Data-Centric Automated Data Mining

Marcos M. Campos
Peter J. Stengard
Boriana L. Milenova

Data Mining Technologies

ORACLE®
Overview

- Data mining complexity
- Proposed design solution
- Application I: Oracle Predictive Analytics
- Application II: Spreadsheet Add-In for PA
Overview

- Data mining complexity
- Proposed design solution
- Application I: Oracle Predictive Analytics
- Application II: Spreadsheet Add-In for PA
Data Mining Complexity

- Knowledge of data mining techniques
  - Which algorithm do I use?
- Algorithm specific data preparation
  - How should I prepare my data?
- Model parameter tuning
  - What kernel function should I use?
- Deployment
  - I’ve deployed a model, now what data can I score with it?
Industry Future

“Predictive analytics builds on the data mining multistep process and statistical modeling techniques to add a layer of automation and self-directed built-in intelligence. Business users (and not just Ph.D. statisticians) can now analyze large amounts of customer, supplier, employee and product data for patterns and trends.”

-Kent Bauer
Design Approach

• Goal: “Good results with minimum effort”

• Data-centric focus
  – familiar to database and business intelligence communities

• Process automation
  – ease-of-use for non-expert users
Data-centric Design

- Eliminates concepts of models or complex methodologies
- Requires only knowledge of the data source
- Supporting objects are either removed or linked to the data source
- Users see only predictive or descriptive results
- Goal-oriented tasks
Goal-oriented Tasks

- **Explain**  - attribute importance
- **Predict**  - classification or regression
- **Group**   - clustering/segmentation
- **Detect**  - anomaly/outlier detection
- **Map**     - project data to lower dimensionality
- **Profile** - supervised segmentation
Process Automation

- Statistics computation
- Sampling
- Attribute type identification
- Attribute selection
- Algorithm selection
- Data transformation
- Model creation and selection
- Output generation
Process Automation (cont.)

- **Statistics computation**
  - number of records, number of attributes, attribute ranges and cardinality
  - used to make decisions about target and attribute type and guide data transformations

- **Sampling (random & stratified)**
  - improve training times for large datasets
  - ensure sufficient rare target value/range representation
Process Automation (cont.)

- Attribute type identification
  - categorical versus numeric types
  - essential for correct data preparation and algorithm performance

- Attribute selection
  - enhance performance (speed and accuracy)
  - improve explanatory power
  - filter methods preferable over wrapper methods in the context of automation
Process Automation (cont.)

- Algorithm selection
  - data-driven (e.g., classification vs. regression)
- Data transformation
  - choice and sequence of transformations based on
    - selected algorithm
    - data characteristics (e.g., attribute type, attribute range/cardinality, percentage of missing values)
  - common transformations: binning, normalization, missing value imputation, outlier treatment
Process Automation (cont.)

• Model creation and selection
  – model creation across different algorithms or via parameter tuning on a single algorithm
  – quality assessment and selection
  – figure of merit provided to the user

• Output generation
  – scoring (e.g., prediction)
  – descriptive information (e.g., cluster description)
  – explanations must be compatible with original data values and ranges (transformation reversal)
Oracle Predictive Analytics

- PL/SQL API
- Targets database users
- Emphasis on ease-of-use
- Data in database (table/view)
- Results presented in tables
- Actionable results
Explain

- Embedded data preparation
- No intermediary objects persisted
- Produce figure of merit and rank for each attribute

```sql
DBMS_PREDICTIVE_ANALYTICS.EXPLAIN (  
data_table_name       IN VARCHAR2,  
explain_column_name   IN VARCHAR2,  
result_table_name     IN VARCHAR2,  
data_schema_name      IN VARCHAR2  
DEFAULT NULL);  
```
Explain Methodology

Data preparation

- Compute statistics
- Large data?
- Yes
  - Sample
- No
  - Remove outliers
  - Treat missing values
  - Normalize / Discretize

Build attribute importance model

Analyze model

Produce key attributes
Predict

- Automatically determine problem type
- Embedded data preparation
- No intermediary objects persisted
- Produce prediction for each record in data

```sql
DBMS_PREDICTIVE_ANALYTICS.PREDICT (  
    accuracy              OUT NUMBER,  
    data_table_name       IN VARCHAR2,  
    case_id_column_name   IN VARCHAR2,  
    target_column_name    IN VARCHAR2,  
    result_table_name     IN VARCHAR2,  
    data_schema_name      IN VARCHAR2 DEFAULT NULL);  
```
Predict Methodology

1. Compute statistics
2. Treat missing values
3. Normalize / Discretize
4. Remove outliers

Large data?
- Yes: Sample
- No: Go to next step

Data preparation

Split the data

Target type?
- Categorical: Build classification model
- Numeric: Build regression model

Measure performance
Score the whole data

Target type?
Spreadsheet Add-In for Predictive Analytics

- Excel front-end
- Targets business analysts
- Emphasis on ease-of-use
- Data in database or Excel spreadsheets
- Results presented in Excel
- Familiar environment for evaluation and presentation
Explain Results

<table>
<thead>
<tr>
<th>ATTRIBUTE_NAME</th>
<th>EXPLANATORY_VALUE</th>
<th>RANK</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBERSAUT</td>
<td>0.05894324</td>
<td>1</td>
</tr>
<tr>
<td>APERSAUT</td>
<td>0.03786425</td>
<td>2</td>
</tr>
<tr>
<td>PBAND</td>
<td>0.02365205</td>
<td>3</td>
</tr>
<tr>
<td>AWPART</td>
<td>0.01362232</td>
<td>4</td>
</tr>
<tr>
<td>PAWPART</td>
<td>0.00924178</td>
<td>5</td>
</tr>
<tr>
<td>APLEZIER</td>
<td>0.00621631</td>
<td>6</td>
</tr>
<tr>
<td>PLEZIER</td>
<td>0.0065774</td>
<td>7</td>
</tr>
<tr>
<td>ABYSTAND</td>
<td>0.00301782</td>
<td>8</td>
</tr>
<tr>
<td>ABRAND</td>
<td>0.0003645</td>
<td>9</td>
</tr>
</tbody>
</table>

![Graph showing EXPLANATORY_VALUE distribution]
Task: determine which attributes are useful in predicting the target (caravan insurance policy buyer)

The four most important attributes found were in agreement with submissions to the CoIL 2000 Challenge, including the competition’s winner.
Predict Results – Classification
Predict Results – CoIL 2000

- **Data:** CoIL 2000
- **Task:** predict which customers are interested in buying a caravan insurance policy

<table>
<thead>
<tr>
<th>Actual\Predicted</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2658</td>
<td>1104</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>138</td>
</tr>
</tbody>
</table>

- Results compare favorably to those submitted to the original competition
Predict Results – Regression

<table>
<thead>
<tr>
<th>ID</th>
<th>ID</th>
<th>MEDV</th>
<th>PREDICTION</th>
<th>% Deviation</th>
<th>Threshold (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>25.75</td>
<td>25.69</td>
<td>0.31</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>21.6</td>
<td>25.41</td>
<td>16.36</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>34.7</td>
<td>24.93</td>
<td>19.66</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>33.4</td>
<td>24.72</td>
<td>18.94</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>36.2</td>
<td>22.15</td>
<td>11.00</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>26.7</td>
<td>28.94</td>
<td>4.33</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>27.1</td>
<td>18.54</td>
<td>18.50</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>16.5</td>
<td>16.07</td>
<td>0.90</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>9</td>
<td>18.9</td>
<td>19.20</td>
<td>0.87</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>15.9</td>
<td>15.93</td>
<td>0.07</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>11</td>
<td>18.9</td>
<td>20.21</td>
<td>8.90</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>21.7</td>
<td>20.96</td>
<td>3.94</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>13</td>
<td>20.4</td>
<td>19.52</td>
<td>5.22</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>14</td>
<td>16.2</td>
<td>19.23</td>
<td>5.97</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>15</td>
<td>19.9</td>
<td>14.74</td>
<td>4.78</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>21.0</td>
<td>21.01</td>
<td>0.00</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>17</td>
<td>14.0</td>
<td>14.31</td>
<td>2.41</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>18</td>
<td>20.2</td>
<td>18.89</td>
<td>8.95</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>19</td>
<td>16.2</td>
<td>15.57</td>
<td>4.27</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>15.6</td>
<td>14.71</td>
<td>10.78</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>19.5</td>
<td>14.75</td>
<td>5.56</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>22</td>
<td>16.2</td>
<td>14.70</td>
<td>3.24</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>23</td>
<td>14.5</td>
<td>14.71</td>
<td>0.38</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>24</td>
<td>16.6</td>
<td>16.22</td>
<td>0.17</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>25</td>
<td>19.9</td>
<td>14.74</td>
<td>5.20</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>26</td>
<td>16.6</td>
<td>16.42</td>
<td>0.19</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>27</td>
<td>16.6</td>
<td>14.72</td>
<td>13.03</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>28</td>
<td>16.4</td>
<td>16.50</td>
<td>0.25</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>29</td>
<td>21</td>
<td>16.61</td>
<td>-2.89</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>30</td>
<td>12.7</td>
<td>14.70</td>
<td>15.71</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>31</td>
<td>1.5</td>
<td>14.74</td>
<td>10.61</td>
<td>0.36</td>
<td>15</td>
</tr>
<tr>
<td>32</td>
<td>12.7</td>
<td>14.70</td>
<td>15.71</td>
<td>0.36</td>
<td>15</td>
</tr>
</tbody>
</table>

**Housing Value Deviation**

Number of properties exceeding 16% value deviation

- # Undervalued
- # Overvalued

Oracle
Predict Results – Boston Housing

- Data: Boston Housing (US Census Service)
- Task: predict the median value of owner-occupied homes (MEDV)

<table>
<thead>
<tr>
<th></th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA</td>
<td>2.55</td>
<td>3.89</td>
</tr>
<tr>
<td>Excel Regression</td>
<td>3.99</td>
<td>5.96</td>
</tr>
</tbody>
</table>

- The results also compare well against other published predictive outcomes
Conclusions

• New design approach offers many benefits
  – ease-of-use with minimum user input
  – high productivity

• Potential to bring data mining to a wider audience
  – database community
  – business intelligence users

• Applications provide out the box competitive results on challenging data