18CS54- Data Mining (DM)

Lecture 3 : Introduction to KDD and DM

Dr. Vani Vasudevan

OUTLINE

- MOTIVATION: WHY DATA MINING?
- WHAT IS DATA MINING?
- DATA MINING TASKS…
- DATA MINING : A KDD PROCESS
- DATA MINING: ON WHAT KINDS OF DATA?
- CHALLENGES OF DATA MINING

Source : Data Mining: Concepts and Techniques, Han & Kamber

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}

3

Source Introduction to Data Mining , Tan

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!

Source Introduction to Data Mining , Tan

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!

Source Introduction to Data Mining , Tan

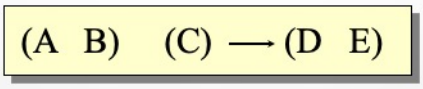
ASSOCIATION RULE DISCOVERY: APPLICATION 3

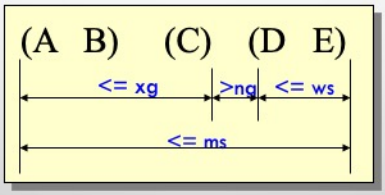
- **Inventory Management:**
  - **Goal:** A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - **Approach:** process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Source Introduction to Data Mining , Tan

SEQUENTIAL PATTERN DISCOVERY: DEFINITION

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.
- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.





7

Source Introduction to Data Mining , Tan

SEQUENTIAL PATTERN DISCOVERY: EXAMPLES

- In telecommunications alarm logs,
  - (inverter_problem excessive_line_current)

    (rectifier_alarm) --> (fire_alarm)
- In point-of-sale transaction sequences,
  - Computer bookstore:

    (intro_to_visual_c) (c++_primer) -->
    (perl_for_dummies,tcl_tk)
  - Athletic apparel store:

    (shoes) (racket, racketball) --> (sports_jacket)

Source Introduction to Data Mining , Tan

REGRESSION

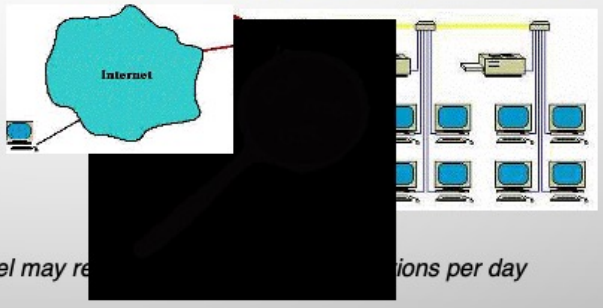
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- **Examples:**
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Source Introduction to Data Mining , Tan

DEVIATION/ANOMALY DETECTION

- DETECT SIGNIFICANT DEVIATIONS FROM NORMAL BEHAVIOR
- APPLICATIONS:
  - CREDIT CARD FRAUD DETECTION

  - NETWORK INTRUSION DETECTION



Typical network traffic at University level may re ions per day

Source Introduction to Data Mining , Tan

CHAPTER 1 - EXERCISES

- 1.7. EXERCISES- **PG. NO.14 & 15** PANG-NING TAN, VIPIN KUMAR, MICHAEL STEINBACH: **INTRODUCTION TO DATA MINING,** PEARSON, 2012.