Association Rules

Charles Sutton
Data Mining and Exploration
Spring 2012

Based on slides by Chris Williams and Amos Storkey

The Goal

- Find "patterns": local regularities that occur more often than you would expect. Examples:
  - If a person buys wine at a supermarket, they also buy cheese. (confidence: 20%)
  - If a person likes Lord of the Rings and Star Wars, they like Star Trek (confidence: 90%)
- Look like they could be used for classification, but
  - There is not a single class label in mind. They can predict any attribute or a set of attributes. They are *unsupervised*
  - Not intended to be used together as a set
- Often mined from very large data sets

Example Data

Market basket analysis, e.g., supermarket

Item

Transactions
trip to market

Chicken	Onion	Rocket	Caviar	Haggis	
1		1		1	
	1	1		1	
1	1			1	
				1	
		1		1	
1				1	

. . . .

These are databases that companies have already.

Other Examples

- Collaborative-filtering type data: e.g., Films a person has watched
- Rows: patients, columns: medical tests (Cabena et al, 1998)
- Survey data (Impact Resources, Inc., Columbus OH, 1987)

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

Toy Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	False	No
D2	Sunny	Hot	High	True	No
D3	Overcast	Hot	High	False	Yes
D4	Rain	Mild	High	False	Yes
D5	Rain	Cool	Normal	False	Yes
D6	Rain	Cool	Normal	True	No
D7	Overcast	Cool	Normal	True	Yes
D8	Sunny	Mild	High	False	No
D9	Sunny	Cool	Normal	False	Yes
D10	Rain	Mild	Normal	False	Yes
D11	Sunny	Mild	Normal	True	Yes
D12	Overcast	Mild	High	True	Yes
D13	Overcast	Hot	Normal	False	Yes
D14	Rain	Mild	High	True	No

Itemsets, Coverage, etc

- Call each column an attribute $A_1, A_2, \ldots A_m$
- An item set is a set of attribute value pairs

$$(A_{i_1} = a_{j_1}) \wedge (A_{i_2} = a_{j_2}) \wedge \ldots (A_{i_k} = a_{j_k})$$

- Example: In the Play Tennis data

$$\text{Humidity} = \text{Normal} \wedge \text{Play} = \text{Yes} \wedge \text{Windy} = \text{False}$$

- The support of an item set is its frequency in the data set
  - Example:
    - support ($\text{Humidity} = \text{Normal} \wedge \text{Play} = \text{Yes} \wedge \text{Windy} = \text{False}$) = 4
- The confidence of an association rule if Y=y then Z=z is $P(Z = z | Y = y)$
  - Example:

$$P(\text{Windy} = \text{False} \wedge \text{Play} = \text{Yes} | \text{Humidity} = \text{Normal}) = 4/7$$

Item sets to rules

- First: We will find frequent item sets
- Then: We convert them to rules
- An itemset of size k can give rise to 2^k-1 rules
- Example: itemset

```
Windy=False, Play=Yes, Humidity=Normal
```

- Results in 7 rules including:

```
IF Windy=False and Humidity=Normal THEN Play=Yes           (4/4)
IF Play=Yes THEN Humidity=Normal and Windy=False           (4/9)
IF True THEN Windy=False and Play=Yes and Humidity=Normal  (4/14)
```

- We keep rules only whose confidence is greater than a threshold

Finding Frequent Itemsets

- Task: Find all item sets with support
- Insight: A large set can be no more frequent than its subsets, e.g.,

$$\text{support}(\text{Wind} = \text{False}) \geq \text{support}(\text{Wind} = \text{False}, \text{Outlook} = \text{Sunny})$$

- So search through itemsets in order of number of items
- An efficient algorithm for this is APRIORI (Agarwal and Srikant, 1994; Mannila et al, 1994)

APRIORI Algorithm

(for binary variables)

$i = 1$
$C_i = \{\{A\} | A \text{ is a variable}\}$
while C_i is not empty
    database pass:
        for each set in C_i test if it is frequent
        let L_i be collection of frequent sets from C_i
    candidate formation:
        let C_{i+1} be those sets of size $i + 1$
        all of whose subsets are frequent
end while

Single database pass is linear in $|C_i|n$, make a pass for each i until C_i is empty

Candidate formation

- Find all pairs of sets $\{U, V\}$ from L_i such that $U \cup V$ has size $i + 1$ and test if this union is really a potential candidate. $O(|L_i|^3)$

Example: 5 three-item sets
(ABC), (ABD), (ACD), (ACE), (BCD)
Candidate four-item sets
(ABCD) ok
(ACDE) not ok because (CDE) is not present above

Comments

- Some association rules will be trivial, some interesting. Need to sort through them
  - Example: pregnant => female (confidence: 1)
- Also can miss "interesting but rare" rules
  - Example: vodka --> caviar (low support)
- Really this is a type of exploratory data analysis
- For rule A -->B, can be useful to compare P(B|A) to P(B)
- APRIORI can be generalised to structures like subsequences and subtrees