

ORACLE DATABASE 11G

ОПЦИЯ DATA MINING

КЛЮЧЕВЫЕ ФУНКЦИИ И УСОВЕРШЕНСТВОВАНИЯ

ORACLE DATA MINING

- Аналитика в пределах базы данных
- Модели интеллектуального анализа данных – объекты базы данных первого класса
- Полный набор алгоритмов
- Обнаружение аномалий
- Автоматическая подготовка данных
- Руководства по процессам интеллектуального анализа данных
- Подготовка данных
- Анализ текста
- Программные интерфейсы на Java и PL/SQL
- Генерация кода
- Прогнозирующая аналитика
- Платформа Oracle Database

Опция Data Mining пакета Oracle Database 11g (версия 1) дает предприятиям возможность получать информацию, позволяющую делать точные прогнозы и принимать правильные решения, а также создавать интегрированные приложения для бизнес-аналитики. Благодаря функциям интеллектуального анализа данных, встроенным в Oracle 11g Database, бизнес-аналитики способны обнаруживать закономерности в имеющихся данных и получать углубленное понимание ситуации. Разработчики приложений, в свою очередь, могут быстро автоматизировать обнаружение и распространение новых знаний, представляющих ценность для бизнеса, – прогнозов, закономерностей и открытий – в рамках всей организации.

Аналитика в пределах базы данных

Oracle Data Mining (ODM) предоставляет всеобъемлющие возможности интеллектуального анализа данных, встроенные в СУБД Oracle Database. Благодаря этому исчезает необходимость извлекать данные из базы данных во внешние аналитические системы для выполнения анализа данных. Все функции Oracle Data Mining встроены в Oracle 11g Database. С Oracle Data Mining, данные никогда не покидают базу данных. Сами данные, подготовка данных, построение моделей и действия по оценке моделей – все это остается в пределах Oracle Database. Это также несет с собой значительные преимущества с точки зрения безопасности, масштабируемости, управляемости, разработки приложений и пользовательского доступа.

Встроенные функции интеллектуального анализа данных в Oracle Data Mining в базе данных означает, что в пределах базы данных не только остаются данные, но и выполняются преобразования данных и задачи по интеллектуальному анализу данных. Они могут выполняться автоматически, асинхронно и независимо от каких-либо пользовательских интерфейсов.

Масштабируемость Oracle 11g Database позволяет средству Oracle Data Mining анализировать большие объемы данных с целью обнаружения трудноуловимых закономерностей и зависимостей, а также извлечения новых знаний, скрытых в данных и представляющих ценность для бизнеса. Результаты анализа и прогнозы, получаемые с помощью Oracle Data Mining, хранятся в таблицах базы данных и доступны для обращения из средств и приложений генерации запросов, отчетов и анализа – как работающих на основе Oracle, так и других.

Полный набор алгоритмов интеллектуального анализа данных

Oracle Data Mining поддерживает широкий диапазон методов построения и оценки моделей интеллектуального анализа данных, среди которых классификация, регрессия, кластеризация, ассоциации, обнаружение аномалий, анализ текста, поиск существенных атрибутов и выделение признаков.

Для решения задач классификации поддерживаются следующие алгоритмы: деревья принятия решений, наивные байесовские классификаторы, обобщенные линейные модели (GLM) и метод опорных векторов (SVM).

Кластеризация выполняется с помощью улучшенного алгоритма k-средних (на основе метрики расстояния) либо метода «O-cluster» (на основе плотности).

Для решения проблемы регрессии предлагаются обобщенные линейные модели и метод опорных векторов.

Для анализа текста и обнаружения аномалий применяется метод опорных векторов, для поиска существенных атрибутов – принцип минимальной длины описания (MDL), для ассоциаций – метод A Priori, а для выделения признаков – неотрицательная матричная факторизация (NMF).

Обнаружение аномалий

Обычно для классификации требуется знание всех конечных классов. Версия SVM позволяет построить профиль одного класса и затем при применении отмечать случаи, так или иначе отличающиеся от этого профиля (т.е. «ненормальные» или «подозрительные»). Это позволяет обнаруживать редкие случаи, необязательно связанные друг с другом, выявить которые с помощью классификации практически невозможно.

Автоматическая подготовка данных

В обычном случае те преобразования данных, которых требует модель интеллектуального анализа данных, необходимо выполнять вручную в рамках процедур интеллектуального анализа данных. В Oracle 11g Data Mining все преобразования данных, требуемые для того или иного алгоритма, выполняются автоматически в рамках выполнения модели.

Процессы интеллектуального анализа данных

Применяемые в графическом пользовательском интерфейсе (GUI) средства Oracle Data Miner процессы интеллектуального анализа данных не только предписывают надлежащий порядок операций и выполняют все преобразования данных, требуемые в соответствии с алгоритмами, но также имеют интеллектуальные настройки и варианты оптимизации всех параметров. Эти параметры, тем не менее, можно раскрыть с целью модификации установленных по умолчанию значений.

Процесс построения модели включает в себя оценку модели, когда это необходимо, а также ряд методов тестирования, включая ROC-анализ (Receiver Operating Characteristics – ROC) для классификации и график остатков для регрессии.

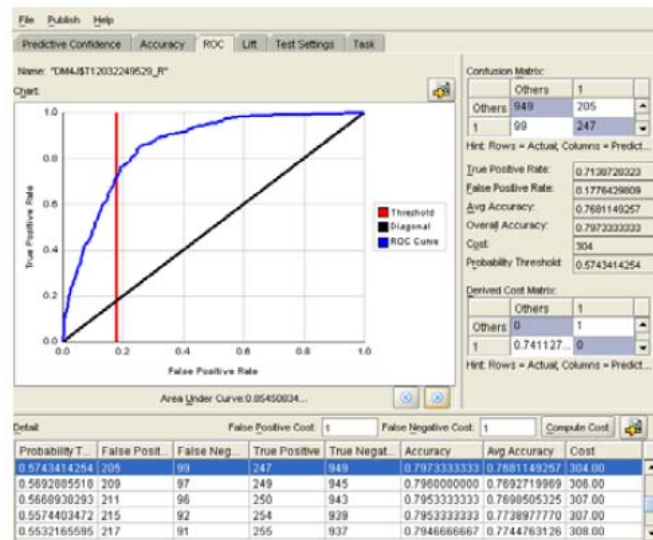


Рисунок 1. ROC-анализ

Процесс построения модели «запоминает» все преобразования данных и значения параметров, поэтому, когда подходит время оценки данных по оптимальной модели, метаданные построения бесшовно передаются в процесс применения модели для автоматического выполнения.

Подготовка данных

Oracle Data Miner способен принимать в качестве входных данных множество таблиц или представлений, а также выполнять соответствующие соединения и преобразования, необходимые для моделирования. ODM умеет интеллектуально анализировать транзакционные данные и вложенные таблицы данных. Благодаря управлению агрегацией, преобразованиями и подготовкой данных в пределах базы данных, развертывание модели и разработка приложений ускоряются.

Анализ текста

Такие алгоритмы, как метод опорных векторов, ассоциативные правила, алгоритм k-средних, а также неотрицательная матричная факторизация, способны принимать в качестве входного атрибута текст (неструктурированные данные). В результате столбец, содержащий, например, записи врача, технический документ или полицейский отчет, может обрабатываться точно так же, как любая другая входная переменная, за счет чего повышается ценность прогнозирующей модели.

Программные интерфейсы на Java и PL/SQL

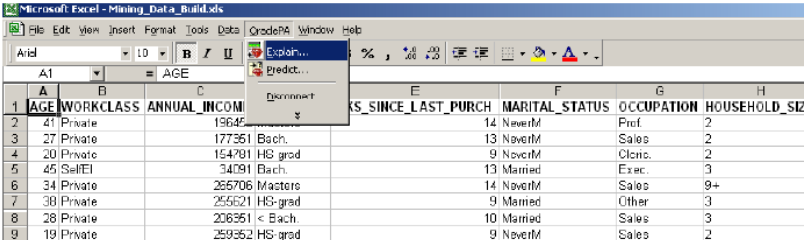
Разработчикам приложений доступны программные интерфейсы Oracle Data Mining на Java и PL/SQL, позволяющие им интегрировать результаты анализа и прогнозы в бизнес-приложения. Более того, построив модель интеллектуального анализа данных с использованием одного программного интерфейса, например PL/SQL, можно затем применять модель с использованием любого из двух интерфейсов. Примеры программ с кодом, необходимым для большинства операций интеллектуального анализа данных, поставляются вместе с СУБД Oracle Database.

Генерация кода

При выполнении в графическом пользовательском интерфейсе Oracle Data Miner операции той или иной процедуры генерируется код на PL/SQL, что позволяет упаковать построение, тестирование или применение прогнозирующей модели для выполнения в другой среде Oracle Database. После получения доступа к коду посредством JDeveloper или SQLDeveloper его можно использовать в создании того или иного приложения. Таким образом, модель, построенная и оптимизированная на одной системе, может быть применена к данным в качестве компонента приложения на другой системе.

Прогнозирующая аналитика

SQL-функции PREDICT, EXPLAIN и PROFILE являются полностью самостоятельными пакетами для построения модели классификации или модели поиска существенных атрибутов. Всем параметрам задаются оптимизированные значения, а промежуточные данные не сохраняются. Результаты – т.е. спрогнозированные оценки (PREDICT), ранжированный список атрибутов (EXPLAIN) либо оценки и правила (PROFILE) – могут использоваться в рамках операционной магистрали либо быть отображены в командной строке или электронной таблице.



	A	B	C	E	F	G	H
	AGE	WORKCLASS	ANNUAL_INCOM	KS SINCE LAST PURCH	MARITAL STATUS	OCCUPATION	HOUSEHOLD SIZE
2	41	Private	19642		14 NeverM	Prof.	2
3	27	Private	177361	Bach.	13 NeverM	Sales	2
4	20	Private	154781	HS-grad	9 NeverM	Exec.	2
5	45	SellFI	34091	Bach.	13 Married	Exec.	3
6	34	Private	265706	Maectere	14 NeverM	Sales	9+
7	30	Private	255621	HS-grad	9 Married	Other	3
8	28	Private	206351	< Bach.	10 Married	Sales	3
9	19	Private	259352	HS-grad	9 NeverM	Sales	2

Рисунок 2. Выполнение функции PREDICT в электронной таблице

Платформа Oracle Database

Благодаря Oracle Data Mining предприятия получают выгоду от полностью интегрированной среды, включающей хранилище Oracle Data Warehouse и бизнес-аналитику. Все функции Oracle Data Mining интегрированы с лидирующей в отрасли платформой Oracle Database, нацеленной на безопасность, масштабируемость и управление данными.

© Oracle, 2007. Все права защищены.

Данный документ предоставляется исключительно в информационных целях. Его содержимое может изменяться без предварительного уведомления. Компания Oracle не гарантирует, что документ не содержит ошибок, а также не предоставляет иных гарантий либо положений, как изложенных в устной форме, так и неявно определяемых законодательством – в том числе неявных гарантий и положений относительно товарного состояния или пригодности для конкретной цели. В частности, компания Oracle не несет никакой ответственности в связи с настоящим документом и заявляет, что настоящий документ не создает каких-либо явных или неявных контрактных обязательств. Настоящий документ запрещается воспроизводить или передавать с какой-либо целью, в какой-либо форме и какими-либо средствами, в том числе электронными и механическими, без предварительного письменного согласия компании Oracle.

Oracle, JD Edwards, PeopleSoft и Siebel являются зарегистрированными товарными знаками корпорации Oracle и/или ее дочерних предприятий. Прочие наименования могут являться товарными знаками соответствующих владельцев.