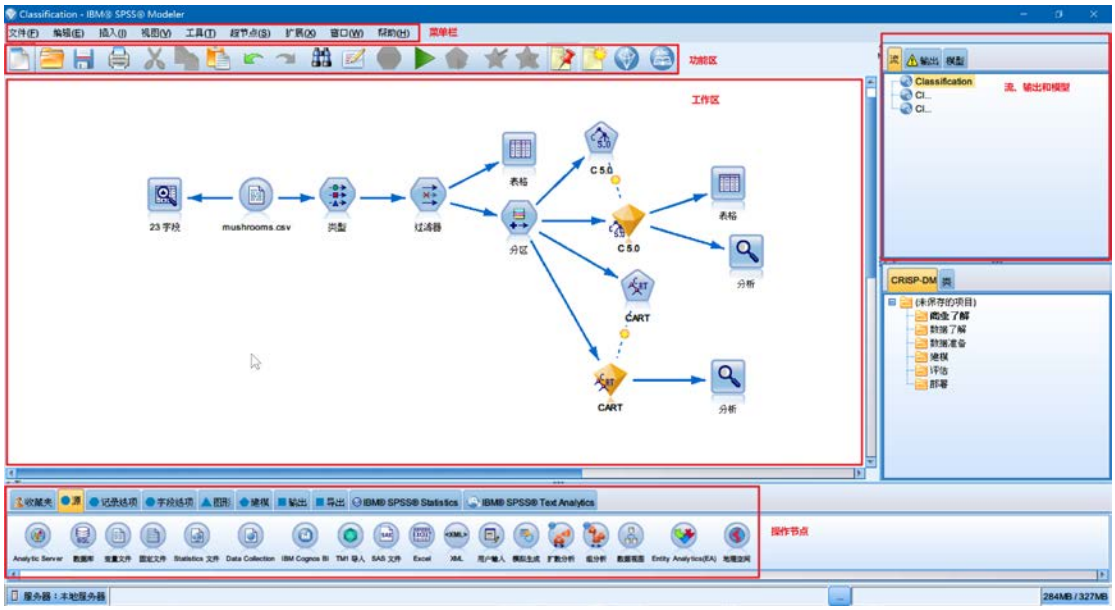
在 SPSS 中应用分类和聚类算法

本次实践分别用两个数据集来进一步学习如何在 SPSS 中对两个数据集进行分类和聚类。两个数据集均来自 Kaggle 网站的 UCI Machine Learning。

一、 蘑菇数据集

蘑菇数据集来自于 Mushroom Classification | Kaggle，该数据集总数据量为 8124 条，涉及特征包括菌盖形状、菌盖表面、气味等等 22 个字段，目标是对蘑菇是否有毒进行二分类，即有毒、无毒。所有的特征都是离散分类特征，在 Kaggle 上可以看到各特征的具体含义和属性取值。

本次实践使用的软件是 IBM SPSS Modeler 18.0，其主界面如下：



在本次实践过程主要涉及两个区域：工作区和节点选项板。工作区放置操作节点和数据流，节点选项板则提供了不同功能的节点。首先介绍两个基本概念：节点和数据流。SPSS Modeler 进行的数据挖掘重点关注通过一系列节点运行数据的过程，我们将这一过程称为数据流。也可以说 SPSS Modeler 是以数据流为驱动的产品。这一系列节点代表要对数据执行的操作，而节点之间的链接指示数据的流动方向。通常，SPSS Modeler 将数据以一条条记录的形式读入，然后通过对数据进行一系列操作，最后将其发送至某个地方（可以是模型，或某种格式的数据输出）。使用 SPSS Modeler 处理数据的三个步骤：

1. 将数据读入 SPSS Modeler。
2. 通过一系列操纵运行数据。
3. 将数据发送到目标位置。

在 SPSS Modeler 中，可以通过打开新的数据流来一次处理多个数据流。会话期间，可以在 SPSS Modeler 窗口右上角的流管理器中管理打开的多个数据流。

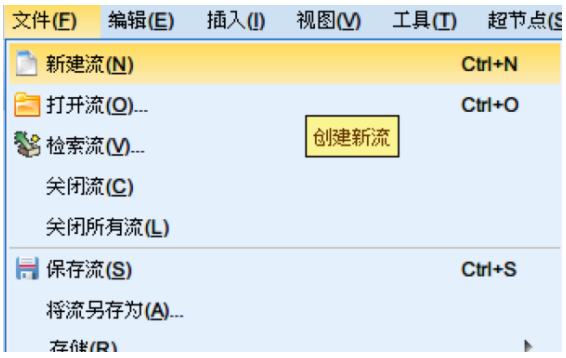
在节点选项板中，每个选项板选项卡均包含一组不同的流操作阶段中使用的相关节点，如：

- **源** 此类节点可将数据导入 SPSS Modeler，如数据库、文本文件、SPSS Statistics 数据文件、Excel、XML 等。
- **记录选项** 此类节点可对数据记录执行操作，如选择、合并和追加等。
- **字段选项** 此类节点可对数据字段执行操作，如过滤、导出新字段和确定给定字段的测量级别等。
- **图形** 此类节点可在建模前后以图表形式显示数据。图形包括散点图、直方图、网络节点和评估图表等。
- **建模** 此类节点可使用 SPSS Modeler 中提供的建模算法，如神经网络、决策树、聚类算法和数据排序等。
- **数据库建模** 节点使用 Microsoft SQL Server、IBM DB2 和 Oracle 数据库中可用的建模算法直接在数据库里进行建模及评估。
- **输出** 节点生成数据、图表和可在 SPSS Modeler 中查看的模型等多种输出结果。
- **导出** 节点生成可在外部应用程序（如 IBM SPSS Data Collection 或 Excel）中查看的多种输出。
- IBM SPSS Statistics 节点将 IBM SPSS Statistics 数据导入或导出为 SPSS Statistics 数据，以及运行 SPSS Statistics 提供的功能。

接下来我们开始对数据集进行操作：

第一步 创建流

在左上角菜单栏选择“文件”→“新建流”，创建一个新的数据流。

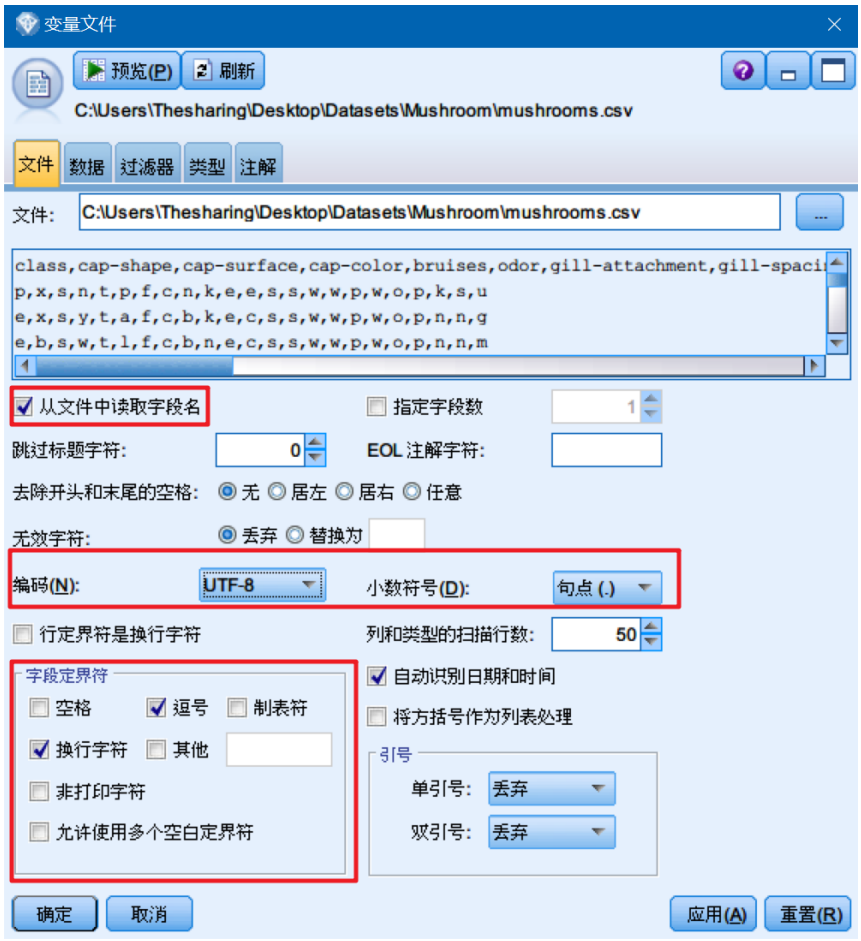


第二步 导入数据

在下方节点选项板中选择“源”→“变量文件”，将其拖入工作区。这里“变量文件”是指从分隔的列文本文件中读取数据，正好我们下载的源数据文件为 csv 文件，一般以逗号（,）为分隔符。

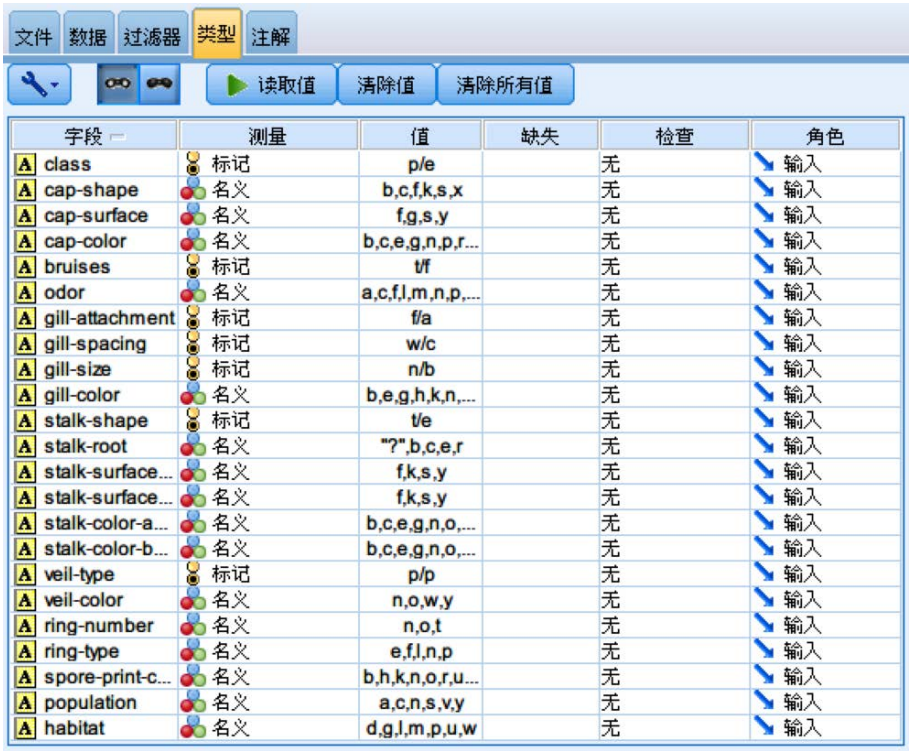
双击打开“变量文件”对话框，选中之前下载好的数据文件，然后对数据格式进行设置。在这里主要说以下几点：

1. 如果文件第一行是字段名，则选中“从文件中读取字段名”，SPSS 会自动提取。
2. 如果文件显示乱码，则在“编码”位置修改成正确的编码。
3. 一些 CSV 不是以逗号为分隔符的，可以在“字段定界符”中选择相应的符号。

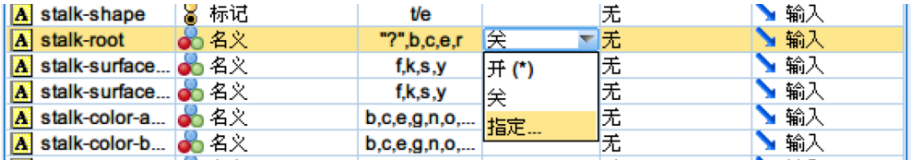


完成基础设置后可以在“数据”选项卡中查看各字段的类型。在“过滤器”中可以选择过滤掉一部分字段，例如下图所示，将“Stalk-root”这个字段过滤掉，只剩 22 个字段。值得注意的是，在节点选项板中也有“过滤”操作，为了显式说明，我们将在后面再设置过滤，因此这里直接跳过即可。

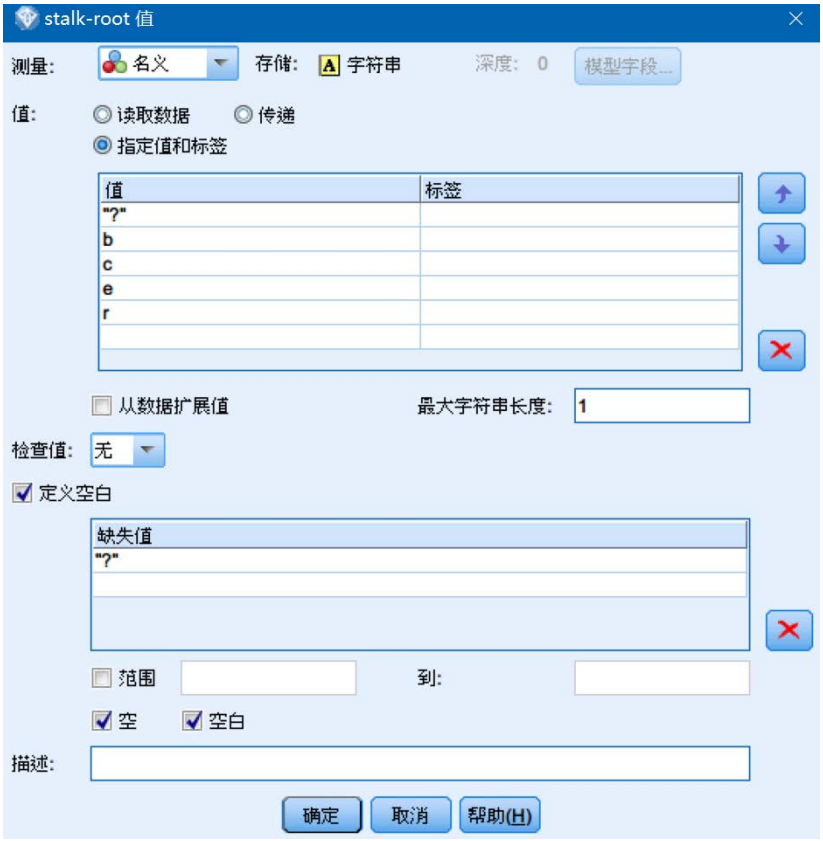
在“类型”选项卡中可以设置各字段的类型和角色，和“过滤器”类似，在节点选项板中可以选择“类型”操作。这里点击中间的“读取值”即可自动识别所有字段的类型和取值范围，如下图所示。



在图中我们可以看到，有两个字段存在异常，一是“stalk-root”中有一个问号（?），二是“veil-type”中显示为 p/p。查阅 kaggle 的字段含义可知，“stalk-root”的问号（?）是指缺失值，因此这里我们为该字段设置缺失值。在该字段的“缺失值”这一栏选择“指定…”。



然后在弹出对话框中点击“定义空白”，在下方输入英文问号（?），点击“确定”，即设定好缺失值。



我们还可以为各数据添加自定义标签，在“class”的“值”一栏选择“指定…”。

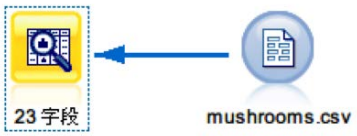
然后在值这一栏根据 kaggle 上对数据的描述，写上标签内容，点击“确定”。

点击“确定”即可以完成对输入数据的设置，你还可以点击“预览”预览前

10 条数据。

第三步 数据审核

在节点选项板中选择“输出”→“数据审核”并拖放至工作区，然后右键点击工作区中的“mushroom.csv”，选择“连接”，然后点击“数据审核”，即可以将数据流引向数据审核节点。如下图所示：



再在“数据审核”节点上点击右键选择“运行”，即可以打开“数据审核”窗口，在这里我们可以看到有三个选项卡，“审核”选项卡中包含各字段的值分布、类型、最小值/最大值等等属性。我们重点来关注三个字段的分布。

双击各柱状图即可查看详情，我们可以看到，对于类别来说，有毒/无毒的分布基本对半，说明类别平衡，不需要进一步进行处理即可用于预测。

值	比例	%	计数
e		51.8	4208
p		48.2	3916

然后是之前看到的两个字段，“stalk-root”和“veil-type”。先看 veil-type，可以看到 p 占了全部，也就是说该字段没有任何信息量，因此之后会删掉。

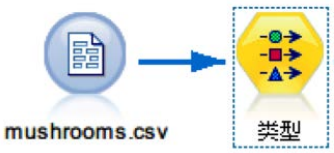
值	比例	%	计数
p		100.0	8124

然后我们看“质量”选项卡，包括了离群值、极值、空值、完成度等等数据质量。我们可以看到，“stalk-root”的完成度只有 69%，有近 2480 个我们定义的“空白值”，因此该字段也需要删掉。

利用数据审核我们可以查看数据哪些字段需要做预处理。

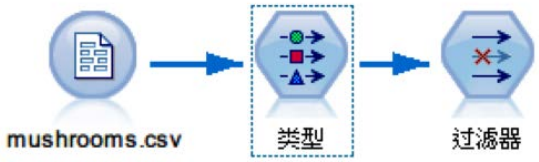
第四步 数据预处理

首先我们定义数据各字段的类型。在节点选项板中选择“字段选项”→“类型”，拖入工作区并和“mushroom.csv”连接起来（右键菜单中）。



然后双击编辑属性，在其中将“class”的角色设置为“目标”，即我们预测的目标值，如下图所示：

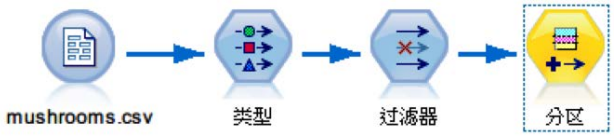
然后点击“确定”。再在节点选项板中选择“字段选项”→“过滤器”，将之前说的两个字段删除。将过滤器拖至工作区并和“类型”连接起来。



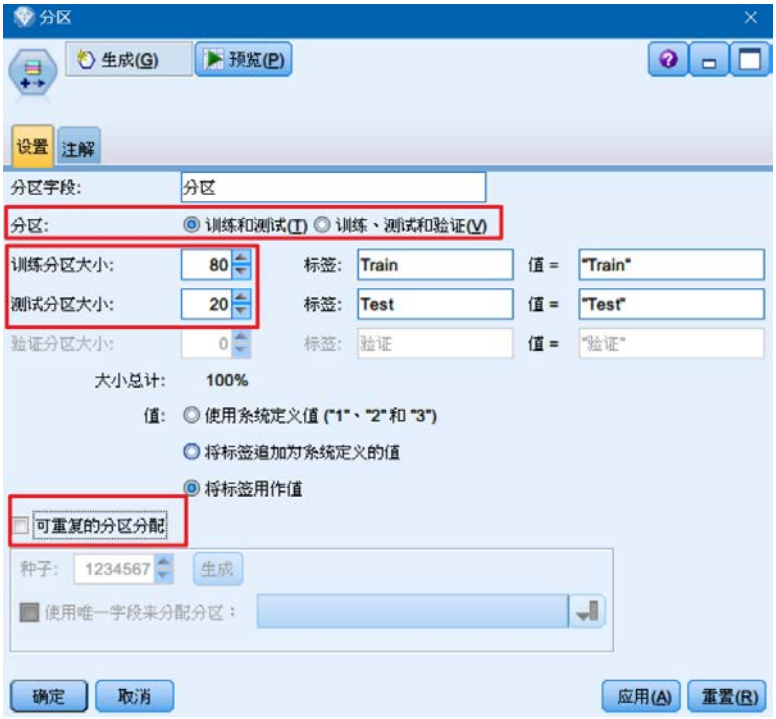
双击“过滤器”，在其中找到“stalk-root”和“veil-type”，点击中间的箭头，过滤掉这两个字段。

第五步 数据分区

数据在训练前需要分为训练集和测试集，在训练集上进行训练，然后测试集上进行测试。我们在节点选项板中选择“字段选项”→“分区”，将其拖至工作区并和“过滤器”连接起来。



在界面中主要注意以下三点：



首先，选择“训练和测试”，也就是说我们只分训练集和测试集。然后设置训练分区大小和测试分区大小。二者之和不能大于 100%，但是可以小于。如果小于 100%，数据集会被随机丢掉一部分数据。这里我们设置为 80%和 20%。接下来勾掉“可重复的分区分配”，勾选上则表示一条数据可以重复出现于两个分区中，在这里我们选择两个分区数据是不重复的。点击确定即分好两个数据集。

第六步 训练和测试模型

在本次实践中我们将对两个决策树算法进行对比：C5.0 和 CART 树。

C5.0 模型的工作原理是根据提供最大信息增益的字段分割样本。然后通常会根据不同的字段再次分割由第一次分割定义的每个子样本，且此过程会重复下去直到无法继续分割子样本。最后，将重新检查最底层分割，并删除或修剪对模型值没有显著影响的分割。

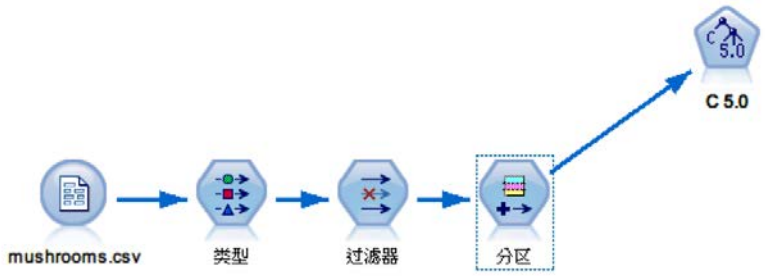
C5.0 可以生成两种模型。决策树是对由算法建立的分割的简单描述。每个终端（或“叶”）节点可描述训练数据的特定子集，而训练数据中的每个观测值都完全属于树中的某个终端节点。换句话说，对于在决策树中显示的任何特定数据记录，仅可能有一个预测。

反过来， 规则集则是尝试对单个记录进行预测的一组规则。规则集源自决策树，并且在某种程度上表示在决策树中建立的经简化或提取的信息版本。通常，规则集可保留完整的决策树中的大部分重要信息，但其使用的模型比较简单。由于规则集的这种工作方式，其属性与决策树的属性不同。最重要的区别是，使用规则集时，可以为任意特定记录应用多个规则，也可以不应用任何规则。如果应

用多个规则，则每个规则将根据与此规则关联的置信度获得一个加权“投票”，并通过组合应用到所讨论记录的所有规则的加权投票来确定最终的预测。如果没有规则可应用，则会将缺省预测分配到该记录。

CART 树是指分类与回归树（Classification and Regression Tree），在 SPSS 中称为 C&T 树。分类和回归 (C&R) 树节点是一种基于树的分类和预测方法。与 C5.0 类似，此方法可使用递归分区将训练记录分割为具有相似输出字段值的段。首先，“C&R 树”节点通过检查输入字段来查找最佳分割（以分割所引起的杂质指标下降情况进行测量）。分割可定义两个子组，其中每个子组随后又被分割为两个子组，依此类推，直到触发其中一个停止标准为止。所有分割都是二元的（仅有两个子组）。

在节点选项板中选择“建模”→“分类”→“C5.0”，然后拖至工作区并与“分区”连接。



双击“class”编辑模型属性。

各设置项的含义：

- **模型名称** 指定要生成的模型的名称。

- **自动** 在选中此选项的情况下，将根据目标字段名称自动生成模型名称。这是缺省选项。
- **定制** 选中此选项可以为此节点将创建的模型块指定定制名称。

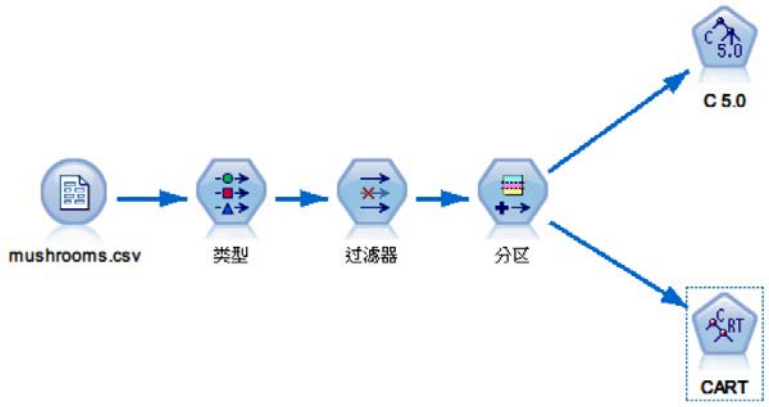
- **使用分区数据** 如果定义了分区字段，那么此选项可确保仅训练分区的数据用于构建模型。
- **创建分割模型** 给指定为分割字段的输入字段的每个可能值构建一个单独模型。有关更多信息，请参阅构建分割模型主题。
- **输出类型** 在此处指定您希望生成的模型块是决策树还是规则集。
- **组符号** 如果选中了此选项，那么 C5.0 将尝试对输出字段具有相似模式的符号值进行组合。如果未选中此选项，C5.0 将为用于分割父节点的符号字段的每个值创建一个子节点。例如，如果 C5.0 分割的是 颜色字段（其值为 红色、 绿色和 蓝色），则它将缺省创建一个三向分割。但是，如果选中此选项，且 颜色 = 红色的记录与 颜色 = 蓝色的记录非常相似，则 C5.0 将创建一个双向分割，其中所有 绿色记录在一个组中，而所有 蓝色记录连同所有 红色记录在另一个组中。
- **使用 boosting** C5.0 算法包含一个用于提高其准确率的特殊方法，称为增强。它的工作原理是在序列中构建多个模型。第一个模型按常规方式进行构建。构建第二个模型时，将焦点集中于由第一个模型误分类的记录。构建第三个模型时，将焦点集中于第二个模型的错误，依此类推。最后，通过将整个模型集应用到观测值，并使用加权投票过程将单独的预测组合为一个总预测来分类观测值。增强方法可以显著提高 C5.0 模型的准确性，但也需要更长的训练时间。通过尝试次数选项，您可以控制用于增强型模型的模型数。
- **交叉验证** 如果选中此选项，那么 C5.0 将使用一组根据训练数据的子集构建的模型来估算根据整个数据集构建的模型的准确性。如果数据集太小以致于无法将其分割为传统的训练集合和测试集合，此选项非常有用。在计算准确性评估后，交叉验证模型将被丢弃。可以指定用于交叉验证的 折叠次数 或模型数。注意，在 IBM® SPSS Modeler 以前的版本中，构建模型和交叉验证模型是两个单独的操作。在当前的版本中，则无需执行单独的模型构建步骤。模型构建和交叉验证将同时执行。
- **方式** 对于简单训练，大多数 C5.0 参数是自动设置的。专家训练允许更直接地控制训练参数。
- **简单模式选项**
  - **偏向** 缺省情况下，C5.0 将尝试尽可能生成最准确的树。在某些情况下，此操作可能会导致过度拟合，从而在将此模型应用于新数据时导致性能偏低。选择普遍性以使用受此问题影响较小的算法设置。
  - **预期的噪声 (%)** 指定训练集合中噪声数据或错误数据所占的预期

比例。

- **专家模式选项**
  - **修剪严重性** 确定决策树或规则集的修剪程度。增加该值可获得一个更简洁的小型树。减小该值可获得一个更精确的树。此设置仅影响本地修剪（请参见下面的“使用全局修剪”）。
  - **每个子分支的最小记录数** 可以使用子组的大小来限制树的任何分支中的分割数。仅当两个或多个生成的子分支中至少包含从训练集合得到的这一最小记录数时，才可分割树的分支。缺省值为 2。增加该值有助于防止使用噪声数据进行 过度训练 。
  - **使用全局修剪** 树的修剪分为两个阶段：第一个阶段是本地修剪，将检查子树并折叠分支以提高模型的准确性。第二个阶段是全局修剪，在此阶段中将把树视作一个整体并折叠虚弱的子树。缺省情况下将执行全局修剪。要忽略全局修剪阶段，请取消选中此选项。
  - **辨别属性** 如果选中此选项，那么 C5.0 将在开始构建模型前检查预测变量的有效性。如果发现不相关的预测变量，则会将其从模型构建过程中排除。此选项对于具有许多预测变量字段的模型非常有用，并且有助于防止过度拟合。

在这里我们不做调整，直接以默认参数进行训练。

以同样的方式我们将 C&RT 节点加入工作区。



然后双击“CART”节点进行编辑。

C&RT 树各设置项含义：

1. 目标

- **构建单个树** 创建单个标准决策树模型。通常，与使用其他目标选项构建的模型相比，标准模型更易于说明并可以更快速地进行评分。
  - **方式** 指定用于构建模型的方法。生成模型可在运行流时自动创建模型。启动交互式会话将打开树构建器，可以通过该构建器在创建模型块之前构建树（一次一级）、编辑分割并根据需要进行修剪。
  - **使用树指令** 选中此选项可以指定从节点中生成交互树时应用的指令。例如，可以指定第一级分割和第二级分割，当启动树构建器时会自动应用这些分割。还可以保存交互树构建会话中的指令，以便将来重新创建树时使用。
- **提高模型准确性（Boosting）** 如果要使用一种名为增强的特殊方法来提高模型准确率，请选择此项。增强的工作原理是在序列中构建多个模型。第一个模型按常规方式进行构建。构建第二个模型时，将焦点集中于由第一个模型误分类的记录。构建第三个模型时，将焦点集中于第二个模型的错误，依此类推。最后，通过将整个模型集应用到观测值，并使用加权投票过程将单独的预测组合为一个总预测来分类观测值。增强方法可以显著提高决策树模型的准确性，但也需要更长的训练时间。

- **提高模型稳定性（Bagging）** 如果要使用一种名为组装的特殊方法来提高模型稳定性并避免过度拟合，请选择此项。此选项将创建多个模型并将其进行组合，以获取更加可靠的预测。与标准模型相比，使用此选项获取的模型构建和评分所花费的时间更长。
- **为超大型数据集创建模型** 如果数据集过大，而无法使用任何上述目标选项构建模型，请选择此项。此选项用于将数据划分为更小的数据块，并对每个块构建一个模型。然后，将自动选择最准确的模型并将它们合并到单一模型块中。如果在此屏幕上选择继续训练现有模型选项，可以执行增量式模型更新。

2. 基本

- **最大树深度指定根节点以下的最大级数（对样本进行递归分割的次数）**缺省值为 5；选择定制，并输入值以指定其他级数。
- **对树进行修剪以避免过度拟合** 修剪包括删除对于树的准确性没有显著影响的底层分割。 修剪有助于简化树，使树更容易被理解，在某些情况下还可提高广义性。如果需要未修剪的完整树，请取消选中此选项。
  - **设置风险最大差分**（在标准误差范围内）通过此选项，您可以指定更自由的修剪规则。 标准误差规则使算法可以选择最简单的树，该树的风险估计接近于（但也可能大于）风险最小的子树的风险估计。该值表示已修剪树和风险最小的树之间所允许的风险估计差异大小。例如，如果指定 2 ，那么将选择其风险估计（2× 标准误差）大于完整树的风险估计的树。
- **最大代用项** 替代项是用于处理缺失值的方法。对于树中的每个分割，算法都会对与选定的分割字段最相似的输入字段进行识别。这些被识别的字段就是该分割的代用项 。 当必须对某个记录进行分类，但此记录中的分割字段中具有缺失值时，可以使用代用项字段的值填补此分割。增加此设置将可以更加灵活地处理缺失值，但也会导致内存使用量和训练时间增加。

3. 中止规则

这些选项可控制树的构建方式。停止规则可确定何时停止分割树的特定分支。设置最小分支大小可阻止通过分割创建非常小的子组。如果节点（父级）中要分割的记录数小于指定值，那么父分支中的最小记录数将阻止进行分割。 如果由拆分创建的任何分支（子级）中的记录数小于指定值，那么子分支中的最小记录数将阻止进行分割。

- **使用百分比** 按总训练数据的百分比指定大小。
- **使用绝对值** 按绝对记录数指定大小。

4. 整体

这些设置用于确定在“目标”中请求增强、组装或超大型数据集时的整体行为。将忽略不适用于选定目标的选项。

- **组装和超大型数据集** 在对整体评分时，此规则用于组合来自基本模型的预测值，以计算整体评分值。
  - **分类目标的缺省组合规则** 可以通过投票、最高概率或最高均值概率来对分类目标的整体预测值进行组合。投票选择在基本模型中最常具有最高概率的类别。最高概率选择在所有基本模型中取得单个最高概率的类别。最高均值概率选择在基本模型中对类别概率取平均值时具有最高值的类别。
  - **连续目标的缺省组合规则** 可以使用基本模型中预测值的平均值或中值来组合连续字段的整体预测值。

  要注意的是，如果目标是增强模型准确性，那么将忽略组合规则选择。增强始终使用加权多数表决对分类目标进行评分，而使用加权中值对连续目标进行评分。
- **增强和组装** 当以增强模型精确性或稳定性为目标时，指定要构建的基本模型数；对于组装方法，此为自助样本数。它应该为正整数。

5. 成本和先验

- **误分类成本**

在某些环境中，特定错误类别的成本高于其他错误的成本。例如，将高风险信贷申请人分类为低风险申请人（一种错误类别）的成本高于将低风险申请人分类为高风险申请人（另一种错误类别）的成本。使用误分类成本可指定不同类别的预测误差的相对重要性。

误分类成本在本质上指应用于特定结果的权重。这些权重可化为模型中的因子，并可能在实际上更改预测（作为避免高成本错误的一种方式）。

除 C5.0 模型之外，在对模型进行评分时，误分类成本是不适用的；在使用自动分类器节点、评估图表或分析节点对模型进行排秩或比较时，误分类成本也不予以考虑。将成本计算在内的模型不比不将成本计算在内的模型产生的误差小，这样的模型不会也不可能按照总体精确性排序到任何更高的级别，但是在实际应用中，这样的模型执行的结果可能更好，因为它有一个内置的偏差，从而有利于将错误的成本降低。

成本矩阵显示了预测类别和实际类别的每个可能的组合的成本。缺省情况下，所有误分类成本都设置为 1.0。要输入定制成本值，可选择 使用误分类成本 并将定制值输入到成本矩阵中。

要更改误分类成本，可选择与所需的预测值和实际值的组合对应的单元格，清除此单元格内现有的内容，然后为其输入所需的成本。成本不会自动均摊。例如，如果将 A 误分类为 B 的成本设置为 2.0，那么将 B 误分类为 A 的成本

将仍是缺省值 1.0，除非也明确地对它进行更改。

- **先验**

通过这些选项可以在预测分类目标字段时为分类指定先验概率。先验概率是对总体（从中抽取训练数据）中每个目标分类的总相对频率的估计。换句话说，先验概率是对预测变量值有任何了解 之前对每个可能的目标值的概率估计。有三种方法用来设置先验概率：

- **基于训练数据** 这是缺省选项。先验概率基于训练数据中分类的相对频率。
- **对于所有类别都相等** 所有类别的先验概率都定义为 1/k，其中 k 是目标分类数。
- **定制** 可以自行指定先验概率。对于所有类，都将先验概率的初值设置为相等。可以将单个分类的概率调整为用户定义的值。要调整特定分类的概率，可在表中对应于所需分类的概率单元格中，先清除其内容，然后输入所需的值。

所有分类的先验概率之和应为 1.0（ 概率约束 ）。如果权重之和不为 1.0，将出现一个警告，显示带有自动标准化这些值的选项。此自动调整操作可在强制执行概率约束时保留分类中的比例。通过单击标准化按钮，可在任何时间执行此调整。将表中所有分类重置为相同的值，可单击均衡按钮。

- **使用误分类成本调整先验概率** 通过此选项可以根据误分类成本（在“成本”选项卡中指定）调整先验概率。从而可为使用两分杂质测量的树将损失信息直接合并到树生成过程中。（未选中此选项时，损失信息仅用于为基于两分测量的树分类记录和计算风险估计。）

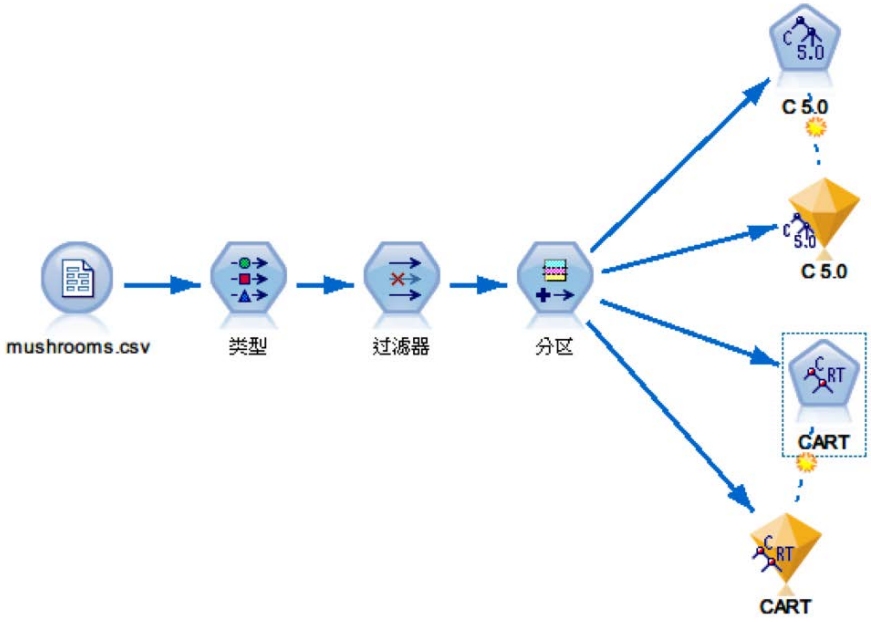
6. 高级

- **最小杂质改变** 指定最小杂质改变以便在树中创建新的分割。杂质是指由树定义的子组在每个组中所具有的输出字段值的广度。对于分类目标，如果节点中 100% 的观测值都落在目标字段的特定类别中，那么该节点被认为是“纯节点”。树构建的目的是创建具有相似输出值的子组 - 换句话说，是为了减少每个节点中的杂质。如果某个分支的最佳分割按小于指定值的数量减少杂质，那么不会进行此分割。
- **分类目标的杂质测量** 对于分类目标字段，指定用于测量树的杂质的方法。（对于连续目标，将忽略此选项，而通常会使用 最小平方差 杂质测量。）
  - **吉尼**是基于分支的类别成员资格概率的一般杂质测量。
  - **两分**是强调二元分割并更有可能导致从分割中生成大小近似相同的分支的杂质测量。
  - **有序**添加了额外的限制，即只有相邻的目标类才可以组成一组，此选项仅适用于有序目标。如果对于名义目标选中此选项，将缺省使用标准的两分测量。

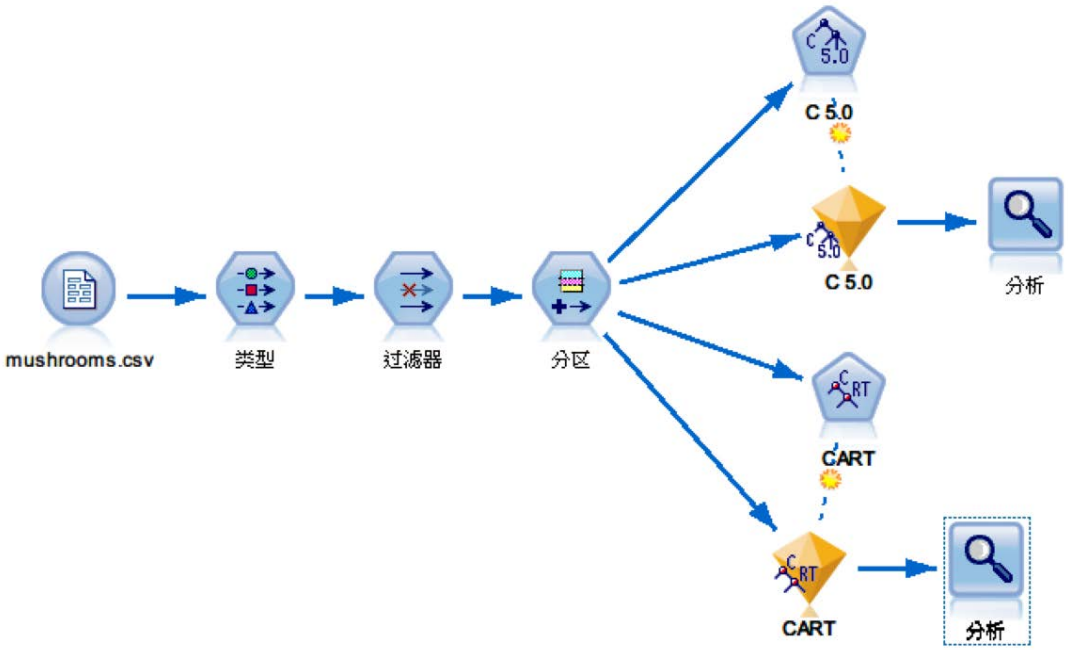
- **防止过度拟合集合** 该算法在内部将记录划分为模型构建集合和防止过度拟合集合，后者作为独立的数据记录集，用于跟踪训练过程中的错误，以防止该方法对数据中的几率变异进行建模。指定记录的百分比。缺省值为 30。
- **复制结果** 通过设置随机种子，您可以复制分析。指定一个整数，或单击生成，这将产生一个介于 1 与 2147483647 之间（包括 1 和 2147483647）的伪随机整数。

在这里我们仍选择默认参数训练。

点击上方“工具栏”中的运行按钮即可运行所有节点。运行完成后会得到两个算法的结果节点。

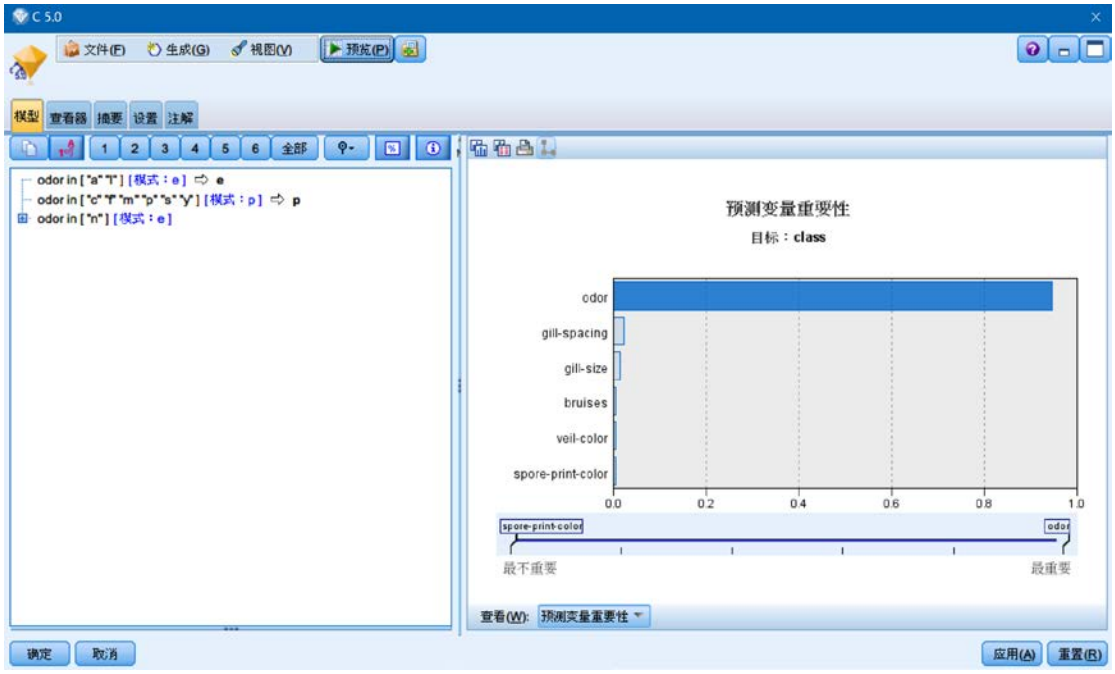


为了进一步分析，我们将两个节点选项板中的“输出”→“分析”节点拖至工作区中并分别和两个结果节点进行连接，得到：



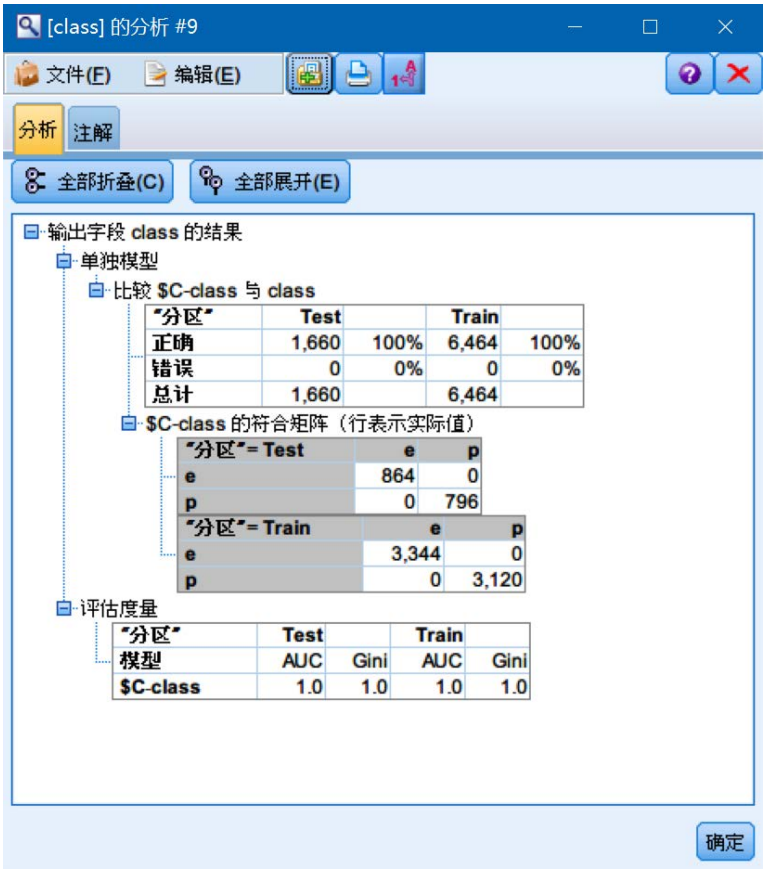
双击“分析”节点，设置要分析的内容，“重合矩阵”和“评估度量”。

然后双击结果节点查看结果。



可以在各选项卡中看到预测变量的重要性、决策树模型、摘要（包括树的深度，输入、参数等等）。

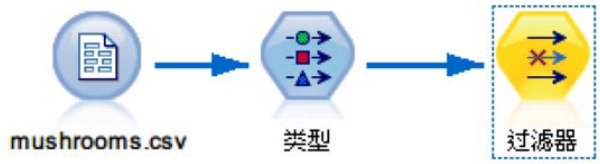
右键点击“分析”节点并点击“运行”即可进行进一步分析。



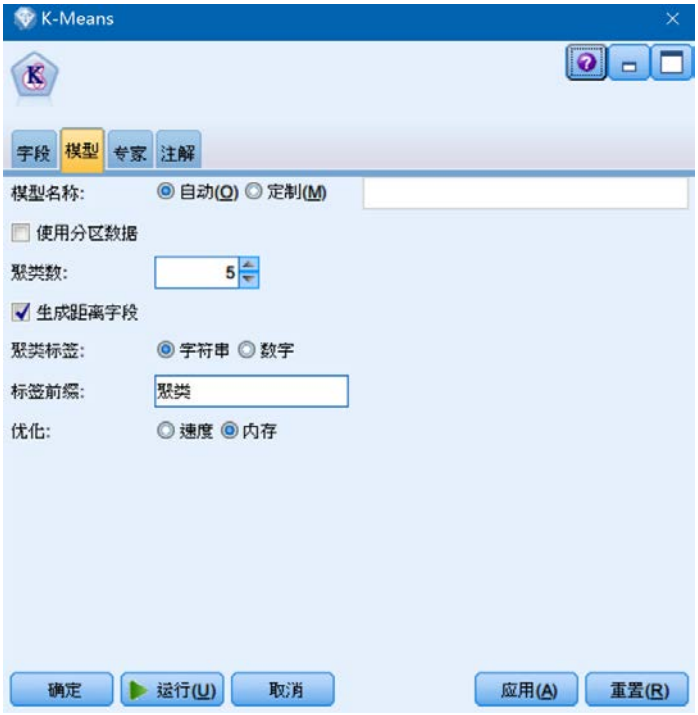
可以看到，正确率达到了 100%。

第七步 聚类模型

接下来我们进行聚类。我们新建一个流，然后将数据节点、类型节点和过滤器节点拷贝进去。



然后在节点选项板中找到“建模”→“细分”→“K-Means”，拖入工作区并与“过滤器”连接。双击“K-Means”节点进行参数配置。



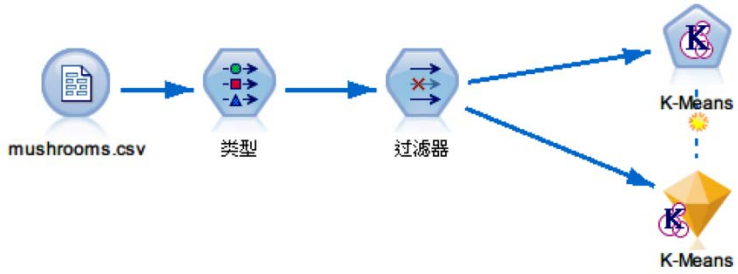
各参数描述如下：

- **模型名称** 用户可根据目标或标识字段自动生成模型名称（未指定此类字段时自动生成模型类型）或指定一个定制名称。
- **使用分区数据** 如果定义了分区字段，那么此选项可确保仅训练分区的数据用于构建模型。
- **指定的聚类数** 指定要生成的聚类数。缺省值是 5。
- **生成距离字段** 如果选中此选项，那么模型块将包括一个字段，该字段包含每条记录与所分配到的聚类的中心之间的距离。
- **聚类标签** 为生成的聚类成员资格字段的值指定格式。聚类成员资格可表示为具有指定标签前缀的字符串（例如，“聚类 1”、“聚类 2”等等），也可以表示为数值。
- **优化** 根据具体需求，选择旨在提高模型构建性能的选项。
  - **选择速度** 可指示算法从不使用磁盘溢出，以便提高性能。

- **选择内存** 可指示算法在合适的时候，以牺牲某些速度为代价使用磁盘溢出。缺省情况下，此选项处于选中状态。

- **中止条件** 指定训练模型时要使用的中止条件。缺省 停止标准为 20 次迭代或差异 <0.000001，以先满足的标准为准。选中定制可指定自己的停止标准。
  - **最大迭代次数** 使用此选项可在指定的迭代次数后中止模型训练。
  - **差异容差** 通过此选项，您可以在某次迭代的聚类中心中的最大差异小于指定的级别时中止模型训练。
- **集合的编码值** 指定 0 到 1.0 之间的值，以用于将集合字段重新编码为数字字段组。缺省值是 0.5 的平方根（大约为 0.707107），它可为重新编码的标志字段提供适当的加权。值越接近 1.0，对集合字段的加权就越高于对数值字段的加权。

这里我们仍旧以默认参数运行，得到结果。



双击结果节点可以看到聚类结果，包括各聚类的分布，以及聚类质量。

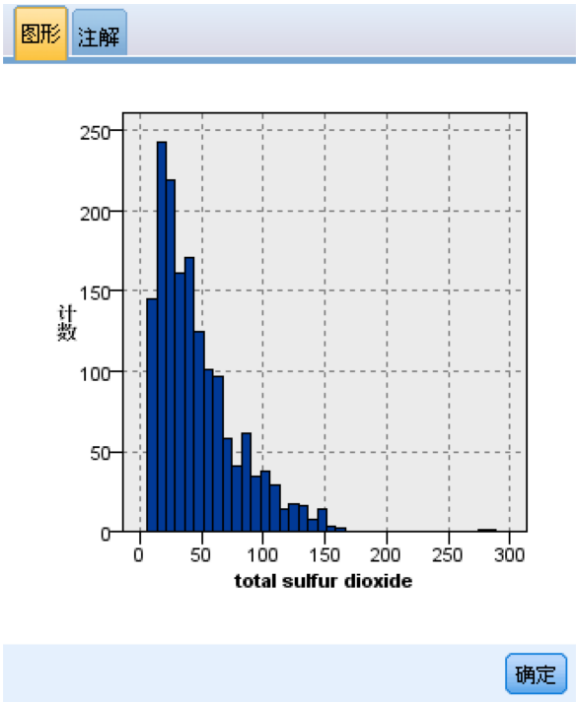
二、 红酒数据集

红酒数据集来自 Red Wine Quality | Kaggle，共有 1599 条数据，通过固定酸度、挥发酸度等 11 个字段预测酒的质量，酒的质量从 1-10 打分，数据中都分布在 3-8。所有的特征都是连续特征，与上个数据集不同。因此我们主要会考虑数据预处理上的不同。

首先将数据导入：

然后用数据审阅对数据进行分析。

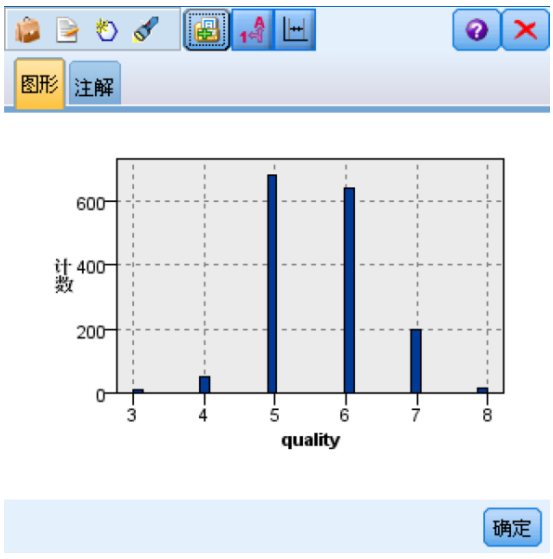
可以看到，对于连续值来说，数据审阅可以计算出各字段的平均值、方差等统计值。我们按照方差来排序，可以看到，方差最大达到了 1082.102，这是由于存在离群点导致的。为了解决这一问题，我们可以通过“选择器”来筛掉离群点。进一步双击“total sulfur dioxide”的条形图我们可以看到：



在 277.53 和 285.18 这两个点的地方存在有两个明显的离群点，为了提升预测的准确性，因此我们需要在之后将这两个点删除。

同样，我们以此类推，可以分析得到“chlorides”、“residual sugar”和“sulphates”字段也需要相同的操作。

我们再来看目标值——质量。



可以看到，几个分类有明显的不平均。在这里我们将问题简化，将原先的预测质量多分类问题替换成预测品质的高低这一二分类问题，并按照质量>5 和质量<=5 分为高质量（High）和低质量（Low），在之后我们需要进行重新分类的操作。

接下来先做类型标注，在工作区添加“字段选项”→“类型”节点，将“quality”字段标记为 “目标”。

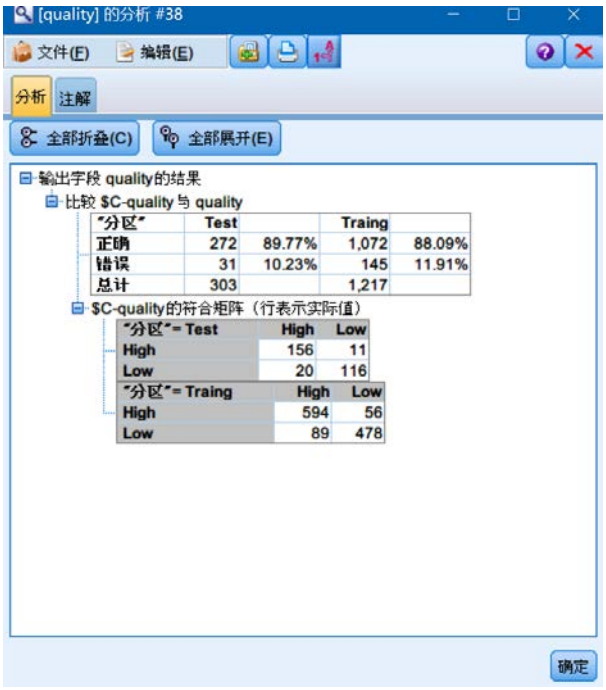
然后在工作区添加“记录选项”→“选择”节点，利用表达式对数据进行筛选，筛除一部分数据。

其中条件表达式为：chlorides > 0.17 or 'residual sugar' > 5 or 'total sulfur dioxide' > 163 or 'free sulfur dioxide' > 42，也就是说符合“chlorides”大于 0.17、“total sulfur dioxide”大于 163 或者“free sulfur dioxide”大于 42 等等条件的所有数据都会被丢弃。这样我们就丢弃了部分离群点。

接下来在工作区添加“字段选项”→“重新分类”，对“quality”字段进行重新分类，对原始值为 3、4 和 5 的，新值替换成“Low”；对于原始值为 6、7 和 8 的，新值替换成“High”。

这样我们就将原问题替换成一个二分类问题。然后和上一个实验一致，我们添加“字段选项”→“分区”节点，将数据集分成训练集和测试集，其比例为 80：20。

完成分区后我们将“建模”→“分类”中的 C5.0 和 C&RT 树节点分别添加至工作区，点击上方功能区的“运行”按钮得到模型结果。然后在“结果”节点上分别添加“输出”→“分析”节点，进一步对结果进行分析。分析结果如下所示：

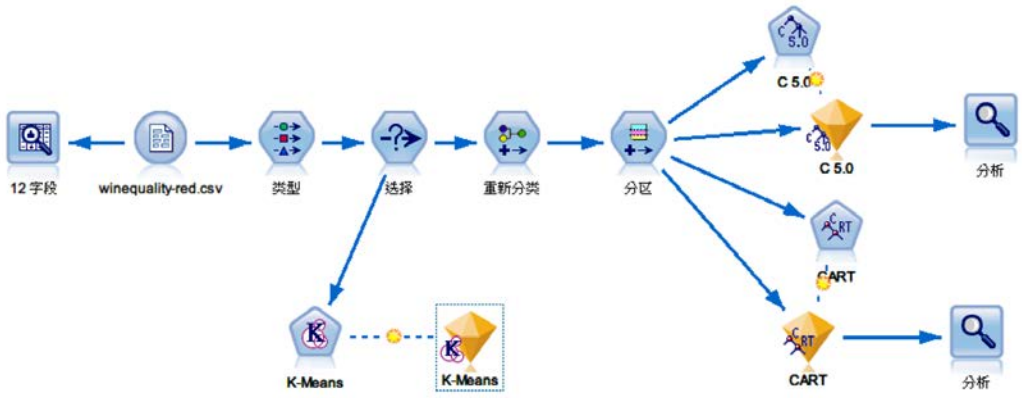


可以看到，C5.0 的预测正确率为 89.77%，可以说对于二分类问题来说正确率尚可。事实上可以通过调整参数来对模型进行进一步调优。

然后我们再对数据进行聚类分析，将“建模”→“细分”→“K-Means”添加至工作区，然后运行，得到聚类结果：

聚类质量平均 Silhouette=0.3，说明聚类效果不是特别好，仍有待改进的地方。

最终整个数据流图如下所示：



接下来你可以尝试：

- 优化 C5.0 和 C&RT 树的表现
- 优化 K-Means 算法的表现
- 对数据进行进一步处理
- 如果将选择节点去掉，模型表现会怎么样呢？修改选择节点的条件表达式呢？
- 尝试更为复杂的数据集和问题