Mining Association Rules

——Mining Various Kinds of Association Rules——

徐华

清华大学 计算机系 智能技术与系统国家重点实验室

xuhua@tsinghua.edu.cn

1

Association and Correlations

- Association and Correlations
- Efficient and Scalable Frequent Itemset Mining Methods
- Mining Various Kinds of Association Rules
- From Association Mining to Correlation Analysis
- Constraint-based Association Mining

2

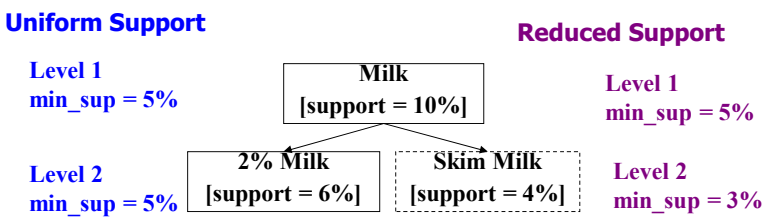
Mining Various Kinds of Association Rules

- Mining multi-level association
  - concept hierarchy
- Mining multi-dimensional association
  - Age, item, occupation
- Mining quantitative association
- Mining interesting correlation patterns

3

Multiple-level Association Rules

- Items often form hierarchy
- Flexible support settings: Items at the lower level are expected to have lower support.
- Transaction database can be encoded based on dimensions and levels.
- Explore shared multi-level mining.



4

Multi-dimensional Association

- Single-dimensional rules:

  buys(X, "milk") ⇒ buys(X, "bread")

- Multi-dimensional rules: ≥ 2 dimensions or predicates
  - Inter-dimension assoc. rules (*no repeated predicates*)

    age(X," 19-25") ∧ occupation(X, "student") ⇒ buys(X, "coke")
  - Hybrid-dimension assoc. rules (*repeated predicates*)

    age(X," 19-25") ∧ buys(X, "popcorn") ⇒ buys(X, "coke")
- Categorical Attributes
  - finite number of possible values, no ordering among values
- Quantitative Attributes
  - numeric, implicit ordering among values

5

ML/MD Associations with Flexible Support Constraints

- Why flexible support constraints?
  - Real life occurrence frequencies vary greatly
    - Diamond, watch, pens in a shopping basket
  - Uniform support may not be an interesting model
- A flexible model
  - The lower-level, the more dimension combination, and the long pattern length, usually the smaller support
  - General rules should be easy to specify and understand
  - Special items and special group of items may be specified individually and have higher priority

6

Multi-level Association: Redundancy Filtering

- Some rules may be redundant due to "ancestor" relationships between items.
- Example
  - Desktop computer ⇒ b/w printer [support = 8%, confidence = 70%]
  - IBM Desktop computer ⇒ b/w printer [support = 2%, confidence = 72%]
- We say the first rule is an ancestor(祖先) of the second rule.
- A rule is redundant if its support is close to the "expected" value, based on the rule's ancestor.

7

Multi-Level Mining: Progressive Deepening

- A top-down, progressive deepening approach:
  - First mine high-level frequent items:
    milk (15%), bread (10%)
  - Then mine their lower-level "weaker" frequent itemsets:
    Nest milk (5%), wheat bread (4%)
- Different min_support threshold across multi-levels lead to different algorithms:
  - If adopting the same *min_support* across multi-levels
    then toss *t* if any of *t*' s ancestors is infrequent.
  - If adopting reduced *min_support* at lower levels
    then examine only those descendents (后裔) whose ancestor's support is frequent/non-negligible.

8

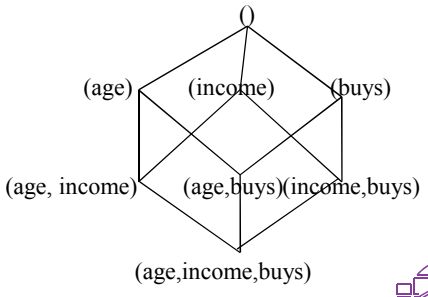
Techniques for Mining MD Associations

- Search for frequent *k*-predicate set:
  - Example: {age, occupation, buys} is a 3-predicate set.
  - Techniques can be categorized by how age is treated.
- Using static discretization of quantitative attributes
  - Quantitative attributes are statically discretized by using predefined concept hierarchies.
- Quantitative association rules
  - Quantitative attributes are dynamically discretized into "bins" based on the distribution of the data.
- Distance-based association rules
  - This is a dynamic discretization process that considers the distance between data points.

9

Static Discretization of Quantitative Attributes

- Discretized prior to mining using concept hierarchy.
- Numeric values are replaced by ranges.
- In relational database, finding all frequent k-predicate sets will require *k* or *k*+1 table scans.
- Data cube is well suited for mining.
- The cells of an n-dimensional
  - cuboid correspond to the predicate sets.
- Mining from data cubes can be much faster.



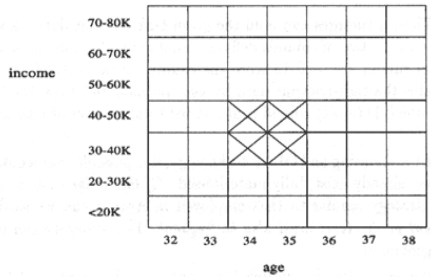
10

Quantitative Association Rules

- Numeric attributes are *dynamically* discretized
  - Such that the confidence or compactness of the rules mined is maximized.
- 2-D quantitative association rules: Aquan1 ∧ Aquan2 ⇒ Acat
- Cluster "adjacent"
  - Association rules to form general rules using a 2-D grid.
- Example:

  age(X," 30-34") ∧ income(X," 24K - 48K")

  ⇒ buys(X," high resolution TV")



11

Mining Distance-based Association Rules

- Binning methods do not capture the semantics of interval data

Price($)	Equi-width (width $10)	Equi-depth (depth 2)	Distance-based
7	[0,10]	[7,20]	[7,7]
20	[11,20]	[22,50]	[20,22]
22	[21,30]	[51,53]	[50,53]
50	[31,40]		
51	[41,50]		
53	[51,60]		

- Distance-based partitioning, more meaningful discretization considering:
  - Density/number of points in an interval
  - "closeness" of points in an interval

12

Thanks !