泰迪智能科技
服务热线:40068-40020

TipDM大数据挖掘建模平台

操　作　手　册

地址：广州市黄埔区科学城凝彩路26号锦昊智谷2005房
网址：http://www.tipdm.com
邮箱：services@tipdm.com
热线：40068-40020
邮编：510000
电话：020-22205718

1 平台介绍

1.1 产品简介

关于TipDM大数据挖掘建模平台

TipDM大数据挖掘建模平台是由广东泰迪智能科技股份有限公司自主研发，研发之初主要是为了满足企业大数据挖掘需要，使用平台配置的开箱即用的算法，帮助企业快速构建大数据分析与挖掘流程应用，研发至今有南方电网、中国电力科学研究院、珠江数码、北京智慧信访、中国石油勘探研究院、轻工业环境保护研究所、交通运输部公路科学研究所等众多企业使用。随着平台上的算法的不断积累与完善，超百所高校使用TipDM大数据挖掘建模平台作为工程平台。近些年TipDM大数挖掘建模平台也作为“泰迪杯”竞赛的线上竞赛辅助平台，立足大数据/人工智能产业实践，为高校教育赋能。

为什么选择TipDM大数据挖掘建模平台?

1、多功能自定义算法配置：框架使用JAVA语言开发，使用R语言、Python、Spark计算引擎，支持使用R，Python，Scala，SparkR，PySpark进行算法自定义开发。

2、简单易用：封装上百种算法，包含机器学习算法、文本挖掘，深度学习等，通过可视化拖拽实现模型训练，为用户提供数据接入，数据处理，模型训练，模型部署，模型预测一站式服务。

3、分布式架构：基于Spring Cloud构建，提供稳定可靠的服务调用、服务治理、服务降级能力。

4、多数据源聚合：支持主流的关系型数据库，支持文本、图像、音视频等非结构化文件。快速实现异构数据源之间的数据同步问题。

5、可靠的任务调度：提供稳定可靠的分布式定时调度系统，调度中心支持HA部署，任务分布式执行。提供流程化的任务编排方式，直观的体现任务之间的依赖关系。

1.2 名词解释

算法：将建模过程涉及到的输入、数据探索、数据预处理、建模等算法分别进行封装，每一个封装好的算法都可称为一个算法。算法分为系统算法和个人算法。系统算法是由平台管理员编辑，提供给所有用户使用的算法。个人算法是由平台个人用户编辑，只提供给本账号使用的算法。

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

参数：每个算法都有提供给用户进行设置的内容，这部分内容称为参数。参数可分为字段参数和算法参数。字段参数是设置该输入的数据字段。算法参数是设置该算法算法提供的部分可选参数。

工程库：对于创建好的工程，可通过工程库分享给其他用户。通过工程库，其他用户能够建立一个无需导入数据和设置参数，就能够快速运行的工程。

Python系统算法使用介绍

8.1 预处理

8.1.1 数据筛选

（1）作用

在数据分析工作中不乏会面对多且密的数据集，要想分析出海量数据中所蕴含的数据价值，则需要对有价值的数据进行筛选，可见数据筛选在整个数据处理流程中处于至关重要的地位。平台上该组件是对数据表数据进行过滤操作，根据自定义的条件筛选掉符合条件的数据记录，提高收集存储的相关数据的可用性，便利于后期数据分析的工作。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	展示经过数据筛选后表格的部分数据。
2	日志	输出数据筛选后数据维度变化。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行筛选并输出的数据列，数值型
2	筛选条件	逻辑运算符	每个过滤条件间的联系，可选项有“与”，“或”
3	筛选条件	表达式	筛选的条件，可选项有大于、大于等于、等于、小于、小于等于、不等于、包含、不包含
4	筛选条件	筛选条件的目标值	筛选的条件，输入需要筛选的值，数值型或字符型，也支持填入“int”、“float”等
5	筛选条件	筛选的特征列	数据中进行筛选的列

（5）示例

对于“iris”数据集进行数据筛选示例，iris的数据集有3类，150个数据样本，每类有50个样本。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行数据筛选，将【数据筛选】组件与输入源连接，在“字段设置”的“特征”中勾选所有字段。

添加筛选条件。点击筛选条件上的加号按钮，筛选条件数目可自定义添加，添加一个筛选条件，选择列为“outcome”，表达式选择“等于”，筛选条件的比较值输入0，继续添加筛选条件，选择列为“outcome”，表达式选择“等于”，筛选条件的比较值输入1，右键单击【数据筛选】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	过滤的列	outcome	对类别进行筛选
2	表达式	等于	筛选等于该类别的数据
3	筛选条件	0,1	筛选0、1类别数据
4	逻辑运算符	或	筛选0、1类别数据

打开日志，查看结果。在日志中可以查看筛选后数据维度的变化。对【数据筛选】组件右击，点击“查看日志”。

序号	名称	作用
1	筛选条件	筛选条件表达式
2	数据筛选前维度	观测数据筛选前维度
3	数据筛选后维度	观测数据筛选后维度及变化

查看日志

数据帅选表达式为

& data.iloc[:, 4] = 0 | data.iloc[:, 4] = 1

数据筛选前数据维度为：

(150, 5)

数据筛选后数据维度为：

(100, 5)

数据筛选后数据内容如下:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	outcome
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0

8.1.2 新增序列

（1）作用

新增序列组件的作用是在原有的数据表上增添新的列。新增列的值可以是指定数值、随机数值或自增数值。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	展示经过新增序列后表格的部分数据。
2	日志	输出数据新增序列后数据维度变化。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要输出的数据列，数值型
2	参数设置	新增序列方式	新增序列方式，可选特定值、随机值、自增序列。
3	特定值设置	特定值类型	新增的特定值数值类型，字符型。
4	特定值设置	特定值	新增的特定值，数值型
5	随机数范围	最小值、最大值	新增的随机值范围，数值型
6	随机数范围	结果保留小数位	新增float类型随机值小数点保留位数

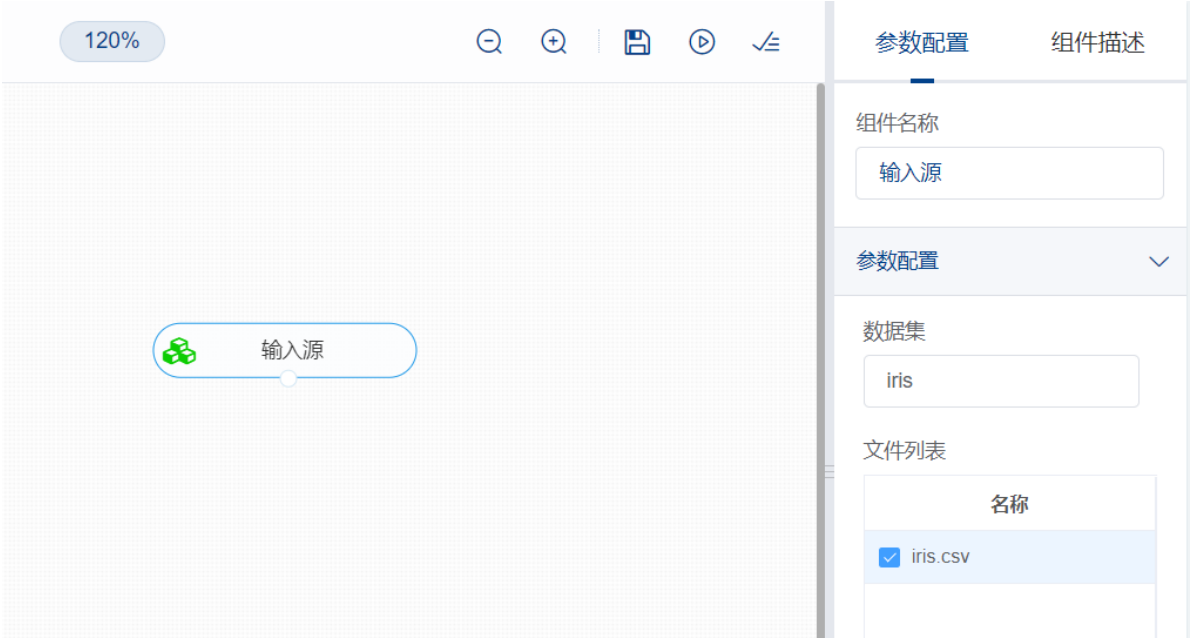
（5）示例

对于“iris”数据集进行新增序列示例。

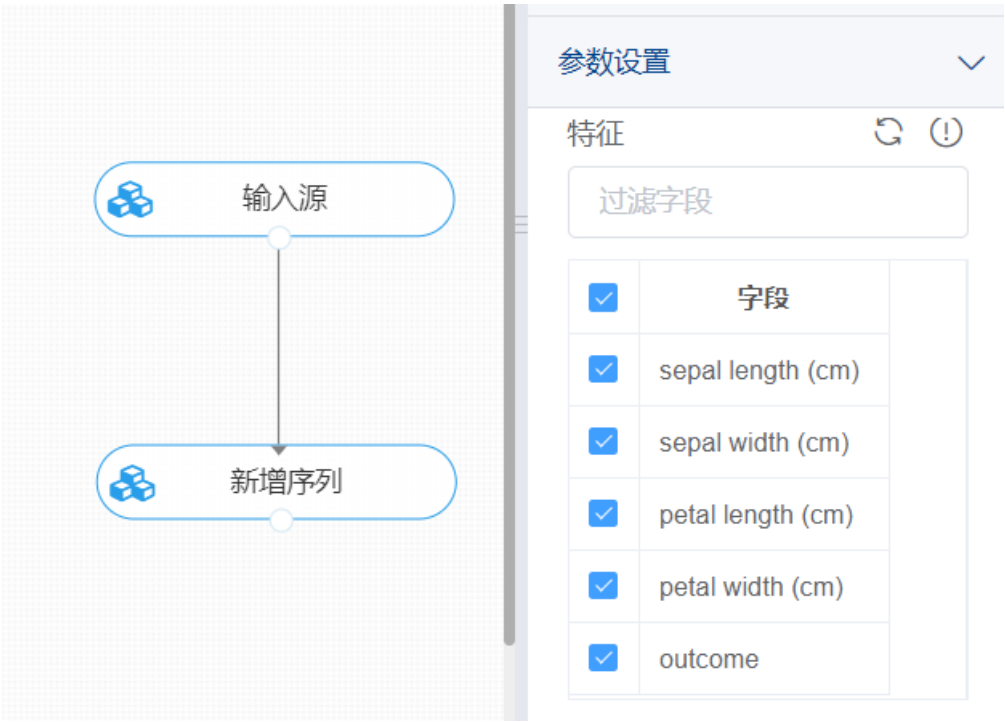
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

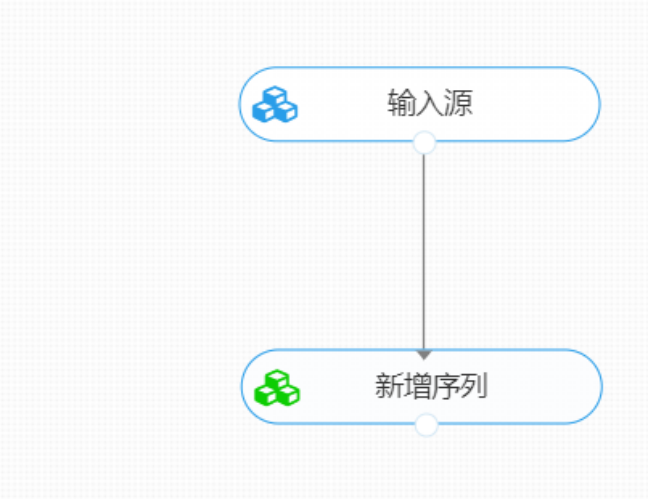
首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行新增序列，将【新增序列】组件与输入源连接，在“字段设置”的“特征”中勾选所有字段，点击新增序列方式，选择特定值，在特定值设置中特定值类型选择int，特定值设置为1，右键单击【新增序列】组件，选择“运行该节点”。



序号	参数名称	数值	原因
1	新增序列方式	特定值	在新一列中填充特定值
2	特定值类型	int	填充int数值
3	特定值	1	填充数值为1



打开日志，查看结果。在日志中可以查看筛选后数据维度的变化。对【新增序列】组件右击，点击“查看日志”。

序号	名称	作用
1	新增序列前维度	观测新增序列前维度
2	新增序列后维度	观测新增序列后维度及变化

8.1.3 数据排序

（1）作用

数据排序组件的作用是将数据表按某一列的数据对行进行升序排序或降序排序。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	展示经过新增序列后表格的部分数据。
2	日志	输出数据新增序列后数据维度变化。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要输出的数据列，数值型
2	参数设置	按哪个字段排序	选择进行排序的列。
3	参数设置	排序方式	进行排序的方式，升序或降序。
4	参数设置	null值位置	排序后空值存放位置，可选顶部或底部

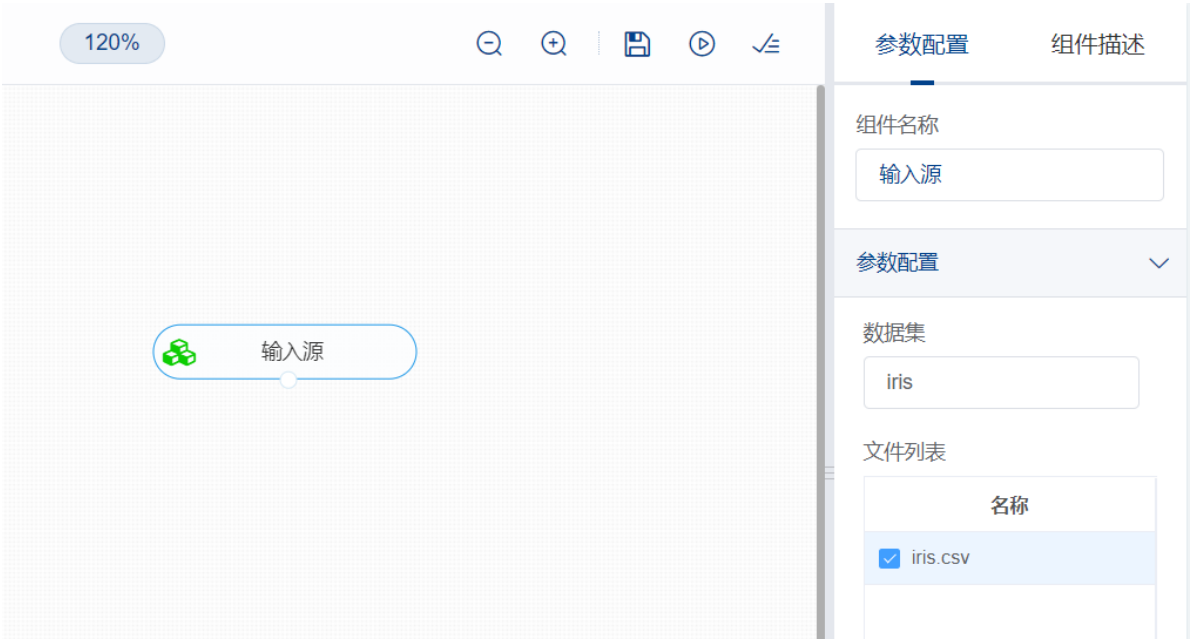
（5）示例

对于“iris”数据集进行数据排序示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行数据排序，将【数据排序】组件与输入源连接，在“字段设置”的“特征”中勾选所有字段，按哪个字段排序中选择“sepal_length”，在参数设置中排序方式选择降序，右键单击【数据排序】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	按哪个字段排序	sepal_length	按sepal_length列排序
2	排序方式	升序	按升序排序

打开日志，查看结果。在日志中可以查看筛选后数据维度的变化。对【数据排序】组件右击，点击“查看日志”。

序号	名称	作用
1	数据	观测排序后数据表

8.1.4 行列转置

（1）作用

行列转置组件的作用是将数据的行转置为列，亦或是将多列转置为行。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	展示经过行列转置后表格的部分数据。

（4）参数

序号	分组	参数	解释
1	字段选择	是否选择转置索引	若选择是，转置结果列为我们的转置索引，若为否，则为简单的行列翻转
2	字段选择	转置索引选择	可根据选择的转置索引进行行列转置
3	字段选择	特征列	需要转置列，数值型

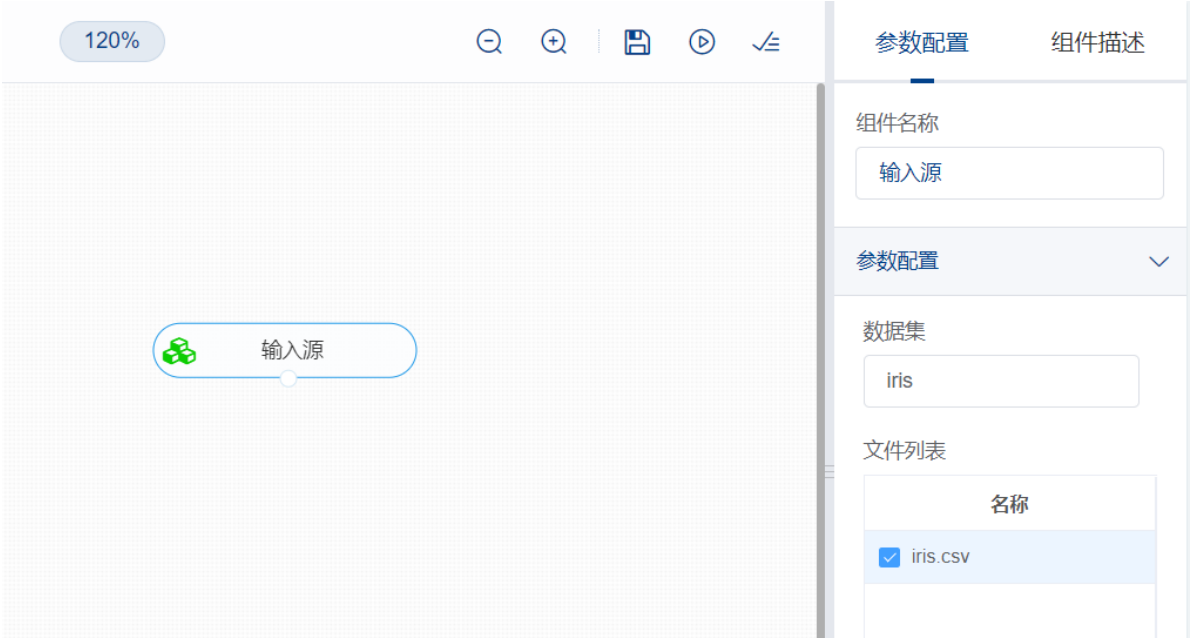
（5）示例

对于“iris”数据集进行行列转置示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



- 情况一：

进行无转置索引的行列转置，将【行列转置】组件与输入源连接，在“字段设置”的“特征”中勾选“sepal_length”，右键单击【行列转置】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	特征列	sepal_length	将sepal_length列转置为行

运行成功后可以查看转置后数据表的变化。对【行列转置】组件右击，点击“查看数据”。

序号	名称	作用
1	数据	观测排序后数据表
2	日志	查看转置后的数据样式

- 情况二：

进行有转置索引的行列转置，将【行列转置】组件与输入源连接，在“字段设置”的“是否选择转置索引”中选择 “是”，并在“转置索引”中选择所需的转置索引名“outcome”，继而在“特征”中选择需要转置的数据，右键单击【行列转置】组件，选择“运行该节点”。

运行成功后可以查看转置后数据表的变化。

8.1.5 行扁平化

（1）作用

行扁平化组件的作用是在原数据表中选取新的索引列作为行，选取特征列和填充新表的列来构造一个新的数据表。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	展示经过行扁平化后表格的部分数据。
2	日志	展示经过行扁平化后表格的部分数据。

（4）参数

序号	分组	参数	解释
1	字段选择	新的索引列	用于新表第一列索引的列
2	字段选择	新的特征列	用于新表作为每列特征的列
3	字段选择	用于填充新框架值的列	填充新表的值

（5）示例

对于“iris”数据集进行行扁平化示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行行扁平化，将【行扁平化】组件与输入源连接，在“字段设置”中，新的索引列选择sepal_length，新的特征列选择species，用于填充新框架的值选择sepal_width，右键单击【行扁平化】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	新的索引列	sepal_length	将sepal_length作为新表的行索引
2	新的特征列	species	将species作为新表每列的特征
3	用于填充新框架值得列	sepal_width	用sepal_width列的值填充新表

打开日志，查看结果。在日志中可以查看行扁平化后部分数据表。对【行扁平化】组件右击，点击“查看日志”。

序号	名称	作用
1	数据	观测排序后数据表

8.1.6 数据采样

（1）作用

数据采样组件的作用是在原数据表中按照指定比例随机选取数据作为新数据表输出。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	展示经过数据采样后表格的部分数据。
2	日志	展示经过数据采样后表格的部分数据。

（4）参数

序号	分组	参数	解释
1	参数设置	采样比例	采取作为新表的数据比例

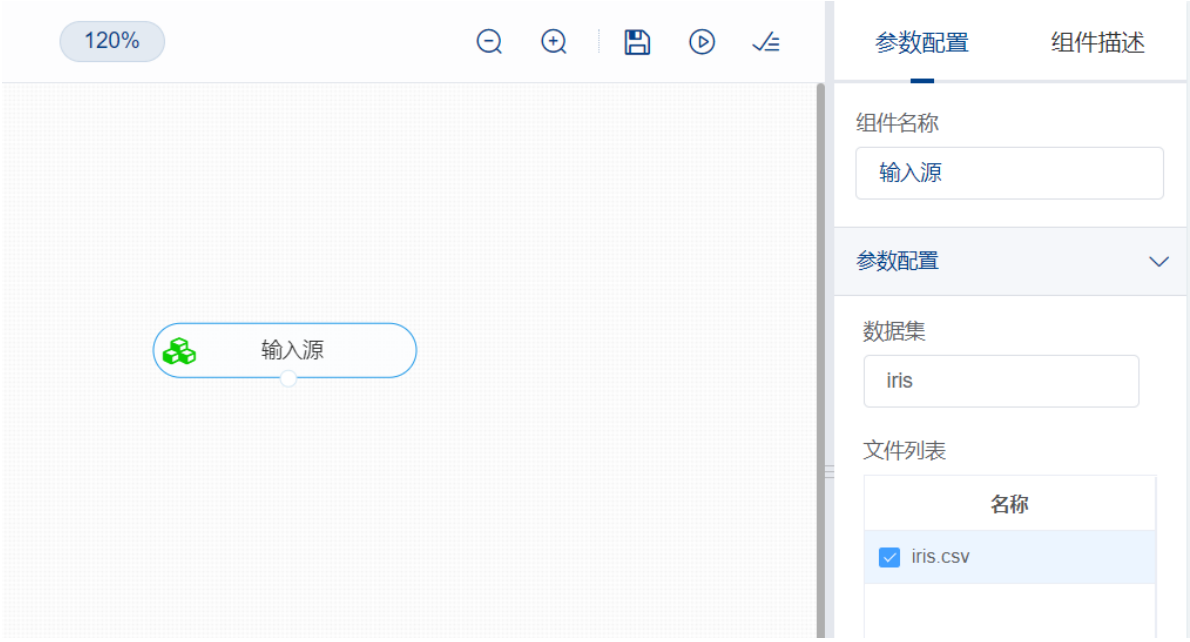
（5）示例

对于“iris”数据集进行数据采样示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

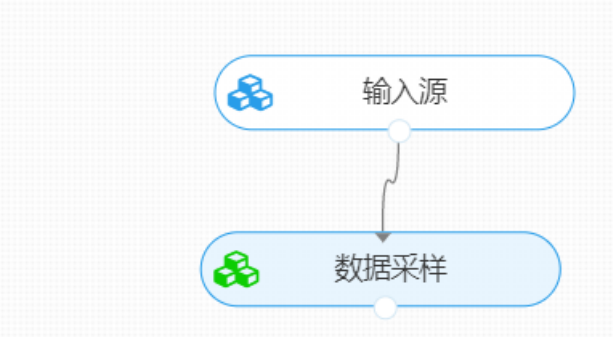
iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行数据采样，将【数据采样】组件与输入源连接，在参数设置中，采样比例设置为0.1，右键单击【数据采样】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	采样比例	0.1	在原数据表中选择10%的数据作为新表



打开日志，查看结果。在日志中可以查看进行采样后的部分数据。对【数据采样】组件右击，点击“查看日志”。

序号	名称	作用
1	数据	观测采样后数据表

8.1.7 数据编码化

（1）作用

数据编码化组件的作用是根据原数据表中数据的特征将数据转化为整数。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	展示经过数据编码化后表格的部分数据。
2	日志	展示经过数据编码化后表格的部分数据。

（4）参数

序号	分组	参数	解释
1	参数设置	编码化特征	需要进行编码化的列

（5）示例

对于“iris”数据集进行数据编码化示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行数据编码化，将【数据编码化】组件与输入源连接，在编码化特征中，选择“species”字段，右键单击【数据编码化】组件，选择“运行该节点”。

参数名称	序号	数值	原因
编码化特征	1	species	将iris数据集标签进行编码化

打开数据，查看结果。对【数据编码化】组件右击，点击“查看数据”，可以查看进行编码化后的部分数据。

序号	名称	作用
1	数据	观测编码化后数据表

8.1.8 修改列名

（1）作用

修改列名组件的作用是修改数据表中列的名称。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	展示经过修改列名后表格的部分数据。
2	日志	展示经过修改列名后表格的部分数据。

（4）参数

序号	分组	参数	解释
1	参数设置	列索引名	需要修改的列名，用英文逗号间隔

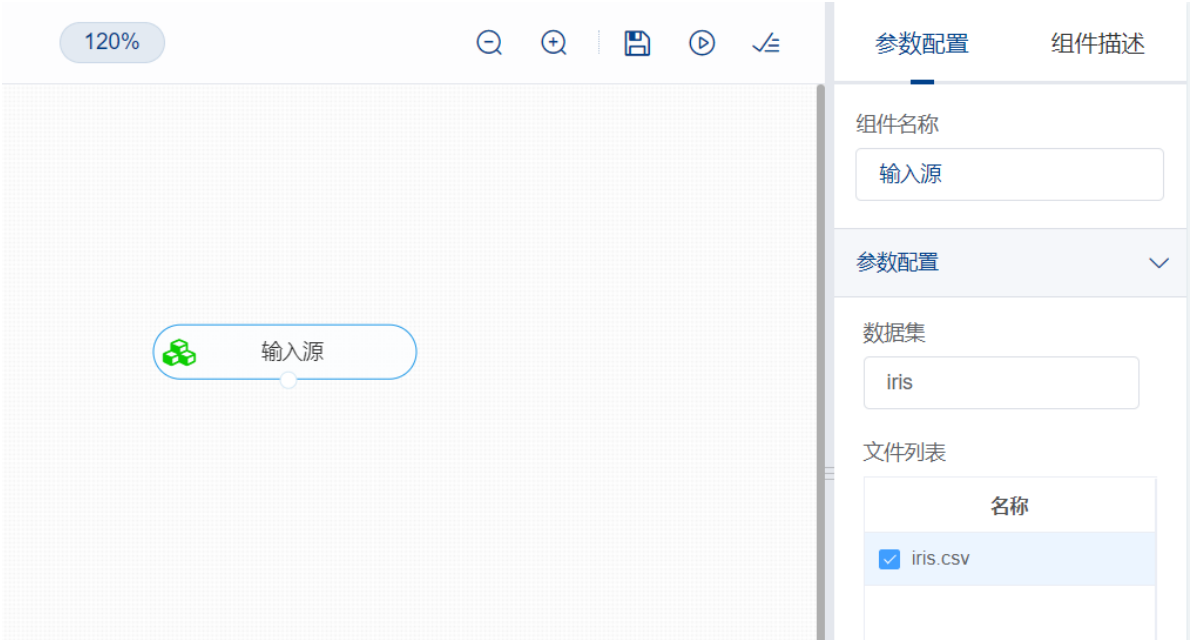
（5）示例

对于“iris”数据集进行修改列名示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

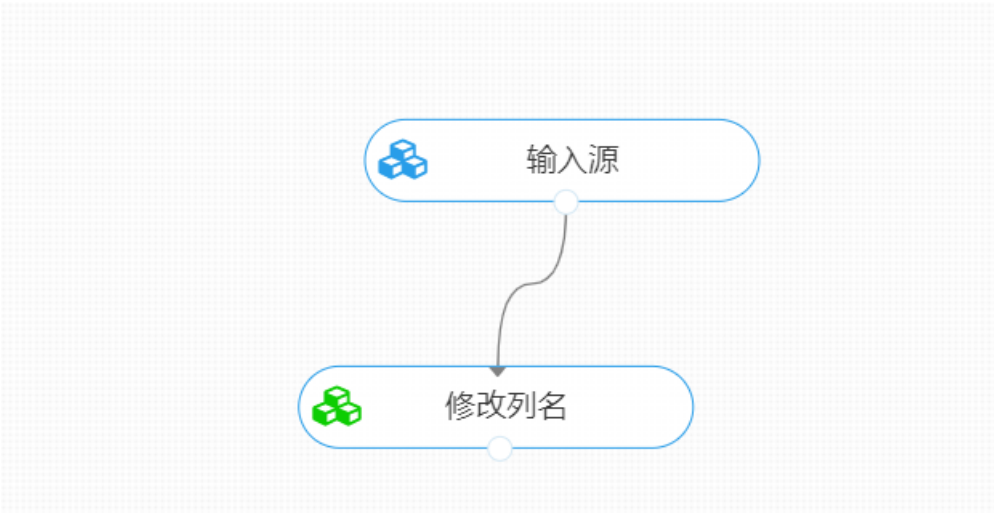
iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行列名修改，将【修改列名】组件与输入源连接，在参数设置中，对原特征列名对应进行修改为“花萼长度,花萼宽度,花瓣长度,花瓣宽度,类别”，右键单击【修改列名】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	新列名	新字段名	将需要修改的字段名进行新字段名填写



打开数据，查看结果。对【修改列名】组件右击，点击“查看数据”，可以查看修改列名后的部分数据。

序号	名称	作用
1	数据	观测修改列名后的数据表
2	日志	查看新旧列名对比与修改后的数据表

8.1.9 记录去重

（1）作用

记录去重组件的作用是根据选中列的主键对数据表进行删除重复项操作。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	展示经过去重后表格的部分数据。
2	日志	展示经过去重后表格的部分数据。

（4）参数

序号	分组	参数	解释
1	字段设置	特征	需要进行去重并输出的数据表列
2	字段设置	去重主键	选择某特征列作为去重标准
3	参数设置	去重方式	有三个可选参数，分别是 first、last、False 默认为 first，表示只保留第一次出现的重复项，删除其余重复项； last 表示只保留最后一次出现的重复项；False 则表示删除所有重复项。

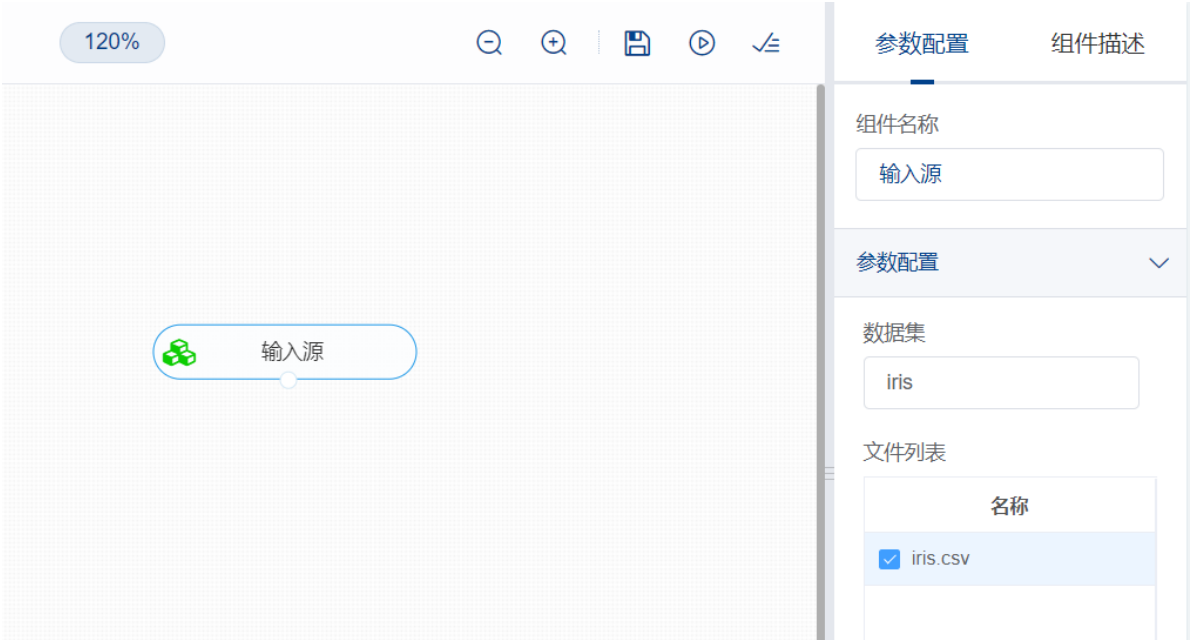
（5）示例

对于“iris”数据集进行记录去重示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行去重操作，将【记录去重】组件与输入源连接，在特征中选择所有字段，去重主键中，选择“sepal length (cm)”，“sepal width (cm)”字段，右键单击【记录去重】组件，选择“运行该节点”。

序号	参数名称	数值	说明
1	去重主键	sepal length (cm)、sepal width (cm)	根据花萼形状对数据去重
2	去重方式	first	只保留第一次出现的重复值

打开数据，查看结果。对【记录去重】组件右击，点击“查看数据”，可以查看去重后的数据。

序号	名称	作用
1	数据	观测去重后的数据表
2	日志	查看数据去重前后的维度变化

8.1.10 数据拆分

（1）作用

数据拆分组件主要用于数据集划分，其作用是将输入数据按指定比例拆分，一般划分为训练集与测试集，提供给模型进行训练以及评估。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	train_out.csv	拆分后的训练集。
2	test_out.csv	拆分后的测试集。
3	日志	展示拆分后数据表变化。

（4）参数

序号	分组	参数	解释
1	字段设置	特征	选择需要划分的数据
2	参数设置	测试集占比	划分为测试集的比例，如果是整数的话就是样本的数量
3	参数设置	随机种子	该组随机数的编号，在需要重复试验的时候，保证得到一组一样的随机数。

（5）示例

对于“iris”数据集进行数据拆分示例。

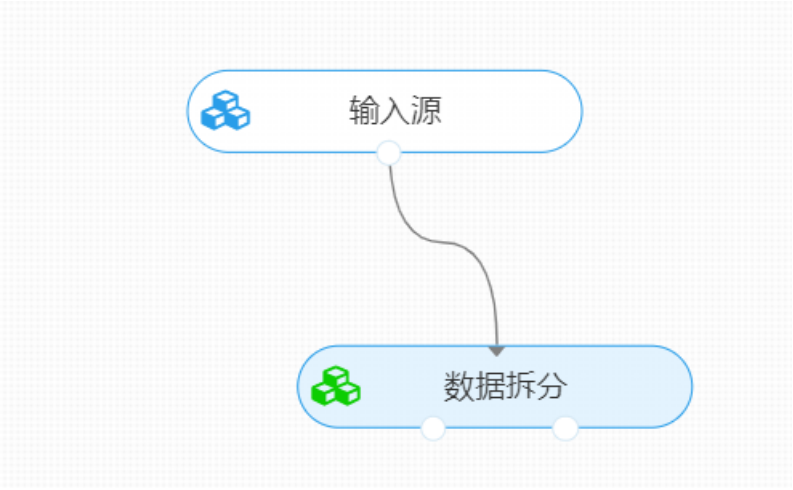
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行数据拆分，将【数据拆分】组件与输入源连接，在特征中选择所有字段，参数设置中，测试集占比设置为0.2，右键单击【数据拆分】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	测试集占比	0.2	划分20%的数据作为测试集



打开日志，查看结果。对【数据拆分】组件右击，点击“查看日志”，可以查看拆分的结果。

序号	名称	作用
1	日志	查看拆分后的数据表维度变化
2	数据	可分别查看train_out、test_out的部分数据


```
原始数据维度为：  (150, 5)
训练数据维度为：  (120, 5)
测试数据维度为：  (30, 5)
```

8.1.11 数据抽取

（1）作用

数据抽取组件的作用是在数据表中抽取指定连续的行作为新表并输出。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	抽取后的数据表。
2	日志	展示抽取后数据表维度变化。

（4）参数

序号	分组	参数	解释
1	字段设置	特征	选择需要抽取的数据列
2	参数设置	起始位置	开始抽取的行数
3	参数设置	终止位置	结束抽取的行数

（5）示例

对于“iris”数据集进行数据抽取示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行数据抽取，将【数据抽取】组件与输入源连接，在特征中选择所有字段，参数设置中，起始位置设置为10，终止位置设置为20，右键单击【数据抽取】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	起始位置	10	从第10行开始抽取
2	终止位置	20	抽取至第20行

打开日志，查看结果。对【数据抽取】组件右击，点击“查看日志”，可以查看抽取后的结果。

序号	名称	作用
1	日志	查看抽取后的数据表维度变化

8.1.12 分组聚合

（1）作用

分组聚合组件的作用是在数据表中选取指定列作为分组标准，并对数据进行计数、求和、计算均值、方差等操作。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	分组聚合后的数据表。
2	日志	展示分局聚合后数据表维度变化。

（4）参数

序号	分组	参数	解释
1	字段设置	特征	选择需要分组聚合的数据列
2	字段设置	分组主键	作为分组主键的列
3	参数设置	聚合函数	选择对数据进行聚合的操作， 具体有计数、最大值、最小值、平均值、标准差、求和以及中位数等聚合方式

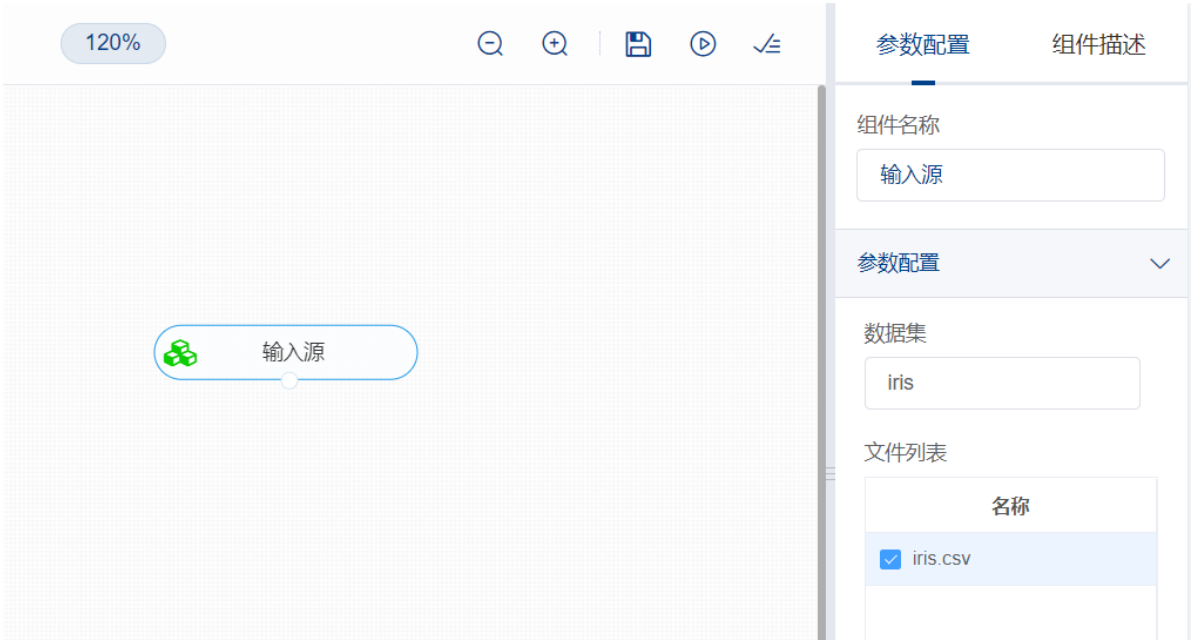
（5）示例

对于“iris”数据集进行分组聚合示例。

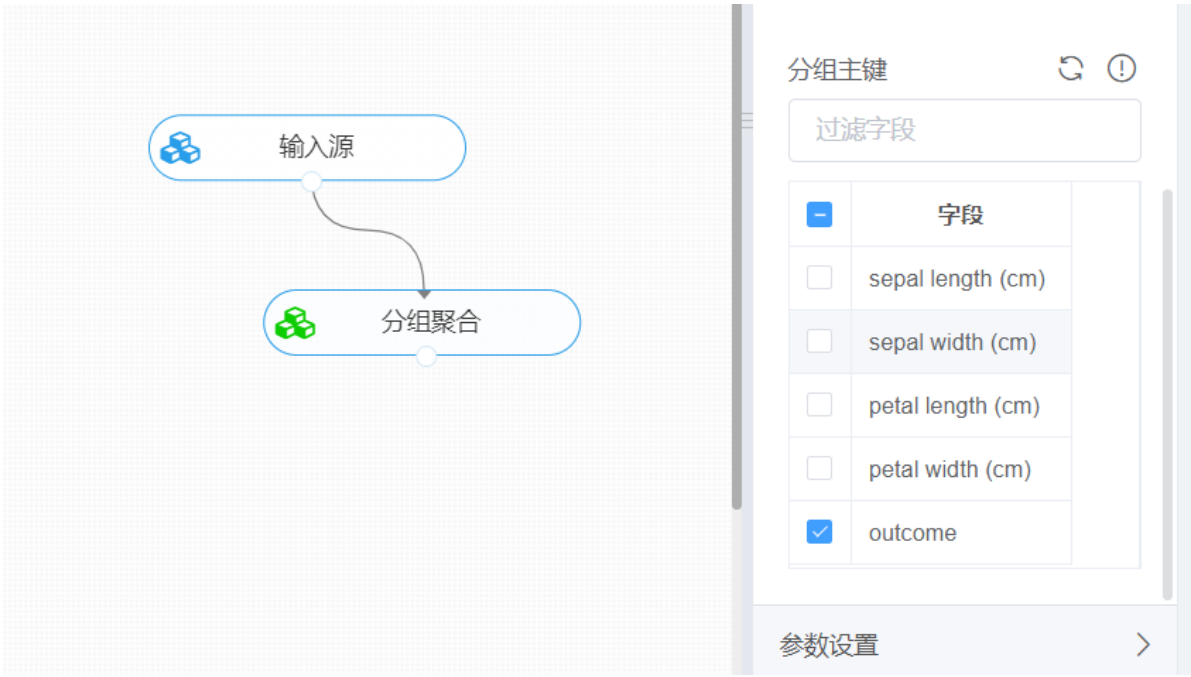
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行分组聚合，将【分组聚合】组件与输入源连接，在特征中选择所有字段，分组主键选择“outcome”，参数设置中，聚合函数选择平均值，右键单击【分组聚合】组件，选择“运行该节点”。



序号	参数名称	数值	原因
1	分组主键	outcome	将花的类别作为分组主键
2	聚合函数	平均值	计算每类花的特征列平均值

打开日志，查看结果。对【分组聚合】组件右击，点击“查看日志”，可以查看分组聚合后的结果。

序号	名称	作用
1	日志	查看分组聚合后的数据表维度变化

查看日志

分组聚合前后数据维度为:

(150, 5)

分组聚合后数据维度为:

(3, 5)

	outcome	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	0	5.006	3.428	1.462	0.246
1	1	5.936	2.770	4.260	1.326
2	2	6.588	2.974	5.552	2.026

8.1.13 字符集转换

（1）作用

字符集转换组件的作用是将选中特征列的编码格式进行转换。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	转换后的数据表。

（4）参数

序号	分组	参数	解释
1	字段设置	特征	选择需要进行转换的数据列
2	参数设置	编码格式	原编码格式
3	参数设置	解码格式	更改后的编码格式

（5）示例

对于“air_data”数据集进行字符集转换示例。

member_nc	ffp_date	load_time	flight_cc	sum_yr_1	sum_yr_2	seg_km_su	last_to_end	avg_disc
54993	2006/11/2	2014/3/31	210	239560	234188	580717	1	0.961639
28065	2007/2/19	2014/3/31	140	171483	167434	293678	7	1.252314
55106	2007/2/1	2014/3/31	135	163618	164982	283712	11	1.254676
21189	2008/8/22	2014/3/31	23	116350	125500	281336	97	1.09087
39546	2009/4/10	2014/3/31	152	124560	130702	309928	5	0.970658
56972	2008/2/10	2014/3/31	92	112364	76946	294585	79	0.967692
44924	2006/3/22	2014/3/31	101	120500	114469	287042	1	0.965347
22631	2010/4/9	2014/3/31	73	82440	114971	287230	3	0.96207
32197	2011/6/7	2014/3/31	56	72596	87401	321489	6	0.828478
31645	2010/7/5	2014/3/31	64	85258	60267	375074	15	0.70801
58877	2010/11/18	2014/3/31	43	69056	91581	262013	22	0.988658
37994	2004/11/13	2014/3/31	145	92975	126821	271438	6	0.952535
28012	2006/11/23	2014/3/31	29	44750	53977	321529	67	0.799127
54943	2006/10/25	2014/3/31	118	105466	119832	179514	3	1.398382
57881	2010/2/1	2014/3/31	50	68941	79076	270067	2	0.921985
1254	2008/3/28	2014/3/31	22	69300	54764	234721	65	1.026085
8253	2010/7/15	2014/3/31	101	93840	93114	172231	7	1.386525
58899	2010/11/10	2014/3/31	40	66239	63260	284160	45	0.837844
26955	2006/4/6	2014/3/31	64	99735	93006	169358	2	1.401596
41616	2011/8/29	2014/3/31	38	60930	52316	332896	24	0.708285
21501	2008/7/30	2014/3/31	106	69566	122763	167113	4	1.369404
41281	2011/6/7	2014/3/31	23	46800	198224	214590	6	1.061631
47229	2005/4/10	2014/3/31	94	59169	74497	305250	23	0.741804
28474	2010/4/13	2014/3/31	20	64258	59600	222380	74	1.004904
58472	2010/2/14	2014/3/31	44	38510	75816	281837	17	0.787308
13942	2010/10/14	2014/3/31	62	72806	83496	243674	17	0.90299

首先将air_data数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“air_data”，勾选文件“air_data.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行字符集转换，将【字符集转换】组件与输入源连接，需要进行编码转换特征中选择member_no字段，参数设置中，编码格式选择UTF-8，解码格式选择ASCII，右键单击【字符集转换】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	需要进行编码转换特征	member_no	转换member_no列
2	编码格式	UTF-8	原格式为UTF-8
3	解码格式	ASCII	转换为ASCII

打开数据，查看结果。对【字符集转换】组件右击，点击“查看数据”，可以查看转换后的结果。

序号	名称	作用
1	数据	查看转换后的数据表

8.1.14 主键合并

（1）作用

主键合并组件的作用是根据选定的主键对两张数据表进行连接。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	合并后的数据表。
2	日志	合并后数据表维度变化。

（4）参数

序号	分组	参数	解释
1	字段设置	左表特征	选择第一个表进行合并的数据列
2	字段设置	右表特征	选择第二个表进行合并的数据列
3	参数设置	连接方式	两个表之间的连接方式，可选左连接、右连接、内连接、外连接
4	参数设置	left_on	第一个表进行连接的主键
5	参数设置	right_on	第二个表进行连接的主键

（5）示例

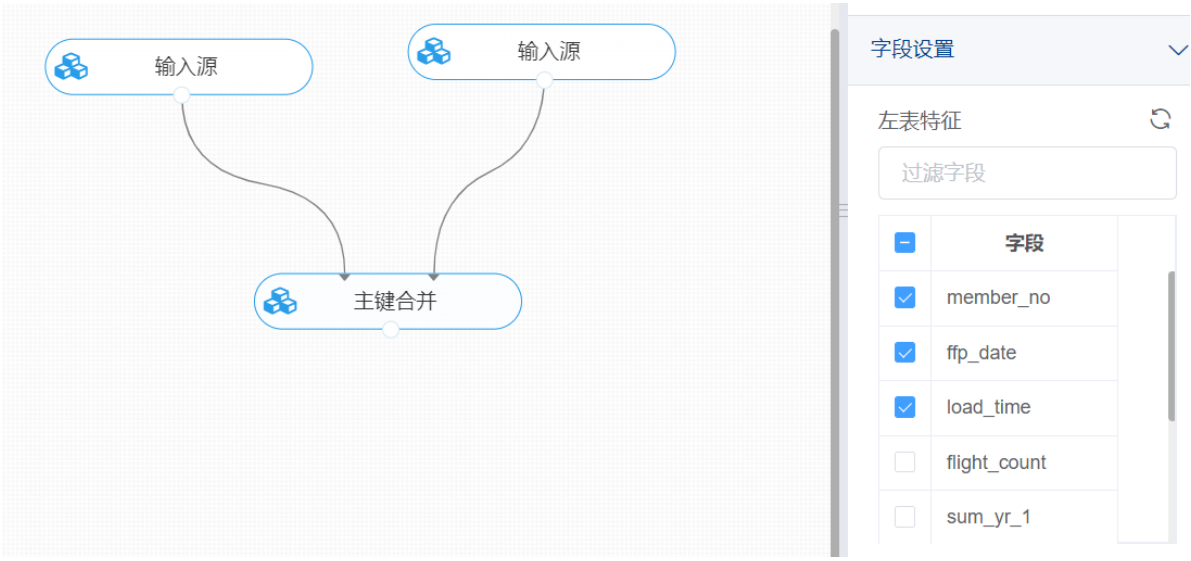
对于“air_data”数据集进行主键合并示例。

member_nc	ffp_date	load_time	flight_cc	sum_yr_1	sum_yr_2	seg_km_su	last_to_end	avg_disc
54993	2006/11/2	2014/3/31	210	239560	234188	580717	1	0.961639
28065	2007/2/19	2014/3/31	140	171483	167434	293678	7	1.252314
55106	2007/2/1	2014/3/31	135	163618	164982	283712	11	1.254676
21189	2008/8/22	2014/3/31	23	116350	125500	281336	97	1.09087
39546	2009/4/10	2014/3/31	152	124560	130702	309928	5	0.970658
56972	2008/2/10	2014/3/31	92	112364	76946	294585	79	0.967692
44924	2006/3/22	2014/3/31	101	120500	114469	287042	1	0.965347
22631	2010/4/9	2014/3/31	73	82440	114971	287230	3	0.96207
32197	2011/6/7	2014/3/31	56	72596	87401	321489	6	0.828478
31645	2010/7/5	2014/3/31	64	85258	60267	375074	15	0.70801
58877	2010/11/18	2014/3/31	43	69056	91581	262013	22	0.988658
37994	2004/11/13	2014/3/31	145	92975	126821	271438	6	0.952535
28012	2006/11/23	2014/3/31	29	44750	53977	321529	67	0.799127
54943	2006/10/25	2014/3/31	118	105466	119832	179514	3	1.398382
57881	2010/2/1	2014/3/31	50	68941	79076	270067	2	0.921985
1254	2008/3/28	2014/3/31	22	69300	54764	234721	65	1.026085
8253	2010/7/15	2014/3/31	101	93840	93114	172231	7	1.386525
58899	2010/11/10	2014/3/31	40	66239	63260	284160	45	0.837844
26955	2006/4/6	2014/3/31	64	99735	93006	169358	2	1.401596
41616	2011/8/29	2014/3/31	38	60930	52316	332896	24	0.708285
21501	2008/7/30	2014/3/31	106	69566	122763	167113	4	1.369404
41281	2011/6/7	2014/3/31	23	46800	198224	214590	6	1.061631
47229	2005/4/10	2014/3/31	94	59169	74497	305250	23	0.741804
28474	2010/4/13	2014/3/31	20	64258	59600	222380	74	1.004904
58472	2010/2/14	2014/3/31	44	38510	75816	281837	17	0.787308
13942	2010/10/14	2014/3/31	62	72806	83496	243674	17	0.90299

首先将air_data数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“air_data”，勾选文件“air_data.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行主键合并，将【主键合并】组件与输入源连接，左表特征选择member_no、ffp_date、load_time字段，右表特征选择member_no、sum_yr_1、sum_yr_2字段，参数设置中，连接方式选择内连接，left_on选择member_no，right_on选择member_no，右键单击【主键合并】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	左表特征	member_no、ffp_date、load_time	将member_no、ffp_date、load_time列进行合并
2	右表特征	member_no、sum_yr_1、sum_yr_2	将member_no、sum_yr_1、sum_yr_2列进行合并
3	连接方式	内连接	采用内连接方式连接两张表
4	left_on	member_no	选择member_no列作为主键
5	right_on	member_no	选择member_no列作为主键



打开数据，查看结果。对【主键合并】组件右击，点击“查看数据”，可以查看合并后的结果。

序号	名称	作用
1	数据	查看合并后的数据表
2	日志	查看合并后数据表的维度变化

8.1.15 数据标准化

（1）作用

在数据分析之前，我们通常需要先将数据标准化（normalization），利用标准化后的数据进行数据分析。数据标准化也就是统计数据的指数化。其处理主要包括数据同趋化处理和无量纲化处理两个方面。

数据同趋化处理主要解决不同性质数据问题，对不同性质指标直接加总不能正确反映不同作用力的综合结果，须先考虑改变逆指标数据性质，使所有指标对测评方案的作用力同趋化，再加总才能得出正确结果。数据无量纲化处理主要解决数据的可比性。

数据标准化的方法有很多种，常用的有“极差标准化”、“Z-score标准化”等。经过标准化处理，原始数据均转换为无量纲化指标测评值，即各指标值都处于同一个数量级别上，可以进行后续综合数据分析。

- Z-score标准化：标准差标准化，也称零均值标准化，将数据转化为标准正态分布（均值为0，方差为1），极差标准化后的数据结果集中在0附近且方差值为1，具体计算公式为：X_scale = (X - X_mean) / X_std

- 极差标准化：将数据在缩放在固定区间，默认缩放到区间 [0, 1]。如果有新数据加入，可能会导致最大值（Xmax）和最小值（Xmin）发生变化，就需要进行重新定义，并重新计算极差，具体公式为：X_scale = (X - X_min) / (X_max - X_min)
- max_abs标准化：数据的缩放比例为绝对值最大值，并保留正负号，即在区间 [-1.0, 1.0] 内。唯一可用于稀疏数据 scipy.sparse的标准化
- Normalization主要思想是对每个样本计算其p-范数，然后对该样本中每个元素除以该范数，这样处理的结果是使得每个处理后样本的p-范数(l1-norm,l2-norm)等于1。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	数据标准化后的数据表。

（4）参数

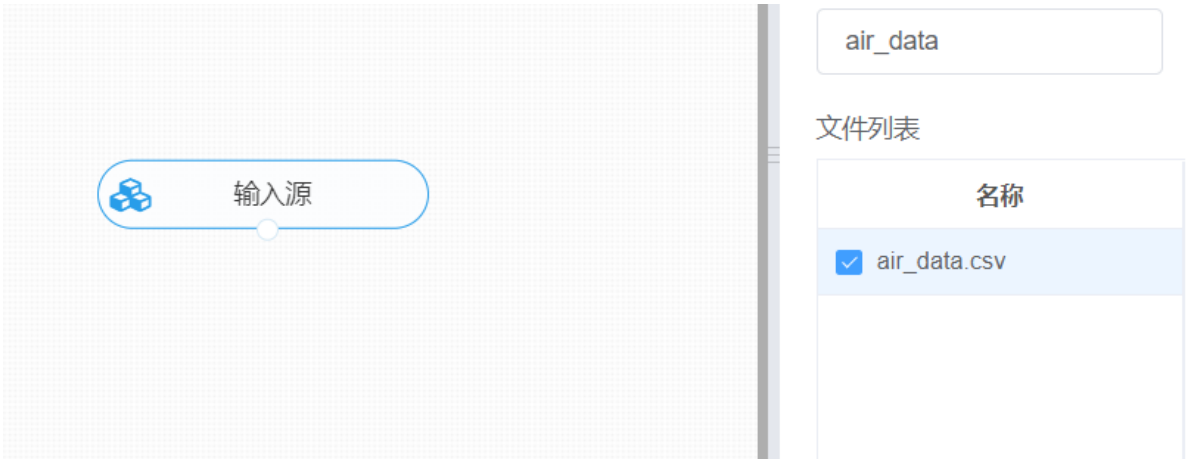
序号	分组	参数	解释
1	字段设置	特征	选择需要进行数据标准化的数据列
2	参数设置	标准化方式	选择对数据进行标准化的方式

（5）示例

对于“air_data”数据集进行分组聚合示例。

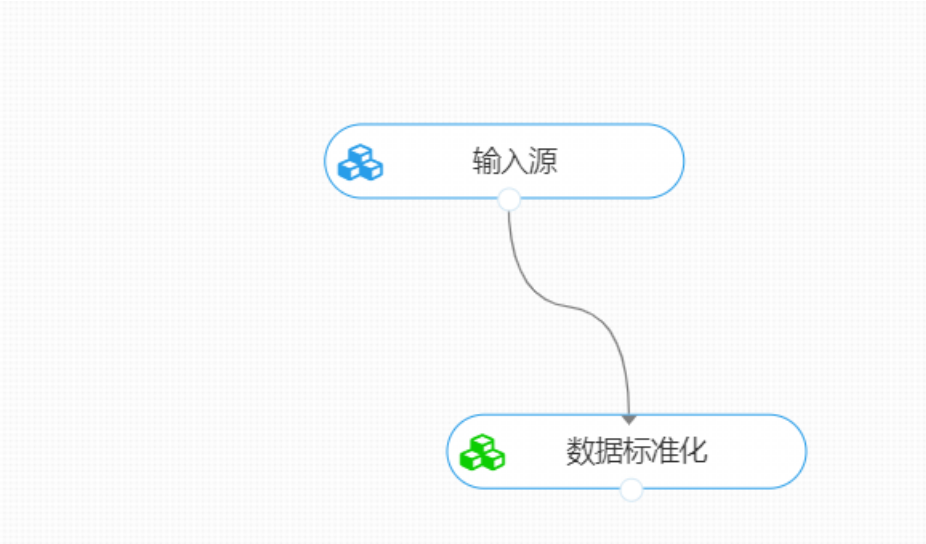
member_nc	ffp_date	load_time	flight_cc	sum_yr_1	sum_yr_2	seg_km_su	last_to_end	avg_disc
54993	2006/11/2	2014/3/31	210	239560	234188	580717	1	0.961639
28065	2007/2/19	2014/3/31	140	171483	167434	293678	7	1.252314
55106	2007/2/1	2014/3/31	135	163618	164982	283712	11	1.254676
21189	2008/8/22	2014/3/31	23	116350	125500	281336	97	1.09087
39546	2009/4/10	2014/3/31	152	124560	130702	309928	5	0.970658
56972	2008/2/10	2014/3/31	92	112364	76946	294585	79	0.967692
44924	2006/3/22	2014/3/31	101	120500	114469	287042	1	0.965347
22631	2010/4/9	2014/3/31	73	82440	114971	287230	3	0.96207
32197	2011/6/7	2014/3/31	56	72596	87401	321489	6	0.828478
31645	2010/7/5	2014/3/31	64	85258	60267	375074	15	0.70801
58877	2010/11/18	2014/3/31	43	69056	91581	262013	22	0.988658
37994	2004/11/13	2014/3/31	145	92975	126821	271438	6	0.952535
28012	2006/11/23	2014/3/31	29	44750	53977	321529	67	0.799127
54943	2006/10/25	2014/3/31	118	105466	119832	179514	3	1.398382
57881	2010/2/1	2014/3/31	50	68941	79076	270067	2	0.921985
1254	2008/3/28	2014/3/31	22	69300	54764	234721	65	1.026085
8253	2010/7/15	2014/3/31	101	93840	93114	172231	7	1.386525
58899	2010/11/10	2014/3/31	40	66239	63260	284160	45	0.837844
26955	2006/4/6	2014/3/31	64	99735	93006	169358	2	1.401596
41616	2011/8/29	2014/3/31	38	60930	52316	332896	24	0.708285
21501	2008/7/30	2014/3/31	106	69566	122763	167113	4	1.369404
41281	2011/6/7	2014/3/31	23	46800	198224	214590	6	1.061631
47229	2005/4/10	2014/3/31	94	59169	74497	305250	23	0.741804
28474	2010/4/13	2014/3/31	20	64258	59600	222380	74	1.004904
58472	2010/2/14	2014/3/31	44	38510	75816	281837	17	0.787308
13942	2010/10/14	2014/3/31	62	72806	83496	243674	17	0.90299

首先将air_data数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“air_data”，勾选文件“air_data.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行数据标准化，将【数据标准化】组件与输入源连接，在特征中选择“sum_yr_1”、“sum_yr_2”、“seg_km_sum”字段，参数设置中，标准化方式选择零均值标准化，右键单击【数据标准化】组件，选择“运行该节点”。

序号	参数名称	数值	说明
1	标准化方式	零均值标准化	对特征列进行零均值标准化操作



打开数据，查看结果。对【数据标准化】组件右击，点击“查看数据”，即可查看标准化后的结果。

序号	名称	作用
1	数据	查看标准化后的数据表
2	日志	查看标准化处理的方法与标准化结果

	member_no	ffp_date	load_time	flight_count	sum_yr_1	sum_yr_2	seg_km_sum	last_to_end	avg_discount
0	54993	2006/11/2	2014/3/31	210	28.880829	26.264073	26.888115	1	0.961639
1	28065	2007/2/19	2014/3/31	140	20.485950	18.594104	13.193949	7	1.252314
2	55106	2007/2/1	2014/3/31	135	19.516082	18.312372	12.718487	11	1.254676
3	21189	2008/8/22	2014/3/31	23	13.687253	13.775929	12.605132	97	1.090870
4	39546	2009/4/10	2014/3/31	152	14.699665	14.373634	13.969210	5	0.970658
...	...	...	...	...	...	...	...	...	...
62983	18375	2011/5/20	2014/3/31	2	-0.660385	-0.643897	-0.762851	297	0.000000
62984	36041	2010/3/8	2014/3/31	4	-0.660385	-0.643897	-0.434522	89	0.000000
62985	45690	2006/3/30	2014/3/31	2	-0.660385	-0.643897	-0.693197	29	0.000000
62986	61027	2013/2/6	2014/3/31	2	-0.660385	-0.643897	-0.629268	400	0.000000

8.1.16 衍生变量

（1）作用

衍生变量组件的作用是根据输入的表达式对特征列的数据进行四则运算得到新的列，并保存到原始数据中成为新的数据列。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	衍生变量后的数据表。

（4）参数

序号	分组	参数	解释
1	字段设置	特征	选择需要进行衍生变量的数据列
2	参数设置	新特征名	衍生出来的新列名
3	参数设置	表达式	对特征列进行运算的表达式

（5）示例

对于“air_data”数据集进行衍生变量示例。

member_nc	ffp_date	load_time	flight_cc	sum_yr_1	sum_yr_2	seg_km_su	last_to_end	avg_disc
54993	2006/11/2	2014/3/31	210	239560	234188	580717	1	0.961639
28065	2007/2/19	2014/3/31	140	171483	167434	293678	7	1.252314
55106	2007/2/1	2014/3/31	135	163618	164982	283712	11	1.254676
21189	2008/8/22	2014/3/31	23	116350	125500	281336	97	1.09087
39546	2009/4/10	2014/3/31	152	124560	130702	309928	5	0.970658
56972	2008/2/10	2014/3/31	92	112364	76946	294585	79	0.967692
44924	2006/3/22	2014/3/31	101	120500	114469	287042	1	0.965347
22631	2010/4/9	2014/3/31	73	82440	114971	287230	3	0.96207
32197	2011/6/7	2014/3/31	56	72596	87401	321489	6	0.828478
31645	2010/7/5	2014/3/31	64	85258	60267	375074	15	0.70801
58877	2010/11/18	2014/3/31	43	69056	91581	262013	22	0.988658
37994	2004/11/13	2014/3/31	145	92975	126821	271438	6	0.952535
28012	2006/11/23	2014/3/31	29	44750	53977	321529	67	0.799127
54943	2006/10/25	2014/3/31	118	105466	119832	179514	3	1.398382
57881	2010/2/1	2014/3/31	50	68941	79076	270067	2	0.921985
1254	2008/3/28	2014/3/31	22	69300	54764	234721	65	1.026085
8253	2010/7/15	2014/3/31	101	93840	93114	172231	7	1.386525
58899	2010/11/10	2014/3/31	40	66239	63260	284160	45	0.837844
26955	2006/4/6	2014/3/31	64	99735	93006	169358	2	1.401596
41616	2011/8/29	2014/3/31	38	60930	52316	332896	24	0.708285
21501	2008/7/30	2014/3/31	106	69566	122763	167113	4	1.369404
41281	2011/6/7	2014/3/31	23	46800	198224	214590	6	1.061631
47229	2005/4/10	2014/3/31	94	59169	74497	305250	23	0.741804
28474	2010/4/13	2014/3/31	20	64258	59600	222380	74	1.004904
58472	2010/2/14	2014/3/31	44	38510	75816	281837	17	0.787308
13942	2010/10/14	2014/3/31	62	72806	83496	243674	17	0.90299

首先将air_data数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“air_data”，勾选文件“air_data.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行衍生变量示例，将【衍生变量】组件与输入源连接，在特征中选择所有字段，参数设置中，新特征名中输入“l”，表达式输入“LOAD_TIME - FFP_DATE”，右键单击【衍生变量】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	新特征列	l	将衍生的新列命名为l
2	表达式	LOAD_TIME - FFP_DATE	用load_time减去ffp_data计算出来的值作为新一列的值

打开数据，查看结果。对【衍生变量】组件右击，点击“查看数据”，可以查看衍生变量后的结果。

序号	名称	作用
1	数据	查看衍生变量后的数据表
2	日志	查看衍生变量后数据表维度变化

8.1.17 表堆叠

（1）作用

表堆叠组件的作用是将两个数据表按照行或列拼接在一起。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	
2	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	衍生变量后的数据表。

（4）参数

序号	分组	参数	解释
1	字段设置	表1特征	选择需要进行表堆叠的数据列
2	字段设置	表2特征	选择需要进行表堆叠的数据列
3	参数设置	合并方式	两张数据表的合并方式

（5）示例

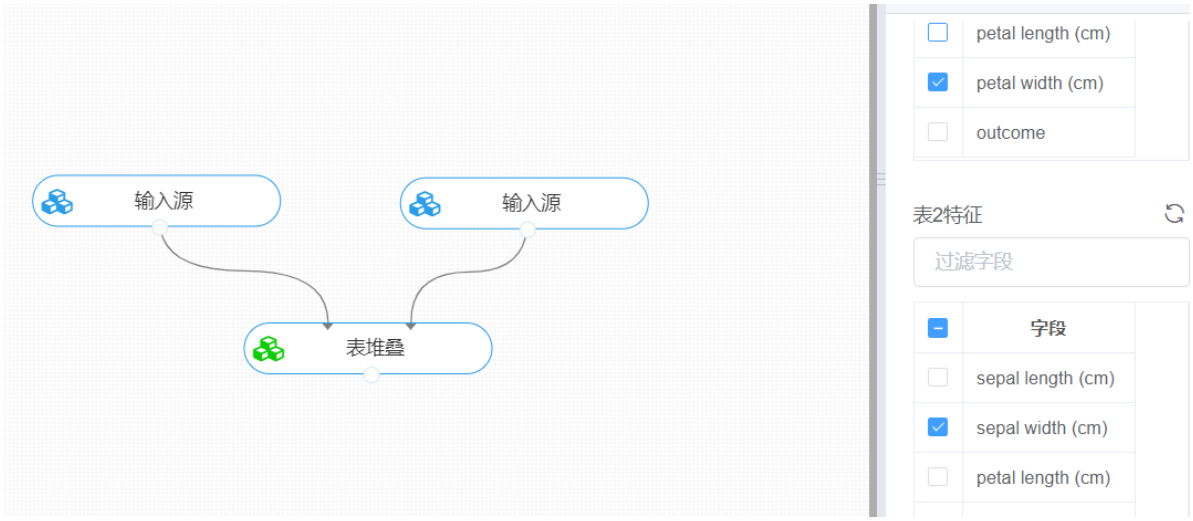
对于“iris”数据集进行表堆叠示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行表堆叠示例，将【表堆叠】组件与输入源连接，在表1特征中选择“petal width”，在表2特征中选择“sepal width”，合并方式选择“按列合并”，右键单击【表堆叠】组件，选择“运行该节点”。



序号	参数名称	数值	原因
1	表1特征	petal width	选择petal width列合并
2	表2特征	sepal width	选择sepal width列合并

打开日志，查看结果。对【表堆叠】组件右击，点击“查看数据”，可以查看表堆叠后的结果。

序号	名称	作用
1	数据	查看合并后的数据表
2	日志	查看合并后数据表维度变化

8.1.18 缺失值处理

（1）作用

在数据分析工作中，若分析过程中存在缺失值可能会导致结果的不可靠，所以缺失值处理是数据处理中的重要环节。缺失值处理组件的作用是查找数据表中空值的位置，并进行按行删除、按列删除或填充等操作。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	缺失值处理后的数据表。

（4）参数

序号	分组	参数	解释
1	字段设置	特征	选择需要进行缺失值处理的数据列
2	参数设置	缺失值处理方式	选择对数据进行缺失值处理的方式
3	参数设置	特定值类型	当选择特定值填充时，填充值的类型
4	参数设置	特定值	当选择特定值填充时，填充的值

（5）示例

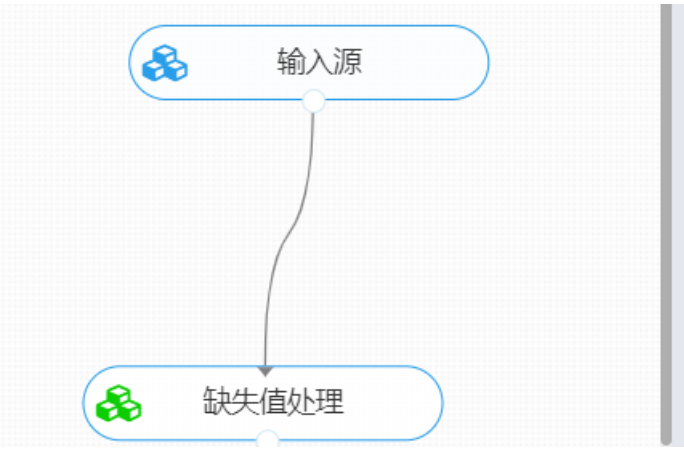
对于“iris”数据集进行缺失值处理示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将air_data数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“air_data”，勾选文件“air_data.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行缺失值处理，将【缺失值处理】组件与输入源连接，在特征中选择所有字段，参数设置中，处理缺失值方式选择按行删除，右键单击【缺失值处理】组件，选择“运行该节点”。



序号	参数名称	数值	原因
1	处理缺失值方式	按行删除	删除含有空值的行

打开数据，查看结果。对【缺失值处理】组件右击，点击“查看数据”，可以查看处理缺失值后的结果。

序号	名称	作用
1	数据	查看处理后的数据表
2	日志	查看处理前后的数据各列缺失值情况以及数据表维度变化

查看日志

缺失值处理

缺失值处理方式：按行删除

缺失值处理前各属性的缺失值个数为：

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	outcome
缺失值个数	0	0	0	0	0

缺失值处理后各属性的缺失值个数为：

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	outcome
缺失值个数	0	0	0	0	0

8.1.19 英文大小写转换

1. 作用

英文大小写转换组件的作用是将选中列的英文字符串进行大小写之间的转换。

2. 输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

3. 输出

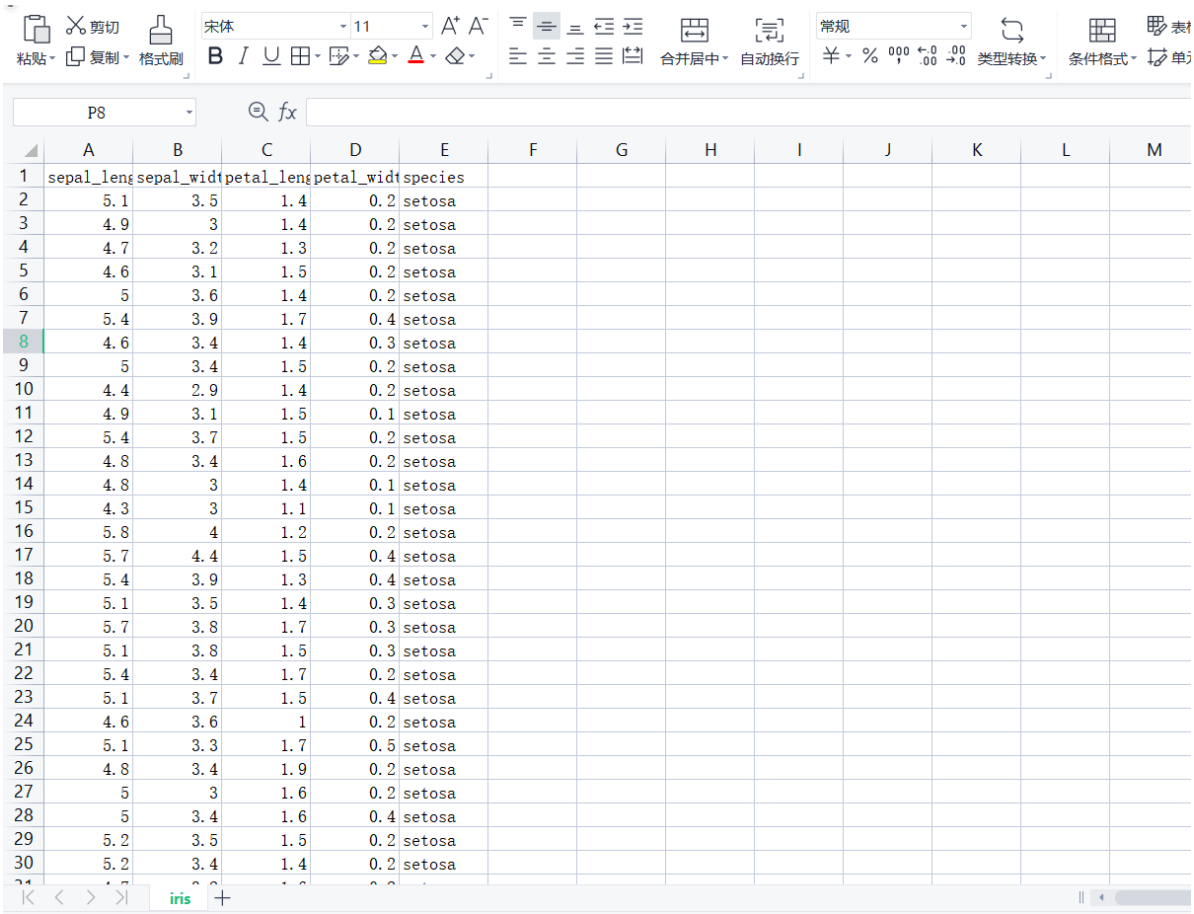
序号	名称	内容
1	data_out.csv	转换后的数据表。

4. 参数

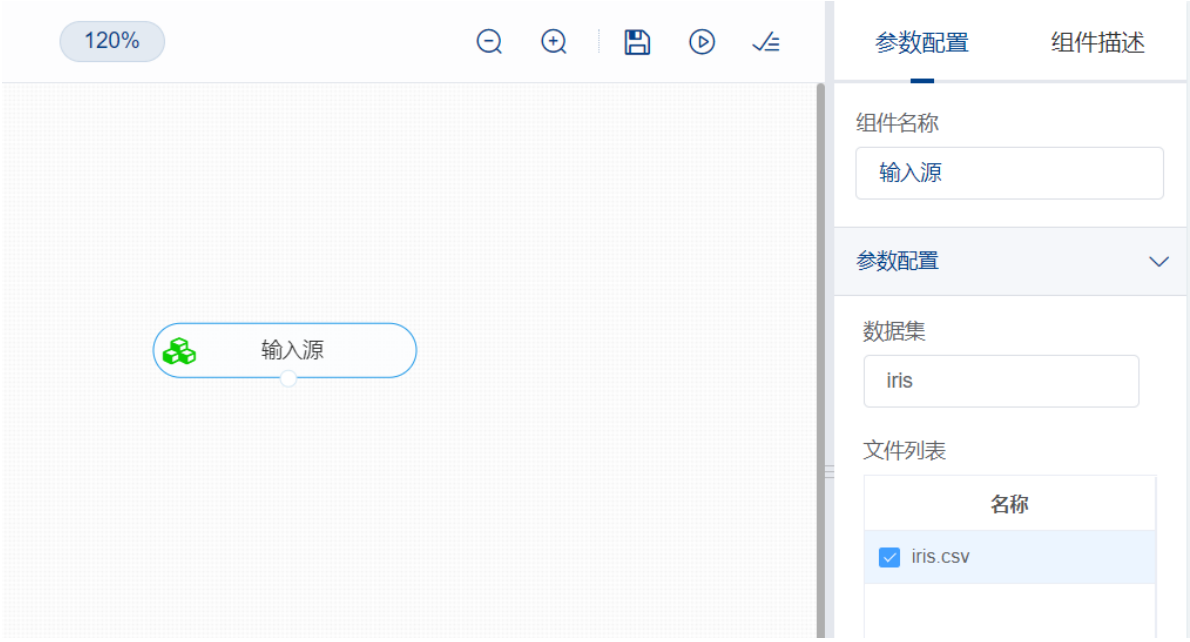
序号	分组	参数	解释
1	字段设置	特征	选择需要进行转换的数据列
2	参数设置	转换操作	对数据进行转换的方式

5. 示例

对于“iris”数据集进行英文大小写转换示例。



首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行英文大小写转换，将【英文大小写转换】组件与输入源连接，在特征中选择“species”字段，参数设置中，转换操作选择全大写，右键单击【英文大小写转换】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	转换操作	全大写	将英文标签转换为全大写英文

打开数据，查看结果。对【英文大小写转换】组件右击，点击“查看数据”，可以查看转换后的结果。

序号	名称	作用
1	数据	查看处理后的数据表
2	日志	查看处理后部分数据表

8.1.20 字符串填充

1. 作用

字符串填充组件的作用是将选中列的字符串进行填充。

2. 输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

3. 输出

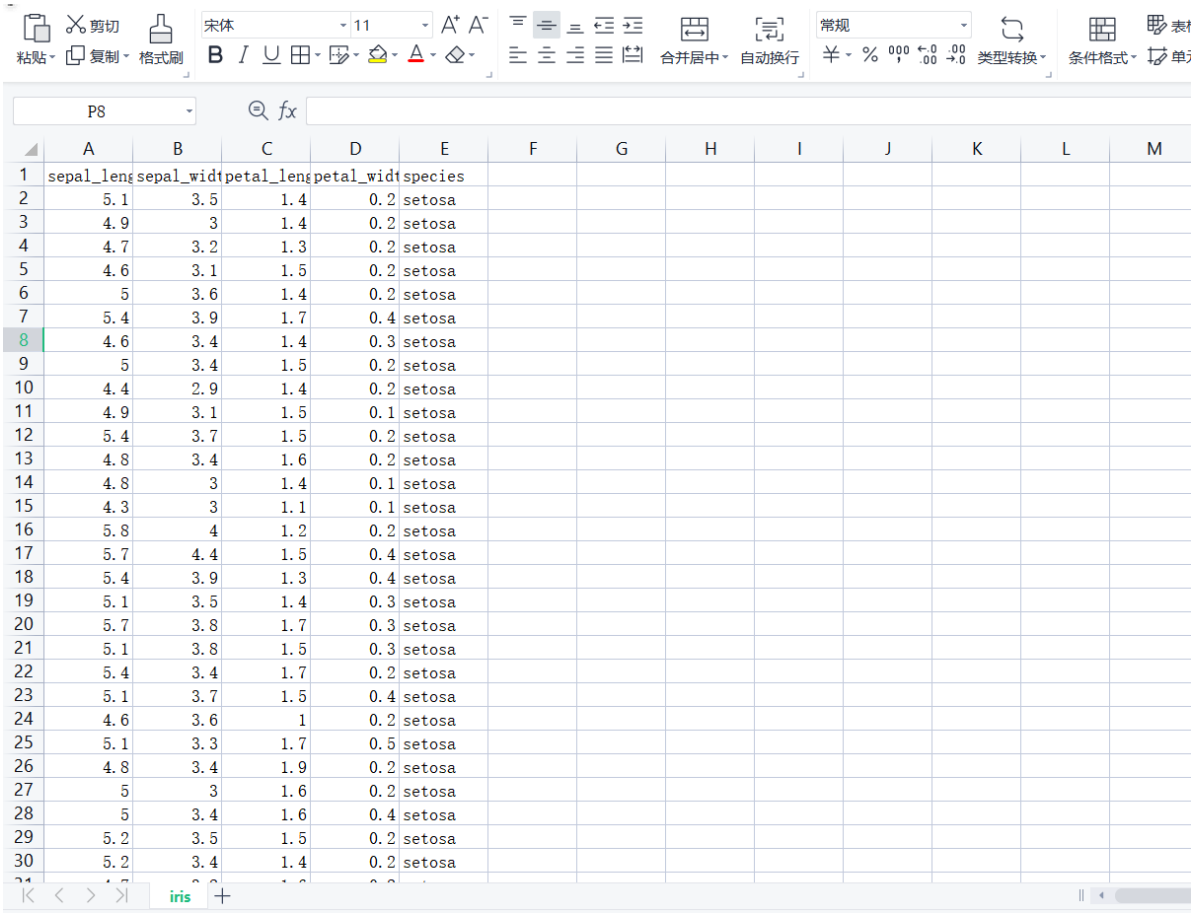
序号	名称	内容
1	data_out.csv	填充后的数据表。

4. 参数

序号	分组	参数	解释
1	字段设置	选择需要进行填充的特征	选择需要进行转换的数据列
2	参数设置	生成的字符串的最小宽度	填充后字符串的长度
3	参数设置	用于填充的字符	填充的字符
4	参数设置	填充位置	填充位置

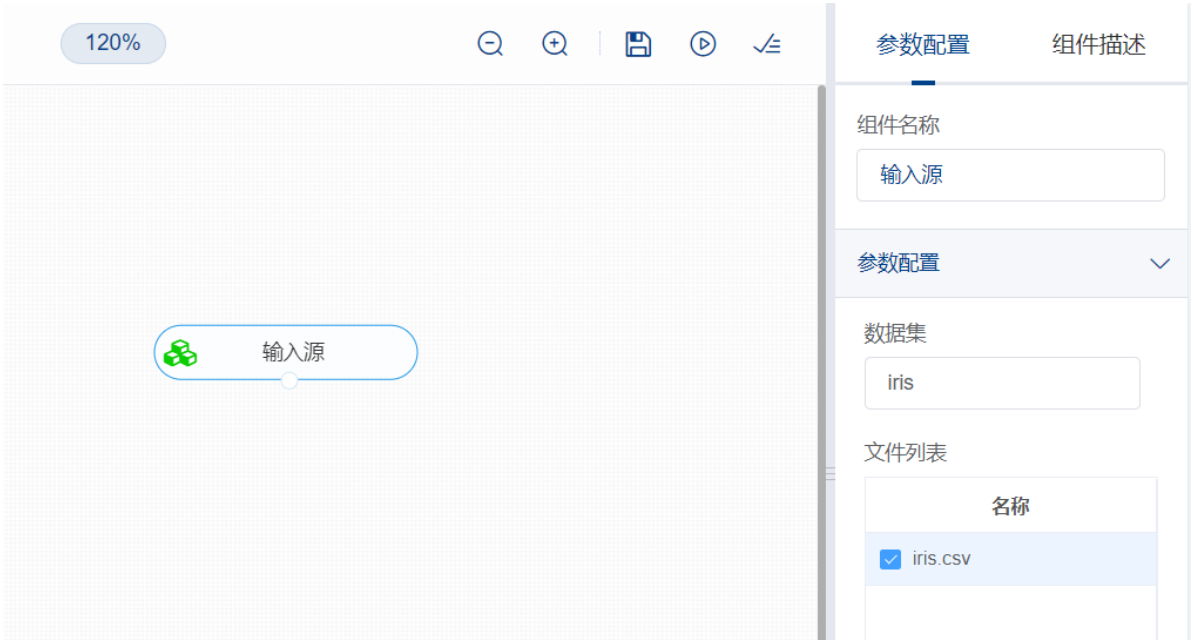
5. 示例

对于“iris”数据集进行字符串填充示例。



	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行字符串填充，将【字符串填充】组件与输入源连接，在特征中选择“species”字段，参数设置中，生成的字符串的最小宽度设置为10，用于填充的字符填入“?”，填充位置选择两端填充，右键单击【字符串填充】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	生成的字符串的最小宽度	10	填充为长达10宽度的字符串
2	用于填充的字符	?	填充?
3	填充位置	两端填充	在字符串的两端填充

打开数据，查看结果。对【字符串填充】组件右击，点击“查看数据”，可以查看填充后的结果。

序号	名称	作用
1	数据	查看处理后的数据表
2	日志	查看处理后部分数据表

8.1.21 字符串截取

1. 作用

字符串截取组件的作用是将选中列的字符串进行截取。

2. 输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

3. 输出

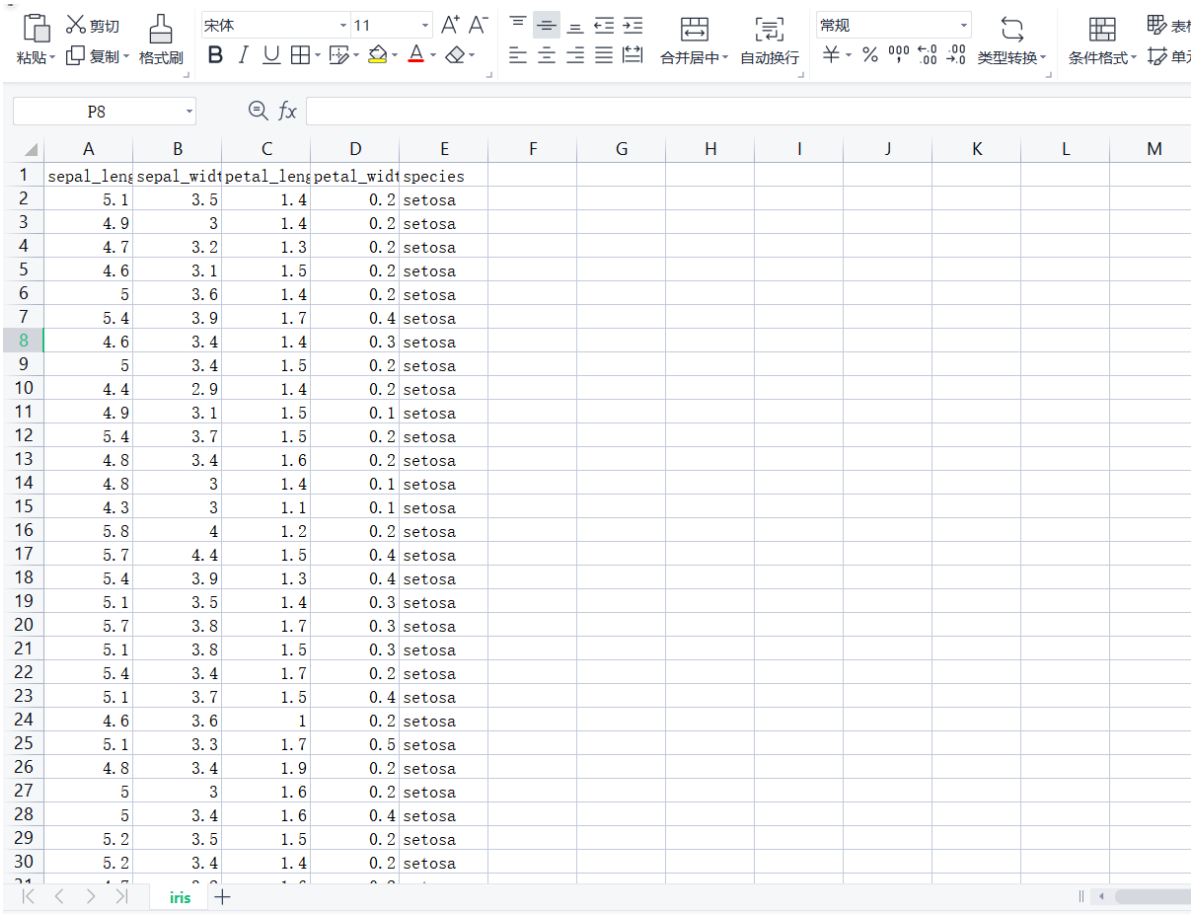
序号	名称	内容
1	data_out.csv	字符串截取后的数据表。

4. 参数

序号	分组	参数	解释
1	字段设置	选择需要进行截取的特征	选择需要进行截取的数据列
2	参数设置	起始位置	字符串开始截取的位置
3	参数设置	终止位置	字符串结束截取的位置
4	参数设置	步长	字符串截取的间隔长度

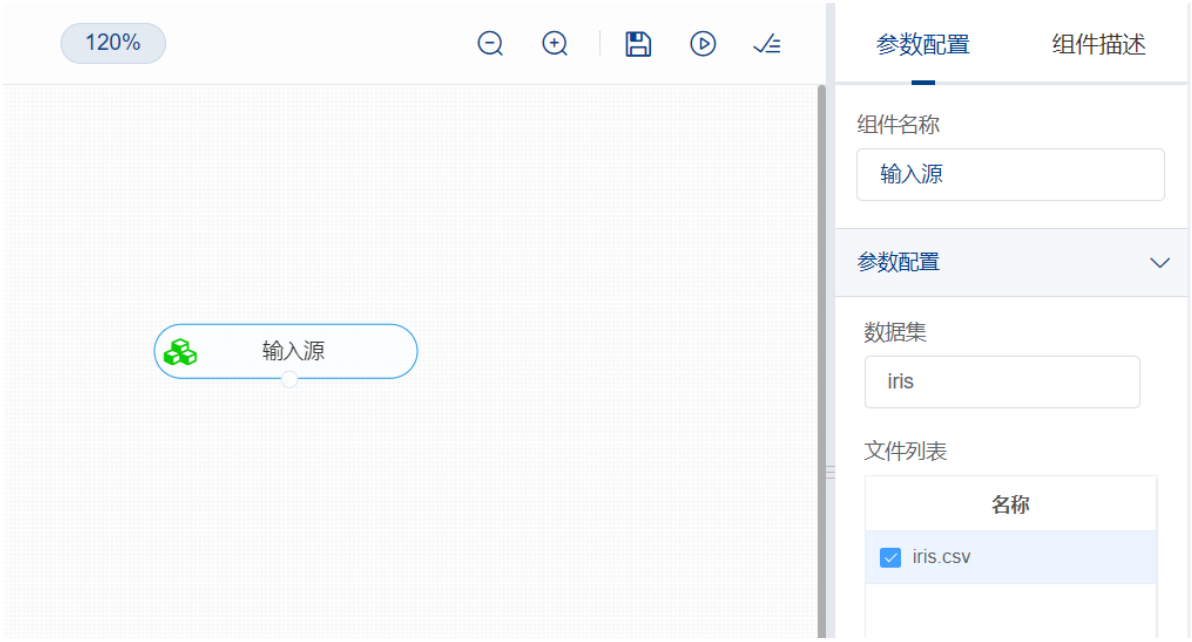
5. 示例

对于“iris”数据集进行字符串截取示例。



	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

（1） 首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 进行字符串截取，将【字符串截取】组件与输入源连接，在特征中选择“species”字段，参数设置中，起始位置设置为0，停止位置设置为7，步长设置为2，右键单击【字符串截取】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	起始位置	0	从第1个字符开始截取
2	终止位置	7	截取至第8个字符
3	步长	2	每2个字符截取一次

（3） 打开数据，查看结果。对【字符串截取】组件右击，点击“查看数据”，可以查看截取后的结果。

序号	名称	作用
1	数据	查看处理后的数据表
2	日志	查看处理后部分数据表

8.1.22 字符串切分

1. 作用

字符串切分组件的作用是将选中列的字符串进行切割操作。

2. 输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

3. 输出

序号	名称	内容
1	data_out.csv	字符串切分后的数据表。

4. 参数

序号	分组	参数	解释
1	字段设置	选择需要进行切分的特征	选择需要进行切分的数据列
2	字段设置	特征列是否需要强转为字符串类型	选择将非字符串类型的数据进行转化
2	参数设置	要分割的字符串或正则表达式	进行分割的间隔字符，或正则表达式
3	参数设置	限制输出的分割数	分割次数上限
4	参数设置	是否展开为数据框	分割后每个字符串是否单独存放

5. 示例

对于“titanic”数据集进行字符切分示例。

	A	B	C	D	E	F
1	Survived	Passenger	Pclass	Sex	Age	
2	0	1	3	male	22	
3	1	2	1	female	38	
4	1	3	3	female	26	
5	1	4	1	female	35	
6	0	5	3	male	35	
7	0	6	3	male		
8	0	7	1	male	54	
9	0	8	3	male	2	
10	1	9	3	female	27	
11	1	10	2	female	14	
12	1	11	3	female	4	
13	1	12	1	female	58	
14	0	13	3	male	20	
15	0	14	3	male	39	
16	0	15	3	female	14	
17	1	16	2	female	55	
18	0	17	3	male	2	
19	1	18	2	male		
20	0	19	3	female	31	
21	1	20	3	female		
22	0	21	2	male	35	
23	1	22	2	male	34	
24	1	23	3	female	15	
25	1	24	1	male	28	
26	0	25	3	female	8	
27	1	26	3	female	38	
28	0	27	3	male		

titanic_data

（1） 首先将titanic数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“titanic”，勾选文件“titanic.csv”，右键单击【输入源】算法，选择“运行该节点”。

（2） 进行字符串截取，将【字符串切分】组件与输入源连接，在特征中选择“sex”字段，参数设置中，要分割的字符串或正则表达式设置为“f”，限制输出的分割数设置为2，是否展开为数据框选择否，右键单击【字符串切分】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	要分割的字符串或正则表达式	o	用f分割字符串
2	限制输出的分割数	2	分割2次
3	是否展开为数据框	否	放在一个列表内

（3） 打开数据，查看结果。对【字符串切分】组件右击，点击“查看数据”，可以查看切分后的结果。

序号	名称	作用
1	数据	查看处理后的数据表
2	日志	查看处理后部分数据表

8.1.23 字符串查找

1. 作用

字符串查找组件的作用是对选中列的字符串进行指定字符查找。

2. 输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

3. 输出

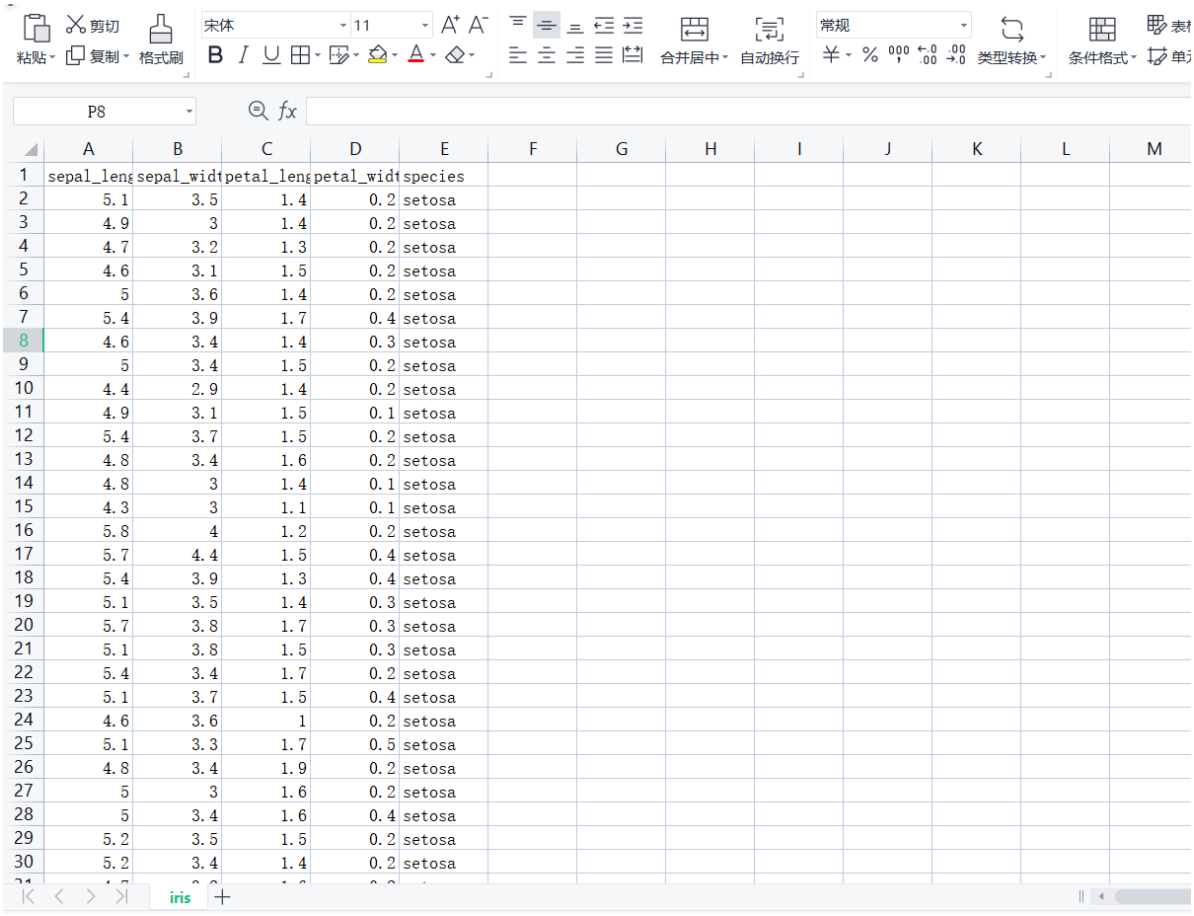
序号	名称	内容
1	data_out.csv	查找字符串后的数据表。

4. 参数

序号	分组	参数	解释
1	字段设置	选择需要进行字符查找的特征	选择需要进行字符查找的数据列
2	参数设置	字符串或正则表达式	需要查找的字符

5. 示例

对于“iris”数据集进行字符查找示例。



	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

（1） 首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

（2） 进行字符串截取，将【字符串查找】组件与输入源连接，在特征中选择“species”字段，参数设置中，字符串或正则表达式输入“s”，右键单击【字符串查找】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	字符串或正则表达式	s	查找字符串中是否存在s

（3） 打开数据，查看结果。对【字符串查找】组件右击，点击“查看数据”，可以查看查找后的结果。

序号	名称	作用
1	数据	查看处理后的数据表
2	日志	查看处理后部分数据表

8.1.24 字符串统计

1. 作用

字符串统计组件的作用是对选中列的字符串进行统计字符出现次数或长度。

2. 输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

3. 输出

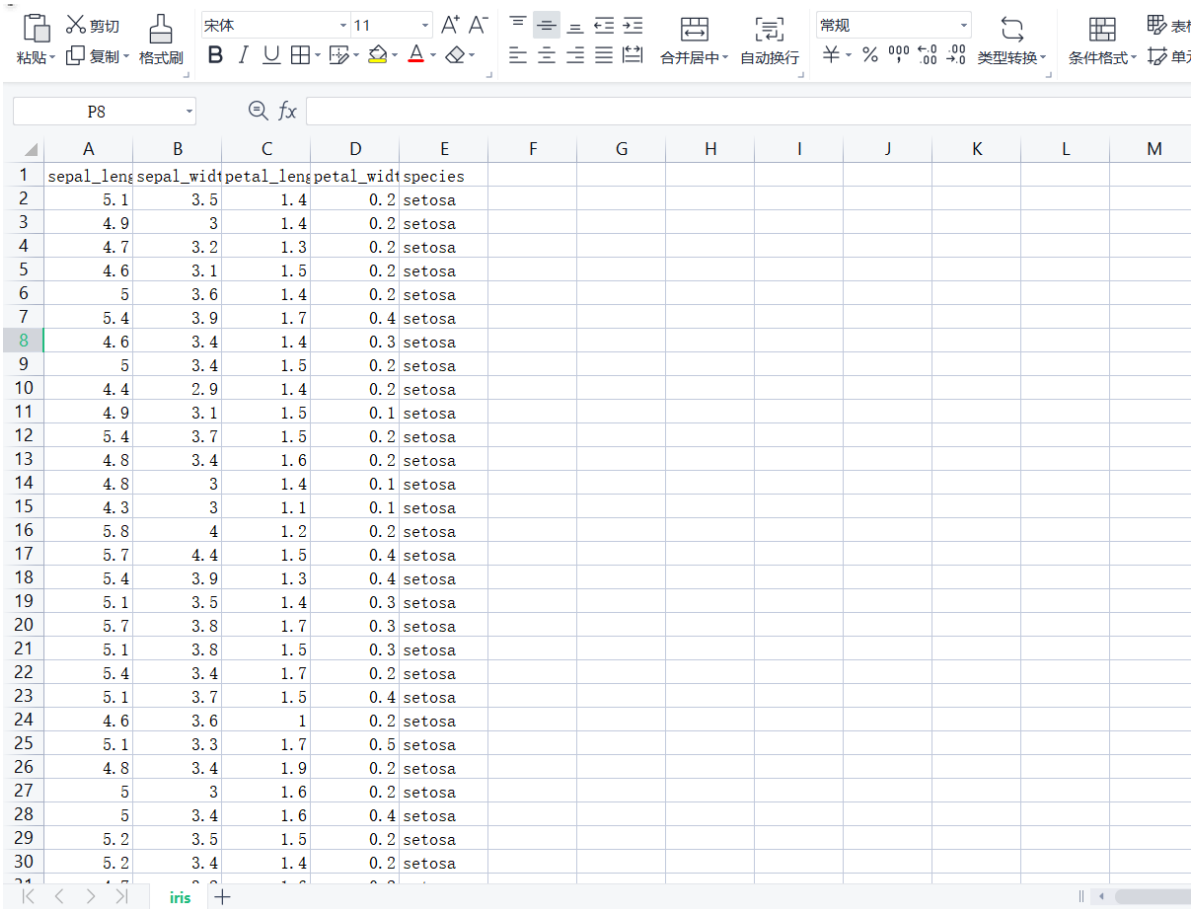
序号	名称	内容
1	data_out.csv	统计字符串后的数据表。

4. 参数

序号	分组	参数	解释
1	字段设置	选择需要进行统计的特征	选择需要进行统计的数据列
2	参数设置	统计类型	统计字符出现次数或长度
3	参数设置	指定的字符或正则表达式	需要统计出现次数的字符或正则表达式

5. 示例

对于“iris”数据集进行字符统计示例。



	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

（1） 首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

（2） 进行字符串统计，将【字符串统计】组件与输入源连接，在特征中选择“species”字段，参数设置中，统计类型选择“统计指定字符出现次数”，字符串或正则表达式输入“s”，右键单击【字符串统计】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	统计类型	统计指定字符出现次数	统计字符串出现次数
2	指定的字符或正则表达式	s	统计字符串s出现次数

（3） 打开数据，查看结果。对【字符串统计】组件右击，点击“查看数据”，可以查看统计后的结果。

序号	名称	作用
1	数据	查看处理后的数据表
2	日志	查看处理后部分数据表

8.1.25 内容判断

1. 作用

内容判断组件的作用是判断输入字符或正则表达式是否在字符串内存在。

2. 输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

3. 输出

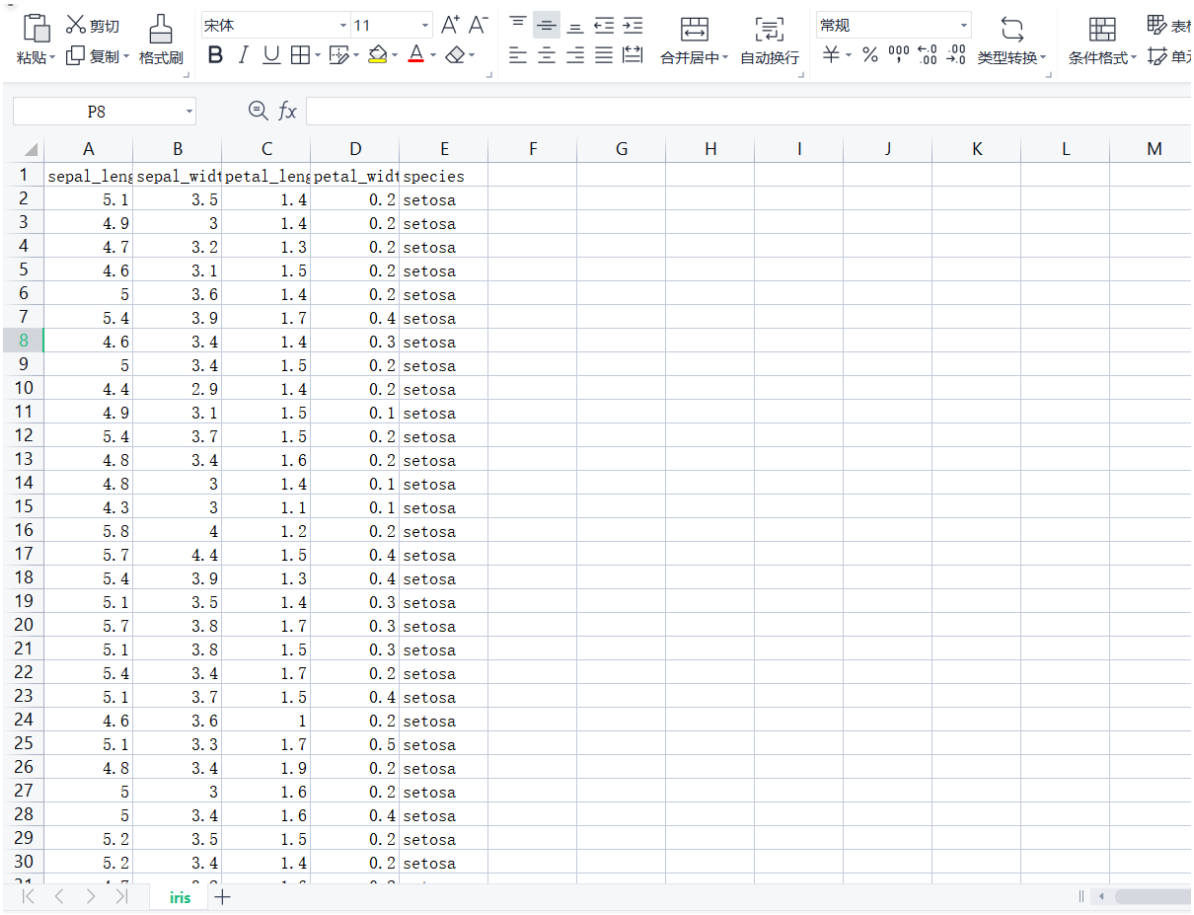
序号	名称	内容
1	data_out.csv	添加判断结果的数据表。

4. 参数

序号	分组	参数	解释
1	字段设置	选择需要进行判断的特征	选择需要进行判断的数据列
2	参数设置	字符串或正则表达式	需要判断存在的字符或正则表达式
3	参数设置	缺失值填充	判断目标为空时的填充值
4	参数设置	是否使用正则表达式	针对特殊符号，默认情况下必须使用转义符，或者设置为否

5. 示例

对于“iris”数据集进行内容判断示例。



	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

（1） 首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

（2） 进行字符串统计，将【内容判断】组件与输入源连接，在特征中选择“species”字段，参数设置中，字符串或正则表达式填入“color”，缺失值填充选择空值，是否使用正则表达式选择否，右键单击【内容判断】组件，选择“运行该节点”。

序号	参数名称	数值	原因
1	字符串或正则表达式	color	判断字符串是否存在“color”字符
2	缺失值填充	空值	缺失值判断填充空值
3	是否使用正则表达式	否	输入项无正则表达式

（3） 打开数据，查看结果。对【内容判断】组件右击，点击“查看数据”，可以查看统计后的结果。

序号	名称	作用
1	数据	查看处理后的数据表
2	日志	查看处理后部分数据表

8.1.26 数据离散化

（1）作用

数据离散化是指将连续的数据进行分段，使其变为一段段离散化的区间。有效的离散化能减小算法的时间和空间开销，提高系统对样本的分类聚类能力和抗噪声能力，同时可以有效的克服数据中隐藏的缺陷，使数据的模型结果更为稳定。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

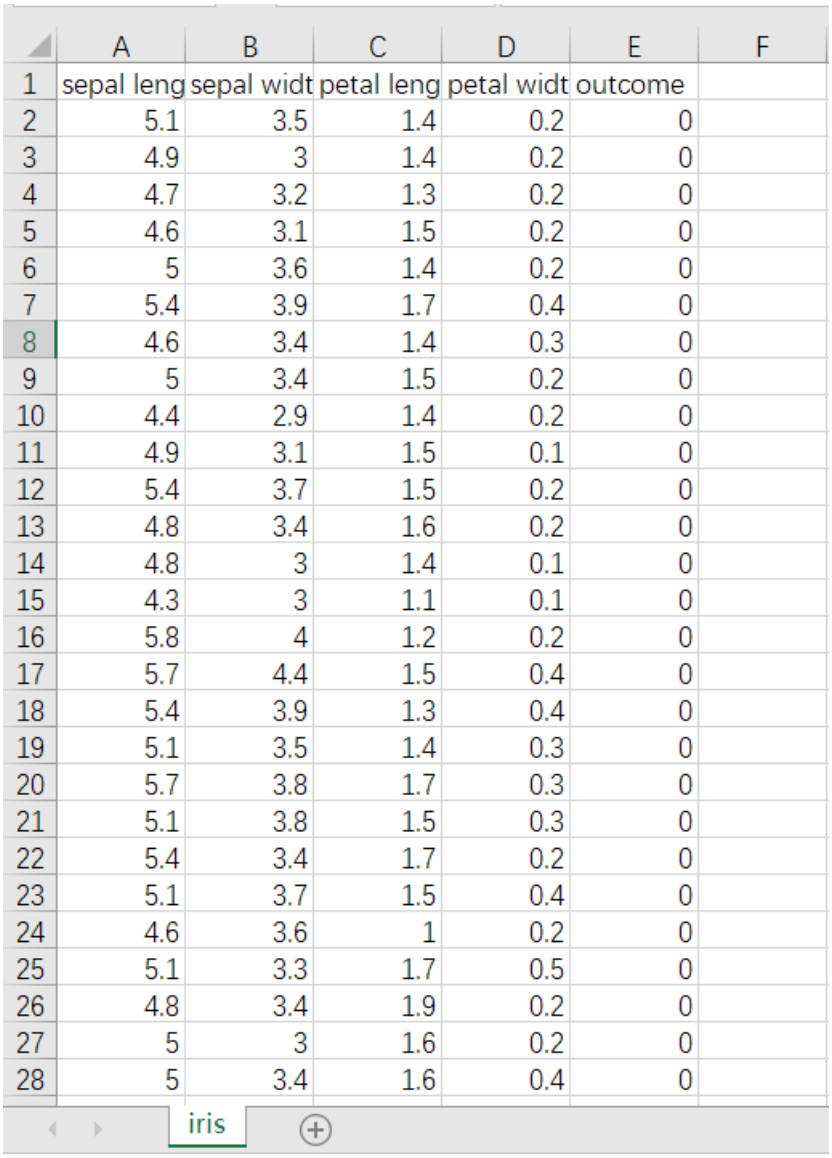
序号	名称	内容
1	data_out.csv	数据离散化后的数据

（4）参数

序号	分组	参数	说明
1	参数设置	离散化数目	选择离散的个数
2	参数设置	离散化方法	等宽、等频、聚类三个离散化的方法
3	字段设置	特征	选择需要进行数据离散化的数据列

（5）示例

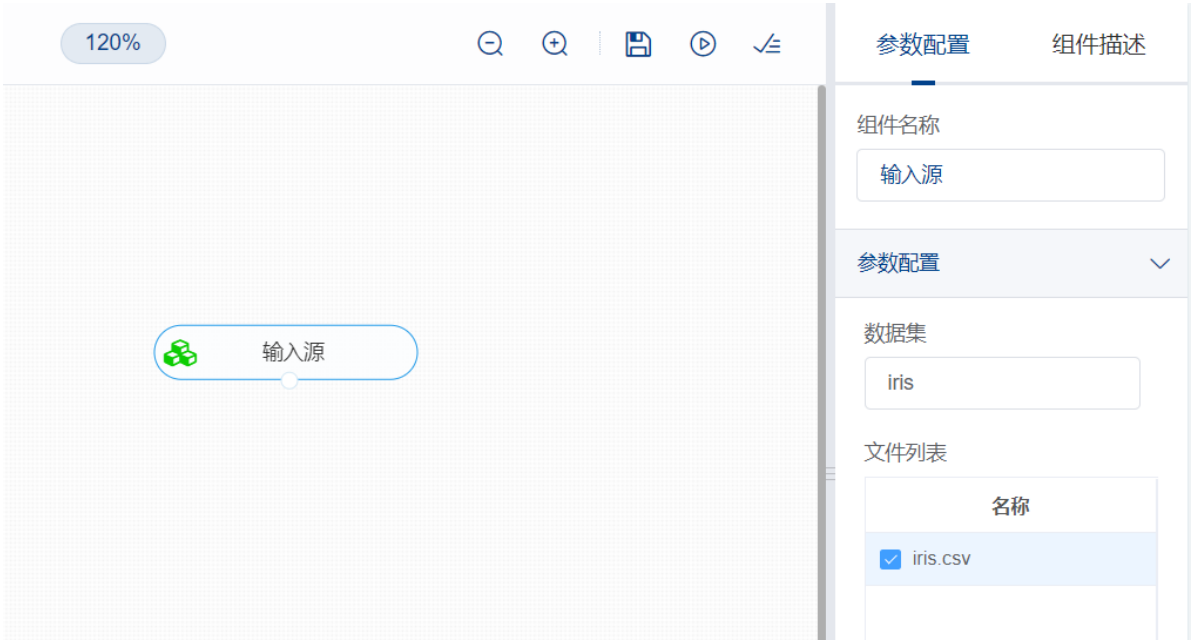
对iris数据集进行数据离散化示例。



	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将iris数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行数据离散化，将【数据离散化】组件与输入源连接，在参数设置中选择“离散个数”与“离散化方法”，在字段设置中选择“sepal length”字段，右键单击【数据离散化】组件，选择“运行该节点”。

打开数据，查看结果。对【数据离散化】组件右击，点击“查看数据”，即可查看离散化后的结果。

序号	名称	作用
1	数据	离散化后的数据表
2	日志	查看离散化方法与离散后的部分数据

8.1.27 时间类型转换

（1）作用

在涉及时间序列的数据分析工作中，难免会遇上关于时间格式处理的情况，该组件算法则主要用于对数据表中的时间列进行格式转换，同时还可自定义转换格式。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	时间类型转换后的数据表。

（4）参数

序号	分组	参数	解释
1	字段设置	特征列	选择需要所需的数据列
2	参数设置	自定义输出格式	时间数据转换后的格式，默认”%Y/%m/%d“
3	参数设置	错误类型	”raise“：抛出异常；“ignore”：忽略；“coerce”：转为NaN

（5）示例

对于“air_data”数据集进行缺失值处理示例。

	A	B	C	D	E	F	G	H	I
1	member_r	ffp_date	load_time	flight_cou	sum_yr_1	sum_yr_2	seg_km_su	last_to_en	avg_discount
2	54993	2006/11/2	2014/3/31	210	239560	234188	580717	1	0.961639
3	28065	2007/2/19	2014/3/31	140	171483	167434	293678	7	1.252314
4	55106	2007/2/1	2014/3/31	135	163618	164982	283712	11	1.254676
5	21189	2008/8/22	2014/3/31	23	116350	125500	281336	97	1.09087
6	39546	2009/4/10	2014/3/31	152	124560	130702	309928	5	0.970658
7	56972	2008/2/10	2014/3/31	92	112364	76946	294585	79	0.967692
8	44924	2006/3/22	2014/3/31	101	120500	114469	287042	1	0.965347
9	22631	2010/4/9	2014/3/31	73	82440	114971	287230	3	0.96207
10	32197	2011/6/7	2014/3/31	56	72596	87401	321489	6	0.828478
11	31645	2010/7/5	2014/3/31	64	85258	60267	375074	15	0.70801
12	58877	2010/11/18	2014/3/31	43	69056	91581	262013	22	0.988658
13	37994	2004/11/13	2014/3/31	145	92975	126821	271438	6	0.952535
14	28012	2006/11/23	2014/3/31	29	44750	53977	321529	67	0.799127
15	54943	2006/10/25	2014/3/31	118	105466	119832	179514	3	1.398382
16	57881	2010/2/1	2014/3/31	50	68941	79076	270067	2	0.921985
17	1254	2008/3/28	2014/3/31	22	69300	54764	234721	65	1.026085
18	8253	2010/7/15	2014/3/31	101	93840	93114	172231	7	1.386525
19	58899	2010/11/10	2014/3/31	40	66239	63260	284160	45	0.837844
20	26955	2006/4/6	2014/3/31	64	99735	93006	169358	2	1.401596
21	41616	2011/8/29	2014/3/31	38	60930	52316	332896	24	0.708285
22	21501	2008/7/30	2014/3/31	106	69566	122763	167113	4	1.369404
23	41281	2011/6/7	2014/3/31	23	46800	198224	214590	6	1.061631
24	47229	2005/4/10	2014/3/31	94	59169	74497	305250	23	0.741804
25	28474	2010/4/13	2014/3/31	20	64258	59600	222380	74	1.004904
26	58472	2010/2/14	2014/3/31	44	38510	75816	281837	17	0.787308
27	13942	2010/10/14	2014/3/31	62	72806	83496	243674	17	0.90299
28	45075	2007/2/1	2014/3/31	213	136769	96568	187917	3	1.146355
29	47114	2005/1/15	2014/3/31	74	101398	83139	148685	11	1.433364
30	54619	2006/1/7	2014/3/31	101	94055	107896	159129	18	1.33817

航空客户

首先将air_data数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“air_data”，勾选文件“航空客户.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行时间类型转换，将【时间类型转换】组件与输入源连接，在字段设置中选择所需的数据特征列以及选择所需要进行格式转换的特征列（若没有显示可选特征，点击右上角的刷新按钮即可），在参数设置中自定义转换格式，右键单击【时间类型转换】组件，选择“运行该节点”。

打开数据，查看结果。对【时间类型转换】组件右击，点击“查看数据”，即可查看时间列转换的结果。

8.2 统计分析

统计分析可用于业务人员或者数据分析人员通过图表分析公司业务的经营状况，发现公司经营过程中潜在的隐患，还可以通过图表挖掘其中潜在的价值。

8.2.1 主成分分析

（1） 作用及原理

主成分分析（Principal Components Analysis, PCA）是一种常用的数据分析手段，是图像处理过程中常用到的降维方法。对于一组不同维度之间可能存在线性相关关系的数据，PCA能够把这组数据通过正交变换变成各个维度之间线性无关的数据，通过剔除方差小的那些维度上的数据，达到数据降维的目的。

PCA的思想是将n维特征映射到k维上（k<n），这k维特征称为主元（主成分），是旧特征的线性组合，这些线性组合最大化样本方差，尽量使用新的k个特征互不相关。这k维是全新的正交特征，是重新构造出来的k维特征，而不是简单地从n维特征中取出其余n-k维特征。

（2） 输入

序号	条件	要求	说明
1	数据是否需要标准化	是	消除明显的量纲关系
2	数据是否允许缺失值	否	影响结果的准确性
3	数据是否需要去除重复值	是	影响结果的准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过		

（3） 输出

序号	名称	内容
1	日志	含有主成分分析后保留的信息数，得到新变量包含的信息数，保留信息数大的变量，得到降维效果

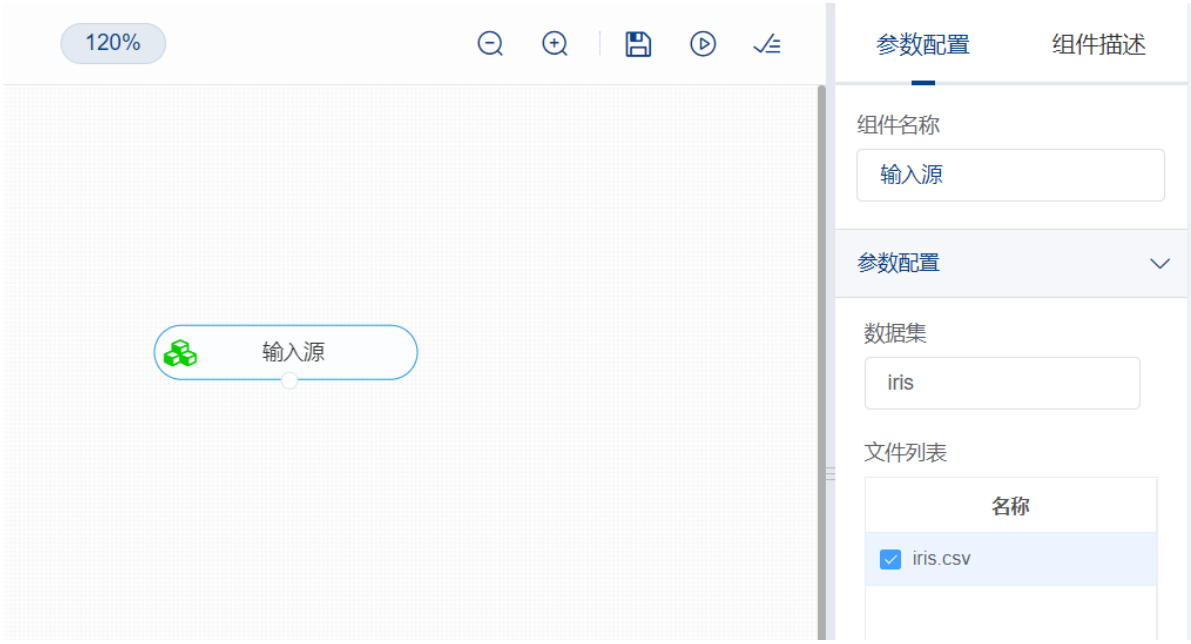
（4） 参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行主成分分析的列，数值型
2	参数设置	需要降为几维	需要降到几维，数值型

（5） 示例

对于“iris”数据集，它没有缺失值和重复值，不需要进行缺失值处理和重复值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行主成分分析。

首先将需要进行主成分分析的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



开始进行主成分分析，将该数据实施降维。拖入【主成分分析】算法，将【输入源】算法和【主成分分析】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击“参数设置”，"需要降为几维"设置为3，右键单击【主成分分析】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	需要降为几维	3	数值在1至变量个数之间

（3） 打开日志，查看结果。在日志中可以得到主成分分析后新的变量保留的信息数。对【主成分分析】算法右击，点击“查看日志”。

查看日志

[-0.58202985 0.59791083 0.07623608 0.54583143]]

解释方差比

[0.92461872 0.05306648 0.01710261]

降维后数据

	comp_1	comp_2	comp_3
0	-2.6841	0.3194	-0.0279
1	-2.7141	-0.1770	-0.2105
2	-2.8890	-0.1449	0.0179
3	-2.7453	-0.3183	0.0316
4	-2.7287	0.3268	0.0901
...	...	...	...
145	1.9441	0.1875	0.1779

序号	名称	作用
1	主成分分析后保留的信息数	可得知各个成分的重要性级别
2	方差解释比	成分对信息的解释程度

8.2.2 卡方检验

（1）作用及原理

卡方检验是一种用途很广的计数资料的假设检验方法。它属于非参数检验的范畴，主要是比较两个及两个以上样本率(构成比）以及两个分类变量的关联性分析。其根本思想就是在于比较理论频数和实际频数的吻合程度或拟合优度问题。它在分类资料统计推断中的应用，包括：两个率或两个构成比比较的卡方检验；多个率或多个构成比比较的卡方检验以及分类资料的相关分析等。

该检验的基本思想是：首先假设*H*0成立，基于此前提计算出χ2值，它表示观察值与理论值之间的偏离程度。根据χ2分布及自由度可以确定在*H*0假设成立的情况下获得当前统计量及更极端情况的概率P。如果P值很小，说明观察值与理论值偏离程度太大，应当拒绝无效假设，表示比较资料之间有显著差异；否则就不能拒绝无效假设，尚不能认为样本所代表的实际情况和理论假设有差别。

（2） 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	否	影响结果的准确性
3	数据是否需要去除重复值	是	影响结果的准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过		

（3） 输出

序号	名称	内容
1	data_out.csv	含有卡方检验的P值，可根据P值得到检验结果是否显著
2	日志	可查看卡方检验的P值

（4） 参数

序号	分组	参数	解释
1	字段设置	特征列	进行检验的数据，数值型

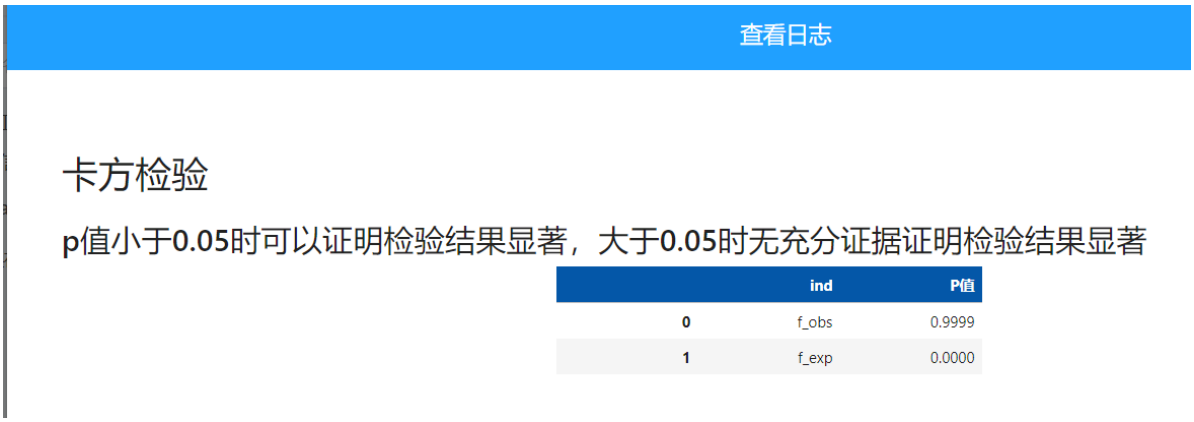
（5） 示例

数据集“Support_for_school”中没有缺失值和重复值，因此不用缺失值处理和重复值处理，因此可直接对数据集进行卡方检验算法。

首先将需要进行卡方检验的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“Support_for_school”，勾选文件“Support_for_school.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行卡方检验，得出检验结果。拖入【卡方检验】算法，将【输入源】算法和【卡方检验】算法相连接，在“字段设置”的“特征列”中勾选“f_obs”，“f_exp”字段，右键单击【卡方检验】算法，选择“运行该节点”。

打开日志，查看结果。在日志中可以查看卡方检验的P值。对【卡方检验】算法右击，点击“查看日志”。



序号	名称	作用
1	P值	p值小于0.05时可以证明检验结果显著，大于0.05时无充分证据证明检验结果显著

8.2.3 相关性分析

（1） 作用及原理

相关性分析是指对两个或多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。相关性的元素之间需要存在一定的联系或者概率才可以进行相关性分析，需要选择数值型的特征列。

- 皮尔逊相关系数：其值介于[-1,1]之间，当r值越接近0，相关度越弱（等于0，线性无关），随着r值往-1或1移动，相关度增强。

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

- kendall相关系数：是秩相关系数的一种，它可以度量两个有序变量之间单调关系强弱。
- 斯皮尔曼相关系数：是秩相关系数的一种。“秩”，即秩序，可以理解为一种顺序或排序，根据变量在数据内的位置进行计算。同时，斯皮尔曼相关系数不受离群值影响，适用于非线性数据。

$$p = 1 - \frac{6\sum d_i^2}{n^3 - n}$$ 其中，di表示顺序的差值，n表示数据个数

（2） 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	自身会将每个变量都转化为标准单位
2	数据是否允许缺失值	否	影响结果的准确性
3	数据是否需要去除重复值	是	影响结果的准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过		

（3） 输出

序号	名称	内容
1	data_out.csv	含有相关系数矩阵，可查看两变量之间的相关系数
2	日志	可查看相关系数矩阵图，可以直观地看到两变量的相关性

（4）参数

序号	分组	参数	解释
1	字段设置	特征列	进行相关性分析的数据，数值型
2	参数设置	相关性系数	考察两个变量之间的相关程度。 有“标准相关系数”，“肯德尔相关性系数”，“斯皮尔曼相关性系数”三种选择。

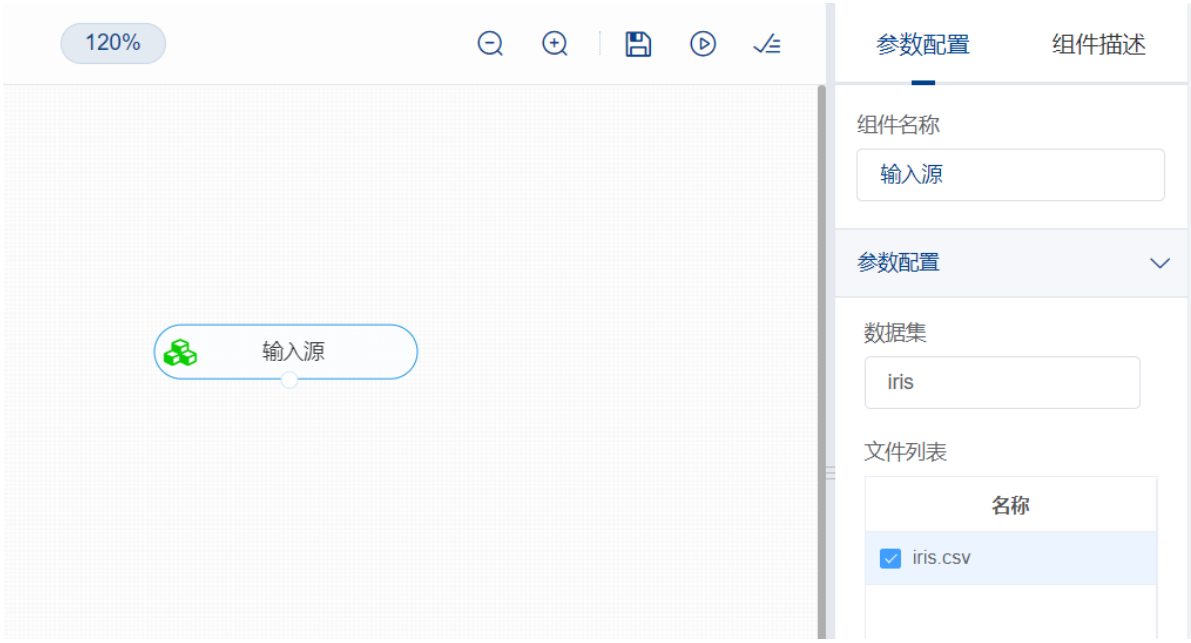
（5） 示例

数据集“iris”中没有缺失值和重复值，因此不用缺失值处理和重复值处理，因此可直接对数据集进行相关性分析算法。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行相关性分析的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行相关性分析，得出检验结果。拖入【相关性分析】算法，将【输入源】算法和【相关性分析】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，在“参数设置”的“相关性系数”设置为“斯皮尔曼相关性系数”，右键单击【相关性分析】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	相关性系数	斯皮尔曼相关性系数	目前不清楚变量的总体分布形态，且斯皮尔曼相关系数不受数据离群值影响，所以选用该相关系数

（3） 打开日志，查看结果。在日志中可以查看各变量的相关系数。对【相关性分析】算法右击，点击“查看日志”。

sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	outcome
sepal length (cm)	1.000	-0.167	0.882	0.834	0.798
sepal width (cm)	-0.167	1.000	-0.310	-0.289	-0.440
petal length (cm)	0.882	-0.310	1.000	0.938	0.935
petal width (cm)	0.834	-0.289	0.938	1.000	0.938
outcome	0.798	-0.440	0.935	0.938	1.000

序号	名称	作用
1	相关系数矩阵图	可更直观的看到变量之间的相关性

8.2.4 正态性检验

（1） 作用及原理

正态性检验主要用于判断计量数据是否服从或近似服从正态分布。因为很多常见的统计学方法都要求数据满足正态性，如常见的t检验、单因素方差分析等。在考虑采用上述方法时，要对数据进行正态性检验。如果数据明显不服从正态分布，但由于我们没有正态性检验的结果，直接使用了t检验、单因素方差分析等参数检验的方法，有可能导致统计效能下降，导致假阴性风险增加。

其原理是生成正态概率图并进行假设检验，以检查观测值是否服从正态分布。对于正态性检验，原假设为H0：数据服从正态分布;备择假设H1：数据不服从正态分布。

（2） 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	否	影响结果的准确性
3	数据是否需要去除重复值	是	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

（3） 输出

序号	名称	内容
1	data_out.csv	正态性检验的P值，据此可以得出数据是否服从正态分布
2	日志	正态性检验的P值，据此可以得出数据是否服从正态分布

（4） 参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行正态性检验的列，数值型
2	参数设置	缺失值处理方式	数据中有缺失值会导致会影响模型的准确性，因此需要处理缺失值； 此处有三种方式可选择，分别为 “结果返回nan值”，”引发错误“，”忽略nan值“

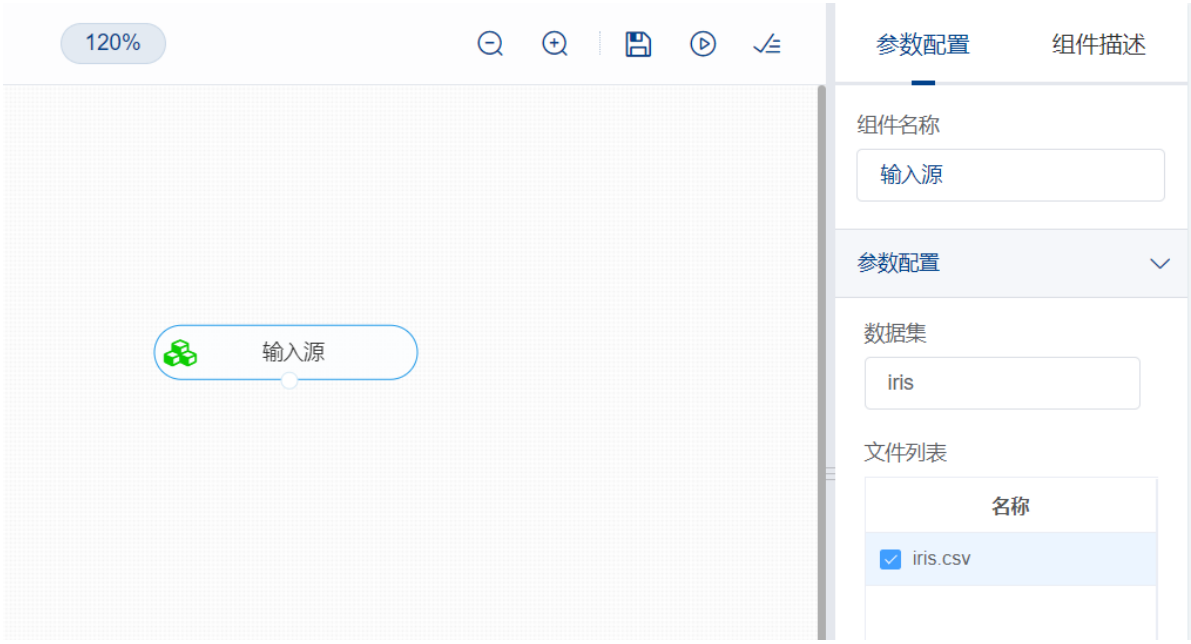
（5） 示例

数据集“iris”中没有缺失值和重复值，因此不用缺失值处理和重复值处理且数据集无明显的量纲差异，无需进行数据标准化。因此可直接对数据集进行正态性检验算法。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行正态性检验的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



开始进行正态性检验，检验数据的正态性。拖入【正态性检验】算法，将【输入源】算法和【正态性检验】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击“参数设置”，”缺失值处理方式“设置为“结果返回nan值”，右键单击【正态性检验】算法，选择“运行该节点”。

序号	参数名称	值	原因
1	缺失值处理方式	结果返回nan值	当特征列中有缺失时，会影响模型的准确性

（3） 打开日志，查看结果。在日志中可以查看正态性检验的P值。对【正态性检验】算法右击，点击“查看日志”。

查看日志

正态性检验

检验结果，当p<0.05时，可以证明数据不服从正态分布

	ind	p值
0	sepal length (cm)	0.0568
1	sepal width (cm)	0.2097
2	petal length (cm)	0.0000
3	petal width (cm)	0.0000

序号	名称	作用
1	P值	根据检验结果，可以得出数据是否服从正态分布

8.2.5 方差齐性检验

1 作用及原理

方差齐性检验(test for homogeneity of variance)，假设检验的一种。关于两个或两个以上总体的方差是否相等的统计检验。根据情况不同，有不同的检验方法。方差齐性检验是方差分析的重要前提，是方差可加性原则应用的一个条件。方差齐性检验是对两样本方差是否相同进行的检验。

其基本原理是先对总体的特征作出某种假设，然后通过抽样研究的统计推理，对此假设应该被拒绝还是接受作出推断。常用方法有:Hartley检验、Bartlett检验、修正的Bartlett检验。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	否	
3	数据是否需要去除重复值	是	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

3 输出

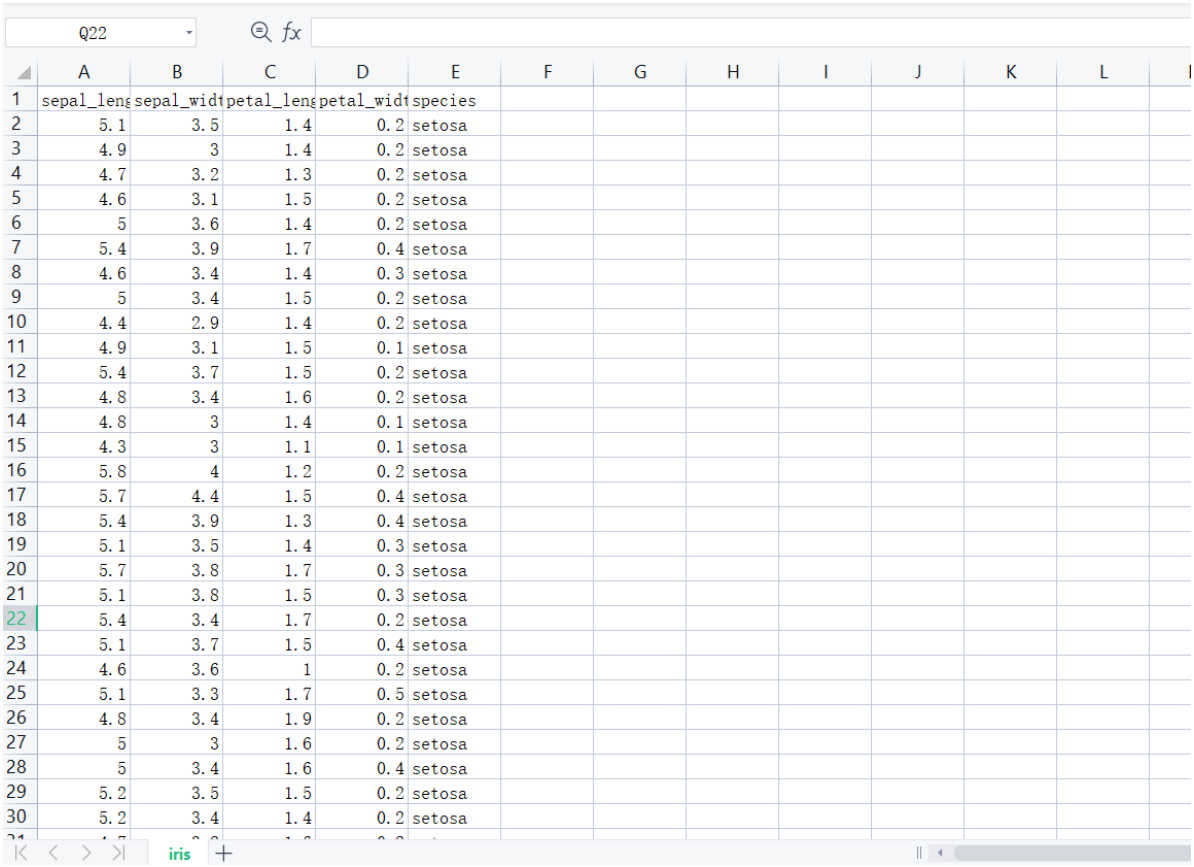
序号	名称	内容
1	日志	根据检验结果，查看两组的方差是否显著性差异

4 参数

序号	分组	参数	解释
1	字段设置	选择待检验特征列	需要进行方差齐性检验的列，数值型
2	参数设置	中心选择	指定离差的计算方式是以组内均值为中心还是以组内中位数为中心。

5 示例

数据集“iris”中没有缺失值和重复值，因此不用缺失值处理和重复值处理且数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接对数据集进行正态性检验算法。

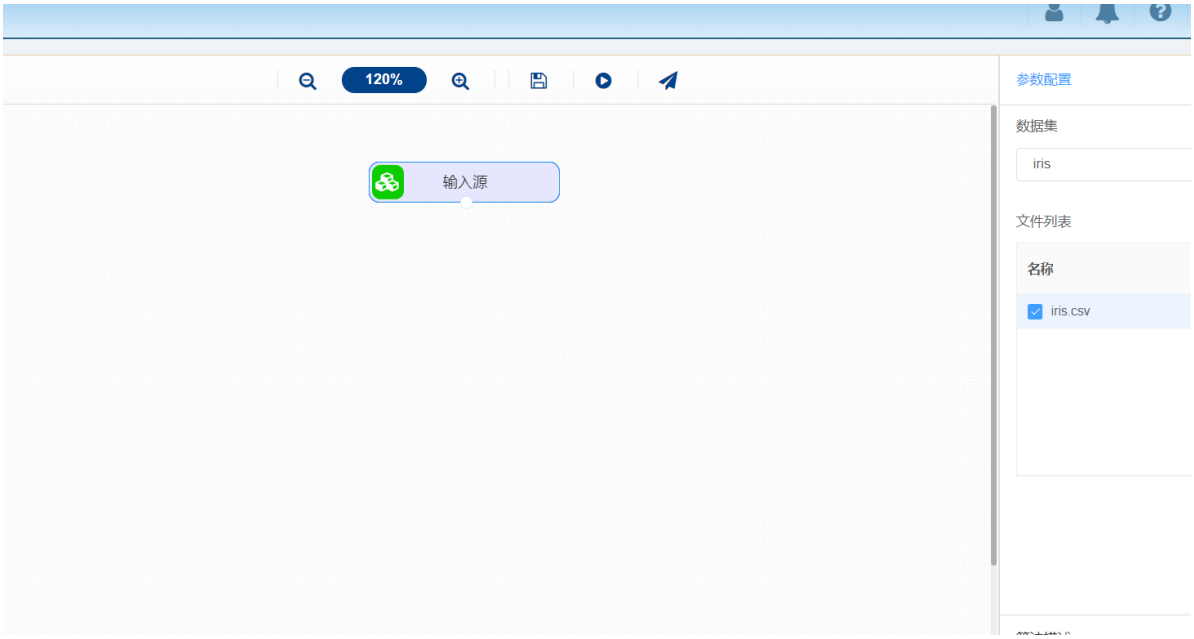


Q22

	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

iris

（1） 首先将需要进行方差齐性检验的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行方差齐性检验，检验总体的方差是否相等。拖入【方差齐性检验】算法，将【输入源】算法和【正态性检验】算法相连接，在“字段设置”的“选择待检验特征1”中选择“sepal_width”字段，“选择待检验特征2”中选择“sepal_length”字段，点击“参数设置”，”中心选择“设置为“中位数”，右键单击【方差齐性检验】算法，选择“运行该节点”。

序号	参数名称	值	原因
1	中心选择	中位数	倾斜非正态分布建议用中位数

（3） 打开日志，查看结果。在日志中可以查看正态性检验的P值。对【方差齐性检验】算法右击，点击“查看日志”。

序号	名称	作用
1	P值	根据检验结果，查看两组的方差是否显著性差异

8.2.6 因子分析

（1） 作用及原理

因子分析法是指从研究指标相关矩阵内部的依赖关系出发，把一些信息重叠、具有错综复杂关系的变量归结为少数几个不相关的综合因子的一种多元统计分析方法。因子分析是基于降维的思想，它可在许多变量中找出隐藏的具有代表性的因子。将相同本质的变量归入一个因子，可减少变量的数目，同时又可反映变量之间的内在联系。

基本思想是根据各个特征列的相关性大小把特征分组，使得同组内的特征之间相关性较高，但不同组的特征变量不相关或相关性较低，每组特征变量代表一个基本结构——公共因子。

（2） 输入

序号	条件	要求	说明
1	数据是否需要标准化	是	消除明显的量纲关系
2	数据是否允许缺失值	否	
3	数据是否需要去除重复值	是	
4	载入文件格式	CSV格式	

（3） 输出

序号	名称	内容
1	data_out.csv	因子得分系数矩阵
2	日志	含有因子得分系数矩阵

（4） 参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行因子分析的列，数值型
2	参数设置	形成多少个因子	形成几个不相关的综合因子，数值型
3	参数设置	迭代次数	迭代的次数，数值型

（5） 示例

数据集“iris”中没有缺失值和重复值，因此不用缺失值处理和重复值处理且数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行正态性检验算法。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

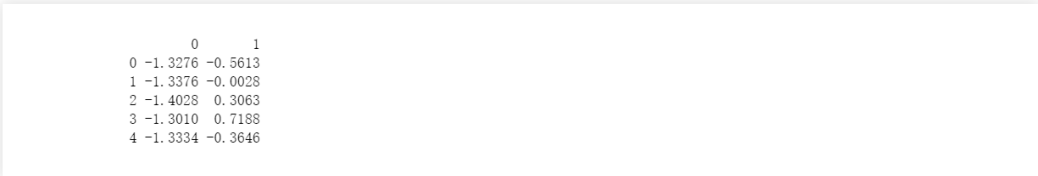
首先将需要进行因子分析的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行因子分析，从变量群中提取共性因子。拖入【因子分析】算法，将【输入源】算法和【因子分析】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击“参数设置”，”形成多少个因子“设置为“2”，”迭代次数“设置为“1000”，右键单击【因子分析】算法，选择“运行该节点”。

序号	参数名称	值	原因
1	形成多少个因子	2	原数据中记录的为花萼与花瓣的数据，尝试采用因子分析进行数据降维
2	迭代次数	1000	迭代次数越多，结果越准确。但到一定次数，基本就没效果了，还可能会产生过拟合

打开数据，查看结果。在数据中可以查看因子得分系数。对【因子分析】算法右击，点击“查看日志”。

查看日志



	0	1
0	-1.3276	-0.5613
1	-1.3376	-0.0028
2	-1.4028	0.3063
3	-1.3010	0.7188
4	-1.3334	-0.3646

序号	名称	作用
1	因子得分系数	含有因子得分系数矩阵，有了因子得分值，则可以在许多分析中使用这些因子 例如以因子的得分做聚类分析的变量，做回归分析中的回归因子。

8.2.7 频数统计

（1）作用

频数是指某个数据表中某个分类或某个数值字符出现的次数。该组件则是将数据表中某一特征的特征值出现情况进行次数统计。

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	是	
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	

（3）输出

序号	名称	内容
1	data_out.csv	含有频数的数据，count为频数
2	日志	含有频数的数据，count为频数

（4） 参数

序号	分组	参数	解释
1	字段设置	需要进行频数统计的特征列	需要进行频数统计的数据

（5）示例

数据集“iris”中没有缺失值和重复值，因此不用缺失值处理和重复值处理，可直接对数据集“iris”进行频数统计。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行因子分析的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行频数统计，计算各个变量的频数。拖入【频数统计】算法，将【输入源】算法和【频数统计】算法相连接，在“字段设置”的“需要进行频数统计的特征”中选择“outcome”字段，右键单击【频数统计】算法，选择“运行该节点”。

打开日志，查看结果。在日志中可以查看因子得分系数矩阵。对【频数统计】算法右击，点击“查看日志”。

	outcome	count
0	0	50
1	1	50
2	2	50

8.2.8 全表统计

（1） 作用

全表统计用于对统计全表，或某些选中的列进行统计信息分析。

（2） 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	是	
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

（3） 输出

序号	名称	内容
1	data_out.csv	含有变量的均值，标准差等的数据

（4） 参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行全表统计的数据，任何值
2	参数设置	分位点	当前估值在历史中处于什么样的位置，默认值为0.5

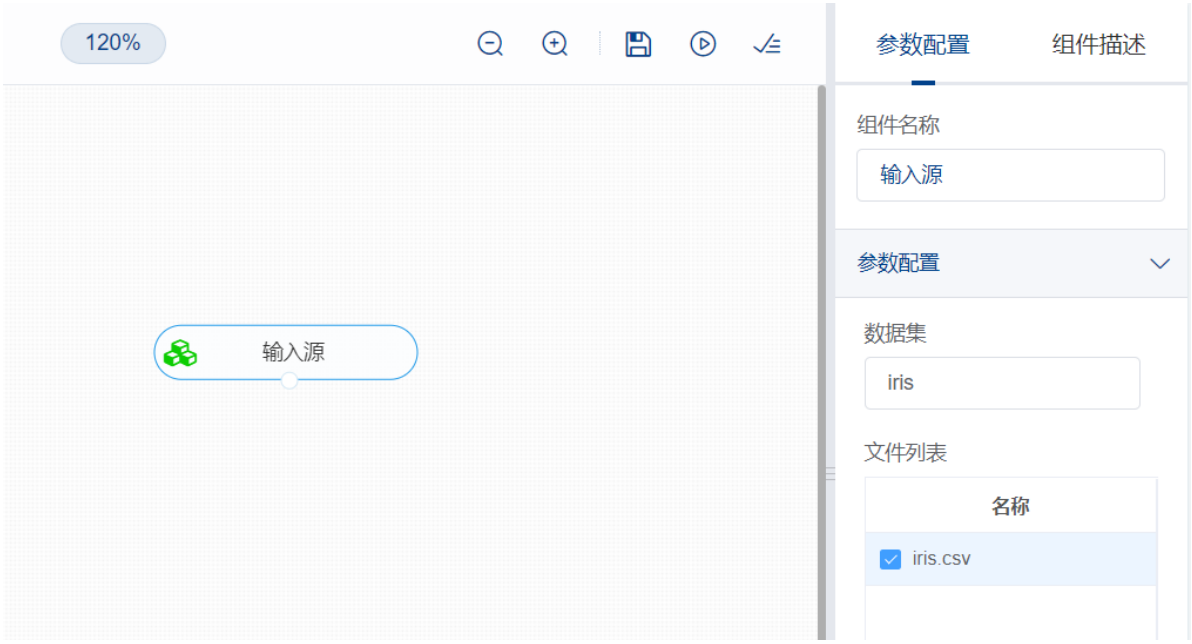
（5） 示例

数据集“iris”中没有缺失值和重复值，因此不用缺失值处理和重复值处理，可直接对数据集“iris”进行频数统计。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行因子分析的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



开始进行全表统计，计算各个变量的频数。拖入【全表统计】算法，将【输入源】算法和【全表统计】算法相连接，在“字段设置”的“特征”中勾选所有字段，右键单击【全表统计】算法，选择“运行该节点”。

序号	参数名称	值	原因
1	分位点	0.4	得到40%分位数以及中位数

8.2.9 时间聚合计算

1 作用

对随着时间变化的数据进行聚合计算。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	是	
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

3 输出

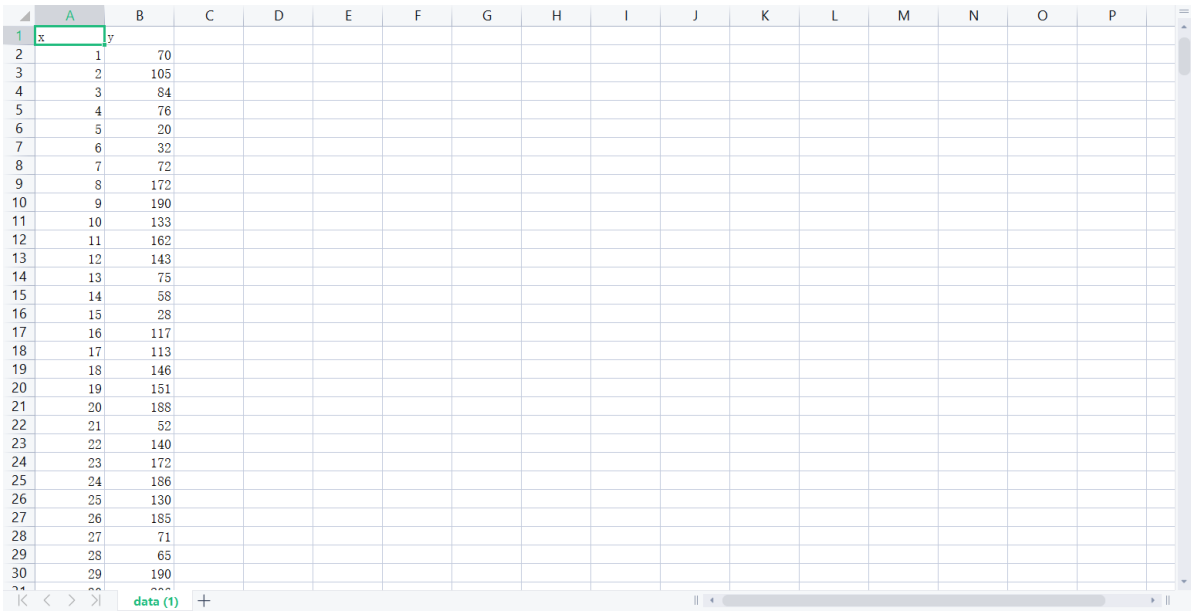
序号	名称	内容
1	data_out.csv	输出相应的计算方式结果，例如求和，中位数

4 参数

序号	分组	参数	解释
1	时间聚合计算	时间列	时间数据。数值型
2	时间聚合计算	数据值	随着时间变化的数据。数值型
3	时间聚合计算	计算方式	有9个选项，例如求和，均值，中位数等
4	时间聚合计算	时间频率单位	有9个选项，例如小时，分钟，秒等
5	时间聚合计算	结果保留小数位	结果保留的小数位数，数值型

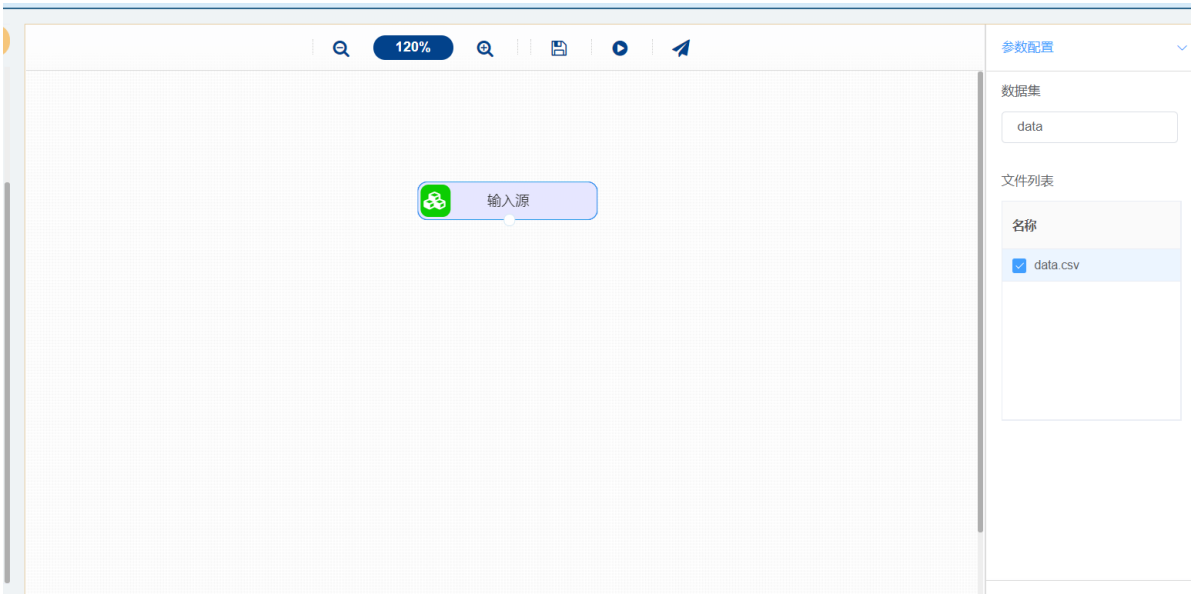
5 示例

数据集“data”中没有缺失值和重复值，因此不用缺失值处理和重复值处理，可直接对数据集“data”进行频数统计。



	A	B
1	x	y
2	1	70
3	2	105
4	3	84
5	4	76
6	5	20
7	6	32
8	7	72
9	8	172
10	9	190
11	10	133
12	11	162
13	12	143
14	13	75
15	14	58
16	15	28
17	16	117
18	17	113
19	18	146
20	19	151
21	20	188
22	21	52
23	22	140
24	23	172
25	24	186
26	25	130
27	26	185
28	27	71
29	28	65
30	29	190

（1） 首先将需要进行因子分析的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“data”，勾选文件“data.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行时间聚合计算，根据相应的计算方式得到想要的计算结果。拖入【时间聚合计算】算法，将【输入源】算法和【时间聚合计算】算法相连接，在“时间聚合计算”的“选择时间列”中选择“x”字段，“选择数值列”中选择“y”字段，右键单击【时间聚合计算】算法，选择“运行该节点”。

8.2.10 移动计算

1 作用

对某一列的固定的前几个数据进行相应的计算方式。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	否	计算过程中含有缺失值的，结果也空缺
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

3 输出

序号	名称	内容
1	data_out.csv	含有相应计算方式得出结果的原始数据

4 参数

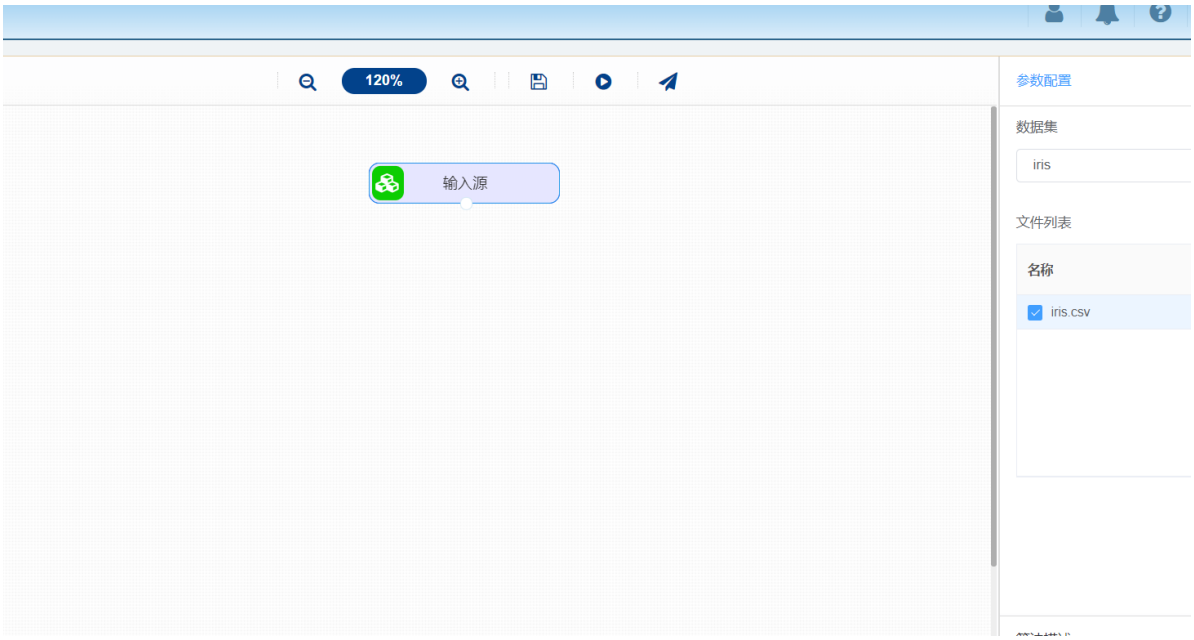
序号	分组	参数	解释
1	移动计算	数值列	进行移动计算的数据列。数值型
2	移动计算	移动窗口大小	需要进行计算的数据个数。数值型
3	移动计算	计算方式	有9个选项，例如求和，平均值，中位数等
4	移动计算	输出形式	有“单独输出”，“与原数据合并”2个选项
5	移动计算	结果保留小数位	结果保留的小数位数，数值型

5 示例

数据集“iris”中没有缺失值，因此不用缺失值处理，可直接对数据集“iris”进行频数统计。

	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

（1） 首先将需要进行移动计算的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行移动计算，对固定的前几个数据进行相应的计算方式。拖入【移动计算】算法，将【输入源】算法和【移动计算】算法相连接，在“移动计算”的“选择数值列”中选择“petal_length”字段，右键单击【移动计算】算法，选择“运行该节点”。

8.2.11 var方差函数

1 作用

函数 VAR 假设其参数是样本总体中的一个样本。如果数据为样本总体，则应使用函数 VARP 来计算方差。其用途是计算基于给定样本的方差。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	是	
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

3 输出

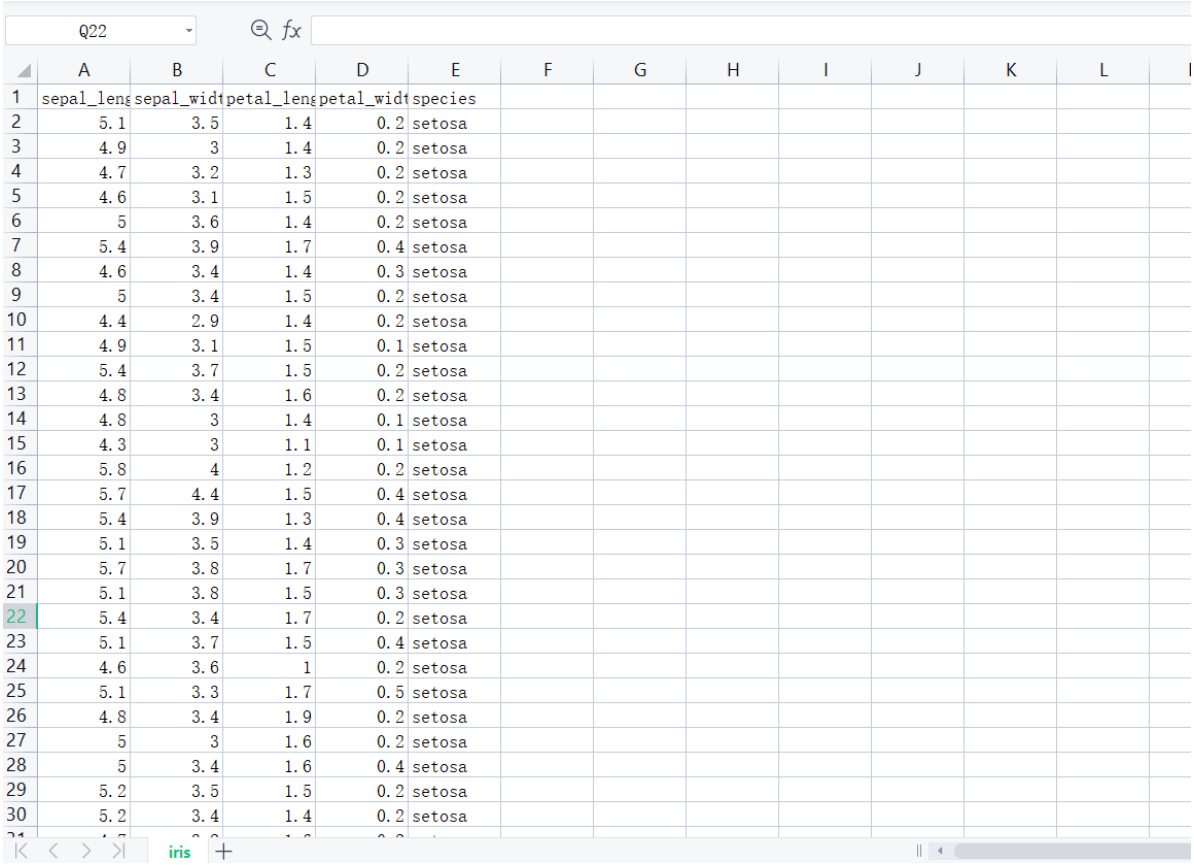
序号	名称	内容
1	data_out.csv	样本的方差

4 参数

序号	分组	参数	解释
1	字段设置	字段列	进行计算方差的数据列。数值型

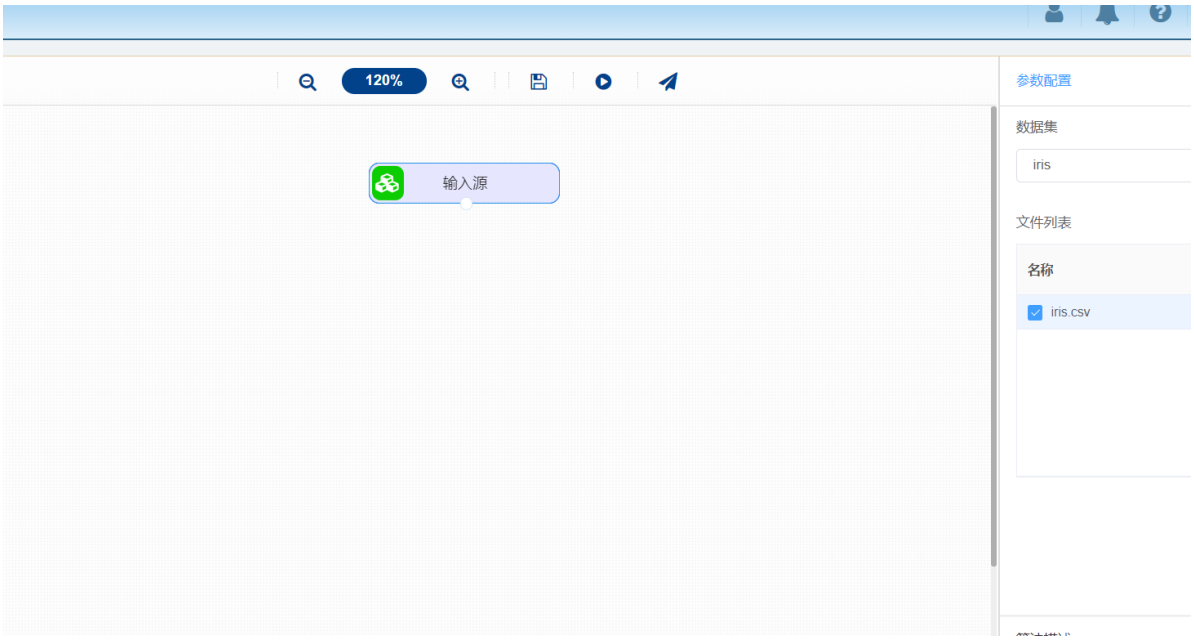
5 示例

数据集“iris”中没有缺失值，因此不用缺失值处理，可直接对数据集“iris”进行var方差函数算法。



	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

（1） 首先将需要进行var方差函数算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行var方差函数算法，计算基于给定样本的方差。拖入【var方差函数】算法，将【输入源】算法和【var方差函数】算法相连接，在“字段设置”的“字段”中选择“petal_length”字段，右键单击【var方差函数】算法，选择“运行该节点”。

8.2.12 分布函数

1 作用

分布函数(distribution function)定义：设X是一个随机变量，x是任意实数，函数F(x)=PX≤x称为X的分布函数。有时也记为X~F(x)。分布函数是一个普遍的函数，正是通过它，我们将能用数学分析的方法来研究随机变量。如果将X看成是数轴上的随机点的坐标，那么，分布函数F(x)在x处的函数值就表示X落在区间(-∞,x）上的概率。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	是	
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

3 输出

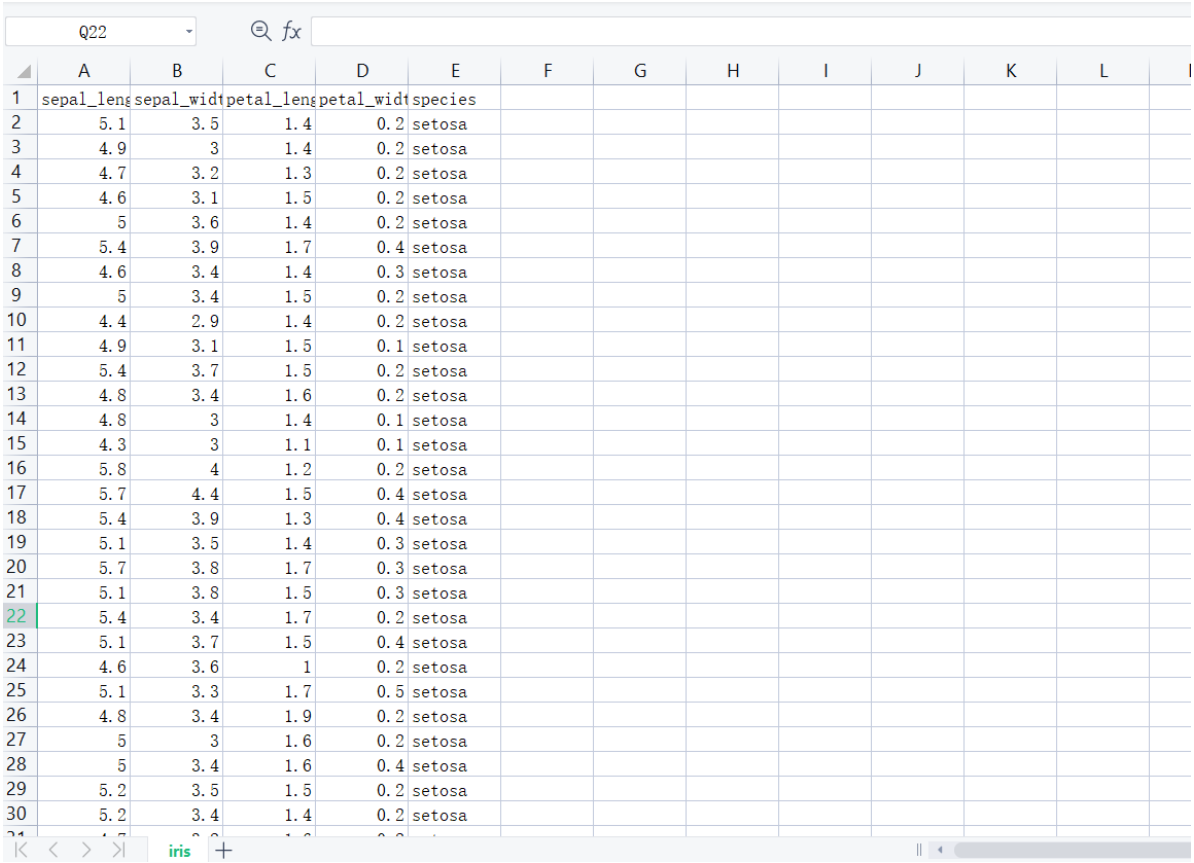
序号	名称	内容
1	data_out.csv	含有变量的个数和百分比，可画频数分布直方图，清楚显示各组频数分布情况又易于显示各组之间频数的差别
2	日志	含有变量的个数和百分比

4 参数

序号	分组	参数	解释
1	字段设置	特征列	进行分布函数算法的数据列。数值型或字符型

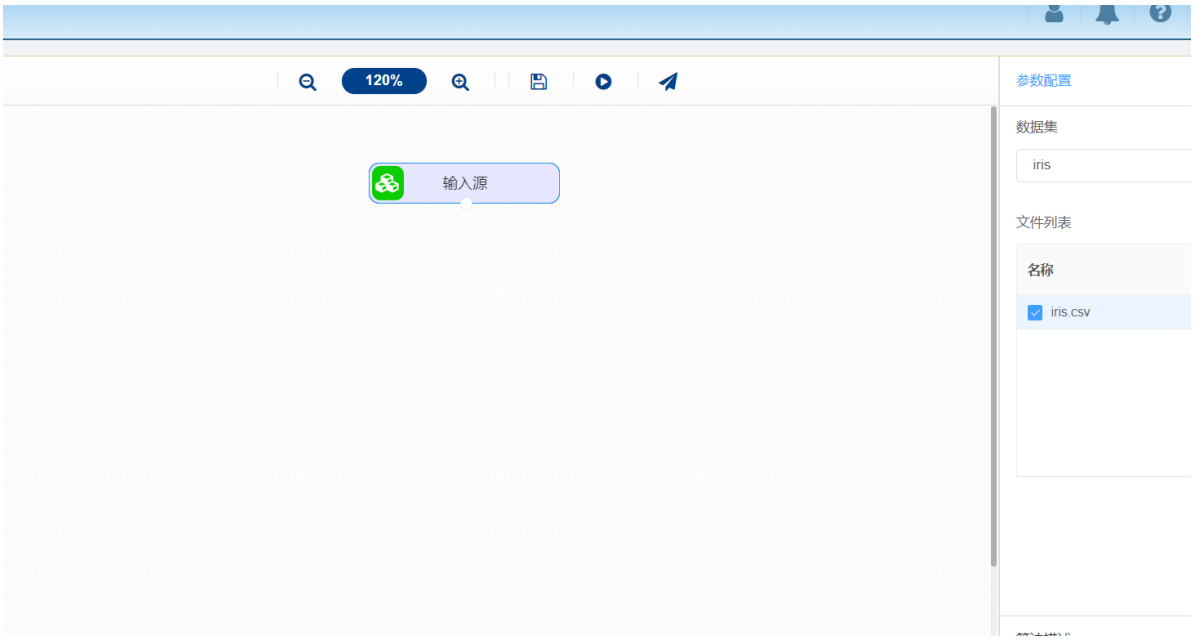
5 示例

对数据集“iris”进行分布函数算法。

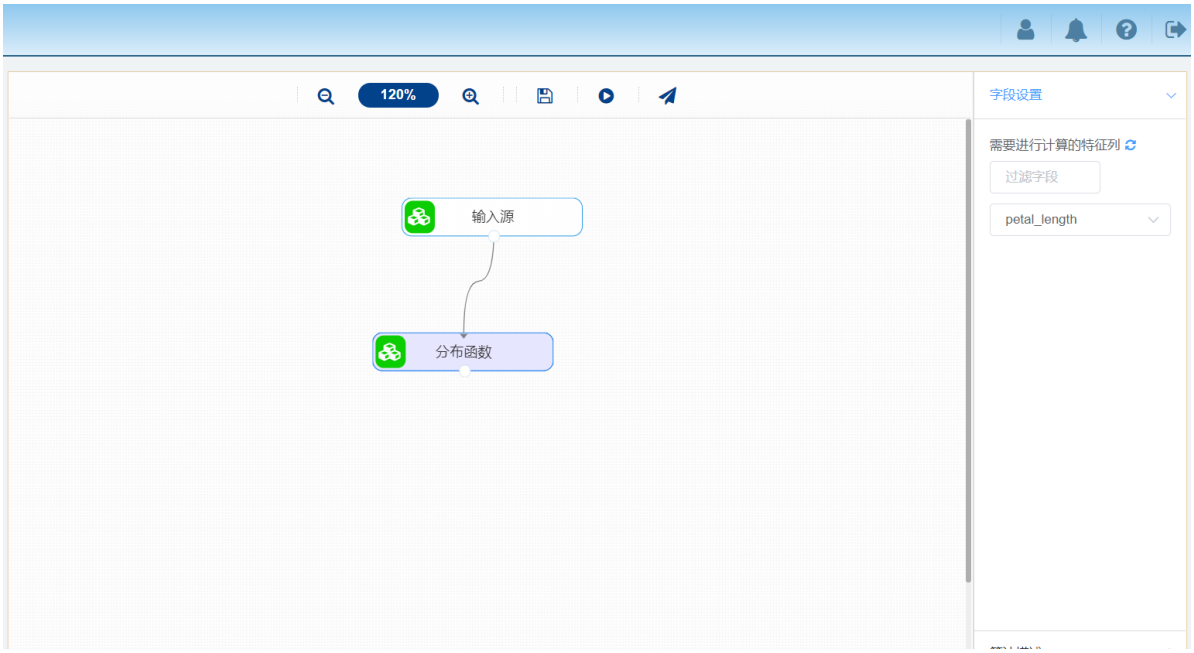


	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

（1） 首先将需要进行分布函数算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行分布函数算法，计算变量的个数和百分比。拖入【分布函数】算法，将【输入源】算法和【分布函数】算法相连接，在“字段设置”的“字段需要进行计算的特征列”中选择“petal_length”字段，右键单击【分布函数】算法，选择“运行该节点”。



（3） 打开日志，查看结果。在日志中可以查看计算变量的个数和百分比。对【分布函数】算法右击，点击“查看日志”。

8.2.13 累计计算

1 作用

连以前的数目合并计算。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	否	含有缺失值的行，没有累加结果
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

3 输出

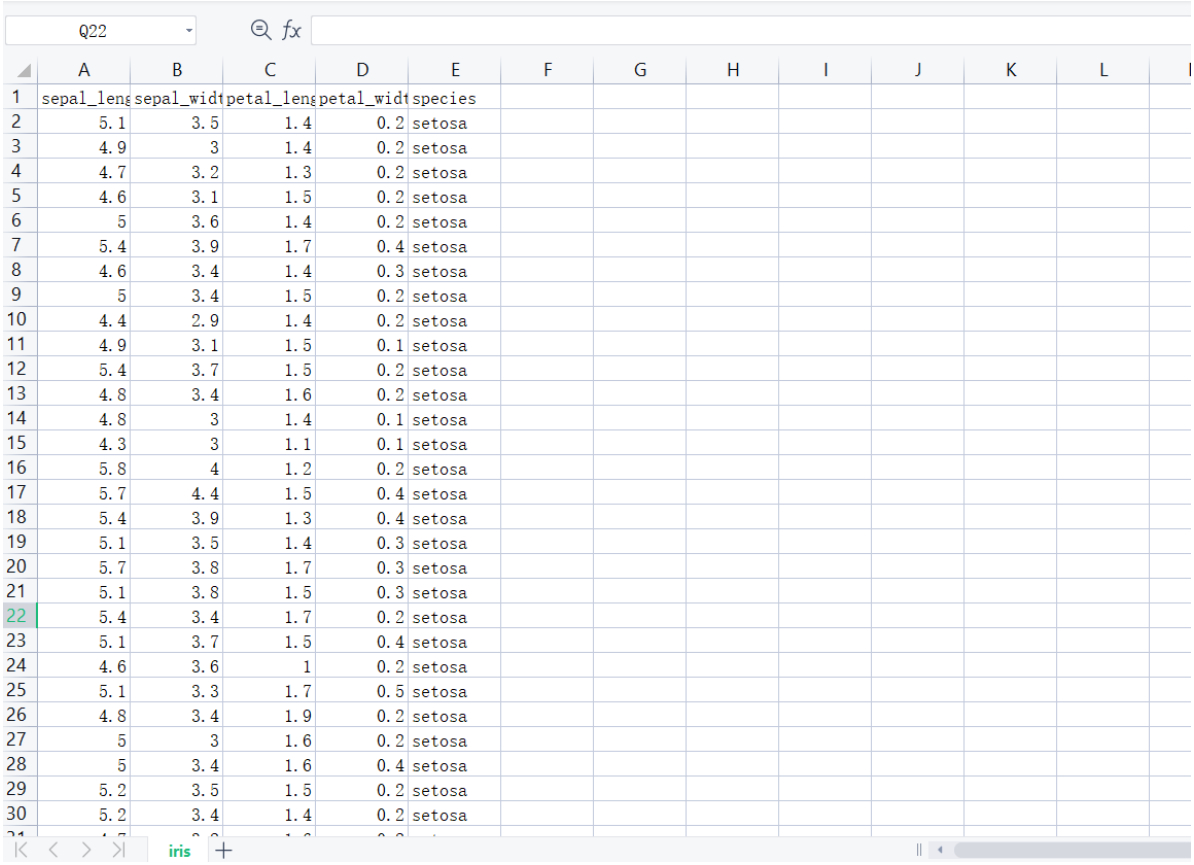
序号	名称	内容
1	data_out.csv	含有累加数据的原始数据，其中sepal_length_累加是累加数据。

4 参数

序号	分组	参数	解释
1	累计计算	数值列	进行累计计算算法的数据列。数值型
2	累计计算	计算方式	有“累加”，“累乘”，“累积最大值”，“累积最小值”选项
3	累计计算	输出形式	有“单独输出”，“与原数据合并”选项
4	累计计算	结果保留小数位	结果保留的小数位。数值型

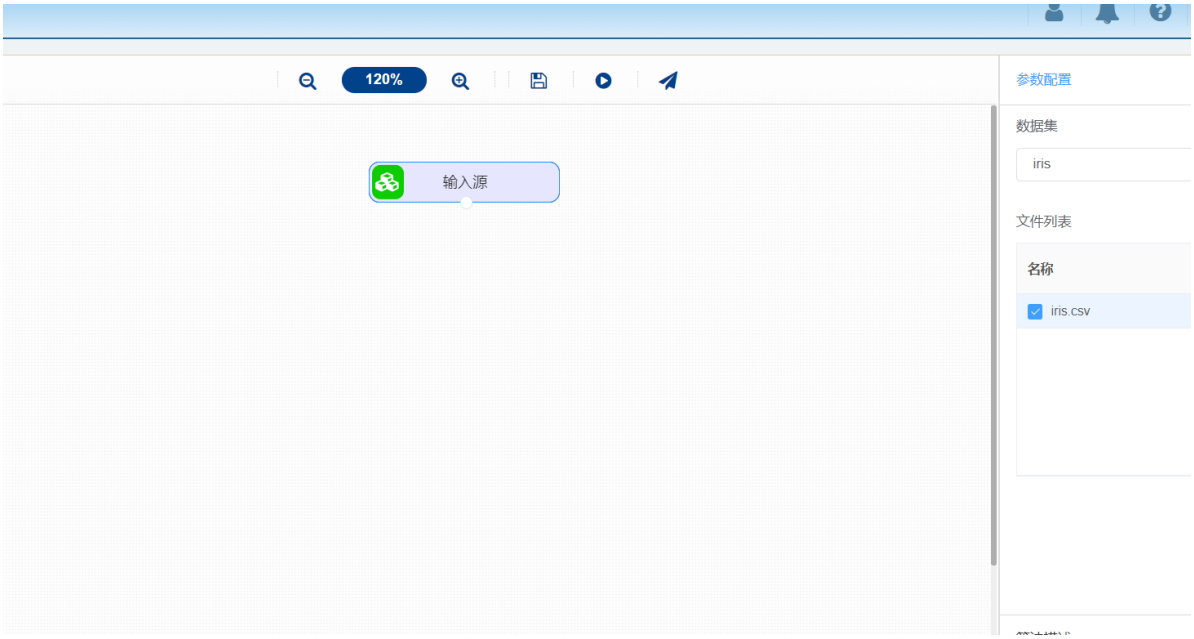
5 示例

对数据集“iris”进行累计计算算法。



	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

（1） 首先将需要进行累计计算算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行累计计算算法，得到累计计算结果。拖入【累计计算】算法，将【输入源】算法和【累计计算】算法相连接，在“累计计算”的“选择数值列”中选择“petal_length”字段，右键单击【累计计算】算法，选择“运行该节点”。

8.2.14 LASSO回归

1 作用及原理

LASSO是由1996年Robert Tibshirani首次提出，全称Least absolute shrinkage and selection operator。该方法是一种压缩估计。它通过构造一个惩罚函数得到一个较为精炼的模型，使得它压缩一些系数，同时设定一些系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。

Lasso回归使得一些系数变小，甚至还是一些绝对值较小的系数直接变为0，因此特别适用于参数数目缩减与参数的选择，因而用来估计稀疏参数的线性模型。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	是	消除明显的量纲关系
2	数据是否允许缺失值	否	影响结果的准确性
3	数据是否需要去除重复值	是	影响结果的准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过		

3 输出

序号	名称	内容
1	data_out.csv	与标签的相关性的特征列以及标签列的原始数据
2	日志	特征与标签的相关性

4 参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行 LASSO回归的列，数值型
2	字段设置	标签列	需要输出与其他数据列相关性的列，数值型
3	参数设置	L1项系数	正则项系数，数值越大，则对复杂模型的惩罚力度越大。数值型
4	参数设置	最大迭代次数	部分求解器需要通过迭代实现，这个参数指定了模型优化的最大迭代次数。数值型

5 示例

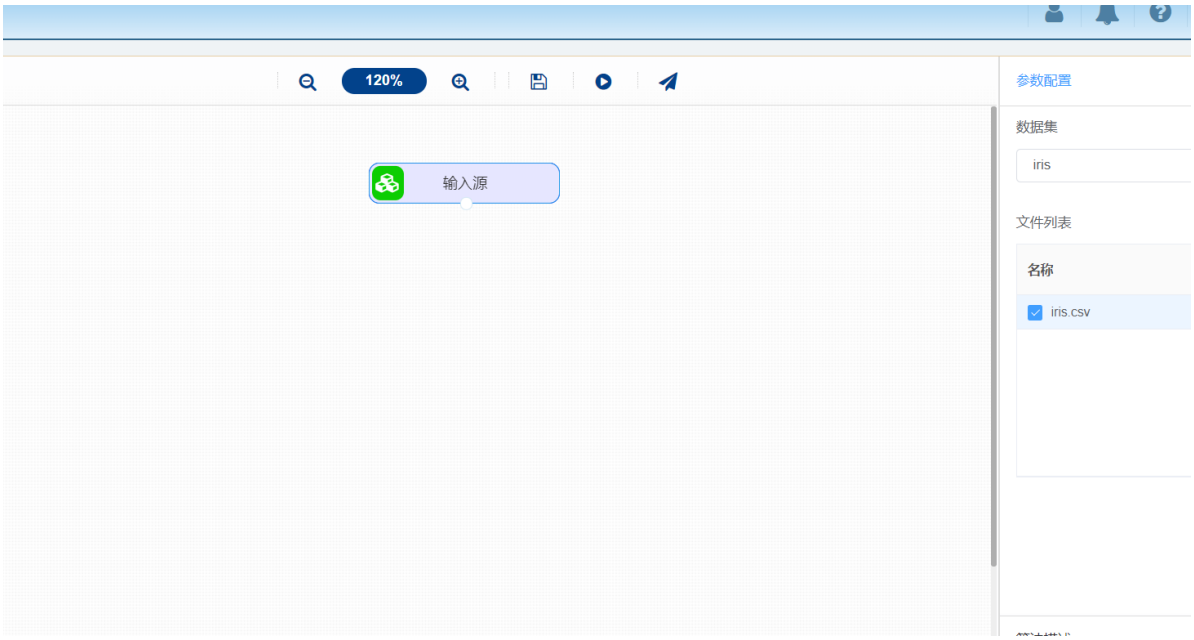
数据集“iris”中没有缺失值和重复值，因此不用缺失值处理和重复值处理且数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行正态性检验算法。

Q22

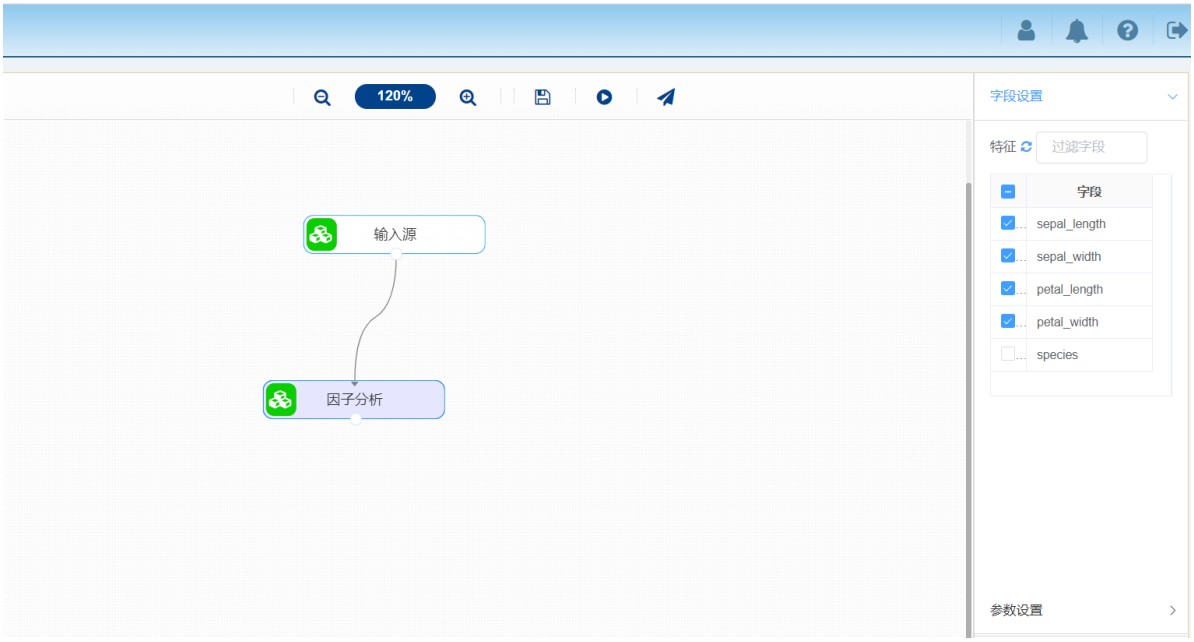
	A	B	C	D	E	F	G	H	I	J	K	L
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species							
2	5.1	3.5	1.4	0.2	setosa							
3	4.9	3	1.4	0.2	setosa							
4	4.7	3.2	1.3	0.2	setosa							
5	4.6	3.1	1.5	0.2	setosa							
6	5	3.6	1.4	0.2	setosa							
7	5.4	3.9	1.7	0.4	setosa							
8	4.6	3.4	1.4	0.3	setosa							
9	5	3.4	1.5	0.2	setosa							
10	4.4	2.9	1.4	0.2	setosa							
11	4.9	3.1	1.5	0.1	setosa							
12	5.4	3.7	1.5	0.2	setosa							
13	4.8	3.4	1.6	0.2	setosa							
14	4.8	3	1.4	0.1	setosa							
15	4.3	3	1.1	0.1	setosa							
16	5.8	4	1.2	0.2	setosa							
17	5.7	4.4	1.5	0.4	setosa							
18	5.4	3.9	1.3	0.4	setosa							
19	5.1	3.5	1.4	0.3	setosa							
20	5.7	3.8	1.7	0.3	setosa							
21	5.1	3.8	1.5	0.3	setosa							
22	5.4	3.4	1.7	0.2	setosa							
23	5.1	3.7	1.5	0.4	setosa							
24	4.6	3.6	1	0.2	setosa							
25	5.1	3.3	1.7	0.5	setosa							
26	4.8	3.4	1.9	0.2	setosa							
27	5	3	1.6	0.2	setosa							
28	5	3.4	1.6	0.4	setosa							
29	5.2	3.5	1.5	0.2	setosa							
30	5.2	3.4	1.4	0.2	setosa							

iris

（1） 首先将需要进行LASSO回归的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行LASSO回归，得到标签列与特征列的相关性。拖入【LASSO回归】算法，将【输入源】算法和【LASSO回归】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”字段，“标签”中勾选“petal_width”字段，点击“参数设置”，”L1项系数“设置为“1 ”，”迭代次数“设置为“1000”，右键单击【LASSO回归】算法，选择“运行该节点”。



序号	参数名称	值	原因
1	L1项系数	1	
2	迭代次数	1000	迭代次数越多，结果越准确。但到一定次数，基本就没效果了，还可能会产生过拟合

（3） 打开日志，查看结果。在日志中可以查看因子得分系数矩阵。对【LASSO回归】算法右击，点击“查看日志”。

特征与标签的相关性

	列名	相关性
0	sepal_length	0.000000
1	sepal_width	-0.000000
2	petal_length	0.092706

序号	名称	作用
1	相关性	两个变量的关联程度。

8.2.15 对比校验

1 作用

同行对比是否有相同值。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	
2	数据是否允许缺失值	是	
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

3 输出

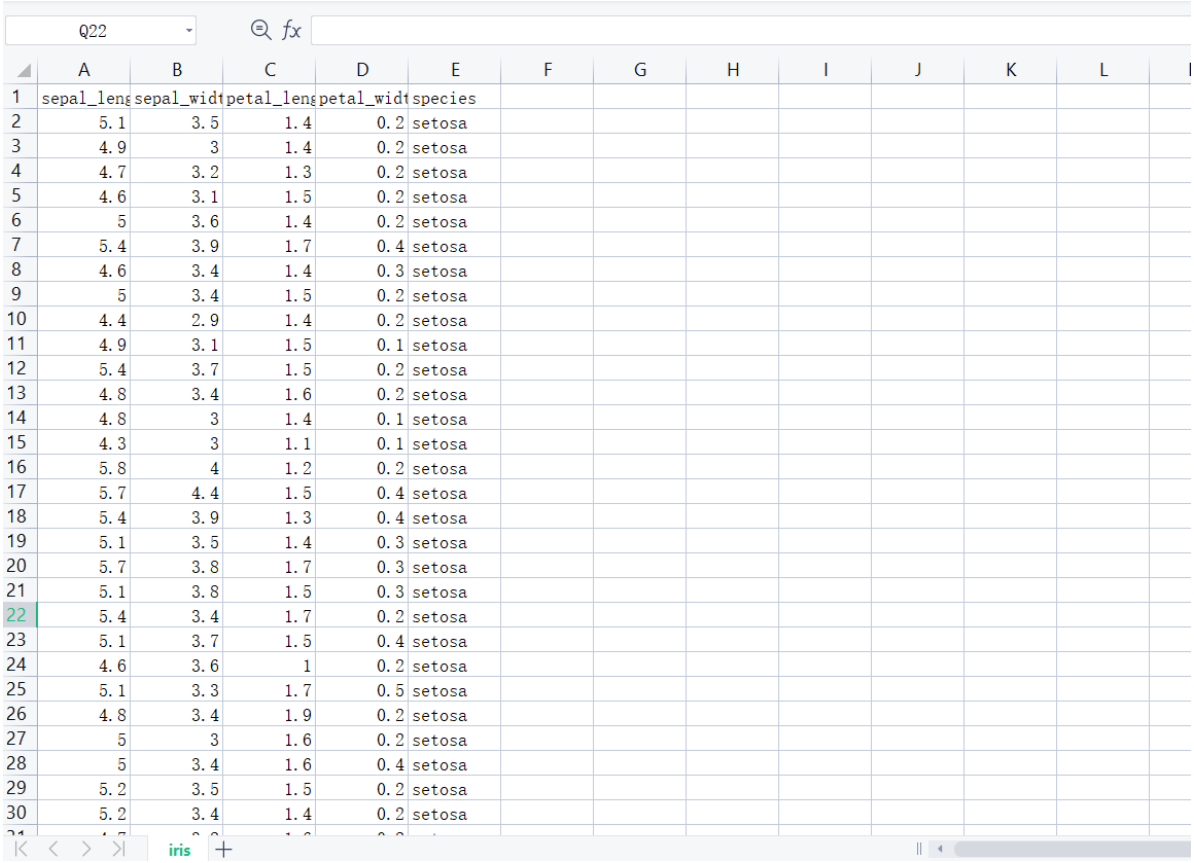
序号	名称	内容
1	data_out.csv	含有校验结果的原始数据
2	日志	含有校验结果

4 参数

序号	分组	参数	解释
1	字段设置	校验特征列	需要进行对比校验的列，数值型

5 示例

数据集“iris”中没有重复值，因此不用重复值处理。因此可直接对数据集进行对比校验算法。

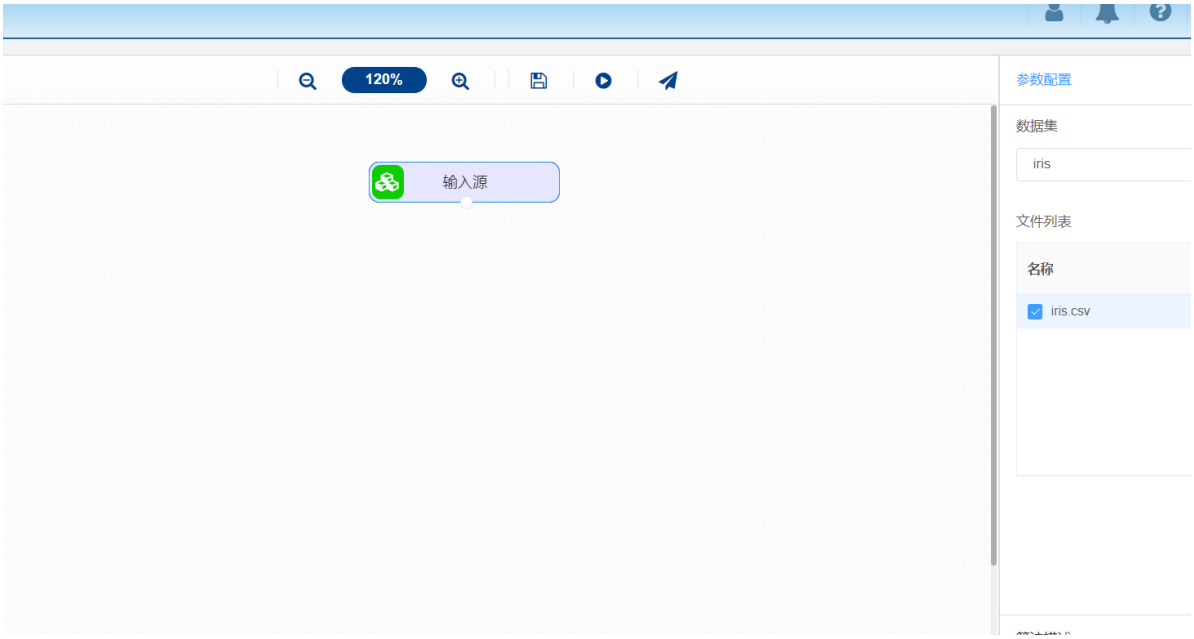


Q22

	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

iris

（1） 首先将需要进行对比校验的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行对比校验，对比同一行是否有相同数据。拖入【对比校验】算法，将【输入源】算法和【对比校验】算法相连接，在“字段设置”的“校验特征1”中勾选“sepal_length”字段，“校验特征2”中勾选“sepal_width”字段，右键单击【对比校验】算法，选择“运行该节点”。

（3） 打开日志，查看结果。在日志中可以查看对比校验的结果。对【对比校验】算法右击，点击“查看日志”。

8.2.16 时序检验

1 作用

对时间序列检验其是否平稳或者是否为白噪声序列。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	是	消除明显的量纲关系
2	数据是否允许缺失值	是	
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议不超过		

3 输出

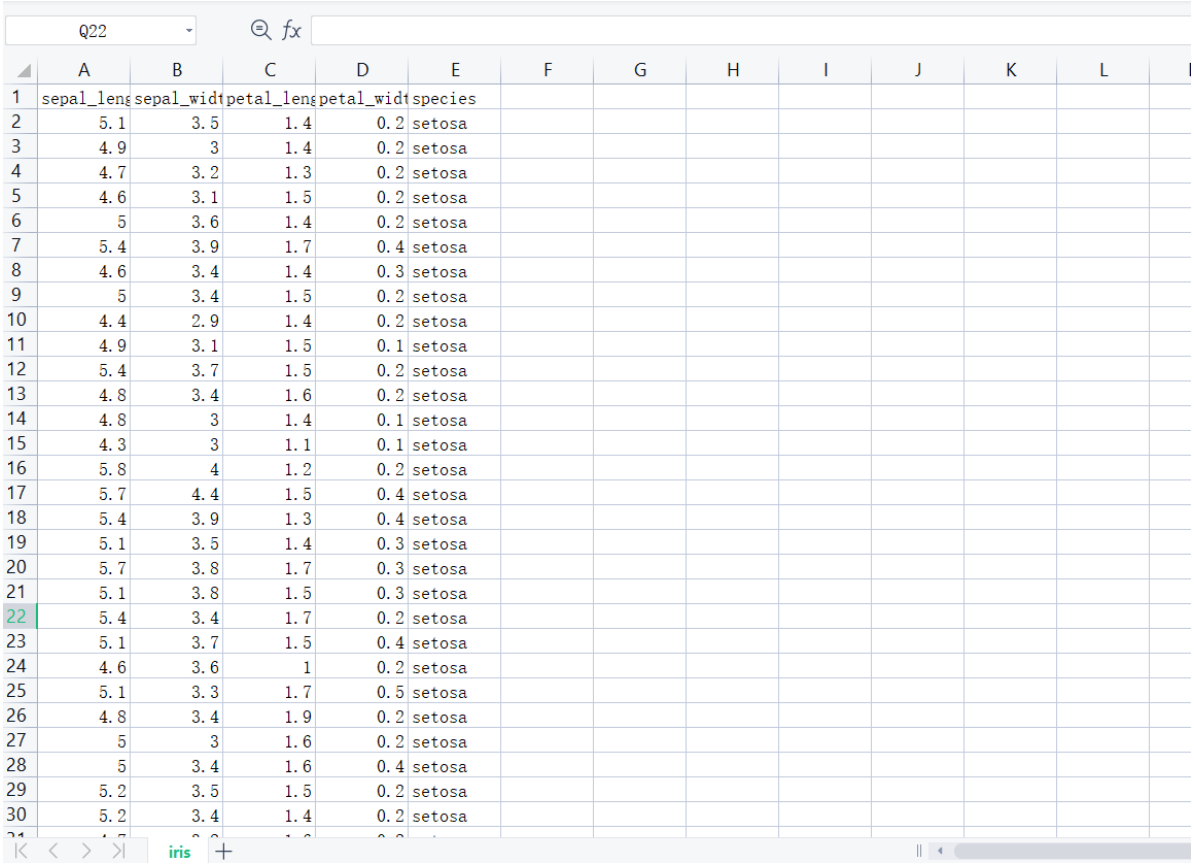
序号	名称	内容
1	日志	得到平稳性检验结果

4 参数

序号	分组	参数	解释
1	字段设置	检验列	进行检验的列，数值型
2	参数设置	检验类型	进行检验的类型，有“白噪声检验”，“平稳性检验“选项

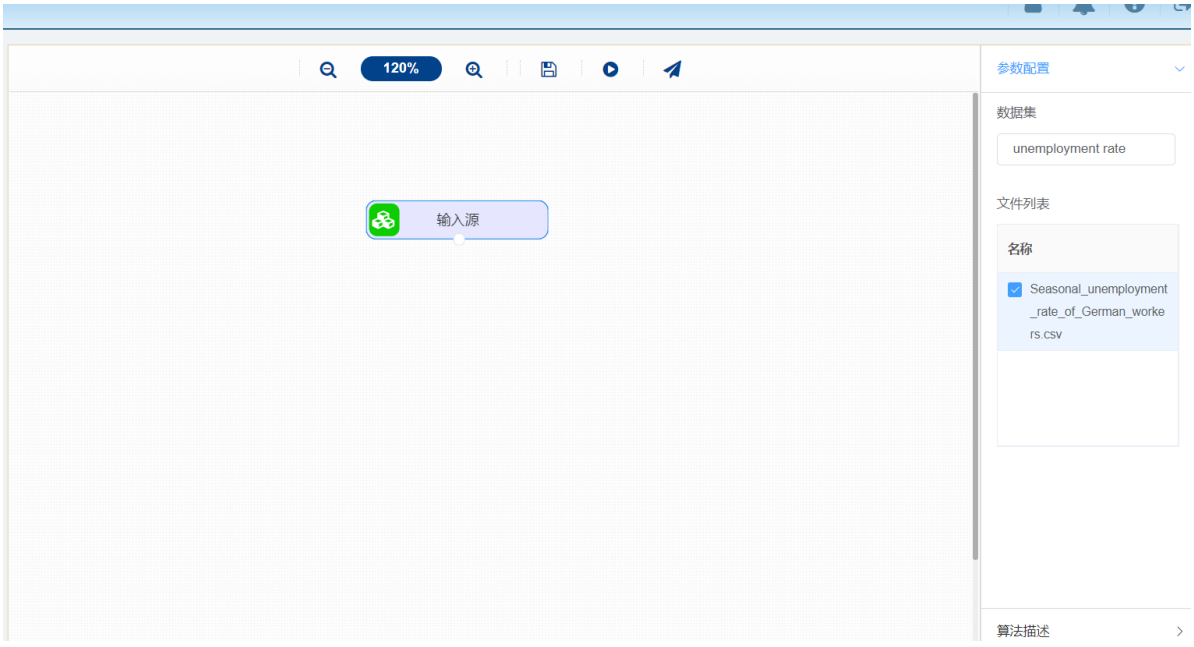
5 示例

数据集“unemployment rate”中数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行时序检验算法。

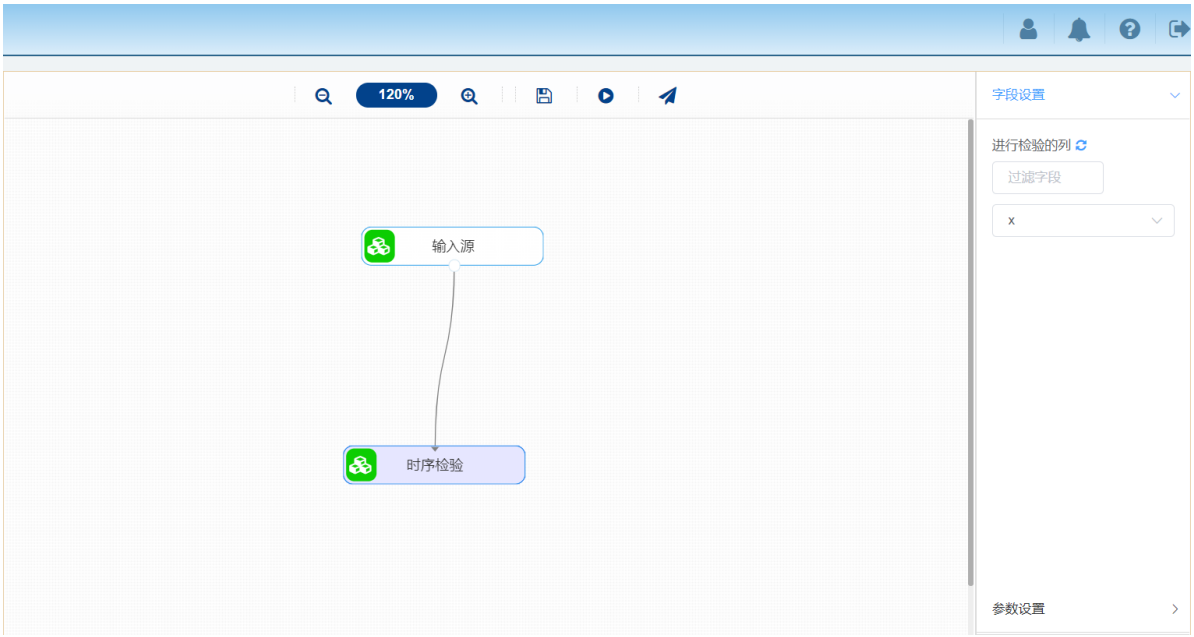


	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

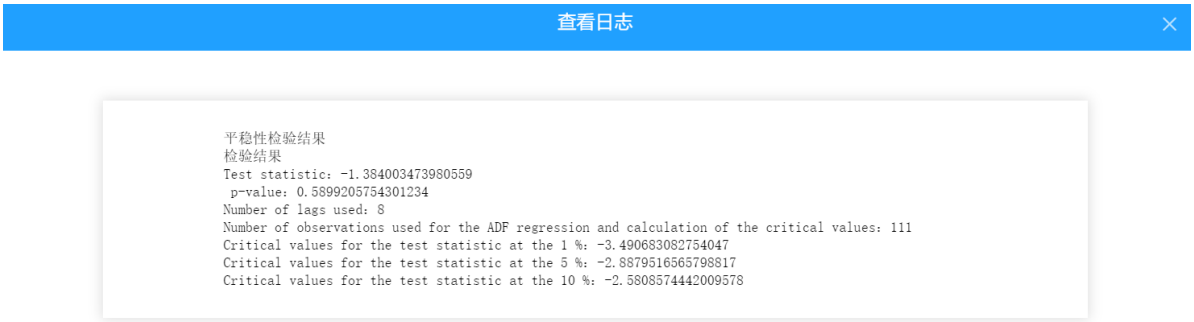
（1） 首先将需要进行时序检验的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“unemployment rate”，勾选文件“Seasonal_unemployment_rate_of_German_workers.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行时序检验，检验其平稳性。拖入【时序检验】算法，将【输入源】算法和【时序检验】算法相连接，在“字段设置”的“进行检验的列”中勾选“x”字段，点击“参数设置”，”检验类型“设置为“平稳性检验”，右键单击【时序检验】算法，选择“运行该节点”。



（3） 打开日志，查看结果。在日志中可以查看因子得分系数矩阵。对【因子分析】算法右击，点击“查看日志”。



序号	名称	作用
1	平稳性检验结果	用于时间序列以及时序模型的残差是否存在自相关性(是否为白噪声)

8.2.17 行列统计

1 作用

行列统计主要是计算数据的行数与列数，可应用于数据预处理阶段下对数据整体的统计分析。

2 输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

3 输出

序号	名称	内容
1	日志	得到数据行列统计结果

4 参数

序号	分组	参数	解释
1	字段设置	特征	

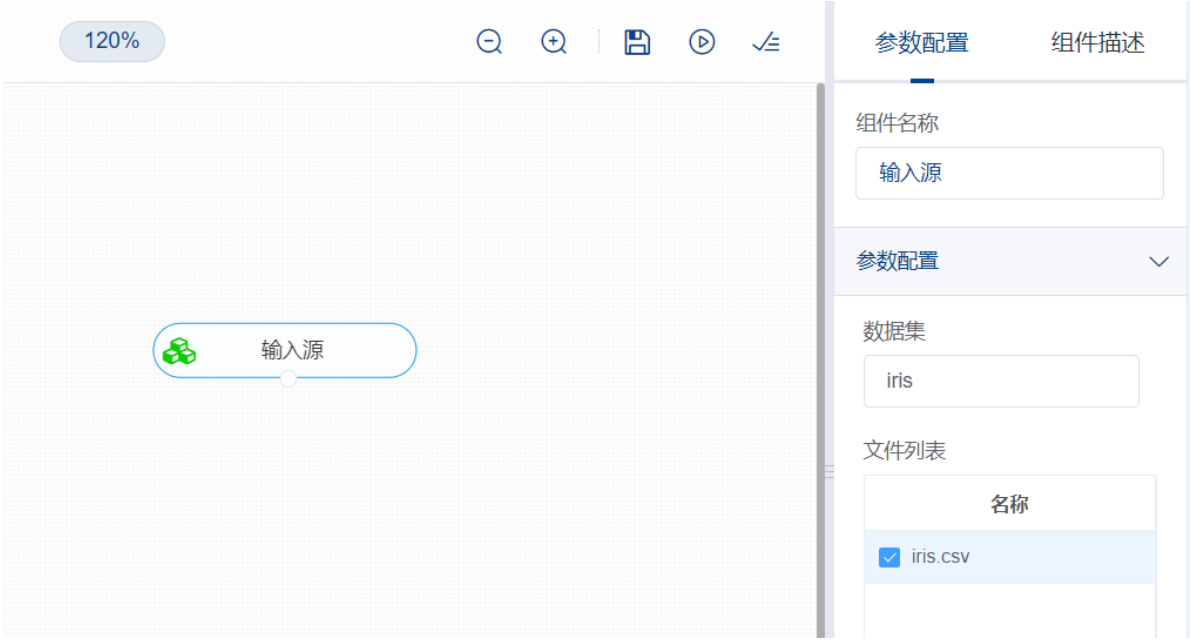
5 示例

对iris数据集进行行列统计示例。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

（1） 首先将需要进行因子分析的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行行列统计，计算各个变量的频数。拖入【行列统计】算法，将【输入源】算法和【行列统计】算法相连接，在“字段设置”的特征中勾选需要的特征列，右键单击【行列统计】算法，选择“运行该节点”。

（3）打开日志，查看结果。在日志中可以查看行列统计结果。对【行列】算法右击，点击“查看日志”。

8.3 聚类

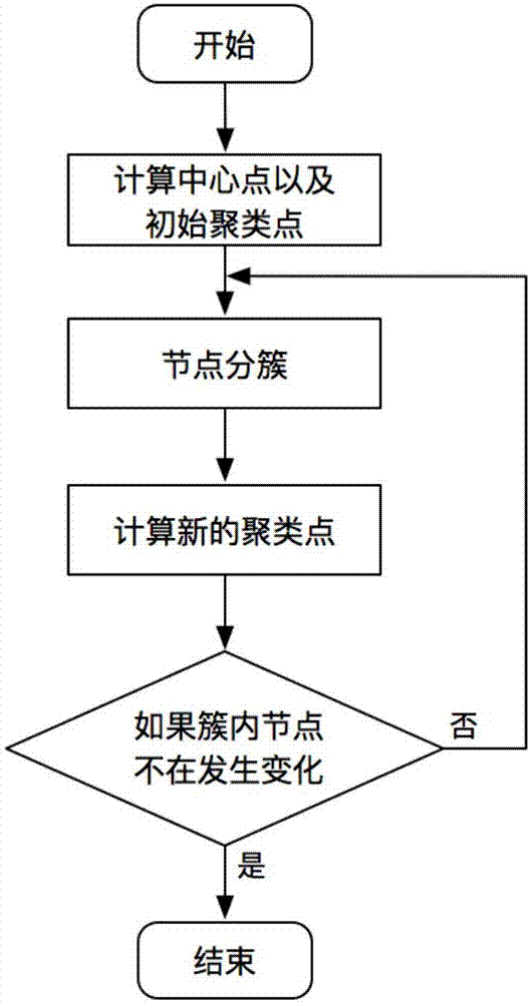
聚类主要应用于模式识别中的语音识别、字符识别等,机器学习中的聚类算法应用于图像分割和机器视觉,图像处理中聚类用于数据压缩和信息检索。聚类的另一个主要应用是数据挖掘(多关系数据挖掘)、时空数据库应用(GIS等)、序列和异类数据分析等，聚类分析对生物学、心理学、考古学、地质学、地理学以及市场营销等研究也都有重要作用。

8.3.1 KMeans

（1）作用及原理

k均值聚类算法（k-means clustering algorithm）是一种“物以类聚”的科学有效的方法，其作用是将没有进行分类的、无规律的、错综复杂的原始数据，要使得这些数据能够反映出一定的规律性或特殊的分类性，需要对数据或变量进行聚类分析，以使数据或变量呈现一定的分门别类的特征。

其原理是一种迭代求解的聚类分析算法，聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。



（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及距离计算
2	数据是否允许缺失值	否	涉及平均值计算
3	数据是否需要去除重复值	否	涉及平均值计算
4	载入文件格式	CSV格式	
5	数据量建议不超过		

（3）输出

序号	名称	内容
1	data_out.csv	增加了add_col列的被标准化后的原始数据，其中，add_col为聚类类标号。
2	cluster_center.csv	为聚类中心数据。可用来做雷达图，根据雷达图分析出对象的特征情况。
3	日志	含有聚类中心，饼图，雷达图和散点图。根据日志可分析各群体的优劣势以及占比。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行聚类的列，数值型
2	基础参数	最大迭代次数	迭代的次数，数值型
3	基础参数	聚类个数	聚类的个数，数值型，默认值为3

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行KMeans算法。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行KMeans聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行KMeans聚类，将数据集分门别类。拖入【KMeans】算法，将【输入源】算法和【KMeans】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击参数设置，聚类个数设置为3，最大迭代次数设置为100，右键单击【KMeans】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	聚类个数	3	数据集有三类品种的花
2	最大迭代次数	100	迭代次数越多，结果越准确。但随着迭代次数不断增大，或是数据量的增加，运行所需要的时间也就越来越久，所以确定一个最佳的迭代次数非常重要。

打开日志，查看结果。在日志中可以查看各个聚类分群的个数与占比、每个分群中各个特征的优势与劣势以及聚类中心。对【KMeans】算法右击，点击“查看日志”。

序号	名称	作用
1	聚类中心	初始时聚类中心是在样本中随机选取的K个对象，所剩下其它对象，则根据它们与这些聚类中心的距离，分别将它们分配给与其最相似的聚类中心所代表的聚类，在每分配一个样本后，聚类中心会根据聚类中现有的对象被重新计算。聚类中心可用来做雷达图，根据雷达图分析出对象的特征情况。
2	饼图	饼图可看出聚类分群个数，以及各个分群的个数和占比。
3	雷达图	雷达图在每个属性上的大小反应的是每个分群中该特征的优势和劣势。
4	散点图	过对数据进行降维，在二维空间中展示的聚类结果。

8.3.2 快速KMeans

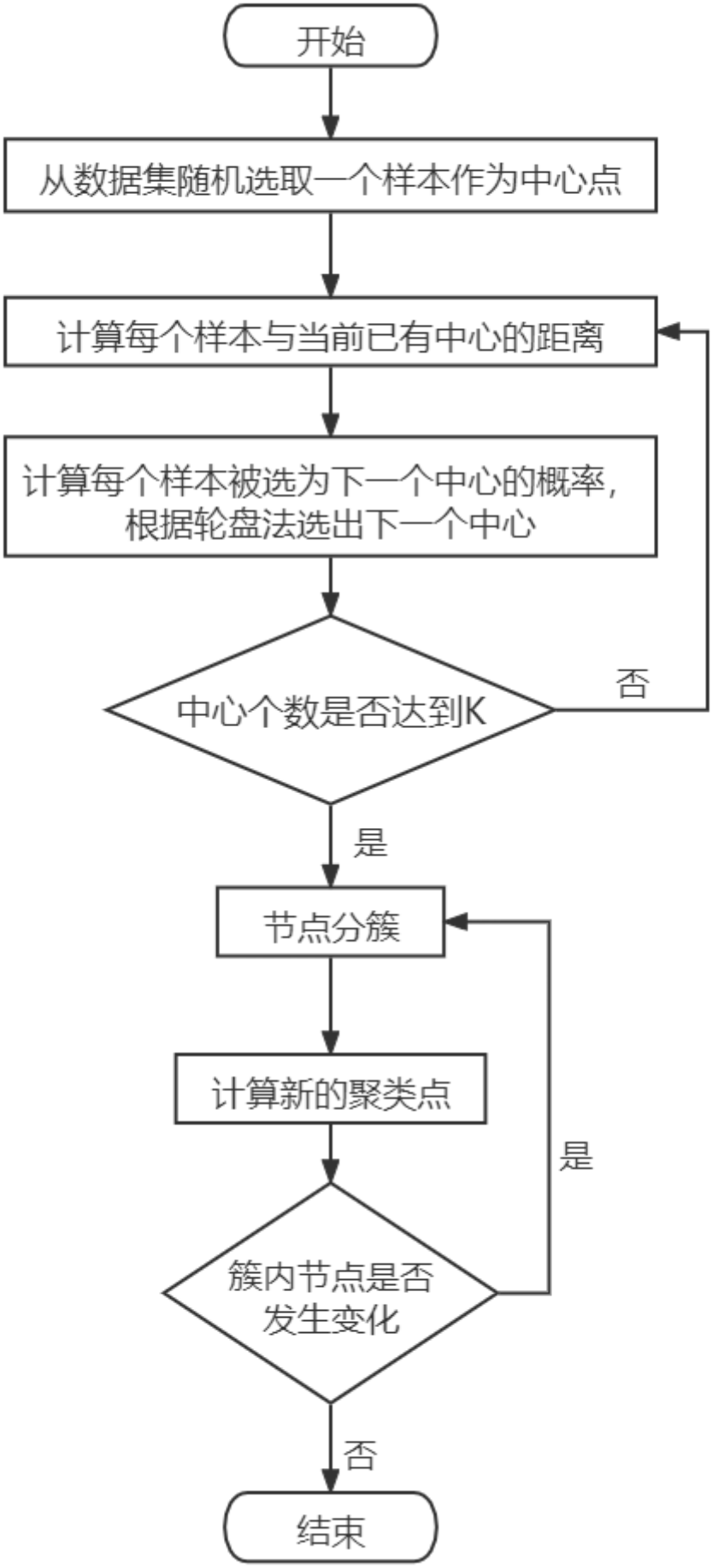
（1）作用及原理

快速KMeans(KMeans++)是KMeans算法的改进，作用与KMeans相同。

但其原理与传统KMeans有所区别。原始KMeans算法最开始随机选取数据集中K个点作为聚类中心，而KMeans++按照如下的思想选取K个聚类中心：

1. 假设已经选取了n个初始聚类中心(0<n<K)，则在选取第n+1个聚类中心时：距离当前n个聚类中心越远的点会有更高的概率被选为第n+1个聚类中心。
2. 在选取第一个聚类中心(n=1)时同样通过随机的方法。

KMeans++能显著的改善分类结果的最终误差。尽管计算初始点时花费了额外的时间，但是在迭代过程中，KMeans本身能快速收敛，因此算法实际上降低了计算时间。


开始
从数据集随机选取一个样本作为中心点
计算每个样本与当前已有中心的距离
计算每个样本被选为下一个中心的概率，
根据轮盘法选出下一个中心
中心个数是否达到K
否
是
节点分簇
计算新的聚类点
簇内节点是否
发生变化
是
否
结束

(2) 输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及距离计算
2	数据是否允许缺失值	否	涉及平均值计算
3	数据是否需要去除重复值	否	涉及平均值计算
4	载入文件格式	CSV格式	
5	数据量建议不超过	少量数据	

（3）输出

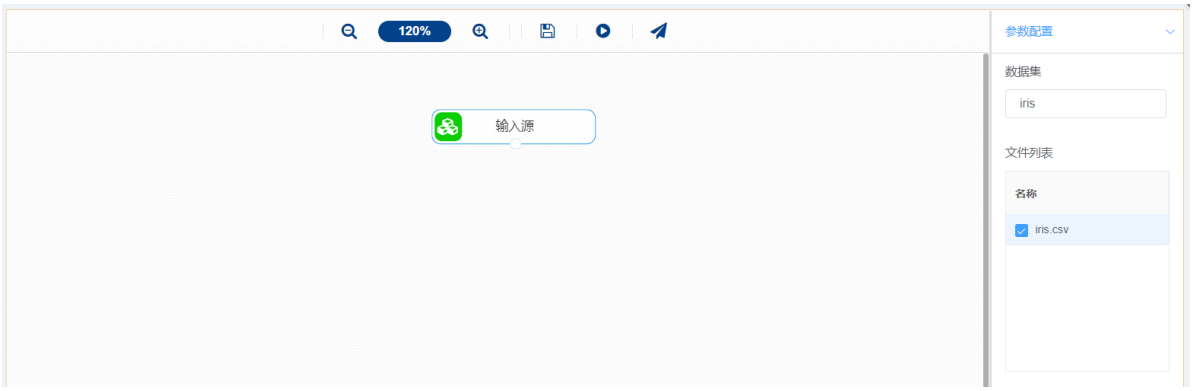
序号	名称	内容
1	data_out.csv	增加了add_col列的被标准化后的原始数据，其中，add_col为聚类类标号。
2	日志	含有聚类中心，饼图，雷达图和散点图。根据日志可分析各群体的优劣势以及占比。

（4）参数

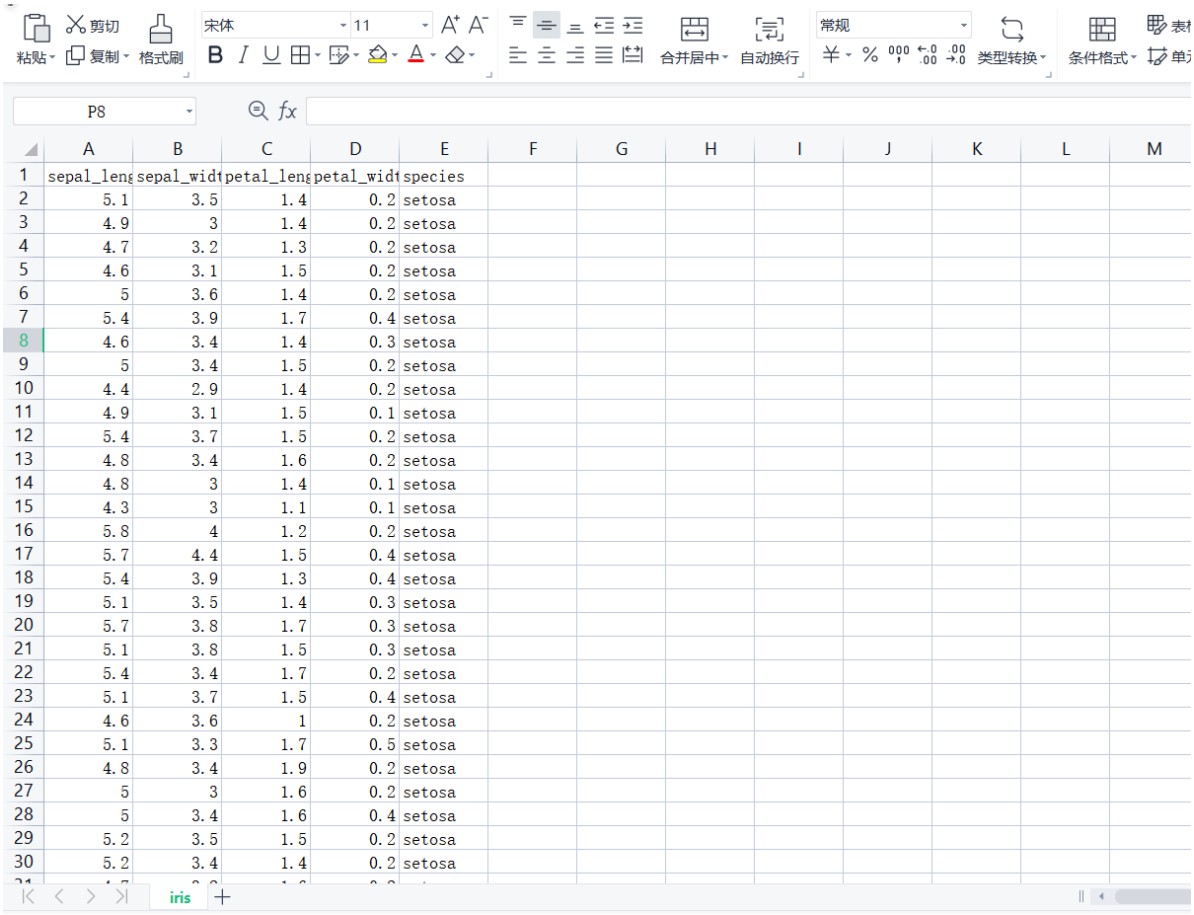
序号	分组	参数	解释
1	字段选择	特征列	需要进行聚类的列，数值型
2	基础参数	最大迭代次数	迭代的次数，数值型
3	基础参数	聚类个数	聚类的个数，数值型，默认值为3

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接对数据集进行快速KMeans聚类。

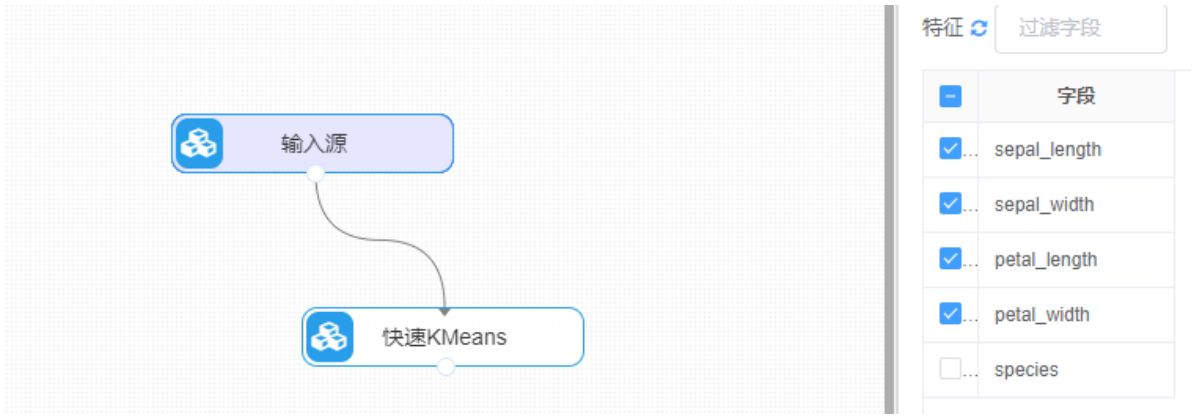


首先将需要进行快速KMeans聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

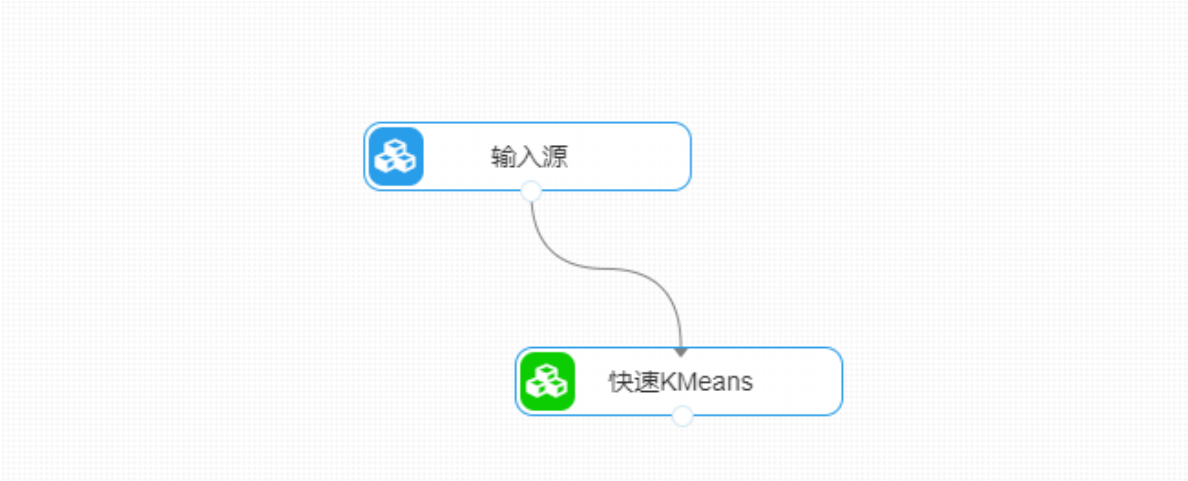


	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

开始进行快速KMeans聚类，将数据集分门别类。拖入【快速KMeans】算法，将【输入源】算法和【快速KMeans】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击参数设置，聚类个数设置为3，最大迭代次数设置为100，右键单击【快速KMeans】算法，选择“运行该节点”。



序号	参数名称	数值	原因
1	聚类个数	3	数据集有三类品种的花
2	最大迭代次数	100	迭代次数越多，结果越准确。但随着迭代次数不断增大，或是数据量的增加，运行所需要的时间也就越来越久，所以确定一个最佳的迭代次数非常重要。



打开日志，查看结果。在日志中可以查看各个聚类分群的个数与占比、每个分群中各个特征的优势与劣势以及聚类中心。对【快速KMeans】算法右击，点击“查看日志”。

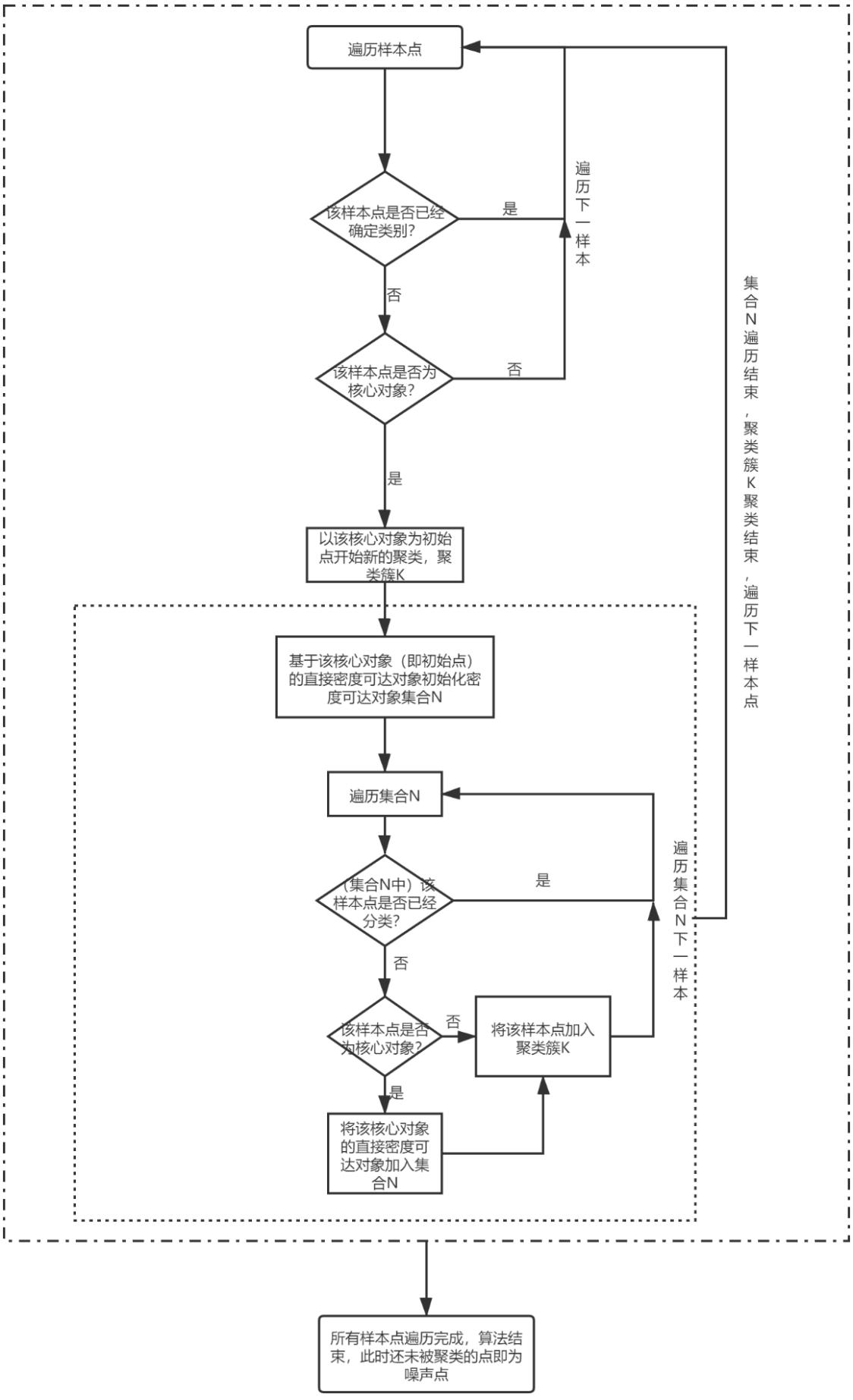
序号	名称	作用
1	聚类中心	初始时聚类中心是在样本中随机选取的K个对象，所剩下其它对象，则根据它们与这些聚类中心的距离，分别将它们分配给与其最相似的聚类中心所代表的聚类，在每分配一个样本后，聚类中心会根据聚类中现有的对象被重新计算。聚类中心可用来做雷达图，根据雷达图分析出对象的特征情况。
2	饼图	饼图可看出聚类分群个数，以及各个分群的个数和占比。
3	雷达图	雷达图在每个属性上的大小反应的是每个分群中该特征的优势和劣势。
4	散点图	通过对数据进行降维，在二维空间中展示的聚类结果。

8.3.3 DBSCAN密度聚类

（1）作用及原理

DBSCAN是一个比较有代表性的基于密度聚类的聚类算法，它对簇的定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在有噪声的数据中发现任意形状的聚类。

其原理是通过检查数据集中的每个对象的邻域来寻找聚类，如果一个点p的邻域包含对于m个对象，则创建一个p作为核心对象的新簇。然后，DBSCAN反复地寻找这些核心对象直接密度可达的对象，这个过程可能涉及密度可达簇的合并。当没有新的点可以被添加到任何簇时，该过程结束，得到最终的所有聚类类别结果。


遍历样本点
该样本点是否已经确定类别?
是
遍历下一样本
否
该样本点是否为核心对象?
否
是
以该核心对象为初始点开始新的聚类，聚类簇K
基于该核心对象（即初始点）的直接密度可达对象初始化密度可达对象集合N
遍历集合N
（集合N中）该样本点是否已经分类?
是
遍历集合N下一样本
否
该样本点是否为核心对象?
否
将该样本点加入聚类簇K
是
将该核心对象的直接密度可达对象加入集合N
集合N遍历结束，聚类簇K聚类结束，遍历下一样本点
所有样本点遍历完成，算法结束，此时还未被聚类的点即为噪声点

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	否	数据量纲不影响计算
2	数据是否允许缺失值	否	涉及平均值计算
3	数据是否需要去除重复值	否	涉及平均值计算
4	载入文件格式	CSV格式	
5	数据量建议不超过		

（3）输出

序号	名称	内容
1	data_out.csv	增加了add_col列的原始数据，其中，add_col为聚类类标号。
2	日志	含有聚类占比饼图和聚类散点图。根据日志可分析各群体的优劣势以及占比。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行聚类的列，数值型
2	基础参数	领域内最小样本数目	确定核心点的参数，数值型
3	基础参数	领域大小	核心点要成为核心对象所需要的邻域样本数阈值，数值型
4	基础参数	叶子大小	最近邻搜索算法参数，影响算法的运行速度和使用内存大小，数值型
5	基础参数	NearestNeighbors模块使用的算法	实现算法选择，默认为auto

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接对数据集进行DBSCAN密度算法。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行DBSCAN密度算法聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行DBSCAN密度算法聚类。拖入【DBSCAN密度算法】算法，将【输入源】算法和【DBSCAN密度算法】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击参数设置，领域内最小样本数目设置为10，领域大小设置为0.5，其他参数保持默认，右键单击【DBSCAN密度算法】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	领域内最小样本数目	10	领域内最小样本数目过大，则核心对象会过少，此时簇内部分本来是一类的样本可能会被标为噪音点，类别数也会变多。反之领域内最小样本数目过小的话，则会产生大量的核心对象，可能会导致类别数过少。
2	领域大小	0.5	领域大小对分类结果影响非常大，若参数设置过小，大部分数据不能聚类；若参数设置过大，多个簇和大部分对象会归并到同一个簇中。
3	叶子大小	30	最近邻搜索算法参数，没有运用保持默认。
4	NearestNeighbors模块使用的算法	auto	auto会在三种算法中做权衡，选择一个拟合最好的最优算法。

在日志中可以查看各个聚类分群的个数与比例、每个分群中个体的聚类散点图。初始时是由一个任意未被访问的点开始，然后探索这个点的ε-邻域，如果ε-邻域里有足够的点，则建立一个新的聚类，否则这个点被标签为噪音。注意这个点之后可能被发现在其它点的ε-邻域里，而该ε-邻域可能有足够的点，届时这个点会被加入该聚类中。由于初始点的确定是随机的，最终得到的结果也是不同的，但是参数相同的情况下大体相似。对【DBSCAN密度算法】算法右击，点击“查看日志”。

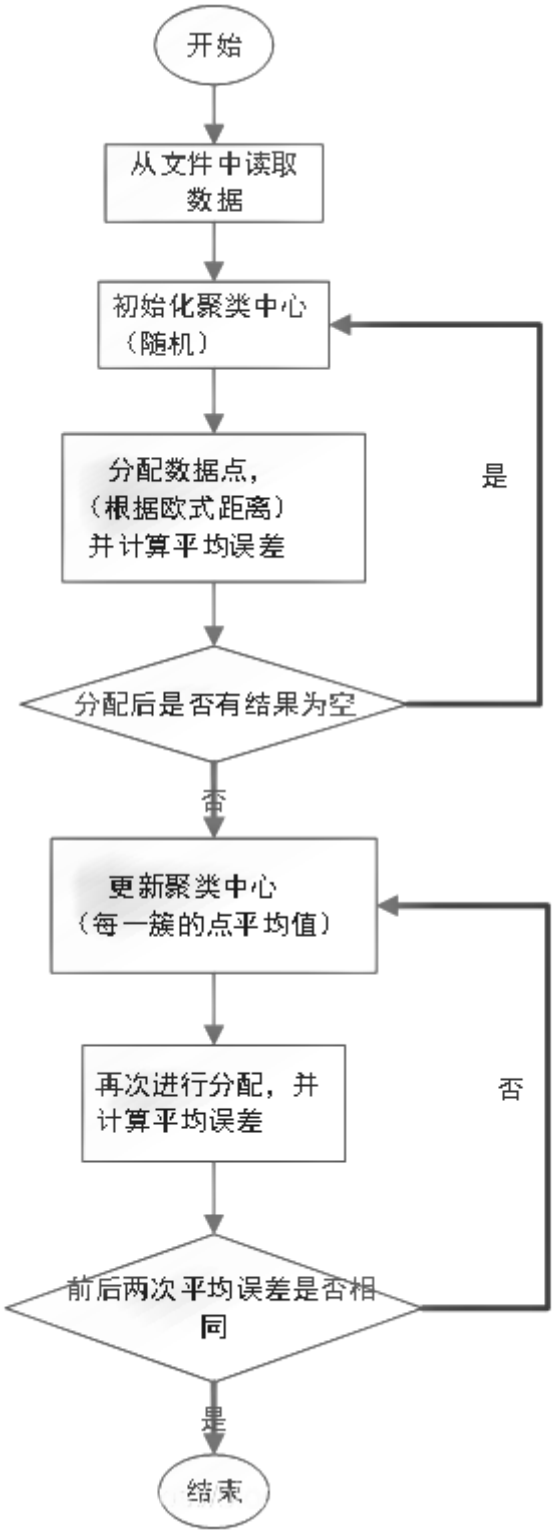
序号	名称	作用
1	饼图	饼图可看出聚类分群个数，以及各个分群的个数和占比。
2	散点图	过对数据进行降维，在二维空间中展示的聚类结果。

8.3.4 K-中心点聚类

（1）作用及原理

K-中心点聚类与KMeans算法类似，作用是将n个样本分成k个聚类，每个聚类里的样本关联性（相似性）较强。K-中心点聚类的基本思想和KMeans的思想相同，实质上是对Kmeans算法的优化和改进。

K-中心聚类算法计算的是某点到其它所有点的距离之和最小的点，通过距离之和最短的计算方式可以减少某些孤立数据对聚类过程的影响。从而使得最终效果更接近真实划分，但是由于上述过程的计算量会相对于Kmeans，大约增加O（n）的计算量，因此一般情况下K-中心算法更加适合小规模数据运算。



(2) 输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及距离计算
2	数据是否允许缺失值	否	涉及平均值计算
3	数据是否需要去除重复值	否	涉及平均值计算
4	载入文件格式	CSV格式	
5	数据量建议不超过		

(3) 输出

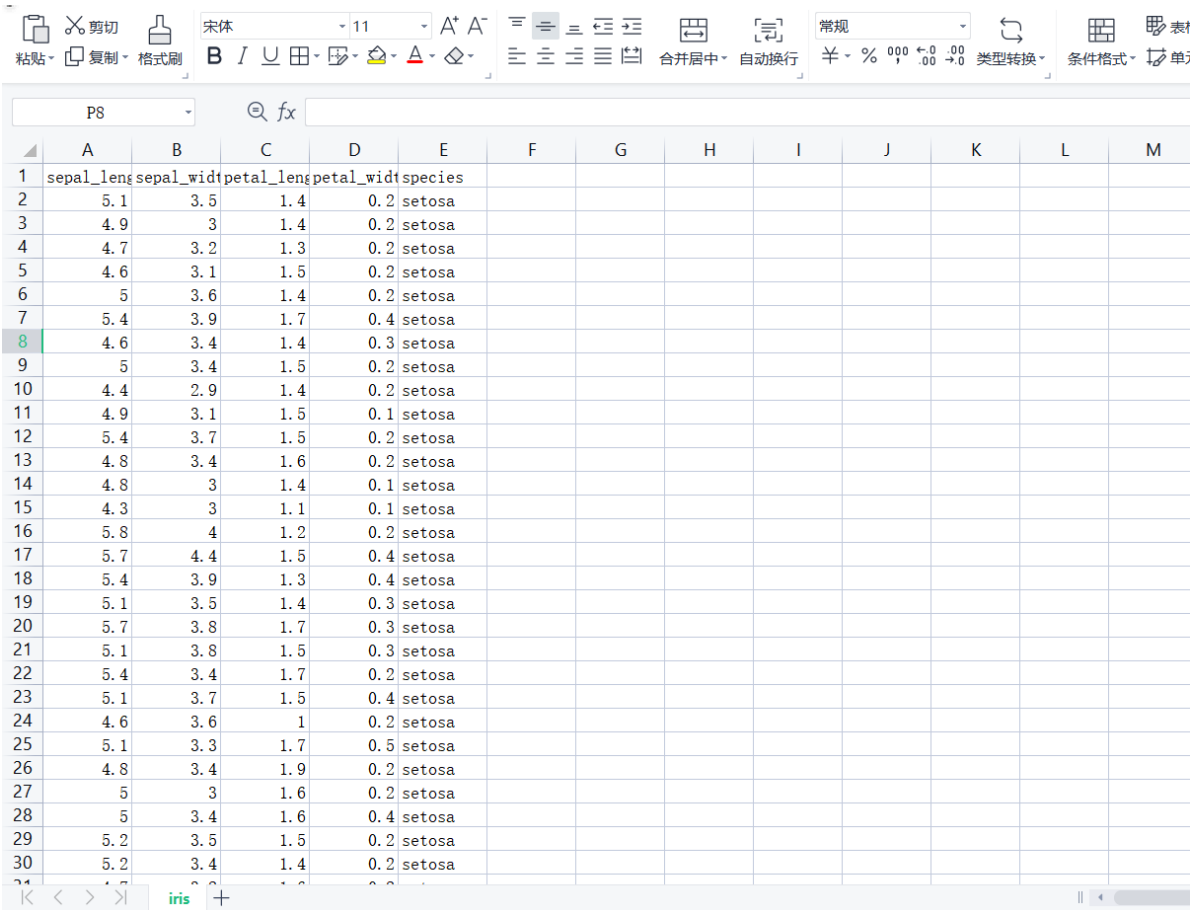
序号	名称	内容
1	data_out.csv	增加了add_col列的原始数据，其中，add_col为聚类类标号。
2	日志	含有聚类占比饼图和聚类散点图。根据日志可分析各群体的优劣势以及占比。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行聚类的列，数值型[O2]
2	基础参数	K值	聚类的个数，数值型，默认值为3
2	基础参数	最大迭代次数	迭代的次数，数值型
3	基础参数	随机质心初始化的数量	初始质心的数量，与聚类个数相关，数值型，默认值为3

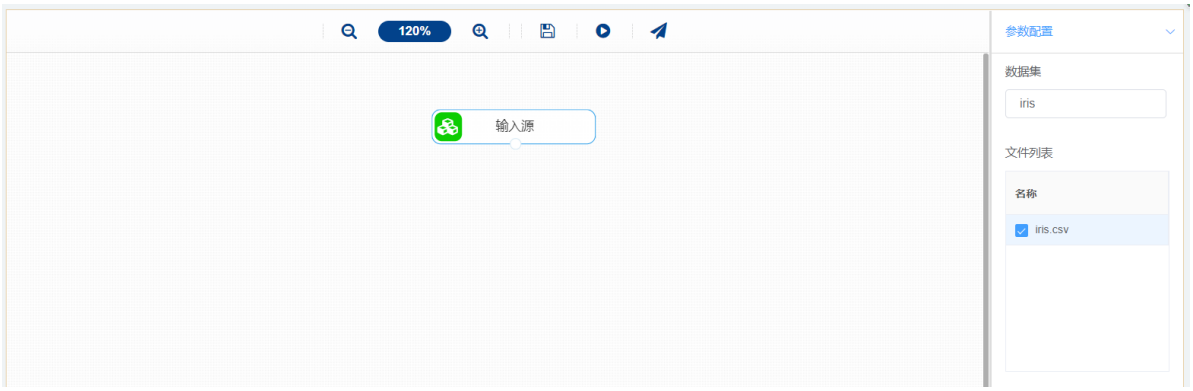
（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接对数据集进行KMeans算法。



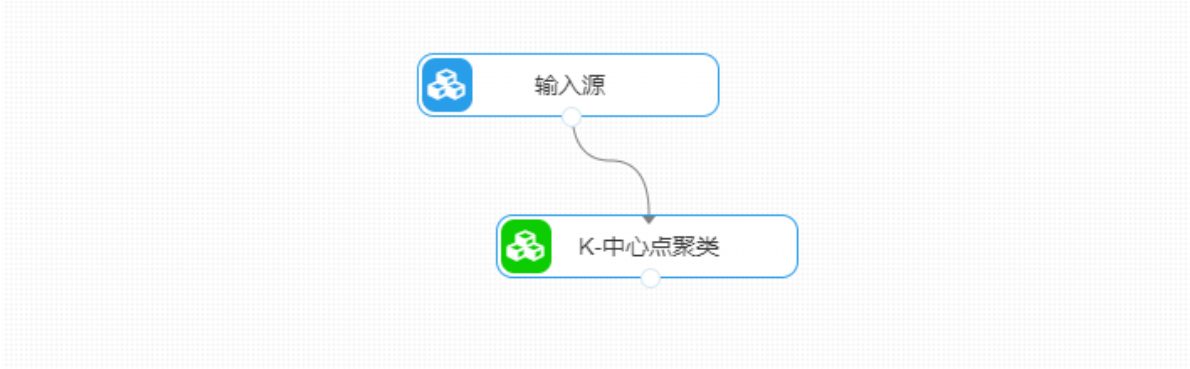
	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

首先将需要进行K-中心点聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



开始进行K-中心点聚类，将数据集分门别类。拖入【K-中心点聚类】算法，将【输入源】算法和【K-中心点聚类】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击参数设置，K值设置为3，最大迭代次数设置为100，随机质心初始化的数量设置为3，右键单击【K-中心点聚类】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	K值	3	数据集有三类品种的花
2	最大迭代次数	100	迭代次数越多，结果越准确。但随着迭代次数不断增大，或是数据量的增加，运行所需要的时间也就越来越久，所以确定一个最佳的迭代次数非常重要。
3	随机质心初始化的数量	3	与K值相关，设置为3



打开日志，查看结果。在日志中可以查看各个聚类分群的个数与占比、每个分群中各个特征的优势与劣势以及聚类中心。对【K-中心点聚类】算法右击，点击“查看日志”。

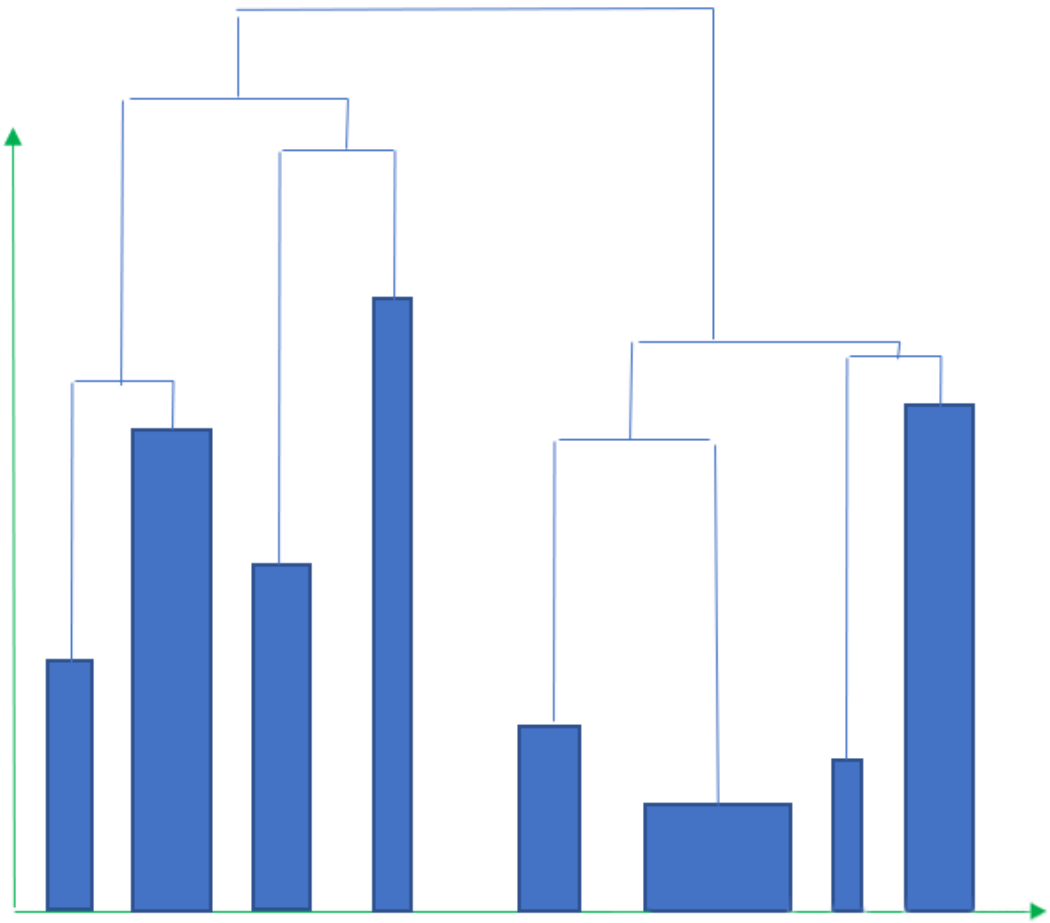
序号	名称	作用
1	聚类中心	初始时聚类中心是在样本中随机选取的K个对象，所剩下其它对象，则根据它们与这些聚类中心的距离，分别将它们分配给与其最相似的聚类中心所代表的聚类，在每分配一个样本后，聚类中心会根据聚类中现有的对象被重新计算。聚类中心可用来做雷达图，根据雷达图分析出对象的特征情况。
2	饼图	饼图可看出聚类分群个数，以及各个分群的个数和占比。
3	雷达图	雷达图在每个属性上的大小反应的是每个分群中该特征的优势和劣势。
4	散点图	通过对数据进行降维，在二维空间中展示的聚类结果。

8.3.5 层次聚类

（1）作用及原理

层次聚类(Hierarchical Clustering)是聚类算法的一种，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点。层次聚类算法相比划分聚类算法的优点之一是可以在不同的尺度上（层次）展示数据集的聚类情况。AggregativeClustering是一种常用的层次聚类算法。

其原理是：通过计算两类数据点间的相似性，对所有数据点中最为相似的两个数据点进行组合，并反复迭代这一过程。简单的说层次聚类的合并算法是通过计算每一个类别的数据点与所有数据点之间的距离来确定它们之间的相似性，距离越小，相似度越高。并将距离最近的两个数据点或类别进行组合，生成聚类树。



（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及距离计算
2	数据是否允许缺失值	否	涉及平均值计算
3	数据是否需要去除重复值	否	涉及平均值计算
4	载入文件格式	CSV格式	
5	数据量建议不超过	少量数据	

（3）输出

序号	名称	内容
1	data_out.csv	增加了add_col列的被标准化后的原始数据，其中，add_col为聚类类标号。
2	日志	含有饼图和散点图。根据日志可分析各群体的优劣势以及占比。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行聚类的列，数值型
2	基础参数	K值	聚类的个数，数值型，默认值为3
3	基础参数	距离度量标准	用于计算连接的度量，字符型，默认为欧氏距离
4	基础参数	链接算法	使用哪种链接标准，字符型，默认为离差平方和

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行层次聚类。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行层次聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行层次聚类，将数据集分门别类。拖入【层次聚类】算法，将【输入源】算法和【层次聚类】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击参数设置，k值设置为3，距离度量标准选择欧氏距离，链接算法选择离差平方和，右键单击【层次聚类】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	k值	3	数据集有三类品种的花
2	距离度量标准	欧氏距离	因为“iris”数据集没有明显量纲差异，所以可以选择效果较好的欧氏距离来度量样本间距。
3	链接算法	离差平方和	离差平方和会合并聚类的差异，效果最佳。

打开日志，查看结果。在日志中可以查看各个聚类分群的个数与占比、每个分群中各个特征的优势与劣势以及聚类中心。对【层次聚类】算法右击，点击“查看日志”。

序号	名称	作用
1	饼图	饼图可看出聚类分群个数，以及各个分群的个数和占比。
2	树状图	展示每个聚类的组成
2	散点图	通过对数据进行降维，在二维空间中展示的聚类结果。

8.3.6 模糊聚类

（1）作用及原理

模糊c均值聚类融合了模糊理论的精髓。相较于k-means的硬聚类，模糊c提供了更加灵活的聚类结果。因为大部分情况下，数据集中的对象不能划分成为明显分离的簇，指派一个对象到一个特定的簇有些生硬，也可能会出错。故对每个对象和每个簇赋予一个权值，指明对象属于该簇的程度。当然，基于概率的方法也可以给出这样的权值，但是有时候我们很难确定一个合适的统计模型，因此使用具有自然地、非概率特性的模糊c均值就是一个比较好的选择。

其原理步骤与传统硬聚类算法类似。第一步初始化中心，通常采用随机初始化，即权值随机地选取，簇数需要人为选定。第二步计算质心，软聚类算法中的质心有别于传统质心的地方在于，它是以隶属度为权重做一个加权平均。第三步更新模糊伪划分，即更新权重（隶属度）。简单地说，如果样本x越靠近质心c，则隶属度越高，反之越低。重复迭代二三步，直至模型收敛或达到迭代次数。

1. *Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$*
2. *At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$*

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update $U^{(k)}$, $U^{(k+1)}$*

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. *If $|| U^{(k+1)} - U^{(k)}|| < \varepsilon$ then STOP; otherwise return to step 2.*

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及距离计算
2	数据是否允许缺失值	否	涉及平均值计算
3	数据是否需要去除重复值	否	涉及平均值计算
4	载入文件格式	CSV格式	
5	数据量建议不超过	少量数据	

（3）输出

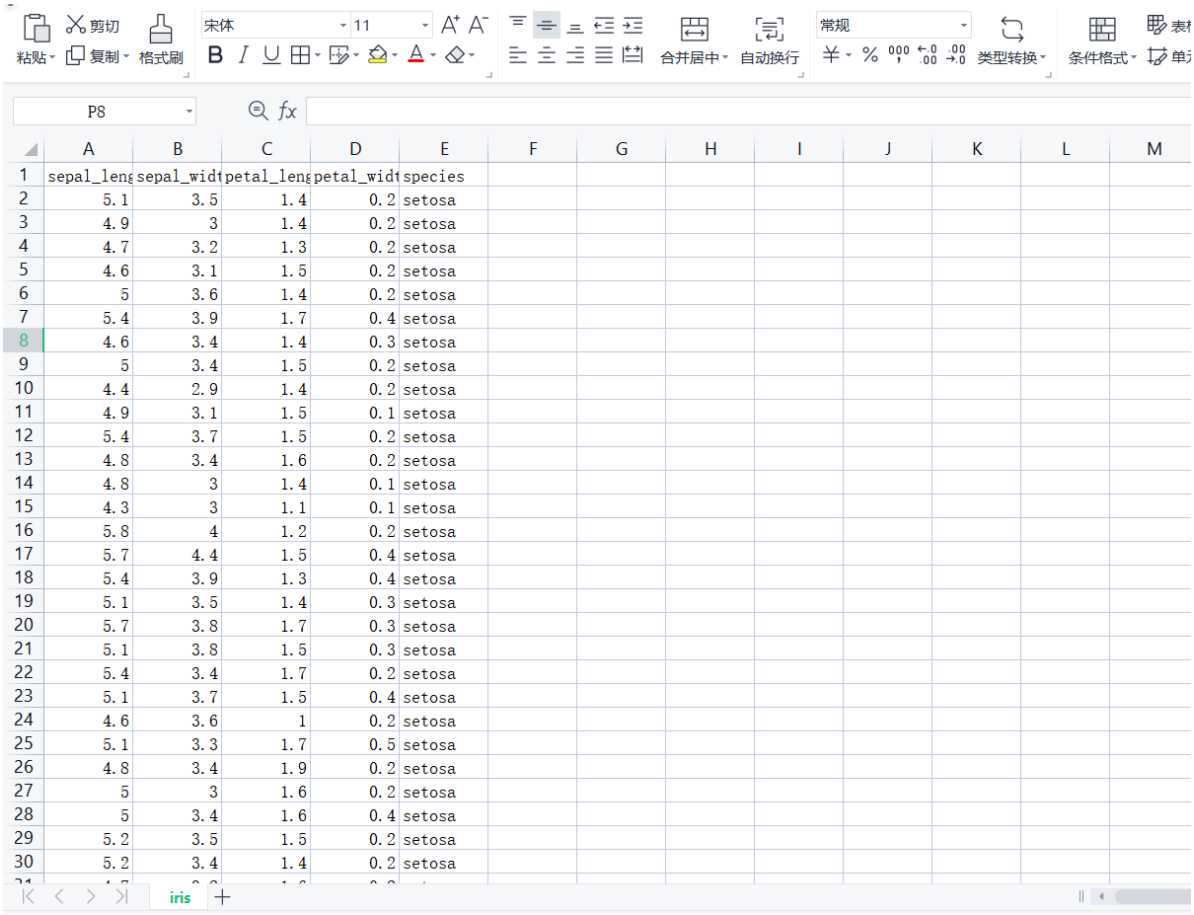
序号	名称	内容
1	data_out.csv	增加了add_col列的被标准化后的原始数据，其中，add_col为聚类类标号。
2	日志	含有散点图。根据日志可分析各群体的分布。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行聚类的列，数值型
2	基础参数	最大迭代次数	迭代的次数，数值型
3	基础参数	聚类个数	聚类的个数，数值型，默认值为3
4	基础参数	模糊参数	计算样本点隶属度的参数，要求大于1，默认值为2

（5）示例

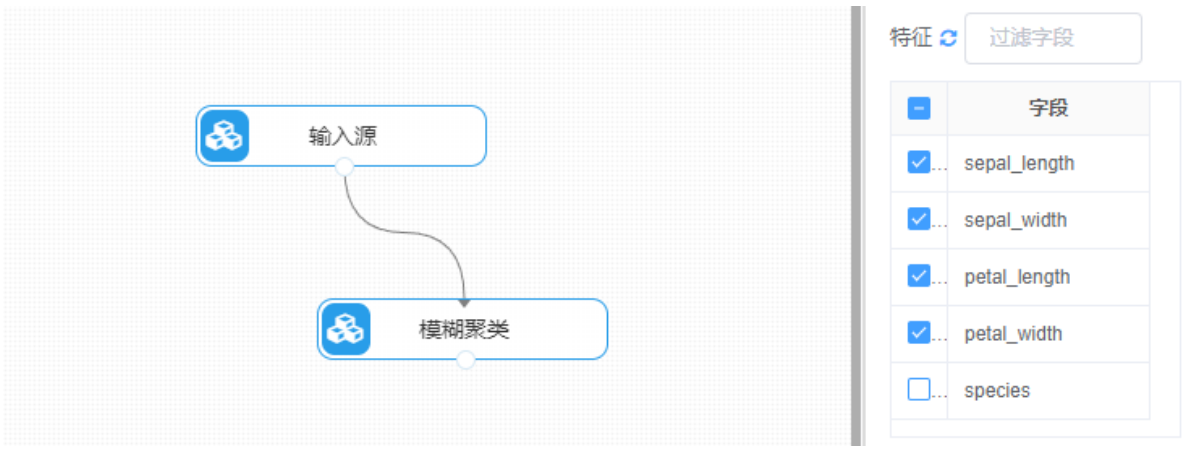
对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行模糊聚类。



	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

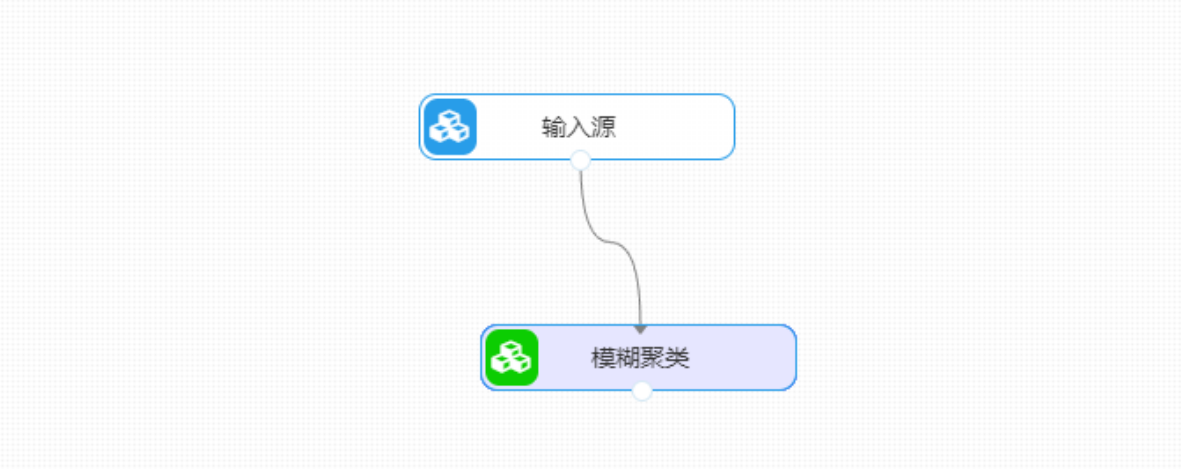
首先将需要进行模糊聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行模糊聚类，将数据集分门别类。拖入【模糊聚类】算法，将【输入源】算法和【模糊聚类】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击参数设置，聚类个数设置为3，最大迭代次数设置为100，右键单击【KMeans】算法，选择“运行该节点”。



序号	参数名称	序号	原因
1	聚类个数	3	数据集有三类品种的花
2	最大迭代次数	100	迭代次数越多，结果越准确。但随着迭代次数不断增大，或是数据量的增加，运行所需要的时间也就越来越久，所以确定一个最佳的迭代次数非常重要。
3	模糊参数	2	计算样本点隶属度的参数，当该参数为2时模糊程度较小，用于示例参考。

打开日志，查看结果。在日志中可以查看各个聚类分群的个数与占比、每个分群中各个特征的优势与劣势以及聚类中心。对【模糊聚类】算法右击，点击“查看日志”。

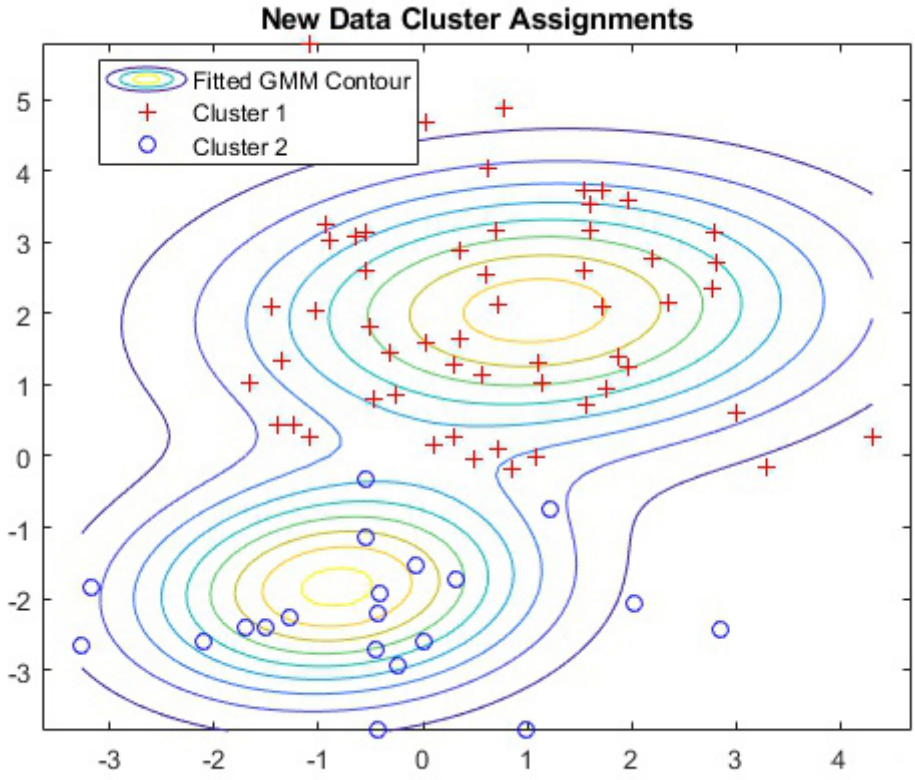


序号	名称	作用
1	饼图	饼图可看出聚类分群个数，以及各个分群的个数和占比。
2	散点图	通过对数据进行降维，在二维空间中展示的聚类结果。

8.3.7 高斯混合模型

（1）作用及原理

高斯混合模型（Gaussian Mixture Model），顾名思义，多个高斯分布的结合组成的概率分布模型，简称为GMM。事实上，GMM和KMeans很像，不过GMM是学习出一些概率密度函数来。简单地说，KMeans的结果是每个数据点被分类到其中某一个聚类，而GMM则给出这些数据点被分配到每个聚类的概率，又称作软聚类。



高斯混合模型的核心思想是，假设数据可以看作从多个高斯分布中生成出来 的。在该假设下，每个单独的分模型都是标准高斯模型，其均值 ui和方差是待估计的参数。此外，每个分模型都还有一个参数 可以理解为权重或生成数据的概率。高斯混合模型的公式为：

$$p(x) = \sum_{i=1}^{k} \pi_i N(x|u_i, \sum_i)$$

通常我们并不能直接得到高斯混合模型的参数，而是观察到了一系列数据点，给出一个类别的数量K后，希望求得最佳的K个高斯分模型。因此，高斯混合模型的计算，便成了最佳的均值μ、方差Σ、权重π的寻找，这类问题通常通过最大似然估计来求解。遗憾的是，此问题中直接使用最大似然估计，得到的是一个复杂的非凸函数，目标函数是和的对数，难以展开和对其求偏导。

在这种情况下，可以用EM算法。EM算法是在最大化目标函数时，先固定一个变量使整体函数变为凸优化函数，求导得到最值，然后利用最优参数更新被固定的变量，进入下一个循环。具体到高斯混合模型的求解，EM算法的迭代过程如下。

首先，初始随机选择各参数的值。然后，重复下述两步，直到收敛。

1. E步骤。根据当前的参数，计算每个点由某个分模型生成的概率。
2. M步骤。使用E步骤估计出的概率，来改进每个分模型的均值，方差和权重。

也就是说，我们并不知道最佳的K个高斯分布的各自3个参数，也不知道每个数据点究竟是哪个高斯分布生成的。所以每次循环时，先固定当前的高斯分布不变，获得每个数据点由各个高斯分布生成的概率。然后固定该生成概率不变，根据数据点和生成概率，获得一个组更佳的高斯分布。循环往复，直到参数的不再变化，或者变化非常小时，便得到了比较合理的一组高斯分布。

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及距离计算
2	数据是否允许缺失值	否	涉及平均值计算
3	数据是否需要去除重复值	否	涉及平均值计算
4	载入文件格式	CSV格式	
5	数据量建议不超过	少量数据	

（3）输出

序号	名称	内容
1	data_out.csv	增加了add_col列的被标准化后的原始数据，其中，add_col为聚类类标号。
2	日志	含有饼图和散点图。根据日志可分析各群体的优劣势以及占比。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行聚类的列，数值型
2	基础参数	K值	聚类的个数，数值型，默认值为3
3	基础参数	指定协方差类型	包括{‘full’,‘tied’,‘diag’,‘spherical’}四种，full指每个分量有各自不同的标准协方差矩阵，完全协方差矩阵（元素都不为零），tied指所有分量有相同的标准协方差矩阵（HMM会用到），diag指每个分量有各自不同对角协方差矩阵（非对角为零，对角不为零），spherical指每个分量有各自不同的简单协方差矩阵，球面协方差矩阵（非对角为零，对角完全相同，球面特性），默认‘full’完全协方差矩阵
4	基础参数	指定初始化权重的策略	初始化参数实现方式，默认用kmeans实现，也可以选择随机产生

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接对数据集进行高斯混合模型聚类。

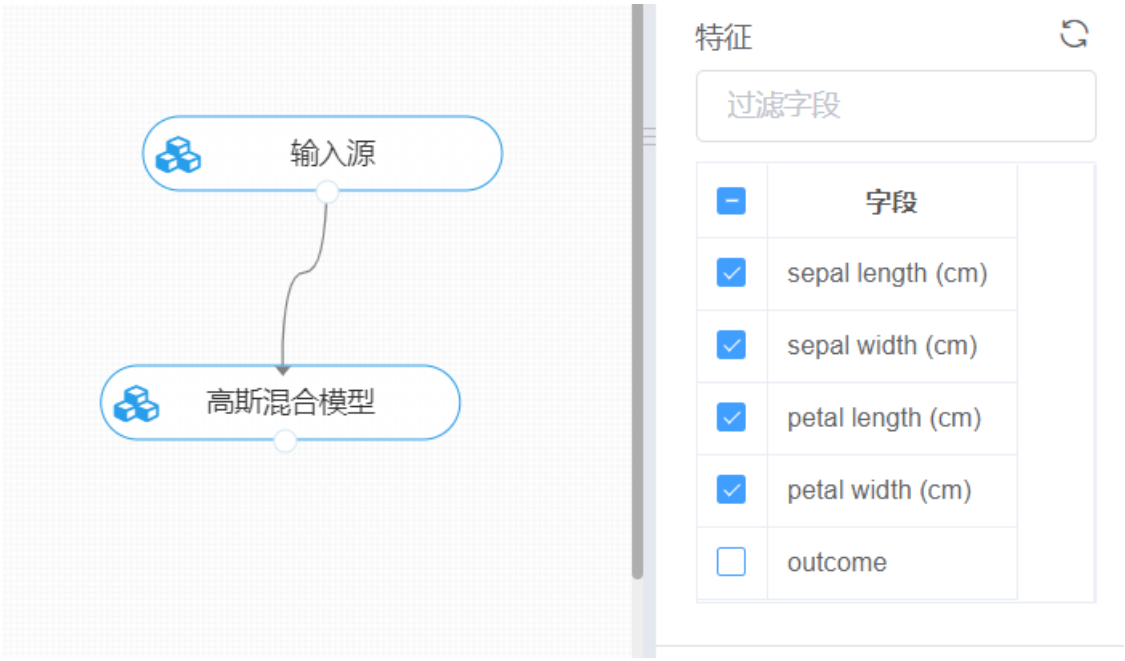
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行高斯混合模型聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行高斯混合模型聚类，将数据集分门别类。拖入【高斯混合模型】算法，将【输入源】算法和【高斯混合模型】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击参数设置，聚类个数设置为3，最大迭代次数设置为100，右键单击【高斯混合模型】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	K值	3	数据集有三类品种的花
2	指定协方差类型	full	通常情况下，完全协方差表现最好
3	指定初始化权重的策略	kmeans	相对于随机生成，kmeans初始化效果更好。



打开日志，查看结果。在日志中可以查看各个聚类分群的个数与占比、每个分群中各个特征的优势与劣势以及聚类中心。对【高斯混合模型】算法右击，点击“查看日志”。

序号	名称	作用
1	饼图	饼图可看出聚类分群个数，以及各个分群的个数和占比。
2	散点图	通过对数据进行降维，在二维空间中展示的聚类结果。

8.3.8 EM聚类

1、作用

EM算法最大期望算法，是一种迭代优化策略，由于它的计算方法中每一次迭代都分两步，其中一个为期望步（E步），另一个为极大步（M步），所以算法被称为EM算法（Expectation Maximization Algorithm）。其基本思想是：首先根据已经给出的观测数据，估计出模型参数的值；然后再依据上一步估计出的参数值估计缺失数据的值，再根据估计出的缺失数据加上之前已经观测到的数据重新再对参数值进行估计，然后反复迭代，直至最后收敛，迭代结束。

可以通过k-Means算法来简单理解EM算法过程：

- E步：在初始化K个中心点后，我们对所有的样本归到K个类别。
- M步：在所有的样本归类后，重新求K个类别的中心点，相当于更新了均值。

EM算法流程：

$$Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)}, \theta^j)$$

$$L(\theta, \theta^j) = \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) logP(x^{(i)}, z^{(i)}|\theta)$$

$$\theta^{j+1} = arg\max_{\theta} L(\theta, \theta^j)$$

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及距离计算
2	数据是否允许缺失值	否	涉及平均值计算
3	数据是否需要去除重复值	否	涉及平均值计算
4	载入文件格式	CSV格式	
5	数据量建议不超过		

（3）输出

序号	名称	内容
1	data_out.csv	增加了add_col列的被标准化后的原始数据，其中，add_col为聚类类标号。
2	日志	含有聚类中心，饼图，树状图和散点图。根据日志可分析各群体的分布。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行聚类的列，数值型
2	参数设置	k值	聚类的个数，数值型，默认值为3

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接对数据集进行EM算法。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行EM聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行EM聚类，将数据集分门别类。拖入【EM】算法，将【输入源】算法和【EM】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击参数设置，聚类个数设置为3，右键单击【EM】算法，选择“运行该节点”。

打开日志，查看结果。在日志中可以查看各个聚类分群的个数与占比、每个分群中各个特征的优势与劣势以及聚类中心。对【EM】算法右击，点击“查看日志”。

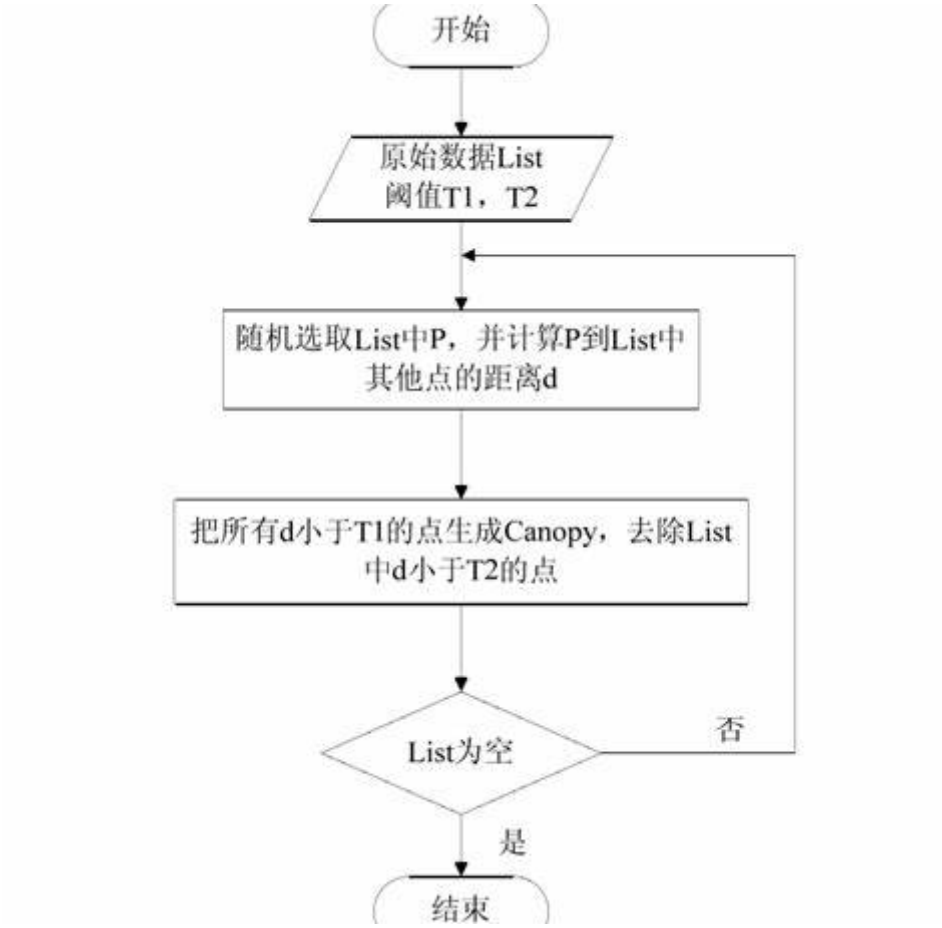
序号	名称	作用
1	饼图	饼图可看出聚类分群个数，以及各个分群的个数和占比。
2	散点图	通过对数据进行降维，在二维空间中展示的聚类结果。
3	树状图	展示每个聚类的组成

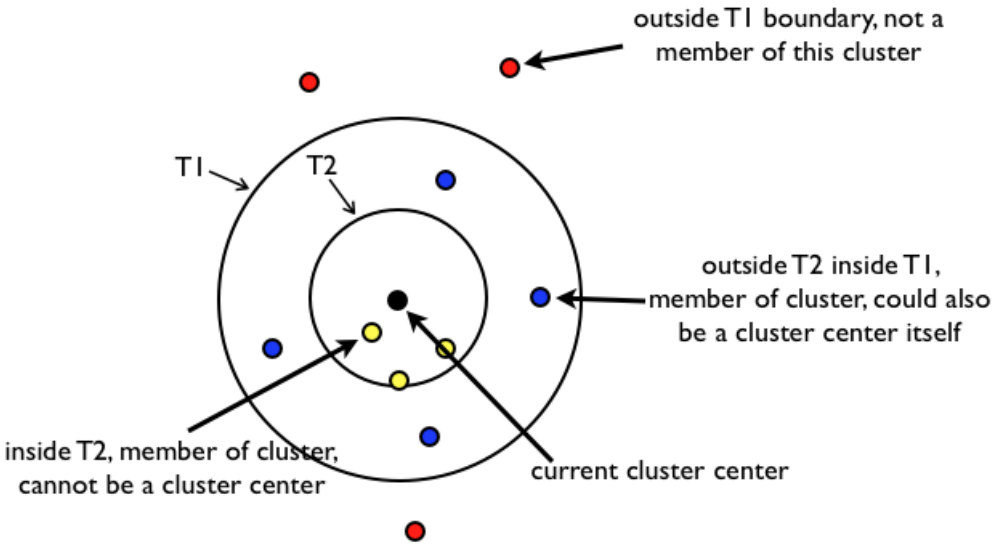
8.3.9 Canopy聚类算法

（1）作用

Canopy聚类算法是一个将对象分组到类的简单、快速、精确地方法。每个对象用多维特征空间里的一个点来表示。这个算法使用一个快速近似距离度量和两个距离阈值T1>T2来处理。

基本的算法是，从一个点集合开始并且随机删除一个，创建一个包含这个点的Canopy，并在剩余的点集合上迭代。对于每个点，如果它的距离第一个点的距离小于T1，然后这个点就加入这个聚集中。除此之外，如果这个距离<T2，然后将这个点从这个集合中删除。这样非常靠近原点的点将避免所有的未来处理，不可以再做其它Canopy的中心。这个算法循环到初始集合为空为止，聚集一个集合的Canopies，每个可以包含一个或者多个点。每个点可以包含在多于一个的Canopy中。





（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及距离计算
2	数据是否允许缺失值	否	涉及平均值计算
3	数据是否需要去除重复值	否	涉及平均值计算
4	载入文件格式	CSV格式	
5	数据量建议不超过		

（3）输出

序号	名称	内容
1	data_out.csv	增加了add_col列的被标准化后的原始数据，其中，add_col为聚类类标号。
2	日志	含有饼图和散点图。可根据日志可分析各群体的分布。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行聚类的列，数值型
2	参数设置	T1,T2	聚类过程中的距离阈值，T1>T2

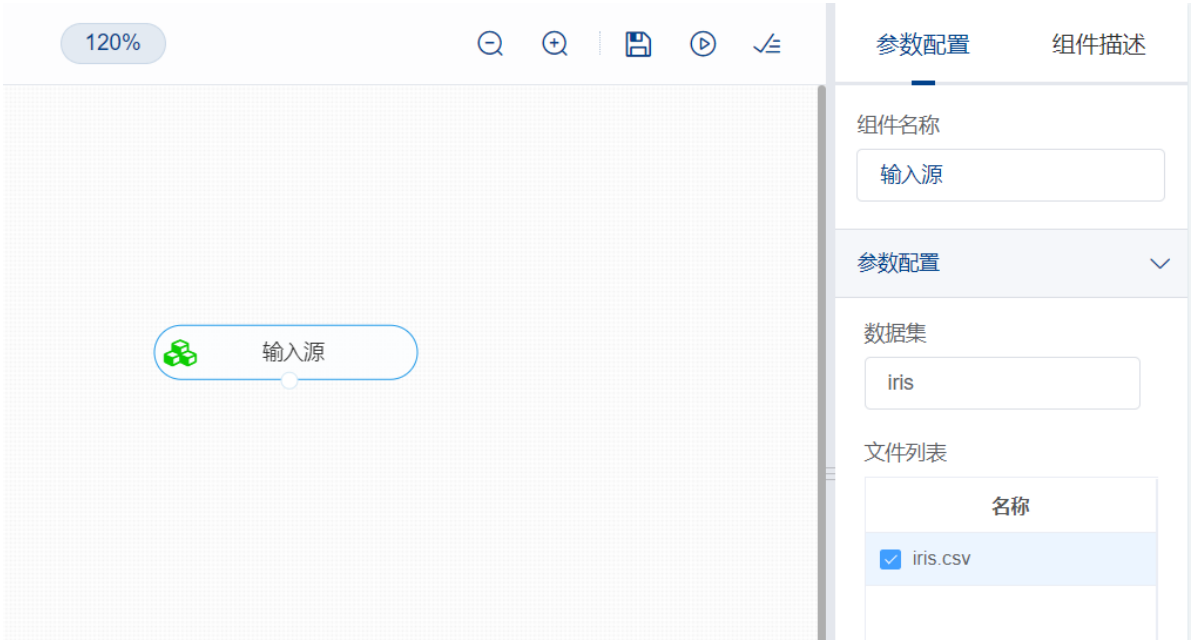
（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接对数据集进行Canopy算法。

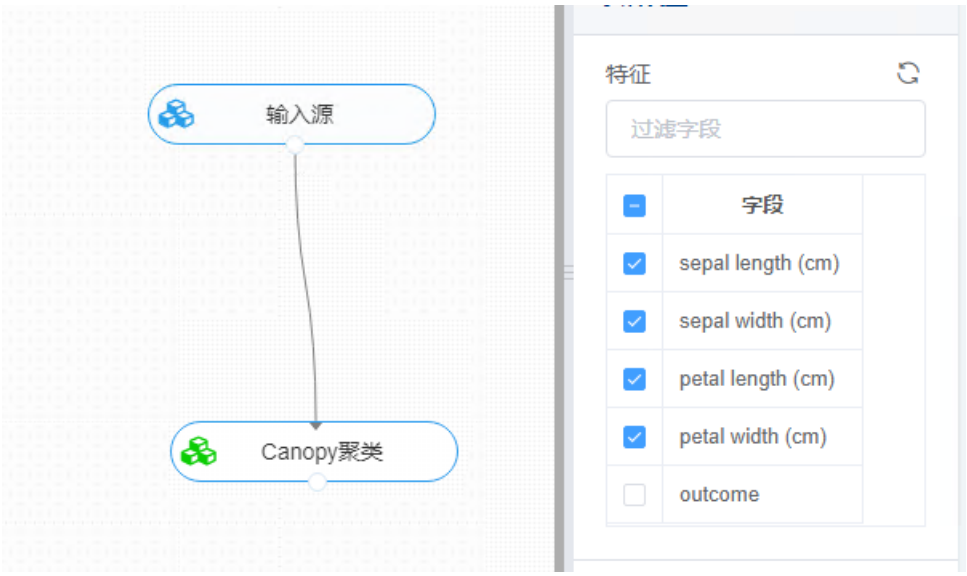
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行EM聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



开始进行Canopy聚类，将数据集分门别类。拖入【Canopy聚类】算法，将【输入源】算法和【Canopy聚类】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，点击参数设置，填写T1和T2的值，右键单击【Canopy聚类】算法，选择“运行该节点”。



序号	参数名称	数值	原因
1	参数设置	T1	T1的值过大，会导致更多的数据会被重复迭代，形成过多的canopy；值过小则导致相反的效果
2	参数设置	T2	T2的值过大，会导致一个canopy中的数据太多，反之则过少

打开日志，查看结果。在日志中可以查看各个聚类分群的个数与占比、每个分群中各个特征的优势与劣势以及聚类中心。对【Canopy】算法右击，点击“查看日志”。

序号	名称	作用
1	饼图	饼图可看出聚类分群个数，以及各个分群的个数和占比。
2	散点图	通过对数据进行降维，在二维空间中展示的聚类结果。

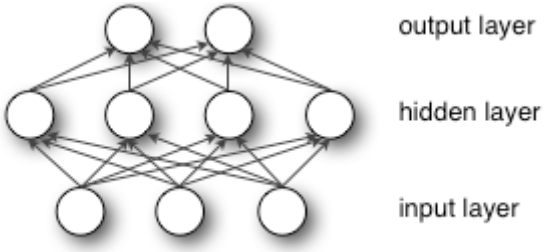
8.4 分类

分类是在一群已经知道类别标号的样本中，训练一种分类器，让其能够对某种未知的样本进行分类。分类算法属于一种有监督的学习。分类算法的分类过程就是建立一种分类模型来描述预定的数据集或概念集，通过分析由属性描述的数据库元组来构造模型。分类的目的就是使用分类对新的数据集进行划分，其主要涉及分类规则的准确性、过拟合、矛盾划分的取舍等。

8.4.1 神经网络

（1）作用及原理

MLP又名多层感知机，也叫人工神经网络（ANN，Artificial Neural Network），它是一个监督算法，可以处理图片、数据等自动分类。除了输入输出层，它中间可以有多个隐藏层，如果没有隐藏层即可解决线性可划分的数据问题。最简单的MLP模型只包含一个隐藏层，即三层的结构，如下图所示。



从上图可以看到，多层感知机的层与层之间是全连接的（全连接的意思就是：上一层的任何一个神经元与下一层的所有神经元都有连接）。多层感知机最底层是输入层，中间是隐藏层，最后是输出层。

输入层即接收输入数据的层。比如输入是一个n维向量，就有n个神经元。

隐藏层的与输入层是全连接的，假设输入层用向量X表示，则隐藏层的输出就是f(W1X+b1)，W1是权重（也叫连接系数），b1是偏置，函数f可以是常用的sigmoid函数或者tanh函数：

因此，MLP所有的参数就是各个层之间的连接权重以及偏置，包括W1、b1、W2、b2。对于一个具体的问题，通常采用梯度下降法（SGD）来确定参数：首先随机初始化所有参数，然后迭代地训练，不断地计算梯度和更新参数，直到满足某个条件为止（比如误差足够小、迭代次数足够多时）。这个过程涉及到代价函数、规则化（Regularization）、学习速率（learning rate）、梯度计算等。

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	影响收敛速度
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	尽量多	

（3）输出

序号	名称	内容
1	data_out.csv	添加了label、predict_label列的原始数据，label为原始类别，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估
3	日志	含有模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	数据元个数设置	隐藏层神经元个数，(100，100)代表有两层
4	基础参数	迭代次数	迭代的次数，数值型
5	基础参数	权重优化器	lbfgs：quasi-Newton方法的优化器 sgd：随机梯度下降 adam：Kingma、Diederik、Jimmy Ba提出的机遇随机梯度的优化器
6	基础参数	激活函数	激活函数,可选identity、logistic、tanh、relu，logistic也就是sigmod。默认为relu
7	基础参数	正则化参数	L2惩罚（正则化项）参数，默认为0.0001

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接将数据集输入神经网络进行训练。

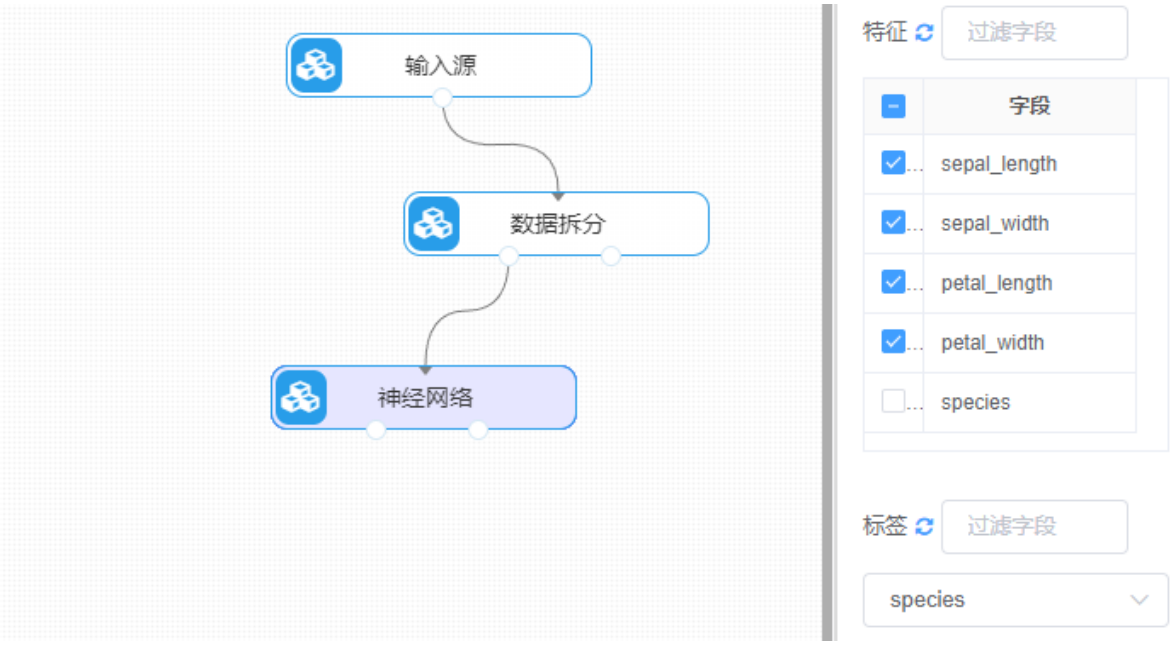
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行训练的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

训练神经网络前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始训练神经网络。拖入【神经网络】算法，将【数据拆分】组件的训练集输出节点和【神经网络】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，数据元个数设置为“100,100”，迭代次数设置为200，权重优化器选择adam，激活函数选择relu，正则化参数设置为0.0001，右键单击【神经网络】算法，选择“运行该节点”。



序号	参数名称	数值	原因
1	数据元个数设置	100,100	神经网络结构，训练数据较为简单设置2层
2	迭代次数	200	迭代次数越多，结果越准确。但随着迭代次数不断增大，或是数据量的增加，运行所需要的时间也就越来越久，所以确定一个最佳的迭代次数非常重要。
3	权重优化器	adam	该优化器效果较好

打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【神经网络】算法右击，点击“查看日志”。

序号	名称	作用
1	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
2	混淆矩阵	可以直观地观测模型每个分类的效果
3	ROC图	计算AUC面积，评价模型训练好坏
4	K-S曲线	可根据曲线距离确定模型的阈值

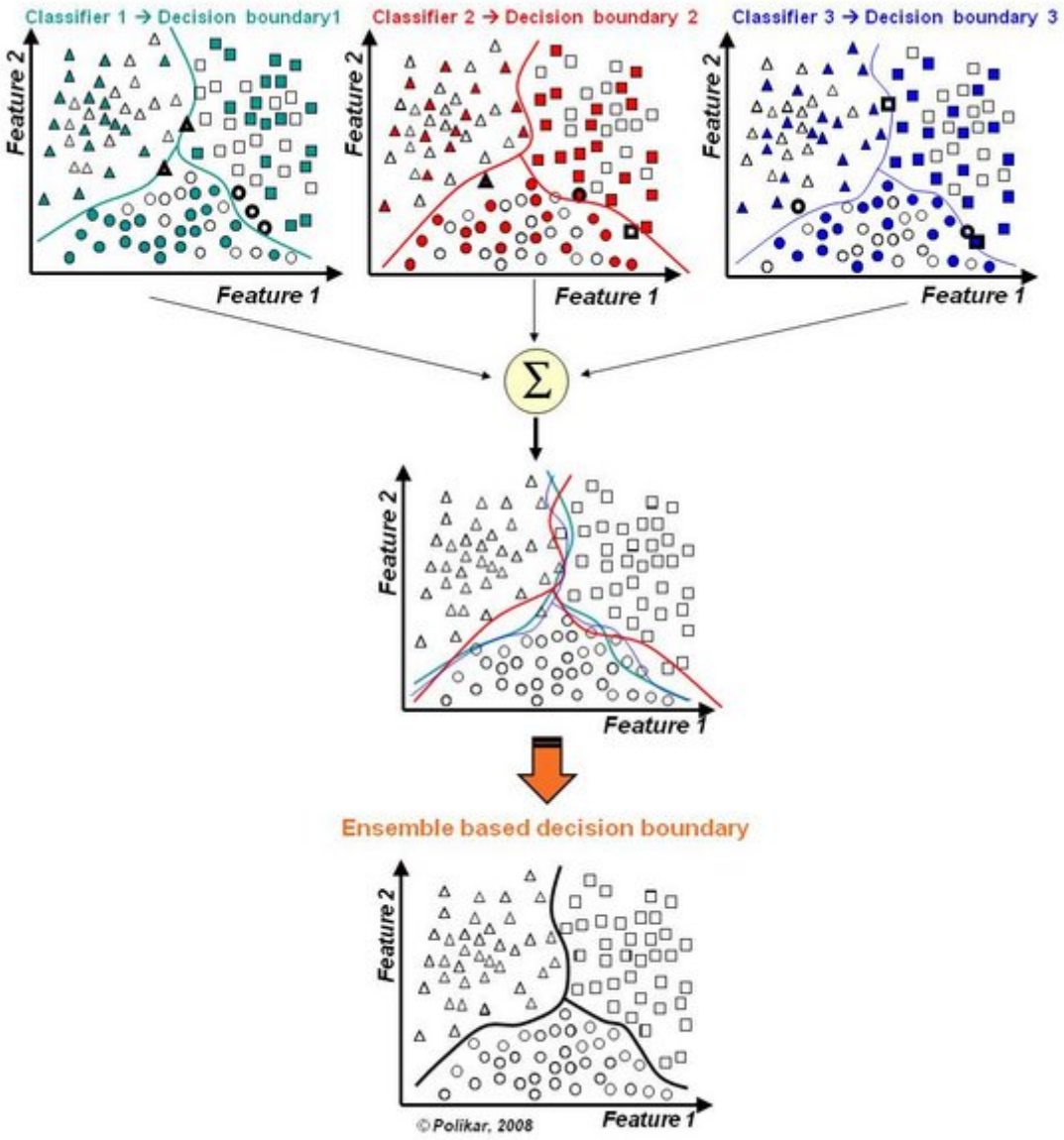
对于训练好的神经网络，还可以使用【模型评估】组件对测试集进行模型评估。【神经网络】的第一个输出节点输出的是训练好的模型，第二个节点输出的是测试集数据。拖入【模型评估】组件，模型输入节点与【神经网络】的第一个输出节点连接，数据输入节点与【数据拆分】的测试集输出节点连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，右键组件运行该节点。

8.4.2 随机森林

（1）作用及原理

随机森林是一个可做能够回归和分类。它具备处理大数据的特性，而且它有助于估计或变量是非常重要的基础数据建模。随机森林是几乎任何预测问题(甚至非直线部分)的固有选择。

随机森林就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树，而它的本质属于机器学习的一大分支——集成学习（Ensemble Learning）方法。从直观角度来解释，每棵决策树都是一个分类器（假设现在针对的是分类问题），那么对于一个输入样本，N棵树会有N个分类结果。而随机森林集成了所有的分类投票结果，将投票次数最多的类别指定为最终的输出，这就是一种最简单的Bagging思想。



（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	否	对特征值大小不敏感
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	尽量多	

（3）输出

序号	名称	内容
1	data_out.csv	添加了label、predict_label列的原始数据，label为原始类别，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估
3	日志	含有模型参数、模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	树的个数	森林里（决策）树的数目，数值型
4	基础参数	树的分裂规则	衡量分裂质量的性能（函数）。受支持的标准是基尼不纯度的"gini"，和信息增益的"entropy"（熵）。
5	基础参数	树的最大深度	（决策）树的最大深度。如果值为None，那么树会生长到所有叶子都分到一个类，或者某节点所代表的样本数已小于min_samples_split
6	基础参数	最小分裂样本数	分割内部节点所需要的最小样本数量
7	基础参数	最小叶子节点样本数	需要在叶子结点上的最小样本数量

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行随机森林分类。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

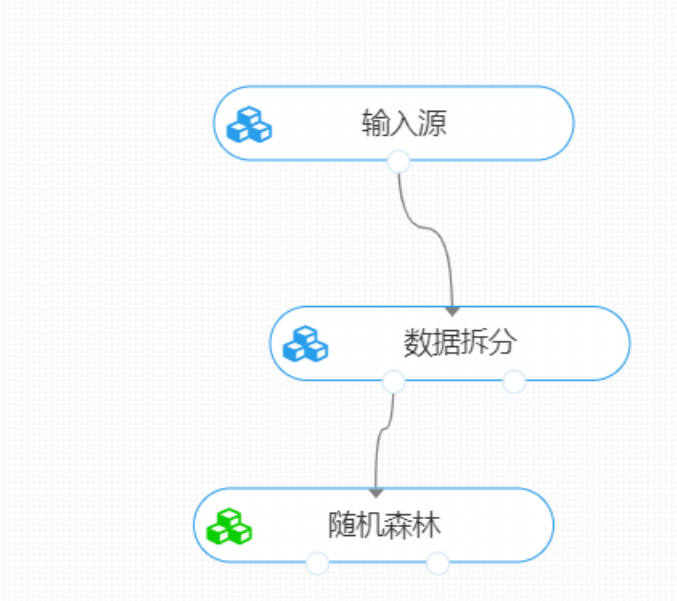
iris

首先将需要进行分类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行分类前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始进行随机森林分类。拖入【随机森林】算法，将【数据拆分】组件的训练集输出节点和【随机森林】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，树的个数设置为10个，树分裂的规则选择gini，树的最大深度设置为0，最小分裂样本数设置为2，最小叶子节点样本数设置为1，右键单击【随机森林】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	树的个数	10	较多的子树可以让模型有更好的性能，但同时让代码运行变慢。应该选择尽可能高的值，只要运行环境能够承受的住，因为这能使预测更好更稳定。
2	树的最大深度	0	设置值为None，树会生长到所有叶子都分到一个类，或者某节点所代表的样本数已小于min_samples_split。
3	最小分裂样本数	2	分裂所需的最小样本数，影响性能。
4	最小叶子节点样本数	1	叶节点最小样本数，影响性能。



打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【随机森林】算法右击，点击“查看日志”。

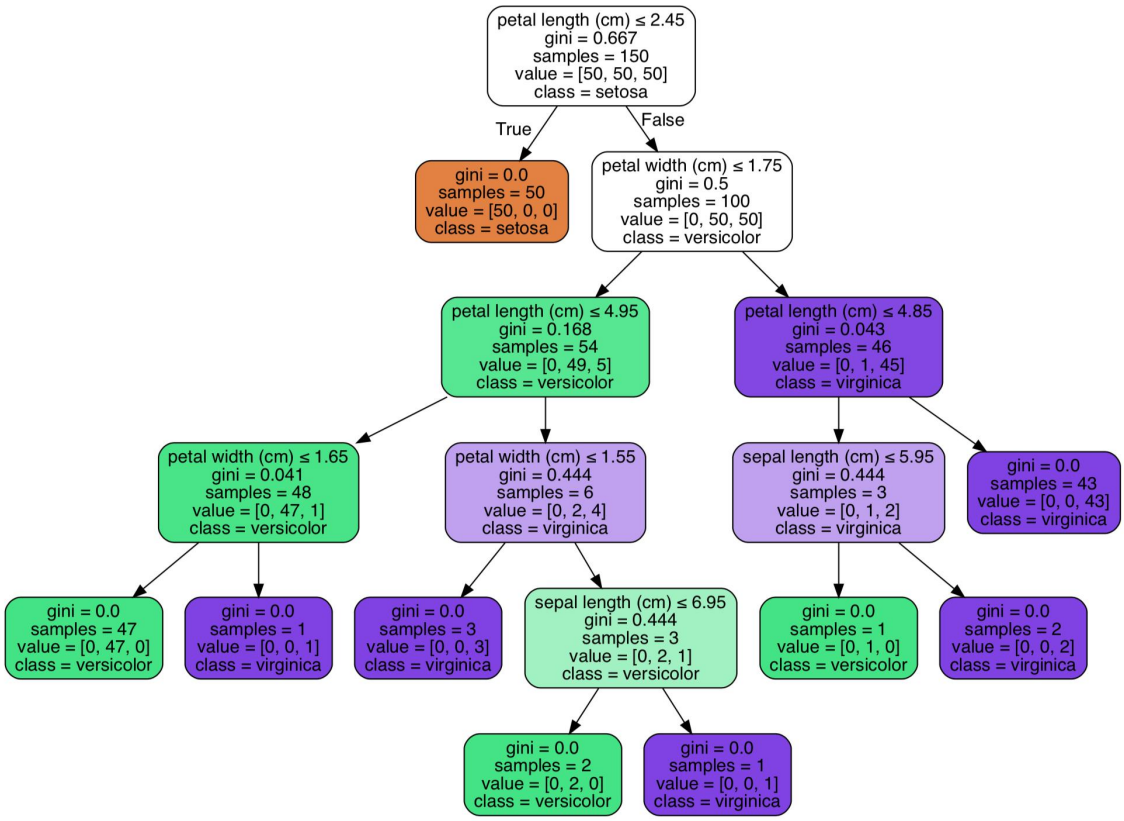
序号	名称	作用
1	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
2	混淆矩阵	可以直观地观测模型每个分类的效果
3	ROC图	计算AUC面积，评价模型训练好坏
4	随机森林决策树	可视化其分类的每个过程

8.4.3 CART分类树

（1）作用及原理

CART全称叫Classification and Regression Tree。它与随机森林类似，都是基于决策树的算法，能对大数据进行分类以及回归。

其原理是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。



（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	否	对特征值大小不敏感
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	尽量多	

（3）输出

序号	名称	内容
1	data_out.csv	添加了label、predict_label列的原始数据，label为原始类别，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估。
3	日志	含有模型参数、模型的特征的重要性信息、模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	特征选择标准	特征选择标准，默认gini，即CART算法。
4	基础参数	特征划分点选择标准	样本的特征划分标准，默认为best
5	基础参数	最大深度	决策树最大深度，数值型
6	基础参数	内部节点最小样本数	内部节点（即判断条件）再划分所需最小样本数，数值型
7	基础参数	叶子节点最小样本数	叶子节点（即分类）最少样本数，数值型
8	基础参数	叶子节点最小的样本权重和	叶子节点（即分类）最小的样本权重和，数值型

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行CART算法分类。

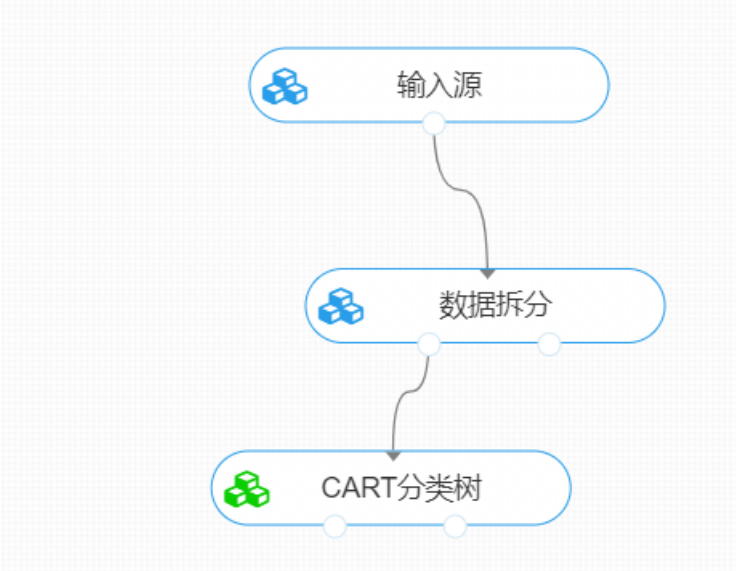
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行分类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行分类前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始进行CART算法分类。拖入【CART分类树】算法，将【数据拆分】组件的训练集输出节点和【CART分类树】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，特征选择标准选择基尼系数，特征划分点选择选择best，最大深度填写None，内部节点最小样本数设置为2，叶子节点最小样本数设置为1，叶子节点最小的样本权重和设置为0，右键单击【CART分类树】算法，选择“运行该节点”。



序号	参数名称	序号	原因
1	特征选择标准	基尼系数	可选项有gini、entropy，前者是基尼系数，后者是信息熵。
2	特征划分点选择标准	best	可选项有best、random，前者是在所有特征中寻找最好的切分点，后者是在部分特征中寻找，默认的“best”适合样本量不大的时候，而如果样本数据量非常大，此时决策树构建推荐“random”。
3	最大深度	None	填写int类型数值或None。设置决策树的最大深度，深度越大，越容易过拟合，推荐树的深度为：5-20之间。
4	内部节点最小样本数	2	设置结点的最小样本数量，当样本数量可能小于此值时，结点将不会在划分。
5	叶子节点最小样本数	1	这个值限制了叶子节点最少的样本数，如果某叶子节点数目小于样本数，则会和兄弟节点一起被剪枝。
6	叶子节点最小的样本权重和	0	这个值限制了叶子节点所有样本权重和的最小值，如果小于这个值，则会和兄弟节点一起被剪枝。默认是0，就是不考虑权重问题。

打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【CART分类树】算法右击，点击“查看日志”。

序号	名称	作用
1	模型的特征的重要性信息	输出模型的特征信息
2	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
3	混淆矩阵	可以直观地观测模型每个分类的效果
4	ROC图	计算AUC面积，评价模型训练好坏

8.4.4 逻辑回归

（1）作用及原理

逻辑回归（Logistic Regression）是一种用于解决监督学习（Supervised Learning）问题的学习算法，一般用于二分类（Binary Classification）问题中，用逻辑回归训练得到一个分类器，对输入的数据进行判断其类型，并进行概率的估计；逻辑回归的目的，是使训练数据的标签值与预测出来的值之间的误差最小化。

logistic回归是一种广义线性回归（generalized linear model），因此与多重线性回归分析有很多相同之处。它们的模型形式基本上相同，都具有wx+b，其中w和b是待求参数，其区别在于他们的因变量不同，多重线性回归直接将wx+b作为因变量，即y=wx+b，而logistic回归则通过函数L将wx+b对应一个隐状态p，p=L(wx+b),然后根据p与1-p的大小决定因变量的值。如果L是logistic函数，就是logistic回归，如果L是多项式函数就是多项式回归。

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	否	对特征值大小不敏感
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	尽量多	

（3）输出

序号	名称	内容
1	data_out.csv	添加了predict_label列的原始数据，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估
3	日志	含有模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	对于多分类问题的策略	分类方式选择参数，str类型，默认为ovr。
4	基础参数	优化算法	solver参数决定了对逻辑回归损失函数的优化方法。
5	基础参数	正则化系数的倒数	正则化系数λ的倒数，float类型，默认为1.0。
6	基础参数	惩罚项	惩罚项，str类型，默认为l2。

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接将数据集输入逻辑回归模型进行训练。

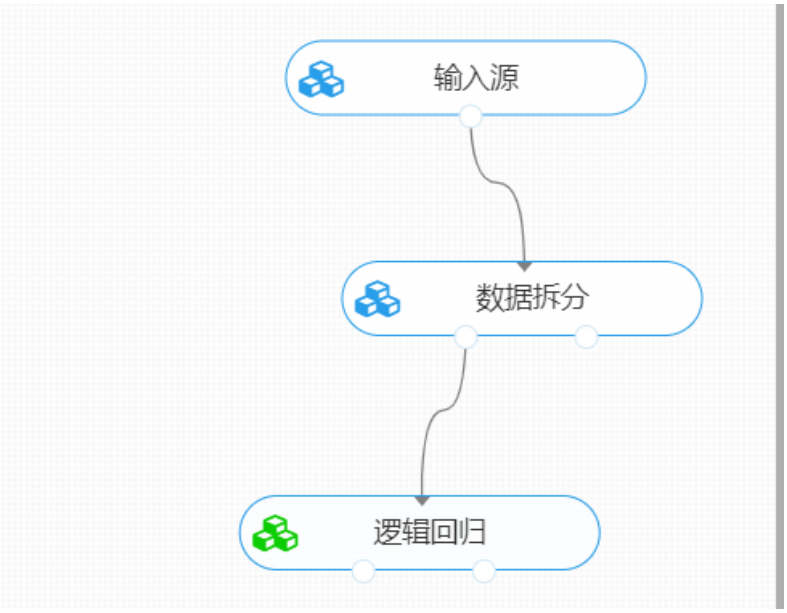
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行训练的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

训练逻辑回归模型前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始训练逻辑回归模型。拖入【逻辑回归】算法，将【数据拆分】组件的训练集输出节点和【逻辑回归】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，对于多分类问题的策略选择ovr，优化算法选择liblinear，正则化系数的倒数设置为1，惩罚项选择L2，右键单击【逻辑回归】算法，选择“运行该节点”。



序号	参数名称	数值	原因
1	对于多分类问题的策略	ovr	分类方式选择参数，可选参数为ovr和multinomial，默认为ovr。ovr即one-vs-rest，multinomial即many-vs-many(MvM)。如果是二元逻辑回归，ovr和multinomial并没有任何区别，区别主要在多元逻辑回归上。
2	优化算法	liblinear	solver参数决定了对逻辑回归损失函数的优化方法，有五个可选参数，即newton-cg,lbfgs,liblinear,sag,saga。默认为liblinear。liblinear：使用了开源的liblinear库实现，内部使用了坐标轴下降法来迭代优化损失函数。lbfgs：拟牛顿法的一种，利用损失函数二阶导数矩阵即海森矩阵来迭代优化损失函数。 newton-cg：也是牛顿法家族的一种，利用损失函数二阶导数矩阵即海森矩阵来迭代优化损失函数。sag：即随机平均梯度下降，是梯度下降法的变种，和普通梯度下降法的区别是每次迭代仅仅用一部分的样本来计算梯度，适合于样本数据多的时候。 saga：线性收敛的随机优化算法的的变重。
3	正则化系数的倒数	1.0	正则化系数λ的倒数，默认为1.0。必须是正浮点型数,越小的数值表示越强的正则化。
4	惩罚项	L2	惩罚项，可选参数为l1和l2，默认为l2。用于指定惩罚项中使用的规范。newton-cg、sag和lbfgs求解算法只支持L2规范。L1规范假设的是模型的参数满足拉普拉斯分布，L2假设的模型参数满足高斯分布，所谓的范式就是加上对参数的约束，使得模型更不会过拟合(overfit)，但是如果要说是不是加了约束就会好，只能说加了约束的情况下，理论上应该可以获得泛化能力更强的结果。

打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【逻辑回归】算法右击，点击“查看日志”。

序号	名称	作用
1	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
2	混淆矩阵	可以直观地观测模型每个分类的效果
3	ROC图	计算AUC面积，评价模型训练好坏

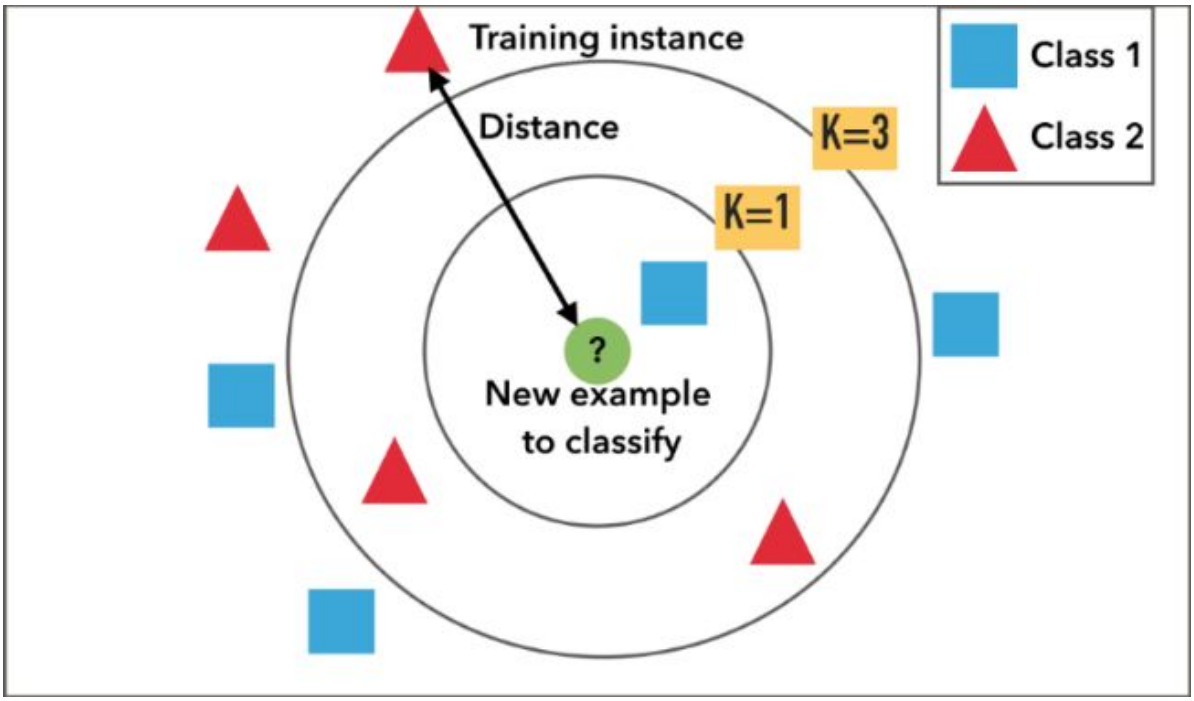
对于训练好的逻辑回归模型，还可以使用【模型评估】组件对测试集进行模型评估。【逻辑回归】的第一个输出节点输出的是训练好的模型，第二个节点输出的是测试集数据。拖入【模型评估】组件，第一个输入节点与【逻辑回归】的模型输出节点连接，第二个输入节点与【数据拆分】的测试集输出节点连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，右键组件运行该节点。

8.4.5 K最近邻

（1）作用及原理

K最近邻（K-Nearest Neighbor，KNN）分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：在特征空间中，如果一个样本附近的k个最近(即特征空间中最邻近)样本的大多数属于某一个类别，则该样本也属于这个类别。

其原理是，给定一个训练数据集，对新的输入实例，在训练数据集中找到与该实例最邻近的K个实例（也就是上面所说的K个邻居），这K个实例的多数属于某个类，就把该输入实例分类到这个类中。



（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	对特征值大小敏感
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	少量	涉及距离计算量

（3）输出

序号	名称	内容
1	data_out.csv	添加了predict_label列的原始数据，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估
3	日志	含有模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	K值	查询的默认邻居的数量，数值型
4	基础参数	投票权重类型	预测中使用的权重函数
5	基础参数	算法	用于计算最近邻居的算法
6	基础参数	树叶的大小	传入BallTree或者KDTree算法的叶子数量，数值型

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接将数据集输入K最近邻模型进行训练。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

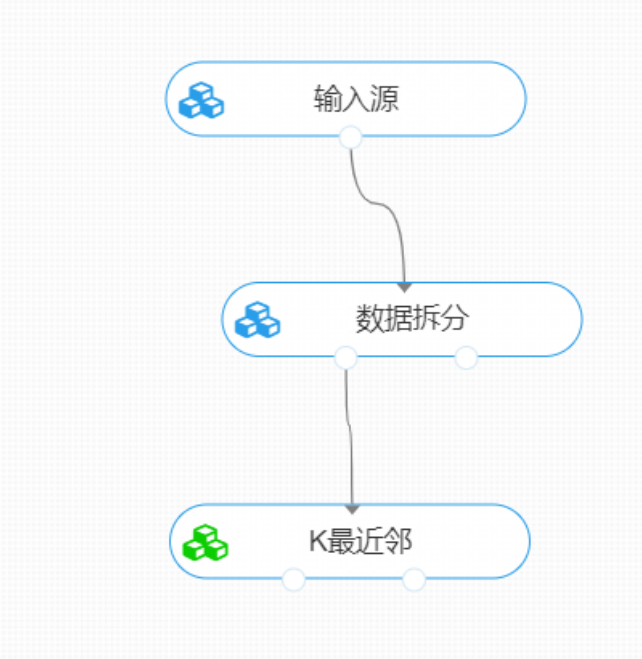
iris

首先将需要进行训练的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

训练K最近邻模型前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始进行K最近邻分类。拖入【K最近邻】算法，将【数据拆分】组件的训练集输出节点和【K最近邻】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，K值设置为5，投票权重类型选择“等权加权”，算法选择auto，树叶的大小设置为30，右键单击【K最近邻】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	K值	5	寻找的邻居数，默认是5。
2	投票权重类型	等权加权	预测中使用的权重函数。可能的取值：“等权加权”：统一权重，即每个邻域中的所有点均被加权。“距离加权”：权重点与其距离的倒数，在这种情况下，查询点的近邻比远处的近邻具有更大的影响力。
3	算法	auto	“ball_tree”将使用BallTree,“kd_tree”将使用KDTree,“brute”将使用暴力搜索。“auto”将尝试根据传递给fit方法的值来决定最合适的算法。注意：在稀疏输入上进行拟合将使用蛮力覆盖此参数的设置。
4	树叶的大小	30	传入BallTree或者KDTree算法的叶子数量，这会影响构造和查询的速度，以及存储树所需的内存。最佳值取决于问题的性质，默认30



打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【K最近邻】算法右击，点击“查看日志”。

序号	名称	作用
1	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
2	混淆矩阵	可以直观地观测模型每个分类的效果
3	ROC图	计算AUC面积，评价模型训练好坏
4	KS曲线	可根据曲线距离确定模型阈值

对于训练好的K最近邻模型，还可以使用【模型评估】组件对测试集进行模型评估。【K最近邻】的第一个输出节点输出的是训练好的模型，第二个节点输出的是测试集数据。拖入【模型评估】组件，第一个输入节点与【K最近邻】的模型输出节点连接，第二个输入节点与【数据拆分】的测试集输出节点连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，右键组件运行该节点。

8.4.6 朴素贝叶斯

（1）作用及原理

朴素贝叶斯是以贝叶斯原理为基础，使用概率统计的知识对样本数据集进行分类。在sklearn中，一共有3个朴素贝叶斯的分类算法类。分别是GaussianNB，MultinomialNB和BernoulliNB。这三个类适用的分类场景各不相同，一般来说，如果样本特征的分布大部分是连续值，使用GaussianNB会比较好。如果样本特征的分布大部分是多元离散值，使用MultinomialNB比较合适。而如果样本特征是二元离散值或者很稀疏的多元离散值，应该使用BernoulliNB。

其原理是，以贝叶斯定理为基础并且假设特征条件之间相互独立，通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入X求出使得后验概率最大的输出Y。

- GaussianNB假设特征的先验概率为正态分布，即如下式：

$$P(X_j|Y=C_k)=\frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left(-\frac{(x_j-\mu_k)^2}{2\sigma_k^2}\right)$$

其中*Ck*为Y的第k类类别。*μk*和*σk^2*为需要从训练集估计的值。

GaussianNB会根据训练集求出*μk*和*σk^2*。*μk*为在样本类别*Ck*中，所有*Xj(j=1,2,3...)*的平均值。*σk^2*为在样本类别*Ck*中，所有*Xj(j=1,2,3...)*的方差。

- MultinomialNB假设特征的先验概率为多项式分布，即如下式：

$$P(X_j=x_{jl}|Y=C_k)=\frac{x_{jl}+\lambda}{m_k+n\lambda}$$

其中，*P*是第*k*个类别的第*j*维特征的第*l*个取值条件概率。*mk*是训练集中输出为第*k*类的样本个数。*λ*为一个大于0的常数，常常取为1，即拉普拉斯平滑，也可以取其他值。

- BernoulliNB假设特征的先验概率为二元伯努利分布，即如下式：

$$P(X_j=x_{jl}|Y=C_k)=P(j|Y=C_k)x_{jl}+(1-P(j|Y=C_k))(1-x_{jl})$$

此时只有两种取值。*xjl*只能取值0或者1。

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	否	通过概率评判
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	无限制	

（3）输出

序号	名称	内容
1	data_out.csv	添加了predict_label列的原始数据，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估
3	日志	含有模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	类函数	选择朴素贝叶斯的分类算法

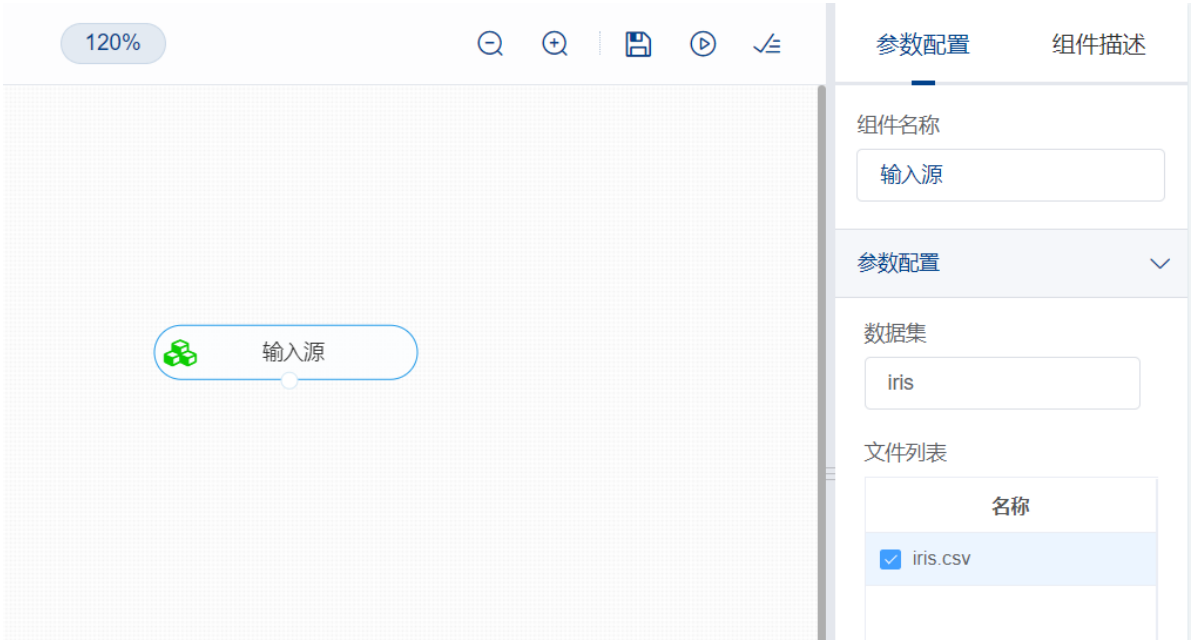
（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接将数据集输入朴素贝叶斯模型进行训练。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行训练的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



进行分类前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始进行朴素贝叶斯分类。拖入【朴素贝叶斯】算法，将【数据拆分】组件的训练集输出节点和【朴素贝叶斯】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，类函数选择高斯朴素贝叶斯，右键单击【朴素贝叶斯】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	类函数	高斯朴素贝叶斯	如果样本特征的分布大部分是连续值，使用GaussianNB会比较好； 如果样本特征的分布大部分是多元离散值，使用MultinomialNB比较合适； 而如果样本特征是二元离散值或者很稀疏的多元离散值，应该使用BernoulliNB。

打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【朴素贝叶斯】算法右击，点击“查看日志”。

序号	名称	作用
1	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
2	混淆矩阵	可以直观地观测模型每个分类的效果
3	ROC图	计算AUC面积，评价模型训练好坏
4	KS曲线	可根据曲线的距离确定模型的阈值

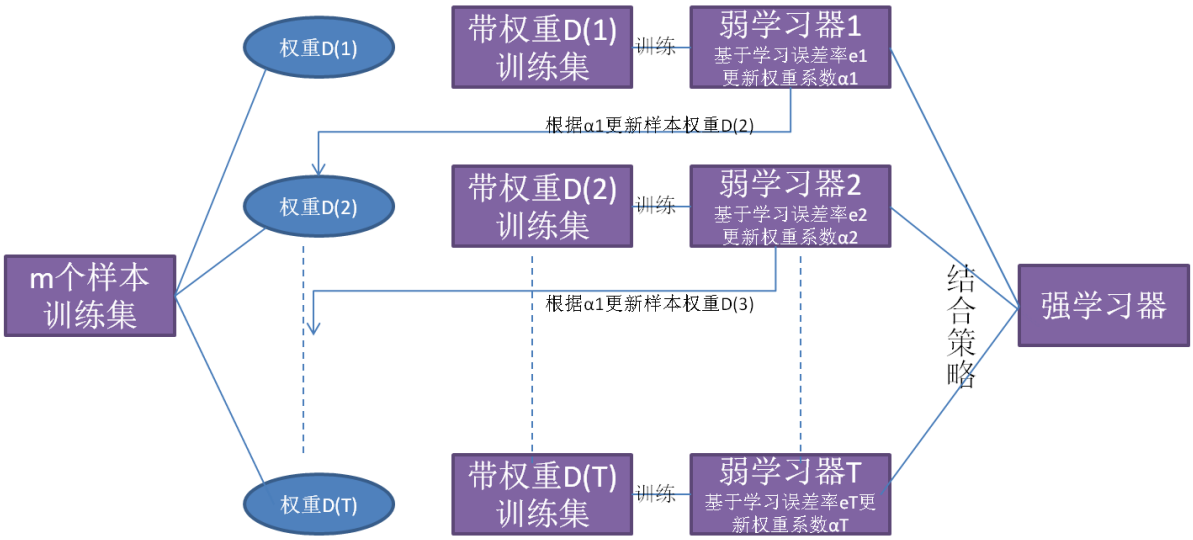
对于训练好的朴素贝叶斯模型，还可以使用【模型评估】组件对测试集进行模型评估。【朴素贝叶斯】的第一个输出节点输出的是训练好的模型，第二个节点输出的是测试集数据。拖入【模型评估】组件，第一个输入节点与【朴素贝叶斯】的模型输出节点连接，第二个输入节点与【数据拆分】的测试集输出节点连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，右键组件运行该节点。

8.4.7 Adaboost

（1）作用及原理

Adaboost(AdaptiveBoosting)是一种迭代算法，通过对训练集不断训练弱分类器，然后把这些弱分类器集合起来，构成强分类器。

Adaboost算法训练的过程中，初始化所有训练样例的具有相同的权值重，在此样本分布下训练出一个弱分类器，针对错分样本加大对其对应的权值，分类正确的样本降低其权值，使前一步被错分的样本得到突显，获得新的样本分布，在新的样本分布下，再次对样本进行训练，又得到一个分类器。依次循环，得到T个分类器，将这些分类器按照一定的权值组合，得到最终的强分类器。训练的关键是针对比较难分的训练样本，在联合弱分类器时，使用加权投票，这样分类效果好的弱分类器得到较大的权重，分类效果差的则权值较小。



（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	使模型更准确
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	无限制	

（3）输出

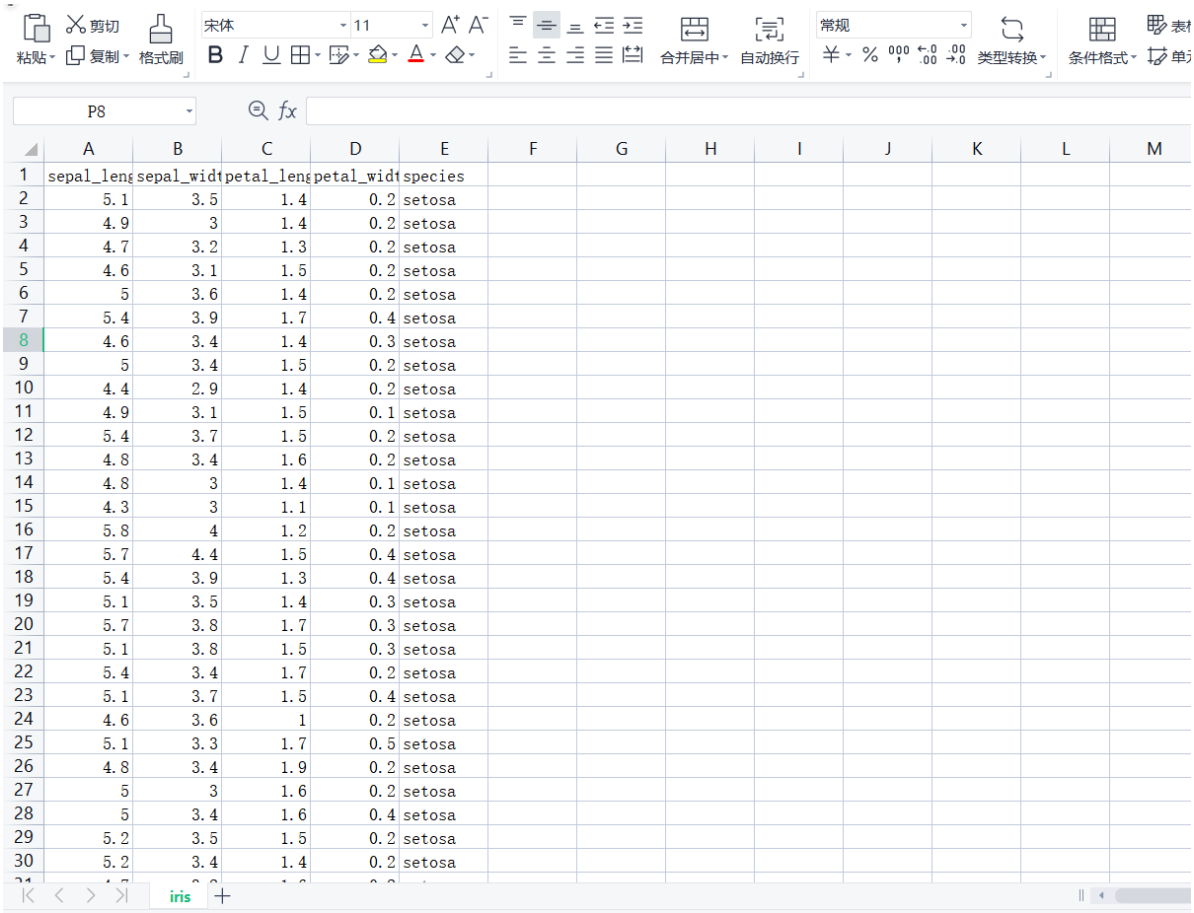
序号	名称	内容
1	data_out.csv	添加了predict_label列的原始数据，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估
3	日志	含有模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	最大迭代次数	弱学习器的最大迭代次数，默认为50，数值型
4	基础参数	模型提升准则	弱学习器权重的度量，有两种方式SAMME和SAMME.R，默认为SAMME.R
5	基础参数	poly函数维度	梯度收敛速度，默认为1，数值型

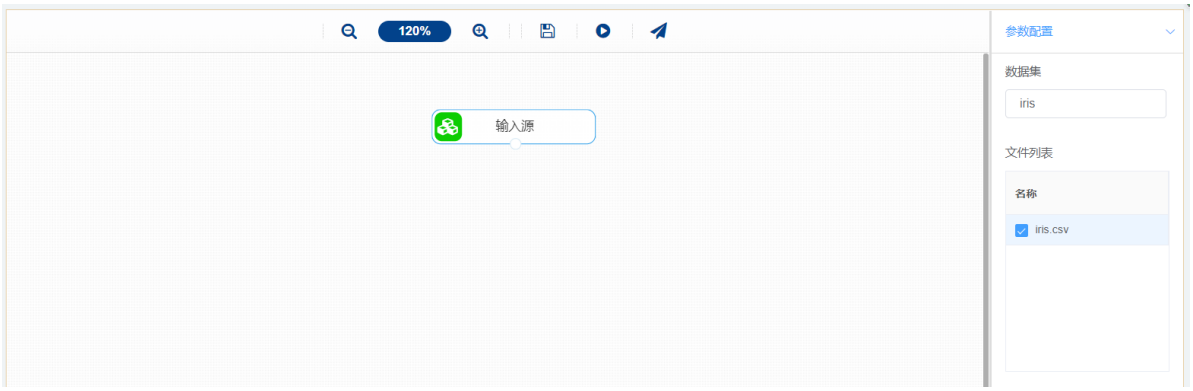
（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接将数据集输入Adaboost模型进行训练。

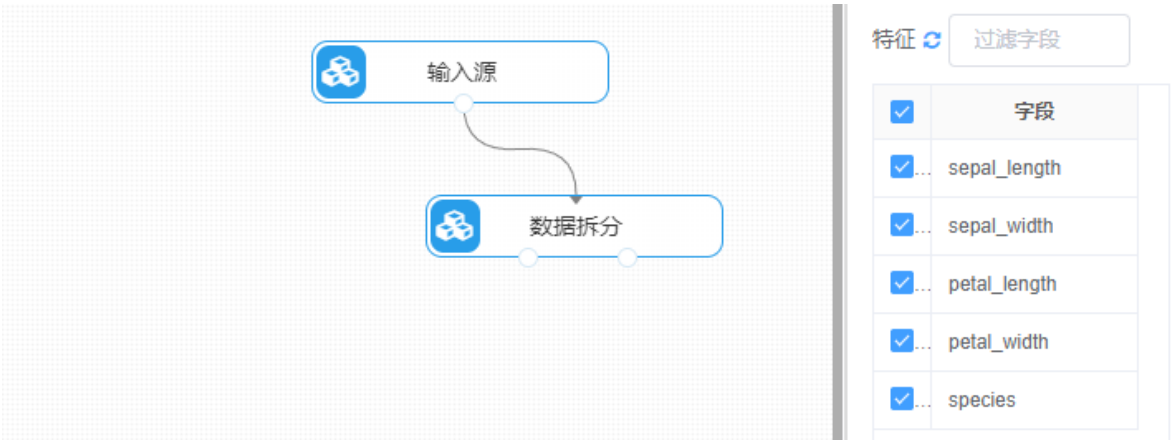


	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa
14	4.8	3	1.4	0.1	setosa
15	4.3	3	1.1	0.1	setosa
16	5.8	4	1.2	0.2	setosa
17	5.7	4.4	1.5	0.4	setosa
18	5.4	3.9	1.3	0.4	setosa
19	5.1	3.5	1.4	0.3	setosa
20	5.7	3.8	1.7	0.3	setosa
21	5.1	3.8	1.5	0.3	setosa
22	5.4	3.4	1.7	0.2	setosa
23	5.1	3.7	1.5	0.4	setosa
24	4.6	3.6	1	0.2	setosa
25	5.1	3.3	1.7	0.5	setosa
26	4.8	3.4	1.9	0.2	setosa
27	5	3	1.6	0.2	setosa
28	5	3.4	1.6	0.4	setosa
29	5.2	3.5	1.5	0.2	setosa
30	5.2	3.4	1.4	0.2	setosa

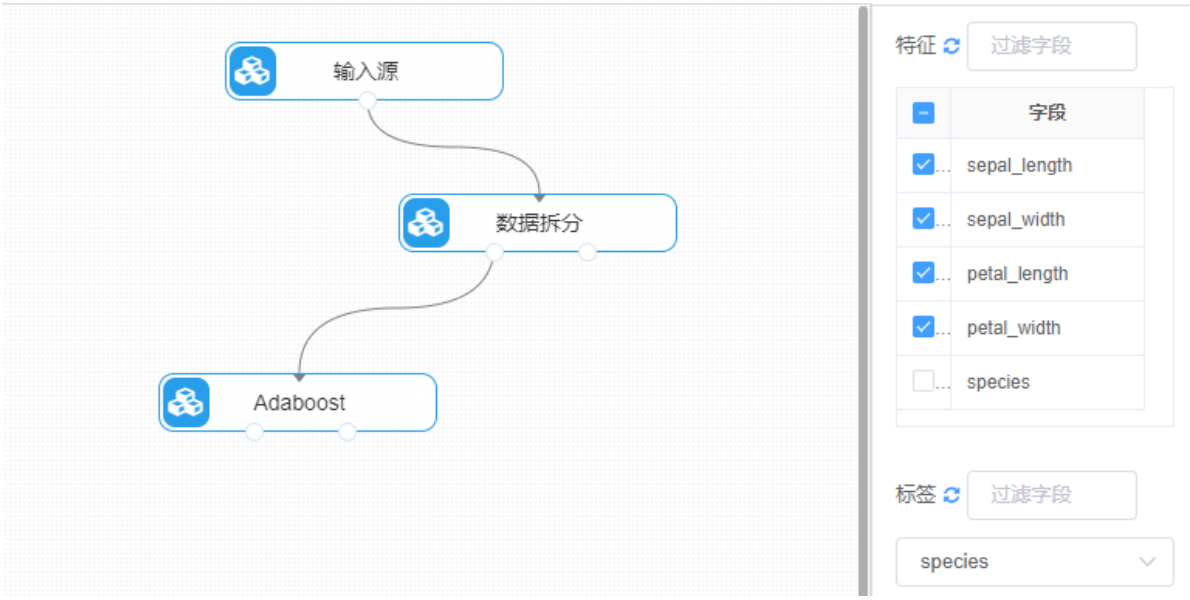
首先将需要进行训练的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



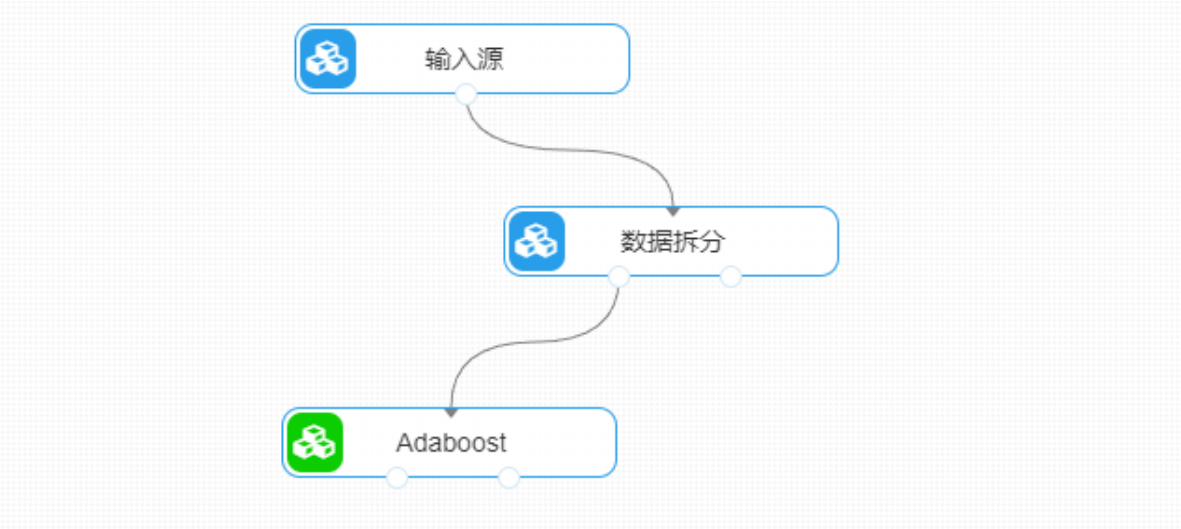
进行分类前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。



开始进行Adaboost分类。拖入【Adaboost】算法，将【数据拆分】组件的训练集输出节点和【Adaboost】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“species”字段，点击参数设置，最大迭代次数设置为50，模型提升准则选择SAMMA.R，poly函数维度设置为1，右键单击【Adaboost】算法，选择“运行该节点”。



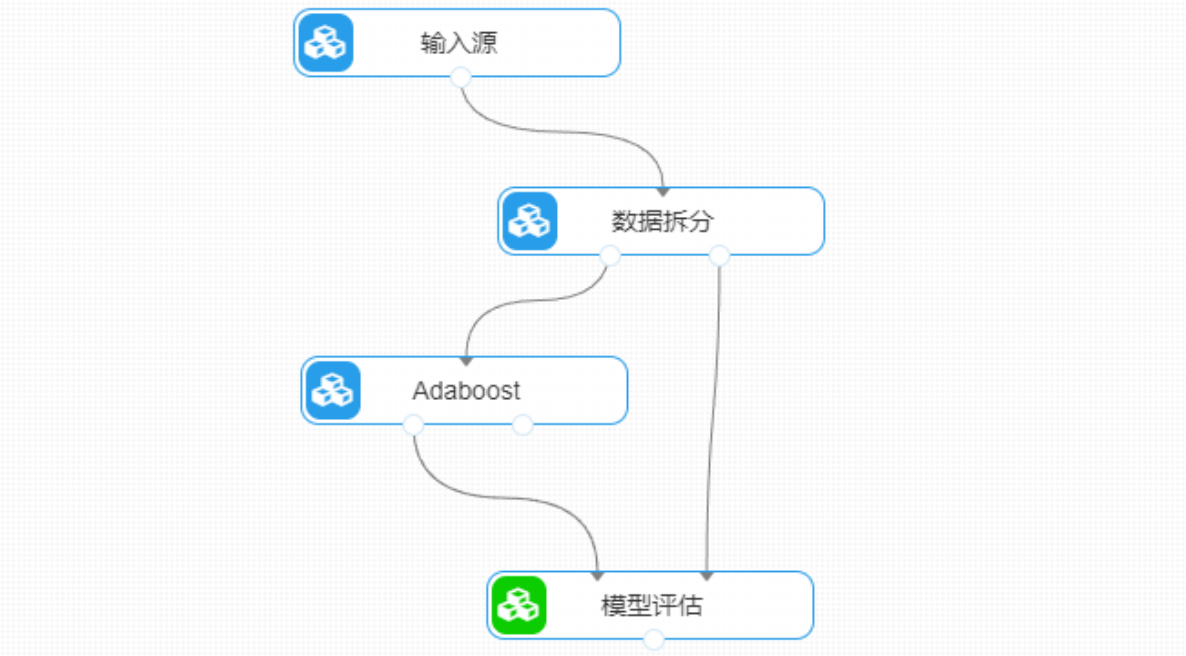
序号	参数名称	数值	原因
1	最大迭代次数	50	弱学习器的最大迭代次数。默认是50。太小容易欠拟合，太大容易过拟合。
2	模型提升准则	MME.R	两者的主要区别是弱学习器权重的度量，SAMME用对样本集分类效果作为弱学习器权重，而SAMME.R使用了对样本集分类的预测概率大小来作为弱学习器权重。由于SAMME.R使用了概率度量的连续值，迭代一般比SAMME快，因此AdaBoostClassifier的默认算法algorithm的值也是SAMME.R。我们一般使用默认的SAMME.R就够了，但是要注意的是使用了SAMME.R，则弱分类学习器参数base_estimator必须限制使用支持概率预测的分类器。SAMME算法则没有这个限制。
3	poly函数维度	1	学习率，表示梯度收敛速度，默认为1，如果过大，容易错过最优值，如果过小，则收敛速度会很慢；该值需要和迭代次数进行一个权衡，当分类器迭代次数较少时，学习率可以小一些，当迭代次数较多时，学习率可以适当放大。



打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【Adaboost】算法右击，点击“查看日志”。

序号	名称	作用
1	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
2	混淆矩阵	可以直观地观测模型每个分类的效果
3	ROC图	计算AUC面积，评价模型训练好坏

对于训练好的Adaboost模型，还可以使用【模型评估】组件对测试集进行模型评估。【Adaboost】的第一个输出节点输出的是训练好的模型，第二个节点输出的是测试集数据。拖入【模型评估】组件，第一个输入节点与【Adaboost】的模型输出节点连接，第二个输入节点与【数据拆分】的测试集输出节点连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“species”字段，右键组件运行该节点。

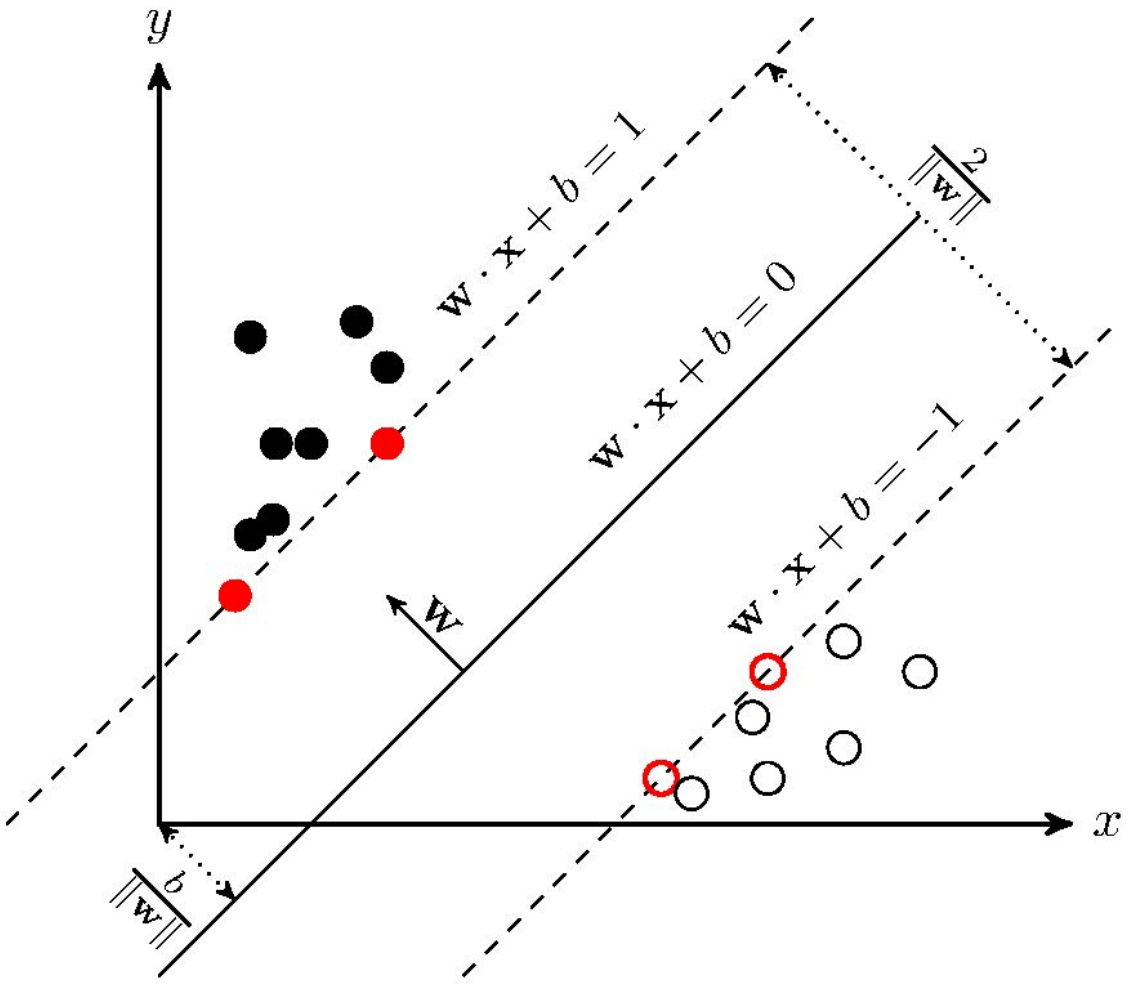


8.4.8 支持向量机

（1）作用及原理

支持向量机（support vector machines，SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM还包括核技巧，这使它成为实质上的非线性分类器。SVM的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM的的学习算法就是求解凸二次规划的最优化算法。

SVM学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下面图所示，即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的分离超平面却是唯一的。



（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	数据异常会影响结果
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	无限制	

（3）输出

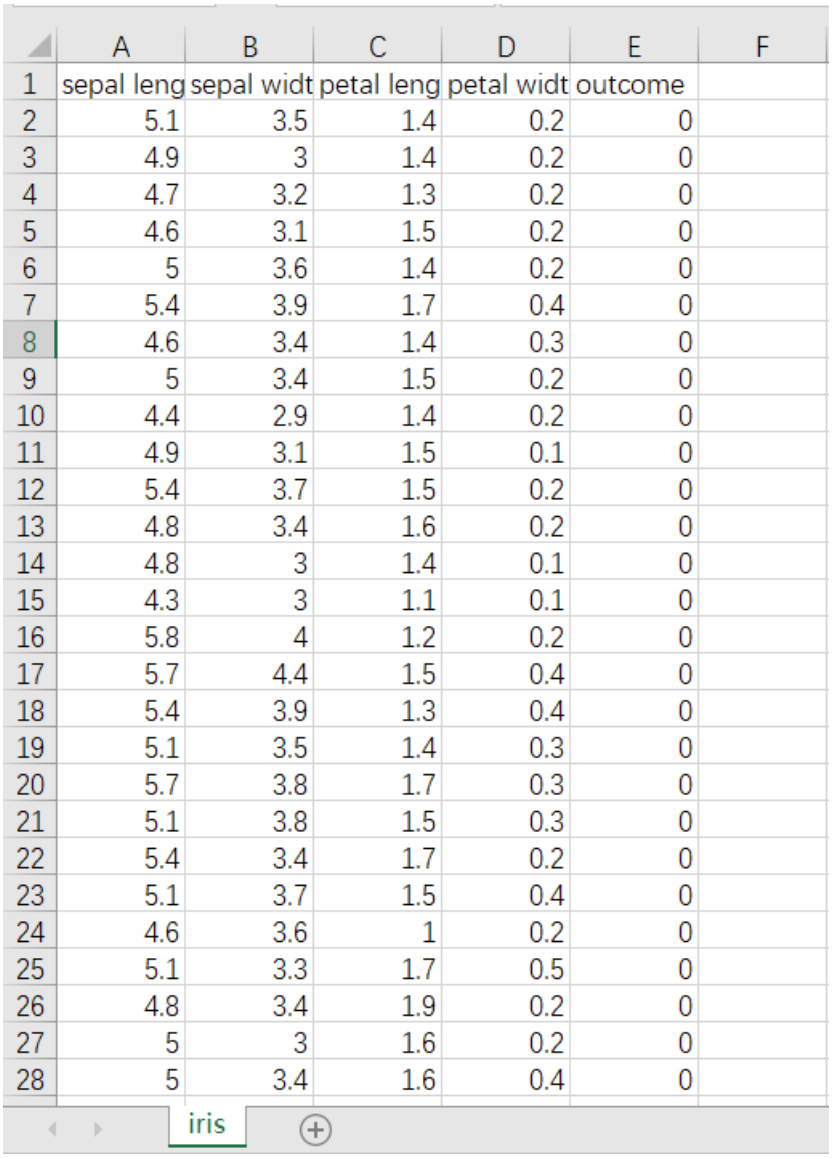
序号	名称	内容
1	data_out.csv	添加了predict_label列的原始数据，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估
3	日志	含有模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	惩罚参数	惩罚系数，用来控制损失函数的惩罚系数,数值型
4	基础参数	核函数	核函数类型，默认为RBF函数
5	基础参数	poly函数维度	多项式poly函数的维度，数值型，默认是3
6	基础参数	核函数参数	rbf、多项式和sigmoid的核函数参数。默认是auto

（5）示例

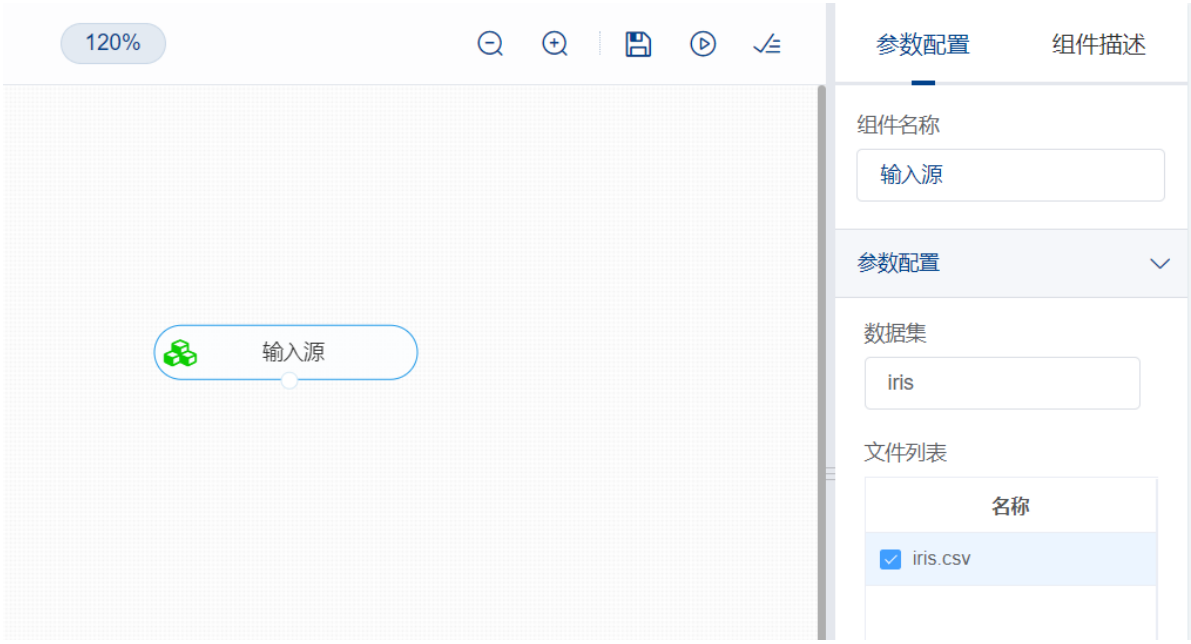
对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接将数据集输入支持向量机模型进行训练。



	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行训练的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



训练支持向量机模型前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始训练支持向量机模型。拖入【支持向量机】算法，将【数据拆分】组件的训练集输出节点和【支持向量机】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，惩罚参数设置为1，核函数选择线性函数，核函数参数输入auto，右键单击【支持向量机】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	惩罚参数	1	这是一个软间隔分类器，对于在边界内的点有惩罚系数C，C的取值在0-1之间，默认值为1.0。C越大代表这个分类器对在边界内的噪声点的容忍度越小，分类准确率高，但是容易过拟合，泛化能力差。所以一般情况下，应该适当减小C，对在边界范围内的噪声有一定容忍。
2	核函数	RBF函数	核函数类型，默认为'rbf'，高斯核函数 其他可选项有：'linear':线性核函数 'poly':多项式核函数 'sigmoid':sigmoid核函数
3	poly函数维度	3	多项式核的阶数，默认为3，对其他核函数不起作用。
4	核函数参数	auto	核函数系数，只对rbf、poly、sigmoid起作用。默认为auto，此时值为样本特征数的倒数，即1/n_features。

打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【支持向量机】算法右击，点击“查看日志”。

序号	名称	作用
1	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
2	混淆矩阵	可以直观地观测模型每个分类的效果
3	ROC图	计算AUC面积，评价模型训练好坏

8.4.9 决策树-ID3

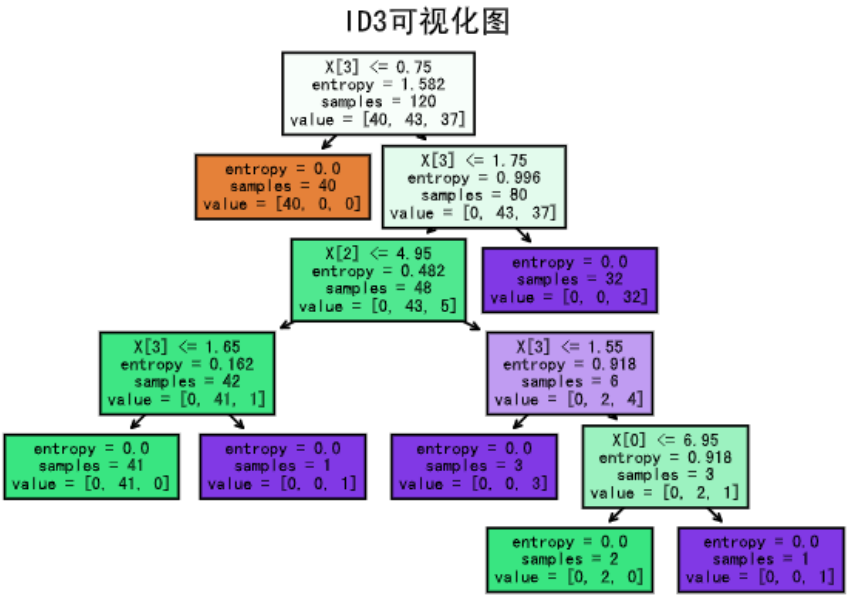
1、作用与原理

决策树是机器学习方法中的一种监督学习算法，表示根据特征对样本进行分类的树形结构，可以用于分类和回归。

从根节点开始，按照训练数据的每个特征进行计算，根据每个特征的不确定性将训练数据分配到其子节点（分支），沿着该分支可能达到叶子节点或者到达另一个内部节点，然后对剩余的特征递归执行下去，直到抵达一个叶子节点。当都到达叶子节点时，我们便得到了最终的分类结果。把这种决策分支画成图形很像一棵树的枝干，也就是决策树。

信息熵是度量随机变量的不确定性。在分类问题中的意义：信息熵表示分类的不确定性。样本集纯度越高，熵越小;反之，成分越复杂，纯度越低，则熵越大。

在决策树中，信息增益作为决策树选择特征（ID3算法）的衡量指标，目的是为了要建立一个能够准确分类而且尽可能矮的树。在建立决策树的过程中，一个特征的信息增益越大，表明特征对样本的熵减少的能力越强，这个特征使得数据由不确定性变成确定性的能力越强。 缺点：信息增益偏向取值较多的特征。



2、输入

序号	条件	要求	说明
1	数据是否需要标准化	否	对特征值大小不敏感
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	尽量多	

3、输出

序号	名称	内容
1	data_out.csv	添加了label、predict_label列的原始数据，label为原始类别，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估。
3	日志	含有模型参数、模型的特征的重要性信息、模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

4、参数

序号	分组	参数	解释
1	字段选择	特征列	进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	特征选择标准	特征选择标准，默认gini，即CART算法。
4	基础参数	特征划分点选择标准	样本的特征划分标准，默认为best
5	基础参数	最大深度	决策树最大深度，数值型
6	基础参数	内部节点最小样本数	内部节点（即判断条件）再划分所需最小样本数，数值型
7	基础参数	叶子节点最小样本数	叶子节点（即分类）最少样本数，数值型
8	基础参数	叶子节点最小的样本权重和	叶子节点（即分类）最小的样本权重和，数值型

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行ID3决策树算法分类。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行分类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行分类前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始进行ID3算法分类。拖入【ID3决策树】算法，将【数据拆分】组件的训练集输出节点和【ID3决策树】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，，特征划分点选择选择best，最大深度填写None，内部节点最小样本数设置为2，叶子节点最小样本数设置为1，叶子节点最小的样本权重和设置为0，右键单击【ID3决策树】算法，选择“运行该节点”。

序号	参数名称	序号	原因
1	特征选择标准	信息熵	可选项有gini、entropy，前者是基尼系数，后者是信息熵。
2	特征划分点选择标准	best	可选项有best、random，前者是在所有特征中寻找最好的切分点，后者是在部分特征中寻找，默认的“best”适合样本量不大的时候，而如果样本数据量非常大，此时决策树构建推荐“random”。
3	最大深度	None	填写int类型数值或None。设置决策树的最大深度，深度越大，越容易过拟合，推荐树的深度为：5-20之间。
4	内部节点最小样本数	2	设置结点的最小样本数量，当样本数量可能小于此值时，结点将不会在划分。
5	叶子节点最小样本数	1	这个值限制了叶子节点最少的样本数，如果某叶子节点数目小于样本数，则会和兄弟节点一起被剪枝。
6	叶子节点最小的样本权重和	0	这个值限制了叶子节点所有样本权重和的最小值，如果小于这个值，则会和兄弟节点一起被剪枝。默认是0，就是不考虑权重问题。

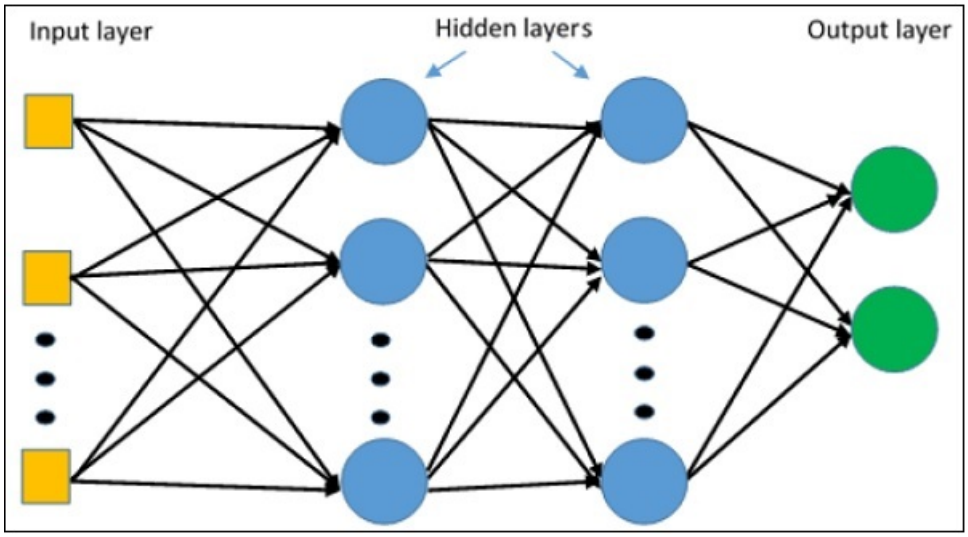
打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【CART分类树】算法右击，点击“查看日志”。

序号	名称	作用
1	模型的特征的重要性信息	输出模型的特征信息
2	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
3	混淆矩阵	可以直观地观测模型每个分类的效果
4	ROC图	计算AUC面积，评价模型训练好坏
5	KS曲线	可根据曲线距离确定最能划分模型的阈值
6	ID3可视化	决策树可视化，展示分类过程

8.4.10 多层感知分类器

（1）作用

多层感知机是有多个全连接层的感知机，而且全连接层之间有激活函数(activation)，可以对训练数据进行非线性处理，非线性模型对复杂数据特征可以更好地拟合其规律。



（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	影响收敛速度
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	尽量多	

（3）输出

序号	名称	内容
1	data_out.csv	添加了label、predict_label列的原始数据，label为原始类别，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估
3	日志	含有模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段选择	特征列	需要进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	基础参数	数据元个数设置	隐藏层神经元个数，(100，100)代表有两层
4	基础参数	迭代次数	迭代的次数，数值型
5	基础参数	权重优化器	lbfgs：quasi-Newton方法的优化器 sgd：随机梯度下降 adam：Kingma、Diederik、Jimmy Ba提出的机遇随机梯度的优化器
6	基础参数	激活函数	激活函数,可选identity、logistic、tanh、relu，logistic也就是sigmod。默认为relu
7	基础参数	正则化参数	L2惩罚（正则化项）参数，默认为0.0001

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接将数据集输入多层感知机分类器进行训练。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行训练的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

训练多层感知分类器前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始训练。拖入【多层感知机】算法，将【数据拆分】组件的训练集输出节点和【多层感知机】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，数据元个数设置为“100,100”，迭代次数设置为200，权重优化器选择adam，激活函数选择relu，正则化参数设置为0.0001，初始学习率设置为0.001，右键单击【多层感知机】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	隐藏层个数	100	神经网络结构，每个隐藏层个数都设置为100
2	迭代次数	200	迭代次数越多，结果越准确。但随着迭代次数不断增大，或是数据量的增加，运行所需要的时间也就越来越久，所以确定一个最佳的迭代次数非常重要。
3	权重优化器	adam	该优化器效果较好
4	学习率	0.001	学习率越大，输出误差对参数的影响就越大，参数更新的就越快，但同时受到异常数据的影响也就越大，很容易发散。初始学习率一般设置在0.01-0.001之间。

打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【神经网络】算法右击，点击“查看日志”。

序号	名称	作用
1	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
2	混淆矩阵	可以直观地观测模型每个分类的效果
3	ROC图	计算AUC面积，评价模型训练好坏
4	K-S曲线	根据曲线距离确定模型的阈值

8.4.11 OVR

（1）解释

OneVsRest：一对一（OvR）的多类/多标签策略，也称为“一对多”，它使用多个两类分类模型进行多类分类。它是针对一些二分类算法（此处我们是基于的二分类算法中的逻辑回归算法）来实现多分类任务的两种最为常用的方式。对于每个分类器，该分类将与所有其他分类进行拟合。这种方法的一个优点是其可解释性。由于每个类别仅由一个和一个分类器表示，因此可以通过检查其对应的分类器来获取有关该类别的知识。

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	影响收敛速度
2	数据是否允许缺失值	否	涉及计算
3	数据是否需要去除重复值	否	
4	载入文件格式	CSV格式	
5	数据量建议	尽量多	

（3）输出

序号	名称	内容
1	data_out.csv	添加了label、predict_label列的原始数据，label为原始类别，predict_label为预测类别。
2	model	训练后的模型，可用作模型预测和模型评估
3	日志	含有模型评价指标、混淆矩阵、ROC图，可以评估模型训练好坏。

（4）参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行分类的列，数值型
2	字段选择	标签	需要分类的标签
3	参数设置	用于计算的数量	进行多分类转换

（5）示例

对于“iris”数据集，它没有缺失值，不需要进行缺失值处理，且因为“iris”数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接将数据集输入OVR进行训练。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

首先将需要进行训练的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行分类前，先将训练数据进行数据拆分为训练集和测试集。拖入【数据拆分】组件，将【输入源】和【数据拆分】连接，在“字段设置”的“特征”中勾选所有字段，参数设置中，测试集占比设置为0.2，右键单击组件，选择“运行该节点”。

开始分类。拖入【OVR】算法，将【数据拆分】组件的训练集输出节点和【OVR】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，“标签”选择“outcome”字段，点击参数设置，填写用于计算的数量（默认为None），右键单击【OVR】算法，选择“运行该节点”。

打开日志，查看结果。在日志中可以查看模型的评价指标、混淆矩阵以及ROC图。对【OVR】算法右击，点击“查看日志”。

序号	名称	作用
1	模型评价指标	输出预测准确率、召回率、F1-score，评价模型效果
2	混淆矩阵	可以直观地观测模型每个分类的效果
3	ROC图	计算AUC面积，评价模型训练好坏
4	K-S曲线	根据曲线距离确定模型的阈值

8.5 回归

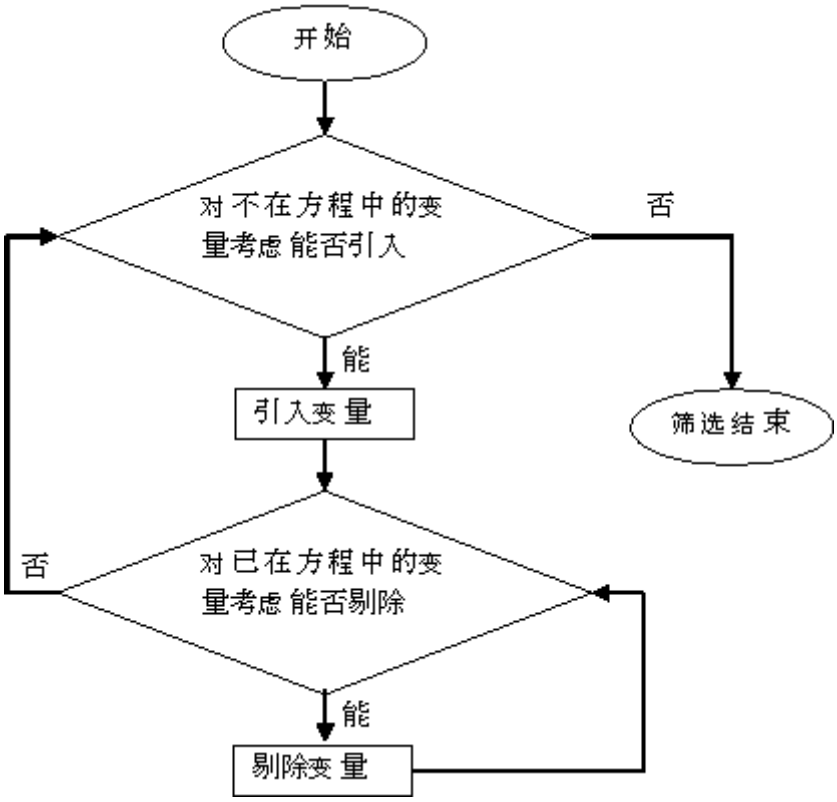
回归的主要应用场景有信用卡申请人风险评估、预测公司业务增长量、预测房价，未来的天气情况等。

8.5.1 线性回归

（1） 作用及原理

线性回归有很多实际用途。分为以下两大类：
一是如果目标是预测或者映射，线性回归可以用来对观测数据集的和X的值拟合出一个预测模型。当完成这样一个模型以后，对于一个新增的X值，在没有给定与它相配对的y的情况下，可以用这个拟合过的模型预测出一个y值。二是给定一个变量y和一些变量X1,...,Xp，这些变量有可能与y相关，线性回归分析可以用来量化y与Xj之间相关性的强度，评估出与y不相关的Xj，并识别出哪些Xj的子集包含了关于y的冗余信息。

回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。 如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。



（2）输入

序列	条件	要求	说明
1	数据是否需要异常值处理	是	涉及平均值和标准差计算
2	数据是否需要标准化	否	不会影响线性回归预测值
3	数据是否允许缺失值	否	涉及平均值和标准差计算
4	载入文件格式	CSV格式	
5	数据量建议	大于30或者大于3（K+1），其中K为解释变量个数	T分布稳定，检验才较为有效

（3）输出

序号	名称	内容
1	data_out.csv	含有预测值的原始数据，其中predict_value为预测值
2	日志	含有模型参数、公式、评价指标以及拟合情况，可以了解模型拟合的好坏以及预测变量

（4）参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行线性回归的列，数值型
2	字段设置	标签列	选择响应变量所在的列，数值型
3	参数设置	归一化	归一化是把所有的数据全部缩放到0-1之间。布尔型，默认为True
4	参数设置	拟合截距	若参数值为True时，代表训练模型需要加一个截距项； 若参数为False时，代表模型无需加截距项。布尔型，默认为True

（5）示例

数据集“iris”中没有缺失值和异常值，因此不用缺失值处理和异常值处理。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

先将需要进行线性回归的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行线性回归，建立模型。拖入【线性回归】算法，将【输入源】算法和【线性回归】算法相连接，在“字段设置”的“特征”中勾选除了“outcome”和“petal width”字段的其余字段，在“标签”中选择“petal width”。点击“参数设置”，拟合截距设置为True，归一化设置为True，右键单击【线性回归】算法，选择“运行该节点”。

序号	参数名称	值	原因
1	拟合截距	True	保留默认值True，增加一个截距项
2	归一化	True	使数据都落在0-1之间

（3）打开日志，查看结果。在日志中可以了解模型拟合的好坏。对【线性回归】算法右击，点击“查看日志”。

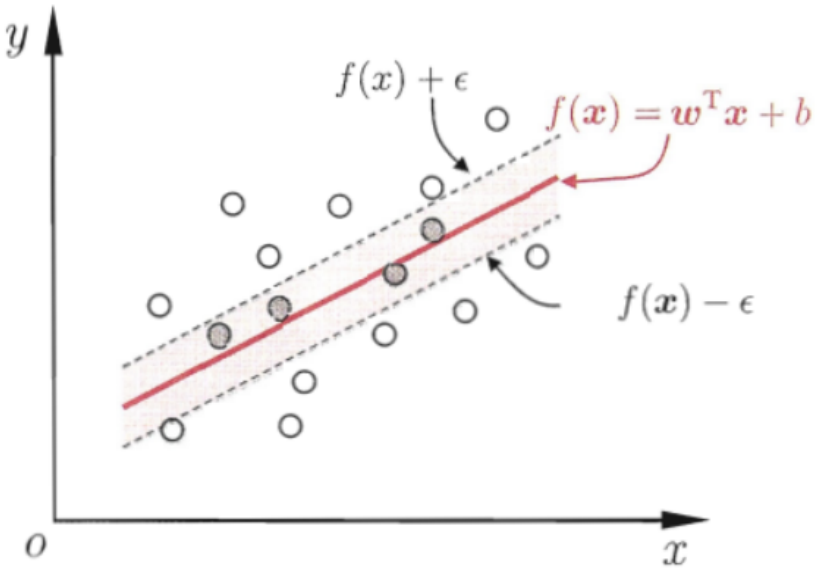
序号	名称	作用
1	模型参数	可查看参数设置
2	模型公式	可用来预测变量
3	模型评价指标	可查看模型拟合的好坏，例如，R-Squared为0.94表示模型拟合效果较好
4	模型拟合情况	对比预测值与真实值，了解模型拟合情况

8.5.2 支持向量回归

（1）作用及原理

支持向量回归可利用少量的预测数值来去直接拟合线。同时可以修改内置的默认配置来获得更好的预测性能。

对于一般的回归问题，给定训练样本D={(x1,y1),(x2,y2),...,(xn,yn)},yi€R,我们希望学习到一个f(x)使得其与y尽可能的接近，w，b是待确定的参数。在这个模型中，只有当f(x)与y完全相同时，损失才为零，而支持向量回归假设我们能容忍的f(x)与y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2ε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。（间隔带两侧的松弛程度可有所不同）



支持向量回归示意图. 红色显示出 ϵ-间隔带, 落入其中的样本不计算损失.

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及距离计算
2	数据是否允许缺失值	否	对缺失数据敏感
3	数据是否需要去除重复值	是	影响回归准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过	2000	

（3）输出

序号	名称	内容
1	data_out.csv	含有预测值的原始数据，其中predict_label为预测值
2	日志	含有模型参数、评价指标以及拟合情况，可以了解模型拟合的好坏和预测变量

（4）参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行 支持向量回归的列，数值型
2	字段设置	标签列	选择响应变量所在的列，数值型
3	参数设置	核函数	核函数有四种内置选择，如果选择了这些核函数， 对应的核函数参数在后面有单独的参数需要调。 默认是高斯核函数。
4	参数设置	poly函数维度	如果我们在核函数参数使用了多项式核函数 ‘poly’，那么我们就需要对这个参数进行调参，默认是3。数值型
5	参数设置	核函数参数	如果我们在核函数参数使用了多项式核函数 ‘poly’，高斯核函数‘rbf’, 或者sigmoid核函数，那么我们就需要对这个参数进行调参。默认为‘auto’。
6	参数设置	惩罚参数	惩罚系数，即对误差的宽容度。惩罚系数越高，说明越不能容忍出现误差，容易过拟合。惩罚系数越小，容易欠拟合。惩罚系数过大或过小，泛化能力变差。默认为1，数值型

（5）示例

数据集“iris”中没有缺失值和重复值，因此不用缺失值处理和重复值处理且数据集无明显的量纲差异，所以不需要进行数据标椎化。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

先将需要进行支持向量回归算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行支持向量回归算法，建立模型。拖入【支持向量回归】，将【输入源】算法和【支持向量回归】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”字段，在“标签”中选择“petal_width”字段。保留默认参数。右键单击【支持向量回归】，选择“运行该节点”。

序号	名称	值	原因
1	核函数	RBF 函数	能够实现非线性映射；参数的数量影响模型的复杂程度，多项式核函数参数较多；RBF 核的数值难度较小。
2	poly函数维度	3	如果核函数参数使用了多项式核函数 'poly'，那么就需要对这个参数进行调参，选择其他核函数时会被忽略。
3	核函数参数	auto	代表其值为样本特征数的倒数。
4	惩罚参数	1	减小惩罚参数的话，容许训练样本中有一些误分类错误样本，泛化能力强。

打开日志，查看结果。在日志中可以了解模型拟合的好坏。对【支持向量回归】算法右击，点击"查看日志"。

序号	名称	作用
1	模型参数	可查看参数设置
2	模型评价指标	可查看模型拟合效果指标，例如R-Squared为0.95，模型拟合好
4	模型拟合情况	真实值与拟合值折线图趋势对比

8.5.3 岭回归

（1）作用及原理

岭回归也是一种用于回归的线性模型，因此它的预测公式与普通最小二乘法相同。但在岭回归中，对系数（w）的选择不仅要在训练数据上得到好的预测结果，而且还要拟合附加约束。我们还希望系数尽量小。换句话说，w 的所有元素都应接近于0。直观上来看，这意味着每个特征对输出的影响应尽可能小（即斜率很小），同时仍给出很好的预测结果。 这种约束是所谓正则化（regularization）的一个例子。正则化是指对模型做显式约束，以避免过拟合。岭回归用到的这种被称为 L2 正则化。

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及欧式距离
2	数据是否允许缺失值	否	涉及距离计算
3	数据是否需要去除重复值	是	影响模型准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过	500000	

（3）输出

序号	名称	内容
1	data_out.csv	含有预测值的原始数据，其中add_col为预测值
2	日志	含有模型参数、评价指标以及拟合情况，可以了解模型拟合的好坏

（4）参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行K最近邻回归的列，数值型
2	字段设置	标签列	选择响应变量所在的列，数值型
3	参数设置	L2项系数	正则化力度，正浮点数，默认值为1； 增大L2项系数会使得系数更加趋向于 0，从而降低训练集性能， 但可能会提高泛化性能；反之则可以让系数受到的限制更小
4	参数设置	最大迭代次数	共轭梯度求解器的最大迭代次数
5	参数设置	是否归一化	如果为True，则在回归前，回归变量X将会进行归一化，减去均值，然后除以L2范数

（5）示例

数据集“data”中没有缺失值和重复值，因此不用缺失值处理和重复值处理且数据集无明显的量纲差异，所以不需要进行数据标椎化。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

先将需要进行 K最近邻回归算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“data”，勾选文件“data.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行K最近邻回归算法，建立模型。拖入【K最近邻回归】，将【输入源】算法和【K最近邻回归】算法相连接，在“字段设置”的“特征”中勾选“sepeal length”,"sepal width","petal length"字段，在“标签”中选择"petal width"。保持默认参数。右键单击【K最近邻回归】，选择“运行该节点”。

打开日志，查看结果。在日志中可以了解模型拟合的好坏。对【K最近邻回归】算法右击，点击“查看日志”。

序号	名称	作用
1	模型参数	可查看参数设置
2	模型公式	预测值与回归变量的线性关系式
2	模型评价指标	可查看模型拟合效果指标，其中R-Squared为0.7，模型拟合的较好
3	模型拟合情况	可对比真实值与预测值

8.5.4 CART回归树

（1）作用及原理

CART分类回归树是一种典型的二叉决策树，可以做分类或者回归。如果待预测结果是离散型数据，则CART生成分类决策树；如果待预测结果是连续型数据，则CART生成回归决策树。数据对象的属性特征为离散型或连续型，并不是区别分类树与回归树的标准。

决策树的生成就是递归地构建二叉决策树的过程，对回归树用平方误差最小化准则，对分类树用基尼指数最小化准则，进行特征选择，生成二叉树。

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	否	不要求对特征标准化和统一量化
2	数据是否允许缺失值	否	影响模型准确性
3	数据是否去除重复值	是	影响模型准确性
4	载入文件格式	CSV格式	
5	数据量建议不少于	100	

（3）输出

序号	名称	内容
1	data_out.csv	含有预测值的原始数据，其中add_col为预测值
2	日志	含有模型参数、属性、评价指标以及拟合情况，可以了解模型拟合的好坏

（4）参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行CART回归树的列，数值型
2	参数设置	树的最大深度	设置决策树的最大深度。默认值为None。数值型
3	参数设置	子树划分所需最小样本数	内部节点包含的最少样本数，可输入整数或者小数； 即如果内部节点包含样本数低于这个值，则不再分裂，直接作为叶子节点。数值型
4	参数设置	叶子节点最少样本数	叶子节点包含的最少样本数。 如果一个节点分裂后，左右子节点包含样本数低于叶子节点最少样本数，则不可以分裂。数值型
5	参数设置	最大特征数	节点分裂时考虑的最大特征数量。
6	参数设置	随机种子	随机种子/随机生成器。
7	参数设置	最大叶子节点数	叶子节点最大个数。数值型
8	参数设置	切分评价准则	衡量生成树的纯度，可选择基尼系数 ‘gini’，或者信息熵 ‘entropy’。
9	参数设置	切分原则	分裂点选择，有两种方式可选择，'best' 与 'random'。
10	参数设置	叶子节点最小的样本权重	叶子节点包含样本的最小权重值。数值型

序号	分组	参数	解释
11	参数设置	预排序	是否对样本进行预分类。 即节点寻找最优分裂点之前，对样本进行预排序，加快找到最优分裂点； 由于增加了一步操作，数据集小的时候可以加快速度； 但数据集大的时候反而会减慢速度。
12	参数设置	节点划分最小减少不纯度	不纯度最小减少值，即分裂节点时必须满足减少的不纯度大于这个值。数值型
13	字段设置	标签列	选择响应变量所在的列，数值型

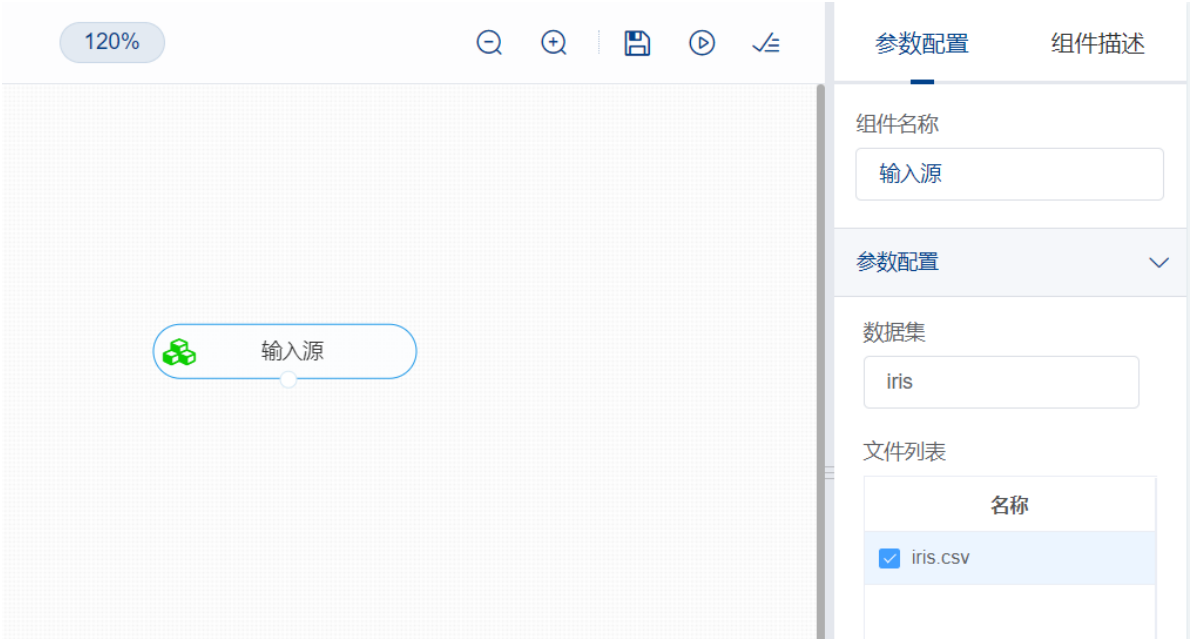
（5）示例

数据集“iris”中没有缺失值和异常值，因此不用缺失值处理和异常值处理。

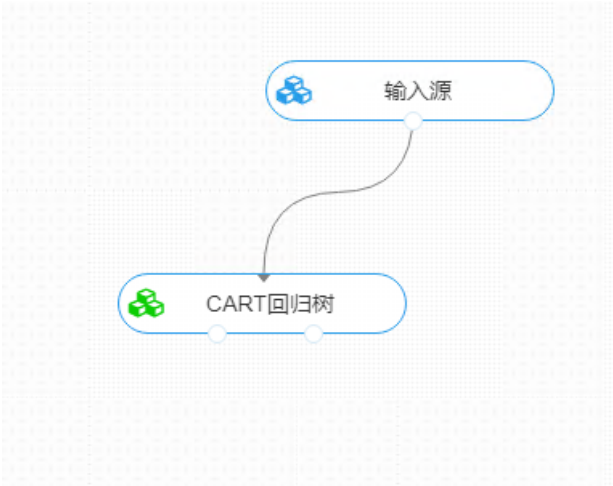
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

先将需要进行CART回归树算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。



开始进行CART回归树算法，建立模型。拖入【CART回归树】，将【输入源】算法和【CART回归树】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，在“标签”中选择“outcome”。保持默认参数。右键单击【CART回归树】，选择“运行该节点”。



打开日志，查看结果。在日志中可以了解模型拟合的好坏。对【CART回归树】算法右击，点击“查看日志”。

序号	名称	作用
1	模型参数	可查看参数设置
2	模型属性	可查看输入输出特征数
3	模型评价指标	可查看模型拟合效果
4	模型拟合情况	可对比预测值和真实值
5	CART可视化图	回归树展示

8.5.5 XGBoost算法

1 作用

XGBoost（extreme Gradient Boosting）全名叫极端梯度提升，XGBoost是集成学习方法的王牌。XGBoost在绝大多数的回归和分类问题上表现的十分顶尖。Boosting法是结合多个弱学习器给出最终的学习结果，不管任务是分类或回归，我们都用回归任务的思想来构建最优Boosting模型 。

其思想是把每个弱学习器的输出结果当成连续值，这样做的目的是可以对每个弱学习器的结果进行累加处理，且能更好的利用损失函数来优化模型。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	标准化不会改变分裂点的位置
2	数据是否允许缺失值	否	补全缺失值可以增加特征的预测能力
3	数据是否需要异常值处理	是	可以增加特征的预测能力
4	载入文件格式	CSV格式	
5	数据量建议不超过	10000	

3 输出

序号	名称	内容
1	data_out.csv	含有预测值的原始数据，其中predict_label为预测值
2	日志	含有模型参数、评价指标以及拟合情况，可以了解模型拟合的好坏和预测变量

4 参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行XGBoost的列，数值型
2	参数设置	最大迭代次数	总共迭代的次数，即决策树的个数。最大迭代次数越大，模型的学习能力就会越强，模型也越容易过拟合。数值型
3	参数设置	树深度	树的最大深度。 这个值也是用来避免过拟合的。树深度越大，模型会学到更具体更局部的样本。数值型
4	参数设置	学习率	通过减少每一步的权重，可以提高模型的鲁棒性。数值型
5	参数设置	惩罚项系数	惩罚项系数，指定节点分裂所需的最小损失函数下降值。 这个参数的值越大，算法越保守。数值型
6	参数设置	训练集比例	控制对于每棵树，随机采样的比例。 减小这个参数的值，算法会更加保守，避免过拟合。但是，如果这个值设置得过小，它可能会导致欠拟合。数值型
7	参数设置	测试集数据比例	用来控制每棵随机采样的列数的占比(每一列是一个特征)。数值型
8	字段设置	标签列	选择响应变量所在的列，数值型

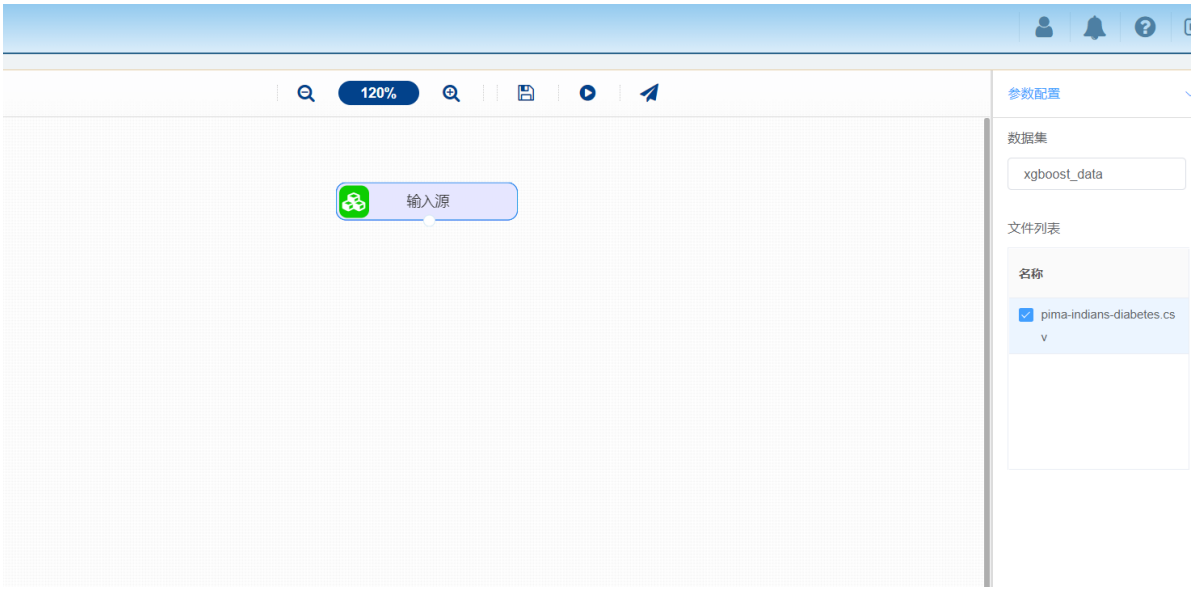
5 示例

数据集“xgboost_data”中没有缺失值和异常值，因此不用缺失值处理和异常值处理。

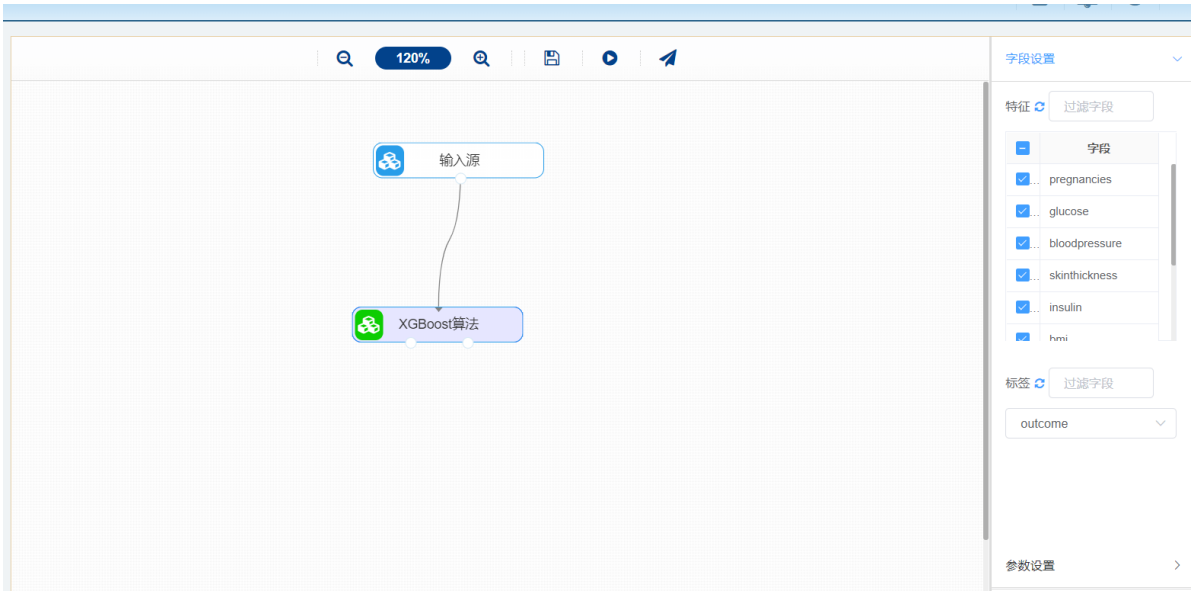
	A	B	C	D	E	F	G	H	I
1	pregnancie	glucose	bloodpress	skinthickr	insulin	bmi	diabetespe	age	outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1
25	9	119	80	35	0	29	0.263	29	1
26	11	143	94	33	146	36.6	0.254	51	1
27	10	125	70	26	115	31.1	0.205	41	1
28	7	147	76	0	0	39.4	0.257	43	1
29	1	97	66	15	140	23.2	0.487	22	0
30	13	145	82	19	110	22.2	0.245	57	0

pima-indians-diabetes

（1）先将需要进行XGBoost算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“xgboost_data”，勾选文件“pima-indians-diabetes.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2）开始进行XGBoost算法，建立模型。拖入【XGBoost算法】，将【输入源】算法和【XGBoost算法】算法相连接，在“字段设置”的“特征”中勾选除了“outcome”字段的其余字段，在“标签”中选择“outcome”。保持默认参数。右键单击【XGBoost算法】，选择“运行该节点”。



序号	名称	值	原因
1	最大迭代次数	150	一般都不会建议一个太大的数目，300以下为佳。
2	树深度	3	值越大，越容易过拟合；值越小，越容易欠拟合。典型值3-10。
3	学习率	0.1	值越小，训练越慢。典型值为0.01-0.2。
4	惩罚项系数	0	值越大，算法越保守。
5	训练集比例	0.8	典型值为0.5-1，值越大，越易过拟合，需避免过拟合。
6	测试集数据比例	0.8	典型值为0.5-1，需避免过拟合。

（3） 打开日志，查看结果。在日志中可以了解模型拟合的好坏。对【XGBoost算法】算法右击，点击“查看日志”。

序列	名称	作用
1	模型参数	可查看参数设置
2	模型评价指标	可查看模型拟合效果指标，例如R-Squared为0.7，模型拟合较好
3	模型拟合情况	可对比真实值与预测值

8.5.6 Lasso回归

1 作用及原理

LASSO是由1996年Robert Tibshirani首次提出，全称Least absolute shrinkage and selection operator。该方法是一种压缩估计。它通过构造一个惩罚函数得到一个较为精炼的模型，使得它压缩一些回归系数，即强制系数绝对值之和小于某个固定值；同时设定一些回归系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。它通过构造一个惩罚函数，可以将变量的系数进行压缩并使某些回归系数变为0，进而达到变量选择的目的。

一般来说，对于高维的特征数据，尤其线性关系是稀疏的，我们会采用Lasso回归。或者是要在一堆特征里面找出主要的特征，那么Lasso回归更是首选了。但是Lasso类需要自己对α调优，所以不是Lasso回归的首选。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	是	涉及正则化计算
2	数据是否允许缺失值	否	导致不可靠的输出
3	数据是否需要去除重复值	是	导致不可靠的输出
4	载入文件格式	CSV格式	
5	数据量建议不少于	100	

3 输出

序号	名称	内容
1	data_out.csv	含有预测值的原始数据，其中add_col为预测值
2	日志	含有模型参数、评价指标以及拟合情况，可以了解模型拟合的好坏

4 参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行Lasso回归的列，数值型
2	参数设置	拟合截距	表示是否计算截距，推荐设置为True。布尔类型，默认为True
3	参数设置	归一化	表示是否对各个特征进行归一化，推荐设置为True。布尔类型，默认为True
4	参数设置	L1项系数	正则项系数，数值越大，则对复杂模型的惩罚力度越大。数值型
5	参数设置	最大迭代次数	部分求解器需要通过迭代实现，这个参数指定了模型优化的最大迭代次数。数值型
6	字段设置	标签列	选择响应变量所在的列，数值型

5 示例

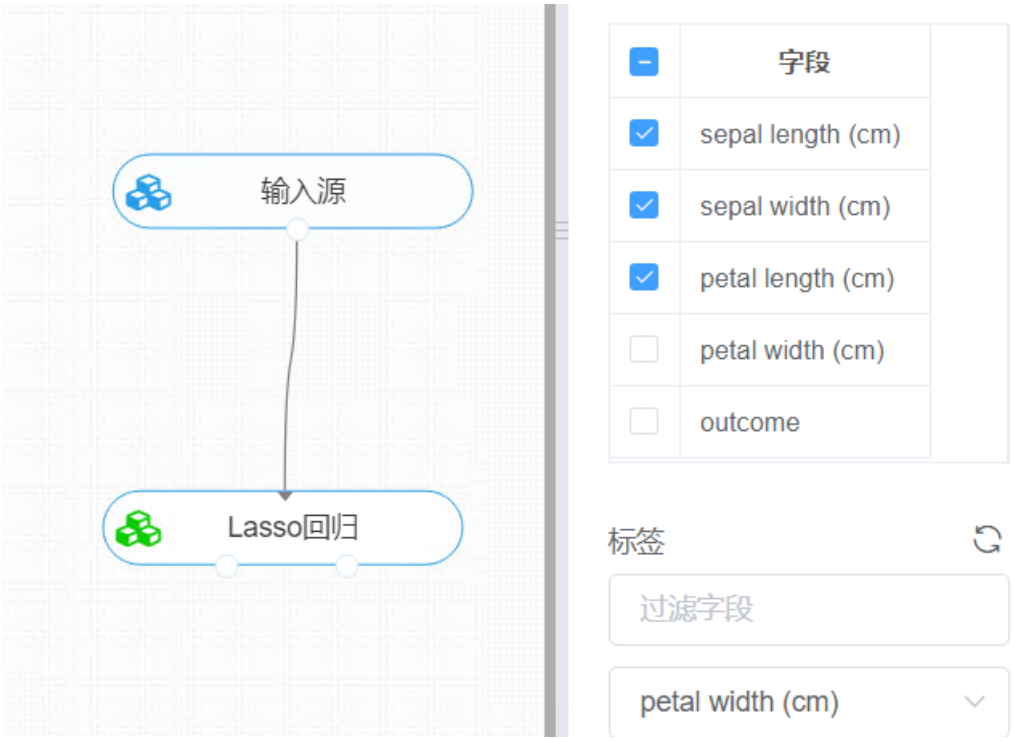
数据集“iris”中没有缺失值和重复值，因此不用缺失值处理和重复值处理且数据集无明显的量纲差异，所以不需要进行数据标椎化。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

（1）先将需要进行 Lasso回归算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

（2）开始进行Lasso回归算法，建立模型。拖入【Lasso回归】，将【输入源】算法和【Lasso回归】算法相连接，在“字段设置”的“特征”中勾选除了“petal_width”与“outcome”字段的其余字段，在“标签”中选择“petal_width”。保持默认参数。右键单击【Lasso回归】，选择“运行该节点”。



序号	参数名称	值	原因
1	拟合截距	True	计算截距
2	归一化	True	原因是正则化是有偏估计，会对权重进行惩罚。在量纲不同的情况，正则化会带来更大的偏差。
3	L1项系数	0.001	模型拟合效果较好
4	最大迭代次数	1000	迭代次数越多，结果越准确。

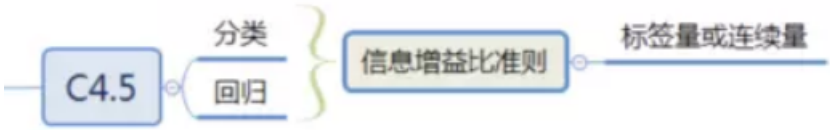
（3） 打开日志，查看结果。在日志中可以了解模型拟合的好坏。对【Lasso回归】算法右击，点击“查看日志”。

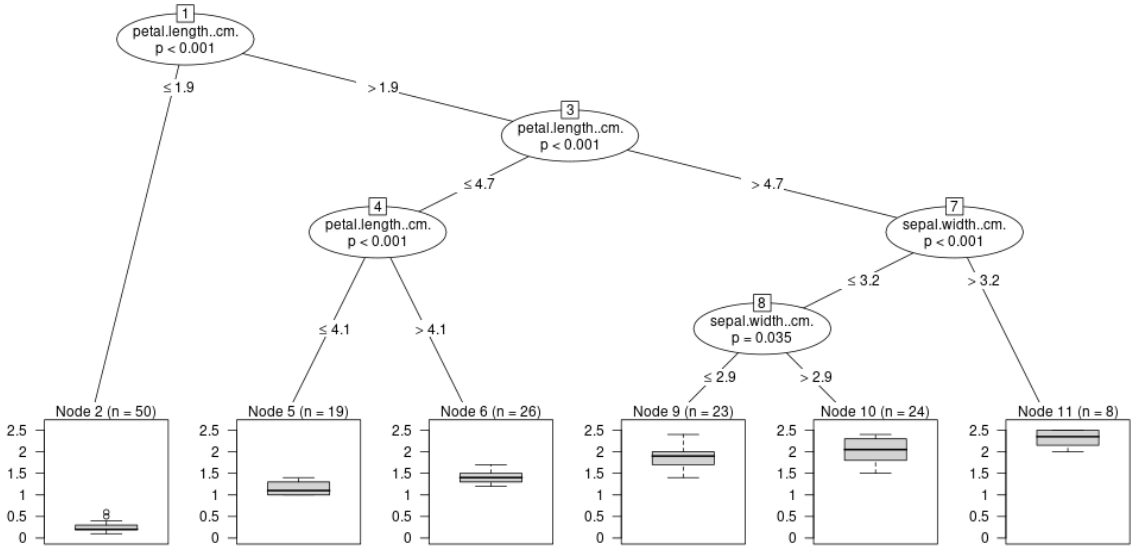
序号	名称	作用
1	模型参数	可查看参数设置
2	模型评价指标	可查看模型拟合效果指标，例如R-Squared为0.93，模型拟合好
3	模型拟合情况	可对比真实值与预测值

8.5.7 C4.5 回归树

（1）作用与原理

为了解决ID3算法的弊端，进而产生了C4.5算法。C4.5算法与ID3算法相似，不同之处仅在于C4.5算法在选择特征的时候采用了信息增益比作为标准，即选择信息增益比最大的特征作为当前样本集合的根节点。





（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	否	标准化不会改变分裂点的位置
2	数据是否允许缺失值	否	补全缺失值可以增加特征的预测能力
3	数据是否需要异常值处理	是	可以增加特征的预测能力
4	载入文件格式	CSV格式	
5	数据量建议不超过	10000	

（3）输出

序号	名称	内容
1	data_out.csv	含有预测值的原始数据，其中predict_label为预测值
2	日志	含有模型信息、回归树、回归结果评价、真实值与预测值散点图分布，可以了解模型的预测情况

（4）参数

序号	分组	参数	解释
1	字段设置	特征列	需要进行C4.5回归树的列，数值型
2	字段设置	标签列	选择响应变量所在的列，数值型

（5）示例

数据集“iris”中没有缺失值和异常值，因此不用缺失值处理和异常值处理。

	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

先将需要进行C4.5回归树算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

开始进行C4.5回归树算法，建立模型。拖入【C4.5回归树】，将【输入源】算法和【C4.5回归树】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，在“标签”中选择“outcome”。保持默认参数。右键单击【C4.5回归树】，选择“运行该节点”。

打开日志，查看结果。在日志中可以了解模型拟合的好坏。对【C4.5回归】算法右击，点击“查看日志”。

序号	名称	作用
1	模型信息	可查看回归树类型
2	回归树	可视化回归过程中的分类情况
2	模型评价指标	可查看模型拟合效果指标（MES,RMES,MAE,MAPE,R-Squared)
3	散点图	可对比真实值与预测值的分布情况

8.6 时间序列

8.6.1 模型残差检验

应用场景有节日客流量预测、季节销量预测、实时股价预测等等。

1 作用及原理

一个模型是否显著有效主要看它提取的信息是否充分。一个好的拟合模型应该能够提取观察值序列中几乎所有的样本相关信息，换言之，拟合残差项中将不再蕴含任何相关信息，即残差序列应该为白噪声序列。这样的模型称为显著有效的模型。

反之，如果残差序列为非白噪声序列，那就意味着残差序列中还残留着相关信息未被提取，这就说明拟合模型不够有效，通常需要选择其他模型，重新拟合。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	是	消除明显的量纲关系
2	数据是否允许缺失值	否	影响结果的准确性
3	数据是否需要去除重复值	是	影响结果的准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过	50000	

3 输出

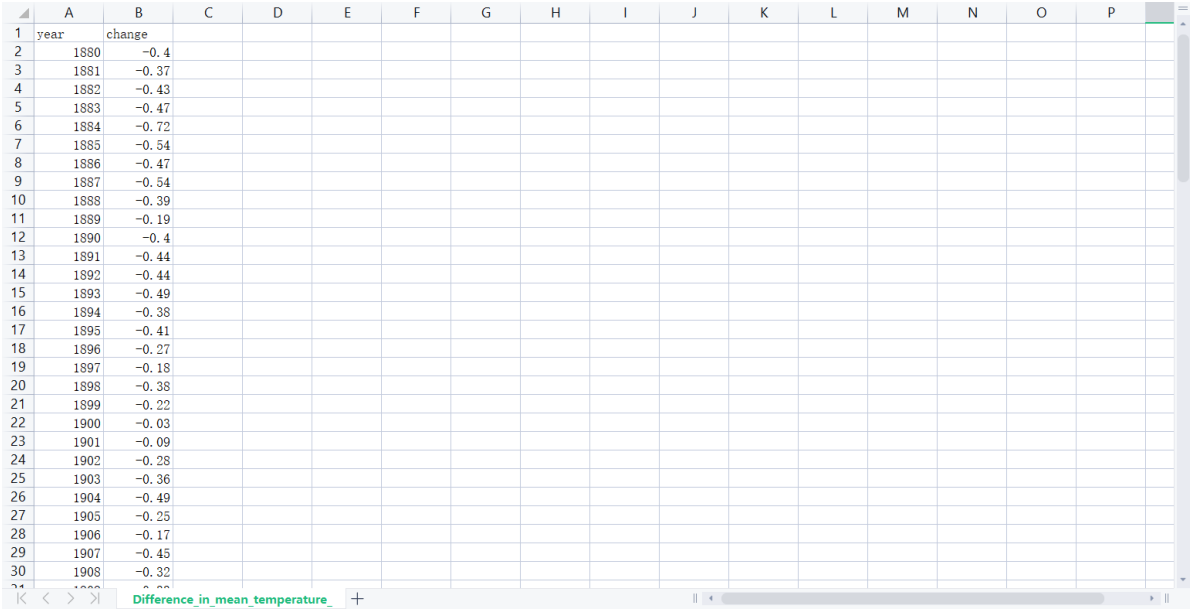
序号	名称	内容
1	日志	含有残差序列自相关图、残差序列平稳性检验结果，可知残差序列是否为平稳序列，从而得出模型是否显著有效。

4 参数

序号	分组	参数	解释
1	字段设置	模型数据列	进行检验的数据列，数值型
2	参数设置	ar阶数	代表预测模型中采用的时序数据本身的滞后数(lags)，数值型
3	参数设置	差分阶数	代表时序数据需要进行几阶差分化，才是稳定的，数值型
4	参数设置	ma阶数	代表预测模型中采用的预测误差的滞后数(lags)，数值性

5 示例

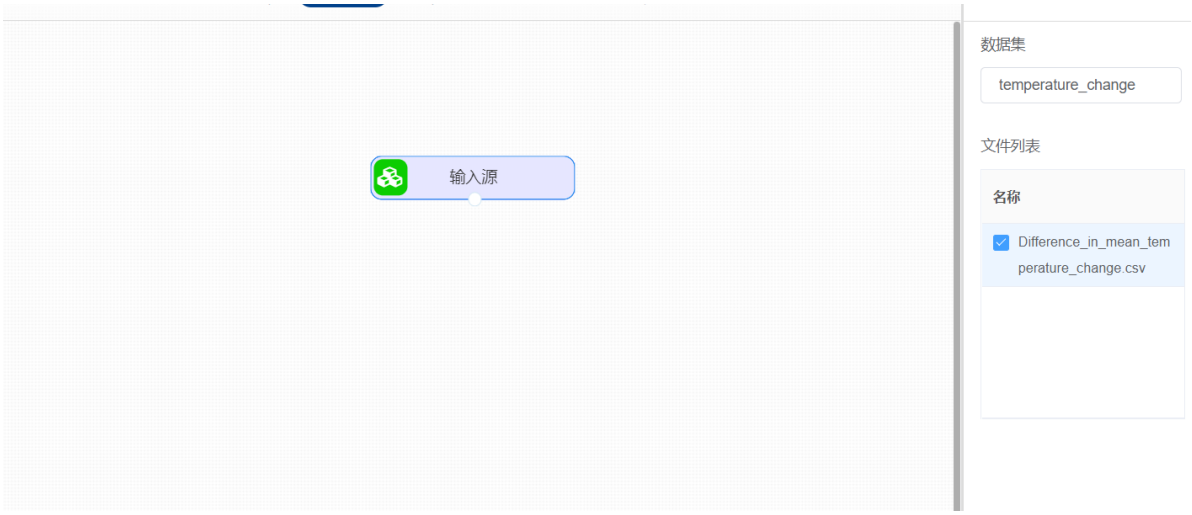
对于“temperature_change”数据集，它没有缺失值和重复值，不需要进行缺失值处理和重复值处理，且因为“temperature_change”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接对数据集进行模型残差检验算法。



	A	B
1	year	change
2	1880	-0.4
3	1881	-0.37
4	1882	-0.43
5	1883	-0.47
6	1884	-0.72
7	1885	-0.54
8	1886	-0.47
9	1887	-0.54
10	1888	-0.39
11	1889	-0.19
12	1890	-0.4
13	1891	-0.44
14	1892	-0.44
15	1893	-0.49
16	1894	-0.38
17	1895	-0.41
18	1896	-0.27
19	1897	-0.18
20	1898	-0.38
21	1899	-0.22
22	1900	-0.03
23	1901	-0.09
24	1902	-0.28
25	1903	-0.36
26	1904	-0.49
27	1905	-0.25
28	1906	-0.17
29	1907	-0.45
30	1908	-0.32

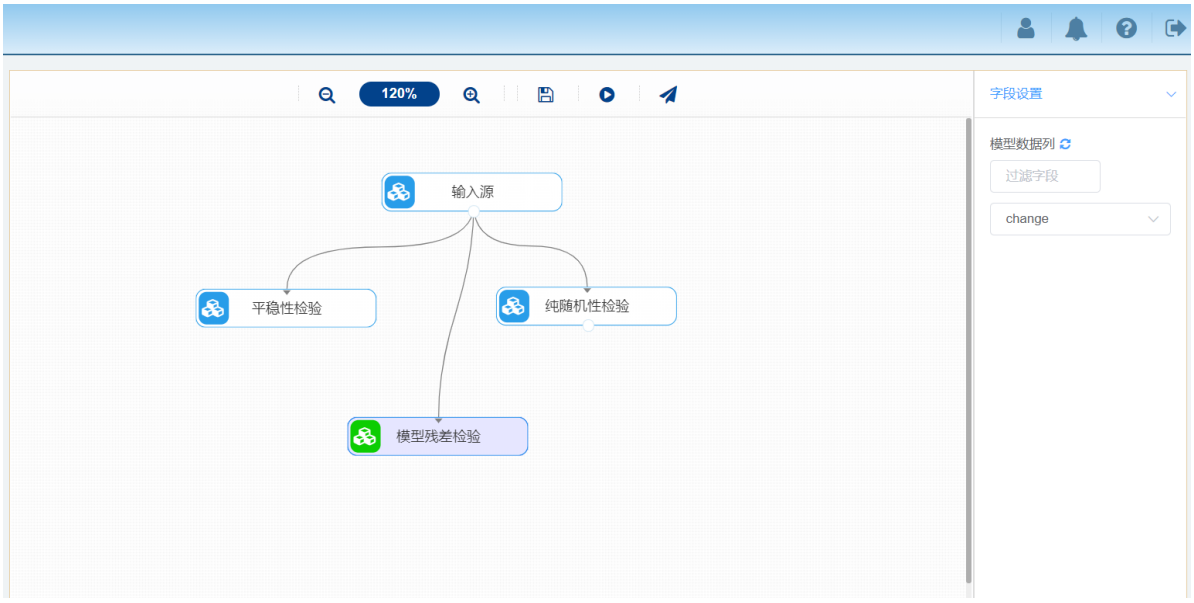
Difference_in_mean_temperature_

（1） 首先将需要进行模型残差检验的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“temperature_change”，勾选文件“Difference_in_mean_temperature_change.csv”，右键单击【输入源】算法，选择“运行该节点”。



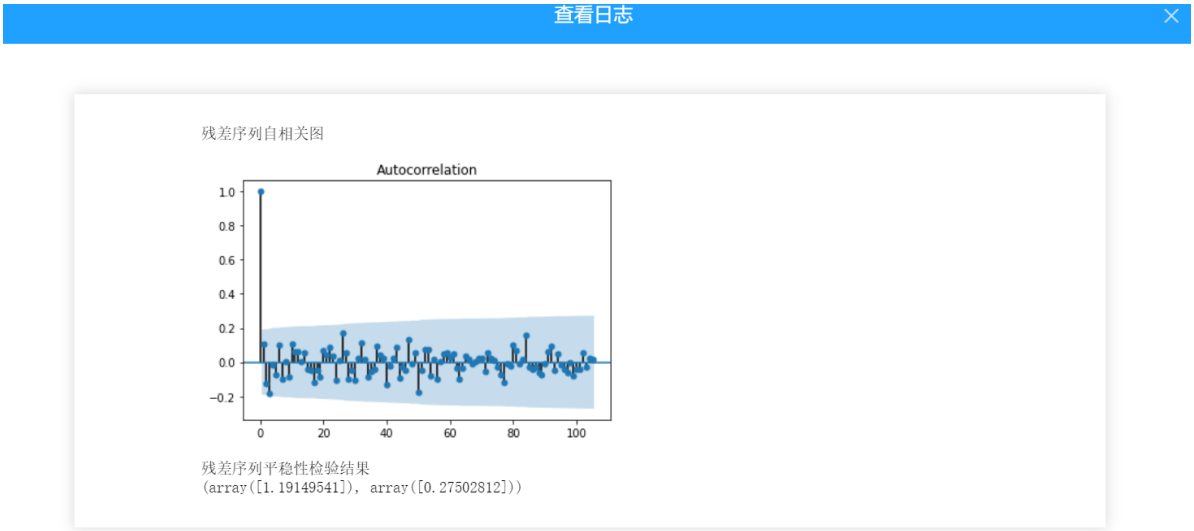
（2）进行平稳性检验和纯随机性检验，检验该序列是否是平稳非白噪声序列。拖入【纯随机检验】算法和【平稳性检验】算法，分别与【输入源】算法连接。右键单击算法，选择“运行该节点”。

（2） 开始进行模型残差检验，检验残差是否存在自相关及是否平稳。拖入【模型残差检验】算法，将【输入源】算法和【模型残差检验】算法相连接，在“字段设置”的“模型数据列”中选择“change”字段，点击“参数设置”，将“差分阶数”设置为0，右键单击【模型残差检验】算法，选择“运行该节点”。



序号	参数名称	数值	原因
1	ar阶数	1	PACF1阶后衰减趋于零
2	差分阶数	0	序列平稳，不需要进行差分运算
3	ma阶数	1	ACF1阶后衰减趋于零

（3） 打开日志，查看结果。在日志中可以查看结果得出残差是否存在自相关。对【模型残差检验】算法右击，点击“查看日志”。



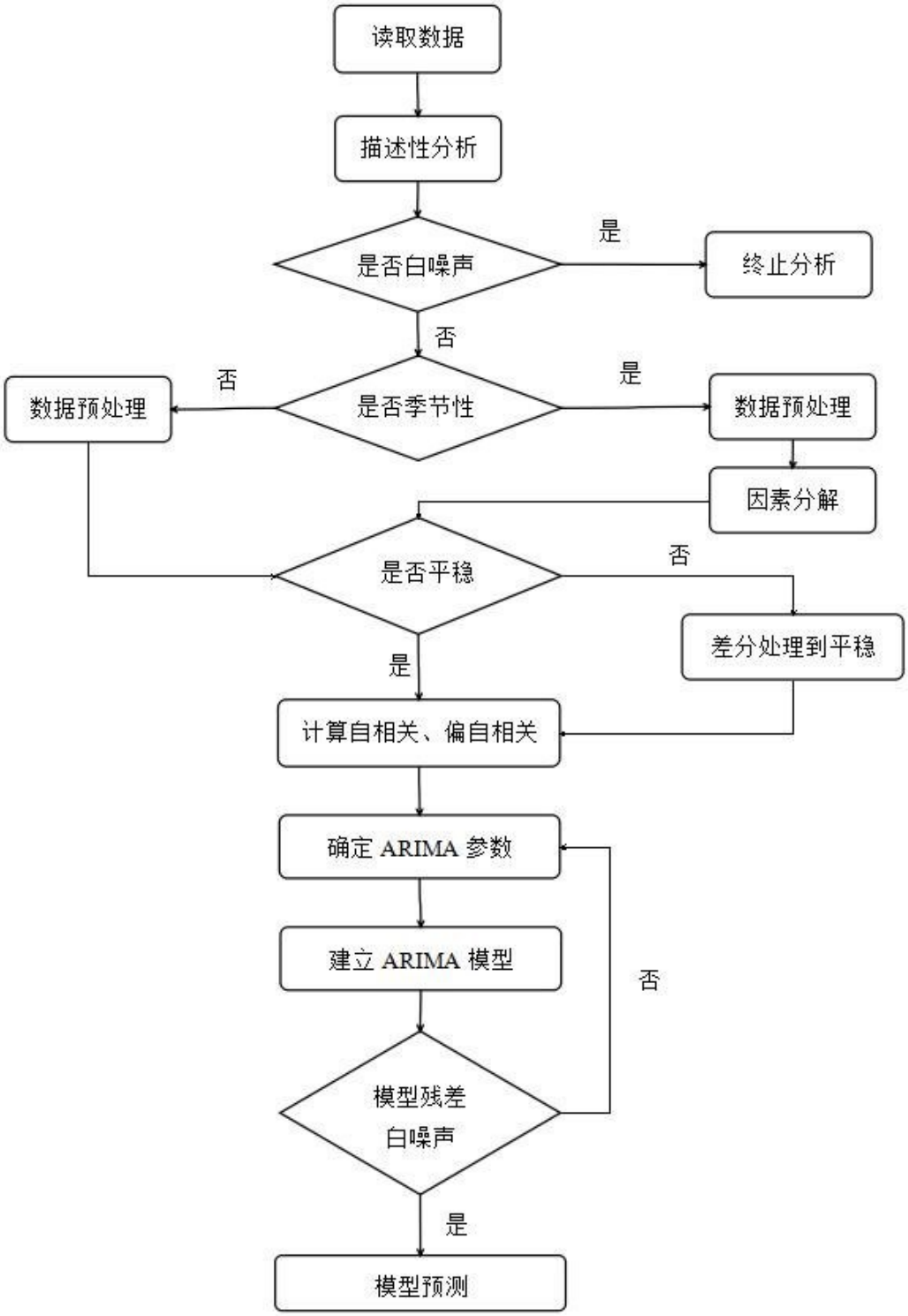
序号	名称	作用
1	残差序列自相关图	可判断残差序列的平稳性。图中，自相关系数始终控制在2倍的标准差范围内，这是平稳序列通常具有的自相关图特征。
2	残差序列平稳性检验结果	可判断残差序列的平稳性。图中，p值为0.28>0.05，因此可以认为不能拒绝原假设，认为该序列是白噪声序列

8.6.2 ARIMA

（1）作用及原理

ARIMA模型是将预测对象随时间推移而形成的数据序列视为一个随机序列，用一定的数学模型来近似描述这个序列。这个模型一旦被识别后就可以从时间序列的过去值及现在值来预测未来值。现代统计方法、计量经济模型在某种程度上已经能够帮助企业对未来进行预测。

ARIMA模型全称为自回归移动平均模型(Autoregressive Integrated Moving Average Model，简记ARIMA)。是由博克思(Box)fFfl詹金斯(Jenkins)于70年代初提出的一著名时间序列预测方法，所以又称为box—enkins模型、博克思一詹金斯法。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。ARIMA模型可分为3种：(1)自回归模型(简称[AR模型)；(2)滑动平均模型(简称MA模型)；(3)自回归滑动平均混合模型(简称ARIMA模型)。以时间序列的自相关分析为基础。ARIMA模型在经济预测过程中既考虑了经济现象在时间序列上的依存性，又考虑了随机波动的干扰性，对于经济运行短期趋势的预测准确率较高，是应用比较广泛的方法之一。


读取数据
描述性分析
是否白噪声
是
终止分析
否
是否季节性
否
数据预处理
是
数据预处理
因素分解
是否平稳
否
差分处理到平稳
是
计算自相关、偏自相关
确定 ARIMA 参数
建立 ARIMA 模型
模型残差
白噪声
否
是
模型预测

(2) 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	涉及乘法运算时，标准化后易出现数据为零
2	数据是否允许缺失值	否	表现出的不确定性更加显著，蕴涵的确定性成分更难把握
3	数据是否需要去除重复值	是	影响模型准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过	4000	

（3） 输出

序号	名称	内容
1	data_out.csv	10期的预测数据。其中y_preValue为预测数据
2	日志	可查看模型的拟合情况，例如R-Squared为0.95，初步认为模型拟合较好

（4）参数

序号	分组	参数	解释
1	参数配置	时序列	随着时间变化的数据。数值型
2	参数配置	时间列	时间数据。数值型或字符型
3	参数设置	预测周期数	预测的周期数。数值型
4	参数设置	自回归项数p	代表预测模型中采用的时序数据本身的滞后数(lags)。数值型
5	参数设置	差分次数d	代表时序数据需要进行几阶差分化，才是稳定的。数值型
6	参数设置	移动平均项数q	代表预测模型中采用的预测误差的滞后数(lags)。数值型
7	参数设置	季节性自回归阶数P	AR模型中季节分量的阶数，等于PACF图中显著滞后的数量。数值型
8	参数设置	季节性差分阶数D	季节性整合阶数。数值型
9	参数设置	季节性移动平均阶数Q	MA模型中季节分量的阶数，等于ACF图中显著滞后的数量。数值型
10	参数设置	单个季节期间的时间步数S	季节周期的长度值。数值型

（5）示例

对于“BOGAMBNS”数据集，它没有缺失值和重复值，不需要进行缺失值处理和重复值处理，且因为“BOGAMBNS”数据集无明显的量纲差异，所以不需要进行数据标准化。因此可直接对数据集进行ARIMA算法。

	A	B	C	D
1	DATE	BOGAMBNS		
2	1959/1/1	40.656		
3	1959/2/1	40.224		
4	1959/3/1	40.205		
5	1959/4/1	40.409		
6	1959/5/1	40.485		
7	1959/6/1	40.649		
8	1959/7/1	41.062		
9	1959/8/1	40.962		
10	1959/9/1	40.967		
11	1959/10/1	40.945		
12	1959/11/1	41.134		
13	1959/12/1	41.766		
14	1960/1/1	41.015		
15	1960/2/1	40.271		
16	1960/3/1	40.212		
17	1960/4/1	40.329		
18	1960/5/1	40.374		
19	1960/6/1	40.597		
20	1960/7/1	40.944		
21	1960/8/1	40.841		
22	1960/9/1	40.991		
23	1960/10/1	41.114		
24	1960/11/1	41.43		
25	1960/12/1	41.864		
26	1961/1/1	41.158		
27	1961/2/1	40.542		
28	1961/3/1	40.449		

BOGAMBNS

首先将需要进行ARIMA算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“BOGAMBNS”，勾选文件“BOGAMBNS.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行平稳性检验和纯随机性检验，检验该序列是否是平稳非白噪声序列。拖入【纯随机检验】算法和【平稳性检验】算法，分别与【输入源】算法连接。右键单击算法，选择“运行该节点”。

开始进行ARIMA算法，建立模型。拖入【ARIMA】算法，将【输入源】算法和【ARIMA】算法相连接，在“参数配置”的“时序列”中勾选“BOGAMBNS”字段，“时间列”中勾选“DATE”字段。点击“参数设置”，“自回归项数p”设置为4，"差分次数d"设置为1，"季节性差分阶数D"设置为1，“单个季节期间的时间步数S”设置为4，其余参数保持默认值。右键单击【ARIMA】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	预测周期数	10	预测后面10期数据
2	自回归项数p	4	偏自相关系数4阶显著非零，4阶之后截尾
3	差分次数d	1	进行1阶差分提取趋势信息，将序列转化成平稳序列
4	移动平均项数q	0	自相关图拖尾
5	季节性自回归阶数P	0	建立加法模型，P=0
6	季节性差分阶数D	1	进行周期为4的一次差分，将序列转化成平稳序列
7	季节性移动平均阶数Q	0	建立加法模型，Q=0
8	单个季节期间的时间步数S	4	一般情况下季节性周期数据为4，月度周期数据为12。

打开日志，查看结果。在日志中可以查看模型拟合效果。对【ARIMA】算法右击，点击“查看日志”。

序号	名称	作用
1	模型参数	可查看参数设置
2	模型具体信息	可查看数据量为120，模型为SARIMAX(4, 1, 0)x(0, 1, 0, 4)等
3	模型评价指标	可得MSE 为0.53，R-Squared为0.95等，MSE越小越好，R-Squared越大越好，初步得出模型拟合效果好
4	模型拟合情况	得到的预测值与实际值对比图，大概了解模型拟合情况
5	模型预测情况	预测后面10期的数据图

8.6.3 模型定阶

1 作用及原理

对于一个时间序列，若已知其服从的模型，关键是定阶。实际上在建模过程中，首先要解决定阶问题，才能进行参数估计。虽然所建模型的阶数越高，则其越能够准确反应时序的特性，但当模型阶数过高时，则要求估计的参数越多，误差也随之增加，则必定会损伤模型函数。因而确定合适的模型阶数至关重要。

在时序建模过程中，考虑到模型中所含待定参数的个数多少直接影响所拟合模型对数据的逼近程度。为了很好的解决了上述问题，日本统计学家 Akaike 提出了 AIC、BIC 准则（最小信息准则）。其定义为：AIC 最小信息准则及 BIC 最小信息准则的计算值达到最小时则认为此时可达到模型的理想阶数。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	是	消除明显的量纲关系
2	数据是否允许缺失值	否	影响结果的准确性
3	数据是否需要去除重复值	是	影响结果的准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过	50000	

3 输出

序号	名称	内容
1	日志	可得到BIC达到最小的阶数

4 参数

序号	分组	参数	解释
1	字段选择	模型数据列	需要进行模型定位的列，数值型
2	参数设置	ar阶数	代表预测模型中采用的时序数据本身的滞后数(lags)，数值型
3	参数设置	ma阶数	代表预测模型中采用的预测误差的滞后数(lags)，数值性

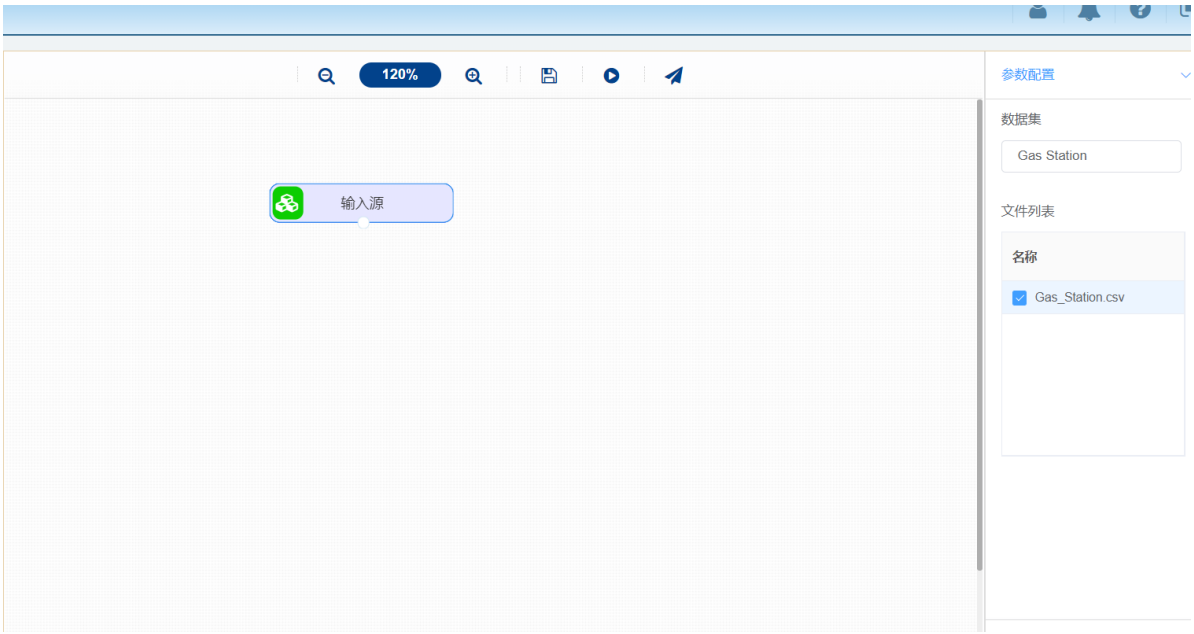
5 示例

对于“Gas Station”数据集，它没有缺失值和重复值，不需要进行缺失值处理和重复值处理，且因为数据集无明显的量纲差异，所以不需要进行数据标椎化。因此可直接对数据集进行模型定阶算法。

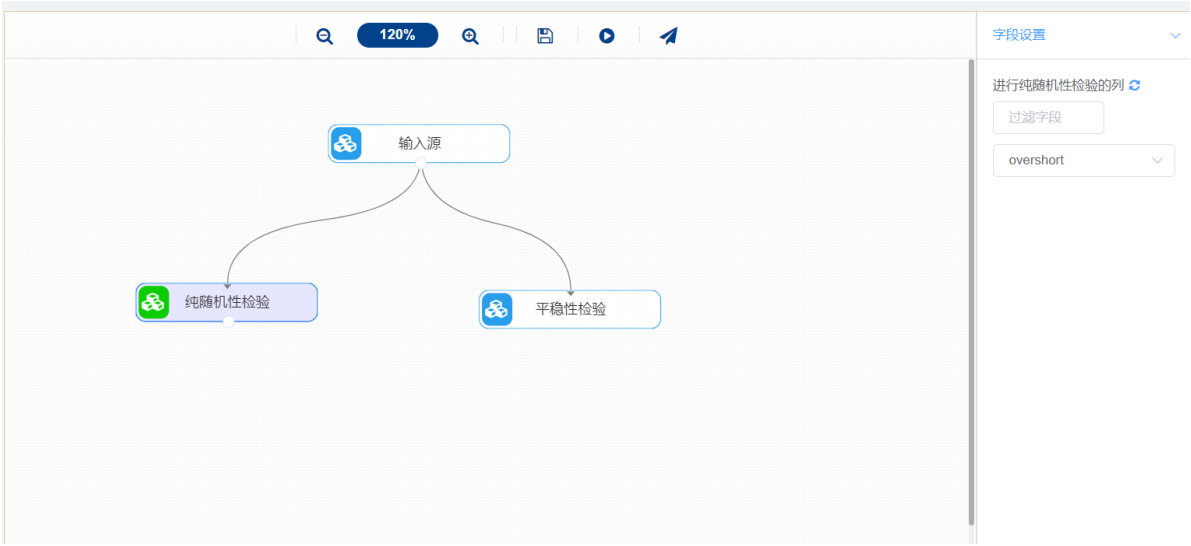
	A	B	C	D	E	F	G	H	I
1	day	overshort							
2	1	78							
3	2	-58							
4	3	53							
5	4	-63							
6	5	13							
7	6	-6							
8	7	-16							
9	8	-14							
10	9	3							
11	10	-74							
12	11	89							
13	12	-48							
14	13	-14							
15	14	32							
16	15	56							
17	16	-86							
18	17	-66							
19	18	50							
20	19	26							
21	20	59							
22	21	-47							
23	22	-83							
24	23	2							
25	24	-1							
26	25	124							
27	26	-106							
28	27	113							
29	28	-76							
30	29	-47							

Gas_Station

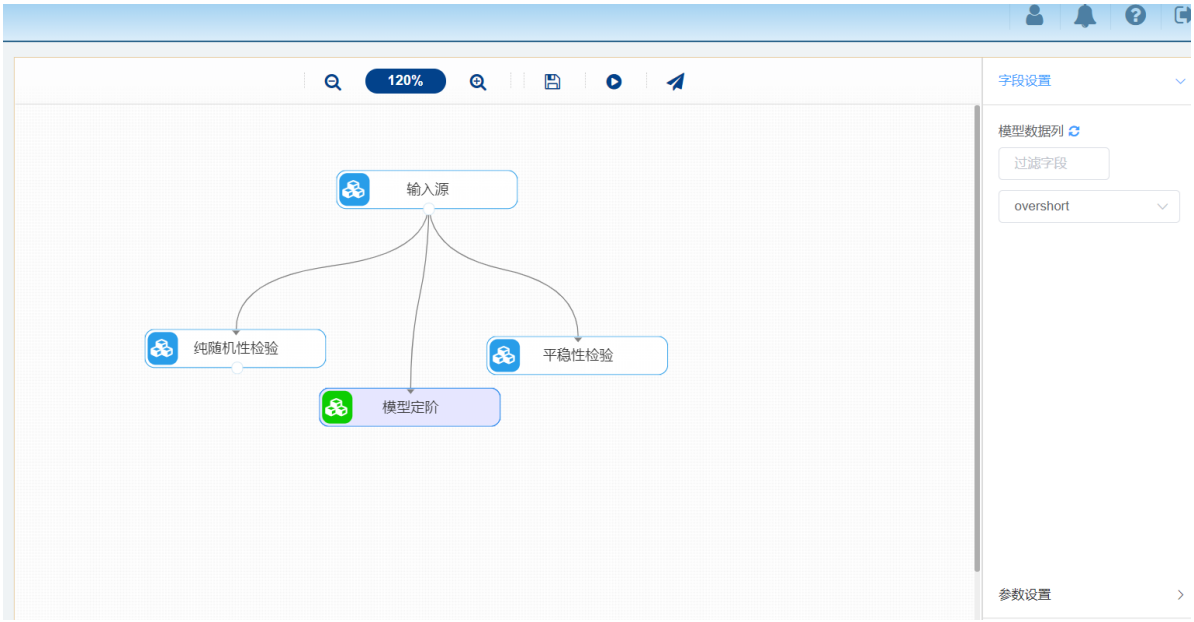
（1） 首先将需要进行模型定阶的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“Gas_Station”，勾选文件“Gas_Station.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2）进行平稳性检验和纯随机性检验，检验该序列是否是平稳非白噪声序列。拖入【纯随机检验】算法和【平稳性检验】算法，分别与【输入源】算法连接。右键单击算法，选择“运行该节点”。



（3） 开始进行模型定位，将该序列自动定阶阶数。拖入【模型定位】算法，将【输入源】算法和【模型定阶】算法相连接，在“字段设置”的“模型数据列”中选择“change”字段，点击“参数设置”，"ar阶数"设置为4，“ma阶数”设置为4，右键单击【模型定阶】算法，选择“运行该节点”。



序号	参数名称	数值	原因
1	ar阶数	4	如果阶数选得太低，会导致得不到理想阶数
2	ma阶数	4	如果阶数选得太低，会导致得不到理想阶数

（3） 打开日志，查看结果。在日志中可以得到BIC达到最小的阶数。对【模型定阶】算法右击，点击“查看日志”。

序号	名称	作用
1	BIC	BIC 值达到最小的可得到理想阶数

8.6.4 指数平滑法

1 作用及原理

指数平滑法是生产预测中常用的一种方法。也用于中短期经济发展趋势预测，所有预测方法中，指数平滑是用得最多的一种。简单的全期平均法是对时间数列的过去数据一个不漏地全部加以同等利用;移动平均法则不考虑较远期的数据，并在加权移动平均法中给予近期资料更大的权重;而指数平滑法则兼容了全期平均和移动平均所长，不舍弃过去的数据，但是仅给予逐渐减弱的影响程度，即随着数据的远离，赋予逐渐收敛为零的权数。

指数平滑法是在移动平均法基础上发展起来的一种时间序列分析预测法，它是通过计算指数平滑值，配合一定的时间序列预测模型对现象的未来进行预测。其原理是任一期的指数平滑值都是本期实际观察值与前一期指数平滑值的加权平均。

2 输入

序号	条件	要求	说明
1	数据是否需要标准化	否	涉及乘法运算时，标准化后易出现数据为零
2	数据是否允许缺失值	否	涉及加权平均数计算
3	数据是否需要去除重复值	是	涉及加权平均数计算
4	载入文件格式	CSV格式	
5	数据量建议不超过	20000	

3 输出

序号	名称	内容
1	data_out.csv	10期的预测数据。其中milk_preValue为预测数据
2	日志	可查看模型的拟合情况，例如R-Squared为0.99，初步认为模型拟合较好

4 参数

序号	分组	参数	解释
1	参数配置	时序列	随着时间变化的数据。数值型
2	参数配置	时间列	时间数据。日期型数据
3	参数设置	预测周期数	预测的周期数。数值型
4	参数设置	指数平滑模型	指数平滑模型是最简单和最常用的时间序列预测模型。有三种常用分类：单指数模型，双指数模型和三指数模型。
5	参数设置	趋势模型	有三种可选项，就是加法趋势和乘法趋势还有None。
6	参数设置	季节模型	有三种可选项，加法、乘法还有None。
7	参数设置	季节性周期	季节性的周期长期。数值型
8	参数设置	阻尼	阻尼系数越小，近期实际值对预测结果的影响越大；反之，阻尼系数越大，近期实际值对预测结果的影响就越小。布尔值

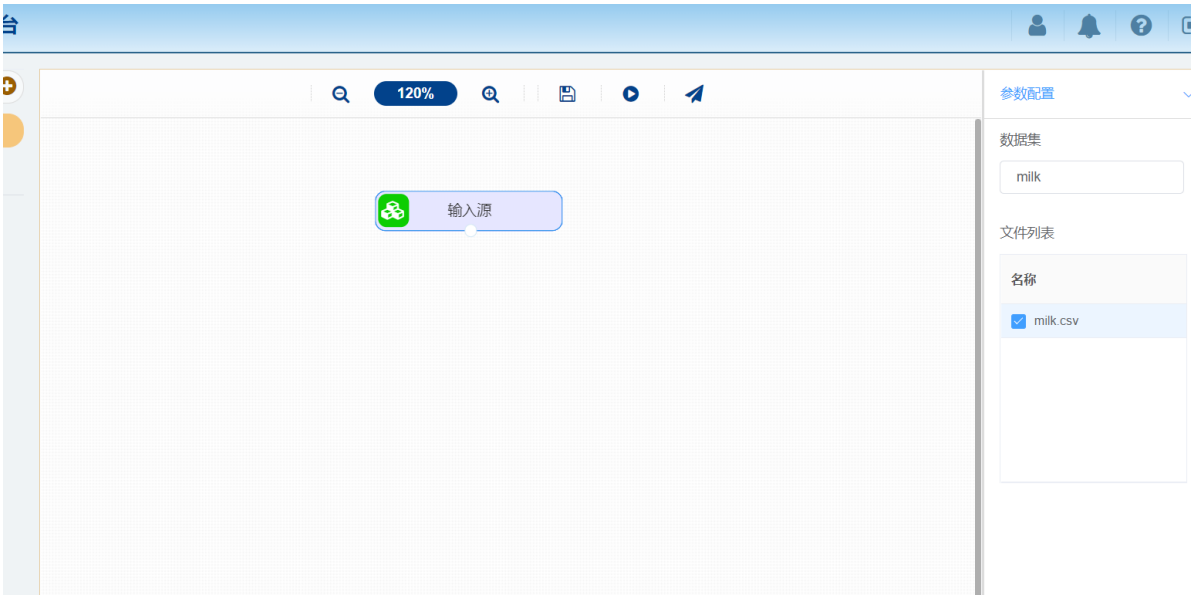
5 示例

数据集“milk”中没有缺失值和重复值，因此不用缺失值处理和重复值处理且数据集无明显的量纲差异，所以不需要进行数据标准化。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	time	milk															
2	Jan-62	589															
3	Feb-62	561															
4	Mar-62	640															
5	Apr-62	656															
6	May-62	727															
7	Jun-62	697															
8	Jul-62	640															
9	Aug-62	599															
10	Sep-62	568															
11	Oct-62	577															
12	Nov-62	553															
13	Dec-62	582															
14	Jan-63	600															
15	Feb-63	566															
16	Mar-63	653															
17	Apr-63	673															
18	May-63	742															
19	Jun-63	716															
20	Jul-63	660															
21	Aug-63	617															
22	Sep-63	583															
23	Oct-63	587															
24	Nov-63	565															
25	Dec-63	598															
26	Jan-64	628															
27	Feb-64	618															
28	Mar-64	688															

Sheet2 Sheet3 +

（1） 首先将需要进行指数平滑法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“milk”，勾选文件“milk.csv”，右键单击【输入源】算法，选择“运行该节点”。



（2） 开始进行指数平滑法，建立模型。拖入【指数平滑法】算法，将【输入源】算法和【指数平滑法】算法相连接，在“参数配置”的“时序列”中勾选“milk”字段，“时间列”中勾选“time”字段，在“参数设置“中将“指数平滑模型”设置为three，”季节性周期“设置为12，"阻尼"设置为True。其他参数保持不变。右键单击【指数平滑法】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	预测周期数	10	预测后面10期数据
2	指数平滑模型	three	三次指数平滑算法可以对同时含有趋势和季节性的时间序列进行预测
3	趋势模型	加法	AIC和BIC是用来比较模型的好坏的，而且是越小越好。选择加法时，AIC和BIC最小
4	季节模型	加法	AIC和BIC是用来比较模型的好坏的，而且是越小越好。选择加法时，AIC和BIC最小
5	季节性周期	12	月度周期数据为12
6	阻尼	True	对于长期预测，使用Holt方法的预测在未来会无限期地增加或减少。 在这种情况下，我们使用具有阻尼参数0<φ<1的阻尼趋势方法来防止预测“失控”

（3） 打开日志，查看结果。在日志中可以查看各个聚类分群的个数与占比、每个分群中各个特征的优势与劣势以及聚类中心。对【指数平滑法】算法右击，点击“查看日志”。

序号	名称	作用
1	模型具体信息	可查看模型的参数设置情况
2	模型评价指标	可查看模型拟合效果指标，例如R-Squared为0.99，初步认为模型拟合较好
3	模型拟合情况	得到的预测值与实际值对比图，大概了解模型拟合情况
4	模型预测情况	预测周期数为10的预测图

8.6.5 ARCH

（1）作用

ARCH模型的英文直译是：自回归条件异方差模型。ARCH是一种用来处理时间序列的模型。在股票中，ARCH可以用来预测股票的波动率，从而控制风险。（在金融领域，波动率与风险直接挂钩，一个资产波动越大，风险越大，而获得更高收益的可能也更大）ARCH模型广泛应用于波动性有关广泛研究领域。包括政策研究、理论命题检验、季节性分析等方面。

ARCH模型的基本思想是指在以前信息集下，某一时刻一个扰动项的发生是服从正态分布。该正态分布的均值为零，方差是一个随时间变化的量（即为条件异方差）。并且这个随时间变化的方差是过去有限项噪声值平方的线性组合（即为自回归）。这样就构成了自回归条件异方差。

（2）输入

序号	条件	要求	说明
1	数据是否需要标准化	否	涉及乘法运算时，标准化后易出现数据为零
2	数据是否允许缺失值	否	表现出的不确定性更加显著，蕴涵的确定性成分更难把握
3	数据是否需要去除重复值	是	影响模型准确性
4	载入文件格式	CSV格式	
5	数据量建议不超过	4000	

（3）输出

序号	名称	内容
1	日志	可查看模型信息、模型的拟合情况

（4）参数

序号	分组	参数	解释
1	参数配置	时序列	随时间变化的数据

（5） 示例

数据集“total_purchase_amt”中没有缺失值和异常值，因此不用缺失值处理和异常值处理。

	A	B	C	D
1	names	report_date	total_purchase_amt	
2	1	2013/7/1 0:00	32488348	
3	2	2013/7/2 0:00	29037390	
4	3	2013/7/3 0:00	27270770	
5	4	2013/7/4 0:00	18321185	
6	5	2013/7/5 0:00	11648749	
7	6	2013/7/6 0:00	36751272	
8	7	2013/7/7 0:00	8962232	
9	8	2013/7/8 0:00	57258266	
10	9	2013/7/9 0:00	26798941	
11	10	2013/7/10 0:00	30696506	
12	11	2013/7/11 0:00	44075197	
13	12	2013/7/12 0:00	34183904	
14	13	2013/7/13 0:00	15164717	
15	14	2013/7/14 0:00	22615303	
16	15	2013/7/15 0:00	48128555	
17	16	2013/7/16 0:00	50622847	
18	17	2013/7/17 0:00	29015682	
19	18	2013/7/18 0:00	24234505	
20	19	2013/7/19 0:00	33680124	
21	20	2013/7/20 0:00	20439079	
22	21	2013/7/21 0:00	21142394	
23	22	2013/7/22 0:00	40448896	
24	23	2013/7/23 0:00	58136147	
25	24	2013/7/24 0:00	48422518	
26	25	2013/7/25 0:00	57433418	
27	26	2013/7/26 0:00	44721817	
28	27	2013/7/27 0:00	17194451	

total_purchase_amt

（1） 首先将需要进行ARCH算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“purchase”，勾选文件“total_purchase_amt.csv”，右键单击【输入源】算法，选择“运行该节点”。

（2） 开始进行ARCH算法，建立模型。拖入【ARCH】算法，将【输入源】算法和【ARCH】算法相连接，在“参数配置”的“时序列”中勾选“total_purchase_amt”字段。右键单击【ARCH】算法，选择“运行该节点”。

（3） 打开日志，查看结果。在日志中可以查看模型拟合效果。对【ARCH】算法右击，点击“查看日志”。

序号	名称	作用
1	模型迭代	可查看模型迭代信息
2	模型具体信息	AECH模型的拟合结果
4	模型拟合情况	得到的预测值与实际值对比图，大概了解模型拟合情况

8.7 文本挖掘

8.7.1文本过滤

（1）作用

此处的文本过滤是指过滤掉数据的某一特征列中不满足用户的部分数据，有效的文本过滤可以克服数据中隐藏的缺陷。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out.csv	文本过滤后的数据表

（4）参数

序号	分组	参数	说明
1	参数配置	是否生成新列	过滤内容作为新列出现
2	参数配置	文本列	选择过滤的文本列
3	参数配置	过滤内容	填写需要被过滤掉的内容，支持正则表达式

（5）示例

对titanic数据集进行数据文本过滤示例。

	A	B	C	D	E	F
1	Survived	Passenger	Pclass	Sex	Age	
2	0	1	3	male	22	
3	1	2	1	female	38	
4	1	3	3	female	26	
5	1	4	1	female	35	
6	0	5	3	male	35	
7	0	6	3	male		
8	0	7	1	male	54	
9	0	8	3	male	2	
10	1	9	3	female	27	
11	1	10	2	female	14	
12	1	11	3	female	4	
13	1	12	1	female	58	
14	0	13	3	male	20	
15	0	14	3	male	39	
16	0	15	3	female	14	
17	1	16	2	female	55	
18	0	17	3	male	2	
19	1	18	2	male		
20	0	19	3	female	31	
21	1	20	3	female		
22	0	21	2	male	35	
23	1	22	2	male	34	
24	1	23	3	female	15	
25	1	24	1	male	28	
26	0	25	3	female	8	
27	1	26	3	female	38	
28	0	27	3	male		

titanic_data

首先将titanic数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“titanic”，勾选文件“titanic.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行文本过滤，将【文本过滤】组件与输入源连接，在参数配置中选择“是否生成新列”与“文本列”，并填写需要过滤的目标内容，右键单击【文本过滤】组件，选择“运行该节点”。

打开数据，查看结果。对【文本过滤】组件右击，点击“查看数据”，即可查看过滤后的数据结果。

序号	名称	作用
1	数据	文本过滤后的数据表

8.7.2 情感分析

（1）作用

情感分析（Sentiment Analysis）是一种常见的自然语言处理（NLP）方法的应用，它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理，利用情感得分来量化定性数据，当情感得分越接近1，说明该文本属于积极性范畴，反之则属于消极性范畴。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out.csv	带有情感分析得分的数据表

（4）参数

序号	分组	参数	说明
1	参数配置	文本列	选择需要进行情感分析的特征列

（5）示例

对movie数据集进行情感分析示例。

	A	B	C	D	E	F
1	movieRow	movieId	title			
2	0	1	Toy Story (1995)			
3	1	2	Jumanji (1995)			
4	2	3	Grumpier Old Men (1995)			
5	3	4	Waiting to Exhale (1995)			
6	4	5	Father of the Bride Part II (1995)			
7	5	6	Heat (1995)			
8	6	7	Sabrina (1995)			
9	7	8	Tom and Huck (1995)			
10	8	9	Sudden Death (1995)			
11	9	10	GoldenEye (1995)			
12	10	11	American President, The (1995)			
13	11	12	Dracula: Dead and Loving It (1995)			
14	12	13	Balto (1995)			
15	13	14	Nixon (1995)			
16	14	15	Cutthroat Island (1995)			
17	15	16	Casino (1995)			
18	16	17	Sense and Sensibility (1995)			
19	17	18	Four Rooms (1995)			
20	18	19	Ace Ventura: When Nature Calls (1995)			
21	19	20	Money Train (1995)			
22	20	21	Get Shorty (1995)			
23	21	22	Copycat (1995)			
24	22	23	Assassins (1995)			
25	23	24	Powder (1995)			
26	24	25	Leaving Las Vegas (1995)			
27	25	26	Othello (1995)			
28	26	27	Now and Then (1995)			

movies

首先将movie数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“movie”，勾选文件“movie.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行情感分析，将【情感分析】组件与输入源连接，在参数配置中选择“文本列”，右键单击【情感分析】组件，选择“运行该节点”。

打开数据，查看结果。对【情感分析】组件右击，点击“查看数据”，即可查看情感分析后的得分结果数据表。

8.7.3 词袋模型

（1）作用

将数据文本向量化并作为词袋模型输出，方便后续对文本的分词预处理操作。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	数据特征列是否需要是列表	列表、字符形式的列表“[]”	

（3）输出

序号	名称	内容
1	data_out	词袋模型

（4）参数

序号	分组	参数	说明
1	参数配置	词汇量大小	分词处理的最大单词数量
2	参数配置	文本列	

（5）示例

对position数据集进行词袋模型示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行词袋模型，将【词袋模型】组件与输入源连接，在参数配置中选择需要进行分词的特征列，以及词汇量数量，右键单击【词袋模型】组件，选择“运行该节点”。

导出模型。对【词袋模型】右击选择导出数据，即可导出组件得出的结果词袋模型。

8.7.4 Jieba 分词

（1）作用

Jieba分词在自然语言处理中发挥着重要的作用。其主要功能是中文分词，组件可对文本数据进行分词处理，同时也可通过添加自定义字典进行分词处理。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	输入字符串序列，如：”数据挖掘工程师“
2	自定义分词词典	有需要时可添加	jieba有内置的分词词典，大多情况我们采用默认的分词词典

（3）输出

序号	名称	内容
1	data_out	分词结果

（4）参数

序号	分组	参数	说明
1	字段配置	特征列	选择需要进行分词的特征列

（5）示例

用position数据集进行Jieba中文分词示例。

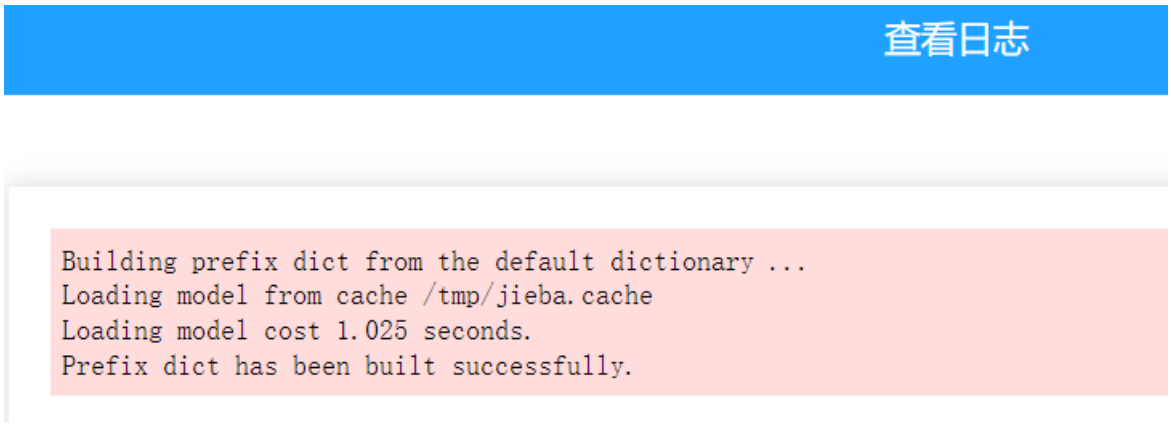
	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行jieba分词，将【jieba分词】组件与输入源连接，在参数配置中选择需要进行分词的特征列，右键单击【jieba分词】组件，选择“运行该节点”。

查看结果。对【jieba分词】右键单击选择查看数据即可预览文本数据的分词结果；右键选择查看日志可以查看组件是否成功对数据进行分词。



8.7.5 Hanlp分词

（1）作用

Hanlp是一个文本工具包，目标是促进自然语言处理在生产环境中的应用。Hanlp中有一系列"开箱即用"的静态分词器，可根据需求选取适当的分词器对文本数据进行分词与词性标注操作。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	- 字符串序列，如："数据挖掘工程师"； - 字符串单层列表，如：["数据挖掘工程师"，"数据分析"] 注意：若输入多层嵌套列表，则组件报错。
2	自定义分词词典	有需要时可添加	若传入自定义词典，则需要选择对词典的载入方式。

（3）输出

序号	名称	内容
1	data_out	分词结果或分词加词性标注结果

（4）参数

序号	分组	参数	说明
1	参数设置	分词结果输出格式	可选项有：“嵌套列表拉直，输出单层列表”、“按原格式输出”。
2	参数设置	是否输出词性标注结果	可选项有：“只输出词”、“同时输出词和词性”。
3	参数设置	对自定义词典的操作	默认“无自定义修改”。选择对自定义词典的加载方式，若无自定义词典输入，则选择“无自定义修改”。可选项有：“无自定义修改”、“动态增加”、“强行插入”，“移除词语”。
4	参数设置	选择分词器	用于词性标注。
5	输入设置	选择分词字段	
6	输入设置	文本编码	输入文件字符编码。

- 分词器选型解释：

序号	分词器	说明
1	HMM词法分词器	
2	感知机词法分词器	感知机分词是所有“由字构词”的分词器实现中最快的。
3	CRF词法分析器	对新词有很好的识别能力，但是无法利用自定义词典，对人名识别精度更高。
4	标准词法分析器	
5	NLP词法分析器	会执行全部命名实体识别和词性标注，由结构化感知机序列标注框架支撑。
6	索引词法分析器	是面向搜索引擎的分词器，能够对长词全切分。
7	极速词典词法分析器	极速分词是词典最长分词，速度极其快，精度一般。
8	最短路径词法分析器	
9	N-最短路径词法分析器	N最短路分词器比最短路分词器慢，但是效果稍微好一些，对命名实体识别能力更强
10	viterbi词法分析器	

- 词性结果的词性对照表如下：

标签	说明	标签	说明	标签	说明
a	形容词	ad	副形词	nis	机构后缀
ag	形容词性语素	al	形容词性惯用语	nit	教育相关机构
an	名形词	b	区别词	nl	名词性惯用语
begin	仅用于始##始	bg	区别语素	nm	物品名
bl	区别词性惯用语	c	连词	nmc	化学品名
cc	并列连词	d	副词	nn	工作相关名词
dg	辄,俱,复之类的副词	dl	连语	nnd	职业
e	叹词	end	仅用于终##终	nnt	职务职称
f	方位词	g	学术词汇	nr	人名
gb	生物相关词汇	gbc	生物类别	nrl	复姓
gc	化学相关词汇	gg	地理地质相关词汇	nr2	蒙古姓名
gi	计算机相关词汇	gm	数学相关词汇	nrf	音译人名
gp	物理相关词汇	h	前缀	nrj	日语人名
i	成语	j	简称略语	ns	地名
k	后缀	l	习用语	nsf	音译地名
m	数词	mg	数语素	nt	机构团体名
Mg	甲乙丙丁之类的数词	mq	数量词	ntc	公司名
n	名词	nb	生物名	ntcb	银行
nba	动物名	nbc	动物纲目	ntcf	工厂
nbp	植物名	nf	食品，比如“薯片”	ntch	酒店宾馆
ng	名词性语素	nh	医药疾病等健康相关名词	nth	医院
nhd	疾病	nhm	药品	nto	政府机构
ni	机构相关（不是独立机构名）	nic	下属机构	nts	中小学
o	拟声词	q	量词	ntu	大学
p	介词	qg	量词语素	nx	字母专名

标签	说明	标签	说明	标签	说明
pba	介词“把”	qt	时量词	nz	其他专名
pbei	介词“被”	qv	动量词	r	代词
rg	代词性语素	Rg	古汉语代词性语素	rr	人称代词
ry	疑问代词	rys	处所疑问代词	ryt	时间疑问代词
ryv	谓词性疑问代词	rz	指示代词	rzs	处所指示代词
rzt	时间指示代词	rzv	谓词性指示代词	s	处所词
t	时间词	tg	时间词性语素	u	助词
ud	助词	udel	的 底	ude2	地
ude3	得	udeng	等 等等 云云	udh	的话
ug	过	uguo	过	uj	助词
ul	连词	ule	了 喽	ulian	连 （“连小学生都会”）
uls	来讲 来说 而言 说来	usuo	所	uv	连词
uyy	一样 一般 似的 般	uz	着	uzhe	着
uzhi	之	v	动词	vd	副动词
vf	趋向动词	vg	动词性语素	vi	不及物动词（内动词）
vl	动词性惯用语	vn	名动词	vshi	动词“是”
vx	形式动词	vyou	动词“有”	w	标点符号
wb	百分号千分号，全角：％ ‰ 半角：%	wd	逗号，全角：， 半角：,	wf	分号，全角：； 半角：;
wh	单位符号，全角：￥ ＄ ￡ ° ℃ 半角：$	wj	句号，全角：。	wky	右括号，全角：） 〕 ］ ｝ 》 】 』 〉 半角：)]{>
wkz	左括号，全角：（ 〔 ［ ｛ 《 【 〖 〈 半角：([{<	wm	冒号，全角：： 半角：:	wn	顿号，全角：、
wp	破折号，全角：—— － -—— - 半角：— —-	ws	省略号，全角：…… …	wt	叹号，全角：！

标签	说明	标签	说明	标签	说明
ww	问号，全角：？	wyy	右引号，全角：”’』	wyz	左引号，全角：“‘『
x	字符串	xu	网址URL	xx	非语素字
y	语气词	yg	语气素语	z	状态词
zg	状态词				

（5）示例

用position数据集进行hanlp中文分词示例。

	A
1	期望职位
2	["数据挖掘工程师","算法工程师"]
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]
5	["数据分析师","数据挖掘工程师","算法工程师"]
6	["数据分析师","数据挖掘工程师"]
7	["算法工程师","数据分析师","机器学习工程师"]
8	["数据分析师","数据挖掘工程师"]
9	["数据分析师","其他","数据挖掘工程师"]
10	["数据分析师","图像处理工程师","机器学习工程师"]
11	["Hadoop大数据开发工程师"]
12	["Hadoop大数据开发工程师"]
13	["Hadoop大数据开发工程师"]
14	["数据分析师","机器学习工程师","图像处理工程师"]
15	["数据分析师"]
16	["数据分析师","机器学习工程师"]
17	["数据分析师","数据挖掘工程师","图像处理工程师"]
18	["数据分析师","数据挖掘工程师","机器学习工程师"]
19	["数据分析师"]
20	["数据分析师","数据挖掘工程师"]
21	["数据分析师"]
22	["数据分析师","其他"]
23	["Hadoop大数据开发工程师"]
24	["Hadoop大数据开发工程师"]
25	["数据分析师","其他"]
26	["Hadoop大数据开发工程师"]
27	["数据分析师","自然语言处理工程师"]
28	["Hadoop大数据开发工程师","数据分析师","其他"]

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行hanlp分词，将【hanlp分词】组件与输入源连接，在参数配置中选择分词输出格式、是否标注词性以及分词器，因为此次操作没有添加自定义词典则无需修改对词典的操作，在输入设置中选择需要进行分词的特征列，右键单击【hanlp分词】组件，选择“运行该节点”。

查看日志。对【hanlp分词与词性】右键选择查看日志，即可观测到文本数据的分词结果与词性标注结果。

8.7.6 分句

（1）作用

该组件用于对字符串数据进行分句，分句的依据为输入的标点符号，格式为：，。？！等。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out	分句结果

（4）参数

序号	分组	参数	说明
1	字段配置	选择分句字段	分词处理的最大单词数量
2	字段配置	分割符	分句的符号依据

（5）示例

对position数据集进行分句示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行分句，将【分句】组件与输入源连接，在参数配置中选择“文本列”并填写分割符号“，”，右键单击【分句】组件，选择“运行该节点”。

打开数据，查看结果。对【分句】组件右击，点击“查看数据”，即可查看根据目标符号分句后的文本数据。

查看日志

	期望职位	期望职位_sentens
0	[数据挖掘工程师，算法工程师]	[[数据挖掘工程师]，[算法工程师]]
1	[数据分析师，数据挖掘工程师，自然语言处理工程师]	[[数据分析师]，[数据挖掘工程师]，[自然语言处理工程师]]
2	[数据分析师，自然语言处理工程师，数据挖掘工程师]	[[数据分析师]，[自然语言处理工程师]，[数据挖掘工程师]]
3	[数据分析师，数据挖掘工程师，算法工程师]	[[数据分析师]，[数据挖掘工程师]，[算法工程师]]
4	[数据分析师，数据挖掘工程师]	[[数据分析师]，[数据挖掘工程师]]
..	...	...
65	[数据分析师，数据挖掘工程师，机器学习工程师]	[[数据分析师]，[数据挖掘工程师]，[机器学习工程师]]
66	[数据分析师，Hadoop大数据开发工程师，其他]	[[数据分析师]，[Hadoop大数据开发工程师]，[其他]]
67	[数据分析师，数据挖掘工程师]	[[数据分析师]，[数据挖掘工程师]]
68	[数据分析师，数据挖掘工程师]	[[数据分析师]，[数据挖掘工程师]]
69	[数据分析师，数据挖掘工程师]	[[数据分析师]，[数据挖掘工程师]]

[70 rows x 2 columns]

8.7.7 停用词过滤

（1）作用

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为Stop Words（停用词）。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	载入停用词文件	txt文件	

（3）输出

序号	名称	内容
1	data_out	过滤停用词后的文本结果

（4）参数

序号	分组	参数	说明
1	字段配置	选择需要过滤的字段	
2	参数配置	停用词过滤后是否允许输出嵌套列表	

（5）示例

对position数据集进行停用词过滤示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行停用词过滤，将【停用词过滤】组件与输入源连接，一个为需要过滤的文本数据输入点，一个为停用词文本输入点，在参数配置中选择“文本列”与文本输出格式，右键单击【停用词过滤】组件，选择“运行该节点”。

打开数据，查看结果。对【停用词过滤】组件右击，点击“查看数据”，即可查看过滤停用词后的文本结果。

8.7.8 正则匹配

（1）作用

该组件主要用于完成对于字符串的匹配、提取、替换、分割任务。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out.csv	执行匹配后的文本结果

（4）参数

序号	分组	参数	说明
1	输入配置	数据输入处理	是否把输入数据当做字符串，只允许字符串或列表形式的字符串
2	参数设置	匹配方式	可选项有“字符提取”，“字符替换”，“字符匹配”，“字符切割”
3	字符替换参数	目标表达式	输入目标表达式，不要使用引号
4	字符替换参数	替换的表达式	
5	字符替换参数	替换次数	
6	字符替换参数	结果返回形式	可选项有：字符串、元组（显示替换的次数）
7	字符匹配参数	匹配位置	可选项有：从匹配字符串位置开始、从任意位置开始
8	字符匹配参数	字符匹配表达式	
9	字符匹配参数	结果返回形式	可选项有：匹配字符的位置、匹配的字符、匹配结束的位置、匹配开始的位置
10	字符匹配参数	匹配方式	
11	字符提取参数	提取表达式	
12	字符分割参数	字符分割表达式	
13	字符分割参数	分割最大次数	
14	数据选型	选择数据文本字段	

（5）示例

对position数据集进行正则匹配示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行正则匹配，将【正则匹配】组件与输入源连接，在数据选型中选择需要进行正则匹配的特征列，并选择数据输入处理，在匹配操作中选择具体的操作，如“字符匹配”操作，对相对应的字符匹配参数进行设置，填写具体匹配的表达式与匹配结果方式，其他参数设置不做处理，右键单击【正则匹配】组件，选择“运行该节点”。

打开数据，查看结果。对【正则匹配】组件右击，点击“查看数据”，即可查看匹配操作后的文本结果。

8.7.9 新词发现

（1）作用

在文本处理中由于新词的不断增加，也使中文分词结果中出现过多的“散串”，进而影响了分词的准确率。新词发现也可称为未登录词识别，严格来讲，新词是指随时代发展而新出现或旧词新用的词语。同时，可以认为特定领域的专有名词也可归属于新词的范畴。

该组件是基于统计方法， 将文本按字符分割后拼接为候选词，通过计算每一个候选词左右邻字丰富程度和内部凝聚程度得到最终短语，即将文本中的新词展示出来，其中包括有各类专有名词（人名、地名、企业名）、缩写词、流行词汇等等。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	自定义新词词典	txt文件	

（3）输出

序号	名称	内容
1	data_out	新词文本结果

（4）参数

序号	分组	参数	说明
1	参数设置	最低词频	预料中新词最低出现的次数
2	参数设置	新词最小字符数	输出的新词长度将大于等于该取值
3	参数设置	是否过滤新词	是否将发现的新词与已知词汇表进行比较，输出词汇表中未出现的词汇。 仅当自定义的新词词典不为空时，该参数有效
4	参数设置	新词最大字符数	,输出的新词长度将小于等于该取值
5	参数设置	新词个数	
6	输入设置	特征	选择所需的文本列

（5）示例

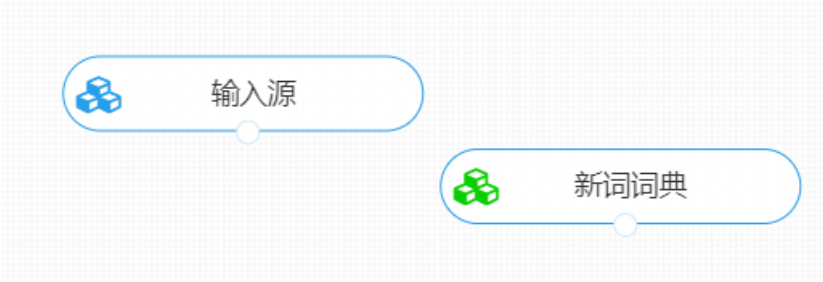
对position数据集进行新词发现示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

拖入新的输入源组件，将自定义的新词词典读入系统，操作同以上读入position数据集一样。为了方便辨认组件，操作中可对其中的组件名称进行修改，右键选择对组件重命名即可。



进行新词发现，将【新词发现】组件与输入源连接，注意要连接相对应的输入点，在参数配置中设置对应的参数，最低词频为2，新词最小字符数为2，最大字符数为4，输出的新词个数设置为10，右键单击【新词发现】组件，选择“运行该节点”。

打开数据，查看结果。对【新词发现】组件右击，点击“查看数据”，即可查看算法得出的新词结果。

预览数据

_c0	new_word
0	处理工程
1	处理工
2	工程师
3	计算机视
4	算机视觉
5	机视觉
6	自然语言

8.7.10 简繁体转换

（1）作用

在日常的NLP处理中，尤其是针对新闻、博客、评论等互联网公开数据，由于输入法和语言使用习惯等差异，通常会出现同一语料下中文文本繁体和简体共存的情况，若不统一字体，将影响后续文本向量化、形似度计算等处理效果，进而影响最终研究目标的准确率。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out	简繁转换结果

（4）参数

序号	分组	参数	说明
1	参数设置	转换模型	繁体转简体、简体转繁体
2	输入设置	选择字段	

（5）示例

对position数据集进行简转繁示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行简繁体转换，将【简繁体转换】组件与输入源连接，在参数配置中选择转换模型“简体转繁体”，在输入设置中选择文本列，右键单击【简繁体转换】组件，选择“运行该节点”。

打开数据，查看结果。对【简繁体转换】组件右击，点击“查看数据”，即可查字体转换后的数据结果。

预览数据

	_c0	期望职位	期望职位_change
0		["数据挖掘工程师","算法工程师"]	['數據挖掘工程師', '算法工程師']
1		["数据分析师","数据挖掘工程师","自然语言处理工程师"]	['數據分析師', '數據挖掘工程師', '自然語言處理工程師']
2		["数据分析师","自然语言处理工程师","数据挖掘工程师"]	['數據分析師', '自然語言處理工程師', '數據挖掘工程師']
3		["数据分析师","数据挖掘工程师","算法工程师"]	['數據分析師', '數據挖掘工程師', '算法工程師']
4		["数据分析师","数据挖掘工程师"]	['數據分析師', '數據挖掘工程師']

8.7.11 文本纠错

（1）作用

文本纠错是自然语言处理的常见任务之一，通常包括以下几种错误类型：

- 谐音字词，如 配副眼睛-配副眼镜；
- 混淆音字词，如 流浪织女-牛郎织女；
- 字词顺序颠倒，如 伍迪艾伦-艾伦伍迪；
- 字词补全，如 爱有天意-假如爱有天意；
- 形似字错误，如 高梁-高粱；
- 中文拼音全拼，如 xingfu-幸福；
- 中文拼音缩写，如 sz-深圳。
- 语法错误，如 想象难以-难以想象。

当然，针对不同业务场景，这些问题并不一定全部存在，比如输入法中需要处理前四种，搜索引擎需要处理所有类型，语音识别后文本纠错只需要处理前两种， 其中'形似字错误'主要针对五笔或者笔画手写输入等。

文本纠错处理通常划分为两个步骤：第一步是错误检测，第二步是错误纠正。该算法实现文本纠错的具体步骤如下：① 错误检测部分先通过结巴中文分词器切词，由于句子中含有错别字，所以切词结果往往会有切分错误的情况，这样从字粒度和词粒度两方面检测错误， 整合这两种粒度的疑似错误结果，形成疑似错误位置候选集；② 错误纠正部分，是遍历所有的疑似错误位置，并使用音似、形似词典替换错误位置的词，然后通过语言模型计算句子困惑度，对所有候选集结果比较并排序，得到最优纠正词。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	自定义词典	txt文件 非必选	通过加载自定义混淆集，支持用户纠正已知的错误。 内容格式为：天真浪慢 天真浪漫 每一行第一个为错误词汇，第二个为正确词汇。
3	自定义模型	.klm文件	支持用户加载自己训练的kenlm语言模型，模型以“.klm”文件形式存储。

（3）输出

序号	名称	内容
1	data_out	进行错误检测或者纠正处理后的文本数据

（4）参数

序号	分组	参数	说明	
1	输入设置	选择字段		
2	参数设置	是否加载自定义模型	当输入自定义语言模型时，该参数有效	
3	参数设置	是否开启字粒度检测或纠正	模型默认进行词粒度的纠正,字粒度的纠正准确率较低	
4	参数设置	是否加载自定义混淆集	当输入自定义混淆集时，该参数有效	
5	参数设置	是否纠正	若不纠正则只进行错误检测	

（5）示例

对position数据集进行文本纠正示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行文本纠错，将【文本纠错】组件与输入源连接，在输入设置中选择所需文本列，在参数配置中保持默认设置，即不加载自定义混淆集和自定义模型，对检测结果进行纠正处理，右键单击【文本纠错】组件，选择“运行该节点”。

打开数据，查看结果。对【文本纠错】组件右击，点击“查看数据”，即可查看文本纠错结果。其中列名尾缀为“*correction_text*”*为文本纠正结果*，“correcrtion_detail”的为纠正详情。

预览数据

_c0	期望职位	期望职位_correction_text	期望职位_correction_details
0	["数据挖掘工程师","算法工程师"]	['数据挖掘工程师', '算法工程师']	[[], []]
1	["数据分析师","数据挖掘工程师","自然语言处理工程师"]	['数据分析师', '数据挖掘工程师', '自然语言处理工程师']	[[], [], []]
2	["数据分析师","自然语言处理工程师","数据挖掘工程师"]	['数据分析师', '自然语言处理工程师', '数据挖掘工程师']	[[], [], []]
3	["数据分析师","数据挖掘工程师","算法工程师"]	['数据分析师', '数据挖掘工程师', '算法工程师']	[[], [], []]

8.7.12 拼音转换

（1）作用

拼音转换可分为拼音转汉字与汉字转拼音两个方面：

- 汉字转为拼音，可用于汉字注音、排序、检索等。应用场景有：垃圾广告中的同音字识别（如公众号、工仲号、公众號）、敏感信息过滤中的同音字过滤（如六合彩、六和彩、溜喝彩）等，将汉字转换成拼音后再进行深度学习分类，以优化同音字的分类效果。

  实现步骤如下：

  1. 对输入的字符串按是否是汉字进行分词；
  2. 对分词结果的每个词条按如下步骤进行获取词条拼音的逻辑（即词语-拼音转换逻辑）：
     ① 检查词条是否是汉字，不是汉字则走无拼音处理逻辑；
     ② 检查词条是否在词语-拼音词典（PHRASES_DICT）中，如果在直接取词典中词条对应的拼

音数据；
③ 如果词条不在词典中，遍历词条包含的字符，每个字符进行单字-拼音转换逻辑处理；

3. 单字-拼音转换逻辑：
① 检查字符是否在单字-拼音词典（PINYIN_DICT）中，若在则取词典中该字符对应的拼音数据；
② 若不在，则走无拼音处理逻辑；
4. 无拼音处理逻辑：根据errors参数的值，返回不同的处理结果；
5. 对上述步骤获得的拼音数据按指定的拼音风格进行转换；
6. 对风格转换后的数据按格式输出。

- 拼音转汉字，可作为拼音输入法的转换引擎，利用拼音获取汉字文本。

（2）输入

- 汉字转拼音

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	数据特征列格式	文本列格式：① 汉字字符串，如：'数据分析师'； ② 汉字字符列表，如：['数据分析师', '数据挖掘师']	字符列表必须为单层列表

- 拼音转汉字

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	数据特征列格式	文本列格式：拼音字符列表，可为单层列表或双层列表 如：['shu', 'ju', 'fen']或[['shu', 'fen'], ['fen', 'xi']]。	字符列表必须为单层列表

（3）输出

序号	名称	内容
1	data_out	拼音转换结果

（4）参数

- 汉字转拼音

序号	分组	参数	说明
1	参数配置	是否启用多音字	当启用多音字时，以嵌套列表形式对多音字输出所有拼音结果。
2	参数配置	无拼音字处理	对非汉字或无效字符可指定其处理方式，可选方式有：（1）保留该原始字符；（2）忽略该字符，即该字符的输出结果为空。
3	参数配置	是否遵守《汉语拼音方案》标准	该参数用于控制处理声母和韵母时是否严格遵循 《汉语拼音方案》 标准。
4	参数配置	自定义单字拼音	非必填。用于自定义输入单字拼音库，以修正输出结果。输入示例：{"数": "shu2","据": "jv"}。
5	参数配置	自定义词语拼音	非必填。用于自定义输入词语拼音库，以修正输出结果。输入示例：{"数据": [["shu2"], ["jv2"]], "分析": [["fen1"], ["xi1"]]}。
6	参数配置	拼音标注风格	非必填。用于指定拼音风格。可选项有： NORMAL：不带声调；TONE：标准声调（默认）；TONE2：声调风格2； TONE3：声调风格3；INITIALS：仅声母；FIRST_LETTER：仅首字母； FINALS：仅韵母；FINALS_TONE：标准韵母；FINALS_TONE2：韵母风格2； FINALS_TONE3：韵母风格3；BOPOMOFO：注音风格； BOPOMOFO_FIRST：注音风格-仅首字母； CYRILLIC：汉语拼音与俄语字母对照风格； CYRILLIC_FIRST：汉语拼音与俄语字母对照风格-仅首字母。
7	输入设置	特征	选择需要进行拼音转换的特征列

- 拼音转汉字

序号	分组	参数	说明
1	参数配置	转换算法选择	可选项有 ① viterbi：基于HMM的转换； ② DAG：基于DAG的转换，应用词库+动态规划算法。
2	参数配置	是否使用对数打分	对算法转换出的多个汉字应用对数进行打分，选取分数最高的汉字作为输出结果。

（5）示例

对position数据集进行拼音转换示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行汉字转拼音，将【汉字转拼音】组件与输入源连接，在参数配置中进行相对应的设置，是否启用多音字，选择"否"，无拼音字符处理方式，选择"忽略该字符"，是否遵守《汉语拼音方案》标准，选择"是"，拼音标注风格，选择"标准声调（默认）"，自定义单字拼音，不填入内容，自定义词语拼音，不填入内容。右键单击【汉字转拼音】组件，选择"运行该节点"。

打开数据，查看结果。对【汉字转拼音】组件右键选择查看数据，即可查看文本中汉字转拼音的结果。

预览数据

_c0	期望职位	期望职位_pinyin
0	['数据挖掘工程师', '算法工程师']	['shù', 'jù', 'wā', 'jué', 'gōng', 'chéng', 'shī', 'suàn', 'fǎ', 'gōng', 'chéng', 'shī']
1	['数据分析师', '数据挖掘工程师', '自然语言处理工程师']	['shù', 'jù', 'fēn', 'xī', 'shī', 'shù', 'jù', 'wā', 'jué', 'gōng', 'chéng', 'shī', 'zi', 'rán', 'yǔ', 'yán', 'chǔ', 'lǐ', 'gōng', 'chéng', 'shī']
2	['数据分析师', '自然语言处理工程师', '数据挖掘工程师']	['shù', 'jù', 'fēn', 'xī', 'shī', 'zi', 'rán', 'yǔ', 'yán', 'chǔ', 'lǐ', 'gōng', 'chéng', 'shī', 'shù', 'jù', 'wā', 'jué', 'gōng', 'chéng', 'shī']

② 进行拼音转汉字，将【拼音转汉字】组件与上面的汉字转拼音的输出结果进行连接，在参数设置中选择转换算法为dag和使用对数打分，输入设置中勾选上面的转换结果。

打开数据，查看结果。对【拼音转汉字】组件右键选择查看数据，即可查看拼音转汉字的结果，其结果用_pinyin_to_hanzi后缀标注。

预览数据

_c0	Unnamed: 0	期望职位	期望职位_pinyin	期望职位_pinyin_to_hanzi
0	0	['数据挖掘工程师', '算法工程师']	['shù', 'jù', 'wā', 'jué', 'gōng', 'chéng', 'shī', 'suàn', 'fǎ', 'gōng', 'chéng', 'shī']	['数据挖掘工程师算法工程师']
1	1	['数据分析师', '数据挖掘工程师', '自然语言处理工程师']	['shù', 'jù', 'fēn', 'xī', 'shī', 'shù', 'jù', 'wā', 'jué', 'gōng', 'chéng', 'shī', 'zi', 'rán', 'yǔ', 'yán', 'chǔ', 'lǐ', 'gōng', 'chéng', 'shī']	['数据分析师数据挖掘工程师自然语言处理工程师']
2	2	['数据分析师', '自然语言处理工程师', '数据挖掘工程师']	['shù', 'jù', 'fēn', 'xī', 'shī', 'zi', 'rán', 'yǔ', 'yán', 'chǔ', 'lǐ', 'gōng', 'chéng', 'shī', 'shù', 'jù', 'wā', 'jué', 'gōng', 'chéng', 'shī']	['数据分析师自然语言处理工程师数据挖掘工程师']

8.7.13 生僻字检测

（1）作用

生僻字，即不常见的或人们不熟悉的汉字。 生僻字无明确的判定标准，但可通过字符编码的转换进行识别。调用该组件时可以通过检测是否为非指定文本编码的字符，或字符的Unicode编码范围是否为扩展库范围，以此判断是否属于生僻字。

生僻字的检测可应用于文章生僻字注音、生僻字存储或读写异常等。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	特征列格式	字符串序列、字符串列表	

（3）输出

序号	名称	内容
1	data_out	生僻字检测结果

（4）参数

序号	分组	参数	说明
1	参数设置	输出格式	可选项有： 以True或False形式输出；仅输出生僻字； 在原文中对生僻字标注拼音
2	输入设置	特征列	
3	输入设置	文本编码	

（5）示例

对position数据集进行生僻字检测示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行生僻字检测，将【生僻字检测】组件与输入源连接，在参数配置中选择“文本列”与结果输出形式，此处选择仅对文本中的生僻字进行输出， 右键单击【生僻字检测】组件，选择“运行该节点”。

打开数据，查看结果。对【生僻字检测】组件右键选择查看数据，即可查看检测结果，其中检测结果用_rare_words后缀标注。

查看日志

	期望职位	期望职位_rare_words
0	[数据挖掘工程师, 算法工程师]	[]
1	[数据分析师, 数据挖掘工程师, 自然语言处理工程师]	[]
2	[数据分析师, 自然语言处理工程师, 数据挖掘工程师]	[]
3	[数据分析师, 数据挖掘工程师, 算法工程师]	[]
4	[数据分析师, 数据挖掘工程师]	[]
..	...	...
65	[数据分析师, 数据挖掘工程师, 机器学习工程师]	[]
66	[数据分析师, Hadoop大数据开发工程师, 其他]	[]
67	[数据分析师, 数据挖掘工程师]	[]
68	[数据分析师, 数据挖掘工程师]	[]
69	[数据分析师, 数据挖掘工程师]	[]

[70 rows x 2 columns]

8.7.14 数据增强-回译

（1）作用

回译数据增强目前是文本数据增强方面效果较好的增强方法，，一般基于google翻译接口，，将文本数据翻译成另外一种语言(一般选择小语种），之后再翻译回原语言，即可认为得到与与原语料同标签的新语料，这种方式不仅有同义词替换，词语增删，还具有对句子结构语序调整的效果，并还能保持与原句子意思相近。新语料加入到原数据集中即可认为是对原数据集数据增强。

目前该组件仅支持中英文回译。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	数据特征列格式	字符串序列、字符串单层列表	例：”数据分析师“ ["数据分析师"，“数据挖掘”] 若输入多层嵌套列表，则组件报错

（3）输出

序号	名称	内容
1	data_out	回译结果

（4）参数

序号	分组	参数	说明
1	输入配置	特征	选择目标文本列
2	输入设置	文本编码	输入文件字符编码，如“utf8"

（5）示例

对position数据集进行文本回译示例。

	A	B	C	D	E	F
1	movieRow	movieId	title			
2	0	1	Toy Story (1995)			
3	1	2	Jumanji (1995)			
4	2	3	Grumpier Old Men (1995)			
5	3	4	Waiting to Exhale (1995)			
6	4	5	Father of the Bride Part II (1995)			
7	5	6	Heat (1995)			
8	6	7	Sabrina (1995)			
9	7	8	Tom and Huck (1995)			
10	8	9	Sudden Death (1995)			
11	9	10	GoldenEye (1995)			
12	10	11	American President, The (1995)			
13	11	12	Dracula: Dead and Loving It (1995)			
14	12	13	Balto (1995)			
15	13	14	Nixon (1995)			
16	14	15	Cutthroat Island (1995)			
17	15	16	Casino (1995)			
18	16	17	Sense and Sensibility (1995)			
19	17	18	Four Rooms (1995)			
20	18	19	Ace Ventura: When Nature Calls (1995)			
21	19	20	Money Train (1995)			
22	20	21	Get Shorty (1995)			
23	21	22	Copycat (1995)			
24	22	23	Assassins (1995)			
25	23	24	Powder (1995)			
26	24	25	Leaving Las Vegas (1995)			
27	25	26	Othello (1995)			
28	26	27	Now and Then (1995)			

movies

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行回译，将【回译】组件与输入源连接，在输入设置中选择“特征”便填写文本编码，右键单击【词袋模型】组件，选择“运行该节点”。

打开数据，查看结果。对【回译（中英）】组件右键选择查看数据，即可查看回译的文本结果，其中回译结果列名以“_translation”进行进行后缀标注。

8.7.15 数据增强-EDA

（1）作用

EDA（easy data augmentation）是一种应用于文本分类的简单的数据增强技术，由4种方法组成，分别是：同义词替换（对非停用词的词语进行同义词替换）、随机插入（不去停用词，直接随机选取词语，做近义词，进行插入操作）、随机替换与随机删除。EDA的文本增强技术对小样本在模型中的学习性能有显著的提高。传入数据需要有原数据列、分词列、去停用词列。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	数据列内容	原数据列、分词列、去停用词列	

（3）输出

序号	名称	内容
1	data_out	文本增强结果

（4）参数

序号	分组	参数	说明
1	字段设置	分词后的数据列	选择目标文本列
2	字段设置	去停用词后的数据列	输入文件字符编码，如“utf8"
3	字段设置	特征	选择需要的文本列
4	参数设置	EDA方式	可选项有： 同义词替换、随机插入、随机转换、随机删除
5	参数设置	进行几次随机插入/随机变换	进行随机插入或随机变换时设置有效

（5）示例

对position数据集进行EDA文本增强示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行分词操作，将【分词】组件与输入源连接，在参数配置中进行相对应设置（具体可查看8.7.5Jieba分词章节的操作），右键单击【分词】组件，选择“运行该节点”，生成带有分词的文本数据。

② 进行去除停用词操作，将【过滤停用词】组件与分词结果进行连接，同时拖入新的输入源，读入停用词数据，与【过滤停用词】进行连接，字段设置中的特征选择①步骤输出的“期望职位_cut_words”，算法引擎选取python，结果格式以原格式输出，右键单击【过滤停用词】组件，选择“运行该节点”，生成过滤停用词后的分词文本数据。现下右键对【过滤停用词】组件查看数据，可得数据中包含的内容如下：

预览数据

??望职位	期望职位_cut_words	期望职位_cut_words_stopwords
["数据挖掘工程师","算法工程师"]	['数据挖掘', '工程师', '算法', '工程师']	['数据挖掘', '工程师', '算法', '工程师']
["数据分析师","数据挖掘工程师","自然语言处理工程师"]	['数据', '分析师', '数据挖掘', '工程师', '自然语言', '处理', '工程师']	['数据', '分析师', '数据挖掘', '工程师', '自然语言', '处理', '工程师']
["数据分析师","自然语言处理工程师","数据挖掘工程师"]	['数据', '分析师', '自然语言', '处理', '工程师', '数据挖掘', '工程师']	['数据', '分析师', '自然语言', '处理', '工程师', '数据挖掘', '工程师']

③ 进行EDA文本增强操作，将【EDA】组件与【过滤停用词组件】进行连接，在字段设置中分词后的数据列选择“期望职位*cut_words*”，*去停用词后的数据列选择“期望职位*cut_words_stopwords*”,特征中勾选“期望职位*”，“*期望职位*”*cut_words”，“期望职位cut_words_stopwords”，参数设置中对于EDA方式选择”同义替换“操作，右键单击【EDA】组件运行该节点。

打开结果，查看数据。对【EDA】组件右键查看数据，即可查看同义词替换的结果，其中文本增强结果以“synonym_replace”列保存。

预览数据

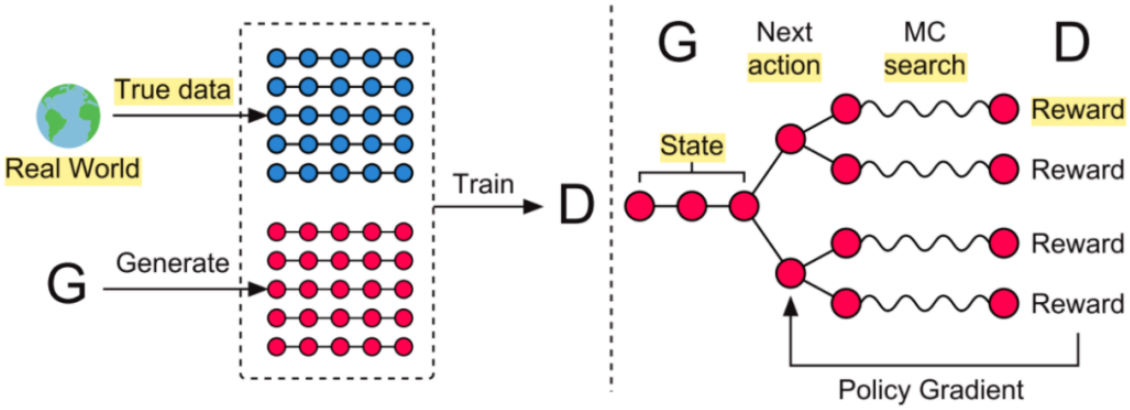
??望职位	期望职位_cut_words	期望职位_cut_words_stop words	synonym_replace
["数据挖掘工程师","算法工程师"]	['数据挖掘', '工程师', '算法', '工程师']	['数据挖掘', '工程师', '算法', '工程师']	['数据挖掘', '项目经理', '算法', '项目经理']
["数据分析师","数据挖掘工程师","自然语言处理工程师"]	['数据', '分析师', '数据挖掘', '工程师', '自然语言', '处理', '工程师']	['数据', '分析师', '数据挖掘', '工程师', '自然语言', '处理', '工程师']	['数据', '分析师', '数据挖掘', '电气工程师', '自然语言', '处理', '电气工程师']
["数据分析师","自然语言处理工程师","数据挖掘工程师"]	['数据', '分析师', '自然语言', '处理', '工程师', '数据挖掘', '工程师']	['数据', '分析师', '自然语言', '处理', '工程师', '数据挖掘', '工程师']	['数据', '分析师', '自然语言', '处理', '工程师', '数据挖掘', '工程师']
["数据分析师","数据挖掘工程师","算法工程师"]	['数据', '分析师', '数据挖掘', '工程师', '算法', '工程师']	['数据', '分析师', '数据挖掘', '工程师', '算法', '工程师']	['数据', '分析师', '数据挖掘', '工程师', '算法', '工程师']

8.7.16 数据增强-seqGAN

GAN属于无监督学习，由两个神经网络组成：一个是生成器（generator），一个是判别模型（discriminator）。生成器的任务是生成看起来逼真与原始数据相似的样本。判别器的任务是判断生成模型生成的样本是真实的还是伪造的。具体而言，生成器（generator）会从潜在空间中随机获取样本，并与真实数据一起作为判别器（discriminator）的输入。判别器是一个经典分类器，作用是把真实数据和生成数据尽量分开。判别器对生成数据的判别结果会返回给生成器，训练生成器生成更多能够成功“骗过”判别器的数据。训练的最终目的是使生成样本的分布与真实数据达到一致，判别器完全不能区分真伪。

该组件使用的GAN模型为seqGAN。

SeqGAN结构如下图所示，已经存在的红色圆点称为现在的状态（state），要生成的下一个红色圆点称作动作（action），因为D需要对一个完整的序列评分，所以就是用MCTS（蒙特卡洛树搜索）将每一个动作的各种可能性补全，D对这些完整的序列产生reward，回传给G，通过增强学习更新G。这样就是用Reinforcement learning的方式，训练出一个可以产生下一个最优的action的生成网络。



（2）输入

序号	条件	要求	说明
1	载入文件格式	txt文件	
2	文本处理要求	传入的数据为经词袋模型处理后的数据	

（3）输出

序号	名称	内容
1	data_out.csv	生成词袋向量结果

（4）参数

序号	分组	参数	说明
1	参数设置	生成器批量大小	
2	参数设置	预训练批量数	
3	参数设置	判别器预训练次数	
4	参数设置	对抗训练的总批次	

（5）示例

对准备好词袋数据的进行GAN示例。

real_cn_number_data.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
0 0 0 0 0 0 0 0 0 0 0 0 10 11 12 13 14 15 16 17
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 18 19 20 21
0 0 0 0 0 0 0 0 0 0 0 0 0 22 23 24 25 26 27 28 4
0 0 0 0 0 0 0 0 0 0 29 30 31 32 33 34 35 36 37 38
0 0 0 0 0 0 0 0 0 0 0 0 0 0 39 40 41 42 43 44 45
5 49 50 1 51 1 52 53 1 54 55 56 57 58 59 1 60 61 62 63
0 0 0 0 0 0 0 0 0 64 6 65 66 67 6 3 68 69 70 71
0 0 0 72 73 2 7 74 8 2 75 76 9 77 78 79 8 2 7 80
0 0 0 0 0 0 0 0 0 0 0 81 82 83 84 85 86 87 88 89
0 90 91 92 93 94 95 96 97 9 4 98 99 100 101 102 103 104 105 106
```

首先将词袋数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“ci”，勾选文件“real_cn_number_data.txt”，右键单击【输入源】算法，选择“运行该节点”。

进行seqGAN，将【seqGAN】组件与输入源连接，在参数配置中进行相应的参数设置，右键单击【seqGAN】组件，选择“运行该节点”。

打开数据，查看结果。对【SeqGAN】组件右键点击查看数据，即可查看经过训练生成的相似文本词向量。

69 30 26 69 30 69 17 69 0 6 9 69 69 69 69 88 69 50 30 3 57 69
69 0 69 0 0 69 0 30 69 69 0 30 106 0 69 69 0 50 0 69
69 69 30 0 69 69 69 4237 6 9 69 30 69 69 1477 69 134 69 26 69 69
1 0 0 0 0 0 0 0 69 0 0 69 0 3 0 69 69 69 69 0 0
50 0 0 69 69 0 17 69 22 69 69 0 69 0 0 0 0 0 69 69

8.7.17 依存句法LTP

（1）作用

句法分析是自然语言处理中的关键技术之一，其基本任务是确定句子的句法结构或者句子中词汇之间的依存关系。主要包括两方面的内容，一是确定语言的语法体系，即对语言中合法的句子的语法结构给与形式化的定义，即语义依存；另一方面是句法分析技术，即根据给定的语法体系，自动推导出句子的句法结构，分析句子所包含的句法单位和这些句法单位之间的关系，即句法依存。

随着自然语言应用的日益广泛，特别是对文本处理需求的进一步增加，句法分析的作用愈加突出，它在机器翻译、信息检索与抽取、问答系统、语音识别等研究领域中都有重要的应用价值。

LTP依存句法通过分析语言单位内成分之前的依存关系解释其句法结构，主张句子中核心动词是支配其他成分的中心成分。而它本身却不受其他任何成分的支配，所有受支配成分都以某种关系从属于支配者。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	文本特征格式	采用ltp分词结果	

（3）输出

序号	名称	内容
1	data_out	句子的全部依存关系是由多个tuple组成的列表，每一个tuple包含五个元素，从左到有分别表示：开始词汇在句子中的索引（从1开始计算）、开始词汇、结束词汇在句子中的索引、结束词汇，开始词汇与结束词汇依存关系，其中索引为0表示该词语为句子的核心词汇。 句子的提存关系的提取是包含多种依存关系的字段，其中key为各依存关系，value为属于这一依存关系的tuple组成的列表，若句子不带有该依存关系，将默认返回空列表。

（4）参数

关系类型	Tag	Description	Example
主谓关系	SBV	subject-verb	我送她一束花 (我 <– 送)
动宾关系	VOB	直接宾语，verb-object	我送她一束花 (送 –> 花)
间宾关系	IOB	间接宾语，indirect-object	我送她一束花 (送 –> 她)
前置宾语	FOB	前置宾语，fronting-object	他什么书都读 (书 <– 读)
兼语	DBL	double	他请我吃饭 (请 –> 我)
定中关系	ATT	attribute	红苹果 (红 <– 苹果)
状中结构	ADV	adverbial	非常美丽 (非常 <– 美丽)
动补结构	CMP	complement	做完了作业 (做 –> 完)
并列关系	COO	coordinate	大山和大海 (大山 –> 大海)
介宾关系	POB	preposition-object	在贸易区内 (在 –> 内)
左附加关系	LAD	left adjunct	大山和大海 (和 <– 大海)
右附加关系	RAD	right adjunct	孩子们 (孩子 –> 们)
独立结构	IS	independent structure	两个单句在结构上彼此独立
核心关系	HED	head	指整个句子的核心

（2）输入

序号	分组	参数	说明
1	参数配置	主谓关系	
2	输入设置	依存句法分析字段	
3	输入设置	LTP分词结果字段	
4	输入设置	LTP词性标注结果字段	
5	输出设置	输出格式设置	是否选择输出嵌套列表

（5）示例

对position数据集进行依存句法示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行ltp分词，将【LTP分词与词性标注】组件与输入源连接，在输入配置中选择需要进行分词的特征列，在参数设置中使用平台默认的分词字典，右键单击【LTP分词与词性标注】组件，选择“运行该节点”。

② 进行依存句法TLP，将【LTP依存句法】组件与①的输出点进行连接，在输入配置中选择相对应的特征列，右键单击【LTP依存句法】组件，选择“运行该节点”。

打开数据，查看结果。对【LTP依存句法】组件右键选择查看数据，即可查看文本的分词之间的依存关系。其中结果以"_arcs"尾缀进行标识。

8.7.18 依存句法Hanlp

（1）作用

依存句法分析（Dependency Parsing，DP）通过分析语言单位内成分之间的依存关系，揭示其句法结构。HanLP依存句法分析应用基于神经网络的高性能依存句法分析器，实现句子的句法结构识别。进而可对识别出的依存关系，进行依存关系提取。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	文本特征格式	字符串序列、字符串列表	

（3）输出

序号	名称	内容
1	data_out	（1）包含的所有关系：以特征选择中选择的待提取字段名+“*all_relation*”进行命名。*以列表嵌套元组的格式输出，元组中含5个元素，依次为：支配词索引、支配词、从属词索引、从属词、依存关系。如：[(1,'我',2,'送','主谓关系')]。* *（2）参数选择中，选择要提取的关系：输出字段名，以特征选择中选择的待提取字段名*+“relation”进行命名。以列表嵌套字典的格式输出，一个字典为一个句子的提取结果，字典中的值为列表嵌套元组的格式。如：[{'主谓关系':[(1,'我',2,'送','主谓关系')]}]。

（4）参数

序号	分组	参数	说明
1	输入配置	特征	分词处理的最大单词数量
2	输入设置	文本编码	
3	输出设置	输出文件类型	
4	参数设置	选择需提取的依存关系	关系类型如下

关系类型	Tag	Description	Example
主谓关系	SBV	subject-verb	我送她一束花 (我 <– 送)
动宾关系	VOB	直接宾语，verb-object	我送她一束花 (送 –> 花)
间宾关系	IOB	间接宾语，indirect-object	我送她一束花 (送 –> 她)
前置宾语	FOB	前置宾语，fronting-object	他什么书都读 (书 <– 读)
兼语	DBL	double	他请我吃饭 (请 –> 我)
定中关系	ATT	attribute	红苹果 (红 <– 苹果)
状中结构	ADV	adverbial	非常美丽 (非常 <– 美丽)
动补结构	CMP	complement	做完了作业 (做 –> 完)
并列关系	COO	coordinate	大山和大海 (大山 –> 大海)
介宾关系	POB	preposition-object	在贸易区内 (在 –> 内)
左附加关系	LAD	left adjunct	大山和大海 (和 <– 大海)
右附加关系	RAD	right adjunct	孩子们 (孩子 –> 们)
独立结构	IS	independent structure	两个单句在结构上彼此独立
核心关系	HED	head	指整个句子的核心

（5）示例

对position数据集进行句法关系提取示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行Hanlp依存句法，将【Hanlp依存句法】组件与输入源连接，在参数配置中选择需要进行分词的特征列，选择需提取的依存关系，右键单击【Hanlp依存句法】组件，选择“运行该节点”。

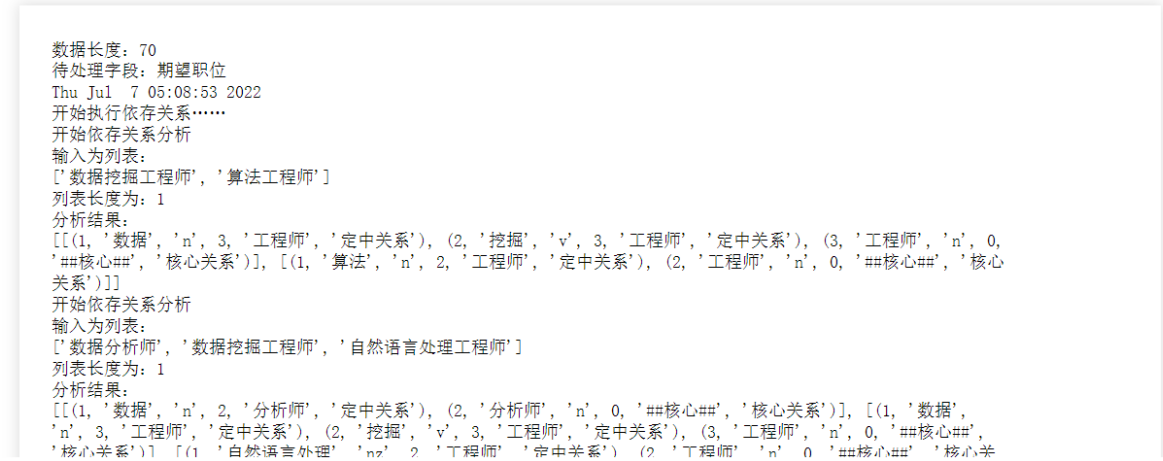
打开数据，查看结果。对【Hanlp依存句法】组件右键选择查看数据，即可查看依存句法分析结果与其之间的依存关系。

预览数据

_c0	期望职位	期望职位_all_relation	期望职位_relation
0	['数据挖掘工程师', '算法工程师']	[[(1, '数据', 'n', 3, '工程师', '定中关系'), (2, '挖掘', 'v', 3, '工程师', '定中关系'), (3, '工程师', 'n', 0, '##核心##', '核心关系')], [(1, '算法', 'n', 2, '工程师', '定中关系'), (2, '工程师', 'n', 0, '##核心##', '核心关系')]]	[{'主谓关系': []}, {'主谓关系': []}]
		[[(1, '数据', 'n', 2, '分析师', '定中关系'), (2, '分析师', 'n', 0, '##核心##', '核心关系')], [(1, '数据', 'n', 3, '工程师', '定中关	

查看日志。对【Hanlp依存句法】组件右键选择查看日志，可以观察到文本数据的依存分析过程。

查看日志



8.7.19 语义角色标注

（1）作用

语义角色标注(Semantic Role Labeling,简称 SRL)是一种浅层的语义分析。给定一个句子， SRL 的任务是找出句子中谓词的相应语义角色成分，包括核心语义角色（如施事者、受事者等） 和附属语义角色（如地点、时间、方式、原因等），其能够对问答系统、信息抽取和机器翻译等应用产生推动作用。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	文本特征格式	tlp分词与词性标注结果	

（3）输出

序号	名称	内容
1	data_out	语义标注结果

（4）参数

序号	分组	参数	说明
1	输出设置	输出文件类型	
2	输入设置	语义角色标注字段	
3	输出设置	目标字段LTP分词结果字段	
4	参数设置	目标字段LTP词性标注结果字段	

（5）示例

对position数据集进行语义角色标注示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行TLP分词与词性标注结果，将【TLP分词与词性标注】组件与输入源连接，在参数配置中选择需要进行分词的特征列，选择输出词性标注，右键单击【TLP分词与词性标注】组件，选择“运行该节点”。

② 进行LTP语义角色标注，将【LTP语义角色标注】与①步骤的分词与标注结果进行连接，在输入设置中选择对应的特征列，右键单击【LTP语义角色标注】组件，选择“运行该节点”。

打开数据，查看结果。对【LTP语义角色标注】组件右键选择查看数据，即可对文本的语义角色标注进行查看。

8.7.20 文本向量化TF-IDF

（1）作用

TF-IDF是Term Frequency-Inverse Document Frequency的简称。她是一种非常常见的用于将文本转化为有意义的数字表示的算法。这个技术被广泛的应用与NLP的各个方面。

TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

该组件使用TF-IDF算法，根据分词结果，训练生成向量转换器，并利用词汇tf-idf值向量化文本。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	停用词文本	txt文件	
3	tfidf_vectorizer模型	不必须	已训练好的词向量模型

（3）输出

序号	名称	内容
1	data_out	文本数据
2	tfidf_matrix_df	词向量结果

（4）参数

序号	分组	参数	说明
1	字段设置	文本特征	
2	字段设置	选取停用词列	
3	参数设置	过滤低于某一百分比的特征	所提取的特征所出现的文档占比小于该占比时，则会从特征中删除
4	参数设置	是否将英字母符统一为小写	
5	参数设置	是否过滤单字符	
6	参数设置	应用子线性tf缩放	用1 + log（tf）替换tf
7	参数设置	特征允许构成的元数（n-gram）,请输入允许的最小值和最大值，中间用英文逗号隔开	数如为元组形式 tuple （min_n, max_n），表示最后得到的特征可以由几个单部分（词/句子等）构成，例如（1,2）表示，得到的特征可以由1个或者2个连续的部分构成。
8	参数设置	最大特征数（默认保存所有过滤后特征）	在大规模语料上训练TFIDF会得到非常多的词语，如果再使用了上一个设置加入了词组，那么我们词表的大小就会爆炸。出于时间和空间效率的考虑，可以限制最多使用多少个词语，模型会优先选取词频高的词语留下。
9	参数设置	是否平滑idf权重	
10	参数设置	过滤大于某一百分比的特征	

（5）示例

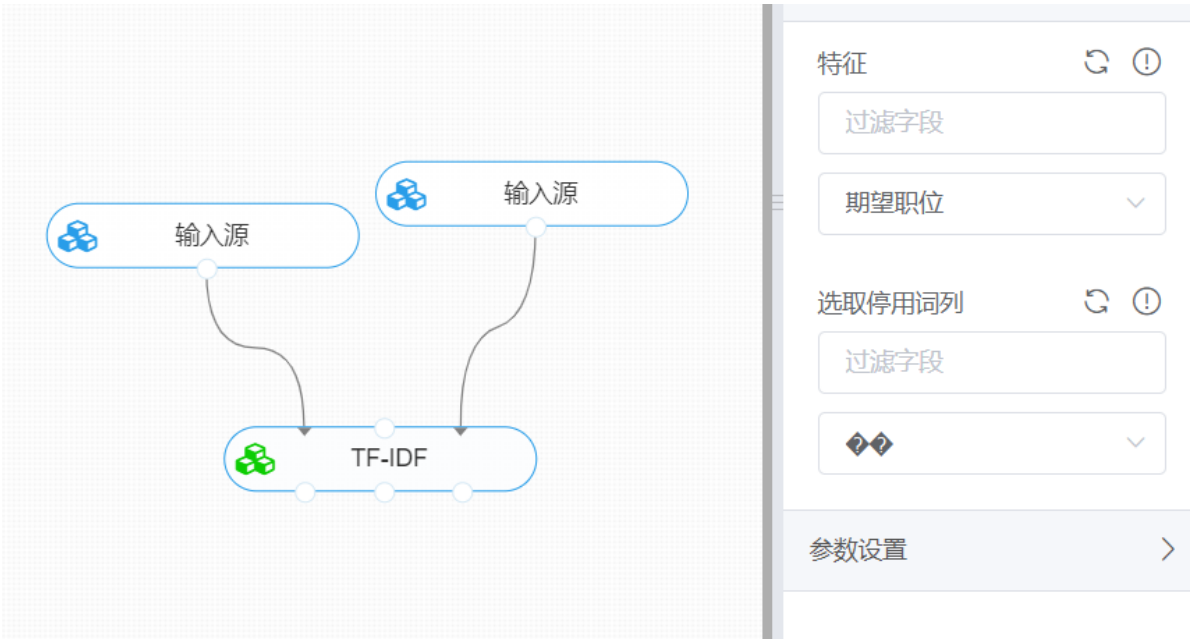
对position数据集进行词向量示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行TF-IDF，将【TF-IDF】组件与输入源连接，在字段配置中选择相对应的特征列，在参数设置中修改过滤低于某一百分比的特征为0，其他采用默认设置，同时拖入新的输入源，读入停用词数据，与【TF-IDF】组件进行连接。右键单击【TF-IDF】组件，选择“运行该节点”。



打开数据，查看结果。对【TF-IDF】组件右键选择查看数据tfidf_matrix_df，即可查看文本向量化后的结果表。

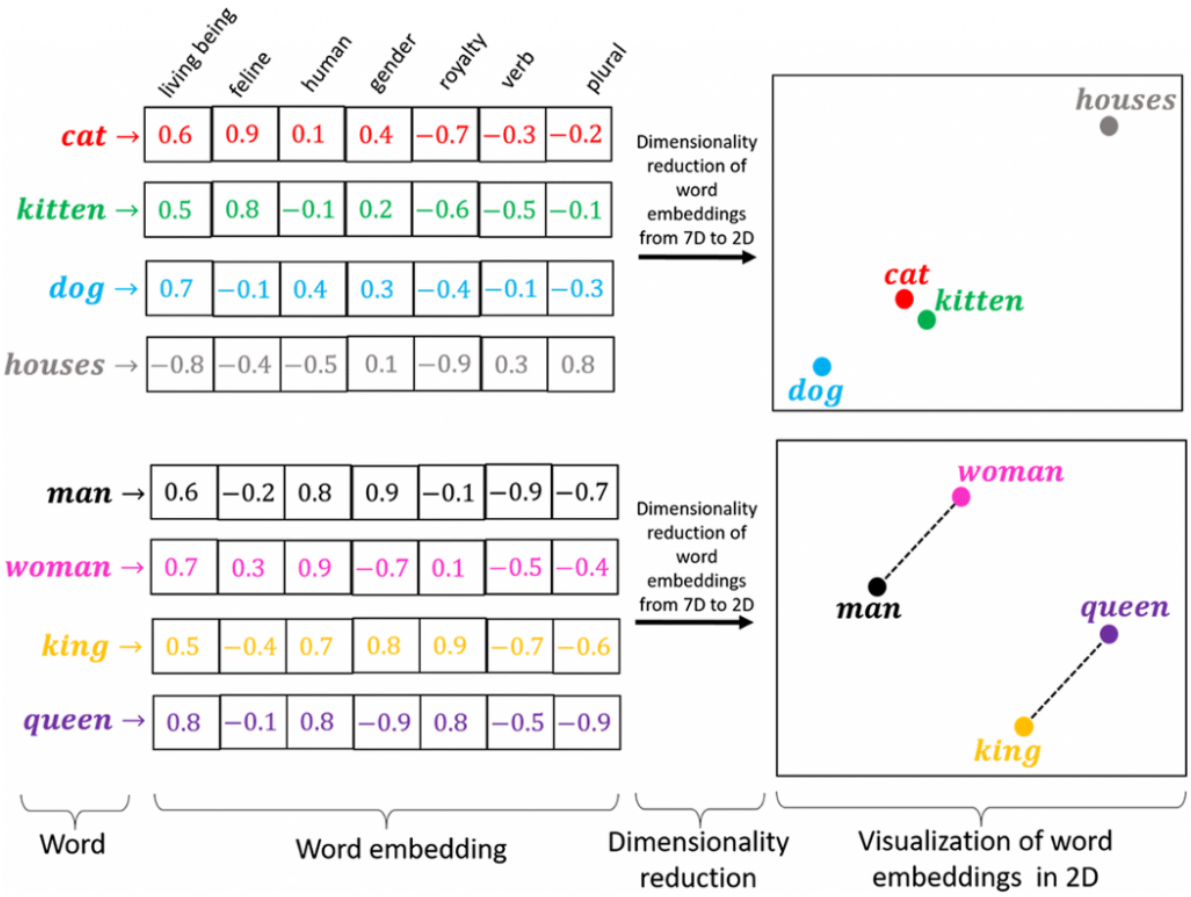
预览数据

0	1	2	3	4	5	
0.0	0.0	0.0	0.0	0.47363151106880075	0.0	0.88072
0.0	0.0	0.0	0.3345122195261065	0.5162916654978894	0.0	0.0
0.0	0.0	0.0	0.3345122195261065	0.5162916654978894	0.0	0.0
0.0	0.0	0.0	0.29336949852219046	0.45279131271460626	0.0	0.84197
0.0	0.0	0.0	0.5437566374219028	0.8392429441226332	0.0	0.0

8.7.21 文本向量化word2vec

文本向量化是自然语言处理中的基础工作，文本的表示直接影响到了整个自然语言处理的性能。Word2vec 是 Word Embedding 的方法之一，他是 2013 年由谷歌的 Mikolov 提出了一套新的词嵌入方法。其主要用于词语的文本向量化。

word2evc文本向量化的过程如下，它可以将文本通过一个低维向量来表达，语义相似的词在向量空间上的值也会比较相近，可以用于不同的任务中。

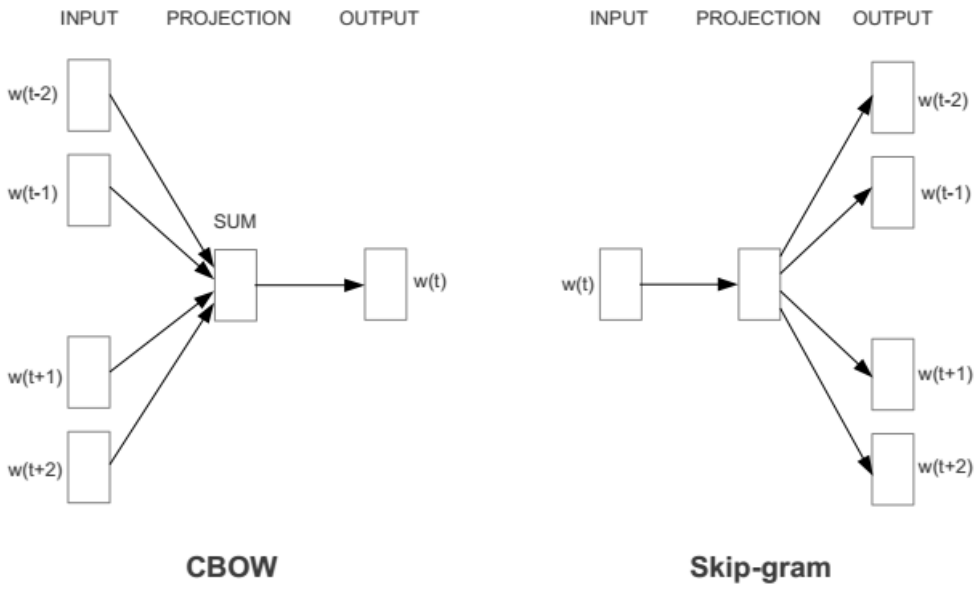


训练算法参数（框架）：Negative Sampling（负采样）、Hierarchical Softmax

- Negative Sampling（负采样）：在word2vec中我们预测的是当前单词与其他单词的一起出现的概率，每一个单词与多个单词形成组合形成了大量的分类，导致计算复杂，为了简化计算，采用负分类的是想，将公式转化，通过判断两个单词的组合是否正确将问题转换为二分类问题，简而言之，负采样就是将多分类的softmax转成二分类的sigmoid。
- Hierarchical Softmax：用霍夫曼树来代替传统神经网络的隐藏层和输出层的神经元，霍夫曼树的叶子节点起到输出层神经元的作用，叶子节点的个数即为词汇表的小大。 而内部节点则起到隐藏层神经元的作用。

训练算法（模型）：CBOW、skip_gram

- CBOW模型的训练输入是某一个特征词的上下文相关的词对应的词向量，而输出就是这特定的一个词的词向量。
- Skip-Gram模型和CBOW的思路是反着来的，即输入是特定的一个词的词向量，而输出是特定词对应的上下文词向量。



(2) 输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	model	非必填	已经训练好的word2vec模型

(3) 输出

序号	名称	内容
1	data_out	文本向量结果

(4) 参数

序号	分组	参数	说明	
1	字段设置	是否文本转向量		
2	字段设置	特征	输入分词后的序列	
3	参数设置	是否进行模型训练	若不进行模型训练，则需要输入已训练的word2vec模型	
4	参数设置	负采样个数		
5	参数设置	词向量维度	若不纠正则只进行错误检测	
6	参数设置	学习率		
7	参数设置	训练算法参数		
8	参数设置	最小词频	对于词频小于最小词频的词语，将不进入模型中	
9	参数设置	训练算法		
10	参数设置	训练并行数		

（5）示例

对position数据集进行词向量示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行word2vec，将【word2vec】组件与输入源连接，在输入设置中选择所需文本列，在参数配置中选择训练模型与模型使用的框架，右键单击【word2vec】组件，选择“运行该节点”。

打开数据，查看结果。对【word2vec】组件右键选择查看数据，即可查看文本向量化的结果。

预览数据

��望职位	word_vector
	[[-0.004352581, -0.0017379174, -0.0014628099, 0.002460897, 0.0007782102, -0.004035845, -0.0029358456, -0.0040689963, -0.0002654155, -0.004203462, -0.0027888725, 0.003392478, -0.0023723627, -0.0013862575, -0.00023911441, 0.0023281765, 0.0029007618, 0.0003639105, -0.0020346665, -0.0010699901, -0.0013835909, -0.0012424755, -0.00163

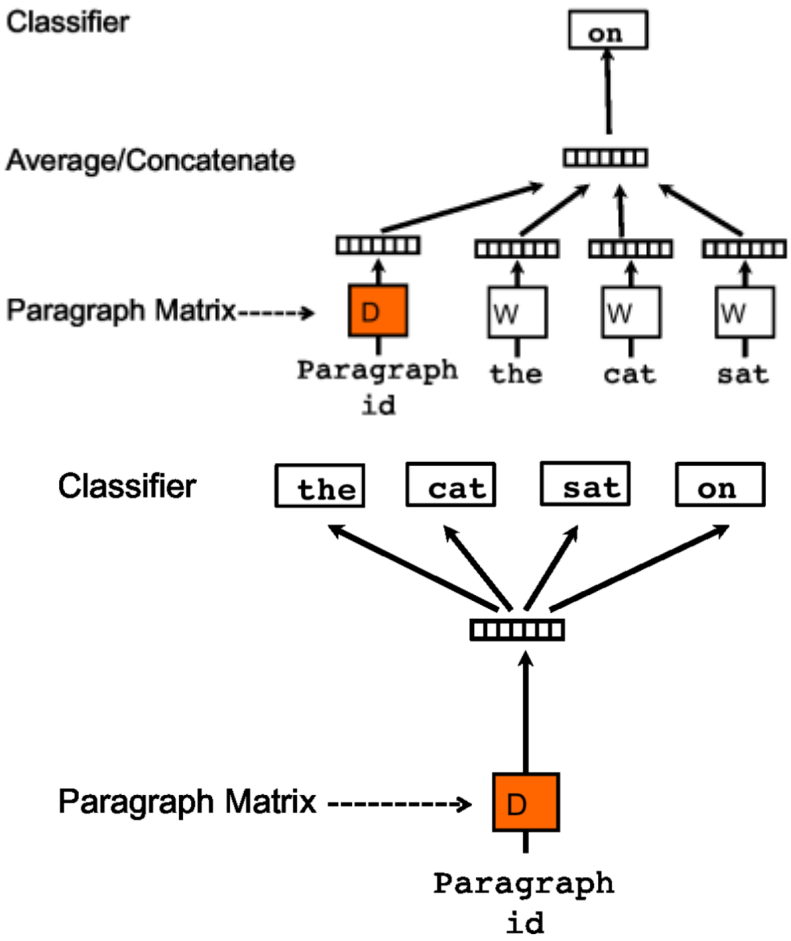
8.7.22 文本向量化doc2vec

（1）作用

Doc2vec方法是一种无监督算法，能从变长的文本（例如：句子、段落或文档）中学习得到固定长度的特征表示。Doc2vec也可以叫做 Paragraph Vector、Sentence Embeddings，它可以获得句子、段落和文档的向量表达，是Word2Vec的拓展，其具有一些优点，比如不用固定句子长度，接受不同长度的句子做训练样本。Doc2vec算法用于预测一个向量来表示不同的文档，该模型的结构潜在的克服了词袋

模型的缺点。例如对于一个句子I want to drink water，如果要去预测句子中的单词want，那么不仅可以根据其他单词生成feature， 也可以根据其他单词和句子来生成feature进行预测。

doc2vec的框架如下：



每个段落/句子都被映射到向量空间中，可以用矩阵的一列来表示。每个单词同样被映射到向量空间，可以用矩阵的一列来表示。然后将段落向量和词向量级联或者求平均得到特征，预测句子中的下一个单词。

训练方式：

- 分布记忆的段落向量（Distributed Memory Model of Paragraph Vectors , PV-DM）：这个段落向量/句向量也可以认为是一个单词，它的作用相当于是上下文的记忆单元或者是这个段落的主题，所以我们一般叫这种训练方法为Distributed Memory Model of Paragraph Vectors(PV-DM)，类似于Word2Vec中的CBOW模型。
- 分布词袋版本的段落向量（Distributed Bag of Words version of Paragraph Vector，PV-DBOW）：另一种训练方法是忽略输入的上下文，让模型去预测段落中的随机一个单词。在每次迭代的时候，从文本中采样得到一个窗口，再从这个窗口中随机采样一个单词作为预测任务，让模型去预测，输入就是段落向量，这种方法为 Distributed Bag of Words version of Paragraph Vector(PV-DBOW)，类似于Word2Vec中的Skip-gram模型。
- PV-DM+PV-DBOW，两种方法相结合。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	model	非必填	已经训练好的doc2vec模型

(3) 输出

序号	名称	内容
1	data_out	文本向量结果

(4) 参数

序号	分组	参数	说明	
1	字段设置	是否文本转向量		
2	字段设置	特征	输入分词后的序列	
3	字段设置	是否进行模型训练	若不进行模型训练或进行增量训练，请输入已训练的doc2vec模型	
4	参数设置	训练算法	PV-DM、PV-DBOW、PV-DM+PV_DBOW	
5	参数设置	负采样个数		
6	参数设置	迭代次数		
7	参数设置	是否将上下文向量与文档向量拼接	PV-DM	
8	参数设置	窗口大小		
9	参数设置	最小词频		
10	参数设置	是否训练词向量	若为Word_vector（词向量）时，则在训练doc_vector（DBOW）的同时训练Word_vector（Skip-gram）； 若只训练doc_vector（文档向量）时，速度更快	

序号	分组	参数	说明	
11	参数设置	特征向量维度		

（5）示例

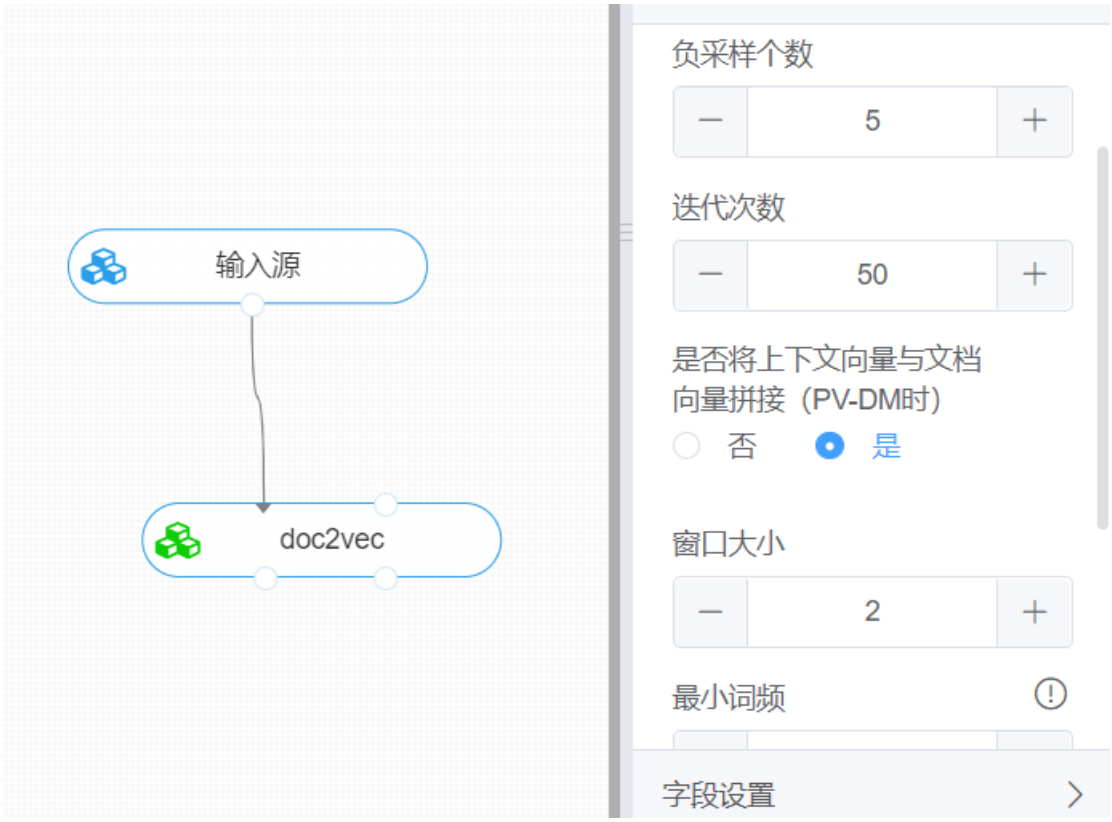
对position数据集进行词向量示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行doc2vec，将【doc2vec】组件与输入源连接，在输入设置中选择所需文本列，在参数配置中训练算法选择DM+DBOW相结合模型，右键单击【doc2vec】组件，选择“运行该节点”。



打开结果，查看数据。对【doc2vec】组件右键选择查看数据，即可查看文本词向量结果。

预览数据

??望职位	doc_vector
	[-0.0028796138, 0.0036719008, 0.005750234, 0.0019357643, 0.0006687495, 0.0047474415, 0.0044825184, -0.0051604533, -0.004471147, -0.0034540668, -0.0019660147, -0.0015346926, 0.0036748294, 0.0010550225, 0.0032867303, 0.003429727, 0.00104511, -0.0034709612, -0.0031946613, -0.0015819347, -0.0026527818, -0.0036444566, -0.001573181

8.7.23 文本独热编码

（1）作用

one-hot编码被称为独热码，在英文文献中称做 one-hot code, 直观来说就是有多少个状态就有多少比特，而且只有一个比特为1，其他全为0的一种码制，简单的来说就是用0和1的编码方式来表示需要处理的一些信息，以达到该信息向量化的一种手段。

缺点是：

会造成维数过高，随着语料的增加，维数越来越大，导致维数灾难；

没有考虑到单词的顺序，忽略了词的语义信息。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	文本特征要求	分词特征列	

（3）输出

序号	名称	内容
1	data_out	独热编码结果

（4）参数

序号	分组	参数	说明	
1	字段设置	分词后的特征列		
2	字段设置	特征		

（5）示例

对position数据集进行文本独热编码示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行分词，将【结巴分词】组件与输入源连接，在输入设置中选择所需文本列，在参数配置中默认采用精确模式进行分词，其中模型采用HMM，右键单击【结巴分词】组件，选择“运行该节点”，得到分词后的数据列，用于下面的独热编码。

② 进行独热编码，将【独热编码】组件与上面的分词组件输出点进行连接，在输入设置中分词特征列选择“期望职位_cut_word”，特征中勾选所需文本列，右键单击【独热编码】组件，选择“运行该节点”。

打开数据，查看结果。对【文本独热编码】组件右键选择查看数据，即可查看文本转化后的独热码。

预览数据

??望职位	期望职位_cut_words	onehot_vec
["数据挖掘工程师","算法工程师"]	['数据挖掘', '工程师', '算法', '工程师']	[[0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0] [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0 0 0 1 0 0 0] [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0]]
["数据分析师","数据挖掘工程师","自然语言处理工程师"]	['数据', '分析师', '数据挖掘', '工程师', '自然语言', '处理', '工程师']	[[0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0] [0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0] [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0] [0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0] [0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0] [0 0

8.7.24 TF-IDF信息提取

（1）作用

TF-IDF（Term Frequency-InversDocument Frequency）是一种常用于信息处理和数据挖掘的加权技术。该技术采用一种统计方法，根据字词的在文本中出现的次数和在整个语料中出现的文档频率来计算一个字词在整个语料中的重要程度。它的优点是能过滤掉一些常见的却无关紧要本的词语，同时保留影响整个文本的重要字词。这种方式能有效避免常用词对文本关键词的影响，提高了关键词与文章之间的相关性。

$$TF-IDF=词频（TF）\times 逆文档频率（IDF）$$

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	model	非必填	已经训练好的doc2vec模型

（3）输出

序号	名称	内容
1	data_out	文本向量结果

（4）参数

序号	分组	参数	说明
1	字段设置	选择目标列	
2	字段设置	特征	
3	参数设置	关键词个数	输出权重值较大的词个数
4	参数设置	输出词性	
5	参数设置	输出权重	TF-IDF值
6	参数设置	目标词性	提取的该词性的词语，下面是词性编码对照表

词性编码	词性名称
Ag	形语素
a	形容词
ad	副形词
an	名形词
b	区别词
c	连词
dg	副语素
d	副词
e	叹词
f	方位词
g	语素
h	前接成分
i	成语
j	简称略语
k	后接成分
l	习用语
m	数词
Ng	名语素
n	名词
nr	人名
ns	地名
nt	机构团体
nz	其他专名
o	拟声词
p	介词
q	量词
r	代词
s	处所词
tg	时语素
t	时间词

词性编码	词性名称
u	助词
vg	动语素
v	动词
vd	副动词
vn	名动词
w	标点符号
x	非语素字
y	语气词
z	状态词
un	未知词

（5）示例

对position数据集进行TF-IDF信息提取示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行TF-IDF信息提取，将【TF-IDF信息提取】组件与输入源连接，在输入设置中选择目标列，在参数配置中填入需要提取的关键词词性与个数，选择输出提取的词性与其权重值，右键单击【TF-IDF信息提取】组件，选择“运行该节点”。

打开数据，查看结果。对【TF-IDF信息提取】组件右键选择查看数据，即可查看文本信息提取结果。

查看日志

TF-IDF信息提取

	期望职位	关键词	词性	权重
0	["数据挖掘工程师","算法工程师"]	工程师, 数据挖掘, 算法	n, n, n	3.9, 3.3, 2.17
1	["数据分析师","数据挖掘工程师","自然语言处理工程师"]	工程师, 数据挖掘, 分析师, 处理, 数据	n, n, n, v, n	2.6, 2.2, 0.96, 0.9, 0.8
2	["数据分析师","自然语言处理工程师","数据挖掘工程师"]	工程师, 数据挖掘, 分析师, 处理, 数据	n, n, n, v, n	2.6, 2.2, 0.96, 0.9, 0.8

8.7.25 TextRank短语

（1）作用

TextRank 算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 PageRank算法, 通过把文本分割成若干组成单元(单词、句子)并建立图模型, 利用投票机制对文本中的重要成分进行排序, 仅利用单篇文档本身的信息即可实现关键词提取、文摘。由于在短文本中词频低、词间共现关系相近即词组重要性区别不大，因此算法更适用于长文本。

该组件主要用于提取关键词的提取。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	stop_words	自定义停用词词典	每行一个停用词，必须为UTF-8编码
3	文本特征格式	字符串序列、字符串列表	

（3）输出

序号	名称	内容
1	data_out	文本中短语提取结果（若无关键词，则以空列表输出）

（4）参数

序号	分组	参数	说明
1	输入设置	特征	
2	输入设置	文本编码	如utf-8
3	参数设置	在前几个关键词中构造短语	获取几个关键词构造可能出现的短语，默认取前10个关键词
4	参数设置	窗口大小	窗口大小，用来构造单词之间的边，默认值为2
5	参数设置	图模型中节点的构造方式	可选项有：['直接分词','分词后过滤停用词','分词后按停用词和词性过滤']
6	参数设置	图模型中边的构造方式	可选项有：['直接分词','分词后过滤停用词','分词后按停用词和词性过滤']
7	参数设置	是否将文本转换成小写	
8	参数设置	分句分隔符	非必填，默认值是`?!;？！。；…\n`，用来将文本拆分
9	参数设置	短语在文本出现的最小次数	短语在原文本中至少出现的次数，默认1次

（5）示例

对position数据集进行TextRank示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行TextRank短语，将【TextRank短语】组件与输入源连接，同时拖入新的输入源。用于传入停用词文本，与【TextRank短语】组件的stop连接点进行连接，在输入设置中选择目标列，在参数配置中选择相应的参数，右键单击【TextRank短语】组件，选择“运行该节点”。

打开数据，查看结果。对【TextRank】组件右键选择查看数据，即可查看提取的关键词短语信息。

预览数据

_c0	期望职位	期望职位_phrases
16	师', '机器学习工程师']	[[]]
17	['数据分析师']	[['数据分析师']]
18	['数据分析师', '数据挖掘工程师']	[[]]
19	['数据分析师']	[['数据分析师']]
20	['数据分析师', '其他']	[[]]
21	['Hadoop大数据开发工程师']	[[]]

8.7.26 LDA-主题分类

（1）作用

LDA主题模型（Topic Model）是以非监督学习的方式对文集的隐含语义结构进行聚类的统计模型。主题模型主要被用于自然语言处理中的语义分析和文本挖掘问题，例如按主题对文本进行收集、分类和降维。该LDA组件就是在给定的分类数下，可以把数据集进行分类，并且给出每个类别所包含的主要的特征。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	特征格式	分词后的特征列	

（3）输出

序号	名称	内容
1	data_out	文本中短语提取结果（若无关键词，则以空列表输出）

（4）参数

序号	分组	参数	说明
1	输入设置	特征	
2	参数设置	语料转化为向量集的方式	可选项有：词频、TF-IDF
3	参数设置	随机种子	
4	参数设置	输出各主题最有可能的主题词个数	窗口大小，用来构造单词之间的边，默认值为2
5	参数设置	主题数	即文本分类数
6	参数设置	最大迭代次数	

（5）示例

对position数据集进行LDA主题分类示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行分词，将【结巴分词】组件与输入源连接，在输入设置中选择目标特征列，在参数配置中选择相应的参数，右键单击【结巴分词】组件，选择“运行该节点”，得到分词后的文本数据。

② 进行LDA主题分类，将【LDA】组件与①的组件进行连接，在输入设置中选择“期望职位_cut_words”，在参数设置中对于语料转化方式选择词频（因为此处提供的测试文本属于短语类型，对其进行词频统计即可），主题数目默认为3，右键单击【LDA】组件，选择“运行该节点”。

打开数据，查看结果。对【LDA】右键选择查看数据，其中data_out为主题分类结果，topic_df为文本的主题词频分布。

预览数据

��望职位	期望职位_cut_words	label
["数据分析师","图像处理工程师","机器学习工程师"]	['数据', '分析师', '图像处理', '工程师', '机器', '学习', '工程师']	0
["Hadoop大数据开发工程师"]	['Hadoop', '大', '数据', '开发', '工程师']	2
["Hadoop大数据开发工程师"]	['Hadoop', '大', '数据', '开发', '工程师']	2
["Hadoop大数据开发工程师"]	['Hadoop', '大', '数据', '开发', '工程师']	2
["数据分析师","机器学习工程	['数据', '分析师', '机器', '学习',	0

预览数据

��题1	主题2	主题3	主题1_词概率	主题2_词概率	主题3_词概率
Hadoop	数据挖掘	其他	0.158	0.186	0.216
大	开发	分析师	0.155	0.115	0.199
开发	大	机器	0.150	0.111	0.086
数据挖掘	Hadoop	学习	0.103	0.104	0.085

8.7.27 左右信息熵-短语提取

（1）作用

信息熵(entropy)指的是某条消息所含的信息量。它反映的是听说某个消息之后，关于该事件的不确定性的减少量。左右信息熵算法是判断文本数据中的每个词是否可以选为标签，也即左右信息熵算法适用于提取语义信息较短的标签，并不能实现词与词进行组合得到具有一定语义信息的较长的标签的提取。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out	文本中短语提取结果

（4）参数

序号	分组	参数	说明
1	输入设置	目标列	
2	输入设置	特征	输入单层列表序列或字符序列
3	参数设置	提取的短语个数	

（5）示例

对position数据集进行左右信息熵短语提取示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行左右信息熵短语提取，将【左右信息熵】组件与输入源连接，在输入设置中选择目标列，在参数配置中填写提取的短语个数，右键单击【左右信息熵】组件，选择“运行该节点”。

打开数据，查看结果。对【左右信息熵】右键选择查看数据，即可查看短语信息提取的文本结果。

左右信息熵-信息提取

	期望职位	期望职位_phrases
0	[数据挖掘工程师, 算法工程师]	[[数据挖掘工程师], [算法工程师]]
1	[数据分析师, 数据挖掘工程师, 自然语言处理工程师]	[[数据分析师], [数据挖掘工程师], [自然语言处理工程师]]
2	[数据分析师, 自然语言处理工程师, 数据挖掘工程师]	[[数据分析师], [自然语言处理工程师], [数据挖掘工程师]]

8.7.28 文本分类-FastText

（1）作用

FastText是facebook开源的一个词向量与文本分类工具，在2016年开源，典型应用场景是"带监督的文本分类问题"。提供简单而高效的文本分类和表征学习的方法，性能比肩深度学习而且速度更快。

fastText结合了自然语言处理和机器学习中最成功的理念。包括：使用词袋以及n-gram袋表征语句，使用子字(subword)信息，并通过隐藏表征在类别间共享信息。它另外采用了一个softmax层级(利用了类别不均衡分布的优势)来加速运算过程。

fasttext两个主要任务为：

- 有效文本分类 ： 有监督学习
- 学习词向量表征：无监督学习

本组件基于fastText 的有效分类功能来实现文本的快速划分。fastText 的模型架构和 word2vec 中的 CBOW 模型的结构很相似。CBOW 模型是利用上下文来预测中间词，而fastText 是利用上下文来预测文本的类别。而且从本质上来说，word2vec是属于无监督学习，fastText 是有监督学习。但两者都是三层的网络（输入层、单层隐藏层、输出层），具体的模型结构如下：

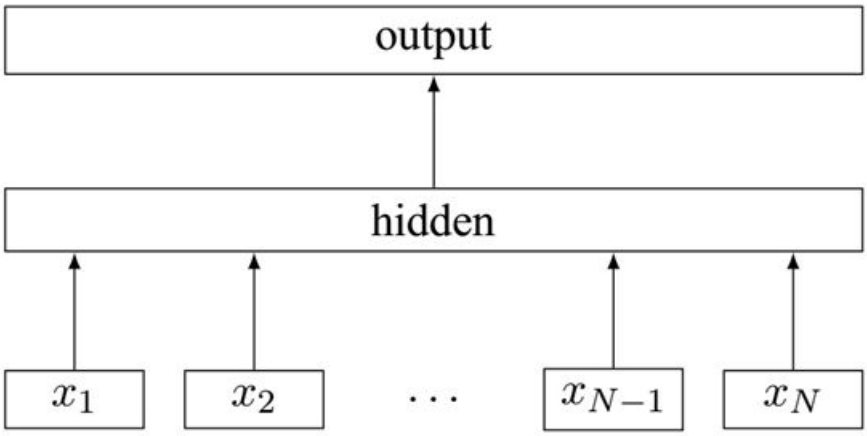


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	特征列格式	输入必须为文本分词结果，列表转换成的字符或者列表，如“["数据","工程师"]”	

（3）输出

序号	名称	内容
1	data_out	训练数据基于其自身训练生成fastText 网络得到到的预测结果
2	model	训练生成的fasttext网络模型
3	out_put_test	FastText模型对测试集的预测结果，保存在原始输入测试集中共同输出，当且仅当输入测试集时生效；

（4）参数

序号	分组	参数	说明
1	模型评估参数	是否引入测试数据	
2	输入设置	训练集输入特征	已分词的文本数据列，列表或者列表类型的字符串，只允许选择单列
3	输入设置	训练集目标输出	目标分类列，离散型数值或文本。
4	输入设置	测试集输入特征	若不进行模型效果测试，可忽略该参数
5	输入设置	测试集目标输出	若不进行模型效果测试，可忽略该参数
6	模型训练参数	学习率	默认为0.1，取值范围为[0,1]，用于决定模型训练过程中目标函数能否收敛到局部最小值以及何时收敛到最小值。
7	模型训练参数	内容窗口大小	表示上下文窗口的大小（当前词与预测词在一个句子中的最大距离是多少）
8	模型训练参数	字符ngram的最小长度	非必填，以单个字为粒度的ngram的最小长度，默认为0
9	模型训练	类别最小样本	样本数小于该取值的类别将被过滤
10	模型训练	字符ngram的最大长度	非必填，以单个字为粒度的ngram的最大长度，默认为0
11	模型训练	迭代次数	
12	模型训练	词汇ngram的最大长度	认是使用 1-gram，也就是单独的词。
13	模型训练	损失函数	用来计算测试集中目标值Y的真实值和预测值的偏差程度
14	模型训练	词向量维度	为避免训练时间过长，此处将词向量维度限制在前1000；引入外部模型，则该参数从外部模型中获取
15	模型训练	最小词频数	小于该词频的词汇将被过滤

（5）示例

对留言文本数据集进行FastText分类示例。

	A	B	C	D	E	F
1	id	user_id	them	time	detail	first_class
2	102738	A0008529	K2区映山	2014/3/24	我家住在K市K2区华源府第小区，	环境保护
3	127391	A0008525	L5县环保	2016/12/2	县环保局噪声测试不按国家规定	环境保护
4	144491	A0003479	请责令M	2012/9/3 7	U优会所早晨排污持续，请责令	环境保护
5	118104	A0005202	K11县县	2018/9/25	我投资200余万元的纸厂被县环保局	环境保护
6	140709	A0006959	M1区环保	2017/7/2 6	您好：日兴砖厂是一家打着环保	环境保护
7	174438	A0001663	为何日月	2012/12/1	邓书记： 您好！日月星城KTV营业	环境保护
8	94729	A0009906	J9县沙田	2014/3/18	尊敬的黄县长： 您好！请您帮帮	环境保护
9	121512	A0007873	L市小型砖	2017/11/1	尊敬的彭书记； 你好！我是	环境保护
10	127245	A0006728	关于L5县	2017/8/27	尊敬的蒙书记 你好！ 百	环境保护
11	170620	A0008893	G8县薮城	2019/11/2	您好！我是G8县薮城石膏实业有限公	环境保护
12	161600	A0001571	西地省盛	2014/8/14	在临G5县合口镇有个大的盛常玻	环境保护
13	181619	A0001904	I3县南洲	2018/6/6 1	I3县南洲镇鑫顺广场A栋一楼开了一家	环境保护
14	182424	A0005128	I市山水华	2016/10/8	我们是I市山水华庭的住户，我们	环境保护
15	10689	A0006174	关于取消	2016/1/8 1	尊敬的领导： 您好！ 我们是	环境保护
16	161993	A0005596	A市和顺	2015/11/9	1．国家《电磁辐射管理办法》规定	环境保护
17	138263	A0007788	M4市数千	2017/4/9 1	河流守望者；接到来自西地省M4	环境保护
18	130815	A0006794	泸阳镇下	2016/12/2	泸阳镇下坪村与壮稻村，采石场	环境保护
19	138105	A0005961	M4市中连	2017/7/14	尊敬的李书记 您好！ 在	环境保护
20	155617	A0009205	L6县巫水	2018/4/15	清明假期，来到阔别多年的西地省，	环境保护
21	117719	A0008602	K市佑康糖	2015/11/9	K市佑康糖尿病专科医院自建院	环境保护
22	139889	A0006086	M2县私人	2018/12/3	呼吁有关部门坚决取缔关停街埠头村	环境保护
23	155419	A0007694	J10县永	2018/8/20	大源水库是J10县全县人民饮水取水	环境保护
24	16857	A0004390	第三次请	2017/12/2	你好，本人于9-11月，有多次关	环境保护
25	30073	A0007499	A6区丁字	2016/8/25	尊敬的孔书记 A6区丁字国土所所	环境保护
26	137103	A0004155	请对M2县	2019/2/23	请求政府有关部门责令相关企业采取	环境保护
27	161596	A0006557	请求领导	2014/8/28	尊敬的领导； 我们是M5市三甲乡	环境保护
28	22293	A0001605	关于岳临	2015/1/21	尊敬的周县长，您好！ 我是A8县	环境保护
29	22687	A0001902	A4区捞刀	2015/9/3 9	今年上半年，我写过关于了篇贴	环境保护
30	142722	A0009365	M3县移动	2017/8/24	M3县移动城西贸易区基站由于道	环境保护
31	119997	A0001117	L市凯通领	2019/9/17	尊敬的领导：凯通领御小区位于湖天	环境保护
32	126447	A0007990	L5县沙溪	2019/6/27	自沙溪炼油厂开厂以来，我村多次举	环境保护
33	126462	A0001081	L5县沙溪	2019/6/15	L5县纪委、监委：自从沙溪村废旧轮	环境保护
34	124323	A0003652	污染触目	2018/4/9 1	各位网友，你们好！我们是L12市的居	环境保护

首先将liuyan数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“liuyan”，勾选文件“留言.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行分词，对留言详情文本进行分词，将【结巴分词】组件与输入源连接，在输入设置中勾选detail特征列进行分词操作，右键单击【结巴分词】算法，选择“运行该节点”。

② 进行fasttext，将【FastText】组件与【结巴分词】组件连接，在输入设置中勾选特征列与文本分类列，在参数配置中设置相对应的参数，右键单击【FastText】组件，选择“运行该节点”。

打开数据，查看结果。对【FastText】右键选择查看数据，即可查看模型训练分类结果。

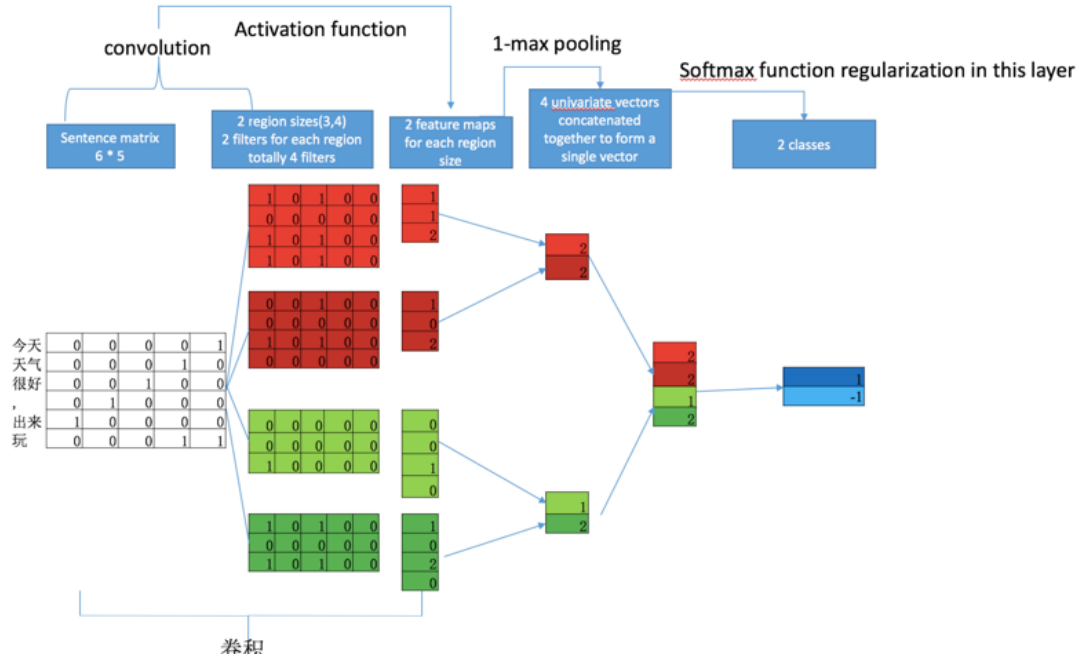
预览数据

	them	time	detail	first_class	detail_cut_words	pre_label
	K2区映山岭通宵作业，华源府第小区居民得不到正常休息	2014/3/24 1:53	站，答复已现场查看，问题确实存在，但处理不好；再次拨打12345，与环境监测站进行三方通话，环境监测站答应需要请示领导后第二天给我答复（难道工作人员处理不了的问题还没有汇报领导？）。18日，拨打环境监测站，接电话的工作人员称不清楚这件事情，我问他昨天值班的人是不是姓张，他说是的，于是我将自己电话告诉值班人员，让张姓工作	环境保护	', ', '环境', '监测站', '答应', '需要', '请示', '领导', '后', '第二天', '给', '我', '答复', ' (', '难道', '工作人员', '处理', '不了', '的', '问题', '还', '没有', '汇报', '领导', '？', '）', '，', '。', '18', '日', ', ', '拨打', '环境', '监测站', ', ', '接电话', '的', '工作人员', '称', '不', '清楚', '这件', '事情', ', ', '我', '问', '他', '昨天', '值班', '的', '人', '是不是', '姓张', ', ', '他', '说', '是', '的', ', ', '于是', '我', '将', '自己', '电话',	环境保护

8.7.29 文本分类-TextCNN

（1）作用

TextCNN是一种基于卷积神经网络的文本分类模型，其利用多个不同的卷积核来提取句子中的关键信息（类似于多窗口大小的n-gram），从而能够更好地捕捉局部相关性。基本流程：先将文本分词做embeeding得到词向量，将词向量经过一层卷积，一层max-pooling。最后将输出外接softmax 来做n分类。



上图为TextCnn的网络结构图，流程可以概括为：先将文本分词做embeeding得到词向量, 将词向量经过一层卷积,一层max-pooling, 最后将输出外接softmax 来做n分类

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	特征列格式	输入必须为文本分词结果，列表转换成的字符或者列表，如“["数据","工程师"]”	
3	model	不必须，传入word2vec或者doc2vec模型用于网络中的词嵌入层	

（3）输出

序号	名称	内容
1	data_out	训练数据基于其自身训练生成Textcnn网络得到到的预测结果
2	model	训练生成的网络模型
3	data_in_test	Textcnn模型对测试集的预测结果，保存在原始输入测试集中共同输出，当且仅当输入测试集时生效；

（4）参数

序号	分组	参数	说明
1	全连接层参数	激活函数	
2	全连接层参数	正则化参数	防止过拟合，0到1之间数值
3	模型训练参数	损失函数	用来计算测试集中目标值Y的真实值和预测值的偏差程度，默认为“categorical_crossentropy”
4	模型训练参数	批量次数	表示一个批次的样本数量
5	模型训练参数	优化器	默认为“adam”，用于通过训练优化参数，来最小化（最大化）损失函数
6	模型训练参数	训练次数	
7	模型训练参数	验证集比例	训练集中用于做验证的数据的比例
8	模型训练参数	特征数（词汇个数）	训练集中样本包含的词汇个数，若不设置特征数，将默认为训练集的最大分词数
9	模型训练参数	评估标准	
10	词嵌入层参数	是否引入外部词向量模型	若选择是,则需输入word2vec或者doc2vec模型
11	词嵌入层参数	词向量维度	为避免训练时间过长，此处将词向量维度限制在前1000；引入外部模型，则该参数从外部模型中获取
12	卷积层参数	卷积窗口的长度	允许输入多个数值，数值与数值间采用英文逗号分隔；数值个数代表并列卷积层的层数，默认采用3个卷积层
13	卷积层参数	填充方式	其中选择“same”表示填充输入以使输出具有与原始输入相同的长度，选择“valid”表示不填充

序号	分组	参数	说明
14	卷积层参数	卷积中滤波器的输出数量	
15	卷积层参数	激活函数	
16	卷积层参数	卷积步长	在文本处理中，该参数常设置为1
17	输入设置	训练集输入特征	已分词的文本数据列，列表或者列表类型的字符串，只允许选择单列
18	输入设置	训练集目标输出	目标分类列，离散型数值或文本。
19	输入设置	测试集输入特征	若不进行模型效果测试，可忽略该参数
20	输入设置	测试机目标输出	若不进行模型效果测试，可忽略该参数
21	池化层参数	填充方式	其中选择“same”表示填充输入以使输出具有与原始输入相同的长度，选择“valid”表示不填充
22	池化层参数	池化步长	作为缩小比例的因数。 例如，2 会使得输入张量缩小一半。 如果不填写，那么取值上一层卷积层输出的维度
23	模型测试参数	是否引入测试数据	

（5）示例

对留言文本数据集进行TextCnn文本分类示例。

	A	B	C	D	E	F
1	id	user_id	them	time	detail	first_class
2	102738	A0008529	K2区映山	2014/3/24	我家住在K市K2区华源府第小区，	环境保护
3	127391	A0008525	L5县环保	2016/12/2	县环保局噪声测试不按国家规定	环境保护
4	144491	A0003479	请责令M	2012/9/3 7	U优会所早晨排污持续，请责令	环境保护
5	118104	A0005202	K11县县	2018/9/25	我投资200余万元的纸厂被县环保局	环境保护
6	140709	A0006959	M1区环保	2017/7/2 6	您好：日兴砖厂是一家打着环保	环境保护
7	174438	A0001663	为何日月	2012/12/1	邓书记： 您好！日月星城KTV营业	环境保护
8	94729	A0009906	J9县沙田	2014/3/18	尊敬的黄县长： 您好！请您帮帮	环境保护
9	121512	A0007873	L市小型砖	2017/11/1	尊敬的彭书记； 你好！我是	环境保护
10	127245	A0006728	关于L5县	2017/8/27	尊敬的蒙书记 你好！ 百	环境保护
11	170620	A0008893	G8县薮城	2019/11/2	您好！我是G8县薮城石膏实业有限公	环境保护
12	161600	A0001571	西地省盛	2014/8/14	在临G5县合口镇有个大的盛常玻	环境保护
13	181619	A0001904	I3县南洲	2018/6/6 1	I3县南洲镇鑫顺广场A栋一楼开了一家	环境保护
14	182424	A0005128	I市山水华	2016/10/8	我们是I市山水华庭的住户，我们	环境保护
15	10689	A0006174	关于取消	2016/1/8 1	尊敬的领导： 您好！ 我们是	环境保护
16	161993	A0005596	A市和顺	2015/11/9	1．国家《电磁辐射管理办法》规定	环境保护
17	138263	A0007788	M4市数千	2017/4/9 1	河流守望者；接到来自西地省M4	环境保护
18	130815	A0006794	泸阳镇下	2016/12/2	泸阳镇下坪村与壮稻村，采石场	环境保护
19	138105	A0005961	M4市中连	2017/7/14	尊敬的李书记 您好！ 在	环境保护
20	155617	A0009205	L6县巫水	2018/4/15	清明假期，来到阔别多年的西地省，	环境保护
21	117719	A0008602	K市佑康糖	2015/11/9	K市佑康糖尿病专科医院自建院	环境保护
22	139889	A0006086	M2县私人	2018/12/3	呼吁有关部门坚决取缔关停街埠头村	环境保护
23	155419	A0007694	J10县永	2018/8/20	大源水库是J10县全县人民饮水取水	环境保护
24	16857	A0004390	第三次请	2017/12/2	你好，本人于9-11月，有多次关	环境保护
25	30073	A0007499	A6区丁字	2016/8/25	尊敬的孔书记 A6区丁字国土所所	环境保护
26	137103	A0004155	请对M2县	2019/2/23	请求政府有关部门责令相关企业采取	环境保护
27	161596	A0006557	请求领导	2014/8/28	尊敬的领导； 我们是M5市三甲乡	环境保护
28	22293	A0001605	关于岳临	2015/1/21	尊敬的周县长，您好！ 我是A8县	环境保护
29	22687	A0001902	A4区捞刀	2015/9/3 9	今年上半年，我写过关于了篇贴	环境保护
30	142722	A0009365	M3县移动	2017/8/24	M3县移动城西贸易区基站由于道	环境保护
31	119997	A0001117	L市凯通领	2019/9/17	尊敬的领导：凯通领御小区位于湖天	环境保护
32	126447	A0007990	L5县沙溪	2019/6/27	自沙溪炼油厂开厂以来，我村多次举	环境保护
33	126462	A0001081	L5县沙溪	2019/6/15	L5县纪委、监委：自从沙溪村废旧轮	环境保护
34	124323	A0003652	污染触目	2018/4/9 1	各位网友，你们好！我们是L12市的居	环境保护

首先将liuyan数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“liuyan”，勾选文件“留言.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行分词，对留言详情文本进行分词，将【结巴分词】组件与输入源连接，在输入设置中勾选detail特征列进行分词操作，右键单击【结巴分词】算法，选择“运行该节点”。

② 进行TextCnn，将【TextCnn】组件与【结巴分词】组件连接，在输入设置中勾选特征列与文本分类列，在参数配置中设置相对应的参数，右键单击【TextCnn】组件，选择“运行该节点”。

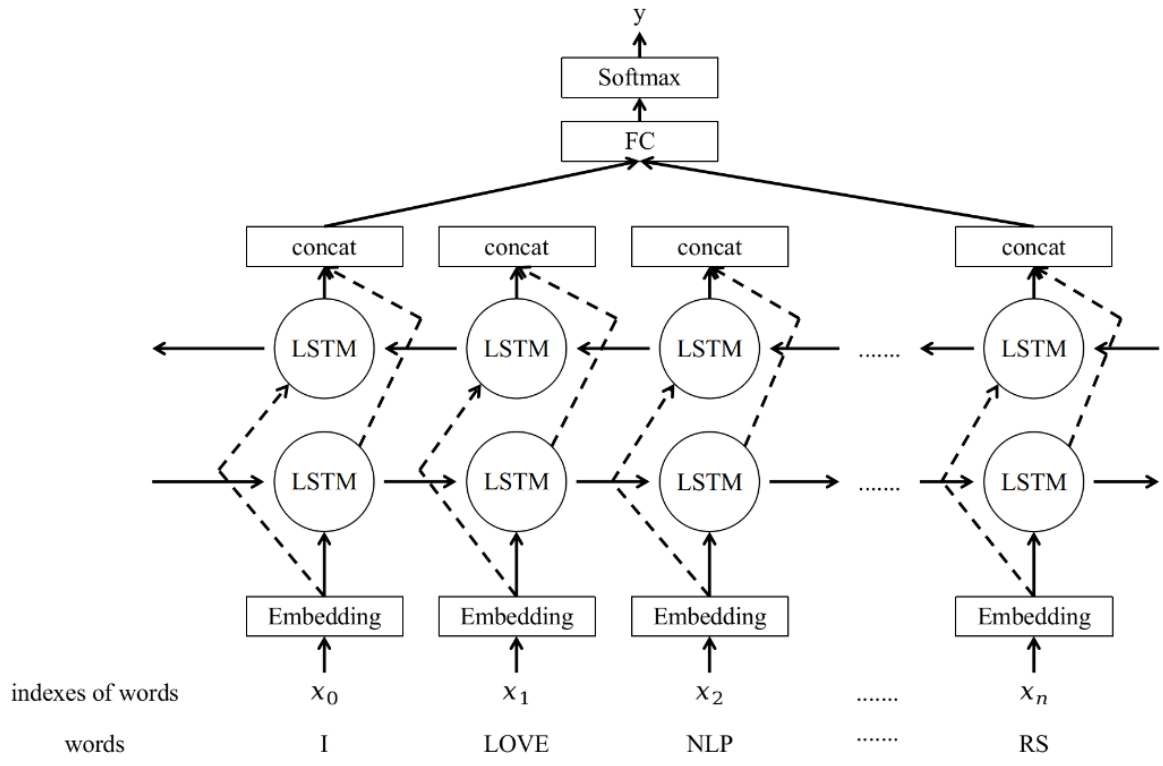
打开数据，查看结果。对【TextCnn】右键选择查看数据，即可查看模型训练分类结果。

8.7.30 文本分类-TextRNN

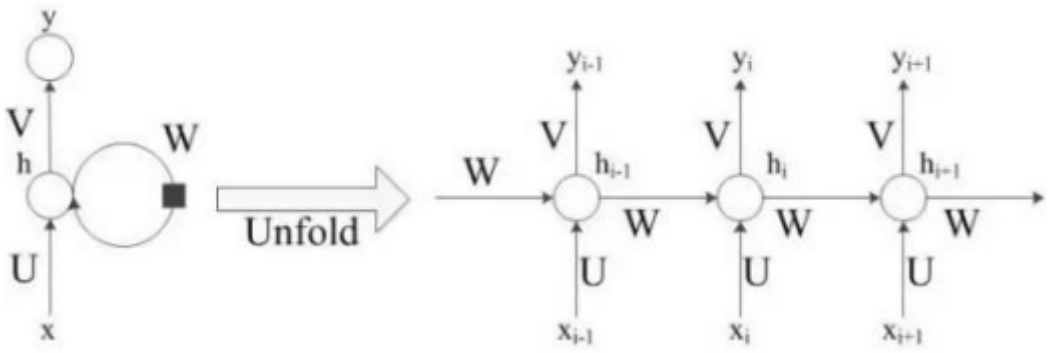
（1）作用

TextRNN指的是利用RNN循环神经网络解决文本分类问题，文本分类是自然语言处理的一个基本任务，试图推断出给定文本(句子、文档等)的标签或标签集合。比如情感分析、新闻主题分类、虚假新闻检测等。文本分类任务中，CNN可以用来提取句子中类似N-Gram的关键信息，适合短句子文本。而TextRNN擅长捕获更长的序列信息。具体到文本分类任务中，从某种意义上可以理解为可以捕获变长、单向的N-Gram信息（Bi-LSTM可以是双向）。

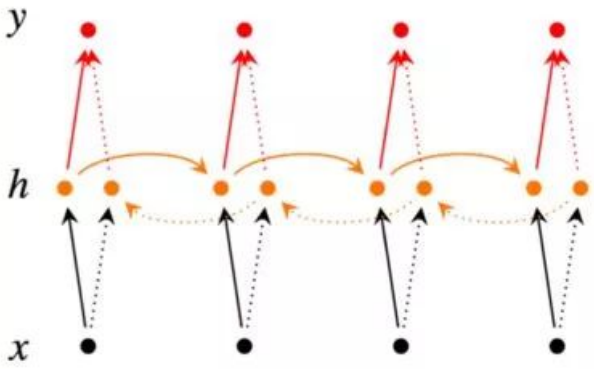
TextRnn网络结构示例图：



该组件对应的是SimpleRnn，SimpleRnn是简单的TextRNN网络，其结构如下：



在SimpleRNN中只考虑了预测词前面的词，即只考虑了上下文中“上文”，并没有考虑该词后面的内容。这可能会错过了一些重要的信息，使得预测的内容不够准确。而双向SimpleRNN不仅从前往后(如下图黄色实箭头)保留该词前面的词的重要信息，而且从后往前(如下图黄色虚箭头)去保留该词后面的词的重要信息，然后基于这些重要信息进行预测该词。双向SimpleRNN模型如下：



（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	特征列格式	输入必须为文本分词结果，列表转换成的字符或者列表，如“["数据","工程师"]”	
3	model	不必须，传入word2vec或者doc2vec模型用于网络中的词嵌入层	

（3）输出

序号	名称	内容
1	data_out	训练数据基于其自身训练生成Textrnn网络得到到的预测结果
2	model	训练生成的网络模型
3	data_in_test	Textrnn模型对测试集的预测结果，保存在原始输入测试集中共同输出，当且仅当输入测试集时生效；

（4）参数

序号	分组	参数	说明
1	全连接层参数	激活函数	
2	全连接层参数	正则化参数	防止过拟合，0到1之间数值
3	模型训练参数	损失函数	用来计算测试集中目标值Y的真实值和预测值的偏差程度，默认为“categorical_crossentropy”
4	模型训练参数	批量次数	表示一个批次的样本数量
5	模型训练参数	优化器	默认为“adam”，用于通过训练优化参数，来最小化（最大化）损失函数
6	模型训练参数	训练次数	
7	模型训练参数	验证集比例	训练集中用于做验证的数据的比例
8	模型训练参数	特征数（词汇个数）	训练集中样本包含的词汇个数，若不设置特征数，将默认为训练集的最大分词数
9	模型训练参数	评估标准	
10	词嵌入层参数	是否引入外部词向量模型	若选择是,则需输入word2vec或者doc2vec模型
11	词嵌入层参数	词向量维度	为避免训练时间过长，此处将词向量维度限制在前1000；引入外部模型，则该参数从外部模型中获取
12	卷积层参数	卷积窗口的长度	允许输入多个数值，数值与数值间采用英文逗号分隔；数值个数代表并列卷积层的层数，默认采用3个卷积层
13	卷积层参数	填充方式	其中选择“same”表示填充输入以使输出具有与原始输入相同的长度，选择“valid”表示不填充

序号	分组	参数	说明
14	卷积层参数	卷积中滤波器的输出数量	
15	卷积层参数	激活函数	
16	卷积层参数	卷积步长	在文本处理中，该参数常设置为1
17	输入设置	训练集输入特征	已分词的文本数据列，列表或者列表类型的字符串，只允许选择单列
18	输入设置	训练集目标输出	目标分类列，离散型数值或文本。
19	输入设置	测试集输入特征	若不进行模型效果测试，可忽略该参数
20	输入设置	测试机目标输出	若不进行模型效果测试，可忽略该参数
21	RNN层参数	模型类型	单向、双向（网络结构中Bi-LSTM可以是双向）
22	RNN层参数	使用偏置向量	
23	RNN层参数	神经元个数	

（5）示例

对留言文本数据集进行TextRnn文本分类示例。

	A	B	C	D	E	F
1	id	user_id	them	time	detail	first_class
2	102738	A0008529	K2区映山	2014/3/24	我家住在K市K2区华源府第小区	环境保护
3	127391	A0008525	L5县环保	2016/12/2	县环保局噪声测试不按国家规定	环境保护
4	144491	A0003479	请责令M	2012/9/3 7	U优会所早晨排污持续，请责令	环境保护
5	118104	A0005202	K11县县	2018/9/25	我投资200余万元的纸厂被县环保局	环境保护
6	140709	A0006959	M1区环保	2017/7/2 6	您好：日兴砖厂是一家打着环保	环境保护
7	174438	A0001663	为何日月	2012/12/1	邓书记： 您好！日月星城KTV营业	环境保护
8	94729	A0009906	J9县沙田	2014/3/18	尊敬的黄县长： 您好！请您帮帮	环境保护
9	121512	A0007873	L市小型砖	2017/11/1	尊敬的彭书记； 你好！我是	环境保护
10	127245	A0006728	关于L5县	2017/8/27	尊敬的蒙书记 你好！ 百	环境保护
11	170620	A0008893	G8县薮城	2019/11/2	您好！我是G8县薮城石膏实业有限公	环境保护
12	161600	A0001571	西地省盛	2014/8/14	在临G5县合口镇有个大的盛常玻	环境保护
13	181619	A0001904	I3县南洲	2018/6/6 1	I3县南洲镇鑫顺广场A栋一楼开了一家	环境保护
14	182424	A0005128	I市山水华	2016/10/8	我们是I市山水华庭的住户，我们	环境保护
15	10689	A0006174	关于取消	2016/1/8 1	尊敬的领导： 您好！ 我们是	环境保护
16	161993	A0005596	A市和顺	2015/11/9	1．国家《电磁辐射管理办法》规定	环境保护
17	138263	A0007788	M4市数千	2017/4/9 1	河流守望者；接到来自西地省M4	环境保护
18	130815	A0006794	泸阳镇下	2016/12/2	泸阳镇下坪村与壮稻村，采石场	环境保护
19	138105	A0005961	M4市中连	2017/7/14	尊敬的李书记 您好！ 在	环境保护
20	155617	A0009205	L6县巫水	2018/4/15	清明假期，来到阔别多年的西地省，	环境保护
21	117719	A0008602	K市佑康糖	2015/11/9	K市佑康糖尿病专科医院自建院	环境保护
22	139889	A0006086	M2县私人	2018/12/3	呼吁有关部门坚决取缔关停街埠头村	环境保护
23	155419	A0007694	J10 县永	2018/8/20	大源水库是J10 县全县人民饮水取水	环境保护
24	16857	A0004390	第三次请	2017/12/2	你好，本人于9-11月，有多次关	环境保护
25	30073	A0007499	A6区丁字	2016/8/25	尊敬的孔书记 A6区丁字国土所所	环境保护
26	137103	A0004155	请对M2县	2019/2/23	请求政府有关部门责令相关企业采取	环境保护
27	161596	A0006557	请求领导	2014/8/28	尊敬的领导； 我们是M5市三甲乡	环境保护
28	22293	A0001605	关于岳临	2015/1/21	尊敬的周县长，您好！ 我是A8县	环境保护
29	22687	A0001902	A4区捞刀	2015/9/3 9	今年上半年，我写过关于了篇贴	环境保护
30	142722	A0009365	M3县移动	2017/8/24	M3县移动城西贸易区基站由于道	环境保护
31	119997	A0001117	L市凯通邻	2019/9/17	尊敬的领导：凯通领御小区位于湖天	环境保护
32	126447	A0007990	L5县沙溪	2019/6/27	自沙溪炼油厂开厂以来，我村多次举	环境保护
33	126462	A0001081	L5县沙溪	2019/6/15	L5县纪委、监委：自从沙溪村废旧轮	环境保护
34	124323	A0003652	污染触目	2018/4/9 1	各位网友，你们好！我们是L12市的居	环境保护

首先将liuyan数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“liuyan”，勾选文件“留言.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行分词，对留言详情文本进行分词，将【结巴分词】组件与输入源连接，在输入设置中勾选detail特征列进行分词操作，右键单击【结巴分词】算法，选择“运行该节点”。

② 进行TextRnn，将【Rnn】组件与【结巴分词】组件连接，在输入设置中勾选特征列与文本分类列，在参数配置中设置相对应的参数，右键单击【Rnn】组件，选择“运行该节点”。

打开数据，查看结果。对【Rnn】右键选择查看日志，即可查看模型训练过程中个网络层的参数与模型评估结果。

查看日志

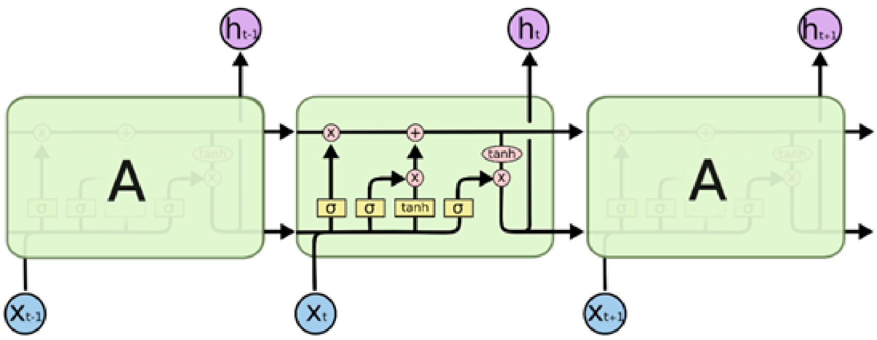
```
7          模型训练   特征数（每一个样本包含的词汇个数                       1227
8          模型训练             损失函数  categorical_crossentropy
9          模型训练              优化器                     adam
10         模型训练             评估标准                       acc
11         模型训练           每批次样本数                         20
12         模型训练             训练次数                       2
13         模型训练            验证集比例                     0.0%
14         模型训练             监视数据                   val_acc
15         模型训练          未进步的训练轮数                          5
三、模型评估
训练集对应的混淆矩阵如下：
训练集中各类型样本的评估指标如下：
          index precision recall f1-score support
0           环境保护       1.00    1.00      1.00      33
1       accuracy       1.00      33      None     None
2      macro avg       1.00    1.00      1.00      33
3   weighted avg       1.00    1.00      1.00      33
/mnt/tipdm-eb-edu/data_1/workspaces/11610/1545314898286878720_0_3
```

8.7.31 文本分类-LSTM

（1）作用

长短时记忆网络(Long Short Term Memory Network, LSTM)，是一种改进之后的循环神经网络，可以解决RNN无法处理长距离的依赖的问题，当前常用于文本生成、机器翻译、语音识别、生成图像描述和视频标记等领域。本组件将LSTM用于文本分类。

LSTM的单个循环结构(又称为细胞)内部有四个状态。相比于RNN，LSTM循环结构之间保持一个持久的单元状态不断传递下去，用于决定哪些信息要遗忘或者继续传递下去，其单层结构和双层结构分别如下：



LSTM_net1

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	特征列格式	输入必须为文本分词结果，列表转换成的字符或者列表，如“["数据","工程师"]”	
3	model	不必须，传入word2vec或者doc2vec模型用于网络中的词嵌入层	

（3）输出

序号	名称	内容
1	data_out	训练数据基于其自身训练生成Textrnn网络得到到的预测结果
2	model	训练生成的网络模型
3	data_in_test	Textrnn模型对测试集的预测结果，保存在原始输入测试集中共同输出，当且仅当输入测试集时生效；

（4）参数

序号	分组	参数	说明
1	全连接层参数	激活函数	
2	全连接层参数	正则化参数	防止过拟合，0到1之间数值
3	模型训练参数	损失函数	用来计算测试集中目标值Y的真实值和预测值的偏差程度，默认为“categorical_crossentropy”
4	模型训练参数	批量次数	表示一个批次的样本数量
5	模型训练参数	优化器	默认为“adam”，用于通过训练优化参数，来最小化（最大化）损失函数
6	模型训练参数	训练次数	
7	模型训练参数	验证集比例	训练集中用于做验证的数据的比例
8	模型训练参数	特征数（词汇个数）	训练集中样本包含的词汇个数，若不设置特征数，将默认为训练集的最大分词数
9	模型训练参数	评估标准	
10	词嵌入层参数	是否引入外部词向量模型	若选择是,则需输入word2vec或者doc2vec模型
11	词嵌入层参数	词向量维度	为避免训练时间过长，此处将词向量维度限制在前1000；引入外部模型，则该参数从外部模型中获取
12	卷积层参数	卷积窗口的长度	允许输入多个数值，数值与数值间采用英文逗号分隔；数值个数代表并列卷积层的层数，默认采用3个卷积层
13	卷积层参数	填充方式	其中选择“same”表示填充输入以使输出具有与原始输入相同的长度，选择“valid”表示不填充

序号	分组	参数	说明
14	卷积层参数	卷积中滤波器的输出数量	
15	卷积层参数	激活函数	
16	卷积层参数	卷积步长	在文本处理中，该参数常设置为1
17	输入设置	训练集输入特征	已分词的文本数据列，列表或者列表类型的字符串，只允许选择单列
18	输入设置	训练集目标输出	目标分类列，离散型数值或文本。
19	输入设置	测试集输入特征	若不进行模型效果测试，可忽略该参数
20	输入设置	测试机目标输出	若不进行模型效果测试，可忽略该参数
21	RNN层参数	模型类型	单向、双向（网络结构中Bi-LSTM可以是双向）
22	RNN层参数	使用偏置向量	
23	RNN层参数	神经元个数	

（5）示例

对留言文本数据集进行LSTM文本分类示例。

	A	B	C	D	E	F
1	id	user_id	them	time	detail	first_class
2	102738	A0008529	K2区映山	2014/3/24	我家住在K市K2区华源府第小区	环境保护
3	127391	A0008525	L5县环保	2016/12/2	县环保局噪声测试不按国家规定	环境保护
4	144491	A0003479	请责令M3	2012/9/3 7	U优会所早晨排污持续，请责令玛	环境保护
5	118104	A0005202	K11县县区	2018/9/25	我投资200余万元的纸厂被县环保局	环境保护
6	140709	A0006959	M1区环保	2017/7/2 6	您好：日兴砖厂是一家打着环保	环境保护
7	174438	A0001663	为何日月	2012/12/1	邓书记： 您好！日月星城KTV营业	环境保护
8	94729	A0009906	J9县沙田	2014/3/18	尊敬的黄县长： 您好！请您帮帮	环境保护
9	121512	A0007873	L市小型砖	2017/11/1	尊敬的彭书记； 你好！我是	环境保护
10	127245	A0006728	关于L5县	2017/8/27	尊敬的蒙书记 你好！ 百	环境保护
11	170620	A0008893	G8县薮城	2019/11/2	您好！我是G8县薮城石膏实业有限公	环境保护
12	161600	A0001571	西地省盛	2014/8/14	在临G5县合口镇有个大的盛常玻	环境保护
13	181619	A0001904	I3县南洲	2018/6/6 1	I3县南洲镇鑫顺广场A栋一楼开了一家	环境保护
14	182424	A0005128	I市山水华	2016/10/8	我们是I市山水华庭的住户，我们	环境保护
15	10689	A0006174	关于取消	2016/1/8 1	尊敬的领导： 您好！ 我们是	环境保护
16	161993	A0005596	A市和顺	2015/11/9	1．国家《电磁辐射管理办法》规定	环境保护
17	138263	A0007788	M4市数千	2017/4/9 1	河流守望者；接到来自西地省M4	环境保护
18	130815	A0006794	泸阳镇下	2016/12/2	泸阳镇下坪村与壮稻村，采石场	环境保护
19	138105	A0005961	M4市中连	2017/7/14	尊敬的李书记 您好！ 在	环境保护
20	155617	A0009205	L6县巫水	2018/4/15	清明假期，来到阔别多年的西地省，	环境保护
21	117719	A0008602	K市佑康糖	2015/11/9	K市佑康糖尿病专科医院自建院	环境保护
22	139889	A0006086	M2县私人	2018/12/3	呼吁有关部门坚决取缔关停街埠头村	环境保护
23	155419	A0007694	J10县永	2018/8/20	大源水库是J10县全县人民饮水取水	环境保护
24	16857	A0004390	第三次请	2017/12/2	你好，本人于9-11月，有多次关	环境保护
25	30073	A0007499	A6区丁字	2016/8/25	尊敬的孔书记 A6区丁字国土所所	环境保护
26	137103	A0004155	请对M2县	2019/2/23	请求政府有关部门责令相关企业采取	环境保护
27	161596	A0006557	请求领导	2014/8/28	尊敬的领导； 我们是M5市三甲乡	环境保护
28	22293	A0001605	关于岳临	2015/1/21	尊敬的周县长，您好！ 我是A8县	环境保护
29	22687	A0001902	A4区捞刀	2015/9/3 9	今年上半年，我写过关于了篇贴	环境保护
30	142722	A0009365	M3县移动	2017/8/24	M3县移动城西贸易区基站由于道	环境保护
31	119997	A0001117	L市凯通邻	2019/9/17	尊敬的领导：凯通领御小区位于湖天	环境保护
32	126447	A0007990	L5县沙溪	2019/6/27	自沙溪炼油厂开厂以来，我村多次举	环境保护
33	126462	A0001081	L5县沙溪	2019/6/15	L5县纪委、监委：自从沙溪村废旧轮	环境保护
34	124323	A0003652	污染触目	2018/4/9 1	各位网友，你们好！我们是L12市的居	环境保护

首先将liuyan数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“liuyan”，勾选文件“留言.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行分词，对留言详情文本进行分词，将【结巴分词】组件与输入源连接，在输入设置中勾选detail特征列进行分词操作，右键单击【结巴分词】算法，选择“运行该节点”。

② 进行LSTM，将【LSTM】组件与【结巴分词】组件连接，在输入设置中勾选特征列与文本分类列，在参数配置中设置相对应的参数，右键单击【LSTM】组件，选择“运行该节点”。

打开数据，查看结果。对【LSTM】右键选择查看日志，即可查看模型训练过程中个网络层的参数与模型评估结果。

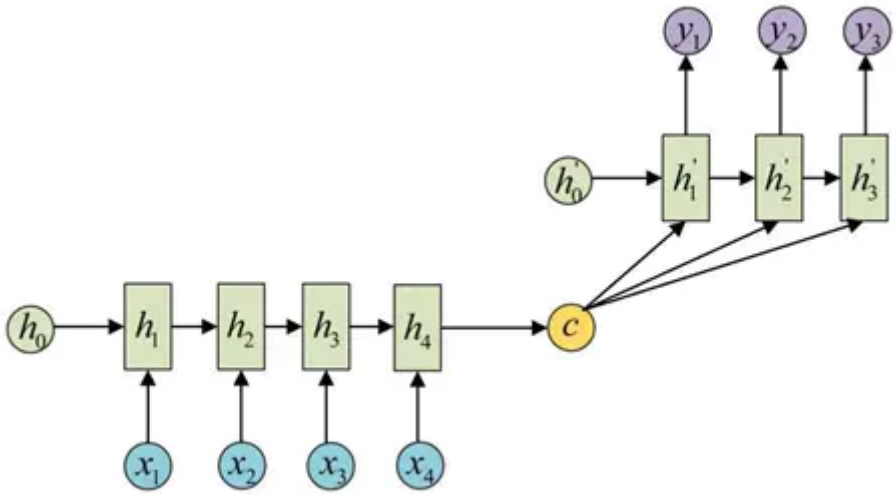
查看日志

```
Epoch 1/2
2/2 [==============================] - ETA: 0s - loss: 0.0000e+00 - acc: 1.0000WARNING:tensorflow:Early stopping conditioned on metric `val_acc` which is not available. Available metrics are: loss,acc
2/2 [==============================] - 1s 384ms/step - loss: 0.0000e+00 - acc: 1.0000
Epoch 2/2
2/2 [==============================] - ETA: 0s - loss: 0.0000e+00 - acc: 1.0000WARNING:tensorflow:Early stopping conditioned on metric `val_acc` which is not available. Available metrics are: loss,acc
2/2 [==============================] - 2s 844ms/step - loss: 0.0000e+00 - acc: 1.0000
('Failed to import pydot. You must `pip install pydot` and install graphviz (https://graphviz.gitlab.io/download/), ', 'for `pydotprint` to work.')
lstm网络训练结果报告
一、数据分布
训练集数据样本量为：33
待划分的样本类型有：环境保护
二、参数选取与网络结构
lstm网络各参数取值如下：
                参数类型          参数名称          参数取值
0   词嵌入层（emdedding）          词向量维度          100
1              RNN层          模型类型          False
2              RNN层          神经元个数          32
```

8.7.32 Seq2seq

（1）作用

Seq2Seq模型是RNN最重要的一个变种，Seq2Seq Model是序列到序列（Sequence to Sequence）模型的简称，也被称为一种编码器-解码器（Encoder-Decoder）模型。



（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	特征列格式	输入必须为文本分词结果，列表转换成的字符或者列表，如“["数据","工程师"]”	
3	model	不必须，传入word2vec或者doc2vec模型用于网络中的词嵌入层	

（3）输出

序号	名称	内容
1	data_out	训练数据基于其自身训练生成Textrnn网络得到到的预测结果
2	model	训练生成的网络模型
3	data_in_test	Textrnn模型对测试集的预测结果，保存在原始输入测试集中共同输出，当且仅当输入测试集时生效；

（4）参数

序号	分组	参数	说明
1	模型评估参数	是否引入测试数据	
2	模型训练参数	损失函数	用来计算测试集中目标值Y的真实值和预测值的偏差程度，默认为“categorical_crossentropy”
3	模型训练参数	批量次数	表示一个批次的样本数量
4	模型训练参数	监视数据	当被监测的数量不再提升，则停止训练
5	模型训练参数	训练次数	
6	模型训练参数	验证集比例	训练集中用于做验证的数据的比例
7	模型训练参数	特征数（词汇个数）	训练集中样本包含的词汇个数，若不设置特征数，将默认为训练集的最大分词数
8	模型训练参数	评估标准	
9	输入设置	训练集输入特征	已分词的文本数据列，列表或者列表类型的字符串，只允许选择单列
10	输入设置	训练集目标输出	目标分类列，离散型数值或文本。
11	输入设置	测试集输入特征	若不进行模型效果测试，可忽略该参数
12	输入设置	测试机目标输出	若不进行模型效果测试，可忽略该参数
13	模型训练参数	优化器	
14	模型训练参数	未进步的训练轮数	没有进步的训练轮数，在这之后训练就会被停止。

（5）示例

对留言文本数据集进行Seq2seq文本分类示例。

	A	B	C	D	E	F
1	id	user_id	them	time	detail	first_class
2	102738	A0008529	K2区映山	2014/3/24	我家住在K市K2区华源府第小区，	环境保护
3	127391	A0008525	L5县环保	2016/12/2	县环保局噪声测试不按国家规定	环境保护
4	144491	A0003479	请责令M	2012/9/3 7	U优会所早晨排污持续，请责令	环境保护
5	118104	A0005202	K11县县	2018/9/25	我投资200余万元的纸厂被县环保局	环境保护
6	140709	A0006959	M1区环保	2017/7/2 6	您好：日兴砖厂是一家打着环保	环境保护
7	174438	A0001663	为何日月	2012/12/1	邓书记： 您好！日月星城KTV营业	环境保护
8	94729	A0009906	J9县沙田	2014/3/18	尊敬的黄县长： 您好！请您帮帮	环境保护
9	121512	A0007873	L市小型砖	2017/11/1	尊敬的彭书记； 你好！我是	环境保护
10	127245	A0006728	关于L5县	2017/8/27	尊敬的蒙书记 你好！ 百	环境保护
11	170620	A0008893	G8县薮城	2019/11/2	您好！我是G8县薮城石膏实业有限公	环境保护
12	161600	A0001571	西地省盛	2014/8/14	在临G5县合口镇有个大的盛常玻	环境保护
13	181619	A0001904	I3县南洲	2018/6/6 1	I3县南洲镇鑫顺广场A栋一楼开了一家	环境保护
14	182424	A0005128	I市山水华	2016/10/8	我们是I市山水华庭的住户，我们	环境保护
15	10689	A0006174	关于取消	2016/1/8 1	尊敬的领导： 您好！ 我们是	环境保护
16	161993	A0005596	A市和顺	2015/11/9	1．国家《电磁辐射管理办法》规定	环境保护
17	138263	A0007788	M4市数千	2017/4/9 1	河流守望者；接到来自西地省M4	环境保护
18	130815	A0006794	泸阳镇下	2016/12/2	泸阳镇下坪村与壮稻村，采石场	环境保护
19	138105	A0005961	M4市中连	2017/7/14	尊敬的李书记 您好！ 在	环境保护
20	155617	A0009205	L6县巫水	2018/4/15	清明假期，来到阔别多年的西地省，	环境保护
21	117719	A0008602	K市佑康糖	2015/11/9	K市佑康糖尿病专科医院自建院	环境保护
22	139889	A0006086	M2县私人	2018/12/3	呼吁有关部门坚决取缔关停街埠头村	环境保护
23	155419	A0007694	J10县永	2018/8/20	大源水库是J10县全县人民饮水取水	环境保护
24	16857	A0004390	第三次请	2017/12/2	你好，本人于9-11月，有多次关	环境保护
25	30073	A0007499	A6区丁字	2016/8/25	尊敬的孔书记 A6区丁字国土所所	环境保护
26	137103	A0004155	请对M2县	2019/2/23	请求政府有关部门责令相关企业采取	环境保护
27	161596	A0006557	请求领导	2014/8/28	尊敬的领导； 我们是M5市三甲乡	环境保护
28	22293	A0001605	关于岳临	2015/1/21	尊敬的周县长，您好！ 我是A8县	环境保护
29	22687	A0001902	A4区捞刀	2015/9/3 9	今年上半年，我写过关于了篇贴	环境保护
30	142722	A0009365	M3县移动	2017/8/24	M3县移动城西贸易区基站由于建	环境保护
31	119997	A0001117	L市凯通领	2019/9/17	尊敬的领导：凯通领御小区位于湖天	环境保护
32	126447	A0007990	L5县沙溪	2019/6/27	自沙溪炼油厂开厂以来，我村多次举	环境保护
33	126462	A0001081	L5县沙溪	2019/6/15	L5县纪委、监委：自从沙溪村废旧轮	环境保护
34	124323	A0003652	污染触目	2018/4/9 1	各位网友，你们好！我们是L12市的居	环境保护

首先将liuyan数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“liuyan”，勾选文件“留言.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行分词，对留言详情文本进行分词，将【结巴分词】组件与输入源连接，在输入设置中勾选detail特征列进行分词操作，右键单击【结巴分词】算法，选择“运行该节点”。

② 进行Seq2seq翻译，将【Seq2seq】组件与【结巴分词】组件连接，在输入设置中勾选特征列与文本分类列，在参数配置中设置相对应的参数，右键单击【Seq2seq】组件，选择“运行该节点”。

8.7.33 命名实体识别-NLTK

（1）作用

命名实体识别作为自然语言处理的子任务之一，旨在通过算法能够自动的识别出一句话中的实体，比如人物、地点、物品、时间、数字等等。

命名实体识别（NER）系统的目标是识别所有文字提及的命名实体。可以分解成两个子任务：确定NE的边界和确定其类型。 命名实体识别非常适用于基于分类器类型的方法来处理的任务。

NLTK有一个已经训练好的可以识别命名实体的分类器，可以使用函数nltk.ne_chunk()进行访问。该组件可以识别所有命名实体或是将所有命名实体识别为各自类型。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out	实体标注结果

（4）参数

序号	分组	参数	说明
1	字段设置	命名实体识别选项	可选型有：识别所有命名实体、将命名实体识别为它们各自的类型
2	字段设置	特征	
3	字段设置	目标列	

（5）示例

对position数据集进行文本独热编码示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行NLTK命名实体识别，将【NLTK命名实体识别】组件与输入源连接，在字段设置中选择目标特征列，命名实体识别选型选择将命名实体识别为它们各自的类型，右键单击【NLTK命名实体识别】组件，选择“运行该节点”。

打开数据，查看结果。对【NLTK命名实体识别】组件右键选择查看数据，即可查看文本实体命名标注结果。

预览数据

��望职位	text_ner
["数据挖掘工程师","算法工程师"]	(S [/NN ``/`` 数据挖掘工程师/NN "/" ,/, "/" 算法工程师/NN "/"]/NN)
["数据分析师","数据挖掘工程师","自然语言处理工程师"]	(S [/NN ``/`` 数据分析师/NN "/" ,/, "/" (ORGANIZATION 数据挖掘工程师/NN) "/" ,/, "/" (ORGANIZATION 自然语言处理工程师/NN) "/"]/NN)
["数据分析师","自然语言处理工程师","数据挖掘工程师"]	(S [/NN ``/`` 数据分析师/NN "/" ,/, "/" (ORGANIZATION 自然语言处理工程师/NN) "/"

8.7.32 命名实体识别-斯坦福

（1）作用

该组件是通过Stanford命名实体识别器对文本数据进行实体识别，同时支持中文与英文两种文本的识别。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out	实体标注结果

（4）参数

序号	分组	参数	说明
1	参数设置	语言	中文、英文
2	字段设置	特征	
3	字段设置	目标列	

（5）示例

对position数据集进行文本独热编码示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行斯坦福命名实体识别，将【斯坦福命名实体识别】组件与输入源连接，在字段设置中选择目标特征列，语言选型选择中文，右键单击【斯坦福命名实体识别】组件，选择“运行该节点”。

打开数据，查看结果。对【斯坦福命名实体识别】组件右键选择查看数据，即可查看文本实体命名标注结果。

预览数据

??望职位	text_ner
["数据挖掘工程师","算法工程师"]	[('[', 'O'), ("", 'O'), ('数据', 'O'), ('挖掘', 'O'), ('工程师', 'TITLE'), ("", 'O'), (',', 'O'), ("", 'O'), ('算法', 'O'), ('工程师', 'TITLE'), ("", 'O'), (']', 'O')]
["数据分析师","数据挖掘工程师","自然语言处理工程师"]	[('[', 'O'), ("", 'O'), ('数据', 'O'), ('分析师', 'TITLE'), ("", 'O'), (',', 'O'), ("", 'O'), ('数据', 'O'), ('挖掘', 'O'), ('工程师', 'TITLE'), ("", 'O'), (',', 'O'), ("", 'O'), ('自然', 'O'), ('语言', 'O'), ('处理', 'O'), ('工程师', 'TITLE'),

8.7.33 命名实体识别-LTP

（1）作用

LTP平台支持人名、地名、机构名三种类型实体的识别。其标注结果采用O-S-B-I-E标注形式，含义为：

标记	含义
O	这个词不是实体
S	这个词单独构成一个实体
B	这个词为一个实体的开始
I	这个词为一个实体的中间
E	这个词位一个实体的结尾

三种实体的含义如下表所示：

标记	含义
Nh	人名
Ni	机构名
Ns	地名

本组件在基于LTP平台开源接口实现命名实体标注的基础上还支持基于标注结果对实体进行提取，提取的规则包括两种：

- 独立实体：以S开头进行标注，如：S-Nh，S-Ni，S-Ns；
- B-I-E形式：由B-I-E标注共同组成实体，如标注[B-Nh,I-Nh,I-Nh,E-Nh]对应的分词直接拼接成一个人名实体。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out	实体标注结果

（4）参数

序号	分组	参数	说明
1	参数设置	提取地名（Ns）	
2	参数设置	提取人名（Nh）	
3	参数设置	提取机构名（Ni）	
4	输入设置	上述字段LTP分词结果字段	
5	输入设置	上述字段LTP词性标注结果字段	
6	输入设置	命名实体识别字段	
7	输出设置	是否允许输出嵌套列表	

（5）示例

对position数据集进行文本独热编码示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行ltp分词与词性标注，将【ltp分词与词性标注】组件与输入源连接，选择输出词性标注结果，右键单击【ltp分词与词性标注】组件，选择“运行该节点”，得到分词后的文本数据。

进行ltp命名实体识别，将【ltp命名实体识别】组件与上面的分词结果进行连接，在字段设置中选择对应的特征列，右键单击【ltp命名实体识别】组件，选择“运行该节点”。

打开数据，查看结果。对【ltp命名实体识别】组件右键选择查看数据，即可查看文本中对人名的提取信息。

8.7.34 命名实体识别-Hanlp

（1）作用

基于HanLP分词与词性标注后的结果，从中提取所需命名实体。HanLP库支持人名、地名、机构名等命名实体的识别，但需在分词与词性标注中进行开启，当前“HanLP分词与词性标注”组件已默认开启所有支持的命名实体识别，因此本组件仅为从从分词与词性标注结果中提取识别的命名实体名称。

注意：建议对命名实体识别要求较高的用户在“HanLP分词与词性标注”组件中，使用”感知机词法分析器“。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out	实体标注结果

（4）参数

序号	分组	参数	说明
1	参数设置	一级实体	选择需要保留的命名实体中的一级分类下的实体名称
2	输入设置	特征	输入基于HanLP分词与词性标注后的序列
3	输出设置	输出文件类型	

一级实体（多选）：必填。请选择需要保留的命名实体中的一级分类下的实体名称。可选项有：

词性	命名实体名称	词性	命名实体名称
nr	人名	nb	生物名
ns	地名	nh	医药疾病等健康相关名词
nt	机构团体名	nf	食品
ni	机构相关（非独立机构名）	nz	其他专名

团体机构名下的二级实体（多选）：非必填，请选择需要保留的机构团体名下的二级分类实体名称。可选项有：

词性	命名实体名称	词性	命名实体名称
ntc	公司名	ntcb	银行
ntcf	工厂	ntch	酒店宾馆
nth	医院	nto	政府机构
nts	中小学	ntu	大学

生物名下的二级实体（多选）：非必填，请选择需要保留的生物名下的二级分类实体名称。可选项有：

词性	命名实体名称
nba	动物名
nbc	动物纲目
nbp	植物名

医药疾病等健康相关名词下的二级实体（多选）：非必填，请选择需要保留的医药疾病等健康相关名词下的二级分类实体名称。可选项有：

词性	命名实体名称
nhd	疾病
nhm	药品

（5）示例

对position数据集进行文本独热编码示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行hanlp分词与词性标注，将【hanlp分词与词性标注】组件与输入源连接，选择输出词性标注结果，右键单击【hanlp分词与词性标注】组件，选择“运行该节点”，得到词性标注的文本数据。

预览数据

_c0	期望职位	期望职位_cutwords	期望职位_cutwordsPos
0	['数据挖掘工程师', '算法工程师']	['数据挖掘', '工程师', '算法', '工程师']	['数据挖掘/gi', '工程师/nnt', '算法/n', '工程师/nnt']
1	['数据分析师', '数据挖掘工程师', '自然语言处理工程师']	['数据', '分析师', '数据挖掘', '工程师', '自然语言处理', '工程师']	['数据/n', '分析师/nnt', '数据挖掘/gi', '工程师/nnt', '自然语言处理/nz', '工程师/nnt']
2	['数据分析师', '自然语言处理工程师', '数据挖掘工程师']	['数据', '分析师', '自然语言处理', '工程师', '数据挖掘', '工程师']	['数据/n', '分析师/nnt', '自然语言处理/nz', '工程师/nnt', '数据挖掘/gi', '工程师/nnt']

进行hanlp命名实体识别，将【hanlp命名实体识别】组件与上面的分词结果进行连接，在字段设置中选择对应的特征列，右键单击【hanlp命名实体识别】组件，选择“运行该节点”。

打开数据，查看结果。对【hanlp命名实体识别】组件右键选择查看数据，即可查看文本中对实体的提取信息。

预览数据

Unnamed: 0	期望职位	期望职位_cutwords	期望职位_cutwordsPos	期望职位_cutwordsPos_ner
0	['数据挖掘工程师', '算法工程师']	['数据挖掘', '工程师', '算法', '工程师']	['数据挖掘/gi', '工程师/nnt', '算法/n', '工程师/nnt']	[{}]
1	['数据分析师', '数据挖掘工程师', '自然语言处理工程师']	['数据', '分析师', '数据挖掘', '工程师', '自然语言处理', '工程师']	['数据/n', '分析师/nnt', '数据挖掘/gi', '工程师/nnt', '自然语言处理/nz', '工程师/nnt']	[{}]
2	['数据分析师', '自然语言处理工程师', '数据挖掘工程师']	['数据', '分析师', '自然语言处理', '工程师', '数据挖掘', '工程师']	['数据/n', '分析师/nnt', '自然语言处理/nz', '工程师/nnt', '数据挖掘/gi', '工程师/nnt']	[{}]

8.7.35 文本相似度-TFIDF

（1）作用

研究短文本相似度的方法基本上都是基于词向量生成句子向量的方法。TF-IDF是最基础的文本相似度计算方法。 TF (Term Frequency)指一篇文档中单词出现的频率，IDF (Inverse Document Frequency)指语料库中出现某个词的文档数，取对数。

TF原理：某个词在一篇文档中出现的频率越多则对这篇文章越重要；

IDF原理：该词在越多的文章中出现，则说明它对文章没有很强的区分度，在文档中所占的权重也就越小，一般采用取词频的逆。 还要考虑一个现象，一些通用词出现的次数可能是低频词的几十倍上百倍，如果只是简单的取逆处理，通用词的权重会变动非常小，稀缺词的权重就显得的过大了。

利用TF-IDF计算文本相似度，需要经过分词、列出所有词、计算词频、写出词频向量等前期，这样就把计算N个文本的相似度变成计算N个向量的相似度。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（3）输出

序号	名称	内容
1	data_out	

（4）参数

序号	分组	参数	说明
1	字段设置	特征	
2	参数设置	相似度计算方法	可选项有： 皮尔森相似度、斯皮尔曼相似度、肯德尔相似度、余弦相似度、欧几里得相似度、曼哈顿距离、马氏距离、杰卡德相似度

（5）示例

对position数据集进行文本独热编码示例。

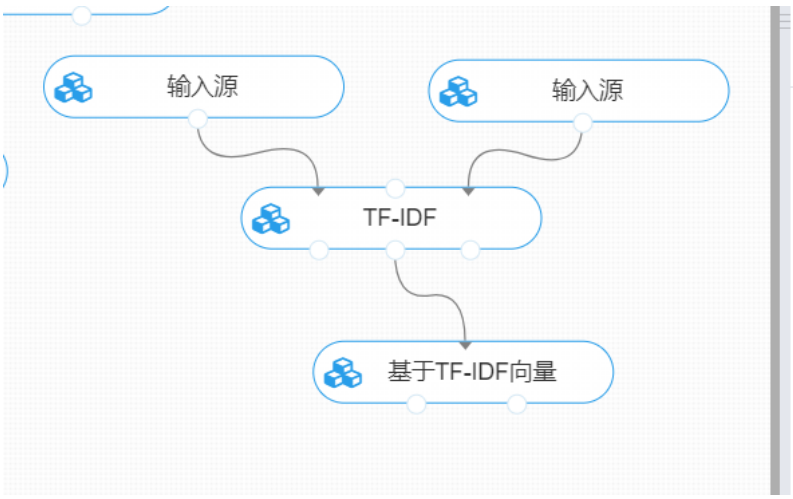
	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行tfidf词向量转化，将【tfidf】组件与输入源连接，选择目标特征列，右键单击【tfidf】组件，选择运行该节点，即可得到文本的词向量结果。

② 进行tfidf文本相似度计算，将【tfidf文本相似度计算】组件与前面的词向量结果连接，选择文本相似度计算方式，右键单击【tfidf文本相似度计算】组件，选择“运行该节点”。



8.7.36 文本相似度-词向量/文档向量

（1）作用

基于word2vec/doc2vec模型进行相似度计算。

利用word2vec进行句子相似度计算，是先将输入query，进行分词，把目标句子的各个词的词向量进行相加取平均，从而把任意长的句子表示为固定维度的向量，然后计算两句子词嵌入之间的余弦相似度，进行相似度比较。

doc2vec是基于word2vec的，word2vec对于计算两个词语的相似度效率比较好，若是文档类型的数据则采用doc2vec模型。

具体的word2vec和doc2vec介绍可查看前面的章节。

（2）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	
2	model	word2vec和doc2vec模型	

（3）输出

序号	名称	内容
1	data_out	原数据和文本间相似度结果

（4）参数

序号	分组	参数	说明
1	字段设置	目标列	分词后的序列
2	字段设置	特征	
3	参数设置	模型选择	word2vec和doc2vec模型
4	参数设置	方式选择	词与词相似度：输入值为词序列，一行一个词； 文档与文档的相似度：输入为分词后的文档序列，每行一个列表，列表中包含多个词，一个列表为一篇文档。

（5）示例

对position数据集进行文本独热编码示例。

	A	B	C	D	E	F
1	期望职位					
2	["数据挖掘工程师","算法工程师"]					
3	["数据分析师","数据挖掘工程师","自然语言处理工程师"]					
4	["数据分析师","自然语言处理工程师","数据挖掘工程师"]					
5	["数据分析师","数据挖掘工程师","算法工程师"]					
6	["数据分析师","数据挖掘工程师"]					
7	["算法工程师","数据分析师","机器学习工程师"]					
8	["数据分析师","数据挖掘工程师"]					
9	["数据分析师","其他","数据挖掘工程师"]					
10	["数据分析师","图像处理工程师","机器学习工程师"]					
11	["Hadoop大数据开发工程师"]					
12	["Hadoop大数据开发工程师"]					
13	["Hadoop大数据开发工程师"]					
14	["数据分析师","机器学习工程师","图像处理工程师"]					
15	["数据分析师"]					
16	["数据分析师","机器学习工程师"]					
17	["数据分析师","数据挖掘工程师","图像处理工程师"]					
18	["数据分析师","数据挖掘工程师","机器学习工程师"]					
19	["数据分析师"]					
20	["数据分析师","数据挖掘工程师"]					
21	["数据分析师"]					
22	["数据分析师","其他"]					
23	["Hadoop大数据开发工程师"]					
24	["Hadoop大数据开发工程师"]					
25	["数据分析师","其他"]					
26	["Hadoop大数据开发工程师"]					
27	["数据分析师","自然语言处理工程师"]					
28	["Hadoop大数据开发工程师","数据分析师","其他"]					

position

首先将position数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“position”，勾选文件“position.csv”，右键单击【输入源】算法，选择“运行该节点”。

① 进行分词，对数据进行分词操作，对【结巴分词】组件与输入源连接，在输入设置中勾选征列进行分词操作，右键单击【结巴分词】算法，选择“运行该节点”。

② 训练word2vec模型，对【word2vec】组件与①的分词结果进行连接，在字段设置中选择“期望职位_cut_words”，右键单击【word2vec】算法，选择“运行该节点”。

③ 进行文本相似度计算，对【词向量/文档向量】组件与上面的分词组件和模型输出进行连接，对相应的参数进行设置，右键单击【词向量/文档向量】算法，选择“运行该节点”。

打开数据，查看结果。对【词向量/文档向量】组件右键选择查看数据，即可查看文本之间的相似度。

预览数据

str_id_x	str_id_y	similarity
2	1	0.687548041343689
2	38	0.5254848003387451
2	20	0.5254848003387451
2	18	0.5254848003387451
2	14	0.5254848003387451
2	13	0.6483370661735535
2	45	0.5179008841514587

8.8 绘图

可视化在数据分析中发挥着重要的作用。Echarts 是个由百度开源的数据可视化，凭借着良好的交互性，精巧的图表设计，得到了许多人的认可。Pyecharts是Python将Echarts结合起来的强大的数据可视化库，帮助我们的绘图组件的得出更具个性化的可视化结果。

8.8.1 面积图

在Pyecharts中本身没有面积图，要实现面积图效果只需要将Line图进行区域填充即可。

（1）输入

序号	条件	要求	说明
1	载入文件格式	CSV文件	

（2）输出

序号	名称	内容
1	日志	面积图展示

（3）参数

序号	分组	参数	说明
1	参数配置	X轴	选择一列作为x轴显示标签
2	参数配置	Y轴	选取一列或多列作为Y轴显示数值
3	样式配置	副标题	不填则不显示
4	样式配置	是否显示数值标签	
5	样式配置	区域透明度	面积区域的透明度，数值型0-1
6	样式配置	翻转X轴	x轴翻转显示
7	样式配置	标题	不填则不显示
8	样式配置	X轴名称	不填则不显示

（4）示例

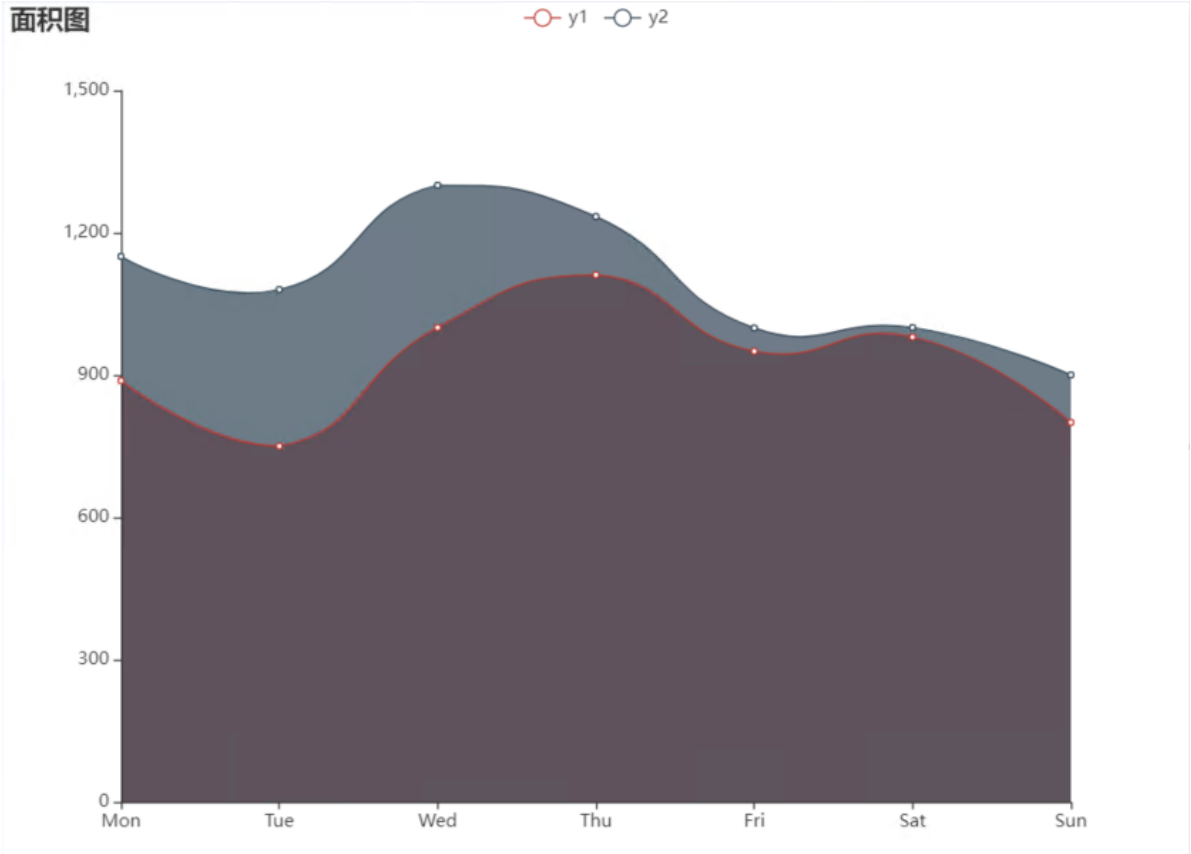
对test数据集进行面积图示例。

	x	y1	y2
2	Mon	888	1150
3	Tue	750	1080
4	Wed	1000	1300
5	Thu	1111	1234
6	Fri	950	999
7	Sat	980	1000
8	Sun	800	900
9			
10			

首先将test数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“tset”，勾选文件“tset.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行面积图，将【面积图】组件与输入源连接，在参数配置中进行X轴与Y轴的特征选择，右键单击【面积图】组件，选择“运行该节点”。

查看日志。对【面积图】右击选择查看日志，即可查看面积图。



8.8.2 树状图

层次化数据的可视化方法，常用于层次聚类中分析聚类情况。

（1）输入

序号	条件	要求	说明
1	载入文件格式	json文件	

（2）输出

序号	名称	内容
1	日志	树状图可视化展示

（3）参数

序号	分组	参数	说明
1	样式配置	字体大小	
2	样式配置	文本角度	
3	样式配置	线性	不同线性展示（曲线、直线）
4	样式配置	结构	树状图的生长方向（左右结构、右左结构、上下结构、下上结构）
5	样式配置	节点形状	
6	样式配置	标题	不填则不显示
7	样式配置	默认显示自己数	如果子集过多，为方便显示可设置显示个数
8	样式配置	布局	正交布局（即水平和垂直 方向）、中心发散（以根节点为圆心向外发散）
9	样式配置	缩放	是否开启鼠标缩放和平移漫游

（4）示例

使用下列文本将其转成.json文件作为示例数据集。

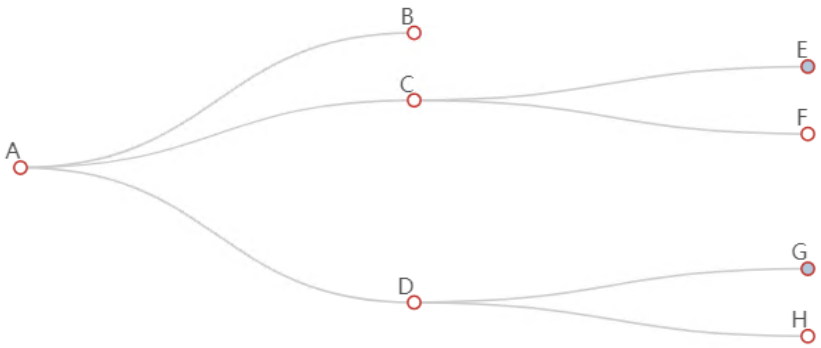
```
{"children": [{"name": "B"},
  {"children": [{"children": [{"name": "I"}], "name": "E"}, {"name": "F"}],
   "name": "C"},
  {"children": [{"children": [{"name": "J"},{"name":"K"}],"name": "G"},
   {"name": "H"}],
   "name": "D"}],
  "name": "A"}
```

首先将tj.json数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“tj”，勾选文件“tj.json”，右键单击【输入源】算法，选择“运行该节点”。

进行树状图，将【树状图】组件与输入源连接，在样式配置中进行样式设置后，右键单击【树状图】组件，选择“运行该节点”。

查看日志。对【树状图】右击选择查看日志，即可查看树状图。

树状图



8.8.3 雷达图

雷达图也称为网络图、蜘蛛图、星图、蜘蛛网图，它被认为是一种表现多维数据的图表。它将多个维度的数据量映射到坐标轴上，每一个维度的数据都分别对应一个坐标轴，这些坐标轴以相同的间距沿着径向排列，并且刻度相同。连接各个坐标轴的网格线通常只作为辅助元素，将各个坐标轴上的数据点用线连接起来就形成了一个多边形。即坐标轴、点、线、多边形共同组成了雷达图。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	
2	各特征维度取值范围	载入csv文件	

（2）输出

序号	名称	内容
1	日志	雷达图可视化展示

（3）参数

序号	分组	参数	说明
1	参数配置	数值列	
2	参数配置	图例列	数值列的列不能包含图例列的列
3	样式配置	副标题	
4	样式配置	底图性质	
5	样式配置	区域透明图	雷达图面积区域透明度
6	样式配置	显示数值标签	

（4）示例

以空气质量评价数据集bj作为示例：

years	AQI	PM2.5	PM10	CO	NO2	SO2
2002	55	9	56	0.46	18	6
2003	25	11	21	0.65	34	9
2004	56	7	63	0.3	14	5
2005	33	7	29	0.33	16	6
2006	42	24	44	0.76	40	16
2007	82	58	90	1.77	68	33
2008	74	49	77	1.46	48	27
2009	78	55	80	1.29	59	29
2010	267	216	280	4.8	108	64
2011	185	127	216	2.52	61	27
2012	39	19	38	0.57	31	15
2013	41	11	40	0.43	21	7

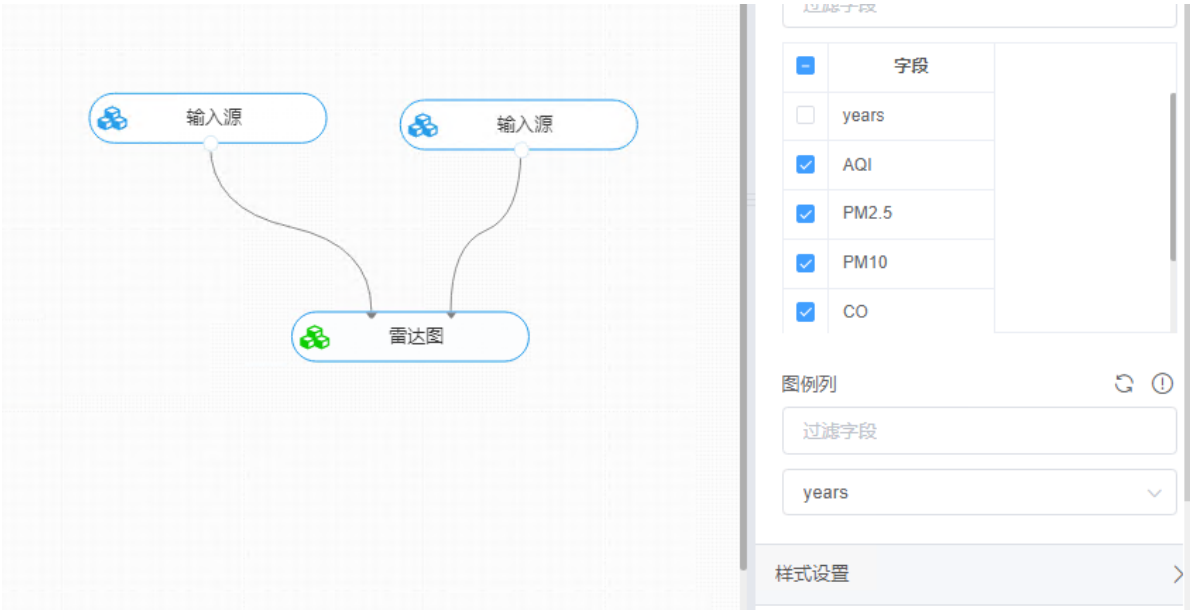
以及各特征维度取值范围表：

name	max	min
AQI	300	5
PM2.5	250	20
PM10	300	5
CO	5	
NO2	200	
SO2	100	

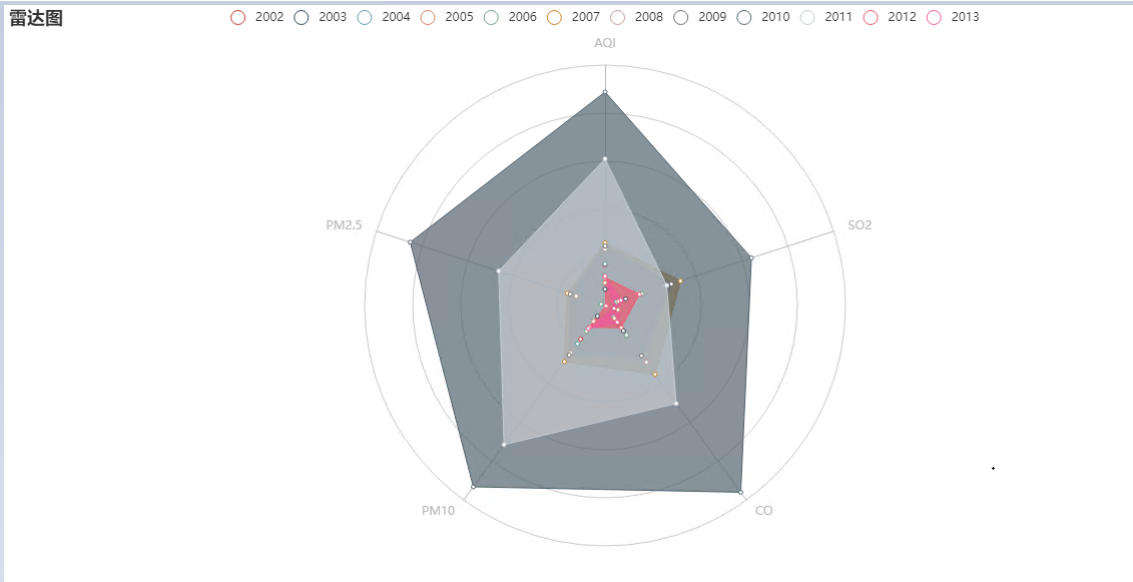
air_quality_schema_data

首先将bj数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“bj”，勾选文件“bj.csv”，同样的操作将air_quality_schema_data（取值范围表）也读入系统。

进行雷达图图，将【雷达图】组件与输入源连接，在参数配置中数值列中勾选所需的数值特征，在图例列中勾选标签特征列，继而在样式配置中进行相应的样式配置，右键单击【雷达图】组件，选择“运行该节点”。



查看日志。对【雷达图】右击选择查看日志，即可查看雷达图。



8.8.4 直方图

直方图(Histogram)是一种二维图表，又称质量分布图，是一种统计报告图，由一系列高度不等的纵向条纹或线段表示数据分布的情况，显示数据之间的差别， 一般用横轴表示数据类型，纵轴表示分布情况。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	
2	数据量	不少于50条	

（2）输出

序号	名称	内容
1	日志	直方图可视化展示

（3）参数

序号	分组	参数	说明
1	参数配置	Y轴	选择目标 数值列 作为y轴数值展示
2	参数配置	X轴	x轴标签显示
3	样式配置	翻转x轴	
4	样式配置	区域缩放	
5	样式配置	标题	
6	样式配置	x轴名称	
7	样式配置	显示数值标签	

（4）示例

以iris数据集作为示例。

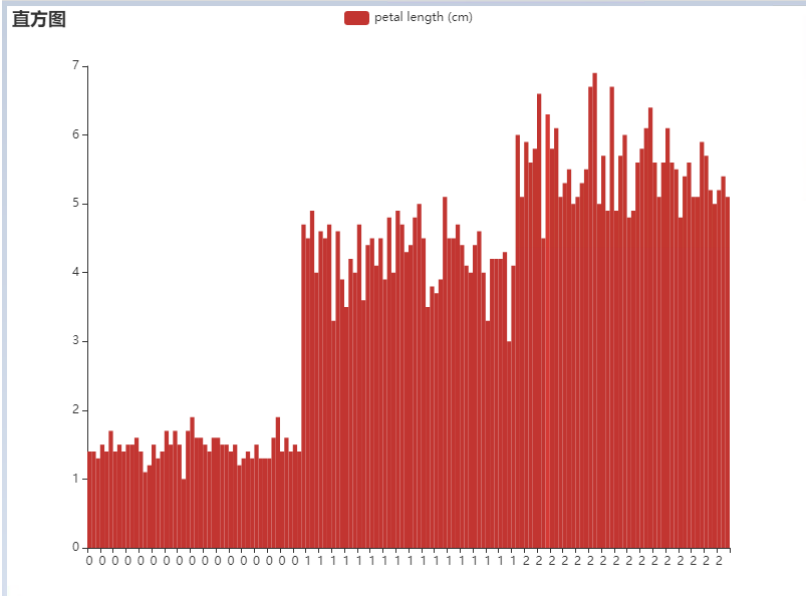
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

先将需要进行直方图的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

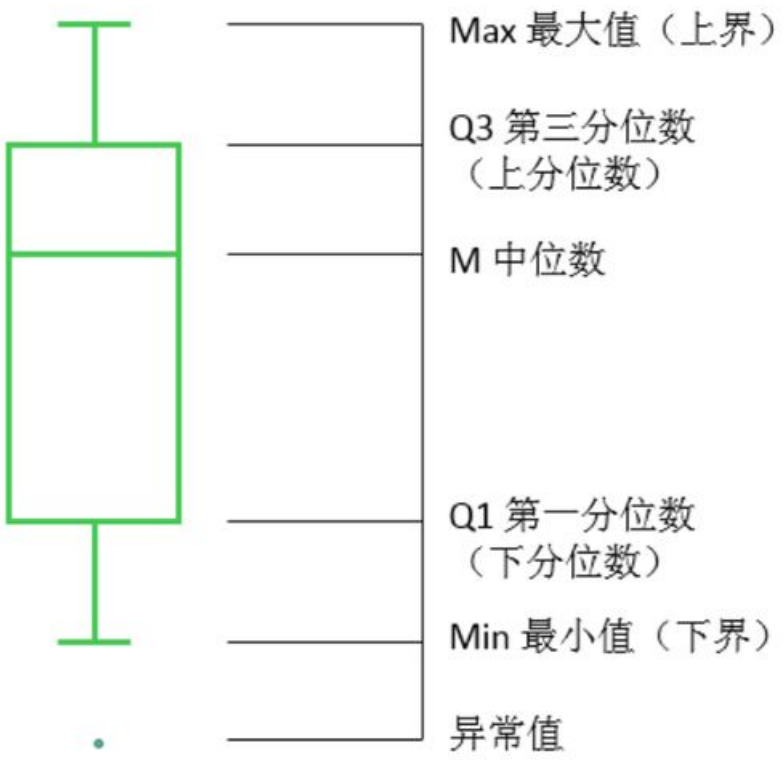
开始进行直方图，将【输入源】和【直方图】相连接，在参数配置中选择y轴与x轴所需的特征列，继而进行相应的样式配置，右键单击【直方图】算法，选择“运行该节点”。

查看日志。对【雷达图】右击选择查看日志，即可查看样本数据的直方图分布情况。



8.8.5 箱线图

箱线图（Box-plot）又称为盒须图、盒式图或箱线图，是一种用作显示一组数据分散情况的统计图。它是利用数据中的五个统计量：最小值、第一四分位数、中位数、第三四分位数、与最大值来描述数据的一种方法。



（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	

（2）输出

序号	名称	内容
1	日志	箱线图可视化展示

（3）参数

序号	分组	参数	说明
1	参数配置	数值列	可以选择多列特征进行数值展示
2	样式配置	副标题	不填则不显示
3	样式配置	y轴最大值/最小值	设置y轴的取值范围
4	样式配置	分割线	在多个数据里用分割线显示

（4）示例

采用iris数据集进行示例。

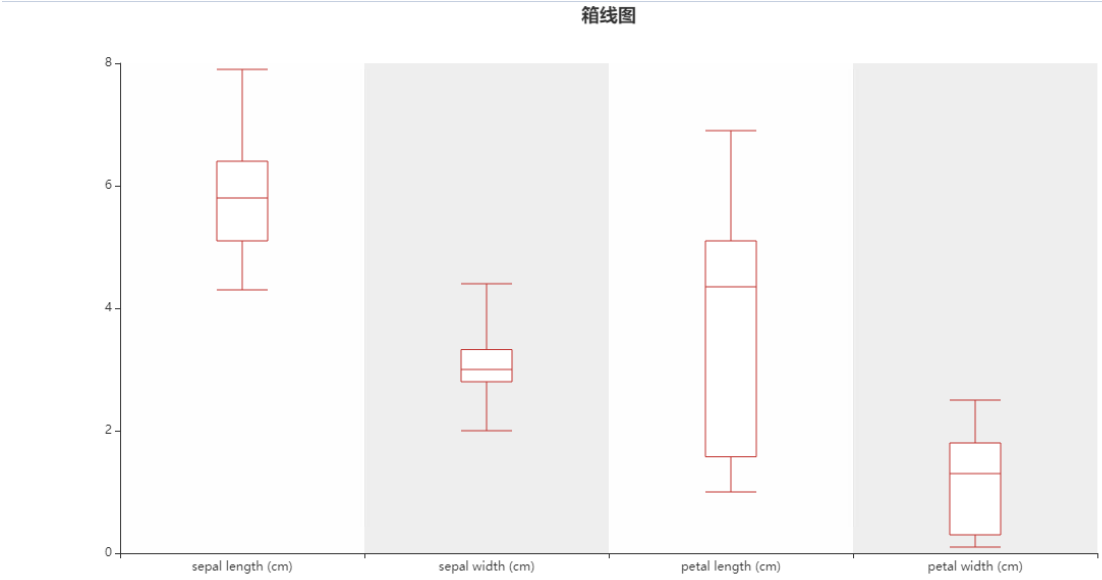
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

先将需要进行箱线图的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点

进行箱线图操作。拖入【箱线图】，将【输入源】和【箱线图】相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段，继而进行所需的样式配置。右键单击【箱线图】，选择“运行该节点”。

查看日志。对【直方图】右击选择查看日志，即可查看样本数据的箱线图分布情况。



8.8.6 饼图

饼图常用于统计学模型，反映各组数据之间的比例。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	

（2）输出

序号	名称	内容
1	日志	饼图可视化展示

（3）参数

序号	分组	参数	说明
1	参数配置	Y轴	数值列
2	参数配置	X轴	标签列
3	样式配置	南丁格尔图类型	radius：扇区圆心角展现数据的百分比，半径展现数据的大小； area：所有扇区圆心角相同，仅通过半径展现数据大小
4	样式配置	中心镂空	

（4）示例

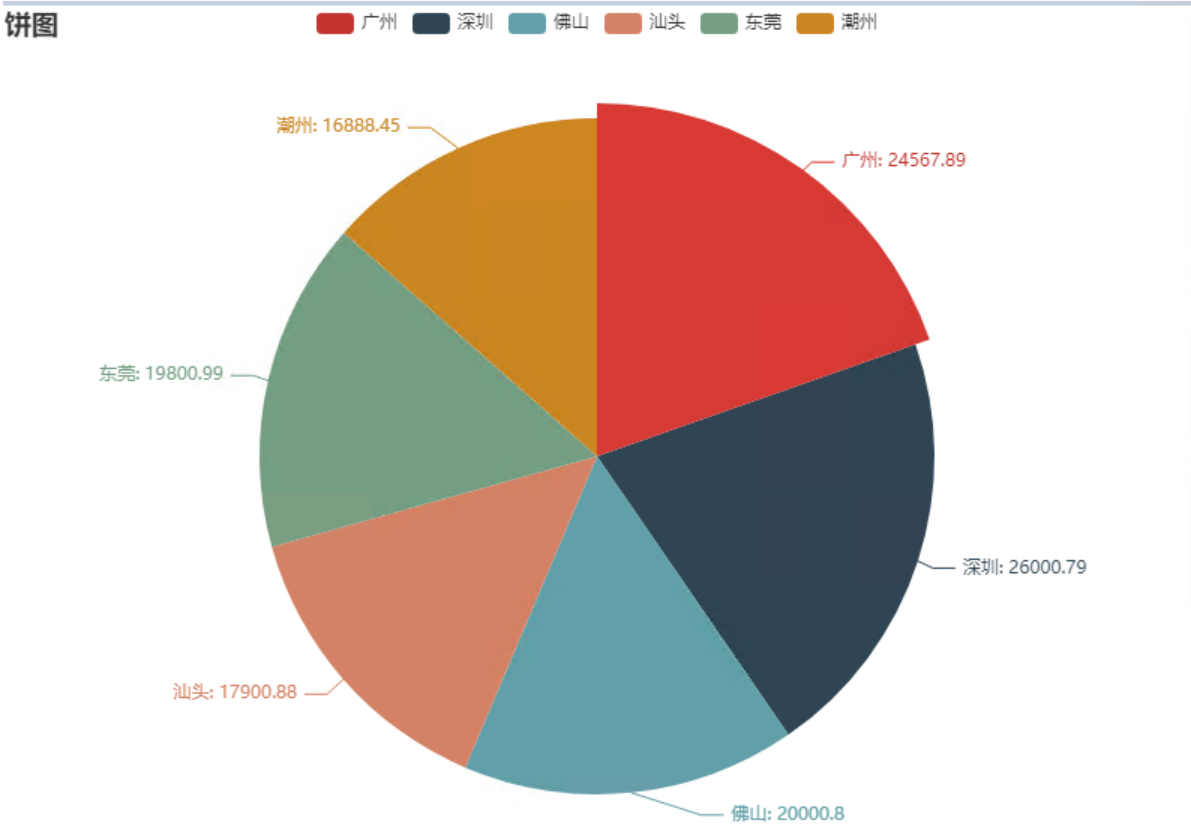
采用sale数据集进行示例。

	A	B	C
1	地点	销售额	
2	广州	24567.89	
3	深圳	26000.79	
4	佛山	20000.8	
5	汕头	17900.88	
6	东莞	19800.99	
7	潮州	16888.45	
8			

先将需要进行饼图的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“sale”，勾选文件“sale.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行饼图操作。拖入【饼图】，将【输入源】和【饼图】相连接，在参数配置的X轴选择目标标签列，Y轴选择目标数值列，继而进行所需的样式配置。右键单击【饼图】，选择“运行该节点”。

查看日志。对【饼图】右击选择查看日志，即可查看样本数据的饼图占比情况。



8.8.7 散点图

散点图是指在数理统计回归分析中，数据点在直角坐标系平面上的分布图，其表示因变量随自变量而变化的大致趋势，由此趋势可以选择合适的函数进行经验分布的拟合，进而找到变量之间的函数关系。散点图适用于大量数据中寻找规律。

散点图主要的构成元素有：数据源，横纵坐标轴，变量名，研究的对象。而基本的要素就是点，也就是我们统计的数据，由这些点的分布我们才能观察出变量之间的关系。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	

（2）输出

序号	名称	内容
1	日志	散点图可视化展示

（3）参数

序号	分组	参数	说明
1	参数配置	Y轴	数值列，可多个特征列（Y轴的所有列不能包含有X轴的列）
2	参数配置	X轴	标签列
3	样式配置	标记线	最大值、最小值和平均值标记线
4	样式配置	中心镂空	
5	样式配置	副标题	
6	样式配置	分割线	
7	样式配置	翻转X轴	
8	样式配置	标题	不填则不显示
9	样式配置	区域缩放	
10	样式配置	X轴名称	不填则不显示
11	样式配置	显示数值标签	

（4）示例

采用iris数据集进行示例，绘制“petal length”和“petal width”特征之间的散点图。

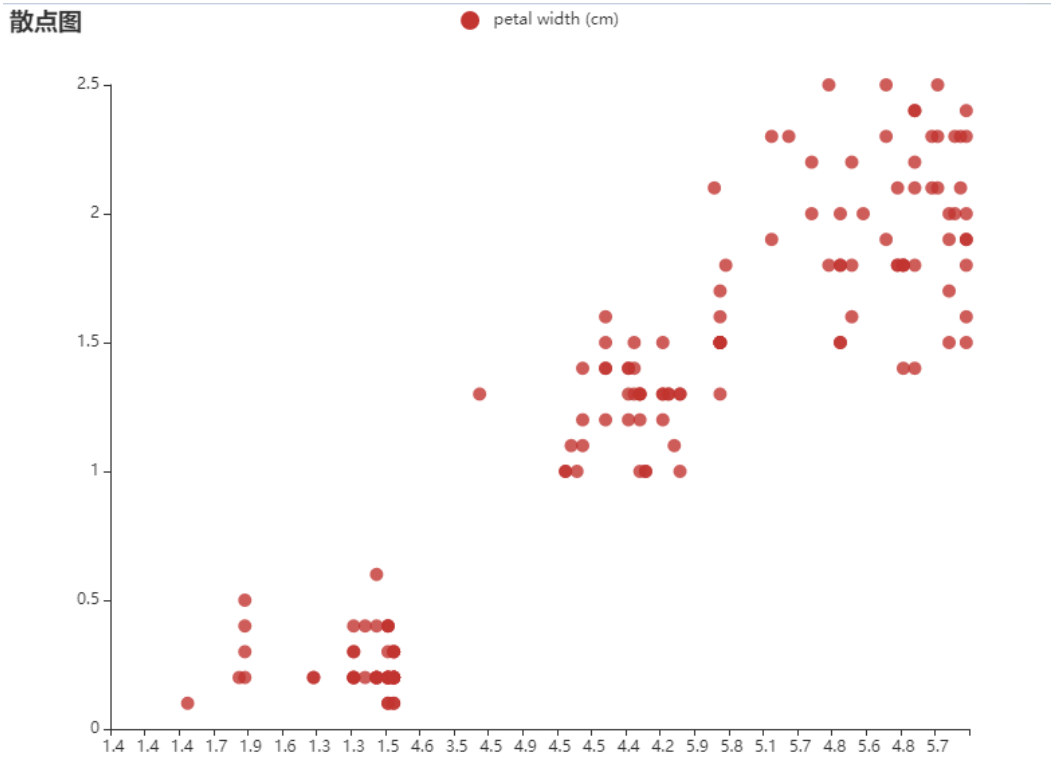
	A	B	C	D	E	F
1	sepal leng	sepal widt	petal leng	petal widt	outcome	
2	5.1	3.5	1.4	0.2	0	
3	4.9	3	1.4	0.2	0	
4	4.7	3.2	1.3	0.2	0	
5	4.6	3.1	1.5	0.2	0	
6	5	3.6	1.4	0.2	0	
7	5.4	3.9	1.7	0.4	0	
8	4.6	3.4	1.4	0.3	0	
9	5	3.4	1.5	0.2	0	
10	4.4	2.9	1.4	0.2	0	
11	4.9	3.1	1.5	0.1	0	
12	5.4	3.7	1.5	0.2	0	
13	4.8	3.4	1.6	0.2	0	
14	4.8	3	1.4	0.1	0	
15	4.3	3	1.1	0.1	0	
16	5.8	4	1.2	0.2	0	
17	5.7	4.4	1.5	0.4	0	
18	5.4	3.9	1.3	0.4	0	
19	5.1	3.5	1.4	0.3	0	
20	5.7	3.8	1.7	0.3	0	
21	5.1	3.8	1.5	0.3	0	
22	5.4	3.4	1.7	0.2	0	
23	5.1	3.7	1.5	0.4	0	
24	4.6	3.6	1	0.2	0	
25	5.1	3.3	1.7	0.5	0	
26	4.8	3.4	1.9	0.2	0	
27	5	3	1.6	0.2	0	
28	5	3.4	1.6	0.4	0	

iris

先将需要进行散点图的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】组件，点击【输入源】组件，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行散点图操作。拖入【散点图】，将【输入源】和【散点图】相连接，在参数配置的X轴选择目标标签列，Y轴选择目标数值列，继而进行所需的样式配置。右键单击【散点图】，选择“运行该节点”。

查看日志。对【散点图】右击选择查看日志，即可查看样本数据的散点图分布趋势。



8.8.8 柱形图

柱形图是一种用矩形柱来表示数据分类的图表，柱形图可以垂直绘制，也可以水平绘制，它的高度与其所表示的数值成正比关系。 柱形图显示了不同类别之间的比较关系和同类别各变量之间的比较情况，图表的水平轴 X 指定被比较的类别，垂直轴 Y 则表示具体的类别值。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	
2	数据量	建议较小数据集	

（2）输出

序号	名称	内容
1	日志	柱形图可视化展示

（3）参数

序号	分组	参数	说明
1	参数配置	Y轴	数值列，可多个特征列
2	参数配置	X轴	标签列
3	样式配置	堆叠模式	若有多个特征列可开启该模型来设置柱形的分布
4	样式配置	横向显示	柱形的生长方向（横向、纵向）
5	样式配置	副标题	
6	样式配置	翻转X轴	
7	样式配置	区域缩放	
8	样式配置	标记线	最大值、最小值和平均值标记线
9	样式配置	显示数值标签	
10	样式配置	X轴名称	不填则不显示

（4）示例

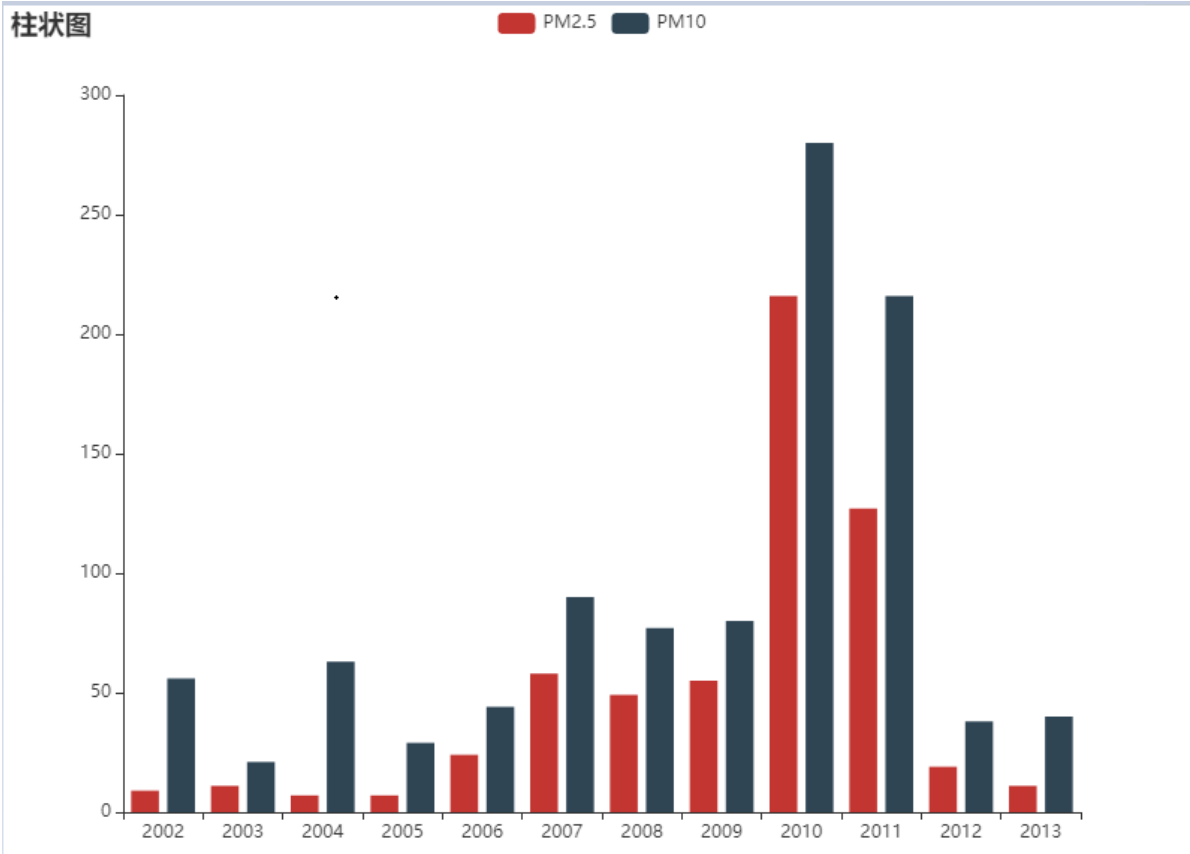
以空气质量评价数据作为示例。

years	AQI	PM2.5	PM10	CO	NO2	SO2
2002	55	9	56	0.46	18	6
2003	25	11	21	0.65	34	9
2004	56	7	63	0.3	14	5
2005	33	7	29	0.33	16	6
2006	42	24	44	0.76	40	16
2007	82	58	90	1.77	68	33
2008	74	49	77	1.46	48	27
2009	78	55	80	1.29	59	29
2010	267	216	280	4.8	108	64
2011	185	127	216	2.52	61	27
2012	39	19	38	0.57	31	15
2013	41	11	40	0.43	21	7

先将需要进行柱形图的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】组件，点击【输入源】组件，填写数据集名称“bj”，勾选文件“bj.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行柱形图操作。拖入【柱形图】，将【输入源】和【柱形图】相连接，在参数配置的X轴选择目标标签列，Y轴选择目标数值列，继而进行所需的样式配置。右键单击【柱形图】，选择“运行该节点”。

查看日志。对【柱形图图】右击选择查看日志，即可查看样本数据的各类别之间的对比情况。



8.8.9 折线图

观察一个或者多个数据指标连续变化的趋势，显示随时间而变化的连续数据，所有值数据沿垂直轴均匀分布。因此非常适用于显示在相等时间间隔下数据的趋势。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	

（2）输出

序号	名称	内容
1	日志	折线图可视化展示

（3）参数

序号	分组	参数	说明
1	参数配置	Y轴	数值列，可多个特征列
2	参数配置	X轴	标签列
3	样式配置	堆叠模式	若选择多个特征列，第一个数据系列和折线图中显示的是一样的，而第二个数据系列的值要和第一个数据系列的值在同一分类（或时间上）进行累计， 这样可以显示两个数据系列在同一分类（或时间上）的值的总和的发展变化趋势
4	样式配置	横向显示	柱形的生长方向（横向、纵向）
5	样式配置	副标题	
6	样式配置	翻转X轴	
7	样式配置	区域缩放	
8	样式配置	标记线	最大值、最小值和平均值标记线
9	样式配置	显示数值标签	
10	样式配置	X轴名称	不填则不显示

（4）示例

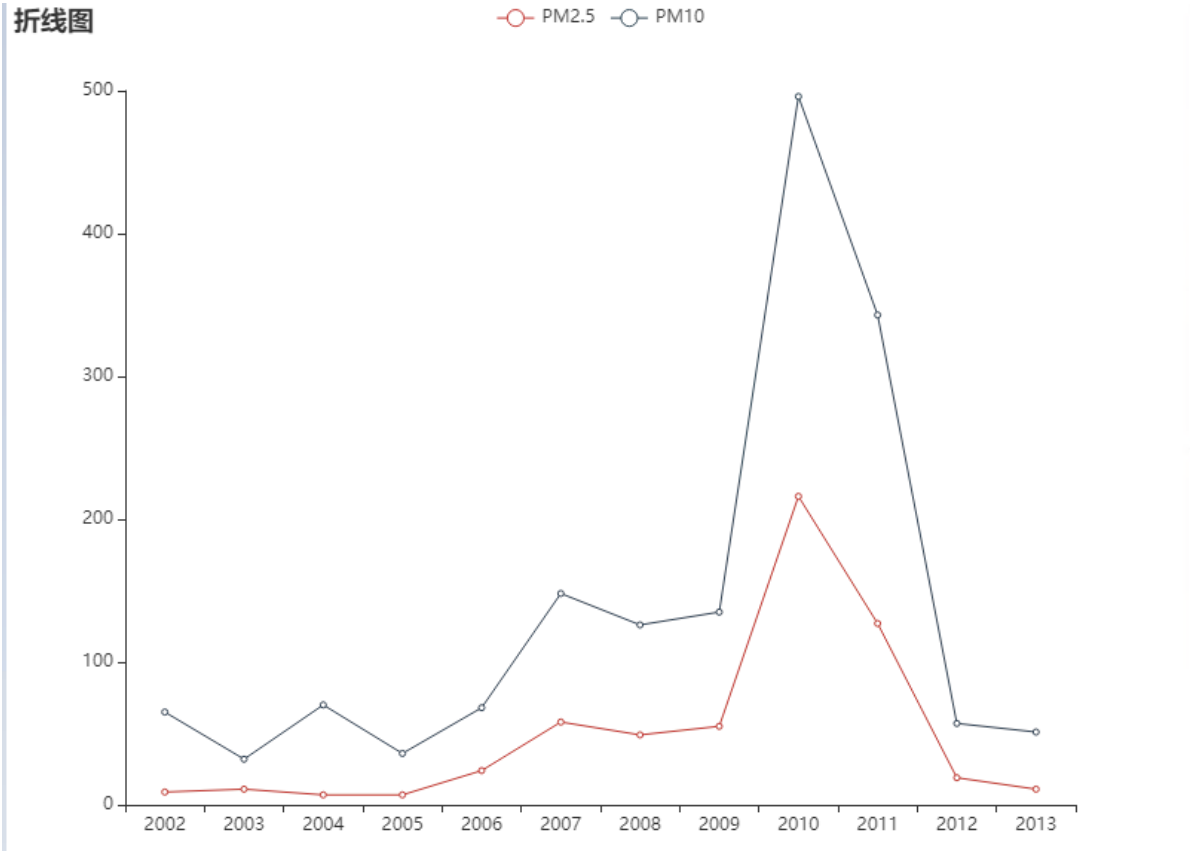
以空气质量评价数据作为示例。

years	AQI	PM2.5	PM10	CO	NO2	SO2
2002	55	9	56	0.46	18	6
2003	25	11	21	0.65	34	9
2004	56	7	63	0.3	14	5
2005	33	7	29	0.33	16	6
2006	42	24	44	0.76	40	16
2007	82	58	90	1.77	68	33
2008	74	49	77	1.46	48	27
2009	78	55	80	1.29	59	29
2010	267	216	280	4.8	108	64
2011	185	127	216	2.52	61	27
2012	39	19	38	0.57	31	15
2013	41	11	40	0.43	21	7

先将需要进行折线图的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】组件，点击【输入源】组件，填写数据集名称“bj”，勾选文件“bj.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行折线图操作。拖入【折线图】，将【输入源】和【折线图】相连接，在参数配置的X轴选择目标标签列，Y轴选择目标数值列，继而进行所需的样式配置。右键单击【折线图】，选择“运行该节点”。

查看日志。对【折线图】右击选择查看日志，即可查看样本数据的随时间的变化情况。



8.8.10 词云图

词云图作为一种分析热度的基本可视化图，在数据分析占据重要地位。其是一种文本出现频率较高的词以视觉化的展现，词云图会过滤掉大量低频低质的文本信息，让某个事物的重要性一目了然。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	词列、词频列
2	图片模板	JPG、PNG	当没有导入图片模板，词云图呈矩形白布

（2）输出

序号	名称	内容
1	日志	词云图展示

（3）参数

序号	分组	参数	说明
1	参数配置	词列	
2	参数配置	词频列	
3	样式配置	最大/小字体字号	
4	样式配置	是否使用图片颜色	
5	样式配置	缩放	

（4）示例

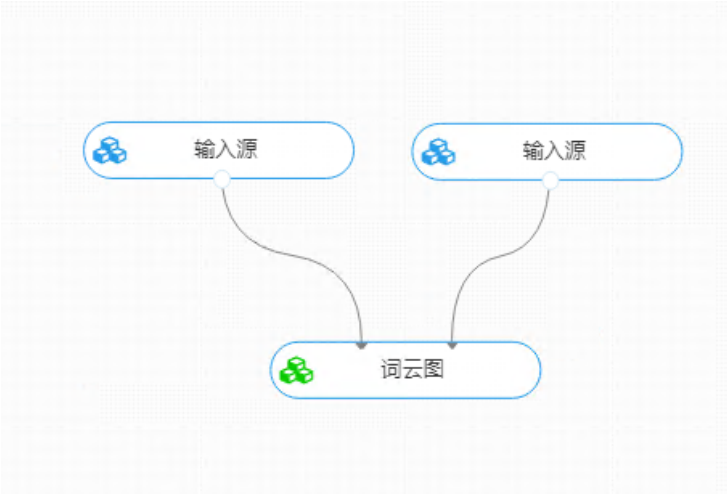
使用下列数据集进行示例展示。

	A	B	C
1	0	1	
2	生活资源	999	
3	供热管理	888	
4	供气质量	777	
5	生活用水	688	
6	一次供水	588	
7	交通运输	516	
8	城市交通	515	
9	环境保护	483	
10	房地产管	462	
11	城乡建设	449	
12	社会保障	429	
13	社会保障	407	
14	文体与教	406	
15	公共安全	406	
16	公交运输	386	
17	出租车运	385	
18	供热管理	375	
19	市容环卫	355	
20	自然资源	355	
21	粉尘污染	335	
22	噪声污染	324	
23	土地资源	304	
24	物业服务	304	
25	医疗卫生	284	
26	粉煤灰污	284	
27	占道	284	
28	供热发展	254	

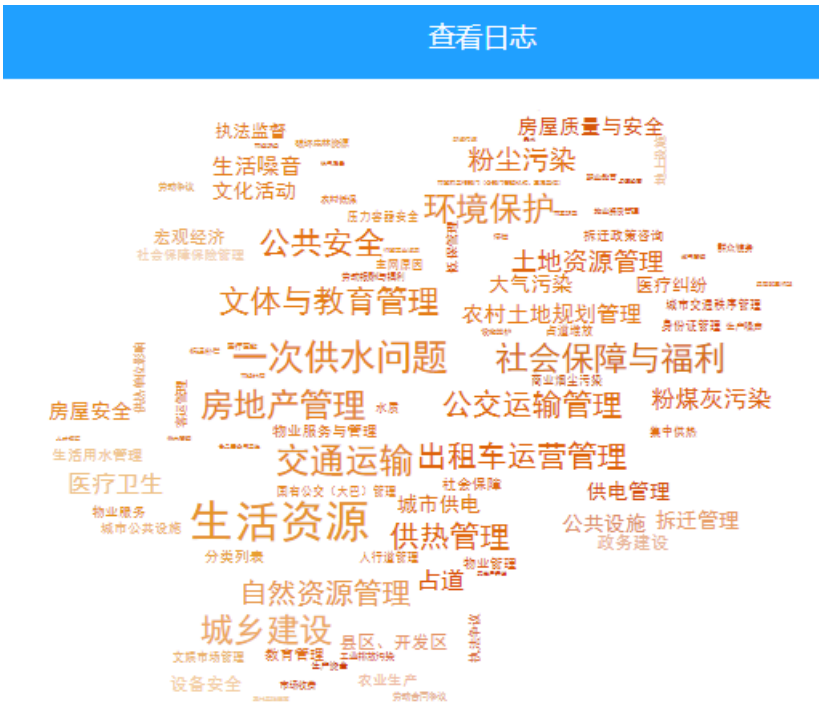
ciyun

先将需要进行词云图的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】组件，点击【输入源】组件，填写数据集名称“ciun”，勾选文件“ciyun.csv”；继而同样的操作将图片模板读入系统。

进行折线图操作。拖入【词云图】，将【输入源】和【词云图】相连接，在参数配置的词列选择数据集的文本列，词频列选择数值列，继而进行所需的样式配置。右键单击【词云图】，选择“运行该节点”。



查看日志。对【词云图】右击选择查看日志，即可查看样本的高频词情况。



8.8.11 漏斗图

漏斗图又叫倒三角图，漏斗图将数据呈现为几个阶段，每个阶段的数据都是整体的一部分，从一个阶段到另一个阶段数据自上而下逐渐下降，其有一定的逻辑顺序，且所有阶段的占比总计100%。

漏斗图适用于业务流程比较规范、周期长、环节多的单流程单向分析，通过漏斗各环节业务数据的比较能够直观地发现和说明问题所在的环节，进而做出决策。漏斗图用梯形面积表示某个环节业务量与上一个环节之间的差异。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	
2	数据量建议	不超过12条数据	

（2）输出

序号	名称	内容
1	日志	漏斗图展示

（3）参数

序号	分组	参数	说明
1	参数配置	X轴	标签列
2	参数配置	Y轴	数值列
3	样式配置	标签位置	当显示数值标签时才作用
4	样式配置	排序	
5	样式配置	显示数值标签	

（4）示例

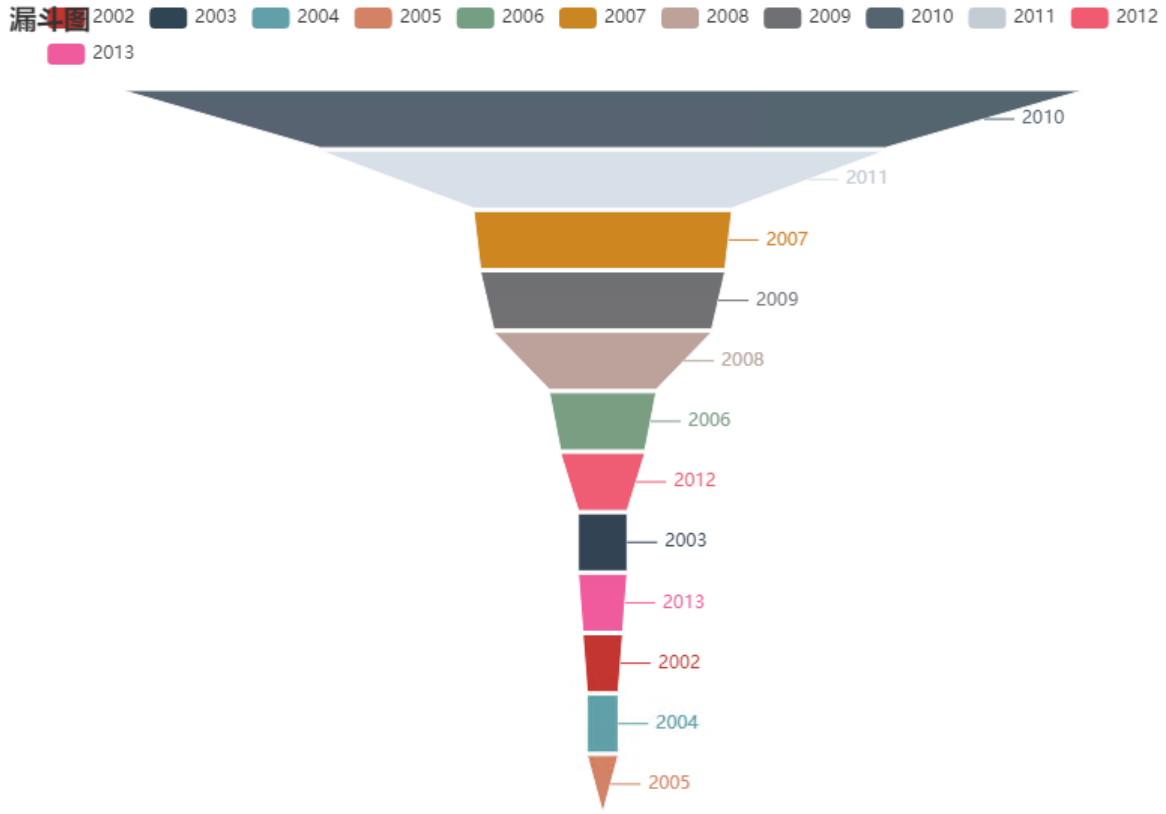
采用空气质量指标作为示例展示。

years	AQI	PM2.5	PM10	CO	NO2	SO2
2002	55	9	56	0.46	18	6
2003	25	11	21	0.65	34	9
2004	56	7	63	0.3	14	5
2005	33	7	29	0.33	16	6
2006	42	24	44	0.76	40	16
2007	82	58	90	1.77	68	33
2008	74	49	77	1.46	48	27
2009	78	55	80	1.29	59	29
2010	267	216	280	4.8	108	64
2011	185	127	216	2.52	61	27
2012	39	19	38	0.57	31	15
2013	41	11	40	0.43	21	7

先将需要进行漏斗图的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】组件，点击【输入源】组件，填写数据集名称“bj”，勾选文件“bj.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行漏斗图操作。拖入【漏斗图】，将【输入源】和【漏斗图】相连接，在参数配置的X轴选择目标标签列，Y轴选择目标数值列，继而进行所需的样式配置。右键单击【漏斗图】，选择“运行该节点”。

查看日志。对【漏斗图】右击选择查看日志，即可查看样本的漏斗图分布差异情况。



8.9 关联规则

关联规则（association rule）用来描述两个或多个事物之间的关联性，其通过一件或多件事物来预测其它事物，可以从大量数据中获取有价值数据之间的联系。

- 关联规则是形如 X => Y 的蕴含式，其中X 称为规则的前提，Y 称为规则的结果。
- 关联规则反映X中的项目出现时，Y中项目也跟着出现的规律。

8.9.1 灰色关联度分析法

对于两个系统之间的因素，其随时间或不同对象而变化的关联性大小的量度，称为关联度。在系统发展过程中，若两个因素变化的趋势具有一致性，即同步变化程度较高，即可谓二者关联程度较高；反之，则较低。因此，灰色关联分析方法，是根据因素之间发展趋势的相似或相异程度，亦即“灰色关联度”，作为衡量因素间关联程度的一种方法。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	

（2）输出

序号	名称	内容
1	日志	关联度热力图展示

（3）参数

序号	分组	参数	说明
1	参数配置	特征列	选择需要进行关联分析的特征列
2	参数配置	分辨系数	范围0~1，一般取0.5；相关系数越小，差异越大，分辨能力越强

（4）示例

以wine数据集作为示例展示，由于该数据集各特征数据存在量钢差异，在关联分析过程中需要对其进行标准化处理。

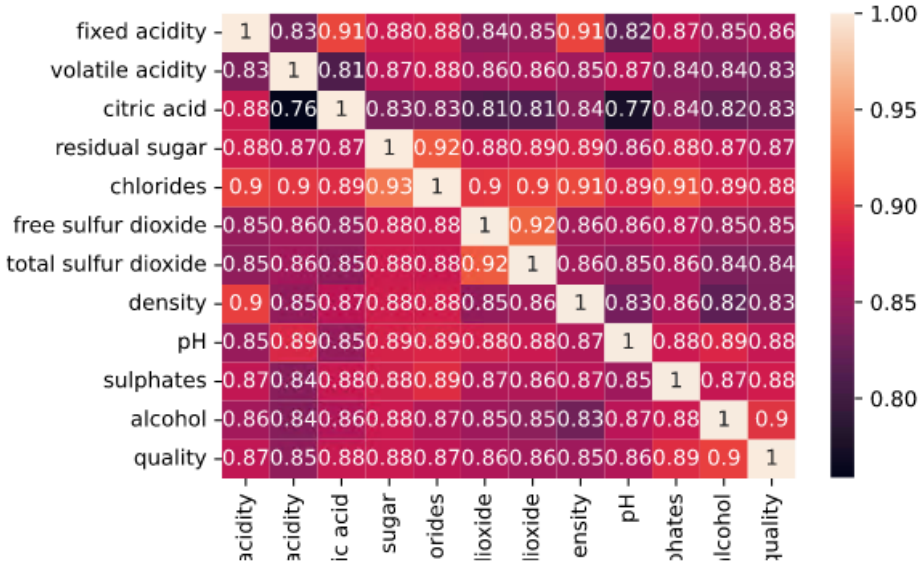
	A	B	C	D	E	F	G	H	I	J	K	L
175	7.4	0.62	0.05	1.9	0.068	24	42	0.9961	3.42	0.57	11.5	6
176	7.3	0.38	0.21	2	0.08	7	35	0.9961	3.33	0.47	9.5	5
177	6.9	0.5	0.04	1.5	0.085	19	49	0.9958	3.35	0.78	9.5	5
178	7.3	0.38	0.21	2	0.08	7	35	0.9961	3.33	0.47	9.5	5
179	7.5	0.52	0.42	2.3	0.087	8	38	0.9972	3.58	0.61	10.5	6
180	7	0.805	0	2.5	0.068	7	20	0.9969	3.48	0.56	9.6	5
181	8.8	0.61	0.14	2.4	0.067	10	42	0.9969	3.19	0.59	9.5	5
182	8.8	0.61	0.14	2.4	0.067	10	42	0.9969	3.19	0.59	9.5	5
183	8.9	0.61	0.49	2	0.27	23	110	0.9972	3.12	1.02	9.3	5
184	7.2	0.73	0.02	2.5	0.076	16	42	0.9972	3.44	0.52	9.3	5
185	6.8	0.61	0.2	1.8	0.077	11	65	0.9971	3.54	0.58	9.3	5
186	6.7	0.62	0.21	1.9	0.079	8	62	0.997	3.52	0.58	9.3	6
187	8.9	0.31	0.57	2	0.111	26	85	0.9971	3.26	0.53	9.7	5
188	7.4	0.39	0.48	2	0.082	14	67	0.9972	3.34	0.55	9.2	5
189	7.7	0.705	0.1	2.6	0.084	9	26	0.9976	3.39	0.49	9.7	5
190	7.9	0.5	0.33	2	0.084	15	143	0.9968	3.2	0.55	9.5	5
191	7.9	0.49	0.32	1.9	0.082	17	144	0.9968	3.2	0.55	9.5	5
192	8.2	0.5	0.35	2.9	0.077	21	127	0.9976	3.23	0.62	9.4	5
193	6.4	0.37	0.25	1.9	0.074	21	49	0.9974	3.57	0.62	9.8	6
194	6.8	0.63	0.12	3.8	0.099	16	126	0.9969	3.28	0.61	9.5	5
195	7.6	0.55	0.21	2.2	0.071	7	28	0.9964	3.28	0.55	9.7	5
196	7.6	0.55	0.21	2.2	0.071	7	28	0.9964	3.28	0.55	9.7	5
197	7.8	0.59	0.33	2	0.074	24	120	0.9968	3.25	0.54	9.4	5
198	7.3	0.58	0.3	2.4	0.074	15	55	0.9968	3.46	0.59	10.2	5
199	11.5	0.3	0.6	2	0.067	12	27	0.9981	3.11	0.97	10.1	6
200	5.4	0.835	0.08	1.2	0.046	13	93	0.9924	3.57	0.85	13	7
201	6.9	1.09	0.06	2.1	0.061	12	31	0.9948	3.51	0.43	11.4	4
202	9.6	0.32	0.47	1.4	0.056	9	24	0.99695	3.22	0.82	10.3	7

wine

先将需要进行关联分析法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】组件，点击【输入源】组件，填写数据集名称“wine”，勾选文件“wine.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行关联分析法操作。拖入【关联分析法】，将【输入源】和【关联分析法】相连接，在参数配置的特征列选择需要进行关联分析的特征，样式配置中对是否进行数据标准化进行选择。右键单击【漏斗图】，选择“运行该节点”。

查看日志。对【关联分析法】右击选择查看日志，即可查看样本的各个特征之间的关联度热力图展示。



8.9.2 Apriori

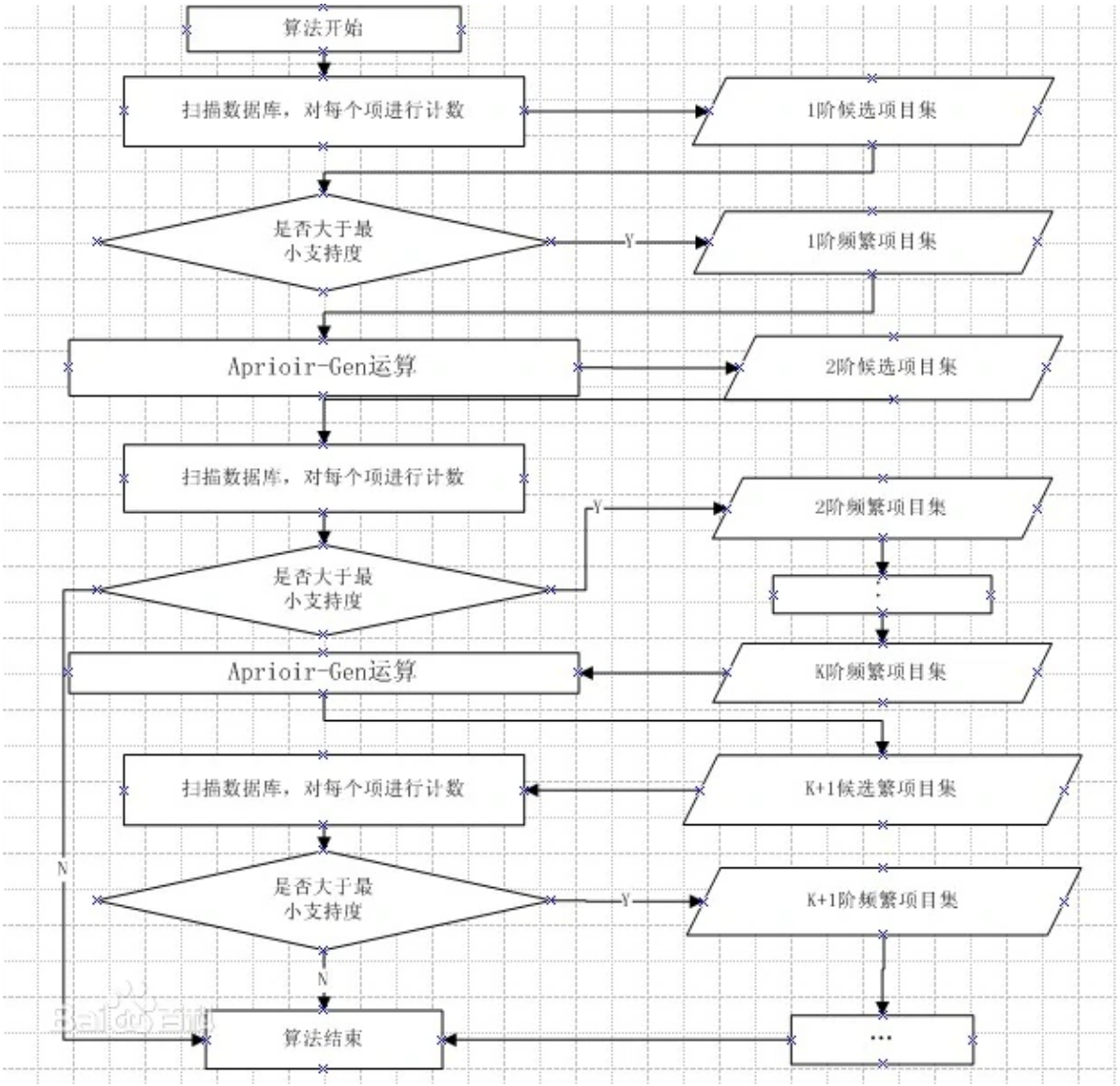
（1）原理

Apriori算法的基本思想是通过对数据库的多次扫描来计算项集的支持度，发现所有的频繁项集从而生成关联规则。Apriori算法对数据集进行多次扫描，第一次扫描得到频繁1-项集L1，第k（k>1）次扫描首先利用第（k-1）次扫描的结果L(k-1)来产生候选k-项集的集合Ck，然后在扫描过程中确定Ck中元素的支持度，最后在每一次扫描结束时计算频繁k-项集的集合Lk，算法在当候选k-项集的集合Ck为空时结束。该算法流程主要分为连接与剪枝两个步骤：

- 连接步。为找到Lk（k≥2），通过L(k-1）与自身连接产生候选k-项集的集合Ck。自身连接时，两个项集对应的想按从小到大顺序排列好，当前除最后一项外的其他项都相等时，两个项集可连接，连接产生的结果为（l1[1], l1[2], ..., l1[k-1], l2[k-2]）。
- 剪枝步。由Apriori算法的性质可知，频繁k-项集的任何子集必须是频繁项集。由连接步产生的集合Ck需进行验证，除去不满足支持度的非频繁k-项集。

Apriori算法已经被广泛的应用到商业、网络安全等各个领域。在消费市场价格分析中，它能够很快的求出各种产品之间的价格关系和它们之间的影响。通过数据挖掘，市场商人可以瞄准目标客户，采用个人股票行市、最新信息、特殊的市场推广活动或其他一些特殊的信息手段，从而极大地减少广告预算和增加收入。百货商场、超市和一些老字型大小的零售店也在进行数据挖掘，以便猜测这些年来顾客的消费

习惯。



（2）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	

（2）输出

序号	名称	内容
1	data_out	频繁项集结果

（3）参数

序号	分组	参数	说明
1	参数配置	最小支持度	表示项集在统计意义上的最低重要性，便于筛选出高支持度的项集
2	参数配置	最小置信度	表示关联规则最低可靠性

（4）示例

以商品购物情况数据集作为示例展示。Apriori关联规则中需要保证每一行数据也即是各记录中的商品是唯一的。

	0	1	2	3	4	5
1	Milk	Onion	Nutmeg	Kidney Beans	Eggs	Yogurt
2	Dill	Onion	Nutmeg	Kidney Beans	Eggs	Yogurt
3	Milk	Apple	Kidney Beans	Eggs	Dill	Yogurt
4	Milk	Unicorn	Corn	Kidney Beans	Yogurt	Eggs
5	Corn	Onion	Eggs	Kidney Beans	Ice cream	
6	Milk	Unicorn	Corn	Yogurt	Eggs	
7	Milk	Unicorn	Eggs	Kidney Beans	Yogurt	
8	Corn	Onion	Eggs	Kidney Beans	Yogurt	
9	Corn	Yogurt	Onion	Kidney Beans	Ice cream	Eggs
10	Milk	Unicorn	Corn	Yogurt	Eggs	

先将需要进行Apriori算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】组件，点击【输入源】组件，填写数据集名称“purchase1”，勾选文件“商品数据集1.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行Apriori算法操作。拖入【Apriori】，将【输入源】和【Apriori】相连接，在其他配置中设置最大项目数、最小支持数以及勾选相应展示内容。右键单击【Apriori】，选择“运行该节点”。

查看日志。对【Apriori】右击选择查看日志，即可查看算法对数据的扫描过程以及关联规则的结果。

8.9.3 FP-Growth

（1）原理

FP-growth(Frequent Pattern Tree, 频繁模式树),是韩家炜老师提出的挖掘频繁项集的方法，是将数据集存储在一个特定的称作FP树的结构之后发现频繁项集或频繁项对，即常在一块出现的元素项的集合FP树。

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。

FP-growth算法比Apriori算法效率更高，在整个算法执行过程中，只需遍历数据集2次，就能够完成频繁模式发现，其发现频繁项集的基本过程如下：

- 构建FP树
- 从FP树中挖掘频繁项集

（2）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	

（3）输出

序号	名称	内容
1	data_out1数据	频繁项集结果
2	data_out2数据	关联规则结果

（4）参数

序号	分组	参数	说明
1	参数配置	最小支持度	表示项集在统计意义上的最低重要性
2	参数配置	最小置信度	表示关联规则最低可靠性

（5）示例

以商品购物情况数据集作为示例展示。

	A	B	C	D	E	F
1	0	1	2	3	4	5
2	Milk	Onion	Nutmeg	Kidney Beans	Eggs	Yogurt
3	Dill	Onion	Nutmeg	Kidney Beans	Eggs	Yogurt
4	Milk	Apple	Kidney Be	Eggs		
5	Milk	Unicorn	Corn	Kidney Beans	Yogurt	
6	Corn	Onion	Onion	Kidney Beans	Ice cream	Eggs
7	Milk	Unicorn	Corn	Yogurt	Eggs	
8	Milk	Unicorn	Eggs	Kidney Beans	Yogurt	
9	Corn	Onion	Onion	Kidney Beans	Yogurt	Eggs
10	Corn	Yogurt	Onion	Kidney Beans	Ice cream	Eggs
11	Milk	Unicorn	Corn	Yogurt	Eggs	
12						

先将需要进行FP-growth算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】组件，点击【输入源】组件，填写数据集名称“pur”，勾选文件“pur.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行FP-growth算法操作。拖入【FP-growth】，将【输入源】和【FP-growth】相连接，在其他配置中设置最大项目数、最小支持数以及勾选相应展示内容。右键单击【FP-growth】，选择“运行该节点”。

查看日志。对【FP-growth】右击选择查看日志，即可查看样本的关联规则结果与频繁项集结果。

8.9.4 Relim算法

FP-growth算法是当前挖掘频繁项集算法中速度最快，应用最广，并且不需要候选项集的一种频繁项集挖掘算法，但是FP-growth也存在着算法结构复杂和空间利用率低等缺点。Relim算法是在FP-growth算法的基础上提出的一种新的不需要候选项集的频繁项集挖掘算法。它具有算法结构简单，空间利用率高，易于实现等显著优点。

Relim算法的主要思想和FP-growth相似，也是基于递归搜索(Recursive Exploration)，但是和FP-growth不同的是：Relim算法在运行时不必创建频繁模式树，而是通过建立一个事务链表组(transaction lists)来找出所有频繁项集。

（1）输入

序号	条件	要求	说明
1	载入文件格式	csv文件	

（2）输出

序号	名称	内容
1	data_out数据	关联规则相关信息输出

（3）参数

序号	分组	参数	说明
1	参数配置	特征列	
2	其他配置	最大项目数	每个项目集的最大项目数
3	其他配置	最小支持数	min_sup描述了关联规则的最低重要程度，此处为百分数
4	其他配置	最小项目数	每个项目集的最小项目数
5	其他配置	评估指标的阈值	评估指标的阈值（默认值：10%
6	其他配置	展示内容	选择最终与项目集一起输出的值，可多选

展示内容：

展示内容	说明
support_itemset_absolute（绝对支持度）	数据集中包含项集A的事物数。
support_itemset_relative（相对支持度）	项集A的绝对支持度与数据集事物总数的比值。
support_itemset_relative_pct（相对支持度百分比）	
confitence（置信度）	同时包含A项和B项的交易数与包含A项的交易数之比
confidence_pct（置信度百分数）	
lift（提升度）	反映了关联规则中的A项与B项的相关性。 其中提升度>1且越高表明正相关性越高，提升度<1且越低表明负相关性越高，提升度=1表明没有相关性；当负值时则表明物品之间具有相互排斥的作用。
lift_pct（提升度百分比）	
support_bodyset_absolute	X => Y中后项绝对支持度
support_bodyset_relative	X => Y中后项相对支持度
support_bodyset_relative_pct	X => Y中后项相对支持度百分数
support_headitem_absolute	X => Y中前项绝对支持度
support_headitem_relative	X => Y中前项相对支持度
support_headitem_relative_pct	X => Y中前项相对支持度百分数
evaluction	项目集评估度量值
Q	支持空集

（4）示例

以商品购物情况数据集作为示例展示。

	A	B	C	D	E	F
1	0	1	2	3	4	5
2	Milk	Onion	Nutmeg	Kidney Beans	Eggs	Yogurt
3	Dill	Onion	Nutmeg	Kidney Beans	Eggs	Yogurt
4	Milk	Apple	Kidney Be	Eggs		
5	Milk	Unicorn	Corn	Kidney Beans	Yogurt	
6	Corn	Onion	Onion	Kidney Beans	Ice cream	Eggs
7	Milk	Unicorn	Corn	Yogurt	Eggs	
8	Milk	Unicorn	Eggs	Kidney Beans	Yogurt	
9	Corn	Onion	Onion	Kidney Beans	Yogurt	Eggs
10	Corn	Yogurt	Onion	Kidney Beans	Ice cream	Eggs
11	Milk	Unicorn	Corn	Yogurt	Eggs	
12						

先将需要进行Relim算法的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】组件，点击【输入源】组件，填写数据集名称“pur”，勾选文件“pur.csv”，右键单击【输入源】算法，选择“运行该节点”。

进行Relim算法操作。拖入【Relim算法】，将【输入源】和【Relim算法】相连接，在参数配置的特征列选择需要进行关联规则分析的特征，其他配置中设置最大项目数、最小支持数以及勾选相应展示内容。右键单击【Relim算法】，选择“运行该节点”。

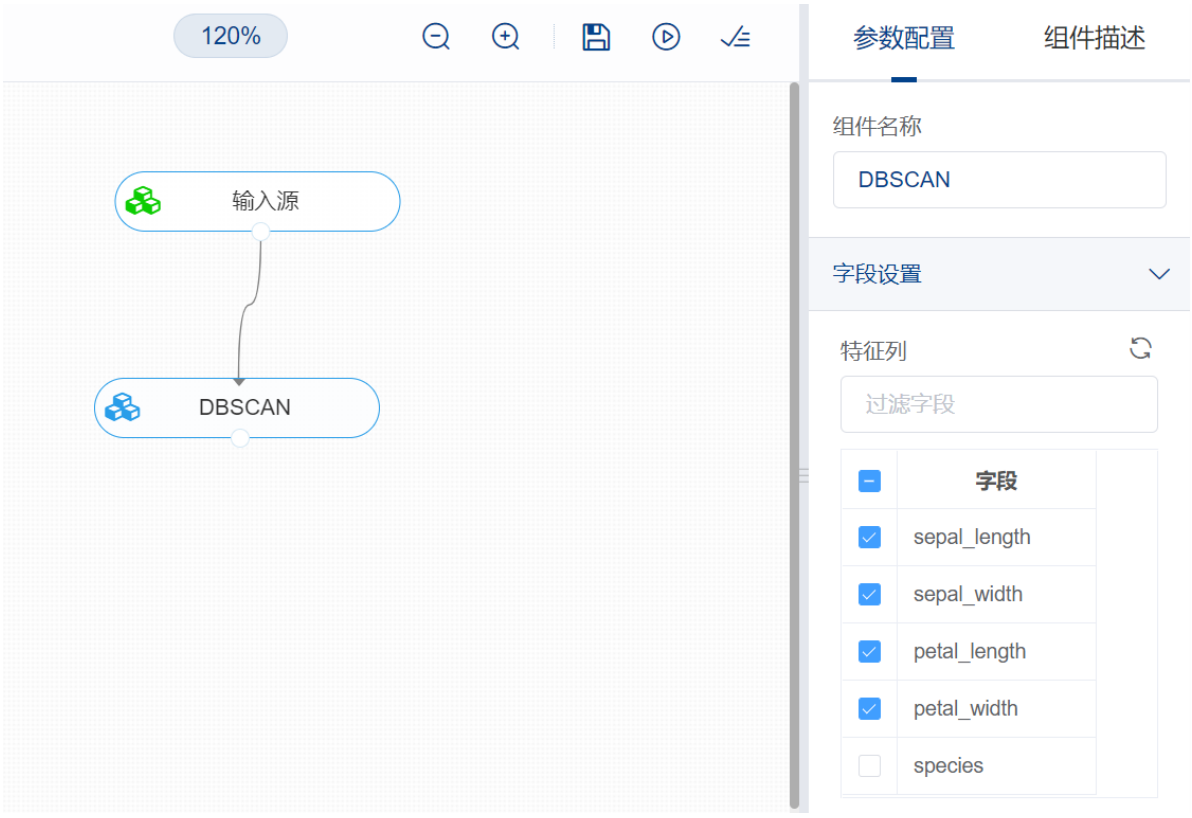
查看日志。对【Relim】右击选择查看日志，即可查看样本的各个项集的关联规则信息。

	A	B	C	D	E
1	sepal_length	sepal_width	petal_length	petal_width	species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa
6	5	3.6	1.4	0.2	setosa
7	5.4	3.9	1.7	0.4	setosa
8	4.6	3.4	1.4	0.3	setosa
9	5	3.4	1.5	0.2	setosa
10	4.4	2.9	1.4	0.2	setosa
11	4.9	3.1	1.5	0.1	setosa
12	5.4	3.7	1.5	0.2	setosa
13	4.8	3.4	1.6	0.2	setosa

iris

（1）首先将需要进行DBSCAN密度算法聚类的数据集读入系统，这里要用到【输入源】组件。拖入【输入源】算法，点击【输入源】算法，填写数据集名称“iris”，勾选文件“iris.csv”，右键单击【输入源】算法，选择“运行该节点”。

（2）开始进行DBSCAN密度算法聚类。拖入【DBSCAN】算法，将【输入源】算法和【DBSCAN】算法相连接，在“字段设置”的“特征”中勾选“sepal_length”，“sepal_width”，“petal_length”，“petal_width”字段。



（3）点击参数设置，领域内最小样本数目设置为10，领域大小设置为0.5，其他参数保持默认，右键单击【DBSCAN密度算法】算法，选择“运行该节点”。

序号	参数名称	数值	原因
1	领域内最小数目个数	5	领域内最小样本数目过大，则核心对象会过少，此时簇内部分本来是一类的样本可能会被标为噪音点，类别数也会变多。反之领域内最小样本数目过小的话，则会产生大量的核心对象，可能会导致类别数过少。
2	领域半径	0.2	领域大小对分类结果影响非常大，若参数设置过小，大部分数据不能聚类；若参数设置过大，多个簇和大部分对象会归并到同一个簇中。

（4）在日志中可以查看各个聚类分群的个数与比例。初始时是由一个任意未被访问的点开始，然后探索这个点的ε-邻域，如果ε-邻域里有足够的点，则建立一个新的聚类，否则这个点被标签为噪音。注意这个点之后可能被发现在其它点的ε-邻域里，而该ε-邻域可能有足够的点，届时这个点会被加入该聚类中。由于初始点的确定是随机的，最终得到的结果也是不同的，但是参数相同的情况下大体相似。对【DBSCAN】算法右击，点击“查看日志”。

查看日志

DBSCAN算法结果

聚类信息

	cluster_id	total_number	percent
1	1	133	88.667%
2	2	10	6.667%
3	3	7	4.667%

聚类各类中心图