Unsupervised Data Mining

Association Rule Learning

Association Rule Analysis

- Popular in mining data bases
- Automated discovery of sets of variables that occur frequently or one(s) leading to other(s)

Feature	Demographic	# values	Type
1	sex	2	categorical
2	marital status	5	categorical
3	age	7	ordinal
4	education	6	ordinal
5	occupation	9	categorical
6	income	9	ordinal
7	years in Bay Area	5	ordinal
8	dual incomes	3	categorical
9	number in household	9	ordinal
10	number of children	9	ordinal
11	householder status	3	categorical
12	type of home	5	categorical
13	ethnic classification	8	categorical
14	language in home	3	categorical

Association Rule Analysis (cont)

Association rule 1: Support 25%, confidence 99.7% and lift 1.03.

$$\left[\begin{array}{rcl} \text{number in household} & = & \mathit{1} \\ \text{number of children} & = & \mathit{0} \end{array}\right]$$

$$\Downarrow$$

$$\text{language in home} = \mathit{English}$$

Association rule 2: Support 13.4%, confidence 80.8%, and lift 2.13.

$$\left[\begin{array}{rcl} \text{language in home} & = & \mathit{English} \\ \text{householder status} & = & \mathit{own} \\ \text{occupation} & = & \{\mathit{professional/managerial}\} \end{array}\right]$$

$$\Downarrow$$

$$\text{income} \geq \mathit{\$40{,}000}$$

Association rule 3: Support 26.5%, confidence 82.8% and lift 2.15.

$$\left[\begin{array}{rcl} \text{language in home} & = & \mathit{English} \\ \text{income} & < & \mathit{\$40{,}000} \\ \text{marital status} & = & \mathit{not\ married} \\ \text{number of children} & = & \mathit{0} \end{array}\right]$$

$$\Downarrow$$

$$\text{education} \notin \{\mathit{college\ graduate,\ graduate\ study}\}$$

Market Basket Analysis

- Retail outlets
  - Placement of merchandises (affinity positioning)
  - Cross advertising
- Banks
- Insurance
  - link analysis for fraud
- Medical
  - symptom analysis

Co-occurrence Matrix

Customer 1: beer, pretzels, potato chips, aspirin

Customer 2: diapers, baby lotion, grapefruit juice, baby food, milk

Customer 3: soda, potato chips, milk

Customer 4: soup, beer, milk, ice cream

Customer 5: soda, coffee, milk, bread

Customer 6: beer, potato chips

	Beer	Pot. Chips	Milk	Diap.	Soda
Beer	3	2	1	0	0
Pot. Chips	2	3	1	0	1
Milk	1	2	4	1	2
Diapers	0	0	1	1	0
Soda	0	1	2	0	2

beer & potato chips - makes sense   milk & soda - probably noise

- Interesting cases can have 10^4 variables and 10^8 of samples
- Co-occurrence gives only pair-wise association

Practical Solutions

- Run up against curse-of-dimensionalities
  - With 10^4 variables, each with many possible values, need very large # of samples to populate the space, "bump" hunting in fine scale is not possible
    - Look for regions in the probability spaces with high density
  - Even for binary variables, there are 2^k (e.g., 2^{1,000} possible 1,0-tuples, must have efficient search algorithms

Simplification

- Assuming binary variables
- If not, force them binaries
  - Instead of 6 different education levels, just 2 (college and above, or below)
- Change of variables
  - Initially (X1,…, Xp)
  - Each with (S1, … Sp) possible values
  - K = S1+ … Sp
  - Create Zk binary variables
    - 1 if the corresponding variable Xi assuming value Sij
    - 0 otherwise

Apriori Algorithm

- Threshold t
- 1^{st} pass:
  - Single-variable set: must have occurrence larger than t
- 2^{nd} pass:
  - Pair-wise variable sets: together must have occurrence large than t
- …
- mth pass:
  - Only those tuples in (m-1)th pass have probability higher than t are considered
- To avoid combinatorial explosion, t cannot be too low

Tuples to Rules

- Tuples {Zk} to A=>B
  - A antecedent
  - B consequent
  - T(A=>B): support, probability of simultaneously observing A and B P(A&B)
  - C(A=>B) = T(A=>B)/T(A): confidence, probability of P(B|A)
  - L(A=>B) = C(A=>B)/T(B): lift, probability of P(A&B)/(P(A)P(B))

Examples

- K={peanut butter, jelly, bread}
- {peanut butter, jelly} => bread
- Support of 0.03: if {peanut butter, jelly, bread} appears in 3% of sample baskets
- Confidence of 82%: if peanut butter and jelly are purchased, then 82% time bread is also
- Lift of 1.9: If bread appear in 43% of sample baskets, then 0.82/0.43=1.9

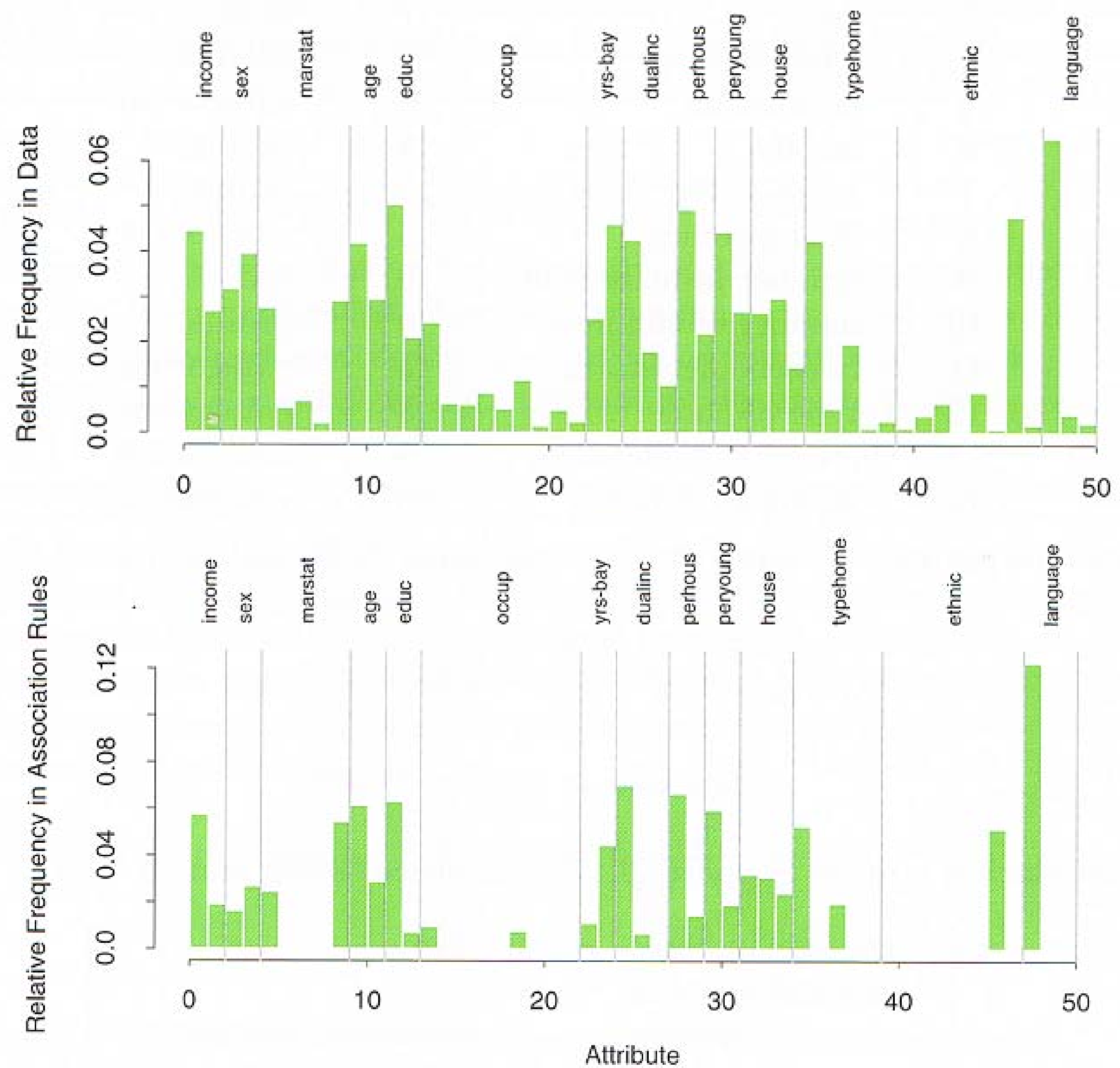

FIGURE 14.2. *Market basket analysis: relative frequency of each dummy variable (coding an input category) in the data (top), and the association rules found by the Apriori algorithm (bottom).*