(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2021 年 10 月 21 日 (21.10.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/208685 A1

(51) 国际专利分类号:
G06F 16/22 (2019.01) *G06N 20/00* (2019.01)
G06F 16/25 (2019.01) *G06K 9/62* (2006.01)

(21) 国际申请号: PCT/CN2021/082518

(22) 国际申请日: 2021 年 3 月 24 日 (24.03.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010307807.5 2020年4月17日 (17.04.2020) CN

(71) 申请人: 第四范式（北京）技术有限公司(THE FOURTH PARADIGM (BEIJING) TECH CO LTD) [CN/CN]; 中国北京市海淀区清河中街66号院1号楼九层LO901-1号, Beijing 100085 (CN)。

(72) 发明人: 吕自荟 (LV, Zihui); 中国北京市海淀区清河中街66号院1号楼九层LO901-1号, Beijing 100085 (CN)。 王昱森 (WANG, Yusen); 中国北京市海淀区清河中街66号院1号楼九层LO901-1号, Beijing 100085 (CN)。 岳凌 (YUE, Ling); 中国北京市海淀区清河中街66号院1号楼九层LO901-1号, Beijing 100085 (CN)。 郭夏玮 (GUO, Xiawei); 中国北京市海淀区清河中街66号院1号楼九层LO901-1号, Beijing 100085 (CN)。 郭朕 (GUO, Zhen); 中国北京市海淀区清河中街66号院1号楼九层LO901-1号, Beijing 100085 (CN)。 张宇 (ZHANG, Yu); 中国北京市海淀区清河中街66号院1号楼九层LO901-1号, Beijing 100085 (CN)。 谷宇晨 (GU, Yuchen); 中国北京市海淀区清河中街66号院1号楼九层LO901-1号, Beijing 100085 (CN)。

(74) 代理人: 北京博雅睿泉专利代理事务所（特殊普通合伙）(BEYOND TALENT PATENT AGENT FIRM); 中国北京市朝阳区永安东里甲3号院1号楼15层1808内F, Beijing 100022 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT,

(54) **Title:** METHOD AND APPARATUS FOR EXECUTING AUTOMATIC MACHINE LEARNING PROCESS, AND DEVICE

(54) **发明名称**: 执行自动机器学习过程的方法、装置及设备

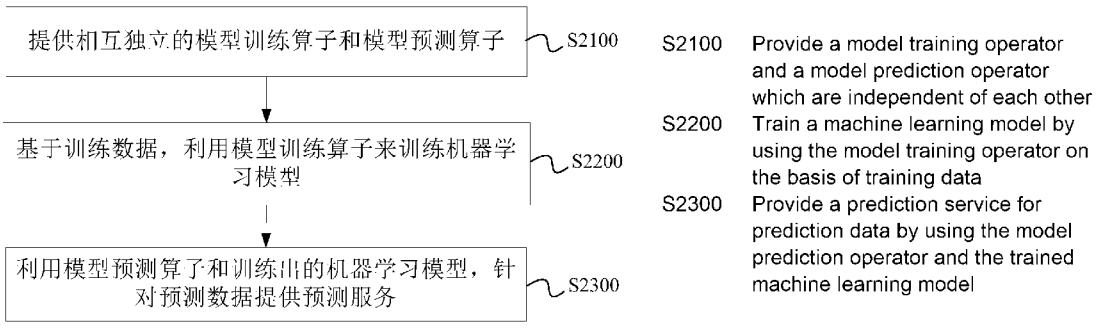


图 2

(57) **Abstract:** A method and apparatus for executing an automatic machine learning process, and a device, the method comprising: providing a model training operator and a model prediction operator which are independent of each other (S2100); training a machine learning model by using the model training operator on the basis of training data (S2200); and providing a prediction service for prediction data by using the model prediction operator and the trained machine learning model (S2300).

(57) **摘要**: 一种执行自动机器学习过程的方法、装置及设备，该方法包括：提供相互独立的模型训练算子和模型预测算子(S2100)；基于训练数据，利用所述模型训练算子来训练机器学习模型(S2200)；以及，利用所述模型预测算子和训练出的所述机器学习模型，针对预测数据提供预测服务(S2300)。

WO 2021/208685 A1

[见续页]

JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

执行自动机器学习过程的方法、装置及设备

本公开要求于 2020 年 04 月 17 日提交中国专利局，申请号为 202010307807.5，申请名称为“执行自动机器学习过程的方法、装置及设备”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

本公开涉及人工智能领域，更具体地，涉及一种执行自动机器学习过程的方法、一种执行自动机器学习过程的装置、一种包括至少一个计算装置和至少一个存储装置的设备、及一种计算机可读存储介质。

背景技术

随着机器学习技术的快速发展和应用，自动机器学习技术大大降低了机器学习的门槛，并且节省了机器学习的人力成本。然而，现有的自动机器学习工具功能过于简单片面，不能覆盖机器学习模型构建及应用的整个流程，即，现有的自动机器学习工具仅能完成如何基于积累的历史数据训练出一个机器学习模型，而无法有效地实现机器学习模型的后续生产和应用（例如，难以利用机器学习模型提供在线服务），换句话说，现有技术存在建模方案或成果与模型应用过程之间割裂严重的问题，而且，其不具备友好的交互方式，使得用户有一定的编程基础才可使用，没有做到真正的低门槛。

发明内容

本公开实施例的一个目的是提供一种执行自动机器学习过程的新的技术方案。

根据本公开的第一方面，提供了一种执行自动机器学习过程的方法，其包括：提供相互独立的模型训练算子和模型预测算子；基于训练数据，利用所述模型训练算子来训练机器学习模型；以及，利用所述模型预测算子和训练出的所述机器学习模型，针对预测数据提供预测服务。

根据本公开的第二方面，还提供一种执行自动机器学习过程的装置，其包括：交互模块，被配置为提供相互独立的模型训练算子和模型预测算子；机器学习模型训练模块，被配置为基于保存的训练数据，利用所述模型训练算子来训练机器学习模型；以及，数据预测模块，被配置为利用所述模型预测算子和训练出的所述机器学习模型，针对收集到的预测数据提供预测服务。

根据本公开的第三方面，还提供一种包括至少一个计算装置和至少一个存储装置的设备，其中，所述至少一个存储装置被配置为存储指令，所述指令被配置为在被所述至少一个计算装置运行时，促使所述至少一个计算装置执行根据以上第一方面所述的方法。

根据本公开的第四方面，还提供一种计算机可读存储介质，其中，其上存储有计算机程序，所述计算机程序在被处理器执行时实现如以上第一方面所述的方法。

根据本公开实施例的方法，其能够提供相互独立的模型训练算子和模型预测算子，并通过模型训练算子完成机器学习模型的训练，及通过模型预测算子提供预测服务，从而使得模型生产和模型应用等多个过程实现全流程循环作业，进而大大降低了机器学习的门槛和成本。

附图说明

被结合在说明书中并构成说明书的一部分的附图示出了本公开的实施例，并且连同其说明一起用于解释本公开的原理。

图 1 是显示可用于实现本公开实施例的电子设备的硬件配置的例子的框图；

图 2 示出了本公开实施例的执行自动机器学习过程的方法的流程图；

图 3 至图 7 示出了根据本公开实施例的执行自动机器学习过程的界面显示示意图；

图 8 示出了本公开实施例的执行自动机器学习过程的装置的原理框图；

图 9 示出了本公开实施例的设备的原理框图。

具体实施方式

现在将参照附图来详细描述本公开的各种示例性实施例。应注意到：除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本公开的范围。

以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本公开及其应用或使用的任何限制。

对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为说明书的一部分。

在这里示出和讨论的所有例子中，任何具体值应被解释为仅仅是示例性的，而不是作为限制。因此，示例性实施例的其它例子可以具有不同的值。

应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

下面，参照附图描述根据本公开实施例的各个实施例和例子。

<硬件配置>

本公开实施例的方法可以由至少一台电子设备实施，即，用于实施该方法的装置 8000 可以布置在该至少一台电子设备上。图 1 示出了任意电子设备的硬件结构。图 1 所示的电子设备可以是便携式电脑、台式计算机、工作站、服务器等，也可以是任意的具有处理器等计算装置和存储器等存储装置的其他设备，在此不做限定。

如图 1 所示，该电子设备 1000 可以包括处理器 1100、存储器 1200、接口装置 1300、通信装置 1400、显示装置 1500、输入装置 1600、扬声器 1700、麦克风 1800 等等。其中，处理器 1100 用于执行计算机程序。该计算机程序可以采用比如 x86、Arm 、RISC、MIPS、SSE 等架构的指令集编写。存储器 1200 例如包括 ROM（只读存储器）、RAM（随机存取存储器）、诸如硬盘的非易失性存储器等。接口装置 1300 例如包括 USB 接口、耳机接口等。通信装置 1400 例如能够进行有线或无线通信，具体地可以包括 Wifi 通信、蓝牙通信、2G/3G/4G/5G 通信等。显示装置 1500 例如是液晶显示屏、触摸显示屏等。输入装置 1600 例如可以包括触摸屏、键盘、体感输入等。电子设备 1000 可以通过扬声器 1700 输出语音信息，及可以通过麦克风 1800 采集语音信息等。

图 1 所示的电子设备仅仅是说明性的并且决不意味着对本公开、其应用或使用的任何限制。应用于本公开的实施例中，电子设备 1000 的所述存储器 1200 用于存储指令，所述指令用于控制所述处理器 1100 进行操作以执行本公开实施例的执行自动机器学习过程的方法。技术人员可以根据本公开所公开方案设计指令。指令如何控制处理器进行操作，这是本领域公知，故在此不再详细描述。

在一个实施例中，提供了一种包括至少一个计算装置和至少一个存储装置的设备，该至少一个存储装置被配置为存储指令，该指令被配置为控制该至少一个计算装置执行根据本公开任意实施例的方法。

该设备可以包括至少一台图 1 所示的电子设备 1000，以提供至少一个例如是处理器的计算装置和至少一个例如是存储器的存储装置，在此不做限定。

<方法实施例>

在本实施例中，提供一种执行自动机器学习过程的方法，该执行自动机器学习过程的方法可以是由电子设备实施，该电子设备可以是如图 1 所示的电子设备 1000。

根据图 2 所示，本实施例的执行自动机器学习过程的方法可以包括如下步骤 S2100～S2300：

步骤 S2100，提供相互独立的模型训练算子和模型预测算子。

模型训练算子是用于对输入的训练数据进行数据预处理，并针对进行数据预处理后的训练数据进行特征工程，及根据特征工程的结果进行模型训练得到机器学习模型的工具。本实施例中，建模人员可以将模型训练算子内容进行提前编辑，然后提供编辑好的模型训练算子的标识信息，例如可以是模型训练算子的名称，如图 3 所示，可以直接为“模型训练算子”，在使用时，用户可以直接将“模型训练算子”进行拖曳。

模型预测算子是用于对输入的预测数据进行数据预处理，并针对进行数据预处理后的预测数据进行特征工程，及利用机器学习模型对特征工程的结果进行预测得到预测结果的工具。本实施例中，建模人员将模型预测算子内容进行提前编辑，然后提供编辑好的模型预测算子的标识信息，例如可以是模型预测算子的名称，如图 3 所示，可以直接为“模型预测算子”，在使用时，用户可以直接将“模型预测算子”进行拖曳。

在本实施例中，可以提供图形用户界面，并在图形用户界面的算子节点中分别提供“模型训练算子”节点和“模型预测算子”节点。如图 3 所示，在需要对训练数据进行模型训练时，可以是将“模型训练算子”拖曳至执行自动机器学习过程的图形用户界面中，得到“模型训练算子”节点。以及，在需要对预测数据进行模型预测时，可以是将“模型预测算子”拖曳至执行自动机器学习过程的图形用户界面中，得到“模型预测算子”节点。

在提供相互独立的模型训练算子和模型预测算子之后，进入：

步骤 S2200，基于训练数据，利用模型训练算子来训练机器学习模型。

在本实施例中，针对不同的机器学习问题提供不同的训练数据。例如可以是在执行本公开实施例的电子设备的指定位置中预先存储对应多种应用场景的训练数据集，每一训练数据集包括多个训练数据，该训练数据例如可以是带标注的图像数据、带标注的一张或多张文本数据表或者带标注的语音数据等。

本实施例中，可以提供图形用户界面，并在图形用户界面的数据节点中选择对应应用场景的训练数据。如图 3 所示，在需要对训练数据进行模型训练时，将该训练数据拖曳至执行自动机器学习过程的图形用户界面中，得到“训练数据”节点，并将该“训练数据”节点与“模型训练算子”节点连接，右键点击“模型训练算子”，选择运行，进而利用模型训练算子基于训练数据来训练机器学习模型。

需要说明的是，本公开实施例中的机器学习模型可被用于预测图像类别、文本类别、语音情感、欺诈交易、广告点击率等。机器学习模型旨在针对相关场景中的对象或事件有关的问题进行预测。例如，可用于预测图像类别、预测图像中文字、预测文本类别、预测语音情感类别、预测欺诈交易、预测广告点击率、预测商品价格等等，使得预测结果可直接作为决策依据或进一步结合其他规则而成为决策依据。

在一个实施例中，本公开实施例中的机器学习模型可被用于的场景包括但不限于以下场景：

图像处理场景，包括：光学字符识别（OCR）、人脸识别、物体识别和图片分类；更具体地举例来说，OCR 可应用于票据（如发票）识别、手写字识别等，人脸识别可应用安防等领域，物体识别可应用于自动驾驶场景中的交通标志识别，图片分类可应用于电商平台

的“拍照购”、“找同款”等。

语音识别场景，包括可通过语音进行人机交互的产品，如手机的语音助手（如苹果手机的 Siri）、智能音箱等。

自然语言处理场景，包括：审查文本（如合同、法律文书和客服记录等）、垃圾内容识别（如垃圾短信识别）和文本分类（情感、意图和主题等）。

自动控制场景，包括：矿井组调节操作预测、风力发电机组调节操作预测和空调系统调节操作预测；具体的对于矿井组可预测开采率高的一组调节操作，对于风力发电机组可预测发电效率高的一组调节操作，对于空调系统，可以预测满足需求的同时节省能耗的一组调节操作。

智能问答场景，包括：聊天机器人和智能客服。

业务决策场景，包括：金融科技领域、医疗领域和市政领域的场景，其中：金融科技领域包括：营销（如优惠券使用预测、广告点击行为预测、用户画像挖掘等）与获客、反欺诈、反洗钱、承保和信用评分、商品价格预测。

医疗领域包括：疾病筛查和预防、个性化健康管理和辅助诊断。

市政领域包括：社会治理与监管执法、资源环境和设施管理、产业发展和经济分析、公众服务和民生保障、智慧城市（公交、网约车、共享单车等各类城市资源的调配和管理）。

推荐业务场景，包括：新闻、广告、音乐、咨询、视频和金融产品（如理财、保险等）的推荐。

搜索场景，包括：网页搜索、图像搜索、文本搜索、视频搜索等。

异常行为检测场景，包括：国家电网客户用电异常行为检测、网络恶意流量检测、操作日志中的异常行为检测等。

在本实施例中，该步骤 S2200 中基于训练数据，利用模型训练算子来训练机器学习模型可以进一步包括如下步骤 S2210~S2230：

步骤 S2210，响应于针对模型训练算子的触发操作，提供进行关于模型训练的配置的配置界面。

本步骤 S2210 中，例如可以是针对模型训练算子实施点击操作，电子设备响应于该点击操作，提供进行关于模型训练的配置的配置界面。

该配置界面包括以下至少一个配置项：机器学习模型的输入源配置项、机器学习模型的适用问题类型配置项、训练机器学习模型的算法模式配置项、机器学习模型的优化目标配置项、及机器学习模型的预测目标字段的字段名配置项。

以上输入源在一般情况下默认为 1 个输入源，即，代表训练数据为 1 张数据表，此时模型训练算子只有 1 个输入节点，如图 4 所示，通过点击“添加输入源”，模型训练算子的输入节点增加，便可接入多张数据表作为训练数据，在此，该模型训练算子还用于将多张数据表按照拼接主键进行拼接。

以上适用问题类型可以包括二分类问题、回归问题以及多分类问题之中的任意一项。例如可以是如图 4 所示提供用于选择适用问题类型的下拉菜单框，以便获取用户通过该下拉菜单框选择的问题类型作为机器学习模型所适用问题类型。

以上算法模式可以包括快速模式、标准模型和精细模型之中的任意一项，该算法模式默认为标准模式。例如可以如图 4 所示提供用于选择算法模式的下拉菜单框，以便获取用户通过该下拉菜单框选择的算法模式作为机器学习模型所适用算法模式。需要注意的是，快速模式不会生成特征，只是使用极小数据快速的生成机器学习模型，标准模式会利用大部分训练数据进行数据预处理，但是只使用一些在大多数训练数据上效果好的特征规则生成特征，精细模式会使用不溢出内存的全量训练数据并利用所有特征规则进行探索以生成机器学习模型。

以上优化目标包括均方误差（Mean Squared Error，MSE）、平均绝对误差（Mean Absolute

Error ，MAE）、平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）、R2、AUC、KS、召回率 Recall、准确率 Precision、精确率 Accuracy、f1、Logloss 之中的至少一项。其中，针对于不同问题类型，提供不同选项，例如对于回归问题，可选 MSE、MAE、MAPE、R2 等，对于二分类问题，可选 AUC、KS、Recall、Precision、Accuracy、f1、Logloss 等。

以上预测目标字段的字段名用于表示模型预测目标所在字段的字段名。例如可以是如图 4 所示提供用于输入预测目标字段的字段名的输入框，以便用户手动输入模型预测目标所在字段的字段名。

步骤 S2220，根据通过配置界面输入的配置信息，对训练数据进行数据预处理和特征工程处理，获得训练样本。

继续以图 4 所示的配置界面为例，输入源默认为 1 个输入源，问题类型选择二分类问题，优化目标选择 AUC，算法模式选择标准模式，目标值字段由用户手动输入（图 4 中未示出所输入的目标值字段）。

本步骤 S2220 中对训练数据进行数据预处理可以包括以下各项之中的至少一项：

第一项，对训练数据进行数据类型转换。

本项中，由于实际输入的业务数据包括的输入数据类型多种多样，格式不一，在此，例如可以是将不同的数据类型统一转换为被广泛使用的 PandasDataFrame 格式。

第二项，对训练数据进行采样。

本项中，可以是对输入的全量数据进行总体的下采样，仅留下算法预先设置的样本个数，留下的样本个数由算法根据开发环境的环境自动配置。例如是分类任务，可以对样本进行分层采样，而对于其他任务，则采用随机采样。

第三项，将训练数据标注为有标签数据和无标签数据。

本项中，可以是利用有标签数据进行模型训练，并利用有标签数据和无标签数据进行特征生成。

第四项，自动识别训练数据的数据字段类型。

本项中，可以是将训练数据包含的每一属性信息的数据字段类型转换为后续特征工程中需要用到的业务类型，该业务类型是根据数据特征的物理意义划分的，并在数据上事先标注好，该业务类型例如可以是时间类型，离散值类型，连续值类型，数组类型和字典类型。通常，如果用户没有自己定义业务类型，算法会将浮点数类型转换为连续值类型，将非浮点数类型转换为离散值类型等。

第五项，对训练数据进行缺失值填充。

本项中，对于某一列 A，会对应生成一个 A’，在第 i 个样本上，如果 A_i 为空，则其值为 A'_i = 1，否则 A'_i = 0。例如，对于任意一个属性信息 A，会对应映射出一个新的属性信息 A’，该新的属性信息 A’ 中的取值规则为：对于一个属性信息 A 中的具体值，若该值为空，则将该值映射出的新的属性信息 A’ 的具体值为 1。若该值不为空值，则将该值映射出的新的属性信息 A’ 的具体值为 0。

第六项，对训练数据中的初始时间字段进行解析，获得并添加新的时间字段，且删除初始时间字段。

本项中，可以是将不同格式的时间类型列转换成统一的数据格式 Date，解析 Date 列得到年，月，日，周和小时的信息，并将其分别作为新的离散列和新的连续值列补充到原训练数据当中，同时将 Date 列的时间戳当做一列新的连续值特征，并删除原数据中的初始时间类型特征。

第七项，自动识别训练数据中的非数值数据，并对非数值数据进行哈希处理。

本项中，可以是判断是否有数据存储类型不属于整数，浮点数的列，如果有，将其用哈希算法映射成整数字符串，模型可以利用新生成的整数字符串学习原始数据列中的信息。

在本实施例中，该步骤 S2220 中对训练数据进行数据预处理和特征工程处理，获得训

练样本可以进一步包括如下步骤 S2221~S2224：

步骤 S2221，对进行预处理后的训练数据进行采样。

本步骤 S2221 中，可以是对进行预处理后的训练数据进行下采样，例如进行随机采样，可降低训练数据的数据量，以便提升后续特征重要性值的计算速度。

步骤 S2222，对进行采样后的训练数据进行前特征选择，获得基础特征。

通过本步骤 S2222 中对进行采样后的训练数据进行前特征选择，可以筛选出特征重要性值较高的特征作为基础特征。

在本实施例中，该步骤 S2222 中对进行采样后的训练数据进行前特征选择，获得基础特征可以进一步包括如下步骤 S2222-1~S2222-3：

步骤 S2222-1，提取进行采样后的训练数据中包括的所有属性信息。

该属性信息用于形成特征。例如，训练数据可包括用户期望向消费者推荐的资讯、消费者的基本属性信息（例如，资讯主题、资讯展示位置、消费者标识符、性别、年龄、身高、体重、兴趣爱好等）中的至少一种。

步骤 S2222-2，获取每一属性信息的特征重要性值。

本步骤 S2222-2 中，特征重要性值例如可以是 hellinger 距离，随机森林特征切分增益以及梯度提升决策树特征切分增益等之中的任意一项。例如对于分类任务可以计算 hellinger 距离作为每一属性信息的特征重要性值。又例如对于回归任务可以计算随机森林特征切分增益作为每一属性信息的特征重要性值。

本步骤 S2222-2 中，例如可以是计算以上资讯主题、资讯展示位置、消费者标识符、性别、年龄、身高、体重、兴趣爱好等的特征重要性值。

步骤 S2222-3，根据特征重要性值，获得基础特征。

本实施例中，该步骤 S2222-3 中根据特征重要性值，获得基础特征可以进一步包括如下步骤 S2222-31~S2222-32：

步骤 S2222-31，对所有特征重要性值进行排序，获得排序结果。

本步骤 S2222-31 中，例如可以是对以上资讯主题、资讯展示位置、消费者标识符、性别、年龄、身高、体重、兴趣爱好等的特征重要性值按照从大到小的顺序进行排序，进而获得排序结果。

步骤 S2222-32，根据排序结果，获取第一预定数量的特征重要性值所对应的属性信息作为基础特征。

第一预定数量可以是根据具体应用场景或者是仿真试验设置的数值。例如对于不同应用场景可以是设置对应该应用场景的数值，对应不同应用场景的数值可以相同，也可以不同。又例如也可以是对于所有应用场景均设置相同的数值，本实施例并不限定第一预定数量的具体设置方式。该第一设定数量还可以根据计算资源来进行动态调整。

本步骤 S2222-32 中，例如可以是根据以上从大到小的排序结果，获取前第一预定数量的特征重要性值所对应的属性信息作为基础特征。

步骤 S2223，对基础特征进行特征衍生，获得衍生特征。

本实施例中，该步骤 S2223 中对基础特征进行特征衍生，获得衍生特征可以进一步包括：利用设定的特征生成规则，对基础特征进行统计计算、特征组合中的至少一个，获得衍生特征。

以上设定的特征生成规则例如可以包括 Count, Nunique，NumAdd，NumSubtract，NumMultip， NumDivision， CatNumMean， CatNumStd， CatNumMax， CatNumMin，TumeSubtract，NumOutlier，CatTimeDiff 之中的任意一项或者多项。

本步骤 S2223 中，可以保存计算生成特征所需要参数，便于加速模型预测算子的特征生成过程。

步骤 S2224a，根据基础特征和衍生特征，生成训练样本。

本实施例中，还可以是在根据以上步骤 S2223 获得衍生特征之后，进一步包括如下步骤 S2224b-1~S2224b-2：

步骤 S2224b-1，对基础特征和衍生特征进行后特征选择。

本实施例中，该步骤 S2224b-1 中对基础特征和衍生特征进行后特征选择可以进一步包括如下步骤 S2224b-11~S2224b-13：

步骤 S2224b-11，获取每一基础特征和每一衍生特征的特征重要性值。

本步骤 S2224b-11 中，特征重要性值例如可以是 hellinger 距离，随机森林特征切分增益以及梯度提升决策树特征切分增益等之中的任意一项。例如对于回归任务可以计算随机森林特征切分增益作为每一特征的特征重要性值。

步骤 S2224b-12，对所有特征重要性值进行排序，获得排序结果。

本步骤 S2224b-12 中，例如可以是对获取到的每一基础特征和每一衍生特征的特征重要性值按照从大到小的顺序进行排序，进而获得排序结果。

步骤 S2224b-13，根据排序结果，获取第二预定数量的特征作为生成训练样本所需要的特征。

第二预定数量可以是根据具体应用场景或者是仿真试验设置的数值。例如对于不同应用场景可以是设置对应该应用场景的数值，对应不同应用场景的数值可以相同，也可以不同。又例如也可以是对于所有应用场景均设置相同的数值，本实施例并不限定第二预定数量的具体设置方式。

本步骤 S2224b-13 中，例如可以是根据以上从大到小的排序结果，获取前第二预定数量的特征重要性值所对应的特征作为生成训练样本所需要的特征。

本实施例中，还可以是设置一个阈值参数 r，根据得到的特征重要性值组成特征重要性值集合，并获取该集合的中位数 m，在该集合中，如果一个特征重要性值大于 r*m，则将该特征重要性值对应的特征保留。

步骤 S2224b-2，根据进行后特征选择得到的特征，生成训练样本。

步骤 S2230，利用至少一种模型训练算法，基于训练样本来训练出机器学习模型。

模型训练算法例如可以是梯度提升决策树，随机森林，因子分解机，域敏感的因子分解机，线性回归之中的至少一项。

在实际运行时，本实施例还支持提前停止（early stop）策略，具体来说，当多种算法同时进行训练时，可以根据某种策略来提前判定训练数据更适合哪种算法，从而暂停不适合的算法的探索，把时间和资源都花费在更适合的算法上。

在基于训练数据，利用模型训练算子来训练机器学习模型之后，进入：

步骤 S2300，利用模型预测算子和训练出的机器学习模型，针对预测数据提供预测服务。

预测数据可以是用户（例如，用于推荐资讯的资讯服务商）期望获得相关预测结果的数据。例如，当用户想要获知期望向其客户（例如，终端消费者）推荐的资讯是否会被接受（即，是否会被消费者点击或阅读）的预测结果时，该预测数据即为关于期望推荐的资讯的属性信息数据。

如图 3 所示，在需要针对预测数据提供预测服务时，可以是将“模型预测算子”拖曳至执行自动机器学习过程的图形用户界面中，得到“模型预测算子”节点，以及，将预测数据拖曳至执行自动机器学习过程的图形用户界面中，得到“预测数据”节点，并将该“预测数据”节点与“模型预测算子”节点连接，并右键点击“模型预测”，选择运行，进而利用模型预测算子基于预测数据提供批量预测服务。

在本实施例中，该步骤 S2300 中利用模型预测算子和训练出的机器学习模型，针对预测数据提供预测服务可以进一步包括如下步骤 S2310~S2330：

步骤 S2310，响应于针对模型预测算子的触发操作，提供进行关于批量预测服务的配

置的配置界面。

本步骤 S2310 中，例如可以是针对模型预测算子实施点击操作，电子设备响应于该点击操作，提供进行关于批量预测服务的配置的配置界面。

如图 5 所示的配置界面，该配置界面包括预测结果中字段选择的配置项、模拟实时预测服务的开关状态的配置项中的至少一项。

以上预测结果中字段选择可以包括只保留样本目标值和预测分值，及自定义之中的至少一项。例如，如果选择“只保留样本目标值和预测分值”，则在预测结果中展示样本的目标值和模型预测结果。又例如，如果选择“自定义”，则提供一个文本框供用户输入字段名称，同时提供所有可供选择字段的名称供用户勾选，勾选字段将自动出现在文本框中，如图 6 所示。

以上模拟实时预测服务的开关状态为关闭状态时为批量预估模式，预测数据整体参与预估，但是各样本的预测结果可能会相互影响。模拟实时预测服务的开关状态为开启状态时预测样本之间不会互相影响，预测结果与实时预估完全一致。

步骤 S2320，根据通过配置界面输入的配置信息，对预测数据进行数据预处理和特征更新处理，获得预测样本。

继续以图 5 所示的配置界面为例，预测结果中字段选择配置为“只保留样本目标值和预测分值”，并且，模拟实时预测服务的开关状态为关闭状态。

本步骤 S2320 中对预测数据进行数据预处理可以包括以下各项之中的至少一项：

第一项，对预测数据进行数据类型转换。

第二项，对预测数据进行分块。

本项中，可以先判断预测数据的数据量，当数据量不足以在内存中进行一次处理时，可以将预测数据分成多份存储至硬盘中，以便分批次进行之后的数据预处理，特征工程和结果预估。

第三项，将预测数据中的属性信息与对数据预处理后的训练数据中的属性信息对齐。

本项中，当预测数据读入时，会将预测数据每一列的类型和训练数据对应列的类型进行对齐。

第四项，自动识别预测数据的数据字段类型。

第五项，对预测数据进行缺失值填充。

第六项，对预测数据中的初始时间字段进行解析，获得并添加新的时间字段，且删除初始时间字段。

在本实施例中，该步骤 S2320 中对预测数据进行数据预处理和特征更新处理，获得预测样本可以进一步包括如下步骤 S2321~S2324：

步骤 S2321，从特征工程的结果中筛选出特征集合。

特征集合中包括基础特征和衍生特征。

步骤 S2322，识别衍生特征对应的特征生成规则。

步骤 S2323，将对齐后的预测数据中的属性信息中不属于基础特征的属性信息删除，以得到预测数据的基础特征。

步骤 S2324，将删除后的预测数据中的属性信息，按照特征生成规则，生成预测数据的衍生特征。

步骤 S2325，根据预测数据的基础特征和预测数据的衍生特征，生成预测样本。

步骤 S2330，利用训练出的机器学习模型，针对预测样本提供预测结果。

在本实施例中，可以是利用训练出的多个机器学习模型，分别针对预测样本提供预测结果，并将多个机器学习模型的预测结果的平均值，作为预测数据对应的最终预测结果。

根据本实施例的方法，其能够提供相互独立的模型训练算子和模型预测算子，并通过模型训练算子完成机器学习模型的训练，及通过模型预测算子提供预测服务，从而使得模

型生产和模型应用等多个过程实现全流程循环作业，进而大大降低了机器学习的门槛和成本。

而且，其能够适应于不同的结构化数据场景，例如可以是支持二分类场景、多分类场景、回归类场景以及聚类问题场景。

在一个实施例中，可以提供实时预测服务的配置界面，以便针对预测数据提供实时预测服务。该实施例中，该执行自动机器学习过程的方法还可以包括如下步骤 S7100~S7300：

步骤 S7100，根据配置实时预测服务的操作，提供进行关于实时预测服务的配置的配置界面。

配置界面包括用于从训练出的机器学习模型之中选择线上机器学习模型的模型选取规则的配置项、应用资源的配置项中的至少一项。

步骤 S7200，通过配置界面中设定的 API 地址接收包括预测数据的预测服务请求。

步骤 S7300，响应于接收到的预测服务请求，利用选择出的机器学习模型得到针对预测数据的预测结果，并将预测结果通过 API 地址发出。

在一个实施例中，可以提供人机交互接口，以支持建模人员进行操作，进而获得模型训练算子。该实施例中，该用于执行自动机器学习过程的方法还包括如下步骤 S8100~S8300：

步骤 S8100，根据编辑模型训练算子的操作，提供编辑界面。

本实施例中，电子设备可以响应于编辑模型训练算子的操作，提供编辑界面。

该编辑界面中可以包括编辑入口，该编辑入口可以是输入框、下拉列表、语音输入等。

步骤 S8200，获取通过编辑界面输入的算子内容。

算子内容包括对输入的训练数据进行数据预处理的操作命令、针对数据预处理后的训练数据进行特征工程的操作命令、及根据特征工程的结果进行模型训练的操作命令。

本实施例中，建模人员可以通过编辑界面提供的编辑入口输入算子内容，供电子设备获取该算子内容。

步骤 S8300，封装算子内容，得到模型训练算子。

本实施例中，可以是封装对输入的训练数据进行数据预处理的操作命令、针对数据预处理后的训练数据进行特征工程的操作命令、及根据特征工程的结果进行模型训练的操作命令后，获得模型训练算子。

在一个实施例中，还提供可视化界面将模型训练过程中得到的模型训练方案进行可视化展示。该实施例中，该用于执行自动机器学习过程的方法还包括以下步骤 S9100~S9200：

步骤 S9100，基于训练出的机器学习模型，获得模型训练方案。

该模型训练方案包括用于训练机器学习模型的算法、机器学习模型的超参数、机器学习模型的效果以及特征信息之中的任意一项或者多项。

该算法包括但不限于是以上梯度提升决策树，随机森林，因子分解机，域敏感的因子分解机，线性回归之中的任意一项。

该超参数可以包括模型超参数和训练超参数。

以上模型超参数是用于定义模型的超参数，例如但不限于是激活函数（例如恒等函数、S 型函数以及截断斜波函数等）、隐含层节点数量、卷积层通道数量以及全连接层节点数量等。

以上训练超参数是用于定义模型训练过程的超参数，例如但不限于是学习率、批尺寸以及迭代次数等。

该特征信息包括特征数量、特征生成方式以及特征重要性分析结果之中的任意一项或者多项。

步骤 S9200，将模型训练方案进行可视化展示。

本步骤 S9200 中，该可视化展示结果可以是如图 7 所示的图形用户界面。

在本实施例中，该用于执行自动机器学习过程的方法还可以包括根据可视化展示的预览结果重新训练机器学习模型的步骤。

本实施例中，如果预览结果未达要求，可返回至模型训练步骤，通过修改关于模型训练的配置的配置界面中的配置信息，重新进行模型训练。

<装置实施例>

在本实施例中，提供一种执行自动机器学习过程的装置 8000，如图 8 所示，包括交互模块 8100、机器学习模型训练模块 8200 及数据预测模块 8300。

该交互模块 8100，被配置为提供相互独立的模型训练算子和模型预测算子。

该机器学习模型训练模块 8200，被配置为基于保存的训练数据，利用所述模型训练算子来训练机器学习模型。

该数据预测模块 8300，被配置为利用所述模型预测算子和训练出的所述机器学习模型，针对收集到的预测数据提供预测服务。

在一个实施例中，该装置 8000 还包括模型训练算子获取模块（图中未示出），该模型训练算子获取模块被配置为：

根据编辑所述模型训练算子的操作，提供编辑界面。

获取通过所述编辑界面输入的算子内容，其中，所述算子内容包括对输入的训练数据进行数据预处理的操作命令、针对数据预处理后的训练数据进行特征工程的操作命令、及根据特征工程的结果进行模型训练的操作命令。

封装所述算子内容，得到所述模型训练算子。

在一个实施例中，该机器学习模型训练模块 8200 具体被配置为：

响应于针对所述模型训练算子的触发操作，提供进行关于模型训练的配置的配置界面。

根据通过所述配置界面输入的配置信息，对所述训练数据进行数据预处理和特征工程处理，获得训练样本。

利用至少一种模型训练算法，基于所述训练样本来训练出机器学习模型。

在一个实施例中，所述配置界面包括以下至少一个配置项：机器学习模型的输入源配置项、机器学习模型的适用问题类型配置项、训练机器学习模型的算法模式配置项、机器学习模型的优化目标配置项、及机器学习模型的预测目标字段的字段名配置项。

在一个实施例中，该机器学习模型训练模块 8200 具体被配置为对所述训练数据进行以下各项之中的至少一项：

第一项，对所述训练数据进行数据类型转换.

第二项，对所述训练数据进行采样。

第三项，将所述训练数据标注为有标签数据和无标签数据。

第四项，自动识别所述训练数据的数据字段类型。

第五项，对所述训练数据进行缺失值填充。

第六项，对所述训练数据中的初始时间字段进行解析，获得并添加新的时间字段，且删除所述初始时间字段。

第七项，自动识别所述训练数据中的非数值数据，并对所述非数值数据进行哈希处理。

在一个实施例中，该机器学习模型训练模块 8200 具体被配置为：

对进行所述预处理后的训练数据进行采样.

对进行所述采样后的训练数据进行前特征选择，获得基础特征。

对所述基础特征进行特征衍生，获得衍生特征。

根据所述基础特征和所述衍生特征，生成训练样本。

在一个实施例中，该机器学习模型训练模块 8200 具体被配置为：

提取进行所述采样后的训练数据中包括的所有属性信息，其中，所述属性信息用于形成特征。

获取每一所述属性信息的特征重要性值。

根据所述特征重要性值，获得基础特征。

在一个实施例中，该机器学习模型训练模块 8200 具体被配置为：

对所有所述特征重要性值进行排序，获得排序结果。

根据所述排序结果，获取第一预定数量的属性信息作为基础特征。

在一个实施例中，该机器学习模型训练模块 8200 具体用于：

利用设定的特征生成规则，对所述基础特征进行统计计算、特征组合中的至少一个，获得衍生特征。

在一个实施例中，该机器学习模型训练模块 8200 具体被配置为：

对所述基础特征和所述衍生特征进行后特征选择。

根据进行所述后特征选择得到的特征，生成训练样本。

在一个实施例中，该机器学习模型训练模块 8200 具体被配置为：

获取每一所述基础特征和每一所述衍生特征的特征重要性值。

对所有所述特征重要性值进行排序，获得排序结果。

根据所述排序结果，获取第二预定数量的特征作为生成训练样本所需要的特征。

在一个实施例中，该装置 8000 还包括模型训练方案展示模块（图中未示出），该模型训练方案展示模块被配置为：

基于训练出的所述机器学习模型，获得模型训练方案。

将所述模型训练方案进行可视化展示。

其中，所述模型训练方案包括用于训练机器学习模型的算法、机器学习模型的超参数、机器学习模型的效果以及特征信息之中的任意一项或者多项。

其中，所述特征信息包括特征数量、特征生成方式以及特征重要性分析结果之中的任意一项或者多项。

在一个实施例中，该机器学习模型训练模块 8200 具体被配置为：

根据可视化展示的预览结果重新训练机器学习模型。

在一个实施例中，该数据预测模块 8300 包括批量预测单元（图中未示出），该批量预测单元被配置为：

响应于针对所述模型预测算子的触发操作，提供进行关于批量预测服务的配置的配置界面；

根据通过所述配置界面输入的配置信息，对所述预测数据进行数据预处理和特征更新处理，获得预测样本；

利用训练出的所述机器学习模型，针对所述预测样本提供预测结果。

在一个实施例中，所述配置界面包括预测结果中字段选择的配置项、模拟实时预测服务的开关状态的配置项中的至少一项。

在一个实施例中，该数据预测模块 8300 包括实时预测单元（图中未示出），该实时预测单元被配置为：

根据配置实时预测服务的操作，提供进行关于实时预测服务的配置的配置界面。

通过所述配置界面中设定的 API 地址接收包括所述预测数据的预测服务请求。

响应于接收到的所述预测服务请求，利用选择出的机器学习模型得到针对所述预测数据的预测结果，并将预测结果通过所述 API 地址发出。

在一个实施例中，所述配置界面包括用于从训练出的机器学习模型之中选择线上机器学习模型的模型选取规则的配置项、应用资源的配置项中的至少一项。

<设备实施例>

如图 9 所示，本实施例提供了一种至少一个计算装置 9100 和至少一个存储装置 9200 的设备 9000。该至少一个存储装置 9200 被配置为存储指令，该指令被配置为在被至少一个计算装置运行时，促使至少一个计算装置 9100 执行根据本公开任意实施例的执行自动机器学习过程的方法。

<存储介质实施例>

本实施例提供了一种计算机可读存储介质，其中，其上存储有计算机程序，所述计算机程序在被处理器执行时实现根据上述方法实施例中任一项所述的方法。

本公开可以是设备、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于使处理器实现本公开的各个方面的计算机可读程序指令。

计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子（非穷举的列表）包括：便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、静态随机存取存储器（SRAM）、便携式压缩盘只读存储器（CD-ROM）、数字多功能盘（DVD）、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波（例如，通过光纤电缆的光脉冲）、或者通过电线传输的电信号。

这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

用于执行本公开操作的计算机程序指令可以是汇编指令、指令集架构（ISA）指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言—诸如 Smalltalk、C++等，以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。在一些实施例中，通过利用计算机可读程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列（FPGA）或可编程逻辑阵列（PLA），该电子电路可以执行计算机可读程序指令，从而实现本公开的各个方面。

这里参照根据本公开实施例的方法、装置（系统）和计算机程序产品的流程图和/或框图描述了本公开的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理

装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

附图中的流程图和框图显示了根据本公开的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。对于本领域技术人员来说公知的是，通过硬件方式实现、通过软件方式实现以及通过软件和硬件结合的方式实现都是等价的。

以上已经描述了本公开的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。本公开的范围由所附权利要求来限定。

工业实用性

通过本公开实施例，其能够使得模型生产和模型应用等多个过程实现全流程循环作业，大大降低了机器学习的门槛和成本。

权利要求书

1、一种执行自动机器学习过程的方法，包括：

提供相互独立的模型训练算子和模型预测算子；

基于训练数据，利用所述模型训练算子来训练机器学习模型；以及，

利用所述模型预测算子和训练出的所述机器学习模型，针对预测数据提供预测服务。

2、根据权利要求 1 所述的方法，其中，所述方法还包括获得所述模型训练算子的步骤，包括：

根据编辑所述模型训练算子的操作，提供编辑界面；

获取通过所述编辑界面输入的算子内容，其中，所述算子内容包括对输入的训练数据进行数据预处理的操作命令、针对数据预处理后的训练数据进行特征工程的操作命令、及根据特征工程的结果进行模型训练的操作命令；

封装所述算子内容，得到所述模型训练算子。

3、根据权利要求 1 或 2 所述的方法，其中，所述基于训练数据，利用所述模型训练算子来训练机器学习模型，包括：

响应于针对所述模型训练算子的触发操作，提供进行关于模型训练的配置的配置界面；

根据通过所述配置界面输入的配置信息，对所述训练数据进行数据预处理和特征工程处理，获得训练样本；

利用至少一种模型训练算法，基于所述训练样本来训练出机器学习模型。

4、根据权利要求 3 所述的方法，其中，所述配置界面包括以下至少一个配置项：机器学习模型的输入源配置项、机器学习模型的适用问题类型配置项、训练机器学习模型的算法模式配置项、机器学习模型的优化目标配置项、及机器学习模型的预测目标字段的字段名配置项。

5、根据权利要求 3 所述的方法，其中，所述对所述训练数据进行数据预处理包括以下各项之中的至少一项：

第一项，对所述训练数据进行数据类型转换；

第二项，对所述训练数据进行采样；

第三项，将所述训练数据标注为有标签数据和无标签数据；

第四项，自动识别所述训练数据的数据字段类型；

第五项，对所述训练数据进行缺失值填充；

第六项，对所述训练数据中的初始时间字段进行解析，获得并添加新的时间字段，且删除所述初始时间字段；

第七项，自动识别所述训练数据中的非数值数据，并对所述非数值数据进行哈希处理。

6、根据权利要求 3 所述的方法，其中，所述对所述训练数据进行数据预处理和特征工程处理，获得训练样本，包括：

对进行所述预处理后的训练数据进行采样；

对进行所述采样后的训练数据进行前特征选择，获得基础特征；

对所述基础特征进行特征衍生，获得衍生特征；

根据所述基础特征和所述衍生特征，生成训练样本。

7、根据权利要求 6 所述的方法，其中，所述对进行所述采样后的训练数据进行前特征选择，获得基础特征，包括：

提取进行所述采样后的训练数据中包括的所有属性信息，其中，所述属性信息用于形成特征；

获取每一所述属性信息的特征重要性值；

根据所述特征重要性值，获得基础特征。

8、根据权利要求 7 所述的方法，其中，所述根据所述特征重要性值，获得基础特征，包括：

对所有所述特征重要性值进行排序，获得排序结果；

根据所述排序结果，获取第一预定数量的属性信息作为基础特征。

9、根据权利要求 6 所述的方法，其中，所述对所述基础特征进行特征衍生，获得衍生特征，包括：

利用设定的特征生成规则，对所述基础特征进行统计计算、特征组合中的至少一个，获得衍生特征。

10、根据权利要求 6 所述的方法，其中，所述方法在获得衍生特征后，还包括：

对所述基础特征和所述衍生特征进行后特征选择；

根据进行所述后特征选择得到的特征，生成训练样本。

11、根据权利要求 10 所述的方法，其中，所述对所述基础特征和所述衍生特征进行后特征选择，包括：

获取每一所述基础特征和每一所述衍生特征的特征重要性值；

对所有所述特征重要性值进行排序，获得排序结果；

根据所述排序结果，获取第二预定数量的特征作为生成训练样本所需要的特征。

12、根据权利要求 1 至 11 中任一项所述的方法，其中，所述方法还包括：

基于训练出的所述机器学习模型，获得模型训练方案；

将所述模型训练方案进行可视化展示；

其中，所述模型训练方案包括用于训练机器学习模型的算法、机器学习模型的超参数、机器学习模型的效果以及特征信息之中的任意一项或者多项；

其中，所述特征信息包括特征数量、特征生成方式以及特征重要性分析结果之中的任意一项或者多项。

13、根据权利要求 12 所述的方法，其中，所述方法还包括：

根据可视化展示的预览结果重新训练机器学习模型的步骤。

14、根据权利要求 1 至 13 中任一项所述的方法，其中，所述利用所述模型预测算子和训练出的所述机器学习模型，针对预测数据提供预测服务，包括：

响应于针对所述模型预测算子的触发操作，提供进行关于批量预测服务的配置的配置界面；

根据通过所述配置界面输入的配置信息，对所述预测数据进行数据预处理和特征更新处理，获得预测样本；

利用训练出的所述机器学习模型，针对所述预测样本提供预测结果。

15、根据权利要求 14 所述的方法，其中，所述配置界面包括预测结果中字段选择的配置项、模拟实时预测服务的开关状态的配置项中的至少一项。

16、根据权利要求 14 所述的方法，其中，所述方法还包括：

根据配置实时预测服务的操作，提供进行关于实时预测服务的配置的配置界面；

通过所述配置界面中设定的 API 地址接收包括所述预测数据的预测服务请求；

响应于接收到的所述预测服务请求，利用选择出的机器学习模型得到针对所述预测数据的预测结果，并将预测结果通过所述 API 地址发出。

17、根据权利要求 16 所述的方法，其中，所述配置界面包括用于从训练出的机器学习模型之中选择线上机器学习模型的模型选取规则的配置项、应用资源的配置项中的至少一项。

18、一种执行自动机器学习过程的装置，包括：

交互模块，被配置为提供相互独立的模型训练算子和模型预测算子；

机器学习模型训练模块，被配置为基于保存的训练数据，利用所述模型训练算子来训

练机器学习模型；以及

数据预测模块，被配置为利用所述模型预测算子和训练出的所述机器学习模型，针对收集到的预测数据提供预测服务。

19、根据权利要求 18 所述的装置，其中，所述装置还包括模型训练算子获取模块，所述模型训练算子获取模块被配置为：

根据编辑所述模型训练算子的操作，提供编辑界面；

获取通过所述编辑界面输入的算子内容，其中，所述算子内容包括对输入的训练数据进行数据预处理的操作命令、针对数据预处理后的训练数据进行特征工程的操作命令、及根据特征工程的结果进行模型训练的操作命令；

封装所述算子内容，得到所述模型训练算子。

20、根据权利要求 18 或 19 所述的装置，其中，所述机器学习模型训练模块具体被配置为：

响应于针对所述模型训练算子的触发操作，提供进行关于模型训练的配置的配置界面；

根据通过所述配置界面输入的配置信息，对所述训练数据进行数据预处理和特征工程处理，获得训练样本；

利用至少一种模型训练算法，基于所述训练样本来训练出机器学习模型。

21、根据权利要求 20 所述的装置，其中，所述配置界面包括以下至少一个配置项：机器学习模型的输入源配置项、机器学习模型的适用问题类型配置项、训练机器学习模型的算法模式配置项、机器学习模型的优化目标配置项、及机器学习模型的预测目标字段的字段名配置项。

22、根据权利要求 20 所述的装置，其中，所述机器学习模型训练模块具体被配置为对所述训练数据进行以下各项之中的至少一项：

第一项，对所述训练数据进行数据类型转换；

第二项，对所述训练数据进行采样；

第三项，将所述训练数据标注为有标签数据和无标签数据；

第四项，自动识别所述训练数据的数据字段类型；

第五项，对所述训练数据进行缺失值填充；

第六项，对所述训练数据中的初始时间字段进行解析，获得并添加新的时间字段，且删除所述初始时间字段；

第七项，自动识别所述训练数据中的非数值数据，并对所述非数值数据进行哈希处理。

23、根据权利要求 20 所述的装置，其中，所述机器学习模型训练模块具体被配置为：

对进行所述预处理后的训练数据进行采样；

对进行所述采样后的训练数据进行前特征选择，获得基础特征；

对所述基础特征进行特征衍生，获得衍生特征；

根据所述基础特征和所述衍生特征，生成训练样本。

24、根据权利要求 23 所述的装置，其中，所述机器学习模型训练模块具体被配置为：

提取进行所述采样后的训练数据中包括的所有属性信息，其中，所述属性信息用于形成特征；

获取每一所述属性信息的特征重要性值；

根据所述特征重要性值，获得基础特征。

25、根据权利要求 24 所述的装置，其中，所述机器学习模型训练模块具体被配置为：

对所有所述特征重要性值进行排序，获得排序结果；

根据所述排序结果，获取第一预定数量的属性信息作为基础特征。

26、根据权利要求 23 所述的装置，其中，所述机器学习模型训练模块具体被配置为：

利用设定的特征生成规则，对所述基础特征进行统计计算、特征组合中的至少一个，

获得衍生特征。

27、根据权利要求 23 所述的装置，其中，所述机器学习模型训练模块具体被配置为：

对所述基础特征和所述衍生特征进行后特征选择；

根据进行所述后特征选择得到的特征，生成训练样本。

28、根据权利要求 27 所述的装置，其中，所述机器学习模型训练模块具体被配置为：

获取每一所述基础特征和每一所述衍生特征的特征重要性值；

对所有所述特征重要性值进行排序，获得排序结果；

根据所述排序结果，获取第二预定数量的特征作为生成训练样本所需要的特征。

29、根据权利要求 18 至 28 中任一项所述的装置，其中，所述装置还包括模型训练方案展示模块，所述模型训练方案展示模块被配置为：

基于训练出的所述机器学习模型，获得模型训练方案；

将所述模型训练方案进行可视化展示；

其中，所述模型训练方案包括用于训练机器学习模型的算法、机器学习模型的超参数、机器学习模型的效果以及特征信息之中的任意一项或者多项；

其中，所述特征信息包括特征数量、特征生成方式以及特征重要性分析结果之中的任意一项或者多项。

30、根据权利要求 29 所述的装置，其中，所述机器学习模型训练模块具体被配置为：

根据可视化展示的预览结果重新训练机器学习模型。

31、根据权利要求 18 至 30 中任一项所述的装置，其中，所述数据预测模块包括批量预测单元，所述批量预测单元被配置为：

响应于针对所述模型预测算子的触发操作，提供进行关于批量预测服务的配置的配置界面；

根据通过所述配置界面输入的配置信息，对所述预测数据进行数据预处理和特征更新处理，获得预测样本；

利用训练出的所述机器学习模型，针对所述预测样本提供预测结果。

32、根据权利要求 31 所述的装置，其中，所述配置界面包括预测结果中字段选择的配置项、模拟实时预测服务的开关状态的配置项中的至少一项。

33、根据权利要求 31 所述的装置，其中，所述数据预测模块包括实时预测单元，所述实时预测单元被配置为：

根据配置实时预测服务的操作，提供进行关于实时预测服务的配置的配置界面；

通过所述配置界面中设定的 API 地址接收包括所述预测数据的预测服务请求；

响应于接收到的所述预测服务请求，利用选择出的机器学习模型得到针对所述预测数据的预测结果，并将预测结果通过所述 API 地址发出。

34、根据权利要求 33 所述的装置，其中，所述配置界面包括用于从训练出的机器学习模型之中选择线上机器学习模型的模型选取规则的配置项、应用资源的配置项中的至少一项。

35、一种包括至少一个计算装置和至少一个存储装置的设备，其中，所述至少一个存储装置被配置为存储指令，所述指令被配置为在被所述至少一个计算装置运行时，促使所述至少一个计算装置执行关于执行自动机器学习过程的如下步骤：

提供相互独立的模型训练算子和模型预测算子；

基于训练数据，利用所述模型训练算子来训练机器学习模型；以及，

利用所述模型预测算子和训练出的所述机器学习模型，针对预测数据提供预测服务。

36、根据权利要求 35 所述的设备，所述指令被配置为在被所述至少一个计算装置运行时，促使所述至少一个计算装置执行关于执行自动机器学习过程的如下步骤：

根据编辑所述模型训练算子的操作，提供编辑界面；

获取通过所述编辑界面输入的算子内容，其中，所述算子内容包括对输入的训练数据进行数据预处理的操作命令、针对数据预处理后的训练数据进行特征工程的操作命令、及根据特征工程的结果进行模型训练的操作命令；

封装所述算子内容，得到所述模型训练算子。

37、根据权利要求 35 或 36 所述的设备，其中，所述基于训练数据，利用所述模型训练算子来训练机器学习模型，包括：

响应于针对所述模型训练算子的触发操作，提供进行关于模型训练的配置的配置界面；

根据通过所述配置界面输入的配置信息，对所述训练数据进行数据预处理和特征工程处理，获得训练样本；

利用至少一种模型训练算法，基于所述训练样本来训练出机器学习模型。

38、根据权利要求 37 所述的设备，其中，所述配置界面包括以下至少一个配置项：机器学习模型的输入源配置项、机器学习模型的适用问题类型配置项、训练机器学习模型的算法模式配置项、机器学习模型的优化目标配置项、及机器学习模型的预测目标字段的字段名配置项。

39、根据权利要求 37 所述的设备，其中，所述对所述训练数据进行数据预处理包括以下各项之中的至少一项：

第一项，对所述训练数据进行数据类型转换；

第二项，对所述训练数据进行采样；

第三项，将所述训练数据标注为有标签数据和无标签数据；

第四项，自动识别所述训练数据的数据字段类型；

第五项，对所述训练数据进行缺失值填充；

第六项，对所述训练数据中的初始时间字段进行解析，获得并添加新的时间字段，且删除所述初始时间字段；

第七项，自动识别所述训练数据中的非数值数据，并对所述非数值数据进行哈希处理。

40、根据权利要求 37 所述的设备，其中，所述对所述训练数据进行数据预处理和特征工程处理，获得训练样本，包括：

对进行所述预处理后的训练数据进行采样；

对进行所述采样后的训练数据进行前特征选择，获得基础特征；

对所述基础特征进行特征衍生，获得衍生特征；

根据所述基础特征和所述衍生特征，生成训练样本。

41、根据权利要求 40 所述的设备，其中，所述对进行所述采样后的训练数据进行前特征选择，获得基础特征，包括：

提取进行所述采样后的训练数据中包括的所有属性信息，其中，所述属性信息用于形成特征；

获取每一所述属性信息的特征重要性值；

根据所述特征重要性值，获得基础特征。

42、根据权利要求 41 所述的设备，其中，所述根据所述特征重要性值，获得基础特征，包括：

对所有所述特征重要性值进行排序，获得排序结果；

根据所述排序结果，获取第一预定数量的属性信息作为基础特征。

43、根据权利要求 40 所述的设备，其中，所述对所述基础特征进行特征衍生，获得衍生特征，包括：

利用设定的特征生成规则，对所述基础特征进行统计计算、特征组合中的至少一个，获得衍生特征。

44、根据权利要求 40 所述的设备，其中，所述指令被配置为在被所述至少一个计算装置运行时，促使所述至少一个计算装置执行关于执行自动机器学习过程的如下步骤：

对所述基础特征和所述衍生特征进行后特征选择；

根据进行所述后特征选择得到的特征，生成训练样本。

45、根据权利要求 44 所述的设备，其中，所述对所述基础特征和所述衍生特征进行后特征选择，包括：

获取每一所述基础特征和每一所述衍生特征的特征重要性值；

对所有所述特征重要性值进行排序，获得排序结果；

根据所述排序结果，获取第二预定数量的特征作为生成训练样本所需要的特征。

46、根据权利要求 35 至 45 中任一项所述的设备，其中，所述指令被配置为在被所述至少一个计算装置运行时，促使所述至少一个计算装置执行关于执行自动机器学习过程的如下步骤：

基于训练出的所述机器学习模型，获得模型训练方案；

将所述模型训练方案进行可视化展示；

其中，所述模型训练方案包括用于训练机器学习模型的算法、机器学习模型的超参数、机器学习模型的效果以及特征信息之中的任意一项或者多项；

其中，所述特征信息包括特征数量、特征生成方式以及特征重要性分析结果之中的任意一项或者多项。

47、根据权利要求 46 所述的设备，其中，所述指令被配置为在被所述至少一个计算装置运行时，促使所述至少一个计算装置执行关于执行自动机器学习过程的如下步骤：

根据可视化展示的预览结果重新训练机器学习模型。

48、根据权利要求 35 至 47 中任一项所述的设备，其中，所述利用所述模型预测算子和训练出的所述机器学习模型，针对预测数据提供预测服务，包括：

响应于针对所述模型预测算子的触发操作，提供进行关于批量预测服务的配置的配置界面；

根据通过所述配置界面输入的配置信息，对所述预测数据进行数据预处理和特征更新处理，获得预测样本；

利用训练出的所述机器学习模型，针对所述预测样本提供预测结果。

49、根据权利要求 48 所述的设备，其中，所述配置界面包括预测结果中字段选择的配置项、模拟实时预测服务的开关状态的配置项中的至少一项。

50、根据权利要求 48 所述的设备，其中，所述指令被配置为在被所述至少一个计算装置运行时，促使所述至少一个计算装置执行关于执行自动机器学习过程的如下步骤：

根据配置实时预测服务的操作，提供进行关于实时预测服务的配置的配置界面；

通过所述配置界面中设定的 API 地址接收包括所述预测数据的预测服务请求；

响应于接收到的所述预测服务请求，利用选择出的机器学习模型得到针对所述预测数据的预测结果，并将预测结果通过所述 API 地址发出。

51、根据权利要求 50 所述的设备，其中，所述配置界面包括用于从训练出的机器学习模型之中选择线上机器学习模型的模型选取规则的配置项、应用资源的配置项中的至少一项。

52、一种计算机可读存储介质，其中，其上存储有计算机程序，所述计算机程序在被处理器执行时实现如权利要求 1 至 17 中任一项所述的方法。

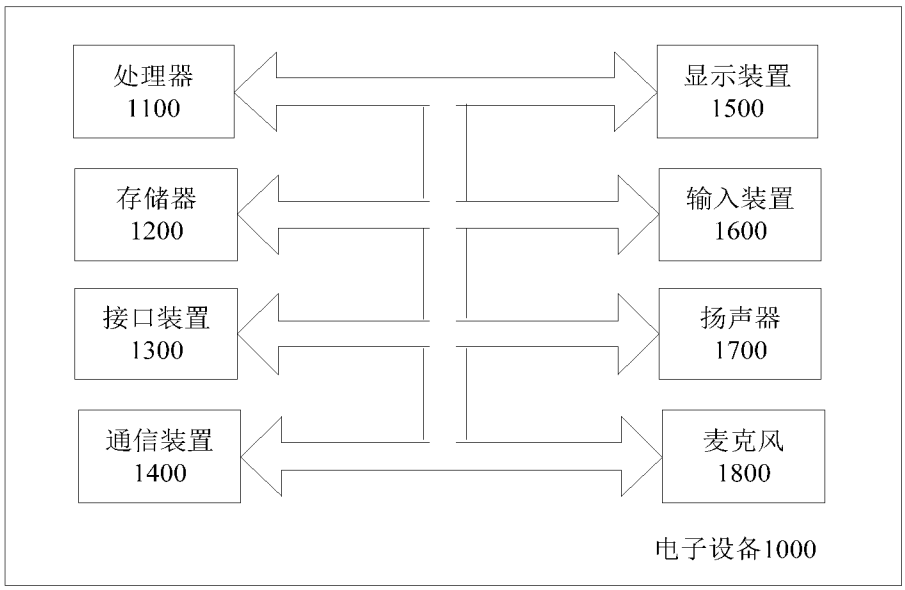

处理器
1100
显示装置
1500
存储器
1200
输入装置
1600
接口装置
1300
扬声器
1700
通信装置
1400
麦克风
1800
电子设备1000

图 1

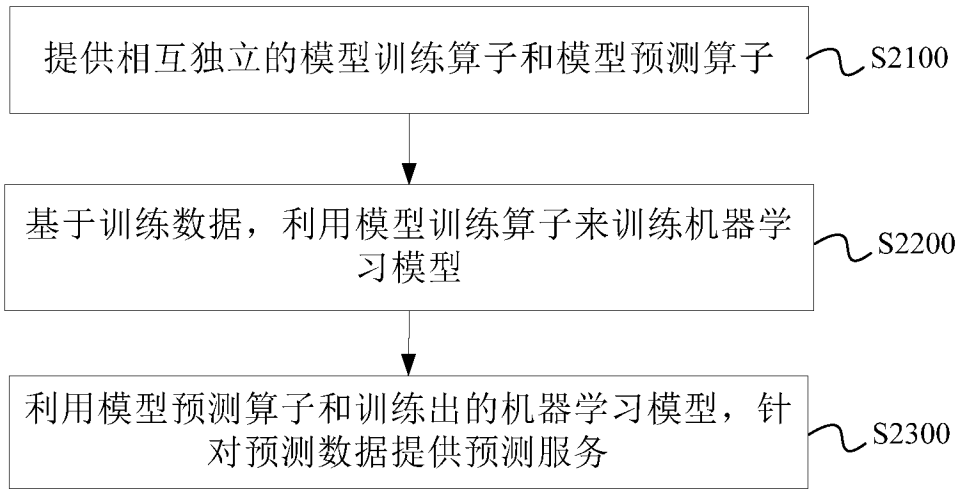

提供相互独立的模型训练算子和模型预测算子
S2100
基于训练数据，利用模型训练算子来训练机器学习模型
S2200
利用模型预测算子和训练出的机器学习模型，针对预测数据提供预测服务
S2300

图 2

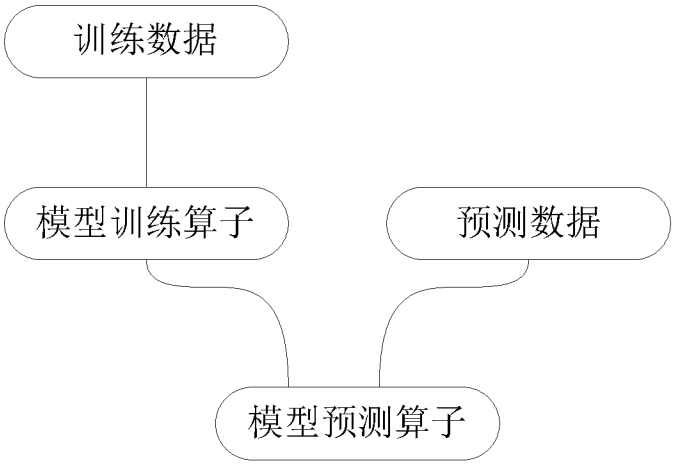


图 3



图 4

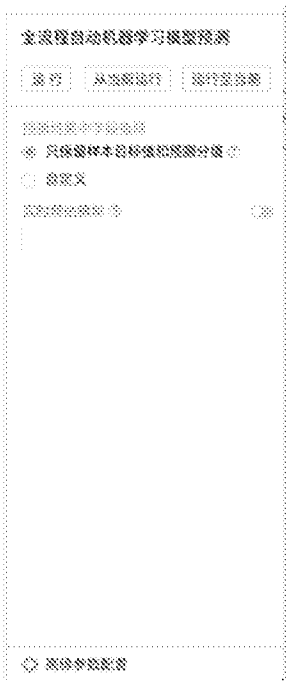


图 5



图 6

∨ 模型详情

基本信息

创建时间：2019-09-23 5:06:51　　创建人：lvzihui　　权重非零特征维度：475594

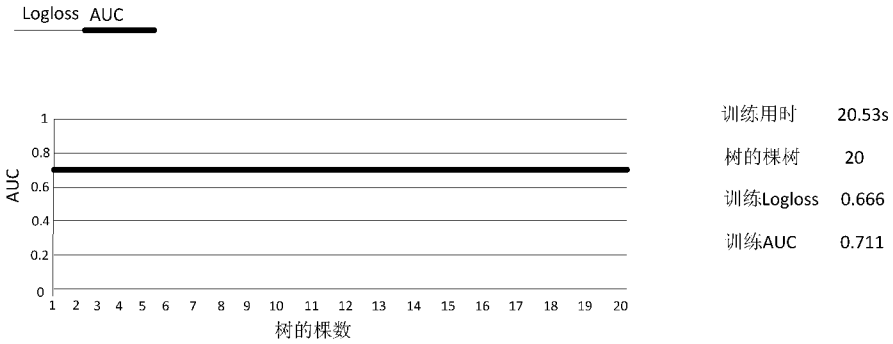
算法参数配置

算法名称：gbdt　　学习率：0.01　　树的棵树：20

单棵树的最大深度：5　　L0正则项系数：0　　L2正则项系数：0

叶子节点最小权重：0　　叶子节点最小划分增益：0

训练过程概要



训练用时	20.53s
树的棵树	20
训练Logloss	0.666
训练AUC	0.711

∨ 特征详情

行数:20678　　列数:25

序号	1 action_type String	2 combined_shot_type String	3 game_event_id String	4 game_id String	5 lat Double	6 loc_x Int
1	Alley Oop Dunk Shot	Dunk	110	20600969	34.0443	0

图 7


执行自动机器学习过程的装置8000
交互模块8100
机器学习模型训练模块8200
数据预测模块8300

图 8

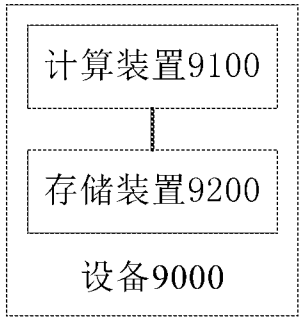

计算装置9100
存储装置9200
设备9000

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/082518

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/22(2019.01)i; G06F 16/25(2019.01)i; G06N 20/00(2019.01)i; G06K 9/62(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G06K; G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: 机器学习, 模型, 训练, 预测, 算子, 界面, 配置, 特征工程, 样本, machine learning, model, train, predict, operator, interface, configuration, characteristic project, sample

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 111611240 A (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 01 September 2020 (2020-09-01) description, paragraphs 4-123	1-52
Y	CN 108875045 A (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 23 November 2018 (2018-11-23) description paragraphs 6-7, 13-14, 39-40, 60-64, 82-83	1-52
Y	CN 111008707 A (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 14 April 2020 (2020-04-14) description paragraphs 29-60, 65-81	1-52
A	CN 109242040 A (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 18 January 2019 (2019-01-18) entire document	1-52
A	CN 107844837 A (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 27 March 2018 (2018-03-27) entire document	1-52

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 June 2021	**23 June 2021**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/082518

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 107590254 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 16 January 2018 (2018-01-16) entire document	1-52

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/082518

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 111611240 A	01 September 2020	None	
CN 108875045 A	23 November 2018	None	
CN 111008707 A	14 April 2020	None	
CN 109242040 A	18 January 2019	None	
CN 107844837 A	27 March 2018	CN 111652380 A	11 September 2020
CN 107590254 A	16 January 2018	None	

国际检索报告

国际申请号
PCT/CN2021/082518

A. 主题的分类

G06F 16/22(2019.01)i; G06F 16/25(2019.01)i; G06N 20/00(2019.01)i; G06K 9/62(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F; G06K; G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT, CNKI, WPI, EPODOC: 机器学习, 模型, 训练, 预测, 算子, 界面, 配置, 特征工程, 样本, machine learning, model, train, predict, operator, interface, configuration, characteristic project, sample

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 111611240 A (第四范式北京技术有限公司) 2020年 9月 1日 (2020-09-01) 说明书第4-123段	1-52
Y	CN 108875045 A (第四范式北京技术有限公司) 2018年 11月 23日 (2018-11-23) 说明书第6-7、13-14、39-40、60-64、82-83段	1-52
Y	CN 111008707 A (第四范式北京技术有限公司) 2020年 4月 14日 (2020-04-14) 说明书第29-60、65-81段	1-52
A	CN 109242040 A (第四范式北京技术有限公司) 2019年 1月 18日 (2019-01-18) 全文	1-52
A	CN 107844837 A (第四范式北京技术有限公司) 2018年 3月 27日 (2018-03-27) 全文	1-52
A	CN 107590254 A (华南理工大学) 2018年 1月 16日 (2018-01-16) 全文	1-52

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2021年 6月 10日	2021年 6月 23日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	杨春雨
传真号 (86-10)62019451	电话号码 86-(10)-53961437

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/082518

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 111611240 A	2020年 9月 1日	无	
CN 108875045 A	2018年 11月 23日	无	
CN 111008707 A	2020年 4月 14日	无	
CN 109242040 A	2019年 1月 18日	无	
CN 107844837 A	2018年 3月 27日	CN 111652380 A	2020年 9月 11日
CN 107590254 A	2018年 1月 16日	无	