Data Mining
Session 5 – Main Theme
Characterization

Dr. Jean-Claude Franchitti
New York University
Computer Science Department
Courant Institute of Mathematical Sciences

Adapted from course textbook resources
Data Mining Concepts and Techniques (2nd Edition)
Jiawei Han and Micheline Kamber

Agenda

1 Session Overview
2 Characterization
3 Summary and Conclusion
Characterization in Brief

- What is Concept Description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases

Icons / Metaphors

- Information
- Common Realization
- Knowledge/Competency Pattern
- Governance
- Alignment
- Solution Approach
Agenda

1. Session Overview
2. Characterization
3. Summary and Conclusion

Concept Description: Characterization and Comparison

- What is Concept Description?
  - Data generalization and summarization-based characterization
  - Analytical characterization: Analysis of attribute relevance
  - Mining class comparisons: Discriminating between different classes
  - Mining descriptive statistical measures in large databases
What is Concept Description?

- Descriptive vs. predictive data mining
  - **Descriptive mining**: describes concepts or task-relevant data sets in concise, summarative, informative, discriminative forms
  - **Predictive mining**: Based on data and analysis, constructs models for the database, and predicts the trend and properties of unknown data
- Concept description:
  - **Characterization**: provides a concise and succinct summarization of the given collection of data
  - **Comparison**: provides descriptions comparing two or more collections of data

Concept Description: Characterization and Comparison

- What is Concept Description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
Data Generalization and Summarization-based Characterization

- **Data generalization**
  - A process which abstracts a large set of task-relevant data in a database from a low conceptual levels to higher ones.

![Conceptual levels diagram]

Approaches:
- Data cube approach (OLAP approach)
- Attribute-oriented induction approach

**Characterization: Data Cube Approach**

- Perform computations and store results in data cubes
- **Strength**
  - An efficient implementation of data generalization
  - Computation of various kinds of measures
    - e.g., count( ), sum( ), average( ), max( )
  - Generalization and specialization can be performed on a data cube by roll-up and drill-down
- **Limitations**
  - handle only dimensions of *simple nonnumeric data* and measures of *simple aggregated numeric values*.
  - Lack of intelligent analysis, can’t tell which dimensions should be used and what levels should the generalization reach
Attribute-Oriented Induction

- Proposed in 1989 (KDD ‘89 workshop)
- Not confined to categorical data nor particular measures.
- How it is done?
  - Collect the task-relevant data (initial relation) using a relational database query
  - Perform generalization by attribute removal or attribute generalization.
  - Apply aggregation by merging identical, generalized tuples and accumulating their respective counts.
  - Interactive presentation with users.

Basic Principles of Attribute-Oriented Induction

- **Data focusing**: task-relevant data, including dimensions, and the result is the initial relation.
- **Attribute-removal**: remove attribute A if there is a large set of distinct values for A but (1) there is no generalization operator on A, or (2) A’s higher level concepts are expressed in terms of other attributes.
- **Attribute-generalization**: If there is a large set of distinct values for A, and there exists a set of generalization operators on A, then select an operator and generalize A.
- **Attribute-threshold control**: typical 2-8, specified/default.
- **Generalized relation threshold control**: control the final relation/rule size.
### Example

- Describe general characteristics of graduate students in the Big-University database
  ```
  use Big_University_DB
  mine characteristics as "Science_Students" in relevance to name, gender, major, birth_place, birth_date, residence, phone#, gpa
  from student
  where status in "graduate"
  ```
- Corresponding SQL statement:
  ```
  Select name, gender, major, birth_place, birth_date, residence, phone#, gpa
  from student
  where status in {"Msc", "MBA", "PhD" }
  ```

### Class Characterization: An Example

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Major</th>
<th>Birth-Place</th>
<th>Birth_date</th>
<th>Residence</th>
<th>Phone #</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jim Woodman</td>
<td>M</td>
<td>CS</td>
<td>Vancouver, BC, Canada</td>
<td>8-12-76</td>
<td>3511 Main St., Richmond</td>
<td>687-4958</td>
<td>3.67</td>
</tr>
<tr>
<td>Scott Lachance</td>
<td>M</td>
<td>CS</td>
<td>Montreal, Que, Canada</td>
<td>28-7-75</td>
<td>345 1st Ave., Richmond</td>
<td>253-9106</td>
<td>3.70</td>
</tr>
<tr>
<td>Laura Lee</td>
<td>F</td>
<td>Physics</td>
<td>Seattle, WA, USA</td>
<td>25-8-70</td>
<td>125 Austin Ave., Burnaby</td>
<td>420-5232</td>
<td>3.83</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Major</th>
<th>Birth_region</th>
<th>Age_range</th>
<th>Residence</th>
<th>GPA</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Science</td>
<td>Canada</td>
<td>20-25</td>
<td>Richmond</td>
<td>Very-good</td>
<td>16</td>
</tr>
<tr>
<td>F</td>
<td>Science</td>
<td>Foreign</td>
<td>25-30</td>
<td>Burnaby</td>
<td>Excellent</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Country</th>
<th>Foreign</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>16</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>22</td>
<td>32</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>36</td>
<td>62</td>
</tr>
</tbody>
</table>
Concept Description: Characterization and Comparison

- What is Concept Description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases

Characterization vs. OLAP

- Similarity:
  - Presentation of data summarization at multiple levels of abstraction.
  - Interactive drilling, pivoting, slicing and dicing.
- Differences:
  - Automated desired level allocation.
  - Dimension relevance analysis and ranking when there are many relevant dimensions.
  - Sophisticated typing on dimensions and measures.
  - Analytical characterization: data dispersion analysis.
Attribute Relevance Analysis

- Why?
  - Which dimensions should be included?
  - How high level of generalization?
  - Automatic vs. interactive
  - Reduce # attributes; easy to understand patterns

- What?
  - Statistical method for preprocessing data
    - Filter out irrelevant or weakly relevant attributes
    - Retain or rank the relevant attributes
  - Relevance related to dimensions and levels
  - Analytical characterization, analytical comparison

Attribute relevance analysis (continued)

- How?
  - Data Collection
  - Analytical Generalization
    - Use information gain analysis (e.g., entropy or other measures) to identify highly relevant dimensions and levels.
  - Relevance Analysis
    - Sort and select the most relevant dimensions and levels.
  - Attribute-oriented Induction for class description
    - On selected dimension/level
  - OLAP operations (e.g. drilling, slicing) on relevance rules
Relevance Measures

- Quantitative relevance measure determines the classifying power of an attribute within a set of data.
- Methods
  - information gain (ID3)
  - gain ratio (C4.5)
  - $\chi^2$ contingency table statistics
  - uncertainty coefficient

Information-Theoretic Approach

- Decision tree
  - each internal node tests an attribute
  - each branch corresponds to attribute value
  - each leaf node assigns a classification
- ID3 algorithm
  - build decision tree based on training objects with known class labels to classify testing objects
  - rank attributes with information gain measure
  - minimal height
    - the least number of tests to classify an object
**Top-Down Induction of Decision Tree**

Attributes = \{Outlook, Temperature, Humidity, Wind\}
PlayTennis = \{yes, no\}

---

**Entropy and Information Gain**

- S contains $s_i$ tuples of class $C_i$ for $i = \{1, \ldots, m\}$
- Information measures info required to classify any arbitrary tuple

$$I(S_1, S_2, \ldots, S_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log \frac{s_i}{s}$$
- Entropy of attribute A with values \{a_1, a_2, \ldots, a_v\}

$$E(A) = \sum_{j=1}^{v} \frac{S_j}{s} I(S_1, \ldots, S_m)$$
- Information gained by branching on attribute A

$$Gain(A) = I(S_1, S_2, \ldots, S_m) - E(A)$$
Example: Analytical Characterization

- Task
  - Mine general characteristics describing graduate students using analytical characterization

- Given
  - attributes *name*, *gender*, *major*, *birth_place*, *birth_date*, *phone*#, and *gpa*
  - $Gen(a_i) =$ concept hierarchies on $a_i$
  - $U_i =$ attribute analytical thresholds for $a_i$
  - $T_i =$ attribute generalization thresholds for $a_i$
  - $R = $ attribute statistical relevance threshold

Example: Analytical Characterization (continued)

- 1. Data collection
  - target class: graduate student
  - contrasting class: undergraduate student

- 2. Analytical generalization using $U_i$
  - attribute removal
    - remove *name* and *phone*#
  - attribute generalization
    - generalize *major*, *birth_place*, *birth_date* and *gpa*
    - accumulate counts
  - candidate relation: *gender*, *major*, *birth_country*, *age_range* and *gpa*
Example: Analytical characterization
(continued)

<table>
<thead>
<tr>
<th>gender</th>
<th>major</th>
<th>birth_country</th>
<th>age_range</th>
<th>gpa</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Science</td>
<td>Canada</td>
<td>20-25</td>
<td>Very_good</td>
<td>16</td>
</tr>
<tr>
<td>F</td>
<td>Science</td>
<td>Foreign</td>
<td>25-30</td>
<td>Excellent</td>
<td>22</td>
</tr>
<tr>
<td>M</td>
<td>Engineering</td>
<td>Foreign</td>
<td>25-30</td>
<td>Excellent</td>
<td>18</td>
</tr>
<tr>
<td>F</td>
<td>Science</td>
<td>Foreign</td>
<td>25-30</td>
<td>Excellent</td>
<td>25</td>
</tr>
<tr>
<td>M</td>
<td>Science</td>
<td>Canada</td>
<td>20-25</td>
<td>Excellent</td>
<td>21</td>
</tr>
<tr>
<td>F</td>
<td>Engineering</td>
<td>Canada</td>
<td>20-25</td>
<td>Excellent</td>
<td>18</td>
</tr>
</tbody>
</table>

Candidate relation for Target class: Graduate students (Σ=120)

<table>
<thead>
<tr>
<th>gender</th>
<th>major</th>
<th>birth_country</th>
<th>age_range</th>
<th>gpa</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Science</td>
<td>Foreign</td>
<td>&lt;20</td>
<td>Very_good</td>
<td>18</td>
</tr>
<tr>
<td>F</td>
<td>Business</td>
<td>Canada</td>
<td>&lt;20</td>
<td>Fair</td>
<td>20</td>
</tr>
<tr>
<td>M</td>
<td>Business</td>
<td>Canada</td>
<td>&lt;20</td>
<td>Fair</td>
<td>22</td>
</tr>
<tr>
<td>F</td>
<td>Science</td>
<td>Canada</td>
<td>20-25</td>
<td>Fair</td>
<td>24</td>
</tr>
<tr>
<td>M</td>
<td>Engineering</td>
<td>Foreign</td>
<td>20-25</td>
<td>Very_good</td>
<td>22</td>
</tr>
<tr>
<td>F</td>
<td>Engineering</td>
<td>Canada</td>
<td>&lt;20</td>
<td>Excellent</td>
<td>24</td>
</tr>
</tbody>
</table>

Candidate relation for Contrasting class: Undergraduate students (Σ=130)

Example: Analytical characterization
(continued)

3. Relevance analysis
   » Calculate expected info required to classify an arbitrary tuple
   \[ I(s_1, s_2) = I(120, 130) = \frac{120}{250} \log_2 \frac{120}{250} + \frac{130}{250} \log_2 \frac{130}{250} = 0.9988 \]

   » Calculate entropy of each attribute: e.g. major

   For major = "Science": \[ s_{11}=84 \quad s_{21}=42 \quad I(s_{11}, s_{21})=0.9183 \]
   For major = "Engineering": \[ s_{12}=36 \quad s_{22}=46 \quad I(s_{12}, s_{22})=0.9892 \]
   For major = "Business": \[ s_{13}=0 \quad s_{23}=42 \quad I(s_{13}, s_{23})=0 \]

   Number of grad students in “Science”
   Number of undergrad students in “Science”
Example: Analytical Characterization
(continued)

- Calculate expected info required to classify a given sample if $S$ is partitioned according to the attribute

$$E(\text{major}) = \frac{126}{250} I(s_{11}, s_{21}) + \frac{82}{250} I(s_{12}, s_{22}) + \frac{42}{250} I(s_{13}, s_{23}) = 0.7873$$

- Calculate information gain for each attribute

$$\text{Gain(major)} = I(s_1, s_2) - E(\text{major}) = 0.2115$$

  » Information gain for all attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>0.0003</td>
</tr>
<tr>
<td>birth_country</td>
<td>0.0407</td>
</tr>
<tr>
<td>major</td>
<td>0.2115</td>
</tr>
<tr>
<td>gpa</td>
<td>0.4490</td>
</tr>
<tr>
<td>age_range</td>
<td>0.5971</td>
</tr>
</tbody>
</table>

Example: Analytical characterization
(continued)

- 4. Initial working relation derivation
  » $R = 0.1$
  » remove irrelevant/weakly relevant attributes from candidate relation => drop gender, birth_country
  » remove contrasting class candidate relation

<table>
<thead>
<tr>
<th>major</th>
<th>age_range</th>
<th>gpa</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>20-25</td>
<td>Very_good</td>
<td>16</td>
</tr>
<tr>
<td>Science</td>
<td>25-30</td>
<td>Excellent</td>
<td>47</td>
</tr>
<tr>
<td>Science</td>
<td>20-25</td>
<td>Excellent</td>
<td>21</td>
</tr>
<tr>
<td>Engineering</td>
<td>20-25</td>
<td>Excellent</td>
<td>18</td>
</tr>
<tr>
<td>Engineering</td>
<td>25-30</td>
<td>Excellent</td>
<td>18</td>
</tr>
</tbody>
</table>

Initial target class working relation: Graduate students

- 5. Perform attribute-oriented induction
**Concept Description: Characterization and Comparison**

- What is Concept Description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
  - Mining class comparisons: Discriminating between different classes
  - Mining descriptive statistical measures in large databases

**Mining Class Comparisons**

- Comparison: Comparing two or more classes.
- Method:
  - Partition the set of relevant data into the target class and the contrasting class(es)
  - Generalize both classes to the same high level concepts
  - Compare tuples with the same high level descriptions
  - Present for every tuple its description and two measures:
    - support - distribution within single class
    - comparison - distribution between classes
  - Highlight the tuples with strong discriminant features
- Relevance Analysis:
  - Find attributes (features) which best distinguish different classes.
**Example: Analytical Comparison**

- **Task**
  - Compare graduate and undergraduate students using discriminant rule.
  - **DMQL query**

  ```
  use Big_University_DB
  mine comparison as "grad_vs_undergrad_students"
  in relevance to name, gender, major, birth_place, birth_date, residence, phone#, gpa
  for "graduate_students"
  where status in "graduate"
  versus "undergraduate_students"
  where status in "undergraduate"
  analyze count%
  from student
  ```

**Example: Analytical comparison**
(continued)

- **Given**
  - attributes name, gender, major, birth_place, birth_date, residence, phone# and gpa
  - $Gen(a_i) = \text{concept hierarchies on attributes } a_i$
  - $U_i = \text{attribute analytical thresholds for attributes } a_i$
  - $T_i = \text{attribute generalization thresholds for attributes } a_i$
  - $R = \text{attribute relevance threshold}$
Example: Analytical comparison
(continued)

- 1. Data collection
  » target and contrasting classes

- 2. Attribute relevance analysis
  » remove attributes name, gender, major, phone#

- 3. Synchronous generalization
  » controlled by user-specified dimension thresholds
  » prime target and contrasting class(es) relations/cuboids

<table>
<thead>
<tr>
<th>Birth_country</th>
<th>Age_range</th>
<th>Gpa</th>
<th>Count%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>20-25</td>
<td>Good</td>
<td>5.53%</td>
</tr>
<tr>
<td>Canada</td>
<td>25-30</td>
<td>Good</td>
<td>2.32%</td>
</tr>
<tr>
<td>Canada</td>
<td>Over_30</td>
<td>Very_good</td>
<td>5.86%</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Other</td>
<td>Over_30</td>
<td>Excellent</td>
<td>4.68%</td>
</tr>
</tbody>
</table>

Prime generalized relation for the target class: Graduate students

<table>
<thead>
<tr>
<th>Birth_country</th>
<th>Age_range</th>
<th>Gpa</th>
<th>Count%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>15-20</td>
<td>Fair</td>
<td>5.53%</td>
</tr>
<tr>
<td>Canada</td>
<td>15-20</td>
<td>Good</td>
<td>4.53%</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Canada</td>
<td>25-30</td>
<td>Good</td>
<td>5.02%</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Other</td>
<td>Over_30</td>
<td>Excellent</td>
<td>0.68%</td>
</tr>
</tbody>
</table>

Prime generalized relation for the contrasting class: Undergraduate students
4. Drill down, roll up and other OLAP operations on target and contrasting classes to adjust levels of abstractions of resulting description

5. Presentation
   » as generalized relations, crosstabs, bar charts, pie charts, or rules
   » contrasting measures to reflect comparison between target and contrasting classes
      • e.g. count%

---

**Concept Description: Characterization and Comparison**

- What is Concept Description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
Mining Data Dispersion Characteristics

- **Motivation**
  - To better understand the data: central tendency, variation and spread

- **Data dispersion characteristics**
  - median, max, min, quantiles, outliers, variance, etc.

- **Numerical dimensions correspond to sorted intervals**
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals

- **Dispersion analysis on computed measures**
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- **Mean**
  - $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - Weighted arithmetic mean
  - $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

- **Median: A holistic measure**
  - Middle value if odd number of values, or average of the middle two values otherwise
  - estimated by interpolation
  - $median = L_h + \left( \frac{n/2 - \sum f l}{f_{median}} \right) c$

- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: $Q_1$ (25th percentile), $Q_3$ (75th percentile)
  - Inter-quartile range: $IQR = Q_3 - Q_1$
  - Five number summary: min, $Q_1$, $M$, $Q_3$, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - Outlier: usually, a value higher/lower than $1.5 \times IQR$

- Variance and standard deviation
  - Variance $s^2$: (algebraic, scalable computation)
  - Standard deviation $s$ is the square root of variance $s^2$

\[
 s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2 \right]
\]

Boxplot Analysis

- Five-number summary of a distribution: Minimum, $Q_1$, $M$, $Q_3$, Maximum
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is $IQR$
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum
A Boxplot

unit price (S)

Agenda

1 Session Overview
2 Characterization
3 Summary and Conclusion
### Summary

- Concept description: characterization and discrimination
- OLAP-based vs. attribute-oriented induction
- Efficient implementation of AOI
- Analytical characterization and comparison
- Mining descriptive statistical measures in large databases
- Discussion
  - Incremental and parallel mining of description
  - Descriptive mining of complex types of data

### References

References (continued)


Assignments & Readings

- Readings
  - Chapter 3
- Individual Project #1
  - Due March 11 2010
Next Session: Mining Frequent Patterns, Association, and Correlations