ICS 35.020
L70

YD

中华人民共和国通信行业标准

YD/T 3762—2020

大数据 数据挖掘平台技术要求与测试方法

Big data - technical specification and test methods on data mining platform

2020-08-31 发布　　2020-10-01 实施

中华人民共和国工业和信息化部　发布

目　　次

前　言

本标准是大数据系列标准之一，该系列标准名称和结构如下：

——大数据 分布式批处理平台技术要求与测试方法

——大数据 分布式分析型数据库技术要求与测试方法

——大数据 分布式事务型数据库技术要求与测试方法

——大数据 分布式流处理平台技术要求与测试方法

——大数据 时序数据库技术要求与测试方法

——大数据 商务智能分析工具技术要求与测试方法

——大数据 数据管理平台技术要求与测试方法

——大数据 数据集成工具技术要求与测试方法

——大数据 数据挖掘平台技术要求与测试方法

本标准按照 GB/T 1.1—2009 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。

本标准由中国通信标准化协会提出并归口。

本标准起草单位：中国信息通信研究院、中兴通讯股份有限公司、中国电信集团有限公司、星环信息科技（上海）有限公司、中移（苏州）软件技术有限公司、腾讯云计算（北京）有限责任公司、华为技术有限公司、中国科学院计算技术研究所。

本标准主要起草人：魏凯、姜春宇、马鹏玮、王卓、刘蔚、常宝玲、李新龙、林智、裴非、张勇、刘俊良、朱松、詹剑锋、王磊、查礼。

大数据 数据挖掘平台技术要求与测试方法

1 范围

本标准规定了用于对大数据进行数据挖掘的软件平台或服务应具有的技术要求及相关的测试方法。

本标准适用于数据挖掘平台产品的设计、研发、测试、评估和验收、科学大数据系统的测试，用于金融、电信、能源、公共安全等行业数据挖掘平台的测试和选型。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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 5271.31—2006 信息技术 词汇 第31部分：人工智能 机器学习

GB/T 35295—2017 信息技术 大数据 术语

YD/T 1212—2016 大数据 分布式批处理平台技术要求与测试方法

3 术语定义及缩略语

3.1 术语定义

3.1.1

数据挖掘 data mining

从大量的数据中通过算法搜索隐藏于其中信息的过程。

注：数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

3.1.2

数据挖掘平台 data mining platform

集成了常见数据挖掘应用场景以及技术组件的平台化产品，从而使业务人员能够便捷创建数据挖掘业务。

3.1.3

分布式文件系统 distributed file system

能够管理分布在多个节点上的文件的文件管理系统，节点间通过分布式系统中的网络进行通信和数据传输。

[YD/T 1212—2016，定义 3.1.7]

3.1.1

结构化数据 structured data

一种数据表示形式，按此种形式，由数据元素汇集而成的每个记录的结构都是一致的并且可以使用关系模型予以有效描述。

[GB/T 35295—2017，定义 2.2.13]

3.1.2

非结构化数据 unstructured data

不具有预定义模型或未以定义方式组织的数据。

[GB/T 35295—2017，定义 2.1.25]

3.1.3

机器学习 machine learning

功能单元通过获取新知识或技能，或通过整理已有的知识或技能来改进其性能的过程。

[GB/T 5271.31—2006 定义 31.1.2]

3.2 缩略语

下列缩略语适用于本标准。

CNN	卷积神经网络	Convolutional Neural Networks
CSV	字符分隔符	Comma-Separated Values
CPU	中央处理器	Central Processing Unit
DNN	深度神经网络	Deep Neural Networks
FPGA	现场可编程门阵列	Field－Programmable Gate Array
GAN	生成式对抗网络	Generative Adversarial Networks
GPU	图形处理器	Graphics Processing Unit
kNN	K 最近邻	k-Nearest Neighbor
LSTM	长短期记忆网络	Long Short-Term Memory
RNN	循环神经网络	Recurrent Neural Network
SVM	支持向量机	Support Vector Machine

4 总体要求

数据挖掘平台及其组件主要功能是提供数据挖掘业务的自动化、智能化创建及管理服务。

4.1 参考架构

数据挖掘平台应包括如下功能模块：

a） 数据导入、导出功能组件；

b） 用户权限管理模块；

c） 数据及模型管理组件；

d） 算法管理组件；

e） 工作流管理模块；

f） 平台运维模块。

4.2 功能要求

数据挖掘平台应满足如下功能和性能要求：

a） 具备基本的管理能力，包括数据源支撑、用户管理、训练结果和流程的导出复用、训练结果通知等能力；

a） 具备核心算法能力，包括支持预处理、数据挖掘、文本分析、特征工程等算法的能力，以及建模案例、算法详细说明等支持能力；

b） 具备故障告警、高可用、日志等运维常用功能；

c） 具备并行化及兼容能力，支持开发接口，包括算法并行化、任务并行化、GPU 加速和异构硬件兼容能力。

4.3 性能要求

数据挖掘平台应满足执行不同日常数据挖掘任务的性能要求，衡量性能的指标包括算法效果评价指标、算法任务执行时间、鲁棒性等。

5 技术要求

5.1 管理能力要求

数据挖掘平台应具备基本的管理能力，具体包括：

a） 支持将结构化和非结构化的数据库、文件等多种形式数据作为数据挖掘任务数据源的能力；

b） 支持对于不同用户进行不同授权，用户各自的数据及模型相互隔离、训练过程相互独立，实验模板可进行共享等能力；

c） 支持将用户数据、训练结果、实验模板、训练模型等进行保存并能够通过多种方式进行导出的能力；

d） 支持对于用户数据、训练结果、实验模板、训练模型等进行存储、查找、删除、重命名、另存为等管理能力；

e） 支持在实验执行完成时通过邮件或短信等方式对用户进行通知的能力。

5.2 核心能力要求

数据挖掘平台应具备完整且强大的算法能力，同时对于算法执行过程的周边功能也需要有完备的支撑，具体包括：

a） 支持通过如空值插补、去重、字段过滤等在内的多种预处理算法对于数据进行预处理的能力；

b） 支持通过如协方差矩阵、方差、标准差等在内的多种统计分析类算法对用户数据进行统计分析的能力；

c） 支持使用如 Kmeans、线性回归、决策树等在内的多种机器学习算法执行机器学习任务等能力，算法类型包括但不限于文本分析、分类、聚类、回归、推荐、关联分析等；

d） 支持通过特征工程类算法对训练数据进行处理的能力，算法类型包括但不限于特征离散、特征向量切割、随机欠抽样等；

e） 支持通过可视化拖拽的方式进行建模的能力；

f） 支持在可视化建模的基础上能够对建模中关键信息进行修改和配置的调优能力；

g） 支持通过多种评价指标，对于算法模型结果进行评估的能力；

h） 支持通过多种可视化展示方式，对于已有模型结果进行展示的能力；

i） 支持通过平台提供的模板、案例来进行模型快速创建的能力；

j） 支持在进行建模时，可以提供算法的教程式解释信息以协助进行建模的能力；

k） 支持在可视化建模的基础上，可以对模型中的算法算子进行如顺序执行、并行执行、判断执行等方式编排的能力；

l） 支持对模型训练任务进行调度配置实现自动周期性执行的能力。

数据挖掘平台宜具备其他算法能力以及周边支持能力，具体包括：

a） 支持使用如 CNN、RNN、GAN 等在内的深度学习算法执行深度学习任务的能力；

b） 支持通过直接编写或接口调用的形式实现执行用户自定义算法的能力；

c） 支持对于模型训练任务执行过程进行日志记录，并对于执行日志能够进行查询、展示的能力；

d） 支持可以使用如 python、java、R 等在内的编程语言进行自定义算法编写的二次开发语言兼容能力。

5.3 运维能力要求

数据挖掘平台应具备一定的运维能力，具体包括：

a） 支持在模型训练或执行过程中出现错误时，通过邮件或短信进行远程通知的能力；

b） 支持对于用户数据、模型及执行结果进行备份并在数据意外失去时进行恢复的能力；

c） 支持对于数据挖掘平台日常登录、数据和模型的新增及删除、系统异常等信息的日志记录能力。

数据挖掘平台宜具备训练过程高可用能力，具体为：

a） 支持在模型训练过程中途出现中断，系统能够自动重启或继续完成训练的能力。

5.4 并行化及兼容能力要求

数据挖掘平台应具备算法和任务的并行化能力，具体包括：

a） 支持在实验中包含多个不同算法，算法之间属于不同分支，可以同时并行执行的能力；

b） 支持可以同时并行执行多个实验任务的能力。

数据挖掘平台宜具备使用 GPU 加速计算以及同时兼容异构计算资源的兼容能力，具体包括：

a） 支持能够应用 GPU 对机器学习算法的模型训练过程进行加速的能力；

b） 支持对于如 CPU、GPU、FPGA 等在内的多种异构计算资源进行兼容使用的能力。

5.5 性能要求

数据挖掘平台应具备正常执行不同日常数据挖掘任务的性能水准，具体包括：

a） 在执行机器学习算法时应达到一定执行速率，即在一定数据量的数据集上执行特定算法的执行时间；

b） 在执行机器学习算法时算法结果应达到一定准确率，即在特定数据集上执行特定算法的算法结果中结果正确的数据项占总数据项的比例；

c） 在执行机器学习算法时能够保持一定的鲁棒性，即在具有一定数据量的特定数据集上多次执行特定算法的执行速率和算法结果准确率维持在同一水平，不会出现较大幅度波动。

6 测试方法

本章规定了数据挖掘平台各类技术要求的测试方法，包括管理能力、核心能力、运维能力、并行化和兼容能力以及性能。

6.1 管理能力测试方法

6.1.1 数据源支持

测试编号：6.1.1
测试项目：数据源支持
测试目的：验证数据挖掘平台支持以包括结构化和非结构化数据、分布式文件系统、文件等多种形式的数据存储形式作为模型训练数据源导入使用
预置条件： 1）数据挖掘平台测试环境； 2）预置多种数据源
测试步骤：以读取本地 CSV 文件作为数据源为例，其他方式数据源类似 1）从本地 CSV 文件读取数据； 2）查看数据读取结果； 3）对比数据读取结果与原数据相同
预期结果： 1）从本地 CSV 文件读取数据成功，数据集可用于模型训练与测试； 2）结果展示正确，能够展示用户数据

6.1.2 用户权限管理

测试编号：6.1.2
测试项目：用户权限管理
测试目的：验证数据挖掘平台支持授权管理，各用户数据隔离及共享
预置条件： 当前系统对接对应认证服务器，并且存在多用户并可登录

测试编号：6.1.2
测试步骤： 1）以 user1 用户登录数据挖掘平台系统； 2）以 user2 使用其他浏览器或测试机登录相同数据挖掘平台系统； 3）user1、user2 查看各自能够使用的数据； 4）user1、user2 创建各自模型训练过程； 5）user1、user2 同时执行各自模型训练过程； 6）user1、user2 创建各自的实验模板
预期结果： 1）user1、user2 只能查看和使用自己用户的数据； 2）user1、user2 创建的流程只有当前用户可以查看和执行； 3）user1、user2 流程和执行相互独立不受影响； 4）user1、user2 生成的模型只有各自用户才能看到； 5）user1、user2 创建的实验模板可共享，在所有的用户都展示(可选)

6.1.3 数据及模型输出

测试编号：6.1.3
测试项目：数据及模型输出
测试目的：验证数据挖掘平台支持： 1）用户数据、训练结果及模型的多种导出形式； 2）用户数据集支持：页面直接下载、另存为表、数据表保存为 HDFS 文件等； 3）用户实验支持另存为模板，模板支持导入导出功能 4）训练好模型型支持导入、导出、另存为其他模型
预置条件： 1）数据挖掘平台测试环境； 2）当前系统已存在创建好的实验，并且已执行成功； 3）当前系统已存在训练好的模型； 4）当前系统已存在导入或保存的模板
测试步骤： 1）进入已经训练完成的实验； 2）选择 Spark 表或其他入口算子，下载用户的数据； 3）选择其他过程或结果算子，下载结果文件； 4）实验另存为模板； 5）选择某已有模板，可选择下载模板； 6）选择某已有模型界面，下载模型

测试编号：6.1.3
预期结果： 1）用户数据集可下载至本地，并且文件正确； 2）过程数据集和结果数据集可下载至本地，并且文件正确； 3）模板可下载存储； 4）模型可下载

6.1.4 数据及模型管理

测试编号：6.1.4
测试项目：数据及模型管理
测试目的：用户数据、训练结果及模型的存储、查找、删除等
预置条件： 1）当前系统已存在创建好的实验，并且已执行成功； 2）当前系统已存在训练好的模型； 3）当前系统已存在导入或保存的模板
测试步骤： 1）可以查找需要的表数据； 2）用户任意数据，通过存储为某种格式； 3）进入任意模型界面，查找、删除模型； 4）重命名任意模型并另存为
预期结果： 1）用户数据可正确查找、存储等； 2）模型可正确查找、存储、删除

6.1.5 结果通知

测试编号：6.1.5
测试项目：结果通知
测试目的：验证数据挖掘平台支持执行结果通过邮件或短信进行通知
预置条件： 当前数据挖掘平台可用，并且已对接对应通知（电子邮件/短信）服务器
测试步骤： 1）新建实验并执行，实验中添加结果通知算子； 2）等待执行完成并查看是否收到对应通知信息（电子邮件/短信）
预期结果： 1）从本地 CSV 文件读取数据成功，数据集可用于模型训练与测试； 2）结果展示正确，能够展示用户数据

6.2 核心能力测试方法

6.2.1 数据预处理算法

测试编号：6.2.1
测试项目：数据预处理算法
测试目的：验证数据挖掘平台支持基本的数据预处理算法：如空值插补、去重、字段过滤、行列转换、排序等，至少支持 3 种算法
预置条件： 1）数据挖掘平台测试环境； 2）含有噪声数据的数据集
测试步骤： 1）进行模型配置； 2）选择含有噪声的数据集； 3）选择数据预处理算法：空值插补、去重、字段过滤、行列转换等； 4）分别运行算法； 5）记录数据的预处理方法和结果
预期结果： 输出的数据已按照数据预处理算法进行处理

6.2.2 统计分析

测试编号：6.2.2
测试项目：统计分析
测试目的：验证数据挖掘平台支持集成数据统计分析类算法，包括但不限于协方差矩阵、方差、标准差、卡方独立性检验、皮尔森相关系数等
预置条件： 1）数据挖掘平台测试环境正常运行； 2）具有满足统计分析类型的初始数据
测试步骤： 1）平台算法中能够展示统计分析类算法； 2）选择链接对应的统计分析类初始数据； 3）页面使用统计分析类算法； 4）配置统计分析类算法使用规则； 5）点击运行统计分析类算法
预期结果： 1）统计分析类算法运行成功； 2）查看数据输出结果，是否与预期统计分析类算法输出结果一致

6.2.3 支持基本的机器学习算法

测试编号：6.2.3
测试项目：支持基本的机器学习算法
测试目的：验证数据挖掘平台支持基本的机器学习算法，包括但不限于文本分析、分类、聚类、回归、推荐、关联分析等。如 kmeans、朴素贝叶斯、随机森林、线性回归、逻辑回归、决策树、SVM、KNN（K 最近邻算法）等至少 5 种
预置条件： 1）数据挖掘平台测试环境； 2）训练数据集和验证数据集
测试步骤： 1）进行模型配置； 2）选择训练数据集和验证数据集； 3）选择机器学习算法； 4）运行算法； 5）记录数据的算法模型和结果
预期结果： 能够成功运行至少 5 种机器学习算法

6.2.4 深度学习算法

测试编号：6.2.4
测试项目：深度学习算法
测试目的：验证数据挖掘平台支持集成深度学习能力，例如深度信念网络、CNN、RNN、DNN、LSTM、GAN 中至少一种以上
预置条件： 具有满足机器深度类型的初始数据
测试步骤： 1）平台算法中能够展示深度学习类算法； 2）选择链接对应的深度学习类初始数据； 3）页面使用深度学习类算法； 4）配置深度学习类算法使用规则； 5）点击运行深度学习类算法； 6）查看深度学习计算平台上是否运行该深度学习算法
预期结果： 1）深度学习类算法运行成功； 2）查看数据输出结果，是否与预期深度学习类算法输出结果一致

6.2.5 自定义算法

测试编号：6.2.5
测试项目：自定义算法
测试目的：验证数据挖掘平台支持用户自定义的算法，直接编写或接口形式
预置条件： 数据挖掘平台测试环境
测试步骤： 1）进入自定义算法模块； 2）直接编写或上传自定义的算法模型； 3）保存自定义算法； 4）选择数据集，运行自定义算法； 5）记录数据的模型结果
预期结果： 能够成功运行自定义编写的算法

6.2.6 特征工程算法

测试编号：6.2.6
测试项目：特征工程算法
测试目的：验证数据挖掘平台支持集成特征工程类算法，包括但不限于特征离散、特征向量切割、随机欠抽样等
预置条件： 1）数据挖掘平台测试环境； 2）具有满足特征工程类型的初始数据
测试步骤： 1）平台算法中能够展示特征工程类算法； 2）选择链接对应的特征工程类初始数据； 3）页面使用特征工程类算法； 4）配置特征工程类算法使用规则； 5）点击运行特征工程类算法
预期结果： 1）特征工程类算法运行成功； 2）查看数据输出结果，是否与预期特征工程类算法输出结果一致

6.2.7 可视化建模

测试编号：6.2.7
测试项目：可视化建模

测试编号：6.2.7
测试目的：验证数据挖掘平台支持可视化建模，能够为用户提供便捷、快速的建模方式
预置条件： 数据挖掘平台测试环境
测试步骤： 1）进入平台建模页面； 2）记录页面的各种建模方式
预期结果： 页面提供可视化建模方式，能够很好的与算法进行融合，提供给建模工程师快速建模的可视化工具

6.2.8 可视化模型调优

测试编号：6.2.8
测试项目：可视化模型调优
测试目的：验证数据挖掘平台支持在可视化建模功能的基础之上，对建模中所用到的关键信息进行修改和配置，从而影响模型输出结果
预置条件： 数据挖掘平台测试环境
测试步骤： 1）进入平台建模页面； 2）选择已建立好的数据模型； 3）查看页面提供出来的参数配置； 4）修改页面参数配置并执行； 5）对比不同参数情况下的执行结果
预期结果： 1）对于模型中所用到的参数，可视化建模支持在线调优； 2）调优参数能够影响模型的数据结果集

6.2.9 评估

测试编号：6.2.9
测试项目：评估
测试目的：验证数据挖掘平台支持算法模型结果评估
预置条件： 1）数据挖掘平台测试环境； 2）已训练模型结果的输出和原始数据真实值

测试编号：6.2.9
测试步骤： 1）模型配置； 2）运行模型； 3）查看评估算法模型结果： 分类模型支持至少 1 种指标：混淆矩阵、分类正确率、覆盖率、命中率等；回归模型至少支持 1 种指标：相关系数、平均绝对误差、相对误差、最大误差等； 4）记录评估指标
预期结果： 能够通过至少 1 种指标对算法模型结果进行评估

6.2.10 算法结果可视化

测试编号：6.2.10
测试项目：算法结果可视化
测试目的：验证数据平台支持对于已有模型结果集，进行多种可视化展示
预置条件： 1）数据挖掘平台测试环境； 2）具有已运行出结果的模型结果集
测试步骤： 1）选择已有模型结果的输出结果集； 2）点击查看结果按钮； 3）查看页面上的结果展示
预期结果： 平台针对不同的结果集，提供多种维度的结果可视化展示

6.2.11 建模案例

测试编号：6.2.11
测试项目：建模案例
测试目的：验证数据挖掘平台支持提供模板、案例形式的快速创建
预置条件： 数据挖掘平台测试环境
测试步骤： 1）通过提供的模板或案例创建模型； 2）配置模型； 3）运行算法； 4）记录运行结果

测试编号：6.2.11
预期结果： 1）从本地 CSV 文件读取数据成功，数据集可用于模型训练与测试； 2）结果展示正确，能够展示用户数据

6.2.12 算法详细信息

测试编号：6.2.12
测试项目：算法详细信息
测试目的：验证数据平台支持提供算法的教程式解释信息
预置条件： 数据挖掘平台测试环境
测试步骤： 1）选择所需算法； 2）查看算法解释信息； 3）根据算法解释信息，创建模型； 4）运行算法并记录结果
预期结果： 根据算法解释信息成功创建模型并运行成功

6.2.13 执行日志

测试编号：6.2.13
测试项目：执行日志
测试目的：验证数据平台支持执行日志记录、查询及展示
预置条件： 数据挖掘平台测试环境
测试步骤： 1）任选模型运行； 2）查看执行日志； 3）进行日志查看视图，执行条件查询； 4）记录查询结果
预期结果： 能够对日志记录的内容进行查看、查询，日志内容正确无误

6.2.14 二次开发语言兼容性

测试编号：6.2.14
测试项目：二次开发语言兼容性
测试目的：验证数据挖掘平台支持多语言编写及调用，如 python、java、R 等至少 2 种
预置条件： 数据挖掘平台测试环境
测试步骤： 1）进入自定义算法页面； 2）选择一种编程语言； 3）页面编写算法或导入某该语言编写的算法； 4）选择数据集并运行上述 3 编写的算法； 5）记录运行结果； 6）另选一种编程语言，重复上述步骤
预期结果： 支持多种编程语言编写及调用

6.2.15 可视化算子编排调度

测试编号：6.2.15
测试项目：可视化算子编排调度
测试目的：验证数据平台支持数据挖掘算法的可视化编排
预置条件： 数据挖掘平台测试环境
测试步骤： 1）进入模型配置页面； 2）选择多种算子进行编排，例如顺序执行关系、并行执行关系、判断执行关系等； 3）保存并运行模型； 4）记录运行结果
预期结果： 可视化编排各算法并成功执行

6.2.16 调度流程编排

测试编号：6.2.16
测试项目：调度流程编排
测试目的：验证数据平台具备周期性工作流功能

测试编号：6.2.16
预置条件： 1）数据挖掘平台测试环境； 2）具有配置完整的若干个模型
测试步骤： 1）进入周期调度页面； 2）点击新建，创建周期性调度任务； 3）配置周期性调度规则； 4）添加周期性调度的子任务模型； 5）点击开始运行周期任务
预期结果： 1）能够根据规则，创建周期性调度任务； 2）任务能够正常运行，运行结果能够查看； 3）任务能够在周期完成后自动结束

6.3 运维能力测试方法

6.3.1 故障告警

测试编号：6.3.1
测试项目：故障告警
测试目的：验证数据挖掘平台支持对故障的远程通知能力
预置条件： 1）数据挖掘平台测试环境； 2）测试环境对接相应的邮件服务器（或告警箱等）
测试步骤： 1）执行错误模型或流程，系统异常触发告警（如进程异常等）； 2）查看是否收到告警邮件（或短信）
预期结果： 1）系统能够正确告警； 2）能够正常进行告警通知

6.3.2 备份恢复能力

测试编号：6.3.2
测试项目：备份恢复能力
测试目的：验证数据挖掘平台支持结果以及模型的备份恢复能力

测试编号：6.3.2
预置条件： 数据挖掘平台测试环境
测试步骤： 1）系统安全备份功能已启用； 2）系统启动后添加实验并执行； 3）执行系统备份； 4）删除原有数据，使用备份的数据恢复； 5）恢复完成后查看恢复结果
预期结果： 1）数据恢复成功； 2）可以正常查看备份之前的实验以及执行结果； 3）继续创建实验并可正常执行

6.3.3 平台操作日志记录

测试编号：6.3.3
测试项目：平台操作日志记录
测试目的：验证数据挖掘平台支持平台操作的日志记录以及管理
预置条件： 数据挖掘平台测试环境
测试步骤： 1）平台登陆、启动等日志记录； 2）数据、模型的新增、删除等日志记录； 3）系统异常信息等的记录
预期结果： 各日志可查看

6.3.4 训练过程的高可用性

测试编号：6.3.4
测试项目：训练过程的高可用性
测试目的：验证数据挖掘平台如果训练过程中途中断，系统能够自动重启或者继续训练
预置条件： 数据挖掘平台测试环境

测试编号：6.3.4
测试步骤： 1）新建实验并执行； 2）中途停止实验执行； 3）点击继续执行
预期结果： 1）实验执行成功，点击停止后停止成功； 2）点击从此处开始执行或执行到此处时，实验可继续执行； 3）实验最终结果展示正确

6.4 并行化及兼容能力测试方法

6.4.1 支持并行化算法

测试编号：6.4.1
测试项目：支持并行化算法
测试目的：验证数据挖掘平台支持算法的分布式从而方便后续扩展
预置条件： 数据挖掘平台测试环境
测试步骤： 1）新建实验，试验中包含多个算法，并且算法属于不同分支，可并行执行； 2）对不同算法分配执行资源； 3）执行实验进行训练； 4）查看实验算法的执行情况； 5）查看结果
预期结果： 1）实验执行成功，不同分支算法之间相互独立不受影响； 2）每个算法提交的任务以分布式方式提交到集群，在多个节点上并行执行； 3）实验执行成功，分配的分布式节点不影响算法结果

6.4.2 支持并行化任务

测试编号：6.4.2
测试项目：支持并行化任务
测试目的：验证数据挖掘平台支持任务的分布式化从而方便后续扩展
预置条件： 数据挖掘平台测试环境

测试编号：6.4.2
测试步骤： 1）新建三个实验并执行； 2）查看任务在集群中执行节点
预期结果： 1）三个实验同时执行成功； 2）每个任务以分布式方式提交到集群，在多个节点上并行执行； 3）实验以分布式集群的方式提交成功； 4）由 Spark 控制节点分配到不同的节点执行

6.4.3 GPU 加速能力

测试编号：6.4.3
测试项目：GPU 加速能力
测试目的：验证数据挖掘平台能够应用到 GPU 进行计算加速
预置条件： 基于 GPU 的数据挖掘平台测试环境 注：GPU 型号建议使用 2 块 TITAN X (Pascal)（12G 显存/3584 CUDA Cores）
测试步骤： 1）配置基于 CPU 的 K-Means 聚类实验； 2）配置基于 GPU 的 K-Means 聚类实验； 3）使用相同数据集进行训练； 4）使用训练出的模型进行预测
预期结果： 1）基于 CPU 与 GPU 计算任务执行成功； 2）训练出的模型通过数据挖掘平台训练成功； 3）基于 GPU 的计算实验效率比如基于 CPU 的计算效率显著提升

6.4.4 异构计算资源兼容

测试编号：6.4.4
测试项目：异构计算资源兼容
测试目的：验证数据挖掘平台支持多种类型异构计算资源兼容，例如 CPU、GPU、FPGA 等
预置条件： 配置基于 GPU 的数据挖掘平台测试环境

测试编号：6.4.4
测试步骤： 1）配置基于 CPU 的 K-Means 聚类实验； 2）配置基于 GPU 的 K-Means 聚类实验； 3）任务执行成功
预期结果： 1）K-Means 聚类任务执行成功； 2）K-Means 聚类任务可基于 CPU 环境执行； 3）K-Means 聚类任务可基于 GPU 环境执行； 4）聚类结果正确

6.5 性能测试方法

6.5.1 算法执行速率

测试编号：6.5.1
测试项目：算法执行速率
测试目的：验证数据挖掘平台机器学习算法的执行速率
预置条件： 1）数据挖掘平台测试环境； 2）1GB 的特征值数据集
测试步骤： 1）进入模型配置页面； 2）选择朴素贝叶斯算法到模型训练页面； 3）选择 1GB 的特征值数据集； 4）运行算法； 5）记录算法的执行时间； 6）查看算法执行期间的硬件环境情况
预期结果： 算法能够快速，正常的运行

6.5.2 算法准确率

测试编号：6.5.2
测试项目：算法准确率
测试目的：验证数据挖掘平台机器学习算法的准确率

测试编号：6.5.2
预置条件： 1）数据挖掘平台测试环境； 2）1GB 的特征值数据集
测试步骤： 1）进入模型配置页面； 2）选择朴素贝叶斯算法到模型训练页面； 3）选择 1GB 的特征值数据集； 4）运行算法； 5）查看算法运行结果的正确性
预期结果： 朴素贝叶斯算法能够精准计算出结果

6.5.3 算法鲁棒性

测试编号：6.5.3
测试项目：算法鲁棒性
测试目的：验证数据挖掘平台机器学习算法多次执行中运行速率和准确率的稳定性
预置条件： 1）数据挖掘平台测试环境； 2）1GB 的特征值数据集
测试步骤： 1）进入模型配置页面； 2）选择朴素贝叶斯算法到模型训练页面； 3）选择 1GB 的特征值数据集； 4）反复运行共计 5 次模型，分别记录每一次的运行速率、结果集
预期结果： 1）平台的每一次的计算结果均一致； 2）每一次的运行速率呈现均态分布，不会出现大幅波动