

## 基于 IBM SPSS Modeler 14.2 的客户数据挖掘

IBM SPSS Modeler 14.2 是一个从大量数据中挖掘有用模式的企业级数据分析平台，遵循跨行业数据挖掘流程标准（CRISP-DM）。从数据源到数据建模，IBM SPSS Modeler 14.2 提供了丰富的数据挖掘流程各个阶段需要的组件。

IBM SPSS Modeler 14.2 包含数据获取、数据预处理、数据建模、评估和部署等一系列步骤，分析人员可通过拖放方式组合节点完成数据挖掘流程(以下简称数据流)。

IBM SPSS Modeler 14.2 主界面如图 1 所示，包括流工作区、节点选项卡、管理器和 IBM SPSS Modeler 工程。其中流工作区主要是用于创建数据流，用户可以把节点选项卡下的组件直接拖放到流工作区。节点选项卡有多种节点：数据源、记录选项、字段选项、图形、建模、输出和导出等。管理器主要用于管理输出和模型，用户可以对这些输出和模型进行打开、重命名、保存和删除等操作。IBM SPSS Modeler 工程允许用户以 CRISP-DM 模式管理数据流。

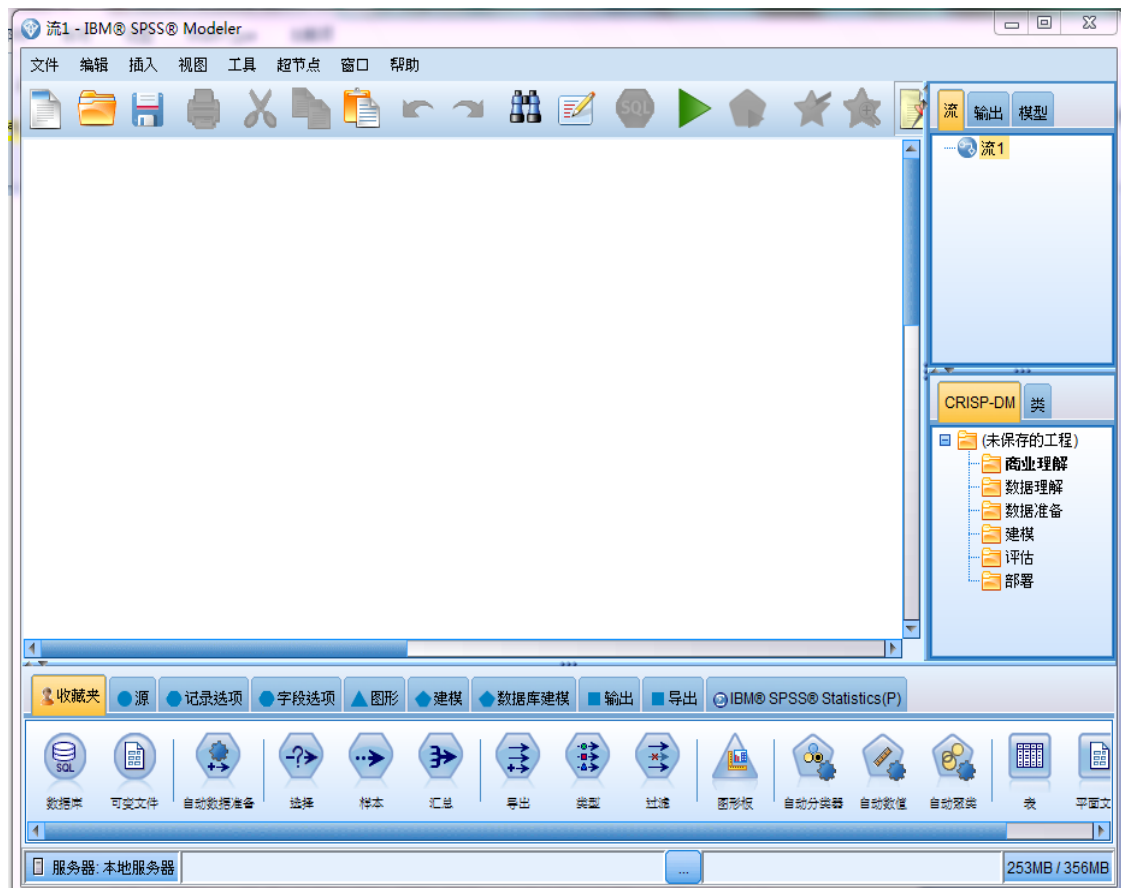


图 1 IBM SPSS Modeler 主界面

IBM SPSS Modeler 14.2 允许用户直接手动输入数据和把可变文件、Statistics 文件、SAS 文件、Excel 和 XML 等多种数据导入，以供数据分析。在导入数据后，需要对数据进行预处理。IBM SPSS Modeler 14.2 提供丰富的数据预处理组件，主要包括记录预处理和字段预处理。其中在记录预处理中，提供了选择、抽样、汇总、排序、合并和追加等组件。字段预处理包括类型、过滤、导出、分箱、字段重排、自动数据准备和分区等组件。

IBM SPSS Modeler 14.2 提供了各种来自机器学习和统计学的建模方法，如分类、关联、

聚类、序列和回归等模型。

本章应用 IBM SPSS Modeler 14.2 平台的几种常用数据挖掘算法，对客户交易的数据进行分析，获取客户管理有用的知识。

## 1 聚类分析

IBM SPSS Modeler 14.2 提供了多种聚类算法，其中主要包括 k-means 聚类算法、两步聚类算法和 Kohonen 聚类算法等，下面对这几种聚类算法结合实例进行说明。

### 1.1 K-means 聚类

首先应用 k-means 算法进行聚类分析。启动 SPSS Modeler 14.2。选择“开始”→“所有程序”→“IBM SPSS Modeler 14.2”→“IBM SPSS Modeler 14.2”，启动 SPSS Modeler 程序，如图 1 所示。

添加数据源节点。首先选择窗口底部节点选项卡中的“源”选项卡，再点击“Statistics 文件”节点，单击工作区的合适位置，把“Statistics 文件”的源添加到流中，如图 2 所示。

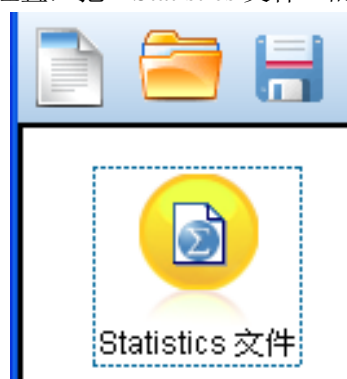


图 2 工作区中的“Statistics 文件”节点


添加数据源。右键单击工作区的“Statistics 文件”节点，选择“编辑”，打开如图 3 的编辑窗口，其中有许多选项可供选择，例如数据选项卡主要是设置变量名称和值。此处均选择默认设定。点击“导入文件”右侧的  按钮，弹出文件选择对话框，选择安装路径下“Demos”文件夹中的 tree\_credit.sav 文件，点击“打开”，如图 4 所示。单击“应用”，点击“确定”按钮，关闭编辑窗口。



图 3 “Statistics 文件”节点编辑窗口

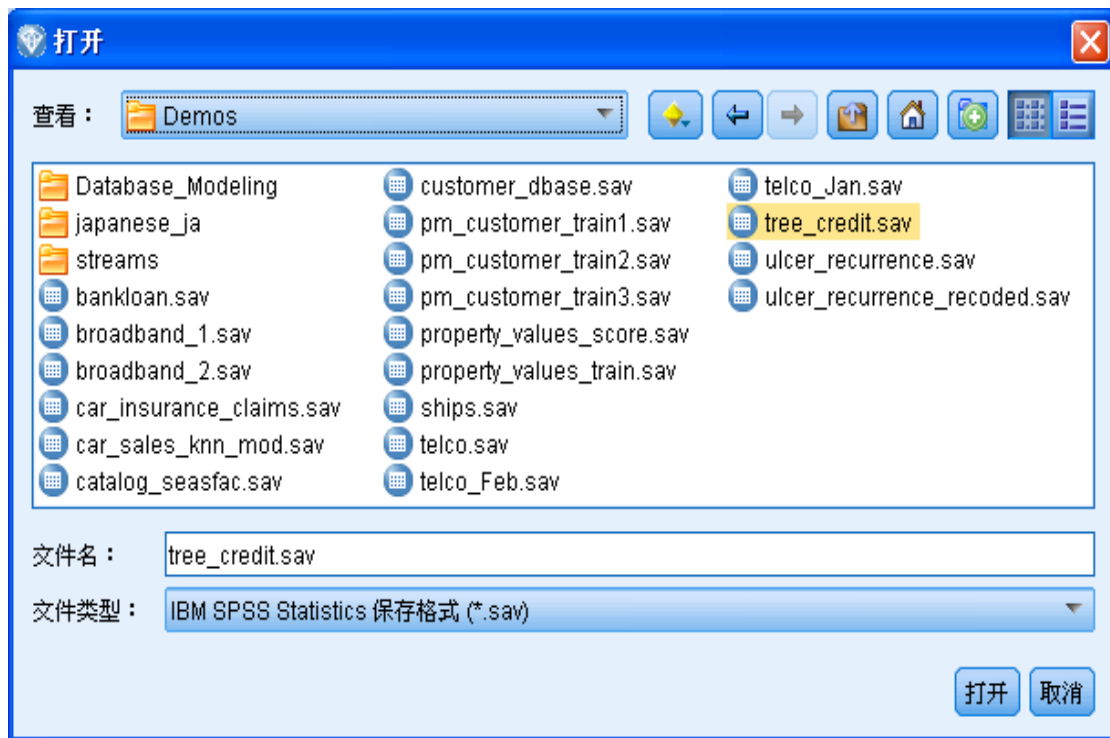


图 4 文件选择对话框

添加“表”节点。选中工作区的“tree\_credit.sav”节点，双击“输出”选项卡中的“表”节点，如图 5 所示。

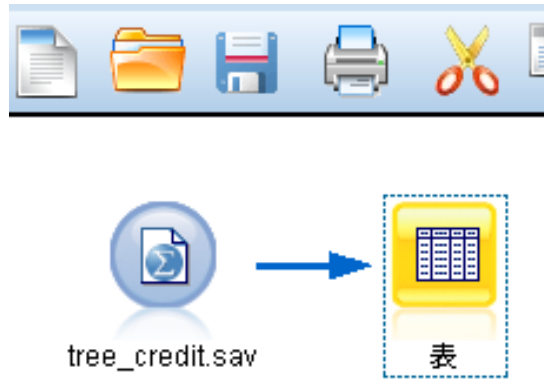


图5 工作区中的“表”节点

运行“表”节点，可以看到图6中有关银行客户信用卡的数据，包含6个字段：Credit\_rating（信用等级）、Age（年龄）、Income（收入水平）、Credit\_Cards（信用卡数量）、Education（教育水平）和Car\_loans（车贷），共1000条记录。

	Credit_rating	Age	Income	Credit_cards	Education	Car_loa...
1	0.00	36.22	2.00	2.00	2.00	2.00
2	0.00	21.99	2.00	2.00	2.00	2.00
3	0.00	29.17	1.00	2.00	1.00	2.00
4	0.00	32.75	1.00	2.00	2.00	1.00
5	0.00	36.77	2.00	2.00	2.00	2.00
6	0.00	39.32	2.00	2.00	2.00	2.00
7	0.00	31.70	2.00	2.00	2.00	2.00
8	0.00	34.72	1.00	2.00	1.00	2.00
9	0.00	31.53	1.00	2.00	1.00	2.00
10	0.00	24.78	2.00	2.00	2.00	2.00
11	0.00	22.76	1.00	2.00	2.00	2.00

图6 查看数据

添加类型节点。选中“tree\_credit.sav”节点，在“字段选项”选项卡中双击“类型”节点，“类型”节点出现在工作区中，如图7所示。

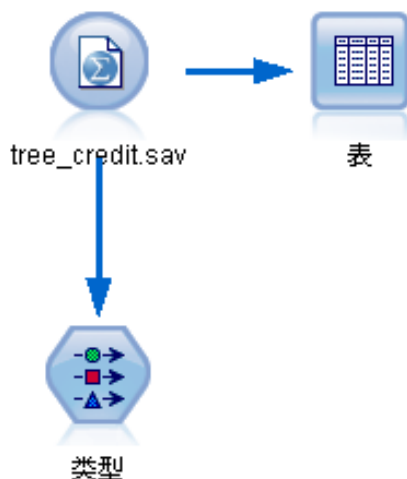


图7 工作区中的“类型”节点

使用“类型”节点选择聚类分析的字段。右键单击“类型”节点，选择“编辑”，可以看到一张含有多个字段的表，如图8所示。把所有字段的角色项设为“输入”，表示把所有字段进行聚类分析。点击“确定”按钮。



图8 “类型”节点编辑窗口

添加“K-Means”节点。使用 k-means 算法进行聚类分析。选择工作区的“类型”，在“建模”选项卡中，找到“K-Means”节点，并双击。在工作区中，得到一个“K-Means”模型，如图9所示。

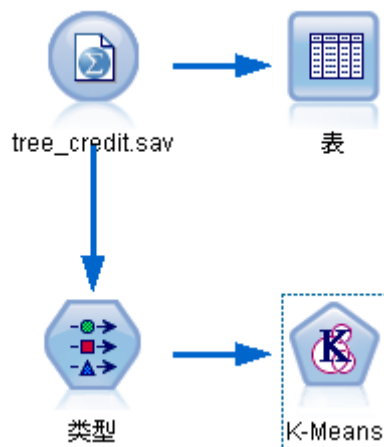


图9 “K-Means”模型

编辑 K-Means 节点。右键单击工作区的“K-Means”，选择“编辑”。在字段选项卡选择“使用类型节点设置”，表示继承上一个节点“类型设置”。在模型选项卡，可以自定义生成模型的名称，也可以采用默认的名称，“使用分区数据”指是否根据前面的分区设置进行建模，“聚类数”是指最后聚类类别的个数，“生成距离字段”是指生成每条记录与其聚类中心的距离，其他的采用默认设置，如图 10。



图10 “K-Means”模型编辑窗口

在“专家”选项卡中，选择“简单”，如图 11。“停止”是指聚类的停止条件，可直接设为默认值，也可以进行自定义。其中“最大迭代数”是指聚类最大的迭代次数，“更改容忍度”是指聚类更改的最大容忍性，“集合编码值”是指聚类过程中编码值。



图 11 专家模式

运行“K-Means”节点。右键“K-Means”节点并运行该节点，生成 K-Means 聚类模型。

查看 K-Means 聚类分析结果。在窗口右上侧区域的“模型”选项卡中，可以看到 K-Means 的模型，右键单击“编辑”，可以看到 K-Means 聚类分析图，聚类的质量属于“尚好”。把左侧的“视图”选择为“聚类”，如图 12 所示。可以发现，聚类分析把数据分成了五个类别，每个类占总数的比例分别为 28.1%、26.2%、22.2%、12.7%和 10.8%。把右侧的“视图”选为“预测变量重要性”。其中，对分类字段的依赖性从车贷、信用卡数量、信用等级、收入水平、教育水平和年龄逐渐递减。在“汇总”选项卡可以看到，本次聚类进行了 8 次迭代。

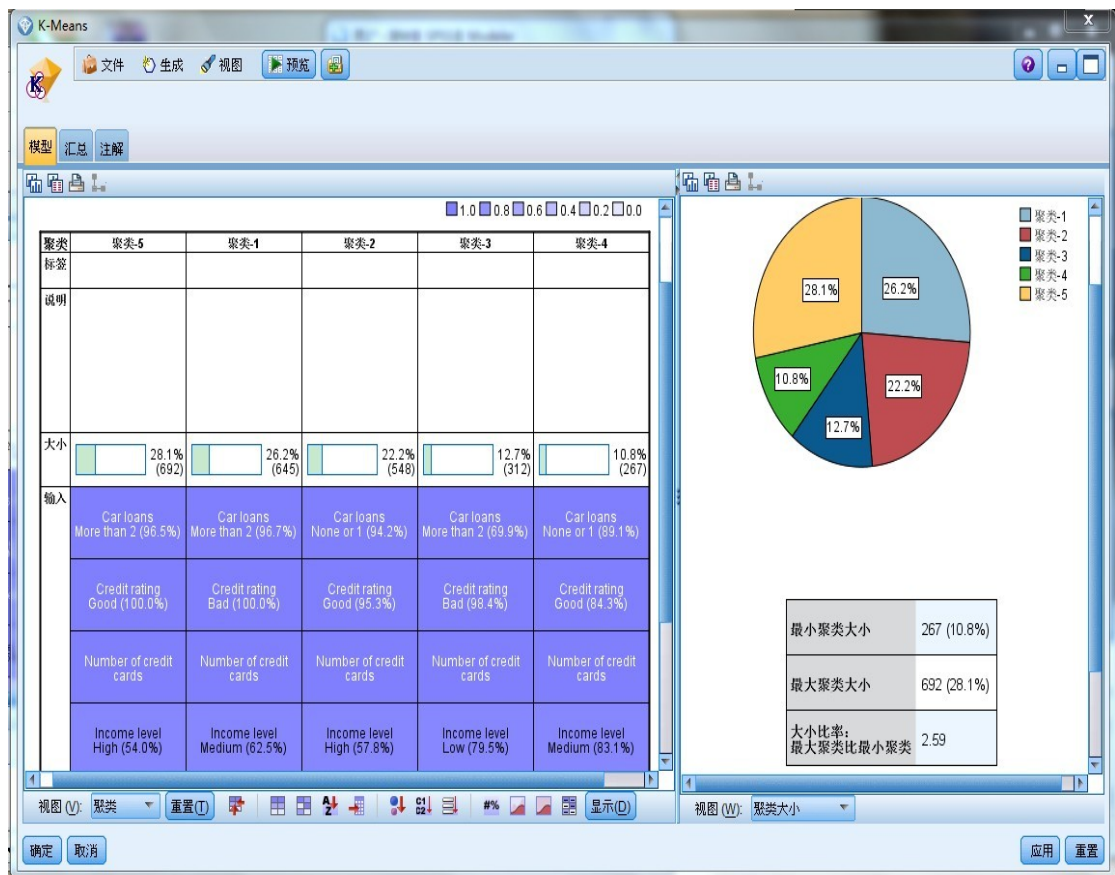


图 12 k-means 聚类结果

## 1.2 两步聚类

添加“两步”聚类节点。选中工作区的“类型”节点，在“建模”选项卡中，找到“两步”模型并双击，如图 13 所示。

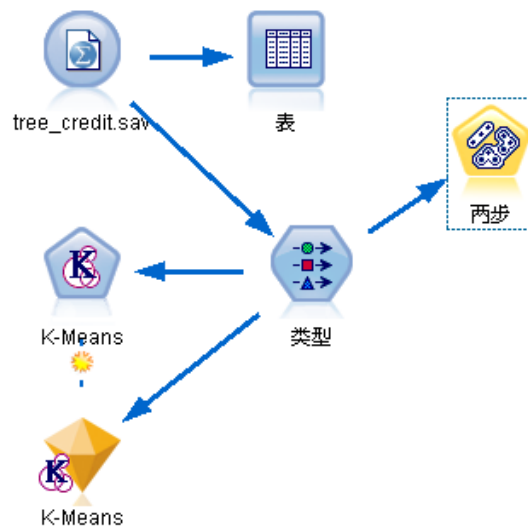


图 13 工作区中的“两步”模型

编辑“两步”节点。右键编辑“两步”节点，如图 14。在“模型”选项卡下，“标准化数值字段”指是否把字段数据标准化，“排除离群值”是指踢出距离聚类中心较远的记录（百



分比)，“自动计算聚类数”指系统自动根据需要确定聚类的个数(可设置最大值与最小值)，“指定聚类数”确定聚类的个数，“聚类距离”可选择对数似然距离或欧几里得距离，聚类准则可从 SBC 准则或 AIC 信息准则选择。

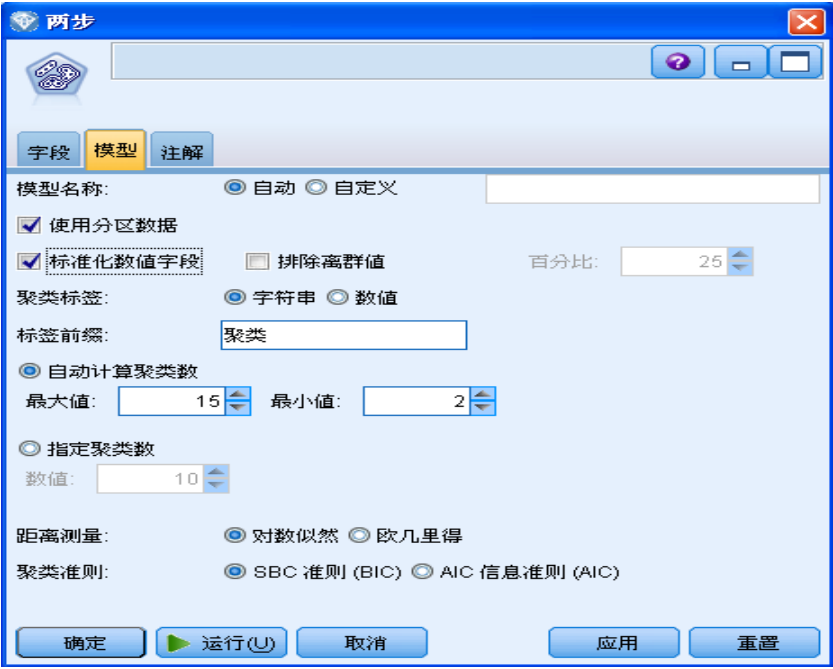


图 14 两步聚类设置

运行并查看模型。右键运行“两步”聚类节点。运行完成后，可以在窗口右上侧框中出现“两步”聚类模型，右键单击该模型，选择“浏览”，可得到图 15 的聚类分析图。可以发现，“两步”聚类分析得到三个类。分类预测字段的重要性从 Car\_loan、Credit\_cards、Credit\_rating、Income、Age 和 Education 递减。此外，还可以看出聚类质量属于“尚好”。

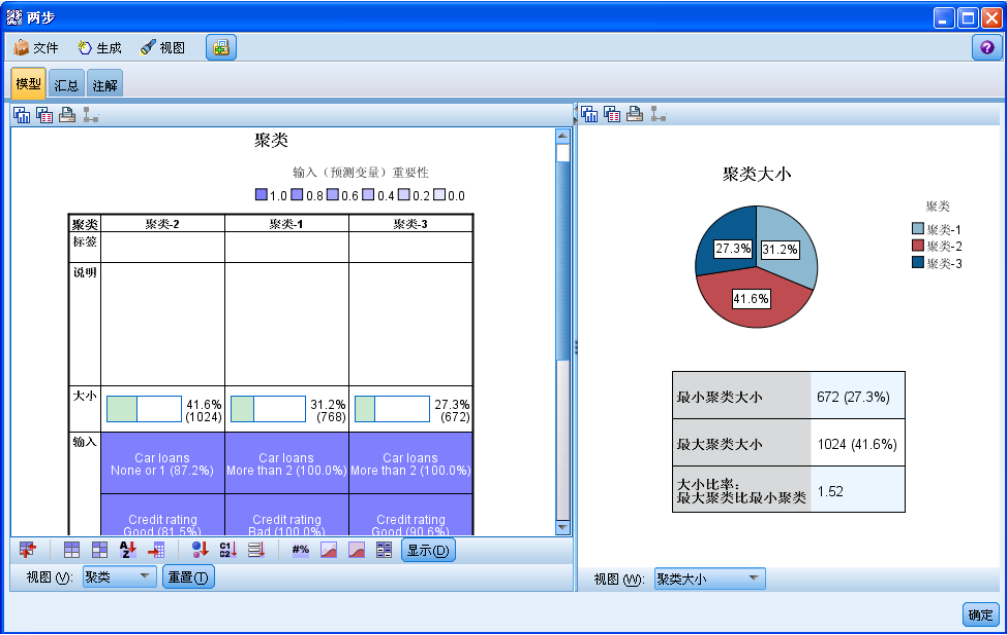


图 15 “两步”聚类分析图

### 1. 3Kohonen 聚类

添加“Kohonen”节点。选中工作区的“类型”节点，在“建模”选项卡中，双击“Kohonen”

模型。在工作区中，得到一个“Kohonen”模型节点，如图 16 所示。

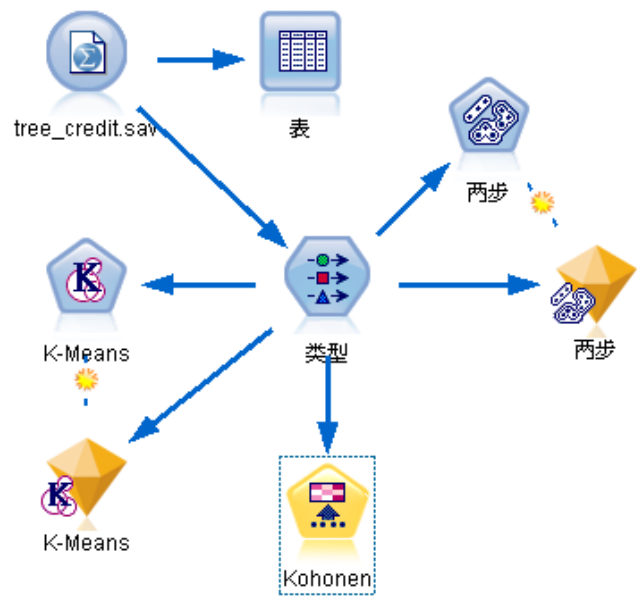


图 16 工作区中的“Kohonen”节点

编辑“Kohonen”节点。右键编辑“Kohonen”的属性，如图 17。可以设置模型基本参数：是否使用分区数据，是否继续训练现有模型，是否显示反馈图形、停止条件以及设置随机种子等。



图 17 Kohonen 节点的属性选项卡

运行并查看“Kohonen”节点。这里采用默认设置。右键运行“Kohonen”节点，在右上方会生成 Kohonen 模型。把模型重命名为 Kohonen1。右键浏览“Kohonen1”模型，如图 18。可以发现，数据分成 12 类。各字段的重要性按以下逐渐递减：Car\_loan、Education、

Credit\_cards、Credit\_rating、Income 和 Age。

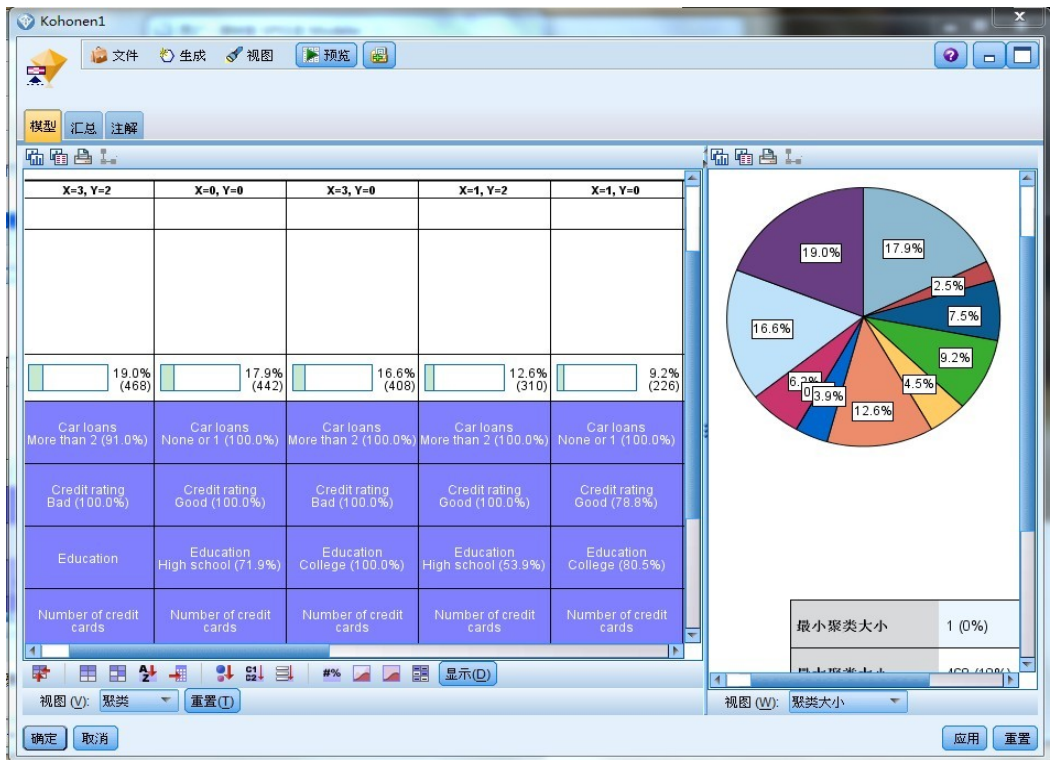


图 18 Kohonen1 聚类结果

选择“专家”模式。Kohonen 专家编辑选项卡如图 19，宽度和长度表示输出图的大小，“学习速率衰减”选择线性或指数学习速率衰减。Kohonen 网络训练分为两个阶段：阶段 1 和阶段 2，每个阶段都有以下三个参数：近邻（设置近邻的大小）、初始 Eta（为学习速率设置起始值）和周期（为训练的每个阶段设置周期数）。



图 19 Kohonen 模型的“专家”模式设置

运行并查看“Kohonen”节点（专家模式）。在这里，为了比较专家模式和简单模式两

者的差别，对 Kohonen 的专家模式进行建模。运行“Kohonen”节点，生成 Kohonen 模型。把模型命名为“Kohonen2”。右键浏览“Kohonen2”，如图 20 所示。可以得到，模型共有 33 个聚类，且聚类质量为好。预测变量重要性从 Car\_loan、Education、Credit\_cards、Income、Credit\_rating 和 Age 逐渐递减。

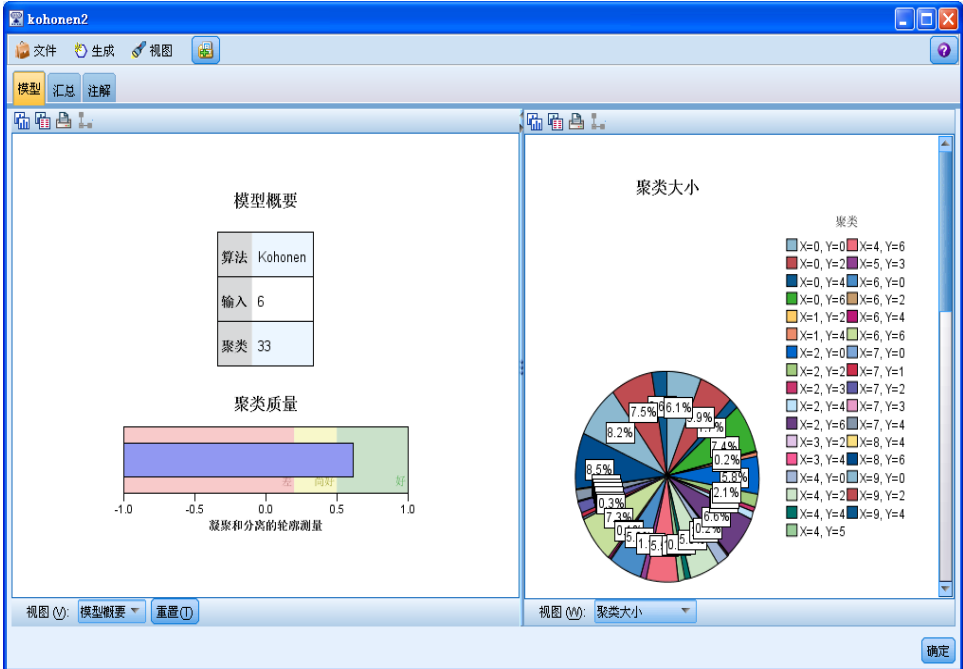


图 20 Kohonen2 模型

下面分析 K-Means 聚类的结果。查看 K-Means 模型。添加“表”节点对聚类的结果进行查看，如图 21。可以看出每个客户所属的类型，表中最后一列“\$KND-K-Means”指每一个客户距离类中心的距离，距离越小，表示效果越好。

表 (8 个字段, 2,464 条记录)

	Credit_rating	Age	Income	Credit_cards	Education	Car_loa...	\$KM-K-Means	\$KMD-K...
1	0.00	36.22	2.00	2.00	2.00	2.00	聚类-1	0.497
2	0.00	21.99	2.00	2.00	2.00	2.00	聚类-1	0.514
3	0.00	29.17	1.00	2.00	1.00	2.00	聚类-3	0.370
4	0.00	32.75	1.00	2.00	2.00	1.00	聚类-3	1.163
5	0.00	36.77	2.00	2.00	2.00	2.00	聚类-1	0.500
6	0.00	39.32	2.00	2.00	2.00	2.00	聚类-1	0.521
7	0.00	31.70	2.00	2.00	2.00	2.00	聚类-1	0.478
8	0.00	34.72	1.00	2.00	1.00	2.00	聚类-3	0.385
9	0.00	31.53	1.00	2.00	1.00	2.00	聚类-3	0.371
10	0.00	24.78	2.00	2.00	2.00	2.00	聚类-1	0.493
11	0.00	22.76	1.00	2.00	2.00	2.00	聚类-1	0.754
12	0.00	45.97	1.00	2.00	1.00	2.00	聚类-3	0.522

图 21 用于查看“K-means”的表

采用其他形式查看聚类结果。除了“表”外，还可以利用其它查看方式。这里，采用“图

形”选项卡下的“分布”节点进行展示。选择“K-Means”节点，双击“分布”，可得到“分布节点”。然后，双击该节点，可以为图选择字段，如图 22 所示。把“字段”选为“\$KM-K-Means”，颜色选为“Car\_loans”，点击运行。



图 22 “分布”节点编辑窗口

分析分布图。如图 23 所示，从分布图中可以看到每个类别样本所占的比例和记录数。



图 23 “K-Means”聚类分布图

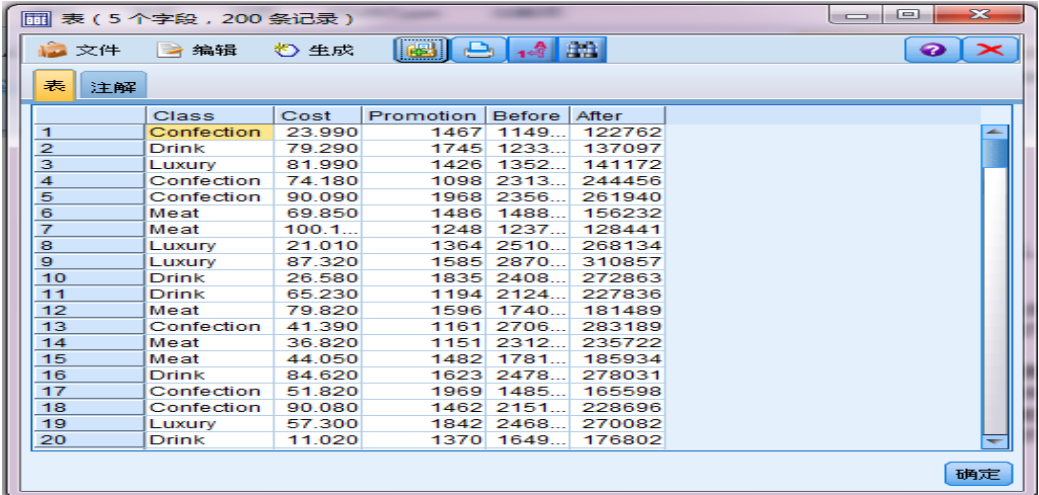
## 2 神经网络

利用神经网络模型分析哪些营销策略适合做促销，可以帮助销售部门提出合理有效的促销方案。

11.2.1 神经网络模型的生成

添加源节点，选择“可变文件”，并点击工作区的合适位置，完成添加。然后读入 Demos 文件夹中的 GOODS1n 数据文件。用 GOODS1n 文件建立神经网络模型，然后利用 GOODS2n 数据文件对生成的模型进行测试。

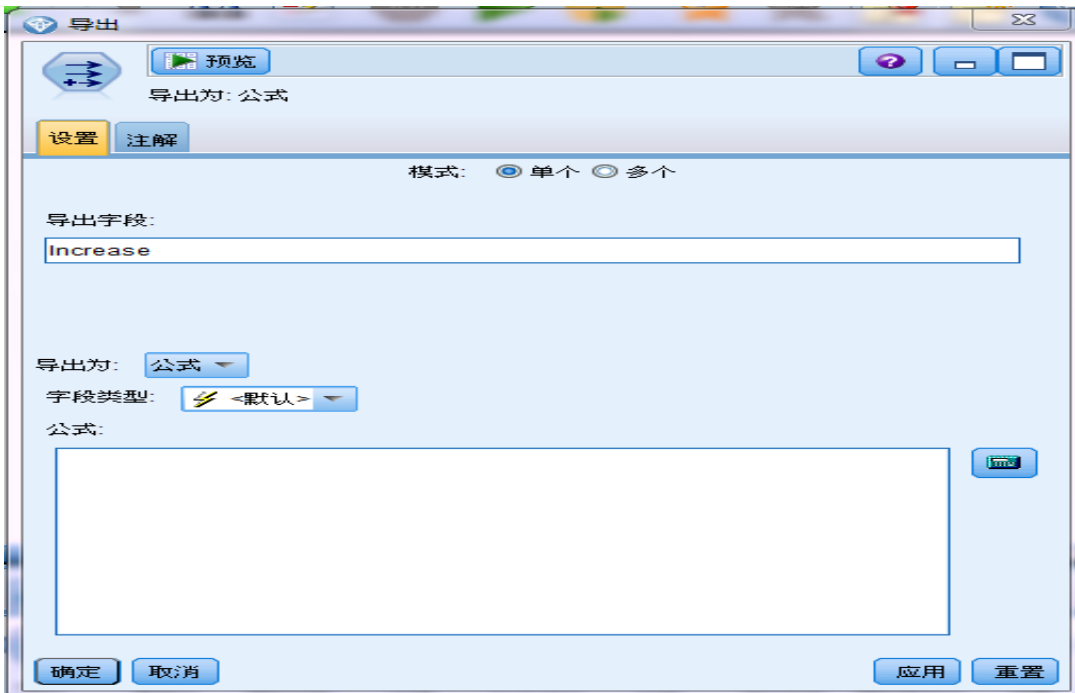
在读入 GOODS1n 数据文件后，利用“输出”选项卡下的“表”节点查看数据，得到图 26 所示的表格。该数据源是某商场在促销活动中各种商品的数据信息。可以看出 GOODS1n 共有五个字段：Class、Cost、Promotion、Before 和 After，共 200 条记录。



	Class	Cost	Promotion	Before	After
1	Confection	23.990	1467	1149...	122762
2	Drink	79.290	1745	1233...	137097
3	Luxury	81.990	1426	1352...	141172
4	Confection	74.180	1098	2313...	244456
5	Confection	90.090	1968	2356...	261940
6	Meat	69.850	1486	1488...	156232
7	Meat	100.1...	1248	1237...	128441
8	Luxury	21.010	1364	2510...	268134
9	Luxury	87.320	1585	2870...	310857
10	Drink	26.580	1835	2408...	272863
11	Drink	65.230	1194	2124...	227836
12	Meat	79.820	1596	1740...	181489
13	Confection	41.390	1161	2706...	283189
14	Meat	36.820	1151	2312...	235722
15	Meat	44.050	1482	1781...	185934
16	Drink	84.620	1623	2478...	278031
17	Confection	51.820	1969	1485...	165598
18	Confection	90.080	1462	2151...	228696
19	Luxury	57.300	1842	2468...	270082
20	Drink	11.020	1370	1649...	176802

图 26 探测数据

添加“导出”节点。读入数据后，增加一个“字段”选项卡下的“导出”节点作为促销后比促销前增加收入的比例。双击“导出”节点，可以看到一个编辑窗口，如图 27。同时，把“导出字段”取名为 Increase，设定“导出为”为“公式”。然后，点击“公式”右侧的按钮，对公式进行编辑。



导出

预览

导出为: 公式

设置 注解

模式: ☒ 单个 ☐ 多个

导出字段:

Increase

导出为: 公式

字段类型: <默认>

公式:

确定 取消 应用 重置

图 27 “导出”节点编辑

编辑公式。在编辑公式的对话框中编辑公式:  $(After - Before) / Before * 100$ ，即促销后

比促销前增加的收入百分比，如图 28 所示。在编辑完成后需要点击“检查”，以确认公式的正确性，检查无误后，确认完成编辑。



图 28 编辑公式

选择工作区的“Increase”节点，利用“表”节点查看数据，如图 29。可以看出增加了一列新的数据，即促销后比促销前增加的收入百分比（Increase）。把 Increase 作为神经网络的输出，其他四个字段作为输入。

	Class	Cost	Promotion	Before	After	Increase
1	Confection	23.990	1467	1149...	122762	6.789
2	Drink	79.290	1745	1233...	137097	11.119
3	Luxury	81.990	1426	1352...	141172	4.382
4	Confection	74.180	1098	2313...	244456	5.647
5	Confection	90.090	1968	2356...	261940	11.157
6	Meat	69.850	1486	1488...	156232	4.935
7	Meat	100.1...	1248	1237...	128441	3.782
8	Luxury	21.010	1364	2510...	268134	6.796
9	Luxury	87.320	1585	2870...	310857	8.296
10	Drink	26.580	1835	2408...	272863	13.313
11	Drink	65.230	1194	2124...	227836	7.264
12	Meat	79.820	1596	1740...	181489	4.291
13	Confection	41.390	1161	2706...	283189	4.640
14	Meat	36.820	1151	2312...	235722	1.920
15	Meat	44.050	1482	1781...	185934	4.376
16	Drink	84.620	1623	2478...	278031	12.161
17	Confection	51.820	1969	1485...	165598	11.441
18	Confection	90.080	1462	2151...	228696	6.320
19	Luxury	57.300	1842	2468...	270082	9.396
20	Drink	11.020	1370	1649...	176802	7.163

图 29 查看增加“增加率”后的数据

添加“类型”节点。在“Increase”节点之后，接入“类型”节点，如图 30 所示。



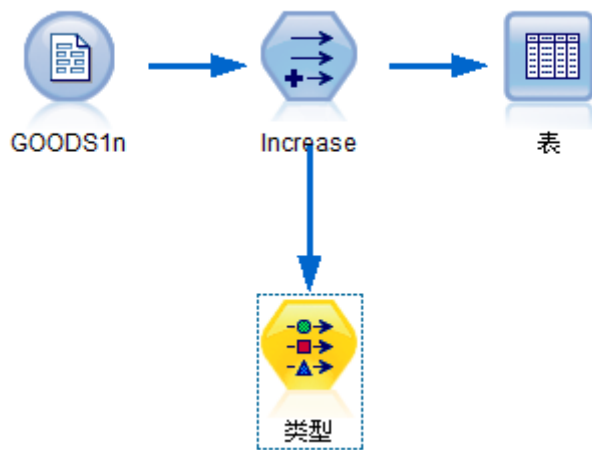


图 30 工作区中的“类型”节点

设定“类型”节点的角色。点击“读取值”，把字段 After 设为无，把 Increase 设为目标，如图 31 所示。



图 31 设定“类型”节点的角色

添加神经网络节点。选择“类型”节点，双击“建模”选项卡下的“神经网络”节点，取名为 Increase，如图 32 所示。双击“Increase”神经网络节点，对神经网络模型进行设置，选择默认值，如图 33 所示。



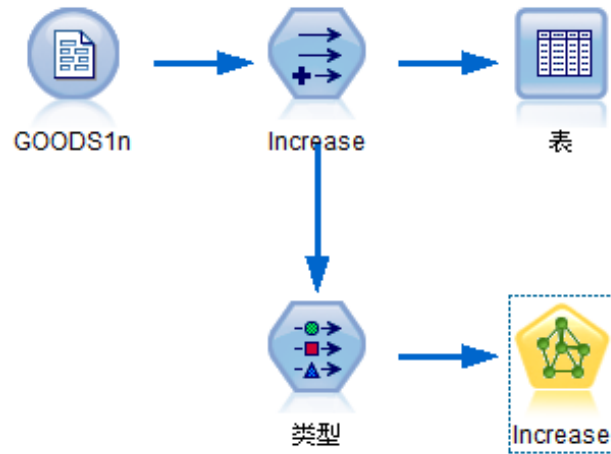


图 32 增加神经网络节点

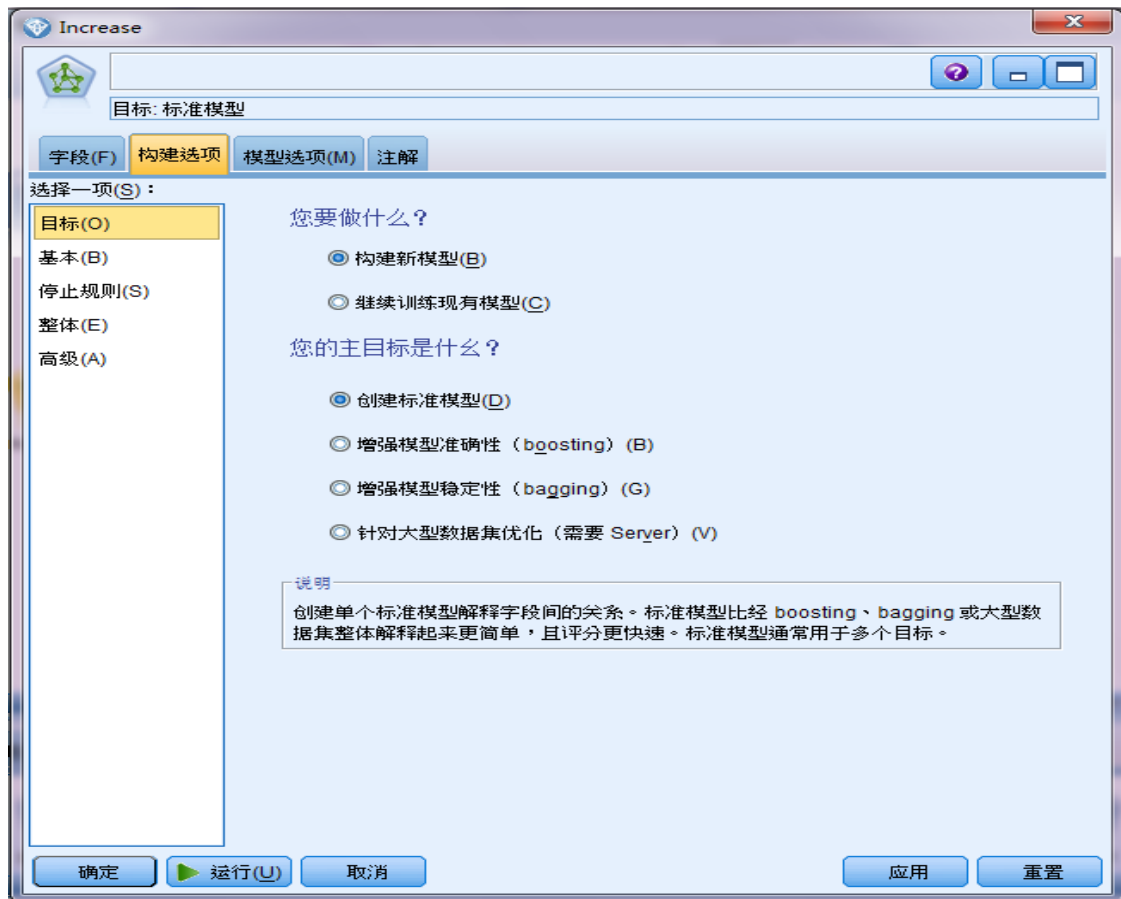


图 33 Increase 设置

生成神经网络模型。运行工作区的数据流，得到神经网络模型。

查看神经网络模型。在右上侧的“模型”选型卡中选择名为 **Increase** 的神经网络模型。右键“浏览”，如图 34 所示。模型描述内容包括“模型概要”、“预测变量重要性”、“由观察预测”和“网络”等。

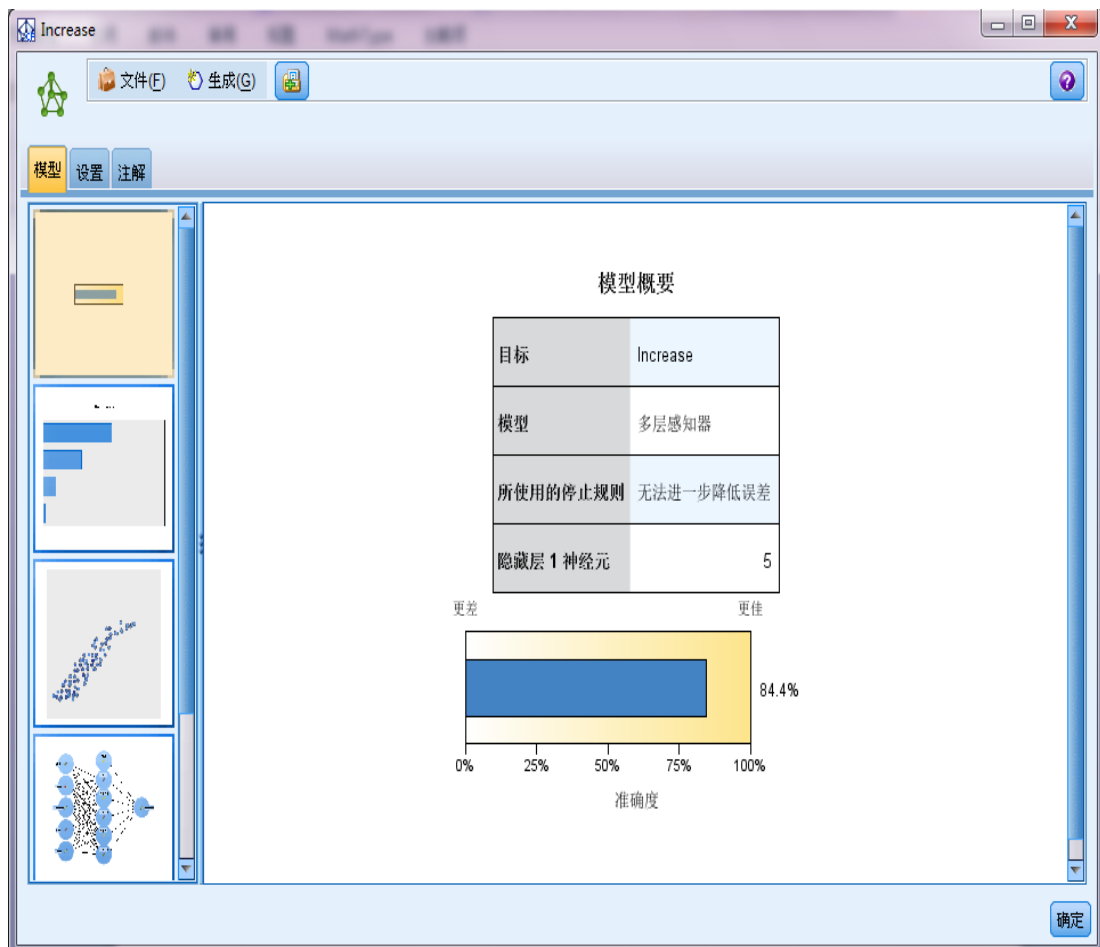


图 34 查看神经网络模型

下面利用“分析”节点对模型进行分析。添加“分析”节点。选择上一步生成的神经网络模型，双击“输出”选项卡下的“分析”节点，如图 35 所示。

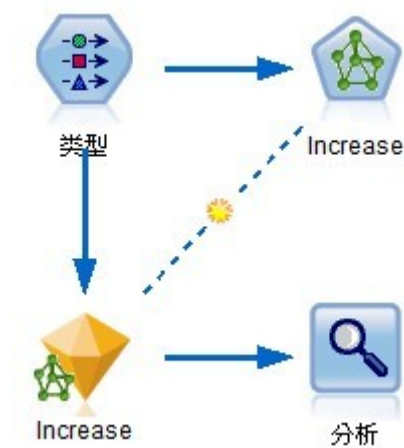


图 35 添加分析节点

运行“分析”节点。右键运行“分析”节点，如图 36。从中可以得到模型的最小误差、最大误差、平均误差和标准差等，可以据此判断模型的有效性、正确性和可靠性等。



图 36 Increase 模型的分析

神经网络训练完成，下面需要根据模型预测其他促销方案是否可行。

## 2.2 神经网络模型的评估

在以上步骤的基础上，导入新的样本数据作为测试数据。在同一个数据流中，利用“可变文件”导入 Demos 文件夹下的 GOODS2n 数据作为测试数据。

利用“输出”选项卡下的“表”节点查看“GOODS2n”中的数据，如图 37 所示。为了对 GOODS2n 中的数据进行预测，需要把 After 字段过滤掉。

	Class	Cost	Promotion	Before	After
1	Luxury	31.2...	1467	223360	238333
2	Drink	82.5...	1316	198980	219791
3	Luxury	10.4...	1734	248095	266357
4	Drink	40.4...	1002	215989	235013
5	Drink	20.2...	1127	289011	305659
6	Meat	59.3...	1884	234745	241302
7	Meat	71.1...	1655	208719	216708
8	Drink	62.7...	1108	192209	204458
9	Drink	98.2...	1075	234269	248692
10	Drink	34.6...	1644	110999	121988
11	Luxury	87.4...	1105	136104	140323
12	Drink	92.7...	1828	209199	239858
13	Luxury	66.4...	1137	121856	126166
14	Meat	5.810	1446	206271	214172
15	Meat	92.9...	1260	157496	159442
16	Luxury	34.7...	1644	237554	248668
17	Meat	69.9...	1398	228312	238146
18	Conf...	80.3...	1007	189862	198010
19	Luxury	20.4...	1389	252244	267280
20	Meat	17.4...	1084	270672	271425

图 37 查看数据

过滤 GOODS2n 中的数据。双击 GOODS2n 的源节点，选择其中的“过滤”选项卡，点击 After 行过滤“箭头”，点击“确定”完成过滤，如图 38 所示。此外，读者也可以直接利用“字段”选项卡下的“过滤”节点进行数据的过滤。



图 38 过滤 GOODS2n 的数据

接入神经网络。选择工作区的 GOODS2n 节点，双击右上侧生成的 Increase 神经网络模型，完成接入，如图 39。

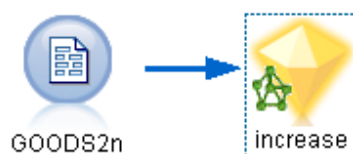


图 39 构建测试流

查看神经网络模型预测结果。选择预测的“Increase”节点，添加“表”节点查看神经网络的预测，如图 40 所示。其中\$N-Increase 字段为神经网络模型预测输出的促销前后收入百分比。

	Class	Cost	Promotion	Before	\$N-Increase
1	Luxury	31.2...	1467	223360	5.348
2	Drink	82.5...	1316	198980	9.553
3	Luxury	10.4...	1734	248095	6.331
4	Drink	40.4...	1002	215989	6.972
5	Drink	20.2...	1127	289011	7.214
6	Meat	59.3...	1884	234745	3.632
7	Meat	71.1...	1655	208719	3.264
8	Drink	62.7...	1108	192209	7.889
9	Drink	98.2...	1075	234269	8.205
10	Drink	34.6...	1644	110999	10.745
11	Luxury	87.4...	1105	136104	5.034
12	Drink	92.7...	1828	209199	12.861
13	Luxury	66.4...	1137	121856	4.994
14	Meat	5.810	1446	206271	2.251
15	Meat	92.9...	1260	157496	3.166
16	Luxury	34.7...	1644	237554	6.155
17	Meat	69.9...	1398	228312	2.924
18	Conf...	80.3...	1007	189862	6.436
19	Luxury	20.4...	1389	252244	4.907
20	Meat	17.4...	1084	270672	2.070

图 40 查看由神经网络模型预测输出的数据

把预测的数据规范到 0-1 之间。这里，利用“字段”选项卡下的“导出”进行规范化。先删除由上一步骤新添的表，选择“Increase”模型，双击“导出”。然后，对“导出”进行编辑，导出字段取名为 Format，“导出为”下拉列表选择“公式”，编写公式  $1/(1+\log(1+1/(\$N-Increase)))$ ，如图 41 所示。



图 41 把预测输出的数据规范到 0-1 之间

查看规范化后的预测值。添加“表”节点查看，如图 42 所示。预测值已规范到 0-1 之间。

表 ( 6 个字段, 400 条记录 )

	Class	Cost	Promotion	Before	\$N-Increase	Format
1	Luxury	31.2...	1467	2233...	5.348	0.854
2	Drink	82.5...	1316	1989...	9.553	0.909
3	Luxury	10.4...	1734	2480...	6.331	0.872
4	Drink	40.4...	1002	2159...	6.972	0.882
5	Drink	20.2...	1127	2890...	7.214	0.885
6	Meat	59.3...	1884	2347...	3.632	0.804
7	Meat	71.1...	1655	2087...	3.264	0.789
8	Drink	62.7...	1108	1922...	7.889	0.893
9	Drink	98.2...	1075	2342...	8.205	0.897
10	Drink	34.6...	1644	1109...	10.745	0.918
11	Luxury	87.4...	1105	1361...	5.034	0.847
12	Drink	92.7...	1828	2091...	12.861	0.930
13	Luxury	66.4...	1137	1218...	4.994	0.846
14	Meat	5.810	1446	2062...	2.251	0.731
15	Meat	92.9...	1260	1574...	3.166	0.785
16	Luxury	34.7...	1644	2375...	6.155	0.869
17	Meat	69.9...	1398	2283...	2.924	0.773
18	Conf...	80.3...	1007	1898...	6.436	0.874
19	Luxury	20.4...	1389	2522...	4.907	0.844
20	Meat	17.4...	1084	2706...	2.070	0.717

图 42 查看规范化后的预测值



利用“记录”选项卡下的“选择”项对预测值进行选择。把上一步中的“表”节点删除，选择“导出”节点，然后双击“记录”选项卡下的“选择”节点，双击该节点进行公式的编辑。如图 43 所示，其中表达式为“Format >= 0.9”。



图 43 “选择”节点编辑窗口

查看选择后的预测数据。添加“表”节点查看，如图 44 所示。可以看出，满足条件 (Format 值都大于 0.9) 的记录共有 126 条。这些都是比较合适的促销方案样本。

	Class	Cost	Promotion	Before	\$N-Increase	Format
1	Drink	82.560	1316	1989...	9.553	0.909
2	Drink	34.620	1644	1109...	10.745	0.918
3	Drink	92.760	1828	2091...	12.861	0.930
4	Confection	51.980	1804	1580...	9.875	0.912
5	Drink	98.150	1706	2234...	12.335	0.928
6	Confection	77.470	1914	2481...	11.001	0.920
7	Drink	47.660	1318	2079...	8.874	0.904
8	Drink	44.540	1938	1718...	12.689	0.929
9	Drink	103.3...	1904	2447...	13.225	0.932
10	Drink	76.190	1888	2579...	12.759	0.930
11	Confection	95.500	1910	1627...	11.406	0.922
12	Drink	23.110	1640	1579...	10.598	0.917

图 44 查看选择后的预测数据

IBM SPSS Modeler 14.2 支持多种格式的数据导出，包括 Excel 表、数据库、Statistics 文件等。导出的方法也比较简单，可以用一个专门的“导出”选项卡。

下面通过神经网络模型预测 After，并把该预测值与真实值进行比较，分析预测的正确性。

创建数据源。在工作区中，新建一个“可变文件”，并把 Demos 文件夹下的 GOODS2n 的数据导入到源节点中，并把其重命名为 DiffGOODS2n。添加神经网络模型。选择 DiffGOODS2n 源数据节点，双击上面生成的神经网络模型，加入神经网络模型到数据流中。添加促销后商品收入的预测值节点。选择上一步的神经网络模型节点，双击位于“字段”选项卡中的“导出”，完成添加，如图 48 所示。



图 48 添加“导出”的数据流图

编辑促销后商品收入的预测值节点。右键编辑上一步的导出节点，并把导出字段命名为 PredictedAfter，导出公式为： $\text{Before} * (1 + \text{'\$N-Increase'}/100)$ ，如图 49 所示。



图 49 PredictedAfter 的导出公式

添加预测值与真实值的“误差（百分比）”节点。选择上一步的“神经网络模型”节点，双击位于“字段”选项卡中的“导出”节点，完成添加。

编辑预测值与真实值的误差（百分比）节点。右键编辑上一步的“导出”节点，并把导出字段命名为 DiffPercentage，导出公式为： $(\text{PredictAfter} - \text{After}) / \text{After} * 100$ 。

查看预测值和误差。在上一步的“DiffPercentage”节点后添加一个“表”节点，并右键运行，如图 50 所示。可以看出，预测值与真实值之间的差距并不大，误差都在 3% 以内。神经网络模型对这些数据的预测比较好。

表 ( 8 个字段, 400 条记录 )

	Class	Cost	Promotion	Before	After	\$N-Increase	PredictedAfter	DiffPercentage
1	Luxury	31.2...	1467	2233...	238333	5.348	235305.339	-1.270
2	Drink	82.5...	1316	1989...	219791	9.553	217987.989	-0.820
3	Luxury	10.4...	1734	2480...	266357	6.331	263801.558	-0.959
4	Drink	40.4...	1002	2159...	235013	6.972	231047.353	-1.687
5	Drink	20.2...	1127	2890...	305659	7.214	309859.824	1.374
6	Meat	59.3...	1884	2347...	241302	3.632	243270.698	0.816
7	Meat	71.1...	1655	2087...	216708	3.264	215531.597	-0.543
8	Drink	62.7...	1108	1922...	204458	7.889	207372.809	1.426
9	Drink	98.2...	1075	2342...	248692	8.205	253491.777	1.930
10	Drink	34.6...	1644	1109...	121988	10.745	122926.024	0.769
11	Luxury	87.4...	1105	1361...	140323	5.034	142955.112	1.876
12	Drink	92.7...	1828	2091...	239858	12.861	236103.654	-1.565
13	Luxury	66.4...	1137	1218...	126166	4.994	127940.969	1.407
14	Meat	5.810	1446	2062...	214172	2.251	210915.165	-1.521
15	Meat	92.9...	1260	1574...	159442	3.166	162481.842	1.907
16	Luxury	34.7...	1644	2375...	248668	6.155	252175.941	1.411
17	Meat	69.9...	1398	2283...	238146	2.924	234988.744	-1.326
18	Conf...	80.3...	1007	1898...	198010	6.436	202080.608	2.056
19	Luxury	20.4...	1389	2522...	267280	4.907	264621.044	-0.995
20	Meat	17.4...	1084	2706...	271425	2.070	276275.836	1.787

确定

图 50 预测的误差

### 3 关联分析和分类

首先进行关联分析，然后建立一个决策树。

#### 3.1 关联分析

打开并查看数据文件。利用“可变文件”节点把 Demos 下的数据 BASKETS1n 读入节点中。然后使用“输出”选项卡下的“表”节点查看数据，如图 51 所示。这里的数据是某商场中的交易记录，共 18 个字段，1000 条记录，其中“T”表示客户已购买该商品，“F”表示没有购买该商品。

表 ( 18 个字段, 1,000 条记录 )

	cardid	value	pmethod	sex	homeown	income	age	fruitveg	freshmeat	dairy	can...	cannedmeat	froz
1	39808	42.712	CHEQUE	M	NO	27000	46 F	T	T	F	F	F	F
2	67362	25.357	CASH	F	NO	30000	28 F	T	F	F	F	F	F
3	10872	20.618	CASH	M	NO	13200	36 F	F	F	T	F	F	T
4	26748	23.688	CARD	F	NO	12200	26 F	F	T	F	F	F	F
5	91609	18.813	CARD	M	YES	11000	24 F	F	F	F	F	F	F
6	26630	46.487	CARD	F	NO	15000	35 F	T	F	F	F	F	F
7	62995	14.047	CASH	F	YES	20800	30 T	F	F	F	F	F	F
8	38765	22.203	CASH	M	YES	24400	22 F	F	F	F	F	F	F
9	28935	22.975	CHEQUE	F	NO	29500	46 T	F	F	F	F	F	T
10	41792	14.569	CASH	M	NO	29600	22 T	F	F	F	F	F	F
11	59480	10.328	CASH	F	NO	27100	18 T	T	T	T	F	F	F
12	60755	13.780	CASH	F	YES	20000	48 T	F	F	F	F	F	F
13	70998	36.509	CARD	M	YES	27300	43 F	F	T	F	T	F	T
14	80617	10.201	CHEQUE	F	YES	28000	43 F	F	F	F	F	F	F
15	61144	10.374	CASH	F	NO	27400	24 T	F	T	F	F	F	F
16	36405	34.822	CHEQUE	F	YES	18400	19 F	F	F	F	F	F	T
17	76567	42.248	CARD	M	YES	23100	31 T	F	F	T	F	F	F
18	85699	18.169	CASH	F	YES	27000	29 F	F	F	F	F	F	F
19	11357	10.753	CASH	F	YES	23100	26 F	F	F	F	F	F	F

确定

图 51 探查交易数据



确定关联分析字段。除了分析客户购买商品之间的关联行为外（读者可以做一下，这里不再累述），还可以确定客户的某些属性是否与商品的购买是否存在一定的关联。为了简单起见，这里考虑购买购买者的性别 sex 与商品种类 fruitveg 是否存在关联。

读入分析字段的类型。在工作区生成“类型”节点，并点击“读取值”，把上一步选择的 sex 和 fruitveg 字段的角色设定为两者，如图 52。



图 52 类型节点编辑窗口

添加关联模型节点。在“类型”之后添加“Apriori”节点和“Carma”节点，分别命名为 Apriori 和 Carma，如图 53 所示。

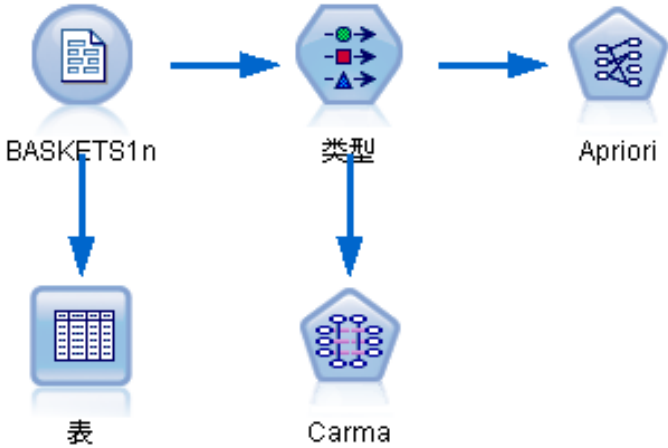


图 53 添加 Apriori 和 Carma 节点

修改“Apriori”模型设置。右键编辑“Apriori”节点，在“模型”选项卡下，把最低条件支持度和最低规则置信度都设为 10.0，最大前项数为 5，如图 54 所示。



图 54 Apriori 节点属性设置

运行并查看“Apriori”关联分析结果。在右上侧的模型管理区生成数据模型，右键查看，如图 55 所示。可以看出，客户的性别 sex 和购买 fruitveg 之间有一定的关联。这意味着如果一个客户已经购买了 fruitveg，该客户为男性的支持度为 29.9%，置信度为 45.819%。如果该客户是男性，他购买 fruitveg 的支持度为 48.8%，支持度为 28.074%。



图 55 Apriori 关联分析结果

“Carma”模型设置。右键编辑“Carma”节点，在“模型”选项卡下，把最低条件支持度和最低规则置信度都设为 10.0，最大前项数为 5。

运行并查看“Carma”。右键运行“Carma”模型的数据流，生成关联模型，如图 56 所示。可以看出，如果一个客户已经购买了 fruitveg，该客户为男性的支持度为 46.0%，置信度为 45.819%。如果该客户是男性，他购买 fruitveg 的支持度为 75.077%，置信度为 28.074%。



Carma 关联分析结果窗口显示了关联分析的结果。窗口顶部有“文件”、“生成”、“预览”等按钮。下方有“模型”、“设置”、“汇总”、“注解”等选项卡。当前显示的是“按以下内容进行排序：置信度 %”，排序后的结果如下表所示：

后项	前项	支持度 %	置信度 %
homeown	wine	29.019	45.993
sex	fruitveg	30.233	45.819
homeown	frozenmeal	30.536	45.033
homeown	cannedveg	30.637	43.234
frozenmeal	sex	49.343	42.828
cannedveg	sex	49.343	40.779
beer	sex	49.343	40.164
sex	wine	29.019	37.282
homeown	fruitveg	30.233	35.117
cannedveg	sex	49.343	31.148
frozenmeal	sex	49.343	30.738
cannedveg	sex	49.343	30.328
beer	sex	49.343	28.893
beer	homeown	49.949	28.138
fruitveg	sex	49.343	28.074
sex	sex	49.343	28.074

窗口底部有“确定”、“取消”、“应用”、“重置”按钮。

图 56 Carma 关联分析结果

除了利用关联分析对客户的性别 sex 和购买商品 fruitveg 之间的关联进行分析，还可以对客户的其他属性进行分析。例如客户的 homeown 属性与购买商品之间的关联，客户选择支付方式与客户的性别之间的关系等。

### 3.2 建立决策树

下面再根据用户的 age、sex、income、homeown 以及 value，利用 C5.0 决策树对客户的支付方式进行分析。

(1) 导入 BASKETS1n 数据。首先添加一个“可变文件”的节点，把 Demos 下的 BASKETS1n 数据导入。

(2) 添加“类型”节点。编辑字段的角色，如图 57 所示。其中 sex、homeown、income、age 和 value 等字段都设置为输入，而 pmethod 改为目标，其他的字段选择无。



类型窗口显示了字段的角色设置。窗口顶部有“预览”按钮。下方有“类型”、“格式”、“注解”等选项卡。当前显示的是“类型”选项卡，下方有“读取值”、“清除值”、“清除所有值”按钮。下方表格显示了字段的角色设置：

字段	测量	值	缺失	检查	角色
value	连续	[10.007,4...		无	输入
sex	标志	M/F		无	输入
homeown	标志	YES/NO		无	输入
income	连续	[10200,30...		无	输入
age	连续	[16,50]		无	输入
pmethod	名义	CARD,CA...		无	目标
cardid	连续	[10150,10...		无	无

窗口底部有“查看当前字段”、“查看未使用的字段设置”、“确定”、“取消”、“应用”、“重置”按钮。

图 57 字段角色设置

(3) 添加 C5.0 决策树模型。双击“建模”选项卡下的“C5.0”模型“pmethod”，如图 58 所示。

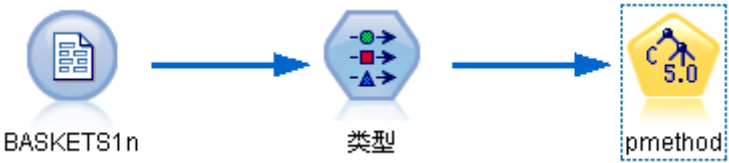


图 58 接入 C5.0 模型

(4) 选择专家模式。把“修剪严重性”设为 75，如图 59 所示。“修剪严重性”指修剪的程度。

(5) 其他属性的讨论。如图 59，组符号指当有多个字段在同一个分支时，把这几个字段放在一组。使用 **boosting** 指使用部分数据再次生成决策树，最后综合这些决策树提高决策树的精度。交互验证是指一部分数据用来生成决策树，一部分作为测试。折叠次数指把样本集分成的组数。

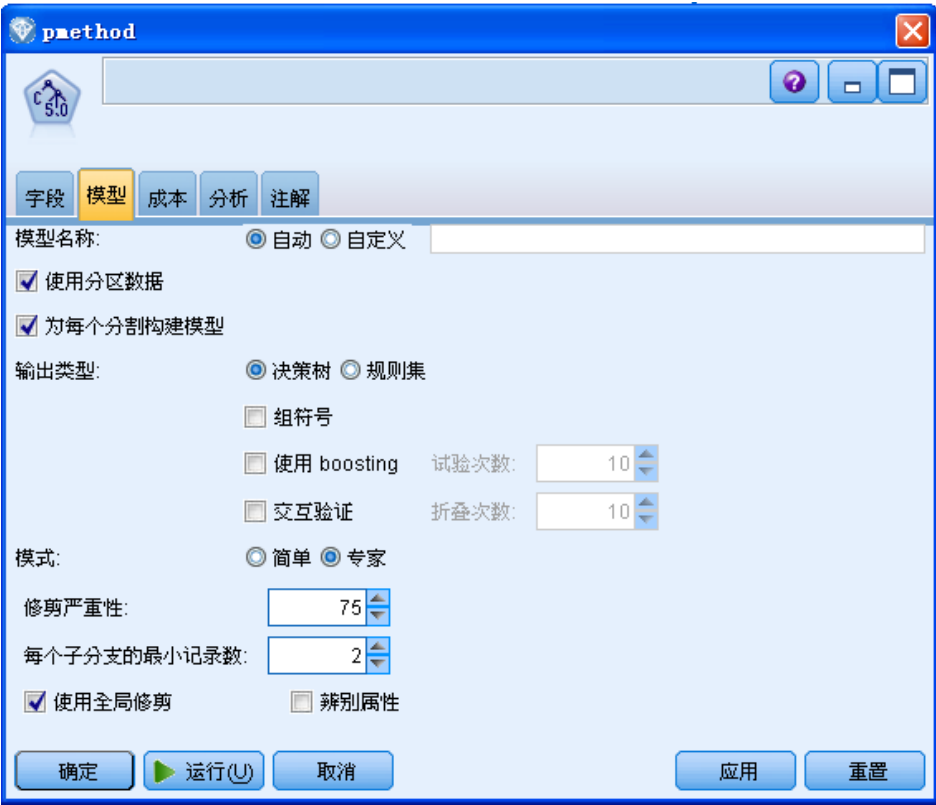


图 59 其他属性设置

运行决策树模型。把生成的决策树模型重命名为 C5.0Tree，如图 60 所示。

双击查看 C5.0Rule 模型。点击“生成”->“规则集”，重命名为 C5.0Rule，如图 61 所示。由 C5.0Tree 生成的结果可以看出：对支付方式影响的重要性从 income、sex、age、homeown 和 value 逐渐递减。由决策树也可以知道哪些特征的客户会选择信用卡支付，哪些客户会选择现金支付。由 C5.0Rule 生成的结果可以看出，对支付方式影响的重要性从 income、sex、homeown、age 和 value 逐渐递减。类似地，该模型可以根据客户的属性和金额对客户支付方式进行预测。

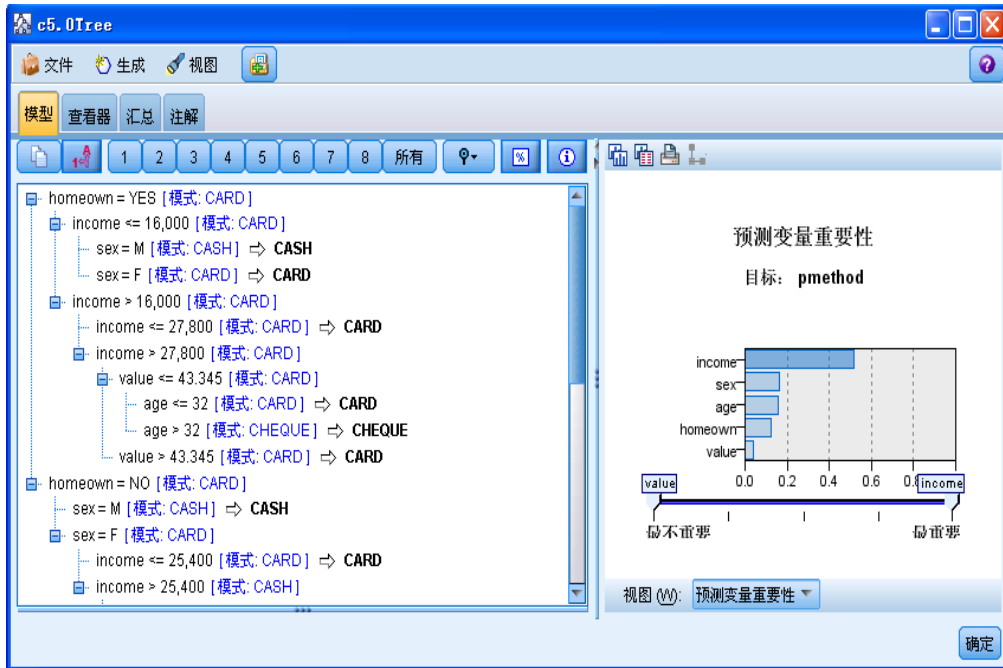


图 60 C5.0Tree 生成的决策树

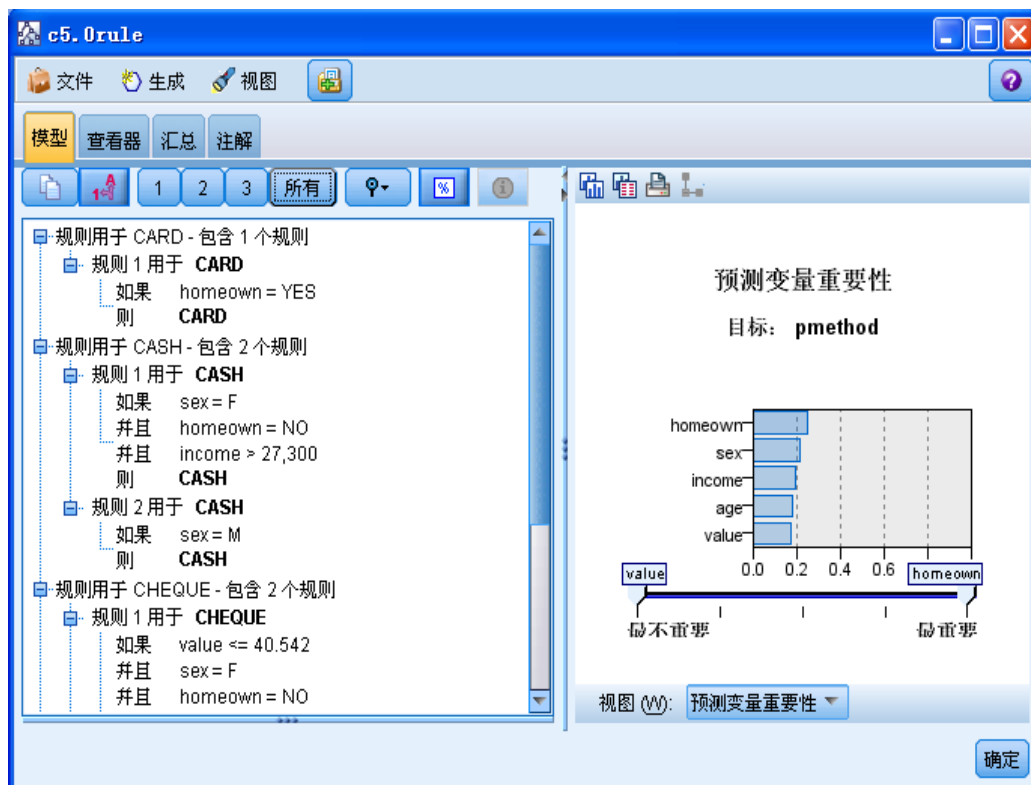


图 61 C5.0Rule 规则集

把生成的决策树模型加入数据流。选择“字段”选项卡下的“类型”节点，分别双击右上侧的 c5.0Tree 和 c5.0Rule。再分别添加“分析”节点，得到如图 62 所示的数据流。

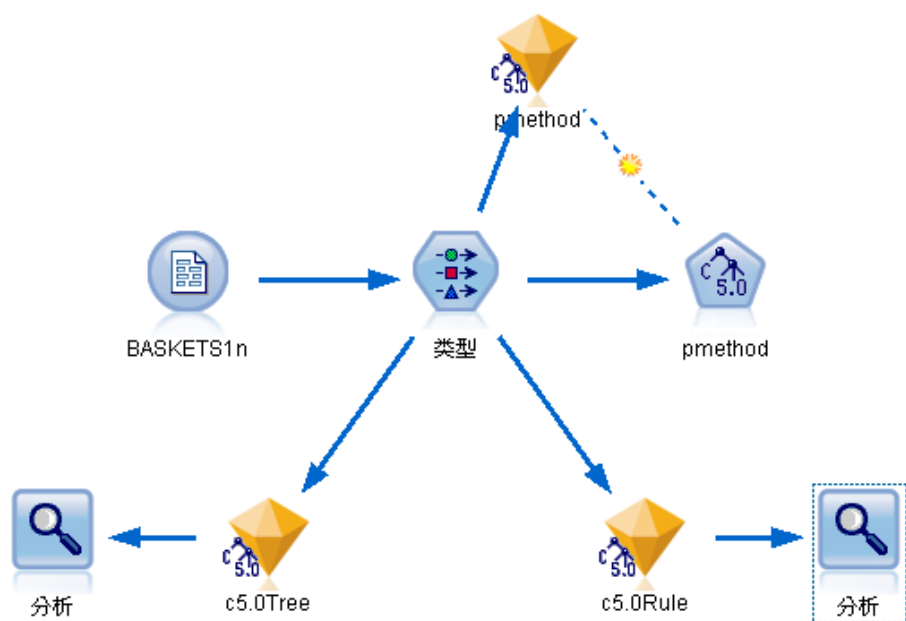


图 62 添加“分析”节点

分析 c5.0Tree 生成决策树的正确性。运行 c5.0Tree 下的“分析”节点，在图 63 中可以看出，c5.0Tree 生成决策树的正确率为 47.8%。



图 63 C5.0Tree 生成决策模型的正确性

分析 c5.0Rule 生成决策树的正确性。运行 c5.0Rule 下“分析”节点后，如图 64 所示，可以看出决策树分类的正确率为 46.2%，两者的正确率均不是很高，这可能是由于客户属性与支付方式本身的关联不是很强，而且目标选项有三个：CASH、CARD 和 CHEQUE。



图 64 C5.0Rule 的正确率

决策树检测的“成本”如图 65 所示。这里，把 CHEQUE 判为错的成本改为 2.0，修剪严重性设定为 75%，运行该决策树模型，并把生成的模型重命名为 c5.0TreePay，如图 66 所示。然后与原来生成的 c5.0Tree 模型（采用成本的默认值是 1，“修剪严重性”为 75%）进行比较。可以看出，把 CHEQUE 判为错的成本改为 2.0 后，预测的重要性也从原来的 income、sex、age、homeown 和 value 逐渐递减变成了从 income、age、homeown、sex 和 value 逐渐递减，决策树层次也发生了变化。但决策树模型改变的方向有规律的，即朝着成本最低的方向移动。



图 65 修改决策树模型的成本选项



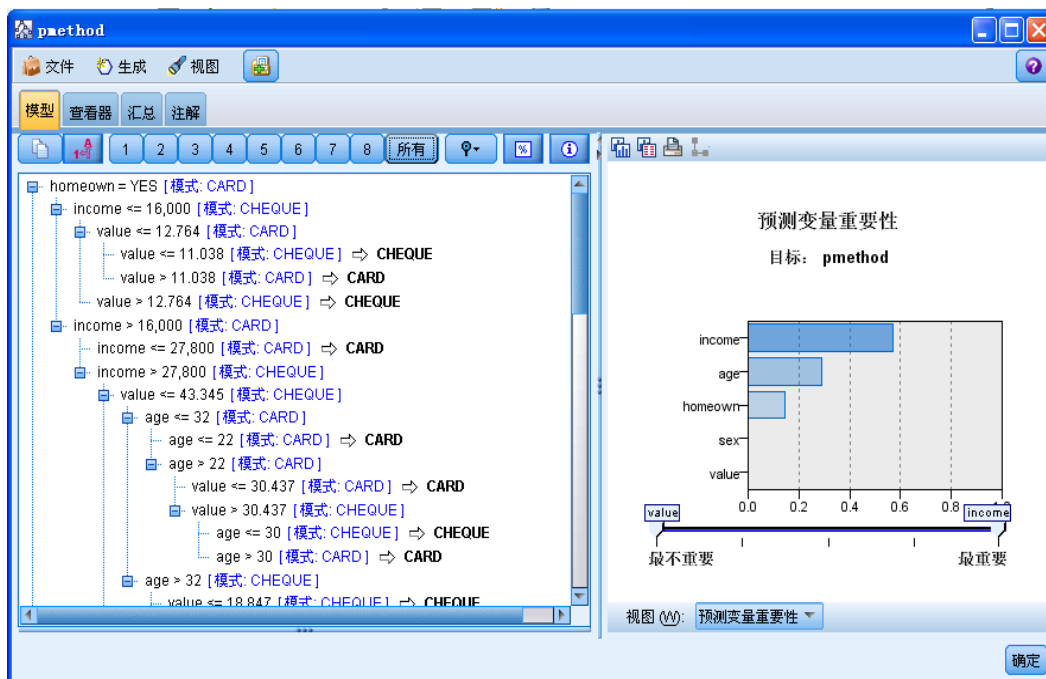


图 66 pmethod 运行结果

添加“输出”选项卡下的“矩阵”节点。把 c5.0TreePay 决策树模型添加到数据流中，然后在 c5.0Tree 和 c5.0TreePay 决策树模型后添加矩阵节点，如图 67 所示。

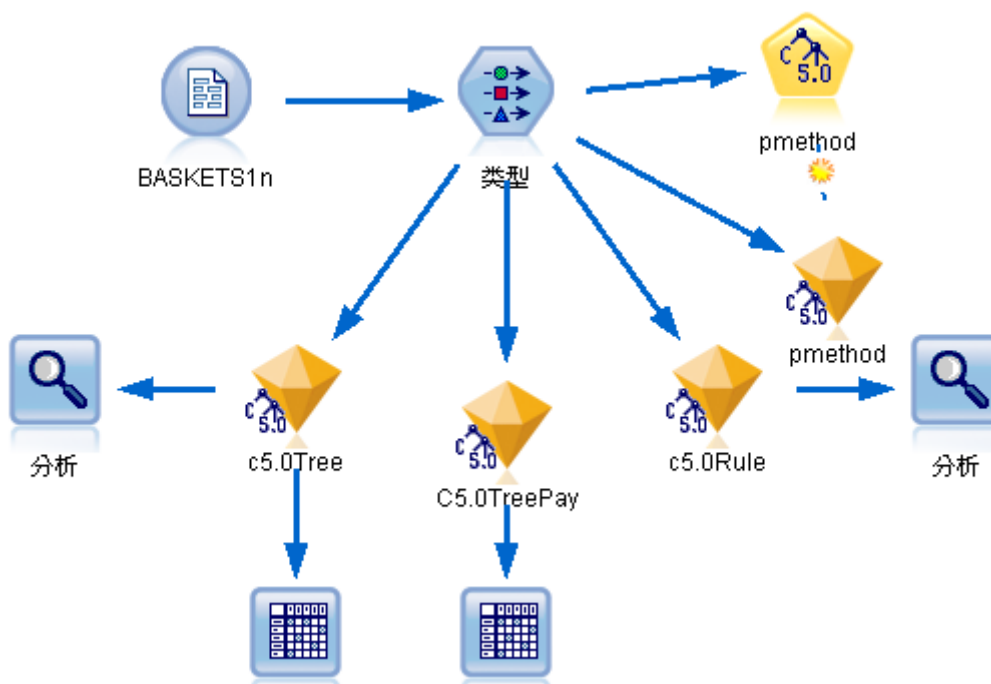


图 67 工作区的矩阵节点

运行“矩阵”节点。在运行之前，需要把这两个“矩阵”节点的行设为 pmethod，列设置为 \$C-pmethod。运行后得到图 68，可以比较 pmethod 的真实值与预测值 \$C-pmethod 之间的关系。



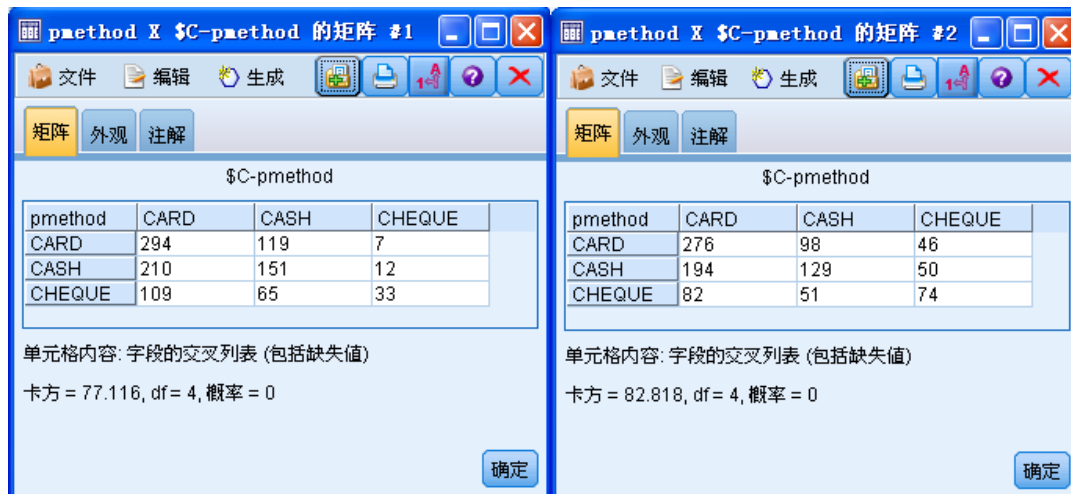


图 68 矩阵节点运行后结果