DATA ANALYTICS

Unit 4: Rule Generation (Apriori Algorithm) + Evaluation of Recommender Systems

Gowri Srinivasa

Department of Computer Science and Engineering

Apriori Algorithm for Frequent Itemset Generation

Method:

- Let k=1
- Generate frequent itemsets of length 1
- Repeat until no new frequent itemsets are identified
  - Generate length (k+1) candidate itemsets from length k frequent itemsets
  - Prune candidate itemsets containing subsets of length k that are infrequent
  - Count the support of each candidate by scanning the DB
  - Eliminate candidates that are infrequent, leaving only those that are frequent

Rule Generation

Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement

If {A,B,C,D} is a frequent itemset, candidate rules:

$ABC \rightarrow D$,	$ABD \rightarrow C$,	$ACD \rightarrow B$,	$BCD \rightarrow A$,
$A \rightarrow BCD$,	$B \rightarrow ACD$,	$C \rightarrow ABD$,	$D \rightarrow ABC$
$AB \rightarrow CD$,	$AC \rightarrow BD$,	$AD \rightarrow BC$,	$BC \rightarrow AD$,
$BD \rightarrow AC$,	$CD \rightarrow AB$		

If $|L| = k$, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \varnothing$ and $\varnothing \rightarrow L$)

Rule Generation

How to efficiently generate rules from frequent itemsets?

In general, confidence does not have an anti-monotone property

$c(ABC \rightarrow D)$ can be larger or smaller than $c(AB \rightarrow D)$

But confidence of rules generated from the same itemset has an anti-monotone property e.g., L = {A,B,C,D}:

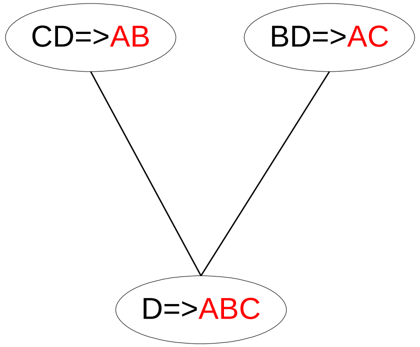
$$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

Confidence is anti-monotone w.r.t. number of items on the RHS of the rule

Candidate rule is generated by merging two rules that share the same prefix in the rule consequent

join(CD=>AB,BD=>AC)
would produce the candidate rule D => ABC

Prune rule D=>ABC if its
subset AD=>BC does not have high confidence



Slide courtesy of Tan,Steinbach, Kumar, Introduction to Data Mining

Support and Confidence

$$support(A \Rightarrow B) = P(A \cup B)$$

$$confidence(A \Rightarrow B) = P(B|A).$$

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

Minimum support (minsup)

- Note that the itemset support defined is sometimes referred to as *relative support*, whereas the occurrence frequency is called the **absolute support**.
- If the relative support of an itemset *I* satisfies a prespecified **minimum support threshold** (i.e., the absolute support of *I* satisfies the corresponding **minimum support count threshold**), then *I* is a **frequent** itemset.
- The set of frequent *k*-itemsets is commonly denoted by L_k

Applying multiple minimum support

How to apply multiple minimum support?

MS(i): minimum support for item i

e.g.: MS(Milk)=5%, MS(Coke) = 3%,
MS(Broccoli)=0.1%, MS(Salmon)=0.5%

MS({Milk, Broccoli}) = min (MS(Milk), MS(Broccoli))
= 0.1%

Challenge: Support is no longer anti-monotone

Suppose: Support(Milk, Coke) = 1.5% and
Support(Milk, Coke, Broccoli) = 0.5%

{Milk,Coke} is infrequent but {Milk,Coke,Broccoli} is frequent

Order the items according to their minimum support (in ascending order)

e.g.: MS(Milk)=5%, MS(Coke) = 3%,
MS(Broccoli)=0.1%, MS(Salmon)=0.5%

Ordering: Broccoli, Salmon, Coke, Milk

Need to modify Apriori such that:

L_1 : set of frequent items

F_1 : set of items whose support is $\geq$ MS(1)
where MS(1) is $\min_i$(MS(i))

C_2 : candidate itemsets of size 2 is generated from F_1 instead of L_1

Multiple minimum support and modified Apriori

Order the items according to their minimum support (in ascending order)

e.g.: MS(Milk)=5%, MS(Coke) = 3%,
MS(Broccoli)=0.1%, MS(Salmon)=0.5%

Ordering: Broccoli, Salmon, Coke, Milk

Need to modify Apriori such that:

L_1 : set of frequent items

F_1 : set of items whose support is $\geq$ MS(1)
where MS(1) is $\min_i$(MS(i))

C_2 : candidate itemsets of size 2 is generated from F_1
instead of L_1

Modifications to Apriori: In traditional Apriori, A candidate (k+1)-itemset is generated by merging two frequent itemsets of size k

The candidate is pruned if it contains any infrequent subsets of size k

Pruning step has to be modified:

Prune only if subset contains the first item

e.g.: Candidate={Broccoli, Coke, Milk} (ordered according to minimum support)

{Broccoli, Coke} and {Broccoli, Milk} are frequent but {Coke, Milk} is infrequent

Candidate is not pruned because {Coke, Milk} does not contain the first item, i.e., Broccoli.

Evaluation of an association rule

Contingency table for $X \rightarrow Y$

	Y	Y	
X	f_{11}	f_{10}	f_{1+}
X	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	$\|T\|$

f_{11}: support of X and Y
f_{10}: support of X and $\overline{\text{Y}}$
f_{01}: support of $\overline{\text{X}}$ and Y
f_{00}: support of $\overline{\text{X}}$ and $\overline{\text{Y}}$

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

$$PS = P(X,Y) - P(X)P(Y)$$

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

Limitation of Confidence

	Coffee	Not Coffee	
Tea	15	5	20
Not Tea	75	5	80
	90	10	100

Association Rule: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75 (75% of those who drink tea also drink coffee)

but P(Coffee) = 0.9 (90% of the people in our sample drink coffee (most of them do!))

$\Rightarrow$ Although confidence is high, rule is misleading

$\Rightarrow$ P(Coffee|NotTea) = 0.9375 (more interesting/ meaningful that nearly 94% of those who do not drink tea, drink coffee)

$\Rightarrow$ One is more likely to drink coffee if they do not drink tea (than if they do drink tea)

Computing Confidence

- In confidence of rule equation A => B can be easily derived from the support counts of A and A∪B.
- That is, once the support counts of A, B, and A ∪B are found, it is straightforward to derive the corresponding association rules A =>B and B =>A and check whether they are strong.
- Thus, the problem of mining association rules can be reduced to that of mining frequent itemsets.

Continuous and Categorical Attributes

How to apply association analysis formulation to non-asymmetric binary variables?

Session Id	Country	Session Length (sec)	Number of Web Pages viewed	Gender	Browser Type	Buy
1	USA	982	8	Male	IE	No
2	China	811	10	Female	Netscape	No
3	USA	2125	45	Female	Mozilla	Yes
4	Germany	596	4	Male	IE	Yes
5	Australia	123	9	Male	Mozilla	No
…	…	…	…	…	…	…

Example of Association Rule:

{Number of Pages $\in$[5,10) $\wedge$ (Browser=Mozilla)} $\rightarrow$ {Buy = No}

Transform categorical attribute into asymmetric binary variables

Introduce a new "item" for each distinct attribute-value pair

Example: replace Browser Type attribute with

Browser Type = Internet Explorer

Browser Type = Mozilla

Browser Type = Mozilla

Handling of Categorical Attributes

Potential Issues

What if an attribute has many possible values?

Example: attribute country has more than 200 possible values

Many of the attribute values may have very low support

Potential solution: Aggregate the low-support attribute values

What if distribution of attribute values is highly skewed?

Example: 95% of the visitors have Buy = No

Most of the items will be associated with (Buy=No) item

Potential solution: drop the highly frequent items

Multiple minimum support also comes in handy in both cases

Handling of Continuous Attributes

Different kinds of rules:

- $Age \in [21,35) \wedge Salary \in [70k,120k) \rightarrow Buy$
- $Salary \in [70k,120k) \wedge Buy \rightarrow Age: \mu=28, \sigma=4$

Different methods:

- Discretization-based
- Statistics-based (mean, median, standard deviation, etc.)
- Non-discretization based minApriori (concept hierarchy)

Discretization-based

Unsupervised:

- Equal-width binning
- Equal-depth binning
- Clustering

Supervised:

Attribute values, v

Class	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
Anomalous	0	0	20	10	20	0	0	0	0
Normal	150	100	0	0	0	100	100	150	100

$\underbrace{v_1\ v_2}_{bin_1}$ $\underbrace{v_3\ v_4\ v_5}_{bin_2}$ $\underbrace{v_6\ v_7\ v_8\ v_9}_{bin_3}$

Evaluation – objective measures

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2 + P(\overline{B}\|A)^2] + P(\overline{A})[P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2]$ $-P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\|B)^2 + P(\overline{A}\|B)^2] + P(\overline{B})[P(A\|\overline{B})^2 + P(\overline{A}\|\overline{B})^2]$ $-P(A)^2 - P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\Big(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\Big)$
13	Conviction (V)	$\max\Big(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\Big)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B) - P(A)P(B)$
17	Certainty factor (F)	$\max\Big(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\Big)$
18	Added Value (AV)	$\max(P(B\|A) - P(B), P(A\|B) - P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$

It is sufficient if we understand the idea behind the measures and are able to use some of these, such as, support, confidence, lift (or interest), phi-coefficient to evaluate a confidence rule or test for independence of (or correlation) between itemsets

Slide courtesy of Tan,Steinbach, Kumar, Introduction to Data Mining

Evaluation – subjective measures

Objective measure:

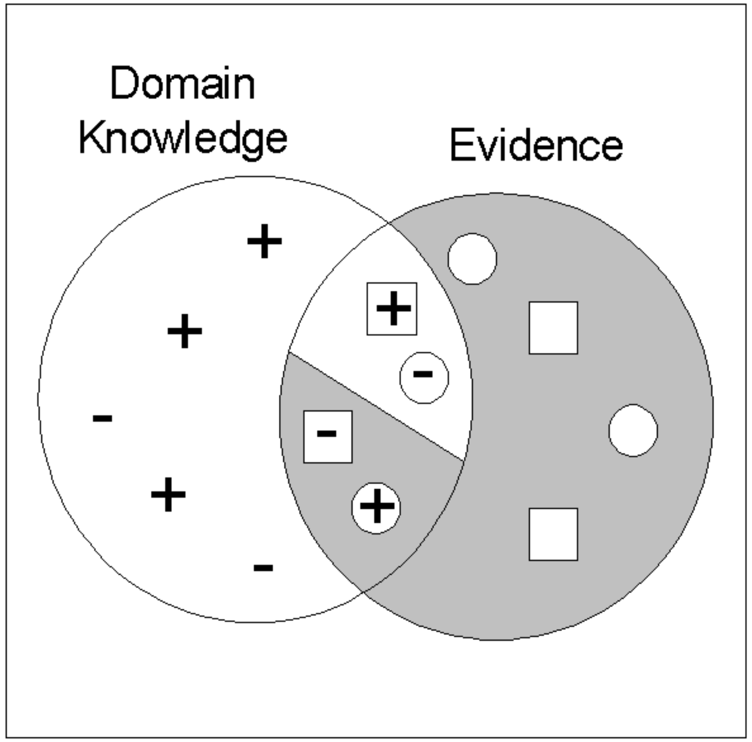
- Rank patterns based on statistics computed from data
- e.g., 21 measures of association (support, confidence, Laplace, Gini, mutual information, Jaccard, etc).

Subjective measure:

- Rank patterns according to user's interpretation
  - A pattern is subjectively interesting if it contradicts the expectation of a user (Silberschatz & Tuzhilin)
  - A pattern is subjectively interesting if it is actionable (Silberschatz & Tuzhilin)

Interestingness via unexpectedness

Need to model expectation of users (domain knowledge)



\+ Pattern expected to be frequent

\- Pattern expected to be infrequent

Pattern found to be frequent

Pattern found to be infrequent

\+ - Expected Patterns

\- + Unexpected Patterns

Need to combine expectation of users with evidence from data
(i.e., extracted patterns)

Slide courtesy of Tan,Steinbach, Kumar, Introduction to Data Mining

Additional References

R1 Data Mining: Concepts and Techniques by Han, Kamber and Pei (Morgan Kaufman)

Introduction to Data Mining by Tan, Steinbach and Kumar (Pearson – First Edition) Chapters 6 and 7

Recommender Systems – The Textbook by Charu C. Agarwal (Chapter 7)

THANK YOU

Gowri Srinivasa
Professor,
Department of Computer Science
gsrinivasa@pes.edu