國立高雄科技大學電腦與通訊工程系
資料探勘與最佳化實驗室

巨量資料分析應用與實作

(Big Data Analytics in Practice)

Overview of Data Mining

授課教師：江傳文

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in hypothesis formation

Why Mine Data? Commercial Viewpoint

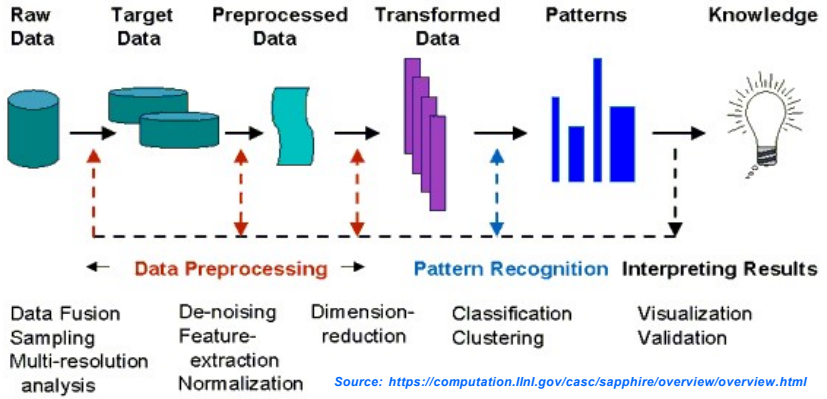
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/grocery stores
  - bank/credit card transactions
- Computers have become cheaper and more powerful
- Competitive pressure is strong
  - provide better, customized services for an edge (e.g. in Customer Relationship Management)

Mining Large Data Sets: Motivation

- There is often information "hidden" in the data that is not readily evident.
- Human analysts may take weeks to discover useful information.
- Much of the data is never analyzed at all.

What is Data Mining? (1/8)

- Data mining and the knowledge discovery process:



What is Data Mining? (2/8)

- Steps of a KDD process: (1/4)
  - Developing and understanding the application domain. This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge.
  - Creating a target data set. Here the data miner selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset.

What is Data Mining? (3/8)

- Steps of a KDD process: (2/4)
  - Data cleaning and preprocessing. This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes.
  - Data reduction and projection. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.

What is Data Mining? (4/8)

- Steps of a KDD process: (3/4)
  - Choosing the data mining task. Here the data miner matches the goals defined in Step 1 with a particular DM method, such as classification, regression, clustering, etc.
  - Choosing the data mining algorithm. The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate.
  - Data mining. This step generates patterns in a particular representational form, such as classification rules, decision trees, regression models, trends, etc.

What is Data Mining? (5/8)

- Steps of a KDD process: (4/4)
  - Interpreting mined patterns. Here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models.
  - Consolidating discovered knowledge. The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge..

What is Data Mining? (6/8)

- More on the KDD process: (1/2)
  - A data mining project should always start with an analysis of the data with traditional query tools
    - » 80% of the interesting information can be extracted using SQL
      - how many transactions per month include item number 15?
      - show me all the items purchased by Sandy Smith.

What is Data Mining? (7/8)

- More on the KDD process: (2/2)
  - A data mining project should always start with an analysis of the data with traditional query tools
    - » 20% of hidden information requires more advanced techniques
      - which items are frequently purchased together by my customers?
      - how should I classify my customers in order to decide whether future loan applicants will be given a loan or not?

What is Data Mining? (8/8)

- Many definitions:
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

What is (not) Data Mining?

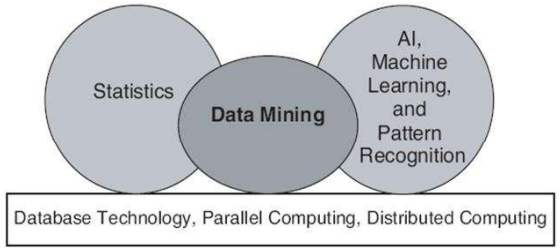
- What is not data mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"
- What is data mining?
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com)

Origins of Data Mining (2/3)

- Statistics, machine learning and data mining (1/2)
  - Statistics
    - » more theory-based
    - » more focused on testing hypotheses
  - Machine learning
    - » more heuristic
    - » focused on improving performance of a learning agent
    - » also looks at real-time learning and robotics—areas not part of data mining

Origins of Data Mining (1/3)

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems



- Traditional techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data

Origins of Data Mining (3/3)

- Statistics, machine learning and data mining (2/2)
  - Data mining and knowledge discovery
    - » integrates theory and heuristics
    - » focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
  - Distinctions are fuzzy

Data Mining Tasks (1/2)

- Prediction methods
  - Use some variables to predict unknown or future values of other variables.
- Description methods
  - Find human-interpretable patterns that describe the data.

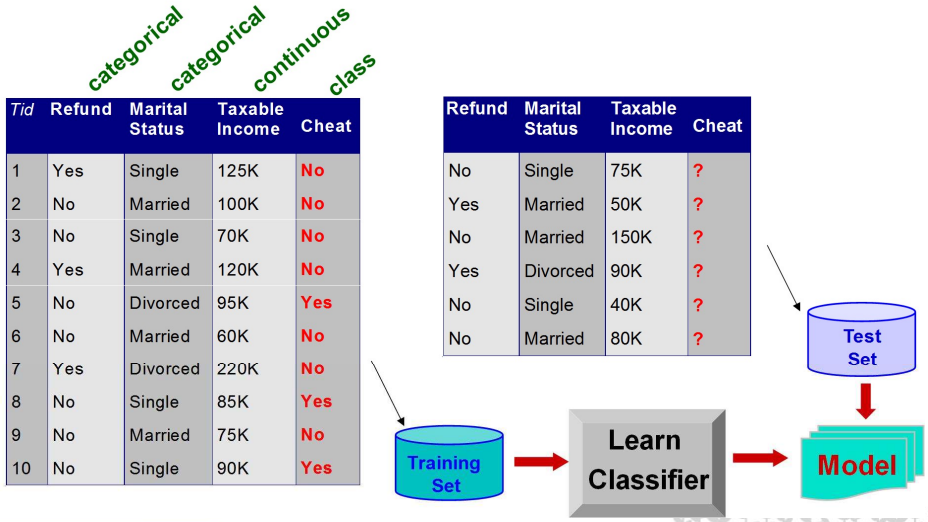
Classification: Definition

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Data Mining Tasks (2/2)

- Classification [predictive]
- Clustering [descriptive]
- Association rule discovery [descriptive]
- Sequential pattern discovery [descriptive]
- Regression [predictive]
- Deviation detection [predictive]

Classification: An Example



Classification: Application #1

- Direct marketing
  - Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - » Use the data for a similar product introduced before.
    - » We know which customers decided to buy and which decided otherwise. This {buy, don't buy} decision forms the class attribute.
    - » Collect various demographic, lifestyle, and company-interaction related information about all such customers.
    - » Use this information as input attributes to learn a classifier model.

Classification: Application #2

- Fraud detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - » Use credit card transactions and the information on its account-holder as attributes. (When does a customer buy, what does he buy, how often he pays on time, etc.)
    - » Label past transactions as fraud or fair transactions. This forms the class attribute.
    - » Learn a model for the class of the transactions.
    - » Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application #3

- Customer attrition/churn
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - » Use detailed record of transactions with each of the past and present customers, to find attributes.
      - * How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - » Label the customers as loyal or disloyal.
    - » Find a model for loyalty.

Clustering: Definition

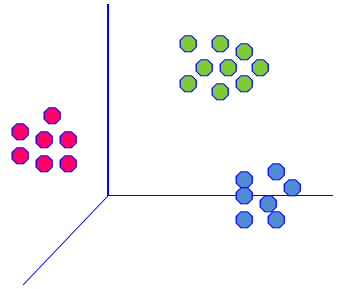
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity measures:
  - Euclidean distance if attributes are continuous.
  - Other problem-specific measures.

Illustrating Clustering

☒Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application #2

- Document clustering
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Clustering: Application #1

- Market segmentation
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - » Collect different attributes of customers based on their geographical and lifestyle related information.
    - » Find clusters of similar customers.
    - » Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection.
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application #1

- Marketing and sales promotion
  - Let the rule discovered be
    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Example Analytics Applications

Marketing	Risk Management	Government	Web	Logistics	Other
Response modeling	Credit risk modeling	Tax avoidance	Web analytics	Demand forecasting	Text analytics
Net lift modeling	Market risk modeling	Social security fraud	Social media analytics	Supply chain analytics	Business process analytics
Retention modeling	Operational risk modeling	Money laundering	Multivariate testing		
Market basket analysis	Fraud detection	Terrorism detection			
Recommender systems					
Customer segmentation					

Association Rule Discovery: Application #2

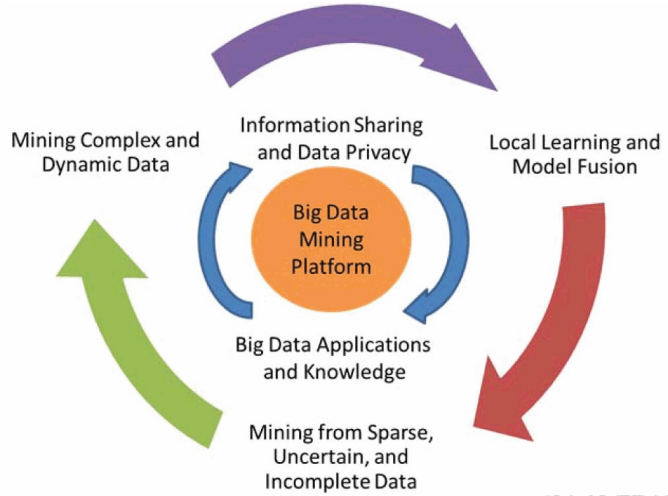
- Supermarket shelf management
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - » If a customer buys diaper and milk, then he is very likely to buy beer.
    - » So, don't be surprised if you find six-packs stacked next to diapers!

Data Mining Challenges with Big Data (1/7)

- For an intelligent learning database system to handle Big Data, the essential key is to
  - scale up to the exceptionally large volume of data and
  - provide treatments for the characteristics of Big Data.
- In general, a conceptual view of the Big Data processing framework includes three tiers from inside out with considerations on
  - data accessing and computing (Tier I),
  - data privacy and domain knowledge (Tier II), and
  - Big Data mining algorithms (Tier III).

Data Mining Challenges with Big Data (2/7)

- A Big Data processing framework



Data Mining Challenges with Big Data (3/7)

- The challenges at Tier I focus on data accessing and arithmetic computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing.
- For example, typical data mining algorithms require all data to be loaded into the main memory, this, however, is becoming a clear technical barrier for Big Data because moving data across different locations is expensive, even if we do have a super large main memory to hold all data for computing.

Data Mining Challenges with Big Data (4/7)

- The challenges at Tier II center around semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III).
- For example, depending on different domain applications, the data privacy and information sharing mechanisms between data producers and data consumers can be significantly different. Sharing sensor network data for applications like water quality monitoring may not be discouraged, whereas releasing and sharing mobile users' location information is clearly not acceptable for majority, if not all, applications.

Data Mining Challenges with Big Data (5/7)

- In addition to the above privacy issues, the application domains can also provide additional information to benefit or guide Big Data mining algorithm designs. For example:
  - In market basket transactions data, each transaction is considered independent and the discovered knowledge is typically represented by finding highly correlated items, possibly with respect to different temporal and/or spatial restrictions.
  - In a social network, on the other hand, users are linked and share dependency structures. The knowledge is then represented by user communities, leaders in each group, and social influence modeling, and so on.

Data Mining Challenges with Big Data (6/7)

- At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics.
- In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing.

References

1. X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data Mining with Big Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97-107, January 2014.

Data Mining Challenges with Big Data (7/7)

- The circle at Tier III contains three stages.
  - First, sparse, heterogeneous, uncertain, incomplete, and multisource data are preprocessed by data fusion techniques.
  - Second, complex and dynamic data are mined after preprocessing.
  - Third, the global knowledge obtained by local learning and model fusion is tested and relevant information is feedback to the preprocessing stage. Then, the model and parameters are adjusted according to the feedback.

國立高雄科技大學電腦與通訊工程系
資料探勘與最佳化實驗室

Q & A