ICS 35.240
L 70

团 体 标 准

T/BSIA 00X-2026

衍生数据认定方法 第 2 部分 通信行业

Method standard for testing the identification of derivative data
part 2: telecommunications industry

（征求意见稿）

2026-xx-xx 发布　　2026-xx-xx 实施

北京软件和信息服务业协会　发布

目　次

前言

本文件按照 GB/T 1.1—2020《标准化工作导则　第 1 部分：标准化文件的结构和起草规则》的规定起草。

《衍生数据认定方法》分为5个部分：

——第1部分：总则；

——第2部分：通信行业；

——第3部分：矿山行业；

——第4部分：医疗健康行业；

——第5部分：情感认知行业。

本部分为《衍生数据认定方法》第 2 部分。

本文件由北京市大数据中心提出，由北京软件和信息服务业协会归口。

本文件起草单位：智慧足迹数据科技有限公司、中国联合网络通信有限公司北京市分公司、北京市大数据中心、北京软件和信息服务业协会。

本文件主要起草人：施含章、赵华、吕振国、袁梦童、潘泼、刘文昊、王顺方、季爱生、赵章界、赵莹、方方。

衍生数据认定方法 第2部分 通信行业

1 范围

本文件规定了通信行业中衍生数据认定的认定环境、认定数据集要求、具体认定程序、判定准则以及认定报告的格式。

本文件适用于第三方检验检测或认证机构，以及通信运营商及其合作伙伴，在执行衍生数据认定认定任务时使用。

本文件覆盖的典型认定场景包括但不限于基于通信数据的位置洞察、用户画像构建、网络智能运维和客户关系管理等高敏感数据应用的场景。

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 35273 信息安全技术 个人信息安全规范

YD/T 3348-2018 电信网和互联网数据分类分级指南

XXX XXXX 衍生数据认定方法标准 第1部分：总则

3 术语和定义

下列术语和定义适用于本文件。

3.1

用户信令数据 User Signaling Data

用户终端与通信网络之间为建立、维持和释放通信连接而交互的控制类信息。

注：包括但不限于用户鉴权、位置更新、呼叫建立、短信收发、数据会话（PDP/PDU Session）激活等信令记录，通常包含时间戳、用户标识（如IMSI）、位置标识（如小区ID）等字段。

3.2

网络流量数据 Network Traffic Data

用户通过通信网络进行数据传输时产生的数据，也称DPI（Deep Packet Inspection）数据。

注：通常记录了用户上网行为的流量信息，如访问的应用/域名、协议类型、上/下行流量、访问时长等。

3.3

客户关系管理数据 Customer Relationship Management (CRM) Data

通信运营商在服务用户的过程中记录的与用户身份、业务、消费及互动相关的数据。

注：包括用户基本信息（年龄、性别）、套餐信息、账单、缴费记录、终端型号、入网时长、投诉记录等。

3.4

认定对象 Test Object

认定中需要被判定是否属于衍生数据的数据集、数据产品或数据服务输出。

3.5

源数据 Source Data

数据处理者对其享有使用权，在保护各方合法权益前提下，用于生成衍生数据的全部初始数据。

注：源数据包括数据处理者声称的及潜在的数据。

4 认定要点

衍生数据认定认定应遵循《衍生数据认定方法标准 第1部分：总则》中规定的客观性、可追溯性、全面性和审慎性原则。认定流程的核心是处理过程分析与量化计算分析相结合。

4.1 处理过程分析优先

应首先分析从源数据到认定对象的处理逻辑，识别是否存在《总则》7.1.2.2中定义的关键数据转换操作。

4.2 量化计算验证

应采用信息熵量化分析法，计算并评估认定对象相较于源数据的“显著差异”程度。

4.3 结论一致性

处理过程分析的定性结论与量化计算的定量结论应相互印证。若出现矛盾，应启动人工复核程序。

5 认定条件

5.1 认定环境要求

认定环境应满足执行数据处理和量化分析的需要，具体配置要求如下表所示。当受测方数据无法出域时，应由受测方提供可访问认定数据的认定环境。

表 1 认定环境要求

类别	项目	最低要求
硬件	CPU	8核或以上
	内存（RAM）	32 GB 或以上
	存储空间	1 TB 或以上可用空间
软件	操作系统	Linux (CentOS 7+, Ubuntu 18.04+) 或 Windows Server 2016+

类别	项目	最低要求
	数据库	支持SQL查询的关系型数据库（如PostgreSQL, MySQL）或大数据平台（如Hadoop/Spark）
	编程环境	Python 3.8+
	核心库	pandas, numpy, scipy, scikit-learn

5.2 典型认定数据集要求

认定所用的数据集应由被测方提供，并满足真实性和代表性要求。认定机构应对数据集进行抽样和核验。

表 2 典型认定场景数据集要求

认定场景	源数据（示例字段）	认定对象（示例形态）	数据集要求
位置洞察	信令数据: user_pseudo_id, timestamp, cell_id	区域人流热力图: grid_id, time_window, user_count	源数据应包含至少10万用户在24小时内的连续信令记录。
用户画像	CRM: user_pseudo_id, ARPU, tenure DPI: user_pseudo_id, domain_category, traffic	用户标签库: user_pseudo_id, tag_list	源数据应覆盖至少1万名用户的多维度信息，标签体系定义清晰。
网络智能运维	每秒采样：cpu_usage 累计值：in_bytes/out_bytes 单个告警：alarm_id 单次测量：latency	统计聚合趋势分析表：avg_cpu_5min，实时带宽利用率，alarm_count_hour，failure_pred_prob	至少覆盖业务高峰、平峰、低谷等不同场景，覆盖工作日、周末和节假日，样本点不少于1000个。
客户流失预警	CRM: user_pseudo_id, tenure, complaint_count 通话记录: user_pseudo_id, call_duration_monthly_avg	流失风险列表: user_pseudo_id, churn_probability	源数据应包含历史已流失和未流失用户样本，总量不少于5万条。

6 认定程序

6.1 认定准备

a) 任务接收与范围界定：接收认定委托，明确认定对象及其边界。

b) 资料收集：向被测方获取生成认定对象所需的全部资料，包括但不限于：

1) 源数据的描述、来源和样本。

2) 数据处理的完整脚本、代码或 ETL 工具的配置截图。

3) 相关的业务逻辑、算法模型说明文档。

c) 数据集准备：将被测方提供的源数据和认定对象数据导入认定环境，并进行随机抽样以备后续计算。

6.2 核心认定步骤

6.2.1 处理过程分析

a) 解析处理逻辑：审查 6.1 中收集的代码和文档，绘制或还原数据从输入到输出的完整处理流程图。

b) 识别关键操作：在流程图中标记出所有符合《总则》7.1.2.2 定义的关键数据转换操作（如 GROUP BY 聚合、机器学习模型应用、多源数据 JOIN 等）。

c) 形成初步结论：

1) 若识别出至少一项关键数据转换操作，则初步判定“符合衍生数据特征”。

2) 若未识别出任何关键数据转换操作，则初步判定“不符合衍生数据特征”。

6.2.2 量化计算分析

a) 本步骤旨在为步骤一的结论提供客观数据支撑。

b) 数据预处理：根据数据类型，对源数据和认定对象进行处理，使其适于概率分布计算。具体方法遵循《总则》附录 F.2.2。

c) 构建概率分布：分别计算源数据和认定对象的概率分布 P_orig 和 P_der。

d) 执行量化计算：根据规范性附录 B 中的代码，计算以下指标：

1) 信息熵 H_orig 和 H_der。

2) 信息熵变化率 ECR。

3) 相对熵 D_KL(P_der || P_orig)（若适用）。

注：当源数据与认定对象的事件空间（如分析维度）因聚合、映射等操作发生根本性改变时，D_KL不适用。此时，ECR和处理过程分析是主要判定依据。

6.2.3 结果判定与复核

a) 比对阈值：将计算出的 ECR 和 D_KL 与第 7 章定义的判定准则进行比较。

b) 综合判定：结合步骤一和步骤二的结果，形成最终判定结论。

1) 若处理过程分析为“符合”，且量化指标超过阈值，则最终判定为“认定为衍生数据”。

2) 若处理过程分析为“不符合”，且量化指标未超过阈值，则最终判定为“认定为非衍生数据”。

c) 启动复核：若定性与定量分析结论不一致，或认定对象涉及高敏感度数据，必须启动人工专家复核程序。

7 判定准则

7.1 定性判定准则

若认定对象的生成过程包含了《衍生数据认定方法标准 第1部分：总则》7.1.3.2节中定义的一种或多种关键数据转换操作，则初步定性判定为“符合衍生数据特征”。

7.2 定量判定准则

根据不同认定场景，量化指标（ECR或D_KL）超过下表中任一对应阈值，即定量判定为“存在显著差异”。

表 3 不同认定场景的定量判定准则

认定场景	主要处理技术	适用量化指标	ECR 阈值 δ	D_KL 阈值 ε	说明
位置洞察	时空聚合、统计计算	ECR	> 40%	不适用	输入输出的事件空间不同，ECR是核心指标。
用户画像	规则引擎、多源融合	ECR / D_KL	> 25%	> 0.3	可对比融合前后某一共同维度的分布变化。
网络智能运维	异常检测、关联规则挖掘	ECR / D_KL	> 20%	> 0.2	判定网络状态或根因，信息内涵发生改变。
客户流失预警	机器学习模型（分类/回归）	D_KL	不适用	> 0.15	对比输入特征分布与输出概率分数的分布差异。

注：y以上为基准阈值。阈值调整机制：针对特殊数据场景类型，可根据实际数据特征在±10%范围内动态调整，调整需经过3名及以上行业专家签字确认并记录调整理由。对于涉及个人信息、具有重大商业价值或社会影响的认定对象，认定机构应采取更严格的（即更低的）阈值，并应在认定报告中说明理由。

8 认定记录与报告

认定完成后，必须出具正式的认定报告。报告应至少包含以下内容，并可根据需要扩展。

8.1 基本信息

涵盖报告编号、委托单位、认定对象名称、认定日期等基本要素。

8.2 认定结论

应涵盖认定结论（认定为衍生数据 / 认定为非衍生数据 / 尚待进一步认定）和简要判定依据。例如，“经处理过程分析，识别出聚合运算关键操作；经量化计算，ECR为52.3%，超过40%的阈值”。

8.3 认定环境与数据集

包括硬件环境、软件环境、源数据描述和认定对象描述。

8.4 认定过程与结果

8.4.1 处理过程分析

应包括处理逻辑简述和识别的关键转换操作。其中处理逻辑简述宜画出流程图或给出关键代码片段，识别的关键转换操作宜列出具体操作，GROUP BY, ML Model Application等。

8.4.2 量化计算分析

应列举计算维度/特征、计算结果和采用的判定阈值。

8.5 附录

如适用，应涵盖复核记录，涵盖复核专家、意见和复核日期。

8.6 签章

认定报告应具备认定工程师、审核人签名及认定机构公章。

附 录 A
（规范性）
认定案例

A.1 案例一：基于用户信令数据的位置洞察报告认定

A.1.1 认定对象：“静安寺商圈工作日人流热力图”数据。

A.1.2 认定准备

a) 源数据：某工作日上海市静安寺区域内所有基站的信令日志样本（user_pseudo_id, timestamp, cell_id）。
b) 认定对象数据：按小时、按 100m×100m 地理网格聚合后的人流统计数据（grid_id, hour, user_count）。
c) 处理逻辑：被测方提供的 Python 脚本，显示主要操作为 GROUP BY grid_id, hour 和 COUNT(DISTINCT user_pseudo_id)。

A.1.3 认定执行

A.1.3.1 处理过程分析

识别出“空间映射”（cell_id到grid_id）、“分组聚合”（GROUP BY）和“统计计算”（COUNT DISTINCT）三项关键数据转换操作。初步判定“符合衍生数据特征”。

A.1.3.2 量化计算

a) 预处理：源数据以 cell_id 为维度，认定对象以 grid_id 为维度。
b) 构建概率分布：P_orig 为用户随机出现在各基站的概率分布。P_der 为用户随机出现在各网格的概率分布。
c) 计算：假设计算得出 H_orig = 10.2 bits（基站数量多，分布较均匀），H_der = 4.8 bits（聚合后的网格数量少，分布更集中）。
d) ECR = |4.8 - 10.2| / 10.2 = 5.4 / 10.2 ≈ 52.9%。
e) D_KL 不适用，因事件空间不同。

A.1.4 结果判定

ECR (52.9%) > 判定准则中的阈值 δ (40%)。
定性与定量结论一致。
认定结论：“静安寺商圈工作日人流热力图” 认定为衍生数据。

A.2 案例二：基于 CRM 与网络使用数据的客户流失预警模型输出认定

A.2.1 认定对象：高价值客户月度流失概率评分列表（user_pseudo_id, churn_probability）。

A.2.2 认定准备

a) 源数据：包含高价值客户的 CRM 数据（tenure, ARPU）和网络使用数据（avg_monthly_traffic）的特征表。
b) 认定对象数据：每个用户对应的流失概率值（0 到 1 之间）。
c) 处理逻辑：被测方提供文档，说明使用了梯度提升树（GBT）分类模型进行预测。

A.2.3 认定执行

A.2.3.1 处理过程分析

识别出“应用机器学习模型生成预测值”这一关键数据转换操作。初步判定“符合衍生数据特征”。

A.2.3.2 量化计算

a) 预处理

1) 源数据：选取最重要的输入特征 ARPU 进行分析。将其值域进行等频分箱，划分为 10 个等级（arpu_bin_1 到 arpu_bin_10）。

2) 认定对象：将输出的 churn_probability 值域也划分为 10 个等级（prob_bin_0.0-0.1 到 prob_bin_0.9-1.0）。

b) 构建概率分布

P_orig为用户在10个ARPU分箱中的分布。P_der为同一批用户在10个概率分箱中的分布。

c) 计算

假设计算得出 D_KL(P_der || P_orig) = 0.45。

ECR在此场景下意义不大，因为熵的变化可能不显著，但分布的“扭曲”是核心。

A.2.4 结果判定

D_KL (0.45) > 判定准则中的阈值 ε (0.15)。

定性与定量结论一致。

认定结论：“高价值客户月度流失概率评分列表” 认定为衍生数据。

附 录 B
（资料性）
量化计算代码示例

B.1 信息熵计算

```
import numpy as np
from collections import Counter

def calculate_entropy(data_series):
  """
  计算一维离散数据的香农信息熵。
  :param data_series: pandas Series或list, 包含离散的类别数据。
  :return: 信息熵值 (bits)。
  """
  if len(data_series) == 0:
    return 0.0

  counts = Counter(data_series)
  total_count = len(data_series)
  probabilities = [count / total_count for count in counts.values()]

  entropy = -np.sum([p * np.log2(p) for p in probabilities if p > 0])
  return entropy
```

B.2 相对熵（KL 散度）计算

```
import numpy as np
from collections import Counter

def calculate_kl_divergence(p_series, q_series):
  """
  计算两个离散概率分布的KL散度 D_KL(P || Q)。
  注意：两个series的事件空间（类别）必须相同。
  :param p_series: 代表P分布的原始数据序列。
  :param q_series: 代表Q分布的原始数据序列。
  :return: KL散度值 (bits)。
  """
  # 确保两个数据集大小相同以进行直接比较
  assert len(p_series) == len(q_series), "Data series must have the same length."

  p_counts = Counter(p_series)
  q_counts = Counter(q_series)

  # 获取所有事件（类别）的集合
  all_events = set(p_counts.keys()) | set(q_counts.keys())

  p_total = len(p_series)
  q_total = len(q_series)
```

```
    kl_divergence = 0.0
    for event in all_events:
      p_prob = p_counts.get(event, 0) / p_total
      q_prob = q_counts.get(event, 0) / q_total

      # 避免log(0)和除以0的错误
      if p_prob > 0:
        if q_prob == 0:
          return np.inf
        else:
          kl_divergence += p_prob * np.log2(p_prob / q_prob)
    return kl_divergence
```

参 考 文 献

[1] XXXXXXX
[2] XXXXXXX