4Paradigm 第四范式

第四范式先知

产品说明书

PRODUCT INTRODUCTION

产品简介

第四范式“先知”是一个大规模分布式机器学习的全流程平台，平台提供数据处理、特征工程、模型构建、模型发布应用和模型自学习等功能，帮助用户完成从数据到价值提升的全过程。

PRODUCT ADVANTAGES

产品优势

第四范式先知平台产品优势主要体现在以下三方面：

一、 使用便捷，易理解、易操作

平台提供包括交互式图形界面在内的多种交互模式，并包含了多种模型自动优化技术，降低用户使用机器学习的难度的同时，满足各类用户的使用习惯。平台内置多类业务建模模板，供用户参考与学习，快速构建出适合自己业务的模型。

二、 性能强劲的算法引擎

自主研发的高性能机器学习计算引擎 GDBT，算法运算时间是 spark 数百倍以上。引擎可支持万亿样本量、万亿级特征量数据建模，实现真正的“大数据”建模。

三、 最前沿科技保证最佳模型效果

先知平台内封装了数十项最新科技与专利，蕴含世界顶尖科学家数十年的人工智能研究与应用经验。数据免清洗、专利算法优化、特征优化、自动参数调优等技术，帮助您用低于以往数十倍甚至数百倍时间就能获得较好的模型效果。

PRODUCT ARCHITECTURE

产品架构

先知产品功能架构如图 1 所示。从功能分层的角度，先知由以下主要的功能域组成：

图 1：第四范式先知产品功能架构图

- **人机接口域**

平台提供图形界面交互和命令行交互两种交互模式。

图形界面交互模式下，用户可以通过点选拖拽完成建模过程，并通过任务流程图了解模型生成的过程和相关配置。（参见图 2）

命令行交互模式适合有一定编程基础的用户。该模式下，用户通过统一的交互接口定义数据处理和机器学习建模的各个步骤。命令行模式更适合有深度配置和复杂实验需求的深度用户。

- **任务逻辑域**

机器学习建模是一个复杂的数据计算过程，一般包含多个数据处理与建模任务，这些任务根据业务需求以特定顺序构成了一个个工作流。任务逻辑域的各个功能模块主要完成工作流的分析管理，以及任务的科学调度等工作，确保建模过程以高效的方式完成。

其中部分功能将在后面进行详细介绍。

- **平台功能域**

平台功能域的各功能模块覆盖先知平台的核心功能，即围绕机器学习的数据接入，数据处理，特征处理，模型训练，模型应用，以及对应的扩展能力和系统级的优化能力等。第四范式将大量独有的行业经验沉淀和专利技术融入至平台功能域的各功能中，降低使用门槛的同时，显著提升模型效果。

平台功能域各功能将在后面进行详细介绍。

- **基础设施域：**

用以支持平台业务运作的底层基础设施，包括了灾备管理，基础设施和服务的监控，计算负载和服务负载均衡，数据元数据管理和存储管理，计算集群管理等功能。

- **系统管理域：**

平台为系统管理角度提供多项管理功能，包括用户权限管理；资源管理；数据管理；任务管理；日志、服务和模块监控；任务活动信息管理，系统配置等功能。帮助管理员进行平台运维，确保平台稳定运转。

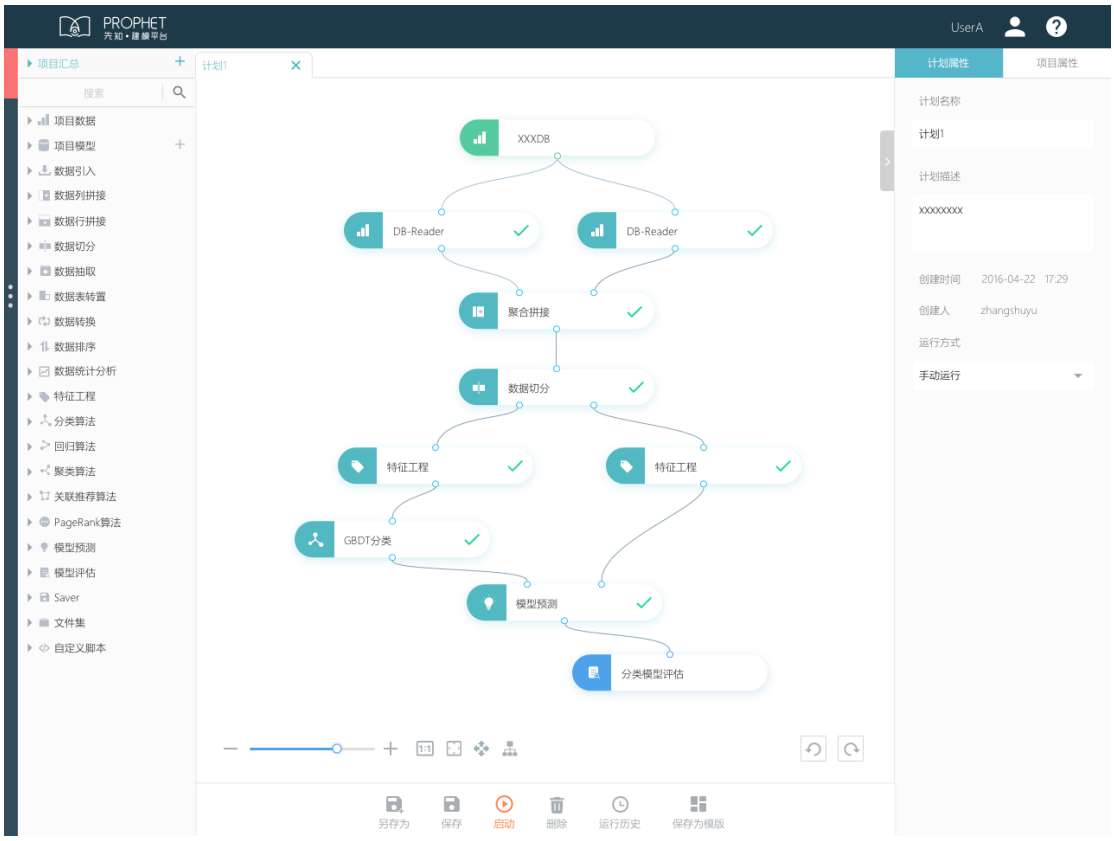

图 2 第四范式先知 图形交互界面

PRODUCT ARCHITECTURE

平台核心功能与特性介绍

内容索引

平台功能域主要功能及其特性介绍：

数据接入：

数据接入模块组主要完成从外部的数据源将数据导入到第四范式先知平台内部的过程，导入过程中平台会根据用户定义进行数据预处理，并将数据转为平台标准格式，供后续计算使用。

第四范式先知平台的数据接入和预处理充分的考虑了大数据场景和大数据系统特点，可对接业界常见数据存储平台与系统，符合主流数据平台兼容性要求和处理效率要求；内部的数据存储机制符合通用平台标准，支持用户进行自定义数据处理与分析。

平台支持的数据接入方式包括：

- 本地数据源接入

对于本地数据源，平台支持基于 Web 的本地文件上传和 FTP 文件传输机制，用户可直接上传数据文件或通过第三方 FTP 工具将大数据文件上传到平台。

- 分布式数据源接入

平台支持 HDFS 分布式存储系统上的数据源直接接入，以及主流公有云平台上数据源的接入。

- 数据库数据源接入

平台支持接入 JDBC 的数据库的数据源，用户可直接将数据库中数据导入至第四范式先知平台。

- 数据预处理

平台可以对接多种常见数据类型和常见数据格式，包括文本文件、CSV、Parquet，也支持处理开启压缩的文件。

平台支持自动识别接入数据的 schema，包括数据的各变量名和变量类型；并支持对异常数据（空值和异常值）的识别和处理，充分提升数据接入的效率。

数据处理：

数据处理与转换是机器学习建模过程中非常重要，将业务逻辑准确表达为数据定义，对数据分析的价值和数据建模的准确性，都有巨大的意义，但数据计算定义也最繁冗耗费时间的工作。为了帮助用户更轻松高效的完成数据处理过程，第四范式先知平台基于数据科学家多年的业界经验，将最常用、最有效的数据处理过程进行了产品封装，提供包括数据格式转换，数据表的简单及复杂拼接、数据表的分割和抽样，以及数据统计分析等功能。用户仅通过简单拖拽和定义一系列的数据处理过程，即可将一个或多个原始的数据表处理为一个包含样本属性（和样本 label）的样本基础表，并可根据实际需求对数据表进行抽样或分割，供后续的机器学习建模和模型预估使用。

数据计算过程全部运行于分布式计算引擎之上，平台对引擎进行了多项数据处理执行优化，包括数据格式优化、数据落盘优化等，数据处理速度可达单机平台的数十倍，大大减少数据科学家在数据处理工作上所耗费时间。

- 数据表转换

对数据表以列为目标定义转化过程，包括时间类型数据的格式转化，对数据中指定内容进行删除或内容替换处理等。

- 数据表分割

平台支持以多种方式对数据表进行按行的分割处理，从而可以将原数据表分割相同 schema 的两个数据表，分别用于模型训练与模型测试评估。

支持的分割方式包括：按比例分割、随机抽样后的按比例分割、分层随机抽样后的按比

例分割、按 SQL 语法的用户自定义逻辑的分割、按用户指定的列排序后的按比例分割。

多种分割方式可帮助用户维持分割后数据集间的分布同质性 ,以及避免建模过程中发生数据穿越（用较新的数据建模用于预测较历史的数据）。

- 数据表抽样

平台支持对数据表按指定的抽样方法和比例进行抽样。抽样方法包括全部随机抽样和分层随机抽样。

- 数据表直接拼接

对一个主数据表和一个（或以上）的从数据表按指定的拼接键值进行直接拼接。解决当主表表项和从表表项是一对一（或主表多 对 从表一）的关系时的数据表拼接问题。

- 数据表聚合拼接

对一个主数据表和一个（或以上）的从数据表按指定的拼接键值进行聚合拼接，解决主表表项和从表表项是一对多（或多对多）的关系下的数据表拼接问题，提供的聚合函数包括了：avg、count、sum、字符串拼接、first、last、max、min。

- 数据表时序拼接

数据表时序（时间序列）拼接是数据表聚合拼接的加强版。在很多业务问题中，样本主体的短期历史信息对识别预测目标有一定的帮助（如用户最近 30 天餐饮类消费总额），这类隐含的数据信息对机器学习建模有效，但是定义复杂，计算代价和复杂度都非常高。为了解决这一问题，第四范式先知开创性提供时序特征拼接方法。用户仅通过简单配置，即可定义和计算出复杂的时序特征用于数据分析与建模。

数据表时序拼接是指，对一个主表和一个（或以上）的从表按指定的拼接键值进行聚合拼接，并在拼接时，为从表中被聚合的数据列指定一个计算时间窗口。平台支持用户自定义主从表的时间游标、计算的时间窗口大小，时间窗口内是否按类别分别聚合，以及聚合函数

（avg、count、sum、字符串拼接、first、last、max、min）。

- 统计分析：

建模过程是从数据中发掘信息的过程，而常用的统计分析能够机器学习建模前帮助用户建立对数据全面和概括的认知，帮助用户更合理的准备数据和处理数据。

平台提供对数据表的基本统计分析功能，包括以列为单位对表项统计：数据的分布、最大最小值、中位数、分位数、分段统计、唯一值&缺失值统计等统计指标，并提供分析结果的可视化。

自定义扩展：

数据科学家都有自己的独门绝技，为了满足数据科学家的个性化数据处理需求和数据建模习惯，让用户将其独门绝技与平台功能结合运用，在享受平台内置先进算法和计算架构的同时充分发挥个人经验，平台支持用户通过 SQL、PySpark 等语言，以及平台 SDK 等，自定义数据计算逻辑和处理过程。

- 自定义 SQL 脚本

SQL query 是数据处理领域通用的处理数据查询数据的方式。第四范式先知在平台内集成了基于SparkSQL的数据处理查询技术，熟悉SQL语言的用户可以用自己熟悉的方式，自定义数据表的处理逻辑。

- 自定义 PySpark 脚本

平台提供内建的符合PySpark框架的脚本定义功能，用户可以通过平台提供的PySpark模板自定义脚本来进行数据处理，用户只需要定义处理逻辑，整体的运行和后续处理由系统自动托管。

- SDK

平台提供两类 SDK，一类为对应所有界面功能操作的 SDK，即命令行的平台接口，用户可通过程序调用的方式完成对平台功能的使用；二类为对平台功能和算法二次开发的 SDK,用户可以基于先知分布式数据计算引擎和高性能机器学习框架平台进行定制开发。

特征处理：

特征处理的过程是指，对包含样本属性信息的基础数据表进行加工、扩展和衍生，形成更易被算法学习出规律的特征变量，并将数据表转化为更适合于机器学习算法框架处理的数据格式。

- 特征定义：

平台提供多种特征定义方法帮助客户进行特征加工和扩展，包括时间类信息的提取方法、数值类的特征处理方法（上取整、取 LOG、线性变换等）、文本类的特征处理方法（文本切词），以及对特征进行组合。

平台会在完成特征定义的计算后，对数据表进行存储转化，生成适合为机器学习算法处理的样本表。

平台的特征抽取在海量数据下也能保证高效的处理效率，具有“稀疏数据无需特征筛选”的优势。

- 特征重要性分析：

特征重要性分析功能可以帮助用户了解各个特征对预测目标值的贡献程度，从而对数据有更全面的认知和把握，同时也可以提前发现异常数据，减少无谓的时间消耗。

平台提供 Slot-Wise LR 和 互信息 两种特征重要性分析功能，用于对样本表中的特征和目标值进行特征重要性分析，衡量每个特征对目标值的影响程度，帮助用户进行特征选择和特征评估。

由于算法原理不同，经常会出现特征分析过程得出的特征重要性结果，与实际模型训练过程中评估的特征重要性不一致的情况。而第四范式先知平台独有的 Slot-wise LR 的计算方式，能够从直接影响模型效果的角度衡量特征重要性，从而避免了上述不一致的问题。

模型生成：

模型生成是指基于训练样本集和测试样本集（带标注的样本数据集）进行模型训练与评估，生成机器学习模型，并明确其预估能力的过程。模型生成共涉及模型训练、模型预估、模型评估等过程。

- 模型训练：

模型训练是指机器学习算法基于训练样本学习出模型的过程。第四范式先知平台支持多种机器学习领域经典和领先的算法，解决分类、回归等常见机器学习问题，并在算法中融入了范式科学家多年学术和业界应用经验，进行了多项专利性算法调优，算法效率与效果相比于主流的并行计算框架都有明显优势。

核心算法包括：

- Logistic Regression
- GBDT
- DNN（全连接 NN 网络，支持四种不同的激活函数和两种损失函数以及四类优化方法）

注：决策树类算法，对离散变量支持不够友好，无法支持包含大量离散变量，或离散变量取值空间较大的训练样本集。而在实际业务中，很多真实属性无法用连续值表达，只能够以离散值来记录，如教育程度、兴趣标签等。对于传统决策树算法，只能在建模过程中丢弃离散特征，或将离散特征人工定制为模型外的规则，这种方式不仅损失大量信息影响预测效

果，也为模型维护带来很大难度。第四范式先知平台对决策树算法进行了独有的专利性优化，算法能有效利用所有离散特征，对数据中的信息更充分挖掘，实现更优的模型效果。

- 模型预估：

基于训练好的模型，对数据样本进行批量的计算，得出样本集上每个实例的目标值预测评分的过程。

- 模型评估：

模型评估是指通过多种专业效果衡量指标，评价模型预测能力的过程。对于训练好的模型，使用一份有实际标注的样本数据进行模型预估，再通过模型评估功能，基于实际标注和模型预估的标注对模型效果进行整体衡量。

不同的业务在衡量模型效果时考虑的因素是不同的，很多平台在模型效果评价时都存在指向单一的问题，先知的模型评估提供了多维度、多度量体系的评估系统，无论用户是什么业务导向，都能在平台中找到适合的指标进行模型评价，更好的进行模型挑选。

衡量指标体系包括：AUC、ROC 图、LIFT 图、Gain 图、K-S 图、各判定阈值下的混淆矩阵以及对应的准确率、精确率、召回率，以及分段的各指标统计等。

模型应用：

对于机器学习模型，一般在实际业务中会有两类应用方式：离线批量预估和在线实时预估。第四范式先知平台通过工作流、API 等方式对上述两类应用方式均提供支持。

- 离线批量预估

用户可通过指定数据或数据源，手动或按时间计划定期执行模型应用功能。平台支持用户对模型预估结果进行配置，如自定义预估结果表的输出表项、结果表的排序方式等。用户可以根据自身业务需求和业务统设计，定期从平台导出结果至其他文件系统或数据库。

- 在线实时预估

平台支持模型实时服务化，即用户可针对训练好的模型发布专有 API，从而可以通过 API 调用的方式，实时发送预估请求，并获取待预估样本的预估结果。

模型发布时，平台自动集成模型应用程序和预估服务容器，并将其自动部署至预估服务器。实时模型服务支持线上模型版本控制和服务管理，预估服务的硬件资源支持弹性伸缩，确保模型预估服务的高响应率和快速响应，满足各类复杂业务的需求。

- 模型全生命周期管理

实际业务应用中的模型需要不断自学习，自我迭代，以确保模型效果稳定（参见下一小节有关模型自学习的介绍）。先知独创性支持模型自学习，并对模型的全生命周期进行管理，支持更新后的模型自动上线，确保业务效果持续稳定，同时维护每次从新数据到新模型的过程与版本信息，确保历史可追溯。真正实现端到端的模型应用效果，解决传统机器学习软件和平台落地应用的难题。

自动优化：

为了帮助刚刚入门机器学习的初级用户也能享受机器学习带来的价值和效率提升，第四范式先知平台在提供各类业务建模模板的同时，还提供了多种模型自动优化功能。用户无需进行复杂的配置操作，即可快速获得有效的机器学习模型。不仅真正降低了机器学习的门槛，对于资深的数据科学家，也可以大大提升建模的工作效率。

- 模型自学习

先知平台区别传统机器学习平台或工具的一大特点是，平台可以基于定期基于最新产生的样本数据，进行模型的自学习，以修正自身错误，及时适应数据中体现出来的业务变化，

更动态的满足业务需求，确保效果稳定。平台基于定时模型生成数据流，和动态数据处理机制实现模型自我更新的效果。自更新过程无需用户参与。

- 一键建模

先知独创的基于自动数据分析和模型推断技术的一键建模功能，提供了更简化的机器学习任务定义方式，显著降低了初接触机器学习的用户在应用人工智能时的门槛，无需太多领域专业知识也能快速搞定有效模型。

- 自动调参：

通过算法配置参数、进行大量模型试验以得到最优模型结果的过程，是一个严重耗时而且很难用经验去判断和优化的过程，特别的，如果是复杂的算法如 DNN，通过人工调参获得最优模型的可操作性明显降低。为解决这一问题，先知平台提供了多种自行探索最优参数配置的策略，在不同算法特点和数据特点的情况下都能有很好的参数搜索效率，解决调参难的问题。用户仅需定义参数取值范围，即可获得系统自动为其定位的最优参数配置。

高性能分布式机器学习计算框架：

平台内置第四范式独家实现自主研发的大规模分布式机器学习计算框架 GDBT，在模型训练过程中可以高效利用计算资源，快速的完成模型训练过程。GDBT 的运算时间随数据量的上升，接近线性增长，效率与效果相比于主流的并行计算框架都有明显优势。在大数据量场景下，计算效率可达 Spark 数百甚至数千倍（见表 1 GDBT 与 spark 在机器学习方面的效率对比）。

	1w样本		5万样本		25万样本		125万样本		625万样本		3125万样本	
	内存占用	时间(秒)	内存占用	时间(秒)	内存占用	时间(秒)	内存占用	时间(秒)	内存占用	时间(秒)	内存占用	时间(秒)
GDBT	32M	30.86	38M	31.11	74M	29.86	170M	40.45	262M	110.59	276M	460.54
SPARK	1000M	31.86	1000M	35.96	1000M	70.38	1600M	399.93	4800M	7927.97	20000M	183640
对比	31倍	1倍	26倍	1倍	14倍	2倍	9倍	10倍	18倍	72倍	72倍	399倍

任务逻辑域主要功能及其特性介绍

任务调度：

从数据处理，到模型训练和预估，每一步操作，都会形成一个计算任务。对任务的科学调度可以大幅提升建模过程的运算效率，和系统资源利用率。

针对不同特性的数据和不同计算负载，平台提供多种任务调度策略和任务执行管理方式，确保底层计算资源的有效利用。

任务调度中包含调度定时器，用户可以自行定义各种任务定时计划，让系统按要求进行自动运行；任务的运行支持多种模式，包括单步运行、指定起止点运行、按时间计划运行等。

工作流：

机器学习建模是一个复杂的数据计算过程，一般包含多个数据处理与建模任务，这些任务根据业务需求以特定顺序构成了一个个工作流。

平台通过 DAG（有向无环图）的方式来定义数据活动工作流，即依据数据任务的输入输出关系来定义任务之间的执行顺序和关联关系，从而形成一个包含所有操作及其执行关系的图。平台以 DAG 为工作任务的组织单位。

- 业务模板：

平台针对常见机器学习问题以及常见机器学习算法，内置多种业务建模模板，并以业务 DAG 模板的方式呈现。这些模板都是根据典型的企业业务问题对应的机器学习过程而沉淀下来的，用户可以根据自己需要解决问题的分类，找到最贴合的模板作为参考，学习建模过

程和操作，或直接基于模板结合自己的数据进行建模，仅需简单配置调整即可进行解决自己业务问题的机器学习过程。

- 工作流分析：

平台内置工作流分析功能。系统会记录下工作流中的各个任务状态和数据状态，分析工作流的可用性，并进行工作流运行优化。对于已经运行过，上下游数据和配置均未发生变化的任务，系统会自动进行识别，避免重复运行的计算开销，提升系统整体运行效率。

- 动态工作流：

部分业务对模型自学习的频率要求可能是小时计甚至分钟级，这不仅是对模型自学习运算效率的高要求，也是对数据的更新机制、以及工作流和数据管理机制的高要求。

第四范式先知平台提供动态工作流管理机制，快速获取最新数据，执行已定义好的工作流完成模型更新，输出最新训练结果，并在整个过程中持续管控任务的运行，以及数据和模型的存储和版本变化。动态工作流是支持模型快速更新的核心基础。

服务调度：

对于人工智能应用系统的开发者，平台还提供计算任务服务化的功能。平台提供多种SDK，供开发者在自己的系统中调用和监控先知平台上的计算过程，实现先知平台功能与业务应用系统的有效整合。

为满足服务的高可用性，平台提供针对服务的资源管理、接口管理，服务发布管理和状态控制等功能。