Clementine 数据挖掘快速上手

Version1.0

Prepared by 高处不胜寒

QQ 群：14094415

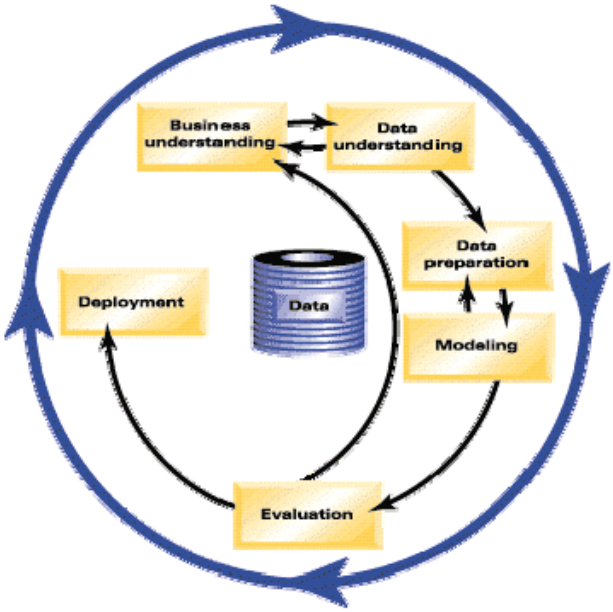
2009-10-15

一、Clementine数据挖掘的基本思想

数据挖掘（Data Mining）是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程，它是一种深层次的数据分析方法。随着科技的发展，数据挖掘不再只依赖在线分析等传统的分析方法。

它结合了人工智能（AI）和统计分析的长处，利用人工智能技术和统计的应用程序，并把这些高深复杂的技术封装起来，使人们不用自己掌握这些技术也能完成同样的功能，并且更专注于自己所要解决的问题。

Clementine为我们提供了大量的人工智能、统计分析的模型（神经网络，关联分析，聚类分析、因子分析等），并用基于图形化的界面为我们认识、了解、熟悉这个软件提供了方便。除了 这些Clementine还拥有优良的数据挖掘设计思想，正是因为有了这个工作思想，我们每一步的工 作也变得很清晰。（如图一所示）



CRISP-DM process model

如图可知，CRISP-DM Model包含了六个步骤，并用箭头指示了步骤间的执行顺序。这些顺序并不严格，用户可以根据实际的需要反向执行某个步骤，也可以跳过某些步骤不予执行。通过对这些步骤的执行，我们也涵盖了数据挖掘的关键部分。

商业理解(Business understanding)：商业理解阶段应算是数据挖掘中最重要的一个部分，在这个阶段里我们需要明确商业目标、评估商业环境、确定挖掘目标以及产生一个项目计划。

数据理解(Data understanding)：数据是我们挖掘过程的“原材料”，在数据理解过程中我们要知道都有些什么数据，这些数据的特征是什么，可以通过对数据的描述性分析得到数据的特点。

数据准备(Date preparation)：在数据准备阶段我们需要对数据作出选择、清洗、重建、合并等工作。 选出要进行分析的数据，并对不符合模型输入要求的数据进行规范化操作。

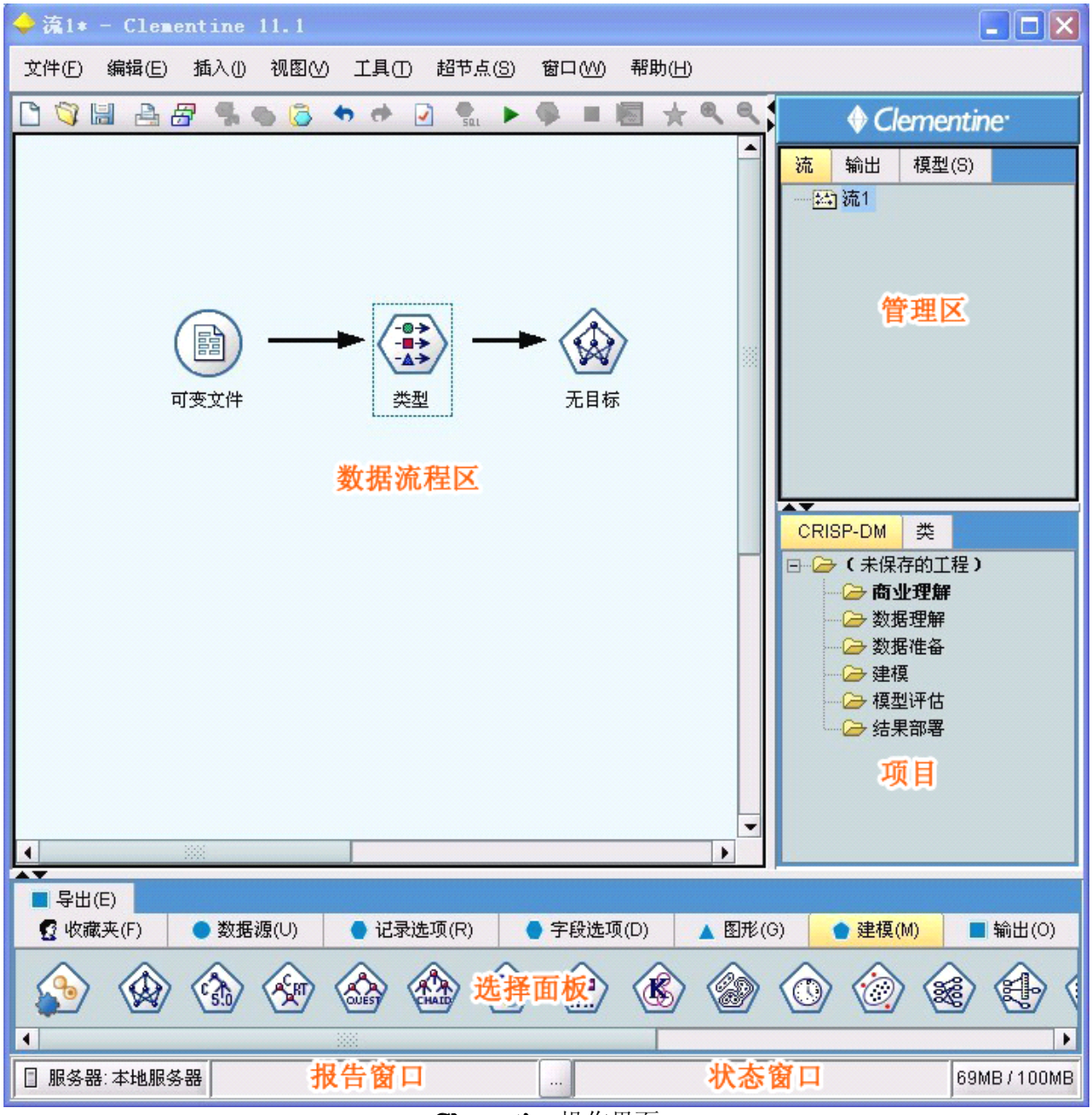
建模(Modeling)：建模过程也是数据挖掘中一个比较重要的过程。我们需要根据分析目的选出适合的模型工具，通过样本建立模型并对模型进行评估。

模型评估(Evaluation)：并不是每一次建模都能与我们的目的吻合，评价阶段旨在对建模结果进行评估，对效果较差的结果我们需要分析原因，有时还需要返回前面的步骤对挖掘过程重新定义。

结果部署(Deployment)：这个阶段是用建立的模型去解决实际中遇到的问题，它还包括了监督、维持、产生最终报表、重新评估模型等过程。

二、Clementine的基本操作方法

1.操作界面的介绍



Clementine 操作界面

1.1数据流程区

Clementine在进行数据挖掘时是基于数据流程形式，从读入数据到最后的结果显示都是由流程图的形式显示在数据流程区内。数据的流向通过箭头表示，每一个结点都定义了对数据的不同操作，将各种操作组合在一起便形成了一条通向目标的路径。数据流程区是整个操作界面中最大的部分，整个建模过程以及对模型的操作都将在这个区域内执行。我们可以通过"文件"(File）—"新建流"(new stream)新建一个空白的数据流，也可以打开已有的数据流。所有在一个运行期内打开的数据流都将保存在管理器的Stream栏下。

1.2选项面板

选项面板横跨于Clementine操作界面的下部，它被分为收藏夹（Favorites）、数据源（Sources）、记录选项（Record Ops）、字段选项（Fields Ops）、图形（Graphs）、建模（Modeling）、输出（Output）、导出八个栏，其中每个栏目包含了具有相关功能的结点。 结点是数据流的基本组成部分，每一个结点拥有不同的数据处理功能。设置不同的栏是为了将不同功能的结点分组，下面我们介绍各个栏的作用。

数据源(Sources)：该栏包含了能读入数据到Clementine的结点。例如Var. File结点读取自由格式的文本文件到Clementine，SPSS File读取spss文件到Clementine。

记录选项(Record Ops)：该栏包含的结点能对数据记录进行操作。例如筛选出满足条件的记录（select)、将来自不同数据源的数据合并在一起（merge)、向数据文件中添加记录(append)等。

字段选项(Field Ops)：该栏包含了能对字段进行操作的结点。例如过滤字段（filter）能让被过滤的字段不作为模型的输入、导出（derive）结点能根据用户定义生成新的字段，同时我们还可以定义字段的数据格式。

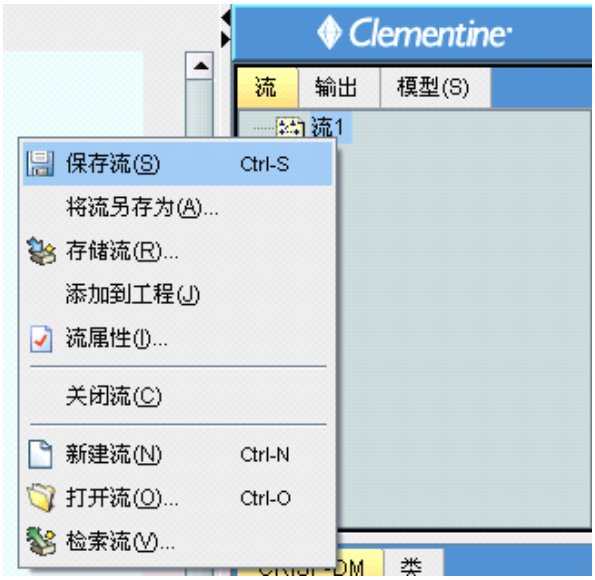
图形(Graphs)：该栏包含了众多的图形结点，这些结点用于在建模前或建模后将数据由图形形式输出。

建模(Modeling)：该栏包含了各种已封装好的模型，例如神经网络（Neural Net)、决策树（C5.0）等。这些模型能完成预测（Neural Net，Regression，Logistic ）、分类（C5.0，C&R Tree，Kohonen，K-means，Twostep)、关联分析(Apriori，GRI，Sequece)等功能。

输出(Output)：该栏提供了许多能输出数据、模型结果的结点，用户不仅可以直接在Clementine中查看输出结果，也可以输出到其他应用程序中查看，例如SPSS和Excel。

收藏夹(Favorites)：该栏放置了用户经常使用的结点，方便用户操作。用户可以自定义其Favorites 栏，操作方法为：选中菜单栏的工具(Tools)，在下拉菜单中选择收藏夹(Favorites)，在弹出的Palette Manager 中选中要放入Favorites栏中的结点。

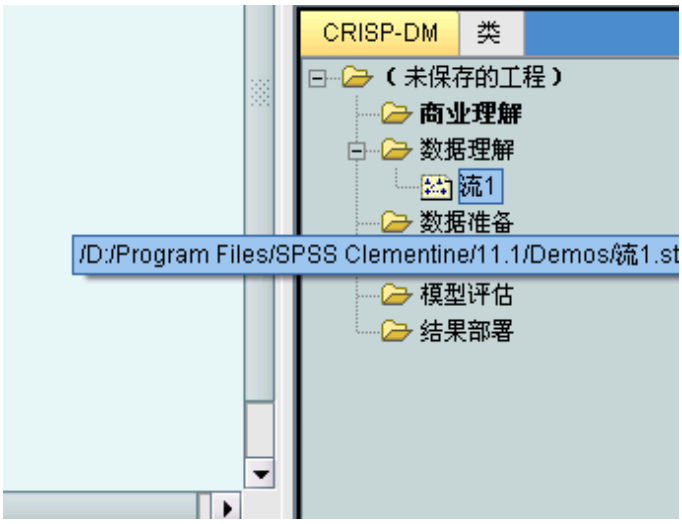
1.3管理器



管理器中共包含了流(Streams)、输出(Outputs)、模型(Models)三个栏。其中流（Streams）中放置了运行期内打开的所有数据流，可以通过右键单击数据流名对数据流进行保存、设置属性等操作。输出(Outputs)中包含了运行数据流时所有的输出结果，可以通过双击结果名查看输出的结果。模型(Models)中包含了模型的运行结果，我们可以右键单击该模型从弹出的浏览(Browse)中查看模型结果，也可以将模型结果加入数据流中。

1.4项目窗口的介绍

项目窗口含有两个选项栏，一个是CRISP-DM，一个是类(Classes)。



CRISP-DM的设置是基于CRISP-DM Model的思想，它方便用户存放在挖掘各个阶段形成的 文件。由右键单击阶段名，可以选择生成该阶段要拥有的文件，也可以打开已存在的文件将其放入该阶段。这样做的好处是使用户对数据挖掘过程一目了然，也有利于对它进行修改。

类(Classes)窗口具有同CRISP-DM窗口相似的作用，它的分类不是基于挖掘的各个过程，而是基于存储的文件类型。例如数据流文件、结点文件、图表文件等。

2、数据流基本操作的介绍

2.1生成数据流的基本过程

数据流是由一系列的结点组成，当数据通过每个结点时，结点对它进行定义好的操作。我们在建立数据流是通常遵循以下四步：

①、向数据流程区增添新的结点；

②、将这些结点连接到数据流中；

③、设定数据结点或数据流的功能；

④、运行数据流。

2.2向数据流程区添/删结点 当向数据流程区添加新的结点时，我们有下面三种方法遵循：

①、双击结点面板中待添加的结点；

②、左键按住待添加结点，将其拖到数据流程区内；

③、选中结点面板中待添加的结点，将鼠标放入数据流程区，在鼠标变为十字形时单击数据流程区。

通过上面三种方法我们都将发现选中的结点出现在了数据流程区内。当我们不再需要数据流程区内的某个结点时，可以通过以下两种方法来删除：

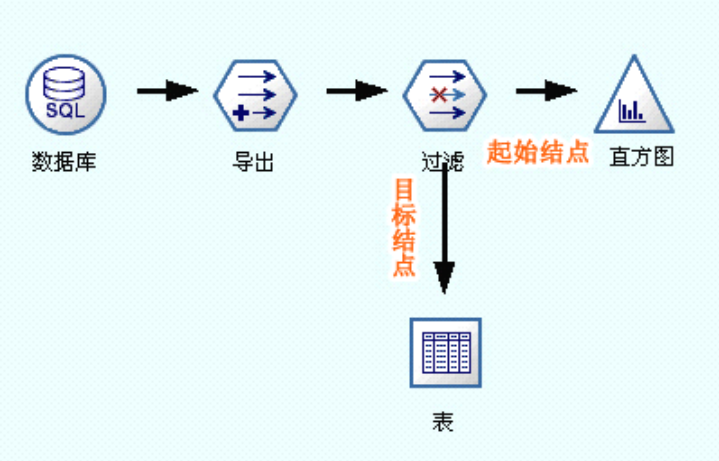
①左键单击待删除的结点，用删除（delete）；

②右键单击待删除的结点，在出现的菜单中选择删除（delete）。

2.3将结点连接到数据流中上面我们介绍了将结点添加到数据流程区的方法，然而要使结点真正发挥作用，我们需要

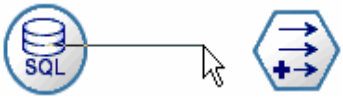
把结点连接到数据流中。以下有三种可将结点连接到数据流中的方法：

①、双击结点　左键选中数据流中要连接新结点的结点（起始结点），双击结点面板中要连接入数据



流的结点（目标结点），这样便将数据流中的结点与新结点相连接了；　图六双击目标结点以加入数据流

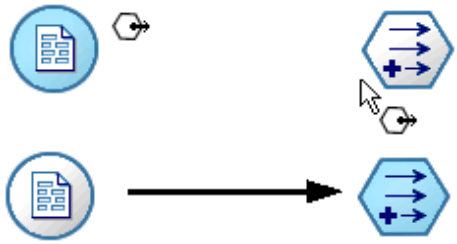
②、通过鼠标滑轮连接



在工作区内选择两个待连接的结点，用左键选中连接的起始结点，按住鼠标滑轮将其拖曳到目标结点放开，连接便自动生成。（如果鼠标没有滑轮也选用alt键代替）由滑轮连接两结点

③、手动连接右键单击待连接的起始结点，从弹出的菜单栏中选择连接(Connect)。选中连接(Connect)后鼠标和起始

结点都出现了连接的标记，用鼠标单击数据流程区内要连接的目标结点，连接便生成。



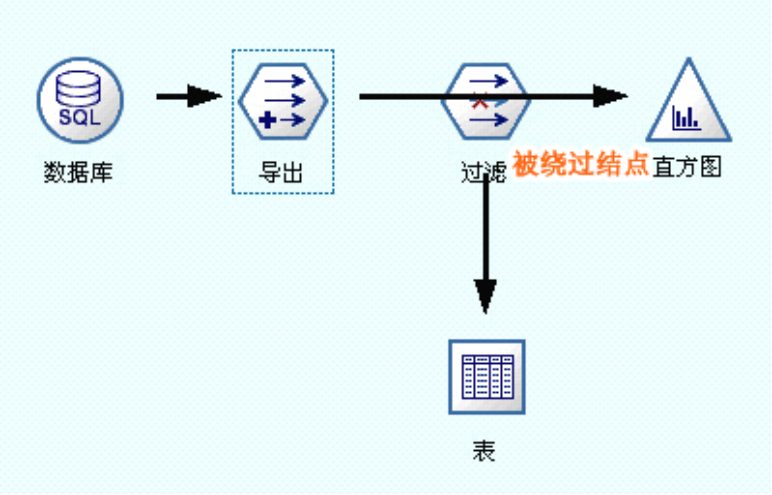
图八　选择菜单栏中的**连接connect**

图九点击要连入的结点　注意：①、第一种连接方法是将选项面板中的结点与数据流相连接，后两种方法是将已在

数据流程区中的结点加入到数据流中　②、数据读取结点（如SPSS　File）不能有前向结点，即在　连接时它只能作为起始结点而不能作为目标结点。

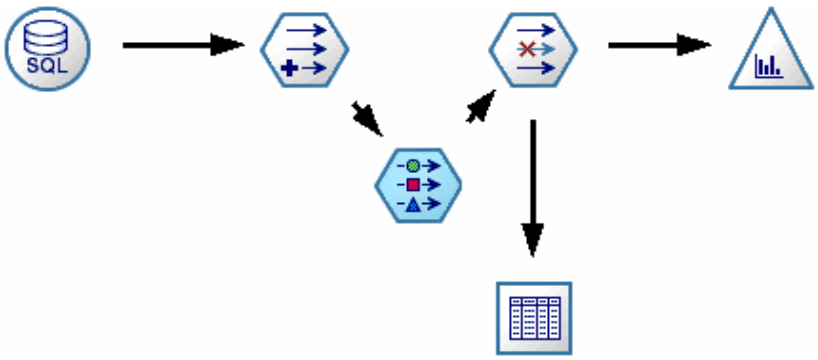
2.4绕过数据流中的结点　当我们暂时不需要数据流中的某个结点时我们可以绕过该结点。在绕过它时，如果该结点

既有输入结点又有输出结点那么它的输入节点和输出结点便直接相连；如果该结点没有输出结　点，那么绕过该结点时与这个结点相连的所有连接便被取消。

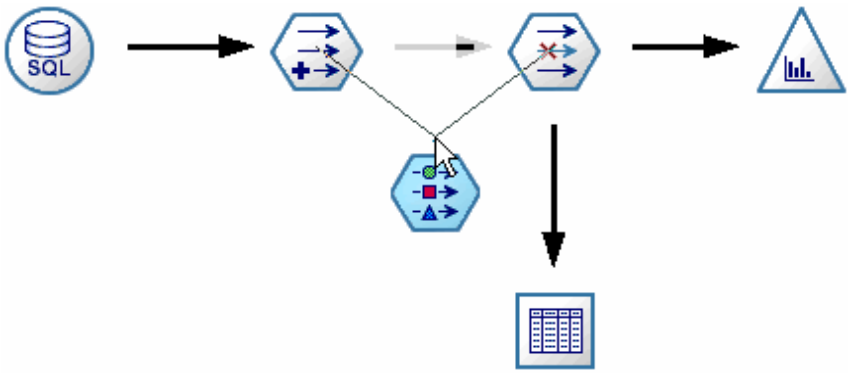


方法：用鼠标滑轮双击需要绕过的结点或者选择按住alt键，通过用鼠标左键双击该结点来 完成。

2.5将结点加入已存在的连中当我们需要在两个已连接的结点中再加入一个结点时，我们可以采用这种方法将原来的连接变成两个新的连接。



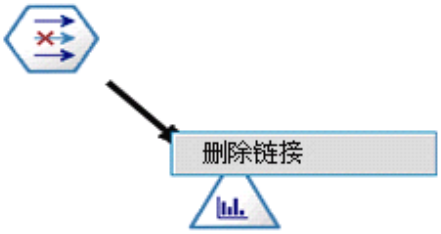
方法：用鼠标滑轮单击欲插入新结点的两结点间的连线，按住它并把他拖到新结点时放手， 新的连接便生成。（在鼠标没有滑轮时亦可用alt键代替）



2.6删除连接当某个连接不再需要时，我们可以通过以下三种方法将它删除：

①、选择待删除的连接，单击右键，从弹出菜单中选择Delete Connection；

②、选择待删除连接的结点，按F3键，删除了所有连接到该结点上的连接；



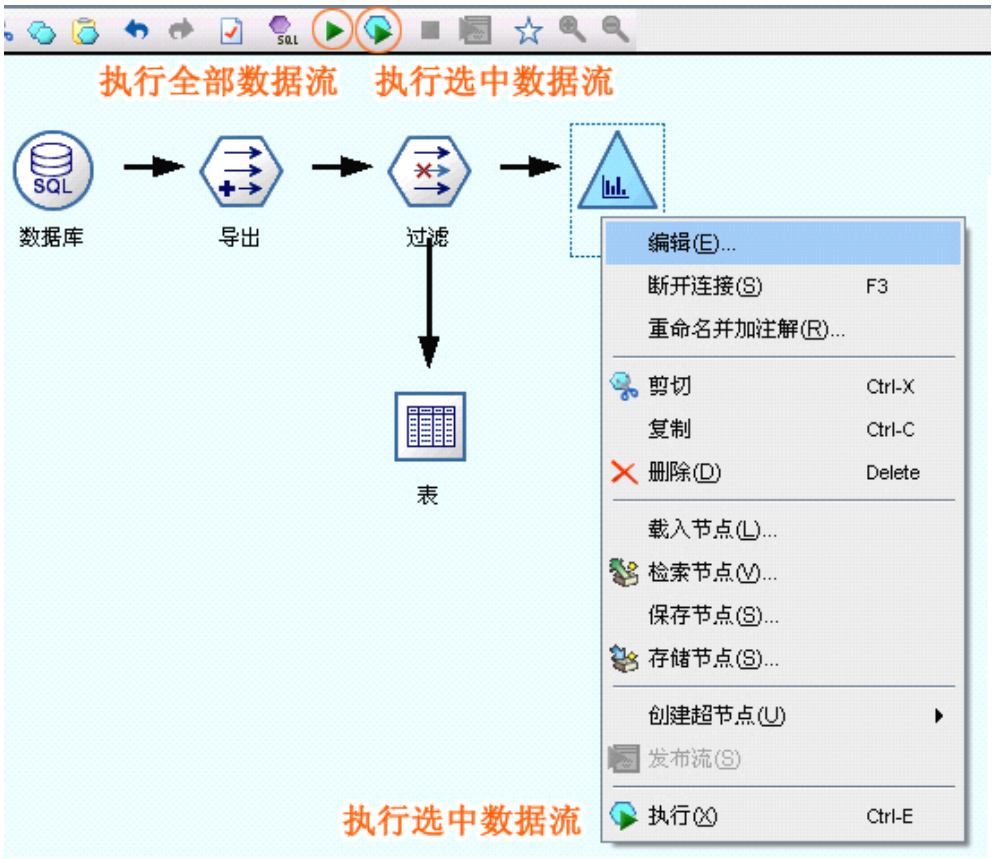
③、选择待删除连接的结点，从主菜单中选择断开连接(Edit Node Disconnect)。

2.7数据流的执行数据流结构构建好后要通过执行数据流数据才能从读入开始流向各个数据结点。执行数据流的方法有以下三种：

①、选择菜单栏中的按钮，数据流区域内的所有数据流将被执行；

②、先选择要输出的数据流，再选择菜单栏中的按钮，被选的数据流将被执行；

③、选择要执行的数据流中的输出结点，单击鼠标右键，在弹出的菜单栏中选择Execute选项，执行被选中的数据流。



三、模型建立

在这部分我们将介绍五种分析方法的建立过程，它们分别是因子分析、关联分析、聚类分析、决策树分析和神经网络。为了方便大家练习，我们将采用Clementine自带的示例，这些示例在demos文件夹中均可找到，它们的数据文件也在demos文件夹中。在模型建立过程中我们将介绍各个结点的作用。

1、因子分析(factor. str)

研究从变量群中提取共性因子的统计技术。最早由英国心理学家 C. E. 斯皮尔曼提出。他发现学生的各科成绩之间存在着一定的相关性，一科成绩好的学生，往往其他各科成绩也比较好，从而推想是否存在某些潜在的共性因子，或称某些一般智力条件影响着学生的学习成绩。因子分析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，还可检验变量间关系的假设。

因子分析的主要目的是用来描述隐藏在一组测量到的变量中的一些更基本的，但又无法直接测量到的隐性变量 (latent variable, latent factor)。比如，如果要测量学生的学习积极性(motivation)，课堂中的积极参与，作业完成情况，以及课外阅读时间可以用来反应积极性。而学习成绩可以用期中，期末成绩来反应。在这里，学习积极性与学习成绩是无法直接用一个测度(比如一个问题)测准，它们必须用一组测度方法来测量，然后把测量结果结合起来，才能更准确地来把握。换句话说，这些变量无法直接测量。可以直接测量的可能只是它所反映的一个表征(manifest)，或者是它的一部分。在这里，表征与部分是两个不同的概念。表征是由这个隐性变量直接决定的。隐性变量是因，而表征是果，比如学习积极性是课堂参与程度(表征测度)的一个主要决定因素。

那么如何从显性的变量中得到因子呢？因子分析的方法有两类。一类是探索性因子分析，另一类是验证性因子分析。探索性因子分析不事先假定因子与测度项之间的关系，而让数据“自己说话”。主成分分析是其中的典型方法。验证性因子分析假定因子与测度项的关系是部分知道的，即哪个测度项对应于哪个因子，虽然我们尚且不知道具体的系数。

示例factor.str是对孩童的玩具使用情况的描述，它一共有76个字段。过多的字段不仅增添了分析的复杂性，而且字段之间还可能存在一定的相关性，于是我们无需使用全部字段来描述样本信息。下面我们将介绍用Clementine进行因子分析的步骤：

Step一：读入数据

数据源(Source)栏中的结点提供了读入数据的功能，由于玩具的信息存储为toy_train.sav，所以我们

需要使用SPSS文件(SPSS File)结点来读入数据。双击SPSS文件(SPSS File)结点使之添加到数据流程区内，双击添加到数据流程区里的SPSS文件(SPSS File)结点，由此来设置该结点的属性。

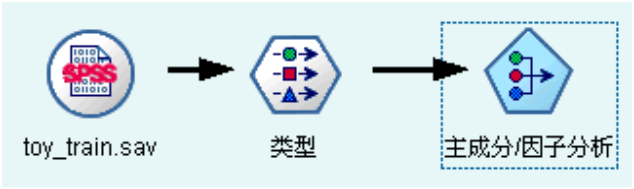
在属性设置时，单击导入文件(Import file)栏右侧的按钮，选择要加载到数据流中进行分析的文件，这里选择toy_train.sav。单击注解(Annotations)页，在名称(name)栏中选择定制(custom)选项并在其右侧的文本框中输入自定义的结点名称。这里我们按照原示例输入toy_train。

Step二：设置字段属性

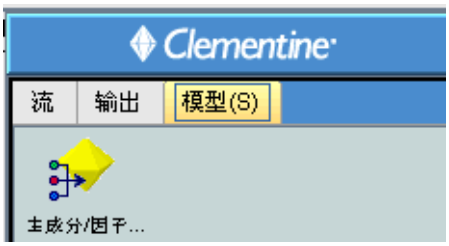
进行因子分析时我们需要了解字段间的相关性，但并不是所有字段都需要进行相关性分析，比如“序号”字段，所以需要我们将要进行因子分析的字段挑选出来。字段选项(Field Ops)栏中的类型(Type)结点具有设置各字段数据类型、选择字段在机器学习中的的输入/输出属性等功能，我们利用该结点选择要进行因子分析的字段。首先，将类型(Type)结点加入到数据流中，双击该结点对其进行属性设置：

由上图可看出数据文件中所有的字段名显示在了字段(Field)栏中，类型(Type)表示了每个字段的数据类型。我们不需要为每个字段设定数据类型，只需从Values栏中的下拉菜单中选择<Read>项，然后选择读取值(Read Value)键，软件将自动读入数据和数据类型；缺失(Missing)栏是在数据有缺失 时选择是否用空(Blank)填充该字段；检查(Check)栏选择是否判断该字段数据的合理性；而方向(Direction) 栏在机器学习模型的建立中具有相当重要的作用，通过对它的设置我们可将字段设为输入/ 输出/输入且输出/非输入亦非输出四种类型。在这里我们将前19个字段的方向(Direction) 设置为无(none)，这表明在因子分析我们不将这前19个字段列入考虑，从第20个字段起我们将以后字段的方向(direction)设置为输入(In)，对这些字段进行因子分析。

Step三：对数据行因子分析因子分析模型在建模(Modeling)栏中用主成分/因子分析(PCA/Factor)表示。在分析过程中模型需要有大于或等于两个的字段输入，上一步的Type结点中我们已经设置好了将作为模型输入的字段，这里我们将主成分/因子分析(PCA/Factor)结点连接在类型(Type)结点之后不修改它的属性，默认采用主成分分析方法。



在建立好这条数据流后我们便可以将它执行。右键单击主成分/因子分析(PCA/Factor)结点，在弹出的菜 单栏中选择执行(Execute)命令。执行结束后，模型结果放在管理器的模型(Models)栏中，其标记为名称为主成分/因子分析(PCA/Factor)的黄色结点。



右键单击该结果结点，从弹出的菜单中选择浏览(Browse)选项查看输出结果。由结果可知参与因子分析的字段被归结为了五个因子变量，其各个样本在这五个因子变量里的得分也在结果中显示。

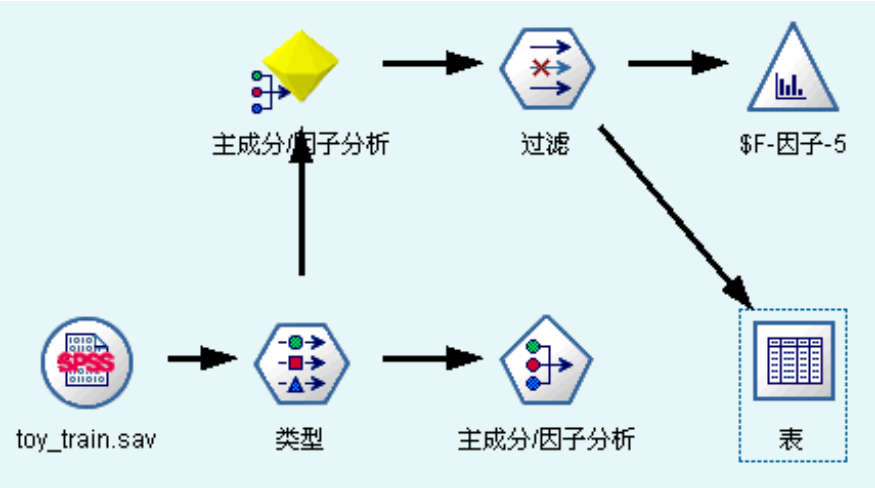
Step四：显示经过因子分析后的数据表

模型的结果结点也可以加入到数据流中对数据进行操作。我们在数据流程区内选中类型(Type)结点，然后双击管理器模型(Models)栏中的PCA/Factor结点，该结点便加入到数据流中。为了显示经过因子分析后的数据我们可以采用表格(Table)结点，该结点将数据由数据表的形式输出。

4.1为因子变量命名在将PCA/Factor（结果）结点连接到表格(Table)结点之前，用户可以设置不需要显示的字段，也可以更改因子变量名，为了达到这个目的我们可以添加字段选项(Field Ops)栏中的字段(filter)结点。

在对过滤(filter)结点进行属性设置时，过滤(filter)项显示了字段的过滤与否，如果需要将某个字段过滤，只需用鼠标单击Filter栏中的箭头，当箭头出现红“×”时该字段便被过滤。 第一个字段(Field)栏结点表明数据在读入过滤(filter)结点时的字段名，第二个字段(Field)栏表示数据经过过滤(filter)结点后的字段名。由于因子分析生成的因子变量都由系统自动命名，用户可以通过修改这些因子变量的第二个字段(Field)的值来重新设定其字段名。

4.2数据输出显示,在对数据进行输出时我们选择了输出(Output)栏中的表格(Table)结点和图形(Graph)栏中的柱状图（Histogram）结 点。这两个结点一个通过数据表的形式输出，一个通过柱装图的形式输出。 对柱装图我们设置其显示store play字段的数据（store_play为第五个因子变量的新名）。通过“执行”按钮分别执行两条数据流，将经过因子分析后的数据显示。



P.S. ：在这个因子分析的案例中我们用到了SPSS**文件(SPSS File)、类型(Type)、过滤(Filter)、表格(Table)、柱状图(Histogram)、PCA/Factor**结点。

2.关联分析、决策树分析（**baskrule.str**）

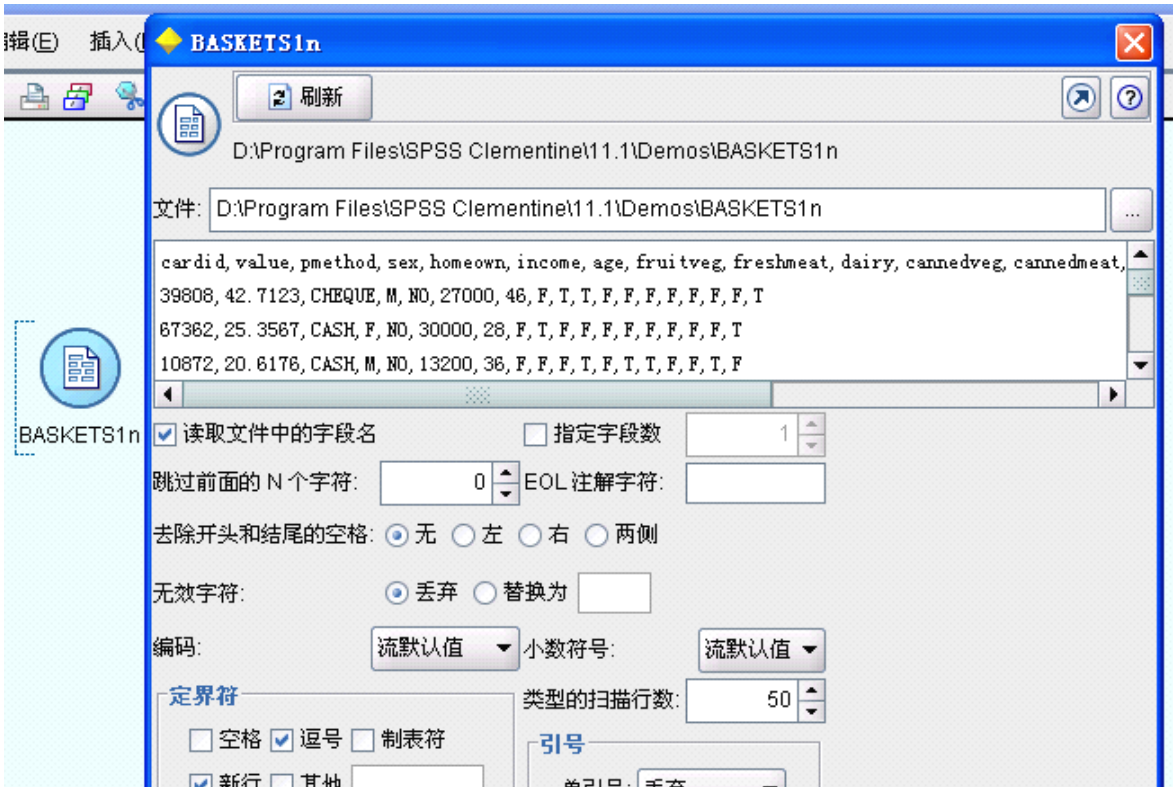
关联分析是指如果两个或多个事物之间存在一定的联系,那么其中一个事物就能通过其他事物进行预测.它的目的是为了挖掘隐藏在数据间的相互关系

在数据挖掘的基本任务中关联(ass()ciation)和顺序序贯模型(seqtlencing)关联分析是指搜索事务数据库(trarisactionaldatabases)中的所有细节或事务,从中寻找重复出现概率很高的模式或规则。

示例baskrule.str是针对某商场的购物资料对数据进行分析。为了找出商品在出售时是否存在某种联系，我们将使用关联分析方法；为了得到购买某种商品的顾客特征，我们将采用决策树方法对顾客分类。

Step一： 读入数据

该模型的数据文件存储为BASKETS1n，我们选择Source栏的Var. File（自由格式文本文件）结点作为数据读入结点，双击该结点进行属性设置。



Step二：关联分析从数据源读入数据后我们需要根据要进行的分析对字段进行设置。关联分析是分析多个量之间的关系，所以需要将进行分析的字段既设置为模型的输入又设置为模型的输出，对字段的设置可以通过Type结点进行。

2.1为数据设置字段格式

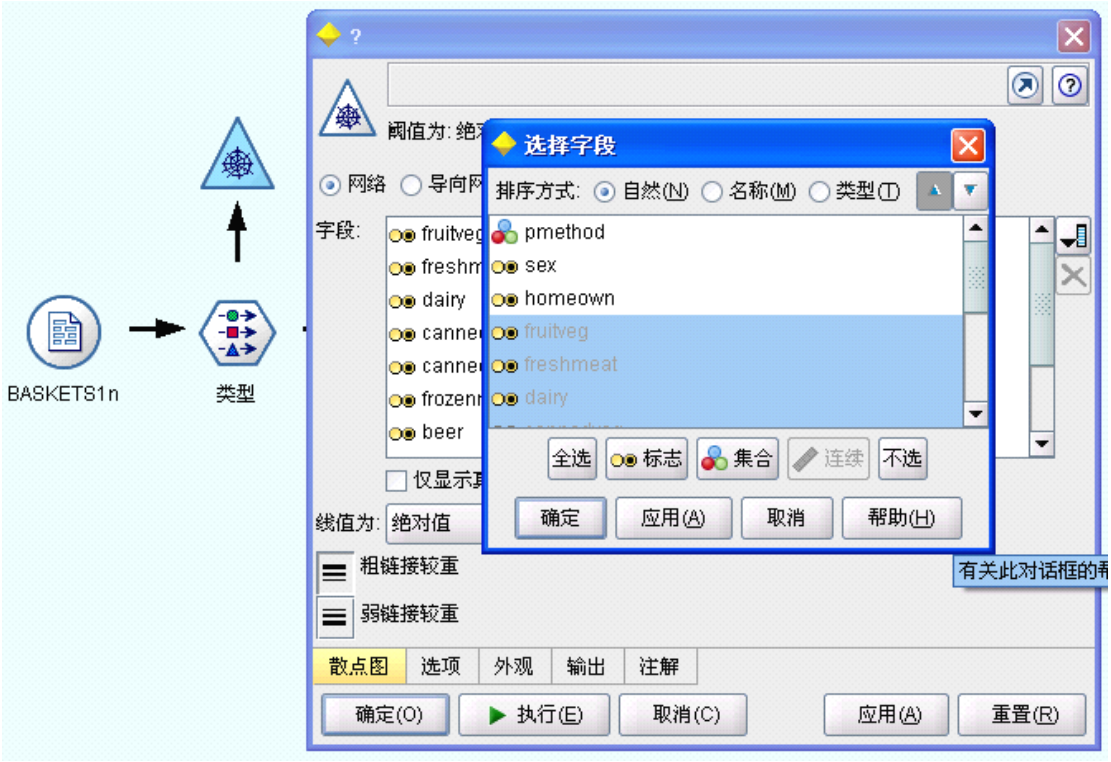
在数据流程区内选中已存在的Var. File结点，双击文件选择(File OPs)栏中的类型(Type)结点，将类型(Type)结点加入到数据流中。由于我们的分析是对商品进行，与顾客的个人信息无关，所以在类型(Type)中将顾客个人信息字段的方向(Direction)设为空(none)，其他商品字段的方向(Direction)设为双向(Both)。同时我们也将读入字段类型和字段取值。

2.2生成关联分析数据流

Clementine提供了三个可以进行关联分析的模型，他们分别是Apriori、GRI、Sequence，在这里我们选择GRI结点加入到数据流中。执行该数据流，它的结果将在在管理器的Models栏中以与模型同名的结点显示，右键选择浏览该结点，结果如下图：

*结果数据表显示了各种商品间的关系，该表的每一行表明了当某种商品被购买时还有哪些产品可能被同时购买，它是居于关联分析中的支持度和可信度来分析的。

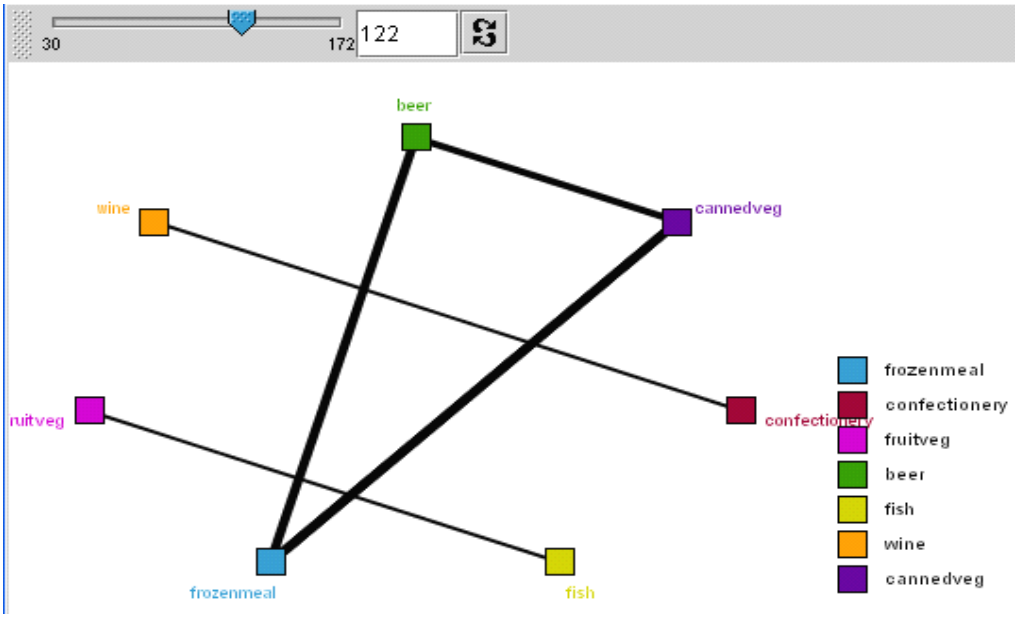
Step三：图形化显示各商品之间的关系对数据进行关联分析除了利用模型外，我们还可以利用Graphs栏中的Web结点将它们之间的关系通过网状图显示。 选中Web结点将它连接到Type结点上，对Web结点的属性设置如下图所示：

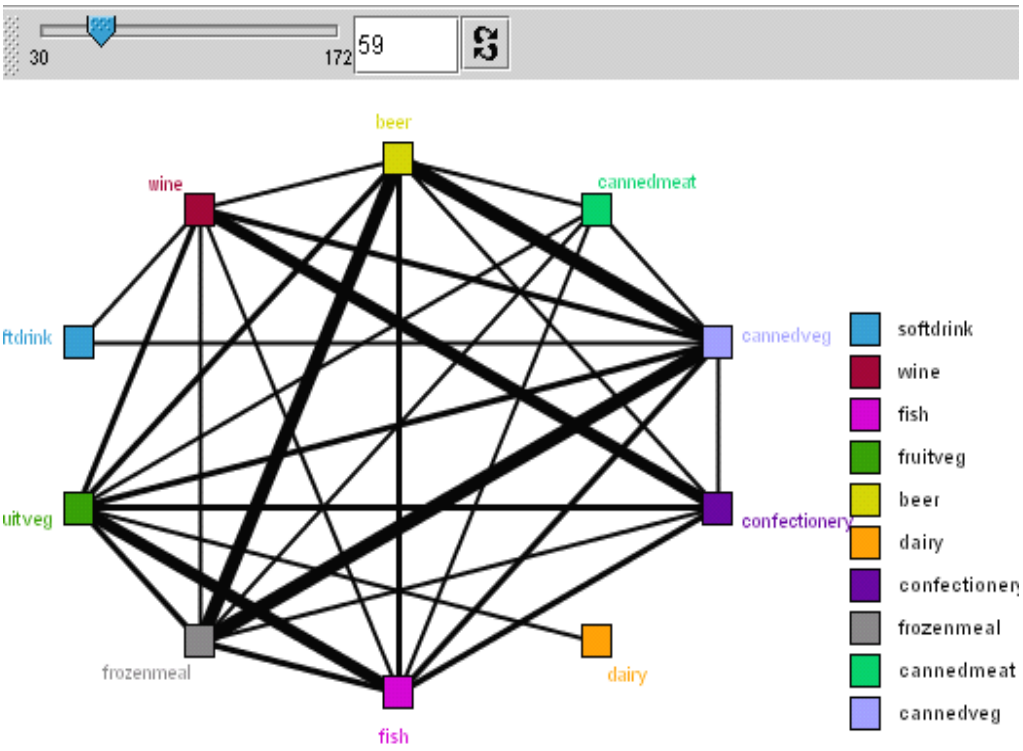


选择Fields栏右边的打开对话框按钮，弹出如上图所示的选择字段(Select Fields)对话框。选出将要作关联分析的项，确定后返回Web属性菜单。

在plot面板中选中“仅显示真值标志(show true tag only)”栏可帮我们简化输出网络。在Web结点的属性设置好后我们可以运行这条数据流，运行结果如下左图所示。

*各色的结点代表了各种不同的商品，任两点的连线越粗表明这两点间的关系越强烈。我们还可以通过改变浮标值设置不同的显示，当浮标值越大时web图将显示拥有越强关系的点（如下右图所示）。





决策树(decision tree)一般都是自上而下的来生成的。每个决策或事件（即自然状态）都可能引出两个或多个事件，导致不同的结果，把这种决策分支画成图形很像一棵树的枝干，故称决策树。 决策树就是将决策过程各个阶段之间的结构绘制成一张箭线图。

选择分割的方法有好几种，但是目的都是一致的：对目标类尝试进行最佳的分割。

从根到叶子节点都有一条路径，这条路径就是一条“规则”。

决策树可以是二叉的，也可以是多叉的。

对每个节点的衡量：

1) 通过该节点的记录数

2) 如果是叶子节点的话，分类的路径

3) 对叶子节点正确分类的比例

有些规则的效果可以比其他的一些规则要好。

决策树对于常规统计方法的优缺点

优点：

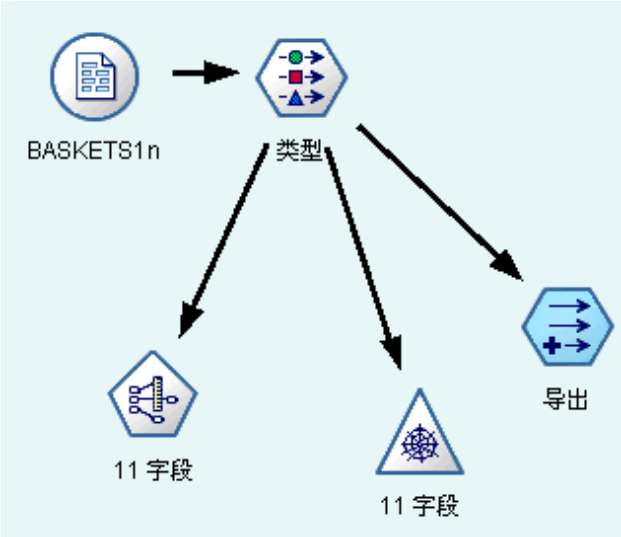
1) 可以生成可以理解的规则；
2) 计算量相对来说不是很大；
3) 可以处理连续和种类字段；
4) 决策树可以清晰的显示哪些字段比较重要。

缺点：

1) 对连续性的字段比较难预测；
2) 对有时间顺序的数据，需要很多预处理的工作；
3) 当类别太多时，错误可能就会增加的比较快；
4) 一般的算法分类的时候，只是根据一个字段来分类。

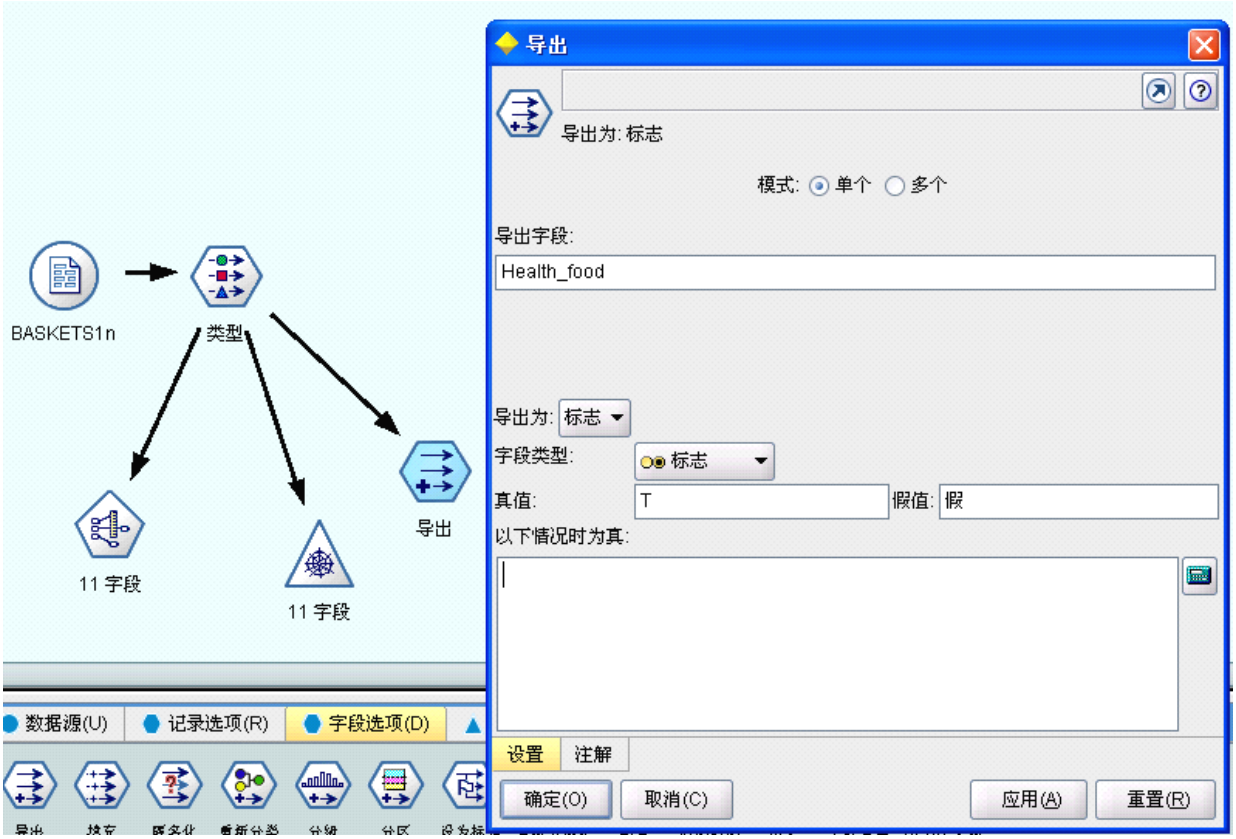
Step四：用决策树进行分类分析在本例中我们运用决策树对购买某样商品的客户进行分类，通过分析他的个人信息（例如年龄、收入等）判断怎样的人会购买健康食品。在用决策树建模时我们需要设置一个输出结点，模型根据样本在该结点的不同取值构造出决策树。

4.1将**导出(Derive)**结点连接到**Type**结点后



Derive结点在字段选项(Field OPs)栏中，可选用任何一种结点连入数据流的方法将这个结点连接；

4.2设置**Drive**结点的属性双击Drive结点打开属性对话框，如下图所示：

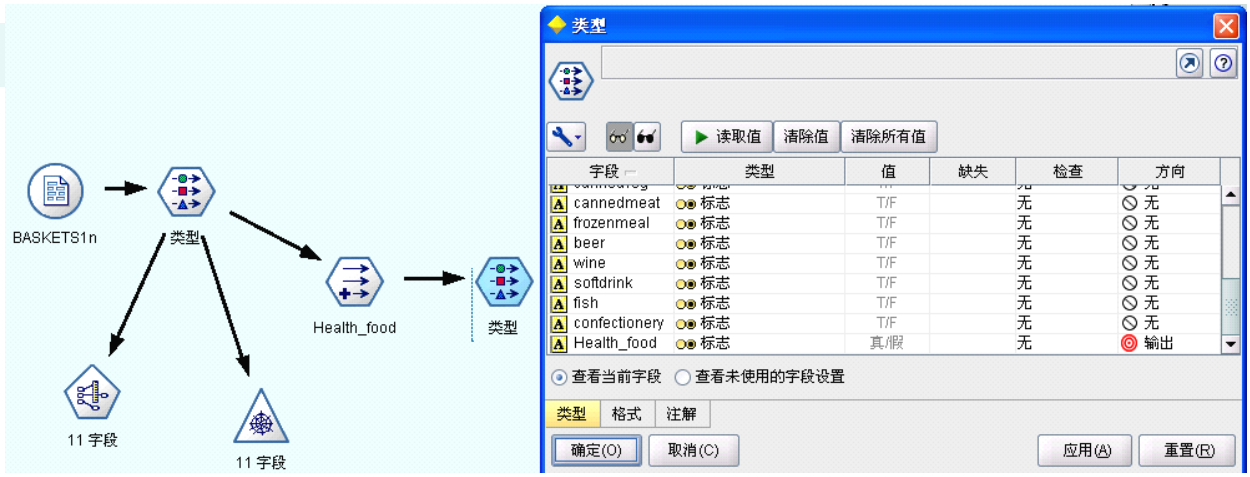


在Drive Field栏中将该结点命名为health_food，在导出为(Drive as)栏中选择Flag，这表明新生成的health_food字段将存储两值类型的数据。在真值(True value)和假值(False value)栏中分别填写新字段的两种数据值，其中真值(True value)表示当条件满足时该字段的值，假值(False value)表明当条件不满足时该字段的值。

对判断条件的设置我们可以通过单击True when栏右边的按钮进行。在表达式构建器(Expression Builder)中我们可以选择数据的任一字段，通过设计表达式建立结果为真时的条件。这里我们设置表达式为fruitveg = 'T' and fish = 'T'，这表明当顾客购买了fruitveg 和fish时该顾客便购买了健康食物。

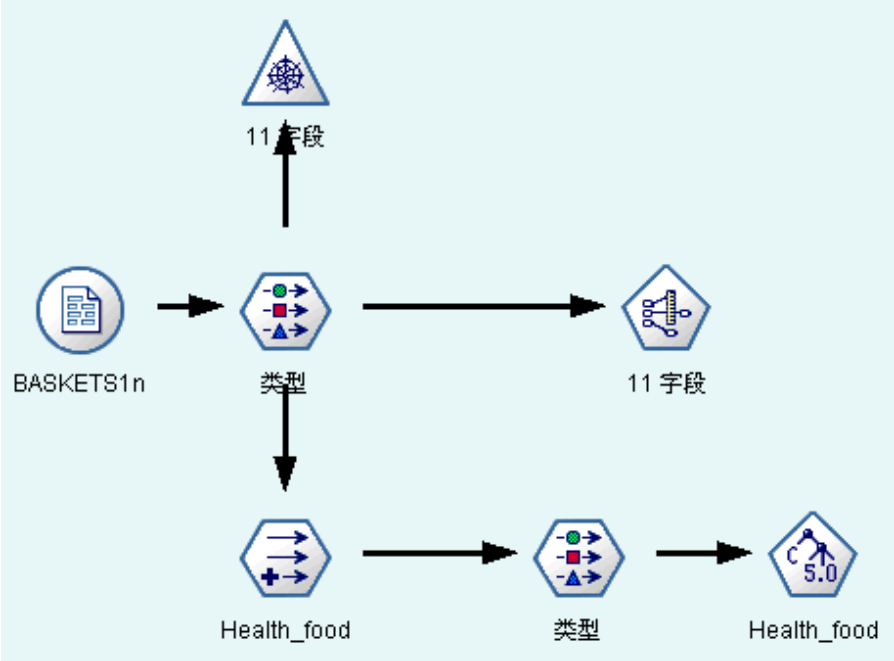
4.3设置字段的输入/输出方向

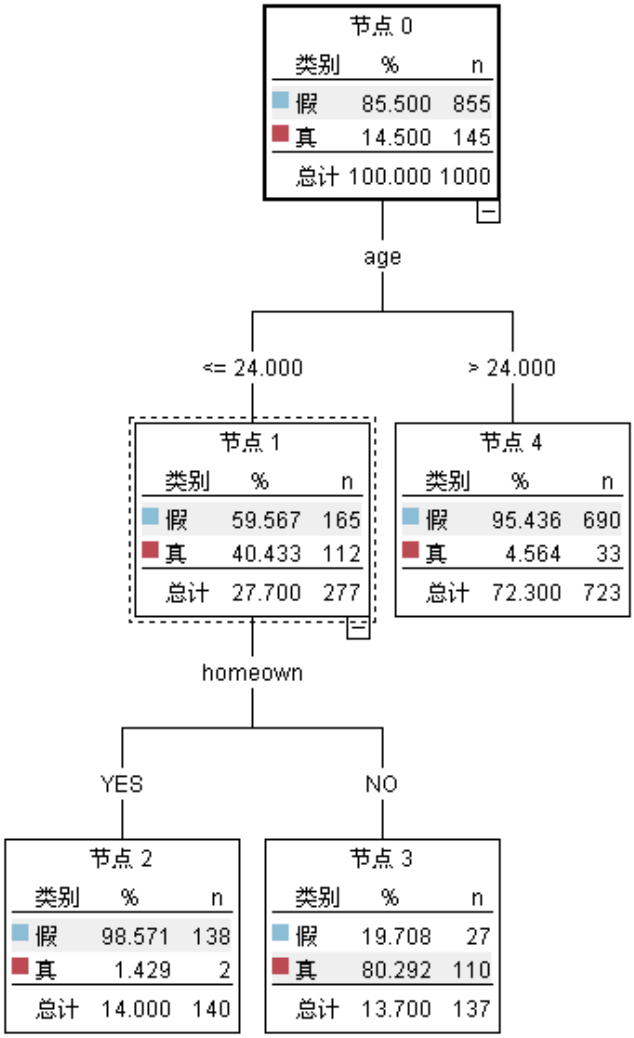
要用决策树模型建模就需要在数据载入模型前定义一个输出字段，这里我们通过在health_food结点后添加一Type结点来定制字段的输入/输出方向。由于我们要分析购买健康食物的顾客特征，所以我们将health_food字段的Direction选项设置为输出(Out)，将顾客的各个特征设置为输入(In)，将其他商品设置为无(None)。



4.4数据流的最终建立

在对字段定义结束后，我们将C5.0（决策树模型）结点加入到数据流。其数据流建立如下图：





运行建立了决策树的数据流，我们可得到输出结果如下树形图所示。该树的叶结点表明了怎样的顾客将选择健康食品，怎样的顾客将拒绝健康食品，我们也可以根据该树的将客户按是否购买健康食品进行分类。

P.S. :在这个关联分析/决策树分析的案例中我们用到了**Var. File**、**Derive**、**Web**、**GRI**和**C5.0**结点。

3.聚类分析（**cluster.str**）

聚类分析指将物理或抽象对象的集合分组成为由类似的对象组成的多个类的分析过程。它是一种重要的人类行为。

聚类与分类的不同在于，聚类所要求划分的类是未知的。

聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。

聚类分析的目标就是在相似的基础上收集数据来分类。聚类源于很多领域，包括数学，计算机科学，统计学，生物学和经济学。在不同的应用领域，很多聚类技术都得到了发展，这些技术方法被用作描述数据，衡量不同数据源间的相似性，以及把数据源分类到不同的簇中。

从统计学的观点看，聚类分析是通过数据建模简化数据的一种方法。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类等。采用k-均值、k-中心点等算法的聚类分析工具已被加入到许多著名的统计分析软件包中，如SPSS、SAS等。

从机器学习的角度讲，簇相当于隐藏模式。聚类是搜索簇的无监督学习过程。与分类不同，无监督学习不依赖预先定义的类或带类标记的训练实例，需要由聚类学习算法自动确定标记，而分类学习的实例或数据对象有类别标记。聚类是观察式学习，而不是示例式的学习。

从实际应用的角度看，聚类分析是数据挖掘的主要任务之一。而且聚类能够作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的聚簇集合作进一步地分析。聚类分析还可以作为其他算法（如分类和定性归纳算法）的预处理步骤。

Clementine提供了多种可用于聚类分析的模型，包括Kohonen，Kmeans，TwoStep方法。示例Cluster.str是对人体的健康情况进行分析，通过测量人体类胆固醇、Na、Ka等的含量将个体归入不同类别。示例中采用了三种方法对数据进行分类，这里我们重点讨论Kmeans聚类方法。

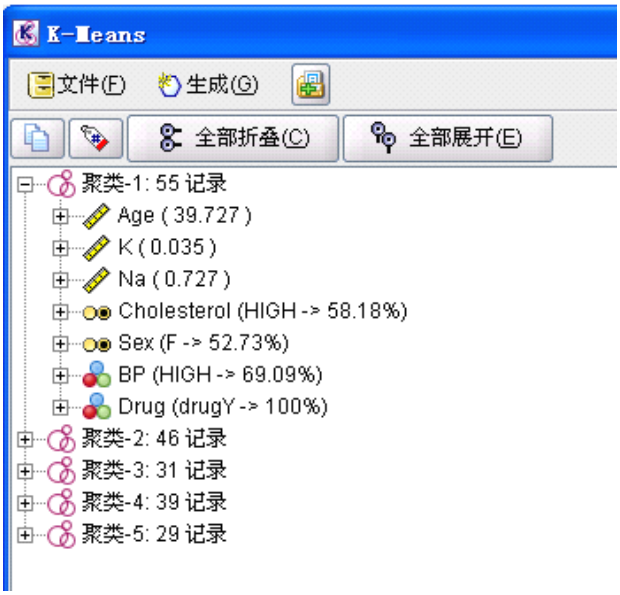
Step一：读入数据和前两步一样，在建立数据流时首先应读入数据文件。该示例中数据文件存储为DRUG1n，我们向数据流程区内添加Var. File结点读入数据。

Step二：为数据设置字段格式将Type结点连入数据流，通过编辑该结点对数据字段进行设置。

在机器学习方法中聚类被称为无导师的学习。所谓无导师的学习是指事先并不知道数据的分类情况，就像在决策树方法中我们通过已知的某个结点值来建立模型，在聚类方法中所有参与聚类的字段在设置字段格式时其方向(Direction)都将被设置为输入(In)。

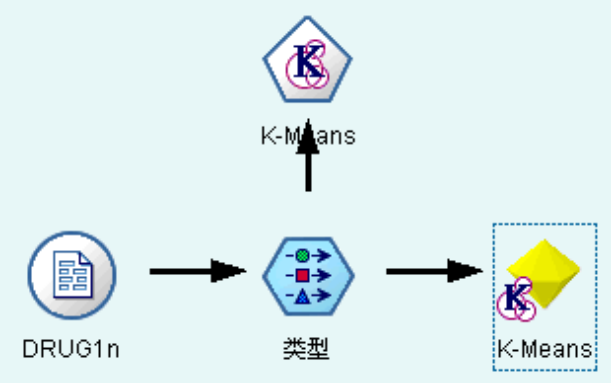
Step三：生成聚类分析数据流设置好字段格式后我们将Kmeans结点加入到数据流。在编辑Kmeans结点时我们重点需要定义将要其分成的类别数，这个属性在聚类数(Specified number of cluster)中设定。

数据流建立好后，右键单击Kmeans结点选择执行该数据流。执行结果以与Kmean同名的结 点显示在管理器的Models窗口中，浏览该结点我们能够得到关于分类的信息，如下图所示：



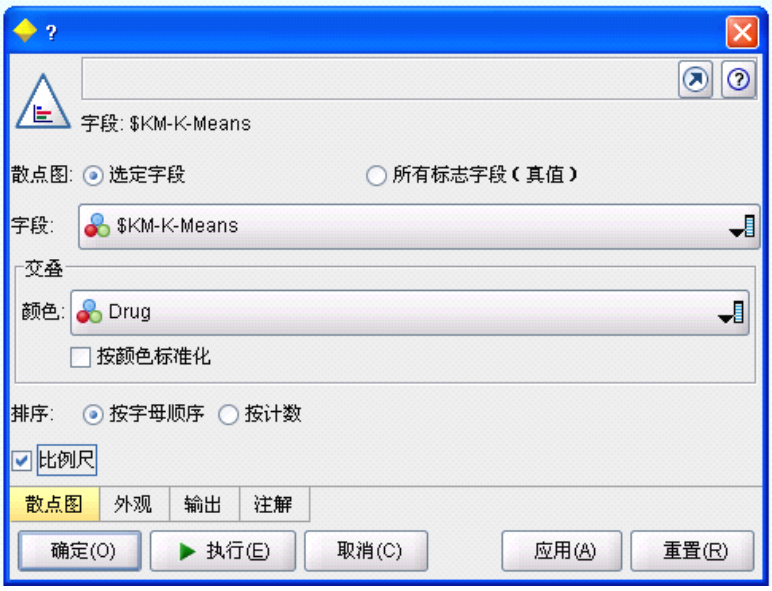
Step四：图形化输出各个类的组成情况查看各类中的情况除了浏览结果结点外，我们还可以选择用图形将结果显示出来。

4.1将模型的结果结点连入数据流。



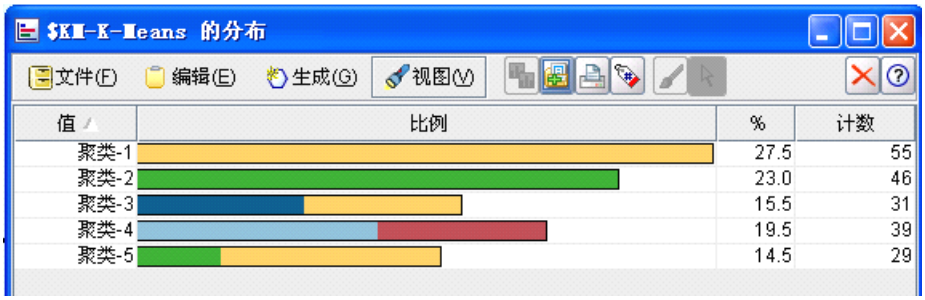
选中Type结点，双击Models窗口中的Kmeans结果结点将该结点连接到Type后

4.2设置图形输出结点 选择Graph栏中的Distribution结点将它连接到Kmeans结果结点后，双击该结点对它进行设置。



在Field栏中选择$KM-Kmeans选项，该选项保存了分类结果，即每个样本在聚类后所属的 类别。Distribution结点要求Field栏为非数据结点。在Overlay选项中我们选择Drug项，这是为了 研究在不同的分类类别里Drug的各个取值的所占比例。

运行该数据流我们可得到下图,图中详细的显示了不同Drug类型在各个类别里的分布情况。同样道理，我们也可以对其他属性进行研究。



P.S. :在这个聚类分析的案例中我们用到了**Kmeans**、**Distribution**结点。

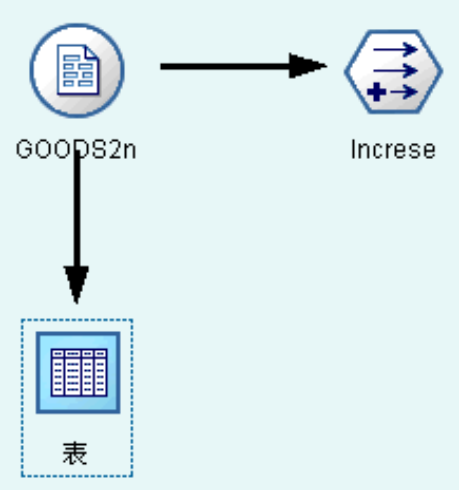
4、神经网络（**goodlearn.str**）

神经网络是一种仿生物学技术，通过建立不同类型的神经网络可以对数据进行预存、分类等操作。

示例goodlearn.str通过对促销前后商品销售收入的比较，判断促销手段是否对增加商品收益有关。Clementine提供了多种预测模型，包括Nerual Net、Regression和Logistic。这里我们用神经网络结点建模，评价该模型的优良以及对新的促销方案进行评估。

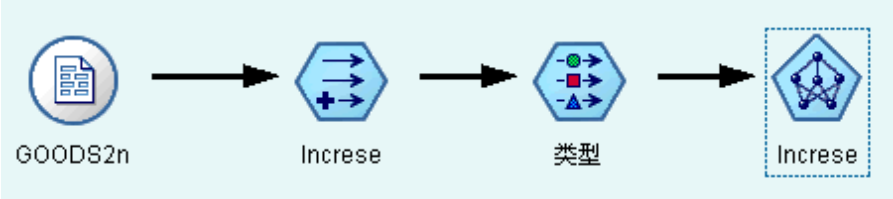
Step一：读入数据，本示例的数据文件保存为GOODS1n，我们向数据流程区添加Var. File结点，并将数据文件读入该结点。

Step二、计算促销前后销售额的变化率向数据流增加一个Derive结点，将该结点命名为Increse。在公式栏中输入(After - Before) / Before * 100.0以此来计算促销前后销售额的变化



Step三：为数据设置字段格式添加一个Type结点到数据流中。由于在制定促销方案前我们并不知道促销后商品的销售额，所以将字段After的Direction属性设置为None；神经网络模型需要一个输出，这里我们将Increase字段的Direction设置为Out，除此之外的其它结点全设置为In。

Step四：神经网络学习过程



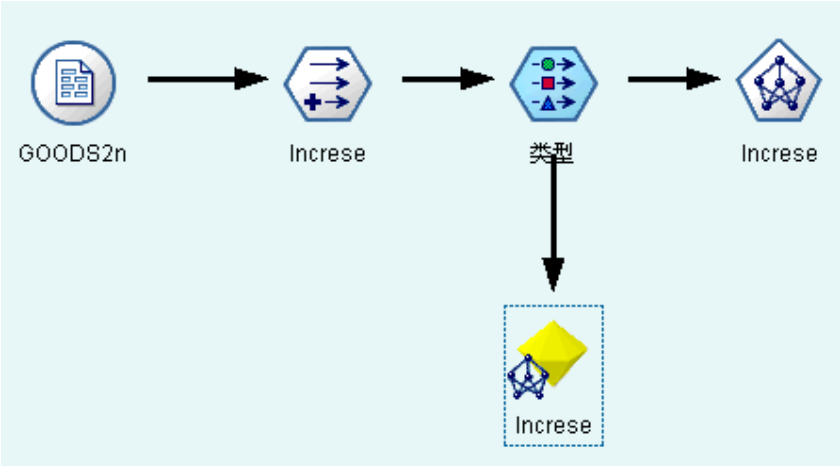
在设置好各个字段的Direction方向后我们将Neural Net结点连接入数据流。

在对Neural Net进行设置时我们选择快速建模方法（Quick），选中防止过度训练(Prevent overtraining)。同时我们还可以根据自己的需要设置训练停止的条件。

在建立好神经网络学习模型后我们运行这条数据流，结果将在管理器的Models栏中显示。 选择查看该结果结点，我们可以对生成的神经网络各个方面的属性有所了解。

Step四：为训练网络建立评估模型

4.1将模型结果结点连接到数据流将Increase结果结点连接在数据流中的Type结点后；



4.2添加字段比较预测值与实际值 向数据流中增加Derive结点并将它命名为ratio，然后将它连接到Increase结果结点。设置该结点属性，将增添的字段的值设置为(abs(Increase - '$N-Increase') /Increase) * 100，其中$N-Increase是由神经网络生成的预测结果。通过该字段值的显示我们可以看出预测值与实际值之间的差异大小。

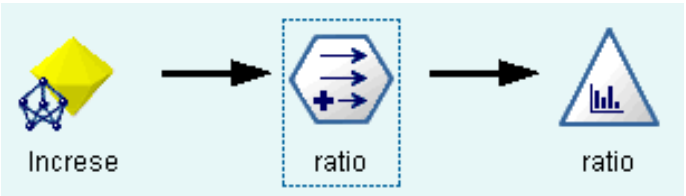
表（8 个字段，400 条记录） #1

文件(F) 编辑(E) 生成(G)

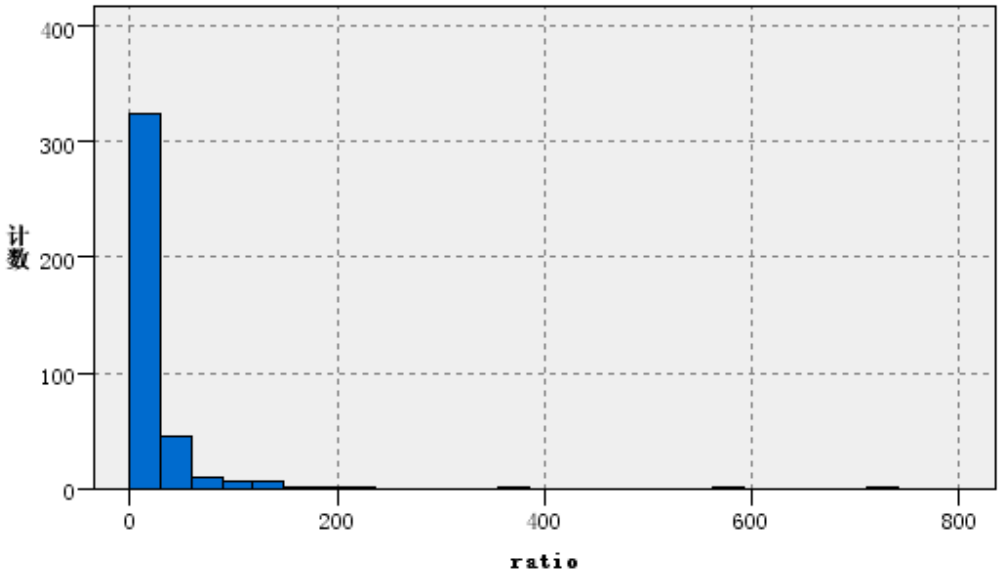
	Class	Cost	Promotion	Before	After	Increse	$N-Increse
1	Luxury	31.2...	1467	2233...	238333	6.704	5.789
2	Drink	82.5...	1316	1989...	219791	10.459	9.166
3	Luxury	10.4...	1734	2480...	266357	7.361	7.186
4	Drink	40.4...	1002	2159...	235013	8.808	7.269
5	Drink	20.2...	1127	2890...	305659	5.760	8.118
6	Meat	59.3...	1884	2347...	241302	2.793	4.348
7	Meat	71.1...	1655	2087...	216708	3.828	3.483
8	Drink	62.7...	1108	1922...	204458	6.373	7.900
9	Drink	98.2...	1075	2342...	248692	6.157	7.280
10	Drink	34.6...	1644	1109...	121988	9.900	12.005
11	Luxury	87.4...	1105	1361...	140323	3.100	4.053
12	Drink	92.7...	1828	2091...	239858	14.655	12.027
13	Luxury	66.4...	1137	1218...	126166	3.537	4.166
14	Meat	5.810	1446	2062...	214172	3.830	2.966
15	Meat	92.9...	1260	1574...	159442	1.236	2.468
16	Luxury	34.7...	1644	2375...	248668	4.679	6.634
17	Meat	69.9...	1398	2283...	238146	4.307	2.835
18	Conf...	80.3...	1007	1898...	198010	4.292	4.846
19	Luxury	20.4...	1389	2522...	267280	5.961	5.486
20	Meat	17.4...	1084	2706...	271425	0.278	2.339

表 注解

4.3评价模型 可以通过观察预测值与实际值之间的差异来评价模型的优劣。从Graph栏中选择histogram结点连接到ratio结点。



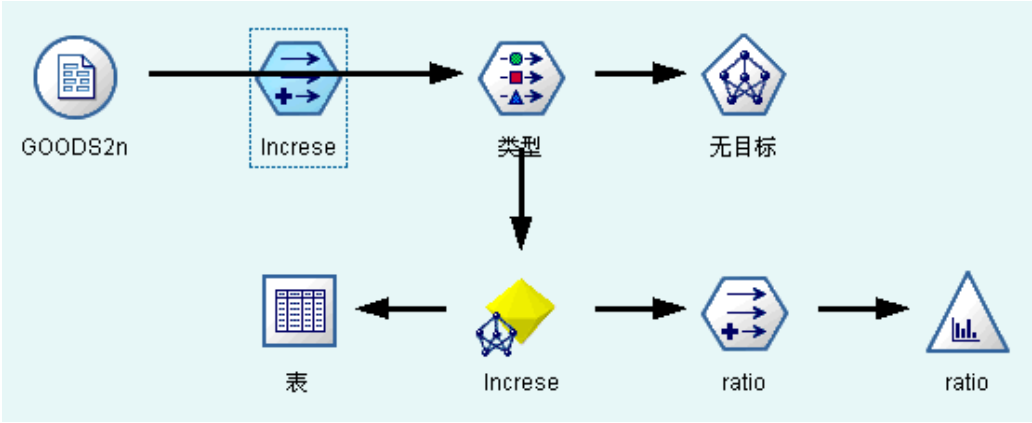
设置该结点，使其输出显示ratio的值（在field的下拉列表中选择ratio），输出结果如下图所 示：



该图形的横坐标为ratio的值，纵坐标表示一共有多少个样本的ratio值落在相对应的横坐 标上。从ratio的定义公式我们知道ratio越小表明预测值与实际值的差别越小，所以我们希望更多的ratio值处于一个比较小的范围。因此由输出图形我们可以看出该模型达到了一定的精度。

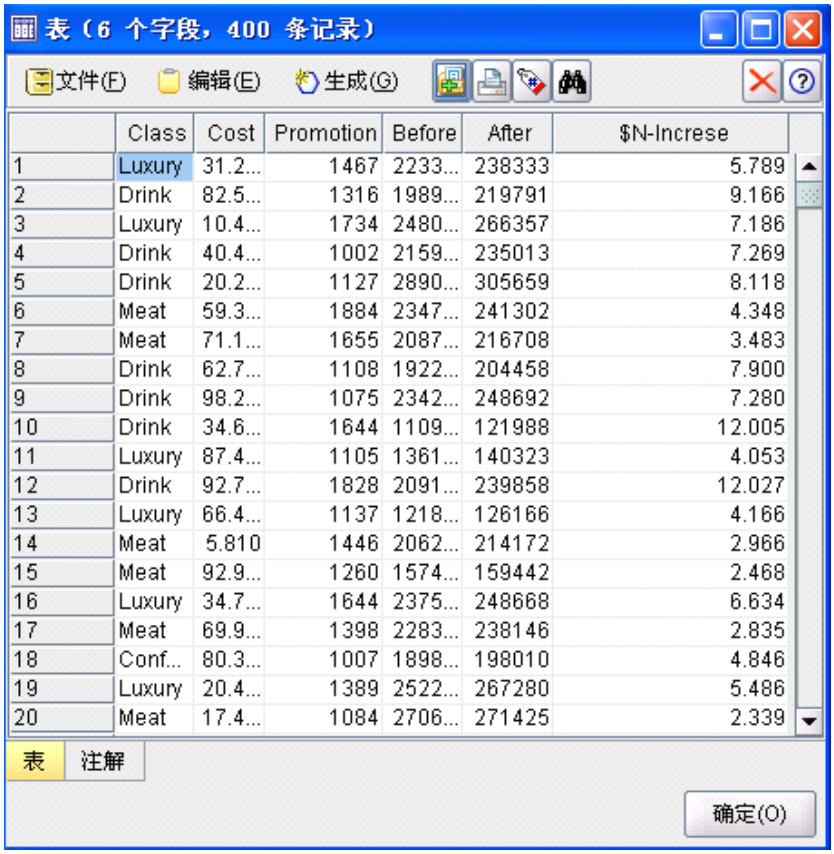
Step五：模型预测

5.1预测模型建立



该模型的建立就是为了预测新样本。我们现将数据源的文件改为GOODS2n；然后按alt 键双击Increase结点以此来绕过该结点；断开Increase结果结点与Ratio结点之间的连接，再增添 一个Table结点观察Increase结果结点的输出。在Type结点中我们只设置字段after的Direction属性 为None，其余的都为In。通过这种方法建立好的数据流如下图所示：

右键单击Table结点，选择运行数据流。运行生成的结果如下，其中$N-Increase为预测结果：



	Class	Cost	Promotion	Before	After	$N-Increse
1	Luxury	31.2...	1467	2233...	238333	5.789
2	Drink	82.5...	1316	1989...	219791	9.166
3	Luxury	10.4...	1734	2480...	266357	7.186
4	Drink	40.4...	1002	2159...	235013	7.269
5	Drink	20.2...	1127	2890...	305659	8.118
6	Meat	59.3...	1884	2347...	241302	4.348
7	Meat	71.1...	1655	2087...	216708	3.483
8	Drink	62.7...	1108	1922...	204458	7.900
9	Drink	98.2...	1075	2342...	248692	7.280
10	Drink	34.6...	1644	1109...	121988	12.005
11	Luxury	87.4...	1105	1361...	140323	4.053
12	Drink	92.7...	1828	2091...	239858	12.027
13	Luxury	66.4...	1137	1218...	126166	4.166
14	Meat	5.810	1446	2062...	214172	2.966
15	Meat	92.9...	1260	1574...	159442	2.468
16	Luxury	34.7...	1644	2375...	248668	6.634
17	Meat	69.9...	1398	2283...	238146	2.835
18	Conf...	80.3...	1007	1898...	198010	4.846
19	Luxury	20.4...	1389	2522...	267280	5.486
20	Meat	17.4...	1084	2706...	271425	2.339

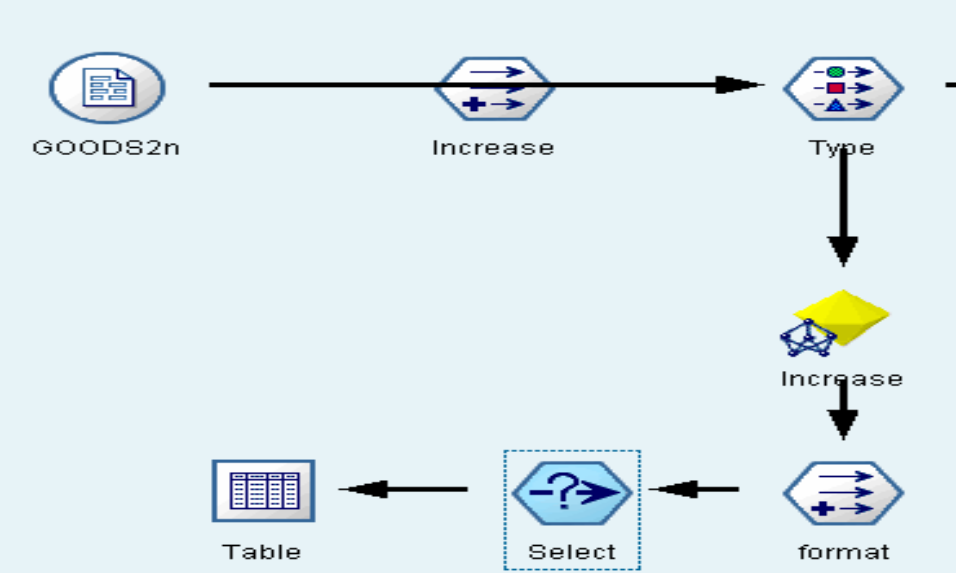
5.2输出规范化

$N-Increase栏表示促销后销售额可能增减的比率。由于神经网络的最终输出需要规范到$[0,1]$区间，所以我们选择输出值在$(0,1)$内连续的S形函数将结果规范化。S型函数表达式为$f(x)=\dfrac{1}{1+e^{-x}}$。我们通过增加Derive结点将结果其规范化。

Class	Cost	Promotion	Before	After	$N-Increase	format
Luxury	31.2...	1467	2233...	238333	5.359	0.995
Drink	82.5...	1316	1989...	219791	9.903	1.000
Luxury	10.4...	1734	2480...	266357	6.531	0.999
Drink	40.4...	1002	2159...	235013	7.865	1.000
Drink	20.2...	1127	2890...	305659	8.335	1.000
Meat	59.3...	1884	2347...	241302	4.338	0.987
Meat	71.1...	1655	2087...	216708	3.567	0.973
Drink	62.7...	1108	1922...	204458	8.668	1.000
Drink	98.2...	1075	2342...	248692	8.574	1.000
Drink	34.6...	1644	1109...	121988	11.419	1.000
Luxury	87.4...	1105	1361...	140323	4.465	0.989
Drink	92.7...	1828	2091...	239858	12.096	1.000

5.3选择促销方案 根据神经网络模型的预测输出，我们可以选出GOODS2n文件中包含的可执行促销方案。假定预测结果经规范化后结值1的方案为可执行方案，我们需要增加一个结点来选出满足这些条件

的结点。Clementine为我们提供了Select结点，它可以从数据集中筛选出满足预定条件的记录。



从Record OPs栏内选择Select结点连接到Format结点后，在它的属性设置中选择包含format＝1.000的结点，整个流程图由下所示：

	Class	Cost	Promotion	Before	After	$N-Increase	format
1	Drink	92.760	1828	2091...	239858	12.096	1.000
2	Drink	98.150	1706	2234...	247907	11.666	1.000
3	Drink	44.540	1938	1718...	196179	12.394	1.000
4	Drink	103.3...	1904	2447...	281376	12.309	1.000
5	Drink	76.190	1888	2579...	288452	12.225	1.000
6	Drink	102.4...	1718	1537...	170421	11.782	1.000
7	Drink	53.880	1902	2354...	265308	12.254	1.000
8	Drink	40.870	1720	1434...	161531	11.697	1.000
9	Drink	48.270	1690	2009...	224928	11.540	1.000
10	Drink	87.110	1824	1233...	140228	12.148	1.000
11	Drink	26.970	1711	2261...	250797	11.536	1.000
12	Drink	36.960	1945	2545...	287404	12.335	1.000
13	Drink	70.150	1714	2001...	227041	11.679	1.000
14	Drink	90.530	1697	1252...	139851	11.720	1.000
15	Drink	50.300	1783	1555...	175344	11.932	1.000

运行数据流后我们将得到可用于促销的方案。结果图如下所示：

如果我们只需要得到这些方案的某些字段，而不想知道它的全部细节，则可以在Select和Table键中增设Filter结点，将不需要的字段过滤。

P.S. :在神经网络示例的学习中，我们用到了Neural Net、Select结点。