永洪深度分析 Z-Advanced Analytics

用户使用手册

版权声明

免责声明

通讯方式

北京永洪商智科技有限公司

北京市朝阳区光华路 9 号光华路 SOHO 二期 C 座 9 层（100020）

电话：(86-10)-58430919

邮箱：public@yonghongtech.com

网站：http://www.yonghongtech.com

目录

第 1 章：Z-Advanced Analytics 简介

对数据的深度分析，Yonghong 是通过集成 R 来实现的。R 是开源、免费的统计语言，提供了多种统计，例如，线性和非线性建模、经典的统计检验、时间序列分析、分类、聚类 …… 在 Yonghong 中，通过调用 R 语言，针对不同的用户需求，例如业务人员和分析人员，提供不同的 R 使用方案。业务人员对 R 语言不了解，我们提供预定义的一些常用的操作（一元线性回归、LDA 线性分类、K-means 聚类和 HoltWinters 时序分析）中所用的 R Script。而分析用户是 R 的深度用户，我们也提供用户定制的 R 脚本。用户定制的 R 脚本中可以引用自带的 R 包中的函数（library 下），也可以调用第三方 R 包中的函数。

第 2 章：如何集成

产品自带 Rserve

RServe 默认随安装包一并发布，安装产品的过程中也会安装 RServe，并支持用户自己设置 R 的安装路径。

安装 Yonghong 产品之后，系统会自动在 bi.properties 中配置：r.serve.local=true 和 r.serve.path=RServe。在 Windows 系统中，会自动设置环境变量 PATH 和 R_HOME，PATH 指 R.dll 所在路径，R_HOME 指向整个 R 的安装包。当启动 YonghongBI 产品后，本地 RServe 会自动启动。用户也可以在管理系统 -> 系统设置中设置远程 R Serve 连接。

第三方 R 包

可以在 R 的控制台里通过 install.packages 命令导入，也可以在 RGui 中打开 R 的菜单栏 -> 程序包 -> “从本地 zip 文件安装程序包 ...” 导入。

如果用户已经装有 R，并且导入了第三方包，可以通过在 R 里安装 RServe 来提供 R 计算的服务。我们产品可以通过远程连接 RServe 来使用 R。

安装完包后，需要加载才能使用其中的函数，在定制 R 脚本中通过 require() 载入。如：

载入 MASS 包

require(MASS)

另：如果要调用其他地方的 R 脚本，可以通过 source （ ）调用。如：

调用 D 盘下的 test.R 脚本

source("D:\\test.R")

R 环境配置

如果使用第三方 R 包，用户需要进行 R 环境配置。

（1）已安装 R 软件。

（2）在环境变量 - 用户变量。

Path, 设置 R 的工作目录。例如：D:\Program Files (x86)\R\R-3.1.1\bin\x64

新建 R_HOME，设置为 R 的安装路径，例如：D:\Program Files (x86)\R\R-3.1.1

（3）下载 Rserve，解压后在 Rserve 启动目录（例如：“D:\Program Files (x86)\Rserve_1.8-0\Rserve\libs\x64”）中新建一个文件 “Rserv.cfg”，文件的内容如下：

port=6311

remote=enable

control=enable

encoding=utf8

（4）启动 Rserve。

（5）在产品 - 系统设置中设置 Rserve 连接。

第 3 章：使用深度分析

在产品中使用深度分析的地方有：

1. 在数据集上新建分析算法。在数据集的元数据界面上，右键选择 “ 新建分析算法 ” ，在弹出的 “ 分析算法 ” 窗口中输入名称，选择分析算法类型，选择并设置算法需要的数据列或属性值，或自定义 R 脚本，返回新的 R 字段（输出值）。根据分析算法生成 R 字段，作用域是当前数据集，所有使用该数据集的报表都能使用此数据集上的 R 字段。

2. 在报表组件绑定的数据集上新建分析算法。新生成的 R 字段作用域是当前报表，当前报表上的所有组件都可以使用此数据集上的 R 字段。

3. 图表组件上可以用四种分析算法进行快速绘图。支持一元线性回归、LDA 线性分类、K-means 聚类、HoltWinters 时序分析。把分析算法拖入图表区域，在弹出的 “ 分析算法 ” 窗口中选择并设置算法需要的数据列或属性值，点击 “ 确定 ” 之后，R 字段会自动绑定到图表上，绘制出用户期望的图形。

第 4 章：分析算法

一元线性回归

回归分析是一种应用非常广泛的统计工具，主要用来建立两个变量之间的关系模型。其中一个变量被称为自变量，其值是通过实验收集的。另一变量称为因变量，其值是根据自变量计算而得。线性回归这两个变量满足一个等式，其中这两个变量是指数（幂）相关。在数学上，一元一次线性关系的图形表示为直线。一元 N 次线性关系的图形表示为曲线。一元一次线性回归的一般数学方程为 y = ax+b，其中 y 是因变量，x 是自变量，a 和 b 称为系数常数。

【自变量】x，从下拉列表中选出需要作为自变量的字段。

【因变量】y，从下拉列表中选出需要作为因变量的字段。

【多项式次方】表示自变量与因变量是一个几次的线性方程关系，默认为 1，如果是 2 则表示一元二次方程。

【输出值】【拟合值】：被勾选时，会得到一个拟合值字段。其结果是得到相应的线性回归模型后，对给定的样本值 (x1,x2,...,xn) 做预测，也就是 (y1,y2,...,yn) 的估计值。

【输出值】【残差】：被勾选上时，会得到一个残差字段。其结果是 y 的实际值减去拟合值。

【输出值】【置信区间】：被勾选上时，根据 Level 算出估计值得到上界和下界，默认 Level 是 95%。

举例说明

对一系列身高和体重的值进行线性回归分析。收集一系列身高和体重的值，身高为自变量，体重为因变量，使用一元线性回归分析算法找出所创建模型的数学方程。根据数学方程，计算体重的拟合值。

如图在数据集上创建分析算法：

输出结果如图：

LDA 线性分类

LDA 线性分类 (Linear Discriminant Analysis, LDA) 的基本思想是将高维的模式样本投影到最佳鉴别矢量空间，以达到抽取分类信息和压缩特征空间维数的效果。投影后保证模式样本在新的子空间有最大的类间距离和最小的类内距离，即模式在该空间中有最佳的可分离性。LDA 线性分类通过分类标签列和训练集数据列，调用 lda 函数得到 lda 模型，根据模型对新样本数据进行预测分类。

【分类标签列】样本数据标签列，每个样本数据都有对应的标签值。

【训练集】样本数据列。从左侧的可配置列中选择需要作为训练集的字段直接拖入到训练集框中。

【新样本】新样本数据列。从左侧的可配置列中选择需要作为新样本的字段直接拖入到新样本框中。新样本的数据行数可以和训练集数据行数不一致。

【输出值】【分类结果】新样本数据的预测标签值。

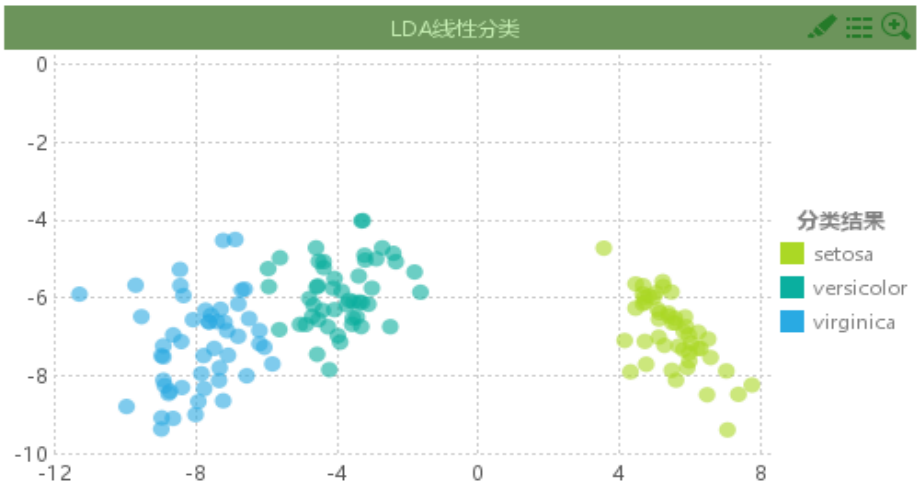
【输出值】【主成分】对分类的特征做主成分分析，取最重要的两个成分。

举例说明

已知三种花的样本数据，根据样本数据（花瓣和萼片的长宽值）建立合适的线性分类函数，根据线性分类函数对新样本进行预测分类。

在图表上创建 LDA 线性分类分析，如图：

分类结果如图：



K-means 聚类

K-means 聚类要指定聚类的分类个数 N，随机取 N 个样本作为初始类的中心，计算各样本与类中心的距离并进行归类，所有样本划分完成后重新计算类中心，重复这个过程直到类中心不再变化。

在 R 中使用 kmeans 函数进行 K-means 聚类，kmeans(data,centers=3,nstart=10)，

其中 centers 参数用来设置分类个数，nstart 参数用来设置取随机初始中心的次数，即运行 kmeans 方法的次数，我们在用 kmeans 函数时，默认取 10。

【聚类维度】聚类的样本集。从左侧的可配置列中选择需要作为聚类维度的字段直接拖入到聚类维度框中。

【设置 K 值】分类的个数。可以手动输入分类个数，也可以输入最大 K 值，系统根据轮廓系数计算出最佳的 K 值。

【输出值】【聚类标签】每个样本所属的类别。

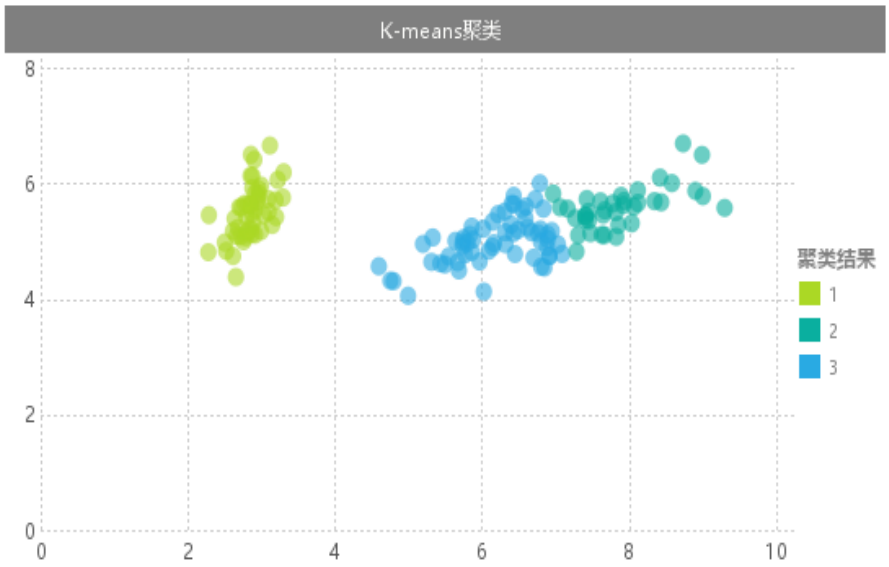
【输出值】【主成分】对聚类的维度做主成分分析，取最重要的两个成分。

举例说明

假设去掉类别列，根据四种属性对三种花进行分类。

在图表上创建 K-means 聚类分析，如图：

聚类结果如图：



HoltWinters 时序分析

HoltWinters 时序分析通过考虑水平趋势和季节性趋势，对一段时间内、等时间间隔的采样数据进行分析，以预测未来一段时间的数据。即根据已知的历史数据，预测未来的数据。

【时间列】选择时间字段。根据选择的时间字段的数据，自动算出时间间隔。

【数据列】选择数据字段。在报表组件绑定的数据集上新建分析算法，或使用图表的快速分析算法时，需要选择聚合函数，这样将按时间列进行分组，数据列进行聚合，在此分组聚合之后的数据上进行时序分析。

【周期】需填入时间间隔的整数倍，根据周期和时间间隔（周期 / 时间间隔）算出频率，即单位时间内的观测数。根据时间间隔，系统会自动往周期填入一个合理的数值，此数值也可手动修改。

【往后预测跨度】往后预测的时间跨度，需填入时间间隔的整数倍。选择时间列后，系统会自动填入一个合理的数值，此数值也可进行手动修改。

【趋势 (beta)】是否考虑水平趋势。默认是被勾选，表示按水平趋势拟合。

【输出值】预测值被勾选上时，表示会得出一个拟合值字段。置信区间被勾选上时，表示会得出一个上界和一个下界字段。

【季节因子 (gamma)】是否考虑季节性趋势。如果设置为不勾选（FALSE），则非季节性模型拟合。如果设置为勾选，则进行季节性模型拟合。季节性模式可以是加法效应（additive）和乘法效应（multiplicative）。加法效应默认勾选，表示按季节性加法的趋势增长。当乘法效应被勾选时，表示按

季节性乘法趋势增长。季节性模型拟合时，需满足一个周期内至少有两个数据点，即频率大于等于 2，且时间序列至少包含 2 个周期。

【输出值】【预测值】被勾选时，会得到一个拟合值字段。其结果是根据得到的模型，对往后预测的时间跨度做预测，算出预测值。

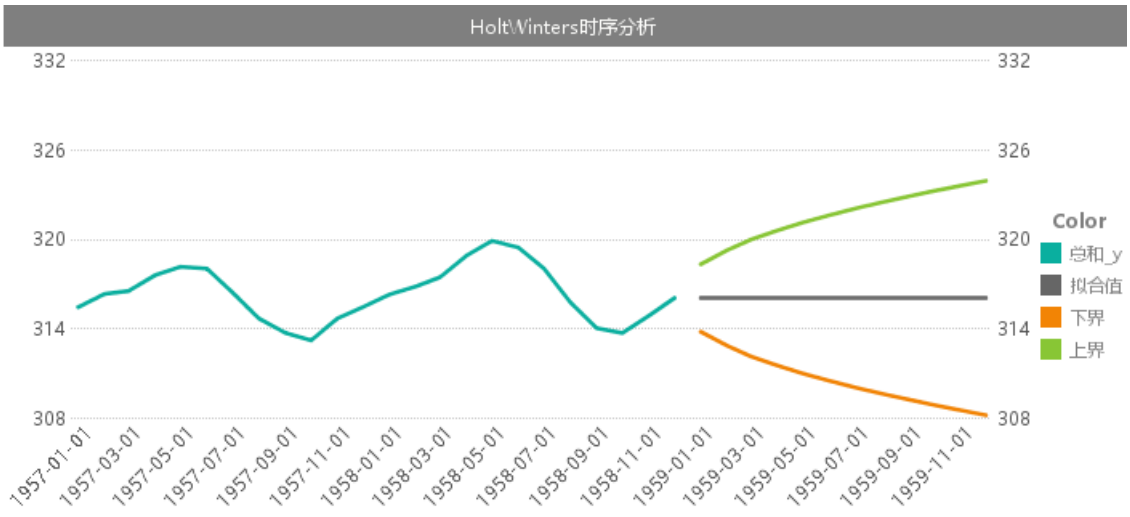
【输出值】【置信区间】被勾选上时，根据 Level 算出估计值的上界和下界，默认 Level 是 95%。

举例说明

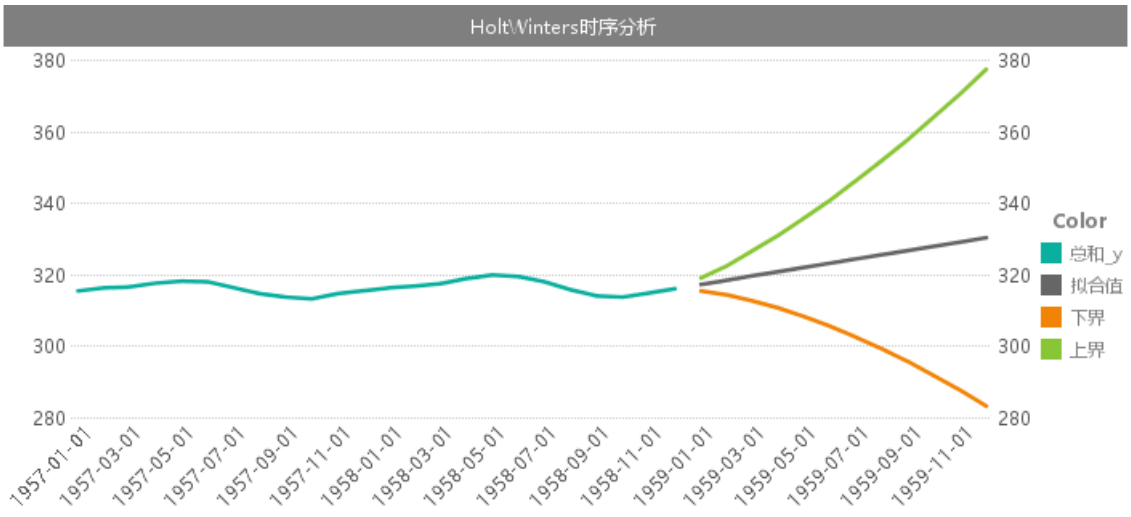
假设数据是从 1957 年 1 月到 1958 年 12 月的数据，时间间隔为 1 月，用 HoltWinters 时序分析，选择周期 6 个月，往后预测跨度 12 个月。

如果趋势（beta）选择否，季节因子（gamma）不勾选：

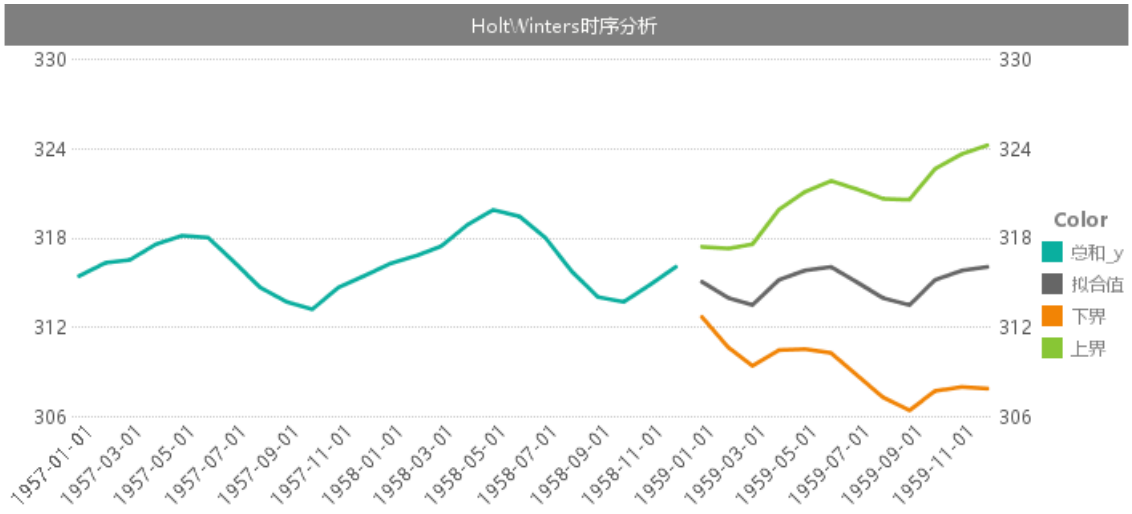
预测结果如图：



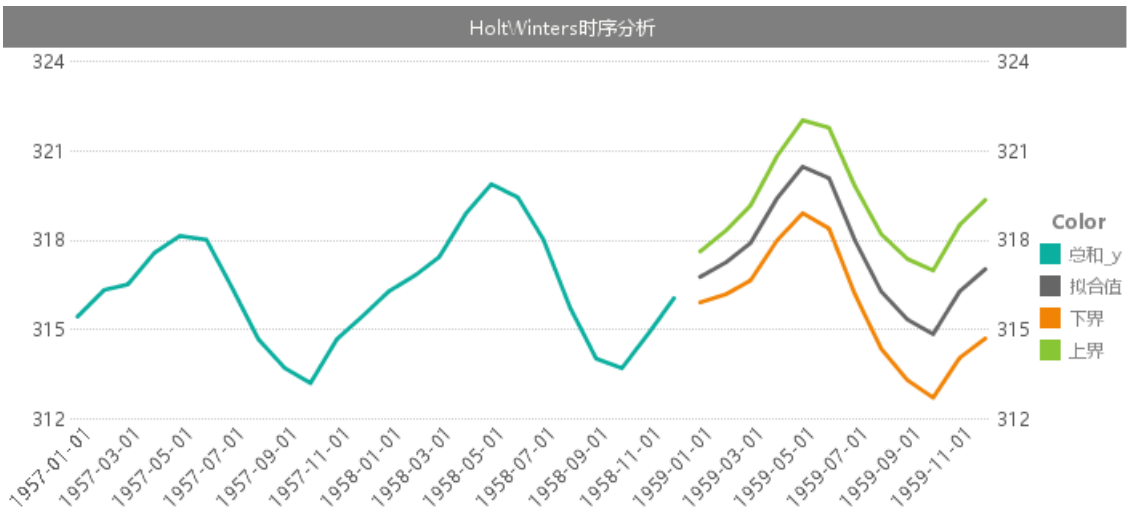
如果趋势（beta）选择是，季节因子（gamma）不勾选：



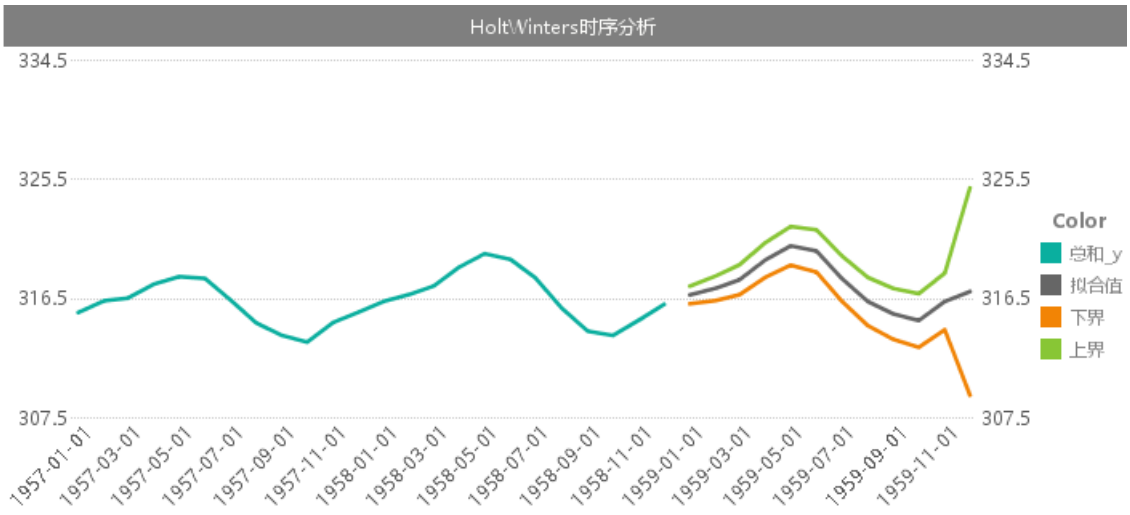
如果趋势（beta）选择否，季节因子（gamma）勾选，选择加法效应：



如果趋势（beta）选择是，季节因子（gamma）勾选，选择乘法效应：



如果趋势（beta）选择是，季节因子（gamma）勾选，选择加法效应：



定制

【计算类型】在连接数据的元数据区，默认细节计算处于置灰状态。在组件绑定数据集，默认细节计算处于选中状态。其下拉框中有细节计算和聚合计算，细节计算和聚合计算的区别是：聚合计算出来的 R 字段为聚合字段。

【脚本】输入脚本内容。

可以通过 col[[“xxx”]] 来传入数据集中对应列的值 ,xxx 为列的名称；也可以通过 param[[“xxx”]] 来传入参数值，xxx 为参数名称。

对于定制脚本，R 将最后执行的代码行的结果作为返回值返回。Yonghong 产品中要求返回值必须是 list 对象，包含若干返回值列，如 list(out1=a, out2=b)，其中 out1,out2 为返回值列的名称，而 a，b 为相应返回值列的取值，可以是常数或向量。

其他 R 脚本请参考 R 官网。

举例说明


分析算法
名称(N)：
定制1
分析算法(A)：
定制
计算类型(L)：
细节计算
脚本(S)：
1 a<-col[['height']];
2 b<-col[['weight']];
3 c<-a
4 d<-b/c
5 e<-c*param[['a']];
6 list(out1=e,out2=d)
收集输出值(U)
输出值：
名称
数据类型
out2
双精度浮点数
out1
整数
名称
确定(O)
取消(C)

第 5 章：动态更新 R 脚本生成的图片

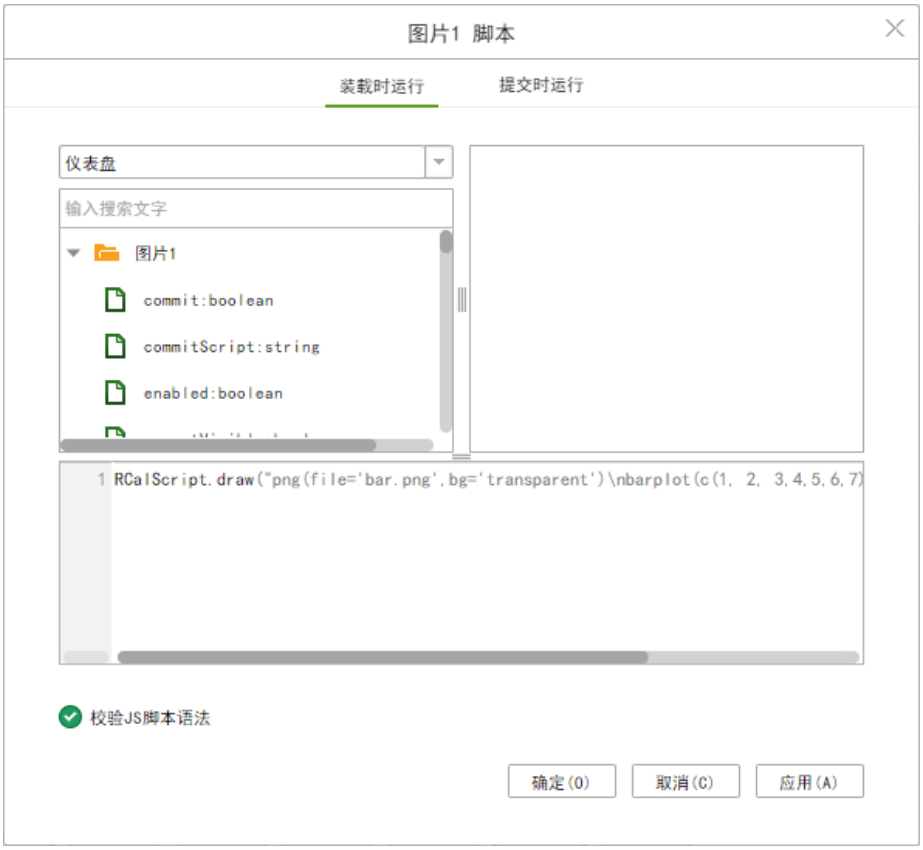
R 语言有丰富的绘图函数，系统为用户提供使用 R 脚本绘图的接口。通过 js 脚本在后台运行 R 脚本以生成图片，用户可通过 Image 组件动态加载生成的图片。

【步骤】

1. 使用 RCalScript 定义 Scriptable 对象，有如下方法：

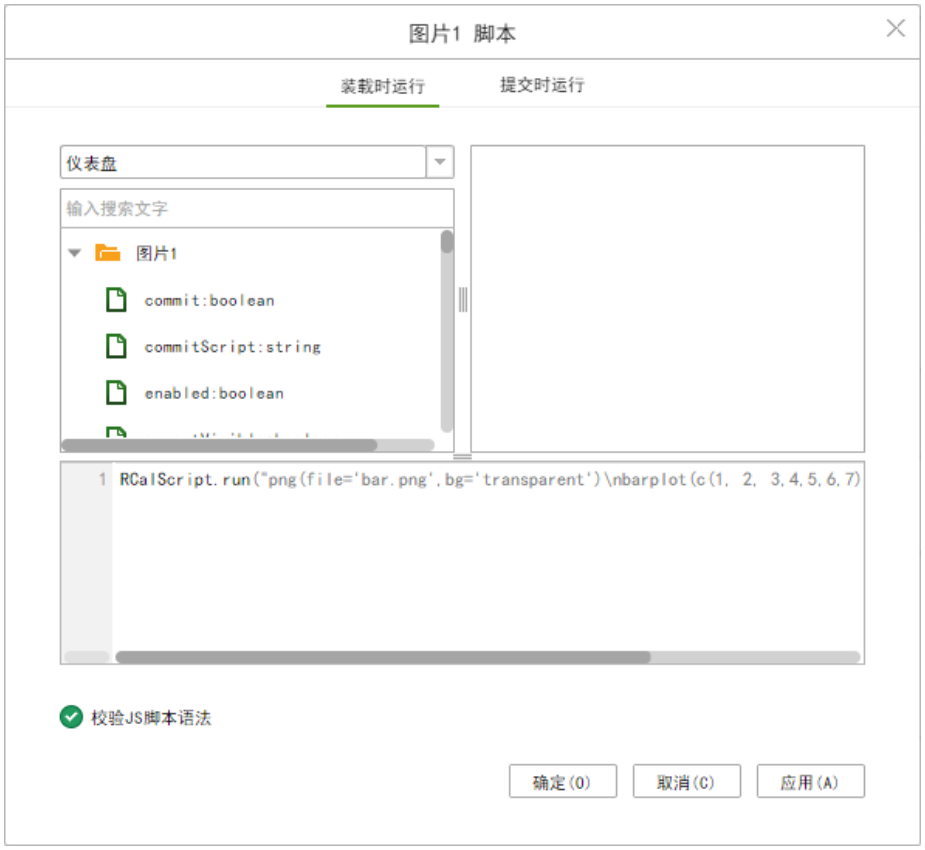
（1）draw(String script, String path, boolean refresh):

其中 script 指的是 R 脚本； path 指的是图片路径（包括名称、后缀，图片默认保存在 bihome/image 下）； refresh 指的是是否刷新，如果是则每次都生成新的图片，如果否则只在图片不存在的情况下生成图片。如果 R 脚本绘制失败，前台返回出错原因。



（2）run(String script)：

其中 script 指的是 R 脚本。如果 R 脚本绘制失败，前台返回出错原因。



2. 执行完绘图方法后，图片保存路径下就有相应的图片资源，在图片组件的属性对话框中刷新导入即可。

【图片类型的支持】

产品目前支持三种类型：png 格式，jpg 格式，bmp 格式。