I 65.020.20
C

团 体 标 准

T/CACM ****—20**

道地药材图谱检测数据处理基础信息模型

Fundamental Information Model for Daodi Medicinal Materials' Spectroscopic Profiling Data Processing

20**-**-**发布 20**-**-**实施

目　　录

1　范围 4
2　规范性引用文件 4
3　术语和定义 4
4　信息模型的核心实体 4
5　信息模型实体及属性字段定义 6

前　　言

《道地药材图谱检测数据处理基础信息模型》（以下简称“本标准”） 按照 GB/ T1.1-2009《标准化工作导则第 1 部分：标准的结构和编写》给出的规则起草。

本标准由 XX 提出。

本标准由 XX 归口。

本标准起草单位：XXX。

本标准主要起草人：XXX。

道地药材图谱检测数据处理基础信息模型

1 范围

本标准规定了道地药材图谱检测数据处理中所涉及的信息实体及语义关系，用于指导设计数据驱动的图谱检测和分析系统的底层数据存储结构。

2 规范性引用文件

无

3 术语和定义

下列术语和定义适用于本标准。

3.1 道地药材 Dodi Medicinal Materials

指经过中医临床长期应用优选出来的，产在特定地域，与其他地区所产同种中药材相比，品质和疗效更好，且质量稳定，具有较高知名度的中药材。

3.2 图谱检测 Spectroscopic Profiling

包括质谱（及其联用）、振动光谱（如拉曼、红外、紫外光谱）及核磁共振谱等检测方式。

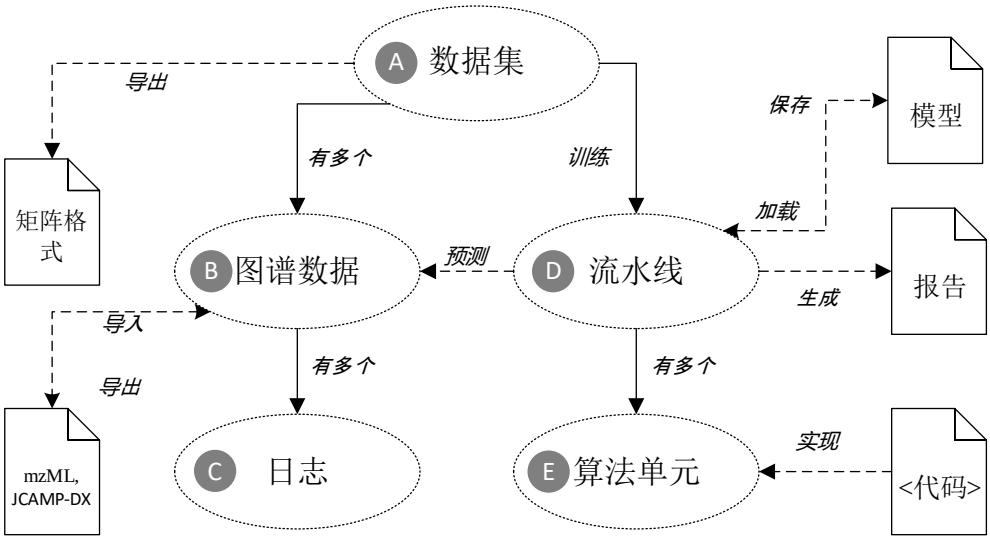
3.3 本体 Ontology

本体（ontology）起源于哲学，指组成现实（reality）的各类实体（entity）。在信息学领域，本体是对目标领域内客观实体的规范化表示。本体实现了计算机系统对于领域知识的一致性“理解”，为构建各种应用提供基础的语义支撑，是一种信息模型。

3.4 实体 Entity

客观物理世界的事物在计算机系统中的表征。

4 信息模型的核心实体



实线箭头表示底层数据库中具体的外键引用，虚线表示外部引用，如 URL 或资源路径。椭圆表示实体，文档图标表示外部或中间文件对象。

4.1 数据集 dataset

“数据集”是多个“图谱”实例的集合。一个数据集的图谱是面向同一检测主题的 (例如，分类牛奶品牌和识别特定的地道药材)，通过相同的检测模态（拉曼或 MALDI-TOF-MS），使用相同的数据预处理方法（如过滤、平均、基线校准) ，并具有相同的数据维数(如峰值数)。在机器学习情境中，“数据集”对应训练集，用于训练特定的预测模型。

“数据集”可以导出为矩阵或表格形式，供主要的科学数据分析平台导入，如 MATLAB、R 或 Python。在实际的系统操作中，这种中间数据格式更易于驱动整个图谱数据的分析过程。

4.2 图谱数据 spectroscopic profiling data

“图谱”表示一份图谱数据。一个图谱对象包含一个 x 值数组（如，用于拉曼的波数，或用于 MALDI-TOF-MS 的 m/z)和一个可选的 y 标签(在有监督数据分析的情况下)。图谱数据是信息模型的核心实体。

每个“图谱”实例可以序列化为第三方标准文件格式，如 mzML (MS)或 JCAMP-DX。对于第三方仪器系统，如 Agilent, Bruker, Horiba, Shimadzu, Thermo, Waters 等，这些标准文件格式可以用于交换和共享图谱数据。

4.3 日志

每个“图谱”实例有多个“日志”项，用于追踪数据状态的变化。该实体定义了图谱数据生命周期的几

个阶段，包括生成、预处理、审查、分析和报告。

4.4 流水线

“流水线”是一组算法单元组织起来的流程序列。每个“流水线”都针对于特定的数据集和分析目的。一个典型的图谱数据流水线通常包含若干预处理单元（如过滤、归一化、降维）及一个回归器/分类器。流水线在运行时（runtime）环境中被实例为复合模型（如特征选择+逻辑回归、支持向量机或神经网络），并由目标数据集训练。训练后的模型可以持久化到文件中(如 MATLAB 的.mat 文件或 python 的.pkl 文件)。此后，模型文件反序列化后可以加载回运行时环境中，对新样本进行预测分析后，可以生成人可读的报告和计算机可处理的结构化报告形式，服务于进一步的决策支持。

4.5 算法单元

算法单元包括基线漂移去除、平均滤波、特征缩放、特征选择、分类器、回归器等。每个算法单元需提供实现代码或伪代码。不同的算法单元针对不同的数据科学平台和编程语言可以有多种实现。算法工程师既可以直接调用使用现有的库，也可以上传编译后的二进制代码来实现。附录 A 列举了基本的算法单元，应内置到相关分析系统中。

5 信息模型实体及属性字段定义

5.1 数据集

描述：为相同的目的而生成的谱数据集合。具有相同的检测模态（拉曼或 MALDI-TOF-MS），采用相同的数据预处理方法（过滤、平均、识别、基线漂移去除等），具有相同的数据维度。

字段	类型	描述
数据集 ID	属性	唯一的 ID，主键。
数据集名称	属性	数据集的名称。
数据集检索代码	属性	拼音首字母缩写，用于快速检索。
检测对象	属性	样本来源，例如，婴儿奶、马肉、特定草药等，建议使用公共术语对对象进行编码，如 FOODON
检测主题	属性	如，品牌鉴别、产地鉴别、有害物质检测等
SOP	属性	产生该检测数据集的标准操作规程，包括采用的样品预处理步骤、仪器参数等

检测模态	属性	从以下取值中选择：
拉曼	值	拉曼光谱。
MS	值	质谱。
MALDI_TOF_MS	值	基质辅助激光解析/电离化飞行时间质谱。
SELDI_TOF_MS	值	表面增强型激光解吸/电离化时间飞行质谱。
IMS	值	离子迁移率光谱法。
NIRS	值	近红外光谱法。
FIRS	值	远红外光谱法。
SPI_MS	值	单光子电离质谱。
设备	属性	使用的仪器和客户端软件版本。
原始文件路径	属性	设备端导出的原始数据文件路径。
图谱数据	属性	该数据集所有图谱数据的二进制存储。
X 标签	属性	数据集的 X 标签。
Y 标签说明	属性	数据集的 Y 标签映射表，如 1 – 合格，0 – 不合格。
Y 标签样本分布	属性	每个 Y 标签的样本数量分布，使用 json 格式。
时间戳	属性	时间戳。

5.2 图谱数据

描述：表示一个图谱数据，通常是预处理后的。

字段	类型	描述
图谱 ID	属性	统一属性 ID，主键。
导出文件的缓存路径	属性	矩阵、表格文件或 mzML、JCAMP-DX 等格式。服务端缓存的（如果已经存在，则不会重新创建）文件路径，可以供外部科学数据分析平台导入，如 MATLAB、R 或 Python。
摘要	属性	数据的数字指纹或摘要。
Y 标签	属性	此数据的类别或 Y 标签。用于有监督学习。
序列	属性	图谱数据的压缩字节数组。
检测模态	属性	测试/检测方式。
X 轴含义	值	X 轴的物理化学意义。
X 轴单位	属性	X 轴单位。如拉曼的 cm^{-1}。

日志记录	属性	日志集合的导航属性，用于跟踪数据的历史状态变化。
元数据	属性	该数据相关的其他元数据。可以序列化为 json 或 xml 对象。
时间戳	属性	时间戳。

5.3 日志

描述：跟踪图谱数据状态变化。

字段	类型	描述
日志 ID	属性	唯一的 ID，主键。
操作员	属性	导致数据状态变更的操作者，如审核人。
操作	属性	下列枚举值之一。
生成/获取	值	
预处理	值	
审核	值	
分析	值	
报告	值	
操作场所	属性	执行该操作的机构或实验室。
附加信息	属性	附加消息或补充说明。
图谱 ID 外键	属性	指向相关图谱对象的外键。
日志的时间戳	属性	日志条目的创建时间戳。

5.4 流水线

描述：流水线是一组算法模块组成的数据分析流程，如“特征提取+分类+可视化”。流水线用于构建一个串行化的机器学习模型，经过数据集训练后可用于新样本的预测分析。

字段	类型	描述
流水线 ID	属性	唯一的 ID，主键。
流水线名称	属性	流水线的名称。
流水线检索代码	属性	首字母缩写，用于快速搜索。
描述	属性	流水线的文献或文档。
流水线 URL	属性	该算法流水线的知识库 URL
流水线说明	属性	对流水线的描述。

流水线元数据	属性	流水线元数据。可以序列化的 json 对象。
流水线模板	属性	可以用快速实例化的流水线模板，如 ipynb 格式的文件
流水线时间戳	属性	最新修订的时间戳。

5.5 算法单元

描述：算法单元或模块，是构造流水线的基本元素。

字段	类型	描述
算法	实体	唯一的 ID，主键。
算法 ID	属性	该算法来自公共实现还是自定义实现。
算法来源	属性	
算法名称	属性	算法的名称
算法检索代码	属性	首字母缩写，用于快速搜索。
算法类别	属性	算法的类别。取值为以下之一
预处理	值	
降维	值	
特征选择	值	
回归	值	
分类	值	
聚类	值	
可视化	值	
算法标签	属性	该算法附加的自定义标签。
算法参考文献	属性	说明算法内部设计原理和实现细节等。
算法的 URL	属性	知识库链接。
算法描述	属性	对该算法的简要描述。
算法元数据	属性	关于该算法的元数据。序列化的 json 或 xml 对象。
算法的实现	属性	用于算法实现的编程语言或脚本。
Python	值	
C/C++	值	
C#	值	
Javscript	值	

R	值	
Java	值	
Matlab	值	
Octave	值	
Go	值	
算法代码	属性	该算法的代码片段或伪代码。
算法时间戳	属性	最新修订的时间戳。

附录 A 内置算法单元

算法单元	类别	默认实现
standard scaler 标准化缩放	preprocessing 预处理	sklearn.preprocessing.StandardScaler
re-scaler 一般缩放		sklearn.preprocessing.MinMaxScaler
imputer 异常值处理		sklearn.preprocessing.Imputer
LASSO (least absolute shrinkage and selection operator) 最小绝对收缩和选择操作符	feature selection 特征选择	sklearn.linear_model.Lasso
elastic net 弹性网络		sklearn.linear_model.ElesticNet
linear regressor 线性回归	regression 回归	sklearn.linear_model.LinearRegression
ridge regressor 岭回归		sklearn.linear_model.Ridge
ANOVA (analysis of variance) 方差分析	mean test 均值检验	scipy.stats.f_oneway
MANOVA (multivariant ANOVA) 多元方差分析		statsmodels.multivariate.manova
logistic regression 逻辑回归	classification 分类	sklearn.linear_model.LogisticRegression
support vector classifier 支持向量机		sklearn.svm.SVC
k-neighbors classifier K 最近邻		sklearn.neighbors.KNeighborsClassifier
linear discriminant analysis 线性判别分析		sklearn.discriminant_analysis.LinearDiscriminantAnalysis

autoencoder 自编码器	dimension reduction 降维	keras
VAE (variational autoencoder) 变分自编码器		keras
PCA (principal component analysis) 主成分分析		sklearn.decomposition.PCA
NMF (non-negative matrix factorization) 非负矩阵分解		sklearn.decomposition.NMF
MDS (multidimensional scaling) 多维缩放		sklearn.manifold.MDS
t-SNE (t-distributed stochastic neighbor embedding) t 分布随机邻域嵌入		sklearn.manifold.TSNE
boxplot 箱图	visualization 可视化	matplotlib.pyplot.boxplot
histogram 频率直方图		matplotlib.pyplot.hist