Comp 135
Introduction to Machine Learning and Data Mining

Fall 2016

Professor: Roni Khardon

Computer Science
Tufts University

Association Rules

- Unsupervised learning but complementary to data exploration in clustering.
- The goal is to find "weak implications" in the data that have "non-negligible coverage"
- Useful in marketing, in understanding application data, as feature generator for supervised learning.

Association Rules

- Find all rules that have "Diet Coke" as consequent. These rules may help plan what the store should do to boost the sale of Diet Coke.
- Find all rules that have "bagels" in the antecedent. These rules may help determine what products may be impacted if the store discontinues selling bagels.
- Find all rules that have "sausage" in the antecedent and "mustard" in the consequent. This query can be phrased alternatively as a request for the additional items that have to be sold together with sausage in order to make it highly likely that mustard will also be sold.

Text from paper by [AIS93] that introduced the topic

Association Rules

- Find all the rules relating items located on shelves A and B in the store. These rules may help shelf planning by determining if the sale of items on shelf A is related to the sale of items on shelf B.
- Find the "best" k rules that have "bagels" in the consequent. Here, "best" can be formulated in terms of the confidence factors of the rules, or in terms of their support, i.e., the fraction of transactions satisfying the rule.

Text from paper by [AIS93] that introduced the topic

Data Model

- Following the market-basket application
- We assume a table where
  - Each row is a "transaction"
  - Each column is an "item"
- Table entries are in {0,1} i.e., discrete
- A transaction can be seen to represent the corresponding set of items

Association Rules

- What are useful rules?
- At least …% coverage: support

$$support(X) = \#\text{transactions including } X$$

$$frequency(X) = support(X)/\#\text{transactions}$$

- At least …% predictive: confidence

$$confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)}$$

Association Rules

- Applications and data characteristics from some early papers:

Data	#transact.	#items	transact. size	Avg size
Market Basket	50K	13K	1-100	10
Web Clicks	50K	500	1-267	2.5
Census	30K	2000	70	70

- In more demanding applications data does not fit in memory

Association Rules

- More data characteristics

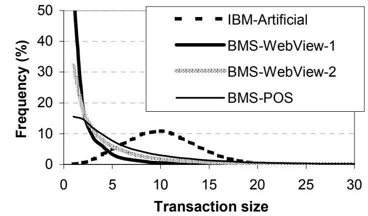


Figure from [ZKM01]

Association Rules

- Too many sets and rules …

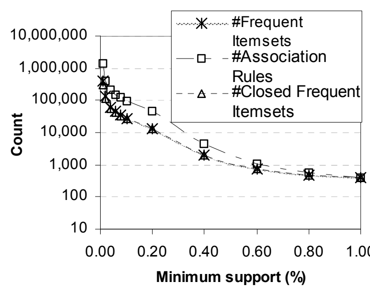


Figure from [ZKM01]

Association Rules

- Huge data: technology challenge making use of memory hierarchy
- Huge data: algorithmic challenge to process it efficiently
- Huge output: conceptual challenge to identify "most interesting" rules

Association Rules

- A concrete task 1:
- find all rules with frequency at least *f* and confidence at least *c*.
- How can we do this?
- If (X→Y) satisfies conditions then (X+Y) must also have frequency at least *f*.
- A concrete task 2:
- find all sets Z with frequency at least *f*.

Association Rules

- From frequent sets to rules
- Given frequent set Z
- for example {A,B,C,D,E}
- Remove potential conclusion W
- for example {D}
- And check the confidence of (Z\W→W)
- of (ABCE→D)

Frequent Set Mining

- find all sets Z with frequency at least *f*
- How can we do this?
- Main insight: anti-monotonicity
  if set Z is frequent then all its subsets are also frequent
- Algorithmic ideas?

Lattice Structure of Freq Sets

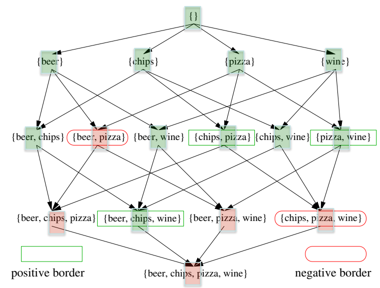


Figure 1: The lattice for the itemsets of Example 1 and its border.

Figure from [Goethals 2003]

Lattice Structure of Freq Sets

- Notice the notions of
  - positive border
  - negative border
- that are implicit in the monotonicity property and in the view via the lattice
- The borders capture all the frequent sets. Some algorithms attempt to find these directly.

Level-wise (Apriori) Algorithm

- level=1
- cands[1] = Sets with single items
- While cands[level] not empty
  1. Calc support for cands[level]
  2. freq[level] = cands[level] with high support
  3. cands[level+1]=generate from freq[level]
  4. level=level+1

Calculating support for candidates

- Basic implementation of step 1
- For each row R
- For each candidate X
- if X subset of R then: count[X]+=1
- One pass over database
- Improve run time via trie data structure that captures set of candidates

Generating candidates

- Monotonicity can be used to generate and prune potential candidates
- Use trie or lexicographical ordering to identify potential candidates
- Prune via subset relation
- Prune via upper/lower bounds on frequency (we skip details of this idea)

Vertical View

- View each item as a set of transactions
- This directly captures support
- Support of a set is the intersection of support of parents
- This incurs large space cost for storing support
- DFS recursive exploration avoids this and leads to efficient algorithm (we skip the details of this)

Which Rules are Interesting?

- Confidence can often be misleading

$$confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)}$$

- If p(B) is large
- p(B|A)=p(B) i.e., independent
- Confidence(A→B) is still large

Which Rules are Interesting?

- Lift measures dist from independence

$$Lift(X \Rightarrow Y) = \frac{freq(X \cup Y)}{freq(X)freq(Y)}$$

- Conviction aims at "implication"

$$Conviction(X \Rightarrow Y) = \frac{freq(X)(1 - freq(Y))}{freq(X) - freq(X \cup Y)}$$

- Interpret as inverse of
  Lift(X and Not Y)

Which Rules are Interesting?

conviction	implication rule
∞	five year olds don't work
∞	unemployed people don't earn income from work
∞	men don't give birth
50	people who are not in the military and are not looking for work and had work this year (1990, the year of the census) currently have civilian employment
10	people who are not in the military and who worked last week are not limited in their work by a disability
2.94	heads of household do not have personal care limitations
1.5	people not in school and without personal care limitations have worked this year
1.4	African-American women are not in the military
1.28	African-Americans reside in the same state they were born
1.28	unmarried people have moved in the past five years

Table 3: Sample Implication Rules From Census Data

Table from [BMUT97]

Positive and Negative Borders

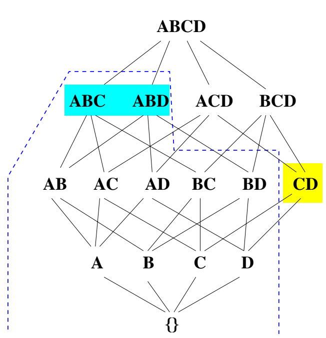


Figure from [GKMSTS 2003]

AMSS Alg: Find one maximal set

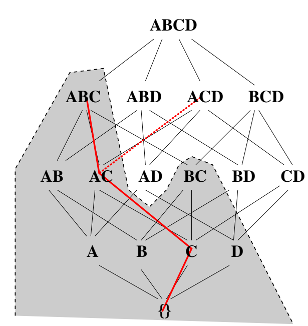
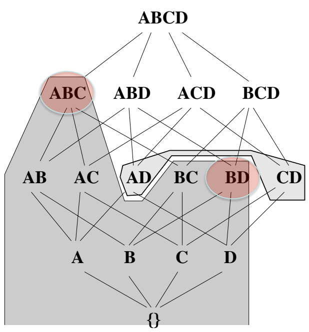


Figure from [GKMSTS 2003]

Dualize and Advance Algorithm



ABC and BD are max set found so far

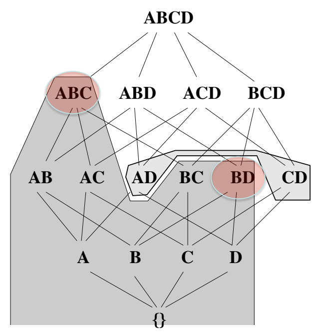
AD and CD are the negative border of this set

Any undiscovered frequent set must lie above the negative border.

Figure from [GKMSTS 2003]

comp135 Roni Khardon, Tufts University

Dualize and Advance Algorithm



ABC and BD are max set found so far

D and AC are their complements

Computing Transversals (apply distributive law) for (D and AC) yields (AD and CD) Which is the negative border

Figure from [GKMSTS 2003]

comp135 Roni Khardon, Tufts University

Frequent Sets as Features

- One way to generate enriched features for supervised learning is to generate frequent sets (because they occur and thus have a chance of making a difference)
- Very successful in frequent sub-graph mining, which extends the topic of this lecture to graphs, and its application to classifying molecules

comp135 Roni Khardon, Tufts University

Summary

- Association rules: a novel form of data exploration with different goals from previous supervised and unsupervised learning
- Algorithmic/computational challenge
- Frequent set mining as an important subtask
- Property: Anti-monotonicity
- Level-wise algorithm
- Many alternative algorithms

comp135 Roni Khardon, Tufts University