IBM SPSS Modeler 14.2 建模节点

注意：使用本信息以及其支持的产品之前，请阅读 注意事项第 页码 下的常规信息。

前言

IBM® SPSS® Modeler 是 IBM Corp. 企业级数据挖掘工作平台。SPSS Modeler通过深入的数据分析帮助组织改进与客户和市民的关系。组织通过借助源自 SPSS Modeler 的洞察力可以留住优质客户，识别交叉销售机遇，吸引新客户，检测欺诈，降低风险，促进政府服务交付。

SPSS Modeler’的可视化界面让用户可以应用他们自己的业务专长，这将生成更强有力的预测模型，缩减实现解决方案所需的时间。 SPSS Modeler 提供了多种建模技术，例如预测、分类、细分和关联检测算法。模型创建成功后，通过 IBM® SPSS® Modeler Solution Publisher，在广泛的企业内交付给决策者，或通过数据库交付。

关于 IBM Business Analytics

IBM Business Analytics 软件为决策者提供可信赖的完整、一致和准确信息，以帮助其提升业务绩效。这一涵盖商务智能、预测分析、财务绩效与战略管理以及分析应用程序的全面组合可提供有关当前业务表现的清晰、立即和切实可行的深入见解，并能够有效预测未来结果。其中整合了丰富的行业解决方案、经过验证的做法与专业服务，以帮助各种规模的组织提升生产效率、自动化决策并取得卓越成果。

作为该软件组合的一部分，IBM SPSS Predictive Analytics 软件能够帮助各类组织有效地预测未来事件，并针对所得到的深入见解提前采取行动，以取得更优秀的业务成果。全球企业、政府和学院客户依赖 IBM SPSS 技术作为吸引、留住和增加客户数量的竞争优势，并降低欺诈和转移风险。通过将 IBM SPSS 软件融入其日常运营中，这些组织将成为“预测型”企业，即能够指引并自动化决策，以实现业务目标和取得可衡量的竞争优势。有关详细信息，或联系我们的代表，请访问 http://www.ibm.com/spss。

技术支持

我们提供有技术支持服务以维护客户。客户可就 IBM Corp. 产品使用或某一受支持硬件环境的安装帮助寻求技术支持。要获得技术支持，请访问 IBM Corp. 网站 http://www.ibm.com/support。在请求帮助时，请做好准备，以便识别您自己、您的组织以及您的支持协议。

内容

10 统计模型 220

章 1

关于 IBM SPSS Modeler

IBM® SPSS® Modeler 是一组数据挖掘工具，通过这些工具可以采用商业技术快速建立预测性模型，并将其应用于商业活动，从而改进决策过程。SPSS Modeler 参照行业标准 CRISP-DM 模型设计而成，可支持从数据到更优商业成果的整个数据挖掘过程。

SPSS Modeler 提供了各种借助机器学习、人工智能和统计学的建模方法。通过建模选项板中的方法，您可以根据数据生成新的信息以及开发预测模型。每种方法各有所长，同时适用于解决特定类型的问题。

SPSS Modeler 可以作为独立产品购买，或与 SPSS Modeler Server 一起使用。同时提供了大量其他选项，以下各节将对这些选项进行概述。 有关详细信息，请参阅 http://www.ibm.com/software/analytics/spss/products/modeler/。

IBM SPSS Modeler Server

SPSS Modeler 使用客户端/服务器体系结构将资源集约型操作的请求分发给功能强大的服务器软件，因而使大数据集的传输速度大大加快。除了此处所列的产品和更新，也可能还有其他可用的产品和更新。 有关详细信息，请参阅 http://www.ibm.com/software/analytics/spss/products/modeler/。

SPSS Modeler。SPSS Modeler 是具有完整功能的产品，它安装并运行于用户的台式计算机上。它既可以在本机模式下独立运行，也可以与 IBM® SPSS® Modeler Server 一起联机使用，从而提高了对大数据集的处理速度。

SPSS Modeler Server. SPSS Modeler Server 与一个或多个 IBM® SPSS® Modeler 安装程序一起在分布式分析模式下不间断运行，这种运行方式大大提高了对大数据集的处理速度，因为在服务器上可以完成内存集约型操作，且无需将数据下载至客户端计算机。SPSS Modeler Server 还提供对 SQL 优化和数据库内建模功能的支持，从而在性能和自动化方面带来更多优势。要运行分析，必须至少安装有一个 SPSS Modeler。

IBM SPSS Modeler 选项

还可以单独购买以下组件和功能并获得使用许可，以用于 SPSS Modeler。请注意，还可能提供其他产品或更新。 有关详细信息，请参阅 http://www.ibm.com/software/analytics/spss/products/modeler/。

- SPSS Modeler Server 访问权限，可针对大型数据集提供更高的可扩展性和性能，并提供对 SQL 优化以及数据库内建模功能的支持。

- SPSS Modeler Solution Publisher，用于在 SPSS Modeler 环境外执行实时或自动评分。 有关详细信息，请参阅第 2 章中的IBM SPSS Modeler Solution Publisher中的IBM SPSS Modeler 14.2 解决方案发布者。
- 允许部署到 IBM SPSS Collaboration and Deployment Services 或精简客户端应用程序 IBM SPSS Modeler Advantage 的适配器。 有关详细信息，请参阅第 9 章中的存储和部署 IBM SPSS Collaboration and Deployment Services Repository 对象中的IBM SPSS Modeler 14.2 用户 指南。

IBM SPSS Text Analytics

IBM® SPSS® Text Analytics 是一个 SPSS Modeler 完全集成内插式插件，它采用了先进语言技术和 Natural Language Processing (NLP)，以快速处理大量无结构文本数据，抽取和组织关键概念，以及将这些概念分为各种类别。抽取的概念和类别可以和现有结构化数据中进行组合（例如人口统计学），并且可用于借助 IBM® SPSS® Modeler 的一整套数据挖掘工具来进行建模，以此实现更好更集中的决策。

- 文本挖掘节点提供了概念、类型建模以及交互式工作平台，通过此平台，可以完成文本链接和聚类的高级探索，创建自己的类别和改进语言资源模板。
- 支持多种导入格式，其中包括“博客”和其他基于 Web 的资源。
- 还包括定制模板、库和指定域的词典，例如 CRM 和神经网络。

注：访问此组件需要单独许可证。 有关详细信息，请参阅 http://www.ibm.com/software/analytics/spss/products/modeler/。

IBM SPSS Modeler 文档

可以从 SPSS Modeler 的帮助菜单中获取在线帮助格式的完整文档。此文档包括 SPSS Modeler、SPSS Modeler Server 和 SPSS Modeler Solution Publisher 的文档以及《应用程序指南》和其他支持材料。

每个产品的完整文档（PDF 格式）也位于每个产品 DVD 的 \Documentation 文件夹下。

- **IBM SPSS Modeler 用户指南。**使用 SPSS Modeler 的一般使用介绍，包括如何构建数据流、处理缺失值、生成 CLEM 表达式、处理项目和报告以及将用于部署的流打包为 IBM SPSS Collaboration and Deployment Services、预测应用程序或 IBM SPSS Modeler Advantage。
- **IBM SPSS Modeler 源、处理和输出节点。** 介绍用于以不同的格式读取、处理和输出数据的所有节点。实际上这表示所有节点而非建模节点。
- **IBM SPSS Modeler 建模节点。** 有关用于创建数据挖掘模型的所有节点的描述。IBM® SPSS® Modeler 可提供各种借助机器学习、人工智能和统计学的建模方法。 有关详细信息，请参阅第 21 页码第 3 章中的建模节点概述。
- **IBM SPSS Modeler 算法指南。**介绍 SPSS Modeler 中所用建模方法的数学基础。
- **IBM SPSS Modeler 应用程序指南。**本指南中的示例旨在为具体的建模方法和技术提供具有针对性的简介。还可以在“帮助”菜单中查阅本指南的在线版本。 有关详细信息，请参阅应用程序示例中的IBM SPSS Modeler 14.2 用户 指南。

- **IBM SPSS Modeler 脚本编写与自动化。** 通过编写脚本实现系统自动化的相关信息，包括用于操作节点和流的属性信息。
- **IBM SPSS Modeler 部署指南。** 有关在 IBM® SPSS® Collaboration and Deployment Services Deployment Manager 中以处理作业的步骤形式运行 SPSS Modeler 流和方案的信息。
- **IBM SPSS Modeler CLEF 开发人员指南**CLEF 提供了将第三方程序（例如，数据处理例程或建模算法）作为节点集成到 SPSS Modeler 的功能。
- **IBM SPSS Modeler 数据库内数据挖掘指南。**有关如何利用数据库的功能通过第三方算法来改进性能并增强分析功能的信息。
- **IBM SPSS Modeler Server 和性能指南。** 有关如何配置和管理 IBM® SPSS® Modeler Server 的信息。
- **IBM SPSS Modeler Administration Console 用户指南。**有关安装和使用控制台用户界面以监视和配置 SPSS Modeler Server 的信息。控制台实现为 Deployment Manager 应用程序的插件。
- **IBM SPSS Modeler Solution Publisher 指南。**SPSS Modeler Solution Publisher 是一个附加式组件，通过它组织可发布在标准 SPSS Modeler 环境之外使用的流。
- **IBM SPSS Modeler CRISP-DM 指南。** 借助 CRISP-DM 方法进行 SPSS Modeler 数据挖掘的分步指南。

应用程序示例

SPSS Modeler 中的数据挖掘工具可以帮助解决很多业务和组织问题，应用程序示例将提供有关特定建模方法和技术的简要的针对性说明。此处使用的数据集比某些数据挖掘器管理的大量数据存储要小得多，但涉及的概念和方法应可扩展到实际的应用程序。

可以通过在 SPSS Modeler 中的“帮助”菜单中单击**应用程序示例**来访问示例。数据文件和样本流安装在产品安装目录下的 Demos 文件夹中。有关详细信息，请参阅 Demos 文件夹中的IBM SPSS Modeler 14.2 用户 指南。

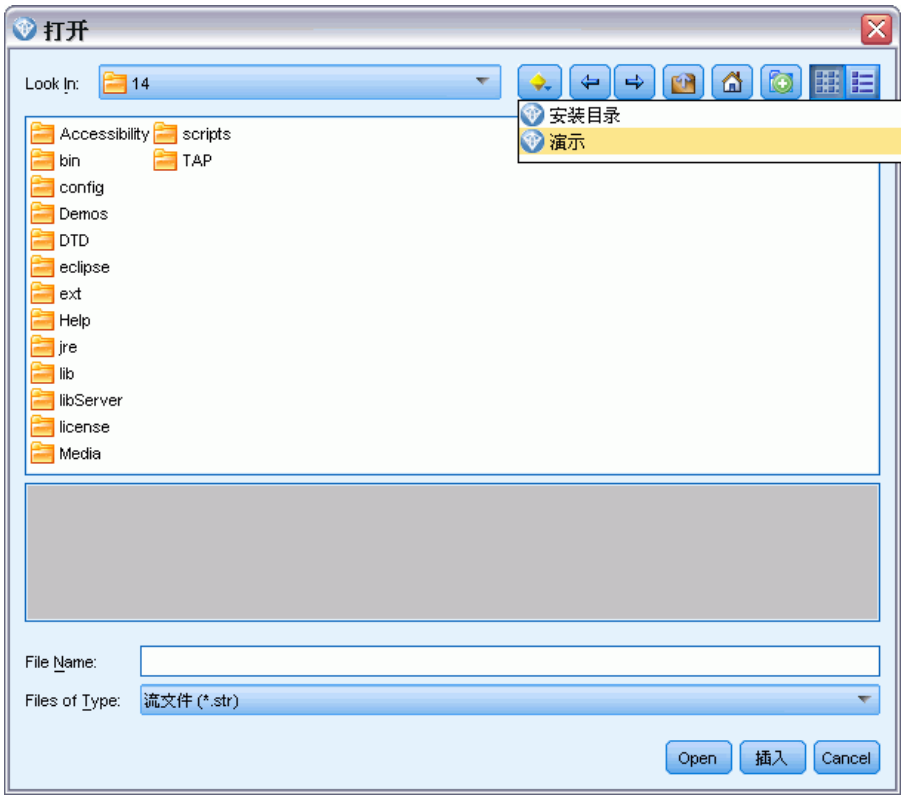
数据库建模示例。请参阅 IBM SPSS Modeler 数据库内挖掘指南 中的示例。

编写示例脚本。请参阅 IBM SPSS Modeler 脚本编写和自动化指南 中的示例。

Demos 文件夹

与应用程序示例一起使用的数据文件和样本流安装在产品安装目录下的 Demos 文件夹中。可从 Windows 的“开始”菜单中 IBM SPSS Modeler 14.2 程序组访问该文件夹，也可以在“文件打开”对话框中最近目录的列表中单击 Demos。

图片 1-1
在最近使用的目录列表中选择 Demos 文件夹



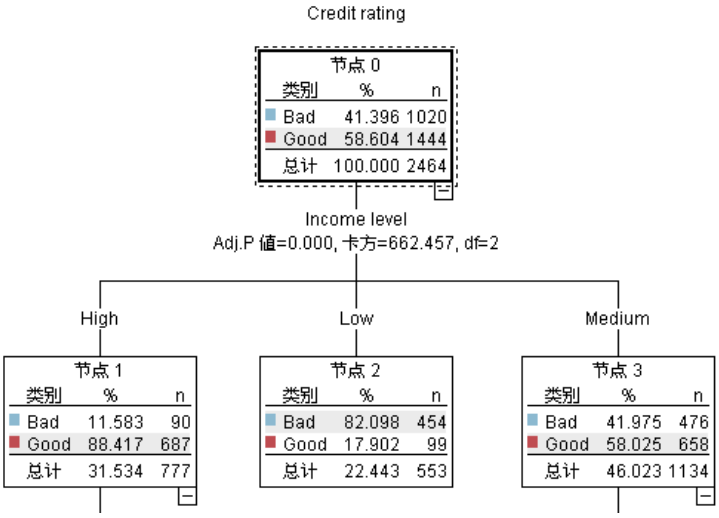
章
2

建模简介

模型是一组规则、公式或方程式，可以用它们根据一组输入或变量来预测输出。例如，一家财务机构可根据对过往申请人的已知信息，使用模型预测贷款申请人可能存在优良还是不良风险。

能够预测结果是预测性分析的中心目标，并且了解建模过程是使用 IBM® SPSS® Modeler 的关键。

图片 2-1
简单的决策树模型



本示例使用**决策树**模型，该模型使用一系列决策规则对记录进行分类（并预测响应），例如：

如果收入 = 中等
并且卡 <5
则 -> “优良”

本示例使用 CHAID（卡方自动交互效应检测）模型时，旨在进行常规的介绍，大部分概念会广泛应用于 SPSS Modeler 中的其他建模类型。

无论要了解哪种模型，均需要首先了解进入该模型的数据。此示例中的数据包含有关银行客户的信息。其中使用了下列字段：

字段名	描述
Credit_rating	信用评价：0=不良，1=优良，9=丢失值
年龄	Age in years
收入	收入水平：1=低，2=中，3=高

字段名	描述
Credit_cards	持有的信用卡数量：1=少于五张，2=五张或更多
教育	教育程度：1=高中，2=大学
Car_loans	贷款的汽车数量：1=没有或一辆，2=超过两辆

银行可维护一个包含银行贷款客户历史信息，包括这些客户是正在还贷（信用评价 = 优良）还是在拖欠贷款（信用评价 = 不良）的数据库。银行希望使用现有的数据建立一个模型，允许他们预测未来贷款申请人拖欠贷款的可能性。

使用决策树模型，您可分析两组客户的特征，并预测拖欠贷款的可能性。

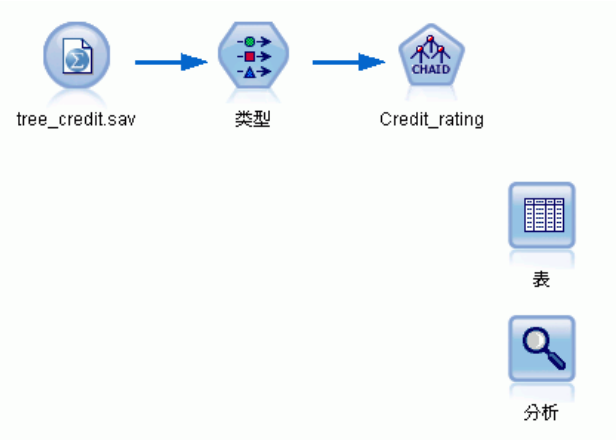
本示例使用了名为 modelingintro.str 的流，该流位于 streams 子文件夹下的 Demos 文件夹中。数据文件是 tree_credit.sav。有关详细信息，请参阅第 1 章中的Demos 文件夹中的IBM SPSS Modeler 14.2 用户 指南。

我们来看一下流。

- 从主菜单中选择下列选项：
  文件 > 打开流
- 单击“打开”对话框的工具栏上的金色模型块图标，然后选择 Demos 文件夹。
- 双击 streams 文件夹。
- 双击名为 modelingintro.str 的文件。

构建流

图片 2-2
建模流



要构建流以创建模型，至少需要三个元素：

- 一个从某些外部源读取数据的源节点，在本示例中为 IBM® SPSS® Statistics 数据文件。
- 一个指定字段属性的源节点或“类型”节点，字段属性包括测量级别（字段包含的数据类型）以及每个字段在建模过程中的角色是目标还是输入等。
- 一个在运行流时生成模型块的建模节点。

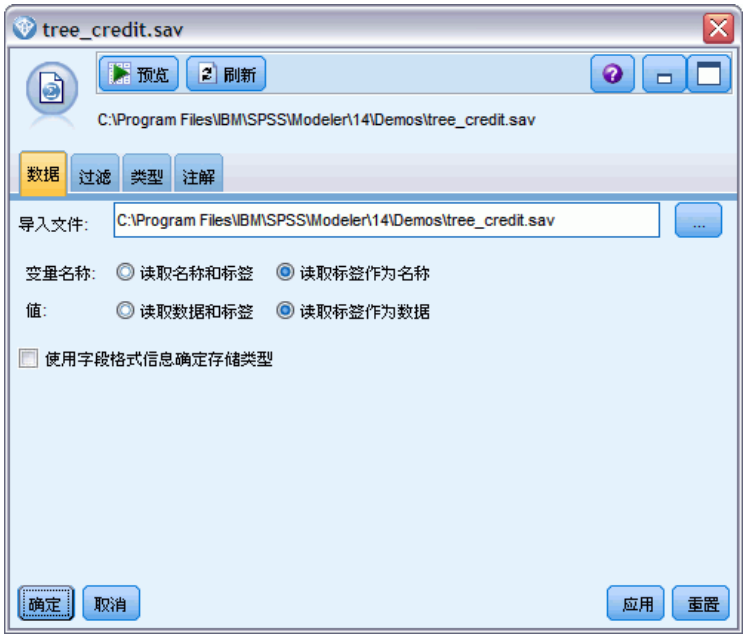
在本例中，我们使用 CHAID 建模节点。CHAID，或卡方自动交互效应检测，是一种通过使用称作卡方统计量的特定统计类型识别决策树中的最优分割来构建决策树的分类方法。

如果在源节点中指定了测量级别，则可以去掉单独的“类型”节点。从功能上来说，结果是一样的。

该流中还包含“表”节点和“分析”节点，创建模型块并将其添加到流中后，将使用这两个节点查看评分结果。

Statistics 文件源节点从 tree_credit.sav 数据文件读取 SPSS Statistics 格式数据，该文件安装在 Demos 文件夹中。（名为 $CLEO_DEMOS 的特殊变量用于引用位于当前 IBM® SPSS® Modeler 安装下的该文件。这样，无论当前的安装文件夹或版本是什么，均可以确保路径有效。）

图片 2-3
使用 Statistics 文件源节点读取数据



类型节点指定每个字段的**测量级别**。测量级别是一种指示字段中数据类型的类别。我们的源数据文件使用三种不同的测量级别。

连续字段（例如年龄字段）包含连续的数字值，而**名义**字段（例如信用评价字段）有两个或多个不同值，例如不良、优良或无信用历史。**有序**字段（例如收入水平字段）用于描述具有顺序固定的不同值的数据，在本例中为低、中和高。

图片 2-4
用类型节点设置目标和输入字段

对于每个字段，类型节点还指定**角色**，以指示每个字段在建模中扮演的部分。将字段信用评价的角色设置为目标，此字段指示指定的客户是否拖欠贷款。这是**目标**，或者是要预测其值的字段。

对于其他字段，将角色设置为输入。输入字段有时也称为**预测变量**，或建模算法用其值来预测目标字段值的字段。

CHAID 建模节点生成模型。

在建模节点的“字段”选项卡中，已选中**使用预定义角色**，这意味着将按在类型节点中的指定使用目标和输入。我们可以在此处更改字段角色，但本例中我们不做任何更改使用这些角色。

▶ 单击“构建选项”选项卡。

图片 2-5
CHAID 建模节点、“字段”选项卡

此处包含的选项可以用于指定要构建的模型类型。

由于我们想要一个全新的模型，因此使用默认选项**构建新模型**。

我们还要求它为单个标准决策树模型，并且不包含任何增强，因此保留默认目标选项**构建单个树**。

我们可以选择启动允许对模型进行微调的交互建模会话，本示例只使用默认设置**生成模型**来生成模型。

图片 2-6
CHAID 建模节点、“构建选项”选项卡

对于此示例，我们希望保持树的结构简单，因此通过增加用于父节点和子节点的最小个案数限制树的增长。

- 在“构建选项”选项卡上，从左侧的导航器窗格选择**停止规则**。
- 选择**使用绝对值**选项。
- 将**父分支中的最小记录数**设置为 400。

▶ 将子分支中的最小记录数设置为 200。

图片 2-7
为构建决策树设置停止标准

在本例中，我们可以使用所有其他默认选项，因此单击运行以创建模型。（另外，也可以右键单击该节点，然后从上下文菜单中选择运行，或选择节点，并从“工具”菜单中选择运行。）

浏览模型

执行完成后，模型块将添加到应用程序窗口右上角的“模型”选项板中，它还会置于流工作区中，并带有指向创建它的建模节点的链接。要查看模型的详细信息，右键单击模型块并选择**浏览**（在模型选项板上）或**编辑**（在工作区上）。

图片 2-8
“模型”选项板

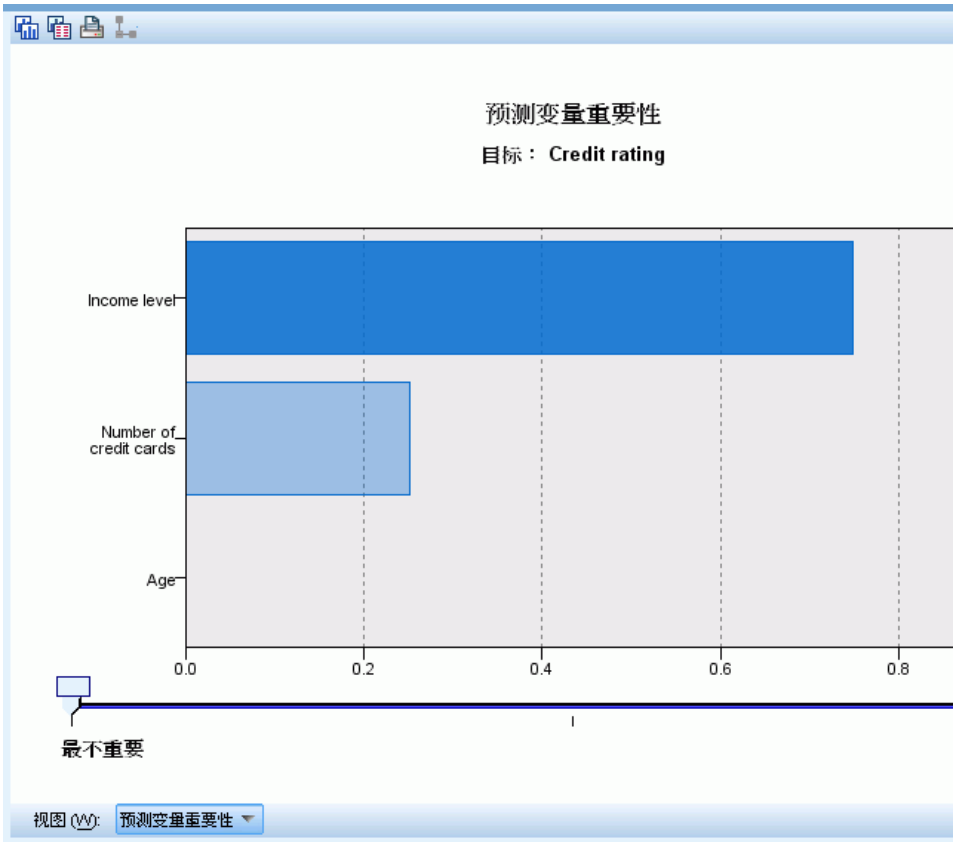
对于 CHAID 模型块，“模型”选项卡以规则集的形式显示详细信息，规则集实际上是可根据不同输入字段的值将各个记录分配给子节点的一组规则。

图片 2-9
CHAID 模型块、规则集

对于每个决策树终端节点—意味着那些树节点没有进一步拆分—返回优良或不良的预测值。对于落在该节点内的记录，所有个案中的预测均由**模式**或最常见的响应决定。

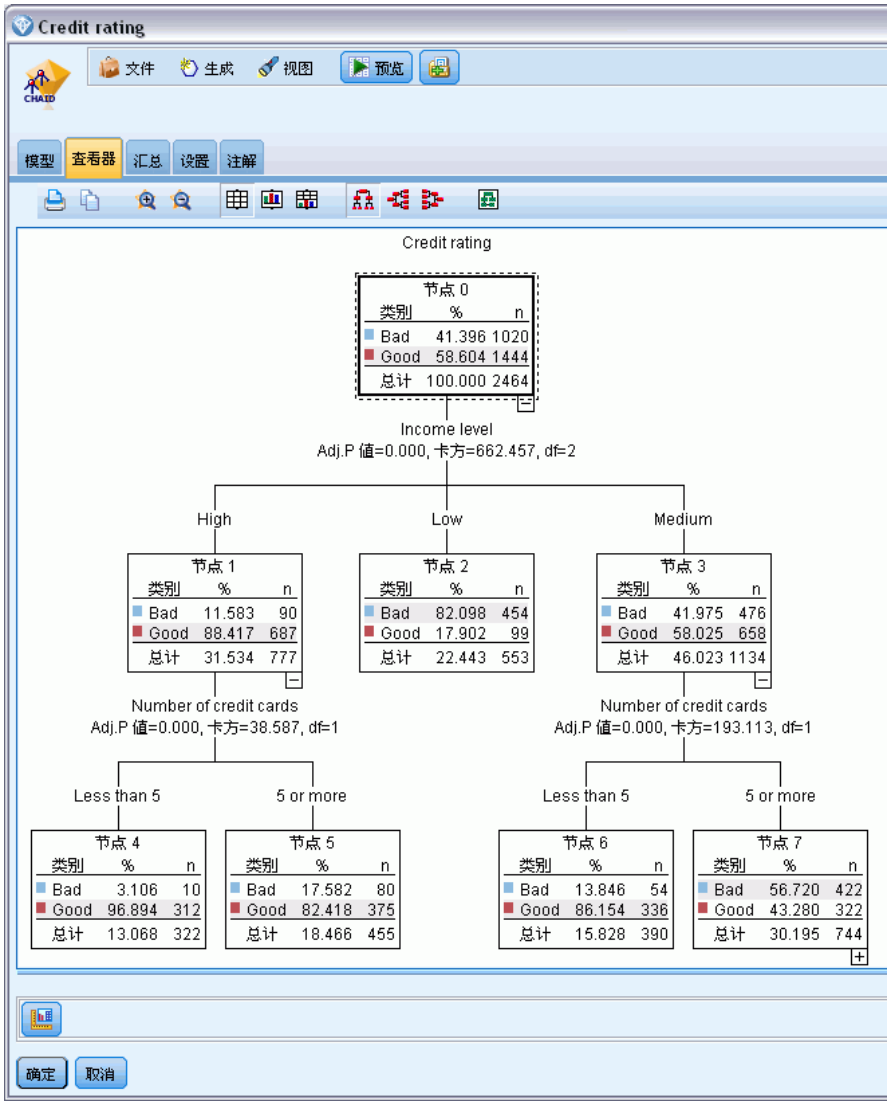
在规则集的右侧，“模型”选项卡显示预测变量重要性图表，该图表显示评估模型时每个预测变量的相对重要性。通过这一点，我们看到收入水平在此个案中最显著，而其他唯一显著的因子是信用卡数量。

图片 2-10
预测变量重要性图表



模型块中的“查看器”选项卡以树的形式显示相同的模型，每个决策点上都有一个节点。可使用工具栏上的缩放控件放大特定节点，或缩小节点以查看更完整的树。

图片 2-11
模型块中的查看器选项卡，已选择缩小



查看树的上部分，第一个节点（节点 0）为我们提供数据集中所有记录的摘要。数据集中超过 40% 的个案分类为不良风险。这是相当高的比例，因此让我们看看树是否能为我们提供哪些因子负责的任何线索。

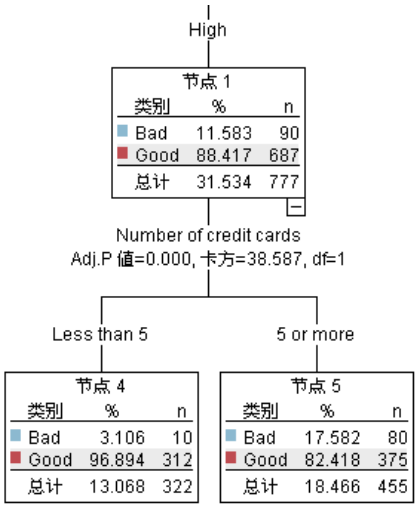
我们可以看到第一个分割是根据收入水平。收入水平位于低类别的记录被指定到节点 2，并且看到此类别包含贷款拖欠人的最高百分比不足为奇。我们可以很明显地了解，此类别中的客户具有高风险。

但是，此类别中的 16% 客户实际上没有拖欠，因此预测并非始终准确。没有模型能够预测每一个响应，但好的模型能够根据可用数据预测对每一个记录作出的最常见的响应。

同样，如果我们查看高收入客户（节点 1），我们看到绝大部分（89%）是优良风险。但是在这些客户中 10 位中有超过 1 位也会拖欠。我们能精炼自己的贷款标准以便将此处的风险最小化吗？

注意模型如何根据持有的信用卡数量，将这些客户分成两个子类别（节点 4 和节点 5）。对于高收入客户，如果我们只向那些信用卡少于 5 张的客户贷款，则可以将我们的成功率从 89% 提高到 97%—甚至更满意的结果。

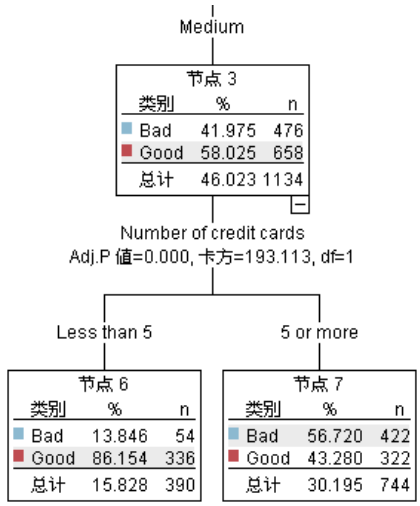
图片 2-12
高收入客户的树状视图



但中等收入类别（节点 3）中的那些客户是什么情况？他们更加均匀地划分为优良和不良评价。

子类别（此情况中是节点 6 和 7）仍然能帮助我们。这次，只向那些信用卡少于 5 张的中等收入客户贷款，可将优良评价的百分比从 58% 提高到 85%，这是显著的改进。

图片 2-13
中等收入客户的树状视图



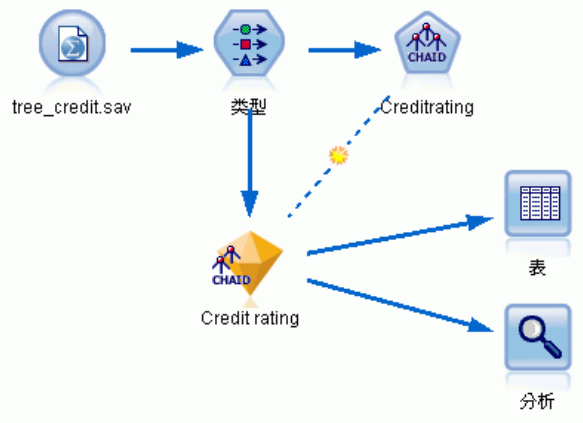
因此，我们了解到输入此模型的每项记录都将被分配到一个特定节点，并且根据该节点最常见的响应分配优良或不良的预测值。

为各个记录分配预测值的这一过程称为**评分**。通过对用于评估该模型的相同记录进行评分，可以评估该模型执行训练数据（我们知道结果的数据）的准确度。让我们看看如何做到这一点。

评估模型

我们浏览了模型以了解评分方式。但是，如果要评估模型的准确度，则需要对一些记录进行评分，并将模型预测的响应与实际结果进行比较。接下来对用于评估该模型的相同记录进行评分，以将观察到的响应与预测响应进行比较。

图片 2-14
将模型块附加到输出节点用于模型评估

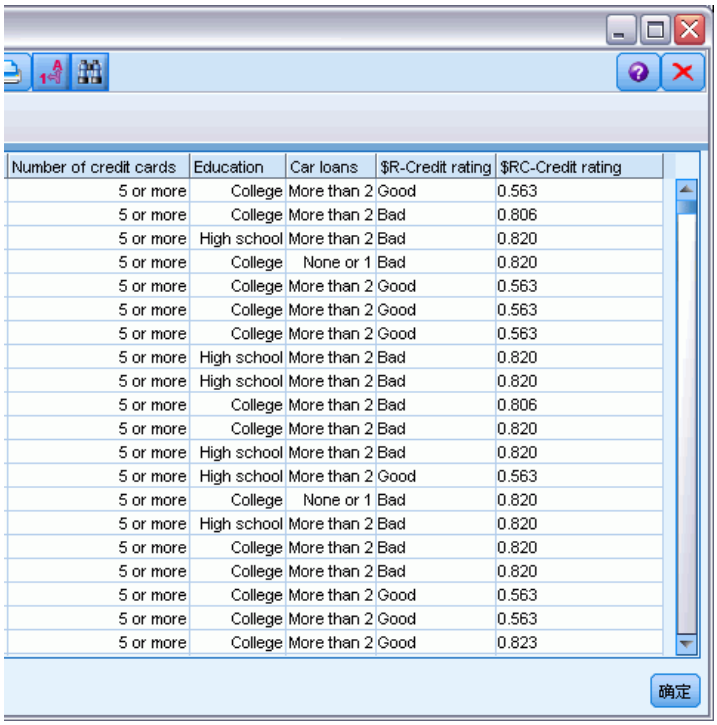


- 要查看分数或预测值，请将表节点添加到模型块，然后双击“表”节点，并单击**运行**。

  表在名为 $R-Credit rating 的字段中显示预测分数，该字段由模型创建。我们可以将这些值与包含实际响应的原始信用评价字段进行比较。

按照惯例，在评分过程中生成的字段的名称基于目标字段，但是要加上标准前缀，例如 $R- 表示预测值，$RC- 表示置信度值。不同的模型类型使用不同的前缀集。**置信度值**是模型自己的评估，尺度从 0.0 到 1.0，表示每个预测值的精确程度。

图片 2-15
表格显示生成的分数和置信度值



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Bad	0.806
5 or more	High school	More than 2	Bad	0.820
5 or more	College	None or 1	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.806
5 or more	College	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Good	0.563
5 or more	College	None or 1	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.823

与预期的一样，预测值与大多数（并非全部）记录的实际响应相匹配。原因是每个 CHAID 终端节点均有混合响应。预期值与最常见 的响应相匹配，但对于该节点中的其他响应，该预期值是错误的。（记住，16% 的少部分低收入客户没有拖欠。）

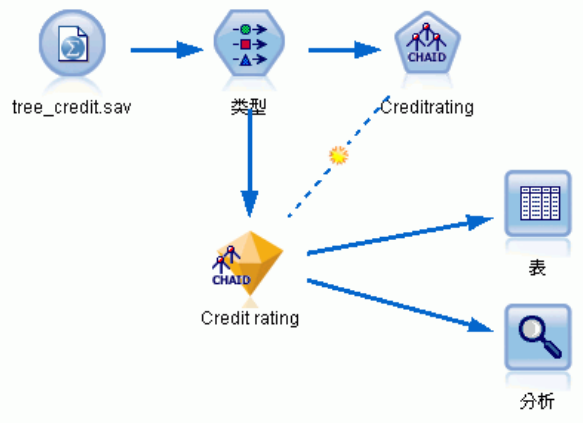
若要避免出现此情况，应继续将树分割为更小的分支，直到每个节点都是不含混合响应的 100% 纯节点为止—即全部为优良或不良。但是，这样的模型可能会非常复杂，并且不易推广到其他数据集。

要查看具体有多少预测值正确，我们可通读表格，并计算预测字段 $R-Credit rating 的值匹配信用评价的值的记录数量。幸运的是，这里有更简单的方式—我们可使用分析节点，它自动进行此项操作。

- 将模型块连接到分析节点。

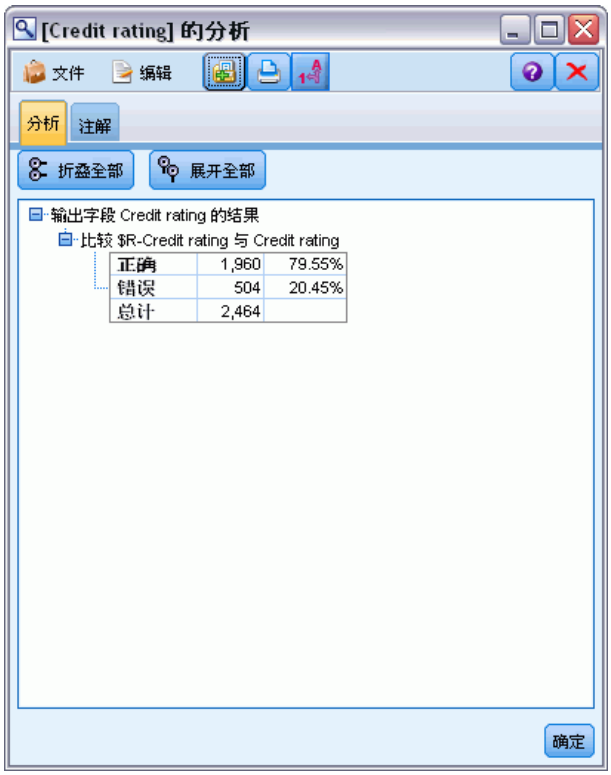
▶ 双击“分析”节点，然后单击运行。

图片 2-16
添加分析节点



分析表明，2464 个记录中有 1899 个记录（超过 77%）的模型预测值与实际响应相匹配。

图片 2-17
观察到的响应与预测的响应的比较分析结果



此结果受到评分的记录和用于评估模型的记录相同的事实的限制。在真实情况中，可使用分区节点将数据分割为培训和评估的单独示例。

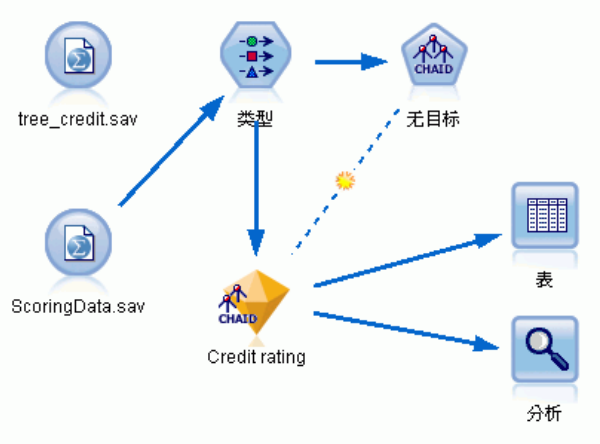
通过使用一个样本分区生成模型并使用另一个样本对模型进行检验，您会得到该模型推广到其他数据集的情况。

通常分析节点，可以针对已知道其实际结果的记录来检验模型。下一阶段介绍如何使用模型对我们不知道结果的记录进行评分。例如，这可能包括当前不是银行客户的人员，但他们是促销邮寄的潜在目标。

对记录评分

之前，我们对用于评估模型的相同记录进行了评分，以评价模型准确程度。现在，我们要查看如何对和用于创建模型不同的记录集进行评分。这是使用目标字段进行建模的目标：研究已知道其结果的记录，以标识您可以从中预测未知结果的模式。

图片 2-18
附加用于评分的新数据



可以更新 Statistics 文件源节点，使它指向其他数据文件，也可以添加一个新的源节点，从它读取要评分的数据。无论采用哪种方式，新数据集包含的输入字段必须与模型（年龄、收入水平、教育等）所使用的相同，但不包含目标字段信用评价。

另外，也可以将模型块添加到包含预期的输入字段的任何流中。无论是读取文件还是数据库，只要字段名和类型与模型使用的相匹配，源类型都无关紧要。

也可以将模型块保存为单独的文件、将模型导出为 PMML 格式以用于其他支持此格式的应用程序，或将模型存储到 IBM® SPSS® Collaboration and Deployment Services 存储库中，这样可以在企业范围对模型进行部署、评分和管理。

无论使用何种基础结构，模型自身都按相同的方式工作。

摘要

本示例演示创建、评估模型以及对模型评分的基本步骤。

- 建模节点通过研究已知道其结果的记录来评估模型，并创建模型块。这有时称为训练模型。
- 可将模型块添加到包含预期字段的任何流中，以对记录进行评分。通过对已知道其结果的记录（如现有客户）进行评分，可以评估模型的运行情况。
- 如果您对模型的运行情况感到满意，则可以对新数据（如准客户）进行评分，以预测他们的响应。
- 用于训练或评估模型的数据可以称为分析数据或历史数据；评分数据也可以称为操作数据。

章 3

建模概述

建模节点概述

IBM® SPSS® Modeler 提供了各种借助机器学习、人工智能和统计学的建模方法。通过建模选项板中的方法，您可以根据数据生成新的信息以及开发预测模型。每种方法各有所长，同时适用于解决特定类型的问题。

SPSS Modeler 应用程序指南 为上述多种方法提供了示例以及建模过程的一般介绍。本指南既作为联机教程提供，还有 PDF 格式。有关详细信息，请参阅第 1 章中的应用程序示例中的IBM SPSS Modeler 14.2 用户 指南。

建模方法划分为三种类别：

- Classification
- 关联
- 细分

分类模型

分类模型使用一个或多个**输入**字段的值来预测一个或多个输出（或**目标**）字段的值。这些技术的部分示例为：决策树（C&R 树、QUEST、CHAID 和 C5.0 算法）、回归（线性、logistic、广义线性和 Cox 回归算法）、神经网络、Support Vector Machine (SVM) 和贝叶斯网络。

分类模型可帮助组织预测已知的结果，例如顾客是否购买、流失或某交易是否符合某种已知的犯罪模式。其建模方法包括机器学习、规则归纳、子组标识、统计方法和多模型生成。

分类节点

“自动分类器”节点用于创建和对比二元结果（是或否，流失或不流失等）的若干不同模型，使用户可以选择给定分析的最佳处理方法。由于支持多种建模算法，因此可以对用户希望使用的方法、每种方法的特定选项以及对比结果的标准进行选择。节点根据指定的选项生成一组模型并根据用户指定的标准排列最佳候选项的顺序。 有关详细信息，请参阅第 81 页码第 5 章中的自动分类器节点。

自动数值节点使用多种不同方法估计和对比模型的连续数字范围结果。此节点和自动分类器节点的工作方式相同，因此可以选择要使用和要在单个建模传递中使用多个选项组合进行测试的算法。受支持的算法包括神经网络、C&R 树、CHAID、线性回归、广义线性回归以及 Support Vector Machine (SVM)。可基于相关度、相对错误或已用变量数对模型进行对比。 有关详细信息，请参阅第 90 页码第 5 章中的自动数值节点。

分类和回归（C&R）树节点生成可用于预测或分类未来观测值的决策树。该方法通过在每个步骤最大限度降低不纯洁度，使用递归分区来将训练记录分割为组。如果节点中 100% 的观测值都属于目标字段的一个特定类别，则树中的该节点将被认定为“纯洁”。目标和输入字段可以是数字范围或分类（名义、有序或标志）；所有分割均为二元分割（即仅分割为两个子组）。 有关详细信息，请参阅第 130 页码第 6 章中的C&R 树节点。

QUEST 节点可提供用于构建决策树的二元分类法，此方法的设计目的是减少大型 C&R 树分析所需的处理时间，同时也减少在分类树方法中发现的趋势以便支持允许有多个分割的输入。输入字段可以是数字范围（连续），但目标字段必须是分类。所有分割都是二元的。 有关详细信息，请参阅第 131 页码第 6 章中的QUEST 节点。

CHAID 使用卡方统计量来生成决策树，以确定最佳的分割。CHAID 与 C&R 树和 QUEST 节点不同，它可以生成非二元树，这意味着有些分割将有多于两个的分支。目标和输入字段可以是数字范围（连续）或分类。Exhaustive CHAID 是 CHAID 的修正版，它对所有分割进行更彻底的检查，但计算时间比较长。 有关详细信息，请参阅第 130 页码第 6 章中的CHAID 节点。

C5.0 节点构建决策树或规则集。该模型的工作原理是根据在每个级别提供最大信息收获的字段分割样本。目标字段必须为分类字段。允许进行多次多于两个子组的分割。 有关详细信息，请参阅第 146 页码第 6 章中的C5.0 节点。

决策列表节点可标识子组或段，显示与总体相关的给定二元结果的似然度的高低。例如，您或许在寻找那些最不可能流失的客户或最有可能对某个商业活动作出积极响应的客户。通过定制段和并排预览备选模型来比较结果，您可以将自己的业务知识体现在模型中。决策列表模型由一组规则构成，其中每个规则具备一个条件和一个结果。规则依顺序应用，相匹配的第一个规则将决定结果。 有关详细信息，请参阅第 188 页码第 9 章中的决策表。

线性回归模型根据目标与一个或多个预测变量间的线性关系来预测连续目标。 有关详细信息，请参阅第 221 页码第 10 章中的线性模型。

线性回归是一种通过拟合直线或平面以实现汇总数据和预测的普通统计方法，它可使预测值和实际输出值之间的差异最小化。

因子/主成分分析节点提供了用于降低数据复杂程度的强大数据缩减技术。主成份分析（PCA）可找出输入字段的线性组合，该组合最好地捕获了整个字段集合中的方差，且组合中的各个成分相互正交（相互垂直）。因子分析则尝试识别底层因素，这些因素说明了观测的字段集合内的相关性模式。这两种方式的目标都是找到有效概括原始字段集中的信息的一小部分导出字段。 有关详细信息，请参阅第 255 页码第 10 章中的主成份分析/因子节点。

“特征选择”节点会根据某组条件（例如缺失值百分比）筛选可删除的输入字段；对于保留的输入，将相对于指定目标对其重要性进行排序。例如，假如某个给定数据集有上千个潜在输入，那么哪些输入最有可能用于对患者结果进行建模呢？ 有关详细信息，请参阅第 64 页码第 4 章中的特征选择节点。

判别式分析所做的假设比 logistic 回归的假设更严格，但在符合这些假设时，判别式分析可以作为 logistic 回归分析的有用替代项或补充。 有关详细信息，请参阅第 263 页码第 10 章中的判别式节点。

Logistic 回归是一种统计方法，它可根据输入字段的值对记录进行分类。它类似于线性回归，但采用的是类别目标字段而非数字范围。 有关详细信息，请参阅第 239 页码第 10 章中的逻辑节点。

“广义线性”模型对一般线性模型进行了扩展，这样因变量通过指定的关联函数与因子和协变量线性相关。另外，该模型允许因变量呈非正态分布。它包括统计模型大部分的功能，其中包括线性回归、logistic 回归、用于计数数据的对数线性模型以及区间删失生存模型。 有关详细信息，请参阅第 270 页码第 10 章中的GenLin 节点。

使用 Cox 回归节点，您可以在已有的检查记录中建立时间事件的生存模型。该模型会生成一个生存函数，该函数可预测在给定时间 (t) 内对于所给定的输入变量值相关事件的发生概率。 有关详细信息，请参阅第 282 页码第 10 章中的Cox 节点。

使用 Support Vector Machine (SVM) 节点，可以将数据分为两组，而无需过度拟合。SVM 可以与大量数据集配合使用，如那些含有大量输入字段的数据集。 有关详细信息，请参阅第 391 页码第 15 章中的SVM 节点。

通过贝叶斯网络节点，你可以利用对真实世界认知的判断力并结合所观察和记录的证据来构建概率模型。该节点重点应用了树扩展简单贝叶斯 (TAN) 和马尔可夫毯网络，这些算法主要用于分类问题。 有关详细信息，请参阅第 165 页码第 7 章中的贝叶斯网络节点。

自学响应模型（SLRM）节点可用于构建一个包含单个新观测值或少量新观测值的模型，通过此模型，无需使用全部数据对模型进行重新训练即可对模型进行重新评估。 有关详细信息，请参阅第 381 页码第 14 章中的SLRM 节点。

时间序列节点估计时间序列数据的指数平滑模型、单变量自回归整合移动平均 (ARIMA) 模型和多变量 ARIMA（即变换函数）模型，并生成未来性能的预测数据。在时间序列节点之前必须有时间区间节点。 有关详细信息，请参阅第 361 页码第 13 章中的时间序列建模节点。

The k-最近相邻元素 (KNN) 节点将新的个案关联到预测变量空间中与其最邻近的 k 个对象的类别或值（其中 k 为整数）。类似个案相互靠近，而不同个案相互远离。 有关详细信息，请参阅第 396 页码第 16 章中的KNN 节点。

关联模型

关联模型查找您数据中的模式，其中一个或多个实体（如事件、购买或属性）与一个或多个其他实体相关联。这些模型构建定义这些关系的规则集。数据中的字段可以作为输入和目标。您可以手动查找这些关联，但关联规则算法可以更快速地完成，并能探索更

多复杂的模式。Apriori 和 Carma 模型是使用此类算法的示例。另一种类型的关联模型是序列检测模型，后者可以在按时间建立结构的数据中查找顺序模式。

关联模型在预测多个结果时非常有用，例如，购买了产品 X 的顾客也购买了产品 Y 和 Z。关联模型可以将特定结论（如购买某些产品的决策）与一组条件关联起来。关联规则算法相对于更标准的决策树算法（C5.0 和 C&RT）的优势在于，它可以找到任何属性间存在的关联。决策树算法只使用单一结论来构建规则，而关联算法则试图找到更多规则，且每个规则具有不同的结论。

关联节点

Apriori 节点从数据抽取一组规则，即抽取信息内容最多的规则。Apriori 节点提供五种选择规则的方法并使用复杂的索引模式来高效地处理大数据集。对于较大的问题，Apriori 训练的速度通常较快；它对可保留的规则数量没有任何限制，而且可处理最多带有 32 个前提条件的规则。Apriori 要求输入和输出字段均为分类型字段，但因为它专为处理此类型数据而进行优化，因而处理速度快得多。 有关详细信息，请参阅第 321 页码第 12 章中的Apriori 节点。

CARMA 模型在不要求用户指定输入或目标字段的情况下从数据抽取一组规则。与 Apriori不同的是：CARMA 节点提供构建规则设置支持（前项和后项支持），而不仅仅是前项支持。这就意味着生成的规则可以用于更多应用程序，例如用于找到后项为想在节日期间促销的商品的产品或服务（前项）的列表。 有关详细信息，请参阅第 324 页码第 12 章中的CARMA 节点。

序列节点可发现连续数据或与时间有关的数据中的关联规则。序列是一系列可能会以可预测顺序发生的项目集合。例如，一个购买了剃刀和须后水的顾客可能在下次购物时购买剃须膏。序列节点基于 CARMA 关联规则算法，该算法使用有效的两步法来发现序列。 有关详细信息，请参阅第 343 页码第 12 章中的序列节点。

细分模型

细分模型将数据划分为具有类似输入字段模式的记录段或聚类。细分模型只对输入字段感兴趣，没有输出或目标字段的概念。细分模型的示例为 Kohonen 网络、K-Means 聚类、两步聚类和异常检测等。

在不知道特定结果的情况下（例如，需要识别新犯罪模式或在客户群中识别利益群体时），细分模型（也称为“聚类模型”）非常有用。聚类模型主要用来确定相似记录的组并根据它们所属的组来为记录添加标签。此方法的优点在于，不用提前了解这些组及其特征就可以使用，它使聚类模型（其中没有需要模型预测的预定义输出或目标字段）区别于其他的建模技术。对于这些模型来说，没有正确或错误的结果之分。模型的值由模型捕获数据中感兴趣的分组并提供这些分组的有用说明信息的能力来确定。聚类模型通常用于创建在后续分析中用作输入的聚类或段（例如，将潜在用户分成几个相似的子组）。

细分节点

自动聚类节点估算和比较识别具有类似特征记录组的聚类模型。节点工作方式与其他自动建模节点相同，使您在一次建模运行中即可试验多个选项组合。模型可使用基本测量进行比较，以尝试过滤聚类模型的有效性以及对其进行排序，并提供一个基于特定字段的重要性的测量。 有关详细信息，请参阅第 95 页码第 5 章中的自动聚类节点。

K-Means 节点将数据集聚类到不同分组（或聚类）。此方法将定义固定的聚类数量，将记录迭代分配给聚类，以及调整聚类中心，直到进一步优化无法再改进模型。k-means 节点作为一种非监督学习机制，它并不试图预测结果，而是揭示隐含在输入字段集中的模式。 有关详细信息，请参阅第 298 页码第 11 章中的K-Means 节点。

Kohonen 节点会生成一种神经网络，此神经网络可用于将数据集聚类到各个差异组。此网络训练完成后，相似的记录应在输出映射中紧密地聚集，差异大的记录则应彼此远离。您可以通过查看模型块中每个单元所捕获观测值的数量来找出规模较大的单元。这将让您对聚类的相应数量有所估计。 有关详细信息，请参阅第 293 页码第 11 章中的Kohonen 节点。

TwoStep 节点使用两步聚类方法。第一步完成简单数据处理，以便将原始输入数据压缩为可管理的子聚类集合。第二步使用层级聚类方法将子聚类一步一步合并为更大的聚类。TwoStep 具有一个优点，就是能够为训练数据自动估计最佳聚类数。它可以高效处理混合的字段类型和大型的数据集。有关详细信息，请参阅第 302 页码第 11 章中的两步聚类节点。

“异常检测”节点确定不符合“正常”数据格式的异常观测值（离群值）。即使离群值不匹配任何已知格式或用户不清楚自己的查找对象，也可以使用此节点来确定离群值。 有关详细信息，请参阅第 70 页码第 4 章中的异常检测节点。

数据库内数据挖掘模型

SPSS Modeler 支持对数据库提供商的数据挖掘工具和建模工具进行整合，其中包括 Oracle Data Miner、IBM DB2 InfoSphere Warehouse 和 Microsoft Analysis Services。您可以使用 SPSS Modeler 应用程序在数据库中构建、评分和存储模型。有关详细信息，请参阅产品 DVD 上的《SPSS Modeler 数据库内数据挖掘指南》。

IBM SPSS Statistics 模型

如果您在计算机上拥有 IBM® SPSS® Statistics 安装和许可的一个副本，您可以从 SPSS Modeler 访问和运行某些 SPSS Statistics 例程以构建模型和给模型评分。有关详细信息，请参阅第 8 章中的IBM SPSS Statistics 节点 - 概述中的IBM SPSS Modeler 14.2 源、过程和输出节点。

其他信息

此外还有一些有关建模算法的详细文档。有关详细信息，请参阅产品 DVD 上的《SPSS Modeler 算法指南》。

构建分割模型

分割模型能够使用一个流为标志、名义或连续输入字段的每个可能值构建单独的模型，可从一个模型块访问全部得出模型。输入字段的可能值可能对模型具有非常不同的影响。使用分割建模，您可以容易地在流的一次执行中为每个可能的字段值构建最佳拟合模型。

请注意，交互建模会话不能使用分割。您通过互动建模单独指定每个模型，而使用分割会自动构建多个模型，所以使用分割没有优势。

分割建模会指定某个输入字段为分割字段。在“类型”规范中设置**分割**的字段角色完成此操作：

图片 3-1
指定输入字段为分割字段

您仅可将测量级别为**标志**、**名义**、**有序**或**连续**的字段指定为分割字段。

您可以将多个输入字段分配为分割字段。但是这种情况下，所创建模型数量可能大增。给所选分割字段值的每个可能组合构建一个模型。例如，如果三个输入字段指定为分割字段，每个字段具有三个可能值，则结果会创建 27 个不同模型。

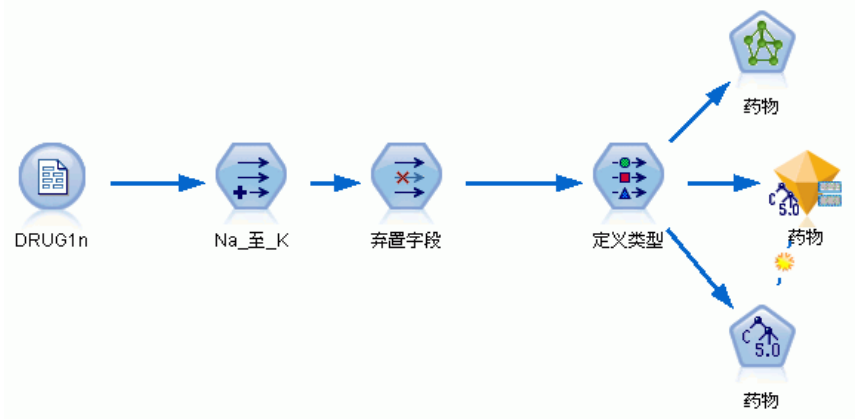
即使您指定一个或多个字段为分割字段后，您仍可通过建模节点对话框上的复选框设置选择创建多个分割模型还是一个模型：

图片 3-2
选择构建分割模型

如果定义了分割字段但未选择复选框，则只生成一个模型。同样，如果选择了复选框但未定义分割字段，则分割被忽略，生成一个模型。

当您执行流时，在后台为分割字段的每个可能值构建单独的模型，但只有一个模型块置于模型选项板和流工作区中。以下分割符号表示分割模型块：

图片 3-3
流中的分割模型块



当您浏览分割模型块时，您会看到一个已经创建的所有单独模型的列表：

图片 3-4
拆分模型浏览器

您可以通过在浏览器中双击块从列表中查看单个模型。这样打开单个模型的标准浏览器窗口。当块位于工作区中时，双击缩略图打开标准大小的图形。 有关详细信息，请参阅第 55 页码拆分模型浏览器。

一旦将模型创建为分割模型之后，就不能删除其分割处理，也不能从分割建模节点或模型块下游撤销分割。

示例。一家全国销售商希望按照其全国每一家店铺的产品类别估算销售情况。则其通过使用分割建模，将其输入数据的“店铺”字段指定为分割字段，这样能在一次操作中为每个店铺的每个分类构建单独的模型。其然后可以使用所得信息比只使用一个模型更加准确地控制库存水平。

分割和分区

分割与分区共有某些特征，但其使用方式截然不同。

分区将数据集随机分成两部分或三部分：训练、测试和（可选）验证，并用于测试单个模型的性能。

分割将数据集分成与分割字段可能值一样多的部分，并用于构建多个模型。

分区和分割工作方式彼此完全不同。您可以在建模节点中选择一个、两个或一个也不选。

支持拆分模型的建模节点

大量建模节点可创建分割模型。例外的情况是自动聚类、时间序列、主成分分析/因子、特征选择、SLRM、关联模型（Apriori、Carma 和序列）、聚类模型（K-Means、Kohonen、两步和异常）、Statistics 模型以及用于数据库内建模的节点。

支持拆分建模的建模节点是：


C&R 树


贝叶斯网络


QUEST


GenLin


CHAID


KNN


C5.0


Cox


神经网络


自动分类器


决策表


自动数值


回归


Logistic


判别


SVM

受分割影响的特征

使用拆分模型以各种方式影响大量 IBM® SPSS® Modeler 特征。此部分指导在流中使用拆分模型连同其他节点。

记录选项节点

当在包含**样本**节点的流中使用拆分模型，按拆分字段分层记录，以实现记录的平均抽样。当您选择“复杂”作为样本方法时，此选项可用。 有关详细信息，请参阅第 3 章中的聚类和分层设置中的IBM SPSS Modeler 14.2 源、过程和输出节点。

如果流包含**平衡**节点，注意平衡适用于输入记录的整体集合，而非拆分内的记录子集。

当通过**汇总**节点分类汇总记录时，如果您想计算每个拆分的汇总，请将拆分字段设置为关键字段。 有关详细信息，请参阅第 3 章中的汇总节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

字段选项节点

类型节点是指定将那个或哪些字段用作拆分字段的地方。 有关详细信息，请参阅第 4 章中的类型节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

注意，尽管**整体**节点用于结合两个或多个模型块，但是不能用于颠倒拆分操作，因为拆分模型包含在单个模型块内。

建模节点

拆分模型不支持预测变量重要性（估算模型时预测变量输入字段的相对重要性）计算。构建拆分模型时会忽略预测变量重要性设置。

KNN（最近相邻元素）节点只有在设置预测目标字段时，才能支持拆分模型。其他设置（只标识最近相邻元素）不创建模型。如果选择选项“自动选择 k”，每个拆分模型可能有不同数量的最近相邻元素。因此，整体模型生成的列数等于所有拆分模型找到的最近相邻元素的最大数。对于那些最近邻元素数小于此最大值的拆分模型，将有填充 $null 值的相应列数。 有关详细信息，请参阅第 396 页码第 16 章中的KNN 节点。

数据库建模节点

数据库内建模节点不支持拆分模型。

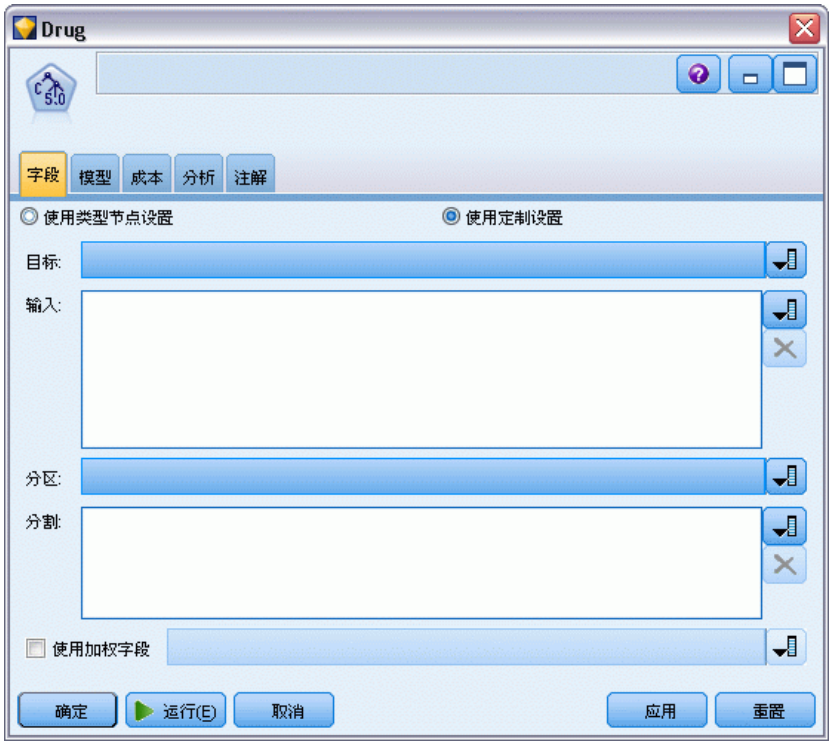
模型块

不可能从拆分模型块**导出到 PMML**，因为块包含多个模型，而 PMML 不支持这种包装。但是可能导出到文本或 HTML。

建模节点字段选项

所有建模节点均有一个“字段”选项卡，在此选项卡中指定的字段将用于构建模型。

图片 3-5
C5.0 建模节点、“字段”选项卡



在构建模型之前，需要指定要将哪些字段用作目标和输入。某些特殊情况下，所有建模节点将采用上游的“类型”节点的字段信息。如果正在使用类型节点选择输入和目标字段，则不必在此选项卡上做任何更改。（特殊情况包括序列节点和“文本抽取”节点，这两个节点需要在建模节点中指定字段设置。）

使用类型节点设置。 该选项通知节点使用来自上游类型节点的字段信息。这是默认值。

使用自定义设置。该选项通知节点使用在此处指定的字段信息，而不是在任何上游类型节点中给出的字段信息。选中此选项后，请根据需要指定下面的字段。

注意：未显示所有节点的所有字段。

- **使用交易格式（仅 Apriori、CARMA、MS 关联规则、ISWAssociation 和 Oracle Apriori 节点）。**如果源数据为**交易格式**，则选中此复选框。此格式的记录具有两个字段，一个为 ID 字段，一个为内容字段。每条记录代表单个交易或单个项，关联项通过相同的 ID 得以链接。如果数据为**表格格式**，则取消选中此复选框，表格格式中项目由独立标志代表，其中每个标志字段代表某个特定项是否存在，且每个记录代表关联项的完整集合。 有关详细信息，请参阅第 320 页码第 12 章中的表格格式数据与事务处理格式数据。

- **ID。**对于事务处理格式的数据，请从列表中选择 ID 字段。数字字段或符号字段可用作 ID 字段。此字段的每个唯一值都应该表明一个特定的分析单元。例如，在市场购物篮的应用中，每个 ID 可能表示一个客户。对于 Web 日志分析应用，每个 ID 可能代表一个计算机（以 IP 地址表示）或一个用户（以登录数据表示）。
- **ID 是连续的。** （仅 Apriori 和 CARMA 节点）如果您的数据进行了预先排序，以便所有 ID 相同的记录在数据流中分组在一起，那么选择此选项可以加快处理速度。如果您的数据未经预先排序（或者您不确定），请将此选项保持未选中状态，则该节点将自动对数据进行排序。

  注意：如果您的数据未经排序而您选择了此选项，则可能会在模型中得到无效结果。
- **内容。**指定模型的内容字段。这些字段包含与关联建模有关的项目。您可以指定多个标志字段（如果数据为表格格式）或者一个名义字段（如果数据为事务格式）。
- **目标。** 对于需要一个或多个目标字段的模型，请选择目标字段或字段。此操作与在“类型”节点中将字段的角色设置为目标类似。
- **评估。**（仅适合自动聚类模型。）不为聚类模型指定目标，但可选择一个评估字段以确定其重要性等级。此外，还可评估聚类区分此字段值的程度，从而指示是否可使用聚类来预测此字段。
- **输入。**选择输入字段或字段。此操作与在“类型”节点中将字段的角色设置为输入类似。
- **分区。**该字段允许您使用指定字段将数据分割为几个不同的样本，分别用于模型构建过程中的训练、测试和验证阶段。通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。如果已使用“类型”或“分区”节点定义了多个分区字段，则必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。（如果仅有一个分区字段，则将在启用分区后自动引入此字段。）有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。同时请注意，要在分析时应用选定分区，同样必须启用节点“模型选项”选项卡中的分区功能。（取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）
- **分割。**对于分割模型，选择分割字段或字段。此操作与在“类型”节点中将字段的角色设置为分割类似。您仅可将测量级别为**标志**、**名义**、**有序**或**连续**的字段指定为分割字段。选为分割字段的字段无法用作目标、输入、分区、频率或权重字段。 有关详细信息，请参阅第 26 页码构建分割模型。
- **使用频率字段。**此选项允许您选择某个字段作为频率权重。如果训练数据中的每条记录代表多个单元（例如，您正在使用聚合的数据），则可采用此项。字段值应是每个记录代表的单位的数量。 有关详细信息，请参阅第 33 页码使用频率和权重字段。

注意：如果您看到错误消息**元数据（在输入/输出字段上）无效**，确保您已经指定了所有必填字段，如“频率”字段。

- **使用权重字段。**此选项允许您选择某个字段作为案例权重。个案权重将作为对输出字段各个水平上方差的差异的一种考量。 有关详细信息，请参阅第 33 页码使用频率和权重字段。
- **结果。** 对于规则归纳节点（Apriori），请选择在最终规则集中用作结果的字段。（这对应于“类型”节点中角色为目标或双向的字段。）
- **条件。** 对于规则归纳节点（Apriori），请选择在最终规则集中用作条件的字段。（这对应于“类型”节点中角色为输入或双向的字段。）

某些模型的“字段”选项卡与本节所述“字段”选项卡不同。

- 有关详细信息，请参阅第 344 页码第 12 章中的序列节点字段选项。
- 有关详细信息，请参阅第 325 页码第 12 章中的CARMA 节点字段选项。

使用频率和权重字段

频率和权重字段用于赋予某些记录高于其他记录的附加重要性，例如，因为您知道一部分人未在训练数据（权重）中表示出来，或者因为一个记录代表多个相同观测值（频率）。

- 频率字段的值应为正整数。频率权重小于或等于零的记录将排除在分析之外。非整数频率权重将四舍五入为最近的整数。
- 安全权重值应为正数但不一定是整数值。案例权重小于或等于零的记录将排除在分析之外。

评分频率和权重字段

频率和权重字段用于训练模型，但不用于评分，因为每个记录的分数基于该记录的特征，而与它代表的观测值个数无关。例如，假设您拥有以下数据：

已婚	已响应
Yes	Yes
Yes	Yes
Yes	Yes
Yes	No
No	Yes
No	No
No	No

基于上表，可以得出这样的结论：四分之三的已婚者对促销作出响应；而三分之二的未婚者对此未作出响应。因此，您可以相应地作出如下的新记录：

已婚	$-已响应	$RP-已响应
Yes	Yes	0.75 (3/4)
No	No	0.67 (2/3)

此外，还可以使用频率字段更加细致地存储训练数据：

已婚	已响应	Frequency
Yes	Yes	3
Yes	No	1
No	Yes	1
No	No	2

因为此表完全代表同一数据集，因此可以构建相同的模型并仅根据婚姻状况预测响应率。如果评分数据中有十位已婚者的记录，则无论这十个人是代表十条独立的记录，还是频率为 10 的一个人，都可预测他们每位的回答均为是。虽然通常情况下权重不是整数，但可以认为它近似表示记录的重要性。这就是对记录进行评分时不使用频率和权重字段的原因。

评估和比较模型

某些模型类型可支持频率字段，某些可支持权重字段，还有一些可同时支持这两种字段。但在使用这两种字段的所有情况中，它们仅用于构建模型，在使用“评估”节点或“分析”节点对模型进行评估时，或者在使用受“自动分类器”节点和“自动数值”节点支持的大部分方法进行模型排序时，均不考虑使用这两种字段。

- 例如，在使用评估图表比较模型时将忽略频率和权重值。因此可以在使用频率和权重字段的模型与不使用这两种字段的模型之间进行级别比较。但同时意味着，若要获得准确评估，必须使用不依赖频率字段或权重字段且能准确代表总体的数据集。在实际应用中，要执行此操作，就要确保使用频率字段值或权重字段值始终为空或 1 的检验样本评估模型。（这种限制仅适用于评估模型；如果训练样本和检验样本的频率值或权重值始终为 1，则首次不必使用这两种字段。）
- 如果使用“自动分类器”基于“利润”对模型进行排序，则可考虑频率，在这种情况下推荐使用此方法。
- 若有必要，可以使用“分区”节点，将数据分割为训练样本和检验样本。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

建模节点分析选项

许多建模节点都包括“分析”选项卡，使用该选项卡可获取预测变量重要性信息以及原始和调整后的倾向得分。

图片 3-6
建模节点中的“分析”选项卡

模型评估

计算预测变量重要性。对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。请注意，对于某些模型，计算预测变量重要性（特别对较大数据集进行操作时）可能需要花较长时间，因此默认情况下，对于某些模型，预测变量重要性均处于关闭状态。预测变量重要性对于决策列表模型不可用。 有关详细信息，请参阅第 45 页码预测变量重要性。

倾向得分

可以在建模节点中和模型块的“设置”选项卡上启用倾向得分。该功能仅在所选目标为标志字段时才可用。有关详细信息，请参阅第 36 页码倾向得分。

计算原始的倾向得分。原始的倾向得分仅从基于训练数据的模型中导出。如果模型预测值为真（将响应），则倾向与 P 相同，其中 P 为预测的可能性。如果模型预测的值为假，则计算出的倾向为 (1 - P)。

- 如果构建模型时选择了此选项，则默认情况下将在模型块中启用倾向得分。不过，无论是否在建模节点中选择了原始倾向得分，都可以始终在模型块中选择启用原始倾向得分。
- 对模型进行评分时，原始倾向得分将被添加到将 RP 字母附加到标准前缀的字段中。例如，如果预测位于名为 $R-churn 的字段中，则倾向得分字段的名称将是 $RRP-churn。

计算调整后的倾向得分。原始倾向仅仅基于由可能过度拟合的模型给定的估计上，这样会导致过于乐观地评估倾向。调整后的倾向尝试通过查看模型在检验或验证分区的性能或通过调整倾向来弥补，以相应地给作出更好的估计。

- 此设置要求流中存在有效的分区字段。有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。
- 与原始置信度分数不同，调整后的倾向得分必须在构建模型时计算；否则，对模型块进行评分时该分数将不存在。
- 对模型进行评分时，在将 AP 字母附加到标准前缀的字段中添加调整后的倾向得分。例如，如果预测位于名为 $R-churn 的字段中，则倾向得分字段的名称将是 $RAP-churn。调整后的倾向得分不适用于 logistic 回归模型。
- 在计算调整后的倾向得分时，必须尚未平衡用于计算的检验或验证分区。为避免这一点，请确保在任何上游平衡节点中选中**仅平衡训练数据**选项。有关详细信息，请参阅第 3 章中的为平衡节点设置选项中的IBM SPSS Modeler 14.2 源、过程和输出节点。此外，如果已在上游获取了复杂样本，则会导致调整后的倾向得分无效。
- 调整后的倾向得分不适用于“增强型”树和规则集模型。有关详细信息，请参阅第 158 页码第 6 章中的增强型 C5.0 模型。

  基于。对于有待计算的调整后的倾向得分，流中必须存在一个分区字段。可以指定是使用检验分区还是验证分区进行此计算。为获取最佳结果，检验或验证分区包含的记录数量应至少与用于训练原始模型的分区所包含的记录数相同。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

倾向得分

对于返回预测为是或否的模型，您除了可以要求标准预测和置信度值以外，还可要求倾向得分。倾向得分指示特定结果或响应的可能性。例如：

表 3-1
倾向得分

客户	要响应的倾向
Joe Smith	35%
Jane Smith	15%

倾向得分仅适用于有标志目标的模型，并且指示为字段定义的值为真的可能性，如在源节点或类型节点中指定的那样。 有关详细信息，请参阅第 4 章中的指定标志的值中的IBM SPSS Modeler 14.2 源、过程和输出节点。

倾向得分与置信度得分

倾向得分与置信度得分不同，无论当前预测是是还是否，置信度得分都能应用于当前预测。例如，当预测为否 时，较高的置信度实际上意味着不 响应的可能性较大。倾向性得分避开此限制，以便轻松比较所有记录。例如，置信度为 0.85 的否 预测将转换为 0.15（或 1 减 0.85）的原始倾向。

表 3-2
置信度得分

客户	预测	置信度
Joe Smith	会响应	.35
Jane Smith	不会响应	.85

获得倾向得分

- 可以在建模节点中的“分析”选项卡或模型块中的“设置”选项卡上启用倾向得分。该功能仅在所选目标为标志字段时才可用。 有关详细信息，请参阅第 34 页码建模节点分析选项。
- 也可以通过整体节点计算倾向得分，具体取决于所用的整体方法。 有关详细信息，请参阅第 4 章中的整体节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

计算调整后的倾向得分

计算调整后的倾向得分将作为构建模型过程的一部分，否则没有可用的调整后的倾向得分。构建模型后，则可使用检验或验证分区中的数据对模型进行评分，同时通过在该分区上分析原始模型的性能，构建一个提供调整后的倾向得分的新模型。根据模型的类型，可以使用两种方法之一来计算调整后的倾向得分。

- 对于规则集模型和树模型，要生成调整后的倾向分数，可通过重新计算每个树节点上每个类别的频率（适用于树模型）或重新计算每个规则的支持和置信度（适用于规则集模型）。这样一来，请求调整后的倾向得分时将使用与原始模型一起存储的新规则集模型或树模型。每次将原始模型应用到新数据时，都会随之将新模型应用到原始倾向分数以生成调整后的分数。
- 对于其他模型，通过对检验或验证分区上的原始模型进行评分而生成的记录将按其原始倾向得分进行分级。接着，对定义非线性函数的神经网络模型进行训练，该函数从每个分级的平均原始倾向中映射到相同分级的平均观测倾向中。正如之前对树模型的说明，得出的神经网络模型将与原始模型一起存储，并且在请求调整后的倾向分数时应用到原始倾向得分。

关于测试分区中缺失值的警告说明。测试/验证分区中缺失输入值的处理方法根据模型而不同（请参阅各个模型评分算法以了解详细信息）。有缺失输入值时，C5 模型无法计算调整倾向。

模型块

图片 3-7
模型块

模型块是模型的容器，其中包含一组规则、公式或方程式，它们代表在 IBM® SPSS® Modeler 中模型构建操作的结果。模型块的主要用途为对数据进行评分以生成预测，或允许进一步分析模型属性。在屏幕上打开模型块后，可以查看有关模型的各类详细信息，例如，在模型创建中输入字段的相对重要性。要查看预测，则需要进一步添加并执行处理或输出节点。 有关详细信息，请参阅第 57 页码使用流中的模型块。

图片 3-8
从建模节点到模型块的模型链接

在成功地执行建模节点后，会在流工作区上放置对应的模型块，并以金色钻石形图标表示（因此称之为“块”）。在流工作区上显示的模型块，带有到位于建模节点之前的最近合适节点的连接（实线），以及到建模节点本身的链接（虚线）。

此外，模型块也放置在位于 SPSS Modeler 窗口右上角的“模型”选项板中。从任一位置均可选中模型块，并浏览模型的详细信息。

在建模节点成功执行后，模型块始终位于“模型”选项板中。可以设置用户选项来控制是否也将模型块置于流工作区上。 有关详细信息，请参阅第 12 章中的设置通知选项中的IBM SPSS Modeler 14.2 用户 指南。

以下主题提供了使用 SPSS Modeler 中模型块的相关信息。要深入了解所用到的算法，请参阅SPSS Modeler算法指南（可从 IBM® SPSS® ModelerDVD 的 \Documentation 文件夹中获取）。

模型链接

默认情况下，在流工作区上显示的模型块带有指向创建它的建模节点的链接。这在具有多个模型块的复杂流中特别有用，它使您能够识别将被每个建模节点更新的模型块。每个链接包含一个指示当建模节点执行时是否替换模型的符号。 有关详细信息，请参阅第 40 页码替换模型。

定义和删除模型链接

您可以在工作区上手动定义和删除模型链接。在定义新的链接后，光标将变成链接光标。

图片 3-9
链接光标

定义新链接（上下文菜单）

- 右键单击要作为链接起点的建模节点。
- 从上下文菜单中选择**定义模型链接**。
- 单击要作为链接终点的模型块。

定义新链接（主菜单）

- 单击要作为链接起点的建模节点。
- 在主菜单中，选择：
  编辑 > 节点 > 定义模型链接
- 单击要作为链接终点的模型块。

删除现有链接（上下文菜单）

- 右键单击位于链接终点的模型块。
- 从上下文菜单中选择**删除模型链接**。

  或者：
- 右键单击位于链接中部的符号。
- 从上下文菜单中选择**删除链接**。

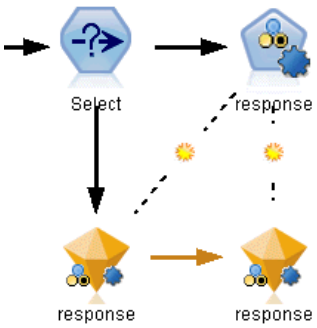
删除现有链接（主菜单）

- 单击要删除其链接的建模节点或模型块。
- 在主菜单中，选择：
  编辑 > 节点 > 删除模型链接

复制和粘贴模型链接

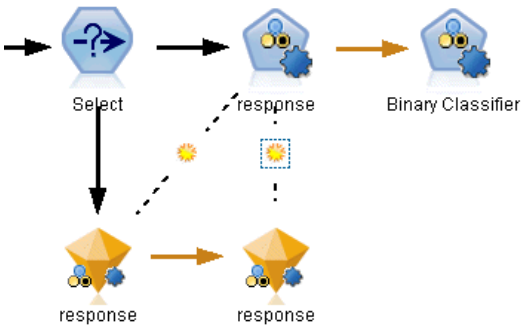
如果复制了带链接的模型块，但未包括其建模节点，则当将其粘贴到同一流中时，粘贴后的模型块将具有到建模节点的链接。新链接具有与原始链接相同的模型替换状态（请参阅替换模型第 40 页码）：

图片 3-10
复制和粘贴带有链接的模型块



如果将模型块连同其链接的建模节点一起复制和粘贴，则不论对象复制到同一流还是新流中，都会保留链接：

图片 3-11
复制和粘贴带有链接的模型块



注意：如果复制了带链接的模型块，但未包括其建模节点，则当将其粘贴到新流中（或不包含建模节点的超节点中）时，链接将被破坏，并且只会粘贴模型块。

模型链接和超节点

如果定义超节点包含链接模型的建模节点或模型块（但未同时包含），链接将被破坏。展开超节点不会恢复链接，只能通过撤销创建超节点来完成此操作。

替换模型

您可以选择在重新执行创建模型块的建模节点时是否替换（即更新）现有模型块。如果关闭替换选项，则重新执行建模节点时将创建新的模型块。

注意：替换模型不同于刷新模型，它是指在方案中更新模型。 有关详细信息，请参阅第 9 章中的模型刷新中的IBM SPSS Modeler 14.2 用户 指南。

每个从建模节点到模型块的链接包含一个指示当建模节点重新执行时是否替换模型的符号。

图片 3-12
模型替换处于打开的模型链接

初始显示链接时，模型替换处于打开，并通过链接中的小旭日形符号指示。在此状态下，重新执行位于链接一端的建模节点就会更新另一端的模型块。

图片 3-13
模型替换处于关闭的模型链接

如果模型替换处于关闭，则链接符号替换为灰色点。在此状态下，重新执行位于链接一端的建模节点会在工作区上新增一个更新后的模型块。

在任一情况下，在“模型”选项卡中是更新现有模型块还是新增模型块，取决于**替换原有模型**系统选项的设置。 有关详细信息，请参阅第 12 章中的设置通知选项中的 IBM SPSS Modeler 14.2 用户 指南。).

执行顺序

当执行具有包含模型块的多个分支的流时，首先对流进行评估，以确保先执行模型替换处于打开的分支，然后再执行使用结果模型块的任何分支。

如果您的需求更为复杂，则可通过脚本手动设置执行顺序。

更改模型替换设置

要更改模型替换设置：

- 右键单击链接上的符号。
- 根据情况选择**打开（关闭）模型替换**。

注意：在模型链接上的模型替换设置将覆盖在“用户选项”对话框的“通知”选项卡上的设置（工具 > 选项 > 用户选项）。

模型选项板

模型选项板（位于管理器窗口“模型”选项卡中）允许您以各种方式使用、检查和修改模型块。

图片 3-14
“模型”选项板

右键单击模型选项板中的模型块，打开带有以下选项的上下文菜单：

图片 3-15
模型块上下文菜单

- **添加到流。**将模型块添加到当前活动流。如果流中存在选定节点，当可以连接时，模型块将连接到选定节点，否则链接到最近的可能节点。如果创建模型的建模节点仍然在流中，则显示的模型块将带有到建模节点的链接。
- **浏览。**打开模型块的模型浏览器。
- **重命名并注解。**允许重命名模型块和/或修改模型块的注解。
- **生成建模节点。**如果要修改或更新某个模型块，但无法使用用于创建该模型的流，则可以使用此选项与创建原始模型相同的选项来重新生成一个建模节点。
- **保存模型，将模型另存为。**将此模型块保存到外部生成模型（.gm）二进制文件。
- **存储模型。**在IBM® SPSS® Collaboration and Deployment Services Repository中保存模型块。 有关详细信息，请参阅第 9 章中的关于 IBM SPSS Collaboration and Deployment Services Repository中的IBM SPSS Modeler 14.2 用户 指南。
- **导出 PMML。**以预测模型标记语言（PMML）格式导出模型块，其可用于 IBM® SPSS® Modeler 之外的新数据评分。**导出 PMML** 对所有生成的模型节点可用。注意：需要 IBM® SPSS® Modeler Server 许可证才能使用此功能。 有关详细信息，请参阅第 12 章中的设置 PMML 导出选项中的IBM SPSS Modeler 14.2 用户 指南。

- **添加到工程。**保存模型块并将其添加到当前工程。在“类别”选项卡上，模型块将添加到“生成的模型”文件夹中。在 CRISP-DM 选项卡上，此节点将被添加到默认工程阶段。（请参阅设置默认工程阶段以获取更改默认工程阶段的相关信息。）
- **删除。**从选项板中删除此模型块。

图片 3-16
模型选项板上下文菜单

右键单击模型选项板中的未占用区域，打开带有以下选项的上下文菜单：

- **打开模型。**载入之前在 SPSS Modeler 中创建的模型块。
- **检索模型。**从 IBM SPSS Collaboration and Deployment Services 存储库检索保存的模型。
- **载入选项板。**从外部文件载入保存的模型选项板。
- **检索选项板。**从 IBM SPSS Collaboration and Deployment Services 存储库检索保存的模型选项板。
- **保存选项板。**将模型选项板的所有内容保存到外部生成模型选项板 (.gen) 文件。
- **存储选项板。**将模型选项板的所有内容保存到 IBM SPSS Collaboration and Deployment Services 存储库中。
- **清除选项板。**从选项板中删除所有模型块。
- **将选项板添加到工程。**保存模型选项板并将其添加到当前工程。在“类别”选项卡上，模型块将添加到“生成的模型”文件夹中。在 CRISP-DM 选项卡上，此节点将被添加到默认工程阶段。
- **导入 PMML。**从外部文件载入模型。可以打开、浏览由 IBM® SPSS® Statistics 或其他支持此格式的应用程序所创建的 PMML 模型并对其进行计分。 有关详细信息，请参阅导入和导出 PMML 模型中的IBM SPSS Modeler 14.2 用户 指南。

浏览模型块

使用模型块浏览器可以检查和使用模型结果。在浏览器中，您可以保存、打印或导出生成模型，检查模型摘要，查看或编辑模型注释。对于某些类型的模型块，还可以生成新的节点，例如“过滤”节点或“规则集”节点。对于某些模型，您还可以查看模型参数，如规则或聚类中心。对于某些类型的模型（基于树的模型和聚类模型），您可以查看其模型结构的图表显示。使用模型块浏览器的控件如下所述。

菜单

“文件”菜单。所有模型块均有一个“文件”菜单，其中包括以下选项的子集：

- **保存节点。**将模型块保存到某个节点 (.nod) 文件。
- **存储节点。**在 IBM SPSS Collaboration and Deployment Services 存储库中保存模型块。
- **页眉和页脚。**允许从模型块打印时对页面的页眉和页脚进行编辑。
- **页面设置。**允许从模型块打印时更改页面设置。
- **打印预览。**显示模型块的打印预览。从子菜单中选择要预览的信息。
- **打印。**打印模型块的内容。从子菜单中选择要打印的信息。
- **打印视图。**打印当前视图或所有视图。
- **导出文本。**将模型块内容导出到某个文本文件。从子菜单中选择要导出的信息。
- **导出 HTML。**将模型块内容导出到 HTML 文件。从子菜单中选择要导出的信息。
- **导出 PMML。**以预测模型标记语言 (PMML) 格式导出模型，导出的文件可在其它 PMML 兼容软件中使用。注意：需要 IBM® SPSS® Modeler Server 许可证才能使用此功能。 有关详细信息，请参阅第 12 章中的设置 PMML 导出选项中的IBM SPSS Modeler 14.2 用户 指南。
- **导出 SQL。**以 SQL（结构化查询语言）格式导出模型，导出的文件可在其它数据库中编辑使用。

  注意：“SQL 导出”仅在以下模型中可用：C5、C&RT、CHAID、QUEST、线性回归、Logistic 回归、神经网络、主成分分析 / 因子以及决策列表模型。 有关详细信息，请参阅第 6 章中的支持 SQL 生成的节点中的IBM SPSS Modeler Server 14.2 管理和性能指南。

“生成”菜单。多数模型块还具有一个“生成”菜单，通过此菜单可以生成基于模型块的新节点。此菜单中的可用选项取决于您所浏览模型的类型。请查看具体的模型块类型，以详细了解您可从特定模型中生成的内容。

“视图”菜单。在模型块的“模型”选项卡上，此菜单允许您显示或隐藏在当前模式下可用的各类直观表示工具栏。要使全部工具栏可用，可从“常规”工具栏中选择“编辑模式”（画笔图标）。

预览按钮。某些模型块具有“预览”按钮，允许您查看模型数据的样本，包括由建模过程创建的额外字段。默认显示的行数为 10，不过可以在流属性中更改此值。 有关详细信息，请参阅第 5 章中的设置流选项中的IBM SPSS Modeler 14.2 用户 指南。

“添加到当前工程”按钮。保存模型块并将其添加到当前工程。在“类别”选项卡上，模型块将添加到“生成的模型”文件夹中。在 CRISP-DM 选项卡上，此节点将被添加到默认工程阶段。（请参阅设置默认工程阶段以获取更改默认工程阶段的相关信息。）

模型块概要/信息

模型块的“概要”选项卡或“信息”视图显示了关于字段、构建设置和模型评估过程的信息。结果以树状视图显示，通过单击指定项可以扩展或合并树状视图。

分析。显示模型相关信息。具体详细信息因模型类型而异，这些信息可在每种模型块的相应章节中找到。此外，如果已执行附加到该建模节点的分析节点，则还会在此部分显示该分析中的信息。 有关详细信息，请参阅第 6 章中的分析节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

字段。 列出构建模型时用作目标和输入的字段。对于分割模型，也列出确定分割的字段。

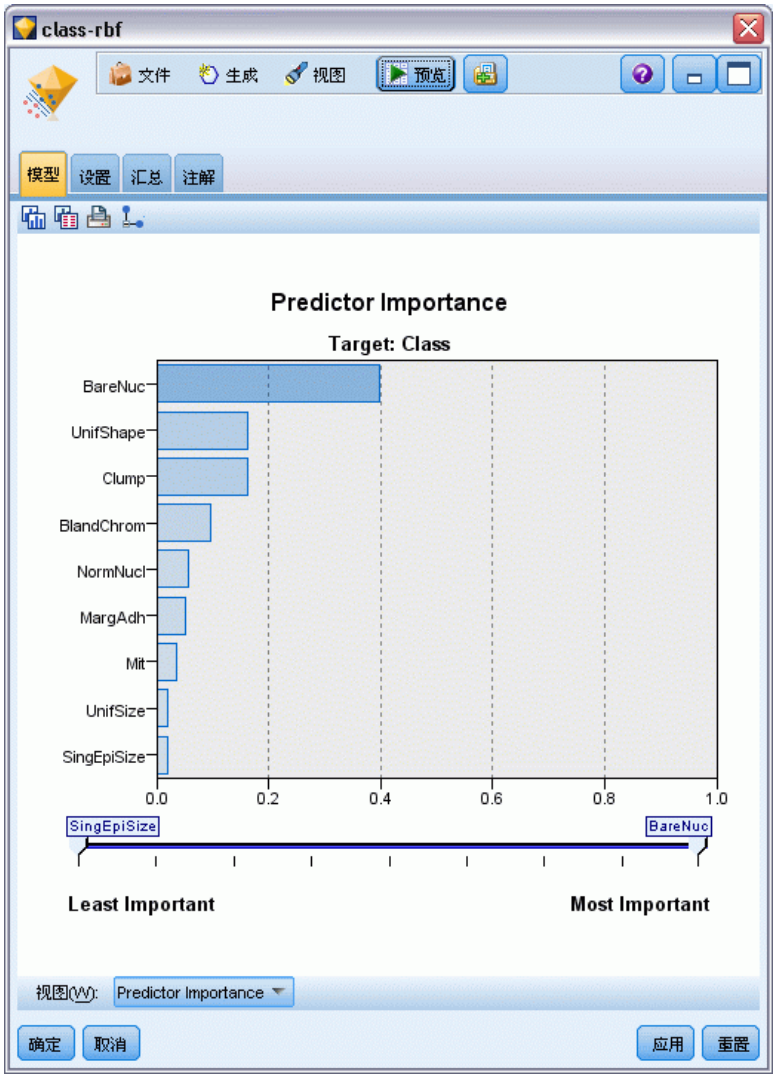
构建设置/选项。 包含有关在构建模型中使用的设置的信息。

训练概要。 显示模型类型、用于创建模型的流、模型创建者、模型构建完成时间和模型构建所用时间。

预测变量重要性

通常，您将需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的变量。预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。由于它们是相对值，因此显示的所有预测变量的值总和为1.0。预测变量的重要性与模型精确性无关。它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

图片 3-17
预测变量重要性图表



预测变量重要性对于可生成相应重要性统计标准的模型可用，包括神经网络模型、决策树（C&R 树、C5.0、CHAID 和 QUEST）、贝叶斯网络模型、判别式模型、SVM 和 SLRM 模型、线性和 logistic 回归模型、广义线性模型以及最近相邻元素 (KNN) 模型。对于这些模型中的大部分而言，可以在建模节点的“分析”选项卡上启用预测变量重要性。有关详细信息，请参阅第 34 页码建模节点分析选项。 有关 KNN 模型，请参阅相邻元素第 400 页码。

注意：拆分模型不支持预测变量重要性。构建拆分模型时会忽略预测变量重要性设置。有关详细信息，请参阅第 26 页码构建分割模型。

计算预测变量重要性所用的时间远远大于建构模型的用时，特别当使用较大数据集时。对于 SVM 和 logistic 回归模型，计算变量重要性的用时比对其他模型执行此操作的用时都要长，所以默认情况下这两种模型均禁用此功能。使用一个包含许多预测变量的数据集时，使用“特征选择”节点进行初始筛选可以较快地生成结果（请参阅以下内容）。

- 如果适用，可以从检验分区计算出预测变量重要性。否则，就使用训练数据。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。
- 预测变量重要性也适用于 SLRM 模型，但需要使用 SLRM 算法进行计算。 有关详细信息，请参阅第 385 页码第 14 章中的SLRM 模型块。
- 可以使用 IBM® SPSS® Modeler 的图表工具进行交互、编辑，并保存图表。
- 还可以根据预测变量重要性图表中的信息生成“过滤”节点。 有关详细信息，请参阅第 47 页码基于重要性过滤变量。

预测变量重要性和特征选择

在某些情况下，模型块中显示的预测变量重要性图表可能似乎给出与“特征选择”节点相似的结果。当特征选择基于每个输入字段与特定目标（与其他输入无关）的关系强度对输入字段进行排序时，预测变量重要性图表将显示此特定模型中各个输入的相对重要性。因此，在筛选输入时使用特征选择可能较为保守。例如，如果工作职务和工作类别与薪资的关系强度相同，特征选择就会指示这两者都很重要。但在建模时，还需考虑交互性和相关性。这样，当两个输入的大部分信息都相同时，您可能会发现仅使用了两个输入之一。在实际应用中，特征选择对预筛选最有用，特别是处理包含大量变量的较大数据集时，而预测变量重要性在微调模型时更为有用。

基于重要性过滤变量

还可以根据预测变量重要性图表中的信息生成“过滤”节点。

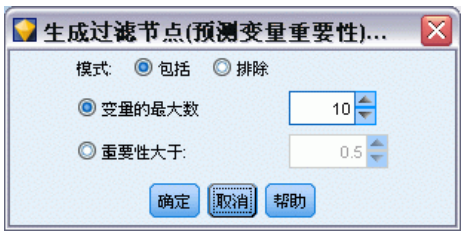
标记要包括在图表上的预测变量（若适用），然后从菜单中选择：

生成 > 过滤节点（预测变量重要性）

或

> 字段选择（预测变量重要性）

图片 3-18
基于重要性过滤预测变量



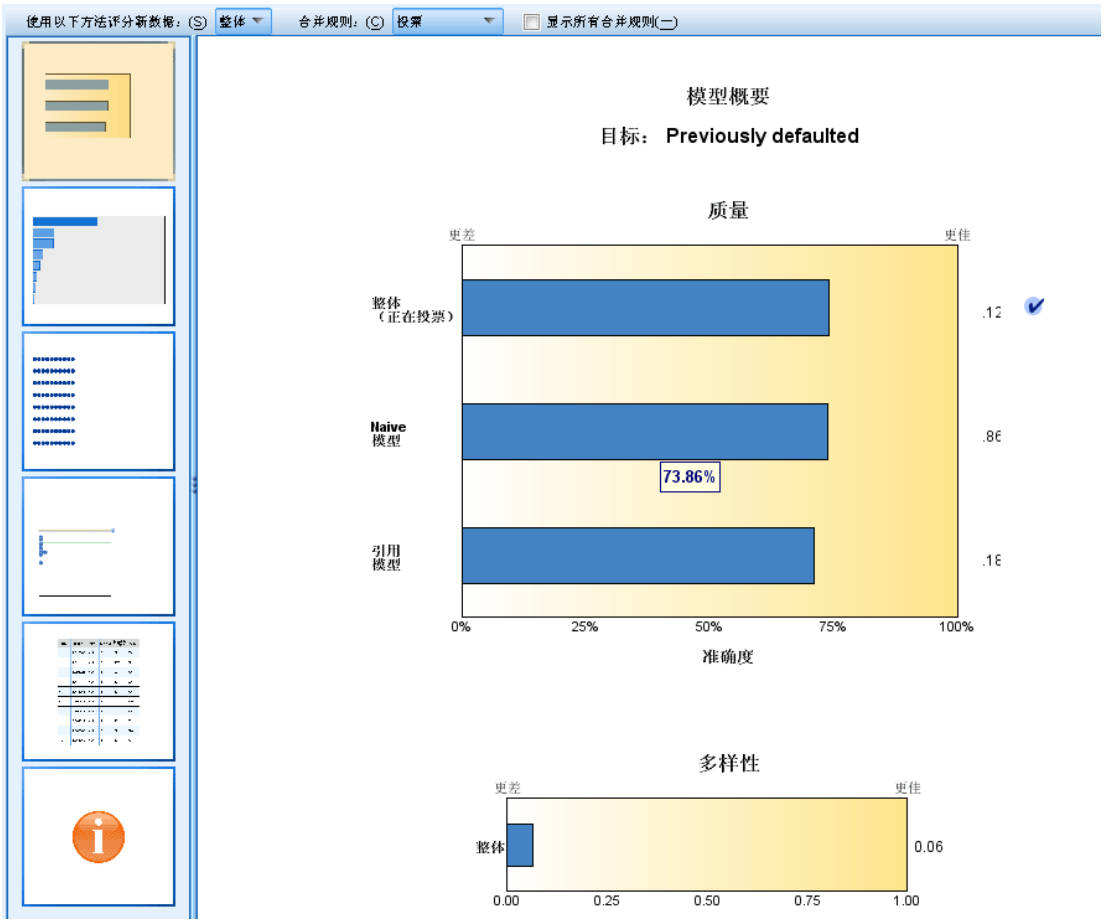
变量的最大数。包括或排除等于指定数字的最重要预测变量。

重要性大于。包括或排除所有相对重要性高于指定值的预测变量。

整体模型

整体模型提供了有关整体中的组件模型和整体性能的信息。

图片 3-19
模型摘要视图



主（独立视图）工具栏允许您选择使用整体或参考模型来进行评分。如果使用整体进行评分，您还可以选择组合规则。这些更改不需要重新执行模型；但是，这些选择将保存到模型（块）以供评分和/或下游模型评估。它们也会影响从整体查看器导出的 PMML。

组合规则。在对整体评分时，此规则用于组合来自基本模型的预测值，以计算整体得分值。

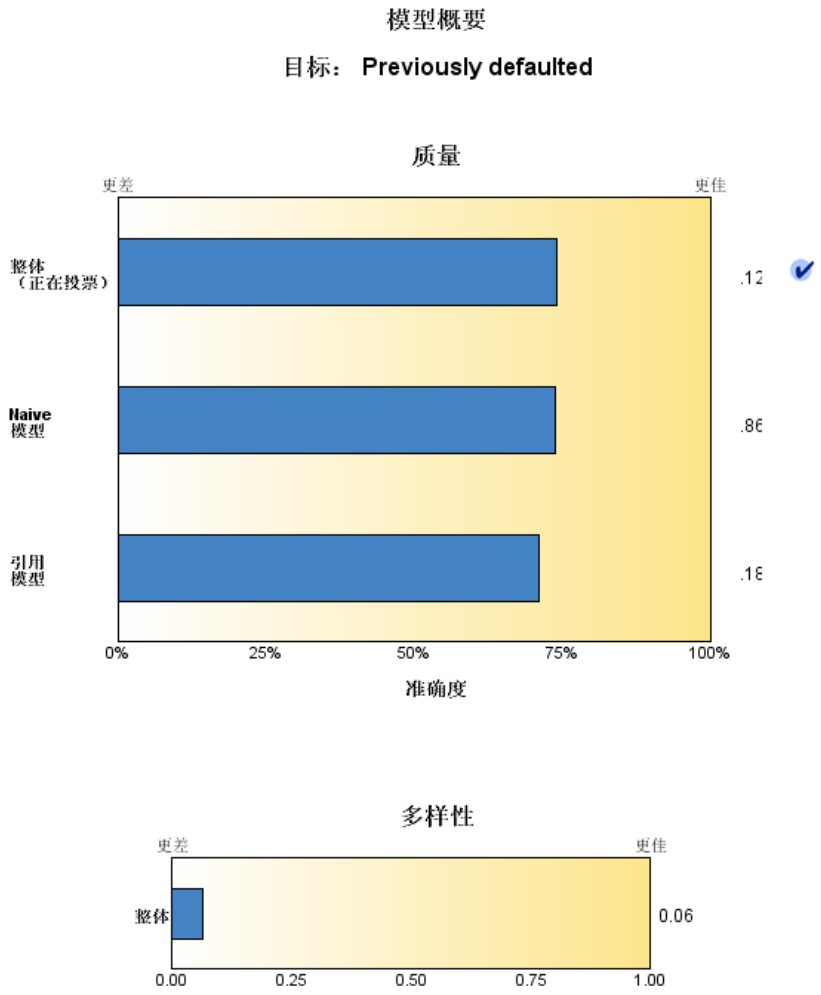
- 可以使用投票、最高概率或最高平均概率来组合**分类**目标的整体预测值。**投票**选择在基本模型中最常具有最高概率的类别。**最高概率**选择在所有基本模型中达到单一最高概率的类别。**最高平均概率**选择当类别概率在基本模型中取平均值时具有最高值的类别。
- 可以通过对来自基本模型的预测值取平均值或中位数，对**连续**目标的整体预测值进行组合。

默认值取自在建模过程中生成的指定。更改组合规则会重新计算模型精确性并更新模型精确性的所有视图。也会更新预测变量重要性图表。如果选择参考模型用于评分，则此控件将被禁用。

显示所有组合规则。 选择该选项时，所有可用组合规则的结果将显示在模型质量图表中。组件模型精确性图表也将更新以显示每种投票方式的参考线。

模型摘要

图片 3-20
模型摘要视图



“模型摘要”视图是整体质量和差异性的快照摘要。

质量。该图表显示与参考模型和 navie 模型相比较的最终模型精确性。精确性越大，模型越好的格式；“最佳”模型将具有最高精确性。对于分类目标，精确性就是预测值与观测值匹配的记录百分比。对于连续目标，精确性为 1 减去预测中的平均绝对误差（预测值的绝对平均值减去观测值）与预测值范围（最大预测值减去最小预测值）的比率。

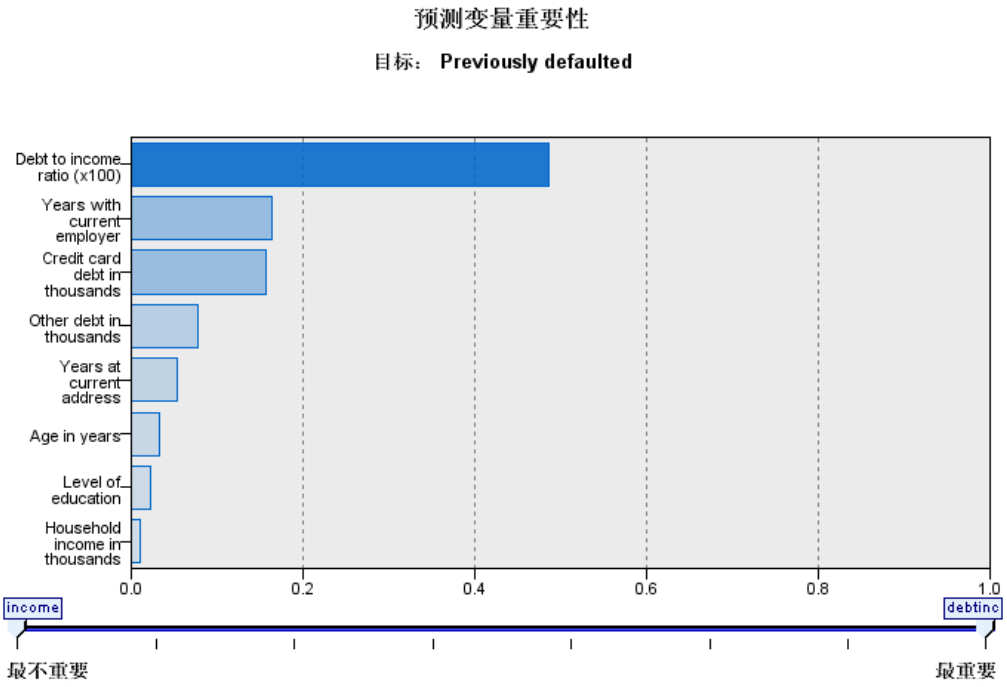
对于 bagging 整体，参考模型是构建在整个培训分区上的标准模型。对于 boosted 整体，参考模型是第一个组件模型。

如果未构建模型，则由 Naive 模型代表精确性，并将所有记录分配给模态类别。不会为连续目标计算 Naive 模型。

差异性。该图表显示用于构建整体的组件模型间的“观点差异性”，以越大则差异性越大格式表示。这是一种基本模型间预测差异程度的测量。差异性对 boosted 整体模型不可用，同时也不会对连续目标显示。

预测变量重要性

图片 3-21
预测变量重要性视图

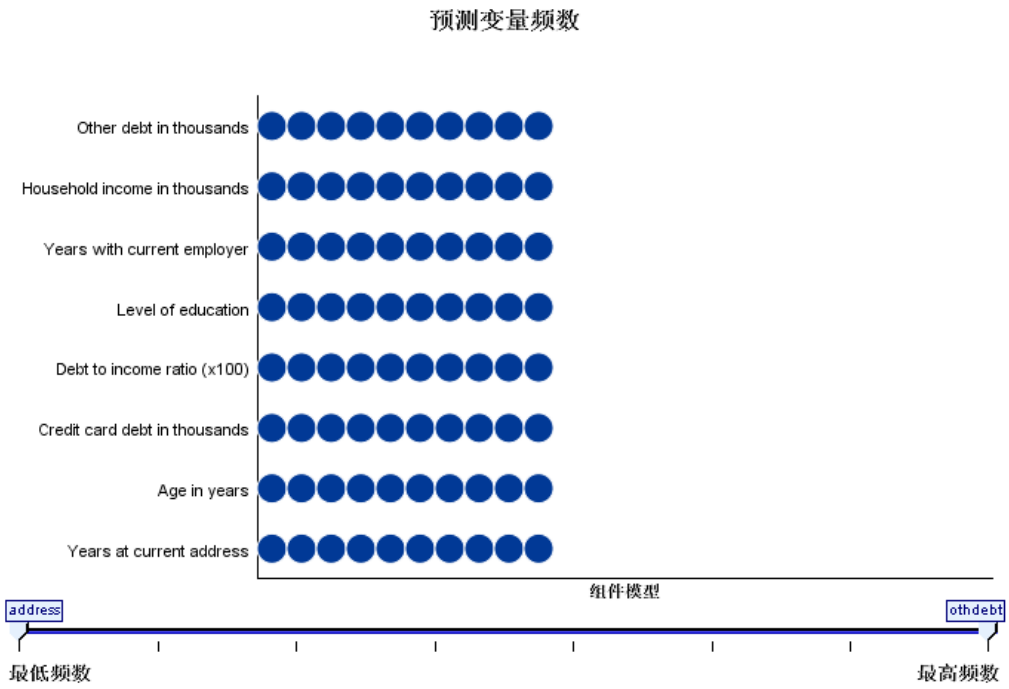


通常，您将需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的变量。预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。预测变量的重要性与模型精确性无关。它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

预测变量重要性对所有整体模型均不可用。预测变量集在组件模型之间可能会有所不同，但可以为至少在一个组件模型中使用的预测变量计算重要性。

预测变量频率

图片 3-22
预测变量频率视图

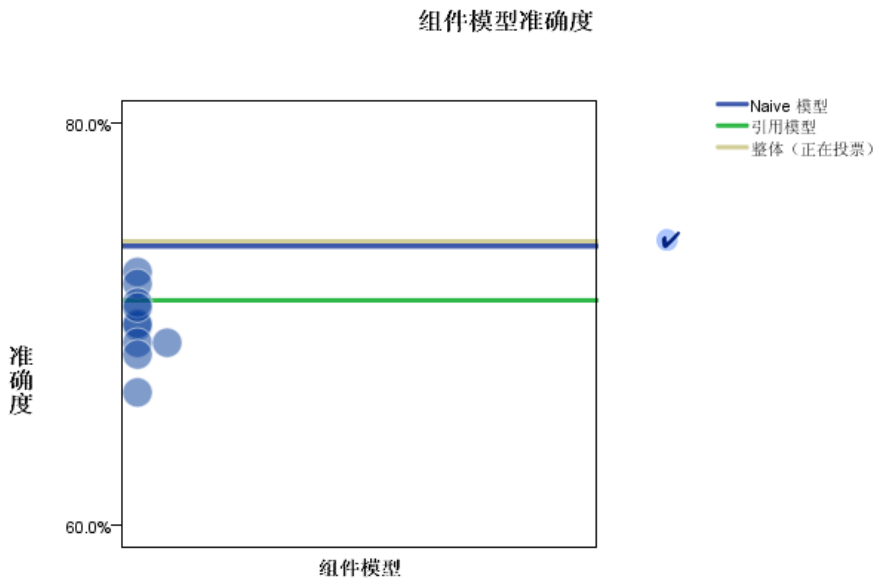


由于选择的建模方法或预测变量选择不同，预测变量集在组件模型间也可能不同。预测变量频率图是一个点图，显示了预测变量在整体组件模型中的分布。每个点代表包含预测变量的一个或多个组件模型。预测变量绘制在 y 轴上，并以频率的降序排序；因此，最顶端的是在最多组件模型中使用的预测变量，而最低端的是在最少组件模型中使用的预测变量。将显示排在前 10 位的预测变量。

出现频率最高的预测变量通常是最重要的。此图对于使预测变量集在组件模型间保持一致的方法没用。

组件模型精确性

图片 3-23
组件模型精确性视图



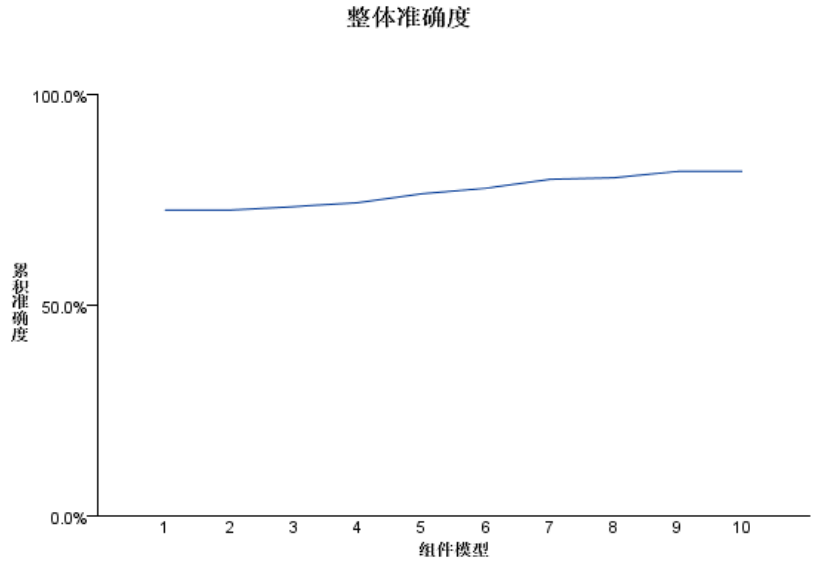
该图表是组件模型预测精确性的点图。每个点代表在 y 轴上绘制了精确性水平的一个或多个组件模型。悬停在任意点上可获得对应的单独组件模型的信息。

参考线。该图显示整体的颜色编码线以及参考模型和 naïve 模型。对应于要用于评分的模型的线的旁边会显示一个复选标记。

互动。该图表会在您更改组合规则时更新。

Boosted 整体。为 boosted 整体显示一个线图。

图片 3-24
整体精确性视图，boosted 整体



组件模型详细信息

图片 3-25
组件模型详细信息视图

组件模型详细信息

模型	准确度	方法	预测变量	模型大小（突触）	记录
1	72.6%		8	77	700
2	72.0%		8	107	692
3	69.9%		8	92	708
4	70.0%		8	107	685
5	71.1%		8	107	706
6	69.1%		8	92	690
7	69.1%		8	92	696
8	70.8%		8	122	703
9	66.6%		8	62	726
10	68.5%		8	107	701

该表显示关于组件模型的信息，按行列出。默认情况下，组件模型按模型编号的升序排序。您可以按任意列的值对这些行进行升序或降序排序。

模型。代表组件模型创建顺序的数字。

精确性。百分比形式的整体精确性。

方法。建模方法。

预测变量。组件模型中使用的预测变量数。

模型大小。模型大小取决于建模方法：对于树模型，它是树中节点的数量；对于线性模型，它是系数的数量；对于神经网络，它是神经元的数量。

记录。训练样本中输入记录的加权数量。

自动数据准备

图片 3-26
自动数据准备视图

此视图显示在自动数据准备（ADP）步骤中排除了哪些字段，以及转换字段的派生方式等信息。对于每个转换或排除字段，在此表中列出了字段名、在分析中的角色，以及 ADP 步骤所采取的操作。这些字段按其名称的字母升序排列。

操作 `Trim outliers`（如果显示）表示位于截断值（平均值的 3 个标准差）之外的连续预测变量值被设为截断值。

分割模型的模型块

分割模型的模型块可以访问分割创建的所有单独模型。

分割模型块包含：

- 创建的所有拆分模型列表，连同每个模型的统计量集合
- 有关整体模型的信息

从拆分模型列表中，您可以打开单个模型以进一步检查。

拆分模型浏览器

“模型”选项卡列出块中包含的所有模型，以各种形式提供有关拆分模型的统计量。它有以下两种一般形式，具体取决于建模节点。

图片 3-27
拆分模型浏览器

排序方式。使用此列表选择列出模型的顺序。您可以根据任何显示列的值将列表按升序或降序排序。或者，单击列标题，按该列将列表排序。默认是总精确性的降序。

显示/隐藏列菜单。单击此按钮，以显示菜单，以便选择单个列以显示或隐藏。

查看。如果您正在使用分区，您可以选择查看培训数据或测试数据的结果。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

对于每个拆分，详细信息显示如下：

图形。指示此模型数据分布的缩略图。当块位于工作区中时，双击缩略图打开标准大小的图形。

模型。模型类型图标。双击图标打开此特定分割的模型块。

分割字段。建模节点中指定为分割字段的字段及其各个可能值。

分割中记录数。此特定分割中涉及的记录数。

使用的字段编号。基于所用输入字段的数量排序拆分模型。

总准确度 (%)。在该拆分中与记录总数有关的拆分模型正确预测出的记录百分比。

图片 3-28
拆分模型查看器

分割组

ed	准确度	模型大小（突触）	记录
Did not complete high school	68.3%	62	372
High school degree	66.8%	42	198
Some college	72.3%	42	87
College degree	59.2%	12	38
Post-undergraduate degree	.	.	.

不能为一个或多个分割组构件模型。

拆分。列标题显示用于创建拆分的字段，而单元格则是拆分值。双击任意拆分打开“模型查看器”以用于该拆分的建模。

精确性。百分比形式的整体精确性。

模型大小。模型大小取决于建模方法：对于树模型，它是树中节点的数量；对于线性模型，它是系数的数量；对于神经网络，它是神经元的数量。

记录。训练样本中输入记录的加权数量。

使用流中的模型块

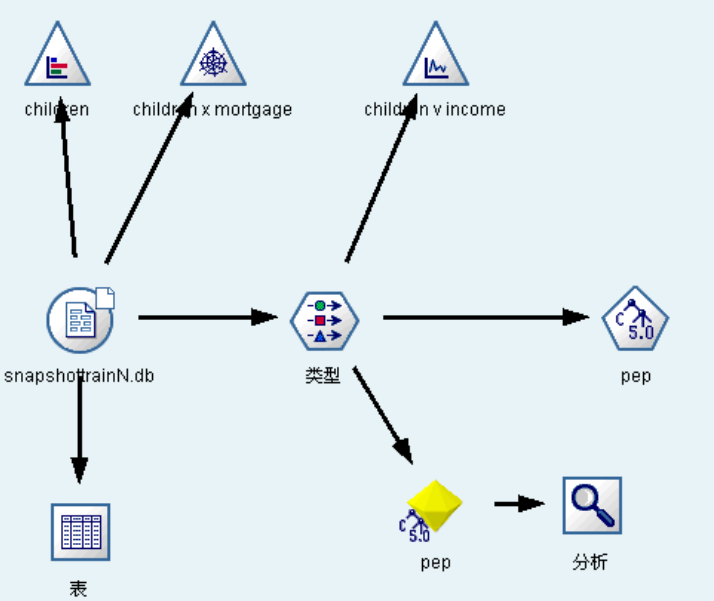
模型块置于流中，允许您对新数据进行评分并生成新节点。**计分**数据允许使用从模型构建中获得的数据来创建新记录的预测。要查看评分结果，需要为模型块添加终端节点（即处理或输出节点）并执行终端节点。

对于某些模型而言，还可从模型块中获得有关预测质量的其他信息，例如置信度值或到聚类中心的距离。生成新节点时允许基于生成模型的结构来方便地创建新节点。例如，执行输入字段选择的多数模型允许生成“过滤”节点，此过滤节点仅传递模型所标识的重要输入字段。

使用模型块对数据进行评分

- 将模型块连接到向其传递数据的数据源或流。

图片 3-29
使用模型块评分



- 将一个或多个处理或输出节点（如“表”或“分析”节点）添加或连接到模型块。
- 执行模型块中的某个下游节点。

注意：不能使用“非精练规则”节点对数据进行计分。要对基于关联规则模型的数据进行计分，请使用“非精练规则”节点来生成“规则集”模型块，并使用“规则集”模型块进行计分。 有关详细信息，请参阅第 338 页码第 12 章中的从关联模型块生成规则集。

使用模型块生成处理节点

- 在此选项板中浏览模型，或者在流工作区中编辑模型。
- 在“模型块浏览器”窗口的“生成”菜单中选择所需节点类型。可用选项将因模型块类型的不同而有所不同。请查看具体的模型块类型，以详细了解您可从特定模型中生成的内容。

重新生成建模节点

如果要修改或更新某个模型块，但无法使用用于创建该模型的流，则可以使用与创建原始模型相同的选项来重新生成一个建模节点。

- 要重新构建模型，右键单击模型选项板中的模型，然后选择**生成建模节点**。
- 此外，当浏览模型时，请选择“生成”菜单中的**生成建模节点**。

多数情况下，重新生成的建模节点应与创建原始模型的建模节点在功能上一致。

- 对决策树模型而言，还可以将交互式会话过程中的其它设置存储到节点，重新生成建模节点的过程中将启用**使用树型指令**选项。

- 对于决策列表模型而言，将启用**使用保存的交互会话信息**选项。 有关详细信息，请参阅第 192 页码第 9 章中的决策列表模型选项。
- 对于时间序列模型而言，将启用**使用现有模型继续评估**选项，使用该选项可以重新生成包含当前数据的原有模型。 有关详细信息，请参阅第 364 页码第 13 章中的时间序列模型选项。

导入和导出 PMML 模型

PMML（也称为预测模型标记语言）是一种 XML 格式，用于描述数据挖掘和统计模型，包括模型的输入、用于为数据挖掘准备数据的变换，以及定义模型自身的参数。IBM® SPSS® Modeler 可导入和导出 PMML，使得其能够与其他支持此格式的应用程序（如 IBM® SPSS® Statistics）共享模型。

注意：需要 IBM® SPSS® Modeler Server 许可证才能导出 PMML。

有关 PMML 的详细信息，请参阅数据挖掘组网站 (http://www.dmg.org)。

导出模型

PMML 导出支持大多数模型类型，这些模型类型生成在 SPSS Modeler 中。 有关详细信息，请参阅支持 PMML 的模型类型中的IBM SPSS Modeler 14.2 用户 指南。

▶ 右键单击模型调色板上的模型块。（或者，双击工作区上的模型块并选择“文件”菜单。）

▶ 在菜单上，单击**导出 PMML**。

图片 3-30
以 PMML 格式导出模型

▶ 在“导出”（或“保存”）对话框中，指定此模型的目标目录及唯一名称。

注意：可在“用户选项”对话框中为 PMML 导出更改选项。在主菜单中，单击：**工具 > 选项 > 用户选项**

然后单击 PMML 选项卡。

有关详细信息，请参阅第 12 章中的设置 PMML 导出选项中的IBM SPSS Modeler 14.2 用户 指南。

导入以 PMML 格式保存的模型

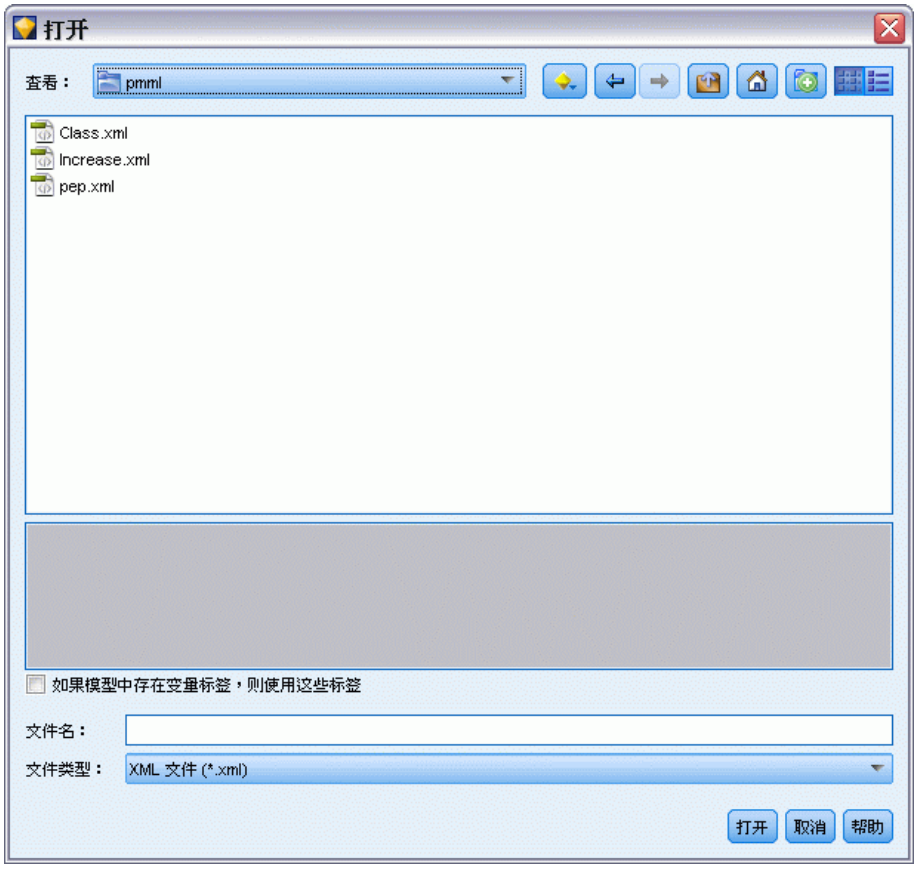
以 PMML 格式从 SPSS Modeler 或其他应用程序中导出的模型可以导入到模型调色板中。有关详细信息，请参阅支持 PMML 的模型类型中的IBM SPSS Modeler 14.2 用户 指南。

- 在模型调色板上，右键单击调色板并从菜单中选择**导入 PMML**。

图片 3-31
以 PMML 格式导入模型

- 选择要导入的文件并按需要为变量标签指定选项。
- 单击**打开**。

图片 3-32
为使用 PMML 格式保存的模型选择 XML 文件



如果模型中存在变量标签，则使用这些标签。 PMML 可为数据字典中的变量同时指定变量名和变量标签（例如 Referrer ID，简称 RefID）。如果在最初导出的 PMML 中存在变量标签，则选中此选项可以使用这些变量标签。

如果已选中变量标签选项但在 PMML 中没有变量标签，则按常规使用变量名。

支持 PMML 的模型类型

PMML 导出

SPSS Modeler 模型。在 IBM® SPSS® Modeler 中创建的下列模型都可导出为 PMML 4.0 格式：

- C&R 树
- QUEST
- CHAID
- 线性回归
- 神经网络

- C5.0
- Logistic 回归
- Genlin
- SVM
- 贝叶斯网络
- Apriori
- Carma
- 序列
- K-Means
- Kohonen
- 两步
- KNN
- Statistics模型

在 SPSS Modeler 中创建的下列模型都可导出为 PMML 3.2 格式：

- 决策表

数据库本地模型。 对于使用数据库本地算法生成的模型，仅在 IBM InfoSphere Warehouse 模型中可使用 PMML 导出。无法导出使用 Microsoft 的 Analysis Services 或 Oracle Data Miner 创建的模型。另外请注意，以 PMML 格式导出的 IBM 模型无法重新导入到 SPSS Modeler 中。 有关详细信息，请参阅第 2 章中的数据库建模概述中的IBM SPSS Modeler 14.2 数据库内数据挖掘指南。

PMML 导入

SPSS Modeler 可以导入并评分由所有 IBM® SPSS® Statistics 产品的当前版本生成的 PMML 模型，包括从 SPSS Modeler 导出的模型和由 SPSS Statistics 17.0 或以后版本生成的模型或转换 PMML。这实质上意味着评分引擎可评分的任何 PMML，以下除外：

- 无法导入 Apriori、CARMA 及异常检测模型。
- 将 PMML 模型导入到 SPSS Modeler 中后，虽然可以对其进行评分，但不能进行浏览。（注意，其中包括最初从 SPSS Modeler 中导出的模型。为避免此限制，可将模型按生成的模型文件 (*.gm) 导出而不是按 PMML 导出。）
- 以 PMML 格式导出的 IBM InfoSphere Warehouse 模型无法再导入。
- 在导入时会执行有限的验证，但在试图对模型评分时会执行全面验证。因此有可能导入成功，但评分却失败或产生不正确的结果。

非精练模型

非精炼 模型包含从数据中抽出的信息，但并不用于直接生成预测。即这些模型不能添加到流。非精练的模型在生成模型选项板上显示为“未打磨的钻石”。

图片 3-33
非精练模型的图标

要查看非精练规则模型的详细信息，右键单击模型，然后选择上下文菜单中的浏览。像其他在 IBM® SPSS® Modeler 中生成的模型一样，各种选项卡将提供所创建模型的相关概要和规则信息。

生成节点。“生成“菜单允许您基于规则创建新节点。

- **选择节点。**生成选择节点来选择当前选定规则所要应用的记录。如果未选择任何规则，此选项则禁用。
- **规则集。**生成规则集节点预测单个目标字段的值。 有关详细信息，请参阅第 338 页码第 12 章中的从关联模型块生成规则集。

章 4

筛选模型

筛选字段和记录

分析的预备阶段中可以使用多个建模节点来查找对建模最有用的字段和记录。可使用特征选择节点来按照重要性筛选字段并为之排序，以及使用异常检测节点来查找不符合“正常”数据已知模式的异常记录。

“特征选择”节点会根据某组条件（例如缺失值百分比）筛选可删除的输入字段；对于保留的输入，将相对于指定目标对其重要性进行排序。例如，假如某个给定数据集有上千个潜在输入，那么哪些输入最有可能用于对患者结果进行建模呢？有关详细信息，请参阅第 64 页码特征选择节点。

“异常检测”节点确定不符合“正常”数据格式的异常观测值（离群值）。即使离群值不匹配任何已知格式或用户不清楚自己的查找对象，也可以使用此节点来确定离群值。 有关详细信息，请参阅第 70 页码异常检测节点。

注意：异常检测并不考虑任何特定的目标（相关）字段，也不考虑这些字段是否与正在预测的模式相关，只是通过基于模型中所选字段集的聚类分析确定异常记录或观测值。由于上述原因，您可能想将异常检测与特征选择或字段筛选和排序的其他方法结合使用。例如，您可以使用特征选择来确定与某个特定目标相关的最重要的字段，然后使用异常检测寻找针对这些字段而言最异常的记录。（另外一个方法是构建一个决策树模型，然后将所有错误分类的记录视为可能的异常进行检查。但是此方法很难用于进行大批量的复制和自动化。）

特征选择节点

数据挖掘问题可能包括成百甚至上千个可用作输入的备选字段。从而花费大量的时间和精力来检查模型究竟应该包含哪些字段或变量。为了缩小选择范围，可以使用特征选择算法来识别对某给定分析最为重要的字段。例如，如果你试着根据多种因素来预测患者结果，那么哪些因素最为重要呢？

特征选择由以下三个步骤组成：

- **筛选。**删除不重要或有问题的输入、记录或个案（例如输入字段含有过多缺失值，或者输入字段的变异太大或太少而变得无用）。
- **秩。**对剩余输入进行排序并根据重要性进行分级。
- **选择。**识别在后续模型中使用的功能子集，例如通过仅保留最重要的输入，过滤或排除所有其它输入。

当下，许多组织的数据均已超载，因此简化和加快建模过程是特征选择的根本优势。通过将注意力迅速集中到最重要的字段上，可以降低所需的计算量，并且可以方便地找到因某种原因被忽略的小而重要的关系，最终获得更简单、精确和易于解释的模型。通过减少模型中的字段数量，可以减少评分时间以及未来迭代中所收集的数据量。

减少字段数量特别有利于 Logistic 回归这样的模型（字段数量限制在 350 个）。

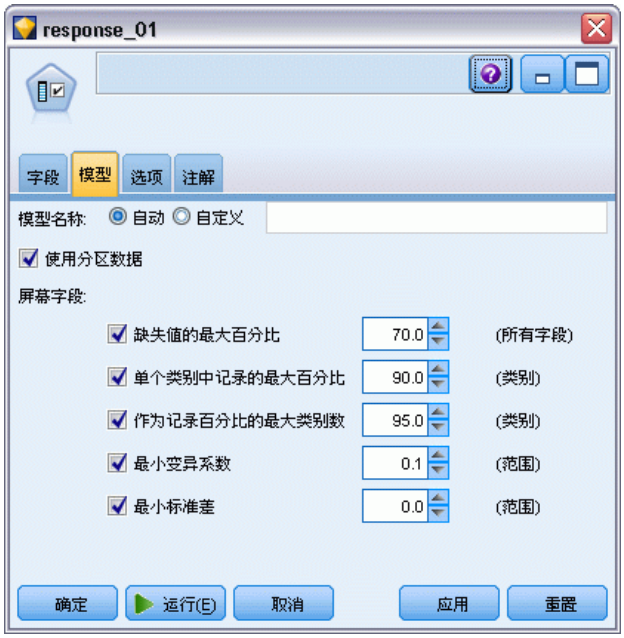
示例。 有个电话公司拥有一个数据仓库，该数据仓库包含 5000 名公司客户对某次促销活动的响应的信息。数据包含有客户年龄、职业、收入、电话使用情况的统计数据等大量数据。三个目标字段表示客户是否对三个报价做出响应。该公司想利用这些数据来预测哪些客户最可能在将来对类似报价做出响应。有关详细信息，请参阅第 10 章中的筛选预测变量（特征选择）中的IBM SPSS Modeler 14.2 应用程序 指南。

要求。单个目标字段（其角色设置为目标），以及要根据目标进行筛选或排序的多个输入字段。目标和输入字段均具有连续（数值范围）或分类的测量级别。

特征选择模型设置

“模型”选项卡上的设置含有标准模型选项以及用于调整输入字段筛选条件的设置选项。

图片 4-1
特征选择的“模型”选项卡



模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

筛选输入字段

筛选就是剔除不提供关于输入/目标关系的任何有用信息的输入或观测值。筛选选项只依据在问题中使用字段的属性，而不考虑该字段针对于选定目标字段的预测能力。被筛选出来的字段将不参与有关输入排序的计算，同时还可选择将这些字段过滤掉，或是从用于建模的数据中删除。

可根据以下标准筛选字段：

- **缺失值的最大百分比。**筛选具有过多缺失值的字段，用占记录总数的百分比来表示。缺失值百分比大的字段几乎不提供任何预测信息。

- **单个类别中的记录最大百分比** 筛选相对于记录总数而言同个类别中具有过多记录的字段。例如，如果数据库中 95% 的客户开同一类型的车，则此信息无助于区分客户。任何超过指定最大值的字段都将被筛选掉。此选项仅适用于分类字段。
- **表示为记录百分比的最大类别数。** 筛选掉相对于记录总数而言具有过多类别的字段。如果很高百分比的类别只含有一个观测值，则该字段用处有限。例如，如果每名客户都戴不同的帽子，则此信息在建立行为模式模型时就不太可能有用。此选项仅适用于分类字段。
- **最小变异系数**筛选变异系数小于或等于指定最小值的字段。此度量值是输入字段标准偏差与输入字段均值之间的比值。如果此值接近 0，则变量值的变异性就不高。此选项仅适用于连续（数字范围）字段。
- **最小标准差。**筛选标准差小于或等于指定最小值的字段。此选项仅适用于连续（数字范围）字段。

带有缺失数据的记录。目标字段具有缺失值或所有输入都具有缺失值的记录或观测值将被从用于排序的计算式中排除。

特征选择选项

“选项”选项卡用于指定在模型块中选择或排除输入字段的默认设置。然后可以将模型添加到流，以选择用于后续模型构建的字段子集。或者，也可以通过在生成模型后在模型浏览器中选择或弃选其他字段，以覆盖这些设置。但是，默认设置下，无需更多修改即可应用模型块，这点在脚本编写方面特别有用。

有关详细信息，请参阅第 68 页码特征选择模型结果。

图片 4-2
特征选择的“选项”选项卡

可用选项有：

所有已排序字段。 根据字段的重要、一般 或 不重要的排序等级来选择字段。可编辑每项排序的标签及用于指派记录的排序等级的截断值。

前几个字段。 请根据重要性选择前 n 个字段。

重要性大于。 请选择重要性大于指定值的所有字段。

不管如何选择，目标字段总是被保留。

重要性排序选项

均为分类字段。所有输入和目标均为分类字段时，可以依据以下任何一个度量来排列重要性顺序：

- **Pearson 卡方。**无需现有关系的强度或方向即可检验目标和输入的独立性。
- **似然比卡方。**与 Pearson 卡方类似，也用于检验目标 - 输入的独立性。
- **Cramer’s V。** 关联的基于 Pearson 卡方统计量的度量。值范围为 0 到 1，0 表示无关联，1 表示完全关联。
- **Lambda。**一个关联的度量，反映变量用于预测目标值时错误减少的比例。值为 1 表示输入字段完美地预测了目标，值为 0 则表示输入未提供目标的任何有用信息。

部分为分类字段。当部分—但并非所有—输入为分类字段且目标也为分类字段时，可根据 Pearson 或似然比卡方进行重要性排序。（除非所有输入均为分类变量，否则 Cramer’s V 和 lambda 均不可用。）

分类与连续。针对连续目标来为分类输入排序或与之相反的情形时（即其中之一为分类字段，但不能两者均为分类字段），则使用 F 统计量。

两者均为连续字段。针对连续目标来为连续输入排序时，将使用基于相关系数的 t 统计量。

特征选择模型块

“特征选择”模型块显示每个输入相对于选定目标的重要性（遵循“特征选择”节点的排序）。排序前已筛选掉的所有字段也将被列出。 有关详细信息，请参阅第 64 页码特征选择节点。

运行含有特“特征选择”型块的流时，模型行为将如同过滤器，仅保留“模型”选项卡上当前选中的输入。例如，可以选择评定为“重要”的所有字段（默认选项之一）或在“模型”选项卡上手动选择一个字段子集。不管如何选择，目标字段总是被保留。所有其他字段将被排除。

过滤仅基于字段名称；例如，如果选择年龄和收入，则匹配其中一个名称的任何字段都将被保留。该模型不是基于新数据更新字段排序，而只是根据选定的名称来过滤字段。所以，将模型应用到新的或更新过的数据时应多加注意。存有疑问时，最好重新生成模型。

特征选择模型结果

“特征选择”模型块的“模型”选项卡在顶部窗格显示所有输入的排序和重要性，且可以通过左栏中的复选框选择用于过滤的字段。运行流时，将只保留选中的字段。其他字段将被丢弃。默认选择是基于模型构建节点中指定的选项，但可以根据需要选择或弃选其他字段。

底部窗格列出依据缺失值百分比或建模节点中指定的其他标准而从排序中排除的输入。与其他排序字段一样，可以通过左栏复选框来选择包含或丢弃这些字段。 有关详细信息，请参阅第 65 页码特征选择模型设置。

图片 4-3
特征选择模型结果

- 要按照排序、字段名称重要性或任何其他显示的列来排列该列表的顺序，可单击列标题。如果要使用工具栏，则可以从“排序方式”列表选择需要的项，并使用“向上”和“向下”箭头来更改排序方向。
- 可使用工具栏来选中或弃选所有字段和访问“选中字段”对话框，可在该对话框上根据排序或重要性来选择字段。也可以按住 Shift 和 Ctrl 键并单击字段，以选择更多的字段，并使用空格键来切换选定的字段组。 有关详细信息，请参阅第 69 页码按照重要性选择字段。
- 将输入评定为“重要”、“一般”和“不重要”的阈值显示在表格下方的注释中。这些值在建模节点中指定。 有关详细信息，请参阅第 66 页码特征选择选项。

按照重要性选择字段

使用“特征选择”模型块对数据进行评分时，由排序或筛选字段选中的所有字段都将被保留，如左栏复选框所示。其他字段将被丢弃。要更改选择，可以使用工具栏访问“选中字段”对话框，并在该对话框上根据排序或重要性来选择字段。

图片 4-4
“选中字段”对话框

所有标记字段。 选择标记为“重要”、“一般”和“不重要”的所有字段。

前几个字段。 用于根据重要性选择前 n 个字段。

重要性大于。 请选择重要性大于指定阈值的所有字段。

从特征选择模型中生成过滤器

可根据特征选择模型的结果，生成一个或多个过滤节点，该节点根据相对于指定目标的重要性包含或排除字段子集。虽然模型块也可以用于过滤，但使用此方法可以在不复制或不修改模型的情况下自由地尝试不同的字段子集。不管是选择包含还是选择排除，过滤时将总是保留目标字段。

图片 4-5
生成过滤节点

包含/排除 可选择包含或排除字段—例如包含前 10 个字段或排除所有标记为“不重要”的字段。

选定字段。 包含或排除表中中当前选定的所有字段。

所有标记字段。 选择标记为“重要”、“一般”和“不重要”的所有字段。

前几个字段。 用于根据重要性选择前 n 个字段。

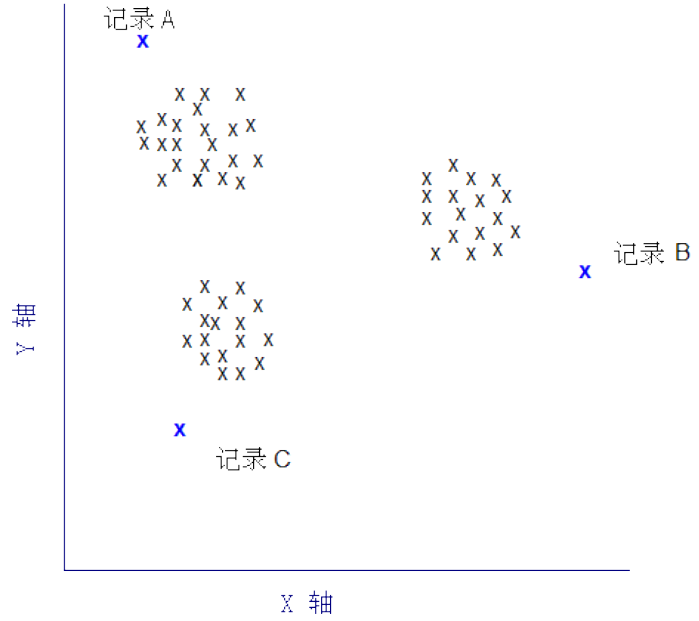
重要性大于。 请选择重要性大于指定阈值的所有字段。

异常检测节点

异常检测模型用于识别数据中的离群值或异常观测值。与存储有关异常观测值的规则的其他建模方法不同，异常检测模型存储有关正常行为的信息。因此即使在离群值不符合任何已知模式的情况下，异常检测模型也使识别离群值成为可能，在新模式可能不断涌现的应用（如缺陷检测）中，该模型可能尤其有用。异常检测是一种不受监督的方法，这就意味着它不需要包含已知缺陷观测值的训练数据集作为开始点。

识别离群值的传统方法通常是一次检查一个或两个变量，而异常检测可以检查大量字段以识别相似记录所属的聚类或对等组。然后，可将每个记录与其对等组中的其他记录进行比较，以识别出可能的异常值。观测值与正常中心值离得越远，它越有可能是异常观测值。例如，该算法可能会将记录聚合为三个不同的聚类，并对离任何一个聚类的中心值较远的那些记录进行标记。

图片 4-6
使用聚类识别潜在异常



每个记录都指定了一个异常指数，该指数是组偏差指数与该观测值所属聚类中平均值的比。此指数的值越大，观测值与平均值的偏差就越大。通常情况下，异常指数值小于 1 甚至小于 1.5 的观测值都不会被视为异常值，因为该偏差与平均值相同或者只是大一点。但是，指数值大于 2 的观测值有可能是异常观测值，因为该偏差至少是平均值的两倍。

异常检测是一种探索性方法，它是为对应该进行进一步分析的可能异常观测值或记录进行快速检测而设计的。这些观测值应视为疑似 异常值，在进行进一步检查后，可以证明它们是或不是真正的异常值。您可能会发现某个记录完全有效，但无法选择从数据中将其筛选出来用于模型构建。另外，如果算法重复检测出虚假异常值，则可能表示数据收集过程中存在错误或假象。

注意：异常检测并不考虑任何特定的目标（相关）字段，也不考虑这些字段是否与正在预测的模式相关，只是通过基于模型中所选字段集的聚类分析确定异常记录或观测值。由于上述原因，您可能想将异常检测与特征选择或字段筛选和排序的其他方法结合使用。例如，您可以使用特征选择来确定与某个特定目标相关的最重要的字段，然后使用异常检测寻找针对这些字段而言最异常的记录。（另外一个方法是构建一个决策树模型，然后将所有错误分类的记录视为可能的异常进行检查。但是此方法很难用于进行大批量的复制和自动化。）

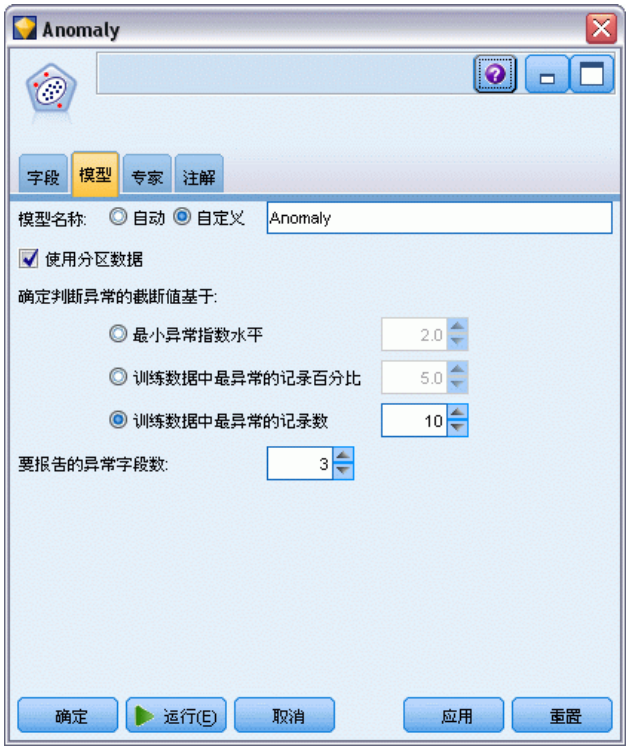
示例。在筛选可能存在农业发展补贴诈骗的案例时，异常检测可用于发现悖于常态的偏差，突出显示那些异常的、值得进一步研究的记录。特别值得关注的是那些看起来相对农场类型和规模而言要求过多（或过少）补助金的补贴申请。

要求。 一个或多个输入字段。注意：只有通过使用源或“类型”节点将角色设置为**输入**的字段才能用作输入。目标字段（角色设置为**目标**或**双向**）将被忽略。

强度。 通过标记不符合已知规则集，而不是符合已知规则集的观测值，异常检测模型即使在观测值不符合先前已知的样式时也能确定异常观测值。当与特征选择组合使用时，异常检测可以用于筛选大量数据，以更快地确定最有用的记录。

异常检测模型选项

图片 4-7
异常检测模型选项卡



模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

确定异常截断值的根据。 指定用于确定标记异常的截断值的方法。可用选项有：

- **最小异常索引等级。**指定标记异常的最小截断值。达到或超过此阈值的记录将被标记。
- **训练数据中最异常记录百分比。** 自动设置一个阈值，其水平标记为训练数据中记录的指定百分比。所得出的截断值作为参数包含在模型中。注意：此选项决定截断值的设置方式，而不 是决定评分期间被标记记录的百分比。实际评分结果可能根据数据的不同而有所不同。
- **训练数据中异常记录的数量。** 自动设置一个阈值，其水平标记为训练数据中记录的指定数量。所得出的临界值作为参数包含在模型中。注意：此选项决定截断值的设置方式，而不 是决定评分期间被标记记录的具体数量。实际评分结果可能根据数据的不同而有所不同。

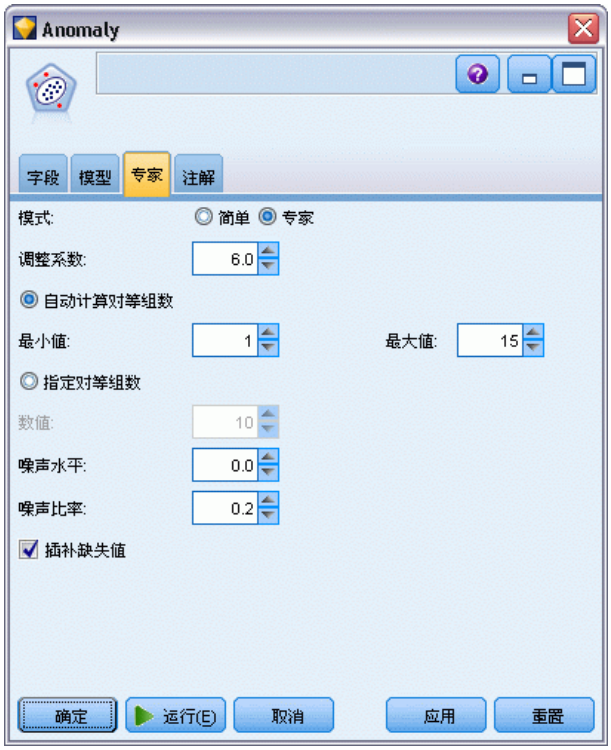
注意：不管如何确定截断值，都不会影响每条记录报告的潜在异常索引值。它只是在对模型进行估算和评分时指定将记录标记为异常的阈值。如果想稍后检查更大或更小数量的记录，则可以使用选择节点来根据异常索引值（$O-AnomalyIndex > X）确定记录子集。

要报告的异常字段数。 指定要报告的异常字段数，报告中说明记录被标记为异常的原因。将报告最异常字段，最异常字段指的是与记录所属聚类的字段标准偏差最大的字段。

异常检测专家选项

要指定缺失值和其他设置的选项，请在“专家”选项卡上将模式设置为**专家**。

图片 4-8
异常检测专家选项卡



调整系数。用于平衡在计算距离时赋予连续（数字范围）和分类字段的相对权重的数值。值越大，连续字段的影响也越大。它必须为非 0 值。

自动计算对等组数。 异常检测可用于快速分析大量可能的解决方案，以选择训练数据的最佳对等组数。可通过设置对等组的最大数和最小数来扩大或缩小数值范围。较大的值可使系统在更大范围内搜索可能的解决方案，但是，消耗的处理时间也随之增加。

指定对等组数。 如果知道模型中应包含聚类的数量，请选中此选项并输入对等组数。一般而言，选中此选项可提高性能。

噪声水平和比率。 这两个设置决定了两阶段聚类期间离群值的处理方式。第一阶段将使用聚类特征 (CF) 树来将数据从大量零散记录浓缩成可管理数量的聚类。该树基于相似性度量值构建，树的某个节点中记录过多时，它将分裂子节点。第二阶段将从 CF 树的终端节点开始创建分层聚类。第一阶段时开启噪声处理，第二阶段时则关闭。第一阶段噪声聚类中的观测值将被分配到第二阶段的常规聚类中。

- **噪声水平。** 指定值必须处于 0 到 0.5 之间。此设置只有在以下情况中才有用：CF 树在增长期间被充满，即该树的叶节点无法接收更多的观测值且叶节点无法分裂。

如果 CF 树被充满且噪声水平设置为 0，则阈值将增大且 CF 树将用所有观测值重新生长。最终聚类之后，不能分配到聚类的变量标记为离群值。离群值聚类将被赋予一个识别号 -1。离群值聚类不包含在聚类数的计数中；也就是说，如果指定 n 个聚类和噪声处理，则算法将输出 n 个聚类和一个噪声聚类。实际应用中，增大此值可使算法更有可能将异常记录匹配到树，而不是将它们分配到独立的离群值聚类。

如果 CF 树被充满且噪声水平大于 0，则 CF 树将在稀疏叶片中的所有数据放到其自身的噪声叶片后重新生长。如果叶片中的观测值数量与最大叶片中的观测值数量的比率小于噪声水平，则该叶片将被认定为稀疏叶片。树创建完成后，可能的话，离群值将被放置在 CF 树中。如果未放在树中，第二步聚类中的离群值将被丢弃。

- **噪声比。**指定分配给用于噪声缓冲的组件的内存量。此值必须处于 0.0 到 0.5 之间。如果将特定观测值插入树的叶片后，所产生的紧性小于阈值，叶片将不再分裂。如果紧性超过阈值，叶片将分裂，结果将把另一个小聚类添加到 CF 树。实际上，提高此设置值将可能导致算法容易更快速地创建较简单的树。

为缺失值归因。对于连续字段，请用字段均值代替缺失值。对于分类字段，不同缺失值类别将被组合为一个有效分类进行处理。如果取消选中此选项，则任何带有缺失值的记录都将从分析中剔除。

异常检测模型块

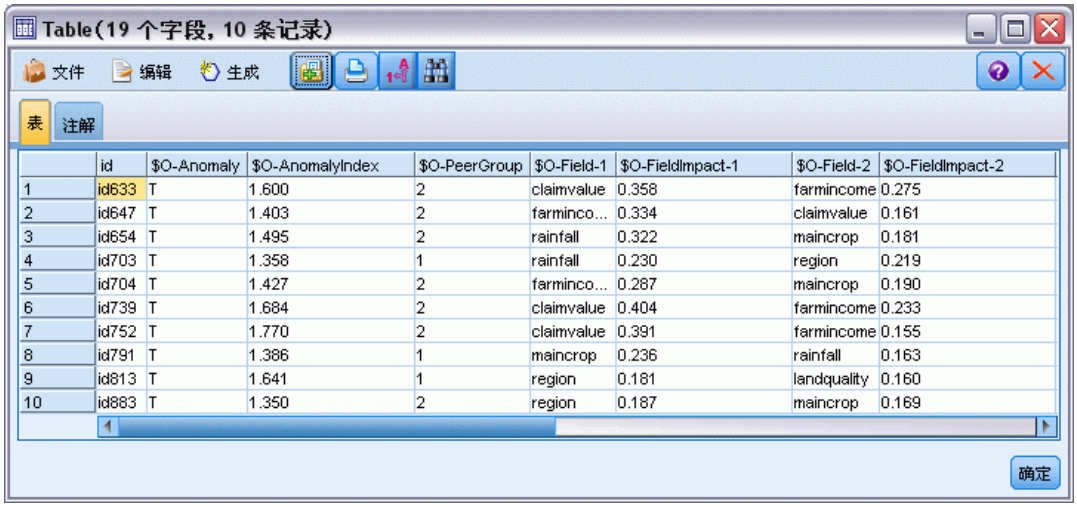
“异常检测”模型块含有“异常检测”模型捕获的所有信息以及有关训练数据和估算过程的信息。

运行含有“异常检测”模型块的流时，若干新字段将按照模型块中“设置”选项卡上的选择添加到流中。有关详细信息，请参阅第 77 页码异常检测模型设置。 新字段名称基于模型名称，并带有前缀 $O，如下表之概括：

$O-Anomaly	指明记录是否异常的标志字段。
$O-AnomalyIndex	记录的异常索引值。
$O-PeerGroup	指定记录分配到哪个对等组。
$O-Field-n	与聚类标准偏差最大的第 n 个异常字段的名称。
$O-FieldImpact-n	字段的变量偏差指数。此值衡量与记录分配到的聚类字段标准的偏差。

也可以选择抑止非异常记录的得分，以使结果更易于读取。

图片 4-9
抑制非异常记录的评分结果

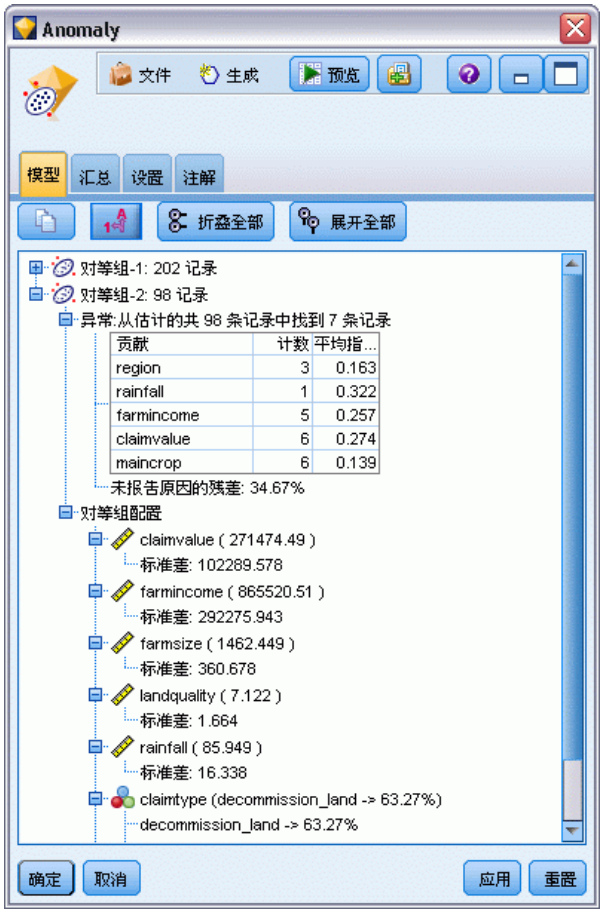


	id	$O-Anomaly	$O-AnomalyIndex	$O-PeerGroup	$O-Field-1	$O-FieldImpact-1	$O-Field-2	$O-FieldImpact-2
1	id633	T	1.600	2	claimvalue	0.358	farmincome	0.275
2	id647	T	1.403	2	farminco...	0.334	claimvalue	0.161
3	id654	T	1.495	2	rainfall	0.322	maincrop	0.181
4	id703	T	1.358	1	rainfall	0.230	region	0.219
5	id704	T	1.427	2	farminco...	0.287	maincrop	0.190
6	id739	T	1.684	2	claimvalue	0.404	farmincome	0.233
7	id752	T	1.770	2	claimvalue	0.391	farmincome	0.155
8	id791	T	1.386	1	maincrop	0.236	rainfall	0.163
9	id813	T	1.641	1	region	0.181	landquality	0.160
10	id883	T	1.350	2	region	0.187	maincrop	0.169

异常检测模型详细信息

已生成异常检测模型的“模型”选项卡显示模型中对等组的有关信息。

图片 4-10
异常检测模型块详细信息



注意：报告的对等组大小和统计信息是基于训练数据的估算值，即使在同样的数据上运行，可能也与实际评分结果稍微有点不同。

异常检测模型概要

“异常检测”模型块的“概要”选项卡显示字段、构建设置和估算过程的有关信息。同时还显示对等组数量以及用于将记录标记为异常的截断值。

图片 4-11
异常检测模型块概要

异常检测模型设置

“设置”选项卡用于指定对模型块进行评分的选项。

图片 4-12
异常检测模型块的评分选项

使用以下方式表示异常记录。 指定输出中异常记录的处理方式。

- **标志和指数。**创建标志字段，模型中包含的任何记录超过截断值时，记录的标志字段将被设置为真。同时将单独使用一个字段中报告每条记录的异常索引。 有关详细信息，请参阅第 72 页码异常检测模型选项。
- **仅标志。** 创建标志字段，但不报告每条记录的异常索引。
- **仅指数。** 报告异常索引但不创建标志字段。

要报告的异常字段数。 指定要报告的异常字段数，报告中说明记录被标记为异常的原因。将报告最异常字段，最异常字段指的是与记录所属聚类的字段标准偏差最大的字段。

丢弃记录. 选择此选项，以从流丢弃所有非异常记录，以便在下游节点中集中关注于潜在异常。此外，也可以丢弃所有异常记录，以便将后续分析限制在那些没有被模型标记为潜在异常的记录上。

注意：由于取整造成的轻微差异，即使在同样的数据上运行，评分期间被标记记录的实际数量也可能与训练模型期间的情况有所不同。

章 5

自动建模节点

自动建模节点估算和比较多个不同的建模方法，使您在一次建模运行中即可尝试各种方法。您可以选择所使用的建模算法，以及每个建模算法的具体选项，包括可能互斥的组合。例如，您无需为神经网络选择快速、动态或修剪之中的某个方式，完全可以全部尝试。节点研究选项的每个可能组合，根据您指定的测量为每个候选模型排序，并保存最佳模型用于评分或将来的分析。

您可以根据分析需要从三个自动建模节点中进行选择：

“自动分类器”节点用于创建和对比二元结果（是或否，流失或不流失等）的若干不同模型，使用户可以选择给定分析的最佳处理方法。由于支持多种建模算法，因此可以对用户希望使用的方法、每种方法的特定选项以及对比结果的标准进行选择。节点根据指定的选项生成一组模型并根据用户指定的标准排列最佳候选项的顺序。 有关详细信息，请参阅第 81 页码自动分类器节点。

自动数值节点使用多种不同方法估计和对比模型的连续数字范围结果。此节点和自动分类器节点的工作方式相同，因此可以选择要使用和要在单个建模传递中使用多个选项组合进行测试的算法。受支持的算法包括神经网络、C&R 树、CHAID、线性回归、广义线性回归以及 Support Vector Machine (SVM)。可基于相关度、相对错误或已用变量数对模型进行对比。 有关详细信息，请参阅第 90 页码自动数值节点。

自动聚类节点估算和比较识别具有类似特征记录组的聚类模型。节点工作方式与其他自动建模节点相同，使您在一次建模运行中即可试验多个选项组合。模型可使用基本测量进行比较，以尝试过滤聚类模型的有效性以及对其进行排序，并提供一个基于特定字段的重要性的测量。 有关详细信息，请参阅第 95 页码自动聚类节点。

最佳模型保存在一个复合模型块中，可对其进行浏览和比较，并选择评分中使用哪个模型。

- 只有对于二元、名义和数字目标，您才可以选择多个评分模型，并将得分组合在一个模型整体中。通过结合多个模型的预测，可以避免单个模型的局限性，使所得的整体准确性通常比从任一模型中获得的准确性要高。
- 您还可以选择向下浏览结果，或为要使用或进一步探索的所有单独模型生成建模节点或模型块。

模型和执行时间

根据模型的数据集和数量，自动建模节点执行时间可能为数小时或甚至更长。在选择选项时，请注意正在生成的模型个数。如果现实条件允许，您可能希望将建模运行的时间安排在夜晚或周末，因为此时对系统资源的需求可能比较小。

- 必要的话，可以使用分区节点或样本节点减少包括在初始训练传递中的记录数。一旦将选择限制在几个生成的候选模型内，就可以恢复全部数据集。 有关详细信息，请参阅样本节点 或。

- 要减少输入字段数，请使用特征选择。有关详细信息，请参阅第 64 页码第 4 章中的特征选择节点。或者，可以使用初始建模运行来识别需要进一步探索的字段和选项。例如，如果性能最佳的模型似乎都使用了相同的三个字段，则有力地说明这些字段值得保留。
- 您还可以限制评估任一模型所需的时间并且指定用于过滤和排序模型的评估尺度。

自动建模节点算法设置

对于每个模型类型，可以使用默认设置，或为每个模型类型选择选项。这些特定选项类似于独立建模节点中可用的选项，不同之处在于并非只能选择一种设置而是大多数情况下可以根据应用需要选择多种。例如，如果对比神经网络节点，可以选择几种不同的训练方法，并且尝试具有随机种子和不具有随机种子的每种方法。选定选项的所有可能组合都将使用，从而使得在单次遍历中生成许多不同模型变得更容易。但是，使用时要小心，因为选择多个设置会引起模型数非常快速地增加。

图片 5-1
为自动建模选择算法设置

要为每个模型类型选择选项：

- ▶ 在自动建模节点上，选择**专家**选项卡。
- ▶ 单击模型类型的**模型参数**列。
- ▶ 从下拉菜单中，选择**指定**。

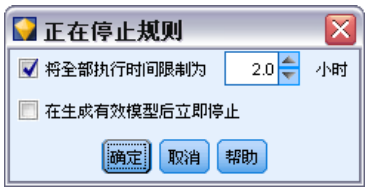
- 在**算法设置**对话框上，从**选项**列中选择选项。

  注意：在**算法设置**对话框的“专家”选项卡上提供了进一步的选项。

自动建模节点停止规则

为自动建模节点指定的停止规则不仅与节点所构建的个别模型的停止有关，还与所有节点执行有关。

图片 5-2
正在停止规则



总执行时间限制。 （仅神经网络、K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和 C&R 模型）在指定小时数后停止执行。所有在该时间点之前（包括该点）生成的模型都将包括在模型块中，但这之后不会再生成模型。

生成有效的模型后立即停止。当模型传递了所有在“丢弃”选项卡（自动分类器或自动聚类节点的）和“模型”选项卡（自动数值节点的）上指定的标准时将停止执行。有关详细信息，请参阅第 87 页码自动分类器节点丢弃选项。有关详细信息，请参阅第 99 页码自动聚类节点丢弃选项。

自动分类器节点

“自动分类器”节点使用多种不同方法估算和比较名义（集合）或二元（是/否）目标的模型，使您在一次建模运行中即可尝试各种方法。您可以选择所用算法，并试验选项的多个组合。例如，您无需为神经网络选择快速、动态或修剪之中的某个方式，完全可以全部尝试。节点研究选项的每个可能组合，根据您指定的测量为每个候选模型排序，并保存最佳模型用于评分或将来的分析。有关详细信息，请参阅第 79 页码自动建模节点。

图片 5-3
自动分类器建模结果

示例。零售公司拥有追踪以前营销活动中向特定客户报价的历史数据。公司现在希望通过向每个客户提供合适的报价，来获取更多的利润。显示

要求。 一个测量级别为名义或标志（角色设置为**目标**）的目标字段和至少一个输入字段（角色设置为**输入**）。对于“标志”字段，假定为目标字段定义的真值表示计算利润、提升和相关统计量时的匹配项。输入字段的测量级别可以是连续或分类，但具有限制，即某些输入可能不适合一些模型类型。例如，在 C&R 树、CHAID 和 QUEST 模型中用作输入的有序字段必须是数字存储类型（而不是字符串），如果指定了其他类型，将被这些模型忽略。类似地，在某些情况下可对连续输入字段进行分级。这和使用单个建模节点时的要求一样；例如，不管是从贝叶斯网络节点还是自动分类器节点生成，贝叶斯网络模型都以同样的方式工作。

频数和加权字段。频数和加权用于增强某些记录的重要性，以超过其他记录，原因可能是用户知道构建数据集省略父总体的一部分（加权）或一个记录代表一个相同观测值数（频数）等。如果指定了频数字段，则可以将其用于 C&R 树、CHAID、QUEST、决策列表和贝叶斯网络模型。加权字段可用于 C&RT、CHAID 和 C5.0 模型。其他模型类型将省略这些字段并以任意方式构建模型。频数和加权字段仅用于模型构建，并且在评估和评分模型时不予以考虑。 有关详细信息，请参阅第 33 页码第 3 章中的使用频率和权重字段。

支持的模型类型

支持的模型类型包括神经网络、C&R 树、QUEST、CHAID、C5.0、Logistic 回归、决策列表、贝叶斯网络、判别式、最近邻元素和 SVM。 有关详细信息，请参阅第 84 页码自动分类器节点专家选项。

自动分类器节点模型选项

使用自动分类器节点的“模型”选项卡可以指定要创建的模型数和用于比较模型的标准。

图片 5-4
自动分类器节点：“模型”选项卡

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

模型排序依据。指定用于比较和排序模型的标准。选项包括总体精确性、ROC 曲线下的区域、利润、提升和字段的数量。请注意，无论在此处选定哪些尺度，所有这些尺度都能在汇总报告中使用。

注意：对于名义（集合）目标，排序限制为**总体精确性**或**字段数**。

计算利润、提升和相关统计量时，将假定目标字段定义为真 值以表示匹配项。

-
-
-
-
-

模型排序方式。 如果正在使用分区，则可以指定基于训练数据集排序或是基于检验数据集排序。对于大型数据集，使用分区对模型进行预筛选将大大提高执行能力。

要使用的模型数。指定要在节点生成的模型块中列出的最大模型数。按照指定的排序标准将顺次列出排序靠前的模型。注意，增加此最大模型数将减缓执行速度。允许的最大值为 100。

计算预测变量重要性。对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。请注意，预测变量重要性可能会增加计算某些模型所需的时间，如果仅仅希望对许多不同的模型进行广泛对比，则不建议评估变量重要性。将分析限制在要进一步探索的几个模型上会更有用。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

利润标准。注意：仅适合标志目标。利润等于每个记录的收入减去该记录的成本。也就是说，分位数的利润就是位于该分位数内的所有记录的利润总和。这里假定利润仅应用于匹配项，但成本可应用于所有的记录。

- **成本。**指定与每个记录相关联的成本。您可以选择**固定**或**可变**成本。对于固定成本，请指定成本值。对于可变成本，请单击“字段选择器”按钮，将某个字段选择为成本字段。
- **收入。**指定与表示匹配项的每个记录相关联的收入。您可以选择**固定**或**可变**成本。对于固定收入，请指定收入值。对于可变收入，请单击“字段选择器”按钮，将某个字段选择为收入字段。
- **加权。**如果数据中的记录代表多个单元，则可以使用频数加权调整结果。使用**固定**或**可变**加权，指定与每个记录相关联的加权。对于固定加权，请指定加权值（每个记录的单元数）。对于可变加权，请单击“字段选择器”按钮，将某个字段选择为加权字段。

提升标准。注意：仅适合标志目标。指定提升计算使用的百分位数。注意，在比较结果时也可以更改此值。 有关详细信息，请参阅第 100 页码自动模型块。

自动分类器节点专家选项

使用自动分类器节点的“专家”选项卡，可以应用分区（如果可用）、选择使用的算法及指定停止规则。

图片 5-5
自动分类器节点：“专家”选项卡

已使用模型。使用左侧列中的复选框选择要在比较中包括的模型类型（算法）。选择的类型越多，创建的模型就会越多，且处理的时间就会越长。

模型类型。列出可用的算法（请参阅下面的内容）。

模型参数。对于每个模型类型，可以使用默认设置，或选择**指定**为每个模型类型选择选项。这些特定选项类似于独立建模节点中可用的选项，不同之处在于可以选择多个选项或组合。例如，比较神经网络模型时，与其选择六种训练方法之一，还不如一次选中全部六种方法以在一次传递中训练六种模型。

模型数。 列出基于当前设置为每个算法生成的模型数。当组合选项时，模型数会激增，因此强烈建议密切关注该模型数，尤其在使用大型数据集时。

限制单模型最长构建时间。（仅 K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和决策列表模型）为任意一个模型设置最长时间限制。例如，如果由于某些复杂的交互效应，某个特定模型所需的训练时间长得出乎意料，则您大概不希望它使得整个的建模运行停滞。

注意：如果目标为名义（集合）字段，“决策列表”选项不可用。

支持的算法

神经网络节点使用的模型是对人类大脑处理信息的方式简化了的模型。此模型通过模拟大量类似于神经元的抽象形式的互连简单处理单元而运行。神经网络是功能强大的一般函数估计器，只需要最少的统计或数学知识就可以对其进行训练或应用。

C5.0 节点构建决策树或规则集。该模型的工作原理是根据在每个级别提供最大信息收获的字段分割样本。目标字段必须为分类字段。允许进行多次多于两个子组的分割。 有关详细信息，请参阅第 146 页码第 6 章中的C5.0 节点。

分类和回归 (C&R) 树节点生成可用于预测或分类未来观测值的决策树。该方法通过在每个步骤最大限度降低不纯洁度，使用递归分区来将训练记录分割为组。如果节点中 100% 的观测值都属于目标字段的一个特定类别，则树中的该节点将被认定为“纯洁”。目标和输入字段可以是数字范围或分类（名义、有序或标志）；所有分割均为二元分割（即仅分割为两个子组）。 有关详细信息，请参阅第 130 页码第 6 章中的C&R 树节点。

QUEST 节点可提供用于构建决策树的二元分类法，此方法的设计目的是减少大型 C&R 树分析所需的处理时间，同时也减少在分类树方法中发现的趋势以便支持允许有多个分割的输入。输入字段可以是数字范围（连续），但目标字段必须是分类。所有分割都是二元的。 有关详细信息，请参阅第 131 页码第 6 章中的QUEST 节点。

CHAID 使用卡方统计量来生成决策树，以确定最佳的分割。CHAID 与 C&R 树和 QUEST 节点不同，它可以生成非二元树，这意味着有些分割将有多于两个的分支。目标和输入字段可以是数字范围（连续）或分类。Exhaustive CHAID 是 CHAID 的修正版，它对所有分割进行更彻底的检查，但计算时间比较长。 有关详细信息，请参阅第 130 页码第 6 章中的CHAID 节点。

Logistic 回归是一种统计方法，它可根据输入字段的值对记录进行分类。它类似于线性回归，但采用的是类别目标字段而非数字范围。 有关详细信息，请参阅第 239 页码第 10 章中的逻辑节点。

决策列表节点可标识子组或段，显示与总体相关的给定二元结果的似然度的高低。例如，您或许在寻找那些最不可能流失的客户或最有可能对某个商业活动作出积极响应的客户。通过定制段和并排预览备选模型来比较结果，您可以将自己的业务知识体现在模型中。决策列表模型由一组规则构成，其中每个规则具备一个条件和一个结果。规则依顺序应用，相匹配的第一个规则将决定结果。 有关详细信息，请参阅第 188 页码第 9 章中的决策表。

通过贝叶斯网络节点，你可以利用对真实世界认知的判断力并结合所观察和记录的证据来构建概率模型。该节点重点应用了树扩展简单贝叶斯 (TAN) 和马尔可夫毯网络，这些算法主要用于分类问题。 有关详细信息，请参阅第 165 页码第 7 章中的贝叶斯网络节点。

判别式分析所做的假设比 logistic 回归的假设更严格，但在符合这些假设时，判别式分析可以作为 logistic 回归分析的有用替代项或补充。 有关详细信息，请参阅第 263 页码第 10 章中的判别式节点。

The k-最近相邻元素 (KNN) 节点将新的个案关联到预测变量空间中与其最邻近的 k 个对象的类别或值（其中 k 为整数）。类似个案相互靠近，而不同个案相互远离。 有关详细信息，请参阅第 396 页码第 16 章中的KNN 节点。

使用 Support Vector Machine (SVM) 节点，可以将数据分为两组，而无需过度拟合。SVM 可以与大量数据集配合使用，如那些含有大量输入字段的数据集。 有关详细信息，请参阅第 391 页码第 15 章中的SVM 节点。

误分类损失

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错误的相对重要性。

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序或比较时，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。

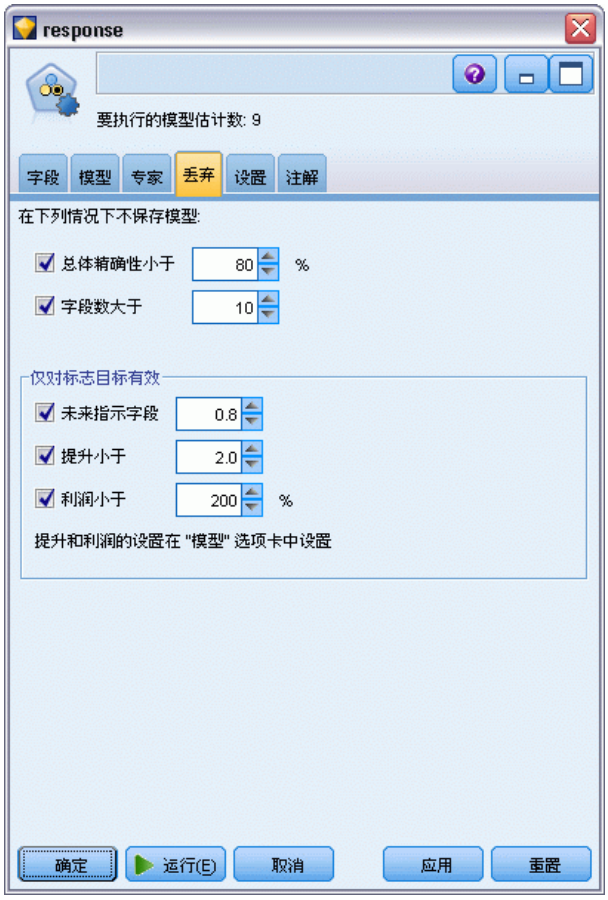
成本矩阵显示了预测类别和实际类别的每个可能的组合的成本。默认情况下，所有误分类成本都设置为 1.0。要输入自定义成本值，可选择**使用误分类成本**并将自定义值输入到成本矩阵中。

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，则将 B 误分类为 A 的成本将仍是默认值 1.0，除非也明确地对它进行更改。

自动分类器节点丢弃选项

使用自动分类器节点的“丢弃”选项卡，可以自动丢弃不符合特定标准的模型。这些模型将不会列在汇总报告中。

图片 5-6
自动分类器节点：“丢弃”选项卡



可以为总准确性指定最小阈值，为模型中使用的变量数指定最大阈值。此外，对于标志目标，可以为提升、利润和曲线下方区域指定最小阈值，提升和利润由在“模型”选项卡上指定的内容所确定。 有关详细信息，请参阅第 83 页码自动分类器节点模型选项。

或者，可以将节点配置为在首次生成满足所有指定标准的模型时停止执行。 有关详细信息，请参阅第 81 页码自动建模节点停止规则。

自动分类器节点设置选项

“自动分类器”节点的“设置”选项卡允许您预配置块上可用的得分时间选项。

图片 5-7
自动分类器节点：“设置”选项卡

整体方法。对于目标，您可以从以下整体方法选择：

- 投票
- 置信度加权投票
- 原始倾向加权投票（仅适用于标志目标）
- 赢得最高置信度
- 平均原始倾向（仅适用于标志目标）。

有关详细信息，请参阅第 4 章中的整体节点设置中的IBM SPSS Modeler 14.2 源、过程和输出节点。

如果约束投票，请使用以下选项选择值。根据投票方法，可以指定解决投票同数的方法：

- **随机选择。**随机选择其中一个同数值。
- **最高置信度。**选择使用最高置信度进行预测的同数值。请注意，该置信度值无需与所有预测值的最高置信度值相同。
- **原始倾向。**（仅适合标志目标）使用最大绝对倾向预测的同数值，其中绝对倾向的计算方法如下：

```
abs(0.5 - propensity) * 2
```

自动数值节点

自动数值节点使用多种不同方法估算和比较模型得出连续数字范围结果，使您在一次运行中即可尝试各种方法。您可以选择所用算法，并试验选项的多个组合。例如，可以使用神经网络、线性回归、C&RT 和 CHAID 模型预测住房值，以查看哪种模型的性能最好，并且可以尝试逐步、前进和后退回归法的不同组合。节点研究选项的每个可能组合，根据您指定的测量为每个候选模型排序，并保存最佳模型用于评分或将来的分析。 有关详细信息，请参阅第 79 页码自动建模节点。

图片 5-8
自动数值结果

示例。市Õþ当局需要更准确地估计房地产税以及无需检查每个属性就可以按需要调整特定属性的值。使用自动数值节点，分析师能产生和对比许多基于构建类型、近邻、大小和其他已知因素来预测属性值的模型。有关详细信息，请参阅第 6 章中的属性值（自动数值）中的IBM SPSS Modeler 14.2 应用程序 指南。

要求。 一个目标字段（角色设置为**目标**）和至少一个输入字段（角色设置为**输入**）。目标必须为连续（数值范围）字段，如年龄或收入。输入字段可以是连续或分类，但具有限制，即某些输入可能不适合一些模型类型。例如，C&R 树模型能将分类字符串字段作为输入使用，而线性回归模型不能使用这些字段并将在指定这些字段后省略它们。这和使用单独建模节点时的要求相同。例如，不管 CHAID 模型是在 CHAID 节点中还是在自动数值节点中生成，其工作方式都相同。

频数和加权字段。频数和加权用于增强某些记录的重要性，以超过其他记录，原因可能是用户知道构建数据集省略父总体的一部分（加权）或一个记录代表一个相同观测值数（频数）等。如果指定频数字段，它就可以用于 C&R 树和 CHAID 算法。加权字段可用于 C&RT、CHAID 回归和 GenLin 算法。其他模型类型将省略这些字段并以任意方式构建

模型。频数和加权字段仅用于模型构建，并且在评估和评分模型时不予以考虑。 有关详细信息，请参阅第 33 页码第 3 章中的使用频率和权重字段。

支持的模型类型

支持的模型类型包括神经网络、C&R 树、CHAID、回归、GenLin、最近相邻元素和 SVM。有关详细信息，请参阅第 92 页码自动数值节点专家选项。

自动数值节点模型选项

使用自动数值节点的“模型”选项卡可以指定要保存的模型数，以及用于比较模型的标准。

图片 5-9
自动数值节点：“模型”选项卡

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

模型排序依据。指定用于比较模型的标准。

- **相关。**每条记录的观测值和模型预测的值之间的 Pearson 相关性。相关性是两种变量之间的线性关联尺度，值越接近 1 说明变量之间的关系越强。（相关性的值在 -1 和 +1 之间，-1 代表完全负关系，+1 代表完全正关系。值为 0 表示无线性关系，但具有负相关性的模型将排在最后。）
- **字段数。**模型中用作预测变量的字段的数量。在某些情况下，选择使用较少字段的模型可简化数据准备过程并提高性能。
- **相对错误。**相对错误是模型预测值的观测值的方差与平均值的观测值方差的比率。在实际应用的角度，它对比模型相对于**空**或**截距**模型（仅返回目标字段的平均值作为预测值）的性能。对于好的模型，此值应小于 1，说明此模型比空模型更精确。相对错误大于 1 的模型不如空模型精确，因此这样的模型没有意义。对于线性回归模型，相对错误等同于相关性的平方并且未添加任何新的信息。对于非线性模型，相对错误与相关性无关并且为评估模型性能提供了附加尺度。

模型排序方式。如果正在使用分区，则可以指定基于训练分区排序还是基于测试分区排序。对于大型数据集，使用分区对模型进行预筛选将大大提高执行能力。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

要使用的模型数。指定要在节点生成的模型块中显示的最大模型数。按照指定的排序标准将顺次列出排序靠前的模型。增加此限制可以对比更多模型的结果，但是可能会降低性能。允许的最大值为 100。

计算预测变量重要性。对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。请注意，预测变量重要性可能会增加计算某些模型所需的时间，如果仅仅希望对许多不同的模型进行广泛对比，则不建议评估变量重要性。将分析限制在要进一步探索的几个模型上会更有用。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

在下列情况下不保存模型。指定相关性、相对错误和所用字段数的临界值。无法满足这些标准中的任意一个的模型将被丢弃，并且不会在汇总报告中列出。

- **相关小于。**要包含在汇总报告中的模型的最小相关性（按绝对值）。
- **所用字段数大于。**要包含的任意模型要使用的字段的最大数。
- **相对错误大于。**要包含的任意模型的最大相对错误。

或者，可以将节点配置为在首次生成满足所有指定标准的模型时停止执行。 有关详细信息，请参阅第 81 页码自动建模节点停止规则。

自动数值节点专家选项

使用自动数值节点的“专家”选项卡可以选择要使用和要指定停止规则的算法和选项。

图片 5-10
自动数值节点：“专家”选项卡

已使用模型。使用左侧列中的复选框选择要在比较中包括的模型类型（算法）。选择的类型越多，创建的模型就会越多，且处理的时间就会越长。

模型类型。列出可用的算法（请参阅下面的内容）。

模型参数。对于每个模型类型，可以使用默认设置，或选择**指定**为每个模型类型选择选项。这些特定选项类似于独立建模节点中可用的选项，不同之处在于可以选择多个选项或组合。例如，比较神经网络模型时，与其选择六种训练方法之一，还不如一次选中全部六种方法以在一次传递中训练六种模型。

模型数。 列出基于当前设置为每个算法生成的模型数。当组合选项时，模型数会激增，因此强烈建议密切关注该模型数，尤其在使用大型数据集时。

限制单模型最长构建时间。（仅 K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和决策列表模型）为任意一个模型设置最长时间限制。例如，如果由于某些复杂的交互效应，某个特定模型所需的训练时间长得出乎意料，则您大概不希望它使得整个的建模运行停滞。

支持的算法

神经网络节点使用的模型是对人类大脑处理信息的方式简化了的模型。此模型通过模拟大量类似于神经元的抽象形式的互连简单处理单元而运行。神经网络是功能强大的一般函数估计器，只需要最少的统计或数学知识就可以对其进行训练或应用。

分类和回归 (C&R) 树节点生成可用于预测或分类未来观测值的决策树。该方法通过在每个步骤最大限度降低不纯洁度，使用递归分区来将训练记录分割为组。如果节点中 100% 的观测值都属于目标字段的一个特定类别，则树中的该节点将被认定为“纯洁”。目标和输入字段可以是数字范围或分类（名义、有序或标志）；所有分割均为二元分割（即仅分割为两个子组）。 有关详细信息，请参阅第 130 页码第 6 章中的C&R 树节点。

CHAID 使用卡方统计量来生成决策树，以确定最佳的分割。CHAID 与 C&R 树和 QUEST 节点不同，它可以生成非二元树，这意味着有些分割将有多于两个的分支。目标和输入字段可以是数字范围（连续）或分类。Exhaustive CHAID 是 CHAID 的修正版，它对所有分割进行更彻底的检查，但计算时间比较长。 有关详细信息，请参阅第 130 页码第 6 章中的CHAID 节点。

线性回归是一种通过拟合直线或平面以实现汇总数据和预测的普通统计方法，它可使预测值和实际输出值之间的差异最小化。

“广义线性”模型对一般线性模型进行了扩展，这样因变量通过指定的关联函数与因子和协变量线性相关。另外，该模型允许因变量呈非正态分布。它包括统计模型大部分的功能，其中包括线性回归、logistic 回归、用于计数数据的对数线性模型以及区间删失生存模型。 有关详细信息，请参阅第 270 页码第 10 章中的GenLin 节点。

The k-最近相邻元素 (KNN) 节点将新的个案关联到预测变量空间中与其最邻近的 k 个对象的类别或值（其中 k 为整数）。类似个案相互靠近，而不同个案相互远离。 有关详细信息，请参阅第 396 页码第 16 章中的KNN 节点。

使用 Support Vector Machine (SVM) 节点，可以将数据分为两组，而无需过度拟合。SVM 可以与大量数据集配合使用，如那些含有大量输入字段的数据集。 有关详细信息，请参阅第 391 页码第 15 章中的SVM 节点。

线性回归模型根据目标与一个或多个预测变量间的线性关系来预测连续目标。 有关详细信息，请参阅第 221 页码第 10 章中的线性模型。

自动数值节点设置选项

“自动数值”节点的“设置”选项卡允许您预配置块上可用的得分时间选项。

图片 5-11
自动数值节点：“设置”选项卡

计算标准误。对于连续（数值范围）目标，默认情况下会运行标准误计算以计算测量或估算值与真值之间的差值；并显示这些估算值的相近匹配程度。

自动聚类节点

自动聚类节点估算和比较识别具有类似特征记录组的聚类模型。节点工作方式与其他自动建模节点相同，使您在一次建模运行中即可试验多个选项组合。模型可使用基本测量进行比较，以尝试过滤聚类模型的有效性以及对其进行排序，并提供一个基于特定字段的重要性的测量。

聚类模型常常用于识别在后续分析中可用作输入的组。例如，您可能希望基于如收入的统计特征来针对客户群，或基于客户过去购买的服务而针对客户群。可以在不了解客户群及其特征的情况下进行此操作 — 您可能不知道要寻找多少个客户群，或该用什么特征去定义客户群。聚类模型常称作不受监督的学习模型，因为其不使用目标字段，且不返回可估算为真或假的具体预测。聚类模型的值由模型捕获数据中感兴趣的分组并提供这些分组的有用说明信息的能力来确定。 有关详细信息，请参阅第 292 页码第 11 章中的聚类模型。

图片 5-12
自动聚类结果

要求。定义兴趣特征的一个或多个字段。聚类模型使用目标字段的方式与其他模型不同，因为其不作出能被评估为真或假的特定预测。相反，其用于识别可能相关的个案组。例如，您无法使用预测给定客户会流失还是对预订作出积极响应的聚类模型。但您可以使用基于客户对此类事物的倾向性将客户分组的聚类模型。也不使用加权和频率字段。

评估字段。虽然不使用目标，您可以指定在比较模型中使用的一个或多个评估字段。可通过测量聚类区分这些字段的好坏情况来评估聚类模型的有效性。

支持的模型类型

支持模型类型包括两步、K 均值及 Kohonen 类型。

自动聚类节点模型选项

使用自动聚类节点的“模型”选项卡可以指定要保存的模型数，以及用于比较模型的标准。

图片 5-13
自动聚类节点：“模型”选项卡

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

模型排序依据。指定用于比较和排序模型的标准。

- **Silhouette。**索引测量聚类结合和分离。有关详细信息，请参阅下面的 Silhouette 排序测量。
- **聚类数。**模型中的聚类数。
- **最小聚类的大小。**最小聚类的大小。
- **最大聚类的大小。**最大聚类的大小。
- **最小/最大聚类。**最小聚类与最大聚类的大小比率。
- **重要性。**字段选项卡上的评估字段的重要性。注意只有在评估字段已指定时，才能计算。

模型排序方式。 如果正在使用分区，则可以指定基于训练数据集排序或是基于检验数据集排序。对于大型数据集，使用分区对模型进行预筛选将大大提高执行能力。

要保留的模型数。指定要在节点生成的块中列出的最大模型数。按照指定的排序标准将顺次列出排序靠前的模型。注意，增加此最大模型数将减缓执行速度。允许的最大值为 100。

Silhouette 排序测量

默认排序测量，Silhouette，默认值为 0，这是因为小于 0 的值（即负值）表示其分配的聚类中的观测值与点之间的平均距离大于观测值与另一个聚类中点的最小平均距离。因此，具有负 Silhouette 值的模型可以安全地丢弃。

排序测量实际上为修改的 silhouette 系数，它结合了聚类结合（偏向包含紧密结合聚类的模型）和聚类分离（偏向包含高度分离聚类的模型）的概念。平均 Silhouette 系数是在所有观测值上的简单平均，每个单独观测值应用下列计算：

(B - A) / max(A, B)

其中 A 为从观测值到其所属聚类的矩心的距离，B 为从观测值到每个其他聚类矩心的最小距离。

Silhouette 系数（及其平均值）大小在 -1（表示极差的模型）与 1（表示极好的模型）之间。可以在总体观测值级别上求平均值（得到总体 Silhouette），也可在聚类级别上求平均值（得到聚类 Silhouette）。距离可以使用 Euclidean 距离进行计算。

自动聚类节点专家选项

使用自动聚类节点的“专家”选项卡，可以应用分区（如果可用）、选择使用的算法及指定停止规则。

图片 5-14
自动聚类节点：“专家”选项卡

已使用模型。使用左侧列中的复选框选择要在比较中包括的模型类型（算法）。选择的类型越多，创建的模型就会越多，且处理的时间就会越长。

模型类型。列出可用的算法（请参阅下面的内容）。

模型参数。对于每个模型类型，可以使用默认设置，或选择**指定**为每个模型类型选择选项。这些特定选项类似于独立建模节点中可用的选项，不同之处在于可以选择多个选项或组合。例如，比较神经网络模型时，与其选择六种训练方法之一，还不如一次选中全部六种方法以在一次传递中训练六种模型。

模型数。 列出基于当前设置为每个算法生成的模型数。当组合选项时，模型数会激增，因此强烈建议密切关注该模型数，尤其在使用大型数据集时。

限制单模型最长构建时间。（仅 K-Means、Kohonen、TwoStep、SVM、KNN、Bayes Net 和决策列表模型）为任意一个模型设置最长时间限制。例如，如果由于某些复杂的交互效应，某个特定模型所需的训练时间长得出乎意料，则您大概不希望它使得整个的建模运行停滞。

支持的算法

K-Means 节点将数据集聚类到不同分组（或聚类）。此方法将定义固定的聚类数量，将记录迭代分配给聚类，以及调整聚类中心，直到进一步优化无法再改进模型。k-means 节点作为一种非监督学习机制，它并不试图预测结果，而是揭示隐含在输入字段集中的模式。 有关详细信息，请参阅第 298 页码第 11 章中的K-Means 节点。

Kohonen 节点会生成一种神经网络，此神经网络可用于将数据集聚类到各个差异组。此网络训练完成后，相似的记录应在输出映射中紧密地聚集，差异大的记录则应彼此远离。您可以通过查看模型块中每个单元所捕获观测值的数量来找出规模较大的单元。这将让您对聚类的相应数量有所估计。 有关详细信息，请参阅第 293 页码第 11 章中的Kohonen 节点。

TwoStep 节点使用两步聚类方法。第一步完成简单数据处理，以便将原始输入数据压缩为可管理的子聚类集合。第二步使用层级聚类方法将子聚类一步一步合并为更大的聚类。TwoStep 具有一个优点，就是能够为训练数据自动估计最佳聚类数。它可以高效处理混合的字段类型和大型的数据集。有关详细信息，请参阅第 302 页码第 11 章中的两步聚类节点。

自动聚类节点丢弃选项

使用自动聚类节点的“丢弃”选项卡，可以自动丢弃不满足某些条件的模型。这些模型将不会列在模型块中。

图片 5-15
自动聚类节点：“丢弃”选项卡

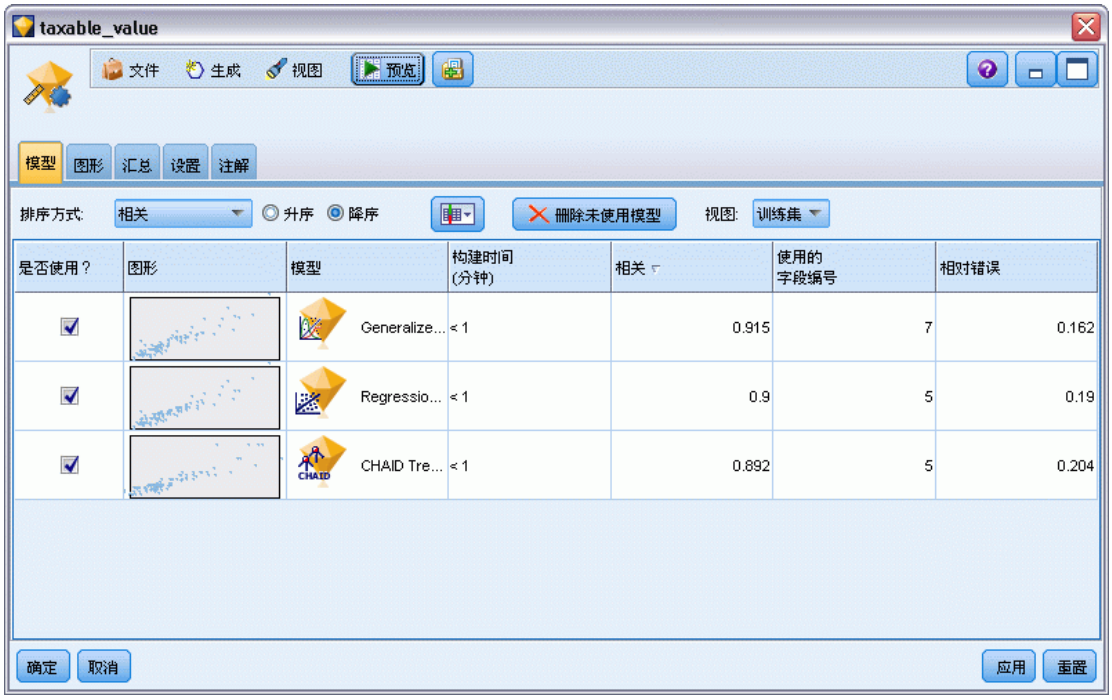
您可以指定最小 silhouette 值、聚类数、聚类大小和模型中所用评估字段的重要性。应根据建模节点中的指定来决定 Silhouette 以及聚类的数量和大小。 有关详细信息，请参阅第 96 页码自动聚类节点模型选项。

或者，可以将节点配置为在首次生成满足所有指定标准的模型时停止执行。 有关详细信息，请参阅第 81 页码自动建模节点停止规则。

自动模型块

执行自动建模节点时，节点评估每个可能选项组合的候选模型，基于您指定的测量为每个候选模型排序，并将最佳模型保存在复合自动模型块中。此模型块实际上包含该节点生成的一个或多个模型的集合，其中模型可单独被浏览或选中用于评分。每个模型列有模型类型和构建时间，以及适合该模型类型的多个其他测量。可以按照这些列中的任意一列对表进行排序，以便快速确定最关注的模型。

图片 5-16
自动数值结果



- 要浏览任何一个单独的模型块，请双击模型块图标。然后，可以从这里生成该模型的建模节点到流工作区，或生成模型块副本到模型选项板。
- 使用缩略图图形可以快速而直观地评估每个模型类型，总结如下。可以双击缩略图生成标准大小的图形。标准大小的散点图可以最多显示 1000 个点并且会在数据集包含更多点时基于样本。（仅对于散点图，图表每显示一次就重新生成一次，所以上游数据中的任意更改（例如在未选中**设置随机数种子** 时更新随机样本或分区）在每次重新绘制散点图时都会反映出来。）
- 使用工具栏在“模型”选项卡上显示或隐藏特定的列或更改用于对表排序的列。（也可以通过单击列标题更改排序列。）
- 使用“删除”按钮以永久删除任何未用的模型。
- 要重新为列排序，请单击列标题并将该列拖放到所需位置。
- 如果正在使用分区，则可选择查看可应用的训练分区或检验分区的结果。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

特定的列取决于要对比的模型的类型，下文已详细列出。

二元目标

- 对于二元模型，缩略图图表显示实际值的分布和与预测值的交叠，来快速直观地表示每个类别中正确预测的记录条数。
- 排序标准与“自动分类器”建模节点中的选项匹配。 有关详细信息，请参阅第 83 页码自动分类器节点模型选项。

- 对于最大利润，还会报告产生的最大数的百分位数。
- 对于累积提升，可以使用工具栏更改选定的百分位数。

名义目标

- 对于名义（集合）模型，缩略图图表显示实际值的分布和与预测值的交叠，来快速直观地表示每个类别中正确预测的记录条数。
- 排序标准与“自动分类器”建模节点中的选项匹配。 有关详细信息，请参阅第 83 页码自动分类器节点模型选项。

连续目标

- 对于连续（数值范围）模型，将根据每个模型的观测值预测图形散点，从而快速直观地表示模型之间的相关性。对于好的模型，点应趋向于聚集在对角线周围，而不是在整个图形中随机分布。
- 排序标准与“自动数值”建模节点中的选项匹配。 有关详细信息，请参阅第 91 页码自动数值节点模型选项。

聚类目标

- 对于聚类模型，将根据每个模型的聚类计算图形散点，从而快速直观地表示聚类分布。
- 排序标准与“自动聚类”建模节点中的选项匹配。 有关详细信息，请参阅第 96 页码自动聚类节点模型选项。

选择评分模型

使用？列可选择评分中使用的模型。

- 对于二元、名义和数字目标，您可以选择多个评分模型，并将得分组合在一个整体模型块中。通过结合多个模型的预测，可以避免单个模型的局限性，使所得的整体准确性通常比从任一模型中获得的准确性要高。
- 对于聚类模型，一次只能选择一个评分模型。默认情况下，首先选择顶级模型。

生成节点和模型

可以从复合自动模型块的构建位置生成其副本，或自动建模节点。例如，当您没有从中构建自动模型块的原始流时，这可能非常有用。此外，还可以为自动模型块中列出的任何单独模型生成模型块或建模节点。

自动建模块

▶ 从?生成?菜单中，选择**模型至选项板**将自动模型块添加到模型选项板上。可对生成的模型进行保存，或者在不重新运行流的情况下使用它。

▶ 或者，可以从“生成”菜单中选择**生成建模节点**以便将建模节点添加到流工作区。可以不用重复完整的建模运行，而使用此节点重新估计选定的模型。

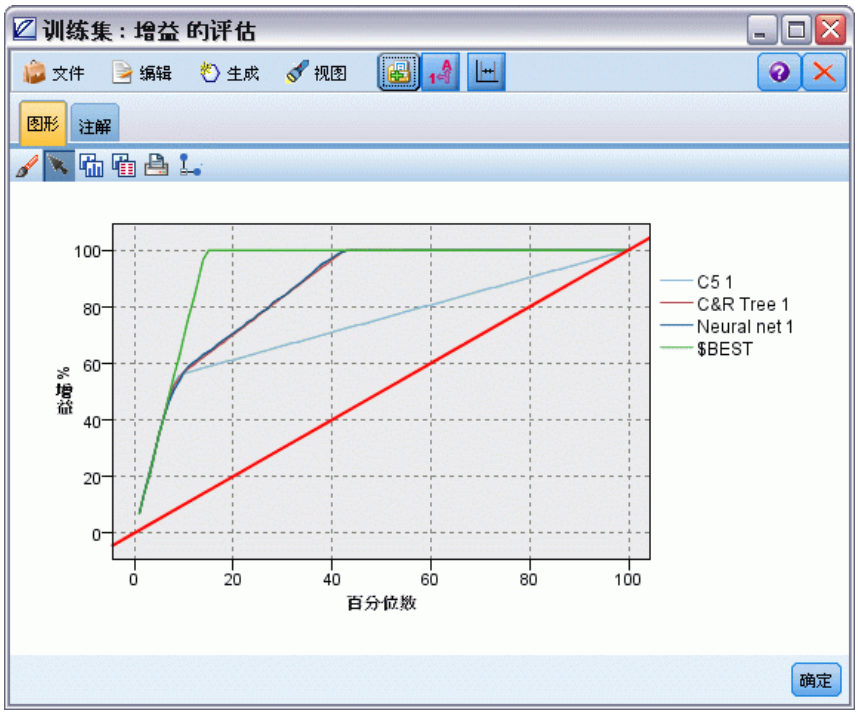
单独模型块

- 在**模型**菜单中，双击所需的单独模型块。块副本在新的对话框中打开。
- 从新对话框中的“生成”菜单中，选择**模型至选项板**将单独建模块添加到模型选项板上。
- 或者，可以从新对话框中的“生成”菜单中选择**生成建模节点**以便将单独建模节点添加到流工作区。

生成评估图表

对于二元模型，可以生成评估图表以直观评价和对比每个模型的性能。评估图表不适用于自动数值或自动聚类节点生成的模型。 有关详细信息，请参阅第 5 章中的评估节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

图片 5-17
具有最佳线和基线的响应图表（累积）



- 在自动分类器自动模型块的使用？列下，选择要评估的模型。
- 从“生成”菜单中，选择**评估图表**。

图片 5-18
生成评估图表

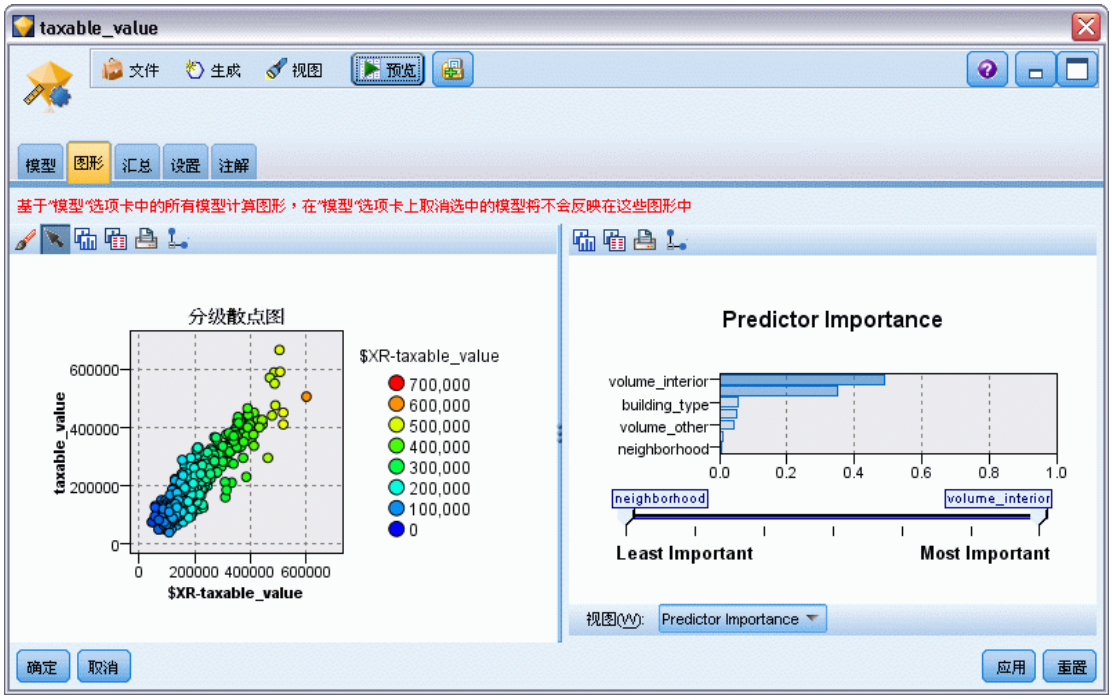
- 选择图表类型和其他需要的选项。 有关详细信息，请参阅第 5 章中的评估散点图选项卡中的IBM SPSS Modeler 14.2 源、过程和输出节点。

评估图形

在自动模型块的“模型”选项卡上，可以向下浏览以显示所示每个模型的单独图形。对于自动分类器和自动数值块，“图形”选项卡同时显示反映所有模型组合结果的图形和预测变量重要性。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

对于自动分类器，则显示分布图；而对于自动数值则显示多重散点图（也称为“散点图”）。有关详细信息，请参阅第 5 章中的通用图形节点功能中的IBM SPSS Modeler 14.2 源、过程和输出节点。

图片 5-19
自动数值 - 自动模型块中整体模型的多重散点图形



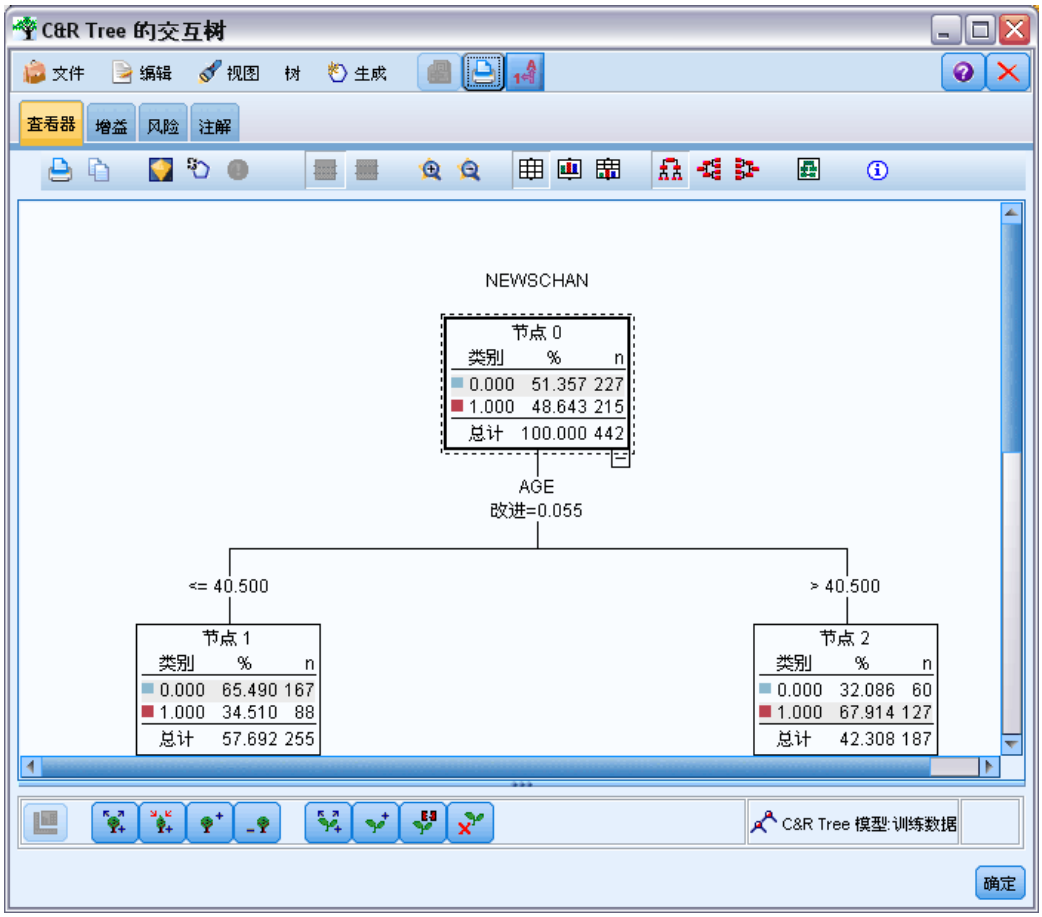
章 6

决策树

决策树模型

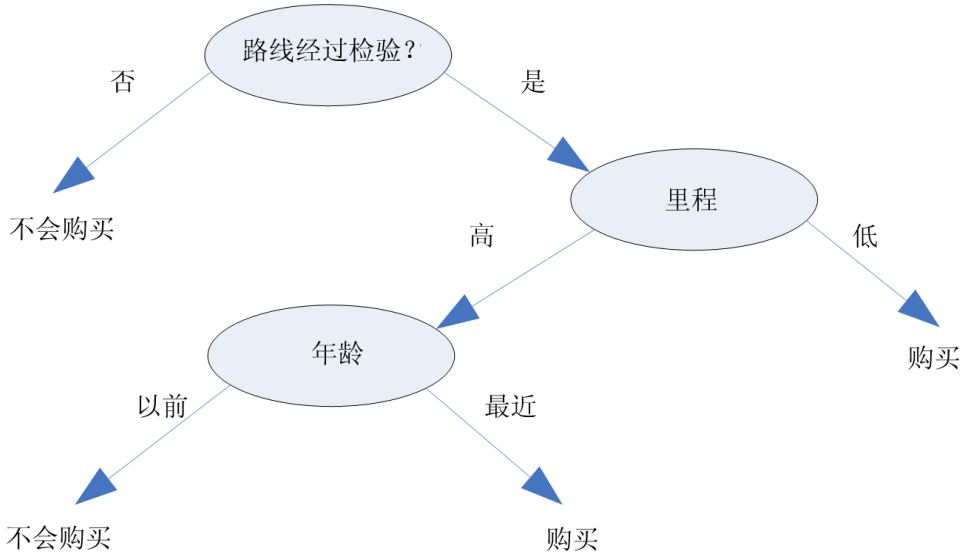
决策树模型允许您开发分类系统，此分类系统可以基于一组决策规则来预测或分类未来的观测值。如果将数据分成您关注的类别（例如，高风险和低风险贷款、用户和非用户、投票人和非投票人或细菌类型），则您可以使用自己的数据来构建规则，借此对新案例或旧案例进行准确性最大的分类。例如，可以基于年龄和其他因素构建对信用风险或购买意向进行分类的树。

图片 6-1
交互树窗口



此方法（有时称为**规则归纳**）有若干优点。首先，浏览树的同时可以明显地看出模型背后的推论过程。这与其它“黑箱”建模技术不同的地方，在其它“黑箱”建模技术中，你很难了解其内部逻辑。

图片 6-2
购买车的简单决策树



其次，此过程将在其规则中自动包含仅能真正影响决策的属性。不会提高树的准确性的属性将被忽略。此方法可获得非常有用的数据信息，并且可用于在培训其他学习方法（如神经网络）之前将数据缩减到相关字段。

决策树模块可转换成 if-then 规则的集合（**规则集**），在多数情况下此规则集以更为复杂的形式显示信息。决策树表示法可以让您知道数据属性是如何将总体**分割**或**分区**成与问题相关的子集。规则集表示法可以让您知道特定项目组与具体结论是如何关联的。例如，以下规则就给出了关于值得购买的一组汽车的**概要**：

```
IF tested = 'yes'
AND mileage = 'low'
THEN -> 'BUY'.
```

树构建算法

四种算法可用于执行分类和段分析。这些算法的执行内容基本相同 - 检查数据集中的所有字段，以找出一个字段，该字段可通过将数据分割成若干子组实现最佳分类或预测。此过程将重复应用以将子组分割成越来越小的单位，直到树结束生长（由特定的停止条件所定义）。构建树的过程中所用的目标和输入字段可以是连续（数字范围）或分类（这取决于所采用的算法）。如果使用的是连续目标，则生成回归树；如果使用的是分类目标，则生成分类树。

分类和回归 (C&R) 树节点生成可用于预测或分类未来观测值的决策树。该方法通过在每个步骤最大限度降低不纯洁度，使用递归分区来将训练记录分割为组。如果节点中 100% 的观测值都属于目标字段的一个特定类别，则树中的该节点将被认定为“纯洁”。目标和输入字段可以是数字范围或分类（名义、有序或标志）；所有分割均为二元分割（即仅分割为两个子组）。 有关详细信息，请参阅第 130 页码C&R 树节点。

CHAID 使用卡方统计量来生成决策树，以确定最佳的分割。CHAID 与 C&R 树和 QUEST 节点不同，它可以生成非二元树，这意味着有些分割将有多于两个的分支。目标和输入字段可以是数字范围（连续）或分类。Exhaustive CHAID 是 CHAID 的修正版，它对所有分割进行更彻底的检查，但计算时间比较长。 有关详细信息，请参阅第 130 页码CHAID 节点。

QUEST 节点可提供用于构建决策树的二元分类法，此方法的设计目的是减少大型 C&R 树分析所需的处理时间，同时也减少在分类树方法中发现的趋势以便支持允许有多个分割的输入。输入字段可以是数字范围（连续），但目标字段必须是分类。所有分割都是二元的。 有关详细信息，请参阅第 131 页码QUEST 节点。

C5.0 节点构建决策树或规则集。该模型的工作原理是根据在每个级别提供最大信息收获的字段分割样本。目标字段必须为分类字段。允许进行多次多于两个子组的分割。 有关详细信息，请参阅第 146 页码C5.0 节点。

基于树的分析的一般用法

以下为一些基于树的分析的若干用法：

分段。 识别出可能成为特定分类的成员的人员。

层次。将案例归入若干类别中的一种，例如高、中和低风险组。

预测。 创建规则并用其预测未来事件。预测还可能意味着尝试将预测属性与连续变量值相关联。

数据缩减和 变量筛选。从 大型变量集合中选择有用的预测变量子集以构建正式的参数模型。

交互 识别。识别那些只适用于 具体子组的关系并在正式的参数模型中指定这些关系。

类别合并 和带状化连续变量。对组预测变量 类别和连续变量以信息丢失最少的方式进行重编码。

交互树构建器

可以自动生成树模型，然后运用算法选择每一级的最佳分割，也可以使用交互树构建器来控制模型的生成，并在保存模型块之前运用专业知识精练或简化树。

- 创建流并添加以下任一决策树节点：C&R 树、CHAID 或 QUEST。

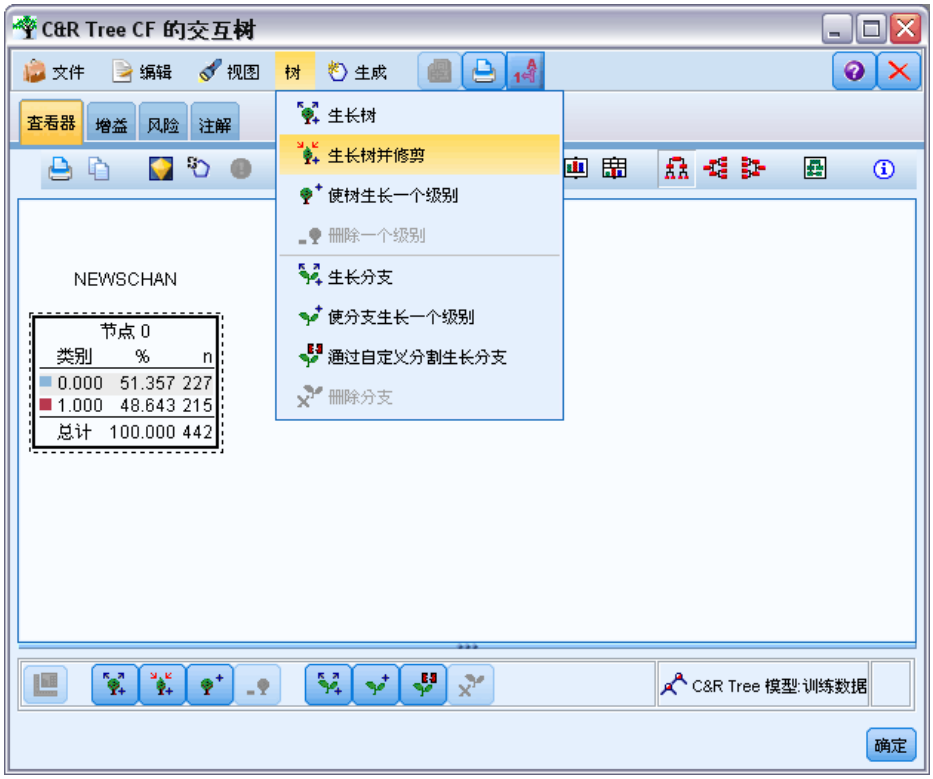
  注意：C5.0 树不支持使用交互树构建。

- 打开节点，在“字段”选项卡上，选择目标和预测变量字段，并在需要时指定其他模型选项。有关具体说明，请参阅各树构建节点文档。

- 在“构建选项”选项卡的“目标”面板上，选择**启动交互会话**。

- 单击**运行**以启动树构建器。

图片 6-3
交互树构建器窗口



其中显示了从根节点开始的当前树。可以逐层编辑和修剪树，并在生成一个或多个模型之前访问收益、风险和相关的信息。

注释

- 使用 C&R 树、CHAID 和 QUEST 节点时，模型中使用的所有有序字段的存储类型都必须是数字（而非字符串）。必要的话，可以使用重新分类节点对存储类型进行转换。 有关详细信息，请参阅第 4 章中的重新对节点分类中的IBM SPSS Modeler 14.2 源、过程和输出节点。
- 还可以选择使用分区字段将数据分隔到训练样本和测试样本中。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。
- 作为使用树构建器的另一种替代方法，也可以直接从建模节点中生成树模型或其他 IBM® SPSS® Modeler 模型。 有关详细信息，请参阅第 127 页码直接构建树模型。

生成和修剪树

使用树构建器的“查看器”选项卡可以查看从根节点开始的当前树。

- 要生成树，请从菜单中选择：
  树 > 生长树

  系统将通过递归分割每个分支直到符合一个或多个停止标准来构建树。然后，可根据使用的建模方法在每个分割处自动选择最合适的预测变量。

- 也可以选择**生成树的第一层**添加一个层。

- 要在一个特定节点下添加分支，可选择该节点，然后选择**生成分支**。

- 要选择某个分割所使用的预测变量，请选择所需的节点，然后选择**使用自定义分割生成分支**。 有关详细信息，请参阅第 110 页码定义自定义分割。

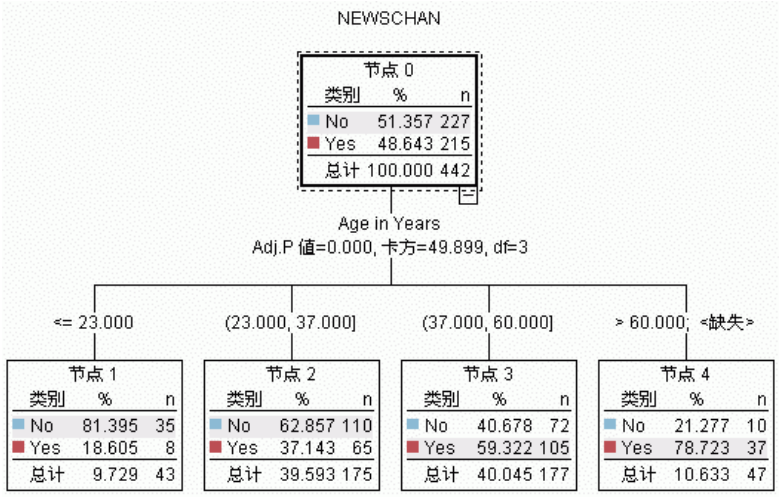
- 要修剪分支，可选择某个节点，然后选择**移除分支**以清除所选择的节点。

- 要移除树的最底层，可选择**移除第一层**。

- 仅对于 C&R 树和 QUEST 树，可选择**生成树和修剪**根据成本复杂性算法（此算法可根据终端节点数调整风险评估）进行修剪，通常会生成一个较简单的树。 有关详细信息，请参阅第 130 页码C&R 树节点。

在查看器选项卡上读取分割规则

图片 6-4
查看器选项卡上显示的分割规则



查看“查看器”选项卡上的分割规则时，方括号表示临界值包含在范围中，而圆括号表示临界值不包含在范围中。因此，表达式 (23, 37] 表示从 23（不含）到 37（含）；即从 23 以上到 37。在“模型”选项卡上，相同的情况显示为：

```
Age > 23 and Age <= 37
```

中断树增长。要中断树增长操作（例如，如果此操作所用的时间比预期的长），可单击工具栏上的“停止执行”按钮。

图片 6-5
“停止执行”按钮

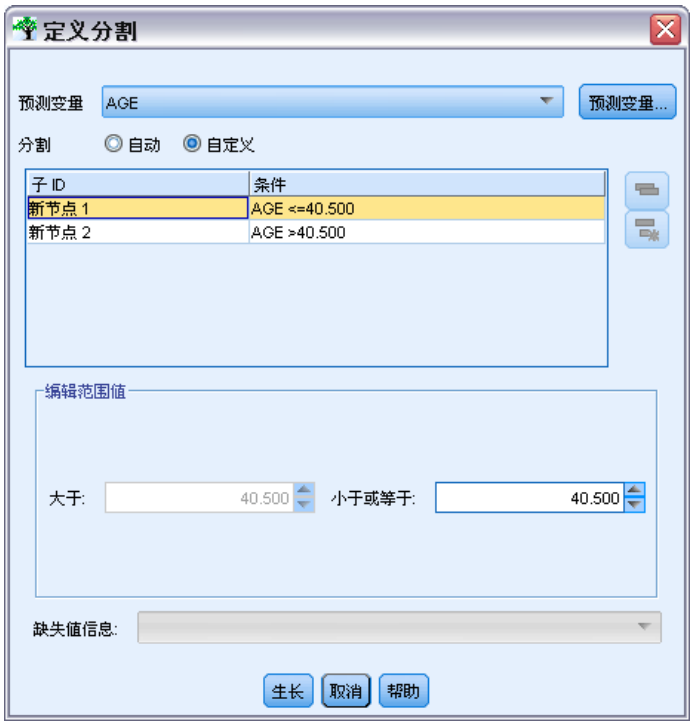
此按钮仅在树增长期间启用。它会使当前的增长操作停止在其当前点上，保留所有已添加的节点，但不保存所做的更改，也不关闭该窗口。树构建器将保持打开状态，以便生成模型、更新指令，或在需要时以适当的格式导出输出。

定义自定义分割

通过“定义分割”对话框，可以选择预测变量并为每个分割指定条件。

- 在树构建器的“查看器”选项卡上选择一个节点，然后从菜单中选择：
  树 > 使用自定义分割生成分支

图片 6-6
“定义分割”对话框



- 从下拉列表中选择所需的预测变量，或单击**预测变量**按钮，以查看每个预测变量的详细信息。 有关详细信息，请参阅第 111 页码查看预测变量详细信息。
- 可接受为每个分割选择的默认条件，或选择**自定义**为分割指定适当的条件。
  - 对于连续（数值范围）的预测变量，可以使用**编辑范围值**字段以指定落在每个新节点中的值的范围。
  - 对于分类预测变量，可使用**编辑集合值**或**编辑有序值**字段，以指定映射到每个新节点的特定值（如果是有序预测变量，则指定值的范围）。

- 选择**生成**，使用选定的预测变量重新生成分支。

在不考虑停止规则的情况下，通常可使用任何预测变量分割树。唯一的例外情况是当节点是纯节点（即所有观测值都落在相同的目标类中，从而没有可分割的观测值），或所选择的预测变量是常数（即没有可分割的预测变量）时无法分割树。

缺失值信息。仅对于 CHAID 树，如果给定的预测变量中有缺失值，则可以在定义自定义分割时选择将这些缺失值分配给特定的子节点。（对于 C&R 树和 QUEST，可使用代用项按算法中所定义的方式处理缺失值。 有关详细信息，请参阅第 111 页码分割的详细信息和代用项。）

查看预测变量详细信息

“选择预测变量”对话框中显示了可用于当前分割的预测变量（有时称为“代替变量”）的统计量。

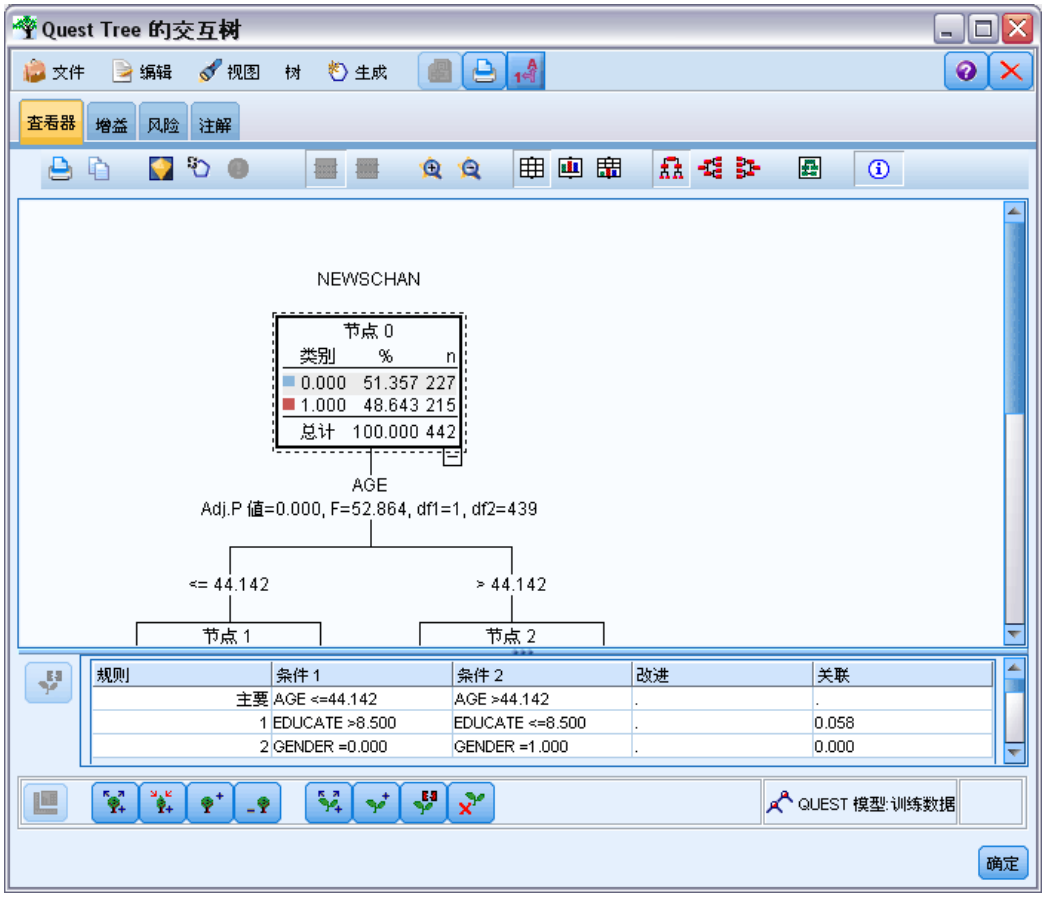
图片 6-7
选择预测变量对话框

- 对于 CHAID 和 Exhaustive CHAID，列出了每个分类预测变量的卡方统计量；如果预测变量是数字范围类型，则显示 F 统计量。卡方统计量可用来测量目标字段与分割字段的不相关程度。较高的卡方统计量通常与较低的概率有关，这意味着两个字段间不相关的机率较低 – 表示此分割情况良好。这里也将自由度包括在内，因为自由度考虑了以下事实，即与双向分割相比，三向分割更易具有较高的统计量和较低的概率。
- 对于 C&R 树和 QUEST，显示了每个预测变量的改进值。如果使用此预测变量，则改进值越大，父节点和子节点间的纯度差异越大。（纯节点指其中所有的观测值都落在一个目标类别中的节点；树中的杂质越少，此模型拟合数据的效果就越好。）换句话说，较高的改进值通常表示对此类型的树进行了有用的分割。所使用的杂质测量在树构建节点中指定。

分割的详细信息和代用项

可在“查看器”选项卡中选择任意节点，然后选择位于工具栏右侧的分割信息按钮查看有关该节点的分割详细信息。此时将显示所使用的分割规则及相关的统计量。对于 C&R 树分类树，将显示改进值和关联值。关联值可用于测量代用项与原始分割字段间的一致性，其中“最佳”代用项通常是对分割字段模拟得最像的字段。对于 C&R 树和 QUEST，也列出了所有用于代替原始预测变量的代用项。

图片 6-8
显示分割信息的交互树构建器窗口



- 要编辑选定节点的分割，可单击位于代用项面板左侧的图标以打开“定义分割”对话框。（作为快捷方式，可以在单击图标选择代用项作为原始分割字段之前，从列表中选择此代用项。）

  代用项。如果适用，则会针对所选节点显示主要分割字段的所有代用项。代用项是在给定记录的主要预测变量值缺失时使用的替代字段。给定分割允许的最大代用项数在树构建节点中指定，但实际数量取决于训练数据。一般来讲，缺失数据越多，可能使用的代用项越多。对于其他决策树模型，此选项卡为空。

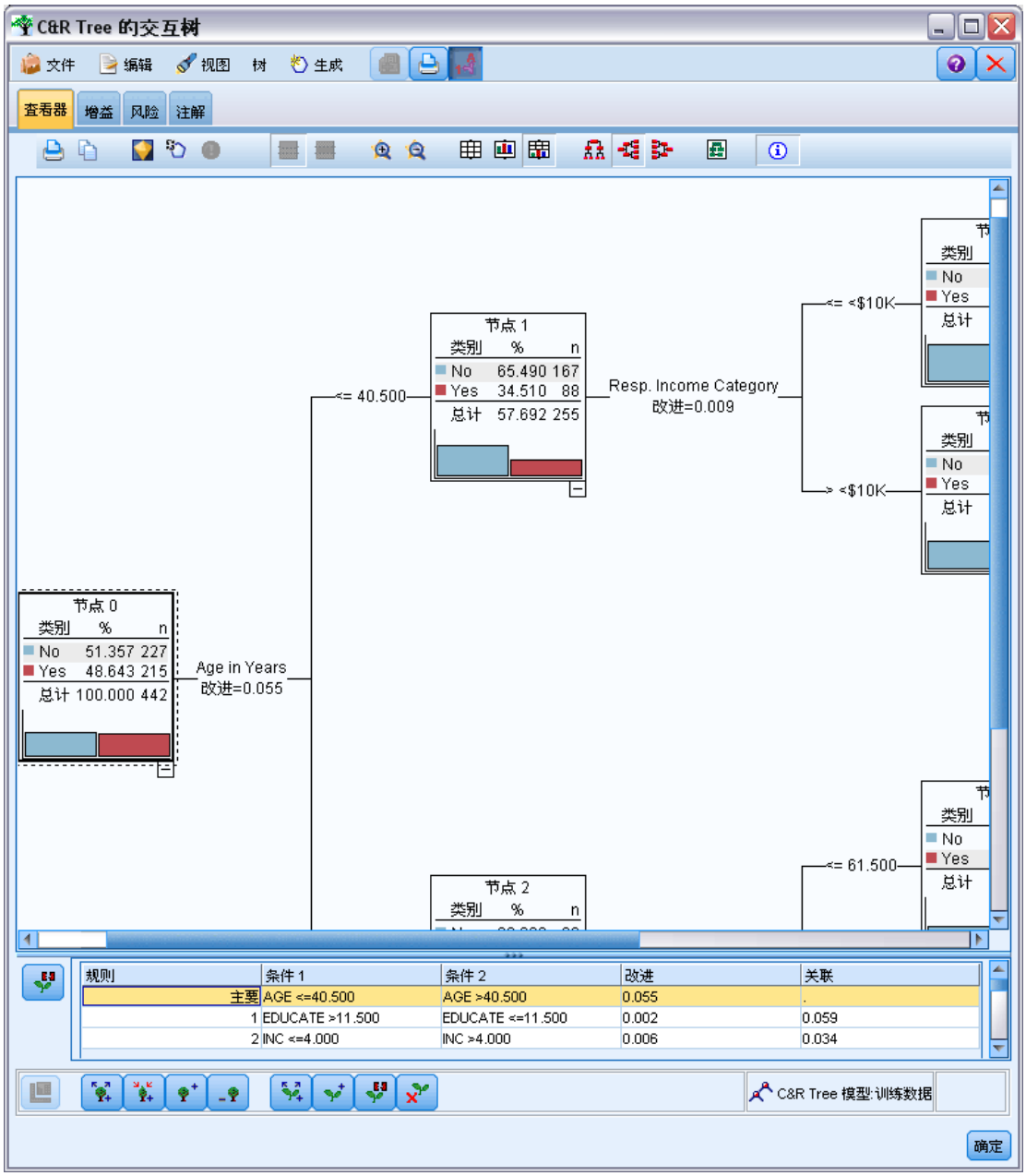
  注：要在模型中包含代用项，必须在训练阶段对其进行标识。如果训练样本没有缺失值，则不会标识任何代用项；在测试或评分过程中遇到的具有缺失值的所有记录将自动落入记录数最大的子节点。如果在测试或评分过程中预期出现缺失值，请确保值在训练样本中也处于缺失状态。代用项对于 CHAID 树不可用。

  虽然 CHAID 树中不使用代用项，但当定义自定义分割时，仍可选择将这些代用项分配给特定的子节点。 有关详细信息，请参阅第 110 页码定义自定义分割。

自定义树状视图

在树构建器的“查看器”选项卡中显示当前的树。默认情况下，将展开树中所有的分支，但也可以按照需要展开和折叠分支并自定义其他设置。

图片 6-9
由左至右分别显示分割规则详细信息、节点图形和标签的视图



- 单击父节点右下角的减号（-）隐藏其所有子节点。单击父节点右下角的加号（+）显示其子节点。
- 使用“视图”菜单或工具栏更改树的方向（由上至下、由左至右或由右至左）。

- 单击主工具栏上的“显示字段和值标签”按钮可以显示或隐藏字段和值标签。
- 使用放大镜按钮放大或缩小视图，或单击工具栏右侧的树状图按钮查看完整的树图表。
- 如果正在使用分区字段，则可在树状视图的训练分区和测试分区之间进行交换（选择**视图** > **分区**）。显示测试样本时，可以查看但不能编辑树。（将在窗口右下角的状态栏中显示当前分区。）
- 单击分割信息按钮（工具栏最右侧的“i”按钮）以查看当前分割的详细信息。 有关详细信息，请参阅第 111 页码分割的详细信息和代用项。
- 将在每个节点中显示统计量、图形或同时显示两者（请参阅下文）。

显示统计量和图形

节点统计量。对于分类目标字段，每个节点中的表可显示每个分类中的记录数和百分比以及该节点代表的整个样本的百分比。对于连续（数值范围）目标字段，该表可显示目标字段的平均值、标准差、记录数和预测值。

节点图形。对于分类目标字段，图形为目标字段的每个类别中的百分比条形图。表中每行的前面是一个颜色样本，其对应的颜色表示该节点图形中的每个目标字段类别。对于连续（数值范围）目标字段，该图形可显示该节点中记录的目标字段的直方图。

Gains

“收益”选项卡可显示树中所有终端节点的统计量。收益可用于测量给定节点上的平均值或比例与总平均值之间的差异大小。一般来说，此差异越大，作为决策工具的树就越有效。例如，某个节点的指数或“提升”值为 148% 表示，该节点中的记录落在目标类别中的可能性大概是其作为一个整体用于数据集的可能性的 1.5 倍。

对于 C&R 树和指定过度拟合集的 QUEST 节点，显示两组统计信息：

- 树生长组 – 移除过度拟合集的培训样本
- 过度拟合集

对于其他 C&R 树和 QUEST 交互树以及所有 CHAID 交互树，只显示树生长组统计信息。

图片 6-10
“收益”选项卡

使用“收益”选项卡可以执行下列操作：

- 显示每个节点统计量、累积数统计量或分位数统计量。

- 显示收益或利润。
- 将视图在表和图表间进行交换。
- 选择目标类别（仅分类目标）。
- 根据指数百分比对表按升序或降序排序。如果显示的是多个分区的统计量，则通常将排序应用于训练样本而不是测试样本。

一般来说，在收益表中选定的内容也会在树状视图中得到更新，反之亦然。例如，如果在表中选择某个行，则也会在树中选中相应的节点。

分类收益

对于分类树（指使用分类目标变量的树），从收益指数百分比可看出每个节点上给定目标类别的比例与总比例间的差异有多大。

依次显示节点统计量

在此视图的表中，将为每个终端节点显示一行。例如，如果直邮活动的总响应是 10%，但有 20% 的记录落在节点 X 内并且做出积极的响应，则该节点的指数百分比应为 200%，表示该组中的响应者进行购买的可能性大概是总人数的两倍。

对于 C&R 树和指定过度拟合集的 QUEST 节点，显示两组统计信息：

- 树生长组 – 移除过度拟合集的培训样本
- 过度拟合集

对于其他 C&R 树和 QUEST 交互树以及所有 CHAID 交互树，只显示树生长组统计信息。

图片 6-11
依次显示节点收益统计量

节点。当前节点的 ID（显示在“查看器”选项卡上）。

节点：n。该节点上的总记录数。

节点 (%)。数据集中所有落在该节点上的记录的百分比。

收益：n。落在该节点上的选定目标类别的记录数。换句话说，在数据集的所有落在目标类别的记录中，有多少记录落在该节点？

收益 (%)。在整个数据集中，所有落在该节点的目标类别中的记录的百分比。

响应（%）。落在当前节点的目标类别中的记录的百分比。该上下文中的响应有时也称为“匹配项”。

指数（%）。当前节点的响应百分比，可表述为响应百分比相对于整个数据集的百分比。例如，指数值为 300% 表示该节点中的记录落在目标类别中的可能性大概是其作为一个整体用于数据集的可能性的三倍。

累积统计量

在累积视图中，表的每行显示一个节点，但统计量是累积的，并按指数百分比以升序或降序顺序排序。例如，如果按降序排序，则首先列出指数百分比最高的节点，并且接下来的行中的统计量是对该行及上面的行的累积数。

图片 6-12
以指数百分比的降序顺序排序的累积收益

节点	节点:n	节点 (%)	收益:n	收益 (%)	响应 (%)	索引 (%)
2	73.00	48.34	45.00	64.29	61.64	132.97
1	151.00	100.00	70.00	100.00	46.36	100.00

节点	节点:n	节点 (%)	收益:n	收益 (%)	响应 (%)	索引 (%)
2	37.00	52.86	23.00	69.70	62.16	131.86
1	70.00	100.00	33.00	100.00	47.14	100.00

随着所添加节点的响应百分比越来越低，累积指数百分比将逐行降低。最后一行的累积指数通常是 100%，因为此时将包括整个数据集。

分位数

在此视图中，表中的每一行都表示一个分位数而不是节点。分位数可以是四分位数 (4)、五分位数 (5)、十分位数 (10)、二十分位数 (20) 或百分位数 (100)。如果需要多个节点以补足此百分比（例如，如果显示四分位数时，而前两个节点包含的观测值不到所有观测值的 50%），则可在一个分位数中列出多个节点。可以对表的其余部分进行累积，且与累积视图的解释方式相同。

图片 6-13
以指数百分比的降序顺序列出显示四分位数的收益

节点	百分位数	百分位数:n	收益:n	收益 (%)	响应 (%)	索引 (%)
2	25.00	38.00	23.00	33.46	61.64	132.97
2,1	50.00	76.00	46.00	65.66	60.48	130.45
1	75.00	113.00	58.00	82.60	51.17	110.38
1	100.00	151.00	70.00	100.00	46.36	100.00

节点	百分位数	百分位数:n	收益:n	收益 (%)	响应 (%)	索引 (%)
2	25.00	18.00	11.00	33.91	62.16	131.86
2	50.00	35.00	22.00	65.93	62.16	131.86
2,1	75.00	53.00	28.00	84.39	52.54	111.46
1	100.00	70.00	33.00	100.00	47.14	100.00

分类利润和投资回报率

对于分类树，收益统计量也可按利润和投资回报率显示。通过“定义利润”对话框可以为每个类别指定收入和支出。

- ▶ 在“收益”选项卡上，单击工具栏上的“利润”按钮（标记为 $/$）访问该对话框。

图片 6-14
定义“利润”对话框

- ▶ 为目标字段的每个类别输入收入和支出值。

例如，如果为每个客户邮寄报价的成本是 $0.48，而从接受三个月的订阅的积极响应中获得的收入是 $9.95，则每个 no 响应将花费 $0.48，而每个 yes 响应将赚取 $9.47（计算方式为 9.95 - 0.48）。

在收益表中，**利润**的计算方式为终端节点的每个记录中的总收入减去支出。**投资回报率**为某个节点的总利润除以总支出。

注释

- 利润值仅影响在收益表中显示的平均利润和投资回报率，可以明确查看统计量，尤其适合查看利润。但它们不影响树模型的基础结构。不应将利润与误分类损失相混淆，误分类损失在树构建节点中指定，且可化为模型中的因子（作为避免高成本错误的一种方式）。
- 在两个交互树构建会话之间不会保留利润说明。

回归收益

对于回归树，可以选择依次显示节点视图、累积节点视图和分位数视图。表中可显示平均值。只有在分位数视图中才可使用图表。

收益图表

在“收益”选项卡上，图表可作为表的替代项显示。

- ▶ 在“收益”选项卡上，选择“分位数”图标（工具栏从左数第三个图标）。（对于依次显示节点统计量或累积统计量，不可使用图表。）
- ▶ 选择“图表”图标。
- ▶ 按照需要从下拉列表中选择所显示的单位（百分位数、十分位数等等）。

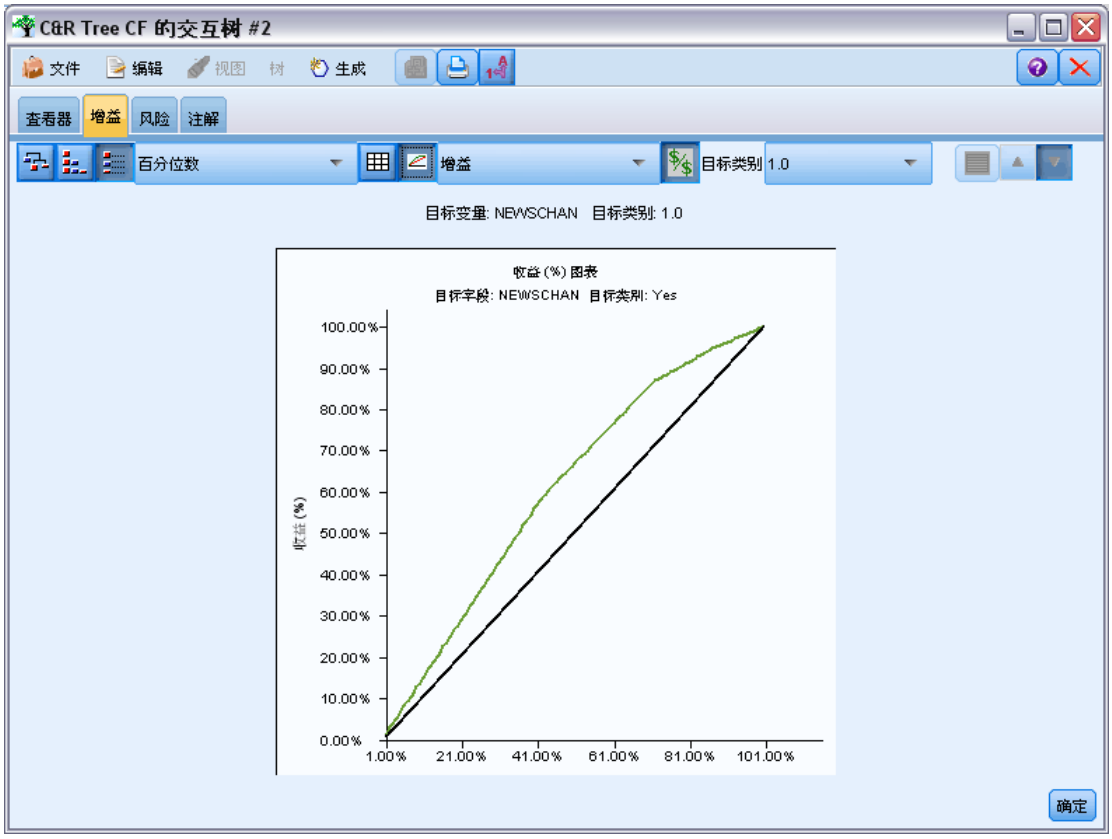
▶ 选择**收益**、**响应**或**提升**更改所显示的测量量。

收益图表

收益图绘制的是表中收益（%）列值的散点图。收益定义为每个增量中匹配项数与树中匹配项总数的比例，它使用下列等式：

（增量中匹配项数/匹配项总数）x 100%

图片 6-15
收益图



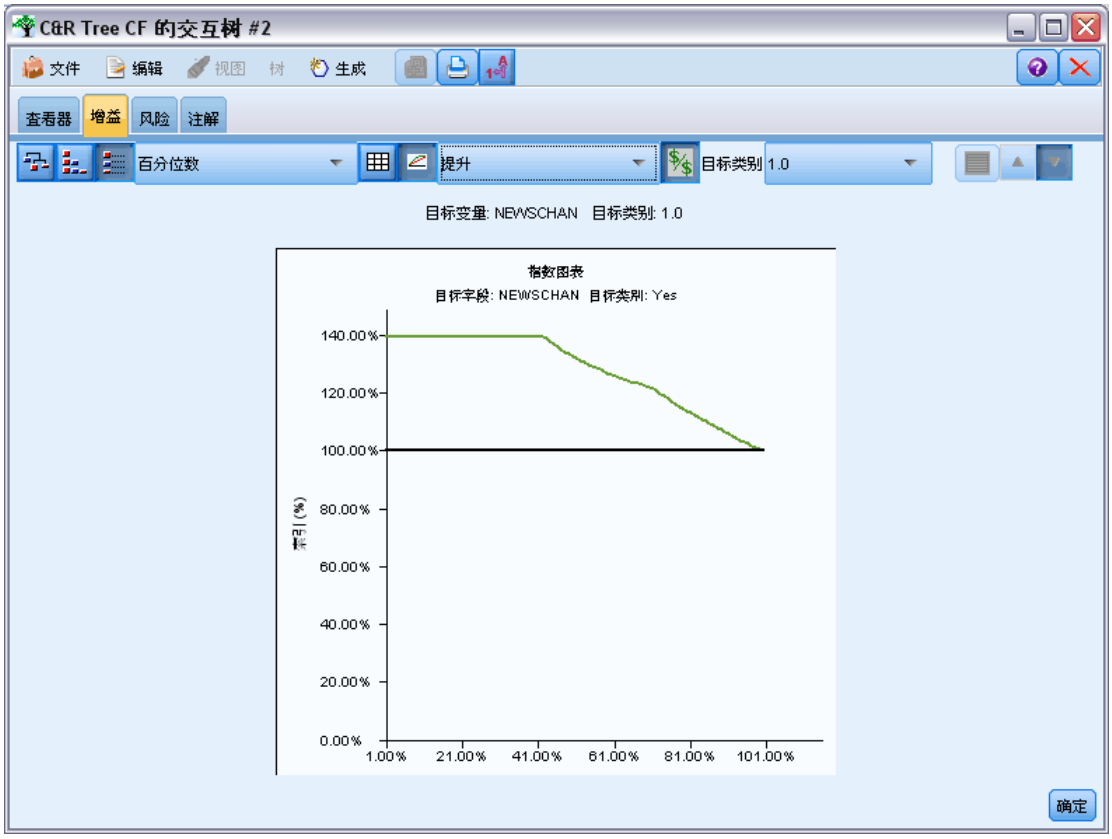
该图有效说明了您需要撒出多大范围的网络，才能获取树中所有匹配项的给定百分比。对角线绘制的是整个样本的预期响应（如果未使用模型的话）。这种情况下，响应率应该为常量，因为一个人响应的可能性与另一个人相同。为了使您的收益加倍，您需要询问两倍数量的人。曲线表明通过将那些秩（基于收益排序）位于较高百分比的人员包括在内，您可以使得响应得到多大程度的改善。例如，包括最高的 50% 可能会网罗超过 70% 的正面响应。该曲线越陡，收益越高。

提升图表

提升图表对表中指数（%）列中的值进行了绘制。此图表将每个增量中具有积极响应的记录的百分比与训练数据集中具有积极响应的记录的总百分比作了比较，其方程式为：

（增量中具有积极响应的记录/增量中的记录）/（具有积极响应的总记录数/总记录数）

图片 6-16
增益图

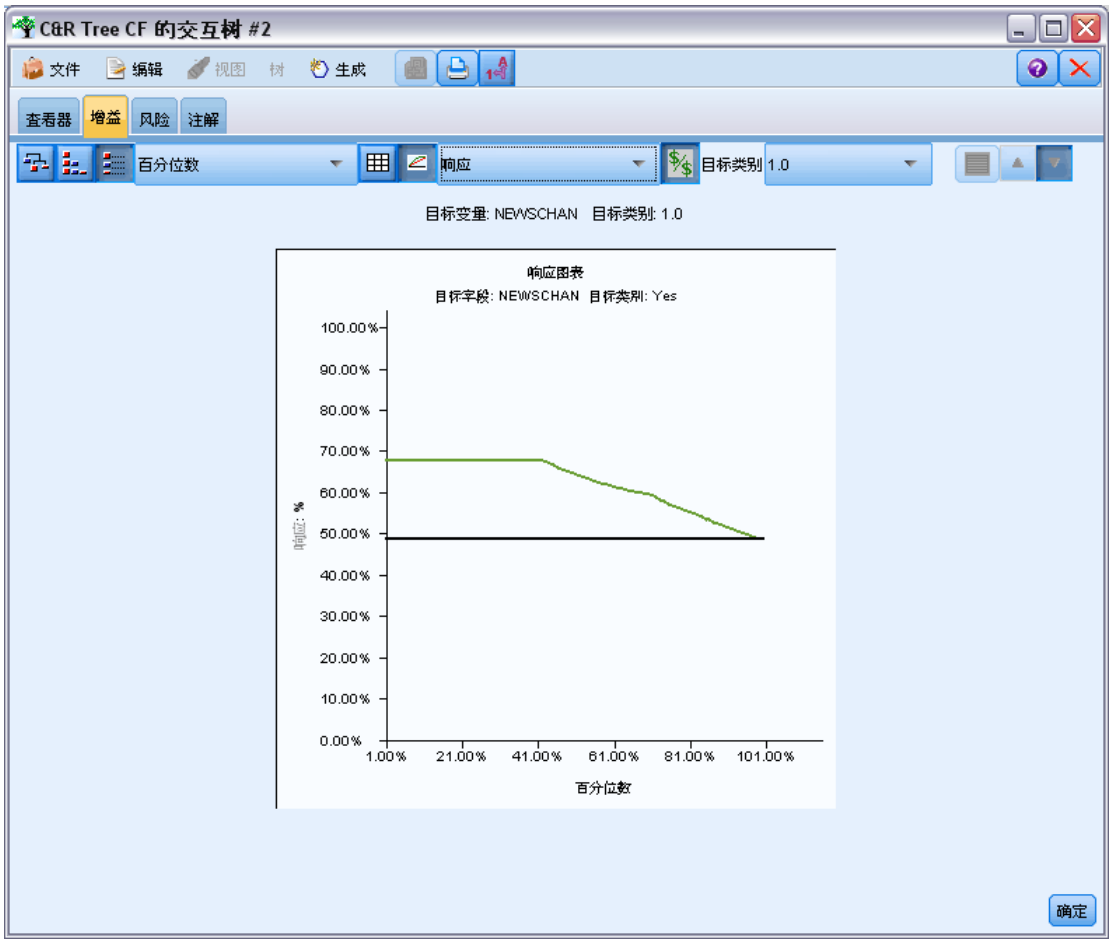


响应图表

响应图表对表中响应（%）列中的值进行了绘制。响应是增量中具有积极响应的记录的百分比，其方程式为：

（增量中具有积极响应的记录/增量中的记录） x 100%

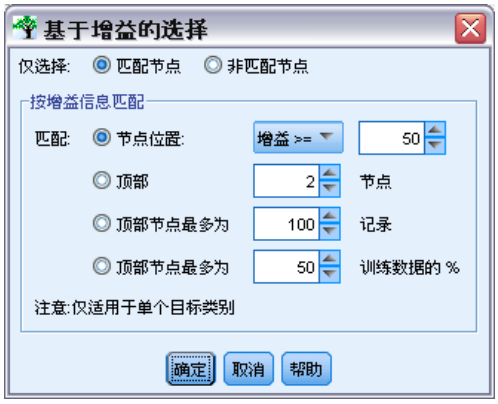
图片 6-17
响应图表



基于收益的选择

使用“基于收益的选择”对话框，可以根据指定的规则或阈值自动选择具有最佳（或最差）收益的终端节点。然后可以根据该选择生成一个选择节点。

图片 6-18
基于收益的选择对话框



- 在“收益”选项卡上，选择依次显示节点视图或累积视图，然后选择该选择所基于的目标类别。（该选择基于当前的表显示，不可用于分位数视图。）
- 从“收益”选项卡的菜单中选择以下项：
  编辑 > 选择终端节点 > 基于收益的选择

  仅选择。可以选择匹配节点或不匹配节点 — 例如，选择前 100 个记录以外的所有节点。

  收益信息匹配。基于当前目标类别的收益统计量的匹配节点，包括：
  - 其收益、响应或提升（指数）与指定的阈值相匹配的节点—例如，响应大于或等于 50%
  - 基于目标类别的收益的顶部 n 个节点。
  - 上限为指定记录数的顶部节点。
  - 上限为指定训练数据百分比的顶部节点。
- 单击**确定**更新“查看器”选项卡上的选择。
- 要根据“查看器”选项卡上的当前选择新建“选择”节点，请从“生成”菜单中选择**选择节点**。 有关详细信息，请参阅第 126 页码生成过滤节点和选择节点。

  注意：因为实际上选择的是节点而不是记录或百分比，因此通常不可能取得与选择标准完全匹配的结果。系统选择上限为指定等级的完整节点。例如，如果选择顶部 12 个观测值，而第一个节点中有 10 个观测值，第二个节点中有 2 个观测值，则将只选择第一个节点。

风险

风险指任意等级上误分类的机率。“风险”选项卡可显示某点的风险评估和（分类输出的）误分类表。

图片 6-19
分类目标的误分类表

- 对于数字预测，风险是每个终端节点上的合并方差评估。
- 对于分类预测，风险是错误分类观测值的比例，可根据任意先验分布或误分类损失进行调整。

保存树模型和结果

可以用以下多种方式保存或导出交互树构建会话的结果：

- 基于当前树生成模型（**生成** > **生成模型**）。
- 保存用于生成当前树的指令。下次执行树构建节点时，将自动重新生成当前树（包括已定义的任何自定义分割）。
- 导出模型、收益和风险信息。 有关详细信息，请参阅第 125 页码导出模型、收益和风险信息。

通过树构建器或树模型块，可以执行下列操作：

- 根据当前的树生成过滤节点或选择节点。有关详细信息，请参阅第 126 页码生成过滤节点和选择节点。
- 生成一个规则集块，该节点将树结构表示成一组定义了树的终端分支的规则。有关详细信息，请参阅第 126 页码从决策树中生成规则集。

- 此外，还可以按 PMML 格式导出模型（仅限于树模型块）。有关详细信息，请参阅第 41 页码第 3 章中的模型选项板。如果模型包含定制分割，则不会在导出的 PMML 中保留此信息。（保留分割，但不保留它是定制分割而不是通过算法选择的分割这一事实。）
- 基于当前树的所选部分生成图形。注：这仅在块附加到流中的其他节点时才有效。有关详细信息，请参阅第 158 页码生成图形。

注意：不能保存交互树本身。为了避免丢失所执行的操作，请在关闭树构建器窗口之前生成模型和/或更新树指令。

从树构建器生成模型

要基于当前树生成模型，可从树构建器菜单中选择以下项：

生成 > 模型

图片 6-20
生成决策树模型

您可以从下列选项中进行选择：

模型名称。可以指定自定义名称或根据建模节点的名称自动生成模型名称。

创建节点位置。可以在工作区、GM 选项板或同时在这两者上添加节点。

包括树指令。要在生成模型中包括来自当前树的指令，选择此选项。这允许您在需要时重新生成树。 有关详细信息，请参阅第 123 页码树生长指令。

树生长指令

对于 C&R 树、CHAID 和 QUEST 模型，树指令可指定生成树（一次一级）的条件。每当从节点中启动交互树构建器时，都会应用指令。

- 指令可作为一种最安全的方法用来重新生成在以前的交互会话中创建的树。有关详细信息，请参阅第 125 页码更新树指令。也可以手动编辑指令，但操作时需要格外小心。
- 指令与其所描述的树结构高度相关。因此，对原始数据或建模选项的任何更改都可能会导致以前有效的一组指令失效。例如，如果 CHAID 算法基于更新的数据将双向分割更改为三向分割，则基于以前的双向分割的所有指令都将失效。

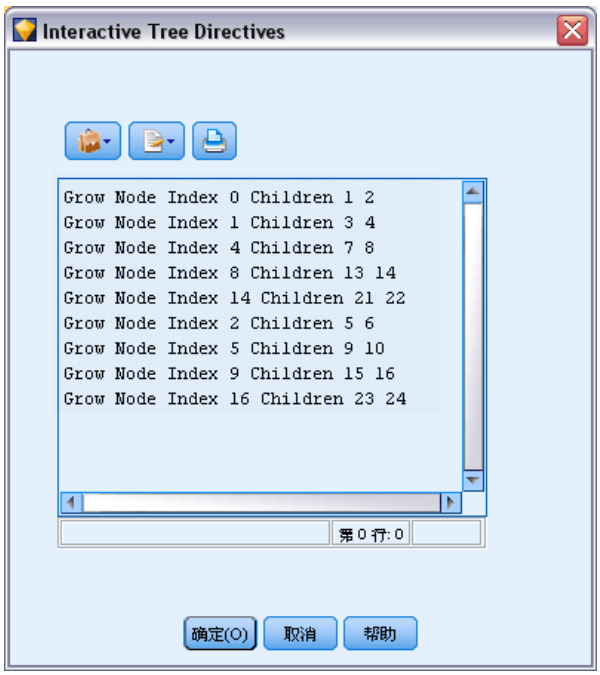
注意：如果选择直接生成模型（不使用树构建器），则将忽略所有的树指令。

编辑指令

▶ 要查看或编辑已保存的指令，请打开树构建节点，并选择“构建选项”选项卡的“目标”面板。

▶ 选择**启动交互会话**以启用控件，选中**使用树指令**，然后单击**指令**。

图片 6-21
树生长指令



指令语法

指令可指定从根节点开始生成树的条件。例如，生成树的第一层：

```
Grow Node Index 0 Children 1 2
```

由于未指定任何预测变量，算法将选择最佳分割。

注意，通常必须在根节点（Index 0）上进行第一次分割，且必须指定两个子节点的索引值（在本例中为 1 和 2）。除非已首先生成创建节点 2 的根节点，否则指定 Grow Node Index 2 Children 3 4 是无效的。

要生成树，请使用：

```
生长树
```

要生成并修剪树（仅 C&R 树），请使用：

```
Grow_And_Prune Tree
```

要为连续预测变量指定自定义分割，请使用：

```
Grow Node Index 0 Children 1 2 Spliton
  ( "EDUCATE", Interval ( NegativeInfinity, 12.5)
     Interval ( 12.5, Infinity ))
```

对具有两个值的名义预测变量进行分割，可使用：

```
Grow Node Index 2 Children 3 4 Spliton
  ( "GENDER", Group( "0.0" )Group( "1.0" ))
```

对具有多个值的名义预测变量进行分割，可使用：

```
Grow Node Index 6 Children 7 8 Spliton
   ( "ORGS", Group( "2.0","4.0" )
             Group( "0.0","1.0","3.0","6.0" ))
```

对有序预测变量进行分割，可使用：

```
Grow Node Index 4 Children 5 6 Spliton
  ( "CHILDS", Interval ( NegativeInfinity, 1.0)
              Interval ( 1.0, Infinity ))
```

注意：指定自定义分割时，字段名和值（EDUCATE、GENDER 和 CHILDS 等等）都区分大小写。

CHAID 树的指令

CHAID 树的指令对数据或模型中的更改非常敏感，因为这些指令与 C&R 树和 QUEST 中的不同，它们不只能使用二元分割。例如，下面的语法看起来很有效，但如果算法将根节点分割为两个以上的子节点时，这些语法将失效：

```
Grow Node Index 0 Children 1 2
Grow Node Index 1 Children 3 4
```

对于 CHAID，节点 0 可能具有 3 个或 4 个子节点，这种情况将使上述第二行语法失效。

在脚本中使用指令

也可使用三重引号将指令嵌入到脚本中。 有关详细信息，请参阅第 3 章中的文字文本块中的IBM SPSS Modeler 14.2 脚本编写 和自动化指南。

更新树指令

要保留在交互树构建会话中执行的操作，可以保存用于生成当前树的指令。与保存不能进一步编辑的模型块不同的是，保存指令可以按树的当前状态重新生成树以进一步进行编辑。

- 要更新指令，请从树构建器菜单中选择以下项：
  文件 > 更新指令

  指令保存在用于创建树（C&R 树、QUEST 或 CHAID）的建模节点中，并可用于重新生成当前树。 有关详细信息，请参阅第 123 页码树生长指令。

导出模型、收益和风险信息

可以从树构建器中根据需要以文本、HTML 或图像格式导出模型、收益和风险统计量。

- 在树构建器窗口中，选择要导出的选项卡或视图。

▶ 从菜单中选择：
文件 > 导出

▶ 根据需要选择**文本**、**HTML** 或**图形**，并从子菜单中选择要导出的特定项目。

在适用的情况下，导出基于当前的选择。

导出文本或 HTML 格式。可以为训练分区或测试分区（如果已定义）导出收益统计量或风险统计量。导出基于“收益”选项卡上的当前选择 - 例如，可以选择依次显示节点统计量、累积统计量或分位数统计量。

导出图形。可以导出在“查看器”选项卡上显示的当前树，或为训练分区或测试分区（如果已定义）导出收益图表。可用的格式包括 .JPEG、.PNG 和 .BMP。对于收益，导出基于“收益”选项卡上的当前选择（仅当显示图表时可用）。

生成过滤节点和选择节点

▶ 在树构建器窗口中，或在浏览决策树模型块时，从菜单中选择以下项：
生成 > 过滤节点

或
> 选择节点

过滤节点。生成的节点可过滤当前树未使用的任何字段。此方法可以快速削减数据集，使其仅包括那些算法选择为重要字段的字段。如果此决策树节点的上游存在“类型”节点，则“过滤”模型块将传递所有角色为目标的字段。

选择节点。生成的节点可选择所有落在当前节点中的记录。此选项需要在“查看器”选项卡中选择一个或多个树分支。

该模型块位于流工作区中。

从决策树中生成规则集

生成的规则集模型块可作为定义树的终端分支的一组规则来表示树的结构。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。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。例如，可以看到所有预测结果为否的规则，紧随其后是所有预测为是的规则。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，则会将默认预测分配到该记录。

仅可从具有分类目标字段的树（不是回归树）中生成规则集。

▶ 在树构建器窗口中，或在浏览决策树模型块时，从菜单中选择以下项：
生成 > 规则集

图片 6-22
“生成规则集”对话框

规则集名称。可以指定新的规则集模型块的名称。

创建节点位置。控制新的规则集模型块的位置。选择**工作区**、**GM 选项板**或**两者**。

最小实例。指定在规则集模型块中保留的最小实例数（已应用规则的记录数）。支持度小于指定值的规则不会包括在新的规则集中。

最小置信度。指定���则集模型块中要保留的规则的最小置信度。置信度小于指定值的规则不会包括在新的规则集中。

直接构建树模型

作为使用交互式树构建器的另一种替代方法，也可以在运行流时直接从节点中构建决策树模型。这与多数其他模型构建节点一致。对于交互树构建器所不支持的 C5.0 树模型来说，这是唯一可以使用的方法。

- 创建流并添加其中一个决策树节点 - C&R 树、CHAID、QUEST 或 C5.0。

图片 6-23
直接构建 C5.0 树

- 对于 C&R 树、QUEST 或 CHAID，在“构建选项”选项卡的“目标”面板上，选择一个主目标。如果您选择“构建单个树”，请确保将“模式”设为**生成模型**。

  对于 C5.0，在“模型”选项卡上，将**输出类型**设为**决策树**。

- 选择目标字段和预测变量字段，并在需要时指定其他模型选项。有关具体说明，请参阅各树构建节点文档。

- 运行流以生成模型。

注释

- 使用此方法生成树时，会忽略树生长指令。
- 无论使用交互模式还是直接模式，这两种创建决策树的方法最终都会生成相似的模型。只需考虑希望在此过程中执行多大程度的控制。

决策树节点

IBM® SPSS® Modeler 中的决策树节点提供对前面介绍的树构建算法的访问：

- C&R 树
- QUEST
- CHAID
- C5.0

有关详细信息，请参阅第 105 页码决策树模型。

这些算法的共同点是它们都通过将数据递归分割成越来越小的子组，从而构建决策树。不过，有一些重要的不同之处：

输入字段。输入字段（预测变量）可以是任何以下类型（测量级别）：连续、分类、标志、名义或有序。 有关详细信息，请参阅第 4 章中的测量级别中的IBM SPSS Modeler 14.2 源、过程和输出节点。

目标字段。仅可指定一个目标字段。对于 C&R 树和 CHAID，目标可以是连续、分类、标志、名义或有序。对于 QUEST，它可以是分类、标志或名义。对于 C5.0，目标可以是标志、名义或有序。

分割类型。 C&R 树和 QUEST 仅支持二元分割（即，每个树节点不能分割成两个以上的分支）。相反，CHAID 和 C5.0 支持一次分割成两个以上的分支。

用于分割的方法。不同算法在用于确定分割的标准上有所不同。C&R 树在预测分类输出时使用离差测量（默认为 Gini 系数，不过您可以进行更改）。对于分类目标，使用最小平方差方法。CHAID 使用卡方检验；QUEST 对分类预测变量使用卡方检验，对连续输入使用方差分析。对于 C5.0，使用一个信息理论度量，信息收益比率。

缺失值处理。所有算法均允许预测变量字段缺失值，但它们使用不同的缺失值处理方法。C&R 树和 QUEST 根据需要使用替代预测字段，以确保具有缺失值的记录在训练期间通过树。CHAID 将缺失值作为单独的类别，并允许在树构建过程中使用它们。C5.0 使用切分方法，将记录的某个切分部分从节点（基于具有缺失值的字段进行分割）向下传递到每个树分支。

修剪。 C&R 树、QUEST 和 C5.0 提供的选项允许完全生成树，然后删除对于树的精确性没有显著贡献的底层分割以进行修剪。不过，所有决策树算法均允许您控制最小子组大小，这有助于避免数据记录过少的分支。

交互树构建。 C&R 树、QUEST 和 CHAID 提供了启动交互式会话的选项。这允许您在创建模型之前构建树（一次一级）、编辑分割并修剪树。C5.0 未提供交互式选项。

先验概率。 C&R 树和 QUEST 支持在预测分类目标字段时为类别指定先验概率。先验概率是对总体（从中可提取训练数据）中的每个目标分类的总相对频率的估计。换句话说，先验概率是对预测值有任何了解之前对每个可能的目标值的概率估计。CHAID 和 C5.0 不支持指定先验概率。

规则集。对于具有分类目标字段的模型，决策树节点提供了以规则集形式创建模型的选项，这有时比复杂决策树更容易解释。对 C&R 树、QUEST 和 CHAID，您可以从交互式会话中生成规则集；对于 C5.0，可以在建模节点上指定此选项。此外，所有决策树模型均允许您从模型块生成规则集。 有关详细信息，请参阅第 126 页码从决策树中生成规则集。

C&R 树节点

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。可通过检查输入字段找到最佳分割来启动 C&R 树，并根据由分割导致的纯度指数降低情况进行测量。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）。

修剪

通过 C&R 树的选项可以首先生成树，然后根据成本复杂性算法（该算法可根据终端节点数调整风险评估）修剪树。通过此方法（此方法可以使树在长大后再根据更复杂的标准进行修剪）可生成交叉验证属性更佳的小型树。增加终端节点数通常会降低当前（训练）数据的风险，但当模型扩展为适用不可见数据时，实际的风险可能会更大。假设在一种极端的情况下，训练集合中的每个记录都有一个单独的终端节点。此时的风险评估可能是 0%，因为每个记录都落在了它自己的节点内，但对于不可见的（测试）数据，误分类的风险几乎肯定大于 0。成本复杂性测量将试图弥补这种风险。

示例。某有线电视公司委托进行市场研究，来确定有意预订有线电视交互服务的用户。使用研究中得来的数据可创建流，其中的目标字段为有意预订有线电视服务，预测变量字段则包括年龄、性别、教育、收入类别、每天看电视的时间和子女数。通过将 C&R 树节点应用到流，可以预测并对响应分类以获得活动的最高响应率。

要求。要训练 C&R 树模型，需要一个或多个输入字段和唯一一个目标字段。目标和输入字段可以是连续（数值范围）或分类。设置为双向或无的字段将忽略。必须将模型中使用的字段的类型完全实例化，且模型中使用的所有有序（有序集合）字段的存储类型必须是数值（而不是字符串）。必要的话，可以使用重新分类节点对存储类型进行转换。 有关详细信息，请参阅第 4 章中的重新对节点分类中的IBM SPSS Modeler 14.2源、过程和输出节点。

强度。对于所存在的问题，例如缺失数据和大量字段，C&R 树模型十分稳健。这些模型通常不需要花费很长的训练时间用于估计。此外，C&R 树模型与某些其他模型类型相比似乎更容易理解 – 源自模型的规则解释起来更简明易懂。与 C5.0 不同的是，C&R 树可同时兼容连续字段和分类输出字段。

CHAID 节点

CHAID 或卡方自动交互效应检测是一种通过使用卡方统计量识别最优分割来构建决策树的分类方法。

CHAID 首先检查每个输入字段和结果之间的交叉列表，然后使用卡方独立性测试来检验显著性。如果以上多个关系具有显著的统计意义，则 CHAID 将选择最重要（p 值最小）的输入字段。如果输入具有两个以上的类别，将会对这些类别进行比较，然后将结果中未显示出差异的类别合并在一起。此操作通过将显示的显著性差异最低的类别对相继合并在一起来实现。当所有剩余类别在指定的检验级别上存在差异时，此类别合并过程将终止。对于名义输入字段，可以合并任何类别；对于有序集合，只能合并连续的类别。

Exhaustive CHAID 是 CHAID 的修正版，它可对每个预测变量的所有可能分割进行更彻底的检查，但计算时间比较长。

要求。目标和输入字段可以是连续字段，也可以是分类字段；节点在每一层上都可以分割为两个或多个子组。模型中使用的所有顺序字段的存储类型都必须是数字类型（不是字符串）。必要的话，可以使用重新分类节点对存储类型进行转换。 有关详细信息，请参阅第 4 章中的重新对节点分类中的IBM SPSS Modeler 14.2 源、过程和输出节点。

强度。CHAID 与 C&R 树和 QUEST 节点不一样，它可以生成非二元树，这意味着有些分割将有多于两个的分支。因此，与二元生成方法相比，CHAID 倾向于创建范围更广的树。CHAID 可使用各种类型的输入，并且可接受观测值加权和频率变量。

QUEST 节点

QUEST，或称快速、无偏倚、高效率统计树，是一种用于构建决策树的二元分类法。开发此方法的一个主要目的是减少包含很多变量或观测值的大型 C&R 树分析所需的处理时间。QUEST 的第二个目的是减少在分类树法中发现的趋势以便支持允许有多个分割的输入，即连续（数值范围）输入或具有多个类别的输入。

- QUEST 可基于显著性检验使用序列规则来评估节点上的输入字段。为了进行选择，可能需要对节点的每个输入执行一次尽可能简单的检验与 C&R 树不同，所有的分割都不用检查，而与 C&R 树和 CHAID 都不同的是，在评估输入字段以供选择时不会检验类别组合。因此可加快分析的速度。
- 通过使用由目标类别形成的组中选定的输入来运行二次判别分析可以确定分割。使用此方法可再次使速度较穷举搜索（C&R 树）得到提高以便确定最优分割。

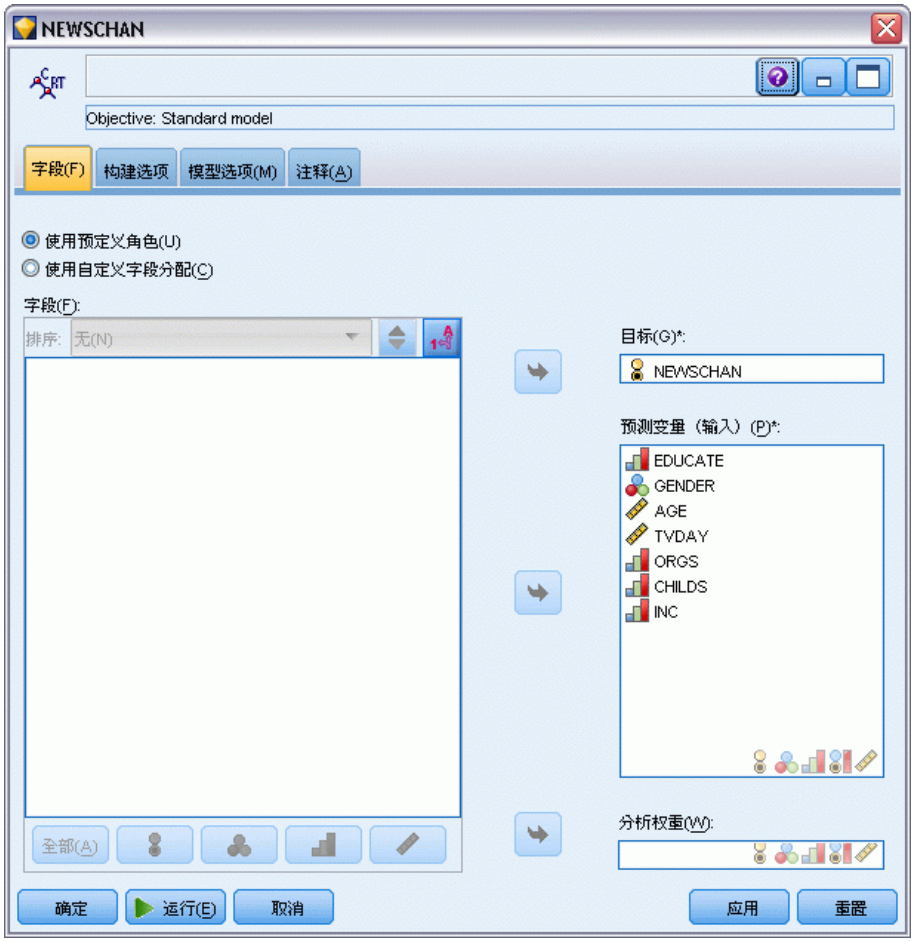
要求。输入字段可以是连续（数值范围）的，但目标字段必须是分类的。所有分割都是二元的。不能使用加权字段。模型中使用的所有有序（有序集合）字段的存储类型都必须是数字类型（不是字符串）。必要的话，可以使用重新分类节点对存储类型进行转换。 有关详细信息，请参阅第 4 章中的重新对节点分类中的IBM SPSS Modeler 14.2 源、过程和输出节点。

强度。与 CHAID 相似但与 C&R 树不同的是，QUEST 可使用统计检验确定是否使用输入字段。QUEST 还可将输入的选择与分割问题分开，分别为其应用不同的标准。不过在 CHAID 中，确定变量选择的统计检验结果还可生成分割。同样，C&R 树也可采用杂质更改测量在选择输入字段的同时确定分割。

决策树节点字段选项

在“字段”选项卡上，可以选择是要使用在上游节点中定义的字段角色设置，还是手动进行字段分配。

图片 6-24
C&R 树节点，“字段”选项卡



使用预定义角色。此选项使用上游类型节点（或上游源节点的“类型”选项卡）的角色设置（目标、预测变量等等）。有关详细信息，请参阅第 4 章中的设置字段角色中的 IBM SPSS Modeler 14.2 源、过程和输出节点。

使用自定义字段分配。如果您要在此屏幕中手动分配目标、预测变量和其他角色，请选择此项。

字段。使用箭头按钮可以从列表中将项目手动分配到屏幕右侧的各类角色字段。图标表示每个角色字段的有效测量级别。

单击**全部**按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

目标。选择单个字段作为预测目标。

预测变量（输入）。选择一个或多个字段作为预测输入。

分析权重。（仅 CHAID 和 C&RT）要使用字段作为个案权重，在此处指定。个案权重将作为对输出字段各个水平上方差的差异的一种考量。 有关详细信息，请参阅第 33 页码第 3 章中的使用频率和权重字段。

决策树节点构建选项

通过“构建选项”选项卡，您可以设置构建模型的所有选项。当然，您只需单击运行按钮，即可采用所有默认选项来构建模型；不过，通常您需要根据具体用途自定义构建选项。

您可以在此选择是构建新模型还是更新现有模型。您还可以设置节点的主目标：构建标准模型、构建具有增强精确性或稳定性的模型，还是构建用于大型数据集的模型。

图片 6-25
C&R 树节点，“构建选项”选项卡

您希望做什么？

新建模型。（默认）每次运行包含此建模结点的流时，就会创建一个全新模型。

继续训练现有模型。默认情况下，每当执行一个建模节点时，就会创建一个完整的新模型。如果选中该选项，则会继续训练该节点成功生成的最后一个模型。这样就可以在无需访问原始数据的情况下更新或刷新现有的模型，并可能会显著提升性能，这是因为只有 新的或更新后的记录被反馈到流中。上一个模型的详细信息与建模节点存储在一起，这样即使先前的模型块在流或模型选项板中不再可用的情况下，也可以使用该选项。

注意：此选项仅当您选择**为超大型数据集创建模型**作为目标时才会激活。

您的主要目标是什么？

- **构建单个树。**创建单个标准决策树模型。一般来说，与使用其他目标选项构建的模型相比，标准模型更易于解释，并能更快地进行评分。

  众数。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。**启动交互会话**可打开树构建器，通过该构建器可在创建模型块之前构建树（一次一级）、编辑分割并按照需要进行修剪。

  使用树指令。选中此选项可指定从节点中生成交互树时所应用的指令。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。 有关详细信息，请参阅第 125 页码更新树指令。

- **增强模型准确性（推进）。**如果您要使用一种名为**推进**的特殊方法来提高模型准确率，请选择此项。推进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。推进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。

- **增强模型稳定性（bagging）。**如果您要使用一种名为 **bagging**（Bootstrap 汇总）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。此选项创建多个模型并加以合并，从而获得更可靠的预测结果。与标准模型相比，通过此选项获得的模型需要更长的构建与评分时间。

- **为超大型数据集创建模型。**如果您的数据集过大，而无法使用任何上述目标选项构建模型，请选择此项。此选项将数据划分为较小的数据块，并在每个块上构建模型。这将自动选择最准确的模型，并合并成单个模型块。如果您在此屏幕上选择**继续训练现有模型**选项，可以执行增量式模型更新。注意：此选项适合大型数据集，需要到 IBM® SPSS® Modeler Server 的连接。 有关详细信息，请参阅第 3 章中的连接到 IBM SPSS Modeler Server中的IBM SPSS Modeler 14.2 用户 指南。

决策树节点 - 基本

您可在此指定有关如何构建决策树的基本选项。

图片 6-26
决策树基本选项

树生长算法。（仅 CHAID）选择您要使用的 CHAID 算法类型。Exhaustive CHAID 是 CHAID 的修正版，它可对每个预测变量的所有可能分割进行更彻底的检查，但计算时间比较长。

最大树深度。指定根节点以下的最大级数（递归分割样本的次数）。默认值为 5；选择自定义，输入值以指定其他级数。

修剪（仅 C&RT 和 QUEST）

修剪树以防止过拟合。修剪包括删除对于树的精确性没有显著贡献的底层分割。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。如果需要完整的未经修剪的树，请保持此项处于未选中状态。

- **最大风险差值（标准误）：**通过此选项可指定更自由的修剪规则。标准误规则使算法能够选择最简单的树，该树的风险评估接近于（但也可能大于）风险最小的子树的风险评估。此值表示在风险评估中已修剪树和风险最小的树之间所允许的风险评估差异大小。例如，如果指定 2，则将选择其风险评估（2 × 标准误）大于完整树的风险评估的树。

最大代用项。 代用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。

决策树节点 - 停止规则

图片 6-27
停止规则的选项

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**）中要分割的记录数小于指定值，则**父分支中的最小记录数**将阻止进行分割。如果由分割创建的任意分支（**子**）中的记录数小于指定值，则**子分支中的最小记录数**将阻止进行分割。

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

决策树节点 - 整体

图片 6-28
整体的选项

这些设置决定了在“目标”中请求 boosting、bagging 或超大型数据集时发生的整体行为。对选定目标不适用的选项将被忽略。

Bagging 和大型数据集。在对整体评分时，此规则用于组合来自基本模型的预测值，以计算整体得分值。

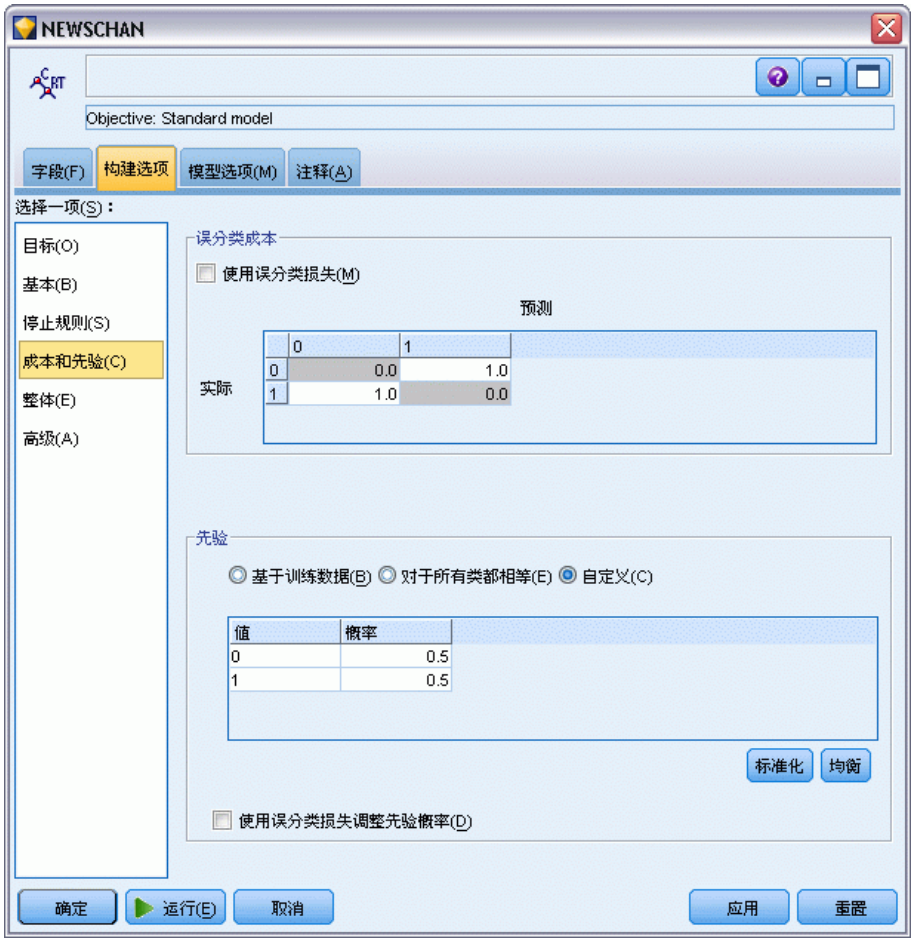
- **分类目标的默认组合规则。**可以通过投票、最高概率或最高平均值概率来对分类目标的整体预测值进行组合。**投票**选择在基本模型中最常具有最高概率的类别。**最高概率**选择在所有基本模型中取得单个最高概率的类别。**最高平均值概率**选择在基本模型中对类别概率取平均值时具有最高值的类别。
- **连续目标的默认组合规则。**可以通过对来自基本模型的预测值取平均值或中位数，对连续目标的整体预测值进行组合。

注意，如果以增强模型精确性为目标，则组合规则选择将被忽略。Boosting 方法始终使用加权大多数投票来对分类目标进行评分，而使用加权中位数对连续目标进行评分。

Boosting 和 Bagging。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 bagging 方法，此为 bootstrap 样本数。它应为正整数。

C&R 树和 QUEST 节点 - 成本和先验

图片 6-29
设置误分类成本和先验概率



误分类损失

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错误的相对重要性。

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序或比较时，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。默认情况下，所有误分类成本都设置为 1.0。要输入自定义成本值，可选择**使用误分类成本**并将自定义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，则将 B 误分类为 A 的成本将仍是默认值 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（从中可提取训练数据）中的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：

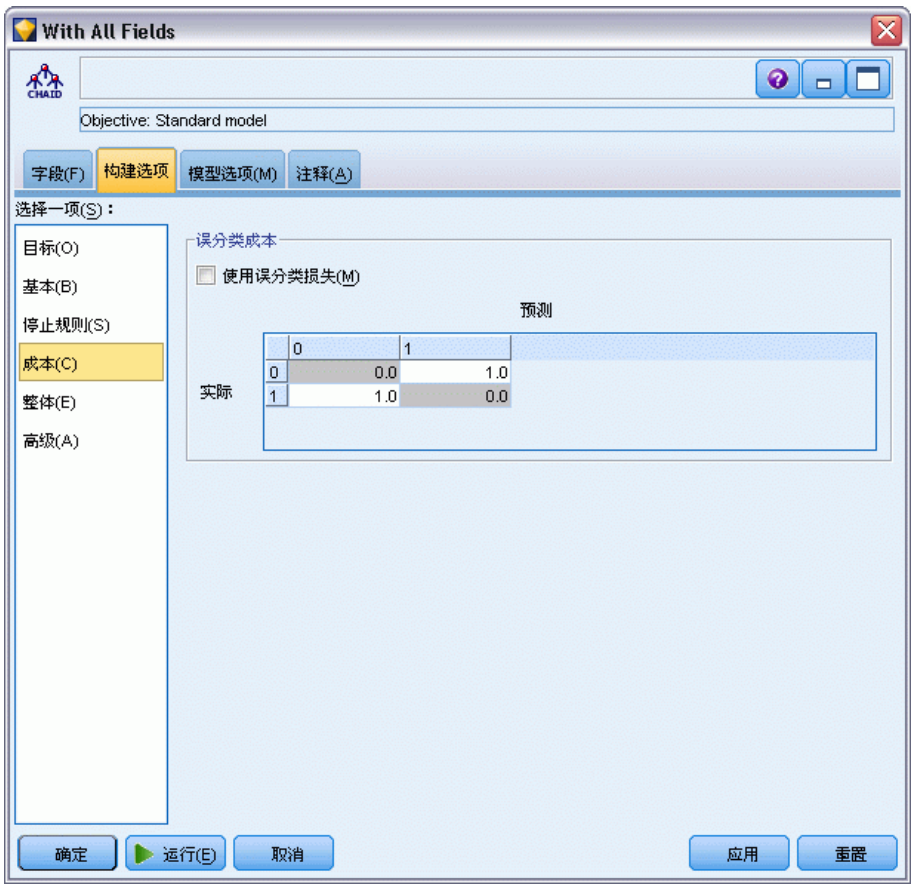
- **基于训练数据。** 这是默认选项。先验概率基于训练数据中分类的相对频率。
- **对于所有类都相等。** 所有分类的先验概率都定义为 1/k，其中 k 是目标分类数。
- **自定义。** 可以自己指定先验概率。对于所有类，都将先验概率的初值设置为相等。可以将单个分类的概率调整为用户定义的值。要调整特定分类的概率，可在表中对应于所需分类的概率单元格中，先清除其内容，然后输入所需的值。

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。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评估。）

CHAID 节点 - 成本

图片 6-30
在 CHAID 节点中的误分类成本



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错误的相对重要性。

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序或比较时，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。默认情况下，所有误分类成本都设置为 1.0。要输入自定义成本值，可选择**使用误分类成本**并将自定义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，则将 B 误分类为 A 的成本将仍是默认值 1.0，除非也明确地对它进行更改。

C&R 树节点 - 高级

使用“高级”选项可对树构建过程进行微调。

图片 6-31
设置 C&R 树节点的高级选项

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**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，则该节点被认为是“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，则不会进行此分割。

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**是基于分支的类别归属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，即只有连续的目标类才可以组成一组，此选项仅适用于顺序目标。如果对于名义目标选中此选项，将默认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。默认值为 30。

重复结果。设置随机种子允许您复制分析。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。

QUEST 节点 - 高级

使用“高级”选项可对树构建过程进行微调。

图片 6-32
设置 QUEST 节点的高级选项

用于分割的显著性水平。指定用于分割节点的显著性水平（alpha）。该值必须位于 0 和 1 之间。值越小，生成的树的节点也会越少。

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。默认值为 30。

重复结果。设置随机种子允许您复制分析。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。

CHAID 节点 - 高级

使用“高级”选项可对树构建过程进行微调。

图片 6-33
设置 CHAID 节点的高级选项

用于分割的显著性水平。指定用于分割节点的显著性水平（alpha）。该值必须位于 0 和 1 之间。值越小，生成的树的节点也会越少。

用于合并的显著性水平。指定用于合并类别的显著性水平（alpha）。该值必须大于 0 并小于或等于 1。为阻止对类别进行任何合并，可将值指定为 1。对于连续目标，这意味着最终树中变量的类别数与指定的时间间隔数相匹配。此选项对于 Exhaustive CHAID 不适用。

使用 Bonferroni 方法调整显著性值。检验预测变量的各种类别组合时调整显著相关值。显著相关值可基于检验次数进行调整，而检验次数直接与预测变量的类别数及测量等级相关。通常需要选中此选项，因为它可以更好地控制假阳性错误率。禁用此选项将提高您的分析能力以找到真差分，但以增加假阳性率为代价。建议您禁用此选项，尤其对于较小的样本。

在节点内允许重新分割合并类别。CHAID 算法试图合并类别以生成最简单的树来描述模型。如果选中此选项，且合并后的结果能够比较好地描述模型，则可重新分割已合并的类别。

类别目标的卡方值。对于类别目标，可指定用于计算卡方统计量的方法。

- **Pearson。**此方法提供更快的计算，但是对于小样本应该谨慎使用它。
- **似然比。**此方法与 Pearson 方法相比更加稳健，但计算时间比较长。对于小样本，这是首选的方法。对于连续目标，通常使用此方法。

期望单元格频率的最小更改。（为名义模型和行效应顺序模型）估计单元格频率时，迭代过程 (epsilon) 用于对最优估计（在特定分割的卡方检验中使用）进行收敛。Epsilon 可确定必须对迭代进行多大的更改才可使其继续；如果对最后一个迭代的更改小于指定的值，则迭代将停止。如果因算法中存在问题而无法收敛，则可以增加该值或增加最大迭代次数，直到发生收敛为止。

收敛的最大迭代次数。无论是否已进行收敛都指定停止前的最大迭代次数。

重复结果。设置随机种子允许您复制分析。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。

决策树节点模型选项

在“模型选项”选项卡上，您可以选择是指定模型名称，还是自动生成名称。您还可以选择获得预测变量重要性信息，以及标志目标的原始和调整倾向得分。

图片 6-34
设置决策树节点的模型选项

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

模型评估

计算预测变量重要性。对于生成相应重要性测量的模型，可以显示一个图表来说明评估模型中每个预测变量的相对重要性。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。请注意，对于某些模型，计算预测变量重要性（特别对较大数据集进行操作时）可能需要花较长时间，因此默认情况下，对于某些模型，预测变量重要性均处于关闭状态。预测变量重要性对于决策列表模型不可用。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

倾向得分

可以在建模节点中和模型块的“设置”选项卡上启用倾向得分。该功能仅在所选目标为标志字段时才可用。有关详细信息，请参阅第 36 页码第 3 章中的倾向得分。

计算原始的倾向得分。原始的倾向得分仅从基于训练数据的模型中导出。如果模型预测值为真（将响应），则倾向与 P 相同，其中 P 为预测的可能性。如果模型预测的值为假，则计算出的倾向为 (1 - P)。

- 如果构建模型时选择了此选项，则默认情况下将在模型块中启用倾向得分。不过，无论是否在建模节点中选择了原始倾向得分，都可以始终在模型块中选择启用原始倾向得分。
- 对模型进行评分时，原始倾向得分将被添加到将 RP 字母附加到标准前缀的字段中。例如，如果预测位于名为 $R-churn 的字段中，则倾向得分字段的名称将是 $RRP-churn。

计算调整后的倾向得分。原始倾向仅仅基于由可能过度拟合的模型给定的估计上，这样会导致过于乐观地评估倾向。调整后的倾向尝试通过查看模型在检验或验证分区的性能或通过调整倾向来弥补，以相应地给作出更好的估计。

- 此设置要求流中存在有效的分区字段。有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。
- 与原始置信度分数不同，调整后的倾向得分必须在构建模型时计算；否则，对模型块进行评分时该分数将不存在。
- 对模型进行评分时，在将 AP 字母附加到标准前缀的字段中添加调整后的倾向得分。例如，如果预测位于名为 $R-churn 的字段中，则倾向得分字段的名称将是 $RAP-churn。调整后的倾向得分不适用于 logistic 回归模型。
- 在计算调整后的倾向得分时，必须尚未平衡用于计算的检验或验证分区。为避免这一点，请确保在任何上游平衡节点中选中**仅平衡训练数据**选项。有关详细信息，请参阅第 3 章中的为平衡节点设置选项中的IBM SPSS Modeler 14.2 源、过程和输出节点。此外，如果已在上游获取了复杂样本，则会导致调整后的倾向得分无效。
- 调整后的倾向得分不适用于“增强型”树和规则集模型。有关详细信息，请参阅第 158 页码增强型 C5.0 模型。

  基于。对于有待计算的调整后的倾向得分，流中必须存在一个分区字段。可以指定是使用检验分区还是验证分区进行此计算。为获取最佳结果，检验或验证分区包含的记录数量应至少与用于训练原始模型的分区所包含的记录数相同。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

C5.0 节点

该节点使用 C5.0 算法构建**决策树**或**规则集**。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**对模型值没有显著贡献的分割。

注意：C5.0 节点只能预测分类目标。分析包含分类（名义或有序）字段的数据时，与 11.0 版以前的 C5.0 版本相比将类别组合在一起的可能性**□□**大。

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，则会将默认预测分配到该记录。

示例。医学研究员已收集了一组患有同一疾病的患者的数据。在治疗过程中，每位患者均对五种药物中的一种有明显反应。可以使用 C5.0 模型连同其他节点来为以后患有同一疾病的患者寻找最适合其的药物。有关详细信息，请参阅第 9 章中的药物治疗（勘察表/C5.0）中的IBM SPSS Modeler 14.2 应用程序 指南。

要求。要训练 C5.0 模型，必须有一个分类（即名义或有序）目标字段和一个或多个任意类型的输入字段。设置为双向或无的字段将忽略。必须对模型中使用的字段的类型完全实例化。还可以指定加权字段。

强度。对于所存在的问题，例如缺失数据和大量输入字段，C5.0 模型十分稳健。这些模型通常不需要花费很长的训练时间用于估计。此外，C5.0 模型与某些其他模型类型相比似乎更容易理解，因为源自模型的规则解释起来更简明易懂。C5.0 还提供功能强大的**推进**方法来提高分类的准确性。

注：启用并行处理有益于提高 C5.0 模型构建速度。 有关详细信息，请参阅第 12 章中的设置优化选项中的IBM SPSS Modeler 14.2 用户 指南。

C5.0 节点模型选项

图片 6-35
C5.0 节点模型选项

模型名称。指定要生成的模型的名称。

- **自动。**选中此选项将根据目标字段名称自动生成模型名称。这是默认值。
- **自定义。**选中此选项可以为由此节点创建的模型块指定自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

输出类型。在此指定希望结果模型块是**决策树**还是**规则集**。

群体字符。如果选中此选项，C5.0 将试图组合输出字段中具有相似样式的符号值。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），则它将默认创建一个三向分割。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。

使用推进。C5.0 算法有一个特殊的方法用于提高其准确率，称为**推进**。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。推进可以显著提高 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**选项可以控制在推进模型过程中使用的模型数目。此功能基于对 Freund & Schapire 的研究，使用其中的专有改进功能可以更好地处理噪声数据。

交互验证。如果选中此选项，C5.0 将使用一组模型（根据训练数据的子集构建）来估计某个模型（根据全部数据集构建）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，交互验证模型将被丢弃。可以指定用于交互验证的**折叠次数**或模型数。注意，在 IBM® SPSS® Modeler 以前的版本中，构建模型和交互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。模型构建和交互验证将同时执行。

众数。对于**简单**训练，大部分 C5.0 参数都自动设置。**专家**训练允许您更直接地控制训练参数。

简单模式选项

支持。默认情况下，C5.0 将试图生成尽可能精确的树。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。

注意：选中**普遍性**选项后，不能保证所构建模型的通用性一定会比其他模型好。当普遍性问题比较严重时，通常可使用保留检验样本验证模型。

预期的噪声（%）。指定训练集合中噪声数据或错误数据的预期比例。

专家模式选项

修剪严重性。确定对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（请参阅下面的“使用全局修剪”）。

每个子分支的最小记录数。可使用子组的大小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。默认值为 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。分两个阶段修剪树：第一个阶段是本地修剪，此时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。默认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，C5.0 将在开始构建模型之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。

注：启用并行处理有益于提高 C5.0 模型构建速度。 有关详细信息，请参阅第 12 章中的设置优化选项中的IBM SPSS Modeler 14.2 用户 指南。

决策树模型块

决策树模型块表示用于预测特定输出字段的树结构，该树结构可由以下决策树建模节点之一发现：IBM® SPSS® Modeler 以前版本中的 C&R 树、CHAID、QUEST、C5.0或构建规则）。树模型可以直接从树构建节点中生成，也可以从交互树构建器中间接生成。 有关详细信息，请参阅第 107 页码交互树构建器。

评分树模型

运行包含树模型块的流时，特定的结果取决于树的类型。

- 对于分类树（分类目标），会将两个新字段（其中分别包含每个记录的预测值和置信度）添加到数据中。预测取决于为其分配记录的终端节点的使用最频繁的类别；如果在给定节点中大多数响应为是，则对分配到该节点的所有记录的预测也为是。
- 对于回归树，仅生成预测值；而不会分配置信度。
- 另外，对于 CHAID、QUEST 和 C&R 树模型，也可以添加表示节点 ID 的附加字段，每个记录都将分配到此节点中。

新的字段名称将通过为模型名称添加前缀生成。对于 C&R 树、CHAID 和 QUEST，预测字段的前缀是 $R-，置信度字段的前缀是 $RC-，节点标识符字段的前缀是 $RI-。对于 C5.0 树，预测字段的前缀是 $C-，置信度字段的前缀是 $CC-。如果存在多个树模型节点，则必要时可在新字段名称的前缀中添加数字以进行区分 - 例如，$R1-、$RC1- 和 $R2- 等等。

使用树模型块

可以多种方式保存或导出与模型相关的信息。

注意：其中的许多选项也适用于树构建器窗口。

通过树构建器或树模型块，可以执行下列操作：

- 根据当前的树生成过滤节点或选择节点。有关详细信息，请参阅第 126 页码生成过滤节点和选择节点。
- 生成一个规则集块，该节点将树结构表示成一组定义了树的终端分支的规则。有关详细信息，请参阅第 126 页码从决策树中生成规则集。
- 此外，还可以按 PMML 格式导出模型（仅限于树模型块）。有关详细信息，请参阅第 41 页码第 3 章中的模型选项板。如果模型包含定制分割，则不会在导出的 PMML 中保留此信息。（保留分割，但不保留它是定制分割而不是通过算法选择的分割这一事实。）
- 基于当前树的所选部分生成图形。注：这仅在块附加到流中的其他节点时才有效。有关详细信息，请参阅第 158 页码生成图形。
- 仅在增强型 C5.0 模型中，可以选择**单一决策树（工作区）**或**单一决策树（GM 选项板）**以根据当前选定的规则创建一个新的规则集。 有关详细信息，请参阅第 158 页码增强型 C5.0 模型。

注意：虽然构建规则节点已由 C&R 树节点所替代，但现有流中最初使用构建规则节点创建的决策树节点仍可正常工作。

单个树模型块

如果在建模节点上选择**构建单个树**作为主目标，则结果模型块包含下列选项卡。

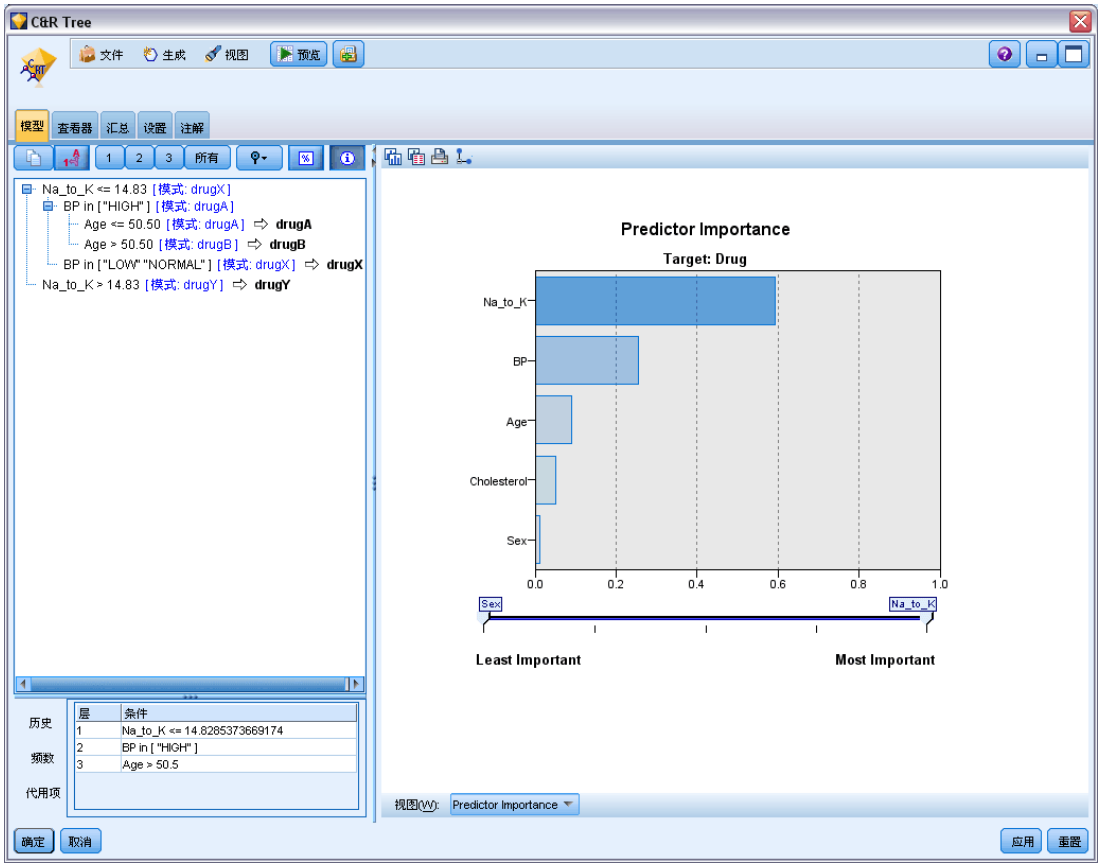
Tab	描述	其他信息
模型	显示定义模型的规则。	有关详细信息，请参阅第 151 页码 决策树模型规则。
查看器	显示模型的树视图。	有关详细信息，请参阅第 154 页码 决策树模型查看器。
摘要	显示有关字段、构建设置和模型评估过程的信息。	有关详细信息，请参阅第 44 页码 第 3 章中的模型块概要/信息。
设置	允许您指定模型评分期间的置信度与 SQL 生成相关选项。	有关详细信息，请参阅第 155 页码 决策树/规则集模型块设置。
注解	允许您添加描述性注解，指定自定义名称，添加工具提示文本，以及指定模型的搜索关键字。	有关详细信息，请参阅第 5 章中的注解中的IBM SPSS Modeler 14.2 用户 指南。

决策树模型规则

决策树模型块的“模型”选项卡显示定义该模型的规则。此外，还可以显示预测变量重要性的图形和包含有关历史、频率和代用项信息的第三个面板。

注意：如果您在 CHAID 节点“构建选项”选项卡（“目标”面板）上选中**为超大型数据集创建模型**选项，则“模型”选项卡只显示树规则详细信息。

图片 6-36
决策树模型块



树规则

左侧面板显示了条件列表，这些条件定义算法发现的数据的分区 - 本质上是一系列规则，可基于不同预测变量的值将单个记录分配给子节点。

决策树基于输入字段值的对数据进行递归分区。数据分区称为**分支**。初始分支（有时称为**根**）包含所有数据记录。根将根据特定输入字段的值被分成若干子集或**子分支**。每个子分支可以进一步分割成次级子分支，次级子分支还可进一步分割，如此类推。不再分割的分支是树的最底层分支。这样的分支称为**终端分支**（或**叶片**）。

规则浏览器显示了输入值，输入值定义了每个分区或分支以及这些分割中记录的输出字段值概要。有关使用模型浏览器的一般信息，请参阅浏览模型块。

对于基于数值型字段的分割，分支将以下行所示的形式显示：

```
fieldname relation value [summary]
```

这里的 relation 是数值型关系。例如，由 revenue 字段大于 100 的值所定义的分支将显示为如下形式：

```
revenue > 100 [summary]
```

对于基于符号型字段的分割，分支将以下行所示的形式显示：

```
fieldname = value [summary] or fieldname in [values] [summary]
```

这里的 values 表示定义分支的字段值。例如，包含 region 字段值为 North、West 或 South 的记录的分支将以如下形式表示：

```
region in ["North" "West" "South"] [summary]
```

终端分支也将进行预测，同时会在规则条件的尾部添加箭头和预测值。例如，定义时 revenue > 100 且预测输出字段值为 high 的叶片将显示如下：

```
revenue > 100 [Mode:high] →     high
```

数值型和符号型输出字段的分支**概要**定义有所不同。对于含有数值型输出字段的树，分支的**平均**值便是概要，分支的**效应**便是分支平均值与其父分支平均值的差。对于含有符号型输出字段的树，分支中记录的**中位数**（或出现频率最高的值）便是概要。

要完全描述分支，需要包含定义分支的条件以及定义树中更深层分割的条件。例如，在树中：

```
revenue > 100
  region = "North"
  region in ["South" "East" "West"]
    revenue <= 200
```

第二行所表示的分支由条件 revenue > 100 和 region = “North” 进行定义。

如果单击工具栏上的 **显示实例/置信度**，则每条规则还将显示其所适用的记录数（**实例数**）和规则为真的**记录所占的比例（**置信度**）。

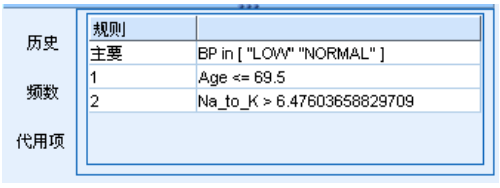
预测变量重要性

另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。注意，只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

其他模型信息

如果单击工具栏中的**显示其他信息面板**，您将在窗口底部看到显示选定规则详细信息的面板。信息面板包含三个选项卡。

图片 6-37
信息面板上显示的“代用项”



历史。此选项卡追踪从根节点至选定节点的分割条件。从而给出了一个条件列表，据此可以判断出记录何时分配给了选定节点。所有条件均为真的记录将分配给此节点。

频率。对于含符号目标字段的模型而言，此选项卡（为每个可能的目标值）显示了分配给包含目标值（训练数据中）节点的记录数。还将显示频率图（显示为最多三位小数的百分比）。对于含数值型目标的模型，此选项卡为空。

代用项。如果适用，则会针对所选节点显示主要分割字段的所有代用项。代用项是在给定记录的主要预测变量值缺失时使用的替代字段。给定分割允许的最大代用项数在树构建节点中指定，但实际数量取决于训练数据。一般来讲，缺失数据越多，可能使用的代用项越多。对于其他决策树模型，此选项卡为空。

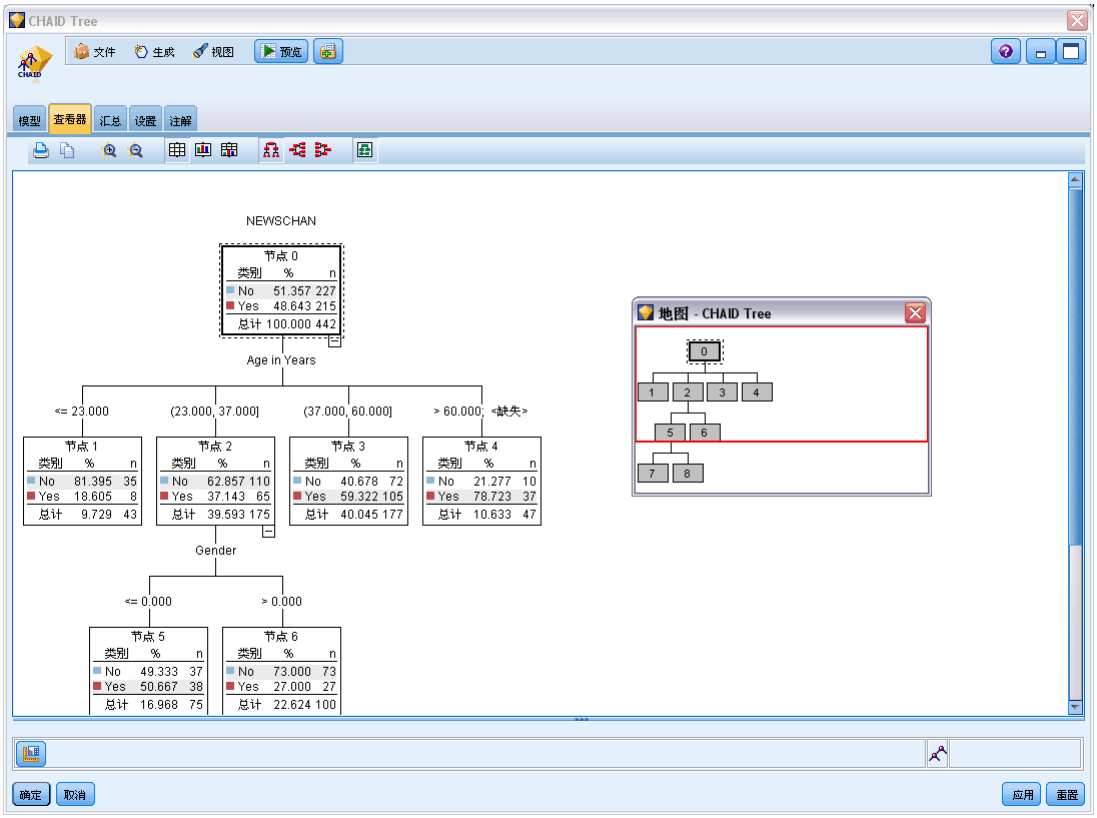
注：要在模型中包含代用项，必须在训练阶段对其进行标识。如果训练样本没有缺失值，则不会标识任何代用项；在测试或评分过程中遇到的具有缺失值的所有记录将自动落入记录数最大的子节点。如果在测试或评分过程中预期出现缺失值，请确保值在训练样本中也处于缺失状态。代用项对于 CHAID 树不可用。

决策树模型查看器

决策树模型块的“查看器”选项卡类似于树构建器中的显示。主要的区别是当浏览模型块时，无法生成或修改树。两个组件中用于查看和自定义显示的其他选项都类似。有关详细信息，请参阅第 113 页码自定义树状视图。

注意：对于您在“构建选项”选项卡的“目标”面板上选中**为超大型数据集创建模型**选项时构建的 CHAID 模型块，不显示“查看器”选项卡。

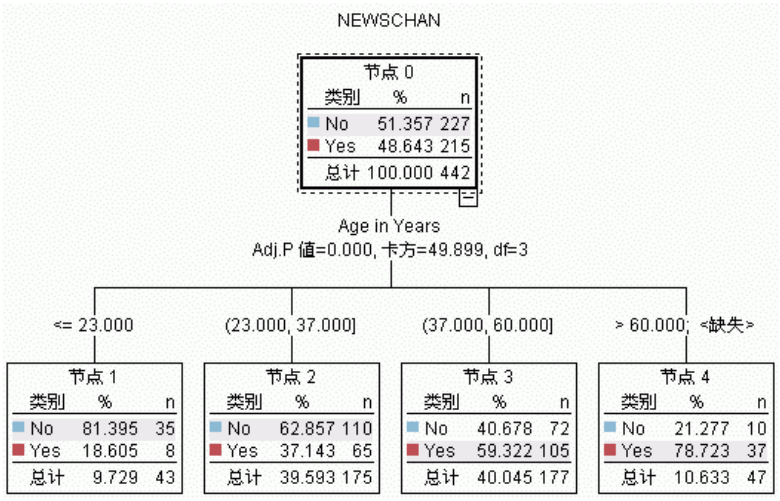
图片 6-38
带树映射窗口的决策树“查看器”选项卡



查看“查看器”选项卡上的分割规则时，方括号表示临界值包含在范围中，而圆括号表示临界值不包含在范围中。因此，表达式 (23,37] 表示从 23（不含）到 37（含）；即从 23 以上到 37。在“模型”选项卡上，相同的情况显示为：

```
Age > 23 and Age <= 37
```

图片 6-39
查看器选项卡上显示的分割规则



决策树/规则集模型块设置

使用决策树或“规则集”模型块的“设置”选项卡，可以在模型评分期间为置信度及 SQL 生成指定选项。只有将模型块添加到流之后，此选项卡才可用。

图片 6-40
决策树模型块设置

计算置信度。选中此选项以便在评分操作中包括置信度。在数据库中评分模型时，排除置信度有助于生成更有效的 SQL。不会为回归树分配置信度。

注意：如果您在 CHAID 模型“构建选项”选项卡的“方法”面板上选中**为超大型数据集创建模型**选项，此复选框仅在名义或标志分类目标的模型块中可用。

计算原始的倾向得分。对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向得分，这些得分指示为目标字段指定结果为真的可能性。除了这些得分，还有其他在评分过程中生成的预测值和置信度值。

注意：如果您在 CHAID 模型“构建选项”选项卡的“方法”窗格上选中**为超大型数据集创建模型**选项，此复选框仅在标志分类目标的模型块中可用。

计算调整后的倾向得分。原始倾向得分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该得分可能会过度优化。调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。此选项要求在流中定义分区字段并且在生成模型之前在建模节点中启用调整的倾向得分。

注意：调整后的倾向得分不适用于增强型树和规则集模型。 有关详细信息，请参阅第 158 页码增强型 C5.0 模型。

规则 ID。对于 CHAID、QUEST 和 C&R 树模型，此选项可在得分输出中添加表示终端节点 ID 的字段，每个记录都将分配到此终端节点中。

注意：选中此选项时，SQL 生成不可用。

生成此模型的 SQL。使用数据库中的数据时，SQL 代码可传回到数据库中执行，从而大大提高许多操作的处理速度。有关详细信息，请参阅第 6 章中的SQL 优化中的IBM SPSS Modeler Server 14.2 管理和性能指南。

选中下列其中一个选项可启用或禁用 SQL 生成。

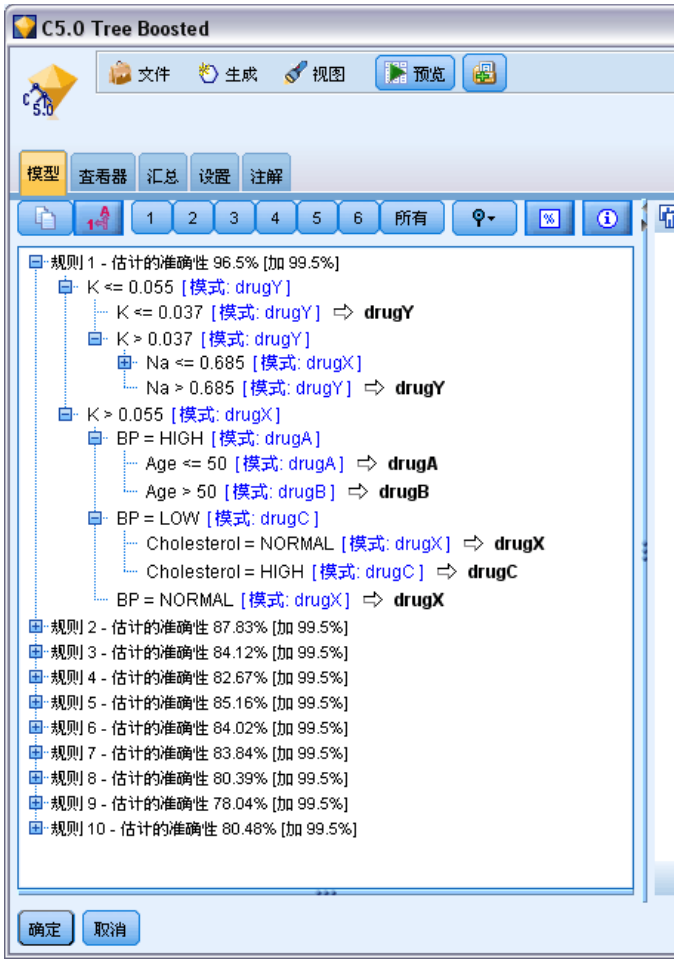
- **不生成。**选中此选项为模型禁用 SQL 生成。
- **在无缺失值支持的情况下生成。**选中此选项可以启用 SQL 生成，而不必经常处理缺失值。如果在评分观测值时遇到缺失值，则此选项会将预测设置为 Null 值（$null$）。

  注意：此选项对于 CHAID 模型不适用。对于其他模型类型，此选项仅适用于决策树（而非规则集）。
- **在缺失值支持的情况下生成。**对于 CHAID、QUEST 和 C&R 树模型，可在支持全部缺失值的情况下启用 SQL 生成。因此，生成 SQL 意味着已按模型中指定的方式处理缺失值。例如，C&R 树可使用代用项规则和最大子退回。

  注意：对于 C5.0 模型，此选项仅适用于规则集（而非决策树）。

增强型 C5.0 模型

图片 6-41
增强型 C5.0 模型块“模型”选项卡



创建增强型 C5.0 模型（规则集或决策树）时，实际上创建了一组相关模型。增强型 C5.0 模型的模型规则浏览器可显示位于层次顶层的模型的列表，以及每个模型的估计准确性和增强型模型整体的总体准确性。要检查特定模型的规则或分割，可选择并根据在单模型中扩展规则或分支的方式扩展该模型。

也可以从增强型模型集中提取特定的模型并创建恰好包含此模型的新规则集模型块。要从增强型 C5.0 模型中创建新的规则集，可选择所需规则集或树，并从“生成”菜单中选择**单一决策树（GM 选项板）**或**单一决策树（工作区）**。

生成图形

树节点提供了大量信息，但对商业用户来说，它可能并不始终是一种方便访问的格式。要使提供的数据便于纳入商业报表和演示文稿，您可生成所选数据的图形。例如，从模型块的“模型”或“查看器”选项卡，或从交互树的“查看器”选项卡，您可以生成树的选定部分的图形，因此只创建选定树或分支节点中个案的图形。

注意：当模型块附加到流中的其他节点时，您只能从模型块生成图形。

生成图形

第一步是选择要显示在图形上的信息：

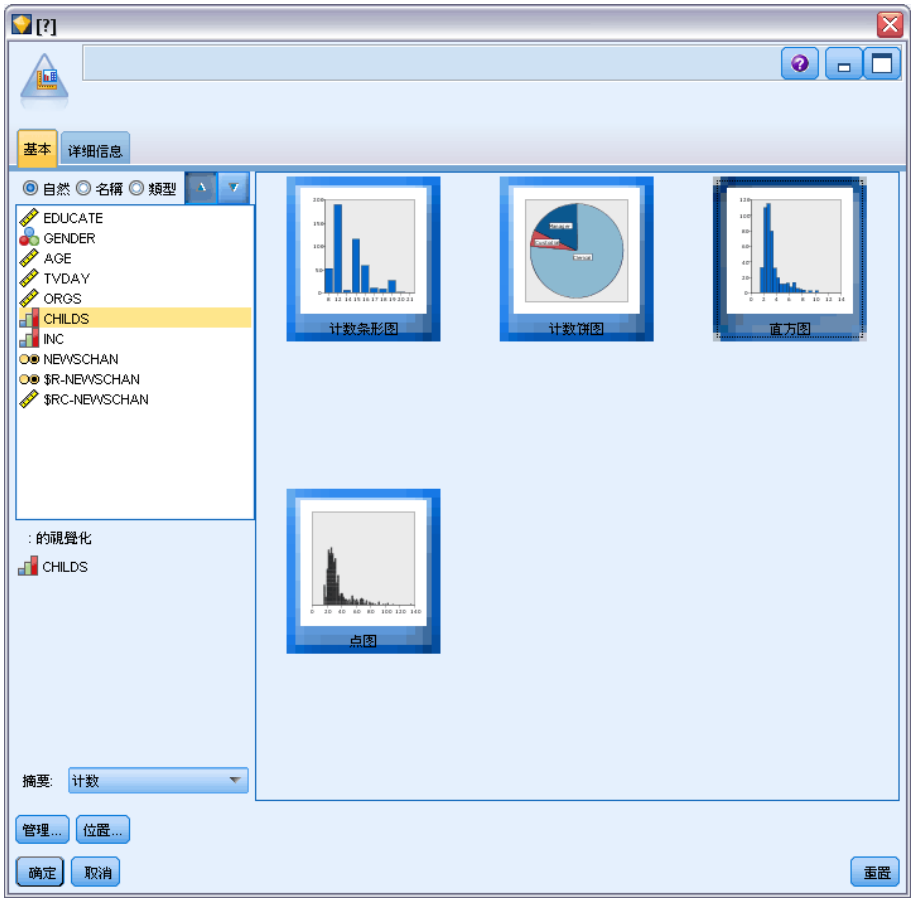
- 在块的“模型”选项卡上，展开左侧窗格中的条件和规则列表，并选择所需的一项。
- 在块的“查看器”选项卡上，展开分支列表，并选择所需的节点。
- 在交互树的“查看器”选项卡上，展开分支列表，并选择所需的节点。

  注意：您无法选择两个“查看器”选项卡上的顶部节点。

不管您选择以何种方式显示数据，创建图形的方式都是相同的：

▶ 从“生成”菜单选择**图形（从选择）**；或者在“查看器”选项卡上单击左下角处的**图形（从选择）**按钮。显示图形板“基本”选项卡。

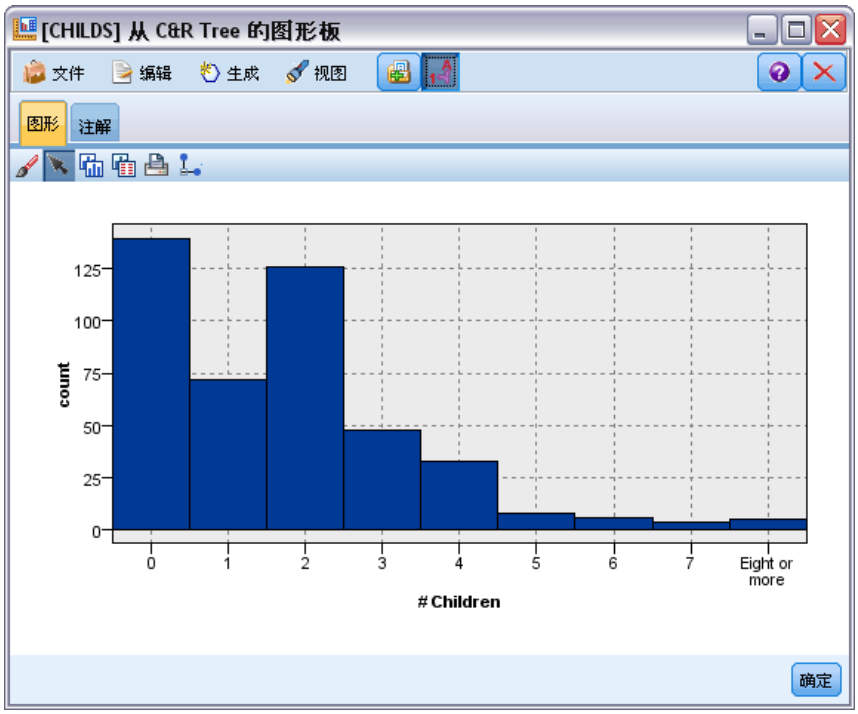
图片 6-42
“图形板”节点对话框，“基本”选项卡



注意：当您以此方式显示“图形板”时，只有“基本”和“详细”选项卡可用。 有关详细信息，请参阅第 5 章中的图形板节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

- 使用“基本”或“详细”选项卡设置指定在图形上显示的详细信息。
- 单击“确定”生成图形。

图片 6-43
从“图形板基本”选项卡生成的直方图



图形标题标识所包含的选定节点或规则。

用于推进、bagging 和超大型数据集的模型块

如果在建模节点上选择**增强模型准确性（推进）**、**增强模型稳定性（bagging）**或**为超大型数据集创建模型**作为主目标，则 IBM® SPSS® Modeler 会构建多个模型的整体。 有关详细信息，请参阅第 48 页码第 3 章中的整体模型。

结果模型块包含下列选项卡。“模型”选项卡提供模型的多个不同视图。

Tab	视图	描述	其他信息
模型	模型概要	显示整体质量和（增强型模型与连续目标除外）多样性概要，多样性是对不同模型之间预测结果变化情况的测量。	有关详细信息，请参阅第 49 页码第 3 章中的模型摘要。
	预测变量重要性	显示一个图表，以指示在估计模型时所使用的各个预测变量（输入字段）的相对重要性。	有关详细信息，请参阅第 50 页码第 3 章中的预测变量重要性。

Tab	视图	描述	其他信息
	预测变量频率	显示一个图表，以显 ❖❖❖在一组模型中与每个预测变量配合使用的相对频率图表。	有关详细信息，请参阅第 51 页码第 3 章中的预测变量频率。
	组件模型准确性	绘制整体中每个不同模型的预测准确性图表。	
	组件模型详细信息	显示整体中每个不同模型的相关信息。	有关详细信息，请参阅第 54 页码第 3 章中的组件模型详细信息。
	信息	显示有关字段、构建设置和模型评估过程的信息。	有关详细信息，请参阅第 44 页码第 3 章中的模型块概要/信息。
设置		允许您在评分操作中包括置信度。	有关详细信息，请参阅第 155 页码决策树/规则集模型块设置。
注解		允许您添加描述性注解，指定自定义名称，添加工具提示文本，以及指定模型的搜索关键字。	有关详细信息，请参阅第 5 章中的注解中的IBM SPSS Modeler 14.2 用户指南。

规则集模型块

规则集模型块表示用于预测特定输出字段的规则，该规则可由 Apriori 中的关联规则建模节点发现，也可由 C&R 树、CHAID、QUEST 或 C5.0 中的任一树构建节点发现。对于关联规则，必须从非精练规则块中生成规则集。对于树，可以从树构建器、C5.0 模型构建节点或任何树模型块中生成规则集。与非精练规则块不同，可将规则集块放置在流中以生成预测。

运行包含规则集块的流时，会将两个新字段（其中分别包含每个记录的预测值和置信度）添加到流中。新的字段名称将通过为模型名称添加前缀生成。对于关联规则集，预测字段的前缀是 $A-，置信度字段的前缀是 $AC-。对于 C5.0 规则集，预测字段的前缀是 $C-，置信度字段的前缀是 $CC-。对于 C&R 树规则集，预测字段的前缀是 $R-，置信度字段的前缀是 $RC-。在一个序列（可预测相同的输出字段）中具有多个规则集块的流中，新的字段名称将包括数字前缀，以便将这些名称区别开来。流中的第一个关联规则集块将使用普通的名称，第二个节点将使用以 $A1- 和 $AC1- 开头的名称，第三个节点使用以 $A2- 和 $AC2- 开头的名称，依此类推。

如何应用规则。从关联规则中生成的规则集与其他模型块不同，因为对于任何特定记录，都可以生成多个预测，且这些预测可能并不一致。可使用两种方法从规则集中生成预测。

注意：不论选择哪种方法，从决策树中生成的规则集都会返回相同的结果，因为从决策树中生成的规则会相互排斥。

- **投票。**此方法试图组合对应用到记录的所有规则的预测。对于每个记录，会检查所有的规则，并使用应用于该记录的每个规则生成一个预测和一个关联置信度。为每个输出值计算置信度图表的总和，具有最大置信度总和的值将被选作最终预测。最终预测的置信度是该值（由应用于该记录的规则数划分）的置信度总和。
- **第一个匹配项。**此方法仅按顺序测试规则，应用到记录的第一个规则也即用于生成预测的规则。

可在流选项中控制所使用的方法。 有关详细信息，请参阅第 5 章中的设置流选项中的IBM SPSS Modeler 14.2 用户 指南。

生成节点。通过“生成“菜单，可基于规则集创建新节点。

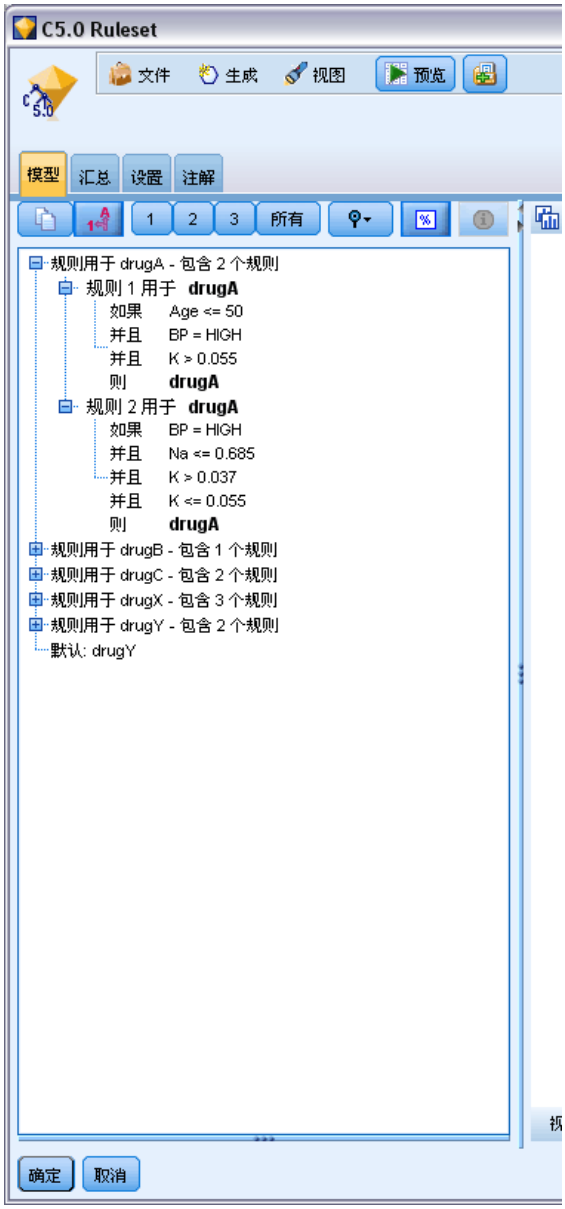
- **过滤节点。**创建新的过滤节点以过滤规则集中的规则所不使用的字段。
- **选择节点。**创建新的选择节点来选择选定规则要应用的记录。生成的节点将选择所应用规则的所有条件均为真的记录。此选项需要选定一个规则。
- **规则追踪节点。**创建可计算字段的新超节点，用来表示对每个记录进行预测时所使用的规则。当使用第一个匹配方法评估规则集时，仅用一个表明将触发第一个规则的符号来表示。当使用投票方法评估规则集时，则用一个显示投票机制的输入的复杂字符串来表示。
- **单一决策树（工作区）/单一决策树（GM 选项板）。**根据当前选定的规则创建一个新规则集块。仅适用于**增强型** C5.0 模型。 有关详细信息，请参阅第 158 页码增强型 C5.0 模型。
- **模型到调色板。**将模型返回到模型选项板。当有同事发给您包含模型的流而不是模型本身时，该功能很有用。

注意：规则集块中的“设置”和“汇总”选项卡与决策树模型中的这两个选项卡完全相同。

规则集模型选项卡

规则集块的“模型”选项卡中显示由算法从数据中提取的规则列表。

图片 6-44
规则集模型块，“模型”选项卡



规则按后项（预测类别）划分，并按下列格式显示：

```
如果antecedent_1
andantecedent_2
. . .
andantecedent_n
thenpredicted  value
```

其中 consequent 和 antecedent_1 - antecedent_n 都是条件。该规则可解释为“对于 antecedent_1 - antecedent_n 都为真的记录，consequent 也可能为真。”如果单击工具栏上的**显示实例/置信度**按钮，则每个规则还将显示有关应用该规则的条件为真的记录的数目信息（**实例**），及整个规则为真的记录的比例信息（**置信度**）。

注意，对于 C5.0 规则集，置信度的计算方式有些不同。C5.0 使用下列公式计算规则的置信度：

```
(1 + number of records where rule is correct) / (2 + number of records
for which the rule's antecedents are true)
```

这一置信度估计计算方式可调整从决策树中生成规则（即 C5.0 创建规则集时所执行的操作）的过程。

从 AnswerTree 3.0 中导入工程

IBM® SPSS® Modeler 可使用标准的“文件”>“打开”对话框导入在 AnswerTree 3.0 或 3.1 中保存的工程，示例如下：

- ▶ 从 SPSS Modeler 菜单中选择：
  文件 > 打开流
- ▶ 从文件类型下拉列表中选择 **AT 工程文件 (*.atp, *.ats)**。

  使用下列节点将导入的每个工程转换到 SPSS Modeler 流中：
  - 一个源节点，它可定义所使用的数据源（例如，IBM® SPSS® Statistics 数据文件或数据库源）。
  - 对于工程中的每个树（可能有多个树），将创建一个类型节点，该节点可为每个字段（变量）定义属性，包括类型、角色（对于预测变量字段为输入，对于预测字段为输出）、缺失值及其他选项。
  - 对于工程中的每个树，将创建一个“分区”节点，该节点可将数据分区为训练样本或测试样本，还将创建一个树构建节点，该节点可定义生成树（C&R 树、QUEST 或 CHAID 节点）的参数。
- ▶ 要查看生成的树，请运行该流。

注释

- 不能将在 SPSS Modeler 中生成的决策树导出到 AnswerTree 中；从 AnswerTree 导入 SPSS Modeler 是一个单向过程。
- 将工程导入到 SPSS Modeler 时，无法保留在 AnswerTree 中定义的利润。

章
7

贝叶斯网络模型

贝叶斯网络节点

通过**贝叶斯网络**节点，您可以利用对真实世界认知的判断力并结合所观察和记录的证据，通过使用看似不相关的属性建立事件发生的几率，从而构建概率模型。该节点重点应用了树扩展简单贝叶斯（TAN）和马尔可夫毯网络，这些算法主要用于分类问题。

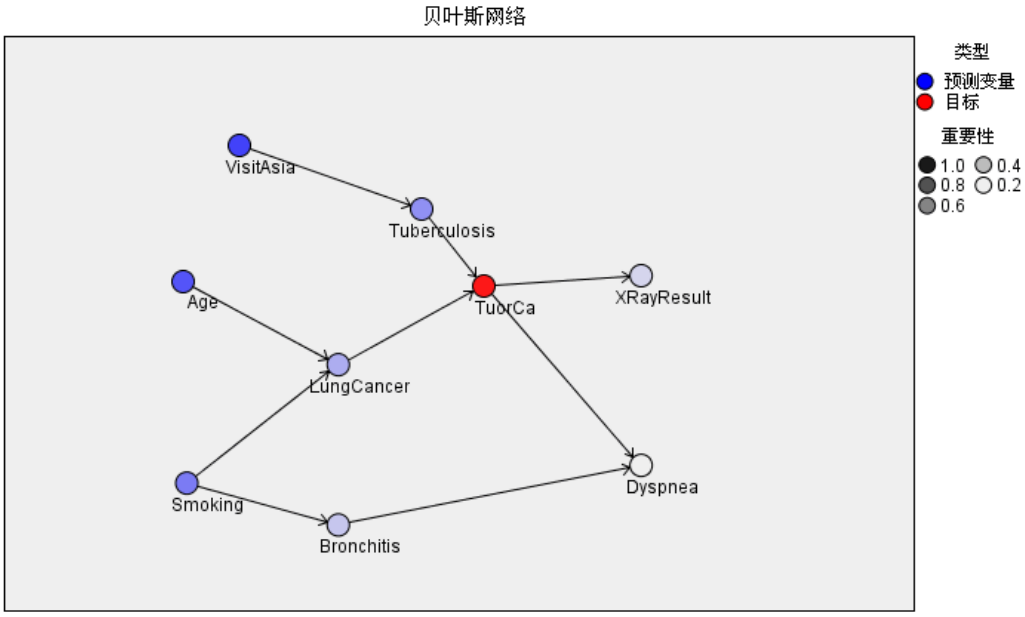
贝叶斯网络可用于在许多不同的情况下进行预测，示例如下：

- 选择违约风险较低的贷款时机。
- 根据传感器输入数据和现有记录，估算设备是否需要维修、增加零配件或更换。
- 借助在线故障排除工具解决客户问题。
- 实时诊断并排除移动电话网络故障。
- 评估研发项目的潜在风险和回报，以在最佳时机集中资源。

贝叶斯网络是一种图形模型，可显示数据集中的变量（通常称之为**节点**）以及概率，还可以显示这些变量之间的条件和独立性。贝叶斯网络可呈现节点之间的因果关系；但是，网络中的链接（也称为 **arcs**）没有必要呈现直接因果关系。例如，当指出是否存在某些症状并提供其他的相关数据时，如果图形中所显示的症状和疾病之间的概率独立性属实，则贝叶斯网络可用来计算患者患有某种特殊疾病的几率。这种网络非常稳健，即使在信息缺失时，也可以利用现有的任何信息作出最佳预测。

标准的基础贝叶斯网络示例由 Lauritzen 和 Spiegelhalter 于 1988 年创建。该网络示例是一种简化的网络版本，通常称作“Asia”模型，医生可用它来诊断新患者的病情，所有链接的方向可大体指示因果关系。每个节点代表与患者状况相关的一个方面，例如“吸烟”表示这些患者确为吸烟者，而“VisitAsia”表示他们最近是否去过亚洲。概率关系由所有节点之间的链接指示，例如，吸烟会增大患者患有支气管炎和肺癌的几率，而年龄仅与肺癌的患病率相关。同样地，肺部 x 光检查的异常结果可能是由肺结核或肺癌引起。同时，如果患者本身患有肺结核或肺癌，则其更有可能呈现出呼吸短促（呼吸困难）的症状。

图片 7-1
Lauritzen 和 Spegelhalter 的 Asia 网络示例



以下是您有可能决定使用贝叶斯网络的几点原因：

- 它可帮助您了解因果关系。由此，您可以了解出现问题的地方并可预测任何干涉可能引发的后果。
- 该网络可提供避免数据过度拟合的有效方法。
- 可以轻松地观测到所涉及关系的清晰视图。

要求。目标字段必须为分类且测量级别为名义、有序或标志。输入内容可以为任何类型的字段。连续（数值范围）输入字段将自动分级；但是，如果分布出现不对称，则可使用贝叶斯网络节点之前的分级节点对字段进行手动分级，从而获得更佳的效果。例如，在**主管字段**与贝叶斯网络节点**目标**字段相同的位置处，使用最优分级。 有关详细信息，请参阅第 4 章中的分级节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

示例。一位银行分析师希望能够预测可能拖欠偿还贷款的客户或潜在客户。您可使用贝叶斯网络模型标识最有可能拖欠还款的客户的特征，并由此构建几种不同类型的模型，以确定哪种类型可以最好地预测潜在的贷款拖欠者。有关详细信息，请参阅第 18 章中的预测贷款拖欠者（贝叶斯网络）中的IBM SPSS Modeler 14.2 应用程序 指南。

示例。一位电信运营商希望减少中断服务（又称为“流失”）的客户数量，并使用上一个月的数据对模型每月进行更新。您可以使用贝叶斯网络模型标识最有可能流失的客户的特征，然后每月使用新数据继续训练该模型。有关详细信息，请参阅第 19 章中的每个月重新训练模型（贝叶斯网络）中的IBM SPSS Modeler 14.2 应用程序 指南。

贝叶斯网络节点模型选项

图片 7-2
贝叶斯网络节点：“模型”选项卡

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

为每个分割构建模型。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。有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

分区。该字段允许您使用指定字段将数据分割为几个不同的样本，分别用于模型构建过程中的训练、测试和验证阶段。通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。如果已使用“类型”或“分区”节点定义了多个分区字段，则必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。（如果仅有一个分区字段，则将在启用分区后自动引入此字段。）有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。同时请注意，要在分析时应用选定分区，同样必须启用节点“模型选项”选项卡中的分区功能。（取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）

分割。对于分割模型，选择分割字段或字段。此操作与在“类型”节点中将字段的角色设置为分割类似。您仅可将测量级别为**标志**、**名义**、**有序**或**连续**的字段指定为分割字段。选为分割字段的字段无法用作目标、输入、分区、频率或权重字段。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

继续训练现有模型。如果选择此选项，则在模型块“模型”选项卡上显示的结果，将在每次运行模型时重新生成和更新。例如，如果已为现有模型添加新的或更新的数据源，则需要执行此操作。

注意：此操作只能更新现有网络，而不能添加或删除节点或连接。每次重新训练模型时，网络的形状都将保持不变，只会更改条件概率和预测变量重要性。如果新数据与旧数据大致相似也无妨，因为您所期望的是关注相同的内容；但是，如果您希望检查或更新重要的内容（针对其重要程度），则需要构建新模型，即构建新网络。

结构类型。选择构建贝叶斯网络时使用的结构：

- **TAN。**树扩展朴素贝叶斯模型（TAN）可创建简单的贝叶斯网络模型，它是一种基于标准朴素贝叶斯模型的改进模型。这是由于该模型允许每一个预测变量除了依赖于目标变量之外，还依赖于其他预测变量，由此增加分类的准确度。
- **马尔可夫覆盖。**可以在包含目标变量的父项、子项及其子项的父项的数据集中选择节点集。马尔可夫覆盖基本上标识了需要预测目标变量的网络中的所有变量。用户认为这种构建网络的方法更为准确；但是，当处理大型数据集时，由于所包含的变量数较多，所以可能会消耗许多处理时间。要减少处理工作量，可以使用“专家”选项卡上的**特征选择**选项，选择与目标变量有重大相关性的变量。

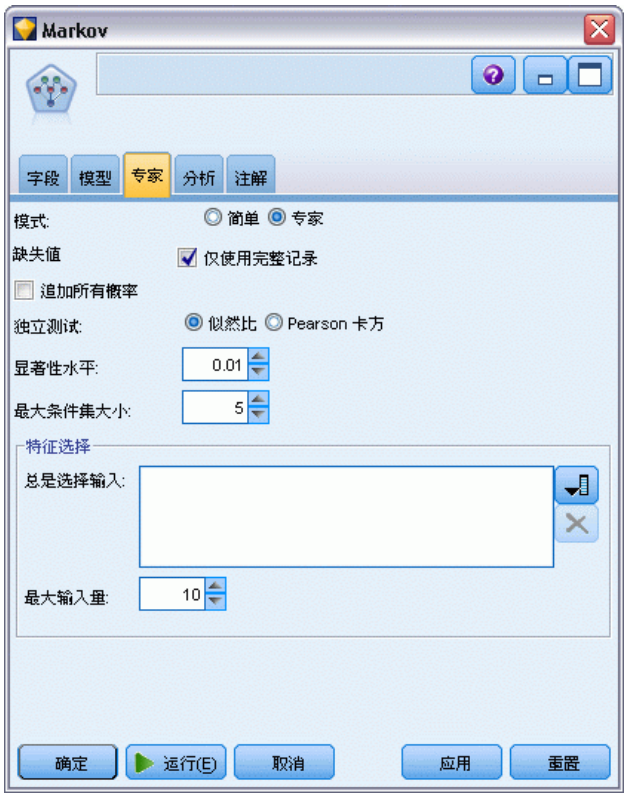
包括特征选择预处理步骤。选择该框，您可以使用“专家”选项卡上的**特征选择**选项。

参数学习方法。对于值为其父项值的每个节点来说，贝叶斯网络参数是指条件概率。有两种可能的选择，您可以用来控制估算节点（此处父项值已知）间条件概率表这一任务。

- **最大似然法。**使用大型数据集时，请选择此框。这是默认选项。
- **对小单元格计数的贝叶斯调整。**对于较小的数据集，可能存在模型过度拟合的风险以及出现大量零计数的可能性。选中此选项可通过应用平滑来减少任何零计数以及不可靠的估计结果带来的影响，从而解决这些问题。

贝叶斯网络节点专家选项

图片 7-3
贝叶斯网络节点：“专家”选项卡



使用节点专家选项可微调模型构建过程。要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

缺失值。默认情况下，IBM® SPSS® Modeler 将仅使用对于模型中使用的所有字段均具有有效值的记录。（这种方式有时称为缺失值的**成列删除**。）如果有很多缺失数据，您可能会发现这种方式去除的记录过多，剩余记录不足以生成较好的模型。在这种情况下，您可以取消选中**仅使用完整记录**选项。SPSS Modeler 随后将尝试使用尽可能多的信息对模型进行估计，其中包括部分字段存在缺失值的记录。（这种方式有时称为缺失值的**成对删除**。）但在某些情形下，以这种方式使用不完整记录可能会在模型的估计过程中产生计算问题。

追加所有概率。指定是否将输出字段每个类别的概率添加到该节点所处理的每个记录。如果未选中此选项，则仅添加预测类别的概率。

独立性检验。一种独立评估检验，可估计两个变量中成对的观测值是否彼此独立。请从以下可用选项中选择要使用的检验类型：

- **似然比。**通过计算两种不同假设下结果概率的最大值之间的比率来检验目标-预测变量的独立性。
- **Pearson 卡方。**通过使用原假设（所观察事件的相对出现频率遵循特定的频率分布）来检验目标-预测变量的独立性。

贝叶斯网络模型可在检验对之外使用附加变量执行独立性的条件检验。此外，模型不仅可以研究目标和预测变量之间的关系，还可研究预测变量自身之间的关系。

注意：只有在“模型”选项卡上选中马尔可夫覆盖的**包括特征选择预处理步骤**或**结构类型**时，才能使用独立性检验选项。

显著性水平。可以与独立性检验设置结合使用，通过此选项，您可以在执行检验时设置要使用的分隔值。该值越小，网络中的链接就越少；默认水平值为 0.01。

注意：只有在“模型”选项卡上选中马尔可夫覆盖的**包括特征选择预处理步骤**或**结构类型**时，才能使用该选项。

最大条件集大小。该算法可用于创建马尔可夫覆盖结构，它可使用增加大小的条件集执行独立性检验，并从网络中删除不需要的链接。由于包含大量条件变量的检验需要更多的时间和内存进行处理，因此您可以限制要包括的变量数目。在处理众多变量间具有较强独立性的数据时，这种操作非常有用。但请注意，最终形成的网络可能包含一些多余链接。

指定执行独立性检验时要使用的条件变量的最大数目。默认设置为 5。

注意：只有在“模型”选项卡上选中马尔可夫覆盖的**包括特征选择预处理步骤**或**结构类型**时，才能使用该选项。

功能选择。使用这些选项，您可以限制在处理模型时所使用的输入量以加速模型构建过程。由于在创建马尔可夫覆盖结构时存在大量的潜在输入，因此该操作特别有用；通过此项操作，您可以选择与目标变量有重大关联的输入。

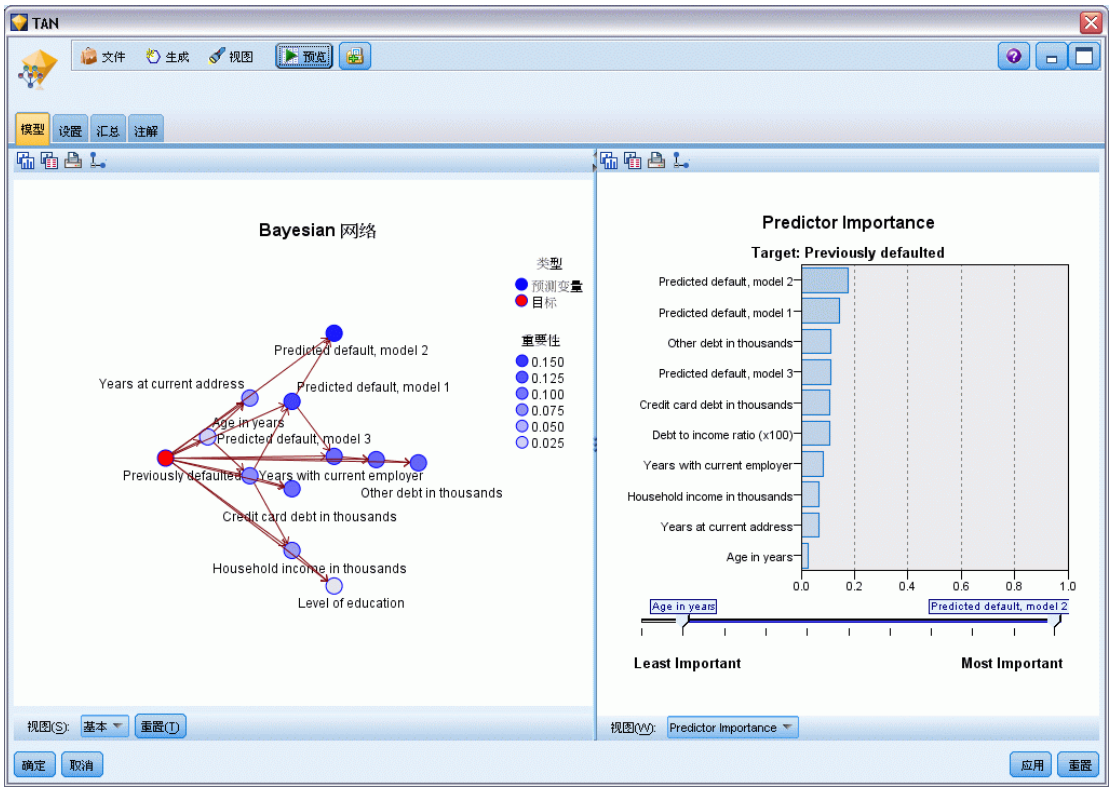
注意：只有在“模型”选项卡上选中**包括特征选择预处理步骤**时，才能使用特征选择选项。

- **始终选择输入**通过使用“字段选择器”（文本字段右侧的按钮），从数据集中选择在构建贝叶斯网络模型时始终使用的字段。注意，请始终选择目标字段。
- **最大输入量。**在数据集中指定构建贝叶斯网络模型时要使用的总输入量。您可以输入的最大数目为数据集中的总输入量。

  注意：如果在**总是选择输入**中选择的字段数超过**最大输入量**的值，则会显示一条错误消息。

贝叶斯网络模型块

图片 7-4
贝叶斯网络及关联预测变量重要性模型详细信息



注意：如果在“模型”选项卡中选中了**继续训练现有参数**，则会在每次重新生成模型时更新模型块“模型”选项卡上显示的信息。

模型块“模型”选项卡分为两个窗格：

左窗格

基本。该视图包含节点网络图，可显示目标与其最重要预测变量之间的关系，以及预测变量自身之间的关系。各预测变量的重要性可通过其颜色的深浅显示；颜色越深表示变量越重要，反之亦然。

当您将鼠标指针悬停在节点上时，弹出式工具提示中会显示代表范围的节点的分级值。

可以使用 IBM® SPSS® Modeler 的图表工具进行交互、编辑，并保存图表。例如，可以在其他应用程序如 MS Word 中使用图表。

提示：如果网络包含大量节点，则可以点击某个节点，然后拖动它以使图形更加清晰。

分布。该视图将以微型图形的格式显示网络中各个节点的条件概率。将鼠标悬停在图形上方，可在弹出式工具提示中显示图形值。

右窗格

预测变量重要性。这将显示一个图表，以指示在估计模型时所使用的各个预测变量的相对重要性。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

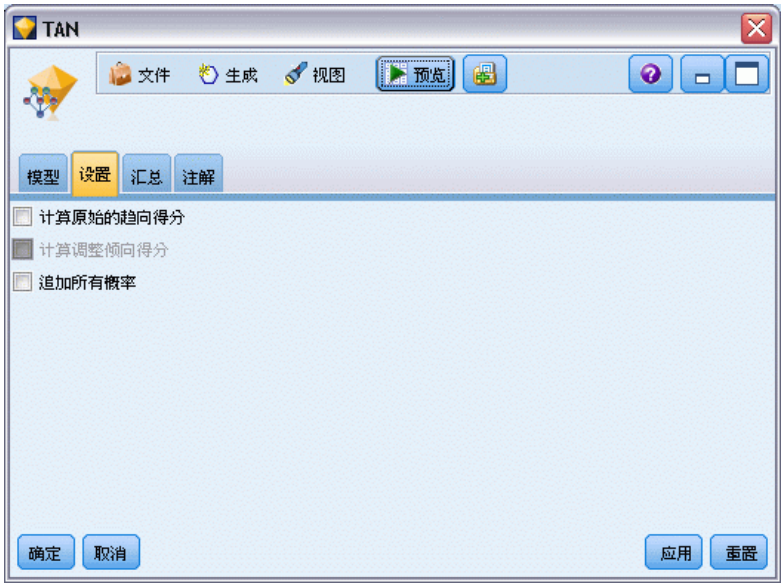
条件概率。当在左窗格中选择了某个节点或微型分布图时，右窗格则会显示相关的条件概率表。该表包含各个节点值的条件概率值，以及各节点的父节点中的值组合。此外，该表还包含为每个记录值和父节点中各个值组合所观测的记录数量。

贝叶斯网络模型设置

在贝叶斯网络模型块的“设置”选项卡中可指定选项以修改已构建的模型。例如，可以通过贝叶斯网络节点使用相同的数据和设置构建几个不同的模型，然后使用每个模型中的此选项卡对设置稍做修改以查看其对结果的影响。

注意：只有将模型块添加到流中之后，此选项卡才可用。

图片 7-5
贝叶斯网络模型的“设置”选项卡



计算原始的倾向得分。对于含标志目标（返回?是?或?否?预测）的模型，您可以请求倾向得分，这些得分指示为目标字段指定结果为真的可能性。除了这些得分，还有其他在评分过程中生成的预测值和置信度值。

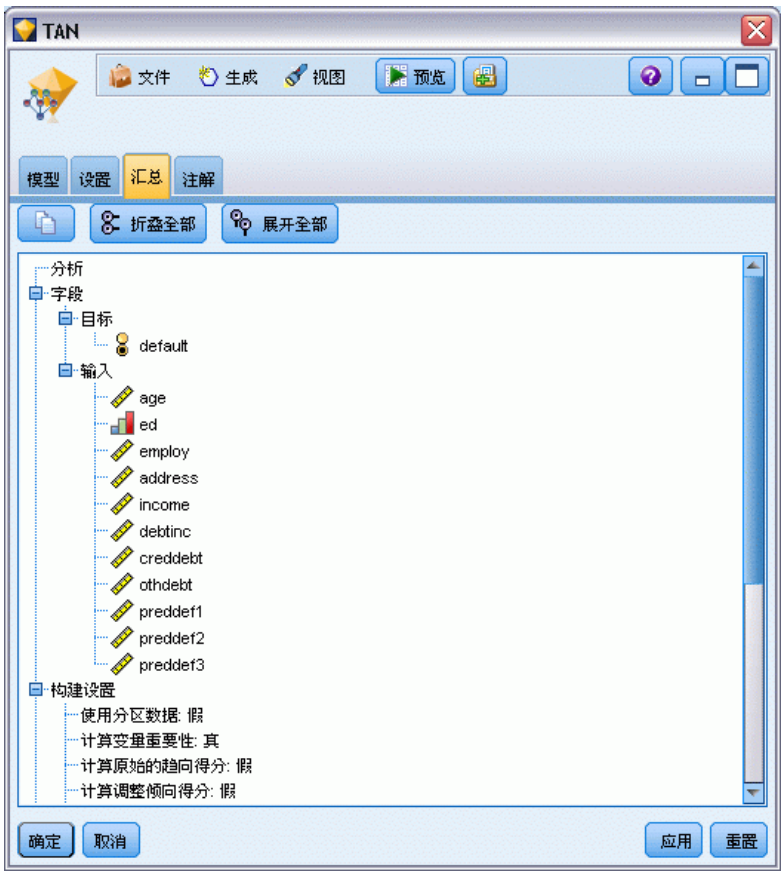
计算调整后的倾向得分。原始倾向得分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该得分可能会过度优化。调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。此选项要求在流中定义分区字段并且在生成模型之前在建模节点中启用调整的倾向得分。

追加所有概率。指定是否将输出字段每个类别的概率添加到该节点所处理的每个记录。如果未选中此选项，则仅添加预测类别的概率。

此复选框的默认设置由建模节点的“专家”选项卡上的相应复选框确定。 有关详细信息，请参阅第 169 页码贝叶斯网络节点专家选项。

贝叶斯网络模型摘要

图片 7-6
贝叶斯网络模型的“摘要”选项卡



模型块的“汇总”选项卡显示了有关模型的下列信息：模型本身（分析）、模型中使用的字段（字段）、构建模型时使用的设置（构建设置）和模型训练（训练概要）。

当第一次浏览此节点时，“汇总”选项卡的结果是折叠起来的。要查看感兴趣的结果，可使用项目左侧的展开控件展开项目，或单击**全部展开**按钮显示所有结果。当结束对项目的查看时，为了隐藏结果，可使用展开控件折叠要隐藏的特定结果，或单击**全部折叠**按钮折叠所有结果。

分析。 显示指定模型的相关信息。

字段。 列出构建模型时用作目标和输入的字段。

构建设置。 包含有关在构建模型中使用的设置的信息。

训练概要。 显示模型类型、用于创建模型的流、模型创建者、模型构建完成时间和模型构建所用时间。

章 8

神经网络

神经网络可以近似多种预测模型，而对模型结构和假设只有最小需求。关系形式在学习过程中确定。如果目标与预测变量间的线性关系适当，神经网络结果会非常接近传统线性模型的结果。如果非线性关系更为适当，神经网络会自动接近“正确”的模型结构。

伴随这种灵活性的缺点是神经网络往往不太容易解释。因此，如果您试图解释生成目标与预测变量间关系的底层过程，最好使用更传统的统计模型。不过，如果模型的可解释性并不重要，您可以使用神经网络以获得良好的预测结果。

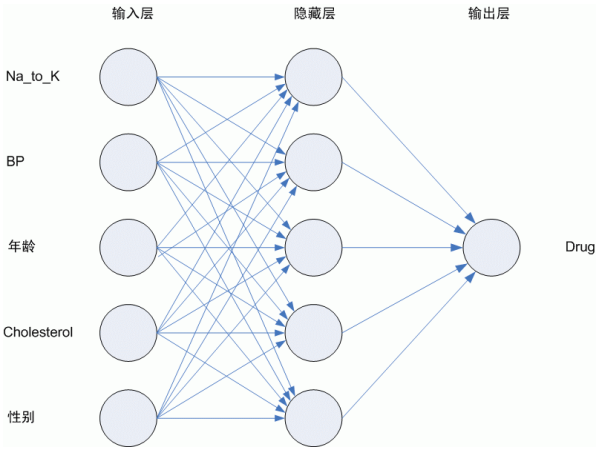
图片 8-1
“字段”选项卡

字段要求。必须至少有一个目标和一个输入。设为“两者”或“无”的字段将被忽略。对目标或预测变量（输入）没有测量级别限制。 有关详细信息，请参阅第 30 页码第 3 章中的建模节点字段选项。

神经网络模型

神经网络是神经系统运转方式的简单模型。其基本单元是**神经元**，通常将其组织到**层**中，如下面的图所示。

图片 8-2
神经网络的结构



神经网络是模拟人类大脑处理信息方式的简化模型。此模型通过模拟大量类似于神经元的抽象形式的互连处理单元而运行。

这些处理单元都位于层中。通常在神经网络中有三个部分：一个**输入层**，其中的单元表示输入字段；一个或多个**隐藏层**；一个**输出层**，带有一个或多个表示目标字段的单元。这些单元通过可变的连接强度（或**权重**）连接。输入数据显示在第一层，其值从每个神经元传播到下一层的每个神经元。最终从输出层中输出结果。

该网络可通过以下过程进行学习，即检查单个记录，然后为每个记录生成预测，并且当生成的预测不正确时，对权重进行调整。在满足一个或多个停止标准之前，此过程会不断重复，而网络会持续提高其预测准确度。

最初，所有的权重都是随机生成的，并且从网络输出的结果很可能没有意义的。网络可通过**训练**来学习。向该网络重复应用已知道结果的示例，并将网络给出的结果与已知的结果进行比较。从此比较中得出的信息会传递回网络，并逐渐改变权重。随着训练的进行，该网络对已知结果的复制会变得越来越准确。一旦训练完毕，就可以将网络应用到未知结果的未来案例中。

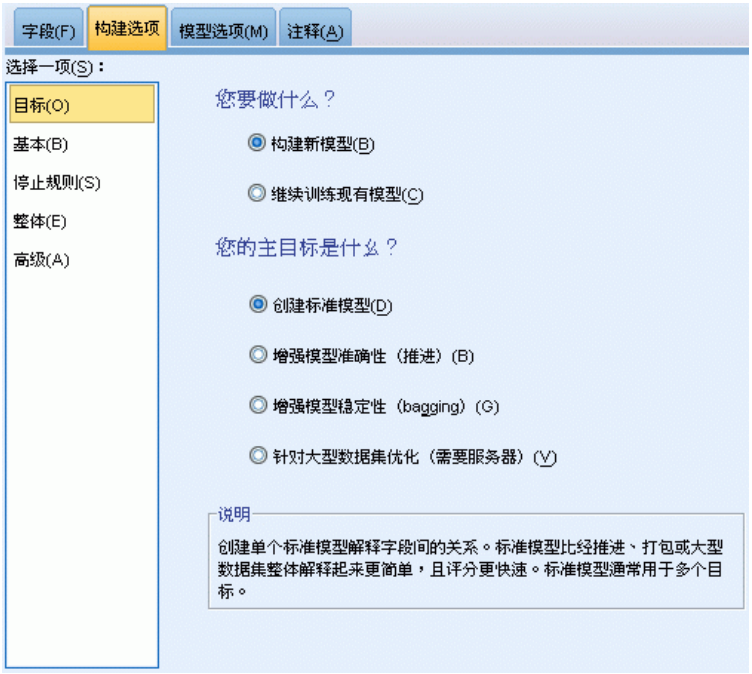
对遗存流使用神经网络

IBM® SPSS® Modeler 版本 14 引入了新的神经网络节点，支持推进和 bagging 技术，并可针对大型数据集进行优化。在此发行版中，包含旧节点的现有流仍然可以构建模型并对其评分。不过，在将来的发行版中将去掉这一支持，因此建议您从现在起使用新的版本。

从版本 13 以后，带有未知值的字段（即，在培训数据中不存在值）不再被自动处理为缺失值，而使用 $null$ 值来进行评分。因此，如果您要在版本 13 或以后版本中使用旧的（13 以前版本）神经网络模型，以便在对字段进行评分时将未知值处理为非空值，则应将未知值标记为缺失值（例如，使用类型节点）。

目标

图片 8-3
目标设置



您希望做什么？

- **构建新的模型。**构建全新的模型。这是该节点的常用操作。
- **继续训练现有模型。**继续训练此节点成功生成的最后一个模型。这样就可以在无需访问原始数据的情况下更新或刷新现有的模型，并可能会显著提升性能，因为只有新的或更新后的记录被传入流中。上一个模型的详细信息与建模节点存储在一起，这样即使先前的模型块在流或模型选项板中不再可用的情况下，也可以使用该选项。

  注意：在启用此选项后，“字段”和“构建选项”选项卡上的所有其他控件将被禁用。

您的主要目标是什么？

- **创建标准模型。**此方法将构建单个模型，以使用预测变量来预测目标。一般来说，与推进、bagged 或大数据集整体模型相比，标准模型更易于解释，并能更快地进行评分。
- **增强模型准确性（推进）。**此方法采用推进方式构建整体模型，这将生成一系列模型以获得更精确的预测结果。与标准模型相比，此整体模型需要更长的构建与评分时间。

  推进方法产生一系列“成分模型”，其中每个模型在整个数据集上构建。在构建每个后续成分模型之前，将根据前一成分模型的残差对记录进行加权。具有较大残差的个案将被给予较高的分析权重，因此下一个成分模型将较好地侧重于这些记录。

这些成分模型共同构成一个整体模型。该整体模型采用组合规则对新记录进行评分。可用的规则取决于目标的测量级别。

- **增强模型稳定性（bagging）。**此方法采用 bagging（Bootstrap 汇总）方式构建整体模型，这将生成多个模型以获得更可靠的预测结果。与标准模型相比，此整体模型需要更长的构建与评分时间。

  Bootstrap 汇总 (bagging) 通过对原始数据集进行放回抽样，产生训练数据集的副本。这将创建大小与原始数据集相同的 bootstrap 样本。然后，在每个副本上构建“成分模型”。这些成分模型共同构成一个整体模型。该整体模型采用组合规则对新记录进行评分。可用的规则取决于目标的测量级别。

- **创建适用于大型数据集（需要 IBM SPSS Modeler Server）的模型。**此方法将数据集划分为多个单独数据块，以构建整体模型。如果您的数据集过大，而无法构建上述任何模型或进行增量式建模，请选择此项。与标准模型相比，此选项的构建时间较短，但评分时间更长。该选项需要 SPSS Modeler Server连接。

如果存在多个目标，则此方法将只创建标准模型，而不考虑所选的目标。

基本

图片 8-4
基本设置

神经网络模型。此类模型确定神经网络如何通过隐藏层将预测变量连接到目标。**多层感知器(MLP)**允许构建较为复杂的关系，但代价是更长的训练与评分时间。**径向基函数(RBF)**可以缩短训练与评分时间，但与 MLP 相比其预测能力要差些。

隐藏层。神经网络的隐藏层包含无法观察到的单元。每个隐藏单元的值均为预测变量的某个函数；函数的确切形式部分取决于网络类型。多层感知器可以有一个或两个隐藏层；径向基函数可以有一个隐藏层。

- **自动计算单元数。**此选项构建具有单个隐藏层的网络，并计算隐藏层中的“最佳”单元数。
- **自定义单元数。**此选项允许您指定每个隐藏层中的单元数。第一个隐藏层必须至少有一个单元。如果为第二个隐藏层指定 0 个单元，则会构建具有单个隐藏层的多层感知器。

  注意：在选择值时，应确保节点数不超过连续预测变量数加上所有分类（标志、名义和有序）预测变量间的类别总数之和。

停止规则

图片 8-5
停止规则设置

这些规则用于确定何时停止训练多层感知器网络；如果使用径向基函数，将忽略这些设置。训练在持续至少一个周期（数据传递）后，可以按照下列条件被停止。

使用最长训练时间（每个成分模型）。选择是否指定算法运行的最大分钟数。请指定一个大于 0 的数字。当构建整体模型时，此为其中每个成分模型的允许训练时间。请注意，为了完成当前周期，训练可能会比指定的时间限制延长一点。

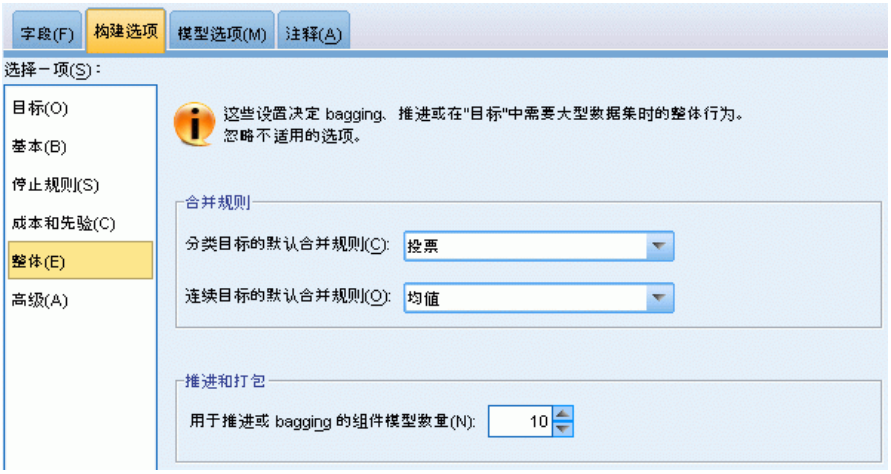
自定义最大训练周期数。允许的最大训练周期数。如果超过最大周期数，则停止训练。指定大于 0 的整数。

使用最小精确性。如果使用此选项，训练则会一直继续，直到达到指定的精确性。这种情况可能永远不会出现，但您可以随时中断训练，以截止到目前所达到的最佳精确性保存该网络。

如果每个周期后防止过度拟合集合中错误未出现减少，训练错误的相对变化较小，或当前训练错误的比率相对于初始错误较低，训练算法也会停止。

整体

图片 8-6
整体设置



这些设置决定了在“目标”中请求 boosting、bagging 或超大型数据集时发生的整体行为。对选定目标不适用的选项将被忽略。

Bagging 和大型数据集。在对整体评分时，此规则用于组合来自基本模型的预测值，以计算整体得分值。

- **分类目标的默认组合规则。**可以通过投票、最高概率或最高平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。**最高平均值概率**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。
- **连续目标的默认组合规则。**可以通过对来自基本模型的预测值取平均值或中位数，对连续目标的整体预测值进行组合。

注意，如果以增强模型精确性为目标，则组合规则选择将被忽略。Boosting 方法始终使用加权大多数投票来对分类目标进行评分，而使用加权中位数对连续目标进行评分。

Boosting 和 Bagging。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 bagging 方法，此为 bootstrap 样本数。它应为正整数。

高级

图片 8-7
高级设置

高级设置允许用户控制与其他设置组不完全吻合的选项。

防止过度拟合集合。神经网络方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。默认值为 30。

重复结果。设置随机种子允许您复制分析。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。默认情况下，使用种子 229176228 来复制分析。

预测变量缺失值。这将指定如何处理缺失值。**成列删除**将在预测变量上存在缺失值的记录从模型构建中排除。**插补缺失值**将替换预测变量中的缺失值，并在分析中使用这些记录。连续字段会插补最小与最大观测值的平均值；分类字段则插补最常出现的类别。请注意，在“字段”选项卡指定的任何其他字段上均具有缺失值的记录始终会从模型构建中排除。

模型选项

图片 8-8
“模型选项”选项卡

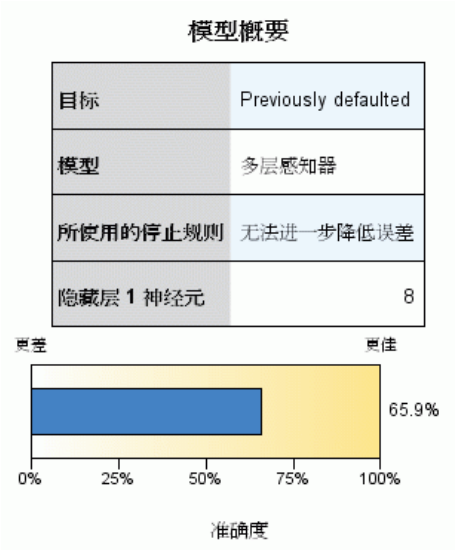
模型名称。可以基于目标字段来自动生成模型名称，或指定自定义名称。自动生成的名称为目标字段名。如果存在多个目标，则模型名称将由这些字段名按顺序排列组成，且字段名之间通过“与”（&）符号连接。例如，如果目标分别为 field1、field2 和 field3，则模型名称为：field1 & field2 & field3。

可用于评分。在对模型评分时，应生成此组中的选定项目。在对模型评分时，始终会计算预测值（适合所有目标）和置信度（适合分类目标）。计算的置信度可以基于预测值的概率（最高预测概率）或最高预测概率与第二高预测概率之间的差值。

- **分类目标的预测概率。**这将生成分类目标的预测概率。为每个类别创建一个字段。
- **标志目标的倾向得分。**对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向得分，这些得分指示为目标字段指定结果为真的可能性。该模型产生原始倾向得分；如果分区处于有效，则模型还会根据测试分区产生调整后的倾向得分。有关详细信息，请参阅第 36 页码第 3 章中的倾向得分。

模型摘要

图片 8-9
神经网络模型摘要视图



“模型摘要”视图是神经网络预测或分类精确性的快照摘要。

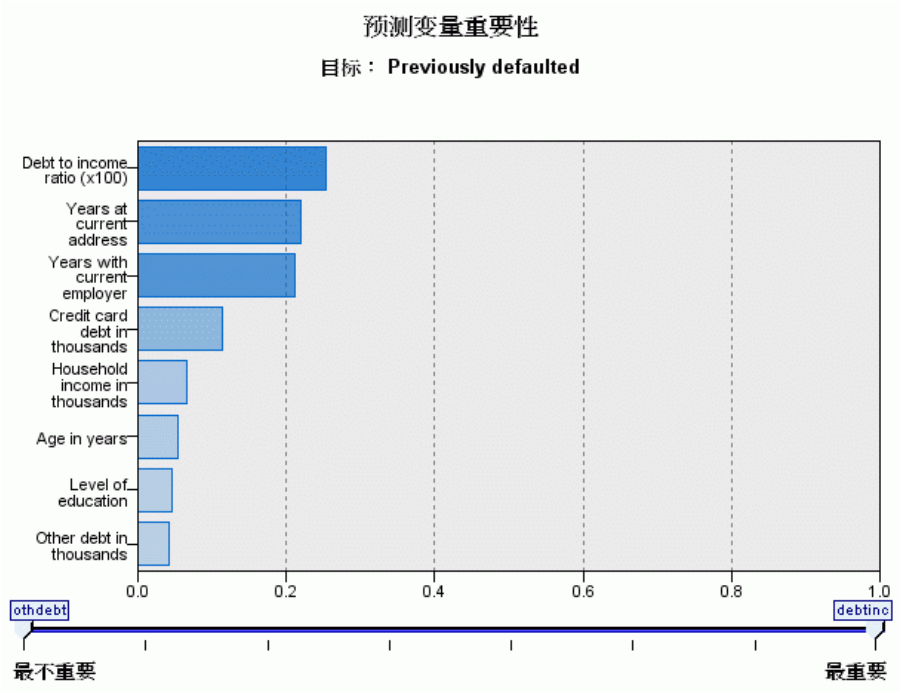
模型摘要。此表标识目标、已训练的神经网络类型、停止训练的停止规则（已训练多层感知器网络时显示），以及网络的每个隐藏层中的神经元数。

神经网络质量。此图表显示最终模型的精确性，数值越大越好。对于分类目标，它仅仅是预测值与观测值相符的记录百分比。对于连续目标，它等于 1 减去预测结果中的绝对平均误差（预测值减去观测值的绝对值的平均值）与预测值范围（最大预测值减去最小预测值）的比率。

多个目标。如果存在多个目标，则每个目标显示在表的**目标**行中。在图表中显示的精确性是单独目标精确性的平均值。

预测变量重要性

图片 8-10
预测变量重要性视图

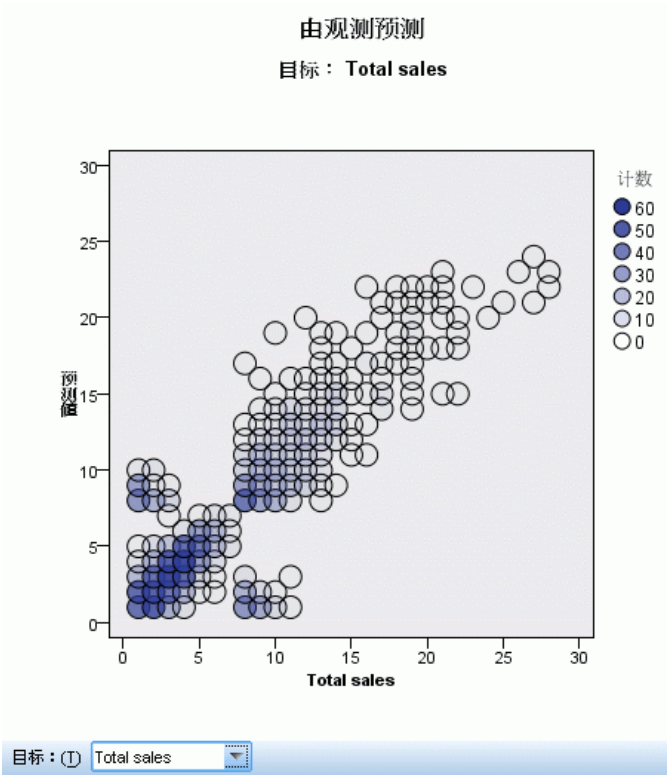


通常，您将需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的变量。预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。预测变量的重要性与模型精确性无关。它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

多个目标。如果存在多个目标，则每个目标显示在单独图表中，并提供有**目标**下拉列表，以控制显示哪个目标。

按已观测进行预测

图片 8-11
按已观测进行预测视图

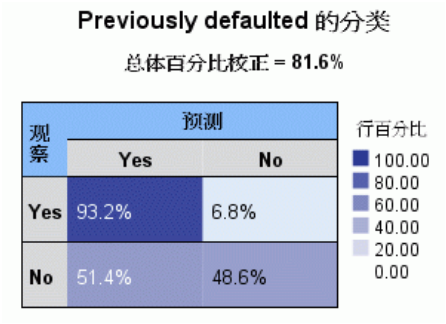


对于连续目标，这将显示一个分级散点图，其中预测值位于垂直轴上，而观测值位于水平轴上。

多个目标。如果存在多个连续目标，则每个目标显示在单独图表中，并提供有**目标**下拉列表，以控制显示哪个目标。

分类

图片 8-12
分类视图，行百分比样式



对于分类目标，这将在一个热图中显示已观测与已预测值的交叉分类，以及整体正确百分比。

表样式。有多种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

- **行百分比。**这将在单元格中显示行百分比（单元格计数表示为行总数的百分比）。这是默认值。
- **单元格计数。**这将在单元格中显示单元格计数。热图中的阴影仍然基于行百分比。
- **热图。**这将只显示阴影，不会在单元格中显示值。
- **压缩。**这将不会在单元格中显示行或列标题，也不会显示值。它在目标具有较多类别时非常有用。

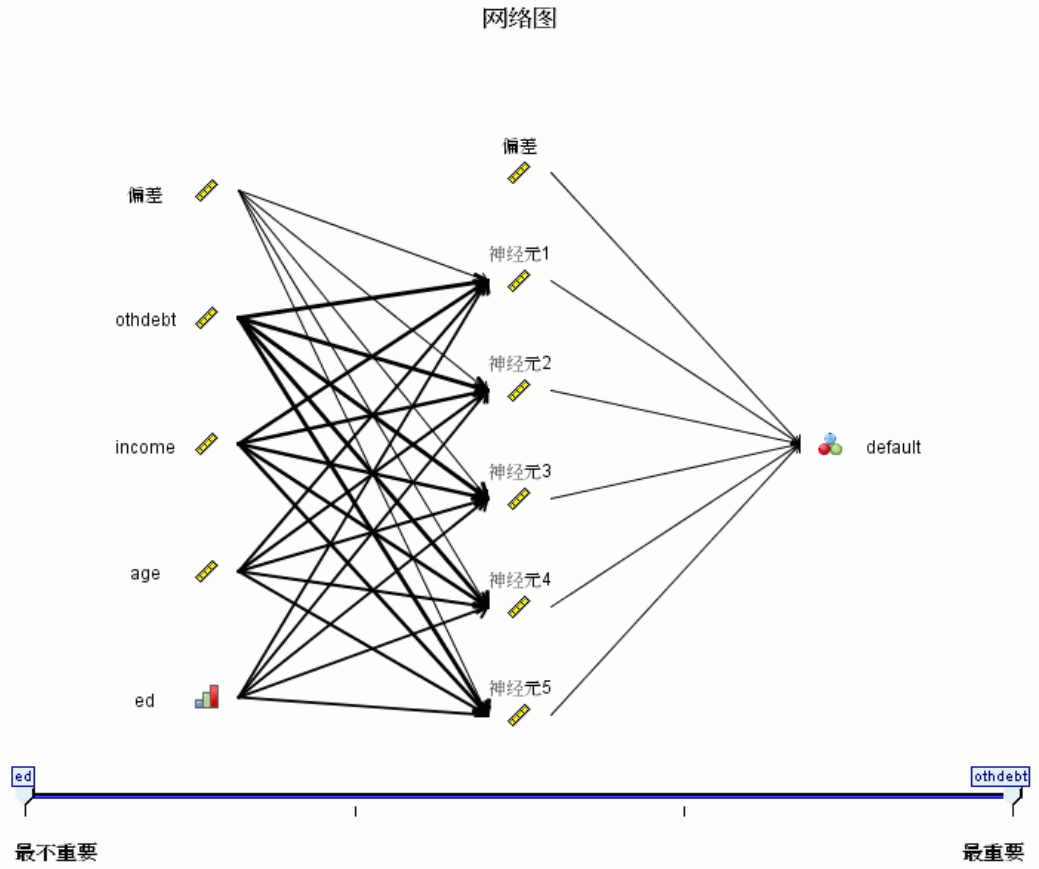
缺失。如果任何记录在目标上具有缺失值，则它们会显示在所有有效行下方的（**缺失**）行中。具有缺失值的记录对整体正确百分比没有贡献。

多个目标。如果存在多个分类目标，则每个目标显示在单独表中，并提供有**目标**下拉列表，以控制显示哪个目标。

大型表。如果显示的目标具有超过 100 个类别，则不会显示表。

网络

图片 8-13
网络视图，输入位于左侧，效果样式



这将显示神经网络的图形表示。

图表样式。有两种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

- **效果。**这会在图表中将每个预测变量与目标显示为单个节点，不论测量尺度是连续还是分类。这是默认值。
- **系数。**这将为分类预测变量与目标显示多个指示节点。在系数样式图表中，连接线条根据估计的键结值显示为不同颜色。

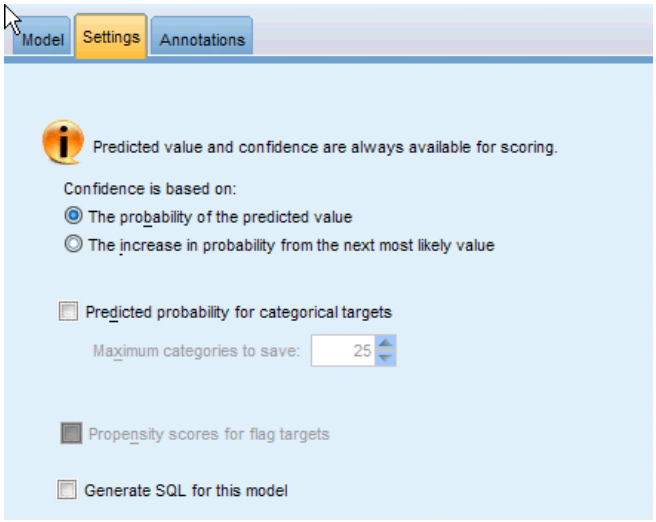
图表方向。默认情况下，输入位于网络图表中的左侧，而目标位于右侧。通过工具栏控件可以更改图表方向，以使输入位于顶部，目标位于底部，反之亦可。

预测变量重要性。在图表中，连接线条根据预测变量的重要性进行加权，粗线条表示重要性较高。在工具栏中提供有一个“预测变量重要性”滑块，以控制在网络图表中显示哪些预测变量。这不会改变模型，只是帮助您重点关注最重要的预测变量。

多个目标。如果存在多个目标，所有目标都将显示在图表中。

设置

图片 8-14
“设置”选项卡



在对模型评分时，应生成此选项卡中的选定项目。在对模型评分时，始终会计算预测值（适合所有目标）和置信度（适合分类目标）。计算的置信度可以基于预测值的概率（最高预测概率）或最高预测概率与第二高预测概率之间的差值。

- **分类目标的预测概率。**这将生成分类目标的预测概率。为每个类别创建一个字段。
- **标志目标的倾向得分。**对于含标志目标（返回“是”或“否”预测）的模型，您可以请求倾向得分，这些得分指示为目标字段指定结果为真的可能性。该模型产生原始倾向得分；如果分区处于有效，则模型还会根据测试分区产生调整后的倾向得分。有关详细信息，请参阅第 36 页码第 3 章中的倾向得分。

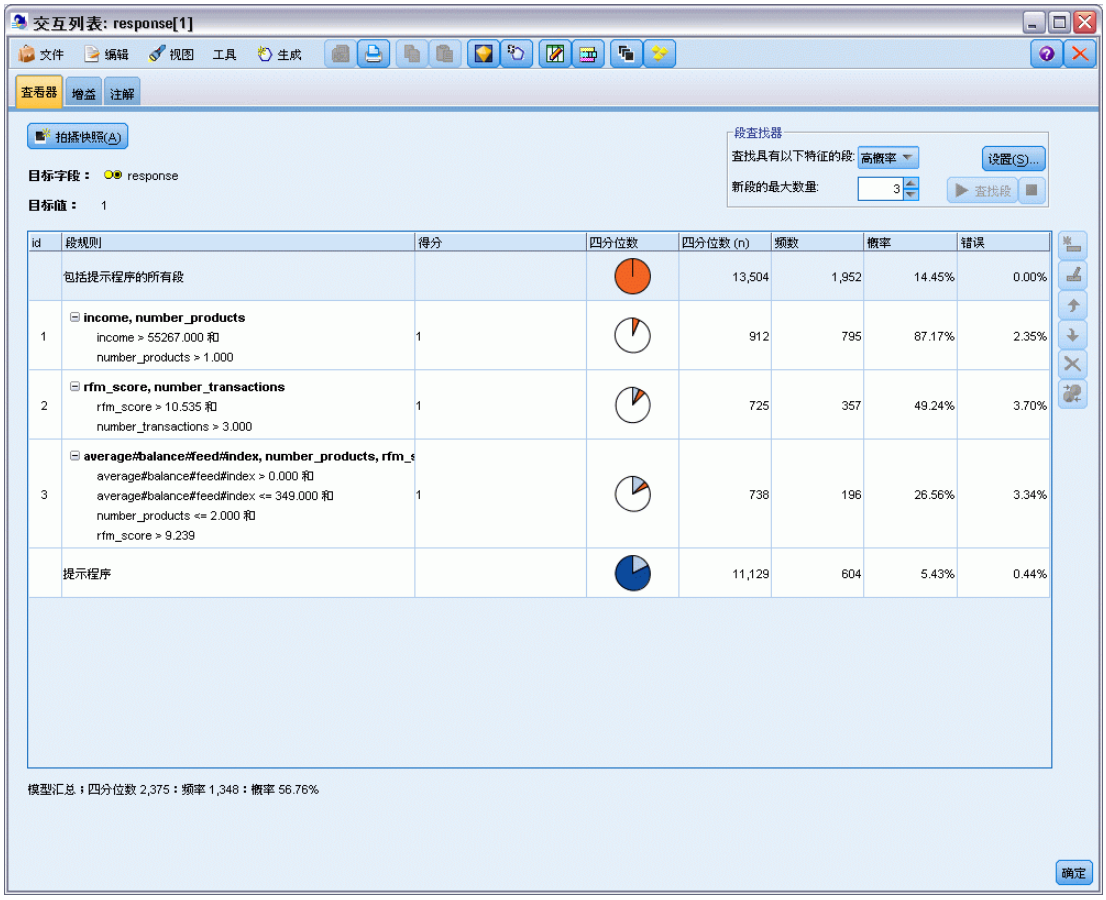
生成此模型的 SQL。使用数据库中的数据时，SQL 代码可传回到数据库中执行，从而大大提高许多操作的处理速度。有关详细信息，请参阅第 6 章中的SQL 优化中的IBM SPSS Modeler Server 14.2 管理和性能指南。

章 9

决策表

Decision List 模型标识了子组或**段**，即，显示了与整体样本相关的二值（yes 或 no）结果的似然度的高低。例如，您或许在寻找那些最不可能流失的客户或最有可能对某个商业活动作出积极响应的客户。通过 Decision List Viewer 可以实现对模型的完全控制，它允许您编辑段、添加自己的业务规则、指定每个段的评分方式，以及采用其他多种方式对所有段的匹配比例进行优化。因此，它尤其适用于生成邮件列表，或确定作为特定活动目标的记录。此外，还可以使用多个**挖掘任务**对不同建模方法进行组合，例如，确定同一模型中性能较高和较低的段，并根据需要在评分阶段包含或排除每个段。

图片 9-1
决策列表模型



段、规则和条件

模型由段列表组成，每个段由选择匹配记录的规则进行定义。给定的规则可以有多个条件，例如：

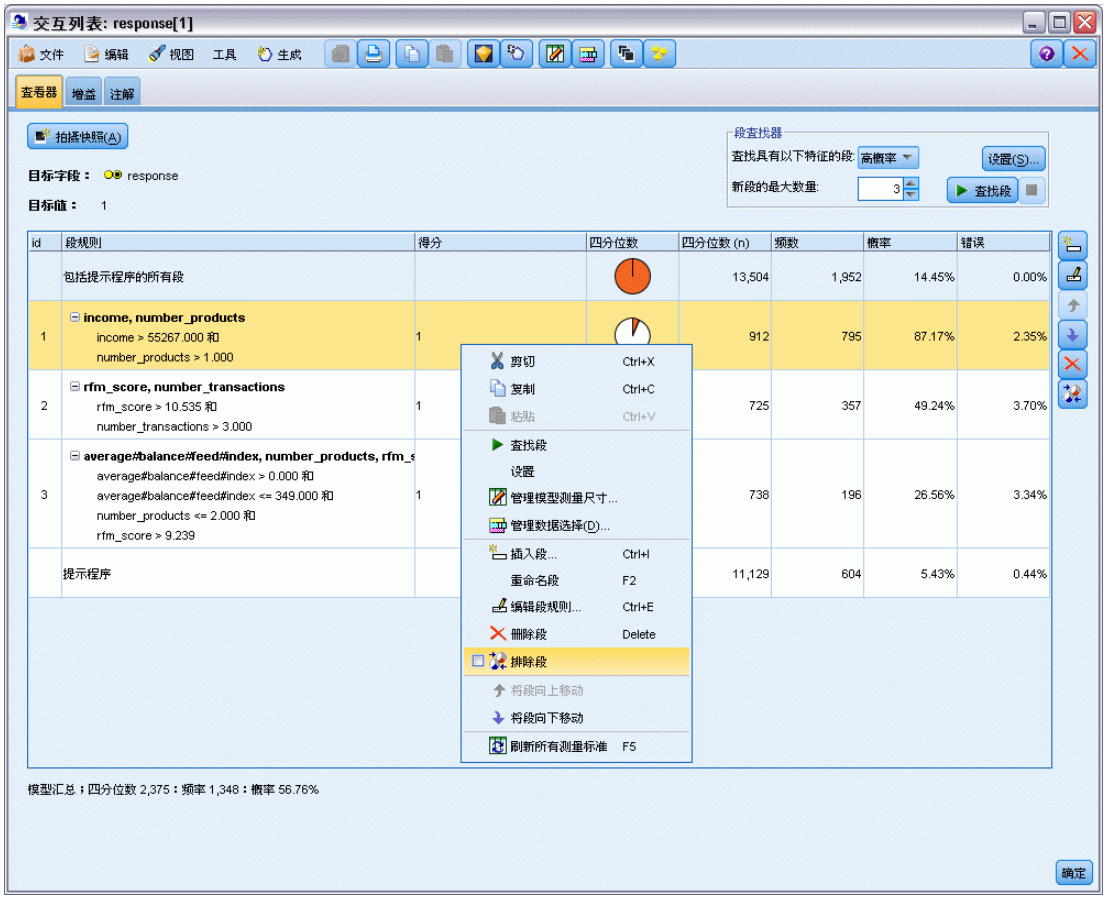
```
RFM_SCORE > 10 and
MONTHS_CURRENT <= 9
```

规则的列表顺序即为应用顺序，第一个匹配规则将决定给定记录的输出结果。如果单独采用，则规则或条件可能会发生重叠，但规则的顺序排除了二义性。如果规则不匹配，则记录将会分配给其余规则。

完全控制评分

Decision List Viewer 允许您查看、修改和重组段，并且可以评分为目的来选择包含或排除段。例如，您可以选择在未来报价中排除某组客户和包含其他客户，并且可以立即查看这对于整体匹配率的影响。Decision List 模型为被包含的段返回得分 Yes，为其它段（包括剩余段）返回 $null$。对评分的这种直接控制使得 Decision List 模型成为生成邮件发送清单的理想工具，而这些模型被广泛应用于客户关系管理中，包括呼叫中心或市场应用方面。

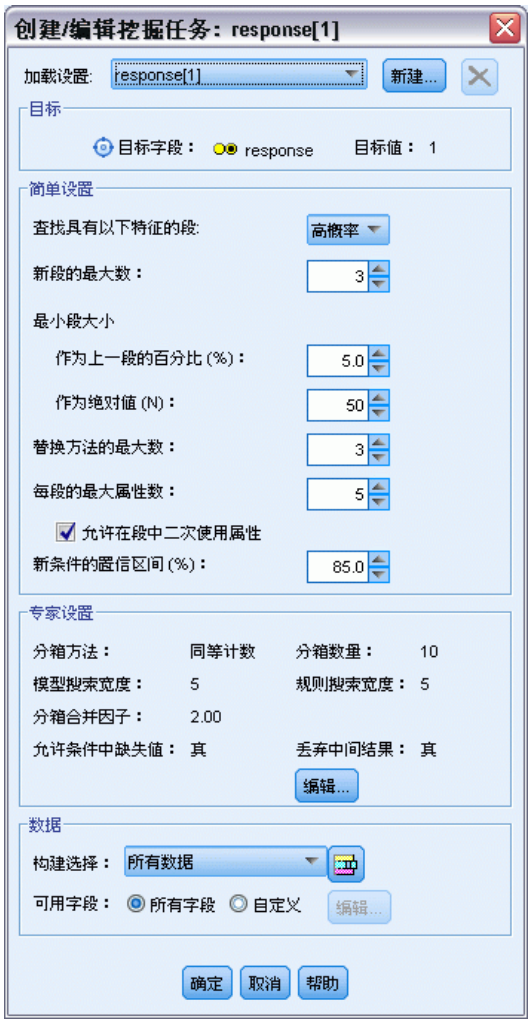
图片 9-2
决策列表模型



挖掘任务、测量和选择

建模过程由**挖掘任务**实现。每项挖掘任务可以有效地启动一次新的建模，并且会返回一组新的备选模型。默认任务基于 Decision List 节点的初始规范，您可以定义任意数量的自定义任务。还可以重复应用任务，例如，您可以在整个训练集合中运行高概率搜索，然后在剩余集合中运行低概率搜索来除去性能较低的段。

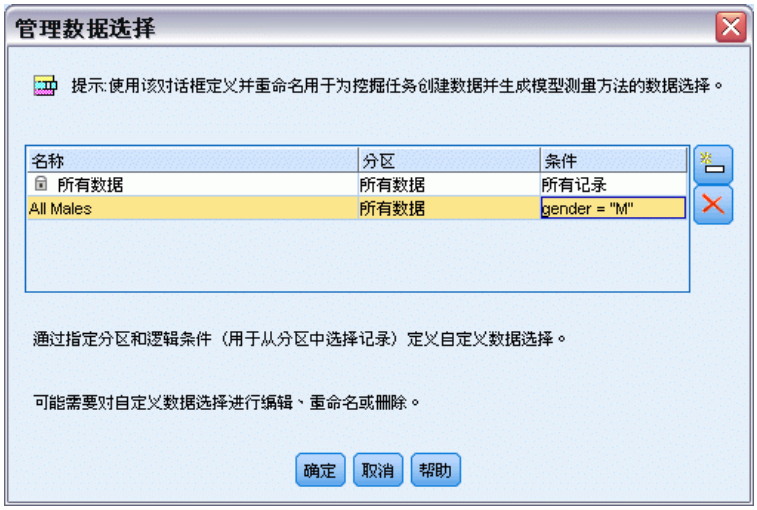
图片 9-3
创建挖掘任务



数据选择

您可以定义数据选择和自定义模型测量以进行模型构建和评估。例如，可以在挖掘任务中指定数据选择以裁剪模型，使之符合具体区域的要求，并且可以创建自定义测量以评估其就整个国家范围而言的性能优劣。不同于挖掘任务的是，测量并不改变底层模型而是以其它视角对其性能进行评估。

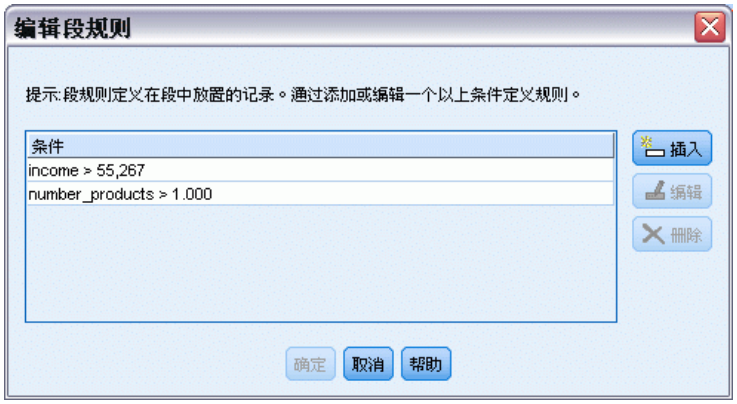
图片 9-4
创建数据选择



添加您的业务知识

通过微调或扩展算法识别的段，Decision List Viewer 允许您将业务知识并入模型。您可以编辑模型所生成的段或添加基于指定规则的其它段。然后可以应用更改并预览结果。

图片 9-5
指定规则



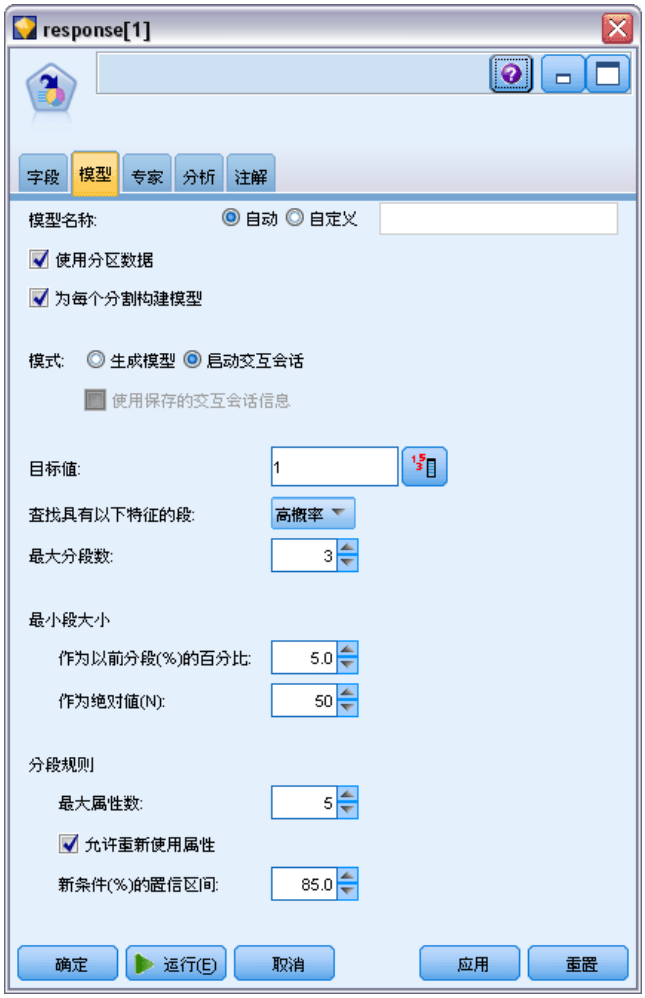
要了解更多详情，Excel 动态链接允许您将数据导出到 Excel，用这些数据可以在 Excel 中创建演示图表和计算定制测量（例如综合利润和 ROI），您可在构建模型的同时在 Decision List Viewer 中查看这些自定义测量。

示例。 某金融机构的市场部门希望通过向每个客户提供最适合他的报价在未来的商业竞争中获取更有益的结果。使用决策列表模型可以根据以前的销售活动，识别会对当前活动积极响应的客户应具备哪些特征，并根据识别的结果生成邮件发送清单。有关详细信息，请参阅第 12 章中的对客户响应建模（决策列表）中的IBM SPSS Modeler 14.2 应用程序 指南。

要求。 一个表示要预测的二元结果（是/否）且测量级别为标志或名义的分类目标字段和至少一个输入字段。当目标字段类型为名义时，必须手动选择一个值作为**匹配**或**响应**；所有其他值集中在一起作为**不匹配**。还可以指定一个可选的频数字段。连续日期/时间字段将被忽略。使用在建模节点的“专家”选项卡上指定的算法对连续数字范围的输入自动分级。为了更好地控制分级，可添加上游分级节点并使用已分级的字段作为测量级别为有序的输入。

决策列表模型选项

图片 9-6
决策列表节点：“模型”选项卡



模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

众数。指定用于构建模型的方法。

- **生成模型。** 运行节点时自动在模型选项板上生成模型。可将生成的模型添加到流中以便评分，但是此模型无法继续编辑。
- **启动交互会话。**打开 Decision List Viewer 交互建模（输出）窗口，您可选取多个选项并重复应用具有不同设置的算法以逐渐构建或修改模型。 有关详细信息，请参阅第 197 页码Decision List Viewer。
- **使用保存的交互会话信息。**使用以前保存的设置启动交互会话。可以使用 Decision List Viewer 中的“生成”菜单（用于创建模型或建模节点）或“文件”菜单（用于更新从中启动会话的节点）保存交互设置。

目标值。指定表示要建模的结果的目标字段的值。例如，如果目标字段“流失”编码为 0 = no 和 1 = yes，指定 1 可标识指明哪些记录可能流失的规则。

查找段。表示搜索目标变量是否应该查找出现的**高概率**或**低概率**。查找和排除这些段可能对于改善您的模型非常有帮助，当剩下的段为低概率段时尤其有用。

最大分段数。指定要返回的最大段数。创建顶部的 N 个段，其中最好的段是概率最高的段，如果多个模型具有相同的概率，则为覆盖率最高的段。允许的最小设置为 1；没有最大设置。

最小段大小。下面的两个设置指定最小段大小。两个值中的较大者优先。例如，如果百分比值等于比绝对值高的数字，则百分比设置优先。

- **以上一个段的百分比表示 (%)。**以记录的百分比指定最小组大小。允许的最小设置为 0；允许的最大设置为 99.9。
- **以绝对值表示 (N)。**以记录的绝对数指定最小组大小。允许的最小设置为 1；没有最大设置。

段规则。

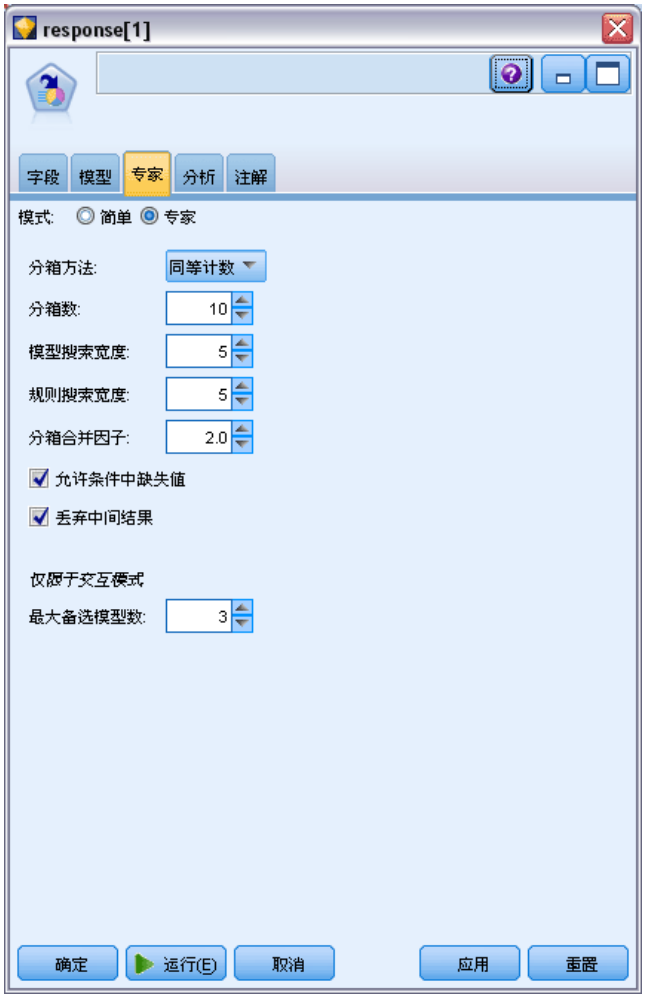
最大属性数。指定每个段规则的最大条件数。允许的最小设置为 1；没有最大设置。

- **允许重新使用属性。**如果启用，则每个周期可以使用所有属性，即使以前的周期已使用过这些属性。段的条件是在周期内构建的，每个周期都会增加一个新条件。周期数使用**最大属性数**设置定义。

新条件的置信区间 (%)。指定用于检验段显著性的置信水平。此设置在返回的段数（如果存在）以及每个段规则的条件数中具有非常重要的作用。值越高，返回的结果集越小。允许的最小设置为 50；允许的最大设置为 99.9。

决策列表节点专家选项

图片 9-7
决策列表节点：“专家”选项卡



使用“专家”选项可调整模型构建过程。

分级方法。用于对连续字段进行间隔的方式（相等计数或相等宽度）

图条数。要为连续字段创建的间隔数。允许的最小设置为 2；没有最大设置。

模型搜索宽度。每个周期可用于下一周期的最大模型结果数。允许的最小设置为 1；没有最大设置。

规则搜索宽度。每个周期可用于下一周期的最大规则结果数。允许的最小设置为 1；没有最大设置。

间隔合并因子。段与其相邻段合并时必须增加的最小量。允许的最小设置为 1.01；没有最大设置。

- **允许条件中使用缺失值。**True 表示允许规则中的 IS MISSING 检验。
- **丢弃中间结果。** 如果为 True，则只返回搜索过程的最终结果。最终结果是不在搜索过程中进行任何进一步细化的结果。如果为 False，则还要返回中间结果。

最大替代值数。指定运行挖掘任务后可以返回的最大替代值数。允许的最小设置为 1；没有最大设置。

注意，挖掘任务将只返回替代值的实际数量，最大为指定的最大数量。例如，如果最大数量设为 100，但只找到 3 个替代值，则只显示这 3 个替代值。

决策列表模型块

模型包括一个**段**列表，每个段都由**规则**进行定义，从而可以选择匹配的记录。在生成模型前可轻松查看或修改这些段，并选择包括哪些段或不包括哪些段。用于评分时，决策列表模型对于包含的段返回是，对于所有其他段（包括余数）返回 $null$。对评分的这种直接控制使得决策列表模型成为生成邮件发送清单的理想工具，而这些模型被广泛应用于客户关系管理中，包括呼叫中心或市场应用方面。

图片 9-8
决策列表模型块

运行包含决策列表模型的流时，节点将添加三个新字段，包括得分字段，其中对于包含的字段得分为 1（表示是），对于不包含的字段得分为 $null，用于其中含有记录的段的概率（匹配率）字段，及段的 ID 编号字段。新字段的名称来自正在预测的输出字段的名称，前缀为 $D- 表示得分字段，前缀为 $DP- 表示概率字段，前缀为 $DI- 表示段的 ID。

按照构建模型时指定的目标值对模型进行评分。可以手动去除某些段以便使它们的得分为 $null$。例如，如果运行低概率搜索以查找低于平均匹配率的段，则这些“低匹配率”段的得分将为是，除非您手动去除这些段。如果必要，可以使用导出节点或过滤节点将空值重新编码为否。

PMML

使用“第一个匹配”选择标准可将决策列表模型评分为 PMML RuleSetModel。但是，希望所有的规则具有相同的得分。为允许对目标字段或目标值进行更改，可将多个规则集模型存储到一个文件中按顺序进行应用，无法与第一个模型匹配的案例将传递到第二个模型，依此类推。算法名称 DecisionList 用于表示此非标准的行为，且仅具有该名称的规则集模型可被识别为决策列表模型并如上所述进行评分。

决策列表模型块设置

通过决策列表模型块的“设置”选项卡，您可以获取倾向得分，还可以启用或禁用 SQL 优化。只有将模型块添加到流之后，才可以使用此选项卡。

图片 9-9
决策列表模型块设置

计算原始的倾向得分。对于含标志目标（返回?是?或?否?预测）的模型，您可以请求倾向得分，这些得分指示为目标字段指定结果为真的可能性。除了这些得分，还有其他在评分过程中生成的预测值和置信度值。

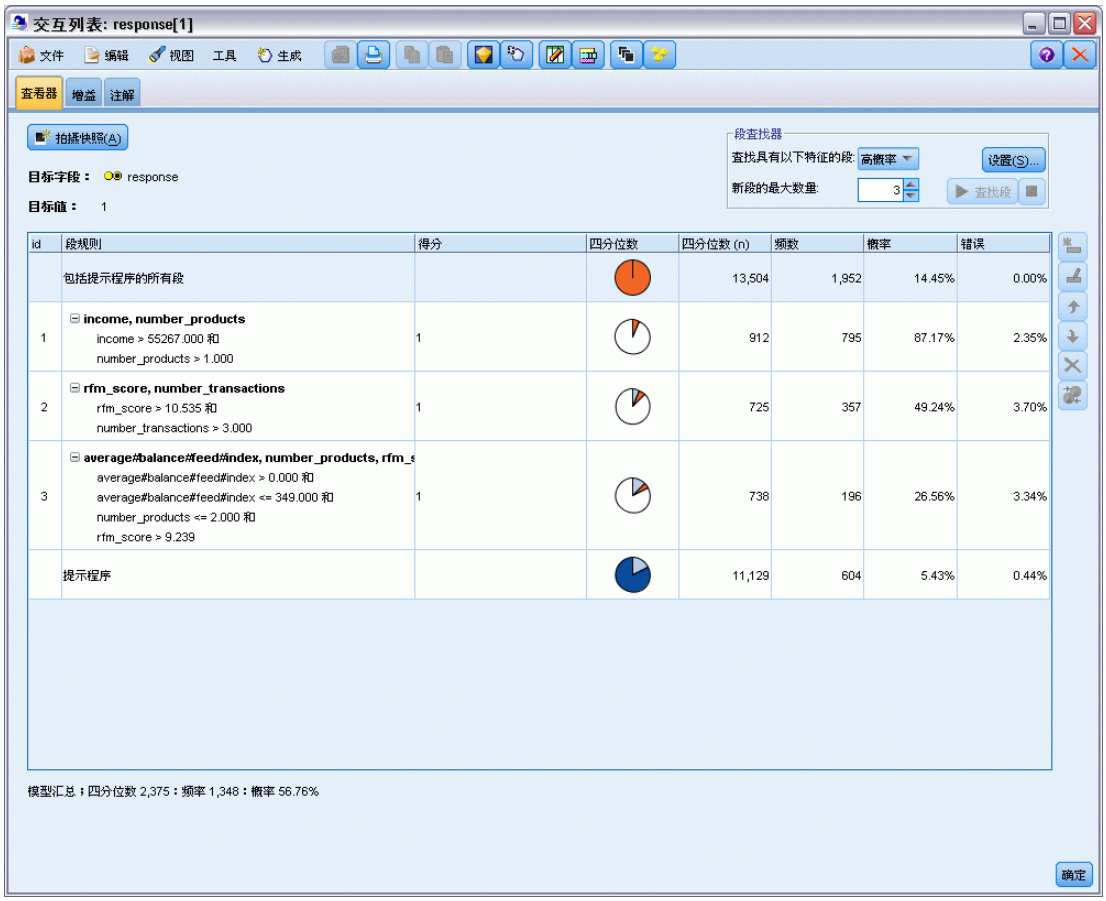
计算调整后的倾向得分。原始倾向得分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该得分可能会过度优化。调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。此选项要求在流中定义分区字段并且在生成模型之前在建模节点中启用调整的倾向得分。

生成此模型的 SQL。使用数据库中的数据时，SQL 代码可传回到数据库中执行，从而大大提高许多操作的处理速度。有关详细信息，请参阅第 6 章中的SQL 优化中的IBM SPSS Modeler Server 14.2 管理和性能指南。

Decision List Viewer

基于任务的 Decision List Viewer 图形界面简单易用，可消除模型构建过程的复杂性，使您免于接触数据挖掘技术的低层详细信息，并且可以将全部精力投入到需要用户参与的分析内容上，如设置目标、选择目标组、分析结果，以及选择最优模型。

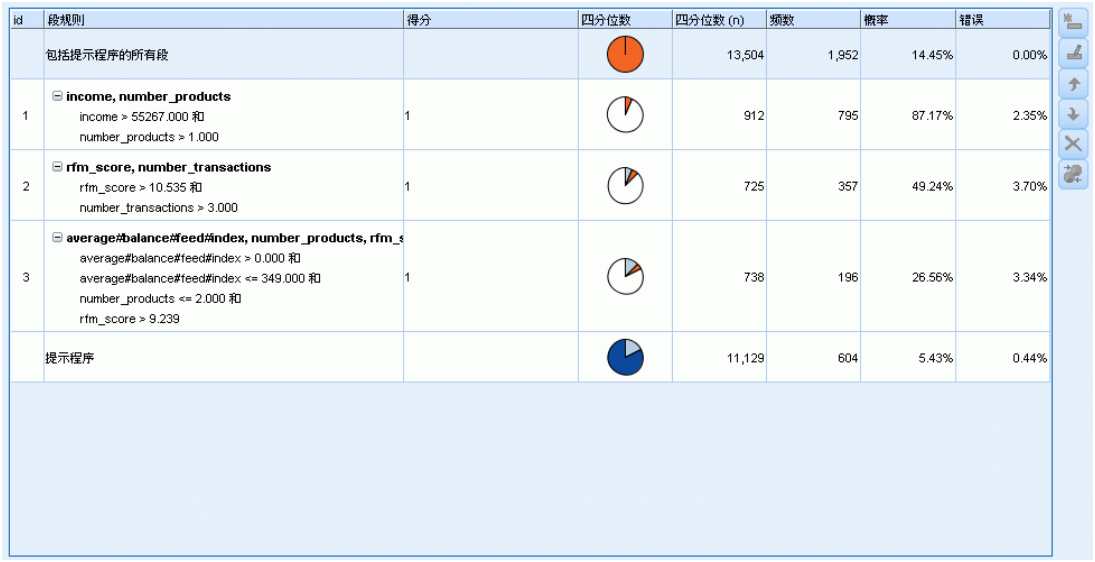
图片 9-10
决策列表交互查看器



工作模型窗格

工作模型窗格将显示当前模型，包括挖掘任务和适用于该工作模型的其他操作。

图片 9-11
工作模型窗格



id	段规则	得分	四分位数	四分位数 (n)	频数	概率	错误
	包括提示程序的所有段			13,504	1,952	14.45%	0.00%
1	**income, number_products** income > 55267.000 和 number_products > 1.000	1		912	795	87.17%	2.35%
2	**rfm_score, number_transactions** rfm_score > 10.535 和 number_transactions > 3.000	1		725	357	49.24%	3.70%
3	**average#balance#feed#index, number_products, rfm_s** average#balance#feed#index > 0.000 和 average#balance#feed#index <= 349.000 和 number_products <= 2.000 和 rfm_score > 9.239	1		738	196	26.56%	3.34%
	提示程序			11,129	604	5.43%	0.44%

ID。标识连续段顺序。模型段根据其 ID 号按顺序进行计算。

段规则。提供段名称和已定义的段条件。默认情况下，段名称是字段名或条件中使用的连接字段名（以逗号为分隔符）。

得分。表示要预测的字段，假定其值与其他字段的值（预测变量）有关。

注意：以下选项可切换为通过组织模型测量对话框显示。

涉及范围。该饼图直观地标识出每个段的涉及范围与整个涉及范围的对比情况。

涉及范围 (n)。列出每个段相对于整个涉及范围的涉及范围量。

频率。列出收到的相对于涉及范围的匹配项的数量。例如，如果涉及范围为 79，频数为 50，则表示在 79 个之中有 50 个对所选段进行了响应。

概率。指明段的概率。例如，如果涉及范围为 79，频数为 50，则表示该段的概率为 63.29%（50 除以 79）。

错误。指明段的错误。

窗格底部的信息显示整个模型的涉及范围、频数和概率。

工作模型工具栏

工作模型窗格的工具栏提供了以下功能。

注意：右键单击模型段也可访问其中某些功能。

表 9-1
工作模型工具栏按钮

	启动生成新的型对话框，该对话框提供用于创建新模型块的选项。
	保存交互会话的当前状态。这会将“决策列表”建模节点更新为当前设置，包括挖掘任务、模型快照、数据选择和自定义测量量。要将会话恢复至此状态，选中建模节点的“模型”选项卡中的**使用保存的会话信息**对话框，然后单击**运行**。
	显示“组织模型测量”对话框。 有关详细信息，请参阅第 213 页码组织模型测量。
	显示“组织数据选择”对话框。 有关详细信息，请参阅第 207 页码组织数据选择。
	显示“快照”选项卡。 有关详细信息，请参阅第 201 页码“快照”选项卡。
	显示“替代”选项卡。 有关详细信息，请参阅第 199 页码“替代”选项卡。
	获取当前模型结构的快照。快照显示在“快照”选项卡中，通常用于模型比较。
	启动插入段对话框，该对话框提供用于创建新模型段的选项。
	启动编辑段规则对话框，该对话框提供的选项可用于将条件添加到模型段，或更改先前定义的模型段条件。
	在模型层次中将所选段上移。
	在模型层次中将所选段下移。
	删除所选段。
	在模型中包括/排除所选段的情况之间进行切换。排除时，段结果将计入余数。不同于删除段的是，排除段允许您选择重新激活段。

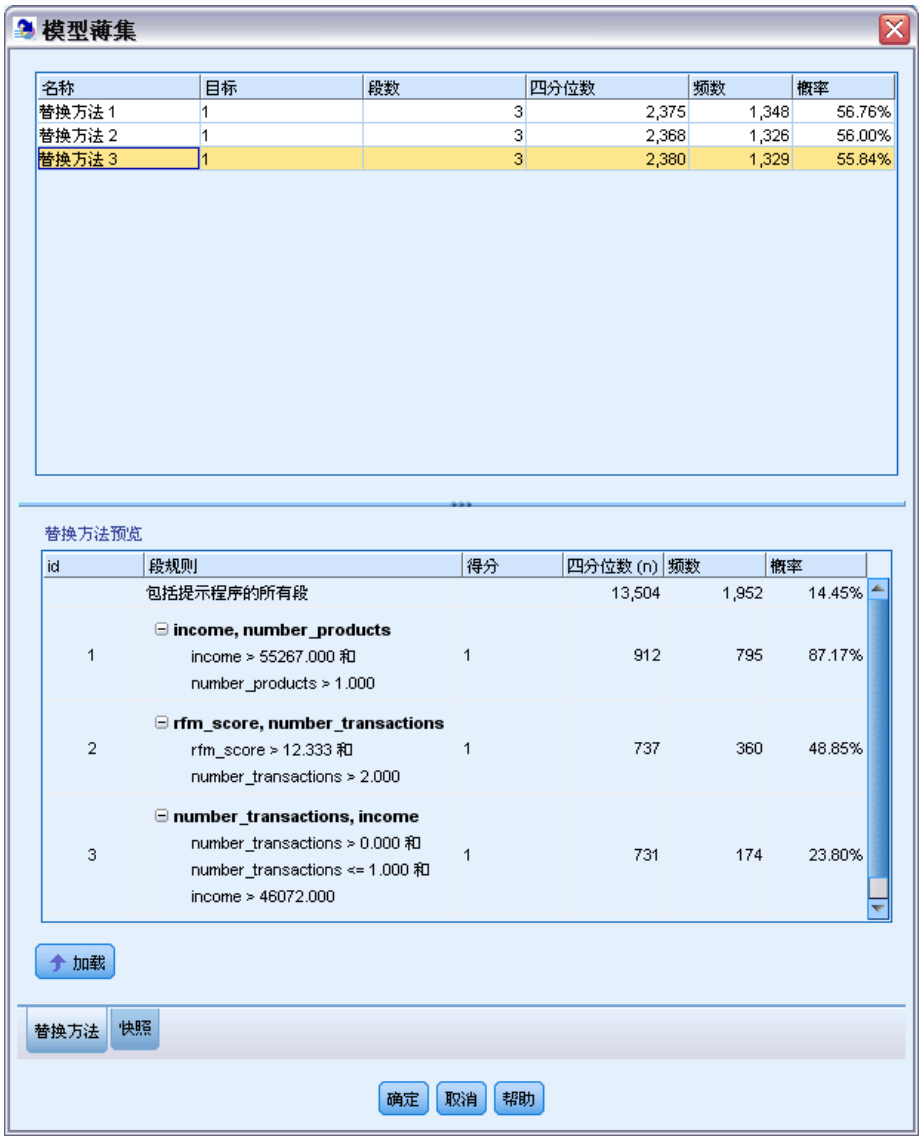
“替代”选项卡

单击**查找段**生成“替代”选项卡，该选项卡将针对工作模型窗格中的选定模型或段列出所有替代挖掘结果。

- 要将替代模型提升为工作模型，突出显示所需替代模型并单击**加载**；则替代模型显示在工作模型窗格中。

注意：只有当您已在决策列表建模节点“专家”选项卡上设置了**最大替代值数**时，才会显示“替代”选项卡以创建多个替代值。

图片 9-12
“替代”选项卡



每个已生成的模型替代项会显示特定的模型信息：

名称。每个替代模型都有顺序编号。第一个替代项通常包含最佳结果。

目标。指明目标值。例如：1，等于“真”。

段数。替代模型中所使用的段规则数。

涉及范围。替代模型的涉及范围。

频率。相对于涉及范围的匹配项的数量。

概率。指明替代模型的概率百分比。

注意：替代结果不会随模型保存；结果只在活动会话中有效。

“快照”选项卡

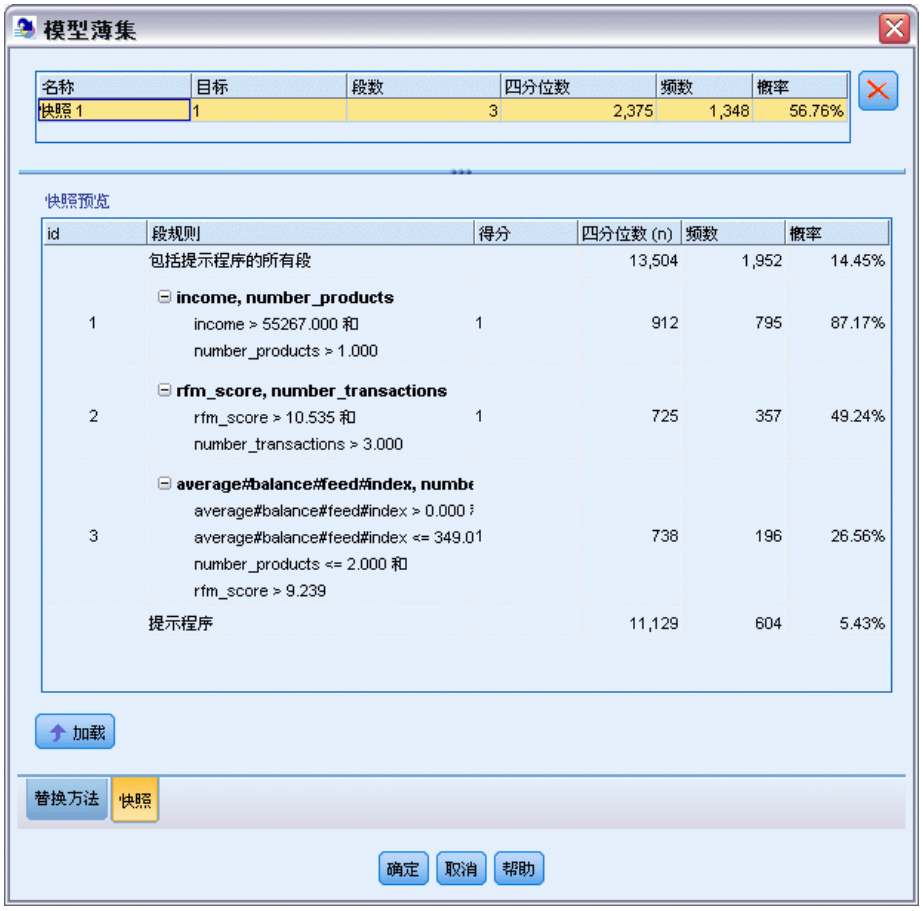
快照是模型在特定时间点的视图。例如，如果您需要将另一个替代模型载入工作模型窗格、但不希望失去当前模型的相关工作，则可以获取模型快照。“快照”选项卡将列出在任意数量的工作模型状态下手动获取的所有模型快照。

注意：快照将随模型保存。我们建议在您加载首个模型时执行快照。该快照用于保存原始模型结构，从而确保您可随时返回原始模型状态。生成的快照名称显示为时间戳，指示其生成时间。

创建模型快照

- 选择要在工作模型窗格中显示的适当的模型/替代项。
- 对该工作模型进行必要的更改。
- 单击**执行快照**。此时将在“快照”选项卡中显示一个新快照。

图片 9-13
快照选项卡



名称。快照名称。您可以双击快照名称对其进行更改。

目标。指明目标值。例如：1，等于“真”。

段数。模型中所使用的段规则数。

涉及范围。模型的涉及范围。

频率。相对于涉及范围的匹配项的数量。

概率。指明模型的概率百分比。

- 要将快照提升为工作模型，突出显示所需快照并单击**加载**；则快照模型显示在工作模型窗格中。
- 可通过以下方法删除快照：单击**删除**，或右键单击快照，然后在菜单中选择**删除**。

使用 Decision List Viewer

将以最佳方式预测客户响应和行为的模型是通过多个阶段进行构建的。启动 Decision List Viewer 时，工作模型将填入已定义的模型段和测量量，并且准备就绪，等待您启动挖掘任务、根据需要修改段/测量量，并生成新的模型或建模节点。

您可添加一个或多个段规则，直到获得满意的模型。可以通过运行挖掘任务或使用**编辑段规则**功能为模型添加段规则。

在模型构建过程中，您可以对模型的性能进行评估，方法是根据测量数据验证模型、在图表中对图形进行可视化处理，或生成自定义 Excel 测量量。

肯定模型的质量后，您可以生成新模型并将其置于 IBM® SPSS® Modeler 工作区或模型选项板中。

挖掘任务

挖掘任务是确定新规则生成方式的参数的集合。其中某些参数是可以选择的，以便为您提供使模型适应新状况的灵活性。任务由任务模板（类型）、目标和构建选择（挖掘数据集）组成。

下列各部分详细介绍各种挖掘任务操作：

- 运行挖掘任务
- 创建和编辑挖掘任务
- 组织数据选择

运行挖掘任务

Decision List Viewer允许您为模型手动添加段规则，方法是运行挖掘任务或在模型之间复制和粘贴段规则。挖掘任务包含有关如何生成新段规则的信息（数据挖掘参数设置，如搜索策略、源属性、搜索宽度、置信水平等）、待预测的客户行为，以及要调查的数据。挖掘任务的目标是搜索可能的最佳段规则。

要通过运行挖掘任务生成模型段规则，请执行下列操作：

▶ 单击**余数**行。如果工作模型窗格中已有显示的段，您也可以选择其中某一个，根据所选段查找其他规则。选择余数或段之后，可采用下列方法之一生成模型或替代模型：

- 从“工具”菜单选择**查找段**。
- 右键单击**余数**行/段，然后选择**查找段**。
- 单击工作模型窗格上的**查找段**按钮。

在任务处理过程中，进度将在工作区底部显示，并在任务完成时提示您。任务完成所用的时间完全取决于挖掘任务的复杂性以及数据集的大小。如果结果中只有一个模型，则任务完成后它将立即显示在工作模型窗格上；但是，如果结果包含多个模型，则模型显示在“替代”选项卡上。

注意：任务结果将为：完成并更新模型或完成但不更新模型抑或失败。

可以重复查找新段规则的过程，直到不再有新规则添加到模型中。这表示已找到所有有意义的客户组。

可以对任何现有的模型段运行挖掘任务。如果对任务的结果不满意，您可以选择对同一模型段启动另一个挖掘任务。此操作将基于所选段提供找到的其他规则。位于所选段“下方”的段（即，在所选段之后添加到模型的段）将被新段替代，因为每个段都取决于其前项。

创建和编辑挖掘任务

挖掘任务是搜索组成数据模型的规则集合的机制。除所选模板中定义的搜索条件外，任务还会定义目标（激发分析的实际问题，如有多少客户可能对邮件做出响应），并标识要使用的数据集。挖掘任务的目标是搜索可能的最佳模型。

创建挖掘任务

要创建挖掘任务，请执行下列操作：

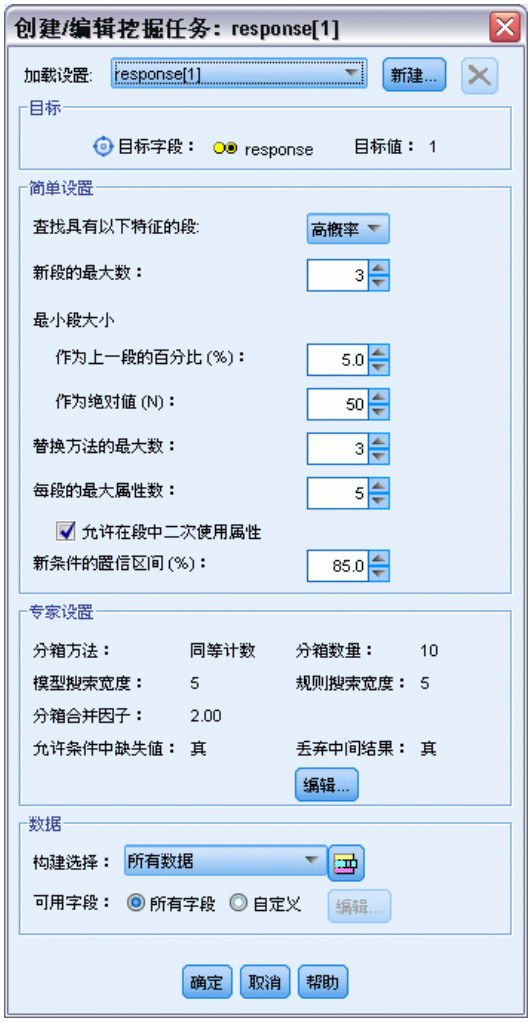
- 选择要在其中挖掘其他段条件的段。
- 单击**设置**。此时将打开“创建/编辑挖掘任务”对话框。该对话框提供用于定义挖掘任务的选项。
- 进行必要的更改并单击**确定**返回到工作模型窗格。Decision List Viewer 使用默认设置运行每个任务，直到选择了替代任务或设置。
- 单击**查找段**以启动选定段上的挖掘任务。

编辑挖掘任务

“创建/编辑挖掘任务”对话框提供的选项可用于定义新的挖掘任务或编辑现有挖掘任务。

可用于挖掘任务的大部分参数与决策列表节点中提供的参数类似。例外显示如下。有关详细信息，请参阅第 192 页码决策列表模型选项。

图片 9-14
“创建/编辑挖掘任务”对话框



加载设置：当您创建了多个挖掘任务时，选择所需任务。

新建...单击以基于当前显示任务的设置新建挖掘任务。

Target

目标字段：表示要预测的字段，假定其值与其他字段的值（预测变量）有关。

目标值。指定表示要建模的结果的目标字段的值。例如，如果目标字段“流失”编码为 0 = no 和 1 = yes，指定 1 可标识指明哪些记录可能流失的规则。

简单设置

最大替代值数。指定运行挖掘任务后将显示的替代值数。允许的最小设置为 1；没有最大设置。

专家设置

编辑... 打开编辑高级参数对话框，您可在其中定义高级设置。 有关详细信息，请参阅第 206 页码编辑高级参数。

Data

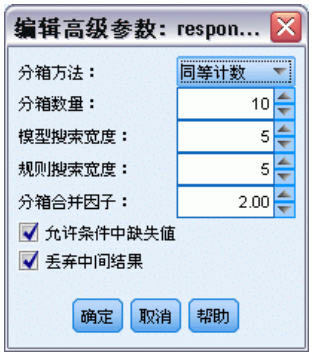
构建选择。提供的选项用于指定Decision List Viewer应对其进行分析以查找新规则的评估尺度。列出的评估尺度在“组织数据选择”对话框中进行创建/编辑。

可用字段。提供显示所有字段或手动选择要显示的字段的选项。

编辑...如果已选择自定义选项，会打开自定义可用字段对话框，您可在其中选择通过挖掘任务找到的可用作段属性的字段。 有关详细信息，请参阅第 207 页码自定义可用字段。

编辑高级参数

图片 9-15
高级参数



“编辑高级参数”对话框提供以下配置选项。

分级方法。用于对连续字段进行间隔的方式（相等计数或相等宽度）

图条数。要为连续字段创建的间隔数。允许的最小设置为 2；没有最大设置。

模型搜索宽度。每个周期可用于下一周期的最大模型结果数。允许的最小设置为 1；没有最大设置。

规则搜索宽度。每个周期可用于下一周期的最大规则结果数。允许的最小设置为 1；没有最大设置。

间隔合并因子。段与其相邻段合并时必须增加的最小量。允许的最小设置为 1.01；没有最大设置。

- **允许条件中使用缺失值。**True 表示允许规则中的 IS MISSING 检验。
- **丢弃中间结果。** 如果为 True，则只返回搜索过程的最终结果。最终结果是不在搜索过程中进行任何进一步细化的结果。如果为 False，则还要返回中间结果。

自定义可用字段

图片 9-16
“自定义可用字段”对话框

使用“自定义可用字段”对话框，可以选择通过挖掘任务找到的可用作段属性的字段。

可用。列出当前可用作段属性的字段。要从列表中删除字段，请选择适当的字段，然后单击**删除 >>**。此时所选字段将从“可用”列表移至“不可用”列表。

不可用。列出不可用作段属性的字段。要将字段包括在“可用”列表中，请选择适当的字段，然后单击**<< 添加**。此时所选字段将从“不可用”列表移至“可用”列表。

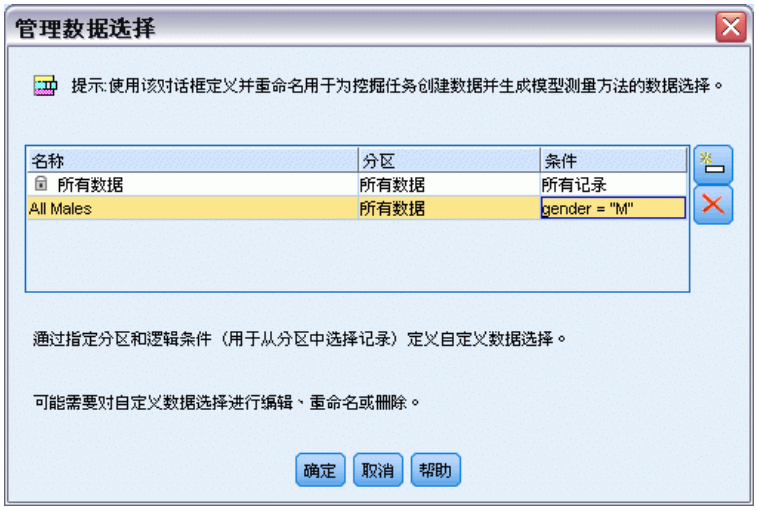
组织数据选择

通过组织数据选择（挖掘数据集），可以指定Decision List Viewer应对哪些评估尺度进行分析以查找新规则，并选择要用作尺度基准的数据选择。

要组织数据选择，请执行下列操作：

- 从“工具”菜单中选择**组织数据选择**，或右键单击某个段并选择该选项。此时将打开“组织数据选择”对话框。

图片 9-17
“组织数据选择”对话框



注意：“组织数据选择”对话框也可用于编辑或删除现有的数据选择。

- 单击**添加新的数据选择**按钮。此时会将一个新的数据选择条目添加到现有的表中。
- 单击**名称**并输入适当的选择名称。
- 单击**分区**并选择适当的分区类型。
- 单击**条件**并选择适当的条件选项。如果选择**指定**，则会打开“指定选择条件”对话框，其中包含定义特定字段条件的选项。

图片 9-18
“指定选择条件”对话框

- 定义适当的条件，然后单击**确定**。

通过“创建/编辑挖掘任务”对话框中的“构建选择”下拉列表可访问这些数据选择。使用该列表可选择用于特定挖掘任务的评估尺度。

段规则

通过运行基于任务模板的挖掘任务，可以查找模型段规则。您可以使用“插入段”或“编辑段规则”功能手动为模型添加段规则。

如果选择挖掘新的段规则，结果（如果有）将在“交互列表”对话框的“查看器”选项卡中显示。通过从“模型作品集”对话框中选择替代结果，并单击**加载**，可以快速精练您的模型。这样，您可以尝试不同结果，直到准备好构建出准确描述最佳目标组的模型。

插入段

您可以使用“插入段”功能手动为模型添加段规则。

要将段规则条件添加到模型，请执行下列操作：

- 在交互列表对话框中，选择您要添加新段的位置。新段将直接插在所选段的上方。
- 在“编辑”菜单中，选择**插入段**或通过右键单击段访问此选项。

  此时将打开“插入段”对话框，通过该对话框可插入新的段规则条件。
- 单击**插入**。此时将打开“插入条件”对话框，通过该对话框可以定义新规则条件的属性。
- 从下拉列表中选择字段和运算符。

  注意：如果选择**不在**运算符，则所选条件将作为排除条件，并在“插入规则”对话框中显示为红色。例如，当条件 region = 'TOWN' 显示为红色时，表示结果集中不包括 TOWN。
- 输入一个或多个值，或者单击**插入值**图标，以显示“插入值”对话框。对话框允许您选择为选定字段定义的值。例如，字段**已婚**将提供值**是**和**否**。
- 单击**确定**返回“插入段”对话框。再次单击**确定**将所创建的段添加到模型中。

  此时该新段将显示在指定的模型位置。

编辑段规则

使用“编辑段规则”功能，可以添加、更改或删除段规则条件。

要更改段规则条件，请执行下列操作：

- 选择要编辑的模型段。

- 从“编辑”菜单选择**编辑段规则**，或右键单击规则以访问此选项。

  此时将打开“编辑段规则”对话框。

- 选择适当的条件，然后单击**编辑**。

  此时将打开“编辑条件”对话框，通过该对话框可以定义所选规则条件的属性。

- 从下拉列表中选择字段和运算符。

  注意：如果选择**不在**运算符，则所选条件将作为排除条件，并在“编辑段规则”对话框中显示为红色。例如，当条件 region = 'TOWN' 显示为红色时，表示结果集中不包括 TOWN。

- 输入一个或多个值，或单击**插入值**按钮以显示“插入值”对话框。对话框允许您选择为选定字段定义的值。例如，字段**已婚**将提供值**是**和**否**。

- 单击**确定**返回到“编辑段规则”对话框。再次单击**确定**返回工作模型。

  此时所选择的段将与更新的规则条件一起显示。

删除段规则条件

要删除段规则条件，请执行下列操作：

- 选择包含要删除的规则条件的模型段。
- 从“编辑”单中选择**编辑段规则**，或右键单击段以访问此选项。

  此时将打开“编辑段规则”对话框，您可在其中删除一个或多个段规则条件。

- 选择适当的规则条件，然后单击**删除**。
- 单击**确定**。

  删除一个或多个段规则条件将使工作模型窗格刷新其测量度量。

复制段

Decision List Viewer为您提供了一种复制模型段的简便方法。如果要将一个模型中的段应用于另一个模型时，只需将该段从一个模型复制（或剪切）并粘贴到另一个模型中即可。此外，您还可以从“替代预览”窗格中显示的模型复制段并将其粘贴到工作模型窗格中显示的模型中。这些剪切、复制和粘贴功能使用系统剪贴板存储或检索临时数据。这意味着将在剪贴板中复制条件和目标。剪贴板内容不仅仅保留用于Decision List Viewer，也可以粘贴在其他应用程序中。例如，在文本编辑器中粘贴剪贴板内容时，会以 XML 格式粘贴条件和目标。

要复制或剪切模型段，请执行下列操作：

- 选择要在其他模型中使用的模型段。
- 从“编辑”菜单中选择**复制**（或**剪切**），或右键单击模型段并选择**复制**或**剪切**。
- 打开适当的模型（将在其中粘贴模型段的模型）。

- 选择某个模型段，然后单击**粘贴**。

  注意：除了**剪切**、**复制**和**粘贴**命令外，还可以使用组合键：Ctrl+X、Ctrl+C 和 Ctrl+V。

  复制（剪切）的段将插入先前选择的模型段上方。粘贴的段和下方段的测量量将重新计算。

  注意：此过程中的两个模型必须以同一基础模型模板为基准，并包含相同的目标，否则将显示错误消息。

替代模型

当有多个结果时，“替代”选项卡显示每个挖掘任务的结果。每个结果包含所选数据中与目标最接近匹配的条件，以及所有“相当匹配”的替代项。显示的替代项总数取决于分析过程中采用的搜索条件。

要查看替代模型，请执行下列操作：

- 单击“替代”选项卡上的替代模型。在“替代预览”窗格中，替代模型段显示或替代当前模型段。
- 要在工作模型窗格中使用替代模型，在“替代预览”窗格中选择模型并单击**加载**，或在“替代”选项卡上右键单击替代模型名称并选择**加载**。

  注意：生成新模型时，不会保存替代模型。

自定义模型

数据不是静态的。客户会迁移、结婚和更换工作。产品会随之失去市场焦点并作废。Decision List Viewer为商业用户提供了方便迅速地使模型适应新状况的灵活性。您可通过编辑、设置优先级、删除或停用特定模型段来更改模型。

为段设置优先级

您可选择任意顺序，对模型规则进行排列。默认情况下，模型段按优先级顺序显示，第一个段具有最高优先级。当您为一个或多个段指定不同的优先级时，模型会发生相应的更改。您可以根据需要通过将段移至较高或较低的优先级位置来更改模型。

要为模型段设置优先级，请执行下列操作：

- 选择要为其指定不同优先级的模型段。
- 单击工作模型窗格工具栏中的两个箭头按钮之一，将所选模型段在列表中上移或下移。

  设置优先级后，会重新计算先前的所有评估结果，并显示新值。

删除段

要删除一个或多个段，请执行下列操作：

- 选择模型段。
- 从“编辑”菜单中选择**删除段**，或在工作模型窗格的工具栏中单击删除按钮。

  测量量将针对修改后的模型重新计算，模型也会发生相应的更改。

排除段

在搜索特定组时，您可能会将一部分模型段作为商业操作的基准。部署模型时，您可能会选择排除模型中的某些段。排除的段作为空值进行评分。排除某个段并不代表不使用该段，而是从邮件列表中排除与该规则匹配的所有记录。该规则仍在应用，但方式不同。

要排除特定的模型段，请执行下列操作：

- 在工作模型窗格中选择一个段。
- 在工作模型窗格的工具栏中单击**切换段排除**按钮。此时将在所选段的所选“目标”列中显示**已排除**。

  注意：与删除的段不同，已排除的段在最终模型中仍可供重复使用。已排除的段仍将影响图表结果。

更改目标值

使用“更改目标值”对话框，可以更改当前目标字段的目标值。

与工作模型具有不同目标值的快照和会话结果会通过将该行的表背景变为黄色进行标识。这表示该快照/会话结果已过时。

创建/编辑挖掘任务对话框将显示当前工作模型的目标值。该目标值不会随挖掘任务保存，而是取自工作模型的值。

当您将某个与当前工作模型具有不同目标值的已保存模型提升为工作模型（例如，通过编辑替代结果或编辑快照副本）时，已保存模型的目标值将更改为工作模型的目标值（工作模型窗格中显示的目标值不会更改）。模型度量将根据新目标重新计算。

生成新的型

“生成新模型”对话框提供的选项可用于命名模型并选择创建新节点的位置。

模型名称。选择**自定义**可调整自动生成的名称，或为流工作区中显示的节点创建唯一名称。

创建节点位置。选择**工作区**会将新模型置于工作区中；选择 **GM 选项板**会将新模型置于“模型”选项板中；选择**两者**会将新模型同时置于工作区和“模型”选项板中。

包括交互会话状态。如果启用此选项，则会在生成的模型中保留交互会话状态。稍后根据模型生成建模节点时，该状态将继续传递并用于初始化交互会话。无论是否选择此选项，模型本身对新数据的评分方式都是相同的。如果未选择此选项，模型仍然可以创建构建节点，但该节点将更为一般化，它会启动新的交互会话而不是从原有会话停

止的位置继续前进。如果更改节点设置但以保存的某种状态执行，则会忽略已更改的设置以采用保存状态的设置。

注意：标准度量是唯一随模型保留的度量。其他度量将保留在交互状态。生成的模型不会显示已保存的交互挖掘任务状态。启动Decision List Viewer时，它会显示通过查看器所做的初始设置。

有关详细信息，请参阅第 58 页码第 3 章中的重新生成建模节点。

模型评估

成功的建模需要在生产环境中执行实施之前进行谨慎的模型评估。Decision List Viewer提供了可用于评估模型实际应用效果的多种统计测量量和商业测量量。其中包括收益图表和与 Excel 的全面互操作，从而实现成本/收益方案的模拟，以便评估部署的作用。

您可采用以下方式评估自己的模型：

- 使用 Decision List Viewer 中提供的预定义的统计测量量和商业模型测量量（概率、频数）。
- 评估从 Microsoft Excel 中导入的测量量。
- 使用收益图表对模型进行可视化处理。

组织模型测量

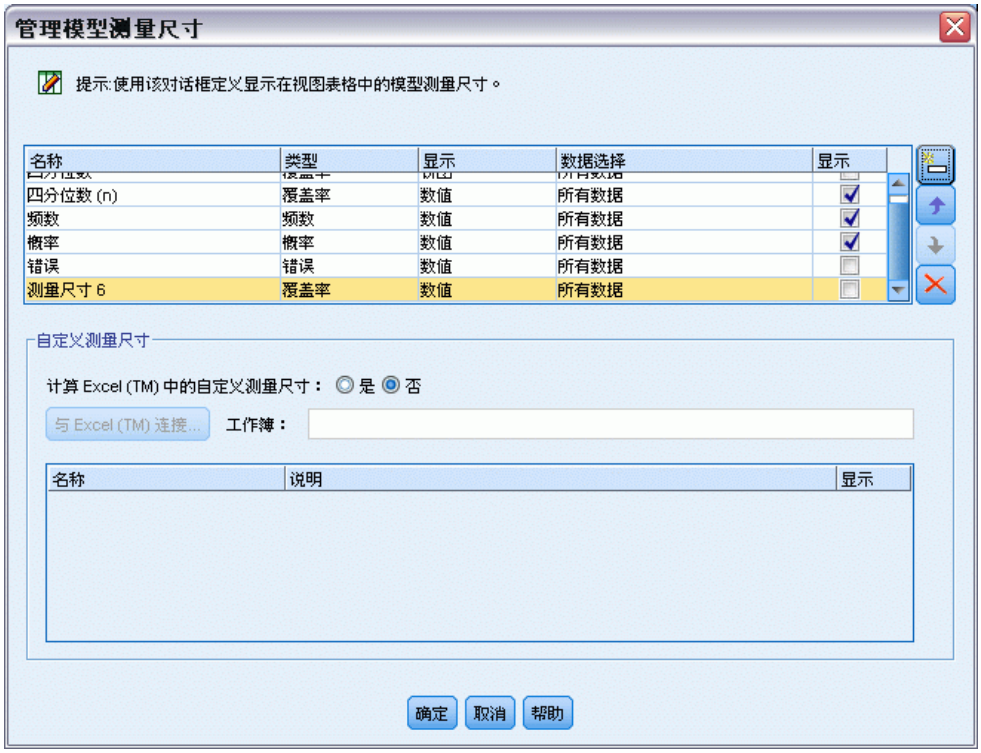
Decision List Viewer提供了用于定义按列计算并显示的测量量的选项。每个段可包括默认的涉及范围、频数、概率和错误等测量量，按列显示。此外，您也可以创建将按列显示的新测量量。

定义模型测量

要为模型添加测量量或定义现有的测量量，请执行下列操作：

▶ 从“工具”菜单中选择**组织模型测量**，或右键单击模型以选择此选项。此时将打开“组织模型测量”对话框。

图片 9-19
“组织模型测量”对话框



- 单击**添加新的模型测量**按钮（位于“显示”列右侧）。此时将在表中显示一个新的测量量。
- 提供测量量名称，并选择适当的类型、显示选项和选择。“显示”列指示是否为工作模型显示测量。定义现有测量量时，请选择适当的度量和选择，并指定该度量是否将在工作模型中显示。
- 单击**确定**返回Decision List Viewer工作区。如果已选中新测量的“显示”列，则会为工作模型显示该新测量。

Excel 中的自定义度量

有关详细信息，请参阅第 215 页码Excel 中的评估。

刷新测量

在某些特定情况下，可能需要重新计算模型测量，例如对一组新客户应用现有模型时。

要重新计算（刷新）模型测量，请执行下列操作：

- 在“编辑”菜单中选择**刷新所有测量量**。

或

- 按 F5。

此时将重新计算所有测量量，并针对工作模型显示新值。

Excel 中的评估

Decision List Viewer可与 Microsoft Excel 集成，使您可以在模型构建过程中直接使用自己的值计算和利润公式，以模拟成本/收益方案。与 Excel 的链接使您可以将数据导出至 Excel（数据在其中可用于创建演示图表）、计算自定义测量量（如复杂利润和 ROI 测量量），并且可以在构建模型时通过Decision List Viewer查看这些测量量。

有关详细信息，请参阅第 12 章中的使用 Excel 计算自定义测量量中的IBM SPSS Modeler 14.2 应用程序 指南。

注意：要使用 Excel 电子表格，必须由 CRM 分析专家针对Decision List Viewer 与 Microsoft Excel 的同步定义配置信息。该配置包含于 Excel 电子表格文件中，用于指明Decision List Viewer 与 Excel 之间相互传输的信息。

以下步骤仅在已安装 MS Excel 的情况下有效。如果未安装 Excel，则不会显示使模型与 Excel 同步的选项。

要使模型与 MS Excel 同步，请执行下列操作：

- 打开模型，运行交互会话，并从“工具”菜单中选择**组织模型测量**。
- 为**计算 Excel 中的自定义测量量**选项选择**是**。此时将激活**工作簿**字段，您可在其中选择预先配置的 Excel 工作簿模板。
- 单击**连接到 Excel** 按钮。此时将打开“打开”对话框，您可在其中导航至预先配置的模板在本地或网络文件系统中的位置。
- 选择适当的 Excel 模板，然后单击**打开**。此时将启动所选的 Excel 模板；使用 Windows 任务栏（或按 Alt+Tab）返回到“选择自定义测量的输入”对话框。
- 在 Excel 模板中定义的度量名称与模型度量名称之间选择适当的映射，然后单击**确定**。

建立链接后，Excel 将立即采用预先配置的 Excel 模板启动，该模板以电子表格显示模型规则。Excel 中的计算结果在Decision List Viewer中显示为新列。

注意：保存模型时，不会保留 Excel 度量；度量只在活动会话中有效。但是，您可以创建包括 Excel 度量的快照。在快照视图中保存的 Excel 度量仅适用于历史比较，在重新打开时不会刷新。有关详细信息，请参阅第 201 页码“快照”选项卡。Excel 度量不会在快照中显示，直到重新建立与 Excel 模板的连接为止。

MS Excel 集成设置

Decision List Viewer 与 Microsoft Excel 的集成是通过使用预先配置的 Excel 电子表格模板实现的。该模板由以下三个工作表组成：

模型测量。显示导入的Decision List Viewer测量量、自定义 Excel 测量量，以及计算总计（在“设置”工作表中定义）。

设置。提供用于基于导入的Decision List Viewer测量量和自定义 Excel 测量量生成计算的变量。

配置。提供用于指定从Decision List Viewer导入哪些测量量以及用于定义自定义 Excel 测量量的选项。

警告：“配置”工作表的结构已严格定义。请**勿**编辑绿色阴影区域中的任何单元。

- **来自模型的度量。**指明在计算中使用哪些Decision List Viewer度量。
- **导入到模型的度量。**指明哪些 Excel 生成的度量将被返回Decision List Viewer。Excel 生成的度量在Decision List Viewer中显示为新的测量量列。

注意：生成新模型时，模型不会保留 Excel 度量；度量只在活动会话中有效。

更改模型测量

下列示例演示如何通过多种方法更改模型测量：

- 更改现有测量。
- 从模型导入其他标准测量。
- 将其他自定义测量导出到模型。

更改现有测量

▶ 打开模板并选择“配置”工作表。

▶ 通过突出显示并重写名称或说明来编辑任何**名称**或**说明**。

请注意，如果要更改测量（例如，为了提示用户概率而非频数），只需更改**来自模型的度量**中的名称和说明 – 该名称和说明随后将显示在模型中并且用户可以选择要映射的恰当测量。

从模型导入其他标准测量

▶ 打开模板并选择“配置”工作表。

▶ 从菜单中选择：
工具 > 保护 > 不受保护的表单

▶ 选择 A5 单元格，该单元格有黄色阴影且包含**结束**字。

▶ 从菜单中选择：
插入 > 行(W)

▶ 在新测量的**名称**和**说明**中键入相应内容。例如，**错误**和**段的相关错误**。

▶ 在 C5 单元格中输入公式 `=COLUMN('Model Measures'!N3)`。

▶ 在 D5 单元格中输入公式 `=ROW('Model Measures'!N3)+1`。

这些公式会使新的测量显示在模型测量工作表的 N 列中，此列目前为空。

▶ 从菜单中选择：
工具 > 保护 > 保护表单

▶ 单击**确定**。

▶ 在模型测量工作表中，确保 N3 单元格已将**错误**作为新列的标题。

▶ 选择整个 N 列。

- 从菜单中选择：
  格式 > 单元格
- 默认情况下，所有单元均有一个**一般**数字类别。单击**百分比**可更改数字显示的方式。此方法可帮助您检查 Excel 中的数字；此外，也提供给您另外一种使用数字的方法，例如可将数字用作图表的输出。
- 单击**确定**。
- 将电子表格保存为 Excel 2003 模板，该模板具有唯一的名称且文件扩展名为 .xlt。为了易于定位新模板，建议您将其保存在本地或网络文件系统上的预先配置的模板中。

将其他自定义测量导出到模型

- 打开之前示例中已添加“错误”列的模板；选择“配置”工作表。
- 从菜单中选择：
  工具 > 保护 > 不受保护的表单
- 选择 A14 单元格，该单元格有黄色阴影且包含**结束**字。
- 从菜单中选择：
  插入 > 行(W)
- 在新测量的**名称**和**说明**中键入相应内容。例如，**定比变换错误**和**应用于 Excel 错误的定比变换**。
- 在 C14 单元格中输入公式 =COLUMN(‘Model Measures’!03)。
- 在 D14 单元格中输入公式 =ROW(‘Model Measures’!03)+1。

  这些公式指定 0 列将提供模型的新测量。
- 选择“设置”工作表。
- 在 A17 单元格中输入说明’- **定比变化错误**。
- 在 B17 单元格中输入 10 的定比变换因子。
- 在“模型测量”工作表中，在 03 单元格中输入说明**定比变换错误**作为新列的标题。
- 在 04 单元格中输入公式 =N4*Settings!B17。
- 选择 04 单元格的右下角并将其向下拖动到 022 单元格，以将公式复制到每一个单元格中。
- 从菜单中选择：
  工具 > 保护 > 保护表单
- 单击**确定**。
- 将电子表格保存为 Excel 2003 模板，该模板具有唯一的名称且文件扩展名为 .xlt。为了易于定位新模板，建议您将其保存在本地或网络文件系统上的预先配置的模板中。

  当使用该模板连接 Excel 时，“错误”值可用作新的自定义测量。

对模型进行可视化处理

了解模型作用的最佳方式是对其进行可视化处理。使用收益图表，可以通过研究多个替代项的实际效果深入掌握有关模型商业收益和技术收益的有价值的日常信息。收益图表部分显示了某个模型在随机决策过程中的收益，并可于存在替代模型时实现对多个图表的直接比较。

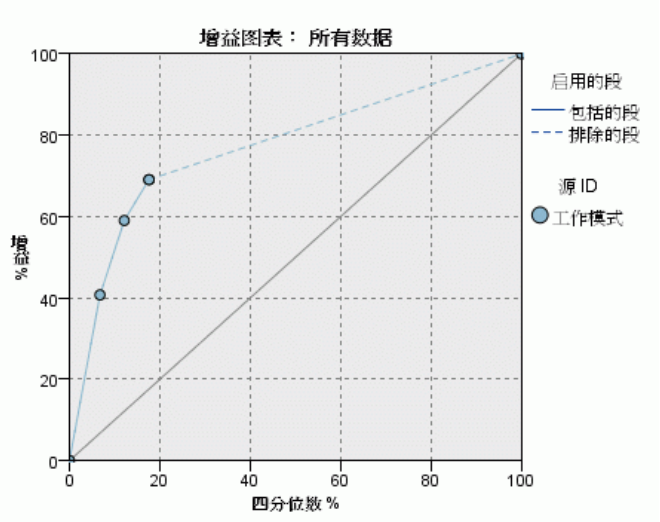
收益图表

收益图绘制的是表中收益 % 列值的散点图。收益定义为每个增量中匹配项数与树中匹配项总数的比例，它使用下列等式：

（增量中匹配项数/匹配项总数）x 100%

收益图表有效地为您说明需要怎样的撒网广度才能捕获树中所有匹配项的给定百分比。斜线绘制整个样本在未使用模型的情况下的预期响应。这种情况下，响应率应该为常量，因为一个人响应的可能性与另一个人相同。为了使您的收益加倍，您需要询问两倍数量的人。曲线表明通过将那些秩（基于收益排序）位于较高百分比的人员包括在内，您可以使得响应得到多大程度的改善。例如，包括最高的 50% 可能会网罗超过 70% 的正面响应。该曲线越陡，收益越高。

图片 9-20
“收益”选项卡



要查看收益图表，请执行下列操作：

- 打开包含决策列表节点的流，并从该节点启动一个交互会话。
- 单击**收益**选项卡。根据指定的分区，您会看到一个或两个图表（例如，如果同时为模型测量定义了训练分区和检验分区，则会显示两个图表）。

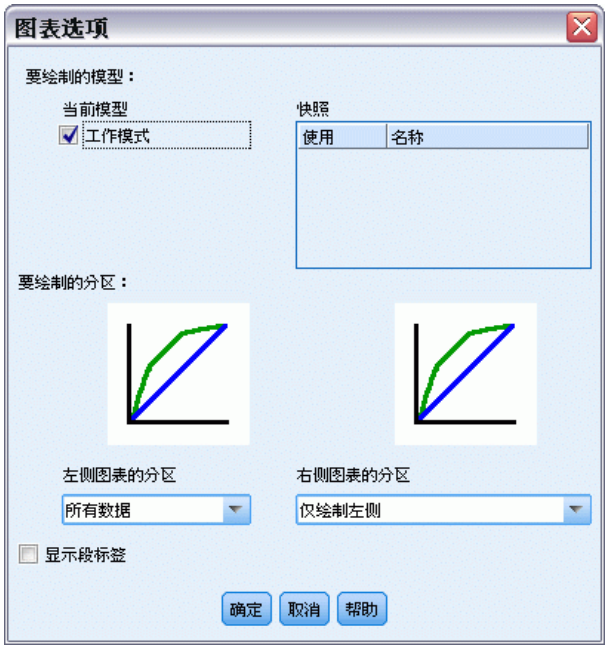
  默认情况下，图表会显示为段。您可以将图表切换为分位数显示，方法是选择**分位数**，然后在下拉菜单中选择适当的分位数方法。

  注意：有关使用图形的信息，请参阅编辑直观表示。

图表选项

“图表选项”功能提供的选项可用于选择以图表显示哪些模型和快照、绘制哪些分区，以及是否显示段标签。

图片 9-21
“图表选项”对话框



要绘制的模型

当前模型。允许您选择要以图表显示的模型。您可以选择工作模型或任何已创建的快照模型。

要绘制的分区

左侧图的分区。该下拉列表提供显示所有已定义分区或所有数据的选项。

右侧图的分区。该下拉列表提供显示所有已定义分区、所有数据或仅显示左侧图表的选项。如果选择**只绘制左侧图**，则仅显示左侧图表。

显示段标签。如果启用此选项，则会在图表中显示全部的段标签。

章
10

统计模型

统计模型使用数学方程式对从数据中提取的信息进行编码。在某些情况下，统计建模技术能非常快速地提供合适的模型。甚至对于那些只有更加灵活的机器学习技术（例如神经网络）才能最终给出更好结果的问题，仍然可以将统计模型作为基线预测模型以判断更先进技术的性能。

以下为可用的统计建模节点。

线性回归模型根据目标与一个或多个预测变量间的线性关系来预测连续目标。 有关详细信息，请参阅第 221 页码线性模型。

Logistic 回归是一种统计方法，它可根据输入字段的值对记录进行分类。它类似于线性回归，但采用的是类别目标字段而非数字范围。 有关详细信息，请参阅第 239 页码逻辑节点。

因子/主成分分析节点提供了用于降低数据复杂程度的强大数据缩减技术。主成份分析（PCA）可找出输入字段的线性组合，该组合最好地捕获了整个字段集合中的方差，且组合中的各个成分相互正交（相互垂直）。因子分析则尝试识别底层因素，这些因素说明了观测的字段集合内的相关性模式。这两种方式的目标都是找到有效概括原始字段集中的信息的一小部分导出字段。 有关详细信息，请参阅第 255 页码主成份分析/因子节点。

判别式分析所做的假设比 logistic 回归的假设更严格，但在符合这些假设时，判别式分析可以作为 logistic 回归分析的有用替代项或补充。 有关详细信息，请参阅第 263 页码判别式节点。

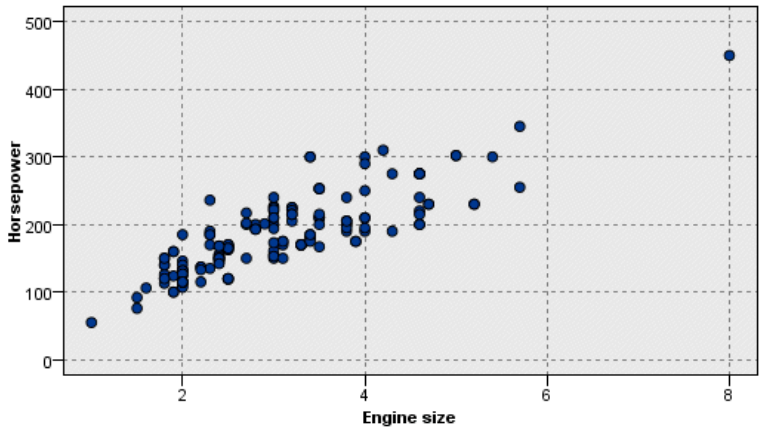
“广义线性”模型对一般线性模型进行了扩展，这样因变量通过指定的关联函数与因子和协变量线性相关。另外，该模型允许因变量呈非正态分布。它包括统计模型大部分的功能，其中包括线性回归、logistic 回归、用于计数数据的对数线性模型以及区间删失生存模型。 有关详细信息，请参阅第 270 页码GenLin 节点。

使用 Cox 回归节点，您可以在已有的检查记录中建立时间事件的生存模型。该模型会生成一个生存函数，该函数可预测在给定时间 (t) 内对于所给定的输入变量值相关事件的发生概率。 有关详细信息，请参阅第 282 页码Cox 节点。

线性节点

线性回归是一种常用统计方法，它可根据数值输入字段的值对记录进行分类。线性回归拟合将预测输出值与实际输出值之间的差异最小化的直线或平面。

图片 10-1
简单线性回归图



要求。在线性回归模型中只能使用数值字段。必须有且仅有一个目标字段（角色设置为目标），但可以有一个或多个预测变量（角色设置为输入）。角色为两者或无的字段将被忽略，就像对待非数值字段一样。（如有必要，可以使用导出节点对非数字字段进行重新编码。 有关详细信息，请参阅第 4 章中的使用派生节点对值进行重新编码中的IBM SPSS Modeler 14.2 源、过程和输出节点。）

强度。线性回归模型相对简单，用来形成预测的数学公司易于解释。由于线性回归是一种由来已久的统计方法，因此线性回归模型的属性已广为人所熟知。而且线性模型训练起来也非常快。线性节点提供了可排除方程式中无意义输入字段的自动字段选择方法。

注意：对于目标字段为类别（如 yes/no 或 churn/don’t churn）而非连续范围的情况，可以将 Logistic 回归用作替代方法。Logistic 回归还支持非数值输入，因而无需对这些字段进行重新编码。 有关详细信息，请参阅第 239 页码逻辑节点。

线性模型

线性模型根据目标与一个或多个预测变量间的线性关系来预测连续目标。

线性模型相对简单，用于评分的数学公式也易于解释。这些模型的属性比较好理解，与同一数据集上的其他模型类型（如神经网络或决策树）相比能够非常快速构建。

示例。在调查业主保险理赔方面拥有有限资源的保险公司希望构建一个模型来估计理赔成本。通过在服务中心部署该模型，客服代表可以在接听客户电话的同时输入理赔信息，并立即获得基于以往数据的“预期”成本。

图片 10-2
“字段”选项卡

字段要求。必须有一个目标和至少一个输入。默认情况下，不使用带“两者”或“无”预定义角色的字段。目标必须为连续（刻度）。对预测变量（输入）没有测量级别限制。分类（标记、名义、有序）字段用作模型中的因子，同时连续字段用作协变量。有关详细信息，请参阅第 30 页码第 3 章中的建模节点字段选项。

目标

图片 10-4
目标设置

您希望做什么？

- **构建新的模型。**构建全新的模型。这是该节点的常用操作。
- **继续训练现有模型。**继续训练此节点成功生成的最后一个模型。这样就可以在无需访问原始数据的情况下更新或刷新现有的模型，并可能会显著提升性能，因为只有新的或更新后的记录被传入流中。上一个模型的详细信息与建模节点存储在一起，这样即使先前的模型块在流或模型选项板中不再可用的情况下，也可以使用该选项。

  注意：在启用此选项后，“字段”和“构建选项”选项卡上的所有其他控件将被禁用。

您的主要目标是什么？

- **创建标准模型。**此方法将构建单个模型，以使用预测变量来预测目标。一般来说，与推进、bagged 或大数据集整体模型相比，标准模型更易于解释，并能更快地进行评分。
- **增强模型准确性（推进）。**此方法采用推进方式构建整体模型，这将生成一系列模型以获得更精确的预测结果。与标准模型相比，此整体模型需要更长的构建与评分时间。

  推进方法产生一系列“成分模型”，其中每个模型在整个数据集上构建。在构建每个后续成分模型之前，将根据前一成分模型的残差对记录进行加权。具有较大残差的个案将被给予较高的分析权重，因此下一个成分模型将较好地侧重于这些记录。

这些成分模型共同构成一个整体模型。该整体模型采用组合规则对新记录进行评分。可用的规则取决于目标的测量级别。

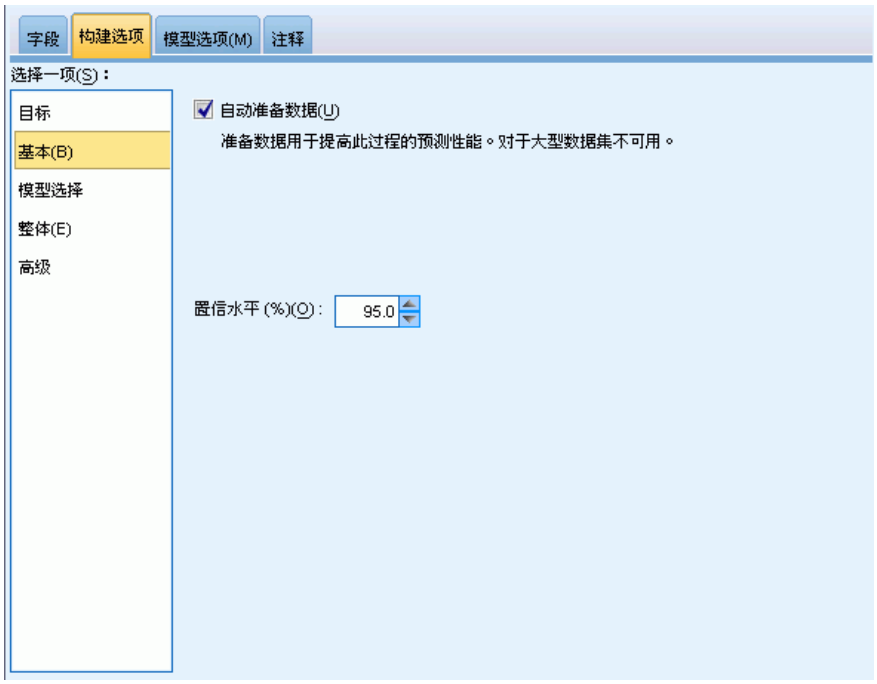
- **增强模型稳定性（bagging）。**此方法采用 bagging（Bootstrap 汇总）方式构建整体模型，这将生成多个模型以获得更可靠的预测结果。与标准模型相比，此整体模型需要更长的构建与评分时间。

  Bootstrap 汇总（bagging）通过对原始数据集进行放回抽样，产生训练数据集的副本。这将创建大小与原始数据集相同的 bootstrap 样本。然后，在每个副本上构建“成分模型”。这些成分模型共同构成一个整体模型。该整体模型采用组合规则对新记录进行评分。可用的规则取决于目标的测量级别。

- **创建适用于大型数据集（需要 IBM SPSS Modeler Server）的模型。**此方法将数据集划分为多个单独数据块，以构建整体模型。如果您的数据集过大，而无法构建上述任何模型或进行增量式建模，请选择此项。与标准模型相比，此选项的构建时间较短，但评分时间更长。该选项需要 SPSS Modeler Server连接。

基本

图片 10-5
基本设置



自动准备数据。 该选项允许在内部转换目标和预测变量，以使模型的预测能力最大化；将保存模型的任何转换并应用到新数据用于评分。转换字段的原始版本将从模型中排除。默认情况下，执行以下自动数据准备。

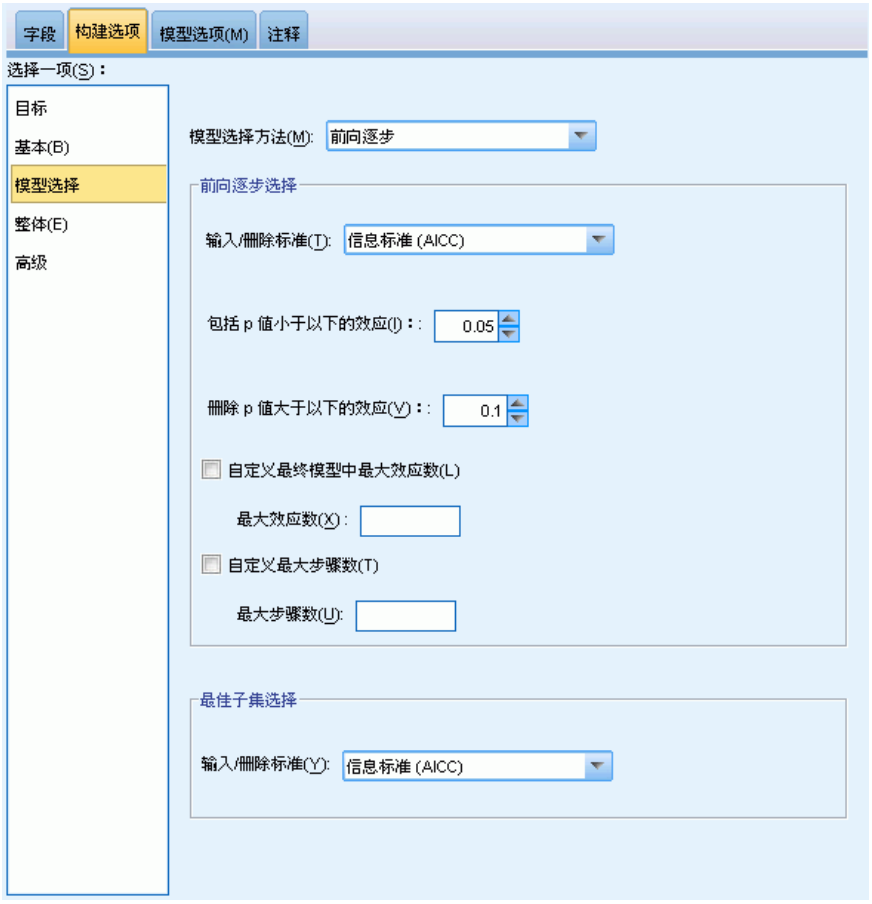
- **日期与时间处理。**每个日期预测变量被转换成新的连续预测变量，其中包含自参考日期（1970-01-01）以来经过的时间。每个时间预测变量被转换成新的连续预测变量，其中包含自参考时间（00:00:00）以来经过的时间。

- **调整测量级别。**具有少于 5 个不同值的连续预测变量将被重新设计成有序预测变量。具有多于 10 个不同值的有序预测变量将被重新设计成连续预测变量。
- **离群值处理。**如果连续预测变量的值位于截断值（平均值的 3 个标准差）之外，则将其设为截断值。
- **缺失值处理。**名义预测变量的缺失值被替换为训练分区的众数。有序预测变量的缺失值被替换为训练分区的中位数。连续预测变量的缺失值被替换为训练分区的平均值。
- **受监督的合并。**这将减少与目标关联的需处理的字段数，得到更简约的模型。通过输入与目标间的关系可以确定类似的类别。无显著差异（即 p 值大于 0.1）的类别则被合并。如果所有类别合并为一个类别，则字段的原始和派生版本将从模型中排除，因为它们没有作为预测变量的值。

置信水平。此为用于在系数视图中计算模型系数的区间估计值的置信水平。指定一个大于 0 且小于 100 的值。默认值为 95。

模型选择

图片 10-6
模型选择设置



模型选择方法。 选择一种模型选择方法（下面将详细介绍）或**包括所有预测变量**，后者简单地输入所有可用预测变量作为主效应模型项。默认使用**前向逐步**。

前向逐步选择。在开始时模型中没有任何效应，然后在每个步骤中添加和删除效应，直到根据逐步选择标准不能再添加或删除效应为止。

- **纳入/移除标准。**此为用于决定是将某个效应添加到还是剔除出模型的统计量。**信息准则 (AICC)** 基于模型中给定训练集合的似然估计，并可调整以惩罚过度复杂模型。**F 统计量**基于有关模型错误改进情况的某个统计量检验。**调整 R 方**基于训练集合的拟合度，并可调整以惩罚过度复杂模型。**防止过度拟合准则 (ASE)** 基于防止过度拟合集的拟合度（平均方差，或 ASE）。防止过度拟合集是不用于训练模型且大约为原始数据集 30% 的随机子样本。

  如果选择了 **F 统计量**以外的标准，则在每步中将对应于选择标准的最大正增长的效应添加到模型。对应于标准中减少情况的任何模型效应将被移除。

  如果选择了 **F 统计量**作为标准，则在每步中将具有低于指定阈值（**纳入 p 值小于此值的效应**）的最小 p 值的效应添加到模型。默认值为 0.05。任何具有大于指定阈值**移除 p 值大于此值的效应**的 p 值的模型效应将被移除。默认值为 0.10。
- **自定义最终模型中的最大效应数。**默认情况下，所有可用效应都将被输入模型中。或者，如果逐步选择算法在具有指定最大效应数的某个步骤结束，则此算法将以当前效应集合结束。
- **自定义最大步骤数。**逐步选择算法在达到特定步骤数后停止。此值默认为可用效应数的 3 倍。或者，指定一个正整数作为最大步骤数。

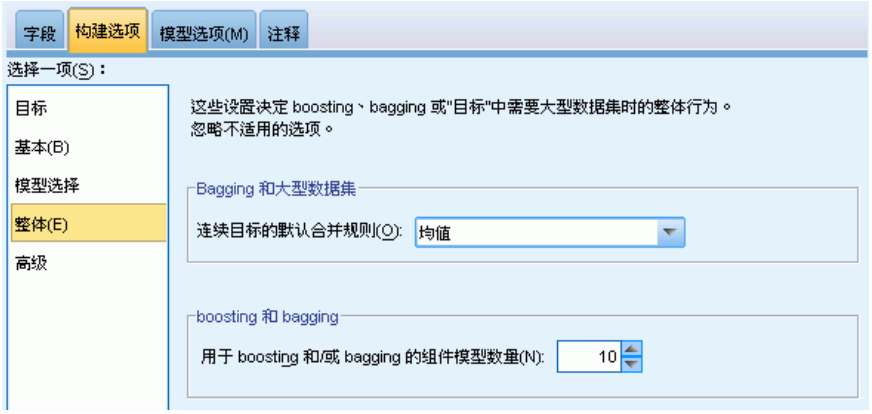
最佳子集选择。这将检查“所有可能的”模型，或至少检查可能模型的较大子集（大于“前向逐步”方法），以选择满足相应标准的最佳子集。**信息准则 (AICC)** 基于模型中给定训练集合的似然估计，并可调整以惩罚过度复杂模型。**调整 R 方**基于训练集合的拟合度，并可调整以惩罚过度复杂模型。**防止过度拟合准则 (ASE)** 基于防止过度拟合集的拟合度（平均方差，或 ASE）。防止过度拟合集是不用于训练模型且大约为原始数据集 30% 的随机子样本。

选择具有最大标准值的模型作为最佳模型。

注意：与向前逐步选择相比，最佳子集选择涉及更密集的计算。在与 boosting、bagging 或超大型数据集配合执行最佳子集时，花费的时间比使用向前逐步选择构建标准模型要长得多。

整体

图片 10-7
整体设置



这些设置决定了在“目标”中请求 boosting、bagging 或超大型数据集时发生的整体行为。对选定目标不适用的选项将被忽略。

Bagging 和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，以计算整体得分值。

- **连续目标的默认组合规则。**可以通过对来自基本模型的预测值取平均值或中位数，对连续目标的整体预测值进行组合。

注意，如果以增强模型精确性为目标，则组合规则选择将被忽略。Boosting 方法始终使用加权大多数投票来对分类目标进行评分，而使用加权中位数对连续目标进行评分。

Boosting 和 Bagging。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 bagging 方法，此为 bootstrap 样本数。它应为正整数。

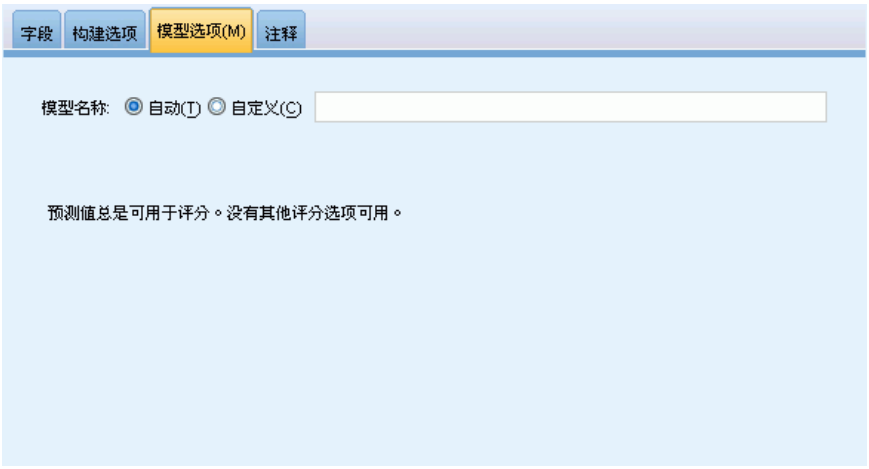
高级

图片 10-8
高级设置

重复结果。设置随机种子允许您复制分析。随机数生成器用于选择哪个记录在过度拟合集中。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。默认值为 54752075。

模型选项

图片 10-9
“模型选项”选项卡

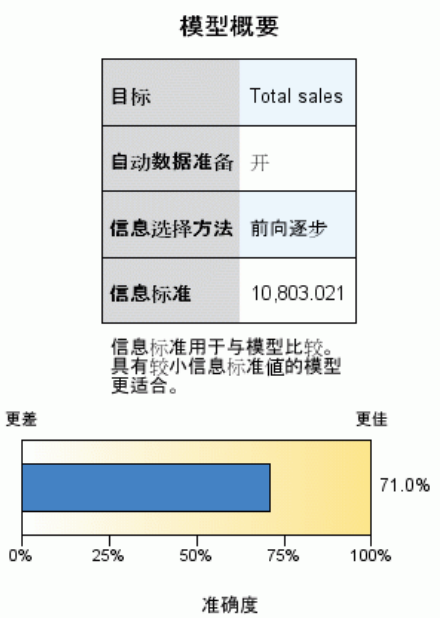


模型名称。可以基于目标字段来自动生成模型名称，或指定自定义名称。自动生成的名称为目标字段名。

请注意，在对模型评分时，始终会计算预测值。新字段的名称为目标字段的名称，加上前缀 $L-。例如，对于名为 sales 的目标字段，新字段将命名为 $L-sales。

模型摘要

图片 10-10
模型摘要视图



“模型摘要”视图是模型及其拟合的快照一览摘要。

表。该表标识一些高级模型设置，包括：

- 目标名称，在字段选项卡上指定；
- 是否执行自动数据准备，在基本设置中指定；
- 模型选择方法和选择标准，在模型选择设置中指定。还显示了最终模型的选择标准值，并以较小、较佳的格式显示。

图表。此图表显示最终模型的精确性，数值越大越好。对于最终模型，此值为 100 × 调整后的 R^2。

自动数据准备

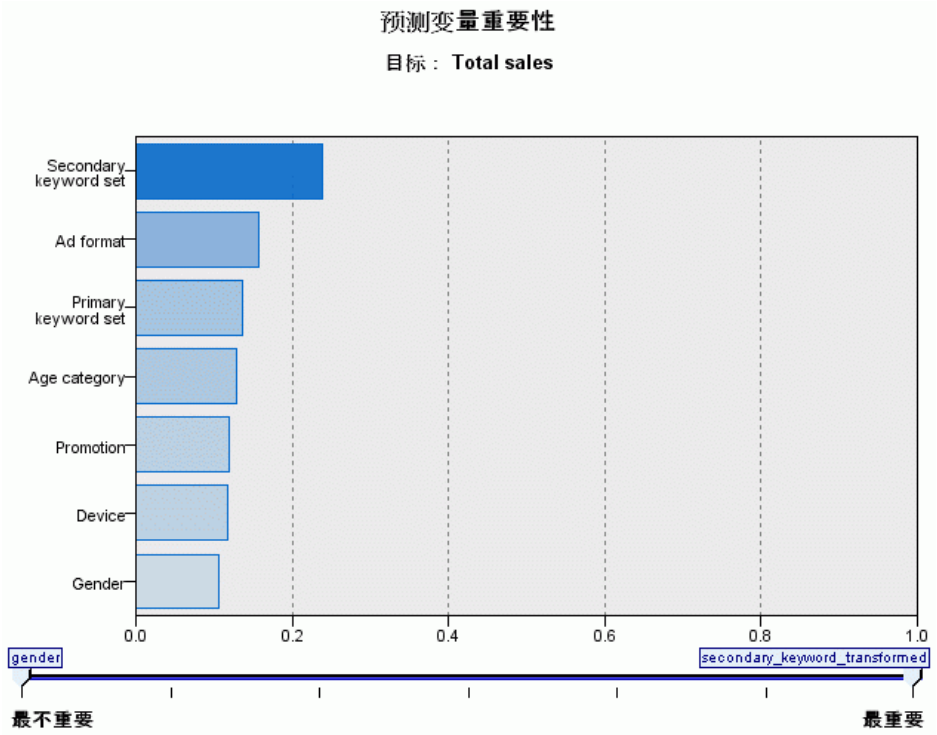
图片 10-11
自动数据准备视图

此视图显示在自动数据准备 (ADP) 步骤中排除了哪些字段，以及转换字段的派生方式等信息。对于每个转换或排除字段，在此表中列出了字段名、在分析中的角色，以及 ADP 步骤所采取的操作。这些字段按其名称的字母升序排列。对每个字段可能执行的操作包括：

- **导出持续时间：月份**以月为单位，计算从包含日期的字段值到当前系统日期所经过的时间。
- **导出持续时间：小时**以小时为单位，计算从包含时间的字段值到当前系统时间所经过的时间。
- **将测量级别从连续改为有序**将不到 5 个唯一值的连续字段重新设计为有序字段。
- **将测量级别从有序改为连续**将超过 10 个唯一值的有序字段重新设计为连续字段。
- **删除离群值**如果连续预测变量的值位于截断值（平均值的 3 个标准差）之外，则将其设为截断值。
- **替换缺失值**分别使用众数、中位数和平均值替换名义字段、有序字段和连续字段的缺失值。
- **合并类别以最大化与目标的关联**根据输入与目标间的关系确定“类似”的预测变量类别。无显著差异（即 p 值大于 0.05）的类别则被合并。
- **排除常量预测变量/在离群值处理之后/在合并类别之后**删除具有单个值的预测变量，可能在执行其他 ADP 操作之后。

预测变量重要性

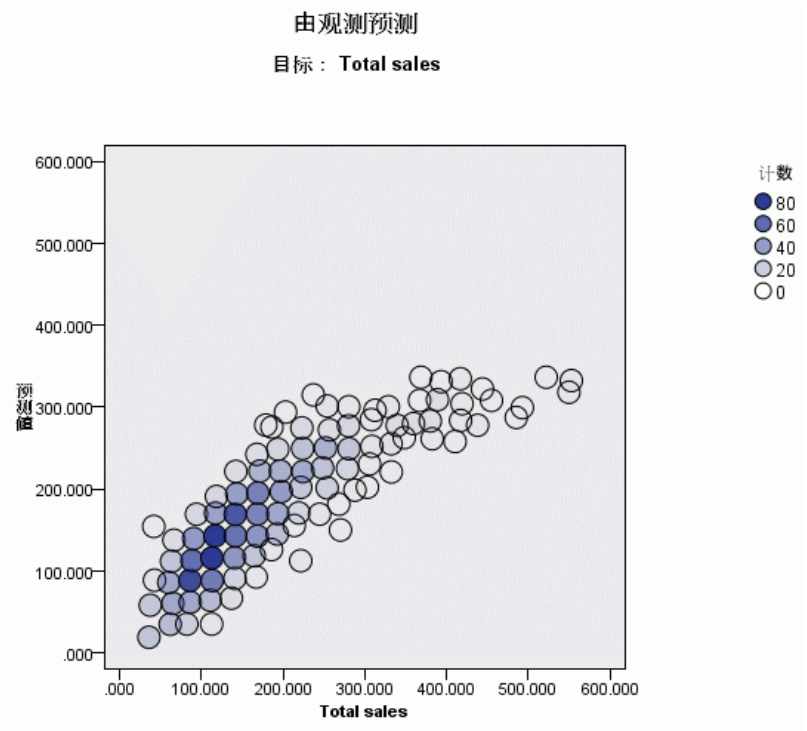
图片 10-12
预测变量重要性视图



通常，您将需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的变量。预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。由于它们是相对值，因此显示的所有预测变量的值总和为1.0。预测变量的重要性与模型精确性无关。它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

按已观测进行预测

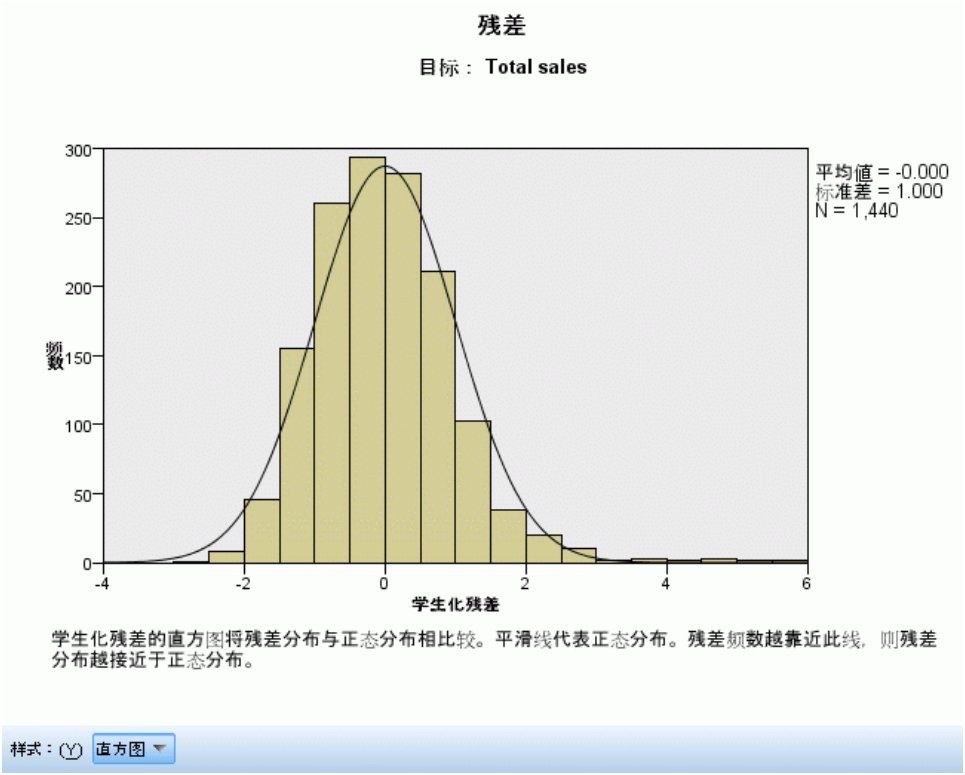
图片 10-13
按已观测进行预测视图



这将显示一个分级散点图，其中预测值位于垂直轴上，而观测值位于水平轴上。理想情况下，该点应在 45 度线上；您可以从该视图上判断出任何被模型预测为较差的纪录。

残差

图片 10-14
残差视图，直方图样式



这将显示模型残差的诊断图表。

图表样式。有多种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

- **直方图。**此为学生化残差的分级直方图，并带有正态分布交叠。线性模型假设残差具有正态分布，因此理想情况下直方图应相当接近平滑线。
- **P-P 图。**此为分级概率-概率 (P-P) 图，将学生化残差与正态分布进行对比。如果绘制点的坡度比正态线更平缓，则残差显示出比正态分布更显著的变异性；如果更陡峭，则残差的变异性低于正态分布。如果绘制点呈 S 型曲线，则残差为偏斜分布。

离群值

图片 10-15
离群值视图

离群值

目标： Total sales

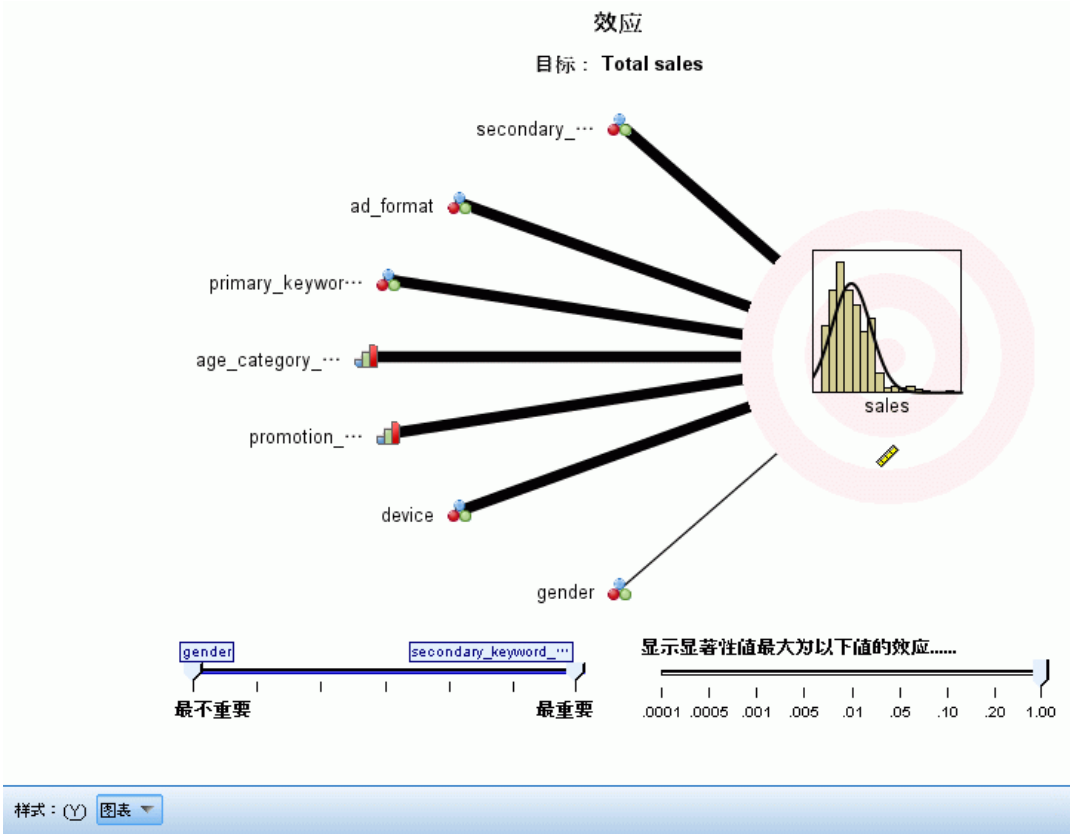
Total sales	Cook's 距离
560.040	0.026
566.440	0.025
548.990	0.018
539.630	0.018
485.430	0.014
543.240	0.014

此表列出对模型施加过度影响的记录，并显示记录 ID（如果在“字段”选项卡上指定）、目标值，以及 Cook 距离。Cook 距离是在特定记录从模型系数的计算中排除的情况下，所有记录的残差变化幅度的测量。较大的 Cook 距离表示在排除记录后系数会发生显著变化，因此应被视为有一定影响。

应仔细检查有影响的记录，以确定是在模型估计中给予较低权重，按照特定可接受阈值截断离群值，还是彻底移除有影响的记录。

效应

图片 10-16
效应视图，图表样式



此视图显示模型中每个效应的大小。

样式。有多种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

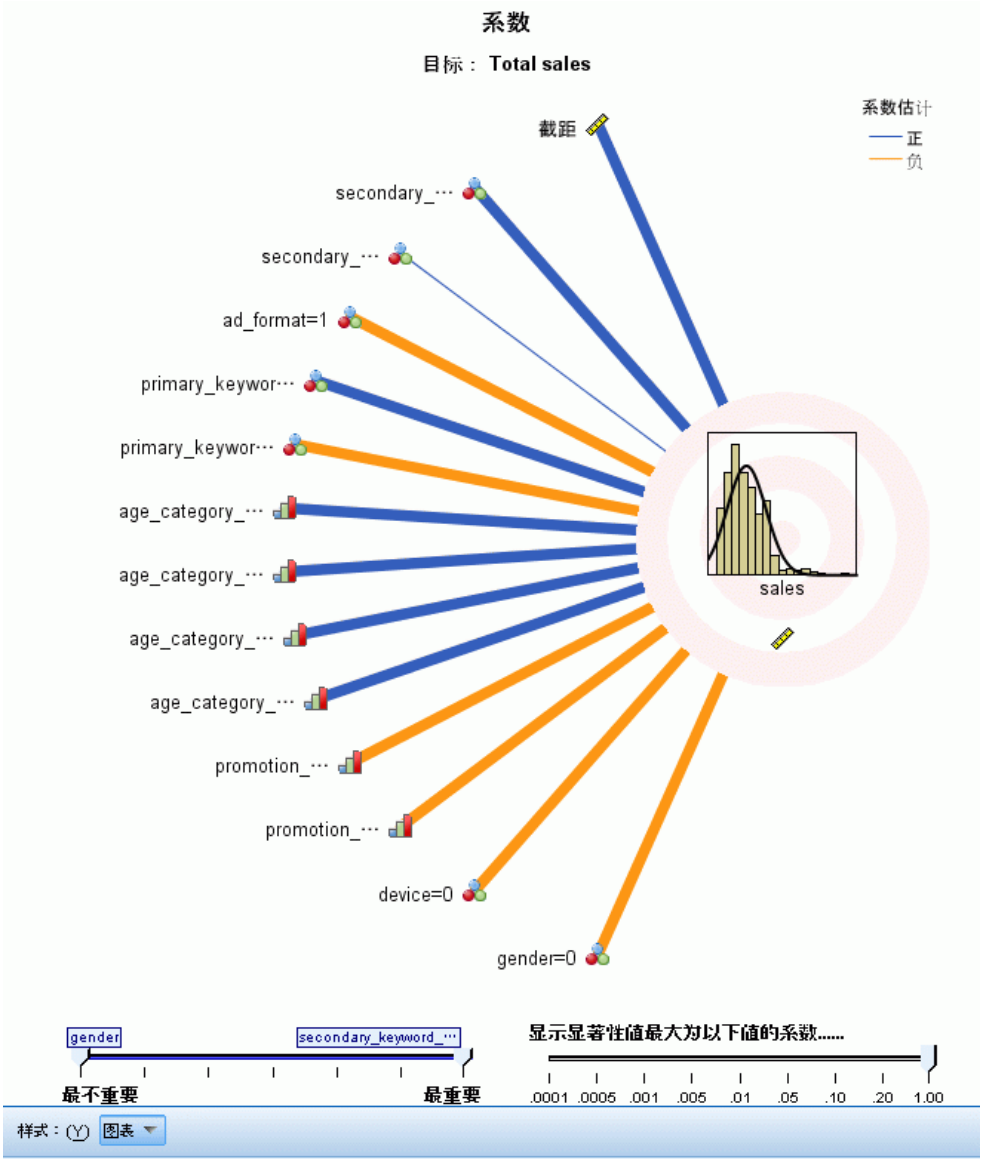
- **图表。**在此图表中，将按预测变量重要性递减顺序，从上到下排列显示效应。在图表中，连接线条根据效应的显著性进行加权，粗线条表示较显著的效应（p 值较小）。悬停在连接线条上将显示工具提示，以指示效应的 p 值和重要性。这是默认值。
- **表**此为总体模型与单独模型效应的 ANOVA 表。各个效应将按预测变量重要性递减顺序，从上到下排列显示。注意，在默认情况下，此表处于折叠状态，只显示总体模型结果。要查看单独模型效应的结果，在表中单击**校正的模型**单元格。

预测变量重要性。提供有一个“预测变量重要性”滑块，以控制在视图中显示哪些预测变量。这不会改变模型，只是帮助您重点关注最重要的预测变量。默认显示前 10 个效应。

显著性。提供有一个“显著性”滑块，以便在按预测变量重要性显示效应的基础上，进一步控制在视图中显示哪些效应。显著性值大于滑块值的效应将被隐藏。这不会改变模型，只是帮助您重点关注最重要的效应。默认情况下此值为 1.00，因此不会根据显著性来过滤效应。

系数

图片 10-17
系数视图，图表样式



此视图显示模型中每个系数的值。注意，由于因子（分类预测变量）在模型内部经过指示符编码，因此包含因子的**效应**通常具有多个关联**系数**；每种类别一个关联系数，但对应于冗余（参考）参数的类别除外。

样式。有多种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

- **图表。**在此图表中，首先显示截距，然后按预测变量重要性递减顺序，从上到下排列显示效应。在包含因子的效应中，系数按照数据值的升序进行排列。在图表中，连接线条根据系数的显著性（参见图表键）而具有不同颜色，粗线条表示较

显著的系数（p 值较小）。悬停在连接线条上将显示工具提示，以指示与参数关联的效应的系数值、p 值和重要性。这是默认样式。

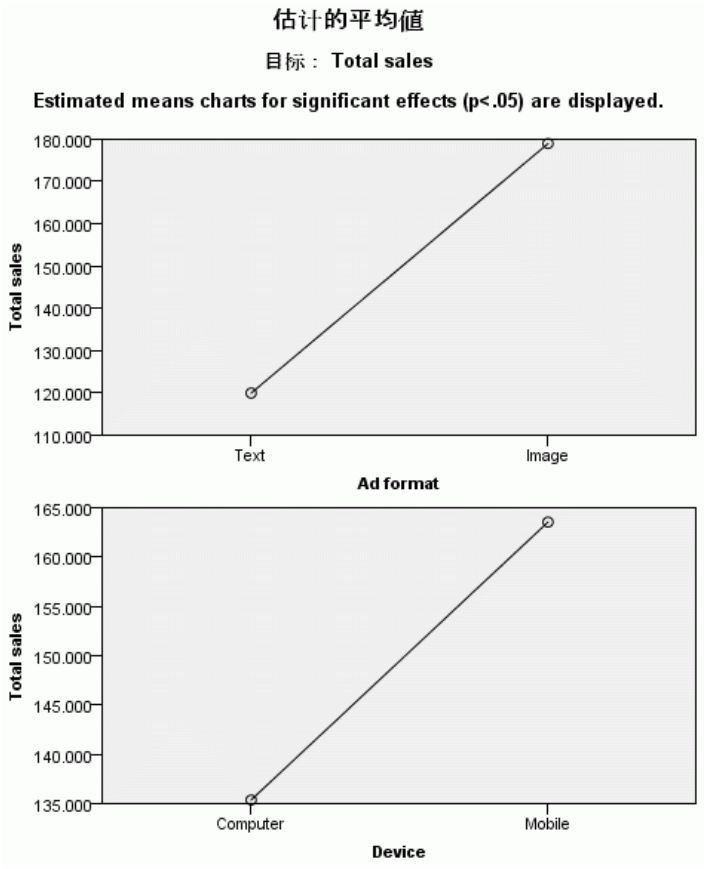
- **表**这将显示单独模型系数的值、显著性检验，以及置信区间。在截距后面，各个效应将按预测变量重要性递减顺序，从上到下排列显示。在包含因子的效应中，系数按照数据值的升序进行排列。注意，在默认情况下，此表处于折叠状态，只显示每个模型参数的系数、显著性和重要性。要查看标准误、t 统计量和置信区间，在表中单击**系数**单元格。悬停在表中的模型参数名称上，将显示工具提示，以指示参数名称、与参数关联的效应以及与模型参数关联的值标签（对于分类预测变量）。当自动数据准备合并分类预测变量的相似类别时，这尤其适合用于查看新创建的类别。

预测变量重要性。提供有一个“预测变量重要性”滑块，以控制在视图中显示哪些预测变量。这不会改变模型，只是帮助您重点关注最重要的预测变量。默认显示前 10 个效应。

显著性。提供有一个“显著性”滑块，以便在按预测变量重要性显示系数的基础上，进一步控制在视图中显示哪些系数。显著性值大于滑块值的系数将被隐藏。这不会改变模型，只是帮助您重点关注最重要的系数。默认情况下此值为 1.00，因此不会根据显著性来过滤系数。

估计平均值

图片 10-18
估计平均值视图

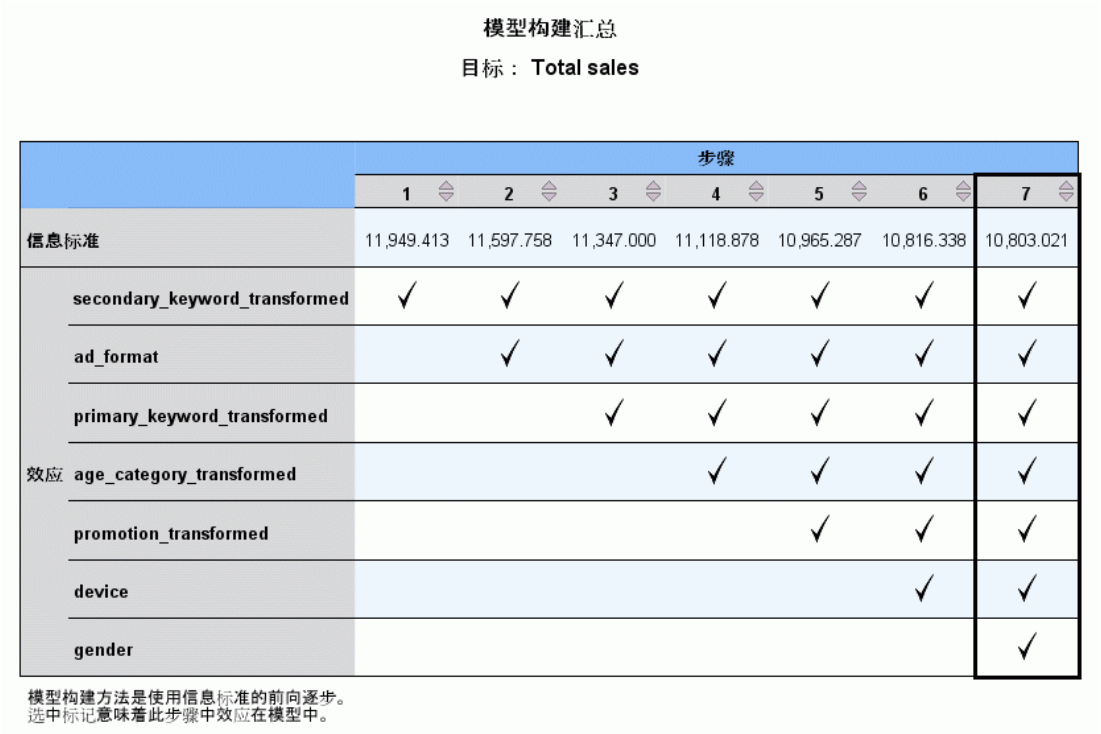


只为显著的预测变量显示这些图表。在图表中，目标的模型估计值位于垂直轴上，预测变量的每个值位于水平轴上，所有其他预测变量保持恒定。它提供了有关每个预测变量系数在目标上的效应的直观表示，非常有用。

注意：如果没有显著的预测变量，则不会生成估计平均值。

模型构建摘要

图片 10-19
模型构建摘要视图，前向逐步算法



模型构建汇总

目标： Total sales

		步骤						
		1	2	3	4	5	6	7
信息标准		11,949.413	11,597.758	11,347.000	11,118.878	10,965.287	10,816.338	10,803.021
效应	secondary_keyword_transformed	✓	✓	✓	✓	✓	✓	✓
	ad_format		✓	✓	✓	✓	✓	✓
	primary_keyword_transformed			✓	✓	✓	✓	✓
	age_category_transformed				✓	✓	✓	✓
	promotion_transformed					✓	✓	✓
	device						✓	✓
	gender							✓

模型构建方法是使用信息标准的前向逐步。
选中标记意味着此步骤中效应在模型中。

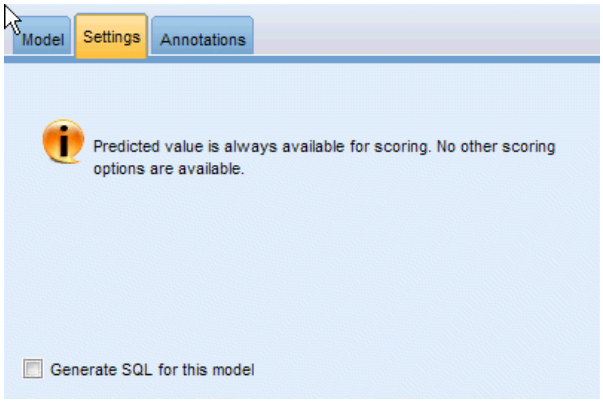
如果在“模型选择”设置中选择了**无**以外的模型选择算法，这将提供有关模型构建过程的一些详细信息。

前向逐步。如果选择算法为前向逐步，此表将显示逐步选择算法中的最近 10 步。对于其中每个步骤，显示在此步骤上选择标准的值与模型中的效应。这允许您了解每个步骤对模型的贡献大小。每列允许您对行进行排序，因此可以方便地看到在给定步骤上模型中有哪些效应。

最佳子集。如果选择算法为最佳子集，此表将显示前 10 个模型。对于每个模型，显示选择标准的值与模型中的效应。您可以从中了解这些最佳模型的稳定性；如果它们倾向于具有存在少量差异的相似效应，那么您可以充分确信它们的确是“最佳”模型；如果它们倾向于具有迥异的效应，那么某些效应可能太相似，需要进行合并（或删除一些）。每列允许您对行进行排序，因此可以方便地看到在给定步骤上模型中有哪些效应。

设置

图片 10-20
“设置”选项卡



请注意，在对模型评分时，始终会计算预测值。新字段的名称为目标字段的名称，加上前缀 $L-。例如，对于名为 sales 的目标字段，新字段将命名为 $L-sales。

生成此模型的 SQL。使用数据库中的数据时，SQL 代码可传回到数据库中执行，从而大大提高许多操作的处理速度。有关详细信息，请参阅第 6 章中的SQL 优化中的IBM SPSS Modeler Server 14.2 管理和性能指南。

逻辑节点

Logistic 回归（也称为**名义回归**）是一种用于依据输入字段的值对记录进行分类的统计技术。这种技术与线性回归类似，但用分类目标字段代替了数值字段。同时支持二项模型（用于具有两种离散类别的目标）和多项模型（用于具有两种以上类别的目标）。

Logistic 回归的工作原理是构建一组方程式，使输入字段值与每个输入字段类别所关联的概率相关。生成模型后，便可以用它来估计新数据的概率。对于每条记录，将计算每种可能输出类别的归属概率。具有最高概率的目标类别将被指定为该记录的预测输出值。

二项模型示例。某电信服务提供商希望了解流失到竞争对手那里的客户数量。使用服务利用率数据，可以创建二项模型以预测哪些客户有可能转向其他提供商，并自定义服务以保留尽可能多的客户。采用二项模型的原因是目标分为两种不同类别（可能转向或可能不转向）。有关详细信息，请参阅第 14 章中的电信客户流失（二项 Logistic 回归）中的IBM SPSS Modeler 14.2 应用程序 指南。

注意：字符串字段的大小必须限制为 8 个字符（仅适用于二项模型）。如有必要，可使用“重新分类”节点对较长的字符串进行重新编码。有关详细信息，请参阅第 4 章中的重新对节点分类中的IBM SPSS Modeler 14.2 源、过程和输出节点。有关详细信息，请参阅第 11 章中的减少输入数据字符串长度（重新分类）中的IBM SPSS Modeler 14.2 应用程序 指南。

多项模型示例。 电信提供商按照服务用途模式划分客户群，将客户分类成四组。通过使用人口统计数据预测组成员，可以创建多项模型，从而将预期客户分为几组，然后针对各个客户自定义服务。有关详细信息，请参阅第 13 章中的电信业客户分类（多项 Logistic 回归）中的IBM SPSS Modeler 14.2 应用程序 指南。

要求。一个或多个输入字段和唯一一个具有两个或多个类别的分类目标字段。对于二项模型，目标必须具有标志测量级别。于多项模型，目标可以具有标志，或名义的测量级别，以及两个或多个类别。设置为双向或无的字段将忽略。必须对模型中使用的字段的类型完全实例化。

强度。Logistic 回归模型通常相当准确。它们可处理符号和数字类型的输入字段。它们可以给出所有目标类别的预测概率，从而能够轻松识别出第二最佳推测值。当组成员关系是真正分类字段时，Logistic 模型最为有效；如果组成员关系基于连续范围字段的值（例如，高 IQ 与低 IQ），则应考虑使用线性回归，以利用整个范围的值所提供的更丰富的信息。Logistic 模型也可以执行自动字段选择，但其他方式（如树模型或特征选择）在对大型数据集执行此操作时可能更迅速。最后，由于 Logistic 模型被很多分析人员和数据挖掘人员所熟知，因此他们可能会将其用作比较其他建模技术的基准。

处理大型数据集时，可以禁用高级输出选项似然比检验，从而显著提高性能。 有关详细信息，请参阅第 247 页码Logistic 回归高级输出。

Logistic 节点模型选项

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

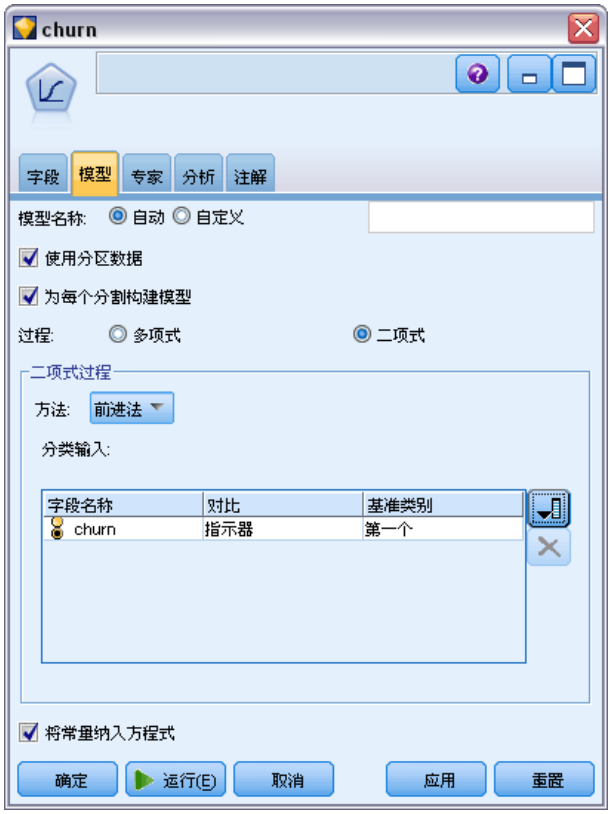
过程。指定是创建二项模型还是创建多项模型。对话框中提供的选项会因所选建模过程的类型而异。

- **二项式。**当目标字段是具有两个离散（二分）值（如是/否、启动/关闭或男/女）的标志或名义字段时使用。
- **多项式。**当目标字段是具有两个以上值的名义字段时，应使用此选项。可以指定**主效应**、**全析因**或**自定义**。

在等式中包含常量。此选项用于确定结果方程式是否将包含常数项。在大多数情况下，应将此选项保持为选中状态。

二项模型

图片 10-21
Logistic 节点：二项模型选项



对于二项模型，可用的方法和选项如下：

方法。指定要用于构建 Logistic 回归模型的方法。

- **按 Enter 键。**这是默认方法，可将所有项直接输入方程式。构建模型时不进行字段选择。
- **前进法。**字段选择的前进法通过逐步向前移动来构建模型。采用这种方法，初始模型是最简单的模型，只能向模型中添加常量和项。每个步骤会对尚未纳入到模型中的项进行检验，看它们对模型的改进起多大作用，然后将其中的最佳项添加到模型中。当无法再添加任何项、或最佳备选项无法对模型产生足够的改进时，最终模型便已生成。
- **后退法。**从本质上说，后退法与前进法是相反的。但采用这种方法时，初始模型包含作为预测变量的所有项，只能从模型中删除项。对模型贡献较小的模型项将被逐一删除，直到无法再删除任何项而不对模型功能造成重大损害，从而生成最终模型。

分类输入。列出标识为分类字段的字段，即具有标志、名义或有序的测量级别。可以为每个分类字段指定对比和基准类别。

- **字段名。**此列包含分类输入的字段名称，并预填入了数据中的所有标志值和名义值。要向此列中添加连续输入字段或数值输入字段，请单击列表右边的“添加字段”图标，然后选择所需输入字段。
- **对比。**分类字段的回归系数的解释取决于所用的对比。对比决定如何设定假设检验以比较估计均值。例如，如果已知某个分类字段具有隐含顺序（如模式或分组），则可以使用对比为该顺序建模。可用的对比如下：

  指示符。这些对比指示类别成员资格是否存在。这是默认方法。

  简单。将预测字段的每个类别（参考类别除外）与参考类别进行比较。

  差分。将预测字段的每个类别（第一个类别除外）与前面类别的平均效果进行比较。也称为逆 Helmert 对比。

  Helmert。将预测字段的每个类别（最后一个类别除外）与后续类别的平均效果进行比较。

  重复。将预测字段的每个类别（第一个类别除外）与前一个类别进行比较。

  多项式。正交多项式对比。假设类别均匀分布。多项式对比仅适用于数值字段。

  偏差。将预测字段的每个类别（参考类别除外）与总体效果进行比较。
- **基准类别。**指定如何针对所选对比类型确定参考类别。选择**第一个**使用输入字段的第一个类别（按字母顺序排列），或选择**最后一个**使用最后一个类别。默认值为“第一个”。

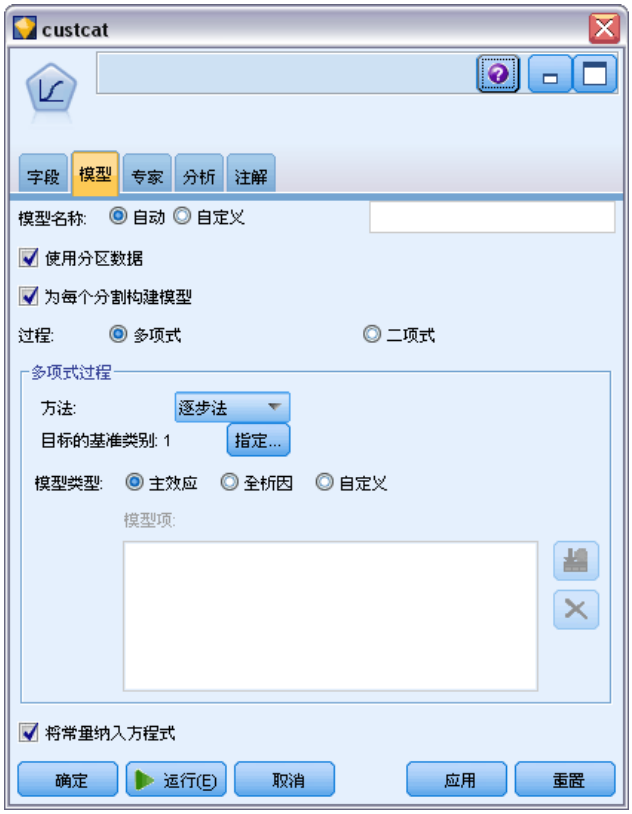
  注意：如果对比设置为“差分”、“Helmert”、“重复”或“多项式”，此字段将不可用。

每个字段对整体响应影响的估计，可以计算为其他各个类别相对于参考类别的似然增量或减量。这有助于确定比较有可能给出特定响应的字段和值。

基准类别在输出中显示为 0.0。这是因为将其与自己进行比较会产生空的结果。所有其他类别均显示为与基准类别相关的方程式。 有关详细信息，请参阅第 250 页码Logistic 模型块详细信息。

多项模型

图片 10-22
Logistic 节点：多项模型选项



对于多项模型，可用的方法和选项如下：

方法。指定要用于构建 Logistic 回归模型的方法。

- **按 Enter 键。**这是默认方法，可将所有项直接输入方程式。构建模型时不进行字段选择。
- **逐步法。**顾名思义，逐步字段选择法就是分步构建方程式。初始模型是可能的最简单模型，其方程式中不含任何模型项（除常量外）。在每个步骤中，对尚未添加到模型的项进行评估，如果其中的最佳项能够显著增加模型预测能力，则将该项添加到模型中。此外，还会重新评估当前包含在模型中的项，以确定能否在不对模型功能造成重大减损的情况下删除其中任何项。如果可以，则会将其删除。然后重复此过程，添加并/或删除其他项。当无法再添加任何项来改进模型、且无法再删除任何项而不对模型功能造成减损时，最终模型便已生成。
- **前进法。**字段选择的前进法与分步构建模型的逐步法类似。但采用这种方法时，初始模型是最简单的模型，只能向模型中添加常量和项。每个步骤会对尚未纳入到模型中的项进行检验，看它们对模型的改进起多大作用，然后将其中的最佳项添加到模型中。当无法再添加任何项、或最佳备选项无法对模型产生足够的改进时，最终模型便已生成。

- **后退法。**从本质上说，后退法与前进法是相反的。但采用这种方法时，初始模型包含作为预测变量的所有项，只能从模型中删除项。对模型贡献较小的模型项将被逐一删除，直到无法再删除任何项而不对模型功能造成重大损害，从而生成最终模型。
- **后退逐步法。**从本质上说，后退逐步法与逐步法是相反的。采用这种方法时，初始模型将包含作为预测变量的所有项。每个步骤会评估模型中的项，并且将可以删除而不对模型功能造成重大减损的项删除。此外，还会对先前删除的项进行重新评估，以确定其中的最佳项是否对模型的预测功能起到显著作用。如果是，则会将其重新添加到模型中。当无法再删除任何项而不对模型功能造成重大减损、且无法再添加任何项以改进模型时，最终模型便已生成。

注意：自动方法（包括逐步法、前进法和后退法）是适应性强的学习方法，并且特别容易过度拟合训练数据。使用这些方法时，用新数据或使用分区节点创建的保留测试样本对结果模型的有效性进行验证尤为重要。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

目标的基准类别。指定如何确定参考类别。这将用作对目标中所有其他类别的回归方程式进行估计的基准。选择**第一个**使用当前目标字段的第一个类别（按字母顺序排列），或选择**最后一个**使用最后一个类别。或者，可以选择**指定**以选择特定类别，并从列表中选择所需的值。可以在类型节点中为每个字段定义可用值。 有关详细信息，请参阅第 4 章中的使用值对话框中的IBM SPSS Modeler 14.2 源、过程和输出节点。

通常应将关注程度最低的类别指定为基准类别，例如低价促销产品。然后再以相对方式将其他类别与该基准类别相关，从而确定什么使它们更有可能自成类别。这有助于确定比较有可能给出特定响应的字段和值。

基准类别在输出中显示为 0.0。这是因为将其与自己进行比较会产生空的结果。所有其他类别均显示为与基准类别相关的方程式。 有关详细信息，请参阅第 250 页码Logistic 模型块详细信息。

模型类型。用于定义模型中的项的选项共有三种。**主效应**模型仅分别包括各个输入字段，而不检验输入字段之间的交互效应（乘法效应）。**全析因**模型包括所有交互效应，以及输入字段主效应。全析因模型捕获复杂关系的能力较强，但也比较难以解释，而且更有可能出现过度拟合情况。由于有可能出现大量可能组合，因此对于全析因模型，自动字段选择方法（进入法以外的方法）处于禁用状态。**自定义**模型仅包括您指定的项（主效应和交互效应）。选择此选项时，应使用“模型项”列表在模型中添加或删除项。

模型项。构建自定义模型时，将需要明确指定模型中的项。此列表显示了模型项的当前集合。“模型项”列表右边的按钮可用于添加或删除模型项。

▶ 要将项添加到模型中，请单击添加新的模型项按钮。

▶ 要删除项，请选定所需项，然后单击删除选定模型项按钮。

将项添加到 Logistic 回归模型

在请求 Logistic 回归模型时，可以通过单击“Logistic 回归模型”选项卡中的添加新的模型项按钮将项添加到模型中。此时将打开一个新的对话框，您可在其中指定项。

图片 10-23
Logistic 回归“新建项”对话框

要添加的项类型。有几种将项添加到模型的方法，具体取决于在“可用字段”列表中对输入字段的选择。

- **单向交互效应。**插入表示所有选定字段的交互效应的项。
- **主效应。**针对每个选定的输入字段插入一个主效应项（该字段本身）。
- **所有双向交互效应。**针对每个可能的选定输入字段对插入一个双向交互效应项（输入字段的积）。例如，如果已在“可用字段”列表中选定输入字段 A、B 和 C，此方法将插入项 A * B、A * C 和 B * C。
- **所有三向交互效应。**针对每个可能的选定输入字段组合（一次取三个）插入一个三向交互效应项（输入字段的积）。例如，如果已在“可用字段”列表中选定输入字段 A、B、C 和 D，此方法将插入项 A * B * C、A * B * D、A * C * D 和 B * C * D。
- **所有四向交互效应。**针对每个可能的选定输入字段组合（一次取四个）插入一个四向交互效应项（输入字段的积）。例如，如果已在“可用字段”列表中选定输入字段 A、B、C、D 和 E，此方法将插入项 A * B * C * D、A * B * C * E、A * B * D * E、A * C * D * E 和 B * C * D * E。

可用字段。列出要用于构造模型项的可用输入字段。

预览。根据上述所选字段和项类型，显示单击**插入**时将添加到模型中的项。

插入。将项插入模型（根据当前选择的字段和项类型）并关闭对话框。

Logistic 节点专家选项

如果具备 Logistic 回归的深入知识，则可以使用专家选项对训练过程进行调整。要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

图片 10-24
Logistic 回归“专家”选项卡

尺度（仅限多项模型）。您可以指定将用于更正参数协方差矩阵的估计值的离差尺度值。Pearson使用 Pearson 卡方统计量估计尺度值。偏差使用偏差函数（似然比卡方）统计量估计尺度值。您也可以指定自己的用户定义尺度值。必须是正数值。

追加所有概率。如果选中此选项，则会将输出字段的每个类别的概率添加到节点所处理的每个记录中。如果未选中此选项，则仅添加预测类别的概率。

例如，包含具有三个类别的多项模型结果的表将包括五个新列。一个列将列出预测正确的结果的概率，第二个列将显示该预测准确或失误的概率，第三个列将显示每个类别的预测失误或准确的概率。 有关详细信息，请参阅第 250 页码Logistic 模型块。

注意：对于二项模型，此选项始终处于选中状态。

奇异性容许误差。指定用于检查异常值的容差。

收敛。这些选项可用于控制模型收敛的参数。当您执行模型时，收敛设置将控制重复运行不同参数以观察其拟合程度的次数。参数的尝试次数越多，结果将越接近（即，结果将会收敛）。 有关详细信息，请参阅第 246 页码Logistic 回归收敛选项。

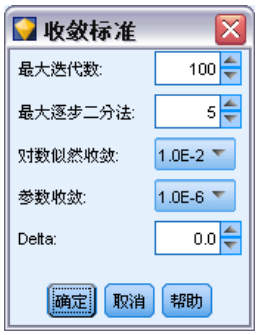
输出。 通过这些选项，可以请求将出现在由节点构建的模型块的高级输出中的附加统计量。 有关详细信息，请参阅第 247 页码Logistic 回归高级输出。

步进。这些选项可用于控制采用逐步、前进、后退或后退逐步估计法添加和删除字段的标准。（如果已选择进入法，该按钮将处于禁用状态。） 有关详细信息，请参阅第 249 页码Logistic 回归步进选项。

Logistic 回归收敛选项

您可设置用于 Logistic 回归模型估计的收敛参数。

图片 10-25
Logistic 回归收敛选项



最大迭代次数。指定用于估计模型的最大迭代次数。

最大步骤对分。逐步二分法是 Logistic 回归用于处理估计过程中的复杂情况的一种技术。在通常情况下，应使用默认设置。

对数似然估计收敛。如果对数似然的相对变化小于此值，迭代将停止。如果值为 0，则不使用该标准。

参数收敛。如果参数的绝对变化或相对变化小于此值，迭代将停止。如果值为 0，则不使用该标准。

Delta（仅用于多项模型）。可以指定要添加到每个空单元格（输入字段和输出字段值的组合）的值，该值介于 0 和 1 之间。当相对于数据中的记录数有许多可能的字段值组合时，此值有助于估计算法处理数据。默认值为 0。

Logistic 回归高级输出

选择要在回归模型块的高级输出中显示的可选输出。要查看高级输出，请浏览模型块并单击高级选项卡。 有关详细信息，请参阅第 254 页码Logistic 模型块高级输出。

二项式选项

图片 10-26
Logistic 回归二项输出选项

选择要为模型生成的输出的类型。 有关详细信息，请参阅第 254 页码Logistic 模型块高级输出。

显示。选择是要在每个步骤中显示结果还是等到所有步骤均已完成时再显示结果。

exp(B) 的 CI。选择表达式中每个系数（显示为 Beta）的置信区间。指定置信区间的水平（默认值为 95%）。

残差诊断。请求残差的“观测值诊断”表。

- **离群值极限（标准差）。**仅列出这样的残差观测值：所列变量的绝对标准化值至少与您指定的值一样大。默认值为 2。
- **全部个案。**在残差的“观测值诊断”表中包括所有观测值。

  注意：由于此选项将列出每个输入记录，因此可能在报告中产生异常巨大的表，其中每个记录占一行。

分类标准值。此选项可用于确定对观测值进行分类的分割点。具有大于分类界限值的预测值的个案被分类为正，具有小于分类界限值的预测值的个案分类为负。要更改默认值，请输入一个 0.01 到 0.99 之间的值。

多项选项

图片 10-27
Logistic 回归：多项输出选项

选择要为模型生成的输出的类型。 有关详细信息，请参阅第 254 页码Logistic 模型块高级输出。

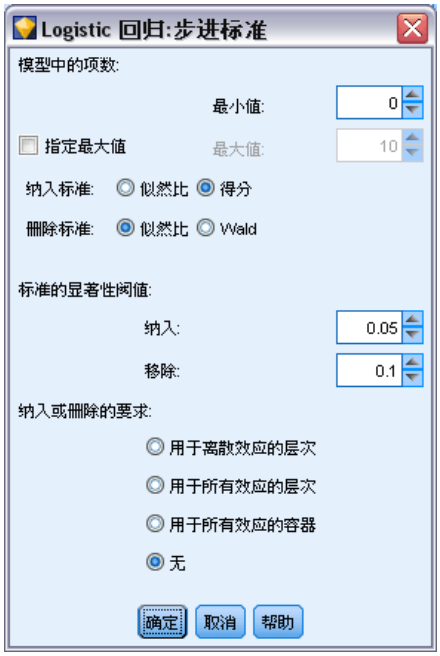
注意：选择**似然比检验**选项会大大增加构建 Logistic 回归模型所需的处理时间。如果模型构建时间过长，可以考虑禁用此选项，或利用 Wald 统计量和得分统计量。 有关详细信息，请参阅第 249 页码Logistic 回归步进选项。

迭代历史间隔。选择在高级输出中打印迭代状态的分步间隔。

置信区间。方程式中系数的置信区间。指定置信区间的水平（默认值为 95%）。

Logistic 回归步进选项

图片 10-28
Logistic 回归步进标准



模型中的项数（仅用于多项模型）。可以指定模型中的最小项数（针对后退法和后退逐步法模型）和最大项数（针对前进法和逐步法模型）。如果指定大于 0 的最小值，模型将包括该数量的项，即使根据统计标准应将其中某些项删除也是如此。对于前进法、逐步法和进入法模型，将忽略最小值设置。如果指定最大值，可能会省略模型中的某些项，即使根据统计标准应将其选中也是如此。对于后退法、后退逐步法和进入法模型，将忽略**指定最大值**设置。

纳入标准（仅适用于多项模型）。选择 **得分**以最大化处理速度。**似然比**选项可能会稍微多提供一些有力的估计值，但所需的计算时间较长。默认设置是使用得分统计量。

移除标准。选择**似然比**可得到更稳健的模型。要缩短构建模型所需的时间，可以尝试选择 Wald。但是，如果数据中有完全或半完全分隔（可使用模型块的“高级”选项卡确定），Wald 统计量将变得极不可靠，不应采用。默认设置是使用似然比统计量。对于二项模型，还有附加选项**条件**。此选项提供以基于条件参数估计值的似然比统计量的概率为依据的移除检验。

标准的显著性阈值。使用此选项可基于每个字段关联的统计概率（p 值）指定选择标准。仅当关联的 p 值小于**纳入标准**值时，才会将字段添加到模型中；仅当 p 值大于**剔出标准**值时，才会将字段删除。**纳入标准**值必须小于**剔出标准**值。

纳入或移除的要求（仅适用于多项模型）。对于某些应用程序，除非模型也包含交互效应项所涉及字段的低阶项，否则将交互效应项添加到模型中在数学上是没有意Òuclå摹例如，除非 A 和 B 也纳入到模型中，否则将 A * B 纳入到模型中没有意义。使用这些选项，可以确定如何在逐步模型项选择过程中处理这些依赖关系。

- **用于离散效果的层次。**仅当相关字段的低阶效应（涉及较少字段的主效应或交互效应）均已进入模型中时，高阶效应（涉及较多字段的交互效应）才会进入模型，而且只要高阶效应仍在模型中，便不会删除涉及相同字段的低阶效应。此选项仅适用于分类字段。 有关详细信息，请参阅第 4 章中的测量级别中的IBM SPSS Modeler 14.2 源、过程和输出节点。
- **用于所有效果的层次。** 此选项的工作原理与上一选项相同，但它适用于所有输入字段。
- **用于所有效果的容器。** 仅当效应中包含的所有效应也纳入到模型中时，该效应才能纳入到模型中。此选项与**用于所有效果的层次**选项类似，只是连续字段的处理方式略有不同。要让一个效应包含另一个效应，被包含（低阶）效应必须包括包含（高阶）效应中涉及的所有连续字段，且被包含效应的分类字段必须是包含效应中离散字段的子集。例如，如果 A 和 B 是分类字段，X 是连续字段，那么项 A * B * X 将包含项 A * X 和 B * X。
- **无。**没有任何强制关系；模型中项的添加和删除是各自独立的。

Logistic 模型块

Logistic 模型块表示由 Logistic 节点估计的方程式。其中包含 Logistic 回归模型捕获的所有信息，以及有关模型结构和性能的信息。这种类型的方程式也可以通过其他模型（如 Oracle SVM）生成。

运行包含 logistic 模型块的流时，该节点可添加包含模型预测和关联概率在内的两个新字段。新字段的名称来自正被预测的输出字段的名称，前缀 $L- 表示预测类别，前缀 $LP- 表示相关概率。例如，对于名为 colorpref 的输出字段，新字段将命名为 $L-colorpref 和 $LP-colorpref。此外，如果在 Logistic 节点中选中了**追加所有概率**选项，则会针对输出字段的每个类别添加一个附加字段，其中包含属于每个记录对应类别的概率。这些附加字段基于输出字段的值进行命名，带有前缀 $LP-。例如，如果 colorpref 的合法值为 Red、Green 和 Blue，则会添加以下三个新字段：$LP-Red、$LP-Green 和 $LP-Blue。

生成过滤节点。使用“生成”菜单可以创建新的过滤节点，用于根据模型结果传递输入字段。因多重共线性而从模型中删除的字段以及模型中未使用的字段将被生成的节点过滤。

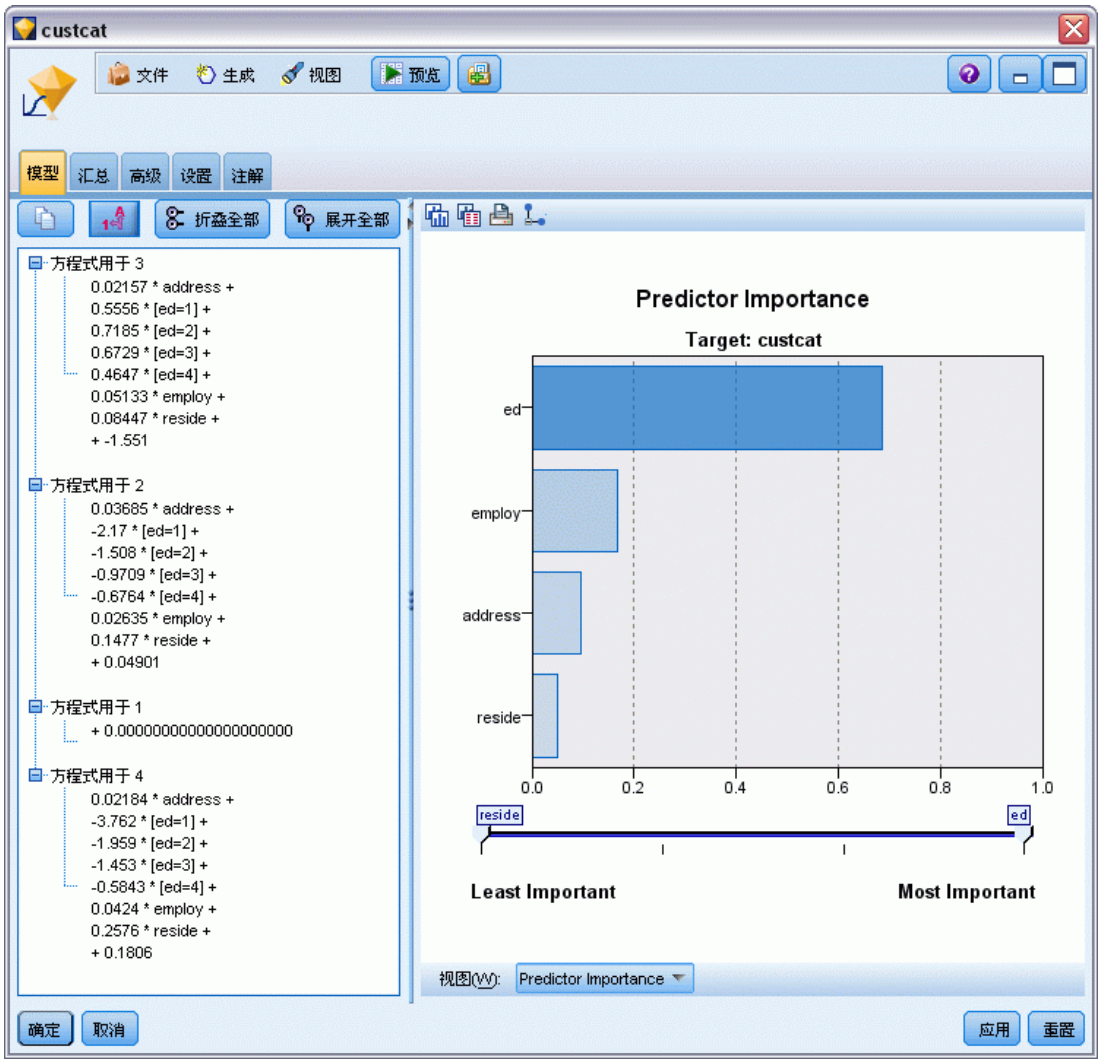
Logistic 模型块详细信息

对于多项模型，Logistic 模型块中的模型选项卡有一个分割显示，在左侧窗格中显示模型方程式，在右侧窗格中显示预测变量重要性。而二项模型的选项卡中只显示预测变量重要性。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

模型方程式

对于多项模型，左窗格显示为 logistic 回归模型估计的实际方程式。在目标字段中，除基准类别之外，每种类别均有一个方程式。这些方程式以树格式显示。这种类型的方程式也可以通过某些其他模型（如 Oracle SVM）生成。

图片 10-29
显示预测变量重要性的 Logistic 块模型详细信息



等式用于。显示用于在给定一组预测变量值的情况下推导出目标类别概率的回归方程式。目标字段的最后一个类别将被视为**基准类别**；显示的方程式将针对一组特定预测变量值给出其他类别相对于基准类别的对数优势比。给定预测变量模式的每个类别的预测概率根据这些对数优势比值推导得出。

如何计算概率

每个方程式会计算一个特定目标类别相对于基准类别的对数优势比。**对数优势比**（也称为**对数**）是指定目标类别相对于基准类别的概率比，并对结果取自然对数。对于基准类别，类别相对于自身的优势比为 1.0，因此其对数优势比为 0。可以将这种情况视为基准类别的隐含方程式，其中所有系数均为 0。

要根据特定目标类别的对数优势比推导出概率，需要取该类别的方程式计算的 logit 值，并应用以下公式：

$P(group_i) = exp(g_i) / \Sigma_k exp(g_k)$

其中 g 是计算的对数优势比，i 是类别参考号，k 为 1 至目标类别数之间的数字。

预测变量重要性

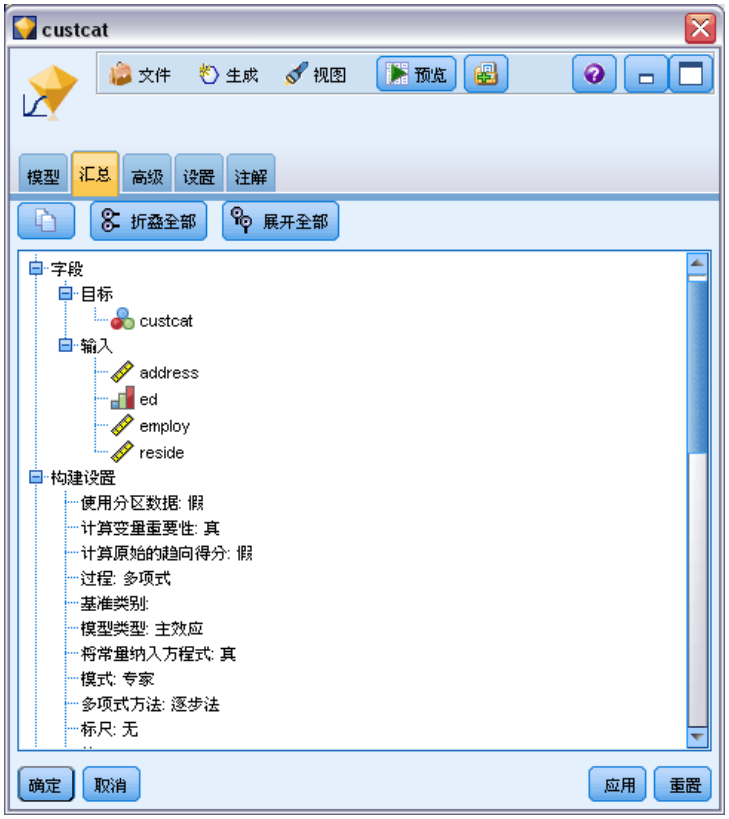
另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。注意，只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

注意：与其他类型的模型相比，计算 logistic 回归的预测变量重要性所用时间更长，所以默认情况下，在“分析”选项卡中不选中预测变量重要性。选中该选项可能会降低性能，对大数据集尤为明显。

Logistic 模型块概要

Logistic 回归模型的概要显示用于生成该模型的字段和设置。此外，如果已执行附加到该建模节点的分析节点，则还会在此部分显示该分析中的信息。有关详细信息，请参阅第 6 章中的分析节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。有关使用模型浏览器的一般信息，请参阅浏览模型块第 43 页码。

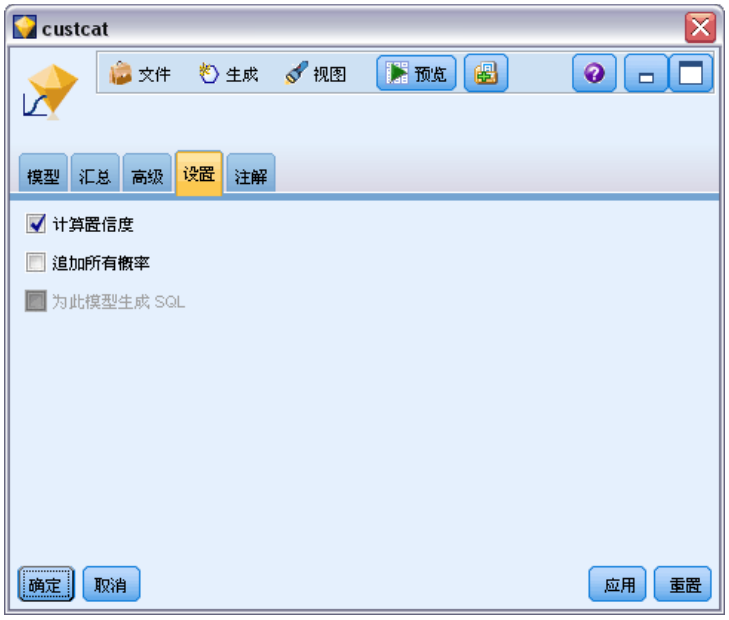
图片 10-30
Logistic 回归模型块：“汇总”选项卡



Logistic 模型块设置

Logistic 模型块中的“设置”选项卡用于指定模型评分过程中的置信度、概率、倾向得分和 SQL 生成选项。该选项卡仅在已将模型块添加到流中之后才可用，而且可以根据模型和目标的类型显示不同选项。

图片 10-31
具有名义目标的多项模型的“设置”



多项模型

对于多项模型，可用的选项如下：

计算置信度。指定是否在评分时计算置信度。

计算原始倾向得分（仅适用于标志目标）。 只有对于具有标志目标的模型，才可以请求原始倾向得分，这些得分指示为目标字段指定的似然值为真的结果。除此之外，标准预测及置信值也是如此。调整后的倾向得分不可用。 有关详细信息，请参阅第 34 页码第 3 章中的建模节点分析选项。

追加所有概率。指定是否将输出字段每个类别的概率添加到该节点所处理的每个记录。如果未选中此选项，则仅添加预测类别的概率。例如，对于具有三种类别的名义目标，得分输出针对三种类别的每一种仅包括一列，其中第四列指示任何时候预测类别的概率。例如，如果类别红色、绿色和蓝色的概率分别是 0.6、0.3 和 0.1，则预测类将为红色，其中概率为 0.6。

生成此模型的 SQL。使用数据库中的数据时，SQL 代码可传回到数据库中执行，从而大大提高许多操作的处理速度。有关详细信息，请参阅第 6 章中的SQL 优化中的IBM SPSS Modeler Server 14.2 管理和性能指南。

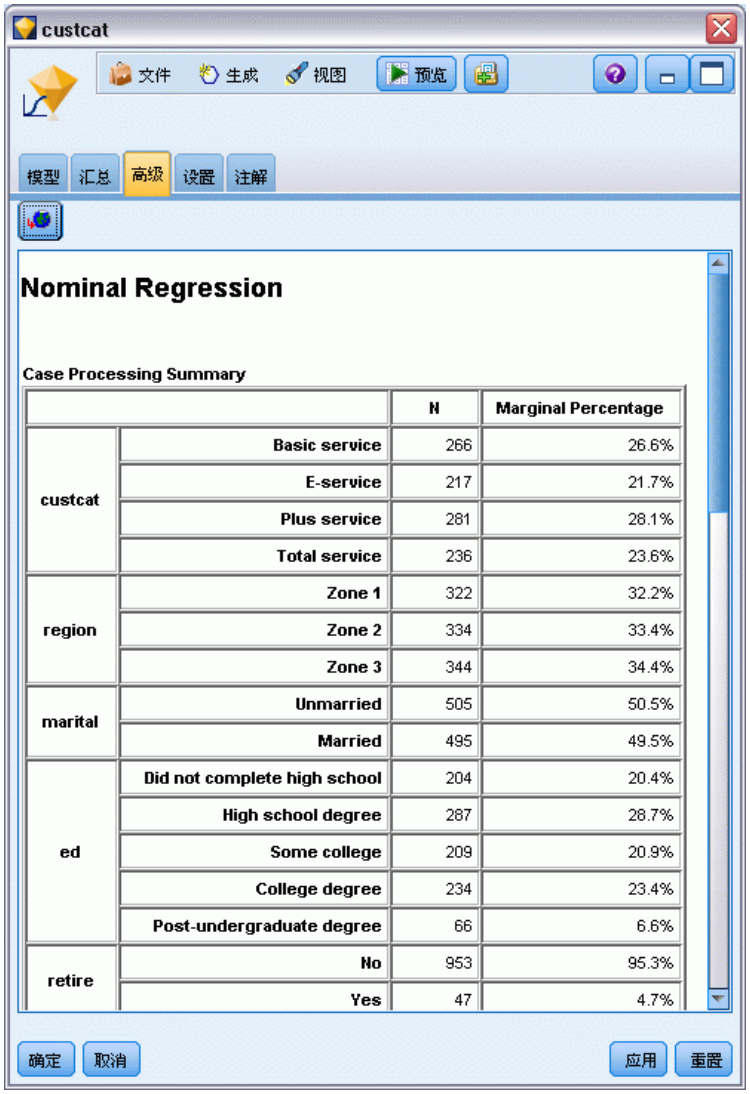
注意：对于多项模型，如果已选中**追加所有概率**，则 SQL 生成不可用；或者，对于具有名义目标的模型，如果选中**计算置信度**，则 SQL 生成不可用。仅仅对具有标志目标的多项模型，支持具有置信度计算的 SQL 生成。SQL 生成不适用于二项模型。

二项模型

对于二项模型，始终启用置信度和概率，并且用于禁用这两个选项的设置不可用。SQL 生成不适用于二项模型。对于二项模型，唯一可以更改的设置是计算原始倾向得分的功能。正如之前针对多项模型的说明，这一项内容适用于只具有标志目标的模型。 有关详细信息，请参阅第 34 页码第 3 章中的建模节点分析选项。

Logistic 模型块高级输出

图片 10-32
Logistic 回归方程式节点“高级”选项卡示例



		N	Marginal Percentage
custcat	Basic service	266	26.6%
	E-service	217	21.7%
	Plus service	281	28.1%
	Total service	236	23.6%
region	Zone 1	322	32.2%
	Zone 2	334	33.4%
	Zone 3	344	34.4%
marital	Unmarried	505	50.5%
	Married	495	49.5%
ed	Did not complete high school	204	20.4%
	High school degree	287	28.7%
	Some college	209	20.9%
	College degree	234	23.4%
	Post-undergraduate degree	66	6.6%
retire	No	953	95.3%
	Yes	47	4.7%

Logistic 回归（也称为**名义回归**）的高级输出将提供有关估计模型及其性能的详细信息。高级输出包含的大部分信息技术含量很高，需要具备 Logistic 回归分析方面的广泛知识才能正确理解该输出。

警告。指明结果中存在的任何警告或潜在问题。

个案处理摘要。列出由模型中的每个符号字段处理和细分的记录数。

步骤概要（可选）。列出在采用自动字段选择的情况下，模型创建的每个步骤添加或删除的效应。

注意：仅针对逐步法、前进法、后退法或后退逐步法显示此选项。

迭代历史（可选）。显示每 n 次迭代的参数估计值的迭代历史（从初始估计值开始），其中 n 是打印间隔值。默认设置是打印每次迭代 (n=1)。

模型拟合信息（多项模型）。显示该模型（最终模型）相对于其中所有参数系数均为 0（仅有截距）的模型的似然比检验。

分类（可选）。显示输出字段预测值和实际值的百分比矩阵。

拟合优度卡方统计量（可选）。显示 Pearson 和似然比卡方统计量。这些统计量可检验模型对训练数据的总体拟合度。

Hosmer-Lemeshow 拟合优度（可选）。显示将观测值分组为风险的十分位数并对每个十分位数内的观测概率与预期概率进行比较的结果。此拟合优度统计量比多项模型中采用的传统拟合优度统计量更为稳健，尤其适用于具有连续协变量的模型和小样本的研究。

伪 R 平方（可选）。显示模型拟合度的 Cox 和 Snell、Nagelkerke 和 McFadden R 平方度量。这些统计量在某些方面与线性回归中的 R 平方统计量类似。

单调性测量（可选）。显示数据中一致对、不一致对和约束对的数量，以及每种对占总对数的百分比。此表中还显示 Somers 的 D、Goodman 和 Kruskal 的 Gamma、Kendall 的 tau-a 以及协调索引 C。

信息标准（可选）。显示 AIC 信息准则和 Schwarz BIC 信息准则。

似然比检验（可选）。显示模型效应系数是否在统计上不等于 0 的统计量检验。有意义的输入字段是输出的显著性水平很低（标示为 Sig.）的输入字段。

参数估计（可选）。显示方程式系数、这些系数的检验、衍生自标示为 Exp(B) 的系数的优势比及其置信区间的估计值。

渐近协方差/相关矩阵（可选）。显示系数估计值的渐近协方差和/或相关性。

已观测到的频数和预测的频数（可选）。对于每个协变量模式，为每个输出字段的值显示已观测到的频数和预测的频数。此表可能很大，对于具有数字输入字段的模型来说尤其如此。如果结果表太大无法应用，则将省略该表，并显示一条警告。

主成份分析/因子节点

因子/主成分分析节点提供了用于降低数据复杂程度的强大数据缩减技术。该技术提供以下两种相似但不同的方法。

- **主成分分析 (PCA)** 可找出能在整个字段集中最好地捕获方差的输入字段的线性组合，其中成分相互正交（垂直）。主成分分析集中关注所有方差，包括共享方差和独有方差。
- **因子分析**尝试找出可解释一组被观测字段中的相关模式的基本概念或**因子**。因子分析只集中关注共享方差。估计模型时不考虑特定字段独有的方差。因子/主成分分析节点提供几种因子分析方法。

这两种方式的目标都是找到有效概括原始字段集中的信息的一小部分导出字段。

要求。主成分分析因子模型中只能使用数值字段。要估计因子分析或主成分分析，需要一个或多个角色设置为输入字段的字段。角色设置为目标、双向或无的字段将被忽略，就像对待非数值字段一样。

强度。因子分析和主成分分析可以在不牺牲太多信息内容的情况下有效降低数据复杂程度。这些技术可帮助您构建更稳健的模型，并实现比原始输入字段更高的执行速度。

主成分分析/因子节点模型选项

图片 10-33
主成分分析（PCA）/因子模型选项卡

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

提取方法。指定要用于数据降维的方法。

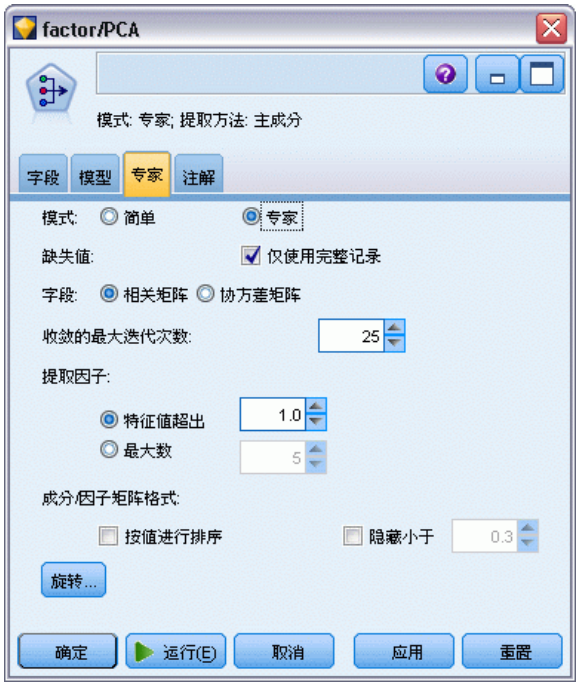
- **主成分。**这是默认方法，将使用主成分分析来找出概括输入字段的成分。
- **未加权最小二乘。**此因子分析方法的工作原理是找出最有能力重现输入字段之间关系（相关）模式的因子集合。
- **广义最小二乘。**此因子分析方法与未加权最小二乘方法类似，区别在于它利用加权降低具有大量独有（非共享）方差的字段的重要程度。
- **最大似然法。**此因子分析方法将产生最有可能生成输入字段中观测到的关系（相关）模式的因子方程式，它以对这些关系的形式的假定为基础。特别是，该方法假定训练数据服从多元正态分布。
- **主轴因子法。**此因子分析方法与主成分方法十分类似，区别在于它只集中关注共享方差。

- **Alpha 因子法。**此因子分析方法将分析中的字段视为潜在输入字段范围内的样本。它会将因子的统计可靠性最大化。
- **映像因子法。**此因子分析方法使用数据估计来隔离通用方差，并找出描述该方差的因子。

主成份分析（PCA）/因子节点专家选项

如果具备因子分析和主成分分析的深入知识，则可以使用专家选项对训练过程进行调整。要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

图片 10-34
主成分分析（PCA）/因子专家选项卡



缺失值。默认情况下，IBM® SPSS® Modeler 将仅使用对于模型中使用的所有字段均具有有效值的记录。（这种方式有时称为缺失值的**成列删除**。）如果有很多缺失数据，您可能会发现这种方式去除的记录过多，剩余记录不足以生成较好的模型。在这类情况下，可以取消选中**仅使用完整记录**选项。然后，SPSS Modeler 将尝试使用尽可能多的信息估计模型，包括其中一些字段具有缺失值的记录。（这种方式有时称为缺失值的**成对删除**。）但在某些情形下，以这种方式使用不完整记录可能会在模型的估计过程中产生计算问题。

字段。指定估计模型时是使用输入字段的相关矩阵（默认设置）还是使用其协方差矩阵。

收敛的最大迭代次数。指定用于估计模型的最大迭代次数。

提取因子。选择要从输入字段中提取的因子数的方法共有两种。

- **特征值超出。**此选项将保留特征值大于指定标准的所有因子或成分。**特征值**用于度量每个因子或成分概括输入字段集合中方差的能力。使用相关矩阵时，模型将保留特征值大于指定值的所有因子或成分。使用协方差矩阵时，标准是指定的乘以平均特征值。该尺度变换使此选项对于两种类型的矩阵具有类似的意义。
- **最大数。**此选项将保留指定数量的因子或成分，按特征值的降序排列。换言之，将保留 n 个最高特征值所对应的因子或成分，其中 n 为指定标准。默认提取标准为五个因子/成分。

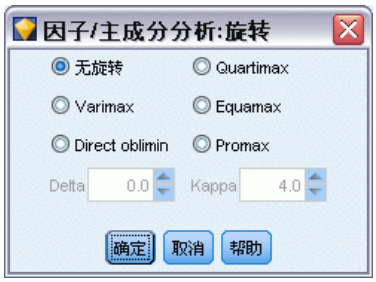
成分/因子矩阵格式。这些选项用于控制因子矩阵（对应主成分分析模型为成分矩阵）的格式。

- **按值进行排序。**如果选中此选项，则会按数字顺序对模型输出中的因子载入进行排序。
- **隐藏小于。**如果选中此选项，则会在矩阵中隐藏低于指定阈值的得分，以便于查看矩阵中的模式。

旋转。这些选项可用于控制模型的旋转方法。 有关详细信息，请参阅第 258 页码主成分分析（PCA）/因子节点旋转选项。

主成分分析（PCA）/因子节点旋转选项

图片 10-35
主成分分析（PCA）/因子旋转选项



许多情况下，对保留的因子集合进行数学旋转可提高其实用性，尤其可以降低其解释难度。选择一种旋转方法：

- **无旋转。**这是默认选项。不使用旋转。
- **Varimax。**这种正交旋转方法可将每个因子上载荷较高的字段的数量降至最低。它简化了因子的解释过程。
- **斜交旋转。**斜交（非正交）旋转的方法。当 Delta 等于 0（默认值）时，解将采用斜交法。delta 负得越厉害，因子的斜交度越低。要覆盖默认的 delta 值 0，请输入小于等于 0.8 的数。
- **Quartimax。**这种正交旋转方法可将解释每个字段所需的因子的数量降至最低。它简化了被观测字段的解释过程。
- **Equamax。**这种旋转方法是 Varimax 法与 Quartimax 法的组合，前者用于简化因子，后者用于简化字段。可将某个因子上载荷较高的字段数量和解释某个字段所需的因子数量降至最低。
- **Promax。**这种斜交旋转方法允许因子之间相关。它计算起来比斜交旋转更快，因此适用于大型数据集。Kappa 用于控制解的倾斜度（因子相关的程度）。

主成分分析（PCA）/因子模型块

主成分分析（PCA）/因子模型块表示由主成分分析（PCA）/因子节点创建的因子分析和主成分分析（PCA）模型。其中包含被训练模型捕获的所有信息，以及有关模型性能和特征的信息。

当您运行包含因子方程式模型的流时，节点会为模型中的每个因子或成分添加一个新字段。新字段的名称来自模型名称，加上前缀 $F- 和后缀 -n，其中 n 是因子或成分的编号。例如，如果模型名为 Factor 且包含三个因子，新字段将命名为 $F-Factor-1、$F-Factor-2 和 $F-Factor-3。

为更好地了解因子模型的编码内容，可以进一步执行一些下游分析。查看因子模型结果的一种实用方法是使用统计量节点查看因子与输入字段之间的相关性。这种方法可显示哪些输入字段对哪些因子的载荷较重，并帮助您发现因子是否具有潜在的意义或解释。 有关详细信息，请参阅第 6 章中的统计量节点中的IBM SPSS Modeler 14.2源、过程和输出节点。

您还可以使用高级输出中提供的信息对因子模型进行评估。要查看高级输出，请单击模型块浏览器的**高级**选项卡。高级输出包含大量详细信息，适合于在因子分析或主成分分析方面具有广泛知识的用户。 有关详细信息，请参阅第 262 页码主成分分析/因子模型块高级输出。

主成分分析/因子模型块方程式

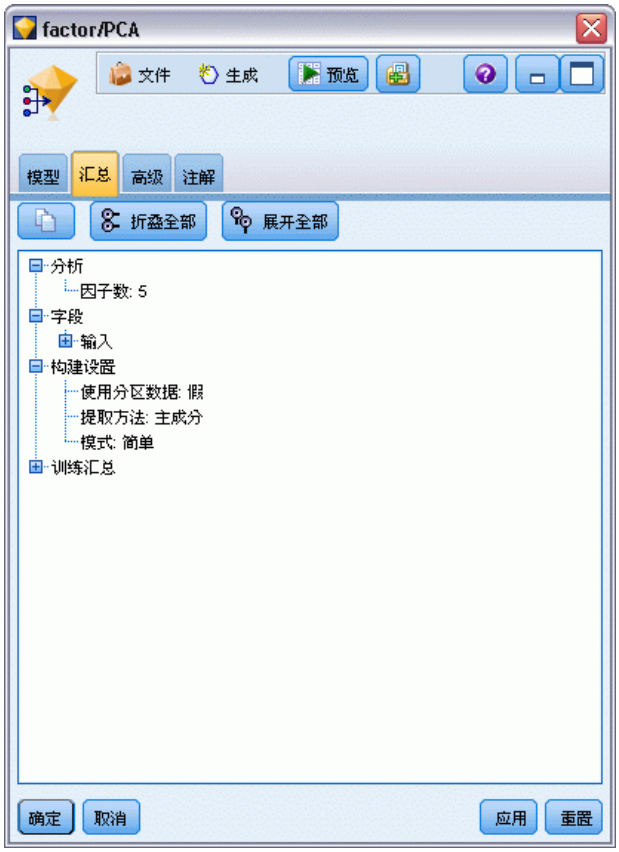
因子模型快的“模型”选项卡显示每个因子的因子得分方程式。因子或成分的得分是通过将每个输入字段值乘以其系数并将结果相加计算得出的。

图片 10-36
主成分分析（PCA）/因子模型块“模型”选项卡

主成分分析/因子模型块概要

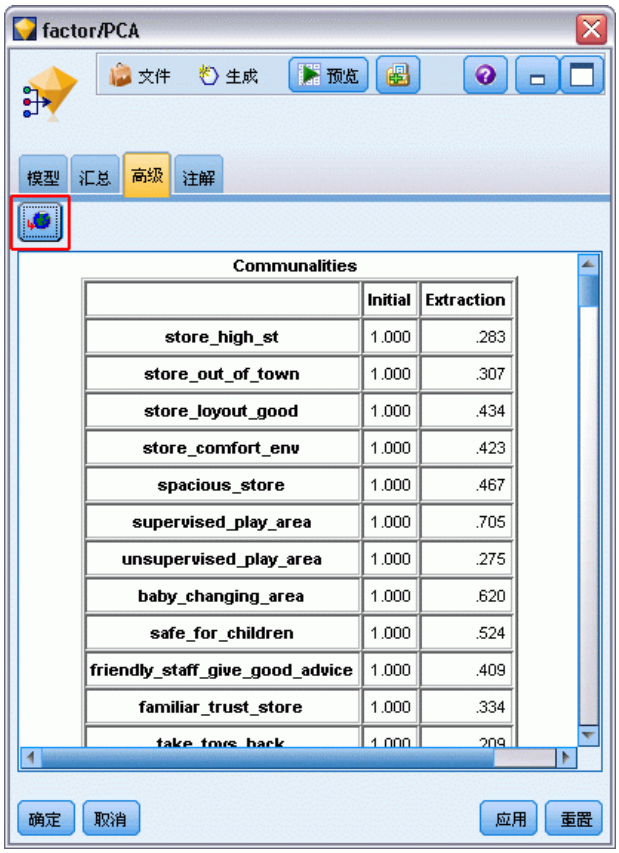
因子模型的“概要”选项卡显示因子/主成分分析模型中保留的因子数，以及有关用于生成模型的字段和设置的其他信息。 有关详细信息，请参阅第 43 页码第 3 章中的浏览模型块。

图片 10-37
因子方程式节点“概要”选项卡示例



主成分分析/因子模型块高级输出

图片 10-38
因子方程式节点“高级”选项卡示例



因子分析的高级输出提供有关所估计模型及其性能的详细信息。高级输出中包含的大部分信息技术含量很高，需要具备因子分析方面的广泛知识才能正确理解该输出。

警告。指明结果中存在的任何警告或潜在问题。

公因子方差。显示因子或成分占每个字段的方差的比例。初始给出整个因子集合（以与输入字段相同数量的因子开始的模型）的公因子方差，提取给出基于保留因子集的公因子方差。

解释的总方差。显示能被模型中因子解释的总方差。初始特征值显示可由整个初始因子集解释的方差。提取载入平方和显示由模型中保留的因子解释的方差。旋转载入平方和显示由旋转因子解释的方差。请注意，对于斜交旋转，旋转载入平方和仅显示载入平方和，而不显示方差百分比。

因子（或成分）矩阵。显示输入字段与非旋转因子之间的相关性。

旋转因子（或成分）矩阵。显示输入字段与正交旋转的旋转因子之间的相关性。

模式矩阵。显示输入字段与斜交旋转的旋转因子之间的偏相关。

结构矩阵。显示输入字段与斜交旋转的旋转因子之间的简单相关性。

因子相关矩阵。显示斜交旋转的因子之间的相关性。

判别式节点

判别式分析用于为组成员关系构建预测模型。该模型由一个判别函数组成（如果有两个以上组，则由一组判别函数组成），判别函数是预测变量的线性组合，提供组之间的最佳判别。这些函数通过组成员关系已知的观测值样本生成；然后，可以将这些函数应用于具有预测变量测量值、但组成员关系未知的新观测值。

示例。 根据使用数据，电信公司可以用判别式分析对用户进行分组。此操作使电信公司可对潜在的用户进行评分，并确定哪些用户最有可能属于最有价值的组。有关详细信息，请参阅第 22 章中的电信客户分类（判别式分析）中的IBM SPSS Modeler 14.2 应用程序 指南。

要求。 您需要一个或多个输入字段，且只需要一个目标字段。目标必须为带有字符串或整数存储的分类字段（测量级别为标志或名义）。（如有必要，可以使用过滤节点或导出节点转换存储类型。有关详细信息，请参阅第 4 章中的使用填充节点进行存储类型转换中的IBM SPSS Modeler 14.2 源、过程和输出节点。）将忽略设置为双向或无的字段。必须对模型中使用的字段的类型完全实例化。

强度。 判别式分析和 Logistic 回归都是适用于分类的模型。然而，“判别式”分析会对输入字段进行更多的假设—例如，假设这些字段为正态分布且为连续，则当满足这些要求时它们能提供更好的结果，尤其是当样本量比较小时。

判别式节点模型选项

图片 10-39
判别式节点对话框：“模型”选项卡

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

方法。 下列选项用于向模型中输入预测变量：

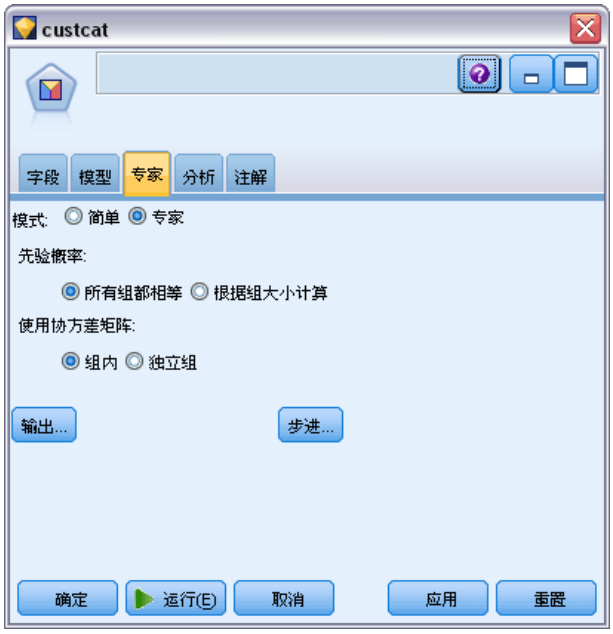
- **按 Enter 键。**这是默认方法，可将所有项直接输入方程式。不能显著增加模型预测能力的项将不被添加。
- **逐步法。** 初始模型可能是最简单的模型，其方程式中不含任何模型项（除常量外）。在每个步骤中，对尚未添加到模型的项进行评估，如果其中的最佳项能够显著增加模型预测能力，则将该项添加到模型中。

注意：逐步法对训练数据有强烈地过度拟合的趋势。当使用这些方法时，用保留测试样本或新数据对结果模型的有效性进行验证尤其重要。

判别式节点专家选项

如果对判别式分析有详尽了解，可用专家选项调整训练过程。要访问专家选项，请在“专家”选项卡中将**模式**设置为**专家**。

图片 10-40
判别式节点对话框：“专家”选项



先验概率此选项决定是否根据已有的组成员知识调整分类系数。

- **所有组均相等。**假定所有组的先验概率均相等；此选项对系数没有影响。
- **根据组大小计算。**样本中观测到的组大小决定组成员的先验概率。例如，如果分析中包括的观测值有 50% 属于第一组，25% 属于第二组，25% 属于第三组，则会调整分类系数，以提高第一组成员相对于其他两组成员的似然性。

使用协方差矩阵。您可以选择使用组内协方差矩阵或类协方差矩阵对观测值进行分类。

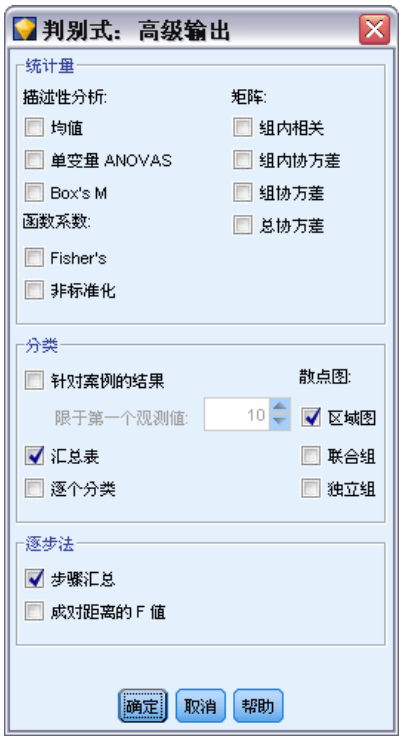
- **组内.** 汇聚的组内协方差矩阵用来对个案分类。
- **独立组.** 分组协方差矩阵用于分类。由于分类基于判别函数（而非基于原始变量），因此该选项并不总是等同于二次判别。

输出。通过这些选项，可以请求将显示在由节点构建的模型块的高级输出中的附加统计量。 有关详细信息，请参阅第 265 页码判别式节点输出选项。

步进。 通过这些选项，可以使用逐步评估法对添加和删除字段的标准进行控制。（如果已选择进入法，该按钮将处于禁用状态。） 有关详细信息，请参阅第 267 页码判别式节点步进选项。

判别式节点输出选项

图片 10-41
判别式节点高级输出选项



选择要在 Logistic 回归模型块的高级输出中显示的可选输出。要查看高级输出，请浏览模型块并单击高级选项卡。 有关详细信息，请参阅第 268 页码判别式模型块高级输出。

描述性统计量。可用的选项有均值（包括标准差）、单变量 ANOVA 和 Box’s M 检验。

- **均值.** 显示自变量的总均值、组均值和标准差。

- **单变量 ANOVA.** 每个自变量的组均值的等同性执行单因子方差分析检验。
- **Box's M.** 组协方差矩阵的等同性检验。对于足够大的样本，不显著的 p 值表示断定矩阵不同的证据不足。该检验对于偏离多变量正态性很敏感。

函数系数。可用的选项有 Fisher 分类系数和非标准化系数。

- **Fisher.** 显示可以直接用于分类的 Fisher 分类函数系数。为每个组获得一组单独的分类函数系数，将一个个案分配给该组，该个案对此组具有最大判别分数（分类函数值）。
- **未标准化.** 显示未标准化的判别函数系数。

矩阵。自变量系数的可用矩阵为类内相关矩阵、类内协方差矩阵、类协方差矩阵和总协方差矩阵。

- **组内相关.** 显示汇聚的组内相关矩阵，获取该矩阵的方法是在计算相关性之前，求得所有组的单个协方差矩阵的平均值。
- **类内协方差.** 显示汇聚的组内协方差矩阵，该矩阵与总协方差矩阵可能不同。获取该矩阵的方法是，求得所有组的单个协方差矩阵的平均值。
- **组协方差.** 显示每个组的分离协方差矩阵。
- **总协方差.** 显示来自所有个案的协方差矩阵，就好像它们来自一个样本一样。

分类。 下列输出属于分类结果。

- **个案结果.** 为每个个案显示实际组的代码、预测组、后验概率和判别得分。
- **摘要表.** 基于判别分析，正确地和不正确地指定给每个组的个案数。有时称为“混乱矩阵”。
- **不考虑该个案时的分类.** 分析中的每个个案由除该个案之外的所有个案生成的函数来进行分类。这也称为“U 方法”。
- **区域图.** 用于基于函数值将个案分类到组的边界图。其个数对应于个案分类到的组数。每个组的均值在其边界内用一个星号表示。如果只有一个判别函数，则该图不会显示。
- **合并组.** 创建前两个判别函数值的所有组散点图。如果只有一个函数，则转而显示一个直方图。
- **分组.** 创建前两个判别函数值的分组散点图。如果只有一个函数，则转而显示直方图。

逐步法。步骤汇总显示执行每个步骤后所有变量的统计量；**成对距离的 F 值**显示每两个组中成对 F 比的矩阵。F 比可用于组之间马氏距离的显著性检验。

判别式节点步进选项

图片 10-42
判别式节点“逐步法”选项

方法。选择要用于输入或删除新变量的统计量。可用替代选项有 Wilks 的 lambda、无法解释的方差、马氏距离、最小 F 比以及 Rao 的 V。使用 Rao 的 V 时，可以为要输入的变量指定 V 的最小增量。

- **Wilk 的 lambda.** 一种用于逐步判别分析的变量选择方法，它基于变量能在多大程度上降低 Wilks 的 lambda 来选择要输入到方程中的变量。在每一步，均是输入能使总体 Wilks 的 lambda 最小的变量。
- **无法解释的方差.** 在每一步中输入使组间未解释变动合计最小的变量。
- **马氏距离.** 自变量上个案的值与所有个案的平均值相异程度的测量。大的 Mahalanobis 距离表示个案在一个或多个自变量上具有极值。
- **最小 F 比.** 一种逐步分析中的变量选择方法，它基于使从组间 Mahalanobis 距离计算得到的 F 比最大。
- **Rao 的 V.** 组均值之间的差分的测量。也称为 Lawley-Hotelling 轨迹。在每一步，能使 Rao 的 V 增加最大的变量被选进来。选择此选项之后，请输入要进入分析，变量必须具有的最小值。

标准。可用替代选项为**使用 F 值**和**使用 F 的概率**。输入用于输入和删除变量的值。

- **使用 F 值.** 如果变量的 F 值大于“输入”值，则该变量输入模型，如果 F 值小于“剔除”值，则该变量从模型中移去。“输入”值必须大于“剔除”值，且两者均必须为正数。要将更多的变量选入到模型中，请降低“输入”值。要将更多的变量从模型中移去，请增大“剔除”值。
- **使用 F 的概率.** 如果变量的 F 值的显著性水平小于“输入”值，则将该变量选入到模型中，如果该显著性水平大于“剔除”值，则将该变量从模型中移去。“输入”值必须小于“剔除”值，且两者均必须为正数。要将更多的变量选入到模型中，请增加“输入”值。要将更多的变量从模型中移去，请降低“剔除”值。

判别式模型块

“判别式”模型块表示由判别式节点估计的方程式。这些方程式包含由判别式模型所捕获的所有信息及有关模型结构和性能的信息。

当运行包含“判别式”模型块的流时，该节点可添加包含模型预测和关联概率的两个新字段。新字段的名称来自正在预测的输出字段的名称，前缀为 $D- 表示预测的类别，前缀为 $DP- 表示关联的概率。例如，对于名称为 colorpref 的输出字段，新字段的名称应是 $D-colorpref 和 $DP-colorpref。

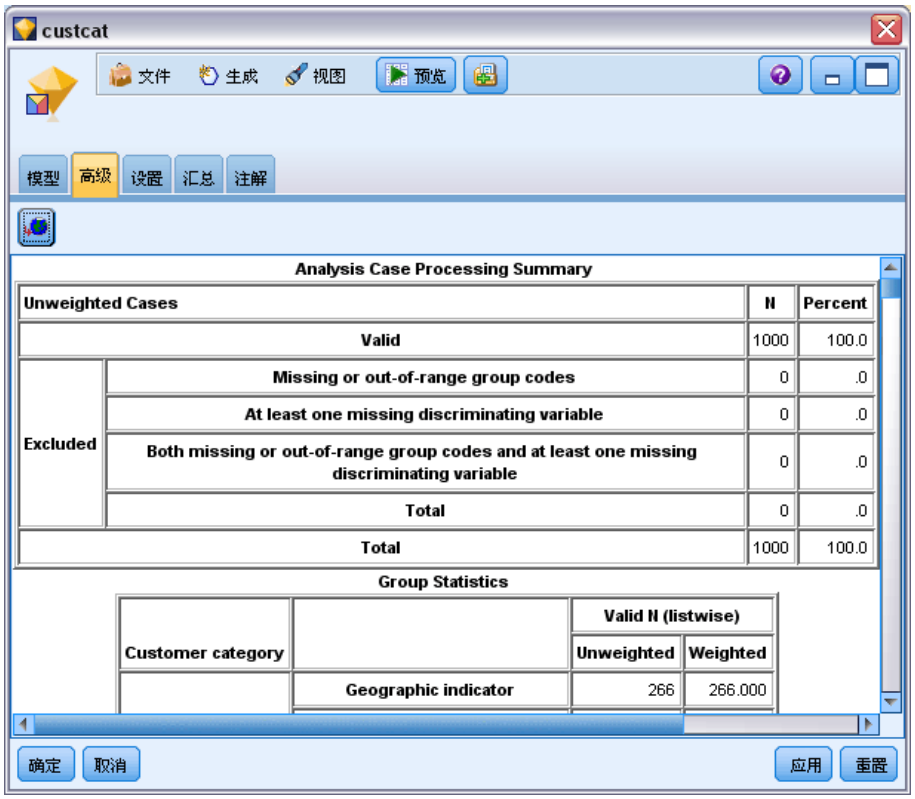
生成过滤节点。使用“生成”菜单可以创建新的过滤节点，用于根据模型结果传递输入字段。

预测变量重要性

另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。注意，只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

判别式模型块高级输出

图片 10-43
判别式模型块：“高级”选项卡



判别式分析的高级输出给出了有关估计模型及其性能的详细信息。在高级输出中包含的多数信息具有很强的技术性，需要具有广泛的判别式分析方面的知识才能够对此输出作出正确地解释。 有关详细信息，请参阅第 265 页码判别式节点输出选项。

判别式模型块设置

通过判别式模型块中的“设置”选项卡，您可以在对模型进行评分时获取倾向得分。此选项卡在只带有标志目标的模型中提供，并且仅在已将模型块添加到流中后可用。

图片 10-44
判别式模型块，标志目标的“设置”选项卡

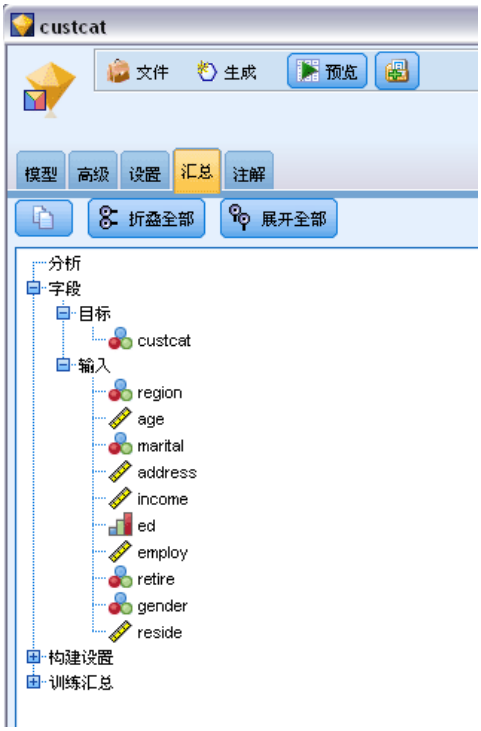
计算原始的倾向得分。对于含标志目标（返回?是?或?否?预测）的模型，您可以请求倾向得分，这些得分指示为目标字段指定结果为真的可能性。除了这些得分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向得分。原始倾向得分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该得分可能会过度优化。调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。此选项要求在流中定义分区字段并且在生成模型之前在建模节点中启用调整的倾向得分。

判别式模型块汇总

判别式模型块的“汇总”选项卡显示了用于生成模型的字段和设置。此外，如果已执行附加到该建模节点的分析节点，则还会在此部分显示该分析中的信息。有关详细信息，请参阅第 6 章中的分析节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。有关使用模型浏览器的一般信息，请参阅浏览模型块第 43 页码。

图片 10-45
判别式模型块：“汇总”选项卡



GenLin 节点

广义线性模型扩展了一般线性模型，使因变量通过指定的链接函数与因子和协变量线性相关。而且，该模型还允许因变量为非正态分布。它涵盖了广泛使用的统计模型，如用于正态分布响应的线性回归、用于二进制数据的 logistic 模型、用于计数数据的对数线性模型、用于区间删失生存数据的互补重对数模型以及使用其非常通用的模型公式的其他许多统计模型。

示例。 运输公司可以使用广义线性模型，对在不同期间建造的一些轮船类型的损坏统计采用泊松回归，其结果模型可帮助确定哪些轮船类型最容易损坏。有关详细信息，请参阅第 24 章中的使用泊松回归来分析船只损坏率（广义线性模型）中的IBM SPSS Modeler 14.2 应用程序 指南。

汽车保险公司可以使用广义线性模型，对汽车损坏理赔采用 gamma 回归，其结果模型可帮助确定对理赔额度贡献最大的因素。有关详细信息，请参阅第 25 章中的将 Gamma 回归拟合至汽车保险理赔（广义线性模型）中的IBM SPSS Modeler 14.2 应用程序 指南。

医疗研究人员可以使用广义线性模型，对间隔检查生存数据采用互补双对数回归，以预测医疗条件再次出现的时间。有关详细信息，请参阅第 23 章中的分析区间型删失的生存数据（广义线性模型）中的IBM SPSS Modeler 14.2 应用程序 指南。

广义线性模型的工作原理是构建一个方程式，从而使输入字段值与输出字段值关联起来。生成模型后，便可以将其用于为新数据估值。对于每条记录，将计算每种可能输出类别的归属概率。具有最高概率的目标类别将被指定为该记录的预测输出值。

要求。您需要一个或多个输入字段，同时有且仅有一个具有两个或多个类别的目标字段（其测量级别可以为连续或标记）。必须对模型中使用的字段的类型完全实例化。

强度。广义线性模型极为灵活，但选择模型结构的过程并未自动化，因此您需要对数据有一定的了解（这在“黑盒”算法中是不需要的）。

GenLin 节点字段选项

图片 10-46
GenLin 节点对话框，“字段”选项卡

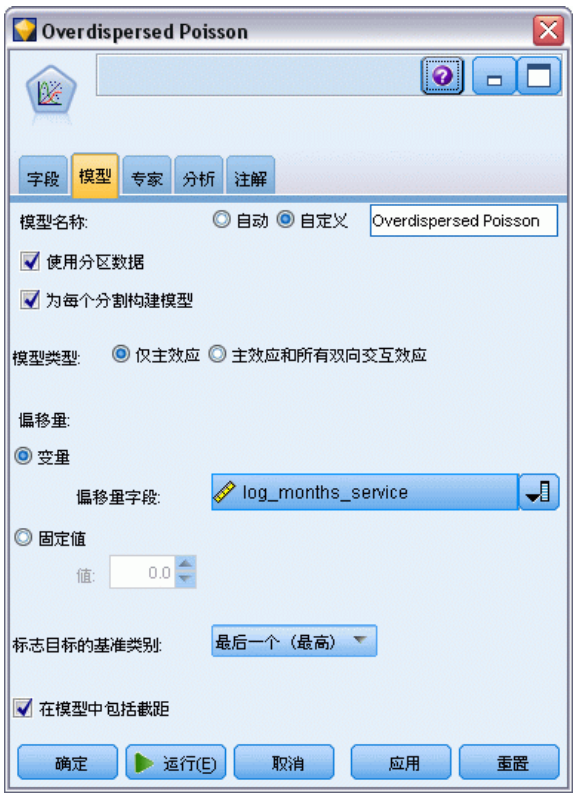
除建模节点的“字段”选项卡通常提供的目标、输入和分区等自定义选项外（请参阅建模节点字段选项第 30 页码），GenLin 节点还提供以下附加功能。

使用加权字段。尺度参数是与响应方差相关的估计模型参数。尺度权重是“已知”值，可能因观察值的不同而异。如果指定了刻度权重变量，则对每个观察值，都会用与响应方差相关的尺度参数除以该尺度权重变量。分析中不会使用尺度权重值小于等于 0 或缺失的记录。

目标字段表示一组试验中发生的事件的数量。如果响应是一组试验中发生的事件的数量，目标字段将包含该事件数量，您可选择包含试验次数的附加变量。或者，如果试验数在所有主体中都相同，则可以使用固定值指定试验。对于每条记录，试验次数应大于或等于事件数量。事件应为非负整数，试验应为正整数。

GenLin 节点模型选项

图片 10-47
GenLin 节点对话框，“模型”选项卡



模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

模型类型。有两个选项用于要构造的模型类型。**仅主效应** 使模型仅分别包括各个输入字段，而不检验输入字段之间的交互效应（乘法效应）。**主效应和所有双向交互**包括所有双向交互以及输入字段主效应。

偏移量。偏移量项是一个“结构”预测变量。它的系数不通过模型估计而假定其值为 1；因此，偏移量的值只与因变量的线性预测变量简单相加。这对于泊松回归模型尤其有用，在这种模型中，每个观测值对于相关事件可以具有不同的揭示级别。例如，为各个驾驶员的事故率建模时，有三年驾驶经验的驾驶员在一次事故中的过错率与有 25 年驾驶经验的驾驶员在一次事故中的过错率存在重大差别。如果将驾驶员经历纳入偏移项，则事故数可以建模为泊松响应。

注意：如果使用变量偏移量，则指定字段不应同时也用作输出。如果需要，可在上游源节点或“类型”节点中将偏移量字段的角色设置为**无**。 有关详细信息，请参阅第 4 章 中的设置字段角色中的IBM SPSS Modeler 14.2 源、过程和输出节点。

标志目标的基准类别。

对于二元响应，您可以为因变量选择参考类别。这样可能会影响某些输出，如参数估计值和保存值，但应该不会改变模型拟合度。例如，如果二元响应的值为 0 和 1：

- 默认情况下，该过程会使得最后一个（值最高的）类别（即 1）成为参考类别。在此种情况下，模型保存的概率会估计出给定观测值采用值 0 的几率，参数估计应解释为与类别 0 的似然相关。
- 如果将第一个（值最低的）类别（即 0）指定为参考类别，则模型保存的概率会估计出给定观测值采用值 1 的几率。
- 如果指定自定义类别而且变量定义了标签，则可以通过从列表中选择值来设置参考类别。这样在指定模型过程中可能比较方便，您不必确切记得某个特定变量是如何编码的。

在模型中包含截距。模型中通常包含截距。如果您可以假设数据穿过原点，则可以排除截距。

GenLin 节点专家选项

如果具备广义线性模型的深入知识，则可以使用专家选项对训练过程进行微调。要访问专家选项，请在“专家”选项卡中将**模式**设置为**专家**。

图片 10-48
Genlin 节点对话框，“专家”选项卡

目标字段分布和连接函数

分布。

此选项指定因变量的分布。能够指定非正态分布和非恒等链接函数是广义线性模型相对于一般线性模型的重大改进。存在许多可能的分布-链接函数组合，其中有若干组合可适用于任何给定的数据集，因此您可以根据先验理论考虑的事项或哪个组合看起来拟合得最好来指导您的选择。

- **二项。**此分布仅适用于表示二元响应或事件数的变量。
- **Gamma。**此分布适用于具有向更大正值偏斜的正尺度值的变量。如果某个数据值小于等于 0 或者缺失，则在分析中不使用相应的观测值。
- **逆高斯。**此分布适用于具有向更大正值偏斜的正尺度值的变量。如果某个数据值小于等于 0 或者缺失，则在分析中不使用相应的观测值。

- **负二项。**此分布可视为观测到 k 次成功所需进行的试验次数，它适用于具有非负整数值的变量。如果某个数据值为非整数、小于 0 或者缺失，则在分析中不使用相应的观测值。负二项式分布辅助参数的固定值可以是大于等于 0 的任何值。辅助参数设置为 0 时，使用此分布相当于使用泊松分布。
- **正态。**此分布适用于其值围绕中心值（均值）呈对称钟形分布的尺度变量。因变量必须是数值型。
- **泊松。**此分布可视为某个相关事件在某个固定时段的发生次数，它适用于具有非负整数值的变量。如果某个数据值为非整数、小于 0 或者缺失，则在分析中不使用相应的观测值。
- **Tweedie。**该分布适用于可以由泊松分布和伽玛分布混合表示的变量；从某种意义上说该分布属于“混合型”分布，是因为该分布同时具有连续分布（采用非负实数值）和离散分布（正概率群位于单个值 0 上）的属性。因变量必须为数字，且数据值大于等于 0。如果数据值小于 0 或者缺失，则在分析中将不使用相应的观测值。Tweedie 分布参数的固定值可以为大于 1 且小于 2 的任意数字。
- **多项式。**此分布适用于表示顺序响应的变量。因变量可以为数字或字符串，且至少有两个不同的有效数据值。

连接函数。

链接函数是允许进行模型估计的因变量的变换。可用链接函数如下：

- **恒等式。**f(x)=x。不对因变量进行变换。此链接函数可用于任何分布。
- **互补重对数。**f(x)=log(−log(1−x))。此函数仅适用于二项分布。
- **累积 Cauchit。**f(x) = tan(π (x − 0.5))，应用于响应的每一个类别的累积概率。此函数仅适用于多项分布。
- **累积互补双对数。**f(x)=ln(−ln(1−x))，适用于每个响应类别的累积概率。此函数仅适用于多项分布。
- **累积分对数。**f(x)=ln(x / (1−x))，应用于响应的每一个类别的累积概率。此函数仅适用于多项分布。
- **累积负重对数。**f(x) = −ln(−ln(x))，应用于响应的每一个类别的累积概率。此函数仅适用于多项分布。
- **累积 probit。**$f(x)=\Phi^{-1}(x)$，适用于每个响应类别的累积概率，其中 Φ^{-1} 是逆标准正态累积分布函数。此函数仅适用于多项分布。
- **对数。**f(x)=log(x)。此链接函数可用于任何分布。
- **对数补码。**f(x)=log(1−x)。此函数仅适用于二项分布。
- **分对数。**f(x)=log(x / (1−x))。此函数仅适用于二项分布。
- **负二项。**$f(x)=\log(x / (x+k^{-1}))$，其中 k 是负二项分布的辅助参数。此函数仅适用于负二项分布。
- **负重对数。**f(x)=−log(−log(x))。此函数仅适用于二项分布。
- **优势幂。**$f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$，如果 $\alpha \neq 0$. $f(x)=\log(x)$，如果 $\alpha=0$. α 为必需的数字指定，且必须为实数。此函数仅适用于二项分布。

- **概率值。**f(x)=Φ−1(x)，其中 Φ−1 是逆标准正态累积分布函数。此函数仅适用于二项分布。
- **幂。**f(x)=xα，如果 α ≠ 0. f(x)=log(x)，如果 α=0. α 为必需的数字指定，且必须为实数。此链接函数可用于任何分布。

参数。通过此组中的控件，可以在选中某些分布选项时指定参数值。

- **负二项式的参数。**对于负二项式分布，选择以指定一个值或允许系统提供估计值。
- **Tweedie 参数。**对于 Tweedie 分布，给固定值指定在 1.0 与 2.0 之间的一个数字。

参数估计。通过此组中的控件，可以指定估计方法，以及为参数估计提供初始值。

- **方法。**您可以选择参数估计方法。您可以选择 Newton-Raphson 方法、Fisher 评分方法或混合方法，在混合方法中，首先会执行 Fisher 评分迭代，然后再切换到 Newton-Raphson 方法。如果在该混合方法的 Fisher 评分阶段，在达到最大 Fisher 迭代次数之前实现了收敛，则该算法将继续执行 Newton-Raphson 方法。
- **尺度参数方法。**您可以选择尺度参数估计方法。最大似然法可联合估计尺度参数和模型效应；请注意，如果响应具有负二项式、泊松或二项式分布，则此选项无效。偏差和 Pearson 卡方选项根据这些统计量的值估计尺度参数。另外，您还可以为尺度参数指定固定值。
- **协方差矩阵。**基于模型的估计是 Hessian 矩阵的广义逆负矩阵。健壮性（也称为 Huber/White/sandwich）估计是“改正”的基于模型的估计，即使错误地指定了方差和关联函数，也能提供对协方差的一致估计。

迭代。这些选项可用于控制模型收敛的参数。 有关详细信息，请参阅第 276 页码广义线性模型迭代。

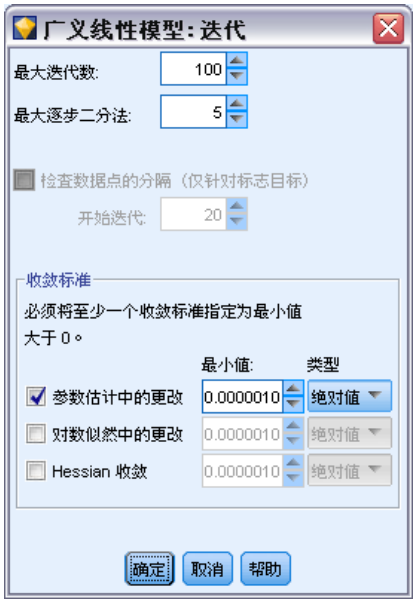
输出。通过这些选项，可以请求将显示在由节点构建的模型块的高级输出中的附加统计量。 有关详细信息，请参阅第 278 页码广义线性模型高级输出。

异常值容差。异常（或非可逆）矩阵具有线性相关列，这样可能会导致估计算法出现严重问题。即使是近似异常的矩阵也可能导致糟糕的结果，因此该过程会将决定因子小于容差的矩阵视为异常矩阵。请指定一个正值。

广义线性模型迭代

您可设置用于对广义线性模型进行估计的收敛参数。

图片 10-49
广义线性模型的迭代选项



迭代。

- **最大迭代次数。**算法将执行的最大迭代次数。指定一个非负整数。
- **最大逐步二分法。**每次迭代时，步长都会缩短一半，直到对数似然增加或达到最大步长二分。请指定一个正整数。
- **检查数据点的分离。**如果选中，该算法将执行检验，以确保参数估计具有唯一值。如果该过程能够生成正确分类每个观测值的模型，则对数据点进行分离。此选项可用于二元格式的二项式响应。

收敛标准。

- **参数收敛。**如果选中，该算法会在参数估计值的绝对或相对变化小于指定值（该值必须为正）的迭代之后停止。
- **对数似然收敛。**如果选中，该算法会在对数似然函数中的绝对或相对变化小于指定值（该值必须为正）的迭代之后停止。
- **Hessian 收敛。**对于绝对指定，如果基于 Hessian 收敛的统计量小于指定的正值，则认为是收敛的。对于相对指定，如果统计量小于指定正值与对数似然绝对值的乘积，则认为是收敛的。

广义线性模型高级输出

图片 10-50
广义线性模型的高级输出选项

选择要在广义线性模型块的高级输出中显示的可选输出。要查看高级输出，请浏览模型块并单击高级选项卡。 有关详细信息，请参阅第 280 页码GenLin 模型块高级输出。

可用输出如下：

- **观测值处理概要。**显示分析以及“相关数据概要”表中所包括和排除的观测值的数量和百分比。
- **描述性统计量。**显示关于因变量、协变量和因子的描述性统计量和摘要信息。
- **模型信息。**显示数据集名称、因变量或事件和试验变量、偏移变量、尺度权重变量、概率分布和链接函数。
- **拟合优度统计量。**显示偏差和定比变换偏差、Pearson 卡方和定比变换 Pearson 卡方、对数似然、AIC 信息准则、有限样本校正 (AICC)、贝叶斯信息准则 (BIC) 和相容 AIC (CAIC)。
- **模型摘要统计量。**显示模型拟合检验，包括模型拟合公用检验的似然比统计量以及每个效应的类型 I 或 III 对比的统计量。
- **参数估计值。**显示参数估计值以及对应的检验统计量和置信区间。除原始参数估计值外，还可以显示取幂的参数估计值。
- **参数估计值的协方差矩阵。**显示估计参数的协方差矩阵。
- **参数估计值的相关矩阵。**显示估计参数的相关矩阵。
- **对比系数 (L) 矩阵。**显示默认效应的对比系数，如果在“EM 均值”选项卡上请求了，则还会显示估计边缘均值的对比系数。

- **一般可估函数。**显示用于生成对比系数 (L) 矩阵的矩阵。
- **迭代历史。**显示参数估计值和对数似然的迭代历史，输出梯度向量和 Hessian 矩阵的最终值。迭代历史表从第 0 次迭代（初始估计值）开始每隔 n次迭代显示参数估计值，其中 n 为输出间隔的值。如果请求显示迭代历史，则无论 n 为多少都会显示最后一次迭代。
- **Lagrange 乘数检验。**显示用于针对标准、伽玛和逆高斯分布评估尺度参数有效性的拉格朗日乘数检验统计量，该参数使用偏差或 Pearson 卡方计算得出，或设置为固定值。对于负二项式分布，它检验固定辅助参数。

模型效应。

- **分析类型。**指定要生成的分析的类型。类型 I 分析一般适用于您有先验理由对模型中的预测变量进行排序的情况，而类型 III 则适用于更一般的情况。Wald 或似然比统计量是根据卡方统计量组中的选择而计算的。
- **置信区间。** 请指定大于 50 小于 100 的置信水平。Wald 区间基于这样的假设，参数为渐近正态分布；剖面似然置信区间更准确，但可能计算花费高昂。剖面似然置信区间的误差等级是用于停止计算置信区间的迭代算法的条件。
- **对数似然函数。**此选项控制对数似然函数的显示格式。完整的函数包括一个相对于参数估计值来说不变的附加项；它对参数估计没有影响，在某些软件产品中不显示。

GenLin 模型块

GenLin 模型块表示由 GenLin 节点估计的方程式。这些方程式包含由模型所捕获的所有信息及有关模型结构和性能的信息。

当您运行包含 GenLin 模型块的流时，该节点会添加一些新字段，这些字段的内容取决于目标字段的性质：

- **标志目标。**添加的字段包含预测类别和相关概率，以及每个类别的概率。前两个新字段的名称派生自所预测的输出字段的名称，前缀 $G- 表示预测类别，前缀 $GP- 表示相关概率。例如，对于名为 default 的输出字段，新字段将命名为 $G-default 和 $GP-default。后两个附加字段基于输出字段的值进行命名，带有前缀 $GP-。例如，如果 default 的有效值为 Yes 和 No，则新字段会以 $GP-Yes 和 $GP-No 命名。
- **连续目标。**添加的字段包含预测均值和标准误。
- **连续目标，表示一系列试验中发生的事件的数量。**添加的字段包含预测均值和标准误。

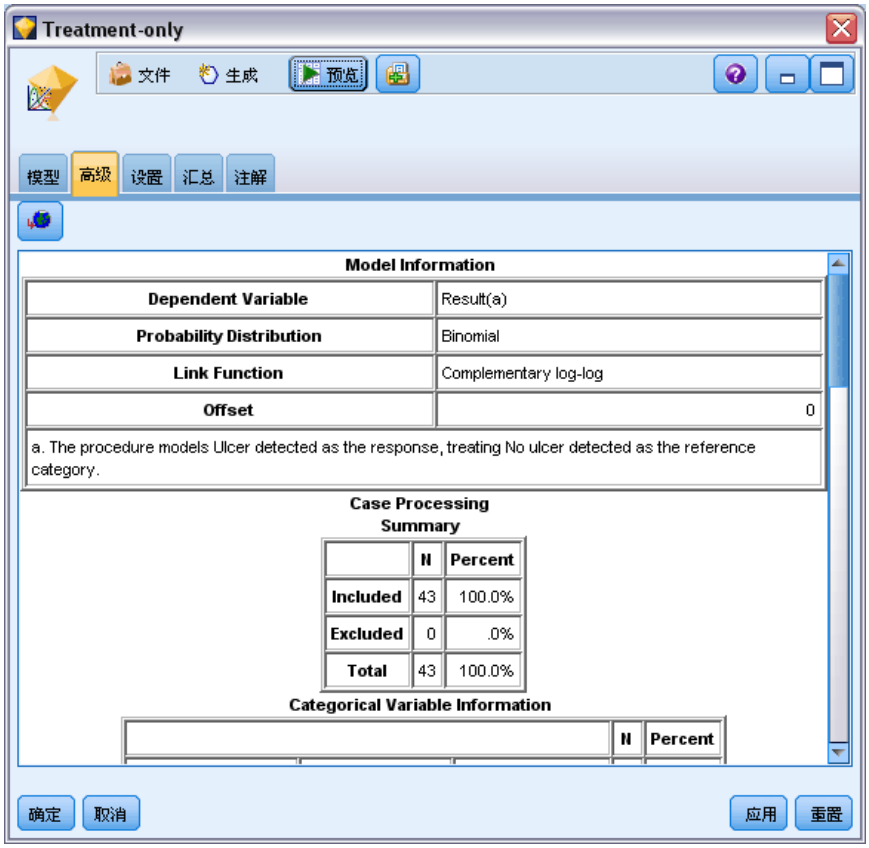
生成过滤节点。使用“生成”菜单可以创建新的过滤节点，用于根据模型结果传递输入字段。

预测变量重要性

另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。注意，只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

GenLin 模型块高级输出

图片 10-51
GenLin 模型块，“高级”选项卡



广义线性模型的高级输出可提供有关估计模型及其性能的详细信息。高级输出中包含的大部分信息的技术性含量都很高，需要进行此类分析所需的丰富知识才能够对此输出作出正确解释。 有关详细信息，请参阅第 278 页码广义线性模型高级输出。

GenLin 模型块设置

使用 GenLin 模型块的“设置”选项卡，您可以在对模型进行评分时获取倾向得分。此选项卡在只带有标志目标的模型中提供，并且仅在已将模型块添加到流中后可用。

图片 10-52
GenLin 模型块，标志目标的“设置”选项卡

计算原始的倾向得分。对于含标志目标（返回?是?或?否?预测）的模型，您可以请求倾向得分，这些得分指示为目标字段指定结果为真的可能性。除了这些得分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向得分。原始倾向得分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该得分可能会过度优化。调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。此选项要求在流中定义分区字段并且在生成模型之前在建模节点中启用调整的倾向得分。

GenLin 模型块汇总

GenLin 模型块的“汇总”选项卡显示了用于生成模型的字段和设置。此外，如果已执行附加到该建模节点的分析节点，则还会在此部分显示该分析中的信息。有关详细信息，请参阅第 6 章中的分析节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。有关使用模型浏览器的一般信息，请参阅浏览模型块第 43 页码。

图片 10-53
GenLin 模型块，“汇总”选项卡

Cox 节点

Cox 回归可为时间事件数据构建预测模型。该模型会生成一个生存函数，该函数可预测在给定时间 t 内对于所给定的预测变量值相关事件的发生概率。从观测主项对生存函数的形状以及预测变量的回归系数进行评估；然后可以将该模型应用于具有预测变量测量的新案例中。请注意，已删失主项（即观测期没有经历相关事件的主项）的相关信息对于评估模型十分有用。

示例。作为减少客户流失计划的一部分，电信公司对建模“流失时间”很感兴趣，以便确定客户快速切换到其他服务的相关因素。为此，随机选取了一些客户样本，和他们作为客户所花费的时间（无论他们是否仍为活动客户）以及从数据库中抽取的各种人口统计字段。有关详细信息，请参阅第 27 章中的将 Cox 回归用于客户流失时间模型中的 IBM SPSS Modeler 14.2 应用程序 指南。

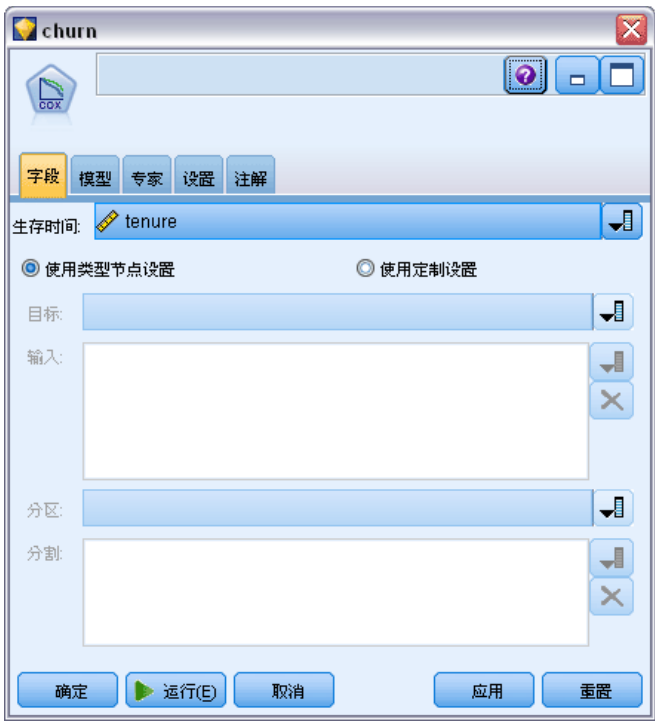
要求。您需要一个或多个输入字段，只需一个目标字段，且必须在 Cox 节点中指定生存时间字段。应对目标字段进行编码，使得“false”值表示生存时间，“true”值表示所关注事件已发生；目标字段的测量级别必须为标志，且带有字符串或整数存储。（如有必要，可以使用过滤节点或导出节点转换存储类型。有关详细信息，请参阅第 4 章中的使用填充节点进行存储类型转换中的IBM SPSS Modeler 14.2 源、过程和输出节点。）将忽略设置为双向或无的字段。必须对模型中使用的字段的类型完全实例化。生存时间可以是任意数字字段。

日期 & 时间。“日期和时间”字段不能直接用于定义生存时间；如果有“日期和时间”字段，则应根据输入研究的日期和观测日期之间的差值，使用这些字段创建包含生存时间的字段。 有关详细信息，请参阅第 7 章中的时间和日期的处理中的IBM SPSS Modeler 14.2 用户 指南。

Kaplan-Meier 分析。可以在没有输入字段的情况下执行 Cox 回归。这等效于 Kaplan-Meier 分析。

Cox 节点字段选项

图片 10-54
Cox 节点对话框，“字段”选项卡



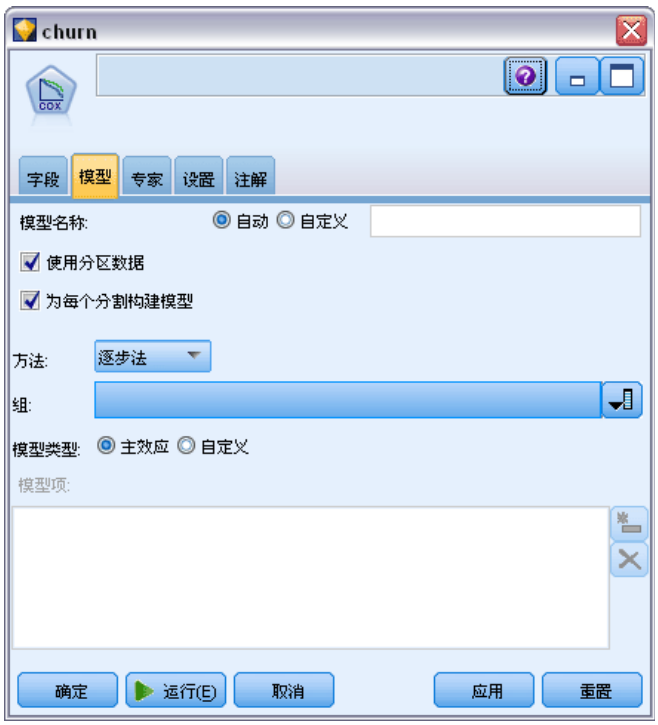
生存时间。选择数值字段（测量级别为连续的字段）以使节点可执行。生存时间表示所预测记录的寿命。例如，当模型化客户流失时间时，它可能是记录客户在组织内的时间长度的字段。客户加入公司或离开公司的日期不会影响该模型；只有客户工龄的持续时间与其相关。’

生存时间为无单位的持续时间。您必须确保输入字段与生存时间相匹配。例如，在按月测量流失的研究中，您可将月销售量而非年销售量用作输入。如果您的数据具有开始日期和结束日期而不是持续时间，您必须在 Cox 代码上游将这些日期重新编码为持续时间。

此对话框中的剩余字段是整个 IBM® SPSS® Modeler 中通用的标准字段。 有关详细信息，请参阅第 30 页码第 3 章中的建模节点字段选项。

Cox 节点模型选项

图片 10-55
Cox 节点对话框，“模型”选项卡



模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

方法。 下列选项用于向模型中输入预测变量：

- **按 Enter 键。**这是默认方法，可将所有项直接输入模型。构建模型时不进行字段选择。
- **逐步法。**顾名思义，逐步字段选择法就是分步构建模型。初始模型可能是最简单的模型，其模型中不含任何模型项（除常量外）。在每个步骤中，对尚未添加到模型的项进行评估，如果其中的最佳项能够显著增加模型预测能力，则将该项添加到模型中。此外，还会重新评估当前包含在模型中的项，以确定能否在不对模型功能造成重大减损的情况下删除其中任何项。如果可以，则会将其删除。然后重复此过程，添加并/或删除其他项。当无法再添加任何项来改进模型、且无法再删除任何项而不对模型功能造成减损时，最终模型便已生成。
- **后退逐步法。**从本质上说，后退逐步法与逐步法是相反的。采用这种方法时，初始模型将包含作为预测变量的所有项。每个步骤会评估模型中的项，并且将可以删除而不对模型功能造成重大减损的项删除。此外，还会对先前删除的项进行重新评

估，以确定其中的最佳项是否对模型的预测功能起到显著作用。如果是，则会将其重新添加到模型中。当无法再删除任何项而不对模型功能造成重大减损、且无法再添加任何项以改进模型时，最终模型便已生成。

注意：自动方法（包括逐步法和后退逐步法）是适应性强的学习方法，并且特别容易过度拟合训练数据。使用这些方法时，用新数据或使用分区节点创建的保留测试样本对结果模型的有效性进行验证尤为重要。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

组。指定组字段会导致节点为每个字段类别计算单独的模型。该字段可以是存储类型为字符串或整数的分类字段（标志或名义）。

模型类型。用于定义模型中的项的选项共有两种。**主效应**模型仅分别包括各个输入字段，而不检验输入字段之间的交互效应（乘法效应）。**自定义**模型仅包括您指定的项（主效应和交互效应）。选择此选项时，应使用“模型项”列表在模型中添加或删除项。

模型项。构建自定义模型时，将需要明确指定模型中的项。此列表显示了模型项的当前集合。“模型项”列表右边的按钮可用于添加或删除模型项。

- 要将项添加到模型中，请单击添加新的模型项按钮。
- 要删除项，请选定所需项，然后单击删除选定模型项按钮。

将项添加到 Cox 回归模型

在请求自定义模型时，可以通过单击“模型”选项卡中的添加新的模型项按钮将各项添加到模型中。此时将打开一个新的对话框，您可在其中指定项。

图片 10-56
“新建项”对话框

要添加的项类型。有几种将项添加到模型的方法，具体取决于在“可用字段”列表中对输入字段的选择。

- **单向交互效应。**插入表示所有选定字段的交互效应的项。
- **主效应。**针对每个选定的输入字段插入一个主效应项（该字段本身）。
- **所有双向交互效应。**针对每个可能的选定输入字段对插入一个双向交互效应项（输入字段的积）。例如，如果已在“可用字段”列表中选定输入字段 A、B 和 C，此方法将插入项 A * B、A * C 和 B * C。
- **所有三向交互效应。**针对每个可能的选定输入字段组合（一次取三个）插入一个三向交互效应项（输入字段的积）。例如，如果已在“可用字段”列表中选定输入字段 A、B、C 和 D，此方法将插入项 A * B * C、A * B * D、A * C * D 和 B * C * D。
- **所有四向交互效应。**针对每个可能的选定输入字段组合（一次取四个）插入一个四向交互效应项（输入字段的积）。例如，如果已在“可用字段”列表中选定输入字段 A、B、C、D 和 E，此方法将插入项 A * B * C * D、A * B * C * E、A * B * D * E、A * C * D * E 和 B * C * D * E。

可用字段。列出要用于构造模型项的可用输入字段。请注意，列表中可能包含非法输入字段，因此务必确保所有的模型项都只包含输入字段。

预览。根据上述所选字段和项类型，显示单击**插入**时将添加到模型中的项。

插入。将项插入模型（根据当前选择的字段和项类型）并关闭对话框。

Cox 节点专家选项

图片 10-57
Cox 节点对话框，“专家”选项卡

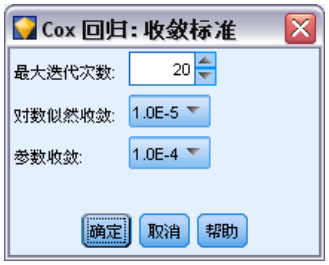
收敛。这些选项可用于控制模型收敛的参数。当您执行模型时，收敛设置将控制重复运行不同参数以观察其拟合程度的次数。参数的尝试次数越多，结果将越接近（即，结果将会收敛）。 有关详细信息，请参阅第 287 页码Cox 节点收敛标准。

输出。通过这些选项，可以请求将显示在由节点构建的生成模型的高级输出中的附加统计量和散点图（包括生存曲线）。有关详细信息，请参阅第 287 页码Cox 节点高级输出选项。

步进。 通过这些选项，可以使用逐步评估法对添加和删除字段的标准进行控制。（如果已选择进入法，该按钮将处于禁用状态。） 有关详细信息，请参阅第 288 页码Cox 节点步进标准。

Cox 节点收敛标准

图片 10-58
“Cox 回归收敛标准”对话框



最大迭代次数。允许您指定模型的最大迭代次数，用于控制过程求解的时间。

对数似然估计收敛。如果对数似然的相对变化小于此值，迭代将停止。如果值为 0，则不使用该标准。

参数收敛。如果参数的绝对变化或相对变化小于此值，迭代将停止。如果值为 0，则不使用该标准。

Cox 节点高级输出选项

图片 10-59
“Cox 回归高级输出”对话框

统计量。您可以获得模型参数的统计量，包括 exp(B) 的置信区间和估计值的相关性。您可以在每一步或者仅在最后一步请求这些统计量。

显示基线函数。允许您显示协变量均值下的基线风险函数和累积生存。

图

图有助于评估估计的模型和解释结果。您可以对生存函数、危险函数、负对数累积生存函数的对数和 1 减去生存函数绘图。

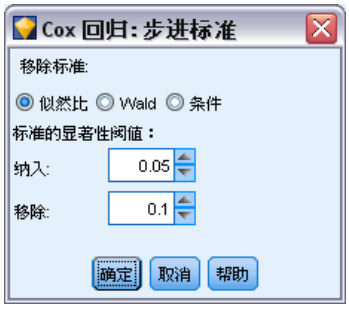
- **生存函数.** 在线性刻度上显示累积生存函数。
- **危险函数(H).** 在线性刻度上显示累积风险函数。
- **对数减对数。**在将 ln(-ln) 转换应用于估计值后显示累积生存估计值。
- **1 减去生存函数(M).** 以线性尺度绘制 1 减生存函数。

为每个值绘制单独线。此选项仅适用于分类字段。

用于散点图的值。由于这些函数取决于预测变量的值，因此您必须使用预测变量的常数值绘制函数与时间的比值。默认情况下，将使用各个预测变量的平均数作为常数值，但您可以使用网格为散点图输入自己的值。对于分类输入，使用指示符编码，因此每个类别都具有回归系数（最后一个类别除外）。因此，分类输入具有每个指示符对比度的平均值，等于类别中对应于指示符对比度的观测值比例。

Cox 节点步进标准

图片 10-60
“Cox 回归步进标准”对话框



移除标准。选择似然比可得到更稳健的模型。要缩短构建模型所需的时间，可以尝试选择 Wald。还有附加选项条件，此选项提供以基于条件参数估计值的似然比统计量的概率为依据的移除检验。

标准的显著性阈值。使用此选项可基于每个字段关联的统计概率（p 值）指定选择标准。仅当关联的 p 值小于纳入标准值时，才会将字段添加到模型中；仅当 p 值大于剔出标准值时，才会将字段删除。纳入标准值必须小于剔出标准值。

Cox 节点设置选项

图片 10-61
Cox 节点对话框，“设置”选项卡

以未来时间预测生存时间。指定一个或多个未来时间。即在未发生终端事件的情况下，无论每个观测值是否可能至少在此时间长度（从现在开始）内生存，都将在每个时间值为每条记录预测生存时间，一个时间值对应一个预测值。请注意，生存时间为目标字段的“false”值。

- **规则区间。**生存时间值从指定的**时间区间**和**要对其评分的时段数**生成。例如，如果请求 3 个时段，时间区间为 2，则对未来时间的生存时间将为 2、4 和 6。以相同时间值评估每条记录。
- **时间字段。**在所选的时间字段中为每条记录提供生存时间（生成一个预测字段），因此可以在不同的时间评估各条记录。

过去生存时间。将目前为止记录的生存时间指定为一个字段—例如，将现有客户的保有期作为一个字段。在未来时间对生存的似然进行评分取决于过去生存时间。

注意：未来和过去生存时间的值必须在用于训练模型的数据中的生存时间范围内。时间超出此范围的记录将标记为空。

追加所有概率。指定是否将输出字段每个类别的概率添加到该节点所处理的每个记录。如果未选中此选项，则仅添加预测类别的概率。为每个未来时间计算概率。

计算累积风险函数指定是否将累积风险的值添加到每条记录中。为每个未来时间计算累积风险。

Cox 模型块

Cox 回归模型表示由 Cox 节点所估计的方程式。这些方程式包含由模型所捕获的所有信息及有关模型结构和性能的信息。

运行包含生成的 Cox 回归模型的流时，该节点可添加包含模型预测和关联概率在内的两个新字段。新字段的名称派生自所预测的输出字段的名称，前缀 $C- 表示预测类别，前缀 $CP- 表示相关概率。后缀为未来时间区间的数量或定义时间区间的时间字段的名称。例如，对于名为 churn 的输出字段，以及以规则区间定义的两个未来时间区间，新字段命名为 $C-churn-1、$CP-churn-1、$C-churn-2 和 $CP-churn-2。如果使用时间字段 tenure 定义未来时间，则新字段为 $C-churn_tenure 和 $CP-churn_tenure。

如果在 Cox 节点中选中了**追加所有概率**设置选项，则会针对每个未来时间添加两个附加字段，其中包含每条记录生存和失败的概率。这些附加字段基于输出字段的值进行命名，其中前缀 $CP-<false value>- 表示生存的概率，$CP-<true value>- 表示事件已发生的概率，后缀为未来时间区间的数量。例如，对于“false”值为 0，“true”值为 1 的输出字段和以规则区间定义的两个未来时间区间，新字段命名为 $CP-0-1、$CP-1-1、$CP-0-2 和 $CP-1-2。如果使用单个时间字段 tenure 定义未来时间，由于存在单个的未来区间，则新字段为 $CP-0-1 和 $CP-1-1

如果在 Cox 节点中选中了**计算累积风险函数**设置选项，则会针对每个未来时间添加附加字段，其中包含每条记录的累计风险函数。这些附加字段基于输出字段的名称进行命名，前缀为 $CH-，后缀为未来时间区间的数量或定义时间区间的时间字段的名称。例如，对于名为 churn 的输出字段，以及以规则区间定义的两个未来时间区间，新字段命名为 $CH-churn-1 和 $CH-churn-2。如果使用时间字段 tenure 定义未来时间，则新字段为 $CH-churn-1。

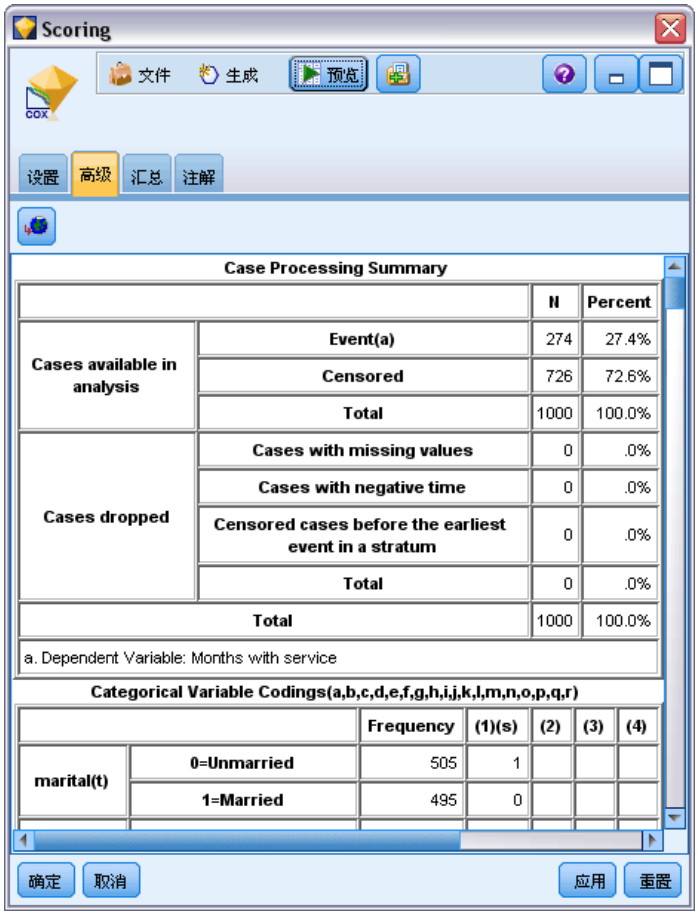
Cox 回归输出设置

块的“设置”选项卡与模型节点的“设置”选项卡包含相同的控件。块控件的默认值由模型节点中设置的值决定。 有关详细信息，请参阅第 289 页码Cox 节点设置选项。

Cox 回归高级输出

Cox 回归的高级输出可提供有关所估计模型及其性能的详细信息，其中包括生存曲线。高级输出中包含的大部分信息的技术含量都很高，需要具备 Cox 回归方面的广泛知识才能正确理解该输出。

图片 10-62
Cox 模型块，“高级”选项卡



Case Processing Summary

		N	Percent
Cases available in analysis	Event(a)	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

Categorical Variable Codings(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r)

		Frequency	(1)(s)	(2)	(3)	(4)
marital(t)	0=Unmarried	505	1			
	1=Married	495	0			

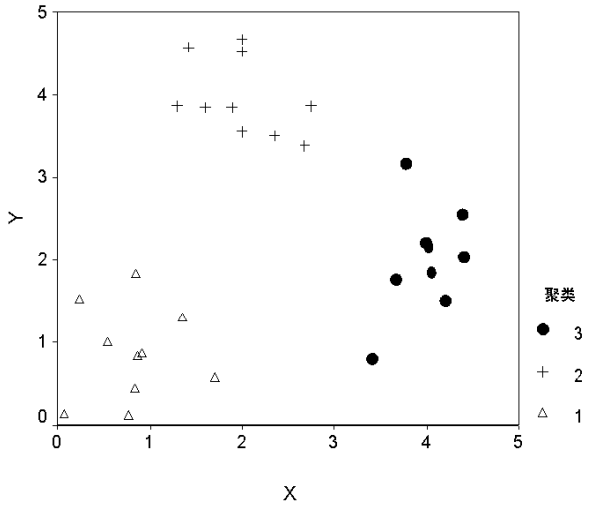
章
11

聚类模型

聚类模型主要用来确定相似记录的组并根据它们所属的组来为记录添加标签。不需事先了解组信息及组特征即可完成该操作。事实上，甚至无法确切知道要查找多少个组。这点将聚类模型与其他机器学习方法区别开来 — 即不存在供模型预测的预定义输出或目标字段。由于不存在用于判断模型的分类效果的外部标准，因而这些模型通常被称作**不受监督学习**模型。对于这些模型而言，不存在对或错的答案。模型的值由模型捕获数据中感兴趣的分组并提供这些分组的有用说明信息的能力来确定。

聚类方法基于对记录间距离和聚类间距离的测量。将记录指派给聚类时将尽量缩短属于同一个聚类的记录之间的距离。

图片 11-1
简单聚类模型



提供有三种聚类方法：

K-Means 节点将数据集聚类到不同分组（或聚类）。此方法将定义固定的聚类数量，将记录迭代分配给聚类，以及调整聚类中心，直到进一步优化无法再改进模型。k-means 节点作为一种非监督学习机制，它并不试图预测结果，而是揭示隐含在输入字段集中的模式。 有关详细信息，请参阅第 298 页码K-Means 节点。

TwoStep 节点使用两步聚类方法。第一步完成简单数据处理，以便将原始输入数据压缩为可管理的子聚类集合。第二步使用层级聚类方法将子聚类一步一步合并为更大的聚类。TwoStep 具有一个优点，就是能够为训练数据自动估计最佳聚类数。它可以高效处理混合的字段类型和大型的数据集。有关详细信息，请参阅第 302 页码两步聚类节点。

Kohonen 节点会生成一种神经网络，此神经网络可用于将数据集聚类到各个差异组。此网络训练完成后，相似的记录应在输出映射中紧密地聚集，差异大的记录则应彼此远离。您可以通过查看模型块中每个单元所捕获观测值的数量来找出规模较大的单元。这将让您对聚类的相应数量有所估计。 有关详细信息，请参阅第 293 页码Kohonen 节点。

通常使用聚类模型来创建聚类或段，然后将聚类或段用作后续分析的输入。常见例子如营销人员常使用市场分段来将整个市场划分为多个类似的子组。每个市场分段都有自己的特征，该特性将影响到针对该分段的市场营销努力是否能取得成功。如果您使用数据挖掘来优化市场营销战略，通常可以通过识别合适的市场分段和在预测模型中使用分段信息来显著改进模型。

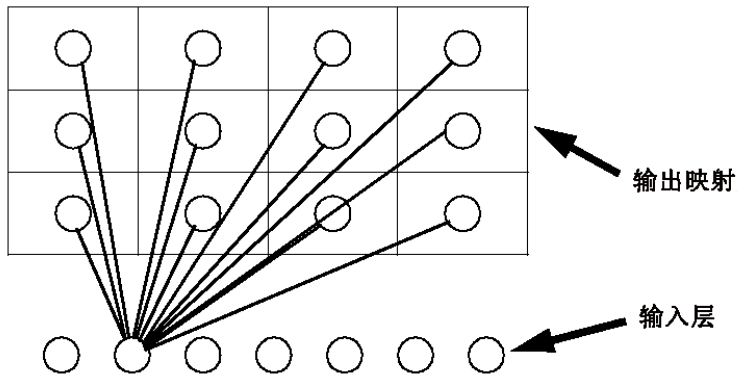
Kohonen 节点

Kohonen 网络是一种执行聚类的神经网络类型，也称为 knet 或**自组织映射**。如果在开始时没有分组的相关信息，则可使用此类型的网络将数据集聚类到有明显区别的不同分组。对记录进行分组，以便组或聚类中的记录趋于相似，而不同组中的记录则有所差异。

基本单元为**神经元**，神经元将分作两层：**输入层**和**输出层**（又称为**输出映射**）。所有输入神经元都和所有输出神经元相连接，这些连接有与其相关的**强度**或**权重**。训练过程中，每个单元会与所有其它单元进行竞争以“赢得”每条记录。

输出映射是神经元的二维网络（单元之间无连接）。A 3 × 4 映射显示如下（虽然一般情况下映射比这要大）。

图片 11-2
Kohonen 网络的结构



所有输入神经元都已与输出映射中的所有神经元相连。（为清楚起见没有显示所有连接）

输入数据会显示在输入层，相应值将传播到输出层。响应最强的输出神经元将称为**胜利者**并且会成为输入的结果。

最初的权重随机产生。如果某个单元赢得一条记录，则其权重（与其附近单元的权重一起统称为**近邻**）将作调整以尽可能地与此条记录的预测变量值的模式相匹配。显示所有输入记录，并且权重将相应更新。将重复此过程，直到变化非常小为止。当进行训练时，网格单元的权重将作调整从而形成聚类的一个二维“映射”（所以会有术语**自组织映射**）。

此网络训练完成后，相似的记录应在输出映射中紧密地聚集，差异很大的记录则应彼此远离。

与 IBM® SPSS® Modeler 中的大多数学习方法不同的是，Kohonen 网络不 使用目标字段。这种没有目标字段的学习称为**无监督学习**。Kohonen 网络试图揭示输入字段集中的模式而不是预测结果。通常，Kohonen 网络最终会形成几个汇总许多观测数据的单元（**强**单元），以及几个实际不对应任何观测数据的单元（**弱**单元）。强单元（有时也包括网格中与其邻近的其他单元）代表可能的聚类中心。

Kohonen 网络的另一种用途是**降维**。二维网格的空间特性可提供从 k 个原始预测变量到保留了原始预测变量中相似性关系的两个派生特征的映射。在某些情况下，此方法的作用与因子分析或主成分分析的作用相同。

请注意，计算输出网格默认大小的方法与 SPSS Modeler 以前的版本相比已发生了变化。通常，新方法将生成更小的输出层，这些输出层训练起来更快且通用性更强。如果您发现使用默认大小得到的结果不理想，可以尝试在“专家”选项卡上增加输出网格的大小。 有关详细信息，请参阅第 296 页码Kohonen 节点专家选项。

要求。 要训练 Kohonen 网络，您需要一个或多个角色设置为输入的字段。角色设置为目标、两者或无的字段将被忽略。

强度。构建 Kohonen 网络模型不需要有组成员关系数据。您甚至不需要知道要寻找的组的个数。Kohonen 网络刚开始会有大量的单元，随着训练的进行，这些单元会向数据中的自然聚类集中。可通过查看模型块中每个单元捕获的观测值数来识别强单元，进而了解适当的聚类数。

Kohonen 节点模型选项

图片 11-3
Kohonen 节点模型选项

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

继续训练现有模型。默认情况下，每次执行 Kohonen 节点时，就会创建一个全新的网络。如果选中此选项，则会继续训练该节点成功生成的最后一个网络。

显示反馈图形。如果选中此选项，则会在训练期间显示二维数组的直观表示。每个节点的强度用颜色表示。红色表示聚集了许多记录的单元（**强**单元），白色表示聚集的记录较少或没有记录的单元（**弱**单元）。如果构建模型所花费的时间相对较短，���能不会显示反馈。注意，此功能会减慢训练进度。要加快训练进度，请取消选中此选项。

图片 11-4
Kohonen 反馈图形

停止于。默认停止标准会基于内部参数停止训练。也可以指定时间作为停止标准。以分钟为单位输入网络训练的时间。

设置随机数种子。如果不设置随机数种子，则每次执行节点时用于初始化网络权重的随机值的序列都会不同。这将导致即使节点设置和数据值都完全相同，节点也会在不同的运行中创建不同的模型。通过选择该选项，可以将随机种子设置为特定值，从而使结果模型具有精确的可再现性。特定的随机种子通常会生成相同的随机值序列，在这种情况下执行节点通常会产生相同的生成模型。

注：为从数据库中读取的记录选择**设置随机数种子**选项时，可能在抽样前需要使用排序节点以确保每次执行节点时能得到相同的结果。这是因为随机种子依赖于记录的顺序，而在关系数据库中不能保证记录具有这种顺序。 有关详细信息，请参阅第 3 章中的排序节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

注意：如果要在模型中包括名义（集合）字段，但在构建模型时遇到内存问题，或构建模型所需的时间太长，则可以考虑重新编码大型集合字段以减少值的数量，或考虑使用值较少的其他字段作为该大型集合的代理。例如，如果包含个别产品值的 product_id 字段出现问题，可以考虑将其从模型中删除并改为添加信息不是很详细的 product_category 字段代替。

优化。根据您的具体需求，选择为了提高建模性能而设计的选项。

- 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。此选项默认为选中状态。

  注：以分布式模式运行时，此设置可能会被 options.cfg 中指定的管理员选项覆盖。 有关详细信息，请参阅第 4 章中的使用 options.cfg 文件中的IBM SPSS Modeler Server 14.2 管理和性能指南。

追加聚类标签。默认对新模型选中此选项，但对从较早版本的 IBM® SPSS® Modeler 加载的模型取消选中。该选项会创建一个由 K-Means 和“两步聚类”节点共同创建的相同类型的分类得分字段。在计算不同模型类型的排序测量量时，该字符串字段用于“自动聚类”节点。 有关详细信息，请参阅第 95 页码第 5 章中的自动聚类节点。

Kohonen 节点专家选项

对于对 Kohonen 网[illegible]有详尽了解的用户，可使用专家选项对训练过程进行微调。要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

图片 11-5
Kohonen 专家选项

宽度和长度。将二维输出图的大小（宽度和长度）指定为每个维上的输出单元数。

学习速率衰减。选择线性或指数学习速率衰减。**学习速率**是随时间递减的加权因子，使得网络可以从数据的大尺度特征开始进行编码，然后逐渐集中于更细微的数据信息。

阶段 1 和阶段 2。Kohonen 网络训练分为两个阶段。阶段 1 是粗略估计阶段，用于捕获数据中的大致模式。阶段 2 是调整阶段，用于调整图以便为数据更精细的特征建模。每个阶段都有以下三个参数：

- **近邻。**设置近邻的起始大小（半径）。此参数确定在训练期间与赢得单元一起被更新的“邻近”单元数。在阶段 1，近邻大小以阶段 1 近邻为起始值，然后减少到（阶段 2 近邻 + 1）。在阶段 2，近邻大小起始为阶段 2 近邻，然后减少到 1.0。阶段 1 近邻应大于阶段 2 近邻。
- **初始 Eta。**为学习速率 eta 设置起始值。在阶段 1，eta 起始为阶段 1 初始 Eta，然后减少到阶段 2 初始 Eta。在阶段 2，eta 起始为阶段 2 初始 Eta，然后减少到 0。阶段 1 初始 Eta应大于阶段 2 初始 Eta。
- **周期。**为训练的每个阶段设置周期数。每个阶段均会进行指定次数的数据处理。

Kohonen 模型块

Kohonen 模型块包含由经过训练的 Kohonen 网络捕获的所有信息，还包含有关网络体系结构的信息。

当运行包含 Kohonen 模型块的流时，节点将添加两个新字段，这两个字段包含 Kohonen 输出网格中对该记录反应最强烈的单元的 X 坐标和 Y 坐标。新字段名得自模型名称，即在模型名称前加上前缀 $KX- 和 $KY-。例如，如果模型名称为 Kohonen，则新字段的名称应是 $KX-Kohonen 和 $KY-Kohonen。

为了更好地了解 Kohonen 网络编码的内容，可单击模型块浏览器上的“模型”选项卡。此时会显示聚类查看器，提供聚类、字段和重要性等级的图形表示。 有关详细信息，请参阅第 306 页码聚类浏览器 - 模型选项卡。

如果更愿意以网格形式显现聚类，则可以通过使用散点图节点绘制 $KX- 和 $KY- 字段来查看 Kohonen 网络的结果。（应在散点图节点中选择 X-Agitation 和 Y-Agitation 以防止每个单元的记录彼此覆盖。）在散点图中，也可以交叠符号字段以调查 Kohonen 网络是如何聚类数据的。

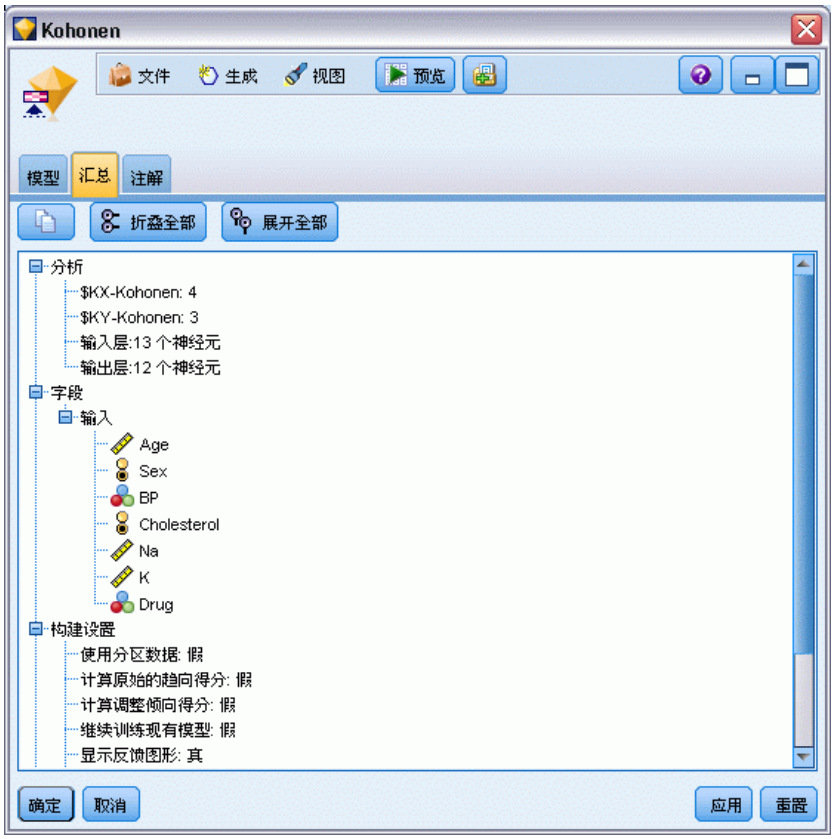
深入了解 Kohonen 网络的另一个有力方法是使用规则归纳来揭示用于区分通过网络发现的聚类的特征。 有关详细信息，请参阅第 146 页码第 6 章中的C5.0 节点。

有关使用模型浏览器的一般信息，请参阅 浏览模型块

Kohonen 模型汇总

Kohonen 模型块的“汇总”选项卡显示有关网络的体系结构或拓扑结构的信息。二维 Kohonen 特征图（输出层）的长度和宽度显示为 $KX-model_name 和 $KY-model_name。对于输入层和输出层，将列出该层的单元数。

图片 11-6
Kohonen 模型块：“汇总”选项卡



K-Means 节点

K-Means 节点提供一种进行**聚类分析**的方法。它可以用于在最初不知道有哪些组时，将数据集聚类为不同的组。与 IBM® SPSS® Modeler 中的大多数学习方法不同的是，K-Means 模型不 使用目标字段。这种没有目标字段的学习称为**无监督学习**。K-Means 模型试图揭示输入字段集的模式而不是预测结果。对记录进行分组，以使一个组或聚类中的记录彼此相似，而不同组中的记录则互不相同。

K-Means 的工作原理是根据数据定义一组起始聚类中心。然后根据记录的输入字段值，将每个记录分配到与其最相似的聚类中。在分配完所有记录后，更新聚类中心以反映分配到每个聚类的新记录集。然后再次检查记录，以确定是否应将这些记录重新分配到不同的聚类中，这个记录分配/聚类迭代过程将一直持续，直到达到最大迭代次数或一次迭代与下次迭代之间的改变不超过指定阈值为止。

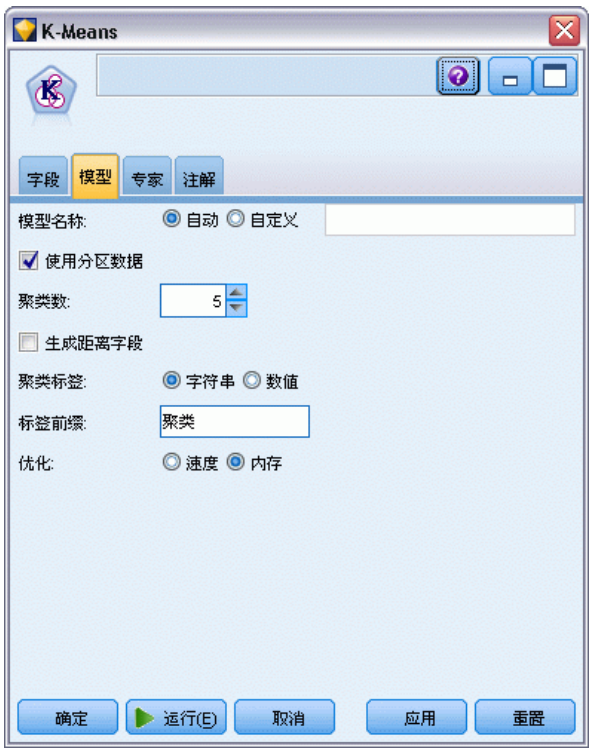
注意：得到的模型一定程度上取决于训练数据的顺序。重排数据顺序并重新构建模型有可能得到不同的聚类模型。

要求。 要训练 K-Means 模型，您需要一个或多个角色设置为输入的字段。角色设置为输出、两者或无的字段将被忽略。

强度。构建 K-Means 模型不需要有组成员关系数据。通常，K-Means 模型是进行大型数据集聚类的最快方法。

K-Means 节点模型选项

图片 11-7
K-Means 节点模型选项



模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

指定的聚类数。指定要生成的聚类数。默认值为 5。

生成距离字段。如果选中此选项，则模型块将包括一个字段，该字段包含每个记录与所分配到的聚类的中心之间的距离。

聚类标签。为生成的聚类成员关系字段的值指定格式。聚类成员关系可表示为具有指定**标签前缀**的**字符串**（例如，"Cluster 1"、"Cluster 2" 等等），也可以表示为**数值**。

注意：如果要在模型中包括名义（集合）字段，但在构建模型时遇到内存问题，或构建模型所需的时间太长，则可以考虑重新编码大型集合字段以减少值的数量，或考虑使用值较少的其他字段作为该大型集合的代理。例如，如果包含个别产品值的 product_id 字段出现问题，可以考虑将其从模型中删除并改为添加信息不是很详细的 product_category 字段代替。

优化。根据您的具体需求，选择为了提高建模性能而设计的选项。

- 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。此选项默认为选中状态。

  注：以分布式模式运行时，此设置可能会被 options.cfg 中指定的管理员选项覆盖。 有关详细信息，请参阅第 4 章中的使用 options.cfg 文件中的IBM SPSS Modeler Server 14.2 管理和性能指南。

K-Means 节点专家选项

对于对 k-means 聚类有详尽了解的用户，可使用专家选项对训练过程进行微调。要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

图片 11-8
K-Means 专家选项

停止于。指定训练模型时要使用的停止标准。**默认**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**自定义**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停止模型训练。
- **差异容忍度。**使用此选项可在某次迭代的聚类中心中的最大差异小于指定的水平时停止模型训练。

集合编码值。指定 0 到 1.0 之间的值用于对作为数值字段组的集合字段进行重新编码。默认值是 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。

K-Means 模型块

K-Means 模型块包含由聚类模型捕获的所有信息，还包含有关训练数据和估计过程的信息。

当运行包含 K-Means 模型块的流时，该节点将添加两个新字段，这两个字段包含聚类成员以及与该记录所分配到的聚类中心的距离。新字段名得自模型名称，即为聚类成员加上 $KM- 前缀，为与聚类中心的距离加上 $KMD- 前缀。例如，如果模型名称为 Kmeans，则新字段的名称应是 $KM-Kmeans 和 $KMD-Kmeans。

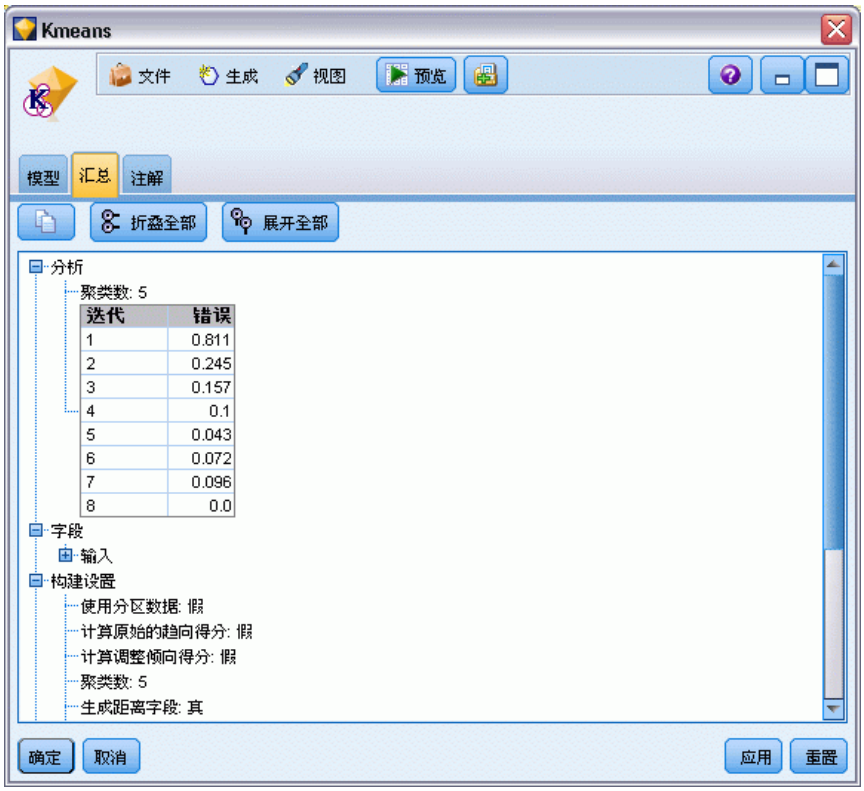
深入了解 K-Means 模型的一种有力方法是使用规则归纳来揭示用于区分通过模型发 ◆◆◆ 的聚类的特征。有关详细信息，请参阅第 146 页码第 6 章中的C5.0 节点。也可以单击模型块浏览器上的“模型”选项卡以显示聚类查看器，它提供聚类、字段和重要性等级的图形表示。 有关详细信息，请参阅第 306 页码聚类浏览器 - 模型选项卡。

有关使用模型浏览器的一般信息，请参阅 浏览模型块

K-Means 模型汇总

K-Means 模型块的“汇总”选项卡包含有关训练数据、估计过程和由模型定义的聚类的信息。显示的信息有聚类数，还有迭代历史。如果已执行附加到此建模节点的分析节点，则分析信息也将显示在此选项卡上。 有关详细信息，请参阅第 6 章中的分析节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

图片 11-9
K-Means 模型块：“汇总”选项卡



两步聚类节点

“两步聚类”节点提供一种形式的**聚类分析**。它可以用于在最初不知道有哪些组时，将数据集聚类为不同的组。与 Kohonen 节点和 K-Means 节点一样，“两步聚类”模型也不使用目标字段。“两步聚类”模型试图揭示输入字段集的模式而不是预测结果。对记录进行分组，以使一个组或聚类中的记录彼此相似，而不同组中的记录则互不相同。

“两步聚类”是一种分两步进行聚类的方法。第一步，完成简单数据处理，这个过程将原始输入数据压缩为若干易处理的子聚类。第二步，采用分层聚类方法逐渐将这些子聚类合并成越来越大的聚类，不需要再次进行数据处理。分层聚类的优点在于不需要事先选择聚类数。许多分层聚类方法刚开始都将单个记录作为最初的聚类，然后递归合并这些记录以不断生成更大的聚类。虽然此类方法常因数据数量巨大而失败，但“两步聚类”的初始预聚类使得分层聚类即使数据集巨大速度也非常快。

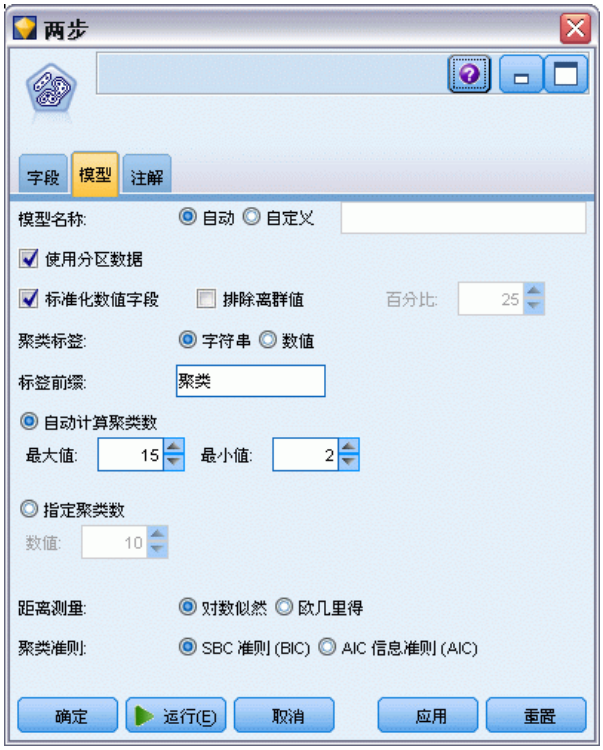
注意：得到的模型一定程度上取决于训练数据的顺序。重排数据顺序并重新构建模型有可能得到不同的聚类模型。

要求。 要训练“两步聚类”模型，您需要一个或多个角色设置为输入的字段。角色设置为目标、两者或无的字段将被忽略。“两步聚类”算法不处理缺失值。构建模型时将忽略任意输入字段包含空白的记录。

强度。“两步聚类”可以处理混合字段类型并能有效处理大型数据集。它还能检验多种聚类解决方案并选择其中最有效的一种，因此不必知道开始时应有多少个聚类。可将“两步聚类”设置为自动排除**离群值**或能对结果造成损害的极其异常情况。

两步聚类节点模型选项

图片 11-10
“两步聚类”节点模型选项



模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

标准化数值字段。默认情况下，“两步聚类”会对所有数值输入字段进行标准化，使它们具有相同的尺度，即均值为 0 且方差为 1。要保留数值字段的原始尺度，可取消选中此选项。符号字段不受影响。

排除离群值。如果选中此选项，则那些与主要聚类似乎格格不入的记录将自动排除在分析之外。这样可以防止此类情况歪曲结果。

离群值检测在预聚类步骤进行。选中此选项时，会将相对于其他子聚类具有较少记录的子聚类视为潜在离群值，且重新构建不包括这些记录的子聚类树。子聚类被视为包含潜在离群值的下限大小由**百分比**选项控制。如果其中某些潜在离群值记录与任何新子聚类配置足够相似，则可将其添加到重新构建的子聚类中。将其余无法合并的潜在离群值视为离群值添加到“噪声”聚类中并排除在分层聚类步骤之外。

使用经过离群值处理的“两步”模型对数据进行评分时，会将与最近主要聚类的距离大于特定阈值距离（基于对数似然）的新观测值视为离群值分配到“噪声”聚类中，名称为 -1。

聚类标签。为生成的聚类成员关系字段指定格式。聚类成员关系可表示为具有指定**标签前缀**的**字符串**（例如，"Cluster 1"、"Cluster 2" 等等），也可以表示为**数值**。

自动计算聚类数。“两步聚类”可以非常迅速地对大量聚类解决方案进行分析并为训练数据选择最佳聚类数。通过设置**最大**聚类数和**最小**聚类数指定要尝试的聚类解决方案的范围。“两步聚类”通过一个两阶段过程确定最佳聚类数。在第一个阶段，随着所添加聚类的增多，可基于贝叶斯信息准则 (BIC) 中的差异选择模型中聚类数的上限。在第二个阶段，为聚类数比最小 BIC 解决方案还少的所有模型找出聚类间最小距离的差异。距离的最大差异用于识别最终聚类模型。

指定聚类数。如果知道模型中要包括的聚类数，请选中此选项并输入聚类数。

距离测量。此选项确定如何计算两个聚类之间的相似性。

- **对数相似性。**该似然度量假设变量服从某种概率分布。假设连续变量是正态分布，而假设分类变量是多项分布。假设所有变量均是独立的。
- **欧几里德距离。**欧几里德距离测量是两个聚类之间的“直线”距离。它只能用于所有变量连续的情况。

聚类准则。此选项确定自动聚类算法如何确定聚类数。可以指定 Bayesian 信息准则 (BIC) 或 Akaike 信息准则 (AIC)。

两步聚类模型块

“两步聚类”模型块包含由聚类模型捕获的所有信息，还包含有关训练数据和估计过程的信息。

当运行包含“两步聚类”模型块的流时，节点将为该记录添加包含聚类成员的新字段。新字段名得自模型名称，即在模型名称前加上 $T- 前缀。例如，如果模型名称为 TwoStep，则新字段的名称应是 $T-TwoStep。

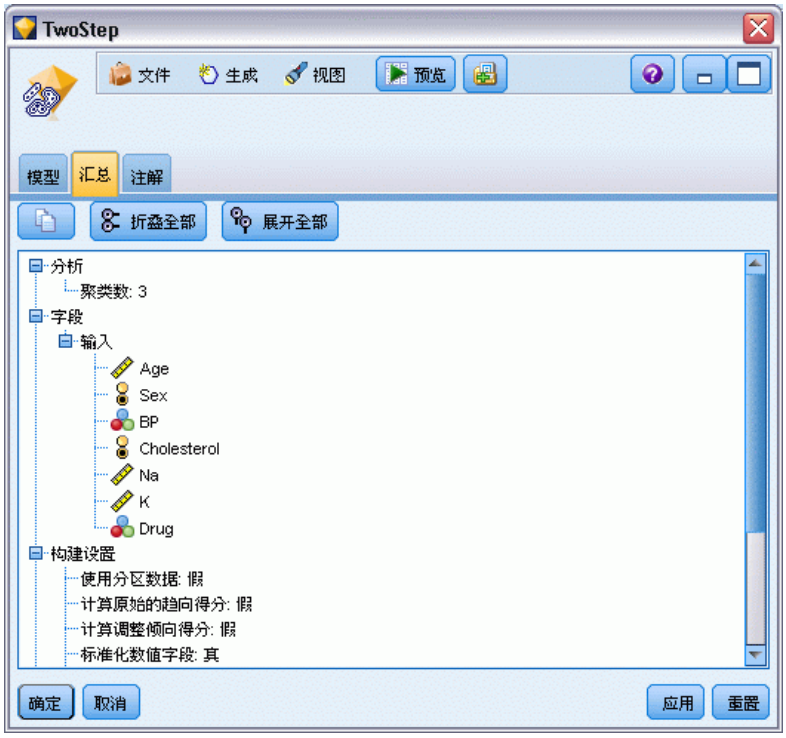
深入了解“两步模型”的一种有力方法是使用规则归纳来揭示用于区分通过模型发现的聚类的特征。有关详细信息，请参阅第 146 页码第 6 章中的C5.0 节点。也可以单击模型块浏览器上的“模型”选项卡以显示聚类查看器，它提供聚类、字段和重要性等级的图形表示。 有关详细信息，请参阅第 306 页码聚类浏览器 - 模型选项卡。

有关使用模型浏览器的一般信息，请参阅 浏览模型块

两步模型汇总

“两步”聚类模型块的“汇总”选项卡显示找出的聚类数以及有关训练数据、估计过程和所使用的构建设置的信息。

图片 11-11
“两步”聚类模型块“汇总”选项卡示例



有关详细信息，请参阅第 43 页码第 3 章中的浏览模型块。

聚类浏览器

聚类模型通常用于根据所检查变量查找具有类似记录的组（聚类），其中同组成员间的相似性高而不同组成员间的相似性低。结果可用于识别原本不明显的关联。例如，通过对客户偏好、收入水平和购物习惯的聚类分析，可以识别出对某种市场营销活动更可能做出反应的客户类型。

有两种方法可以解释聚类显示中的结果：

- 检查聚类以确定该聚类的唯一特征。是否有一个聚类包含所有高收入借款人？此聚类是否包含比其他聚类更多的记录？
- 检查各聚类上的字段以确定值在聚类间的分布情况。个人的教育水平是否决定其在聚类中的成员资格？高信用得分是否在一个聚类或另一个聚类的成员资格之间加以区分？

使用“聚类浏览器”中的主视图和各个链接视图，可以清楚回答这些问题。

可在 IBM® SPSS® Modeler 中生成以下聚类模型块：

- Kohonen 网络模型块
- K-均值模型块
- 二阶聚类模型块

要查看有关聚类模型块的信息，右键单击模型节点并从上下文菜单中选择**浏览**（或选择流中节点的**编辑**）。或者，如果您正使用“自动聚类”建模节点，双击“自动聚类”模型块中的所需聚类模型块。有关详细信息，请参阅第 95 页码第 5 章中的自动聚类节点。

聚类浏览器 - 模型选项卡

聚类模型的“模型”选项卡显示各聚类之间字段的摘要统计和分布的图形显示，也称为**聚类浏览器**。

注意：“模型”选项卡对于使用 IBM® SPSS® Modeler 13 之前版本构建的模型不可用。

图片 11-12
具有默认显示的“聚类浏览器”

“聚类浏览器”包含两个面板，主视图位于左侧，链接或辅助视图位于右侧。有两个主视图：

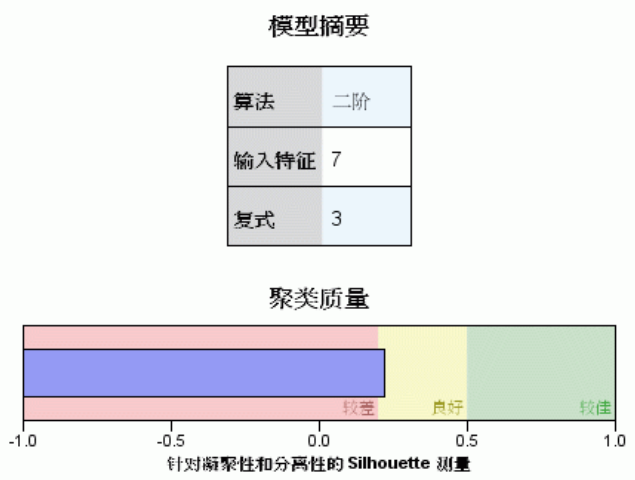
- 模型摘要（默认视图）。 有关详细信息，请参阅第 307 页码模型摘要视图。
- 聚类。 有关详细信息，请参阅第 308 页码聚类视图。

有四个链接/辅助视图：

- 预测变量重要性。 有关详细信息，请参阅第 310 页码聚类预测变量重要性视图。
- 聚类大小（默认视图）。 有关详细信息，请参阅第 311 页码聚类大小视图。
- 单元格分布。 有关详细信息，请参阅第 312 页码单元格分布视图。
- 聚类比较。 有关详细信息，请参阅第 313 页码聚类比较视图。

模型摘要视图

图片 11-13
主面板中的“模型摘要”视图



“模型摘要”视图显示聚类模型的快照或摘要，包括加阴影以表示结果较差、尚可或良好的聚类结合和分离的 Silhouette 测量。该快照可让您快速检查质量是否较差，如果较差，您可返回建模节点修改聚类模型设置以生成较好的结果。

结果较差、尚可和良好是基于 Kaufman 和 Rousseeuw (1990) 关于聚类结构解释的研究成果来判定的。在“模型摘要”视图中，良好的结果表示数据将 Kaufman 和 Rousseeuw 的评级反映为聚类结构的合理迹象或强迹象，尚可的结果将其评级反映为弱迹象，而较差的结果将其评级反映为无明显迹象。

Silhouette 测量所有记录的平均值，(B–A) / max(A,B)，其中 A 是记录与其聚类中心的距离，B 是记录与其非所属最近聚类中心的距离。Silhouette 系数为 1 表示所有个案直接位于其聚类中心上。– 值为 1 表示所有个案位于某些其他聚类的聚类中心上。值为 0 表示在正常情况下个案到其自身聚类中心与到最近其他聚类中心是等距的。

摘要所包含的表格具有以下信息：

- **算法。**所使用的聚类算法，例如“二阶”。
- **输入特征。**字段数量，也称为**输入**或**预测变量**。
- **聚类。**解中聚类的数量。

聚类视图

图片 11-14
主面板中的“聚类中心”视图

“聚类”视图包含一个聚类-特征网格，其中包括每个聚类的名称、大小和概要文件。

网格中的列包含以下信息：

- **聚类。**算法生成的聚类编号。
- **标签。**应用于每个聚类的任何标签（默认为空白）。双击单元格输入描述聚类内容的标签，例如“豪华汽车买家”。
- **描述。**聚类内容的任何描述（默认为空白）。双击单元格输入聚类描述；例如“年龄超过 55 岁、专业人员、收入超过 100,000 美元”。
- **大小。**每个聚类的大小，表示为总体聚类样本的百分比。网格中的每个大小单元格显示一个垂直条，其中显示聚类中的大小百分比、数值格式的大小百分比和聚类个案计数。
- **特征。**单个输入或预测变量，默认按总体重要性排序。如果有列的大小相等，则其以聚类编号的升序显示。

  总体特征重要性由单元格背景阴影的颜色表示；最重要的特征颜色最深；最不重要的特征则没有阴影。表格上方的向导指示与每个特征单元格颜色关联的重要性。

当鼠标悬停在单元格上时，会显示特征的全名/标签和单元格的重要性值。根据视图和特征类型，可能会显示其他信息。在“聚类中心”视图中，这包括单元格统计量和单元格值；例如：“均值：4.32”。对于类别特征，单元格显示最常见（模态）类别的名称及其百分比。

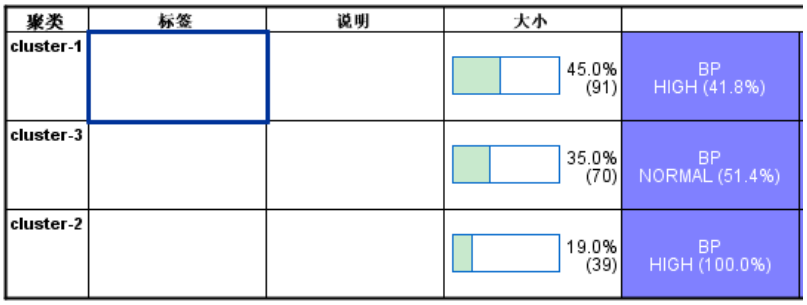
在“聚类”视图中，您可以选择多种显示聚类信息的方式：

- 转置聚类和特征。 有关详细信息，请参阅第 309 页码转置聚类和特征。
- 排序特征。 有关详细信息，请参阅第 309 页码排序特征。
- 排序聚类。 有关详细信息，请参阅第 309 页码排序聚类。
- 选择单元格内容。 有关详细信息，请参阅第 310 页码单元格内容。

转置聚类和特征

默认情况下，聚类显示为列，而特征显示为行。为翻转这种显示，单击**特征排序方式**按钮左侧的**转置聚类和特征**按钮。例如，当显示许多聚类时，您可能想要进行此操作，以减少查看数据所需的水平滚动量。

图片 11-15
主面板中的转置聚类



排序特征

特征排序方式按钮可使您选择特征单元格的显示方式：

- **总体重要性。**这是默认的排序方式。特征以总体重要性的升序进行排序，排序方式在各聚类间相同。如果有特征具有同数重要性值，则按照特征名称的升序列出同数特征。
- **聚类内重要性。**特征按照其相对于每个聚类的重要性进行排序。如果有特征具有同数重要性值，则按照特征名称的升序列出同数特征。当选中此选项时，排序顺序通常因聚类而异。
- **名称。**特征按照名称的字母顺序进行排序。
- **数据顺序。**特征按照其在数据集中的顺序进行排序。

排序聚类

默认情况下，聚类按照大小的降序排序。**聚类排序方**按钮可使您按照名称的字母顺序对其进行排序，或如果您创建了唯一标签，则按照标签的字母顺序对其进行排序。

具有相同标签的特征按照聚类名称排序。如果聚类按照标签排序且您编辑了聚类的标签，则自动更新排序顺序。

单元格内容

单元格按钮使您能够更改特征和评估字段的单元格内容的显示。

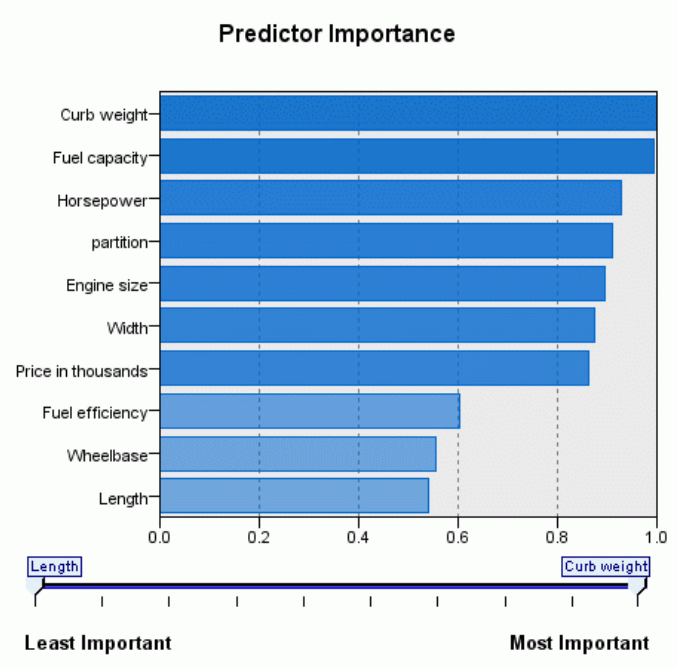
- **聚类中心。**默认情况下，单元格显示特征名称/标签和每个聚类/特征组合的集中倾向。对于连续字段和具有分类字段的类别百分比的模式（最频繁出现的类别）显示均值。
- **绝对分布。**显示特征名称/标签和每个聚类中特征的绝对分布。对于类别特征，显示条形图，其中叠放了按数据值的升序排序的类别。对于连续特征，显示平滑密度图，其对每个聚类使用相同的端点和间隔。

  实心红色显示表示聚类分布，而颜色较淡的显示则表示总体数据。
- **相对分布。**显示特征名称/标签和单元格中的相对分布。总体而言，显示类似于绝对分布的显示，不同之处在于所显示的是相对分布。

  实心红色显示表示聚类分布，而颜色较淡的显示则表示总体数据。
- **基本视图。**如果聚类很多，不滚动很难看到所有详细信息。要减少滚动量，选择此视图将显示更改为更紧凑的表格。

聚类预测变量重要性视图

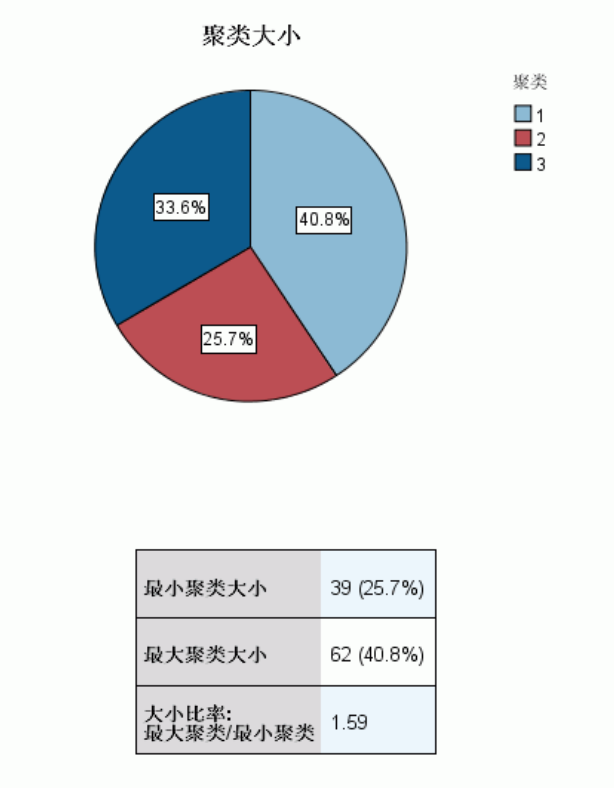
图片 11-16
链接面板中的“聚类预测变量重要性”视图



“预测变量重要性”视图显示评估模型时每个字段的相对重要性。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

聚类大小视图

图片 11-17
链接面板中的“聚类大小”视图



最小聚类大小	39 (25.7%)
最大聚类大小	62 (40.8%)
大小比率: 最大聚类/最小聚类	1.59

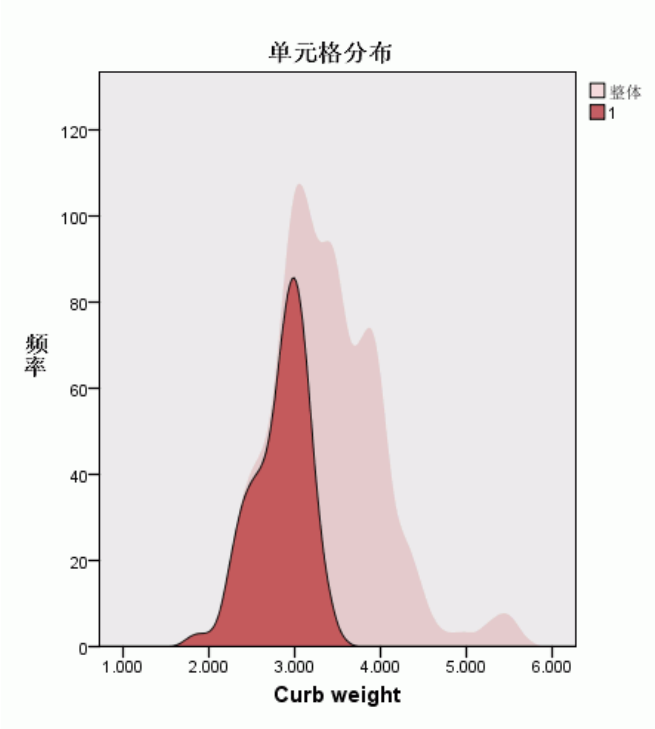
“聚类大小”视图显示包含每个聚类的饼图。每个聚类的百分比大小显示在每个分区上；鼠标悬停在每个分区上显示该分区中的计数。

图表下方的表格列出以下大小信息：

- 最小聚类的大小（总体计数和百分比）。
- 最大聚类的大小（总体计数和百分比）。
- 最大聚类与最小聚类的大小比率。

单元格分布视图

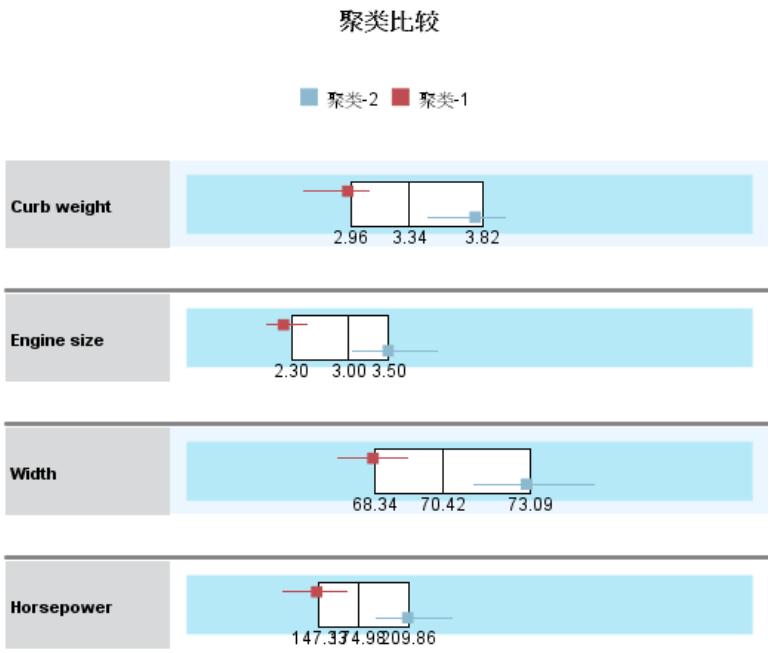
图片 11-18
链接面板中的“单元格分布”视图



“单元格分布”视图显示您在“聚类”主面板的表格中选择的任意特征单元格数据分布的展开的详图。

聚类比较视图

图片 11-19
链接面板中的“聚类比较”视图



“聚类比较”视图由网格式布局构成，行中为特征，列中为选定聚类。此视图帮助您更好地理解组成聚类的因素；同时使您能够看到各聚类间的差异，不但与总体数据比较，而且还在彼此之间比较。

选择要显示的聚类，单击“聚类”主面板中聚类列的顶部。使用 Ctrl+单击或 Shift+单击选择或取消选择多个聚类进行比较。

注意：您可以选择最多五个聚类用于显示。

聚类以选择时的顺序显示，而字段顺序则由**特征排序方式**选项决定。当您选择**聚类内重要性**时，将始终按总体重要性顺序排序字段。

背景图显示每个特征的总体分布：

- 类别特征显示为点图，其中点的大小代表每个聚类最频繁出现的（模态）类别（按特征）。
- 连续特征显示为箱图，其显示整体中位数和四分位距。

叠放在这些背景视图上的是所选聚类的箱图：

- 对于连续特征，方点标记和水平线表示每个聚类的中位数和四分位数间距。
- 每个聚类由不同颜色表示，显示在视图顶部。

浏览聚类浏览器

“聚类浏览器”为交互式显示。您可以：

- 选择字段或聚类以查看更多详细信息。
- 比较聚类以选择感兴趣的项目。
- 更改显示。
- 转置轴。
- 使用“生成”菜单生成“派生”节点、“过滤”节点和“选择”节点。

使用工具栏

您可使用工具栏选项控制在左右两侧面板中显示的信息。您可使用工具栏控件更改显示的方向（从上至下、从左至右或从右至左）。另外，您还可以将浏览器重置为默认设置，并打开对话框以在主面板中指定“聚类”视图的内容。

图片 11-20
用于控制“聚类浏览器”上显示的数据的工具栏

仅当您在主面板中选择**聚类**视图时，**特征排序方式**、**聚类排序方式**、**单元格**和**显示**选项才可用。 有关详细信息，请参阅第 308 页码聚类视图。

	请参阅转置聚类和特征第 309 页码
	请参阅特征排序方式第 309 页码
	请参阅聚类排序方式第 309 页码
	请参阅单元格第 310 页码

从聚类模型生成节点

“生成”菜单可基于聚类模型新建节点。可从生成模型的“模型”选项卡访问该选项，它可基于当前显示或选择（即所有可见聚类或所有选定聚类）生成节点。例如，您可选择一个特征，然后生成“过滤”节点以丢弃所有其他（非可见）特征。生成的节点放置在画布上（未连接）。另外，您还可以在模型调色板上生成模型块的副本。记住，在执行之前连接节点并进行所需编辑。

- **生成建模节点。**在流画布上创建建模节点。例如，如果您想在某个流中使用这些模型设置但您不再拥有用来生成这些设置的建模节点，该功能会很有用。
- **模型到调色板。**在“模型”调色板上创建模型块。当有同事发给您包含模型的流而不是模型本身时，该功能很有用。
- **过滤节点。**创建新的“过滤”节点以过滤聚类模型不使用的过滤字段和/或当前“聚类浏览器”显示中不可见的字段。如果此聚类节点上游有“类型”节点，则所生成的“过滤”节点会丢弃具有角色目标的任何字段。

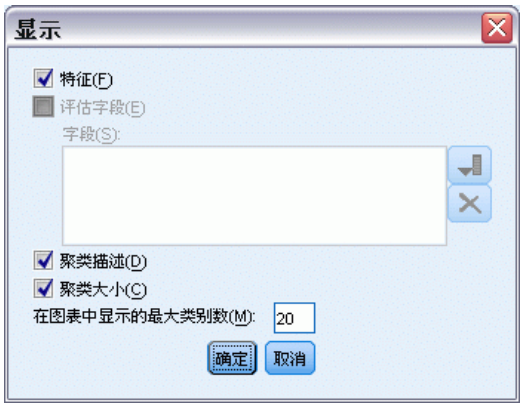
- **过滤节点（从选择创建）。**基于“聚类浏览器”中的选择创建用于过滤字段的新“过滤”节点。使用 Ctrl+单击的方法选择多个字段。在下游丢弃“聚类浏览器”中选择的字段，但您可在执行之前通过编辑“过滤”节点更改此行为。
- **选择节点。**创建新的“选择”节点以基于在当前“聚类浏览器”显示中可见的任一聚类中的成员资格选择记录。自动生成选择条件。
- **选择节点（从选择创建）。**创建新的“选择”节点以基于在“聚类浏览器”中选择的聚类中的成员资格选择记录。使用 Ctrl+单击的方式选择多个聚类。
- **派生节点。**创建新的“派生”节点，其派生出标记字段，该字段基于“聚类浏览器”中所有可见聚类的成员资格分配给记录 True 或 False 值。自动生成派生条件。
- **派生节点（从选择创建）。**创建新的“派生”节点，该节点基于“聚类浏览器”中选择的聚类中的成员资格派生出标记字段。使用 Ctrl+单击的方式选择多个聚类。

除了生成节点之外，您还可以从“生成”菜单创建图形。 有关详细信息，请参阅第 316 页码从聚类模型生成图形。

控制聚类视图显示

要控制主面板的聚类视图中显示的内容，单击显示按钮；打开“显示”对话框。

图片 11-21
“聚类浏览器 - 显示”选项



特征。 默认选定。要隐藏所有输入特征，取消选择该复选框。

评估字段。选择要显示的评估字段（不用于创建聚类模型的字段，但被发送至模型浏览器以评估聚类）；默认不显示任何字段。注意：如果无评估字段可用，则此复选框不可用。

聚类描述。 默认选定。要隐藏所有聚类描述单元格，取消选择该复选框。

聚类大小。 默认选定。要隐藏所有聚类大小单元格，取消选择该复选框。

最大类别数。指定在类别特征图表中显示的最大类别数量；默认值是 20。

从聚类模型生成图形

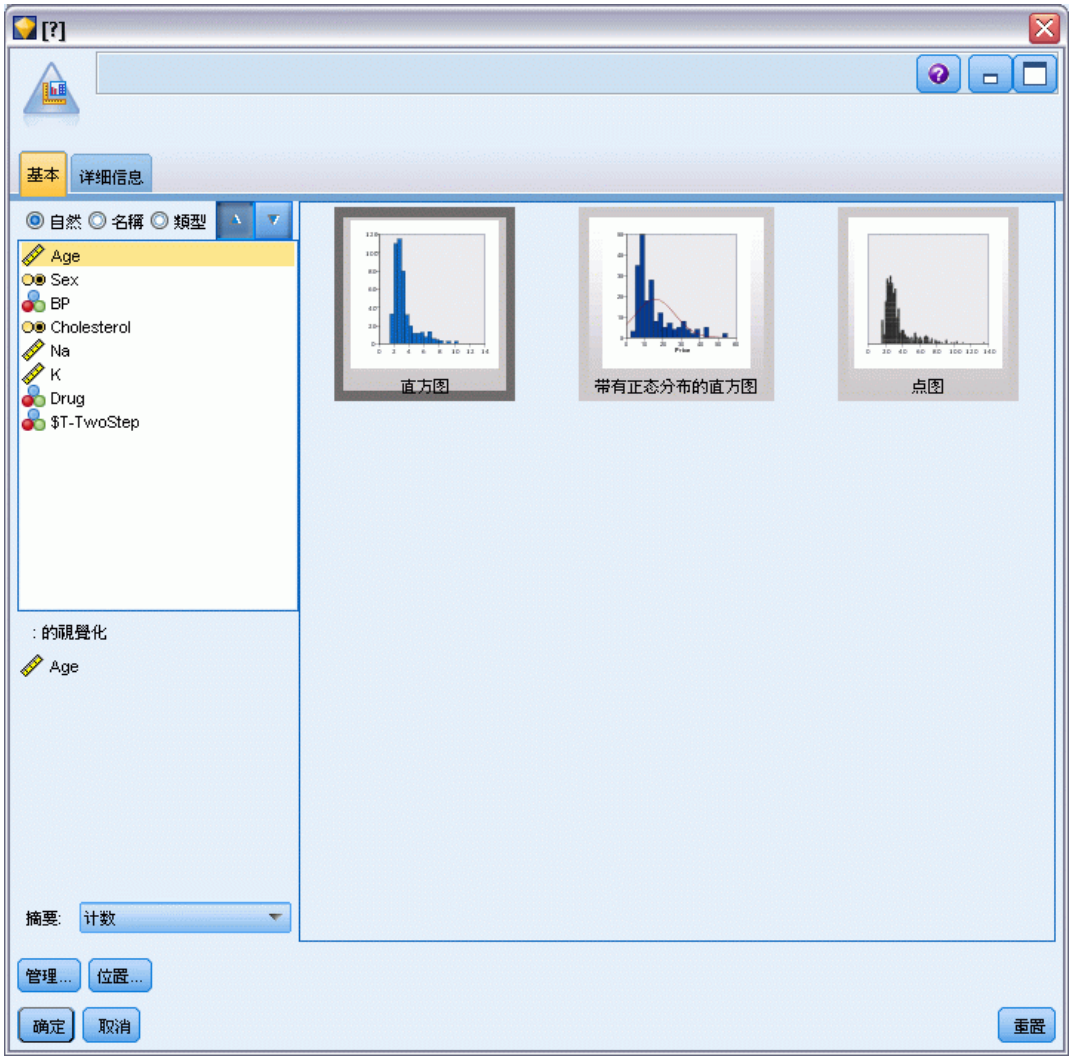
聚类模型提供许多信息，但其格式有时不便于商业用户访问。要使提供的数据便于纳入商业报表和演示文稿，您可生成所选数据的图形。例如，可从“聚类浏览器”生成所选聚类的图形，这样可以只创建该聚类中个案的图形。

注意：仅当模型块连接到流中的其他节点时，您才能从“聚类浏览器”生成图形。

生成图形

- 打开包含“聚类浏览器”的模型块。
- 在“模型”选项卡上，从**视图**下拉列表选择聚类。
- 在主面板上，选择您要为其生成图形的一个或多个聚类。
- 从“生成”菜单，选择**图形（从选择创建）**；显示“图形板基本”选项卡。

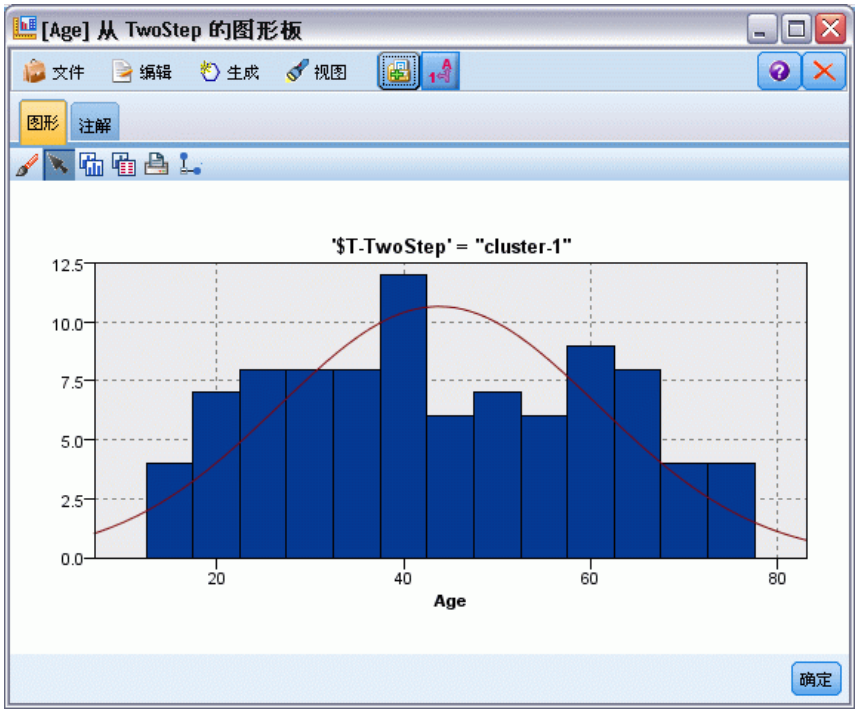
图片 11-22
“图形板”节点对话框，“基本”选项卡



注意：当您以此方式显示“图形板”时，只有“基本”和“详细”选项卡可用。 有关详细信息，请参阅第 5 章中的图形板节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

- 使用“基本”或“详细”选项卡设置指定在图形上显示的详细信息。
- 单击“确定”生成图形。

图片 11-23
从“图形板基本”选项卡生成的直方图



图形标题标识模型类型和选择包含在内的一个或多个聚类。

章
12

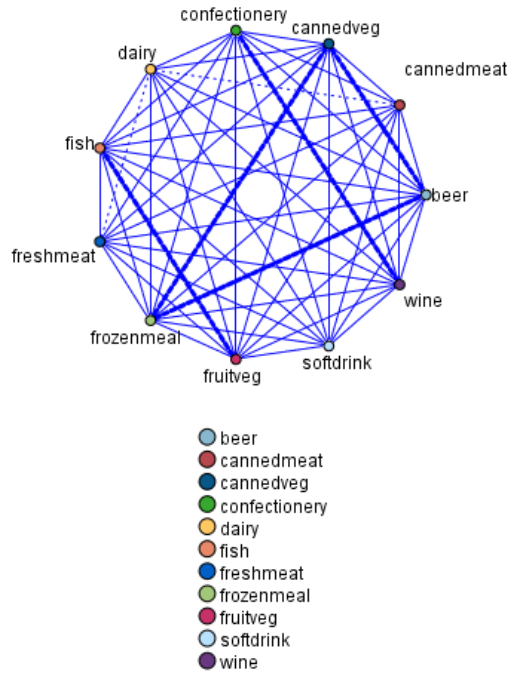
关联规则

关联规则将特定结论（特定产品的购买）与一组条件（若干其他产品的购买）关联起来。例如，规则

啤酒 <= 罐装蔬菜 & 冷冻食品 (173, 17.0%, 0.84)

表述的是：啤酒经常与罐装蔬菜和冷冻食品一起成对出现。该规则可靠率为 84% 并适用于 17% 的数据或 173 条记录。关联规则算法自动找到可使用可视方法（比如 Web 节点）手动找到的关联。

图片 12-1
Web 节点显示各商品分类项目之间的关联



与标准的决策树算法（C5.0 和 C&R 树）相比，关联规则算法的优点是任何属性之间都可以存在关联。决策树算法只使用单一结论来构建规则，而关联算法则试图找到更多规则，且每个规则具有不同的结论。

关联算法的缺点是试图在可能非常大的搜索空间中查找规则，因而运行时间比决策树算法长得多。关联算法使用**生成与检验**方法来查找规则（简单规则将初始生成）并对照数据集来验证这些规则。符合条件的规则将被保存，然后规范所有遵循各种限制条件的规则。**规范**是将条件添加到规则的过程。然后这些新规则将对照数据进行验证，并且验证过程中将迭代保存最符合条件和最有用的规则。用户通常会对允许进入规则的前项的可能的数量给出一定限制，并根据信息理论和高效索引方式使用各种方法来缩小原来可能很大的搜索空间。

处理结束后，将给出最符合条件的规则的列表。此组关联规则不能直接用于做出预测，这点与标准的模型（比如决策树或神经网络）不同。这是由于规则可能有许多不同的结论。需要将关联规则转换为分类规则集的另外一层转换。因此，关联算法生成的关联规则被称作**非精练模型**。虽然用户可以浏览这些非精练模型，但除非用户指令系统从非精练模型生成分类模型，否则无法明确地将这些模型用作分类模型。用户可通过浏览器的“生成”菜单选项来完成这种转换。

支持两种关联规则算法：

Apriori 节点从数据抽取一组规则，即抽取信息内容最多的规则。Apriori 节点提供五种选择规则的方法并使用复杂的索引模式来高效地处理大数据集。对于较大的问题，Apriori 训练的速度通常较快；它对可保留的规则数量没有任何限制，而且可处理最多带有 32 个前提条件的规则。Apriori 要求输入和输出字段均为分类型字段，但因为它专为处理此类型数据而进行优化，因而处理速度快得多。 有关详细信息，请参阅第 321 页码Apriori 节点。

序列节点可发现连续数据或与时间有关的数据中的关联规则。序列是一系列可能会以可预测顺序发生的项目集合。例如，一个购买了剃刀和须后水的顾客可能在下次购物时购买剃须膏。序列节点基于 CARMA 关联规则算法，该算法使用有效的两步法来发现序列。 有关详细信息，请参阅第 343 页码序列节点。

表格格式数据与事务处理格式数据

关联规则模型使用的数据可能是交易格式，也可能表格格式，如下所述。下面的内容是一般描述；具体的要求可能有所不同，请参见每种模型类型文档中的讨论。请注意，对模型进行评分时，要评分的数据必须反映用于构建该模型的数据格式。使用表格数据构建的模型只能用于对表格数据进行评分；使用交易数据构建的模型只能对交易数据进行评分。

交易格式

交易数据对于每个交易或项目具有一个单独的记录。例如，如果客户进行了多次采购，则每次采购都会有一个单独的记录，并且相关联的商品与客户 ID 相链接。这种格式有时称为**行穷尽**格式。

客户	采购
1	jam
2	milk
3	jam
3	bread
4	jam
4	bread
4	milk

Apriori、CARMA 和序列节点都可使用交易数据。

表格数据

表格数据（也称为**篮子**数据或**真值表**数据）由单独的标志表示项目，其中每个标志字段表示一个特定项目的存在或不存在。每个记录表示一个相关项目的完整集合。标志字段可以是分类的也可以是数字的，但某些模型具有更具体的要求。

客户	Jam	Bread	Milk
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

Apriori、CARMA、 和序列节点都可使用表格数据。

Apriori 节点

Apriori 节点会发现数据中的关联规则。Apriori 提供了五种用来选择规则的方法，它使用一种复杂的指数模式来有效处理大型数据集。

要求。要创建 Apriori 规则集，您需要一个或多个输入字段和一个或多,ö目标字段。输入字段和输出字段（角色为输入、目标或两者的字段）必须是符号型字段。角色为无的字段将被忽略。执行节点之前字段类型必须完全实例化。数据可以是表格格式，也可以是事务格式。 有关详细信息，请参阅第 320 页码表格格式数据与事务处理格式数据。

强度。 对于较大的问题，Apriori 训练的速度通常 处理速度快。它对于可以包含的规则数也没有任何限制，可以处理最多带有 32 个预条件的规则。Apriori 提供了五种不同的训练方法，因此将数据挖掘方法与当前问题相匹配时可以实现更强的灵活性。

Apriori 节点模型选项

图片 12-2
Apriori 节点模型选项

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

最低条件支持度。您可以指定在规则集中保留规则的支持度标准。**支持度**指的是训练数据中条件（规则中的“if”部分）为真的记录的百分比。（请注意，此支持度定义与 CARMA 和序列节点中使用的定义不同。有关详细信息，请参阅第 346 页码序列节点模型选项。）如果您获得的规则适用于非常小的数据子集，请尝试增加此设置。

注意：Apriori 的支持度定义基于带有条件的记录的数量。这与 CARMA 和序列算法不同，对于这两种算法，支持度定义基于具有规则中所有项（即条件和结果）的记录的数量。关联模型的结果显示（条件）支持度和规则支持度两个测量。

最小规则置信度。您还可以指定置信度标准。**置信度**基于其规则条件为真的记录，指的是其结果也为真的那些记录的百分比。换句话说，置信度是基于规则的正确预测的百分比。置信度低于指定标准的规则将被放弃。如果您获得的规则太多，请尝试增加此设置。如果您获得的规则太少（甚至根本无法获得规则），请尝试降低此设置。

最大条件数。您可以为任何规则指定最大预条件数。这是一种用来限制规则复杂性的方式。如果规则太复杂或者太具体，请尝试降低此设置。此设置对于训练时间也具有很大的影响。如果规则集训练所需的时间过长，请尝试降低此设置。

只显示值为真的标志变量。如果对于表格（数据表）格式的数据选择了此选项，则在生成的规则中只会包括真值。这样可能有助于使得规则更容易理解。该选项不适用于事务格式的数据。 有关详细信息，请参阅第 320 页码表格格式数据与事务处理格式数据。

优化。根据您的具体需求，选择为了提高建模性能而设计的选项。

- 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。此选项默认为选中状态。注意：以分布式模式运行时，此设置可能会被 options.cfg 中指定的管理员选项覆盖。有关更多信息，请参阅《IBM® SPSS® Modeler Server 管理员指南》。

Apriori 节点专家选项

对于那些详细了解 Apriori 操作的人员来说，通过下列专家选项可以对归纳过程进行微调。要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

图片 12-3
Apriori 专家选项

评估尺度。Apriori 支持五种用来评估潜在规则的方法。

- **规则置信度。**该默认方法使用规则置信度（或准确性）来评估规则。对于此评估尺度，**评估尺度下限**为禁用状态，因为此选项对于“模型”选项卡上的**最小规则置信度**选项来说是多余的。 有关详细信息，请参阅第 322 页码Apriori 节点模型选项。
- **置信度差。**（也称为**与先验相比的绝对置信度差**。）此评估尺度是规则的置信度与其先验置信度之间的绝对差。此选项会防止出现偏差，即结果分布不均匀。因此有助于防止保留“很明显的”规则。例如，可能会出现这样的情况，80% 的客户会购买您最受欢迎的产品。某个以 85% 的准确性预测购买该受欢迎产品的规则不会使您的了解加深，尽管 85% 的准确性对于绝对尺度来说似乎已经相当好了。请将该评估尺度下限设置为您希望保留的规则的置信度最小差。
- **置信度比率。**（也称为**置信度商数与 1 之间的差**。）此评估尺度为 1 减去规则置信度与先验置信度之间的比（如果该比率大于一，则减去其倒数）。与置信度差相似，此方法会考虑不均匀分布。此方法尤其适用于找出预测不常发生事件的规则。例如，假设有一种非常少见的医疗情况只在 1% 的病人中出现。如果一个规则能有

10% 的机会预测出这种医疗情况，那么它与随机猜测相比就是一种很大的提高，尽管从绝对尺度角度来看，10% 的准确性好像非常不起眼。请将该评估尺度下限设置为您希望保留的规则的最小差。

- **信息差。**（也称为**与先验的信息差**。）此评估尺度基于**信息收益**测量。如果某个特定结果的概率被视为一个逻辑值（一个**数位**），则信息收益为基于条件可以确定的该数位的比例。信息差是给定条件的情况下信息收益与只给定了结果的先验置信度的情况下信息收益之间的差。此方法的一个重要特征在于，它考虑了支持度，因此对于给定水平的置信度，它倾向于覆盖更多记录的规则。请将该评估尺度下限设置为您希望保留的规则的信息差。

  注意：因为此评估尺度的尺度与其他尺度相比在某种程度上直观性较差，所以您可能需要试验各种下限才能获得满意的规则集。

- **标准化卡方。**（也称为**标准化卡方评估尺度**。）此评估尺度是条件与结果之间关联的一个统计学指数。此评估尺度进行了标准化，采用 0 和 1 之间的值。此测量尺度甚至比信息差评估尺度更依赖于支持度。请将该评估尺度下限设置为您希望保留的规则的信息差。

  注意：与信息差评估尺度相同，此评估尺度的尺度与其他尺度相比在某种程度上直观性较差，所以您可能需要试验各种下限才能获得满意的规则集。

允许没有条件的规则。选择此选项可允许规则只包括结果（项目或项目集合）。如果您对确定常见项目或项目集合感兴趣，则此选项非常有用。例如，cannedveg 是一个没有条件的单项规则，它表明采购 cannedveg 在数据中经常出现。在某些情况下，如果您只对最可信的预测感兴趣，则可能希望包括这样的规则。此选项默认为关闭状态。按照惯例，没有条件的规则的条件支持度表示为 100%，规则支持度与置信度相同。

CARMA 节点

CARMA 节点使用关联规则发现算法来发现数据中的关联规则。关联规则是下列形式的语句：

如果 条件 **则** 结果

例如，如果某个 Web 客户购买了无限网卡和高端无线路由器，那么该客户还可能购买无线音乐播放器（如果提供该产品的话）。CARMA 模型在不要求用户指定输入或目标字段的情况下从数据抽取一组规则。这就意味着生成的规则可用于很多种应用。例如，您可以使用此节点生成的规则来查找一系列产品或服务（条件），其结果是您要在此假期内进行促销的项目。使用 IBM® SPSS® Modeler，您可以确定哪些客户购买了这些条件产品，然后举办一个旨在促销这些结果产品的营销活动。

要求。与 Apriori 不同，CARMA 节点不需要输入字段或目标字段。这是该算法工作方式的重要组成部分，相当于在将所有字段设置为双向的情况下构建 Apriori 模型。您可以在构建了模型之后过滤该模型，从而限制哪些项目仅列为条件或结果。例如，您可以使用模型浏览器来查找一系列产品或服务（条件），其结果是您要在此假期内进行促销的项目。

要创建 CARMA 规则集，您需要指定一个 ID 字段以及一个或多个内容字段。该 ID 字段可以是任意角色或测量级别。角色为无的字段将被忽略。执行节点之前字段类型必须完全实例化。与 Apriori 相似，数据可以是表格格式，也可以是事务格式。 有关详细信息，请参阅第 320 页码表格格式数据与事务处理格式数据。

强度。CARMA 节点基于 CARMA 关联规则算法。与 Apriori 不同，CARMA 节点为规则支持度（条件和结果的支持度）提供构建设置，而不仅为条件支持度提供构建设置。CARMA 还允许带有多个结果的规则。与 Apriori 相似，CARMA 节点生成的模型可以插入到数据流中用来创建预测。 有关详细信息，请参阅第 38 页码第 3 章中的模型块。

CARMA 节点字段选项

执行 CARMA 节点之前，必须在 CARMA 节点的“字段”选项卡上指定输入字段。虽然大多数建模节点的字段选项卡选项都相同，但 CARMA 节点有几个独特的选项。所有选项均在下面讨论。

图片 12-4
CARMA 节点字段选项

使用类型节点设置。此选项通知节点使用上游类型节点中的字段信息。这是默认值。

使用自定义设置。该选项通知节点使用在此处指定的字段信息，而不是在任何上游类型节点中给出的字段信息。选择了此选项之后，请根据您要读取事务格式的数据还是表格格式的数据来指定下面的字段。

使用交易格式。 此选项将根据您的数据是交易格式还是表格格式来更改此对话框中的其他字段控件。如果您使用带有事务处理格式的数据的多个字段，则认为在某个特定记录中，这些字段中指定的项目表示着可以在一个带有时间戳的事务中找到的项目。 有关详细信息，请参阅第 320 页码表格格式数据与事务处理格式数据。

表格数据

如果未选中**使用事务格式**，则显示以下字段。

- **输入。**选择输入字段或字段。此操作与在“类型”节点中将字段的角色设置为输入类似。
- **分区。**该字段允许您使用指定字段将数据分割为几个不同的样本，分别用于模型构建过程中的训练、测试和验证阶段。通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。如果已使用“类型”或“分区”节点定义了多个分区字段，则必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。（如果仅有一个分区字段，则将在启用分区后自动引入此字段。）有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。同时请注意，要在分析时应用选定分区，同样必须启用节点“模型选项”选项卡中的分区功能。（取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）

交易数据

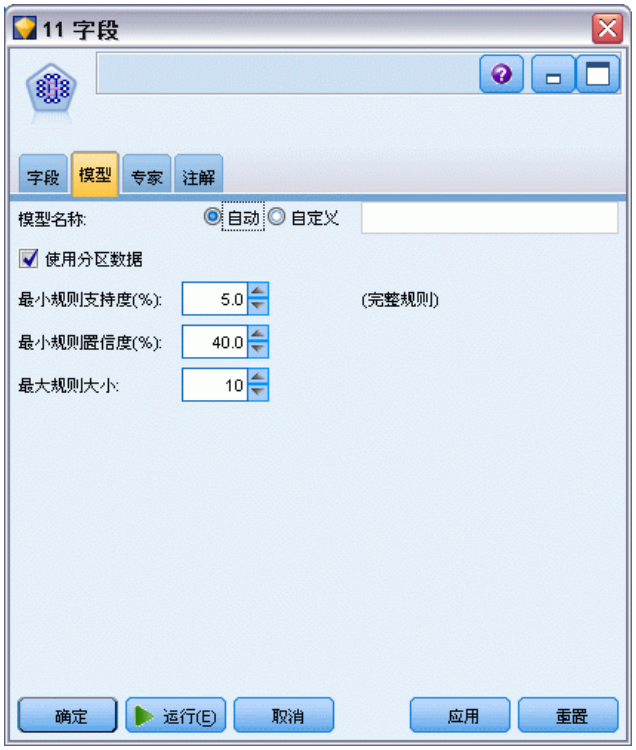
如果选中了**使用事务格式**，则显示以下字段。

- **ID。**对于事务处理格式的数据，请从列表中选择 ID 字段。数字字段或符号字段可用作 ID 字段。此字段的每个唯一值都应该表明一个特定的分析单元。例如，在市场购物篮的应用中，每个 ID 可能表示一个客户。对于 Web 日志分析应用，每个 ID 可能代表一个计算机（以 IP 地址表示）或一个用户（以登录数据表示）。
- **ID 是连续的。** （仅 Apriori 和 CARMA 节点）如果您的数据进行了预先排序，以便所有 ID 相同的记录在数据流中分组在一起，那么选择此选项可以加快处理速度。如果您的数据未经预先排序（或者您不确定），请将此选项保持未选中状态，则该节点将自动对数据进行排序。

  注意：如果您的数据未经排序而您选择了此选项，则可能会在模型中得到无效结果。
- **内容。**指定模型的内容字段。这些字段包含与关联建模有关的项目。您可以指定多个标志字段（如果数据为表格格式）或者一个名义字段（如果数据为事务格式）。

CARMA 节点模型选项

图片 12-5
CARMA 节点模型选项



模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

最低规则支持度 (%)。您还可以指定支持度标准。**规则支持度**指的是训练数据中包含整个规则的 ID 的比例（请注意，此支持度定义与 Apriori 节点中使用的条件支持度不同）。如果您要关注更常见的规则，请增加此设置。

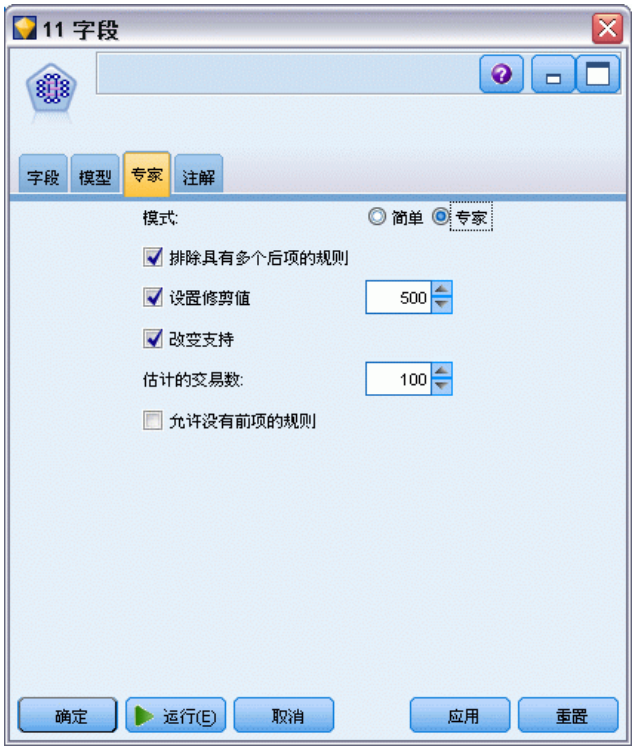
最低规则置信度 (%)。您可以指定在规则集中保留规则的置信度标准。**置信度**指的是得到正确预测的 ID 在所有使用规则进行预测的 ID 中所占的百分比。基于训练数据，该百分比的计算如下：包含整个规则的 ID 数量除以其中包含条件的 ID 数量。置信度低于指定标准的规则将被放弃。如果您获得的规则无关或者太多，请尝试增加此设置。如果您获得的规则太少，请尝试降低此设置。

最大规则大小。您可以设置规则中不同项目集合（与项目 不同） 的最大数量。如果相关规则相对较短，则可以降低此设置，以加快规则集构建速度。

CARMA 节点专家选项

对于那些详细了解 Apriori 操作的人员来说，通过下列专家选项可以对建模过程进行微调。要访问专家选项，请将“专家”选项卡上的“模式”设置为**专家**。

图片 12-6
CARMA 节点专家选项



排除具有多结果的规则。选择该选项可排除“双头”结果，即包含两个项目的结果。例如，规则 bread & cheese & fish -> wine&fruit 包含一个双头结果，即 wine&fruit。默认情况下，这样的规则包括在内。

设置修剪值。为了节省内存，使用的 CARMA 算法在处理期间会从其潜在项目集合列表中定期删除（**修剪**）不频繁的项目集合。选择此选项可调整修剪频率，您指定的数字将决定修剪频率。输入较小的值可降低该算法的内存要求（但可能会延长所需的训练时间），输入较大的值会加快训练速度（但可能会提高内存要求）。默认值为 500。

改变支持度。选择该选项会排除因为纳入不平均而好像表现为非常频繁的不频繁项目集合，从而提高效率。这是通过这样的方式实现的：首先从较高的支持度水平开始，然后逐渐下降到“模型”选项卡上指定的水平。对于**事务的估计数量**输入一个值可指定支持度水平应采用的下降速度。

允许没有条件的规则。选择此选项可允许规则只包括结果（项目或项目集合）。如果您对确定常见项目或项目集合感兴趣，则此选项非常有用。例如，cannedveg 是一个没有条件的单项规则，它表明采购 cannedveg 在数据中经常出现。在某些情况下，如果您只对最可信的预测感兴趣，则可能希望包括这样的规则。此选项默认为不选中状态。

关联规则模型块

关联规则模型块代表由下列关联规则建模节点之一所发现的规则：

- Apriori
- CARMA

模型块包含建模期间从数据提取的规则的相关信息。

查看结果

您可以使用该对话框上的“模型”选项卡浏览关联模型（Apriori 和 CARMA）以及序列模型生成的规则。在生成新节点或对模型评分之前浏览模型块会使您看到规则的相关信息，还会提供用于过滤结果和对结果进行排序的选项。

模型评分

精炼模型块（Apriori、CARMA 和序列）可以添加到流中，用于进行评分。有关详细信息，请参阅第 57 页码第 3 章中的使用流中的模型块。用于评分的模型块在其各自的对话框中包括一个额外的“设置”选项卡。 有关详细信息，请参阅第 335 页码关联规则模型块设置。

非精练模型块无法以其原始格式进行评分。而您可以生成一个规则集，并将该规则集用于评分。 有关详细信息，请参阅第 338 页码从关联模型块生成规则集。

关联规则模型块详细信息

在关联规则模型块的“模型”选项卡上，您可以看到一个表，其中包含了该算法提取的规则。表中的每行都代表一个规则。第一列代表结果（规则的“then”部分），而下一列代表条件（规则的“if”部分）。后面的列包含规则信息，如置信度、支持度和提升。

图片 12-7
关联规则模型块“模型”选项卡

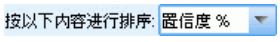
关联规则通常以下列格式显示：

结果	条件
Drug = drugY	Sex = F BP = HIGH

该示例规则的解释为如果 Sex = “F” and BP = “HIGH”，则 Drug 很可能为 drugY；或者以另一种方式解释对于 Sex = “F” and BP = “HIGH” 的记录，Drug 很可能为 drugY。使用对话框工具栏，可以选择显示其他信息，如置信度、支持度和实例数。

排序菜单。工具栏上的“排序”菜单按钮控制着规则的排序。排序的方向（升序或降序）可以使用排序方向按钮（上箭头或下箭头）进行更改。

图片 12-8
用于排序的工具栏选项



您可以按照下列条件对规则进行排序：

- 支持度
- 置信度

- 规则支持
- 结果
- Lift
- 部署能力

显示/隐藏菜单。显示/隐藏菜单（标准工具栏按钮）控制着规则的显示选项。

图片 12-9
显示/隐藏按钮

可用的显示选项如下：

- **规则 ID** 显示建模期间指定的规则 ID。通过规则 ID，可以标识哪些规则要应用于某个给定的预测。通过规则 ID，还可以在以后合并附加的规则信息，如部署能力、产品信息或条件。
- **实例数**显示规则所适用的唯一 ID（即条件为真的 ID）数量的相关信息。例如，假设规则为 **bread -> cheese**，训练数据中包含条件 bread 的记录数量称为**实例数**。
- **支持度**显示条件支持度 — 即其条件为真的 ID 在训练数据中的比例。例如，如果 50% 的训练数据包括 bread（面包）的购买，那么规则 **bread -> cheese** 的条件支持度为 50%。注意：此处定义的支持度与实例数相同，但以百分比的形式表示。
- **置信度**显示规则支持度与条件支持度的比。此比值表明了带有指定条件、并且其结果也为真的 ID 的比例。例如，如果 50% 的训练数据包含 bread（面包）（表明条件支持度），但只有 20% 既包含 bread（面包）又包含 cheese（奶酪）（表明规则支持度），则规则 **bread -> cheese** 的置信度为**Rule Support / Antecedent Support**，在这里为 40%。
- **规则支持度**显示其中整个规则、条件和结果均为真的 ID 的比例。例如，如果 20% 的训练数据既包含 bread（面包）又包含 cheese（奶酪），那么规则 **bread -> cheese** 的规则支持度为 20%。
- **提升**显示规则置信度与具有结果的先验概率的比。例如，如果整个人口统计中 10% 购买了 bread（面包），那么预测人们是否购买 bread（面包）、置信度为 20% 的规则具有的提升将为 20/10 = 2。如果另一个规则告诉您人们将购买 bread（面包），并且置信度为 11%，则该规则的提升接近 1，这就意味着具有条件对于具有结果的概率不会造成太大的影响。总之，提升不为 1 的规则比提升接近 1 的规则的相关性更强。
- **部署能力**是一个有关训练数据中满足了条件而未满足结果的百分比的度量。在产品购买领域，它的意思大致为：总的客户群中有多少百分比拥有了（或已经购买了）条件，但尚未购买结果。部署能力统计量定义为**((Antecedent Support in # of Records - Rule Support in # of Records) / Number of Records) * 100**，其中 Antecedent Support（条件支持度）表示其条件为真的记录数，Rule Support（规则支持度）表示条件和结果都为真的记录数。

过滤按钮。菜单上的“过滤器”按钮（漏斗图标）会扩展对话框的底部，从而显示一个面板，其中将显示活动的规则过滤器。过滤器用于减少“模型”选项卡上显示的规则数量。

图片 12-10
过滤按钮

要创建过滤器，请单击位于扩展面板右侧的过滤器图标。这样将打开一个单独的对话框，您可以在其中指定用于显示规则的约束条件。请注意，“过滤器”按钮通常与“生成”菜单一起使用，以便首先过滤规则，然后生成一个包含部分规则的模型。有关详细信息，请参阅下面的 为规则指定过滤器。

查找规则按钮。通过查找规则按钮（望远镜图标），可以搜索为某个指定的规则 ID 显示的规则。相邻的显示框表明可用数量中当前显示的规则数量。规则 ID 由模型按照发现时间的顺序指定，并且会在评分期间添加到数据中。

图片 12-11
查找规则按钮

要对规则 ID 重新排序：

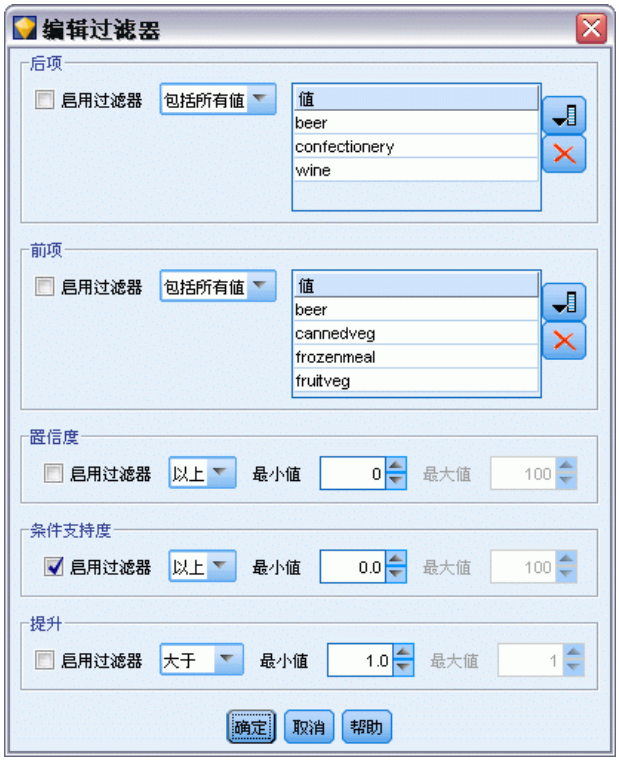
- 您可以在 IBM® SPSS® Modeler 中对规则 ID 进行重新排序，方法是，首先根据所需的测量标准（如置信度或提升）对规则显示表进行排序。
- 然后使用“生成”菜单中的选项，创建一个经过过滤的模型。
- 在“已过滤的模型”对话框中，选择**重新进行连续编号的起始号码**，然后指定一个开始号码。

有关详细信息，请参阅第 339 页码生成已过滤的模型。

为规则指定过滤器

默认情况下，规则算法（如 Apriori、CARMA 和序列）可能会生成非常大量的规则。为了在浏览时增强明确度，或者为了简化规则评分，您应该考虑过滤规则，以便更加显著地显示相关的结果和条件。使用规则浏览器“模型”选项卡上的过滤选项，可以打开一个用于指定过滤条件的对话框。

图片 12-12
规则浏览器过滤器对话框



结果。选择**启用过滤器**可激活基于包括还是排除指定结果的过滤规则的选项。选择**包括任意**可创建一个过滤器，该过滤器中的规则至少包含一个指定结果。另外，选择**排除**可创建一个排除指定结果的过滤器。您可以使用列表框右侧的选取器图标选择结果。这样将打开一个对话框，其中列出生成的规则中包含的所有结果。

注意：结果可能包含多个项目。过滤器只会检查结果是否包含一个指定项目。

条件。选择**启用过滤器**可激活基于包括还是排除指定条件的过滤规则的选项。您可以使用列表框右侧的选取器图标选择项目。这样将打开一个对话框，其中列出生成的规则中包含的所有条件。

- 选择**包括所有**可将过滤器设置为一个包含过滤器，其中的规则必须包括指定的所有条件。
- 选择**包括任意**可创建一个过滤器，该过滤器中的规则至少包含一个指定条件。
- 选择**排除**可创建一个排除包含指定条件的规则的过滤器。

置信度。选择**启用过滤器**可激活基于规则的置信水平过滤规则的选项。您可以使用**最小**和**最大**控件来指定置信度范围。当您浏览生成的模型时，置信度将以百分比的形式列出。当您对输出评分时，置信度则表示为一个介于 0 和 1 之间的数字。

条件支持度。选择**启用过滤器**可激活基于规则的条件支持度水平过滤规则的选项。条件支持度指的是训练数据中与当前规则包含相同条件的比例，因此与普及性指数有点类似。您可以使用**最小**和**最大**控件，根据支持度水平来指定过滤规则的范围。

提升。选择**启用过滤器**可激活基于规则的提升测量量过滤规则的选项。注意：提升过滤只可用于 8.5 版本之后构建的关联模型或之前版本中包含提升测量量的模型。序列模型不包含此选项。

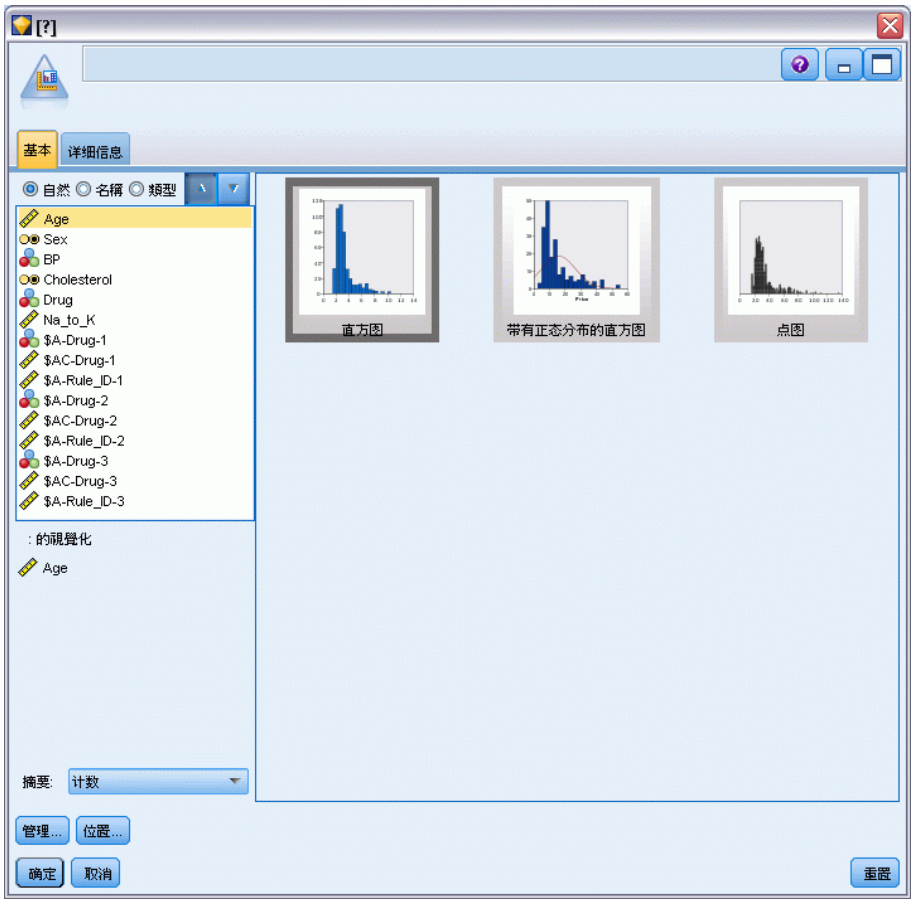
单击**确定**可应用已在此对话框中启用的所有过滤器。

为规则生成图形

关联节点提供了大量信息，但对商业用户来说，它可能并不始终是一种方便访问的格式。要使提供的数据便于纳入商业报表和演示文稿，您可生成所选数据的图形。从“模型”选项卡上，可以为选定规则生成图形，从而只为该规则中的个案创建图形。

- 在“模型”选项卡上，选择感兴趣的规则。
- 从“生成”菜单中，选择**图形（从选定内容）**。显示图形板“基本”选项卡。

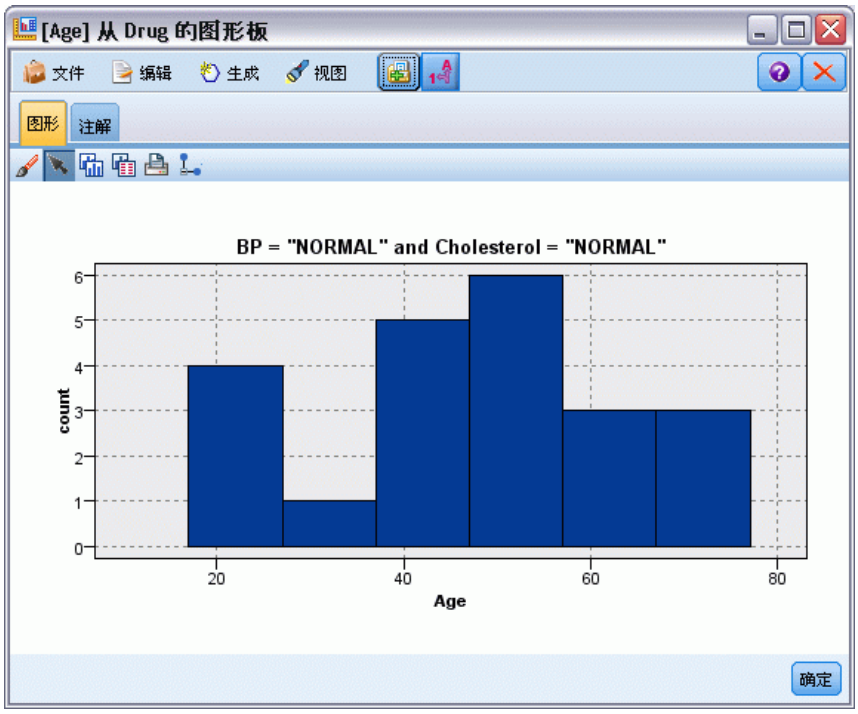
图片 12-13
“图形板”节点对话框，“基本”选项卡



注意：当您以此方式显示“图形板”时，只有“基本”和“详细”选项卡可用。 有关详细信息，请参阅第 5 章中的图形板节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

- 使用“基本”或“详细”选项卡设置指定在图形上显示的详细信息。
- 单击“确定”生成图形。

图片 12-14
“图形板”节点对话框，“基本”选项卡



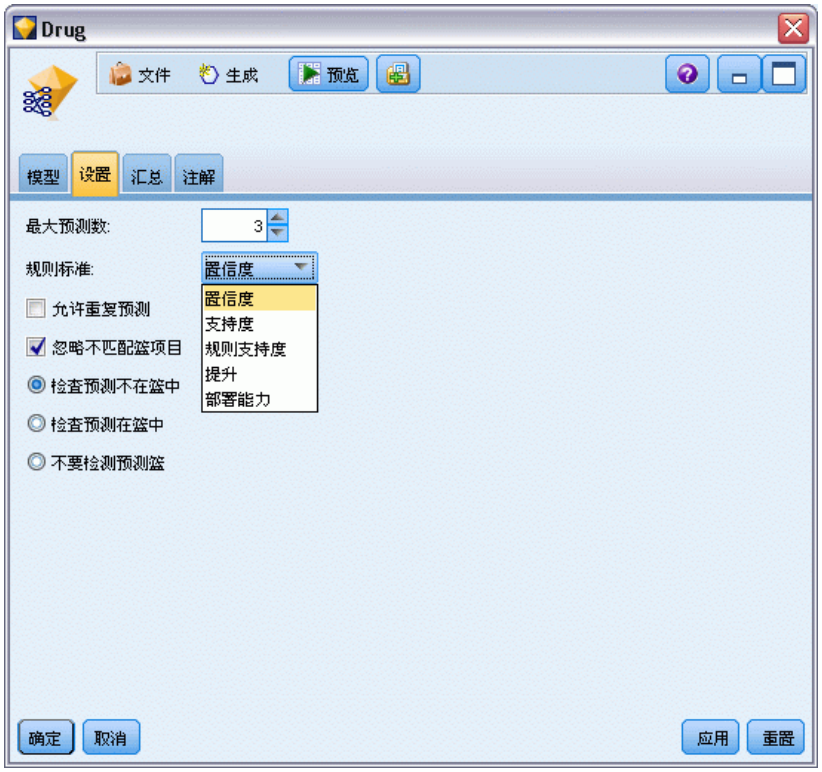
图形标题标识所包含的选定规则和条件详细信息。

关联规则模型块设置

此“设置”选项卡用于为关联模型（Apriori 和 CARMA）指定评分选项。此选项卡仅在模型块添加到用于评分的流后才可用。

注意：用于浏览非精炼模型的对话框不包括“设置”选项卡，因为不能对它进行评分。要对“非精练”模型进行评分，您必须首先生成一个规则集。 有关详细信息，请参阅第 338 页码从关联模型块生成规则集。

图片 12-15
关联规则模型块“设置”选项卡



预测的最大数量。指定每个购物篮项目集合包括的最大预测数。此选项与下面的“规则标准”一起使用可生成“最佳”预测，其中最佳指的是置信度、支持度、提升等的最高水平，如下面的内容所述。

规则标准。选择用于确定规则强度的测量。规则按照此处选择的标准强度进行排序，以便返回项目集合的最佳预测。可用标准有：

- 置信度
- 支持度
- 规则支持度（支持度 * 置信度）
- Lift
- 部署能力

允许重复预测。选择此选项可在评分时包括具有相同结果的多个规则。例如，选择此选项可允许对下列规则进行评分：

```
bread & cheese -> wine
cheese & fruit -> wine
```

关闭此选项可在评分时排除重复的预测。

注意：带有多个结果的规则（bread & cheese & fruit -> wine & pate）仅在所有结果（wine & pate）均在以前经过预测的情况下才会视为重复预测。

忽略不匹配的购物篮项目。选择此选项可忽略项目集合中附加项目的存在。例如，如果对于包含 [tent & sleeping bag & kettle] 的购物篮选择了此选项，规则 tent & sleeping bag -> gas_stove 应用时则会忽略该购物篮中存在的额外项目 (kettle)。

可能存在一些情况应该排除额外的项目。例如，很可能出现这样的情况，某人购买了 tent（帐篷）、sleeping bag（睡袋）和 kettle（水壶），而此人已经拥有了 gas stove（燃气炉），这点通过 kettle（水壶）的存在表明。换句话说，gas stove（燃气炉）可能不是最佳预测。这种情况下，您应该取消选择**忽略不匹配的购物篮项目**以确保规则条件与购物篮内容精确匹配。默认情况下，不匹配的项目将被忽略。

检查购物篮不存在预测值。选择此选项可确保结果也不存在于购物篮中。例如，如果进行评分的目的是为了进行一项家具产品推荐，那么已经包含餐桌的购物篮可能不会购买另一个这样的家具。这种情况下，您应该选择此选项。另一方面，如果产品易腐烂或者是一次性的（如奶酪、婴儿代乳品或者卫生纸），那么其中结果已存在于购物篮的规则可能有些价值。在后面一种情况下，最有用的选项可能是下面的**不检查购物篮中是否存在预测值**。

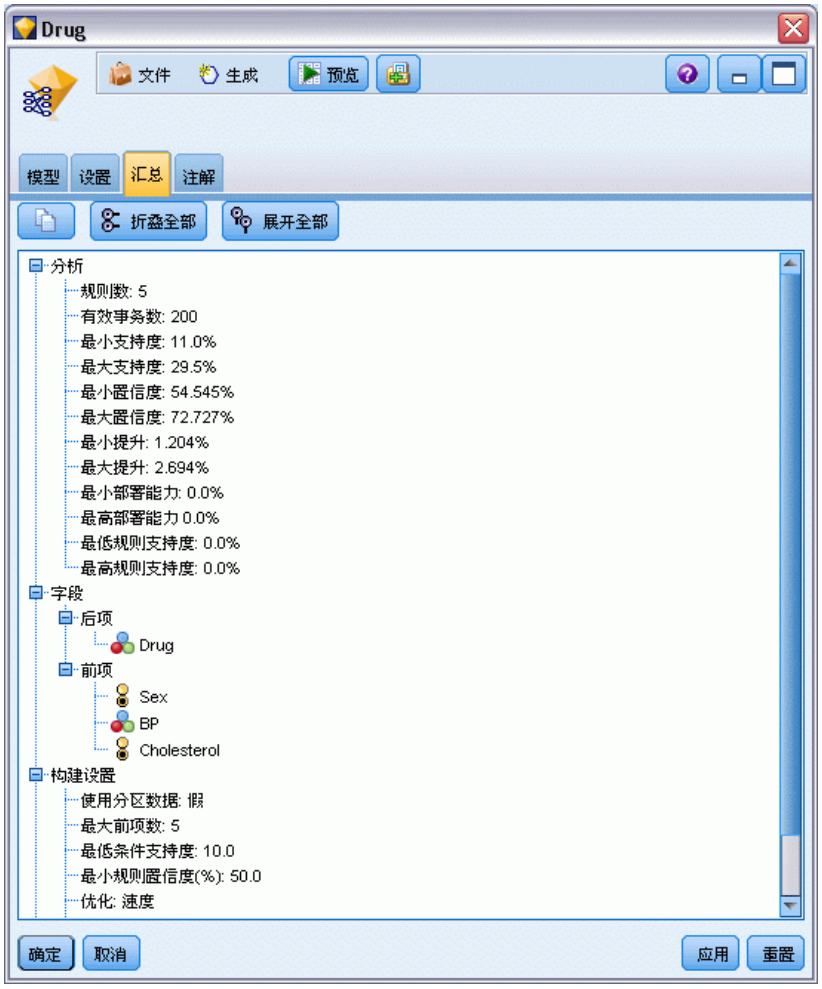
检查购物篮中存在预测值。选择此选项可确保结果也存在于购物篮中。当您尝试深入了解现有的客户或事务时，此方法非常有用。例如，您可能希望确定提升最高的规则，然后探索哪些客户符合这些规则。

不检查购物篮中是否存在预测值。选择此选项可在评分时包括所有规则，而不管购物篮中是否存在结果。

关联规则模型块概要

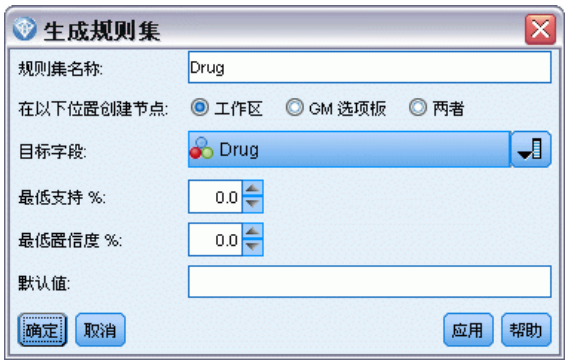
关联规则模型块的“概要”选项卡显示发现的规则数量，以及规则集中规则的最大和最小支持度、提升值、置信度和部署能力。

图片 12-16
关联规则模型块“概要”选项卡



从关联模型块生成规则集

图片 12-17
“生成规则集”对话框



关联模型块（如 Apriori 和 CARMA）可用于直接对数据评分，您也可以首先生成一个规则子集，称为**规则集**。当您对非精练模型进行操作时，因为该模型不能直接用于评分，所以规则集尤其有用。 有关详细信息，请参阅第 62 页码第 3 章中的非精练模型。

要生成规则集，请从模型块浏览器的“生成”菜单中选择**规则集**。您可以指定下列选项，将规则转换为规则集：

规则集名称。使您能够指定新生成规则集节点的名称。

创建节点位置。控制新生成规则集节点的位置。选择**工作区**、**GM 选项板**或**两者**。

目标字段。确定哪个输出字段将用于生成的规则集节点。从列表中选择一个输出字段。

最小支持度。指定生成的规则集中要保留的规则的最小支持度。支持度小于指定值的规则不会包括在新的规则集中。

最小置信度。指定生成的规则集中要保留的规则的最小置信度。置信度小于指定值的规则不会包括在新的规则集中。

默认值。使您能够为分配到不会触发任何规则的已评分记录的目标字段指定默认值。

生成已过滤的模型

图片 12-18
“生成新模型”对话框

要从关联模型块（如 Apriori、CARMA 或序列规则集节点）生成已过滤的模型，请从模型块浏览器的“生成”菜单中选择**已过滤的模型**。这样将创建一个子集模型，其中只包含浏览器中当前显示的那些规则。注意：不能为非精练模型生成已过滤的模型。

您可以指定下列用于过滤规则的选项：

新模型名称。使您能够指定新的已过滤模型节点的名称。

创建节点位置。控制新的已过滤模型节点的位置。选择**工作区**、**GM 选项板**或**两者**。

规则编号。指定规则 ID 在包括在已过滤模型中的规则子集中的编号方式。

- **保留初始规则 ID 号。**选择此选项可保持原始的规则编号。默认情况下，会为规则提供一个与算法发现它们的顺序相对应的 ID。该顺序可能会因所采用算法的不同而有所差别。
- **重新进行连续编号的起始号码。**选择此选项可为过滤的规则指定新的规则 ID。新的 ID 将根据“模型”选项卡上规则浏览器表中显示的排序顺序进行指定，从您在此处指定的数字开始。您可以使用右侧的箭头指定 ID 的开始号码。

关联规则评分

通过关联规则模型块运行新数据生成的得分会返回到不同的字段中。对于每个预测会添加三个新字段，其中 P 表示预测，C 表示置信度，I 表示规则 ID。这些输出字段的排列取决于输入数据是事务格式还是表格格式。请参阅 表格格式数据与事务处理格式数据第 320 页码 大致了解这些格式。

例如，假设您要使用一个基于下面三个规则生成预测的模型对购物篮数据进行评分：

```
Rule_15  bread&wine -> meat (confidence 54%)
Rule_22  cheese -> fruit (confidence 43%)
Rule_5  bread&cheese -> frozveg (confidence 24%)
```

表格数据。对于表格数据，这三个预测（3 为默认值）会返回到一个记录中。

表 12-1
表格格式的得分

ID	Bread	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat	0.54	15	fruit	0.43	22	frozveg	.24	5

事务处理格式数据。对于事务处理格式的数据，对于每个预测都会生成一个单独的记录。预测仍然会添加到单独的列中，但得分在计算时返回。这样会生成带有不完整预测的记录，如下面的示例输出所示。第二个和第三个预测（P2 和 P3）在第一个记录中是空值，同时还会显示相关的置信度和规则 ID。但返回得分时，最后一个记录将包含所有三个预测。

表 12-2
事务处理格式的得分

ID	项目	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	bread	meat	0.54	14	$null$	$null$	$null$	$null$	$null$	$null$
Fred	Cheese	meat	0.54	14	fruit	0.43	22	$null$	$null$	$null$
Fred	wine	meat	0.54	14	fruit	0.43	22	frozveg	0.24	5

要只包括用于报告或部署目的的完整预测，请使用选择节点选择完整的记录。

注意：为了明确起见，在这些示例中使用的字段名称都是缩写。在实际应用中，关联模型的结果字段将按照下列方式命名：

新字段	字段名示例
预测	$A-TRANSACTION_NUMBER-1
置信度（或其他标准）	$AC-TRANSACTION_NUMBER-1
规则 ID	$A-Rule_ID-1

带有多个结果的规则

CARMA 算法允许带有多个结果的规则，例如：

```
bread -> wine&cheese
```

当您对这样的“双头”规则进行评分时，预测将以下表显示的格式返回：

表 12-3
对包括带有多个结果预测的结果进行评分

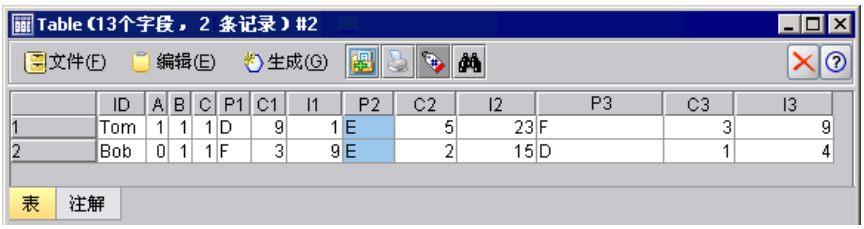
ID	Bread	Wine	Cheese	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	meat&veg	0.54	16	fruit	0.43	22	frozveg	.24	5

在某些情况下，您可能需要在部署之前分割这样的得分。要分割带有多个结果的预测，您需要使用 CLEM 字符串功能解析该字段。 有关详细信息，请参阅第 8 章中的字符串函数中的IBM SPSS Modeler 14.2 用户 指南。

部署关联模型

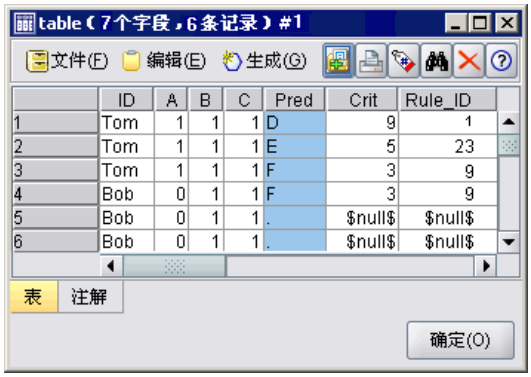
对关联模型进行评分时，预测和置信度将输出到单独的列中（其中 P 表示预测，C 表示置信度，I 表示规则 ID）。这种情况要区分输入数据是表格格式还是事务格式。 有关详细信息，请参阅第 340 页码关联规则评分。

图片 12-19
预测位于列中的表格得分



准备得分进行部署时，您可能会发现您的应用程序需要将输出数据转换为预测位于行中的格式，而不是位于列中的格式（每行一个预测，有时称为“行穷尽”格式）。

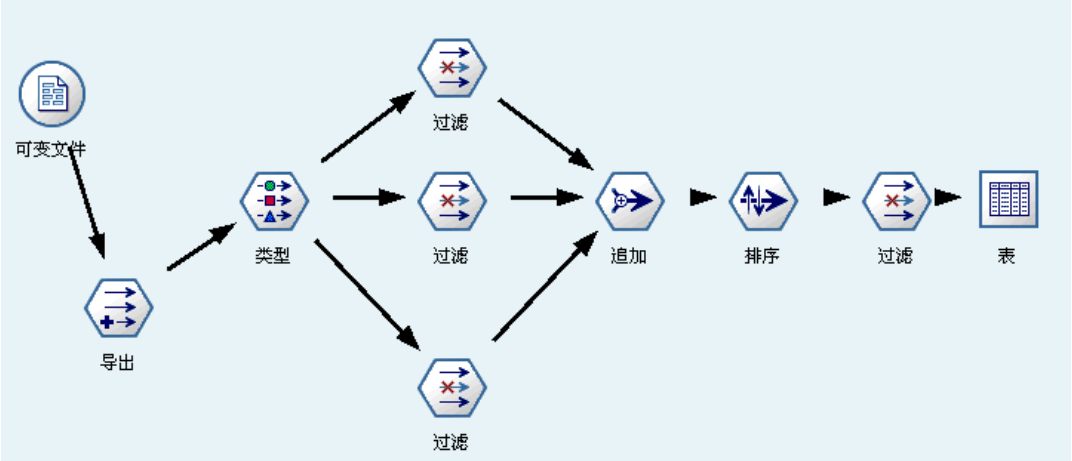
图片 12-20
预测位于行中的已转置得分



转置表格得分

您可以使用 IBM® SPSS® Modeler 中的一些步骤将表格得分从列转置为行，如下面的步骤所示。

图片 12-21
用于将表格数据转置为行穷尽格式的流示例



- 在导出节点中使用 @INDEX 函数可确定预测的当前顺序，并将此指示保存在一个新字段中，如 Original_order。
- 添加一个类型字段，确保所有字段均实例化。
- 使用过滤节点将默认的预测、置信度和 ID 字段（P1、C1、I1）重命名为普通字段，如 Pred、Crit 和 Rule_ID，这些字段将用于在以后追加记录。对于每个生成的预测都需要一个过滤节点。

图片 12-22
重命名预测 2 的字段时过滤预测 1 和预测 3 的字段。

- 使用追加节点追加共享 Pred、Crit 和 Rule_ID 的值。

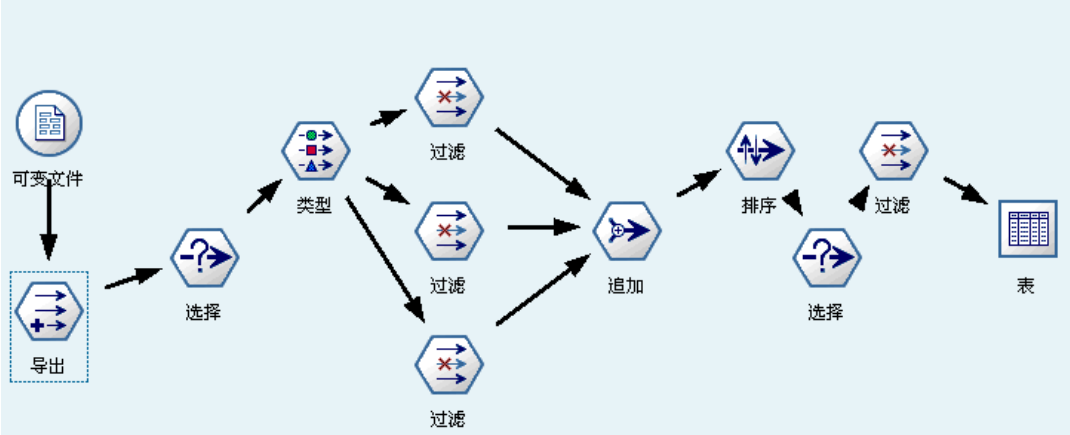
- 连接一个排序节点，以便按照字段 Original_order 的升序对记录进行排序，按照 Crit 的降序对记录进行排序，后面一个字段是用于按标准（如置信度、提升和支持度）对预测进行排序的字段。
- 使用另一个过滤节点将字段 Original_order 从输出中过滤掉。

此时，数据就可以进行部署了。

转置事务得分

转置事务得分的过程与上面的过程相似。例如，下面显示的流会根据部署需要，将得分转置为每行一个预测的格式。

图片 12-23
用于将事务处理格式的数据转置为行穷尽格式的流示例



除了添加两个选择节点之外，该过程与前面用于表格数据的过程完全相同。

- 第一个选择节点用于对相邻记录的规则 ID 进行比较，以便只包括唯一的或非精练的记录。此选择节点使用该 CLEM 表达式选择记录： `ID /= @OFFSET(ID,-1) or @OFFSET(ID,-1) = undef`.
- 第二个选择节点用于放弃多余的规则，或者 Rule_ID 为 Null 值的规则。此选择节点使用下列 CLEM 表达式放弃记录： `not(@NULL(Rule_ID))`.

有关转置得分进行部署的详细信息，请联系技术支持部门。

序列节点

序列节点会发现连续数据或面向时间的数据中的模式，其格式为 bread -> cheese。序列的元素为组成一个事务的**项目集合**。例如，如果某人进入商店，购买了面包和牛奶，几天之后返回了该商店，购买了一些奶酪，那么这个人的购买活动可以表示为两个项目集合。第一个项目集合包含面包和牛奶，第二个包含奶酪。**序列**是一系列可能会以可预测顺序发生的项目集合。序列节点会检测频繁出现的序列，并创建一个可用于生成预测的生成模型节点。

要求。要创建序列规则集，您需要指定一个 ID 字段以及一个可选的时间字段，以及一个或多个内容字段。请注意，这些设置必须在建模节点的“字段”选项卡上进行；不能从上游“类型”节点中读取。该 ID 字段可以是任意角色或测量级别。如果指定时间字段，则该字段可以是任意角色，但其存储必须是数字、日期、时间或时间戳。如果不指定时间字段，序列节点则会使用隐含的时间戳，实际上是使用行号作为时间值。内容字段可具有任意测量级别和角色，但所有内容字段的类型必须相同。如果这些字段是数字型的，则必须为整数范围（不是实数范围）。

强度。序列节点基于 CARMA 关联规则算法，该算法使用一个有效的两次传递方法查找 ��列。另外，序列节点创建的生成的模型节点可以插入到数据流中来创建预测。生成的模型节点还可生成超节点用于检测或计数特定的序列，以及基于特定的序列作出预测。

序列节点字段选项

图片 12-24
序列节点字段选项

执行序列节点之前，必须在序列节点的“字段”选项卡上指定 ID 字段和内容字段。如果您要使用时间字段，也需要在此处指定。

ID 字段。从列表中选择 ID 字段。数字字段或符号字段可用作 ID 字段。此字段的每个唯一值都应该表明一个特定的分析单元。例如，在市场购物篮的应用中，每个 ID 可能表示一个客户。对于 Web 日志分析应用，每个 ID 可能代表一个计算机（以 IP 地址表示）或一个用户（以登录数据表示）。

- **ID 是连续的。**如果您的数据进行了预先排序，以便所有 ID 相同的记录在数据流中分组在一起，那么选择此选项可以加快处理速度。如果您的数据未经预先排序（或者您不确定），请将此选项保持不选中状态，序列节点将自动对该数据进行排序。

注意：如果您的数据未经排序而您选择了此选项，则可能会在序列模型中得到无效结果。

时间字段。如果您要在数据中使用字段来表明事件时间，请选择**使用时间字段**并指定要使用的字段。时间字段必须是数字、日期、时间或时间戳型的如果不指定时间字段，则假设记录按照从数据源出发的顺序到达，记录号将用作时间值（第一个记录发生在时间 "1"；第二个记录发生在时间 "2"；依此类推）。

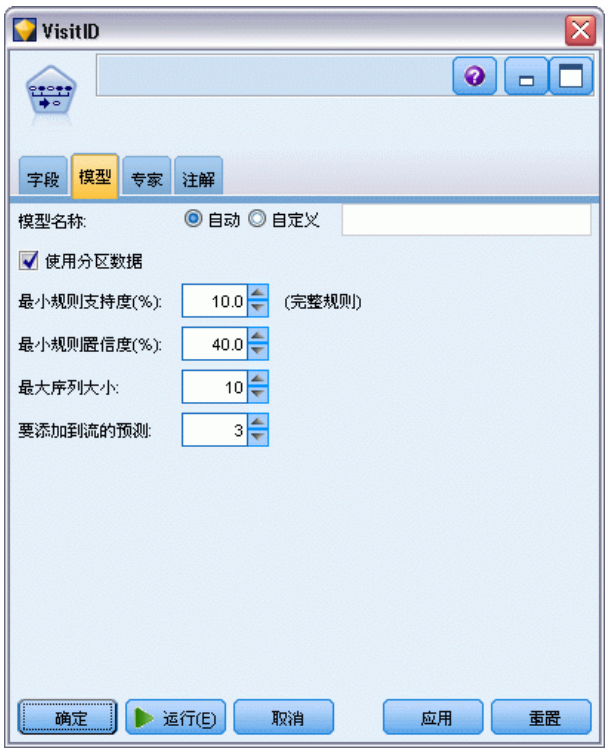
内容字段。指定模型的内容字段。这些字段包含与序列建模有关的事件。

序列节点可以处理表格格式的数据，也可以处理事务格式的数据。如果您使用带有事务处理格式的数据的多个字段，则认为在某个特定记录中，这些字段中指定的项目表示着可以在一个带有时间戳的事务中找到的项目。 有关详细信息，请参阅第 320 页码表格格式数据与事务处理格式数据。

分区。该字段允许您使用指定字段将数据分割为几个不同的样本，分别用于模型构建过程中的训练、测试和验证阶段。通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。如果已使用“类型”或“分区”节点定义了多个分区字段，则必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。（如果仅有一个分区字段，则将在启用分区后自动引入此字段。）有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。同时请注意，要在分析时应用选定分区，同样必须启用节点“模型选项”选项卡中的分区功能。（取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）

序列节点模型选项

图片 12-25
序列节点模型选项



模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

最低规则支持度（%）。您还可以指定支持度标准。**规则支持度**指的是训练数据中包含整个序列的 ID 的比例。如果您要关注更常见的序列，请增加此设置。

最低规则置信度（%）。您可以指定在序列集中保留序列的置信度标准。**置信度**指的是得到正确预测的 ID 在所有使用规则进行预测的 ID 中所占的百分比。基于训练数据，该百分比的计算如下：包含整个序列的 ID 数量除以其中包含条件的 ID 数量。置信度低于指定标准的序列将被放弃。如果您获得的序列太多或者不是非常相关，请尝试增加此设置。如果您获得的序列太少，请尝试降低此设置。

最大序列容量。您可以设置序列中不同项目集合（与项目 不同） 的最大数量。如果相关序列相对较短，则可以降低此设置，以加快序列集构建速度。

要添加到流的预测。指定生成的结果模型节点要添加到流中的预测数量。 有关详细信息，请参阅第 348 页码序列模型块。

序列节点专家选项

对于那些详细了解序列节点操作的人员来说，通过下列专家选项可以对建模过程进行微调。要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

图片 12-26
序列节点专家选项

设置最大持续时间。如果选择了此选项，序列将被限制为小于或等于指定值的一个持续时间（第一个项目集合和最后一个项目集合之间的时间）。如果没有指定时间字段，该持续时间则以原始数据中的行数（记录数）表示。如果使用的时间字段为时间、日期或时间戳型字段，该持续时间则表示为秒数。对于数字字段，持续时间则使用与字段相同的单位数表示。

设置修剪值。为了节省内存，序列节点中使用的 CARMA 算法会在处理期间定期从其潜在项目集合列表中删除（**修剪**）不常使用的项目集合。选择此选项可调整修剪的频率。指定的数字决定了修剪频率。输入较小的值可降低该算法的内存要求（但可能会延长所需的训练时间），输入较大的值会加快训练速度（但可能会提高内存要求）。

设置内存中的最大序列。如果选择了此选项，CARMA 算法则会将建模期间备选序列的内存存储限制为指定的序列数。如果 IBM® SPSS® Modeler 在序列建模期间使用的内存过多，请选择此选项。请注意，您在此处指定的最大序列值指的是在构建模型期间进行内部跟踪的备选序列数。此数字应该比最终模型中预期的序列数大很多。

项目集之间的约束间距。通过此选项可以针对不同项目集合的时间间距指定约束。如果选择了此选项，则不会考虑时间间距小于您所指定的**最小间距**或大于**最大间距**的项目集合作为序列的组成部分。使用此选项可避免考虑包括较长时间区间或者在很短的时间跨度内发生的那些序列。

注意：如果使用的时间字段为时间、日期或时间戳型字段，该时间间距则表示为秒数。对于数字型字段，时间间距则使用与时间字段相同的单位数表示。

例如，以下面的事务列表为例：

ID	Time	内容
1001	1	apples
1001	2	bread
1001	5	Cheese
1001	6	dressing

如果您针对这些数据建模时指定的最小间距为 2，则会得到下列序列：

apples -> cheese

apples -> dressing

bread -> cheese

bread -> dressing

您不会看到像 apples -> bread 这样的序列，因为 apples 和 bread 之间的间距小于最小间距。与之相似，如果是下面的数据：

ID	Time	内容
1001	1	apples
1001	2	bread
1001	5	Cheese
1001	20	dressing

并且最大间距设置为了 10，您则不会看到任何带有 dressing 的序列，因为 cheese 和 dressing 之间的间距太大，不考虑它们作为同一序列的组成部分。

序列模型块

“序列”模型块表示“序列”节点针对某个特定输出字段发现的序列，可以添加到流中以生成预测。

当您运行包含“序列”节点的流时，“序列”节点会将包含预测的一对字段，以及序列模型中每个预测的相关置信度值添加到数据中。默认情况下，会添加包含三个最佳预测的三对字段（以及它们相关联的置信度值）。您既可以通过在构建时设置序列节点模型选项更改构建模型时生成的预测数，也可以在将模型块添加到流之后在“设置”选项卡上更改此数量。 有关详细信息，请参阅第 352 页码序列模型块设置。

新的字段名称派生自模型名称。预测字段的字段名称为 $S-sequence-n（其中 n 表示第 n 个预测）置信度字段的字段名称为 $SC-sequence-n。在一个序列中具有多个序列规则节点的流中，新的字段名称将包括数字前缀，以便将它们区别开来。流中的第一个序列集节点使用普通的名称，第二个节点将使用以 $S1- 和 $SC1- 开头的名称，第三个节点使用以 $S2- 和 $SC2- 开头的名称，依此类推。预测按照置信度的顺序显示，因此 $S-sequence-1 所包含预测的置信度最高，$S-sequence-2 所包含预

测的置信度次高，依此类推。对于其中可用预测数量小于所请求预测数量的记录，剩余的预测包含值 $null$。例如，如果对于某个特定的记录只能进行两个预测，那么 $S-sequence-3 和 $SC-sequence-3 的值将为 $null$。

对于每个记录，会将模型中的规则与目前对于当前 ID 已经处理的事务集合（包括当前记录和具有相同 ID 和较早时间戳的所有以前记录）进行比较。将使用适用于此事务集合的、置信度值最高的 k 个规则为该记录生成 k 个预测，其中 k 为模型添加到流之后在“设置”选项卡上指定的预测数。（如果多个规则对于该事务集合预测了相同的结果，则只使用置信度最高的规则。） 有关详细信息，请参阅第 352 页码序列模型块设置。

与其他类型的关联规则模型相同，数据格式必须与构建序列模型时使用的格式相匹配。例如，使用表格数据构建的模型只能用于对表格数据进行评分。 有关详细信息，请参阅第 340 页码关联规则评分。

注意：在流中使用生成的序列集节点对数据进行评分时，您在建模时选择的任何容差或间距设置都将被忽略，不会用于评分目的。

根据序列规则进行的预测

该节点以与时间相关（如果在构建模型时未使用时间戳字段的话，则与顺序相关）的方式处理记录。记录应该按照 ID 字段和时间戳字段（如果存在的话）排序。但是，预测与添加到其中的记录的时间戳没有关系。它们只是在给出到当前记录为止当前 ID 的事务历史的情况下，指出最可能在将来的某个时间出现的项目。

请注意，每个记录的预测不一定与该记录的事务相关。如果当前记录的事务不触发某个特定的规则，则会根据当前 ID 的以前事务选择规则。换句话说，如果当前记录不向序列添加任何有用的预测信息，则会将此 ID 的最后一个有用事务中的预测转到当前记录。

例如，假设您拥有的序列模型具有一个规则

```
Jam -> Bread (0.66)
```

然后您将其传递到了下列记录：

ID	采购	预测
001	jam	bread
001	milk	bread

请注意，与您的预期相同，第一个记录生成了预测 bread。第二个记录也包含 bread 预测，因为没有规定 jam 后紧跟 milk；因此，milk 事务不会增加任何有用信息，所以规则 Jam -> Bread 仍然适用。

生成新节点

通过“生成”菜单可以基于序列模型创建新的超节点。

- **规则超节点。**创建一个可以检测和计算已评分数据中序列发生次数的超节点。如果未选择任何规则，此选项则禁用。 有关详细信息，请参阅第 353 页码从序列模型块生成规则超节点。
- **模型到调色板。**将模型返回到模型选项板。当有同事发给您包含模型的流而不是模型本身时，该功能很有用。

序列模型块详细信息

序列模型块的“模型”选项卡显示算法提取的规则。表中的每行都代表一个规则，其中条件（规则“if”部分）位于第一列，结果（规则的“then”部分）位于后面的第二列。

图片 12-27
序列模型块“模型”选项卡

每个规则都以下列格式显示：

条件	结果
beer and cannedveg	beer
fish fish	fish

第一个规则示例解释为对于在同一个事务中具有“beer”和“cannedveg”的 ID，很可能后面会出现“beer”。第二个规则示例可以解释为对于在一个事务中具有“fish”，在另一个事务中也是“fish”的 ID，很可能后面会出现“fish”。请注意在第一个规则中，beer 和 cannedveg 是同时购买的；在第二个规则中，fish 是在两个不同的事务中购买的。

排序菜单。工具栏上的“排序”菜单按钮控制着规则的排序。排序的方向（升序或降序）可以使用排序方向按钮（上箭头或下箭头）进行更改。

图片 12-28
用于排序的工具栏选项

您可以按照下列条件对规则进行排序：

- 支持度 %
- 置信度 %
- 规则支持 %
- 结果
- 第一个条件
- 最后一个条件
- 项目数（条件）

例如，下表按照项目数，以降序进行排序。条件集合中具有多个项目的规则排在条件集合中项目数较少的规则前面。

条件	结果
beer and cannedveg and frozenmeal	frozenmeal
beer and cannedveg	beer
fish fish	fish
softdrink	softdrink

显示/隐藏标准菜单。显示/隐藏标准菜单按钮（网格图标）控制着规则的显示选项。可用的显示选项如下：

- **实例数**显示其中发生完整序列（有条件也有结果）的唯一 ID 的数量的相关信息。（请注意，此内容与关联模型不同，后者的实例数指的是其中仅 条件适用的 ID 数。例如，假设规则为 bread -> cheese，训练数据中同时包含 bread 和 cheese 的 ID 数称为**实例数**。
- **支持度**显示训练数据中条件为真的 ID 的比例。例如，如果 50% 的训练数据中包括条件 bread，那么规则 bread -> cheese 的支持度为 50%（与关联模型不同，支持度不基于实例数，如前面所述）。
- **置信度**显示的是得到正确预测的 ID 在所有使用规则进行预测的 ID 中所占的百分比。基于训练数据，该百分比的计算如下：包含整个序列的 ID 数量除以其中包含条件的 ID 数量。例如，如果 50% 的训练数据包含 cannedveg（表明条件支持），但只有 20% 既包含 cannedveg 又包含 frozenmeal，则规则 cannedveg -> frozenmeal 的置信度为 Rule Support / Antecedent Support，在这里为 40%。
- 序列模型的**规则支持度**基于实例数，显示其中整个规则、条件和结果均为真的训练记录的比例。例如，如果 20% 的训练数据既包含 bread 也包含 cheese，那么规则 bread -> cheese 的规则支持度为 20%。

请注意，这些比例基于有效事务（至少具有一个观测项或真值的事务），而不基于总的事务。在这些计算中不会考虑无效事务（没有项目或真值的事务）。

过滤按钮。菜单上的“过滤器”按钮（漏斗图标）会扩展对话框的底部，从而显示一个面板，其中将显示活动的规则过滤器。过滤器用于减少“模型”选项卡上显示的规则数量。

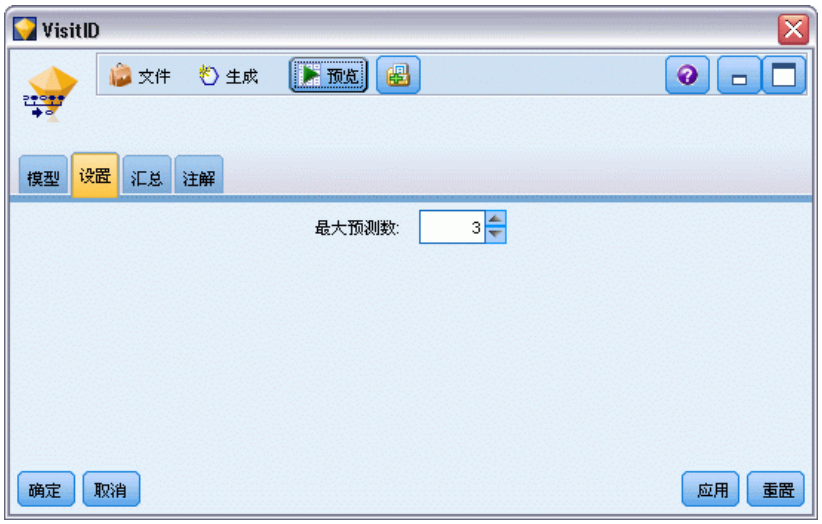
图片 12-29
过滤按钮

要创建过滤器，请单击位于扩展面板右侧的过滤器图标。这样将打开一个单独的对话框，您可以在其中指定用于显示规则的约束条件。请注意，“过滤器”按钮通常与“生成”菜单一起使用，以便首先过滤规则，然后生成一个包含部分规则的模型。有关详细信息，请参阅下面的 为规则指定过滤器。

序列模型块设置

序列模型块的“设置”选项卡显示模型的评分选项。此选项卡仅在模型添加到流工作区用于评分之后可用。

图片 12-30
序列模型块“设置”选项卡

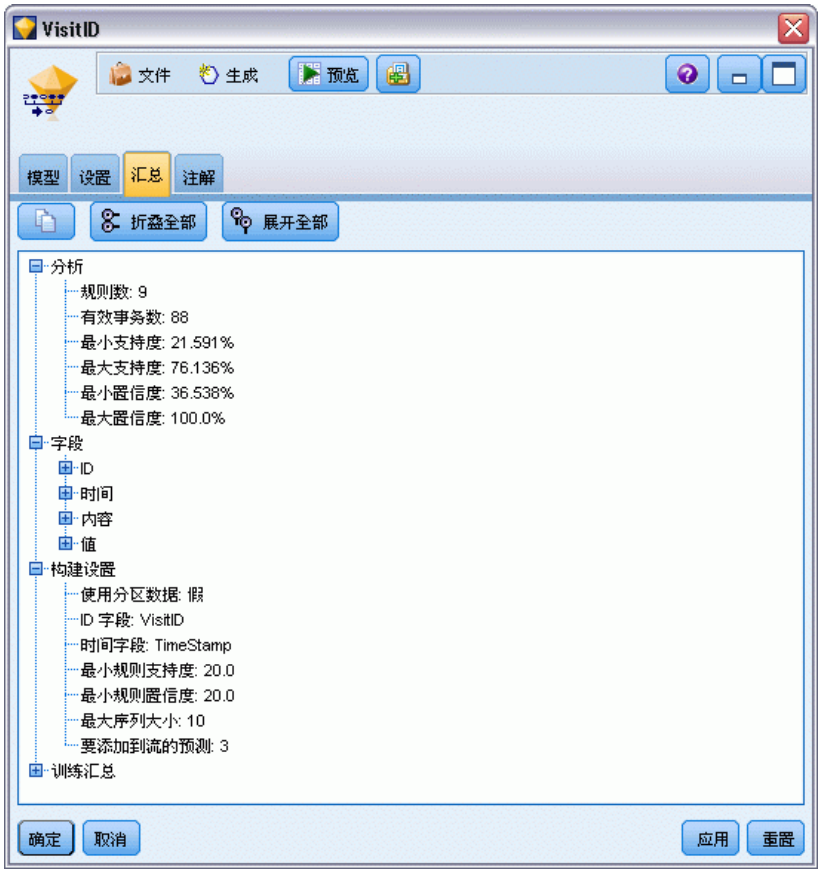


预测的最大数量。指定每个购物篮项目集合包括的最大预测数。适用于此事务集合的、置信度值最高的规则将用于为记录生成预测，预测的数量不超过指定的上限。

序列模型块概要

序列规则模型块的“概要”选项卡显示发现的规则数量，以及规则的最大和最小支持度和置信度。如果已执行附加到此建模节点的分析节点，则分析信息也将显示在此选项卡上。 有关详细信息，请参阅第 6 章中的分析节点中的IBM SPSS Modeler 14.2源、过程和输出节点。

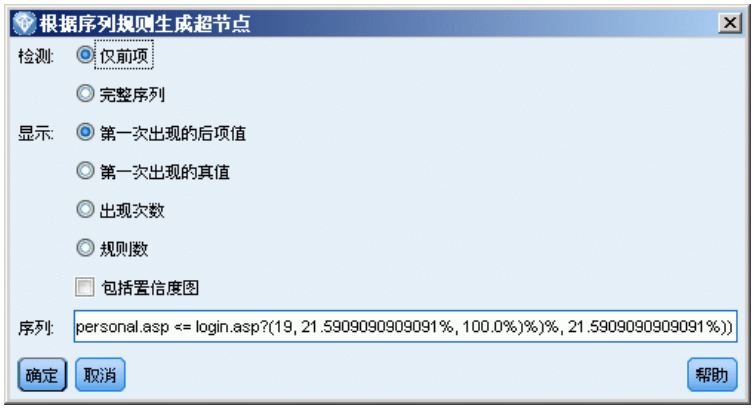
图片 12-31
序列模型块“概要”选项卡



有关详细信息，请参阅第 43 页码第 3 章中的浏览模型块。

从序列模型块生成规则超节点

图片 12-32
“生成规则超节点”对话框



要基于序列规则生成规则超节点：

- 在序列规则模型块的“模型”选项卡上，单击表中的某行以选择所需的规则。
- 从规则浏览器菜单中选择：
  生成 > 规则超节点

  重要事项：要使用生成的超节点，必须在将数据传递到超节点之前按照 ID 字段（和时间字段，如果有的时间字段的话）对数据进行排序。超节点无法在未排序的数据中正确检测序列。

  您可以指定下列用于生成规则超节点的选项：

  检测。指定传递到超节点的数据的匹配定义方式。

  - **仅条件。**每当超节点在具有相同 ID 的一系列记录中发现以正确顺序排列的选中规则的条件时，它都会确定一个匹配，而不管是否同时发现了结果。请注意，此选项不考虑原始序列建模节点中的时间戳容差或项目间距约束设置。在流中检测到最后一个条件项目集合（所有其他条件均以正确顺序发现）后，具有当前 ID 的所有后续记录都将包含下面选择的概要。
  - **整个序列。**每当超节点在具有相同 ID 的一系列记录中发现以正确顺序排列的选中规则的条件和结果时，它都会确定一个匹配。此选项不考虑原始序列建模节点中的时间戳容差或项目间距约束设置。在流中检测到最后一个结果（所有条件均以正确顺序发现）后，当前记录和具有当前 ID 的所有后续记录都将包含下面选择的概要。

  显示。控制匹配概要将以何种方式添加到规则超节点输出的数据中。

  - **首次发生的结果值。**添加到数据中的值为基于第一次发生的匹配预测的结果值。这些值将作为一个名为 rule_n_consequent 的新字段进行添加，其中 n 为规则编号（基于流中规则超节点的创建顺序）。
  - **首次发生的真值。**如果对于该 ID 至少存在一个匹配，添加到数据的值为真；如果没有任何匹配，添加的值则为假。这些值将作为一个名为 rule_n_flag 的新字段添加。
  - **计数发生次数。**添加到数据的值为该 ID 的匹配数。这些值将作为一个名为 rule_n_count 的新字段添加。
  - **规则编号。**添加的值为选中规则的规则编号。**规则编号**是基于超节点添加到流的顺序指定的。例如，第一个规则超节点被视为规则 1，第二个规则超节点被视为规则 2，依此类推。当您要在流中包括多个规则超节点时，此选项最有用。这些值将作为一个名为 rule_n_number 的新字段添加。
  - **包括置信度图表。**如果选中此选项，则会将规则置信度以及选定概要添加到数据流中。这些值将作为一个名为 rule_n_confidence 的新字段添加。

章
13

时间序列模型

为什么要进行预测？

预测的意思就是对一个或多个序列在一定时间内的值进行预言。例如，您可能希望预测某个系列产品或服务的预期需求，以便分配资源进行制造或配送。因为计划决策的实施需要时间，所以预测在很多计划过程中都是一个必不可少的工具。

时间序列建模方法假定历史总会自我重演 — 即使不是完全一样也会非常接近，足以通过研究过去对将来作出更好的决策。例如，为了预测下一年的销售量，您可能得从分析今年的销售量开始，看看近年来都有哪些发展趋势或模式（如果存在的话）。但模式可能很难测量。例如，如果您的销售量在几周之内连续上升，那么这是季节性原因呢还是一种长期趋势的开始？

使用统计建模技术，可以分析过去数据中存在的模式并加以预测，以确定该序列的未来值可能属于的范围。其结果是您的决策所依据的预测更为准确。

时间序列数据

时间序列是以规律的时间间隔采集的测量值的有序集合，例如，每日的股票价格或每周的销售数据。测量值可以是您感兴趣的任何内容，每个序列通常可以归为下列类别之一：

- **依存量。**要预测的序列。
- **预测变量。**可能有助于解释目标的序列 — 例如，使用广告预算来预测销售量。预测变量只能用于 ARIMA 模型。
- **事件。**一种特殊的预测变量序列，用于说明可预测的重复发生事件 — 例如，促销活动。
- **干预。**一种特殊的预测变量序列，用于说明过去的一次性事件 — 例如，停电或员工罢工。

时间间隔可以代表任何时间单位，但所有测量值的时间间隔必须相同。而且，没有测量值的任何时间间隔必须设置为缺失值。因此，有测量值的时间间隔数（包括测量值为缺失值的时间间隔）定义数据历史范围的时间长度。

时间序列的特征

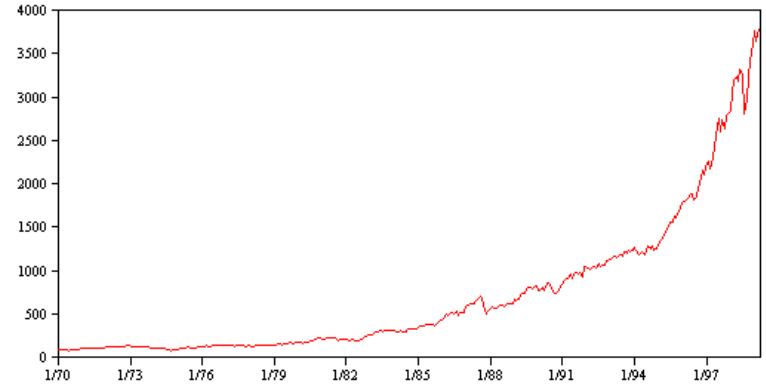
研究序列过去的行为有助于辨别其中的模式从而作出更好的预测。将其绘制成图时，许多时间序列就会表现出下列一种或多种特征：

- 趋势
- 季节周期和非季节周期
- 脉冲和步进
- 界外值

趋势

趋势是指序列水平的逐渐上升或下降或序列值随时间的推移而增大或减小的趋势。

图片 13-1
趋势



趋势既可以是**局部**的，也可以是**全局**的，而一个序列可以同时体现这两种趋势。从历史来看，股票市场指数的序列图总的趋势是上升的。经济萧条时期所表现出的是局部下降趋势，而经济繁荣时期表现出的是局部上升趋势。

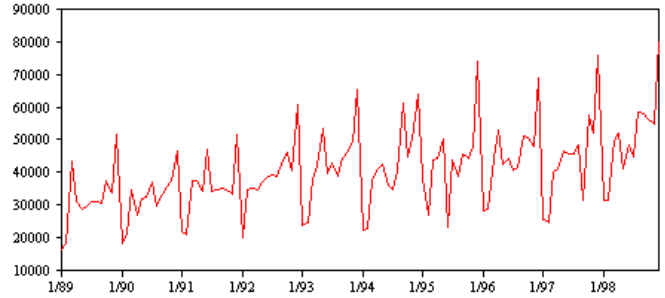
趋势既可以是**线性**的，也可以是**非线性**的。线性趋势是指序列水平表现为正增加或负增加，就和本金以单利计息差不多。非线性趋势通常表现为倍增，即相对于以前的序列值成比例地增长。

全局线性趋势可通过指数平滑模型和 ARIMA 模型很好地拟合和预测。在构建 ARIMA 模型的过程中，通常会对表现出趋势的序列进行区分，以消除趋势的影响。

季节周期

季节周期是序列值中可预测的重复模式。

图片 13-2
季节周期



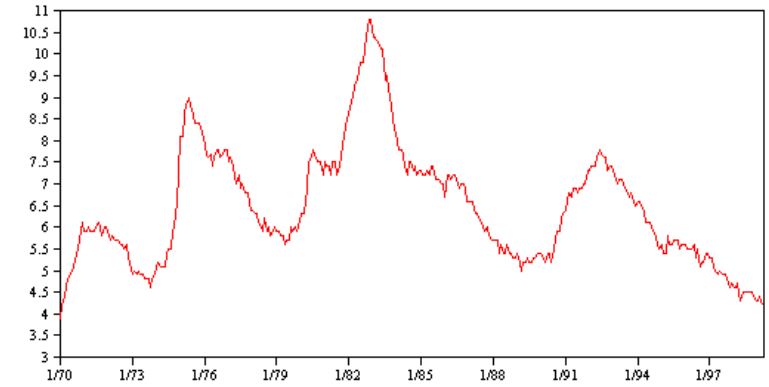
季节周期与序列的时间间隔相联系。例如，月度数据通常会随季度和年度而循环。月度序列可能会表现出第一个季度较低的明显季度周期或每年十二月份都出现峰值的年度周期。表现出季节周期的序列称之为具有**季节性**。

季节模式对于获取良好的拟合和预测非常有用，用来捕获季节性的有指数平滑模型和 ARIMA 模型。

非季节周期

非季节周期是序列值中可能无法预测的重复模式。

图片 13-3
非季节周期



某些序列（如失业率）明显地表现出周期性行为；但这种周期性的周期会随时间而变化，因此很难预测何时高何时低。其他序列可能具有可预测的周期，但可能与阳历并不完全吻合，或者其周期比一年长。例如，潮汐遵循阴历，与奥林匹克运动会相关的国际旅游和贸易每隔四年膨胀一次，还有许多宗教节日，其阳历日期每年都会变化。

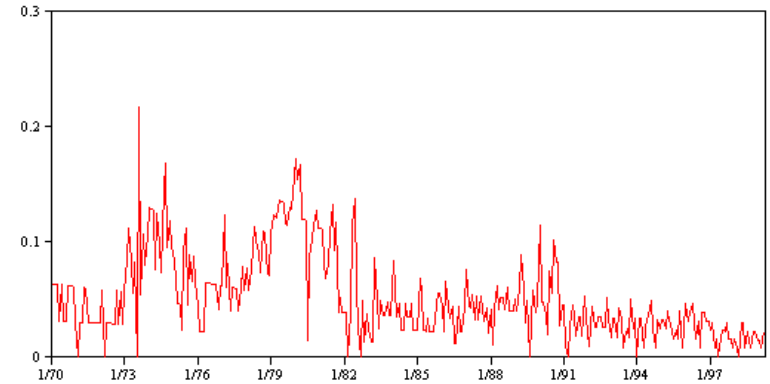
非季节周期模式很难建模，通常会增加预测的不确定性。例如，股票市场的许多序列实例就常使预测者的努力无功而返。即便如此，当存在非季节模式时，还是有必要加以说明。在许多情况下，您仍然可以找出与历史数据拟合得很好的模型，从而最大限度地减小预测中的不确定性。

脉冲和步进

许多序列都会出现水平突变。它们通常分为两种类型：

- 序列水平突然、临时性的变动，或称**脉冲**
- 序列水平突然、永久性的变动，或称**步进**

图片 13-4
脉冲序列



观测到步进或脉冲时，找到一种貌似合理的解释很重要。时间序列模型是用来说明渐变而非突变的。因此，它们往往低估脉冲并为步进所瓦解，导致模型拟合差强人意，增加预测的不确定性。（某些季节性实例可能表现为突然的水平变化，但该水平在不同的季节周期之间则保持稳定。）

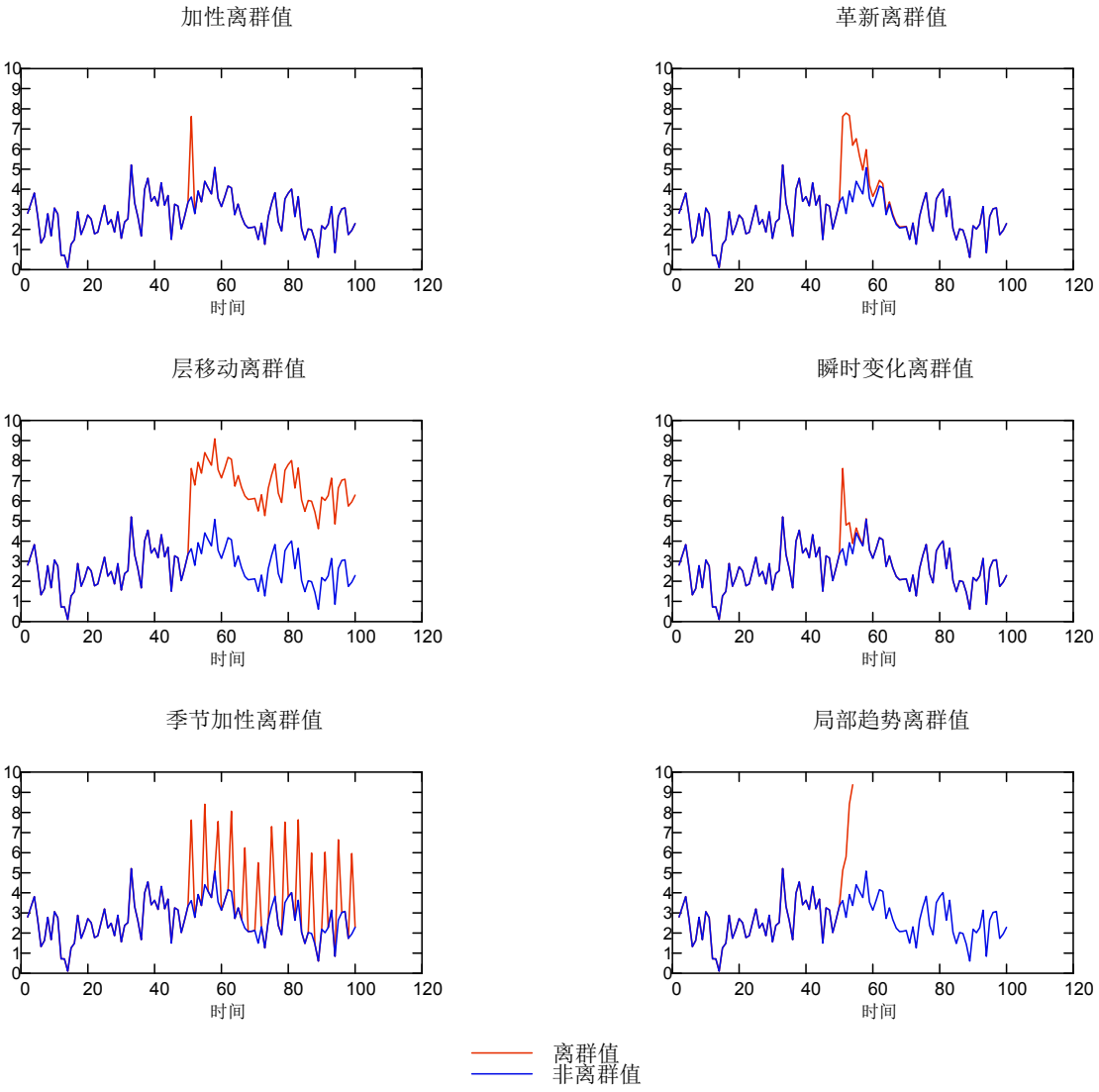
如果扰动是可以解释的，则可以使用**干预**或**事件**为其建模。例如，1973 年 8 月，石油输出国组织 (OPEC) 颁布的石油禁运导致了通货膨胀率的急剧变化，经过数月之后才恢复到正常水平。通过为该禁运月指定一个**点干预**，可以改善模型的拟合度，因此可以间接提高预测的准确性。例如，某个零售店可能会发现，所有商品均标记降价 50% 的当天销售量比平时高出很多。通过将降价 50% 的促销指定为一个定期的**事件**，可以改善模型的拟合度，估计将来重复该项促销措施的影响。

离群值

时间序列水平中无法解释的变动称为**离群值**。这些观测值与序列中的其他值不一致，可能会显著影响分析，从而影响时间序列模型的预测能力。

下图显示了时间序列中常见的几种离群值。蓝线表示没有离群值的序列。红线表示如果序列包含离群值情况下可能存在的模式。这些离群值全部归为**确定性**离群值，因为它们只影响序列的均值水平。

图片 13-5
离群值类型



- **加性离群值。**加性离群值表现为一次观测中出现的异常大或异常小的值。后续观测不受加性离群值的影响。连续的加性离群值通常称为**加法离群值修补**。
- **革新离群值。**革新离群值的特征为初始影响一直对后续观测产生作用。这些离群值的影响可能会随着时间的推移而不断增强。
- **水平变动离群值。**对于水平变动，离群值之后出现的所有观测值均会移到一个新的水平。与加性离群值相反，水平变动离群值会影响许多观测值，并且具有永久性影响。
- **瞬时变化离群值。**瞬时变化离群值类似水平变动离群值，只是这种离群值对后续观测的影响呈指数递减。最终，该序列会恢复到正常水平。
- **季节加性离群值。**季节加性离群值表现为以固定时间间隔重复出现的异常大或异常小的值。

- **局部趋势离群值。**局部趋势离群值会在出现初始离群值之后，在序列中产生一个由离群值中的模式所导致的整体漂移。

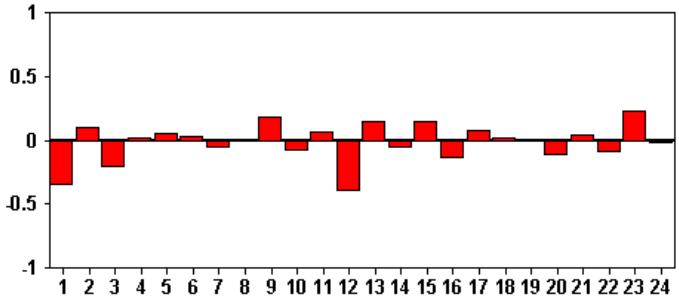
时间序列中的离群值检测包括确定存在的任何离群值的位置、类型和大小。Tsay (1988) 提出了一个用于检测均值水平变化以识别出确定性离群值的迭代过程。此过程是将一个假设不存在离群值的时间序列模型与另一个具有离群值的模型进行比较。从两个模型之间的差异得到将任何给定点视为离群值的影响的估计。

自相关函数和部分自相关函数

自相关和部分自相关是当前序列值和过去序列值之间关联度的测量，表明在预测将来值时过去的哪些序列值最有用。了解了此内容，您就可以确定 ARIMA 模型中过程的顺序。更具体来说，

- **自相关函数 (ACF)。**在延迟为 k 时，这是相距 k 个时间间隔的序列值之间的相关。
- **部分自相关函数 (PACF)。**在延迟为 k 时，这是相距 k 个时间间隔的序列值之间的相关，同时考虑了两个间隔之间的值。

图片 13-6
序列的 ACF 图



ACF 图的 x 轴表示计算自相关处的延迟；y 轴表示相关值（介于 −1 和 1 之间）。例如，ACF 图中延迟 1 处的峰值表示每个序列值与前面的值强相关，延迟 2 处的峰值表示每个值与以前两个点之间的值强相关，依此类推。

- 正相关表示较大的当前值与指定延迟处较大的值相对应；负相关表示较大的当前值与指定延迟处较小的值相对应。
- 相关的绝对值是关联强度的测量，绝对值越大表明关系越强。

序列变换

变换对在模型估计之前稳定序列常常有用。这对 ARIMA 模型尤其重要，因为估计这类模型之前需要序列保持**稳定**。如果在整个序列中，全局水平（均值）以及与该水平的平均偏差（方差）保持不变，则该序列是稳定的。

尽管多数令人感兴趣的序列都不稳定，但只要能够通过应用变换（如，自然对数、差分或季节差分）使序列保持稳定，则 ARIMA 就是有效的。

方差稳定变换。方差随时间变化的序列通常可以使用自然对数变换或平方根变换使其保持稳定。这些变换也称为函数变换。

- **自然对数。**对序列值取自然对数。
- **平方根。**对序列值应用平方根函数。

自然对数变换和平方根变换不能用于具有负值的序列。

水平稳定变换。ACF 中值的缓慢下降表示每个序列值都与上一个值具有很强的相关性。通过分析序列值的变化，您可以获得一个稳定水平。

- **简单差分。**计算序列中每个值与上一个值之间的差，序列中最旧的值除外。这意味着经过差分的序列将比原始序列少一个值。
- **季节差分。**除计算每个值与上一个季节值之间的差值外，其他均与简单差分相同。

将简单差分或季节差分同时用于对数变换或平方根变换时，总是先应用方差稳定变换。同时使用简单差分和季节差分时，无论首先应用简单差分还是季节差分，得到的序列值均相同。

预测变量序列

预测变量序列包括可能有助于解释要预测序列的行为的相关数据。例如，一个网上零售商或目录零售商可能会根据邮寄的目录数量、开通的电话数量或公司网页的点击次数来预测销售量。

任何序列都可以作为预测变量，条件是该序列须延伸到要预测的将来时间，并且具有不存在缺失值的完整数据。

向模型中添加预测变量时以慎重为宜。添加大量预测变量会增加估计模型所需的时间。虽然添加预测变量可以提高模型拟合历史数据的能力，但并不意味着该模型就一定能产生更好的预测结果，因为增加的复杂怕有可能及不上所造成的麻烦。理想的目标是，找出的模型既是最简单的，同时又能作出很好的预测。

一般而言，建议预测变量的数量应小于样本大小除以 15（即最多每 15 个观测值一个预测变量）。

有缺失数据的预测变量。具有不完整数据或缺失数据的预测变量不能在预测中使用。这适用于历史数据和将来值。在某些情况下，可通过设置模型的估计范围以便在估计模型时排除最旧数据来避免上述限制。

时间序列建模节点

时间序列节点可为时间序列估计指数平滑模型、单变量综合自回归移动平均 (ARIMA) 模型和多变量 ARIMA（或变换函数）模型并基于时间序列数据生成预测。

指数平滑是一种使用以前的序列观察的加权值来预测未来值的预测方法。因此，指数平滑不是以对数据的理论理解为基础的。指数平滑每次预测一个点，在输入新数据时可调整其预测。此技术有助于预测可展示趋势和/或季节性的序列。您可以从各种指数平滑法模型中进行选择，它们在处理趋势和季节性上有所不同。

ARIMA 模型比起指数平滑模型在对趋势和季节组件建模方面可提供更成熟的方法，特别是，增加了可在模型中包括自变量（预测变量）的优势。这包括明确指定自回归阶数和移动平均阶数以及差分次数。可以包含预测变量并为任意或所有预测变量定义变换函数以及指定对离群值的自动检测或精确设置。

注意：实际上，如果想要包括预测变量（该变量有助于解释正在预测的序列的行为，例如邮寄的目录数或某公司网页的点击数），ARIMA 模型会非常有用。而指数平滑模型在说明时间序列的行为时，并不试图去了解其行为的原因。例如，过去每隔 12 个月就会达到最大值的序列有可能继续保持该行为，即使您不了解其原因。

还可使用 **Expert Modeler**，它可自动识别和估计对一个或多个目标变量拟合得最好的 ARIMA 模型或指数平滑模型，从而不需要通过试错来识别适当的模型。在所有案例中，Expert Modeler 都可为指定的每个目标变量选择最适合的模型。如果有疑问，请使用 Expert Modeler。

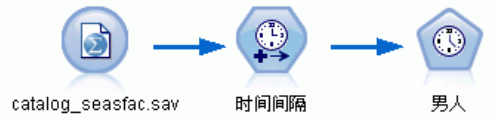
如果已指定预测变量，则 Expert Modeler 会将那些与相关序列具有统计意义下的显著关系的变量包括在 ARIMA 模型中。适当时，使用差分和/或平方根或自然对数变换对模型变量进行转换。默认情况下，Expert Modeler 会考虑所有指数平滑模型和所有 ARIMA 模型并为每个目标字段选择其中最适合的模型。不过，可以将 Expert Modeler 限制为仅选择最适合的指数平滑模型或仅选择最适合的 ARIMA 模型。还可以指定自动检测离群值。

示例。 一家全国宽带提供商要求分析师预测用户注册数量，以推测带宽普及程度。分析师需要对各地市场进行预测，才能得出全国注册用户数量。可使用时间序列建模对各地市场未来三个月注册用户数量进行预测。有关详细信息，请参阅第 15 章中的使用时间序列节点进行预测中的IBM SPSS Modeler 14.2 应用程序 指南。

要求

“时间序列”节点与其他 IBM® SPSS® Modeler 节点不同，在时间序列节点中，不能简单地将节点插入流并运行流。通常在“时间序列”节点之前，必须先插入“时间区间”节点，该节点可指定如下信息，例如所使用的时间区间（年、季度、月等）、用于估计的数据以及预测所延伸到的未来时间的范围（如果已使用）。

图片 13-7
通常在插入时间序列节点之前先插入时间区间节点



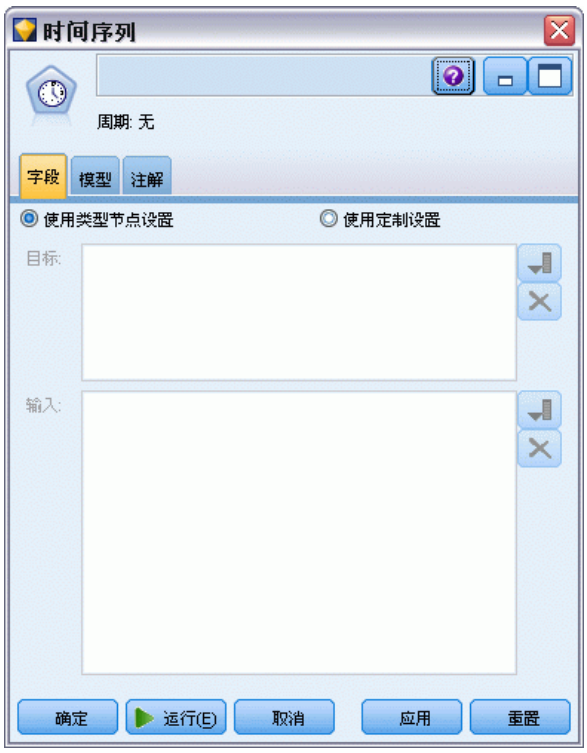
时间序列数据必须是均匀间隔的。时间序列数据建模法需要在每个测量之间有一致的区间，并由空行表示所有缺失值。如果数据尚未满足此需求，则“时间区间”节点会按需要对值进行变换。有关详细信息，请参阅第 4 章中的时间区间节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

有关时间序列节点的其他注意事项有：

- 字段必须是数字型的
- 日期字段不能作为输入使用
- 忽视分区

字段选项

图片 13-8
时间序列节点对话框，“字段”选项卡



在“字段”选项卡中可指定将用于构建模型的字段。在构建模型之前，需要指定要将哪些字段用作目标和输入。通常，时间序列节点会使用来自上游类型节点的字段信息。如果正在使用类型节点选择输入和目标字段，则不必在此选项卡上做任何更改。

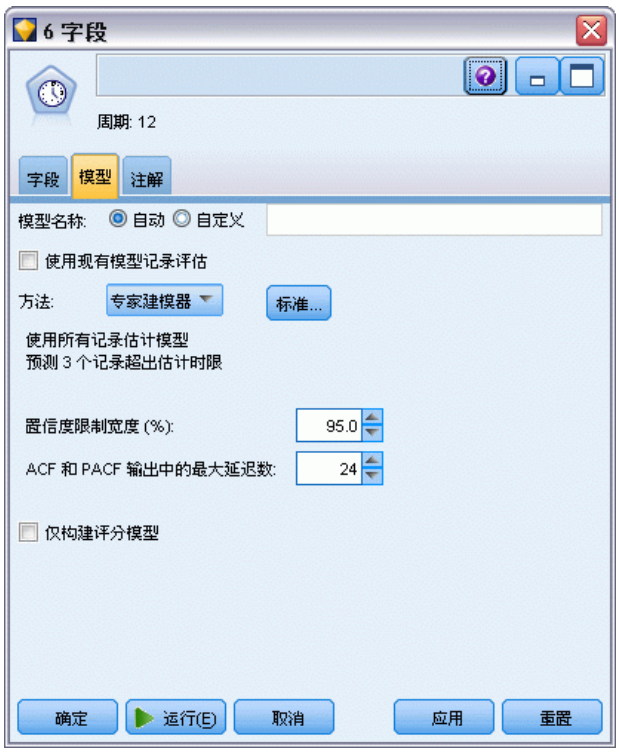
使用类型节点设置。 该选项通知节点使用来自上游类型节点的字段信息。这是默认值。

使用自定义设置。该选项通知节点使用在此处指定的字段信息，而不是在任何上游类型节点中给出的字段信息。选中此选项后，请指定下面的字段。注意，用于存储日期的字段不能用作目标字段或输入字段。

- **目标。** 选择一个或多个目标字段。此操作与在“类型”节点中将字段的角色设置为目标类似。时间序列模型的目标字段必须具有连续的测量级别。可为每个目标字段创建单独的模型。目标字段会将除自身外的所有指定的输入字段作为可能的输入。因此，同一种字段可以包含在两个列表中；除了以此字段作为目标的模型外，此字段可用作所有模型的可能的输入。
- **输入。**选择输入字段。此操作与在“类型”节点中将字段的角色设置为输入类似。时间序列模型的输入字段必须是数值型的。

时间序列模型选项

图片 13-9
“时间序列”节点对话框，“模型”选项卡



模型名称。 指定分配给执行节点时生成的模型的名称。

- **自动。** 基于目标或 ID 字段名自动生成模型名称，在未指定目标的情况下（例如聚类模型），基于模型类型名称自动生成模型名称。
- **自定义。**可为模型块指定自定义名称。

使用现有模型继续评估。如果已生成一个时间序列模型，则选择此选项可以重新使用为该模型指定的标准设置，并在模型选项板中生成一个新的模型节点，而不必从头构建一个新模型。使用此方法（即基于与以前相同的模型设置，使用最新数据重新估计并生成新的预测）可以节省时间。例如，若特定时间序列的原始模型是 Holt’s 线性趋势，则会使用相同的模型类型重新估计和预测该数据；系统不会为新数据重新尝试查找最适合的模型类型。选择此选项可禁用**方法**和**标准**控件。 有关详细信息，请参阅第 374 页码重新估计和预测。

方法。 可以选择 Expert Modeler、指数平滑或 ARIMA。有关详细信息，请参阅第 361 页码时间序列建模节点。选择**标准**可为选定的方法指定选项。

- **Expert Modeler。** 选择此选项以使用 Expert Modeler，将自动为每个相关序列查找拟合得最好的模型。
- **指数平滑法。**使用此选项可指定自定义的指数平滑法模型。
- **ARIMA。**使用此选项可指定自定义的 ARIMA 模型。

时间区间信息

对话框的此部分包含了有关在时间区间节点上进行估计和预测所使用的规范信息。注意，如果选择**使用现有模型继续评估**选项，则此部分不会显示。

该信息的第一行表示在模型中是否有任何记录被排除或保留。 有关详细信息，请参阅第 4 章中的估计时限中的IBM SPSS Modeler 14.2 源、过程和输出节点。

第二行提供了有关在时间区间节点上指定的任何预测时限的信息。 有关详细信息，请参阅第 4 章中的预测中的IBM SPSS Modeler 14.2 源、过程和输出节点。

如果第一行显示为**未定义时间区间**，则表示未连接时间区间节点。此情况在试图运行流时将引发错误；必须在时间序列节点的上游包括时间区间节点。

其他杂项信息

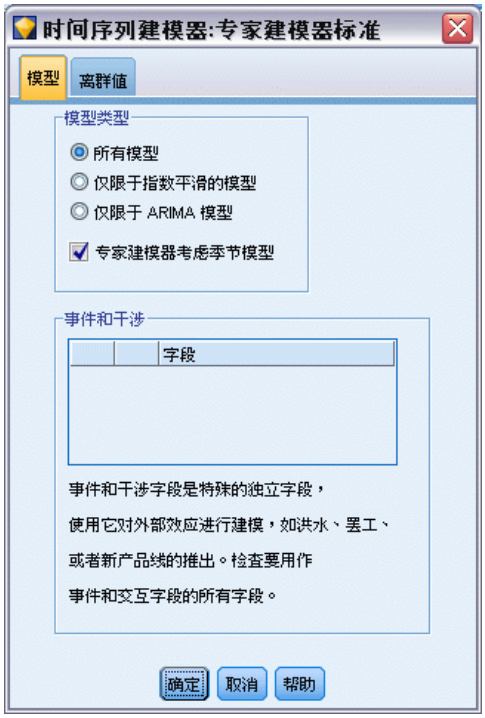
置信限宽度 (%)。将为模型预测值和残差自相关计算置信区间。可以指定小于 100 的任何正数。默认情况下使用 95% 的置信区间。

ACF 和 PACF 输出中的最大延迟数。可以设置在自相关和偏自相关表和图中显示的最大延迟数。

仅构建评分模型。选中此复选框以减少模型中存储的数据量。当构建具有大量（数万个）时间序列的模型时，这样做有助于提高性能。如果选择了此项，“模型”、“参数”和“残差”选项卡不会显示在“时间序列”模型块中，但您仍然可以按常规方式对数据进行评分。

时间序列 Expert Modeler 标准

图片 13-10
Expert Modeler 标准对话框，“模型”选项卡



模型类型。可用选项包括：

- **所有模型。**Expert Modeler 既考虑 ARIMA 模型也考虑指数平滑法模型。
- **仅指数平滑法模型。**Expert Modeler 仅考虑指数平滑法模型。
- **仅限 ARIMA 模型。**Expert Modeler 仅考虑 ARIMA 模型。

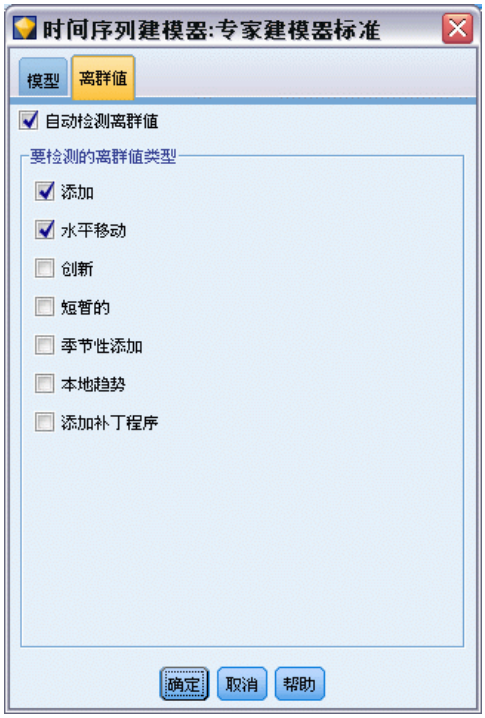
Expert Modeler 考虑季节性模型。只有在为活动数据集定义了周期性时才启用此选项。选中此选项时，Expert Modeler 将同时考虑季节模型和非季节模型。如果未选择此选项，则 Expert Modeler 仅考虑非季节性模型。

事件和干预。可将特定输入字段指定为事件字段或干预字段。此操作可将字段标识为包含受事件（可预期的循环，如促销活动）影响的时间序列数据，或是包含受干预（一次性事件，如停电或雇员罢工）影响的时间序列数据。对于标识为事件字段或干预字段的输入，Expert Modeler 将仅考虑简单回归而不是任意变换函数。

包括在此列表中的输入字段必须具有标志、名义或有序的测量级别，并且必须是数字（例如，对于标志字段，是 1/0 而不是真/假）。 有关详细信息，请参阅第 357 页码脉冲和步进。

界外值

图片 13-11
“专家建模器条件”对话框，“离群值”选项卡



自动检测离群值。 默认情况下，不自动检测离群值。选中此选项以执行离群值自动检测，然后选择所需的离群值类型。 有关详细信息，请参阅第 358 页码离群值。

时间序列指数平滑标准

图片 13-12
“指数平滑法条件”对话框

模型类型。指数平滑模型分类为季节模型或非季节模型。只有使用时间区间节点定义的周期性为季节时，才可使用季节模型。季节周期性有：循环周期、年、季度、月、一周中的天、一天中的小时、一天中的分钟和一天中的秒。 有关详细信息，请参阅第 4 章中的时间区间节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

- **简单。**此模型适合于其中没有趋势或季节性的序列。其唯一的相关平滑参数是水平。简单的指数平滑模型非常类似于自回归阶数为零、差分阶数为一、移动平均阶数为一且没有常量的 ARIMA 模型。
- **Holt’s 线性趋势。**此模型适合于其中有线性趋势但没有季节性的序列。其相关的平滑参数是水平和趋势，并且在此模型中，这些参数的值不会彼此限制。Holt’s 模型比 Brown’s 模型更加常用，但在计算大型序列的估计值时会花费更多的时间。Holt’s 指数平滑模型非常类似于自回归阶数为零、差分阶数为二且移动平均阶数为二的 ARIMA 模型。
- **Brown’s 线性趋势。**此模型适合于其中有线性趋势但没有季节性的序列。其相关的平滑参数是水平和趋势，但在此模型中，这些参数的值假设相等。因此，Brown 模型是 Holt 模型的特例。Brown’s 指数平滑模型非常类似于自回归阶数为零、差分阶数为二且移动平均阶数为二的 ARIMA 模型，其第二阶移动平均的系数等于第一阶的系数的平方的一半。
- **阻尼趋势。**此模型适合于具有逐渐消失的线性趋势但没有季节性的序列。其相关的平滑参数是水平、趋势和阻尼趋势。阻尼指数平滑模型非常类似于自回归阶数为一、差分阶数为一且移动平均阶数为二的 ARIMA 模型。
- **简单季节。**此模型适合于其中没有趋势且季节效应不随时间变化的序列。其相关的平滑参数是水平和季节。季节指数平滑模型非常类似于自回归阶数为零、差分阶数为一、季节差分阶数为一且移动平均阶数为 1、p 和 p+1 的 ARIMA 模型，其中 p 是一个季节区间中的周期数。对于以月为时间单位的数据，p = 12。

- **Winters 加法。**此模型适合于具有线性趋势且季节效应不随时间变化的序列。其相关的平滑参数是水平、趋势和季节。Winters 加法指数平滑模型非常类似于自回归阶数为零、差分阶数为一、季节差分阶数为一且移动平均阶数为 p+1 的 ARIMA 模型，其中 p 是一个季节区间中的周期数。对于以月为时间单位的数据，p = 12。
- **Winters 乘法。**此模型适合于具有线性趋势且季节效应随序列的大小变化的序列。其相关的平滑参数是水平、趋势和季节。Winters 的可乘指数平滑法与任何 ARIMA 模型都不相似。

目标变换。 可指定在对每个因变量建模前对其执行的变换。 有关详细信息，请参阅第 360 页码序列变换。

- **无。**不执行任何转换。
- **平方根。** 执行平方根变换。
- **自然对数。** 执行自然对数变换。

时间序列 ARIMA 标准

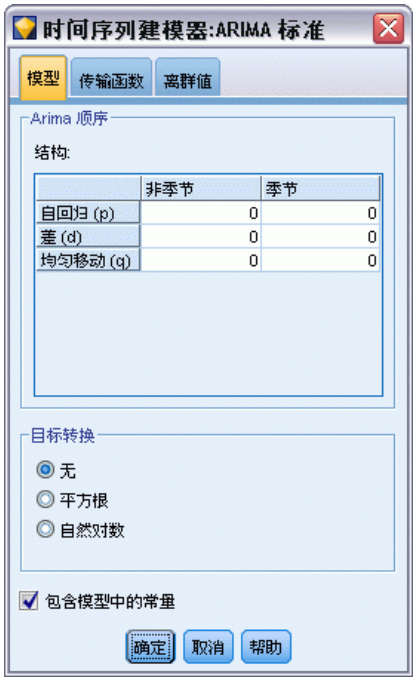
通过时间序列节点可以选择在构建自定义非季节或季节 ARIMA 模型（也称为 Box-Jenkins 模型）时，是否使用确定的输入变量（预测变量）集。可以为任何或所有输入变量定义变换函数并指定对离群值的自动检测或精确设置。

所有指定的输入变量都明确包括在模型中。这与使用 Expert Modeler 有所不同，后者只将那些与目标变量具有统计意义下的显著关系的输入变量包括在模型中。

模型

“模型”选项卡用于指定自定义的 ARIMA 模型的结构。

图片 13-13
ARIMA“条件”对话框，“模型”选项卡



ARIMA 阶数。在“结构”网格的相应单元格中，输入模型的各个 ARIMA 成分的值。所有值都必须为非负整数。对于自回归和移动平均数成分，该值表示最大阶。模型中将包含所有正的较低阶。例如，如果指定 2，则模型包括的阶数是 2 和 1。只有在已为活动数据集定义了周期性时，才会启用季节列中的单元格。

- **自回归 (p)。**模型中的自回归阶数。自回归阶指定要使用序列中以前的哪些值来预测当前值。例如，自回归阶为 2 时，指定序列中过去两个时段的值用于预测当前值。
- **差分 (d)。**指定在估计模型之前应用于序列的差分的阶。在出现趋势（具有趋势的序列通常是不稳序列，而 ARIMA 建模假定其稳定）时需要差分，并将其用于去除其影响。差分的阶与序列趋势度相对应，一阶差分导致线性趋势，二阶差分导致二次趋势，等等。
- **移动平均数 (q).**模型中的移动平均数的阶数。移动平均数的阶指定如何使用先前值的序列平均数的偏差来预测当前值。例如，如果移动平均数的阶为 1 和 2，则指定在预测序列的当前值时将考虑上两个时段的每个时段中的序列的平均值的偏差。

季节性阶。季节性自回归成分、移动平均数成分和差分成分与其非季节性对应成分起着相同的作用。但对于季节性的阶，当前序列值受以前的序列值的影响，序列值之间间隔一个或多个季节性周期。例如，对于月数据（季节性周期为 12），季节性 1 阶表示当前序列值受自当前周期起 12 个周期之前的序列值的影响。因此，对于月数据，指定季节性 1 阶等同于指定非季节性 12 阶。

目标变换。 可指定在对每个目标变量建模前对其执行的变换。 有关详细信息，请参阅第 360 页码序列变换。

- **无。**不执行任何转换。

- **平方根。** 执行平方根变换。
- **自然对数。** 执行自然对数变换。

在模型中包含常数。 除非您确定整个平均数序列值为 0，否则包含常数是标准方法。当应用差分时，建议不包含常数。

传输函数

图片 13-14
ARIMA 标准对话框，“变换函数”选项卡

使用“变换函数”选项卡可以为任何或所有的输入字段定义变换函数。通过变换函数可以指定以何种方式使用这些字段的过去值来预测目标序列的未来值。

只有在“类型”节点或”时间序列”节点的“字段”选项卡上（选择**使用自定义设置—输入**）指定输入字段（角色设置为输入）时，此选项卡才会显示。 有关详细信息，请参阅第 4 章中的设置字段角色中的IBM SPSS Modeler 14.2 源、过程和输出节点。

顶部的列表显示了所有的输入字段。此对话框中其余的信息则特定于上述列表中已选定的输入字段。

转换函数的阶数。 在“结构”网格的相应单元格中，输入转换函数的各个成分的值。所有值都必须为非负整数。对于分子和分母成分，该值代表最大阶。模型中将包含所有正的较低阶。此外，0 阶始终包括在分子成分中。例如，如果将分子阶数指定为 2，则该模型包括的阶数为 2、1 和 0。如果将分母阶数指定为 3，则该模型包括的阶数为 3、2 和 1。只有在已为活动数据集定义了周期性时，才会启用季节列中的单元格。

分子。 变换函数的分子阶数指定选定的独立（预测变量）序列中有哪些以前的值用于预测相关序列的当前值。例如，分子的阶为 1 时，指定独立序列过去一个时间段的值（以及独立序列的当前值）用于预测每个相依序列的当前值。

分母。 变换函数的分母阶数指定如何使用与选定独立（预测变量）序列以前值均值之间的偏差来预测相关序列的当前值。例如，分母的阶为 1 时，指定在预测每个相依序列的当前值时考虑独立序列过去一个时间段的平均值偏差。

差分。 指定在估计模型之前应用于所选独立（预测）序列的差分的阶数。存在趋势时必须使用差分来去除其效果。

季节性阶。 季节性分子、分母和差分成分与其非季节性对应成分起着相同的作用。但对于季节性的阶，当前序列值受以前的序列值的影响，序列值之间间隔一个或多个季节性周期。例如，对于月数据（季节性周期为 12），季节性 1 阶表示当前序列值受自当前周期起 12 个周期之前的序列值的影响。因此，对于月数据，指定季节性 1 阶等同于指定非季节性 12 阶。

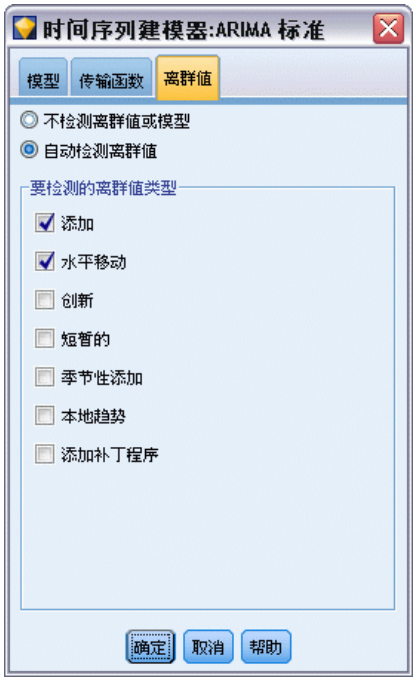
延迟。 设置延迟会使输入字段的影响在指定数目的时间间隔后才产生。例如，如果延迟设置为 5，则输入字段在时间 t 不会产生影响，直到此后五个时限后（t + 5）才会对预测产生影响。

转换。 为自变量指定的变换函数也可对这些变量上执行（可选）变换。

- **无。**不执行任何转换。
- **平方根。** 执行平方根变换。
- **自然对数。** 执行自然对数变换。

处理离群值

图片 13-15
ARIMA“条件”对话框，“离群值”选项卡



“离群值”选项卡为处理数据中的离群值提供了多种选择。

不检测离群值或为其建模。 默认情况下，不检测离群值，也不为其建模。选择此选项可禁用任何检测离群值或为其建模的功能。

自动检测离群值。 选中此选项以对离群值执行自动检测，并选择一个或多个显示的离群值类型。

要检测的离群值类型。 选择要检测的离群值类型。支持的类型有：

- 加法（默认）
- 层移动（默认）
- 创新的
- 瞬时的
- 季节性可加的
- 局部趋势
- 可加的修补

有关详细信息，请参阅第 358 页码离群值。

生成时间序列模型

本章介绍关于生成时间序列模型的某些方面的一般信息：

- 生成多个模型
- 使用时间序列模型进行预测
- 重新估计和预测

生成的模型块在单独主题中介绍。 有关详细信息，请参阅第 374 页码时间序列模型块。

生成多个模型

IBM® SPSS® Modeler 中的时间序列建模会为每个目标字段生成一个模型（ARIMA 模型或指数平滑模型）。因此，如果有多个目标字段，则 SPSS Modeler 会在一个操作中生成多个模型，这样不仅节省了时间，还可以对每个模型的设置进行比较。

如果要比较相同目标字段的 ARIMA 模型和指数平滑模型，可以分别运行时间序列节点，每次指定一个不同的模型。

使用时间序列模型进行预测

时间序列构建操作使用有序案例的特定序列（也称为估计范围），来构建可用于预测序列的未来值的模型。此模型包含有关使用的时间范围（包括区间）的信息。为了使用此模型进行预测，对于目标变量和预测变量，必须使用相同的时间范围和区间信息及相同的序列。

例如，假设在一月初时要预测产品 1 在该年前三个月中每月的销售情况。可使用产品 1 在上一年的一月到十二月（我们称为年份 1）之间的实际每月销售数据来构建模型，并将时间区间设置为“月”。然后可使用此模型预测产品 1 在年份 2 的前三个月的销售情况。

实际上可以对未来任何月数内的情况进行预测，不过可以肯定的是，试图预测的未来时间越远，模型的预测效果将会越差。但是，无法对年份 2 的前三个星期的情况进行预测，因为用于构建模型的区间为“月”。使用此模型预测产品 2 的销售同样是没有意义的，因为一个时间序列模型只与用于定义此模型的数据相关。有关详细信息，请参阅第 15 章中的使用时间序列节点进行预测中的IBM SPSS Modeler 14.2 应用程序 指南。

重新估计和预测

已将估计时限硬编码到生成的模型中。这表示如果将当前模型应用于新数据，估计时限之外的所有值都将被忽略。因此，每当有可用的新数据时，都必须对时间序列模型进行重新估计，这与其他 IBM® SPSS® Modeler 模型不同，在后者中，不用更改就可将模型重新应用于评分。

继续上述示例，假设在年份 2 的四月初，已具有从一月到三月的实际每月销售数据。但是，如果重新应用在一月初生成的模型，该模型会忽视此时限内已知的销售数据而再次预测一月到三月的销售情况。

解决方案为基于更新的实际数据生成新的模型。假设不更改预测参数，新模型用于预测下一个三个月，即四月到六月。如果仍可访问用于生成原始模型的流，则仅需使用包含已更新数据的文件引用替换该流中的源文件引用，并重新运行流以生成新模型。但是，如果可以使用的只有保存在文件中的原始模型，则仍可使用该模型生成“时间序列”节点，然后将此节点添加到包含已更新源文件引用的新流中。假设在此新流中“时间区间”节点（其中已将区间设置为“月”）在“时间序列”节点之前插入，则运行此新流会生成需要的新模型。有关详细信息，请参阅第 15 章中的重新应用时间序列模型中的IBM SPSS Modeler 14.2 应用程序 指南。

时间序列模型块

时间序列建模操作使用如下所示的前缀 $TS- 创建多个新字段：

$TS-colname	每个目标序列模型预测的值。
$TSLCI-colname	每个已预测序列的置信区间下限值。*
$TSUCI-colname	每个已预测序列的置信区间上限值。*
$TSNR-colname	每列生成的模型数据中的噪声残差值。*
$TS-Total	此行中 $TS-colname 的总值。
$TSLCI-Total	此行中 $TSLCI-colname 的总值。*
$TSUCI-Total	此行中 $TSUCI-colname 的总值。*
$TSNR-Total	此行中 $TSNR-colname 的总值。*

* 这些字段的可见性（例如，在添加表格节点的输出中）取决于时间序列模型块的“设置”选项卡上的选项。 有关详细信息，请参阅第 380 页码时间序列模型设置。

图片 13-16
时间序列模型块，“模型”选项卡

对于输入到时间序列构建节点中的每个序列，时间序列模型块将显示为这些序列选择的不同模型的详细信息。可以输入多个序列（例如有关产品线、地区或仓库的数据），且对于每个目标序列可生成一个独立的模型。例如，如果认为东部地区的收入可拟合 ARIMA 模型，但西部地区的收入仅能拟合简单移动平均模型，则会使用适合的模型对每个地区进行评分。

对于每个构建的模型，默认的输出有：模型类型、指定的预测变量数及拟合优度测量（默认为平稳的 R 平方）。如果已指定离群值方法，则会有一列用于显示检测到的离群值数。默认输出还包括 Ljung-Box Q 列、自由度列和显著相关值列。

还可以选择高级输出，其中将显示以下附加列：

- R 方
- RMSE（均方根误差）
- MAPE（平均绝对百分误差）
- MAE（平均绝对误差）
- MaxAPE（最大绝对百分比误差）

- MaxAE（最大绝对误差）
- 标准化BIC（标准化贝叶斯信息准则）

生成。可以回到流或选项板的模型块生成“时间序列”建模节点。

- **生成建模节点。**使用用于创建此模型集的设置将“时间序列”建模节点放置到流中。举例而言，此操作的意义在于，如果要使用这些模型设置，却找不到用于生成这些设置的建模节点时，可以在流中获得这些设置。
- **模型到调色板。**将包含所有目标的模型块放置到模型管理器中。

模型

图片 13-17
“选中所有”按钮和“取消选中所有”按钮

复选框。 选择要用于评分的模型。默认情况下会选中所有复选框。**选中所有**和**取消选中所有**按钮会以一个简单的操作作用于所有的复选框。

排序依据。可以按升序或降序顺序对指定显示列的输出行进行排序。“选定”选项可根据由复选框选定的一行或多行对输出进行排序。举例而言，此操作的意义在于，使名称为从“Market_1”到“Market_9”之间的目标字段显示在“Market_10”之前，因为在默认排序顺序下，“Market_10”将直接显示在“Market_1”之后。

视图。 默认视图（简单）将显示基本的输出列的集合。高级选项将显示拟合优度测量的附加列。

估计中使用的记录数。 原始源数据文件中的行数。

目标。“类型”节点中标识为目标字段（角色为目标的字段）的字段。

模型。 用于此目标字段的模型类型。

预测变量。用于此目标字段的预测变量（角色为输入的字段）数。

离群值。 只有在已请求（在 Expert Modeler 或 ARIMA 标准中）自动检测离群值时，才会显示此列。显示的值为检测到的离群值数。

固定的 R 方. 将模型的平稳部分与简单均值模型相比较的测量。当具有趋势或季节性模式时，该度量适用于普通 R 方。固定的 R 方可以是负无穷大到 1 范围中的负值。负值表示考虑中的模型比基线模型差。正值表示考虑中的模型比基线模型好。

R 方. 线性模型的拟合优度，有时称为判定系数。它是因变量的变动中，由回归模型解释的比例。其值的范围为 0 到 1。如果值较小，则表示该模型与数据的拟合度不好。

RMSE. 均方根误差。均方误差的平方根。度量因变量序列与其模型预测水平的相差程度，用和因变量序列相同的单位表示。

MAPE. 平均绝对误差百分比。度量因变量序列与其模型预测水平的相差程度。它与使用的单位无关，因此可用于比较具有不同单位的序列。

MAE. 平均绝对误差。度量序列与其模型预测水平的差别程度。MAE 以原始序列单位报告。

MaxAPE. 最大绝对误差百分比。最大的预测误差，以百分比表示。该度量对于想象预测的最坏情况方案很有用。

MaxAE. 最大绝对误差。最大的预测误差，以和因变量序列相同的单位表示。与 MaxAPE 相同，它对于想象预测的最坏情况方案很有用。最大绝对误差和最大绝对误差百分比可能发生在不同的序列点上，例如，当较大序列的绝对误差比较小值的绝对误差稍微大一些时。在此情况下，最大绝对误差将发生在较大序列值处，而最大绝对误差百分比将发生在较小序列值处。

标准化的 BIC. 标准化的 BIC（BIC 准则）。尝试代表模型复杂性的模型整体拟合的一般度量。它是基于均方误差的分数，包括模型中参数数量的罚分和序列长度。罚分去除了具有更多参数的模型优势，从而可以容易地比较相同序列的不同模型的统计量。

Q。Ljung-Box Q 统计量。该模型中残差错误的随机测试。

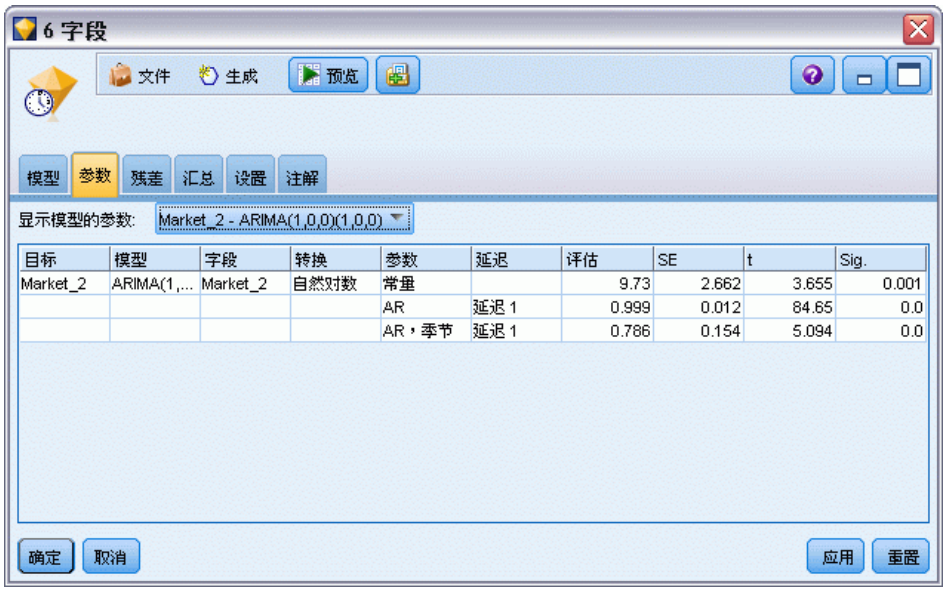
df。自由度。估计特定目标时自由变化的模型参数量。

显著水平Ljung-Box 统计量的显著性值。显著性值小于 0.05 表示残差错误不是随机的。

摘要统计量。 此部分包含了不同列的各种汇总统计量，包括均值、最小值、最大值和百分位数值。

时间序列模型参数

图片 13-18
时间序列模型，“参数”选项卡



“参数”选项卡列出用于构建选定模型的各种参数的详细信息。

显示模型的参数。选择为其显示参数详细信息的模型。

目标。该模型预测的目标字段（角色为目标）的名称。

模型。 用于此目标字段的模型类型。

字段（仅 ARIMA 模型）。包含模型中使用的每个变量的一个条目，第一个是目标，接着是预测变量（如果有）。

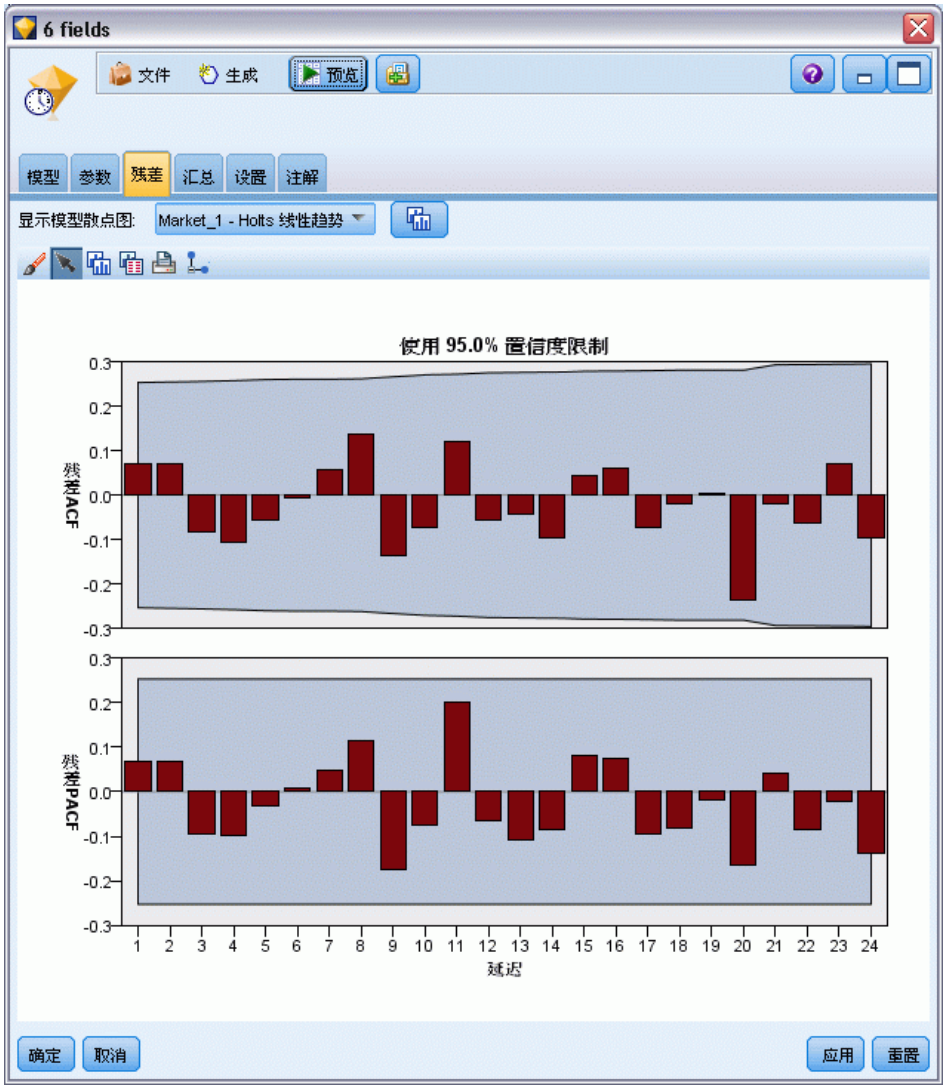
转换。指明在构建模型前为这个字段指定哪一种变换类型（如果有）。

参数。为其显示下列详细信息的模型参数：

- **延迟(仅 ARIMA 模型)。**指示认为是模型中该参数的延迟（如果有）。
- **估计值。**参数估计值。该值用于计算预测值和目标字段的置信区间。
- **SE。**参数估计值的标准误。
- **t。**参数估计值除以标准误后的值。
- **显著水平**参数估计值的显著性级别。0.05 以上的值视为没有显著的统计意义。

时间序列模型残差

图片 13-19
时间序列模型，“残差”选项卡（显示 ACF 和 PACF）

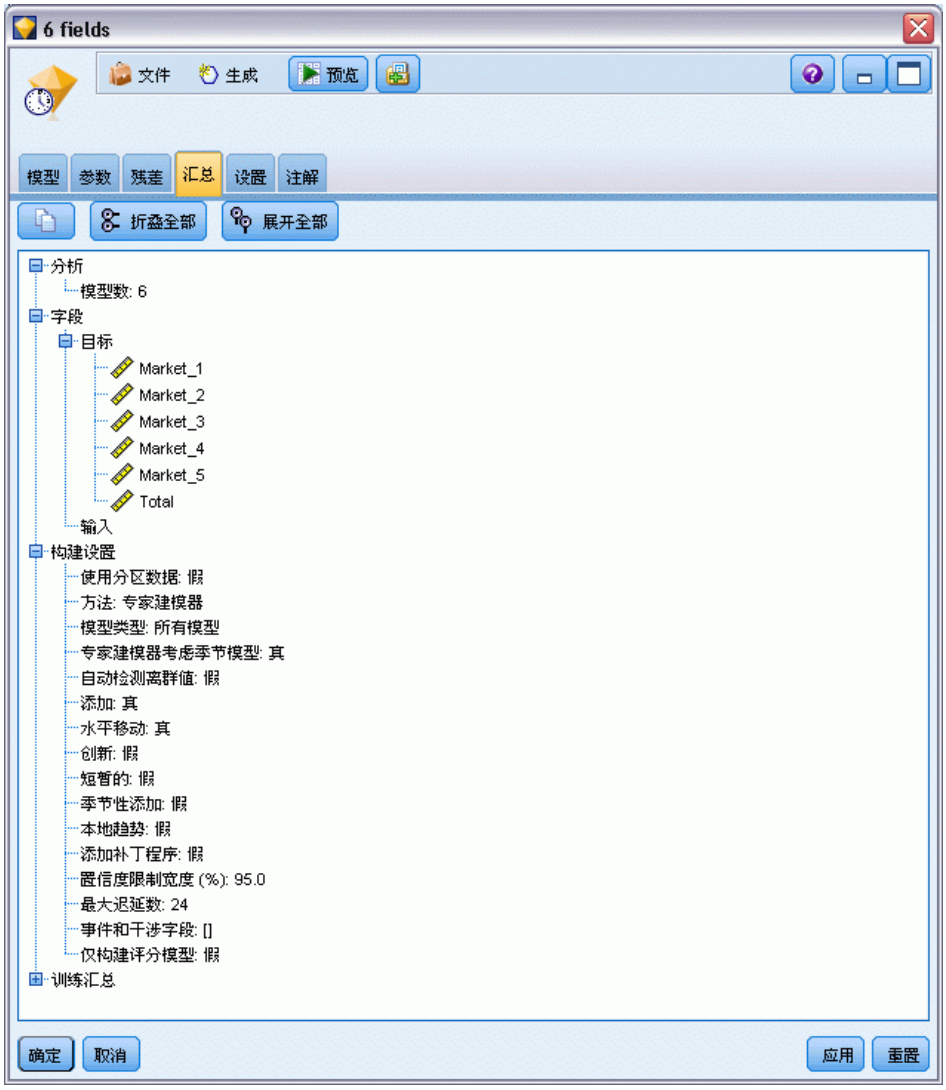


“残差”选项卡为每个构建的模型显示了残差（指期望值和实际值之间的差值）的自相关函数（ACF）和偏自相关函数（PACF）。 有关详细信息，请参阅第 360 页码自相关函数和部分自相关函数。

显示模型散点图。选择要为其显示残差 ACF 和残差 PACF 的模型。

时间序列模型汇总

图片 13-20
时间序列模型，“汇总”选项卡



模型块的“汇总”选项卡显示了有关模型的下列信息：模型本身（分析）、模型中使用的字段（字段）、构建模型时使用的设置（构建设置）和模型训练（训练概要）。

当第一次浏览此节点时，“汇总”选项卡的结果是折叠起来的。要查看感兴趣的结果，可使用项目左侧的展开控件展开项目，或单击**全部展开**按钮显示所有结果。当结束对项目的查看时，为了隐藏结果，可使用展开控件折叠要隐藏的特定结果，或单击**全部折叠**按钮折叠所有结果。

分析。 显示指定模型的相关信息。

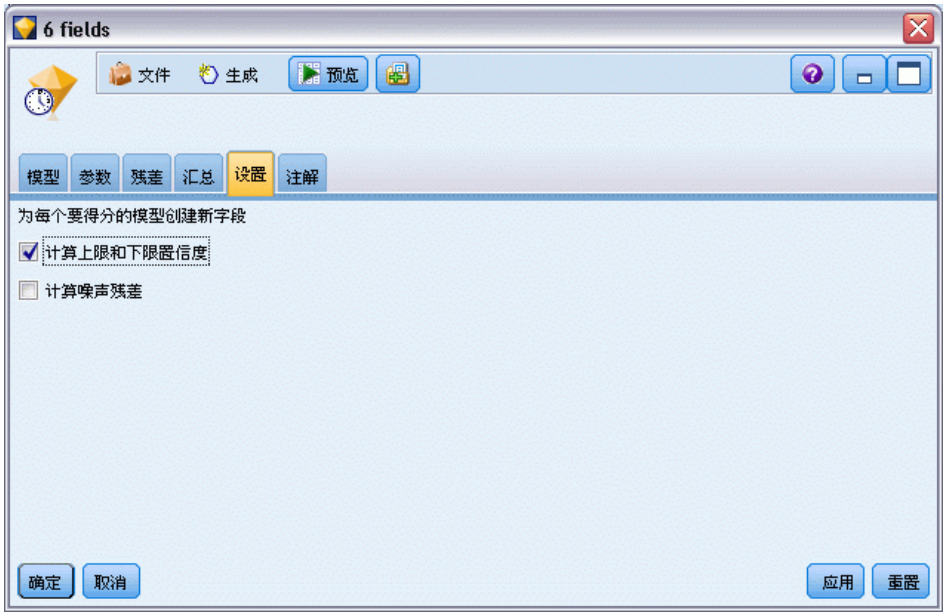
字段。 列出构建模型时用作目标和输入的字段。

构建设置。 包含有关在构建模型中使用的设置的信息。

训练概要。 显示模型类型、用于创建模型的流、模型创建者、模型构建完成时间和模型构建所用时间。

时间序列模型设置

图片 13-21
时间序列模型，“设置”选项卡



使用“设置”选项卡，可以指定通过建模操作创建哪些附加字段。

为每个待评分的模型创建新字段。 可以指定为每个待评分的模型创建的新字段。

- **计算置信上限和下限。** 如果选中，则对于每个目标字段，将分别为置信上限和下限创建新字段（默认前缀为 $TSLCI- 和 $TSUCI-），并同时创建这些值的合计字段。
- **计算噪声残差。**如果选中，则对于每个目标字段，将为模型残差创建新字段（默认前缀为 $TSNR-），并同时创建这些值的合计字段。

章
14

自学响应节点模型

SLRM 节点

使用**自学响应模型**（SLRM）节点，可以构建这样的模型：随着数据集的增长，可以不断对其进行更新或重新估计，而不必每次使用整个数据集重新构建该模型。例如，如果有若干产品，而您希望确定某位客户获得报价后最有可能购买的产品，那么这种模型将十分有用。此模型可用于预测最适合客户的报价，以及该报价被接受的概率。

最初构建模型时，可以使用较小的数据集，其中的报价和对这些报价的响应可以随机选择。随着数据集的增长，模型可得到更新，从而越发能够根据其他输入字段（如年龄、性别、职业和收入）预测最适合客户的报价以及这些客户接受报价的概率。可以通过在节点对话框中添加或删除这些可用报价对其进行更改，而不必更改数据集的目标字段。

如果与 IBM® SPSS® Collaboration and Deployment Services 一起使用，则可以为模型设立自动定期更新。该过程不需要人工监督或操作就可以为不可能或没必要由数据挖掘者自定义干预的组织和应用程序提供灵活且成本低的解决方案。

示例。某金融机构希望为每位客户提供最有可能接受的报价，以获取更大收益。可以使用自学模型基于以前的促销活动确定最有可能对活动作出积极响应的客户的特征，并根据最近的客户响应实时更新该模型。有关详细信息，请参阅第 17 章中的向客户报价（自学）中的IBM SPSS Modeler 14.2 应用程序 指南。

SLRM 节点字段选项

图片 14-1
SLRM 节点对话框，“字段”选项卡

执行 SLRM 节点之前，必须在节点的“字段”选项卡上同时指定目标字段和目标响应字段。

目标字段。从列表中选定目标字段；例如，包含要为客户提供的不同产品的名义（集合）字段。

注意：目标字段的存储方式必须采用字符串存储而不是数字型存储。

目标响应字段。从列表中选定目标响应字段。例如，接受或拒绝。

注意：此字段必须是标志字段。标志的真值表示报价接受，假值表示报价拒绝。

此对话框中的剩余字段是整个 IBM® SPSS® Modeler 中通用的标准字段。 有关详细信息，请参阅第 30 页码第 3 章中的建模节点字段选项。

注意：如果源数据包括要用作连续（数值范围）输入字段的范围，则必须确保元数据包括每个范围的最小值和最大值。

SLRM 节点模型选项

图片 14-2
SLRM 节点对话框，“模型”选项卡

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

继续训练现有模型。默认情况下，每当执行一个建模节点时，就会创建一个完整的新模型。如果选中该选项，则会继续训练该节点成功生成的最后一个模型。这样就可以在无需访问原始数据的情况下更新或刷新现有的模型，并可能会显著提升性能，这是因为只有 新的或更新后的记录被反馈到流中。上一个模型的详细信息与建模节点存储在一起，这样即使先前的模型块在流或模型选项板中不再可用的情况下，也可以使用该选项。

目标字段值 默认情况下会将此选项设置为**使用全部**，表示将构建包含所选目标字段值的每个关联报价的模型。如果希望生成仅包含目标字段的某些报价的模型，请单击**指定**，并使用**添加**、**编辑**和**删除**按钮输入或修改要为其构建模型的报价的名称。例如，如果选择的目标是列出提供的所有产品，则可以使用此字段将提供报价的产品限制为在此输入的产品。

模型评估。此面板中的字段与模型不相关，这些字段不会影响所在模型的评分。不过，这些字段有助于形成一个直观表示，显示模型预测结果的准确程度。

注意：要在模型块中显示模型评估结果，还必须选中**显示模型评估**复选框。

- **包括模型评估。**选中此框可创建针对每项所选报价显示模型预测准确性的图形。
- **设置随机数种子。**根据随机百分比估计模型准确性时，此选项可用于在另一会话中复制相同结果。通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。输入所需的种子值。如果未选中该选项，则每次执行节点时会生成不同的抽样。
- **模拟样本大小。**指定评估模型时样本中要使用的记录数。默认值为 100。
- **迭代数。**使用此选项可在经过指定的迭代数后停止构建模型评估。指定最大迭代次数；默认值为 20。

  注意：记住，样本大小较大及迭代次数较多会增加构建模型所用的时间。

显示模型评估。选中此选项可使用图形显示模型块中的结果。

SLRM 节点设置选项

图片 14-3
SLRM 节点对话框，“设置”选项卡

使用节点设置选项可微调模型构建过程。

每条记录的最大预测数。使用此选项，可以限制对数据集中每条记录进行的预测的次数。默认值为 3。

例如，您有六项报价（如储蓄、抵押、汽车贷款、退休金、信用卡和保险），但只想了解最适于推荐的两项；这时应将此字段设为 2。当您构建模型并将其附加到表中时，会看到每条记录有两个预测列（以及接受的报价的相关置信概率）。预测可以由六种可能报价中的任意报价组成。

随机化等级。 为避免任何偏差（例如，在小型或不完整的数据集中）并平等对待所有可能的报价，可以为选择的报价及其作为推荐报价被纳入的概率添加随机化等级。随机化表示为百分比，以 0.0（无随机化）与 1.0（完全随机化）之间小数值的形式显示。默认值为 0.0。

设置随机数种子。为选择的报价添加随机化等级时，可使用此选项在另一个会话中复制相同结果。通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。输入所需的种子值。如果未选中该选项，则每次执行节点时会生成不同的抽样。

注意：为从数据库中读取的记录选择**设置随机数种子**选项时，可能在抽样前需要使用排序节点以确保每次执行节点时能得到相同的结果。这是因为随机种子依赖于记录的顺序，而在关系数据库中不能保证记录具有这种顺序。有关详细信息，请参阅第 3 章中的排序节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

排列次序。选择报价在所构建模型中的显示顺序：

- **降序。**模型首先显示得分最高的报价。这些报价被接受的概率最高。
- **升序。**模型首先显示得分最低的报价。这些报价被拒绝的概率最高。例如，在决定要从某种特定报价的营销活动中删除哪些客户时，这种顺序相当实用。

目标字段的首选项。构建模型时，您可能希望对数据的某些方面进行主动推广或删除。例如，如果构建用于选择为某个客户推荐的最佳财务报价的模型，您可能需要确保始终包含一种特定报价（无论其对于每个客户的得分如何）。

要在此面板中包含某项报价并编辑其首选项，请单击**添加**，键入报价的名称（例如，“储蓄”或“抵押”），然后单击**确定**。

- **值。**此选项将显示您添加的报价的名称。
- **首选度。**指定要应用于报价的首选度等级。首选度表示为百分比，以 0.0（非首选）与 1.0（最首选）之间小数值的形式显示。默认值为 0.0。
- **始终包括。**要确保某项特定报价始终包含于预测中，请选中此框。

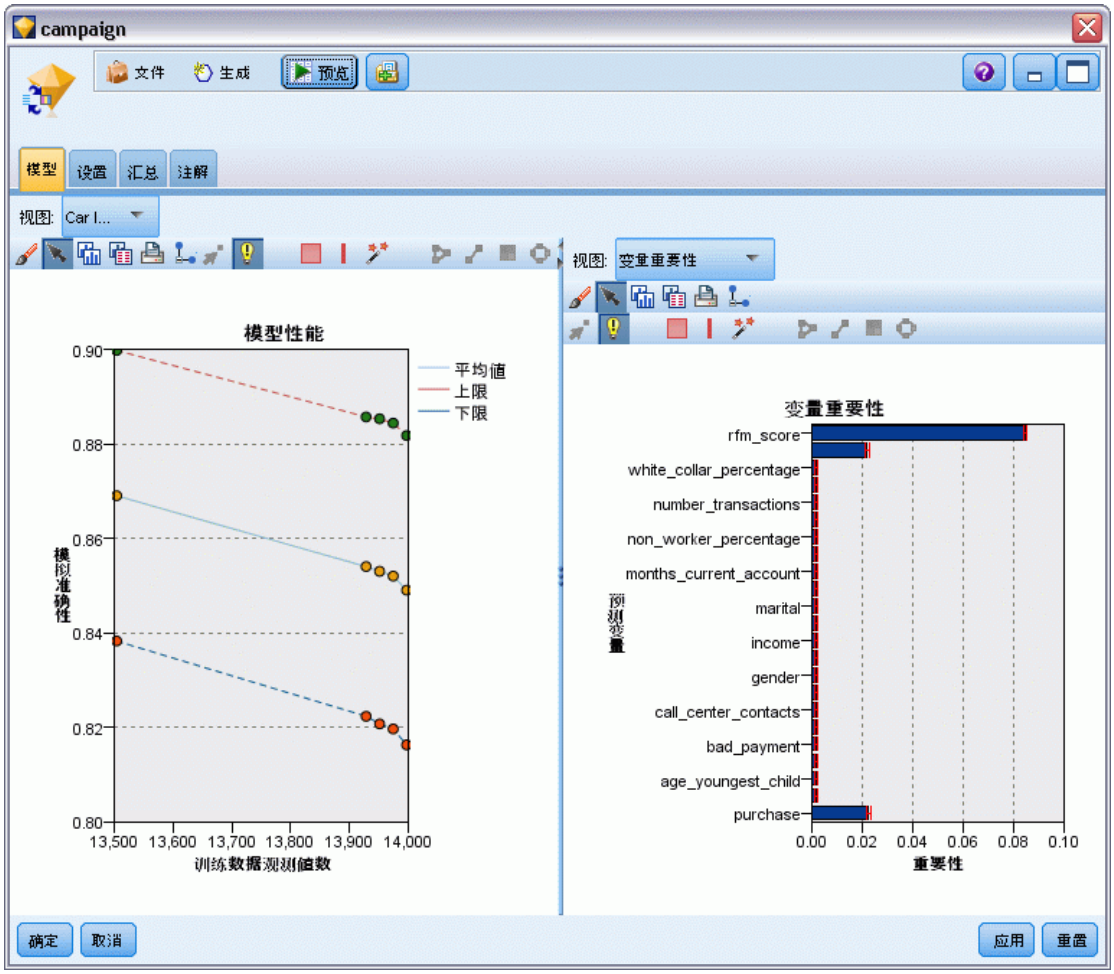
  注意：如果将**首选度**设为 0.0，将忽略**始终包括**设置。

考虑模型可靠性。与数据极少的全新模型相比，已通过多次重新生成进行微调的结构良好、数据丰富的模型应当始终产生更精确的结果。要利用较成熟模型具有的较高可靠性，请选中此框。

SLRM 模型块

注意：如果在“模型选项”选项卡上同时选中**包括模型评估**和**显示模型评估**，则仅在此选项卡上显示结果。

图片 14-4
SLRM 模型块显示



在运行包含 SLRM 模型的流时，该节点会估计每个目标字段值（报价）的预测准确性，以及所用的每个预测变量的重要性。

注意：如果您选中了建模节点“模型”选项卡中的**继续训练现有模型**，则每次生成模型时都会更新模型块上显示的信息。

对于使用 IBM® SPSS® Modeler 12.0 或更高版本构建的模型，模型块的“模型”选项卡分为两列：

左列。

- **视图。**有多项报价时，选择要显示其结果的一项报价。
- **模型性能。**显示针对每个报价所估计的模型的准确性。测试集合通过模拟生成。

右列。

- **视图。**选择是否要显示**与响应之间的关联**或**变量重要性**的详细信息。

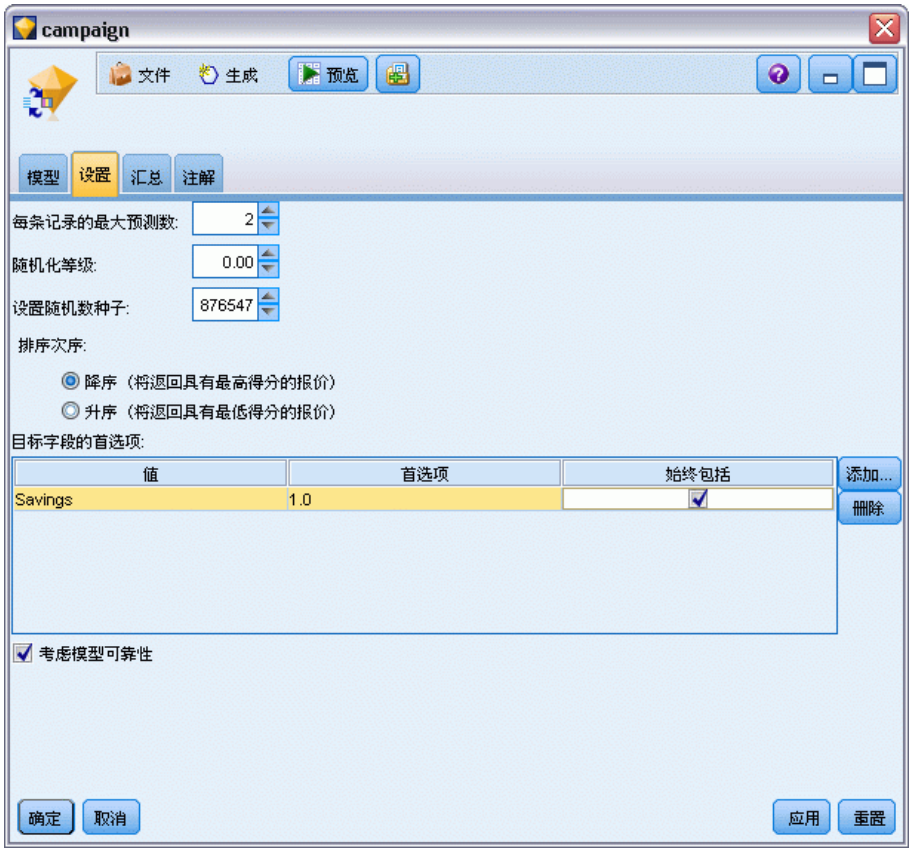
- **与响应之间的关联。**显示每个预测变量与目标变量之间的关联（相关）。
- **预测变量重要性。**表示在估计模型过程中每个预测变量的相对重要性。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。虽然在使用 SLRM 的情况下，图形是由 SLRM 算法模拟生成的，但该图表可用解释其他显示预测变量重要性的模型的方式进行解释。方法是：依次从模型中删除每个预测变量，然后查看此操作对模型准确性的影响如何。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

SLRM 模型设置

在 SLRM 模型块的“设置”选项卡中可指定选项以修改已构建的模型。例如，可以通过 SLRM 节点使用相同的数据和设置构建几个不同的模型，然后使用每个模型中的此选项卡对设置稍做修改以查看其对结果的影响。

注意：只有将模型块添加到流中之后，此选项卡才可用。

图片 14-5
SLRM 模型块对话框，“设置”选项卡



每条记录的最大预测数。使用此选项，可以限制对数据集中每条记录进行的预测的次数。默认值为 3。

例如，您有六项报价（如储蓄、抵押、汽车贷款、退休金、信用卡和保险），但只想了解最适于推荐的两项；这时应将此字段设为 2。当您构建模型并将其附加到表中时，会看到每条记录有两个预测列（以及接受的报价的相关置信概率）。预测可以由六种可能报价中的任意报价组成。

随机化等级。 为避免任何偏差（例如，在小型或不完整的数据集中）并平等对待所有可能的报价，可以为选择的报价及其作为推荐报价被纳入的概率添加随机化等级。随机化表示为百分比，以 0.0（无随机化）与 1.0（完全随机化）之间小数值的形式显示。默认值为 0.0。

设置随机数种子。为选择的报价添加随机化等级时，可使用此选项在另一个会话中复制相同结果。通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。输入所需的种子值。如果未选中该选项，则每次执行节点时会生成不同的抽样。

注意：为从数据库中读取的记录选择**设置随机数种子**选项时，可能在抽样前需要使用排序节点以确保每次执行节点时能得到相同的结果。这是因为随机种子依赖于记录的顺序，而在关系数据库中不能保证记录具有这种顺序。有关详细信息，请参阅第 3 章中的排序节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

排列次序。选择报价在所构建模型中的显示顺序：

- **降序。**模型首先显示得分最高的报价。这些报价被接受的概率最高。
- **升序。**模型首先显示得分最低的报价。这些报价被拒绝的概率最高。例如，在决定要从某种特定报价的营销活动中删除哪些客户时，这种顺序相当实用。

目标字段的首选项。构建模型时，您可能希望对数据的某些方面进行主动推广或删除。例如，如果构建用于选择为某个客户推荐的最佳财务报价的模型，您可能需要确保始终包含一种特定报价（无论其对于每个客户的得分如何）。

要在此面板中包含某项报价并编辑其首选项，请单击**添加**，键入报价的名称（例如，“储蓄”或“抵押”），然后单击**确定**。

- **值。**此选项将显示您添加的报价的名称。
- **首选度。**指定要应用于报价的首选度等级。首选度表示为百分比，以 0.0（非首选）与 1.0（最首选）之间小数值的形式显示。默认值为 0.0。
- **始终包括。**要确保某项特定报价始终包含于预测中，请选中此框。

  注意：如果将**首选度**设为 0.0，将忽略**始终包括**设置。

考虑模型可靠性。与数据极少的全新模型相比，已通过多次重新生成进行微调的结构良好、数据丰富的模型应当始终产生更精确的结果。要利用较成熟模型具有的较高可靠性，请选中此框。

章
15

Support Vector Machine 模型

关于 SVM

Support Vector Machine (SVM) 是一项功能强大的分类和回归技术，可最大化模型的预测准确度，而不会过度拟合训练数据。SVM 特别适用于分析预测变量字段非常多（如数千个）的数据。

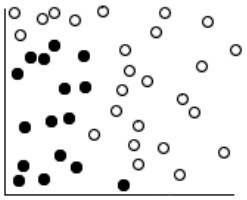
SVM 适用于多个学科，例如客户关系管理 (CRM)、面部图像和其他图像识别、生物信息学、文本挖掘概念提取、入侵检测、蛋白质结构预测以及语音识别。

SVM 如何运行

SVM 的工作原理是将数据映射到高维特征空间，这样即使数据不是线性可分，也可以对该数据点进行分类。找到类别之间的分隔符，然后以将分隔符绘制成超平面的方式变换数据。之后，可用新数据的特征预测新记录所属的组。

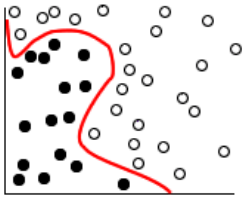
例如，请看下图，图中的数据点落到了两个不同的类别中：

图片 15-1
原始数据集



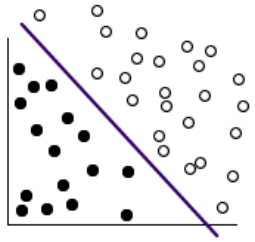
可以用一条曲线分隔这两个类别：

图片 15-2
添加了分隔符后的数据



变换后，可以用超平面定义这两个类别之间的边界：

图片 15-3
变换后的数据



用于变换的数学函数称为**核**函数。IBM® SPSS® Modeler 中的 SVM 支持下列核类型：

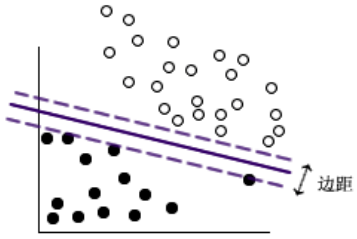
- 线性
- 多项式
- 径向基函数 (RBF)
- Sigmoid

如果数据的线性分隔比较简单，则建议使用线性核函数。在其他情况下，应当使用其他函数。在所有情况下，您都需要尝试使用不同的函数才能获得最佳模型，因为每一个函数均使用不同的算法和参数。

调整 SVM 模型

除了类别之间的分隔线，分类 SVM 模型还会查找定义两个类别之间的空间的边际线。

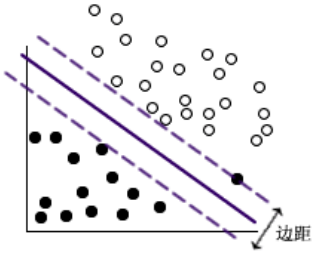
图片 15-4
初具模型的数据



位于边距上的数据点称为**支持向量**。

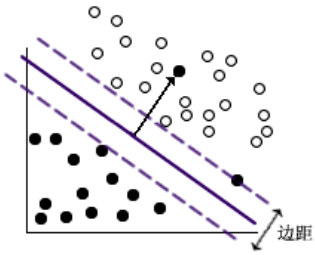
两个类别之间的边距越宽，模型在预测新记录所属的类别方面性能越佳。在上一个示例中，边距不是很宽，因此称该模型**过度拟合**。为了使边界变宽，可以接受少量的误分类，例如：

图片 15-5
模型改进后的数据



在某些情况下，线性分隔难度较大，例如：

图片 15-6
线性分隔存在的问题



在类似这种情况中，目标是找到宽边距和少量误分类数据点之间的最佳平衡。核函数有一个**规则化参数**（称为 C），该参数控制这两个值之间的平衡。若要获得最佳模型，您可能需要对该参数和其他核参数尝试使用不同的值。

SVM 节点

通过 SVM 节点，可以使用 Support Vector Machine 对数据进行分类。SVM 特别适合于大型数据集，即具有大量预测变量字段的数据集。可以对节点使用默认设置以相对较快地生成基本模型，也可以使用“专家”设置以尝试使用不同类型的 SVM 模型。

生成模型后，您可以：

- 浏览模型块，以显示生成模型过程中相对比较重要的输入字段。
- 将表节点附加到模型块中，以查看模型输出。

示例。一位医学研究人员获得了一个包含大量人体细胞样本的特征的数据集，这些样本是从极有可能患上癌症的患者身上提取的。通过对原始数据进行分析，发现良性样本与恶性样本之间的许多特征显著不同。该研究人员希望开发一个 SVM 模型，使该模型可以使用其他患者样本中的相似细胞特征的值，以尽早发现他们的样本是良性的还是恶性。有关详细信息，请参阅第 26 章中的细胞样本分类（SVM）中的IBM SPSS Modeler 14.2 应用程序 指南。

SVM 节点模型选项

图片 15-7
SVM 节点模型选项

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

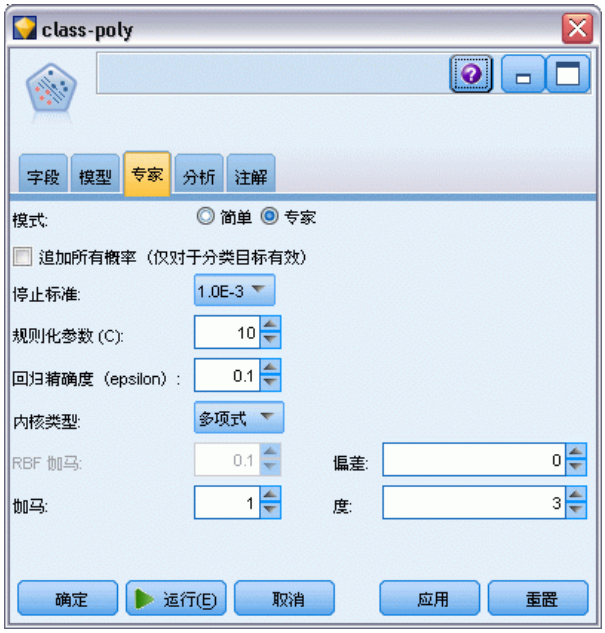
使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

SVM 节点专家选项

如果您对 Support Vector Machine 具有深入了解，则可以使用专家选项对训练过程进行调整。要访问专家选项，请在“专家”选项卡上将“模式”设置为**专家**。

图片 15-8
SVM 节点专家选项



追加所有概率（仅对分类目标有效）。如果选中该选项，则指定为由节点处理的每个记录显示名义或标志目标字段的各个可能值的概率。如果未选中该选项，则仅为名义或标志目标字段显示预测值的概率。该复选框的设置将决定模型块上的相应复选框的默认状态。

停止标准。确定何时停止优化算法。值的范围从 1.0E－1 到 1.0E－6；默认值为 1.0E－3。减小该值会生成更准确的模型，但模型的训练时间也要相应增加。

规则化参数(C)。控制最大化边距和最小化训练错误项之间的平衡。通常情况下，值应当介于 1 和 10（含）之间；默认值为 10。增加该值会提高训练数据的分类准确度（或减少回归错误），但这也可以导致过度拟合。

回归精确度（正数）。仅当目标字段的测量级别为连续时才使用。如果错误数小于此处指定的值，则可以接受错误数。增加该值可能会加快建模速度，但要以准确度为代价。

核类型。确定用于变换的核函数的类型。核类型不同，计算分隔符的方法也将不同，因此建议尝试使用不同的选项。默认值为 RBF（径向基函数）。

RBF 伽马。仅在核类型设置为 RBF 时才启用。通常情况下，值应当介于 3/k 和 6/k 之间，其中 k 为输入字段的数量。例如，如果有 12 个输入字段，则应当尝试使用介于 0.25 和 0.5 之间的值。增加该值会提高训练数据的分类准确度（或减少回归错误），但这也可以导致过度拟合。

伽马。仅在核类型设置为 多项式或 Sigmoid 时才启用。增加该值会提高训练数据的分类准确度（或减少回归错误），但这也可以导致过度拟合。

偏差。仅在核类型设置为 多项式或 Sigmoid 时才启用。在核函数中设置 coef0 值。大多数情况下可以使用默认值 0。

度。仅在核类型设置为多项式时才启用。控制映射空间的复杂性（维度）。通常情况下，不使用大于 10 的值。

SVM 模型块

SVM 模型会创建许多新字段。其中最重要的是 $S-fieldname 字段，该字段显示由模型预测的目标字段值。

模型创建的新字段的数量和名称取决于目标字段的测量级别（此字段在下表中由字段名指示）。

要查看这些字段及其值，请将“表”节点添加到 SVM 模型块中，然后执行“表”节点。

表 15-1
目标字段测量级别为“名义”或“标记”

新字段名	描述
$S-fieldname	目标字段预测值。
$SP-fieldname	预测值概率。
$SP-value	名义或标志的各个可能值的概率（仅在选中模型块中“设置”选项卡上的**追加所有概率**时才显示）。
$SRP-value $SAP-value	（仅适用于标志目标）原始 (SRP) 和调整后的 (SAP) 倾向得分，表示目标字段结果为“真”的可能性。仅当在生成模型之前选中 SVM 建模节点的“分析”选项卡上的相应复选框之后，才显示这些得分。 有关详细信息，请参阅第 34 页码第 3 章中的建模节点分析选项。

表 15-2
目标字段测量级别为“连续”

新字段名	描述
$S-fieldname	目标字段预测值。

预测变量重要性

另外，“模型”选项卡上还可能显示表示评估模型时每个预测变量相对重要性的图表。通常您要将建模的主要精力放在最重要的预测变量上，并考虑丢弃和删除那些最不重要的预测变量。注意，只有在生成模型之前选中“分析”选项卡上的**计算预测变量重要性**，才可以使用此图表。 有关详细信息，请参阅第 45 页码第 3 章中的预测变量重要性。

注意：与其他类型的模型相比，SVM 模型在计算预测变量重要性时可能会花费更长的时间，默认情况下不在“分析”选项卡中选中该选项。选中该选项可能会降低性能，对大数据集尤为明显。

SVM 模型设置

图片 15-9
SVM 模型，“设置”选项卡

通过“设置”选项卡可以指定在查看结果时显示的附加字段（例如，通过执行表节点附加到块）。通过选择这些选项可以查看每个选项的效果，并且单击“预览”按钮（滚动至“预览”输出右侧）可以查看附加字段。

追加所有概率（仅对分类目标有效）。 如果选中该选项，则为由节点处理的各个记录显示名义或标志目标字段的各个可能值的概率。如果未选中该选项，则仅为名义或标志目标字段显示预测值及其概率。

此复选框的默认设置由建模节点的相应复选框确定。

计算原始的倾向得分。对于含标志目标（返回?是?或?否?预测）的模型，您可以请求倾向得分，这些得分指示为目标字段指定结果为真的可能性。除了这些得分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向得分。原始倾向得分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该得分可能会过度优化。调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。此选项要求在流中定义分区字段并且在生成模型之前在建模节点中启用调整的倾向得分。

章
16

最近相邻元素模型

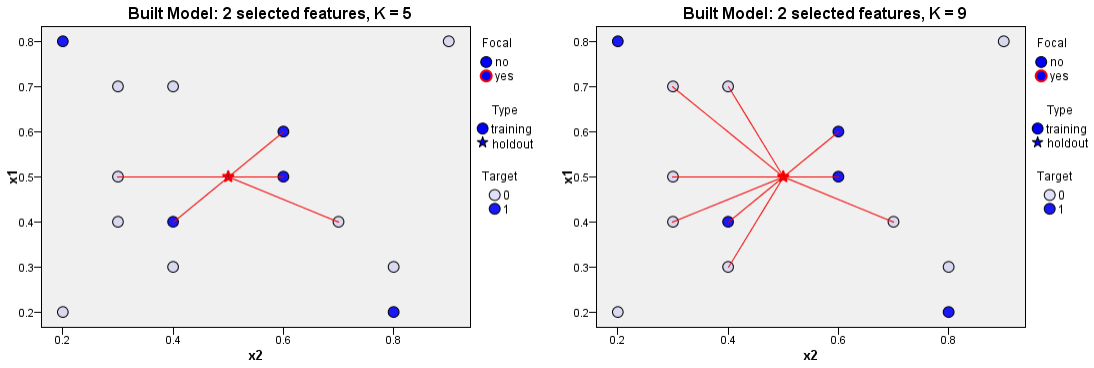
KNN 节点

“最近相邻元素分析”是根据观测值与其他观测值的类似程度分类观测值的方法。在机器学习中，将其开发为识别数据模式的一种方法，而不需要与任何存储模式或观测值完全匹配。相似个案相互邻近，非相似个案则相互远离。因此，两个观测值之间的距离是其不相似性的测量。

将靠近彼此的个案视为“相邻元素。”当提出新的观测值（保留观测值）时，计算其到模型中每个观测值的距离。计算最相似观测值 - 最近相邻元素 - 的分类并将新观测值放在包含最多最近相邻元素的类别中。

您可以规定需要检验的最近相邻元素的数量；此值叫做k。图片显示如何使用两个不同的 k 值分类新观测值。当 k = 5 时，新观测值将被置于类别 1 中，因为大多数最近相邻元素属于类别 1。但当 k = 9 时，新观测值将被置于类别 0 中，因为大多数最近相邻元素属于类别 0。

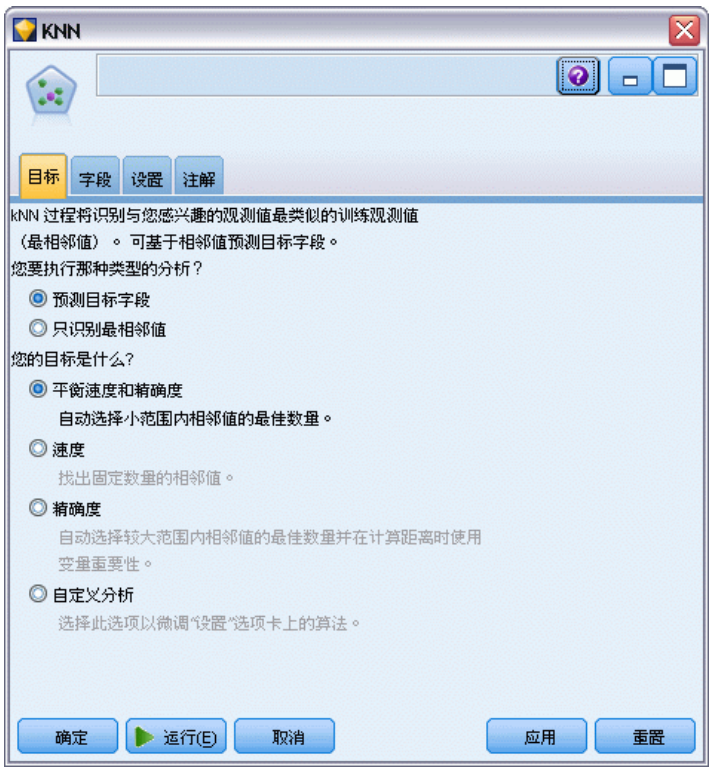
图片 16-1
更改 k 对分类的影响



最近相邻元素分析也可用于计算连续目标的值。在此情况下，最近相邻元素的平均值或中间目标值用于获得新观测值的预测值。

KNN 节点目标选项

图片 16-2
KNN 节点目标选项



您可以在“对象”选项卡输入数据中根据最近相邻元素的值选择构建预测目标字段值的模型，或者只是查找特定感兴趣个案的最近相邻元素。

您要执行哪种类型的分析？

预测目标字段。如果您想根据最近相邻元素的值预测目标字段的值，请选择此选项。

只识别最近相邻元素。如果您只想看到特定字段的最近相邻元素，请选择此选项。

如果您选择只识别最近相邻元素，在此选项卡上与准确性和速度相关的剩余选项将被禁用，因为其只与预测目标相关。

您的目标是什么？

此组选项让您决定当预测目标字段时，速度、准确性或二者是否是最重要的因素。或者您可以选择自己自定义设置。

如果您选择平衡、速度或准确性选项，则算法预先选择该选项的最合适设置组合。高级用户可能希望覆盖这些选择；可在“设置”选项卡上的各个窗格上进行此操作。

均衡速度和精确度。选择小范围内相邻元素的最佳数量。

速度。查找固定数量的相邻元素。

准确性。选择较大范围内的相邻元素的最佳数量，并在计算距离时使用预测变量重要性。

自定义分析。选择该选项以微调“设置”选项卡上的算法。

注意：所得 KNN 模型的大小与多数其他模型不同，随着训练数据量的增加呈线性增加。如果在尝试构建 KNN 模型时看到报告“内存溢出”错误的出错信息，则尝试增加 IBM® SPSS® Modeler 所使用的最大系统内存。要进行此操作，请选择
工具 > 选项 > 系统选项

并在**最大内存**字段中输入新大小。“系统选项”对话框中所作的更改要在重新启动 SPSS Modeler 之后才能生效。

KNN 节点设置

在“设置”选项卡上您可以指定最近相邻元素分析特有的选项。屏幕左侧的侧栏列出了用于指定选项的面板。

模型

图片 16-3
KNN 节点模型选项

“模型”窗格提供控制如何构建模型的选项，例如是否使用分区或分割模型、是否变换数值输入字段以使其落入相同范围内和如何管理感兴趣个案。您也可以给模型选择一个自定义名称。

模型名称。用户可根据目标或 ID 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）或指定一个自定义的名称。

使用分区数据。 如果定义了分区字段，则此选项可确保仅训练分区的数据用于构建模型。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

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。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

手动选择字段...默认情况下，节点使用来自“类型”节点的分区与分割字段设置（如果存在），但此处您可以覆盖这些设置。要激活**分区**与**分割**字段，请选择**字段**选项卡，并选择**使用定制设置**，然后返回此处。

- **分区字段。**该字段允许您使用指定字段将数据分割为几个不同的样本，分别用于模型构建过程中的训练、测试和验证阶段。通过用某个样本生成模型并用另一个样本对模型进行测试，您可以预判出此模型对类似于当前数据的大型数据集的拟合优劣。如果已使用类型或分区节点定义了多个分区字段，则必须在每个用于分区的建模节点的“字段”选项卡中选择一个分区字段。（如果仅有一个分区字段，则将在启用分区后自动引入此字段。）有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。同时请注意，要在分析时应用选定分区，同样必须启用节点“模型选项”选项卡中的分区功能。（取消此选项，则可以在不更改字段设置的条件下禁用分区功能。）
- **分割。**对于分割模型，选择分割字段或字段。此操作与在“类型”节点中将字段的角色设置为分割类似。您可以仅将类型为**标志**、**名义**或**有序**的字段指定为分割字段。选为分割字段的字段无法用作目标、输入、分区、频率或权重字段。 有关详细信息，请参阅第 26 页码第 3 章中的构建分割模型。

标准化范围输入。选中此复选框为连续输入字段标准化值。标准化特征具有相同的值范围，这可改进估计算法的性能。使用经调整后的标准化 [2*(x–min)/(max–min)]–1。调整后的标准化值介于 –1 和 1 之间。

使用个案标签。选中此复选框以启用下拉列表，从这里您可以选择字段并将其值用作标签，以在“模型浏览器”中标识在预测变量空间图表、对等图表和象限图中所需的个案。您可以选择测量级别为名义、有序或标志的任何字段用作标签字段。如果您不在这里选择字段，则用以源数据中行号标识的最近相邻元素在“模型浏览器”图表中显示记录。如果您在构建模型之后要操作数据，可使用个案标签，以避免每次需要参考源数据在显示中标识个案。

识别焦点记录。选中此复选框启用下拉列表，允许您标记感兴趣的输入字段（仅针对标志字段）。如果在此处指定了一个字段，则当构建模型时会在模型浏览器中初始选中代表该字段的点。在此处选择焦点记录是可选的；任何点都可以暂时成为焦点记录，只要在“模型浏览器”中手动选中它。

相邻元素

图片 16-4
KNN 节点相邻元素选项

“相邻元素”窗格具有一组控制如何计算最近相邻元素数量的选项。

最近邻元素的数目 (k)。指定特定个案的最近相邻元素数量。注意，使用大量的邻元素不一定会得到更准确的模型。

如果目标是预测目标，则您具有两个选择：

- **指定固定值 k。**如果您希望指定要查找的最近相邻元素的固定数量，则使用该选项。
- **自动选择 k。**您也可以使用**最小值**和**最大值**字段以指定一个数值范围，并允许该过程选择该范围内相邻元素的“最佳”数量。确定最近相邻元素数目的方法依赖于“特征选择”窗格上要求的特征选择。

  如果特征选择有效，则针对请求范围中每个 k 值执行特征选择，并选择具有最低误差率（如果目标为连续，则为最低平方和误差）的 k 值和特征集。

  如果特征选择无效，则使用 V 折交叉验证来选择“最佳”的邻元素数目。请参阅“交叉验证”窗格以控制折叠指定。

距离计算。该度规用于指定在测量个案相似性中使用的距离度规。

- **Euclidean 度规。**两个个案 x 和 y 之间的距离，为个案值之间的平方差在所有维度上之和的平方根。
- **城市街区度规。**两个个案之间的距离是个案值之间绝对差在所有维度上之和。又称为 Manhattan 距离。

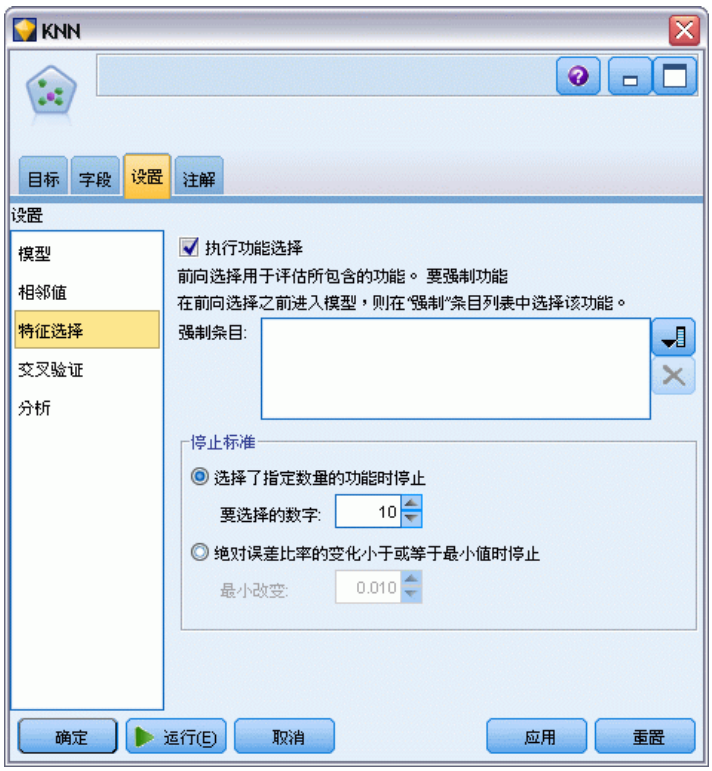
或者，如果目标是预测目标，您可以选择在计算距离时按照其标准化重要性计算特征权重。预测变量的特征重要性的计算方法为：不含预测变量的模型的误差率或平方和误差与完整模型的误差率或平方和误差之比。通过重新对特征重要性值指定权重，来计算标准化的重要性，因此其总和为 1。

计算距离时按照重要性计算特征权重。（只有当目标是预测目标时才显示。）选中此复选框，当计算相邻元素之间距离时，使用预测变量重要性。预测变量重要性将在模型块中显示，并用于预测（因此影响记分）。 有关详细信息，请参阅第 45 页码 第 3 章中的预测变量重要性。

范围目标预测。（只有当目标是预测目标时才显示。）如果指定了连续（数值范围）目标，这可指定预测值是基于最近相邻元素的均值还是中值来计算的。

特征选择

图片 16-5
KNN 节点特征选择选项



只有在目标是预测目标时才激活此窗格。使您能够为特征选择请求和指定选项。默认情况下，特征选择会考虑所有特征，但可以选择特征子集以强制纳入模型。

执行特征选择。选中此复选框启用特征选择选项。

- **强制输入。**单击此框旁的字段选择按钮并选择一个或多个特征以强制纳入模型。

中止准则。在每一步上，如果添加特征可以使误差最小（计算为分类目标的误差率和连续目标的平方和误差），则考虑将其纳入模型中。继续向前选择，直到满足指定的条件。

- **当已选择指定数量的特征时停止。**除了那些强制纳入模型的特征外，算法还会添加固定数目的特征。指定一个正整数。减少所选择的数目值可以创建更简约的模型，但存在缺失重要特征的风险。增加所选择的数目值可以涵盖所有重要特征，但又存在因特征添加而增加模型误差的风险。
- **当绝对误差比率变化小于或等于最小值时停止。**当绝对误差比率变化表明无法通过添加更多特征来进一步改进模型时，算法会停止。指定一个正数。减少最小变化值将倾向于包含更多特征，但存在包含对模型价值不大的特征的风险。增加最小变化值将倾向于排除更多特征，但存在丢失对模型较重要的特征的风险。最小变化的“最佳”值将取决于您的数据和具体应用。请参阅输出中的“特征选择误差日志”，以帮助您评估哪些特征最重要。 有关详细信息，请参阅第 411 页码预测变量选择错误日志。

交叉验证

图片 16-6
KNN 节点交叉验证选项

只有在目标是预测目标时才激活此窗格。该窗格上的选项控制计算最近相邻元素时是否使用交叉验证。

交叉验证将样本划分为许多子样本，或**折叠**。然后，生成最近邻元素模型，并依次排除每个子样本中的数据。第一个模型基于第一个样本折的个案之外的所有个案，第二个模型基于第二个样本折的个案之外的所有个案，依此类推。对于每个模型，估计其错误的方法是将模型应用于生成它时所排除的子样本。“最佳”最近邻元素数为在折中产生最小误差的数量。

交叉验证折叠。V 折交叉验证用于确定“最佳”邻元素数目。因性能原因，它无法与特征选择结合使用。

- **随机分配个案到折。**指定应当用于交叉验证的折数。该过程将个案随机分配到折，从 1 编号到 V（折数）。
- **设置随机数种子。**根据随机百分比估计模型准确性时，此选项可用于在另一会话中复制相同结果。通过指定随机数生成器所使用的起始值，可以确保在每次执行节点时都会分配相同的记录。输入所需的种子值。如果未选中该选项，则每次执行节点时会生成不同的抽样。
- **使用字段分配个案。**指定一个将活动数据集中的每个个案分配到折中的数值字段。字段必须为数值，其值为从 1 到 V 的数字。如果此范围中的任何值缺失，且位于任何分割字段上（如果分割模型有效），这将导致误差。

分析

图片 16-7
KNN 节点分析选项

只有在目标是预测目标时才激活“分析”窗格。您可以使用它指定模型是否要纳入附加变量以包含：

- 每个可能目标字段值的概率
- 个案和最近相邻元素之间的距离
- 原始和调整后的倾向得分（仅适用于标志目标）。

追加所有概率。 如果选中该选项，则为由节点处理的各个记录显示名义或标志目标字段的各个可能值的概率。如果未选中该选项，则仅为名义或标志目标字段显示预测值及其概率。

保存观测值和 k 最近相邻元素之间的距离。对于每个焦点记录，为其 k 个最近相邻元素（来自培训样本）和相应的 k 个最近距离创建单独的变量。

倾向得分

可以在建模节点中和模型块的“设置”选项卡上启用倾向得分。该功能仅在所选目标为标志字段时才可用。有关详细信息，请参阅第 36 页码第 3 章中的倾向得分。

计算原始的倾向得分。原始的倾向得分仅从基于训练数据的模型中导出。如果模型预测值为真（将响应），则倾向与 P 相同，其中 P 为预测的可能性。如果模型预测的值为假，则计算出的倾向为 (1 - P)。

- 如果构建模型时选择了此选项，则默认情况下将在模型块中启用倾向得分。不过，无论是否在建模节点中选择了原始倾向得分，都可以始终在模型块中选择启用原始倾向得分。
- 对模型进行评分时，原始倾向得分将被添加到将 RP 字母附加到标准前缀的字段中。例如，如果预测位于名为 $R-churn 的字段中，则倾向得分字段的名称将是 $RRP-churn。

计算调整后的倾向得分。原始倾向仅仅基于由可能过度拟合的模型给定的估计上，这样会导致过于乐观地评估倾向。调整后的倾向尝试通过查看模型在检验或验证分区的性能或通过调整倾向来弥补，以相应地给作出更好的估计。

- 此设置要求流中存在有效的分区字段。有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。
- 与原始置信度分数不同，调整后的倾向得分必须在构建模型时计算；否则，对模型块进行评分时该分数将不存在。
- 对模型进行评分时，在将 AP 字母附加到标准前缀的字段中添加调整后的倾向得分。例如，如果预测位于名为 $R-churn 的字段中，则倾向得分字段的名称将是 $RAP-churn。调整后的倾向得分不适用于 logistic 回归模型。
- 在计算调整后的倾向得分时，必须尚未平衡用于计算的检验或验证分区。为避免这一点，请确保在任何上游平衡节点中选中**仅平衡训练数据**选项。有关详细信息，请参阅第 3 章中的为平衡节点设置选项中的IBM SPSS Modeler 14.2 源、过程和输出节点。此外，如果已在上游获取了复杂样本，则会导致调整后的倾向得分无效。
- 调整后的倾向得分不适用于“增强型”树和规则集模型。有关详细信息，请参阅第 158 页码第 6 章中的增强型 C5.0 模型。

基于。对于有待计算的调整后的倾向得分，流中必须存在一个分区字段。可以指定是使用检验分区还是验证分区进行此计算。为获取最佳结果，检验或验证分区包含的记录数量应至少与用于训练原始模型的分区所包含的记录数相同。 有关详细信息，请参阅第 4 章中的分区节点中的IBM SPSS Modeler 14.2 源、过程和输出节点。

KNN 模型块

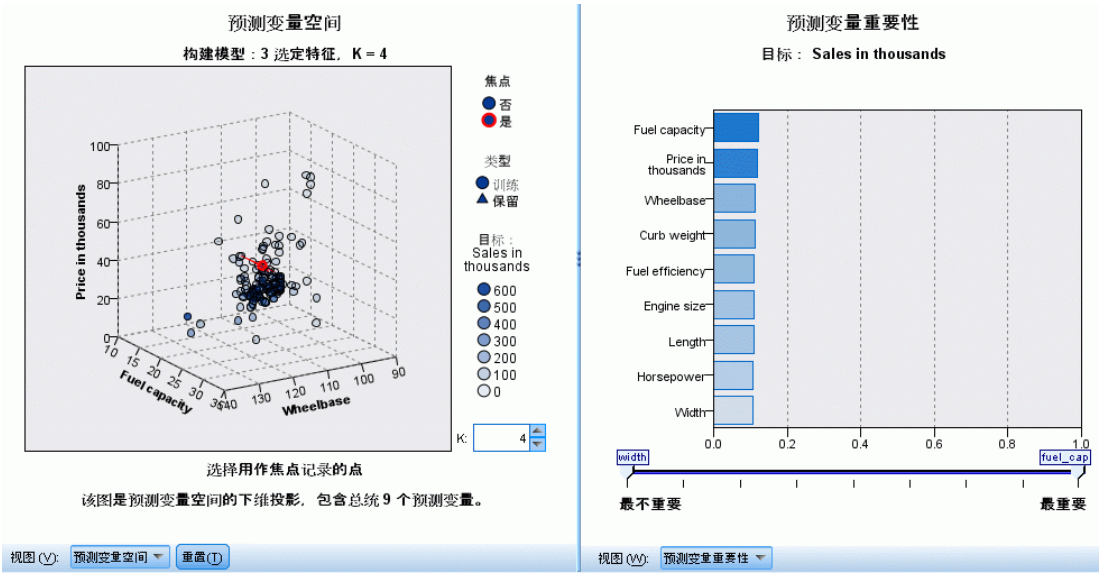
KNN 模型会创建许多新字段，如下表所示。要查看这些字段及其值，请将表节点添加到 KNN 模型块中，然后执行表节点，或单击模型块上的“预览”按钮。

表 16-1
KNN 模型字段

新字段名	描述
$KNN-fieldname	目标字段预测值。
$KNNP-fieldname	预测值概率。
$KNNP-value	名义或标志字段的每个可能值的概率。只有在模型块的“设置”选项卡上选中了**追加所有概率**才会被纳入。
$KNN-neighbor-n	焦点记录的第 n 个最近相邻元素名称。只有当模型块的“设置”选项卡上的**显示最近**设为非零值时才会被纳入。
$KNN-distance-n	焦点记录第 n 个最近相邻元素到焦点记录的相对距离。只有当模型块的“设置”选项卡上的**显示最近**设为非零值时才会被纳入。

模型视图

图片 16-8
最近邻元素分析模型视图



此模型视图有 2 个面板窗口：

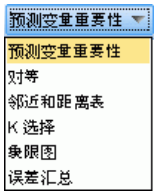
- 第一个面板显示模型概览，称为主视图。
- 第二个面板显示两种视图类型之一：

辅助模型视图显示有关模型的更多信息，但并不专注于模型本身。

当用户深入查看主视图某个部分时，链接视图显示有关某个模型特征的详细信息。

默认情况下，第一个面板显示预测变量空间，第二个面板显示预测变量重要性图表。如果预测变量重要性图表不可用；即在“设置”选项卡的“相邻元素”面板上未选中**按照重要性计算特征权重**时，显示“视图”下拉列表中的第一个可用视图。

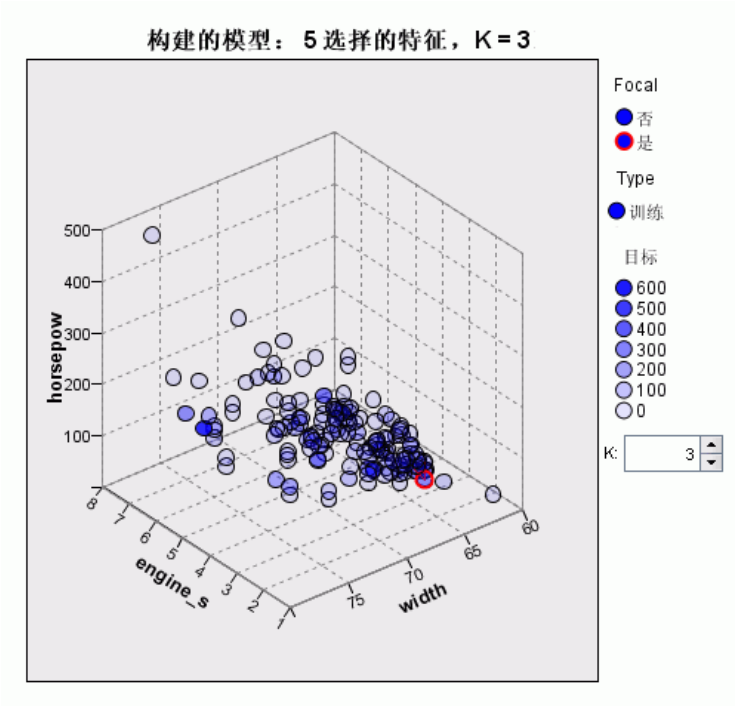
图片 16-9
“最近邻元素分析模型视图”下拉列表



如果视图不具有可用信息，它将从“视图”下拉列表中省略。

预测变量空间

图片 16-10
预测变量空间



预测变量空间图表是有关预测变量空间（如果存在 3 个以上预测变量，则为子空间）的交互式图形。每条轴代表模型中的某个预测变量，图表中的点位置显示个案这些预测变量在训练和坚持分区中的值。

键。除了预测变量值外，图中的点还传递其他信息。

- 其形状表示点所属的分区，即训练或坚持分区。
- 点的颜色/阴影表示该个案的目标值，不同的颜色值等于分类目标的类别，阴影则表示连续目标的值范围。训练分区的指示值为观测值；对于坚持分区，则为预测值。如果未指定目标，则不会显示此键。
- 较粗的轮廓表示个案为焦点个案。显示的焦点记录链接到它们的 k 个最近邻元素。

控制和互动。使用图表中的一些控件可以探索预测变量空间。

- 可以选择在图表中显示哪个预测变量子集，还可更改在维度上表示哪些预测变量。
- “焦点记录”仅仅是在预测变量空间图表中所选的点。如果指定了焦点记录变量，则初始情况下会选中代表焦点记录的点。不过,任何点都可以暂时成为焦点记录，只要您选中它.可以使用点选择的“常规”控制，即单击某个点以选择该点并取消选中其它点；按住 Ctrl 键并单击某个点可将其添加到所选的点集中。链接的视图，如对等图表，将根据在预测变量空间中选择的个案自动更新。
- 您可以更改为焦点记录显示的最近邻元素数目 (k)。
- 在图表中的点上方悬停，可以显示工具提示以及个案标签值，或个案编号（如果未定义个案标签），以及观察和预测目标值。
- 使用“重置”按钮可以将预测变量空间恢复到其原始状态。

更改预测变量空间图表上的轴

您可以控制在预测变量空间图表的轴上显示的特征。

要更改轴设置：

▶ 单击左侧面板上的“编辑模式”按钮（画笔图标），为预测变量空间选择编辑模式。

▶ 在右侧面板中更改视图。在两个主面板之间出现**显示区域**面板。

▶ 单击**显示区域**复选框。

▶ 单击预测变量空间中的任何数据点。

▶ 要使用具有相同数据类型的预测变量替换某个轴：

将新预测变量拖到您要替换的预测变量的区域标签（带有小 X 按钮）上。

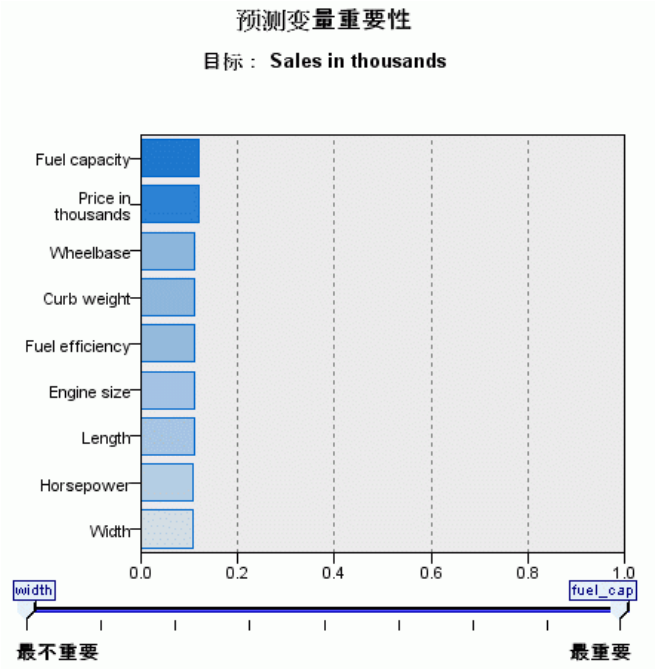
▶ 要使用具有不同数据类型的预测变量替换某个轴：

- 在您要替换的预测变量的区域标签上，单击小 X 按钮。预测变量空间变为二维视图。
- 将新预测变量拖到**添加维度**区域标签上。

▶ 单击左侧面板上的“探索模式”按钮（箭头图标），退出编辑模式。

预测变量重要性

图片 16-11
预测变量重要性



通常，您将需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的变量。预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。预测变量的重要性与模型精确性无关。它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

最近邻元素距离

图片 16-12
最近邻元素距离

k 个最近相邻和距离

对**初始焦点**记录显示

焦点记录	最近相邻				最近距离			
	1	2	3	4	1	2	3	4
72	157	10	67	71	0.079	0.088	0.107	0.117

该表只显示焦点记录的 k 个最近邻元素与距离。如果焦点记录标识符指定在建模节点上，则它为可用，且只显示此变量标识的焦点记录。

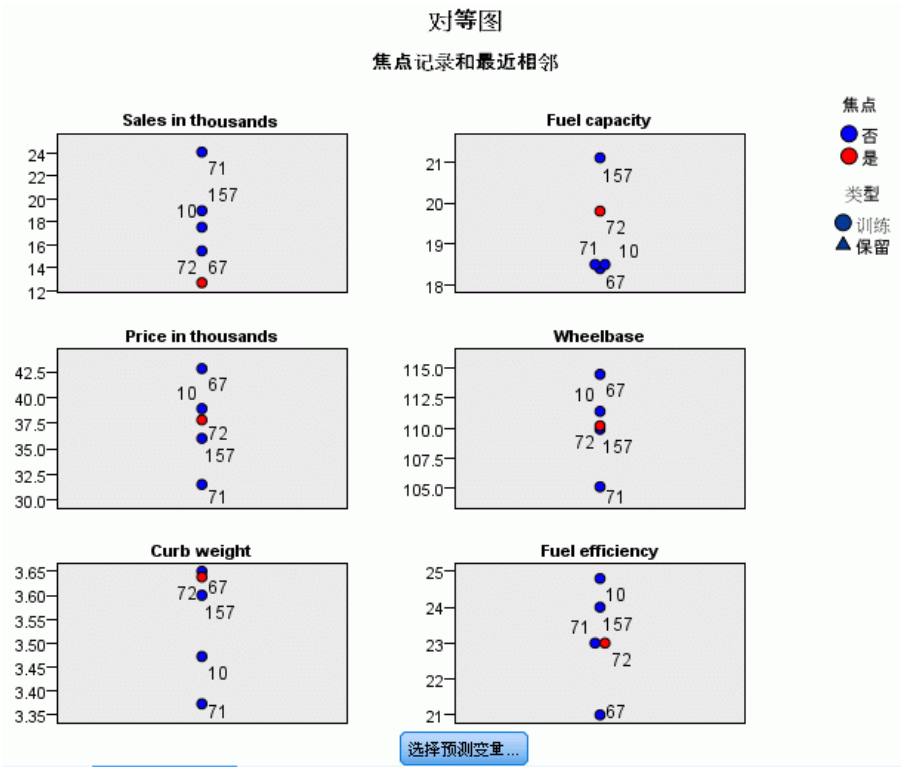
每行：

- **焦点记录**列包含焦点记录的个案标签变量值；如果未定义个案标签，则此列包含焦点记录的个案编号。

- 在**最近相邻元素**组下的第 i 列包含焦点记录的第 i 个最近相邻元素的个案标签变量值；如果未定义个案标签，则此列包含焦点记录第 i 个最近相邻元素的个案号。
- 在**最近距离**组下的第 i 列包含第 i 个最近相邻元素与焦点记录的距离。

对等

图片 16-13
对等图表



该图表显示焦点个案及其在每个预测变量和目标上 k 个最近邻元素。它仅在预测变量空间图表中选择了焦点个案时可用。

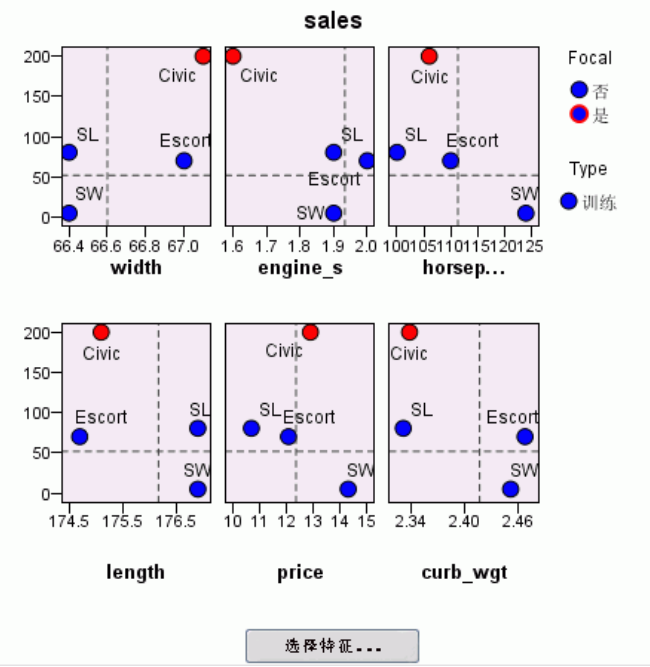
对等图表以两种方式链接到预测变量空间。

- 在预测变量空间中所选的个案（焦点个案）显示在对等图表中，也包括其 k 个最近邻元素。
- 在对等图表中使用在预测变量空间中所选的 k 值。

选择预测变量。使您可选择在对等图表中显示的预测变量。

象限图

图片 16-14
象限图



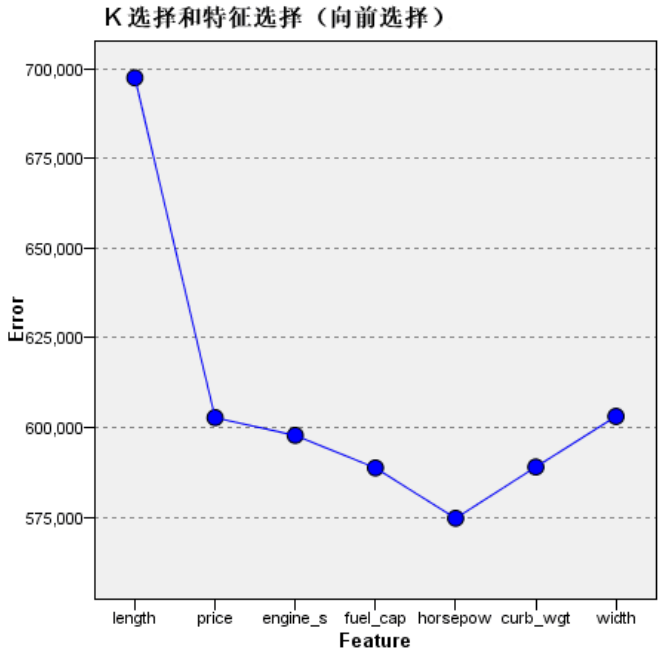
该图表显示焦点个案及其在散点图（或点图，取决于目标的测量级别）上 k 个最近邻元素。目标在 y 轴上，刻度预测变量在 x 轴上，按预测变量划分面板。它仅当存在目标，且在预测变量空间图表中选择了焦点个案时可用。

- 在训练分区的变量均值处，为连续变量绘制了参考线。

选择预测变量。使您可选择在象限图中显示的预测变量。

预测变量选择错误日志

图片 16-15
预测变量选择



对于该图表上的点，其 y 轴值为模型的误差（误差率或平方和误差，取决于目标的测量级别），x 轴上列出模型的预测变量（加上 x 轴左侧的所有特征）。该图表仅当存在目标，且特征选择有效时可用。

分类表

图片 16-18
分类表

分区		预测值		
		0	1	正确率百分比
训练	0	111	1	99.11%
	1	7	33	82.50%
	总计百分比	77.64%	22.37%	94.74%

该表显示按分区对目标观察与预测值的交叉分类。它仅当存在分类目标（标志、名义或有序）时可用。

- 坚持分区中的（缺失）行包含在目标上具有缺失值的坚持个案。这些个案对“坚持样本：整体百分比”有贡献，但对“正确百分比”无影响。

误差摘要

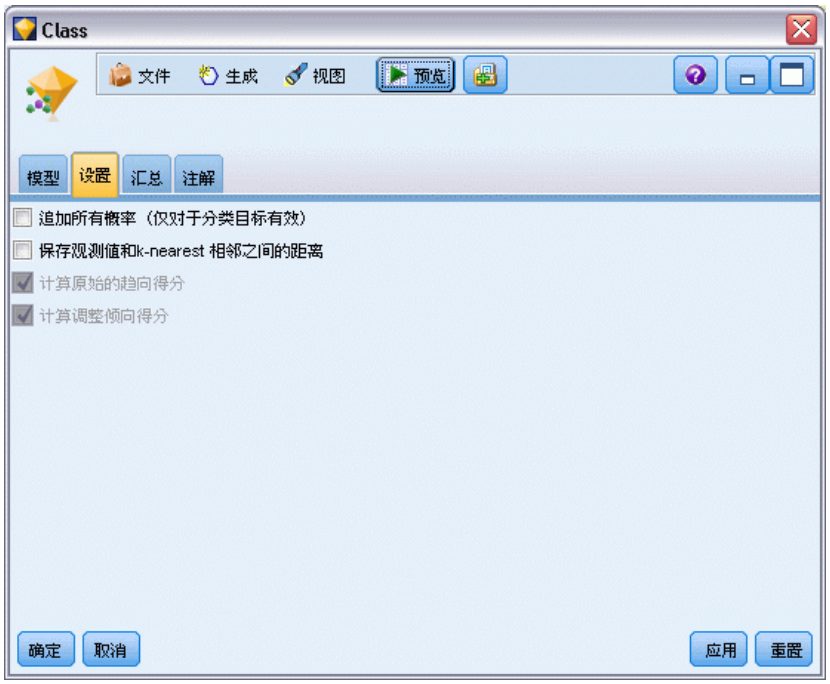
图片 16-19
误差摘要

分区	平方和错误
训练	622043

它仅当存在目标变量时可用。它显示模型相关误差，对于连续目标为平方和误差，对于分类目标为误差率（100% – 整体正确百分比）。

KNN 模型设置

图片 16-20
KNN 模型块设置



通过“设置”选项卡可以指定在查看结果时显示的附加字段（例如，通过执行表节点附加到块）。通过选择这些选项可以查看每个选项的效果，并且单击“预览”按钮（滚动至“预览”输出右侧）可以查看附加字段。

追加所有概率（仅对分类目标有效）。 如果选中该选项，则为由节点处理的各个记录显示名义或标志目标字段的各个可能值的概率。如果未选中该选项，则仅为名义或标志目标字段显示预测值及其概率。

此复选框的默认设置由建模节点的相应复选框确定。

计算原始的倾向得分。对于含标志目标（返回?是?或?否?预测）的模型，您可以请求倾向得分，这些得分指示为目标字段指定结果为真的可能性。除了这些得分，还有其他在评分过程中生成的预测值和置信度值。

计算调整后的倾向得分。原始倾向得分仅依赖于训练数据，并且由于许多模型过度拟合此数据的倾向，该得分可能会过度优化。调整后的倾向会尝试通过针对检验或验证分区对模型性能进行评估进行弥补。此选项要求在流中定义分区字段并且在生成模型之前在建模节点中启用调整的倾向得分。

显示最近。如果您将此值设为 n，其中 n 是非零正整数，则焦点记录的第 n 个最近相邻元素与其到焦点记录的相对距离一起纳入在模型中。

附录
A

注意事项

This information was developed for products and services offered worldwide.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing, Legal and Intellectual Property Law, IBM Japan Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242-8502 Japan.

以下段落不适用于英国或任何其他此类条款与其当地法律不一致的国家： SPSS INC.,IBM COMPANY 一员，“按原样”提供本出版物，不包含任何类型的保证，无论是明示或默示的，包括但不只限于不侵权的默示保证、适销性或适用特定目的。一些国家不允许某些交易中明示或默示保证的免责声明，因此本声明可能不适用于您。

本信息可能包含技术不准确性或印刷错误。我们将定期对以上信息进行更改；这些更改将出现在本出版物的最新版本中。SPSS Inc. 可能在任何时候对本出版物中介绍的产品和/或程序进行改进而不另行通知。

本信息中引用的任何非 SPSS 和非 IBM 网站只用于参考目的，在任何情况下都不作为对这些网站的背书。这些网站上的资料不是本 SPSS Inc. 产品资料的一部分，同时您要自行承担使用这些网站的风险。

当您发送信息给 IBM 或 SPSS 时，您将授予非独占权利给 IBM 和 SPSS，允许它以其认为合适的任何方式使用或分发这些信息而不承担任何责任。

有关非 SPSS 产品的信息分别来自这些产品的供应商、已出版的公告或其它公开的来源。SPSS 尚未测试这些产品，同时无法确认性能的准确性、兼容性或与非 SPSS 产品相关的任何其他声明。如果对非 SPSS 产品的性能有任何疑问，请咨询这些产品的供应商。

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group, Attention: Licensing, 233 S. Wacker Dr., Chicago, IL 60606, USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

本信息包含用于日常商业运营的数据和报告示例。为了尽可能完整的阐明，这些示例包含个人姓名、公司、品牌和产品名称。所有这些名称都是虚构的，任何与实际公司名称和地址类似的情况实属巧合。

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

商标

IBM、IBM 徽标、和 ibm.com 是 IBM Corporation 在全球多个国家注册的商标。有关 IBM 商标的当前列表，请访问公司网站，网址为 http://www.ibm.com/legal/copytrade.shmtl。

SPSS 是 ，已在全球多个国家注册。

Adobe、Adobe 徽标、PostScript 和 PostScript 徽标是 Adobe Systems Incorporated 在美国和/或其他国家的注册商标或商标。

IT Infrastructure Library 是中央计算机与电信总局的注册商标，该局目前是英国商务部的一部分。

索引