1. Introduction

2. Data

3. Exploring Data

4. Classification

5. Association analysis

6. Cluster analysis

7. Anomaly Detection

8. addition ~~for~~ research topic
   - text ming
   - graph ming
   - web ming

1. Introduction to Data Mining
   Pang Ning Tan / Vipin Kumar
   Michael Steinbach

2. Data Mining Concepts ad Techniques
   Jiawei Han / Micheline Kamber / Jian Pei

Introduction

Rapid advances in data collection and storage technology have enabled organizations to accumulate vast amount of data.

However, extracting useful information has proven extremely challenging.

Often, traditional data analysis tools and techniques cannot be used because of the massive size of a data set.

Sometimes, the non-traditional nature of the data means that traditional approaches cannot be applied even if the data set is relatively small.

Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data.

It has also opened up exciting opportunities for exploring and analyzing new types of data and for analyzing old types of data in new ways.

Business applications

Retailers are allowed to collect up-to-the minute data about customer purchases at the checkout counters of their stores.

web logs from e-commerce.

to help ~~customers~~ customers better understand the needs of their customers and make more informed business decisions.

Data mining techniques can be used to support a wide range of business intelligence applications.

such as
- customer profiling
- targeted marketing
- work-flow management
- store layout,
- fraud detection.

Help retailers:
- who are the most profitable customers?
- what products can be cross ~~sto~~ sold?
- what is the revenue outlook of the company for next year?

Medicine, Science and Engineering.

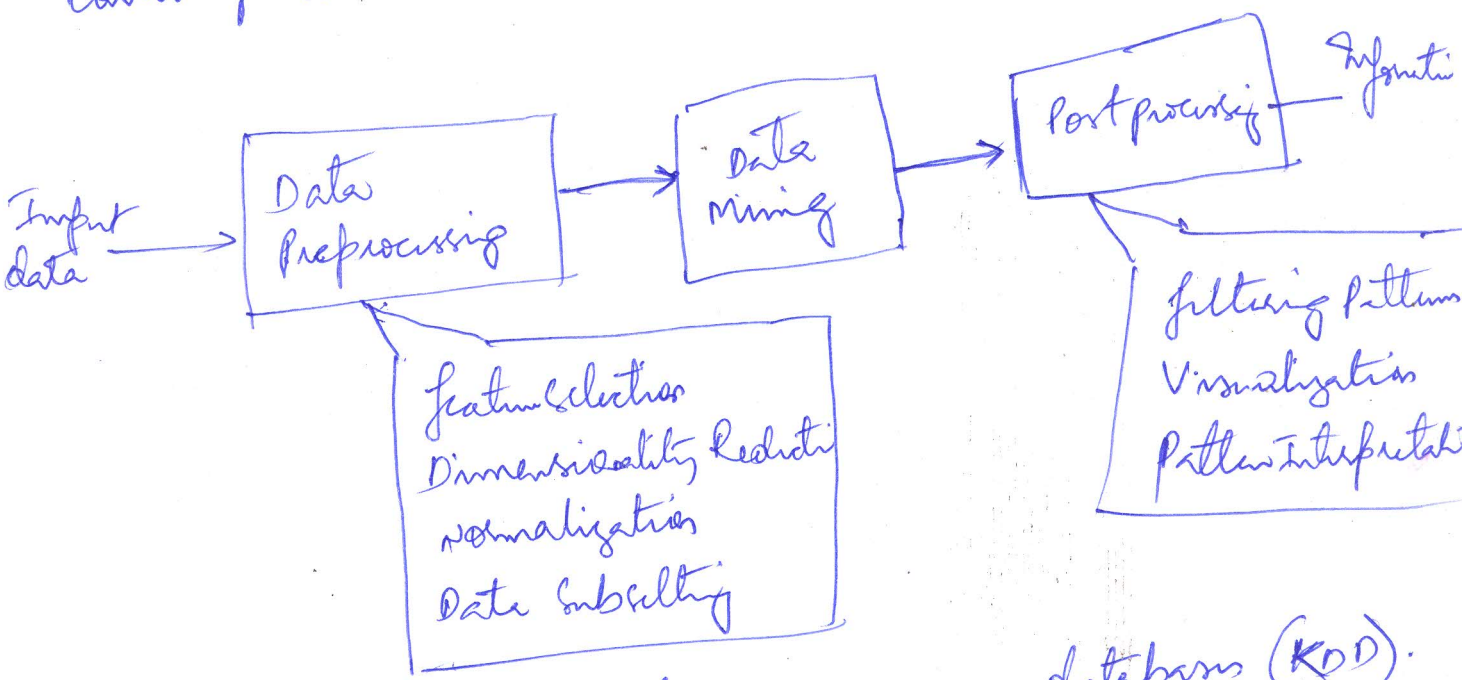
- An important step towards improving our understanding of the Earth's Climate System.
- Data mining can and
  - → what is the relationship between the frequency and intensity of ecosystem disturbances such as droughts and hurricanes to global warming?
  - → How is land surface Precipitation and temperature affected by ocean surface temperature?.
- How well can we predict the beging and end of the growing season for a region?
- Researchers in molecular biology hope to use the large amounts of genomic data currently being gathered to better understand the structure and function of genes.
- Many data mining can also be used to address other important biological challenges such as Protein structure Prediction, multiple sequence alignment, the modeling of biochemical pathways, and phylogenetics.

What is Data Mining?

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

Data mining and Knowledge Discovery

Data Mining is an integral part of knowledge discovery in databases (KDD), which is the overall process of converting raw data into useful information.



The Process of knowledge discovery in databases (KDD).

The input data — Variety of formats
- flat files
- spread sheets
- relational tables.

→ Central repository of distributed across multiple sites.

The Purpose ~~for~~ of Preprocessing is to transform the raw input data into an appropriate format for subsequent analysis.

Motivating challenges:

1. Scalability:

Because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes, or even Petabytes are becoming common.

If data mining algorithms are to handle these massive data sets, then they must be scalable.

2. High Dimensionality:

~~Few~~ ~~handful~~

- Thousands of attributes instead of the handful common a few decades ago.
- For bioinformatics, progress in microarray technology has produced gene expression data involving thousand of features.
  - Temporal data set
  - Spatial data set

Heterogeneous and Complex data :

Heterogeneous attributes.

- webpages containing semi-structured text and hyperlinks;
- DNA data with sequential and three dimensional structure :

Data ownership and Distribution :

The data is geographically distributed among resources belonging to multiple entities.

The Origins of Data Mining :

by the goal of meeting the challenges.

researchers from different disciplines began to focus on developing more efficient and scalable tools that could handle diverse types of data.

This work, which culminated in the field of data Mining

~~built upon the methodology and~~

in particular, data mining draws upon ideas, such as

(1) sampling, estimation, and hypothesis testing from statistics and

(2) search algorithm, modelling techniques.

learning theories from artificial intelligence, pattern recognition, and machine learning.

Data mining has also been quick to adopt ideas from other areas, including optimization, evolutionary computing, information theory, signal processing, visualization and information retrieval.

Data mining as a confluence of many disciplines

- Database systems are needed to provide support for efficient storage, indexing and query processing
- Techniques from high performance (Parallel) computing are often important in addressing the massive size of some of data sets.
- Distributed techniques can also help address the issue of size and are essential when the data cannot be gathered in one location.

Data mining tasks

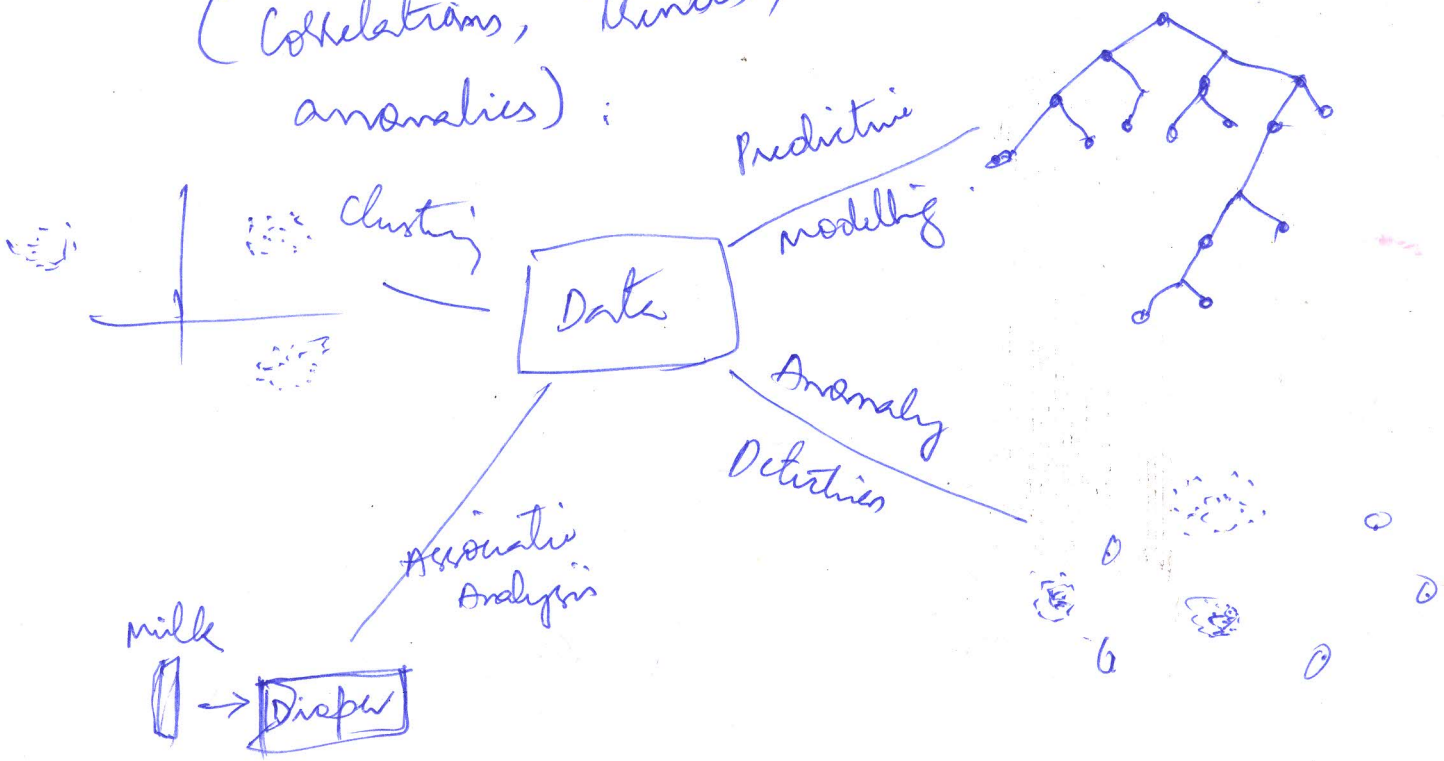
two major categories

Predictive tasks

Predict the value of a particular ~~Feat~~ attribute based on the values of other attributes.

Descriptive tasks

The objective is to derive patterns that summarize the underlying relationship in data. (Correlations, trends, clusters, trajectories and anomalies):



Predictive modeling

Refers to the task of building a model for the target variable as a function of the explanatory variables.

classification : which is used for discrete target variables.

Regression : which is used for continuous target variables.

for example:

Predicting whether a web user will make a purchase at an online bookstore. (classification)

— forecasting the future price of a stock is a regression task. (because price is a continuous-valued attribute).

Association analysis :

used to discover patterns that describe strongly associated features in the data.

Eg: → identifying web pages that are accessed together.

→ ~~select~~ to find items that are frequently bought together.

Cluster analysis

Seeks to find groups of closely related ~~observation so that~~

Eg:- find areas of the ocean that have a significant impact on the earth's climate.

- Document clustering (collection of news articles)

Anomaly detection:

is the task of identifying observations whose characteristics are significantly different from the rest of the data.

A good ~~anomaly~~ detector must have a high detection rate and a low false alarm rate.

Applications - fraud detection
- network intrusions
- unusual patterns of disease
- ecosystem disturbances.

Exercises

1. Discuss whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their gender. : (NO) database query

(b) Dividing the customer of a company according to their profitability. : (NO)

(c) Computing the total sales of a company. (NO)

(d) Sorting a student database based on student identification numbers. (NO)

(e) Predicting the outcomes of tossing a (fair) pair of dice (NO)

(f) Predicting the future stock price of a company using historical records. (yes)

(g) Monitoring the heart rate of a patient for abnormalities. (yes)

(h) Monitoring seismic waves for earthquake activities (yes) (unusual heartbeat)

(i) Extracting the frequencies of a sound wave. (NO) (Signal processing) (Science data)