

IBM SPSS Modeler 15 数据库内 数据挖掘指南



注意：使用本信息及其支持的产品之前，请阅读注意事项第 页码下的一般信息。

此版本适用于 IBM SPSS Modeler 15 及所有后续发布和修订，除非在新版本中另有说明。

Adobe 产品屏幕截图重印已获得 Adobe Systems Incorporated 的许可。

Microsoft 产品屏幕截图重印已获得 Microsoft Corporation 的许可。

受许可保护材料 - IBM 所有

Copyright IBM Corporation 1994, 2012.

美国政府用户受限权利 - 使用、复制或披露受与 IBM Corp. 签订的 GSA ADP Schedule Contract 的限制。

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

1	关于 IBM SPSS Modeler	1
	IBM SPSS Modeler 产品	1
	IBM SPSS Modeler	1
	IBM SPSS Modeler Server	2
	IBM SPSS Modeler Administration Console	2
	IBM SPSS Modeler Batch	2
	IBM SPSS Modeler Solution Publisher	2
	IBM SPSS Modeler Server 适配器用于 IBM SPSS Collaboration and Deployment Services	2
	IBM SPSS Modeler 版本	3
	IBM SPSS Modeler 文档	3
	SPSS Modeler Professional 文档	4
	SPSS Modeler Premium 文档	4
	应用程序示例	5
	Demos 文件夹	6
2	数据库内数据挖掘	7
	数据库建模概述	7
	您的需要	8
	建立模型	9
	数据准备	9
	模型评分	9
	输出和保存数据库模型	10
	模型一致性	11
	查看和导出生成的 SQL	12
3	使用 Microsoft Analysis Services 进行数据库建模	13
	IBM SPSS Modeler 与 Microsoft Analysis Services	13
	与 Microsoft Analysis Services 集成的要求	14
	启用与 Analysis Services 的集成	15
	使用 Analysis Services 构建模型	18
	管理 Analysis Services 模型	19
	对所有算法节点通用的设置	20
	MS 决策树专家选项	22
	MS 聚类专家选项	23
	MS Naive Bayes 专家选项	24

MS 线性回归专家选项	25
MS 神经网络专家选项	26
MS Logistic 回归专家选项	27
MS 关联规则节点	27
MS 时间序列节点	28
MS 序列聚类节点	31
对 Analysis Services 模型评分	33
对所有 Analysis Services 模型通用的设置	34
MS 时间序列模型块	36
MS 序列聚类模型块	40
导出模型和生成节点	40
Analysis Services 数据挖掘示例	40
示例流：决策树	40

4 使用 Oracle Data Mining 构建数据库模型 48

关于 Oracle Data Mining	48
集成 Oracle 的要求	48
启用 Oracle 集成	49
使用 Oracle Data Mining 构建模型	51
Oracle 模型服务器选项	52
误分类损失	54
Oracle Naive Bayes	55
Naive Bayes 模型选项	55
Naive Bayes 专家选项	56
Oracle Adaptive Bayes	56
Adaptive Bayes 模型选项	57
Adaptive Bayes 专家选项	58
Oracle Support Vector Machine (SVM)	59
Oracle SVM 模型选项	59
Oracle SVM 专家选项	61
Oracle SVM 权重选项	62
Oracle 广义线性模型 (GLM)	62
Oracle GLM 模型选项	63
Oracle GLM 专家选项	64
Oracle GLM 权重选项	65
Oracle 决策树	65
决策树模型选项	66
决策树专家选项	67

Oracle O-Cluster	68
O-Cluster 模型选项	68
O-Cluster 专家选项	69
Oracle k-Means	69
k-Means 模型选项	70
k-Means 专家选项	71
Oracle 非负矩阵分解 (NMF).	71
NMF 模型选项	72
NMF 专家选项	73
Oracle Apriori	73
Apriori 字段选项	74
Apriori 模型选项	76
Oracle 最小描述符长度 (MDL).	77
MDL 模型选项	78
Oracle 属性重要性 (AI)	78
AI 模型选项	79
AI 选择选项	79
AI 模型块模型选项卡	80
管理 Oracle 模型	81
Oracle 模型块服务器选项卡	82
Oracle 模型块汇总选项卡	82
Oracle 模型块设置选项卡	83
列出 Oracle 模型	83
Oracle Data Miner.	84
准备数据	86
Oracle Data Mining 示例	87
示例流: 上载数据	87
示例流: 探索数据	88
示例流: 构建模型	89
示例流: 评估模型	90
示例流: 部署模型	93

5 使用 IBM InfoSphere Warehouse 进行数据库建模 94

IBM InfoSphere Warehouse 和 IBM SPSS Modeler.	94
集成 IBM InfoSphere Warehouse 的要求	94
启用 IBM InfoSphere Warehouse 集成	95
使用 IBM InfoSphere Warehouse Data Mining 构建模型.	102
模型评分和部署	102
管理 DB2 模型.	103
列出数据库模型	103

浏览模型	104
导出模型和生成节点	105
对所有算法通用的节点设置	105
ISW 决策树	108
ISW 决策树模型选项	109
ISW 决策树专家选项	110
ISW 关联	110
ISW 关联字段选项	111
ISW 关联模型选项	113
ISW 关联专家选项	114
ISW 分类法选项	114
ISW 序列	117
ISW 序列模型选项	118
ISW 序列专家选项	119
ISW 回归	119
ISW 回归模型选项	120
ISW 回归专家选项	121
ISW 聚类	124
ISW 聚类模型选项	124
ISW 聚类专家选项	126
ISW Naive Bayes	128
ISW Naive Bayes 模型选项	128
ISW Logistic 回归	129
ISW Logistic 回归模型选项	129
ISW 时间序列	129
ISW 时间序列字段选项	130
ISW 时间序列模型选项	131
ISW 时间序列专家选项	132
显示 ISW 时间序列模型	132
ISW Data Mining 模型块	133
ISW 模型块服务器选项卡	133
ISW 模型块“设置”选项卡	134
ISW 模型块汇总选项卡	136
ISW Data Mining 示例	136
示例流：上载数据	137
示例流：探索数据	137
示例流：构建模型	139
示例流：评估模型	139
示例流：部署模型	141

6 采用 IBM Netezza Analytics 进行数据库建模 143

IBM SPSS Modeler and IBM Netezza Analytics	143
集成 IBM Netezza Analytics 的要求.	143
启用 IBM Netezza Analytics 集成.	144
配置 IBM Netezza Analytics.	144
为 IBM Netezza Analytics 创建 ODBC 源.	144
在 IBM SPSS Modeler 中启用 IBM Netezza Analytics 集成.	146
启用 SQL 生成和优化.	146
采用 IBM Netezza Analytics 构建模型.	147
Netezza 模型 - 字段选项.	149
Netezza 模型 - 服务器选项.	150
Netezza 模型 - 模型选项.	151
Netezza 决策树.	152
实例权重和类权重.	152
Netezza 决策树字段选项.	153
Netezza 决策树构建选项.	154
Netezza K-Means.	158
Netezza K-Means 字段选项.	158
Netezza K-Means 构建选项.	159
Netezza 贝叶斯网络.	160
Netezza 贝叶斯网络字段选项.	160
Netezza 贝叶斯网络构建选项.	161
Netezza Naive Bayes.	162
Netezza KNN.	162
Netezza KNN 模型选项 - 常规.	163
Netezza KNN 模型选项 - 评分选项.	164
Netezza 分裂式聚类.	165
Netezza 分裂式聚类字段选项.	166
Netezza 分裂式聚类构建选项.	167
Netezza PCA.	168
Netezza PCA 字段选项.	168
Netezza PCA 构建选项.	169
Netezza 回归树.	170
Netezza 回归树构建选项 - 树增长.	170
Netezza 回归树构建选项 - 树修剪.	171
Netezza 线性回归.	172
Netezza 线性回归构建选项.	172
Netezza 时间序列.	173
Netezza 时间序列值的插值.	174
Netezza 时间序列字段选项.	176

Netezza 时间序列建构选项	177
Netezza 时间序列模型选项	182
Netezza 广义线性	183
Netezza 广义线性模型选项 - 常规	183
Netezza 广义线性模型选项 - 交互	185
Netezza 广义线性模型选项 - 评分选项	187
管理 IBM Netezza Analytics 模型.	187
IBM Netezza Analytics 模型评分	187
Netezza 模型块服务器选项卡	187
Netezza 决策树模型块.	188
Netezza K-Means 模型块	190
Netezza 贝叶斯网络模型块	192
Netezza Naive Bayes 模型块	193
Netezza KNN 模型块	194
Netezza 分裂式聚类模型块	196
Netezza PCA 模型块	196
Netezza 回归树模型块.	197
Netezza 线性回归模型块.	199
Netezza 时间序列模型块.	199
Netezza 广义线性模型块.	200

附录

A 注意事项	203
---------------	------------

索引	205
-----------	------------

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，或作为客户端与 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 产品

IBM® SPSS® Modeler 系列产品及其相关软件包括如下成员：

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server 适配器用于 IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler 是在个人电脑上安装并运行的完整功能版本产品。您可以在本地模式下将 SPSS Modeler 作为独立产品来运行；也可以在分布式模式下与 IBM® SPSS® Modeler Server 协同使用，从而提高了对大数据集的处理速度。

使用 SPSS Modeler，您可以快速直观地建构精确的预测模型，无需编程。使用其独特的可视界面，您可以轻松地将数据挖掘过程可视化。通过内嵌在该产品中的高级分析支持，您可以发现之前在您的数据中隐藏的模式与趋势。您可以建构输出模型，并理解其影响因子，让您可以更好地利用业务机会、降低风险。

SPSS Modeler 提供两个版本：SPSS Modeler Professional 和 SPSS Modeler Premium。有关详细信息，请参阅 [IBM SPSS Modeler 版本中的 IBM SPSS Modeler 15 用户指南](#)。

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。

SPSS Modeler Server 是一项独立许可的产品，可以协同一份或多份 IBM® SPSS® Modeler 安装，以分布式分析模式在服务器主机上持续运行。因此，SPSS Modeler Server 在处理大型数据集时具有卓越的性能，因为内存密集型的操作可在服务器完成而无需将数据下载到客户端计算机。IBM® SPSS® Modeler Server 还提供了对 SQL 优化的支持与数据库内建模能力，有助于进一步提高性能与自动化程度。

IBM SPSS Modeler Administration Console

Modeler Administration Console 是一个图形化的应用程序，可管理 SPSS Modeler Server 的多项配置选项，配置也可通过编辑一个选项文件来进行。该应用程序提供了一个控制台用户界面，用以监控和配置所安装的 SPSS Modeler Server，SPSS Modeler Server 的现用户可以免费使用该应用程序。应用程序只能安装在 Windows 计算机上；但是它可以管理安装在任何受支持平台上的服务器。

IBM SPSS Modeler Batch

虽然数据挖掘通常是交互式的过程，但是也可以通过命令行来运行 SPSS Modeler 而无需图形用户界面。例如，您可能有些任务需长期或重复性地运行而无用户干预。SPSS Modeler Batch 是该产品一个特殊版本，无需通过常规的用户界面即可完整地实现 SPSS Modeler 的分析功能。使用 SPSS Modeler Batch 需要具备 SPSS Modeler Server 许可。

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher 是一个工具，让您可以创建 SPSS Modeler 流的打包版本，由外部运行时引擎运行，或嵌入一个外部应用程序。以此方式，您可以发布与部署完整的 SPSS Modeler 流，在未安装 SPSS Modeler 的环境下也能使用。SPSS Modeler Solution Publisher 是作为 IBM SPSS Collaboration and Deployment Services - Scoring 服务的一部分来发行的，需另行购买许可。获得许可后，您会收到 SPSS Modeler Solution Publisher Runtime，让您可以执行已发布的流。

IBM SPSS Modeler Server 适配器用于 IBM SPSS Collaboration and Deployment Services

有若干种 IBM® SPSS® Collaboration and Deployment Services 适配器可以让您通过 SPSS Modeler 和 SPSS Modeler Server 与 IBM SPSS Collaboration and Deployment Services 存储库交互。以此方式，部署在存储库中的 SPSS Modeler 流即可实现多用户共享，或通过瘦客户端应用程序 IBM SPSS Modeler Advantage 访问。适配器须安装在存储库的主机系统中。

IBM SPSS Modeler 版本

SPSS Modeler 提供下列版本。

SPSS Modeler Professional

SPSS Modeler Professional 提供您在处理大多数类型的结构化数据（例如 CRM 系统中跟踪的行为或交互活动、人口统计、购买行为与销售数据）时所需的所有工具。

SPSS Modeler Premium

SPSS Modeler Premium 是一项单独许可的产品，它扩展了 SPSS Modeler Professional 的功能，使其可以处理如实体分析或社会网络专门数据，以及非结构化的文本数据。SPSS Modeler Premium 包含下列组件。

IBM® SPSS® Modeler Entity Analytics 在 IBM® SPSS® Modeler 预测分析的基础上添加了全新的维度。预测分析会尝试根据过去数据预测未来行为，而实体分析侧重于通过解析记录自身的身份冲突，提高当前数据的连贯性和一致性。身份可以指个人、组织、对象或可能不确定的任何其他实体的身份。身份解析在许多领域中都非常重要，包括客户关系管理、检测、反洗钱以及国家与国际安全。

IBM SPSS Modeler Social Network Analysis 将关于关系的信息转换为字段，这些字段可描述个人和组社交行为的特征。使用介绍社交网络之下关系的数据，IBM® SPSS® Modeler Social Network Analysis 可识别影响网络中他人行为的社交领导。此外，可确定受其他网络参与者影响最大的人员。通过结合这些结果和其他测量，您可创建个人的综合配置文件，作为预测模型的基础。包括此社交信息的模型比不包括的模型执行效果更好。

Text Analytics for IBM® SPSS® Modeler 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 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。

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（包括安装说明）也在每个产品 DVD 的 \Documentation 文件夹下以 PDF 格式提供。安装文档也可从以下网页中下载：
<http://www-01.ibm.com/support/docview.wss?uid=swg27023172>。

两种格式的文档均可从 SPSS Modeler 信息中心获取，其网址如下：
<http://publib.boulder.ibm.com/infocenter/spssmodl/v15r0m0/>。

SPSS Modeler Professional 文档

SPSS Modeler Professional 的文档套件（不含安装说明）如下：

- **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、Predictive Application 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 3 章中的建模节点概述中的 IBM SPSS Modeler 15 建模节点。](#)
- **IBM SPSS Modeler 算法指南。** 介绍 SPSS Modeler 中所用建模方法的数学基础。此指南仅提供 PDF 版。
- **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 15 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法来改进性能并增强分析功能的信息。
- **IBM SPSS Modeler Server 管理和性能指南。** 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。
- **IBM SPSS Modeler Batch 用户指南。** 在批处理模式下使用 IBM SPSS Modeler 的完整指南，包括批处理模式的执行与命令行参数的详细信息。此指南仅提供 PDF 版。

SPSS Modeler Premium 文档

SPSS Modeler Premium 的文档套件（不含安装说明）如下：

- **IBM SPSS Modeler Entity Analytics 用户指南。** 关于通过 SPSS Modeler 使用实体分析的信息，涵盖存储库的安装与配置、实体分析节点以及管理任务。

- **IBM SPSS Modeler Social Network Analysis 用户指南。** 通过 SPSS Modeler 进行社会网络分析的指南，包括群组分析与传播分析。
- **Text Analytics for SPSS Modeler 用户指南。** 关于通过 SPSS Modeler 使用文本分析的信息，涵盖文本发掘节点、交互式工作台、模板及其他资源。
- **Text Analytics for IBM SPSS Modeler Administration Console 用户指南。** 有关安装和使用控制台用户界面监视和配置 IBM® SPSS® Modeler Server 用于 Text Analytics for SPSS Modeler 的信息。控制台实现为 Deployment Manager 应用程序的插件。

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明的针对性说明。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 15 用户指南。](#)

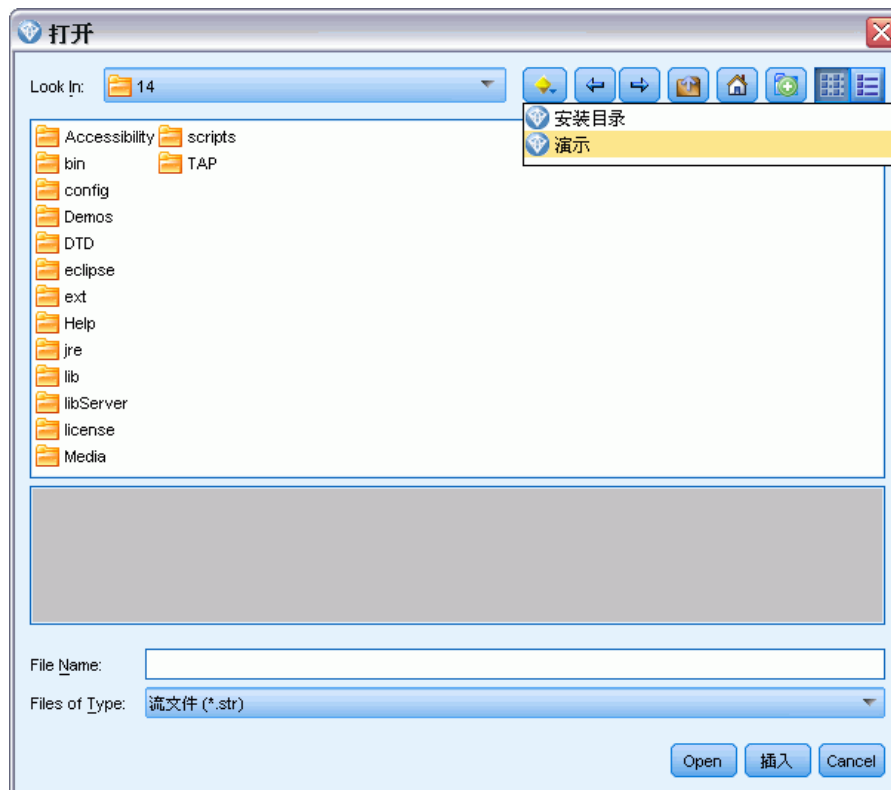
数据库建模示例。 请参阅 IBM SPSS Modeler 数据库内挖掘指南 中的示例。

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 15 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



数据库内数据挖掘

数据库建模概述

IBM® SPSS® Modeler Server 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 IBM Netezza、IBM DB2 InfoSphere Warehouse、Oracle Data Miner 和 Microsoft Analysis Services。您可以使用 IBM® SPSS® Modeler 应用程序在数据库中构建、评分和存储模型。通过集成，可将 SPSS Modeler 的分析功能和易用性将与数据库的功能和性能相结合，同时还兼备数据库提供商提供的数据库自有算法。模型在数据库内创建，然后可以借助 SPSS Modeler 界面以正常方式浏览模型并为之评分，必要时还可使用 IBM® SPSS® Modeler Solution Publisher 来对模型进行部署。在 SPSS Modeler 的“数据库建模”选项板中列出了支持的算法。

使用 SPSS Modeler 访问数据库自有算法的若干优势：

- 数据库内的算法常常与数据库服务器紧密结合，通常可提高性能。
- 在“数据库内”构建和存储的模型不仅由可访问该数据库的应用程序共享，且更易于在这些应用程序之中部署。

SQL 生成。数据库内建模与 SQL 生成（又称为“SQL 回送”）存在明显区别。使用此功能可以生成本地 SPSS Modeler 操作的 SQL 语句，这些语句可“回送”至数据库（即，在其中执行）以提高性能。例如，合并、聚类和选择节点均可生成 SQL 代码，此类代码可以通过上述方式回送至数据库。将 SQL 生成与数据库建模结合使用可以使流自始至终在数据库中运行，相比于在 SPSS Modeler 中运行流，前者具有极大的性能优势。有关详细信息，请参阅第 6 章中的 SQL 优化中的 IBM SPSS Modeler Server 15 管理和性能指南。

注意：数据库建模和 SQL 优化需要在 IBM® SPSS® Modeler 计算机上启用 SPSS Modeler Server 连接。通过启用此设置，您可以访问数据库算法，直接从 SPSS Modeler 回送 SQL 以及访问 SPSS Modeler Server。要验证当前许可证的状态，请从 SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项服务器启用。

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 15 用户指南。

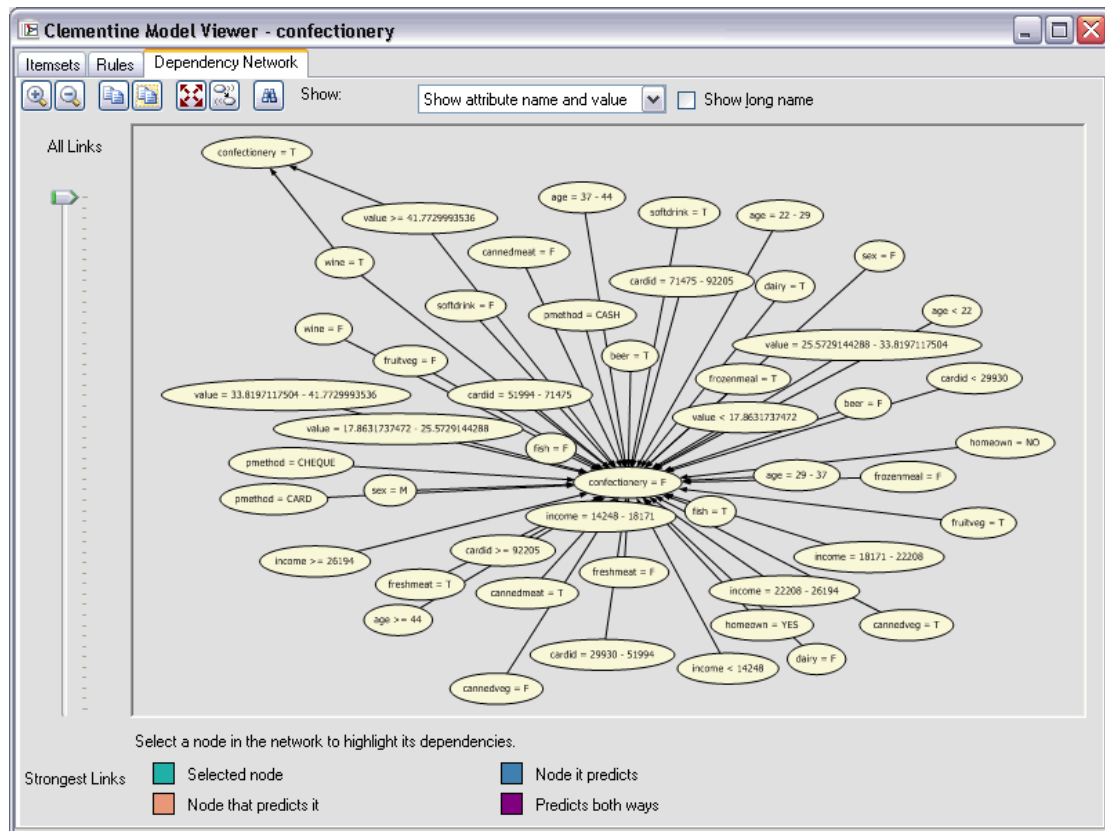
图片 2-1
数据库建模选项板



关于所支持的算法的更多信息，请参阅针对指定提供商的后续章节。

图片 2-2

提供 Microsoft Analysis Services 关联规则模型结果的图形视图的查看器



您的需要

进行数据库建模，需要进行以下设置：

- 在已安装所需分析组件（Microsoft Analysis Services、Oracle Data Miner 或 IBM DB2 InfoSphere Warehouse）的前提下，建立到相应数据库的 ODBC 连接。
- 在 IBM® SPSS® Modeler 中，必须在“辅助应用程序”对话框（工具 > 辅助应用程序）中启用数据库建模。
- 应该启用 IBM® SPSS® Modeler 以及 IBM® SPSS® Modeler Server（如果采用）中“用户选项”对话框内的生成 SQL 和 SQL 优化设置。有关详细信息，请参阅第 4 章中的性能/优化中的 IBM SPSS Modeler Server 15 管理和性能指南。注意，进行数据库建模时不一定要启用 SQL 优化，但如果考虑到性能，则强烈建议启用此项。

注意：数据库建模和 SQL 优化需要在 SPSS Modeler 计算机上启用 SPSS Modeler Server 连接。通过启用此设置，您可以访问数据库算法，直接从 SPSS Modeler 回送 SQL 以及访问 SPSS Modeler Server。要验证当前许可证的状态，请从 SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项服务器启用。

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 15 用户指南](#)。

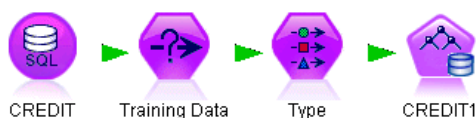
关于详细信息，请参阅针对指定供应商的后续章节。

建立模型

采用数据库算法构建模型和对模型评分的过程类似于 IBM® SPSS® Modeler 中其他类型的数据挖掘。节点和建模“块”的一般处理过程类似于 SPSS Modeler 中其他的流处理过程。唯一的不同是，处理和模型构建实际上是在数据库内进行的。

例如，以下流在概念上与 SPSS Modeler 中的其他数据流相同。但是，此流的所有操作均在数据库中完成（采用 Microsoft 决策树节点），其中包括模型构建。运行流时，SPSS Modeler 会指示数据库构建和存储最终模型，而且详细信息将下载到 SPSS Modeler。

图片 2-3
数据库建模示例流中的紫色节点表示在数据库内执行



数据准备

无论是否采用数据库自有算法，为了提高性能，只要有可能，数据准备均应返回至数据库完成。

- 如果原始数据存储在数据库中，则目标就是通过确保所有需要的上游操作均可转换成 SQL 语句来将数据保存在数据库中。这样可以防止数据被下载至 IBM® SPSS® Modeler—避免可能抵消收益的瓶颈—，并且允许在数据库中运行整个流。[有关详细信息，请参阅第 6 章中的 SQL 优化中的 IBM SPSS Modeler Server 15 管理和性能指南](#)。
- 如果原始数据没有存储于数据库，则仍可使用数据库建模。此种情况下，将在 SPSS Modeler 中进行数据准备，所准备的数据集将自动上载到此数据库并进行模型构建。

模型评分

采用数据库内数据挖掘在 IBM® SPSS® Modeler 中生成的模型与常规的 SPSS Modeler 模型不同。虽然这些模型作为“块”，显示在模型管理器中，但它们实际是保存在远程数据挖掘或数据库服务器中的远程模型。您在 SPSS Modeler 中所看到的其实是对这些远程模型的引用。换言之，您所看到的 SPSS Modeler 模型是“空”模型，其中包含数据库服务器主机名、数据库名和模型名等信息。当对采用数据库自有算法创建的模型进行浏览和评分时，您应当清楚这个明显差别。

图片 2-4
生成的 Microsoft 决策树模型 “nugget”



创建模型后，您可以将其添加到流并像其他所有在 SPSS Modeler 中生成的模型一样进行评分。所有评分均在数据库中进行，即便没有上游操作。（如果可以提高性能，上游操作仍可能会被推回数据库，但评分时并不一定要求这样。）大多数情况下，您还可使用数据库提供商所给予的标准浏览器来浏览生成的模型。

为了同时进行浏览和评分，需要连接运行 Oracle Data Miner、IBM DB2 InfoSphere Warehouse 或 Microsoft Analysis Services 的服务器。

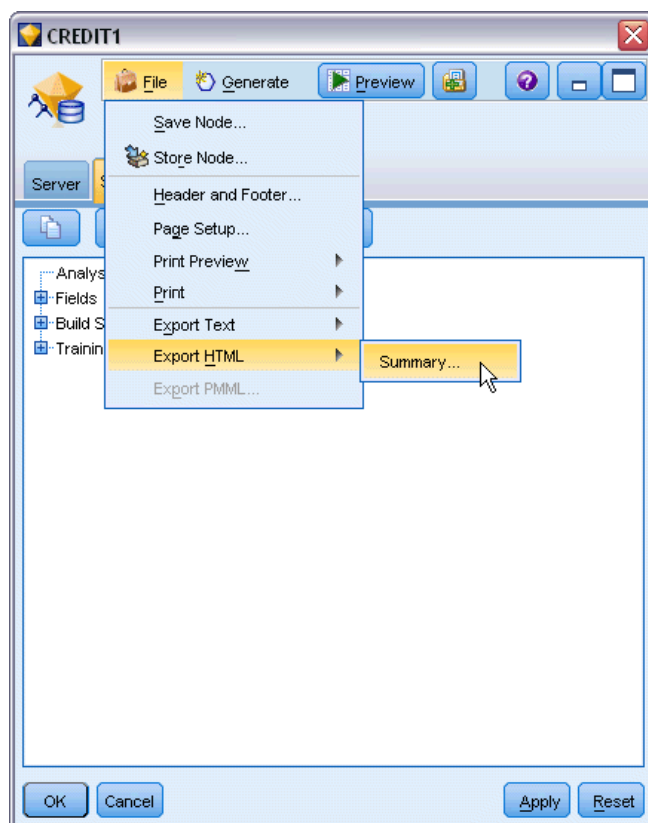
查看结果和指定设置

要查看结果和指定评分设置，请双击流工作区上的模型。您还可以选择右键单击此模型，然后选择浏览或编辑。具体设置取决于模型的类型。

输出和保存数据库模型

借助“文件”菜单中的选项，可以从模型浏览器中导出数据库模型和摘要，就像导出在 IBM® SPSS® Modeler 中创建的模型一样。

图片 2-5
以 HTML 格式导出 Microsoft 决策树模型摘要



- ▶ 在模型浏览器的“文件”菜单中，选择以下某项：
 - 导出文本将模型摘要导出到文本文件
 - 导出 HTML将模型摘要导出到 HTML 文件
 - 导出 PMML（仅支持 IBM DB2 IM 模型）以 PMML（预测模型标记语言）格式导出模型，导出的文件可在其它 PMML 兼容软件中使用。有关详细信息，请参阅第 10 章中的导入和导出 PMML 模型中的 IBM SPSS Modeler 15 用户指南。

注意：还可通过从“文件”菜单中选择保存节点来保存某个生成的模型。有关详细信息，请参阅第 3 章中的浏览模型块中的 IBM SPSS Modeler 15 建模节点。

模型一致性

对于生成的每个数据库模型，IBM® SPSS® Modeler 会存储一个模型结构说明，同时会以数据库中的模型名称来保存一个模型引用。生成模型的“服务器”选项卡将显示为此模型所生成的唯一关键字，此关键字与数据库中的实际模型相匹配。

图片 2-6
生成的模型关键字和检查选项



The screenshot shows a software interface with a light blue background. At the top, there are three tabs: '服务器' (Server), '汇总' (Summary), and '注解' (Annotation). Below the tabs, there are four input fields and two buttons. The first field is labeled '分析服务器主机:' (Analysis server host:) and contains the text 'gb1w2k3dbhost1'. The second field is labeled '分析服务器数据库:' (Analysis server database:) and contains 'dan', with a blue button containing three dots to its right. The third field is labeled 'SQL Server 连接:' (SQL Server connection:) and contains 'SQL Server Stndrd', also with a blue button containing three dots to its right. The fourth field is labeled '模型 GUID:' (Model GUID:) and contains the GUID '{2D5DB0DF-5888-43EC-BBC2-1218F057F7AD}'. At the bottom of the form, there are two buttons: a blue button with a checkmark icon and the text '检查' (Check), and a blue button with a magnifying glass icon and the text '视图' (View).

SPSS Modeler 采用随机生成关键字来检查模型是否仍然一致。此关键字会在创建模型时存储于模型说明中。最好在运行部署流之前检查关键字匹配情况。

- ▶ 单击检查按钮，比较模型说明与 SPSS Modeler 所存储的随机关键字，以此检查数据库中模型的一致性。如果未找到数据库模型或关键字不匹配，则系统将报错。

查看和导出生成的 SQL

可以在执行前预览所生成的 SQL 节点，这会对调试有所帮助。有关详细信息，请参阅第 6 章中的预览生成的 SQL 中的 IBM SPSS Modeler Server 15 管理和性能指南。

使用 Microsoft Analysis Services 进行数据库建模

IBM SPSS Modeler 与 Microsoft Analysis Services

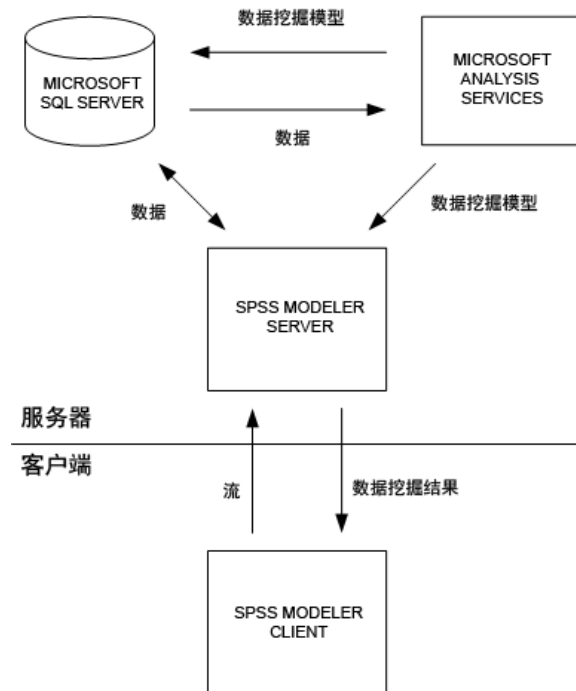
IBM® SPSS® Modeler支持与 Microsoft SQL Server Analysis Services 的集成。此功能作为 SPSS Modeler 中的建模节点实现，并且可以从“数据库建模”选项板上使用此功能。如果该选项板不可见，可通过启用 MS Analysis Services 集成（位于“辅助应用程序”对话框的“Microsoft”选项卡上）将其激活。 [有关详细信息，请参阅第 15 页 码启用与 Analysis Services 的集成。](#)

SPSS Modeler 支持集成以下 Analysis Services 算法：

- 决策树
- 聚类
- 关联规则
- Naive Bayes
- 线性回归
- 神经网络
- Logistic 回归
- 时间序列
- 序列聚类

下图说明了从客户端到服务器的数据流，其中数据库内挖掘由 IBM® SPSS® Modeler Server 管理。模型构建使用 Analysis Services 进行。得到的模型由 Analysis Services 存储。对此模型的引用在 SPSS Modeler 流中维护。然后，该模型从 Analysis Services 下载到 Microsoft SQL Server 或 SPSS Modeler 中进行评分。

图片 3-1
模型构建过程中，IBM SPSS Modeler、Microsoft SQL Server 与 Microsoft Analysis Services 之间的数据流



注意：SPSS Modeler Server 虽然可以使用，但不是必需的。IBM® SPSS® Modeler 客户端自身就能够处理数据库内挖掘计算。

与 Microsoft Analysis Services 集成的要求

以下是在 IBM® SPSS® Modeler 中使用 Analysis Services 算法执行数据库内建模的必备条件。您可能需要咨询数据库管理员以确保满足这些条件。

- 在 Windows 上安装 IBM® SPSS® Modeler Server 后（分布式模式）运行 IBM® SPSS® Modeler。与 Analysis Services 的集成不支持 UNIX 平台。
重要事项：SPSS Modeler 用户必须使用从位于其他 SPSS Modeler Server 要求下的下列 URL 获取的 Microsoft SQL Native Client 驱动程序来配置 ODBC 连接。此处不推荐使用 IBM® SPSS® Data Access Pack（一般推荐用于 SPSS Modeler 中的其他用途）中所提供的驱动程序。驱动程序应配置为在启用与 Windows 验证集成的条件下使用 SQL Server，因为 SPSS Modeler 不支持 SQL Server 验证。如果您有关于创建或设置 ODBC 数据源权限的问题，请联系您的数据库管理员。
- 必须安装 SQL Server 2005 或 2008，但不一定要与 SPSS Modeler 安装在同一台主机上。SPSS Modeler 用户必须有足够的权限来读、写数据以及删除和创建表与视图。
注意：推荐使用 SQL Server Enterprise Edition。Enterprise Edition 借助高级参数来调节算法结果，并以此提供了更多的灵活性。Standard Edition 版本提供了相同的参数但不允许用户编辑某些高级参数。
- Microsoft SQL Server Analysis Services 必须安装在与 SQL Server 相同的主机上。

其他 IBM SPSS Modeler Server 要求

要在 SPSS Modeler Server 中使用 Analysis Services 算法，则必须在 SPSS Modeler Server 主机上安装以下组件。

注意：如果 SQL Server 安装在与 SPSS Modeler Server 相同的主机上，则这些组件已经可用。

- Microsoft .NET Framework V 2.0 Redistributable Package (x86)
- Microsoft Core XML Services (MSXML) 6.0
- Microsoft SQL Server 2008 Analysis Services 10.0 OLE DB Provider（确保选择适合您操作系统的正确版本）
- Microsoft SQL Server 2008 Native Client（确保选择适合您操作系统的正确版本）

要下载这些组件，转到 www.microsoft.com/downloads，搜索 **.NET Framework** 或（对于所有其他组件）**SQL Server Feature Pack**，并选择您的 SQL Server 版本的最新软件包。

这些组件可能需要首先安装其他软件包，此类软件包也可从 Microsoft 下载站点获得。

其他 IBM SPSS Modeler 要求

要在 SPSS Modeler 中使用 Analysis Services 算法，必须安装以上组件，同时在客户端添加以下组件：

- Microsoft SQL Server 2008 Datamining Viewer Controls（确保选择了适合您操作系统的正确版本）- 这还需要：
- Microsoft ADOMD.NET

要下载这些组件，转到 www.microsoft.com/downloads，搜索 **SQL Server Feature Pack**，并选择您的 SQL Server 版本的最新软件包。

注意：数据库建模和 SQL 优化需要在 SPSS Modeler 计算机上启用 SPSS Modeler Server 连接。通过启用此设置，您可以访问数据库算法，直接从 SPSS Modeler 回送 SQL 以及访问 SPSS Modeler Server。要验证当前许可证的状态，请从 SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项服务器启用。

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 15 用户指南](#)。

启用与 Analysis Services 的集成

要启用 Analysis Services 的 IBM® SPSS® Modeler 集成，需要配置 SQL Server 和 Analysis Services，创建 ODBC 源，在 SPSS Modeler 的“辅助应用程序”对话框中启用集成，并启用 SQL 生成和优化。

注意：Microsoft SQL Server 和 Microsoft Analysis Services 必须可用。有关详细信息，请参阅第 14 页码与 [Microsoft Analysis Services 集成的要求](#)。

配置 SQL Server

配置 SQL Server 以便可以在数据库内进行评分。

- ▶ 在 SQL Server 主机上创建以下注册表键：

HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP

- ▶ 为该键添加如下 **DWORD** 键值：

AllowInProcess 1

- ▶ 完成上述更改后，重新启动 SQL Server。

配置 Analysis Services

必须首先在分析服务器的“属性”对话框中手动配置两项设置后，SPSS Modeler 才能与 Analysis Services 进行通信。

- ▶ 通过 MS SQL Server Management Studio 登录到分析服务器。
- ▶ 要访问“属性”对话框，请右键单击服务器名称，然后选择属性。
- ▶ 选中显示高级（所有）属性复选框。
- ▶ 更改以下属性：
 - 将 DataMining\AllowAdHocOpenRowsetQueries 的值更改为 True（默认值为 False）。
 - 将 DataMining\AllowProvidersInOpenRowset 的值更改为 [all]（无默认值）。

为 SQL Server 创建 ODBC DSN

要读取或写入到数据库中，您必须为相关数据库安装并配置 ODBC 数据源，并根据需要配置读取或写入权限。Microsoft SQL Native Client ODBC 驱动程序是必需的，SQL Server 会自动安装该驱动程序。此处不推荐使用 IBM® SPSS® Data Access Pack（一般推荐用于 SPSS Modeler 中的其他用途）中所提供的驱动程序。如果 SPSS Modeler 和 SQL Server 驻留在不同的主机上，可以下载 Microsoft SQL Native Client ODBC 驱动程序。[有关详细信息，请参阅第 14 页码与 Microsoft Analysis Services 集成的要求。](#)

如果您有关于创建或设置 ODBC 数据源权限的问题，请联系您的数据库管理员。

- ▶ 使用 Microsoft SQL Native Client ODBC 驱动程序，创建指向数据挖掘过程中所使用的 SQL Server 数据库的 ODBC DSN。余下的驱动程序设置应使用默认设置。
- ▶ 对于此 DSN，请确保选中了使用集成的 Windows 身份验证。
 - 如果 IBM® SPSS® Modeler 和 IBM® SPSS® Modeler Server 运行在不同的主机上，请在每个主机上创建相同的 ODBC DSN。确保每台主机上使用的 DSN 名称相同。

在 IBM SPSS Modeler 中启用 Analysis Services 集成

要使 SPSS Modeler 能够使用 Analysis Services，首先必须在“辅助应用程序”对话框中输入服务器规范。

- ▶ 从 SPSS Modeler 菜单中选择：
工具 > 选项 > 辅助应用程序
- ▶ 单击 Microsoft 选项卡。
 - **启用 Microsoft Analysis Services 集成。** 启用 SPSS Modeler 窗口底部的“数据库建模”选项板（如尚未显示）并添加 Analysis Services 算法的建模节点。

图片 3-2
“数据库建模”选项卡



- **分析服务器主机。** 指定运行 Analysis Services 的计算机的名称。
- **分析服务器数据库。** 通过单击省略号 (...) 按钮打开一个子对话框，从其中的可用数据库中选择所需的数据库。列表中的数据库均是可用于指定分析服务器的数据库。由于 Microsoft Analysis Services 在指定数据库中存储数据挖掘模型，因此应选择在其中存储了由 SPSS Modeler 构建的 Microsoft 模型的相应数据库。
- **SQL Server 连接。** 指定 SQL Server 数据库使用的 DSN 信息以存储传递到分析服务器中的数据。选择将用来提供用于构建 Analysis Services 数据挖掘模型的数据的 ODBC 数据源。如果您要根据平面文件或 ODBC 数据源中的数据构建 Analysis Services 模型，则此类数据将自动上载到在此 ODBC 数据源所指向的 SQL Server 数据库中创建的临时表。
- **覆盖数据挖掘模型时发送警告。** 选中此项可确保在没有发出警告的情况下 SPSS Modeler 不会覆盖数据库中存储的模型。

注意：可以在各个 Analysis Services 节点内覆盖在“辅助应用程序”对话框中所做的设置。

启用 SQL 生成和优化

- ▶ 从 SPSS Modeler 菜单中选择：
工具 > 流属性 > 选项

图片 3-3
优化设置



- ▶ 在导航窗格中单击优化选项。
- ▶ 确认是否已启用生成 SQL 选项。要使数据库建模正常发挥作用，此设置是必需的。
- ▶ 选中优化 SQL 生成和优化其他执行（非严格必需但强烈推荐使用，以使性能更优）。

有关详细信息，请参阅第 5 章中的设置流的优化选项中的 IBM SPSS Modeler 15 用户指南。

使用 Analysis Services 构建模型

Analysis Services 模型构建要求训练数据集位于 SQL Server 数据库内的表或视图中。如果数据不在 SQL Server 中，或者需要在 IBM® SPSS® Modeler 中作为不能在 SQL Server 中执行的数据准备的一部分来处理，则此类数据将在模型构建前自动上载到 SQL Server 中的临时表。

管理 Analysis Services 模型

通过 IBM® SPSS® Modeler 构建 Analysis Services 模型会在 SPSS Modeler 中创建一个模型，然后在 SQL Server 数据库中创建一个模型或替换其中一个模型。SPSS Modeler 模型会引用存储在数据库服务器中的数据库模型的内容。SPSS Modeler 可通过将完全相同的生成模型关键字字符串存储在 SPSS Modeler 模型和 SQL Server 模型中执行一致性检查。



MS 决策树 建模节点可同时用于分类属性和连续属性的预测建模。对于分类属性，该节点根据数据集中输入列之间的关系进行预测。例如，某方案要预测哪些顾客最有可能购买自行车，如果在年轻顾客中购买自行车的比例是十分之九，而在年纪较大的顾客中购买比例仅为十分之二，则该节点可推断出年龄是有关自行车购买行为的良好预测变量。决策树可根据此特定结果的趋势做出预测。对于连续属性，算法将使用线性回归确定决策树分割的位置。如果有一个以上的列被设置为可预测的列，或如果输入数据包含一个被设置为可预测的嵌套表，则该节点可为每个可预测的列构建单独的决策树。



MS Clustering 建模节点采用迭代技术将某个数据集中的个案分组归入具有类似特征的聚类。此类分组可用于探索数据、识别数据异常和创建预测。聚类模型可以识别您无法通过表面观测进行逻辑推导而获得的数据集中的关系。例如，您可能从逻辑上看出乘自行车上下班的人通常其住地距离工作地不远。但通过算法可以找出乘自行车上班者的其它并不明显的特征。聚类节点区别于其它未指定目标字段的数据挖掘节点。聚类节点将通过数据中的关系和节点所识别聚类的关系对模型进行严格训练。



MS 关联规则 建模节点对于推荐引擎十分有用。推荐引擎将根据客户已购买的项目或其已表示出兴趣的项目来向客户推荐产品。将根据同时包含个别个案 ID 和此类个案所含项目 ID 的数据集来创建关联模型。个案中的一组项目称为**项目集**。关联模型由一系列项目集和个案中的项目分组规则构成。算法所发现的规则可用于根据客户已决定购买的项目来预测客户未来可能购买哪些产品。



MS Naive Bayes 建模节点可计算目标字段和预测变量字段之间的条件概率，并假定这些列是相互独立的。该模型被称为 naïve 是因为它将所有被提议的预测变量视为相互独立的。此方法比其他 Microsoft 算法的计算量小，因此对于在建模初期迅速发现关系非常有用。可使用此节点对数据进行初始探索，然后将探索结果应用于由其他节点创建的附加模型，这些节点用于计算的时间会更长，但会得出更准确的结果。



MS 线性回归 建模节点是决策树节点的变异，其中 `MINIMUM_LEAF_CASES` 参数被设置为大于或等于节点用来训练挖掘模型的数据集中的案例总数。如果按上述方法设置参数，则该节点将永远不会创建分割，因此可执行线性回归。



MS 神经网络 建模节点类似于 MS 决策树节点，即，当给定可预测属性的每个状态时，MS 神经网络节点会为输入属性的每个可能的状态计算概率。之后，可以根据已输入的属性，使用这些概率预测属性的结果。



MS Logistic 回归 建模节点是 MS 神经网络节点的变异，其中 `HIDDEN_NODE_RATIO` 参数设置为 0。此设置可创建不包含隐藏层的神经网络模型，因此相当于 logistic 回归。



The **MS 时间序列**建模节点提供的回归算法对连续值（如产品销售）在时间上的预测进行了优化。而其他 Microsoft 算法，如决策树，需要额外的新信息列作为输入才能预测趋势，时间序列模型则不是这样。时间序列模型可以只基于用于创建模型的原始数据集预测趋势。在进行预测时，您还可以添加新数据到模型，并在趋势分析中自动包含新数据。 [有关详细信息，请参阅第 28 页码MS 时间序列节点。](#)



MS 序列聚类建模节点标识数据中的顺序序列，并将此分析的结果与聚类技术结合以基于序列和其他属性生成聚类。 [有关详细信息，请参阅第 31 页码 MS 序列聚类节点。](#)

您可以从 SPSS Modeler 窗口底部的“数据库建模”选项板中访问每个节点。

对所有算法节点通用的设置

以下设置通用于所有 Analysis Services 算法。

服务器选项

在“服务器”选项卡上，可以配置分析服务器主机、数据库和 SQL Server 数据源。此处指定的选项将覆盖在“辅助应用程序”对话框的“Microsoft”选项卡上指定的选项。[有关详细信息，请参阅第 15 页码启用与 Analysis Services 的集成。](#)

图片 3-4
服务器选项

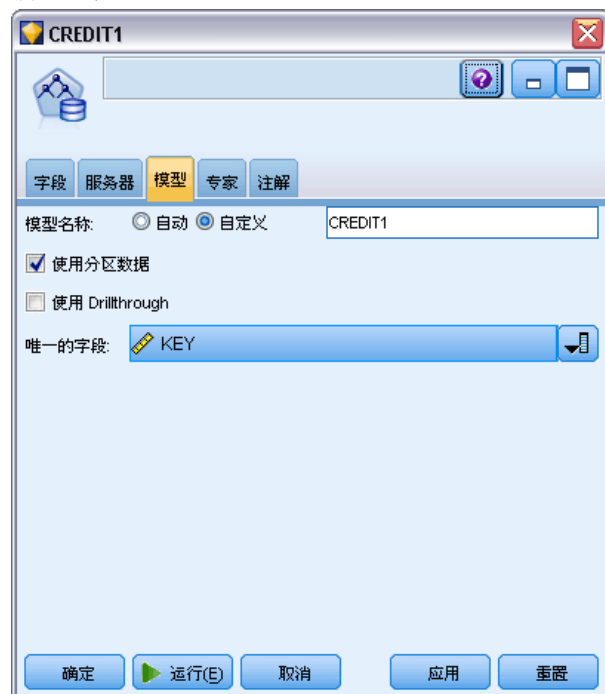


注意：当对 Analysis Services 模型进行评分时，还可以使用此选项卡的变体。有关详细信息，请参阅第 34 页码 Analysis Services 模型块服务器选项卡。

模型选项

要构建最基本的模型，需要在处理前在“模型”选项卡上指定选项。评分方法和其他高级选项可在“专家”选项卡上找到。

图片 3-5
模型选项



提供以下基本建模选项：

模型名称。指定分配给执行节点时所创建模型的名称。

- **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。**允许您为所创建模型指定自定义名称。

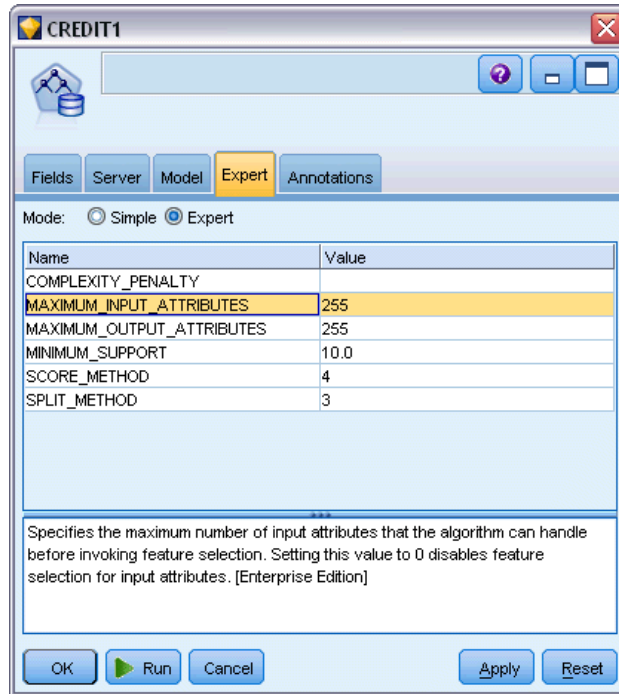
使用分区数据。将数据分割成若干独立的子集或样本，以根据当前分区字段进行训练、检验和验证。使用一个样本创建模型并用另一个样本对其进行检验，这将揭示出该模型在多大程度上可以推广到与当前数据类似的更大数据集上。如果未在流中指定分区字段，将忽略此选项。有关详细信息，请参阅第 4 章中的分区节点中的 IBM SPSS Modeler 15 源、过程和输出节点。

有 Drillthrough。如果显示此选项，则您可以查询模型以了解模型中所包含个案的详细信息。

唯一字段。从下拉列表中，选择唯一标识每种情况的字段。通常，这个字段为 ID 字段，例如 Customer ID。

MS 决策树专家选项

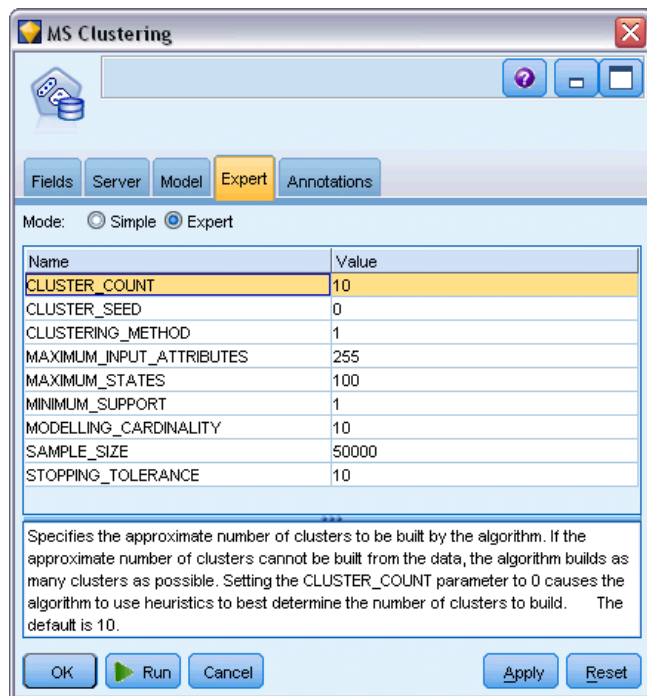
图片 3-6
MS 决策树专家选项



“专家”选项卡上的可用选项会依据选定流的结构而变化。有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 聚类专家选项

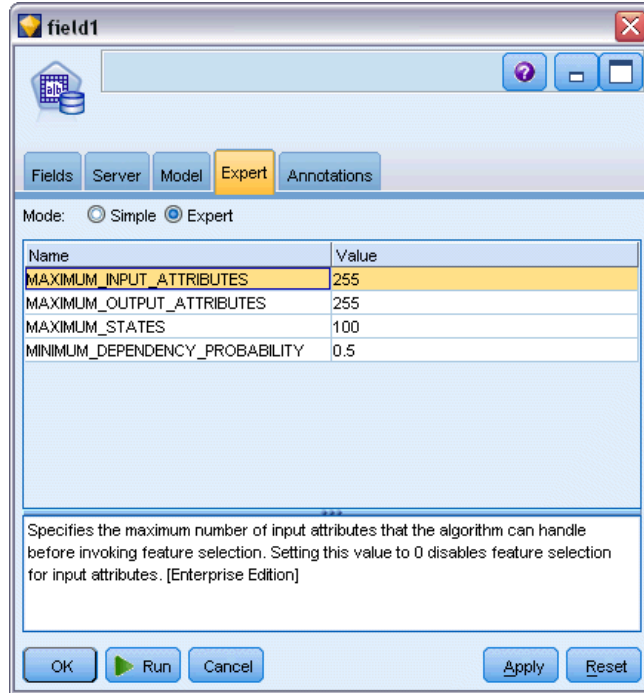
图片 3-7
MS 聚类专家选项



“专家”选项卡上的可用选项会依据选定流的结构而变化。有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS Naive Bayes 专家选项

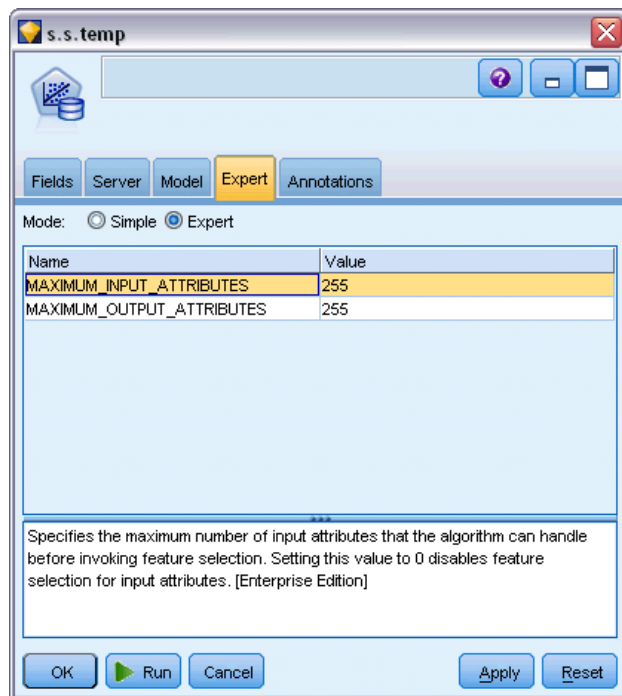
图片 3-8
MS Naive Bayes 专家选项



“专家”选项卡上的可用选项会依据选定流的结构而变化。有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 线性回归专家选项

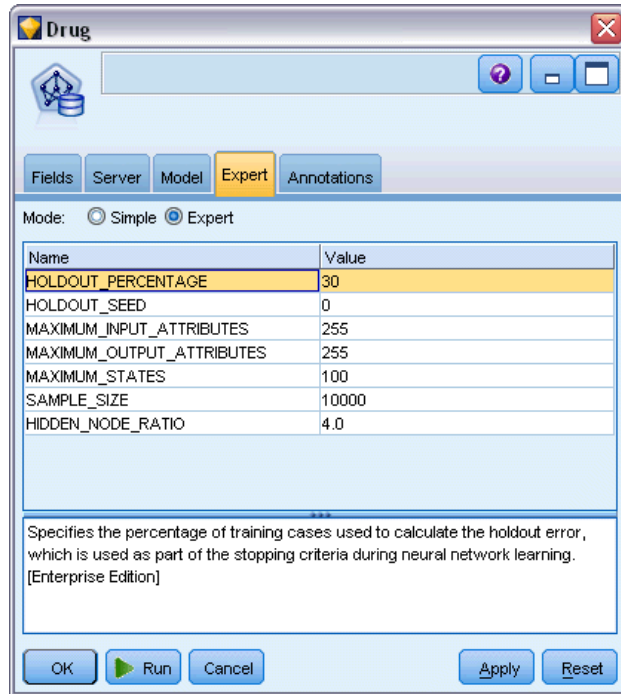
图片 3-9
MS 线性回归专家选项



“专家”选项卡上的可用选项会依据选定流的结构而变化。有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 神经网络专家选项

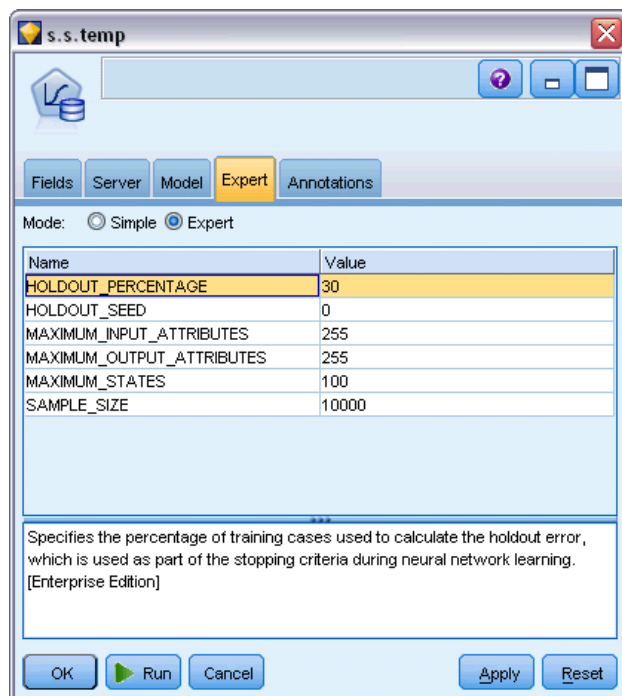
图片 3-10
MS 神经网络专家选项



“专家”选项卡上的可用选项会依据选定流的结构而变化。有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS Logistic 回归专家选项

图片 3-11
MS Logistic 回归专家选项



“专家”选项卡上的可用选项会依据选定流的结构而变化。有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 关联规则节点

MS 关联规则建模节点对于推荐引擎十分有用。推荐引擎将根据客户已购买的项目或其已表示出兴趣的项目来向客户推荐产品。将根据同时包含个别个案 ID 和此类个案所含项目 ID 的数据集来创建关联模型。个案中的一组项目称为**项目集**。

关联模型由一系列项目集和个案中的项目分组规则构成。算法所发现的规则可用于根据客户已决定购买的项目来预测客户未来可能购买哪些产品。

对于表格格式数据，该算法创建代表每个生成推荐 (\$M-field) 的概率 (\$MP-field) 的得分。对于交易格式数据，为支持 (\$MS-field)、每个生成推荐 (\$M-field) 的概率 (\$MP-field) 和调整概率 (\$MAP-field) 创建得分。有关详细信息，请参阅第 12 章中的表格格式数据与事务处理格式数据中的 IBM SPSS Modeler 15 建模节点。

要求

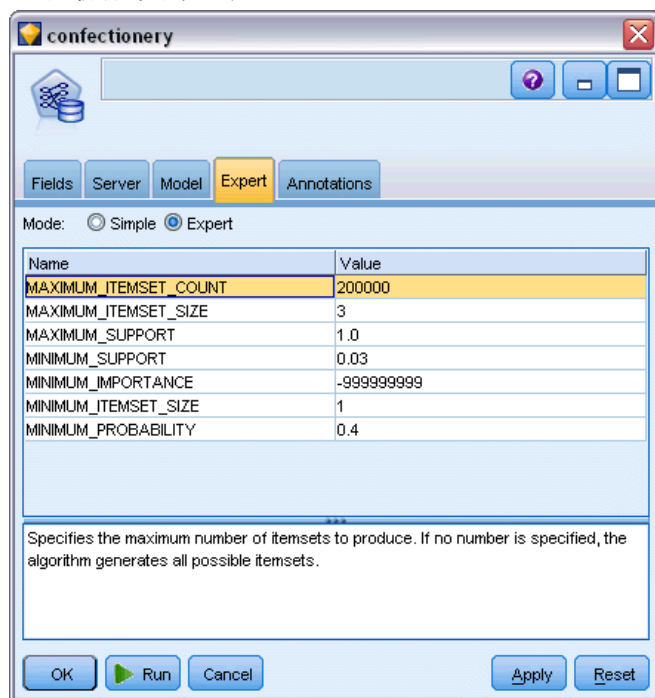
交易关联模型的要求如下：

- **唯一字段**。关联规则模型需要唯一标识记录的关键字。

- **ID 字段。**当构建具有交易格式数据的 MS 关联规则模型时，标识每个交易的 ID 字段为必填。ID 字段可设为与唯一字段相同。
- **至少一个输入字段。**关联规则算法需要至少一个输入字段。
- **目标字段。**当构建具有交易数据的 MS 关联模型时，目标字段必须为交易字段，例如用户购买的产品。

MS 关联规则专家选项

图片 3-12
MS 关联规则专家选项



“专家”选项卡上的可用选项会依据选定流的结构而变化。有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

MS 时间序列节点

MS 时间序列建模节点支持两种类型的预测：

- 未来
- 历史

未来预测评估在历史数据结束之外若干指定时间段的目标字段值，并总是得到执行。**历史预测**是在历史数据中具有实际值的若干指定时间段的评估目标字段值。您可使用历史预测，通过将实际历史值与预测值进行比较来评估模型质量。预测起点的值决定是否执行历史预测。

与 IBM® SPSS® Modeler 时间序列节点不同，MS 时间序列节点不需要提前的时间区间节点。另外一个不同是，默认情况下只对预测行生成得分，不对时间序列数据中的所有历史行生成得分。

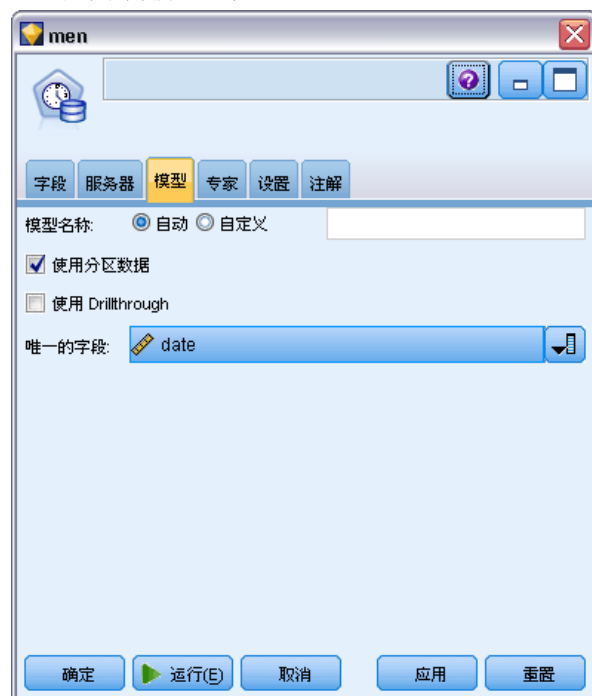
要求

MS 时间序列模型的要求如下：

- **单关键字时间字段。** 每个模型必须包含一个数值或日期字段，该字段将用作个案序列并定义模型使用的时间块。关键字时间字段的数据类型可为日期时间数据类型或数值数据类型。但该字段必须包含连续值，且值必须对每个序列唯一。
- **单目标字段。** 您在每个模型中只能指定一个目标字段。目标字段的数据类型必须为连续值。例如，您可预测数值属性（如收入、销售或温度）如何随时间变化。但您无法将包含某些分类值（如购买状态或教育水平）的字段用作目标字段。
- **至少一个输入字段。** MS 时间序列算法需要至少一个输入字段。输入字段的数据类型必须具有连续值。构建模型时忽略非连续输入字段。
- **数据集必须排序。** 输入数据集必须排序（在关键字时间字段上），否则模型构建会因错误而中断。

MS 时间序列模型选项

图片 3-13
MS 时间序列模型选项



模型名称。 指定分配给执行节点时所创建模型的名称。

- **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。** 允许您为所创建模型指定自定义名称。

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 15 源、过程和输出节点。

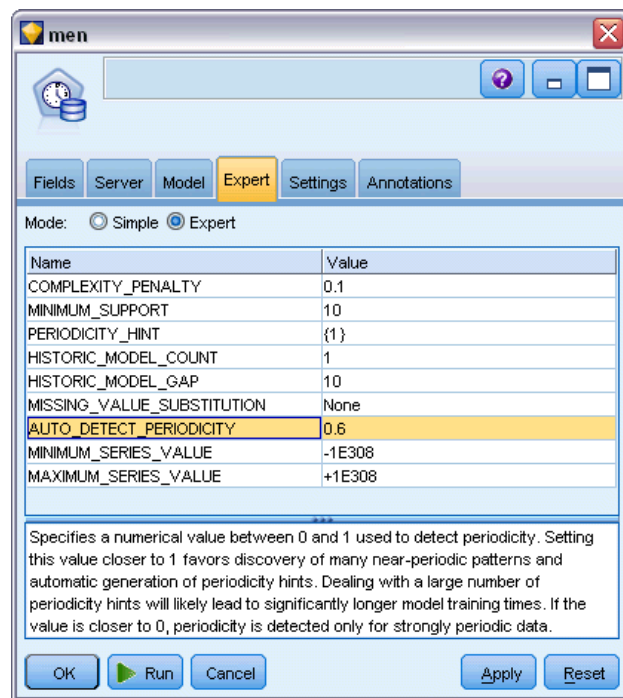
有 Drillthrough。 如果显示此选项，则您可以查询模型以了解模型中所包含个案的详细信息。

唯一字段。 从下拉列表选择关键字时间字段，该字段用于构建时间序列模型。

MS 时间序列专家选项

图片 3-14

MS 时间序列专家选项



“专家”选项卡上的可用选项会依据选定流的结构而变化。有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

如果您正在进行历史预测，则在得分结果中包含的历史步骤的数量由 $(\text{HISTORIC_MODEL_COUNT} * \text{HISTORIC_MODEL_GAP})$ 的值决定。默认情况下，此限制为 10，意味着将只进行 10 次历史预测。此时，例如当您在模型块的“设置”选项卡上为历史预测输入小于 -10 的值时，会发生错误（参见 MS 时间序列模型块设置选项卡第 39 页码）。如果你想看到更多的历史预测，可增加 HISTORIC_MODEL_COUNT 或 HISTORIC_MODEL_GAP 的值，但这会增加模型的构建时间。

MS 时间序列设置选项

图片 3-15
MS 时间序列设置选项



开始估计。 指定预测开始的时间段。

- **开始于：新预测。** 未来预测开始的时间段，表示为最后一个历史数据时间段的偏移值。例如，如果您的历史数据结束于 12/99，且您想在 01/00 开始预测，则应使用值 1；但如果您想在 03/00 开始预测，则应使用值 3。
- **开始于：历史预测。** 历史预测开始的时间段，表示为最后一个历史数据时间段的负偏移值。例如，如果您的历史数据结束于 12/99，且您想对数据的最后五个时间段进行历史预测，则使用值 -5。

结束估计。 指定预测停止的时间段。

- **预测的结束步骤。** 预测停止的时间段，表示为最后一个历史数据时间段的偏移值。例如，如果您的历史数据结束于 12/99，且您想在 6/00 停止预测，则应在这里使用值 6。对于未来预测，值必须总是大于或等于开始于值。

MS 序列聚类节点

MS 序列聚类节点使用一种序列分析算法，该算法探索包含可由以下路径链接的事件的数据或序列。这种类型的示例可以是用户在导航或浏览网站时创建的单击路径，或客户在在线零售商添加商品到购物车的顺序。算法按照分组或聚类找出最常见的序列和等同的序列。

要求

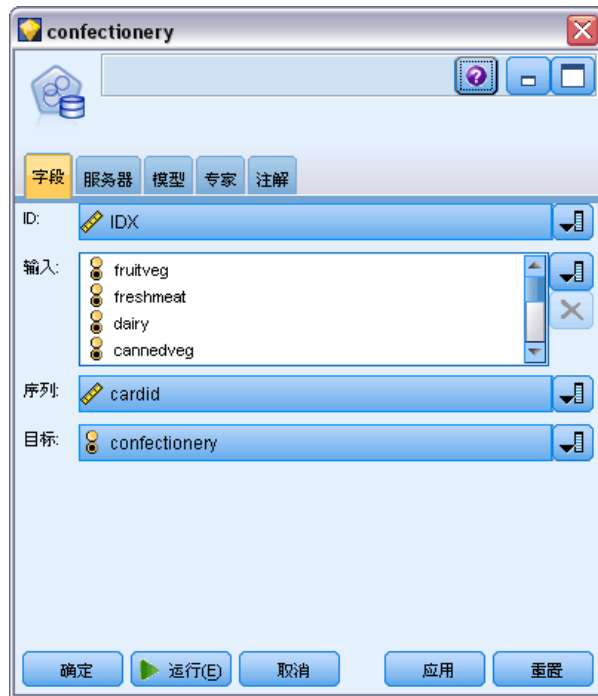
Microsoft 序列聚类模型的要求：

- **ID 字段。**Microsoft 序列聚类算法要求序列信息以交易格式存储（参见 [表格格式数据与事务处理格式数据第 页码](#)）。为此，标识每个交易的 ID 字段为必填。
- **至少一个输入字段。**算法需要至少一个输入字段。
- **序列字段。**算法还需要序列标识符字段，该字段必须具有“连续”测量水平。例如，您可使用网页标识符、整数或文本字符串，只要字段标识序列中的事件。一个序列只允许一个序列标识符，每个模型只允许一种序列类型。序列字段必须与 ID 字段和唯一字段不同。
- **目标字段。**构建序列聚类模型时，目标字段为必填。
- **唯一字段。**序列聚类模型需要唯一标识记录的关键字字段。您可将唯一字段设为与 ID 字段相同。

MS 序列聚类字段选项

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 3-16
指定用于 MS 序列聚类的字段



在您构建序列聚类模型之前，需要指定哪些字段将用作目标和输入。注意，对于 MS 序列聚类节点，您无法使用上游类型节点的字段信息；您必须在这里指定字段设置。

ID。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）。

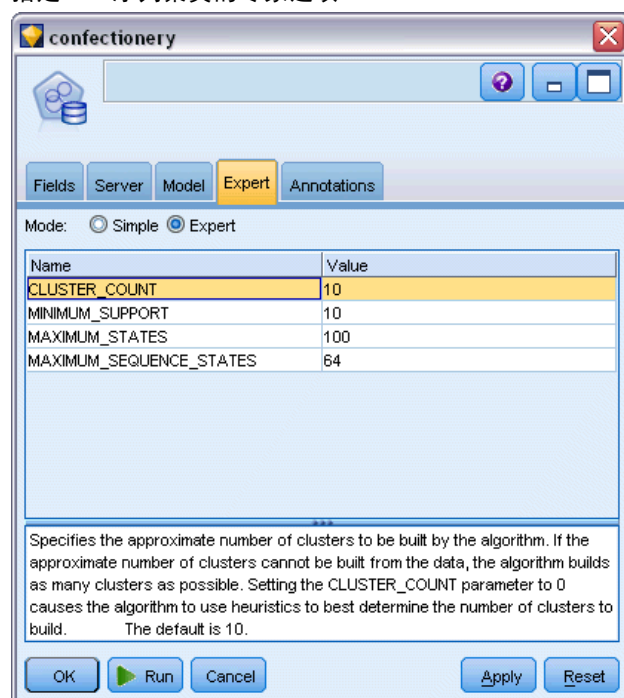
输入。选择模型的输入字段或字段。这些字段包含与序列建模有关的事件。

序列。从列表选择一个字段用作序列标识符字段。例如，您可使用网页标识符、整数或文本字符串，只要字段标识序列中的事件。一个序列只允许一个序列标识符，每个模型只允许一种序列类型。序列字段必须与 ID 字段（在此选项卡上指定）和唯一字段（在“模型”选项卡上指定）不同。

目标。选择一个字段用作目标字段，即您将基于序列数据尝试预测其值的字段。

MS 序列聚类专家选项

图片 3-17
指定 MS 序列聚类的专家选项



“专家”选项卡上的可用选项会依据选定流的结构而变化。有关选定的 Analysis Services 模型节点的专家选项的详细信息，请参阅用户界面现场帮助。

对 Analysis Services 模型评分

模型评分发生在 SQL Server 内并由 Analysis Services 执行。如果数据源自 IBM® SPSS® Modeler 内或需要在 SPSS Modeler 内准备，则可能需要将数据集上载到临时表。您使用数据库内挖掘从 SPSS Modeler 创建的模型实际是保存在远程数据挖掘或

数据库服务器上的远程模型。当对采用 Microsoft Analysis Services 算法创建的模型进行浏览和评分时，这是一个需要了解的重要区别。

在 SPSS Modeler 中，通常只提供一次预测以及关联的概率或置信度。

欲了解模型评分示例，请参阅[Analysis Services 数据挖掘示例第 40 页码](#)。

对所有 Analysis Services 模型通用的设置

以下设置通用于所有 Analysis Services 模型。

Analysis Services 模型块服务器选项卡

“服务器”选项卡用于为数据库内挖掘指定连接。此选项卡还提供唯一的模型关键字。此关键字是在构建模型时随机生成的，并存储于 IBM® SPSS® Modeler 的模型中及 Analysis Services 数据库中存储的模型对象的说明中。

图片 3-18
MS 决策树模型块的服务器选项

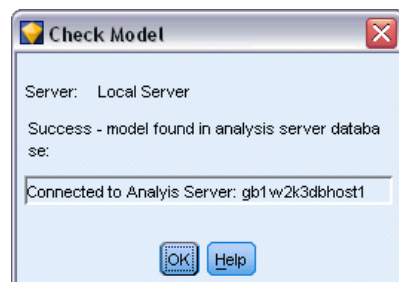


在“服务器”选项卡上，可以为评分操作配置分析服务器主机和数据库及 SQL Server 数据源。在 IBM® SPSS® Modeler 中，此处指定的选项将覆盖那些在“辅助应用程序”或“构建模型”对话框中指定的选项。[有关详细信息，请参阅第 15 页码启用与 Analysis Services 的集成。](#)

模型 GUID。 在此处显示模型关键字。此关键字是在构建模型时随机生成的，并存储于 SPSS Modeler 的模型中及 Analysis Services 数据库中存储的模型对象的说明中。

检查。 单击此按钮对照存储于 Analysis Services 数据库中的模型的关键字检查此模型关键字。此操作有助于验证模型是否仍存在于分析服务器中，并表示模型的结构未更改。

图片 3-19
检查模型关键字的结果

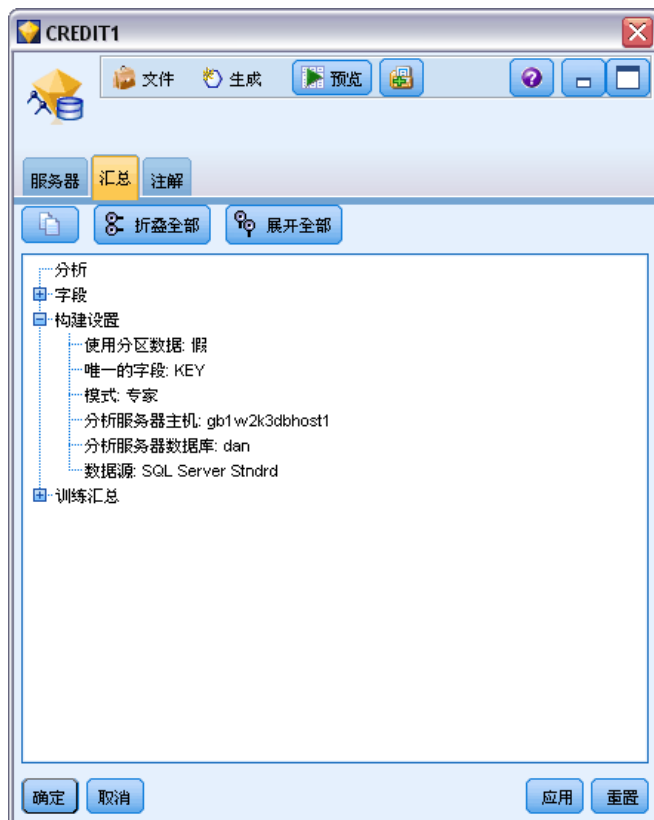


注意：仅在添加到流工作区的准备评分的模型中，才可使用“检查”按钮。如果检查失败，可调查此模型是否已被删除或被服务器上的其他模型替换。

视图。 单击以打开决策树模型的图形视图。决策树查看器由 SPSS Modeler 中的其他决策树算法所共享，且功能相同。[有关详细信息，请参阅第 6 章中的决策树模型块中的 IBM SPSS Modeler 15 建模节点。](#)

Analysis Services 模型块汇总选项卡

图片 3-20
MS 决策树模型块的汇总选项



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。查看完成后要隐藏结果时，请使用展开控件来折叠想要隐藏的具体结果，或者单击全部折叠按钮来折叠所有结果。

分析。 显示指定模型的相关信息。如果已执行附加到该模型块的分析节点，则还会在此部分显示通过分析获得的信息。[有关详细信息，请参阅第 6 章中的分析节点中的 IBM SPSS Modeler 15 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。

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。

MS 时间序列模型块

MS 时间序列模型只生成预测时间段内的得分，不生成历史数据的得分。

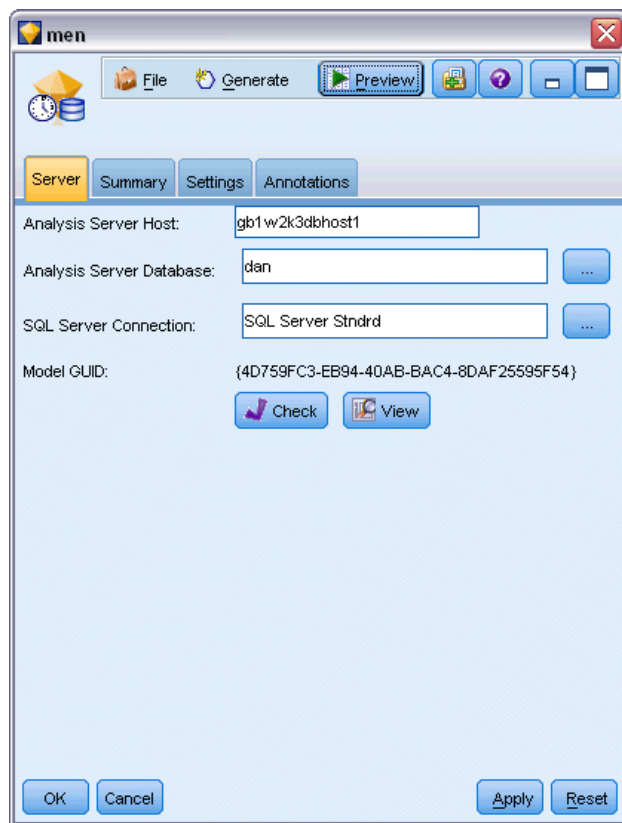
以下字段被添加到模型：

字段名称	描述
\$M-field	field 的预测值
\$Var-field	field 的计算方差
\$Stdev-field	field 的标准差

MS 时间序列模型块服务器选项卡

“服务器”选项卡用于为数据库内挖掘指定连接。此选项卡还提供唯一的模型关键字。此关键字是在构建模型时随机生成的，并存储于 IBM® SPSS® Modeler 的模型中及 Analysis Services 数据库中存储的模型对象的说明中。

图片 3-21
MS 时间序列模型块的服务器选项

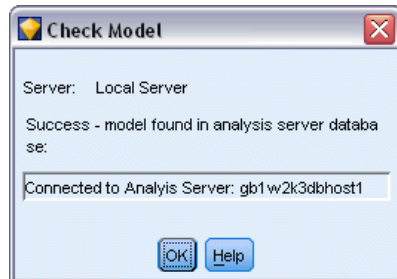


在“服务器”选项卡上，可以为评分操作配置分析服务器主机和数据库及 SQL Server 数据源。在 IBM® SPSS® Modeler 中，此处指定的选项将覆盖那些在“辅助应用程序”或“构建模型”对话框中指定的选项。[有关详细信息，请参阅第 15 页码启用与 Analysis Services 的集成。](#)

模型 GUID。 在此处显示模型关键字。此关键字是在构建模型时随机生成的，并存储于 SPSS Modeler 的模型中及 Analysis Services 数据库中存储的模型对象的说明中。

检查。 单击此按钮对照存储于 Analysis Services 数据库中的模型的关键字检查此模型关键字。此操作有助于验证模型是否仍存在于分析服务器中，并表示模型的结构未更改。

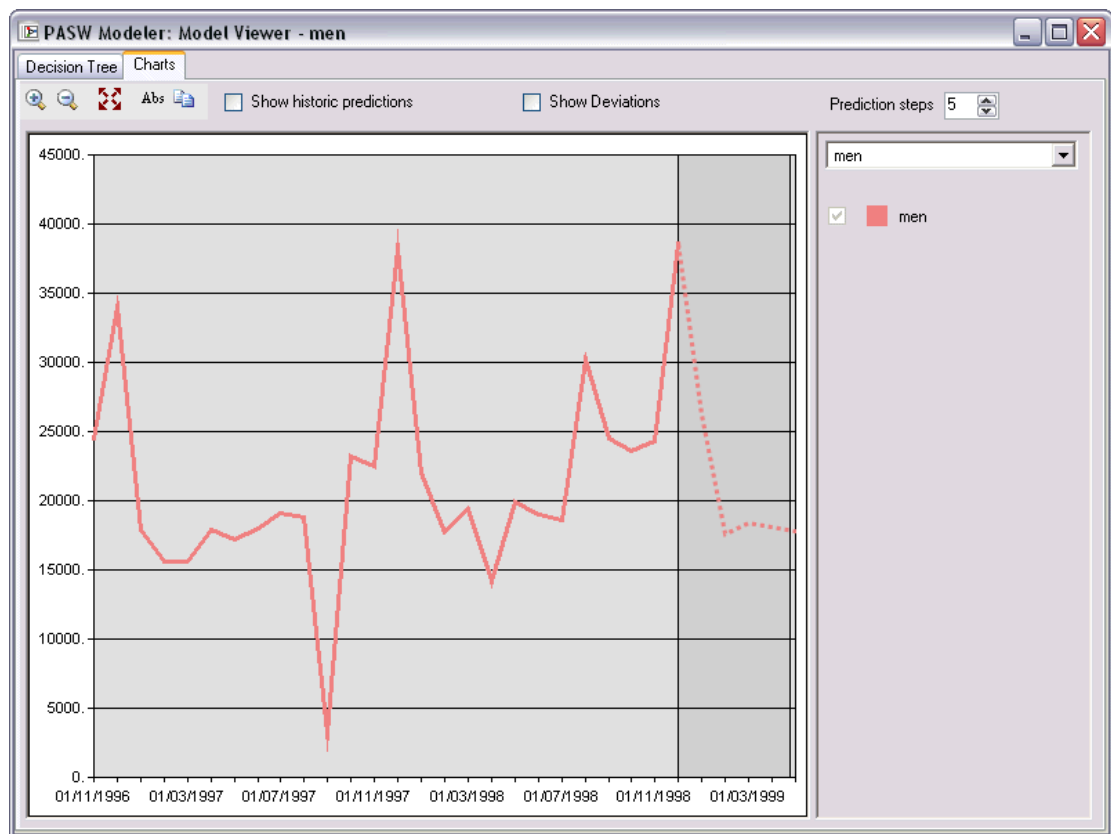
图片 3-22
检查模型关键字的结果



注意：仅在添加到流工作区的准备评分的模型中，才可使用“检查”按钮。如果检查失败，可调查此模型是否已被删除或被服务器上的其他模型替换。

视图。 单击以打开时间序列模型的图形视图。Analysis Services 将完成的模型显示为树。您也可查看显示目标字段的历史值和预测将来值随时间变化的图形。

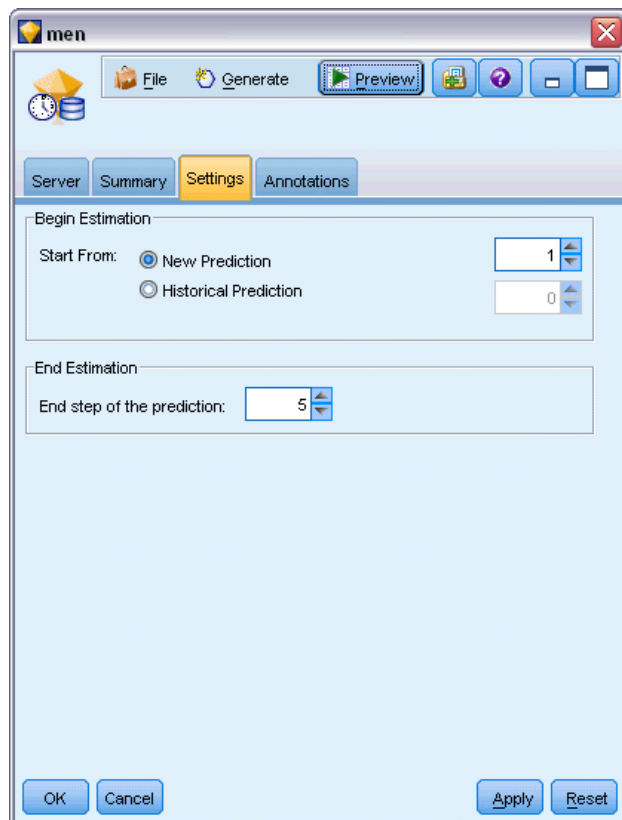
图片 3-23
具有历史值（实线）和预测将来值（虚线）的 MS 时间序列查看器



有关更多信息，请参阅 MSDN 库中对时间序列查看器的说明，位置在 <http://msdn.microsoft.com/en-us/library/ms175331.aspx>。

MS 时间序列模型块设置选项卡

图片 3-24
MS 时间序列模型块的设置选项



开始估计。 指定预测开始的时间段。

- **开始于：新预测。** 未来预测开始的时间段，表示为最后一个历史数据时间段的偏移值。例如，如果您的历史数据结束于 12/99，且您想在 01/00 开始预测，则应使用值 1；但如果您想在 03/00 开始预测，则应使用值 3。
- **开始于：历史预测。** 历史预测开始的时间段，表示为最后一个历史数据时间段的负偏移值。例如，如果您的历史数据结束于 12/99，且您想对数据的最后五个时间段进行历史预测，则使用值 -5。

结束估计。 指定预测停止的时间段。

- **预测的结束步骤。** 预测停止的时间段，表示为最后一个历史数据时间段的偏移值。例如，如果您的历史数据结束于 12/99，且您想在 6/00 停止预测，则应在这里使用值 6。对于未来预测，值必须总是大于或等于开始于值。

MS 序列聚类模型块

以下字段被添加到 MS 序列聚类模型（其中 field 是目标字段的名称）：

字段名称	描述
\$MC-field	此序列所属的聚类的预测。
\$MCP-field	此序列属于预测聚类的概率。
\$MS-field	field 的预测值
\$MSP-field	\$MS-field 值正确的概率。

导出模型和生成节点

可以将模型汇总和结构导出到文本文件和 HTML 文件。需要时还可以生成相应的选择和过滤节点。有关详细信息，请参阅第 3 章中的浏览模型块中的 IBM SPSS Modeler 15 建模节点。

与 IBM® SPSS® Modeler 中的其他模型块类似，Microsoft Analysis Services 模型块支持直接生成记录和字段操作节点。使用模型块“生成”菜单项，您可以生成以下节点：

- 选择节点（仅当在“模型”选项卡上选中某项时）
- 过滤节点

Analysis Services 数据挖掘示例

其中包含多个样本流，这些样本流演示了如何使用 IBM® SPSS® Modeler 进行 MS Analysis Services 数据挖掘。这些流位于 SPSS Modeler 安装文件夹中，该文件夹目录为：

\Demos\Database_Modelling\Microsoft

注意：可以从 Windows “开始”菜单 IBM SPSS Modeler 程序组中访问这些 Demos 文件夹。

示例流：决策树

下列流按顺序一起使用可作为使用由 MS Analysis Services 提供的决策树算法的数据库挖掘过程的示例。

流	说明
1_upload_data.str	用于净化数据和将数据从平面文件上载到数据库。
2_explore_data.str	提供关于 IBM® SPSS® Modeler 数据探索的示例
3_build_model.str	采用数据库自有算法构建模型。
4_evaluate_model.str	用作 SPSS Modeler 模型评估的示例
5_deploy_model.str	部署用于数据库内评分的模型。

注：要运行此示例，必须按此顺序执行各个流。此外，必须更新每个流中的源节点和建模节点，以便将想使用的数据库作为有效数据源供您引用。

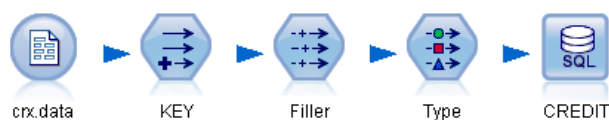
这些示例流中使用的数据集与信用卡申请有关，演示了同时带有分类和连续预测变量的分类问题。关于此数据集的更多信息，请参阅示例流中同一文件夹下的 `crx.names` 文件。

此数据集可从位于 `ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/` 的 UCI Machine Learning Repository 中获得。

示例流：上载数据

第 1 个示例流，即 `1_upload_data.str`，用于清除数据和将数据从纯文本文件上传到 SQL 服务器。

图片 3-25
用于上传数据的示例流



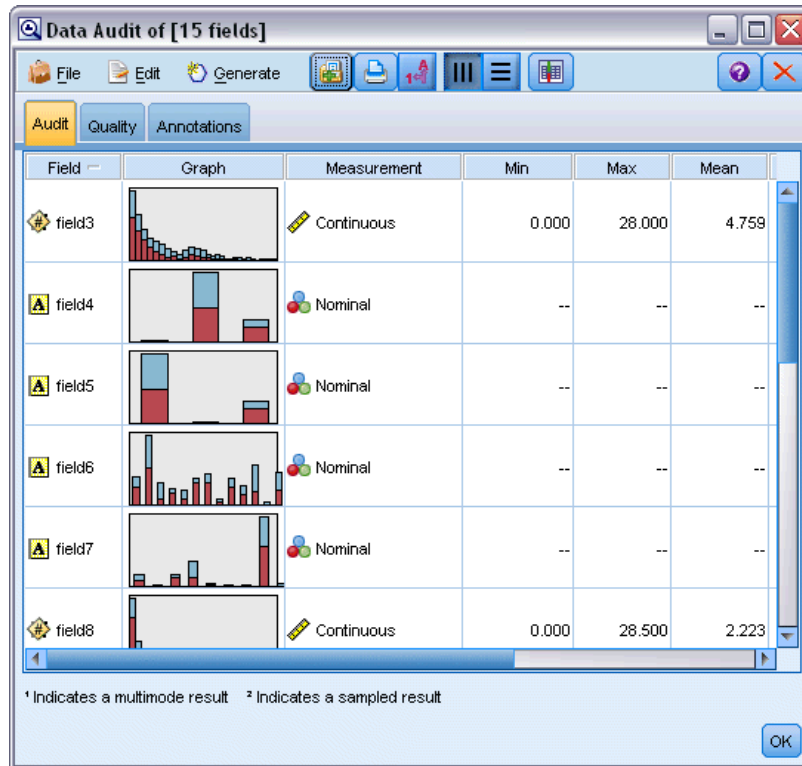
由于 Analysis Services 数据挖掘需要一个关键字字段，因此此初始流使用导出节点将新字段添加到名为 KEY 的数据集中（唯一值为 1,2,3，使用 IBM® SPSS® Modeler@INDEX 函数）。

随后的填充节点用于缺失值处理，并将从文本文件 `crx.data` 中读取的空字段替换为 NULL 值。

示例流：探索数据

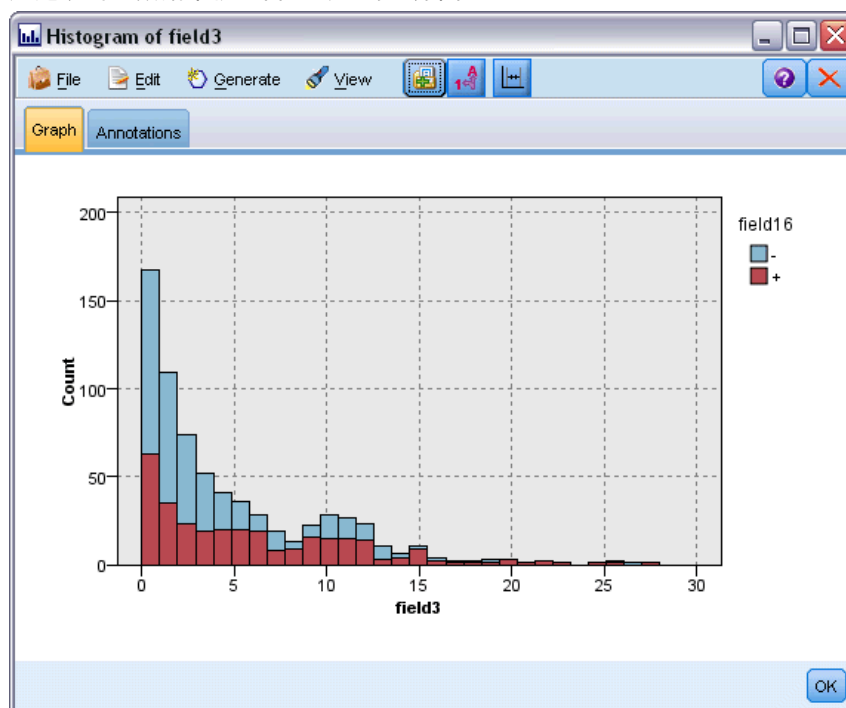
第二个示例流 `2_explore_data.str` 用于演示如何使用数据审核节点获取数据（包括汇总统计量和图形）的一般概述。有关详细信息，请参阅第 6 章中的数据审核节点中的 IBM SPSS Modeler 15 源、过程和输出节点。

图片 3-26
数据审核结果



双击“数据审核报告”中的图形可显示一个更为详细的图形，用于更深入地探索给定字段。

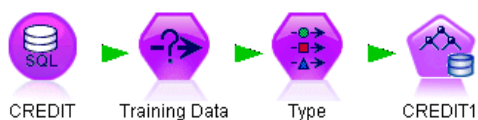
图片 3-27
通过双击“数据审核”窗口创建的直方图



示例流：构建模型

第 3 个示例流，即 3_build_model.str，演示 IBM® SPSS® Modeler 中的模型构建。可将数据库模型附加到流并通过双击指定构建设置。

图片 3-28
数据库建模示例流中的紫色节点表示在数据库内执行



在此对话框的“模型”选项卡上，可以指定以下设置：

- 选择 Key 字段作为唯一标识字段。

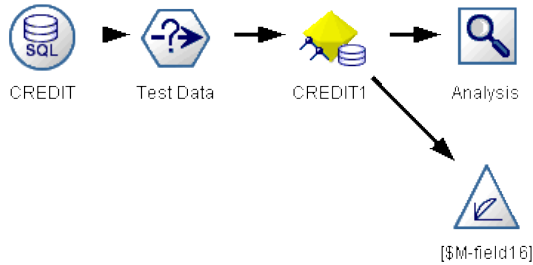
在“专家”选项卡上，可以微调设置以构建模型。

运行前，请确保已为模型构建指定了正确的数据库。使用“服务器”选项卡调整设置。

示例流：评估模型

第 4 个示例流，即 4_evaluate_model.str，演示构建数据库内模型时使用 IBM® SPSS® Modeler 的优点。执行该模型之后，可以将其添加回数据流，并使用 SPSS Modeler 中提供的几个工具评估该模型。

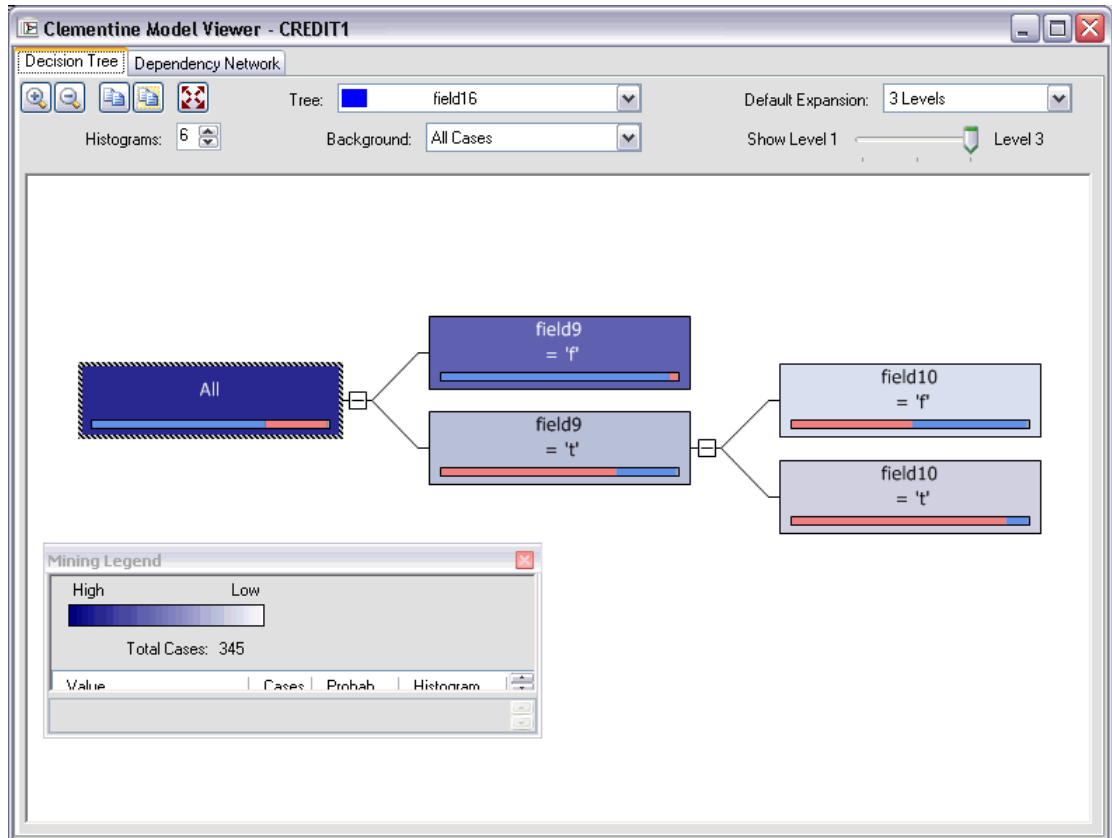
图片 3-29
用于模型评估的示例流



查看建模结果

您可以双击模型块探索结果。“汇总”选项卡提供了结果的规则树视图。还可以单击视图按钮（位于“服务器”选项卡上）来查看决策树模型的图形视图。

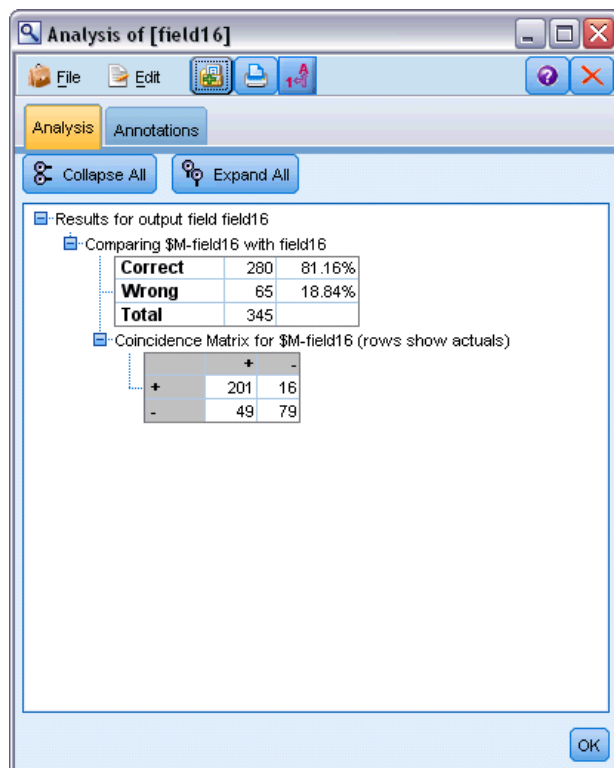
图片 3-30
提供 MS 决策树模型结果的图形视图的查看器



评估建模结果

样本流中的“分析”节点创建说明每个预测字段及其目标字段之间的匹配模式的符合矩阵。执行分析节点以查看结果。

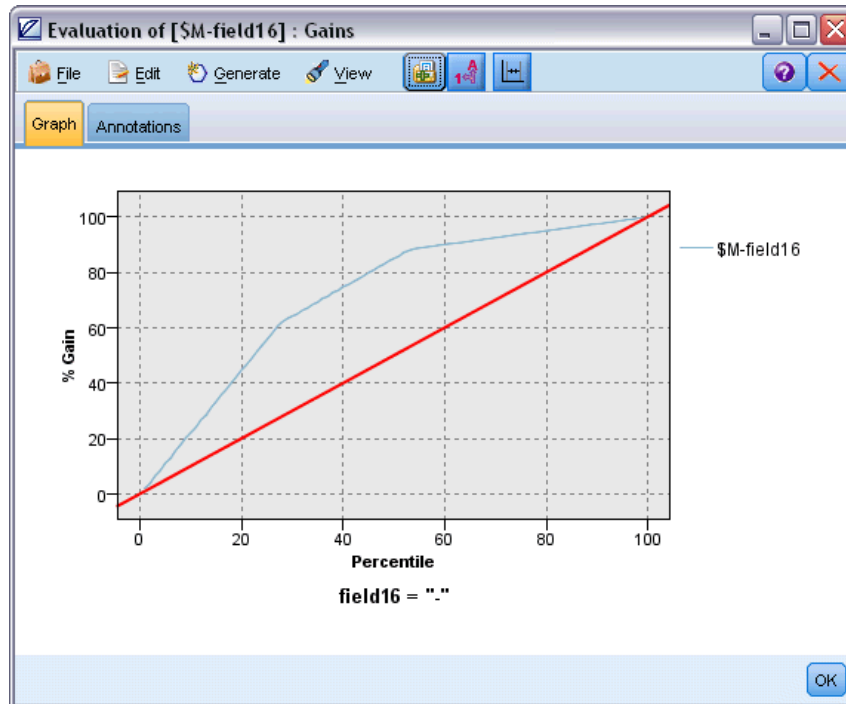
图片 3-31
分析节点结果



该表格表明 MS 决策树算法生成的预测中 81.16% 是正确的。

样本流中的评估节点创建收益图表，以显示模型对准确率提高。执行评估节点以查看结果。

图片 3-32
使用评估节点生成的收益图表



示例流：部署模型

一旦对模型准确率感到满意，即可部署该模型，以用于外部应用程序或往回发布到数据库。在最后一个示例流 5_deploy_model.str 中，将从表 CREDIT 中读取数据，然后使用数据库导出节点对数据进行评分，并将数据发布到表 CREDITSCORES。

图片 3-33
用于部署模型的示例流



运行流后生成以下 SQL：

```
DROP TABLE CREDITSCORES
```

```
CREATE TABLE CREDITSCORES ( "field1" varchar(1),"field2" varchar(255),"field3" float,"field4" varchar(1),"field5" varchar(2),"field6" varchar(2),"field7" varchar(2),"field8" float,"field9" varchar(1),"field10" varchar(1),"field11" int,"field12" varchar(1),"field13" varchar(1),"field14" int,"field15" int,"field16" varchar(1),"KEY" int,"$M-field16" varchar(9),"$MC-field16" float )
```

```
INSERT INTO CREDITSCORES ("field1", "field2", "field3", "field4", "field5", "field6", "field7", "field8",
```



```

"field9", "field10", "field11", "field12", "field13", "field14", "field15", "field16",
"KEY", "$M-field16", $MC-field16")
SELECT TO. C0 AS C0, TO. C1 AS C1, TO. C2 AS C2, TO. C3 AS C3, TO. C4 AS C4, TO. C5 AS C5,
    TO. C6 AS C6, TO. C7 AS C7, TO. C8 AS C8, TO. C9 AS C9, TO. C10 AS C10,
    TO. C11 AS C11, TO. C12 AS C12, TO. C13 AS C13, TO. C14 AS C14,
    TO. C15 AS C15, TO. C16 AS C16, TO. C17 AS C17, TO. C18 AS C18
FROM (
    SELECT CONVERT(NVARCHAR, [TA].[field1]) AS C0, CONVERT(NVARCHAR, [TA].[field2]) AS C1,
        [TA].[field3] AS C2, CONVERT(NVARCHAR, [TA].[field4]) AS C3,
        CONVERT(NVARCHAR, [TA].[field5]) AS C4, CONVERT(NVARCHAR, [TA].[field6]) AS C5,
        CONVERT(NVARCHAR, [TA].[field7]) AS C6, [TA].[field8] AS C7,
        CONVERT(NVARCHAR, [TA].[field9]) AS C8, CONVERT(NVARCHAR, [TA].[field10]) AS C9,
        [TA].[field11] AS C10, CONVERT(NVARCHAR, [TA].[field12]) AS C11,
        CONVERT(NVARCHAR, [TA].[field13]) AS C12, [TA].[field14] AS C13,
        [TA].[field15] AS C14, CONVERT(NVARCHAR, [TA].[field16]) AS C15,
        [TA].[KEY] AS C16, CONVERT(NVARCHAR, [TA].[$M-field16]) AS C17,
        [TA].[$MC-field16] AS C18
    FROM openrowset('MSOLAP',
        'Datasource=localhost;Initial catalog=FoodMart 2000',
        'SELECT [T].[C0] AS [field1], [T].[C1] AS [field2], [T].[C2] AS [field3],
            [T].[C3] AS [field4], [T].[C4] AS [field5], [T].[C5] AS [field6],
            [T].[C6] AS [field7], [T].[C7] AS [field8], [T].[C8] AS [field9],
            [T].[C9] AS [field10], [T].[C10] AS [field11], [T].[C11] AS [field12],
            [T].[C12] AS [field13], [T].[C13] AS [field14], [T].[C14] AS [field15],
            [T].[C15] AS [field16], [T].[C16] AS [KEY], [CREDIT1].[field16] AS [$M-field16],
            PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
        FROM [CREDIT1] PREDICTION JOIN
            openrowset('MSDASQL',
                'Dsn=LocalServer;Uid=;pwd=', 'SELECT TO. "field1" AS C0, TO. "field2" AS C1,
                TO. "field3" AS C2, TO. "field4" AS C3, TO. "field5" AS C4, TO. "field6" AS C5,
                TO. "field7" AS C6, TO. "field8" AS C7, TO. "field9" AS C8, TO. "field10" AS C9,
                TO. "field11" AS C10, TO. "field12" AS C11, TO. "field13" AS C12,
                TO. "field14" AS C13, TO. "field15" AS C14, TO. "field16" AS C15,
                TO. "KEY" AS C16 FROM "dbo". CREDITDATA TO') AS [T]
        ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
            and [T].[C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
            and [T].[C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
            and [T].[C14] = [CREDIT1].[field15]') AS [TA]
    ) TO

```

使用 Oracle Data Mining 构建数据库模型

关于 Oracle Data Mining

IBM® SPSS® Modeler 支持与 Oracle Data Mining (ODM) 的集成，ODM 提供了紧密内嵌于 Oracle RDBMS 中的一系列数据挖掘算法。这些功能可通过访问 SPSS Modeler 的图形用户界面和面向工作流的开发环境加以使用，使客户可以充分利用 ODM 提供的数据挖掘算法。

SPSS Modeler 可集成 Oracle Data Mining 中所含的下列算法：

- Naive Bayes
- Adaptive Bayes
- Support Vector Machine (SVM)
- 广义线性模型 (GLM)*
- 决策树
- O-Cluster
- K-Means
- 非负矩阵分解 (NMF)
- Apriori
- 最小描述符长度 (MDL)
- 属性重要性 (AI)

* 11仅限 g R1

集成 Oracle 的要求

以下是使用 Oracle Data Mining 执行数据库内建模的必备条件。您可能需要咨询数据库管理员以确保满足这些条件。

- 以本地模式或在 Windows 或 UNIX 上安装 IBM® SPSS® Modeler Server 后运行 IBM® SPSS® Modeler。
- 带有 Oracle Data Mining 选项的 Oracle 10gR2 或 11gR1 (10.2 或更高版本的数据库)。

注意：10gR2 支持广义线性模型（需要 11gR1）以外的所有数据库建模算法。

- 连接 Oracle（如下所述）的 ODBC 数据源。

注意：数据库建模和 SQL 优化需要在 SPSS Modeler 计算机上启用 SPSS Modeler Server 连接。通过启用此设置，您可以访问数据库算法，直接从 SPSS Modeler 回送 SQL 以及访问 SPSS Modeler Server。要验证当前许可证的状态，请从 SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项服务器启用。

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 15 用户指南](#)。

启用 Oracle 集成

要启用 Oracle Data Mining 的 IBM® SPSS® Modeler 集成，需要配置 Oracle 并创建 ODBC 源，启用 SPSS Modeler “辅助应用程序”对话框中的集成功能，并启用 SQL 生成和优化。

配置 Oracle

要安装和配置 Oracle Data Mining，请参阅 Oracle 文档—特别是 Oracle 管理员指南一，以获得更多详细信息。

为 Oracle 创建 ODBC 源

要启用 Oracle 和 SPSS Modeler 之间的连接，您需要创建 ODBC 系统数据源名称 (DSN)。

在创建 DSN 之前，您应当对 ODBC 数据源和驱动程序，以及 SPSS Modeler 中的数据库支持有基本的了解。有关详细信息，请参阅第 2 章中的[数据访问中的 IBM SPSS Modeler Server 15 管理和性能指南](#)。

如果以分布式方式运行 IBM® SPSS® Modeler Server，请在服务器计算机上创建 DSN。如果以本地（客户机）模式运行，请在客户计算机上创建 DSN。

- ▶ 安装 ODBC 驱动程序。您可在随此版本附带的 IBM® SPSS® Data Access Pack 安装盘上找到这些驱动程序。运行 setup.exe 文件以启动安装程序，并选择所有相关的驱动程序。按照屏幕说明操作以安装驱动程序。
- ▶ 创建 DSN。

注意：菜单排列顺序取决于 Windows 版本。

 - **Windows XP。**从“开始”菜单中选择控制面板。双击管理工具，然后双击数据源 (ODBC)。
 - **Windows Vista。**从“开始”菜单中选择控制面板，然后选择系统维护。双击管理工具，选择数据源 (ODBC)，然后单击打开。
 - **Windows 7。**从“开始”菜单中选择控制面板，选择系统和安全，然后选择管理工具。选择数据源 (ODBC)，然后单击打开。
- ▶ 单击系统 DSN 选项卡，然后单击添加。

- ▶ 选择 SPSS OEM 6.0 Oracle Wire Protocol 驱动程序。
- ▶ 单击完成。
- ▶ 在 ODBC Oracle Wire Protocol 驱动程序安装屏幕中，输入选择的数据源名称、Oracle 服务器的主机名、连接端口号及使用的 Oracle 示例的 SID。
如果已使用 tnsnames.ora 文件配置了 TNS，则可以从服务器计算机的 tnsnames.ora 文件获取主机名、端口和 SID。要获取更多信息，请联系 Oracle 管理员。
- ▶ 请单击 测试 按钮，以测试连接。

在 IBM SPSS Modeler 中启用 Oracle Data Mining 集成

- ▶ 从 SPSS Modeler 菜单中选择：
工具 > 选项 > 辅助应用程序
- ▶ 单击 Oracle 选项卡。

启用 Oracle Data Mining 集成。 启用 SPSS Modeler 窗口底部的“数据库建模”选项板（如尚未显示）并添加 Oracle Data Mining 算法的建模节点。

Oracle 连接。 请指定用于构建和保存模型的默认 Oracle ODBC 数据源，以及有效的用户名和密码。可在各个建模节点和模型块上覆盖此设置。

注意：用于建模的数据库连接可以与用于访问数据的连接相同，也可以不相同。例如，可能有一个流可以用于访问一个 Oracle 数据库的数据，将数据下载到 SPSS Modeler 以进行清理或执行其他操作，然后将数据上传到另一个 Oracle 数据库，用于建模。此外，也可以将原始数据放置在平面文件或其他（非 Oracle）源中，但这种情况下需要将数据上传到 Oracle，才能用于建模。所有情况下数据都将自动上传到在用于建模的数据库中创建的一个临时表格中。

覆盖 Oracle Data Mining 模型时发出警告。 选中此选项，可确保只有在发出警告的情况下，数据库中存储的模型才会被 SPSS Modeler 覆盖。

列出 Oracle Data Mining 模型。 显示可用数据挖掘模型。

启用 Oracle Data Miner 的启动。（可选） 启用该选项后 SPSS Modeler 即可启动 Oracle Data Miner 应用程序。详细信息请参阅 [Oracle Data Miner 第 84 页码](#)。

Oracle Data Miner 可执行程序路径。（可选） 用于指定 Oracle Data Miner for Windows 可执行文件的物理位置（例如 C:\odm\bin\odminerw.exe）。Oracle Data Miner 未与 SPSS Modeler 一起安装，必须从 Oracle 网站 (<http://www.oracle.com/technology/products/bi/odm/odminer.html>) 下载正确的版本并在客户端进行安装。

启用 SQL 生成和优化

- ▶ 从 SPSS Modeler 菜单中选择：
工具 > 流属性 > 选项

图片 4-1
优化设置



- ▶ 在导航窗格中单击优化选项。
- ▶ 确认是否已启用生成 SQL 选项。要使数据库建模正常发挥作用，此设置是必需的。
- ▶ 选中优化 SQL 生成和优化其他执行（非严格必需但强烈推荐使用，以使性能更优）。

有关详细信息，请参阅第 5 章中的设置流的优化选项中的 IBM SPSS Modeler 15 用户指南。

使用 Oracle Data Mining 构建模型

Oracle 建模节点的工作方式与 IBM® SPSS® Modeler 中其他建模节点的一样，不过也有几个例外。可通过横向显示在 SPSS Modeler 窗口底部的数据库建模选项板来访问这些节点。

图片 4-2
数据库建模选项板



数据注意事项

Oracle 要求分类数据应以字符串格式 (CHAR 或 VARCHAR2) 存储。因此, SPSS Modeler 不允许将测量级别为标志或名义 (分类) 的数字存储字段指定为 ODM 模型的输入。如有必要, 可在 SPSS Modeler 中使用“重新分类”节点将数字转换为字符串。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 15 源、过程和输出节点。

目标字段。 只能选择一个字段作为 ODM 分类模型的输出字段。

模型名称。 从 Oracle 11gR1 之后, 名称 `unique` 已成为关键字, 不能用作自定义模型名称。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如, ID 字段, 如客户 ID。SPSS Modeler 限定此关键字字段必须为数值。

注: 除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外, 此字段对所有 Oracle 节点都是可选的。

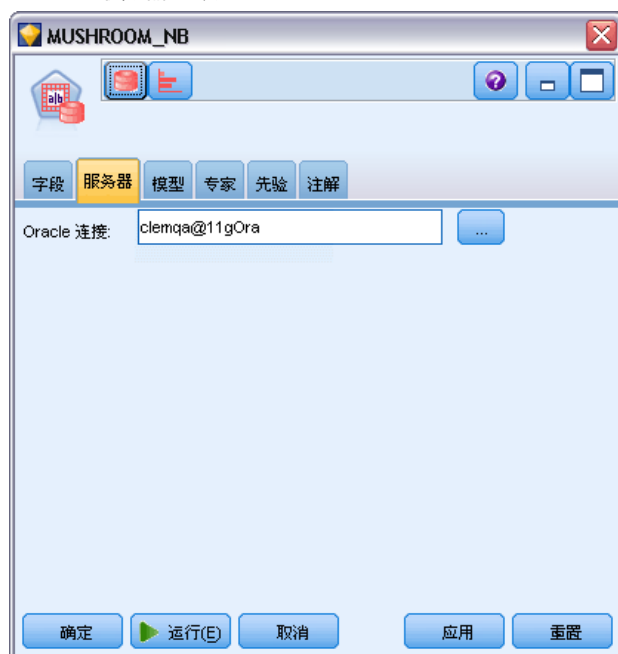
一般评论

- 对于 Oracle Data Mining 创建的模型, SPSS Modeler 不提供 PMML 导出/导入功能。
- 模型评分总是在 ODM 内执行。如果数据来自于 SPSS Modeler 或需要在其中准备数据, 则需要将数据集上载到临时表。
- 在 SPSS Modeler 中, 通常只提供一次预测以及关联的概率或置信度。
- SPSS Modeler 将模型构建和评分中的可用字段的数量限制为 1000。
- SPSS Modeler 可以从使用 IBM® SPSS® Modeler Solution Publisher 发布用于执行的流来对 ODM 模型进行评分。有关详细信息, 请参阅第 2 章中的 IBM SPSS Modeler Solution Publisher 的工作原理中的 IBM SPSS Modeler 15 Solution Publisher。

Oracle 模型服务器选项

指定用于上传建模数据的 Oracle 连接。必要的话, 可以在“服务器”选项卡上为每个建模节点选择一个连接, 以覆盖“辅助应用程序”对话框中指定的默认 Oracle 连接。有关详细信息, 请参阅第 49 页码启用 Oracle 集成。

图片 4-3
Oracle 服务器选项

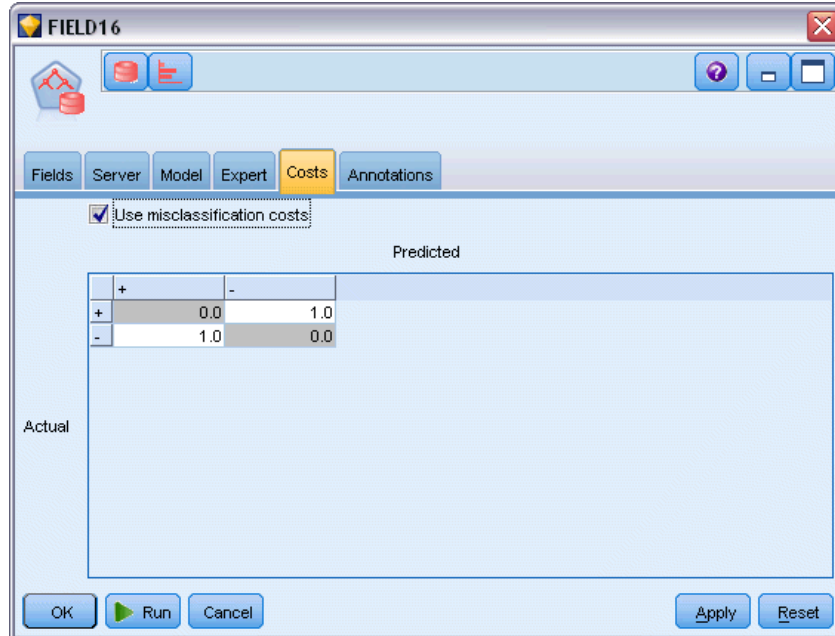


注释

- 用于建模的连接可以与流的源节点中使用的连接相同，也可以不相同。例如，可能有一个流可以用于访问一个 Oracle 数据库的数据，将数据下载到 IBM® SPSS® Modeler 以进行清理或执行其他操作，然后将数据上传到另一个 Oracle 数据库，用于建模。
- ODBC 数据源名称可有效地内嵌于每个 SPSS Modeler 流中。如果在一个主机创建的流在另一个主机上执行，则各个主机上的数据源名称必须一样。此外，也可以在各个源或建模节点中的“服务器”选项卡上选择另一个数据源。

误分类损失

图片 4-4
Oracle 损失选项



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。

注意：仅“决策树”模型允许在构建时指定损失。

Oracle Naive Bayes

Naive Bayes 是广泛用于处理分类问题的算法。因为该模型将所有给出的预测变量视为互相独立的，因而取名为 naïve。Naive Bayes 是一种快速的、可伸缩的算法，用于计算属性和目标属性组合的条件概率。使用训练数据确定独立的概率。给定来自每个输入变量的所有值分类的发生率，使用此概率可计算出每个目标类的似然值。

- 交叉验证用于检验模型拟合（用于构建模型的）数据的准确性。如果可用于构建模型的观测值的数量很小，则该交叉验证特别有用。
- 模型输出可用矩阵格式浏览。矩阵中的数字为条件概率，这些条件概率与预测的类（列）和预测变量值的组合（行）相关联。

Naive Bayes 模型选项

图片 4-5
Naive Bayes 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 15 源、过程和输出节点。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

Naive Bayes 专家选项

图片 4-6
Naive Bayes 专家选项



除非给定的值或值对在训练数据中具有足够高的发生率，否则在模型构建后，单个预测变量属性值或值对将被忽略。基于训练数据中的记录数而计算出来的分数值，可指定用于忽略值的发生率临界值。调整此临界值可减少噪声并改进模型拟合其他数据集的能力。

- **单临界值。** 指定给定的预测变量属性值的临界值。给定值的出现次数必须等于或大于指定的分数，否则该值将被忽略。
- **双临界值。** 指定给定属性和预测变量值对的临界值。给定值对的出现次数必须等于或大于指定的分数，否则该值对将被忽略。

预测概率。 允许模型包含目标字段可能结果的正确预测概率。要启用此功能，选择选择，单击指定按钮，选择一个可能结果并单击插入。

使用预测集合。 生成目标字段所有可能结果（outcome）的所有可能结果（result）的表格。

Oracle Adaptive Bayes

Adaptive Bayes Network (ABN) 使用最小描述符长度 (MDL) 和自动特征选择来构造 Bayesian Network 分类符。尽管 ABN 的执行速度慢些，但在 Naive Bayes 表现糟糕的某些情况中它仍有良好表现，而在其他大多数情况下也至少不比 Naive Bayes 差。

ABN 算法能够用于构建三种高级的、基于 Bayesian 的模型，包括简化的决策树（单功能）、修剪的 Naive Bayes 和增强型多功能模型。

已生成的模型

在单功能构建模式中，ABN 可根据一组人可读规则生成一个简化的决策树，使商业用户或分析人员可以了解模型预测的基础并据此向其他人演示或解说。相比于 Naive Bayes 和多功能模型，这是一个突出的优势。这些规则可以像 IBM® SPSS® Modeler 中的标准规则集一样进行浏览。如下所示的是一个简单的规则集：

```
IF MARITAL_STATUS = "Married"  
AND EDUCATION_NUM = "13-16"  
THEN CHURN= "TRUE"  
Confidence = .78, Support = 570 cases
```

修剪的 Naive Bayes 和多功能模型无法在 SPSS Modeler 中浏览。

Adaptive Bayes 模型选项

图片 4-7
Adaptive Bayes 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 15 源、过程和输出节点。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

模型类型

构建模型时有三种不同模式可供选择。

- **多功能。** 构建和对比若干个模型，包括 NB (Naive Bayes) 模型、单功能产品概率模型和多功能产品概率模型。这是最详尽的模式，但通常所需的计算时间也最长。只有单功能模型胜出而成为最佳模型时，才会产生规则。如果选择了多功能或 NB 模型，则不会生成任何规则。
- **单功能。** 根据规则集创建简化决策树。每个规则均含有一个条件以及与每个结果关联的概率。各规则互相排斥且其为人可读格式，这可能是相比于 Naive Bayes 和多功能模型的重要优点。
- **Naive Bayes。** 构建单一 NB 模型并将它与全局样本先验分布进行对比（全局样本中目标值的分布）。只有 NB 模型胜出而成为比全局先验分布更好的目标值预测变量时，才产生 NB 模型作为输出。否则，将不会输出任何模型。

Adaptive Bayes 专家选项

图片 4-8
Adaptive Bayes 专家选项



限制执行时间。 请选择此选项来指定以分钟表示的最长构建时间。此选项可用于缩短模型生成时间，不过这样一来，所生成的模型准确性较差。该算法将在建模过程的每个重要步骤检验是否能够在指定的时间内完成下一个重要步骤，然后再继续下一步，并在达到限制时返回可用的最佳模型。

预测变量的最大数量。 此选项可用于通过限制使用的预测变量的数量，来限制模型的复杂性和提高执行速度。预测变量将根据预测变量与目标相关性的 MDL 度量值来进行排序，此排序度量了预测变量包含在模型中的可能性。

Naive Bayes 预测变量的最大数量。 此选项指定 Naive Bayes 模型中使用的预测变量的最大数。

Oracle Support Vector Machine (SVM)

Support Vector Machine (SVM) 是一种分类和回归算法，它使用机器学习理论在不过度拟合数据的同时，最大限度地提高预测准确性。SVM 使用训练数据的可选非线性转换方法，然后搜索已转换数据中的回归方程，以分类（对于分类目标）或匹配目标（对于连续目标）。Oracle 上配置了 SVM 后，就可以使用两个可用核函数（线性和高斯）的其中一个来构建模型。线性核函数完全忽略了非线性转换，使得生成的模型本质上为回归模型。

详细信息请参阅《Oracle Data Mining 应用程序开发人员指南》和《Oracle Data Mining 概念》。

Oracle SVM 模型选项

图片 4-9
SVM 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。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

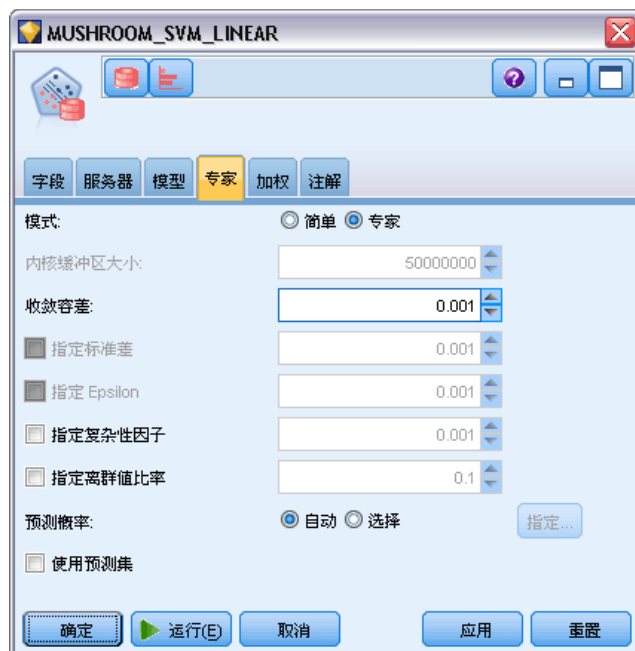
主动学习。 提供处理大型建模数据集的方法。算法可使用主动学习，根据小样本创建一个初始模型，随后将初始模型应用到完整的训练数据集中，再根据结果递增地更新样本和模型。更新循环将不断重复，直到模型在训练数据上收敛，或支持向量的数量达到了允许的最大值。

核函数。 选择线性或高斯，或保留默认的系统已确定允许系统选择最适合的内核。Gaussian 核函数模拟更复杂的关系，但一般来说，耗费的计算时间更长。可先使用线性核函数，然后如果线性核函数未能找到合适的拟合，再尝试使用高斯核函数。这种方法在回归模型中更常用，因为回归模型中核函数的选择更重要。同时请注意，用高斯核函数构建的 SVM 模型在 SPSS Modeler 中无法浏览。用线性核函数构建的模型则可以像浏览标准回归模型一样在 SPSS Modeler 中进行浏览。

正态化方法。 指定连续输入和目标字段的标准化方法。可选择 Z-Score、最值法 或 无。如果选中自动数据准备复选框，Oracle 将自动执行标准化。取消选中此复选框以选择手动标准化方法。

Oracle SVM 专家选项

图片 4-10
SVM 专家选项



核心缓存大小。 指定以字节表示的缓存大小，该缓存用于保存构建操作期间计算的核函数。如所预期，较大的缓存通常构建速度更快。默认值为 50MB。

收敛容差。 指定模型构建终止前允许的容差值。该值必须处于 0 到 1 之间，默认值为 0.001。值较大，构建速度也较快，但模型准确率较低。

指定标准差。 指定高斯核函数使用的标准差参数。此参数影响着模型的复杂度和拓展到其他数据集的能力（即数据的过度拟合和失度拟合）之间的平衡。标准差值越高，越容易倾向于失度拟合。此参数值默认通过训练数据估算得出。

指定 Epsilon。 仅适用于回归模型，用于指定构建对 epsilon 不敏感的模型时可允许错误的区间的值。换言之，它用于区分小错误（忽略）与大错误（不可忽略）。该值必须处于 0 到 1 之间。默认情况下，该值将通过训练数据估算得出。

指定复杂性因子。 指定复杂性因子，复杂性因子用于平衡模型错误（通过训练数据测量出）和模型复杂度，以防止数据的过度拟合和失度拟合。该值越高则对错误的罚分就越高，数据过度拟合的风险也越高；值越低则对错误的罚分就越低，也就越容易数据的失度拟合。

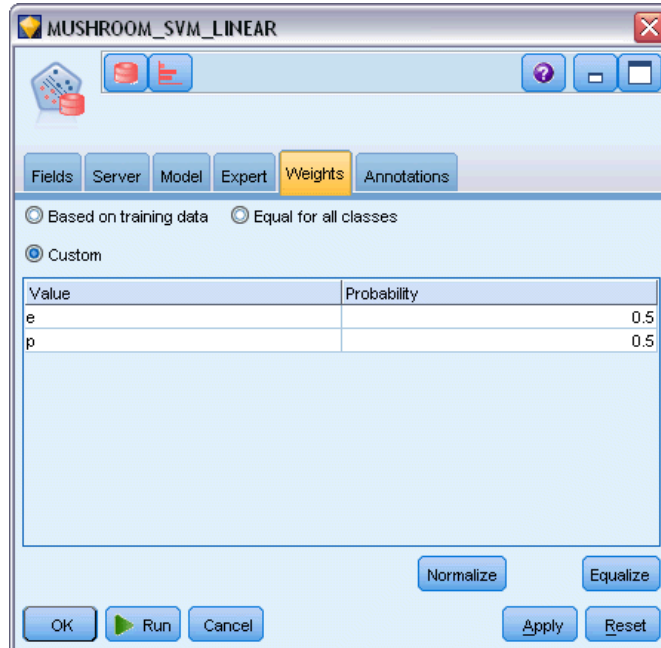
指定离群值比率。 指定训练数据中期望的离群值比率。只对一级 SVM 模型有效。不能与**指定复杂性因子**设置一起使用。

预测概率。 允许模型包含目标字段可能结果的正确预测概率。要启用此功能，选择选择，单击指定按钮，选择一个可能结果并单击插入。

使用预测集合。 生成目标字段所有可能结果（outcome）的所有可能结果（result）的表格。

Oracle SVM 权重选项

图片 4-11
SVM 权重选项



在分类模型中，通过使用权重您可以指定各种可能的目标值的相对重要性。这样做可能非常有用，例如，如果训练数据中的数据点没有实际分布到各个类别中。权重可以使模型产生偏差，以便弥补那些在数据中没有得到合适描述类别。增加目标值的权重会增加该类别获得正确预测的百分比。

有三种方法可用来设置权重：

- **基于训练数据。** 这是默认选项。权重以训练数据中类别的相对频率为基础。
- **对于所有类都相等。** 所有类别的权重都定义为 $1/k$ ，其中 k 是目标类别数。
- **自定义。** 您可以自己指定权重。对于所有类，都将权重的初始值设置为相等。可以将各个分类的权重调整为用户定义的值。要调整特定分类的权重，可在表中对应于所需类别的权重单元格中，先清除其内容，然后输入所需的值。

所有类别的权重之和应为 1.0。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权重约束时保留类别中的比例。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。

Oracle 广义线性模型 (GLM)

(仅 11g) “广义线性模型”放宽了线性模型所作的限制假设。例如，这些包括目标变量有正态分布的假设，以及预测值对目标变量的效应本质上是线性的假设。广义线性模型适合目标分布可能拥有非正态分布的预测值，如多项或泊松分布。类似地，广义线性模型用于预测值与目标之间的关系或链接可能是非线性的个案。

详细信息请参阅《Oracle Data Mining 应用程序开发人员指南》和《Oracle Data Mining 概念》。

Oracle GLM 模型选项

图片 4-12
GLM 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。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

正态化方法。 指定连续输入和目标字段的标准化方法。可选择 Z-Score、最值法 或 无。如果选中自动数据准备复选框，Oracle 将自动执行标准化。取消选中此复选框以选择手动标准化方法。

缺失值处理。 指定如何处理输入数据中的缺失值：

- 替换为均值或众数将数值属性的缺失值替换为均值，并将分类属性的缺失值替换为众数。
- 仅使用完整记录忽略带有缺失值的记录。

Oracle GLM 专家选项

图片 4-13
GLM 专家选项



使用行权重。 选中此复选框以激活相邻下拉列表，从中可以为行选择包含权重因子的列。

将行诊断保存到表格。 选中此复选框以激活相邻文本字段，在此可以输入表格名称以包含行级别诊断。

系数置信级别。 目标的预测值在模型计算的置信区间内的确定性程度，从 0.0 到 1.0。置信边界连同系数统计一起返回。

目标的参考类别。 选择自定义为用作参考类别的目标字段选择值或保留默认值自动。

Ridge 回归。 Ridge 回归是一种补偿在变量中有太高相关性程度的情况的方法。您可以使用自动选项，允许算法控制此方法的使用，或者也可通过禁用和启用选项手动控制。如果您选择手动启用 Ridge 回归，您可以通过在相邻字段中输入一个值覆盖 Ridge 参数的系统默认值。

为 Ridge 回归生成 VIF。 如果您想当 Ridge 正在用于线性回归时生成方差膨胀因子 (VIF) 统计量，选中此复选框。

预测概率。 允许模型包含目标字段可能结果的正确预测概率。要启用此功能，选择选择，单击指定按钮，选择一个可能结果并单击插入。

使用预测集合。 生成目标字段所有可能结果 (outcome) 的所有可能结果 (result) 的表格。

Oracle GLM 权重选项

图片 4-14
GLM 权重选项



在分类模型中，通过使用权重您可以指定各种可能的目标值的相对重要性。这样做可能非常有用，例如，如果训练数据中的数据点没有实际分布到各个类别中。权重可以使模型产生偏差，以便弥补那些在数据中没有得到合适描述类别。增加目标值的权重会增加该类别获得正确预测的百分比。

有三种方法可用来设置权重：

- **基于训练数据。** 这是默认选项。权重以训练数据中类别的相对频率为基础。
- **对于所有类都相等。** 所有类别的权重都定义为 $1/k$ ，其中 k 是目标类别数。
- **自定义。** 您可以自己指定权重。对于所有类，都将权重的初始值设置为相等。可以将各个分类的权重调整为用户定义的值。要调整特定分类的权重，可在表中对应于所需类别的权重单元格中，先清除其内容，然后输入所需的值。

所有类别的权重之和应为 1.0。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权重约束时保留类别中的比例。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。

Oracle 决策树

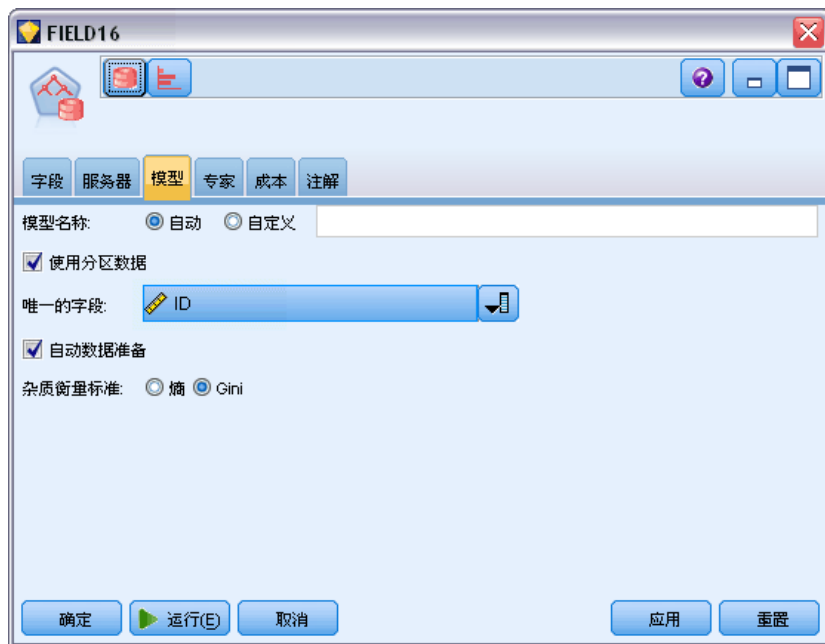
Oracle Data Mining 根据常用的分类和回归树算法，提供了一种经典的决策树功能。ODM 决策树模型含有每个节点的完整信息，包括置信、支持和分割标准。可以显示每个节点的完整规则，而且还提供每个节点的替代属性，该替代属性用于在将模型应用到具有缺失值的观测数据时作为替代。

决策树的广泛应用是因为它适用性广、便于应用及易于理解。决策树将对所有可能的输入属性进行筛选，以查找最佳“分割器”，即属性切割点（例如，AGE > 55），以将下游数据记录分割成若干更均质的总体。每次分割决策后，ODM 将重复长出整个树和创建终端“叶子”的过程，该叶子代表具有类似记录、项目或人员的总体。从树节点的根部往下看（例如，总人口），决策树提供人可读规则的 IF A, then B 语句。这些决策树规则还提供每个树节点的支持和置信。

Adaptive Bayes Networks 也可以提供用于解释每项预测的简单规则，但每个分割决策的 Oracle Data Mining 完整规则是由决策树提供。决策树还可以用于生成最佳客户、已恢复健康病人、与欺骗关联的因子等的详细配置信息。

决策树模型选项

图片 4-15
决策树模型选项



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。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

杂质度量。 指定寻求分割每个节点数据的最佳测试问题时使用的度量。最佳分割器和分隔值是那些能最大限度提高节点中各实体的目标值均一性的分割器和分隔值。均一性通过一个度量值来衡量。受支持的度量值为 **基尼** 和 **熵**。

决策树专家选项

图片 4-16
决策树专家选项



最大深度。 设置要构建的树模型的最大深度。

节点中记录的最小百分比。 设置节点中记录的最小百分比。

进行分割的记录最小百分比。 设置父节点中记录的最小数，该最小数以用于训练模型的记录总数的百分比表示。如果记录数小于此百分比，则不会尝试进行任何分割。

节点中的最小记录数。 设置返回记录的最小数。

用于分割的记录的最小数。 设置父节点中记录的最小数，该最小数以数字表示。如果记录数小于此值，则不会尝试进行任何分割。

规则 ID。 如果选中，模型中会包含一个字符串以在已进行特定分割的树中标识节点。

预测概率。 允许模型包含目标字段可能结果的正确预测概率。要启用此功能，选择选择，单击指定按钮，选择一个可能结果并单击插入。

使用预测集合。 生成目标字段所有可能结果 (outcome) 的所有可能结果 (result) 的表格。

Oracle O-Cluster

Oracle O-Cluster 算法确定数据总体中自然发生的分组。正交分区聚类 (O-Cluster) 是 Oracle 专有的聚类算法，它创建基于分层网格的聚类模型，也就是说，它在输入属性空间中创建轴平行（正交）分区。该算法递归式地运行。所产生的分层结构为一个不规则的网格，该网格将属性空间分割成各个聚类。

O-Cluster 算法可处理数字属性和分类属性，且 ODM 将自动选择最佳的聚类定义。ODM 提供聚类详细信息，聚类规则和聚类矩心值，并可以用于根据总体的聚类成员资格对总体进行评分。

O-Cluster 模型选项

图片 4-17
O-Cluster “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。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

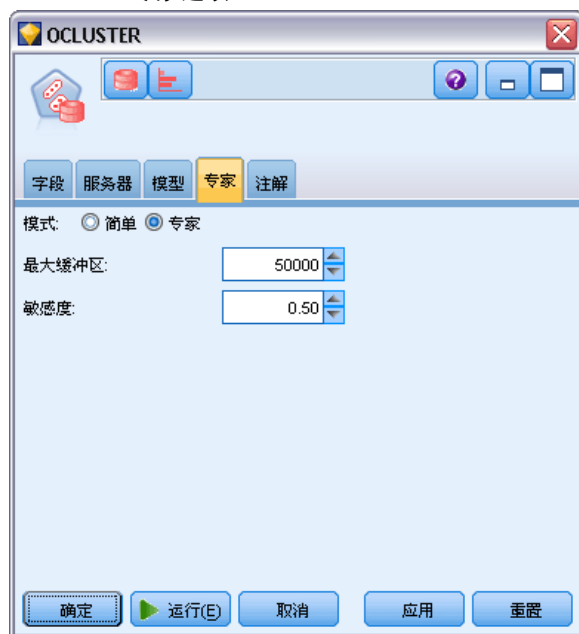
注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

最大聚类数。 设置生成聚类的最大数量。

0-Cluster 专家选项

图片 4-18
0-Cluster 专家选项



最大缓冲区。 设置最大缓冲区大小。

敏感度。 设置一个分数，该分数指定分割新聚类所要求的最高密度。该分数与全局均匀密度相关联。

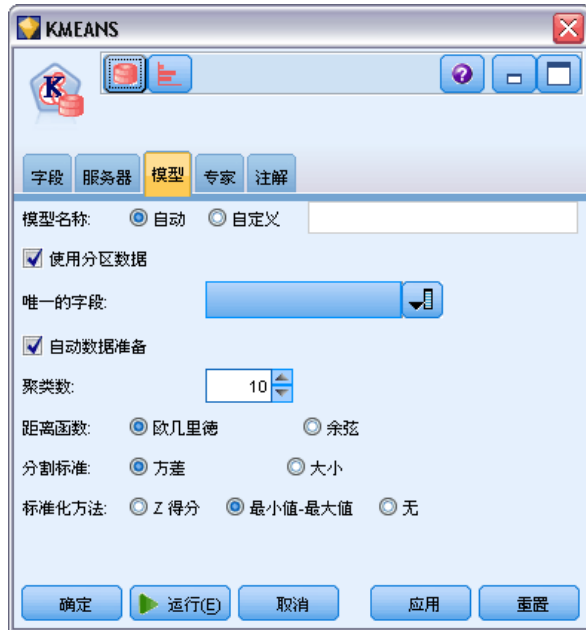
Oracle k-Means

Oracle 0-Cluster 算法确定数据中自然发生的分组。k-Means 算法是基于距离的聚类算法，该算法将数据分区为预定数量的聚类（条件是存在足够的不同观测值）。基于距离的算法根据距离度量（函数）来衡量数据点之间的相似性。根据所使用的距离度量，数据点被指派到与之距离最近的聚类。ODM 提供增强版的 k-Means。

k-Means 算法支持分层聚类，处理数字和分类属性并将总体分割为用户指定数量的聚类。ODM 提供聚类详细信息，聚类规则和聚类矩心值，并可以用于根据总体的聚类成员资格对总体进行评分。

k-Means 模型选项

图片 4-19
k-Means “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。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

聚类数。 设置生成聚类的数量。

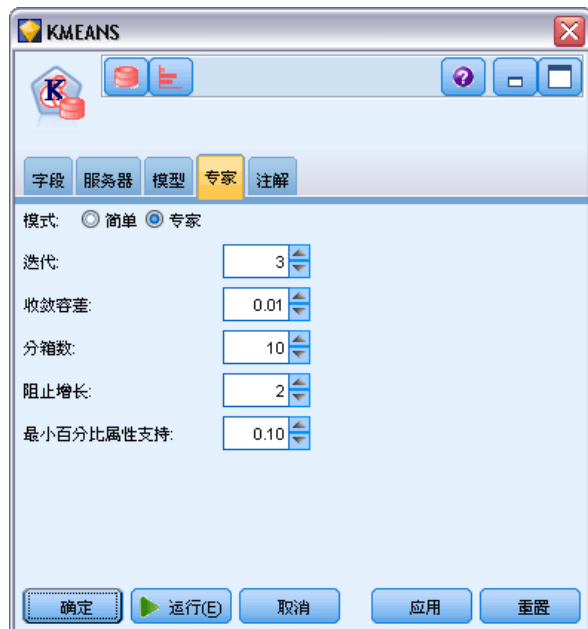
距离函数。 指定 k-Means 聚类使用的距离函数。

分割标准。 指定 k-Means 聚类使用的分割标准。

正态化方法。 指定连续输入和目标字段的标准化方法。可选择 Z-Score、最值法 或 无。

k-Means 专家选项

图片 4-20
k-Means “专家” 选项



迭代。 设置 k-Means 算法的迭代次数。

收敛容差。 设置 k-Means 算法的收敛容差。

图条数。 指定 k-Means 生成的属性直方图中的图条数。每个属性的图条边界都是通过对整个训练数据集进行全局计算得到的。图条方法为等宽法。具有单一值的属性只有一个分类，除此以外，其他所有属性均具有同样数量的图条。

块增长。 设置分配用于容纳聚类数据的内存的增长因子。

最小百分比属性支持度。 设置属性值分数，该属性值必须为非 NULL，才能使该属性包含在聚类的规则说明中。如果参数值在具有缺失值的数据中设置得过高，则可能导致规则过短，或甚至为空。

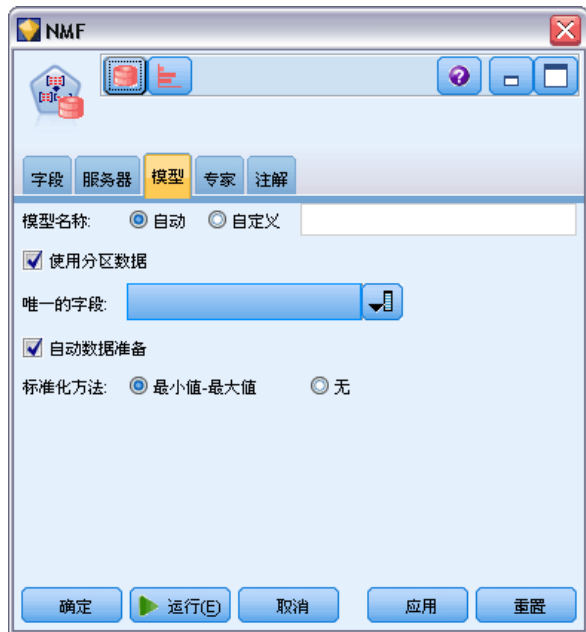
Oracle 非负矩阵分解 (NMF)

非负矩阵分解 (NMF) 用于将大数据集简化为若干具有代表性的属性。它与主成分分析 (PCA) 的原理类似，但可以处理更大量的属性，在加法表示模型中，NMF 是功能强大的先进数据挖掘算法，而且用途广泛。

NMF 可以用于将大量数据（比如文本数据）简化为小的、稀疏得多的表示，NMF 降低了数据的维度，即用少得多的变量保存了等量的信息。NMF 模型的输出可用有监督的学习方法（比如 SVM）或没有监督的学习方法（比如 聚类）来进行分析。Oracle Data Mining 用 NMF 和 SVM 算法来挖掘尚未结构化的文本数据。

NMF 模型选项

图片 4-21
NMF “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。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

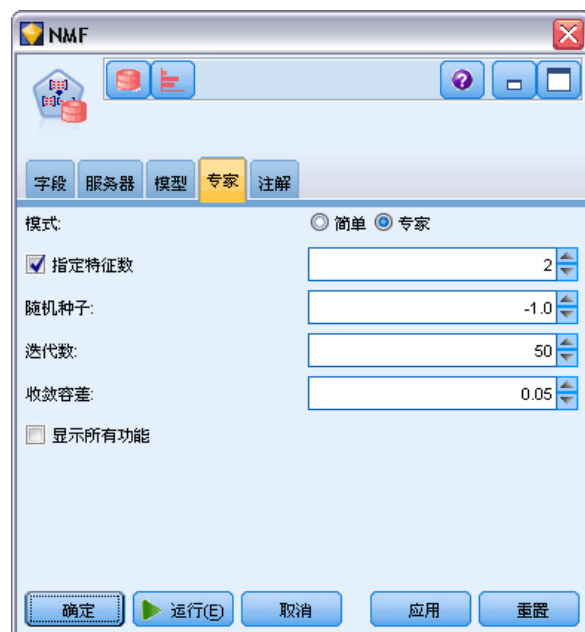
注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

正态化方法。 指定连续输入和目标字段的标准化方法。可选择 Z-Score、最值法 或 无。如果选中自动数据准备复选框，Oracle 将自动执行标准化。取消选中此复选框以选择手动标准化方法。

NMF 专家选项

图片 4-22
NMF “专家” 选项



指定特征数。 指定要提取的特征的数量。

随机种子。 设置 NMF 算法的随机种子。

迭代数。 设置 NMF 算法的迭代数。

收敛容差。 设置 NMF 算法的收敛容差。

显示所有特征。 显示所有特征的特征 ID 和置信度，而不是仅显示最佳特征的特征 ID 和置信度。

Oracle Apriori

Apriori 算法会发现数据中的关联规则。例如，“如果客户购买剃须刀和须后产品，则该客户还会购买剃须膏，并且置信度为 80%。” 关联挖掘问题可以分解为两个子问题：

- 找到所有称为频繁项集合的项组合，即支持度大于最小支持度的项组合。
- 使用频繁项集合来生成所需要的规则。举例说明规则的生成原理，例如，ABC 和 BC 为频繁项，如果 $\text{support}(ABC)$ 与 $\text{support}(BC)$ 的比例大于等于最小置信度时，则可使用“从规则 A 推导出 BC”。注意：如果 ABCD 为频繁项，该规则将具有最小支持度。ODM 关联仅支持单一后项规则（从 ABC 推导出 D）。

频繁项集合的数量取决于最小支持度参数。生成规则的数量取决于频繁项集合的数量和置信参数。如果置信参数设得过高，则关联模型中可能存在频繁项集合，但不存在规则。

ODM 将基于 SQL 来执行 Apriori 算法。候选生成和支持计数步骤使用 SQL 查询来执行。不使用专门的内存数据结构。SQL 查询将使用各种提示进行优化，以便能在数据库服务器中高效运行。

Apriori 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。

在构建 Apriori 模型之前，需要指定要将哪些字段用作与关联建模有关的项目。

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。选择此选项后，根据是否正在使用交易格式来指定对话框中的剩余字段。

图片 4-23
默认的自定义字段设置



如果没有 使用交易格式，请指定：

- **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。
- **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。 [有关详细信息，请参阅第 4 章中的分区节点中的 IBM SPSS Modeler 15 源、过程和输出节点。](#)

如果正在 使用交易格式，请指定：

使用交易格式。 如果希望将每个项目行中的数据转换为每个观测值行中的数据，请使用此选项。

选择此选项会更改该对话框下半部分中的字段控件：

图片 4-24
事务格式的字段设置

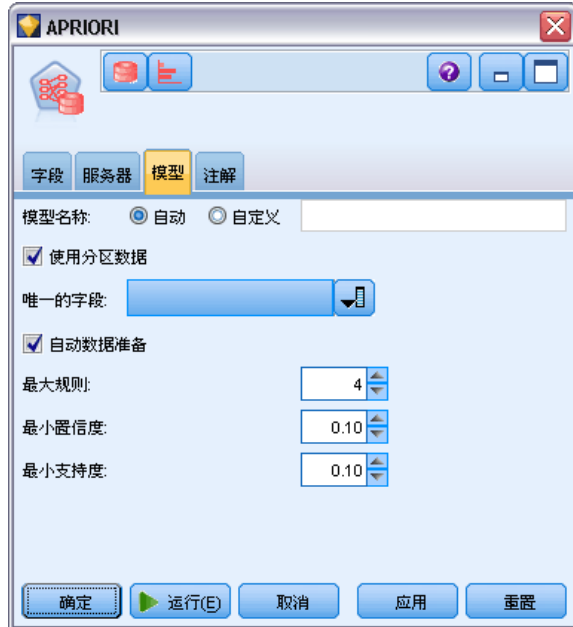


对于事务格式，请指定：

- **ID**。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）。
- **内容**。指定模型的内容字段。该字段包含与关联建模有关的项目。
- **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 15 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。）

Apriori 模型选项

图片 4-25
Apriori “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。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

最大规则长度。 为任何规则设置最大预条件数，该值为从 2 到 20 的整数。这是一种用来限制规则复杂性的方法。如果规则太复杂或者太具体，或者如果您的规则集培训时间太长，请尝试降低此设置。

最小置信度。 设置最小置信级别，该值介于 0 和 1 之间。置信度低于指定标准的规则将被放弃。

最小支持度。 设置最小支持度阈值，该值介于 0 和 1 之间。“先验”发现频率高于最小支持度阈值的模式。

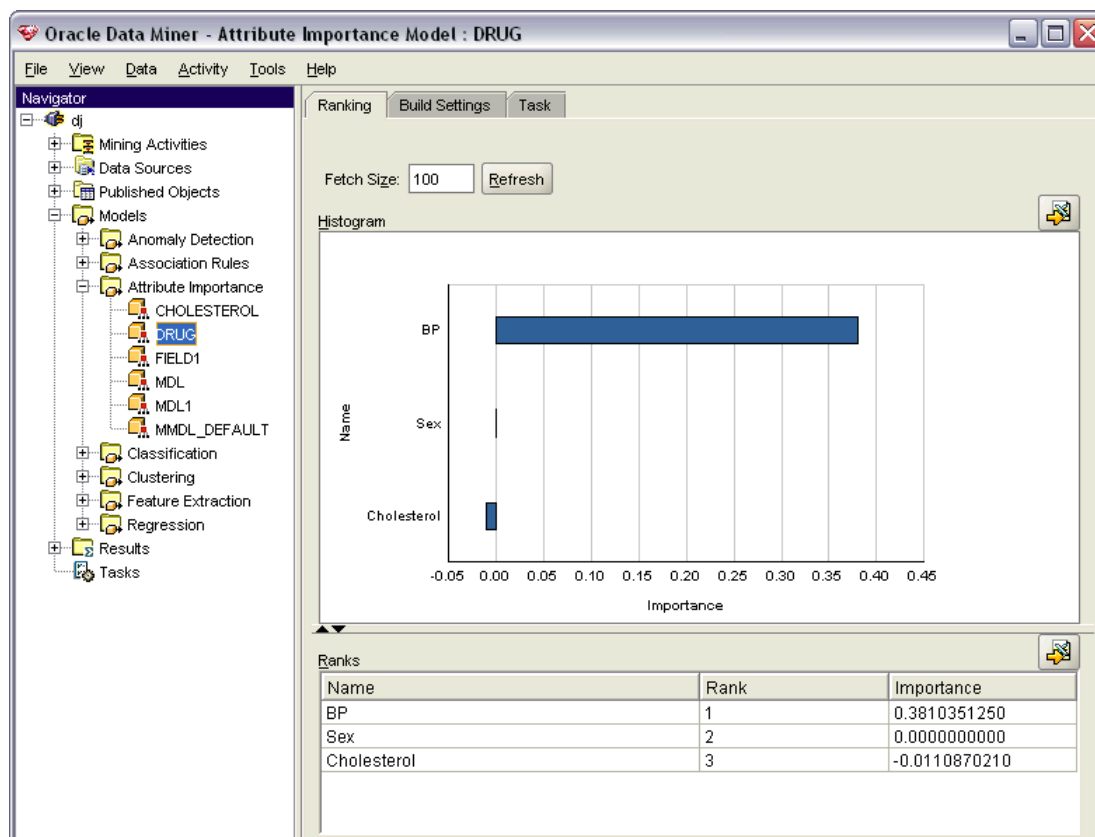
Oracle 最小描述符长度 (MDL)

Oracle 最小描述符长度 (MDL) 算法帮助识别对目标属性影响最大的属性。通常情况下，知道哪个是最有影响的属性可以更好地了解和管理业务并且有助于简化建模操作。另外，这些属性可以指示为扩大模型而希望添加的数据的类型。例如，可使用 MDL 找到与预测已制造部件的质量最相关的工艺属性、与流失相关的因素或最可能用于治疗特定疾病的基因等。

Oracle MDL 丢弃在预测目标时视为不重要的输入字段。然后使用剩余的输入字段构建与 Oracle 模型关联的非精练模型块（在 Oracle Data Miner 中可见）。在 Oracle Data Miner 中浏览模型会显示一张图表，其中显示剩余的输入字段（按照其对预测目标的重要性顺序排序）。

图片 4-26

使用显示输入字段在预测目标中的相对重要性的 Oracle MDL 图表



负秩数指示噪声。秩数为零或更小的输入字段对于预测没有贡献，应从数据中删除。

要显示图表

- ▶ 在“模型”选项板中右键单击非精练模型块并选择浏览。
- ▶ 从模型窗口单击按钮启动 Oracle Data Miner。

- ▶ 连接到 Oracle Data Miner。有关详细信息，请参阅第 84 页码 Oracle Data Miner。
- ▶ 在 Oracle Data Miner 导航面板中，展开模型，然后展开属性重要性。
- ▶ 选择相关的 Oracle 模型（其名称与您在 IBM® SPSS® Modeler 中指定的目标字段名称相同）。如果您不确定哪个正确，选择“属性重要性”文件夹并按照创建日期查找模型。

MDL 模型选项

图片 4-27
MDL “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。

唯一字段。 指定用于作为各个观测值唯一标识的字段。例如，ID 字段，如客户 ID。IBM® SPSS® Modeler 限定此关键字字段必须为数值。

注：除 Oracle Adaptive Bayes、Oracle O-Cluster 和 Oracle Apriori 节点外，此字段对所有 Oracle 节点都是可选的。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

Oracle 属性重要性 (AI)

属性重要性的目标是找出数据集中的哪些属性与结果相关，以及其影响最终结果的程度。Oracle 属性重要性节点分析数据，找出模式，并通过关联置信水平预测结果。

AI 模型选项

图片 4-28
AI 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 15 源、过程和输出节点。

自动数据准备。（仅 11g）启用（默认）或禁用 Oracle Data Mining 的自动数据准备模式。如选中此框，ODM 自动按照算法执行所需的数据转换。有关详细信息，请参阅 Oracle Data Mining 概念。

AI 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。

图片 4-29
AI 选择选项



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。

AI 模型块模型选项卡

Oracle AI 模型块的“模型”选项卡显示所有输入的排序和重要性，并允许您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和目标预测。其他输入字段将被丢弃。默认选择基于建模节点中指定的选项，但可以根据需要选择或取消选择其他字段。

图片 4-30
AI 模型块



- 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。或者，从靠近“排序方式”按钮的列表选择所需项目，并使用上下箭头更改排序方向。
- 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。也可在单击字段时按下 Shift 或 Ctrl 键以扩展选择。有关详细信息，请参阅第 4 章中的按照重要性选择字段中的 IBM SPSS Modeler 15 建模节点。
- 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。

管理 Oracle 模型

Oracle 模型添加到模型选项板的方式与其他 IBM® SPSS® Modeler 模型的添加方式一样，而且使用方法也大致相同。但是，也有几点重大差异，比如 SPSS Modeler 中生成的每个 Oracle 模型实际引用的是存储在数据库服务器上的模型。

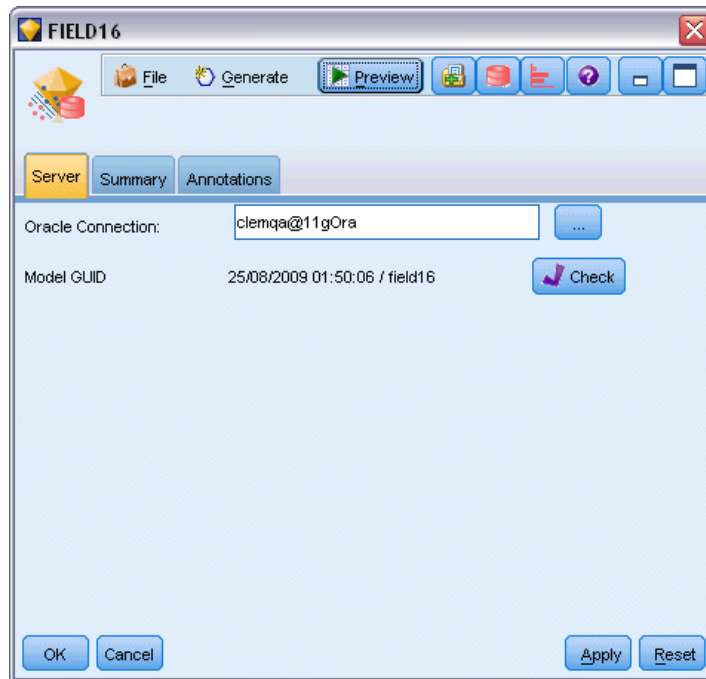
Oracle 模型块服务器选项卡

通过 IBM® SPSS® Modeler 构建 ODM 模型即可在 SPSS Modeler 中创建一个模型，并创建或替代 Oracle 数据库中的一个模型。这种 SPSS Modeler 模型引用数据库服务器上存储的数据库模型的内容。SPSS Modeler 可通过将完全相同的生成**模型关键字**字符串存储在 SPSS Modeler 模型和 Oracle 模型中执行一致性检查。

每个 Oracle 模型的关键字字符串显示在“列出模型”对话框中的模型信息列下。SPSS Modeler 模型的关键字字符串在 SPSS Modeler 模型的“服务器”选项卡上显示为**模型关键字**（放置在流中时）。

模型块“服务器”选项卡上的“检验”按钮，可用于检验 SPSS Modeler 模型中的模型关键字和 Oracle 模型是否匹配。如果 Oracle 中无法找到名称相同的模型，或者模型关键字不匹配，则 Oracle 模型已被删除或在 SPSS Modeler 模型构建后重新构建。

图片 4-31
Oracle 模型块服务器选项卡选项



Oracle 模型块汇总选项卡

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。查看完成后要隐藏结果时，请使用展开控件来折叠想要隐藏的具体结果，或者单击**全部折叠**按钮来折叠所有结果。

分析。 显示指定模型的相关信息。如果已执行附加到该模型块的分析节点，则还会在此部分显示通过分析获得的信息。[有关详细信息，请参阅第 6 章中的分析节点中的 IBM SPSS Modeler 15 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。

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。

Oracle 模型块设置选项卡

模型块上的“设置”选项卡允许您覆盖建模节点上某些选项的设置以达到得分目的。

Oracle 决策树

使用误分类损失。 确定是否在 Oracle 决策树模型中使用误分类损失。 [有关详细信息，请参阅第 54 页码误分类损失。](#)

规则 ID。 如果选择（选中），将规则标识符列添加到 Oracle 决策树模型中。规则标识符在进行特定拆分的树中识别节点。

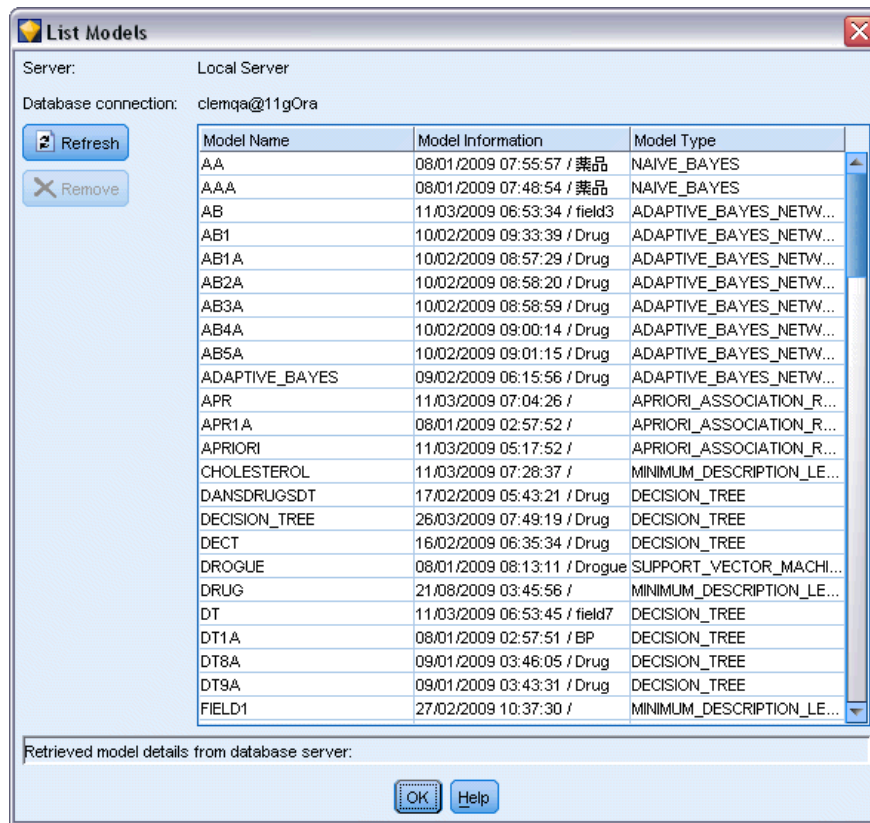
Oracle NMF

显示所有特征。 如果选择（选中），显示所有特征的特征 ID 和置信度，而不是仅在 Oracle NMF 模型中显示最佳特征的特征 ID 和置信度。

列出 Oracle 模型

“列出 Oracle Data Mining 模型”按钮用于启动一个对话框，该对话框列出现有数据库模型并允许删除模型。此对话框可通过“辅助应用程序”对话框启动，也可以通过 ODM 关联节点的构建、浏览和应用对话框启动。

图片 4-32
Oracle “列出模型” 对话框



将显示各模型的如下信息：

- **模型名称。** 模型的名称，用于对列表进行排序
- **模型信息。** 模型的关键信息，比如构建日期/时间和目标列名称
- **模型类型。** 构建此模型的算法的名称

Oracle Data Miner

Oracle Data Miner 是 Oracle Data Mining (ODM) 的用户界面，并替代以前 IBM® SPSS® Modeler 的 ODM 用户界面。Oracle Data Miner 旨在提高分析人员在使用 ODM 算法方面的成功率。该目标通过以下方式来实现：

- 用户在应用能同时处理数据准备和算法选择的方法学方面需要更多帮助。Oracle Data Miner 通过提供可引导应用使用正确方法学的数据挖掘操作，解决此用户需求。
- Oracle Data Miner 为模型构建提供改进的和扩展的试探法，为指定模型和转换设置提供可降低错误几率的转换向导。

定义 Oracle Data Miner 连接

- ▶ Oracle Data Miner 可通过任何版本的 Oracle 进行启动，可通过启动 Oracle Data Miner 按钮应用节点和输出对话框。

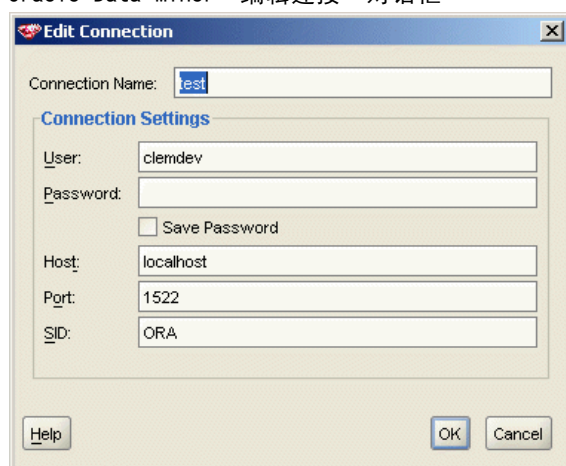
图片 4-33
启动 Oracle Data Miner 按钮



- ▶ 如果正确设置的“辅助应用程序”选项，则 Oracle Data Miner 的**编辑连接**对话框将在 Oracle Data Miner 外部应用程序启动之前显示在用户面前。

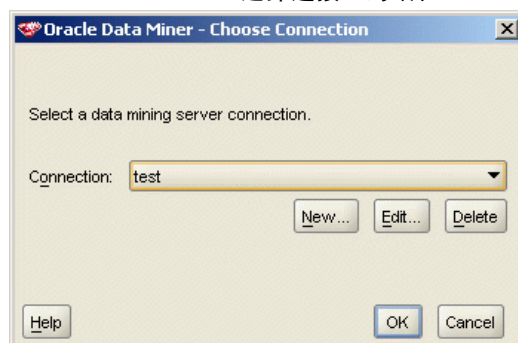
注意：此对话框仅在不存在已定义连接名称时显示。

图片 4-34
Oracle Data Miner “编辑连接”对话框



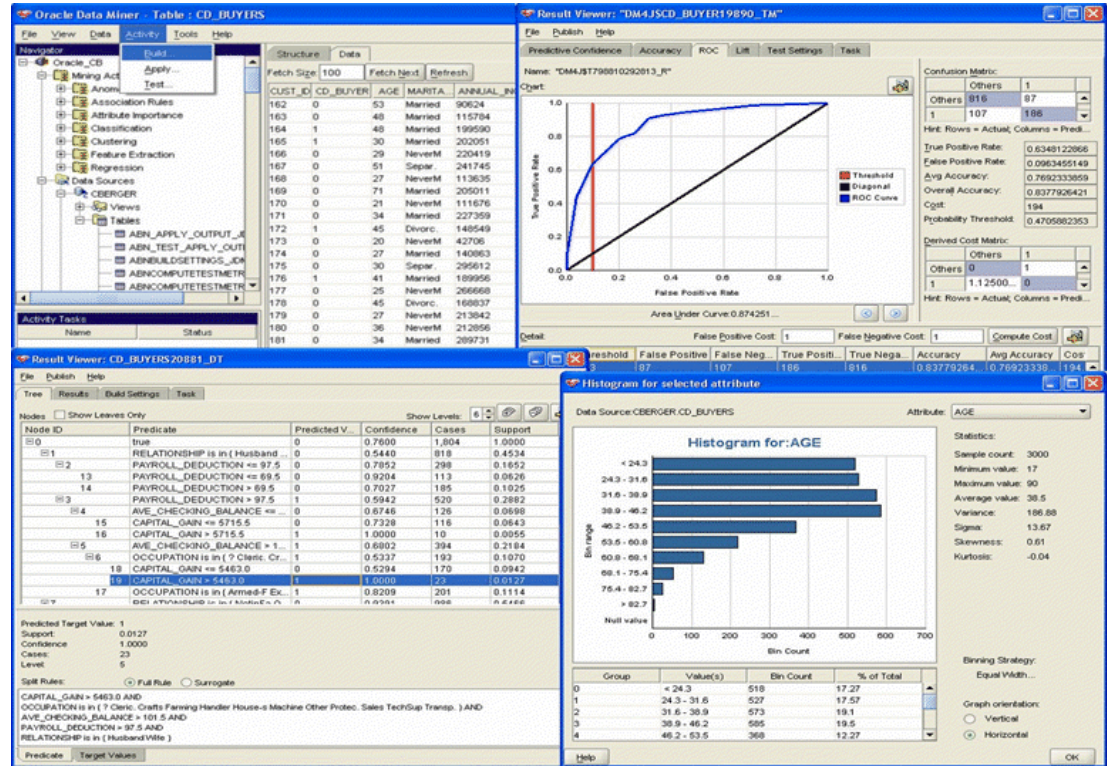
- 提供一个 Data Miner 连接名称并输入对应的 Oracle 10gR1 或 10gR2 服务器信息。Oracle 服务器应与 SPSS Modeler 中指定的服务器一样。
- ▶ Oracle Data Miner 的**选择连接**对话框提供用于指定使用哪个（以上步骤中定义的）连接名称的选项。

图片 4-35
Oracle Data Miner “选择连接”对话框



关于 Oracle Data Miner 需求、安装和使用的详细信息，请参阅 Oracle Web 站点上的 [Oracle Data Miner](http://www.oracle.com/technology/products/bi/odm/odminer/odminer_install_102.htm) (http://www.oracle.com/technology/products/bi/odm/odminer/odminer_install_102.htm)。

图片 4-36 Oracle Data Miner 用户界面



准备数据

使用 Oracle Data Mining 算法的 Naive Bayes、Adaptive Bayes 和 Support Vector Machine 来建模时，可以使用两种类型的数据准备：

- **分类**，对于无法接受连续数据的算法，将连续数字范围字段转换为其可接受的类别。
- **标准化**，也就是应用于数字范围的转换，以使这些数字范围具有类似的均值和标准差。

离散化

IBM® SPSS® Modeler 的“分级”节点提供执行分级操作的若干种方法。将定义可应用于一个或多个字段的分类操作。如在数据集上执行分级操作，则将创建临界值并允许创建 SPSS Modeler 的“派生”节点。“派生”操作可转换为 SQL 并模型构建和评分前被应用。此方法将在模型和执行分类的“派生”节点之间创建依存关系，且允许多个建模任务重复使用分类规范。

标准化

用作 Support Vector Machine 模型的输入值的连续（数字范围）字段应在构建模型前进行标准化。如果是回归模型，则还必须反转标准化，以通过模型输入重建构建评分结构。SVM 模型设置用于选择 Z-Score、最值法 或无。通过 Oracle 构建标准化系数是模型

构建过程中的一个步骤，这些系数将被上传到 SPSS Modeler 并保存在模型中。应用时，这些系数将被转换为 SPSS Modeler 派生表达式，并用于准备（评分时使用的）数据，然后再将数据传输到模型。此情况中，标准化与建模任务紧密关联。

Oracle Data Mining 示例

提供若干样本流，以演示如何在 IBM® SPSS® Modeler 中使用 ODM。这些流位于 \Demos\Database_Modelling\Oracle Data Mining\ 目录下的 SPSS Modeler 安装文件夹中。

注意：可以从 Windows “开始” 菜单 SPSS Modeler 程序组中访问这些 Demos 文件夹。

以下样本流是数据库挖掘过程的示例，通过使用 Oracle Data Mining 提供的 Support Vector Machine (SVM) 算法，依次使用这些样本流。

流	说明
1_upload_data.str	用于净化数据和将数据从平面文件上载到数据库。
2_explore_data.str	提供关于 SPSS Modeler 数据探索的示例
3_build_model.str	采用数据库自有算法构建模型。
4_evaluate_model.str	用作 SPSS Modeler 模型评估的示例
5_deploy_model.str	部署用于数据库内评分的模型。

注：要运行此示例，必须按此顺序执行各个流。此外，必须更新每个流中的源节点和建模节点，以便将想使用的数据库作为有效数据源供您引用。

这些示例流中使用的数据集与信用卡申请有关，演示了同时带有分类和连续预测变量的分类问题。关于此数据集的更多信息，请参阅示例流中同一文件夹下的 crx.names 文件。

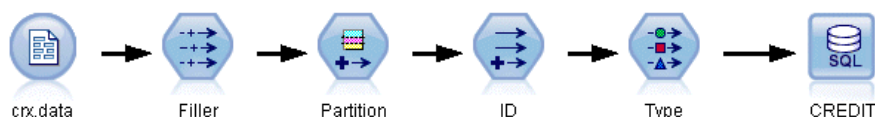
此数据集可从位于

<ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/> 的 UCI Machine Learning Repository 中获得。

示例流：上传数据

第一个示例流 1_upload_data.str 用于清理平面文件中的数据并将其上载到 Oracle。

图片 4-37
用于上传数据的示例流



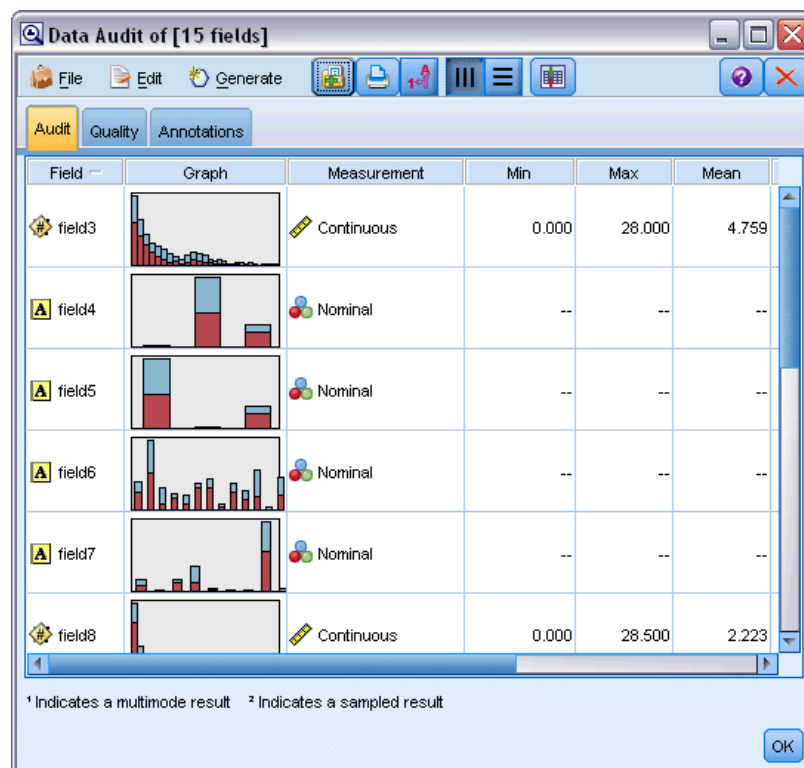
由于 Oracle Data Mining 要求有唯一 ID 字段，因而此初始流通过 IBM® SPSS® Modeler 的 @INDEX 函数，使用“派生”节点来将新字段添加到名称为 ID 的数据集，其唯一值为 1、2 和 3。

“填充”节点用于处理缺失值，并将从文本文件 crx.data 读取的空字段替换为 NULL 值。

示例流：探索数据

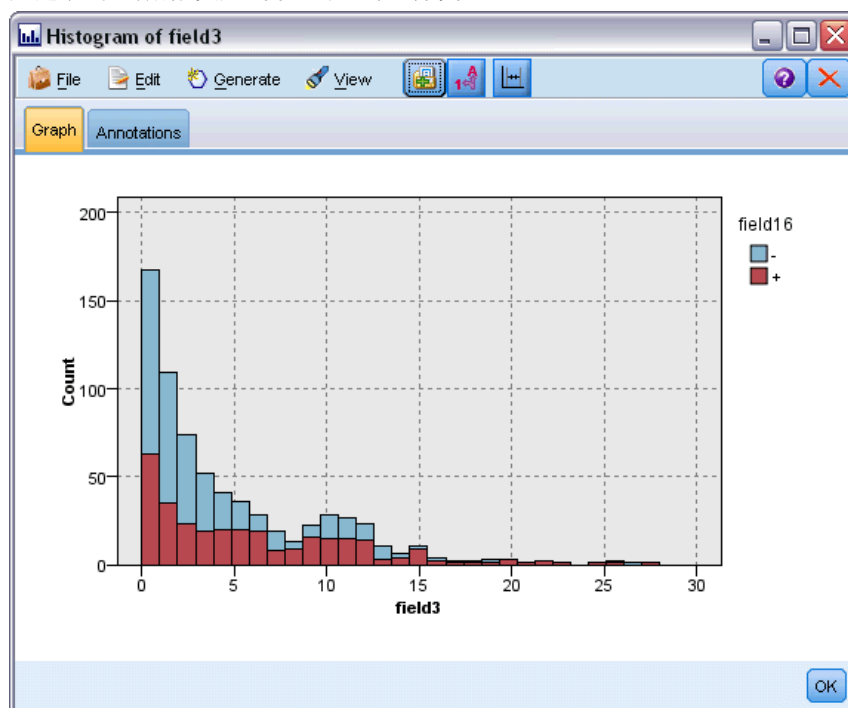
第二个示例流 2_explore_data.str 用于演示如何使用数据审核节点获取数据（包括汇总统计量和图形）的一般概述。有关详细信息，请参阅第 6 章中的数据审核节点中的 IBM SPSS Modeler 15 源、过程和输出节点。

图片 4-38
数据审核结果



双击“数据审核报告”中的图形可显示一个更为详细的图形，用于更深入地探索给定字段。

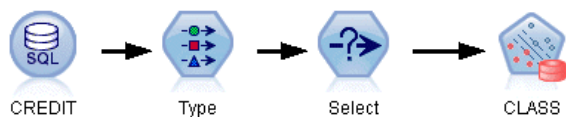
图片 4-39
通过双击“数据审核”窗口创建的直方图



示例流：构建模型

第 3 个示例流，即 3_build_model.str，演示 IBM® SPSS® Modeler 中的“模型”型构建。双击数据库源节点（标注为 CREDIT）以指定数据源。要指定构建设置，请双击构建节点（最初标注为 CLASS，指定数据源后将更改为 FIELD16）。

图片 4-40
数据库建模示例流



在对话框的“模型”选项卡上：

- ▶ 确保选择 ID 作为唯一字段。
- ▶ 确保选择线性作为核函数，选择 Z 得分作为标准化方法。

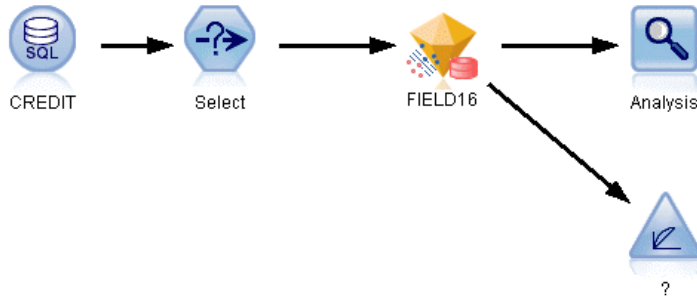
图片 4-41
Oracle SVM 模型选项



示例流：评估模型

第 4 个示例流，即 4_evaluate_model.str，演示构建数据库内模型时使用 IBM® SPSS® Modeler 的优点。一旦执行完模型，即可将它添加回数据流中并使用 SPSS Modeler 提供的多种工具来评估模型。

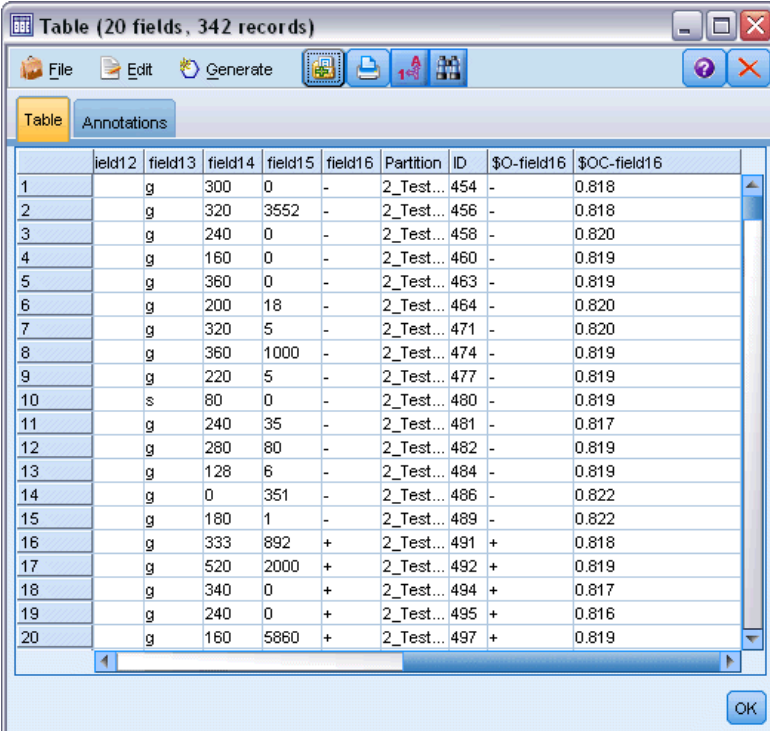
图片 4-42
用于模型评估的示例流



查看建模结果

将表节点附加到模型块以研究结果。\$0-field16 字段显示每个观测值中 field16 的预测值，而 \$0C-field16 字段显示该预测的置信值。

图片 4-43
具有已生成预测相关信息的表

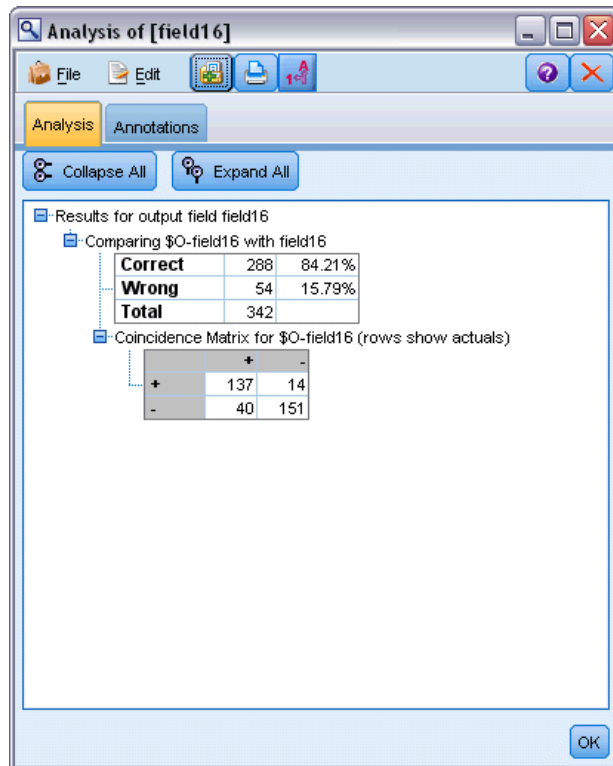


	field12	field13	field14	field15	field16	Partition	ID	\$O-field16	\$OC-field16
1		g	300	0	-	2_Test...	454	-	0.818
2		g	320	3552	-	2_Test...	456	-	0.818
3		g	240	0	-	2_Test...	458	-	0.820
4		g	160	0	-	2_Test...	460	-	0.819
5		g	360	0	-	2_Test...	463	-	0.819
6		g	200	18	-	2_Test...	464	-	0.820
7		g	320	5	-	2_Test...	471	-	0.820
8		g	360	1000	-	2_Test...	474	-	0.819
9		g	220	5	-	2_Test...	477	-	0.819
10		s	80	0	-	2_Test...	480	-	0.819
11		g	240	35	-	2_Test...	481	-	0.817
12		g	280	80	-	2_Test...	482	-	0.819
13		g	128	6	-	2_Test...	484	-	0.819
14		g	0	351	-	2_Test...	486	-	0.822
15		g	180	1	-	2_Test...	489	-	0.822
16		g	333	892	+	2_Test...	491	+	0.818
17		g	520	2000	+	2_Test...	492	+	0.819
18		g	340	0	+	2_Test...	494	+	0.817
19		g	240	0	+	2_Test...	495	+	0.816
20		g	160	5860	+	2_Test...	497	+	0.819

评估建模结果

可以使用“分析”节点创建说明每个预测字段及其目标字段之间的匹配模式的符合矩阵。运行分析节点以查看结果。

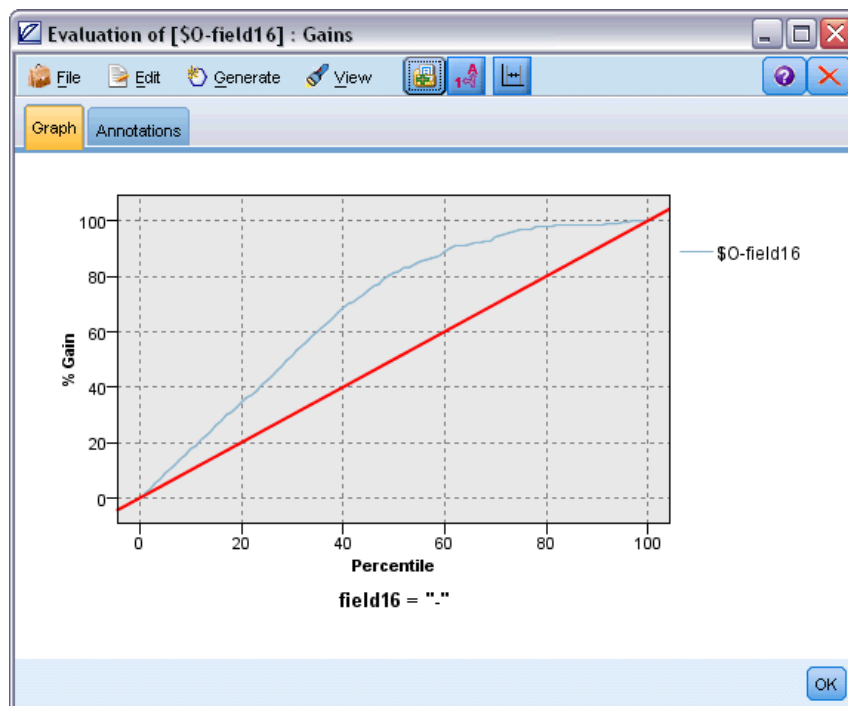
图片 4-44
具有分析结果的有关信息的“分析”选项卡



该表格表明 Oracle SVM 算法生成的预测中 84.21% 是正确的。

您可以使用评估节点创建收益图表，以显示模型对准确率的提高。运行评估节点以查看结果。

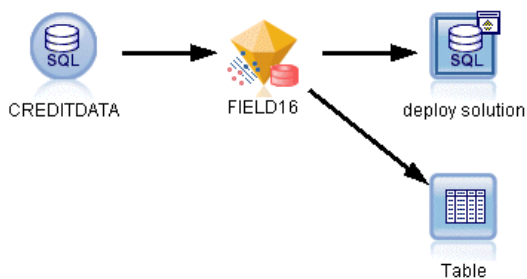
图片 4-45
具有模型准确率提高的相关信息的收益图表



示例流：部署模型

一旦对模型准确率感到满意，即可部署该模型，以用于外部应用程序或往回发布到数据库。在最后一个示例流 5_deploy_model.str 中，将从表 CREDITDATA 中读取数据，然后使用名称为部署解决方案的 Publisher 节点对数据进行评分，并将数据发布到表 CREDITSCORES。

图片 4-46
数据库建模示例流



有关详细信息，请参阅第 2 章中的 IBM SPSS Modeler Solution Publisher 的工作原理中的 IBM SPSS Modeler 15 Solution Publisher。

使用 IBM InfoSphere Warehouse 进行数据库建模

IBM InfoSphere Warehouse 和 IBM SPSS Modeler

IBM InfoSphere Warehouse (ISW) 提供了一系列数据挖掘算法，这些算法嵌入在 IBM DB2 RDBMS 中。IBM® SPSS® Modeler 提供的节点支持与下列 IBM 算法集成：

- 决策树
- 关联规则
- 人口统计聚类
- Kohonen 聚类
- 序列规则
- 变换回归
- 线性回归
- 多项回归
- Naive Bayes
- Logistic 回归
- 时间序列

有关这些算法的详细信息，请参阅 IBM InfoSphere Warehouse 安装随附的文档。

集成 IBM InfoSphere Warehouse 的要求

以下是使用 InfoSphere Warehouse Data Mining 执行数据库内建模的必备条件。您可能需要咨询数据库管理员以确保满足这些条件。

- 在 Windows 或 UNIX 上安装 IBM® SPSS® Modeler Server 后运行 IBM® SPSS® Modeler。
- IBM DB2 Data Warehouse Edition Version 9.1
或
- IBM InfoSphere Warehouse Version 9.5 Enterprise Edition
- 连接如下所述 DB2 的 ODBC 数据源。

注意：数据库建模和 SQL 优化需要在 SPSS Modeler 计算机上启用 SPSS Modeler Server 连接。通过启用此设置，您可以访问数据库算法，直接从 SPSS Modeler 回送 SQL 以及访问 SPSS Modeler Server。要验证当前许可证的状态，请从 SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项服务器启用。

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 15 用户指南](#)。

启用 IBM InfoSphere Warehouse 集成

要启用 IBM InfoSphere Warehouse (ISW) Data Mining 的 IBM® SPSS® Modeler 集成，需要配置 ISW，创建 ODBC 源，启用 SPSS Modeler “辅助应用程序”对话框中的集成功能，并启用 SQL 生成和优化。

配置 ISW

要安装和配置 ISW，按照 InfoSphere Warehouse 安装指南中的说明操作。

为 ISW 创建 ODBC 源

要启用 ISW 和 SPSS Modeler 之间的连接，您需要创建 ODBC 系统数据源名称 (DSN)。

在创建 DSN 之前，您应当对 ODBC 数据源和驱动程序，以及 SPSS Modeler 中的数据库支持有基本的了解。[有关详细信息，请参阅第 2 章中的数据访问中的 IBM SPSS Modeler Server 15 管理和性能指南](#)。

如果 IBM® SPSS® Modeler Server 和 IBM InfoSphere Warehouse Data Mining 运行在不同的主机上，请在每个主机上创建相同的 ODBC DSN。确保对每个主机上的这个 DSN 使用相同的名称。

- ▶ 安装 ODBC 驱动程序。您可在随此版本附带的 IBM® SPSS® Data Access Pack 安装盘上找到这些驱动程序。运行 setup.exe 文件以启动安装程序，并选择所有相关的驱动程序。按照屏幕说明操作以安装驱动程序。
- ▶ 创建 DSN。
 - 注意: 菜单排列顺序取决于 Windows 版本。
 - **Windows XP。** 从“开始”菜单中选择控制面板。双击管理工具，然后双击数据源 (ODBC)。
 - **Windows Vista。** 从“开始”菜单中选择控制面板，然后选择系统维护。双击管理工具，选择数据源 (ODBC)，然后单击打开。
 - **Windows 7。** 从“开始”菜单中选择控制面板，选择系统和安全，然后选择管理工具。选择数据源 (ODBC)，然后单击打开。
- ▶ 单击系统 DSN 选项卡，然后单击添加。
- ▶ 选择 SPSS OEM 6.0 DB2 Wire Protocol 驱动程序。
- ▶ 单击完成。
- ▶ 在“ODBC DB2 Wire Protocol 驱动程序设置”对话框中：
 - 指定数据源名称。
 - 对于 IP 地址，请指定 DB2 RDBMS 所在服务器的主机名。

- 接受默认的 TCP 端口 (50000)。
- 指定要连接的数据库的名称。
- ▶ 单击测试连接。
- ▶ 在“登录 DB2 Wire Protocol”对话框中，输入数据库管理员提供给您用户名和密码，然后单击确定。

此时会显示连接已建立！的消息。

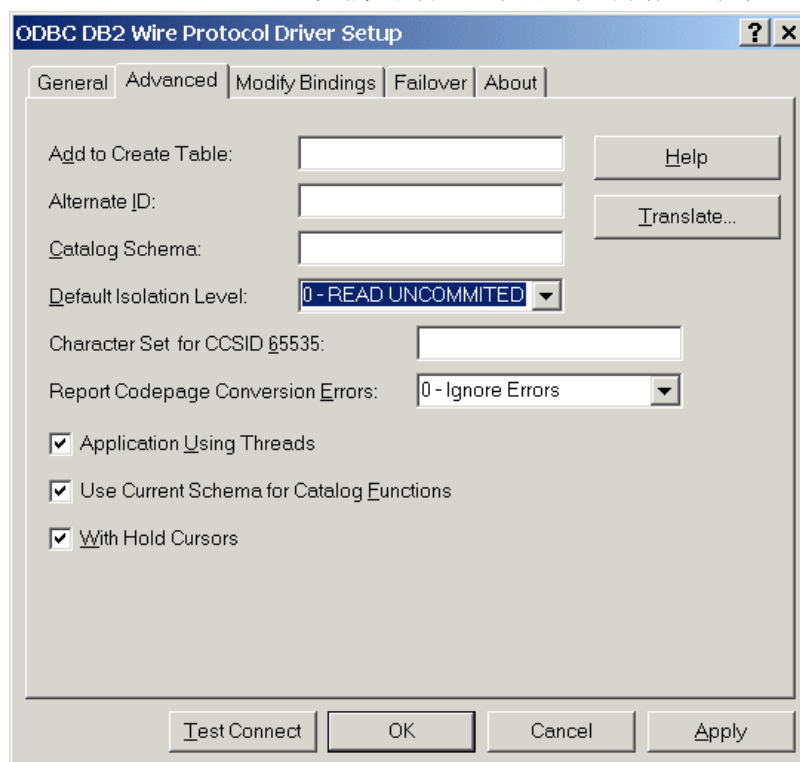
IBM DB2 ODBC 驱动程序。如果您的 ODBC 驱动程序是 IBM DB2 ODBC DRIVER，请按照下列步骤创建 ODBC DSN：

- ▶ 在“ODBC 数据源管理器”中，单击系统 DSN 选项卡，然后单击添加。
 - ▶ 选择 IBM DB2 ODBC DRIVER，然后单击完成。
 - ▶ 在“IBM DB2 ODBC DRIVER—添加”窗口中，输入数据源名称，然后输入数据库别名，单击添加。
 - ▶ 在“CLI/ODBC 设置—<数据源名称>”窗口的“数据源”选项卡上，输入数据库管理员提供给您用户 ID 和密码，然后单击 TCP/IP 选项卡。
 - ▶ 在“TCP/IP”选项卡上，输入：
 - 要连接的数据库的名称。
 - 数据库别名（不超过八个字符）。
 - 要连接的数据库服务器的主机名。
 - 该连接的端口号。
 - ▶ 单击安全选项选项卡，选择指定安全选项（可选），然后接受默认设置（使用服务器的 DBM 配置中的身份验证值）。
 - ▶ 单击数据源选项卡，然后单击连接。
- 此时会显示连接测试成功的消息。

配置用于反馈的 ODBC（可选）

要在模型构建期间从 IBM InfoSphere Warehouse Data Mining 得到反馈，并使得 SPSS Modeler 能够取消模型构建，请按下列步骤配置在上一节中创建的 ODBC 数据源。请注意，此配置步骤使得 SPSS Modeler 能够读取可能未通过并发执行事务处理提交到数据库的 DB2 数据。如果对此更改的涵义有怀疑，请咨询数据库管理员。

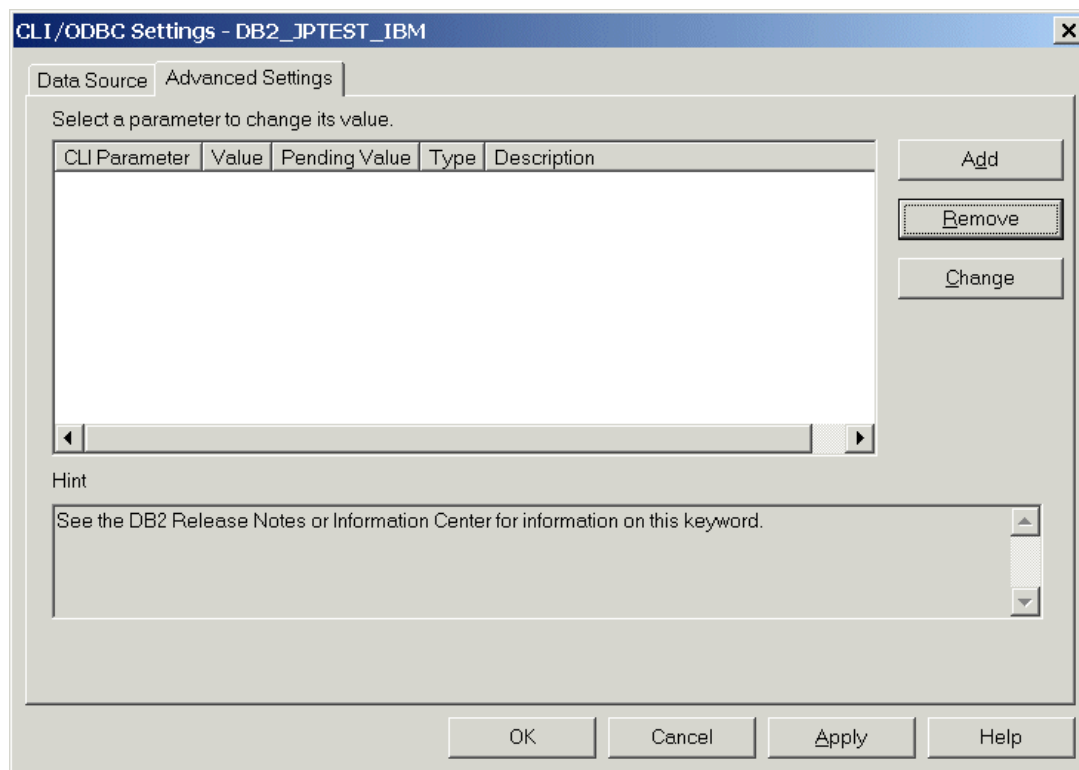
图片 5-1
“ODBC DB2 Wire Protocol 驱动程序设置”对话框的“高级”选项卡



SPSS OEM 6.0 DB2 Wire Protocol 驱动程序。对于 Connect ODBC 驱动程序，请按下列步骤操作：

- ▶ 启动“ODBC 数据源管理器”，选择上一节中创建的数据源，然后单击配置按钮。
- ▶ 在“ODBC DB2 Wire Protocol 驱动程序设置”对话框中，单击高级选项卡。
- ▶ 将默认隔离级别设置为 0-READ UNCOMMITTED，然后单击确定。

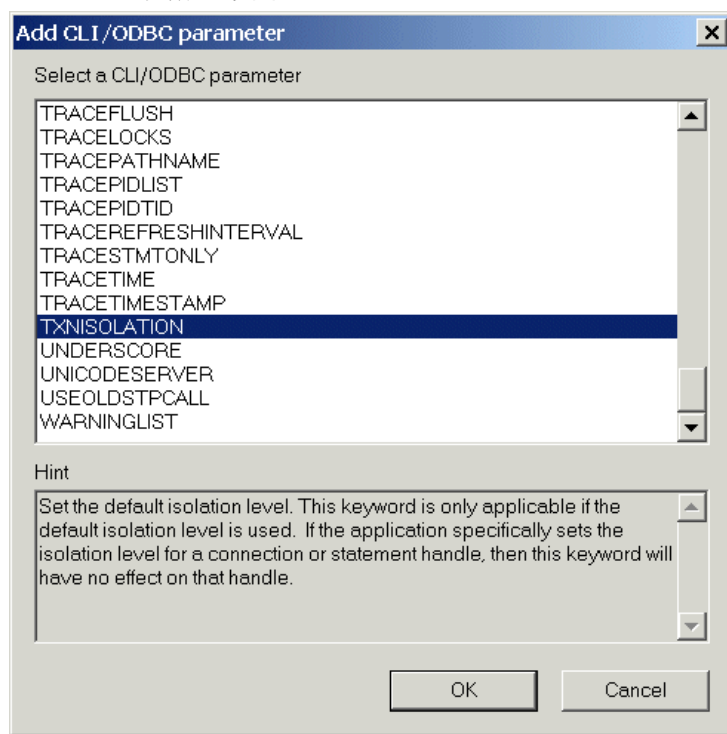
图片 5-2
“CLI/ODBC 设置”对话框的“高级设置”选项卡



IBM DB2 ODBC 驱动程序。对于 IBM DB2 驱动程序，请按下列步骤操作：

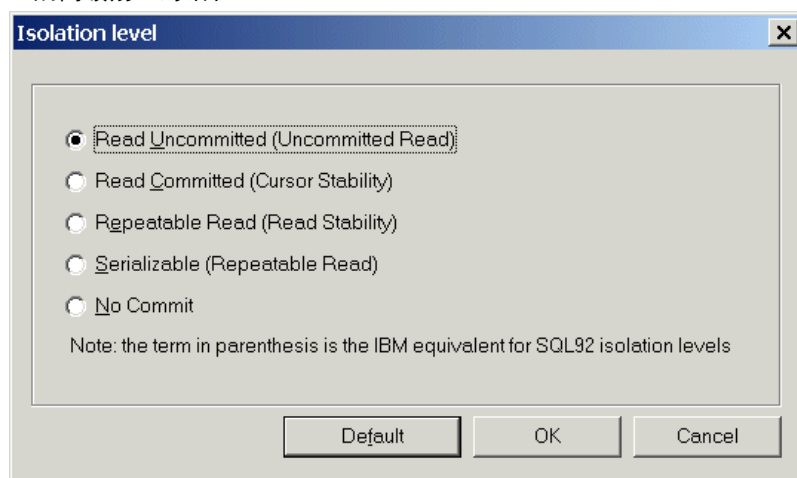
- ▶ 启动“ODBC 数据源管理器”，选择上一节中创建的数据源，然后单击配置按钮。
- ▶ 在“CLI/ODBC 设置”对话框中，单击高级设置选项卡，然后单击添加按钮。

图片 5-3
“CLI/ODBC 参数”对话框



- ▶ 在“添加 CLI/ODBC 参数”对话框中，选择参数 TXNISOLATION，然后单击确定。

图片 5-4
“隔离级别”对话框



- ▶ 在“隔离级别”对话框中，选择 Read Uncommitted，然后单击确定。
- ▶ 在“CLI/ODBC 设置”对话框中，单击确定完成配置。

请注意，IBM InfoSphere Warehouse Data Mining 报告的反馈显示为下列格式：

<ITERATIONNO> / <PROGRESS> / <KERNELPHASE>

其中：

- <ITERATIONNO> 表示数据的当前处理次数，从 1 开始。
- <PROGRESS> 表示当前迭代的进度，为 0.0 到 1.0 之间的数字。
- <KERNELPHASE> 描述挖掘算法当前所处的阶段。

在 IBM SPSS Modeler 中启用 IBM InfoSphere Warehouse Data Mining 集成

要使 SPSS Modeler 能够将 DB2 用于 IBM InfoSphere Warehouse Data Mining，首先必须在“辅助应用程序”对话框中指定一些选项。

- ▶ 从 SPSS Modeler 菜单中选择：
工具 > 选项 > 辅助应用程序

- ▶ 单击 IBM InfoSphere Warehouse 选项卡。

启用 InfoSphere Warehouse Data Mining 集成。 启用 SPSS Modeler 窗口底部的“数据库建模”选项板（如尚未显示）并添加 ISW 数据挖掘算法的建模节点。

DB2 连接。 指定用于构建和存储模型的默认 DB2 ODBC 数据源。在单个模型建模和生成的模型节点上，此设置可以被覆盖。单击省略号 (...) 按钮，选择数据源。

用于建模的数据库连接可以与用于访问数据的连接相同，也可以不相同。例如，可以有这样一个流，该流访问一个 DB2 数据库中的数据，将数据下载到 SPSS Modeler 进行清理或其他操作，然后再将数据上载到另一个 DB2 数据库用于建模。另外，原始数据可能会驻留在平面文件或其他（非 DB2）源中，这种情况下需要将数据上载到 DB2 用于建模。在所有情况下，如果需要，数据都将自动上载到在数据库中创建的、用于建模的一个临时表中。

覆盖 InfoSphere Warehouse Data Mining 集成模型时发出警告。 选中此选项，可确保只有在发出警告的情况下，数据库中存储的模型才会被 SPSS Modeler 覆盖。

列出 InfoSphere Warehouse Data Mining 模型。 通过此选项可列出和删除存储在 DB2 中的模型。 [有关详细信息，请参阅第 103 页码列出数据库模型。](#)

启动 InfoSphere Warehouse Data Mining 可视化。 如果安装了可视化模块，则必须在此处启用才能用于 SPSS Modeler。

可视化的可执行文件路径。 可视化模块的可执行文件（如已安装）的位置，例如 C:\Program Files\IBM\ISWarehouse\Im\IMVisualization\bin\imvisualizer.exe。

时间序列可视化插件目录。 时间序列可视化 Flash

插件（如已安装）的位置，例如 C:\Program

Files\IBM\ISWShared\plugins\com.ibm.datatools.datamining.imvisualization.flash_2.2.1.v2009111

启用 InfoSphere Warehouse Data Mining Power 选项。 您可以针对数据库内挖掘算法设置内存使用限制，以命令行形式为特定模型指定其他任意选项。通过内存限制，可以控制内存使用量，为功能选项 -buf 指定值。其他功能选项可以在此处以命令行形式指定，并传递到 IBM InfoSphere Warehouse Data Mining。 [有关详细信息，请参阅第 106 页码功能选项。](#)

检查 InfoSphere Warehouse 版本。 检查当前使用的 IBM InfoSphere Warehouse 的版本，如果您尝试使用当前版本不支持的数据挖掘功能，则报告错误。

启用 SQL 生成和优化

- ▶ 从 SPSS Modeler 菜单中选择：
工具 > 流属性 > 选项

图片 5-5
优化设置



- ▶ 在导航窗格中单击优化选项。
- ▶ 确认是否已启用生成 SQL 选项。要使数据库建模正常发挥作用，此设置是必需的。
- ▶ 选中优化 SQL 生成和优化其他执行（非严格必需但强烈推荐使用，以使性能更优）。

有关详细信息，请参阅第 5 章中的设置流的优化选项中的 IBM SPSS Modeler 15 用户指南。

使用 IBM InfoSphere Warehouse Data Mining 构建模型

IBM InfoSphere Warehouse Data Mining 模型构建需要将训练数据集位于 DB2 数据库的表或视图中。如果数据不在 DB2 中，或需要作为数据准备过程（该过程无法在 DB2 中进行）的一部分在 IBM® SPSS® Modeler 中进行处理，则该数据会在模型构建之前自动上载到 DB2 的一个临时表中。

模型评分和部署

模型评分始终发生在 DB2 内部，且始终由 IBM InfoSphere Warehouse Data Mining 执行。如果数据源自或需要在 IBM® SPSS® Modeler 内准备，则可能需要将数据集上载到临时表。对于 SPSS Modeler 中的决策树、回归和聚类模型，一般只提供一个预测以及关联的概率或置信度。此外，用于显示每个可能结果的置信度的用户选项（与 logistic 回归中的选项类似）是模型块“设置”选项卡上可用的得分时间选项（包括所有类的置信度复选框）。对于 SPSS Modeler 中的关联模块和序列模块，有多个值可供选择。SPSS Modeler 可以对 IBM InfoSphere Warehouse Data Mining 模型进行评分，该模型来自使用 IBM® SPSS® Modeler Solution Publisher 进行发布以供执行的流。

下列字段是由评分模型生成的：

表 5-1
模型评分字段

模型类型	得分列	含义
决策树	\$I-field	field的最佳预测。
	\$IC-field	field的最佳预测的置信度。
	\$IC-value1, ..., \$IC-valueN	（可选）fieldN 个可能值中每一个可能值的置信度。
回归	\$I-field	field的最佳预测。
	\$IC-field	field的最佳预测的置信度。
聚类	\$I-model_name	输入记录的最佳聚类分配。
	\$IC-model_name	输入记录的最佳聚类分配的置信度。
关联	\$I-model_name	匹配规则的标识符。
	\$IH-model_name	头项目。
	\$IHN-model_name	头项目的名称。
	\$IS-model_name	匹配规则的支持值。
	\$IC-model_name	匹配规则的置信度值。
	\$IL-model_name	匹配规则的提升值。
	\$IMB-model_name	匹配主体项目或主体项目集的数目（由于所有主体项目或主体项目集都必须与此数值相匹配，因此该值与主体项目或主体项目集的数量相等）。

模型类型	得分列	含义
序列	\$I-model_name	匹配规则的标识符
	\$IH-model_name	匹配规则的头项目集
	\$IHN-model_name	匹配规则的头项目集中的项目名称
	\$IS-model_name	匹配规则的支持值
	\$IC-model_name	匹配规则的置信度值
	\$IL-model_name	匹配规则的提升值
	\$IMB-model_name	匹配主体项目或主体项目集的数目（由于所有主体项目或主体项目集都必须与此数值相匹配，因此该值与主体项目或主体项目集的数量相等）
Naive Bayes	\$I-field	field的最佳预测。
	\$IC-field	field的最佳预测的置信度。
Logistic 回归	\$I-field	field的最佳预测。
	\$IC-field	field的最佳预测的置信度。

管理 DB2 模型

通过 IBM® SPSS® Modeler 构建 IBM InfoSphere Warehouse Data Mining 模型会在 SPSS Modeler 中创建一个模型，然后在 DB2 数据库中创建或替换一个模型。这种 SPSS Modeler 模型可引用数据库服务器上存储的数据库模型的内容。SPSS Modeler 可通过将完全相同的生成模型关键字字符串存储在 SPSS Modeler 模型和 DB2 模型中执行一致性检查。

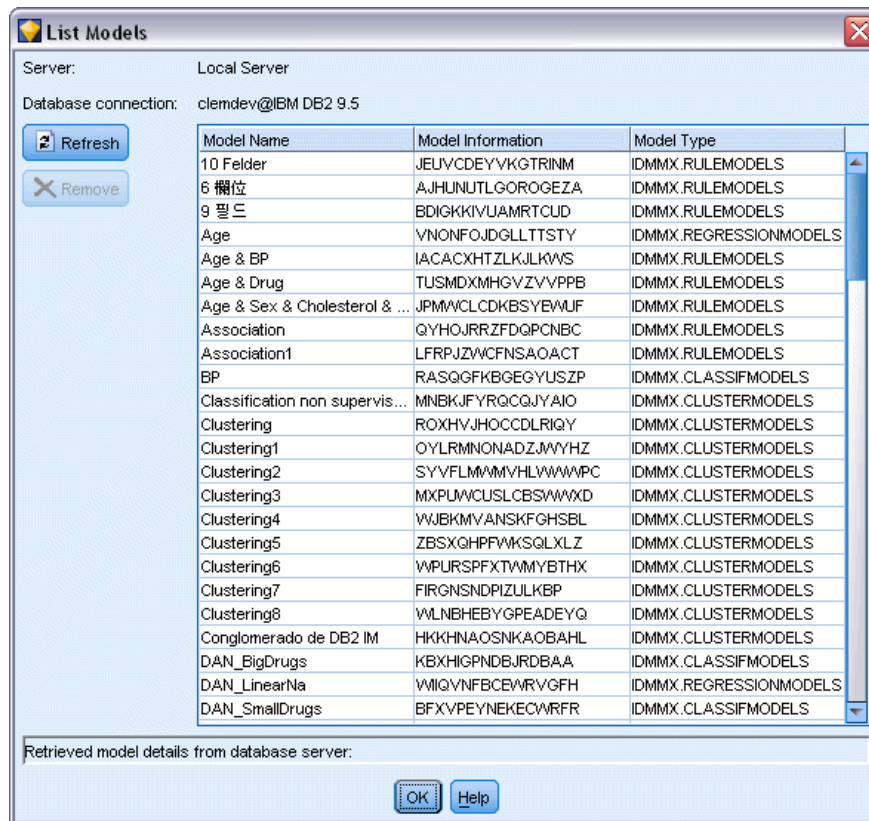
每个 DB2 模型的关键字字符串都显示在“列出数据库模型”对话框的模型信息列下面。SPSS Modeler 模型的关键字字符串显示为 SPSS Modeler 模型“服务器”选项卡上的“模型关键字”（放入流中时显示）。

“检查”按钮可用于检查 SPSS Modeler 模型和 DB2 模型中的模型关键字是否匹配。如果在 DB2 中找不到具有相同名称的模型，或模型关键字不匹配，则说明构建了 SPSS Modeler 模型之后删除或重新构建了该 DB2 模型。[有关详细信息，请参阅第 133 页码 ISW 模型块服务器选项卡。](#)

列出数据库模型

IBM® SPSS® Modeler 提供了一个对话框用于列出存储在 IBM InfoSphere Warehouse Data Mining 中的模型，并允许删除模型。

图片 5-6
DB2 “列出模型” 对话框



此对话框可从 IBM “辅助应用程序” 对话框以及 IBM InfoSphere Warehouse Data Mining 相关节点的构建、浏览和应用对话框中访问。将显示各模型的如下信息：

- 模型名称（模型的名称，用于对列表进行排序）。
- 模型信息（SPSS Modeler 构建模型时生成的随机关键字中的模型关键字信息）。
- 模型类型（IBM InfoSphere Warehouse Data Mining 存储该模型的 DB2 表）。

浏览模型

可视化工具是浏览 InfoSphere Warehouse Data Mining 模型的唯一方法。该工具可选择与 InfoSphere Warehouse Data Mining 一起安装。[有关详细信息，请参阅第 95 页码启用 IBM InfoSphere Warehouse 集成。](#)

- 单击**查看**可启动该可视化工具。该工具显示的内容取决于生成的节点类型。例如，从 ISW 决策树模型块启动时，该可视化工具将返回一个预测类视图。
- 单击**测试结果**（仅用于决策树和序列）可启动该可视化工具，查看所生成模型的整体质量。

导出模型和生成节点

您可以对 IBM InfoSphere Warehouse Data Mining 模型执行 PMML 导入和导出操作。导出的 PMML 是 IBM InfoSphere Warehouse Data Mining 生成的原始 PMML。该导出功能以 PMML 格式返回模型。

可以将模型汇总和结构导出到文本文件和 HTML 文件。需要时还可以生成相应的过滤、选择和导出节点。有关更多信息，请参阅《IBM® SPSS® Modeler 用户指南》中的“导出模型”。

对所有算法通用的节点设置

以下设置通用于许多 IBM InfoSphere Warehouse Data Mining 算法：

目标和预测变量。您可以通过使用类型节点指定目标和预测变量，也可以通过使用模型构建器节点的“字段”选项卡手动指定目标和预测变量，这些在 IBM® SPSS® Modeler 中都是标准设置。

ODBC 数据源。此设置使得用户能够覆盖当前模型的默认 ODBC 数据源。（该默认值在“辅助应用程序”对话框中指定。） [有关详细信息，请参阅第 95 页码启用 IBM InfoSphere Warehouse 集成。](#)

ISW 服务器选项卡选项

您可以指定用于上载建模数据的 DB2 连接。如果需要，您可以在“服务器”选项卡上为每个建模节点都选择一个连接，以覆盖在“辅助应用程序”对话框中指定的默认 DB2 连接。 [有关详细信息，请参阅第 95 页码启用 IBM InfoSphere Warehouse 集成。](#)

图片 5-7
ISW 服务器选项卡



用于建模的连接可以与流的源节点中使用的连接相同，也可以不相同。例如，可以有这样一个流，该流访问一个 DB2 数据库中的数据，将数据下载到 IBM® SPSS® Modeler 进行清理或其他操作，然后再将数据上载到另一个 DB2 数据库用于建模。

ODBC 数据源名称会有效地嵌入到每个 SPSS Modeler 流中。如果在一个主机上创建的流在另一个主机上执行，则该数据源在两个主机上的名称必须相同。此外，也可以在各个源或建模节点中的“服务器”选项卡上选择另一个数据源。

通过使用下列选项，可以在构建模型时获得反馈：

- **启用反馈。**选择此选项可在模型构建期间获得反馈（默认为关闭状态）。
- **反馈间隔（以秒数记）。**指定模型构建期间 SPSS Modeler 检索反馈的频率。

启用 InfoSphere Warehouse Data Mining Power 选项。选择此选项以启用功能选项按钮，此按钮允许您指定大量高级选项，如内存限制和自定义 SQL。[有关详细信息，请参阅第 106 页码功能选项。](#)

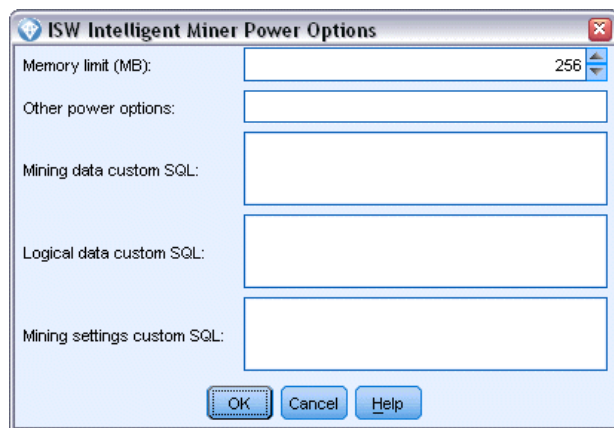
生成节点的“服务器”选项卡包括一个选项，该选项可通过将完全相同的生成模型关键字字符串存储在 SPSS Modeler 模型和 DB2 模型中执行一致性检查。[有关详细信息，请参阅第 133 页码 ISW 模型块服务器选项卡。](#)

功能选项

所有算法的“服务器”选项卡都包括一个用于启用 ISW 建模功能选项的复选框。当您单击功能选项按钮时，将显示“ISW 功能选项”对话框，其中包括下列选项：

- 内存限制。
- 其它功能选项。
- Mining data custom SQL。
- Logical data custom SQL。
- Mining settings custom SQL。

图片 5-8
ISW 功能选项设置



内存限制。限制模型构建算法的内存使用。请注意，标准功能选项会对分类数据中离散值的数量设置限制。

其它功能选项。用于以命令行形式为特定模型或解决方案指定任意功能选项。这些特定内容可能因实施或解决方案的不同而有所差别。用户可以手动扩展 IBM® SPSS® Modeler 生成的 SQL 以定义模型构建任务。

Mining data custom SQL。用户可添加方法调用以修改 `DM_MiningData` 对象。例如，输入下列 SQL 会向构建模型时使用的数据中添加一个基于名为 `Partition` 的过滤器的过滤器：

```
..DM_setWhereClause('Partition' = 1')
```

Logical data custom SQL。用户可添加方法调用以修改 `DM_LogicalDataSpec` 对象。例如，下列 SQL 会从用于模型构建的字段集中删除一个字段：

```
..DM_remDataSpecFld('field6')
```

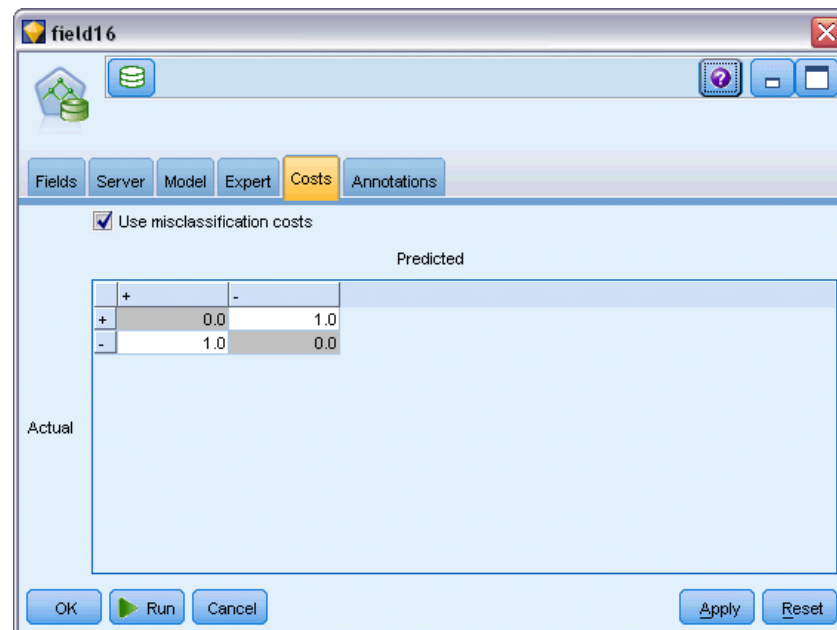
Mining settings custom SQL。用户可添加方法调用以修改 `DM_GlasSettings/DM_RuleSettings/DM_GlusSettings/DM_RegSettings` 对象。例如，输入下列 SQL 会命令 IBM InfoSphere Warehouse Data Mining 将字段 `Partition` 设置为活动状态（这就意味着它总是包括在生成的模型中）：

```
..DM_setFldUsageType('Partition',1)
```

ISW 成本选项

在“成本”选项卡上，您可以调整误分类成本，从而使您能够指定各种不同预测错误的相对重要性。

图片 5-9
ISW 成本选项卡



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。

ISW 决策树

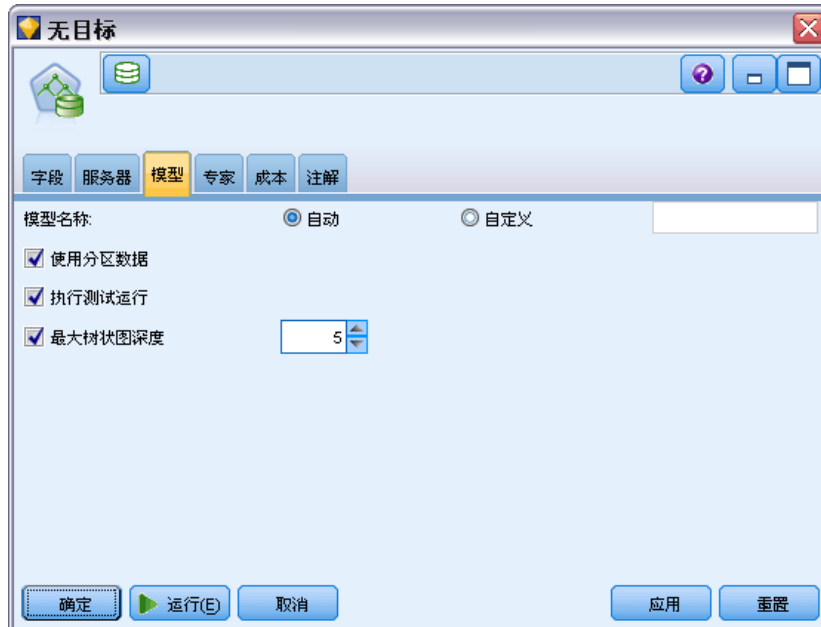
决策树模型允许您开发分类系统，此分类系统可以基于一组决策规则来预测或分类未来的观测值。如果将数据分成您关注的类别（例如，高风险和低风险贷款、用户和非用户、投票人和非投票人或细菌类型），则您可以使用自己的数据来构建规则，借此对新案例或旧案例进行准确性最大的分类。例如，可以基于年龄和其他因素构建对信用风险或购买意向进行分类的树。

ISW 决策树算法会根据分类输入数据构建分类树。生成的决策树是二元性的。可以应用各种不同的设置（包括误分类成本）来构建模型。

ISW 可视化工具是浏览 IBM InfoSphere Warehouse Data Mining 模型的唯一方法。

ISW 决策树模型选项

图片 5-10
ISW 决策树节点模型选项卡



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。 如果定义分区字段，请选择使用分区数据。

执行测试运行。 您可以选择执行测试运行。在训练分区上构好模型之后，执行 IBM InfoSphere Warehouse Data Mining 测试运行。这一步针对测试分区执行一次传递，以建立模型质量信息、提升图表等。

最大树深度。 您可以指定最大树深度。此设置会将树的深度限制为指定的级数。如果不选择此选项，则不强制限制。为避免模型过于复杂，建议一般不要选择大于 5 的值。

ISW 决策树专家选项

图片 5-11
ISW 决策树节点“专家”选项卡



最高纯度。此选项会设置内部节点的最高纯度。如果分割节点会导致其中一个子节点超过指定的纯度测量值（例如，超过 90% 的观测值属于某个指定类别），则该节点不会分割。

每内部节点的最小观测值。如果分割某个节点会生成观测值数少于指定最小值的节点，将不会分割该节点。

ISW 关联

您可使用 ISW 关关节点查找存在于组集合中的项目间的关联规则。关联规则会将某个特定的结论（如某个特定产品的购买）与一系列条件（如若干其他产品的购买）相关联。

您可通过指定**约束**选择从模型包含或排除关联规则。如果您选择包含某个特定的输入字段，则会在模型中包括含有至少一个指定项目的关联规则。如果您排除了一个输入字段，则会在结果中丢弃包含任一指定项目的关联规则。

ISW 关联和序列算法可使用**分类法**。分类法将单个值映射到更高级别的概念。例如，钢笔和铅笔可以映射到文具类别。

关联规则具有一个后项（结论）和多个前项（条件集合）。示例如下：

[面包, 果酱] ⇨ [黄油]

[Bread, Jam]
⇨ [Margarine]

这里，Bread 和 Jam 是前项（也称为**规则体**）且 Butter 或 Margarine 各是后项的一个示例（也称为**规则头**）。第一个规则表示购买面包和果酱的人还会同时购买黄油。第二个规则标识在商店购物期间购买相同组合（面包和果酱）的同时还会购买人造黄油的客户。

可视化工具是浏览 IBM InfoSphere Warehouse Data Mining 模型的唯一方法。

ISW 关联字段选项

在“字段”选项卡中可指定将用于构建模型的字段。

图片 5-12
ISW 关联节点“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。默认设置为使用类型节点来选择输入和目标字段，在此选项卡中仅可更改非交易数据的表布局设置。

使用类型节点设置。 该选项指定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。 该选项指定使用在此处输入的字段信息，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。

使用交易格式。 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 12 章中的表格格式数据与事务处理格式数据中的 IBM SPSS Modeler 15 建模节点。

- **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）。
- **内容**。指定模型的一个或多个内容字段。这些字段包含与关联建模有关的项目。如果数据为交易格式，则可指定一个名义字段。

使用表格格式。 如果源数据为表格格式，则取消选中使用交易格式复选框。

- **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 15 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。）

用于非交易数据的表布局。 对于表格数据，可选择标准表布局（默认）或限制项目长度布局。

在默认布局中，列数由关联项总数决定。

表 5-2
默认表布局

组 ID	支票帐户	储蓄帐户	信用卡	贷款	保管帐户
Smith	Y	Y	Y	-	-
Jackson	Y	-	Y	Y	Y
Douglas	Y	-	-	-	Y

在限制项目长度布局中，列数由所有行中的最大关联项数决定。

表 5-3
限制项目长度表布局

组 ID	Item1	Item2	Item3	Item4
Smith	支票帐户	储蓄帐户	信用卡	-
Jackson	支票帐户	信用卡	贷款	保管帐户
Douglas	支票帐户	保管帐户	-	-

ISW 关联模型选项

图片 5-13
ISW 关联节点“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 15 源、过程和输出节点。

最低规则支持度 (%)。 关联规则或序列规则的最低支持水平。只有达到至少该支持水平的规则才会包含在模型中。结果计算为 $A/B*100$ ，其中 A 是包含在规则中出现的所有项目的组的数量，B 是所考虑的所有组的总数。如果您要关注更常见的关联或序列，请增加此设置。

最低规则置信度 (%)。 关联规则或序列规则的最低置信水平。只有达到至少该置信水平的规则才会包含在模型中。值计算为 $m/n*100$ ，其中 m 是包含连接规则头（后项）和规则体（前项）的组的数量，n 是包含规则体的组的数量。如果您获得的关联或序列太多或者不是非常相关，请尝试增加此设置。如果您获得的关联或序列太少，请尝试降低此设置。

最大规则大小。 规则中允许的最大项数，其中包括后项。如果相关关联或序列相对较短，则可以降低此设置，以加快集的构建速度。

注意：只对具有交易输入格式的节点进行评分；真值表（表格数据）格式保持非精炼。

ISW 关联专家选项

在关联节点的“专家”选项卡上，您可以指定要在结果中包含的关联规则或从结果中排除的关联规则。如果您想包含指定项目，则在模型中包括含有至少一个指定项目的规则。如果您想排除指定项目，则从结果丢弃含有任一指定项目的规则。

图片 5-14
ISW 关联节点“专家”选项卡



选择了使用项目约束之后，您添加到约束列表的任何项目将根据约束类型的设置从结果包含或排除。

约束类型。 选择您是从结果包含还是排除含有指定项目的关联规则。

编辑约束。 要添加项目到约束项目列表，在项目列表中选中并单击右箭头按钮。

ISW 分类法选项

ISW 关联和序列算法可使用**分类法**。分类法将单个值映射到更高级别的概念。例如，钢笔和铅笔可以映射到文具类别。

在“分类法”选项卡上，您可以定义类别映射，以表示数据中的分类。例如，某种分类法可能会创建两个类别（**Staple** 和 **Luxury**），然后为每个类别指定基本项目。例如，**wine** 分配给 **Luxury**，**bread** 分配给 **Staple**。该分类法具有如下父子结构：

子	父
wine	奢侈品
bread	常用品

有了此分类法，您可以构建一个关联或序列模型，其中包括涉及类别以及基本项目的规则。

注意：要激活此选项卡上的选项，源数据必须为交易格式，且您必须在字段选项卡上选择使用交易格式，然后选择此选项卡上的使用分类法。有关详细信息，请参阅第 12 章中的表格格式数据与事务处理格式数据中的 IBM SPSS Modeler 15 建模节点。

图片 5-15
ISW 关联节点“分类法”选项卡



表名称。此选项指定用于存储分类法详细信息的 DB2 表的名称。

子列。此选项指定分类法表中子列的名称。该子列包含项目名或类别名。

父列。此选项指定分类法表中父列的名称。该父列包含类别名。

向表格加载详细信息。如果 IBM® SPSS® Modeler 中存储的分类信息应在构建模型时上载到分类法表中，请选择此选项。请注意，如果分类法表已经存在，则会将其丢弃。分类法信息使用模型构建节点进行存储，使用“编辑类别”和“编辑分类法”按钮进行编辑。

类别编辑器

通过“编辑类别”对话框，可以向排序列表中添加类别以及从中删除类别。

图片 5-16
分类法类别编辑器



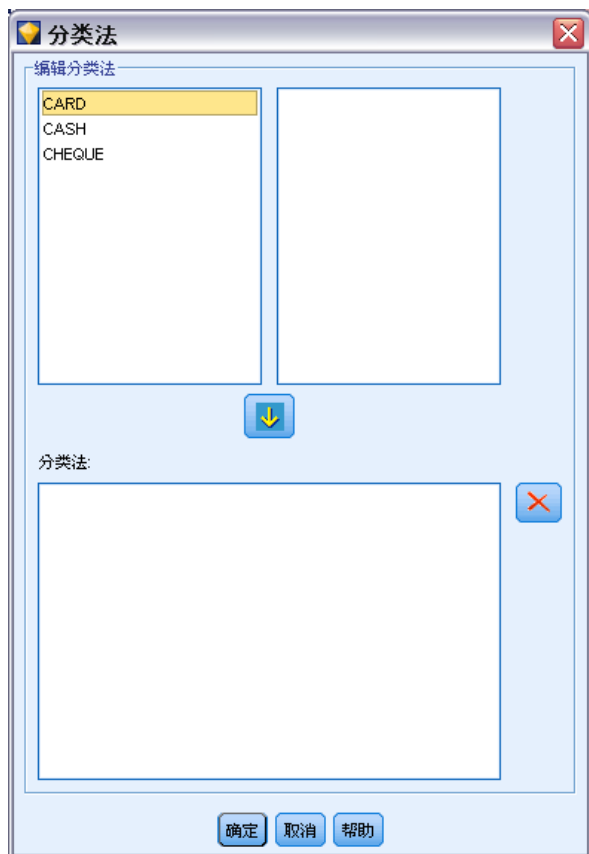
要添加类别，在新类别字段中键入其名称，并单击箭头按钮将其移到类别列表。

要删除类别，在类别列表中选择该类别并单击相邻的“删除”按钮。

分类法编辑器

通过“编辑分类法”对话框，可以将数据中定义的基本项目集和类别集组合在一起构建一个分类法。要向分类法中添加条目，请从左侧列表选择一个或多个项目或类别，从右侧列表选择一个或多个类别，然后单击箭头按钮。请注意，如果向分类法中添加内容会导致冲突（例如，指定 `cat1 -> cat2` 同时又指定与之相反的 `cat2 -> cat1`），则这样的添加不会成功。

图片 5-17
分类法编辑器



ISW 序列

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。

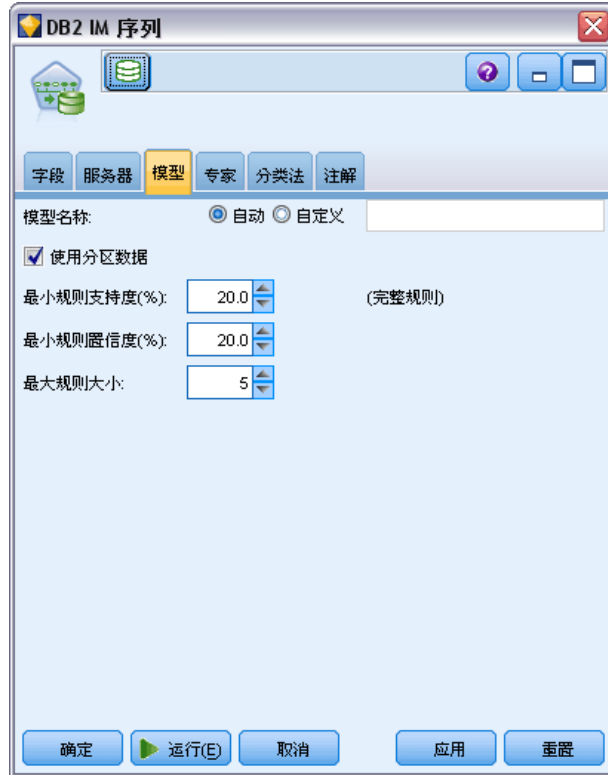
您可以在各种业务领域使用序列规则挖掘功能。例如，在零售行业，您可以发现典型购买序列。这些序列显示出不同的客户、产品和购买时间组合。利用这些信息，您可以识别出目前尚未购买某个特定产品的潜在客户。而且，您可以在适当时间向潜在客户提供产品。

序列是一个项目集合的有序集。序列包含下列分组级别：

- 同时发生的事件组成一个事务处理或一个项目集合。
- 每个项目或每个项目集合属于一个事务处理组。例如，所购买的物品属于某个客户，特定页面点击来自某个网络冲浪者，某个部件属于已生产的汽车等。发生在不同时间且属于同一个事务处理组的若干项目集合形成一个序列。

ISW 序列模型选项

图片 5-18
ISW 序列节点“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 15 源、过程和输出节点。

最低规则支持度 (%)。 关联规则或序列规则的最低支持水平。只有达到至少该支持水平的规则才会包含在模型中。结果计算为 $A/B*100$ ，其中 A 是包含在规则中出现的所有项目的组的数量，B 是所考虑的所有组的总数。如果您要关注更常见的关联或序列，请增加此设置。

最低规则置信度 (%)。 关联规则或序列规则的最低置信水平。只有达到至少该置信水平的规则才会包含在模型中。值计算为 $m/n*100$ ，其中 m 是包含连接规则头（后项）和规则体（前项）的组的数量，n 是包含规则体的组的数量。如果您获得的关联或序列太多或者不是非常相关，请尝试增加此设置。如果您获得的关联或序列太少，请尝试降低此设置。

最大规则大小。 规则中允许的最大项数，其中包括后项。如果相关关联或序列相对较短，则可以降低此设置，以加快集的构建速度。

注意：只对具有交易输入格式的节点进行评分；真值表（表格数据）格式保持非精炼。

ISW 序列专家选项

您可指定应在结果中包含或从结果中排除的序列规则。如果您想包含指定项目，则在模型中包含含有至少一个指定项目的规则。如果您想排除指定项目，则从结果丢弃含有任一指定项目的规则。

图片 5-19
ISW 序列节点“专家”选项卡



选择了使用项目约束之后，您添加到约束列表的任何项目将根据约束类型的设置从结果包含或排除。

约束类型。 选择您是从结果包含还是排除含有指定项目的关联规则。

编辑约束。 要添加项目到约束项目列表，在项目列表选中并单击右箭头按钮。

ISW 回归

ISW 回归节点支持以下回归算法：

- 变换（默认）
- 线性
- Polynomial
- RBF

变换回归

ISW 变换回归算法构建在树叶处具有回归方程的决策树模型。请注意，IBM 的 Visualizer 不会显示这些模型的结构。

IBM® SPSS® Modeler 浏览器显示设置和注解。但是无法浏览模型结构。用户可配置的构建设置相对较少。

线性回归

ISW 线性回归算法假定说明字段与目标字段之间存在线性关系。它会生成用方程式表示的模型。预测值应与观测值有所不同，因为回归方程是目标字段的近似值。这种差异称为残差。

IBM InfoSphere Warehouse Data Mining 建模认可没有说明值的字段。为了确定某个字段是否有说明值，除自动变量选择外，线性回归算法还执行统计检验。如果您知道字段没有说明值，则可以自动选择说明字段子集，以缩短运行时间。

线性回归算法提供下列可用来自动选择说明字段子集的方法：

逐步回归。对于逐步回归，您必须指定最低的显著性水平。只有显著性水平高于指定值的字段才会被线性回归算法使用。

R 平方回归。R 平方回归方法通过优化模型质量测量值来确定最佳模型。使用的质量测量值为下列其中一个：

- 平方 Pearson 相关系数
- 调整的平方 Pearson 相关系数。

默认情况下，线性回归算法通过使用调整的平方 Pearson 相关系数自动选择说明字段的子集，以优化模型的质量。

多项回归

ISW 多项回归算法假定存在多项式关系。多项回归模型是一个由以下几部分组成的方程：

- 多项回归的最高次数
- 目标字段的近似值
- 说明字段。

RBF 回归

ISW RBF 回归算法假定说明字段与目标字段之间存在某种关系。这种关系可以表示为高斯函数的线性组合。高斯函数为特定的径向基函数。

ISW 回归模型选项

在 ISW 回归节点的“模型”选项卡上，您可指定要使用的回归算法类型，以及：

- 是否使用分区的数据
- 是否执行测试运行

- R^2 值的限制
- 执行时间的限制

图片 5-20
ISW 回归节点“模型”选项卡



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 15 源、过程和输出节点。

回归方法。 选择您要执行的回归类型。有关详细信息，请参阅第 119 页码 ISW 回归。

执行测试运行。 您可以选择执行测试运行。然后将在训练分区上构建模型之后，执行 InfoSphere Warehouse Data Mining 测试运行。此运行会针对测试分区执行一次传递，以建立模型质量信息、提升图表等。

限制 R 平方。 此选项指定允许的最大系统错误数（平方 Pearson 相关系数， R^2 ）。此系数测量验证数据的预测错误与实际目标值之间的相关性。其值在 0（不相关）与 1（完美正相关或负相关）之间。您在此处定义的值设置了模型的可接受系统错误上限。

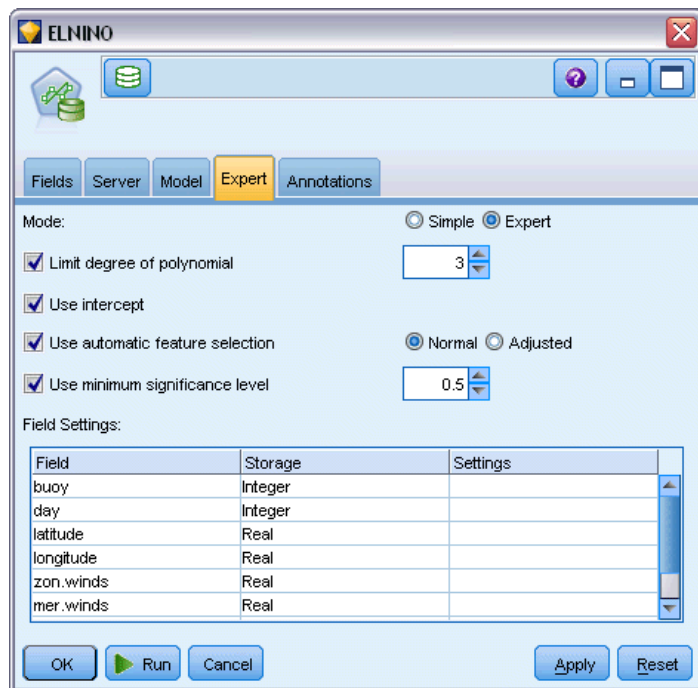
限制执行时间。 指定要求的最长执行时间（分钟）。

ISW 回归专家选项

在 ISW 回归节点的“专家”选项卡上，您可以为线性回归、多项回归或 DBF 回归指定多个高级选项。

线性回归或多项回归专家选项

图片 5-21
线性回归或多项回归的 ISW 回归节点“专家”选项卡



限制多项式次数。 设置多项回归的最高次数。如果将多项回归的最高次数设置为 1，则多项回归算法与线性回归算法完全相同。如果将多项回归的最高次数指定为一个较高的值，则多项回归算法有可能过度拟合。这意味着生成的模型会精确近似训练数据，但应用于非训练数据时则会失败。

使用截距。 如果启用此选项，则会强制回归曲线穿越原点。这就意味着该模型将不包含常项。

使用自动特征选择。 如果启用此选项，算法则会在您未指定最低显著性水平的情况下，尝试确定可能预测变量的最优化子集。

使用最低显著性水平。 如果指定了最低显著性水平，则使用逐步回归确定可能的预测变量的子集。只有其显著性高于指定值的独立字段才对回归模型的计算有贡献。

字段设置。 要指定单个输入字段的选项，单击“字段设置”表“设置”列中的相应行，并选择<指定设置>。 [有关详细信息，请参阅第 123 页码指定回归的字段设置。](#)

RBF 回归专家选项

图片 5-22

RBF 回归的 ISW 回归节点“专家”选项卡



使用输出样本大小。定义 N 中取 1 样本以供模型验证和测试。

使用输入样本大小。定义 N 中取 1 样本以供训练。

使用最大中心数。在每次传递中构建的最大中心数。由于在一次传递中，中心数可以增加至初始数目的两倍，因此实际中心数可能高于您指定的数目。

使用最小区域大小。指定给某个区域的最小记录数。

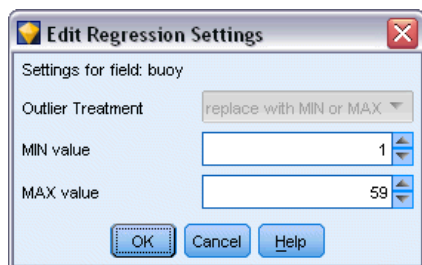
使用最大数据传递数。算法通过输入数据执行的最大传递数。如果指定此值，则它必须大于或等于最小传递数。

使用最小数据传递数。算法通过输入数据执行的最小传递数。仅当您拥有足够的训练数据并且确信存在良好的模型时，才可指定较大值。

指定回归的字段设置

您可在这里为单个输入字段指定值范围。

图片 5-23
指定输入字段的回归设置



最小值。此输入字段的最小有效值。

最大值。此输入字段的最大有效值。

ISW 聚类

聚类挖掘函数会搜索输入数据，以找出最常见的公共特征。它会将输入数据组合为聚类。每个聚类内的成员都具有相似的属性。对于数据内存在哪些模式，没有预先形成的概念。聚类是一个发现过程。

ISW 聚类节点提供了以下聚类方法：

- 人口统计
- Kohonen
- 增强 BIRCH（使用层次的平衡迭代减少及聚类）

demographic 聚类算法技术为基于分布的技术。基于分布的聚类提供对巨型数据库快速且自然的聚类。聚类数是自动选择的（您可以指定最大聚类数）。有大量用户可以配置的参数。

Kohonen 聚类算法技术是基于中心的技术。Kohonen 特征图尝试将聚类中心放在会使记录与聚类中心之间的整体距离最小的位置。聚类的可分离性未考虑在内。中心向量被组织在一个具有特定列数和行数的图中。这些向量互相连接，因此不仅使得与训练记录最接近的中选向量得到调整，而且也使得与之邻近的那些向量也得到调整。但是，其他中心离得越远，它们调整得就越少。

增强 BIRCH 聚类算法技术是基于分布的技术，它尝试使记录与聚类之间的整体距离最小。默认使用对数似然距离来确定记录与聚类之间的距离；或者，如果所有活动字段均为数值，还可选择 Euclidean 距离。BIRCH 算法执行两个独立的步骤。首先，它在聚类特征树中排列输入记录，这样类似记录成为了相同树节点的一部分；然后，它在内存中聚类该树的树叶，以产生最终聚类结果。

ISW 聚类模型选项

在“聚类”节点的“模型”选项卡上，可指定用于创建聚类的方法以及一些相关选项。

图片 5-24
ISW 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 15 源、过程和输出节点。

聚类方法。 选择用于创建聚类的方法：人口统计、Kohonen 或增强 BIRCH。有关详细信息，请参阅第 124 页码 ISW 聚类。

聚类数量限制。 通过防止很多小聚类的生成，聚类数量限制节省了运行时间。

行数/列数。（仅 Kohonen 方法）指定 Kohonen 特征图的行数和列数。（仅当选中限制 kohonen 传递数并取消选中限制聚类数时才可用。）

限制 kohonen 传递数。（仅 Kohonen 方法）指定聚类算法在训练运行期间对数据进行的传递次数。对于每次传递，中心向量都会进行调整，以使得聚类中心和记录之间的总距离最短。另外，向量的调整量也会减小。在第一次传递中，调整比较大。在最后一次传递中，中心的调整量则相当小。只进行细微的调整。

距离测量。（仅增强 BIRCH 方法）选择 BIRCH 算法要使用的记录到聚类距离测量。可以选择对数似然距离（默认）或欧几里德距离。注意：如果所有活动字段均为数值，则只能选择欧几里德距离。

最大叶节点数。（仅增强 BIRCH 方法）您希望聚类特征树可以具有的最大叶节点数。聚类特征树是增强 BIRCH 算法中第一步的结果，通过以树形式排列数据记录，以便使类似记录属于相同叶节点。该算法的运行时间随叶节点数量的增加而增加。默认值为 1000。

Birch 传递数。（仅增强 BIRCH 方法）为了精练聚类结果，算法对数据进行的传递次数。传递数会影响训练运行的处理时间（每次传递需要扫描整个数据）与模型质量。较低的传递数会导致处理时间较短，但模型质量也较低。较高的传递数会导致处理时间较长，但通常会产生更好的模型。通常传递数为 3 或以上可以带来较好的结果。默认值为 3。

ISW 聚类专家选项

在聚类节点的“专家”选项卡上，您可以指定如相似性阈值、执行时间限制和字段权重这些高级选项。

图片 5-25
ISW 聚类节点“专家”选项卡



限制执行时间。选中此复选框以启用可控制创建模型的用时的选项。您可按分钟、需处理训练数据的最小百分比或同时使用以上两种方式指定时间。此外，对于 Birch 方法，还可以指定要在 CF 树中创建的最大叶节点数。

指定相似性阈值。（仅 Demographic 聚类）属于同一聚类的两个数据记录的相似度下限。例如，一个 0.25 的值意味着值的相似度为 25% 的记录有可能分配给同一聚类。值 1.0 意味着记录必须相同才能出现在同一聚类中。

字段设置。要指定单个输入字段的选项，单击“字段设置”表“设置”列中的相应行，并选择<指定设置>。

指定聚类的字段设置

您可在这里指定单个输入字段的选项。

图片 5-26
指定输入字段的聚类设置



字段权重。模型构建过程期间分配或多或少的权重给字段。例如，如果您认为此字段与其他字段相比对模型不太重要，则相对于其他字段降低其字段权重。

值权重。分配或多或少的权重给此字段的特定值。某些字段值可能比其他值更常见。字段中出现频率低的值的符合性与出现频率高的值的符合性相比对聚类更重要。您可选择以下一种方法确定此字段值的权重（不论哪种情况，出现频率低的值具有较大权重，出现频率高的值具有较小权重）：

- **对数。**根据其在输入数据中概率的对数分配权重给每个值。
- **概率。**根据其在输入数据中的概率分配权重给每个值。

无论哪种方法，您都可以选择具有补偿选项来补偿应用于每个字段的值加权。如果您补偿值加权，则加权字段的总体重要性等于未加权字段的总体重要性。无论可能值的数量有多少都将如此。补偿后的加权只影响可能值集合内符合性的相对重要性。

使用相似性尺度。如果您想使用相似性尺度控制字段相似性测量的计算，则选中此复选框。将相似性尺度指定为绝对值数。只对活动数值字段考虑进行该指定。如果不指定相似性尺度，则使用默认值（标准差的一半）。要获得更多数量的聚类，通过降低数值字段的相似性尺度来降低各对聚类之间的平均相似性。

离群值处理。离群值是在字段的指定值范围之外的字段值，由最小值和最大值定义该范围。您可选择此字段离群值的处理方式。

- 默认为无，意味着不对离群值进行特殊操作。
- 如果您选择替换为最小值或最大值，则小于最小值的字段值或大于最大值的字段值将被相应替换为最小值或最大值。此时您可设置最小值和最大值。
- 如果您选择视为缺失值，则离群值被视为缺失值且被忽略。此时您可设置最小值和最大值。

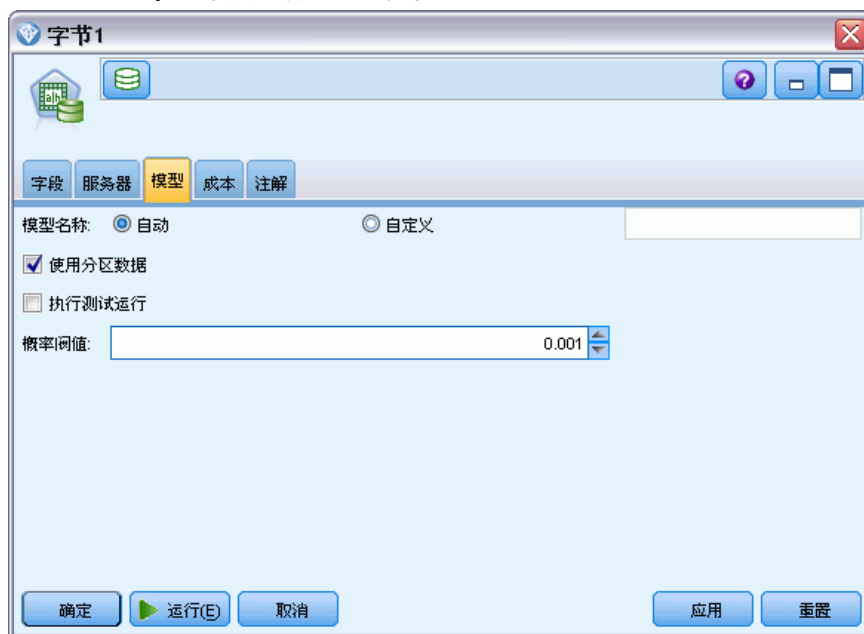
ISW Naive Bayes

Naive Bayes 是广泛用于处理分类问题的算法。因为该模型将所有给出的预测变量视为互相独立的，因而取名为 naïve。Naive Bayes 是一种快速的、可伸缩的算法，用于计算属性和目标属性组合的条件概率。使用训练数据确定独立的概率。给定来自每个输入变量的所有值分类的发生率，使用此概率可计算出每个目标类的似然值。

ISW Naive Bayes 分类算法是一种概率分类器。它基于概率模型，采用强独立性假定。

ISW Naive Bayes 模型选项

图片 5-27
ISW Naive Bayes 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 15 源、过程和输出节点。

执行测试运行。 您可以选择执行测试运行。在训练分区上构好模型之后，执行 IBM InfoSphere Warehouse Data Mining 测试运行。这一步针对测试分区执行一次传递，以建立模型质量信息、提升图表等。

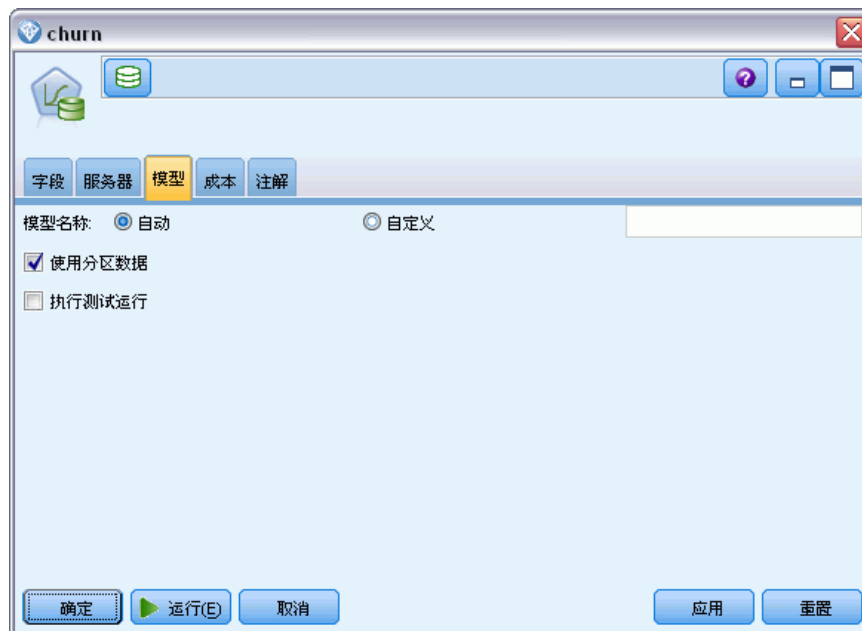
概率阈值。 概率阈值定义在训练数据中不可见的任何预测变量与目标值组合的概率。此概率必须处于 0 到 1 之间，默认值为 0.001。

ISW Logistic 回归

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。这种回归相当于线性回归，但 ISW Logistic 回归算法用标志（二元）目标字段代替了数值字段。

ISW Logistic 回归模型选项

图片 5-28
ISW Logistic 回归节点“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 15 源、过程和输出节点。

执行测试运行。 您可以选择执行测试运行。在训练分区上构好模型之后，执行 IBM InfoSphere Warehouse Data Mining 测试运行。这一步针对测试分区执行一次传递，以建立模型质量信息、提升图表等。

ISW 时间序列

ISW 时间序列算法可以基于过去的已知事件预测未来的事件。

与一般回归方法类似，时间序列算法也是预测数值。但与一般回归方法相比，时间序列预测侧重于某个有序序列的未来值。这些预测（Predictions）通常称为预测（Forecasts）。

时间序列算法为单变量算法。这意味着自变量为时间列或顺序列。预测是基于过去的值。它们并不基于其他自变量列。

时间序列算法不同于一般回归算法，因为它们不仅预测未来值，而且还在预测中融入季节周期。

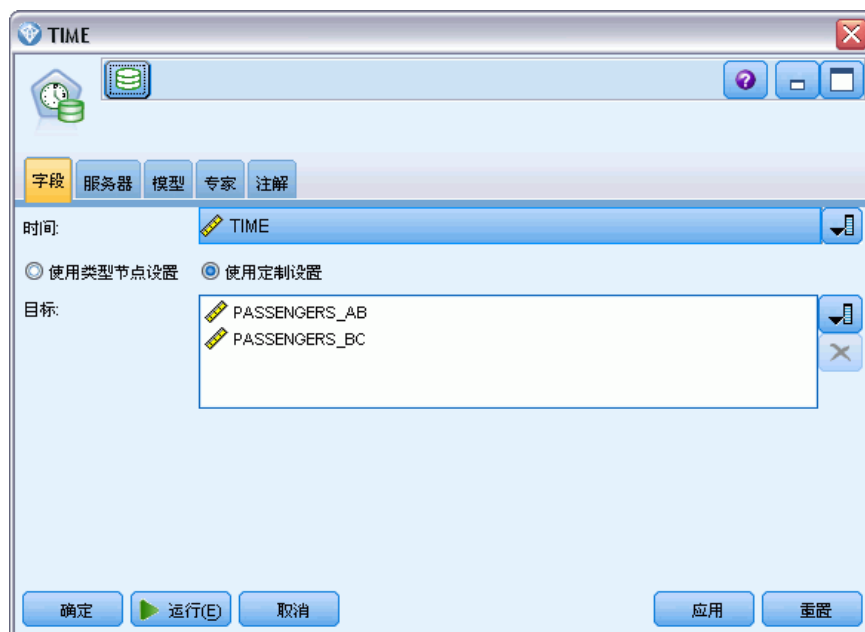
时间序列挖掘函数提供下列算法以预测未来趋势：

- 求和自回归移动平均数 (ARIMA)
- 指数平稳
- 季节性趋势分解

该算法根据不同的模型假定，为您的数据生成最佳预测。可以同时计算所有预测。该算法计算详细的预测，包括原始时间序列的季节性行为。如果安装了 IBM InfoSphere Warehouse Client，则可使用时间序列可视化工具来评估和比较结果曲线。

ISW 时间序列字段选项

图片 5-29
ISW 时间序列节点“字段”选项卡



时间。 选择包含时间序列的输入字段。这必须是存储类型为日期、时间、时间戳、实数或整数的字段。

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。

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。

ISW 时间序列模型选项

图片 5-30
ISW 时间序列节点“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。

预测算法。 选择要用于建模的算法。可以选择以下方法中的一种或多种：

- ARIMA
- 指数平滑
- 季节性趋势分解

预测结束时间。 指定是自动计算还是手动指定预测结束时间。

时间字段值。 如果将预测结束时间设为手动指定，则在此输入预测结束时间。根据时间字段类型输入相应的值。例如，如果类型为代表小时的整数，则可以输入 48 以表示在处理 48 小时数据后停止预测。或者，该字段还可能提示您输入日期或时间作为预测结束时间值。

ISW 时间序列专家选项

图片 5-31
ISW 时间序列节点“专家”选项卡



使用所有记录构建模型。这是默认设置。当模型构建时将分析所有记录。

使用记录子集构建模型。如果只打算从部分可用数据中创建模型，选择此选项。例如，当您拥有超大量重复性数据时，可能需要使用此选项。

输入开始时间值和结束时间值以确定要使用的数据。请注意，在这些字段中可以输入的值取决于时间字段的类型。例如，它可能是小时或天数，或者是特定日期或时间。

缺失目标值的插值方法。如果要处理的数据存在一个或多个缺失值，则在此选择要用来计算它们的方法。可以选择以下选项之一：

- 线性
- 指数样条
- 三次样条

显示 ISW 时间序列模型

ISW 时间序列模型为非精练模型形式的输出，其中包含从数据中抽取的信息，但不能直接用于产生预测。

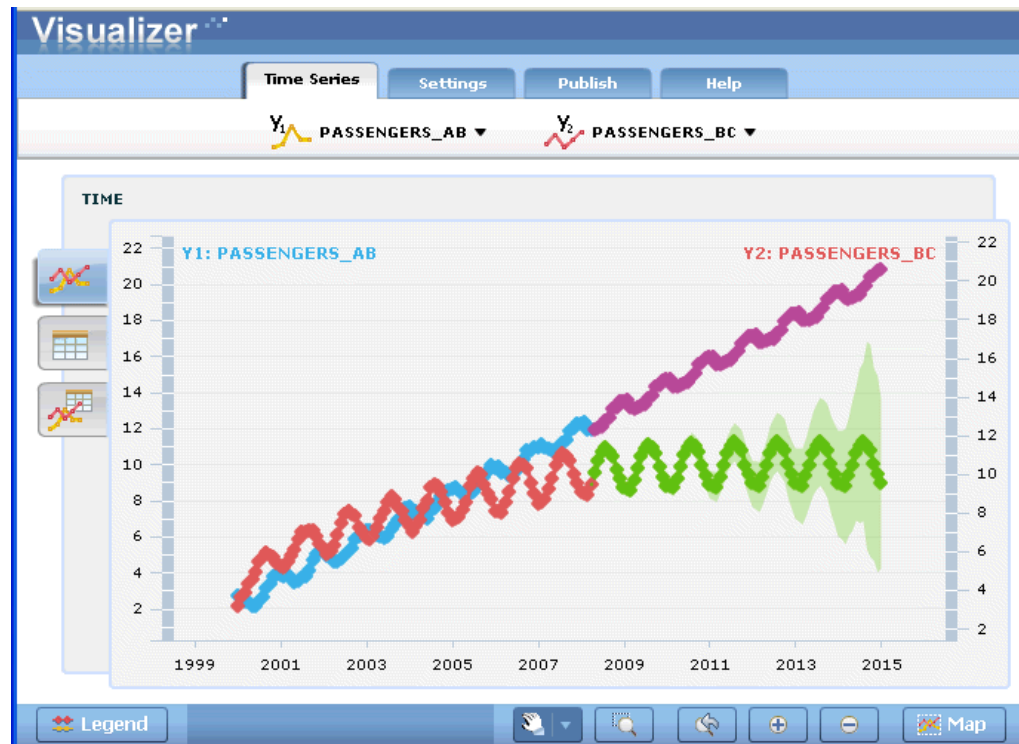
图片 5-32
非精练模型的图标



有关详细信息，请参阅第 3 章中的非精练模型中的 IBM SPSS Modeler 15 建模节点。

如果安装了 IBM InfoSphere Warehouse Client，则可使用时间序列可视化工具来获得时间序列数据的图形显示。

图片 5-33
在可视化工具中显示的 ISW 时间序列模型



使用时间序列可视化工具：

- ▶ 确保已经完成将 IBM® SPSS® Modeler 与 IBM InfoSphere Warehouse 集成的任务。 [有关详细信息，请参阅第 95 页码启用 IBM InfoSphere Warehouse 集成。](#)
- ▶ 在“模型”选项板中，双击非精练模型图标。
- ▶ 在对话框的“服务器”选项卡上，单击“查看”按钮，以在默认的网络浏览器中显示可视化工具。

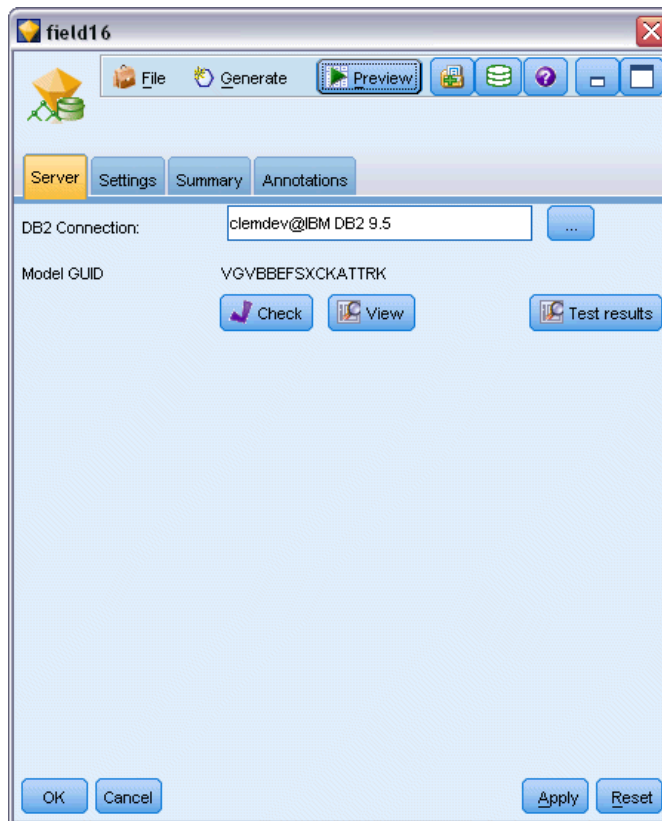
ISW Data Mining 模型块

您可以从 IBM® SPSS® Modeler 包含的 ISW 决策树节点、关关节点、序列节点、回归节点和聚类节点创建模型。

ISW 模型块服务器选项卡

“服务器”选项卡提供了用于执行一致性检查和启动 IBM Visualizer 工具的选项。

图片 5-34
ISW 模型块“服务器”选项卡



IBM® SPSS® Modeler 可通过将完全相同的生成模型关键字字符串存储在 SPSS Modeler 模型和 ISW 模型中执行一致性检查。一致性检查通过单击“服务器”选项卡上的检查按钮执行。有关详细信息，请参阅第 103 页码管理 DB2 模型。

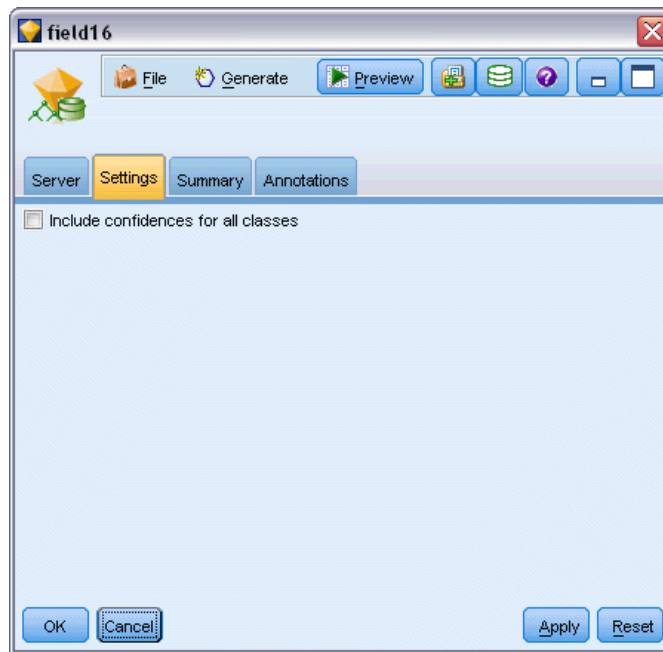
可视化工具是浏览 InfoSphere Warehouse Data Mining 模型的唯一方法。该工具可选择与 InfoSphere Warehouse Data Mining 一起安装。有关详细信息，请参阅第 95 页码启用 IBM InfoSphere Warehouse 集成。

- 单击查看可启动该可视化工具。该工具显示的内容取决于生成的节点类型。例如，从 ISW 决策树模型块启动时，该可视化工具将返回一个预测类视图。
- 单击测试结果（仅用于决策树和序列）可启动该可视化工具，查看所生成模型的整体质量。

ISW 模型块“设置”选项卡

在 IBM® SPSS® Modeler 中，通常只会提供一个预测以及关联的概率或置信度。此外，用于显示每个结果的概率的用户选项（与 logistic 回归中的选项类似）是模型块“设置”选项卡上可用的得分时间选项。

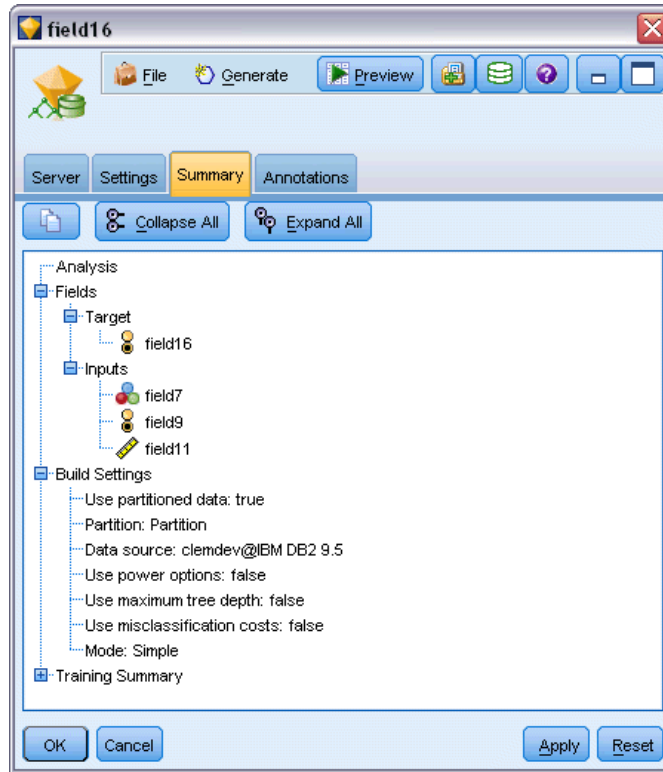
图片 5-35
ISW 模型块“设置”选项卡



包括所有类的置信度。对于目标字段的每个可能结果，添加表示置信水平的一列。

ISW 模型块汇总选项卡

图片 5-36
ISW 模型块“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。查看完成后要隐藏结果时，请使用展开控件来折叠想要隐藏的具体结果，或者单击全部折叠按钮来折叠所有结果。

分析。 显示指定模型的相关信息。如果已执行附加到该模型块的分析节点，则还会在此部分显示通过分析获得的信息。[有关详细信息，请参阅第 6 章中的分析节点中的 IBM SPSS Modeler 15 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。

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。

ISW Data Mining 示例

用于 Windows 的 IBM® SPSS® Modeler 附带了一些用于说明数据库挖掘过程的演示流。这些流位于 IBM® SPSS® Modeler 安装文件夹中，该文件夹目录为：

\Demos\Database_Modeling\IBM DB2 ISW

注意：可以从 Windows “开始” 菜单 SPSS Modeler 程序组中访问这些 Demos 文件夹。

下列流按顺序一起使用可作为数据库挖掘过程的一个示例：

- 1_upload_data.str—用于清理平面文件中的数据并将其上载到 DB2。
- 2_explore_data.str—用作使用 SPSS Modeler 进行数据探索的示例。
- 3_build_model.str—用于构建 ISW 决策树模型。
- 4_evaluate_model.str—用作使用 SPSS Modeler 进行模型评估的示例。
- 5_deploy_model.str—用于部署进行数据库内评分的模型。

这些示例流中使用的数据集与信用卡申请有关，演示了同时带有分类和连续预测变量的分类问题。有关此数据集的更多信息，请参阅 SPSS Modeler 安装文件夹下的下列文件：

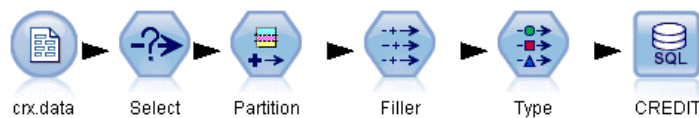
\Demos\Database_Modeling\IBM DB2 ISW\crx.names

此数据库可从位于 <http://archive.ics.uci.edu/ml/> 的 UCI Machine Learning Repository 中获得。

示例流：上传数据

第一个示例流 1_upload_data.str 用于清理平面文件中的数据并将其上载到 DB2。

图片 5-37
用于上传数据的示例流

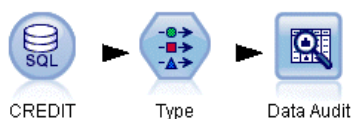


填充节点用于缺失值处理，它会将从文本文件 crx.data 中读取的空字段替换为 NULL 值。

示例流：探索数据

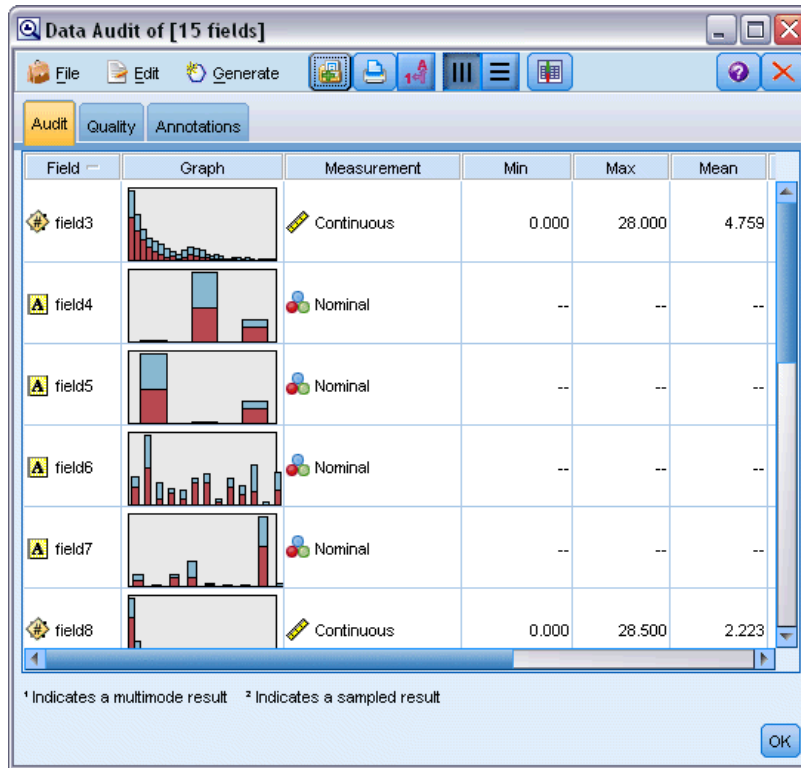
第二个示例流 2_explore_data.str 用于演示如何在 IBM® SPSS® Modeler 中探索数据。

图片 5-38
探索数据的示例流



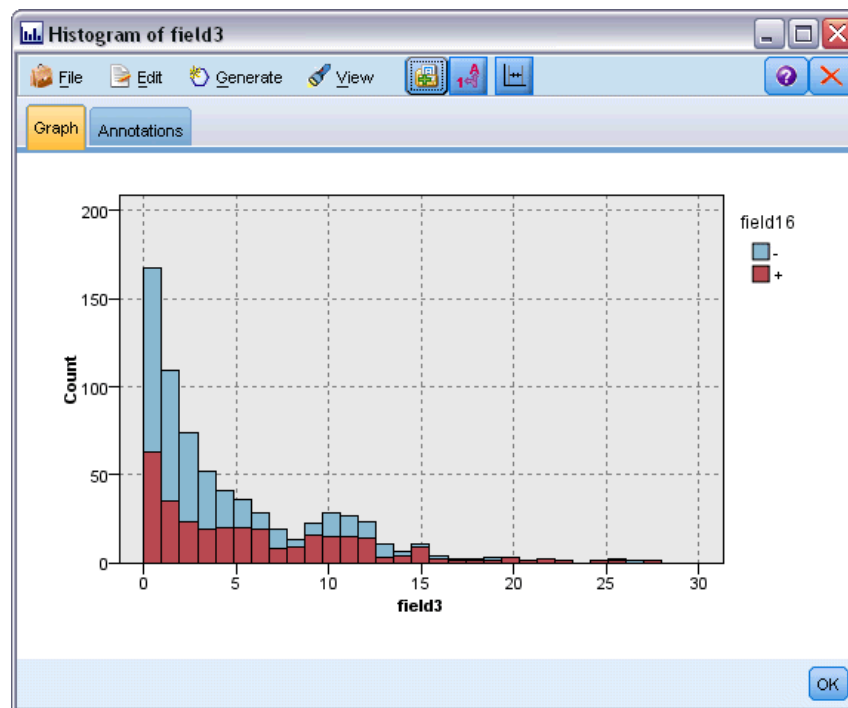
数据探索过程的一个典型步骤是将数据审核节点附加到数据。可在输出节点选项板中找到数据审核节点。

图片 5-39
数据审核结果



您可以使用数据审核节点的输出来概览字段和数据分布。双击“数据审核”窗口中的图形可显示一个更为详细的图形，用于更深入地探索某个给定字段。

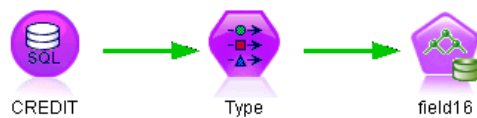
图片 5-40
通过双击“数据审核”窗口创建的直方图



示例流：构建模型

第 3 个示例流，即 3_build_model.str，演示 IBM® SPSS® Modeler 中的“£”型构建。您可以将数据库建模节点连接到该流，然后双击该节点以指定构建设置。

图片 5-41
数据库建模示例流中的紫色节点表示在数据库内执行

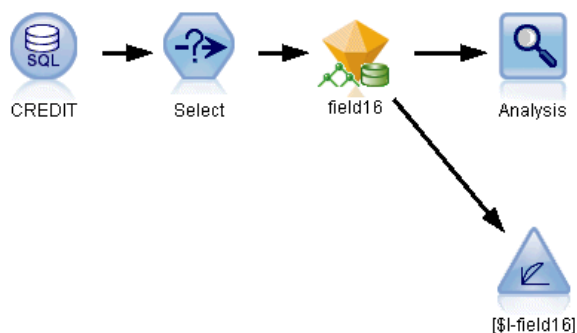


使用建模节点的“模型”和“专家”选项卡，可以调整最大树深度，通过设置最高纯度和每内部节点的最小观测值数，可以停止从构建初始决策树时进行的进一步节点分割。有关详细信息，请参阅第 108 页码 ISW 决策树。

示例流：评估模型

第 4 个示例流，即 4_evaluate_model.str，演示构建数据库内模型时使用 IBM® SPSS® Modeler 的优点。一旦执行完模型，即可将它添加回数据流中并使用 SPSS Modeler 提供的多种工具来评估模型。

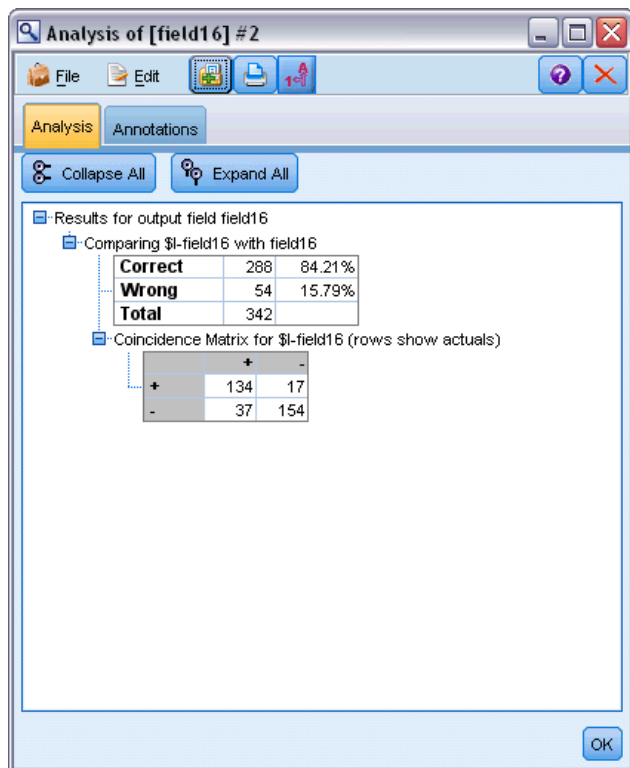
图片 5-42
用于模型评估的示例流



首次打开流时，模型块（field16）不包含在流中。打开 CREDIT 源节点并确保您已指定了数据源。接下来，假如您已运行 3_build_model.str 流来创建“模型”选项板中的 field16 模型块，则可通过单击工具栏上的 运行 按钮（具有绿色三角形的按钮）运行断开连接的节点。这样，运行的脚本会将 field16 模型块复制到流中，将其连接到现有节点，然后运行流中的终端节点。

可附加“分析”节点（位于“输出”选项板），以创建说明每个生成的字段和其目标字段之间的匹配模式的符合矩阵。运行分析节点以查看结果。

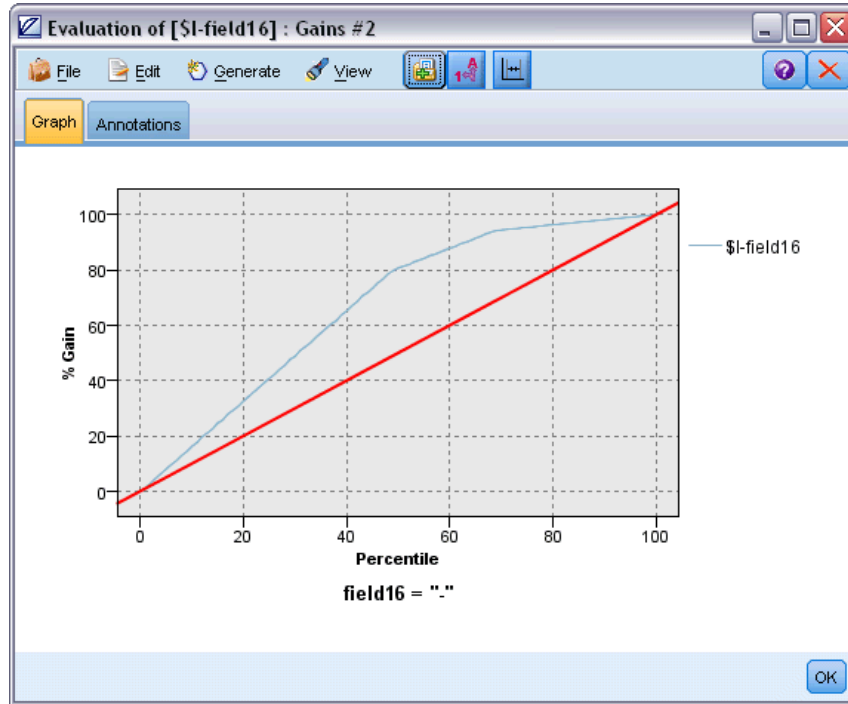
图片 5-43
分析节点结果



生成的表表明，ISW 的决策树算法所生成的 84.21% 的预测是正确的。

您还可以创建收益图表来显示模型带来的准确性提高。将“评估”节点附加到生成的模型，然后运行流，以查看结果。

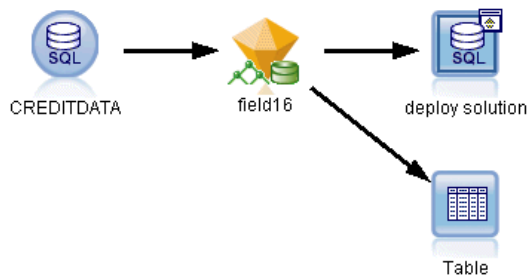
图片 5-44
使用评估节点生成的收益图表



示例流：部署模型

如果满意模型的准确度，可以将其部署到外部应用程序中使用或用于将得分写回到数据库中。在示例流 5_deploy_model.str 中，数据是从表 CREDIT 中读取的。当运行部署解决方案数据库导出节点时，实际上不对数据评分。相反，流会创建发布的映像文件 credit_scorer.pim 和发布的参数文件 credit_scorer.par。

图片 5-45
用于部署模型的示例流



如前例所示，流运行的脚本会将 field16 模型块从“模型”选项板复制到流中，将其连接到现有节点，然后运行流中的终端节点。此时您必须首先在数据库源和导出节点中指定数据源。

采用 IBM Netezza Analytics 进行数据库建模

IBM SPSS Modeler and IBM Netezza Analytics

IBM® SPSS® Modeler 支持 IBM® Netezza® Analytics 集成，这提供了在 IBM Netezza 服务器上运行高级分析的能力。这些功能可通过访问 SPSS Modeler 图形用户界面和面向工作流的开发环境加以使用，使您可以在 IBM Netezza 环境中运行数据挖掘算法。

SPSS Modeler 支持集成来自 Netezza Analytics 的以下算法。

- 决策树
- K-Means
- 贝叶斯网络
- Naive Bayes
- KNN
- 分裂式聚类
- 主成分分析 (PCA)
- 回归树
- 线性回归

有关算法的更多信息，请参阅 Netezza Analytics 开发人员指南和 Netezza Analytics 参考指南。

集成 IBM Netezza Analytics 的要求

以下是使用 IBM® Netezza® Analytics 执行数据库内建模的必备条件。您可能需要咨询数据库管理员以确保满足这些条件。

- IBM® SPSS® Modeler 以本地模式或在 Windows 或 UNIX 上（不包括 zLinux，未为其提供 IBM Netezza ODBC 驱动程序）安装 IBM® SPSS® Modeler Server 后运行。
- 运行 IBM® SPSS® In-Database Analytics 数据包的 IBM Netezza Performance Server 6.0 或更高版本。
- 连接到 IBM Netezza 数据库所需的 ODBC 数据源。 [有关详细信息，请参阅第 144 页码启用 IBM Netezza Analytics 集成。](#)
- SPSS Modeler 中启用的 SQL 生成和优化。 [有关详细信息，请参阅第 144 页码启用 IBM Netezza Analytics 集成。](#)

注意：数据库建模和 SQL 优化需要在 IBM® SPSS® Modeler 计算机上启用 SPSS Modeler Server 连接。通过启用此设置，您可以访问数据库算法，直接从 SPSS Modeler 回送 SQL 以及访问 SPSS Modeler Server。要验证当前许可证的状态，请从 SPSS Modeler 菜单中选择以下项目。

帮助 > 关于 > 其他详细信息

如果启用了连接，您可以在“许可证状态”选项卡中看到选项服务器启用。

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 15 用户指南](#)。

启用 IBM Netezza Analytics 集成

启用与 IBM® Netezza® Analytics 的集成包括以下步骤。

- 配置 Netezza Analytics
- 创建 ODBC 源
- 在 IBM® SPSS® Modeler 中启用集成
- 在 SPSS Modeler 中启用 SQL 生成和优化

在以下部分中将介绍这些内容。

配置 IBM Netezza Analytics

要安装和配置 IBM® Netezza® Analytics，请参阅 Netezza Analytics 文档，特别是《Netezza Analytics 安装指南》，以获得更多详细信息。该指南中的设置数据库权限部分包含需要运行以允许 IBM® SPSS® Modeler 流读取数据库的脚本的详细信息。

注意：如果您要使用依赖于矩阵计算的节点（Netezza PCA 和 Netezza 线性回归），则必须通过运行 `CALL NZM..INITIALIZE()`；来初始化 Netezza 矩阵引擎，否则存储程序执行将失败。对于每个数据库，该初始化为一次性设置步骤。

为 IBM Netezza Analytics 创建 ODBC 源

要启用 IBM Netezza 数据库和 IBM® SPSS® Modeler 之间的连接，您需要创建 ODBC 系统数据源名称（DSN）。

在创建 DSN 之前，您应当对 ODBC 数据源和驱动程序，以及 SPSS Modeler 中的数据库支持有基本的了解。[有关详细信息，请参阅第 2 章中的数据访问中的 IBM SPSS Modeler Server 15 管理和性能指南](#)。

如果以分布式方式运行 IBM® SPSS® Modeler Server，请在服务器计算机上创建 DSN。如果以本地（客户机）模式运行，请在客户计算机上创建 DSN。

Windows 客户端

- ▶ 从您的 Netezza Client CD 上，运行 `nzodbcsetup.exe` 文件以启动安装程序。按照屏幕说明操作以安装驱动程序。有关详细说明，请参阅《IBM Netezza ODBC、JDBC 和 OLE DB 安装和配置指南》。
- ▶ 创建 DSN。
 - 注意：菜单排列顺序取决于 Windows 版本。
 - **Windows XP。**从“开始”菜单中选择控制面板。双击管理工具，然后双击数据源 (ODBC)。
 - **Windows Vista。**从“开始”菜单中选择控制面板，然后选择系统维护。双击管理工具，选择数据源 (ODBC)，然后单击打开。
 - **Windows 7。**从“开始”菜单中选择控制面板，选择系统和安全，然后选择管理工具。选择数据源 (ODBC)，然后单击打开。
- ▶ 单击系统 DSN 选项卡，然后单击添加。
- ▶ 从列表中选择 **NetezzaSQL**，然后单击完成。
- ▶ 在 Netezza ODBC 驱动程序设置屏幕的 DSN 选项选项卡上，键入选择的数据源名称、IBM Netezza 服务器的主机名或 IP 地址、连接端口号、使用的 Netezza 实例的数据库，以及用于数据库连接的用户名和密码信息。单击帮助按钮获得字段说明。
- ▶ 单击测试连接按钮并确保您连接到数据库。
- ▶ 在成功连接后，重复单击确定以退出 ODBC 数据源管理器屏幕。

Windows 服务器

对于 Windows Server，该程序与 Windows XP 客户端的程序相同。

UNIX 或 Linux 服务器

以下程序适用于 UNIX 或 Linux 服务器（不包括 zLinux，未提供适用的 IBM Netezza ODBC 驱动程序）。

- ▶ 从您的 Netezza Client CD 上，将对应的 `<platform>cli.package.tar.gz` 文件复制到服务器上的临时位置。
- ▶ 通过 `gunzip` 和 `untar` 命令，提取存档内容。
- ▶ 为提取的 `unpack` 脚本添加执行权限。
- ▶ 运行脚本，并在屏幕提示时给出回答。
- ▶ 编辑 `modelersrv.sh` 文件以包括以下行。

```
. /usr/IBM/SPSS/SDAP61_notfinal/odbc.sh
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64
NZ_ODBC_INI_PATH=/usr/IBM/SPSS/SDAP61_notfinal; export NZ_ODBC_INI_PATH
```

- ▶ 找到文件 `/usr/local/nz/lib64/odbc.ini` 并将其内容复制到随 SDAP 6.1 安装的 `odbc.ini` 文件（由环境变量 `$ODBCINI` 定义）中。

注意：对于 64 位 Linux 系统，Driver 参数错误地引用了 32 位驱动程序。当您在上一步骤中复制 `odbc.ini` 内容时，应相应地编辑该参数中的路径，例如：

```
/usr/local/nz/lib64/libnzodbc.so
```

- ▶ 编辑 Netezza DSN 定义中的参数，以反映要使用的数据库。
- ▶ 重新启动 SPSS Modeler Server，并在客户端上测试使用 Netezza 数据库内挖掘节点。

在 IBM SPSS Modeler 中启用 IBM Netezza Analytics 集成

- ▶ 在 IBM® SPSS® Modeler 主菜单中，选择
工具 > 选项 > 辅助应用程序。
- ▶ 单击 IBM Netezza 选项卡。

启用 Netezza Data Mining 集成。 启用 SPSS Modeler 窗口底部的“数据库建模”选项板（如尚未显示）并添加 Netezza Data Mining 算法的建模节点。

Netezza 连接。 单击 **编辑** 按钮，并选择之前在创建 ODBC 源时设置的 Netezza 连接字符串。有关详细信息，请参阅第 144 页码为 [IBM Netezza Analytics 创建 ODBC 源](#)。

启用 SQL 生成和优化

由于使用超大型数据集的可能性，出于性能的原因，您应在 IBM® SPSS® Modeler 中启用 SQL 生成和优化选项。

- ▶ 从 SPSS Modeler 菜单中选择：
工具 > 流属性 > 选项

图片 6-1
优化设置



- ▶ 在导航窗格中单击优化选项。
- ▶ 确认是否已启用生成 SQL 选项。要使数据库建模正常发挥作用，此设置是必需的。
- ▶ 选中优化 SQL 生成和优化其他执行（非严格必需但强烈推荐使用，以使性能更优）。

有关详细信息，请参阅第 5 章中的设置流的优化选项中的 IBM SPSS Modeler 15 用户指南。

采用 IBM Netezza Analytics 构建模型

每种受支持的算法均具有对应的建模节点。您可从节点选项板上的“数据库建模”选项卡访问 IBM Netezza 建模节点。有关详细信息，请参阅第 3 章中的节点选项板中的 IBM SPSS Modeler 15 用户指南。

数据注意事项

数据源中的字段可以包含不同数据类型的变量，具体取决于建模节点。在 IBM® SPSS® Modeler 中，数据类型称为**测量级别**。在建模节点的“字段”选项卡上，通过图标来指示其输入和目标字段的允许测量级别类型。[有关详细信息，请参阅第 4 章中的测量级别中的 IBM SPSS Modeler 15 源、过程和输出节点。](#)

目标字段。目标字段是您打算预测其值的字段。在可以指定目标的情况下，只能选择一个源数据字段作为目标字段。

记录 ID 字段。指定用于作为各个观测值唯一标识的字段。例如，ID 字段，比如客户 ID。如果源数据不包含 ID 字段，您可以通过“导出”节点来创建此字段，如下所示。

- ▶ 选择源节点。
- ▶ 在节点选项板的“字段选项”选项卡中，双击“导出”节点。
- ▶ 在工作区上双击“导出”节点的图标可将其打开。
- ▶ 在 导出字段字段中，输入（例如） ID。
- ▶ 在公式字段中，输入 @INDEX 并单击确定。
- ▶ 将“导出”节点连接到流的其余部分。

处理 Null 值

如果输入数据包含 Null 值，使用某些 Netezza 节点可能导致错误消息或流长时间运行，因此我们建议删除包含 Null 值的记录。请使用以下方法。

- ▶ 将“选择”节点附加到源节点。
- ▶ 将“选择”节点的模式选项设置为丢弃。
- ▶ 在条件字段中输入以下内容：
`@NULL(field1) [or @NULL(field2) [... or @NULL(fieldN)]]`

确保包括每个输入字段。

- ▶ 将“选择”节点连接到流的其余部分。

模型输出

包含 Netezza 建模节点的流在每次运行时可能产生略微不同的结果。这是因为数据在建模之前被读入临时表，因此节点读取源数据的顺序并不始终相同。但是，这种影响产生的差异可以忽略不计。

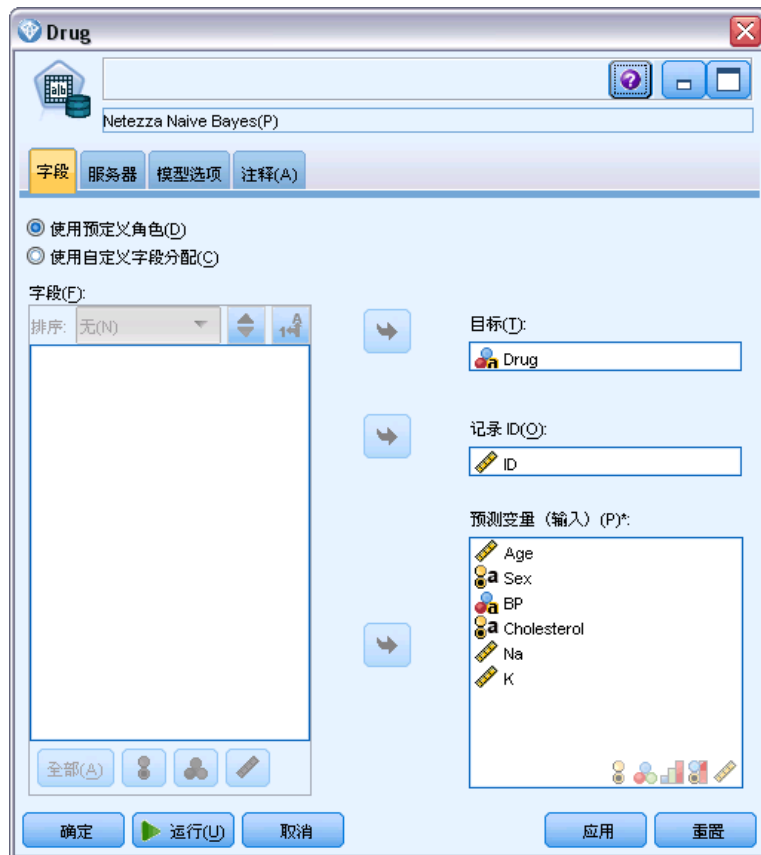
一般评论

- 在 IBM® SPSS® Collaboration and Deployment Services 中，不能使用包含 IBM Netezza 数据库建模节点的流来创建评分配置。
- Netezza 节点构建的模型无法进行 PMML 导出或导入。

Netezza 模型 - 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 6-2
Netezza 字段选项示例



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 15 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。对于广义线性模型，请另查看此屏幕中的试验字段。

记录 ID。此字段将用作唯一记录标识符。

预测变量（输入）。 选择一个或多个字段作为预测输入。

Netezza 模型 – 服务器选项

在该选项卡上，指定到用于存储模型的 IBM Netezza 数据库。

图片 6-3
Netezza 服务器选项示例



Netezza DB Server 详细信息。 您可在该处指定要为模型使用的数据库的连接详细信息。

- **使用上游连接。**（默认）使用上游节点中指定的连接详细信息，例如数据库源节点。注意：该选项仅在所有上游节点能够使用 SQL 回送的情况下有用。在此情况下，无需将数据移出数据库，因为 SQL 完全实现所有的上游节点。
- **移动数据到连接。** 将数据移动到您在此处指定的数据库。这样，当数据位于另一个 IBM Netezza 数据库，或其他供应商的数据库，甚至位于平面文件中时，建模仍可工作。此外，如果由于某个节点未执行 SQL 回送而导致数据已被提取，则数据将移回到此处指定的数据库中。单击编辑按钮以浏览并选择连接。警告：IBM® Netezza® Analytics 通常用于大型数据集。在数据库之间传输大量数据，或者从数据库中取出或存入大量数据，可能非常耗时，应尽可能避免。

表名称。用于存储模型的数据库表的名称。注意：它必须为新表；您不得使用现有表进行此操作。

注释

- 用于建模的连接无需与流的源节点中使用的连接相同。例如，可能有一个流可以用于访问一个 IBM Netezza 数据库的数据，将数据下载到 IBM® SPSS® Modeler 以进行清理或执行其他操作，然后将数据上传到另一个 IBM Netezza 数据库，用于建模。不过，请注意，此类配置对性能有负面影响。
- ODBC 数据源名称可有效地内嵌于每个 SPSS Modeler 流中。如果在一个主机创建的流在另一个主机上执行，则各个主机上的数据源名称必须一样。此外，也可以在各个源或建模节点中的“服务器”选项卡上选择另一个数据源。

Netezza 模型 - 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。您还可以设置评分选项的默认值。

图片 6-4
Netezza 模型选项示例



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。

可用于评分。 可在此设置在模型块的对话框上显示的评分选项的默认值。有关这些选项的详细信息，请参阅特定模型块的“设置”选项卡的帮助主题。

Netezza 决策树

决策树是一个代表分类模型的层次结构。使用决策树模型，您可以开发分类系统，以便从一组训练数据来预测或分类未来的观测值。分类采用树结构的形式，其中分支表示分类中的分割点。分割将数据递归分割为子组，直至到达停止点。停止点处的树节点称为**叶片**。每片树叶分配一个标签（称为**类标签**）给其子组或类成员。

模型输出采用树的文本表示形式。文本的每一行对应于一个节点或一片树叶，缩进反映树的层级。对于节点，将显示分割条件；对于树叶，则显示所分配的类标签。

实例权重和类权重

默认情况下，假定所有输入记录和类拥有相同的相对重要性。您可分配单个权重给单个或这两个项目来更改其相对重要性。这样做可能非常有用，例如，如果训练数据中的数据点没有实际分布到各个类别中。权重可以使模型产生偏差，以便弥补那些在数据中没有得到合适描述类别。增加目标值的权重会增加该类别获得正确预测的百分比。

在“决策树”建模节点中，可指定两种类型的权重。**实例权重**分配一个权重给每一行输入数据。在大部分情况下，权重通常指定为 1.0，同时只在那些个案比大部分个案更加重要或更加不重要的情况下分配更大或更小的值，例如：

记录 ID	目标	实例权重
1	drugA	1.1
2	drugB	1.0
3	drugA	1.0
4	drugB	0.3

类权重分配一个权重给目标字段的每个类别，例如：

类	类权重
drugA	1.0
drugB	1.5

可同时使用这两种类型的权重，在这种情况下，它们将相乘并作为实例权重使用。因此，如果将之前的两个示例一起使用，算法将使用以下实例权重。

记录 ID	计算	实例权重
1	1.1*1.0	1.1
2	1.0*1.5	1.5
3	1.0*1.0	1.0
4	0.3*1.5	0.45

Netezza 决策树字段选项

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 6-5
决策树字段选项



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 15 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。

记录 ID。此字段将用作唯一记录标识符。该字段的值对于每个记录必须是唯一的（例如，客户 ID 号）。

实例权重。在此指定一个字段将允许您使用实例权重（每行输入数据的权重），而不是默认的分类权重（目标字段每个类别的权重），或这两个权重。在此指定的字段必须是包含每行输入数据的数值权重的字段。 [有关详细信息，请参阅第 152 页码实例权重和类权重。](#)

预测变量（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。

Netezza 决策树构建选项

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。

图片 6-6
树增长的决策树构建选项



可以设置以下构建选项：

- 树增长
- 类标签权重
- 树修剪

本节将介绍树增长选项。

增长测量。这些选项可控制树增长的测量方式。如果您不想使用默认值，单击自定义并进行更改。

- **杂质测量。**杂质测量，用于评估分割树的最佳位置。**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。

受支持的测量为**熵**（默认）和**基尼**。这些是基于分支的类别归属概率的两种流行杂质测量。

- **最大树深度。**在根节点以下树可以增长到的最大级数（即，递归分割样本的次数）。默认值为 62，此为用于建模的最大允许树深度。但请注意，模型块的查看器最多可以显示 10 个级别。

分割标准。这些选项可控制何时停止分割树。如果您不想使用默认值，单击自定义并进行更改。

- **分割最小改进。**在树中创建新的分割之前，必须减少的最小杂质量。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分割标准所指定的数量来减少杂质，则不会进行此分割。
- **用于分割的最小实例数。**可分割的最小记录数。如果剩余的未分割记录低于此数目，则不会执行进一步分割。您可以使用此字段来防止在树中创建非常小的子组。

Netezza 决策树节点 - 类权重

您可以在此为各个类分配权重。在默认情况下，将为所有类分配值 1，使它们具有相同权重。通过为不同类标签指定不同的数值权重，将引导算法相应地对特定类的训练集进行加权。

图片 6-7
决策树类权重选项



要更改权重，在权重列双击权重并进行所需更改。

值。类标签集源自目标字段的可能值。

权重。分配给特定类的权重。如果为某个类分配较高权重，则模型将对此类比其他类更为敏感。

您可将类权重与实例权重一起使用。 [有关详细信息，请参阅第 152 页码实例权重和类权重。](#)

Netezza 决策树节点 - 树修剪

您可以使用修剪选项来指定决策树的修剪标准。修剪的目的在于，通过去掉那些过度增长而又不会提升新数据的预期精确度的子组，以降低过度拟合的风险。

图片 6-8
决策树修剪选项



修剪测量。默认的修剪测量为精确度，它确保在从树上去掉一个树叶后，模型的估计精确度仍保持在可接受的限度内。如果您要在应用修剪时将类权重考虑在内，则可以使用加权精确度选项。

用于修剪的数据。您可以使用部分或全部训练数据来估计新数据的预期精确度。或者，您还可为此专门使用来自指定表的单独修剪数据集。

- **使用所有训练数据。**此选项（默认）使用所有训练数据来估计模型精确度。
- **使用特定百分比的训练数据来进行修剪。**使用此选项将数据分为两个集合，分别用于训练和修剪，在此处指定修剪数据的百分比。

如果您要指定随机种子，以确保在您每次运行流时，数据以相同方式分区，请选择复制结果。您可以在用于修剪的种子字段中指定一个整数，或单击生成来创建伪随机整数。

- **使用现有表中的数据。**指定用于估计模型精确度的单独修剪数据集的表名称。这种做法被视为比使用训练数据更为可靠。不过，此选项可能导致从删除训练集中去除较大的数据子集，因而会降低决策树的质量。

Netezza K-Means

K-Means 节点实现 k-means 算法，这提供了聚类分析的方法。您可使用该节点来聚类数据集为不同的组。

该算法是基于距离的聚类算法，依赖于距离度量（函数）来测量数据点之间的相似性。根据所使用的距离度量，数据点被指派到与之距离最近的聚类。

算法会执行基本过程相同的几个迭代，并在其中将每个训练实例分配到最近的聚类（相对于适用于实例和聚类中心的指定距离函数）。然后，重新计算所有聚类中心，作为分配给特定聚类实例的平均属性值向量。

Netezza K-Means 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 6-9
K-Means 字段选项



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 15 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。

记录 ID。此字段将用作唯一记录标识符。

预测变量（输入）。选择一个或多个字段作为预测输入。

Netezza K-Means 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。

图片 6-10
K-Means 构建选项



距离测量。用于测量数据点之间距离的方法，更大的距离表示更大的相异性。选项为：

- **欧几里德距离。**（默认）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **Manhattan。**两点之间的距离计算为其坐标之间的绝对差总和。
- **Canberra。**类似于 Manhattan 距离，但对更加靠近原点的点更加敏感。
- **最大值。**两点之间的距离计算为任何坐标尺寸之差的极大值。

聚类数 (k)。指定要创建的聚类数。

最大迭代次数。算法会执行过程相同的几个迭代。使用此选项可在指定的迭代次数后停止模型训练。

重复结果。如果您要设置随机种子，请选中该复选框，这将允许您重复分析。可指定一个整数或单击生成来创建伪随机整数。

Netezza 贝叶斯网络

贝叶斯网络是一种模型，可显示数据集中的变量以及概率，还可以显示这些变量之间的条件和独立性。使用贝叶斯网络节点，可以通过将观察到并记录下的证据与实际常识结合起来构建概率模型，以通过使用表面看上去不相关的属性确定发生的可能性。

Netezza 贝叶斯网络字段选项

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。

对于该节点，目标字段只需用于评分，因此它不会显示在此选项卡上。您可以在类型节点上、该节点的“模型选项”选项卡上或模型块的“设置”选项卡上设置或更改目标。有关详细信息，请参阅第 192 页码 [Netezza 贝叶斯网络块 - 设置选项卡](#)。

图片 6-11
贝叶斯网络字段选项



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 15 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。

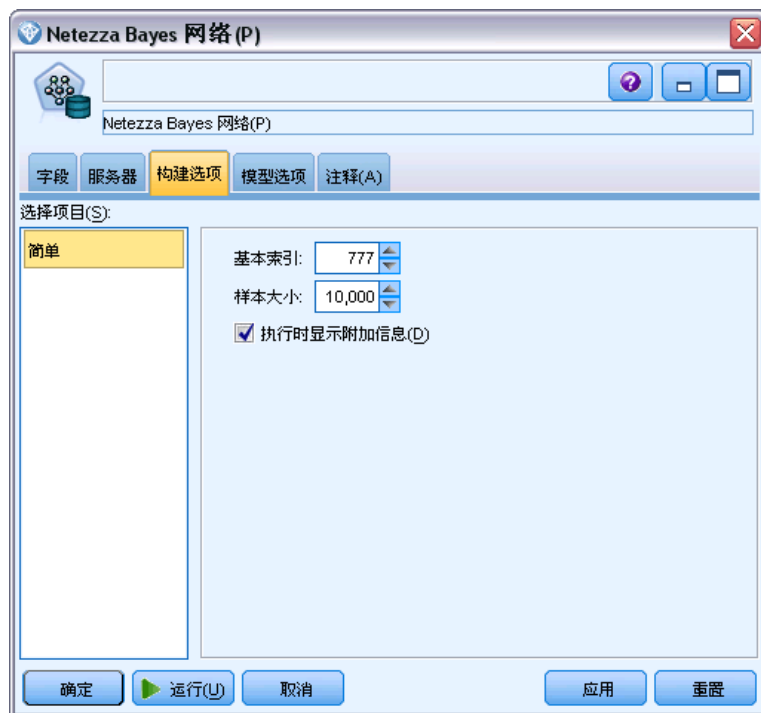
单击全部按钮可以选择列表中的所有字段，或单击单独的测量级别按钮以选择具有此测量级别的所有字段。

预测变量（输入）。选择一个或多个字段作为预测输入。

Netezza 贝叶斯网络构建选项

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。

图片 6-12
贝叶斯网络构建选项



基本索引。为第一个属性（输入字段）分配的数字标识符，以方便内部管理。

样本大小。当属性数量过多并可能导致处理时间过长时，要采用的样本大小。

在执行期间显示更多信息。 如果选中（默认）此复选框，将在消息对话框中显示更多进度信息。

Netezza Naive Bayes

Naive Bayes 是广泛用于处理分类问题的算法。因为该模型将所有给出的预测变量视为互相独立的，因而取名为 naïve。Naive Bayes 是一种快速的、可伸缩的算法，用于计算属性和目标属性组合的条件概率。使用训练数据确定独立的概率。给定来自每个输入变量的所有值分类的发生率，使用此概率可计算出每个目标类的似然值。

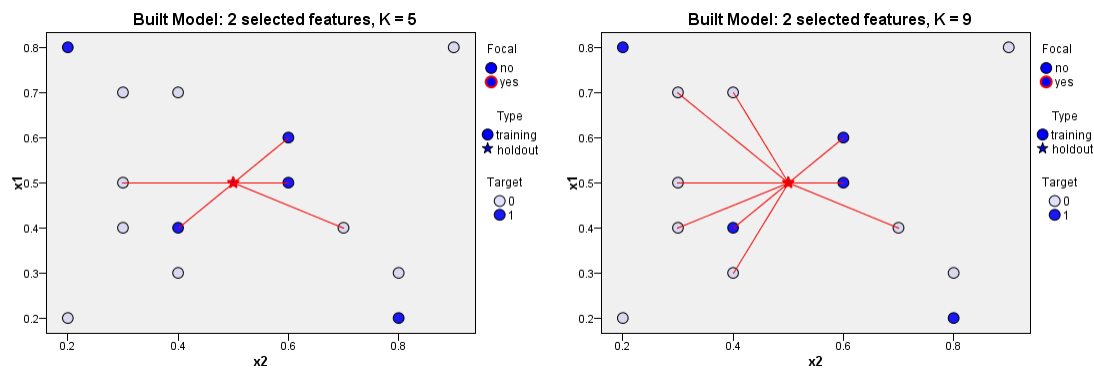
Netezza KNN

“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 6-13
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。

Netezza KNN 模型选项 - 常规

在“模型选项 - 常规”选项卡上，您可以选择是指定模型名称，还是自动生成名称。您还可以设置那些控制如何计算最近相邻元素数量的选项，并设置相关选项以获得增强的模型性能和准确度。

图片 6-14
KNN 常规模型选项



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。

相邻元素

距离测量。 用于测量数据点之间距离的方法，更大的距离表示更大的相异性。选项为：

- **欧几里德距离。**（默认）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **Manhattan。** 两点之间的距离计算为其坐标之间的绝对差总和。
- **Canberra。** 类似于 Manhattan 距离，但对更加靠近原点的点更加敏感。
- **最大值。** 两点之间的距离计算为任何坐标尺寸之差的绝对值。

最近邻元素的数目 (k)。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。

通过选择 k，您可以控制在防止过度拟合（这可能很重要，尤其对于“噪声”数据）和求解（针对类似实例产生不同预测结果）之间的平衡。您通常需要针对每个数据集来调整 k 值，其典型值在 1 至几十之间。

增强性能和准确度

在计算距离之前标准化测量结果。如果选中，该选项将标准化连续输入字段的测量结果，然后再计算距离值。

使用核心集以提升大型数据集性能。如果选中，该选项将针对大型数据集采用核心集抽样以加快计算过程。

Netezza KNN 模型选项 - 评分选项

在“模型选项 - 评分选项”选项卡上，您可以设置评分选项的默认值，并为单独类指定相对权重。

图片 6-15
KNN 常规模型选项



可用于评分

包括输入字段。 指定是否默认在评分中纳入输入字段。

类权重

如果您要更改单独类在构建模型中的相对重要性，请使用此选项。

注意：该选项仅在您使用 KNN 来分类的情况下启用。如果您要执行回归（即，目标字段类型为连续），此选项将被禁用。

在默认情况下，将为所有类分配值 1，使它们具有相同权重。通过为不同类标签指定不同的数值权重，将引导算法相应地对特定类的训练集进行加权。

要更改权重，在**权重**列双击权重并进行所需更改。

值。 类标签集源自目标字段的可能值。

权重。 分配给特定类的权重。如果为某个类分配较高权重，则模型将对此类比其他类更为敏感。

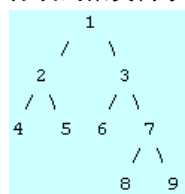
Netezza 分裂式聚类

分裂式聚类是一种聚类分析方法，它通过重复运行算法，使聚类分裂为子聚类，直至达到规定的停止点。

聚类形成从包含全部训练实例（记录）的单个聚类开始。算法的第一次迭代将数据集分为两个子聚类，后续的迭代将进一步细分这些子聚类。停止条件指定为最大迭代次数、数据集细分的最大层级数和用于进一步分区的所需最小实例数。

产生的层次聚类树可以用于将实例从根聚类向下传播，以便对它们进行分类，如下例所示。

图片 6-16
分裂式聚类树示例



在每个层级上，根据从子聚类中心到实例的距离来选择最佳匹配的子聚类。

当通过所应用的层次层级 -1（默认）来对实例评分时，将只返回一个树叶聚类，因为树叶是采用负数来指定。在本例中，这将是聚类 4、5、6、8 或 9 中之一。不过，如果将层次层级设置为 2，则评分会返回根聚类下方第二层级上的聚类（4、5、6 或 7）之一。

Netezza 分裂式聚类字段选项

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 6-17
分裂式聚类字段选项



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 15 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。

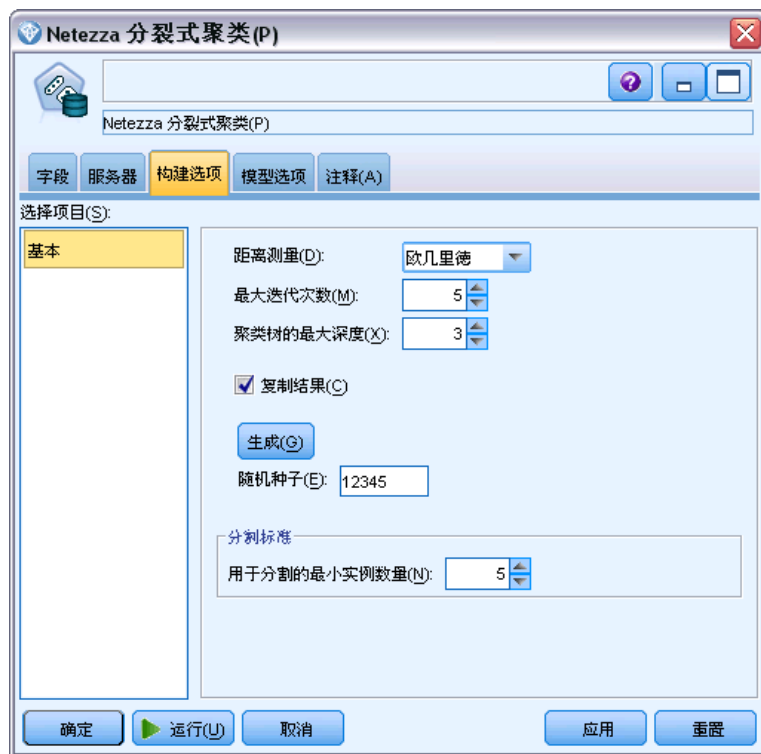
记录 ID。此字段将用作唯一记录标识符。

预测变量（输入）。选择一个或多个字段作为预测输入。

Netezza 分裂式聚类构建选项

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。

图片 6-18
分裂式聚类构建选项



距离测量。用于测量数据点之间距离的方法，更大的距离表示更大的相异性。选项为：

- **欧几里德距离。**（默认）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **Manhattan。**两点之间的距离计算为其坐标之间的绝对差总和。
- **Canberra。**类似于 Manhattan 距离，但对更加靠近原点的点更加敏感。
- **最大值。**两点之间的距离计算为任何坐标尺寸之差的最大值。

最大迭代次数。算法会执行过程相同的几个迭代。使用此选项可在指定的迭代次数后停止模型训练。

聚类树的最大深度。数据集可以细分的最大层级数。

重复结果。如果您要设置随机种子，请选中该复选框，这将允许您重复分析。可指定一个整数或单击生成来创建伪随机整数。

用于分割的最小实例数。可分割的最小记录数。如果剩余的未分割记录低于此数目，则不会执行进一步分割。您可以使用此字段来防止在聚类树中创建非常小的子组。

Netezza PCA

主成分分析 (PCA) 是用于降低数据复杂性的强大数据降维技术。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中的方差，且组合中的各个成分相互正交（不相关）。其目标在于找到有效概括原始输入字段集中的信息的一小部分导出字段（主要成分）。

Netezza PCA 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 6-19
PCA 字段选项



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 15 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。

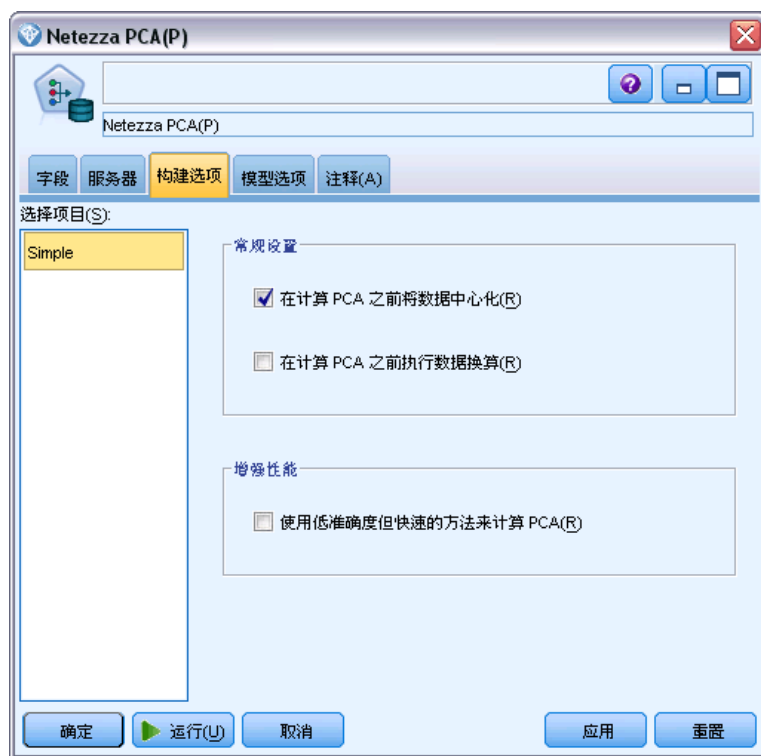
记录 ID。 此字段将用作唯一记录标识符。

预测变量（输入）。 选择一个或多个字段作为预测输入。

Netezza PCA 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。

图片 6-20
PCA 构建选项



在计算 PCA 之前将数据中心化。 如果选中（默认），该选项将在分析之前执行数据中心化（也称为“平均值消去法”）。数据中心化是必要的，它可以确保第一主成分描述最大方差的方向，否则该成分可能更接近于数据的平均值。如果已采用这种方式来准备数据，您通常可以取消选中此选项以提升性能。

在计算 PCA 之前执行数据换算。 该选项将在分析之前执行数据换算。在不同变量采用不同单位时，这样可以降低分析的任意性。作为最简单的形式，可通过将每个变量除以其标准差来执行数据换算。

使用低准确度但快速的方法来计算 PCA。 该选项将导致算法使用低准确度但快速的方法（forceEigensolve）来寻找主成分。

Netezza 回归树

回归树是一种基于树结构的算法，它根据数字目标字段值来重复分割数据集样本，以导出同类子集。与决策树一样，回归树将数据分解为子集，其中树叶对应于足够小或足够均匀的子集。通过选择分割来降低目标属性值的离差，以便采用树叶处的平均值来足够合理地预测它们。

模型输出采用树的文本表示形式。文本的每一行对应于一个节点或一片树叶，缩进反映树的层级。对于节点，将显示分割条件；对于树叶，则显示所分配的类标签。

Netezza 回归树构建选项 - 树增长

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。

图片 6-21
树增长的回归树构建选项



最大树深度。在根节点以下树可以增长到的最大级数（即，递归分割样本的次数）。默认值为 62，此为用于建模的最大允许树深度。但请注意，模型块的查看器最多可以显示 12 个级别。

分割标准。这些选项可控制何时停止分割树。如果您不想使用默认值，单击自定义并进行更改。

- **分割评估尺度。**类杂质测量，用于评估分割树的最佳位置。注意：目前方差是唯一可能的选项。
- **分割最小改进。**在树中创建新的分割之前，必须减少的最小杂质质量。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分割标准所指定的数量来减少杂质，则不会进行此分割。
- **用于分割的最小实例数。**可分割的最小记录数。如果剩余的未分割记录低于此数目，则不会执行进一步分割。您可以使用此字段来防止在树中创建非常小的子组。

Netezza 回归树构建选项 - 树修剪

您可以使用修剪选项来指定回归树的修剪标准。修剪的目的在于，通过去掉那些过度增长而又不会提升新数据的预期精确度的子组，以降低过度拟合的风险。

图片 6-22
树修剪的回归树构建选项



修剪测量。修剪测量确保在从树上去掉一个树叶后，模型的估计精确度仍保持在可接受的限度内。可以选择以下测量之一：

- **mse。**均方误差 - （默认）测量拟合线与数据点的接近程度。

- **r2**。R 平方 - 测量因变量的变动中，由回归模型解释的比例。
- **Pearson**。Pearson 相关系数 - 测量正态分布的线性因变量之间的关系强度。
- **Spearman**。Spearman 相关系数 - 检测根据 Pearson 相关性看起来较弱，但实际可能较强的非线性关系。

用于修剪的数据。您可以使用部分或全部训练数据来估计新数据的预期精确度。或者，您还可为此专门使用来自指定表的单独修剪数据集。

- **使用所有训练数据**。此选项（默认）使用所有训练数据来估计模型精确度。
- **使用特定百分比的训练数据来进行修剪**。使用此选项将数据分为两个集合，分别用于训练和修剪，在此处指定修剪数据的百分比。

如果您要指定随机种子，以确保在您每次运行时，数据以相同方式分区，请选择复制结果。您可以在用于修剪的种子字段中指定一个整数，或单击生成来创建伪随机整数。

- **使用现有表中的数据**。指定用于估计模型精确度的单独修剪数据集的表名称。这种做法被视为比使用训练数据更为可靠。不过，此选项可能导致从删除训练集中去除较大的数据子集，因而会降低决策树的质量。

Netezza 线性回归

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。线性回归模型仅限于直接建模线性关系，但它相对简单，用于评分的数学公式也易于解释。与其他更精练的回归算法产生的模型相比，线性模型快速、高效，并且简单易用，但其应用范围有限。

Netezza 线性回归构建选项

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。

图片 6-23
线性回归构建选项



使用奇异值分解来求解方程。 使用奇异值分解矩阵而不是原始矩阵，不但能够更有效地应对数字误差，并且可以加快计算过程。

在模型中包含截距。 包含截距可以提高解的总体准确性。

计算模型诊断。 该选项将导致在模型上计算大量诊断结果。这些结果存储在矩阵或表中。以供后续查阅。诊断选项包括 r 平方、残差平方和、估计方差、标准差、 p 值和 t 值。

这些诊断与模型的有效性和可用性相关。您应当针对底层数据运行其他诊断，以确保其满足线性假设。

Netezza 时间序列

时间序列 是一个数值序列，以时间上前后接续的（但不必是规律的）点计量 - 例如，每日股票价格或每周销售数据。分析此类数据有时会很有用，例如，用于突显某些行为，像是趋势或季节性变动（一项重复性的模式），或是通过过去的事件预测未来的行为的时候。

Netezza 时间序列支持下列时间序列算法。

- 光谱分析
- 指数平滑法
- 求和自回归移动平均数 (ARIMA)
- 季节性趋势分解

这些算法将时间序列分解成一个趋势和一个季节性组件。再对这些组件进行分析，以构建出一个可用于预测的模型。

光谱分析用于识别时间序列中的周期性行为。对于包含多个底层周期性的时间序列，或数据中存在大量随机噪声，光谱分析提供了最为清晰的手段来判别周期性组件。该方法将序列从时间域转换为频率域，从而能够识别周期性行为的频率。

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。使用指数平滑法，观察本身所造成的影响程度随时间推移而以指数级减少。该方法一次预测一个点，当有新数据进入时再对预测作出调整，对资料的加入、趋势以及季节性变化作出整体性考虑。

ARIMA 模型提供了比指数平滑模型更复杂的方法给趋势和季节性组件建模。这方法包括明确指定自回归阶数和移动平均阶数以及差分次数。

注意：在实际应用上，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，并不试图解释其行为原因。

季节性趋势分解先将周期性行为从时间序列中删除，以便进行趋势分析，之后再为趋势选择一个基本形状，例如一个二次函数。这些基本形状带有若干参数，应为这些参数值确定一个值，以尽量减少平均残差均方误差（即时间序列拟合值与观测值之间的差异）。

Netezza 时间序列值的插值

插值是估算时间序列中缺失的数据并插补一个值的过程。

如果一个时间序列具有规则的时间间隔，只是某些值缺失了，那么这些缺失值可以用线性插值估算出来。考虑如下示例序列中某机场航厦每月的乘客抵达人数。

表 6-1
某航厦的每月抵达人数

月	乘客
3	3,500,000
4	3,900,000
5	-
6	3,400,000
7	4,500,000
8	3,900,000
9	5,800,000
10	6,000,000

在此例中，通过线性插值可以估算出第 5 个月的缺失值为 3,650,000（第 4 个月与第 6 个月的中间点）。

对非规则性的间隔则有不同的处理方法。考虑如下序列中的温度读数。

表 6-2
温度读数

日期	时间	温度
2011-07-24	7:00	57
2011-07-24	14:00	75
2011-07-24	21:00	72
2011-07-25	7:15	59
2011-07-25	14:00	77
2011-07-25	20:55	74
2011-07-27	7:00	60
2011-07-27	14:00	78
2011-07-27	22:00	74

这里，我们有 3 天内从 3 个点所取得的一系列读数，但除了少数读数外，大部分读数的获取时间并不相同。此外，其中只有 2 天是连续的。

该状况可以通过以下两种方法之一来处理：计算汇总值，或者确定出步长值。

汇总值可能是根据对数据的语义上的了解，使用公式计算出来的每日汇总值。执行这一步会得到如下的数据集。

表 6-3
温度读数（汇总）

日期	时间	温度
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	Null
2011-07-27	24:00	72

另外，该算法可以将序列当成一个不同的序列来确定适当的步长值。在此例中，算法所确定的步长值可能是 8 个小时，这样会得到如下结果。

表 6-4
使用步长值计算的温度读数

日期	时间	温度
2011-07-24	6:00	
2011-07-24	14:00	75
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	77
2011-07-25	22:00	
2011-07-26	6:00	
2011-07-26	14:00	
2011-07-26	22:00	

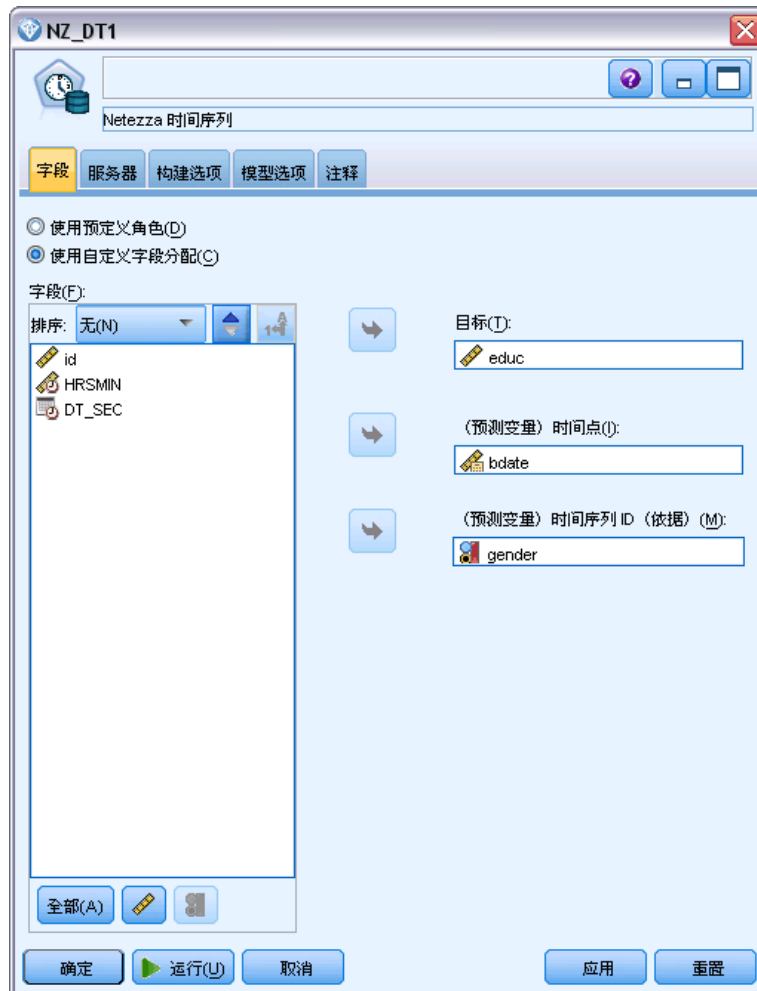
日期	时间	温度
2011-07-27	6:00	
2011-07-27	14:00	78
2011-07-27	22:00	74

在这里，只有 4 个读数与原始测量值对应，但借着原始序列中的其他已知值，缺失的值可再次通过插值计算出来。

Netezza 时间序列字段选项

在“字段”选项卡上，指定源数据输入字段的角色。

图片 6-24
时间序列字段选项



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。有关详细信息，请参阅第 4 章中的测量级别中的 IBM SPSS Modeler 15 源、过程和输出节点。

目标。 选择单个字段作为预测目标。必须是测量级别设为“连续”的字段。

（预测变量）时间点。 （要求）输入字段包含时间序列的日期或时间值。必须是测量级别设为“连续”或“分类”的字段，其数据储存类型为日期、时间、时间戳、或者数值。您在此指定的数据存储类型同时也定义了该建模节点的其他选项卡上的某些字段输入类型。 [有关详细信息，请参阅第 2 章中的设置字段存储类型和格式中的 IBM SPSS Modeler 15 源、过程和输出节点。](#)

（预测变量）时间序列 ID（按）。 包含时间序列 ID 的字段。如果输入项包含一个以上的时间序列，则使用这个字段。

Netezza 时间序列建构选项

建构选项分两个级别：

- 基本 - 设定算法选择、插值以及所采用的时间范围。
- 高级 - 设置预测

本节描述基本选项。

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。

图片 6-25
时间序列基本建构选项



算法

这些是有关所要采用的时间序列算法的设置。

算法名称。 选择您要使用的时间序列算法。可选的算法包括光谱分析、指数平滑法（默认）、ARIMA 或季节趋势分解。有关详细信息，请参阅第 173 页码 [Netezza 时间序列](#)。

趋势。（仅指数平滑法）如果时间序列呈现出一项趋势，则简单的指数平滑法效果不佳。若有趋势，使用该字段来指定它，以使算法可将它纳入考量。

- **系统决定。**（默认）系统尝试为该参数找到最佳值。
- **无(N)。** 时间序列未呈现趋势。
- **加法(A)。** 随着时间推移而稳定增加的趋势。
- **衰减加法(DA)。** 随着时间推移最终会消失的加法趋势。

- **乘法 (M)**。该趋势也是随时间而增加，但速度通常比稳定加法趋势快。
- **衰减乘法 (DM)**。随着时间推移最终会消失的乘法趋势。

季节性。（仅指数平滑法）使用该字段指定时间序列中的数据是否呈现季节性特征。

- **系统决定**。（默认）系统尝试为该参数找到最佳值。
- **无 (N)**。时间序列未呈现季节性模式。
- **加法 (A)**。季节性浮动模式呈现随时间推移稳定上行的趋势。
- **乘法 (M)**。具有与加法季节性相同的特点，但除此之外，其季节性浮动的振幅（高低点间的距离）围绕着总体的上行趋势而上下浮动。

为 ARIMA 采用系统决定的设置。（仅 ARIMA）如果您希望由系统来确定 ARIMA 算法的设置，请选择此选项。

指定。（仅 ARIMA）选择此选项并单击按钮，手动指定 ARIMA 设置。

插值

时间序列源数据有遗缺时，选用一个方法来插入估计值进行填补。[有关详细信息，请参阅第 174 页码 Netezza 时间序列值的插值。](#)

- **线性**。时间序列的间隔有规律，仅仅是某些值缺失时，请选择此方法。
- **指数样条**。把数据值以高速增加或减少的已知点拟合成一条平滑曲线。
- **三次样条**。将已知数据点拟合成一条平滑曲线来估算缺失的值。

时间范围

在此可选择是否使用全范围的时间序列数据，或时间序列数据的一个连续的子集来建立模型。这些字段的输入何为有效，是通过在“字段”选项卡上为时间点指定其字段的数据存储类型来定义的。[有关详细信息，请参阅第 176 页码 Netezza 时间序列字段选项。](#)

- **使用数据中最早与最晚时间**。如果您想要使用全范围的时间序列数据，请选择此选项。
- **指定时间窗口**。如果您希望只使用时间序列的一部分，请选择此选项。使用**最早时间（自）**与**最晚时间（至）**字段来界定边界。

ARIMA 结构

图片 6-26
时间序列的 ARIMA 设置



指定 ARIMA 模型中各种非季节性及季节性组件的值。在每一种情况下，均先将运算符设置为 <（小于）、=（等于）或 <=（小于等于），然后指定相邻字段的值。指定度数的所有值都必须为非负整数。

非季节性。模型中各非季节性组件的值。

- **自相关度 (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价值用于预测当前值。
- **派生 (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。

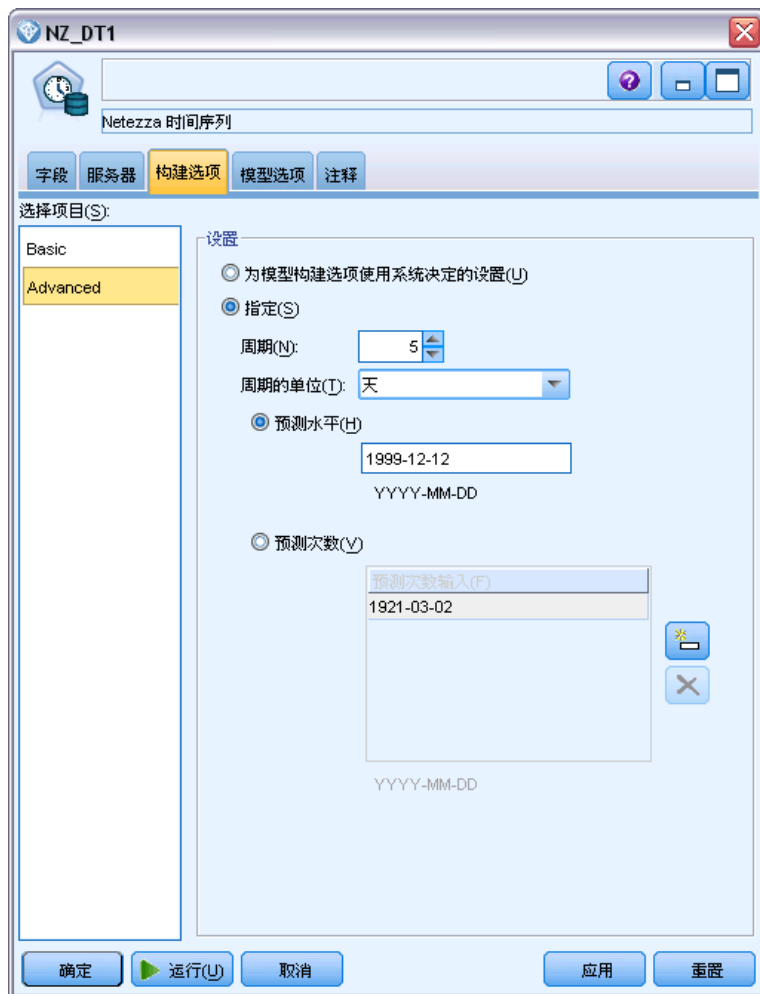
季节性。季节性自相关 (SP)、派生 (SD) 以及移动平均 (SQ) 组件承担与它们的非季节性同类相同的角色。但对于季节性的阶，当前序列值受以前的序列值的影响，序列值之间间隔一个或多个季节性周期。例如，对于月数据（季节性周期为 12），季节性 1 阶表示当前序列值受自当前周期起 12 个周期之前的序列值的影响。因此，对于月数据，指定季节性 1 阶等同于指定非季节性 12 阶。

仅当在数据中检测到季节性趋势时，或您从“高级”选项卡中指定了“周期设置”时，才需用到季节性设置。

Netezza 时间序列建构选项 - 高级

可以使用高级设置来指定预测选项。

图片 6-27
时间序列高级建构选项



模型建构选项采用系统决定的设置值。 如果您希望由系统来作高级设置，请选择此选项。

指定。 如果您希望手动设置高级设置，请选择此选项。（算法为光谱分析时，该选项不可选。）

- **周期/周期单位。** 经过一个时间周期之后，时间序列中的某些特征行为自动重复。例如，对于一个每周销售数字，您可以指定周期为 1，单位为 **星期**。周期必须为非负整数；周期单位可以是毫秒、秒、分、小时、天、星期、季或者年之一。如果未设置周期，或时间类型不为数字，请勿设置周期单位。但是，如果您指定周期，您也必须也指定周期单位。

预测设置。 您可以选择预测到某特定的时间点之前的整个时段，或在某特定时间点上。这些字段的输入何为有效，是通过在“字段”选项卡上为时间点指定其字段的数据存储类型来定义的。有关详细信息，请参阅第 176 页码 [Netezza 时间序列字段选项](#)。

- **预测范围。** 仅当您想要指定预测结束点的时候选择此选项。预测将到此时间点为止。
- **预测时间点。** 选择此选项指定一个或多个时间点，作为预测的时间点。单击添加在时间点的表中增加一行。要删除一行，请选定该行，再单击删除。

Netezza 时间序列模型选项

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。您还可以设置模型输出选项的默认值。

图片 6-28
时间序列模型选项



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。

可用于评分。 可在此设置在模型块的对话框上显示的评分选项的默认值。

- **在输出中包含历史值。**按照默认，模型输出不包含数据的历史值（之前用来进行预测的值）。选择此复选框以包含这些值。
- **在输出中包含插补的值。**如果您选择在输出中包含历史值，且希望同时包含插补的值（如果有的话）时，选择此选框。请注意，插值仅对历史数据起作用，所以如果未选择在输出中包含历史值，则此框不可用。 [有关详细信息，请参阅第 174 页码 Netezza 时间序列值的插值。](#)

Netezza 广义线性

线性回归是一种广为接受的统计方法，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。线性模型由于训练简单、应用方便，在构建真实现象方面用途甚广。然而，线性模型假设因变量（对象）呈正态分布，且自变量（预测变量）对因变量的影响是线性的。

有许多场合都需要用到线性回归，只是不符合上述的假设条件。例如，在对顾客从给定数量的商品中做选择的行为建模时，因变量可能呈多项分布。同样的，当对年龄与收入关系建模时，一般而言，随着年龄增长，收入也会增加，但这二者的关联却不像一条直线那么简单。

针对这些情况，就需要用到广义线性模型。广义线性模型扩展了线性回归模型，使因变量与预测变量之间通过特定的联结函数建立关联，由预测变量选择合适的函数。此外，该模型允许因变量呈非正态分布，例如泊松、二项式等。

算法会以迭代的方式求出最切合的模型，迭代次数可以指定。在计算最佳拟合时，误差由因变量的预测值和实际值之间的差异的平方和来表示。

Netezza 广义线性模型选项 - 常规

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。您可以进行有关模型的多项设置，像是联结函数、输入字段的交互（如果有的话）以及设置评分选项的默认值。

图片 6-29
广义线性常规模型选项



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。

常规设置。 这些设置关系到算法的停止标准。

- **最大迭代次数。** 算法最多进行迭代的次数；最小值为 1 次，默认为 20 次。
- **最大误差 (1e)。** 最大误差值（以科学计数法），达到此值后，算法停止寻找最佳拟合模型。最小值为 0，默认值为 -3，表示 $1E-3$ 或 0.001。
- **不显著误差值的阈值 (1e)。** 设置数值（用科学计数法），低于此值的所有误差均被视为 0 值。最小值为 -1，默认值为 -7，表示误差值若低于 $1E-7$ （或 0.0000001），则被视为不显著。

分布设置。 这些设置关系到因变量（目标）的分布

- **反应变量的分布。** 分布类型；为下列类型之一：伯努利（默认）、高斯、泊松、二项式、负二项分布、Wald（逆高斯）以及伽马。
- **试验。**（在必需的情况下为仅二项分布）当目标响应为发生在一组试验中的事件数时，目标字段包含事件数，“试验”字段包含试验数。例如，试验一种新型杀虫剂，则要使蚂蚁样本接触不同的杀虫剂浓聚物，然后纪录死掉的蚂蚁数目和每种样本中的蚂蚁数目。在这里，记录死掉蚂蚁数目的字段应指定为目标（事件）字段，

记录每种样本中蚂蚁数目的字段应指定为试验字段。对于每条记录，试验次数应为正整数，且大于或等于事件数量。

- **参数。**（仅负二项分布）分布为负二项式时，您可指定参数值。选择是指定一个值，还是使用默认的 -1。

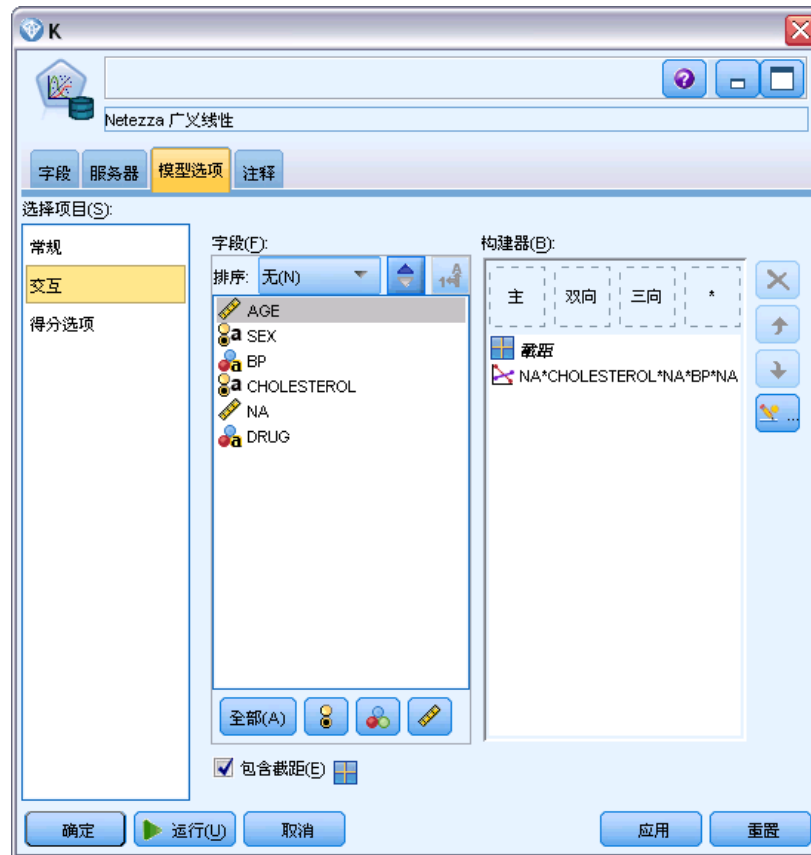
联结函数设置。这些设置与链接函数有关，用以将因变量与预测变量建立关联。

- **联接函数。**可用的函数为下列之一：恒等式、逆、逆否、逆平方、平方根、幂、Oddspower、对数、Clog、重对数、Cloglog、Logit（默认）、概率值、高斯、Cauchit、Canbinom、Cannegbinom。
- **参数。**（仅幂或 Oddspower 这两个联结函数），若联结函数为幂或 Oddspower，您可以指定其参数值。选择是指定一个值，还是使用默认的 1。

Netezza 广义线性模型选项 - 交互

“交互”控制面板包含了选项，可以指定交互行为（即，输入字段间的乘积效应）。

图片 6-30
广义线性模型交互选项



列交互。选择此选项框来指定输入字段的交互性。若无交互行为，请将选项框留空。

要在模型中输入交互，可在源列表中选择一个或多个字段，并拖动至交互列表。所创建的交互类型取决于将选项拖放到何种热点值。

- **主效应。**拖入的字段作为单独的主要交互，显示在交互列表的底部。
- **双向。**所有可能配对的拖入字段作为双向交互，显示在交互列表的底部。
- **三向。**所有可能配成三元组的拖入字段，均作为三向交互，显示在交互列表的底部。
- *****。所拖入的全部字段组合起来，作为单一的交互显示在交互列表的底部。

显示在右侧的按钮让您可以：



要从模型中删除某项，可选取要删除的项，然后单击删除按钮



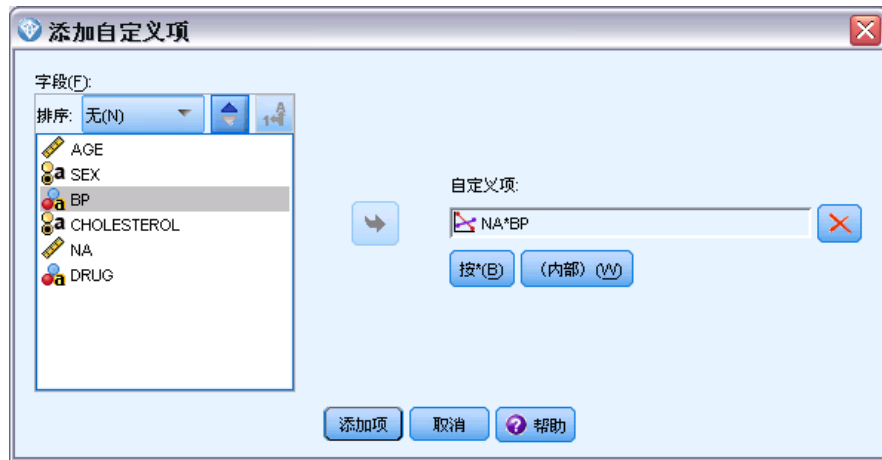
要将模型中的项重新排序，可选取需重新排序的项，然后单击向上或向下箭头



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。

添加自定义项

图片 6-31
添加自定义项对话



您可以 $n_1 \times x_1 \times x_1 \times x_1 \dots$ 形式指定自定义交互。从字段列表中选择一字段，单击右方向箭头按钮将字段添加到自定义项，单击按*，选择下一个字段，再单击右方向箭头按钮，以此类推。当您已完成构建自定义交互，可单击添加项将其返回“交互”面板。

Netezza 广义线性模型选项 - 评分选项

可用于评分。可在此设置在模型块的对话框上显示的评分选项的默认值。 [有关详细信息，请参阅第 202 页码Netezza 广义线性模型块 - 设置选项卡。](#)

- **包括输入字段。**如果您需要将该输入字段连同预测值一同显示在模型输出中，请选择该选框。

管理 IBM Netezza Analytics 模型

IBM® Netezza® Analytics 可以使用与其他 IBM® SPSS® Modeler 模型相同的方式添加 Netezza 模型到工作区和模型选项板中，并以相同的方式来使用。但是，也有几点重大差异，比如 SPSS Modeler 中创建的每个 Netezza Analytics 模型实际引用的是存储在数据库服务器上的模型。因此，要使流正常工作，必须将其连接到创建模型所在的数据库，并且模型表未被外部进程修改。

IBM Netezza Analytics 模型评分

在工作区上使用金色模型块图标来代表模型。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。得分以一个或多个额外数据字段的形式添加，通过将表节点附加到模型块并运行此流分支（如本部分后面所述）以使这些字段可见。某些模型块对话框（例如，决策树或回归树的模型块对话框）还设有“模型”选项卡，其中提供了模型的直观表示。

这些额外字段通过在目标字段的名称上添加前缀 `$<id>-` 以加以区分，其中 `<id>` 取决于模型，并标识所添加的信息类型。在每个模型块的主题中描述了不同的标识符。

要查看得分，按以下步骤操作：

- ▶ 将表节点附加到模型块。
- ▶ 打开表节点。
- ▶ 单击运行。
- ▶ 滚动到表输出窗口的右侧，以查看附加字段及其得分。

Netezza 模型块服务器选项卡

在“服务器”选项卡上，可以设置模型评分的服务器选项。您可以继续使用在上游指定的服务器连接，也可将数据移动到在此指定的其他数据库。

图片 6-32
Netezza 模型块服务器选项示例



Netezza DB Server 详细信息。您可在指定要为模型使用的数据库的连接详细信息。

- **使用上游连接。**（默认）使用上游节点中指定的连接详细信息，例如数据库源节点。注意：该选项仅在所有上游节点能够使用 SQL 回送的情况下有用。在此情况下，无需将数据移出数据库，因为 SQL 完全实现所有的上游节点。
- **移动数据到连接。**将数据移动到您在此处指定的数据库。这样，当数据位于另一个 IBM Netezza 数据库，或其他供应商的数据库，甚至位于平面文件中时，建模仍可工作。此外，如果由于某个节点未执行 SQL 回送而导致数据已被提取，则数据将移回到此处指定的数据库中。单击编辑按钮以浏览并选择连接。警告：IBM® Netezza® Analytics 通常用于大型数据集。在数据库之间传输大量数据，或者从数据库中取出或存入大量数据，可能非常耗时，应尽可能避免。

表名称。用于存储模型的数据库表的名称。此名称仅用作信息说明，无法在此处更改。

Netezza 决策树模型块

决策树模型块显示建模操作的输出，还允许您设置一些选项来为模型评分。

在您运行包含决策树建模节点的流时，该节点会默认添加一个新的字段，其名称将从模型名称导出。

表 6-5
决策树的模型评分字段

新增字段的名称	含义
\$I-model_name	当前记录的预测值。

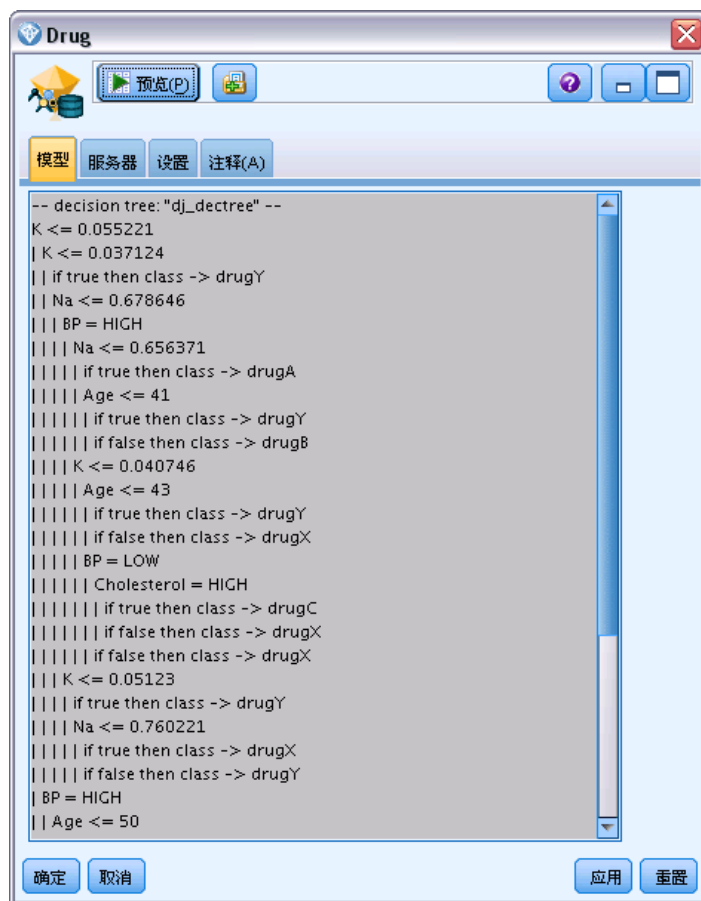
如果您在建模节点或模型块上选择选项计算所分配类用于记录评分的概率，并运行流，则会再添加一个字段。

表 6-6
决策树的模型评分字段 - 更多

新增字段的名称	含义
\$IP-model_name	预测结果的置信度值 (0.0 - 1.0)。

Netezza 决策树块 - 模型选项卡

图片 6-33
决策树模型输出



模型输出采用树的文本表示形式。文本的每一行对应于一个节点或一片树叶，缩进反映树的层级。对于节点，将显示分割条件；对于树叶，则显示所分配的类标签。

Netezza 决策树块 - 设置选项卡

通过“设置”选项卡，可以设置模型评分的某些选项。

包括输入字段。如果选中，该选项将向下游传递所有原始输入字段，并将额外建模字段附加到每行数据。如果您取消选中该复选框，则只会传递记录 ID 字段和额外建模字段，而使流能够更加快速地运行。

计算所分配类用于记录评分的概率。（仅决策树和 Naive Bayes）如果选中，该选项表示附加建模字段包括置信度（即，概率）字段和预测字段。如果您取消选中该复选框，将只生成预测字段。

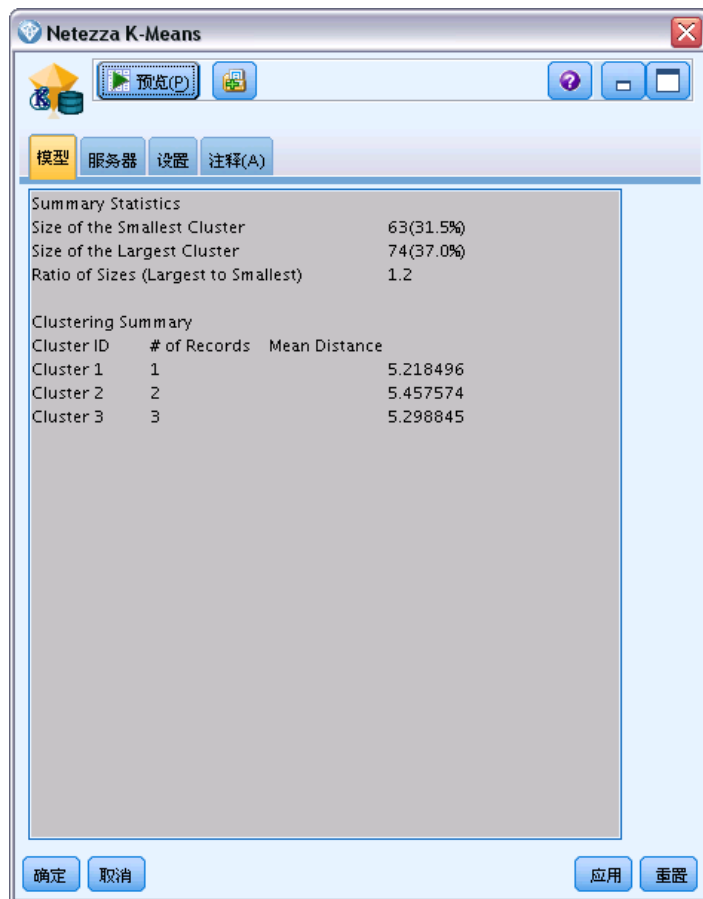
Netezza K-Means 模型块

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 建模节点的流时，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。

Netezza K-Means 块 - 模型选项卡

图片 6-34
K-Means 模型输出



模型输出将显示在“模型”选项卡下，如下所示。

摘要统计量。对于最小和最大的聚类，显示这些聚类拥有的记录数量以及数据集百分比。该列表还显示了最大聚类与最小聚类的比值。

聚类摘要。列出算法生成的聚类。对于每个聚类，该表显示了该聚类中的记录数量，以及这些记录离聚类中心的平均距离。

Netezza K-Means 块 - 设置选项卡

通过“设置”选项卡，可以设置模型评分的某些选项。

包括输入字段。如果选中，该选项将向下游传递所有原始输入字段，并将额外建模字段附加到每行数据。如果您取消选中该复选框，则只会传递记录 ID 字段和额外建模字段，而使流能够更加快速地运行。

距离测量。用于测量数据点之间距离的方法，更大的距离表示更大的相异性。选项为：

- **欧几里德距离**。（默认）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **Manhattan**。两点之间的距离计算为其坐标之间的绝对差总和。
- **Canberra**。类似于 Manhattan 距离，但对更加靠近原点的点更加敏感。
- **最大值**。两点之间的距离计算为任何坐标尺寸之差的绝对值。

Netezza 贝叶斯网络模型块

贝叶斯网络模型块提供了一种设置模型评分选项的方法。

在您运行包含贝叶斯网络建模节点的流时，该节点会添加一个新的字段，其名称将从模型名称导出。

表 6-7
贝叶斯网络的模型评分字段

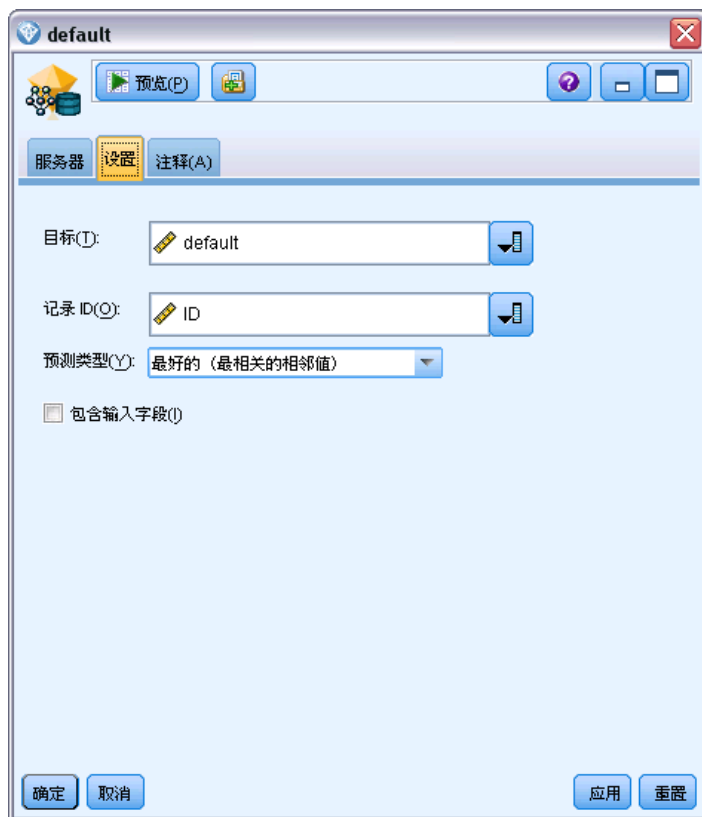
新增字段的名称	含义
\$BN-model_name	当前记录的预测值。

您可以将表节点附加到模型块并运行表节点，以便查看该额外字段。 [有关详细信息，请参阅第 187 页码IBM Netezza Analytics 模型评分。](#)

Netezza 贝叶斯网络块 - 设置选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

图片 6-35
贝叶斯网络模型设置



目标。如果您要对不同于当前目标的目标字段评分，在此选择新的目标。

记录 ID。如果未指定记录 ID 字段，在此选择要使用的字段。

预测类型。您要使用的预测算法的变异：

- **最佳（最相关近邻元素）。**（默认）使用最相关的近邻元素节点。
- **近邻元素（近邻元素的加权预测）。**使用所有近邻元素节点的加权预测。
- **NN-近邻元素（非 Null 近邻元素）**与上一选项相同，不同之处在于它将忽略具有 Null 值的节点（即，对于要计算预测结果的实例，该节点对应的属性存在缺失值）。

包括输入字段。如果选中，该选项将向下游传递所有原始输入字段，并将额外建模字段附加到每行数据。如果您取消选中该复选框，则只会传递记录 ID 字段和额外建模字段，而使流能够更加快速地运行。

Netezza Naive Bayes 模型块

Naive Bayes 模型块提供了一种设置模型评分选项的方法。

在您运行包含 Naive Bayes 建模节点的流时，该节点会默认添加一个新的字段，其名称将从模型名称导出。

表 6-8
Naive Bayes 的模型评分字段 - 默认

新增字段的名称	含义
\$I-model_name	当前记录的预测值。

如果您在建模节点或模型块上选择选项计算所分配类用于记录评分的概率，并运行此流，则会再添加两个字段。

表 6-9
Naive Bayes 的模型评分字段 - 更多

新增字段的名称	含义
\$IP-model_name	实例类的 Bayesian 分子（即，先验类概率与条件实例属性值概率的乘积）。
\$ILP-model_name	后者的自然对数。

您可以将表节点附加到模型块并运行表节点，以便查看这些额外字段。 [有关详细信息，请参阅第 187 页码 IBM Netezza Analytics 模型评分。](#)

Netezza Naive Bayes 块 - 设置选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

包括输入字段。如果选中，该选项将向下游传递所有原始输入字段，并将额外建模字段附加到每行数据。如果您取消选中该复选框，则只会传递记录 ID 字段和额外建模字段，而使流能够更加快速地运行。

计算所分配类用于记录评分的概率。（仅决策树和 Naive Bayes）如果选中，该选项表示附加建模字段包括置信度（即，概率）字段和预测字段。如果您取消选中该复选框，将只生成预测字段。

- **针对较小或严重失衡数据集提高概率准确度。**在计算概率时，该选项将调用 m 估计技术，以避免在估计期间出现零概率。这种类型的概率估计可能速度较慢，但可针对较小或严重失衡数据集提供更好的结果。

Netezza KNN 模型块

KNN 模型块提供了一种设置模型评分选项的方法。

在您运行包含 KNN 建模节点的流时，该节点会添加一个新的字段，其名称将从模型名称导出。

表 6-10
KNN 模型评分字段

新增字段的名称	含义
\$KNN-model_name	当前记录的预测值。

您可以将表节点附加到模型块并运行表节点，以便查看该额外字段。 [有关详细信息，请参阅第 187 页码 IBM Netezza Analytics 模型评分。](#)

Netezza KNN 块 - 设置选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

图片 6-36
KNN 模型设置



距离测量。用于测量数据点之间距离的方法，更大的距离表示更大的相异性。选项为：

- **欧几里德距离。**（默认）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **Manhattan。**两点之间的距离计算为其坐标之间的绝对差总和。
- **Canberra。**类似于 Manhattan 距离，但对更加靠近原点的点更加敏感。
- **最大值。**两点之间的距离计算为任何坐标尺寸之差的极大值。

最近邻元素的数目 (k)。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。

通过选择 k，您可以控制在防止过度拟合（这可能很重要，尤其对于“噪声”数据）和求解（针对类似实例产生不同预测结果）之间的平衡。您通常需要针对每个数据集来调整 k 值，其典型值在 1 至几十之间。

包括输入字段。如果选中，该选项将向下游传递所有原始输入字段，并将额外建模字段附加到每行数据。如果您取消选中该复选框，则只会传递记录 ID 字段和额外建模字段，而使流能够更加快速地运行。

在计算距离之前标准化测量结果。 如果选中，该选项将标准化连续输入字段的测量结果，然后再计算距离值。

使用核心集以提升大型数据集性能。 如果选中，该选项将针对大型数据集采用核心集抽样以加快计算过程。

Netezza 分裂式聚类模型块

分裂式聚类模型块提供了一种设置模型评分选项的方法。

在您运行包含分裂式聚类建模节点的流时，该节点会添加两个新的字段，其名称将从模型名称导出。

表 6-11
分裂式聚类模型评分字段

新增字段的名称	含义
\$DC-model_name	当前记录所分配到的子聚类的标识符。
\$DCD-model_name	从当前记录到子聚类中心的距离。

您可以将表节点附加到模型块并运行表节点，以便查看这些额外字段。 [有关详细信息，请参阅第 187 页码 IBM Netezza Analytics 模型评分。](#)

Netezza 分裂式聚类块 - 设置选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

包括输入字段。 如果选中，该选项将向下游传递所有原始输入字段，并将额外建模字段附加到每行数据。如果您取消选中该复选框，则只会传递记录 ID 字段和额外建模字段，而使流能够更加快速地运行。

距离测量。 用于测量数据点之间距离的方法，更大的距离表示更大的相异性。选项为：

- **欧几里德距离。**（默认）通过将两个点用一条直线连接起来计算得出的两点之间的距离。
- **Manhattan。** 两点之间的距离计算为其坐标之间的绝对差总和。
- **Canberra。** 类似于 Manhattan 距离，但对更加靠近原点的点更加敏感。
- **最大值。** 两点之间的距离计算为任何坐标尺寸之差的绝对值。

所应用的层次层级。 应用于数据的层次层级。

Netezza PCA 模型块

PCA 模型块提供了一种设置模型评分选项的方法。

在您运行包含 PCA 建模节点的流时，该节点会默认添加一个新的字段，其名称将从模型名称导出。

表 6-12
PCA 模型评分字段

新增字段的名称	含义
\$F-model_name	当前记录的预测值。

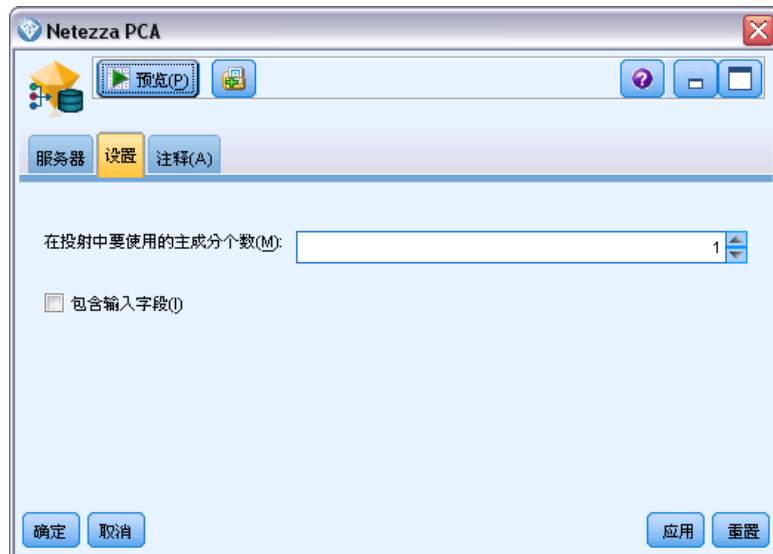
如果您在建模节点或模型块上的主成分个数 ... 字段中指定大于 1 的值，并运行此流，该节点将为每个成分添加一个新的字段。在此情况下，字段名称带有后缀 -n，其中 n 是成分的编号。例如，如果模型名为 pca 且包含三个成分，则新字段将命名为 \$F-pca-1、\$F-pca-2 和 \$F-pca-3。

您可以将表节点附加到模型块并运行表节点，以便查看这些额外字段。 [有关详细信息，请参阅第 187 页码 IBM Netezza Analytics 模型评分。](#)

Netezza PCA 块 - 设置选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

图片 6-37
PCA 模型设置



在投射中要使用的主成分个数。您要用来减小数据集的主成分个数。该值不得超过属性（输入字段）数量。

包括输入字段。如果选中，该选项将向下游传递所有原始输入字段，并将额外建模字段附加到每行数据。如果您取消选中该复选框，则只会传递记录 ID 字段和额外建模字段，而使流能够更加快速地运行。

Netezza 回归树模型块

回归树模型块提供了一种设置模型评分选项的方法。

在您运行包含回归树建模节点的流时，该节点会默认添加一个新的字段，其名称将从模型名称导出。

表 6-13
回归树的模型评分字段

新增字段的名称	含义
\$I-model_name	当前记录的预测值。

如果您在建模节点或模型块上选择选项计算估计方差，并运行流，则会再添加一个字段。

表 6-14
回归树的模型评分字段 - 更多

新增字段的名称	含义
\$IV-model_name	所分配类的估计方差。

您可以将表节点附加到模型块并运行表节点，以便查看这些额外字段。 [有关详细信息，请参阅第 187 页码 IBM Netezza Analytics 模型评分。](#)

Netezza 回归树块 - 模型选项卡

图片 6-38
回归树模型输出

```

-- regression tree: "dj_regtree" --
Time <= 52
| Temperature <= 259
|| Time <= 51
||| if true then class value -> 0
||| Uptime <= 143
||| | Power <= 1050
||| | | Power <= 973
||| | | | if true then class value -> 101
||| | | | if false then class value -> 0
||| | | | if false then class value -> 202
||| | if false then class value -> 0
|| | if false then class value -> 202
| Uptime <= 284
| | Temperature <= 252
| | | Power <= 1084
| | | | Power <= 1080
| | | | | Power <= 1061
| | | | | | Time <= 53
| | | | | | | Temperature <= 251
| | | | | | | | if true then class value -> 101
| | | | | | | | | Power <= 920
| | | | | | | | | if true then class value -> 101
| | | | | | | | | if false then class value -> 0
| | | | | | | | | if false then class value -> 101
| | | | | | | if false then class value -> 202
| | | | | | if false then class value -> 0
| | | | | if false then class value -> 0
  
```


模型输出采用树的文本表示形式。文本的每一行对应于一个节点或一片树叶，缩进反映树的层级。对于节点，将显示分割条件；对于树叶，则显示所分配的类标签。

Netezza 回归树块 - 设置选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

包括输入字段。如果选中，该选项将向下游传递所有原始输入字段，并将额外建模字段附加到每行数据。如果您取消选中该复选框，则只会传递记录 ID 字段和额外建模字段，而使流能够更加快速地运行。

计算估计方差。表示是否应在输出中包含所分配类的方差。

Netezza 线性回归模型块

线性回归模型块提供了一种设置模型评分选项的方法。

在您运行包含线性回归建模节点的流时，该节点会添加一个新的字段，其名称将从模型名称导出。

表 6-15
线性回归的模型评分字段

新增字段的名称	含义
\$LR-model_name	当前记录的预测值。

Netezza 线性回归块 - 设置选项卡

在“设置”选项卡上，可以设置模型评分相关选项。

包括输入字段。如果选中，该选项将向下游传递所有原始输入字段，并将额外建模字段附加到每行数据。如果您取消选中该复选框，则只会传递记录 ID 字段和额外建模字段，而使流能够更加快速地运行。

Netezza 时间序列模型块

通过该模型块，让您可以访问时间序列建模操作的输出项。输出项由下列字段组成。

表 6-16
时间序列模型输出字段

字段	描述
TSID	时间序列的标识符；在建模节点的“字段”选项卡里为时间序列 ID 所指定的字段内容。 有关详细信息，请参阅第 176 页码 Netezza 时间序列字段选项。
时间	当前时间序列内的时间周期。
历史值	数据曾使用过的历史数据值（曾用于预测）。仅当模型块的“设置”选项卡中的在输出中包含历史值选项被选中时，该字段才被包含在内。

字段	描述
\$TS-INTERPOLATED	插补值，仅应用于插补的地方适用。仅当模型块的“设置”选项卡中的在输出中包含插补的值选项被选中时，该字段才被包含在内。插值是建模节点的“建构选项”选项卡中的一个选项。
\$TS-FORECAST	时间序列的预测值。

要查看模型的输出，须将一个表节点（从节点选项板的”输出“选项卡上）附加到模型块上，并运行此表节点。常见的输出看上去如下所列：

图片 6-39
常见的时序序列模型输出

	TSID	TIME	HISTORY	\$TS-INTERPOLATED	\$TS-FORECAST
22	m	1959-11-02	\$null\$	9.810	\$null\$
23	m	1960-07-17	15.000	\$null\$	\$null\$
24	m	1961-05-20	\$null\$	19.591	\$null\$
25	m	1962-07-18	15.000	\$null\$	\$null\$
26	m	1962-08-29	12.000	\$null\$	\$null\$
27	m	1962-12-07	\$null\$	3.401	\$null\$
28	m	1964-06-25	\$null\$	5.399	\$null\$
29	m	1964-11-17	12.000	\$null\$	\$null\$
30	m	1966-01-11	8.000	\$null\$	\$null\$
31	m	1967-07-31	\$null\$	\$null\$	0.590
32	m	1969-02-16	\$null\$	\$null\$	0.719
33	m	1970-09-04	\$null\$	\$null\$	0.667
34	m	1972-03-23	\$null\$	\$null\$	0.619
35	m	1973-10-10	\$null\$	\$null\$	0.574
36	m	1975-04-28	\$null\$	\$null\$	0.532
37	m	1976-11-14	\$null\$	\$null\$	0.494
38	m	1978-06-03	\$null\$	\$null\$	0.458
39	m	1979-12-20	\$null\$	\$null\$	0.425
40	m	1981-07-08	\$null\$	\$null\$	0.394
41	m	1983-01-25	\$null\$	\$null\$	0.366

Netezza 时间序列块 - “设置”选项卡

在“设置”选项卡中，您可以指定选项来自订模型输出。

模型名称。 模型名称，如建模节点上的“模型选项”选项卡中所定义。

其他选项与建模节点的“模型选项”选项卡上的相同。

Netezza 广义线性模型块

通过该模型块，让您可以访问建模操作的输出项。

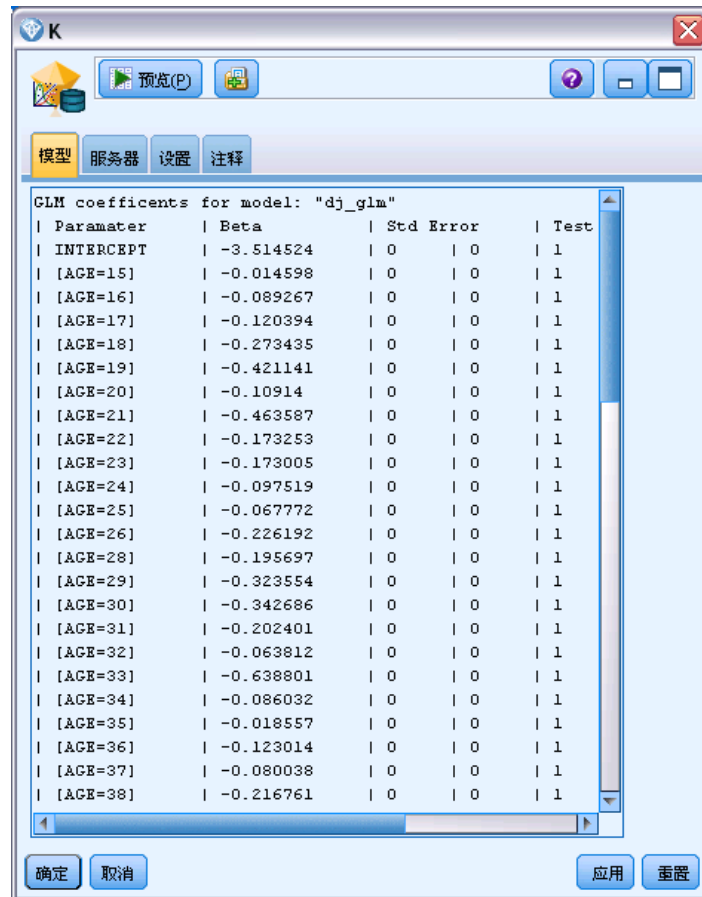
在您运行包含广义线性建模节点的流时，该节点会添加一个新的字段，其名称是从模型名称派生。

表 6-17
广义线性的模型评分字段

新增字段的名称	含义
\$GLM-model_name	当前记录的预测值。

“模型”选项卡显示各种与模型有关的统计量。

图片 6-40
广义线性模型输出



输出项由下列字段组成。

表 6-18
广义线性模型输出字段

输出字段	描述
参数	模型使用的参数（即，预测变量）。这些是数值和名义列，以及截距（回归模型中的常数项）。
Beta	相关系数（即，模型的线性成分）。
标准误	Beta 的标准差。
Test	用于评估参数有效性的检验统计量。
p 值	假定参数显著时，误差的概率。

输出字段	描述
残差汇总	
残差类型	显示汇总值的预测残差类型。
RSS	残差的值。
df	残差的自由度。
p 值	误差的概率。高值表示拟合度差的模型；低值表示拟合度佳。

Netezza 广义线性模型块 - 设置选项卡

在“设置”选项卡中，您可以自订模型的输出。

该选项与建模节点上的计分选项中所显示的相同。 [有关详细信息，请参阅第 187 页码 Netezza 广义线性模型选项 - 评分选项。](#)

注意事项

这些信息开发用于在全球提供的产品和服务。

IBM 可能在其他国家/地区中不提供在本文档中讨论的产品、服务或功能。请咨询您当地的 IBM 代表以了解有关您所在地区当前可用产品和服务的信息。任何对 IBM 产品、程序或服务的引用，并不意味着仅可使用这些 IBM 产品、程序或服务。作为替代，可以使用任何功能相当的产品、程序或服务，前提是不侵犯任何 IBM 知识产权。不过，用户应自行负责评估和验证任何非 IBM 产品、程序或服务的工作情况。

在本文档中介绍的主题可能涉及 IBM 的专利或申请中的专利。您获得该文档并不意味着授予您任何这些专利许可。您可以将书面许可查询函件发送至：

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY
10504-1785, U. S. A.

要查询双字节字符集 (DBCS) 相关许可证信息，请联系所在国家/地区中的 IBM 知识产权部门，或者以书面形式将查询函件发送至：

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此类条款与当地法律不符的其他国家/地区： INTERNATIONAL BUSINESS MACHINES 公司“按原样”提供本出版物，不保证任何明示或暗示，包括但不限于对非侵权性、适销性或对特定用途适用性的暗示担保。有些州/省不允许针对某些交易的明示或暗示免责条款，因此本声明可能不适用于您。

本信息可能含有技术误差或排版错误。此处的信息会定期进行更改；这些更改会体现在本出版物的新版本中。IBM 可能随时对本出版物中所述的产品和/或程序进行改进和/或更改，恕不另行通知。

在本信息材料中对任何非 IBM 网站的引用仅为了方便用户，并不以任何方式表明对这些网站的认可。这些网站上的材料并非本 IBM 产品材料的一部分，您对这些网站的使用需自担风险。

IBM 可以自认为适当并且不会对您构成任何约束的任何方式使用或分发您提供的任何信息。

如果本程序的受许可方试图了解有关程序的信息以启用：(i) 在独立创建的程序和其他程序（包括本程序）之间交换信息；(ii) 相互使用交换的信息，则应联系：

IBM Software Group, Attention:Licensing, 233 S. Wacker Dr., Chicago, IL 60606,
USA.

此类信息的提供应遵照相关条款和条件，其中包括在某些情况下支付适当费用。

在本文档中介绍的受许可保护程序，及其所有受许可保护材料由 IBM 在双方签署的“IBM 客户协议”、“IBM 国际程序许可证协议”或任何其他等同协议下提供。

此处所含的性能数据均在受控环境下决定。因此，在其他操作环境中获得的结果可能差异较大。有些测量可能在开发级的系统中进行，不保证这些测量结果与常用系统上的测量结果相同。此外，有些测量结果可能通过推断来估计得出。实际结果可能有所差异。此文档的用户应针对其具体环境验证适用的数据。

非 IBM 产品的相关信息来自这些产品的供应商，及其发布的公告或其他公开来源。IBM 并未测试过这些产品，无法确认有关非 IBM 产品的性能准确性、兼容性或任何其他声明。有关非 IBM 产品功能的问题应由这些产品的供应商负责。

有关 IBM 未来方向或意向的所有声明均可能未经通知即变更或撤销，并且仅代表目标和目的。

本信息包含日常业务运营中使用的数据和报告的示例。为了尽可能详尽地对其进行说明，示例中包含了人员的姓名、公司、品牌和产品的名称。所有这些名称均为虚构，与真实商业企业使用的名称和地址的任何雷同纯属巧合。

如果您正在查阅此信息的软拷贝，照片和彩色插图可能不会显示。

商标

IBM、IBM 徽标、ibm.com 和 SPSS 是 IBM Corporation 的商标，在全球许多司法辖区注册。有关最新的 IBM 商标列表，请访问网页 <http://www.ibm.com/legal/copytrade.shtml>。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国、其他国家或地区或两者的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国、其他国家或地区或两者的商标。

UNIX 是 The Open Group 在美国和其他国家或地区的注册商标。

Java 和所有基于 Java 的商标和徽标是 Sun Microsystems, Inc. 在美国、其他国家或地区或两者的商标。

其他产品和服务名称可能是 IBM 或其他公司的商标。



索引

- Adaptive Bayes Network
 - Oracle Data Mining, 56 - 58
- Analysis Services
 - 与 IBM SPSS Modeler 集成, 7, 14
 - 决策树, 40
 - 示例, 40
 - 管理模型, 19
- Apriori
 - Microsoft, 27
 - Oracle Data Mining, 73 - 74, 76
- ARIMA 模型
 - IBM Netezza Analytics, 174, 180
- DB2
 - 管理模型, 103
- DSN
 - 配置, 15
- epsilon
 - Oracle Support Vector Machine, 61
- hostname
 - Oracle 连接, 50
- IBM
 - Kohonen 聚类建模, 94
 - Logistic 回归建模, 94
 - Naive Bayes 建模, 94
 - 人口统计聚类建模, 94
 - 关联建模, 94
 - 决策树建模, 94
 - 回归建模, 94
 - 多项回归建模, 94
 - 序列建模, 94
 - 时间序列建模, 94
 - 管理模型, 103
 - 线性回归建模, 94
- IBM InfoSphere Warehouse (ISW)
 - 与 IBM SPSS Modeler 集成, 7
- IBM Netezza Analytics, 143
 - K-Means, 158
 - K-Means 字段选项, 158
 - K-Means 构建选项, 159
 - K-均值模型块, 190 - 191
 - KNN 模型块, 194 - 195
 - KNN 模型选项, 163 - 164
 - Naive Bayes, 162
 - Naive Bayes 模型块, 193 - 194
 - PCA 字段选项, 168
 - PCA 构建选项, 169
 - PCA 模型块, 196 - 197
 - 主成分分析 (PCA), 168
 - 决策树, 152
 - 决策树字段选项, 153
 - 决策树构建选项, 154 - 156
 - 决策树模型块, 188 - 190
 - 分裂式聚类, 165
 - 分裂式聚类字段选项, 166
 - 分裂式聚类构建选项, 167
 - 分裂式聚类模型块, 196
 - 回归树, 170
 - 回归树构建选项, 170 - 171
 - 回归树模型块, 197 - 199
 - 字段选项, 149
 - 广义线性, 183
 - 广义线性模型块, 200, 202
 - 广义线性模型选项, 183, 185
 - 时间序列, 173
 - 时间序列字段选项, 176
 - 时间序列建构选项, 177, 180
 - 时间序列模型块, 199 - 200
 - 时间序列模型选项, 182
 - 最近相邻元素 (KNN), 162
 - 模型选项, 151
 - 管理模型, 187
 - 线性回归, 172
 - 线性回归构建选项, 172
 - 线性回归模型块, 199
 - 贝叶斯网络, 160
 - 贝叶斯网络字段选项, 160
 - 贝叶斯网络构建选项, 161
 - 贝叶斯网络模型块, 192
 - 配置 IBM SPSS Modeler, 143 - 144, 147, 150
- IBM SPSS Modeler, 1
 - 数据库挖掘, 8
 - 文档, 3
- IBM SPSS Modeler Solution Publisher
 - Oracle Data Mining 模型, 52
- InfoSphere Warehouse (IBM), 参阅 ISW, 95
- InfoSphere Warehouse Data Mining
 - 关联建模, 110
 - 决策树, 108
 - 分类系统, 114
 - 回归节点, 119
 - 序列节点, 117
 - 模型块, 133
 - 示例流, 136
- ISW
 - ODBC 连接, 95
 - 与 IBM SPSS Modeler 集成, 95
 - “服务器”选项卡, 105
- K-Means
 - IBM Netezza Analytics, 158 - 159, 190 - 191
 - Oracle Data Mining, 69 - 71
- KNN 模型
 - IBM Netezza Analytics, 194 - 195
- Logistic 回归
 - 专家选项, 27
 - 服务器选项, 20
 - 模型选项, 21
 - 评分 - 服务器选项, 34
 - 评分 - 汇总选项, 36
- Logistic 回归节点
 - InfoSphere Warehouse Data Mining, 129
- MDL, 56

索引

- Microsoft
 - Analysis Services, 13, 15, 33
 - Logistic 回归, 13
 - Logistic 回归建模, 15, 33
 - Naive Bayes 建模, 13, 15, 33
 - 关联规则建模, 13, 15, 33
 - 决策树建模, 13, 15, 33
 - 序列聚类, 13
 - 神经网络, 13
 - 神经网络建模, 15, 33
 - 管理模型, 19
 - 线性回归, 13
 - 线性回归建模, 15, 33
 - 聚类建模, 13, 15, 33
- Microsoft Analysis Services, 36 - 37, 39 - 40
 - 与 IBM SPSS Modeler 集成, 14
- Microsoft SQL Server
 - 与 IBM SPSS Modeler 集成, 14
- naive bayes
 - 专家选项, 24
 - 服务器选项, 20
 - 模型选项, 21
 - 评分 - 服务器选项, 34
 - 评分 - 汇总选项, 36
- Naive Bayes
 - IBM Netezza Analytics, 162, 193
 - InfoSphere Warehouse Data Mining, 128
 - Oracle Data Mining, 55 - 56
- Naive Bayes 模型
 - IBM Netezza Analytics, 194
 - Oracle Adaptive Bayes Network, 58
- NMF
 - Oracle Data Mining, 71 - 73
- 0-Cluster
 - Oracle Data Mining, 68 - 69
- ODBC
 - IBM Netezza Analytics 配置, 143 - 144, 147, 150
 - 为 Oracle 配置, 48 - 49, 51 - 52
 - 配置, 15
 - 配置 ISW, 95
 - 配置 SQL Server, 16
- ODM. 请参阅 Oracle Data Mining, 48
- Oracle Data Miner, 84
 - 与 IBM SPSS Modeler 集成, 7
- Oracle Data Mining, 48
 - Adaptive Bayes Network, 56 - 58
 - Apriori, 73 - 74, 76
 - K-Means, 69 - 71
 - Naive Bayes, 55 - 56
 - NMF, 71 - 73
 - 0-Cluster, 68 - 69
 - Support Vector Machine, 59, 61
 - 一致性检验, 82
 - 决策树, 65 - 67
 - 准备数据, 86
 - 属性重要性 (AI), 78 - 80
 - 广义线性模型 (GLM), 62 - 65
 - 最小描述符长度 (MDL), 77 - 78
 - 示例, 87 - 90, 93
 - 管理模型, 81 - 83
 - 误分类成本, 83
 - 配置 IBM SPSS Modeler, 48 - 49, 51 - 52
- PCA 模型
 - IBM Netezza Analytics, 168 - 169, 196 - 197
- SID
 - Oracle 连接, 50
- Solution Publisher
 - Oracle Data Mining 模型, 52
- SPSS Modeler Server, 2
- SQL Server, 20, 34, 36
 - ODBC 连接, 16
 - 与 IBM SPSS Modeler 集成, 14
 - 配置, 15
 - SQL 优化。请参阅 SQL 生成, 7
 - SQL 回送。请参阅 SQL 生成, 7
 - SQL 生成, 7, 9
- Support Vector Machine
 - Oracle Data Mining, 59, 61
- SVM. 请参阅 Support Vector Machine, 59
- tnsnames.ora 文件, 50
- z 得分
 - 标准化数据, 60, 86
- 不纯度度量
 - Netezza 决策树, 155
- 事务处理格式数据
 - ISW 关联节点, 111
- 交互验证
 - Oracle Naive Bayes, 55
- 优化
 - SQL 生成, 7
- 修剪的 Naive Bayes 模型
 - Oracle Adaptive Bayes Network, 58
- 先验概率
 - Oracle Data Mining, 62
- 光谱分析, IBM Netezza Analytics, 174
- 关联建模
 - InfoSphere Warehouse Data Mining, 110
- 关联规则
 - 专家选项, 28
 - 服务器选项, 20
 - 模型选项, 21
 - 评分 - 服务器选项, 34
 - 评分 - 汇总选项, 36

- 关联规则模型
 - Microsoft, 27
- 关键字
 - 模型关键字, 11
- 决策树
 - IBM Netezza Analytics, 152 - 156, 188 - 190
 - Microsoft Analysis Services, 13, 15, 33
 - Oracle Data Mining, 65 - 67
 - 专家选项, 22
 - 服务器选项, 20
 - 模型选项, 21
 - 评分 - 服务器选项, 34
 - 评分 - 汇总选项, 36
- 决策树模型
 - InfoSphere Warehouse Data Mining, 108
- 分割标准
 - Oracle k-Means, 70
- 分区, 112
 - 模型构建, 30, 55, 57, 79, 113, 118, 121, 125, 128 - 129
 - 选择, 112
- 分区字段
 - 选择, 75
- 分区数据, 75
- 分类系统
 - InfoSphere Warehouse Data Mining, 114
- 分裂式聚类
 - IBM Netezza Analytics, 165 - 167, 196
- 功能选项
 - ISW Data Mining, 106
- 单临界值
 - Oracle Naive Bayes, 56
- 单功能模型
 - Oracle Adaptive Bayes Network, 58
- 双临界值
 - Oracle Naive Bayes, 56
- 发布者节点
 - Oracle Data Mining 模型, 52
- 吉尼杂质测量, 155
- 唯一的字段
 - Oracle Adaptive Bayes Network, 58
 - Oracle Apriori, 66, 76
 - Oracle Data Mining, 52
 - Oracle k-Means, 70
 - Oracle MDL, 78
 - Oracle Naive Bayes, 55
 - Oracle NMF, 72
 - Oracle O-Cluster, 68
 - Oracle Support Vector Machine, 60
- 商标, 204
- 回归树
 - IBM Netezza Analytics, 170 - 171, 197 - 199
- 回归节点
 - InfoSphere Warehouse Data Mining, 119
- 复杂度罚分, 22 - 28, 30
- 复杂性因子
 - Oracle Support Vector Machine, 61
- 多功能模型
 - Oracle Adaptive Bayes Network, 58
- 字段选项
 - IBM Netezza Analytics, 149, 153, 158, 160, 166, 168 - 169, 176
 - 建模节点, 111
- 季节性趋势分解, IBM Netezza Analytics, 174
- 实例权重, 在 Netezza 树模型中, 152
- 导出
 - Analysis Services 模型, 40
 - DB2 模型, 105
- 属性重要性 (AI)
 - Oracle Data Mining, 78 - 80
- 广义线性模型
 - IBM Netezza Analytics, 183, 185 - 187, 200, 202
- 广义线性模型 (GLM)
 - Oracle Data Mining, 62 - 65
- 序列聚类
 - 模型选项, 21
- 序列聚类 (Microsoft), 31
 - 专家选项, 33
 - 字段选项, 32
- 序列节点
 - InfoSphere Warehouse Data Mining, 117
- 应用程序示例, 3
- 建模节点
 - ISW 的数据库内建模, 95
 - Microsoft Logistic 回归, 18
 - Microsoft Naive Bayes, 18
 - Microsoft 关联规则, 18
 - Microsoft 决策树, 18
 - Microsoft 序列聚类, 18
 - Microsoft 时间序列, 18
 - Microsoft 神经网络, 18
 - Microsoft 线性回归, 18
 - Microsoft 聚类, 18

索引

- 数据库内建模, 9, 13, 15, 18, 33
- 成本
 - Oracle, 54
- 指数平滑法
 - IBM Netezza Analytics, 174
- 探索, 41, 88, 137
- 插值, IBM Netezza Analytics 时间序列, 174
- 收敛容差
 - Oracle Support Vector Machine, 61
- 数据审核节点, 41, 88, 137
- 数据库
 - ISW 的数据库内建模, 95
 - 数据库内建模, 9, 13, 15, 18, 33
 - 数据库内建模, 36
 - Analysis Services, 7
 - IBM InfoSphere Warehouse (ISW), 7
 - Oracle Data Miner, 7
 - 数据库建模
 - Analysis Services, 7
 - IBM InfoSphere Warehouse (ISW), 7
 - IBM Netezza Analytics, 143 - 144, 147, 150
 - Oracle, 48 - 49, 51 - 52
 - Oracle Data Miner, 7
 - 数据库挖掘
 - 优化选项, 9
 - 使用 IBM SPSS Modeler, 8
 - 数据准备, 9
 - 构建模型, 9
 - 示例, 40, 136
 - 配置, 15
- 文档, 3
- 时间序列
 - IBM Netezza Analytics, 176 - 177, 180, 182
 - InfoSphere Warehouse Data Mining, 129 - 132
 - 时间序列 (IBM Netezza Analytics), 173, 199 - 200
 - 时间序列 (Microsoft), 28
 - 专家选项, 30
 - 模型选项, 29
 - 设置选项, 31
- 最小值-最大值
 - 标准化数据, 60, 86
- 最小描述符长度, 56
- 最小描述符长度 (MDL)
 - Oracle Data Mining, 77 - 78
- 最近相邻元素模型
 - IBM Netezza Analytics, 162 - 164, 194 - 195
- 服务器
 - 运行 Analysis Services, 20, 34, 36
- “服务器”选项卡
 - ISW, 105
- 杂志度量
 - Oracle Apriori, 66
- 构建选项
 - IBM Netezza Analytics, 154 - 156, 159, 161, 167, 170 - 172, 177, 180
- 标准化数据
 - Oracle 模型, 86
- 标准化方法
 - Oracle k-Means, 70
 - Oracle NMF, 72
 - Oracle Support Vector Machine, 60
- 标准差
 - Oracle Support Vector Machine, 61
- 树叶, 在 Netezza 树模型中, 152
- 模型
 - 一致性问题, 11
 - 保存, 10
 - 列出 DB2, 103
 - 导出, 10
 - 数据库内模型的构建, 9
 - 浏览 DB2, 104
 - 浏览 Oracle, 57
 - 管理 Analysis Services, 19
 - 管理 DB2, 103
 - 评估, 43, 90, 139
 - 评分的数据内模型, 9
- 模型块
 - IBM Netezza Analytics, 188 - 200, 202
 - InfoSphere Warehouse Data Mining, 133
- 模型评分
 - InfoSphere Warehouse Data Mining, 102
- 模型选项
 - IBM Netezza Analytics, 151, 163 - 164, 182 - 183, 185
- 法律注意事项, 203
- 流
 - InfoSphere Warehouse Data Mining 示例, 136
- 熵杂质测量, 155
- 生成节点, 40
- 示例
 - 应用程序指南, 3
 - 数据库挖掘, 40 - 41, 43, 46, 88, 136 - 137, 139, 141
 - 概述, 5
- 神经网络
 - 专家选项, 26
 - 服务器选项, 20
 - 模型选项, 21
 - 评分 - 服务器选项, 34

- 评分 - 汇总选项, 36
- 离散化数据
 - Oracle 模型, 86
- 端口
 - Oracle 连接, 50
- 类别编辑器
 - ISW 关联节点, 116
- 类权重, 在 Netezza 树模型中, 152
- 类标签, 在 Netezza 树模型中, 152
- 线性回归
 - IBM Netezza Analytics, 170, 172, 199
 - 专家选项, 25
 - 服务器选项, 20
 - 模型选项, 21
 - 评分 - 服务器选项, 34
 - 评分 - 汇总选项, 36
- 线性核函数
 - Oracle Support Vector Machine, 59
- 聚类
 - IBM Netezza Analytics, 196
 - InfoSphere Warehouse Data Mining, 124
 - 专家选项, 23
 - 服务器选项, 20
 - 模型选项, 21
 - 评分 - 服务器选项, 34
 - 评分 - 汇总选项, 36
- 聚类数
 - Oracle k-Means, 70
 - Oracle O-Cluster, 68
- 聚类节点
 - InfoSphere Warehouse Data Mining, 124
- 节点
 - 生成, 40
- 表格数据
 - ISW 关联节点, 111
- 评估, 43, 90, 139
- 评分, 9, 187
- 误分类成本
 - Oracle, 54
 - 决策树, 54, 107
- 贝叶斯网络模型
 - IBM Netezza Analytics, 160 - 161, 192
- 距离函数
 - Oracle k-Means, 70
- 部署, 46, 93, 141
- 高斯核函数
 - Oracle Support Vector Machine, 59