Tecniche di Apprendimento Automatico per Applicazioni di Data Mining

Machine Learning, Data Mining, and Knowledge Discovery: An Introduction

• Prof. Pier Luca Lanzi
• Laurea in Ingegneria Informatica
• Politecnico di Milano
• Polo di Milano Leonardo
Lecture outline

- Why Data Mining?
- What are the typical applications?
- What is Knowledge Discovery?
  What is Data Mining?
- What are the major issues?
Why Data Mining?

• “Necessity is the Mother of Invention”
• There is a data flood issue
  – Large amounts of data easily generated, collected, and stored
  – Commercial data: banking, telecom, e-commerce applications, etc.
  – Scientific data: astronomy, biology, etc.
• Drowning in data, but starving for knowledge!
• Solution: data warehousing and knowledge extraction
Trends leading to data flood

• More data is generated:
  – Bank, telecom, other business transactions ...
  – Scientific data: astronomy, biology, etc
  – Web, text, and e-commerce
Big data examples

• Europe's Very Long Baseline Interferometry
  – 16 telescopes, each of which produces 1 Gbit/s of astronomical data over a 25-day observation session
  – storage and analysis a big problem

• AT&T
  – Billions of calls per day
  – So much data, it cannot be all stored
  – Analysis has to be done “on the fly”, on streaming data
Largest databases in 2003

• Commercial databases
  – France Telecom has largest decision-support DB, ~30TB; AT&T ~ 26 TB (Winter Corp. 2003 Survey)

• Web
  – Alexa internet archive: 7 years of data, 500 TB
  – Google searches 4+ Billion pages, many hundreds TB
  – Internet Archive (www.archive.org), ~ 300 TB
Data growth rate

• UC Berkeley 2003 estimates
  5 exabytes (5 million terabytes) of
  new data was created in 2002.
• Twice as much information was created
  in 2002 as in 1999 (~30% growth rate)
• US produces ~40% of new stored data worldwide
• Other growth rate estimates even higher
• Very little data will ever be looked at by a human
• Knowledge Discovery is **NEEDED** to make sense
  and use of data.
What are the typical applications?

• Market analysis and management
  – target marketing, customer relation management, market basket analysis, cross selling, market segmentation

• Risk analysis and management
  – Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  – Fraud detection and management

• Bioinformatics
  – Analysis of genomic data

• Other Applications
  – Text mining and Web analysis.
  – Intelligent query answering
Customer Attrition

• Situation
  – Attrition rate at for mobile phone customers is around 25-30% a year!

• Task
  – Given customer information for the past N months, predict who is likely to attrite next month.
  – Also, estimate customer value and what is the cost-effective offer to be made to this customer.
Customer Attrition: Results

• Verizon Wireless built a customer data warehouse (Reported in 2003)
• Identified potential attriters
• Developed multiple, regional models
• Targeted customers with high propensity to accept the offer
• Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers)
Assessing Credit Risk

• Situation
  – Person applies for a loan

• Task
  – Should a bank approve the loan?

• Note
  – People who have the best credit don’t need the loans, and people with worst credit are not likely to repay. Bank’s best customers are in the middle
Assessing Credit Risk: Results

• Banks develop credit models using variety of machine learning methods.

• Mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan.

• Widely deployed in many countries.
E-commerce

• Situation
  – A person buys products at Amazon.com.

• Task
  – Recommend other products this person is likely to buy

• Amazon does clustering based on product bought and searched

• Recommendation program is quite successful
Security and Fraud Detection

- Credit Card Fraud Detection
- Detection of Money laundering
  - FAIS (US Treasury)
- Securities Fraud
  - NASDAQ KDD system
- Phone fraud
  - AT&T, Bell Atlantic, British Telecom/MCI
- Bio-terrorism detection at Salt Lake Olympics 2002
Market Analysis and Management (1)

• Where are the data sources for analysis?
  – Credit card transactions, loyalty cards, discount coupons, 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 the association information
Market Analysis and Management (2)

• Customer profiling
  – data mining can tell you what types of customers buy what products (clustering or classification)

• Identifying customer requirements
  – identifying the best products for different customers
  – use prediction to find what factors will attract new customers

• Provides summary information
  – various multidimensional summary reports
  – statistical summary information (data central tendency and variation)
Corporate Analysis and 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 and Management (I)

• Applications
  – health care, retail, credit card services, telecommunications, etc.

• Approach
  – historical data to build models of fraudulent behavior and use data mining to help identify similar instances

• Examples
  – auto insurance: detect a group of people who stage accidents to collect on insurance
  – money laundering: detect suspicious money transactions
  – medical insurance: detect professional patients and ring of doctors and ring of references
Fraud Detection and Management (2)

• Detecting inappropriate medical treatment
  – Australian Health Insurance Commission identifies that in many cases blanket screening tests were requested (save Australian $1m/yr).

• Detecting telephone fraud
  – Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.
  – British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.

• Retail
  – Analysts estimate that 38% of retail shrink is due to dishonest employees.
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.
Unsuccessful e-commerce

• Situation
  – Clickstream and purchase data from Gazelle.com, legwear and legcare e-tailer
  – Dataset of 3,465 purchases, 1,831 customers

• Task
  – Characterize visitors who spend more than $12 on an average order at the site

• Very interesting analysis by KDDCup2000 participants
• Thousands of hours - $X,000,000 (Millions) of consulting
• Total sales -- $Y,000
• Obituary: Gazelle.com out of business, Aug 2000
What are suitable problems?

- require knowledge-based decisions
- have a changing environment
- have sub-optimal current methods
- have accessible, sufficient, and relevant data
- provides high payoff for the right decisions!

- Privacy considerations important if personal data is involved
What is Knowledge Discovery?

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

• Note, it is a process!
Example: Credit Risk

- Loan not repaid
- Loan repaid

- Salary

Diagram shows a scatter plot with points indicating loan status based on salary.
Example: Credit Risk

IF salary < k THEN not repaid
Example: Credit Risk

- Valid
  - The pattern has to be valid with respect to a certainty level (rule true for the 86.7%)
- Novel
  - The value $k$ should be previously unknown or obvious
- Useful
  - The pattern should provide information useful to the bank for assessing credit risk
- Understandable
What are the related fields?

- Visualization
- Machine Learning
- Statistics
- Databases

Knowledge Discovery and Data Mining
Statistics, Machine Learning, and Data Mining

• Statistics:
  – more theory-based, focused on testing hypotheses

• Machine learning
  – more heuristic, focused on building program that learns, more general than Data Mining

• Knowledge Discovery
  – integrates theory and heuristics
  – focus on the entire process of discovery, including data cleaning, learning, integration and visualization

• Data Mining
  – focus on the algorithms to extract patterns from data

• Distinctions are fuzzy!
Knowledge Discovery Process flow, according to CRISP-DM
Knowledge Discovery Process

- raw data
- selection
- cleaning
- transformation
- feedback
- evaluation
- mining
- patterns
- knowledge
- patterns
Steps of the Knowledge Discovery Process

• Learning the application domain to extract relevant prior knowledge and goals of application
• Creating a target data set: data selection
• Data cleaning and preprocessing:
• Data reduction and transformation:
• Choosing data mining approach
  – 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
Increasing potential to support business decisions

Knowledge Discovery and Business Intelligence

End User
Business Analyst
Data Analyst
DBA

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

- Graphical user interface
- Pattern evaluation
- Data mining engine
- Database or data warehouse server
- Data cleaning & data integration
- Databases
- Data Warehouse
- Filtering
- Knowledge base
Historical Note:
Many Names of Data Mining

• Data Fishing, Data Dredging: 1960-
  – used by Statistician (as bad name)
• Data Mining :1990-
  – used DB, business
• Knowledge Discovery in Databases (1989-)
  – used by AI, Machine Learning Community
• also Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, ...
• Currently: Data Mining and Knowledge Discovery are used interchangeably
Major Data Mining Tasks

- Classification: predicting an item class
- Clustering: finding clusters in data
- Associations: frequent occurring events...
- Visualization: to facilitate human discovery
- Summarization: describing a group
- Deviation Detection: finding changes
- Estimation: predicting a continuous value
- Link Analysis: finding relationship
Data Mining Tasks: classification

• Classification and Prediction
  – Finding models (functions) that describe and distinguish classes or concepts for future prediction
  – E.g., classify countries based on climate, or classify cars based on gas mileage
  – Presentation: decision-tree, classification rule, neural network
  – Prediction: Predict some unknown or missing numerical values
Data Mining Tasks: classification
Data Mining Tasks: clustering

• Cluster analysis
  – Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  – Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity
Data Mining Tasks: clustering
Data Mining Tasks: others

- Associations
- Outlier analysis
  - Outlier: a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses
Are All the “Discovered” Patterns Interesting?

• A data mining system/query may generate thousands of patterns, not all of them are interesting.
  – Suggested approach: Human-centered, query-based, focused mining

• Interestingness measures: a pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm

• Objective vs. subjective interestingness measures:
  – Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  – Subjective: based on user’s belief in the data, e.g., unexpectedness, novelty, actionability, etc.
Can we find all and only interesting patterns?

• Find all the interesting patterns: Completeness
  – Can a data mining system find all the interesting patterns?
  – Association vs. classification vs. clustering

• Search for only interesting patterns: Optimization
  – Can a data mining system find only the interesting patterns?
  – Approaches
    • First general all the patterns and then filter out the uninteresting ones.
    • Generate only the interesting patterns—mining query optimization
Data Mining tasks

• General functionality
  – Descriptive data mining
  – Predictive data mining

• Different views, different classifications
  – Kinds of databases to be mined
  – Kinds of knowledge to be discovered
  – Kinds of techniques utilized
  – Kinds of applications adapted
Major Issues in Data Mining (I)

- Mining methodology and user interaction
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem

- Performance and scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods
Major Issues in Data Mining (2)

• Issues relating to the diversity of data types
  – Handling relational and complex types of data
  – Mining information from heterogeneous databases and global information systems (WWW)

• Issues related to applications and social impacts
  – Application of discovered knowledge
    • Domain-specific data mining tools
    • Intelligent query answering
    • Process control and decision making
  – Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
  – Protection of data security, integrity, and privacy
Summary

• Technology trends lead to data flood
• Data mining is needed to make sense of data
• Data Mining has many applications, successful and not
• Knowledge Discovery Process
• Different Data Mining Tasks
Where to Find References?

- Data mining and KDD (SIGKDD member CDROM):
  - Conference proceedings: KDD, and others, such as PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery

- Database field (SIGMOD member CD ROM):
  - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, DASFAA
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.

- AI and Machine Learning:
  - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.

- Statistics:
  - Conference proceedings: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.

- Visualization:
  - Conference proceedings: CHI, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.
References