Introduction

CPS 296.3: Information Management and Mining

Jun Yang
Duke University
January 13, 2008

† Thanks to contents borrowed from Ullman (http://infolab.stanford.edu/~ullman/mining/mining.html), Clifton (http://www.cs.purdue.edu/homes/clifton/cs490d/), and Kumar (http://www-users.cs.umn.edu/~kumar/dmbook/)

Trend: Moore's Law reversed

- Moore's Law: *Processing power doubles every 18 months*
- Amount of data doubles every 9 months
  - Disk sales (# of bits) doubles every 9 months
  - Parkinson's Law: *Data expands to fill the space available for storage*
- Moore's Law reversed:

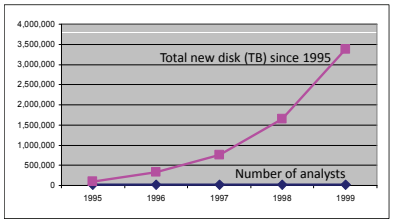
  Time to process all data doubles every 18 months!

- Does your attention span double every 18 months?
  - No, so we need better information management & mining

2

So, why mine data?

- Data explosion: advances in data collection and storage lead to tremendous amount of data waiting to be analyzed
  - Science, commerce, society, …
- We are drowning in data, but starving for knowledge!



Grossman, Kamath, Kumar, "Data Mining for Scientific and Engineering Applications"

3

Data mining

- Data → information → knowledge
- Discovery of useful, possibly unexpected, patterns in data
- Subsidiary issues
  - Data cleansing: detection of bogus data, e.g.:
    - age = 150, gpa = –10
    - Entity resolution: # of Jun Yang's contributing to a DBLP entry
  - Visualization: something better than megabytes of text
  - Data warehousing, parallel processing, etc.

4

What is (not) data mining?

- What is not data mining
  - Look up phone # by name in a phone directory
  - Search for Web pages about "Amazon" on a search engine
- What is data mining
  - Certain names are prevalent in some US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by a search engine—Amazon rainforest, Amazon.com, etc.

5

Origins and cultures

- Databases: concentrate on simple queries over large-scale data (that do not fit in main memory)
- AI/machine learning: concentrate on sophisticated methods (usually on smaller scales)
- Statistics: concentrate on models

6

DB vs. stats example

- To a DB person, data mining is an extreme from of analytical processing—queries that examine large amounts of data to return answers
  - E.g., given a billion numbers, a DB person would compute their average and standard deviation
- To a stats person, data mining is the inference of models that lets you learn about the model parameters
  - E.g., given a billion numbers, a stats person would fit them to the best Gaussian distribution and report the mean and standard deviation

➲Any difference?

7

Meaningfulness of answers

- A big risk with data mining is that you might "discover" patterns that are meaningless
- Statisticians call it Bonferroni's Principle
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap
  - Examples to follow:
    - Rhine Paradox
      - Anyone from CPS 216 remembers?
    - Prof. Ullman's TIA (Total Information Awareness) example

8

The TIA example

- Suppose we believe that certain groups of evil-doers are meeting occasionally in hotels to plot doing evil
- We want to find (unrelated) people who at least twice have stayed at the same hotel on the same day
- 10^9 (1 billion) people being tracked over 1000 days
- Each person stays in a hotel 1% of the time
  - 10 days out of 1000
- Hotels hold 100 people (so 10^5 hotels)

➲If everyone behaves randomly, i.e., no evil-doers, will the data mining detect anything suspicious?

9

Calculations

- Probability that persons p and q will be at the same hotel on day d:
  - $1/100 \times 1/100 \times 10^{-5} = 10^{-9}$
- Probability that p, q will be at the same hotel on given days d_1, d_2:
  - $10^{-9} \times 10^{-9} = 10^{-18}$
- Pairs of days:
  - 5×10^5
- Probability that p and q will be at the same hotel on some two days:
  - $(5 \times 10^5) \times 10^{-18} = 5 \times 10^{-13}$
- Pairs of people:
  - 5×10^{17}
- Expected number of "suspicious" pairs of people:
  - $(5 \times 10^{17}) \times (5 \times 10^{-13}) = 250{,}000$

10

Conclusion?

- Suppose there are (say) 10 pairs of evil-doers who definitely stayed at the same hotel twice
- Analysts have to sift through 250,010 candidates to find the 10 real cases
  - Not gonna happen!
- Moral: When looking for a property (e.g., "two people stayed at the same hotel twice"), make sure there are not so many possibilities that random data will surely produce facts "of interest!"

11

Course roadmap (tentative)

- The fundamentals—mostly lectures
  - Data warehousing
  - Association rules
  - Web indexing, ranking, spam
  - Similarity
  - Clustering
  - Classification
- Advanced stuff—mostly reading and discussing papers
  - Topics to be determined by you and me
  - Possibilities: parallel data analytics, more information retrieval/extraction, social network analysis, streaming data, or working on the new problem of *computational journalism*!

12

Misc. course information

- Web: http://www.cs.duke.edu/courses/spring09/cps296.3/
  - Lecture slides, tentative schedule, reading assignments, etc.
- Book: none required
  - Good reference: Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, 2nd ed.
- Mailing list: cps296.3@cs.duke.edu
  - Announcements, messages of general interest
- Time/location: let's work on that!
  - Proposal: Tuesdays only, 1:30-4pm
    - Ideas to make 2.5 hours less sleep-inducing?
  - Need to finalize today
- Office hours: TBD

13

Course load and grading

- Reading, discussion, and participation (50%)
  - Short reviews for reading assignments (20%)
  - Present and lead discussions, probably twice (20%)
  - Attendance (10%)
- Course project (50%)
  - Work on something that you'll have an urge to share with others!
  - Proposal presentation (15%) after spring recess
  - Short progress report (5%) in the first week of April
  - Final presentation (30%) during the final exam slot
- Class will meet during graduate reading period

➲More details on course website

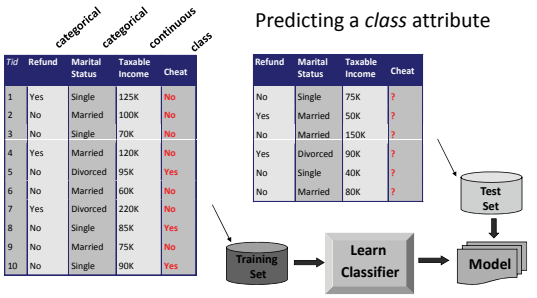
14

Data mining tasks

- Prediction tasks
  - Use some variables to predict unknown or future values of other variables
  - E.g.: classification, regression, deviation/anomaly detection, …
- Description tasks
  - Find human-interpretable patterns to describe the data
  - E.g.: association rules, clustering, pattern discovery, …

15

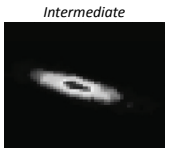
Classification example 1



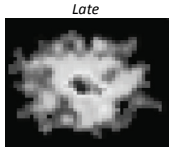
16

Classification example 2

Class:
- Stages of Formation



Attributes:
- Image features
- Characteristics of light waves received
- Etc.



Data size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Courtesy: http://aps.umn.edu

17

Deviation detection examples

- Detecting significant deviations from normal behavior
- Cleanse faulty sensor readings
- Credit card fraud
- Network intrusion detection

18

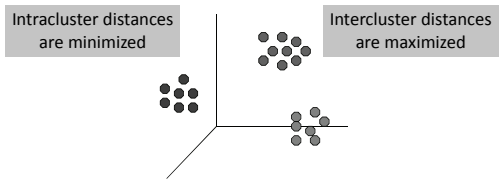
Regression examples

- Predicting a continuously valued variable
- Advertising expenditure on a new product → sales
- Temperature, humidity, air pressure → wind velocity
- Past performance of stock index → future performance

19

Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Points in one cluster are more similar to one another
  - Points in separate clusters are less similar to one another
- E.g.: clustering in 3-d with Euclidean distance



20

Clustering examples

- Market segmentation: divide a market into subsets of customers, where each subset can be targeted with a distinct marketing strategy
  - Collect attributes of customers related to their lifestyle, geographical location, and purchase patterns
- Document clustering: find groups of documents that are similar to each other, so information retrieval can operate on the group level
  - Define a similarity measure based on the frequencies of terms occurring on both documents

21

Sequential pattern examples

- Given a set of sequences of events, find rules that predict strong sequential dependencies among different events
  - E.g.: sequences A(AB)AC(DEF)G, (ABC)BC(DE)F, (AC)BEF...
  - E.g.: rule (AB)C → (DE), often with timing constraints like (AB) and C are no more than 10 units apart
- Mining alarm logs
  - (inverter_problem, excessive_line_current) (rectifier_alarm) → (fire_alarm)
- Mining point-of-sale transaction logs
  - (Intro_to_Visual_C) (C++_Primer) → (Perl_for_Dummies)
  - (shoes) (racket, racketball) → (sports_jacket)

22