Knowledge Discovery with Data and Text Mining

Dursun Delen, Ph.D.
Associate Professor of MSIS
William S. Spears School of Business
Oklahoma State University - Tulsa

Introduction

- Why data mining?
- What is data mining?
- Data Mining Process
- Data Mining Techniques
- Text Mining
- Data & Text Mining Examples
- A Demonstration of Data Mining in Action

Why Now?

Data Deluge

hospital patient registries
electronic point-of-sale data
stock trades
catalog orders
OLTP
telephone calls
bank transactions
remote sensing images
airline reservations
tax returns
credit card charges

Barcodes
RFIDs
Magnetic Strips
Sensors...

Motivation: "Necessity is the Mother of Invention"

- Competitive pressure to make better decisions
- Data explosion problem (or opportunity)
  - Why do we have more data now, then we had before?
    - Technology driven reasons: Automated data collection tools and techniques...
    - Software driven reasons: Mature database technology...
    - Cost driven reasons (both hardware and software)

We are drowning in data, but starving for knowledge!

Solution: Data and data mining

What Is Data Mining?

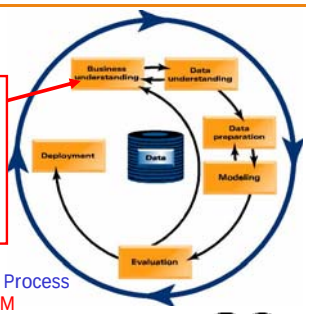
- Data mining :
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information (or patterns) from data
- Alternative names...
  - Data mining: a misnomer?
  - Knowledge discovery, knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Standardized DM Processes

- CRISP-DM
  - Cross-Industry Standard Process for Data Mining
  - www.crisp-dm.org
- SEMMA
  - ✓ Sample the data
  - ✓ Explore the data
  - ✓ Modify the data
  - ✓ Model
  - ✓ Assess
  - http://www.sas.com/technologies/analytics/datamining/miner/semma.html

Standardized Data Mining Processes

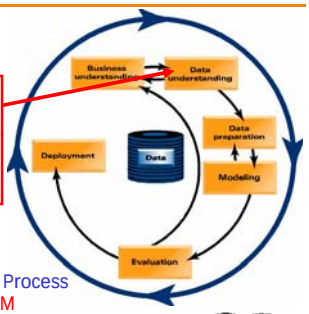
Step 1: Business Understanding
- Determine the business objectives
- Assess the situation
- Determine the data mining goals
- Produce a project plan



Cross-Industry Standard Process for Data Mining CRISP-DM

Standardized Data Mining Processes

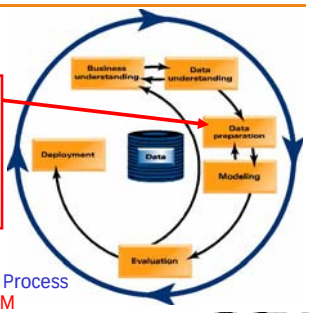
Step 2: Data Understanding
- Collect the initial data
- Describe the data
- Explore the data
- Verify the data



Cross-Industry Standard Process for Data Mining CRISP-DM

Standardized Data Mining Processes

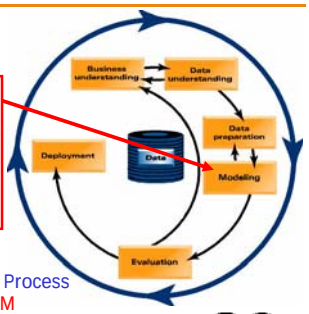
Step 3: Data Preparation
- Select data
- Clean data
- Construct data
- Integrate data
- Format data



Cross-Industry Standard Process for Data Mining CRISP-DM

Standardized Data Mining Processes

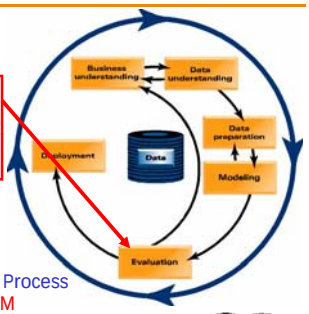
Step 4: Modeling
- Select the modeling technique
- Generate test design
- Build the model
- Assess the model



Cross-Industry Standard Process for Data Mining CRISP-DM

Standardized Data Mining Processes

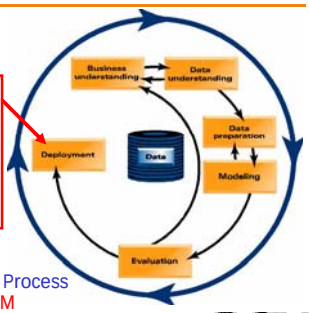
Step 5: Evaluation
- Evaluate results
- Review process
- Determine next step



Cross-Industry Standard Process for Data Mining CRISP-DM

Standardized Data Mining Processes

Step 6: Deployment
- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review the project



Cross-Industry Standard Process for Data Mining CRISP-DM

Data Mining Techniques (1)

- Association (correlation and causality)
  - age(X, "20..29") ^ income(X, "20..29K") → buys(X, "PC") [support = 2%, confidence = 60%]
  - contains(T, "computer") → contains(x, "software") [support = 1%, confidence = 75%]
- Prediction (Classification & Regression)
  - Classification: developing models (functions) that describe and distinguish classes or concepts for future prediction
  - Predicting whether it will rain tomorrow
  - Classifying loan applicant as "good" or "bad"

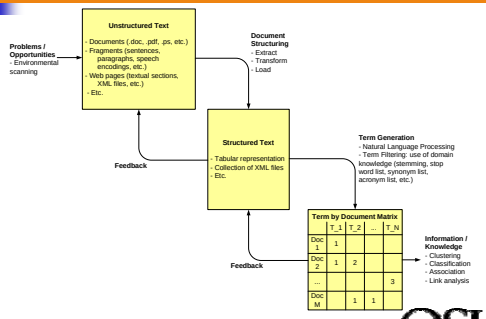
Data Mining Techniques (2)

- Prediction (Classification & Regression)
  - Regression: developing models (functions) that forecast the value of a continuous numerical variable
  - Predicting what the temperature will be tomorrow
  - Forecasting the future value of a stock a year from now
- Cluster analysis
  - Class label is unknown: Group the samples of objects based on their quantifiable characteristics
  - Cluster the customers who share the same characteristics: interest, income level, spending habits, etc.
  - Clustering is done by maximizing the intra-class similarity and minimizing the interclass similarity

Text Mining

- Text mining is a process that employs
  - a set of algorithms for converting unstructured text into structured data objects, and (Statistical Natural Language Processing)
  - the quantitative methods that analyze these data objects to discover knowledge (Data Mining)
- Text Mining = Statistical NLP + DM

Text Mining Process Map



Data Mining Applications... Military Health System (1/3)

- KBSI's Phase II SBIR research project
  - Funded through SBIR program by the Offices of the Secretary of Defense
  - SBIR: Phase I ⇒ Phase II ⇒ Phase III
- DM in Healthcare
  - Managerial
  - Clinical

Reference

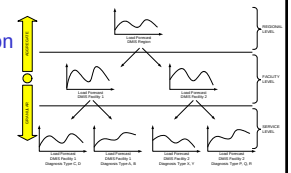
- Delen, D. and S. Ramachandran (2003). A Hybrid Approach to Knowledge Discovery from Military Health Systems. ***Journal of Neural, Parallel and Scientific Computing***, Volume 11, Number 1&2, pp. 161-183.

Data Mining Applications... Military Health System (2/3)

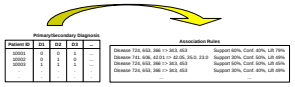
- One of the largest health systems in the US
  - 90+ hospitals
  - 100s of outpatient clinics, treatment facilities
  - 180,000 employees (doctors, nurses, other staff)
  - 8 million beneficiaries
  - $20 Billion/year budget
- Purpose/mission is to
  - Provide healthcare to eligible veterans
  - Provide education and training opportunities to health profession (residency, practical training, etc.)
  - Conduct medical research, create innovation
  - Provide public health service at the time of natural disasters
- Problem: ⇑ demand, ⇓ budget

Data Mining Applications... Military Health System (3/3)

- Classification/prediction
  - Demand forecasting
  - Resource allocation



- Association rules
  - Patient diagnosis



Data Mining Applications... Predicting Breast Cancer Survivability

- Objective: Predicting breast cancer survivability
- Methods/Materials:
  - Used artificial neural networks (MLP), decision trees (C5) and logistic regression
  - Used 10-fold cross validation and more than 200,000 cases
- Results:
  - Decision tree (C5) is the best predictor with 93.6% accuracy, artificial neural networks (MLP) came out to be the second with 91.2% accuracy, and the logistic regression models came out to be the worst of the three with 89.2% accuracy

Reference: Delen, D., G. Walker and A. Kadam (2004). Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods, to appear in ***Artificial Intelligence in Medicine***.

A Brief List of Text Mining Applications

- Exploring trends in research literature
  - We did this for IS literature
- Identifying functional relationships between different genes using large collections of interdisciplinary research literature
- Analyzing records for customer contact centers
  - Both audio as well as text (SAS does this)
- Using text mining to detect deception
  - One of our Ph.D. student working on this
- "Reading", summarizing classifying emails
- Text mining the Web content
- Anti-terrorism initiatives (CIA is using text mining)

Forecasting Box Office Success of Hollywood Movies

Ramesh Sharda and Dursun Delen

Institute for Research in Information Systems
Oklahoma State University

Forecasting Box-Office Receipts: A Tough Problem!

"... No one can tell you how a movie is going to do in the marketplace... not until the film opens in darkened theatre and sparks fly up between the screen and the audience"

Mr. Jack Valenti
President and CEO
of the Motion Picture Association of America

Current work on prediction...

- A lot of research have been done
  - Behavioral models
  - Analytical models
    - Predict after the initial release
- Our approach
  - Use a data mining approach
  - Use as much historical data as possible
  - Make it web-enabled
    - Predict before the initial release

Our Approach – Movie Forecast Guru

- DATA – 849 Movies released between 1998-2006
- Movie Decision Parameters:
  - Intensity of competition rating
  - MPAA Rating
  - Star power
  - Genre
  - Technical Effects
  - Sequel ?
  - Estimated screens at opening
  - ...
- Output: Box office gross receipts (flop → blockbuster)

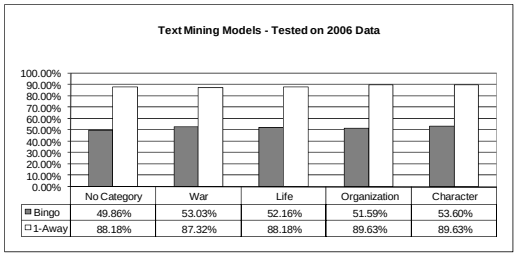
Class No.	1	2	3	4	5	6	7	8	9
Range (in Millions)	< 1 (Flop)	> 1 < 10	> 10 < 20	> 20 < 40	> 40 < 65	> 65 < 100	> 100 < 150	> 150 < 200	> 200 (Blockbuster)

Our Approach – Movie Forecast Guru

PREDICTION MODELS

- Statistical Models:
  - Discriminant Analysis
  - Ordinal Multiple Logistic Regression
- Machine Learning Models:
  - Artificial Neural Networks
  - Decision Tree Induction
    - CART - Classification & Regression Trees
    - C5 - Decision Tree
  - Support Vector Machines
  - Rough Sets
  - ...

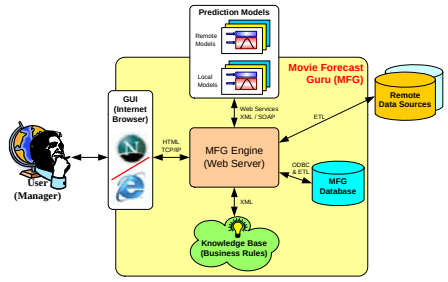
Prediction Results...
Models include plots via text-mining



Sample Predictions...

Movie	Actual	No Category	War	Organization	Life	Character
Spider-Man 3	9	9	9	9	9	9
300	9	5	7	5	5	5
The Simpsons Movie	8	9	9	9	9	9
The Bourne Ultimatum	8	6	7	9	9	7
Knocked Up	7	5	5	5	5	7
Live Free or Die Hard	7	9	7	9	9	9
Evan Almighty	6	6	7	7	7	7
1408	6	5	5	4	5	5
TMNT	5	4	3	4	5	5
Music and Lyrics	5	4	5	5	5	5
Freedom Writers	4	4	3	4	4	5
Reign Over Me	3	4	4	4	4	4
Black Snake Moan	2	3	3	3	3	3
Ta Ra Rum Pum	1	1	1	1	1	1

Web-based DSS



Demonstration of MGF