Introduction

August 2, 2016

Recap

- Looked at differences/similarities between Data mining, machine learning, Artificial Intelligence, Info Retrieval, NLP,...
- Simplest definition given in text:

Data Mining

[1]Data mining is the process of automatically discovering useful information in large data repositories.

[1] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley

Applications

Retail Industry

- Large amount of data collected from sales, customer purchasing history, goods transportation, consumption and services.
- Data in different formats, either structured or unstructured.

Applications

Retail Industry

- Large amount of data collected from sales, customer purchasing history, goods transportation, consumption and services.
- Data in different formats, either structured or unstructured.
  - Structured: Relational DBs.
  - Unstructured: Text.
  - Semistructured: Web pages.
- Use the data to identify customer buying habits for targeted marketing, customer retention, product recommendation, analysis of sales, customers, markets.

Applications

Telecommunication

- Various Telecommunication services generate huge amount of data at network level, as well as at application level
- Analyze them to find Telecommunication patterns, catch fraudulent activities.
- Better use of resources and improve quality of services.
- Visualization tools in tele data analysis.

Applications

Biological Data analysis

- Bioinformatics.
- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Gene, DNA and protein sequencing.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Visualization tools in genetic data analysis.

Health care

- We have medical records of hundreds of thousands of patients.
- Involves each aspect of health, like, pressure, sugar level, blood details, at each and every instance of medical help.
- The scenario not very popular in India, otherwise, the scale would have to be multiplied by 1000!!

Health care

- We have medical records of hundreds of thousands of patients.
- Involves each aspect of health, like, pressure, sugar level, blood details, at each and every instance of medical help.
- The scenario not very popular in India, otherwise, the scale would have to be multiplied by 1000!!
- Use these records to predict the next action in a health care system.
  - Tests to be conducted
  - Diagnosis: Predict possible diseases.
  - Reduce fatality by speedy detection and action.

Data Mining and Knowledge Discovery

An integral step in KDD (Knowledge Discovery in Databases), which is the overall process of converting raw data into useful information.

1. Input Data
2. Data Preprocessing
3. DATA MINING
4. Postprocessing
5. Information.

Why so late?

- The concept of data is not new.
- Then why did, data mining evolve so late? Remember
  - Statistics – 1749
  - Artificial Intelligence – 1940
  - Machine leaning – 1946
  - Data mining – 1980

Why so late?

- The concept of data is not new.
- Then why did, data mining evolve so late? Remember
  - Statistics – 1749
  - Artificial Intelligence – 1940
  - Machine leaning – 1946
  - Data mining – 1980
- Because, we now have DATA and not data
- Various challenges posed by new data sets.

Motivating Challenges

- Scalability
  - MB, GB, TB, PB (Peta),....
  - Data mining algorithms have to be scalable.
  - Special search strategies to handle exponential search problems.
  - How to handle data sets that cannot fit into main memory.
  - Parallel and distributed algorithms.
- High Dimensionality
  - Hundreds or Thousands of features.
  - Doc search can involve lakhs..

Motivating Challenges

- Scalability
  - MB, GB, TB, PB (Peta),....
  - Data mining algorithms have to be scalable.
  - Special search strategies to handle exponential search problems.
  - How to handle data sets that cannot fit into main memory.
  - Parallel and distributed algorithms.
- High Dimensionality
  - Hundreds or Thousands of features.
  - Doc search can involve lakhs..
  - Spatial and temporal components.
  - Computational complexity increases rapidly with the dimensionality.

Motivating Challenges

- Heterogeneous and Complex Data
  - Web pages contain semi-structured text and hyperlinks.
  - DNA data: Sequential and 3-D structure.

Motivating Challenges

- Heterogeneous and Complex Data
  - Web pages contain semi-structured text and hyperlinks.
  - DNA data: Sequential and 3-D structure.
  - Should take into consideration relationships in the data.
    - Temporal and spatial autocorrelation.
    - Graph connectivity, parent-child relationships..
- Data Ownership and Distribution.
  - Data not stored in one location or owned by one organization.
  - Distributed data mining techniques
    - Reduce amount of communication.
    - Effective consolidation of data mining results from multiple sources.
    - Addressing data security issues.
  - Smelling something...

Data Mining Tasks

- Predictive tasks:
  - Predict the value of a particular variable.
  - The variable could denote a class (classification) or a function value (regression).
  - To be done using a set of other variables.
  - $x \in \Re^n$, the task is to find $y = h(x)$.
- Descriptive tasks.
  - Derive patterns (correlations, trends, clusters, trajectories, anomalies).
  - Summarizes the underlying relationships in data.
  - Often exploratory in nature and needs to post-process the results

What tasks do we handle

Text says to handle

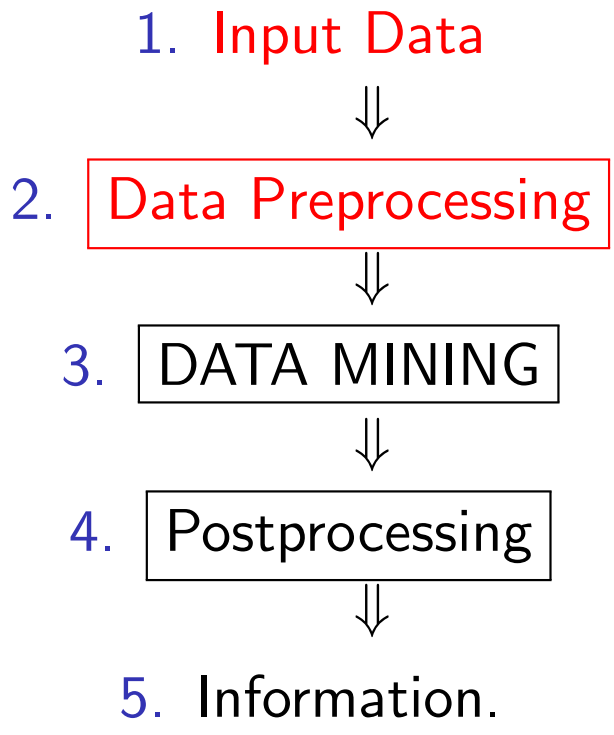
- Predictive Modeling.
- Association Analysis.
- Cluster Analysis.
- Anomaly detection.

2

[2] Exercise: Book1, 1.7: 1,2

DATA

Data Mining and Knowledge Discovery



Data

Question: Take out a few examples from the discussion till now and tell me how you represent the data.

[3] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley

Data

Question: Take out a few examples from the discussion till now and tell me how you represent the data.

- KNOW YOUR DATA.
- Types of Data.
- Quality of Data.
- Preprocessing Steps.
- Analyzing Data in terms of its relationships.

[3] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley

Data

Question: Take out a few examples from the discussion till now and tell me how you represent the data.

- KNOW YOUR DATA.
- Types of Data.
- Quality of Data.
- Preprocessing Steps.
- Analyzing Data in terms of its relationships.
- See Example 2.1[3]

[3] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley