泰迪智能科技
服务热线：40068-40020

Python 数据挖掘建模平台

操 作 手 册

地址：广州市黄埔区科学城开泰大道 36 号
网址： http://www.tipdm.com
邮箱： services@tipdm.com
热线： 40068-40020
邮编： 510000
电话： 020-82039399

目录

文档修订记录

序号	影响版本号	修订内容	修订人	修订时间
1	V1.0	V1.0 初稿	莫芳	2018-3-3

平台介绍

1.1 产品简介

Python 数据挖掘建模平台是由广州泰迪智能科技有限公司自主研发，面向高校数据挖掘课程教学的数据挖掘建模工具。平台使用 JAVA 语言开发，采用 B/S 结构，用户不需要下载客户端，可通过浏览器进行访问。用户可在没有 Python 编程基础的情况下，通过拖拽的方式进行操作，将数据输入输出、数据预处理、挖掘建模、模型评估等环节通过流程化的方式进行连接，以达到数据分析挖掘的目的。

1.2 名词解释

组件：将建模过程涉及到的输入/输出、数据探索及预处理、建模、模型评估等算法分别进行封装，每一个封装好的算法都可称之为组件。

工程：为实现某一数据挖掘目标，将各组件通过流程化的方式进行连接，整个数据流程称为一个工程。

模型：主要针对分类、回归算法而言，使用一部分数据用于训练，会得到一个模型，里面将保存算法的参数，可使用该模型对另一批数据进行验证或预测。

个人组件：用户可按照平台规定的格式编写脚本，配置相关输入、输出、算法参数，可作为平台组件，反复调用。

任务：支持定时同步数据库数据源至平台或定时运行某一工程。

1.3 技术支持

感谢您选择广州泰迪科技公司的数据挖掘产品，在系统的使用过程中如果遇到问题，请通过如下的方式与我们联系，我们将为您提供周到满意的服务。

主页：http://www.tipdm.com

电话：020-22205718

热线：40068-40020

QQ 群：197738983

地址：广州市黄埔区科学城开泰大道 36 号 1 栋 212

邮编：510000

邮箱：services@tipdm.com

2 快速入门

2.1 新建工程入门

实例：针对通讯企业客户数据，应用 BP 神经网络算法预测客户是否流失，使用平台操作如下。

2.1.1 数据准备

- 打开数据源菜单，如图 1 所示。

图 1

- 选择新建数据源→数据来源于文件，如图 2 所示。

图 2

■ 选择本地数据文件并定义表名，如图 3 所示。

图 3

■ 预览数据，如图 4 所示。

图 4

■ 设置字段名称及类型，如图 5 所示。

图 5

■ 点击确定即可上传成功。

2.1.2 新建工程

■ 进入工程界面，如图 6 所示。

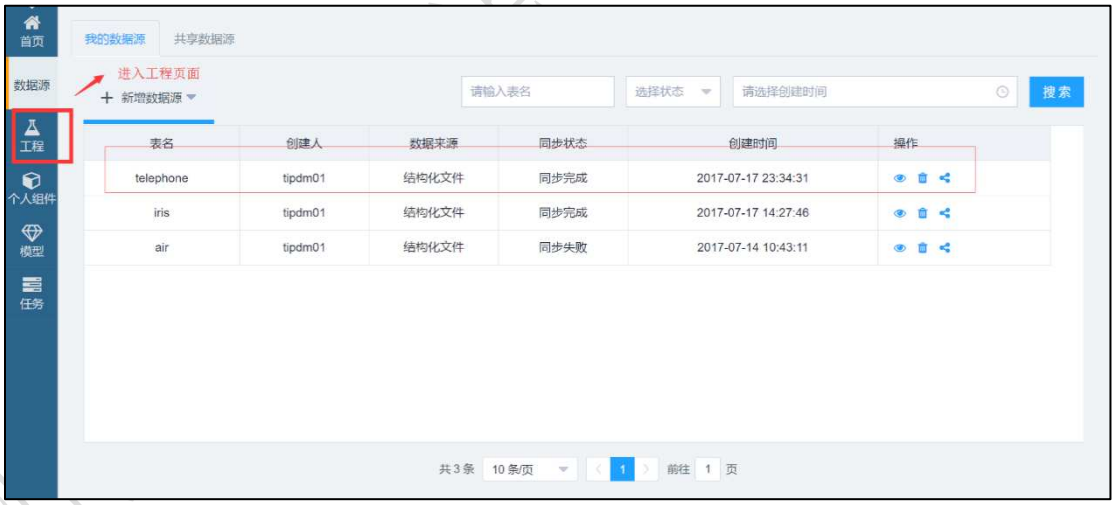


图 6

■ 新建工程，如图 7 所示。

图 7

■ 拖入“输入源”组件，如图 8 所示。

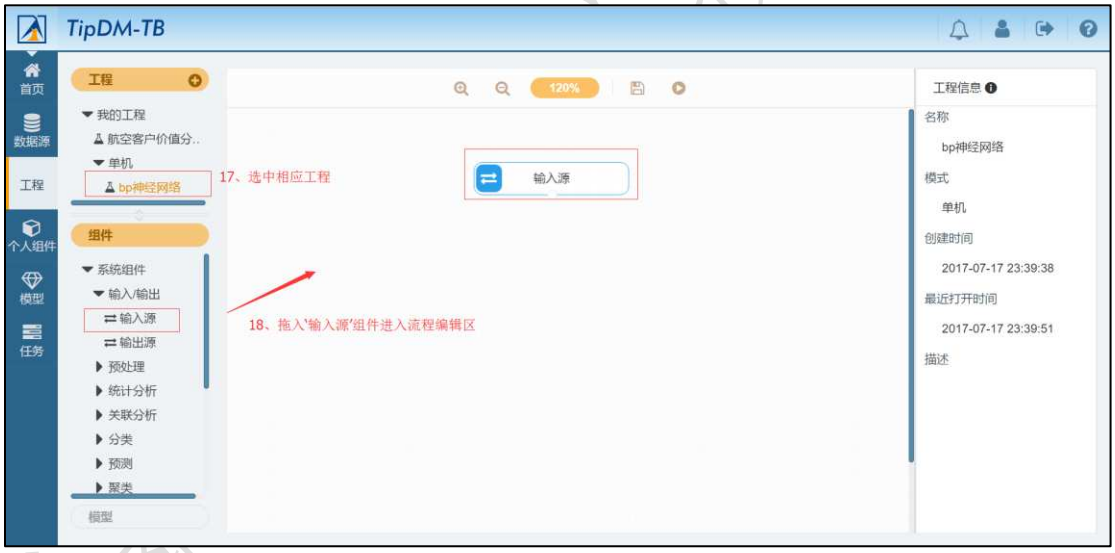


图 8

■ 设置“输入源”参数，如图 9 所示。

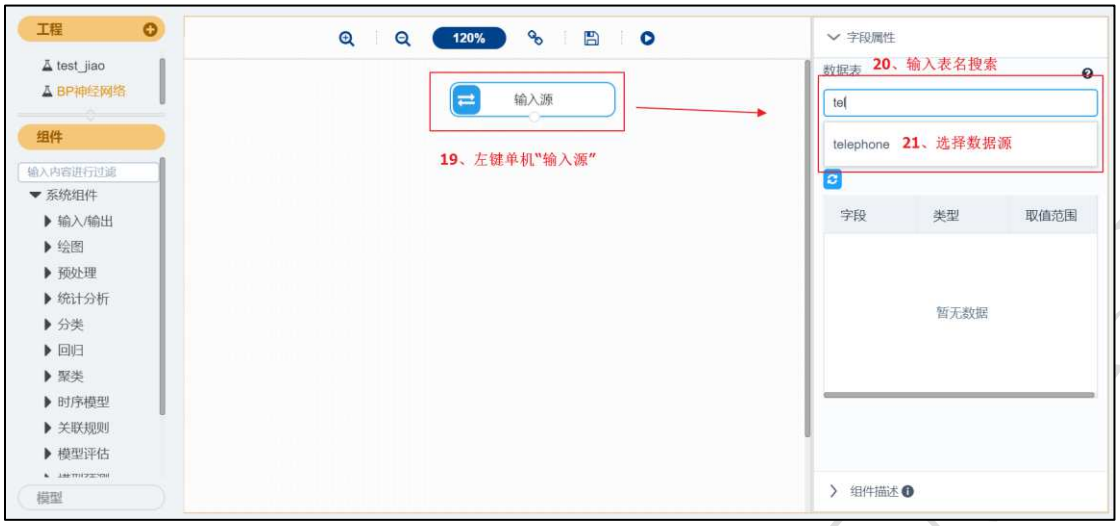


图 9

■ 拖入数据“数据拆分”组件，如图 10 所示。

图 10

■ 设置“数据拆分”组件的参数，如图 11、图 12 所示。

图 11

图 12

2.1.3 模型构建

■ 拖入“BP 神经网络”组件并设置相应参数，如图 13、图 14、图 15 所示。

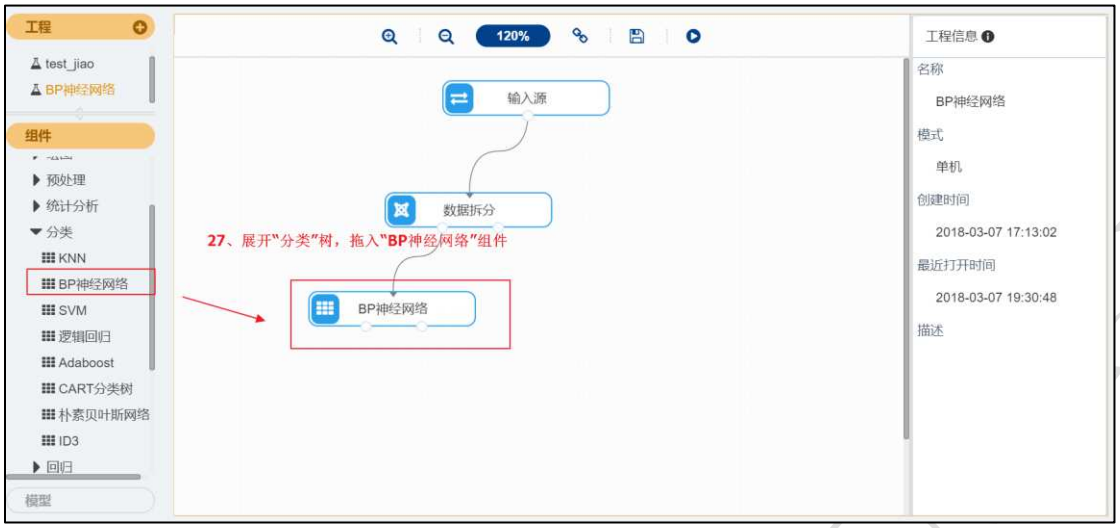

工程
test_jiao
BP神经网络
组件
预处理
统计分析
分类
KNN
BP神经网络
SVM
逻辑回归
Adaboost
CART分类树
朴素贝叶斯网络
ID3
回归
模型
120%
输入源
数据拆分
27、展开"分类"树，拖入"BP神经网络"组件
BP神经网络
工程信息
名称
BP神经网络
模式
单机
创建时间
2018-03-07 17:13:02
最近打开时间
2018-03-07 19:30:48
描述

图 13


120%
输入源
数据拆分
BP神经网络
28、左键单机组件
字段属性
添加字段过滤字符串
29、勾选特征列
字段
类型
wifi
数值
0-51
e_pay
数值
0-1
package_type
数值
1-4
leave
字符
0,1
标签列
30、勾选标签
leave
参数设置
组件描述

图 14

图 15

■ 运行，如图 16、图 17 所示。

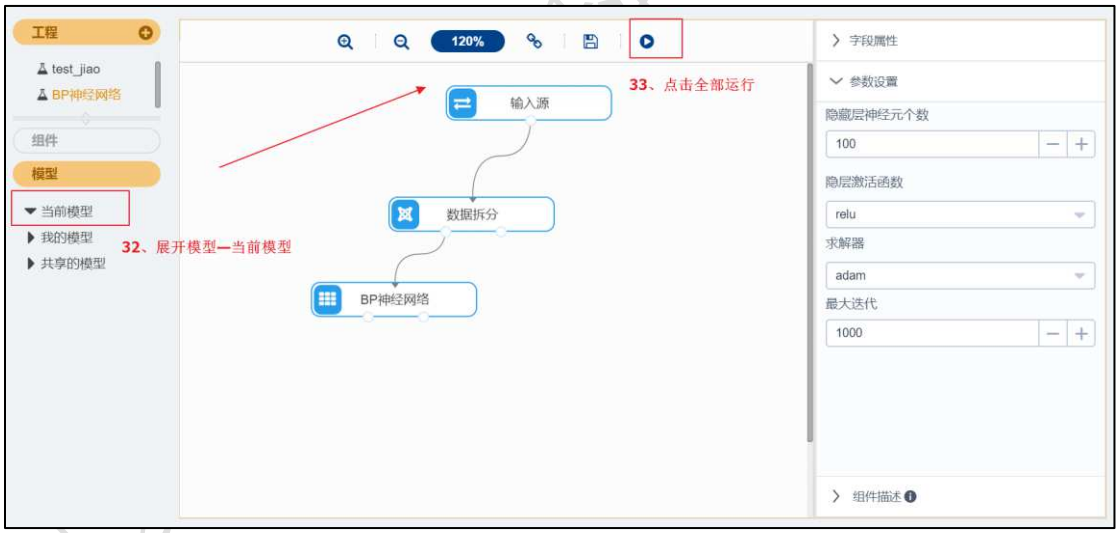


图 16

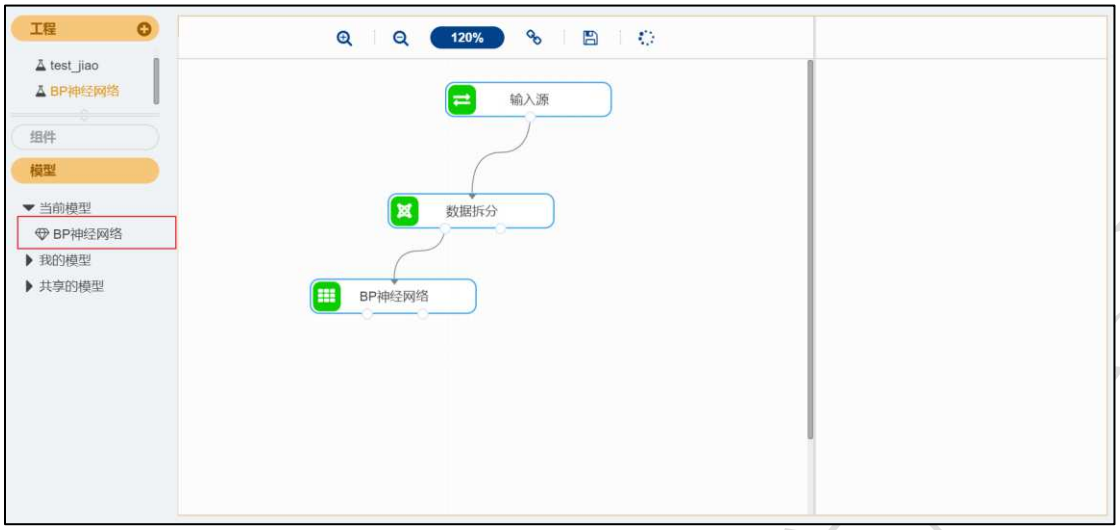


图 17

2.1.4 模型评估

■ 拖入“模型评估”组件，如图 18 所示。

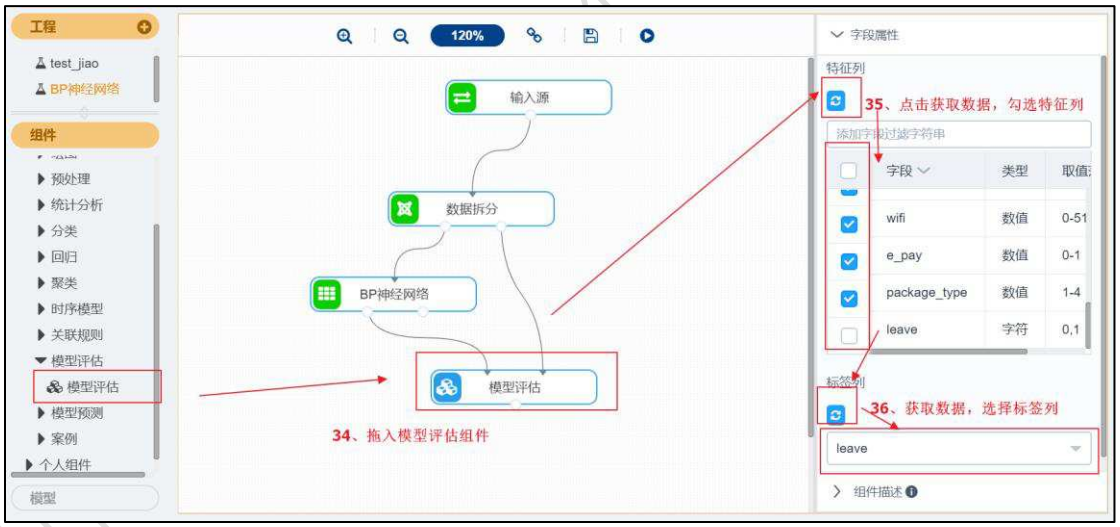


图 18

2.1.5 模型预测

■ 拖入“模型预测”组件，如图 19、图 20 所示。

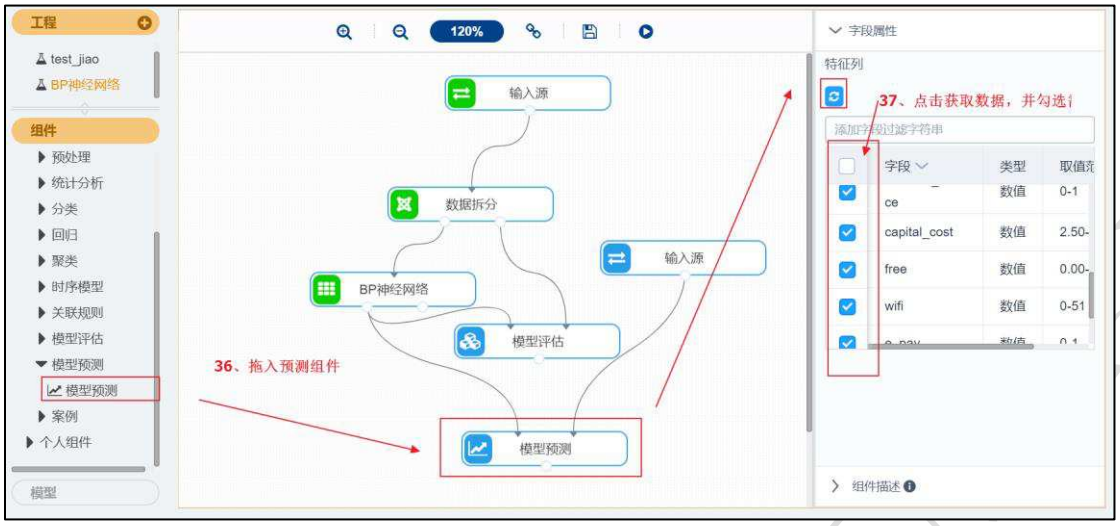


图 19

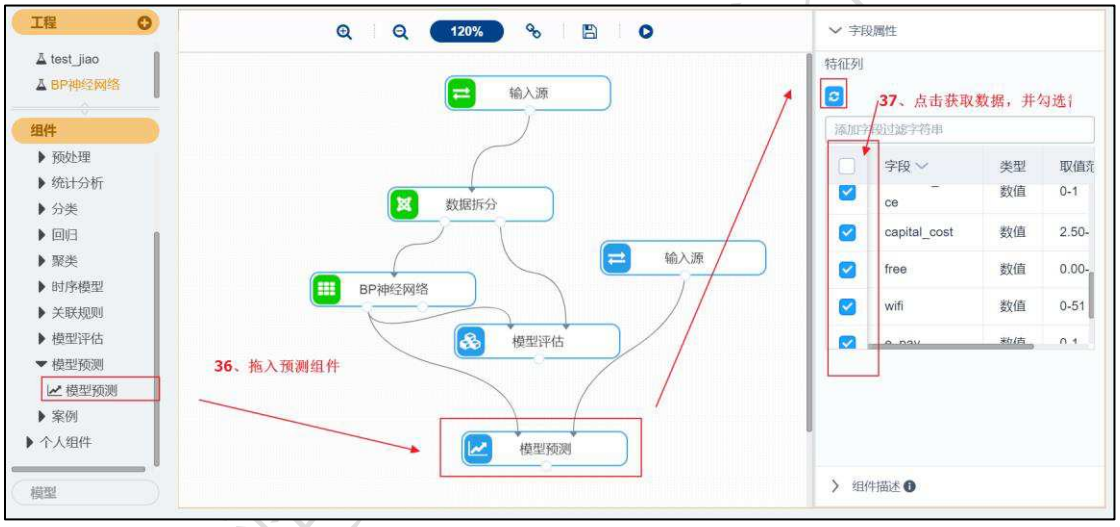


图 20

2.2 使用模板入门

选择一个模板，单击左键，弹出复制模板的对话框，填入工程名称，如图 21 所示。即可直接使用该工程用于学习，如图 21、图 22 所示。

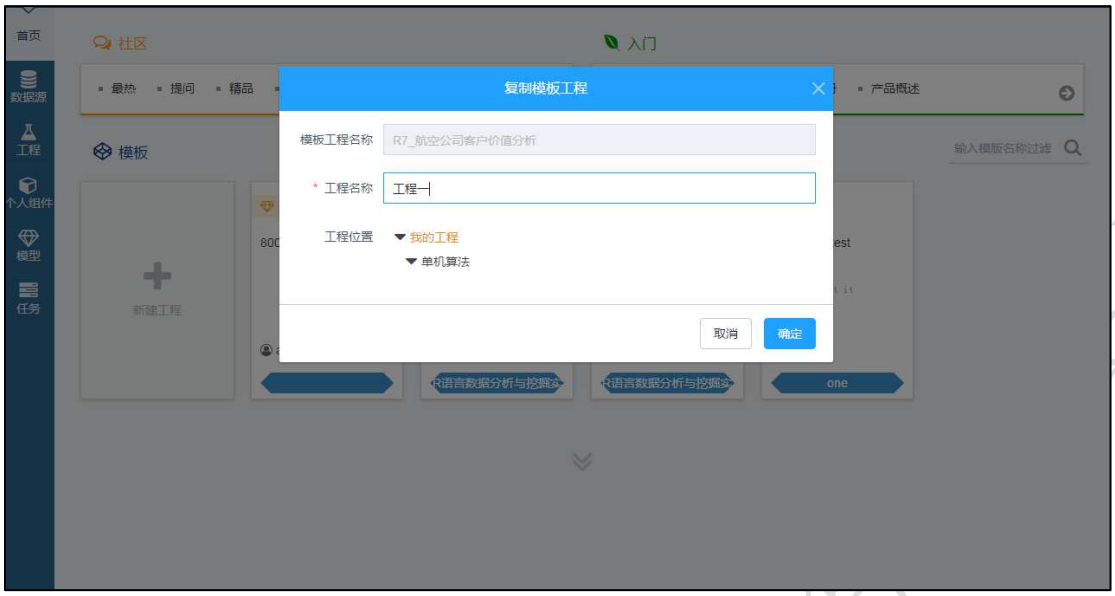


图 21

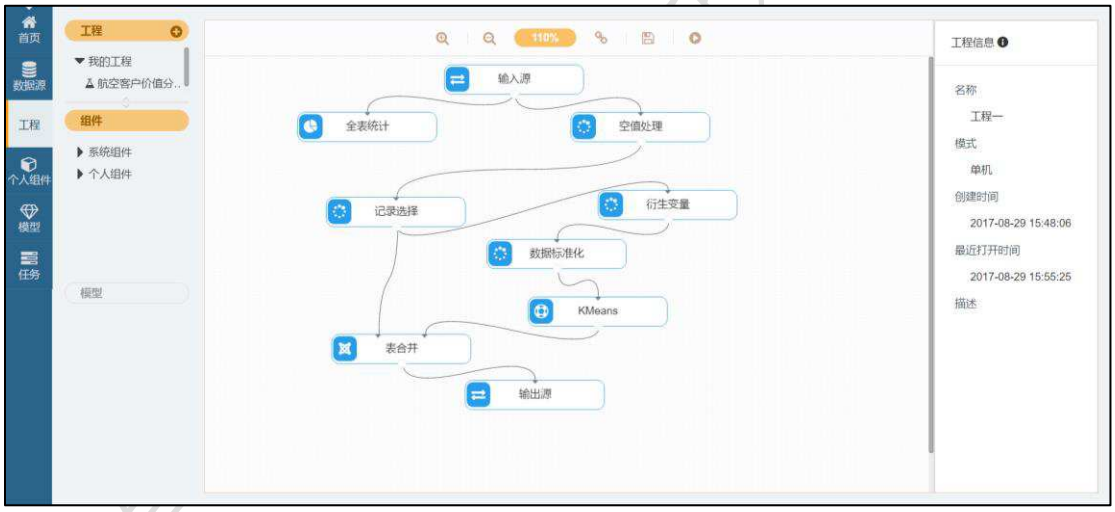


图 22

3 功能说明

3.1 首页

首页主要包括社区、入门、模板三个板块。如图 23 所示。

社区包括最热、提问、精品、话题、笔记、学习资源等链接。

入门包括常见问题、快速入门、操作手册、产品概述等链接。

模板包括已经建立好的工程流程，用户可直接复制工程模板来学习建模平台的使用。

图 23

3.2 数据源

3.2.1 我的数据源

新增数据源有两种方式，分别是：数据来源于文件，数据来源于数据库。如图 24 所示。

3.2.1.1 数据来源于文件

描述：数据来源于文件，支持上传本地 CSV 文件，用户可以自定义列名和数据类型，列名支持小写字母、数字、下划线构成。

操作：

■ 点击新建数据源的下拉菜单，选择数据来源于文件。如图 24 所示。

图 24

■ 在弹出的新增数据源对话框中，点击【数据来源于文件】，选择一个本地 csv 文件，填写表名，如图 25 所示。

图 25

■ 填写完毕后，点击下一步即可预览数据，如图 26 所示。

图 26

■ 预览数据无乱码情况，可以点击下一步，即可修改字段。如果中文数据发生乱码，可返回上一步调整文件编码。如图 27 所示。

图 27

点击【确定】，可在数据源列表中看到新添加的数据源，如图 28 所示。

图 28

3.2.1.2 数据来源于数据库

用户可以使用系统中已存在的数据库与数据来进行数据输入。

本平台支持的数据库有：MySQL、Oracle、SQL Server 、PostgreSQL、 DB2。

下面以 MySQL 为例，进行操作。

■ 点击新建数据源的下拉菜单，选择【数据来源于数据库】。如图 29 所示。

图 29

■ 打开新建数据源对话框，根据实际情况填入信息，如图 30、图 31 所示。

表名：用户自定义数据源的名称。

数据库类型：选择需要连接的数据库。

URL 链接：根据显示的默认链接语句填写链接的数据库的 ip，端口号，数据库名。

用户名：访问数据库的用户名。

密码：访问数据库的密码。

SQL: 输入与执行 sql 语句，执行结果可以在预览结果框中查看。

图 30

图 31

点击【测试连接】，测试成功则在平台右上角弹出提示框提示“测试成功”。此时可以点击下一步预览数据。如图 32 所示。

图 32

预览成功，可以点击下一步进行字段设置，如图 33 所示。

图 33

添加了数据源后，必须执行同步（同步后的数据源才能在流程中使用），如图 34 所示。

图 34

同步成功后，可以看到相应的提示信息，如图 35 所示

图 35

3.2.1.3 分享数据

上传成功后的数据可以分享给其他用户，如图 36 所示，可分享到哪些用户需要系统管理员事先分配可查看的用户，如图 37 所示。

图 36

图 37

3.2.2 共享数据源

共享数据源是其他用户分享给当前用户的数据源，可查看数据，并在工程中使用，如图 38 所示。

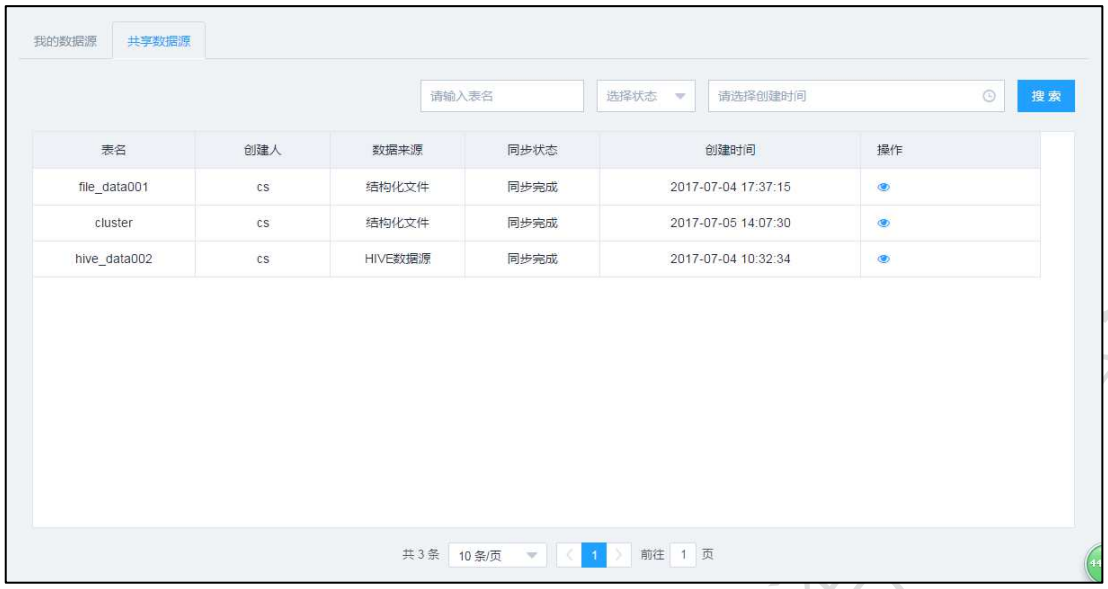


图 38

3.3 工程

工程界面主要包括：工具栏、工程区、组件区、模型区、设计区。如图 39 所示。

工具栏：包括放大/缩小画布功能、保存工程、运行/停止工程。

工程区：用于显示所新建的工程文件或工程。

组件区：包括数据源组件、 数据预处理组件、统计分析组件、机器学习组件、模型评估组件等。

模型区：包括我的模型，保存的模型，分享的模型。

设计区：用于设计工程流程，在运行后能够查看步骤的运行结果，包括模型输出的参数。

图 39

3.3.1 新建工程

新建工程有三种操作方法，主要包括：从首页新建工程、从工程界面点击“+”号新建工程、从工程界面的“我的工程”右键新建工程。

3.3.1.1 首页——新建工程

从首页新建工程如图 40 所示。

图 40

3.3.1.2 工程——点击+号（新建工程）

从工程界面点击“+”号新建工程如图 41 所示。

图 41

3.3.1.3 工程——选择我的工程，单击右键——新建工程

从工程界面的“我的工程”右键新建工程如图 42 所示。

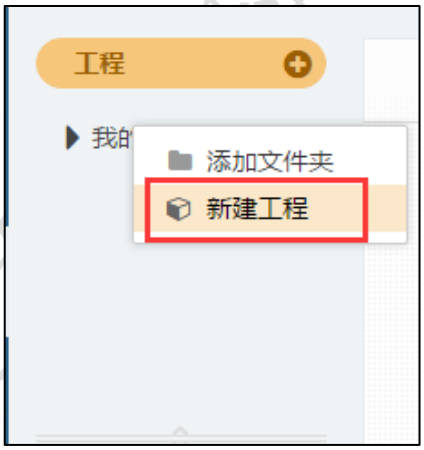


图 42

3.3.2 我的工程

单击我的工程，右键选项包括：添加文件夹、新建工程。如图 43 所示。

图 43

添加文件夹：在我的工程根目录下，新建一个文件夹。

新建工程：新建一个工程，默认目录在根目录下。

3.3.2.1 文件夹

选择工程目录下的某个文件夹，右键选项包括：删除、添加文件夹、导入工程、新建工程、上移（排序）、下移（排序）。如图 44 所示。

图 44

删除：删除文件夹。

添加文件夹：默认在该文件夹目录下新增文件夹。

导入工程：从外部导入工程。

新建工程：默认在该文件夹目录下新增工程，目录可选。

上移：向上移动。

下移：向下移动。

3.3.2.2 工程

选择一个工程，右键选项包括：删除、修改工程信息、另存为工程、另存为模板、创建历史版本、恢复到指定版本、删除历史版本、上移（排序）、下移（排序）。如图 45 所示。

图 45

删除：删除工程。

修改工程信息：可以修改工当前程的描述。

另存为工程：将现有工程另存为新的工程。

另存为模板：可将当前工程另存为模板，在首页展示，相当于共享当前工程给其它用户使用。

创建历史版本：将当前工程定期保存一个版本。

恢复到指定版本：可将工程恢复到保存好的任何一个版本。

删除历史版本：将该工程创建的历史版本删除。

上移：向上移动一个工程。

下移：向下移动一个工程。

3.3.2.3 组件

将组件拖入到设计区，右键单击组件，选项包括：重命名、删除、全部运行、运行到此处、从此节点运行、查看数据、可视化、查看日志、查看源码。如图 46 所示。

图 46

重命名：对该组件重新命名。

删除：删除组件。

全部运行：全部运行流程。

运行到此处：从输入源运行到选择的组件。

运行该节点：只运行该节点。

从此节点运行：从选择的组件开始运行到最后一个组件。

查看数据：查看运行后的数据。

查看日志：可查看该组件的运行日志。

查看源码：根据权限，查看该组件的算法源码。

3.3.2.4 模型

将分类与回归算法生成的模型展示在该界面，如图 47 所示。

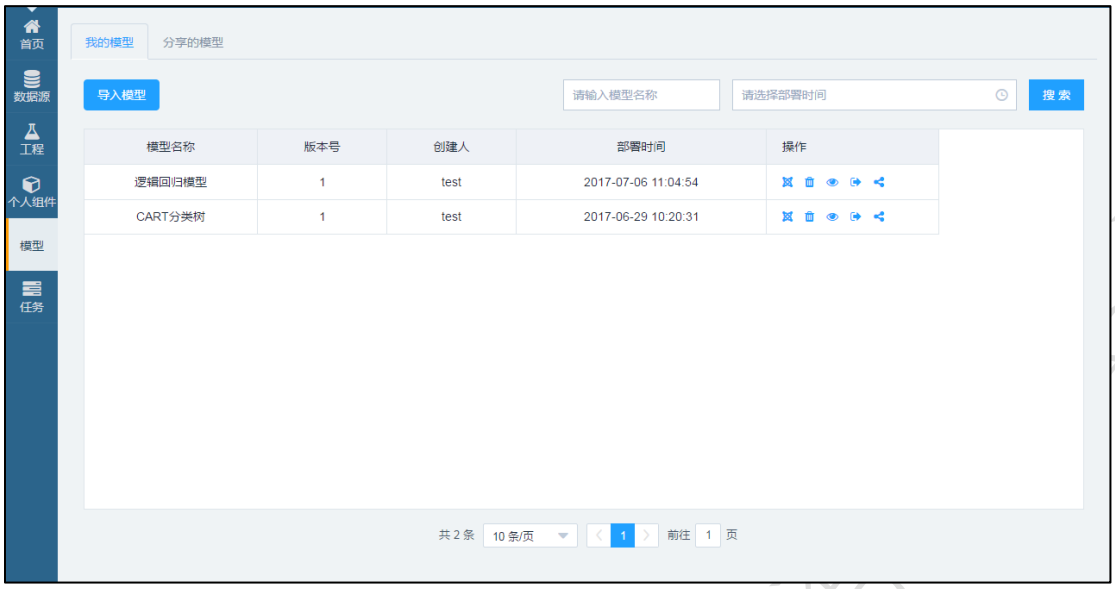


图 47

1 模型生成

“系统组件”栏中分类和回归算法运行后，自动会生成相应的模型，并存在我的模型列表中，如图 48 所示。需用户对生成的模型右键，选择“部署模型”，才能保存到模型管理列表中，如图 49 所示。

模型名的规范为：组件名，可通过重命名组件修改模型名。

平台中每个分类或回归算法节点运行后仅对应唯一一个模型，再次运行节点时，会将原有生产的覆盖。

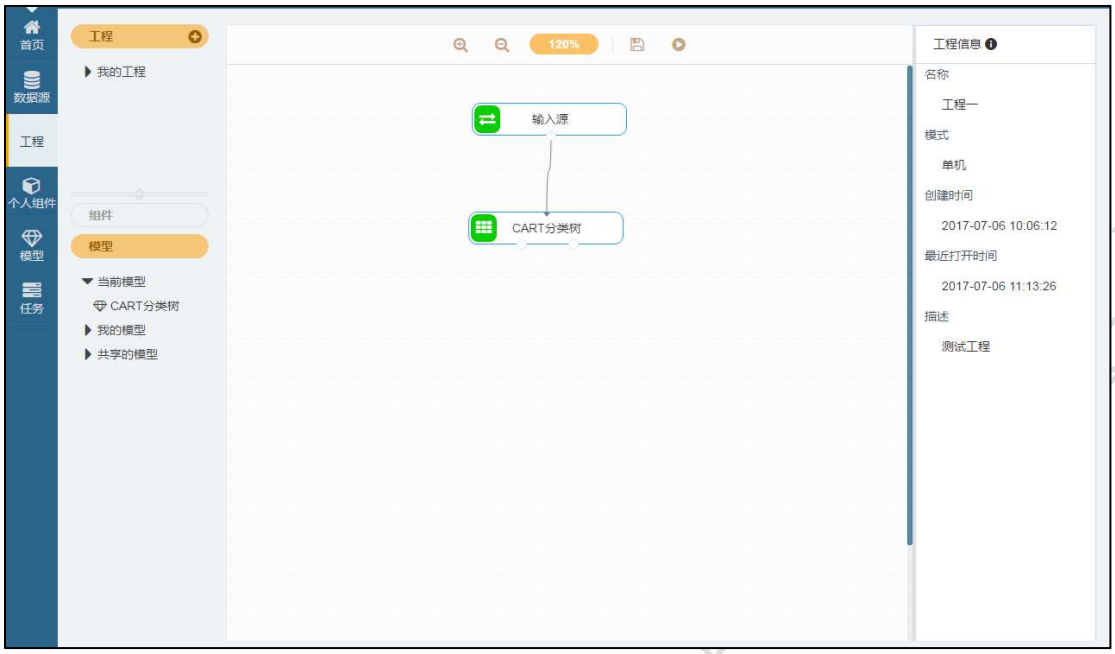


图 48

图 49

2 模型使用

在模型栏中，直接将模型拖拽到画布中就可以使用了，如图 50 所示。

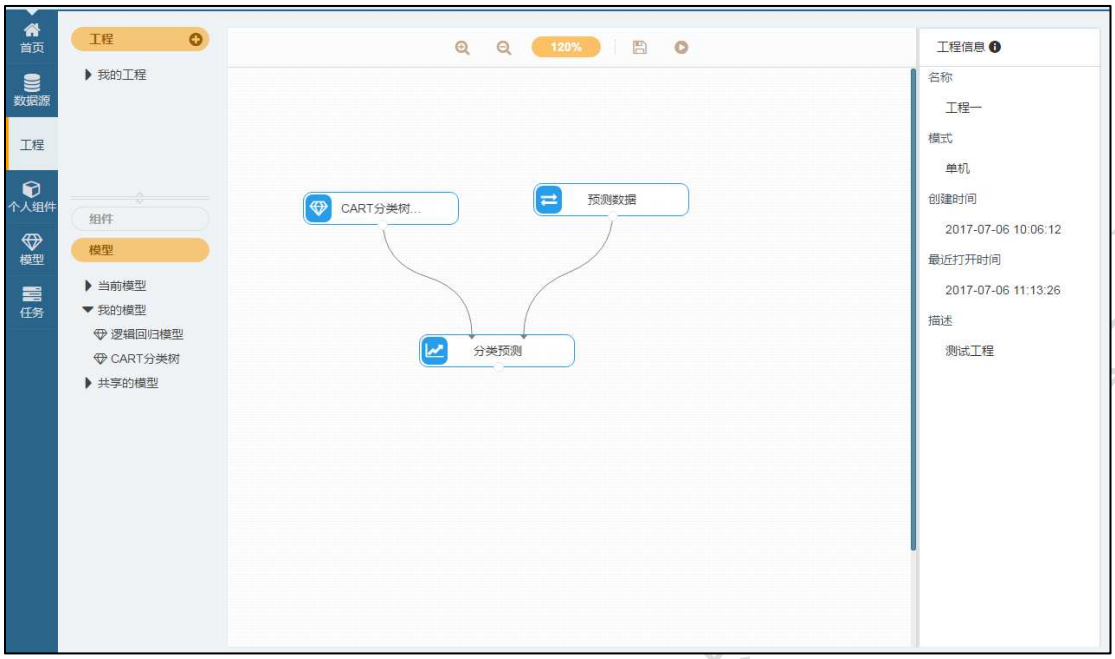


图 50

3.4 系统组件

3.4.1 输入/输出

3.4.1.1 输入源

图标： 

描述：读取表数据组件。当输入表名后，会自动读取表的结构数据，在字段信息中可查看。当数据源是来源于数据库时，表字段修改后，如增加或删除某个字段，在工程内是无法时时更新的，需要用户重新同步该数据。

字段属性

字段属性包括：数据表、字段信息。如图 51 所示。

数据表：输入数据源的表名。

字段信息：查看各字段名称、数据类型及取值范围。

图 51

3.4.1.2 输出源

图标： ⇄ 输出源

描述：将数据表中的数据导出到指定的数据库中。

字段属性

输出字段：勾选要输出的字段。

参数设置

参数设置包括目标表、插入前清空目标表、URL 链接、用户名、密码。如图 52 所示。

目标表：自定义输出数据表的名称

插入前清空表数据：选择是，即导出一份新的数据；选择否，即导出的数据紧接着该表中原有的数据存储。

url 链接格式：

jdbc:mysql://<machine_name>:<port>/<dbname>

jdbc:oracle:thin:@<machine_name>:<port>:<dbname>

jdbc:db2://<machine_name>:<port>/<dbname>

jdbc:postgresql://<hostname>:<port>/<dbname>

jdbc:hive2://<machine_name>:<port>

用户名：数据库用户名。

密码：数据库用户名密码。

图 52

3.4.2 预处理

3.4.2.1 缺失值处理

图标： 

描述：缺失值处理是数据预处理的一部分，由于采集的数据存在一些属性值的缺省，如果不做处理，将直接影响后续算法的挖掘效果，严重时甚至得到错误的结果。处理方法有删除缺失值、中位数插补、众数插补、均值插补、线性插值、多项式插值。

字段属性

特征列：必选。选用中位数插值法、众数插值法、均值插值法时，请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。勾选的列将传入下一个组件。

参数设置

处理方式：选择对该缺失数据列的处理方式，可以选择删除缺失值、中位数插补、众数插补、均值插补、线性插值、多项式插值。

输出

表结果：缺失值处理结果。

报告：无。

示例

下面对某数据进行缺失值处理。

- 勾选需要进行缺失值处理的数据，将会在勾选的数据内查找缺失值，并进行相应的处理。如图 53、**错误!未找到引用源。**所示。
- 选择处理方式为【删除法】，如图 54 所示。
- 运行该组件后可通过查看数据查看结果。

图 53

图 54

3.4.2.2 记录选择

图标： 记录选择

描述：记录选择是对数据表的行进行筛选，只留下满足条件的数据行。

字段属性

特征列：选择需要进行记录选择的列，勾选的列将传入下一个组件。如图 55 所示。

图 55

参数设置

参数设置包括：过滤器的增加和删除、刷新列、运算符、过滤列、过滤条件、过滤值，如图 56 所示。

添加（+）和删除 ：通过点击添加按钮添加一个列的过滤设置，通过删除图标减少一个列的过滤设置

刷新列：想要获取数据，则必须事先通过点击刷新按钮

运算符：提供各条件之间 and 和 or 的选择，

过滤列：获取上级操作单元节点的列信息，供用户选择(单选)

过滤条件：目前操作符支持“=”，“!=”，“>”，“<”， “>=”和“<=”

过滤值：条件值选择，当过滤值为字符类型时，需添加双引号

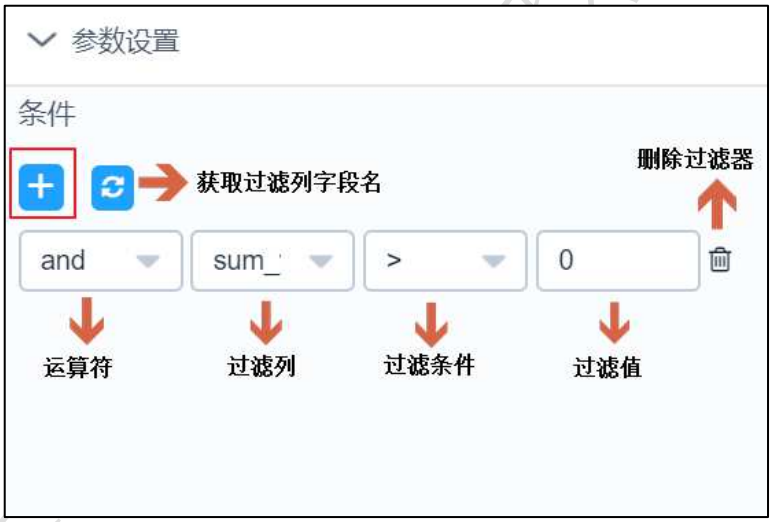


图 56

输出

表结果：记录选择结果。

报告：无。

示例

下面对某数据进行记录选择，数据一共包括四个字段：id、r、f、m。选择满足 r>27 的所有数据。

- 勾选需要进行记录选择的特征列，如图 57 所示。

- 依次点击【加号】及【刷新】按钮，选择字段、运算符、值。如图 58 所示。
- 运行该组件，对组件右击，选择查看数据，结果如图 59 所示。

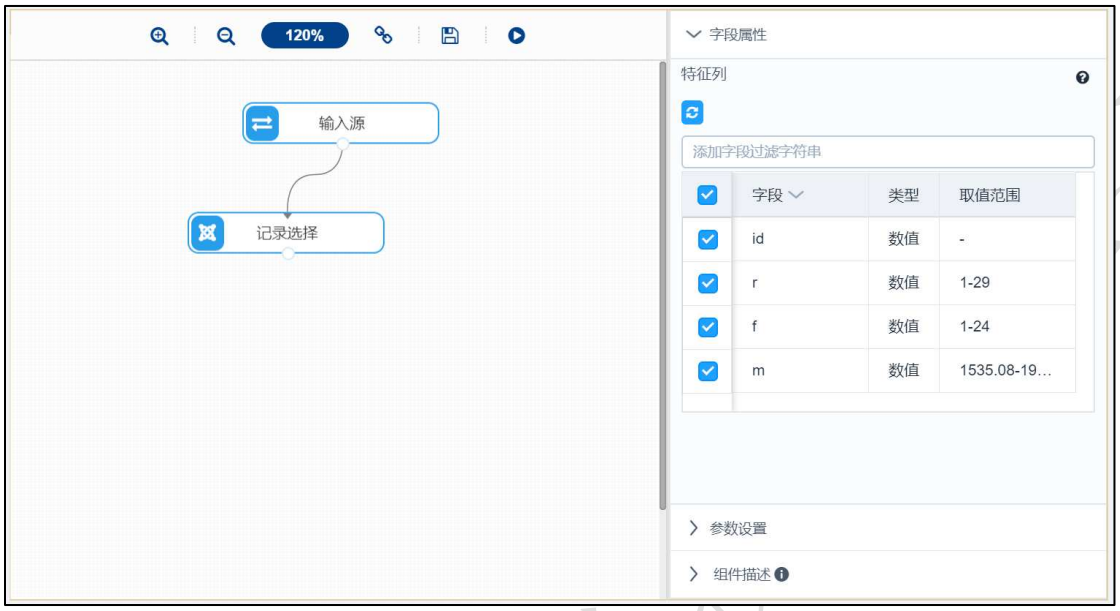


图 57

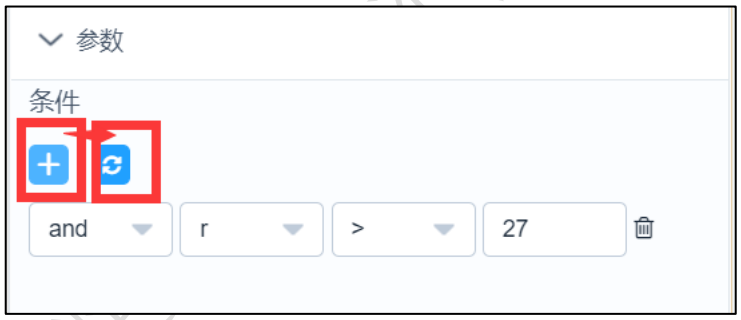


图 58


预览数据（仅显示前100条）

id	f	m	r
14	16	1957.44	30
17	2	1016.34	93
26	21	1628.68	30
30	7	5318.81	60
52	8	1865.99	30
53	8	1791.44	28
58	4	2920.81	66
77	11	1461.63	78
81	2	227.14	28

图 59

3.4.2.3 数学类函数

图标： 

描述： 数学类函数是对勾选的某列运用数学运算函数。

字段属性

特征列：必须包含待运算的列，勾选的列将传入下一个组件。

参数设置

待运算的列：请选择数值型数据。

运算函数：包括向上取整、绝对值、向下取整、平方根、返回整数。

输出

表结果：运算结果。

报告：无。

示例

下面对某数据的一列进行向上取整，原数据如图 60 所示。

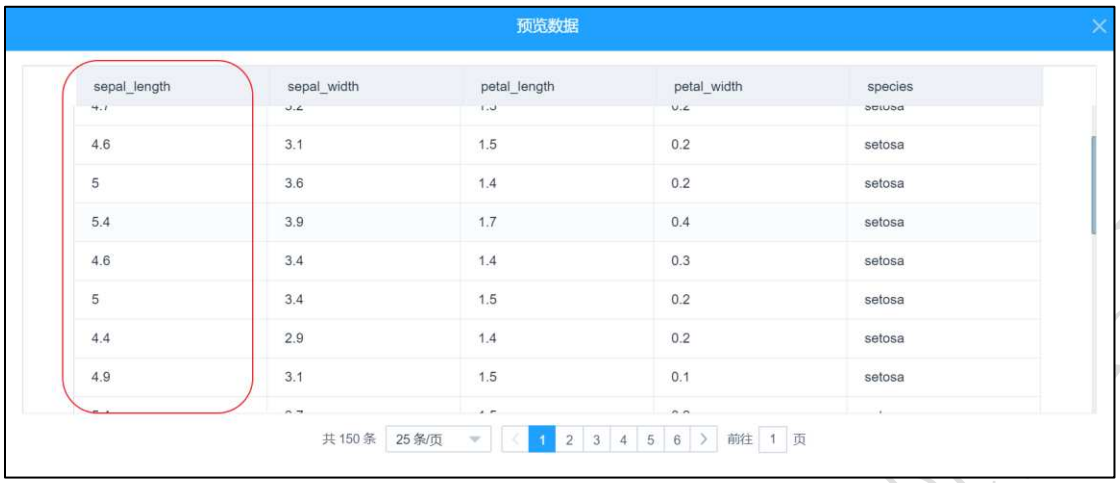


图 60

- 在特征列中勾选需要传入下个组件的数据，必须包含待运算的列。如图 61 所示。

图 61

- 打开参数设置选项卡，在【运算函数】下拉框中选择【向上取整】，如图 62 所示。

图 62

- 运行成功后,点击查看数据,结果如图 63 所示.

图 63

3.4.2.4 表合并

图标： 

描述：表合并是指两张表通过行或列合并成一张表，不需要关键字段。

字段属性

左表特征列：勾选左表需要合并的列，需要注意的是左表特征列与右表特征列列名不能重合。

右表特征列：勾选右表需要合并的列，需要注意的是左表特征列与右表特征列列名不能重合。如图 64 所示。

图 64

参数设置

合并方式：必选。可选择行合并，列合并。需要注意的是，选择列合并时，两个表的行数需要一样；选择行合并时，两个表的列数需要一样，如图 65 所示。

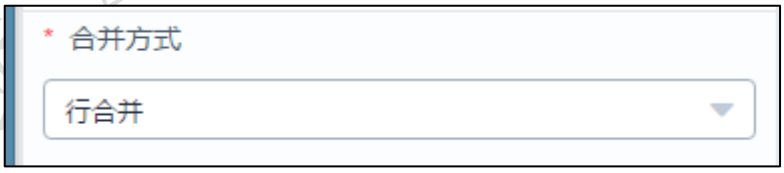


图 65

输出

表结果：合并结果。

报告：无。

示例

下面按列将两个表合并为一个表。

- 勾选需要进行合并的字段，左表选择 f，m 两列，右表选择 r 列，如图 66 所示。
- 合并方式选择列合并，如图 67 所示。
- 运行该组件，右击选择查看数据，结果如图 68 所示。

图 66

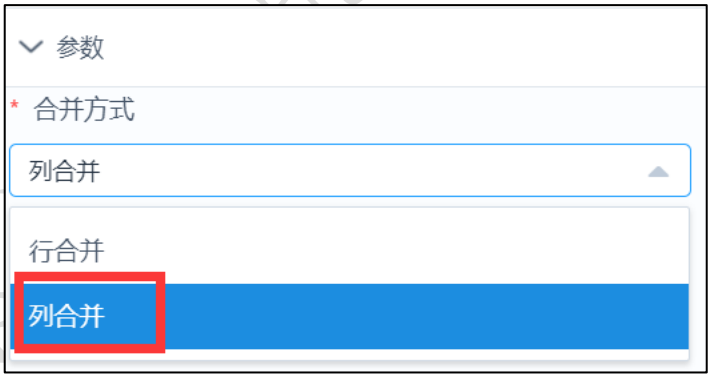


图 67

f	m	r
6	232.61	27
5	1507.11	3
16	817.62	4
11	232.81	3
7	1913.05	14
6	220.07	19
2	615.83	5

图 68

3.4.2.5 表连接

图标： 

描述： 表连接是指两张表通过某列进行关联，合成一张表。

字段属性

左表：必选。选择左表需要关联的列，必须包含左表连接关键字列。

右表：必选。选择右表需要关联的列，必须包含右表连接关键字列，如图 69 所示。

图 69

参数设置

左表连接关键字：必选。选择和右表关联的列。

右表连接关键字：必选。选择和左表关联的列。

连接方式：必选。支持左外连接，内连接，右外连接，全外连接。其中左外连接返回左表中所有的记录以及右表中连接字段相等的记录；右外连接返回右表中所有的记录以及左表中连接字段相等的记录；内连接返回两个表中连接字段相等的记录；全外连接返回两个表中的记录，如图 70、图 71 所示。

图 70

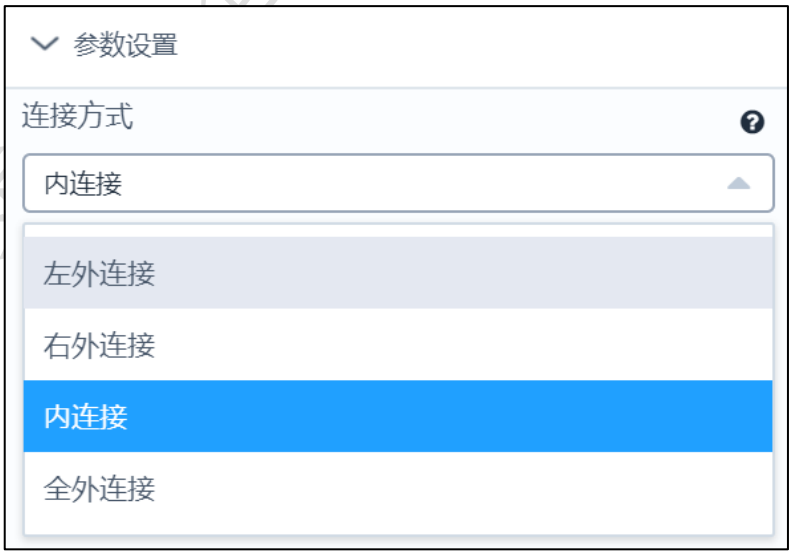


图 71

输出

表结果：表连接结果。

报告：无。

示例

下面对两个数据表进行表连接。

- 勾选需要进行连接的字段，左表及右表的特征列必须包含连接需要用到的连接关键字。该数据关键列为 id，左边选择 id，m 两列，右表选择 id，f 两列。如图 72 所示。
- 勾选左表的关键列为 id，右表的关键列为 id，使用全外连接。如图 73 所示。
- 运行该组件，右击选择查看数据，结果如图 74 所示。

图 72

图 73

id	m	f
1	232.61	6
2	1507.11	5
3	817.62	16
4	232.81	11
5	1913.05	7
6	220.07	6
7	615.83	2

图 74

3.4.2.6 平稳性检验

图标： 平稳性检验

描述：平稳性检验是为了确定序列是否存在确定趋势，否则将会产生“伪回归”问题。伪回归是说，有时数据的高度相关仅仅是因为二者同时随时间有向上或向下的变动趋势，并没有真正联系。这样数据中的趋势项，季节项等无法消除，从而在残差分析中无法准确进行分析。

字段属性

时序列：必选。选择想要进行检验的数据列，请选择数值型数据，如果该列数据含有缺失值，则会自动删除，如图 75 所示。

图 75

参数设置

无

输出

表结果：无。

报告：Test statistic、p-value、Number of lags used、Number of observations used for the ADF regression and calculation of the critical values、Critical values for the test statistic at the 5 %、Critical values for the test statistic at the 1 %、Critical values for the test statistic at the 10 %、自相关图。

示例

下面对某列数据进行平稳性检验。

- 选择时序列，数据必须为数值型。如图 76 所示。

- 运行该组件，对组件右击，选择查看报告，结果如图 77 所示。

图 76

图 77

3.4.2.7 纯随机性检验

图标： 随机性检验

描述：纯随机性检验又称为白噪声检验，是专门用来检验序列是否为纯随机序列的一种方法。纯随机序列的序列值之间没有任何相关关系，也就是没有什么统计规律可言，各项之间也就

没有任何关联，这样的序列没有挖掘的意义。

字段属性

字段属性包括：字段信息、待检验序列，如图 78 所示。

待检验序列：必选。选择想要进行检验的列，请数值型数据。

图 78

输出

表结果：对应各滞后阶数的 P 值。

报告：无。

示例

下面对某列数据进行纯随机性检验。

- 选择待检验序列，数据必须为数值型。如图 79 所示。
- 运行成功后，选择查看数据，如图 80 所示。

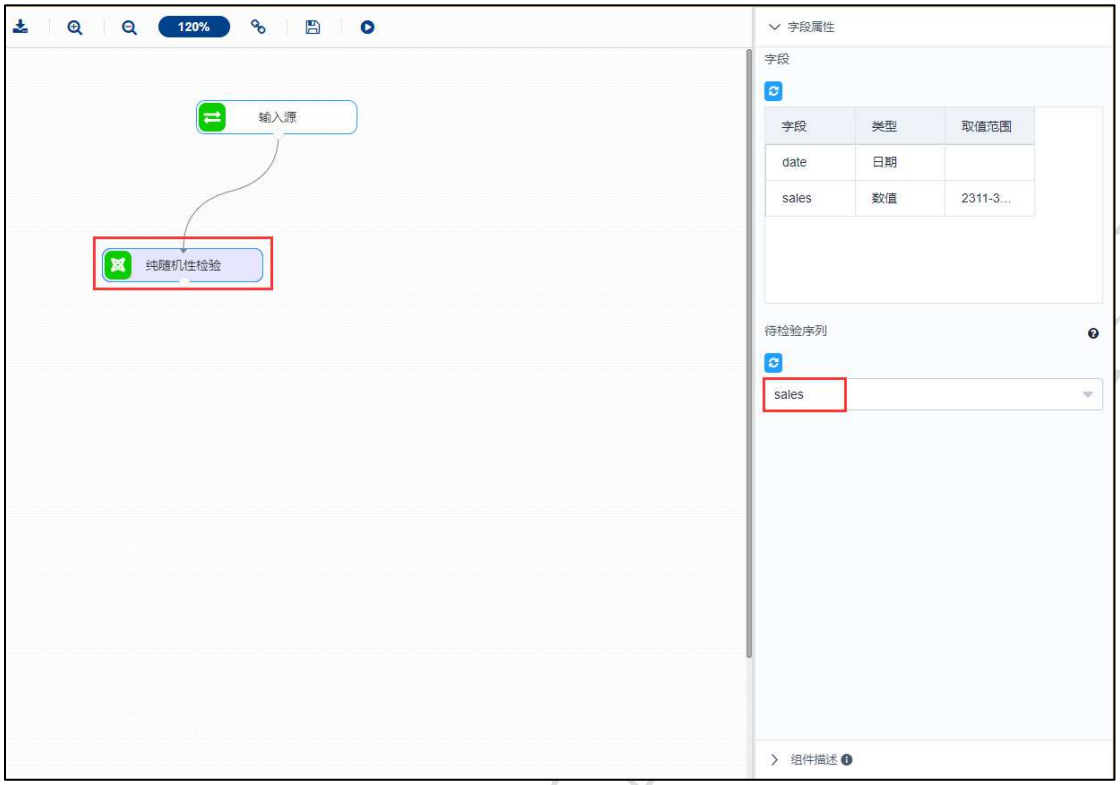


图 79

预览数据

lags	pvalue
1	0.05664850864089642
2	0.08171417014241165
3	0.11638329840952293
4	0.2064018913330824
5	0.3073699633365428
6	0.3672622868506844
7	0.4476144801344415
8	0.27237256179528074

共 40 条 25 条/页 1 2 前往 1 页

图 80

3.4.2.8 记录去重

图标： 

描述：记录去重是去除数据表中的重复的行数据，只保留其中一行数据。

字段属性

特征列：必选。选择需要进行去重的列，勾选的字段可传入下个组件，如图 81 所示。

图 81

参数设置

无

输出

表结果：去重后的数据。

报告：无。

示例

- 勾选需要去重的数据，勾选的列将传入下个组件。如图 82 所示。
- 运行成功后，可右键查看数据，结果如图 83 所示。

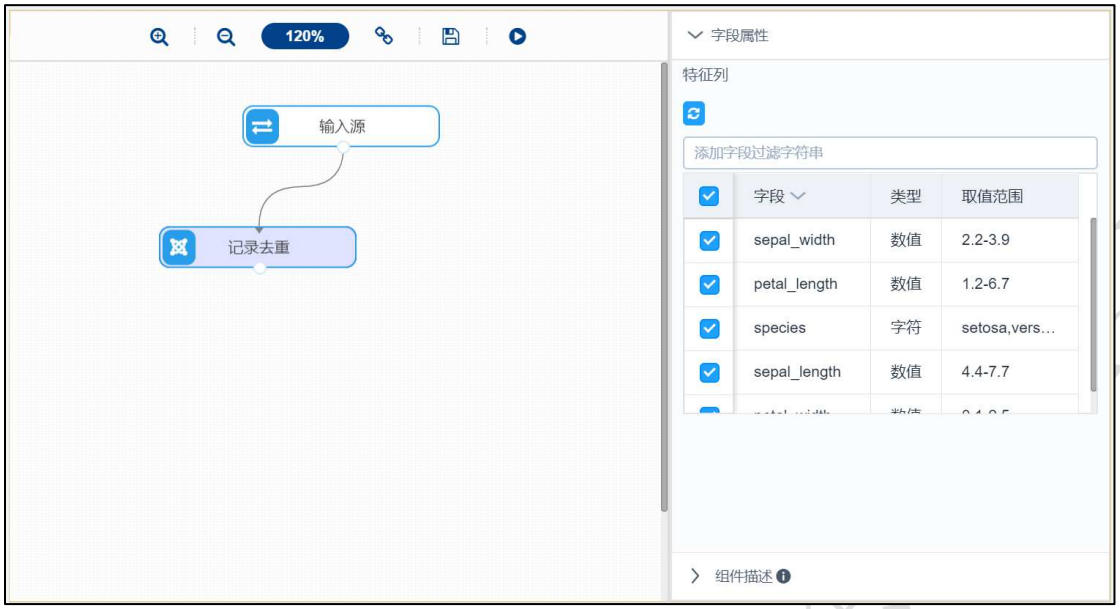


图 82

图 83

3.4.2.9 数据离散化

图标： 数据离散化

描述：某些模型算法，特别是某些分类算法如 ID3 决策树算法和 Apriori 算法等，要求数据是离散的，此时就需要将连续型特征（数值型）变换成离散型特征（类别型），即连续特征离散化。常用的离散化方法主要有三种：等宽法，等频法和通过聚类分析离散化（一维）。

字段属性

待离散化数据：必选。请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。勾选多列时，自动对每一列数据进行离散化如图 84 所示。

图 84

参数设置

离散化方式：选取要使用的离散方式，支持等宽、等频、聚类离散化，默认等宽。

离散个数：离散的个数，默认 2，如图 85 所示。

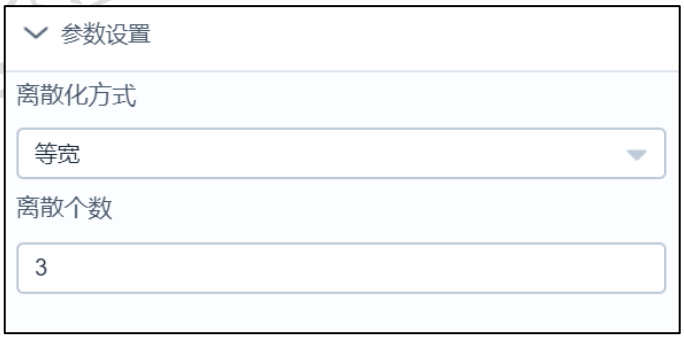


图 85

输出

表结果：对勾选的每一列进行离散化后的结果。

报告：无。

示例

下面对数据进行离散化。原数据如图 86 所示。

图 86

- 勾选需要进行离散化的数据，如图 87 所示。
- 选择离散化方式为等宽，离散个数为 2。如图 88 所示。
- 运行该组件，右击选择查看数据，如图 89 所示。

图 87

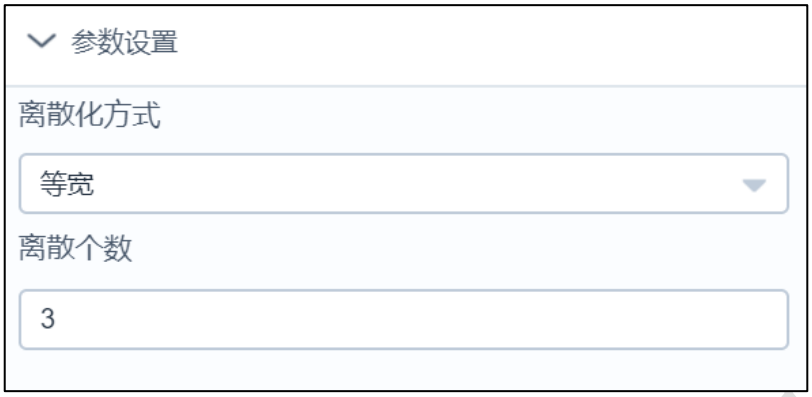


图 88

图 89

3.4.2.10 排序

图标： 

描述： 根据某一列的顺序将所有数据重新排序。

字段属性

特征列：勾选的列必须包含关键字段。勾选的列将传入下一个组件。如图 90 所示。

关键字：必选，由该列值的顺序将待排序的数据重新排序。如图 91 所示。

图 90

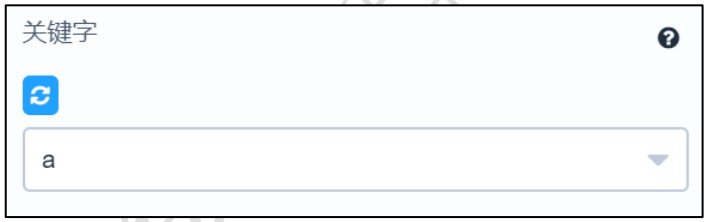


图 91

参数设置

排序方式：将数据按某列的顺序将所有数据按照升序或降序重新排序，如图 92 所示。

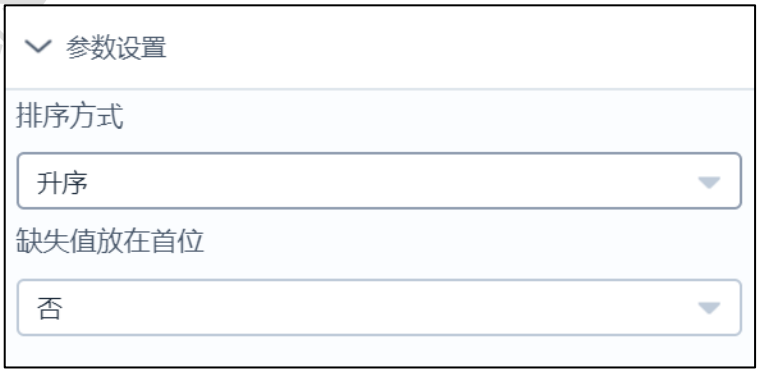


图 92

输出

表结果：按照某列排序后的数据。

报告：无。

示例

下面将按列 m 的值对数据进行升序排序。

- 勾选需要进行排序的特征列，选择关键列 m，特征列必须包含关键列。如图 93 所示。
- 选择升序。如图 94 所示。
- 运行该组件，结果如图 95 所示。

图 93

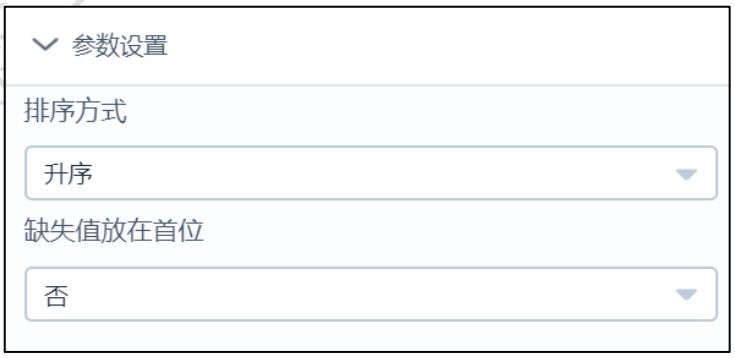


图 94

图 95

3.4.2.11 数据拆分

图标： 

描述：数据拆分对全量数据进行简单随机抽样，将数据拆分为训练数据和测试数据。

字段属性

特征列：必选。选择进行拆分的数据，勾选的列将传入下个组件，如图 96 所示。

图 96

参数设置

随机种子：可以理解为一个序号，这个序号交给一个数列管理器，通过这个序号，从管理器中取出一个数列，这个数列就是通过那个序号得到的随机数。

训练集比例：设置训练数据的比例，范围在 0-1 之间，默认 0.25，如图 97 所示。

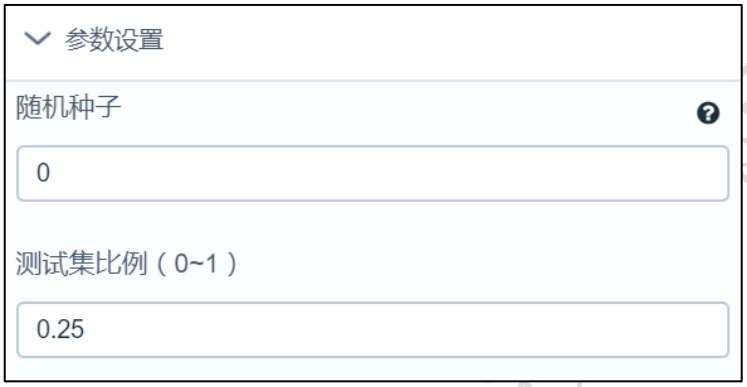


图 97

输出

表结果：训练集与测试集。

报告：无。

示例

下面将某数据拆分为训练集、测试集。

- 勾选需要进行数据拆分的数据，如图 98 所示。
- 保留默认的随机种子，设置测试集比例为 0.25，如图 99 所示。
- 运行该组件，选择查看数据，查看对应的数据集，如图 100 所示。


120%
输入源
数据拆分
字段属性
特征列
添加字段过滤字符串
字段
类型
取值范围
sepal_width
数值
2.2-3.9
petal_length
数值
1.2-6.7
species
字符
setosa,vers...
sepal_length
数值
4.4-7.7
参数设置
组件描述

图 98

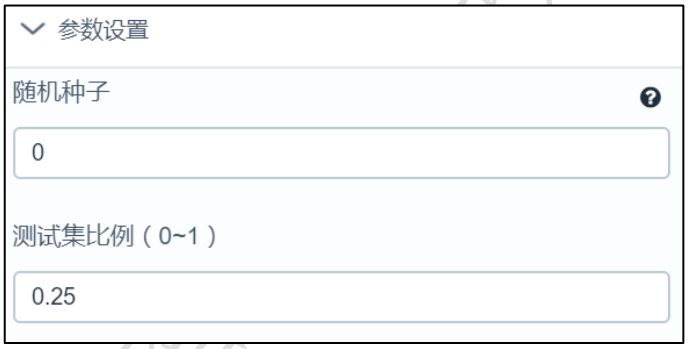

参数设置
随机种子
0
测试集比例（0~1）
0.25

图 99


重命名
删除
全部运行
运行到此处
运行该节点
从此节点运行
查看数据
查看日志
查看源码
输入源
数据拆分

图 100

3.4.2.12 频数统计

图标： 频数统计

描述：频数统计对某种特征的数(标志值)出现的次数进行统计。

字段属性

待统计列：需要计算频数的一列数据。

参数设置

无

输出

表结果：频数表。

报告：无。

示例

对某列数据进行频数统计，数据如图 101 所示。

预览数据

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 101

- 勾选需要进行频数统计的数据，如图 102 所示。
- 运行该节点，右击选择查看数据，如图 103 所示。

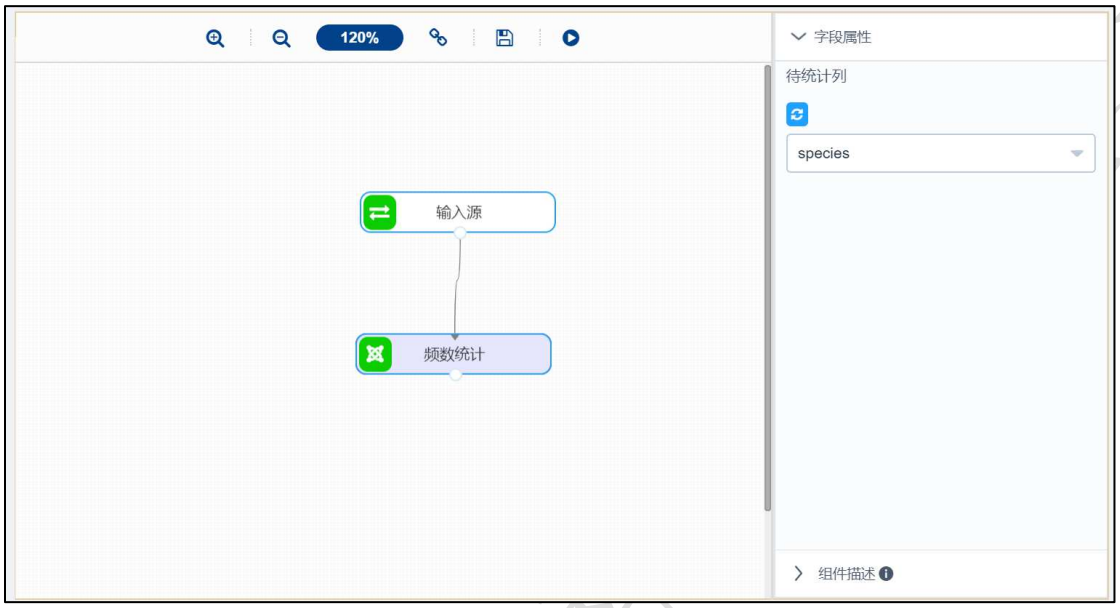


图 102

预览数据

species	count
setosa	50
versicolor	50
virginica	50

图 103

3.4.2.13 新增序列

图标： 新增序列

描述： 新增序列是指在原有数据基础上，新增一列自增序列，作为标识或其他特定作用。

字段属性

特征列：增加的序列会在勾选的字段基础上添加。如图 104 所示。

图 104

参数设置

新增序列列名：定义新增序列的列名。默认 new，如图 105 所示。

图 105

输出

表结果：新增序列后的表。

报告：无。

示例

下面对某数据新增一列自增序列，原表共有四个字段：id，f，m，r。需要新增一列列名为 new 的自增序列。

- 勾选原表字段，新增序列需要在勾选的数据基础上新增。如图 106 所示。
- 保留定义的新增序列列名为 new。如图 107 所示。

- 结果如图 108 所示。

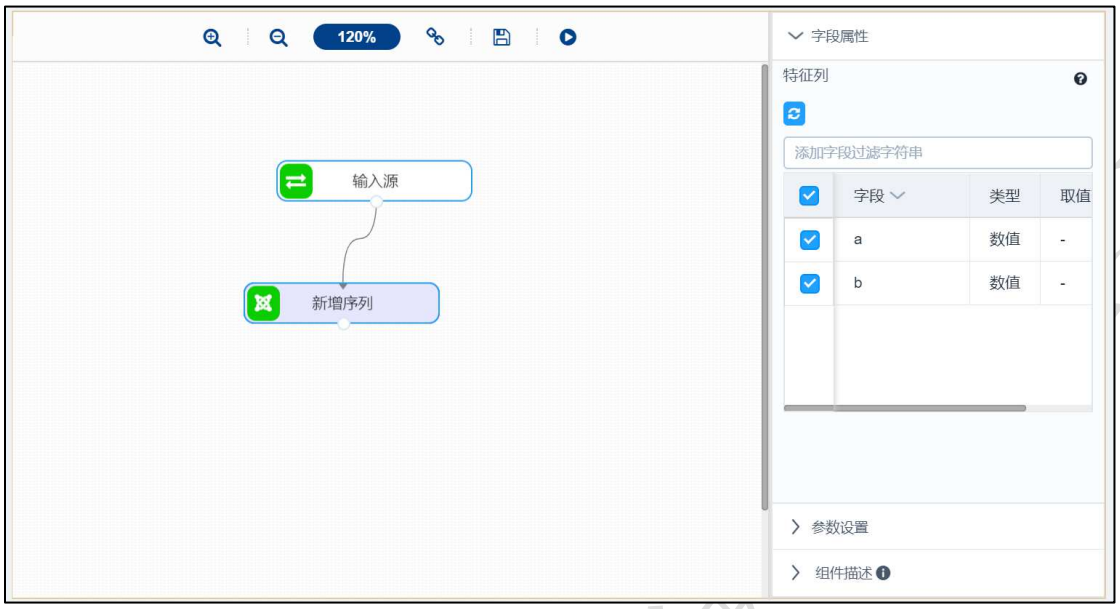


图 106

图 107

图 108

3.4.2.14 K 步差分

图标： K步差分

描述： K 步差分是指相距 k 期的两个序列值之间的减法运算。

字段属性

待差分序列：请选择数值型数据，如图 109 所示。

图 109

参数设置

K 期：整数型，默认 1，如图 110 所示。

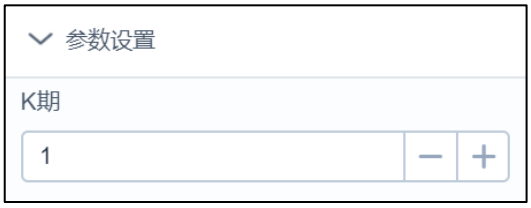


图 110

输出

表结果：差分结果。

报告：无。

示例

下面对某数据进行 1 步差分。原数据如图 111 所示。

预览数据

a	b
1	2
2	3
3	4
4	5
5	6

图 111

- 选择待差分序列，如图 112 所示。
- 设置差分步数 K，如图 113 所示。
- 运行成功后，选择查看数据,结果如图 114 所示。

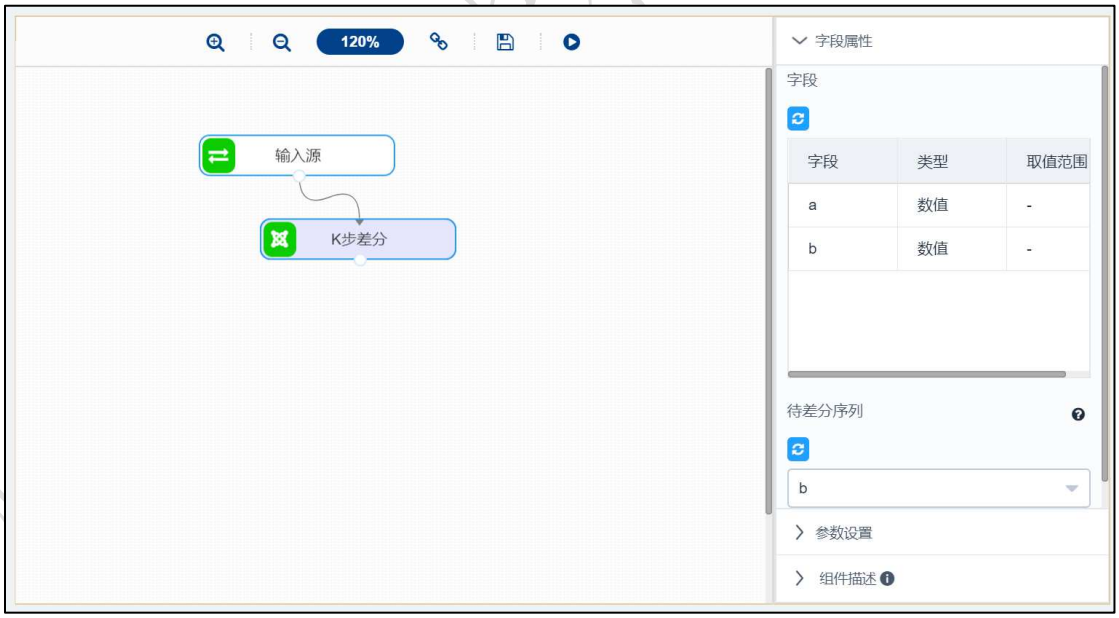


图 112

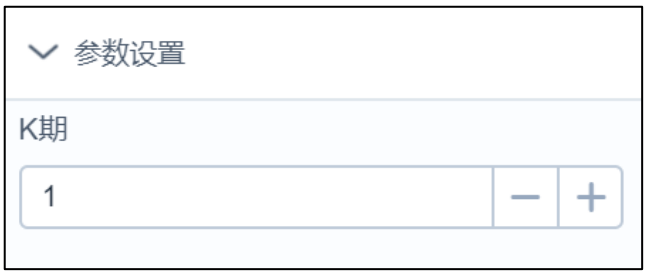


图 113

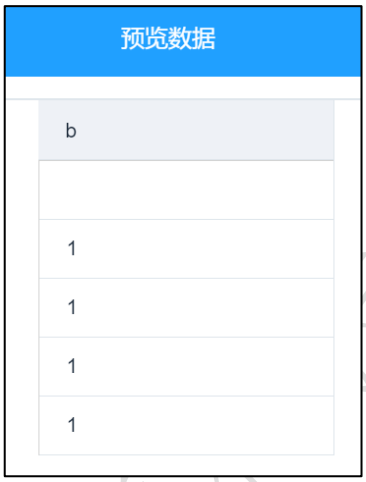


图 114

3.4.2.15 分组聚合

图标： 

描述：分组聚合是指将数据按照某个键拆分为多组，使用一个函数应用到各个分组上产生一个新值，最后将执行的结果合并。

字段属性

待分组与聚合的列：当聚合方式为 count 外，请选择数值型数据。需要注意的是勾选的列不包含键，如图 115 所示。

图 115

键：按照该列的值将数据分组，如图 116 所示。

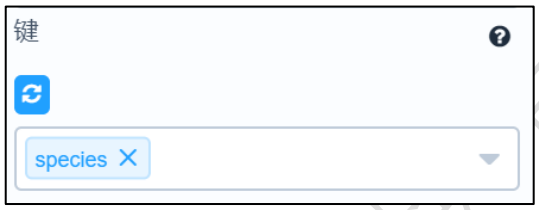


图 116

参数设置

聚合方式：应用到每个分组上的函数。包括 count、max:、mean、median、size、min、std:、sum。

输出

表结果：分组聚合结果。

报告：无。

示例

下面对某数据进行分组聚合,原数据如图 117 所示.

预览数据

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 117

- 勾选待分组与聚合列,将这些数据按照列 species 分组。如图 118 所示。
- 选择聚合方式为 max，如图 119 所示。
- 运行成功，选择查看数据，如图 120 所示。

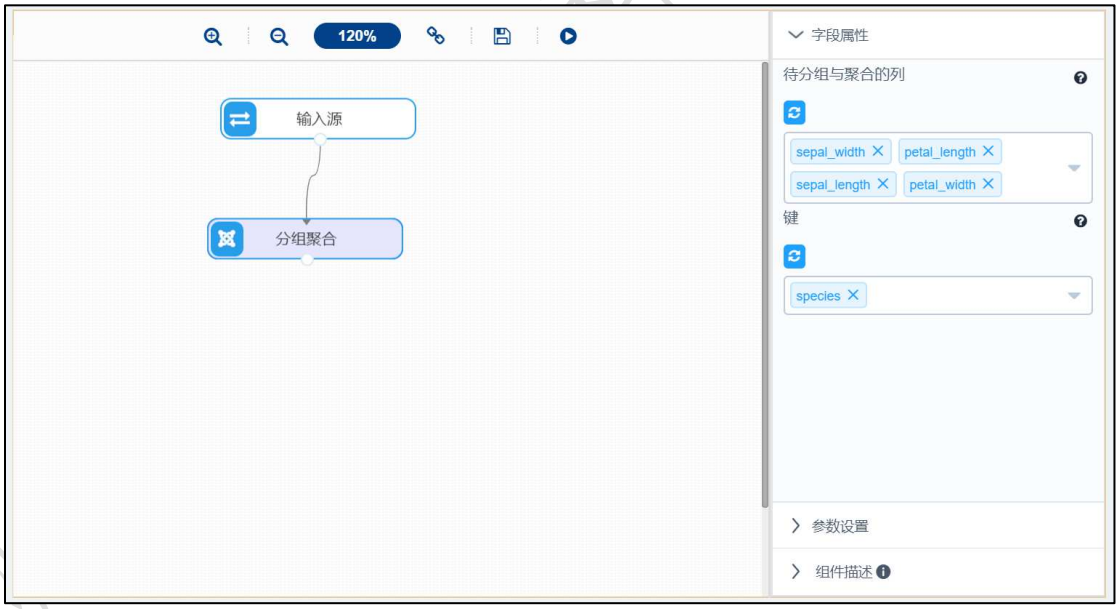


图 118

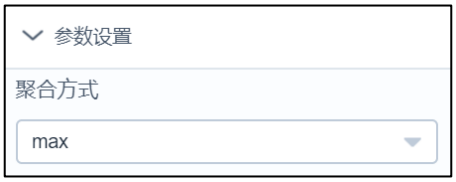


图 119

预览数据

species	sepal_width	petal_length	sepal_length	petal_width
setosa	4.4	1.9	5.8	0.6
versicolor	3.4	5.1	7	1.8
virginica	3.8	6.9	7.9	2.5

图 120

3.4.2.16 数据标准化

图标： 

描述：数据标准化处理是将数据按比例缩放，使之落入一个小的特定区间。

字段属性

特征列：选择进行标准化的列，请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列，如图 121 所示。

图 121

参数设置

标准化方式：标准化方式包括极差标准化、零均值标准化和小数定标标准化，默认零均值标准化。

最小值：选择极差标准化时有效。

最大值：选择极差标准化时有效。如图 122 所示

图 122

输出

表结果：标准化结果。

报告：无。

示例

下面对某数据进行标准化处理。原数据如图 123 所示。

预览数据

id	r	f	m
1	27	6	232.61
2	3	5	1507.11
3	4	16	817.62
4	3	11	232.81
5	14	7	1913.05
6	19	6	220.07
7	5	2	615.83
8	26	2	1059.66

共 940 条 25 条/页 1 2 3 4 5 6 ··· 38 前往 1 页

图 123

- 勾选需要进行数据标准化的数据。如图 124 所示。
- 选择标准化方式为零均值标准化，如图 125 所示。
- 结果如图 126 所示。

图 124

图 125

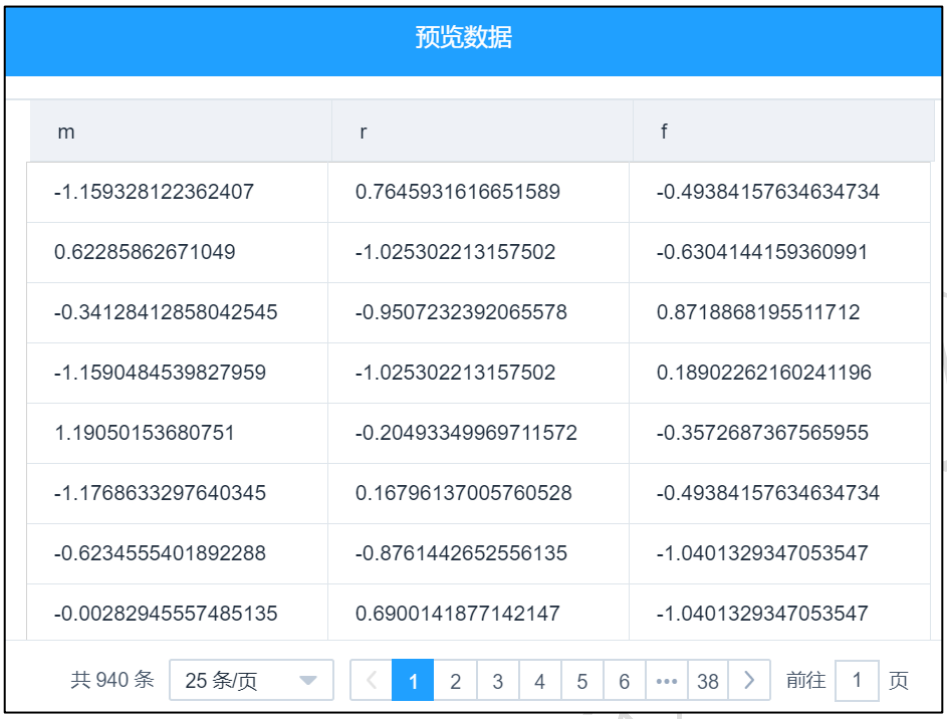


图 126

3.4.2.17 衍生变量

图标： 衍生变量

描述：衍生变量是指将一列或多列通过基本运算生成新列。

字段属性

特征列：必选。选择进行衍生变量的列。请选择数值型数据,增加的序列会在勾选的字段基础上添加，如图 127 所示。

图 127

参数设置

变量名：新增列的列名，输入要求：;1.英文开头;2.小写英文、数字、下划线;3.长度 1-10，默认名称是 new。

表达式：必填。目前只支持四则运算符：+，-，*，/。如图 128 所示。

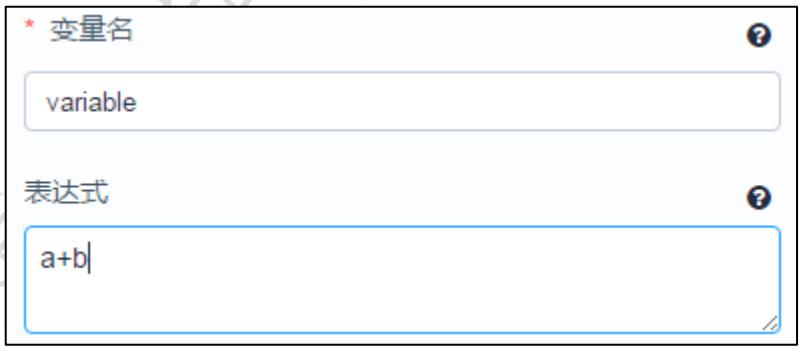


图 128

输出

表结果：勾选的列与衍生出来的列构成的表。

报告：无。

示例

下面对某数据使用衍生变量，将源数据 sepal_width，petal_length 两字段相加（sepal_width

+ petal_length）构成新列 new。

- 勾选需要进行衍生变量的列。如图 129 所示。
- 定义新增列的列名为 new，表达式为”sepal_width + petal_length”，如图 130 所示。
- 运行该节点，右击选择查看数据，如图 131 所示。

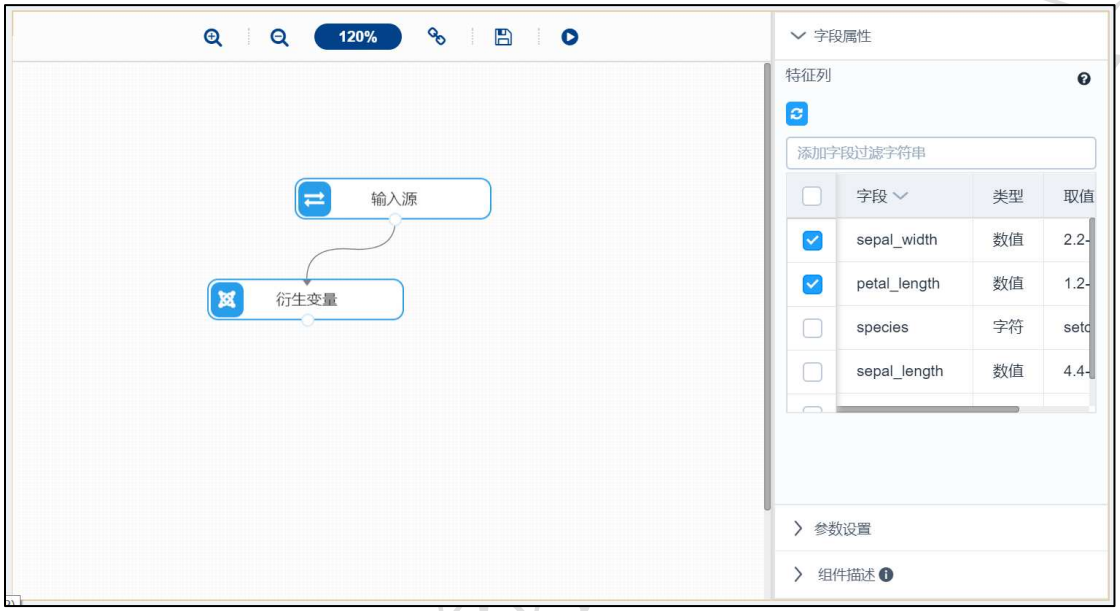


图 129

图 130

图 131

3.4.2.18 修改列名

图标： 修改列名

描述：修改列名是指对数据表中的字段名进行修改。

字段属性

特征列：必选。勾选列将传入下个组件,并且可供该组件修改字段名，如图 132 所示。

图 132

列名转换：必填。修改对对应的列名为需要的列名，如图 133 所示。

图 133

参数设置

无。

输出

表结果：列名修改后的结果。

报告：无。

示例

下面对某数据数据的列名进行修改。

- 勾选需要进行列名修改的数据，如果只勾选 m，r 两列，则只将 m，r 两列数据传入下个组件。如图 134 所示。
- 运行该组件，右击选择查看数据，结果如图 135 所示。

图 134

预览数据

mm	rr
232.61	27
1507.11	3
817.62	4
232.81	3
1913.05	14
220.07	19
615.83	5
1059.66	26

图 135

3.4.2.19 修改类型

图标： 

描述： 修改类型是指对数据表中的字段类型进行修改，目前平台提供的类型有数值型（numeric）、字符型（text）、日期（date）、时间（timestamp）。

字段属性：

字段：必选。勾选需要修改类型的字段，并传入下一个组件，如图 136 所示。

图 136

修改类型：当新类型为数值型时,可以通过调整参数的值设定小数点位数，如图 137 所示。

图 137

参数设置

无。

输出

表结果：修改类型之后的数据。

报告：无。

示例

下面对某数据的类型进行修改，原数据类型如图 138 所示。

图 138

- 修改字段“id”的数据类型为字符型，如图 139 所示。
- 运行成功，下个组件就可获取到字段“id”的数据类型为字符型，如图 140 所示。

图 139

图 140

3.4.2.20 Python 脚本

图标： Python脚本

描述：Python 脚本是指可直接将 Python 脚本按照一定格式粘贴至脚本区作为组件运行。

字段属性

输入列表：当输入节点连入某个数据之后，则会将该数据的表名传值给 input1/input2/input3/input4，如图 141 所示。

图 141

脚本：必填。在脚本区域输入 Python 脚本，格式要求如以下示例。

载入数据代码格式如以下两行脚本。

data_in = db_utils.query(conn, 'select '+ field1 + field2 + ' from ' + inputs['input1'])

return(data_in)

注意每行代码需要缩进 4 个空格

需要输出结构化数据，要求必须为 DataFrame，需要使用 return()。如图 142 所示

图 142

字段属性

无。

输出

表结果：由脚本运算出来的数据。

报告：无。

示例

对某数据执行 SQL 语句，原数据如图 143 所示。

图 143

- 在脚本编辑区内放入 Python 代码，如图 144 所示。
- 运行成功后，查看数据，如图 145 所示。

图 144

图 145

3.4.2.21 Granger 因果检验

图标： Granger因果检验

描述：Granger 因果检验是检验两列时序数据之间是否存在格兰杰因果关系。在时间序列情形下，两个经济变量 X、Y 之间的格兰杰因果关系定义为：若在包含了变量 X、Y 的过去信息的条件下，对变量 Y 的预测效果要优于只单独由 Y 的过去信息对 Y 进行的预测效果，即变量 X 有助于解释变量 Y 的将来变化，则认为变量 X 是引致变量 Y 的格兰杰原因。

字段属性

时间序列 1：请选择数值型数据。

时间序列 2：请选择数值型数据。

参数设置

延迟阶数：默认 2。

输出

表结果：无。

报告：对应各滞后阶数的检验结果。

示例

下面对两个时序数据进行 Granger 因果检验。

- 选择两个时序数据，如图 146 所示。
- 设置延迟阶数为 2，则对每次 1-2 两个延迟阶数分别进行 Granger 因果检验。如图 147 所示。
- 运行成功，选择查看报告，如图 148 所示。

图 146

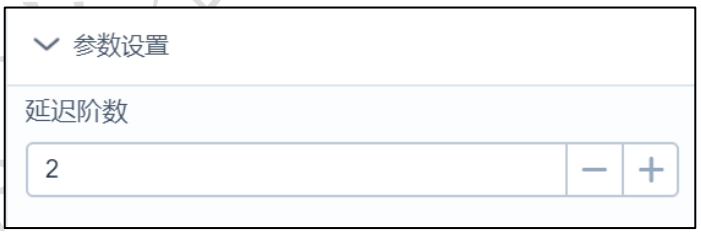


图 147

图 148

3.4.3 统计分析

3.4.3.1 单样本 t 检验

图标： 

描述： 单样本 t 检验是检验一个样本平均数与一个已知的总体平均数的差异是否显著。

字段属性

待检验序列：请选择数值型数据。如**错误!未找到引用源。**所示。

图 149

参数设置

期望值：数值型。如**错误!未找到引用源。**所示。

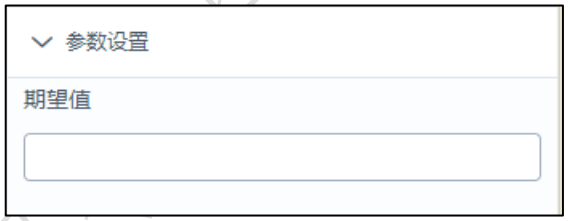


图 150

输出

表结果：无。

报告：p-value。

示例

下面对某列数据进行单样本 t 检验。

- 选择待检验序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 点击参数设置，期望值设置为 5.0。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

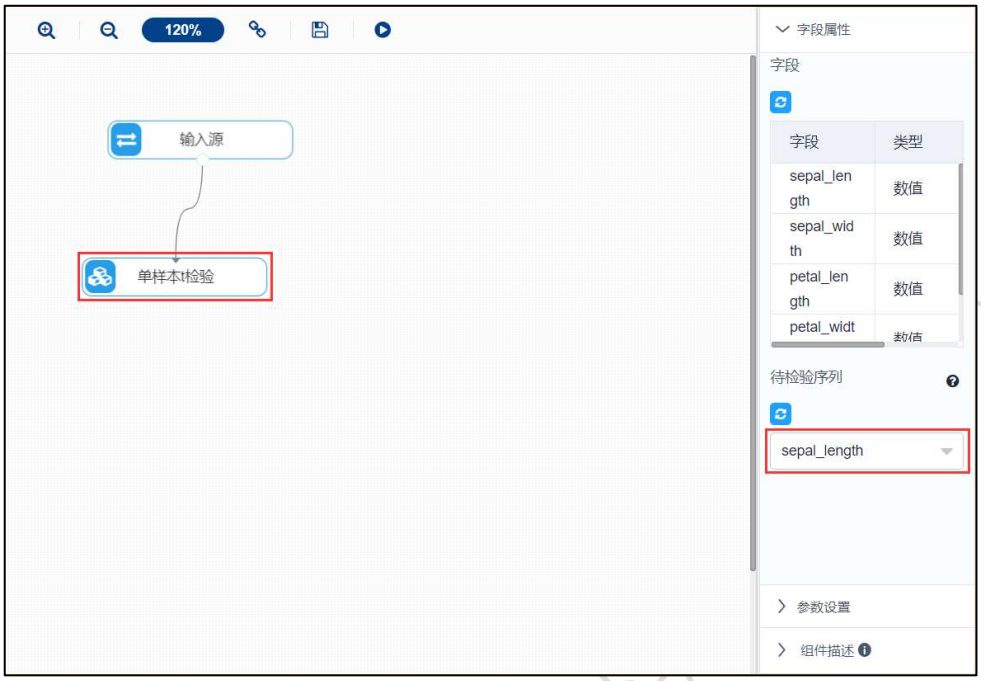


图 151

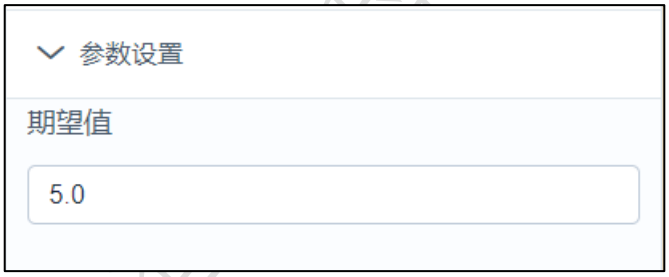


图 152

图 153

3.4.3.2 双样本 t 检验

图标： 

描述： 双样本 t 检验是检验两个样本平均值是否显著不同。

字段属性

样本一：请选择数值型数据。

样本二：请选择数值型数据。

如**错误!未找到引用源。**所示。

图 154

输出

表结果：无。

报告：p-value。

示例

下面对某两列数据进行双样本 t 检验。

- 选择两列待检验序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

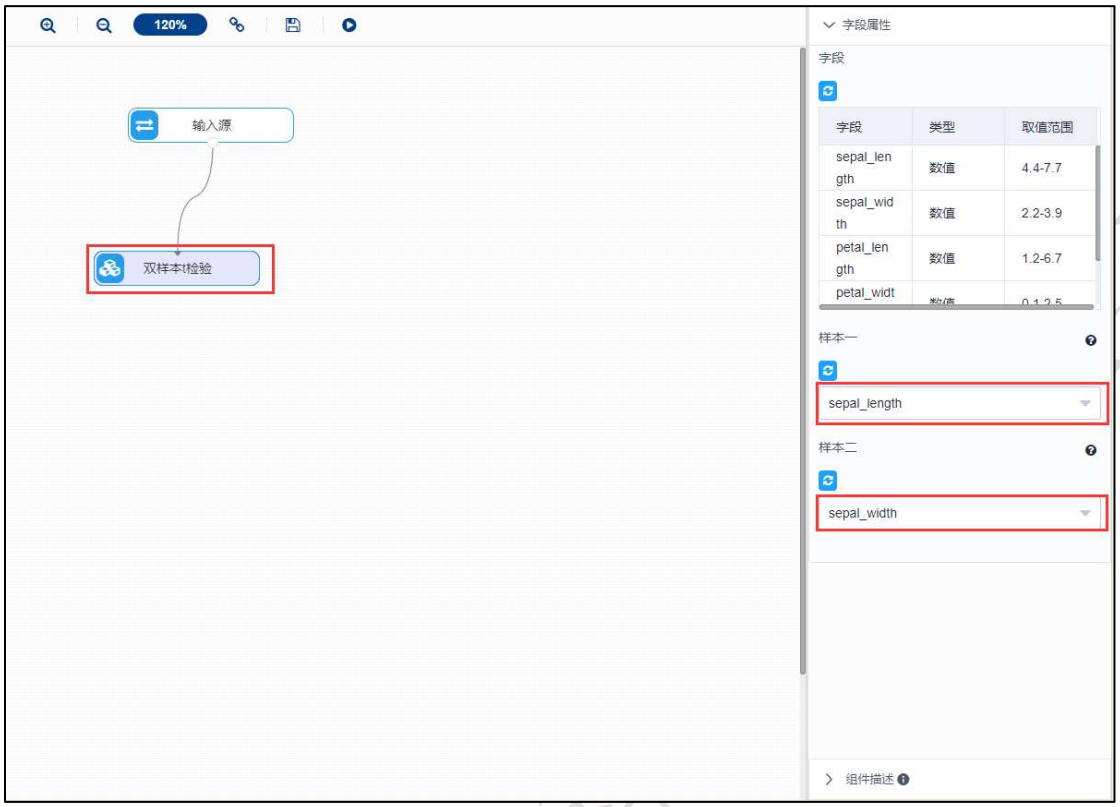


图 155

图 156

3.4.3.3 二项分布检验

图标： 

描述：二项分布检验是为了检验数据是否服从二项分布。现实生活中有很多数据的取值只有两类，如医学中的生与死、患病的有与无、性别中的男性和女性、产品的合格与不合格等。从这种二分类总体中抽取的所有可能结果，要么是对立分类中的这一类，要么是另一类，其频数分布称为二项分布。

字段属性

待检验序列：必须为二分类数据。如**错误!未找到引用源。**所示。

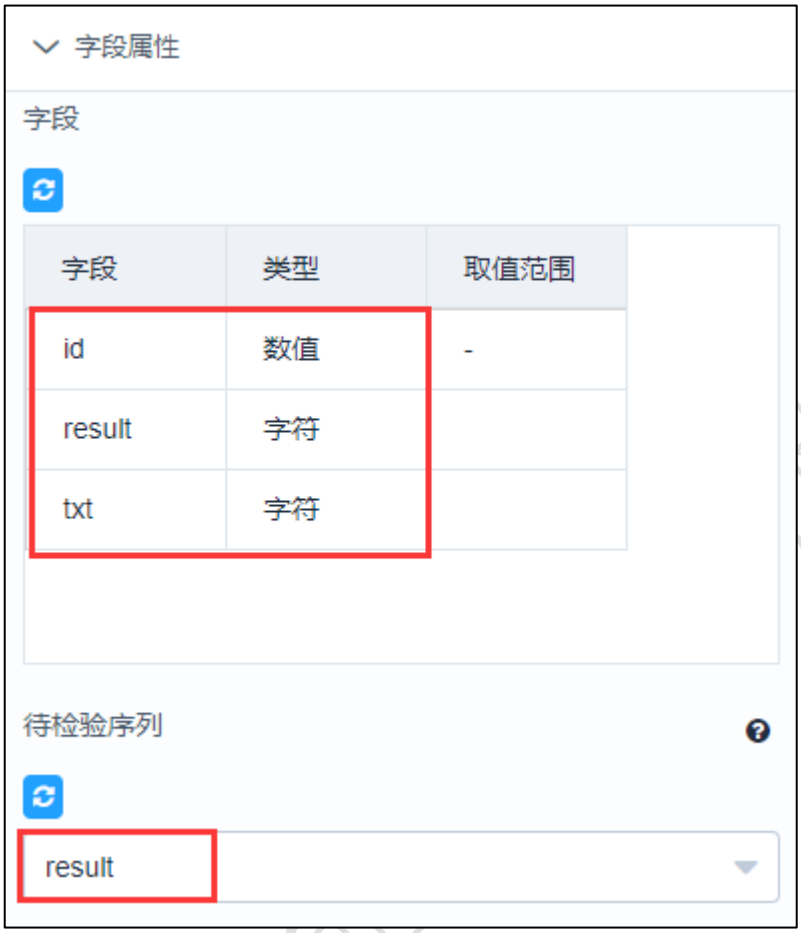


图 157

参数设置

需要统计成功次数的值：待检验序列中的某个值。请填写数值型数据。

假设成功的概率：p 值。范围 0 ~ 1。

检验方法：双侧检验：two-sided；上侧检验：greater；下侧检验：less。

如**错误!未找到引用源。**所示。

图 158

输出

表结果：无。

报告：p-value。

示例

下面对某列数据进行卡方检验。

- 选择待检验序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 点击参数设置，需要统计成功次数的值设置为 0，假设成功的概率设置为 0.7。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

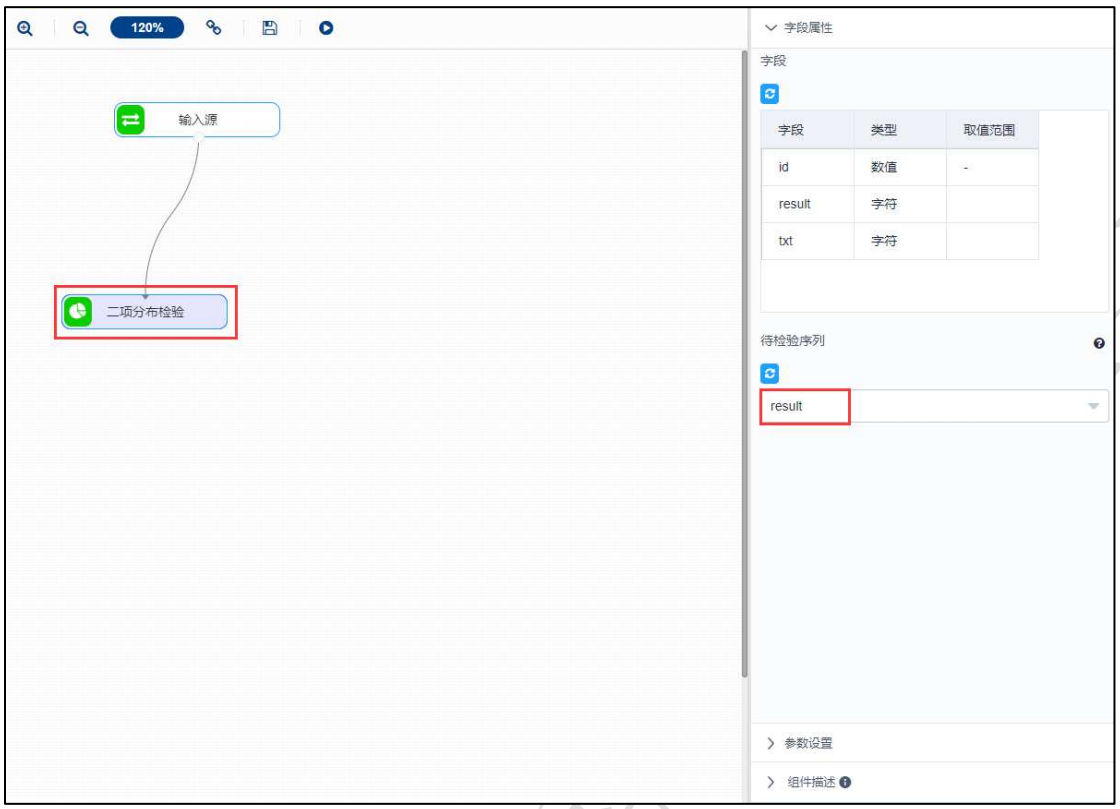

120%
输入源
二项分布检验
字段属性
字段
字段
类型
取值范围
id
数值
-
result
字符
txt
字符
待检验序列
result
参数设置
组件描述

图 159


参数设置
需要统计成功次数的值
0
成功的概率
0.7
检验方法
双侧检验

图 160

图 161

3.4.3.4 卡方检验

图标： 卡方检验

描述： 计算单向卡方检验。

字段属性

待检验序列：请选择数值型数据。如**错误!未找到引用源。**所示。

图 162

输出

表结果：无。

报告：p-value。

示例

下面对某列数据进行卡方检验。

- 选择待检验序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

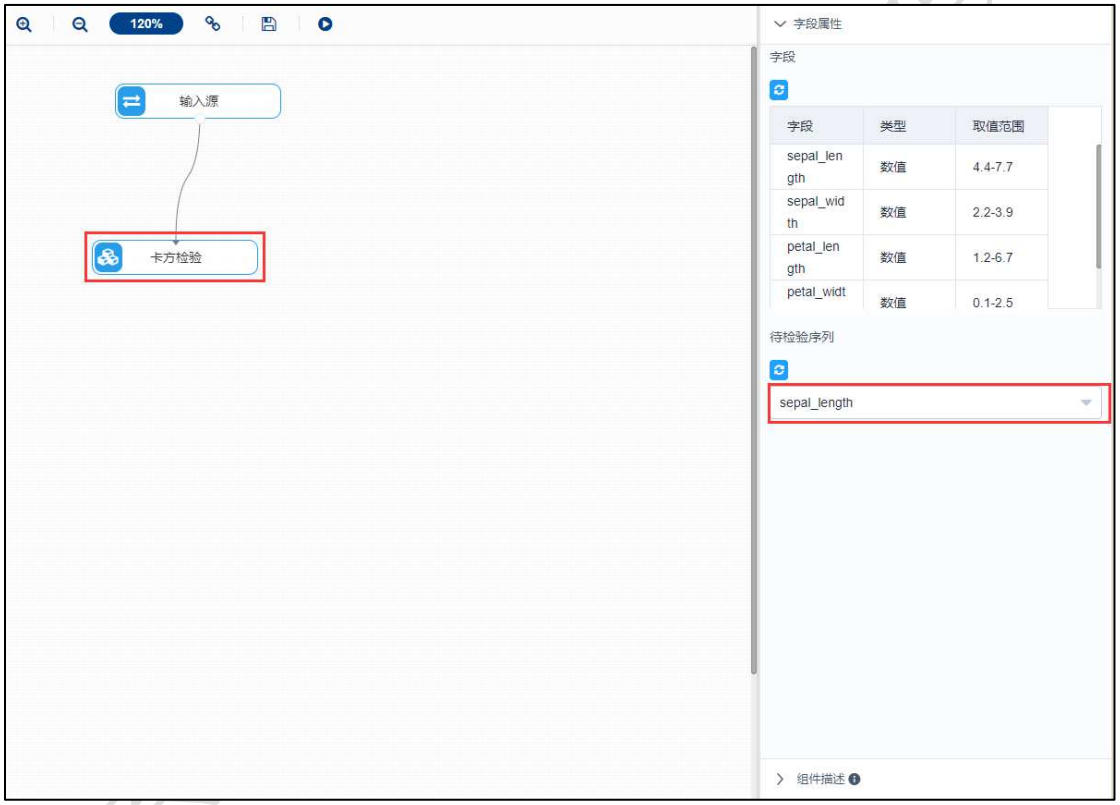


图 163

图 164

3.4.3.5 相关性分析

图标： 

描述：相关性分析是指对两个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度。

字段属性

特征列：对选择的列计算两两之间的相关密切程度。请选择数值型数据。如**错误!未找到引用源。**所示。

图 165

参数设置

检验方法：可以选择 peason、kendail、spearman 方法，默认为 peason。如**错误!未找到引用源。**所示。

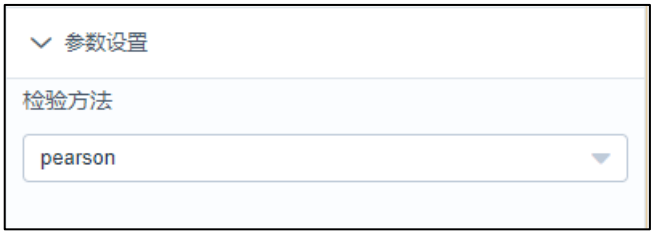


图 166

输出

表结果：相关系数。

报告：无。

示例

下面对某两列数据进行相关性分析。

- 选择两列待检验序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 点击参数设置，检验方法设置为 spearman。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看数据，结果如**错误!未找到引用源。**所示。

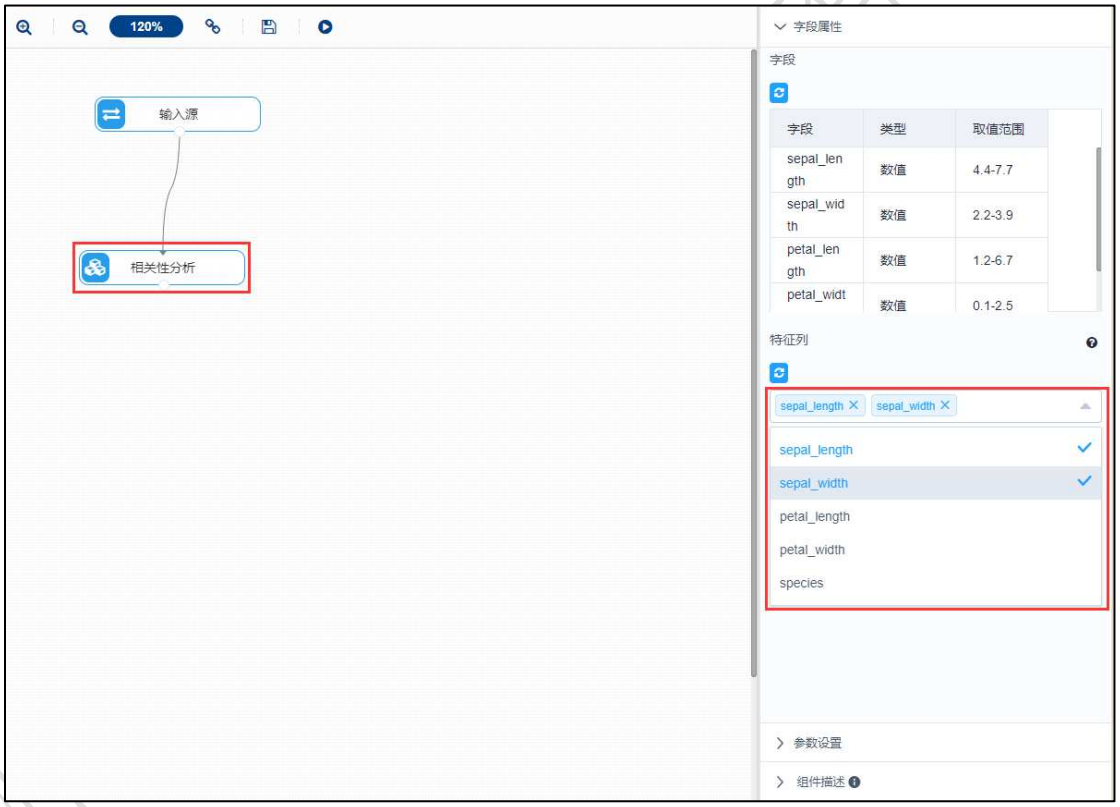


图 167

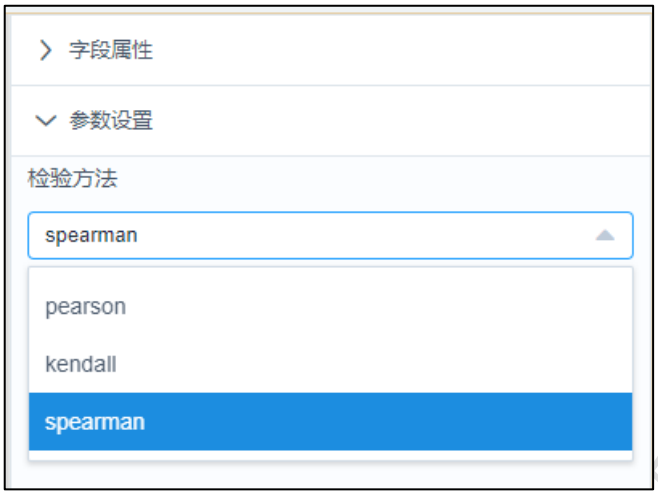


图 168

ind	sepal_length	sepal_width
sepal_length	1	-0.167
sepal_width	-0.167	1

图 169

3.4.3.6 正态性检验

图标： 正态性检验

描述：检验观测值是否服从正态分布。

字段属性

特征列：选择进行正态性检验的序列。选择多列时，默认对每一列进行正态性检验。请选择数值型数据。如**错误!未找到引用源。**所示。

图 170

参数设置：

输入存在空值时的处理方法：结果返回空值：propagate；提示存在空值错误：raise；忽略空值继续检验：omit。默认为结果返回空值。如**错误!未找到引用源。**所示。

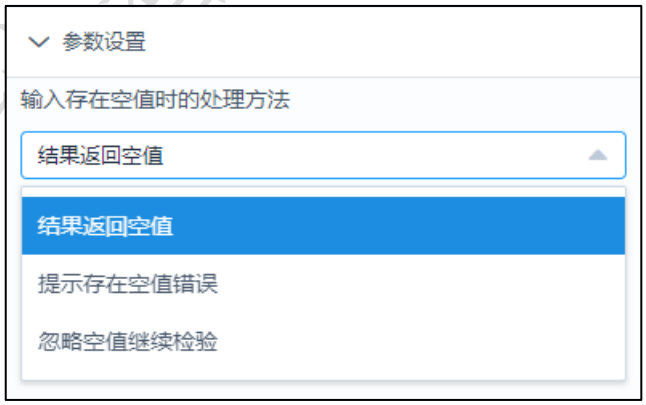


图 171

输出

表结果：无。

报告：p-value。

示例

下面对某两列数据进行正态性检验。

- 选择两列待检验序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 点击参数设置，处理方法设置为结果返回空值。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

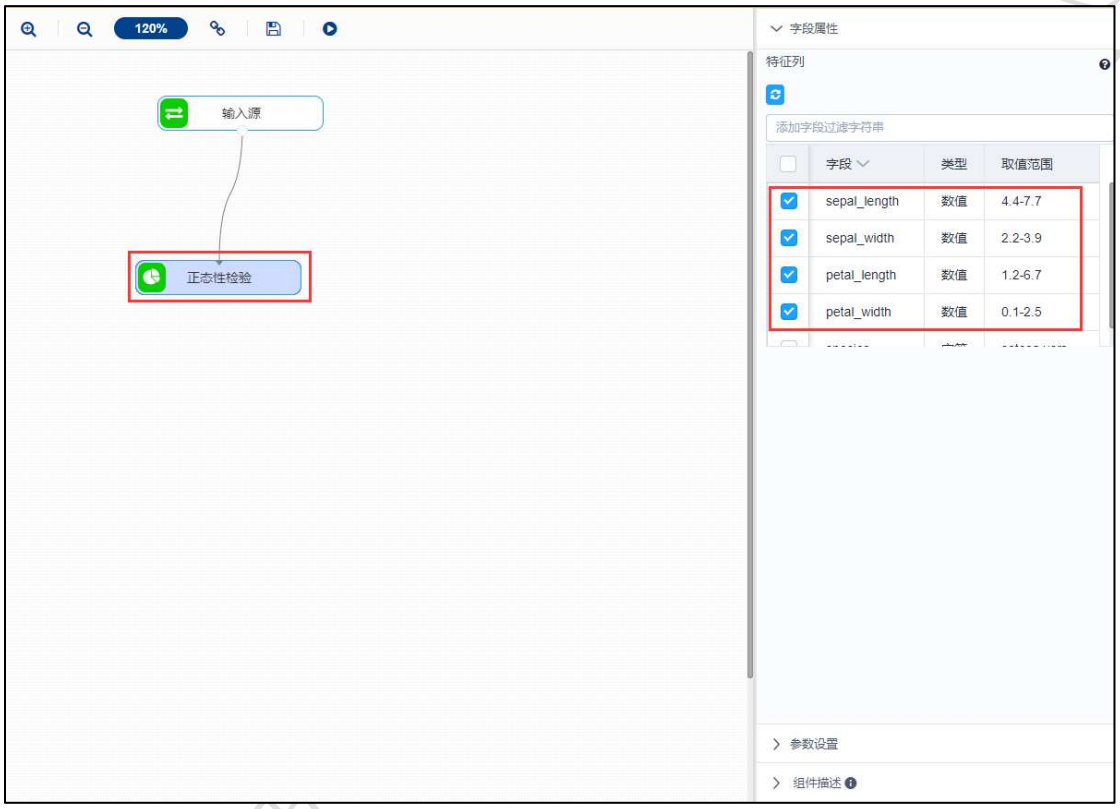


图 172

图 173

图 174

3.4.3.7 方差齐性检验

图标： 方差齐性检验

描述： 方差齐性检验是检验不同样本的总体方差是否相同的一种方法。

字段属性

特征列：选择两列进行方差齐性检验的序列。请选择数值型数据。如**错误!未找到引用源。**所示。

图 175

参数设置

检验方法：均值：mean；中值：median；截尾均值：trimmed。如**错误!未找到引用源。**所示。

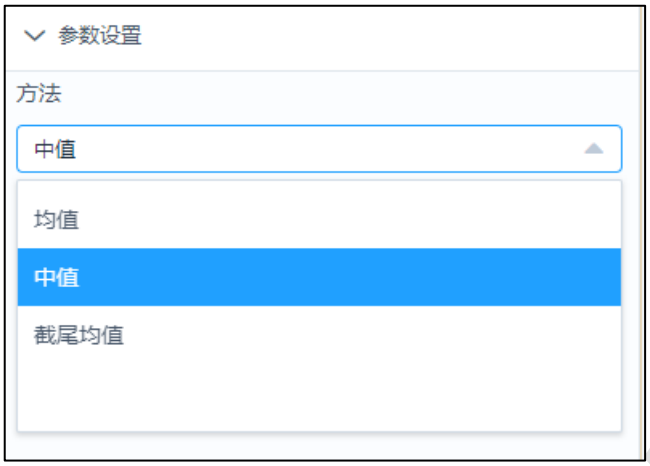


图 176

输出

表结果：无。

报告：p-value。

示例

下面对某两列数据进行方差齐性检验。

- 选择两列待检验序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 点击参数设置，检验方法设置为中值。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

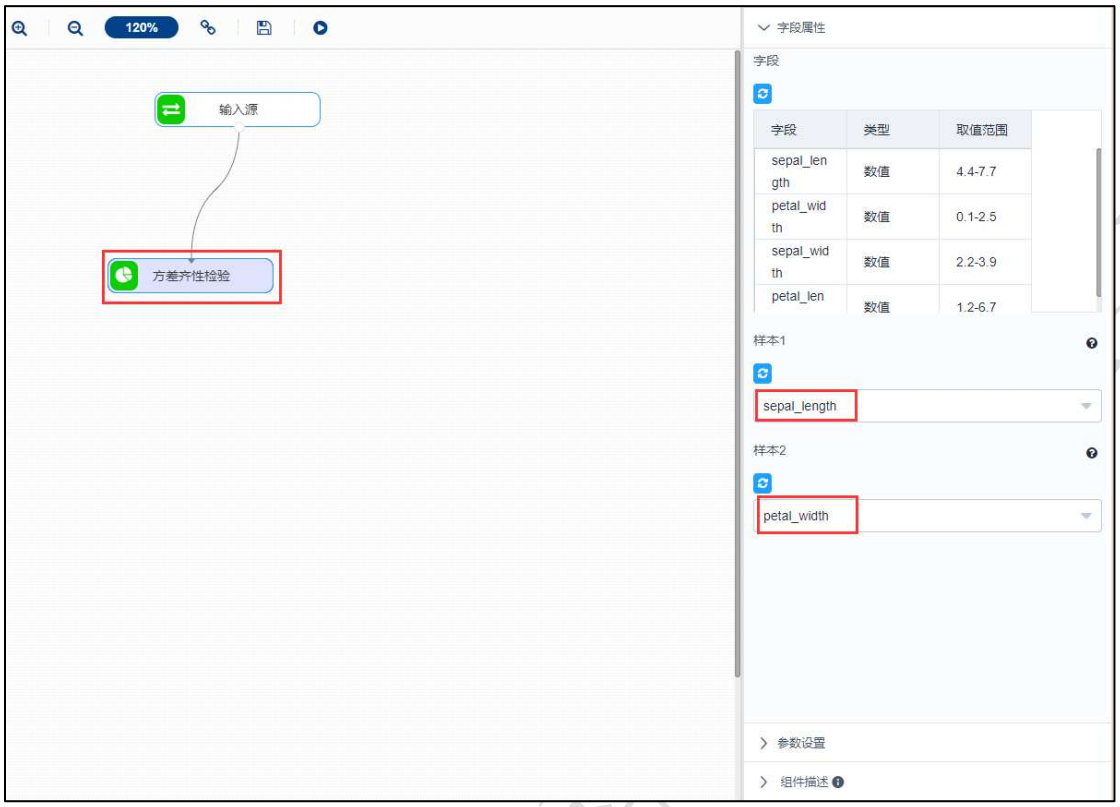


图 177

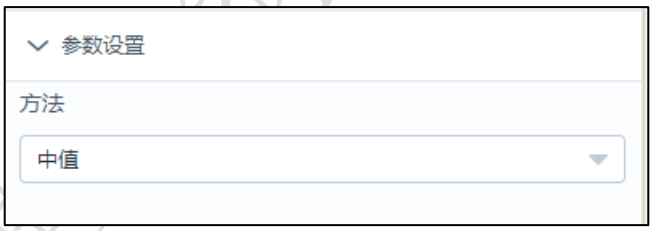


图 178

图 179

3.4.3.8 主成分分析

图标： 主成分分析

描述：指用几个较少的综合指标来代替原来较多的指标，而这些较少的综合指标既能尽可能多地反映原来较多指标的有用信息，且相互之间又是无关的。

字段属性

特征列：必选。选择进行分析的列。请选择数值型数据。如**错误!未找到引用源。**所示。

图 180

参数设置

降维后的维数：不超过所选数据列数。如**错误!未找到引用源。**所示。

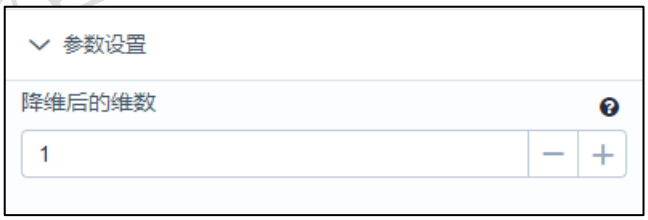


图 181

输出

表结果：主成分分析结果。

报告：无。

示例

下面对某数据进行主成分分析。

- 选择四列待分析序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 点击参数设置，降维后的尾数设置为 1。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看数据，结果如**错误!未找到引用源。**所示。

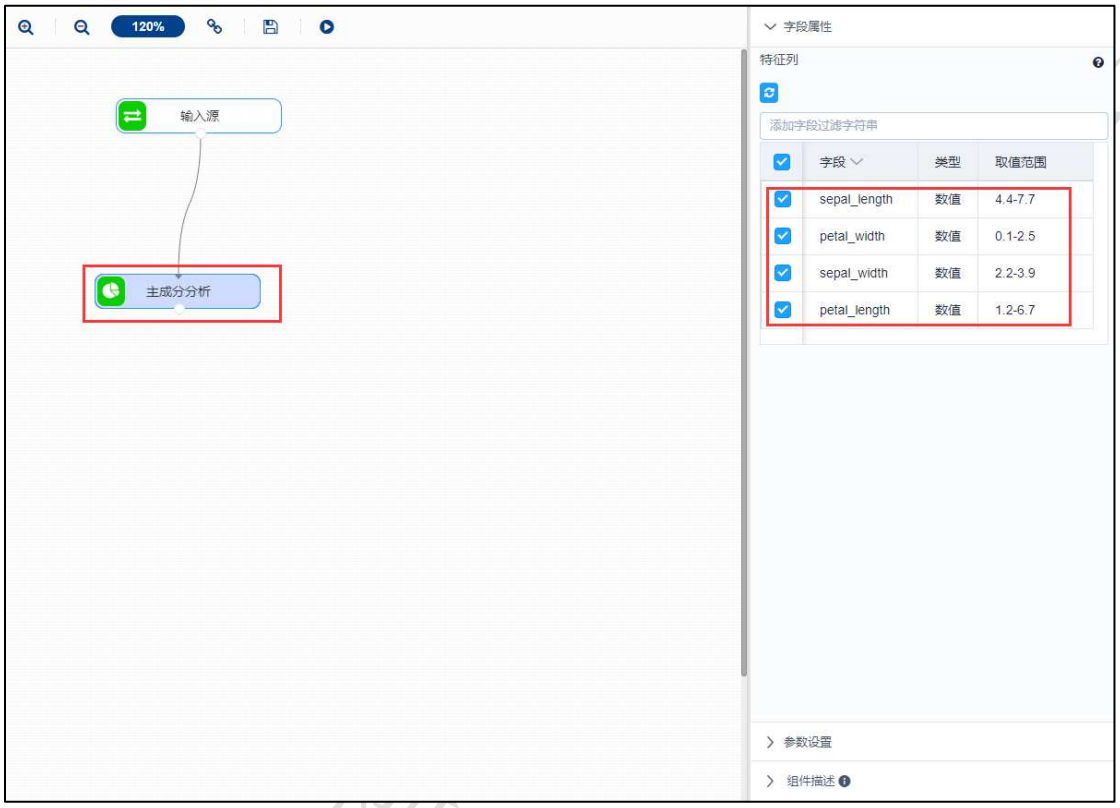


图 182

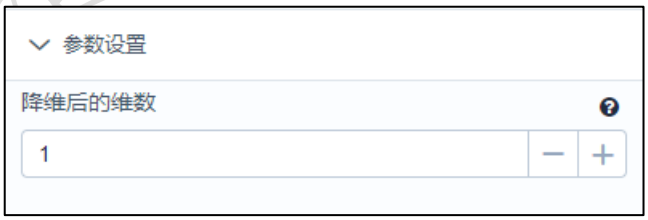


图 183

图 184

3.4.3.9 因子分析

图标： 因子分析

描述：因子分析的主要目的是用来描述隐藏在一组测量到的变量中的一些更基本的，但又无法直接测量到的隐性变量，是用来分析隐藏再表面现象背后的因子作用的一类统计模型。

字段属性

特征列：必选。选择进行分析的列。请选择数值型数据。如**错误!未找到引用源。**所示。

图 185

参数设置

因子个数：因子的个数。默认为 2。

最大迭代次数：默认为 1000。

如**错误!未找到引用源。**所示。

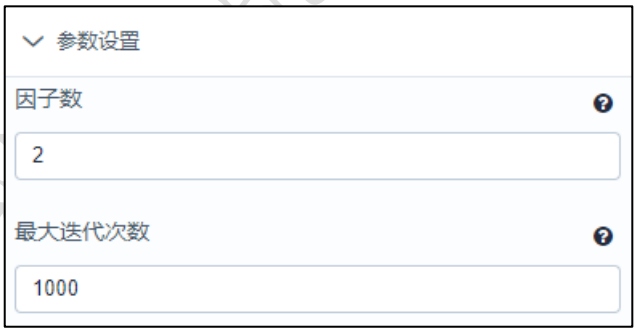


图 186

输出

表结果：因子得分数据。

报告：无。

示例

下面对某数据进行因子分析。

- 选择待检验序列，数据必须为数值型。如**错误!未找到引用源。**所示。

- 点击参数设置，设置如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

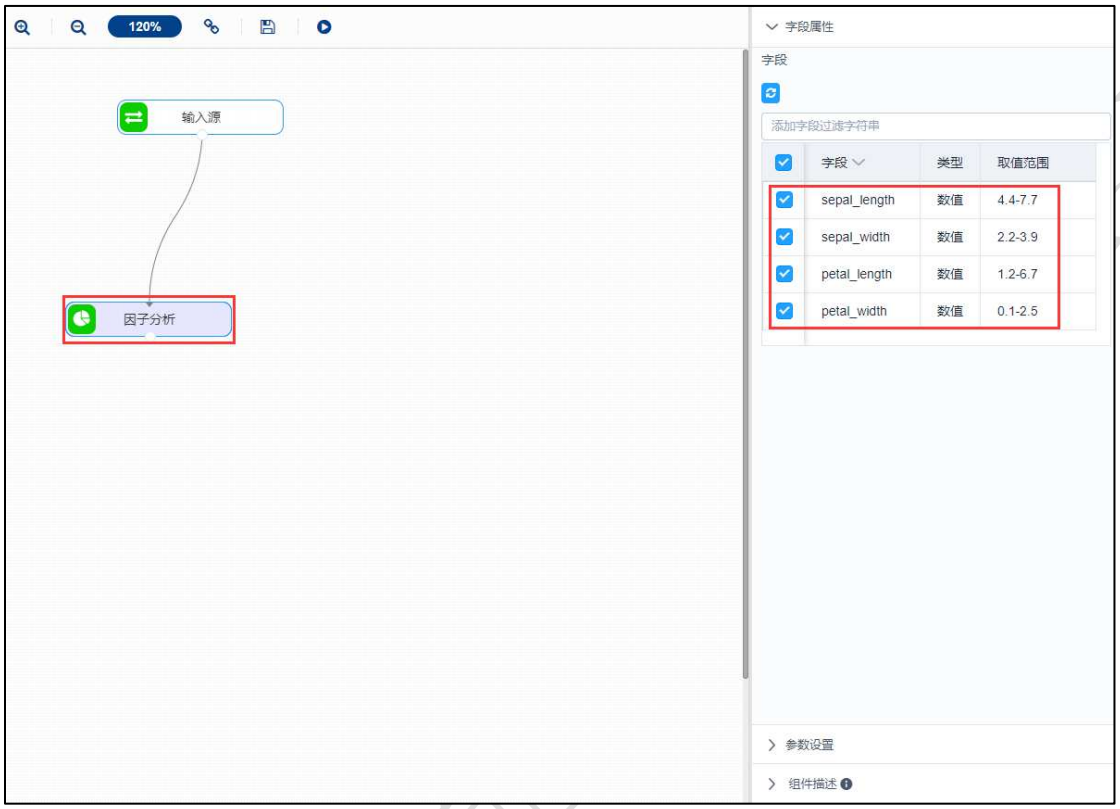


图 187

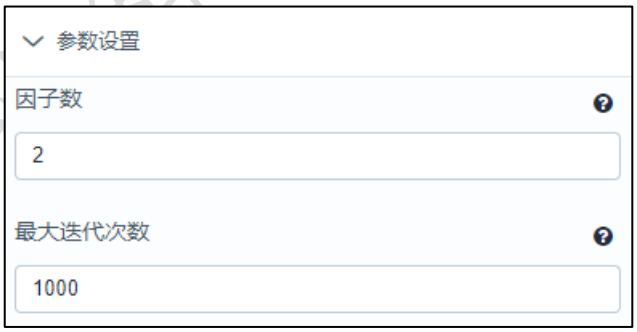


图 188

0	1
-1.3276172693246815	-0.5613107640080864
-1.3376385440508691	-0.0027976490706149624
-1.4028148326810546	0.3063494915081951
-1.3010427438665373	0.7188268345563075
-1.3334243941527006	-0.3645889853124369
-1.1466713352857187	-1.042281934408371
-1.353071777600891	0.5744735580502826
-1.2794072194843107	-0.2331298377724673

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 189

3.4.3.10 Wilcoxon 秩和检验

图标： 

描述：若检验假设成立，则两组的秩和相差不太大。

字段属性

特征列：必选。选择两列进行检验的列。请选择数值型数据。如**错误!未找到引用源。**所示。

图 190

输出

表结果：无。

报告：p-value。

示例

下面对某数据进行 Wilcoxon 秩和检验。

- 选择两列待分析序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

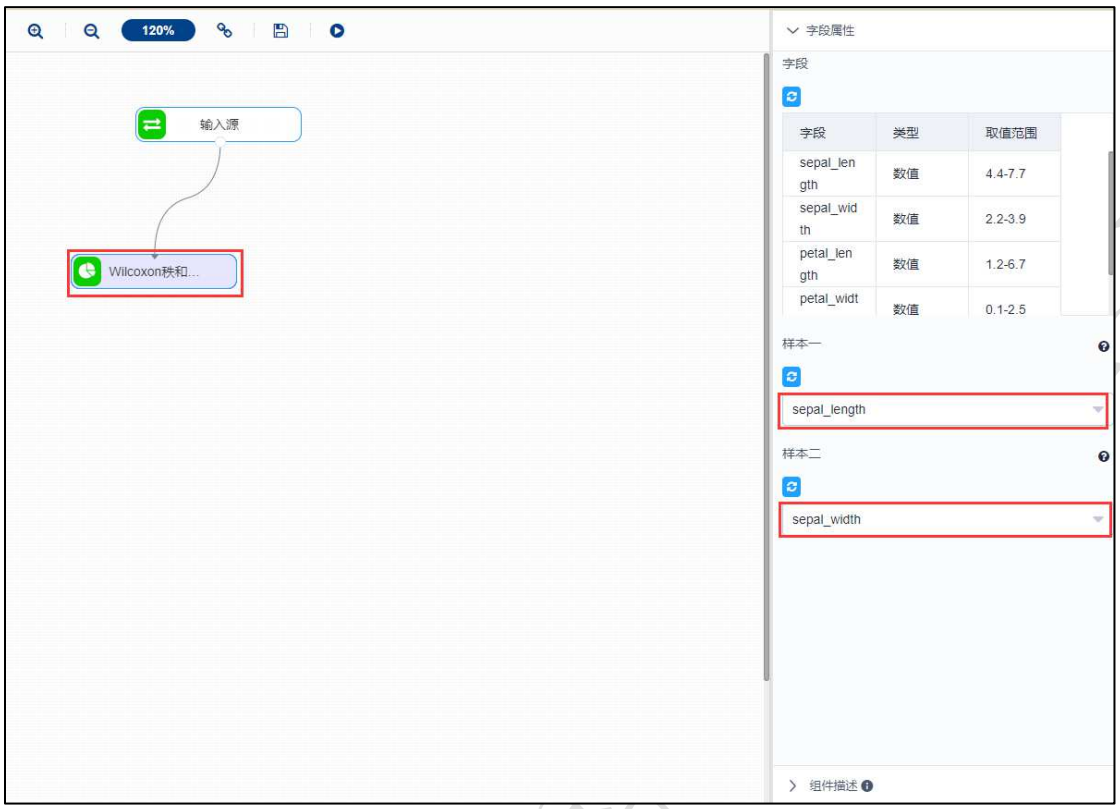
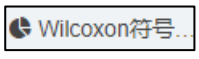


图 191

图 192

3.4.3.11 Wilcoxon 符号秩检验

图标： 

描述： 在 Wilcoxon 符号秩检验中，它把观测值和零假设的中心位置之差的绝对值的秩分别按照不同的符号相加作为其检验统计量。它适用于 T 检验中的成对比较，但并不要求成对数据之差 di 服从正态分布，只要求对称分布即可。检验成对观测数据之差是否来自均值为 0 的总体（产生数据的总体是否具有相同的均值）。

字段属性

特征列：必选。选择进行检验的列。请选择数值型数据。如**错误!未找到引用源。**所示。

图 193

输出

表结果：无。

报告：p-value。

示例

下面对某数据进行 Wilcoxon 符号秩检验。

- 选择待分析序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

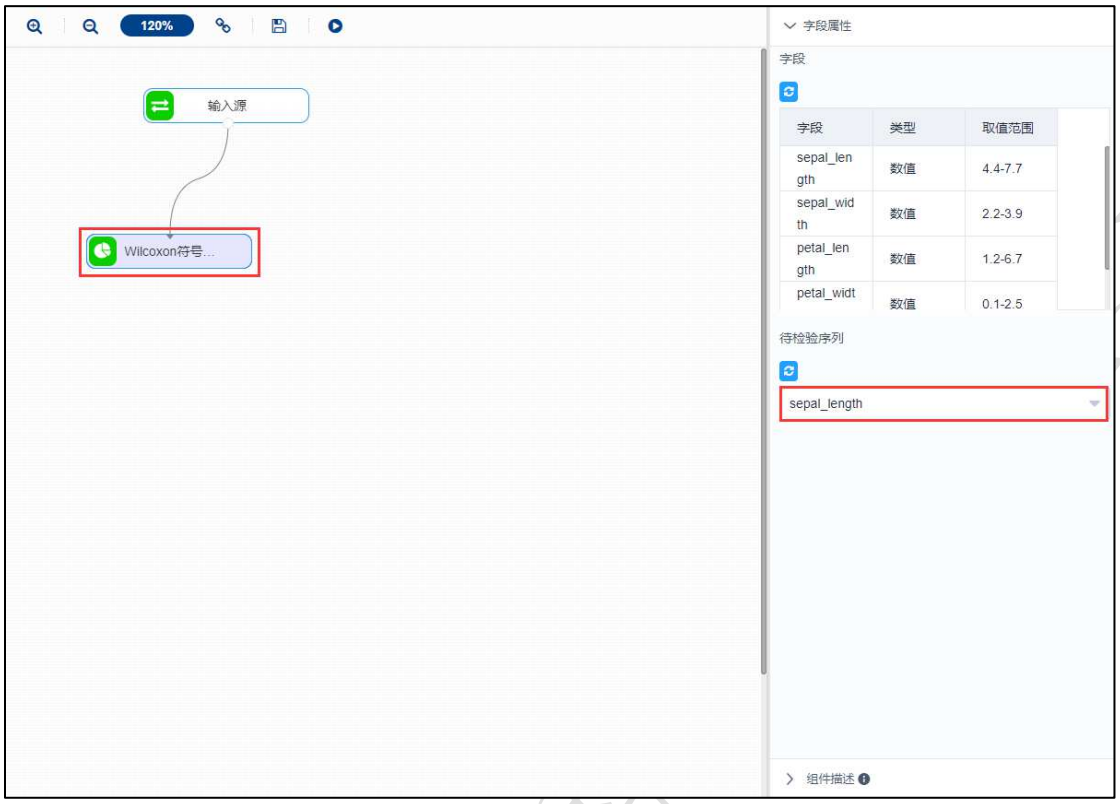


图 194

图 195

3.4.3.12 Mann-Whitney 检验

图标： Mann-Whitney.

描述： Mann-Whitney 检验假设两个样本分别来自除了总体均值以外完全相同的两个总体，目的是检验这两个总体的均值是否有显著的差别。

字段属性

特征列：必选。选择两列进行检验的列。请选择数值型数据。如**错误!未找到引用源。**

所示。

图 196

输出

表结果：无。

报告：p-value。

示例

下面对某数据进行 Mann-Whitney 检验。

- 选择两列待分析序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

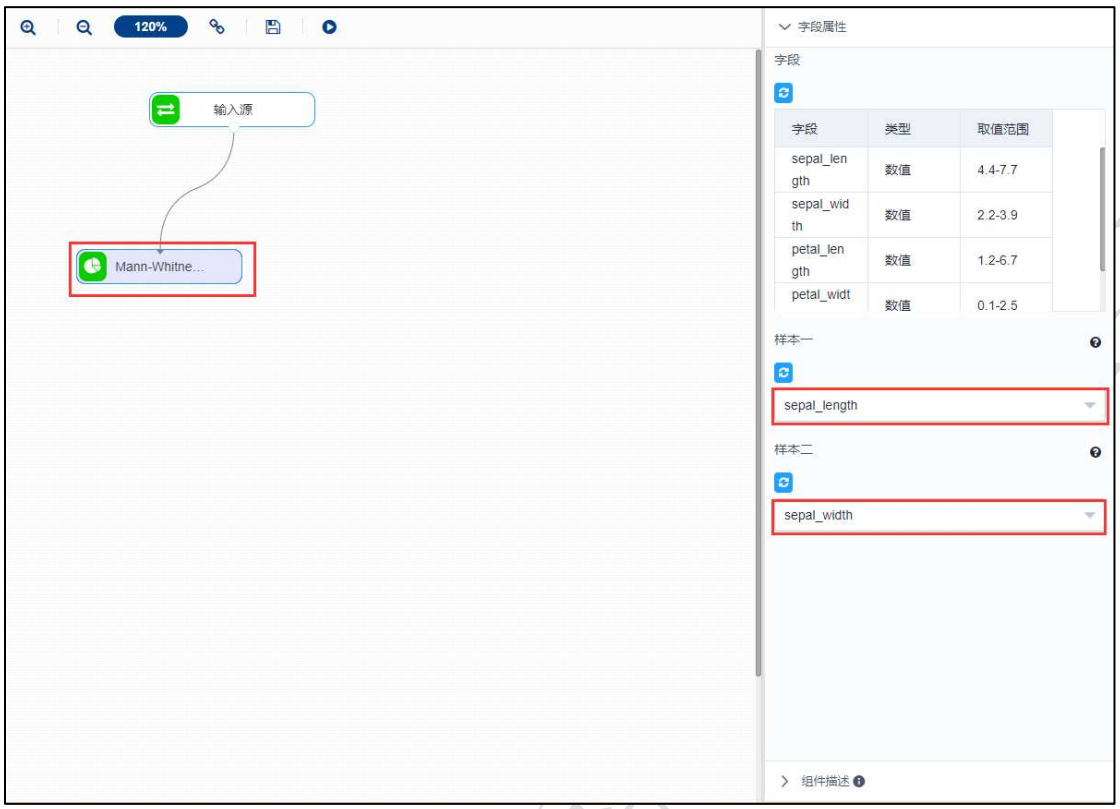
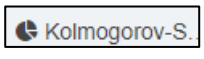


图 197

图 198

3.4.3.13 Kolmogorov-Smirnov 检验

图标： 

描述： Kolmogorov-Smirnov 检验基于累计分布函数，用以检验两个经验分布是否不同或一个经验分布与另一个理想分布是否不同。

字段属性

特征列：必选。选择两列进行检验的列。请选择数值型数据。如**错误!未找到引用源。**

所示。

图 199

输出

表结果：无。

报告：p-value。

示例

下面对某数据进行 Kolmogorov-Smirnov 检验。

- 选择两列待分析序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看报告，结果如**错误!未找到引用源。**所示。

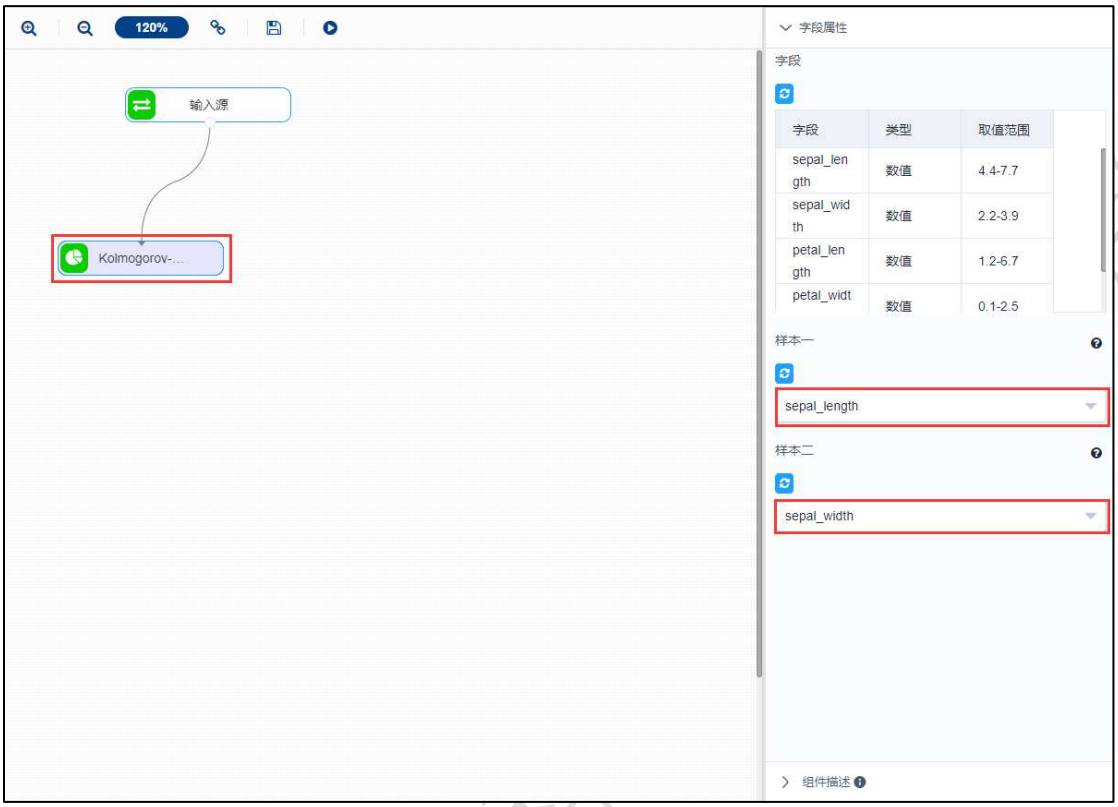


图 200

图 201

3.4.3.14 全表统计

图标： 全表统计

描述： 全表统计是对选取的各列进行描述性统计，并检验是否存在缺失值。

字段属性

特征列：请选择数值型数据。非数值型数据会被自动过滤。如**错误!未找到引用源。**所示。

图 202

输出

表结果：统计表。

报告：无。

示例

下面对某数据进行全表统计。

- 选择待统计的序列，数据必须为数值型。如**错误!未找到引用源。**所示。
- 运行该组件，对组件右击，选择查看数据，结果如**错误!未找到引用源。**所示。

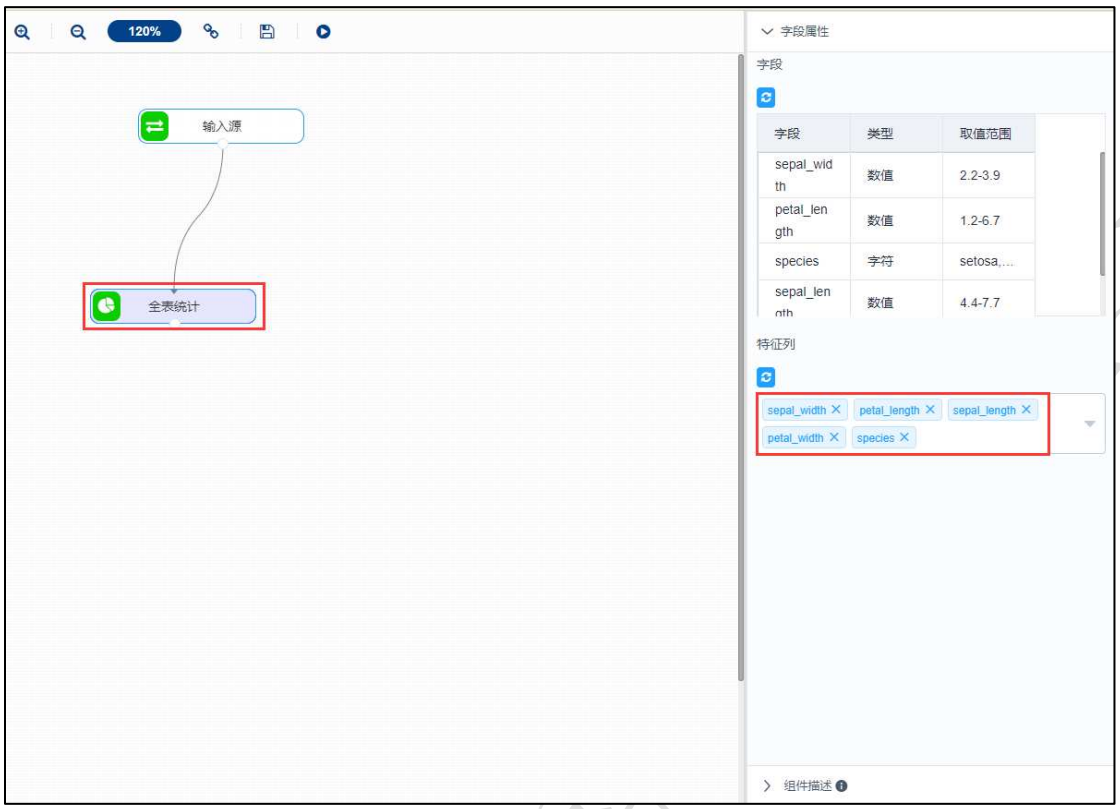


图 203

col	count	mean	std
sepal_width	150	3.06	0.44
petal_length	150	3.76	1.77
sepal_length	150	5.84	0.83
petal_width	150	1.2	0.76

图 204

3.4.4 回归

3.4.4.1 线性回归

图标： ◎ 线性回归

描述：线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量

关系的一种统计分析方法。

字段属性

特征列：请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。

标签列：请选择数值型数据。

输出

表结果：线性回归预测结果。

报告：模型拟合效果。

示例

下列对某数据进行线性回归算法：

- 选择自变量，因变量，均选择数值型数据。如图 205 所示。
- 运行成功后，选择查看数据，如图 206 所示。
- 运行成功后，选择查看报告，如图 207 所示。

图 205

sepal_length	sepal_width	predict_label
5.1	3.5	3.103
4.9	3	3.116
4.7	3.2	3.128
4.6	3.1	3.134
5	3.6	3.11
5.4	3.9	3.085
4.6	3.4	3.134
5	3.4	3.11

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 206

图 207

3.4.4.2 广义最小二乘法

图标： 

描述：广义最小二乘法是一种常见的消除异方差的方法。它的主要思想是为解释变量加上一个权重，从而使得加上权重后的回归方程方差是相同的。因此在 GLS 方法下我们可以得到估计量的无偏和一致估计。

字段属性：

自变量：选择自变量所在列，请选择数值型数据。

因变量：选择响应变量所在的列，请选择数值型数据。

输出

表结果：无。

报告：GLS Regression Results。

示例

下列对某数据使用广义最小二乘法：

- 选择自变量，因变量，均选择数值型数据。如图 208 所示。
- 运行成功后，选择查看报告，如图 209、图 210 所示。

图 208

算法运行报告

广义最小二乘法结果

结果

GLS Regression Results

Dep. Variable:	sepal_width	R-squared:	0.957
Model:	GLS	Adj. R-squared:	0.956
Method:	Least Squares	F-statistic:	3277.
Date:	Wed, 07 Mar 2018	Prob (F-statistic):	2.42e-103
Time:	15:43:20	Log-Likelihood:	-146.83
No. Observations:	150	AIC:	295.7
Df Residuals:	149	BIC:	298.7
Df Model:	1		
Covariance Type:	nonrobust		

图 209

算法运行报告

Df Model: 1

Covariance Type: nonrobust

	coef	std err	t	P>|t|	[0.025	0.975]
sepal_length	0.5118	0.009	57.246	0.000	0.494	0.529

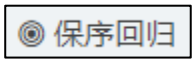
Omnibus:	17.098	Durbin-Watson:	0.433
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7.933
Skew:	0.352	Prob(JB):	0.0189
Kurtosis:	2.121	Cond. No.	1.00

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

图 210

3.4.4.3 保序回归

图标： 

描述：保序回归是定了一个无序的数字序列，通过修改其中元素的值，得到一个非递减的数字序列，要求是使得误差（预测值和实际值差的平方）最小。

字段属性

自变量：请选择数值型数据。

因变量：请选择数值型数据。

输出

模型

表结果：保序回归算法结果。

报告：模型拟合效果、Isotonic regression。

示例

下列对某数据进行保序回归算法：

- 选择自变量，因变量，均选择数值型数据。图 211 所示。
- 运行成功后，选择查看数据，图 212 所示。
- 运行成功后，选择查看报告，如图 213、图 214 所示。
- 模型预测配置如图 215 所示。
- 模型预测结果如图 216 所示。

图 211

预览数据

perf	chmin	predict_label
397	52	22.5
636	16	22.5
36	3	2.611
18	1	1.05
45	3	2.611
80	1	2.611
22	1	1.4
248	12	10.909

共 156 条 25 条/页 1 2 3 4 5 6 7 前往 1 页

图 212

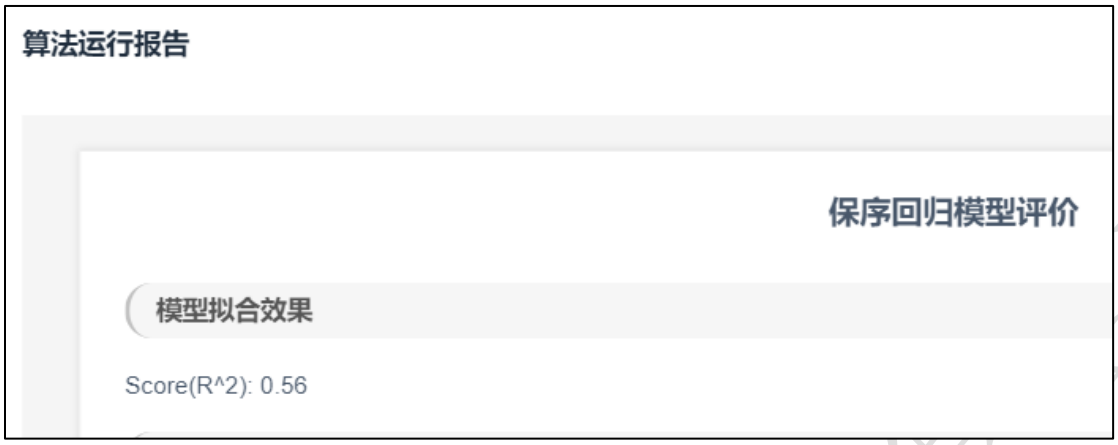

算法运行报告
保序回归模型评价
模型拟合效果
Score(R^2): 0.56

图 213

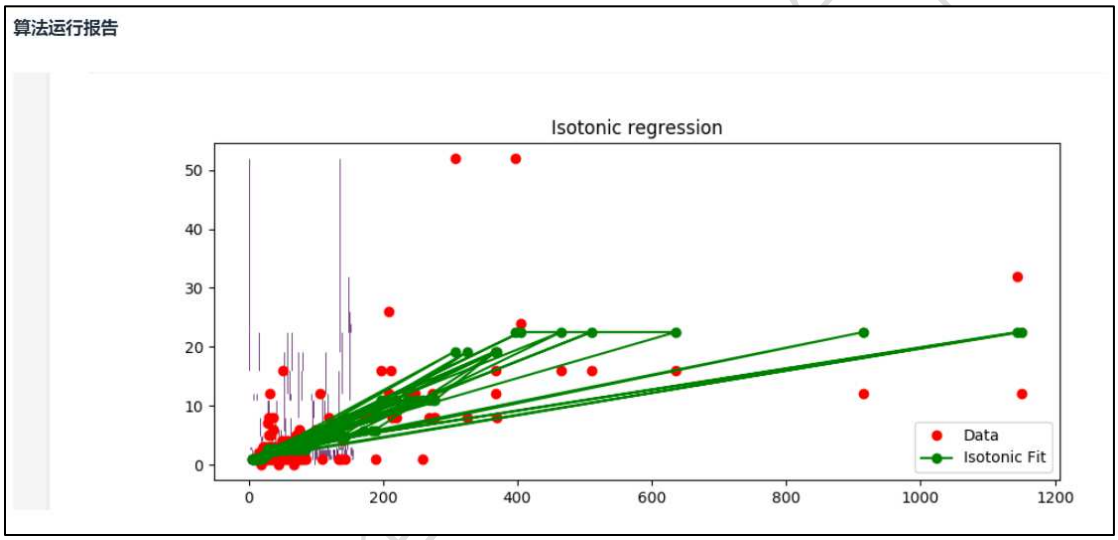

算法运行报告
Isotonic regression
50
40
30
20
10
0
0
200
400
600
800
1000
1200
Data
Isotonic Fit

图 214

图 215

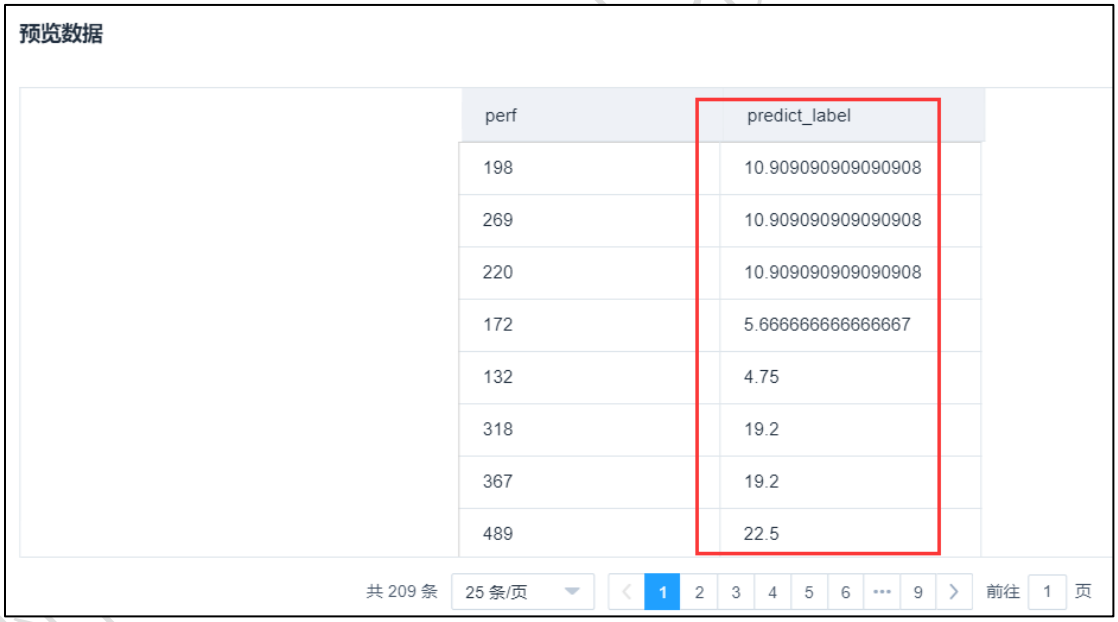


图 216

3.4.4.4 岭回归

图标： ◎ 岭回归

描述：岭回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小

二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。

字段属性

特征列：通过勾选的方式选择特征所在列。。

标签列：选择响应变量所在的列。

输出

模型

表结果：岭回归算法结果。

报告：模型拟合效果。

示例

下列对某数据进行岭回归算法：

- 选择自变量，因变量，均选择数值型数据，如图 217 所示。
- 运行成功后，选择查看数据，如图 218 所示。
- 运行成功后，选择查看报告，如图 219 所示。
- 模型预测配置如图 220 所示。
- 模型预测结果如图 221 所示。


120%
输入源
数据拆分
岭回归
输入源
模型预测
字段属性
添加字段过滤字符串
字段
类型
取值范围
educational_level
数值
1-5
gender
数值
0-1
family_size
数值
1-8
opening_mon
数
因变量
educational_level

图 217

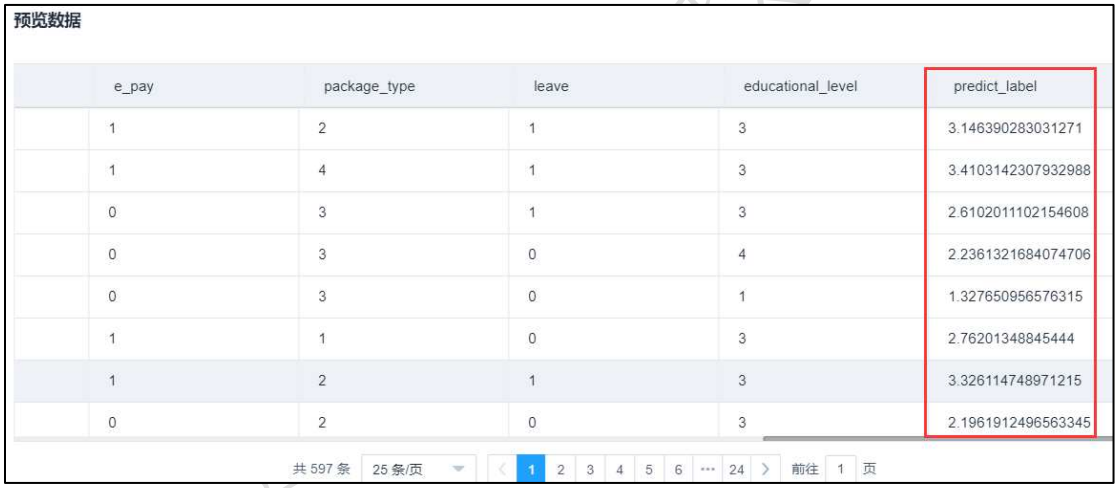

预览数据
e_pay
package_type
leave
educational_level
predict_label
1 2 1 3 3.146390283031271
1 4 1 3 3.4103142307932988
0 3 1 3 2.6102011102154608
0 3 0 4 2.2361321684074706
0 3 0 1 1.327650956576315
1 1 0 3 2.76201348845444
1 2 1 3 3.326114748971215
0 2 0 3 2.1961912496563345
共 597 条
25 条/页
1 2 3 4 5 6 ··· 24
前往 1 页

图 218


算法运行报告
岭回归训练结果
模型拟合效果
Coefficients(系数): 2.7217171635838056
Residual sum of squares(均方误差): 2.7217171635838056
Score(R^2): 2.7217171635838056

图 219

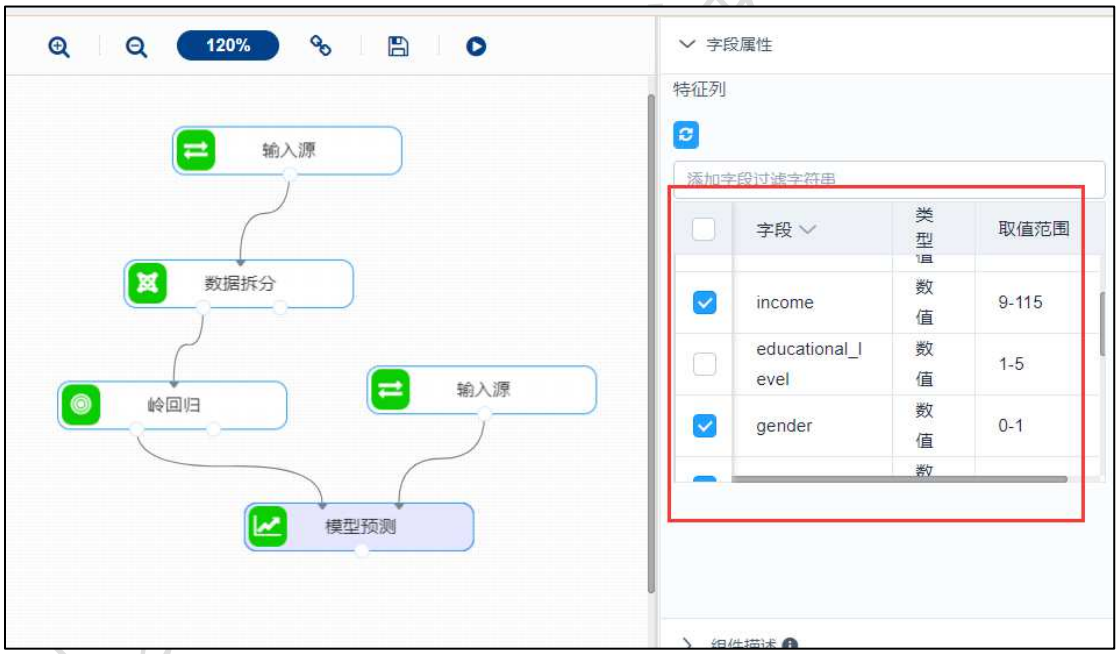
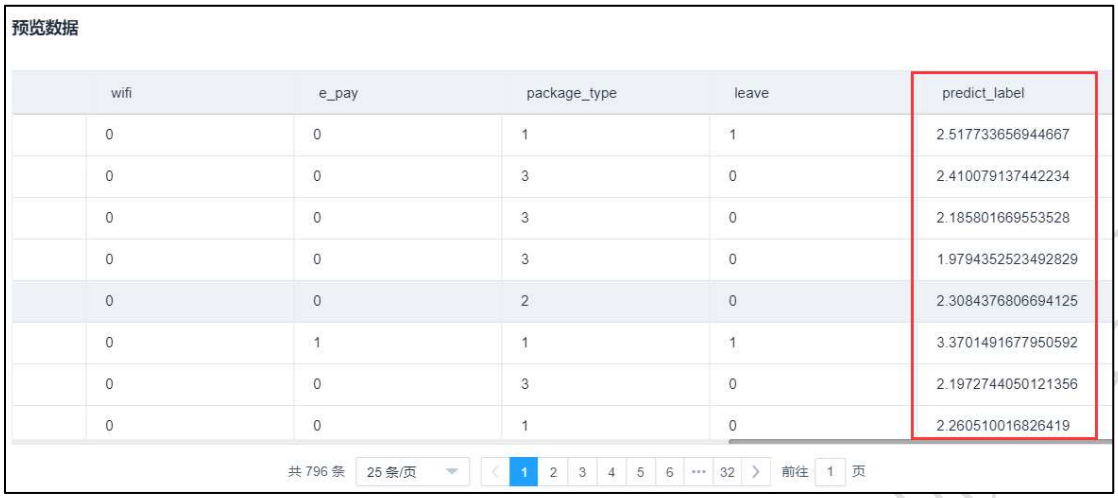

120%
输入源
数据拆分
岭回归
输入源
模型预测
字段属性
特征列
添加字段过滤字符串
字段
类型
取值范围
income
数值
9-115
educational_level
数值
1-5
gender
数值
0-1

图 220


预览数据

	wifi	e_pay	package_type	leave	predict_label
	0	0	1	1	2.517733656944667
	0	0	3	0	2.410079137442234
	0	0	3	0	2.185801669553528
	0	0	3	0	1.9794352523492829
	0	0	2	0	2.3084376806694125
	0	1	1	1	3.3701491677950592
	0	0	3	0	2.1972744050121356
	0	0	1	0	2.260510016826419

共 796 条 25 条/页 1 2 3 4 5 6 ··· 32 前往 1 页

图 221

3.4.4.5 CART 回归树

图标： 

描述： 使用 Cart 决策树算法的回归树。

字段属性

特征列：通过勾选的方式选择特征所在列，仅支持数值型数据。

标签列：选择响应变量所在的列，仅支持数值型数据。

参数设置

切分时的评价准则：包括均方误差、平均绝对误差，默认均方误差。

切分原则：包括选择最优的切分、随机切分，默认选择最优的切分。

输出

模型

表结果：CART 回归树算法结果。

报告：Regression model evaluation。

示例

下列对某数据进行 CART 回归树算法：

- 选择自变量，因变量，均选择数值型数据。如图 222 所示。
- 保留默认参数，切分时的评价准则为均方误差，切分原则为选择最优的切分，如图

223 所示。

- 运行成功后，选择查看数据，如图 224 所示。
- 运行成功后，选择查看报告，如图 225 所示。
- 模型评估配置如图 226 所示。
- 模型评估结果如图 227 所示.
- 模型评估报告如图 228 所示。
- 模型预测配置如图 229 所示。
- 模型预测结果如图 230 所示。

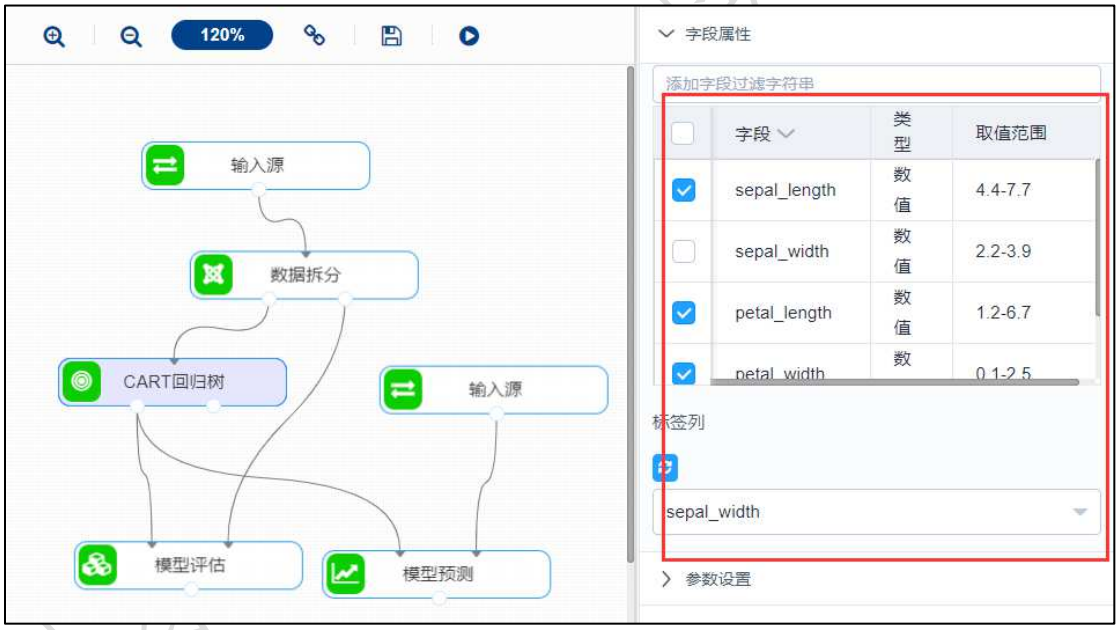


图 222

图 223

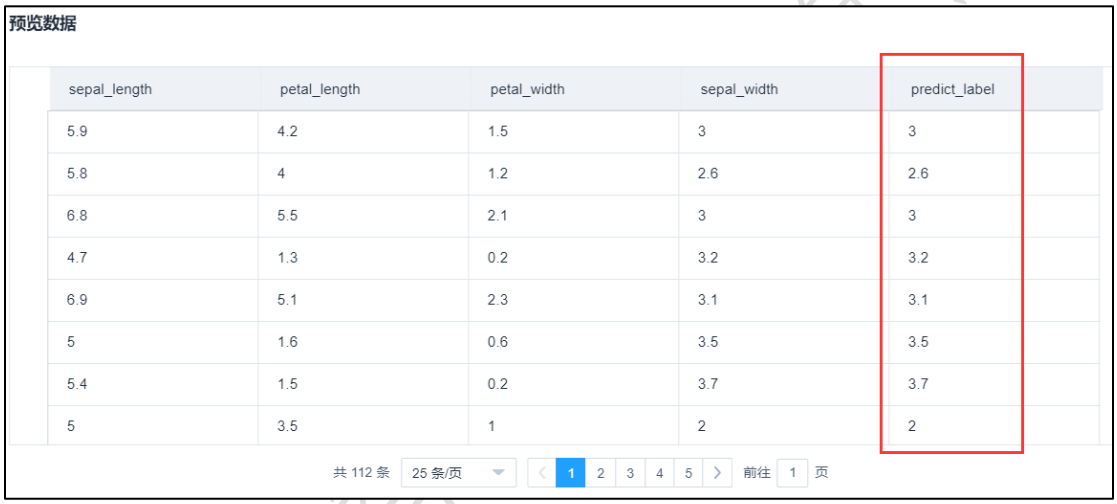


图 224

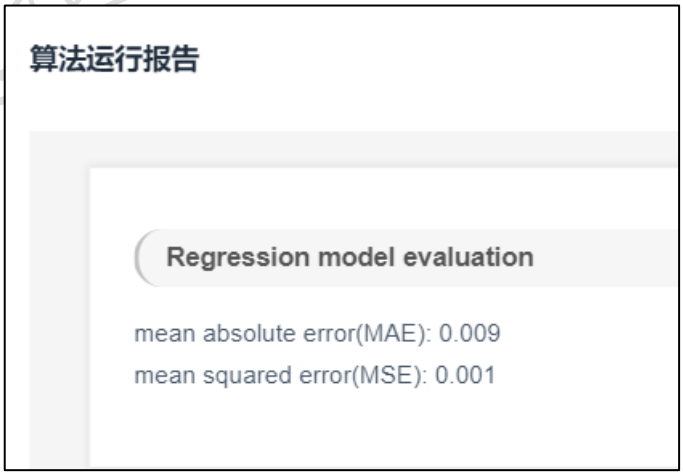


图 225


120%
输入源
数据拆分
CART回归树
输入源
模型评估
模型预测
字段属性
添加字段过滤字符串
字段
类型
取值范围
sepal_length
值
petal_length
数值
1.2-6.7
petal_width
数值
0.1-2.5
predict_label
字符
标签列
sepal_width

图 226

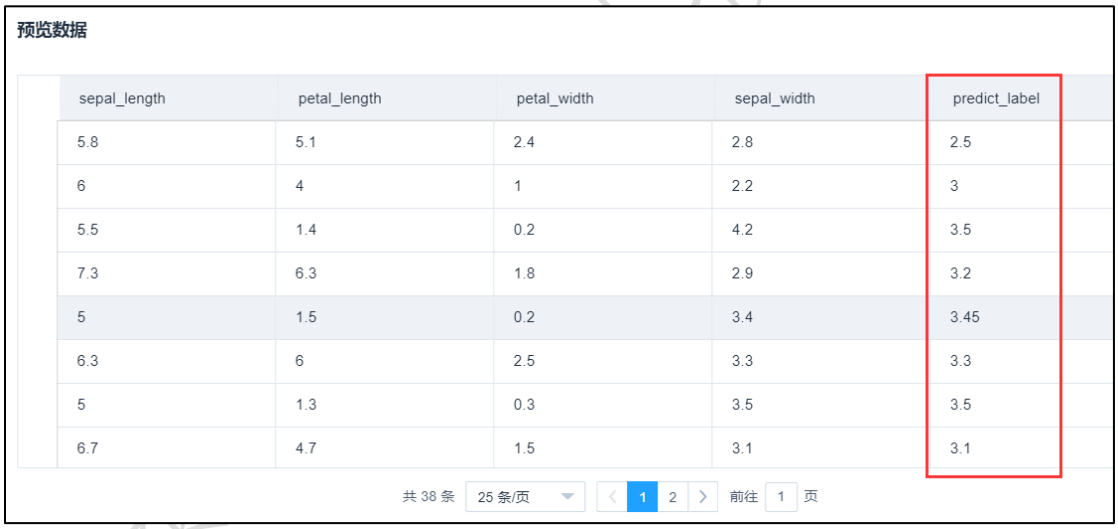

预览数据
sepal_length
petal_length
petal_width
sepal_width
predict_label
5.8 5.1 2.4 2.8 2.5
6 4 1 2.2 3
5.5 1.4 0.2 4.2 3.5
7.3 6.3 1.8 2.9 3.2
5 1.5 0.2 3.4 3.45
6.3 6 2.5 3.3 3.3
5 1.3 0.3 3.5 3.5
6.7 4.7 1.5 3.1 3.1
共 38 条
25 条/页
1
2
前往
1
页

图 227

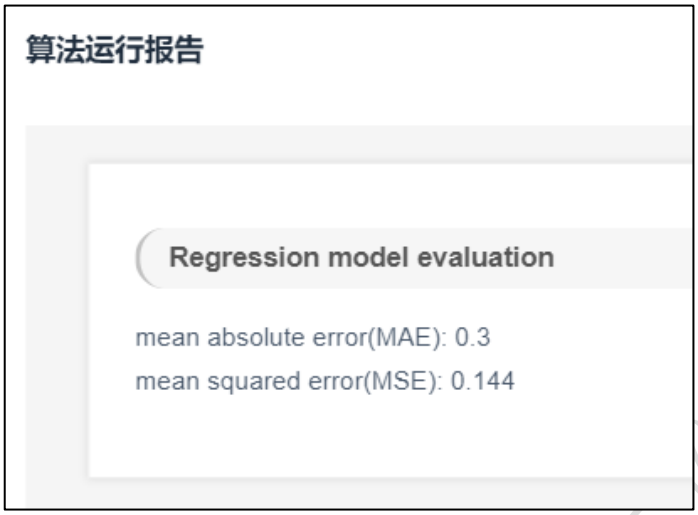

算法运行报告
Regression model evaluation
mean absolute error(MAE): 0.3
mean squared error(MSE): 0.144

图 228


120%
输入源
数据拆分
CART回归树
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
类型
取值范围
sepalwidth
数值
2-4
petallength
数值
1-7
species
字符
setosa,vers...
sepallength
数
4-8

图 229

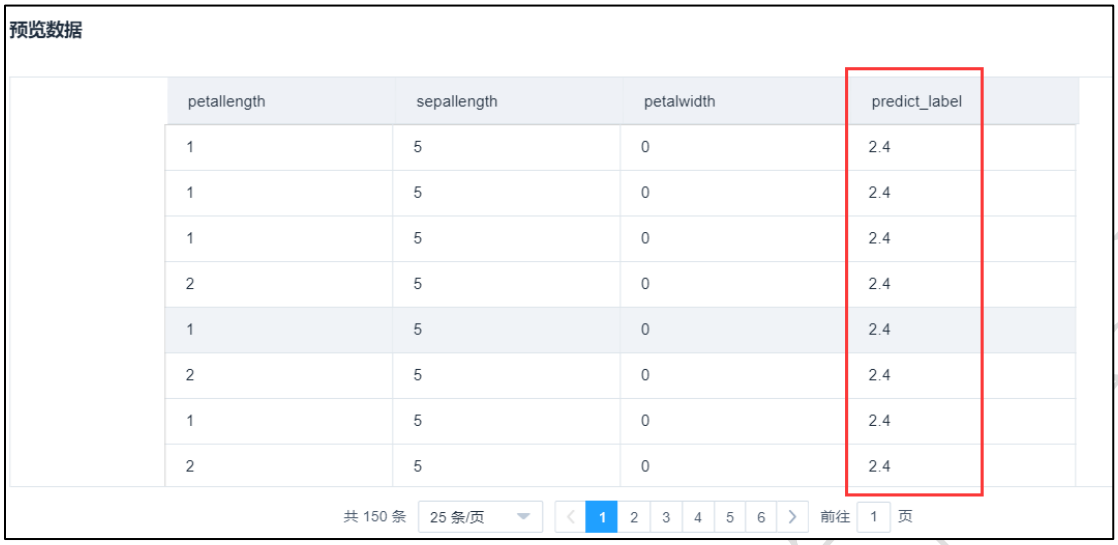
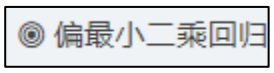


图 230

3.4.4.6 偏最小二乘回归

图标： 

描述：偏最小二乘回归（PLSR）是一种多因变量 Y 对多自变量 X 的回归建模方法，该算法在建立回归的过程中，既考虑了尽量提取 Y 和 X 中的主成分（PCA—Princ ipal Component Analysis，主成分分析的思想），又考虑了使分别从 X 和 Y 提取出的主成分之间的相关性最大化（CCA 的思想）。简单的说，PLSR 是 PCA、CCA 和多元线性回归这三种基本算法组合的产物。

字段属性

自变量：请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。

因变量：请选择数值型数据。

参数设置

保留的主成分数量：整数型，默认为 2。

是否归一化数据：是/否:，默认是。

最大迭代数：整数型，默认 500。

输出

表结果：偏最小二乘回归算法结果。

报告：Regression model evaluation。

示例

下列对某数据进行偏最小二乘回归算法：

- 选择自变量，因变量。如图 231 所示。
- 保留默认参数，保留的主成分数量为 2，设置“是否归一数据”为是，最大迭代数为 500，如图 232 所示。
- 运行成功后，选择查看数据，如图 233 所示。
- 运行成功后，选择查看报告，如图 234 所示。
- 模型评估配置如图 235 所示。
- 模型评估结果数据如图 236 所示。
- 模型评估运行报告如图 237 所示。
- 模型预测配置如图 238 所示。
- 模型预测结果数据如图 239 所示。

图 231


字段属性
参数设置
保留主成分数量
2
是否归一化数据
是
最大迭代数
500

图 232

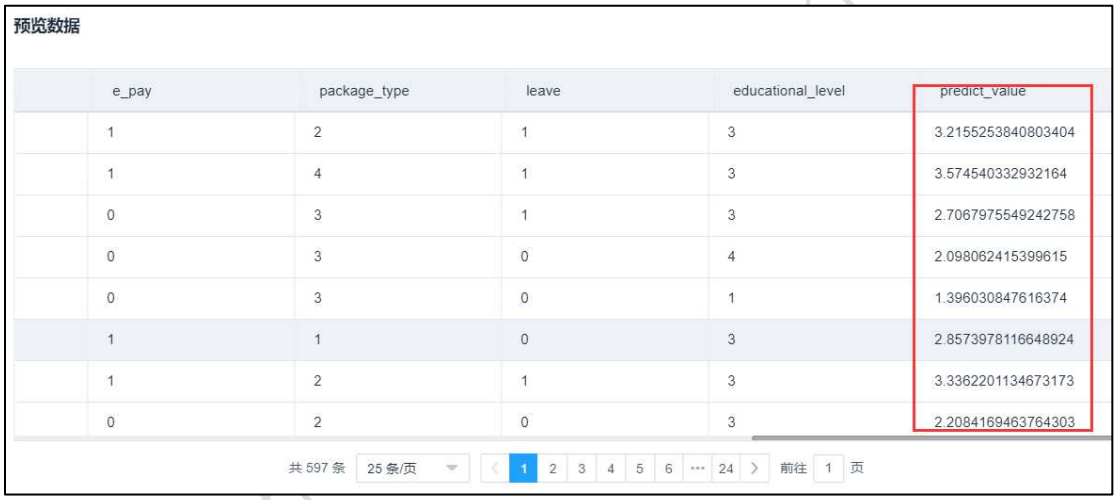

预览数据
e_pay
package_type
leave
educational_level
predict_value
1 2 1 3 3.2155253840803404
1 4 1 3 3.574540332932164
0 3 1 3 2.7067975549242758
0 3 0 4 2.098062415399615
0 3 0 1 1.396030847616374
1 1 0 3 2.8573978116648924
1 2 1 3 3.3362201134673173
0 2 0 3 2.2084169463764303
共 597 条
25 条/页
前往 1 页

图 233

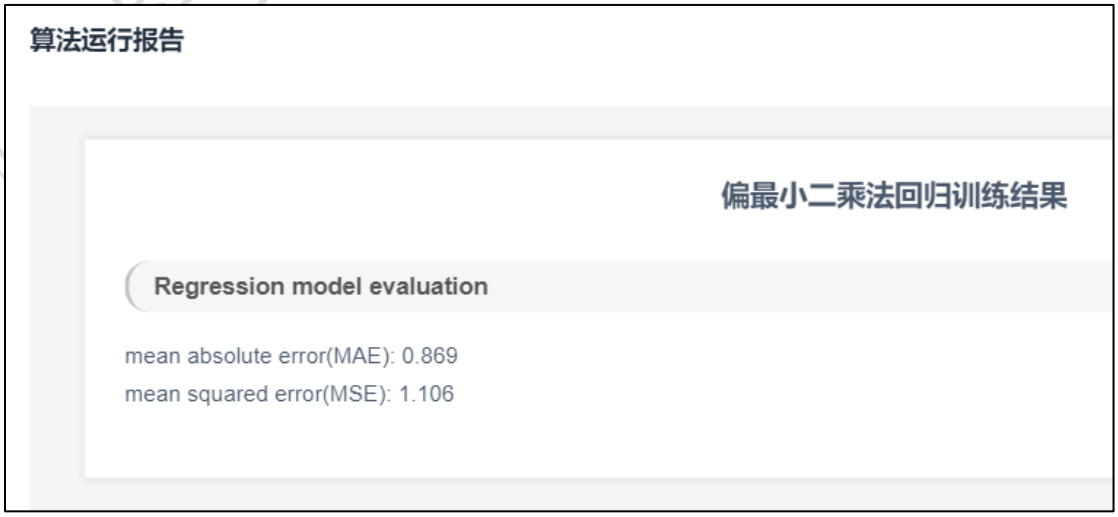

算法运行报告
偏最小二乘法回归训练结果
Regression model evaluation
mean absolute error(MAE): 0.869
mean squared error(MSE): 1.106

图 234

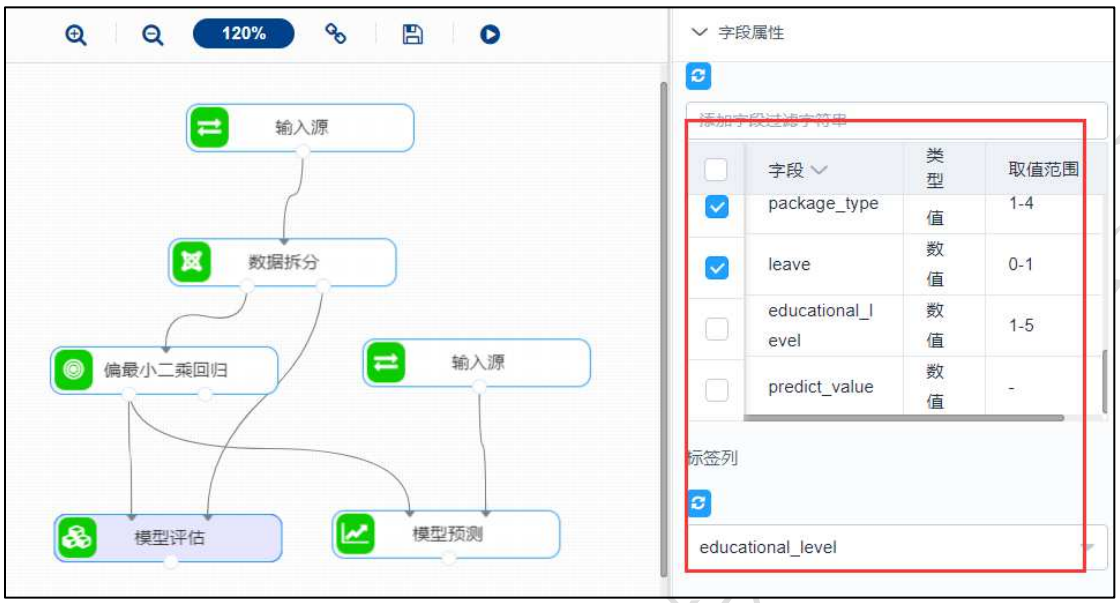


图 235

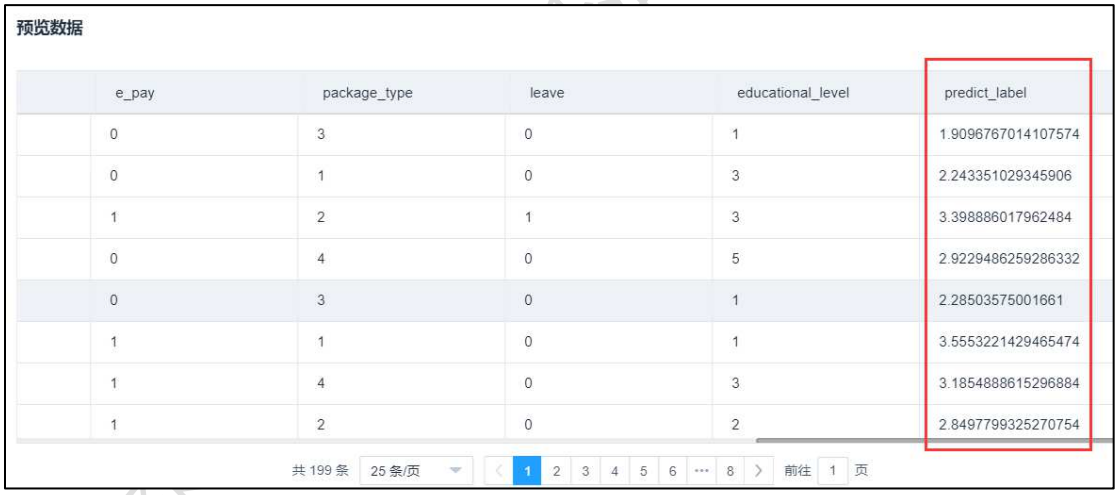


图 236

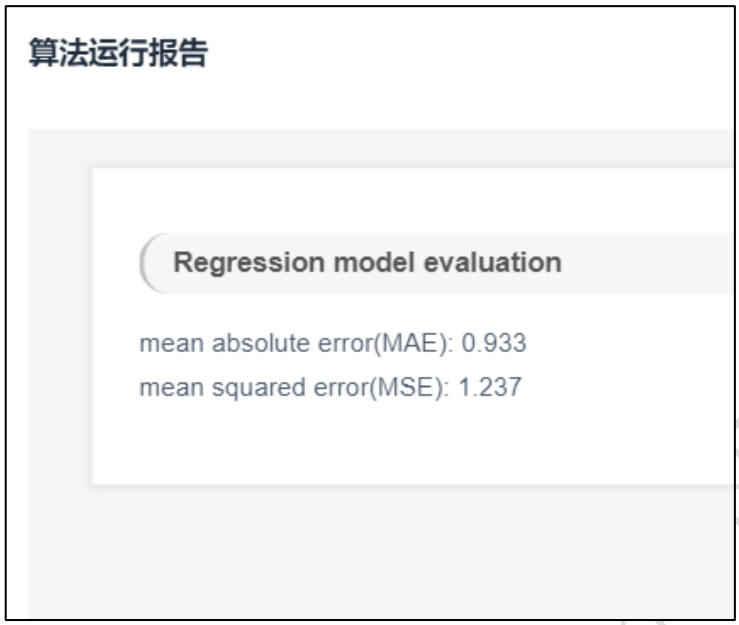

算法运行报告
Regression model evaluation
mean absolute error(MAE): 0.933
mean squared error(MSE): 1.237

图 237


120%
输入源
数据拆分
偏最小二乘回归
输入源
模型评估
模型预测
字段属性
特征列
字段
类型
取值范围
marital_status
数值
0-1
educational_level
数值
1-5
gender
数值
0-1

图 238

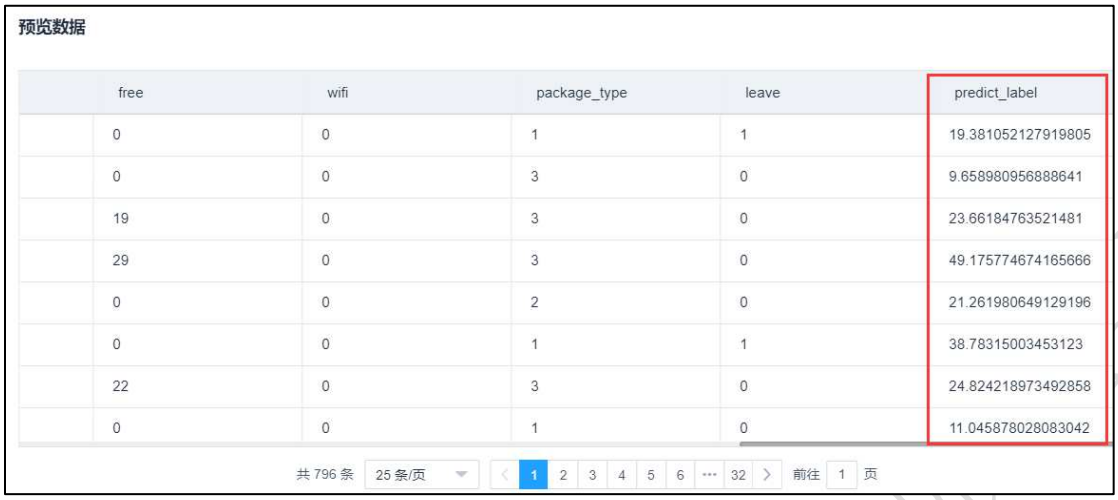


图 239

3.4.4.7 Lasso 回归

图标： 

描述： Lasso 回归是一个用于估计稀疏参数的线性模型，特别适用于参数数目缩减。

字段属性

特征列：通过勾选的方式选择特征所在列。

标签列：选择响应变量所在的列。

参数设置

alpha：浮点型，默认 1.0。

最大迭代次数：整数型，默认 500

输出

表结果：Lasso 回归预测结果。

报告：模型拟合效果。

示例

下列对某数据进行 Lasso 回归算法：

- 选择自变量，因变量。如图 240 所示。
- 保留默认参数，alpha 为 1.0，最大迭代次数为 500，如图 241 所示。

- 运行成功后，选择查看数据，如图 242 所示。
- 运行成功后，选择查看报告，如图 243 所示。
- 模型预测配置如图 244 所示。
- 模型预测结果数据如图 245 所示。

图 240

图 241

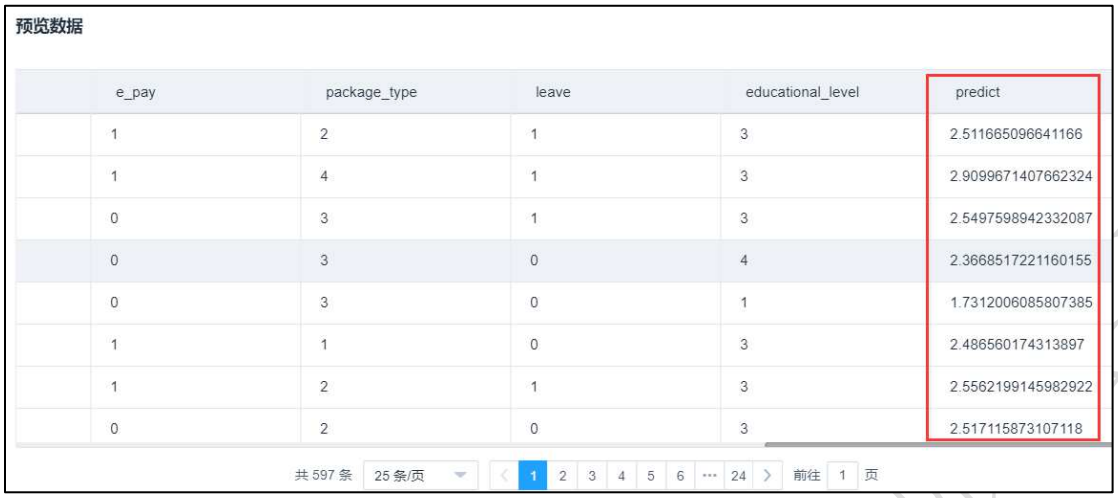

预览数据
e_pay
package_type
leave
educational_level
predict
1 2 1 3 2.511665096641166
1 4 1 3 2.9099671407662324
0 3 1 3 2.5497598942332087
0 3 0 4 2.3668517221160155
0 3 0 1 1.7312006085807385
1 1 0 3 2.486560174313897
1 2 1 3 2.5562199145982922
0 2 0 3 2.517115873107118
共 597 条
25 条/页
1 2 3 4 5 6 ··· 24
前往 1 页

图 242


算法运行报告
Lasso回归模型评价
模型拟合效果
intercept(截距项): 2.8317241284955323
Coefficients(系数): [-0. -0.00658748 -0. 0.00176153 -0. -0. -0.00472689 0. -0. -0.00698292 0.02021693 0. 0. 0.]
Residual sum of squares(均方误差): [-0. -0.00658748 -0. 0.00176153 -0. -0. -0.00472689 0. -0. -0.00698292 0.02021693 0. 0. 0.]
Score(R^2): [-0. -0.00658748 -0. 0.00176153 -0. -0. -0.00472689 0. -0. -0.00698292 0.02021693 0. 0. 0.]

图 243

图 244

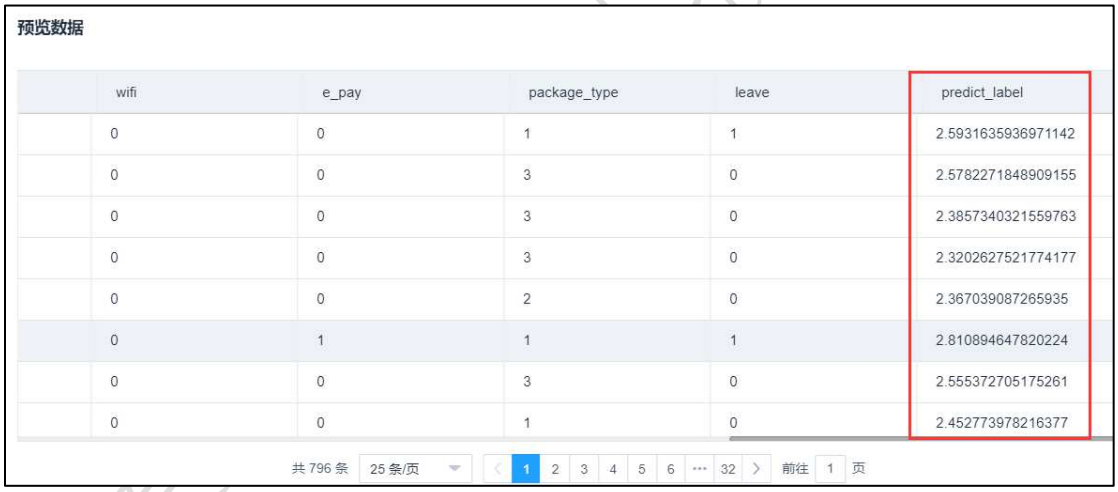


图 245

3.4.4.8 多项式回归

图标： 多项式回归

描述： 多项式回归是研究一个因变量与一个或多个自变量间多项式的回归分析方法。

字段属性

特征列：通过勾选的方式选择特征所在列。如图 246 所示。

图 246

参数设置

degree：多项式的阶数，默认为 2。

interaction_only：是否产生相互影响的特征集，默认为 False。

include_bias：是否包含偏差列，默认为 True。

如图 247 所示。

图 247

输出

表结果：特征矩阵。

报告：无。

示例

下面对某两列数据拟合多项式回归。

- 选择两列待拟合序列，数据必须为数值型。如图 248 所示。
- 点击参数设置，设置如图 249 所示。
- 运行该组件，对组件右击，选择查看报告，结果如图 250 所示。

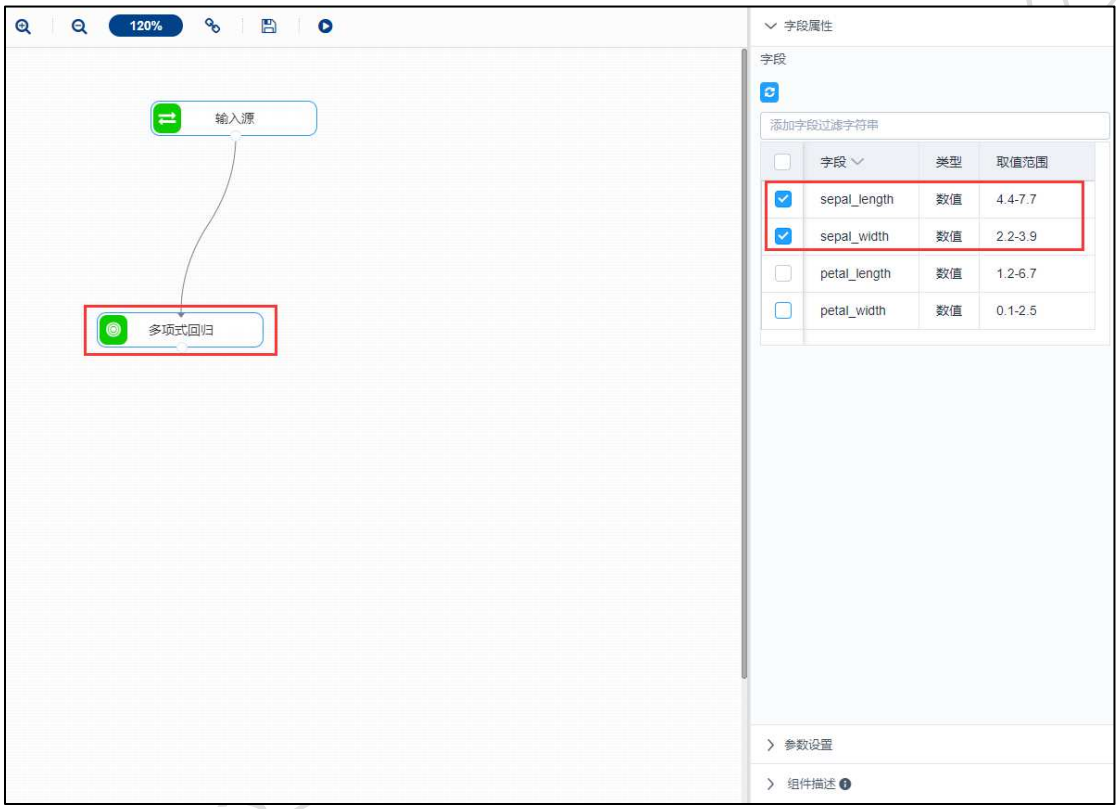


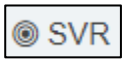
图 248

图 249

0	1	2	3
1	5.1	3.5	17.85
1	4.9	3	14.7
1	4.7	3.2	15.04
1	4.6	3.1	14.26
1	5	3.6	18
1	5.4	3.9	21.06
1	4.6	3.4	15.64
1	5	3.4	17

图 250

3.4.4.9 SVR

图标： 

描述：SVR（支持向量回归）是使用支持向量机解决回归问题。支持向量回归假设我们能容忍的 f(x)与之间最多有 ε 的偏差，当且仅当 f(x)与 y 的差别绝对值大于 ε 时，才计算损失，此时相当于以 f(x)为中心，构建一个宽度为 2 ε 的间隔带，若训练样本落入此间隔带，则认为是被预测正确的。

字段属性

特征列：通过勾选的方式选择特征所在列。

标签列：选择分类标签所在的列。

参数设置

罚项系数：浮点型，默认 1.0。

核函数：支持线性核、多项式核、高斯核、sigmoid，默认高斯核。

输出

表结果：SVR 回归算法结果。

报告：Regression model evaluation。

示例

下列对某数据进行 SVR 回归算法：

- 选择自变量，因变量。如图 251 所示。
- 保留默认参数，罚项系数为 1.0，核函数为高斯核，如图 252 所示。
- SVR 运行成功后，选择查看数据，如图 253 所示。
- SVR 成功后，选择查看报告，如图 254 所示。
- 模型评估配置如图 255 所示。
- 模型评估运行成功后，选择查看数据，如图 256 所示。
- 模型评估运行成功后，选择查看报告，如图 257 所示。
- 模型预测配置如图 258 所示。
- 模型预测运行成功后，选择查看数据，如图 259 所示。

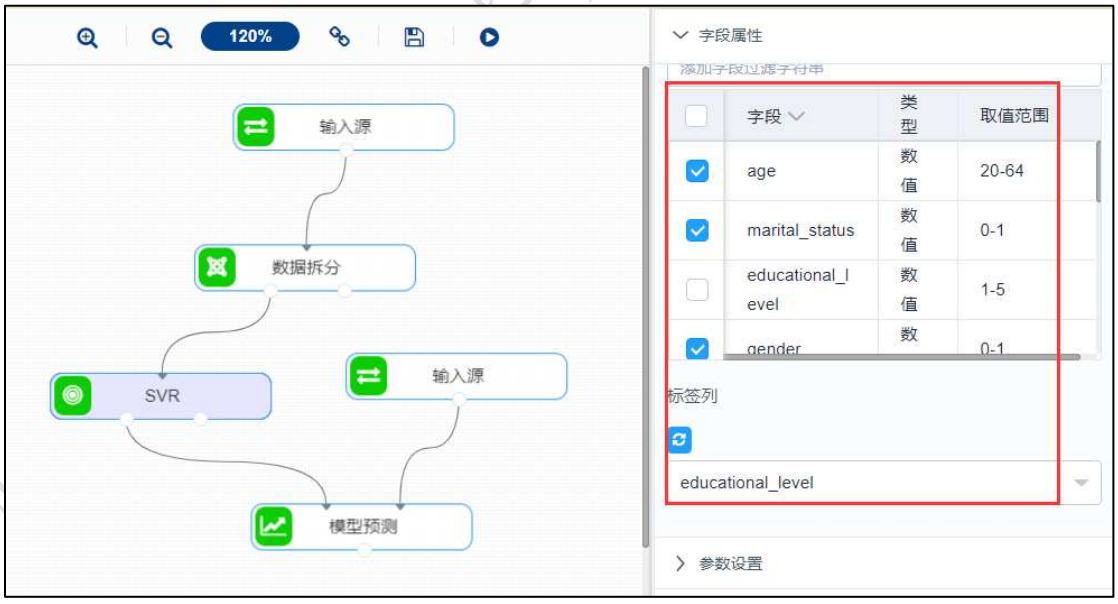


图 251

图 252

预览数据

	wifi	package_type	leave	educational_level	predict_value
	0	2	1	3	2.900089350037783
	28	4	1	3	2.900165396438463
	0	3	1	3	2.899454378940668
	0	3	0	4	3.640831314689791
	0	3	0	1	1.6390969989871498
	0	1	0	3	2.9000916983609377
	0	2	1	3	2.899892927341849
	0	2	0	3	2.900096154261487

共 597 条 25 条/页 1 2 3 4 5 6 ··· 24 前往 1 页

图 253

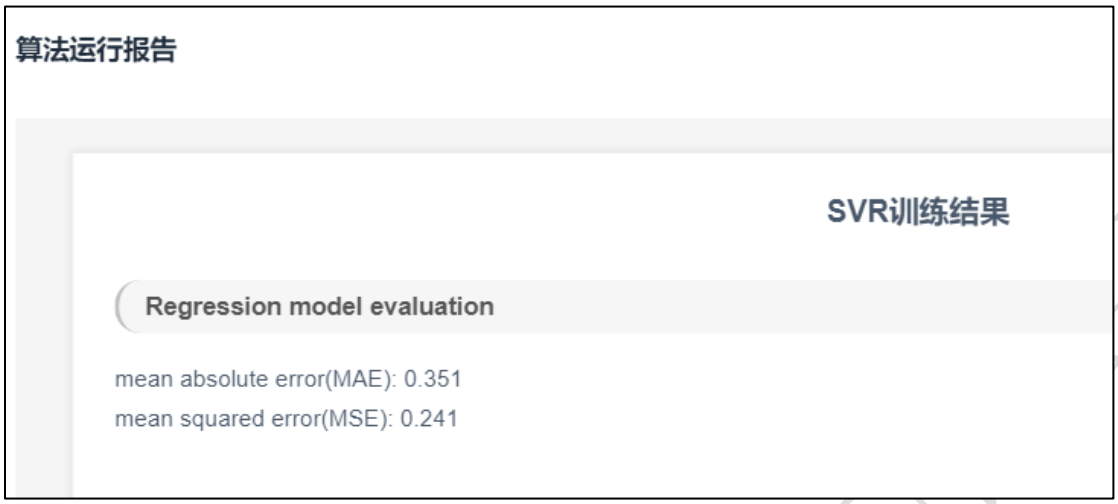
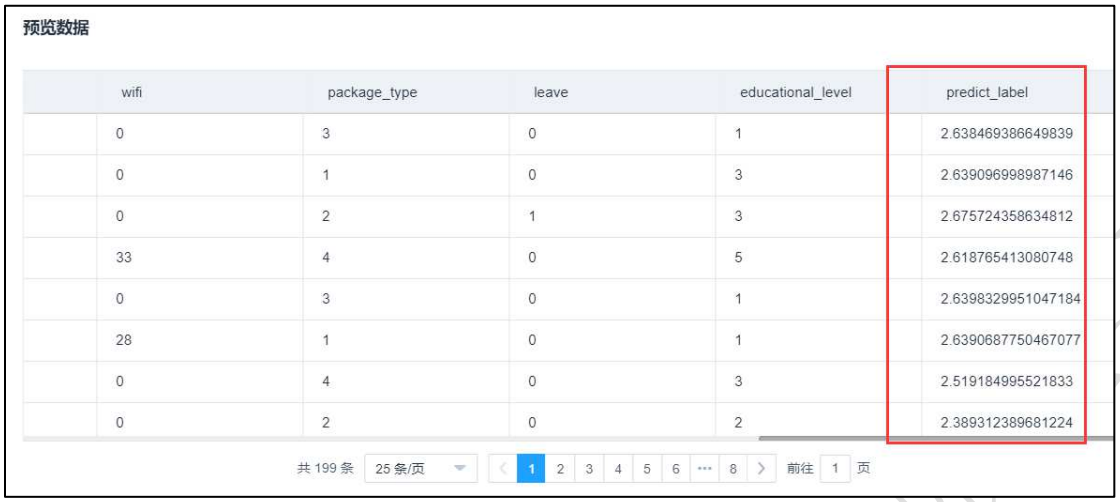

算法运行报告
SVR训练结果
Regression model evaluation
mean absolute error(MAE): 0.351
mean squared error(MSE): 0.241

图 254


120%
输入源
数据拆分
SVR
输入源
模型评估
模型预测
字段属性
字段
类型
取值范围
age
数值
20-64
marital_status
数值
0-1
educational_level
数值
1-5
gender
数
0-1
标签列
educational_level
组件描述

图 255


预览数据

	wifi	package_type	leave	educational_level	predict_label
	0	3	0	1	2.638469386649839
	0	1	0	3	2.639096998987146
	0	2	1	3	2.675724358634812
	33	4	0	5	2.618765413080748
	0	3	0	1	2.6398329951047184
	28	1	0	1	2.6390687750467077
	0	4	0	3	2.519184995521833
	0	2	0	2	2.389312389681224

共 199 条 25 条/页 1 2 3 4 5 6 ··· 8 前往 1 页

图 256

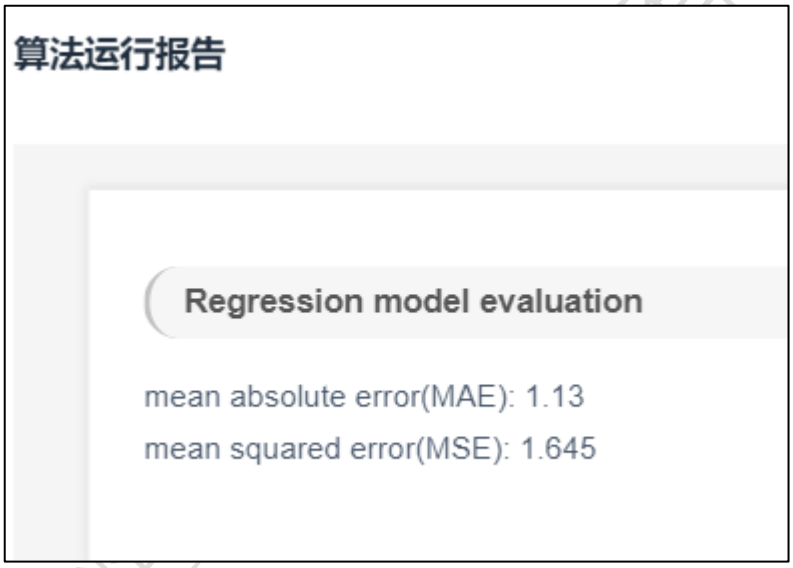


图 257

图 258

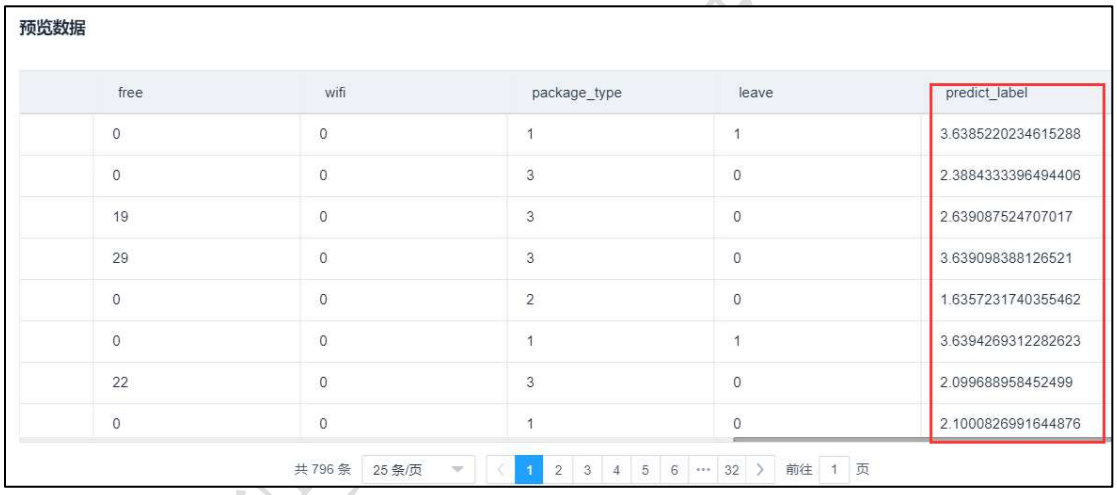


图 259

3.4.4.10 KNN 回归

图标： KNN回归

描述： KNN 进行回归

字段属性：

特征列：通过勾选的方式选择特征所在列。

标签列：仅支持数值型数据。

参数设置：

最近邻个数 K：整数型，通常不大于 20，默认 5.

投票权重类型：权重相等或权重与距离成反比，默认权重相等。

计算最近邻的算法：包括 自动、BallTree、KDTree、暴力搜索法，默认自动。

输出

表结果：KNN 回归算法结果。

报告：Regression model evaluation。

示例

下列对某数据进行 KNN 回归算法：

- 选择自变量，因变量。如图 260 所示。
- 保留默认参数，最近邻个数为 5，投票权重类型为权重相等，计算最近邻的计算为自动，如图 261 所示。
- 运行成功后，选择查看数据，如图 262 所示。
- 运行成功后，选择查看报告，如图 263 所示。
- 模型评估配置如图 264 所示。
- 模型评估运行成功后，选择查看数据，如图 265 所示。
- 模型评估运行成功后，选择查看报告，如图 266 所示。
- 模型预测配置如图 267 所示。
- 模型预测运行成功后，选择查看数据，如图 268 所示。


120%
输入源
数据拆分
KNN
输入源
模型评估
模型预测
字段属性
字段
类型
取值范围
educational_level
数值
1-5
gender
数值
0-1
family_size
数值
1-8
opening_mon
数
1-72
标签列
educational_level

图 260


字段属性
参数设置
最近邻个数K
5
投票权重类型
权重相等
计算最近邻的算法
自动

图 261

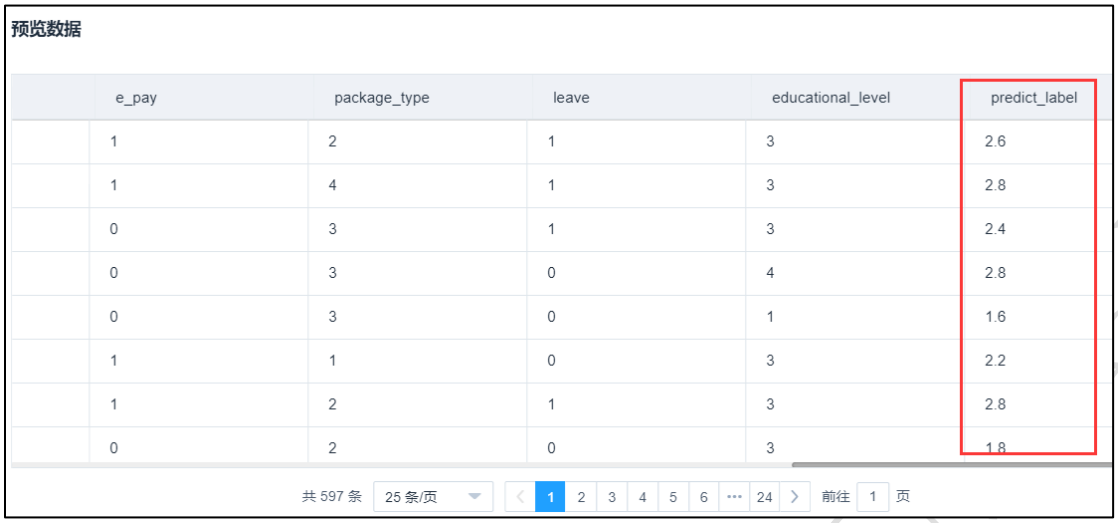

预览数据
e_pay
package_type
leave
educational_level
predict_label
1 2 1 3 2.6
1 4 1 3 2.8
0 3 1 3 2.4
0 3 0 4 2.8
0 3 0 1 1.6
1 1 0 3 2.2
1 2 1 3 2.8
0 2 0 3 1.8
共 597 条
25 条/页
1 2 3 4 5 6 ··· 24
前往 1 页

图 262

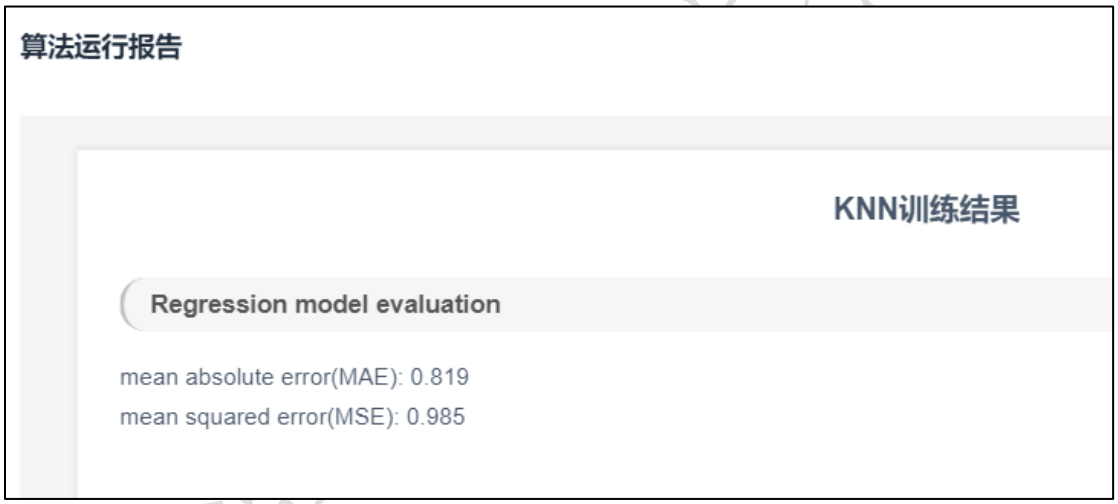

算法运行报告
KNN训练结果
Regression model evaluation
mean absolute error(MAE): 0.819
mean squared error(MSE): 0.985

图 263

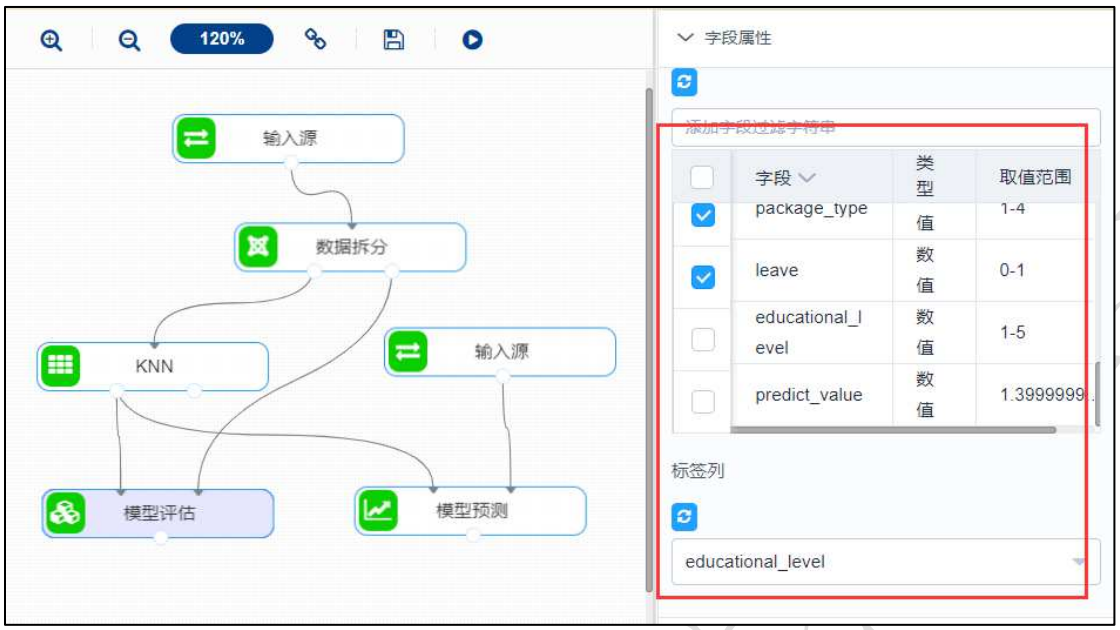

120%
输入源
数据拆分
KNN
输入源
模型评估
模型预测
字段属性
添加字段过滤字符串
字段
类型
取值范围
package_type
值
1-4
leave
数值
0-1
educational_level
数值
1-5
predict_value
数值
1.3999999
标签列
educational_level

图 264

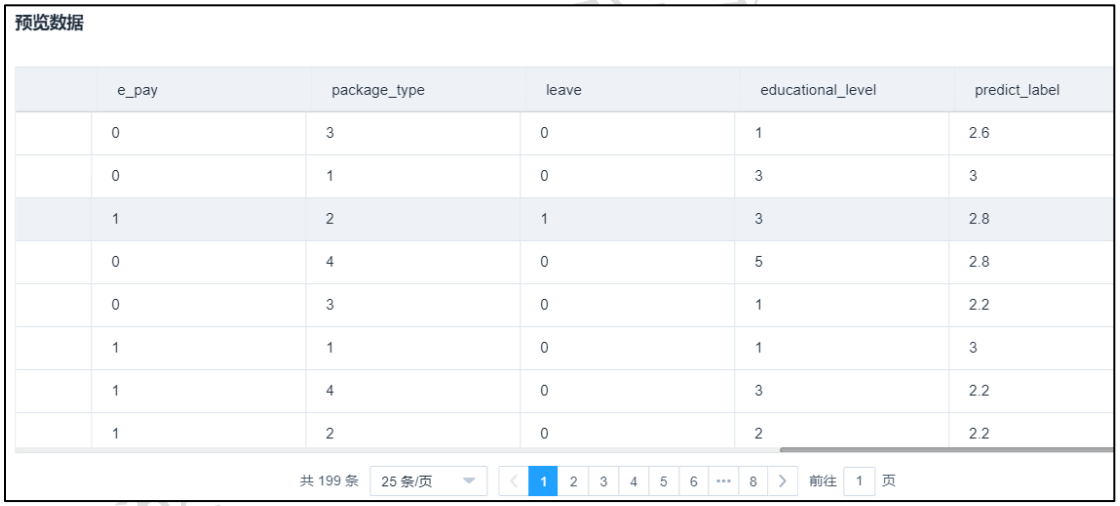

预览数据
e_pay
package_type
leave
educational_level
predict_label
0 3 0 1 2.6
0 1 0 3 3
1 2 1 3 2.8
0 4 0 5 2.8
0 3 0 1 2.2
1 1 0 1 3
1 4 0 3 2.2
1 2 0 2 2.2
共 199 条
25 条/页
1 2 3 4 5 6 ··· 8
前往 1 页

图 265

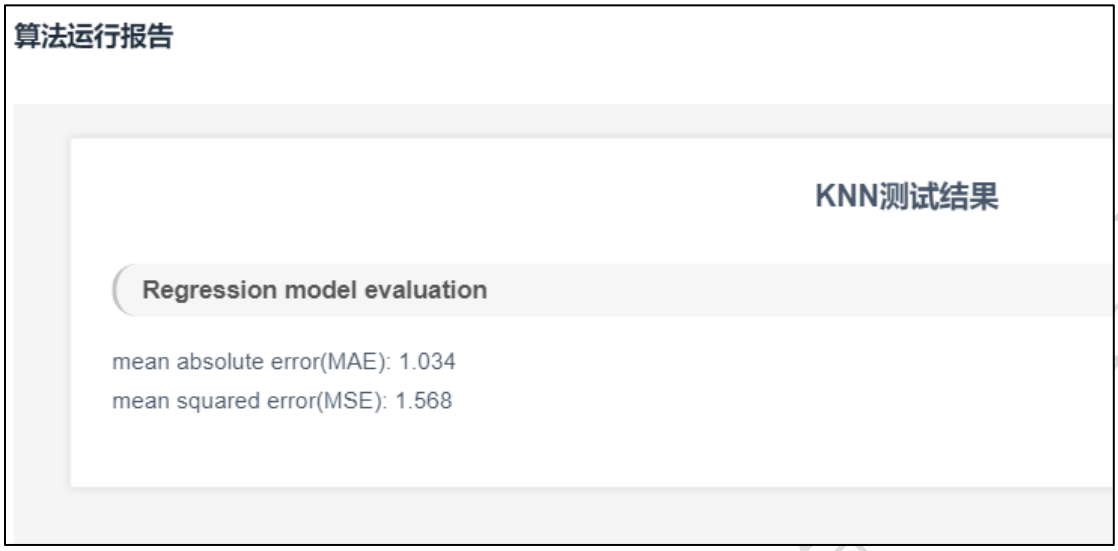

算法运行报告
KNN测试结果
Regression model evaluation
mean absolute error(MAE): 1.034
mean squared error(MSE): 1.568

图 266


120%
输入源
数据拆分
KNN
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
类型
取值范围
educational_level
数值
1-5
gender
数值
0-1
family_size
数值
1-8

图 267

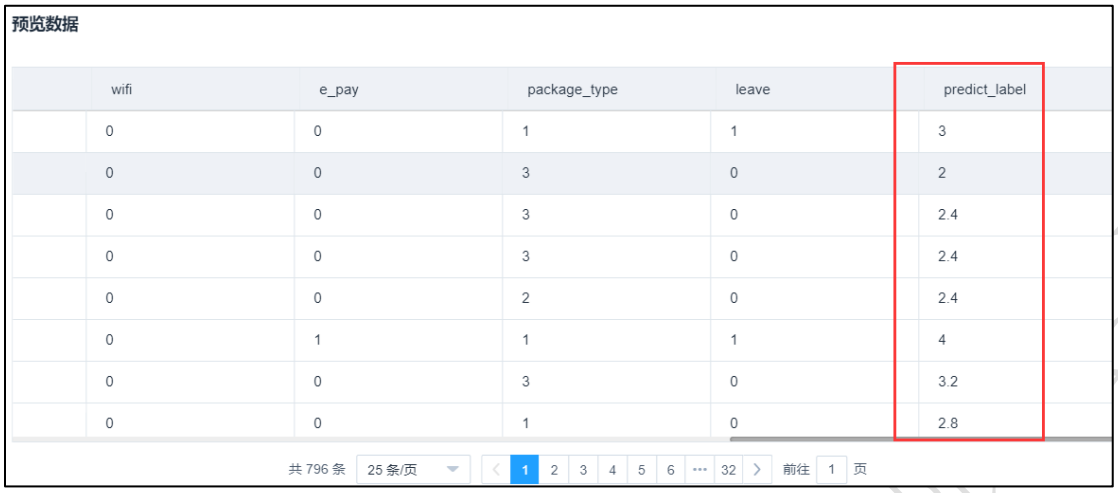


图 268

3.4.5 分类

3.4.5.1 AdaBoost

图标： Adaboost

描述： Adaboost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器）。

字段属性

特征列：通过勾选的方式选择特征所在列。

标签列：选择分类标签所在的列，一般为字符型数据。

参数设置

运算法则：包括 SAMME.R:、SAMME，默认为 SAMME.R。

最大迭代次数：整数型，默认值 50。

权重缩减系数：浮点型，默认值 1.0。

输出

表结果：AdaBoost 算法分类结果。

报告：Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall

示例

下列对某数据进行 AdaBoost 算法分类：

- 选择特征数列，标签列。如图 269 所示。
- 保留默认参数，最大迭代次数为 50，权重缩减系数为 10，运算法则为 SAMME.R，如图 270 所示。
- AdaBoost 运行成功后，可选择查看数据，如图 271 所示。
- AdaBoost 运行成功后，可选择查看报告，如图 272 所示。
- 评估配置如图 273 所示。
- 评估组件运行成功后，可选择查看数据， 如图 274 所示。
- 评估报告： Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall
- 模型预测配置如图 275 所示。
- 模型预测结果如图 276 所示。

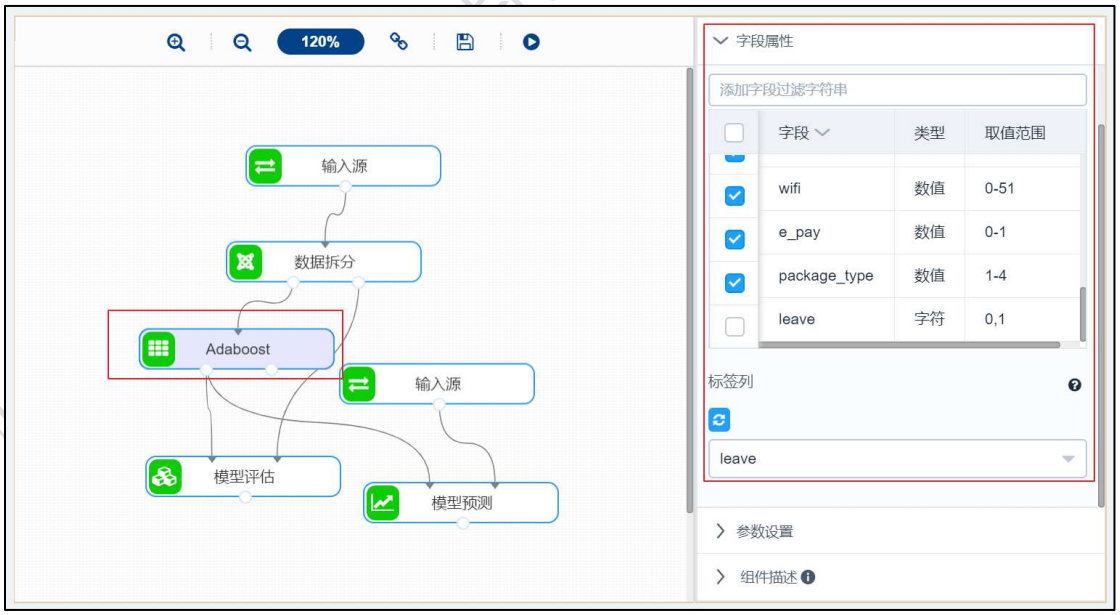


图 269


字段属性
参数设置
运算法则
SAMME.R
最大迭代次数
50
权重缩减系数
1

图 270


预览数据
wifi
e_pay
package_type
leave
predict_label
0 1 2 1 0
28 1 4 1 1
0 0 3 1 1
0 0 3 0 0
0 0 3 0 0
0 1 1 0 0
0 1 2 1 0
0 0 2 0 0
共 597 条
25 条/页
1 2 3 4 5 6 ··· 24
前往 1 页

图 271


算法运行报告
ADABOOST训练结果
Confusion Matrix
400
83
1
True Label
下载

图 272

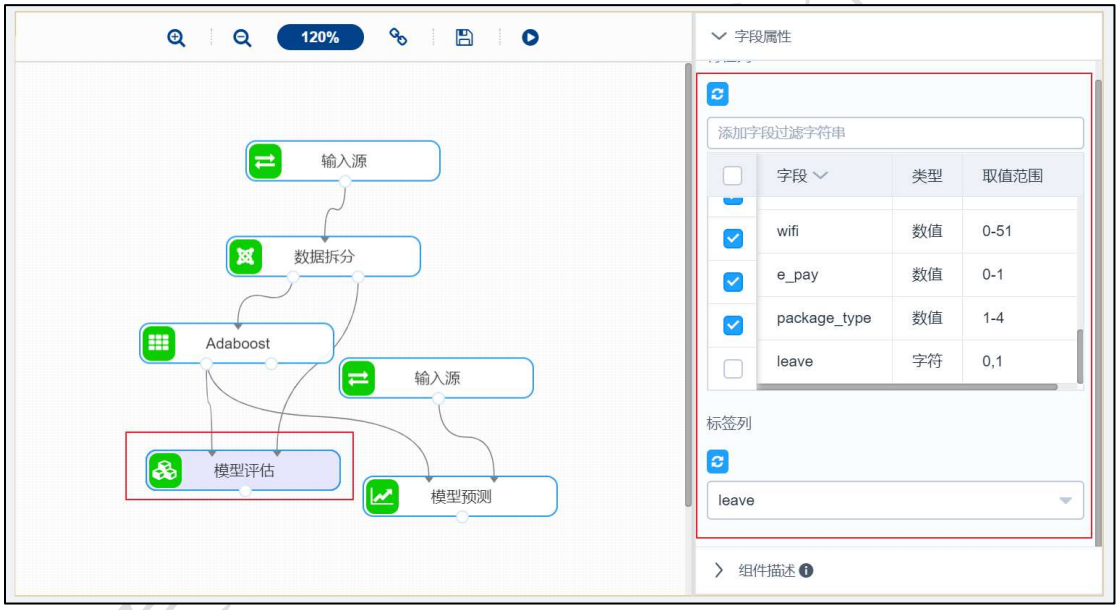

120%
输入源
数据拆分
Adaboost
输入源
模型评估
模型预测
字段属性
添加字段过滤字符串
字段
类型
取值范围
wifi
数值
0-51
e_pay
数值
0-1
package_type
数值
1-4
leave
字符
0,1
标签列
leave
组件描述

图 273


预览数据
wifi
e_pay
package_type
leave
predict_label
0 0 3 0 0
0 0 1 0 0
0 1 2 1 1
33 0 4 0 1
0 0 3 0 0
28 1 1 0 0
0 1 4 0 1
0 1 2 0 0
共 199 条
25 条/页
1 2 3 4 5 6 ··· 8
前往 1 页

图 274

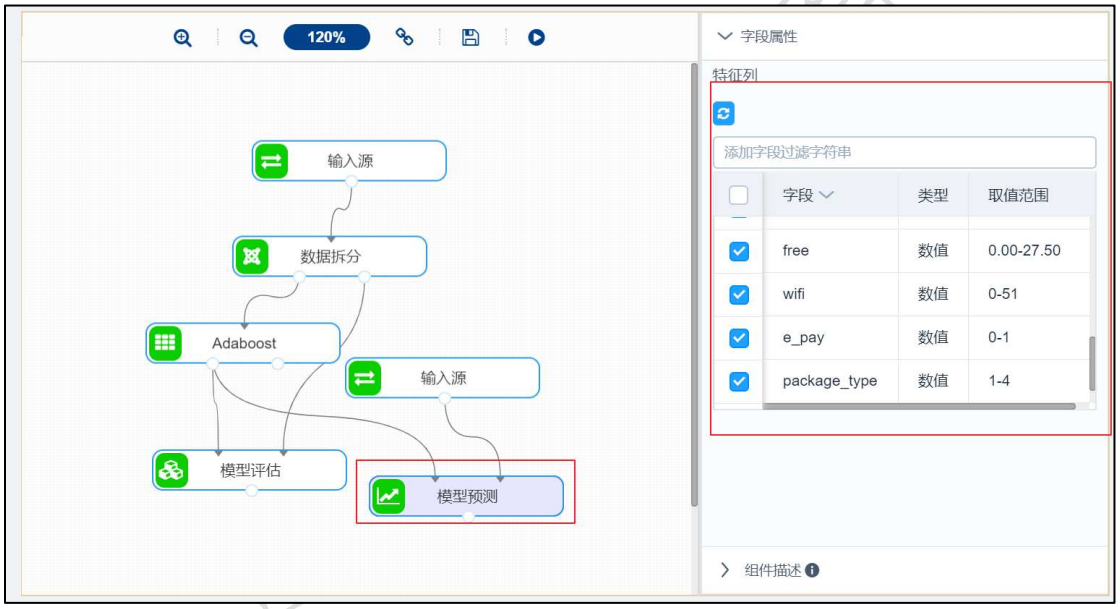
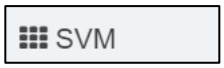

120%
输入源
数据拆分
Adaboost
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
类型
取值范围
free 数值 0.00-27.50
wifi 数值 0-51
e_pay 数值 0-1
package_type 数值 1-4
组件描述

图 275

图 276

3.4.5.2 SVM

图标： 

描述：SVM（支持向量机）方法是通过一个非线性映射，把样本空间映射到一个高维乃至无穷维的特征空间中使得在原来的样本空间中非线性可分的问题转化为在特征空间中的线性可分的问题。

字段属性

特征列：通过勾选的方式选择特征所在列。

标签列：选择分类标签所在的列，仅支持字符型数据。

参数设置

罚项系数：浮点型， 默认 1.0。

核函数：支持线性核、多项式核、高斯核、sigmoid，默认 rbf。

输出

表结果：SVM 算法分类结果。

报告：Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall

示例

下列对某数据进行 SVM 算法分类：

- 选择特征数列，标签列。如图 277 所示。

- 保留默认参数，罚项系数为 1.0，核函数为高斯核，如图 278 所示。
- SVM 运行成功后，可选择查看结果，如图 279 所示。
- SVM 运行成功后，可选择查看报告，如图 280 所示。
- 模型评估配置如图 281 所示。
- 模型评估运行成功后，选择查看数据，如图 282 所示。
- 模型评估报告运行成功后，选择查看报告，如图 283 所示。
- 模型预测配置如图 284 所示。
- 模型预测运行成功后，选择查看数据，如图 285 所示。

图 277

图 278

预览数据

e_pay	package_type	leave	educational_level	predict_label
1	2	1	3	3
1	4	1	3	3
0	3	1	3	3
0	3	0	4	4
0	3	0	1	1
1	1	0	3	3
1	2	1	3	3
0	2	0	3	3

共 597 条 25 条/页 1 2 3 4 5 6 ··· 24 前往 1 页

图 279

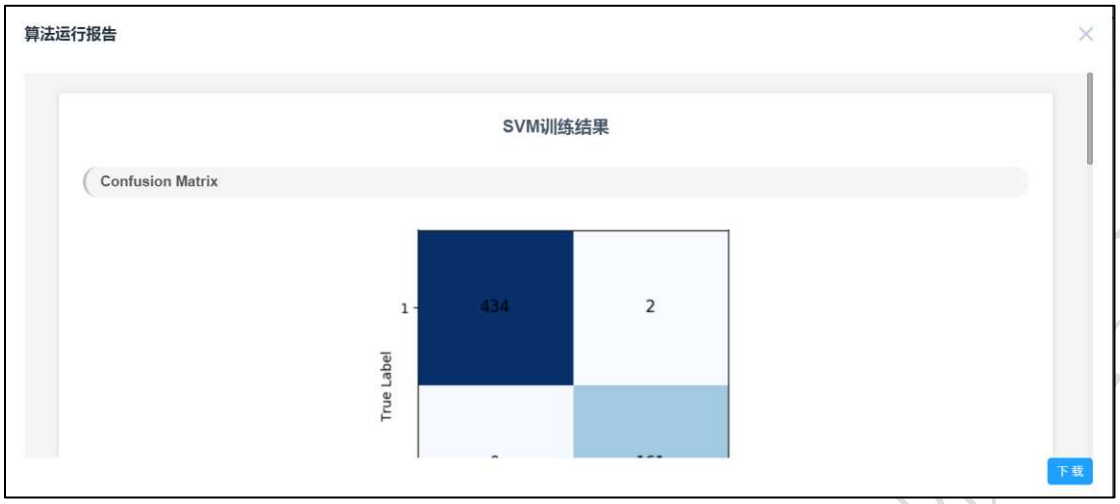

算法运行报告
SVM训练结果
Confusion Matrix
1
434
2
True Label
下载

图 280


120%
输入源
数据拆分
SVM
输入源
模型评估
模型预测
字段属性
添加字段过滤字符串
字段
类型
取值范围
package_type
值
1-4
leave
数值
0-1
educational_level
数值
1-5
predict_label
数值
1-5
标签列
educational_level
组件描述

图 281


预览数据
e_pay
package_type
leave
educational_level
predict_label
0 3 0 1 2
0 1 0 3 2
1 2 1 3 4
0 4 0 5 2
0 3 0 1 2
1 1 0 1 2
1 4 0 3 2
1 2 0 2 2
共 199 条
25 条/页
1 2 3 4 5 6 8
前往 1 页

图 282

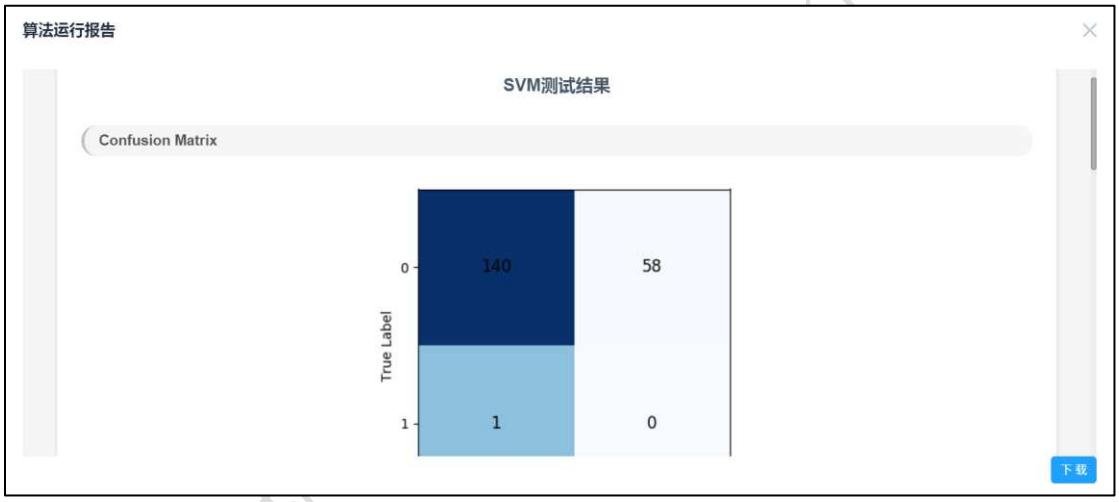

算法运行报告
SVM测试结果
Confusion Matrix
True Label
0
1
140
58
1
0
下载

图 283

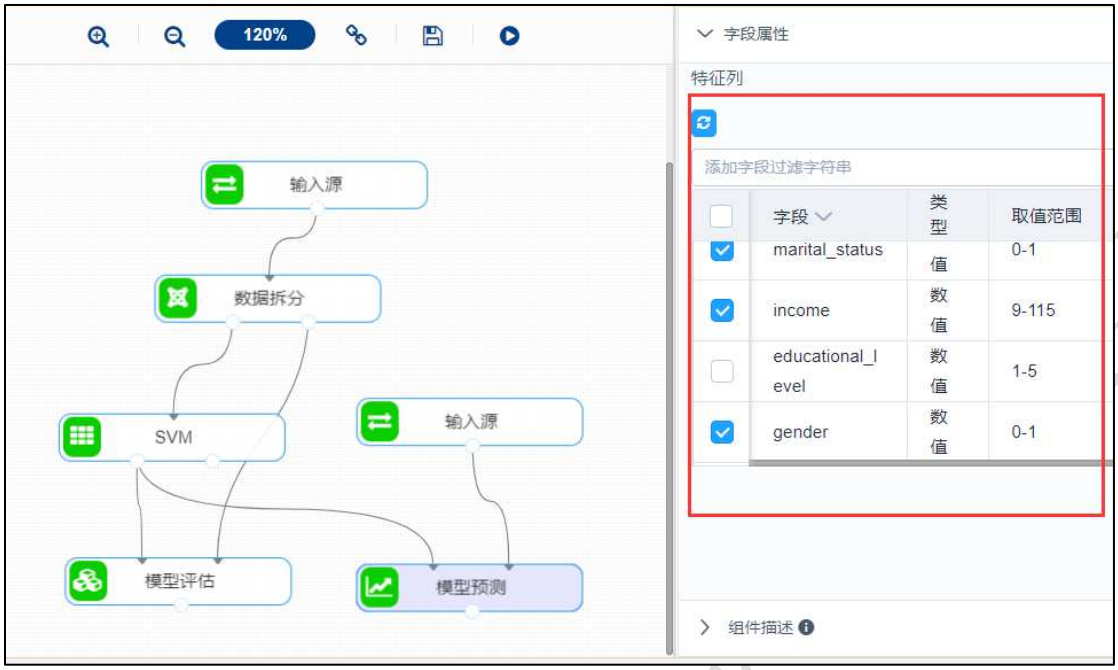


图 284

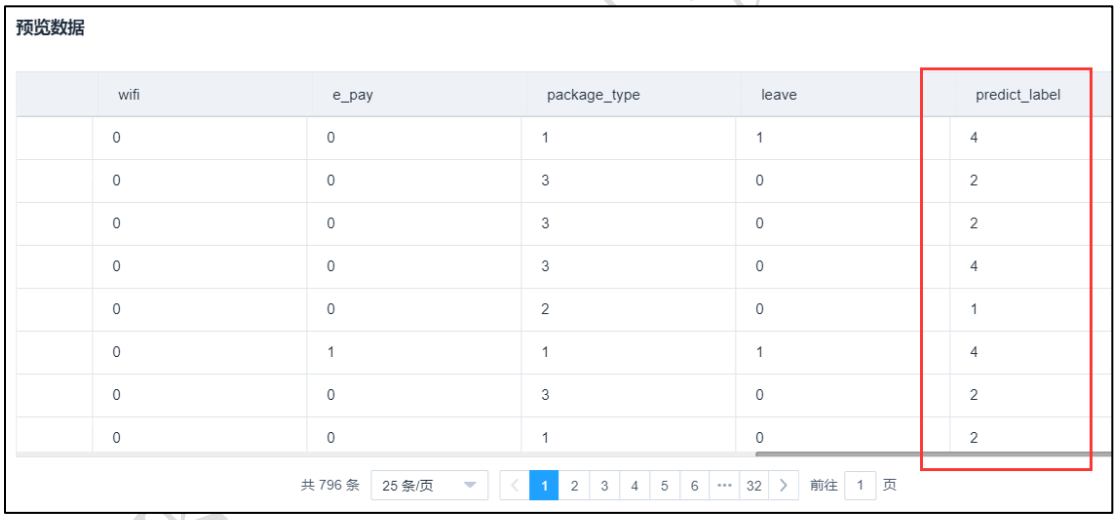


图 285

3.4.5.3 KNN

图标： KNN

描述： KNN 算法的核心思想是如果一个样本在特征空间中的 k 个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。

字段属性

特征列：通过勾选的方式选择特征所在列。

标签列：选择分类标签所在的列，一般为字符型数据。

参数设置：

最近邻个数 K：整数型，通常不大于 20，默认 5.

投票权重类型：权重相等或权重与距离成反比，默认权重相等。

计算最近邻的算法：包括 自动、BallTree、KDTree、暴力搜索法，默认自动。

输出

表结果：KNN 算法分类结果。

报告：Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall

示例

下列对某数据进行 KNN 算法分类：

- 选择特征数列，标签列，如图 286 所示。
- 保留默认参数，最近邻个数 K 为 5，投票权重类型为权重相等，暴力搜索为自动，如图 287 所示。
- KNN 运行成功后，可选择查看结果，如图 288 所示。
- KNN 运行成功后，可选择查看报告，如图 289 所示。
- 模型评估配置如图 290 所示。
- 模型评估运行成功后，选择查看数据，如图 291 所示。
- 模型评估运行成功后，选择查看报告，如所示。
- 模型预测配置如图 292 所示。
- 模型预测结果如图 292 所示。

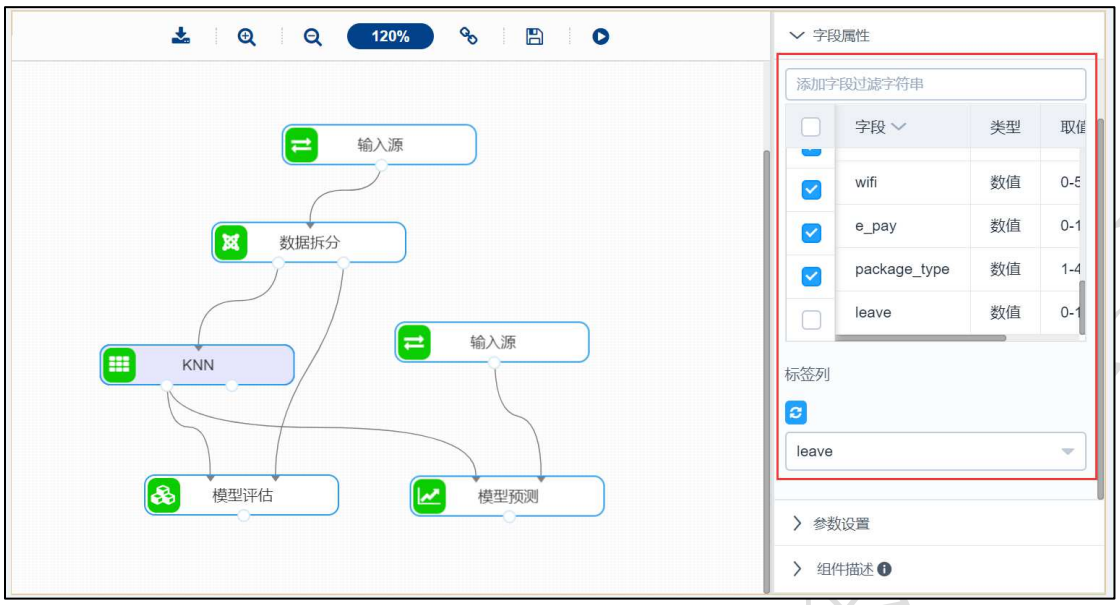

120%
输入源
数据拆分
KNN
输入源
模型评估
模型预测
字段属性
添加字段过滤字符串
字段
类型
wifi
数值
e_pay
数值
package_type
数值
leave
数值
标签列
leave
参数设置
组件描述

图 286


字段属性
参数设置
K值
5
投票权重类型
权重相等
计算最近邻的算法
自动

图 287

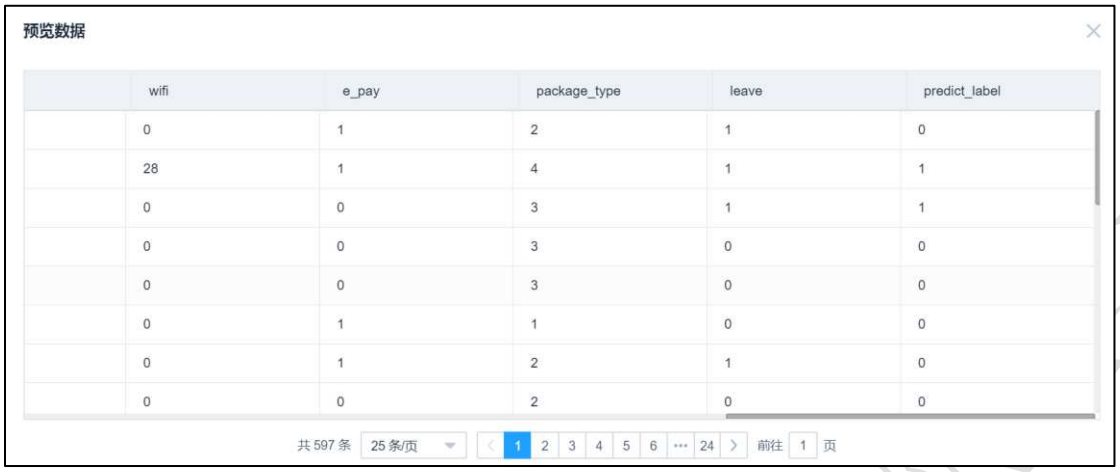

预览数据
wifi
e_pay
package_type
leave
predict_label
0 1 2 1 0
28 1 4 1 1
0 0 3 1 1
0 0 3 0 0
0 0 3 0 0
0 1 1 0 0
0 1 2 1 0
0 0 2 0 0
共 597 条
25 条/页
1 2 3 4 5 6 ··· 24
前往 1 页

图 288

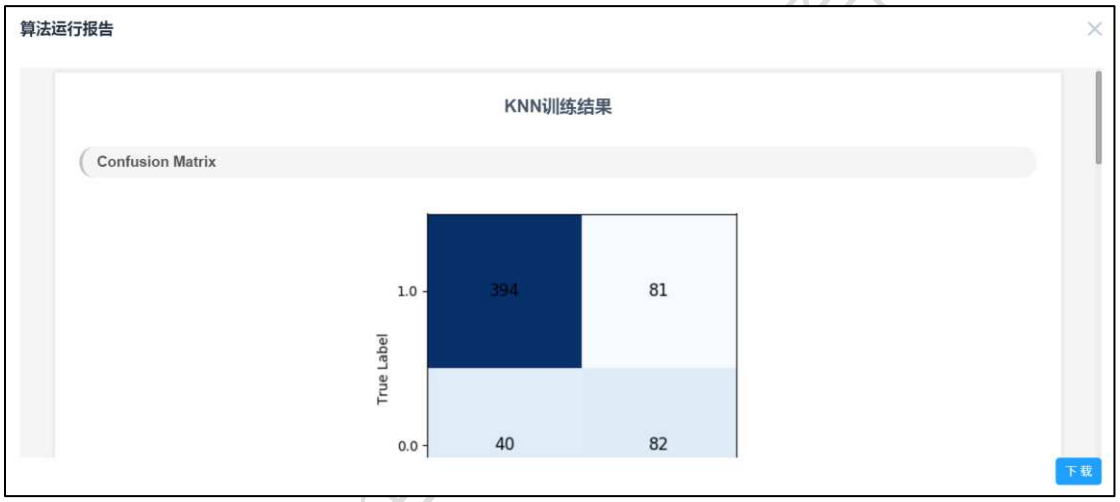

算法运行报告
KNN训练结果
Confusion Matrix
True Label
1.0
0.0
394
81
40
82
下载

图 289

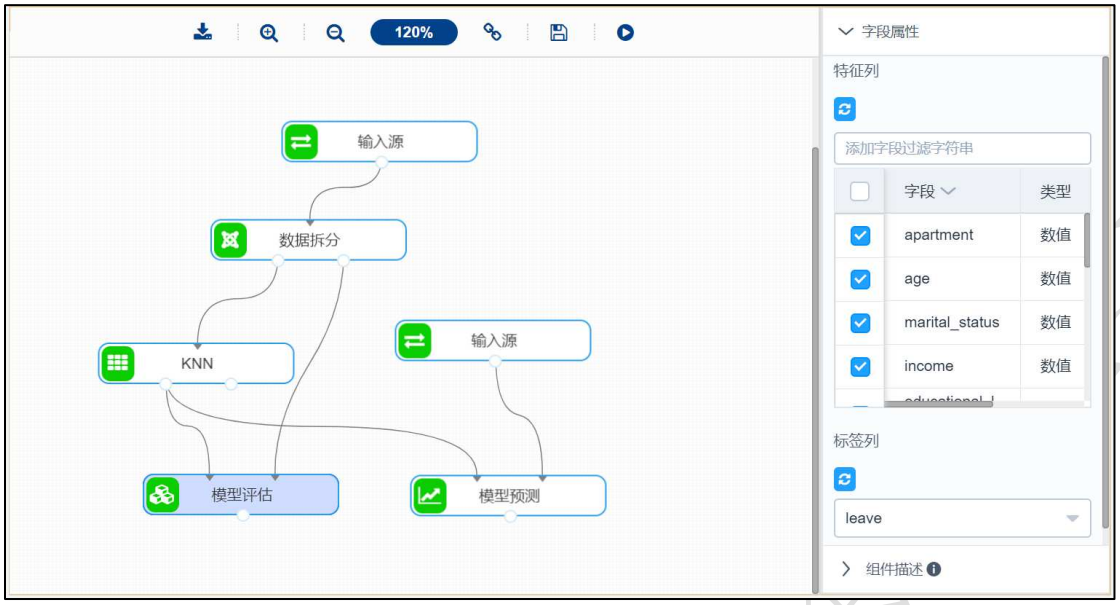

120%
输入源
数据拆分
KNN
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
类型
apartment
数值
age
数值
marital_status
数值
income
数值
标签列
leave
组件描述

图 290

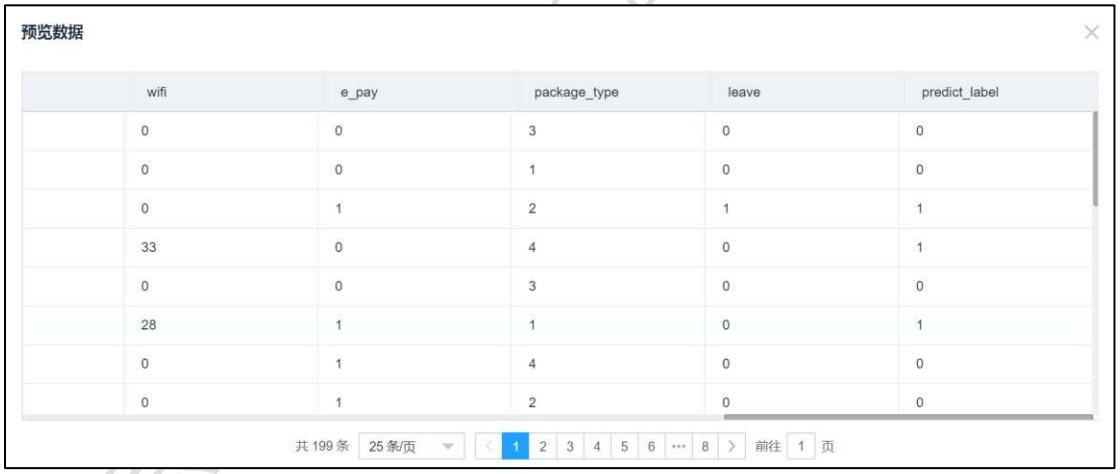

预览数据
wifi
e_pay
package_type
leave
predict_label
0 0 3 0 0
0 0 1 0 0
0 1 2 1 1
33 0 4 0 1
0 0 3 0 0
28 1 1 0 1
0 1 4 0 0
0 1 2 0 0
共 199 条
25 条/页
1 2 3 4 5 6 ··· 8
前往 1 页

图 291

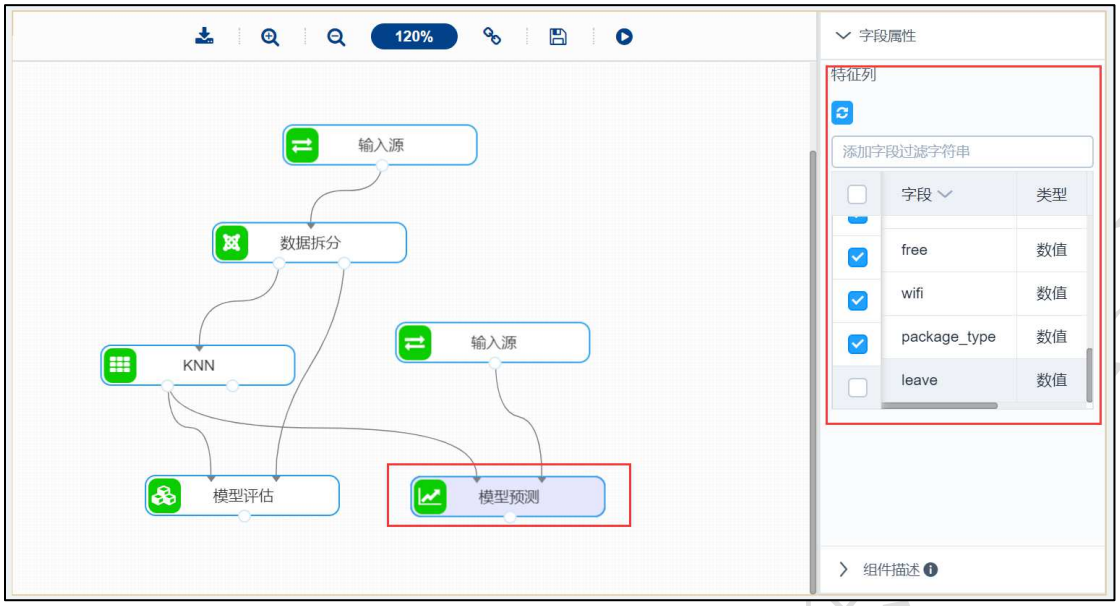


图 292

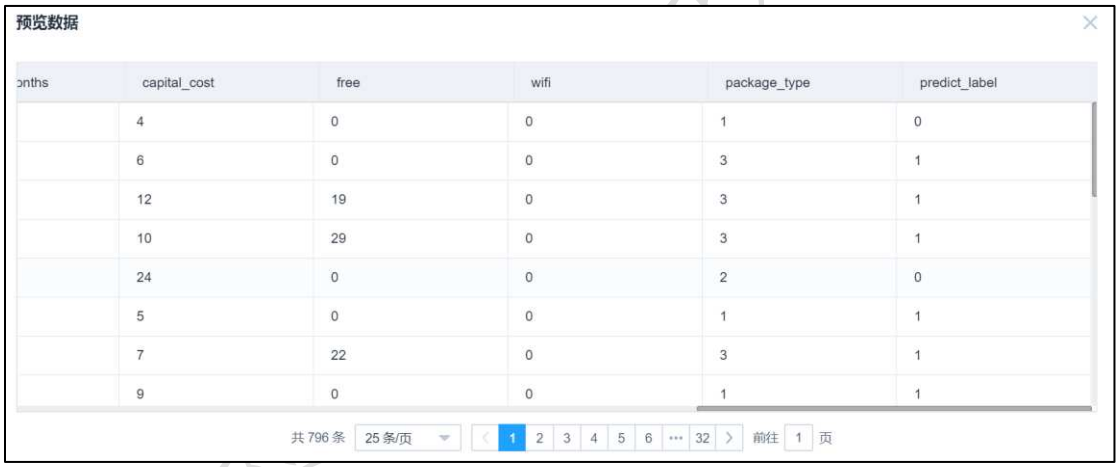


图 293

3.4.5.4 朴素贝叶斯

图标： 

描述： 朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。

字段属性

特征列：通过勾选的方式选择特征所在列。

标签列：选择分类标签所在的列。

参数设置

条件概率分布满足何种分布：包含高斯分布、多项式分布、二项分布，默认高斯分布。

输出

表结果：朴素贝叶斯算法分类结果。

报告：Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall

示例

下列对某数据进行朴素贝叶斯算法分类：

- 选择特征数列，标签列。如图 294 所示。
- 保留默认参数，条件概率分布为高斯分布如图 295 所示。
- 朴素贝叶斯运行成功后，可选择查看结果，如图 296 所示。
- 朴素贝叶斯运行成功后，可选择查看报告，如图 297 所示。
- 模型评估配置如图 298 所示。
- 模型评估运行成功后，选择查看数据，如图 299 所示。
- 模型评估报告：Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall
- 模型预测配置如图 301 所示。
- 模型预测结果如图 302 所示。

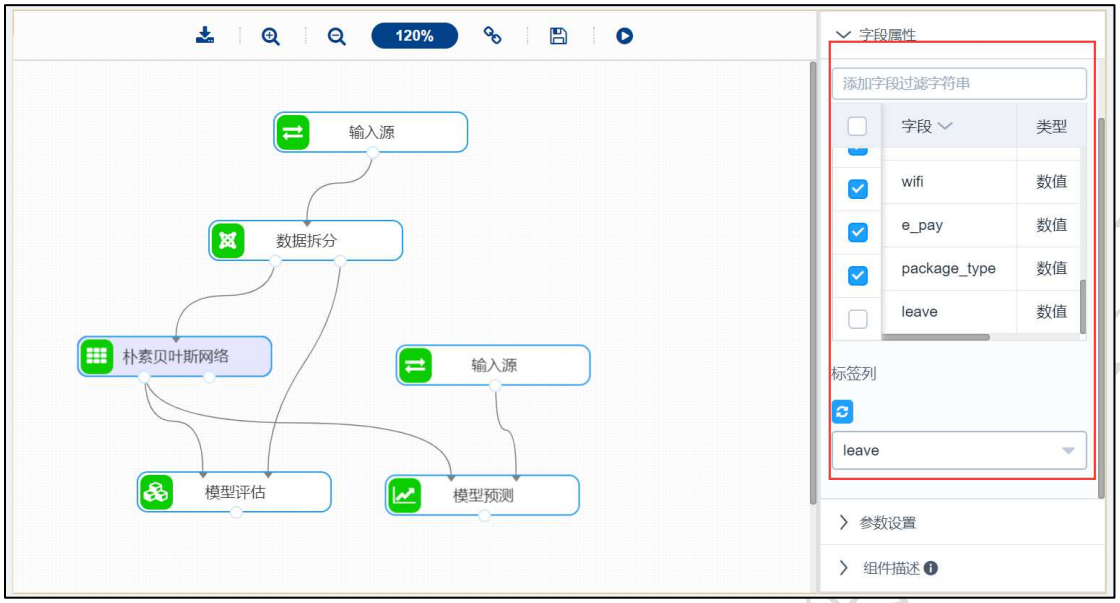

120%
输入源
数据拆分
朴素贝叶斯网络
输入源
模型评估
模型预测
字段属性
添加字段过滤字符串
字段
类型
wifi
数值
e_pay
数值
package_type
数值
leave
数值
标签列
leave
参数设置
组件描述

图 294


字段属性
参数设置
条件概率分布满足何种分布
高斯分布

图 295

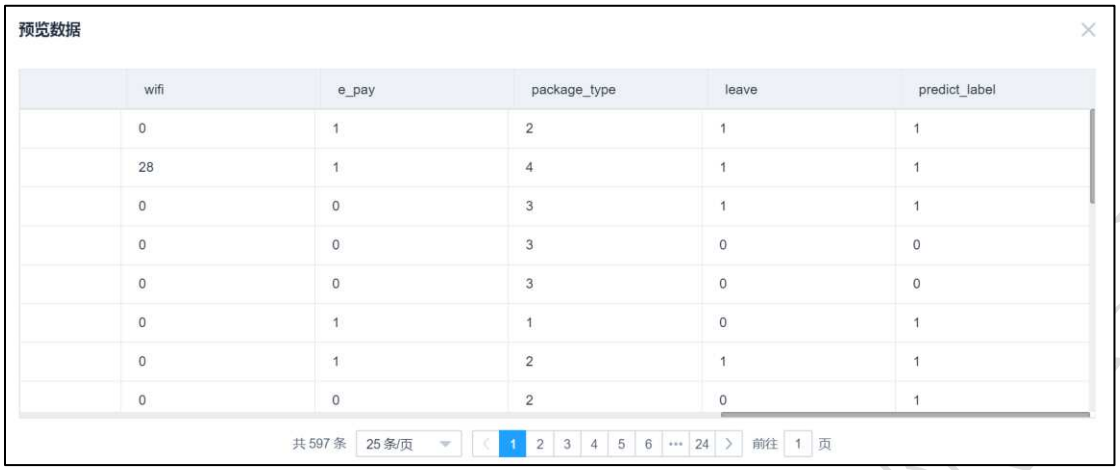

预览数据
wifi
e_pay
package_type
leave
predict_label
0 1 2 1 1
28 1 4 1 1
0 0 3 1 1
0 0 3 0 0
0 0 3 0 0
0 1 1 0 1
0 1 2 1 1
0 0 2 0 1
共 597 条
25 条/页
1 2 3 4 5 6 ··· 24
前往 1 页

图 296

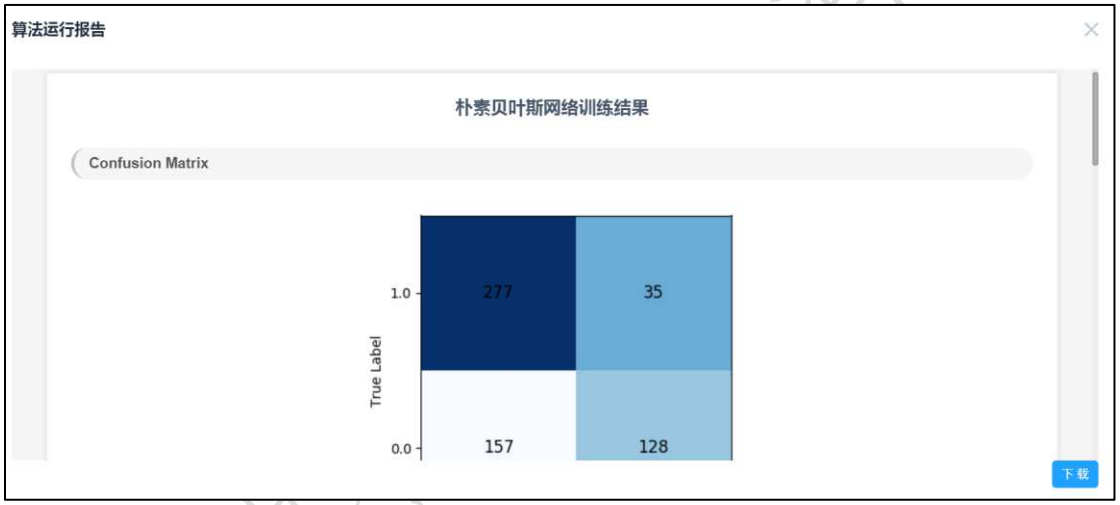

算法运行报告
朴素贝叶斯网络训练结果
Confusion Matrix
True Label
1.0
0.0
277
35
157
128
下载

图 297

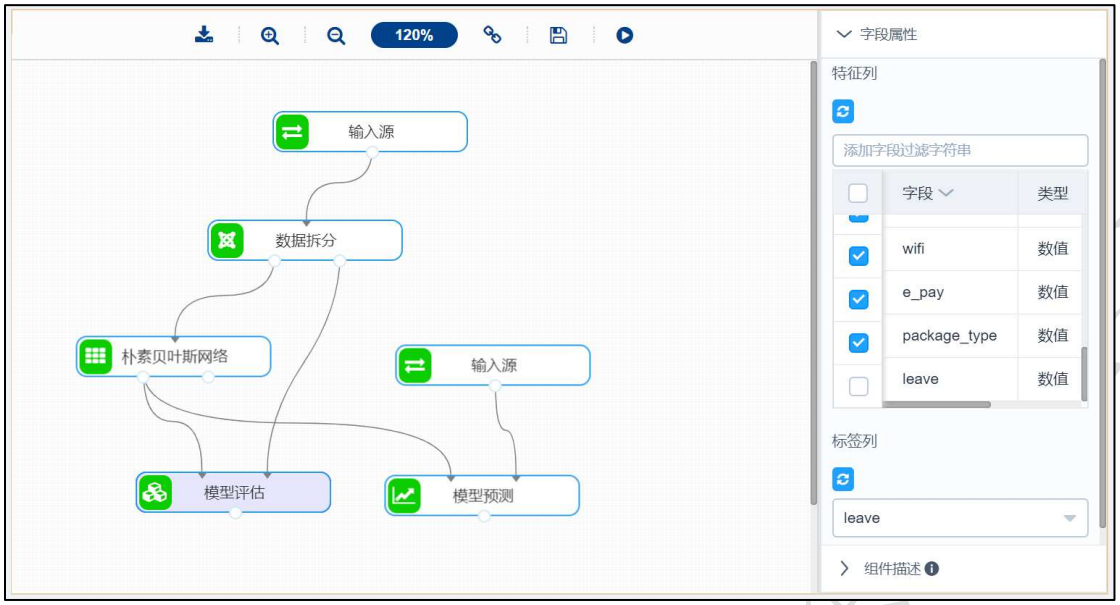

120%
输入源
数据拆分
朴素贝叶斯网络
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
类型
wifi
数值
e_pay
数值
package_type
数值
leave
数值
标签列
leave
组件描述

图 298


预览数据
wifi
e_pay
package_type
leave
predict_label
0 0 3 0 0
0 0 1 0 0
0 1 2 1 1
33 0 4 0 1
0 0 3 0 0
28 1 1 0 1
0 1 4 0 1
0 1 2 0 1
共 199 条
25 条/页
前往
1
页

图 299

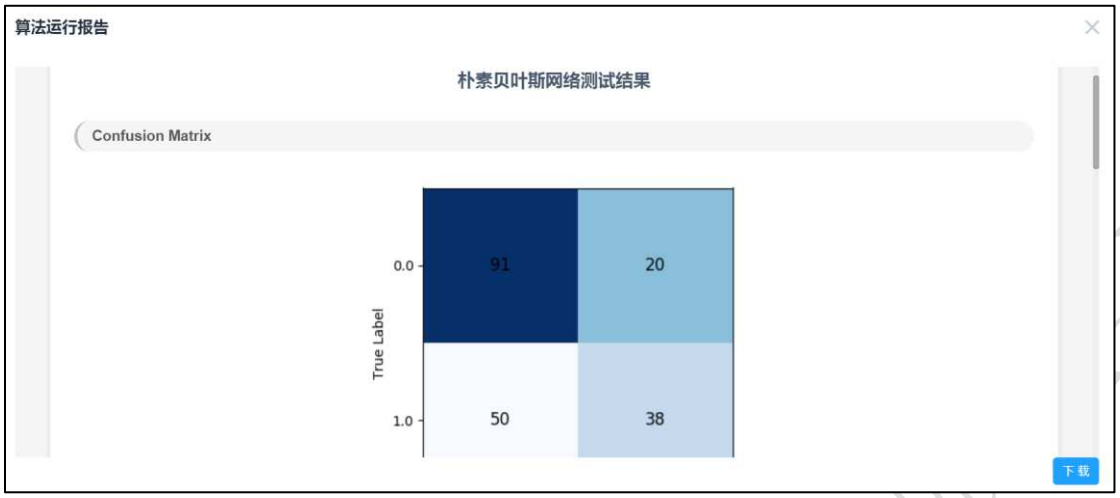

算法运行报告
朴素贝叶斯网络测试结果
Confusion Matrix
0.0
1.0
True Label
91
20
50
38
下载

图 300

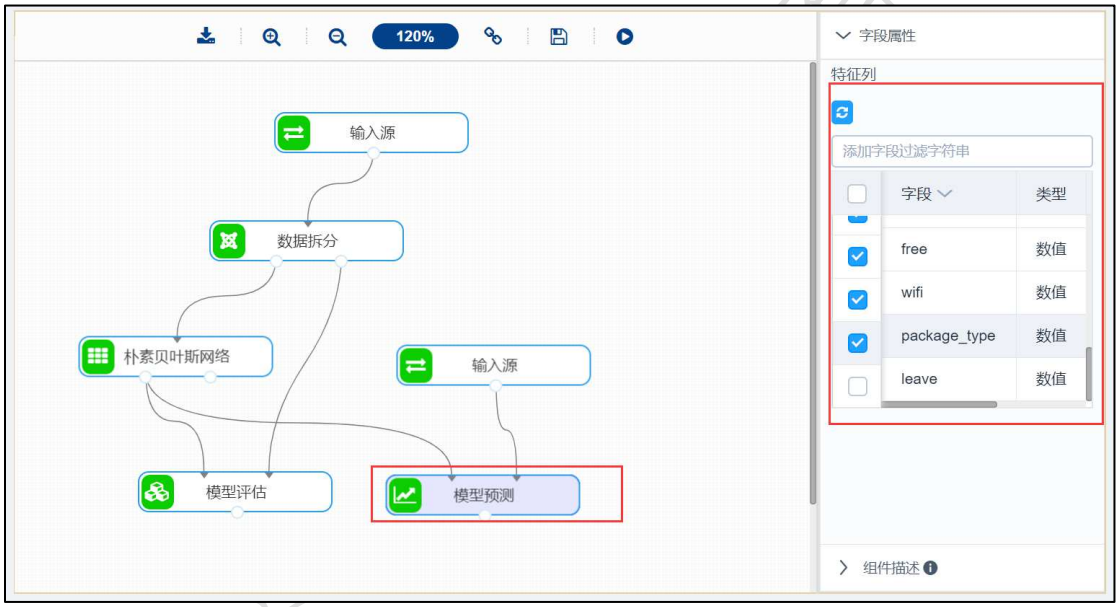

120%
输入源
数据拆分
朴素贝叶斯网络
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
类型
free
数值
wifi
数值
package_type
数值
leave
数值
组件描述

图 301

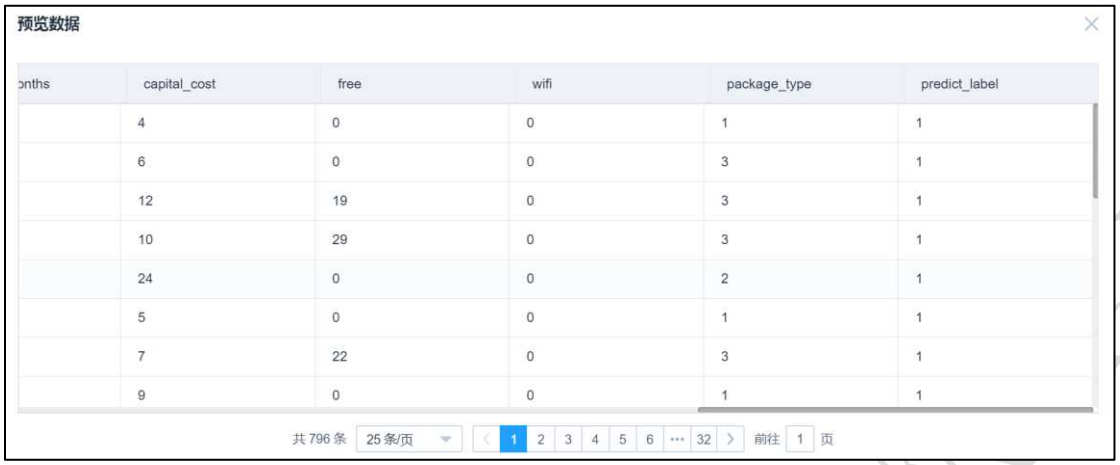


图 302

3.4.5.5 ID3

图标： 

描述：ID3 算法是一种贪心算法，用来构造决策树。以信息熵的下降速度为选取测试属性的标准，即在每个节点选取还尚未被用来划分的具有最高信息增益的属性作为划分标准，然后继续这个过程，直到生成的决策树能完美分类训练样例。ID3 算法可用于划分标称型数据集，不能处理连续分布的数据特征。

字段属性

特征列：请选择标称类数据。

标签列：选择分类标签所在的列。

输出

表结果：ID3 算法分类结果。

报告：Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall

示例

下列对某数据进行 ID3 算法分类：

- 选择特征数列，标签列。如图 303 所示。
- ID3 运行成功后，可选择查看结果，如图 304 所示。
- ID3 运行成功后，可选择查看报告，如图 305 所示。

- 模型评估配置如图 306 所示。
- 模型评估运行成功后，选择查看数据，如图 307 所示。
- 模型评估运行成功后，选择查看报告，如所示。
- 模型预测配置如图 308 所示。
- 模型预测结果如图 309 所示。

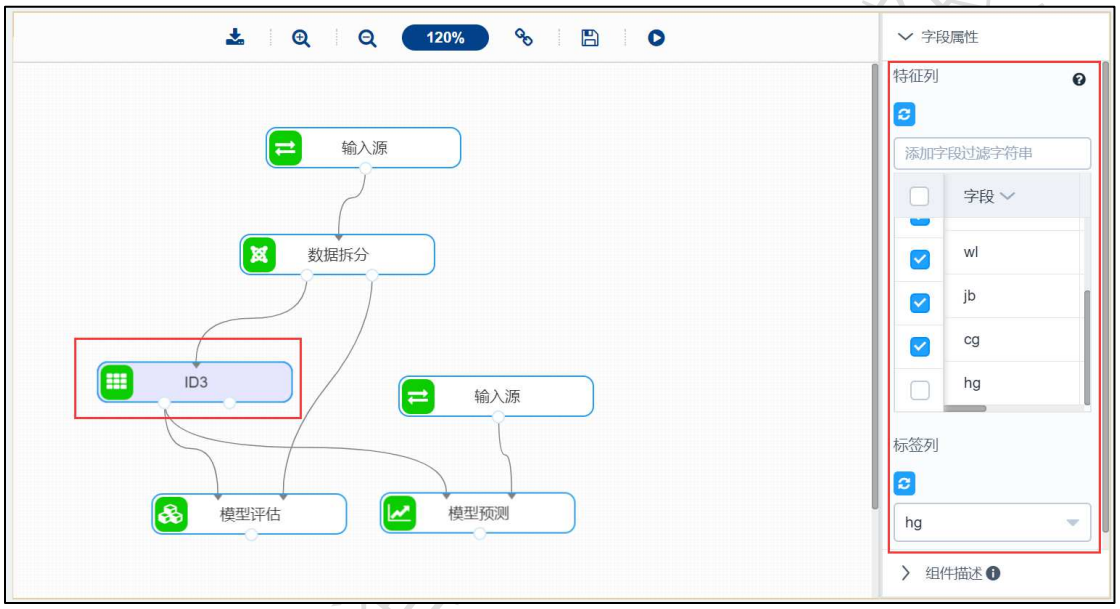


图 303

图 304

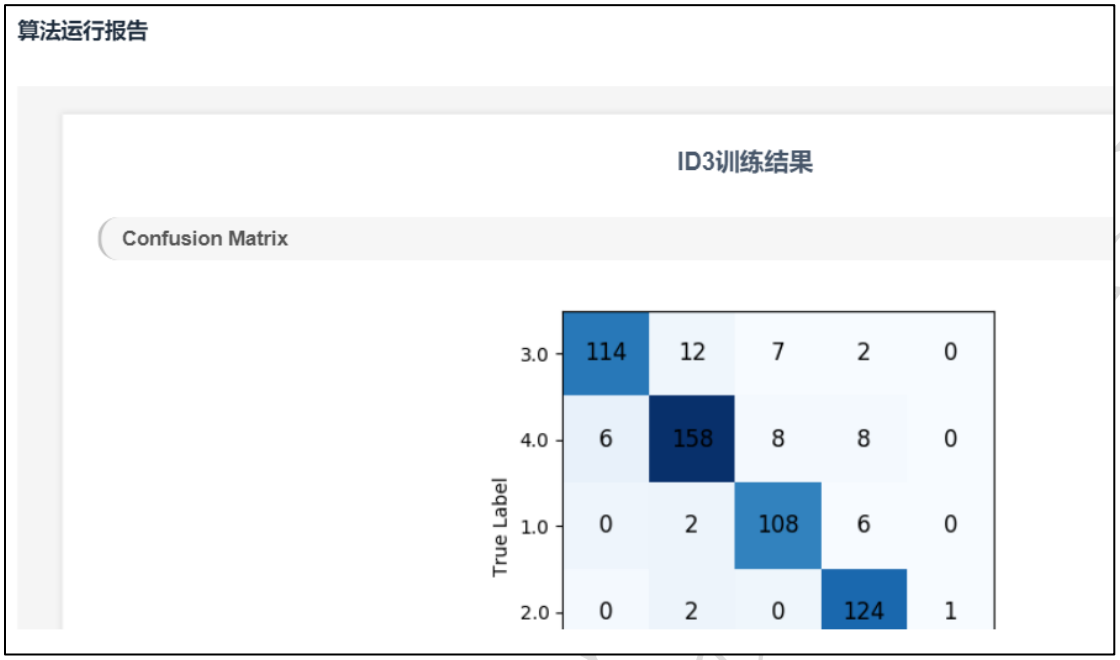


图 305

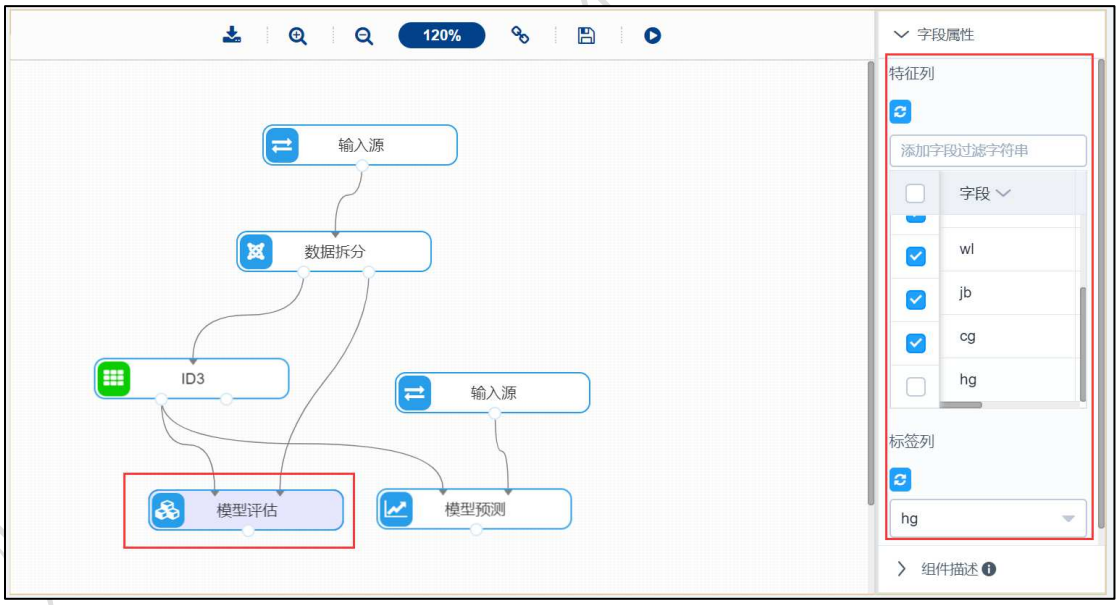


图 306


预览数据
wl
jb
cg
hg
predict_label
清晰 凹陷 硬滑 是 是
稍糊 稍凹 软粘 是 否
稍糊 稍凹 硬滑 否 否
清晰 平坦 软粘 否 是
稍糊 凹陷 硬滑 否 否
共 5 条
25 条/页
1
前往 1 页

图 307

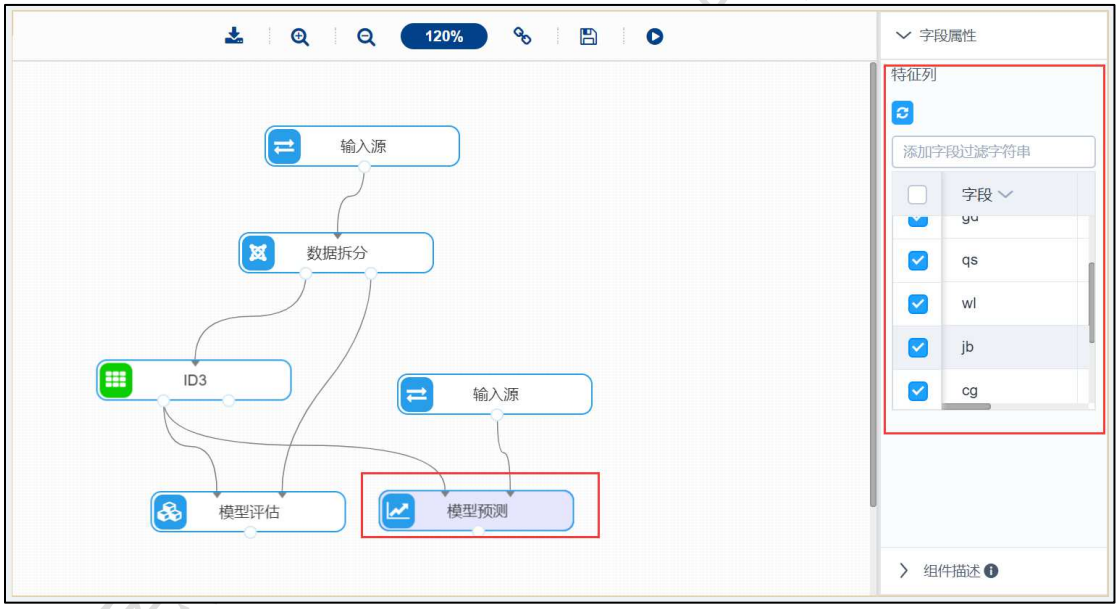

120%
输入源
数据拆分
ID3
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
qs
wl
jb
cg
组件描述

图 308

图 309

3.4.5.6 CART 决策树

图标： 

描述： CART(Classification And Regression Tree)算法是一种决策树分类方法。它采用一种二分递归分割的技术，分割方法采用基于最小距离的基尼指数估计函数，将当前的样本集分为两个子样本集，使得生成的的每个非叶子节点都有两个分支。因此，CART 算法生成的决策树是结构简洁的二叉树。

字段属性

特征列：通过勾选的方式选择特征所在列。

标签列：选择分类标签所在的列，请选择字符型数据。

参数设置

切分时的评价准则：包括 Gini 系数、熵，默认 Gini 系数。

切分原则：包括选择最优的切分、随机切分，默认选择最优的切分。

输出

表结果：CART 决策树算法分类结果。

报告：Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall

示例

下列对某数据进行 CART 决策树算法分类：

- 选择特征数列，标签列。如图 310 所示。
- 保留默认参数，切分时的评价准则为 Gini 系数，切分原则为最优的切分，如图 311 所示。
- CART 决策树运行成功后，可选择查看结果，如图 312 所示。
- CART 决策树运行成功后，可选择查看报告，如图 313 所示。
- 模型评估配置如图 314 所示。
- 模型评估运行成功后，选择查看数据，如图 315 所示。
- 模型评估运行成功后，选择查看报告，如图 316 所示。
- 模型预测配置如图 317 所示。
- 模型预测运行成功后，选择查看数据，如图 318 所示。

图 310


字段属性
参数设置
切分时的评价准则
Gini系数
切分原则
选择最优的切分

图 311


预览数据
wifi
e_pay
package_type
leave
predict_label
0 1 2 1 1
28 1 4 1 1
0 0 3 1 1
0 0 3 0 0
0 0 3 0 0
0 1 1 0 0
0 1 2 1 1
0 0 2 0 0
共 597 条
25 条/页
1 2 3 4 5 6 ··· 24
前往 1 页

图 312

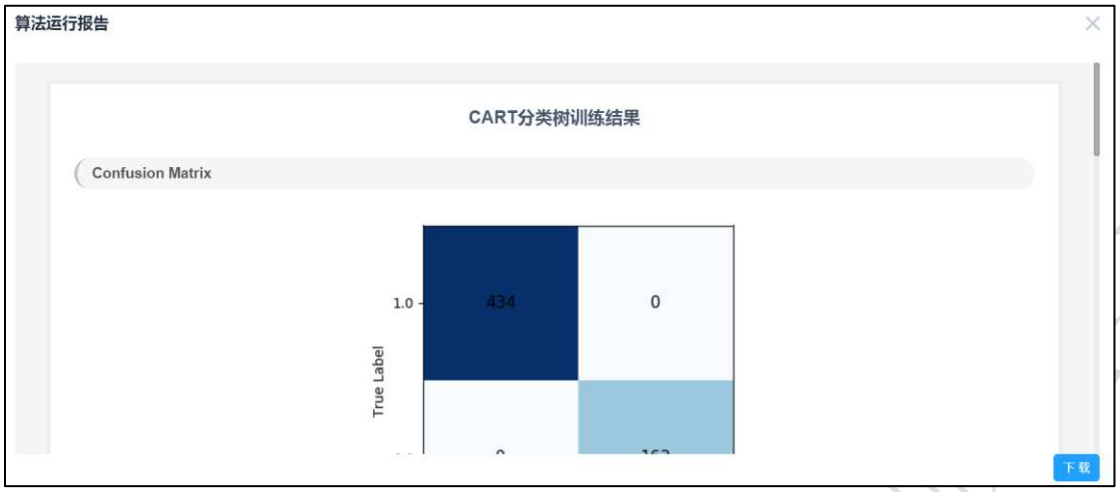

算法运行报告
CART分类树训练结果
Confusion Matrix
1.0
434
0
True Label
下载

图 313


120%
输入源
数据拆分
CART分类树
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
wifi
e_pay
package_type
leave
标签列
leave
组件描述

图 314


预览数据
wifi
e_pay
package_type
leave
predict_label
0 0 3 0 0
0 0 1 0 0
0 1 2 1 0
33 0 4 0 0
0 0 3 0 1
28 1 1 0 0
0 1 4 0 1
0 1 2 0 0
共 199 条
25 条/页
1 2 3 4 5 6 ··· 8
前往 1 页

图 315

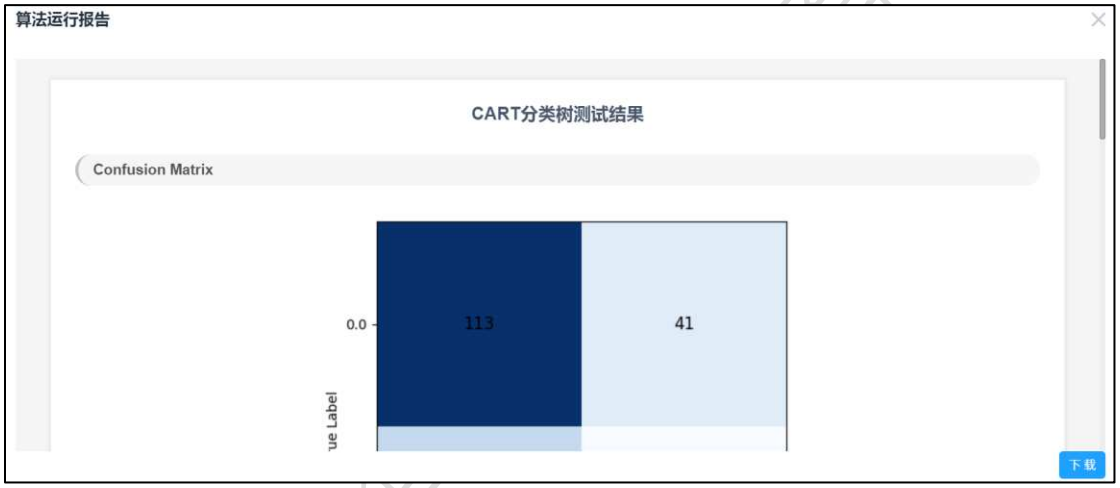

算法运行报告
CART分类树测试结果
Confusion Matrix
0.0
113
41
下载

图 316

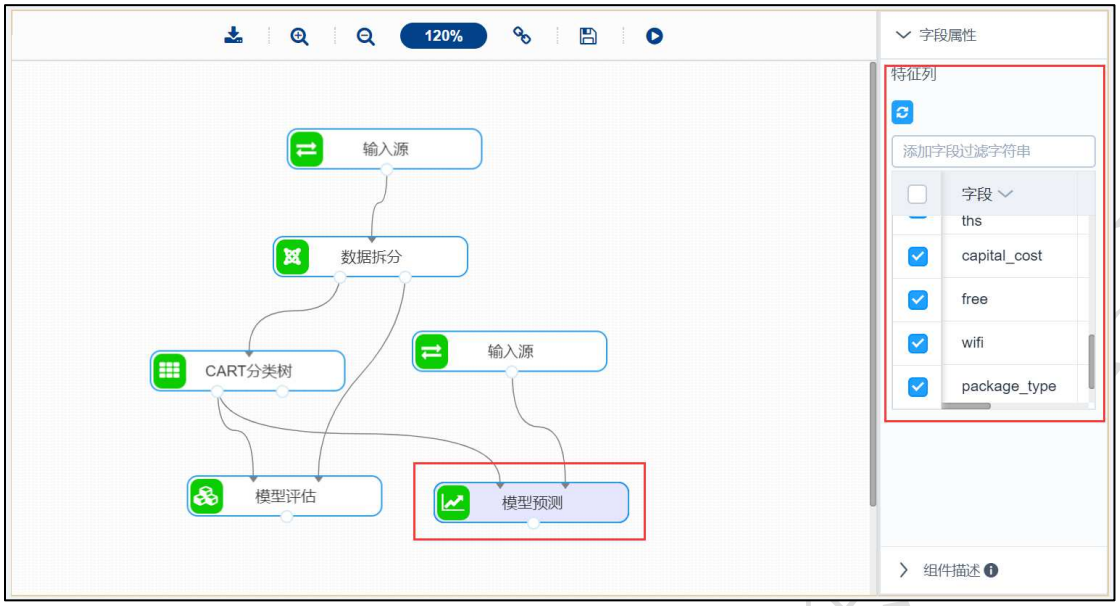


图 317

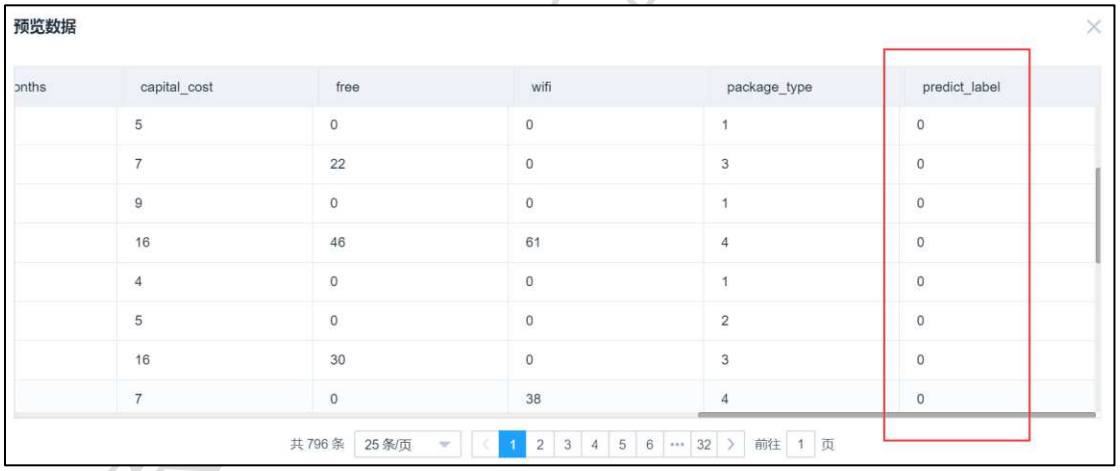


图 318

3.4.5.7 BP 神经网络

图标： 

描述：BP(back propagation)神经网络是一种按照误差逆向传播算法训练的多层前馈神经网络。

字段属性

特征列：必选。通过勾选的方式选择特征所在列。

标签列：必选。选择分类标签所在的列，仅支持字符型数据。

参数设置

隐藏层神经元个数：整数，默认 100。

隐层激活函数：包括 relu、identity、logistic、tanh，默认 relu。

优化器：包括 adam、lbfgs、sgd，默认 adam。

最大迭代：整数，默认 1000。

输出

表结果：BP 神经网络算法分类结果。

报告：Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall

示例

下列对某数据进行 BP 神经网络算法分类：

- 选择特征数列，标签列。如图 319 所示。
- 保留默认参数，隐藏层神经云个数为 100，隐藏激活函数为 relu，优化器为 adam，最大迭代为 1000，如图 320 所示。
- BP 神经网络运行成功后，可选择查看结果，如图 321 所示。
- BP 神经网络运行成功后，可选择查看报告，如图 322 所示。
- 模型评估配置如图 323 所示。
- 模型评估运行成功后，选择查看数据，如图 324 所示。
- 模型评估运行成功后，选择查看报告，如图 325 所示。
- 模型预测配置如图 326 所示。
- 模型预测运行成功后，选择查看数据，如图 327 所示。

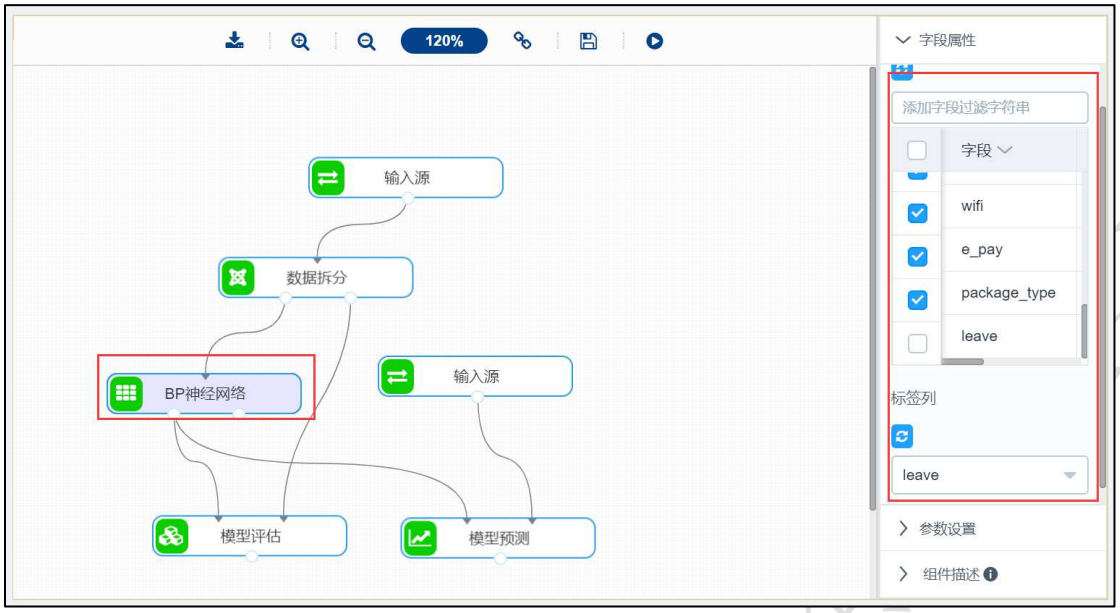

120%
输入源
数据拆分
BP神经网络
输入源
模型评估
模型预测
字段属性
添加字段过滤字符串
字段
wifi
e_pay
package_type
leave
标签列
leave
参数设置
组件描述

图 319


字段属性
参数设置
隐藏层神经元个数
100
隐层激活函数
relu
求解器
adam
最大迭代
1000

图 320

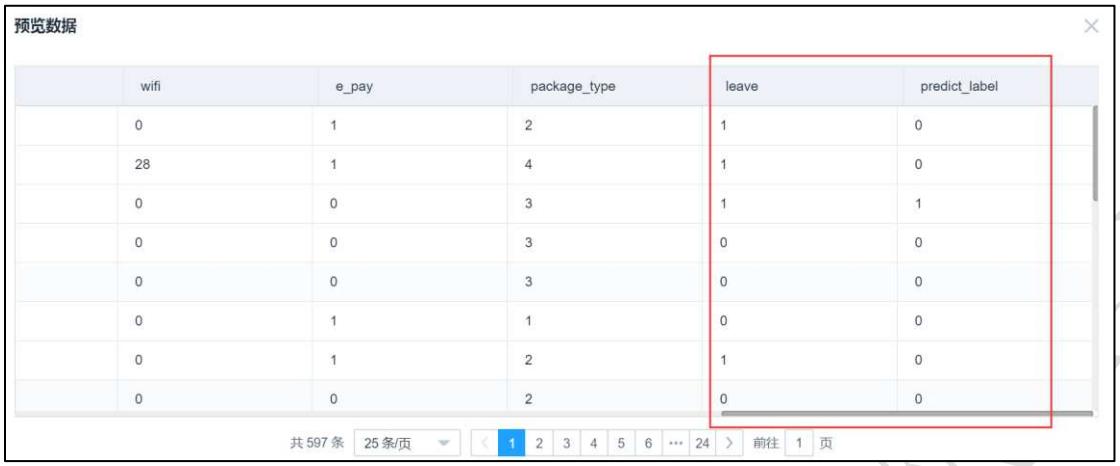

预览数据
wifi
e_pay
package_type
leave
predict_label
0 1 2 1 0
28 1 4 1 0
0 0 3 1 1
0 0 3 0 0
0 0 3 0 0
0 1 1 0 0
0 1 2 1 0
0 0 2 0 0
共 597 条
25 条/页
1 2 3 4 5 6 ··· 24
前往 1 页

图 321

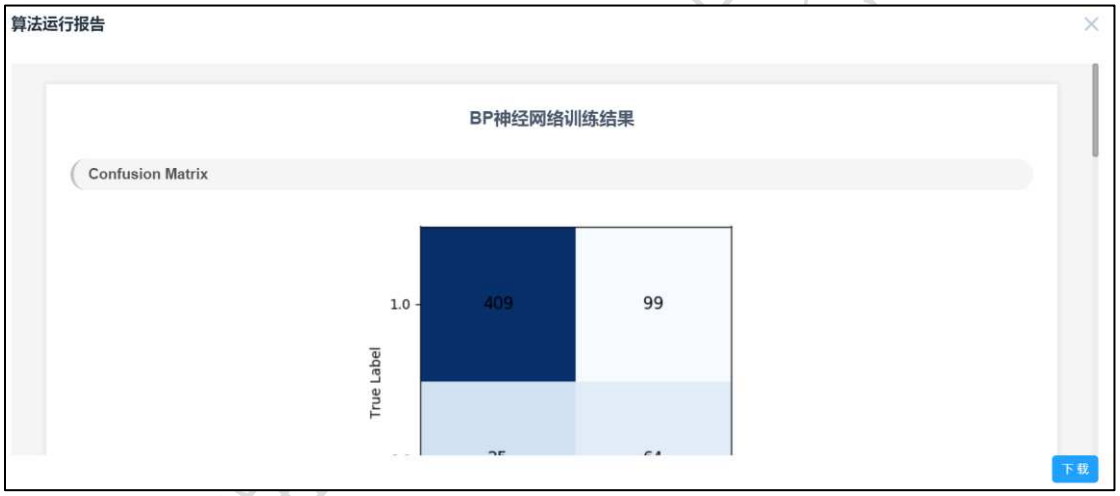

算法运行报告
BP神经网络训练结果
Confusion Matrix
1.0
409
99
True Label
下载

图 322

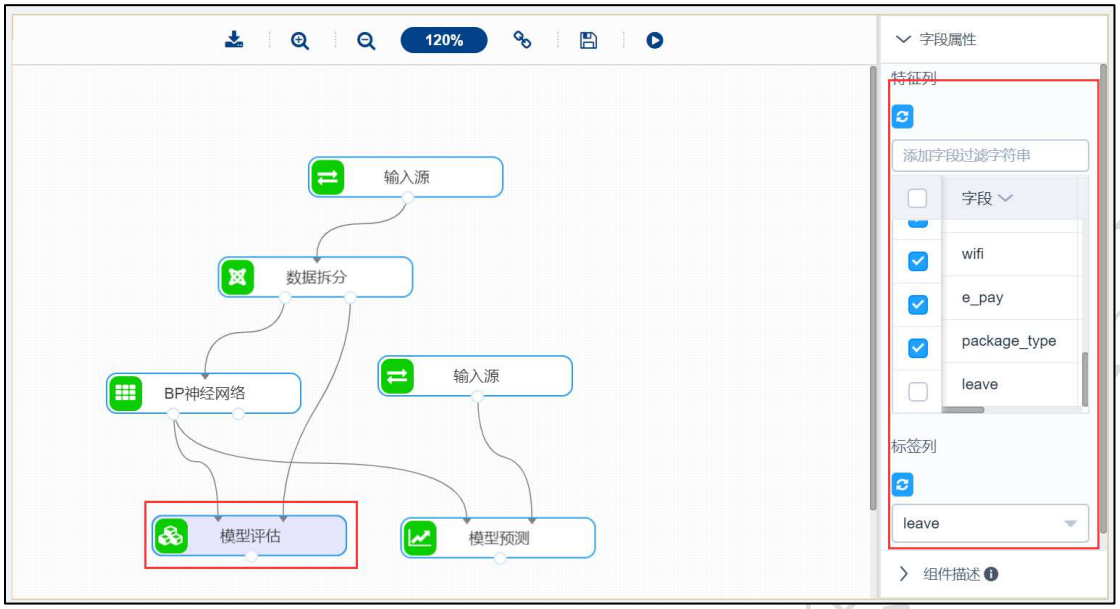

120%
输入源
数据拆分
BP神经网络
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
wifi
e_pay
package_type
leave
标签列
leave
组件描述

图 323

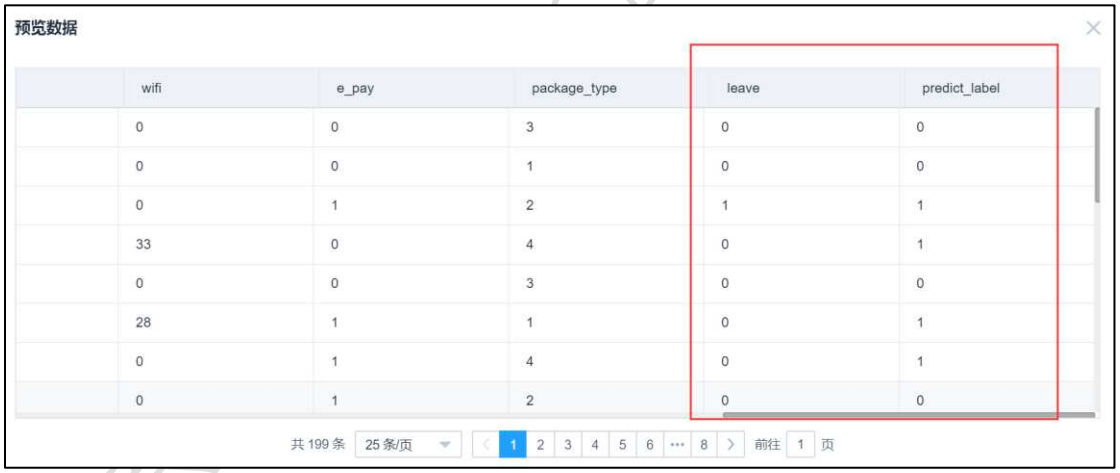

预览数据
wifi
e_pay
package_type
leave
predict_label
0 0 3 0 0
0 0 1 0 0
0 1 2 1 1
33 0 4 0 1
0 0 3 0 0
28 1 1 0 1
0 1 4 0 1
0 1 2 0 0
共 199 条
25 条/页
1 2 3 4 5 6 ··· 8
前往 1 页

图 324

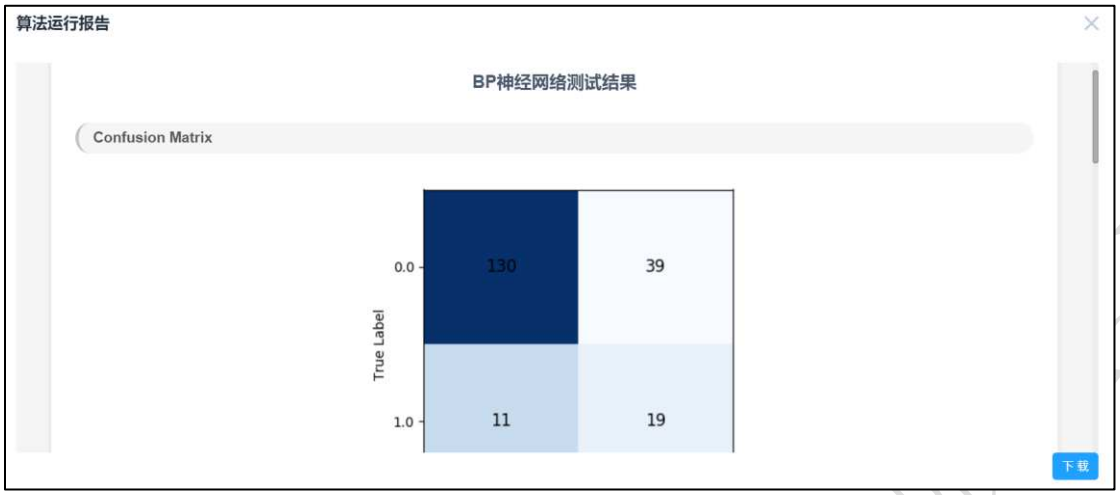

算法运行报告
BP神经网络测试结果
Confusion Matrix
0.0
1.0
True Label
130
39
11
19
下载

图 325

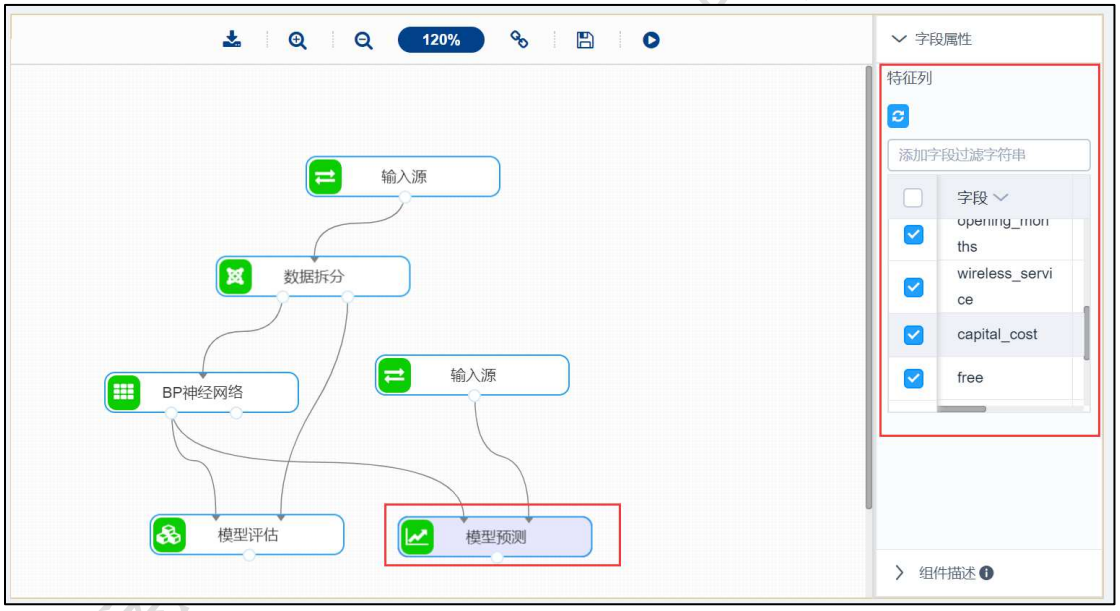

120%
输入源
数据拆分
BP神经网络
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
ths
wireless_service
capital_cost
free
组件描述

图 326

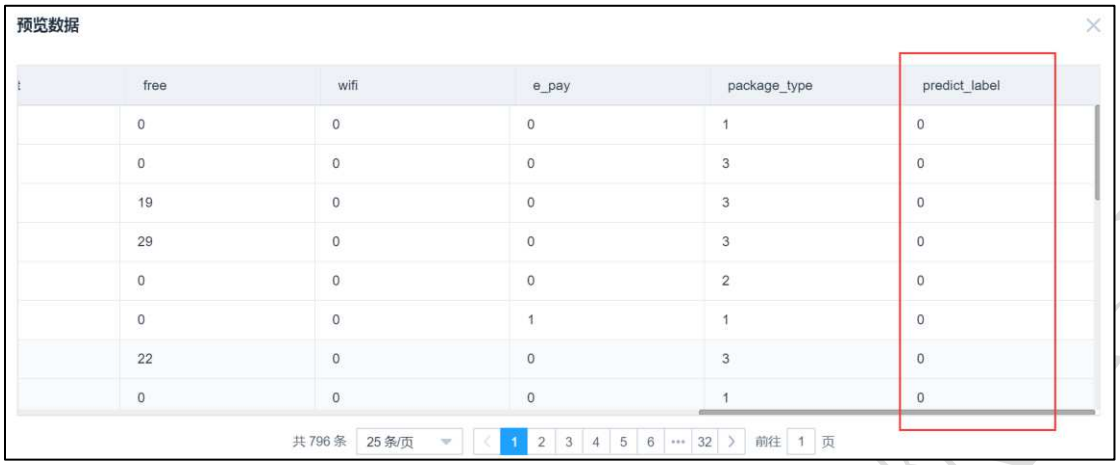


图 327

3.4.5.8 逻辑回归

图标： 

描述：逻辑回归是广义线性模型的一种。广义线性模型是一般线性模型的推广，即因变量均值的函数与解释变量是线性关系，即 $g(E(Y))=\beta X+\varepsilon$。其中 g 被称为连接函数。连接函数为 Logit 函数的广义线性模型就被称为逻辑回归。逻辑回归方程用解释变量预测事件发生的概率，所以可以用来处理分类问题。

字段属性

特征列：必选。通过勾选的方式选择特征所在列。

标签列：必选。选择响应变量所在的列，仅支持字符型数据。

参数设置

对于多分类问题的策略：采用 one-vs-rest 策略、采用多分类逻辑回归策略，默认采用 one-vs-rest 策略。

优化算法选择参数：包括 liblinear、newton-cg、lbfgs、sag、saga，默认 liblinear。

输出

表结果：逻辑回归算法分类结果。

报告：Confusion Matrix、Receiver Operating Characteristic(ROC)、Precision-Recall

示例

下列对某数据进行逻辑回归算法分类：

- 选择特征数列，标签列。如图 328 所示。
- 保留默认参数，策略为 one-vs-rest，优化算法选择为 liblinear，如图 329 所示。
- 逻辑回归运行成功后，可选择查看结果，如图 330 所示。
- 逻辑回归运行成功后，可选择查看报告，如图 331 所示。
- 模型评估配置如图 332 所示。
- 模型评估运行成功后，选择查看数据，如图 333 所示。
- 模型评估运行成功后，选择查看报告如所示。，
- 模型预测配置如图 335 所示。
- 模型预测运行成功后，选择查看数据，如图 336 所示。

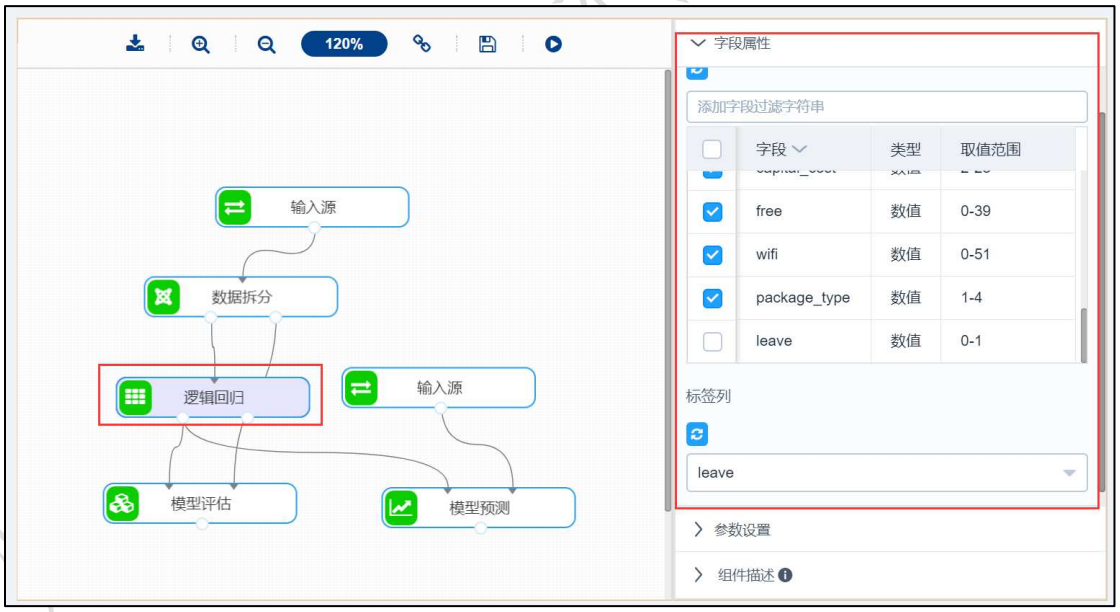


图 328


字段属性
参数设置
对于多分类问题的策略
采用one-vs-rest策略
优化算法选择参数
liblinear

图 329

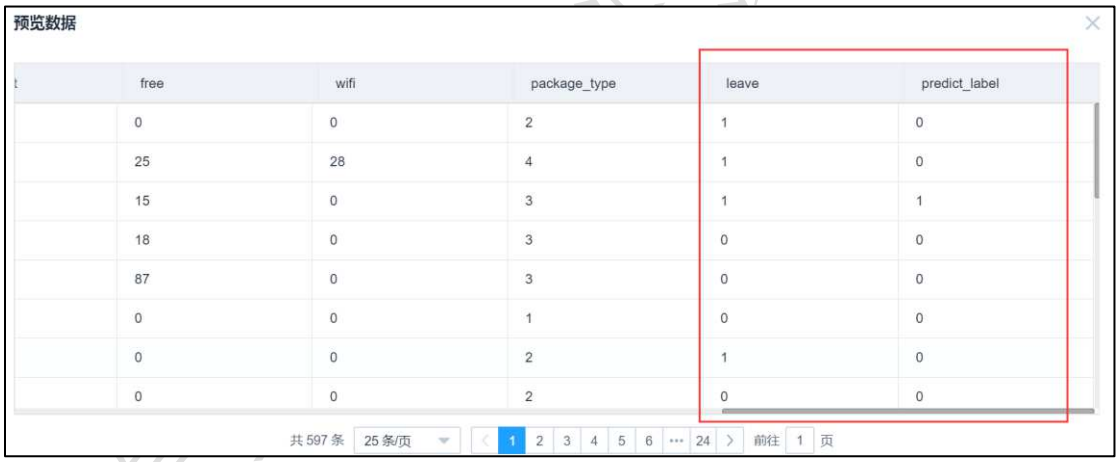

预览数据
free
wifi
package_type
leave
predict_label
0 0 2 1 0
25 28 4 1 0
15 0 3 1 1
18 0 3 0 0
87 0 3 0 0
0 0 1 0 0
0 0 2 1 0
0 0 2 0 0
共 597 条
25 条/页
1 2 3 4 5 6 ··· 24
前往 1 页

图 330

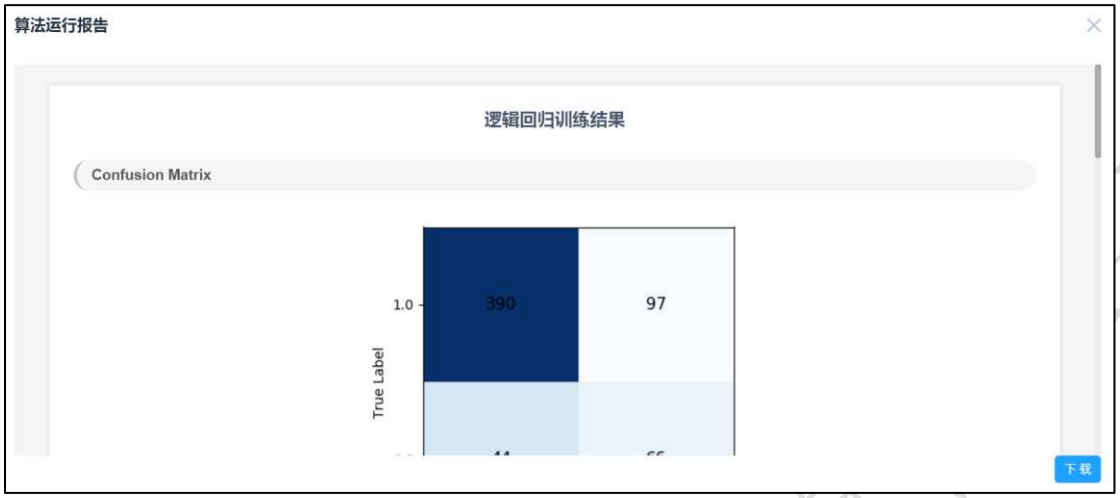

算法运行报告
逻辑回归训练结果
Confusion Matrix
1.0
True Label
390
97
下载

图 331

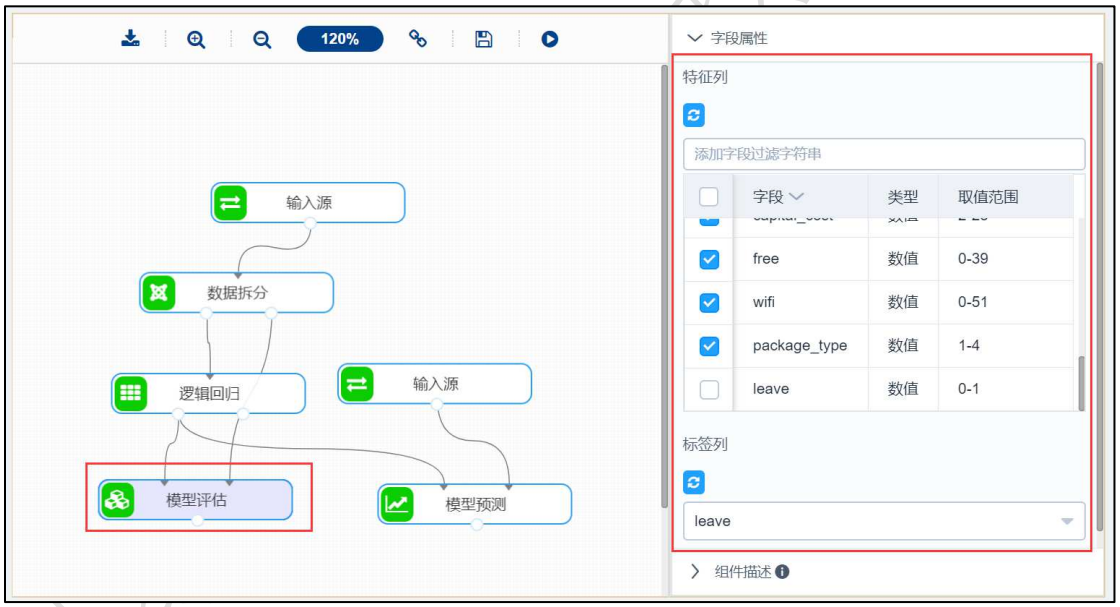

120%
输入源
数据拆分
逻辑回归
输入源
模型评估
模型预测
字段属性
特征列
添加字段过滤字符串
字段
类型
取值范围
free
数值
0-39
wifi
数值
0-51
package_type
数值
1-4
leave
数值
0-1
标签列
leave
组件描述

图 332

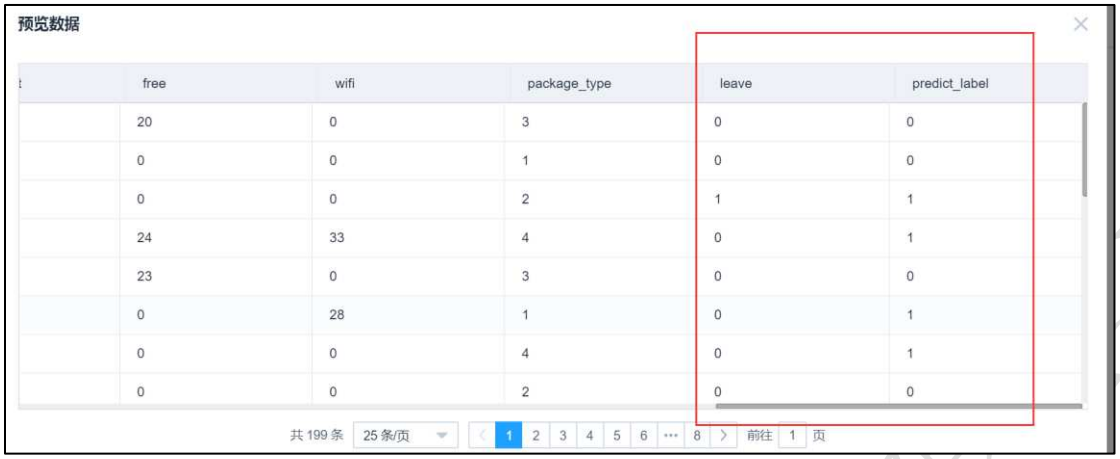

预览数据
free
wifi
package_type
leave
predict_label
20 0 3 0 0
0 0 1 0 0
0 0 2 1 1
24 33 4 0 1
23 0 3 0 0
0 28 1 0 1
0 0 4 0 1
0 0 2 0 0
共 199 条
25 条/页
1 2 3 4 5 6 8
前往 1 页

图 333

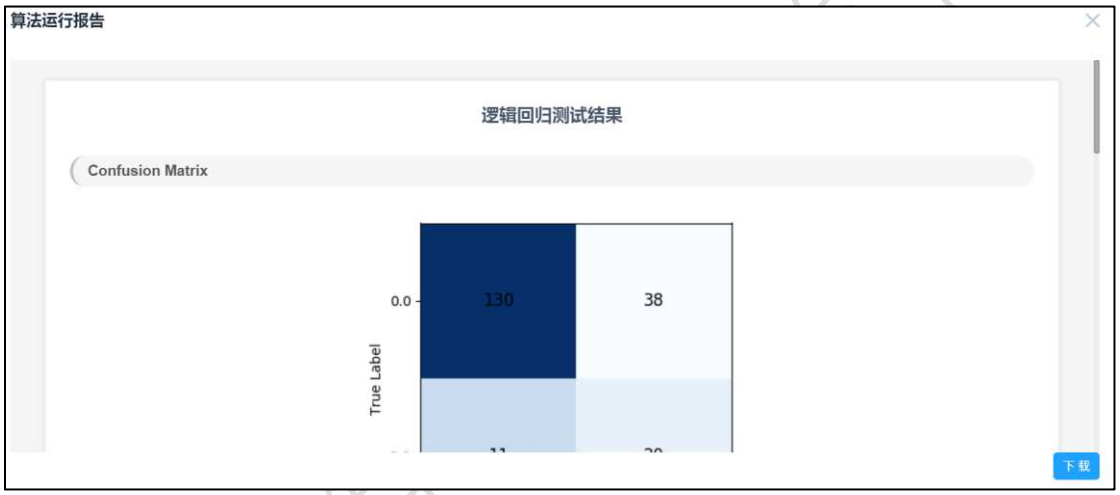

算法运行报告
逻辑回归测试结果
Confusion Matrix
0.0
130
38
True Label
下载

图 334

图 335

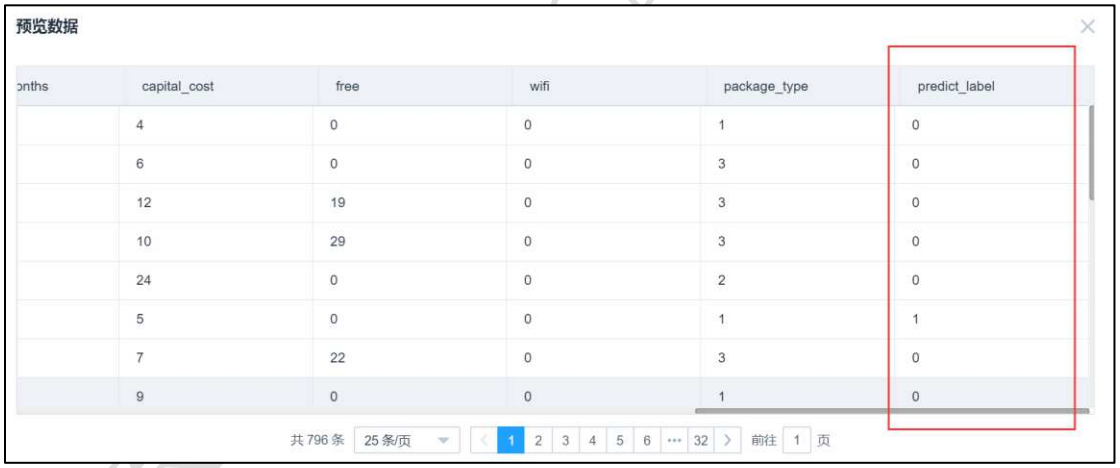


图 336

3.4.6 关联分析

3.4.6.1 Apriori

图标： Apriori

描述： Apriori 是关联规则里一项基本算法，目的就是在一个数据集中找出项与项之间的关

系，也被称为购物蓝分析 (Market Basket analysis)。

字段属性

项：必选。包含若干个项目的集合（一次事务中的），一般会大于 0 个。可为所有项目所在的列，必须要求每一行为一个事务，事务内项与项之间以空格隔开，格式如图 337 所示。

图 337

参数设置

最小支持度：必选。默认 0.5，以[A,B]这个关联规则为例来说明,表示 A、B 同时使用的人数占所有用户数的比例。

最小置信度：必选。默认 0.7，以[A,B]这个关联规则为例来说明,表示使用 A 的用户中同时使用 B 的比例，即同时使用 A 和 B 的人占使用 A 的人的比例。

提升度：默认 2，以[A,B]这个关联规则为例来说明,表示“使用 A 的用户中同时使用 B 的比例”与“使用 B 的用户比例”的比值。

最大项目数：默认为 1。

输出

表结果：规则（前项、后项、支持度、置信度、提升度）。

报告：无。

示例

下面对某数据进行关联分析。原数据如图 338 所示。

图 338

- 首先选择关联分析所需的数据，如图 339 所示。
- 设置相应参数，如图 340 所示。
- 运行成功，可选择查看数据，如图 341 所示。

图 339

图 340

预览数据

lhs	rhs	support	confidence	lift
{F4,A3}	H4	0.07849462365591398	0.8795180722891566	1.9709682101901582
{B2,F4}	H4	0.06236559139784946	0.7945205479452054	1.7804918303350388
{E3,C2}	D2	0.09247311827956989	0.7543859649122807	1.8708771929824564
{F4,C3}	H4	0.07526881720430108	0.875	1.9608433734939759
{D2,H4,F3}	A2	0.06236559139784946	0.7532467532467533	1.9732943113224803

图 341

3.4.6.2 HotSpot

图标： HotSpot

描述：HotSpot 是一种主要用于挖掘数值型特征与类别型标签之间的关联规则的算法。HotSpot 能够发现在某个取值区间内，某个类别发生的概率。

字段属性

前项：请选择数值型数据。

后项：请选择类别特征型数据。

如图 342 所示。

图 342

参数设置

最小支持度：前项与后项同时发生的概率，默认为 0.05。

最小置信度：前项发生的情况下后项发生的概率，默认为 0.2。

步长：步长越大，运行越快，模型效果越低。取值范围为 0-100，建议为 5-20 之间。

如图 343 所示。

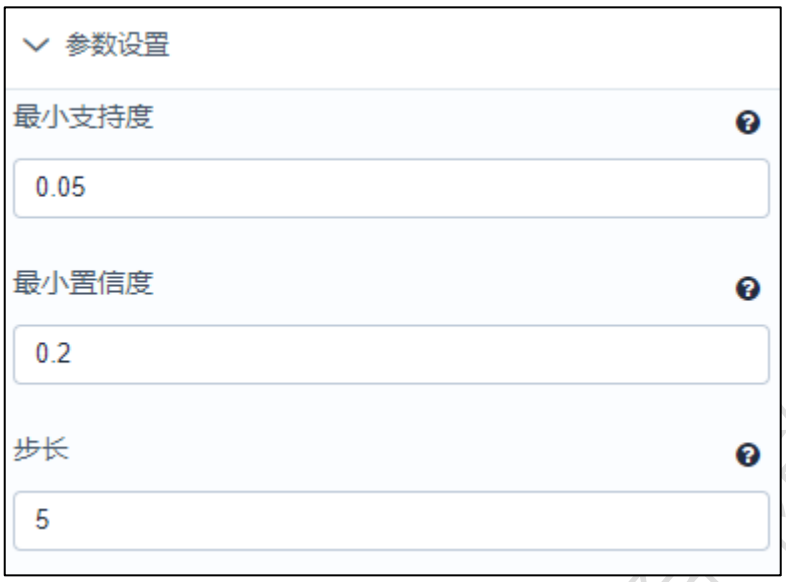


图 343

输出

表结果：规则。

报告：无。

示例

下面对某数据进行 Hotspot 关联规则分析。

- 选择两列序列作为前项，数据必须为数值型；选择类别特征型序列作为后项。如图 344 所示。
- 点击参数设置，设置如图 345 所示。
- 运行该组件，对组件右击，选择查看数据，结果如图 346 所示。

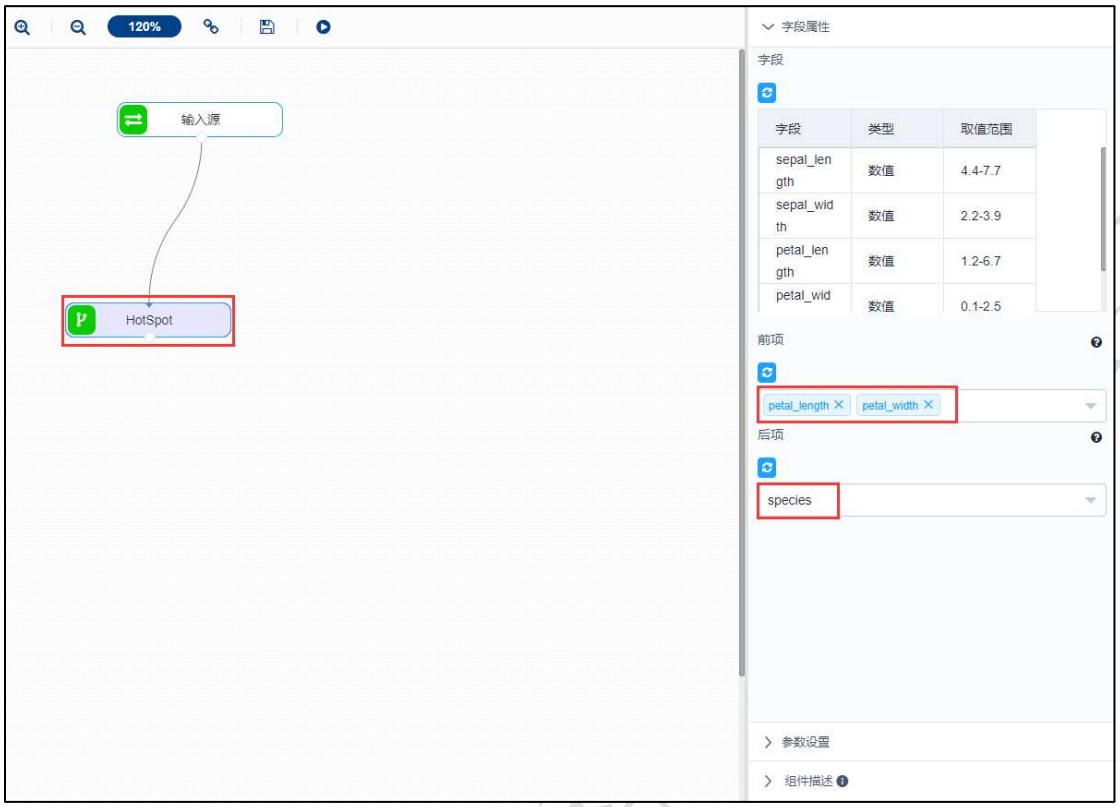

120%
输入源
HotSpot
字段属性
字段
字段
类型
取值范围
sepal_length
数值
4.4-7.7
sepal_width
数值
2.2-3.9
petal_length
数值
1.2-6.7
petal_wid
数值
0.1-2.5
前项
petal_length
petal_width
后项
species
参数设置
组件描述

图 344

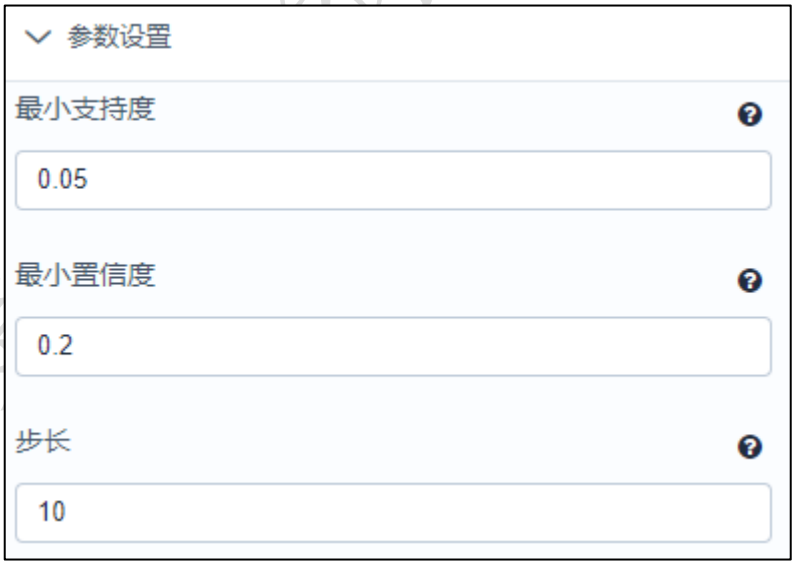

参数设置
最小支持度
0.05
最小置信度
0.2
步长
10

图 345

预览数据

conf	left_obj	right_obj	sup
1	((1.0, 1.42), (0.1, 0.20500000000000002))	setosa	0.12
1	((3.66, 4.29), (0.415000...000000004, 1.5550000000000002))	versicolor	0.11333333333333333
1	((3.66, 4.29), (0.31, 2.08))	versicolor	0.11333333333333333
1	((3.66, 4.29), (0.31, 2.185))	versicolor	0.11333333333333333
1	((3.66, 4.29), (0.31, 2.29))	versicolor	0.11333333333333333
1	((3.66, 4.29), (0.31, 2.3949999999999996))	versicolor	0.11333333333333333

共 32188 条 25 条/页 1 2 3 4 5 6 ··· 1288 前往 1 页

图 346

3.4.6.3 FP-growth

图标： FP-growth

描述： FP-growth 将事务数据表中的各个事务数据项按照支持度排序后，把每个事务中的数据项按降序依次插入到一棵以 NULL 为根结点的树中，同时在每个结点处记录该结点出现的支持度。

字段属性

项：必选。包含若干个项目的集合（一次事务中的），一般会大于 0 个。可为所有项目所在的列，必须要求每一行为一个事务，事务内项与项之间以空格隔开。

参数设置

最小支持度：必选。A、B 同时包含的概率。

最小置信度：必选。发生概率大于多少时认为可信。

输出

表结果：规则、频繁项集。

报告：无。

示例

- 首先选择关联分析所需的数据，如图 347 所示。
- 设置相应参数，如图 348 所示。
- 运行成功，可选择查看数据，如图 349、图 350 所示。

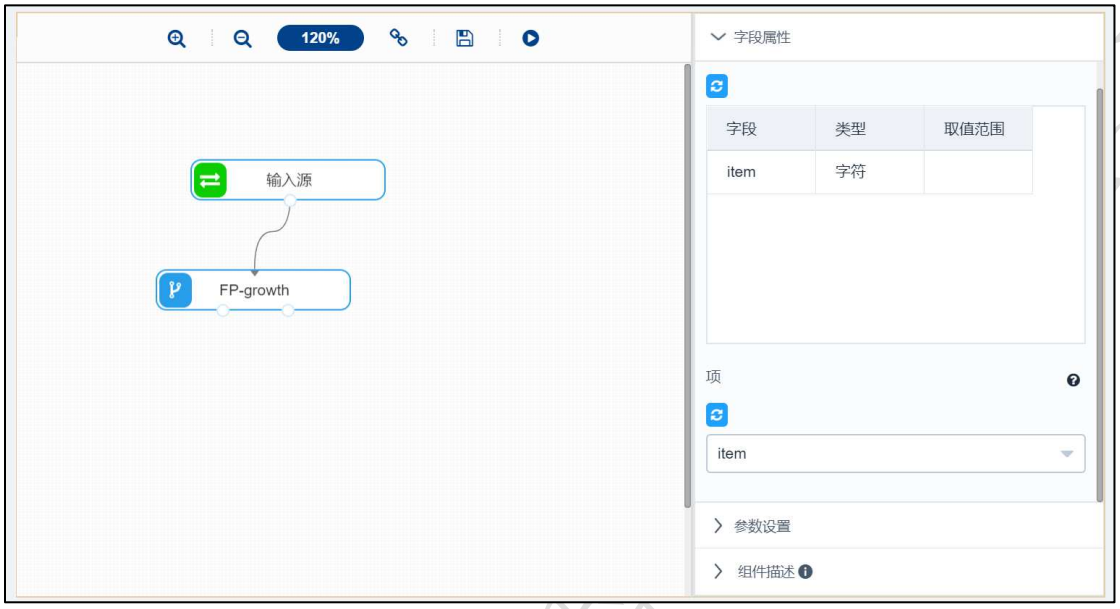


图 347

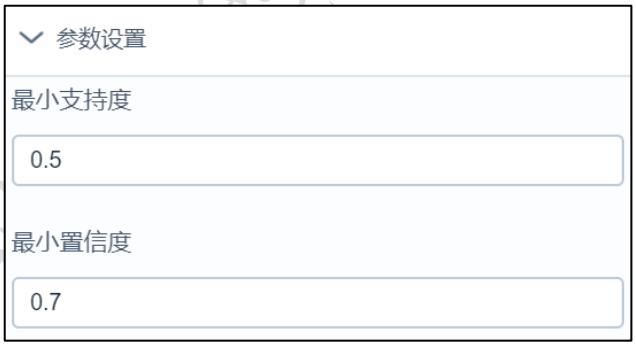


图 348


预览数据

lhs	rhs	confi
A1	E5	1
F6	E5	1
B2	(C3,D4,E5)	1
E5	(B2,C3,D4)	0.75
C3	(B2,D4,E5)	1
(B2,C3)	(D4,E5)	1
(B2,E5)	(C3,D4)	1
(C3,E5)	(B2,D4)	1

共 16 条 25 条/页 1 前往 1 页

图 349


预览数据

item	support
A1	0.5
(A1,E5)	0.5
F6	0.5
(E5,F6)	0.5
B2	0.75
(B2,E5)	0.75
C3	0.75
(B2,C3)	0.75

共 19 条 25 条/页 1 前往 1 页

图 350

3.4.7 聚类

3.4.7.1 K-Means

图标： K-Means

描述： K-Means 是 Mac Queen 提出的一种非监督实时聚类算法，在最小化误差函数的基础上将数据划分为预定的类数 K。

字段属性

特征列：需要进行聚类的列，请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。如图 351 所示。

图 351

参数设置

聚类个数：聚类的个数，默认 3。

最大迭代次数：迭代的次数。

如图 352 所示。

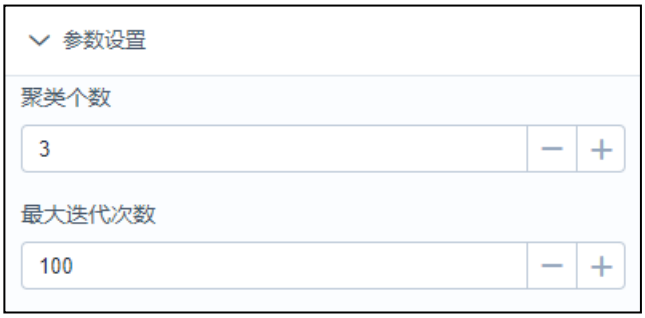


图 352

输出

表结果：包含聚类结果的数据表。

报告：聚类中心、饼图。

示例

下面对某数据进行 K-Means 聚类。

- 选择待聚类的序列，数据必须为数值型。如图 353 所示。
- 点击参数设置，聚类个数设置为 3，最大迭代次数设置为 100。如图 354 所示。
- 运行该组件，对组件右击，选择查看数据与报告，结果如图 355 与图 356 所示。

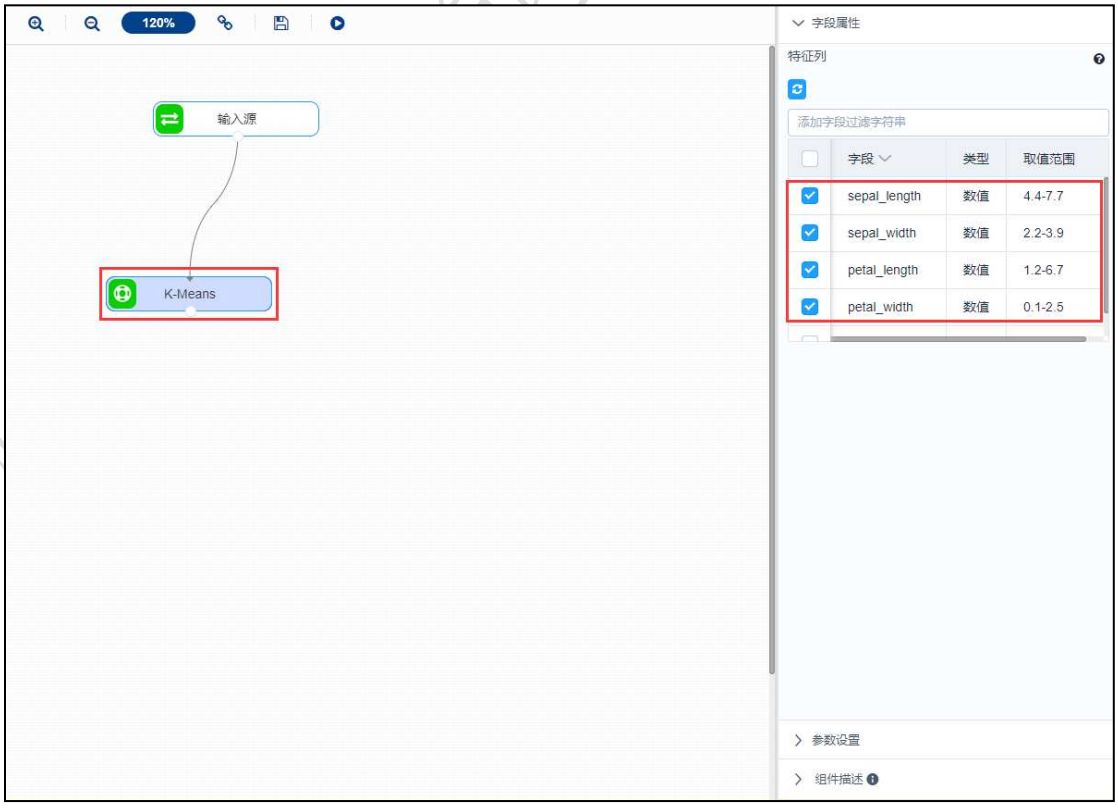


图 353

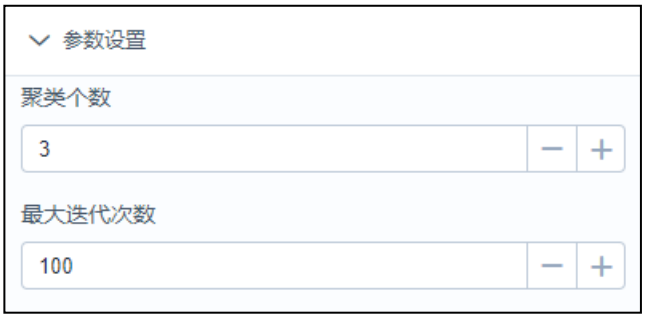


图 354

sepal_length	sepal_width	petal_length	petal_width	cluster_id
5.1	3.5	1.4	0.2	2
4.9	3	1.4	0.2	2
4.7	3.2	1.3	0.2	2
4.6	3.1	1.5	0.2	2
5	3.6	1.4	0.2	2
5.4	3.9	1.7	0.4	2
4.6	3.4	1.4	0.3	2
5	3.4	1.5	0.2	2

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 355

聚类中心

cluster_id	sepal_length	sepal_width	petal_length	petal_width
1	5.901612903225806	2.7483870967741937	4.393548387096774	1.4338709677419355
2	5.006	3.428	1.4619999999999997	0.24600000000000033
3	6.85	3.0736842105263156	5.742105263157894	2.0710526315789473

饼图

图 356

3.4.7.2 GMM(高斯混合模型)

图标： 

描述：GMM（高斯混合模型）是用高斯概率密度函数精确地量化事物，将一个事物分解为若干的基于高斯概率密度函数形成的模型。

字段属性

特征列：需要进行聚类的列，请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。如图 357 所示。

图 357

参数设置

聚类个数：聚类的个数，默认 2。

指定协方差类型：包括球状型、结点型、对角型、全型，其中球状型：所有的分模型的协方差矩阵都是一个标量值；结点型：所有的分模型都共享一个协方差矩阵；对角型：每个分模型的协方差矩阵都是对角矩阵；全型：每个分模型都有自己的协方差矩阵。

指定初始化次数：默认为 1。

指定初始化权重的策略：默认为 kmeans。

如图 358 所示。

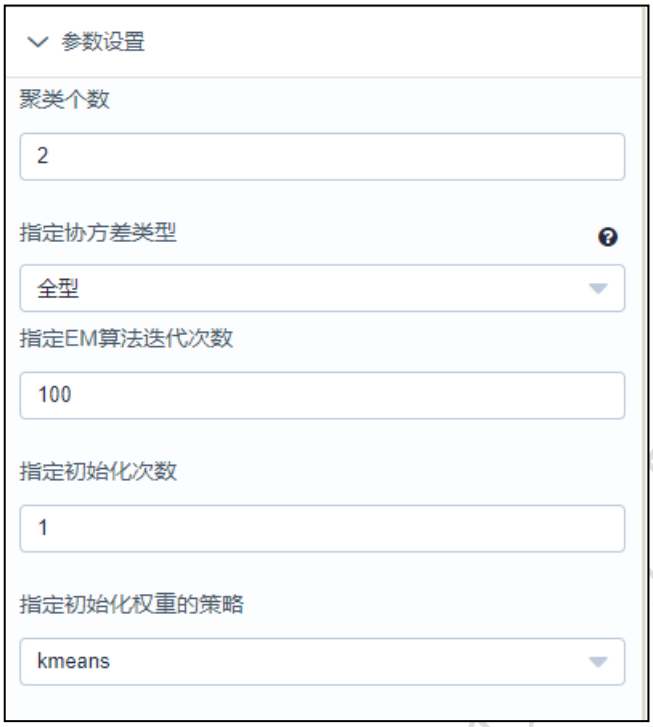


图 358

输出

表结果：包含聚类结果的数据表。

报告：无。

示例

下面对某数据进行 GMM 聚类。

- 选择待聚类的序列，数据必须为数值型。如图 359 所示。
- 运行该组件，对组件右击，选择查看数据，结果如图 360 所示。

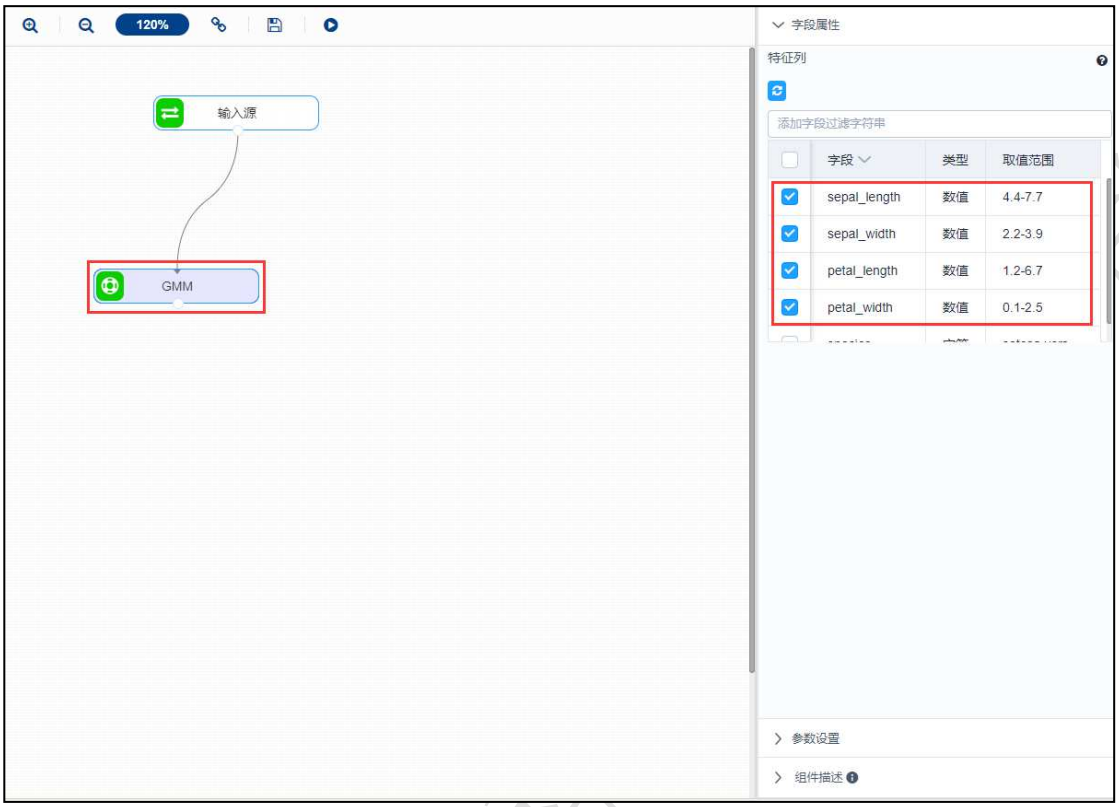


图 359

sepal_length	sepal_width	petal_length	petal_width	cluster_id
5.1	3.5	1.4	0.2	1
4.9	3	1.4	0.2	1
4.7	3.2	1.3	0.2	1
4.6	3.1	1.5	0.2	1
5	3.6	1.4	0.2	1
5.4	3.9	1.7	0.4	1
4.6	3.4	1.4	0.3	1
5	3.4	1.5	0.2	1

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 360

3.4.7.3 密度聚类

图标： 

描述： 密度聚类的核心思想是从某个核心点出发，不断向密度可达的区域扩张，从而得

到一个包含核心点和边界点的最大化区域，区域中任意两点密度相连。对于噪声样本，其簇标记为-1。

字段属性

特征列：需要进行聚类的列，请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。如图 361 所示。

图 361

参数设置

邻域半径：设置某个半径，默认 0.5。

邻域内最小数目：设置邻域内最小点的个数，默认 5。

如图 362 所示。

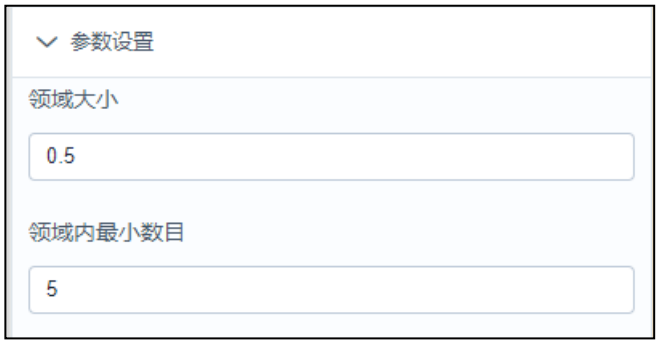


图 362

输出

表结果：包含聚类结果的数据表。

报告：无。

示例

下面对某数据进行密度聚类。

- 选择待聚类的序列，数据必须为数值型，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。。如图 363 所示。
- 运行该组件，对组件右击，选择查看数据，结果如图 364 所示。

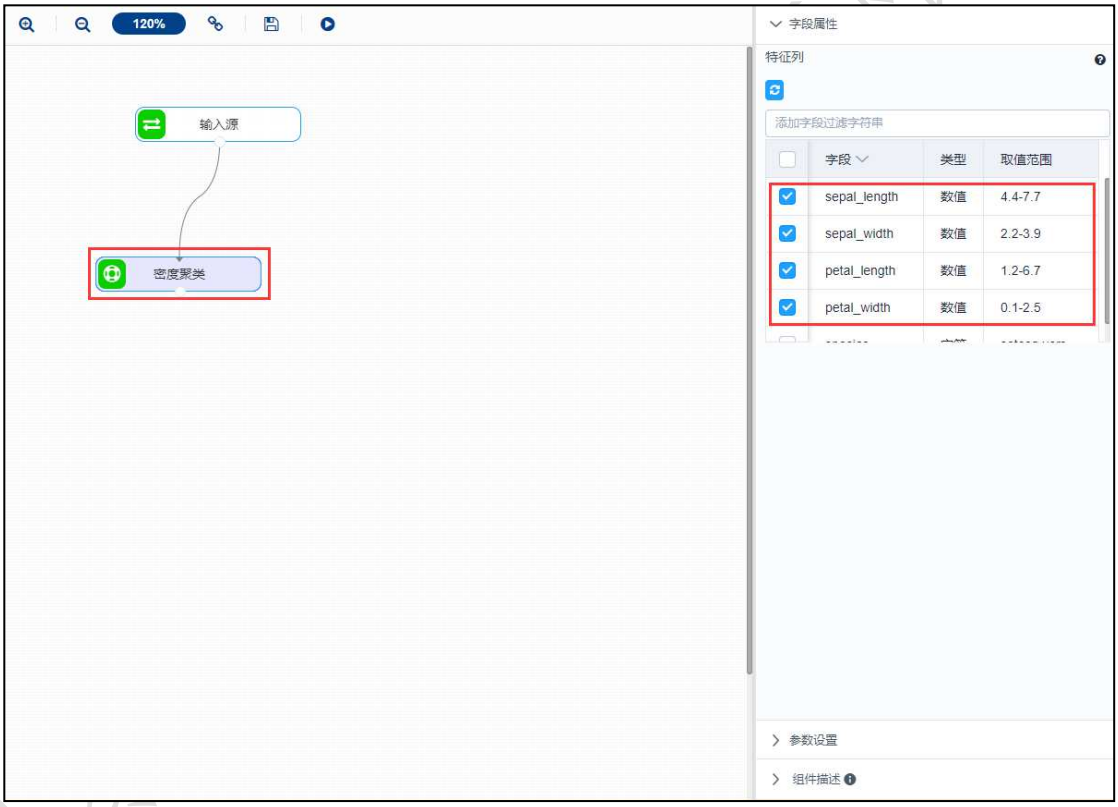


图 363

sepal_length	sepal_width	petal_length	petal_width	cluster_id
5.1	3.5	1.4	0.2	0
4.9	3	1.4	0.2	0
4.7	3.2	1.3	0.2	0
4.6	3.1	1.5	0.2	0
5	3.6	1.4	0.2	0
5.4	3.9	1.7	0.4	0
4.6	3.4	1.4	0.3	0
5	3.4	1.5	0.2	0

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 364

3.4.7.4 K-Medoids

图标： K-Medoids

描述： K-Medioids 是 Kmeans 算法的改进，减轻了 Kmeans 算法对孤立点的敏感性，选用簇中离平均值最近的对象作为簇中心。

字段属性

特征列：需要进行聚类的列，请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。如图 365 所示。

图 365

参数设置

聚类个数：聚类的个数，默认 3。

最大迭代次数：迭代的次数。

如图 366 所示。

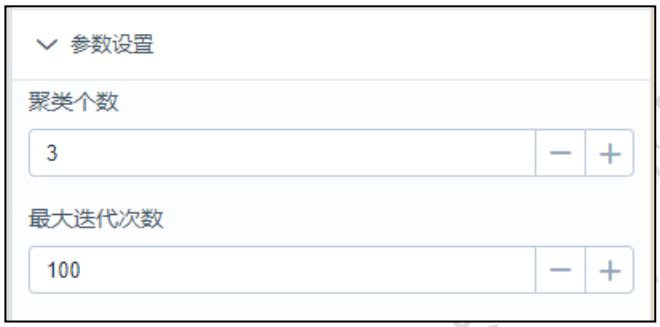


图 366

输出

表结果：包含聚类结果的数据表。

报告：无。

示例

下面对某数据进行 K-Medoids。

- 选择待聚类的序列，数据必须为数值型。如图 367 所示。
- 运行该组件，对组件右击，选择查看数据和报告，结果如图 368 与图 369 所示。

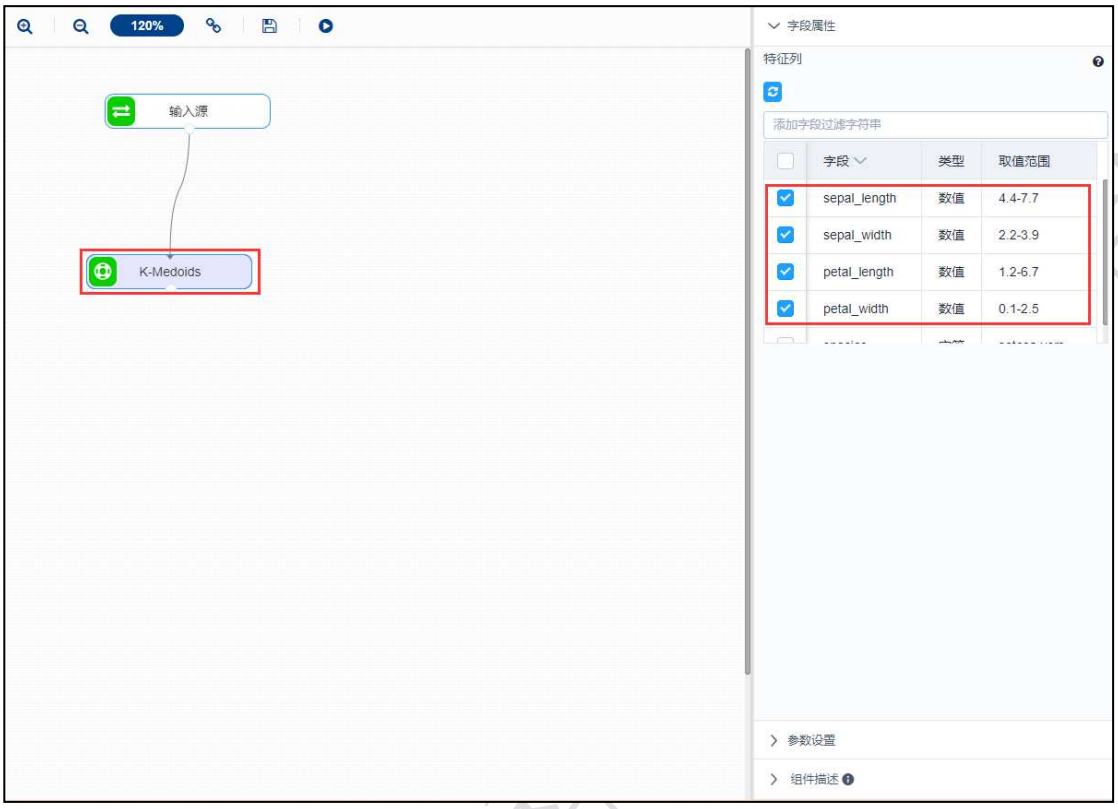


图 367

sepal_length	petal_width	sepal_width	petal_length	cluster_id
5.1	0.2	3.5	1.4	3
4.9	0.2	3	1.4	3
4.7	0.2	3.2	1.3	3
4.6	0.2	3.1	1.5	3
5	0.2	3.6	1.4	3
5.4	0.4	3.9	1.7	3
4.6	0.3	3.4	1.4	3
5	0.2	3.4	1.5	3

共 150 条 25 条/页 1 2 3 4 5 6 前往 1 页

图 368

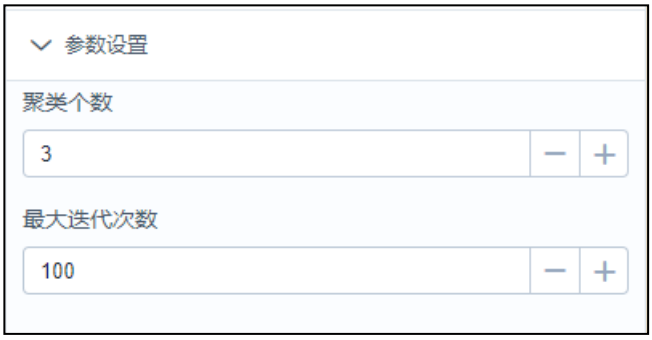


图 369

3.4.7.5 层次聚类

图标： 层次聚类

描述：层次聚类也叫系统聚类，分类单位所处的位置越低，其所包含的个体越少，但这些个体间的共同特征越多。

字段属性

特征列：需要进行聚类的列，请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。如图 370 所示。

图 370

参数设置

输出聚类数：默认显示 2 类。如图 371 所示。

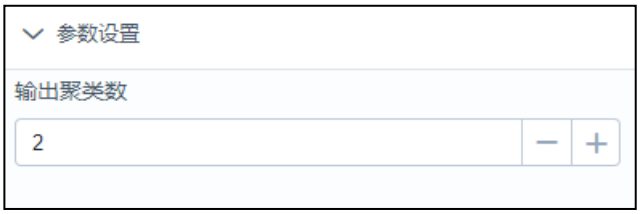


图 371

输出

表结果：包含聚类结果的数据表。

报告：无。

示例

下面对某数据进行层次聚类。

- 选择待聚类的序列，数据必须为数值型。如图 372 所示。
- 运行该组件，对组件右击，选择查看数据，结果如图 373 所示。

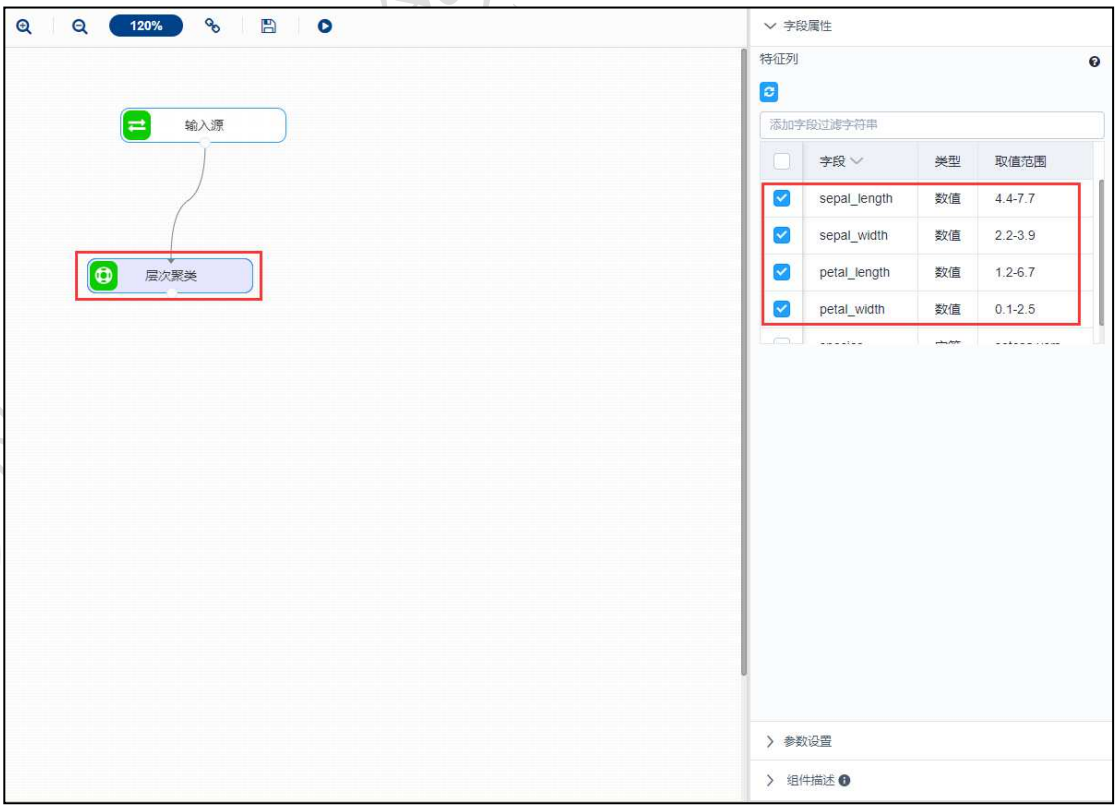


图 372

sepal_length	sepal_width	petal_length	petal_width	cluster_id
5.1	3.5	1.4	0.2	2
4.9	3	1.4	0.2	2
4.7	3.2	1.3	0.2	2
4.6	3.1	1.5	0.2	2
5	3.6	1.4	0.2	2
5.4	3.9	1.7	0.4	2
4.6	3.4	1.4	0.3	2
5	3.4	1.5	0.2	2

图 373

3.4.8 时间序列

3.4.8.1 GM11

图标： 

描述：GM(1,1)模型是灰色系统理论中应用最广泛的一种灰色动态预测模型,该模型由一个单变量的一阶微分方程构成。

字段信息

特征列：请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列，当勾选多列时，则对每一列进行灰色预测，如图 374 所示。

图 374

参数设置

预测个数：数值型，默认 1，如图 375 所示。

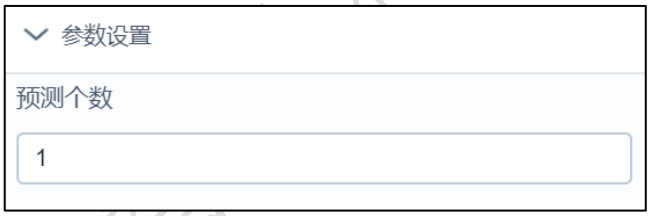


图 375

输出

表结果：预测结果。

报告：各序列方差比、小残差概率。

示例

下面对某数据的各序列进行灰色预测，原数据如图 376 所示。

- 选择需要进行预测的序列。
- 选择各序列预测的个数，如图 377 所示。
- 运行成功可查看报告，如图 378 所示。
- 运行成功可查看数据。

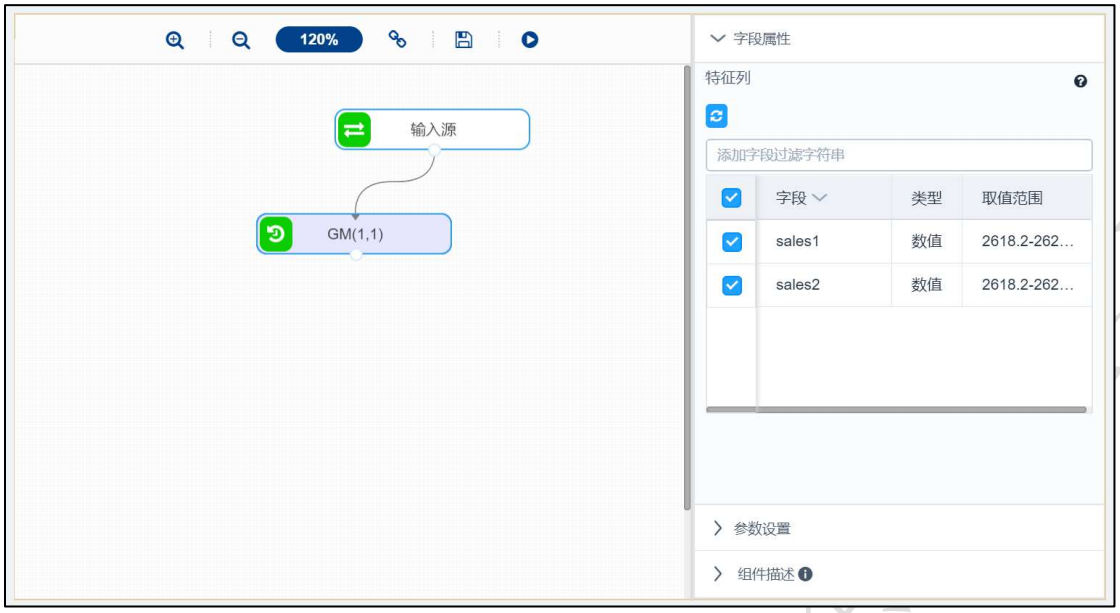


图 376

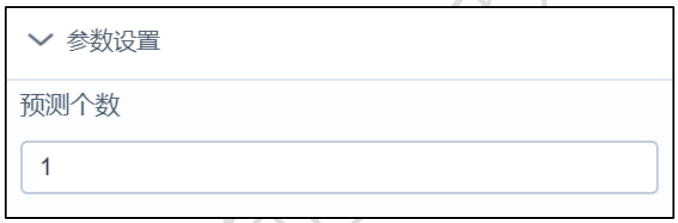


图 377

图 378

3.4.8.2 ARIMA

图标： 

描述： ARIMA 模型全称为自回归移动平均模型(Autoregressive Integrated Moving Average Model，简记 ARIMA)，其中 ARIMA（p，d，q）称为差分自回归移动平均模型，AR 是自回归，p 为自回归项；MA 为移动平均，q 为移动平均项数，d 为时间序列成为平稳时所做的差分次数。

字段属性

时序列：请选择数值型数据，如图 379 所示。

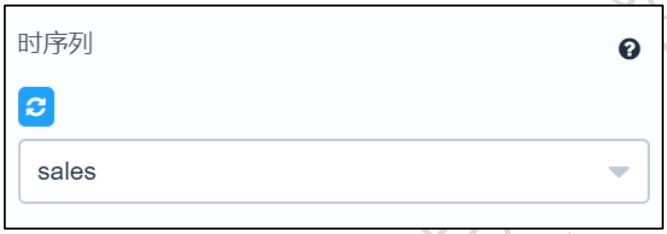


图 379

参数设置

是否构建季节性 ARIMA 模型：默认否。

P：自回归阶数

D：差分阶数

Q：移动平均阶数

P(季节性)：季节性自回归阶数

D(季节性)：季节性移动平均阶数

Q(季节性)：季节性差分阶数

时序周期：一般情况下季节性周期数据为 4，月度周期数据为 12。

输出

表结果：表结果包含预测值、残差、残差的平方。

报告：系数表、模型检验图、预测图、残差及残差平方图。

示例

下面对某数据构造 ARIMA 模型。

- 勾选时序列，如图 380 所示。

- 设置相应参数，如图 381、图 382 所示。
- 运行成功后，选择查看数据，如图 383 所示。
- 运行成功后，选择查看报告，如图 384 所示。

图 380

图 381

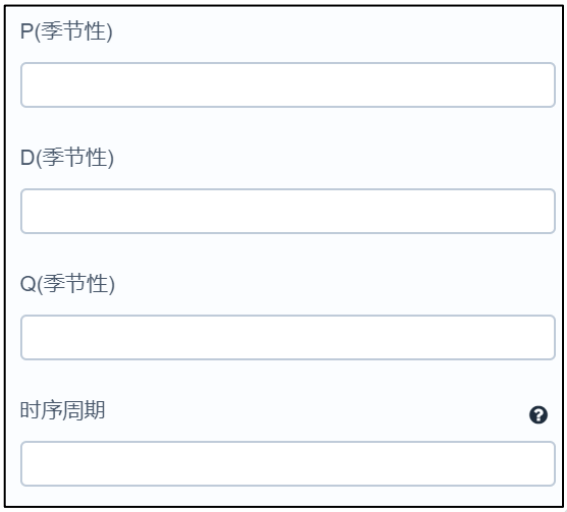


图 382

预览数据

sales	predict	at	at2
51	0	51	2601
2618	47.92111853824536	2570.0788814617545	6605305.456935703
2608	2459.9507516299286	148.04924837007138	21918.579942943084
2652	2450.554453877331	201.44554612266893	40580.30805266034
3442	2491.898163988759	950.1018360112412	902693.4987919314
3393	3234.205686443932	158.79431355606812	25215.63401774288
3137	3188.163827456206	-51.163827456206036	2617.7372399684227
3744	2947.6186049897196	796.3813950102804	634223.3263185202

共 201 条 25 条/页 1 2 3 4 5 6 ••• 9 前往 1 页

图 383

图 384

3.4.8.3 指数平滑

图标： 指数平滑

描述：指数平滑法是平滑法的一种，常用于趋势分析和预测，利用修匀技术，削弱短期随机波动对序列的影响，使序列平滑化。

字段属性

序列：仅支持数值型数据。

参数设置

指数平滑模型：可选择简单指数平滑模型，两参数指数平滑模型，三参数指数平滑模型。

alpha：默认 0.1.

预测值数量：默认 1.

HoltMethod 的 beta 值：默认 0.5，Holt 双参数线性指数平滑法、Winter 线性和季节性指数平滑法有效。

Holt-Winters 的 gamma 值：默认 0.5，Holt 双参数线性指数平滑法、Winter 线性和季节性指数平滑法有效。

季节长度：默认 0，Winter 线性和季节性指数平滑法有效。

输出

表结果：预测结果。

报告：无。

示例

下面对某数据平滑化。

- 勾选时序列，如图 385 所示。
- 设置相应参数，如图 386、图 387 所示。
- 运行成功后，选择查看数据。

图 385

图 386

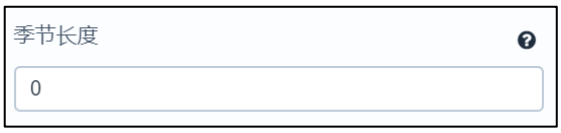


图 387

3.4.8.4 ARCH 模型

图标： ARCH

描述： ARCH 指“自回归条件异方差模型”。

字段属性

时序列：请选择数值型数据，如图 388 所示。

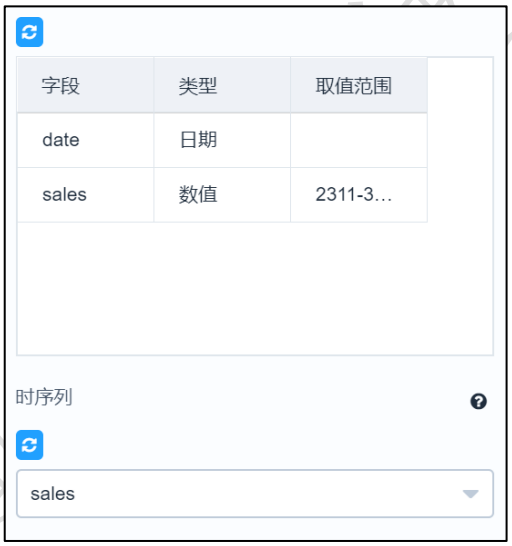


图 388

参数设置

自回归阶数 AR(?)：默认 2.

arch 项阶数 ARCH(?)：默认 1。如图 389 所示。

图 389

输出

表结果：无。

报告：系数表、标准误差、P 值、置信区间、整体的预测拟合图。

示例

下面对某数据 ARCH 模型。。

- 模型配置，如图 390 所示。
- 设置相应参数，如图 391 所示。
- 运行成功后，选择查看报告，如图 392 所示。

图 390

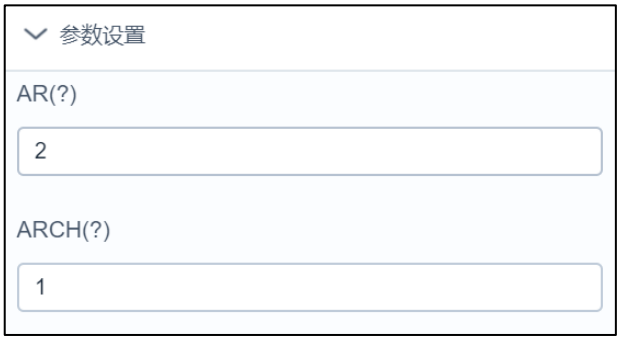


图 391

图 392

3.4.8.5 GARCH 模型

图标： GARCH

描述：GARCH（全称：Generalized AutoRegressive Conditional Heteroskedasticity），又称“广义 ARCH 模型（Generalized ARCH）”、“广义自回归条件异方差模型”。

字段属性

时序列：请选择数值型数据，如图 393 所示。

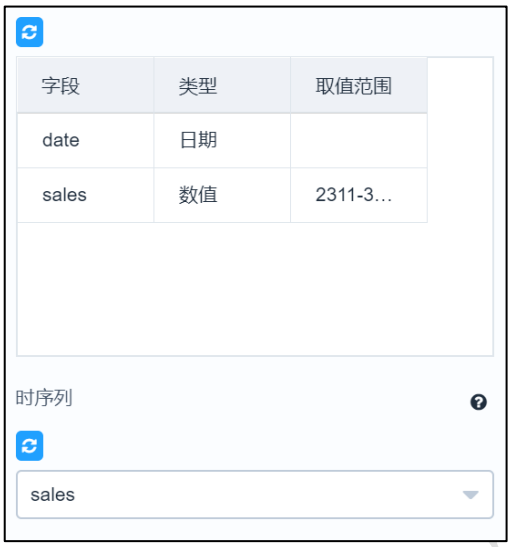


图 393

参数设置

AR(?)：默认 1.

GARCH(?,)：默认 1.

GARCH(,?)：默认 1.。如图 394 所示。

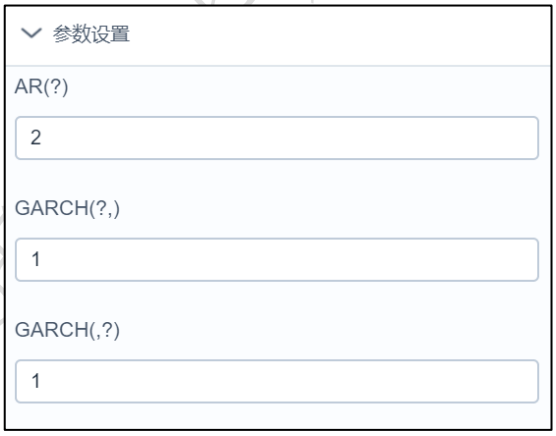


图 394

输出

表结果：无。

报告：系数表、标准误差、P 值、置信区间、整体的预测拟合图。

示例

下面对某数据 CARCH。。

- 模型配置，如图 395 所示。
- 设置相应参数，如图 396 所示。
- 运行成功后，选择查看报告，如图 397 所示。

图 395

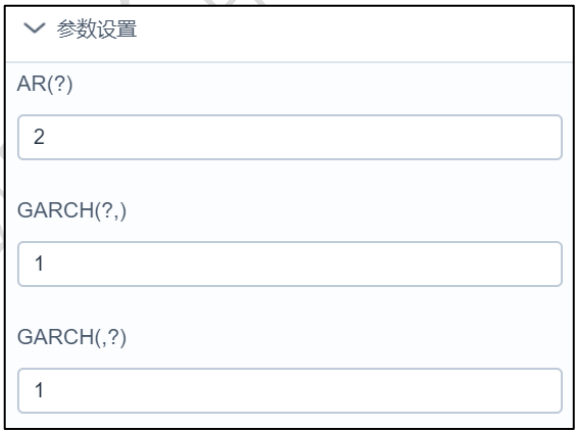


图 396

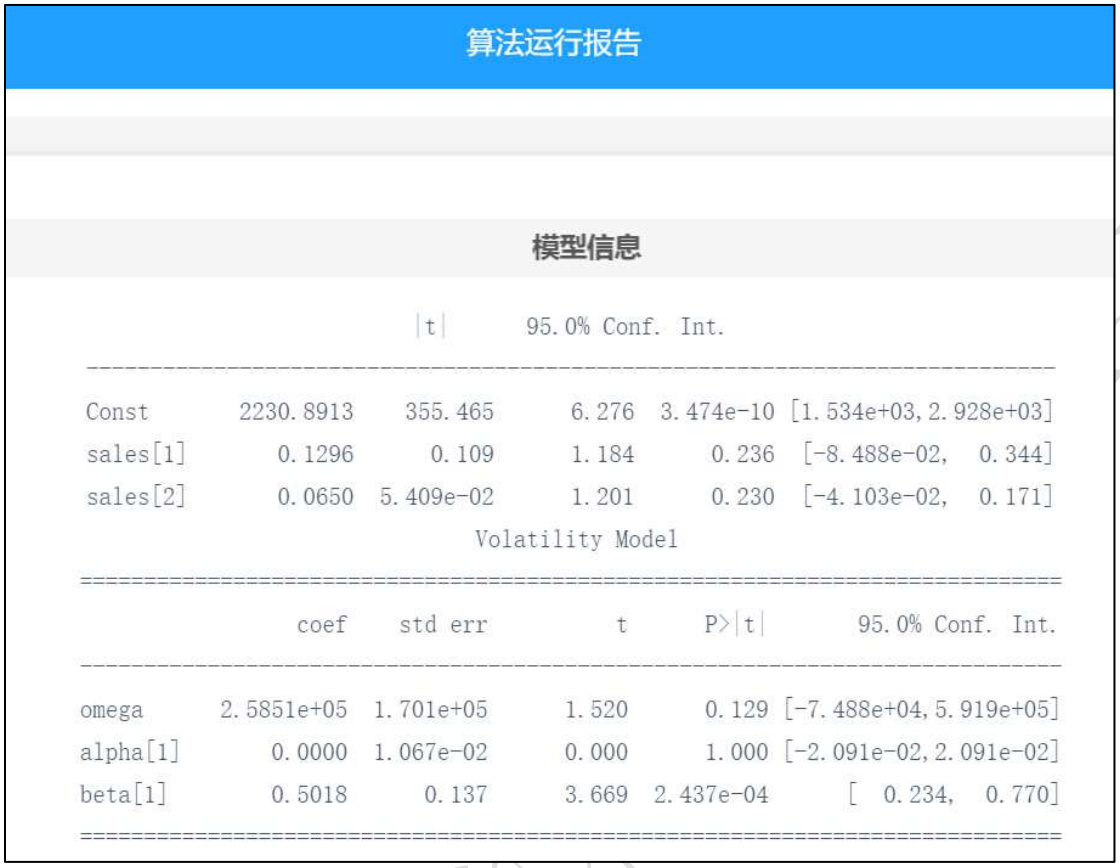


图 397

3.4.9 模型评估

3.4.9.1 模型评估

图标： 模型评估

描述：针对分类算法而言，可对训练集构造的模型使用测试集进行评估。

示例

下列对某数据进行 CART 回归树算法：

- CART 回归树模型评估配置如图 398 所示。
- CART 回归树模型评估结果如图 399 所示。.
- CART 回归树模型评估如图 400 所示。

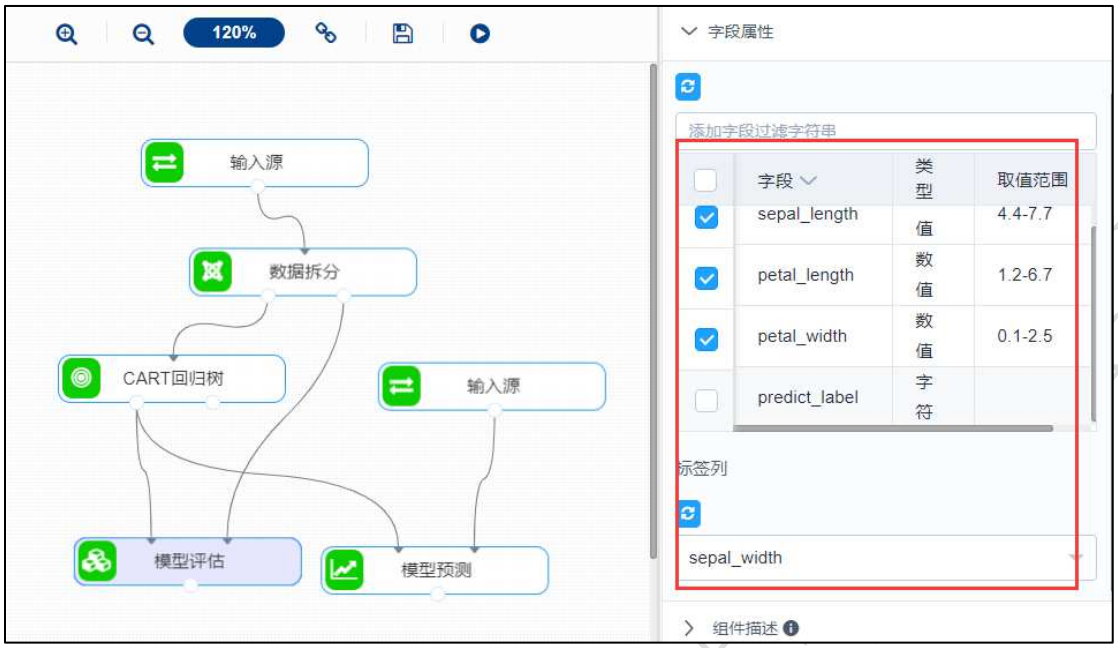

120%
输入源
数据拆分
CART回归树
输入源
模型评估
模型预测
字段属性
添加字段过滤字符串
字段
类型
取值范围
sepal_length
数值
4.4-7.7
petal_length
数值
1.2-6.7
petal_width
数值
0.1-2.5
predict_label
字符
标签列
sepal_width
组件描述

图 398

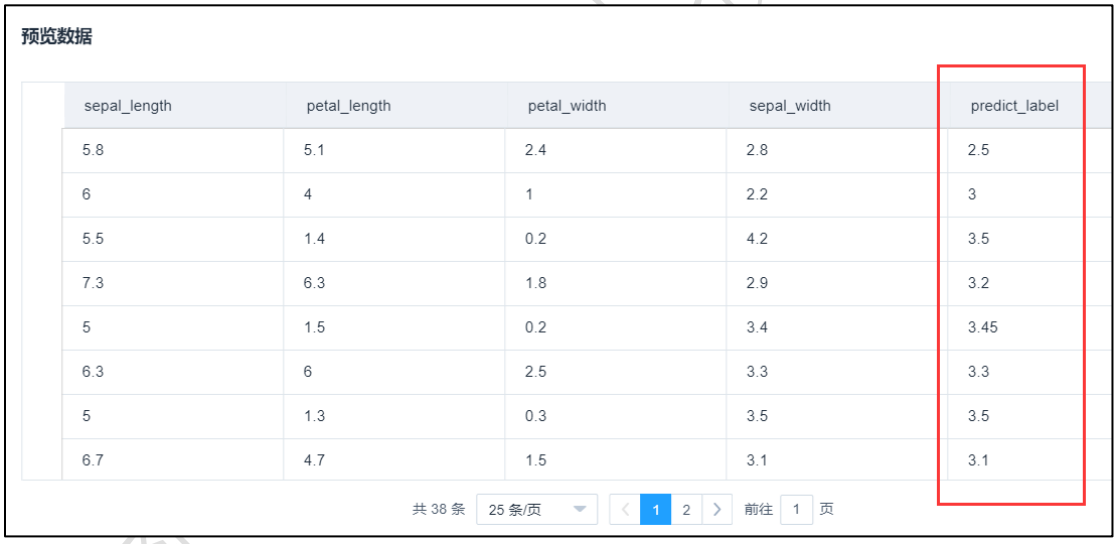

预览数据
sepal_length
petal_length
petal_width
sepal_width
predict_label
5.8 5.1 2.4 2.8 2.5
6 4 1 2.2 3
5.5 1.4 0.2 4.2 3.5
7.3 6.3 1.8 2.9 3.2
5 1.5 0.2 3.4 3.45
6.3 6 2.5 3.3 3.3
5 1.3 0.3 3.5 3.5
6.7 4.7 1.5 3.1 3.1
共 38 条
25 条/页
1
2
前往
1
页

图 399

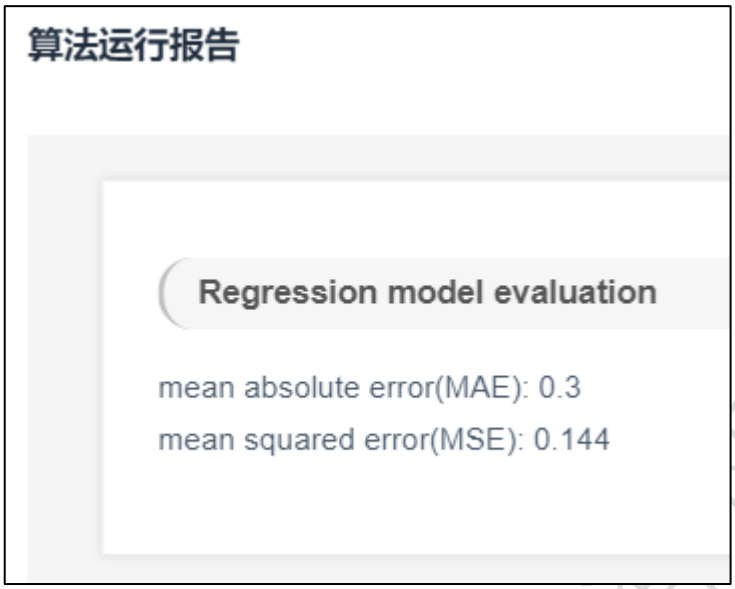
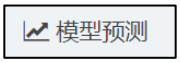


图 400

3.4.10 模型预测

3.4.10.1 模型预测

图标： 

描述： 对分类算法模型进行预测。

字段属性

特征列：勾选进行预测的字段。预测的类型和字段要和模型的字段信息一致

示例

下列对某数据进行 CART 回归树算法：

- CART 回归树模型预测配置如图 401 所示。
- CART 回归树模型预测结果如图 402 所示。

图 401

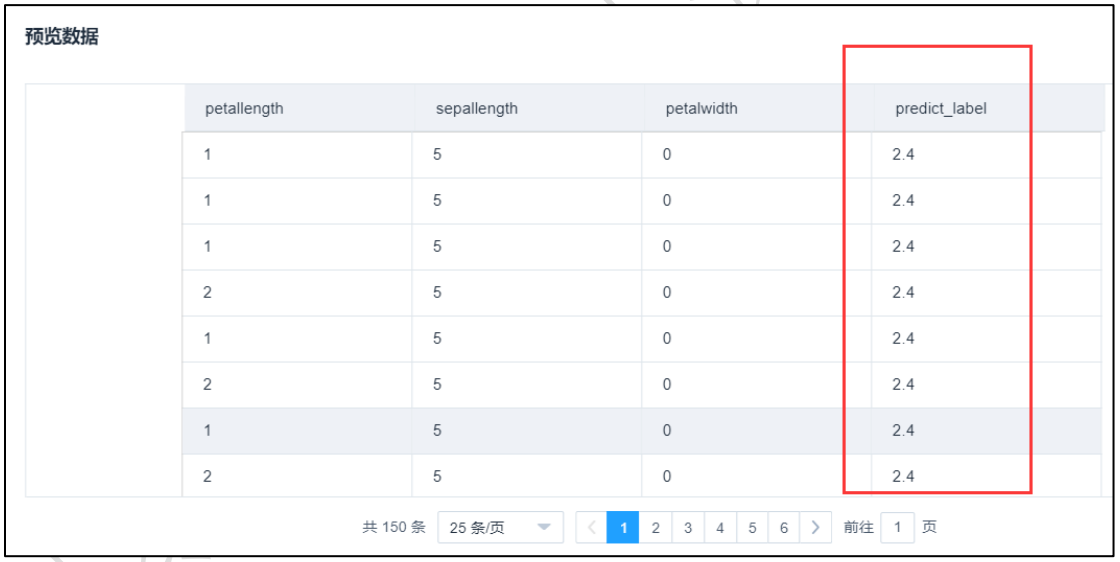


图 402

3.5 个人组件

3.5.1 添加参数说明

3.5.1.1 参数名称

参数名称是与脚本内接收参数的名称对应，如图 403 与图 404 所示。

图 403

```
'''
选择目标数据
'''
data_in = db_utils.query(conn, 'select '+ params['columns'] + ' from ' + inputs['data_in'])
```

图 404

3.5.1.2 表单控件

表单控件是将脚本内各参数以不同的方式展现出来。可通过下拉列表进行选择，如图 405 所示。

图 405

1. 普通输入框：

传入脚本为字符串，例如数据库 character 类型。常用来接收某个算法的参数。效果如图 406 所示。

图 406

2. 数值输入框

传入脚本为整数型数据，但为了保证准确型，建议在脚本内对参数进行类型转换。效果如图 407 所示。

图 407

3. 文本域

传入脚本为长字符串，如数据库的 text 类型。效果如图 408 所示。

图 408

4. 表结构列表

用于勾选分析的所需要的数据，传入后台的值类似“member_no,ffp_date”等列名，列名之间通过英文逗号分隔。下个组件可以获取到上个组件勾选的列名，可以实现在未运行之前去配置整个流程。效果如图 409 所示。

图 409

另外需要配置额外参数，参数的键固定为 key，值在脚本内对应 inputs$后的表名，表示需要从 input 表中勾选字段。如图 410 所示。

图 410

5. 下拉框

实现多选一。效果如图 411 所示。

图 411

另外需要配置选项，选项之间以英文状态下的分号隔开。用户勾选“是”，则将’TRUE’传入脚本，传入的脚本都为字符类型。效果如图 412 所示。

图 412

6. 单选框

实现多选一。效果如图 413 所示。

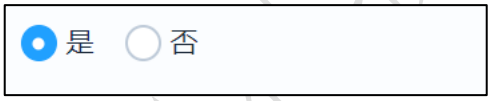


图 413

另外需要配置选项-，选项之间以英文状态下的分号隔开。用户勾选“是”，则将’TRUE’传入脚本，传入的脚本都为字符类型。如图 414 所示。

图 414

7. 文本编辑器-pgsql

编写 sql 脚本，效果如图 415 所示。

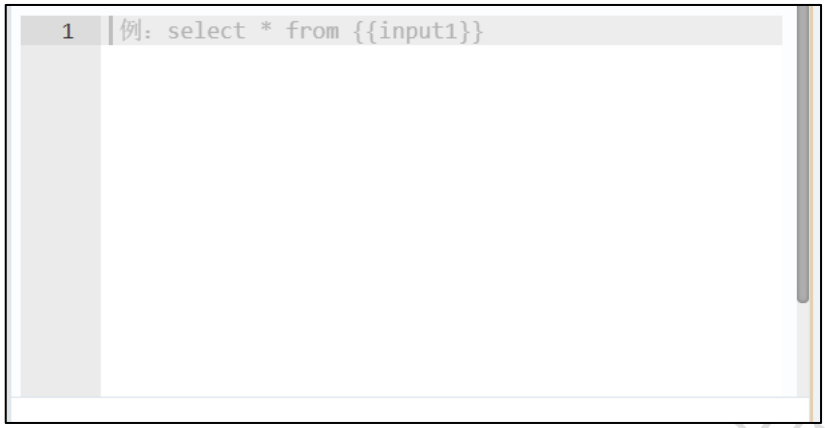


图 415

8. 文本编辑器-Python

编写 Python 脚本。效果如图 416 所示。

图 416

9. 条件列表

仅用于预处理菜单下【记录选择】组件。效果如图 417 所示。

图 417

10. 表结构列表（无复选框）

不会将任何值传入后台脚本，仅用于显示字段数据类型以及范围。效果如图 418 所示。

图 418

另外需要配置额外参数，参数的键固定为 key，值在脚本内对应 inputs$后的表名，表示需要从 input 表中勾选字段。如图 419 所示。

图 419

11. 目标列选择器

可以实现选择某个表的某一列，传入后台脚本为该列的列名，为字符串类型。效果如图 420 所示。

图 420

另外需要配置额外参数，参数的键固定为 key，值在脚本内对应 inputs$后的表名，表示需要从 input 表中勾选字段。如图 421 所示。

图 421

12. 输入列表

用于 R 脚本和 SQL 探索两个组件，用于获取表名。如图 422 所示。

图 422

13. 多选框

实现多选多。效果如图 423 所示。

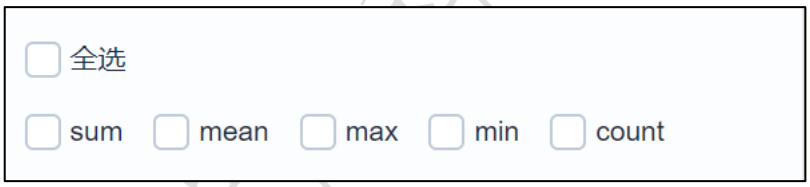


图 423

另外需要配置选项，选项之间以英文状态下的分号隔开。用户勾选“是”，则将’TRUE’传入脚本，传入的脚本都为字符类型。如图 424 所示。

图 424

14. database 数据源搜索下拉框

用于输入源组件，获取某个表数据。效果如图 425 所示。

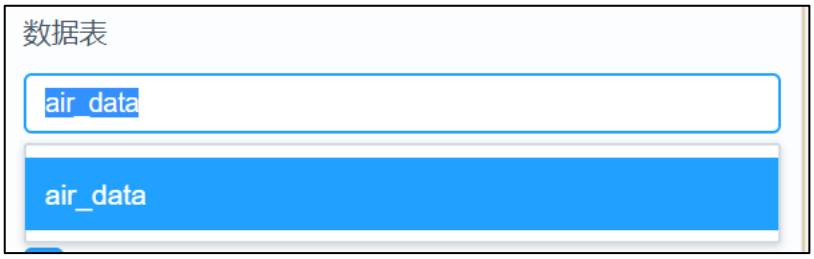


图 425

15. 添加字段文本输入框

定义新增列列名，传入后台脚本为字符串，效果如图 426 所示。

图 426

例如用户在该组件运行时，需要使用 A，B，C 三列数据，输出时需要生成 A,B,C,D 四列，下个组件如果想要在未运行时就能勾选这四列数据，则需要配置这个参数。

16. 修改输出列名称

仅用于修改列名组件。如图 427 所示。

图 427

17. 密码框

仅用于输出源组件，填写数据库密码。效果如图 428 所示。

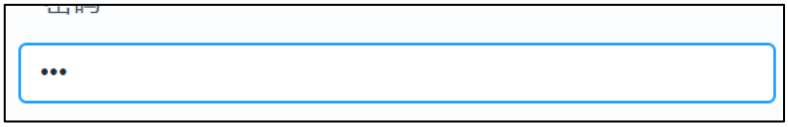


图 428

18. 字段下拉多选框

用于勾选分析的数据，传入后台的值类似“member_no,ffp_date”下个组件如果想获取该组件运行的结果，必须先运行该节点。效果如图 429 所示。

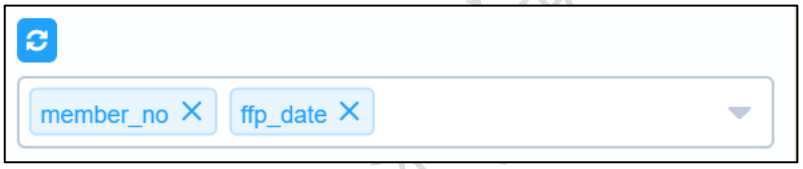


图 429

如果需要从某个数据表勾选字段，需要配置额外参数，参数的键固定为 key，值在脚本对应表名。如图 430 所示。

图 430

19. 修改输出字段类型

仅用于修改类型组件。效果如图 431 所示。

图 431

3.5.1.3 显示名称

显示名称是对应脚本内的特定参数而言，在前端界面显示的名称。例如所示。脚本内用 min 表示最小值，则用户可以在前端界面看到“最小值”这个名称。如图 432 所示。

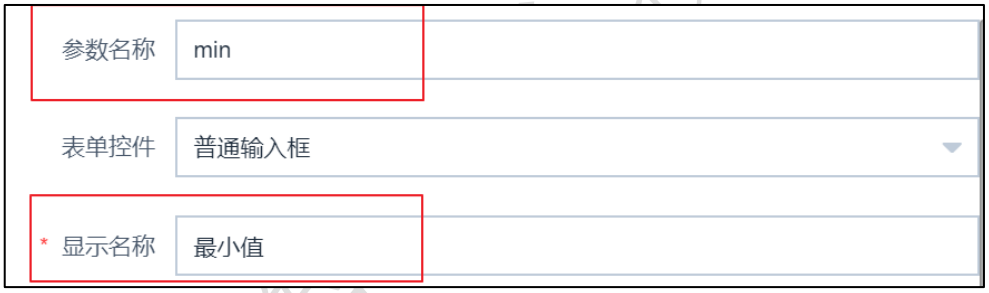


图 432

3.5.1.4 默认值

默认值主要针对表单控件中的普通输入框、数值输入框、文本域、下拉框、单选框、修改输出列名称而言，默认传递该值到脚本中。

3.5.1.5 显示组件

该功能目的是为了控制是否将该参数在前端界面显示，通常与表单控件中的添加字段文本输入框配合。

3.5.1.6 校验输入值

主要对表单控件内需要填写获勾选的内容进行校验，启用“校验输入值”之后，可以选

择对应的校验表达式。如图 433 所示。

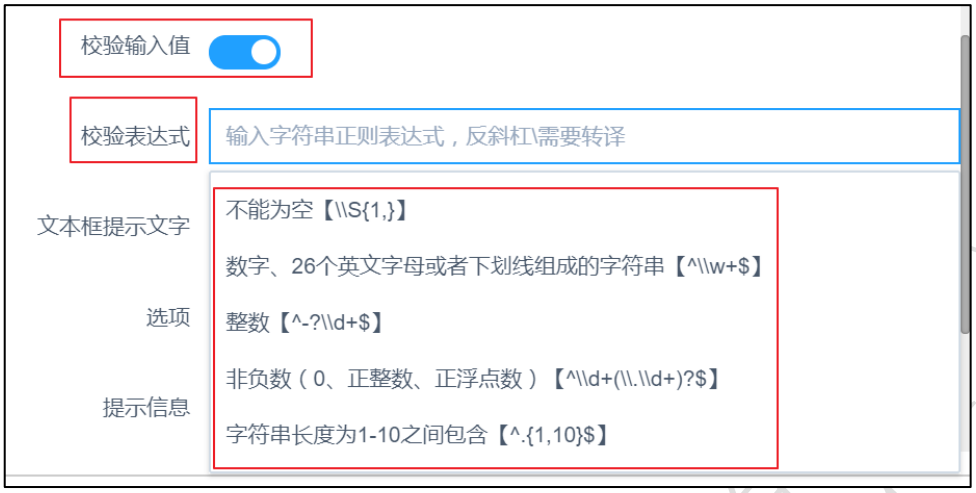


图 433

3.5.1.7 校验表达式

校验表达式主要与校验输入值配合。

3.5.1.8 文本框提示文字

当表单控件选择普通输入框、文本域、文本编辑器-pgsql、文本编辑器-Python 时，可以填写需要提示的文字。

3.5.1.9 选项

当表单控件选择下拉框、单选框、多选框时，选项内容可在此设置，一般格式为：选项一:a;选项二:b，前端界面选择选项“一”时，则将“a”传入脚本。需要注意各选项之间以英文状态下的分号隔开，选项名和对应参数之间以英文状态下的冒号隔开。

3.5.1.10 提示信息

主要对前端界面参数进行描述，如图 434 所示。

图 434

3.5.1.11 顺序

主要对前端界面的参数上下顺序进行调整。

3.5.1.12 额外拓展参数

当表单控件选择表结构列表、表结构列表（无复选框）、目标列选择器、字段下拉多选框时，需要填写额外拓展参数，指定需要从那个输入获取字段列名。如图 435 所示。

图 435

3.5.2 Python 个人组件配置

3.5.2.1 Python 个人算法规则描述

表 1

参数名称	类型	描述
conn	固定值	封装了python与平台数据库连接。
inputs	dict	输入数据集合，存储组件输入节点对应的数据，通过输入节点的key获取数据，例如配置的key为“data_in”，那么脚本对应inputs[‘data_in’],即为该节点对应的数据表。
params	dict	获取的规则与inputs一致。需要注意的是：params中参数的值都是字符类型的，根据参数需要的数据类型，必须要在代码中进行数据,类型转换，比如：聚类数为int(params[‘centers’])。
outputs	dict	输出数据集合，只支持输出DataFrame，存储规则参见inputs。
reportFileName	固定值	算法运行报告文件的存储路径。
model	固定值	分类或回归的模型，在代码中需要使用return返回。

3.5.2.2 示例:standardization 代码封装

输入（inputs）

data_i

参数（params）

columns（特征列）

method（处理方式）:

'o_scale'（零均值标准化）

min_max（极差标准化）：min（zui'xiao）、col（颜色）

输出（outputs）

data_out

```
'''
载入模块
'''
from sklearn import preprocessing
import pandas as pd
import db_utils

'''
选择目标数据
'''
data_in = db_utils.query(conn, 'select '+ params['columns'] + ' from ' + inputs['data_in'])


'''
标准化
'''
data_in = data_in.select_dtypes(include=['number'])  # 筛选数值型数据
data_out = data_in
if (params['method'] == '0_scale'):
    data_out = preprocessing.scale(data_in)
else:
    data_out = preprocessing.minmax_scale(data_in, feature_range = (int(params['min']), int(params['max'])))

data_out = pd.DataFrame(data_out, columns = data_in.columns)


'''
将结果写出
'''
db_utils.dbWriteTable(conn, outputs['data_out'], data_out)
```

图 436

平台效果

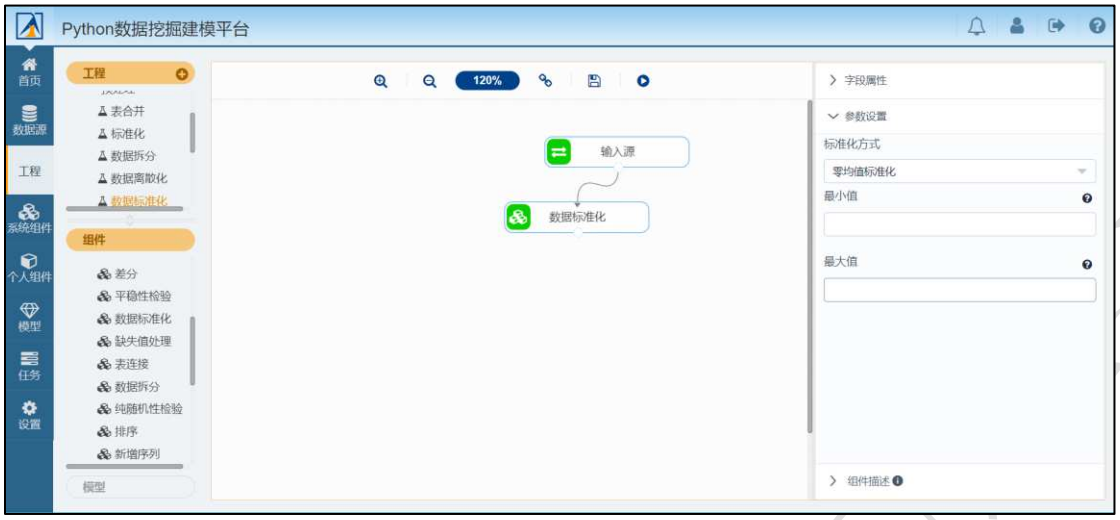


图 437

3.5.2.3 配置步骤

a) 新建个人组件，定义组件名称，设置组件功能

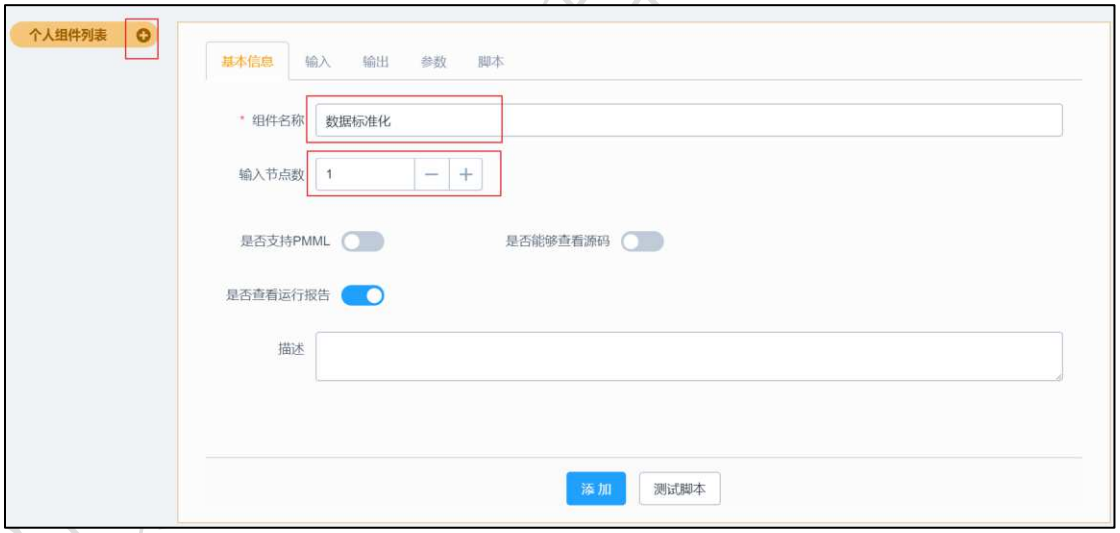


图 438

b) 新建输入参数

图 439

图 440

c) 新建输出参数

图 441

图 442

d) 新建参数—字段属性

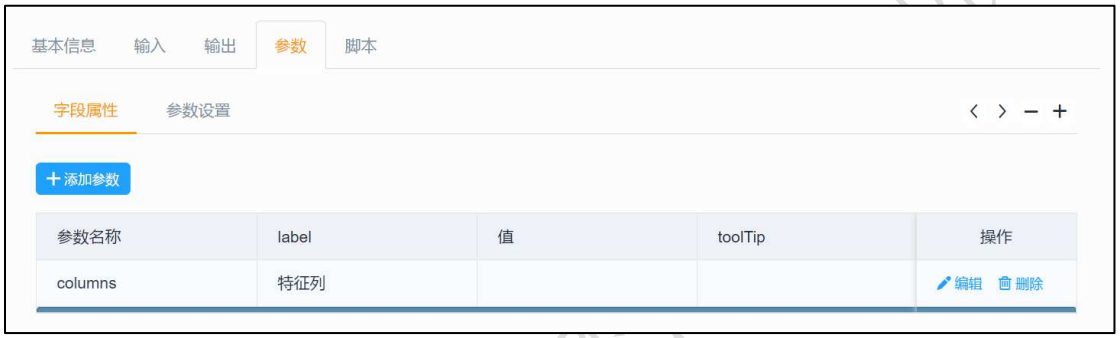


图 443

e) 新建参数—字段信息--columns

图 444

图 445

f) 新建参数—参数设置

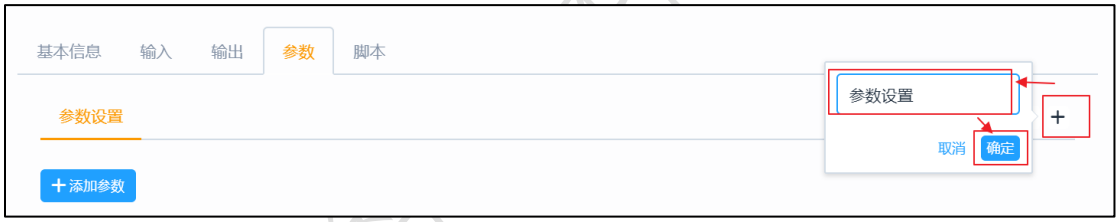


图 446

图 447

g) 新建参数—参数设置—method

图 448

图 449

h) 新建参数—参数设置—min

图 450

图 451

i） 新建参数—参数设置—max

图 452

图 453

j） 粘贴脚本，注意脚本需要每行都缩进四个空格。

图 454

3.6 模型

本平台的模型展示方式是将 数据挖掘算法生成的模型，对应于平台的“系统组件”中分类与回归相关的算法，此种模型存在“模型管理”中。如下图 455 所示。

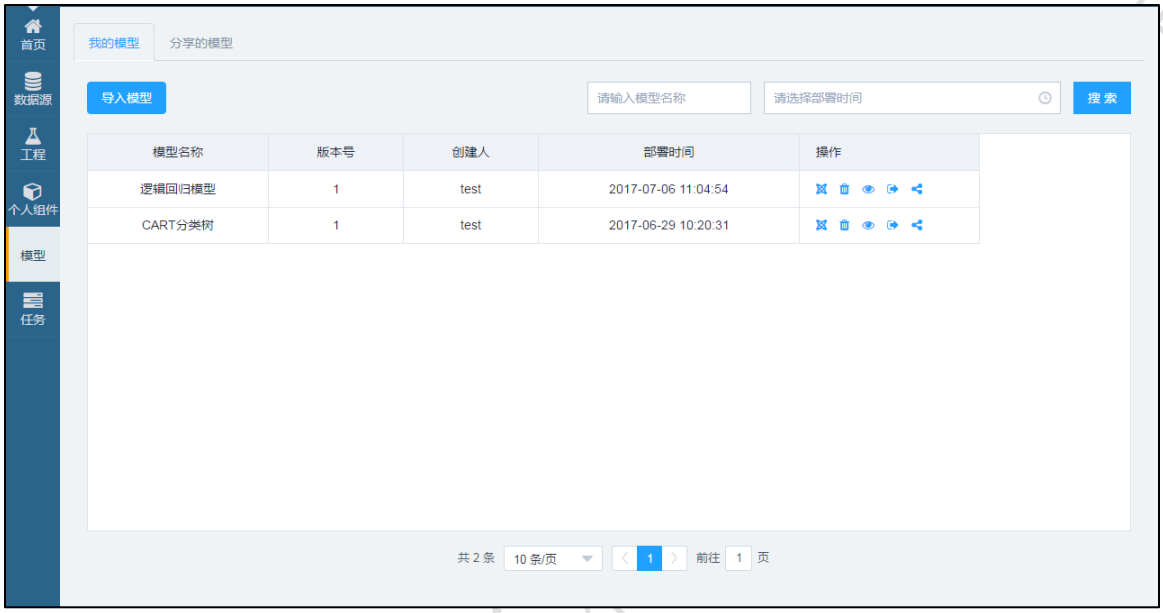


图 455

3.6.1 模型生成

“系统组件”栏中分类和回归算法运行后，自动会生成相应的模型，并存在我的模型列表中；需用户手动操作部署，才能保存到模型管理列表中

模型名的规范为：组件名，可通过重命名组件修改模型名

平台中每个算法节点运行后仅对应唯一一个模型，再次运行节点时，会将原有生产的覆盖，如图 456、图 457 所示。

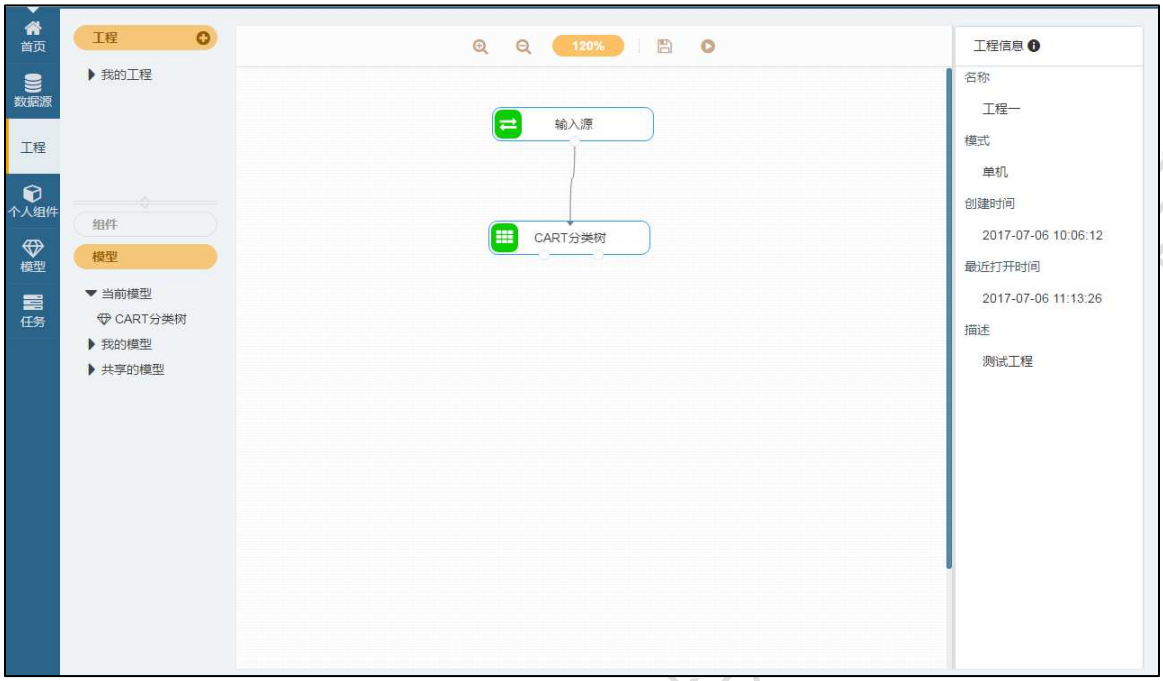


图 456

图 457

3.6.2 模型使用

在模型栏中，直接将模型拖拽到画布中就可以使用，如下图 458 所示。

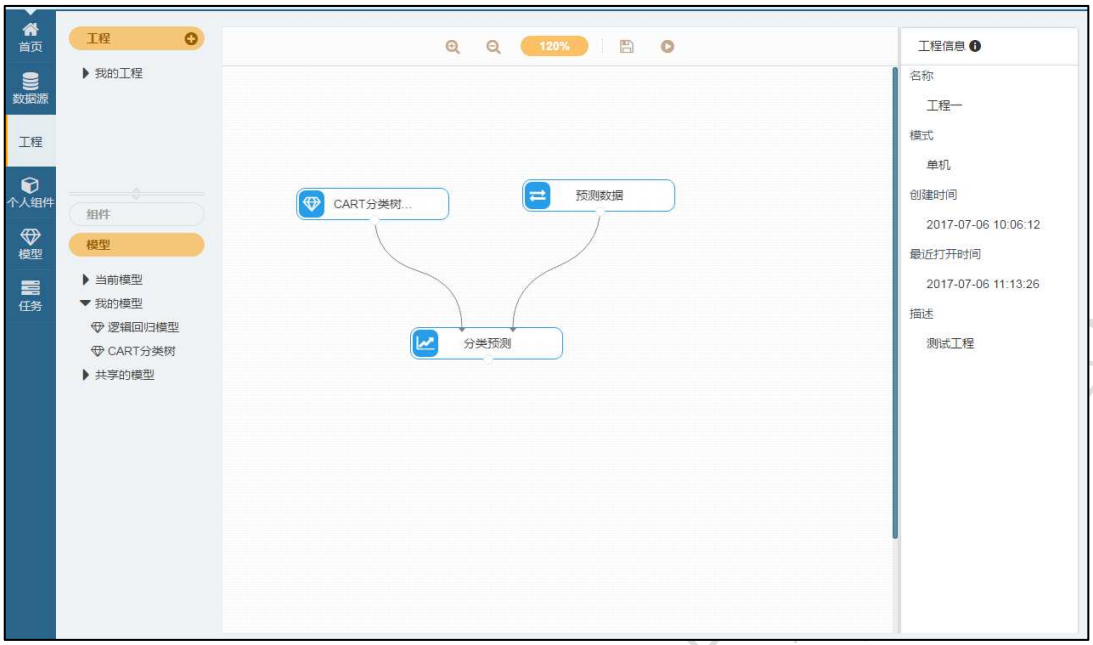


图 458

3.6.3 模型操作

在模型列表中，操作框中，可以进行的操作如下图 459 所示。

图 459

从左至右，依次是：

模型预测：支持单列数据预测，csv 文件预测、数据库预测

卸载模型：将模型从数据表中删除

查看模型：可查看模型的输入列、目标列、输出列的字段信息

模型导出：仅支持 pmml 文件模型导出

模型分享：可以将模型分享给其他用户

3.7 任务

可以把工程的执行或者数据源的导入设置成定时执行。

3.7.1 新建任务

点击右上角的新建任务进行添加，弹出填写框，如图 460、图 461 所示。

图 460

图 461

3.7.2 填写信息

填写任务名称，任务组，选择需要进行定时任务的工程或者数据源，执行的时间为 cron 表达式，cron 表达式的填写点击执行时间输入框查看，如图 462、图 463 所示。

图 462

图 463

Cron 表达式说明：

“*”字符代表所有可能的值，因此“*”在子表达式（月）里表示每个月的含义，“*”在子表达式（天（星期））表示星期的每一天

“/”字符用来指定数值的增量。例如：在子表达式（分钟）里的“0/15”表示从第 0 分钟开始，每 15 分钟 ；

在子表达式（分钟）里的“3/20”表示从第 3 分钟开始，每 20 分钟（它和“3，23，43”）的含义一样;

“？”字符仅被用于天（月）和天（星期）两个子表达式，表示不指定值。当两个子表达式其中之一被指定了值以后，为了避免冲突，需要将另一个子表达式的值设为“？”

“L” 字符仅被用于天（月）和天（星期）两个子表达式，它是单词“last”的缩写,但是它在两个子表达式里的含义是不同的。

在天（月）子表达式中，“L”表示一个月的最后一天 ,在天（星期）子表达式中，“L”表示一个星期的最后一天，也就是 SAT(星期六)

注意：在使用“L”参数时，不要指定列表或范围，因为这会导致问题

下面列出一些例子进行说明：

0 15 10 * * ?* 每天 10 点 15 分触发

0 0 12 ?* WED 表示每个星期三中午 12 点

0 * 14 * * ？每天下午的 2 点到 2 点 59 分每分触发

0 15 10 ?* 6L 每月最后一周的星期五的 10 点 15 分触发

0 15 10 ?* 6#3 每月的第三周的星期五开始触发。

3.7.3 执行任务

点击对应任务的执行按钮开始执行定时任务，如图 464 所示。

图 464

4 用户权限管理

用户权限管理仅限超级管理员使用，对用户使用权限进行管理

4.1 组织架构

4.1.1 新增组织架构

如图 465 所示，点击【新增】按钮，则会弹出如图 466 所示的对话框。

组织架构名称：必填项。

描述：必填项。

上级组织架构：选填项。

排序优先级：必填项。

状态：组织架构是否可用。

图 465

图 466

4.1.2 删除组织架构

在组织架构表的第一列进行勾选，再点击【删除】，即可将勾选的组织架构进行删除。如图 467 所示。

图 467

4.1.3 编辑组织架构

点击【编辑按钮】，可对已有的组织架构进行编辑，可编辑的内容与新建组件架构相同，如图 468 所示。

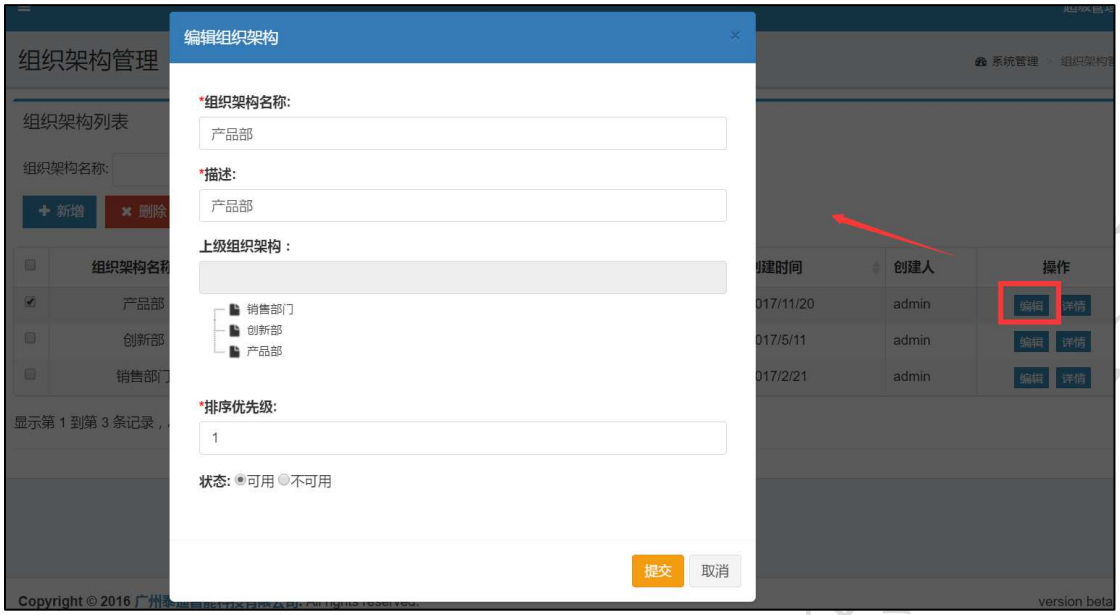


图 468

4.1.4 查看组织架构信息

点击【详情】可查看组织架构的详细信息，包括：组织架构名称、描述、上级组织架构、排序优先级、状态、创建者、创建时间，如图 469 所示。

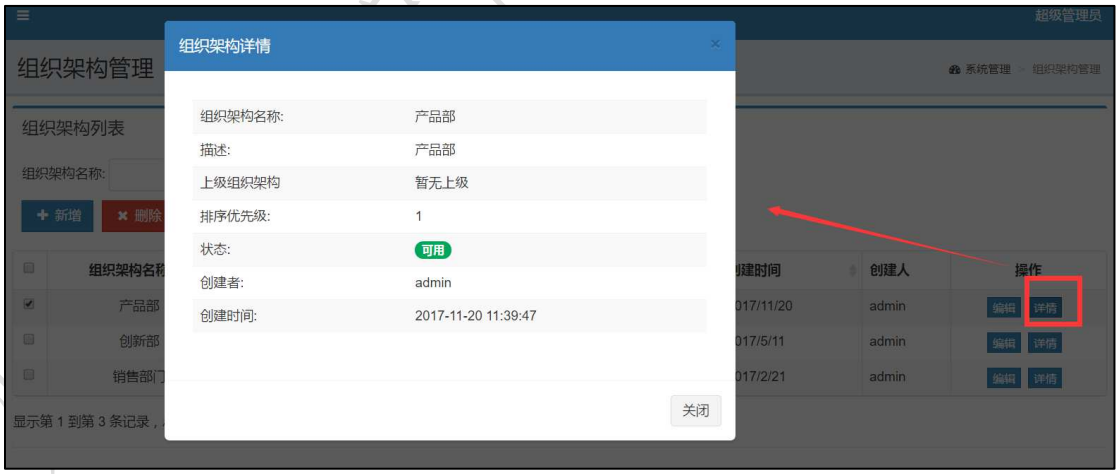


图 469

4.2 角色管理

4.2.1 新增角色

点击【新增】可新增角色，如图 470 所示。新增角色需要填写的信息如图 471 所示。

角色名：必填项，建议使用英文字母。

描述：必填项，对角色进行描述。

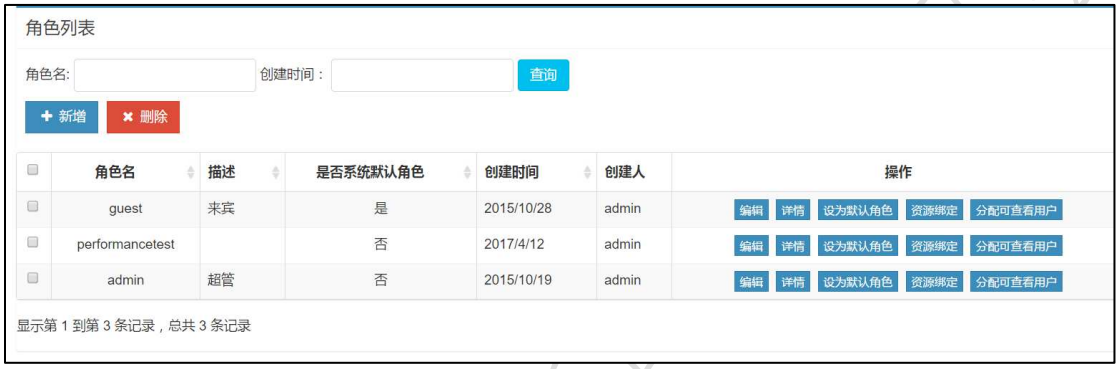


图 470

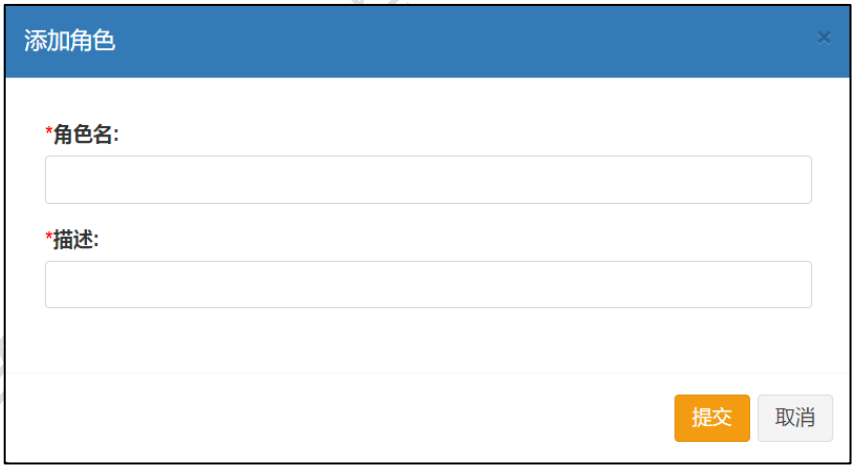


图 471

4.2.2 删除角色

通过勾选角色，点击【删除】，可以删除已经存在的角色。如图 472 所示。

图 472

4.2.3 编辑角色

点击【编辑】按钮可以对角色的描述及状态进行编辑，如图 473 所示。

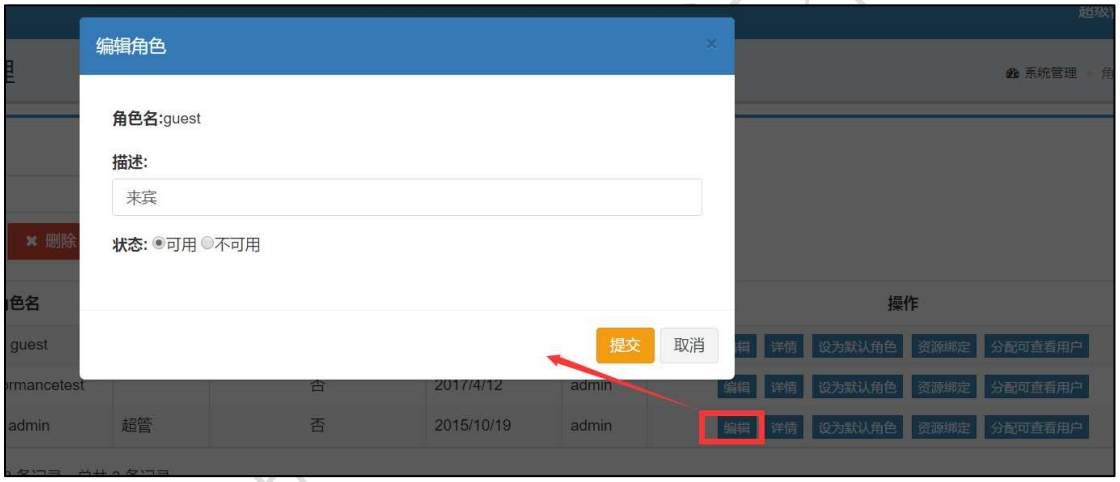


图 473

4.2.4 资源绑定

可给不同的角色绑定不同的资源。如图 476 所示。

图 474

4.2.5 查看角色详情

可查看角色的详情，包括：用户名、描述、角色资源、状态、创建者、创建时间，如图 475 所示。

图 475

4.2.6 设为默认角色

如下图 476 所示。

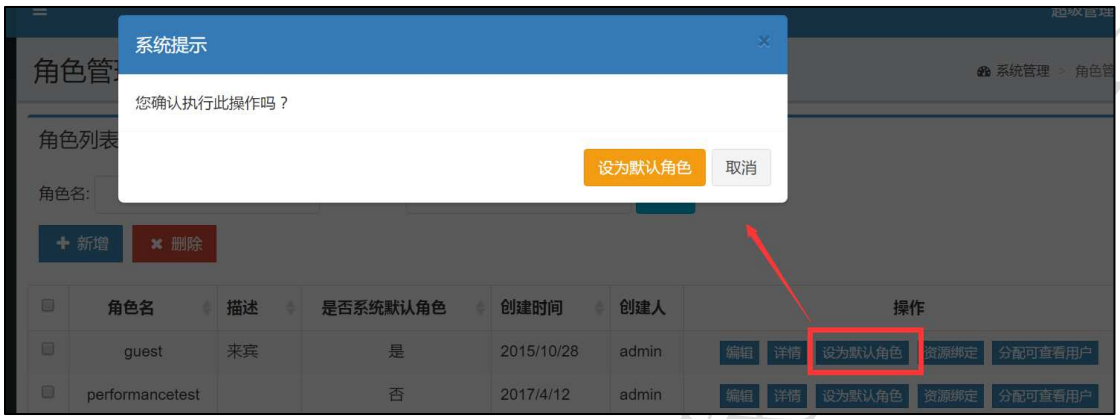


图 476

4.1 用户管理

4.1.1 新增用户

如图 477 所示，点击【新增】按钮可新增用户，则会弹出如图 478 所示的对话框，填写相应信息即可新增用户。

用户名：必填项。

邮箱：必填项。

密码：必填项。

别名：必填项。

组织架构：可选项。

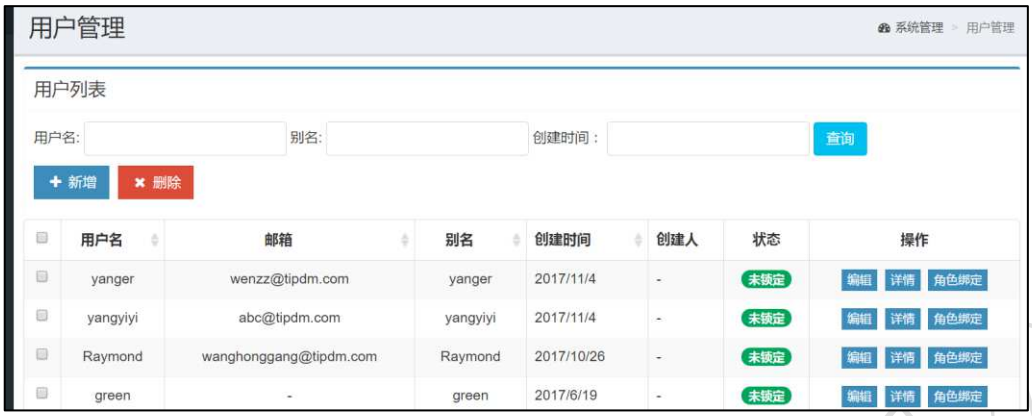


图 477

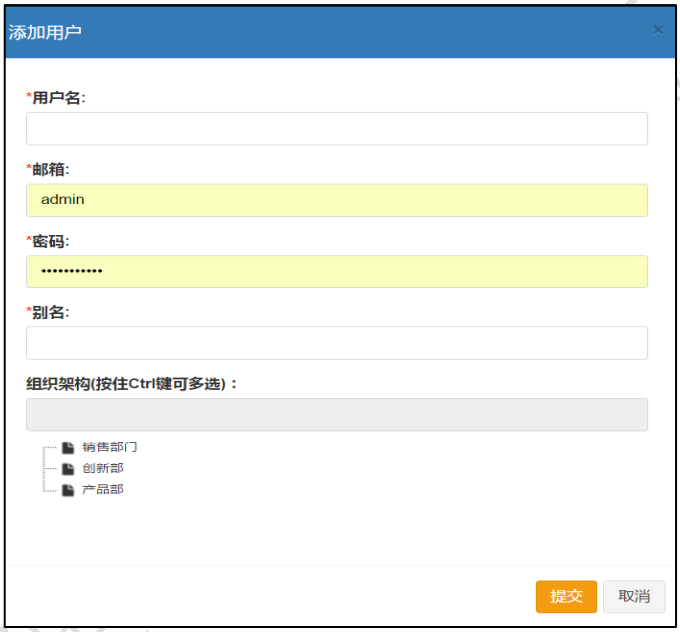


图 478

4.1.2 编辑用户

如图 479 所示。

图 479

4.1.3 查询用户

如图 480 所示。

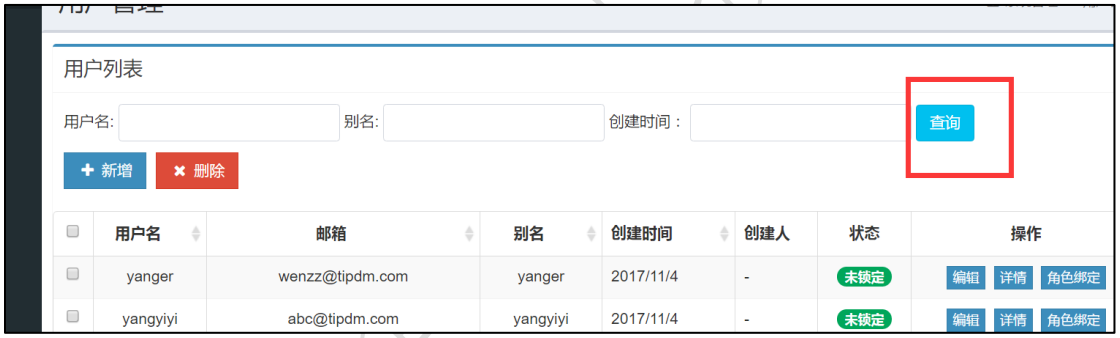


图 480

4.1.4 角色绑定

如图 481 所示。

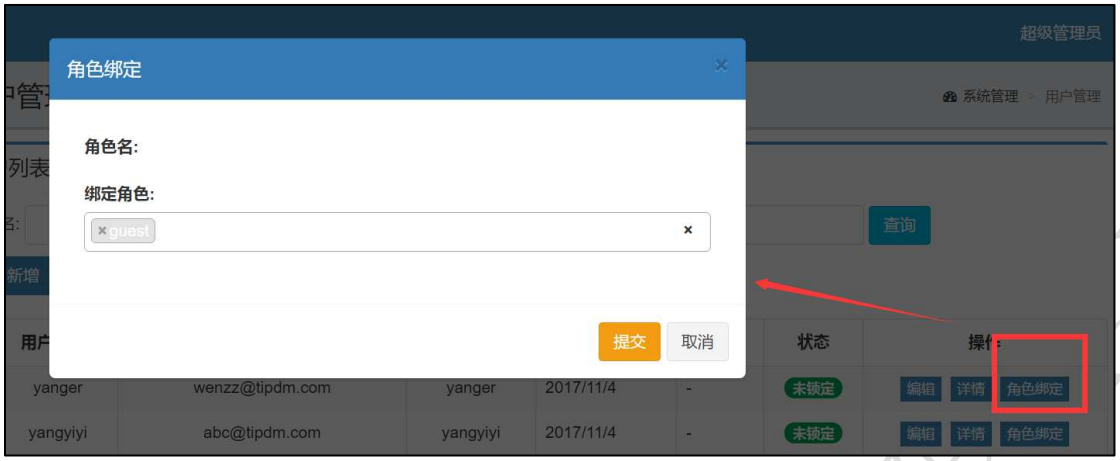


图 481

4.2 资源管理

4.2.1 资源绑定

资源绑定包括基础权限资源及数据分析平台资源。

基础权限资源：主要针对用户管理系统部分的资源进行绑定。

数据分析平台权限资源：主要针对挖掘分析平台部分资源进行绑定，具体见 4.2.1。

如图 482、图 483 所示。

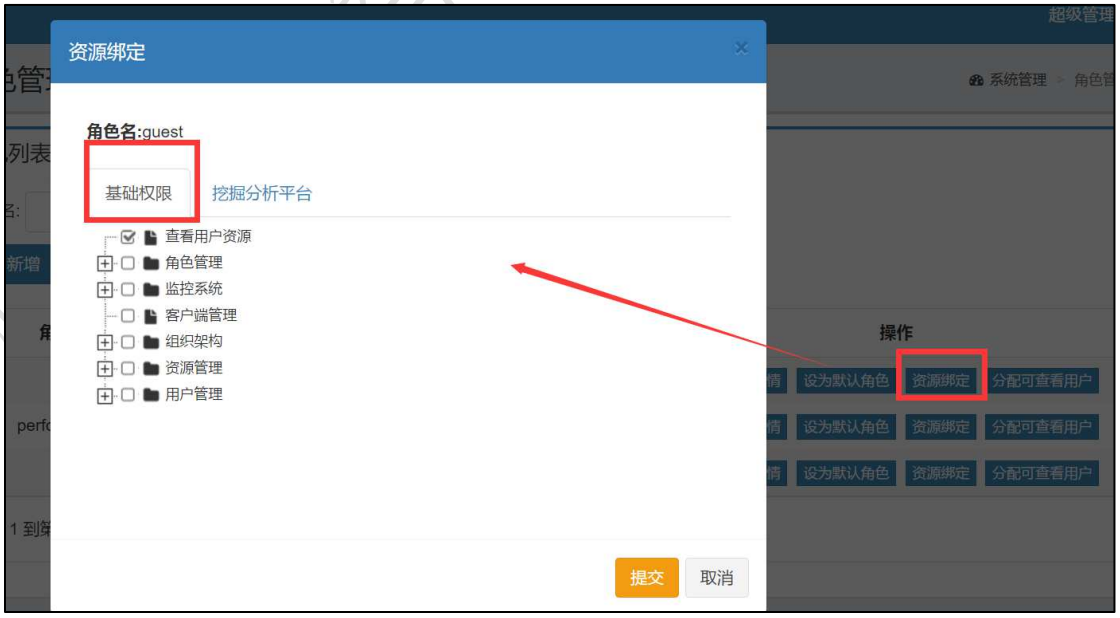


图 482

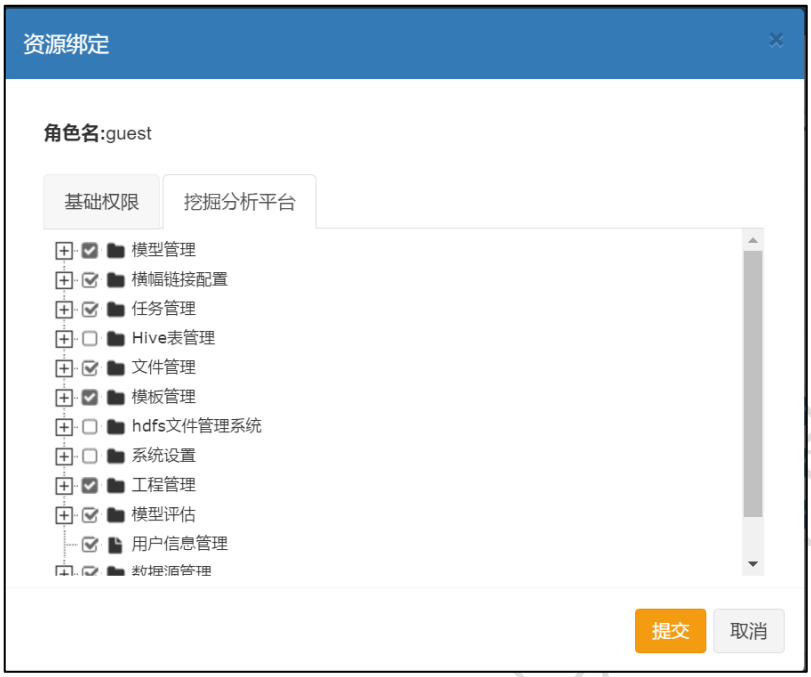


图 483

4.2.1 资源管理

资源管理包括：基础权限资源、挖掘分析平台资源。

基础资源具体包括：

序号	一级菜单	二级菜单
1	角色管理	绑定角色
2		创建角色
3		删除角色
4		编辑角色
5		查询角色
6		设为系统默认角色
7	监控系统	监控系统

8	客户端管理	客户端管理
9	组织架构	创建组织架构
10		编辑组织架构
11		删除组织架构
12		查询组织架构
13	资源管理	创建资源域
14		创建资源
15		编辑资源
16		查询资源
17		删除资源
18	用户管理	绑定角色
19		创建用户
20		删除用户
21		编辑用户
22		查询用户
23	查看用户资源	查看用户资源

挖掘建模平台资源具体包括：

序号	一级菜单	二级菜单	序号	一级菜单	二级菜单
1	模型管理	检验模型	53	工程管理	生成 URL
2		分享模型	54		删除历史版本

3		删除分类/模型	55		另存为模板
4		获取个人模型列表	56		恢复到指定版本
5		分享的模型列表	57		创建历史版本
6		新增模型分类	58		停止流程调度
7		保存模型	59		执行该节点
8		导入模型	60		从此节点开始执行
9		模型导出	61		执行到此处
10		卸载模型	62		查看运行结果
11		获取个人的模型列表	63		查看运行日志
12		获取分享模型列表	64		输出节点数据可视化
13	横幅链接配置	社区	65		查看算法源码
14		帮助文档	66		模板克隆
15		过滤任务数据	67		获取子级节点
16		开始定时任务	68		获取个人模型评估列表
17		添加任务数据	69		新生成模型-重新部署
18	任务管理	更新任务数据	70		部署模型
19		删除任务数据	71	模型评估	新生成模型-部署模型
20		获取任务数据	72		获取个人模型列表
21		介绍定时任务	73		模型预测（CSV）
22	Hive 表管理	获取表的结构信息	74		模型预测（RDBMS）

23		从 HDFS 路径新建 HIVE 表	75		模型预测
24		新建 HIVE 表	76	用户信息管理	用户信息管理
25		过滤表	77	数据源管理	获取表的结构信息
26		测试数据库连接	78		检查表是否存在
27		新建数据库连接信息	79		删除表
28		同步表数据	80		创建表
29	文件管理	检查文件是否存在	81		数据预览
30		上传数据文件	82		测试数据库连接
31		上传算法包文件	83		预览数据
32	模板管理	模板列表	84		预览分享数据
33		模板删除	85		新建数据库连接
34	HDFS 文件管理系统	查看内容的 URL	86		同步表数据
35		判断 URL	87		过滤表
36		删除 URL	88		搜索分享列表
37		获取创建的 URL	89		搜索数据
38		匹配相应字符的路径	90		分享数据源
39	系统设置	修改文件内容	91	组件管理	上移
40		获取内容	92		获取算法集合
41		获取配置文件	93		删除组件参数
42	工程管理	执行流程	94		修改类别名

43		删除工程	95		删除组件分类
44		获取工程	96		获取自分类的算法
45		创建工程	97		删除组件
46		新增工程分类	98		更新组件数据
47		同步工程	99		获取组件数据
48		另存为工程	100		添加组件
49		工程排序上移	101		添加组件分类
50		查看数据	102		添加基础组件类型
51		查看工程的历史版本	103		获取基础组件集合
52		删除工程分类	104		系统组件

4.3 客户端管理

如下图 484 所示。

图 484

4.4 监控系统

如下图 485 所示。

图 485