

阿里云 机器学习PAI

PAI-AutoLearning 自动学习

文档版本：20200628

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。
2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。
3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。
4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。
5. 阿里云文档中所有内容，包括但不限于图片、架构设计、页面布局、文字描述，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。
6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	 禁止： 重置操作将丢失用户配置数据。
	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	 警告： 重启操作将导致业务中断，恢复业务时间约十分钟。
	用于警示信息、补充说明等，是用户必须了解的内容。	 注意： 权重设置为0，该服务器不会再接受新请求。
	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	 说明： 您也可以通过按Ctrl + A选中全部文件。
>	多级菜单递进。	单击 设置 > 网络 > 设置网络类型 。
粗体	表示按键、菜单、页面名称等UI元素。	在 结果确认 页面，单击 确定 。
Courier字体	命令。	执行 <code>cd /d C:/window</code> 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	<code>bae log list --instanceid Instance_ID</code>
[]或者[a b]	表示可选项，至多选择一个。	<code>ipconfig [-all]-t</code>
{ }或者[a b]	表示必选项，至多选择一个。	<code>switch {active stand}</code>

目录

法律声明	I
通用约定	I
1 概述	1
2 OSS授权	2
3 子账号授权	3
4 通用模型训练	5
4.1 推荐召回.....	5
4.2 图像分类.....	8
4.3 模型部署及调用.....	10
4.4 推荐召回示例.....	11
4.5 图像分类示例.....	18

1 概述

PAI-AutoLearning是一个自动机器学习平台，支持在线标注、自动模型训练、超参优化及模型评估。您只需要准备少量标注数据，并设置训练时长，就可以得到深度优化的模型。同时，PAI-AutoLearning与PAI-EAS高效对接，从而可以快速将训练模型部署为RESTful服务。

通用模型训练

PAI-AutoLearning通用模型训练包括**推荐召回**和**图像分类**场景。

推荐召回的核心是召回和排序。召回是指从海量的待推荐候选集中，选取待推荐列表。排序是指对待推荐列表进行排序。您可以将推荐召回和PAI-Studio的召回算法搭配使用，快速构建完整的召回流程。**推荐召回**包括以下模块：

- **召回策略配置**：在表格存储中配置召回策略，支持协同过滤召回、语意召回及自定义召回策略。
- **数据过滤策略配置**：配置召回结果中需要过滤的**User**和**Item**。例如，希望召回列表过滤**001**这款产品，则将**001**配置到表格存储，系统会自动过滤该产品。
- **模型部署及测试**：测试召回模型。如果对推荐结果满意，则可以将该模型部署到PAI-EAS，形成在线服务。

图片分类是指对一组图片添加标签并进行分类，适用于相册图片自动分类及动植物分类等场景。**图片分类**包括数据标注、模型训练评估及模型试用部署模块。

2 OSS授权

PAI-AutoLearning以OSS为数据源，因此需要对OSS角色进行授权。

步骤一：创建RAM角色

1. 登录[RAM控制台](#)。
2. 在左侧导航栏，单击**RAM角色管理**。
3. 在**RAM角色管理**页面，单击**创建RAM角色**。
4. 在右侧参数配置页面，选择**当前可信实体类型**为**阿里云服务**，并单击**下一步**。
5. 在**角色名称**文本框，输入**AutoLearningRole**。
6. 从**选择授信服务**列表，选择**多媒体转码服务**。
7. 单击**完成**。

步骤二：修改RAM角色的信任策略

1. 在**RAM角色管理**页面，单击**RAM角色名称**下的**AutoLearningRole**
2. 在角色**基本信息**页面，单击**信任策略管理**页签。
3. 单击**修改信任策略**。
4. 在**修改信任策略**页面，将信任策略修改为如下内容，并单击**确定**。

```
{
  "Statement": [
    {
      "Action": "sts:AssumeRole",
      "Effect": "Allow",
      "Principal": {
        "Service": [
          "1276057991250096@odps.aliyuncs.com"
        ]
      }
    }
  ],
  "Version": "1"
}
```

步骤三：授权RAM角色

1. 在**RAM角色管理**页面，单击**RAM角色名称**下的**AutoLearningRole**
2. 在角色**基本信息**页面的**权限管理**页签，单击**添加权限**。
3. 在**添加权限**页面，从**选择权限**列表，选择**AliyunOSSFullAccess**授权策略。
4. 单击**确定**。

3 子账号授权

如果使用子账号管理实例及训练模型，则需要主账号对子账号进行授权。

步骤一：创建权限策略

1. 登录RAM控制台。
2. 在左侧导航栏，选择**权限管理 > 权限策略管理**。
3. 在**权限策略管理**页面，单击**创建权限策略**。
4. 在**新建自定义权限策略**页面的**策略名称**文本框，输入**autolearningfullaccess**。
5. 在**新建自定义权限策略**页面，选择**配置模式**为**脚本配置**，并根据实际需要修改**策略内容**。

策略内容的示例如下。

```
{
  "Version": "1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "autolearning:CreateInstance",
        "autolearning:LabelInstance",
        "autolearning:TrainInstance",
        "autolearning>EditInstance",
        "autolearning:ListInstance"
      ],
      "Resource": "*"
    }
  ]
}
```

权限点	权限内容
CreateInstance	实例权限：创建或删除实例。
EditInstance	实例编辑权限。
LabelInstance	实例标注权限：增加、删除或修改标签及图片标注。
TrainInstance	模型训练权限。
ListInstance	实例查看权限。

步骤二：子账号授权

1. 在RAM控制台的左侧导航栏，选择**人员管理 > 用户**。
2. 在**用户**页面，单击待授权子账号所在行的**添加权限**。
3. 在**添加权限**页面，选择**选择权限**为**自定义策略**。

4. (可选) 在**选择权限**区域的文本框, 输入已创建的权限策略名称**autolearningfullaccess**, 搜索该策略。
5. 从**权限策略**列表, 选择**autolearningfullaccess**。
6. 单击**确定**。

4 通用模型训练

4.1 推荐召回

推荐召回的使用流程包括创建实例、配置召回策略、配置数据过滤策略、测试及部署模型。

前提条件

完成OSS授权，详情请参见[OSS授权](#)。

创建OTS实例，详情请参见[#unique_8](#)。

步骤一：创建实例

1. 登录[PAI控制台](#)。
2. 在PAI控制台首页，选择**AutoLearning > 通用模型训练**。
3. 在**AutoLearning**页面，单击**创建实例**。
4. 在**创建实例**页面，配置参数。

参数	描述
实例类型	选择实例类型为 推荐召回 。系统包括以下实例类型： <ul style="list-style-type: none">• 图片分类• 推荐召回
实例名称	长度为1~20个字符，以大写字母、小写字母或中文开头，只能包含中文、字母、下划线（_）或数字。
实例说明	实例的简要说明，便于区分各实例。
储存依赖	使用推荐召回服务，需要将训练数据存储到阿里云表格存储OTS（Table Store，类似于Redis的KV存储数据库）中，详情请参见 OTS创建数据表 。
实例绑定	绑定OTS实例。

5. 单击**确认**。

步骤二：配置召回策略

1. 在**AutoLearning**页面，单击**操作**列下的**打开**。

2. 在召回策略配置的配置向导页面，配置召回策略。

区域	参数	描述
协同过滤召回	策略名称	长度为1~20个字符，以大写字母或小写字母开头，只能包含字母、数字或下划线（_）。
	User-Item数据表	选择召回策略的User数据表。将鼠标悬停至User-Item数据表后的问号图标，查看数据表格式说明。
	Item-Item数据表	选择召回策略的Item数据表。将鼠标悬停至Item-Item数据表后的问号图标，查看数据表格式说明。
	召回数量	该召回策略下，返回的召回数量。取值为正整数。如果召回结果的数量小于配置的召回数量，则以实际召回结果为准。
语义召回	策略名称	长度为1~20个字符，以大写字母或小写字母开头，只能包含字母、数字或下划线（_）。
	User-Topic数据表	选择召回策略的User数据表。将鼠标悬停至User-Topic数据表后的问号图标，查看数据表格式说明。
	Topic-Item数据表	选择召回策略的Item数据表。将鼠标悬停至Topic-Item数据表后的问号图标，查看数据表格式说明。
	召回数量	该召回策略下，返回的召回数量。取值为正整数。如果召回结果的数量小于配置的召回数量，则以实际召回结果为准。
自定义召回策略	策略名称	长度为1~20个字符，以大写字母或小写字母开头，只能包含字母、数字或下划线（_）。
	Item数据表	选择召回策略的数据表。将鼠标悬停至Item数据表后的问号图标，查看数据表格式说明。
	召回数量	该召回策略下，返回的召回数量。取值为正整数。如果召回结果的数量小于配置的召回数量，则以实际召回结果为准。

PAI-AutoLearning支持以下召回策略，您可以根据实际需要选择合适的召回策略：

- 协同过滤召回：推荐领域经典的召回策略，根据人和物品的相关性生成召回结果。
- 语义召回：经常应用于新闻推荐，根据人物感兴趣的文章类型进行召回。
- 自定义召回：基于业务生成User-item的召回策略。

3. 单击添加到策略列表，该召回策略显示在已添加策略列表中。

4. （可选）参考以上步骤，添加多个召回策略。

5. 单击下一步。

步骤三：配置数据过滤策略

1. 在数据过滤策略配置的配置向导页面，配置参数。

区域	参数	描述
自定义过滤策略（U-I过滤）策略配置	策略名称	长度为1~20个字符，以大写字母或小写字母开头，只能包含字母、数字或下划线（_）。
	User-item数据表	选择过滤策略的数据表。将鼠标悬停至User-Item数据表后的问号图标，查看数据表格式说明。
自定义过滤策略（I策略过滤）策略配置	策略名称	长度为1~20个字符，以大写字母或小写字母开头，只能包含字母、数字或下划线（_）。
	item数据表	选择过滤策略的数据表。将鼠标悬停至Item数据表后的问号图标，查看数据表格式说明。

系统支持以下配置策略，您可以根据实际需要选择合适的过滤策略：

- 基于User-item关系对的过滤策略。如果召回结果出现某User，则过滤该User对应的item。数据表格式如下。

输入为user_id和item_id，每一行是一个user与item的对应关系对

字段	类别	类型	说明	示例
user_id	主键	string	用户唯一id	10944750
item_ids	属性	string	item 唯一-id 列表，多个以","分隔	17981042,19412678

- 基于item的过滤策略。如果召回结果出现该item的id，则过滤该item。数据表格式如下。

需要在TableStore中内置该表，按照提示格式

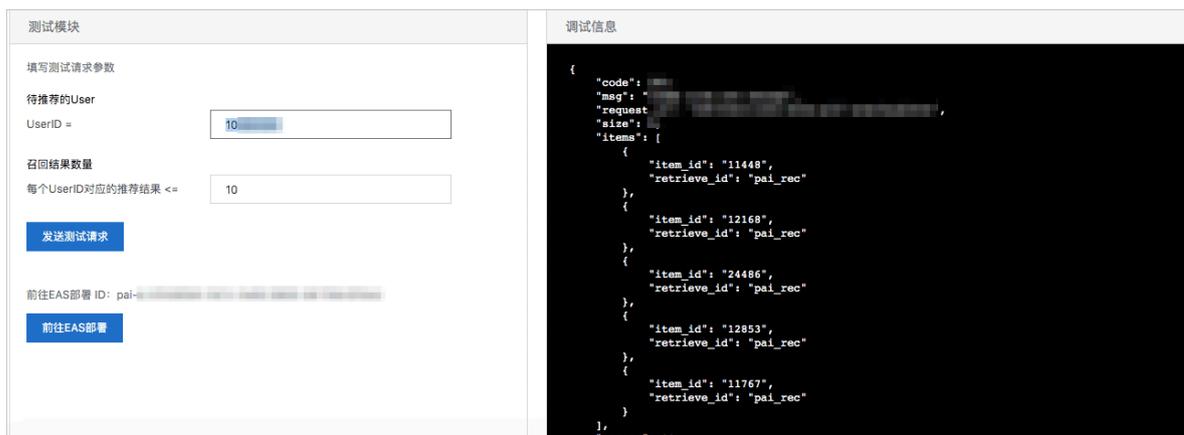
字段	类别	类型	说明	示例
item_id	主键	string	item 唯一-id	10944750

2. 单击**添加到策略列表**，该召回策略显示在**已添加策略列表**中。
3. 单击**部署并测试**。

步骤四：测试及部署

1. 在测试页面，系统会根据配置的召回和过滤策略，将模型部署为服务。

2. 在测试模块区域，输入UserID和召回结果数量。



3. 单击发送测试请求。

4. 在调试信息区域，查看返回的推荐内容。如果对模型结果满意，单击前往EAS部署，将服务部署到PAI-EAS，形成RESTful请求。

4.2 图像分类

图像分类的使用流程包括创建实例、标注数据、训练模型、评估模型及部署模型。

前提条件

如果使用子账号，需要主账号对子账号进行授权，详情请参见[子账号授权](#)。

步骤一：创建实例

1. 登录[PAI控制台](#)。
2. 在PAI控制台首页，选择**AutoLearning > 通用模型训练**。
3. 在**AutoLearning**页面，单击**创建实例**。
4. 在**创建实例**页面，配置参数。

参数	描述
实例类型	选择实例类型为 图片分类 。系统包括以下实例类型： <ul style="list-style-type: none"> • 图片分类 • 推荐召回
实例名称	长度为1~20个字符，以大写字母、小写字母或中文开头，只能包含中文、字母、下划线（_）或数字。
实例说明	实例的简要说明，便于区分各实例。

参数	描述
训练数据	<p>选择图片标注方式，并输入存储图片数据集的OSS路径。系统支持的图片标注方式包括：</p> <ul style="list-style-type: none"> 在线图片标注。如果图片小于50张，可以通过该方式标注图片，详情请参见#unique_11。 <div style="background-color: #f0f0f0; padding: 5px; margin: 5px 0;"> <p> 说明： 所有图片必须在同一文件夹下，支持JPEG、JPG、PNG、BMP及TIFF等图片格式。</p> </div> <ul style="list-style-type: none"> 导入已标注图片文件。如果图片数量大，可以通过该方式标注图片。标注文件的格式如下。 <div style="background-color: #f0f0f0; padding: 5px; margin: 5px 0;"> <pre>id,oss data,label 0,{"tfspath":"","oss://autodl/yuyi/t4.jpeg"},"{"option":"","虎"}" 1,{"tfspath":"","oss://autodl/yuyi/pb5.jpeg"},"{"option":"","北极熊"}" 2,{"tfspath":"","oss://autodl/yuyi/cat4.jpeg"},"{"option":"","猫"}"</pre> </div> <div style="background-color: #f0f0f0; padding: 5px; margin: 5px 0;"> <p> 说明： 所有图片和标注文件必须在同一文件夹下，支持JPEG、JPG、PNG、BMP及TIFF等图片格式。</p> </div>

5. 单击**确认**。

步骤二：标注数据

1. 在**AutoLearning**页面，单击**操作**下的**打开**。
2. 在**数据标注配置向导**页面，标注所有图片，详情请参考[#unique_11](#)。
3. 单击**完成标注**。

步骤三：训练及评估模型

1. 在**开始训练**对话框的**训练设置**区域，输入**最大训练时长**。

PAI-AutoLearning预置的Early Stop机制，可以防止模型训练时间过长而出现拟合现象。



说明：

最大训练时长的取值范围：10 min~60 min。

2. （可选）选中**增量训练**复选框，并从模型版本列表，选择一个模型训练基础版本。
3. 单击**开始训练**。
4. 在**模型训练及评估配置向导**页面，查看**模型训练进度**及**最新评估结果**。

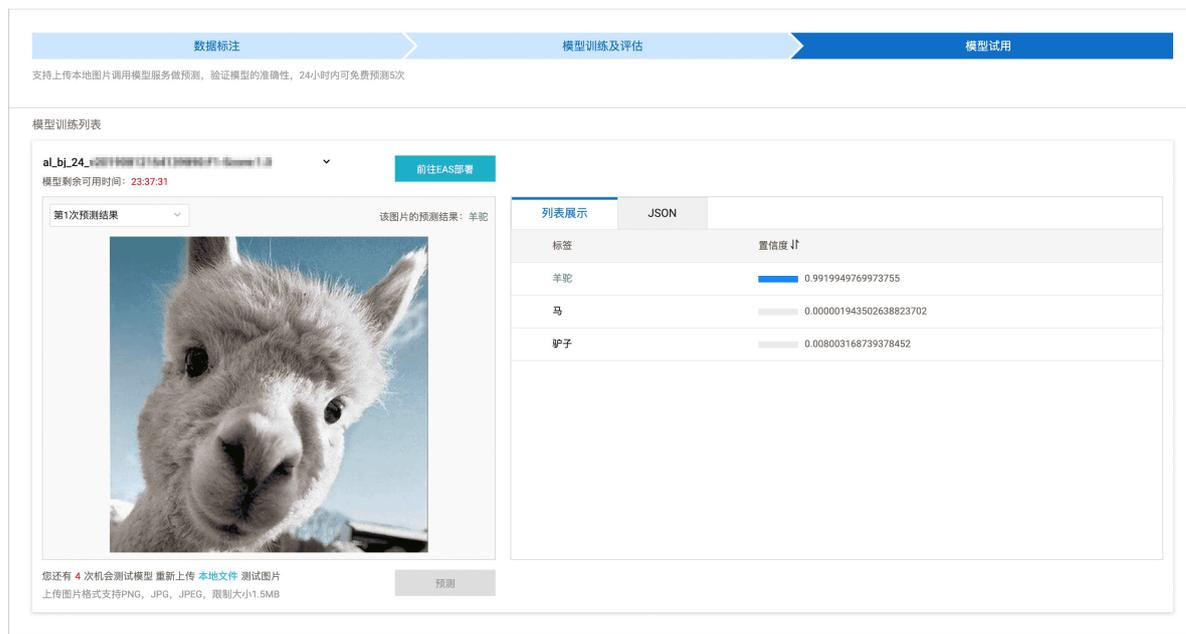
5. (可选) 在**模型训练列表**区域, 单击  图标, 可以查看历史训练模型的评估结果, 包括**准确率、精确率、召回率及F1值**。
6. 在**模型训练及评估**配置向导页面的**操作**区域, 可以管理训练模型。

您可以对训练模型进行以下**操作**:

- **试用**: 生成模型后的24小时内, 可以免费试用模型5次。
- **部署**: 将模型部署到PAI-EAS。
- **删除**: 删除无用的模型。
- **查看日志**: 查看模型训练过程中的日志信息。

步骤四: 试用模型

1. 在**模型训练及评估**配置向导页面的**操作**区域, 单击**试用**。
2. 在**模型试用**配置向导页面, 上传本地文件, 使用模型进行预测。



支持上传本地图片调用模型服务做预测, 验证模型的准确性。24小时内可免费预测5次

模型训练列表

al_bj_24_... 前往EAS部署

模型剩余可用时间: 23:37:31

第1次预测结果 该图片的预测结果: 羊驼

0.9919949769973755

标签	置信度
羊驼	0.9919949769973755
马	0.000001943502638823702
驴子	0.008003168739378452

您还有 4 次机会测试模型 重新上传 本地文件 测试图片

上传图片格式支持PNG, JPG, JPEG, 限制大小1.5MB

预测

首次试用模型, 预计需要3 min~5 min。

3. 如果对模型预测结果满意, 则单击**前往EAS部署**, 将模型部署为RESTful服务。PAI-EAS会收取费用, 详情请参见[#unique_12](#)。

4.3 模型部署及调用

使用PAI-EAS将PAI-AutoLearning训练的深度优化模型部署为在线服务, 从而对外提供API服务。

前提条件

完成模型训练及评估, 详情请参见[推荐召回](#)或[图像分类](#)。

部署模型

1. 进入**模型试用**配置向导页面。

- a) 登录**PAI控制台**。
- b) 在PAI控制台首页，选择**AutoLearning > 通用模型训练**。
- c) 在**AutoLearning**页面，单击**操作**列下的**打开**。
- d) 单击**模型试用**页签。

2. 在**模型试用**配置向导页面，单击**前往EAS部署**。

3. 在**资源和模型**页面，选择**资源种类**，并单击**下一步**。

PAI-EAS预置了PAI-AutoLearning的**AutoLearning Processor**，您无需手动配置**Processor**种类。

4. 在**部署详情及配置确认**页面，输入**自定义模型名称**。

5. 在**实例数和Quota**调整框，单击  或  图标调整资源数量。

6. 单击**部署**。

进入**PAI EAS模型在线服务**页面。如果**模型状态**变为**运行中**，则模型部署成功。

调用模型

1. 在**PAI EAS模型在线服务**页面，单击**服务方式**下的**调用信息**。

2. PAI-EAS支持公网地址或VPC地址调用模型，您可以根据需要选择任何一种调用方式：

- 公网地址调用（常用）
 - a. 在**调用信息**页面的**公网地址调用**页签，查看**访问地址**和**Token**。
 - b. 调用服务API，详情请参见[#unique_14](#)。
- VPC地址调用
 - a. 在**调用信息**页面的**VPC地址调用**页签，查看**访问地址**和**Token**。
 - b. 调用服务API，详情请参见[#unique_15](#)。

4.4 推荐召回示例

本文将推荐召回的协同过滤策略与PAI-Studio推荐算法搭配使用，构建完整的召回流程。

数据说明

使用**协同过滤召回策略**，需要将**User-item**和**Item-item**对应的数据分别导入至表格存储OTS（Table Store）。

- **User-item**数据表存储用户对商品的历史行为。例如购买、单击或收藏。

user_id ▲	item_id ▲	active_type ▲	active_date ▲
10944750	13451	0	6月4日
10944750	13451	2	6月4日
10944750	13451	2	6月4日
10944750	13451	0	6月4日
10944750	13451	0	6月4日
10944750	13451	0	6月4日
10944750	13451	0	6月4日
10944750	13451	0	6月4日
10944750	13451	0	6月4日
10944750	21110	0	6月7日
10944750	8689	0	5月2日
10944750	8689	2	5月2日
10944750	8689	2	5月2日
10944750	8689	0	5月2日
10944750	8689	0	5月2日
10944750	26619	0	6月28日
10944750	18575	0	5月2日
10944750	23662	0	6月19日

- **user_id**: 用户ID。
- **item_id**: 商品ID。
- **active_type**: 行为。**0**表示单击, **1**表示购买, **2**表示收藏。

数据需要存储至OTS, PAI-AutoLearning支持的数据格式如下图所示。其中**user_id**为主键, **item_ids**为每个**user_id**对应的**item**, 多个**item**以英文逗号(,)分隔。

输入为user_id和item_id, 每一行是一个user与item的对应关系对

字段	类别	类型	说明	示例
user_id	主键	string	用户唯一id	10944750
item_ids	属性	string	用户浏览列表	17981042,19412678

- Item-item数据表存储通过协同过滤算法计算得到的item相似度。

item_id	similar_item_ids
13011863	17297683:1,17273476:1,17335402:1,17269664:1,17277077:1,17458472:1,17324042:1,17488275:1,17336239:1,1737
17136593	17538896:1,20629991:1,21523209:1,21653256:0.6666666667,21623893:0.6666666667,17258266:0.5,1758928
17149439	17311760:1,17258266:1,17437753:1,17487869:1,17463871:1,17489382:1,17498471:1,17719916:1,17522998:1,17713
17184154	17817742:1,17442671:1,17710549:1,17350723:1,17637548:1,17818884:1,17699158:1,17850689:1,17717878:1,17720
17187342	19029212:1,20066435:1,17282400:0.5,17719575:0.5,17702334:0.5,17522786:0.5,17584686:0.5,17708789:0.5,177

- **item_id**: 目标商品。
- **similar_item_ids**: **key:value**结构，其中**key**表示商品ID，**value**表示该商品与目标商品的关系。**value**值越大，表示关系越紧密。多个**key:value**数据以英文逗号(,)分隔。

数据需要存储至OTS，PAI-AutoLearning支持的数据格式如下图所示。其中**item_ids**为主键，**similar_item_ids**为每个**item**对应的相似**item**列表，多个**item**对以英文逗号(,)分隔。系统支持**item**不带权重值的情况。

输入Item与Item的相似匹配结果，如果一个Item对应多个Item需要以KV格式列出多个Item以及对应的权重

字段	类别	类型	说明	示例
item_id	主键	string	item唯一id	10944750
similar_item_ids	属性	string	相似的item列表,kv形式,签名为itemId,后面值为权重	7484728:1,16966098:1 或7484728,16966098

步骤一：生成训练数据

1. 进入PAI-Studio控制台页面。
 - a) 登录PAI控制台。
 - b) 在PAI控制台首页，选择模型开发和训练 > Studio-可视化建模。
 - c) 在PAI可视化建模页面，单击进入机器学习。



2. 在左侧菜单栏，单击**首页**。
3. 在【**推荐算法**】**商品推荐**区域，单击**从模板创建**。
4. 在画布中，保留如下组件，删除多余组件。



cf_训练_data为user-item对应数据，**协同过滤**组件为Item-item对应数据。

5. 单击**协同过滤**组件，在右侧**参数设置**页签，配置**TopN**为**5**（对于每个**item**返回5个相关**item**）。
6. 在左侧菜单栏，单击**组件**。
7. 在组件列表，将源/目标下的写数据表两次拖入画布，并分别重命名为**user_item_data**和**item_item_data**。
8. 将**cf训练_data**组件的输出与**user_item_data**连接，**协同过滤**组件的输出与**item_item_data**相连。
9. 单击画布上方的**运行**。

步骤二：将数据导入至OTS

PAI-Studio生成的数据存储在MaxCompute，需要将该数据导入至OTS，才能供PAI-AutoLearning使用。

1. 创建OTS数据表，详情请参见[#unique_9](#)。

创建**user-item**和**item-item**表，表字段和主键名必须与数据说明中的一致。**user-item**表的示例如下。

主键列表	
序号	主键名称
1	user_id

预定义列	
序号	属性列名称
1	item_ids

2. 使用DataWorks将数据导入至OTS。

- 创建OTS数据源，详情请参见[#unique_17](#)。
- 创建离线同步节点，详情请参见[#unique_18](#)。
- 配置**数据来源**和**数据去向**，详情请参见[#unique_19](#)。



因为OTS数据源不支持向导模式，所以需要单击[点击转换为脚本](#)，使用脚本导入。

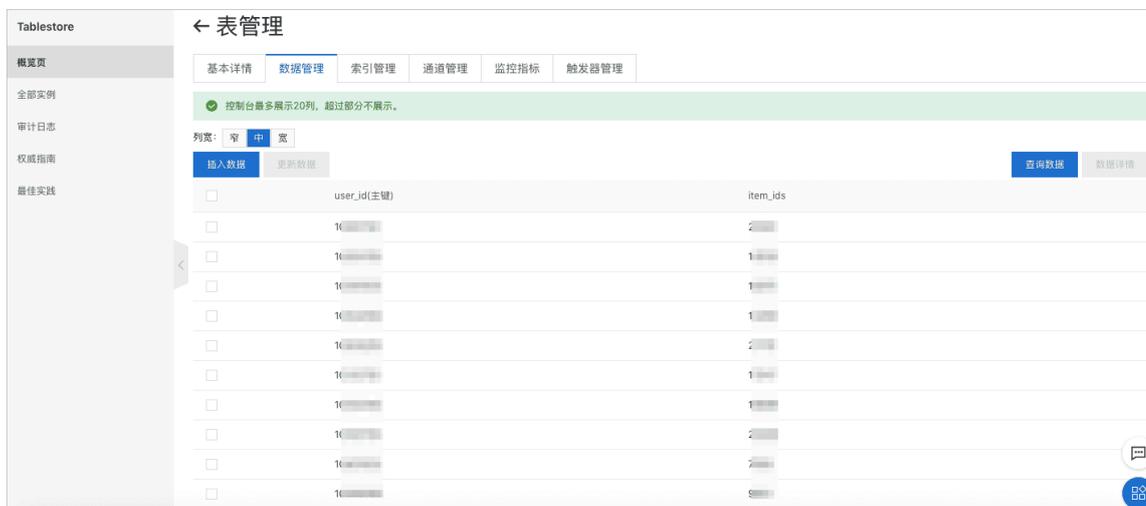
```
{
  "type": "job",
  "steps": [
    {
      "stepType": "odps",
      "parameter": {
        "partition": [],
        "datasource": "odps_first",
        "column": [
          "user_id", //MaxCompute表的列名。
          "item_id" //MaxCompute表的列名。
        ],
        "table": "user_item_data" //MaxCompute表名。
      },
      "name": "Reader",
      "category": "reader"
    }
  ]
}
```

```
    },
    {
      "stepType": "ots",
      "parameter": {
        "datasource": "otc_data", //在数据集成配置的OTS资源名。
        "column": [
          {
            "name": "item_ids", //OTS字段名。
            "type": "STRING"
          }
        ],
        "writeMode": "UpdateRow",
        "table": "user_item", //OTS表名。
        "primaryKey": [
          {
            "name": "user_id", //OTS主键名。
            "type": "STRING"
          }
        ]
      },
      "name": "Writer",
      "category": "writer"
    }
  ],
  "version": "2.0",
  "order": {
    "hops": [
      {
        "from": "Reader",
        "to": "Writer"
      }
    ]
  },
  "setting": {
    "errorLimit": {
      "record": ""
    },
    "speed": {
      "throttle": false,
      "concurrent": 2
    }
  }
}
```

使用示例时，需要删除注释。

3. 查看OTS数据导入情况。

- a) 在**概览页**页面，单击**实例名称**或**操作**列下的**实例管理**。
- b) 在**实例详情页**的**数据表列表**区域，单击**操作**列下的**数据管理**。



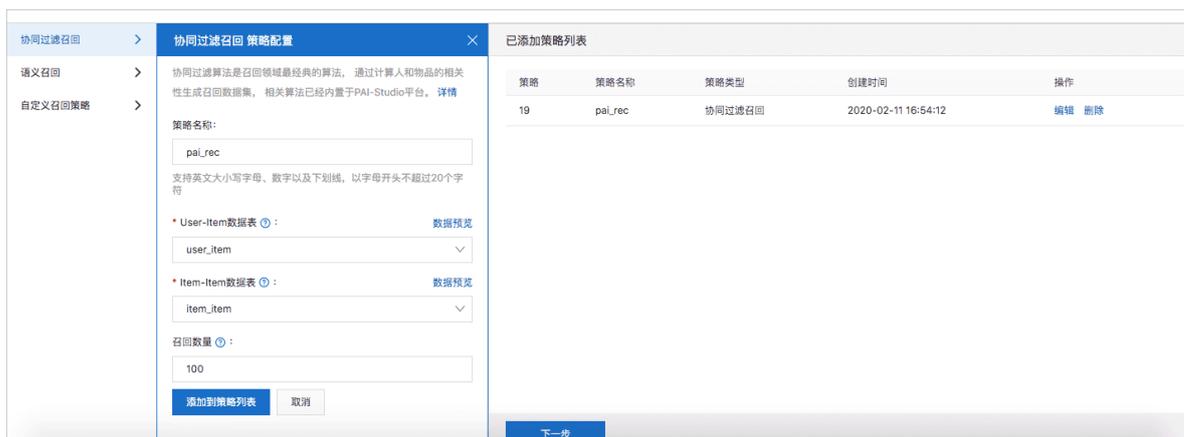
由于PAI-Studio生成的数据是空格分隔（如下图**similarity**），而PAI-Autolearning支持的数据是英文逗号（,）分隔。所以导入数据时，需要先使用DataWorks的SQL工具对数据进行预处理，再导入至OTS。

itemid ▲	similarity ▲
1000	24277:1 18353:1 11849:1 18612:0.5 15784:0.5
10014	26523:0.5 1015:0.25 10584:0.25 12209:0.25 11328:0.25
10066	13888:0.333333333333 3228:0.055555555556

步骤三：配置召回策略

1. 进入**AutoLearning**页面。
 - a) 登录**PAI控制台**。
 - b) 在PAI控制台首页，选择**AutoLearning > 通用模型训练**。
2. 在**AutoLearning**页面，单击**创建实例**。
3. 在**创建实例**页面，选择**实例类型**为**推荐召回**。
4. 在**实例名称**文本框输入**test**，并单击**确认**。
5. 在**AutoLearning**页面，单击**操作**列下的**打开**。
6. 在**协同过滤召回**页面，选择对应的表，并在**召回数量**文本框输入**100**。

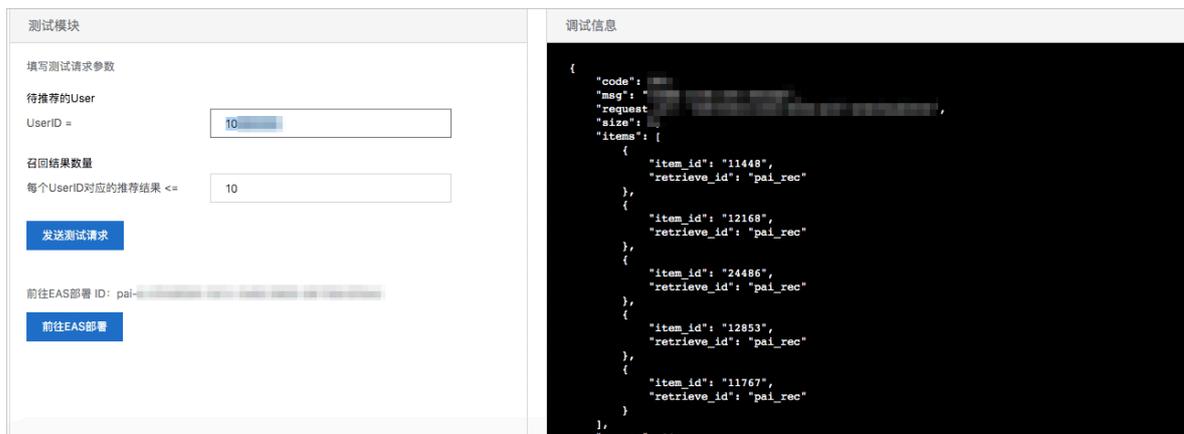
7. 单击添加到策略列表。



8. 单击下一步。

步骤四：部署并测试模型

1. 在数据过滤策略配置页面，单击部署并测试。
2. 在测试模块区域，输入测试数据中的某UserID，并在召回结果数量文本框输入10。
3. 单击发送测试请求。



4. 在调试信息区域，查看返回的推荐内容。如果对模型结果满意，单击前往EAS部署，将服务部署到PAI-EAS，形成RESTful请求。

4.5 图像分类示例

本文以系统提供的骡子、马及羊驼分类模型为例，介绍PAI-AutoLearning全自动机器学习平台的训练流程。

模型介绍

系统提供骡子、马及羊驼图片共33张，通过对该组图片进行训练，得到准确率达80%以上的骡子、马及羊驼的图片分类模型。如果对该模型输入一张图片，则会返回该图片类别。

训练流程

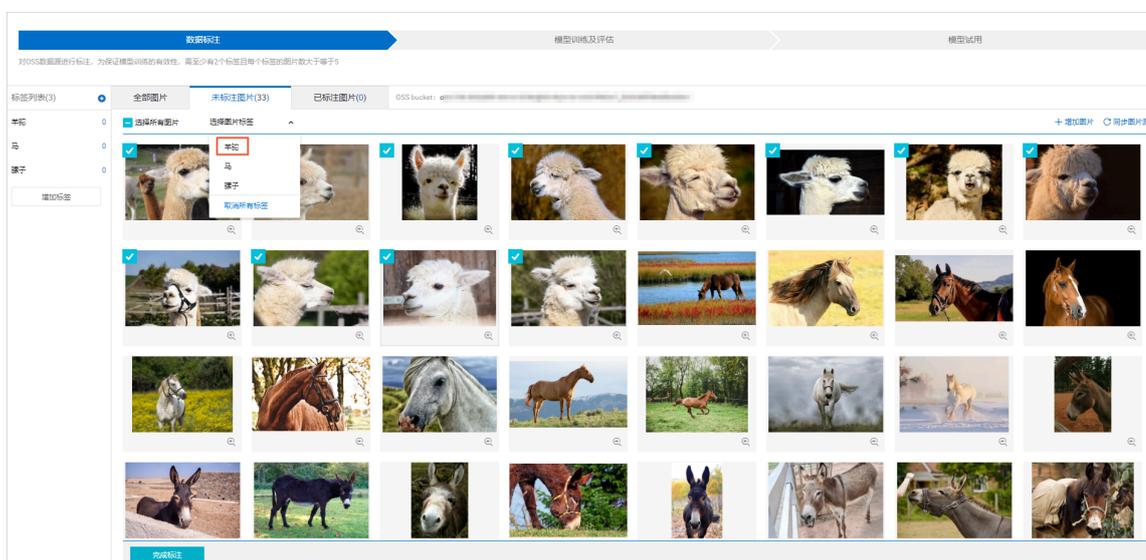
1. 进入AutoLearning页面。

- a) 登录PAI控制台。
- b) 在PAI控制台首页，选择**AutoLearning > 通用模型训练**。

2. 在AutoLearning页面的图片分类_动物分类区域，单击从模板创建。

3. 标注数据。

- a) 在**数据标注配置**向导页面的**标签列表**区域，单击**增加标签**。
- b) 在文本框输入**骡子**，并按下Enter键。
- c) 参考上述步骤，增加**马**和**羊驼**标签。
- d) 选中羊驼图片，并单击**选择图片标签**下的**羊驼**。



- e) 参考上述步骤，标注骡子和马的图片。

- f) 单击**完成标注**。

4. 训练并评估模型。

- a) 在**开始训练**对话框的**最大训练时长**文本框，输入**10**。

PAI-AutoLearning预置的Early Stop机制，可以防止模型训练时间过长而出现拟合现象。



说明：

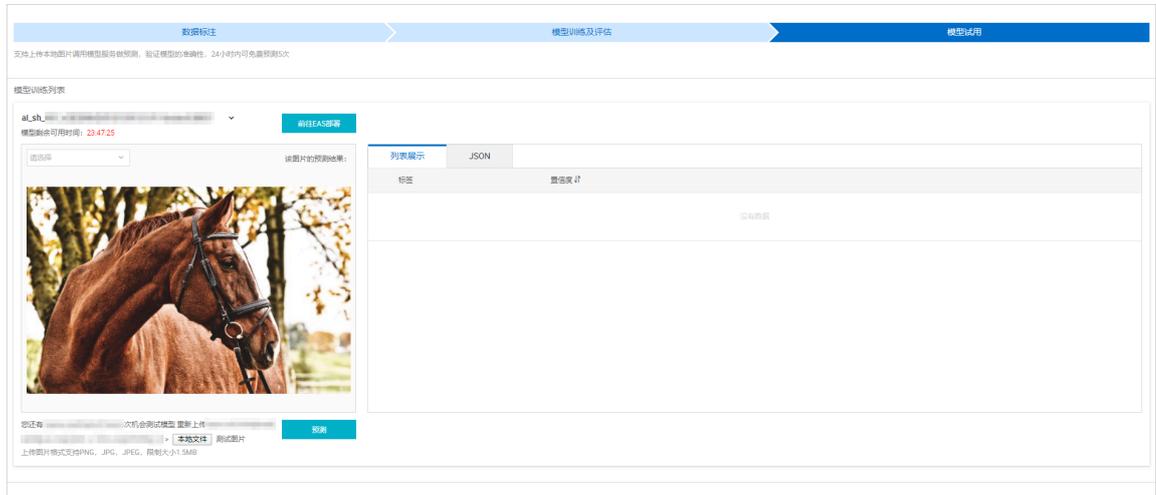
最大训练时长的取值范围：10 min~60 min。

- b) 单击**开始训练**。
- c) 在**模型训练及评估**配置向导页面，查看**模型训练进度**及**最新评估结果**。

如果需要提高模型准确率，可以增加训练数据量或模型训练时长。

5. 试用模型。

- 在**模型训练及评估配置向导页面**的**操作区域**，单击**试用**。
- 在**模型试用配置向导页面**，将本地一张动物图片拖拽至上传文件区域。



- 单击**预测**。
- 在**列表展示**页签，查看模型预测结果的**置信度**。

