Oracle Data Mining for Real-time Analytics
NYOUG
Sep 21, 2006
Shyam Varan Nath
Oracle Corporation
Agenda

- Introduction to Data Mining
- What is Real-time Analytics?
- Overview of Oracle Data Mining (ODM)
- Real-time Analytical Applications
- ODM demo
About me…

• Principal Analytical Consultant, Business Intelligence and Warehousing group at Oracle

• Oracle Certified Professional, DBA track, since 1998 – Oracle database version 7.3 onwards

• Worked in industry domain like Finance industry, Telecomm, Healthcare, Crime Detection project etc.

• Speaker in Oracle Open world (2003), IOUG/Collaborate (2005-06), NYOUG (June), IEEE conferences etc on Data Mining
OVERVIEW

Data Mining
What is Data Mining?

Simply stated Data Mining refers to extracting or “mining” knowledge from large amount of data. The term is actually a misnomer, data mining should be more appropriately named “knowledge mining from the data”. Thus such a misnomer that carries both “data” and “mining” became popular choice. The larger meaning encapsulates a variety of techniques and methodologies that among other things include ‘cluster analysis’, ‘classification’, ‘association rules’: pattern recognition in the data.

- J. Han and M. Kamber
In other words…

• Process of sifting through massive amounts of data to find *hidden* patterns and discover *new* insights

• Data Mining can provide valuable results:
  - Identify factors more associated with a target attribute (*Attribute Importance*)
  - Predict individual behavior (*Classification*)
  - Find profiles of targeted people or items (*Decision Trees*)
  - Segment a population (*Clustering*)
  - Determine important relationships with the population (*Associations*)
  - Find fraud or rare “events” (*Anomaly Detection*)
The Amazon Example (Association Rules)

Your Recent History

Recently Viewed Products
- Introduction to Data Mining, (First Edition) by Pang-Ning Tan

Recently Viewed Categories
- Amazon.com

Recent Searches
- data mining

Customers who bought items in your Recent History also bought:

- Data Mining by Jiawei Han
- Data Mining by Ian H. Witten
Related Fields

Machine Learning

Visualization

Data Mining and Knowledge Discovery

Statistics

Databases
Statistics, Machine Learning and Data Mining

- **Statistics:**
  - more theory-based
  - more focused on testing hypotheses
- **Machine learning**
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- **Data Mining and Knowledge Discovery**
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
- **Distinctions are fuzzy**
## Business Intelligence

### Query and Reporting

- **Extraction of detailed and roll up data**
- **“Information”**

### OLAP

- Summaries, trends and forecasts
- **“Analysis”**

### Data Mining

- Knowledge discovery of hidden patterns
- **Who will re-consolidate loans in the next 6 months and why?**
- **“Insights & Prediction”**

<table>
<thead>
<tr>
<th>How is the interest rate changing in last 2 years?</th>
<th>What is the average <strong>home price</strong> of the condos, by region, by year of construction?</th>
<th>“Insights &amp; Prediction”</th>
</tr>
</thead>
</table>
Results of Data Mining Include:

- Forecasting what may happen in the future
- Classifying people or things into groups by recognizing patterns
- Clustering people or things into groups based on their attributes
- Associating what events are likely to occur together
- Sequencing what events are likely to lead to later events
Data mining is not

• Brute-force crunching of bulk data
• “Blind” application of algorithms
• Going to find relationships where none exist
• Presenting data in different ways
• A database intensive task
• A difficult to understand technology requiring an advanced degree in computer science
Data Mining Is

- A hot buzzword for a class of techniques that find patterns in data
- A user-centric, interactive process which leverages analysis technologies and computing power
- A group of techniques that find relationships that have not previously been discovered
- A relatively easy task that requires knowledge of the business problem/subject matter expertise
### Other Industry Examples

<table>
<thead>
<tr>
<th>Financial Services</th>
<th>Database Marketing</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Combat attrition (churn)</td>
<td>- Buy product x</td>
</tr>
<tr>
<td>- Fraud detection</td>
<td>- More targeted &amp; successful campaigns</td>
</tr>
<tr>
<td>- Loan default</td>
<td>- Identify cross-sell &amp; up-sell opportunities</td>
</tr>
<tr>
<td>- Identify selling opportunities</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Telecommunications</th>
<th>Insurance, Government</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Identify customers likely to leave</td>
<td>- Flag accounting anomalies (Sarbanes-Oxley)</td>
</tr>
<tr>
<td>Target highest lifetime value customers</td>
<td>- Reduce cost of investigating suspicious activity or false claims</td>
</tr>
<tr>
<td>- Identify cross-sell opportunities</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Retail</th>
<th>Life Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Loyalty programs</td>
<td>- Find factors associated with healthy or unhealthy patients</td>
</tr>
<tr>
<td>- Cross-sell</td>
<td>- Discover gene and protein targets</td>
</tr>
<tr>
<td>- Market-basket analysis</td>
<td>- Identify leads for new drugs</td>
</tr>
<tr>
<td>- Fraud detection</td>
<td></td>
</tr>
</tbody>
</table>
What is Real-time Analytics?

- Real-time analytics is the use of, or the capacity to use, all available enterprise data and resources when they are needed. It consists of dynamic analysis, drawing inferences and reporting, based on data entered into a system up to the actual time of use.

- Real-time analytics is also known as real-time data analytics, real-time data integration, and real-time business intelligence.
Examples of Real Time Analytics

- Real-time analytics can be used in CRM (customer relations management) analytics, which includes all programming that analyzes data about an enterprise's customers and presents it so that better and quicker business decisions can be made.

- Another application is in scientific analysis such as the tracking of a hurricane's path, intensity, and wind field, with the intent of predicting these parameters hours or days in advance.
Data Mining Techniques (10g, 10gR2)

- Unsupervised (clustering)
  - K-means
  - O-Cluster
  - Non-Negative Matrix Factorization
  - Anomaly Detection
  - Hierarchical
  - Self Organizing maps
  - Expectation Maximization
  - Many More (pca, mds)

- Supervised (classification and prediction)
  - Adaptive Bayes Network (ABN)
  - Naïve Bayes
  - Support Vector Machine
  - Decision Trees
  - K – Nearest Neighbor
  - Neural Networks
  - Linear Discrimination (PL/SQL code generator)
Attribute Importance
- Identify most influential attributes for a target attribute
  - Factors associated with high costs, responding to an offer, etc.

Classification and Prediction
- Predict customers most likely to:
  - Respond to a campaign or offer
  - Incur the highest costs
- Target your best customers
- Develop customer profiles

Regression
- Predict a numeric value
  - Predict a purchase amount or cost
  - Predict the value of a home
Clustering
- Find naturally occurring groups
  - Market segmentation
  - Find disease subgroups
  - Distinguish normal from non-normal behavior

Association Rules
- Find co-occurring items in a market basket
  - Suggest product combinations
  - Design better item placement on shelves

Feature Extraction
- Reduce a large dataset into representative new attributes
  - Useful for clustering and text mining
Text Mining

- Combine data and text for better models
  - Add unstructured text e.g. physician’s notes to structured data e.g. age, weight, height, etc., to predict outcomes
- Classify and cluster documents
  - Combined with Oracle Text to develop advanced text mining applications e.g. Medline

BLAST

- Sequence matching and alignment
  - Find genes and proteins that are “similar”
Problem: Find customers likely to buy a Buick and their profiles

IF (Age > 55 AND Owns foreign car=no AND Income > 100K...) THEN P(Buy Buick=1) = .77 Support = 250
Oracle Data Mining 10g R2
Anomaly Detection

• “One-Class” SVM Models
  - Fraud, noncompliance
  - Outlier detection
  - Network intrusion detection
  - Disease outbreaks
  - Rare events, true novelty

Problem: Detect rare cases
PL/SQL Code Generator
PL/SQL code sample

/* pl/sql code gen1 */

CREATE PACKAGE "DATAMININGACTIVITY1" AUTHID DEFINER AS
PROCEDURE "MINING_BUILD_TEST"(case_table IN VARCHAR2 DEFAULT "DMUSER1"."MINING_BUILD_TEXT",
additional_table_1 IN VARCHAR2 DEFAULT NULL,
model_name IN VARCHAR2 DEFAULT 'MINING_BUILD_75202_DT',
test_metric_name IN VARCHAR2 DEFAULT ""DM4J$MINING_TEST"",
END;
/
/* pl/sql code gen1 */

CREATE PACKAGE BODY "DATAMININGACTIVITY1" AS
c_long_sql_statement_length CONSTANT INTEGER := 32767;

......
Oracle Data Mining

- Oracle mining platform
  - PL/SQL API
  - Java API (JSR-73)
  - Oracle Data Miner (gui)
  - Spreadsheet Add-In

- Range of algorithms
  - Structured & unstructured data
  - Attribute importance
  - Classification, regression & prediction
  - Anomaly detection
  - Association rules
  - Clustering
  - Nonnegative matrix factorization
  - BLAST
Oracle Spreadsheet Add-In for Predictive Analytics
Oracle Data Miner Support for Text

- ODMr allows one column of the input table for a mining activity to be a text column. You can mine tables with two or more text columns using the ODM programmatic interfaces. If you have text columns only, you can use ODM or Oracle Text.

- Data Preparation for Text Columns - Any text columns must be properly prepared:
  - For the Oracle Data Mining programmatic interfaces, text columns must be converted to nested columns.
  - For an Oracle Data Miner Activity, the text column must be indexed. Oracle Data Miner automatically creates the index during the mining activity.

- Data | Transform | Text lets you prepare a text column for use with the Oracle Data Mining PL/SQL interface. The same transform allows you to do directly the processing done internally by the mining activity.
ODMr Restrictions on Text Mining

- ODMr does not support all of text mining functionality provided by Oracle Data Mining. The following restrictions apply to text mining using ODMr:
- You can include zero or one text columns in a mining operation. If you need to build a model with two or more text columns, you must use one of the ODM programmatic interfaces.
- The mining type of the text column must be text; if the data type of the column is VARCHAR2 or CHAR, you will have to change the mining type from categorical to text.
- The case ID column must be numerical.
- You must have Oracle Text installed with appropriate permissions
Indexing & PL/SQL (Oracle Text)

- Three index types
  - context for traditional information retrieval
  - ctxcat for catalogs
  - ctxrule for classification and/or routing

- Extensions to SQL
  - select ... from ... where contains ...
  - select ... from ... where catsearch ...
  - select ... from ... where matches ...

- Large set of PL/SQL packages
The Inverted Index

document table

<table>
<thead>
<tr>
<th>ID</th>
<th>DOCUMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the cat sat on the mat</td>
</tr>
<tr>
<td>2</td>
<td>the cat in the hat</td>
</tr>
<tr>
<td>3</td>
<td>the hat mat</td>
</tr>
</tbody>
</table>

1 the cat sat on the mat
2 the cat in the hat
3 the hat mat

tokenize

discard stop words

aggregate across documents

<table>
<thead>
<tr>
<th>TOKEN</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAT</td>
<td>DOC 1 POS 2</td>
</tr>
<tr>
<td>SAT</td>
<td>DOC 1 POS 3</td>
</tr>
<tr>
<td>MAT</td>
<td>DOC 1 POS 6</td>
</tr>
<tr>
<td>HAT</td>
<td>DOC 2 POS 5</td>
</tr>
</tbody>
</table>
Text Retrieval

create table foo (text varchar2(80));
insert into foo values ('the cat sat on the mat');
create index foox on foo(text) indextype is ctxsys.context;

select * from foo where contains(text, 'cat')>0;

TEXT
----------------------the cat sat on the mat
Real-Time Applications based on Data Mining

- Why do we need Real time decisions?
- Recent industry changes like “DO NOT CALL LIST”, TIVO, spam blockers, aversion to junk post mail etc
- Importance of “In-Bound Opportunity”
- Finance-banking industry example
Data Mining Flow

Oracle Data Mining

Extract → Clean → Transform

Model → Evaluate → Deploy

Predict → Production Data

Source Data → Mining Table Repository (MTR)
Model Scoring Approach

Gartner's Scoring Approach Classification
Continuum of Scoring Options

Matching the approach to Business requirement

<table>
<thead>
<tr>
<th>Scoring Approach</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch</td>
<td>Most common approach. Can be scheduled, and the results can be shared among different applications. Disadvantages can be incorrect scores due to time lags, amount of time and storage needed to score numerous models with many predictors. In addition, the number of cases affects scoring.</td>
</tr>
<tr>
<td>Real Time</td>
<td>Scoring allows for updates on recently collected data. Especially useful when little new data is collected.</td>
</tr>
<tr>
<td>Dynamic</td>
<td>Scoring allows data to be integrated from current interaction. Useful if there is a significant amount of missing information that is important for the model.</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>Used to gain insight into a new or anonymous customer for whom a previous history or profile is unavailable.</td>
</tr>
</tbody>
</table>
Industry Specific Cases/Applications of Data Mining
The “in-bound” window of opportunity

- In-bound calls present a great opportunity for up-sell and cross sell
- Data mining scoring for recommendations in Real time can provide CSR’s with most likely loan product to offer
- Such an application was built using Oracle Data Mining for finance industry
Intelligent Offer Generation and Retention Management Application

- Integrated User Interface for real time analytical application based on data mining (from Finance Industry)
Demo Scenario: Business Challenges

About National Bank
- Fictional financial services provider
- Customer base: 5 million
- Assets: $69 billion
- Revenue: $4.6 billion
- Large volume Siebel Call Center

Business Challenges
- High customer turnover rate of 14% per year
- Associated replacement cost in millions per year
- Average cost of new customer acquisition: $250
- Currently 2 products per customer, goal of achieving 4 per customer
Demo Scenario: Call Center Solution

- Predict in **real-time** customers **propensity to attrite** and to respond to various retention treatments
- Offer **relevant and timely retention offers** such as free online bill payment only to customers most likely to leave
- Predict in **real-time** customers **propensity to respond** to various cross- & up sell offers
- Target customers with **relevant and timely cross sell offers** at the time of call

**instead of**

Running **costly, less relevant and less timely** outbound retention and cross- & up-sell campaigns
Call Scenario A: Intelligent Cross-Sell

Profile of caller (Linda Johnson):

- Female, 28 years old, single
- Holds checking and savings account at National Bank
- Medium-value customer
- Calls to change address (due to new job after grad school)

Objectives of National Bank:

- Expand customer relationship through real-time intelligent cross- and up-sell offers
Call Scenario A: Upon caller identification, ...

... Linda Johnson is recognized as a student holding two accounts with National.
In addition, upon caller identification, …

... based on Linda’s customer profile, the Decision Server predicts that Linda currently has no significant risk of churning, and therefore no retention treatment is warranted, and …

... that “Auto Insurance” is the marketing offer that is most likely to be accepted by Linda.
Upon noting Linda’s call reason, change of address, ...

... this new in-context information is communicated to the Decision Server in real-time.
Based on the new in-context information, ...

... the Decision Server predicts in real-time that Linda’s churn risk has not increased ...

... but that the most appropriate offer now is “Overdraft Protection”, addressing Linda’s likely increased financial needs.
Linda’s response to the extended offer ...

... is noted by the agent using the offer response buttons. The response information is communicated to the Decision Server in real-time for self-learning.
Linda’s response to the extended offer ...

... is also recorded in the database, enabling offer response tracking and cross channel Marketing Analytics reporting.
Marketing users can analyze offer response behavior ... by viewing the “ideal” profile of customers that are most likely to accept an offer. The ideal profile is useful for evaluating whether an offer appeals to the intended target audience or not, and serves as intelligence for driving outbound marketing activities.
More in-depth reporting on key drivers
Trends in customer response behavior
Overall offer response statistics
Call Scenario B: Real-time Retention

Profile of Caller (Robert Knowles):
- Male, 38 years old, married, homeowner
- Holds several accounts at National Bank
- High-value customer
- Considers closing all accounts (unknown to National Bank)
- Calls to inquire about checking account fees

Objectives of National Bank:
- Retain customer relationship through real-time retention treatment
Call Scenario B: Upon caller identification, ...

... Robert Knowles is recognized as a high-value customer holding several accounts at National.
In addition, upon caller identification, …

... based on Robert’s customer profile, the Decision Server predicts that Robert currently has a low risk of churning, and therefore no retention treatment is warranted ...

... and that “Home Equity” is the offer most likely to be accepted by Robert and ...
Upon noting Robert’s call reason, fees inquiry, …

... this new in-context information is communicated to the Decision Server in real-time.
Based on the new in-context information, the Decision Server predicts in real-time that Robert’s attrition risk is at a critical level, and that the most appropriate retention treatment is an offer for an one-time fee waiver.
Robert’s response to the extended offer... is noted by the agent using the offer response buttons. The response information is communicated to the Decision Server in real-time for self-learning.
Initial offer recommendation after identifying Linda’s call as a service-to-sales opportunity.
Revised offer recommendation in account of call reason (Order Checks)
Automatically created response record, enabling creation of opportunity (call back)
Profile of client who is likely to accept “Web Bill Pay”. Serves as insight for outbound marketing activities.
Performance report across all offers

<table>
<thead>
<tr>
<th>Intelligent Offers</th>
<th>Outcome</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beyond Banking</td>
<td>Presented</td>
<td>2475</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Interested</td>
<td>1036</td>
<td>42%</td>
</tr>
<tr>
<td>Merill+ Credit Card</td>
<td>Presented</td>
<td>2503</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Interested</td>
<td>770</td>
<td>31%</td>
</tr>
<tr>
<td>Visa Signature</td>
<td>Presented</td>
<td>2510</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Interested</td>
<td>785</td>
<td>31%</td>
</tr>
<tr>
<td>Web Bill Pay</td>
<td>Presented</td>
<td>2506</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Interested</td>
<td>1507</td>
<td>60%</td>
</tr>
</tbody>
</table>

**Total Counts**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presented</td>
<td>10000</td>
<td>160%</td>
</tr>
<tr>
<td>Interested</td>
<td>4096</td>
<td>41%</td>
</tr>
</tbody>
</table>

**Pareto of Counts for Interested**

[Graph showing Pareto analysis for interested outcomes]
More detailed reporting on response behavior for “Web Bill Pay”. Serves as insight for outbound marketing activities.
Trend reports on response behavior for "Web Bill Pay". Serves as insight for outbound marketing activities.
Initial attrition risk after identifying Robert’s call as a service-to-sales opportunity
Revised attrition risk in account of call reason (Fees Inquiry)
Automatically created response record, enabling creation of follow up opportunity.
Real-Time Decision Framework

Business Performance Goals
- Increase Customer Revenue
- Reduce Churn

Business Rules
- High Value Customer = Annual Revenue > $100 AND Customer Tenure > 36 Months

Predictive Analytics

<table>
<thead>
<tr>
<th>Alternatives</th>
<th>Likely Success</th>
<th>Impact On Retention</th>
<th>Impact On Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold Card</td>
<td>7%</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Miles Card</td>
<td>17%</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>Platinum Card</td>
<td>3%</td>
<td>+</td>
<td>++</td>
</tr>
</tbody>
</table>
Installing Oracle Data Mining
Architecture for Oracle Data Mining

- Installing Oracle Data Mining
- How to optimize the DB configuration?
- How to understand the schemas necessary to support Data Mining?
Installation Test

- Check to see if the product exists
Install Option

- Default install list for Oracle 9i and 10g RDBMS include ODM
- It’s a separate license, free for development, fee for production
- If ODM not installed, can be added by custom install using Oracle Universal Installer (OUI)
- Companion disk has samples programs for ODM
- ODMr the GUI
- ODM website
Add ODM
ODM...
Installing ODM GUI

- The GUI is called ODMr
- Uses JDBC connection to the database
- Can be used for most data mining tasks such as importing data from flat files, running models, scoring, visually viewing the results etc.
How to optimize the DB configuration
Sizing the DB

- Depends on use of Data warehouse or Data Mart as the ODM schema
- Size of the data to be used for mining
- Batch mode v/s real time response (analogy-optimize DB for total query run time v/s first response)
Config.ora Parameters

- If DB is used for other purposes, parallel DML can be disabled for the Data Mining sessions using the PL/SQL API’s
  - `alter session disable parallel query;`
  - `alter session disable parallel dml;`
  - `alter session disable parallel ddl;`

- Some of the data mining sessions by-pass the underlying table’s parallel settings

- Runs on RAC and Grid seamlessly
Creating mining users

- `CREATE TABLESPACE "ODMPERM" DATAFILE 'C:\ORACLE\PRODUCT\10.2.0\ORADATA\ORCL\odm1.dbf' SIZE 20M REUSE AUTOEXTEND ON NEXT 20M;`

- The next SQL command creates a new temporary tablespace.

  `CREATE TEMPORARY TABLESPACE "ODMTEMP" TEMPFILE 'C:\ORACLE\PRODUCT\10.2.0\ORADATA\ORCL\odmtemp.tmp' SIZE 20M REUSE AUTOEXTEND ON NEXT 20M;`
DB Privileges to the DM user

- **Access Rights:** Data mining users require several `CREATE` privileges. For text mining, users must also have access to the Oracle Text package `ctxsys.ctx_ddl`. The following privileges are required.
  - `CREATE PROCEDURE`
  - `CREATE SESSION`
  - `CREATE TABLE`
  - `CREATE SEQUENCE`
  - `CREATE VIEW`
  - `CREATE JOB`
  - `CREATE TYPE`
  - `CREATE SYNONYM`
  - `EXECUTE ON ctxsys.ctx_ddl`
SQL for Create user

- CREATE USER dmuser1 IDENTIFIED BY change_now DEFAULT TABLESPACE odmperm TEMPORARY TABLESPACE odmtemp QUOTA UNLIMITED on odmperm;

- SQL for Grants
  - GRANT create procedure to DMUSER1;
  - GRANT create session to DMUSER1;
  - GRANT create table to DMUSER1;
  - GRANT create sequence to DMUSER1; ...

- Export / import data mining models
  - SQL> EXECUTE DBMS_DATA_MINING.EXPORT_MODEL('allmodels.dmp','DMTEST');
Contact information:
Shyam.Nath@Oracle.com
(954) 609 2402 cell

DW & BI Special Interest Group
Oracle Business Intelligence

“Predictive Insights Reports”

Response Analysis Report created via Oracle Data Mining and ability to query to details
Oracle Business Intelligence

“Predictive Insights Reports”

Users can query to details to gain insights from “mined data”