Medical Information Management & Mining

You Chen
Jan, 15, 2013
You.chen@vanderbilt.edu
Trees cannot be used to build a house directly. How can we transform trees to building materials?
What we need to know to conduct access logs auditing?

- Data Representation
- Data Normalization
- Similarity Measurements
- Dimensionality Reduction
Data Representation

Information in the real world are abstract and cannot be directly modeled.

Data which can be interpreted or used by designed models.

Data are collected by mapping entities in the domain of interest to symbolic representation by means of some measurement procedure, which associates the value of a variable with a given property of an entity.

An example of data representation in access logs of EHR

Bipartite graph of users and subjects

<table>
<thead>
<tr>
<th>EHR Access Logs</th>
<th>Bipartite graph of users and subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>Patient</td>
</tr>
<tr>
<td>Dr. J</td>
<td>Alice</td>
</tr>
<tr>
<td>Dr. J</td>
<td>Bob</td>
</tr>
<tr>
<td>Nurse S</td>
<td>Bob</td>
</tr>
<tr>
<td>Nurse S</td>
<td>Daniel</td>
</tr>
<tr>
<td>Dr. X</td>
<td>Daniel</td>
</tr>
<tr>
<td>Dr. X</td>
<td>Charles</td>
</tr>
<tr>
<td>Nurse S</td>
<td>Charles</td>
</tr>
</tbody>
</table>

Binary matrix of subjects and users

<table>
<thead>
<tr>
<th></th>
<th>u₁</th>
<th>u₂</th>
<th>u₃</th>
<th>u₄</th>
<th>u₅</th>
<th>u₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>s₁</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s₂</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>s₃</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>s₄</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>s₅</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>s₆</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>s₇</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Various data representations

• Flatted data
  – Data matrix or table
• Time series
• Text
• Image and video
• Workflow data
Distinct Patient Accesses across Time

- Accesses in week days
- Accesses in weekend
A Printed Encounter Note

Subjective

Chief Complaint
Headache

Patient History
- Illnesses - asdsdfsdfsdfsdfsafs; Measles asdf
- Operations - Appendectomy - 1987
- Social History - Married; Employed
- Family History - Denies All

Allergies
- Uncategorized - Peanut-Allergen-Ingredient
- Uncategorized - Advil Liqui-Gel-Allergen-Medication
- Uncategorized - Augmentin-Allergen-Medication
- Uncategorized - Amoxicillin-Allergen-Ingredient
- Uncategorized - Tylenol-Allergen-Medication
- Uncategorized - Vytorin 10-10-Allergen-Medication - Difficulty breathing/constricting of throat, Dizziness, Congestion, Anxiety, Confusion, Drowsiness, Intolerance; comments
- Uncategorized - Dust allergy (disorder)
- Uncategorized - Allergy to animal hair (disorder)

Current Medications
- Vicodin; Date: 01/01/2008; Sig:
Image result stored in EHR
Clinical Workflow
Types of attribute scales

- Nominal Scale
- Ordinal Scale
- Numerical Scale
  - Ratio Scale
  - Interval Scale
Nominal Scale

• The values of the attribute are only “labels”, which is used to distinguish each other
  – Finite number of values
  – No order information
  – No algebraic operation could be conducted, except those related to frequency

• An example
  – \{1,2,3\} -> \{doctor, nurse, pharmacist\}
  -> \{Medical Information Service, Clinical Trials Center, Breast Center\}
A table describing accesses of roles and users on a patient across time

<table>
<thead>
<tr>
<th>Accessed Role</th>
<th>Number of Accesses</th>
<th>Number of Users</th>
<th>Time Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>doctor</td>
<td>10</td>
<td>3</td>
<td>t1-t2</td>
</tr>
<tr>
<td>nurse</td>
<td>20</td>
<td>5</td>
<td>t2-t3</td>
</tr>
<tr>
<td>pharmacist</td>
<td>15</td>
<td>6</td>
<td>t3-t4</td>
</tr>
<tr>
<td>doctor</td>
<td>16</td>
<td>4</td>
<td>t4-t5</td>
</tr>
<tr>
<td>nurse</td>
<td>22</td>
<td>5</td>
<td>t5-t6</td>
</tr>
<tr>
<td>nurse</td>
<td>6</td>
<td>2</td>
<td>t6-t7</td>
</tr>
</tbody>
</table>

Nominal Attribute

Frequency of different values in nominal attribute

- **doctor**: 10 accesses, 3 users, time span t1-t2
- **nurse**: 20 accesses, 5 users, time span t2-t3
- **pharmacist**: 15 accesses, 6 users, time span t3-t4
- **doctor**: 16 accesses, 4 users, time span t4-t5
- **nurse**: 22 accesses, 5 users, time span t5-t6
- **nurse**: 6 accesses, 2 users, time span t6-t7
Ordinal Scale

• The values of the attribute is to indicate certain ordering relationship resided in the attribute
  – Order is more important than value
  – No algebraic operation could be conducted, except those related to sorting

• An example
  – Richter’s scale on earthquake
A heartquake of 5.5 magnitudes is more important than one of 3 but less than one 9. Which indicates that there is an order between the data

<table>
<thead>
<tr>
<th>Richter magnitudes</th>
<th>Description</th>
<th>Earthquake effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 2.0</td>
<td>Micro</td>
<td>Micro earthquakes, not felt.</td>
</tr>
<tr>
<td>2.0-2.9</td>
<td>Minor</td>
<td>Generally not felt, but recorded.</td>
</tr>
<tr>
<td>3.0-3.9</td>
<td>Minor</td>
<td>Often felt, but rarely causes damage.</td>
</tr>
<tr>
<td>4.0-4.9</td>
<td>Light</td>
<td>Noticeable shaking of indoor items, rattling noises. Significant damage unlikely.</td>
</tr>
<tr>
<td>5.0-5.9</td>
<td>Moderate</td>
<td>Can cause major damage to poorly constructed buildings over small regions. At most slight damage to well-designed buildings.</td>
</tr>
<tr>
<td>6.0-6.9</td>
<td>Strong</td>
<td>Can be destructive in areas up to about 160 kilometers (100 mi) across in populated areas.</td>
</tr>
<tr>
<td>7.0-7.9</td>
<td>Major</td>
<td>Can cause serious damage over larger areas.</td>
</tr>
<tr>
<td>8.0-8.9</td>
<td>Great</td>
<td>Can cause serious damage in areas several hundred miles across.</td>
</tr>
<tr>
<td>9.0-9.9</td>
<td>Great</td>
<td>Devastating in areas several thousand miles across.</td>
</tr>
<tr>
<td>10.0+</td>
<td>Epic</td>
<td>Never recorded; see below for equivalent seismic energy yield.</td>
</tr>
</tbody>
</table>
Despite having the same interval of 0.5, the difference from one point to another in the scale in Joule is not uniform.

<table>
<thead>
<tr>
<th>Richter Approximate Magnitude</th>
<th>Joule equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>63.1 kJ</td>
</tr>
<tr>
<td>0.5</td>
<td>355 kJ</td>
</tr>
<tr>
<td>1.0</td>
<td>2.00 MJ</td>
</tr>
<tr>
<td>1.5</td>
<td>11.2 MJ</td>
</tr>
<tr>
<td>2.0</td>
<td>63.1 MJ</td>
</tr>
<tr>
<td>2.5</td>
<td>355 MJ</td>
</tr>
<tr>
<td>3.0</td>
<td>2.00 GJ</td>
</tr>
<tr>
<td>3.5</td>
<td>11.2 GJ</td>
</tr>
</tbody>
</table>
Numerical Scale

• The values of the attribute is to indicate quantity of some predefined unit
  – There should be a basic unit, which can be transferred to another one
  – The value is how many copies of the basic unit
  – Some algebraic operations could be conducted

• Two types of numerical scale
  – Interval scale
  – Ratio scale
Differences of Ratio Scale and Interval Scale

• An interval variable is a measurement **where the difference between two values is meaningful.**
  – The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees.

• A ratio variable, **has all the properties of an interval variable, and also has a clear definition of 0.0.** When the variable equals 0.0, there is none of that variable.
  – Variables like number of accesses on a patient, number of accesses of a user are ratio variables.
  – Temperature, expressed in F or C, is not a ratio variable. A temperature of 0.0 on either of those scales does not mean 'no temperature'.
  – However, temperature in Kelvin is a ratio variable, as 0.0 Kelvin really does mean 'no temperature'.
<table>
<thead>
<tr>
<th>Accessed Role</th>
<th>Number of Accesses</th>
<th>Number of Users</th>
<th>Temperature (F)</th>
<th>Temp (C)</th>
<th>Time Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>doctor</td>
<td>10</td>
<td>3</td>
<td>100.2</td>
<td>38</td>
<td>t1-t2</td>
</tr>
<tr>
<td>nurse</td>
<td>20</td>
<td>5</td>
<td>99.3</td>
<td>37</td>
<td>t2-t3</td>
</tr>
<tr>
<td>pharmacist</td>
<td>15</td>
<td>6</td>
<td>99.1</td>
<td>37</td>
<td>t3-t4</td>
</tr>
<tr>
<td>doctor</td>
<td>16</td>
<td>4</td>
<td>98.2</td>
<td>37</td>
<td>t4-t5</td>
</tr>
<tr>
<td>nurse</td>
<td>22</td>
<td>5</td>
<td>97.5</td>
<td>36</td>
<td>t5-t6</td>
</tr>
<tr>
<td>nurse</td>
<td>6</td>
<td>2</td>
<td>97.8</td>
<td>36</td>
<td>t6-t7</td>
</tr>
</tbody>
</table>

**Ratio Variable**

- Mean: 14.8333
- STD: 6.01383

**Interval Variable**

- Mean: 98.68333
- STD: 1.026483

**Covariance**

- Value: 5.86111
<table>
<thead>
<tr>
<th>Algebraic Operation</th>
<th>Nominal</th>
<th>Ordinal</th>
<th>Interval</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency distribution</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Median and Percentiles</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Add or Subtract</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mean, Standard Deviation</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ratio, or Coefficient of Variation</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
What we need to know to conduct access logs auditing?

- Data Representation
- Data Normalization and Discretization
- Similarity Measurements
- Dimensionality Reduction
Why we need data normalization?

15,000  3.5  1.2

<table>
<thead>
<tr>
<th>User</th>
<th>Access</th>
<th>Age</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25,000</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>40,000</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>55,000</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>27,000</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>53,000</td>
<td>30</td>
<td>5</td>
</tr>
</tbody>
</table>

Euclidean distance

Min-Max Normalization

<table>
<thead>
<tr>
<th>User</th>
<th>Access</th>
<th>Age</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000033</td>
<td>0.100</td>
<td>0.200</td>
</tr>
<tr>
<td>2</td>
<td>0.500000</td>
<td>0.400</td>
<td>0.400</td>
</tr>
<tr>
<td>3</td>
<td>1.000000</td>
<td>0.900</td>
<td>0.800</td>
</tr>
<tr>
<td>4</td>
<td>0.070000</td>
<td>0.200</td>
<td>0.400</td>
</tr>
<tr>
<td>5</td>
<td>0.930000</td>
<td>0.700</td>
<td>0.400</td>
</tr>
</tbody>
</table>

(a) Euclidean distance before normalization

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>15000</td>
<td>30000</td>
<td>2000</td>
</tr>
<tr>
<td>2</td>
<td>15000</td>
<td>0</td>
<td>15000</td>
<td>13000</td>
</tr>
<tr>
<td>3</td>
<td>30000</td>
<td>15000</td>
<td>0</td>
<td>28000</td>
</tr>
<tr>
<td>4</td>
<td>2000</td>
<td>13000</td>
<td>28000</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>28000</td>
<td>13000</td>
<td>2000</td>
<td>26000</td>
</tr>
</tbody>
</table>

(b) Euclidean distance after normalization

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000000</td>
<td>0.616414</td>
<td>1.414166</td>
<td>0.233332</td>
</tr>
<tr>
<td>2</td>
<td>0.616414</td>
<td>0.000000</td>
<td>0.812383</td>
<td>0.477234</td>
</tr>
<tr>
<td>3</td>
<td>1.414166</td>
<td>0.812383</td>
<td>0.000000</td>
<td>1.233286</td>
</tr>
<tr>
<td>4</td>
<td>0.233332</td>
<td>0.477234</td>
<td>1.233286</td>
<td>0.000000</td>
</tr>
<tr>
<td>5</td>
<td>1.127384</td>
<td>0.527022</td>
<td>0.452154</td>
<td>1.000505</td>
</tr>
</tbody>
</table>
Three types of normalization

• The attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, 0.0 to 1.0
  – Min-max normalization
  – Z-score normalization
  – Decimal scaling normalization
Min-Max Normalization

• Performs a linear transformation on the original data

• Support: $\text{min}_A$ and $\text{max}_A$ are the minimum and maximum values of an attribute, $A$.

• Min-max normalization maps a value, $v$, of $A$ to $v'$ in the range $[\text{new}_\text{min}_A, \text{new}_\text{max}_A]$ by computing:

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} \times (\text{new}_\text{max}_A - \text{new}_\text{min}_A) + \text{new}_\text{min}_A$$
An Example of Min-Max Normalization

- Let *income range* $12,000 to $98,000 *normalized to* [0.0, 1.0].
- Then $73,000 is mapped to

\[
\frac{73,600 - 12,000}{98,000 - 12,000} = \frac{61,600}{86,000} = 0.716
\]

\[
= (1.0 - 0) + 0 = 0.716
\]
Z-Score Normalization

• Change the original data quite a bit
• The values for an attribute, A, are normalized based on the mean (\( \bar{A} \)) and standard deviation (\( \sigma_A \)) of A.
• A value, \( v \), of A is normalized to \( v' \) by computing:

\[
v' = \frac{v - \bar{A}}{\sigma_A}
\]
distance is represented by units of standard deviations from the mean

<table>
<thead>
<tr>
<th>Accessed Role</th>
<th>Number of Accesses</th>
<th>Number of Users</th>
<th>Temperature (F)</th>
<th>Temp (C)</th>
<th>Time Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>docotr</td>
<td>10</td>
<td>3</td>
<td>100.2</td>
<td>38</td>
<td>t1-t2</td>
</tr>
<tr>
<td>nurse</td>
<td>20</td>
<td>5</td>
<td>99.3</td>
<td>37</td>
<td>t2-t3</td>
</tr>
<tr>
<td>pharmacist</td>
<td>15</td>
<td>6</td>
<td>99.1</td>
<td>37</td>
<td>t3-t4</td>
</tr>
<tr>
<td>doctor</td>
<td>16</td>
<td>4</td>
<td>98.2</td>
<td>37</td>
<td>t4-t5</td>
</tr>
<tr>
<td>nurse</td>
<td>22</td>
<td>5</td>
<td>97.5</td>
<td>36</td>
<td>t5-t6</td>
</tr>
<tr>
<td>nurse</td>
<td>6</td>
<td>2</td>
<td>97.8</td>
<td>36</td>
<td>t6-t7</td>
</tr>
</tbody>
</table>

\[ v' = \frac{v - \bar{A}}{\sigma_A} \]

-0.8037  0.8591  0.0277  0.1940  1.1917  -1.4688

The role nurse during time t2-t3 has accesses above average— a distance of 0.8591 above the average accesses.
Proportion within each sector

Z-Score  -4  -3  -2  -1  0  +1  +2  +3  +4
Centile  0.003% 0.13% 2.3% 15.9% 50% 84.1% 97.7% 99.87% 99.997%

-0.8037  0.8591  0.0277  0.1940  1.1917  -1.4688
Decimal Scaling Normalization

• normalizes by moving the decimal point of values of attribute A.
• The number of decimal points moved depends on the maximum absolute value of A.
• A value, v, of A is normalized to $v'$ by computing:

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\text{Max}(|v'|) < 1$
An Example of Decimal Scaling

- Suppose that the recorded values of A range from -986 to 917.
- The maximum absolute value of A is 986.
- To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., $j = 3$) so that

\[
\text{Normalized value} = \frac{\text{Original value}}{1,000}
\]
Data Discretization

• Dividing the range of a continuous attribute into intervals
• Interval labels can then be used to replace actual data values
• Reduce the number of values for a given continuous attribute
Why Data Discretization?

• Some classification algorithms only accept categorical attributes
  – many learning methods –like association rules, Bayesian networks can handle only discrete attributes

• This leads to a concise, easy-to-use, knowledge-level representation of mining results

The goal of discretization is to reduce the number of values for a continuous attribute

  grouping them into a number, n, of intervals (bins).
Typical Methods of Discretization

- **Binning**
  - Top-down split, unsupervised

- **Clustering analysis**
  - Either top-down split or bottom-up merge, Supervised

- **Interval merging by $\chi^2$ Analysis**
  - Supervised, bottom-up merge
Illustration of the supervised vs. unsupervised discretization
Binning

- The sorted values are distributed into a number of buckets, or bins, and then replacing each bin value by the bin mean or median
- Binning is a top-down splitting technique based on a specified number of bins
- Binning is an unsupervised discretization technique, because it does not use class information
Two Methods of Binning

• **Equal-width (distance) partitioning**
  – Divides the range into N intervals of equal size
  – if A and B are the lowest and highest values of the attribute, the width of intervals will be: \( W = (B - A)/N \)
  – The most straightforward, but outliers may dominate presentation

• **Equal-depth (frequency) partitioning**
An Example of Equal-width Partitioning

- Sorted data for price (in dollars):
  - 4, 8, 15, 21, 21, 24, 25, 28, 34
- \( W = \frac{(B - A)}{N} = \frac{(34 - 4)}{3} = 10 \)
  - Bin 1: 4-14, Bin 2: 15-24, Bin 3: 25-34
- Equal-width (distance) partitioning:
  - Bin 1: 4, 8
  - Bin 2: 15, 21, 21, 24
  - Bin 3: 25, 28, 34
Distribution of users in hospital on two measurements

Size of 1-nearest neighbor network

Equal Width Partitioning

Size of the network is 6
• Equal-depth (frequency) partitioning
  – Divides the range into N intervals, each containing approximately same number of samples
  – Good data scaling

• Example
  – Sorted data for price (in dollars):
    • 4, 8, 15, 21, 21, 24, 25, 28, 34
  – Equal-depth (frequency) partitioning:
    • Bin 1: 4, 8, 15
    • Bin 2: 21, 21, 24
    • Bin 3: 25, 28, 34
Typical Methods of Discretization

• Binning
  – Top-down split, unsupervised

• Clustering analysis
  – Either top-down split or bottom-up merge, Supervised

• Interval merging by $\chi^2$ Analysis
  – Supervised, bottom-up merge
Clustering Analysis

• A clustering algorithm can be applied to discretize a numerical attribute, A, by partitioning the values of A into clusters or groups.

• Clustering takes the distribution of A into consideration, as well as the closeness of data points, and therefore is able to produce high-quality discretization results.
Generate a Concept Hierarchy for Attribute A

• By following either a **top-down** splitting strategy or a **bottom-up** merging strategy, where each cluster forms a node of the concept hierarchy
  – In the former, each initial cluster or partition may be further decomposed into several sub-clusters, forming a lower level of the hierarchy
  – In the latter, clusters are formed by repeatedly grouping neighboring clusters in order to form higher level concepts.
Example: K-means Clustering Discretization

K-means clustering is an iterative method of finding clusters in multidimensional data; the user must define:

- number of clusters for each feature
- similarity function
- performance index and termination criterion
K-means Clustering Discretization

Example:

cluster centers

interval’s boundaries/midpoints (min value, midpoints, max value)
• **The clustering must be done for all attribute values for each class separately.**
  The final boundaries for this attribute will be all of the boundaries for all the classes.

• **Specifying the number of clusters is the most significant factor influencing the result of discretization:**
  to select the proper number of clusters, we cluster the attribute into several intervals (clusters), and then calculate some measure of goodness of clustering to choose the most “correct” number of clusters
Typical Methods of Discretization

• Binning
  – Top-down split, unsupervised

• Clustering analysis
  – Either top-down split or bottom-up merge, Supervised

• Interval merging by $\chi^2$ Analysis
  – Supervised, bottom-up merge
Chi Merge

• It is a bottom-up method
• Find the best neighboring intervals and merge them to form larger intervals recursively
• The method is supervised in that it uses class information
• The basic notion is that for accurate discretization, the relative class frequencies should be fairly consistent within an interval
• Therefore, if two adjacent intervals have a very similar distribution of classes, then the intervals can be merged. Otherwise, they should remain separate
ChiMerge Technique Example

<table>
<thead>
<tr>
<th>Sample</th>
<th>Feature</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>39</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>59</td>
<td>1</td>
</tr>
</tbody>
</table>

• Interval points for feature F are: 0, 2, 5, 7.5, 8.5, 9.5, etc.

<table>
<thead>
<tr>
<th>Interval [7.5, 8.5]</th>
<th>K=1</th>
<th>K=2</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_{11} = 1</td>
<td></td>
<td></td>
<td>R_1 = 1</td>
</tr>
<tr>
<td>A_{12} = 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interval [8.5, 9.5]</td>
<td></td>
<td></td>
<td>R_2 = 1</td>
</tr>
<tr>
<td>A_{21} = 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A_{22} = 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Σ</td>
<td>C_1 = 2</td>
<td>C_2 = 0</td>
<td>N = 2</td>
</tr>
</tbody>
</table>
ChiMergeTechnique (Example)

<table>
<thead>
<tr>
<th>Interval [7.5, 8.5]</th>
<th>( K=1 )</th>
<th>( K=2 )</th>
<th>( \Sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_{11}=1 )</td>
<td>( A_{12}=0 )</td>
<td>( R_1=1 )</td>
<td></td>
</tr>
<tr>
<td>( A_{21}=1 )</td>
<td>( A_{22}=0 )</td>
<td>( R_2=1 )</td>
<td></td>
</tr>
<tr>
<td>( \Sigma )</td>
<td>( C_1=2 )</td>
<td>( C_2=0 )</td>
<td>( N=2 )</td>
</tr>
</tbody>
</table>

Based on the table’s values, we can calculate expected values:

\[
E_{11} = \frac{2}{2} = 1, E_{12} = \frac{0}{2} \approx 0.1, \\
E_{21} = \frac{2}{2} = 1, \text{ and } E_{22} = \frac{0}{2} \approx 0.1 \\
\]

And corresponding \( \chi^2 \) score:

\[
\chi^2 = \frac{(1-1)^2}{1} + \frac{(0-0.1)^2}{0.1} + \frac{(1-1)^2}{1} + \frac{(0-0.1)^2}{0.1} = 0.2
\]

For the degree of freedom \( d=1 \), and \( \chi^2 = 0.2 < 2.706 \) (P value 0.90)

\[
DF = (\text{rowNum}-1) \times (\text{colNum}-1) = (2-1) \times (2-1)
\]
The table below gives the value $x_0^2$ for which $P[x^2 < x_0^2] = P$ for a given number of degrees of freedom and a given value of $P$.

<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>Values of P</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td>--------------------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>0.005</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
</tr>
</tbody>
</table>
ChiMergeTechnique (Example)

<table>
<thead>
<tr>
<th>Interval [0, 10.0]</th>
<th>K=1</th>
<th>K=2</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$A_{11} = 4$</td>
<td>$A_{12} = 1$</td>
<td>$R_1 = 5$</td>
</tr>
<tr>
<td>Interval [10.0, 42.0]</td>
<td>$A_{21} = 1$</td>
<td>$A_{22} = 3$</td>
<td>$R_2 = 4$</td>
</tr>
<tr>
<td>Σ</td>
<td>$C_1 = 5$</td>
<td>$C_2 = 4$</td>
<td>$N = 9$</td>
</tr>
</tbody>
</table>

- $E_{11} = 2.78$, $E_{12} = 2.22$, $E_{21} = 2.22$, $E_{22} = 1.78$, and $\chi^2 = 2.72 > 2.706$ (NO MERGE !)

- Final discretization: [0, 10], [10, 42], and [42, 60]
The ChiMerge Method

• Initially, each distinct value of a numerical attribute A is considered to be one interval
• Chi-Square tests are performed for every pair of adjacent intervals
• Adjacent intervals with the least Chi-Square values are merged together, since low Chi-Square values for a pair indicate similar class distributions
• This merge process proceeds recursively until a predefined stopping criterion is met (such as significance level, max interval, max inconsistency, etc.)
What we need to know to conduct access logs auditing?

• Data Representation
• Data Normalization and Discretization
• Similarity Measurements
• Dimensionality Reduction
Why care about similarity?

• Represent the internal relationship between data objects

• It is essential to many data mining algorithms
Distance Measurements

• Distance measure can be used to characterize the concept of “similarity”

• Distance or Metric should satisfy
  – Non-negativity: \( d(i, j) \geq 0 \) and \( d(i, j) = 0 \) iff \( i = j \)
  – Symmetry \( d(i, j) = d(j, i) \) for all \( i, j \)
  – Triangle inequality

\[
d(i, j) \leq d(i, k) + d(k, j) \quad \text{for all } i, j \text{ and } k
\]
Minkowski Distance

• Minkowski Distance is a generalization of Euclidean Distance

\[
dist = \left( \sum_{k=1}^{d} |p_k - q_k|^r \right)^{\frac{1}{r}}
\]

Where \( r \) is a parameter, \( d \) is the number of dimensions (attributes) and \( p_k \) and \( q_k \) are, respectively, the kth attributes (components) of data objects \( p \) and \( q \).
• $r = 1$. City block (Manhattan, taxicab, $L_1$ norm) distance

• $r = 2$. Euclidean distance

• $r \to \infty$. “supremum” ($L_{\text{max}}$ norm, $L_\infty$ norm) distance
  – This is the maximum difference between any component of the vectors
Euclidean Distance

• Euclidean Distance

\[ dist = \sqrt{\sum_{k=1}^{d} (p_k - q_k)^2} \]

Normalization is necessary, if scales differ.
Euclidean Distance-Examples

Distance Matrix

<table>
<thead>
<tr>
<th>point</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>p2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>p3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>p4</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\text{Dis}(p1,p2)=\sqrt{(0-2)^2+(2-0)^2} = \sqrt{8} = 2.828
\]
Weighted Minkowski distance

\[ d(x_1, x_2) = \left( \sum_{k=1}^{d} w_k (x_1(k) - x_2(k))^p \right)^{\frac{1}{p}} \]

Reflects the importance of each attribute

In both weighted and unweighted Minkowski distance, each attribute contribute independently to the measure of distance
Mahalanobis Distance

- Mahalanobis distance standardizes data not only in the direction of each attributes but also based on the covariance between attributes

\[
\text{mahalanobis}(p, q) = \sqrt{\left( p - q \right) \Sigma^{-1} \left( p - q \right)^T}
\]

Where \( p \) and \( q \) are two data points in \( d \) dimensions

\( \Sigma \) is the covariance matrix of the input data \( X \), the size of it is \( d \) by \( d \). "\( d \)" is the number of attributes or variables

\[
\Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} \left( X_{ij} - \bar{X}_j \right) \left( X_{ik} - \bar{X}_k \right)
\]
An Example of Mahalanobis Distance for Two Data Points

Step1: There are three data points in two attributes
A: (0.5, 0.5);  
B: (0, 1);  
C: (1.5, 1.5)

Step2: Covariance Matrix Calculation for two variables

\[
\Sigma = \begin{bmatrix}
0.58 & 0.25 \\
0.25 & 0.25 \\
\end{bmatrix} \quad \Sigma^{-1} = \begin{bmatrix}
3 & -3 \\
-3 & 7 \\
\end{bmatrix}
\]

Step3: Distance calculation

\[
Mahal(A, B) = \begin{bmatrix}
0.5 & 0 \\
0.5 & 1 \\
\end{bmatrix} \begin{bmatrix}
3 & 0.5 \\
-3 & -0.5 \\
\end{bmatrix} \begin{bmatrix}
3 & -3 \\
-3 & 7 \\
\end{bmatrix} = 4
\]

\[
\text{mahalanobis}(p, q) = (p - q) \Sigma^{-1} (p - q)^T
\]
An Example of Mahalanobis Distance for One Data Point

Step 1: A single observation $x(410, 400)$ in two dimensions: $d_1$ (named X) and $d_2$ (named Y)

Step 2: Covariance Matrix Calculation of dataset

For $d_1$, the mean value is 500, and standard deviation is 79.32

Covariance Matrix of variables X, and Y on all dataset

Step 3: Distance calculation

$$
\sum_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)
$$
Given that Mahalanobis Distance $D^2 = (x - m)^T C^{-1} (x - m)$

$$(x - m) = \begin{pmatrix} 410 - 500 \\ 400 - 500 \end{pmatrix} = \begin{pmatrix} -90 \\ -100 \end{pmatrix}$$

$$C^{-1} = \begin{pmatrix} 6291.55737 & 3754.32851 \\ 3754.32851 & 6280.77066 \end{pmatrix}^{-1} = \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix}$$

Therefore $D^2 = (-90 \times \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix} \times \begin{pmatrix} -90 \\ -100 \end{pmatrix} = 1.825$
We can also draw actual ellipses at regions of constant Mahalanobis values.

- One interesting feature to note from this figure is that a Mahalanobis distance of 1 unit corresponds to 1 standard deviation along both primary axes of variance.
Similarity Between Binary Vectors

- Common situation is that objects, \( p \) and \( q \), have only binary attributes

- Compute similarities using the following quantities
  \[ M_{01} = \text{the number of attributes where } p \text{ was 0 and } q \text{ was 1} \]
  \[ M_{10} = \text{the number of attributes where } p \text{ was 1 and } q \text{ was 0} \]
  \[ M_{00} = \text{the number of attributes where } p \text{ was 0 and } q \text{ was 0} \]
  \[ M_{11} = \text{the number of attributes where } p \text{ was 1 and } q \text{ was 1} \]

- Simple Matching and Jaccard Distance/Coefficients
  \[ \text{SMC} = \frac{\text{number of matches}}{\text{number of attributes}} = \frac{M_{11} + M_{00}}{M_{01} + M_{10} + M_{11} + M_{00}} \]
  \[ J = \frac{\text{number of value-1-to-value-1 matches}}{\text{number of not-both-zero attributes values}} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \]
SMC versus Jaccard: Example

\[ p = 100000000000 \]
\[ q = 00000001001 \]

\[ M_{01} = 2 \quad \text{(the number of attributes where } p \text{ was 0 and } q \text{ was 1)} \]
\[ M_{10} = 1 \quad \text{(the number of attributes where } p \text{ was 1 and } q \text{ was 0)} \]
\[ M_{00} = 7 \quad \text{(the number of attributes where } p \text{ was 0 and } q \text{ was 0)} \]
\[ M_{11} = 0 \quad \text{(the number of attributes where } p \text{ was 1 and } q \text{ was 1)} \]

\[ SMC = \frac{M_{11} + M_{00}}{M_{01} + M_{10} + M_{11} + M_{00}} = \frac{0 + 7}{2 + 1 + 0 + 7} = 0.7 \]

\[ J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{0}{2 + 1 + 0} = 0 \]
Cosine Similarity

• If $d_1$ and $d_2$ are two document vectors, then
  \[
  \cos(d_1, d_2) = \frac{(d_1 \bullet d_2)}{||d_1|| \cdot ||d_2||},
  \]
  where $\bullet$ indicates vector dot product and $||d||$ is the length of vector $d$.

• Example:

  \[
  d_1 = \begin{bmatrix}
    3 & 2 & 0 & 5 & 0 & 0 & 0 & 2 & 0 & 0
  \end{bmatrix}
  \]
  \[
  d_2 = \begin{bmatrix}
    1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2
  \end{bmatrix}
  \]

  \[
  d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
  \]

  \[
  ||d_1|| = (3^2+2^2+0^2+0^2+5^2+0^2+0^2+2^2+0^2+0^2)^{0.5} = (42)^{0.5} = 6.481
  \]

  \[
  ||d_2|| = (1^2+0^2+0^2+0^2+0^2+0^2+1^2+1^2+0^2+2^2)^{0.5} = (6)^{0.5} = 2.245
  \]

  \[
  \cos(d_1, d_2) = .3150, \text{ distance}=1-\cos(d_1,d_2)
  \]
Distance between two values $x$ and $y$ of an attribute $a$ (*Nominal*)

- **Value Difference Metric (VDM)**- Classes based distance measurements

$$VDM_a(x, y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_a,x} - \frac{N_{a,y,c}}{N_a,y} \right|$$

- The number of instances in $T$ that have value $x$ for attribute $a$ and output class $c$
- The number of output classes
- Constant, usually $1$ or $2$
- The number of instances in the training set $T$ that have value $x$ for attribute $a$

For example, if an attribute *color* has three values *red, green and blue*, and the application is to identify whether or not an object is an apple, *red and green* would be considered closer than *red* and *blue* because the former two both have similar correlations with the output class *apple*. 
Complex Structure

• For distribution: KL divergence, cross entropy, ...

• For trees, graphs: defining graph kernels, ...
What we need to know to conduct access logs auditing?

• Data Representation
• Data Normalization and Discretization
• Similarity Measurements
• Dimensionality Reduction
Principal Component Analysis (PCA)

• summarization of data with many \((p)\) variables by a smaller set of \((k)\) derived (synthetic, composite) variables.

\[
\begin{align*}
\text{n} & \quad \text{A} \\
\text{p} & \\
\text{k} & \quad \text{n} \\
\end{align*}
\]
Principal Component Analysis (PCA)

• takes a data matrix of n objects by p variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original p variables

• the first k components display as much as possible of the variation among objects.
Geometric Rationale of PCA

• objects are represented as a cloud of n points in a multidimensional space with an axis for each of the p variables

• the centroid of the points is defined by the mean of each variable

• the variance of each variable is the average squared deviation of its n values around the mean of that variable.

\[ V_i = \frac{1}{n-1} \sum_{m=1}^{n} (X_{im} - \bar{X}_i)^2 \]
Geometric Rationale of PCA

• degree to which the variables are linearly correlated is represented by their covariance

\[ C_{ij} = \frac{1}{n-1} \sum_{m=1}^{n} (X_{im} - \bar{X}_i)(X_{jm} - \bar{X}_j) \]
Geometric Rationale of PCA

• objective of PCA is to rigidly rotate the axes of this p-dimensional space to new positions (principal axes) that have the following properties:
  – ordered such that principal axis 1 has the highest variance, axis 2 has the next highest variance, .... , and axis \( p \) has the lowest variance
  – covariance among each pair of the principal axes is zero (the principal axes are uncorrelated).
2D Example of PCA

- variables $X_1$ and $X_2$ have positive covariance & each has a similar variance.

$X_2 = 4.91$

$X_1 = 8.35$

$V_1 = 6.67$  
$V_2 = 6.24$  
$C_{1,2} = 3.42$
Configuration is Centered

- each variable is adjusted to a mean of zero (by subtracting the mean from each value).
Principal Components are Computed

- PC 1 has the highest possible variance (9.88)
- PC 2 has a variance of 3.03
- PC 1 and PC 2 have zero covariance.
An Example

<table>
<thead>
<tr>
<th>case</th>
<th>ht ($x_1$)</th>
<th>wt($x_2$)</th>
<th>age($x_3$)</th>
<th>sbp($x_4$)</th>
<th>heart rate ($x_5$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>175</td>
<td>1225</td>
<td>25</td>
<td>117</td>
<td>56</td>
</tr>
<tr>
<td>2</td>
<td>156</td>
<td>1050</td>
<td>31</td>
<td>122</td>
<td>63</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>202</td>
<td>1350</td>
<td>58</td>
<td>154</td>
<td>67</td>
</tr>
</tbody>
</table>

![PCA plot](image-url)
Thanks!