RULE EVALUATION

SYRACUSE UNIVERSITY
School of Information Studies

LIMITATION OF CONFIDENCE MEASURE

100 transactions:
- 75 bought movies
- 60 bought games
- 40 bought both

Both seem to be strong rules.

{movies}->{games}
support 40/100 = 0.4
confidence 40/75 =0.53

{games}->{movies}
support 40/100 = 0.4
confidence 40/60 = 0.67

HOWEVER …

100 transactions:
- 75 bought movies
- 60 bought games
- 40 bought both

P(movies) = 75/100 = 0.75

P(games) = 60/100 = 0.6

P(movies and games) = 40/100 = 0.4

So people tend not to buy movies and games together!

Correlation(movies, games) = P(movies and games) /[P(movies) × P(games)] = 0.4/(0.75 × 0.6) = 0.89

Correlation <1 means negative correlation.

The confidence measure is sometimes misleading.

METRIC: LIFT (CORRELATION)

Measure of dependent or correlated events: Lift

Lift (A => B) = support({A,B})/(support(A) × support(B))

$$lift(A \quad > B) = \frac{P(A \quad B)}{P(A)P(B)}$$

Association rules should have >1 lift to be meaningful.

THE LIFT (CORRELATION) MEASURE

	Game	Not Game	Total
Movie	40	35	75
Not movie	20	5	25
Total	60	40	100

P(buy game) = 0.6
P(not buy movie) = 0.25
P(buy game and not buy movie) = 0.20
Lift (buy game -> not buy movie)
=0.20/(0.6 × 0.25) = 1.33 > 1

Strong rule

ALTERNATIVE MEASURES

Association rule algorithms tend to produce too many rules.

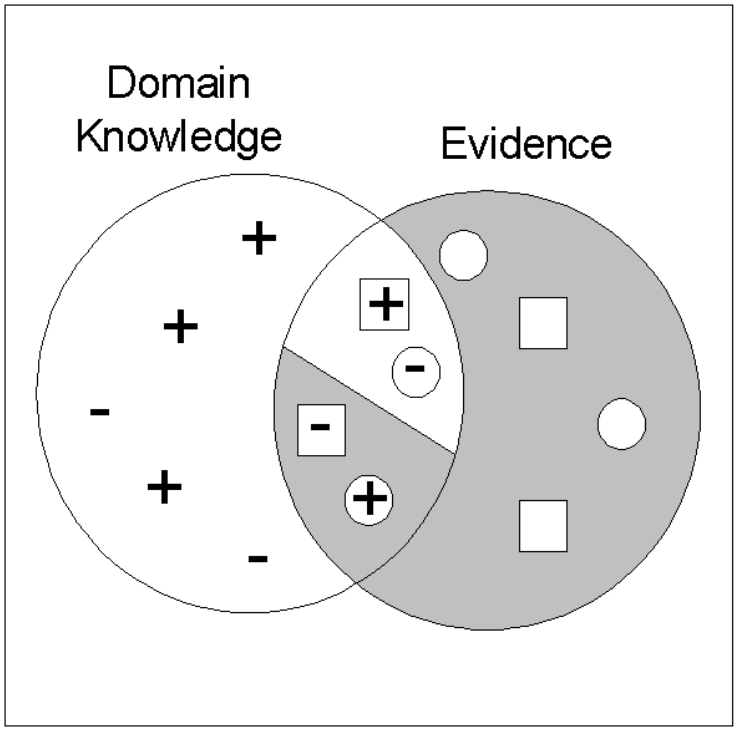
Many of them are uninteresting or redundant.

Uninteresting if it is known knowledge

Redundant if {A,B,C} → {D} and {A,B} → {D}
have same support and confidence

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)

WEKA ASSOCIATION RULES

Implemented a variation of Apriori Algorithm that iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence

Allows mining of “class association rules”: If the data have a class label attribute, the right hand side of a rule can be restricted to that label.

ASSOCIATION RULE MEASURES

In practice, what levels of support, confidence, and lift should we aim for?

Support:

- Depends on data set and business problem
- Common setting is 20–40% of the transactions

Confidence:

- Strong confidence rules ≥ .9, but .6 to .8 range might be OK

Lift:

- Should be above 1.0, the higher the better
- Levels of 2 and above can occasionally be seen but more likely to see around 1.3 to 1.5