2019

智文

V3.20.1

目录

第1章　系统简介

“智文”帮助用户一站式开展数据挖掘和文本挖掘任务。既可以用于科研支撑，帮助研究人员自主和高效地分析数据，挖掘知识；也可以用于教学，培养相关专业的学生数据分析、自然语言处理的能力。使学生形成大数据思维，为科学的量化管理培养基于数据的管理和决策分析技能。

1.1　系统环境和安装

“智文”的设计初衷是将数据挖掘算法模块化，提高数据分析的直观性和可操作性。用拖拽功能节点和制作工作流程的方式，将高度可重复利用的算法组合起来，满足高等院校、科研机构应对各种复杂的数据挖掘，自然语言学习任务的需求。

该系统支持 64 位的 windows10 操作系统和 macOS 操作系统，点击安装文件后按照提示可以简单地完成安装。安装前请关闭主机的杀毒软件，由于用户主机配置不一，在一些主机上安装所用时间可能较长，请耐心等待。

* 安装成功后，运行“智文”，使用商智通账号登录。从“数据”功能节点组中的“SQL 表格”功能节点拖拽至流程画板上。双击进入“SQL 表格”，如果看到窗口右下角提示“想要使用此节点，请安装驱动（backend）。”的提示，请运行与“智文”已同安装的“Ziwen Command Prompt”。输入“pip install mysql-connector-python”，按回车键开始下载安装后端软件。安装过程要保证电脑与外网的网络连接。

1.2　“智文”特点

“智文”把强大算法和直观界面结合起来：算法覆盖各种使用场景；图形界面依靠拖拽式操作。“智文”从数据源头获得数据，然后对数据进行预处理，可视化，建模，评估，导出结论，给用户流畅的数据分析体验。

该系统的特点有如下几点：

1. 完备的算法库

智文数十种各类算法可满足用户所有数据预处理和建模任务。这些算法中既包括常见的神经网络，支持向量机，k 最近邻域等一系列针对一般数据挖掘任务的监督与非监督学习算法，也提供了基于自然语言处理（NLP）技术的各项针对文本分析和挖掘任务的算法，如中文外文分词，词性标注，词频标注，文本情感分析，相似度计算等。

2. 丰富的数据资源，强大的接入能力

可以从各种数据源头轻松读取数据。可读取本地 csv，excel，tab 等格式数据文件，可直接连通数据库，可以利用 opendata API 从商智通数据空间直接接入数据，还可以连接到最多样的实时互联网大数据，包括：互联网金融行业实时数据集；旅游行业实时数据集；汽车行业实时数据集；餐饮行业实时数据集；房地产行业实时数据集等。

3. 满足多种可视化要求

智文提供各种独立或算法一同使用的可视化工具超过 20 种，其中包括：散点图，轮廓图，热力图，毕达哥拉斯树，词云等等。

4. 可靠的模型评估

多种形式的交叉验证（k 折交叉验证，留一法等），自助预测检验等功能可以可靠评估模型的质量，帮助修改模型达到最佳挖掘分析表现。

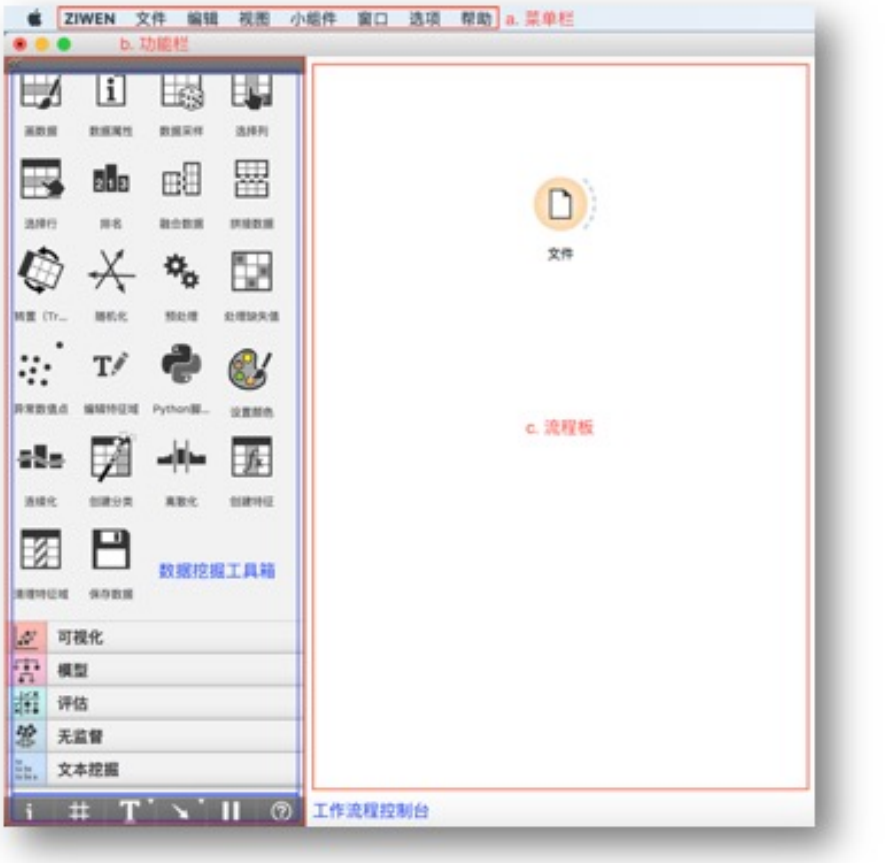
5. 极为直观的操作和图形界面

“智文”将以上提到的数据接入，数据处理，构造模型，可视化和模型评估中的所有算法和操作包装在一个个功能节点中，每个功能节点用来执行特定的具体功能，具有不同的输入输出特性。用户用拖拽的方式添加各个功能节点，根据它们的输入输出特性将它们合理地拖拽连接，就构成了一套类似工厂车间生产线的工作流程。

6. 轻松移植和扩展能力

可以向工作流程和模型添加注释，调试好的流程可直接被存储，导出，完整移植到其他机器。一个项目的工作流程可以轻松扩展到包含几十甚至上百个功能节点，可以完成最为复杂和专业的挖掘任务。

1.3 界面功能简介



运行“智文”后，系统将要求用户使用商智通账号登录。登录成功后用户即可进入主界面。用户将在主页面上完成文本挖掘工作流程全部的设计和搭建工作。主界面分为三部分：

：

1. 菜单栏

   提供一切众多文本挖掘具体工序以外的功能，如保存/读取工作流程，浏览帮助案例等。

2. 功能栏

   展示并提供全部文本挖掘功能。功能栏可分为两部分：占据绝大部分空间的“数据挖掘工具箱”，和功能栏最下方的“工作流程控制台”。智文为用户提供六大数据挖掘工具箱，近百个功能组件，帮助用户应对数据挖掘工作中的几乎所有数据处理需求。“工作流程控制台”则负责对流程板上的数据挖掘工作流程进行控制，检查，标注，暂停等操作。用户可以根据需求点击功能栏左上角箭头标志，展开或收起功能栏。

3. 流程板

   用户在流程板上画出整个数据挖掘工作流程，完成数据挖掘任务。

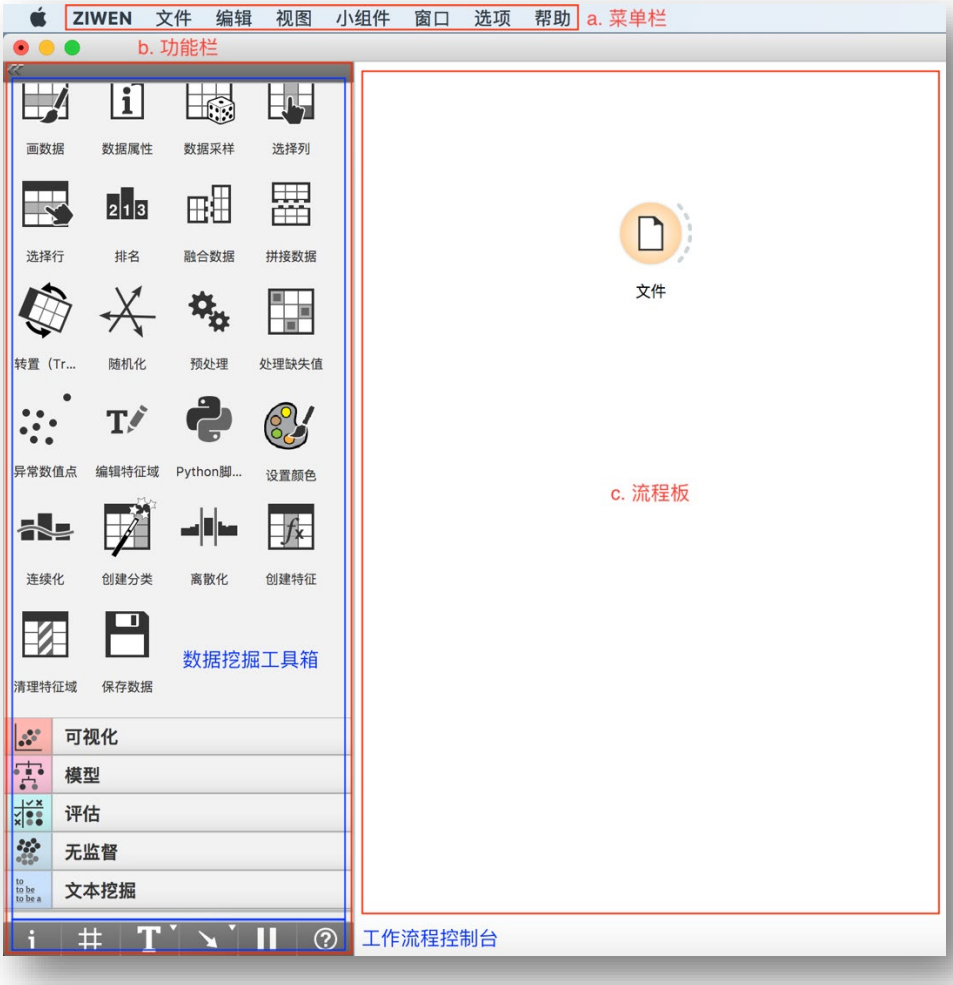
第2章　功能节点

运行智文平台后您将看到的是平台的整个界面。此界面是智文唯一的主界面，用户将在这个主页面上完成文本挖掘工作流程全部的设计和搭建工作。界面分为三部分：

a） 菜单栏：提供一切众多文本挖掘具体工序以外的功能，如保存/读取工作流程，浏览帮助案例等。

b） 功能栏：展示并提供全部文本挖掘功能。功能栏可分为两部分：占据绝大部分空间的“数据挖掘工具箱”，和功能栏最下方的“工作流程控制台”。智文为用户提供六大数据挖掘工具箱，近百个功能组件，帮助用户应对数据挖掘工作中的几乎所有数据处理需求。“工作流程控制台”则负责对流程板上的数据挖掘工作流程进行控制，检查，标注，暂停等操作。用户可以根据需求点击功能栏左上角箭头标志，展开或收起功能栏。

c） 流程板：用户在流程板上画出整个数据挖掘工作流程，完成数据挖掘任务。



2.1　“数据”功能节点组

- 文件：从输入文件或网络上读取数据，输出一个数据表格。
- 数据集：从商智通数据空间载入数据集。

- SQL 表格：从 SQL 数据库读取数据。
- 数据表格：在表格中浏览数据集。
- 画出数据：在平面上画出数据点，以此生成数据。
- 数据信息：展示数据集的基本信息，例如变量的个数和类型，数据集的行数。
- 数据取样：从输入端数据集中随机抽取一个数据点子集。
- 选择列：从数据集中选择列，分配列的角色（特征变量，目标变量或者元变量）。
- 选择行：根据变量的取值从数据中选出满足条件行。
- 排名：根据特征变量的重要性对它们进行排名和过滤。
- 相关系数：计算所有特征变量两两之间的相关系数。
- 合并数据：根据选中的特征变量的取舍，合并数据集。
- 拼接数据：把两个或以上的数据集前后拼接，依次附加成一个数据集。
- 根据数据序号选择：根据序号从数据子集里找出对应的数据实例。
- 转置：把数据表格转置。
- 随机化：对数据表格中的特征变量，目标变量，元变量进行随机化。
- 预处理：构造一个数据预处理流程。
- 转换：转化数据表格。
- 插补缺失值：对数据表格中的缺失值进行插值补值。
- 离群值：离群值识别。
- 编辑域：重命名变量，编辑分类和变量注释。
- Python 脚本：编辑和运行 Python 脚本，处理输入的数据和模型。
- 着色：为变量设置图例中的颜色。
- 连续化：把分类类型的特征变量转型成数值类型，并可以选择将转型后的数值进行归一化。
- 创建分类结果变量：通过字符串特征变量，创建分类结果变量。
- 离散化：把数值类变量离散化。
- 构造特征变量：使用输入数据集里现成的特征变量，构造新的特征变量（数据列）。
- 特征变量统计：展示特征变量的基本统计。
- 邻域：根据引用计算数据中的最近邻域。
- 清理数据域：从数据集中移除多余的数值和特征变量，对取值排序。
- 保存数据：将数据保存至文件。

2.1.1 文件

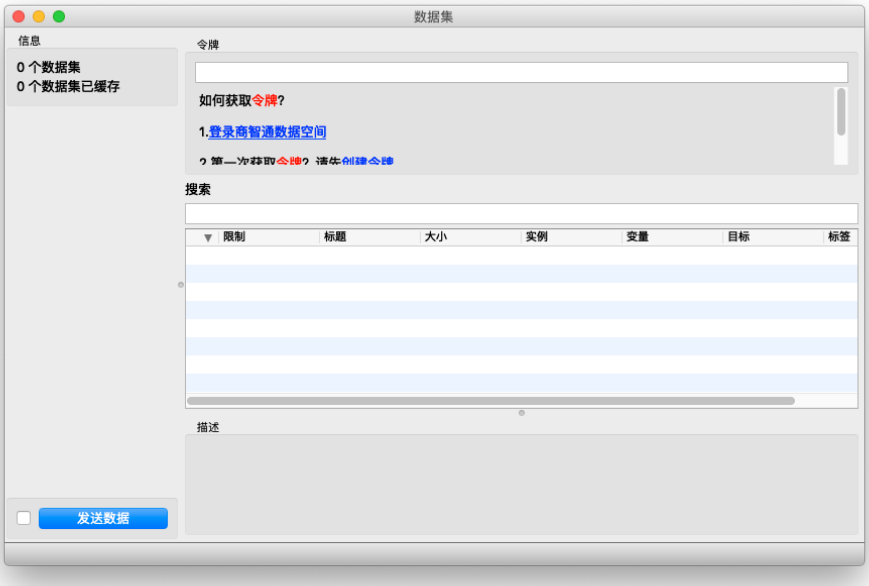
从输入文件或网络上读取数据，输出一个数据表格。

基本功能：

- 设置“文件名”或“URL”，从本地或网络服务器上读取一个数据文件；
- 通过“信息”获得数据文件的基本信息，检查载入是否符合预期；
- 在“列”设置中，设置数据集的域，其中包括每一列的类型和分析角色；
- 点击“浏览自带数据集”可以选择载入智文自带的经典数据集，适合初学或探索式操作；
- 功能节点将自动“应用”：将设置好的数据输送至下游。

2.1.2 数据集

从商智通数据空间载入数据集。

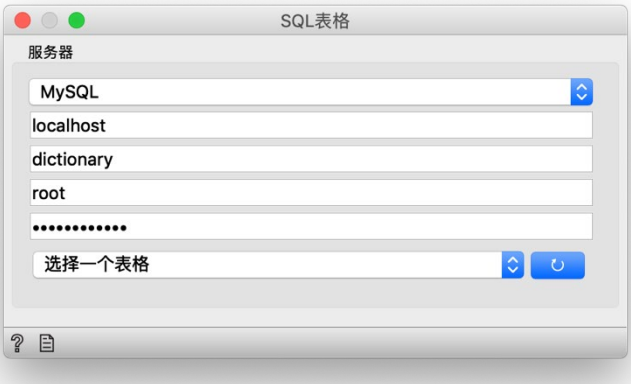


基本功能：

- 根据指示从商智通数据空间获得令牌，将口令输入后，功能节点将自动搜索用户在商智通数据空间上的数据集；
- 在“搜索”栏中，可以查看数据文件的元属性，包括：标题，大小，变量个数等；
- 选中一个数据集后，选择“发送数据”或设置自动发送，数据集将流入下游的分析流程中。

2.1.3 SQL 表格

从 SQL 数据库读取数据。



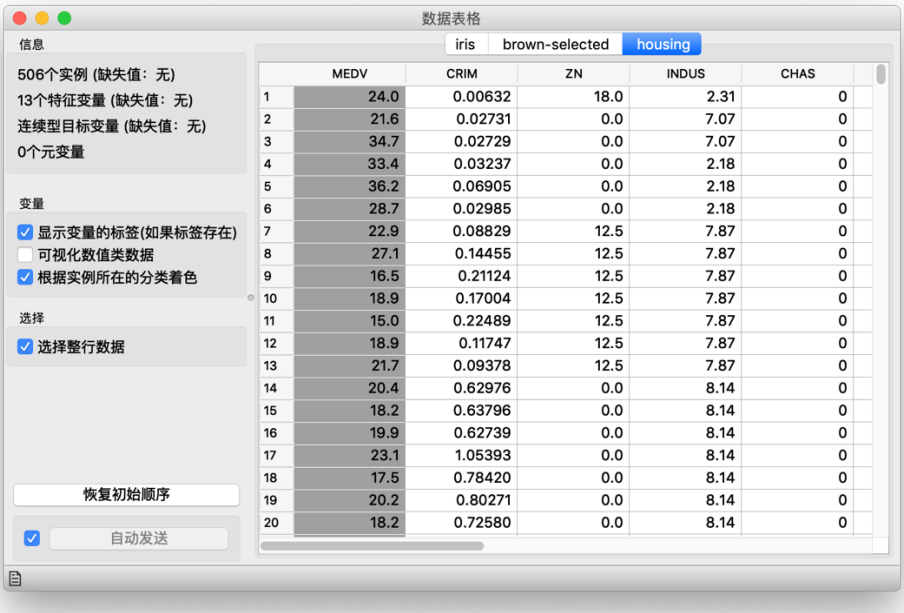
基本功能：

- 可接入“MySQL”，“PostgreSQL”或“SQL Server”数据库服务器，在下拉菜单内选择；
- 根据提示，用户将数据库的服务器位置，数据库名称，用户名和密码输入；
- 在“选择一个表格”中可以看到接入数据库中的表格，选中的表格将自动流入下游的分

析流程中。

2.1.4 数据表格

在表格中浏览数据集。



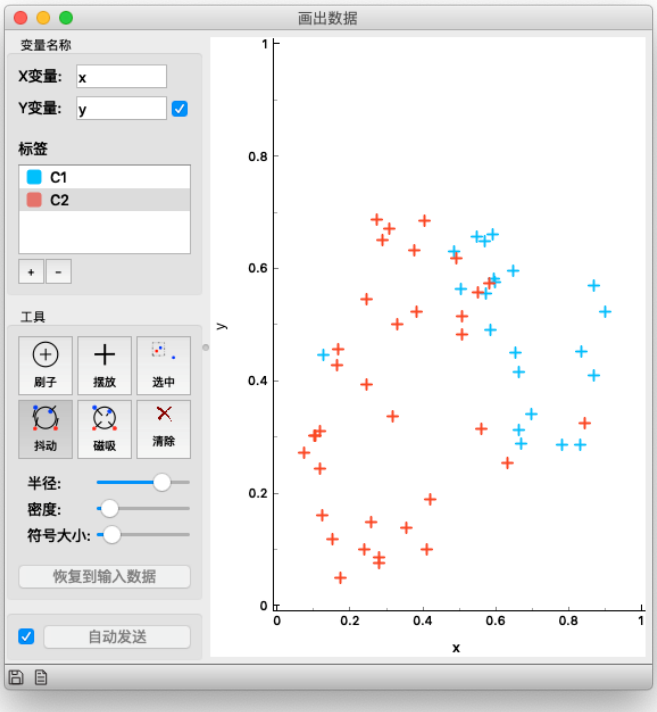
	MEDV	CRIM	ZN	INDUS	CHAS
1	24.0	0.00632	18.0	2.31	0
2	21.6	0.02731	0.0	7.07	0
3	34.7	0.02729	0.0	7.07	0
4	33.4	0.03237	0.0	2.18	0
5	36.2	0.06905	0.0	2.18	0
6	28.7	0.02985	0.0	2.18	0
7	22.9	0.08829	12.5	7.87	0
8	27.1	0.14455	12.5	7.87	0
9	16.5	0.21124	12.5	7.87	0
10	18.9	0.17004	12.5	7.87	0
11	15.0	0.22489	12.5	7.87	0
12	18.9	0.11747	12.5	7.87	0
13	21.7	0.09378	12.5	7.87	0
14	20.4	0.62976	0.0	8.14	0
15	18.2	0.63796	0.0	8.14	0
16	19.9	0.62739	0.0	8.14	0
17	23.1	1.05393	0.0	8.14	0
18	17.5	0.78420	0.0	8.14	0
19	20.2	0.80271	0.0	8.14	0
20	18.2	0.72580	0.0	8.14	0

基本功能：

- 左侧“信息”栏展示数据的元属性；
- 一个“数据表格”功能节点可以接入多个数据源，右侧表格上方可选择要查看的表格名称；
- 左侧“变量”可以勾选多个涉及数据表格显示的选项；
- 输出端可以输出选中的数据实例，在左侧的“选择”设置中，用户可选择是否输出整列数据。

2.1.5 画出数据

在平面上画出数据点，以此生成数据。



基本功能：

- 在“变量名称”中设置两个数轴的变量名称；
- 在“标签”中设置数据的分类，用不同颜色指代不同的分类；
- “工具”提供多种工具和配套的设置选项，用户可以根据情况使用不同的攻击在右侧的平面上生成数据实例点。

2.1.6 数据信息

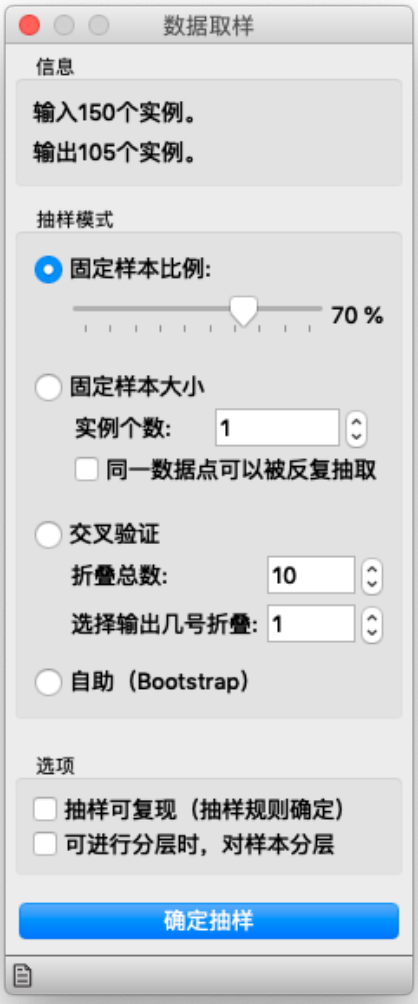
展示数据集的基本信息，例如变量的个数和类型，数据集的行数。

基本功能：

- 显示数据集的基本信息，其中包括：数据集名称，数据集行数和列数，数据集中三种变量的个数等。

2.1.7 数据取样

从输入端数据集中随机抽取一个数据点子集。



基本功能：

- “信息”显示输入和输出两端的实例个数；
- 用户在“抽样模式”设置具体的抽样模式，抽样方法包括：固定样本比例，固定抽样个数，交叉验证和自助 Bootstrap 四种方式；
- 用户可以进一步设置抽样方法，勾选“抽样可复现”则保证功能节点使用固定的随机种子。

2.1.8 选择列

从数据集中选择列，分配列的角色（特征变量，目标变量或者元变量）。

基本功能：

- 界面左侧“可选变量”中显示输入端所有的变量，用户可以在此功能节点中为这些变量分配不同的角色，并重新排列变量的先后顺序。

2.1.9 选择行

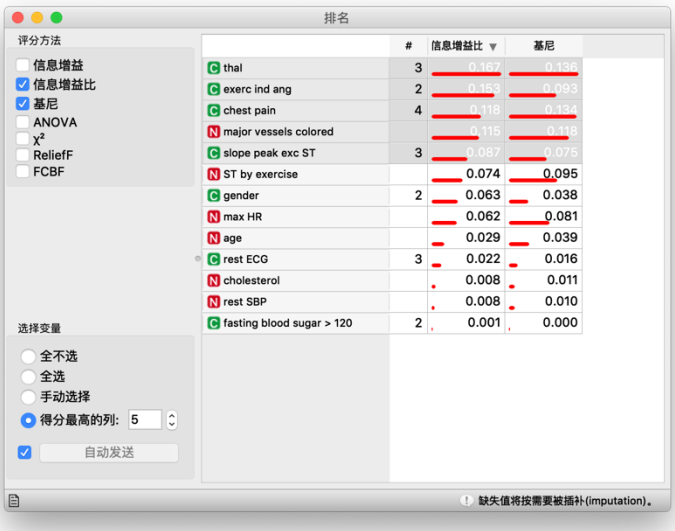
根据变量的取值从数据中选出满足条件行。

基本功能：

- 在“条件”中添加和修改针对变量的取值条件；
- 左下方的“数据”显示抽取数据的效果，满足条件的数据实例将被输出；
- 右下方的“清理”可以将没有在条件中出现过的列或分类取值过滤掉。

2.1.10 排名

根据特征变量的重要性对它们进行排名和过滤。



基本功能：

- 选择“评分方法”对变量的重要性进行排面；
- 评分方法中包括：信息增益，信息增益比，基尼，ANOVA，ReliefF，FCBF；
- 在左下角的“选择变量”中选择输出的列，用户可以根据评分和排名酌情选择输出的列。

2.1.11 相关系数

计算所有特征变量两两之间的相关系数。

基本功能：

- 可选择两种方法计算系数：皮尔逊积矩相关系数（Pearson）和斯皮尔曼等级相关系数（Spearman）。

2.1.12 合并数据

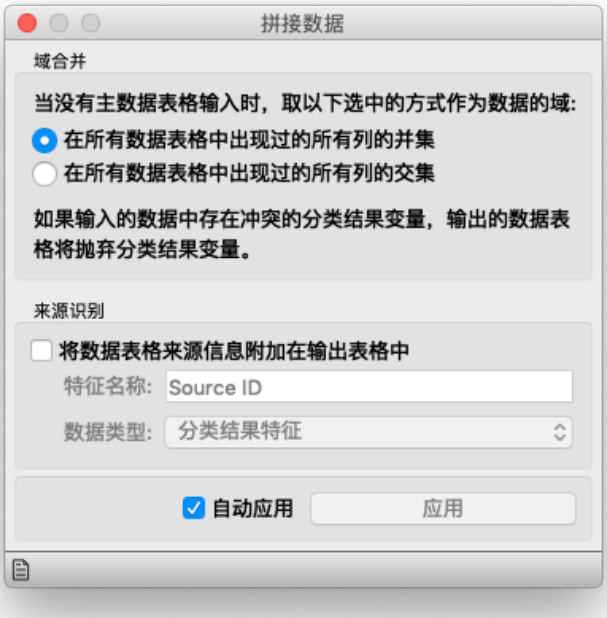
根据选中的特征变量的取舍，合并数据集。

基本功能：

- 通过“数据”和“额外的数据”观察输入数据的基本情况；
- 在“合并”中选择数据合并的方法。

2.1.13 拼接数据

把两个或以上的数据集前后拼接，依次附加成一个数据集。



基本功能：

- 在“域合并”中，设置没有主数据表格输入时的合并模式；
- 在“来源识别”中，可以选择将数据表格来源信息附加在输出的表格。

2.1.14 根据数据序号选择

根据序号从数据子集里找出对应的数据实例。

2.1.15 转置

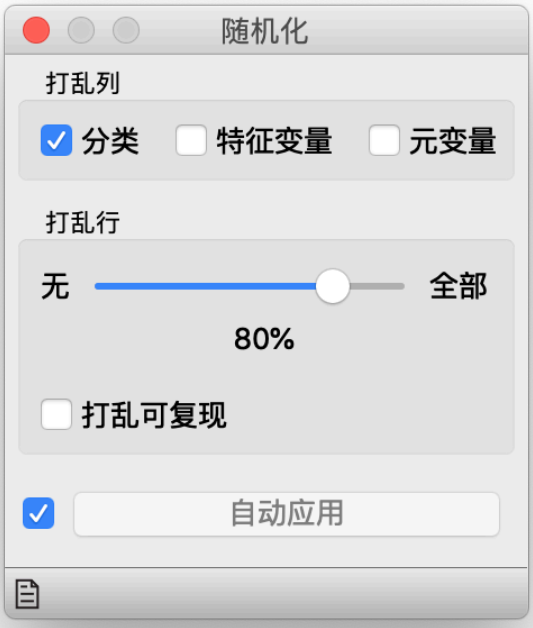
把数据表格转置。

基本功能：

- 选择“特征变量名称”后，功能节点将数据表格转置。

2.1.16 随机化

对数据表格中的特征变量，目标变量，元变量进行随机化。

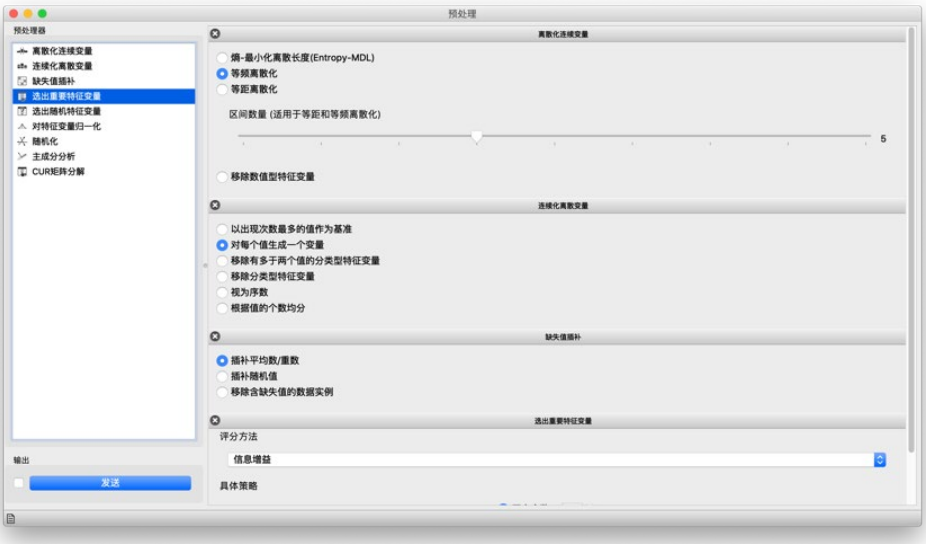


基本功能：

- 在“打乱列”中对不同变量进行随机化，变量分为三种角色：分类目标变量，特征变量，元变量；
- 在“打乱行”中对选定比例的数据进行随机化；
- 随机化使用的随机数种子可以在勾选“打乱可复现”后被固定，帮助用户更好进行分析。

2.1.17 预处理

构造一个数据预处理流程。

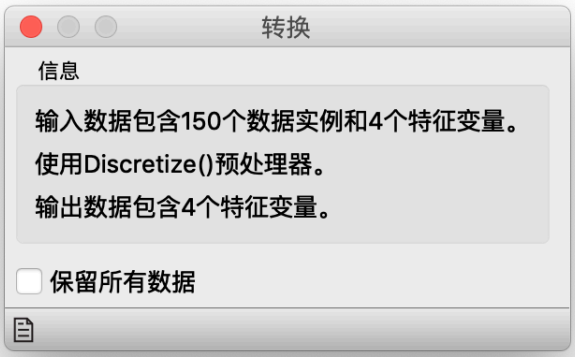


基本功能：

- 从左侧“预处理器”栏中把处理环节拖拽到右侧，形成预处理流程；
- 预处理器包括：离散化连续变量，连续化离散变量，缺失值插补，选出重要特征变量，选出随机特征变量，对特征变量归一化，随机化，主成分分析，CUR 矩阵分解。

2.1.18 转换

转化数据表格。

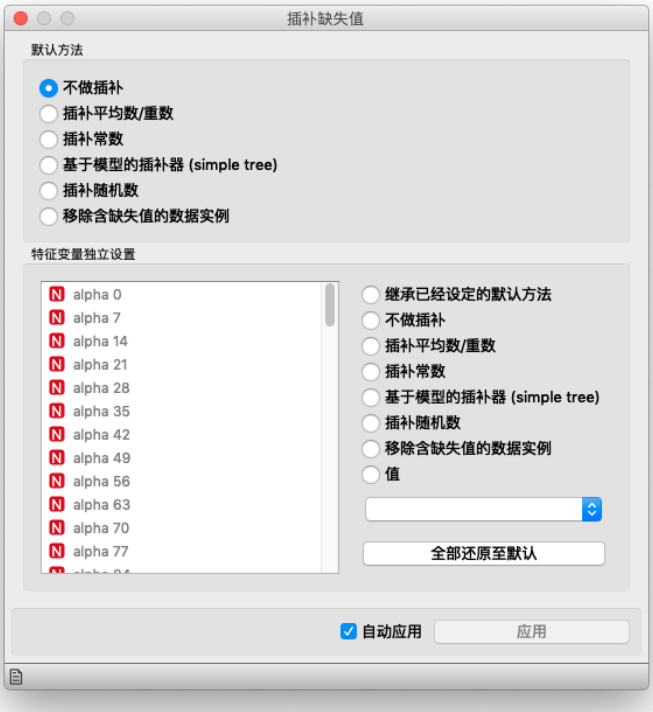


基本功能：

- 显示输入的数据集和预处理器信息，显示输出数据的基本情况。

2.1.19 插补缺失值

对数据表格中的缺失值进行插值补值。

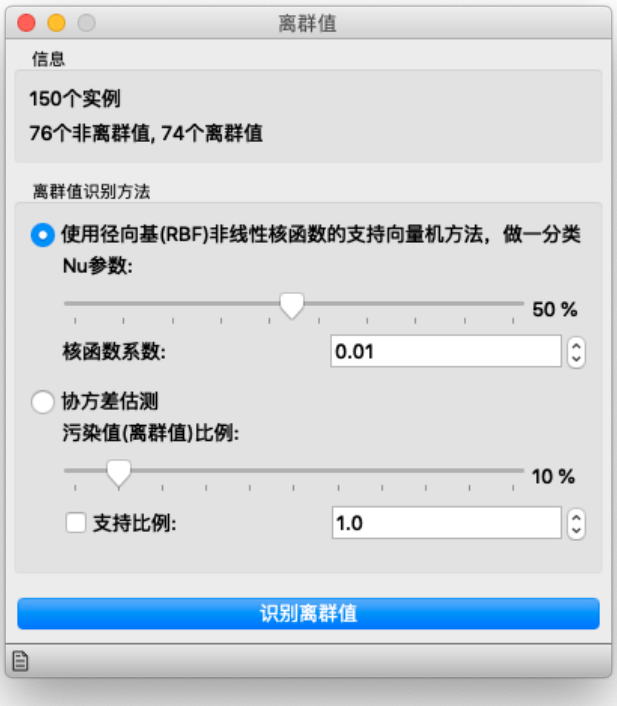


基本功能：

- 在“默认方法”中挑选默认对所有列使用的补值方法，方法包括：不做插补，插补平均数/众数，插补常数，基于模型的插补器（simple tree），插补随机数，移除含缺失值的数据实例；
- 如需要对个别特征变量进行特别的设置，在“特征变量独立设置”中选择特征变量，在右侧选择插补方法；
- 如遇到错误希望将设置还原至初始状态，点击“全部还原至默认”按钮。

2.1.20 离群值

离群值识别。

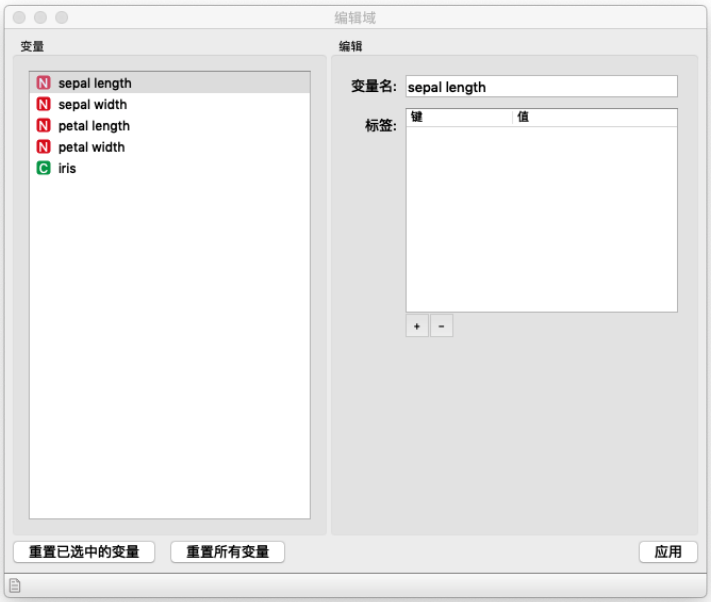


基本功能：

- “信息”一栏展示在选定的方法下，数据实例的总数，离群和非离群数据实例的分布；
- 在“离群值识别方法”中选择方法，可选方法有“径向基非线性核函数的支持向量机方法”和“协方差估测”。

2.1.21 编辑域

重命名变量，编辑分类和变量注释。

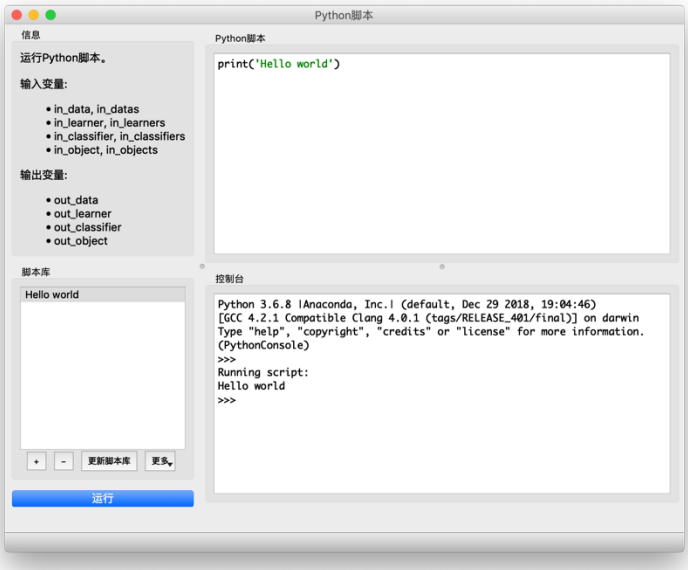


基本功能：

- 在左侧“变量”中选中要编辑的变量；
- 在右侧“编辑”中，修改变量名称和标签信息；
- 如遇到错误需要重置，可选择点击功能节点下方的“重置已选中的变量”按钮和“重置所有变量”按钮。

2.1.22 Python 脚本

编辑和运行 Python 脚本，处理输入的数据和模型。



基本功能：

- “信息”中显示智文 python 编程的数据结构信息；
- 在“Python 脚本”编辑框中编辑脚本；
- 编辑完成的脚本可以被保存至智文外部的文件中，在“脚本库”中设置脚本的读取，载

入，保存；

- “控制台”用来测试和运行脚本。

2.1.23 着色

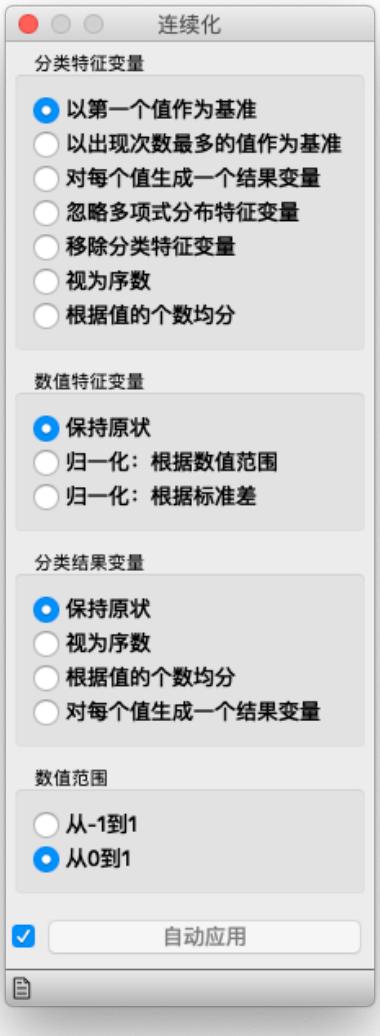
为变量设置图例中的颜色。

基本功能：

- 将所有变量分开为“离散变量”和“数值变量”两种，在两个框中对各个变量的不同取值分配一个颜色，用于接下来的可视化和分析。

2.1.24 连续化

把分类类型的特征变量转型成数值类型，并可以选择将转型后的数值进行归一化。



基本功能：

- 针对“分类特征变量”的连续化，可选择的方法包括：以第一个值作为基准，已出现次数最多的值作为基准，对每个值生成一个结果变量，忽略多项式分布特征变量，移除分类特征变量，视为序数，根据值的个数均分；
- 针对“数值特征变量”，用户可以进一步进行归一化，可选设置包括：保持原状，根据数值范围归一化，根据标准差归一化；
- 针对“分类结果变量”的连续化，可选择的方法包括：保持原状，视为序数，根据值的个数均分，对每个值生成一个结果变量；
- 在“数值范围”中，选择生成连续数值的范围。

2.1.25 创建分类结果变量

通过字符串特征变量，创建分类结果变量。

基本功能：

- 选择要编辑的变量，设置不同分类取值的名称；
- 对新建的分类结果变量，用户可以为其命名；
- 用户可以勾选“只从字符串开头匹配”和“区分大小写”。

2.1.26 离散化

把数值类变量离散化。

基本功能：

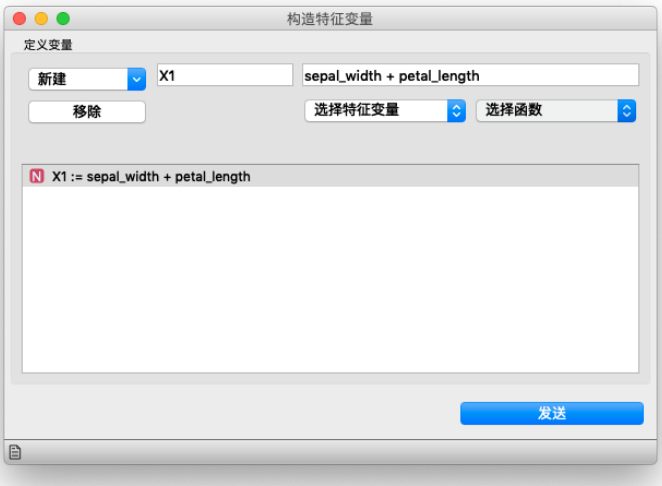
- 在“默认离散化”中，用户选择对所有变量使用的默认离散化方法，其中包括：等频离

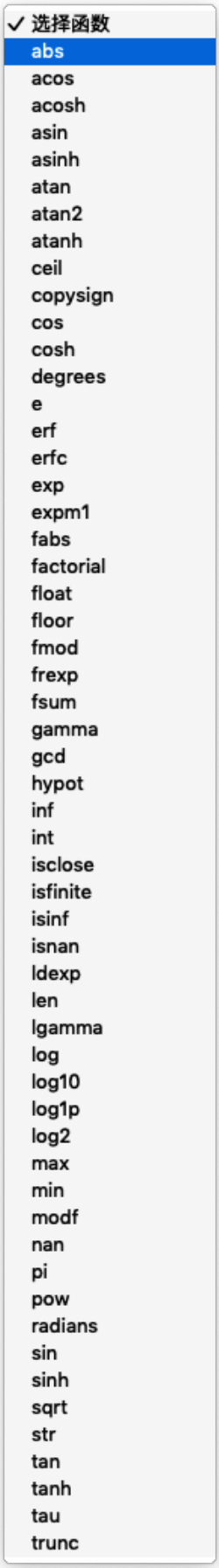
散化，等距离散化，保持数值特征，熵-最小化离散长度，移除数值特征。

- 在“特征变量独立设置”中，可针对个别特征进行单独的离散化设置，从左则的变量列表中选中要离散化的变量，在右侧选择设置其离散化方法。

2.1.27 构造特征变量

使用输入数据集里现成的特征变量，构造新的特征变量（数据列）。



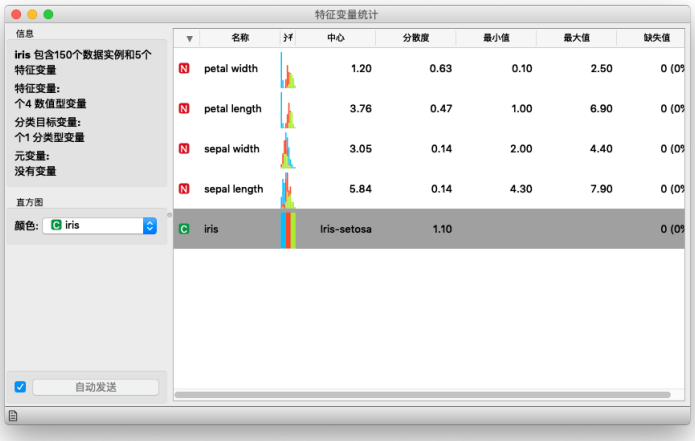

选择函数
abs
acos
acosh
asin
asinh
atan
atan2
atanh
ceil
copysign
cos
cosh
degrees
e
erf
erfc
exp
expm1
fabs
factorial
float
floor
fmod
frexp
fsum
gamma
gcd
hypot
inf
int
isclose
isfinite
isinf
isnan
ldexp
len
lgamma
log
log10
log1p
log2
max
min
modf
nan
pi
pow
radians
sin
sinh
sqrt
str
tan
tanh
tau
trunc

基本功能：

- 选择“新建”，输入新建的特征变量的名称和公式；
- 构造公式时从“选择特征变量”中选择变量，从“选择函数”中选择智文内部的自带函数；
- 新建的特征变量将显示在功能节点下方的表格中，可选中特征变量，点击“移除”移除不满意的特征变量。

2.1.28 特征变量统计

展示特征变量的基本统计。



基本功能：

- 在“信息”中显示数据集的基本信息：数据实例个数和几种变量的个数；
- 在“直方图”中选择着色的依据；
- 右侧显示各个变量的分布和基本统计指标，包括：最大值，最小值等。

2.1.29 邻域

根据引用计算数据中的最近邻域。

基本功能：

- 在“信息”中展示两个输入数据集的基本情况；
- 选择计算距离的方法和生成邻域的个数，功能节点将输出计算出的邻域。

2.1.30 保存数据

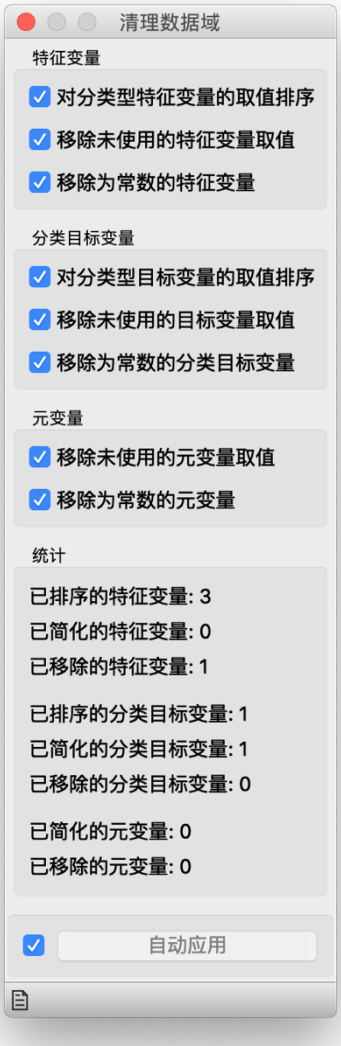
将数据保存至文件。

基本功能：

- 在“格式”中选择保存文件的类型后缀；
- 勾选“压缩”并选择压缩格式，功能节点将生成压缩完成的文件。

2.1.31 清理数据域

从数据集中移除多余的数值和特征变量，对取值排序。



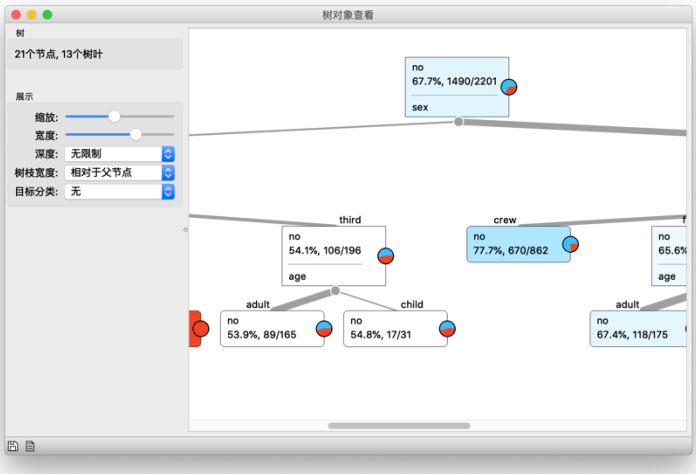
基本功能：

- 针对“特征变量”，用户可以勾选“对分类型特征变量排序”，“移除未使用的特征变量取值”，“移除为常数的特征变量”；
- 针对“分类目标变量”，用户可以选择“对分类型目标变量的取值排序”，“移除未使用的目标变量取值”和“移除为常数的分类目标变量”；
- 针对“元变量”，用户可以选择“移除未使用的元变量取值”和“移除为常数的元变量”；
- 在“统计”中查看特征变量，目标变量和元变量被简化，排序和移除的统计。

2.2 “可视化”功能节点组

- 树对象查看
- 箱线图：用箱线图观察特征变量数值的分布。
- 分布：展示一个特征变量的分布。
- 散点图：可交互散点图，智能提升可视化效果。
- 线图：对数据资料可视化（时间序列等）。
- 筛网图：可视化对比变量组合的期望频率和观察到的频率。
- 马赛克图：用马赛克图展示数据。
- FreeViz：展示 FreeViz 投影。
- 线性投影：在二维平面上的多数轴投影。
- Radviz:展示 Radviz 投影。
- 热力图：对一对特征变量绘制热力图。
- 文氏图：展示数据集之间的重叠关系。
- 轮廓图：视觉评估聚类的品质。
- 毕达哥拉斯树：用毕达哥拉斯树对树类结构进行可视化。
- 毕达哥拉斯森林：用毕达哥拉斯森林对随机森林进行可视化。
- CN2 规则查看：浏览由数据归纳出的规则。
- 列线图：使用列线图观察朴素贝叶斯和逻辑回归分类器。

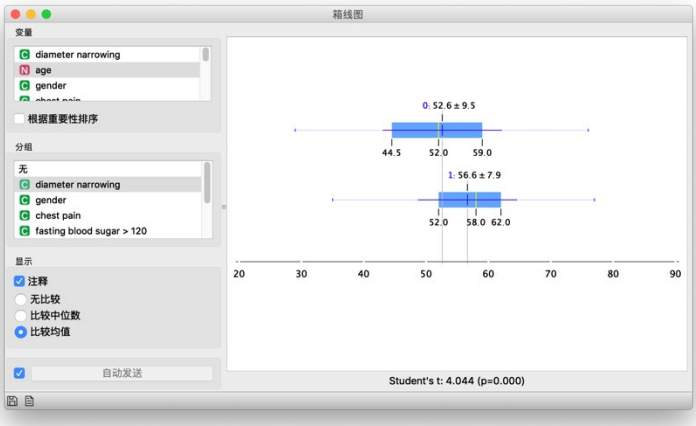
2.2.1 树对象查看



- “树”展示树的信息，显示节点和树叶的个数；
- 在“展示”中设置可视化的参数，包括“缩放”，“宽度”，“深度”，“树枝宽度”和“目标分类”；
- 在右侧的分类树中查看分类情况的条件。

2.2.2 箱线图

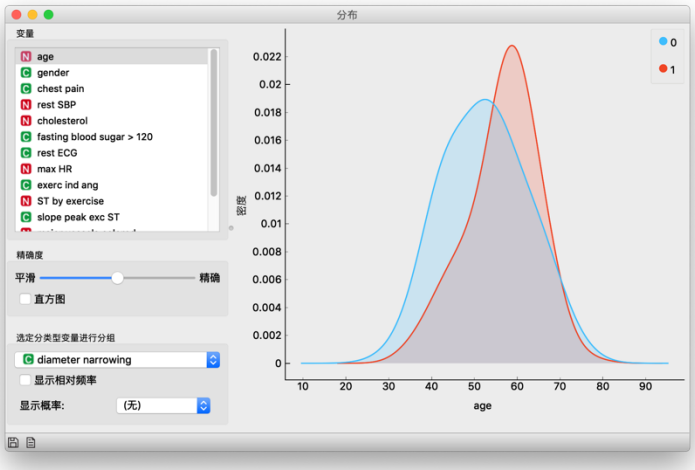
用箱线图观察特征变量数值的分布。



- 在“变量”中选择制作箱线图的变量；
- 在“分类”中选择分类型变量，用来显示和比较不同分类的箱线的不同；
- 在“设置”中可以选择箱线图的显示细节。

2.2.3 分布

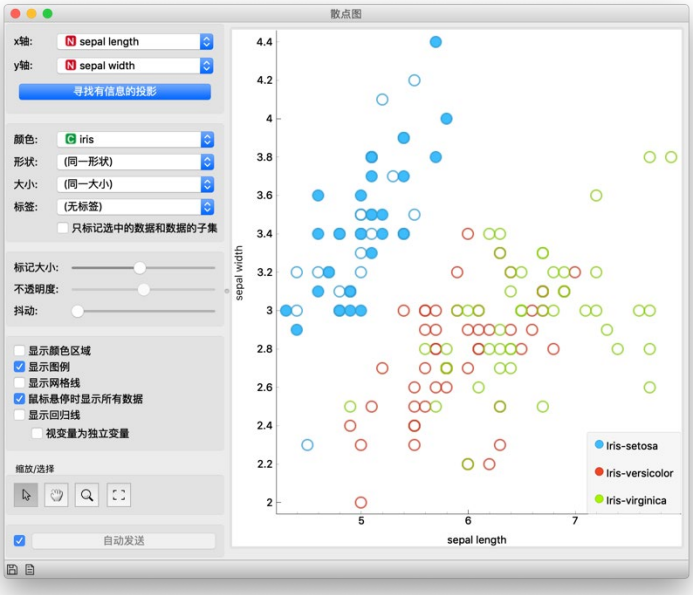
展示一个特征变量的分布。



- 在“变量”中选择用于制作分布图的变量；
- 在“精确度”中调节分布图的平滑/精确度
- 可以在“选定分类型变量进行分组”中选择分类型变量，对应每一个分类生成一个单独的分布，方便比较。

2.2.4 散点图

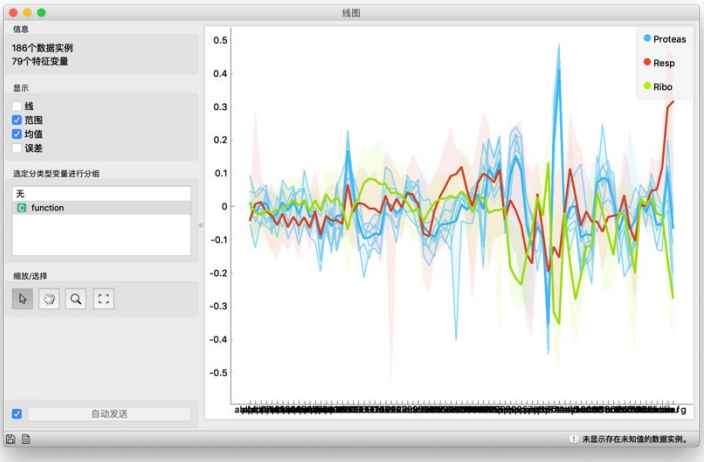
可交互散点图，智能提升可视化效果。



- 在“x 轴”“y 轴”选择二维平面散点图两个数轴代表的变量；点击“寻找有信息的投影”可以计算变量两两之间的关系，帮助用户选择出效果更佳的变量对；
- 用户可以设置散点图显示的细节，包括散点的颜色，形状，大小和标签；
- 用户可以进一步修饰散点图的显示效果，可以勾选“显示颜色区域”，“显示图例”，“显示网格线”，“鼠标悬停时显示所有数据”和“显示回归线”。

2.2.5 线图

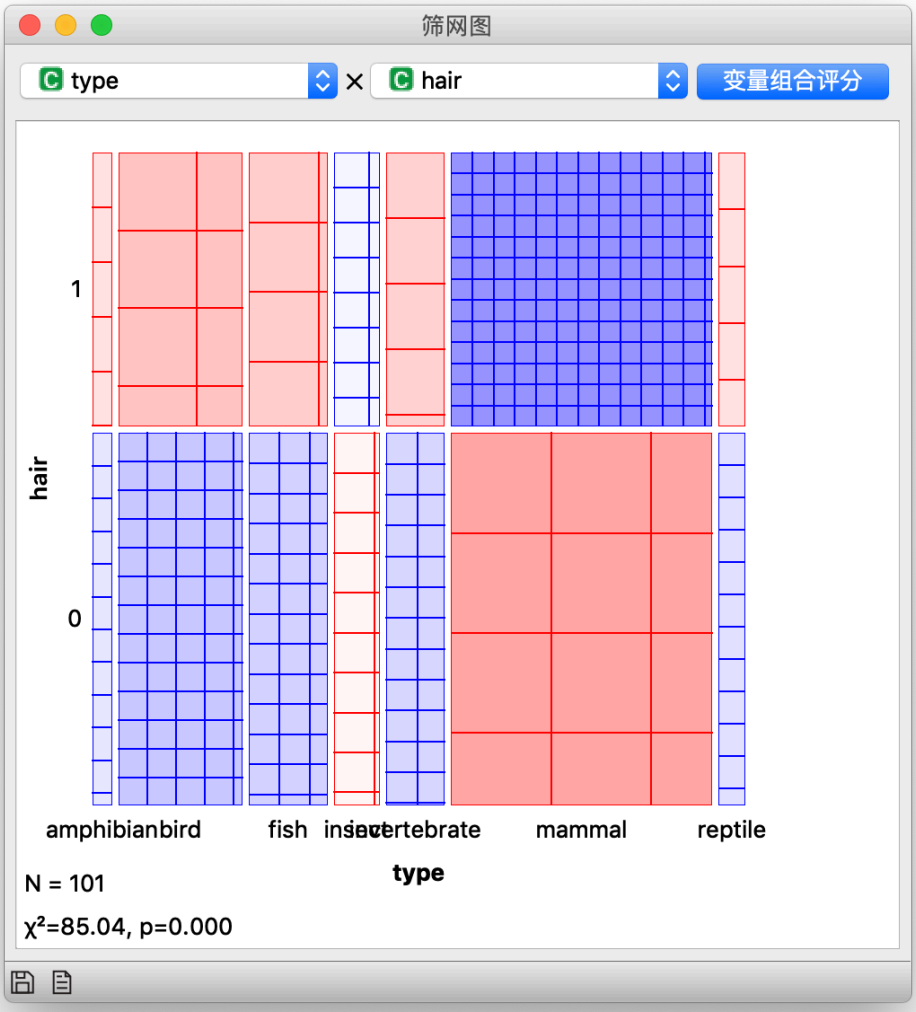
对数据资料可视化（时间序列等）。



- “信息”展示输入数据的基本信息，包括数据实例和特征变量的个数；
- 用户可以选择线图显示的主要内容，包括“线”，“范围”，“均值”和“误差”；
- 用户可以对不同的分类进行单独的线图绘制，在“选定分类型变量进行分组”中选择用于分类的变量。

2.2.6 筛网图

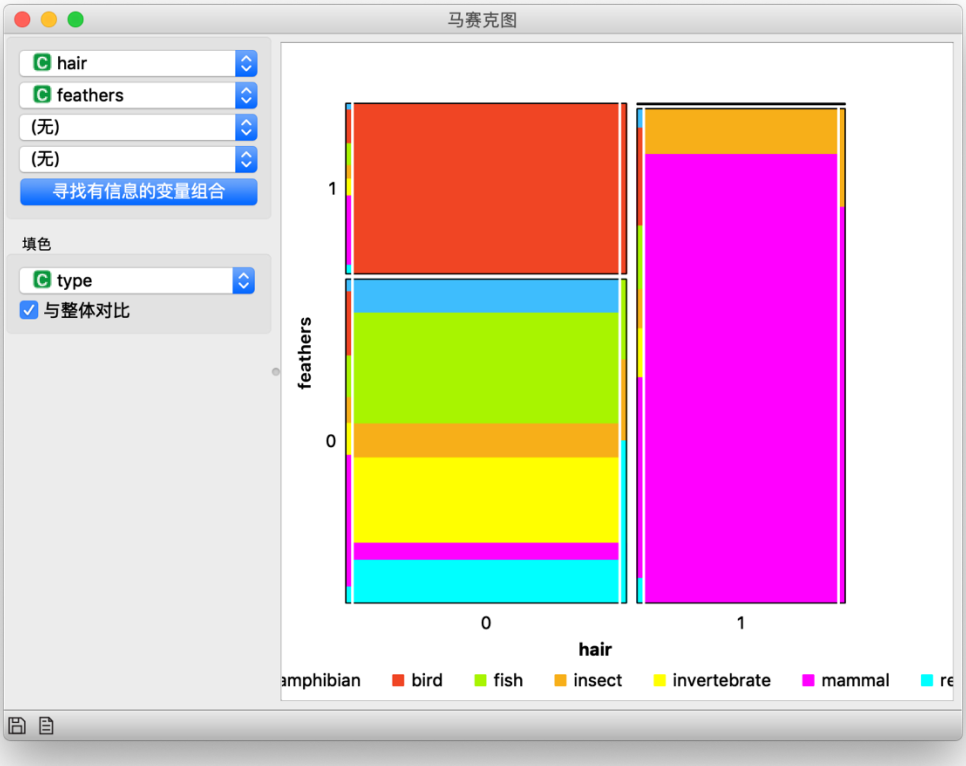
可视化对比变量组合的期望频率和观察到的频率。



- 在功能节点上方选择筛网图的两个维度的变量，点击“变量组合评分”，智文对变量两两组合进行评估，帮助用户选择效果最好的变量对；
- 筛网图用不同密度的网格展示不同分类配对的独立性。

2.2.7 马赛克图

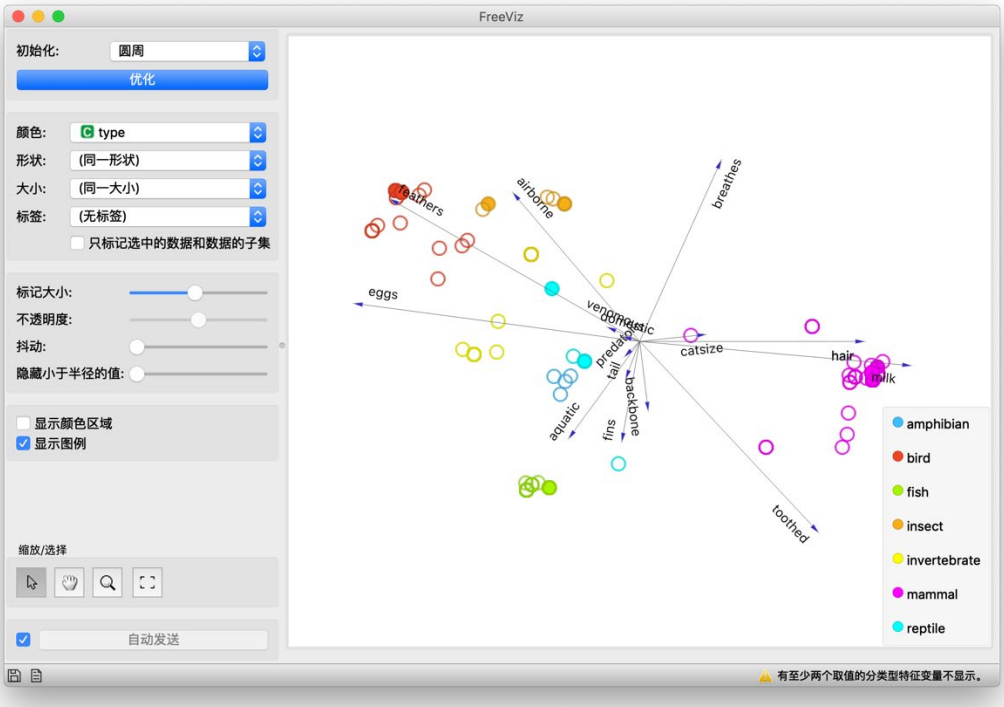
用马赛克图展示数据。



- 点击“寻找有信息的变量组合”，智文帮助用户找到效果最好的变量对，用户也可以自己调整变量组合；
- 在“填色”部分选择一个分类型变来功能，马赛克图将区分此分类变量的不同取值；
- 点击“与整体对比”，每块马赛克格的左右两侧会出现整体分布条。

2.2.8 FreeViz

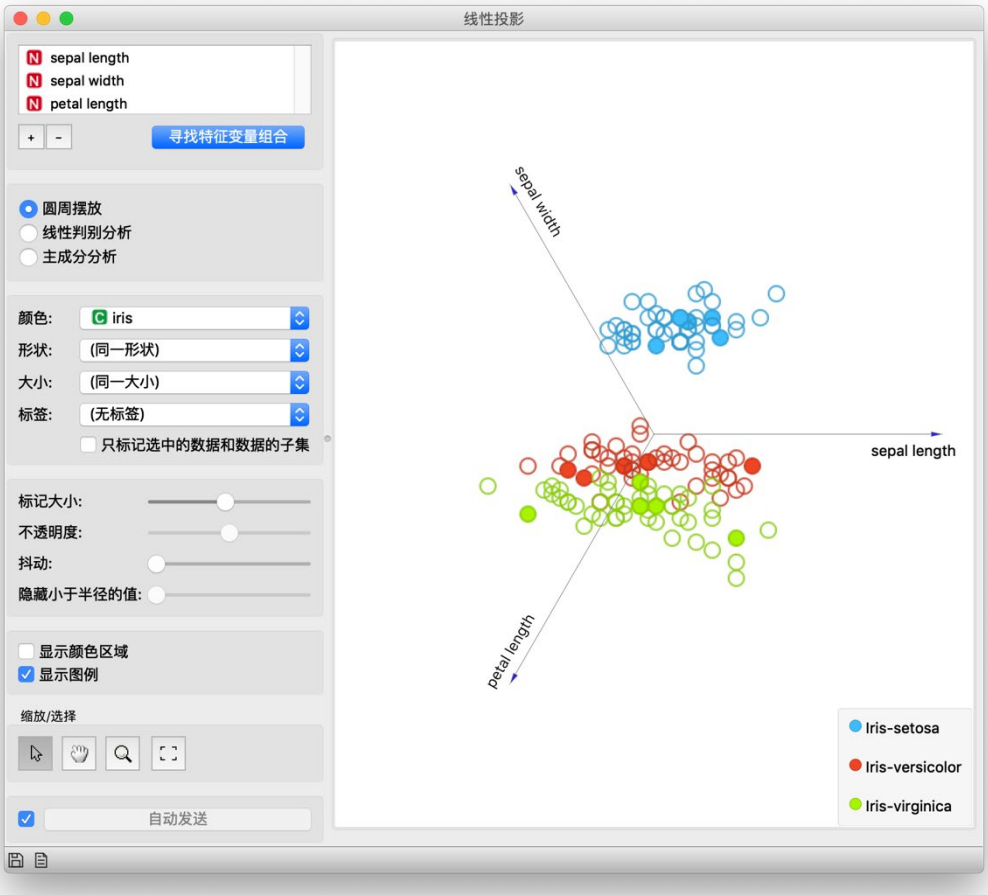
展示 FreeViz 投影。



- “初始化”设置数轴/变量之间的关系，“圆周”给每个数轴一样的权重；
- 点击“优化”，智文将自动对变量的分布进行计算，右侧各个数轴将根据计算结果排斥或吸引，改变相互之间的位置和长度，从而体现不同变量的关系和重要性；
- 可视化设置包括：颜色，形状，大小，标签，抖动，不透明度，隐藏小于半径的值。

2.2.9 线性投影

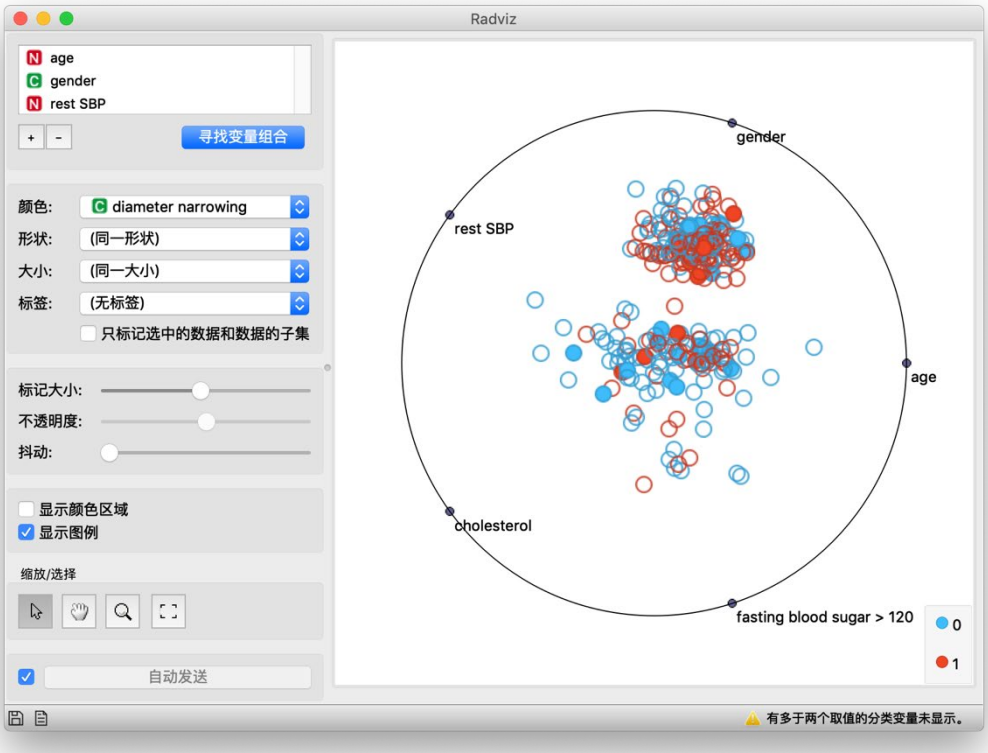
在二维平面上的多数轴投影。



- 选择变量加入到投影图中，成为图中的数轴；
- 点击“寻找特征变量组合”可以帮助用户找到变量之间突出的关系；
- 可选的数轴的摆放方式有“圆周摆放”，“线性判别分析”和“主成分分析”；
- 可视化设置包括：颜色，形状，大小，标签，抖动，不透明度，隐藏小于半径的值。

2.2.10 Radviz

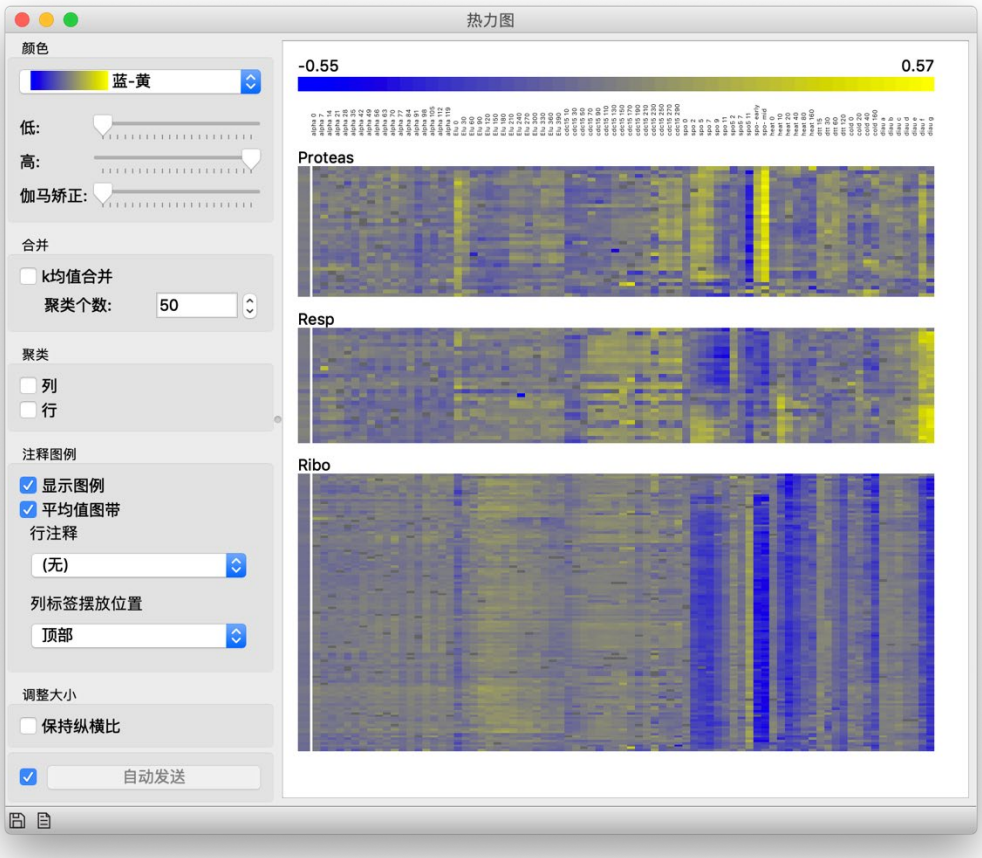
展示 Radviz 投影。



- 选择变量加入到图中，点击“寻找变量组合”可以帮助用户找到变量之间突出的关系；
- 可视化设置包括：颜色，形状，大小，标签，抖动，不透明度，隐藏小于半径的值。

2.2.11 热力图

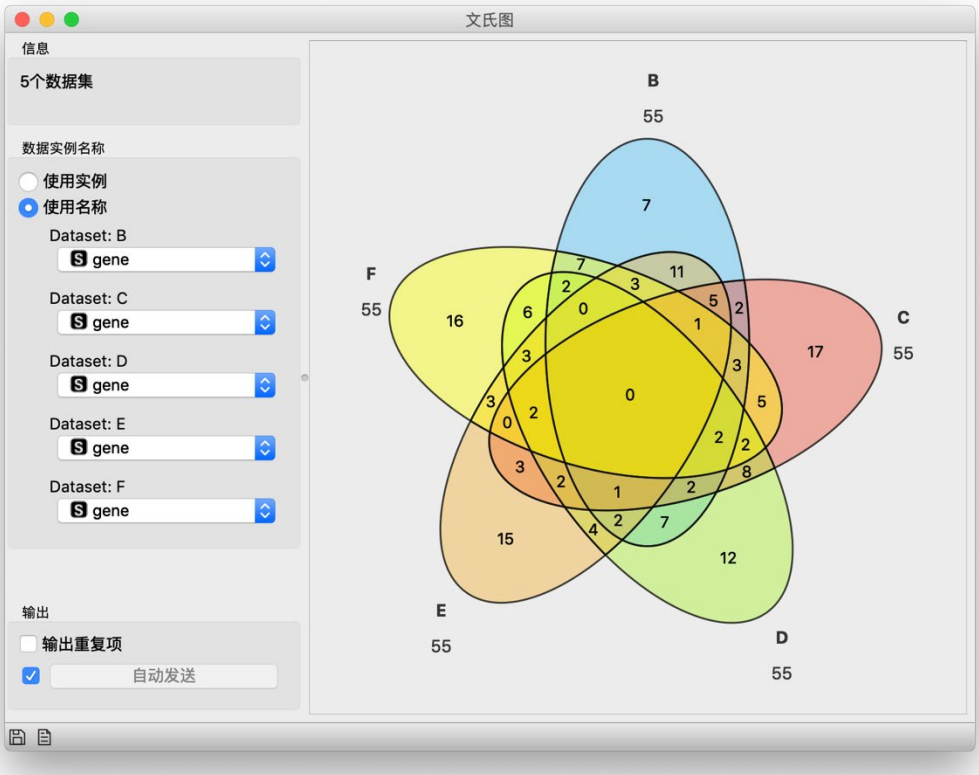
对一对特征变量绘制热力图。



- 通过“颜色”选择热力图的主色调和其他显示效果；
- 通过“合并”进行 k 均值合并，用户可使定聚类的个数；
- 通过“聚类”选择聚类的对象；
- 通过“注释图例”设置热力图的效果。

2.2.12 文氏图

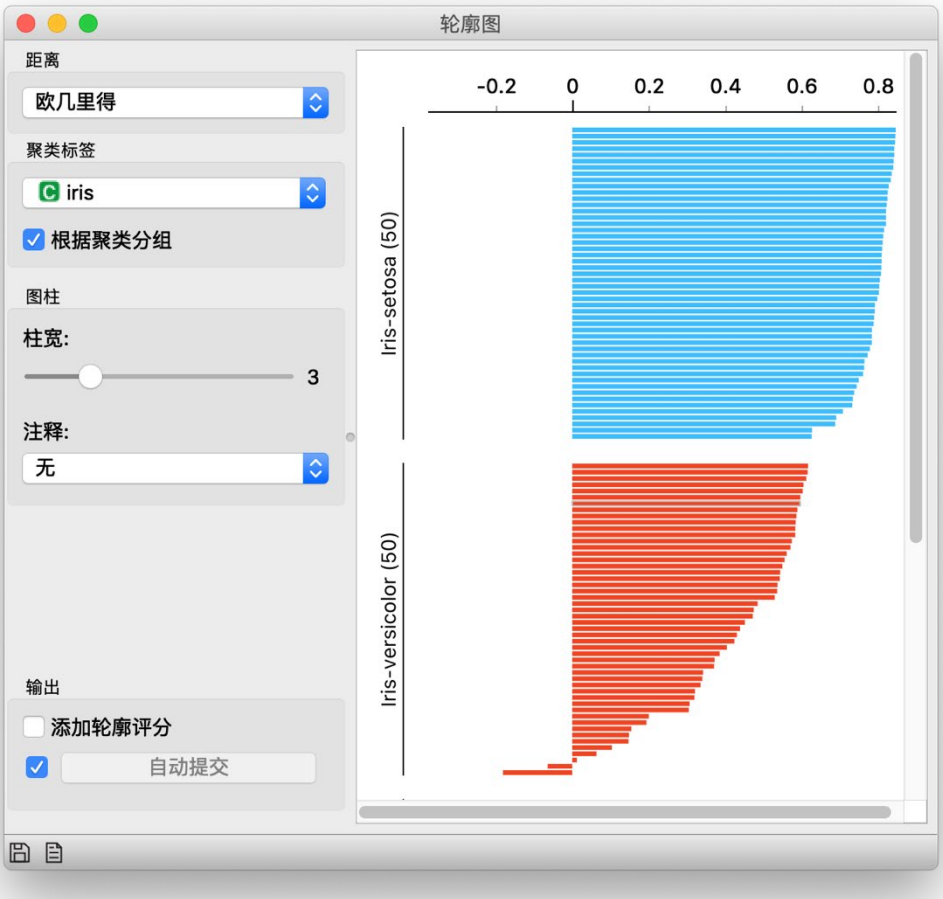
展示数据集之间的重叠关系。



- “信息”显示数据集的个数；
- 在“数据实例名称”中选择数据实例的个体；
- 勾选“输出重复项”，则重叠部分的数据将被输入。

2.2.13 轮廓图

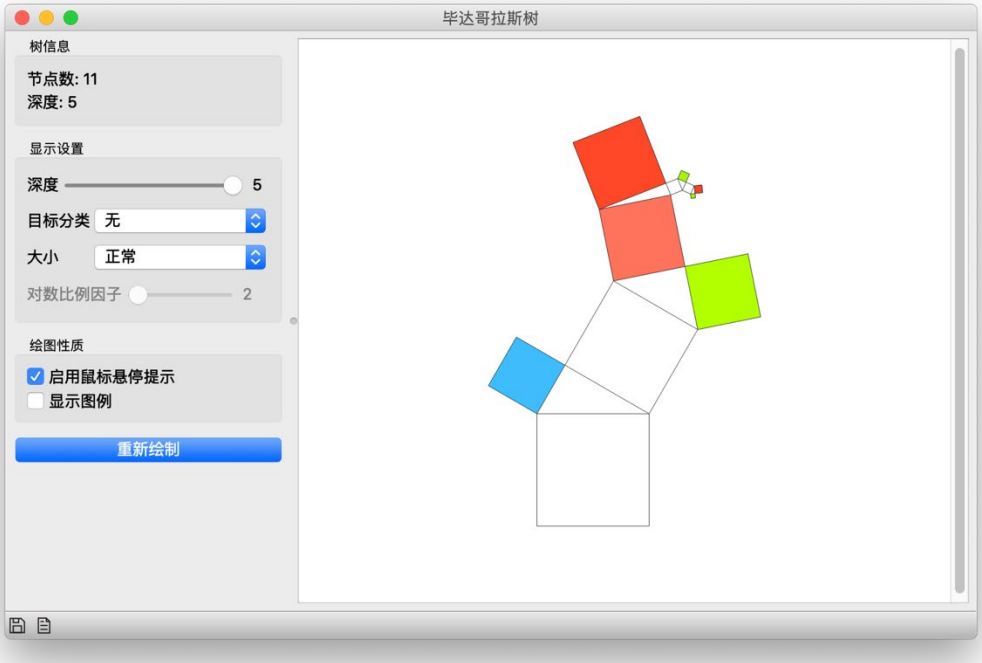
视觉评估聚类的品质。



- 通过“距离”选择计算距离的方法；
- 在“聚类标签”中确定数据集的分类是哪一列；
- 在“图柱”中设置轮廓图中的图形；
- 选择“添加轮廓评分”可以将计算出的轮廓得分一同输出。

2.2.14 毕达哥拉斯树

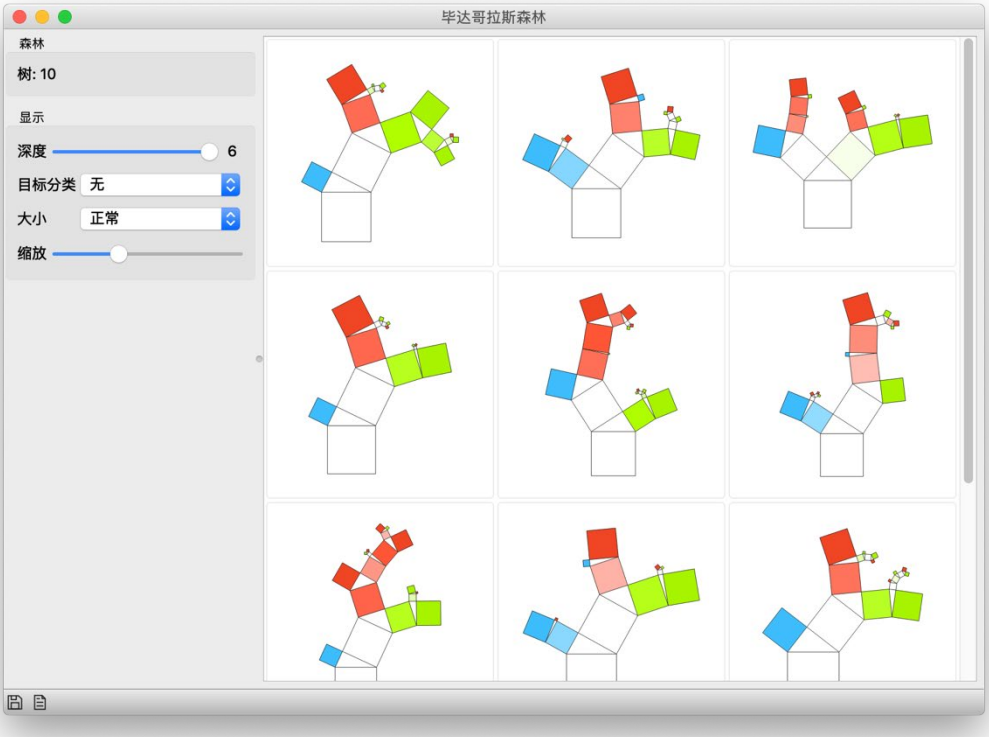
用毕达哥拉斯树对树类结构进行可视化。



- “树信息”显示毕达哥拉斯树的基本信息，包括节点数和深度；
- 在“显示设置”中调整图形。

2.2.15 毕达哥拉斯森林

用毕达哥拉斯森林对随机森林进行可视化。



- “森林”显示毕达哥拉斯森林中，树的个数；
- 通过“显示”中的选项设置毕达哥拉斯森林的可视化效果。

2.2.16 CN2 规则查看

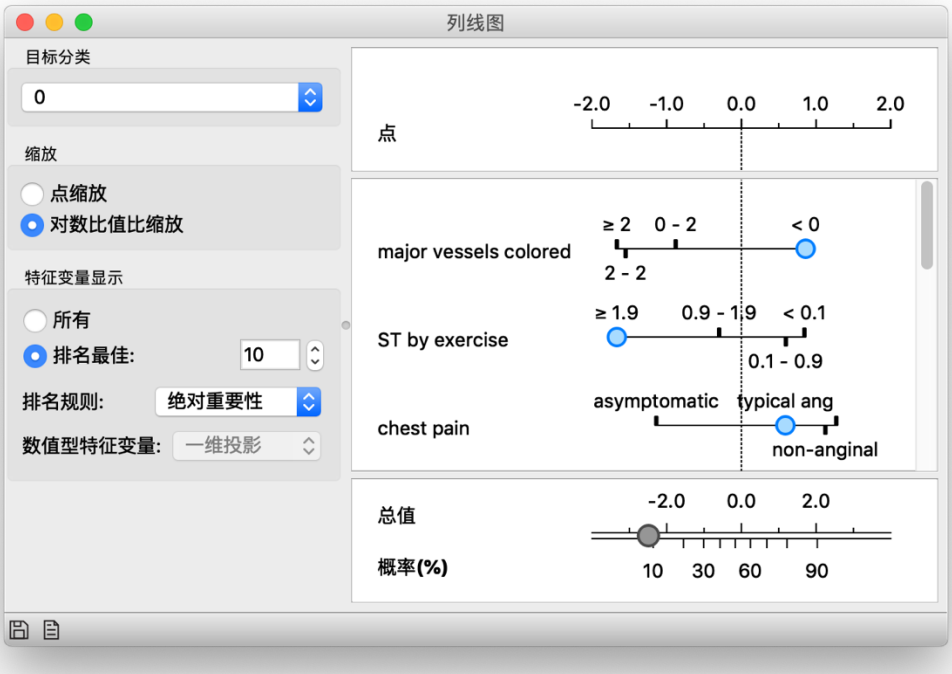
浏览由数据归纳出的规则。

	IF条件		THEN分类	分布	概率[%]	规则质量	规则长度
0	petal length≤3.0 AND sepal width≥2.9	→	iris=Iris-setosa	[49, 0...	96 : 2 : 2	-0.00	2
1	petal width≥1.8 AND sepal length≥6.0	→	iris=Iris-virginica	[0, 0, ...	2 : 2 : 95	-0.00	2
2	sepal length≥4.9 AND sepal width≥3.1	→	iris=Iris-versicolor	[0, 8, ...	9 : 82 : 9	-0.00	2
3	petal length≤4.9 AND petal width≥1.7	→	iris=Iris-virginica	[0, 0, ...	20 : 20 : 60	-0.00	2
4	petal width≥1.8	→	iris=Iris-virginica	[0, 0, ...	12 : 12 : 75	-0.00	1
5	petal length≤5.0 AND sepal width≥2.4	→	iris=Iris-versicolor	[0, 35...	3 : 95 : 3	-0.00	2
6	sepal width≥2.8	→	iris=Iris-virginica	[0, 0, ...	20 : 20 : 60	-0.00	1
7	petal width≤1.0 AND sepal length≥5.0	→	iris=Iris-versicolor	[0, 3, ...	17 : 67 : 17	-0.00	2
8	sepal width≥2.7	→	iris=Iris-versicolor	[0, 1, 0]	25 : 50 : 25	-0.00	1
9	sepal width≥2.6	→	iris=Iris-virginica	[0, 0, 1]	25 : 25 : 50	-0.00	1
10	sepal length≥5.5 AND sepal length≥6.2	→	iris=Iris-versicolor	[0, 2, ...	20 : 60 : 20	-0.00	2
11	sepal length≤5.5 AND petal length≥4.0	→	iris=Iris-versicolor	[0, 1, 0]	25 : 50 : 25	-0.00	2

- 以表格的形式呈现所有 CN2 规则，列分别为“IF 条件”，“THEN 条件”，“分布”，“概率”，“规则质量”，“规则长度”。

2.2.17 列线图

使用列线图观察朴素贝叶斯和逻辑回归分类器。



- 在“目标分类”中选择分类的个数；
- 通过“特征变量显示”设置特征变量的显示效果；
- 可选择两种缩放方式，“点缩放”和“对数比值比缩放”。

2.3 “模型”功能节点组

- 常数：预测训练数据集中频率最高的分类或均值。
- CN2 规则归纳：用 CN2 算法归纳数据的规则。
- kNN：通过最近的训练数据实例做预测。
- 树：应用前剪枝的决策树算法。
- 随机森林：用一整套决策树做预测。
- 支持向量机：支持向量机把输入数据投射到高维度的特征空间。
- 线性回归：线性回归算法，可选择 L1（Lasso），L2（岭回归）或 L1L2（elastic net）正则化。
- 逻辑回归：逻辑回归算法，可选 L1（Lasso）或 L2（岭回归）正则化。
- 朴素贝叶斯：一个快速简单的概率分类器，基于贝叶斯定理，假设特征变量的独立性。
- AdaBoost：自适应增强元算法把一系列较弱的学习器集合起来，以适应每个训练数据集的不同难度。

- 神经网络：基于反向传播的多层感知器算法。
- 随机梯度下降：通过对梯度下降进行随机近似，最小化一个目标函数。
- 模型堆叠：把多个模型堆叠起来。
- 保存模型：将训练好的模型保存至外部文件。
- 载入模型： 从外部文件载入一个模型。

2.3.1 常数

预测训练数据集中频率最高的分类或均值。

- 在“名称”中设置功能节点在智文流程中的名称。

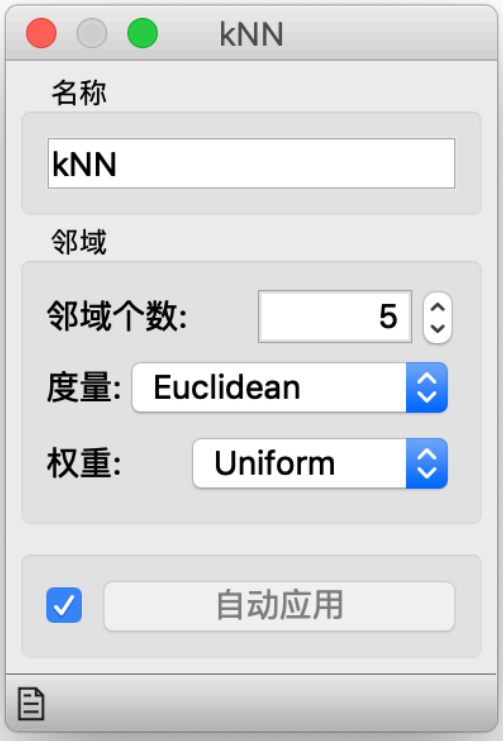
2.3.2 CN2 规则归纳

用 CN2 算法归纳数据的规则。

- 在“规则排序”中选择“已排序”或“未排序”；
- 在“覆盖算法”中选择算法：“排他”或“权重”，如选择权重，需要给出一个权重值；
- 在“规则搜索”中选择对规则的“评估测度”，设置“束宽度”；
- 在“规则过滤”中设置选出符合需求的规则，可以针对规则的覆盖下限，长度上线，统计显著性，相对显著性进行判断，不符合条件的规则不被输出。

2.3.3 kNN

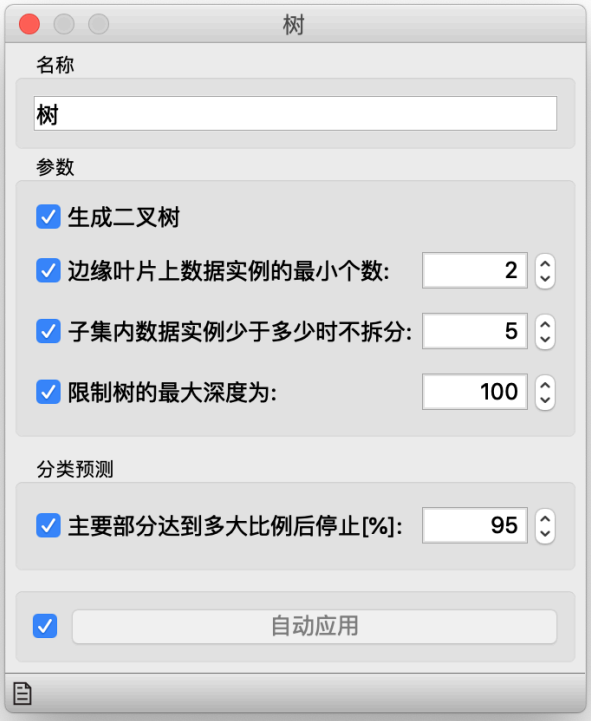
通过最近的训练数据实例做预测。



- 在“邻域”中设置和调试领域的个数，采用不同的度量方法，以不同方式分配权重。

2.3.4 树

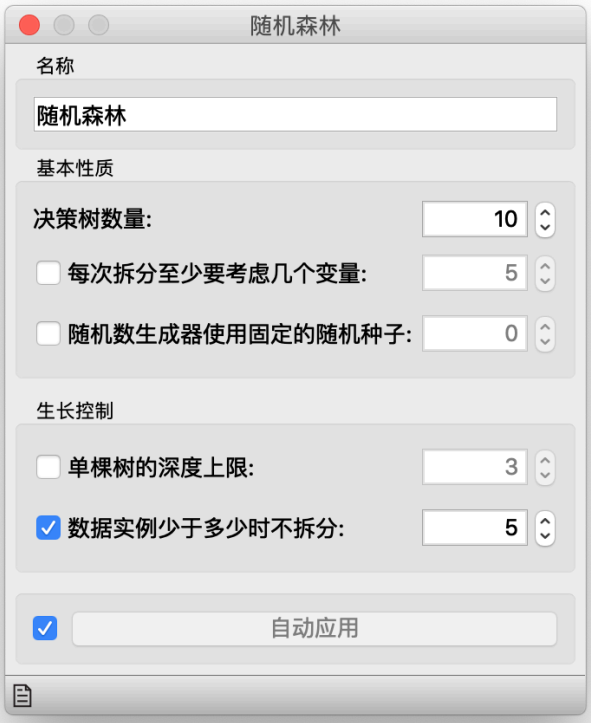
应用前剪枝的决策树算法。



- 在“参数”中设置算法的细节，可设置“生成二叉树”，“边缘叶片上数据实例的最小个数”，“自己哪数据实例少于多少时不拆分”，“显示树的最大深度”；
- 在“分类预测”中设置树的覆盖比例。

2.3.5 随机森林

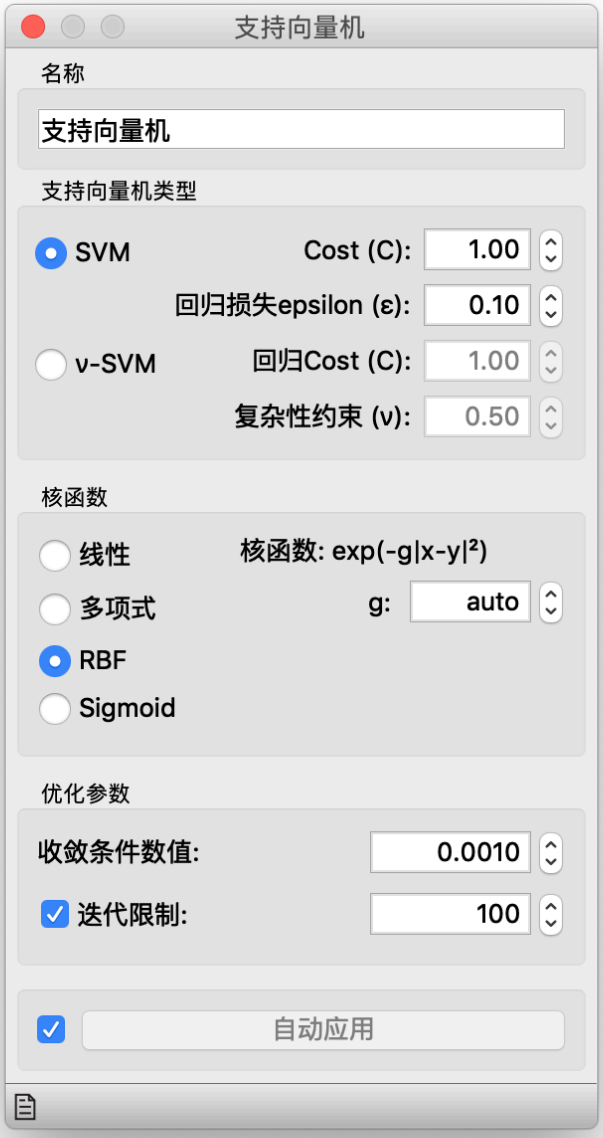
用一整套决策树做预测。



- 在“基本性质”中设置“决策树数量”，“每次拆分至少要考虑的变量数”，“随机数生成器使用固定的随机种子”；
- 在“生长控制”中设置“单棵树的深度上限”，“数据实例少于多少时不拆分”。

2.3.6 支持向量机

支持向量机把输入数据投射到高维度的特征空间。



- 可选择两种支持向量机：“SVM”或“v-SVM”，在“支持向量机类型”中选择并设置相关的参数；
- 在“核函数”中选择和设置支持向量机算法的核函数，可选择线性，多项式，RBF，Sigmoid等多种核函数；
- 在“优化参数”中设定“收敛条件数值”和“迭代限制”。

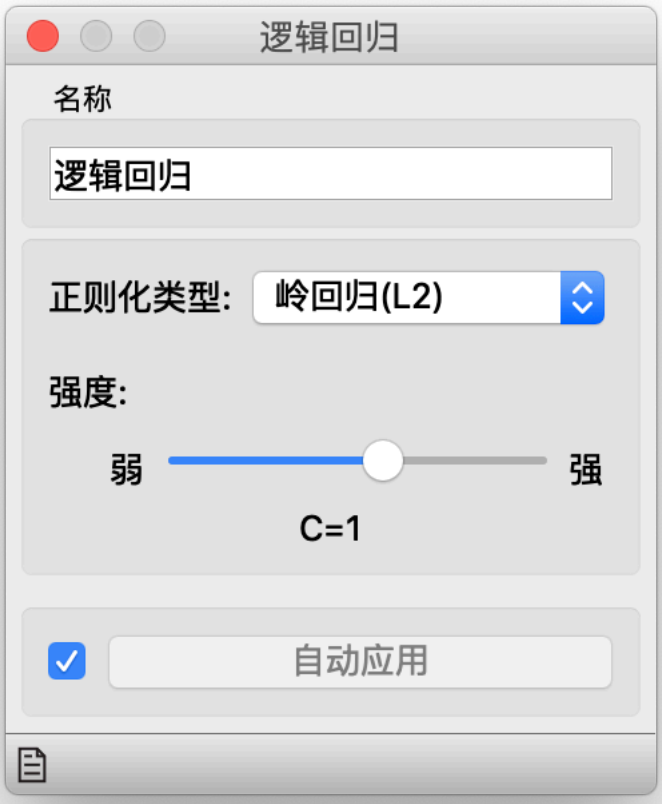
2.3.7 线性回归

线性回归算法，可选择 L1（Lasso），L2（岭回归）或 L1L2（elastic net）正则化。

- 在“正则化”中设置，可选择“不进行正则化”，“岭回归”，“Lasso 回归”，“Elastic net 回归”，用户可根据不同的正则化规则设定“正则化力度”等参数。

2.3.8 逻辑回归

逻辑回归算法，可选 L1（Lasso）或 L2（岭回归）正则化。



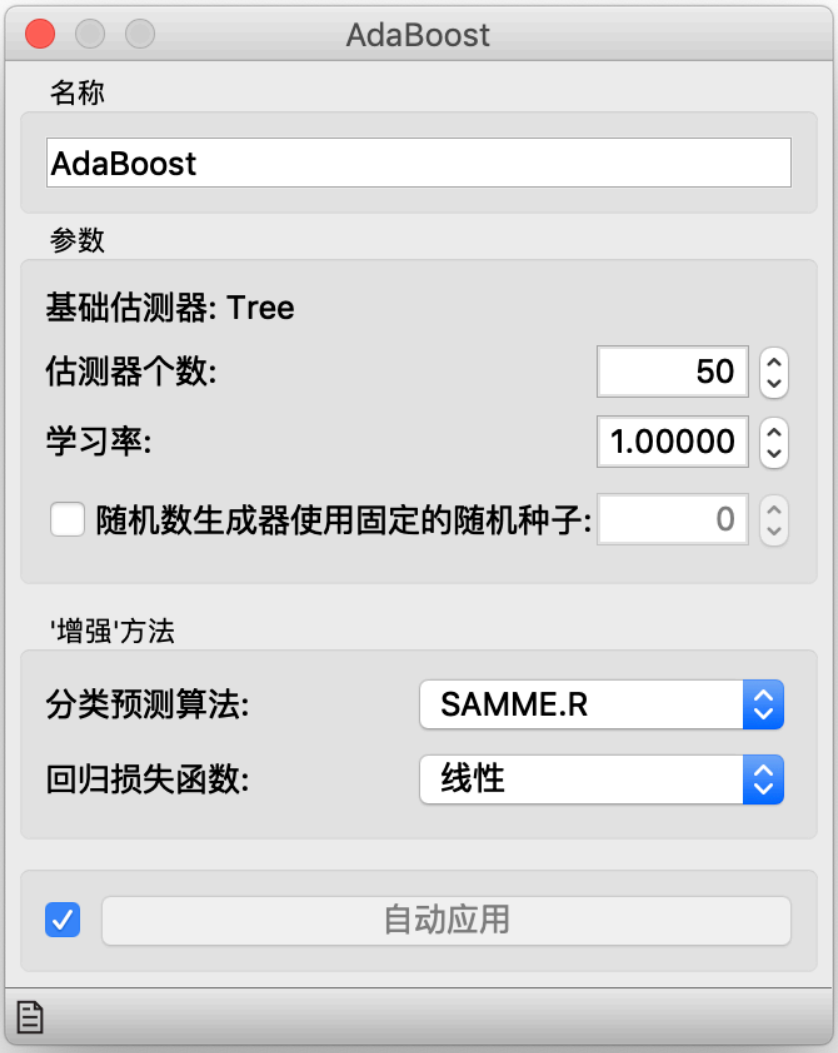
- 可选择“岭回归”等不同的正则化方法，用户可根据不同的正则化规则设定“强度”等参数。

2.3.9 朴素贝叶斯

一个快速简单的概率分类器，基于贝叶斯定理，假设特征变量的独立性。

2.3.10 AdaBoost

自适应增强元算法把一系列较弱的学习器集合起来，以适应每个训练数据集的不同难度。



- 在“参数”中设置“估测器个数”，“学习率”和“随机数生成器使用固定的随机种子”；
- 在“‘增强’方法”中设置“分类预测算法”和“回归损失函数”。

2.3.11 神经网络

基于反向传播的多层感知器算法。

- “隐层中神经节点数量”设置每一个隐层的节点数，每个数字代表每层节点数，用逗号分开
- 用户可选择“激活函数”和“解算方法”，设定正则化参数和迭代次数上限。

2.3.12 随机梯度下降

通过对梯度下降进行随机近似，最小化一个目标函数。


随机梯度下降
名称
SGD
算法
分类预测损失函数:
Hinge
ε:
0.10
回归预测损失函数:
平方损失
ε:
0.10
正则化
正则化方法:
岭回归(L2)
正则化强度(α):
0.00001
混合参数:
0.15
学习参数
学习率:
常数
初始学习率(η0):
0.0100
逆缩放指数(t):
0.2500
迭代次数:
1000
收敛条件:
0.0010
每次迭代时打乱数据
随机打乱使用固定的随机种子:
0
自动应用

- 在“算法”中设置“分类预测损失函数”和“回归预测损失函数”；
- 在“正则化”中选择“正则化方法”，设置“正则化强度”和其他参数；
- 在“学习参数”中设置“学习率”，“初始学习率”，“逆缩放指数”等参数。

2.3.13 模型堆叠

把多个模型堆叠起来。

2.3.14 保存模型

将训练好的模型保存至外部文件。

2.3.15 载入模型

从外部文件载入一个模型。

2.4 “评估”功能节点组

- 测试与评分：估测准确性，交叉验证。
- 预测：使用模型对输入数据集进行预测，并展示。
- 混淆矩阵：用混淆矩阵展示分类器的评估结果。
- ROC 分析：根据对分类器的评估展示 ROC 曲线。
- Lift 曲线：根据对分类器的评估构造展示一条 lift 曲线。
- 校准图：根据对分类器的评估绘制的校准图。

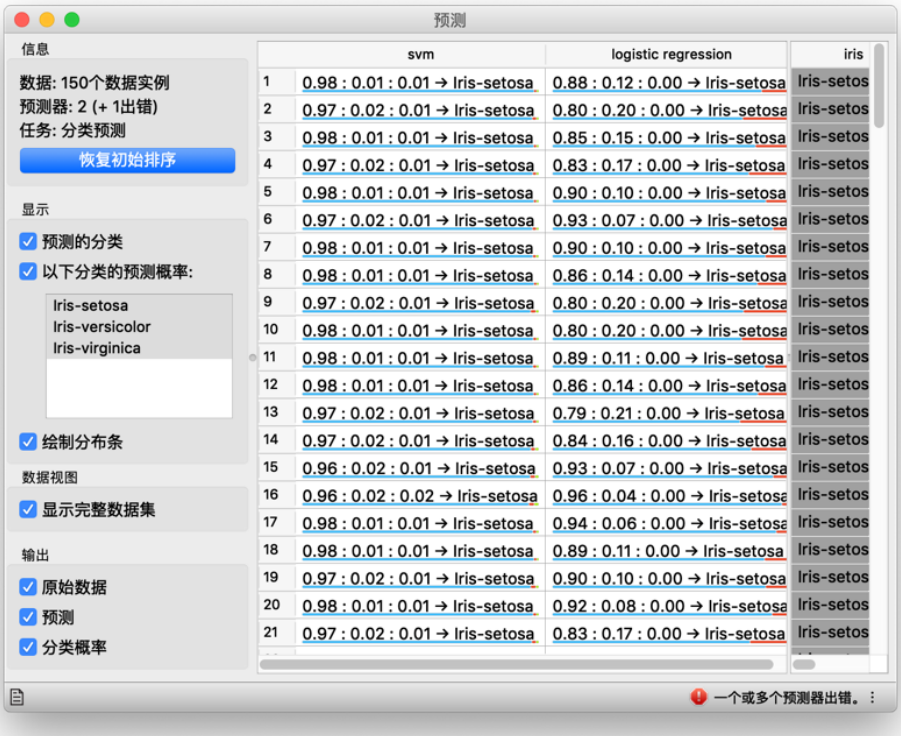
2.4.1 测试与评分

估测准确性，交叉验证。

- 在“抽取样本”抽取数据的方法；
- 在“目标分类”中设置分类变量。

2.4.2 预测

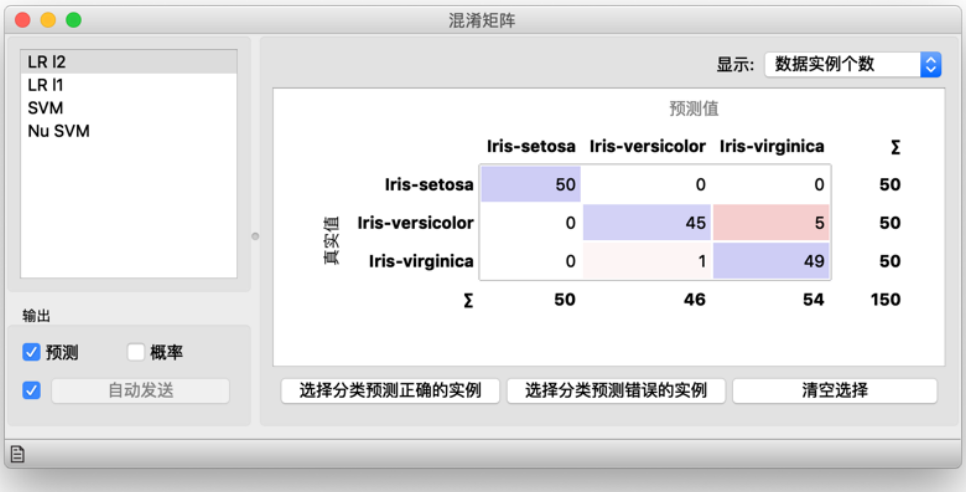
使用模型对输入数据集进行预测，并展示。



- “信息”显示数据的基本信息和预测的性质和效果；
- 在“显示”中可设置查看“预测的分类”，“预测概率”和“绘制分布条”；
- 在“输出”中设置输出的数据结构，用户可选择是否将部分预测数据输出。

2.4.3 混淆矩阵

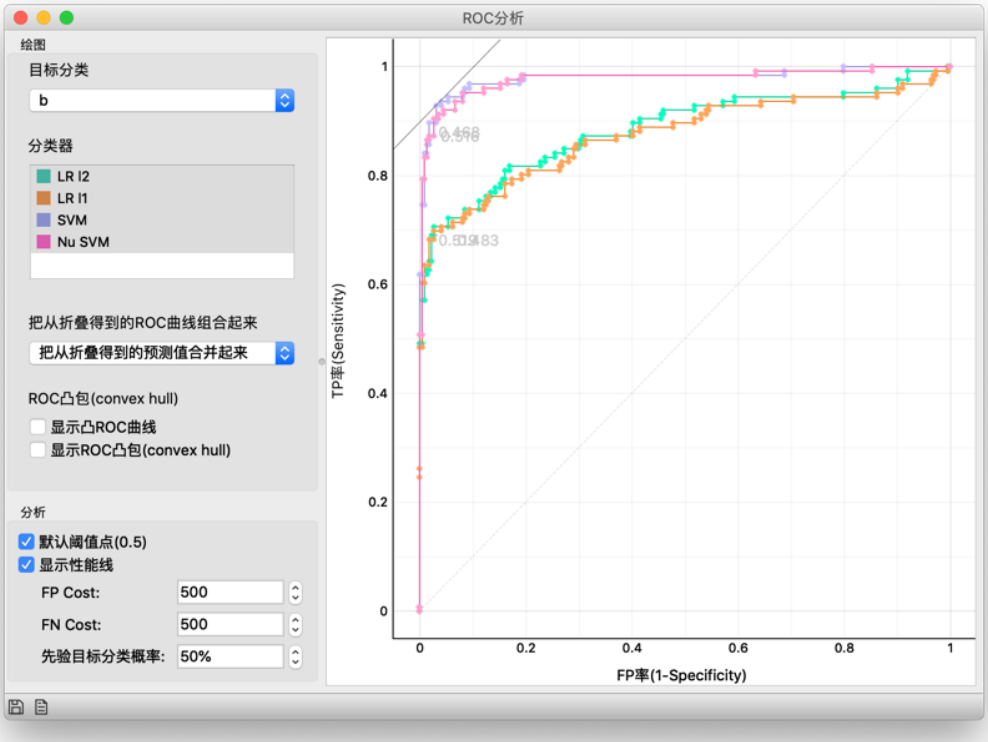
用混淆矩阵展示分类器的评估结果。



- 在界面右侧的矩阵中，可以在“显示”下拉菜单设置矩阵的数据基准，下方的三个按钮可以设置输出的数据内容；
- 在“输出”中设置输出的内容，可选择“预测”数据和“概率”数据。

2.4.4 ROC 分析

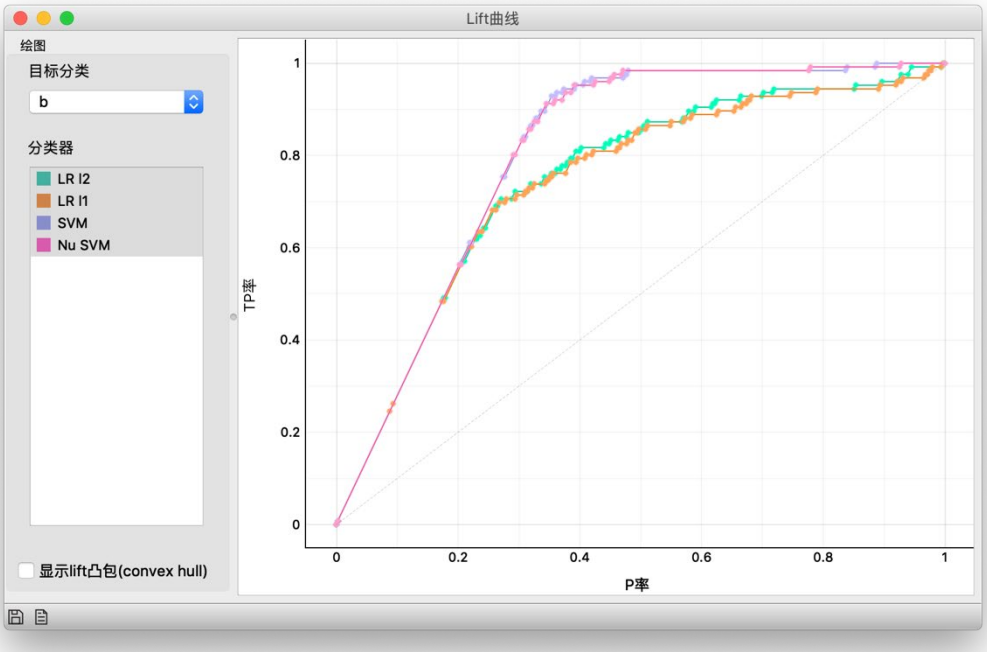
根据对分类器的评估展示 ROC 曲线。



- 在“绘图”中设置分析的目标分类；
- 在“分类器”中选择分类模型，用户可以选择折叠合并，并设置 ROC 凸包曲线；
- 在“分析”中设置“默认阈值点”和“性能线”的相关参数。

2.4.5 Lift 曲线

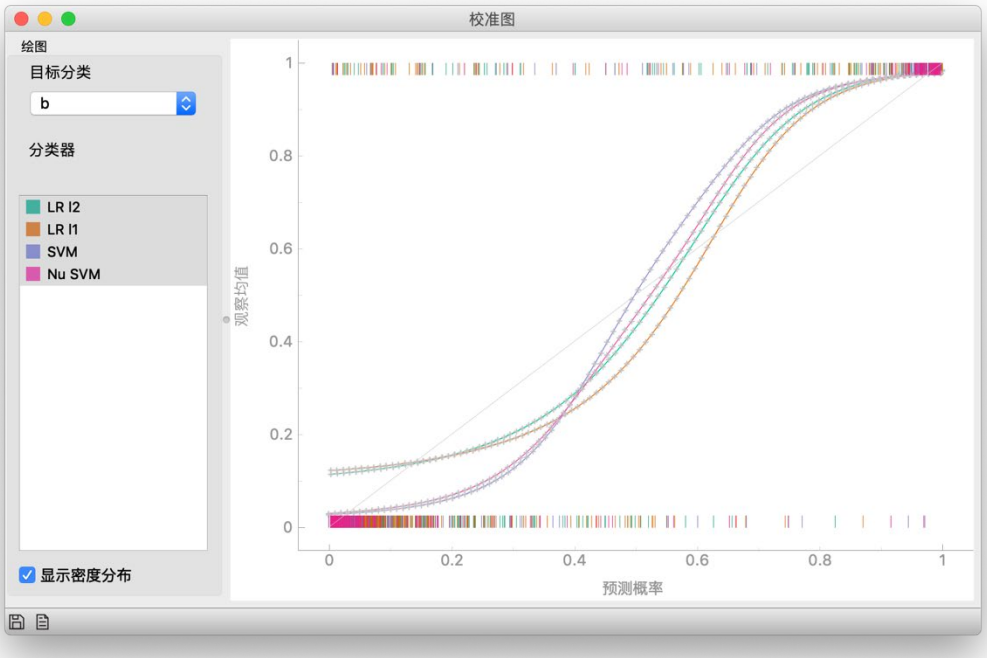
根据对分类器的评估构造展示一条 lift 曲线。



- 在“绘图”中设置分析的目标分类；
- 在“分类器”中选择分类模型。

2.4.6 校准图

根据对分类器的评估绘制的校准图。



- 在“绘图”中设置分析的目标分类；

- 在“分类器”中选择分类模型。

2.5 “无监督”功能节点组

- 距离文件：从一个文件中读取距离。
- 距离矩阵：查看距离矩阵。
- t-SNE：用 t-SNE 方法构造 2 维投影。
- 距离图：对距离矩阵进行可视化。
- 层次聚类：根据输入的距离矩阵构造一个层次聚类树状图。
- K 均值聚类：配合轮廓图进行聚类质量估算的 k 均值聚类算法。
- Louvain 聚类：在最近邻域网络上发现社区。
- 流形学习：非线性降维。
- 主成分分析：用陡坡图做主成分分析。
- 对应分析：对分类性多变量数据进行对应分析。
- 距离：计算距离矩阵。
- 距离转换：根据选中的条件转换距离。
- 多维尺度变换：通过距离矩阵，将多维数据投射到二维平面。
- 保存距离：将距离矩阵保存到一个外部文件。

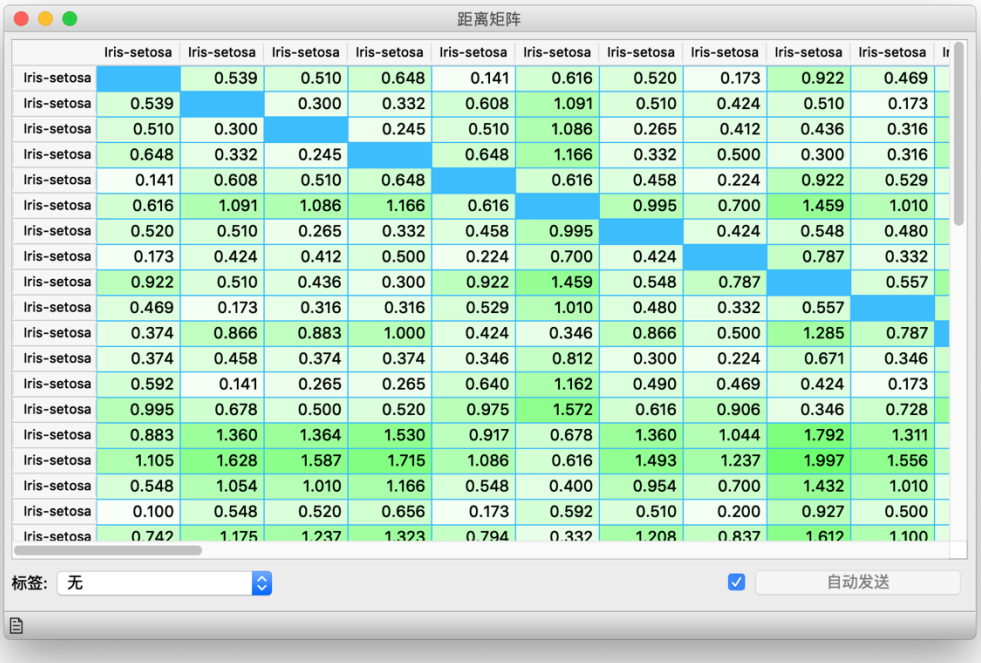
2.5.1 距离文件

从一个文件中读取距离。

- 在“距离文件”中查找本地路径，载入本地距离文件；
- 在“信息”中展示载入数据的基本情况。
- 可点击“浏览自带数据集”按钮，浏览智文自带数据集，并载入。

2.5.2 距离矩阵

查看距离矩阵。

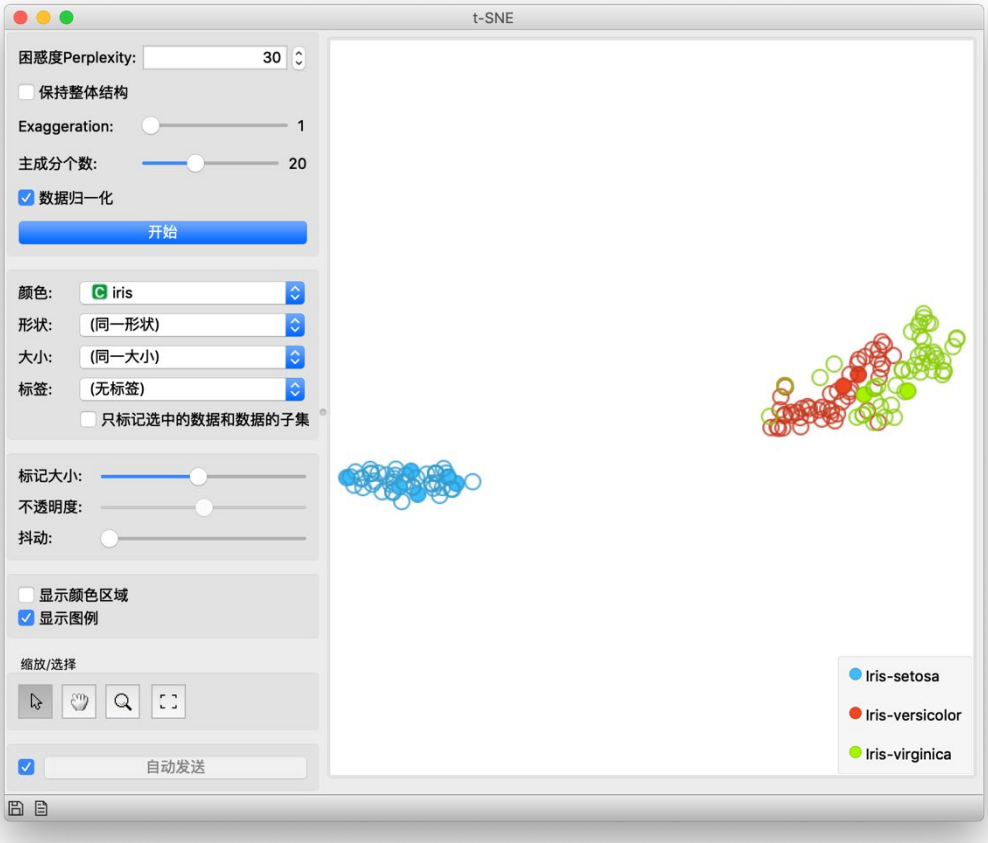


	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa
Iris-setosa		0.539	0.510	0.648	0.141	0.616	0.520	0.173	0.922	0.469
Iris-setosa	0.539		0.300	0.332	0.608	1.091	0.510	0.424	0.510	0.173
Iris-setosa	0.510	0.300		0.245	0.510	1.086	0.265	0.412	0.436	0.316
Iris-setosa	0.648	0.332	0.245		0.648	1.166	0.332	0.500	0.300	0.316
Iris-setosa	0.141	0.608	0.510	0.648		0.616	0.458	0.224	0.922	0.529
Iris-setosa	0.616	1.091	1.086	1.166	0.616		0.995	0.700	1.459	1.010
Iris-setosa	0.520	0.510	0.265	0.332	0.458	0.995		0.424	0.548	0.480
Iris-setosa	0.173	0.424	0.412	0.500	0.224	0.700	0.424		0.787	0.332
Iris-setosa	0.922	0.510	0.436	0.300	0.922	1.459	0.548	0.787		0.557
Iris-setosa	0.469	0.173	0.316	0.316	0.529	1.010	0.480	0.332	0.557	
Iris-setosa	0.374	0.866	0.883	1.000	0.424	0.346	0.866	0.500	1.285	0.787
Iris-setosa	0.374	0.458	0.374	0.374	0.346	0.812	0.300	0.224	0.671	0.346
Iris-setosa	0.592	0.141	0.265	0.265	0.640	1.162	0.490	0.469	0.424	0.173
Iris-setosa	0.995	0.678	0.500	0.520	0.975	1.572	0.616	0.906	0.346	0.728
Iris-setosa	0.883	1.360	1.364	1.530	0.917	0.678	1.360	1.044	1.792	1.311
Iris-setosa	1.105	1.628	1.587	1.715	1.086	0.616	1.493	1.237	1.997	1.556
Iris-setosa	0.548	1.054	1.010	1.166	0.548	0.400	0.954	0.700	1.432	1.010
Iris-setosa	0.100	0.548	0.520	0.656	0.173	0.592	0.510	0.200	0.927	0.500
Iris-setosa	0.742	1.175	1.237	1.323	0.794	0.332	1.208	0.837	1.612	1.100

- 展示两两之间的距离；
- 在“标签”选择显示标签。

2.5.3 t-SNE

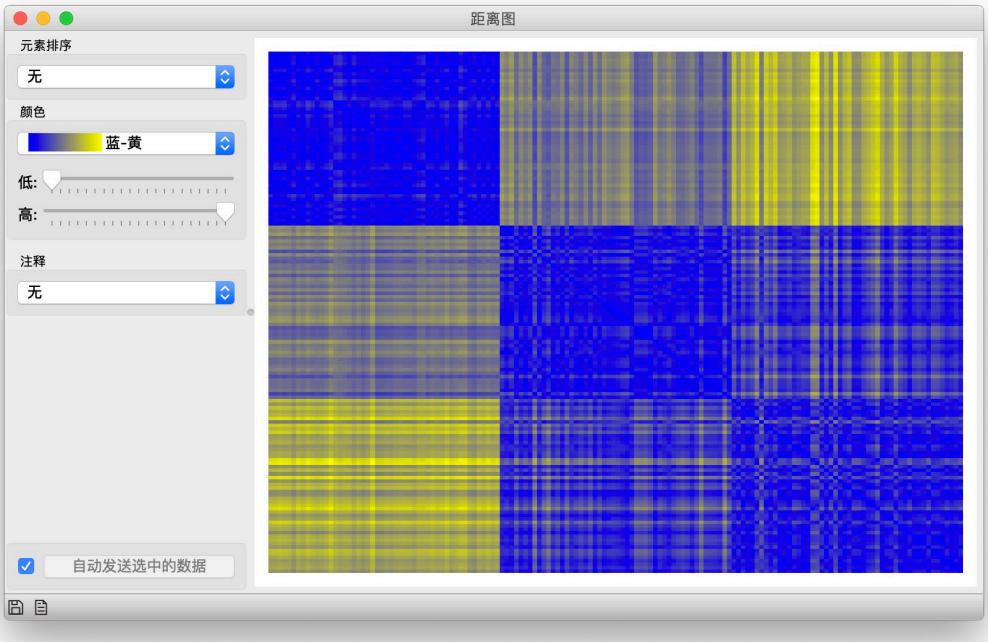
用 t-SNE 方法构造 2 维投影。



- 通过设置“perplexity”，“exaggeration”，“主成分个数”等参数设置 t-SNE 方法；

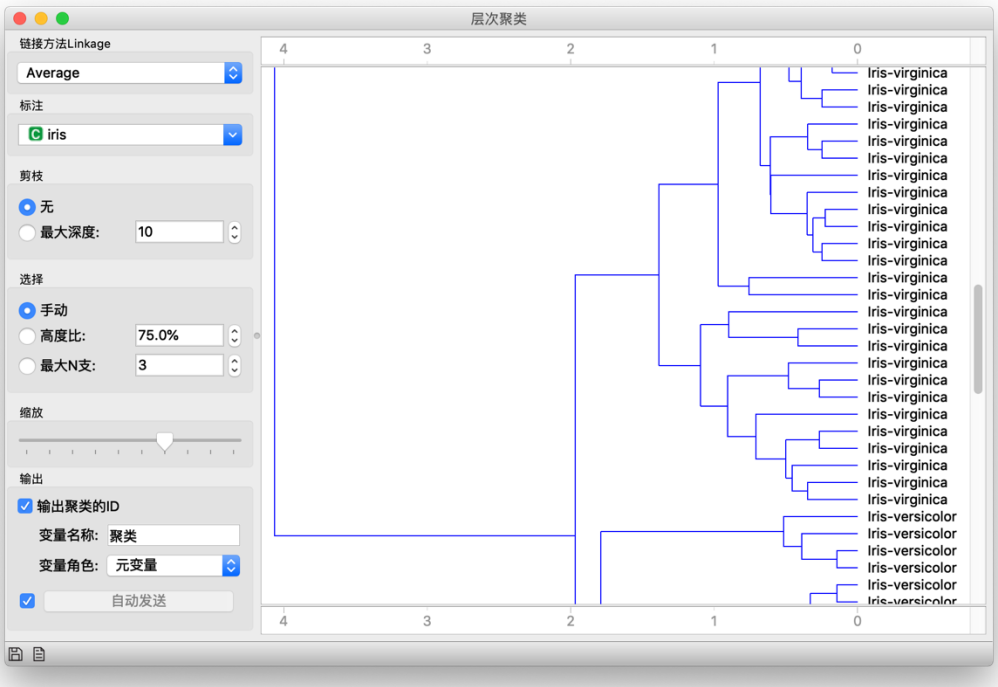
2.5.4 距离图

对距离矩阵进行可视化。



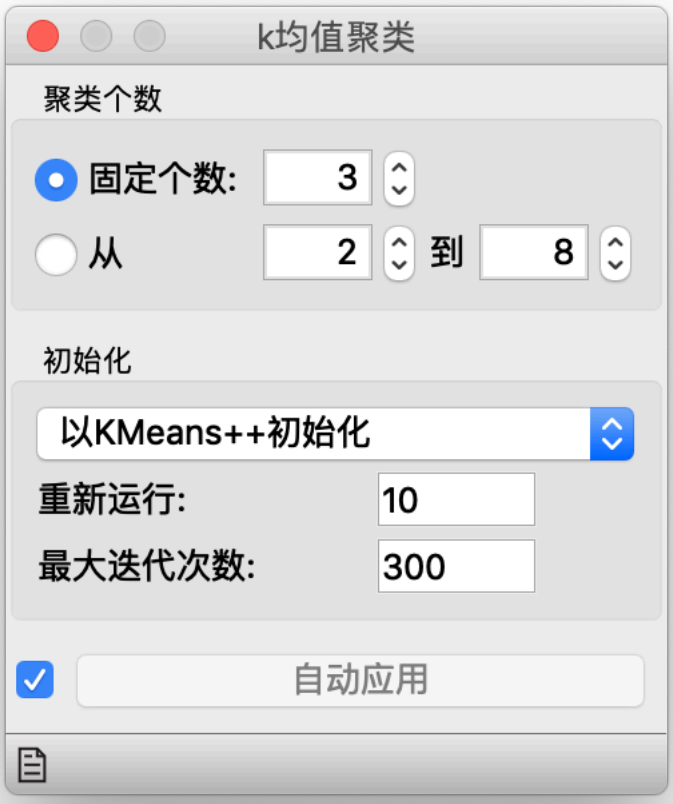
2.5.5 层次聚类

根据输入的距离矩阵构造一个层次聚类树状图。



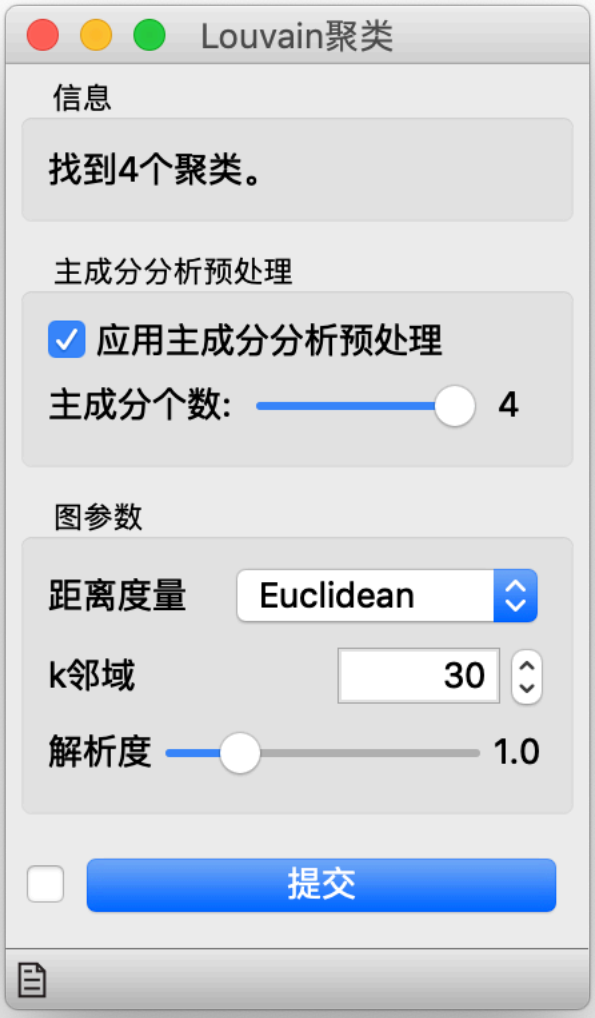
2.5.6 K 均值聚类

配合轮廓图进行聚类质量估算的 k 均值聚类算法。



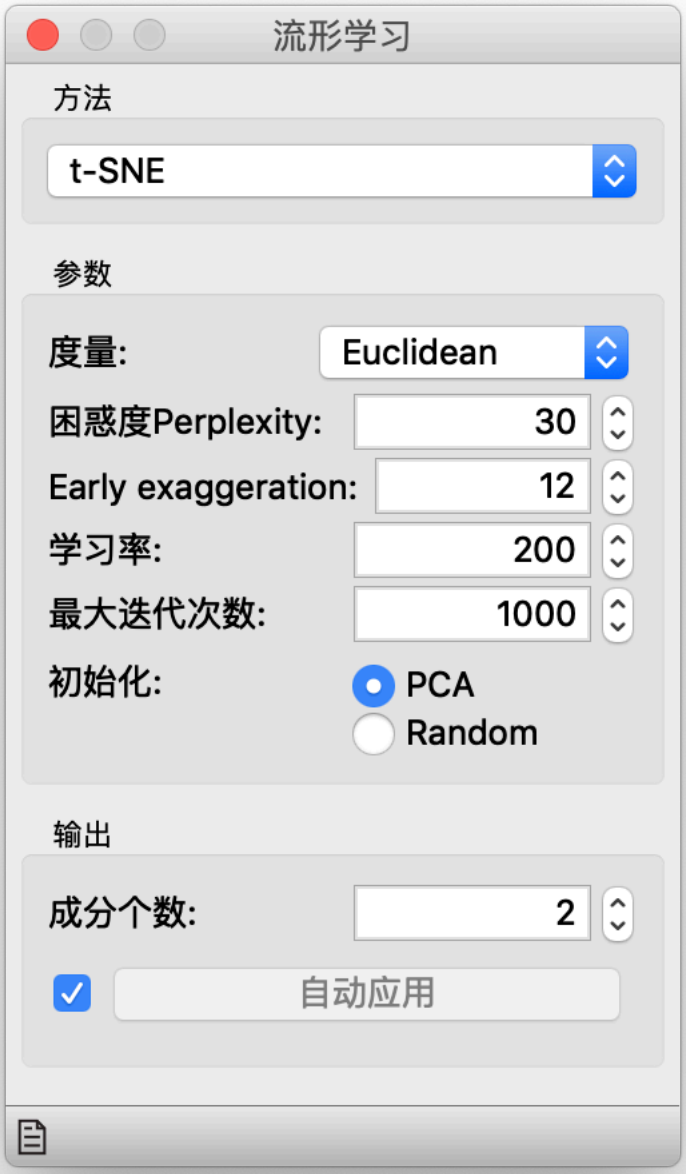
2.5.7 Louvain 聚类

在最近邻域网络上发现社区。



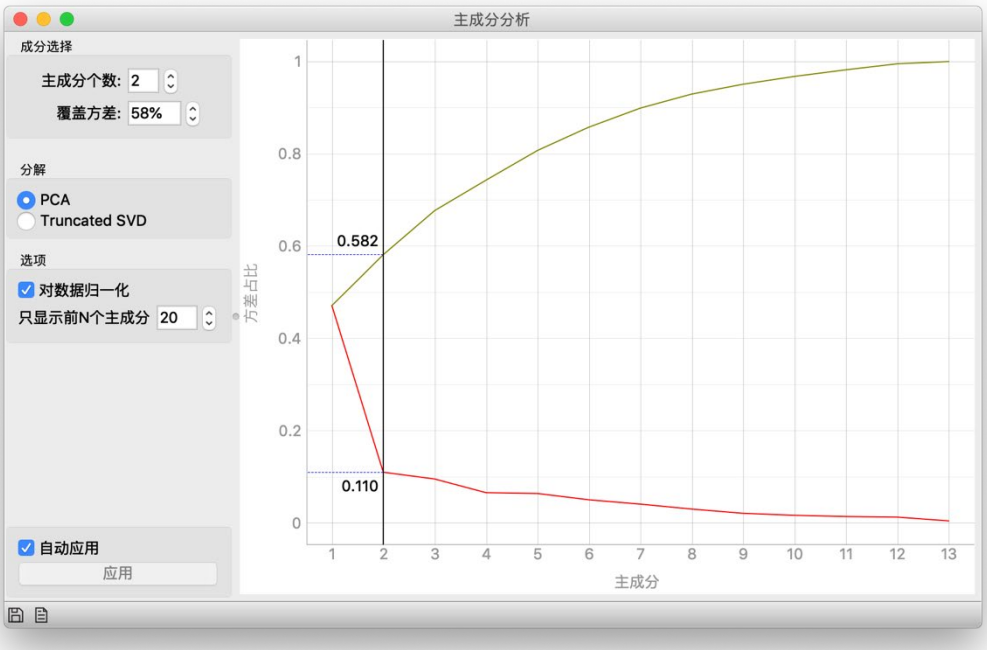
2.5.8 流形学习

非线性降维。



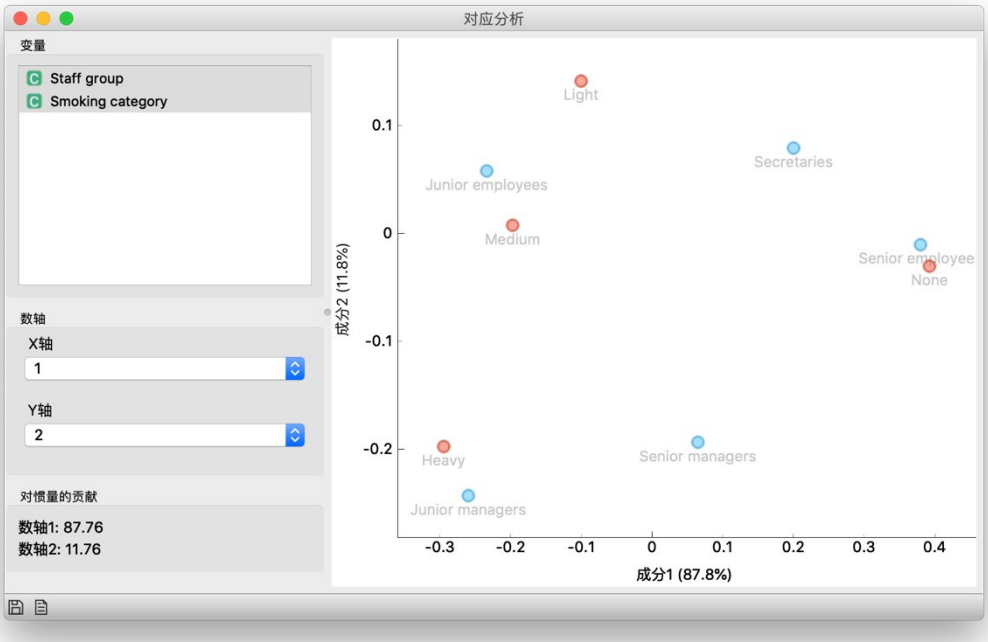
2.5.9 主成分分析

用陡坡图做主成分分析。



2.5.10 对应分析

对分类性多变量数据进行对应分析。

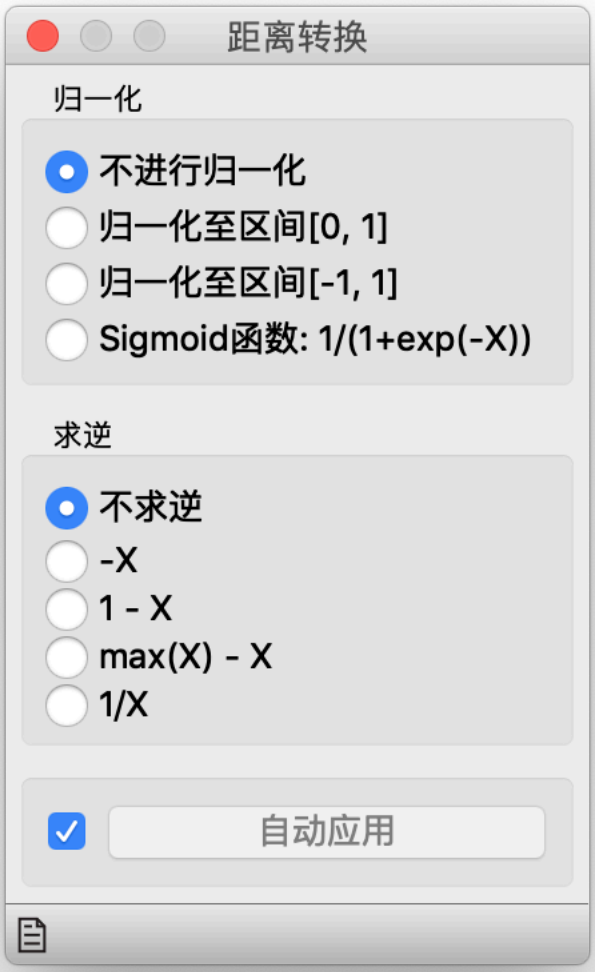


2.5.11 距离

计算距离矩阵。

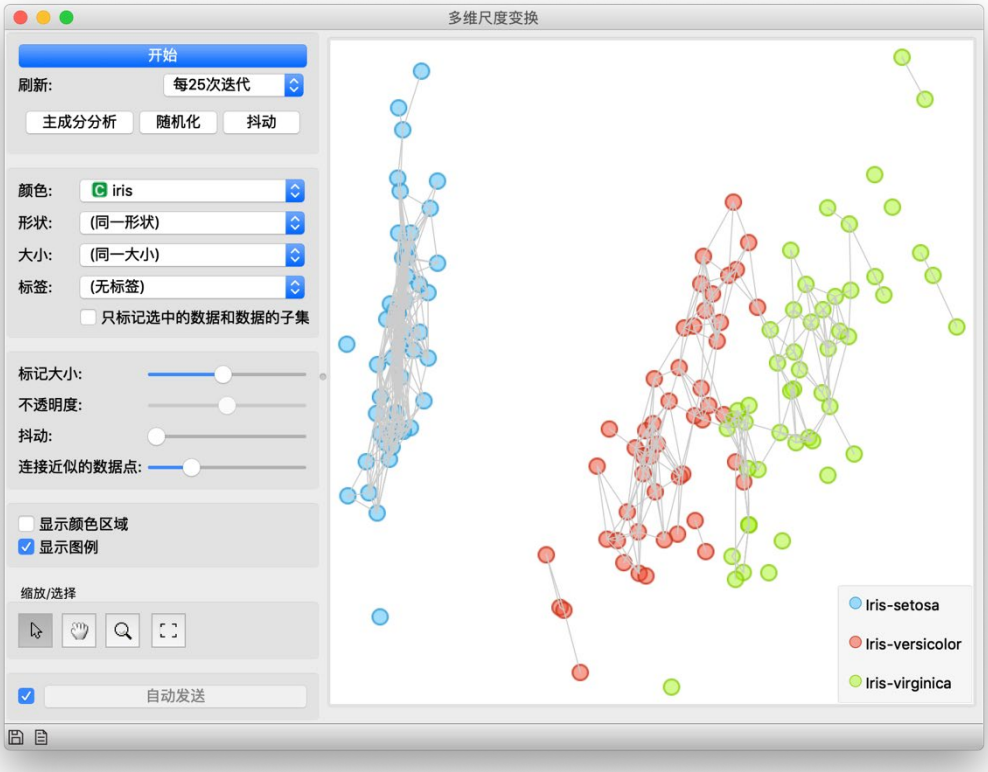
2.5.12 距离转换

根据选中的条件转换距离。



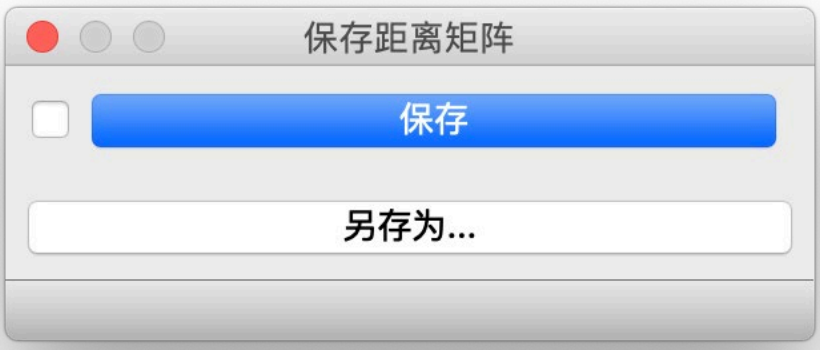
2.5.13 多维尺度变换

通过距离矩阵，将多维数据投射到二维平面。



2.5.14 保存距离

将距离矩阵保存到一个外部文件。



2.6 “文本挖掘”功能节点组

- 语料库：载入语料库中的文档。
- 导入文档：从文件夹中导入文字类文档。
- 卫报：通过卫报 API 抓取文章。
- 纽约时报：通过纽约时报 API 抓取文章。
- Pubmed：从 Pubmed 数据库下载数据。
- 推特：通过推特 API 载入推文。
- 维基百科：通过维基百科 API 载入百科词条。
- 中文文本预处理：创建中文文本预处理流程。
- 外文预处理：构造一个外文文字预处理流程。
- 词袋：将输入语料库生成为一个词袋数据结构。
- 相似性哈希：计算文档哈希。
- 情感分析：从文字预测情感。
- 推文分析器：用 Ekman，Plutchik 或 POMS 方法监测推文的感情。
- 主题模型：发掘语料的隐含主体结构。
- 查看语料库：查看语料库的内容。
- 词云：把词语根据频度制作成词云。
- 一致性：显示词语的上下文。
- 地图：在地图上显示文档来源。
- 词语富集：对选中的文档做词语富集分析。
- 文档查重：从一个语料库中监测并移除重复内容。

2.6.1 语料库

载入语料库中的文档。

2.6.2 导入文档

从文件夹中导入文字类文档。

2.6.3 卫报

通过卫报 API 抓取文章。

2.6.4 纽约时报

通过纽约时报 API 抓取文章。

2.6.5 Pubmed

从 Pubmed 数据库下载数据。

2.6.6 推特

通过推特 API 载入推文。

2.6.7 维基百科

通过维基百科 API 载入百科词条。

2.6.8 中文文本预处理

创建中文文本预处理流程。

2.6.9 外文预处理

构造一个外文文字预处理流程。

2.6.10 词袋

将输入语料库生成为一个词袋数据结构。

2.6.11 相似性哈希

计算文档哈希。

2.6.12 情感分析

从文字预测情感。

2.6.13 推文分析器

用 Ekman，Plutchik 或 POMS 方法监测推文的感情。

2.6.14 主题模型

发掘语料的隐含主体结构。

2.6.15 查看语料库

查看语料库的内容。

2.6.16 词云

把词语根据频度制作成词云。

2.6.17 一致性

显示词语的上下文。

2.6.18 地图

在地图上显示文档来源。

2.6.19 词语富集

对选中的文档做词语富集分析。

2.6.20 文档查重

从一个语料库中监测并移除重复内容。