CS-F415: Data Mining

02 Introduction Continued..

Dr. Kamlesh Tiwari
Associate Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Jan 12, 2024 M/W/F 4:00pm 6101 @ BITS-Pilani [Jan-May 2024]

http://ktiwari.in/dm

KDD

Knowledge Discovering in Databases (KDD) involves

1. **Selection:** collection of data
2. **Preprocessing:** deal with incorrect/missing data
3. **Transformation:** common format and preprocessing
4. **Data Science:** algorithmic tools
5. **Interpretation/Evaluation:** presentation and visualization

Data Mining: Tasks

Two broad categories of data mining models are *Predictive* and *Descriptive*. Some of the related tasks are

- Data Mining
  - A. Predictive
    - 1. Classification
    - 2. Regression
    - 3. Time Series Analysis
    - 4. Prediction
  - B. Descriptive
    - 1. Clustering
    - 2. Summarization
    - 3. Association Rules
    - 4. Sequence Discovery

Classification

Classification maps data into *predefined* labels.

Example: Lots of mails are there in my mail box. Can you tell me which are SPAM?

- Task of supervised learning
- Often based on some patterns or characteristics
- We can use the frequency of words
- Assumption is that some words appears more or less frequently in SPAM

Regression

Regression is used to map data into *real valued* variable.

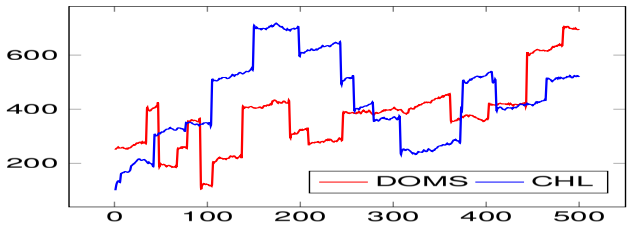
Example: What is the cost of my house?

- Task of supervised learning
- We have data about the cost of house based on features such as
  - location
  - Plot area
  - number of rooms
  - garden available or not
  - how old it is
- Current economical conditions can also matter
- Dimensionality is high

Time Series Analysis

In time series analysis the value of attribute is examined over time.

Example: Which stock is better?



- The values are obtained as evenly spaced time points (daily, weekly, hourly, etc.)
- Distance measures are used to find similarity
- Structural analysis is done

Prediction

Predicting future data states based on current or historical data.

Example: What comes next?

$2, 4, 6, 8, 10, ...?...$

$2, 3, 5, 7, ...?..., 13$

$(10jul, rain), (11jul, rain), (12jul, no - rain), (13jul, ...?...)$

- Predication can sometimes be seen as classification
- Application includes weather, flood, pattern recognition.

Clustering

Clustering is similar to classification except the groups are not pre-defined.

Example: How many kind of files are there in my directory?

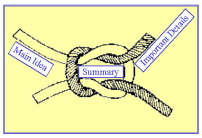
- Unsupervised learning setting
- We can use file name
- Words it has

Example: Who would take my offer?

- The database has information about age, gender, income, location, .. etc.

Summarization

Summarization maps data into subsets with associated simple descriptions. It is also called characterization or generalization.



Example: How to compare two universities?

- Average JEE rank
- Average number of publication
- Student/Faculty ratio
- Combination

Association Rules (Apriori)

Tries to do linked analysis.

Example: Whether sames products are selling together?

- $I = \{i_1, i_2, i_3, ...i_m\}$, $T = \{t_1, t_2, t_3, ..., t_n\}$ and $t_i \subseteq I$
- Minimum support count should be maintained
- Can you see: Subset of frequent items is also frequent

Sequence Discovery

Sequence Discovery is used to discover sequential patterns in the data.

Example: what is my website access pattern?

- Pattern is based on a time sequence of a action
- It is pattern discovery problem

KDD Issues

- Human interaction
- Overfitting, Outliers
- Large dataset,
- High dimension
- Multimedia data
- Missing data
- Irrelevant data
- Noisy data, quality
- Changing data

And much more...

Complex and Heterogeneous Data, Data Ownership and Distribution, Privacy Preservation, Streaming Data

Bonferroni Principle: Big Data can result in highly unlikely outcomes masquerading as statistically sound. [1]

[1] David Rhine, in 1950's tested students for Extra Sensory Perception (ESP) by asking them to guess 10 cards - red or black. He found about 1/1000 of them guessed all 10 (out of random guessing) he declared them to have ESP. At retest they did no better than average.

Applications

In many domains including finance, robotics, bioinformatics, vision, natural language, *etc*.

- Spam filtering
- Speech/handwriting recognition
- Object detection/recognition
- Weather prediction
- Stock market analysis
- Search engines (e.g, Google)
- Ad placement on websites
- Adaptive website design
- Credit-card fraud detection
- Web page clustering (e.g.,Google News)
- Machine Translation (e.g., Google Translate)
- Recommendation systems (e.g., Netflix, Amazon)
- Classifying DNA sequences
- Automatic vehicle navigation
- Performance tuning of computer systems
- Predicting good compilation flags for programs

`and many more...`

Types of Learning

- **Supervised:** "right answers" are provided by teacher for sufficient training examples. Computer tells "right answers" for new input. Performance measure. (Classification and regression)
- **Unsupervised:** "right answers" are NOT provided, computer tries to make sense of the data. How good the spread of items is. (clustering and association rule)
- **Semi-supervised:** "right answers" are provided for few training examples only
- **Active:** computer can ask questions. Needs less training. Opposite is passive learning
- **Lazy:** learner do not consolidate the findings.
- **Reinforced:** hit and trial method to minimize cost. (game playing)
- **Transfer:** Learning a task B to do A. (cycle riding for bike riding)
- **Deep:** processing like human brain

Challenges

1. How good is the model
   Accuracy, CRR, EER, FAR, FRR, ROC ...
2. How do I choose a model
   Decision Tree, SVM, Neural Network, ... ?
3. Do I have enough data
   Pre-processing, augmentation, ...?
4. Is data of sufficient quality
   Error/Noise in data, missing values ... ?
5. How confidence the result is
   Significance, Probability ... ?
6. Am I describing the data correctly
   whether features are correct ?
7. How fair the system is
   if it behaves equally to all?

Thank You!

Thank you very much for your attention!

Queries ?