Руководство по исследованию данных в базе данных IBM SPSS Modeler 16

Примечание

Прежде чем использовать эту информацию и продукт, описанный в ней, прочтите сведения в разделе “Уведомления” на стр. 115.

Информация о продукте

Это издание применимо к версии 16, выпуску 0, модификации 0 IBM(r) SPSS(r) и ко всем последующим выпускам и модификациям до тех пор, пока в новых изданиях не будет указано иное.

Содержание

Предисловие

IBM® SPSS Modeler - это IBM Corp. инструментальная среда масштаба предприятия для анализа данных. SPSS Modeler помогает организациям улучшить взаимосвязи с клиентами и отдельными лицами, обеспечивая глубокое понимание данных. Организации используют приобретенные с помощью SPSS Modeler глубокие знания для сохранения выгодных заказчиков, обнаружения возможностей дополнительных покупок, привлечения новых клиентов, обнаружения ошибок, сокращения рисков и улучшений в обеспечении государственных служб.

Наглядный интерфейс SPSS Modeler дает пользователям возможность применить свой конкретный опыт в бизнесе, что способствует разработке более мощных предсказывающих моделей и сокращает время принятия решения. SPSS Modeler предлагает много способов моделирования, таких как алгоритмы предсказания, классификации, сегментации и ассоциативного обнаружения. Когда моделей IBM SPSS Modeler Solution Publisher поддерживает их распространение на уровне организации для принимающих решение сотрудников или для применения к базе данных.

О бизнес аналитике IBM

Программное обеспечение IBM для бизнес аналитики предоставляет полную, последовательную и точную информацию, которая повышает эффективность ведения бизнеса. Полный набор программного обеспечения для business intelligence, прогностической аналитики, управления финансовой эффективностью и стратегией и аналитических приложений позволяет ясно видеть текущую ситуацию, а также делать прогнозы, позволяющие предпринимать практические действия. В сочетании с решениями для конкретных отраслей, проверенной практикой и услугами бизнес аналитика IBM позволяет организациям любых размеров достигать наивысшей производительности, уверенно автоматизировать процессы принятия решений и добиться лучших результатов.

Как составная часть этого набора, программное обеспечение IBM SPSS Predictive Analytics помогает организациям предсказывать будущие события и предпринимать практические действия непосредственно на основе этих предсказаний. Коммерческие, правительственные и научные организации всего мира, полагаются на технологию IBM SPSS, обеспечивающую конкурентное преимущество в привлечении, удержании клиентов и повышения отдачи от них при уменьшении доли ошибочных решений и сокращении рисков. Включая программное обеспечение IBM SPSS в свои ежедневные операции, организации могут прогнозировать будущие события, направлять и автоматизировать решения для соответствия бизнес-целям и достигать ощутимых конкурентных преимуществ. Чтобы получить дальнейшую информацию или связаться с представителем, зайдите на http://www.ibm.com/spss.

Техническая поддержка

Техническая поддержка предоставляется клиентам, оплачивающим обновительные взносы. Пользователи могут обращаться в службу технической поддержки, если у них возникают какие-либо проблемы с использованием или установкой программного обеспечения IBM Corp.. К службе технической поддержки можно вызывать через сайт IBM Corp. по адресу http://www.ibm.com/support. При обращении за поддержкой будьте готовы назвать себя и организацию, в которой вы работаете.

Глава 1. О программе IBM SPSS Modeler

IBM SPSS Modeler - это комплект инструментов исследования данных, при помощи которого можно быстро разрабатывать прогнозные модели, использующие деловые знания и опыт, и внедрять их в деловые операции для усовершенствования процесса принятия решений. Разработанный на основе модели промышленного стандарта CRISP-DM, IBM SPSS Modeler поддерживает весь процесс исследования данных, от обработки исходных данных до получения лучших деловых результатов.

IBM SPSS Modeler предлагает ряд методов моделирования, взятых из таких областей, как обучение машин, искусственный интеллект и статистика. При помощи методов, доступных на палитре Моделирование, можно извлечь новую информацию из данных и разработать прогнозные модели. У каждого из методов есть свои сильные стороны и типы задач, для решения которых он лучше всего подходит.

SPSS Modeler можно приобрести как отдельный продукт или использовать как клиент в сочетании с сервер SPSS Modeler. Кроме того, доступен ряд дополнительных возможностей, сводка которых дается в следующих разделах. Дополнительную информацию смотрите в разделе http://www.ibm.com/software/analytics/spss/products/modeler/.

Продукты IBM SPSS Modeler

В семейство продуктов IBM SPSS Modeler и связанные с этим семейством программы входят следующие продукты:

- IBM SPSS Modeler
- сервер IBM SPSS Modeler
- IBM SPSS Modeler Administration Console
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- Адаптеры сервер IBM SPSS Modeler для IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

SPSS Modeler - это полнофункциональная версия продукта, устанавливаемая и запускаемая на персональном компьютере. SPSS Modeler можно запустить в локальном режиме, как автономный продукт, или в распределенном режиме вместе с сервер IBM SPSS Modeler, чтобы повысить производительность на больших наборах данных.

Используя SPSS Modeler, можно быстро и интуитивно строить точные прогнозные модели, не прибегая к программированию. Используя уникальный визуальный интерфейс, можно легко визуализировать процесс анализа данных. В продукт встроены расширенные функции аналитики, при поддержке которых можно обнаруживать в данных скрытые структуры и тенденции. Можно моделировать результаты и выяснять, какие факторы на них влияют, чтобы полностью использовать деловые возможности и ограничивать риски.

SPSS Modeler доступен в двух версиях: SPSS Modeler Professional и SPSS Modeler Premium. Дополнительную информацию смотрите в разделе “Выпуски IBM SPSS Modeler” на стр. 2.

сервер IBM SPSS Modeler

SPSS Modeler пользуется архитектурой клиент - сервер, чтобы распределять требования ресурсоемких операций по мощным серверным программам, что повышает производительность для больших наборов данных.

сервер SPSS Modeler - это отдельно лицензируемый продукт, который непрерывно работает в режиме распределенного анализа на хосте сервера совместно с одной или несколькими установками IBM SPSS Modeler. При этом сервер SPSS Modeler обеспечивает высокую производительность для больших наборов данных, поскольку ресурсоемкие операции можно выполнять на сервере без скачивания данных на компьютер клиента. Кроме того, сервер IBM SPSS Modeler обеспечивает поддержку для возможностей оптимизации SQL и моделирования в базе данных, что дает дополнительный выигрыш в производительности и автоматизации.

IBM SPSS Modeler Administration Console

Modeler Administration Console - это графическая программа для управления многочисленными опциями конфигурации сервер SPSS Modeler, который также можно конфигурировать посредством файла опций. Эта прикладная программа содержит консольный пользовательский интерфейс для отслеживания и конфигурирования установок сервер SPSS Modeler installations, and is available free-of-charge сервер SPSS Modeler. Эту прикладную программу можно установить только на компьютерах Windows; однако она может управлять сервером на любой поддерживаемой платформе.

IBM SPSS Modeler Batch

Хотя обычно исследование данных - интерактивный процесс, можно также запустить SPSS Modeler из командной строки, не открывая графический интерфейс. Например, у вас могут быть продолжительные или повторяющиеся задачи, которые желательно выполнить без участия пользователя. SPSS Modeler Batch - это особая версия продукта, предоставляющая поддержку всех аналитических возможностей SPSS Modeler без вызова обычного пользовательского интерфейса. Для использования SPSS Modeler Batch требуется лицензия сервер SPSS Modeler.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher - это инструмент, при помощи которого можно создать пакетную версию потока SPSS Modeler; такую версию можно запускать внешним механизмом времени выполнения или встроить во внешнюю прикладную программу. Этим способом можно публиковать и внедрять полные потоки SPSS Modeler для использования в средах, где SPSS Modeler не установлен. SPSS Modeler Solution Publisher распространяется в составе службы IBM SPSS Collaboration and Deployment Services - Scoring, для которой требуется отдельная лицензия. С этой лицензией вы получаете модуль времени выполнения SPSS Modeler Solution Publisher, при помощи которого можете запускать опубликованные потоки.

Адаптеры сервер IBM SPSS Modeler для IBM SPSS Collaboration and Deployment Services

Для IBM SPSS Collaboration and Deployment Services доступен ряд адаптеров, при посредстве которых SPSS Modeler и сервер SPSS Modeler могут взаимодействовать с репозиторием IBM SPSS Collaboration and Deployment Services. При этом поток SPSS Modeler, внедренный в репозиторий, доступен для совместного использования несколькими пользователями или для обращения из прикладной программы IBM SPSS Modeler Advantage тонкого клиента. Адаптер устанавливается в той системе, в которой находится репозиторий.

Выпуски IBM SPSS Modeler

SPSS Modeler доступен в следующих выпусках.

SPSS Modeler Professional

SPSS Modeler Professional содержит все инструменты, необходимые для работы с большинством типов структурированных данных, таких как трассировка поведения и взаимодействия в системах CRM, демографии, поведения покупателей и данных о продажах.

SPSS Modeler Premium

SPSS Modeler Premium - это отдельно лицензируемый продукт, расширяющий SPSS Modeler Professional для работы с такими специальными данными, как данные в аналитике объектов или социальных сетях, и с неструктурированными текстовыми данными. SPSS Modeler Premium состоит из следующих компонентов.

IBM SPSS Modeler Entity Analytics добавляет дополнительное измерение к прогностической аналитике IBM SPSS Modeler. Прогностическая аналитика пытается предсказать будущее поведение данных из прошлого, а объектная аналитика направлена на улучшение связности и согласованности текущих данных посредством устранения конфликтов идентичности в самих записях. Идентичность может относиться к индивидууму, организации, а также к любому другому объекту, для которого возможна неоднозначность. Разрешение идентичности может оказаться крайне необходимым для ряда полей, в том числе для управления отношениями с клиентами, обнаружения мошенничества, противодействия отмыванию денег или для национальной и международной безопасности.

IBM SPSS Modeler Social Network Analysis преобразует информацию о взаимосвязях в поля, характеризующие социальное поведение отдельных лиц и групп. При помощи данных, описывающих взаимосвязи, в основе которых лежат социальные сети, IBM SPSS Modeler Social Network Analysis определяет социальных лидеров, влияющих на поведение других участников сети. Кроме того, вы можете определить, какие люди наиболее подвержены влиянию других участников сети. Сочетая полученные результаты с результатами других измерений, можно создать исчерпывающие профили отдельных лиц, на которых будут основаны ваши прогнозные модели. Модели, содержащие эту социальную информацию, выполняются лучше моделей, которые ее не содержат.

IBM SPSS Modeler Text Analytics использует новейшие лингвистические технологии и обработку естественного языка (NLP) для быстрой обработки самых разнообразных неструктурированных текстовых данных, для извлечения и организации ключевых понятий и группирования этих понятий в категории. Извлеченные понятия и категории можно сочетать с существующими структурированными данными, такими как демографические, и применять к моделированию при помощи полного комплекта инструментов исследования данных IBM SPSS Modeler для получения более качественных и специализированных решений.

Документация IBM SPSS Modeler

Документация в формате электронной справки доступна через меню Справка SPSS Modeler. Сюда входит документация по SPSS Modeler, сервер SPSS Modeler и SPSS Modeler Solution Publisher, а также Руководство по прикладным программам и другие сопроводительные материалы.

Полная документация по каждому продукту (включая указания по установке) доступна в формате PDF в подпапках *\Documentation* каждого продукта DVD. Документы по установке также можно скачать с веб-сайта по адресу http://www-01.ibm.com/support/docview.wss?uid=swg27038316.

Кроме того, документация в обоих этих форматах доступна в Информационном центре SPSS Modeler по адресу http://publib.boulder.ibm.com/infocenter/spssmodl/v16r0m0/.

Документация SPSS Modeler Professional

В комплект документации SPSS Modeler Professional (включая указания по установке) входят:

- **Руководство пользователя IBM SPSS Modeler.** Общее введение в использование SPSS Modeler, в том числе о создании потоков данных, обработке пропущенных значений, построению выражений CLEM, работе с проектами и отчетами и составлению пакетов потоков для внедрения в IBM SPSS Collaboration and Deployment Services, прогнозирующие прикладные программы или IBM SPSS Modeler Advantage.
- **Узлы источников, обработки и вывода IBM SPSS Modeler.** Описания всех узлов, служащих для чтения, обработки и вывода данных в различных форматах. По существу это все узлы, кроме узлов моделирования.
- **Узлы моделирования IBM SPSS Modeler.** Описания всех узлов, служащих для создания моделей исследования данных. IBM SPSS Modeler предлагает ряд методов моделирования, взятых из таких областей, как обучение машин, искусственный интеллект и статистика.

- **Руководство по алгоритмам IBM SPSS Modeler.** Описание математических основ методов моделирования, используемых в IBM SPSS Modeler. Это руководство доступно только в формате PDF.
- **Руководство по прикладным программам IBM SPSS Modeler.** Примеры в этом руководстве служат кратким специализированным введением к тем или иным методам и технологиям моделирования. Это руководство доступно также в электронном виде в меню Справка. Дополнительную информацию смотрите в разделе “Примеры прикладных программ”.
- **Сценарии и автоматизация IBM SPSS Modeler.** Информация об автоматизации системы путем создания сценариев, включая сценарии свойств, которые могут использоваться для работы с узлами и потоками.
- **Руководство по внедрению IBM SPSS Modeler .** Информация о выполнении IBM SPSS Modeler потоков и сценариев как шагов обработки заданий под управлением IBM SPSS Collaboration and Deployment Services Deployment Manager.
- **Руководство разработчика IBM SPSS Modeler CLEF .** CLEF предоставляет возможности интеграции с программами других производителей, таких как подпрограммы обработки данных или алгоритмы моделирования, как с узлами в IBM SPSS Modeler.
- **Руководство по исследованию данных в базе данных IBM SPSS Modeler.** Информация о том, как использовать мощности вашей базы данных для повышения производительности и расширения диапазона возможностей анализа с привлечением алгоритмов от сторонних производителей.
- **Руководство администратора и руководство по производительности сервер IBM SPSS Modeler .** Информация о том, как сконфигурировать и администрировать сервер IBM SPSS Modeler.
- **Руководство пользователя по консоли администратора IBM SPSS Modeler .** Информация об установке и использовании пользовательского интерфейса консоли для мониторинга и конфигурирования сервер IBM SPSS Modeler. Консоль реализована как подключаемый модуль прикладной программы Диспетчер развертывания .
- **Руководство по CRISP-DM IBM SPSS Modeler.** Пошаговое руководство к использованию методологии CRISP-DM для исследования данных SPSS Modeler.
- **Руководство пользователя IBM SPSS Modeler Batch .** Полное руководство по использованию IBM SPSS Modeler в пакетном режиме, включая подробности выполнения в пакетном режиме и аргументы командной строки. Это руководство доступно только в формате PDF.

Документация SPSS Modeler Premium

В комплект документации SPSS Modeler Premium (включая указания по установке) входят:

- **IBM SPSS Modeler Entity Analytics Руководство пользователя.** Информация об использовании аналитики объектов совместно с SPSS Modeler, в том числе по установке и конфигурированию репозитория, узлам аналитики объектов и задачам управления.
- **IBM SPSS Modeler Social Network Analysis Руководство пользователя.** Руководство по выполнению анализа социальной сети совместно с SPSS Modeler, включая анализ групп и анализ распространения.
- **Руководство пользователя SPSS Modeler Text Analytics .** Информация об использовании аналитики текстов совместно с SPSS Modeler, в том числе по узлам исследования текстов, интерактивной инструментальной среде, шаблонам и другим ресурсам.
- **Руководство пользователя по консоли администратора IBM SPSS Modeler Text Analytics.** Информация об установке и использовании пользовательского интерфейса консоли для мониторинга и конфигурирования сервер IBM SPSS Modeler для использования совместно с SPSS Modeler Text Analytics . Консоль реализована как подключаемый модуль прикладной программы Диспетчер развертывания .

Примеры прикладных программ

Инструменты исследования данных в SPSS Modeler помогают разрешить широкий спектр деловых и организационных проблем, а примеры прикладных программ предоставляют краткие, целевые введения в конкретные методы и способы моделирования. Используемые здесь наборы данных намного меньше огромных складов данных, которыми управляют некоторые исследователи данных, но применяемые понятия и методы должны масштабироваться до реальных прикладных программ.

Обратиться к примерам можно, выбрав **Примеры прикладных программ** в меню Справка в SPSS Modeler. Файлы данных и потоки примеров устанавливаются в папке *Demos* в каталоге установки продукта. Дополнительную информацию смотрите в разделе “Папка demos”.

Примеры моделирования баз данных. Смотрите эти примеры в руководстве *IBM SPSS Modeler In-Database Mining Guide*.

Примеры сценариев. Смотрите эти примеры в руководстве *IBM SPSS Modeler Scripting and Automation Guide*.

Папка demos

Файлы данных и примеры потоков, используемые с примерами прикладных программ, устанавливаются в папке *Demos* в каталоге установки продукта. К этой папке можно также обратиться из группы программ IBM SPSS Modeler в меню Пуск Windows или, щелкнув по *Demos* в списке недавно использовавшихся каталогов в диалоговом окне Открыть файл.

Глава 2. Исследование в базе данных

Обзор моделирования баз данных

сервер IBM SPSS Modeler поддерживает интеграцию с инструментами исследования и моделирования данных, доступными у поставщиков баз данных, в том числе IBM Netezza, IBM DB2 InfoSphere Warehouse, Oracle Data Miner и Microsoft Analysis Services. Построение моделей, их скоринг и сохранение в базе данных - все эти операции возможны в прикладной программе IBM SPSS Modeler. Это позволяет сочетать аналитические возможности и легкость использования IBM SPSS Modeler с мощностью и производительностью базы данных, реализуя одновременно преимущества собственных алгоритмов баз данных, предоставляемых этими поставщиками. Построение моделей выполняется в базе данных, которые можно затем просмотреть и оценить в интерфейсе IBM SPSS Modeler обычным способом, а в случае необходимости - внедрить при помощи IBM SPSS Modeler Solution Publisher. Поддерживаемые алгоритмы находятся на палитре Моделирование баз данных в IBM SPSS Modeler.

При использовании IBM SPSS Modeler для обращения к собственным алгоритмам баз данных реализуется несколько преимуществ:

- Зачастую алгоритмы In-Database полностью интегрируются с сервером баз данных и могут обеспечить улучшенную производительность.
- Модели, встроенные и хранящиеся в базе данных (“In-Database”), можно проще внедрить в любую обращающуюся к ней прикладную программу и совместно использовать с этой прикладной программой.

генерирование SQL. Моделирование In-Database отличается от генерирования SQL, иначе называемого обратным переносом SQL (“SQL Pushback”). Эта функциональная возможность позволяет генерировать операторы SQL для собственных операций IBM SPSS Modeler, которые могут быть "перенесены обратно" в базу данных (то есть выполнены в ней) с целью улучшения производительности. Например, узлы слияния, агрегации, и выбора - все генерируют код SQL, для которого возможен обратный перенос в базу данных этим способом. Применяя генерирование SQL в сочетании с моделированием баз данных, можно получить потоки, обрабатываемые в базе данных от начала до конца, что приведет к значительному росту производительности потоков, запускаемых в IBM SPSS Modeler.

Примечание: Моделирование баз данных и оптимизация SQL требуют, чтобы на компьютере IBM SPSS Modeler была включена возможность соединения с сервером IBM SPSS Modeler. При включенной возможности соединения можно обращаться к алгоритмам баз данных, выполнять обратный перенос SQL непосредственно с клиента сервером IBM SPSS Modeler и обращаться к серверу IBM SPSS Modeler. Проверьте текущее состояние лицензии, для чего в меню клиента сервером IBM SPSS Modeler выберите:

Справка > О программе > Дополнительные подробности

Если возможность соединения включена, на вкладке Состояние лицензии вы увидите опцию **Разрешение для сервера**.

Информацию о поддерживаемых алгоритмах смотрите в следующих разделах для конкретных поставщиков.

Что требуется

Для выполнения моделирования баз данных требуется следующее установленное программное обеспечение:

- Соединение ODBC с соответствующей базой данных с требуемыми установленными аналитическими компонентами (Microsoft Analysis Services, Oracle Data Miner или IBM DB2 InfoSphere Warehouse).
- В IBM SPSS Modeler надо включить моделирование баз данных в диалоговом окне Вспомогательные прикладные программы (**Инструменты > Вспомогательные прикладные программы**).

- Надо включить опции **Генерировать SQL** и **Оптимизация SQL** в диалоговом окне Пользовательские опции в IBM SPSS Modeler, а также на сервер IBM SPSS Modeler (если он используется). Обратите внимание на то, что оптимизация SQL не строго обязательна для моделирования баз данных, но рекомендуется по причинам производительности.

Примечание: Моделирование баз данных и оптимизация SQL требуют, чтобы на компьютере IBM SPSS Modeler была включена возможность соединения с сервером IBM SPSS Modeler. При включенной возможности соединения можно обращаться к алгоритмам баз данных, выполнять обратный перенос SQL непосредственно с клиента сервером IBM SPSS Modeler и обращаться к серверу IBM SPSS Modeler. Проверьте текущее состояние лицензии, для чего в меню клиента сервером IBM SPSS Modeler выберите:

Справка > О программе > Дополнительные подробности

Если возможность соединения включена, на вкладке Состояние лицензии вы увидите опцию **Разрешение для сервера**.

Более подробную информацию смотрите в следующих разделах для конкретных поставщиков.

Построение модели

Процесс построения и оценки моделей с использованием алгоритмов базы данных аналогичен другим типам исследования данных в IBM SPSS Modeler. Общий процесс работы с узлами и "слепками" моделирования похож на любой другой поток при работе в IBM SPSS Modeler. Единственная разница состоит в том, что фактическая обработка данных и построение модели переданы в базу данных.

Поток моделирования базы данных принципиально не отличается от других потоков в IBM SPSS Modeler, но этот поток выполняет все операции в базе данных. К таким операциям относится, например, построение модели с использованием узла дерева решений Microsoft. При запуске этого потока IBM SPSS Modeler передает в базу данных инструкции на построение и сохранение итоговой модели, а подробные данные скачиваются в IBM SPSS Modeler. Факт выполнения в базе данных обозначается затененными фиолетовыми узлами в потоке.

Подготовка данных

Независимо от того, используются ли собственные алгоритмы базы данных, результаты подготовки данных надо при первой возможности передавать в базу данных для повышения производительности.

- Если исходные данные хранятся в базе данных, необходимо сохранять их там, убедившись, что все нужные операции обратного потока можно преобразовать в SQL. Это предотвратит скачивание данных на IBM SPSS Modeler и проявление эффекта узкого горла, который мог бы свести на нет все выгоды, и позволит выполнять весь поток в базе данных.
- Если исходные данные *не* хранятся в базе данных, моделирование базы данных все равно можно использовать. В этом случае подготовка данных проводится в IBM SPSS Modeler, и подготовленный набор данных автоматически закачивается в базу данных для построения модели.

Оценка модели

Модели, сгенерированные в IBM SPSS Modeler с использованием исследования данных в базе данных, отличаются от обычных моделей IBM SPSS Modeler. Хотя эти модели появляются в менеджере моделей как сгенерированные “слепки” моделей, на самом деле они представляют собой удаленные модели, хранящиеся на удаленном сервере баз данных или исследования данных. В IBM SPSS Modeler выводятся просто ссылки на эти удаленные модели. Другими словами, модель IBM SPSS Modeler, которую вы видите, - это “пустая” модель, содержащая такую информацию, как имя хоста сервера баз данных, имя базы данных и имя модели. Это отличие важно понимать при просмотре и оценке моделей, созданных с использованием собственных алгоритмов базы данных.

После создания модели вы можете добавить ее в поток для оценки, как и любую другую сгенерированную в IBM SPSS Modeler модель. Вся оценка производится в базе данных, даже если нет операций обратного

потока. (Операции обратного потока могут использоваться для базы данных, если можно повысить производительность, но они не требуются для проведения оценки). В большинстве случаев вы можете посмотреть сгенерированную модель, используя стандартный браузер поставщика базы данных.

Для просмотра и оценки требуется соединение в реальном времени с сервером, где запущены Oracle Data Miner, IBM DB2 InfoSphere Warehouse или Microsoft Analysis Services.

Просмотр результатов и задание параметров

Для просмотра результатов и задания параметров оценки дважды щелкните по модели на холсте потока. Можно также щелкнуть правой кнопкой мыши по модели и выбрать опцию **Просмотр** или **Изменить**. Конкретные параметры зависят от типа модели.

Экспорт и сохранение моделей баз данных

Модели и сводки баз данных можно экспортировать из браузера моделей таким же образом, как и другие модели, созданные в IBM SPSS Modeler, используя опции в меню Файл.

1. В меню Файл браузера моделей выберите любую из следующих опций:

- **Экспорт текста** - экспортирует сводку модели в текстовый файл
- **Экспорт HTML** - экспортирует сводку модели в файл HTML
- **Экспорт PMML** (поддерживается только для моделей IBM DB2 IM) - экспортирует модели на языке разметки предсказательных моделей (predictive model markup language, PMML), и эти модели можно использовать с другим поддерживающим PMML программным обеспечением.

Примечание: Сгенерированные модели можно также сохранить, выбрав в меню Файл опцию **Сохранить узел**.

Согласованность модели

Для каждой сгенерированной модели базы данных IBM SPSS Modeler хранит описание структуры модели, а также ссылку на модель с тем же именем, которая хранится в базе данных. На вкладке Сервер сгенерированной модели выводится уникальный ключ, созданный для этой модели, который соответствует действующей модели в базе данных.

IBM SPSS Modeler использует этот произвольно сгенерированный ключ для проверки, что модель по-прежнему согласована. Этот ключ сохраняется в описании модели при ее построении. Перед запуском потока внедрения рекомендуется проверить, что эти ключи совпадают.

1. Чтобы проверить согласованность модели, хранимой в базе данных, сравнивая ее описание со случайным ключом, хранимым в IBM SPSS Modeler, нажмите кнопку **Проверить**. Если модель базы данных не удается найти или ключ не совпадает, выводится сообщение об ошибке.

Просмотр и экспорт сгенерированного SQL

Сгенерированный код SQL можно просмотреть перед выполнением, что может быть полезным для отладки.

Глава 3. Требования для интеграции с Microsoft Analysis Services

IBM SPSS Modeler и Microsoft Analysis Services

IBM SPSS Modeler поддерживает интеграцию с Microsoft SQL Server Analysis Services. Эта функциональная возможность реализована как узлы моделирования в IBM SPSS Modeler и доступна из палитры моделирования базы данных. Если эта палитра не показана, ее можно активировать, включив интеграцию MS Analysis Services на вкладке Microsoft в диалоговом окне Вспомогательные прикладные программы. Дополнительную информацию смотрите в разделе “Включение интеграции с Analysis Services” на стр. 13.

IBM SPSS Modeler поддерживает интеграцию следующих алгоритмов Analysis Services:

- деревья решений
- кластеризация
- правила связывания
- наивная модель Байеса
- линейная регрессия
- нейросеть
- логистическая регрессия
- временные ряды
- кластеризация последовательностей

На следующей диаграмме показан поток данных от клиента на сервер, когда исследованием данных в базе данных управляет сервер IBM SPSS Modeler. Построение модели выполняется с помощью Analysis Services. Полученная модель сохраняется Analysis Services. Ссылка на эту модель обрабатывается в потоках IBM SPSS Modeler. Затем эта модель скачивается из Analysis Services для оценки или на Microsoft SQL Server, или в IBM SPSS Modeler.

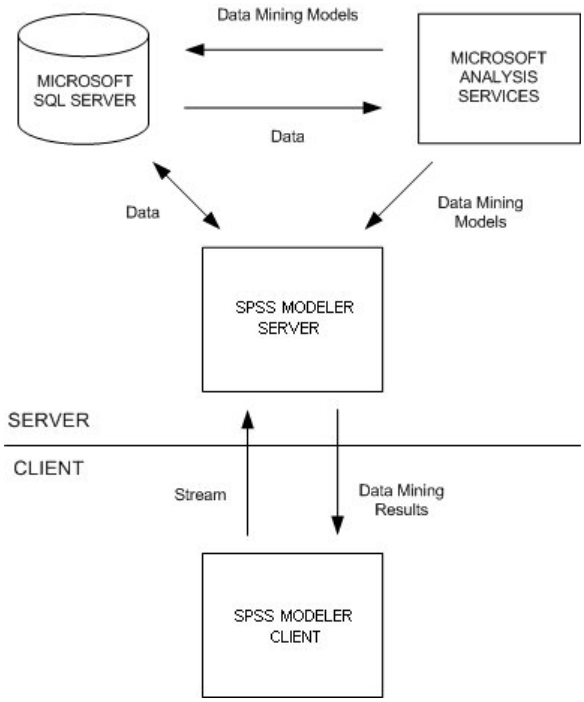

Рисунок 1. Поток данных между IBM SPSS Modeler, Microsoft SQL Server и Microsoft Analysis Services при построении модели

Примечание: сервер IBM SPSS Modeler не обязателен, хотя и может использоваться. Клиент IBM SPSS Modeler может сам обрабатывать вычисления исследования данных в базе данных.

Требования для интеграции с Microsoft Analysis Services

Ниже приведены обязательные предварительные требования для проведения моделирования в базе данных при помощи алгоритмов Analysis Services с IBM SPSS Modeler. Чтобы убедится в их соблюдении, может потребоваться проконсультироваться с администратором баз данных.

- Программа IBM SPSS Modeler, работающая с установкой сервер IBM SPSS Modeler (в распределенном режиме) в Windows. Платформы UNIX в этой интеграции с Analysis Services не поддерживаются.

  Важно: Пользователи IBM SPSS Modeler должны сконфигурировать соединение ODBC при помощи драйвера SQL Native Client, доступного в Microsoft по указанному ниже URL в разделе *Дополнительные требования к сервер IBM SPSS Modeler*. *Драйвер, предоставляемый с IBM SPSS Data Access Pack (и, как правило, рекомендуемый для использования с IBM SPSS Modeler в других задачах), использовать в этих целях не рекомендуется.* Драйвер нужно сконфигурировать для использования SQL Server с включенной опцией **Включить интегрированную проверку подлинности Windows** (Разрешить встроенную проверку подлинности Windows), поскольку IBM SPSS Modeler не поддерживает аутентификацию SQL Server. Для получения дополнительных сведений о создании и настройке разрешений для источников данных ODBC обратитесь к своему администратору базы данных.
- Должен быть установлен SQL Server 2005 или 2008, хотя не обязательно на том же хосте, что и IBM SPSS Modeler. У пользователей IBM SPSS Modeler должны быть достаточные разрешения на чтение и запись данных и на отбрасывание и создание таблиц и производных таблиц.

  Примечание: Рекомендуется SQL Server версии Enterprise Edition. Enterprise Edition обеспечивает дополнительную гибкость, предоставляя дополнительные параметры для настройки результатов алгоритмов. В версии Standard Edition предоставляются те же параметры, но редактировать некоторые дополнительные параметры пользователям не разрешено.
- Программа Microsoft SQL Server Analysis Services должна быть установлена на том же хосте, что и SQL Server.

Дополнительные требования к сервер IBM SPSS Modeler

Для использования алгоритмов Analysis Services с сервер IBM SPSS Modeler на компьютере хоста сервер IBM SPSS Modeler должны быть установлены следующие компоненты.

Примечание: Если SQL Server установлен на том же хосте, что и сервер IBM SPSS Modeler, эти компоненты будут уже доступны.

- Microsoft .NET Framework Version 2.0 Redistributable Package (x86)
- Microsoft Core XML Services (MSXML) 6.0
- Провайдер OLE DB Microsoft SQL Server 2008 Analysis Services 10.0 (обязательно выберите правильный вариант для используемой операционной системы)
- Microsoft SQL Server 2008 Native Client (обязательно выберите правильный вариант для используемой операционной системы)

Чтобы скачать эти компоненты, перейдите на сайт *www.microsoft.com/downloads*, в строке поиска введите **.NET Framework** или **SQL Server Feature Pack** (для всех остальных компонентов) и выберите последний пакет для вашей версии SQL Server.

Сначала может потребоваться установить другие пакеты, которые также должны быть доступны на сайте скачивания Microsoft.

Дополнительные требования к IBM SPSS Modeler

Для использования алгоритмов Analysis Services с IBM SPSS Modeler на клиенте должны быть установлены все приведенные выше, а также следующие компоненты:

- Microsoft SQL Server 2008 Datamining Viewer Controls (обязательно выберите правильный вариант для используемой операционной системы); к этому компоненту также требуется:
- Microsoft ADOMD.NET

Чтобы скачать эти компоненты, перейдите на сайт *www.microsoft.com/downloads*, в строке поиска введите **SQL Server Feature Pack** и выберите поселений пакет для вашей версии SQL Server.

Примечание: Моделирование баз данных и оптимизация SQL требуют, чтобы на компьютере IBM SPSS Modeler была включена возможность соединения с сервером IBM SPSS Modeler. При включенной возможности соединения можно обращаться к алгоритмам баз данных, выполнять обратный перенос SQL непосредственно с клиента сервером IBM SPSS Modeler и обращаться к серверу IBM SPSS Modeler. Проверьте текущее состояние лицензии, для чего в меню клиента сервером IBM SPSS Modeler выберите:

Справка > О программе > Дополнительные подробности

Если возможность соединения включена, на вкладке Состояние лицензии вы увидите опцию **Разрешение для сервера**.

Включение интеграции с Analysis Services

Для включения интеграции IBM SPSS Modeler с Analysis Services потребуется сконфигурировать SQL Server и Analysis Services, создать источник ODBC, включить интеграцию в диалоговом окне Вспомогательные прикладные программы IBM SPSS Modeler и включить поддержку генерирования и оптимизации SQL.

Примечание: Должны быть доступны Microsoft SQL Server и Microsoft Analysis Services. Дополнительную информацию смотрите в разделе “Требования для интеграции с Microsoft Analysis Services” на стр. 12.

Конфигурирование SQL Server

Сконфигурируйте на SQL Server возможность скоринга в базе данных.

1. Создайте на компьютере хоста SQL Server следующий ключ реестра:
   `HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP`
2. Добавьте в этот ключ следующее значение DWORD:
   `AllowInProcess 1`
3. После внесения этого изменения перезапустите SQL Server.

Конфигурирование Analysis Services

Для возможности соединения IBM SPSS Modeler с Analysis Services сначала в диалоговом окне Свойства сервера анализа нужно сконфигурировать вручную два параметра:

1. Войдите в систему сервера анализа через MS SQL Server Management Studio.
2. Откройте диалоговое окно Свойства, щелкнув правой кнопкой по имени сервера и выбрав **Свойства**.
3. Включите переключатель **Показать дополнительные (все) свойства**.
4. Измените следующие свойства:

- Измените для `DataMining\AllowAdHocOpenRowsetQueries` значение на `True` (значение по умолчанию - `False`).
- Измените для `DataMining\AllowProvidersInOpenRowset` значение на `[all]` (значения по умолчанию нет).

Создание DSN ODBC для SQL Server

Для чтения или записи данных из базы данных пользователь должен установить источник данных ODBC, настроить соответствующую базу данных и установить разрешения на чтение и запись. Драйвер ODBC Microsoft SQL Native Client - обязательный, он автоматически устанавливается с SQL Server. *Драйвер, предоставляемый с IBM SPSS Data Access Pack (и как правило, рекомендуемый для использования с IBM SPSS Modeler в других целях), для этой задачи использовать не рекомендуется.* Если IBM SPSS Modeler и SQL Server находятся на разных хостах, драйвер ODBC Microsoft SQL Native Client можно скачать. Дополнительную информацию смотрите в разделе “Требования для интеграции с Microsoft Analysis Services” на стр. 12.

Для получения дополнительных сведений о создании и настройке разрешений для источников данных ODBC обратитесь к своему администратору базы данных.

1. При помощи драйвера ODBC Microsoft SQL Native Client создайте DSN ODBC, указывающее на базу данных SQL Server, используемую в процессе исследования данных. Для остальных параметров драйвера надо использовать значения по умолчанию.
2. Для этого DSN убедитесь, что включена опция **Разрешить встроенную проверку подлинности Windows**.

- Если IBM SPSS Modeler и сервер IBM SPSS Modeler запускаются на разных хостах, создайте на каждом из этих хостов одно и то же DSN ODBC. Убедитесь, что одно и то же имя DSN используется на каждом хосте.

Включение интеграции с Analysis Services в IBM SPSS Modeler

Чтобы включить для IBM SPSS Modeler использование Analysis Services, сначала нужно задать спецификации сервера в диалоговом окне Вспомогательные прикладные программы.

1. Выберите в меню IBM SPSS Modeler:

   Инструменты > Опции > Вспомогательные прикладные программы
2. Щелкните по вкладке **Microsoft**.

- **Включить интеграцию с Microsoft Analysis Services.** Включает поддержку палитры Моделирование баз данных (если она еще не выводится) в нижней части окна IBM SPSS Modeler и добавляет узлы для алгоритмов Analysis Services.
- **Хост сервера анализа.** Задайте имя компьютера, на котором запускается Analysis Services.
- **База данных сервера анализа.** Выберите нужную базу данных, нажав кнопку с многоточием (...), открывающую вспомогательное диалоговое окно, в котором можно выбрать доступные базы данных. В

этом списке выводятся базы данных, доступные для заданного сервера анализа. Поскольку Microsoft Analysis Services сохраняет модели исследования данных в именованных базах данных, следует выбрать подходящую базу данных, в которой хранятся базы данных Microsoft, построенные IBM SPSS Modeler.

- **Соединение с SQL Server.** Задайте информацию DSN, используемую базой данных SQL Server для хранения данных, которые передаются на сервер анализа. Выберите источник данных ODBC, который будет предоставлять данные для построения моделей исследования данных Analysis Services. При построении моделей Analysis Services по данным, предоставляемым в плоских файлах или источниках данных ODBC, эти данные будут автоматически выгружаться в созданную в базе данных SQL Server временную таблицу, на которую указывает источник данных ODBC.
- **Предупреждать о перезаписи модели исследования данных.** Включите эту опцию, чтобы IBM SPSS Modeler не перезаписывал без предупреждения модели, хранимые в базе данных.

Примечание: Опции, сконфигурированные в диалоговом окне Вспомогательные прикладные программы, могут быть переопределены на различных узлах Analysis Services.

Включение поддержки генерирования и оптимизации SQL

1. Выберите в меню IBM SPSS Modeler:

   Инструменты > Свойства потока > Опции
2. На панели навигации щелкните по опции **Оптимизация**.
3. Подтвердите включение опции**Генерировать SQL**. Этот параметр требуется для работы функций моделирования баз данных.
4. Выберите **Оптимизировать построение SQL** и **Оптимизировать другое выполнение** (не требуется строго, но настоятельно рекомендуется для оптимальной производительности).

Построение моделей при помощи Analysis Services

Для построения моделей с помощью Analysis Services требуется, чтобы обучающий набор данных располагался в таблице или в представлении в базе данных SQL Server. Если эти данные расположены не на SQL Server или их нужно обработать в IBM SPSS Modeler как часть подготовки данных, которую нельзя выполнить на SQL Server, перед построением модели эти данные автоматически закачиваются во временную таблицу на SQL Server.

Работа с моделями Analysis Services

При построении модели Analysis Services с использованием IBM SPSS Modeler создается модель в IBM SPSS Modeler и создается или заменяется модель в базе данных SQL Server. Модель IBM SPSS Modeler ссылается на содержимое модели базы данных, хранящейся на сервере баз данных. IBM SPSS Modeler может выполнять проверку согласованности, сохранив одинаковую сгенерированную строку ключа модели и в модели IBM SPSS Modeler, и в модели SQL.

Узел моделирования **Дерево решений MS** служит для прогнозного моделирования атрибутов обоих типов, категориальных и непрерывных. Для категориальных атрибутов этот узел делает прогнозы на основе взаимосвязей между входными столбцами в наборе данных. Например, в сценарии предсказания, какие посетители с высокой вероятностью приобретут велосипед, если велосипеды приобретают девять из десяти молодых посетителей и только два из десяти пожилых, узел заключает, что возраст посетителя - значимый предиктор покупки велосипеда. На основе такого рода тенденции дерево решений делает прогнозы в сторону определенного исхода. Для непрерывных атрибутов алгоритм использует линейную регрессию, чтобы определить расщепление дерева решений. Если в качестве прогнозируемых задано несколько столбцов, или если во входных данных содержится вложенная таблица, заданная как прогнозируемая, узел строит отдельные деревья решений для каждого прогнозируемого столбца.

Узел моделирования **Кластеризация MS** использует итерационные методы для группирования наблюдений из набора данных в кластеры, содержащие сходные характеристики. Такая группировка полезна при просмотре данных, обнаружении аномалий в данных и создании прогнозов. Модели кластеризации находят в наборе данных логически неожиданные взаимосвязи, которые трудно заметить при случайном просмотре. Например, логично предположить, что среди работающих те, кто ездит на работу на велосипеде, обычно живут недалеко от места работы. Однако алгоритм может обнаружить другие, не столь очевидные характеристики тех, кто добирается до места работы на велосипеде. В отличии других узлов исследования данных для узла кластеризации не задается поле назначения. Узел кластеризации обучает модель, исходя непосредственно из того, какие взаимосвязи и кластеры он обнаружил в данных.

Узел моделирования **Правила связывания MS** полезен для механизмов рекомендации. Механизм рекомендации рекомендует продукты посетителям на основе тех продуктов, которые они уже приобрели, или на основе того, к каким продуктам выказали интерес. Ассоциативные модели строятся для наборов данных, которые содержат идентификаторы наблюдений и элементов в этих наблюдениях. Группа элементов в одном наблюдении называется **набором элементов**. Ассоциативная модель состоит из ряда наборов элементов и правил, описывающих группировку этих элементов в наблюдениях. Правила, обнаруживаемые алгоритмом, можно использовать для предсказания вероятных будущих покупок некоторого посетителя на основе уже положенных в корзину продуктов.

Узел моделирования **Наивный Байес MS** вычисляет условную вероятность для полей назначения и полей предикторов и предполагает, что столбцы независимы. Модель называется наивной, поскольку в ней все предложенные переменные прогноза считаются независимыми друг от друга. Для этого метода требуется меньше вычислений, чем для остальных алгоритмов Analysis Services, и поэтому он полезен для быстрого обнаружения взаимосвязей на предварительных стадиях моделирования. Этот узел можно использовать для начального изучения данных, а потом применить результаты при создании дополнительных моделей, используя другие узлы, которые будут вычисляться дольше, но зато дадут более точные результаты.

Узел моделирования **Линейная регрессия MS** - это вариант узла Дерево решений, в котором значение параметра `MINIMUM_LEAF_CASES` больше или равно общему числу наблюдений в наборе данных, используемом этим узлом для обучения модели исследования. При таком значении параметра этот узел никогда не будет расщеплен, и следовательно, выполняет только линейную регрессию.

Узел моделирования **Нейросеть MS** аналогичен узлу Дерево решений MS в том отношении, что вычисляет вероятности всех возможных состояний входного атрибута при каждом данном состоянии прогнозируемого атрибута. В дальнейшем эти вероятности можно использовать для предсказания исхода предсказанного атрибута на основании входных атрибутов.

Узел моделирования **Логистическая регрессия MS** - это вариант узла Нейросеть MS, в котором для параметра `HIDDEN_NODE_RATIO` задано значение 0. При таком значении параметра модель нейросети не содержит скрытого слоя и, следовательно, эквивалентна логистической регрессии.

Узел моделирования **Временные ряды MS** содержит алгоритмы регрессии, оптимизированные для предсказания будущих значений непрерывных величин, таких как продажи продукта. Модель временных рядов не требует, как другие алгоритмы Microsoft (например, деревья решений), задавать дополнительные столбцы новой информации как входные, чтобы спрогнозировать тенденцию. Модель временных рядов способна прогнозировать тенденции на основании только исходного набора данных, использованного при создании модели. Кроме того, вычисляя прогнозы, можно добавлять в модель новые данные, и они автоматически включаются в анализ тенденций. Дополнительную информацию смотрите в разделе “Узел временных рядов MS” на стр. 19.

Узел моделирования **Кластеризация последовательностей MS** обнаруживает в данных упорядоченные последовательности и сочетает результаты этого анализа с методами кластеризации, чтобы сгенерировать кластеры на основе последовательностей и других атрибутов. Дополнительную информацию смотрите в разделе “Узел кластеризации последовательностей MS” на стр. 20.

Все эти узлы доступны на палитре моделирования базы данных у нижнего края окна IBM SPSS Modeler.

Общие параметры для всех узлов алгоритмов

Следующие параметры - общие для всех алгоритмов Analysis Services.

Опции сервера

На вкладке Сервер можно сконфигурировать хост сервера анализа, базу данных и источник данных SQL Server. Заданные здесь опции переопределяют опции, заданные на вкладке Microsoft в диалоговом окне Вспомогательные прикладные программы. Дополнительную информацию смотрите в разделе “Включение интеграции с Analysis Services” на стр. 13.

Примечание: Изменения на этой вкладке доступны также при скоринге моделей Analysis Services. Дополнительную информацию смотрите в разделе “Вкладка Сервер слепков моделей Analysis Services” на стр. 22.

Параметры модели

Чтобы построить самую общую модель, перед прочими действиями нужно задать опции на вкладке Модель. На вкладке Дополнительно доступен метод скоринга и другие расширенные опции.

Доступны следующие основные опции моделирования:

Имя модели. Задает имя, назначаемое модели, которая создается при выполнении узла.

- **Авто.** Автоматически генерирует имя модели на основании имен полей назначения или ID/имени типа модели в случаях, когда не задано назначение (например, в моделях кластеризации).
- **Пользовательская.** Позволяет создать пользовательское имя для создаваемой модели.

Использовать разделенные данные. Разделяет данные на отдельные подмножества или выборки для обучения, испытания и проверки на основании текущего значения поля разделения. Использования одной выборки данных для создания модели и другой для ее испытания позволяет выяснить, насколько хорошо модель обобщается на более крупные наборы данных, аналогичных текущим данным. Если в потоке поле раздела не задано, эта опция игнорируется.

С детализацией. Эта опция, если она выводится, позволяет запросить модель, чтобы узнать подробности о включенных в модель наблюдениях.

Поле уникальности. Выберите в выпадающем списке поле, уникально идентифицирующее каждое наблюдение. Обычно это поле ID, такое как **ID_покупателя**.

Дополнительные опции дерева решений MS

Состав опций, доступных на вкладке Дополнительные, может быть различным в зависимости от структуры выбранного потока. Исчерпывающие подробности о дополнительных опциях для выбранного узла моделирования Analysis Services смотрите в справке для пользовательского интерфейса на уровня полей.

Дополнительные опции кластеризации MS

Состав опций, доступных на вкладке Дополнительные, может быть различным в зависимости от структуры выбранного потока. Исчерпывающие подробности о дополнительных опциях для выбранного узла моделирования Analysis Services смотрите в справке для пользовательского интерфейса на уровня полей.

Дополнительные опции наивного критерия Байеса MS

Состав опций, доступных на вкладке Дополнительные, может быть различным в зависимости от структуры выбранного потока. Исчерпывающие подробности о дополнительных опциях для выбранного узла моделирования Analysis Services смотрите в справке для пользовательского интерфейса на уровня полей.

Дополнительные опции линейной регрессии MS

Состав опций, доступных на вкладке Дополнительные, может быть различным в зависимости от структуры выбранного потока. Исчерпывающие подробности о дополнительных опциях для выбранного узла моделирования Analysis Services смотрите в справке для пользовательского интерфейса на уровня полей.

Дополнительные опции нейросети MS

Состав опций, доступных на вкладке Дополнительные, может быть различным в зависимости от структуры выбранного потока. Исчерпывающие подробности о дополнительных опциях для выбранного узла моделирования Analysis Services смотрите в справке для пользовательского интерфейса на уровня полей.

Дополнительные опции логистической регрессии MS

Состав опций, доступных на вкладке Дополнительные, может быть различным в зависимости от структуры выбранного потока. Исчерпывающие подробности о дополнительных опциях для выбранного узла моделирования Analysis Services смотрите в справке для пользовательского интерфейса на уровня полей.

Узел правил связывания MS

Узел моделирования правил связывания MS полезен для механизмов рекомендаций. Узел рекомендаций предлагает покупателям продукты на основании тех позиций, которые они уже приобрели, или тех, о которых они выразили заинтересованность. Модели связывания строятся для наборов данных, включающих в себя идентификаторы и для отдельных наблюдений, и для содержащихся в них элементов. Группа элементов в наблюдении называется **набором элементов**.

Модель связывания состоит из ряда наборов элементов и правил, описывающих, как эти элементы группируются вместе в наблюдениях. Эти правила, определяемые алгоритмом, можно использовать для предсказания возможных в будущем покупок покупателя на основании элементов, уже присутствующих в его покупательской корзине.

Для табличного формата данных алгоритм создает оценки, представляющие собой вероятности ($MP-*поле*) для каждой из сгенерированных рекомендаций ($M-*поле*). При транзакционном формате данных оценки создаются для поддержки ($MS-*поле*), вероятности ($MP-*поле*) и скорректированной вероятности ($MAP-*поле*) для каждой сгенерированной рекомендации ($M-*поле*).

Технические требования

Технические требования для транзакционной модели связывания:

- **Поле уникальности.** Модель правил связывания требует наличия ключа, однозначно идентифицирующего записи.

- **Поле ID.** При построении модели правил связывания MS с транзакционным форматом данных требуется поле ID, определяющее каждую транзакцию. В качестве поля ID можно задать то же поле, что и в качестве поля уникальности.
- **По крайней мере одно поле ввода.** Для алгоритма правил связывания требуется по крайней мере одно поле ввода.
- **Поле назначения.** При построении модели связывания MS с транзакционными данными поле назначение должно быть полем транзакции, например, предназначенным для указания товаров, которые приобрел покупатель.

Дополнительные опции правил связывания MS

Состав опций, доступных на вкладке Дополнительные, может быть различным в зависимости от структуры выбранного потока. Исчерпывающие подробности о дополнительных опциях для выбранного узла моделирования Analysis Services смотрите в справке для пользовательского интерфейса на уровня полей.

Узел временных рядов MS

Узел моделирования временных рядов MS поддерживает два типа предсказаний:

- будущие
- хронологические

Будущие предсказания оценивают значения в полях назначения через заданное число периодов времени после окончания хронологических данных и выполняются всегда. **Хронологические предсказания** - это оцененные значения в полях назначения на заданное число периодов времени, для которых у вас есть фактические значения в хронологических данных. Хронологические предсказания можно использовать для оценки качества модели, сравнивая фактические хронологические значения с предсказанными. Значение начальной точки времени для предсказаний определяет, будут ли выполняться хронологические предсказания.

В отличие от узла временных рядов IBM SPSS Modeler, узлу временных рядов MS не нужен предшествующий узел Интервалы времени. Другое отличие состоит в том, что по умолчанию оценки делаются только для предсказанных строк, а не для всех хронологических строк в данных временных рядов.

Технические требования

Технические требования для модели временных рядов MS:

- **Одно ключевое поле времени.** Каждая модель должна содержать одно числовое поле или поле даты, используемые как ряд наблюдений, определяющий сектора времени, которые будет использовать модель. Типом данных для ключевого поля времени может быть или тип данных даты-времени, или числовой тип данных. Однако это поле должно содержать количественные значения, и эти значения должны быть уникальными для каждого ряда.
- **Одно поле назначения.** В каждой модели можно задать только одно поле назначения. Типом данных для этого поля должны быть количественные значения. Например, вы можете предсказывать изменение во времени численных атрибутов, таких как доход, объем продаж или температура. Однако в качестве поля назначения нельзя использовать поле, содержащее категориальные значения, такие как состояние покупки или уровень образования.
- **Не меньше одного поля ввода.** Для алгоритма временных рядов MS требуется по крайней мере одно поле ввода. Типом данных для поля ввода должны быть количественные значения. При построении модели неколичественные поля ввода игнорируются.
- **Набор данных должен быть отсортирован.** Набор данных ввода должен быть отсортирован (по ключевому полю времени), в противном случае построение модели будет прервано с выводом сообщения об ошибке.

Опции модели MS Time Series

Имя модели. Задает имя модели, создаваемой при исполнении узла.

- **Авто.** Автоматически генерирует имя модели на основании имен полей назначения или ID/имени типа модели в случаях, когда не задано назначение (например, в моделях кластеризации).

- **Пользовательская.** Позволяет создать пользовательское имя для создаваемой модели.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

С детализацией. Эта опция, если она выводится, позволяет запросить модель, чтобы узнать подробности о включенных в модель наблюдениях.

Поле уникальности. Выберите из выпадающего списка ключевое поле времени, которое используется для построения модели временного ряда.

Опции MS Time Series Expert

Состав опций, доступных на вкладке Дополнительные, может быть различным в зависимости от структуры выбранного потока. Исчерпывающие подробности о дополнительных опциях для выбранного узла моделирования Analysis Services смотрите в справке для пользовательского интерфейса на уровня полей.

Если вы делаете хронологические предсказания, количество хронологических шагов, которые можно включить в результат оценки, определяется произведением (HISTORIC_MODEL_COUNT * HISTORIC_MODEL_GAP). По умолчанию это количество ограничено десятью, то есть можно сделать только 10 хронологических предсказаний. И в этом случае, например, возникнет ошибка, если вы введете для параметра **Хронологическое предсказание** на вкладке Параметры слепка модели (смотрите “Вкладка параметров слепков моделей временных рядов MS” на стр. 23) значение меньше -10. Если вы хотите получить больше хронологических предсказаний, можно увеличить значения параметров HISTORIC_MODEL_COUNT или HISTORIC_MODEL_GAP, но при этом увеличится время построения модели.

Опции параметров временных рядов MS

Начать оценку. Задайте временной период, откуда вы хотите начать предсказания.

- **Начать с: Новое предсказание.** Временной период, в который вы хотите начать будущие предсказания, выраженный в виде смещения относительно последнего периода времени ваших данных хронологии. Например, если данные хронологии закончились 12/99, а вы захотели бы начать предсказания с 01/00, вы использовали бы значение 1; однако если вы захотели бы начать предсказания с 03/00, вы использовали бы значение 3.
- **Начать с: Хронологическое предсказание.** Временной период, в который вы хотите начать хронологические предсказания, выраженный в виде отрицательного смещения относительно последнего периода времени ваших данных хронологии. Например, если данные хронологии закончились 12/99, а вы захотели бы составить хронологические предсказания за последние пять периодов времени ваших данных, вы использовали бы значение -5.

Закончить оценку. Задайте временной период, где вы хотите остановить предсказания.

- **Конечный шаг предсказания.** Временной период, в который вы хотите остановить предсказания, выраженный в виде смещения относительно последнего периода времени ваших данных хронологии. Например, если данные хронологии заканчиваются 12/99, а вы захотели бы остановить предсказания 6/00, вы использовали бы здесь значение 6. Для будущих предсказаний это значение должно быть всегда больше или равно значению **Начать с**.

Узел кластеризации последовательностей MS

Узел кластеризации последовательностей MS использует алгоритм анализа последовательностей, изучающий содержащие данные события, которые можно связать, следуя путям или *последовательностям*. Примерами этого могут быть пути переходов по ссылкам, создаваемые при просмотре Web-сайта или навигации по нему пользователями, или порядок, в котором покупатели добавляют элементы в покупательскую корзину в розничном интернет-магазине. Алгоритм находит наиболее общие последовательности, используя группировку, или *кластеризацию*, идентичных последовательностей.

Требования

Технические требования для модели кластеризации последовательностей Microsoft:

- **Поле ID.** Для алгоритма кластеризации последовательностей Microsoft требуется, чтобы информация последовательностей хранилась в транзакционном формате. Для этого требуется поле ID, идентифицирующее каждую транзакцию.
- **Не меньше одного поля ввода.** Для алгоритма требуется по крайней мере одно поле ввода.
- **Поле последовательности.** Для алгоритма требуется также поле идентификатора последовательностей, у которого должен быть количественный уровень измерения. Например, можно использовать идентификатор Web-страницы, целое число или текстовую строку, лишь бы данное поле определяло события в последовательности. Для каждой последовательности разрешен только один идентификатор, а для каждой модели разрешен только один тип последовательности. Поле последовательности должно отличаться от поля ID и поля уникальности.
- **Поле назначения.** При построении модели кластеризации последовательностей требуется поле назначения.
- **Поле уникальности.** Для модели кластеризации последовательностей требуется поля ключа, который однозначно идентифицирует записи. Поле уникальности можно задать совпадающим с полем ID.

Опции полей кластеризации последовательностей MS

У всех узлов моделирования есть вкладка Поля, где вы задаете поля, которые будут использоваться при построении модели.

Перед построением модели кластеризации последовательности необходимо указать поля, которые должны служить полями назначения и входными полями. Обратите внимание на то, что для узла кластеризации последовательностей MS нельзя использовать информацию из расположенного выше узла Тип; значения в полях нужно указать здесь.

ID. Выберите поле ID из списка. Значения в поле ID могут быть числовыми или символическими. Каждое уникальное значение в этом поле должно обозначать конкретный объект анализа. Например, в прикладной программе Корзина покупок каждый ID может представлять одного покупателя. В прикладной программе Анализ Web-журнала каждый ID может представлять отдельный компьютер (по IP-адресу) или одного пользователя (по регистрационным данным).

Поля ввода. Выберите поле или поля ввода для модели. Это поля, которые содержат интересующие вас события при моделировании последовательностей.

Последовательность. Выберите из списка поле, которое будет использоваться как поле идентификатора последовательности. Например, можно использовать идентификатор Web-страницы, целое число или текстовую строку, лишь бы данное поле определяло события в последовательности. Для каждой последовательности разрешен только один идентификатор, а для каждой модели разрешен только один тип последовательности. Поле последовательности должно отличаться от поля ID (заданного на этой вкладке) и от поля уникальности (заданного на вкладке Модель).

Цель. Выберите поле, которое будет использоваться как поле назначения, то есть такое поле, значение в котором вы пытаетесь предсказать на основе данных последовательности.

Дополнительные опции кластеризации последовательностей MS

Состав опций, доступных на вкладке Дополнительные, может быть различным в зависимости от структуры выбранного потока. Исчерпывающие подробности о дополнительных опциях для выбранного узла моделирования Analysis Services смотрите в справке для пользовательского интерфейса на уровня полей.

Скоринг для моделей Analysis Services

Оценка моделей происходит на SQL Server и выполняется в Analysis Services. Если данные созданы или должны подготавливаться в IBM SPSS Modeler, может потребоваться закачать набор данных во временную таблицу. Модели, создаваемые вами в IBM SPSS Modeler с использованием исследования данных в базе данных, фактически представляют собой удаленные модели, хранящиеся на удаленных серверах

исследования данных или баз данных. Это отличие важно понимать при просмотре и оценке моделей, созданных с использованием алгоритмов Microsoft Analysis Services.

В IBM SPSS Modeler обычно передается только одно предсказание и связанная вероятность или показатель доверия.

Примеры скоринга моделей смотрите в разделе “Примеры исследования Analysis Services” на стр. 24.

Общие параметры для всех моделей Analysis Services

Следующие параметры - общие для всех моделей Analysis Services.

Вкладка Сервер слепков моделей Analysis Services

Вкладка Сервер используется для задания соединений для исследования в базе данных. На этой вкладке предоставляется также ключ уникальности для модели. Этот ключ генерируется случайным образом при построении модели и хранится в модели IBM SPSS Modeler, а также в описании объекта модели, хранимого в базе данных Analysis Services.

На вкладке Сервер можно сконфигурировать хост и базу данных сервера анализа, а также источник данных SQL Server для операции скоринга. Заданные здесь опции перезаписывают соответствующие опции, заданные в IBM SPSS Modeler в диалоговых окнах Вспомогательные программы или или Построить модель. Дополнительную информацию смотрите в разделе “Включение интеграции с Analysis Services” на стр. 13.

GUID модели. Здесь выводится ключ модели. Этот ключ генерируется случайным образом при построении модели и хранится в модели IBM SPSS Modeler, а также в описании объекта модели, хранимого в базе данных Analysis Services.

Проверить. Нажмите эту кнопку, чтобы проверить ключ модели для ключа в модели, хранимой в базе данных Analysis Services. Это позволит убедиться, что модель все еще существует на сервере анализа, и понять, что структура модели не изменена.

Примечание: Кнопка Проверить доступна только для моделей, добавленных на холст потока при подготовке к скорингу. Если проверка завершится неудачно, выясните, не была ли модель удалена или заменена другой моделью на сервере.

Просмотр. Щелкните здесь для получения графического представления модели дерева решений. Программа просмотра дерева решений \используется совместно с другими алгоритмами дерева решений в IBM SPSS Modeler и функционально не отличается.

Вкладка Сводка слепков моделей Analysis Services

На вкладке Сводка слепка модели выводится информация о самой модели (*Анализ*), об используемых в ней полях (*Поля*), значениях параметров, используемых при построении модели, (*Параметры построения*) и об обучении модели (*Сводка по обучению*).

При первом просмотре узла результаты вкладки Сводка свернуты. Чтобы увидеть нужные вам результаты, разверните соответствующие им элементы при помощи элемента управления расширением слева от них или выведите все результаты, нажав кнопку **Развернуть все**. Чтобы скрыть результаты по завершении их просмотра, сверните при помощи элемента управления расширением отдельные результаты, которые вы хотите скрыть, или сверните все результаты, нажав кнопку **Свернуть все**.

Анализ. Выводится информация о конкретной модели. Если в вашем исполнении в указанный слепок модели вложен узел анализа, в этом разделе появится также информация о выполненном анализе.

Поля. Список полей, используемых в качестве полей назначения и входных полей при построении модели.

Параметры компоновки. Содержит информацию об используемых при построении модели параметрах.

Сводная информация по обучению. Выводится тип модели, поток, используемый для ее создания, создавший ее пользователь, отметка времени построения модели и время, затраченное на ее построение.

Слепок модели временного ряда MS

Модель временного ряда MS делает оценки только для предсказанных периодов времени, а не для хронологических данных.

В следующей таблице показаны поля, которые добавлены в модель.

Таблица 1. Добавленные в модель поля

Имя поля	Описание
$M-*поле*	Предсказанное значение *поля*
$Var-*поле*	Вычисленная дисперсия *поля*
$Stdev-*поле*	Среднеквадратичное отклонение *поля*

Вкладка Сервер слепков моделей временных рядов MS

Вкладка Сервер используется для задания соединений для исследования в базе данных. На этой вкладке предоставляется также ключ уникальности для модели. Этот ключ генерируется случайным образом при построении модели и хранится в модели IBM SPSS Modeler, а также в описании объекта модели, хранимого в базе данных Analysis Services.

На вкладке Сервер можно сконфигурировать хост и базу данных сервера анализа, а также источник данных SQL Server для операции скоринга. Заданные здесь опции перезаписывают соответствующие опции, заданные в IBM SPSS Modeler в диалоговых окнах Вспомогательные программы или или Построить модель. Дополнительную информацию смотрите в разделе “Включение интеграции с Analysis Services” на стр. 13.

GUID модели. Здесь выводится ключ модели. Этот ключ генерируется случайным образом при построении модели и хранится в модели IBM SPSS Modeler, а также в описании объекта модели, хранимого в базе данных Analysis Services.

Проверить. Нажмите эту кнопку, чтобы проверить ключ модели для ключа в модели, хранимой в базе данных Analysis Services. Это позволит убедиться, что модель все еще существует на сервере анализа, и понять, что структура модели не изменена.

Примечание: Кнопка Проверить доступна только для моделей, добавленных на холст потока при подготовке к скорингу. Если проверка завершится неудачно, выясните, не была ли модель удалена или заменена другой моделью на сервере.

Просмотр. Щелкните здесь для получения графического представления модели временного ряда. Analysis Services выводят полную модель в виде дерева. Вы можете просмотреть также график, показывающий хронологические значения поля назначения во времени вместе с предсказанными будущими значениями.

Более подробную информацию смотрите в описании инструмента просмотра временных рядов в библиотеке MSDN по адресу http://msdn.microsoft.com/en-us/library/ms175331.aspx.

Вкладка параметров слепков моделей временных рядов MS

Начать оценку. Задайте временной период, откуда вы хотите начать предсказания.

- **Начать с: Новое предсказание.** Временной период, в который вы хотите начать будущие предсказания, выраженный в виде смещения относительно последнего периода времени ваших данных хронологии. Например, если данные хронологии закончились 12/99, а вы захотели бы начать предсказания с 01/00, вы использовали бы значение 1; однако если вы захотели бы начать предсказания с 03/00, вы использовали бы значение 3.

- **Начать с: Хронологическое предсказание.** Временной период, в который вы хотите начать хронологические предсказания, выраженный в виде отрицательного смещения относительно последнего периода времени ваших данных хронологии. Например, если данные хронологии закончились 12/99, а вы захотели бы составить хронологические предсказания за последние пять периодов времени ваших данных, вы использовали бы значение -5.

Закончить оценку. Задайте временной период, где вы хотите остановить предсказания.

- **Конечный шаг предсказания.** Временной период, в который вы хотите остановить предсказания, выраженный в виде смещения относительно последнего периода времени ваших данных хронологии. Например, если данные хронологии заканчиваются 12/99, а вы захотели бы остановить предсказания 6/00, вы использовали бы здесь значение 6. Для будущих предсказаний это значение должно быть всегда больше или равно значению **Начать с**.

Слепок модели кластеризации последовательностей MS

В следующей таблице показаны поля, которые добавлены в модель кластеризации последовательностей MS (здесь *поле* - это имя поля назначения).

Таблица 2. Добавленные в модель поля

Имя поля	Описание
$MC-*поле*	Предсказание кластера, к которому принадлежит эта последовательность.
$MCP-*поле*	Вероятность, что эта последовательность принадлежит к предсказанному кластеру.
$MS-*поле*	Предсказанное значение в *поле*
$MSP-*поле*	Вероятность правильности значения $MS-*поля*.

Экспорт моделей и генерирование узлов

Можно экспортировать сводку и структуру модели в текстовые файлы и файлы формата HTML. Можно генерировать нужные узлы выбора и фильтрации там, где они вам требуются.

Подобно другим слепкам моделей в IBM SPSS Modeler, слепки моделей Microsoft Analysis Services поддерживают непосредственное генерирование узлов операций с записями и полями. Используя пункты меню Генерировать слепок модели, можно сгенерировать такие узлы:

- Узел выбора (только если на вкладке Модель выбран элемент)
- Узел фильтра

Примеры исследования Analysis Services

Дается ряд примеров потока, демонстрирующих, как использовать исследование данных MS Analysis Services совместно с IBM SPSS Modeler. Эти потоки можно найти в подпапке установки IBM SPSS Modeler:

\Demos\Database_Modelling\Microsoft

Примечание: Папку Demos можно открыть из группы программ IBM SPSS Modeler в меню Запуск Windows.

Примеры потоков: деревья решений

Следующие примеры потоков можно использовать вместе в последовательности как пример процесса исследования базы данных, использующего алгоритм деревьев решений, который предоставлен MS Analysis Services.

Таблица 3. Деревья решений - примеры потоков

Поток	Описание
1_upload_data.str	Используется для очистки и закачивания данных из плоского файла в базу данных.
2_explore_data.str	Предоставляет пример изучения данных с помощью IBM SPSS Modeler
3_build_model.str	Строит модель, используя собственный алгоритм базы данных.
4_evaluate_model.str	Используется как пример оценки модели с помощью IBM SPSS Modeler.
5_deploy_model.str	Внедряет модель для оценки базы данных.

Примечание: Чтобы запустить этот пример, надо выполнять потоки по порядку. Кроме этого, узлы источников и моделирования в каждом потоке должны быть изменены, чтобы содержать ссылку на допустимый источник данных для базы данных, который вы хотите использовать.

Набор данных, используемый в этих примерах потоков, относится к прикладным программам обработки кредитных карт и представляет задачу классификации с различными категориальными и непрерывными предикторами. Более подробную информацию об этом наборе данных смотрите в файле *crx.names* в той же папке, где находятся примеры потоков.

Этот набор данных доступен в репозитории UCI Machine Learning по адресу *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*.

Пример потока: Закачка данных

Первый пример потока, *1_upload_data.str*, служит для очистки и закачки данных из плоского файла на SQL Server.

Поскольку Analysis Services для исследования данных требуют наличия ключевого поля, в этом начальном потоке при помощи узла вычислений в набор данных добавляется новое поле *KEY* с уникальными значениями *1,2,3*, использующее функцию IBM SPSS Modeler `@INDEX`.

Далее следует узел заполнения, который обрабатывает пропущенные значения; пустые поля, считанные из текстового файла *crx.data*, заменяются на значение *NULL*.

Пример потока: Просмотр данных

Второй пример потока, *2_explore_data.str*, служит для демонстрации того, как применить узел Аудит данных, чтобы получить обзор данных, включая сводную статистику и диаграммы.

Двойной щелчок по диаграмме в отчете аудита данных выводит более подробную диаграмму для детального изучения данного поля.

Пример потока: построение модели

В третьем примере потока, *3_build_model.str*, иллюстрируется построение модели в IBM SPSS Modeler. Можно присоединить модель базы данных к потоку и дважды щелкнуть, чтобы задать параметры построения.

На вкладке Модель диалогового окна можно задать:

1. Выбрать поле **Ключ** в качестве поля уникального ID.

На вкладке Эксперт доступна точная настройка параметров построения модели.

Перед запуском убедитесь, что правильно указали базу данных для построения модели. Для настройки каких-либо параметров используйте вкладку Сервер.

Пример потока: Оценка модели

Четвертый пример потока, *4_evaluate_model.str*, иллюстрирует преимущества использования IBM SPSS Modeler для моделирования в базе данных. После выполнения модели ее можно снова добавить в поток данных и оценить при помощи ряда инструментов, которыми располагает IBM SPSS Modeler.

Просмотр результатов моделирования

Можно дважды щелкнуть по слепку модели для просмотра результатов. На вкладке Сводка содержится дерево правил результатов. Кроме того, можно нажать кнопку **Вид** на вкладке Сервер и вывести графическое представление модели деревьев решений.

Оценка результатов модели

Узел анализа в примере потока создает матрицу совпадений, содержащую структуру соответствий между каждым предсказанным полем и его полем назначения. Чтобы увидеть результаты, запустите узел анализа.

Узел оценки в примере потока может создать диаграмму выигрышей, предназначенную для демонстрации, насколько эта модель повышает точность. Чтобы увидеть результаты, запустите узел оценки.

Пример потока: Внедрение модели

Когда достигнута достаточная точность модели, эту модель можно внедрить для использования во внешних прикладных программах или для публикации в той же базе данных. В заключительном примере потока, *5_deploy_model.str*, данные считываются из таблицы `CREDIT` и затем оцениваются и публикуются в таблице `CREDITSCORES` при помощи узла Экспорт базы данных.

Выполнение потока генерирует следующий SQL:

```
DROP TABLE CREDITSCORES

CREATE TABLE CREDITSCORES ( "field1" varchar(1),"field2" varchar(255),"field3" f
loat,"field4" varchar(1),"field5" varchar(2),"field6" varchar(2),"field7" varcha
r(2),"field8" float,"field9" varchar(1),"field10" varchar(1),"field11" int,"fiel
d12" varchar(1),"field13" varchar(1),"field14" int,"field15" int,"field16" varch
ar(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 T0.C0 AS C0,T0.C1 AS C1,T0.C2 AS C2,T0.C3 AS C3,T0.C4 AS C4,T0.C5 AS C5,
 T0.C6 AS C6,T0.C7 AS C7,T0.C8 AS C8,T0.C9 AS C9,T0.C10 AS C10,
 T0.C11 AS C11,T0.C12 AS C12,T0.C13 AS C13,T0.C14 AS C14,
 T0.C15 AS C15,T0.C16 AS C16,T0.C17 AS C17,T0.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 T0."field1" AS C0,T0."field2" AS C1,
   T0."field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
   T0."field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
```

```
   T0."field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
   T0."field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
   T0."KEY" AS C16 FROM "dbo".CREDITDATA T0'') 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]
)  T0
```

Глава 4. Моделирование баз данных с использованием Oracle Data Mining

О программе Oracle Data Mining

IBM SPSS Modeler поддерживает интеграцию с Oracle Data Mining (ODM), которая предлагает семейство алгоритмов исследования данных, тесно связанных с реляционной СУБД Oracle. Эти возможности доступны через графический интерфейс IBM SPSS Modeler и в среде разработки, ориентированной на использование рабочих потоков, так что пользователи могут применять алгоритмы исследования данных, предлагаемые ODM.

IBM SPSS Modeler поддерживает интеграцию со следующими алгоритмами из Oracle Data Mining:

- наивная модель Байеса
- Адаптивный Байес
- Метод опорных векторов (Support Vector Machine, SVM)
- Обобщенные линейные модели (Generalized Linear Models, GLM)*
- Деревья решений
- O-кластер
- k-средние
- Разложение матрицы на неотрицательные множители (Nonnegative Matrix Factorization, NMF)
- Априорный анализ
- Метод минимальной длины описания (Minimum Descriptor Length, MDL)
- Важность атрибутов (Attribute Importance, AI)

* Только для 11*g* R1

Требования для интеграции с Oracle

Следующие условия представляют собой обязательные предварительные требования для проведения моделирования In-Database при помощи Oracle Data Mining. Чтобы убедится в их соблюдении, может потребоваться проконсультироваться с администратором баз данных.

- Программа IBM SPSS Modeler, работающая в локальном режиме или с установкой сервер IBM SPSS Modeler в Windows или UNIX
- Oracle 10*g*R2 или 11*g*R1 (10.2 Database или новее) с опцией Oracle Data Mining.

Примечание: 10*g*R2 обеспечивает поддержку всех алгоритмов моделирования баз данных, кроме обобщенных линейных моделей (требуется 11*g*R1).

- Источник данных ODBC для соединения с Oracle, как описано ниже.

Примечание: Моделирование баз данных и оптимизация SQL требуют, чтобы на компьютере IBM SPSS Modeler была включена возможность соединения с сервером IBM SPSS Modeler. При включенной возможности соединения можно обращаться к алгоритмам баз данных, выполнять обратный перенос SQL непосредственно с клиента сервером IBM SPSS Modeler и обращаться к серверу IBM SPSS Modeler. Проверьте текущее состояние лицензии, для чего в меню клиента сервером IBM SPSS Modeler выберите:

Справка > О программе > Дополнительные подробности

Если возможность соединения включена, на вкладке Состояние лицензии вы увидите опцию **Разрешение для сервера**.

Включение интеграции с Oracle

Для включения интеграции IBM SPSS Modeler с Oracle Data Mining потребуется сконфигурировать Oracle, создать источник ODBC, включить интеграцию в диалоговом окне Вспомогательные прикладные программы IBM SPSS Modeler и включить поддержку генерирования и оптимизации SQL.

Конфигурирование Oracle

Информацию об установке и конфигурировании Oracle Data Mining смотрите в документации Oracle; конкретные подробности - в *руководстве администратора Oracle*.

Создание источника ODBC для Oracle

Для включения поддержки соединения между Oracle и IBM SPSS Modeler нужно создать имя системного источника данных ODBC (DSN).

Перед созданием DSN у вас должно быть основное представление об источниках данных и драйверах ODBC и поддержке баз данных в IBM SPSS Modeler.

При работе в распределенном режиме для сервер IBM SPSS Modeler создайте DSN на компьютере сервера. При работе в локальном (клиентском) режиме создайте DSN на компьютере клиента.

1. Установите драйверы ODBC. Они доступны на установочном диске IBM SPSS Data Access Pack, поставляемом с этим выпуском. Запустите файл *setup.exe*, чтобы запустить программу установки, после чего выберите все нужные драйверы. Для установки драйверов следуйте инструкциям на экране.
   a. Создайте DSN.

      Примечание: Последовательность меню зависит от используемой версии Windows.

      - **Windows XP.** В меню Пуск выберите **Панель управления**. Щелкните дважды по значку **Администрирование**, а затем - по значку **Источники данных (ODBC)**.
      - **Windows Vista.** В меню Пуск выберите **Панель управления**, затем выберите **Система**. Щелкните дважды по значку **Администрирование** и выберите **Источники данных (ODBC)**, затем выберите **Открыть**.
      - **Windows 7.** В меню Пуск выберите **Панель управления**, затем **Система и безопасность**, затем **Администрирование**. Выберите **Источники данных (ODBC)**, затем выберите **Открыть**.

   b. Щелкните по вкладке **Системное DSN**, а затем нажмите кнопку **Добавить**.
2. Выберите драйвер **SPSS OEM 6.0 Oracle Wire Protocol**.
3. Нажмите кнопку **Готово**.
4. На экране установки драйвера ODBC Oracle Wire Protocol введите выбранное вами имя источника данных, имя хоста сервера Oracle, номер порта для соединения и SID для используемого вами экземпляра Oracle.

   Имя хоста, номер порта и SID можно получить из файла *tnsnames.ora* на компьютере сервера (если реализовано TNS с файлом *tnsnames.ora*. За дополнительной информацией обращайтесь к администратору Oracle.
5. Нажмите кнопку **Проверить**, чтобы опробовать соединение.

Включение интеграции с Oracle Data Mining в IBM SPSS Modeler

1. Выберите в меню IBM SPSS Modeler:

   Инструменты > Опции > Вспомогательные программы
2. Щелкните по вкладке **Oracle**.

Включить интеграцию с Oracle Data Mining. Включает поддержку палитры Моделирование баз данных (если она еще не выводится) в нижней части окна IBM SPSS Modeler и добавляет узлы для алгоритмов Oracle Data Mining.

Соединение с Oracle. Задайте источник данных ODBC Oracle по умолчанию, используемый для построения и сохранения моделей, а также имя пользователя и пароль. Этот опция может быть переопределена на отдельных узлах моделирования и в слепках моделей.

Примечание: Соединение с базой данных, используемое для целей моделирования, может быть, но может и не быть тем же, что и соединение, используемое для обращения к данным. Например, у вас может быть поток, обращающийся к данным из базы данных Oracle, скачивающий данные в IBM SPSS Modeler для очистки или иных видов обработки, а затем закачивающий эти данные в другую базу данных Oracle для целей моделирования. Но возможен случай, когда исходные данные находятся в плоском файле или другом источнике (не Oracle), и тогда требуется их закачивание в Oracle для моделирования. В всех случаях данные будут автоматически выгружаться во временную таблицу, созданную в используемой для моделирования базе данных.

Предупреждать о перезаписи модели Oracle Data Mining. Включите эту опцию, чтобы IBM SPSS Modeler не перезаписывал без предупреждения модели, хранимые в базе данных.

Получить список моделей Oracle Data Mining. Выводит доступные модели исследования данных.

Включить запуск Oracle Data Miner. (необязательная) Эта опция, если она включена, разрешает IBM SPSS Modeler запуск прикладной программы Oracle Data Miner. Дополнительную информацию смотрите в разделе “Oracle Data Miner” на стр. 48.

Путь для выполняемого файла Oracle Data Miner. (необязательная) Задает физическое положение Oracle Data Miner для выполняемого файла Windows (например: *C:\odm\bin\odminerw.exe*). Программа Oracle Data Miner не устанавливается с IBM SPSS Modeler; вы должны скачать ее правильную версию с сайта Oracle (*http://www.oracle.com/technology/products/bi/odm/odminer.html*) и установить на клиенте.

Включение поддержки генерирования и оптимизации SQL

1. Выберите в меню IBM SPSS Modeler:
   Инструменты > Свойства потока > Опции
2. На панели навигации щелкните по опции **Оптимизация**.
3. Подтвердите включение опции**Генерировать SQL**. Этот параметр требуется для работы функций моделирования баз данных.
4. Выберите **Оптимизировать построение SQL** и **Оптимизировать другое выполнение** (не требуется строго, но настоятельно рекомендуется для оптимальной производительности).

Построение моделей с использованием Oracle Data Mining (ODM)

За несколькими исключениями, узлы построения моделей Oracle работают так же, как остальные узлы моделирования в IBM SPSS Modeler. Эти узлы доступны на палитре моделирования базы данных у нижнего края окна IBM SPSS Modeler.

Данные

Для Oracle требуется, чтобы категориальные данные хранились в строковом формате (CHAR bkb VARCHAR2). В результате IBM SPSS Modeler не разрешит в качестве входных полей для моделей ODM указывать численные поля хранения с категориальным типом измерения *Флаг* или *Номинальный*. При необходимости числа можно преобразовать в строки в IBM SPSS Modeler при помощи узла переклассификации.

Поле назначения. Только одно поле можно выбрать как выходное (поле назначения) в моделях классификации ODM.

Имя модели. Начиная с Oracle 11*g*R1, имя `unique` стало ключевым словом и не может служить именем пользовательской модели.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, О-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Общие комментарии

- Экспорт-импорт PMML из IBM SPSS Modeler не поддерживается для моделей, созданных Oracle Data Mining.
- Скоринг моделей всегда происходит внутри ODM. Иногда набор данных нужно закачать во временную таблицу, если эти данные берутся из IBM SPSS Modeler или проходят в нем подготовку.
- В IBM SPSS Modeler обычно дается только одно прогнозируемое значение и соответствующая вероятность или показатель достоверности.
- В IBM SPSS Modeler для построения и скоринга модели можно использовать не более 1000 полей.
- IBM SPSS Modeler может оценивать модели ODM из потоков, опубликованных для выполнения при помощи IBM SPSS Modeler Solution Publisher.

Опции сервера моделей Oracle

Задайте соединение с Oracle, используемое для закачивания данных для моделирования. При необходимости можно выбрать соединение на вкладке Сервер для каждого узла моделирования, чтобы переопределить соединение с Oracle по умолчанию, заданное в диалоговом окне Вспомогательные прикладные программы. Дополнительную информацию смотрите в разделе “Включение интеграции с Oracle” на стр. 30.

Комментарии

- Соединение, используемое для моделирования, может совпадать или не совпадать с соединением, используемым в узле источника для потока. Например, у вас может быть поток, обращающийся к данным из базы данных Oracle, закачивающий данные в IBM SPSS Modeler для очистки или иных видов обработки, а затем закачивающий эти данные в другую базу данных Oracle для целей моделирования.
- Имя источника данных ODBC эффективно встраивается в каждом потоке IBM SPSS Modeler. Если поток, созданный на одном хосте, выполняется на другом хосте, имя источника данных на этих хостах должно быть одним и тем же. Можно также выбрать другой источник данных на вкладке Сервер каждого узла источника или моделирования.

Стоимости ошибочной классификации

В некоторых контекстах определенные виды ошибок обходятся пользователю дороже других. Например, может оказаться более дорогостоящим классифицировать претендента на кредит с высоким уровнем риска, как с низким уровнем риска (один вид ошибки), чем классифицировать претендента на кредит с низким уровнем риска риском как с высокими уровнем риска (другой вид ошибки). Стоимости ошибочной классификации позволяют задать относительную важность различных видов ошибок предсказания.

Стоимости ошибочной классификации - это по существу веса, применяемые к конкретным исходам. Эти веса факторизуются в модель и могут фактически изменить предсказание (в качестве способа защиты от дорогостоящих ошибок).

За исключением моделей C5.0, стоимости ошибочной классификации при скоринге моделей не применяются, и при ранжировании или сравнении моделей во внимание не принимаются. Модель, включающая в себя стоимости, не может дать меньше ошибок, чем та, которая не ранжируется и не может ранжироваться хоть сколь-нибудь выше в единицах общей точности, но, скорее всего, она будет выполняться лучше на практике, поскольку в ней заложено предусмотренное смещение в пользу *менее дорогостоящих* ошибок.

Матрица стоимостей показывает стоимость для каждого возможного сочетания предсказанной и действительной категорий. По умолчанию для всех стоимостей ошибочной классификации задается

значение 1,0. Чтобы ввести пользовательские значения стоимостей, выберите **Использовать стоимости ошибочной классификации** и введите в матрицу стоимостей нужные вам значения.

Чтобы изменить стоимость ошибочной классификации, выберите ячейку, соответствующую нужному сочетанию предсказанного и действительного значений, удалите существующее содержание ячейки и введите для нее желаемую стоимость. Стоимости не являются автоматически симметричными. Например, если для стоимости ошибочной классификации *A* как *B* задать значение 2,0, у стоимости ошибочной классификации *B* как *A* все равно будет значение по умолчанию 1,0, пока вы не измените также и его явным образом.

Примечание: Задавать стоимости во время построения разрешает только модель Деревья решений.

Наивный критерий Байеса Oracle

Наивный критерий Байеса - это общеизвестный алгоритм для проблем классификации. Модель названа *наивной*, поскольку она рассматривает все предлагаемые переменные предсказания как независимые друг от друга. Наивный критерий Байеса - быстрый, масштабируемый алгоритм, вычисляющий условные вероятности для сочетаний атрибутов и атрибута назначения. На основе обучающих данных оценивается независимая вероятность. Эта вероятность передает правдоподобие каждого класса назначения с учетом вхождения каждой категории значений из каждой входной переменной.

- Перекрестная проверка используется для испытания точности модели на тех же данных, которые использовались для построения этой модели. Эта проверка особенно полезна, если для построения модели доступно небольшое число наблюдений.
- Вывод модели можно просмотреть в форме матрицы. Числа в матрице - это условные вероятности, связывающие предсказанные классы (столбцы), и сочетания предикторов переменная - значение (строки).

Опции адаптивной байесовой модели

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, О-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Дополнительные опции наивного критерия Байеса

При построении модели отдельные значения атрибутов предикторов или пары значений игнорируются, если только не существует достаточного числа вхождений данного значения или пары в обучающих данных. Пороги для игнорирования значений задаются как доли на основе числа записей в обучающих данных. Настройка этих порогов может понизить шум и улучшить возможность обобщения модели на другие наборы данных.

- **Одиночный порог.** Задает порог для данного значения атрибута предиктора. Число вхождений указанного значения должно быть не меньше заданной доли, иначе оно будет проигнорировано.
- **Парный порог.** Задает порог для данной пары значений атрибута и предиктора. Число вхождений указанной пары значений должно быть не меньше заданной доли, иначе она будет проигнорирована.

Вероятность предсказания. Разрешает включать в модель вероятность правильного предсказания для возможного исхода поля назначения. Чтобы включить эту возможность, выберите действие **Выбрать**, нажмите кнопку **Задать**, выберите один из возможных исходов и нажмите кнопку **Вставить**.

Использовать набор предсказания. Генерирует таблицу всех возможных результатов для всех возможных исходов поля назначения.

Адаптивный критерий Байеса Oracle

Адаптивная сеть Байеса (Adaptive Bayes Network, ABN) создает классификаторы байесовских сетей при помощи MDL (Minimum Description Length - минимальная длины описания) и автоматического отбора показателей. ABN дает хорошие результаты в определенных ситуациях, где наивный критерий Байеса не работает достаточно хорошо, и получает не менее хорошие результаты в большинстве остальных случаев, хотя производительность может быть ниже. Алгоритм ABN предоставляет возможность построения трех типов расширенных моделей на основе байесовских, включая упрощенные (однофункциональные) модели дерева решений, сокращенные модели наивного критерия Байеса и многофункциональные модели с применением бустинга.

Сгенерированные модели

ABN в однофункциональном режиме построения генерирует (на основе набора удобочитаемых правил) упрощенное дерево решений, позволяющее бизнес-пользователю или аналитику понять обоснование предсказаний модели и соответственно выполнить действие или объяснить его другим. Это может оказаться существенным преимуществом по сравнению с моделями наивного критерия Байеса и многофункциональными моделями. Указанные правила можно просмотреть в IBM SPSS Modeler подобно стандартному набору правил. Простой набор правил может выглядеть следующим образом:

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

Сокращенные модели наивного критерия Байеса и многофункциональные модели в IBM SPSS Modeler просмотреть нельзя.

Опции адаптивной байесовой модели

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, O-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Тип модели

Для построения модели можно выбрать один из трех различных режимов.

- **Многофункциональный.** Выполняет построение и сравнение ряда моделей, включая модель NB плюс одно- и многофункциональные модели произведения вероятностей. Этот самый всеобъемлющий режим, и поэтому он требует больше всего времени на вычисление. Правила генерируются, только если лучшей оказывается однофункциональная модель. При выборе многофункциональной модели или модели NB никакие правила не генерируются.

- **Однофункциональный.** Создает упрощенное дерево решений на основе набора правил. Каждое правило содержит условие наряду с вероятностями, связанными с каждым исходом. Эти правила исключают друг друга и задаются в формате, удобном для чтения пользователями, что может оказаться существенным преимуществом по сравнению с моделями наивного критерия Байеса и многофункциональными моделями.
- **Наивный Байес.** Выполняет построение модели NB и ее сравнение с априорной вероятностью глобальной выборки (распределением значений назначения в глобальной выборке). Модель NB генерируется как выходная, только если она оказывается лучшим предиктором значений назначения по сравнению с глобальной априорной вероятностью. В противном случае в качестве выходной никакая модель не генерируется.

Дополнительные опции адаптивного критерия Байеса

Ограничить время выполнения. Выберите эту опцию, чтобы задать максимальное время построения в минутах. Это сделает возможным генерирование моделей за меньшее время, хотя итоговая модель может получиться менее точной. В каждой контрольной точке процесса моделирования алгоритм перед продолжением обработки проверяет, сможет ли он выполнить следующий этап за заданное время, и возвращает лучшую модель, доступную на момент достижения предела.

Максимальное число предикторов. Эта опция позволяет ограничить сложность модели и повысить производительность, благодаря ограничению числа используемых предикторов. Предикторы ранжируются на основе меры MDL их корреляции с назначением в качестве меры их правдоподобия, включаемого в модель.

Максимальное число предикторов наивного критерия Байеса. Эта опция задает максимальное число предикторов, используемых в модели наивного критерия Байеса.

Метод опорных векторов (Support Vector Machine, SVM) Oracle

Метод опорных векторов (SVM) - это алгоритм классификации и регрессии, использующий теорию машинного обучения для максимизации точности предсказания без переобучения. SVM использует необязательное нелинейное преобразование обучающих данных с последующим поиском уравнений регрессии в преобразованных данных для разделения классов (в случае категориальных назначений) или подгонки назначения (в случае непрерывных назначений). Реализация Oracle SVM разрешает построение моделей при помощи одного из двух доступных ядер: линейного или гауссово. Линейное ядро полностью опускает нелинейное преобразование, поэтому полученная модель является по существу регрессионной моделью.

Дополнительную информацию смотрите в *руководстве разработчика прикладных программ Oracle Data Mining* и разделе *Понятия Oracle Data Mining*.

Опции модели SVM Oracle

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, O-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Активное обучение. Предоставляет способ работы с большими наборами построения. При активном обучении этот алгоритм создает начальную модель на основе небольшой выборки перед применением к полному обучающему набору данных, а затем выполняет инкрементное обновление выборки и модели на основе полученных результатов. Цикл повторяется, пока не будет получена сходимость модели на обучающих данных или достигнуто максимально допустимое число опорных векторов.

Функция ядра. Выберите **Линейное** или **Гауссово** либо оставьте значение по умолчанию **Определяется системой**, чтобы разрешить системе выбрать наиболее подходящее ядро. Гауссовы ядра пригодны также к изучению более сложных взаимосвязей, но в общем случае требуют больше времени на вычисление. Возможно, вы захотите начать с линейного ядра и попробовать применить гауссово ядро, только если линейному ядру не удастся найти точную подгонку. С большей вероятностью это может произойти с регрессионной моделью, где значение выбора ядра выше. Кроме того, имейте в виду, что модели SVM, построенные с гауссовым ядром, нельзя просмотреть в IBM SPSS Modeler. Модели, подстроенные с линейным ядром, можно просмотреть в IBM SPSS Modeler тем же способом, что и стандартные регрессионные модели.

Метод нормализации. Задает метод нормализации для непрерывных входных полей и полей назначения. Можно выбрать **Z-оценка**, **Мин-Макс** или **Нет**. Если переключатель **Автоматическая подготовка данных** включен, Oracle выполняет нормализацию автоматически. Чтобы выбрать метод нормализации вручную, выключите этот переключатель.

Дополнительные опции SVM Oracle

Размер кэша ядра. Задает размер кэша (в байтах), который будет использоваться для хранения вычисленных ядер во время операции построения. Как можно ожидать, кэши большего размера обычно приводят к более быстрым операциям построения. Значение по умолчанию - 50 Мбайт.

Допуск для сходимости. Задает значение допуска, допустимое перед прекращением построения модели. Это значение должно быть между 0 и 1. Значение по умолчанию - 0,001. С ростом значений ускоряется построение моделей, но понижается их точность.

Задать среднеквадратичное отклонение. Задает параметр среднеквадратичного отклонения, используемый гауссовым ядром. Этот параметр влияет на компромисс между сложностью модели и возможностью ее обобщения на другие наборы данных (переобучение и недообучение данных). Более высокие значения среднеквадратичного отклонения благоприятствует недообучению. По умолчанию этот параметр оценивается по обучающим данным.

Задать эпсилон. (Только для регрессионных моделей) Задает значение интервала допустимой ошибки при построении моделей, нечувствительных к эпсилон. Другими словами, это значение позволяет отличить небольшие (игнорируемые) ошибки от больших (не игнорируемых) ошибок. Это значение должно быть между 0 и 1. Значение по умолчанию оценивается по обучающим данным.

Задать показатель сложности. Задает показатель сложности, устанавливающий компромисс между модельной ошибкой (измеряемой по обучающим данным) и сложностью модели с целью избежать переобучения или недообучения данных. Более высокие значения устанавливают больший штраф на ошибки, с повышенным риском переобучения данных; меньшие значения устанавливают на ошибки меньший штраф и могут привести к недообучению.

Задать норму выбросов. Задает нужную норму выбросов в обучающих данных. Допустимо только для моделей SVM одного класса. С параметром **Задать показатель сложности** использовать нельзя.

Вероятность предсказания. Разрешает включать в модель вероятность правильного предсказания для возможного исхода поля назначения. Чтобы включить эту возможность, выберите действие **Выбрать**, нажмите кнопку **Задать**, выберите один из возможных исходов и нажмите кнопку **Вставить**.

Использовать набор предсказания. Генерирует таблицу всех возможных результатов для всех возможных исходов поля назначения.

Опции весов SVM Oracle

В модели классификации при помощи весов можно задать относительную важность различных возможных значений назначения. Это может оказаться полезным, например, если точки в данных обучения не распределены реалистически по категориям. Веса позволяют сместить модель, чтобы можно было скомпенсировать категории, хуже представленные в данных. С увеличением веса для значения назначения должен возрастать процент правильных предсказаний для данной категории.

Значения веса можно задать тремя способами:

- **На основе обучающих данных.** Это опция по умолчанию. Веса основываются на относительной встречаемости категорий в данных обучения.
- **Равные для всех классов.** Для всех категорий веса определяются как 1/*k*, где *k* - число категорий назначения.
- **Пользовательская.** Можно задать свои собственные веса. Начальные значения для весов задаются равными для всех категорий. Веса для отдельных категорий можно скорректировать с учетом пользовательских значений. Чтобы скорректировать вес конкретной категории, выберите ячейку веса в таблице, соответствующей нужной категории, удалите содержание ячейки и введите желаемое значение.

Веса для всех категорий в сумме должны составлять 1,0. Если их сумма не равна 1,0, выводится предупреждение с опцией автоматической нормализации значений. Такая автоматическая корректировка сохраняет соотношения по категориям одновременно с применением ограничения по весу. Эту корректировку можно выполнить в любое время, нажав кнопку **Нормализовать**. Чтобы восстановить в таблице равные значения для всех категорий, нажмите кнопку **Уравнять**.

Обобщенные линейные модели Oracle (ОЛМ)

(Только для 11g) Обобщенные линейные модели ослабляют ограничивающие допущения, накладываемые линейными моделями. Это касается, например, допущений, что у переменной назначения должно быть нормальное распределение и что влияние предикторов на переменную назначения по своему характеру линейно. Обобщенная линейная модель удобна для предсказаний, где у назначения, скорее всего, будет распределение, отличающееся от нормального распределения, например, полиномиальное или распределение Пуассона. Таким же образом, обобщенная линейная модель полезна в случаях, где взаимосвязь (или связь) между предикторами и назначением, скорее всего, будет отличаться от линейной.

Дополнительную информацию смотрите в *руководстве разработчика прикладных программ Oracle Data Mining* и в разделе *Понятия Oracle Data Mining*.

Опции модели ОЛМ Oracle

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, О-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Метод нормализации. Задает метод нормализации для непрерывных входных полей и полей назначения. Можно выбрать **Z-оценка**, **Мин-Макс** или **Нет**. Если переключатель **Автоматическая подготовка данных** включен, Oracle выполняет нормализацию автоматически. Чтобы выбрать метод нормализации вручную, выключите этот переключатель.

Обработка пропущенных значений. Указывает, как обрабатывать пропущенные значения во входных данных.

- **Заменить на среднее значение и моду** заменяет пропущенные значения числовых атрибутов на среднее значение, а категориальных атрибутов - на моду.
- **Использовать только полные записи** игнорирует записи с пропущенными значениями.

Дополнительные опции ОЛМ Oracle

Использовать веса строк. Включите этот переключатель, чтобы активировать соседний выпадающий список, в котором можно выбрать столбец, содержащий коэффициент взвешивания для строк.

Сохранить диагностику строк в таблице. Включите этот переключатель, чтобы активировать соседнее текстовое поле, где можно ввести имя таблицы для включения в нее диагностики на уровне строк.

Уровень доверия коэффициента. Степень достоверности (от 0,0 до 1,0) нахождения значения, предсказанного для назначения, в доверительном интервале, вычисленном моделью. Доверительные пределы возвращаются со статистикой коэффициента.

Опорная категория для назначения. Выберите **Пользовательское**, чтобы выбрать для поля назначения значение, которое будет использоваться в качестве опорной категории, или оставьте значение п умолчанию **Авто**.

Гребневая регрессия. Гребневая регрессия - это метод, устраняющий ситуацию, где слишком высока степень корреляции в переменных. Вы можете, применив опцию **Авто**, разрешить алгоритму управлять использованием этого метода, либо управлять им вручную посредством опций **Отключить** и **Включить**. Если вы выбрали включение гребневой регрессии вручную, можно переопределить системное значение по умолчанию для параметра гребневой регрессии (ridge), введя значение в смежном поле.

Сгенерировать КРД для гребневой регрессии. Включите этот переключатель, если хотите генерировать статистику Коэффициент разбухания регрессии (КРД) при использовании для линейной регрессии параметра ridge.

Вероятность предсказания. Разрешает включать в модель вероятность правильного предсказания для возможного исхода поля назначения. Чтобы включить эту возможность, выберите действие **Выбрать**, нажмите кнопку **Задать**, выберите один из возможных исходов и нажмите кнопку **Вставить**.

Использовать набор предсказания. Генерирует таблицу всех возможных результатов для всех возможных исходов поля назначения.

Опции весов ОЛМ Oracle

В модели классификации при помощи весов можно задать относительную важность различных возможных значений назначения. Это может оказаться полезным, например, если точки в данных обучения не распределены реалистически по категориям. Веса позволяют сместить модель, чтобы можно было скомпенсировать категории, хуже представленные в данных. С увеличением веса для значения назначения должен возрастать процент правильных предсказаний для данной категории.

Значения веса можно задать тремя способами:

- **На основе обучающих данных.** Это опция по умолчанию. Веса основываются на относительной встречаемости категорий в данных обучения.
- **Равные для всех классов.** Для всех категорий веса определяются как $1/k$, где k - число категорий назначения.

- **Пользовательская.** Можно задать свои собственные веса. Начальные значения для весов задаются равными для всех категорий. Веса для отдельных категорий можно скорректировать с учетом пользовательских значений. Чтобы скорректировать вес конкретной категории, выберите ячейку веса в таблице, соответствующей нужной категории, удалите содержание ячейки и введите желаемое значение.

Веса для всех категорий в сумме должны составлять 1,0. Если их сумма не равна 1,0, выводится предупреждение с опцией автоматической нормализации значений. Такая автоматическая корректировка сохраняет соотношения по категориям одновременно с применением ограничения по весу. Эту корректировку можно выполнить в любое время, нажав кнопку **Нормализовать**. Чтобы восстановить в таблице равные значения для всех категорий, нажмите кнопку **Уравнять**.

Дерево решений Oracle

Oracle Data Mining предлагает классическую возможность Дерево решений на основе общепринятого алгоритма Дерево классификации и регрессии. Модель Дерево решений ODM содержит полную информацию о каждом узле, включая доверительную вероятность, поддержку и критерий разбиения. Для каждого узла может выводиться полное правило, а также предоставляется атрибут идентификатора объекта, подлежащий использованию в качестве атрибута при применении модели к наблюдению с пропущенными значениями.

Деревья решений широко используются из-за их повсеместной применимости, легкости применения и простоты понимания. Деревья решений тщательно проверяют каждый потенциальный входной атрибут, выполняющий поиск лучшего “разделителя” (то есть точки отсечения атрибута, например: `AGE > 55`), который разбивает записи данных нисходящего потока на несколько однородных совокупностей. После каждого решения разбиения ODM повторяет процесс, ведущий к росту всего дерева и созданию конечных “листьев”, представляющих схожие совокупности записей, позиций или людей. Начиная с корневого узла (например, узла общей совокупности), деревья решений обеспечивают удобочитаемые правила операторов типа `IF A, then B`. Эти правила деревьев решений обеспечивают также поддержку и предоставляют доверительную вероятность для каждого узла дерева.

Адаптивные байесовские сети могут также предоставить короткие простые правила, полезные при задании описаний для каждого предсказания, но деревья решений предоставляют одновременно полные правила Oracle Data Mining для каждого решения разбиения. Кроме того, деревья решений полезны для разработки подробных профилей наиболее выгодных заказчиков, платежеспособных пациентов, факторов, связанных с мошенничеством, и так далее.

Опции модели дерева решений

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, O-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Метрика неоднородности. Указывает, какой показатель будет использоваться для поиска лучшего контрольного вопроса для разбиения данных на каждом узле. Лучший разделитель и значение разбиения

приводят к максимальному повышению однородности значений назначения для объектов на узле. Однородность измеряется соответственно показателю. Поддерживаются показатели **gini** и **entropy**.

Опции эксперта по деревьям решений

Максимальная глубина. Задает максимальную глубину построения модели дерева.

Минимальный процент записей в узле. Задает процент минимального числа записей для одного узла.

Минимальный процент записей для разбиения. Задает минимальную долю записей на родительском узле в процентах от общего числа записей, используемых для обучения модели. Если число записей соответствует меньшему проценту, никаких попыток разбиения не предпринимается.

Минимальное число записей в узле. Задает минимальное число возвращаемых записей.

Минимальное число записей для разбиения. Задает минимальное число записей на родительском узле, выраженное значением. Если число записей соответствует меньшему значению, никаких попыток разбиения не предпринимается.

Идентификатор правила. Эта опция, если она включена, включает в модель строку для идентификации узла в дереве, на котором производится конкретное разбиение.

Вероятность предсказания. Разрешает включать в модель вероятность правильного предсказания для возможного исхода поля назначения. Чтобы включить эту возможность, выберите действие **Выбрать**, нажмите кнопку **Задать**, выберите один из возможных исходов и нажмите кнопку **Вставить**.

Использовать набор предсказания. Генерирует таблицу всех возможных результатов для всех возможных исходов поля назначения.

О-кластер Oracle

Алгоритм О-кластер Oracle выявляет естественные группировки в совокупности данных. Кластеризация с ортогональным разделением (О-кластер) - это собственный алгоритм кластеризации Oracle, создающий иерархическую модель кластеризации на основе сетки; то есть создающий в пространстве входных атрибутов разделы, параллельные осям (ортогональные). Этот алгоритм работает рекурсивно. Полученная иерархическая структура представляет нерегулярную сетку, мозаично разбивающую пространство атрибутов на кластеры.

Алгоритм О-кластер обрабатывает как числовые, так и категориальные атрибуты, а ODM автоматически выбирает лучшие определения кластеров. ODM предоставляет подробную информацию о кластерах, правила кластеров, значения центроидов кластеров и может использоваться для оценки совокупности по принадлежности к кластерам.

Опции модели О-кластер

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, О-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM

автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Максимальное число кластеров. Задает максимальное число генерируемых кластеров.

Дополнительные опции O-кластера

Максимальный буфер. Задает максимальный размер буфера.

Чувствительность. Задает долю, определяющую пиковую плотность, которая требуется для отделения нового кластера. Эта доля связана с глобальной плотностью равномерного распределения.

K-средние Oracle

Алгоритм K-средние Oracle выявляет естественные группировки в совокупности данных. Алгоритм k-средних Oracle - это алгоритм кластеризации на основе расстояний, разделяющий данные на заранее заданное число кластеров (при наличии достаточного количества несхожих наблюдений). Алгоритмы на основе расстояний опираются на метрику (функцию) расстояния для получения меры подобия между точками данных. Точки данных назначаются в ближайший кластер в соответствии с используемой метрикой расстояния. ODM предоставляет расширенную версию k-средних.

Алгоритм k-средних поддерживает иерархические кластеры, обрабатывает числовые и категориальные атрибуты и включает совокупность в число кластеров, заданное пользователем. ODM предоставляет подробную информацию о кластерах, правила кластеров, значения центроидов кластеров и может использоваться для оценки совокупности по принадлежности к кластерам.

Опции модели k-средних

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, O-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Число кластеров. Задает число генерируемых кластеров.

Функция расстояния. Указывает, какая функция расстояния будет использоваться для кластеризации методом K-средних.

Критерий разбиения. Указывает, какой критерий разбиения будет использоваться для кластеризации методом K-средних.

Метод нормализации. Задает метод нормализации для непрерывных входных полей и полей назначения. Можно выбрать **Z-оценка**, **Мин-Макс** или **Нет**.

Дополнительные опции k-средних

Итерации. Задает число итераций для алгоритма k-средних.

Допуск для сходимости. Задает допуск для сходимости для алгоритма k-средних.

Число интервалов. Задает число интервалов на гистограмме атрибутов, генерируемой k-средними. Границы интервалов для каждого атрибута вычисляются глобальным образом по всему обучающему набору данных. Метод разделения на интервалы - равноширотный. У всех атрибутов - одно и то же число интервалов, за исключением атрибутов с одним значением, у которых всего один интервал.

Рост блоков. Задает фактор роста для памяти, выделенной под хранение данных кластера.

Минимальный процент поддержки атрибутов. Задает долю значений атрибутов, которые должны быть непустыми, чтобы атрибут был включен в описание правил для каждого кластера. Задание для этого параметра слишком большого значения в данных с пропущенными значениями может привести к очень коротким или даже пустым правилам.

Факторизация неотрицательных матриц (Nonnegative Matrix Factorization, NMF) Oracle

Факторизация неотрицательных матриц (NMF) полезна для сведения большого набора данных к репрезентативным атрибутам. По своему характеру подобный анализу главных компонентов (PCA), но способный обрабатывать большие объемы атрибутов и в модели аддитивного представления, MNF - это мощный, современный алгоритм исследования данных, пригодный для самых разнообразных вариантов использования.

NMF можно использовать для обращения больших объемов данных (таких как текстовые) в меньшие, более разреженные представления, сокращающие размерность данных (одна и та же информация может быть представлена с применением гораздо меньшего числа переменных). Вывод моделей NMF можно проанализировать при помощи контролируемых методов обучения, таких как SVM, или неконтролируемых методов изучения, например, методов кластерного анализа. Oracle Data Mining использует алгоритмы NMF и SVM для исследования неструктурированных текстовых данных.

Опции модели NMF

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, О-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Метод нормализации. Задает метод нормализации для непрерывных входных полей и полей назначения. Можно выбрать **Z-оценка**, **Мин-Макс** или **Нет**. Если переключатель **Автоматическая подготовка данных** включен, Oracle выполняет нормализацию автоматически. Чтобы выбрать метод нормализации вручную, выключите этот переключатель.

Дополнительные опции NMF

Задайте число возможностей. Задает число возможностей, подлежащих извлечению.

Стартовое число генератора псевдослучайных чисел. Задает начальное значение генератора псевдослучайных чисел для алгоритма NMF.

Число итераций. Задает число итераций для алгоритма NMF.

Допуск для сходимости. Задает допуск для сходимости алгоритма NMF.

Показать все возможности. Выводит ID возможности и доверительную вероятность для всех возможностей (вместо вывода этих значений только для лучшей возможности).

Априорный анализ Oracle

Алгоритм априорных значений выполняет обнаружение правил связывания в данных. Например: "если покупатель приобретет бритву и средство после бритья, то он купит крем для бритья с доверительной вероятностью 80%". Эту проблему исследования связывания можно разложить на две составляющие подпроблемы:

- Найти все сочетания товаров (называемые часто встречающимися наборами товаров), поддержка которых выше минимальной.
- Сгенерировать нужные правила при помощи часто встречающихся наборов товаров. Идея в том, что если, например, часто встречаются ABC и BC, правило "A подразумевает BC" сохраняется, если отношение `support(ABC)` к `support(BC)` будет не меньше минимальной доверительной вероятности. Имейте в виду, что у этого правила минимальная поддержка, поскольку часто встречается ABCD. Связывание ODM поддерживает только правила одного консеквента (ABC подразумевает D).

Число часто встречающихся наборов товаров регулируется параметрами минимальной поддержки. Число часто генерируемых правил регулируется числом часто встречающихся наборов товаров и параметром доверительной вероятности. Если для параметра доверительной вероятности задать слишком большое значение, в модели связывания могут оказаться часто встречающиеся наборы товаров, но не будет правил.

В ODM для алгоритма априорных значений используется реализация на основе SQL. Шаги генерирования кандидатов и вычисления поддержки реализуются при помощи запросов SQL. Специализированные структуры данных в памяти не используются. Запросы SQL точно настраиваются для частого запуска на сервере баз данных при помощи разнообразных советов.

Опции полей априорных значений

Все узлы моделирования содержат вкладку Поля, где можно задать поля для использования при построении модели.

Перед построением модели априорных значений нужно указать поля, которые вы хотите использовать в качестве позиций, исследуемых при моделировании связывания.

Использовать параметры узла типа. Эта опция указывает узлу, что следует использовать информацию о полях из узла Тип, расположенного выше. Это опция по умолчанию.

Использовать пользовательские параметры. Эта опция указывает узлу, что следует использовать информацию о полях, заданную здесь, а не ту, что задана на расположенных выше узлах Тип. После выбора этой опции задайте остальные поля в диалоговом окне, которое зависит от того, используется ли транзакционный формат.

Если вы *не используете* транзакционный формат, задайте:

- **Поля ввода.** Выберите входные поля. Это аналогично заданию для поля роли *Входное* на узле Тип.

- **Подмножества.** Это поле позволяет задать поле, с помощью которого данные будут разделяться на отдельные выборки для этапов обучения, испытания и проверки при построении модели.

Если вы *используете* транзакционный формат, задайте:

Использовать транзакционный формат. Используйте эту опцию, если хотите преобразовать данные, содержащие по одной строке на позицию в данные, содержащие по одной строке на наблюдение.

При выборе этой опции изменяются элементы управления полями в нижней части этого диалогового окна:

Для транзакционного формата задайте:

- **ID.** Выберите в списке поле ID. В качестве поля ID могут использоваться числовые или символические поля. Каждое уникальное значение этого поля должно означать конкретную единицу анализа. Например, в прикладной программе Потребительская корзина каждый ID может представлять отдельного покупателя. Для прикладной программы Анализ веб-журналов каждый ID может представлять компьютер (по IP-адресу) или пользователя (по данным входа в систему).
- **Содержимое.** Задайте поле содержимого для модели. Это поле содержит позицию, исследуемую при моделировании связывания.
- **Подмножества.** Это поле позволяет задать поле, с помощью которого данные будут разделяться на отдельные выборки для этапов обучения, испытания и проверки при построении модели. Создав модель при помощи одной выборки и испытав ее при помощи другой, можно получить надежный показатель качества обобщения модели на более крупные наборы данных, подобные текущим. Если при помощи узлов Тип или Раздел были определены несколько полей разделов, на вкладке Поля на каждом узле моделирования, где используется разделение, должно быть выбрано одно поле раздела. (Если присутствует только один раздел, он будет использоваться автоматически при всяком включении разделения.) Кроме того, учтите, что для применения выбранного раздела в анализе на вкладке Параметры модели для узла должно быть также включено разделение. (Выключение этой опции делает возможным отключение разделения без изменения значений параметров полей.)

Опции модели априорных значений

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, O-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Максимальная длина правила. Задает максимальное число предварительных условий для любого правила; целое число от 2 до 20. Это способ ограничения сложности правил. Если правила слишком сложны или специфичны или если ваш набор правил требует слишком много времени на обучение, попробуйте уменьшить значение этого параметра.

Минимальная достоверность. Задает минимальный уровень доверительной вероятности (значение между 0 и 1). Правила с доверительной вероятностью меньше указанного критерия отбрасываются.

Минимальная поддержка. Задает порог минимальной поддержки (значение между 0 и 1). Правила с доверительной вероятностью меньше указанного критерия отбрасываются.

Минимальная длина описания (Minimum Description Length, MDL) Oracle

Алгоритм минимальной длины описания (Minimum Description Length, MDL) Oracle помогает определить атрибуты с наибольшим влиянием на атрибут назначения. Часто знание атрибутов с наибольшим влиянием помогает лучше понять бизнес и управлять им и может помочь упростить операции моделирования. Кроме того, эти атрибуты могут указать на типы данных, которыми, возможно, вы захотите дополнить модели. MDL может использоваться, например, для выявления атрибутов процесса, наиболее важных для предсказания качества комплектующего изделия, факторов, связанных с оттоком клиентов, или генов, которые вероятнее всего связаны с конкретной болезнью.

Алгоритм MDL Oracle отбрасывает входные поля, которые он расценивает как маловажные для предсказания назначения. Затем при помощи остальных полей он строит слепок модели, который связывается с моделью Oracle, выводимой в Oracle Data Miner. При просмотре модели в Oracle Data Miner выводится диаграмма, показывающая остающиеся входные поля, ранжированные в порядке их значимости при предсказании назначения.

Отрицательный ранг указывает на шум. Входные поля с нулевым или отрицательным рангом не вносят вклад в предсказание и, скорее всего, их следует удалить из данных.

Чтобы вывести диаграмму:

1. Щелкните правой кнопкой мыши по слепку модели на палитре моделей и выберите **Просмотр**.
2. В окне модели нажмите кнопку для запуска Oracle Data Miner.
3. Соединитесь с Oracle Data Miner. Дополнительную информацию смотрите в разделе “Oracle Data Miner” на стр. 48.
4. На панели навигации Oracle Data Miner раскройте **Модели**, а затем **Важность атрибутов**.
5. Выберите правильную модель Oracle (у нее будет то же имя, что и у поля назначения, заданного вами в IBM SPSS Modeler). Если вы не уверены, какая модель верна, выберите папку Важность атрибутов и найдите модель по дате ее создания.

Опции модели MDL

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Поле уникальности. Задает поле, используемое для однозначной идентификации каждого наблюдения. Например, им может быть поле ID, такое как *ID покупателя*. IBM SPSS Modeler накладывает ограничение: поле ключа должно быть числовым.

Примечание: Для всех узлов Oracle, кроме полей Адаптивный критерий Байеса Oracle, O-кластер Oracle и Априорный анализ Oracle, это поле - необязательное.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Важность атрибутов Oracle (AI)

Цель важности атрибутов состоит в выявлении атрибутов, которые связаны в наборе данных с результатом, и степени, с которой они влияют на окончательный исход. Узел Важность атрибутов Oracle анализирует данные, находит шаблоны и предсказывает исходы или результаты со связанным доверительным уровнем.

Опции модели AI

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Автоматическая подготовка данных. (Только для 11*g*) Включает (по умолчанию) или отключает режим автоматической подготовки данных Oracle Data Mining. Если этот переключатель включен, ODM автоматически выполняет преобразования данных, требуемые указанным алгоритмом. Дополнительную информацию смотрите в разделе *Понятия Oracle Data Mining*.

Опции выбора AI

Вкладка Опции позволяет задать значения параметров по умолчанию для выбора или исключения входных полей в слепке модели. Затем модель можно добавить в поток для выбора поднабора полей, которые будут использоваться при дальнейшем построении модели. Кроме того, эти значения можно переопределить, выбрав дополнительные поля или отменив выбор таковых в браузере моделей поле генерирования модели. Однако значения параметров по умолчанию делают возможным применение слепка модели без дополнительных изменений, что может оказаться особенно полезным для работы со сценариями.

Доступны следующие параметры:

Все ранжированные поля. Выбирает поля на основе их ранжирования как *важных*, *пограничных* или *маловажных*. Для каждого ранга можно отредактировать метку, а также значения отсечения, при помощи которых будут назначаться записи для того или иного ранга.

Максимальное число полей. Выбирает *n* верхних полей на основе важности.

Важность больше чем. Выбирает все поля с важностью больше указанного значения.

Поле назначения всегда сохраняется, независимо от варианта выбора.

Вкладка Модель слепка модели AI

Вкладка Модель для слепка модели Важность атрибутов (AI) Oracle выводит ранг и важность всех входных полей и позволяет выбрать поля для фильтрации при помощи переключателей в столбце слева. При обработке потока остаются только выбранные поля (вместе с предсказанием назначения). Остальные входные поля отбрасываются. Варианты выбора по умолчанию основаны на опциях, заданных для узла моделирования, но вы можете выбрать дополнительные поля или отменить их выбор нужным вам образом.

- Чтобы отсортировать список по столбцу ранга, имени поля, важности или иному другому из выводимых столбцов, щелкните по его заголовку. Можно также выбрать нужную позицию в списке рядом с кнопкой Сортировать по и при помощи кнопок со стрелками вверх и вниз изменить порядок сортировки.
- При помощи панели инструментов можно включить или выключить все поля и открыть диалоговое окно Включить поля, где можно отобрать поля по рагу или важности. Кроме того, при щелчке по полю можно удерживать нажатой клавишу Shift или Ctrl, чтобы это поле было добавлено к уже выбранным.
- В пояснении под таблицей выводятся значения порогов для ранжирования входных полей как важных, пограничных или маловажных. Эти значения задаются в режиме моделирования.

Управление моделями Oracle

Модели Oracle добавляются на палитру моделей точно также, как и другие модели IBM SPSS Modeler и могут использоваться почти таким же способом. Однако есть несколько важных отличий ввиду того, что каждая модель Oracle, созданная в IBM SPSS Modeler, фактически ссылается на модель, хранящуюся на сервере баз данных.

Вкладка Сервер слепков моделей Oracle

При построении модели ODM через IBM SPSS Modeler создается модель в IBM SPSS Modeler, а также создается или заменяется модель в базе данных Oracle. Модель IBM SPSS Modeler такого рода ссылается на содержимое модели базы данных, хранящейся на сервере баз данных. IBM SPSS Modeler может выполнять проверку согласованности, сохраняя идентичную сгенерированную строку **ключей модели** и в модели IBM SPSS Modeler, и в модели Oracle.

Строка ключа для каждой модели Oracle выводится в столбце *Информация о модели* в диалоговом окне Получить список моделей. Строка ключа для модели IBM SPSS Modeler выводится как **Ключ модели** на вкладке Сервер модели IBM SPSS Modeler (при помещении в поток).

Кнопку Проверить на вкладке Сервер можно использовать для проверки совпадения ключей в модели IBM SPSS Modeler и в модели Oracle. Если модель с таким же именем в Oracle не найдена или ключи этих моделей не совпадают, это означает, что модель Oracle была удалена или повторно построена с момента построения модели в IBM SPSS Modeler.

Вкладка Сводка слепков моделей Oracle

На вкладке Сводка слепка модели выводится информация о самой модели (*Анализ*), об используемых в ней полях (*Поля*), значениях параметров, используемых при построении модели, (*Параметры построения*) и об обучении модели (*Сводка по обучению*).

При первом просмотре узла результаты вкладки Сводка свернуты. Чтобы увидеть нужные вам результаты, разверните соответствующие им элементы при помощи элемента управления расширением слева от них или выведите все результаты, нажав кнопку **Развернуть все**. Чтобы скрыть результаты по завершении их просмотра, сверните при помощи элемента управления расширением отдельные результаты, которые вы хотите скрыть, или сверните все результаты, нажав кнопку **Свернуть все**.

Анализ. Выводится информация о конкретной модели. Если в вашем исполнении в указанный слепок модели вложен узел анализа, в этом разделе появится также информация о выполненном анализе.

Поля. Список полей, используемых в качестве полей назначения и входных полей при построении модели.

Параметры компоновки. Содержит информацию об используемых при построении модели параметрах.

Сводная информация по обучению. Выводится тип модели, поток, используемый для ее создания, создавший ее пользователь, отметка времени построения модели и время, затраченное на ее построение.

Вкладка Параметры слепков моделей Oracle

На вкладке Параметры для слепка модели можно переопределить заданные значения определенных опций на узле моделирования для целей скоринга.

Дерево решений Oracle

Использовать стоимости ошибочной классификации. Определяет, использовать ли стоимости ошибочной классификации в модели Дерево решений Oracle. Дополнительную информацию смотрите в разделе “Стоимости ошибочной классификации” на стр. 32.

Идентификатор правила. Эта опция, если она выбрана (включена), добавляет столбец идентификатора правила в модель Дерево решений Oracle. Идентификатор правила идентифицирует узел в дереве, где производится конкретное разбиение.

NMF Oracle

Показать все возможности. Эта опция, если она выбрана (включена), выводит ID возможности и доверительную вероятность для всех возможностей (вместо вывода этих значений только для лучшей возможности) в модели NMF Oracle.

Список моделей Oracle

Кнопка Список моделей Oracle Data Mining открывает диалоговое окно со списком существующих моделей базы данных; в этом окне модели можно удалять. Это диалоговое окно доступно из диалогового окна Вспомогательные программы и из диалоговых окон построения, просмотра и применения для узлов, связанных с ODM.

Для каждой модели выводится такая информация:

- **Имя модели.** Имя модели; по нему сортируется список
- **Информация о модели.** Основная информация о модели, состоящая из даты и времени сборки и имени столбца назначения
- **Тип модели.** Имя алгоритма, который построил эту модель

Oracle Data Miner

Oracle Data Miner - это пользовательский интерфейс к Oracle Data Mining (ODM); он заменяет прежний пользовательский интерфейс IBM SPSS Modeler к ODM. Oracle Data Miner разработан, чтобы помочь аналитикам правильно применять алгоритмы ODM. Достижение этих целей реализуется несколькими способами:

- Пользователям требуется дополнительное содействие в применении методики, разрешающей и подготовку данных, и выбор алгоритмов. Oracle Data Miner соответствует этой потребности, предоставляя операции по исследованию данных для прохождения пользователями этой отвечающей указанным требованиям методики.
- В состав Oracle Data Miner входят усовершенствованная и расширенная эвристика построения моделей и мастера по преобразованию, уменьшающие вероятность ошибок при задании параметров модели и преобразования.

Определение соединения с Oracle Data Miner

1. Oracle Data Miner можно запустить с любого узла применения или построения Oracle и из любого диалогового окна вывода при помощи кнопки **Запустить Oracle Data Miner**.

Рисунок 2. Кнопка запуска Oracle Data Miner

2. Диалоговое окно Oracle Data Miner **Редактировать соединение** выводится для пользователя перед запуском внешней прикладной программы Oracle Data Miner (при условии, что опция Вспомогательные прикладные программы определена правильно).

   Примечание: Это диалоговое окно появляется, только когда нет заданного имени соединения.

   - Задайте имя соединения Data Miner и введите информацию о подходящем сервере Oracle 10gR1 или 10gR2. Этот сервер Oracle должен быть тем же сервером, который задан в IBM SPSS Modeler.
3. В диалоговом окне Oracle Data Miner **Выберите соединение** предоставляются опции для указания того, какое имя соединения (определенное на шаге выше) следует использовать.

На сайте Oracle по адресу Oracle Data Miner можно найти дополнительную информацию о требованиях к Oracle Data Miner, о его установке и использовании.

Подготовка данных

Два типа подготовки данных могут быть полезны, когда для моделирования используются такие алгоритмы Oracle Data Mining, как Наивный Байес, Адаптивный Байес и метод опорных векторов:

- **Категоризация**, или преобразование непрерывных полей числового диапазона в категории для алгоритмов, которые не принимают непрерывных данных.
- **Нормализация**, или преобразование числовых диапазонов, чтобы получить сходные средние значения и стандартные отклонения.

Категоризация

Узел категоризации IBM SPSS Modeler поддерживает ряд методов категоризации. Можно применить категоризацию к одному полю или к нескольким. Если выполнить категоризацию для набора данных, создаются пороговые значения, и можно создать узел вычисления IBM SPSS Modeler. Операцию вычисления можно преобразовать в SQL и применить перед построением и оценкой модели. При таком подходе создается зависимость между моделью и узлом вычислений, так что не только выполняется категоризация, но и возникает возможность повторно использовать спецификации категоризации в нескольких задачах моделирования.

Нормализация

Прежде чем использовать непрерывные поля (поля числового диапазона) как входные поля в моделях по методу опорных векторов, эти поля сначала нужно нормализовать. Кроме того, для моделей регрессии требуется обратить нормализацию, чтобы реконструировать оценки по выходной информации модели. Параметры модели SVM дают возможность выбрать **Z-значения**, **мин-макс** или **нет**. Коэффициенты нормализации создаются Oracle как шаг в процессе построения модели, закачиваются в IBM SPSS Modeler и сохраняются вместе с моделью. Во время применения коэффициенты преобразуются в вычислительные выражения IBM SPSS Modeler, используемые для подготовки данных к оценке перед передачей данных в модель. В этом случае нормализация тесно связана с задачей моделирования.

Примеры Oracle Data Mining

В состав продукта включен ряд потоков примера, иллюстрирующих использование ODM с IBM SPSS Modeler. Эти потоки можно найти в папке установки IBM SPSS Modeler в подкаталоге *\Demos\Database_Modelling\Oracle Data Mining*

Примечание: К папке Demos можно перейти из группы программ IBM SPSS Modeler в меню Пуск Windows.

Потоки в следующей таблице можно использовать вместе в последовательности как пример процесса исследования базы данных, использующего алгоритм SVM (Support Vector Machine - механизм векторов поддержки), поставляемый с Oracle Data Mining:

Таблица 4. Исследование баз данных - примеры потоков

Поток	Описание
1_upload_data.str	Используется для очистки и закачивания данных из плоского файла в базу данных.
2_explore_data.str	Предоставляет пример изучения данных с помощью IBM SPSS Modeler.
3_build_model.str	Строит модель, используя собственный алгоритм базы данных.
4_evaluate_model.str	Используется как пример оценки модели с помощью IBM SPSS Modeler.
5_deploy_model.str	Внедряет модель для оценки базы данных.

Примечание: Чтобы запустить этот пример, надо выполнять потоки по порядку. Кроме этого, узлы источников и моделирования в каждом потоке надо изменить, включив в них ссылку на допустимый источник данных для базы данных, которую вы хотите использовать.

Набор данных, используемый в этих примерах потоков, относится к прикладным программам обработки кредитных карт и представляет задачу классификации с различными категориальными и непрерывными предикторами. Дополнительную информацию об этом наборе данных смотрите в файле *crx.names* в той же папке, где находятся примеры потоков.

Этот набор данных доступен в репозитории UCI Machine Learning по адресу *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/*.

Пример потока: Закачка данных

Первый пример потока, *1_upload_data.str*, служит для очистки и закачки данных из плоского файла в Oracle.

Поскольку для Oracle Data Mining (ODM) требуется поле уникального ID, в этом начальном потоке при помощи узла вычислений в набор данных добавляется новое поле *ID* с уникальными значениями 1,2,3, использующее функцию IBM SPSS Modeler `@INDEX`.

Для обработки пропущенных значений используется узел заполнения, так что пустые поля при чтении текстового файла *crx.data* заменяются на значения *NULL*.

Пример потока: Просмотр данных

Второй пример потока, *2_explore_data.str*, служит для демонстрации того, как применить узел Аудит данных, чтобы получить обзор данных, включая сводную статистику и диаграммы.

Двойной щелчок по диаграмме в отчете аудита данных выводит более подробную диаграмму для детального изучения данного поля.

Пример потока: построение модели

В третьем примере потока, *3_build_model.str*, иллюстрируется построение модели в IBM SPSS Modeler. Дважды щелкните по узлу источник базы данных (с меткой CREDIT), чтобы указать источник данных. Чтобы задать настройки построения, дважды щелкните по узлу построения (он сначала помечен как CLASS, но после указания источника данных получает метку FIELD16).

На вкладке Модель диалогового окна:

1. Убедитесь, что для поля уникальности выбрано значение **ID**.
2. Убедитесь, что для функции ядра выбрано значение **Линейная**, а для метода нормализации - **Z-оценки**.

Пример потока: Оценка модели

Четвертый пример потока, *4_evaluate_model.str*, иллюстрирует преимущества использования IBM SPSS Modeler для моделирования в базе данных. После выполнения модели ее можно снова добавить в поток данных и оценить при помощи ряда инструментов, которыми располагает IBM SPSS Modeler.

Просмотр результатов моделирования

Присоедините к слепку модели узел таблицы для просмотра результатов. Поле **$O-field16** содержит предсказанное значение для *field16* по каждому наблюдению, а поле **$OC-field16** содержит показатель достоверности этого предсказания.

Оценка результатов модели

При помощи узла анализа можно создать матрицу совпадений, содержащую структуру соответствий между каждым предсказанным полем и его полем назначения. Чтобы увидеть результаты, запустите узел анализа.

При помощи узла оценки можно создать диаграмму выигрышей, предназначенную для демонстрации повышения точности при использовании этой модели. Чтобы увидеть результаты, запустите узел оценки.

Пример потока: Внедрение модели

Когда достигнута достаточная точность модели, эту модель можно внедрить для использования во внешних прикладных программах или для публикации в той же базе данных. В заключительном примере потока, *5_deploy_model.str*, данные считываются из таблицы `CREDITDATA` и затем оцениваются и публикуются в таблице `CREDITSCORES` при помощи узла публикатора *внедрить решение*.

Глава 5. Моделирование баз данных с использованием IBM InfoSphere Warehouse

IBM InfoSphere Warehouse и IBM SPSS Modeler

IBM InfoSphere Warehouse (ISW) предоставляет семейство алгоритмов исследования данных, включенных в реляционную СУБД IBM DB2. IBM SPSS Modeler предоставляет узлы, поддерживающие интеграцию следующих алгоритмов IBM:

- деревья решений
- правила связывания
- Демографическая кластеризация
- Кластеризация Коонена
- Правила последовательности
- Регрессия преобразования
- Линейная регрессия
- Полиномиальная регрессия
- наивная модель Байеса
- логистическая регрессия
- временные ряды

Более подробную информацию об этих алгоритмах смотрите в документации, прилагаемой к установке IBM InfoSphere Warehouse.

Требования для интеграции с IBM InfoSphere Warehouse

Следующие условия - это обязательные предварительные требования для проведения моделирования в базе данных при помощи InfoSphere Warehouse Data Mining. Чтобы убедится в их соблюдении, может потребоваться проконсультироваться с администратором баз данных.

- Программа IBM SPSS Modeler, работающая с установкой сервер IBM SPSS Modeler в Windows или UNIX
- IBM DB2 Data Warehouse Edition Версии 9.1

  или
- IBM InfoSphere Warehouse Enterprise Edition Версии 9.5
- Источник данных ODBC для соединения с DB2, как описано ниже.

Примечание: Моделирование баз данных и оптимизация SQL требуют, чтобы на компьютере IBM SPSS Modeler была включена возможность соединения с сервером IBM SPSS Modeler. При включенной возможности соединения можно обращаться к алгоритмам баз данных, выполнять обратный перенос SQL непосредственно с клиента сервером IBM SPSS Modeler и обращаться к серверу IBM SPSS Modeler. Проверьте текущее состояние лицензии, для чего в меню клиента сервером IBM SPSS Modeler выберите:

Справка > О программе > Дополнительные подробности

Если возможность соединения включена, на вкладке Состояние лицензии вы увидите опцию **Разрешение для сервера**.

Enabling Integration with IBM InfoSphere Warehouse

Чтобы включить IBM SPSS Modeler интеграцию с IBM InfoSphere Warehouse (ISW) Data Mining, надо сконфигурировать ISW, создать источник ODBC, включить интеграцию в диалоговом окне Вспомогательные программы IBM SPSS Modeler, а также включить генерирование и оптимизацию SQL.

Конфигурирование ISW

Чтобы установить и сконфигурировать ISW, следуйте инструкциям в руководстве *Установка InfoSphere Warehouse*.

Создание источника ODBC для ISW

Чтобы включить соединение между ISW и IBM SPSS Modeler, надо создать имя источника данных системы ODBS (data source name - DSN).

Перед созданием DSN вам надо понимать основы источников данных и драйверов ODBC, а также поддержку баз данных в IBM SPSS Modeler.

Если сервер IBM SPSS Modeler и IBM InfoSphere Warehouse Data Mining работают на разных хостах, создайте одинаковые DSN ODBC на каждом из этих хостов. Убедитесь, что вы используете одно и то же имя для этого DSN на каждом хосте.

1. Установите драйверы ODBC. Они доступны на установочном диске IBM SPSS Data Access Pack, поставляемом с этим выпуском. Запустите файл *setup.exe*, чтобы запустилась программа установки, и выберите все нужные драйверы. Следуйте инструкциям на экране, чтобы установить драйверы.
   a. Создайте DSN.

      Примечание: Последовательность меню зависит от используемой версии Windows.

      - **Windows XP.** В меню Пуск выберите **Панель управления**. Щелкните дважды по значку **Администрирование**, а затем - по значку **Источники данных (ODBC)**.
      - **Windows Vista.** В меню Пуск выберите **Панель управления**, затем выберите **Система**. Щелкните дважды по значку **Администрирование** и выберите **Источники данных (ODBC)**, затем выберите **Открыть**.
      - **Windows 7.** В меню Пуск выберите **Панель управления**, затем **Система и безопасность**, затем **Администрирование**. Выберите **Источники данных (ODBC)**, затем выберите **Открыть**.

   b. Щелкните по вкладке **Системное DSN**, а затем нажмите кнопку **Добавить**.
2. Выберите драйвер **SPSS OEM 6.0 DB2 Wire Protocol**.
3. Нажмите кнопку **Готово**.
4. В диалоговом окне Установка драйвера ODBC DB2 Wire Protocol:
   - Задайте имя источника данных.
   - В поле IP-адреса введите имя хоста сервера, на котором находится реляционная СУБД DB2.
   - Примите значение по умолчанию для порта TCP (50000).
   - Задайте имя базы данных, с которой вы будете соединяться.
5. Нажмите кнопку **Проверить соединение**.
6. В диалоговом окне Регистрация в DB2 Wire Protocol введите имя пользователя и пароль, данный вам администратором базы данных, а затем нажмите кнопку **OK**.

   Появится сообщение **Соединение установлено!**.

   Драйвер ODBC IBM DB2. Если ваш ODBC драйвер - это драйвер ODBC IBM DB2, следуйте указаниям ниже, чтобы создать DSN ODBC:
7. В окне Администратор источников данных ODBC откройте вкладку **DSN системы** и нажмите кнопку **Добавить**.
8. Выберите **ДРАЙВЕР ODBC IBM DB2** и нажмите кнопку **Готово**.
9. В окне ДРАЙВЕР ODBC IBM DB2 — Добавить введите имя источника данных, а затем для алиаса базы данных нажмите кнопку **Добавить**.
10. В окне Параметры CLI/ODBC — <Имя источника данных> на вкладке Источник данных введите ID пользователя и пароль, данный вам администратором базы данных, затем откройте вкладку **TCP/IP**.
11. На вкладке TCP/IP введите:

- Имя базы данных, с которой вы хотите соединиться.
- Алиас базы данных (не более восьми символов).
- Имя хоста сервера баз данных, с которым вы хотите соединиться.
- Номер порта для соединения.

12. Щёлкните по вкладке **Опции защиты** и выберите **Задать опции защиты (Необязательно)**, а затем примите опцию по умолчанию (**Использовать значение аутентификации в конфигурации DBM сервера**).
13. Щёлкните по вкладке **Источник данных** и нажмите кнопку **Соединить**.

Появится сообщение**Соединение проверено успешно**.

Сконфигурируйте ODBC для обратной связи (необязательно)

Чтобы получать обратную связь от IBM InfoSphere Warehouse Data Mining во время построения модели и разрешить IBM SPSS Modeler отменять построение модели, следуйте указаниям ниже по конфигурированию источника данных ODBC, который был создан в предыдущем разделе. Заметим, что этот шаг конфигурирования позволяет IBM SPSS Modeler читать данные DB2, которые, возможно, не были приняты на базе данных одновременно выполняющимися транзакциями. Если у вас сомнение по поводу последствий этого изменения, проконсультируйтесь с администратором вашей базы данных.

Драйвер SPSS OEM 6.0 DB2 Wire Protocol. Для драйвера соединения с ODBC выполните следующие действия:

1. Запустите администратор источников данных ODBC, выберите источник данных, который был создан в предыдущем разделе, и нажмите кнопку **Конфигурировать**.
2. В диалоговом окне Установка драйвера ODBC DB2 Wire Protocol щелкните по вкладке **Дополнительно**.
3. Задайте уровень изоляции по умолчанию **0-READ UNCOMMITED**(Чтение непринятого), а затем нажмите кнопку **OK**.

   Драйвер ODBC IBM DB2. Для драйвера IBM DB2 выполните следующие действия:
4. >Запустите администратор источников данных ODBC, выберите источник данных, который был создан в предыдущем разделе, и нажмите кнопку **Конфигурировать**.
5. В диалоговом окне Параметры CLI/ODBC щёлкните по вкладке **Дополнительные параметры**, а затем нажмите кнопку **Добавить**.
6. В диалоговом окне Добавить параметр CLI/ODBC выберите параметр **TXNISOLATION**, а затем нажмите кнопку **OK**.
7. В диалоговом окне Уровень изоляции выберите **Чтение непринятого**, затем нажмите кнопку **OK**.
8. В диалоговом окне Параметры CLI/ODBC нажмите кнопку **OK**, чтобы завершить конфигурацию.

Обратите внимание на то, что обратная связь от IBM InfoSphere Warehouse Data Mining появляется в следующем формате:

```
<номер_итерации> / <ход_выполнения> / <фаза_ядра>
```

где:

- `<номер_итерации>` означает номер текущего прохода по данным, начиная с 1.
- `<ход_выполнения>` означает ход выполнения текущей итерации в виде числа от 0.0 до 1.0.
- `<фаза_ядра>` описывает текущую фазу алгоритма исследования данных.

Включение интеграции IBM InfoSphere Warehouse Data Mining в IBM SPSS Modeler

Чтобы разрешить IBM SPSS Modeler использовать DB2 с IBM InfoSphere Warehouse Data Mining, сначала надо задать некоторые спецификации в диалоговом окне Вспомогательные программы.

1. Выберите в меню IBM SPSS Modeler:

   Инструменты > Опции > Вспомогательные программы

2. Щёлкните по вкладке **IBM InfoSphere Warehouse**.

Включите интеграцию с InfoSphere Warehouse Data Mining. Это включает палитру моделирования баз данных (если она еще не выводится) в нижней части окна IBM SPSS Modeler, а также добавляет узлы для алгоритмов ISW Data Mining.

Соединение DB2. Задаёт источник данных ODBC DB2 по умолчанию, используемый для построения и сохранения моделей. Этот параметр может быть переопределён при построении отдельной модели и сгенерированных узлы модели. Нажмите кнопку с многоточием (**...**), чтобы выбрать источник данных.

Соединение с базой данных, используемое для целей моделирования, может совпадать, но может и не совпадать с соединением, используемым для доступа к данным. Например, у вас мог бы быть поток, который получает доступ к данным из одной базы данных DB2, скачивает данные на IBM SPSS Modeler для их очистки или иных действий, а затем закачивает данные в другую базу данных DB2 для целей моделирования. Другой вариант - исходные данные могли бы находиться в плоском файле или источнике другого формата (не DB2); в таком случае понадобилось бы закачать в DB2 для моделирования. В любом случае данные будут автоматически закачаны во временную таблицу, созданную в базе данных, которая используется для моделирования, если это необходимо.

Предупреждать о перезаписи модели интеграции InfoSphere Warehouse Data Mining. Выберите эту опцию, чтобы быть уверенным, что модели, записанные в базе данных, не переписываются IBM SPSS Modeler без предупреждения.

Список моделей InfoSphere Warehouse Data Mining. Эта опция позволяет вам получать список моделей, записанных в DB2, или удалять эти модели. Дополнительную информацию смотрите в разделе “Перечисление моделей базы данных” на стр. 59.

Разрешить запуск InfoSphere Warehouse Data Mining Visualization. Если вы установили модуль визуализации, надо включить его здесь для использованияIBM SPSS Modeler.

Путь для выполняемого файла модуля визуализации. Положение выполняемого модуля визуализации (если он установлен), например, *C:\Program Files\IBM\ISWarehouse\Im\IMVisualization\bin\imvisualizer.exe*.

Каталог подключаемого модуля визуализации временных рядов. Положение подключаемого флеш-модуля визуализации временных рядов (если он установлен), например, *C:\Program Files\IBM\ISWShared\plugins\com.ibm.datatools.datamining.imvisualization.flash_2.2.1.v20091111_0915*.

Включите Расширенные опции InfoSphere Warehouse Data Mining. Вы можете задать предельный объем памяти для алгоритма исследования данных в базе данных и определить другие опции по вашему выбору в форме командной строки для конкретных моделей. Предел памяти позволяет контролировать выделение памяти и задавать значение для расширенной опции `-buf`. Здесь в командной строке можно задать и другие расширенные опции для передачи в IBM InfoSphere Warehouse Data Mining. Дополнительную информацию смотрите в разделе “Расширенные опции” на стр. 60.

Проверить версию InfoSphere Warehouse. Проверяет версию IBM InfoSphere Warehouse, которую вы используете, и сообщает об ошибке, если вы пытаетесь использовать возможность исследования данных, которая не поддерживается вашей версией.

Включение поддержки генерирования и оптимизации SQL

1. Выберите в меню IBM SPSS Modeler:

   Инструменты > Свойства потока > Опции
2. На панели навигации щелкните по опции **Оптимизация**.
3. Подтвердите включение опции**Генерировать SQL**. Этот параметр требуется для работы функций моделирования баз данных.

4. Выберите **Оптимизировать построение SQL** и **Оптимизировать другое выполнение** (не требуется строго, но настоятельно рекомендуется для оптимальной производительности).

Построение моделей при помощи IBM InfoSphere Warehouse Data Mining

Для построения моделей IBM InfoSphere Warehouse Data Mining требуется, чтобы обучающий набор данных располагался в таблице или в производной таблице в базе данных DB2. Если эти данные расположены не в DB2 или их нужно обработать в IBM SPSS Modeler как часть подготовки данных, которую нельзя выполнить в DB2, перед построением модели эти данные автоматически закачиваются во временную таблицу DB2.

Оценка и внедрение модели

Оценка модели всегда происходит в DB2 и всегда выполняется IBM InfoSphere Warehouse Data Mining. Может потребоваться закачать набор данных во временную таблицу, если данные созданы или должны подготавливаться в IBM SPSS Modeler. Для моделей дерева решений, регрессии или кластеризации из IBM SPSS Modeler обычно передается только одно предсказание и связанная вероятность или доверительный интервал. Кроме этого, пользовательская опция вывода доверительных интервалов для каждого возможного вывода (как и в логистической регрессии) - это опция времени оценки, доступная на вкладке Параметры слепка модели (переключатель **Включить доверительные интервалы для всех классов**). Для моделей Связывание и Последовательность из IBM SPSS Modeler передается несколько значений. IBM SPSS Modeler может оценить модели IBM InfoSphere Warehouse Data Mining изнутри потоков, опубликованных для выполнения, используя IBM SPSS Modeler Solution Publisher.

В следующей таблице объясняются поля, сгенерированные при оценке моделей.

Таблица 5. Поля оценки модели

Тип модели	**Столбцы оценки**	**Значение**
деревья решений	$I-*поле*	Наилучшее предсказание для *поля*.
	$IC-*поля*	Доверительный интервал наилучшего предсказания для *поля*.
	$IC-*значение1*, ..., $IC-*значениеN*	(необязательно) Доверительный интервал каждого из *N* возможных значений для *поля*.
Регрессия	$I-*поле*	Наилучшее предсказание для *поля*.
	$IC-*поле*	Доверительный интервал наилучшего предсказания для *поля*.
Кластеризация	$I-*имя_модели*	Наилучшее назначение кластера для входной записи.
	$IC-*имя_модели*	Доверительный интервал наилучшего назначения кластера для входной записи.
Взаимосвязь	$I-*имя_модели*	Идентификатор правила соответствия.
	$IH-*имя_модели*	Элемент следствия.
	$IHN-*имя_модели*	Имя элемента следствия.
	$IS-*имя_модели*	Значение поддержки правила соответствия.
	$IC-*имя_модели*	Значение достоверности правила соответствия.
	$IL-*имя_модели*	Значение подъема правила соответствия.

Таблица 5. Поля оценки модели (продолжение)

Тип модели	Столбцы оценки	Значение
	$IMB-*имя_модели*	Количество совпадающих элементов условия или наборов элементов условия (так как все элементы условия или наборы элементов условия должны входить в это количество, оно равно числу элементов условия или наборов элементов условия).
Порядковый номер	$I-*имя_модели*	Идентификатор правила соответствия
	$IH-*имя_модели*	Набор элементов следствия соответствующего правила
	$IHN-*имя_модели*	Имена элементов в наборе элементов следствия для соответствующего правила
	$IS-*имя_модели*	Значение поддержки для соответствующего правила.
	$IC-*имя_модели*	Значение достоверности для соответствующего правила
	$IL-*имя_модели*	Значение подъема для соответствующего правила
	$IMB-*имя_модели*	Количество совпадающих элементов условия или наборов элементов условия (так как все элементы условия или наборы элементов условия должны входить в это количество, оно равно числу элементов условия или наборов элементов условия)
наивная модель Байеса	$I-*поле*	Наилучшее предсказание для *поля*.
	$IC-*поля*	Доверительный интервал наилучшего предсказания для *поля*.
логистическая регрессия	$I-*поле*	Наилучшее предсказание для *поля*.
	$IC-*поля*	Доверительный интервал наилучшего предсказания для *поля*.

Управление моделями DB2

При построении модели IBM InfoSphere Warehouse Data Mining через IBM SPSS Modeler создается модель в IBM SPSS Modeler и создается или заменяется модель в базе данных DB2. Модель IBM SPSS Modeler такого рода ссылается на содержимое модели базы данных, хранящейся на сервере баз данных. IBM SPSS Modeler может выполнять проверку соответствия, сохраняя идентичную сгенерированную строку ключа модели и в модели IBM SPSS Modeler, и в модели DB2.

Строка ключа для каждой модели DB2 выводится в столбце *Информация о модели* в диалоговом окне Список моделей базы данных. Строка ключа для модели IBM SPSS Modeler выводится как Ключ модели на вкладке Сервер модели IBM SPSS Modeler (при помещении в поток).

Кнопку Проверить можно использовать для проверки, что ключи в модели IBM SPSS Modeler и в модели DB2 совпадают. Если модель с таким же именем в DB2 не найдена или ключи этих моделей не совпадают, это означает, что модель DB2 была удалена или повторно построена с момента построения модели в IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “Вкладка Сервер слепков моделей ISW” на стр. 75.

Перечисление моделей базы данных

IBM SPSS Modeler содержит диалоговое окно для перечисления моделей, хранимых в IBM InfoSphere Warehouse Data Mining, и позволяет включать и отключать эти модели. Это диалоговое окно доступно из диалогового окна вспомогательных прикладных программ IBM и в диалоговых окнах построения, просмотра и применения для относящихся к IBM InfoSphere Warehouse Data Mining узлов. Для каждой модели выводится следующая информация:

- Имя модели (имя модели, которая используется для сортировки списка).
- Информация о модели (информация о ключе модели, случайно сгенерированном при построении модели IBM SPSS Modeler).
- Тип модели (таблица DB2, в которой IBM InfoSphere Warehouse Data Mining сохранил эту модель).

Просмотр моделей

Инструмент Визуализатор - это единственный способ просмотра моделей исследования данных InfoSphere Warehouse Data Mining. Этот инструмент можно установить дополнительно с InfoSphere Warehouse Data Mining. Дополнительную информацию смотрите в разделе “Enabling Integration with IBM InfoSphere Warehouse” на стр. 53.

- Нажмите кнопку **Просмотр** для запуска инструмента визуализации. Что именно выводит этот инструмент, зависит от типа сгенерированного узла. Например, инструмент визуализации возвратит представление Предсказанные классы при запуске из слепка модели дерева решений ISW.
- Нажмите кнопку **Проверить результаты** (только для узлов Дерево решений и Последовательность), чтобы запустить инструмент визуализации и просмотреть сведения об общем качестве сгенерированной модели.

Экспорт моделей и генерирование узлов

В моделях IBM InfoSphere Warehouse Data Mining можно выполнять экспорт и импорт PMML. Экспортированный PMML - это исходный PMML, сгенерированный IBM InfoSphere Warehouse Data Mining. Функция экспорта возвращает модель в формате PMML.

Сводку модели и ее структуру можно экспортировать в текстовый файл и в файл формата HTML. В подходящих ситуациях вы можете сгенерировать нужные узлы Фильтр, Выбор и Извлечение. Более подробную информацию смотрите в разделе "Экспорт моделей" в *Руководстве пользователя IBM SPSS Modeler*.

Общие параметры узлов для всех алгоритмов

Следующие параметры - общие для многих алгоритмов IBM InfoSphere Warehouse Data Mining:

Назначение и предикторы. Назначение и предикторы можно задать на узле Тип или вручную, используя вкладку Поля узла построителя моделей, что обычно используется в IBM SPSS Modeler.

Источник данных ODBC. Этот параметр позволяет пользователю перезаписать источник данных ODBC по умолчанию для текущей модели. (Значение по умолчанию задано в диалоговом окне Вспомогательные прикладные программы. Дополнительную информацию смотрите в разделе “Enabling Integration with IBM InfoSphere Warehouse” на стр. 53.)

Опции вкладки Сервер ISW

Вы можете задать соединение с DB2, используемое для закачки данных для моделирования. При необходимости можно выбрать соединение на вкладке Сервер для каждого узла моделирования, чтобы перезаписать соединение с DB2 по умолчанию, заданное в диалоговом окне Вспомогательные прикладные программы. Дополнительную информацию смотрите в разделе “Enabling Integration with IBM InfoSphere Warehouse” на стр. 53.

Соединение, используемое для моделирования, может совпадать или не совпадать с соединением, используемым в узле источника для потока. Например, у вас может быть поток, который обращается к

данным одной базы данных DB2, скачивает данные в IBM SPSS Modeler для очистки или других действий и затем закачивает данные в другую базу данных DB2 для целей моделирования.

Имя источника данных ODBC эффективно встроено в каждый поток IBM SPSS Modeler. Если созданный на одном хосте поток выполняется на другом хосте, имена источника данных на этих хостах должны совпадать. Другой вариант - на вкладке Сервер каждого узла источника или моделирования можно выбрать разные источники данных.

При построении модели можно получить обратную связь, используя следующие опции:

- **Включить обратную связь.** Выберите эту опцию для получения обратной связи во время построения модели (значение по умолчанию - выключено).
- **Интервал обратной связи (в секундах).** Укажите, как часто IBM SPSS Modeler будет получать поддержку обратной связи по ходу построения модели.

Включить расширенные опции InfoSphere Warehouse Data Mining. Выберите эту опцию, чтобы выводилась кнопка **Расширенные опции**, позволяющая задавать несколько дополнительных опций, таких как предел памяти и пользовательский SQL. Дополнительную информацию смотрите в разделе “Расширенные опции”.

На вкладке Сервер для сгенерированного узла есть опция для выполнения проверки соответствия, при использовании которой идентичная сгенерированная строка ключа сохраняется и в модели IBM SPSS Modeler, и в модели DB2. Дополнительную информацию смотрите в разделе “Вкладка Сервер слепков моделей ISW” на стр. 75.

Расширенные опции

На вкладке Сервер для всех алгоритмов есть переключатель для включения расширенных опций моделирования ISW. Если нажать кнопку **Расширенные опции**, появится диалоговое окно Расширенные опции ISW, где можно задавать опции, определяющие:

- Предел памяти.
- Другие расширенные опции.
- Пользовательский SQL данных исследования.
- Пользовательский SQL логических данных.
- Пользовательский SQL параметров исследования.

Предел памяти. Ограничивает потребление памяти алгоритмом построения модели. Обратите внимание на то, что стандартная расширенная опция задает предел для числа дискретных значений категориальных данных.

Другие расширенные опции. Позволяет задавать расширенные опции по вашему выбору в виде командной строки для конкретных моделей или решений. Детали могут различаться в зависимости от реализации или решения. Можно вручную расширить сгенерированный IBM SPSS Modeler SQL, чтобы определить задачу построения модели.

Пользовательский SQL данных исследования. Вы можете добавить вызовы методов для изменения объекта `DM_MiningData`. Например, при вводе следующего SQL для данных, используемых при построении модели, добавляется фильтр на основе поля с названием *Partition*:

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

Пользовательский SQL логических данных. Вы можете добавить вызовы методов для изменения объекта `DM_LogicalDataSpec`. Например, следующий SQL удаляет поле из набора полей, используемых для построения модели:

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

Пользовательский SQL параметров исследования. Вы можете добавить вызовы методов для изменения объекта DM_ClasSettings/DM_RuleSettings/DM_ClusSettings/DM_RegSettings. Например, при вводе следующего SQL для IBM InfoSphere Warehouse Data Mining передается инструкция активировать поле *Partition* (то есть оно должно всегда включаться в полученную модель):

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

Опции стоимости ISW

На вкладке Стоимость можно настроить стоимости ошибочных классификаций, что позволит задать относительную важность различных видов ошибок прогнозирования.

В некоторых контекстах определенные виды ошибок обходятся пользователю дороже других. Например, может оказаться более дорогостоящим классифицировать претендента на кредит с высоким уровнем риска, как с низким уровнем риска (один вид ошибки), чем классифицировать претендента на кредит с низким уровнем риска риском как с высокими уровнем риска (другой вид ошибки). Стоимости ошибочной классификации позволяют задать относительную важность различных видов ошибок предсказания.

Стоимости ошибочной классификации - это по существу веса, применяемые к конкретным исходам. Эти веса факторизуются в модель и могут фактически изменить предсказание (в качестве способа защиты от дорогостоящих ошибок).

За исключением моделей C5.0, стоимости ошибочной классификации при скоринге моделей не применяются, и при ранжировании или сравнении моделей во внимание не принимаются. Модель, включающая в себя стоимости, не может дать меньше ошибок, чем та, которая не ранжируется и не может ранжироваться хоть сколь-нибудь выше в единицах общей точности, но, скорее всего, она будет выполняться лучше на практике, поскольку в ней заложено предусмотренное смещение в пользу *менее дорогостоящих* ошибок.

Матрица стоимостей показывает стоимость для каждого возможного сочетания предсказанной и действительной категорий. По умолчанию для всех стоимостей ошибочной классификации задается значение 1,0. Чтобы ввести пользовательские значения стоимостей, выберите **Использовать стоимости ошибочной классификации** и введите в матрицу стоимостей нужные вам значения.

Чтобы изменить стоимость ошибочной классификации, выберите ячейку, соответствующую нужному сочетанию предсказанного и действительного значений, удалите существующее содержание ячейки и введите для нее желаемую стоимость. Стоимости не являются автоматически симметричными. Например, если для стоимости ошибочной классификации *A* как *B* задать значение 2,0, у стоимости ошибочной классификации *B* как *A* все равно будет значение по умолчанию 1,0, пока вы не измените также и его явным образом.

Дерево решений ISW

Модели деревьев решений позволяют создать системы классификации, которые предсказывают или классифицируют будущие наблюдения на основе набора решающих правил. Если данные разделяются на интересующие вас классы (например, ссуды высокого и низкого риска, подписчики и не-подписчики, голосующие и неголосующие или типы бактерий), можно использовать существующие данные для построения правил, которые можно применять для классификации старых и новых наблюдений с максимальной точностью. Например, можно построить дерево, классифицирующее кредитные риски или намерение покупки на основании возраста и других факторов.

Алгоритм дерева решений ISW строит деревья классификаций для категориальных входных данных. В результате получается двоичное дерево решений. Для построения модели можно применить различные параметры, в том числе стоимость ошибочной классификации.

Визуализатор - это единственный инструмент для просмотра моделей IBM InfoSphere Warehouse Data Mining.

Опции модели дерева решений ISW

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если вы определяете поле разделения, выберите опцию **Использовать разделенные данные**.

Выполнить проход тестирования. Можно выбрать возможность выполнения прохода тестирования. В таком случае проход тестирования IBM InfoSphere Warehouse Data Mining выполняется после построения модели для обучающего раздела. При этом выполняется проход по испытательному разделу для установления информации о качестве модели, подъема диаграмм и так далее.

Максимальное количество уровней в дереве. Вы можете задать максимальное количество уровней в дереве. Количество допустимых уровней будет ограничено заданным значением. Если эта опция остается невыбранной, принудительное ограничение не вводится. Чтобы избегать переусложненных моделей, рекомендуется обычно использовать значения не больше пяти.

Опции ISW Decision Tree Expert

Максимальная чистота. Эта опция задает максимальную чистоту для внутренних узлов. Если разделение узла приводит к тому, что для одного из дочерних узлов превышается заданный показатель чистоты (например, больше чем 90% случаев попадают в указанную категорию), такой узел не будет разделен.

Минимальное число наблюдений на внутренний узел. Если разделение узла приводит к узлу с числом наблюдений меньше указанного минимума, такой узел не будет разделен.

Связывание ISW

Узел Связывание ISW можно использовать для поиска правил связывания среди элементов, присутствующих в наборе групп. Правила связывания связывают конкретный вывод (например, покупку некоторого продукта) с набором условий (например, с покупкой нескольких других продуктов).

Задавая **ограничения**, можно выбрать включение правил связывания в модель или их исключение из модели. Если вы выбираете включение конкретного поля ввода, правила связывания, содержащие по крайней мере один из заданных элементов, включаются в модель. Если вы исключаете поле ввода, правила связывания, содержащие любой из заданных элементов, отбрасываются из результатов.

Алгоритмы связывания и последовательности ISW могут использовать **таксономии**. Таксономии отображают отдельные значения на понятия более высокого уровня. Например, ручки и карандаши могут быть отображены в категорию канцелярских товаров.

У правил связывания есть один консеквент (вывод) и несколько антецедентов (набор условий). Пример:

```
[Хлеб, Джем]  ⌂  [Масло]
[Хлеб, Джем]
 ⌂  [Маргарин]
```

Здесь Хлеб и Джем - это антецеденты (их называют также **условием правила**), а Масло или Маргарин - примеры консеквентов (их называют также как **следствием правила**). Первое правило означает, что покупатель хлеба и джема в той же покупке покупает и масло. Второе правило означает, что покупатель при покупке того же сочетания (хлеб и джем) при том же посещении магазина покупает маргарин.

Визуализатор - это единственный инструмент для просмотра моделей IBM InfoSphere Warehouse Data Mining.

Опции полей связывания ISW

На вкладке Поля задаются поля, которые будут использоваться при построении модели.

Перед построением модели необходимо указать поля, которые должны служить полями назначения и входными полями. За немногими исключениями, все узлы моделирования будут использовать информацию о полях из узла Тип, расположенного выше. При использовании опции применения по умолчанию узла Тип для выбора входных полей и полей назначения единственный параметр, который можно изменить на этой вкладке, - это макет таблицы для нетранзакционных данных.

Использовать параметры узла типа. Эта опция указывает на использование информации о полях из узла Тип, расположенного выше. Это опция по умолчанию.

Использовать пользовательские параметры. Эта опция указывает на использование информации о полях, заданной здесь, а не той, что задана на любых расположенных выше узлах Тип. После выбора этого варианта задайте приведенные ниже поля, как это потребуется.

Использование транзакционного формата. Включите этот переключатель, если у исходных данных **транзакционный формат**. У записей в этом формате есть два поля, один для ID и один для содержимого. Каждая запись представляет единственную транзакцию или элемент, и связанные элементы связаны наличием одинакового ID. Выключите этот переключатель, если у данных **табличный формат**, в котором элементы представлены отдельными флагами, где каждое поле флага указывает на наличие или отсутствие конкретного элемента, а каждая запись представляет полный набор связанных элементов.

- **ID.** Для транзакционных данных выберите из списка поле ID. В качестве поля ID могут использоваться числовые или символические поля. Каждое уникальное значение в этом поле должно обозначать конкретный объект анализа. Например, в прикладной программе Корзина покупок каждый ID может представлять одного покупателя. В прикладной программе Анализ Web-журнала каждый ID может представлять отдельный компьютер (по IP-адресу) или одного пользователя (по регистрационным данным).
- **Содержимое.** Задайте поле или поля содержимого для модели. Эти поля содержат нужные элементы при моделировании связывания. Можно задать одно номинальное поле, в котором данные представлены в транзакционном формате.

Использование табличного формата. Выключите переключатель **Использовать транзакционный формат**, если у исходных данных табличный формат.

- **Поля ввода.** Выберите одно или несколько входных полей. Это аналогично заданию для поля роли *Входное* на узле Тип.
- **Подмножества.** Это поле позволяет позволяет задать поле, с помощью которого данные будут разделяться на отдельные выборки для этапов обучения, испытания и проверки при построении модели. Сгенерировав модель при помощи одной выборки и испытав ее при помощи другой, можно получить надежный показатель качества обобщения модели на более крупные наборы данных, подобные текущим. Если при помощи узлов Тип или Раздел были определено несколько полей разделов, на вкладке Поля на каждом узле моделирования, где используется разделение, должно быть выбрано одно поле раздела. (Если представлен только один раздел, он будет использоваться автоматически при всяком включении разделения.) Кроме того, учтите, что для применения выбранного раздела в анализе на вкладке Параметры модели для узла должно быть также включено разделение. (Выключение этой опции делает возможным отключение разделения без изменения значений параметров полей.)

Макет таблицы для нетранзакционных данных. Для табличных данных можно выбрать стандартный макет (по умолчанию) или макет с ограниченной длиной элемента.

В макете по умолчанию количество столбцов определяется общим числом связанных элементов.

Таблица 6. Макет таблицы по умолчанию.

ID группы	Текущий счет	Сберегательный счет	Кредитная карта	Кредит	Депозитный счет
Смит	Д	Д	Д	-	-
Джексон	Д	-	Д	Д	Д

Таблица 6. Макет таблицы по умолчанию (продолжение).

ID группы	Текущий счет	Сберегательный счет	Кредитная карта	Кредит	Депозитный счет
Дуглас	Д	-	-	-	Д

В макете с ограниченной длиной элемента количество столбцов определяется наибольшим количеством связанных элементов в любой из строк.

Таблица 7. Макет таблицы с ограниченной длиной элемента.

ID группы	Элемент1	Элемент2	Элемент3	Элемент4
Смит	текущий счет	сберегательный счет	кредитная карта	-
Джексон	текущий счет	кредитная карта	кредит	депозитный счет
Дуглас	текущий счет	депозитный счет	-	-

Опции моделей связывания ISW

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Минимальная поддержка правила (%). Уровень минимальной поддержки для правил связывания или последовательности. В модель включаются только правила, достигшие по крайней мере этого уровня поддержки. Это значение вычисляется как A/B*100, где A - это количество групп, содержащих все элементы в правиле, а B - общее количество рассматриваемых групп. Если вы хотите сосредоточиться на более общих связях или последовательностях, увеличьте значение этого параметра.

Минимальная достоверность правила (%). Минимальный уровень достоверности для правил связывания или последовательности. В модель включаются только правила, достигшие по крайней мере этого уровня достоверности. Это значение вычисляется как *m*/*n**100, где *m* - это количество групп, содержащих присоединяемый заголовок правила (консеквент) и тело правила (антецедент), а *n* - количество групп, содержащих тело правила. Если у вас получается слишком много связываний или последовательностей (или они не интересны для модели), попробуйте увеличить значение этого параметра. Если связываний или последовательностей получается слишком мало, уменьшите значение этого параметра.

Максимальный размер правила. Максимальное количество разрешенных в правиле элементов, включая консеквент. Если нужные связывания или последовательности относительно короткие, можно уменьшить значение этого параметра для ускорения построения набора.

Примечание: Оцениваются только узлы с транзакционным форматом ввода; настоящие табличные форматы (табличные данные) остаются неуточненными.

Дополнительные опции связывания ISW

На вкладке Дополнительно узла Связывание можно указать, какие правила связывания будут включены в результаты, а какие - исключены. Если вы решили включить некоторые заданные элементы, правила, содержащие по крайней мере один из этих элементов, будут включены в модель. Если вы решили исключить некоторые заданные элементы, правила, содержащие любые из этих элементов, будут исключены из результатов.

Если выбрана опция **Использовать ограничения элементов**, все элементы, которые вы добавили в список ограничений, будут включены в результаты или исключены из них в зависимости от того, как задана опция **Тип ограничения**

Тип ограничения. Выберите, хотите ли вы включить в результаты правила связывания, относящиеся к заданным элементам, или исключить их из результатов.

Редактировать ограничения. Чтобы добавить элемент в список элементов ограничения, выберите его в списке Элементы и нажмите кнопку со стрелкой вправо.

Опции таксономии ISW

Алгоритмы связывания и последовательности ISW могут использовать **таксономии**. Таксономии отображают индивидуальные значения на концепции более высокого уровня. Например, ручки и карандаши могут быть отображены на категорию канцелярских товаров.

На вкладке Таксономия можно определить карты категорий для выражения таксономий через данные. Например, таксономия может создать две категории (`Товары первой необходимости` и `Предметы роскоши`), а затем назначать базовые элементы этим категориям. Например, `вино` назначается в категорию `Предметы роскоши`, а `хлеб` - категории `Товары первой необходимости`. У таксономии структура родительских и дочерних элементов, что показано в следующей таблице.

Таблица 8. Пример структуры таксономии

Дочерний элемент	Родительский элемент
вино	Предметы роскоши
хлеб	Товары первой необходимости

При наличии таксономии можно построить модель связывания или последовательности, включающую в себя правила с категориями, как и с базовыми элементами.

Примечание: Для активации опций на этой вкладке исходные данные должны быть в транзакционном формате, на вкладке **Поля** нужно выбрать опцию **Использовать транзакционный формат**, а затем на этой вкладке выбрать опцию **Использовать таксономию**.

Имя таблицы. Эта опция задает имя таблицы DB2 для хранения подробностей таксономии.

Дочерний столбец. Эта опция задает имя дочернего столбца в таблице таксономии. Дочерний столбец содержит названия элементов или категорий.

Родительский столбец. Эта опция задает имя родительского столбца в таблице таксономии. Родительский столбец содержит имена категорий.

Загрузить подробности в таблицу. Выберите эту опцию, если информация таксономии, хранимая в IBM SPSS Modeler, должна быть закачана в таблицу таксономии во время построения модели. Обратите внимание на то, что если таблица таксономии уже существует, она отбрасывается. Информация таксономии хранится на узле построения модели; ее можно изменять с помощью кнопок Изменить категории и Изменить таксономию.

Редактор категорий

Диалоговое окно Изменить категории позволяет добавлять категории к отсортированному списку и удалять их оттуда.

Чтобы добавить категорию, введите ее имя в поле **Новая категория** и нажмите кнопку со стрелкой, чтобы переместить категорию в список **Категории**.

Для удаления категории выберите ее в списке **Категории** и нажмите рядом с ней кнопку Удалить.

Редактор таксономии

Диалоговое окно Изменить таксономию позволяет объединить для построения таксономии набор основных элементов, определенных в данных, и набор категорий. Чтобы добавить записи в таксономию, выберите один или несколько элементов или категорий из списка слева и одну или несколько категорий из списка справа, а затем нажмите кнопку со стрелкой. Обратите внимание на то, что если какие-либо добавления в таксономию приводят к конфликту (например, задано `cat1 -> cat2` и противоположное `cat2 -> cat1`), эти добавления не производятся.

Последовательность ISW

Узел Последовательность обнаруживает шаблоны в последовательных данных или данных с временной ориентацией в формате `хлеб -> сыр`. Элементы последовательности - это **наборы позиций**, составляющие разовую транзакцию. Например, если человек заходит в магазин и покупает хлеб и молоко, а через несколько дней возвращается и покупает сыр, его покупательская активность может быть представлена двумя наборами товаров. Первый содержит хлеб и молоко, а второй - сыр. **Последовательность** - это список наборов товаров с тенденцией происходить в предсказуемом порядке. Узел Последовательность обнаруживает часто встречающиеся последовательности и создает узел сгенерированной модели, с помощью которого можно делать предсказания.

Функцию исследования данных Правила последовательности можно использовать в различных бизнес-областях. Например, в розничной продаже можно найти типичные последовательности покупок. Эти последовательности показывают различные сочетания покупателей, товаров и времени покупки. С помощью этой информации можно определить потенциальных покупателей конкретного товара, которые еще его не купили. Более того, можно предложить товары потенциальным покупателям в нужное время.

Последовательность - это упорядоченный набор наборов элементов. Последовательности содержат следующие уровни группировки:

- Произошедшие одновременно события образуют одну транзакцию или набор элементов.
- Каждый элемент или набор элементов принадлежит группе транзакций. Например, приобретенный предмет принадлежит покупателю, переход на конкретную страницу принадлежит Web-серверу, а запчасть принадлежит произведенному автомобилю. Несколько наборов элементов, существующих для разного времени и принадлежащие одной группе транзакций, образуют последовательность.

Опции модели последовательности ISW

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Минимальная поддержка правила (%). Уровень минимальной поддержки для правил связывания или последовательности. В модель включаются только правила, достигшие по крайней мере этого уровня поддержки. Это значение вычисляется как A/B*100, где A - это количество групп, содержащих все элементы в правиле, а B - общее количество рассматриваемых групп. Если вы хотите сосредоточиться на более общих связях или последовательностях, увеличьте значение этого параметра.

Минимальная достоверность правила (%). Минимальный уровень достоверности для правил связывания или последовательности. В модель включаются только правила, достигшие по крайней мере этого уровня достоверности. Это значение вычисляется как m/n*100, где m - это количество групп, содержащих присоединяемый заголовок правила (консеквент) и тело правила (антецедент), а n - количество групп, содержащих тело правила. Если у вас получается слишком много связываний или последовательностей (или они не интересны для модели), попробуйте увеличить значение этого параметра. Если связываний или последовательностей получается слишком мало, уменьшите значение этого параметра.

Максимальный размер правила. Максимальное количество разрешенных в правиле элементов, включая консеквент. Если нужные связывания или последовательности относительно короткие, можно уменьшить значение этого параметра для ускорения построения набора.

Примечание: Оцениваются только узлы с транзакционным форматом ввода; настоящие табличные форматы (табличные данные) остаются неуточненными.

Дополнительные опции последовательности ISW

Вы можете задать, какие правила последовательности должны быть включены в результаты или исключены из них. Если вы решили включить некоторые заданные элементы, правила, содержащие по крайней мере один из этих элементов, будут включены в модель. Если вы решили исключить некоторые заданные элементы, правила, содержащие любые из этих элементов, будут исключены из результатов.

Если выбрана опция **Использовать ограничения элементов**, все элементы, которые вы добавили в список ограничений, будут включены в результаты или исключены из них в зависимости от вашего параметра для опции **Тип ограничения**

Тип ограничения. Выберите, хотите ли вы включить в результаты эти правила связывания, относящиеся к заданным элементам, или исключить их из результатов.

Редактировать ограничения. Чтобы добавить элемент в список элементов ограничения, выберите его в списке Элементы и нажмите кнопку со стрелкой вправо.

Регрессия ISW

Узел регрессии ISW поддерживает следующие алгоритмы регрессии:

- Преобразование (по умолчанию)
- Линейный
- Полиномиальный
- RBF

Трансформационной регрессии

Алгоритм трансформационной регрессии ISW строит модели, представляющие из себя деревья решений с уравнениями регрессии на листьях деревьев. Обратите внимание на то, что визуализатор IBM не выводит структуру этих моделей.

Браузер IBM SPSS Modeler показывает параметры и аннотации. Однако просмотреть структуру модели нельзя. Существует относительно немного конфигурируемых пользователем параметров построения.

Линейная регрессия

Алгоритм линейной регрессии ISW предполагает линейную взаимосвязь между значениями в объяснительных полях и в поле назначения. Это приводит к моделям, представляющим уравнения. Предполагается, что предсказанное значение будет отличаться от наблюдаемого, так как уравнение регрессии - это аппроксимация значений в поле назначения. Их разность называется остатком.

При моделировании в IBM InfoSphere Warehouse Data Mining распознаются поля, у которых нет объясняющих значений. Чтобы определить, есть ли у поля объясняющее значение, алгоритм линейной регрессии выполняет статистические тесты в дополнение к выбору независимой переменной. Если известны поля, у которых нет объясняющих значений, можно автоматически выбрать подмножество объясняющих полей для сокращения времени работы.

Алгоритм линейной регрессии предоставляет следующие способы для автоматического выбора подмножеств объяснительных полей:

Пошаговая регрессия. Для пошаговой регрессии необходимо задать минимальный уровень значимости. В алгоритме линейной регрессии используются только поля с уровнем значимости выше заданного значения.

Регрессия R-квадрат. При способе R-квадрат регрессия определяет оптимальную модель, оптимизируя меру качества модели. Используется одна из следующих мер качества:

- Квадрат коэффициента корреляции Пирсона
- Скорректированный квадрат коэффициента корреляции Пирсона.

По умолчанию для оптимизации качества модели алгоритм линейной регрессии автоматически выбирает объяснительные поля, используя скорректированный коэффициент корреляции Пирсона в квадрате.

Полиномиальная регрессия

Алгоритм полиномиальной регрессии ISW предполагает наличие полиномиальной взаимосвязи. Модель полиномиальной регрессии - это уравнение, состоящее из следующих частей:

- Максимальная степень полинома в регрессии
- Приближение для значений в полях назначения
- Объяснительные поля.

Регрессия RBF

Алгоритм регрессии ISW RBF предполагает наличие взаимосвязи между значениями в объяснительных полях и в полях назначения. Эту взаимосвязь можно выразить линейной комбинацией гауссовых функций. Гауссовы функции - это подмножество радиальных базовых функций (Radial Basis Function, RBF).

Опции модели регрессии ISW

На вкладке Модель узла Регрессия ISW можно задать тип используемого алгоритма регрессии, а также:

- Использовать ли многораздельные данные
- Выполнять ли проход тестирования
- Предел для значения R^2
- Предел времени выполнения

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Метод регрессии. Выберите тип регрессии, которую вы хотите использовать. Дополнительную информацию смотрите в разделе “Регрессия ISW” на стр. 67.

Выполнить проход тестирования. Можно выбрать выполнение прохода тестирования. Тогда проход тестирования IBM InfoSphere Warehouse Data Mining выполняется после построения модели для обучающего раздела. При этом выполняется проход по проверочному разделу для установления информации о качестве модели, подъема диаграмм и так далее.

Ограничить R-квадрат. Эта опция задает максимальную допустимую систематическую ошибку (коэффициент корреляции Пирсона в квадрате, R^2). Этот коэффициент измеряет корреляцию между ошибкой предсказания для данных проверки и фактическими целевыми значениями. Это значение должно лежать от 0 (нет корреляции) до 1 (идеальная положительная или отрицательная корреляция). Определенное здесь значение задает верхний предел для приемлемой систематической ошибки модели.

Ограничить время выполнения. Задать желаемое максимальное время выполнения в минутах.

Дополнительные опции регрессии ISW

На вкладке Дополнительно узла Регрессия можно задать несколько расширенных опций для линейной или полиномиальной регрессии, а также для регрессии RBF.

Дополнительные опции для линейной или полиномиальной регрессии

Ограничить степень полиномиальности. Задает максимальную степень полиномиальной регрессии. Если для максимальной степени полиномиальной регрессии задается значение **1**, алгоритм полиномиальной регрессии идентичен алгоритму линейной регрессии. Если задать большое значение для максимальной степени полиномиальной регрессии, алгоритм этой регрессии стремится к сверх-подгонке. Это означает, что полученная модель точно аппроксимирует учебные данные, но оказывается непригодной при применении к данным, не использованным для обучения.

Использовать свободный член. При включении кривая регрессии проходит через начало координат. Это означает, что в модели нет постоянного слагаемого.

Использовать автоматический выбор функций. При включении алгоритм пытается определить оптимальное подмножество возможных предикторов, если вы не задаете минимальный уровень значимости.

Использовать минимальный уровень значимости. Когда задан минимальный уровень значимости, для определения подмножества возможных предикторов используется пошаговая регрессия. В вычисление модели регрессии дают вклад только независимые поля с уровнем значимости выше заданного.

Параметры полей. Чтобы задать опции для индивидуальных входных полей, щелкните по соответствующей строке в столбце Параметры таблицы Параметры полей и выберите **<Задать параметры>**. Дополнительную информацию смотрите в разделе “Задание параметров полей для регрессии”.

Дополнительные опции для регрессии RBF

Использовать размер выходной выборки. Определяет выборку 1-в-N для проверки и тестирования модели.

Использовать размер входной выборки. Определяет выборку 1-в-N для обучения.

Использовать максимальное число центров. Максимальное количество центров, строящихся при каждом проходе. Так как при проходе количество центров может увеличиваться вдвое по сравнению с начальным значением, фактическое количество центров может оказаться больше указанного вами числа.

Использовать минимальный размер региона. Минимальное число записей, назначенных региону.

Использовать максимальное число проходов по данным. Максимальное число проходов алгоритма по входным данным. Если это значение задано, оно должно быть не меньше минимального числа проходов.

Использовать минимальное число проходов по данным. Минимальное число проходов алгоритма по входным данным. Задавайте большое значение только в том случае, если у вас достаточно обучающих данных и есть уверенность, что существует хорошая модель.

Задание параметров полей для регрессии

В диалоговом окне Изменить параметры регрессии можно задать диапазон значений в отдельном входном поле для линейной или полиномиальной регрессии.

Значение MIN. Минимальное допустимое значение для этого входного поля.

Значение MAX. Максимальное допустимое значение для этого входного поля.

Кластеризация ISW

Функция исследования данных Кластеризация ищет входные данные для характеристик, которые обычно встречаются чаще всего. Она группирует входные данные в кластеры. У элементов каждого кластера аналогичные свойства. Не существует заранее принятых предположений, какие именно структуры присутствуют в данных. Кластеризация обнаруживается в процессе анализа.

Узел Кластеризация ISW предоставляет выбор из следующих способов кластеризации:

- Демографический
- Коонена
- Улучшенный BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

Способ алгоритма **демографической кластеризации** основан на распределениях. Кластеризация на основе распределений предоставляет быструю и естественную кластеризацию очень больших баз данных. Количество кластеров выбирается автоматически (вы должны задать максимальное допустимое количество кластеров). Есть много конфигурируемых пользователем параметров.

Алгоритм **кластеризации Коонена** основан на центральном положении кластера. Карта особенностей Коонена пытается расположить центры кластеров так, чтобы минимизировать общее расстояние между записями и центрами кластеров. Возможность разделения кластеров во внимание не принимается. Центральные векторы распределяются в таблице с определенным числом столбцов и строк. Эти векторы взаимосвязаны, так что настраивается не только победивший вектор, который ближе всего к обучающей записи, но и все векторы по соседству с ним. Однако чем дальше другие центры, тем меньше они настраиваются.

Метод улучшенного алгоритма **кластеризации BIRCH** основан на распределениях и пытаются минимизировать общее расстояние между записями и их кластерами. По умолчанию для определения расстояния между записью и кластером используется расстояние логарифмического правдоподобия; можно выбрать также евклидово расстояние, если все активные поля - числовые. Алгоритм BIRCH выполняется за два независимых шага; сначала он распределяет входные записи по дереву возможностей кластеризации, чтобы аналогичные записи были частью одного узла дерева, а затем кластеризует ветви этого дерева в памяти, чтобы сгенерировать окончательный результат кластеризации.

Опции моделей кластеризации ISW

На вкладке Модель узла Кластеризация можно задать способ, используемый для создания кластеров, и сопутствующие опции.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Метод кластеризации. Выберите способ, который вы хотите использовать для создания кластеров: **Демографический**, **Коонена** или **Улучшенный BIRCH**. Дополнительную информацию смотрите в разделе “Кластеризация ISW”.

Ограничить число кластеров. Ограничение числа кластеров экономит время выполнения, не допуская создания большого числа маленьких кластеров.

Число строк/Число столбцов. (только для метода Коонена) Задает количество строк и столбцов для карты особенностей Коонена. (Доступно только в том случае, если выбрана опция **Ограничить количество проходов Коонена** и отменена опция **Ограничить количество кластеров**).

Ограничить количество проходов Коонена. (только для метода Коонена) Задает количество проходов алгоритма кластеризации по данным во время обучающих запусков. При каждом проходе корректируются векторы центрального положения, чтобы минимизировать общее расстояние между центрами кластеров и записями. Уменьшается также степень изменения векторов при каждой итерации. При первом проходе изменения существенны. На последнем проходе степень корректировки центров очень незначительна. Выполняются только тонкие настройки.

Мера расстояния. (только улучшенный метод BIRCH) Выберите меру расстояния от записи до кластера, используемую алгоритмом BIRCH. Можно выбрать или расстояние логарифмического правдоподобия (по умолчанию), или евклидово расстояние. *Примечание*: Если все активные поля числовые, можно выбрать только евклидово расстояние.

Максимальное число конечных узлов. (только для улучшенного метода BIRCH) Максимальное количество конечных узлов, которые вы хотите задать на дереве возможности кластеризации. Дерево возможности кластеризации - это результат первого шага улучшенного алгоритма BIRCH, когда записи данных располагаются в дереве таким образом, чтобы схожие записи принадлежат одному конечному узлу. При росте количества конечных узлов возрастает и время выполнения алгоритма. Значение по умолчанию - 1000.

Проходы Birch. (только для улучшенного метода BIRCH) Количество проходов по данным, выполняемых алгоритмом для уточнения результатов кластеризации. Количество проходов влияет на время обработки обучающих запусков (каждый проход требует полного просмотра данных) и на качество модели. Небольшие значения сократят время обработки, но они же могут привести к понижению качества моделей. Большие значения приводя к возрастанию времени обработки и обычно улучшают качество моделей. В среднем к хорошим результатам приводят три или более проходов. Значение по умолчанию - 3.

Дополнительные опции кластеризации ISW

На вкладке Эксперт узла Кластеризация можно задать расширенные опции, такие как пороги подобия, ограничения времени выполнения и веса полей.

Ограничить время выполнения. Включите этот переключатель для включения опций, позволяющих управлять временем на создание модели. Можно задать время в минутах и/или минимальную процентную долю обработки обучающих данных. В дополнение к методу BIRCH можно задать максимальное количество конечных узлов для создания на дереве CF.

Задать порог подобия. (только для демографической кластеризации) Нижний предел для подобия двух записей данных, принадлежащих одному кластеру. Например, значение 0,25 означает, что записи со значениями, подобными на 25%, могут быть назначены в один кластер. Значение 1,0 означает, что для отнесения к одному кластеру записи должны быть идентичными.

Параметры полей. Чтобы задать опции для отдельных входных полей, щелкните по соответствующей строке в столбце Параметры таблицы Параметры полей и выберите **<Задать параметры>**.

Задание параметров полей для кластеризации

В диалоговом окне Изменить параметры кластера можно задать опции для отдельных входных полей.

Вес поля. В процессе построения модели назначает полю большее или меньшее значение веса. Например, если вы полагаете, что данное поле относительно менее важно для модели, чем другие поля, уменьшите его вес по отношению к другим полям.

Вес значения. Назначает больший или меньший вес конкретным значениям в данном поле. Некоторые значения полей могут быть более распространенными, чем другие значения. Совпадение редких значений в поле может быть более значимым, чем совпадение обычных значений. Вы можете выбрать один из следующих способов для взвешивания значений в данном поле (в любом случае у редких значений будет больший вес, чем у обычных):

- **Логарифмический.** Назначает вес каждому значению по логарифму вероятности его присутствия во входных данных.
- **Вероятностный.** Назначает вес каждому значению по вероятности его присутствия во входных данных.

Для любого из способов вы можете выбрать также опцию **с компенсацией**, чтобы компенсировать значение взвешивания, применяемое к каждому полю. Если проводить компенсацию с учетом взвешивания значений, общая важность взвешенного поля равна тому же значению, что и без взвешивания. Это не зависит от количества возможных значений. Компенсированное взвешивание влияет только на относительную важность совпадений в наборе возможных значений.

Использовать показатель подобия. Включите этот переключатель, если вы хотите использовать показатель подобия для управления вычислением меры подобия для поля. Показатель подобия задается в абсолютных числах. Эта спецификация рассматривается только для активных числовых полей. Если вы не задаете показатель подобия, используется значение по умолчанию (половина среднеквадратичного отклонения). Чтобы получить большее количество кластеров, уменьшите среднее подобие между парами кластеров, задав меньшее значение показателей подобия для числовых полей.

Обработка выбросов. Выбросы - это значения полей, лежащие вне диапазона значений, заданных для поля, как он определен параметрами **значение MIN** и **значение MAX**. Можно выбрать, как обрабатывать значения выбросов для этого поля.

- Значение по умолчанию **нет** означает, что никакие специальные действия для выбросов не предпринимаются.
- Если выбрать значение **заменить на MIN или MAX**, значение поля, меньшее **значения MIN** или большее **значения MAX**, заменяется на значения MIN или MAX соответственно. В этом случае вы можете задать значения MIN и MAX.
- Если выбрать **рассматривать как пропущенное**, выбросы рассматриваются как пропущенные значения и игнорируются. В этом случае вы можете задать значения MIN и MAX.

Наивный байесовский анализ ISW

Наивный критерий Байеса - это общеизвестный алгоритм для проблем классификации. Модель названа *наивной*, поскольку она рассматривает все предлагаемые переменные предсказания как независимые друг от друга. Наивный критерий Байеса - быстрый, масштабируемый алгоритм, вычисляющий условные вероятности для сочетаний атрибутов и атрибута назначения. На основе обучающих данных оценивается независимая вероятность. Эта вероятность передает правдоподобие каждого класса назначения с учетом вхождения каждой категории значений из каждой входной переменной.

Алгоритм наивной классификации Байеса ISW - это вероятностный классификатор. Он основан на моделях вероятности, которые предполагают абсолютную независимость.

Опции наивной модели Байеса ISW

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Выполнить проход тестирования. Можно выбрать возможность выполнения прохода тестирования. В таком случае проход тестирования IBM InfoSphere Warehouse Data Mining выполняется после построения модели для обучающего раздела. При этом выполняется проход по испытательному разделу для установления информации о качестве модели, подъема диаграмм и так далее.

Порог вероятности. Порог вероятности определяет вероятность для любых комбинаций предикторов и целевых значений, которые не видны в обучающих данных. Значение этой вероятности должно быть от 0 до 1. Значение по умолчанию - 0,001.

Логистическая регрессия ISW

Логистическая регрессия (другое название - номинальная регрессия) - это статистический метод для классификации записей на основе значений в полях ввода. Она аналогична линейной регрессии, но логистическая регрессия ISW использует флаговые (двоичные) поля назначения, а не числовые.

Опции модели логистической регрессии ISW

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Использовать разделенные данные. Если определено поле раздела, эта опция гарантирует, что для построения модели будут использоваться данные только из раздела обучения.

Выполнить проход тестирования. Вы можете выбрать выполнение прохода тестирования. Тогда проход тестирования IBM InfoSphere Warehouse Data Mining выполняется после построения модели для обучающего раздела. При этом выполняется проход по испытательному разделу для установления информации о качестве модели, подъема диаграмм и так далее.

Временные ряды ISW

Алгоритмы временных рядов ISW позволяют предсказать будущие события на основе известных событий в прошлом.

Аналогично общим методам регрессии алгоритмы временных рядов предсказывают численные значения. В отличие от общих методов регрессии, предсказания временных рядов нацелены на будущие значения упорядоченных рядов. Такие предсказания в общем случае называются прогнозом.

Алгоритмы временных рядов - это одномерные алгоритмы. Это означает, что независимая переменная - это столбец времени или столбец упорядочивания. Прогнозы основаны на известных значениях для прошлого. Они не основаны на других независимых столбцах.

Алгоритмы временных рядов отличаются от общих алгоритмов регрессии, так как они не только предсказывают будущие значения, но и включают в прогноз сезонные циклы.

Функция анализа данных для временных рядов предоставляет следующие алгоритмы предсказания будущих тенденций:

- ARIMA (Autoregressive Integrated Moving Average - авторегрессивное интегрированное скользящее среднее)
- Экспоненциальное сглаживание
- Декомпозиция сезонных тенденций

Алгоритм, дающий наилучший прогноз для ваших данных, зависит от различных предположений модели. Вы можете вычислить все прогнозы одновременно. Алгоритмы вычисляют подробный прогноз, включающий в себя сезонное поведение исходных временных рядов. Если у вас установлен клиент IBM InfoSphere Warehouse, для оценки и сравнения полученных кривых можно использовать визуализатор временных рядов.

Опции полей ISW Time Series

Время. Выберите входное поле, которое содержит временные ряды. Это должно быть полем с типом хранения Дата, Время, Отметка времени, действительное или целое число.

Использовать параметры узла типа. Эта опция указывает узлу, что следует использовать информацию о полях из узла Тип, расположенного выше. Это опция по умолчанию.

Использовать пользовательские параметры. Эта опция указывает узлу, что следует использовать информацию о полях, заданную здесь, а не ту, что задана на расположенных выше узлах Тип. После выбора этого варианта задайте приведенные ниже поля, как это потребуется.

Поля назначения. Выберите одно или несколько полей назначения. Это аналогично заданию для поля роли *Поле назначения* на узле Тип.

Опции модели ISW Time Series

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Алгоритмы прогнозирования. Выберите алгоритмы, используемые для моделирования. Вы можете выбрать один или несколько из следующих вариантов:

- АРПСС
- Экспоненциальное сглаживание
- Декомпозиция сезонных тенденций.

Конечное время прогнозирования. Укажите, надо ли рассчитывать конечное время прогноза автоматически или же оно задается вручную.

Значение поля времени. Если **Конечное время прогнозирования** задается вручную, введите конечное время для прогноза. Значение, которое можно ввести, зависит от типа поля времени; например, если тип - целое число, представляющее срок в часах, можно ввести 48, чтобы прекратить прогнозирование после обработки данных для 48 часов. Другой вариант - вас могут попросить ввести в качестве конечного времени дату или время.

Опции ISW Time Series Expert

Использовать при построении модели все записи. Это значение по умолчанию; все записи анализируются при построении модели.

Использовать при построении модели подмножество записей. Выберите эту опцию, если хотите создать модель только из части доступных данных. Например, это может понадобиться, если у вас есть большой массив повторяющихся данных.

Введите **Значение времени начала** и **Значение времени окончания** для задания используемых данных. Обратите внимание на то, что значения, которые можно ввести в эти поля, зависят от типа поля времени; например, это может быть срок в часах или днях или определенные дата или время.

Способ интерполяции для отсутствующих значений назначения. При обработке данных с одним или несколькими пропущенными значениями выберите метод для их вычисления. Вы можете выбрать один из следующих вариантов:

- Линейный
- Экспоненциальные сплайны
- Кубические сплайны

Вывод моделей ISW Time Series

Модели ISW Time Series - это выходные данные в форме неуточненной модели, содержащей информацию, которая извлечена из данных, однако не предназначена для непосредственного генерирования прогноза.

Рисунок 3. Значок неуточненной модели

Если у вас установлен клиент IBM InfoSphere Warehouse, можно использовать инструмент Визуализатор временных рядов для графического вывода ваших данных временных рядов.

Чтобы использовать инструмент Визуализатор временных рядов:

1. Убедитесь, что вы выполнили задачи по интеграции IBM SPSS Modeler с IBM InfoSphere Warehouse. Дополнительную информацию смотрите в разделе “Enabling Integration with IBM InfoSphere Warehouse” на стр. 53.
2. Дважды щелкните по неуточненной модели на палитре Модели.
3. На вкладке Сервер диалогового окна нажмите кнопку Вид для вывода визуализатора в браузере по умолчанию.

Слепки моделей исследования данных ISW

Создавать модели можно на узлах Дерево решений ISW, Связывание, Последовательность, Регрессия и Кластеризация, включенных в IBM SPSS Modeler.

Вкладка Сервер слепков моделей ISW

Вкладка Сервер предоставляет возможности для выполнения проверки согласованности и запуска инструмента IBM Visualizer.

IBM SPSS Modeler может выполнять проверку согласованности, сохраняя идентичную сгенерированную строку ключей модели и в модели IBM SPSS Modeler, и в модели ISW. Для выполнения проверки согласованности нажмите кнопку **Проверить** на вкладке Сервер. Дополнительную информацию смотрите в разделе “Управление моделями DB2” на стр. 58.

Инструмент визуализации - это единственный способ для просмотра моделей InfoSphere Warehouse Data Mining. Этот инструмент можно установить дополнительно с InfoSphere Warehouse Data Mining. Дополнительную информацию смотрите в разделе “Enabling Integration with IBM InfoSphere Warehouse” на стр. 53.

- Нажмите кнопку **Просмотр** для запуска инструмента визуализации. Что именно выводит этот инструмент, зависит от типа сгенерированного узла. Например, при запуске из слепка модели дерева решений ISW инструмент визуализации возвратит производную таблицу Предсказанные классы.
- Нажмите кнопку **Проверить результаты** (только для Дерево решений и Последовательность), чтобы запустить инструмент визуализации и просмотреть сведения о качестве сгенерированной модели в целом.

Вкладка Параметры слепков моделей ISW

В IBM SPSS Modeler обычно передается только одно предсказание и связанная вероятность или показатель доверия. Кроме этого, пользовательская опция показа вероятностей для каждого возможного вывода (как и в логистической регрессии) - это опция времени оценки, доступная на вкладке Параметры слепков модели.

Включить показатели доверия для всех классов. Для каждого возможного вывода в поле назначения добавляет столбец с уровнем показателя доверия.

Вкладка Сводка слепков моделей ISW

На вкладке Сводка слепка модели выводится информация о самой модели (*Анализ*), об используемых в ней полях (*Поля*), значениях параметров, используемых при построении модели, (*Параметры построения*) и об обучении модели (*Сводка по обучению*).

При первом просмотре узла результаты вкладки Сводка свернуты. Чтобы увидеть нужные вам результаты, разверните соответствующие им элементы при помощи элемента управления расширением слева от них или выведите все результаты, нажав кнопку **Развернуть все**. Чтобы скрыть результаты по завершении их просмотра, сверните при помощи элемента управления расширением отдельные результаты, которые вы хотите скрыть, или сверните все результаты, нажав кнопку **Свернуть все**.

Анализ. Выводится информация о конкретной модели. Если в вашем исполнении в указанный слепок модели вложен узел анализа, в этом разделе появится также информация о выполненном анализе.

Поля. Список полей, используемых в качестве полей назначения и входных полей при построении модели.

Параметры компоновки. Содержит информацию об используемых при построении модели параметрах.

Сводная информация по обучению. Выводится тип модели, поток, используемый для ее создания, создавший ее пользователь, отметка времени построения модели и время, затраченное на ее построение.

Примеры исследования данных ISW

IBM SPSS Modeler для Windows поставляется с несколькими демонстрационными потоками, иллюстрирующими процесс исследования баз данных. Эти потоки можно найти в папке установки IBM SPSS Modeler по следующему пути:

\Demos\Database_Modeling\IBM DB2 ISW

Примечание: К папке Demos можно перейти из группы программ IBM SPSS Modeler в меню Пуск Windows.

Как пример исследования базы данных можно последовательно совместно использовать следующие потоки:

- *1_upload_data.str* — используется для очистки и закачивания данных из плоского файла в DB2.
- *2_explore_data.str* — используется как пример исследования данных в IBM SPSS Modeler.
- *3_build_model.str* — используется для построения модели дерева решений ISW.
- *4_evaluate_model.str* — используется как пример оценки модели с помощью IBM SPSS Modeler.
- *5_deploy_model.str* — используется для внедрения модели для оценки в базе данных.

Набор данных, используемый в примерах потоков, относится к прикладным программам обработки кредитных карт и представляет задачу классификации с различными категориальными и непрерывными предикторами. Более подробную информацию об этом наборе данных смотрите в следующем файле в папке установки IBM SPSS Modeler по пути:

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

Этот набор данных доступен в репозитории UCI Machine Learning Repository по адресу http://archive.ics.uci.edu/ml/.

Пример потока: закачка данных

Первый пример потока, *1_upload_data.str*, используется для очистки и закачивания данных из плоского файла в DB2.

Узел Filler используется для обработки пропущенных значений и замены значений в пустых полях, прочитанных из текстового файла *crx.data*, на значения *NULL*.

Пример потока: изучение данных

Второй пример потока, *2_explore_data.str*, используется для демонстрации изучения данных в IBM SPSS Modeler.

Обычное действие при изучении данных - это присоединение к данным узла аудита данных. Узел аудита данных доступен на палитре узлов вывода.

Вывод с узла аудита можно использовать для получения общего обзора полей и распределения данных. Двойным щелчком по графику в окне Аудит данных можно вызвать более подробный график для более глубокого изучения данного поля.

Пример потока: построение модели

Третий пример потока, *3_build_model.str*, иллюстрирует построение модели в IBM SPSS Modeler. Можно присоединить узел моделирования базы данных к потоку и дважды щелкнуть по этому узлу, чтобы задать параметры построения.

Используя вкладки узла моделирования Модель и Эксперт, можно настроить максимальную глубину дерева и остановить дальнейшее расщепление узла от точки начального построения дерева решений, задав максимальную чистоту и минимальное число наблюдений для внутреннего узла. Дополнительную информацию смотрите в разделе “Дерево решений ISW” на стр. 61.

Пример потока: оценка модели

Четвертый пример потока, *4_evaluate_model.str*, иллюстрирует преимущества использования IBM SPSS Modeler для моделирования в базе данных. После выполнения модели ее можно добавить обратно в поток данных и оценить модель, используя несколько предложенных в IBM SPSS Modeler инструментов.

При первом открытии потока слепок модели (*field16*) не включается в этот поток. Откройте узел источников CREDIT и убедитесь, что вы задали источник данных. Далее, при условии, что у вас запущен поток *3_build_model.str* для создания слепка *field16* в палитре Модели, можно запустить отсоединенные узлы, нажав кнопку **Выполнить** на панели инструментов (кнопка с зеленым треугольником). При этом запускается сценарий, который копирует слепок *field16* в поток, соединяет его с существующими узлами и затем запускает терминальные узлы в потоке.

Можно присоединить узел Анализ (доступный на палитре Вывод), чтобы создать матрицу совпадений, показывающую структуру совпадений между каждым сгенерированным (предсказанным) полем и его полем назначения. Для просмотра результатов запустите узел Анализ.

Можно создать также диаграмму выигрыша, чтобы показать повышение точности, достигнутое в модели. Присоедините к сгенерированной модели узел Оценка, а затем запустите поток для просмотра результатов.

Пример потока: внедрение модели

Если вы уже удовлетворены точностью модели, можно внедрить ее для использования с внешними прикладными программами или для записи оценок обратно в базу данных. В этом примере потока *5_deploy_model.str* данные считываются из таблицы `CREDIT`. Когда запущен узел экспорта в базу данных *внедрить решение*, данные реально не оцениваются. Вместо этого поток создает файл опубликованного образа *credit_scorer.pim* и файл опубликованных параметров *credit_scorer.par*.

Как и в предыдущем примере, запускается сценарий, который копирует слепок *field16* в поток с палитры Модели, соединяет его с существующими узлами и затем запускает терминальные узлы в потоке. В этом случае вы должны сначала задать источник данных на узлах Источник базы данных и Экспорт.

Глава 6. Моделирование баз данных с использованием IBM Netezza Analytics

IBM SPSS ModelerandIBM Netezza Analytics

IBM SPSS Modeler поддерживает интеграцию с IBM Netezza Analytics, что дает возможность запускать расширенные функции аналитики на серверах IBM Netezza. Эти возможности доступны через графический интерфейс IBM SPSS Modeler и в среде разработки, ориентированной на использование рабочих потоков, так что запускать алгоритмы исследования данных можно непосредственно в среде IBM Netezza.

IBM SPSS Modeler поддерживает интеграцию со указанными ниже алгоритмами из IBM Netezza Analytics.

- деревья решений
- K-средние
- Байесовская сеть
- Наивный Байес
- KNN
- Разделительная кластеризация
- PCA
- дерево регрессии
- линейная регрессия
- временные ряды
- Обобщенная линейная

Дополнительную информацию об алгоритмах смотрите в *Руководстве разработчика IBM Netezza Analytics* и в *Справочном руководстве IBM Netezza Analytics*.

Требования для интеграции с IBM Netezza Analytics

Чтобы выполнять моделирование в базе данных с использованием IBM Netezza Analytics, необходимы приведенные ниже условия. Чтобы проверить выполнение этих условий, можете посоветоваться с администратором базы данных.

- IBM SPSS Modeler в локальном режиме или на установке сервер IBM SPSS Modeler в Windows или UNIX (кроме zLinux, для которой драйверы ODBC IBM Netezza недоступны).
- IBM Netezza Performance Server 6.0 или новее, на котором работает пакет IBM SPSS In-Database Analytics.
- Источник данных ODBC для соединения с базой данных IBM Netezza. Дополнительную информацию смотрите в разделе “Включение интеграции с IBM Netezza Analytics” на стр. 80.
- Должны быть разрешены генерирование и оптимизация SQL в IBM SPSS Modeler. Дополнительную информацию смотрите в разделе “Включение интеграции с IBM Netezza Analytics” на стр. 80.

Примечание: Моделирование баз данных и оптимизация SQL требуют, чтобы на компьютере IBM SPSS Modeler была включена возможность соединения с сервером IBM SPSS Modeler. При включенной возможности соединения можно обращаться к алгоритмам баз данных, выполнять обратный перенос SQL непосредственно с клиента сервером IBM SPSS Modeler и обращаться к серверу IBM SPSS Modeler. Проверьте текущее состояние лицензии, для чего в меню клиента сервером IBM SPSS Modeler выберите:

Справка > О программе > Дополнительные подробности

Если возможность соединения включена, на вкладке Состояние лицензии вы увидите опцию **Разрешение для сервера**.

Включение интеграции с IBM Netezza Analytics

Включение интеграции с IBM Netezza Analytics состоит из следующих действий.

- Конфигурирование IBM Netezza Analytics
- Создание источника ODBC
- Включение интеграции в IBM SPSS Modeler
- Включение генерирования и оптимизации SQL в IBM SPSS Modeler

Эти действия описаны в следующих разделах.

Конфигурирование IBM Netezza Analytics

Чтобы установить и сконфигурировать IBM Netezza Analytics, посмотрите подробности в документации по IBM Netezza Analytics, особенно *Руководство по установке IBM Netezza Analytics*. Раздел *Задание разрешений базы данных* в этом руководстве содержит подробности о сценариях, при помощи которых потоки IBM SPSS Modeler получат доступ к базе данных для записи.

Примечание: Если вы будете использовать узлы, пользующиеся матричными вычислениями (Метод главных компонент Netezza (PCA) и Линейную регрессию Netezza), нужно инициализировать матричный механизм Netezza, для чего запустить `CALL NZM..INITIALIZE()`; если этого не сделать, выполнение хранимых процедур завершится неудачно. Инициализацию нужно выполнить один раз для каждой базы данных.

Создание источника ODBC для IBM Netezza Analytics

Для поддержки соединения между базой данных IBM Netezza и IBM SPSS Modeler нужно создать имя источника данных ODBC (data source name, DSN).

Чтобы создать DSN, сначала нужно познакомиться с основами применения источников данных ODBC и драйверов и поддержкой баз данных в IBM SPSS Modeler.

Если вы работаете в распределенном режиме на сервер IBM SPSS Modeler, создайте DSN на компьютере сервера. Если вы работаете в локальном режиме (на клиенте), создайте DSN на компьютере клиента.

Клиенты Windows

1. С компакт-диска *Netezza Client* запустите файл *nzodbcsetup.exe*, чтобы запустить программу установки. Следуйте указаниям на экране, чтобы установить драйвер. Полные указания смотрите в *Руководстве по установке и конфигурированию IBM Netezza ODBC, JDBC и OLE DB*.
   a. Создайте DSN.

      Примечание: Последовательность меню зависит от используемой версии Windows.

      - **Windows XP.** В меню Пуск выберите **Панель управления**. Щелкните дважды по значку **Администрирование**, а затем - по значку **Источники данных (ODBC)**.
      - **Windows Vista.** В меню Пуск выберите **Панель управления**, затем выберите **Система**. Щелкните дважды по значку **Администрирование** и выберите **Источники данных (ODBC)**, затем выберите **Открыть**.
      - **Windows 7.** В меню Пуск выберите **Панель управления**, затем **Система и безопасность**, затем **Администрирование**. Выберите **Источники данных (ODBC)**, затем выберите **Открыть**.

   b. Щелкните по вкладке **Системное DSN**, а затем нажмите кнопку **Добавить**.
2. Выберите **NetezzaSQL** в списке и нажмите кнопку **Готово**.
3. На вкладке **Опции DSN** экрана Установка драйвера ODBC Netezza введите имя источника данных по своему выбору, имя хоста или IP-адрес сервера IBM Netezza, номер порта для соединения, базу данных используемого экземпляра IBM Netezza и аутентификационные данные для соединения с базой данных - ваше имя пользователя и пароль. Нажмите кнопку **Справка**, чтобы вывести объяснения для полей.
4. Нажмите кнопку **Проверить соединение** и убедитесь, что удается соединиться с базой данных.

5. Дождавшись успешного соединения, нажимайте кнопку **OK**, пока на выйдете из экрана Администратор источников данных ODBC.

Серверы Windows

Порядок действий для сервера Windows такой же, как для клиента Windows XP.

Серверы UNIX или Linux

Описанный ниже порядок действий применим к серверам UNIX или Linux (кроме zLinux, для которой драйверы ODBC IBM Netezza недоступны).

1. Скопируйте с компакт-диска *Netezza Client* соответствующий файл *<платформа>cli.package.tar.gz* во временное положение на сервере.
2. Распакуйте архив при помощи команд `gunzip` и `untar`.
3. Добавьте права выполнения для извлеченного сценария *unpack*.
4. Запустите сценарий, ответив на выведенные приглашения.
5. Отредактируйте файл *modelersrv.sh*, включив в него приведенные ниже строки.

```
. /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
```

6. Найдите файл */usr/local/nz/lib64/odbc.ini* и скопируйте его содержимое в файл *odbc.ini*, установленный вместе с SDAP 6.1 (тот файл, который указан в переменной среды $ODBCINI).

   Примечание: Для 64-битных систем Linux параметр *Driver* некорректно ссылается на 32-битный драйвер. После того, как на предыдущем шаге вы скопировали содержимое файла *odbc.ini*, отредактируйте путь в этом параметре как нужно, например:

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

7. Отредактируйте параметры в определении DSN Netezza, указав нужную базу данных.
8. Перезапустите сервер IBM SPSS Modeler и проверьте, что на клиенте используются узлы исследования данных в базе данных Netezza.

Включение интеграции IBM Netezza Analytics в IBM SPSS Modeler

1. В основном меню IBM SPSS Modeler выберите

   Инструменты > Опции > Вспомогательные программы.
2. Щелкните по вкладке **IBM Netezza**.

Разрешить интеграцию исследования данных Netezza. Включает (если она еще не была выведена) палитру моделирования базы данных в нижней части окна IBM SPSS Modeler и добавляет узлы для алгоритмов исследования данных Netezza.

Соединение Netezza. Нажмите кнопку **Правка** и выберите строку соединения с Netezza, которую сконфигурировали ранее, создавая источник ODBC. Дополнительную информацию смотрите в разделе “Создание источника ODBC для IBM Netezza Analytics” на стр. 80.

Включение генерирования SQL и оптимизации

Если вы собираетесь работать с очень большими наборами данных, из соображений производительности надо разрешить опции генерирования SQL и оптимизации в IBM SPSS Modeler.

1. Выберите в меню IBM SPSS Modeler:

   Инструменты > Свойства потока > Опции
2. На панели навигации щелкните по опции **Оптимизация**.
3. Подтвердите включение опции**Генерировать SQL**. Этот параметр требуется для работы функций моделирования баз данных.

4. Выберите **Оптимизировать построение SQL** и **Оптимизировать другое выполнение** (не требуется строго, но настоятельно рекомендуется для оптимальной производительности).

Построение моделей с использованием IBM Netezza Analytics

У каждого из поддерживаемых алгоритмов есть соответствующий узел моделирования. Узлы моделирования IBM Netezza доступны на вкладке Моделирование базы данных на палитре узлов.

Данные

В зависимости от узла моделирования есть ряд типов данных, которые могут содержаться в полях в источнике данных. В IBM SPSS Modeler типы данных называются **типами измерений**. На вкладке Поля узла моделирования допустимые типы измерений для входных полей и полей назначения показываются значками.

Поле назначения. Поле назначения - это поле, значение которого вы пытаетесь предсказать. В тех случаях, где можно задать поле назначения, в качестве такого поля может быть выбрано только одно из полей данных источника.

Поле ID записи. Задает поле для уникальной идентификации каждого наблюдения. Например, это может быть поле ID, такое как *ID заказчика*. Если в данных источника нет поля ID, вы можете создать такое при помощи узла вычислений, выполнив приведенную ниже последовательность действий.

1. Выберите узел источника.
2. На вкладке Опции поля на палитре узлов дважды щелкните по узлу вычислений.
3. Откройте узел вычислений двойным щелчком по значку на холсте.
4. В поле **Вычисляемое поле** введите, например, `ID`.
5. В поле **Формула** введите `@INDEX` и нажмите кнопку **OK**.
6. Соедините узел вычислений с остальным потоком.

Обработка пустых значений

Если входные данные содержат пустые значения, использование некоторых узлов Netezza может приводить к сообщениям об ошибках или долгому выполнению потоков. Поэтому записи, содержащие пустые значения, рекомендуется удалить. Используйте следующий метод.

1. Присоедините к узлу источника узел выбора.
2. Задайте для опции **Режим** узла выбора значение **Отбрасывание**.
3. Введите следующее в поле **Условие**:
   `@NULL(`*`поле_1`*`) [or @NULL(`*`поле_2`*`)[... or @NULL(`*`поле_N`*`]])`
   Включите в условие все входные поля.
4. Соедините узел выбора с остальным потоком.

Выходная информация модели

Возможно, что от запуска к запуску результаты потока, содержащего узел моделирования Netezza, будут слегка различаться. Причина в непостоянстве порядка, в котором узел считывает данные источника, когда перед построением модели данные считываются во временные таблицы. Впрочем, различия, порожденные этим эффектом, пренебрежимо малы.

Общие комментарии

- В IBM SPSS Collaboration and Deployment Services невозможно создать конфигурации оценивания с использованием потоков, содержащих узлы моделирования базы данных IBM Netezza.
- Для моделей, созданных узлами Netezza, невозможен экспорт или импорт PMML.

Модели Netezza - опции полей

На вкладке Поля указывается, будут ли использоваться значения ролей полей, уже определенные на расположенных выше узлах, или назначение полей будет выполнено вручную.

Использовать заранее заданные роли. Эта опция использует значения ролей (назначения, предикторы и так далее) с расположенного выше узла Тип (или с вкладки Типы лежащего выше узла источника).

Настроить назначения полей. Выберите эту опцию, если хотите назначить объекты назначения, предикторы и другие роли вручную на этом экране.

Поля. При помощи кнопок со стрелками назначьте элементы из этого списка вручную для различных полей ролей в правой части экрана. Значки указывают для каждого поля роли допустимые уровни измерения.

Нажмите кнопку **Все**, чтобы выбрать все поля в списке, или кнопку для отдельного уровня измерения, чтобы выбрать все поля с этим уровнем измерения.

Цель. Выберите одно поле в качестве назначения для предсказания. На этом экране можно также посмотреть информацию об обобщенных линейных моделях в поле **Учебники**.

ID записи. Поле, которое будет использоваться как уникальный идентификатор записи.

Предикторы (входные поля) . Выберите одно или несколько полей в качестве входных для предсказания.

Модели Netezza - опции сервера

На вкладке Сервер задается база данных IBM Netezza, где будет храниться модель.

Подробности сервера баз данных Netezza. Здесь задаются подробности о соединении для базы данных, которую хотите использовать для модели.

- **Использовать восходящее соединение**. (По умолчанию) Использует подробности о соединении, заданные на восходящем узле, например, узле источника базы данных. *Примечание*: Эта опция работает, только если все восходящие узлы могут использовать обратный перенос SQL (SQL pushback). В этом случае перенос данных из базы данных не нужен, поскольку SQL полностью реализует все восходящие узлы.
- **Переместить данные на соединение**. Переносит данные в указанную здесь базу данных. Это действие позволяет моделированию работать, если данные находятся в другой базе данных IBM Netezza, в базе данных от другого поставщика или даже в плоском файле. Кроме того, данные переносятся обратно в указанную здесь базу данных, если они были извлечены из-за того, что узел не выполнил обратный перенос SQL. Нажмите кнопку **Правка**, чтобы найти и выбрать соединение. *Предостережение*: IBM Netezza Analytics обычно используется с очень большими наборами данных. Передача больших объемов данных между базами данных или из базы данных и обратно может занимать очень много времени, и по возможности его следует избегать.

Имя таблицы. Имя таблицы базы данных, в которой будет храниться модель. *Примечание*: Это должна быть новая таблица; вы не можете использовать для этой операции существующую таблицу.

Комментарии

- Соединение, используемое для моделирования, может и совпадать, и не совпадать с соединением, используемым в узле источника для потока. Например, у вас может быть поток, который обращается к данным одной базы данных IBM Netezza, скачивает данные в IBM SPSS Modeler для очистки или других действий и затем закачивает данные в другую базу данных IBM Netezza для целей моделирования. Однако обратите внимание на то, что такая конфигурация может неблагоприятно повлиять на производительность.

- Имя источника данных ODBC эффективно встраивается в каждом потоке IBM SPSS Modeler. Если поток, созданный на одном хосте, выполняется на другом хосте, имя источника данных на этих хостах должно быть одним и тем же. Можно также выбрать другой источник данных на вкладке Сервер каждого узла источника или моделирования.

Модели Netezza - опции моделей

На вкладке Опции модели можно выбрать, задавать ли имя для модели, или сгенерировать имя автоматически. Вы можете задать также значения по умолчанию для опций скоринга.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Сделать доступным для скоринга. Здесь можно задать значения по умолчанию для опций скоринга, которые появятся в диалоговом окне для слепка модели. Подробности об этих опциях смотрите в справочной теме для вкладки Параметры этого конкретного слепка.

Управление моделями Netezza

При построении модели IBM Netezza через IBM SPSS Modeler создается модель в IBM SPSS Modeler, а также создается или заменяется модель в базе данных Netezza. Модель IBM SPSS Modeler такого рода ссылается на содержимое модели базы данных, хранящейся на сервере баз данных. IBM SPSS Modeler может выполнять проверку соответствия, сохраняя идентичную сгенерированную строку ключа модели и в модели IBM SPSS Modeler, и в модели Netezza.

Имя модели для каждой модели Netezza выводится в столбце *Информация о модели* в диалоговом окне Список моделей базы данных. Имя для модели IBM SPSS Modeler выводится как Ключ модели на вкладке Сервер модели IBM SPSS Modeler (при помещении в поток).

Кнопку Проверить можно использовать для проверки, что ключи в модели IBM SPSS Modeler и в модели Netezza совпадают. Если модель с таким же именем в Netezza не найдена или ключи этих моделей не совпадают, это означает, что модель Netezza была удалена или повторно построена с момента построения модели в IBM SPSS Modeler.

Перечисление моделей базы данных

IBM SPSS Modeler содержит диалоговое окно для перечисления моделей, хранимых в IBM Netezza, и позволяет включать и отключать эти модели. Это диалоговое окно доступно из диалогового окна вспомогательных прикладных программ IBM и в диалоговых окнах построения, просмотра и применения для относящихся к узлам, связанным с исследованием данных IBM Netezza. Для каждой модели выводится такая информация:

- Имя модели (имя модели, которая используется для сортировки списка).
- Имя владельца.
- Алгоритм, использованный в модели.
- Текущее состояние модели, например, Завершена.
- Дата создания модели.

Деревья решений Netezza

Дерево решений - это иерархическая структура, представляющая модель классификации. С помощью модели дерева решений можно разрабатывать систему классификации для предсказания и классификации будущих наблюдений на основе набора данных обучения. Классификация принимает форму древообразной структуры, в которой ветви представляют собой точки разветвления в классификации. Эти разветвления рекурсивно делят данные на подгруппы, пока не будет достигнута точка остановки. Узлы дерева в точках остановки называют **листьями** (конечными узлами). Каждый лист назначает метку, называемую **меткой класса** участникам своей подгруппы, то есть класса.

Вывод модели возвращает форму текстового представления дерева. Каждая строка текста соответствует узлу или конечному узлу, а отступ текста отражает уровень дерева. Для узла выводится условие разбиения; для конечного узла - метка назначенного класса.

Веса экземпляров и веса классов

По умолчанию предполагается, что у всех входных записей и классов одинаковая относительная важность. Вы можете изменить это, назначив индивидуальные веса любым записям и/или классам. Это может оказаться полезным, например, если точки в данных обучения не распределены реалистически по категориям. Веса позволяют сместить модель, чтобы можно было скомпенсировать категории, хуже представленные в данных. С увеличением веса для значения назначения должен возрастать процент правильных предсказаний для данной категории.

На узле моделирования дерева решений можно задать два типа весов. **Веса экземпляров** назначают вес каждой строке входных данных. Для большинства наблюдений веса задаются как 1,0, а более высокие или низкие значения присваиваются более или менее важным категориям, как показано в следующей таблице.

Таблица 9. Пример веса экземпляров

ID записи	Назначение	Вес экземпляра
1	drugA	1,1
2	drugB	1,0
3	drugA	1,0
4	drugB	0,3

Веса классов назначаются каждой категории поля назначения, как показано в следующей таблице.

Таблица 10. Пример веса классов

Класс	Вес класса
drugA	1,0
drugB	1,5

Оба типа весов можно использовать одновременно, при этом они перемножаются и используются как веса экземпляров. Таким образом, если два предыдущих примера использовать совместно, алгоритм будет использовать веса экземпляров, показанные в следующей таблице.

Таблица 11. Пример вычисления веса экземпляра

ID записи	Вычисление	Вес экземпляра
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

Опции полей дерева решений Netezza

На вкладке Поля указывается, будут ли использоваться значения ролей полей, уже определенные на расположенных выше узлах, или назначение полей будет выполнено вручную.

Использовать заранее заданные роли. Эта опция применяет параметры ролей (назначений, предикторов и так далее) с восходящего узла Тип (или вкладки Типы восходящего узла источника).

Настроить назначения полей. Выберите эту опцию, если хотите назначить объекты назначения, предикторы и другие роли вручную на этом экране.

Поля. При помощи кнопок со стрелками назначьте элементы из этого списка вручную для различных полей ролей в правой части экрана. Значки указывают для каждого поля роли допустимые уровни измерения.

Нажмите кнопку **Все**, чтобы выбрать все поля в списке, или кнопку для отдельного уровня измерения, чтобы выбрать все поля с этим уровнем измерения.

Цель. Выберите одно поле в качестве назначения для предсказания.

ID записи. Поле, значение в котором будет использоваться как идентификатор уникальной записи. Значение в этом поле должно быть уникальным для каждой записи (например, номера ID покупателей).

Вес экземпляра. Здесь можно задать значение в поле, позволяющее использовать веса экземпляров (вес на строку входных данных) вместо или в дополнение к весам классов (вес на категорию в поле назначения), используемым по умолчанию. Заданное здесь поле должно содержать числовой вес для каждой строки входных данных. Дополнительную информацию смотрите в разделе “Веса экземпляров и веса классов” на стр. 85.

Предикторы (входные поля) . Выберите одно или несколько входных полей. Это аналогично заданию для поля роли *Входное* на узле Тип.

Опции построения дерева решений Netezza

Для построения дерева доступны следующие опции:

Мера роста. Эта опция управляет способом измерения роста деревьев. Если вы не хотите использовать значения по умолчанию, нажмите кнопку **Настроить** и измените значения.

- **Мера неоднородности.** Эта мера служит для оценки лучшего места расщепления дерева. Она измеряет неоднородность в подгруппе или в сегменте данных. Низкая мера неоднородности указывает на то, что ы группе у большинства элементов сходные значения критерия или поля назначения.

  Поддерживаемые измерения - это **Энтропия** и **Джини**. Это измерения основаны на вероятностях принадлежности к категории для ветви.
- **Максимальное количество уровней в дереве.** Максимальное количество уровней, до которого может расти дерево от корневого узла, то есть сколько раз рекурсивно разделена выборка. Значение по умолчанию - 62, это максимальное количество уровней дерева для целей моделирования.

  Примечание: Если программа просмотра в слепке модели выводит текстовое представление модели, может выводиться не более 12 уровней дерева.

Критерии расщепления. Эта опция управляет, когда остановить расщепление дерева. Если вы не хотите использовать значения по умолчанию, нажмите кнопку **Настроить** и измените значения.

- **Значение минимального улучшения для расщеплений**. Минимальное значение уменьшения неоднородности для создания нового расщепления на дереве. Цель построения дерева - создать подгруппы со сходными выходными значениями, чтобы минимизировать неоднородности на каждом узле. Если наилучшее расщепление ветви уменьшает неоднородность меньше, чем на заданное критерием расщепления значение, ветвь не расщепляется.
- **Минимальное количество экземпляров для расщепления.** Минимальное количество записей, которые можно расщепить. Когда число нерасщепленных записей становится меньше заданного значения, дальнейшие расщепления не производятся. Это поле можно использовать для предотвращения создания маленьких подгрупп в дереве.

Статистика. Этот параметр определяет, сколько статистик будет включено в модель. Выберите один из перечисленных ниже вариантов.

- **Все.** Включается вся статистика, связанная со столбцами и со значениями.

  Примечание: Этот параметр включает сбор максимального объема статистики, что может повлиять на производительность вашей системы. Если вы не хотите просматривать модель в графическом формате, задайте **Нет**.
- **Столбцы.** Включается статистика по столбцам.
- **Нет.** Включается только статистика, требуемая для скоринга модели.

Узел дерева решений Netezza - веса классов

Здесь можно назначить веса конкретным классам. По умолчанию всем классам присвоено значение 1, то есть у них одинаковый вес. Задавая различные численные веса для разных меток классов, вы указываете алгоритму соответствующим образом взвешивать обучающие наборы отдельных классов.

Чтобы изменить вес, дважды щелкните по нему в столбце **Вес** и внесите нужные вам изменения.

Значение. Набор меток классов, полученный из возможных значений поля назначения.

Вес. Вес, назначаемый конкретному классу. Назначение классу более высокого веса делает модель более чувствительной к этому классу относительно остальных классов.

Веса классов можно использовать в сочетании с весами экземпляров. Дополнительную информацию смотрите в разделе “Веса экземпляров и веса классов” на стр. 85.

Узел дерева решений Netezza - сокращение дерева

Опции сокращения можно использовать для определения критерия сокращения для дерева решений. Предназначение сокращения - снижение риска переобучения, для чего удаляются излишние подгруппы, которые не повышают ожидаемой точности на новых данных.

Мера сокращения. Мера сокращения по умолчанию (**Точность**) обеспечивает, чтобы оцененная точность модели после удаления листа дерева оставалась в приемлемых пределах. Используйте альтернативную меру **Взвешенная точность**, если, выполняя сокращение, вы хотите принять во внимание веса классов.

Данные для сокращения. Вы можете использовать некоторые или все обучающие данные для оценки ожидаемой точности по новым данных. Для этой цели можно также использовать отдельный набор данных сокращения из заданной таблицы.

- **Использовать все данные обучения**. Эта опция (задаваемая по умолчанию) использует все обучающие данные для оценки точности модели.
- **Использовать % данных обучения для усечения**. Эта опция используется для разбиения данных на два набора: один для обучения, а другой для сокращения (при помощи заданного здесь процента).

  Выберите **Воспроизвести результаты**, если хотите задать начальное значение для генератора псевдослучайных чисел, что гарантирует разделение данных одним и тем же способом при каждом запуске потока. Можно либо задать целое число в поле **Начальное значение, используемое для усечения**, либо выбрать действие **Сгенерировать**, создающее псевдослучайное целое число.
- **Использовать данные из существующей таблицы**. Укажите имя таблицы отдельного набора данных сокращения для оценки точности модели. Выполнение этого действия считают более надежным, чем использование данных обучения. Однако эта опция может привести к удалению из набора обучения большого поднабора данных и тем самым к снижению качества дерева решений.

K-средние Netezza

Узел K-средние реализует алгоритм *k*-средних, представляющий метод кластерного анализа. Этот узел можно использовать, чтобы кластеризовать набор данных в отдельные группы.

Это алгоритм кластеризации на основании расстояний, который основывается на показателе расстояния (функции) для измерения сходства между точками данных. Точки данных назначаются ближайшему кластеру в соответствии с используемым показателем расстояния.

Алгоритм работает, выполняя несколько итераций одного базового процесса, в котором каждый экземпляр обучения назначается ближайшему кластеру (с учетом заданной функции расстояния, применяемой к экземпляру и центру кластера). Затем все центры кластеров вычисляются повторно как средние векторы значений атрибутов экземпляров, назначенных конкретным кластерам.

Опции полей K-средних Netezza

На вкладке Поля указывается, будут ли использоваться значения ролей полей, уже определенные на расположенных выше узлах, или назначение полей будет выполнено вручную.

Использовать заранее заданные роли. Эта опция использует значения ролей (назначения, предикторы и так далее) с расположенного выше узла Тип (или с вкладки Типы лежащего выше узла источника).

Настроить назначения полей. Выберите эту опцию при желании задать назначения, предикторы и другие роли вручную на этом экране.

Поля. При помощи кнопок со стрелками назначьте элементы из этого списка вручную для различных полей ролей в правой части экрана. Значки обозначают допустимые уровни измерения для каждого поля роли.

Нажмите кнопку **Все**, чтобы выбрать все поля в списке, или нажмите кнопку отдельного уровня измерений, чтобы выбрать все поля с этим уровнем.

ID записи. Поле, которое будет использоваться как уникальный идентификатор записи.

Предикторы (входные поля). Выберите одно или несколько полей в качестве входных для предсказания.

Вкладка Опции построения K-средних Netezza

Задавая опции построения, можно настроить построение модели в соответствии с вашими потребностями.

Если вы хотите построить модель с опциями по умолчанию, нажмите кнопку **Запустить**.

Мера расстояния. Этот параметр определяет способ измерения расстояния между точками данных. Большие расстояния соответствую большим отличиям. Выберите один из перечисленных ниже вариантов.

- **Евклидова.** Евклидова мера расстояния - это длина прямой между двумя точками данных.
- **Нормализованная евклидова.** Нормализованная евклидова мера подобна евклидовой, но нормализована по среднеквадратичному отклонению. В отличие от евклидова расстояния, нормализованное евклидово расстояние не зависит от единицы измерения.
- **Махаланобиса.** Мера Махаланобиса - это обобщенное евклидово расстояние, учитывающее корреляции во входных данных. Как и нормализованное евклидово расстояние, расстояние Махаланобиса не зависит от единицы измерения.
- **Манхеттенская.** Манхеттенское расстояние между двумя точками вычисляется как сумма абсолютных разностей между их координатами.
- **Канберра.** Канберрское расстояние подобно манхеттенскому, но больше зависит от близости точек к началу координат.
- **Максимум.** Это расстояние между двумя точками данных вычисляется как наибольшее из различий по любой из их координат.

Число кластеров. Этот параметр определяет количество кластеров, которые будут созданы.

Максимальное количество итераций. Алгоритм выполняет несколько итераций одного процесса. Этот параметр определяет количество итераций, после которого обучение модели прекращается.

Статистика. Этот параметр определяет, сколько статистик будет включено в модель. Выберите один из перечисленных ниже вариантов.

- **Все.** Включается вся статистика, связанная со столбцами и со значениями.

  Примечание: Этот параметр включает сбор максимального объема статистики, что может повлиять на производительность вашей системы. Если вы не хотите просматривать модель в графическом формате, задайте **Нет**.
- **Столбцы.** Включается статистика по столбцам.
- **Нет.** Включается только статистика, требуемая для скоринга модели.

Воспроизвести результаты. Включите этот переключатель, если вы хотите задать начальное значение генератора псевдослучайных чисел для воспроизведения анализа. Можно задать целое число или же создать псевдослучайное целое число, нажав кнопку **Сгенерировать**.

Байесовская сеть Netezza

Байесовская сеть - это модель, выводящая переменные в наборе данных и вероятностные (условные) независимости между ними. Используя узел Байесовская сеть Netezza, можно построить вероятностную модель, комбинируя наблюдаемые и записанные сведения с общеизвестными и очевидными с точки зрения здравого смысла данными, чтобы установить правдоподобие событий с использованием внешне не связанных атрибутов.

Опции полей байесовской сети Netezza

На вкладке Поля указывается, будут ли использоваться значения ролей полей, уже определенные на расположенных выше узлах, или назначение полей будет выполнено вручную.

Для данного узла поле назначения требуется только для скоринга, поэтому оно не выводится на этой вкладке. Задать или изменить поле назначения можно на узле Тип, на вкладке Опции модели этого узла или на вкладке Параметры слепка модели. Дополнительную информацию смотрите в разделе “Слепок байесовской сети - вкладка Параметры” на стр. 107.

Использовать заранее заданные роли. Эта опция использует значения ролей (назначения, предикторы и так далее) с расположенного выше узла Тип (или с вкладки Типы лежащего выше узла источника).

Настроить назначения полей. Выберите эту опцию при желании задать назначения, предикторы и другие роли вручную на этом экране.

Поля. Используйте кнопки со стрелками для назначения вручную элементов из этого списка полям различных ролей справа на экране. Значки обозначают допустимые уровни измерения для каждого поля роли.

Нажмите кнопку **Все**, чтобы выбрать все поля в списке, или нажмите кнопку отдельного уровня измерений, чтобы выбрать все поля с этим уровнем.

Предикторы (входные поля) . Выберите одно или несколько полей как входные поля для предсказания.

Опции построения байесовской сети Netezza

На вкладке Параметры конструкции задаются все опции для построения модели. Можно, конечно, просто нажать кнопку **Выполнить**, чтобы построить модель со всеми опциями по умолчанию, но скорее всего вы захотите настроить конструкцию для своих собственных целей.

Базовый индекс. Числовой идентификатор, присваиваемый первому атрибуту (полю ввода) для простоты внешнего управления.

Объем выборки. Объем выборки, которую нужно использовать, если количество атрибутов настолько велико, что время обработки оказывается неприемлемо большим.

Выводить дополнительную информацию во время выполнения . Если этот переключатель включен (значение по умолчанию), дополнительная информация о ходе выполнения выводится в диалоговом окне сообщений.

Наивный байесовский анализ Netezza

Наивный критерий Байеса - это общеизвестный алгоритм для проблем классификации. Модель названа *наивной*, поскольку она рассматривает все предлагаемые переменные предсказания как независимые друг от друга. Наивный критерий Байеса - быстрый, масштабируемый алгоритм, вычисляющий условные вероятности для сочетаний атрибутов и атрибута назначения. На основе обучающих данных оценивается независимая вероятность. Эта вероятность передает правдоподобие каждого класса назначения с учетом вхождения каждой категории значений из каждой входной переменной.

KNN Netezza

Анализ ближайшего сходства представляет собой метод классификации наблюдений на основе сходства наблюдений. Этот метод машинного обучения был разработан в качестве способа распознавания структуры данных при неточном соответствии имеющих структур или наблюдений. Подобные наблюдения близки друг к другу, а непохожие наблюдения, наоборот, удалены друг от друга. Таким образом, дистанция между двумя наблюдениями является критерием их различия.

Близкие друг к другу наблюдения называются “соседи”. Когда представляется новое наблюдение, обозначенное знаком вопроса, вычисляется его расстояние от всех других наблюдений в модели. Определяется классификация наиболее похожих наблюдений (ближайшее сходство) и новое наблюдение помещается в категорию, в которой содержится наибольшее количество ближайшего сходства.

Вы можете указать количество анализируемых ближайших соседей; это значение обозначается *k*. На рисунках ниже показано, каким образом новое наблюдение будет классифицироваться с использованием двух различных значений *k*. Если $k = 5$, новое наблюдение помещается в категорию *1*, поскольку большинство ближайших соседей принадлежит категории *1*. Однако если $k = 9$, новое наблюдение помещается в категорию *0*, поскольку большинство ближайших соседей принадлежит категории *0* .

Анализ ближайшего сходства также может использоваться для вычисления значений для непрерывного целевого объекта. В этой ситуации среднее целевое значение ближайшего сходства используется для получения предсказанного значения для нового наблюдения.

Опции моделей KNN Netezza - Общие

На вкладке Опции модели - Общее можно выбрать, задавать ли имя для модели, или сгенерировать имя автоматически. Вы можете задать также опции, которые управляют тем, как вычисляется количество ближайших соседей, и задать опции для повышенной производительности и точности модели.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Соседи

Мера расстояния. Метод, используемый для измерения расстояния между точками данных; чем больше расстояния, тем больше различия. Опции:

- **Евклидово**. (по умолчанию) Расстояние между двумя точками вычисляется путем их соединения по прямой.

- **Манхеттенская**. Расстояние между двумя точками вычисляется как сумма абсолютных разностей между их координатами.
- **Канберра**. Аналогична манхеттенской мере расстояния, но более чувствительна к точкам данных, находящимся ближе к источнику.
- **Максимум**. Расстояние между двумя точками вычисляется как наибольшее из различий по любой из их координат.

Количество ближайших соседей (k). Количество ближайших соседей для конкретного наблюдения. Обратите внимание на то, что использование большего числа соседей необязательно приводит к более точной модели.

Выбор *k* управляет соотношением между предотвращением переобучения (оно может оказаться важным особенно для "зашумленных" данных) и разрешения (приводящего к различным предсказаниям для схожих экземпляров). Обычно значение *k* приходится настраивать для каждого набора данных с типичными значениями, ранг которых меняется от 1 до нескольких десятков.

Повысить производительность и точность

Нормализовать изменения перед вычислением расстояния. Эта опция, если она выбрана, стандартизирует измерения для непрерывных входных полей перед вычислением значений расстояний.

Использовать стержневые наборы для повышения производительности для больших наборов данных. Эта опция, если она выбрана, использует выборки стержневых наборов для ускорения вычисления в случае участия в нем больших наборов данных.

Опции моделей KNN Netezza - опции скоринга

На вкладке Опции моделей - опции скоринга можно задать значение по умолчанию для опции скоринга и присвоить относительные веса отдельным классам.

Сделать доступным для скоринга

Включить входные поля. Задает, включаются ли по умолчанию входные поля для скоринга.

Вес класса

Используйте эту опцию, если вы хотите изменить относительную важность конкретных классов при построении модели.

Примечание: Эта опция включается только в том случае, если для классификации вы используете KNN. Если выполняется регрессия (то есть тип полей назначения - количественный), эта опция отключается.

По умолчанию назначается значение 1 для всех классов, то есть их вес одинаков. Задавая различные численные веса для разных меток классов, вы инструктируете алгоритм соответствующим образом взвешивать обучающие наборы отдельных классов.

Чтобы изменить вес, дважды щелкните по нему в столбце **Вес** и внесите нужные вам изменения.

Значение. Набор меток классов, полученный из возможных значений поля назначения.

Вес. Вес, назначаемый конкретному классу. Назначение классу более высокого веса делает модель более чувствительной к этому классу относительно остальных классов.

Разделительная кластеризация Netezza

Разделительная кластеризация - это способ кластерного анализа, в котором алгоритм запускается повторно, чтобы разделять кластер на подкластеры, пока не будет достигнута заданная точка остановки.

Образование кластеров начинается с одного кластера, содержащего все обучающие экземпляры (записи). Первая итерация алгоритма делит набор данных на два подкластера, которые делятся далее последовательными итерациями на последующие подкластеры. Критерии остановки задаются как максимальное число итераций, максимальный уровень деления данных и минимальное требуемое число экземпляров для дальнейшего разделения.

Результирующее иерархическое дерево кластеризации можно использовать для классификации экземпляров, распространяя их вниз от корневого кластера, как в следующем примере.

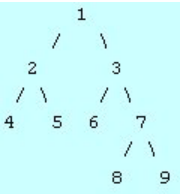

Рисунок 4. Пример дерева разделительной кластеризации

На каждом уровне наилучший подходящий подкластер выбирается по сравнению расстояний экземпляра от центров подкластеров.

Когда экземпляры оцениваются с помощью примененного уровня иерархии -1 (значение по умолчанию), скоринг возвращает только конечный кластер (кластер листа), где листы обозначены отрицательными числами. В данном примере это может быть один из кластеров 4, 5, 6, 8 или 9. Но если для уровня иерархии задано значение 2, скоринг возвратит, например, один из кластеров на втором уровне ниже корневого кластера, то есть 4, 5, 6 или 7.

Опции полей разделительной кластеризации Netezza

На вкладке Поля указывается, будут ли использоваться значения ролей полей, уже определенные на расположенных выше узлах, или назначение полей будет выполнено вручную.

Использовать заранее заданные роли. Эта опция использует значения ролей (назначения, предикторы и так далее) с расположенного выше узла Тип (или с вкладки Типы лежащего выше узла источника).

Настроить назначения полей. Выберите эту опцию при желании задать назначения, предикторы и другие роли вручную на этом экране.

Поля. При помощи кнопок со стрелками назначьте элементы из этого списка вручную для различных полей ролей в правой части экрана. Значки обозначают допустимые уровни измерения для каждого поля роли.

Нажмите кнопку **Все**, чтобы выбрать все поля в списке, или нажмите кнопку отдельного уровня измерений, чтобы выбрать все поля с этим уровнем.

ID записи. Поле, которое будет использоваться как уникальный идентификатор записи.

Предикторы (входные поля). Выберите одно или несколько полей в качестве входных для предсказания.

Опции построения разделительной кластеризации Netezza

На вкладке Параметры конструкции задаются все опции для построения модели. Можно, конечно, просто нажать кнопку **Выполнить**, чтобы построить модель со всеми опциями по умолчанию, но скорее всего вы захотите настроить конструкцию для своих собственных целей.

Мера расстояния. Метод, используемый для измерения расстояния между точками данных; чем больше расстояния, тем больше различия. Опции:

- **Евклидово**. (по умолчанию) Расстояние между двумя точками вычисляется путем их соединения по прямой.
- **Манхеттенская**. Расстояние между двумя точками вычисляется как сумма абсолютных разностей между их координатами.
- **Канберра**. Аналогична манхеттенской мере расстояния, но более чувствительна к точкам данных, находящимся ближе к источнику.
- **Максимум**. Расстояние между двумя точками вычисляется как наибольшее из различий по любой из их координат.

Максимальное количество итераций. Алгоритм работает, выполняя несколько итераций одного процесса. Эта опция позволяет остановить обучение модели после заданного числа итераций.

Максимальная глубина деревьев кластера. Максимальное количество уровней, до которых может разделяться набор данных.

Воспроизвести результаты. Включите этот переключатель, если вы хотите задать стартовое число генератора псевдослучайных чисел, что позволяет воспроизвести результаты. Вы можете или задать целое число, или нажать кнопку **Генерировать**, чтобы сгенерировать псевдослучайное целое число.

Минимальное количество экземпляров для разделения. Минимальное количество записей, которые можно расщепить. Когда число нерасщепленных записей становится меньше заданного значения, дальнейшие расщепления не производятся. Это поле можно использовать для предотвращения создания очень маленьких подгрупп в кластерном дереве.

PCA Netezza

Анализ главных компонентов (principal component analysis, PCA) - это мощное средство сокращения объема данных, разработанное для уменьшения сложности данных. PCA находит линейные комбинации входных полей, которыми главным образом определяются изменения в целом наборе полей, где компоненты ортогональны друг другу (не скоррелированы). Цель PCA - найти небольшое количество производных полей (главных компонентов), которые эффективно суммируют информацию исходного набора входных полей.

Опции полей PCA Netezza

На вкладке Поля указывается, будут ли использоваться значения ролей полей, уже определенные на расположенных выше узлах, или назначение полей будет выполнено вручную.

Использовать заранее заданные роли. Эта опция использует значения ролей (назначения, предикторы и так далее) с расположенного выше узла Тип (или с вкладки Типы лежащего выше узла источника).

Настроить назначения полей. Выберите эту опцию при желании задать назначения, предикторы и другие роли вручную на этом экране.

Поля. При помощи кнопок со стрелками назначьте элементы из этого списка вручную для различных полей ролей в правой части экрана. Значки обозначают допустимые уровни измерения для каждого поля роли.

Нажмите кнопку **Все**, чтобы выбрать все поля в списке, или нажмите кнопку отдельного уровня измерений, чтобы выбрать все поля с этим уровнем.

ID записи. Поле, которое будет использоваться как уникальный идентификатор записи.

Предикторы (входные поля). Выберите одно или несколько полей в качестве входных для предсказания.

Опции построения PCA Netezza

На вкладке Параметры конструкции задаются все опции для построения модели. Можно, конечно, просто нажать кнопку **Выполнить**, чтобы построить модель со всеми опциями по умолчанию, но скорее всего вы захотите настроить конструкцию для своих собственных целей.

Центрировать данные перед вычислением PCA. Если данный переключатель включен (по умолчанию), эта опция перед анализом выполняет центрирование данных (называемое также "извлечение среднего"). Центрирование данных необходимо для обеспечения того, чтобы главный компонент описывал направление максимального изменения, в противном случае компонент может больше соответствовать среднему значению данных. Обычно этот переключатель выключается только для повышения производительности, когда данные уже подготовлены этим способом.

Выполнить масштабирование данных перед вычислением PCA. Эта опция выполняет масштабирование данных перед анализом. Это делает анализ менее произвольным, когда различные переменные измеряются разными единицами. В своей простейшей форме масштабирование данных достигается делением каждой переменной на ее среднеквадратичное отклонение.

Использовать менее точный, но быстрый метод вычисления PCA. При выборе этой опции алгоритм использует менее точный, но быстрый способ нахождения главных компонентов (forceEigensolve).

Дерево регрессии Netezza

Дерево регрессии - это алгоритм типа построения дерева, который несколько раз делит выборку наблюдений для получения подмножеств одного типа на основе значений числового поля назначения. Как и в случае с деревом решений, деревья регрессии разделяют данные на подмножества, конечные из которых (листья) соответствуют относительно небольшим или однородным подмножествам. Разделения выбираются для уменьшения дисперсии значений атрибутов назначения, чтобы они могли достаточно хорошо предсказываться по своим средним значениям на конечных группах разделения (листьях).

Вывод модели возвращает форму текстового представления дерева. Каждая строка текста соответствует узлу или конечному узлу, а отступ текста отражает уровень дерева. Для узла выводится условие разбиения; для конечного узла - метка назначенного класса.

Опции построения дерева регрессии Netezza - рост дерева

Вы можете задать опции построения для роста и усечения дерева.

Если вы не хотите использовать значения по умолчанию, нажмите кнопку **Настроить** и измените значения.

Для построения дерева доступны следующие опции:

Максимальное количество уровней в дереве. Максимальное количество уровней, до которого может расти дерево от корневого узла, то есть сколько раз рекурсивно разделена выборка. Значение по умолчанию - 62, это максимальное количество уровней дерева для целей моделирования.

Примечание: Если программа просмотра в слепке модели выводит текстовое представление модели, может выводиться не более 12 уровней дерева.

Критерии расщепления. Эта опция управляет, когда остановить расщепление дерева. Если вы не хотите использовать значения по умолчанию, нажмите кнопку **Настроить** и измените значения.

- **Показатель оценки расщепления.** Эта мера оценки класса служит для оценки лучшего места расщепления дерева.

  Примечание: В настоящее время единственная возможная опция - дисперсия.

- **Значение минимального улучшения для расщеплений**. Минимальное значение уменьшения неоднородности для создания нового расщепления на дереве. Цель построения дерева - создать подгруппы со сходными выходными значениями, чтобы минимизировать неоднородности на каждом узле. Если наилучшее расщепление ветви уменьшает неоднородность меньше, чем на заданное критерием расщепления значение, ветвь не расщепляется.
- **Минимальное количество экземпляров для расщепления.** Минимальное количество записей, которые можно расщепить. Когда число нерасщепленных записей становится меньше заданного значения, дальнейшие расщепления не производятся. Это поле можно использовать для предотвращения создания маленьких подгрупп в дереве.

Статистика. Этот параметр определяет, сколько статистик будет включено в модель. Выберите один из перечисленных ниже вариантов.

- **Все.** Включается вся статистика, связанная со столбцами и со значениями.

  Примечание: Этот параметр включает сбор максимального объема статистики, что может повлиять на производительность вашей системы. Если вы не хотите просматривать модель в графическом формате, задайте **Нет**.
- **Столбцы.** Включается статистика по столбцам.
- **Нет.** Включается только статистика, требуемая для скоринга модели.

Опции построения дерева регрессии Netezza - сокращение дерева

Опции сокращения можно использовать для определения критерия сокращения для дерева регрессии. Предназначение сокращения - снижение риска переобучения, для чего удаляются излишние подгруппы, которые не повышают ожидаемой точности на новых данных.

Мера сокращения. Мера сокращения обеспечивает, чтобы оцененная точность модели после удаления листа дерева оставалась в приемлемых пределах. Вы можете выбрать одну из следующих мер.

- **mse.** Среднеквадратичная ошибка (mean squared error, MSE) - это опция по умолчанию, она измеряет, насколько подогнанная линия близка к точкам данных.
- **r2.** R-квадрат измеряет долю изменчивости зависимой переменной, объясняемой регрессионной моделью.
- **Пирсона.** Коэффициент корреляции Пирсона измеряет силу взаимосвязи между линейно зависимыми переменными с нормальным распределением.
- **Спирмана.** Коэффициент корреляции Спирмана обнаруживает нелинейные взаимосвязи, которые по корреляции Пирсона могут казаться слабыми, но на самом деле быть сильными.

Данные для сокращения. Вы можете использовать некоторые или все обучающие данные для оценки ожидаемой точности по новым данных. Для этой цели можно также использовать отдельный набор данных сокращения из заданной таблицы.

- **Использовать все данные обучения**. Эта опция (задаваемая по умолчанию) использует все обучающие данные для оценки точности модели.
- **Использовать % данных обучения для усечения**. Эта опция используется для разбиения данных на два набора: один для обучения, а другой для сокращения (при помощи заданного здесь процента).

  Выберите **Воспроизвести результаты**, если хотите задать начальное значение для генератора псевдослучайных чисел, что гарантирует разделение данных одним и тем же способом при каждом запуске потока. Можно либо задать целое число в поле **Начальное значение, используемое для усечения**, либо выбрать действие **Сгенерировать**, создающее псевдослучайное целое число.
- **Использовать данные из существующей таблицы**. Укажите имя таблицы отдельного набора данных сокращения для оценки точности модели. Выполнение этого действия считают более надежным, чем использование данных обучения. Однако эта опция может привести к удалению из набора обучения большого поднабора данных и тем самым к снижению качества дерева решений.

Линейная регрессия Netezza

Линейные модели предсказывают значения непрерывных переменных назначения, основываясь на взаимосвязи между переменной назначения и одним или несколькими предикторами. При ограничении только непосредственно моделируемыми линейными взаимосвязями, модели линейной регрессии относительно просты и дают простые математические формулы для оценок. Линейные модели быстрее, эффективнее и проще в использовании, хотя их применимость ограничена в сравнении с более точными алгоритмами регрессии.

Опции построения линейной регрессии Netezza

На вкладке Параметры конструкции задаются все опции для построения модели. Можно, конечно, просто нажать кнопку **Выполнить**, чтобы построить модель со всеми опциями по умолчанию, но скорее всего вы захотите настроить конструкцию для своих собственных целей.

Использовать декомпозицию отдельного значения для решения уравнений. Использование матрицы декомпозиции отдельного значения вместо исходной матрицы дает преимущество большей устойчивости при численных ошибках, а также может ускорить вычисления.

Включить в модель свободный член. Включение в модель свободного члена повышает общую точность решения.

Рассчитать диагностики модели. Эта опция дает возможность вычисления нескольких диагностик для модели. Результаты хранятся в матрицах или таблицах для последующего изучения. Результаты диагностики включают в себя значения R-квадрат, сумму квадратов остатков, оценку дисперсии, среднеквадратичное отклонение, *p*-значение и *t*-значение.

Эти диагностические показатели относятся к точности и полезности модели. Вы должны запускать диагностики отдельно для соответствующих данных, чтобы убедиться в выполнении предположений о линейной зависимости.

Временные ряды Netezza

Временной ряд - это последовательность числовых значений данных, измеренных в последовательные моменты времени (хотя не обязательно через равные промежутки), например, ежедневные цены акций или еженедельные данные о продажах. Анализ таких данных может быть полезен, например, в выделении такого поведения, как тенденции или сезонность (повторяющиеся структуры), и в предсказании будущего поведения по прошлым событиям.

Временные ряды Netezza поддерживают следующие алгоритмы временных рядов.

- спектральный анализ
- экспоненциальное сглаживание
- авторегрессивное интегрированное скользящее среднее (AutoRegressive Integrated Moving Average, ARIMA)
- декомпозиция сезонных тенденций

Эти алгоритмы разделяют временной ряд на тенденцию и сезонный компонент. Эти компоненты затем используются, чтобы построить модель, которую можно использовать для предсказаний.

Спектральный анализ используется для идентификации периодического поведения во временных рядах. Для временных рядов, состоящих из нескольких внутренних периодических составляющих, или при наличии в данных существенного количества случайного шума спектральный анализ представляет наиболее явный способ выделения периодических компонентов. Этот способ обнаруживает частоты периодического поведения, преобразуя ряд из зависящего от времени в частотный диапазон.

Экспоненциальное сглаживание - это способ предсказания, использующий взвешенные значения предыдущих наблюдений ряда для предсказания будущих значений. При экспоненциальном сглаживании влияние наблюдений убывает во времени по экспоненте. Этот способ прогнозирует некоторую точку во времени, корректируя прогнозы по мере поступления новых данных и принимая во внимание добавки, тенденцию и сезонность.

Модели **ARIMA** предоставляют более сложные способы моделирования тенденций и сезонных компонентов по сравнению с моделями экспоненциального сглаживания. При этом методе непосредственно задаются порядки авторегрессии и скользящего среднего, а также порядок исчисления разностей.

Примечание: С практической точки зрения модели ARIMA наиболее полезны, если вы хотите включить в рассмотрение предикторы, помогающие объяснить поведение прогнозируемого ряда, например, количество отправленных по почте каталогов или число посещений Web-страницы компании. Модели экспоненциального сглаживания описывают поведение временного ряда, не пытаясь объяснить, с чем такое поведение связано.

Декомпозиция сезонных тенденций удаляет периодическое поведение из временных рядов, чтобы выполнить анализ такой тенденции, а затем выбирает основной вид тенденции, например, квадратичную функцию. Для базовой формы тенденции может использоваться много параметров, которые определяются, чтобы минимизировать среднеквадратичную ошибку остатков (то есть, разностей между подогнанными и наблюдаемыми значениями временного ряда).

Интерполяция значений во временных рядах Netezza

Интерполяция - это процесс оценки и вставки пропущенных значений в данных временного ряда.

Если интервалы временного ряда регулярны, но некоторые значения просто не представлены, пропущенные значения можно оценить при помощи линейной интерполяции. Рассмотрим следующий ряд ежемесячных прибытий пассажиров в терминал аэропорта.

Таблица 12. Ежемесячные прибытия в пассажирский терминал

Месяц	Пассажиров
3	3500000
4	3900000
5	-
6	3400000
7	4500000
8	3900000
9	5800000
10	6000000

В этом случае линейная интерполяция оценивает пропущенное значение для месяца 5 как 3650000 (среднее между значениями для месяцев 4 и 6).

Нерегулярные интервалы обрабатываются иначе. Рассмотрим следующий ряд показаний температуры.

Таблица 13. Показания температуры

Дата	Время	Температура
2011-07-24	7:00	15
2011-07-24	14:00	25
2011-07-24	21:00	23

Таблица 13. Показания температуры (продолжение)

Дата	Время	Температура
2011-07-25	7:15	16
2011-07-25	14:00	26
2011-07-25	20:55	24
2011-07-27	7:00	17
2011-07-27	14:00	27
2011-07-27	22:00	24

Здесь присутствуют показания, снятые в трех точках времени в течение трех дней, но это время разное, и только в некоторых случаях совпадает в разные дни. Кроме этого, только два дня последовательны.

Эту ситуацию можно обработать двумя способами - или вычислить агрегаты, или определить размер шага.

Агрегаты могут быть ежедневными агрегатами, вычисленными в соответствии с формулой на основе семантического знания о данных. Выполнение этой процедуры может привести к следующему набору данных.

Таблица 14. Показания температуры (агрегированные)

Дата	Время	Температура
2011-07-24	24:00	22
2011-07-25	24:00	23
2011-07-26	24:00	null
2011-07-27	24:00	24

Как вариант, алгоритм может рассматривать этот ряд как другой отдельный ряд и определить подходящий размер шага. В данном случае определенным алгоритмом шагом может быть значение 8 часов, что приводит к следующему.

Таблица 15. Показания температуры с вычисленным размером шага

Дата	Время	Температура
2011-07-24	6:00	
2011-07-24	14:00	25
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	26
2011-07-25	22:00	
2011-07-26	6:00	
2011-07-26	14:00	
2011-07-26	22:00	
2011-07-27	6:00	
2011-07-27	14:00	27
2011-07-27	22:00	24

Здесь только четыре показания соответствуют исходным измерениям, но с использованием других известных значений исходного ряда отсутствующие значения снова можно вычислить интерполяцией.

Опции полей временных рядов Netezza

На вкладке Поля задаются роли входных полей для исходных данных.

Поля. Используйте кнопки со стрелками для назначения вручную элементов из этого списка полям различных ролей справа на экране. Значки обозначают допустимые уровни измерения для каждого поля роли.

Цель. Выберите одно поле в качестве назначения для предсказания. Это должно быть поле с количественным уровнем измерения.

(Предиктор) Моменты времени. (обязательно) Входное поле, содержащее значения даты или времени для временного ряда. Это должно быть поле с количественным или категориальным уровнем измерения и с типом хранения данных Дата, Время, Отметка времени или Численный. Задаваемый здесь тип хранения данных для поля определяет также тип входных данных для некоторых полей на других вкладках этого узла моделирования.

(Предиктор) ID временных рядов (по). Поле, содержащее значения ID временных рядов; используйте это поле, если во входных данных содержится несколько временных рядов.

Опции построения временных рядов Netezza

Есть два уровня опций построения:

- Базовый - параметры для выбора алгоритма, интерполяции и используемого диапазона времени.
- Расширенный - параметры для прогнозирования

В этом разделе описываются базовые опции.

На вкладке Параметры конструкции задаются все опции для построения модели. Можно, конечно, просто нажать кнопку **Выполнить**, чтобы построить модель со всеми опциями по умолчанию, но скорее всего вы захотите настроить конструкцию для своих собственных целей.

Алгоритм

Эти параметры относятся к алгоритму исследования временных рядов, который будет использоваться.

Название алгоритма. Выберите алгоритм исследования временных рядов, который вы хотите использовать. Доступны следующие алгоритмы - **спектральный анализ**, **экспоненциальное сглаживание** (по умолчанию), **ARIMA** или **декомпозиция сезонных тенденций**. Дополнительную информацию смотрите в разделе “Временные ряды Netezza” на стр. 96.

Тенденция. (только для экспоненциального сглаживания) Если во временном ряду проявляется какая-то тенденция, простое экспоненциальное сглаживание работает не очень хорошо. Используйте это поле для указания тенденции, если такая существует, чтобы она могла быть учтена алгоритмом.

- **Определяется системой.** (по умолчанию) Система сама пытается найти оптимальное значение для этого параметра.
- **Нет (N).** Временной ряд не проявляет тенденции.
- **Аддитивный (A).** Тенденция равномерного нарастания во времени.
- **Демпфированный аддитивный (DA).** Аддитивная тенденция, затухающая со временем.
- **Мультипликативный (M).** Тенденция возрастания во времени, обычно более быстро, чем равномерная аддитивная тенденция.
- **Демпфированный мультипликативный (DM).** Мультипликативная тенденция, затухающая со временем.

Сезонность. (только для экспоненциального сглаживания) Используйте это поле для указания, проявляет ли временной ряд какие-либо сезонные структуры в данных.

- **Определяется системой.** (по умолчанию) Система сама пытается найти оптимальное значение для этого параметра.
- **Нет (N).** Временной ряд не проявляет сезонных структур.
- **Аддитивный (A).** Структура сезонных флуктуаций со временем проявляет равномерную восходящую тенденцию.
- **Мультипликативный (M).** То же, что при аддитивной сезонности, но в дополнение к амплитуде сезонных флуктуаций (расстоянию между высшими и низшими точками) происходит увеличение и по отношению к общему возрастающей тенденции флуктуаций.

Использовать системные параметры для ARIMA. (только для ARIMA) Выберите эту опцию, если вы хотите, чтобы система сама определила параметры для алгоритма ARIMA.

Задать. (только для ARIMA) Выберите эту опцию и нажмите кнопку, чтобы задать параметры ARIMA вручную.

Интерполяция

Если среди исходных данных временного ряда есть пропущенные значения, выберите способ вставки оцененных значений, чтобы заполнить пробелы в данных. Дополнительную информацию смотрите в разделе “Интерполяция значений во временных рядах Netezza” на стр. 97.

- **Линейное.** Выберите этот способ, если интервалы временного ряда регулярны, но некоторые значения просто не представлены.
- **Экспоненциальные сплайны.** Подгоняет гладкую кривую, где известные значения точек данных показывают увеличение или уменьшение с большой скоростью.
- **Кубические сплайны.** Подгоняет известные точки данных гладкой кривой для оценки отсутствующих значений.

Диапазон времени

Здесь можно выбрать, использовать ли для создания модели полный диапазон данных временного ряда или подмножество смежных значений этих данных. Допустимые входные значения для этих полей определяются типом хранения данных для данного поля, заданным для точек времени на вкладке Поля. Дополнительную информацию смотрите в разделе “Опции полей временных рядов Netezza” на стр. 99.

- **Использовать самое раннее и самое позднее время, доступное в данных.** Выберите эту опцию, если вы хотите использовать полный диапазон данных временного ряда.
- **Задать временное окно.** Выберите эту опцию, если вы хотите использовать только часть временного ряда. Чтобы задать границы, используйте поля **Самое раннее время (от)** и **Самое позднее время (до)**.

Структура ARIMA

Задайте значения различных несезонных и сезонных компонентов модели ARIMA. В каждом случае определите операцию **=** (равно) или **<=** (меньше или равно), а затем задайте значение в соседнем поле. Значения должны быть неотрицательными целыми числами, задающими степени.

Несезонная. Значения различных несезонных компонентов модели.

- **Степени автокорреляции (p).** Количество порядков авторегрессии в модели. Порядки авторегрессии задают, какие предыдущие значения из ряда использовались для предсказания текущих значений. Например, порядок авторегрессии 2 означает, что для предсказания текущего значения использовалось на два периода более раннее значение из ряда.
- **Отклонение (d).** Задает порядок исчисления разностей, применимый к ряду до оценки моделей. Вычисление разностей необходимо при наличии тенденций (ряды с тенденциями обычно нестационарные, а моделирование АРПСС предполагает стационарность) и используется для удаления этих эффектов. Порядок исчисления разностей соответствует степени тенденции ряда - разности первого порядка учитывают линейные тенденции, разности второго порядка - квадратичные, и так далее.

- **Скользящее среднее (q).** Количество порядков скользящего среднего в модели. Порядки скользящего среднего задают, как отклонения от среднего значения ряда предыдущих значений используются для предсказания текущих значений. Например, порядки скользящего среднего 1 и 2 указывают, что отклонения от среднего значения ряда для каждого значения за прошлые два периода будут рассматриваться для предсказания текущих значений ряда.

Сезонная. Сезонные компоненты автокорреляции (SP), отклонения (SD) и скользящего среднего (SQ) играют ту же роль, что и их несезонные аналоги. Однако для сезонных порядков на текущие значения ряда влияют предыдущие значения, отделенные одним или несколькими сезонными периодами. Например, для ежемесячных данных (сезонный период 12) сезонный порядок 1 означает, что на текущее значение ряда влияет значение ряда на 12 периодов ранее текущего. Тем самым для ежемесячных данных сезонный порядок 1 - это то же самое, что несезонный порядок 12.

Сезонные параметры рассматриваются только в том случае, если в данных обнаружена сезонность или вы задаете параметр Период на вкладке Дополнительно.

Опции построения временных рядов Netezza - дополнительно

Дополнительные параметры можно использовать, чтобы задать опции прогнозирования.

Использовать системные параметры для опций построения модели. Выберите эту опцию, если вы хотите, чтобы система сама определила дополнительные параметры.

Задать. Выберите эту опцию, если вы хотите задать дополнительные параметры вручную. (Эта опция недоступна, если используется алгоритм спектрального анализа).

- **Период/Единицы для периода.** Период времени, после которого некоторое характеристическое поведение временного ряда повторяет само себя. Например, для временных рядов еженедельных продаж задайте значение 1 для периода и недели для единиц. **Период** должен быть неотрицательным целым числом; **Единицы для периода** могут быть следующие - **Миллисекунды**, **Секунды**, **Минуты**, **Часы**, **Дни**, **Недели**, **Кварталы** или **Годы**. Не задавайте значение **Единицы для периода**, если не задан **Период** или тип значений времени не численный. Однако если задан **Период**, нужно задать и **Единицы для периода**.

Параметры для прогнозирования. Вы можете выбрать вариант создания прогноза - или до конкретной точки во времени, или в конкретных точках времени. Допустимые входные значения для этих полей определяются типом хранения данных для данного поля, заданным для точек времени на вкладке Поля. Дополнительную информацию смотрите в разделе “Опции полей временных рядов Netezza” на стр. 99.

- **Горизонт прогноза.** Выберите эту опцию, если вы хотите задать только конечную точку для прогнозирования. Прогнозы будут делаться только до этой точки времени.
- **Времена прогноза.** Выберите эту опцию, чтобы задать одну или несколько точек времени, для которых будут делаться прогнозы. Нажмите кнопку **Добавить**, чтобы добавить новую строку в таблицу точек времени. Для удаления строки выберите эту строку и нажмите кнопку **Удалить**.

Опции модели Netezza Time Series

На вкладке Опции модели можно выбрать, задавать ли имя для модели, или сгенерировать имя автоматически. Вы можете задать также значения по умолчанию для опций выходных данных модели.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Сделать доступным для скоринга. Здесь можно задать значения по умолчанию для опций скоринга, которые появятся в диалоговом окне для слепка модели.

- **Включать хронологические значения в вывод .** По умолчанию выходные данные модели не включают в себя значения хронологических данных (данных, используемых для предсказаний). Включите этот переключатель, чтобы включить в выходные данные эти значения.

- **Включать интерполированные значения в вывод .** Если вы выбрали включение в выходные данные хронологических значений, включите этот переключатель при желании выводить и интерполированные значения (если такие существуют). Обратите внимание на то, что интерполяция работает только с хронологическими данными, поэтому этот переключатель недоступен, если не выбрана опция **Включить хронологические значения в выходные данные**. Дополнительную информацию смотрите в разделе “Интерполяция значений во временных рядах Netezza” на стр. 97.

Обобщенный линейный анализ Netezza

Линейная регрессия - это традиционный статистический метод для классификации записей на основании значений числовых входных полей. Линейная регрессия подгоняет прямую линию или поверхность, минимизирующую разности между предсказанными и фактическими выходными значениями. Линейные модели полезны при моделировании многих явлений реального мира из-за своей простоты при обучении и применении моделей. Однако линейные модели предполагают нормальное распределение зависимой переменной (переменной назначения) и линейное воздействие независимых переменных (предикторов) на зависимую переменную.

Есть много ситуаций, когда линейная регрессия полезна, но приведенные предположения неприменимы. Например, при моделировании выбора потребителя между дискретным числом продуктов у зависимой переменной будет полиномиальное распределение. Аналогично, при моделировании дохода в зависимости от возраста доход обычно увеличивается с возрастом, но связь между двумя переменными в виде прямой линии маловероятна.

Для таких ситуаций можно использовать обобщенную линейную модель. Обобщенные линейные модели расширяют модель линейной регрессии, так что зависимая переменная связана с предикторными переменными посредством заданной функции связи, которую можно выбрать из числа подходящих функций. Более того, эта модель допускает наличие у зависимой переменной отличного от нормального распределения, например, распределения Пуассона.

Алгоритм итерационно ищет наилучшую подходящую модель до заданного числа итераций. При вычислении наилучшей подгонки ошибка представляется суммой квадратов разностей между предсказанными и фактическими значениями зависимой переменной.

Опции обобщенной линейной модели Netezza - Общие

На вкладке Опции модели можно выбрать, задавать ли имя для модели, или сгенерировать имя автоматически. Можно задать также различные параметры, относящиеся к модели, функцию связи, взаимодействия входных полей (если такое существует) и набор значений по умолчанию для опций скоринга.

Имя модели. Можно сгенерировать имя модели автоматически на основе поля назначения или поля ID (либо типа модели в случае, если никакое из этих полей не задано) либо задать пользовательское имя.

Опции полей. Вы можете задать роли входных полей для построения модели.

Общие параметры. Эти параметры относятся к критериям остановки работы алгоритма.

- **Максимальное количество итераций.** Максимальное количество итераций, которые могут быть выполнены алгоритмом; минимальное значение 1, значение по умолчанию 20.
- **Максимальная ошибка (1e).** Максимальное значение ошибки (в экспоненциальном представлении), при котором алгоритм должен остановиться при поиске модели с наилучшей подгонкой. Минимум равен 0, значение по умолчанию -3, что означает 1E-3, то есть 0,001.
- **Порог значений незначимой ошибки (1e).** Значение (в экспоненциальном представлении), ниже которого ошибки рассматриваются как имеющие нулевое значение. Минимум равен -1, значение по умолчанию -7, то есть значения меньше 1E-7 (или 0,0000001) рассматриваются как незначимые.

Параметры распределения. Эти параметры относятся к распределению зависимой переменной (переменной назначения).

- **Распределение для переменной ответа.** Тип распределения; один из следующих - **Бернулли** (по умолчанию), **Гауссово**, **Пуассона**, **отрицательное биномиальное**, **Вальда** (обратное Гауссово) и **гамма**.
- **Анализ важности независимых переменных.** (только для отрицательного биномиального распределения) Для отрицательного биномиального распределения можно задать значение параметра. Выберите, или задать значение, или использовать значение по умолчанию -1.

Параметры функции связи. Эти параметры относятся к функции связи, связывающей зависимую переменную с предикторными переменными.

- **Функция связи.** Функции, которые будут использоваться; одна из следующих функций - **тождественная**, **обратная**, **обратная негативная**, **обратная квадратичная**, **квадратный корень**, **степенная**, **нечетная степенная**, **логарифмическая**, **C-логарифмическая**, **логарифм-логарифмическая**, **C-логарифм-логарифмическая**, **логит** (по умолчанию), **пробит**, **Гауссит**, **Коши**, **Can-биномиальная**, **Can-геометрическая**, **Can-негативная биномиальная**.
- **Анализ важности независимых переменных.** (только для степенной и нечетной степенной функции связи) Если функция связи - это **степенная** или **нечетная степенная** функция, можно задать значение параметра. Выберите, или задать значение, или использовать значение по умолчанию 1.

Опции обобщенной линейной модели Netezza - взаимодействие

Панель Взаимодействие содержит опции для указания взаимодействий (то есть мультипликативных эффектов между входными полями).

Столбец Взаимодействие. Включите этот переключатель, чтобы задать взаимодействия между входными полями. Если взаимодействий нет, оставьте этот переключатель выключенным.

Введите взаимодействия в модель, выбирая одно или несколько полей в исходном списке и перетаскивая их в список взаимодействий. Тип созданного взаимодействия зависит от того, на какую активную область вы перетащили выбор.

- **Главные.** Поля, которые вы перетащили, выводятся как отдельные главные взаимодействия внизу списка взаимодействий.
- **2-факторные.** Все возможные пары отброшенных полей выводятся как 2-факторные взаимодействия в нижней части списка взаимодействий.
- **3-факторные.** Все возможные триплеты отброшенных полей выводятся как 3-факторные взаимодействия в нижней части списка взаимодействий.
- ***.** Сочетание всех отброшенных полей выводится как одно взаимодействие в нижней части списка взаимодействий.

Включить константу. Обычно в модель включают свободный член. Если вы предполагаете, что данные проходят через начало координат, свободный член можно исключить.

Кнопки в диалоговом окне

Кнопки в правой части экрана позволяют внести изменения в используемые члены модели.

Рисунок 5. Кнопка Удалить

Удалить члены из модели, выбрав члены для удаления и нажав кнопку Удалить.

Рисунок 6. Кнопки Изменить порядок

Изменить порядок членов в модели, выбрав члены, которые нужно переставить, и нажав кнопку со стрелкой вверх или вниз.

Рисунок 7. Кнопка Пользовательское взаимодействие

Добавить задаваемый член

Пользовательские взаимодействия можно задать в форме *n1*x1*x1*x1...* Выберите поле из списка **Поля**, нажмите кнопку со стрелкой направо для добавления поля в **Пользовательский член**, нажмите кнопку **по ***, выберите следующее поле, нажмите кнопку со стрелкой направо и так далее. После построения пользовательского взаимодействия нажмите кнопку **Добавить член**, чтобы вернуться на панель Взаимодействие.

Опции обобщенной линейной модели Netezza - опции скоринга

Сделать доступным для скоринга. Здесь можно задать значения по умолчанию для опций скоринга, которые появятся в диалоговом окне для слепка модели. Дополнительную информацию смотрите в разделе “Слепок обобщенной линейной модели Netezza - вкладка Параметры” на стр. 113.

- **Включить входные поля .** Включите этот переключатель, если вы хотите в выходных данных модели выводить выходные поля, а также предсказания.

Управление моделями IBM Netezza Analytics

Модели IBM Netezza Analytics добавляются на холст и палитру моделей таким же образом, как остальные модели IBM SPSS Modeler, и их можно использовать таким же образом. Однако есть несколько важных особенностей, поскольку все модели IBM Netezza Analytics, созданные в IBM SPSS Modeler, фактически ссылаются на модели, хранящиеся на сервере базы данных. Таким образом, для правильной работы потока требуется соединение с базой данных, в которой модель была создана, и чтобы таблица модели не изменялась внешним процессом.

Скоринг моделей IBM Netezza Analytics

Модели представлены на холсте золотым значком слепка модели. Основная цель слепка - скоринг данных для прогнозирования и дальнейший анализ свойств модели. Оценки добавляются в виде одного или нескольких дополнительных полей данных, которые можно делать видимыми, присоединяя к слепку узел таблицы и выполняя ветвь потока, как описано ниже в этом разделе. Некоторые диалоговые окна слепка, такие как Дерево решений и Дерево регрессии, содержат дополнительную вкладку Модель, на которой выводится наглядное представление модели.

Дополнительные поля отличаются префиксом `$<id>-` перед именем поля назначения, где `<id>` зависит от модели и показывает тип добавляемой информации. Различные идентификаторы описываются в разделах, посвященных слепку той или иной модели.

Чтобы посмотреть оценки, выполните следующие действия:

1. Присоедините к слепку модели узел таблицы.
2. Откройте узел таблицы.

3. Нажмите кнопку **Выполнить**.
4. Прокрутите окно табличного вывода до правого края, чтобы увидеть дополнительные поля и их оценки.

Вкладка сервера слепков модели Netezza

На вкладке Сервер можно задать параметры сервера для скоринга модели. Можно по-прежнему использовать соединение с сервером, заданное выше, а можно переместить данные в другую базу данных, указав ее здесь.

Подробности сервера баз данных Netezza. Здесь задаются подробности о соединении для базы данных, которую хотите использовать для модели.

- **Использовать восходящее соединение**. (По умолчанию) Использует подробности о соединении, заданные на восходящем узле, например, узле источника базы данных. *Примечание*: Эта опция работает, только если все восходящие узлы могут использовать обратный перенос SQL (SQL pushback). В этом случае перенос данных из базы данных не нужен, поскольку SQL полностью реализует все восходящие узлы.
- **Переместить данные на соединение**. Переносит данные в указанную здесь базу данных. Это действие позволяет моделированию работать, если данные находятся в другой базе данных IBM Netezza, в базе данных от другого поставщика или даже в плоском файле. Кроме того, данные переносятся обратно в указанную здесь базу данных, если они были извлечены из-за того, что узел не выполнил обратный перенос SQL. Нажмите кнопку **Правка**, чтобы найти и выбрать соединение. *Предостережение*: IBM Netezza Analytics обычно используется с очень большими наборами данных. Передача больших объемов данных между базами данных или из базы данных и обратно может занимать очень много времени, и по возможности его следует избегать.

Имя модели. Имя модели. Это имя выводится только для информации, изменить его здесь нельзя.

Слепки моделей деревьев решений Netezza

Слепок модели дерева решений показывает выходные данные операции моделирования и позволяет задать некоторые опции для скоринга модели.

При запуске потока, содержащего узел моделирования дерева решений, на этот узел по умолчанию добавляется одно новое поле, имя которого получается из имени модели.

Таблица 16. Поле скоринга модели для дерева решений

Имя добавленного поля	Значение
$I-*имя_модели*	Предсказанное значение для текущей записи.

Если вы выбираете опцию **Вычислить вероятности назначенных классов для записей скоринга** на узле моделирования или в слепке модели и запускаете поток, добавляется следующее поле.

Таблица 17. Поле скоринга модели для дерева решений - дополнительно

Имя добавленного поля	Значение
$IP-*имя_модели*	Значение достоверности (от 0,1 до 1,0) для предсказания.

Слепок дерева решений Netezza - вкладка Модель

На вкладке **Модель** выводится важность предиктора для модели дерева решений в графическом формате. Длина полосы отражает важность предиктора.

Примечание: Когда вы работаете с IBM Netezza Analytics Версии 2.x или более ранних версий, содержимое модели дерева решений выводится только в текстовом формате.

Для этих версий выводится следующая информация:

- Каждая строка текста соответствует узлу или конечному узлу.
- Отступ отражает уровень дерева.
- Для узла выводится условие расщепления.
- Для конечного узла выводится метка назначенного класса.

Слепок дерева решений Netezza - вкладка Параметры

На вкладке Параметры можно задать несколько опций для скоринга модели.

Включить входные поля. Эта опция, если она выбрана, передает весь исходный нисходящий поток входных полей, добавляя к каждой строке данных дополнительные поля моделирования. Если выключить этот переключатель, будут передаваться только поле ID записи и дополнительные поля моделирования, из-за чего поток будет обрабатываться быстрее.

Вычислить вероятности назначенных классов для записей скоринга. (Только для Дерева решений и Наивного критерия Байеса) Эта опция, если она выбрана, означает, что дополнительные поля моделирования будут содержать поле достоверности (то есть, поле вероятностей) также, как и поле предсказаний. Если выключить этот переключатель, будет генерироваться только поле предсказаний.

Слепок модели k-средних Netezza

Слепки моделей K-средних содержат всю информацию, захваченную моделью кластеризации, а также информацию об обучающих данных и процессе оценки.

При запуске потока, содержащего узел моделирования K-средних, этот узел добавляет два новых поля, содержащих информацию о принадлежности к кластеру и расстоянии от назначенного центра кластера для данной записи. Имена новых полей получаются из имени модели с префиксами *$KM-* для принадлежности к кластеру и *$KMD-* для расстояния от центра кластера. Например, если модель называется *Kmeans*, новые поля будут называться *$KM-Kmeans* и *$KMD-Kmeans*.

Вкладка Слепок K-средних Netezza - Модель

Вкладка **Модель** содержит различные графические представления, на которых приводится сводная статистика и распределения для полей в кластерах. Можно экспортировать данные из модели или же экспортировать представление как изображение.

Когда вы работаете с IBM Netezza Analytics Версии 2.x или более ранних версий, или когда вы строите модель с мерой Махаланобиса в качестве меры расстояния, содержимое моделей K-средних выводится только в текстовом формате.

Для этих версий выводится следующая информация:

- **Сводные статистики.** Для самого маленького и самого большого кластера сводные статистики показывают число записей. В сводных статистиках выводится также процент набора данных, занимаемый этими кластерами. Этот список показывает также отношение размеров самого большого и самого маленького кластеров.
- **Сводка кластеризации.** В сводке кластеризации перечисляются кластеры, созданные алгоритмом. Для каждого кластера в таблице показано количество записей в этом кластере вместе со средним расстоянием от центра кластера до этих записей.

Слепок K-средних Netezza - вкладка Параметры

На вкладке Параметры можно задать несколько опций для скоринга модели.

Включить входные поля. Эта опция, если она выбрана, передает весь исходный нисходящий поток входных полей, добавляя к каждой строке данных дополнительные поля моделирования. Если выключить этот переключатель, будут передаваться только поле ID записи и дополнительные поля моделирования, из-за чего поток будет обрабатываться быстрее.

Мера расстояния. Метод, используемый для измерения расстояния между точками данных; чем больше расстояния, тем больше различия. Опции:

- **Евклидово**. (по умолчанию) Расстояние между двумя точками вычисляется путем их соединения по прямой.
- **Манхеттенская**. Расстояние между двумя точками вычисляется как сумма абсолютных разностей между их координатами.
- **Канберра**. Аналогична манхеттенской мере расстояния, но более чувствительна к точкам данных, находящимся ближе к источнику.
- **Максимум**. Расстояние между двумя точками вычисляется как наибольшее из различий по любой из их координат.

Слепки моделей байесовской сети Netezza

Слепок модели байесовской сети предоставляет возможность задания опций для скоринга модели.

При запуске потока, содержащего узел моделирования байесовской сети, на этот узел добавляется одно новое поле, имя которого получается из имени модели.

Таблица 18. Поле скоринга модели для байесовской сети

Имя добавленного поля	Значение
$BN-*имя_модели*	Предсказанное значение для текущей записи.

Вы можете просмотреть дополнительное поле, присоединив к слепку модели узел Таблица и запустив его. Дополнительную информацию смотрите в разделе “Скоринг моделей IBM Netezza Analytics” на стр. 104.

Слепок байесовской сети - вкладка Параметры

На вкладке Параметры можно задать опции для скоринга модели.

Цель. Если вы хотите провести скоринг для поля назначения, отличающегося от текущего назначения, выберите это поле назначения здесь.

ID записи. Если поле ID записи не задано, выберите здесь это поле для использования.

Тип прогноза. Вариант алгоритма предсказания, который вы хотите использовать:

- **Наилучший (самый коррелированный соседний).** (по умолчанию) Использует наиболее коррелированный соседний узел.
- **Соседний (взвешенный прогноз соседних).** Использует взвешенный прогноз от всех соседних узлов.
- **Непустые соседние.** То же, что и в предыдущей опции, но игнорируются узлы с пустыми значениями (то есть узлы, соответствующие атрибутам с пропущенными значениями для экземпляра, прогноз для которого вычисляется).

Включить входные поля. Эта опция, если она выбрана, передает весь исходный нисходящий поток входных полей, добавляя к каждой строке данных дополнительные поля моделирования. Если выключить этот переключатель, будут передаваться только поле ID записи и дополнительные поля моделирования, из-за чего поток будет обрабатываться быстрее.

Слепки наивных моделей Байеса Netezza

Слепок наивной модели Байеса предоставляет возможность задания опций для скоринга модели.

При запуске потока, содержащего узел моделирования наивного Байеса, на этот узел по умолчанию добавляется одно новое поле, имя которого получается из имени модели.

Таблица 19. Поле скоринга наивной модели Байеса - значение по умолчанию

Имя добавленного поля	Значение
$I-*имя_модели*	Предсказанное значение для текущей записи.

Если вы выбираете опцию **Вычислить вероятности назначенных классов для записей скоринга** на узле моделирования или в слепке модели и запускаете поток, добавляются следующие два поля.

Таблица 20. Поля скоринга наивной модели Байеса - дополнительно

Имя добавленного поля	Значение
$IP-*имя_модели*	Байесовский числитель класса для экземпляра (то есть произведение априорной вероятности класса и условных вероятностей значений атрибутов экземпляра).
$ILP-*имя_модели*	Натуральный логарифм последнего.

Дополнительные поля можно просмотреть, присоединив к слепку модели узел Таблица и запустив его. Дополнительную информацию смотрите в разделе “Скоринг моделей IBM Netezza Analytics” на стр. 104.

Слепок наивного Байеса Netezza - вкладка Параметры

На вкладке Параметры можно задать опции для скоринга модели.

Включить входные поля. Эта опция, если она выбрана, передает весь исходный нисходящий поток входных полей, добавляя к каждой строке данных дополнительные поля моделирования. Если выключить этот переключатель, будут передаваться только поле ID записи и дополнительные поля моделирования, из-за чего поток будет обрабатываться быстрее.

Вычислить вероятности назначенных классов для записей скоринга. (Только для Дерева решений и Наивного критерия Байеса) Эта опция, если она выбрана, означает, что дополнительные поля моделирования будут содержать поле достоверности (то есть, поле вероятностей) также, как и поле предсказаний. Если выключить этот переключатель, будет генерироваться только поле предсказаний.

Улучшить точность вероятностей для небольших или сильно несбалансированных наборов данных. При вычислении вероятностей эта функция вызывает способ *m*-оценки для исключения нулевых вероятностей при вычислении. Этот тип вычисления вероятностей может выполняться медленней, но он дает лучшие результаты для небольших и сильно несбалансированных наборов данных.

Слепки моделей KNN Netezza

Слепок модели KNN предоставляет возможность задания опций для скоринга модели.

При запуске потока, содержащего узел моделирования KNN, на этот узел добавляется одно новое поле, имя которого получается из имени модели.

Таблица 21. Поле скоринга модели для KNN

Имя добавленного поля	Значение
$KNN-*имя_модели*	Предсказанное значение для текущей записи.

Вы можете просмотреть дополнительное поле, присоединив к слепку модели узел Таблица и запустив его. Дополнительную информацию смотрите в разделе “Скоринг моделей IBM Netezza Analytics” на стр. 104.

Слепок KNN Netezza - вкладка Параметры

На вкладке Параметры можно задать опции для скоринга модели.

Мера расстояния. Метод, используемый для измерения расстояния между точками данных; чем больше расстояния, тем больше различия. Опции:

- **Евклидово**. (по умолчанию) Расстояние между двумя точками вычисляется путем их соединения по прямой.
- **Манхеттенская**. Расстояние между двумя точками вычисляется как сумма абсолютных разностей между их координатами.
- **Канберра**. Аналогична манхеттенской мере расстояния, но более чувствительна к точкам данных, находящимся ближе к источнику.
- **Максимум**. Расстояние между двумя точками вычисляется как наибольшее из различий по любой из их координат.

Количество ближайших соседей (k). Количество ближайших соседей для конкретного наблюдения. Обратите внимание на то, что использование большего числа соседей необязательно приводит к более точной модели.

Выбор *k* управляет соотношением между предотвращением переобучения (оно может оказаться важным особенно для "зашумленных" данных) и разрешения (приводящего к различным предсказаниям для схожих экземпляров). Обычно значение *k* приходится настраивать для каждого набора данных с типичными значениями, ранг которых меняется от 1 до нескольких десятков.

Включить входные поля. Эта опция, если она выбрана, передает весь исходный нисходящий поток входных полей, добавляя к каждой строке данных дополнительные поля моделирования. Если выключить этот переключатель, будут передаваться только поле ID записи и дополнительные поля моделирования, из-за чего поток будет обрабатываться быстрее.

Нормализовать изменения перед вычислением расстояния. Эта опция, если она выбрана, стандартизирует измерения для непрерывных входных полей перед вычислением значений расстояний.

Использовать стержневые наборы для повышения производительности для больших наборов данных. Эта опция, если она выбрана, использует выборки стержневых наборов для ускорения вычисления в случае участия в нем больших наборов данных.

Слепки моделей разделительной кластеризации Netezza

Слепок модели разделительной кластеризации предоставляет возможность задания опций для скоринга модели.

При запуске потока, содержащего узел моделирования разделительной кластеризации, на этот узел добавляется два новых поля, имена которых получаются из имени модели.

Таблица 22. Поля скоринга модели для разделительной кластеризации

Имя добавленного поля	Значение
$DC-*имя_модели*	Идентификатор подкластера, которому назначена текущая запись.
$DCD-*имя_модели*	Расстояние от центра подкластера для текущей записи.

Дополнительные поля можно просмотреть, присоединив к слепку модели узел Таблица и запустив его. Дополнительную информацию смотрите в разделе “Скоринг моделей IBM Netezza Analytics” на стр. 104.

Слепок разделительной кластеризации Netezza - вкладка Параметры

На вкладке Параметры можно задать опции для скоринга модели.

Включить входные поля. Эта опция, если она выбрана, передает весь исходный нисходящий поток входных полей, добавляя к каждой строке данных дополнительные поля моделирования. Если выключить этот переключатель, будут передаваться только поле ID записи и дополнительные поля моделирования, из-за чего поток будет обрабатываться быстрее.

Мера расстояния. Метод, используемый для измерения расстояния между точками данных; чем больше расстояния, тем больше различия. Опции:

- **Евклидово**. (по умолчанию) Расстояние между двумя точками вычисляется путем их соединения по прямой.
- **Манхеттенская**. Расстояние между двумя точками вычисляется как сумма абсолютных разностей между их координатами.
- **Канберра**. Аналогична манхеттенской мере расстояния, но более чувствительна к точкам данных, находящимся ближе к источнику.
- **Максимум**. Расстояние между двумя точками вычисляется как наибольшее из различий по любой из их координат.

Примененный уровень иерархии. Уровень иерархии, который нужно применить к данным.

Слепки моделей PCA Netezza

Слепок модели PCA предоставляет возможность задания опций для скоринга модели.

При запуске потока, содержащего узел моделирования PCA, на этот узел по умолчанию добавляется одно новое поле, имя которого получается из имени модели.

Таблица 23. Поле скоринга модели для PCA

Имя добавленного поля	Значение
$F-*имя_модели*	Предсказанное значение для текущей записи.

Если вы выбираете большее единицы значение в поле **Число главных компонентов ...** на узле моделирования или в слепке модели и запускаете поток, на узле добавляется новое поле для каждого компонента. В этом случае имена полей дополняются суффиксами *-n*, где *n* - это номер компонента. Например, если модель называется *pca* и содержит три компонента, новые поля будут называться *$F-pca-1*, *$F-pca-2* и *$F-pca-3*.

Дополнительные поля можно просмотреть, присоединив к слепку модели узел Таблица и запустив его. Дополнительную информацию смотрите в разделе “Скоринг моделей IBM Netezza Analytics” на стр. 104.

Слепок PCA Netezza - вкладка Параметры

На вкладке Параметры можно задать опции для скоринга модели.

Число главных компонентов, используемых в проекции. Количество главных компонентов, до которых вы хотите уменьшить набор данных. Это значение должно не превосходить числа атрибутов (входных полей).

Включить входные поля. Эта опция, если она выбрана, передает весь исходный нисходящий поток входных полей, добавляя к каждой строке данных дополнительные поля моделирования. Если выключить этот переключатель, будут передаваться только поле ID записи и дополнительные поля моделирования, из-за чего поток будет обрабатываться быстрее.

Слепки моделей деревьев регрессии Netezza

Слепок модели дерева регрессии предоставляет возможность задания опций для скоринга модели.

При запуске потока, содержащего узел моделирования дерева решений, на этот узел по умолчанию добавляется одно новое поле, имя которого получается из имени модели.

Таблица 24. Поле скоринга модели для дерева регрессии

Имя добавленного поля	Значение
$I-*имя_модели*	Предсказанное значение для текущей записи.

Если вы выбираете опцию **Вычислить оценочную дисперсию** на узле моделирования или в слепке модели и запускаете поток, добавляется следующее поле.

Таблица 25. Поле скоринга модели для дерева регрессии - дополнительно

Имя добавленного поля	Значение
$IV-*имя_модели*	Оцененные дисперсии назначенных классов.

Дополнительные поля можно просмотреть, присоединив к слепку модели узел Таблица и запустив его. Дополнительную информацию смотрите в разделе “Скоринг моделей IBM Netezza Analytics” на стр. 104.

Слепок дерева регрессии Netezza - Вкладка Модель

На вкладке **Модель** выводится важность предиктора для модели дерева регрессии в графическом формате. Длина полосы отражает важность предиктора.

Примечание: Когда вы работаете с IBM Netezza Analytics Версии 2.x или более ранних версий, содержимое модели дерева регрессии выводится только в текстовом формате.

Для этих версий выводится следующая информация:

- Каждая строка текста соответствует узлу или конечному узлу.
- Отступ отражает уровень дерева.
- Для узла выводится условие расщепления.
- Для конечного узла выводится метка назначенного класса.

Слепок дерева регрессии Netezza - вкладка Параметры

На вкладке Параметры можно задать опции для скоринга модели.

Включить входные поля. Эта опция, если она выбрана, передает весь исходный нисходящий поток входных полей, добавляя к каждой строке данных дополнительные поля моделирования. Если выключить этот переключатель, будут передаваться только поле ID записи и дополнительные поля моделирования, из-за чего поток будет обрабатываться быстрее.

Рассчитать оценку дисперсии. Обозначает, должны ли быть включены в выходные данные дисперсии назначенных классов.

Слепки моделей линейной регрессии Netezza

Слепок модели линейной регрессии предоставляет возможность задания опций для скоринга модели.

При запуске потока, содержащего узел моделирования линейной регрессии, на этот узел добавляется одно новое поле, имя которого получается из имени модели.

Таблица 26. Поле скоринга для линейной регрессии

Имя добавленного поля	Значение
$LR-*имя_модели*	Предсказанное значение для текущей записи.

Слепок линейной регрессии Netezza - вкладка Параметры

На вкладке Параметры можно задать опции для скоринга модели.

Включить входные поля. Эта опция, если она выбрана, передает весь исходный нисходящий поток входных полей, добавляя к каждой строке данных дополнительные поля моделирования. Если выключить этот переключатель, будут передаваться только поле ID записи и дополнительные поля моделирования, из-за чего поток будет обрабатываться быстрее.

Слепок модели временных рядов Netezza

Слепок модели предоставляет доступ к выходным данным операции моделирования временного ряда. Выходные данные состоят из следующих полей.

Таблица 27. Выходные поля модели временных рядов

Поле	Описание
TSID	Идентификатор временного ряда; содержимое поля, заданного для ID временных рядов на вкладке Поля узла моделирования. Дополнительную информацию смотрите в разделе “Опции полей временных рядов Netezza” на стр. 99.
TIME	Период времени в текущем временном ряду.
HISTORY	Значения хронологических данных (данные, используемые для предсказания). Это поле включается только в том случае, если выбрана опция **Включить хронологические значения в выходные данные** на вкладке Параметры слепка модели.
$TS-INTERPOLATED	Интерполированные значения, где они используются. Это поле включается только в том случае, если выбрана опция **Включить интерполированные значения в выходные данные** на вкладке Параметры слепка модели. Интерполяция - это одна из опций на вкладке Опции построения узла моделирования.
$TS-FORECAST	Значения прогноза для временного ряда.

Для просмотра выходных данных модели присоедините узел Таблица (с вкладки Выходные данные палитры узла) к слепку модели и запустите узел Таблица.

Слепок временного ряда Netezza - вкладка Параметры

На вкладке Параметры можно задать опции для настройки выходных данных модели.

Имя модели. Имя модели, как оно задано на вкладке Опции модели узла моделирования.

Другие опции совпадают с опциями на вкладке Опции моделирования узла моделирования.

Слепок обобщенной линейной модели Netezza

Слепок модели предоставляет доступ к выходным данным операции моделирования.

При запуске потока, содержащего узел обобщенного линейного моделирования, на этот узел добавляется одно новое поле, имя которого получается из имени модели.

Таблица 28. Поле скоринга для обобщенной линейной модели

Имя добавленного поля	Значение
$GLM-*имя_модели*	Предсказанное значение для текущей записи.

На вкладке Модель выводятся различные статистические данные, относящиеся к модели.

Выходные данные состоят из следующих полей.

Таблица 29. Выходные поля из обобщенной линейной модели

Поле вывода	**Описание**
Параметр	Параметры (то есть предикторные переменные), используемые моделью. Это числовые или номинальные столбцы, а также свободный член (постоянный член в модели регрессии).
Бета	Коэффициент корреляции (то есть линейный компонент модели).
Среднеквадратичная ошибка	Среднеквадратичное отклонение для бета.
Критерий	Статистический критерий, используемый для оценки допустимости параметра.
p-значение	Вероятность ошибки в предположении, что параметр значимый.
Сводка остатков	
Тип остатка	Тип остатка предсказания, для которого показаны сводные значения.
RSS	Значение остатка.
ст.св.	Число степеней свободы остатка.
p-значение	Вероятность ошибки. Большое значение означает плохую подгонку модели; маленькое значение - хорошую подгонку.

Слепок обобщенной линейной модели Netezza - вкладка Параметры

На вкладке Параметры можно настроить выходные данные модели.

Это та же опция, которая показана для опций скоринга на узле моделирования. Дополнительную информацию смотрите в разделе “Опции обобщенной линейной модели Netezza - опции скоринга” на стр. 104.

Уведомления

Эта информация относится к продуктам и сервису, предлагаемым по всему миру.

IBM может не предоставлять в других странах продукты, услуги и аппаратные средства, описанные в данном документе. За информацией о продуктах и услугах, предоставляемых в вашей стране, обращайтесь к местному представителю IBM. Ссылки на продукты, программы или услуги IBM не означают и не предполагают, что можно использовать только указанные продукты, программы или услуги IBM. Разрешается использовать любые функционально эквивалентные продукты, программы или услуги, если при этом не нарушаются права IBM на интеллектуальную собственность. Однако ответственность за оценку и проверку работы любого продукта, программы или сервиса, не произведенного корпорацией IBM, лежит на пользователе.

IBM может располагать патентами или рассматриваемыми заявками на патенты, относящимися к предмету данного документа. Предъявление данного документа не предоставляет какую-либо лицензию на эти патенты. Вы можете послать письменный запрос о лицензии по адресу:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

По поводу лицензий, связанных с использованием наборов двухбайтных символов (DBCS), обращайтесь в отдел интеллектуальной собственности IBM в вашей стране или направьте запрос в письменной форме по адресу:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

Следующий абзац не применяется в Великобритании или в любой другой стране, где подобные заявления противоречат местным законам: INTERNATIONAL BUSINESS MACHINES CORPORATION ПРЕДСТАВЛЯЕТ ДАННУЮ ПУБЛИКАЦИЮ "КАК ЕСТЬ", БЕЗ КАКИХ-ЛИБО ГАРАНТИЙ, КАК ЯВНЫХ, ТАК И ПОДРАЗУМЕВАЕМЫХ, ВКЛЮЧАЯ, НО НЕ ОГРАНИЧИВАЯСЬ ТАКОВЫМИ, ПРЕДПОЛАГАЕМЫЕ ГАРАНТИИ СОБЛЮДЕНИЯ ЧЬИХ-ЛИБО АВТОРСКИХ ПРАВ, ВОЗМОЖНОСТИ КОММЕРЧЕСКОГО ИСПОЛЬЗОВАНИЯ ИЛИ ПРИГОДНОСТИ ДЛЯ КАКИХ-ЛИБО ЦЕЛЕЙ И СООТВЕТСТВИЯ ОПРЕДЕЛЕННОЙ ЦЕЛИ. В некоторых штатах при определенных соглашениях не допускается отказ от выраженных или подразумеваемых гарантий, поэтому данное заявление может к вам не относиться.

Эта информация может содержать технические неточности и типографские ошибки. В представленную здесь информацию периодически вносятся изменения; эти изменения будут включаться в новые издания данной публикации. Фирма IBM может в любое время без уведомления вносить изменения и усовершенствования в продукты и программы, описанные в этой публикации.

Любые приводимые здесь ссылки на web-сайты, не относящиеся к компании IBM, даются исключительно для удобства и ни в коей мере не служат целям поддержки или рекламы этих web-сайтов. Материалы этих Web-сайтов не являются частью данного продукта IBM, и вы можете использовать их только на собственную ответственность.

Любую предоставленную вами информацию IBM может использовать или распространять любым способом, какой сочтет нужным, не беря на себя никаких обязательств по отношению к вам.

Если обладателю лицензии на данную программу понадобятся сведения о возможности: (i) обмена данными между независимо разработанными программами и другими программами (включая данную) и (ii) совместного использования таких данных, он может обратиться по адресу:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Такая информация может быть доступна при соответствующих условиях и соглашениях, включая в некоторых случаях взимание платы.

Описанную в данном документе лицензионную программу и все прилагаемые к ней лицензированные материалы IBM предоставляет на основе положений Соглашения между IBM и Заказчиком, Международного Соглашения о Лицензиях на Программы IBM или любого эквивалентного соглашения между IBM и заказчиком.

Любые данные о выполнении, содержащиеся здесь, были определены в контролируемой среде. Поэтому результаты, полученные в других операционных средах, могут существенно отличаться. Некоторые измерения могли быть сделаны на системах в стадии разработки, и поэтому нет гарантии, что соответствующие показатели останутся теми же на общедоступных системах. Более того, некоторые показатели могли быть оценены путем экстраполяции. Реальные результаты могут отличаться. Пользователи этого документа должны проверить приводимые данные в их конкретной среде.

Информация о продуктах, не принадлежащих компании IBM, была получена от поставщиков этих продуктов, из их опубликованных сообщений или других общедоступных источников. Компания IBM не тестировала эти продукты и не может подтвердить правильность их работы, совместимость и другие утверждения, касающиеся продуктов, не принадлежащих компании IBM. Вопросы о возможностях этих продуктов следует направлять их поставщикам.

Все заявления, касающиеся будущих направлений деятельности или намерений корпорации IBM, подвержены изменению или отмене без предупреждения и являются не более чем выражением целей или намерений.

Эти сведения содержат примеры данных и отчетов, используемых в повседневных деловых операциях. Чтобы проиллюстрировать их настолько полно, насколько это возможно, данные примеры включают имена индивидуумов, названия компаний, брендов и продуктов. Все эти имена и названия являются вымышленными, и любое совпадения с названиями и адресами, используемыми реально действующими компаниями, является чисто случайными.

При просмотре данного электронного информационного документа фотографии и цветные иллюстрации могут не показываться.

Товарные знаки

IBM, логотип IBM, и ibm.com являются товарными знаками или зарегистрированными товарными знаками компании International Business Machines Corp., зарегистрированными во многих странах мира. Прочие наименования продуктов и услуг могут быть товарными знаками, принадлежащими IBM или другим компаниям. Текущий список товарных знаков IBM можно найти в Интернете “Copyright and trademark information” по адресу: www.ibm.com/legal/copytrade.shtml.

Intel, логотип Intel, Intel Inside, логотип Intel Inside, Intel Centrino, логотип Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium и Pentium являются товарными знаками или зарегистрированными товарными знаками компании Intel или ее дочерних компаний в Соединенных Штатах и других странах.

Linux является зарегистрированным товарным знаком Linus Torvalds в Соединенных Штатах и других странах.

Microsoft, Windows, Windows NT и логотип Windows являются товарными знаками корпорации Microsoft в Соединенных Штатах и других странах.

UNIX является зарегистрированным товарным знаком The Open Group в Соединенных Штатах и других странах.

Java и все основанные на Java товарные знаки и логотипы - товарные знаки или зарегистрированные товарные знаки Oracle и/или его филиалов.

Другие названия продуктов и услуг могут являться товарными знаками IBM или других компаний.

Индекс

S

Z

А

Б

В

Г

Д

И

К

Л

М

Н

О

П

Р

С

Т

У

Ф

Ч

Ш

Э

Напечатано в Дании