片分辨率宜不低于500\*500像素,图片格式宜为jpg、png、jpeg、bmp、tif等;
- c) 分类模型与聚类模型测试数据集中所有类别样本数量比例建议相同,可根据具体业务需求进行调整;
- d) 目标检测模型测试数据集中包含目标样本与不包含目标样本比例建议为7:3,且每一个目标类别的样本数量宜不少于500张;
- e) 测试数据集样本标注信息应完备并准确无误,且应避免被人为添加的恶意数据污染。
- f) 语音测试样本格式宜为cpm、speex、speex-wb、mp3等,音频采样率宜不低于8k Hz,音频长度宜不少于3秒且不超过180秒,语言种类包括中文、英文、地方方言等;
- g) 自然语言处理模型的测试数据,宜为UTF-8纯文本格式文件,单次文本长度宜不超过5000字符(一个汉字、英文字母、标点符号,均记为一个字符),文件内宜根据具体需求涵盖单句、段落、文章、诗词等多种文学结构的文本。

### 7.6.3 选择模型评价指标

应结合具体业务应用场景和边端模型类型选择模型评价指标。具体选取规则如下:

- a) 边端模型评价宜包含模型指标评价、安全性、鲁棒性和效率性等内容;
- b) 边端模型指标评价至少应包含7.5中对应的指标;
- c) 安全性评价:基于边端模型安全测试样本,计算功能性指标,观察指标变化评价模型安全性;
- d) 鲁棒性评价:基于边端模型鲁棒性测试样本,计算功能性指标,观察指标变化评价边端模型鲁棒性。

### 7.6.4 模型封装测试

边端模型封装测试通过判断边端模型是否成功部署于边端,并根据边端返回的封装信息来判断模型封装是否成功。

### 7.6.5 模型推理效率测试

模型推理效率测试分为推理速度测试和功耗测试两部分。

#### (1) 边端模型推理速度评价测试

计算模型推理时间在执行模型推理语句前得到时间点,与得到模型推理结果后的时间点进行相减,得到推理时间,时间点可以通过调取系统时间来获得。

#### (2) 边端模型功耗评价测试

模型功耗测试方法如下:

##### (1) 能耗测试接线图

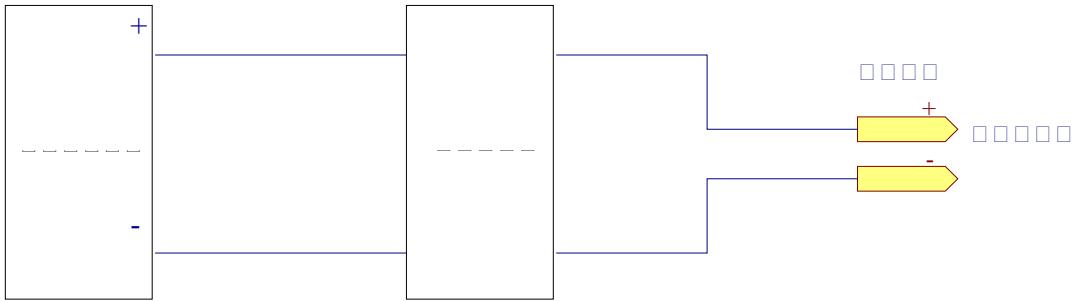


图 7-3 模型功耗测试电路图

可调直流电源为标准的可调稳压电源，输出电压根据待测设备的需求，可设置为3.2V的整数倍（即为3.2V、6.4V、9.6V等）。

功率分析仪全过程记录被测对象功耗情况。

电源接口采用XT60母口。待测前端设备应将内置电池去除，内部电源接口统一更换为XT60公口，接至设备外。

## （2）能耗测试方法

待测前端设备在收到统一的指令后，启动模型推理计算，开始进行不间断的识别运算，每识别完一份数据，立即上报结果，测试此阶段的总能耗。

## 7.7 模型性能等级判定

### 7.7.1 模型等级判定一般规则

- a) 本技术规范适用于边端模型部署应用前的入网评价和部署应用后的应用效果评价；
- b) 本技术规范对边端模型进行等级判定，实际应用中应考虑模型安全性、鲁棒性、效率性等其他因素。
- c) 本技术规范的模型等级是对业务应用功能相近的模型进行归类，不对模型的可用性进行定义，被测模型最终评价结果宜参考具体业务场景的相关规范。

### 7.7.2 分类模型等级

分类模型评价等级参考以下规则：

表 7-1 分类模型评价价值计算

指标判定	模型等级
准确率 $\geq 95\%$ ，精确率 $\geq 95\%$ ，召回率 $\geq 95\%$ ，F1值 $\geq 0.95$ ，对数损失 $\leq 0.7$ ， AUC $\leq 0.7$	C1
准确率 $\geq 85\%$ ，精确率 $\geq 85\%$ ，召回率 $\geq 85\%$ ，F1值 $\geq 0.85$ ，对数损失 $\leq 0.75$ ， AUC $\leq 0.75$	C2
准确率 $\geq 80\%$ 、精确率 $\geq 80\%$ ，召回率 $\geq 80\%$ ，F1值 $\geq 0.8$ ，对数损失 $\leq 0.8$ ， AUC $\leq 0.8$	C3
准确率 $\geq 75\%$ 、精确率 $\geq 75\%$ ，召回率 $\geq 75\%$ ，F1值 $\geq 0.75$ ，对数损失 $\leq 0.85$ ， AUC $\leq 0.85$	C4
准确率 $\geq 70\%$ 、精确率 $\geq 70\%$ ，召回率 $\geq 70\%$ ，F1值 $\geq 0.7$ ，对数损失 $\leq 0.95$ ， AUC $\leq 0.95$	C5

### 7.7.3 回归模型等级

回归模型评价等级参考以下规则:

表 7-2 回归模型评价值计算

指标判定	模型等级
决定系数 $\geq 0.90$	C1
决定系数 $\geq 0.85$	C2
决定系数 $\leq 0.80$	C3
决定系数 $\leq 0.75$	C4
决定系数 $\leq 0.7$	C5

### 7.7.4 聚类模型等级

聚类模型评价等级参考以下规则:

表 7-3 聚类模型评价值计算

指标判定	模型等级
调整兰德系数 $\geq 0.95$ , 调整互信息 $\geq 0.95$ , 轮廓系数 $\geq 0.95$	C1
调整兰德系数 $\geq 0.85$ , 调整互信息 $\geq 0.85$ , 轮廓系数 $\geq 0.85$	C2
调整兰德系数 $\geq 0.8$ , 调整互信息 $\geq 0.8$ , 轮廓系数 $\geq 0.8$	C3
调整兰德系数 $\geq 0.75$ , 调整互信息 $\geq 0.75$ , 轮廓系数 $\geq 0.75$	C4
调整兰德系数 $\geq 0.7$ , 调整互信息 $\geq 0.7$ , 轮廓系数 $\geq 0.7$	C5

### 7.7.5 目标检测模型等级

目标检测模型IOU值宜设置为0.5, 评价等级参考以下规则:

表 7-4 目标检测模型评价值计算

指标判定	模型等级
平均精确率 $\geq 95\%$ , 平均召回率 $\geq 95\%$ , 平均精度 $\geq 0.95$ , 平均精度均值 $\geq 0.95$	C1
平均精确率 $\geq 85\%$ , 平均召回率 $\geq 85\%$ , 平均精度 $\geq 0.85$ , 平均精度均值 $\geq 0.85$	C2
平均精确率 $\geq 80\%$ , 平均召回率 $\geq 80\%$ , 平均精度 $\geq 0.8$ , 平均精度均值 $\geq 0.8$	C3
平均精确率 $\geq 75\%$ , 平均召回率 $\geq 75\%$ , 平均精度 $\geq 0.75$ , 平均精度均值 $\geq 0.75$	C4
平均精确率 $\geq 70\%$ , 平均召回率 $\geq 70\%$ , 平均精度 $\geq 0.7$ , 平均精度均值 $\geq 0.7$	C5

### 7.7.6 光学字符识别模型等级

目标检测模型评价等级参考以下规则:

表 7-5 光学字符识别模型评价值计算

指标判定	模型等级
平均编辑距离, 字符识别准确率 $\geq 95\%$ , 字符识别召回率 $\geq 95\%$ , 文本定位准确率 $\geq 95\%$ , 文本定位召回率 $\geq 95\%$	C1
平均编辑距离, 字符识别准确率 $\geq 85\%$ , 字符识别召回率 $\geq 85\%$ , 文本定	C2

位准确率 $\geq 85\%$ , 文本定位召回率 $\geq 85\%$	
平均编辑距离, 字符识别准确率 $\leq 80\%$ , 字符识别召回率 $\geq 80\%$ , 文本定位准确率 $\leq 80\%$ , 文本定位召回率 $\leq 80\%$	C3
平均编辑距离, 字符识别准确率 $\geq 75\%$ , 字符识别召回率 $\geq 75\%$ , 文本定位准确率 $\geq 75\%$ , 文本定位召回率 $\geq 75\%$	C4
平均编辑距离, 字符识别准确率 $\geq 70\%$ , 字符识别召回率 $\geq 70\%$ , 文本定位准确率 $\geq 70\%$ , 文本定位召回率 $\geq 70\%$	C5

### 7.7.7 语音识别模型等级

语音识别模型评价等级参考以下规则:

表 7-6 语音识别模型评价值计算

指标判定	模型等级
字精确率 $\geq 95\%$ , 句错误率 $\leq 5\%$ , 响应系数 $\leq 1.1$	C1
字精确率 $\geq 85\%$ , 句错误率 $\leq 10\%$ , 响应系数 $\leq 1.2$	C2
字精确率 $\geq 80\%$ , 句错误率 $\leq 15\%$ , 响应系数 $\leq 1.3$	C3
字精确率 $\geq 75\%$ , 句错误率 $\leq 30\%$ , 响应系数 $\leq 1.4$	C4
字精确率 $\geq 70\%$ , 句错误率 $\leq 35\%$ , 响应系数 $\leq 1.5$	C5

### 7.7.8 自然语言处理模型等级

对于自然语言处理的分类任务, 评价等级可参考以下规则:

表 7-7 自然语言处理模型评价值计算

指标判定	模型等级
准确率 $\geq 95\%$ , 召回率 $\geq 95\%$ , F1值 $\geq 0.95$ ,	C1
准确率 $\geq 85\%$ , 召回率 $\geq 85\%$ , F1值 $\geq 0.85$ ,	C2
准确率 $\geq 80\%$ , 召回率 $\geq 80\%$ , F1值 $\geq 0.8$ ,	C3
准确率 $\geq 75\%$ , 召回率 $\geq 75\%$ , F1值 $\geq 0.75$ ,	C4
准确率 $\geq 70\%$ , 召回率 $\geq 70\%$ , F1值 $\geq 0.7$ ,	C5

对于自然语言处理领域的生成式任务, 因文本质量难以量化, 功能性评价可参考BLEU等生成式任务评价指标, 与其他同类型模型进行横向对比。

### 7.7.9 模型推理效率等级

边端模型推理速度等级参考以下规则:

表 7-8 边端模型推理速度评价值计算

指标判定	模型等级
推理速度 $\geq 0.05\text{s}/张$	C1
推理速度 $\geq 0.1\text{s}/张$	C2
推理速度 $\geq 0.2\text{s}/张$	C3
推理速度 $\geq 0.3\text{s}/张$	C4
推理速度 $\geq 0.4\text{s}/张$	C5

边端模型推理功耗等级参考以下规则:

表 7-9 边端模型推理速度评价值计算

指标判定	模型等级
功耗 $\geq 0.1\text{W}$	C1
功耗 $\geq 1\text{W}$	C2
功耗 $\geq 2\text{W}$	C3
功耗 $\geq 3\text{W}$	C4
功耗 $\geq 4\text{W}$	C5

## 附录 A

## (资料性附录)

## 常见框架算子对照表

表 A. 1 基于 Tensorflow 框架的算子对照表

边端算子	TF API	边端算子	TF API
ArgMax	tf.math.argmax	Input	tf.placeholder
BatchNorm	slim.batch_norm	LeakyRelu	tf.maximum(net * alpha, net) # alpha is const
	tf.layers.batch_normalization	Mean	tf.reduce_mean
	tf.nn.batch_normalization	Pad	tf.pad
Concat	tf.concat	Pooling	slim.max_pool2d
Const	tf.constant		tf.layers.average_pooling2d
Conv2D	slim.conv2d		tf.layers.max_pooling2d
	tf.layers.conv2d		tf.nn.max_pool
	tf.nn.conv2d	Relu	tf.nn.relu
DeConv2D	slim.convolution2d_transpose	Relu6	tf.nn.relu6
	tf.layers.convolution2d_transpose	Reshape	tf.reshape
	tf.nn.convolution2d_transpose	Sigmoid	tf.math.sigmoid
Eltwise	+	Softmax	slim.softmax
	tf.math.add		tf.nn.softmax
Flatten	slim.flatten	Upsample	tf.image.resize_bilinear
	tf.reshape		tf.image.resize_images method=BILINEAR
FullyConnected	slim.fully_connected		
	tf.layers.dense		

表 A. 2 Caffe 算子对照表

边端算子	Caffe Layer	边端算子	Caffe Layer
BatchNorm	BatchNorm	Input	Input
Concat	Concat	LRN	LRN
Conv2D	Convolution	Pooling	Pooling
Eltwise	Eltwise	Relu	Relu
FullyConnected	InnerProduct	Softmax	Softmax

表 A.3 Pytorch 算子对照表

边端算子	Pytorch OP	边端算子	Pytorch OP
BatchNorm	batch_norm	Pad	constant_pad_nd
Concat	cat	Pooling	max_pool2d
Conv2D	_convolution		avg_pool2d
Flatten	flatten	Relu	relu or relu_
FullyConnected	matmul	Reshape	view
GlobalAvgPool	adaptive_avg_pool2d	Softmax	softmax
LeakyRelu	leaky_relu_	Upsample	upsample_bilinear2d
Mean	mean		upsample_nearest2d

表 A.4 ONNX 算子对照表

边端算子	ONNX OP	边端算子	ONNX OP
ArgMax	ArgMax	Mean	ReduceMean
BatchNorm	BatchNormalization		GlobalAveragePool
Concat	Concat	Pad	Pad
Conv2D	Conv	Pooling	AveragePool
Eltwise	Add		MaxPool
	Sum	Relu	Relu
Flatten	Reshape	Reshape	Reshape
FullyConnected	Gemm	Softmax	Softmax
	MatMul	Upsample	Resize mode == nearest or bilinear
LeakyRelu	LeakyRelu		

## 附录 B

### (资料性附录)

#### 模型评价指标计算

##### 附录 B.1 分类模型评价指标计算

a) 正确率 *Accuracy*, 按式 (1) 计算:

$$Accuracy = (TN + TP) / (TN + TP + FP + FN) \quad (1)$$

b) 精确率 *Precision*, 按式 (2) 计算:

$$Precision = TP / (TP + FP) \quad (2)$$

c) 召回率 *Recall*, 按式 (3) 计算:

$$Recall = TP / (TP + FN) \quad (3)$$

d) F1 值, 按式 (4) 计算:

$$F1 = (2 \times Precision \times Recall) / (Precision + Recall) \quad (4)$$

e) 对数损失 (Logloss), 按式 (5) 计算:

$$Logloss = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij}) \quad (5)$$

其中, TP 表示预测正确的正样本, TN 预测是正确的负样本, FP 表示预测是错误的正样本, FN 表示预测是错误的负样本, N 为实例总数, M 为类别总数,  $y_{ij}$  是一个二值指标, 表示第 i 个输入实例是否是 j 类别 ( $y_{ij} = 1$  表示是, 反之为否),  $p_{ij}$  为分类模型预测出的第 i 个实例属于 j 类的概率。

f) *AUC*, 按式 (6) 和 (7) 计算:

$$AUC = \frac{\sum I(P_{\text{正样本}}, P_{\text{负样本}})}{M * N} \quad (6)$$

$$I(P_{\text{正样本}}, P_{\text{负样本}}) = \begin{cases} 1, & P_{\text{正样本}} > P_{\text{负样本}} \\ 0.5, & P_{\text{正样本}} = P_{\text{负样本}} \\ 0, & P_{\text{正样本}} < P_{\text{负样本}} \end{cases} \quad (7)$$

其中,  $P_{\text{正样本}}$  表示分类模型预测的正样本的概率,  $P_{\text{负样本}}$  表示分类模型预测的负样本的概率, M 表示测试集中正样本数量, N 表示测试集中负样本数量。

## 附录 B. 2 回归模型评价指标计算

a) 平均绝对误差 MAE, 按式 (8) 计算:

$$MAE = \sum |(\hat{y} - y)| / n \quad (8)$$

b) 均方误差 MSE, 按式 (9) 计算:

$$MSE = \sum (\hat{y} - y)^2 / n \quad (9)$$

c) 均方根误差 RMSE, 按式 (10) 计算:

$$RMSE = \sqrt{\sum (\hat{y} - y)^2 / n} \quad (10)$$

d) 决定系数  $R^2$ , 按式 (11) 计算:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \quad (11)$$

其中,  $y$  表示真实值,  $\hat{y}$  预测值,  $\bar{y}$  表示全部预测值的平均值。

e) 校正决定系数  $R^2_{adjusted}$ , 按式 (12) 计算:

$$R^2_{adjusted} = 1 - \frac{(1-R^2)(n-1)}{n-p-1} \quad (12)$$

其中, 式 (8) - (12) 中,  $y$  表示真实值,  $\hat{y}$  预测值,  $\bar{y}$  表示全部预测值的平均值,  $n$  表示测试集样本数量,  $p$  表示特征数量。

## 附录 B. 3 聚类模型评价指标

a) 调整兰德系数  $ARI$ , 按式 (13) 计算:

$$ARI = (RI - E(RI)) / (\max(RI) - E(RI)) \quad (13)$$

b) 调整互信息  $AMI$ , 按式 (14)、(15) 计算:

$$AMI = \frac{MI - E(MI)}{\max(H(U), H(V)) - E(MI)} \quad (14)$$

$$MI = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i, j) \log \left( \frac{P(i, j)}{P(i)P(j)} \right) \quad (15)$$

其中,  $U$ 、 $V$  是  $N$  个样本标签的两种不同分配情况, 注:  $P(i)$  表示  $U$  中类别为  $U_i$  的样本的概率, 即  $P(i) = \frac{|U_i|}{N}$ ,  $P(j)$  表示  $V$  中类别为  $V_j$  的样本的概率, 即  $P(j) = \frac{|V_j|}{N}$ ,  $P(i, j)$  表示在  $U$  中类别为  $U_i$ , 在  $V$  中类别为  $V_j$  的样本的概率, 即  $P(i, j) = |U_i \cap V_j| / N$ 。 $H(U)$  指的是数据集  $U$  的信息熵,  $H(U) = -\sum_{i=1}^{|U|} P(i) \log(P(i))$ ,  $H(V)$  指的是数据集  $V$  的信息熵。

c) 轮廓系数  $SC$ ，按式 (16) 计算：

$$SC = \frac{b(j)-a(i)}{\max\{a(i), b(j)\}} \quad (16)$$

其中， $a(i) = \text{average}(i)$ ， $i$  表示向量到所有它属于的簇中其它点的距离，计算  $b(j) = \min(j)$ ， $j$  表示向量到某一不包含它的簇内的所有点的平均距离。

#### 附录 B.4 目标检测模型评价指标

a) IoU 大于或等于预设值，判断为真正例；IoU 小于预设值，判断为真反例。IoU 预设值推荐为 0.5。

b) 平均精确率  $MP$ ，按式 (17) 计算：

$$MP = \frac{\sum \text{Precision}}{N} \quad (17)$$

其中， $N$  表示目标类别数，Precision 表示每类目标的精确率。

c) 平均召回率  $MR$ ，按式 (18) 计算：

$$MR = \frac{\sum \text{Recall}}{N} \quad (18)$$

d) 平均精度  $AP$ ，按式 (19) 计算：

$$AP = \int_0^1 p(r)dr \quad (19)$$

其中， $p(r)$  表示模型的P-R曲线函数。

e) 平均精度均值  $MAP$ ，按式 (20) 计算：

$$MAP = \frac{\sum AP}{N} \quad (20)$$

其中， $N$  表示目标类别数。

#### 附录 B.5 光学字符识别模型评价指标

a) 平均编辑距离  $AED$  可按式 (21) 计算：

$$AED = \frac{\sum \text{insert}(\text{result}, \text{label}) + \text{delete}(\text{result}, \text{label}) + \text{replace}(\text{result}, \text{label})}{N} \quad (21)$$

其中， $N$  表示测试数据集中的文本行数， $\text{result}$  表示模型识别出的一行文本， $\text{label}$  表示标准文本， $\text{insert}(\text{result}, \text{label})$  表示将  $\text{result}$  编辑为  $\text{label}$  需要执行的插入操作次数， $\text{delete}(\text{result}, \text{label})$  表示将  $\text{result}$  编辑为  $\text{label}$  需要执行的删除操作次数， $\text{replace}(\text{result}, \text{label})$  表示将  $\text{result}$  编辑为  $\text{label}$  需要执行的替换操作次数。

b) 字符识别准确率  $CRA$ ，按式 (22) 计算：

$$CRA = \frac{\sum C_{right}}{C_{all}} \quad (22)$$

其中,  $C_{right}$  表示识别正确的字符数,  $C_{all}$  表示总识别出的字符数。

c) 字符识别召回率  $CRR$ , 按式 (23) 计算:

$$CRR = \frac{\sum C_{right}}{C_{truth}} \quad (23)$$

其中,  $C_{right}$  表示识别正确的字符数,  $C_{truth}$  表示标准字符数。

d) 文本行定位准确率  $TLPA$ , 按式 (24) 计算:

$$TLPA = \frac{\sum T_{right}}{T_{all}} \quad (24)$$

其中,  $T_{right}$  表示定位正确的文本行数,  $T_{all}$  表示定位出的文本总行数。

e) 文本行定位召回率  $TLPR$ , 按式 (25) 计算:

$$TLPR = \frac{\sum T_{right}}{T_{truth}} \quad (25)$$

其中,  $T_{right}$  表示定位正确的文本行数,  $T_{truth}$  表示标准文本行数。

## 附录 B. 6 语音识别模型评价指标

(1) 设正确文本字数为  $N$ , 识别结果文本字数为  $M$ , 按照识别结果文本与正确文本根据“最小代价匹配”原则运用动态规划算法, 得到正确识别字数  $M_c$ 、删除错误字数  $D$ 、插入错误字数  $I$ 、替换错误字数  $R$ 、出错句子数  $S$  和句子总数  $S_n$ , 则有:  $N=M_c+R+D, M=M_c+R+J$ 。

定义以下性能指标:

a) 字错误率  $CER$ , 按式 (26) 计算:

$$CER = (I + R + D) / N \times 100\% \quad (26)$$

b) 句错误率  $SER$ , 按式 (27) 计算:

$$SER = S / S_n \times 100\% \quad (27)$$

c) 字匹配率  $MCR$ , 按式 (28) 计算:

$$MCR = M_c / N \times 100\% \quad (28)$$

d) 字准确率  $WCR$ , 按式 (29) 计算:

$$WCR = (M_c - 1) / N \times 100\% = 100\% - CER \quad (29)$$

(2) 用于评价声音检测识别类语音识别模型实现的功能是否满足要求, 宜包括下列内容:

该类型模型应兼顾动作的可靠性问题, 指标定义如下:

假设关键词表的词汇量为  $KW$ (个), 检测语音长度为  $HR$ (小时), 出现关键词  $N$ (次),  $C$  为每小时。每个关键词最大容忍的误报个数(一般取 10), 系统报出关键词  $M$ (个), 其中, 正确  $FD$ (个), 错误  $FA$ (个),

FD+FA=M。则：

a) 误报率  $Far$ ，按式 (31) 计算：

$$Far = FA / (KW \times HR \times C) \times 100\% \quad (30)$$

b) 漏报率  $Frr$ ，按式 (32) 计算：

$$Frr = (N - FD) / N \times 100\% \quad (31)$$

c) 检出率  $Fdr$ ，按式 (33) 计算：

$$Fdr = FD / N \times 100\% \quad (32)$$

等错率 EER：DET 曲线上  $Far=Frr$  时， $Far$  或  $Frr$  的值。

质量因数 FOM：ROC 曲线上  $Far=0\%, 10\%, 20\%, 40\%, 60\%, 80\%, 100\%$  时， $Fdr$  的算术平均值一般以 EER 或 FOM 值度量系统性能指标。

(3) 用于评价语音识别类模型响应时间，宜包括下列内容：

假设识别语音用时  $Tr$ ，语音实际时常  $Ts$ ，则：

语音识别模型响应系数  $RF$ ：

$$RF = T_r / T_s \quad (33)$$

## 附录 B.7 自然语言处理模型评价指标

(1) 针对情感分析、词义消歧、词性标注等分类任务，可根据样本的预测结果参照使用式 (1)、(3)、(4) 对模型的精确率、召回率和 F1 值进行评价。

(2) 对于机器翻译、摘要抽取等生成式任务，还可以使用 BLEU 等相对指标对模型性能进行评估，具体计算公式如下：

$$bleu_n = \frac{\sum_{c \in candidates} \sum_{n-gram \in c} Count_{clip}(n-gram)}{\sum_{c \in candidates} \sum_{n-gram \in c} Count(n-gram')} \quad (34)$$

其中， $candidates$  表示模型生成句子的集合， $reference$  表示给定的标准译文， $n$ -gram 表示长度为  $n$  的连续单词切片，对于分子，其第一个求和符号处理模型生成的所有句子，第二个求和符号处理生成句子中的每个  $n$ -gram， $Count_{reference}(n-gram)$  表示  $n$ -gram 切片在  $reference$  中的个数。故分子的含义为在给定句子中有多少  $n$ -gram 出现在标准译文中。分母的含义与分子相同，其统计了所有生成句子中的  $n$ -gram 总数。BLEU 可根据  $n$ -gram 的不同划分为多种评价指标，常见的有 BLEU-1、BLEU-2、BLEU-3、BLEU-4 四种，其中 BLEU-1 衡量的是单词级别的准确性，更高阶的 BLEU 可以衡量句子的流畅性。