Инструментальные средства Oracle Data Mining

Ольга Горчинская
Консультант по технологиям
хранилищ данных и OLAP
Oracle Corporation

ORACLE

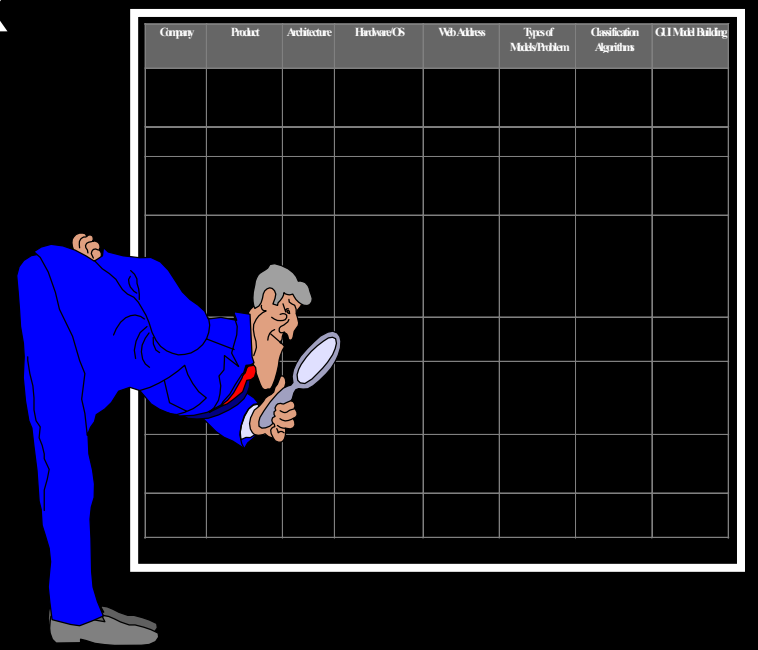
План

- Введение в data mining
- Oracle Data Mining -- общие сведения
- Функциональные возможности
- Технология использования
- Перспективы развития

Введение в Data Mining

Data Mining = Извлечение знаний

- Одна из технологий анализа данных
- Поиск новых общих закономерностей в больших наборах данных
- Использование методов и алгоритмов статистики, распознавания образов, машинного обучения, искусственного интеллекта



Что такое *Data Mining*

Процесс извлечения ранее неизвестной полезной информации из больших баз данных (Knowledge Discovery Process)

Aaron Zornes, META Group

Процесс автоматического выявления новых закономерностей и взаимосвязей в больших наборах данных для использования в процессах принятия решений

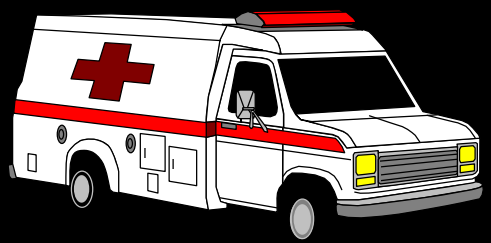
Robert Small, Two Crows

Типичные вопросы

- Какие клиенты обычно отказываются от услуг и переходят в конкурирующую организацию?
- Какие характеристики имеет типичный прибыльный клиент?
- Какие услуги и продукты должны предлагаться различным группам клиентов?
- Какими параметрами характеризуются пользователи пластиковых карточек?
- Какие сочетания параметров говорят о возможных нарушениях?

Области использования

- Банки и финансовые организации
- Телекоммуникация
- Торговля (retail)
- Медицина



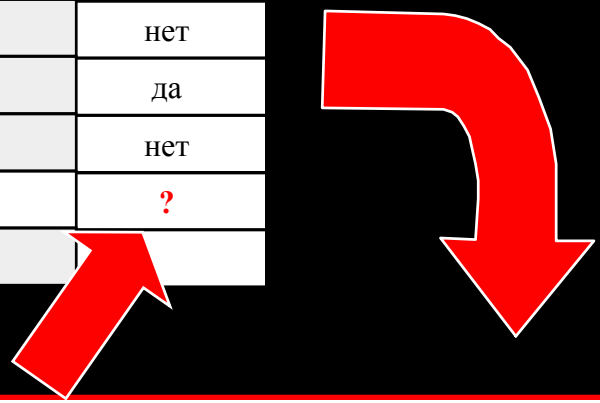
Формальные основы

- Цель извлечения знаний - построение моделей
- Типы моделей :
  - прогнозирующие (predictive);
  - дескриптивные (descriptive)
- Прогнозирующие модели в явном виде содержат информацию для прогноза
- Дескриптивные модели описывают общие закономерности предметной области

Пример построения модели

Информация о клиентах

ИНН	ОКПО	Адрес	Дата рег.	Признак задолжен.	Уставной фонд	….	Надежен ?
077348923	245643560	Москва	23/07/99	да	50000		да
078344864	453453590	С.-Петерб	15/09/01	да	11000		нет
063678454	535643510	Рязань	20/06/99	нет	20000		нет
054778355	563035360	Тула	12/08/00	нет	100000		да
782999634	020432510	С.-Петерб	24/03/01	да	30000		нет
45679/326	**220820100**	**Псков**	**14/10/99**	нет	19000		**?**



Если дата регистрации в Госкомстате России позднее 01.03.2001 и размер уставного фонда менее 20 000 рублей, то клиент - ненадежен с достоверностью 0.92.

Построение моделей

Функциональная зависимость
Y = F(X1, X2, …, Xn)

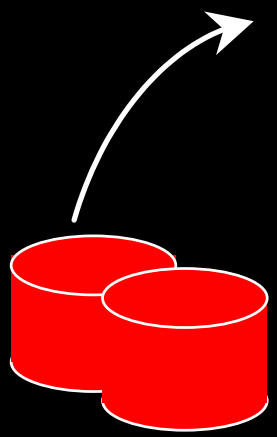
Поля или атрибуты

Name	Income	Age	. . .		Credit Rating 1 = Good, 0 = Bad
Jones	30,000	30			1
Smith	55,000	67			1
Lee	25,000	23			0
Rogers	50,000	44			0

X_1 X_2 ..... X_m Y

Независимые переменные

Зависимая (целевая) переменная



Типы моделей

Прогнозирующие:

- Классификация
- Регрессия
- Дескриптивные:
- Кластерные
- Ассоциативные

Классификационные модели

- На основе характеристик объекта определяется, к какому классу он принадлежит
- Примеры :
  - Подходит ли пациент для данного курса лечения?
  - Выявить ненадежных клиентов
  - Кому следует рассылать новое предложение?
- Для создания модели необходим набор классифицированных случаев

Регрессионные модели

- Прогнозирование новых значений на основе существующих
- В отличие от классификации прогнозируются непрерывные переменные
- В простейшем случае -- стандартные статистические методы (линейная регрессия)

Кластеризация

- Разбиение данных на разные группы (кластеры) по критерию “похожести“ или “близости“
- Основное понятие -- “расстояние“ между различными объектами
- В отличие от классификации признаки и классы, по которому будет проводиться разбиение, заранее неизвестны
- Интерпретация кластеров

Ассоциативные модели

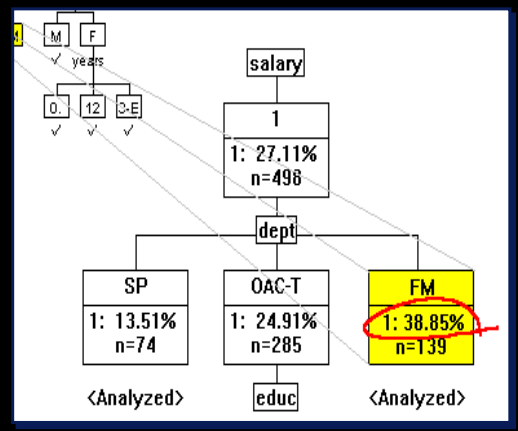
- Правила типа “если…, то…” с использованием коэффициентов уверенности
- “Если клиент покупает молоток, то в 50% случаев он покупает также и гвозди“
- “Если производится импорт товара ***A***, то с вероятностью ***p*** производится импорт товара ***B***”

Алгоритмы извлечения знаний

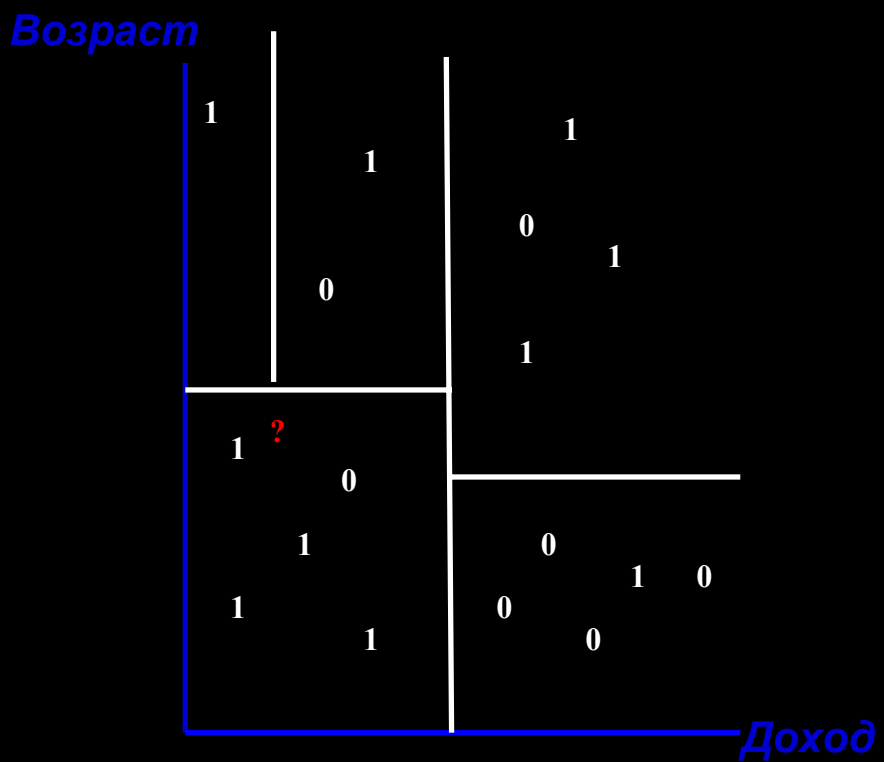
- Для построения моделей используются различные алгоритмы
- Для каждого типа модели -- много алгоритмов (CART, нейронные сети)

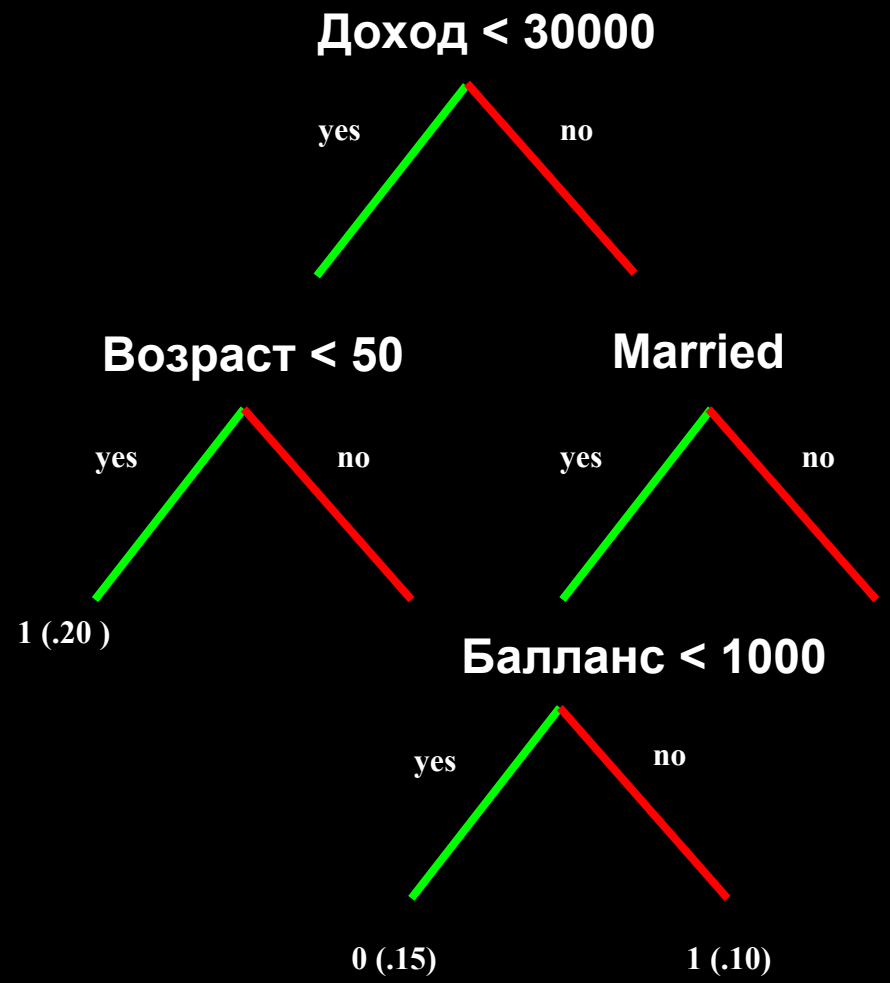
Алгоритмы извлечения знаний

- Классификационные и регрессионные деревья решений
- Нейронные сети
- Вывод правил (Rule induction)
- K-окрестности (Memory based reasoning)
- Генетические алгоритмы
- Поиск ассоциаций
- Поиск ассоциативных последовательностей
- Дискриминантный анализ
- Обобщенные аддитивные модели
- . . .



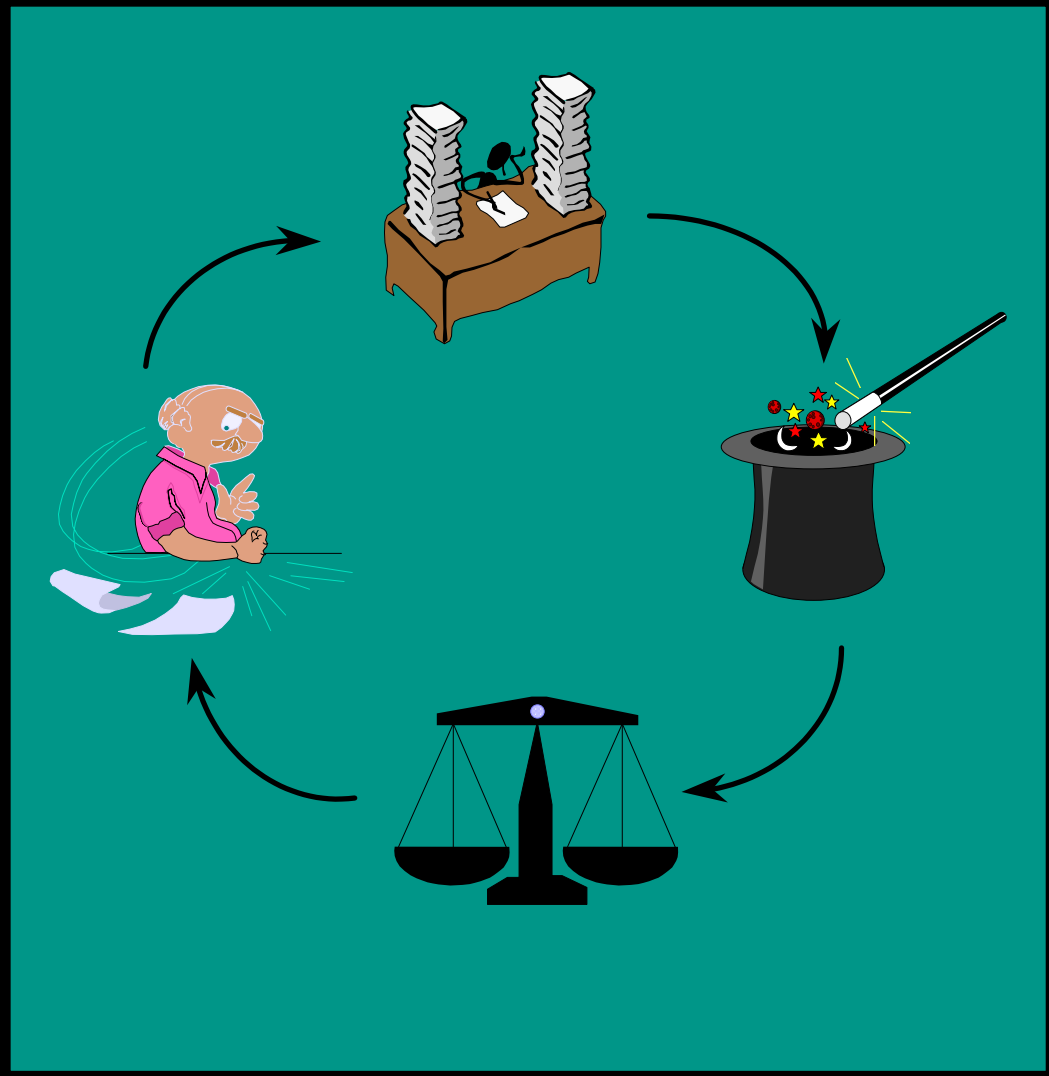
Деревья решений





Технология извлечения знаний

- Постановка задачи
- Подготовка данных
- Построение модели
- Оценка и интерпретация
- Тестирование



Oracle Data Mining -- общие сведения

Oracle & Data mining

- Март, 1998: объявление о совместной деятельности с 7 партнерами -- поставщиками средств извлечения знаний
- Включение в Oracle8i средств поддержки алгоритмов извлечения знаний
- Июнь 1999: Oracle приобретает Darwin (Thinking Machines Corp.)
- 2000 -2001: новые версии Darwin, Oracle Data Mining Suite
- Июнь 2001: Oracle9i Data Mining

Oracle Data Mining

- Встроенные в Oracle Database алгоритмы извлечения знаний (Data Mining Server)
- DM инфраструктура вместо готовой инструментальной среды
- API для разработки

Встроенные средства извлечения знаний

- Упрощение процесса извлечения знаний
- Устранение дополнительного перемещения и хранения данных
- Производительность и масштабируемость

DM инструментальное средства и DM инфраструктура

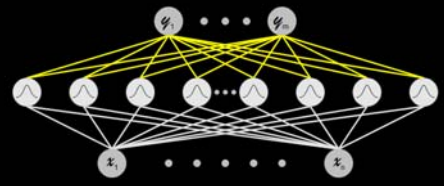
- Инструментальное средства - ad hoc data mining
  - Ориентация на специалистов-аналитиков
  - Дополнительные затраты на извлечение и подготовку данных
- Инфраструктура - Enhance any application
  - Более широкие возможности использования
  - Все данные и процессы – непосредственно в БД

Oracle Data Mining API

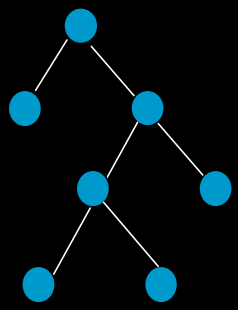
- Использование на уровне программирования
- Java API для разработки на Java
  - основан на принципах JDM (новый стандарт для data mining)
  - будет полностью поддерживать стандарт JDM
- PL/SQL интерфейс:
  - DBMS_DATA_MINING
  - DBMS_MINING_TRANSFORM

Oracle Data Mining -- функциональные возможности

Функции и алгоритмы



- Прогнозирующие модели
  - Классификация (Naïve Bayes, Adaptive Bayes Network, Support Vector Machine *)
  - Регрессия (Support Vector Machine *)
  - Поиск существенных атрибутов (Minimal Descriptor Length)
- Дескрипторные модели
  - Кластеризация (enhanced K-means, O-cluster)
  - Поиск ассоциаций (Apriori Algorithm)
  - Выделение признаков (Feature extraction) (Non-Negative Matrix Factorization *)



* -- *новые алгоритмы Oracle 10g*

Алгоритмы классификации

- **Naïve Bayes** (NB)
  - Быстрее чем ABN (по времени построения модели)
  - Лучше работает для числа атрибутов < 200
  - Точность меньше, чем в ABN
- **Adaptive Bayes Network** (ANB)
  - Для большого числа атрибутов
  - Наглядность модели (генерация правил)
  - Более точные модели, чем в NB
  - Больше параметров настройки
- Support Vector Machine

Регрессия

- Прогнозирование непрерывных величин
- Простейший случай – линейная регрессия
- Support Vector Machine

Поиск существенных атрибутов

- Выявление атрибутов, наиболее важных для прогнозирования целевых значений
- Используется для ускорения процесса построения классификационной модели
- Алгоритм Minimum Descriptor Length (MDL)

Алгоритмы кластеризации

- Enhanced k-means Clustering
  - Число кластеров задается пользователем
  - Только числовые атрибуты
  - Не очень большое число атрибутов
  - Любое число строк
- O-Cluster
  - Автоматически определяется число кластеров
  - Числовые и категорные атрибуты
  - Большое число атрибутов (> 10)
  - Большое число строк (>1000)

Выделение признаков

- Выделение признаков (Feature Extraction) – формирование набора новых атрибутов (свойств), которые
  - достаточно полно и точно представляют исходный набор данных
  - снижают размерность описания данных
  - выявляют важные взаимосвязи в данных

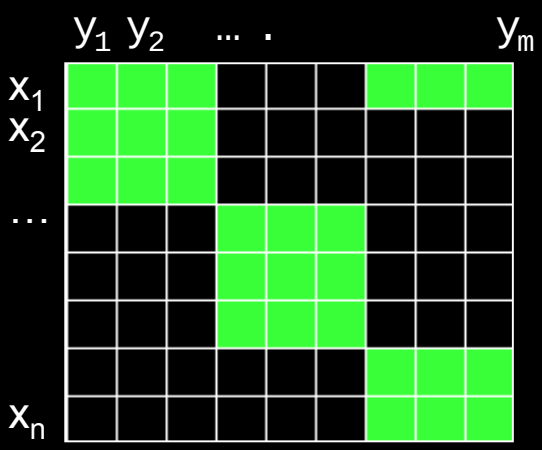
Алгоритм NMF: Non-Negative Matrix Factorization

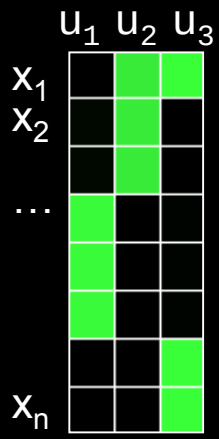
- Пусть D – матрица размерностью M x N и $V = D^T$. Алгоритм NMF итеративно вычисляет апроксимацию в виде произведения двух положительных матриц W and H меньшей размерности.

$$V \sim W \times H$$

Декомпозиция матрицы

V (n x m) ~ W (n x k) x H (k x m)





Базисная модель

Кодирующие коэффициенты

V, W and H > 0

k < n,m

$$\begin{cases} V_{.a} \sim W \times H_{.a} \\ V_{a.} \sim W_{a.} \times H \\ V \sim u_1 \wedge v_1 + \ldots + u_3 \wedge v_3 \end{cases}$$

Базисный вектор (столбцы W)

Кодирующий вектор (строки H)

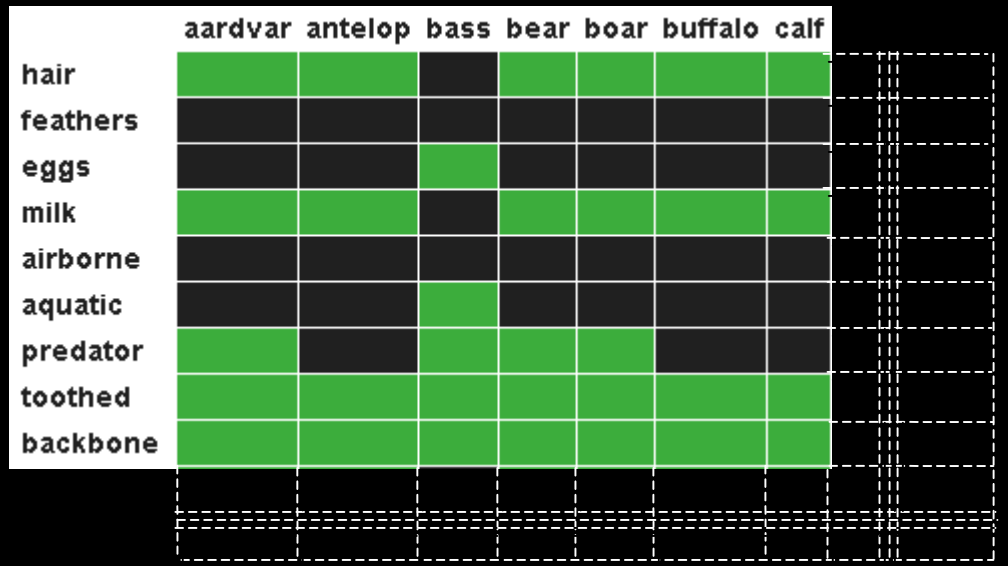
Пример применения

Животные (101) = aardvark, antelope, bear, boar, buffalo, calf, cavy, cheetah, deer, dolphin, elephant, fruitbat, giraffe, girl, goat, gorilla, hamster, hare, leopard, lion, lynx, mink, mole, mongoose, opossum, oryx, platypus, polecat, pony, porpoise, puma, pussycat, raccoon, reindeer, seal, sealion, squirrel, vampire, vole, wallaby,wolf, chicken, crow, dove, duck, flamingo, gull, hawk, kiwi, lark, ostrich, parakeet, penguin, pheasant, rhea, skimmer, skua, sparrow, swan, vulture, wren, pitviper, seasnake, slowworm, tortoise, tuatara, bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse, sole, stingray, tuna, frog, frog, newt, toad, flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp, clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm

17 свойств

- Hair
- Feathers
- Eggs
- Milk
- Airborne
- Aquatic
- Predator
- Toothed
- Backbone
- Breathes
- Venomous
- Fins
- Legs
- Tail
- Domestic
- Catsize
- Type

Исходная матрица V



Матрица W -- выявленные признаки

Discovering Animal Groups

17 характеристик представляются 3 факторами:

Признак 0: описывает понятие «подобные рыбам» (**fish-like)** , в основном состоящее из *fins, predator aquatic, eggs, tail, toothed* and *backbone.*

Признак 1: описывает понятие «подобные птицам» (**bird-like),** в основном состоящее из *feathers, airborne, eggs, breathes, tail and backbone.*

Признак 2: описывает понятие «подобные млекопитающим» (**mammal-like),** в основном состоящее из *legs*, *tail, backbone, breathes, milk, hair* and *toothed*.

Технология работы

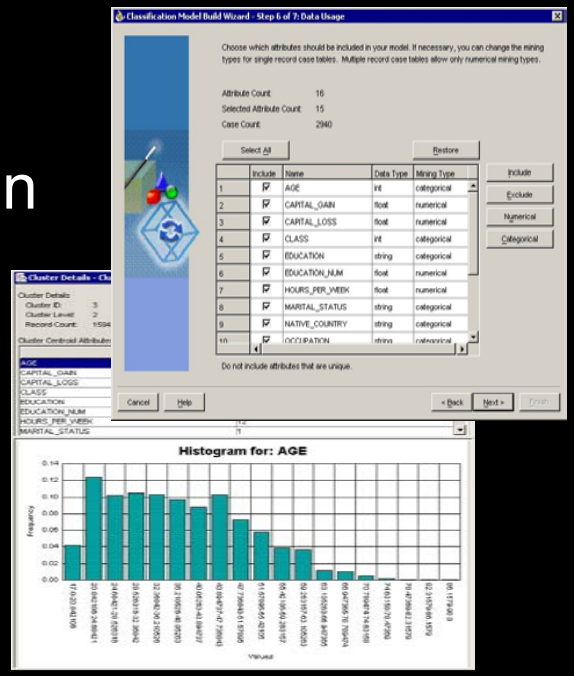
- Создание Java (PL/SQL) модулей для:
  - Идентификации и подготовки данных
  - Построения и тестирования моделей для выбора наилучшей
  - Применение модели к новым данным
- Встраивание Java (PL/SQL) кода в приложение для автоматического использования результатов

Графический интерфейс (DM4J, Data Mining Client)

- Расширения Jdeveloper для Data Mining - приложений (DM4J)
- Data Mining Client
- Генерация Java программ для построения моделей, тестирования, применения

Oracle 10g – новые возможности

- Новые алгоритмы
  - Support Vector Machines
  - Nonnegative Matrix Factorization
- Data Mining для неструктурированной информации
- PL/SQL API (дополнительно к JAVA API)
- Графический интерфейс: ODM Client



Дополнительная информация

- Общая информация

  www.oracle.com
- Техническая информация:
  Oracle Technology Network (OTN)

  http://otn.oracle.com/products/bi/odm/in
  - свободно распространяемое ПО
  - документация, форумы, рекомендации, статьи, презентации