CS-F415: Data Mining

01 Introduction and Logistics

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

Jan 10, 2024 M/W/F 4:00pm 6101 @ BITS-Pilani [Jan-May 2024]

http://ktiwari.in/dm

Welcome to CS-F415

Instructors:
Prof. Kamlesh Tiwari, kamlesh.tiwari@pilani.bits-pilani.ac.in [IC]
Prof. Yashvardhan Sharma, yash@pilani.bits-pilani.ac.in
Ms. Vijay Kumari, p20190065@pilani.bits-pilani.ac.in [LAB]
Ms. Sakshi, p20180437@pilani.bits-pilani.ac.in [LAB]

Schedule: OFFLINE at 6101 NAB, 4-4:50PM Mon/Wed/Fri

Lab: Th/S 2-4PM @ 6018/6017.

Chamber Consultation Hour: Tuesday 4pm to 5pm @6111-C

Book: Tan P. N., Steinbach M & Kumar V. "Introduction to Data Mining" Pearson Education, 2016

Course website: https://nalanda-aws.bits-pilani.ac.in/

Introduction

What is?

Data: Fact or values

Information: Processed output of date

Knowledge: Understanding of information

X1	X2	X3		Y
2	3	4		1
8	7	5		10
9	6	1		14
7	4	9		2
1	5	5		1
5	3	6		2
7	4	4		7
6	3	5		4
3	2	3		

Basically a model.

Perspective: Knowledge Discovery in Databases

What is **data-mining**?

Computation to facilitate Knowledge Discovery in Databases (**KDD**)

Goal of Data Mining (motivation of doing the same?)

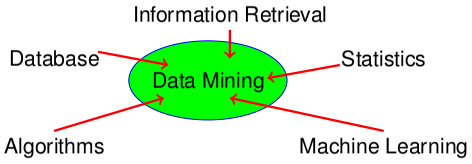
To provide efficient tools and techniques for KDD

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

Data mining is fairly involved discipline. It includes many fields such as database, information retrieval, statistics, and machine learning.



It differs from traditional query processing

- **Query:** not well formed. Miner may not know what he wants.
- **Data:** different version. Preprocessed and modified.
- **Output:** may not a subset. It could be an analysis.

Data Mining Parts

Data Mining has three parts

1. **Model:** is to be fit on data
2. **Search:** technique to evaluate data point
3. **Preference:** criteria to select one model over other

Example:

Assume a credit card company wants to decide whether a transaction should be

1. Authorized
2. Ask for more information
3. Decline

Search requires evaluation of past data. **Model** associates with the criteria to decide for one of the categories. **Preference** is given to criteria that suits the data best (want to reduce number of frauds or amount of fraud).

What is Data Mining and What is Not

What is NOT Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”
- Whether the list of items is sorted?

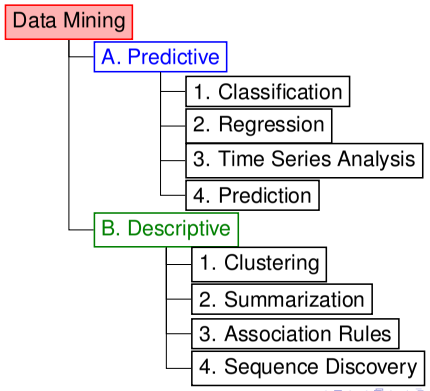
What is Data Mining?

- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com, ...)
- Customers who buy diapers are more likely to buy beer

There should be a pattern. It is fine if we can not explain or get that.

Data Mining: Tasks

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



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.

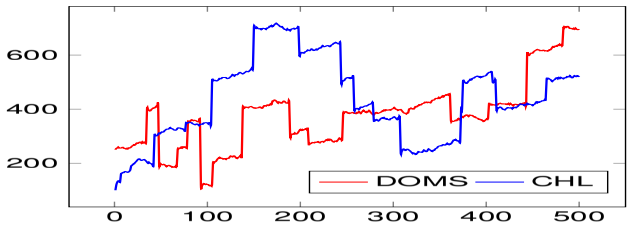
Example: What is the cost of my house?

- 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?

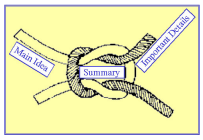
- Unsupervised learning setting
- We can use file name
- Words it has

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?

Contents

Introduction to Data Mining, Motivation, What is Data Mining?, Data Mining Tasks, Issues in Data Mining, Applications, Data Preprocessing, Types of data, Data Quality, Data preprocessing, Similarity and Dissimilarity Measures, Data Exploration, Data Set & its Statistics, Visualization, OLAP & Multidimensional Data Analysis, Classification Alternative Techniques, Rule Based Classifier, Nearest Neighbor Classifier, Bayesian Classification, Support Vector Machine, Ensemble Classifiers, Class Imbalance Problem, Multiclass Problem, Association Rule Mining, Introduction, Applications, Market-Basket Analysis, Frequent Itemsets, Apriori Algorithm, Alternative Methods, Advanced Association Rule Mining, Generalized Association Rules, Multilevel Association Rules, Multidimensional Association Rules, Temporal Association Rules, Infrequent Patterns, Constrained Based Association Rules, Clustering, Introduction, Applications, Partitioning Algorithms, Hierarchical Algorithms, Density based Algorithms, Cluster Evaluation, Clustering: Additional Issues and Algorithms, Characteristics of Data, Clusters and clustering, Algorithms, Graph Based Clustering, Scalable Clustering Algorithms, Anomaly Detection, Preliminaries, Statistical Approaches, Proximity based Outlier Detection, Density based Outlier Detection, Clustering Based Techniques, Advanced Topics, Web Mining, Incremental Algorithms for Data Mining, Stream Data Mining

Learning Goal and Evaluation

Evaluation Scheme (Aug-Dec 2023)

1. Mid Semester Test (Closed Book) Mar 14, 2024	25%
2. Lab [Participation + Coding Exam] 15%	5%+10%
3. Term Paper [task/setup/innovate/result] 20%	4%+6%+6%+4%
5. Comprehensive (Partially Open) May 15, 2024	35%

Course Webpage: `http://ktiwari.in/dm`

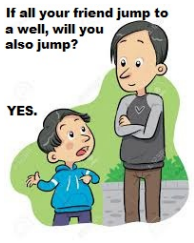
Text Book: *Introduction to Data Mining* by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar 2^{nd} Edition

Solution must always be written independently

Thank You!

Thank you very much for your attention!

Queries ?



`Any sufficiently advanced technology is indistinguishable from magic` - Arthur C. Clarke