Data Informatics

Seon Ho Kim, Ph.D.
seonkim@usc.edu
Data Mining
What Is Data Mining?

• Data mining
  – Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

• Alternative names
  – Knowledge Discovery (mining) in Databases (KDD), knowledge discovery from data, knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
Data Mining—What’s in a Name?

The process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of stored data, using pattern recognition technologies and statistical and mathematical techniques.
Integration of Multiple Technologies

- Machine Learning
- Artificial Intelligence
- Database Management
- Statistics
- Visualization
- Algorithms
- Data Mining
Terms

• **Machine Learning** relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed (definition of Arthur Samuel).

• **Data Mining** can be defined as the process that starting from apparently unstructured data tries to extract knowledge and/or unknown interesting patterns. During this process machine Learning algorithms are used.

• **Data Analysis, Data Mining, Machine Learning and Mathematical Modeling** are **tools**: means towards an end.

• **Analytics, Business Intelligence, Econometrics and Artificial Intelligence** are **application areas**: domains that use the tools above (and others) to produce results within its subject. Among them, Analytics is probably a more generic term (i.e. non domain-specific).

• **Statistics** is a **branch** of Mathematics providing theoretical and practical support to the above tools.

• **Data Science** is a catch-all term to describe using those all tools to provide answers in those all areas (and also in others), specially when dealing with **Big Data**, which is nothing more than a label meaning doing any of the above but when the datasets are huge.
Knowledge Discovery in Databases: Process

adapted from:
Multi-Dimensional View of Data Mining

• Data to be mined
  – Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW

• Knowledge to be mined
  – Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  – Multiple/integrated functions and mining at multiple levels

• Techniques utilized
  – Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.

• Applications adapted
  – Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, Web mining, etc.
“In order to discover anything, you must be looking for something.” Laws of Serendipity
Data Mining: History of the Field

• Knowledge Discovery in Databases workshops started ‘89
  – Now a conference under the auspices of ACM SIGKDD
  – IEEE conference series started 2001

• Key founders / technology contributors:
  – Usama Fayyad, JPL (then Microsoft, then his own company, Digimine, now Yahoo! Research labs)
  – Gregory Piatetsky-Shapiro (then GTE, now his own data mining consulting company, Knowledge Stream Partners)
  – Rakesh Agrawal (IBM Research)

• The term “data mining” has been around since at least 1983 – in the statistics community
Why Data Mining?
Potential Applications

• Data analysis and decision support
  – Market analysis and management
    • Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  – Risk analysis and management
    • Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  – Fraud detection and detection of unusual patterns (outliers)

• Other Applications
  – Text mining (news group, email, documents) and Web mining
  – Stream data mining
  – DNA and bio-data analysis
Market Analysis and Management

• Where does the data come from?
  – Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
• Target marketing
  – Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  – Determine customer purchasing patterns over time
• Cross-market analysis
  – Associations/co-relations between product sales, & prediction based on such association
• Customer profiling
  – What types of customers buy what products (clustering or classification)
• Customer requirement analysis
  – identifying the best products for different customers
  – predict what factors will attract new customers
• Provision of summary information
  – multidimensional summary reports
  – statistical summary information (data central tendency and variation)
Corporate Analysis & Risk Management

• Finance planning and asset evaluation
  – cash flow analysis and prediction
  – contingent claim analysis to evaluate assets
  – cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)

• Resource planning
  – summarize and compare the resources and spending

• Competition
  – monitor competitors and market directions
  – group customers into classes and a class-based pricing procedure
  – set pricing strategy in a highly competitive market
Fraud Detection & Mining Unusual Patterns

• Approaches: Clustering & model construction for frauds, outlier analysis
• Applications: Health care, retail, credit card service, telecomm.
  – Auto insurance: ring of collisions
  – Money laundering: suspicious monetary transactions
  – Medical insurance
    • Professional patients, ring of doctors, and ring of references
    • Unnecessary or correlated screening tests
  – Telecommunications: phone-call fraud
    • Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  – Retail industry
    • Analysts estimate that 38% of retail shrink is due to dishonest employees
  – Anti-terrorism
Other Applications

• Sports
  – IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat

• Astronomy
  – JPL and the Palomar Observatory discovered 22 quasars with the help of data mining

• Internet Web Surf-Aid
  – IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.
Example: Use in retailing

- Goal: Improved business efficiency
  - Improve marketing (advertise to the most likely buyers)
  - Inventory reduction (stock only needed quantities)
- Information source: Historical business data
  - Example: Supermarket sales records

<table>
<thead>
<tr>
<th>Date/Time/Register</th>
<th>Fish</th>
<th>Turkey</th>
<th>Cranberries</th>
<th>Wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/6 13:15 2</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>12/6 13:16 3</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

- Size ranges from 50k records (research studies) to terabytes (years of data from chains)
- Data is already being warehoused

- Sample question – what products are generally purchased together?
- The answers are in the data, if only we could see them
Data Mining applied to Aviation Safety Records (Eric Bloedorn)

- Many groups record data regarding aviation safety including the National Transportation Safety Board (NTSB) and the Federal Aviation Administration (FAA)
- Integrating data from different sources as well as mining for patterns from a mix of both structured fields and free text is a difficult task
- The goal of our initial analysis is to determine how data mining can be used to improve airline safety by finding patterns that predict safety problems
Aircraft Accident Report

• This data mining effort is an extension of the FAA Office of System Safety’s Flight Crew Accident and Incident Human Factors Project

• In this previous approach two database-specific human error models were developed based on general research into human factors
  – FAA’s Pilot Deviation database (PDS)
  – NTSB’s accident and incident database

• These error models check for certain values in specific fields

• Result
  – Classification of some accidents caused by human mistakes and slips.
Problem

- Current model cannot classify a large number of records
- A large percentage of cases are labeled ‘unclassified’ by current model
  - ~58,000 in the NTSB database (90% of the events identified as involving people)
  - ~5,400 in the PDS database (93% of the events)
- Approximately 80,000 NTSB events are currently labeled ‘unknown’
- Classification into meaningful human error classes is low because the explicit fields and values required for the models to fire are not being used
- Models must be adjusted to better describe data
Data mining Approach

• Use information from text fields to supplement current structured fields by extracting features from text in accident reports

• Build a human-error classifier directly from data
  – Use expert to provide class labels for events of interest such as ‘slips’, ‘mistakes’ and ‘other’
  – Use data-mining tools to build comprehensible rules describing each of these classes
Example Rule

• Sample Decision rule using current model features and text features
  - If (person_code_1b= 5150,4105,5100,4100) and
    ((crew-subject-of-intentional-verb = true) or
    (modifier_code_1b = 3114))
    then
    mistake
• “If pilot or copilot is involved and either the narrative, or the modifier code for 1b describes the crew as intentionally performing some action then this is a mistake”
Data Mining Ideas: Logistics

• Delivery delays
  – Debatable what data mining will do here; best match would be related to “quality analysis”: given lots of data about deliveries, try to find common threads in “problem” deliveries

• Predicting item needs
  – Seasonal
    • Looking for cycles, related to similarity search in time series data
    • Look for similar cycles between products, even if not repeated
  – Event-related
    • Sequential association between event and product order (probably weak)
What Can Data Mining Do?

• Cluster
• Classify
  – Categorical, Regression
• Summarize
  – Summary statistics, Summary rules
• Link Analysis / Model Dependencies
  – Association rules
• Sequence analysis
  – Time-series analysis, Sequential associations
• Detect Deviations
Clustering

- Find groups of similar data items
- Statistical techniques require some definition of “distance” (e.g. between travel profiles) while conceptual techniques use background concepts and logical descriptions
- Uses:
  - Demographic analysis
  - Technologies:
    - Self-Organizing Maps
    - Probability Densities
    - Conceptual Clustering
- “Group people with similar travel profiles”
  - George, Patricia
  - Jeff, Evelyn, Chris
  - Rob

Clusters
Classification

• Find ways to separate data items into pre-defined groups
  – We know X and Y belong together, find other things in same group
• Requires “training data”: Data items where group is known

Uses:
• Profiling

Technologies:
• Generate decision trees (results are human understandable)
• Neural Nets

• “Route documents to most likely interested parties”
  – English or non-english?
  – Domestic or Foreign?
Association Rules

• Identify dependencies in the data:
  – X makes Y likely
• Indicate significance of each dependency
• Use Example:
  – Targeted marketing
    – “Find groups of items commonly purchased together”
      – People who purchase fish are extraordinarily likely to purchase wine
      – People who purchase Turkey are extraordinarily likely to purchase cranberries

| Date/Time/Register | Fish | Turkey | Cranberries | Wine |...
|--------------------|------|--------|-------------|------|-----
| 12/6 13:15 2       | N    | Y      | Y           | Y    |...
| 12/6 13:16 3       | Y    | N      | N           | Y    |...  |
Sequential Associations

- Find event sequences that are unusually likely
- Requires “training” event list, known “interesting” events
- Must be robust in the face of additional “noise” events

Uses:
- Failure analysis and prediction

Technologies:
- Dynamic programming (Dynamic time warping)
- “Custom” algorithms

“Find common sequences of warnings/faults within 10 minute periods”
- Warn 2 on Switch C preceded by Fault 21 on Switch B
- Fault 17 on any switch preceded by Warn 2 on any switch

<table>
<thead>
<tr>
<th>Time</th>
<th>Switch</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>21:10</td>
<td>B</td>
<td>Fault 21</td>
</tr>
<tr>
<td>21:11</td>
<td>A</td>
<td>Warn 2</td>
</tr>
<tr>
<td>21:13</td>
<td>C</td>
<td>Warn 2</td>
</tr>
<tr>
<td>21:20</td>
<td>A</td>
<td>Fault 17</td>
</tr>
</tbody>
</table>
Deviation Detection

• Find unexpected values, outliers

Uses:
• Failure analysis
• Anomaly discovery for analysis

Technologies:
• clustering/classification methods
• Statistical techniques
• visualization

“Find unusual occurrences in IBM stock prices”

<table>
<thead>
<tr>
<th>Sample date</th>
<th>Event</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>58/07/04</td>
<td>Market closed</td>
<td>317 times</td>
</tr>
<tr>
<td>59/01/06</td>
<td>2.5% dividend</td>
<td>2 times</td>
</tr>
<tr>
<td>59/04/04</td>
<td>50% stock split</td>
<td>7 times</td>
</tr>
<tr>
<td>73/10/09</td>
<td>not traded</td>
<td>1 time</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
<th>Close</th>
<th>Volume</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>58/07/02</td>
<td>369.50</td>
<td>314.08</td>
<td>.022561</td>
</tr>
<tr>
<td>58/07/03</td>
<td>369.25</td>
<td>313.87</td>
<td>.022561</td>
</tr>
<tr>
<td>58/07/04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58/07/07</td>
<td>370.00</td>
<td>314.50</td>
<td>.022561</td>
</tr>
</tbody>
</table>
Data Mining Complications

• Volume of Data
  – Clever algorithms needed for reasonable performance

• Interest measures
  – How do we ensure algorithms select “interesting” results?

• “Knowledge Discovery Process” skill required
  – How to select tool, prepare data?

• Data Quality
  – How do we interpret results in light of low quality data?

• Data Source Heterogeneity
  – How do we combine data from multiple sources?
Major Issues in Data Mining

• Mining methodology
  – Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  – Performance: efficiency, effectiveness, and scalability
  – Pattern evaluation: the interestingness problem
  – Incorporation of background knowledge
  – Handling noise and incomplete data
  – Parallel, distributed and incremental mining methods
  – Integration of the discovered knowledge with existing one: knowledge fusion

• User interaction
  – Data mining query languages and ad-hoc mining
  – Expression and visualization of data mining results
  – Interactive mining of knowledge at multiple levels of abstraction

• Applications and social impacts
  – Domain-specific data mining & invisible data mining
  – Protection of data security, integrity, and privacy
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 60% 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 Mining and Business Intelligence

- Increasing potential to support business decisions

- Making Decisions
  - Data Presentation
    - Visualization Techniques
  - Data Mining
    - Information Discovery
  - Data Exploration
    - Statistical Analysis, Querying and Reporting
  - Data Warehouses / Data Marts
    - OLAP, MDA
  - Data Sources
    - Paper, Files, Information Providers, Database Systems, OLTP

- End User
- Business Analyst
- Data Analyst
- DBA
Architecture: Typical Data Mining System

- Graphical user interface
- Pattern evaluation
- Data mining engine
- Database or data warehouse server
- Knowledge-base

Data cleaning & data integration

Databases

Filtering

Data Warehouse
State of Commercial/Research Practice

• Increasing use of data mining systems in financial community, marketing sectors, retailing

• Still have major problems with large, dynamic sets of data (need better integration with the databases)
  – COTS data mining packages perform specialized learning on small subset of data

• Most research emphasizes machine learning; little emphasis on database side (especially text)

• People achieving results are not likely to share knowledge
Related Techniques: OLAP
On-Line Analytical Processing

• On-Line Analytical Processing tools provide the ability to pose statistical and summary queries interactively (traditional On-Line Transaction Processing (OLTP) databases may take minutes or even hours to answer these queries)

• Advantages relative to data mining
  – Can obtain a wider variety of results
  – Generally faster to obtain results

• Disadvantages relative to data mining
  – User must “ask the right question”
  – Generally used to determine high-level statistical summaries, rather than specific relationships among instances
Integration of Data Mining and Data Warehousing

• Data mining systems, DBMS, Data warehouse systems coupling
  – No coupling, loose-coupling, semi-tight-coupling, tight-coupling

• On-line analytical mining data
  – integration of mining and OLAP technologies

• Interactive mining multi-level knowledge
  – Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.

• Integration of multiple mining functions
  – Characterized classification, first clustering and then association