Treeminer, Inc.

Training Module – Basic Concepts
Data Mining

• Process of analyzing data in order to uncover new patterns and knowledge
• Knowledge put to use to improve organization effectiveness
• “Data Mining” is broad term covering multiple different types of analysis
Major Categories of Data Mining

• Classification
  – Group data into predetermined groups
  – “Supervised” – examples of different groups are presented to algorithm

• Clustering
  – Group similar data together based upon a notion of similarity
  – “Unsupervised” – algorithm is not presented with any examples

• Association
  – Identify relationships between data
Popular Current Approaches

• Classification
  – SVM – Support Vector Machine
  – kNN – k- Nearest Neighbor

• Clustering
  – K-Means

• Association
  – ARM – Association Rule Mining
Challenges

- Data doubles every two years
- Server performance doubles every four years
- Algorithm execution time grows exponentially as datasets grow

"Extracting Value from Chaos", IDC, June, 2011
"How Much Information 2010", Short, Bohn, Baru, January, 2011
What is Vertical Data Mining?

• Today’s data mining methods operate over “rows” of data to derive insights.
  – As datasets grow, “rows” grow.
  – Example: a corpus of 10,000 documents growing to 100,000 document increases load on algorithm by factor of 10 (remember, non-linear growth in execution time!)

• Vertical data mining methods operate over “attributes”
  – Attributes grow slowly as “rows” grow
  – Increasing a corpus from 10,000 to 100,000 document will not appreciably change the number of attributes
Comparison

TODAY

Horizontal Rows of Data

→

Vertical Strips of Data (pTree)

<table>
<thead>
<tr>
<th>Row by row processing</th>
<th>Bit column by bit column processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured or unstructured</td>
<td>Structured or unstructured</td>
</tr>
<tr>
<td>Rows grow as dataset grows</td>
<td>Columns stay constant as dataset grows</td>
</tr>
<tr>
<td>Complex kernel or distance functions drive computational load</td>
<td>Simple logical operations (AND/OR) implement complex algebra to keep it computationally light</td>
</tr>
</tbody>
</table>
The “Layers” of Vertical Data Mining

- Data Mining Algorithms (like oblique, SPS, vertical k-means)
- Dataset Level Algebra (distance measures like manhattan, cosine similarity, etc)
- Primitive Algebra (addition, subtraction, comparisons, etc)
- pTrees – data structure holding vertical “strands” of data
Terms

- Class – a predefined set of categories that you wish to categorize new data into
- Training Set – a set of data used to “train” the software on how to tell classes apart. We use training sets to build classification models
- Attribute – a column of data from the dataset
- Supervised – algorithm that uses training set
- Unsupervised – algorithm that works “on the fly”
Clustering

• Unsupervised
  – No pre-determined model

• Attributes
  – Dimensions
  – Could be text terms, dataset columns, etc.

• Generally “lightly” iterative approach

• Upcoming release
Classification

- **Supervised**
  - Training model builds predetermined categories as selected by an expert

- **Model**
  - Enables algorithm to predict new samples

- **Classes**
  - “Types” of things you are looking for
    - Classify mortgage applicant as “qualified” or “not qualified”
    - Classify flower as “iris-setosia”, “iris-versicolor”, “iris-virginica”

- **Builds separating hyper-dimensional plane to make a prediction**
Oblique Classifier – Key Concepts

• $X_1$ and $X_2$ are “Attributes” $X_2$
• Black dots and White dots are two “Classes”
• $H_2$ is separating plane
• Each class has a set of statistics associated with it
• The algorithms calculates ideal placement of separating plane, and for every new point predicts which class it is
SPS Anomaly Detection – Key Concepts

- Builds training model of “normal”
- Identifies points that are furthest from “normal”
The Vertical Data Mining Process

Extract
- Conversion to Vertical Data Format

Model
- Create Classification Model and Save for Future Use

Classify
- Use Model to Predict Newly Arriving Samples
Artifacts – Extract Process

Extract

• Conversion to Vertical Data Format

Input: Text Documents, Structured Data
Input: “Attribute Map”, describing metadata
Output: pTree formatted Vertical Data, XML vertical dataset descriptor

• Artifacts

– Attribute Map (XML, input)
– pTree files (Text, output)
– Vertical dataset descriptors (XML, output)
Artifacts – Model Process

Model

• Create Classification Model and Save for Future Use

Input: pTree formatted Vertical Data, XML vertical dataset descriptor
Input: “Attribute Map”, describing metadata
Output: Classification Model

• Artifacts
  – Attribute Map (XML, input)
  – Classification Model (XML, output)
Artifacts – Classification Process

Classify
• Use Model to Predict Newly Arriving Samples

Input: pTree formatted Vertical Data, XML vertical dataset descriptor
Input: “Attribute Map”, describing metadata; “Classification Model”
Output: Data Mining results

• Artifacts
  – Attribute Map (XML, input)
  – Classification Model (XML, input)
  – Results (csv file, output)