CS219: Data Mining

Course Web Page:

www.abulaish.com/dm2019w.html

General Information

- Course Instructor: Muhammad Abulaish
  - Email: abulaish@sau.ac.in
  - Tel: 24195 (148)
  - Office: Room No. 310
- Teaching assistant (TA)
  - Harshita Dalal <sahisnumazumder@gmail.com>
- Lecture:
  - 3:30pm-5:30pm (Tuesday)
  - 3:30pm-4:30pm (Thursday)
- Lab:
  - 4:30pm-6:30pm (Thursday)
- My office hours: 3:30pm-4:30pm, Monday & Wednesday (or by appointment)

2

Course Structure & Evaluation

- The course has three parts
  - **Lectures** (to discuss basic concepts and algorithms)
  - **Lab** (implementation of two data mining algorithms)
  - **Presentations** (self study on some recent advances in data mining and its applications)
- Lecture slides will be available at course web page.
- Evaluation:
  - Quiz: 10% (At least two quizzes)
  - Midterm: 20% (one mid-term)
  - Presentations: 10% (group of two students)
  - Lab assignment: 20%
  - Final Exam: 40%

3

Prerequisite

- Knowledge of
  - Discrete Mathematics & Probability Theory
  - Data Structures and Algorithms

4

Teaching Materials

- Text Book
  - Data mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber, Morgan Kaufmann, ISBN 1-55860-489-8.
- Reference Books
  - Data Mining and Analysis: Fundamental Concepts and Algorithms, M. J. Zaki & Wagner M. Jr., Cambridge Press.
  - Introduction to Data Mining, by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Pearson/Addison Wesley, ISBN 0-321-32136-7.
  - Machine Learning, by Tom M. Mitchell, McGraw-Hill, ISBN 0-07-042807-7

5

Topics

- Introduction
- Data Pre-processing
- Association Rule Mining
- Classification Techniques (Supervised Learning)
- Clustering Techniques (Unsupervised Learning)
- Semi-Supervised Learning
- Applications
  - Social network analysis
  - Opinion mining and sentiment analysis
  - Recommender systems and collaborative filtering
  - Web data extraction

6

Data

- Data is the Latin plural of datum
- Used to represent unprocessed facts and figures without any added interpretation or analysis.
- Generally associated with some entity and often viewed as the lowest level of abstraction from which information and knowledge are derived.
- Data may be unstructured, semi-structured, and structured
- Example: The price of petrol is Rs. 70 per liter

Information

- Information is interpreted (processed) data so that it has meaning for the user.
- "The price of petrol has risen from Rs. 64 to Rs. 70 per liter" – is information for a person who tracks petrol prices.
- Data becomes information when it is processed for some purpose and adds value for the recipient.
- A set of raw sales figures – Data
- Sales report (chart plotting, trend analysis) – Information

Knowledge

- Knowledge is a fluid mix of information, experience and insight that may benefit the individual or the organization.
- "When petrol prices go up by Rs. 6 per liter, it is likely that bus fare will rise by 12%" is knowledge.
- The boundaries between data, information, and knowledge is fuzzy
- What is data to one person is information to someone else.

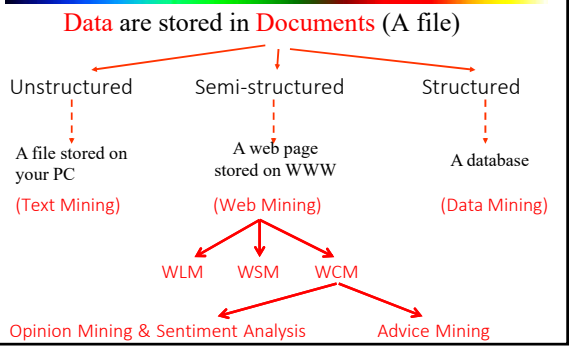
Summarized View

- Data – as in databases, spreadsheets, text files…
- Information – Processed data
- knowledge – **Fluid mix of information, experience, and insight**

OR, knowledge is a meta information about the patterns hidden in the data

The patterns must be discovered automatically!!!

Data Categories & Mining Terminologies



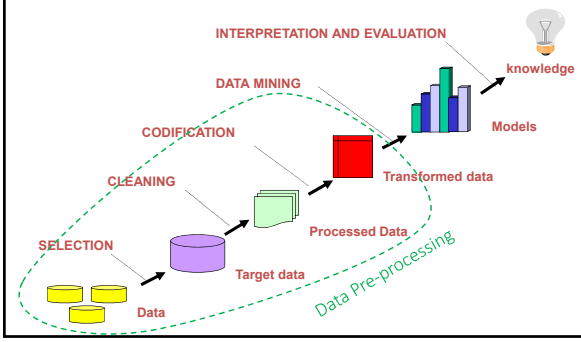
Short History of Data Mining

- 1989 - KDD (Knowledge Discovery in Databases) appeared in (IJCAI Workshop)
- 1991 - A collection of research papers edited by Piatetsky-Shapiro and Frawley
- 1993 – Association Rule Mining Algorithm (Apriori algorithm) was proposed by Agrawal, Imielinski and Swami.
- 1996 – present: KDD evolves as a conjunction of different knowledge areas (data bases, machine learning, statistics, artificial intelligence) and the term Data Mining becomes popular

What is Data Mining?

- An important step of KDD (Knowledge Discovery from Databases) process
- Data mining is the automatic extraction of interesting knowledge (rules, regularities, patterns, constraints) from large data sources, e.g., databases, texts, web, images, etc.
- Identified patterns must be:
  - Valid, novel (non-trivial), potentially useful, and understandable

The KDD Process



Data Mining Objectives

- Identification of data as a source of useful information
- Automatic extraction of valid and novel patters from the data source
- Use of discovered patterns for competitive advantages when working in business environment

Why Data Mining?

- Data Explosion (Information Overload) problem
  - We are drowning in data, but starving for knowledge!
  - Data data everywhere nor any drop of insight!

    (water water everywhere nor any drop to drink)
- Explosive growth of data: from Terabytes to Petabytes
- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, datawarehouses, and other data repositories

Why Data Mining? Cont…

- Major sources of abundant data
  - Business: Web, e-commerce, transactions, stocks, …
  - Science: Remote sensing, bioinformatics, scientific simulation
  - Society and everyone: news, digital cameras,
- The computing power is not an issue.
- Data mining tools are available
- The competitive pressure is very strong.
  - Almost every company is doing (or has to do) it

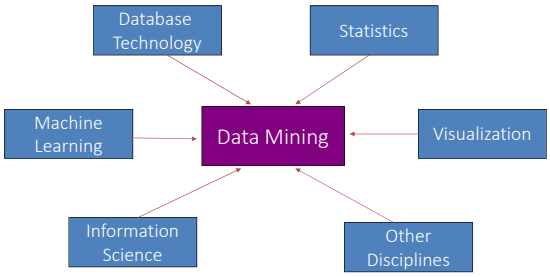
Why Data Mining Important?

- Digitization of businesses produce huge amount of data
  - How to make best use of data?
  - Knowledge discovered from data can be used for competitive advantage.
- E-businesses are generating huge amount of data sets
  - Online retailers (e.g., amazon.com) are largely driving by data mining.
  - Web search engines are information retrieval (text mining) and data mining companies

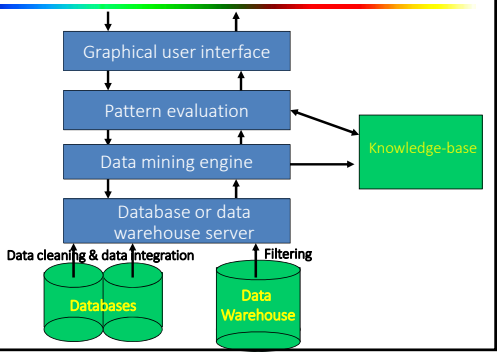
Why is Data Mining Necessary?

- Make use of your data assets (knowledge-based economy)
- Big gap from stored data to knowledge
  - Transition won't occur automatically.
- Many interesting things can't be found using database queries
  - Customers likely to buy my products?
  - Why sale was down after demonitization?
  - Which items should be recommended to a person purchasing computer?

Data Mining: Confluence of Multiple Disciplines



Architecture of a Typical Data Mining System



Data Mining: On what kind of data?

- Relational Databases
- Data Warehouses
- Transactional Databases
- Advanced DB and Data Repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - Web Database

Data Mining Functionalities: Characterization (1)

- A data mining process aims to find rules that describe the properties of a concept.
- Standard form:

If concept then characteristics

- C=1 → A=1 & B=3 (Support: 25%, i.e., there are 25% records for which the rule is true)
- C=1 → A=1 & B=4 (Support: 17%)
- C=1 → A=0 & B=2 (Support: 16%)

Data Mining Functionalities: Discrimination (2)

- A data mining process which aims is to find rules that allow us to discriminate the objects (records) belonging to a given concept (one class) from the rest of records (classes)
- Standard form:

  If characteristics then concept
- A=0 & B=1 → C=1 (Support: 33%, Confidence: 83%)
  - Confidence: The conditional probability of the concept given the characteristics
- A=2 & B=0 → C=1 (27%, 80%)
- A=1 & B=1 → C=1 (12%, 76%)

Data Mining Functionalities: Classification and Prediction (3)

- Finding models (rules) that describe (characterize) and/or distinguish (discriminate) classes or concepts for future prediction.
  - Classify countries based on climate (characteristics)
  - Classify cars based on gas mileage and use it to predict classification of a new car
- Presentation:
  - Decision Tree
  - Classification Rules
  - Neural Network
  - Bayes Network

Data Mining Functionalities: Prediction (statistical) (4)

- A Data Mining process to predict some unknown or missing numerical values.
- Output space: continuous

Data Mining Functionalities: Association Analysis (5)

- A Data Mining process which aims to identify patterns (aka frequent itemsets) in data
- For example:
  - Buy(X, Printer) → Buy (X, Cartridge)
  - Buy (X, Bread) → Buy (X, Butter) ∧ Buy (X, Milk)

Data Mining Functionalities: Cluster Analysis (6)

- Unsupervised learning
- Aims to group data to form new classes
  - Cluster houses to find distribution patterns
- **Basic principle:** Maximizing the intra-class similarity and minimizing the inter-class similarity

Data Mining Functionalities: Outlier Analysis (7)

- **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, etc.

Major issues in Data Mining

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages
- Expression and visualization of data mining results
- Handling noise and incomplete data
- Pattern evaluation: the interestingness problem
- Efficiency and scalability of data mining algorithms
- Parallel, distributed, and incremental mining methods

Major issues in Data Mining (cont…)

- Handling relational and complex types of data
- Mining information from heterogeneous databases and global information systems (WWW)
- 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

Limitations of Computing Machines and Data Deluge

The Huber Taxonomy of Data Set Sizes

Descriptor	Data Set Size In Bytes	Storage Mode
Tiny	10^2	Piece of Paper
Small	10^4	A Few Pieces of Paper
Medium	10^6	A Floppy Disk
Large	10^8	Multipl Floppy Disks
Huge	10^{10}	Hard Disk
Massive	10^{12}	Multiple Hard Disks, e.g. RAID Storage

Algorithmic Complexity

Algorithm	Complexity
Plot a scatterplot	$O(n^{1/2})$
Calculate means, variances, kernel density estimates	O(n)
Calculate fast Fourier transforms	O(n log(n))
Calculate singular value decomposition of an rc matrix; solve a multiple linear regression	O(nc)
Solve most clustering algorithms	$O(n^2)$

No. of Operations for Algorithms of Various Computational Complexities and various Data Set Sizes

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	10	10^2	$2x10^2$	10^3	10^4
small	10^2	10^4	$4x10^4$	10^6	10^8
medium	10^3	10^6	$6x10^6$	10^9	10^{12}
large	10^4	10^8	$8x10^8$	10^{12}	10^{16}
huge	10^5	10^{10}	10^{11}	10^{15}	10^{20}

Computational Feasibility on a Pentium PC (10 MegaFLOPs)

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	10^{-6} seconds	10^{-5} seconds	$2x10^{-5}$ seconds	.0001 seconds	.001 seconds
small	10^{-5} seconds	.001 seconds	.004 seconds	.1 seconds	10 seconds
medium	.0001 seconds	.1 seconds	.6 seconds	1.67 minutes	1.16 days
large	.001 seconds	10 seconds	1.3 minutes	1.16 days	31.7 years
huge	.01 seconds	16.7 minutes	2.78 hours	3.17 years	317,000 years

Computational Feasibility on a Silican Graphics Onyx Workstation (300 MegaFLOPs)

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	$3.3x10^{-8}$ seconds	$3.3x10^{-7}$ seconds	$6.7x10^{-7}$ seconds	$3.3x10^{-6}$ seconds	$3.3x10^{-5}$ seconds
small	$3.3x10^{-7}$ seconds	$3.3x10^{-5}$ seconds	$1.3x10^{-4}$ seconds	$3.3x10^{-3}$ seconds	.33 seconds
medium	$3.3x10^{-6}$ seconds	$3.3x10^{-3}$ seconds	.02 seconds	3.3 seconds	55 minutes
large	$3.3x10^{-5}$ seconds	.33 seconds	2.7 seconds	55 minutes	1.04 years
huge	$3.3x10^{-4}$ seconds	33 seconds	5.5 minutes	38.2 days	10,464 years

Computational Feasibility on an Intel Paragon XP/S A4 (4.2 GigaFLOPs)

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	$2.4x10^{-9}$ seconds	$2.4x10^{-8}$ seconds	$4.8x10^{-8}$ seconds	$2.4x10^{-7}$ seconds	$2.4x10^{-6}$ seconds
small	$2.4x10^{-8}$ seconds	$2.4x10^{-6}$ seconds	$9.5x10^{-6}$ seconds	$2.4x10^{-4}$ seconds	.024 seconds
medium	$2.4x10^{-7}$ seconds	$2.4x10^{-4}$ seconds	.0014 seconds	.24 seconds	4.0 minutes
large	$2.4x10^{-6}$ seconds	.024 seconds	.19 seconds	4.0 minutes	27.8 days
huge	$2.4x10^{-5}$ seconds	2.4 seconds	24 seconds	66.7 hours	761 years

Computational Feasibility on a TeraFLOP Grand Challenge Computer (1000 GigaFLOPs)

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	10^{-11} seconds	10^{-10} seconds	$2x10^{-10}$ seconds	10^{-9} seconds	10^{-8} seconds
small	10^{-10} seconds	10^{-8} seconds	$4x10^{-8}$ seconds	10^{-6} seconds	10^{-4} seconds
medium	10^{-9} seconds	10^{-6} seconds	$6x10^{-6}$ seconds	.001 seconds	1 second
large	10^{-8} seconds	10^{-4} seconds	$8x10^{-4}$ seconds	1 second	2.8 hours
huge	10^{-7} seconds	.01 seconds	.1 seconds	16.7 minutes	3.2 years

Types of Computers for Interactive Feasibility (Response Time < 1 Second)

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Personal Computer
small	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Super Computer
medium	Personal Computer	Personal Computer	Personal Computer	Super Computer	Teraflop Computer
large	Personal Computer	Workstation	Super Computer	Teraflop Computer	---
huge	Personal Computer	Super Computer	Teraflop Computer	---	---

Types of Computers for Feasibility (Response Time < 1 Week)

n	$n^{1/2}$	n	n log(n)	$n^{3/2}$	n^2
tiny	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Personal Computer
small	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Personal Computer
medium	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Personal Computer
large	Personal Computer	Personal Computer	Personal Computer	Personal Computer	Teraflop Computer
huge	Personal Computer	Personal Computer	Personal Computer	Super Computer	---

Data Mining Applications (1)

- Target marketing, customer relation management, market basket analysis, cross selling,
- Forecasting, customer retention, quality control, competitive analysis
- Text mining (news group, email, documents) and Web analysis.
- Intelligent query answering
- Buying patterns
- Decision support
- Fraud detection

Data Mining Applications (2)

- Scientific Applications
  - Networks failure detection
  - Controllers design
  - Geographic Information Systems
  - Genome - Bioinformatics
  - Intelligent robots

Fraud detection and Management (1)

- Applications
  - Widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances

Fraud detection and Management (2)

- Examples
  - Auto Insurance: detect characteristics of group of people who stage accidents to collect on insurance
  - Money Laundering: detect characteristics of suspicious money transactions
  - Medical Insurance: detect characteristics of fraudulent patients and doctors

Market Analysis and Management (1)

- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: when marriage occurs, etc.
- 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 (person-centric recommendations)

Resources

- ACM SIGKDD (ACM Special Interest Group on Knowledge Discovery and Data Mining)
- Data mining related conferences
  - Data mining: KDD, ICDM, SDM, …
  - AI: ICML, NIPS, AAAI, IJCAI, ACL, …
  - Databases: SIGMOD, VLDB, ICDE, …
  - Web: WWW, WSDM, …
  - Information retrieval: SIGIR, CIKM, …
- Kdnuggets: http://www.kdnuggets.com/
  - News and resources. You can sign-up!

48