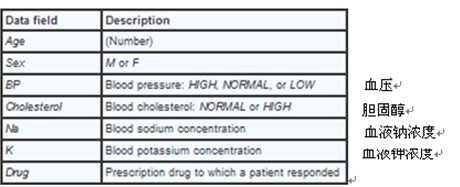
Clementine 应用实例分析--Drug Treatment（Exploratory Graphs/C5.0）

假如你是一个药物研究者，你收集了一些遭受相同疾病病人的数据集，在他们的治疗过程中，每个病人选取5种药物一种。我们的任务就是使用这些数据发现哪种药物针对同一种病情最适合。

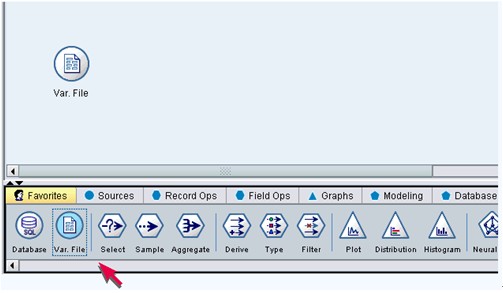
例子使用的数据流为druglearn.str，数据集是DRUG1n。

数据中的字段如下图：

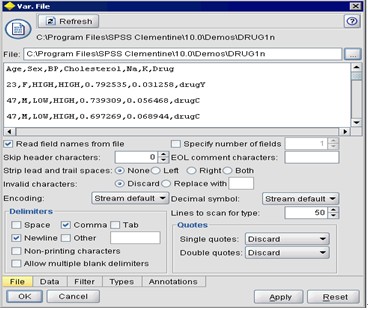
[](http://photo.blog.sina.com.cn/showpic.html#blogid=71884873010116z9&url=http://s8.sinaimg.cn/orignal/71884873gba03b9b15757)

一．数据准备

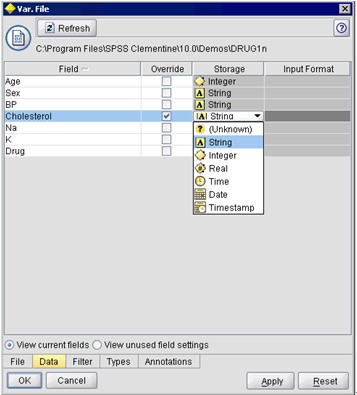
Reading in Text Data（读取文本数据）



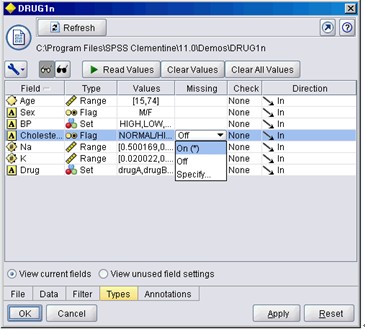
在clementine中，你可以使用Variable File 节点读取有分隔符的文本数据。



点击“Data”标签，重载和改变字段的储存（Storage）---数据字段的使用方式（注意：储存不同于类型type）



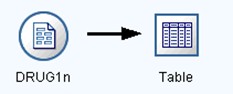
类型标签可以使你获取更多的数据字段的类型。我们可以选择“read values”来查看每个字段的实际值（该过程称为实例化）。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=71884873010116z9&url=http://s7.sinaimg.cn/orignal/71884873gba03b9a2fa96)

二、数据探索

1. Adding a Table

现在我们已经将数据文件加载进clementine，如果你想看一下每一天记录的值，可以通过构建一个包含“Table”节点的流。

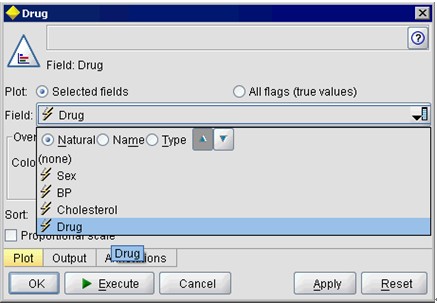


此时点击“执行”即可看到所有的记录。

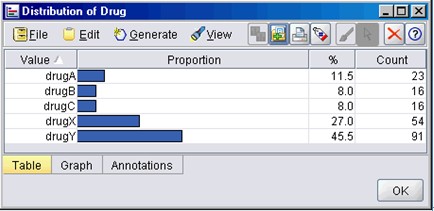
2. Creating a Distribution Graph

在数据挖掘中，经常通过创建可视化的图形来探索数据。Clementine提供了几种不同的图形供我们选择（选择依赖于所总结的数据种类）。例如研究病人选择每种药物的比例，使用一个“Distribution”节点。

增加一个“Distribution”节点至数据流，同数据源节点连接。选择“Drug”作为目标字段（你想显示的字段分布），然后“执行”。



结果图形显示如下，其中大部分病人选择了药物Y，B和C选择的最少。

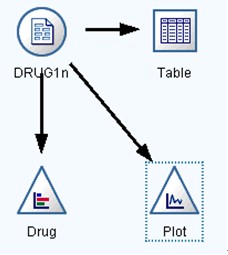
[](http://photo.blog.sina.com.cn/showpic.html#blogid=71884873010116z9&url=http://s13.sinaimg.cn/orignal/71884873gba03d1d0f89c)

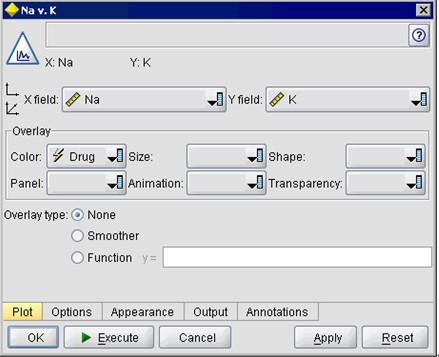
另外，我们也可以增加和执行“Audit”节点，以便快速的看到所有字段的分布和直方图。

3. Creating a Scatterplot（散点图）

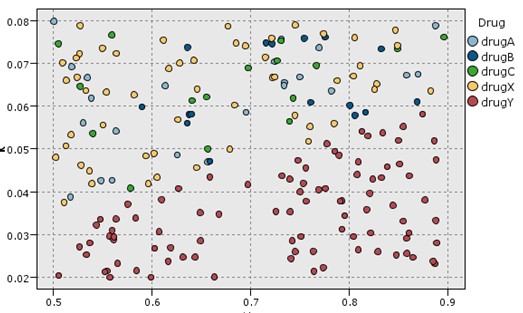
那现在我们分析看看哪些因素影响到最终的结果Drug---目标变量。作为一个研究者，我们知道钠（sodium）和钾（potassium）浓度是重要的因素，因为两者都是数值型，你可以创建钠和钾的散点图，而Drug的分类以颜色分类。

将“Plot”节点添加到数据流中。选择Na作为X字段，K作为Y字段，Drug作为覆盖层字段，点击“执行”。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=71884873010116z9&url=http://s8.sinaimg.cn/orignal/71884873gba03d1cfeb57)

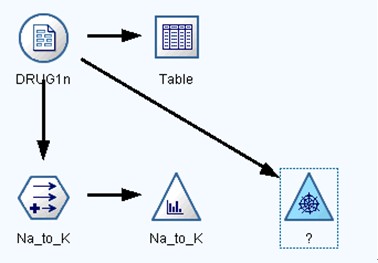
[](http://photo.blog.sina.com.cn/showpic.html#blogid=71884873010116z9&url=http://s5.sinaimg.cn/orignal/71884873gba03d1b17dc4)

显示的散点图如下图，散点图清晰的显示出某个阈值以上都是选择drug y（即图的右下方），而阈值以下都没有选择drug Y（即图的左上方）。这个阈值就是钠（Na）和钾（K）的比率。

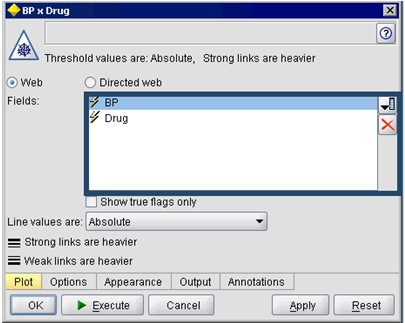


4. Creating a Web Graph

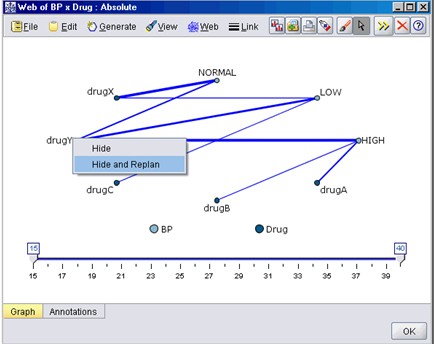
由于许多数据字段都是分类型的，所以我们可以绘制一个网络图，以体现不同种类之间的关系。首先选择一个“Web”节点至数据流。



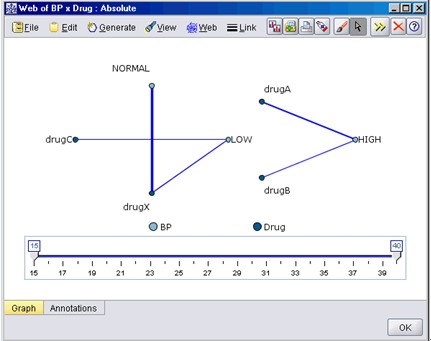
在Web 节点中，选择BP（blood pressure）和Drug，执行。



显示结果如下图，从图中我们看出drug Y同所有水平的血压均有联系，我们认为drug Y是最好的选择，对于其他的药物，我们可以选择隐藏drug Y来观察。



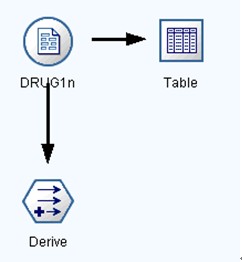
在简化后的图形中，Drug Y和所有它的链接均隐藏。现在我们可以看到仅仅药物A和B同高血压有关，药物X和C与低血压有关，正常血压仅与药物X有关。尽管我们仍然不知道对于一个病人，在药物A和B或者C和X如何选择，但是该模型也有一定的作用。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=71884873010116z9&url=http://s15.sinaimg.cn/orignal/71884873gba03e11aed6e)

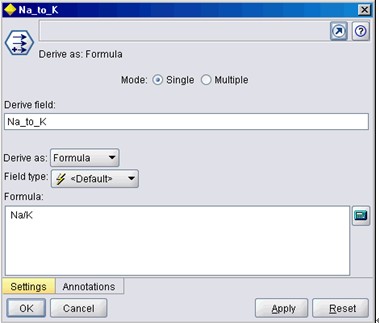
5. Deriving a new field

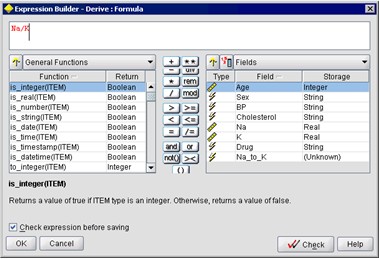
当使用药物Y时的钠与钾的比率是可以计算的，我们就导出一个字段包含每条记录的该比率值。当我们构建模型预测使用各种药物时，可能会用到该字段。

引入“Derive”节点至数据流。

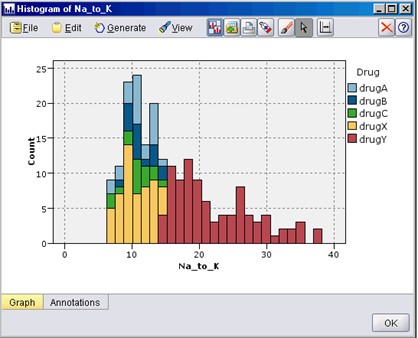


命名新字段为Na\_to\_K，编辑，





我们可以添加直方图节点（Histogram）检查新字段，选择Na\_to\_K作为图形字段，Drug作为覆盖字段。当执行数据流时，得到如下图形，我们可以看到当Na\_to\_K值接近15或以上，选择drug Y。

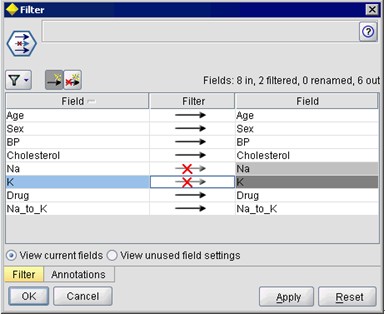


三、构建模型

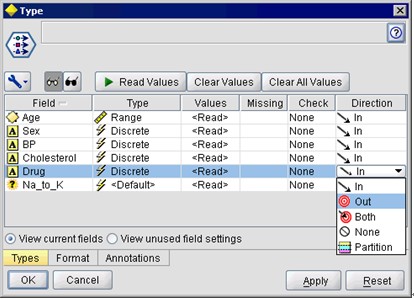
1. Building a Model

通过探索这些数据之后，我们已经形成了一些假定。血液中钠对钾的比率和血压对药物的选择起一定作用，但是我们不能详细的阐释他们之间的关系，而现在要构建的模型就是来解决这个问题。在这个案例中，我们使用规则构建模型C5.0尽可能的使用数据。

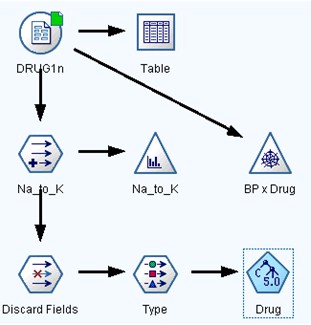
因为我们已经使用了新字段Na\_to\_K，所以我们可以过滤点原始的字段——Na和K。使用“Filter”节点，过滤掉Na和K。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=71884873010116z9&url=http://s3.sinaimg.cn/orignal/71884873gba03ebca1e82)

接下来，增加一个“Type”节点连接在“filter”节点，在该节点设置“Drug”字段为“Out”（暗示Drug为预测字段）。

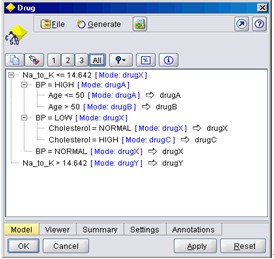


将C5.0节点添加至数据流，右击后点“执行”。



2. Browsing the Model

当C5.0节点执行之后，产生的模型节点增加到窗口的右上角，右击后点“浏览”。



现在我们可以看到

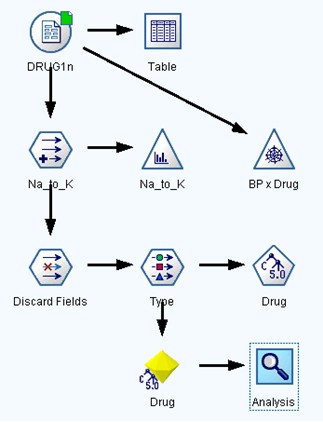
l  Na\_to\_K小于14.642和高血压，年龄决定药物的选择。

对于低血压，cholesterol决定药物的选择。

我们也可以通过Views标签，看具体的分类。

3. Using an Analysis Node

我们可以通过“Analysis”节点获取模型的精确度。首先，将C5.0模型和“Analysis”节点添加至数据流，执行“Analysis”该节点。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=71884873010116z9&url=http://s16.sinaimg.cn/orignal/71884873gba03f4574c7f)

执行之后，结果如下图，该结果显示该模型几乎正确的预测了数据集中的每个记录。在现实数据集中一般不可能看到100%的精确度，但是我们可以使用“Analysis”节点帮助我们决定该模型对我们的应用是否是可接受的精确度。

总结：该应用实例比较全面，包括数据准备、数据探索、构建模型具体的进行数据挖掘。