Data Mining in Cognitive Science

Chen Yu
Indiana University

Acknowledgement: Some of slides are adapted from the slides created by Drs. Balac, Tan, Steinbach, Kumar, Clifton, Tuchinda.

Data

• We are overwhelmed with data. The amount of data in the world, in our lives, seems to be go on and on increasing.
• Ubiquitous electronics record our decisions, our choices in the supermarket, our financial habits, and our comings and goings.
• The WWW overwhelms us with information, but meanwhile every choice we make is recorded.
• The amount of data stored in the world’s databases doubles every 20 months.

From data to information

• Most of the information is in its raw form: data.
• If data is characterized as recorded facts, then information is the set of patterns that underlie the data.

How to look for patterns

• Politicians seek patterns in voter opinions
• Lovers seek patterns in their partner’s responses.
• Entrepreneurs identify opportunities, patterns that can be turned into a profitable business, and exploit them.
• Hunters seek patterns in animal migration behavior.

• A scientist’s job is to make sense of data, to discover the patterns that govern how the physical world works and encapsulate them in theories that can be used for predicting what will happen in new situations.
Data and Knowledge Discovery

• There is a growing gap between the generation of data and our understanding of it.
• Potentially useful information is hidden in the data that is rarely made explicit so that we can take advantage of.
• Very little data will ever be looked at by a human.

Data Mining Definition

• Data mining is defined as the process of discovering patterns in data. The process must be automatic (or semiautomatic).
• The patterns discovered must be meaningful in that they lead to some advantage.

Data Mining Examples

• Grocery store
• NBA: Advanced Scout
  [http://www.springerlink.com/content/r32q737858160r97/](http://www.springerlink.com/content/r32q737858160r97/)
• Personalization & Customer Profiling
• Campaign Management

Necessity Is the Mother of Invention

• Data explosion problem
  – Automated data collection tools and mature database technology lead to tremendous amounts of data accumulated and/or to be analyzed in databases, data warehouses, and other information repositories
• We are drowning in data, but starving for knowledge!
• Solution: data mining
  – Mining interesting knowledge (rules, regularities, patterns, constraints) from data in large databases
Data Mining Tasks

• Prediction Methods
  – Use some variables to predict unknown or future values of other variables.

• Description Methods
  – Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks...

• Classification [Predictive]
• Clustering [Descriptive]
• Association Rule Discovery [Descriptive]
• Sequential Pattern Discovery [Descriptive]
• Regression [Predictive]
• Deviation Detection [Predictive]

Classification: Definition

• Given a collection of records (training set)
  – Each record contains a set of attributes, one of the attributes is the class.
• Find a model for the class attribute as a function of the values of other attributes.
• Goal: previously unseen records should be assigned a class as accurately as possible.
  – A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification

Given training examples of inputs and corresponding outputs \((x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\), produce the “correct” outputs for new inputs.
Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

- Intracluster distances are minimized
- Intercluster distances are maximized

Clustering can be ambiguous

Partitional clustering
Hierarchical clustering

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>

Rules Discovered:
- \{Milk\} \Rightarrow \{Coke\}
- \{Diaper, Milk\} \Rightarrow \{Beer\}

Definition of association rule

Example: \{Milk, Diaper\} \Rightarrow \{Beer\}

\[
\begin{align*}
\text{Support: } s &= \frac{\sigma(X \cup y)}{|T|} = \frac{2}{5} = 0.4 \\
\text{Confidence: } c &= \frac{\sigma(X \cup y)}{\sigma(X)} = \frac{2}{3} = 0.67
\end{align*}
\]
Results

Example of Rules:

- Milk, Diaper → Beer (s=0.4, c=0.67)
- Milk, Beer → Diaper (s=0.4, c=1.0)
- Diaper, Beer → (Milk) (s=0.4, c=0.67)
- (Beer) → Milk, Diaper (s=0.4, c=0.67)
- Diaper → (Milk, Beer) (s=0.4, c=0.5)
- Milk → (Diaper, Beer) (s=0.4, c=0.5)

Observations:
- All the rules above correspond to the same itemset (Milk, Diaper, Beer)
- Rules obtained from the same itemset have identical support but can have different confidence

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Studied in statistics.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Sequential Pattern Discovery: Definition

- Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events.
- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.
- How to deal with noisy data items.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection

Typical network traffic at University level may reach over 100 million connections per day.
Knowledge Discovery Definition

Knowledge Discovery in Data is the *non-trivial* process of identifying
- valid
- novel
- potentially useful
- and ultimately *understandable patterns* in data.

from Advances in Knowledge Discovery and Data Mining, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation
Steps of a KDD Process

- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 50% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
  - Pattern evaluation and knowledge presentation
    - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Data cleaning and preprocessing

- You didn’t collect it yourself.
- Data are not in the right format (e.g., with the sampling rate).
- People make mistakes (typos)
- People are busy (“this is good enough”)
Fourth paradigm: data-intensive science

- Much of the vast volume of scientific data captured by new instruments on a 24/7 basis, along with simulated data from computer models.
- The data is likely to reside forever in a substantially publically accessible, curated state for the purposes of continued analysis.
- This analysis will result in the development of many new theories.

Behavioral studies

- What we think people are doing
- What models tell us what people are doing
- What conclusions we can infer from empirical data collected from well-designed experiments
- What fine-grained data suggest, what patterns can be extracted from such data, and how those patterns relate to principles/theories of human cognition and learning.

Advances in sensing techniques

Data is Information-Rich

- Cognitive processes/mechanisms are revealed by real-time behaviors. E.g. eye movements.
Data reduction

- From data capture to data curation to data analysis
- The published results are just the tip of the data iceberg.

People collect a lot of data and then reduce this down to some number of column inches in Science or Nature.

Hunting the fox

- There are many different ways to reduce data and extract patterns.
  - there is a lot of data that is collected but not curated/coded or analyzed/published.
  - Rewarding: There are many opportunities to extract new patterns and infer new principles even from the same dataset.
  - Frustrating: There are a few ways to do it properly and probably many incorrect ways that lead to nothing.
  - There are also risks/pitfalls of making mistakes.

Example

- Eye gaze data based on Areas of Interest (AOIs) e.g. aabbcdddbccddabbdbbaddbb
- Different ways to look at this data:
  - Proportion of looking time on each region.
  - Average duration, overall or each region
  - # of times looking at each AOI
  - # of switches: transition matrix
  - Histogram of looking durations, overall or each region
  - Frequent sequential patterns
  - Patterns within a certain timing event, e.g. the first 5 seconds
  - Recalculating by filtering small/big events (e.g. short fixations).
  - Within or across participants
  - Grouping participants or instances across participants
  - Relative values vs. absolute values

…”

Extensions from the example

- What if we want to analyze raw coordinate data but not Areas of Interest (AOIs) – from categorical values to continuous values.
- What if the data has several dimensions at a single moment.
- What if we deal with a heterogeneous dataset with other measures.
Key issues

- At what stage to reduce/transform what aspects of the data, e.g. noisy items, from continuous to categorical, when to aggregate the data over time and to what degree.
- After doing that, being aware of limitation and simplification.
- Always go back to raw data (through information visualization).
- Always try to make sense of data (through human involvement).

Human-in-the-loop computation

- Two levels:
  - visualization/human perception
  - Theoretic-driven

Information Visualization

- Visualization uses human perception to recognize patterns in large data sets
- Advantages relative to data mining
  - Perceive "unconsidered" patterns
  - Recognize non-linear relationships
  - E.g. winDirStat
- Disadvantages relative to data mining
  - Hard to recognize "small" patterns
  - Difficult to quantify results

Data Mining and Visualization

- Approaches
  - Visualization to display results of data mining
    - Help analyst to better understand the results of the data mining tool
  - Visualization to aid the data mining process
    - Interactive control over the data exploration process
    - Interactive steering of analytic approaches
- Interactive data mining issues
  - Relationships between the analyst, the data mining tool and the visualization tool
Intelligent data analysis

- Pattern discovery
- Pattern interpretation: theoretical-driven
- Interactive and incremental
- Intelligible data analysis

This course

- Information processing and visualization
- Data mining and machine learning algorithms, focusing on temporal data mining
- Data mining engineering: e.g. data cleaning and data preprocessing

Three Flavors

- Basic math
- Computer skills
  - using matlab to write some simple functions
  - compiling and running some existing software systems.
- Human cognition Theories

Assignment Type I

- Commenting on papers
  Three parts/slides:
  - what are most compelling ideas/techniques in a paper.
  - what are new challenges based on the present work.
  - how to use (and extend) the ideas/techniques in the paper in your data/research, what research topics would be ideal for the techniques in the paper, and what aspects of the data the present techniques cannot handle.
Assignment Type 2

- Presenting new papers
  - Introducing the overall idea of a paper.
  - Leading the discussion.
  - Demonstrating the software with some dataset.

Assignment Type 3

- Installing and running some existing data mining/visualization software applications
- Feeding your data into such system to generate results.

Assignment Type 4

- Writing some Matlab functions to data-mine your data and explore new patterns from the data.

Final Project

- Using what you’ve learned in the class to data-mine a dataset.
- Writing a research report that can be extended to conference/journal submissions.
Grading

• Weekly Assignments: 50%
• Final Project: 30%
• Discussion and Participation: 20%

• http://www.indiana.edu/~dll/p657_dm.html