Knowledge Discovery and Data Mining I

Winter Semester 2018/19
Agenda

1. Introduction

2. Basics
   2.1 Data Representation
   2.2 Data Reduction
   2.3 Visualization
   2.4 Privacy

3. Unsupervised Methods

4. Supervised Methods

5. Advanced Topics
Why data reduction?

- Better perception of patterns
  - Raw (tabular) data is hard to understand
  - Visualization is limited to (hundreds of) thousands of objects
  - Reduction of data may help to identify patterns
- Computational complexity
  - Big data sets cause prohibitively long runtime for data mining algorithms
  - Reduced data sets are useful the more the algorithms produce (almost) the same analytical results

How to approach data reduction?

- Data aggregation (basic statistics)
- Data generalization (abstraction to higher levels)
### Data Reduction Strategies

#### Numerosity Reduction
Reduce number of objects
- Sampling (loss of data)
- Aggregation (model parameters, e.g., center / spread)

#### Dimensionality Reduction
Reduce number of attributes

#### Quantization, Discretization
Reduce number of values per domain

---

<table>
<thead>
<tr>
<th>ID</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54</td>
<td>56</td>
<td>75</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>12</td>
<td>65</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>63</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>86</td>
<td>23</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>A1</th>
<th>A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L</td>
<td>75</td>
</tr>
<tr>
<td>3</td>
<td>XS</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>XL</td>
<td>4</td>
</tr>
</tbody>
</table>
### Data Reduction Strategies

#### Dimensionality reduction

Reduce number of attributes

- **Linear methods:** feature sub-selection, Principal Components Analysis, Random projections, Fourier transform, Wavelet transform
- **Non-linear methods:** Multidimensional scaling (force model)

#### Quantization, discretization

Reduce number of values per domain

- Binning (various types of histograms)
- Generalization along hierarchies (OLAP, attribute-oriented induction)
Data Generalization

- Quantization is a special case of generalization
  - E.g., group age (7 bits) to age_range (4 bits)
- Dimensionality reduction is degenerate quantization
  - Dropping age reduces 7 bits to zero bits
  - Corresponds to generalization of age to "all" = "any age" = no information
Data Aggregation

- Aggregation is numerosity reduction (= less tuples)
- Generalization yields duplicates: Merge duplicate tuples and introduce (additional) counter attribute
Basic Aggregates

- Central tendency: Where is the data located? Where is it centered?
  - Examples: mean, median, mode, etc. (see below)
- Variation, spread: How much do the data deviate from the center?
  - Examples: variance / standard deviation, min-max-range, ...
Distributive Aggregate Measures

Distributive Measures

The result derived by applying the function to \( n \) aggregate values is the same as that derived by applying the function on all the data without partitioning.

Examples

- \( \text{count}(D_1 \cup D_2) = \text{count}(D_1) + \text{count}(D_2) \)
- \( \text{sum}(D_1 \cup D_2) = \text{sum}(D_1) + \text{sum}(D_2) \)
- \( \text{min}(D_1 \cup D_2) = \text{min}(\text{min}(D_1), \text{min}(D_2)) \)
- \( \text{max}(D_1 \cup D_2) = \text{max}(\text{max}(D_1), \text{max}(D_2)) \)
**Algebraic Aggregate Measures**

**Algebraic Measures**

Can be computed by an algebraic function with $M$ arguments (where $M$ is a bounded integer), each of which is obtained by applying a distributive aggregate function.

**Examples**

- $\text{avg}(D_1 \cup D_2) = \frac{\text{sum}(D_1 \cup D_2)}{\text{count}(D_1 \cup D_2)} = \frac{\text{sum}(D_1) + \text{sum}(D_2)}{\text{count}(D_1) + \text{count}(D_2)}$
  
  $\neq \text{avg}(\text{avg}(D_1), \text{avg}(D_2))$

- $\text{standard deviation}(D_1 \cup D_2)$
Holistic Aggregate Measures

Holistic Measures

There is no constant bound on the storage size which is needed to determine/describe a sub-aggregate.

Examples

- **median**: value in the middle of a sorted series of values (=50% quantile)
  \[
  \text{median}(D_1 \cup D_2) \neq \text{simple}\_\text{function}(\text{median}(D_1), \text{median}(D_2))
  \]

- **mode**: value that appears most often in a set of values

- **rank**: $k$-smallest / $k$-largest value (cf. quantiles, percentiles)
Measuring the Central Tendency

Mean – (weighted) arithmetic mean

Well-known measure for central tendency ("average").

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}
\]

Mid-range

Average of the largest and the smallest values in a data set:

\[
(max + min)/2
\]

- Algebraic measures
- Applicable to numerical data only (sum, scalar multiplication)

What about categorical data?
Measuring the Central Tendency

Median

- Middle value if odd number of values
- For even number of values: average of the middle two values (numeric case), or one of the two middle values (non-numeric case)
- Applicable to ordinal data only (an ordering is required)
- Holistic measure

Examples

- never, never, never, rarely, rarely, often, usually, usually, always
- tiny, small, big, big, big, big, big, huge, huge
- tiny, tiny, small, medium, big, big, large, huge

What if there is no ordering?
Mode

- Value that occurs most frequently in the data
- Example: blue, red, blue, yellow, green, blue, red
- Unimodal, bimodal, trimodal, . . .: There are 1, 2, 3, . . . modes in the data (multi-modal in general), cf. mixture models
- There is no mode if each data value occurs only once
- Well suited for categorical (i.e., non-numerical) data
Measuring the Dispersion of Data

Variance

- Applicable to numerical data, scalable computation:

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

- Calculation by two passes: numerically much more stable
- Single pass: calculate sum of squares and square of sum in parallel
- Measures the spread around the mean
- It is zero if and only if all the values are equal
- Standard deviation: Square root of the variance
- Both the standard deviation and the variance are algebraic
Boxplot Analysis

Five-number summary of a distribution

- Minimum, Q1, Median, Q3, Maximum
- Represents 0%, 25%, 50%, 75%, 100%-quantile of the data
- Also called "25-percentile", etc.

Boxplot

- Boundaries: first and third quartiles
- Height: inter-quartile range (IQR)
- The median is marked by a line within the box
- Whiskers: minimum and maximum
- Outliers: usually values more than 1.5 \cdot \text{IQR} below Q1 or above Q3
Boxplot Example

Iris Dataset

sepal length (cm)

setosa versicolor virginica

class

4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Data Generalization

- Which partitions of the data to aggregate?
- All data
  - Overall mean, overall variance: too coarse (overgeneralized)
- Different techniques to form groups for aggregation
  - Binning – histograms, based on value ranges
  - Generalization – abstraction based on generalization hierarchies
  - Clustering (see later) – based on object similarity
Histograms use binning to approximate data distributions

- Divide data into bins and store a representative (sum, average, median) for each bin
- Popular data reduction and analysis method
- Related to quantization problems
Equi-width Histograms

- Divide the range into $N$ intervals of equal size: uniform grid
- If $A$ and $B$ are the lowest and highest values of the attribute, the width of intervals will be $(B - A)/N$

Positive
- Most straightforward

Negative
- Outliers may dominate presentation
- Skewed data is not handled well
Equi-width Histograms

Example

- Sorted data, 10 bins: 5, 7, 8, 8, 9, 11, 13, 13, 14, 14, 14, 15, 17, 17, 17, 18, 19, 23, 24, 25, 26, 26, 26, 27, 28, 32, 34, 36, 37, 38, 39, 97

- Insert 999
Equi-height Histograms

Divide the range into $N$ intervals, each containing approx. the same number of samples (quantile-based approach)

Positive

- Good data scaling

Negative

- If any value occurs often, the equal frequency criterion might not be met (intervals have to be disjoint!)
Equi-height Histograms

**Example**

- Same data, 4 bins: 5, 7, 8, 8, 9, 11, 13, 13, 14, 14, 14, 15, 17, 17, 17, 18, 19, 23, 24, 25, 26, 26, 26, 27, 28, 32, 34, 36, 37, 38, 39, 97

- Median = 50%-quantile
  - More robust against outliers (cf. value 999 from above)
  - Four bin example is strongly related to boxplot
Concept Hierarchies: Examples

No (real) hierarchies

- Name: all
  - A. Abbeck
  - ...
- Gender: all
  - male
  - female
- Phone: all
  - 158932
  - ...
  - 98763

Set grouping hierarchies

- Age: all
  - 15-19
    - 15
    - ...
    - 19
  - 20-24
    - ...
    - ...
    - 25
  - 25-30
    - ...
    - ...
    - 30
Concept Hierarchies: Examples

Schema hierarchies

Place: all
  ├── North America
  │    ├── Canada
  │    └── USA
  │        ├── Vancouver
  │        └── Toronto
  └── Asia
      └── Europe
          └── France

Major: all
  ├── Science
  │    └── CS
  └── Business
      └── EE

Engineering
  └── Civil Eng.
Concept Hierarchy for Categorical Data

- Concept hierarchies can be specified by experts or just by users

- Heuristically generate a hierarchy for a set of (related) attributes
  - based on the number of distinct values per attribute in the attribute set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy

- Fails for counter examples: 20 distinct years, 12 months, 7 days_of_week, but not "year < month < days_of_week" with the latter on top
Summarization-based Aggregation

Data Generalization

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

Conceptual levels:

- 1: all
- 2: federal states
- 3: states
- 4: countries
- 5: cities

Example:

- all
- federal states
- states
- countries
- cities

Approaches:

- Data-cube approach (OLAP / Roll-up) – manual
- Attribute-oriented induction (AOI) – automated
# Basic OLAP Operations

<table>
<thead>
<tr>
<th>Roll up</th>
<th>Summarize data by climbing up hierarchy or by dimension reduction.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drill down</td>
<td><em>Reverse of roll-up.</em> From higher level summary to lower level summary or detailed data, or introducing new dimensions.</td>
</tr>
<tr>
<td>Slice and dice</td>
<td><em>Selection</em> on one (slice) or more (dice) dimensions.</td>
</tr>
<tr>
<td>Pivot (rotate)</td>
<td><em>Reorient</em> the cube, visualization, 3D to series of 2D planes.</td>
</tr>
</tbody>
</table>
Example: Roll up / Drill down

### Query

```
SELECT *  
FROM business  
GROUP BY country, quarter
```

### Roll-Up

```
SELECT *  
FROM business  
GROUP BY continent, quarter
```

```
SELECT *  
FROM business  
GROUP BY country
```

### Drill-Down

```
SELECT *  
FROM business  
GROUP BY city, quarter
```

```
SELECT *  
FROM business  
GROUP BY country, quarter, product
```
Example: Roll up in a Data Cube
Example: Slice Operation

```sql
SELECT income
FROM time t, product p, country c
WHERE p.name = 'VCR'
```

VCR dimension is chosen
Example: Dice Operation

```
SELECT income
FROM  time t, product p, country c
WHERE p.name = 'VCR' OR p.name = 'PC' AND t.quarter BETWEEN 2 AND 3
```

sub-data cube over PC, VCR and quarters 2 and 3 is extracted
Example: Pivot (rotate)

<table>
<thead>
<tr>
<th>year</th>
<th>17</th>
<th>18</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>product</td>
<td>TV</td>
<td>PC</td>
<td>VCR</td>
</tr>
</tbody>
</table>

↓ Pivot (rotate) ↓

<table>
<thead>
<tr>
<th>product</th>
<th>TV</th>
<th>PC</th>
<th>VCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
</tbody>
</table>

Basics Data Reduction October 31, 2018 80
Other operations

- *Drill across*: involving (across) more than one fact table
- *Drill through*: through the bottom level of the cube to its back-end relational tables (using SQL)
Specifying Generalization by a Star-Net

- Each circle is called a *footprint*
- Footprints represent the granularities available for OLAP operations
Discussion of OLAP-based Generalization

▶ Strength
  ▶ 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
  ▶ Handles only dimensions of simple non-numeric data and measures of simple aggregated numeric values
  ▶ Lack of intelligent analysis, can’t tell which dimensions should be used and what levels the generalization should reach
Attribute-Oriented Induction (AOI)

- Apply aggregation by merging identical, generalized tuples and accumulating their respective counts.

- **Data focusing**: task-relevant data, including dimensions, and the result is the initial relation

- **Generalization Plan**: Perform generalization by either *attribute removal* or *attribute generalization*
Attribute-Oriented Induction (AOI)

Attribute Removal

Remove attribute $A$ if:

- there is a large set of distinct values for $A$ but there is no generalization operator (concept hierarchy) on $A$, or
- $A$’s higher level concepts are expressed in terms of other attributes (e.g. street is covered by city, state, country).

Attribute Generalization

If there is a large set of distinct values for $A$, and there exists a set of generalization operators (i.e., a concept hierarchy) on $A$, then select an operator and generalize $A$. 
Attribute Oriented Induction: Example

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Major</th>
<th>Birth place</th>
<th>Birth data</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>8-12-81</td>
<td>3511 Main St., Richmond</td>
<td>687-4598</td>
<td>3.67</td>
</tr>
<tr>
<td>Scott Lachance</td>
<td>M</td>
<td>CS</td>
<td>Montreal, Que, Canada</td>
<td>28-7-80</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>25-8-75</td>
<td>125 Austin Ave., Burnaby</td>
<td>420-5232</td>
<td>3.83</td>
</tr>
</tbody>
</table>

- Name: large number of distinct values, no hierarchy – removed
- Gender: only two distinct values – retained
- Birth_place: many values, hierarchy – generalized, e.g., to country
- Birth_date: many values – generalized to age (or age_range)
- Residence: many streets and numbers – generalized to city
- Phone number: many values, no hierarchy – removed
- Grade_point_avg (GPA): hierarchy exists – generalized to good, ...
- Count: additional attribute to aggregate base tuples
Attribute Oriented Induction: Example

▶ Initial Relation:

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Major</th>
<th>Birth place</th>
<th>Birth data</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>8-12-81</td>
<td>3511 Main St., Richmond</td>
<td>687-4598</td>
<td>3.67</td>
</tr>
<tr>
<td>Scott Lachance</td>
<td>M</td>
<td>CS</td>
<td>Montreal, Que, Canada</td>
<td>28-7-80</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>25-8-75</td>
<td>125 Austin Ave., Burnaby</td>
<td>420-5232</td>
<td>3.83</td>
</tr>
</tbody>
</table>

▶ Prime Generalized Relation:

<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>

▶ Crosstab for generalized relation:

<table>
<thead>
<tr>
<th></th>
<th>Canada</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>
Attribute Generalization Control

Problem: How many distinct values for an attribute?

- Overgeneralization: values are too high-level
- Undergeneralization: level not sufficiently high
- Both yield tuples of poor usefulness

Two common approaches

- Attribute-threshold control: default or user-specified, typically 2-8 values
- Generalized relation threshold control: control the size of the final relation/rule, e.g., 10-30
Next Attribute Selection Strategies for Generalization

- **Aiming at minimal degree of generalization**
  - Choose attribute that reduces the number of tuples the most
  - Useful heuristic: choose attribute with highest number of distinct values.

- **Aiming at similar degree of generalization for all attributes**
  - Choose the attribute currently having the least degree of generalization

- **User-controlled**
  - Domain experts may specify appropriate priorities for the selection of attributes