

COMP37332

Mining Association Rules

Goran Nenadic
School of Computer Science

1

Aims

- Understand the role and basic principles of association rules
- Discuss techniques for mining association rules
- Analyse possible application areas

2

Plan

- Introduction to association rules
- Support and confidence measures
- Generating association rules
- Multidimensional and negative association rules
- Problems and challenges
- Applications

3

Association rule mining

- The process of extracting links among data from large databases/data warehouses
- Also known as “link analysis”
 - discover links between individual data points rather than characterising whole data
- Common example: market-basket analysis
 - which products are bought together in one transaction?
- Examples also in biomedicine, finance, ...

4

Association rules

- Association rules describe relationships between sets of items in data-sets in the form:

$$A \rightarrow B$$

read as “*A implies B*”, where A and B are sets of **items** represented in a data set

- items: e.g. products bought by a customer
- Association rule mining is the process of finding (interesting) association rules in a given DB/DW

5

continued

Association rules

- Example:
 - $\{milk, bread\} \rightarrow \{juice\}$
 - “if a customer buys *milk* and *bread*, he/she is also likely to buy *juice*”
- itemset = a set of items involved in a single transaction
 - a k-itemset contains k items
- $A \rightarrow B$
 - A = left-hand side (LHS)
 - B = right-hand side (RHS)
 - $A \cup B$ = itemset

6

Association rules: measures

- How valid/interesting are rules?
- Measures for itemset distribution
 - support and confidence
- **Support**
 - relates to the frequency with which an itemset occurs in a DB
 - calculated as the percentage of transactions that contain the itemset (relative frequency)
 - if an itemset has *support* higher than some specified threshold (*minsup*), we say that it is *supported* or *frequent* or *large*

7

continued

Association rules: measures

- **Confidence**
 - measures "the implication" shown by the rule
- $$\text{confidence}(A \rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A)}$$
- "conditional probability" that the items in B will be purchased given items in A are purchased
 - we are interested in rules that have confidence above certain threshold (*minconf*)

8

Example

Transaction-id	Items bought
1	A, B, C
2	A, C
3	A, D
4	B, E, F

Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%
{A, B, C}	25%
{A, F}	0%

$\text{confidence}(A \rightarrow C) = \frac{\text{support}(\{A, C\})}{\text{support}(A)} = \frac{50}{75} = 66.67\%$
 $\text{confidence}(A \rightarrow \{B, C\}) = \frac{\text{support}(\{A, B, C\})}{\text{support}(A)} = \frac{25}{75} = 33.33\%$
 $\text{confidence}(A \rightarrow \{F\}) = \frac{\text{support}(\{A, F\})}{\text{support}(A)} = \frac{0}{75} = 0$
 $\text{confidence}(\{A, C\} \rightarrow \{B\}) = \frac{\text{support}(\{A, B, C\})}{\text{support}(\{A, C\})} = \frac{25}{50} = 50\%$
 $\text{confidence}(\{C\} \rightarrow \{B\}) = ?$

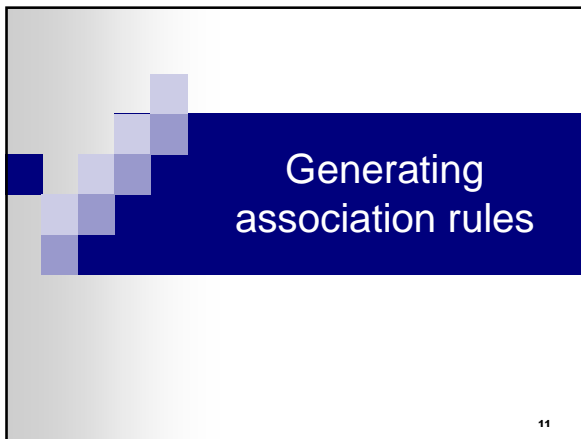
9

continued

Association rules: measures

- High confidence = strong pattern
 - these are of interest
 - gives "association" i.e. link between items
 - typically high (e.g. ~ 80%)
- High support = occurs often
 - less likely to be random occurrence
 - these are of interest as well – larger potential benefit from acting on the rule
 - typically 1-5% (still, huge in databases with millions of transactions)

10



Generating association rules

11

Generating association rules

1. Generate all frequent (supported) itemsets (above a certain threshold *minsup*)
2. For each supported itemset A and its subset B, let C = A – B
 - if $\text{support}(B+C)/\text{support}(C) > \text{minconf}$ then extract rule C → B (note: B+C = A)

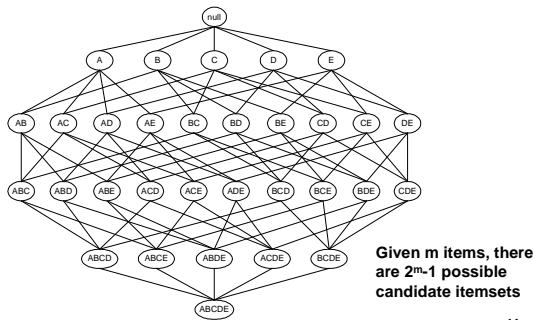
- *minconf* is the minimum confidence level required
- *minsup* (frequency threshold) is typically low

12

Generating association rules

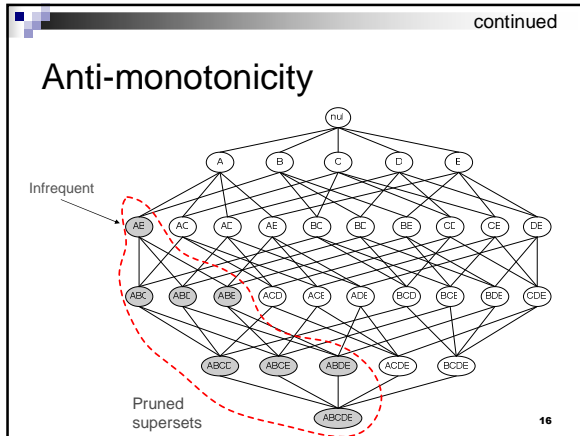
- Generating all supported i.e. frequent itemsets is the major problem
 - hundreds of thousands of different items:
 - if you have m items, then $2^m - 1$ possibilities
 - typical m is in the range of thousands
 - check against millions of transactions
 - Computationally intensive
 - is it possible to reduce the “search” space by reducing the number of potentially frequent itemsets?

Itemset lattice



Anti-monotonicity

- **Downward closure:** any subset of a frequent itemset must be also frequent
 - if $\{A, B, C\}$ is frequent, so is $\{A, C\}$
- Thus, if there is an infrequent itemset, its supersets should not be considered (**anti-monotonicity**)
 - if $\{A\}$ is not frequent, everything that contains $\{A\}$ cannot be frequent, and should not be considered

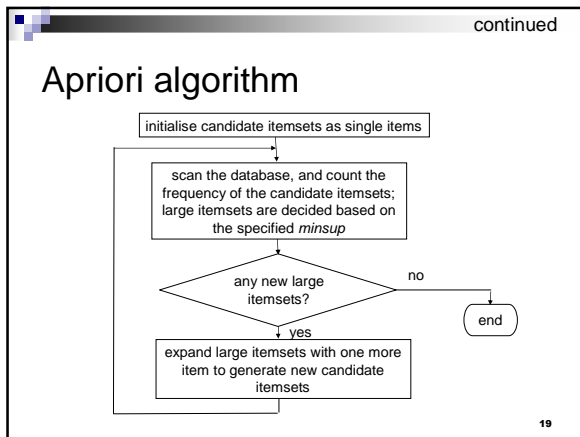


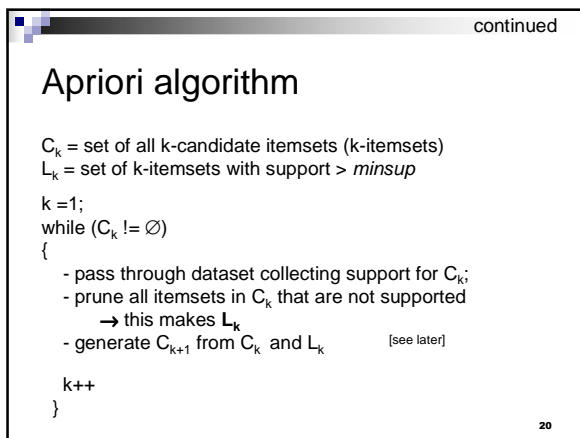
- continued
- ## Generating association rules
- Two steps:
 1. generate all frequent itemsets
 2. generate rules from frequent itemsets
 - Various approaches for step 1
 - e.g. Apriori algorithm and variants
- 17

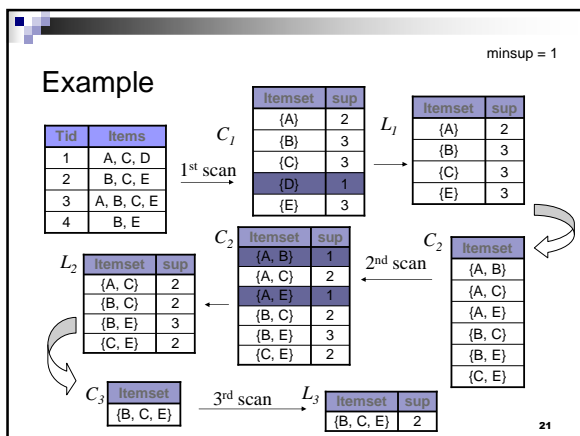
- ## Apriori algorithm
- Input: database of m transactions, $minsup$
 - Output: frequent itemsets L_1, \dots, L_p
 - L_k = frequent k -itemsets
 - Incremental approach:

starting from singleton itemsets ($k = 1$)

 - generate length $k+1$ candidate itemsets from length k frequent itemsets
 - test their support against the DB
- 18







Generating C_{k+1}

- C_{k+1} contains k -itemsets from C_k extended by one item
- C_{k+1} = combine members from C_k so that each subset of size k is in L_k (= frequent itemset of length k)
- Calculated by "self-joining" of L_k
- an example of candidate generation
 - $L_3 = \{abc, abd, acd, ace, bcd\}$ (frequent 3-itemsets)
 - self-joining: $L_3 * L_3$
 - $abcd$ from abc and abd
 - $acde$ from acd and ace
 - $abce$ from abc and ace
 - pruning: $acde$ is removed because ade is not in L_3 , similarly $abce$.
so $C_4 = \{abcd\}$

22

Generating candidates: problems

- When database is scanned to check C_k for creating L_k , a large number of transactions will be scanned even if they do not contain any k -itemset
- One transaction may contain many candidates
- It is costly to handle a huge number of candidate sets.
 - the candidate generation is the inherent cost of the Apriori algorithms, no matter what implementation technique is applied
- To mine a large data sets for long patterns – this algorithm is NOT efficient

23

Step 2: generating rules

if $\text{support}(B+C)/\text{support}(C) > \text{minconf}$ then
extract rule $C \rightarrow B$

Example

$\{B, C, E\}$ is a frequent itemset
(support = $2/4 = 0.5$)

Tid	Items
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E

rule	confidence
$B \rightarrow \{C, E\}$	$\text{support}(\{B, C, E\})/\text{support}(\{B\}) = 2/3$
$C \rightarrow \{B, E\}$	$\text{support}(\{B, C, E\})/\text{support}(\{C\}) = 2/3$
$E \rightarrow \{B, C\}$	$\text{support}(\{B, C, E\})/\text{support}(\{E\}) = 2/3$
$\{B, C\} \rightarrow \{E\}$	$\text{support}(\{B, C, E\})/\text{support}(\{B, C\}) = 2/2 = 100\%$
$\{B, E\} \rightarrow \{C\}$	$\text{support}(\{B, C, E\})/\text{support}(\{B, E\}) = 2/3$
$\{C, E\} \rightarrow \{B\}$	$\text{support}(\{B, C, E\})/\text{support}(\{C, E\}) = 2/2 = 100\%$

24

Example for you

Find all association rules with confidence levels above >0.80

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

25

Criticism of Apriori

■ Search heuristic

- if an itemset is infrequent, then all sets of items containing it must also be infrequent
 - Apriori therefore eliminates infrequent itemsets at an early stage, leaving only rules with high support
- looks for all itemsets (hence, all rules) that exceed the minimum support and confidence thresholds
- assumption is that rules with **high support** are the more 'interesting' ones

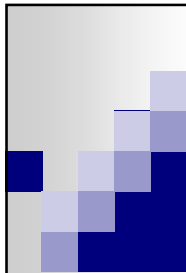
26

continued

Criticism of Apriori

- However, support and confidence are not necessarily indicative of the interestingness of the rule
 - there are other measures
 - E.g. **lift** and **chi-squared statistics** measure how much the actual proportion (of item A bought with item B) differs from the *expected proportion*, and are thus a better indicator of "surprising associations" or "dis-associations" between items!
 - The problem with lift and chi-squared statistics on their own is that there are millions of possible item combinations that can be associated and we have no heuristics for guiding us as to which combinations of items we should test for associations.

27



Application examples

28

“Main” example

- Market basket analysis
 - chain stores keep TBs of customer purchase information
- Benefits/applications
 - **catalogue design, inventory layout, and cross-marketing:** advertise or place complementary products together if people aren't buying them together
 - **loss-leader analysis:** determine which low-margin products would lead people to buy high-margin products

29

continued

“Main” example

- **product pricing and promotion**
 - don't offer simultaneous discounts on two items that are usually purchased together as duplicate discounts don't add any incentive
- **product assortment optimisation**
 - process of adding or removing items to or from your assortment in order to maximize profit and/or customer satisfaction
 - e.g. you may decide against adding Pepsi to your assortment (drinks category) if your profit margin is lower than Coke and people tend to buy Pepsi instead of Coke.
- **customer segmentation based on buying patterns**

30

Fraud detection example

- E.g. fake insurance claims
 - Health Insurance Commission of Australia used association discovery to detect fraud and inappropriate practice in pathology test requests and health insurance claims

31

Document management examples

- Scenario 1: associated words/word networks
 - set of documents (could be the Web or local DW)
 - "baskets" = documents
 - "items" = words in documents
 - frequent word-groups = linked words/concepts
- Scenario 2: associated pages
 - search (or IR) engines: Google, Yahoo
 - "baskets" = web pages
 - "items" = outgoing and incoming links
 - pages with similar references = about same topic

32

Biological example

- Mining rules to help automatic description ("annotation") of gene products with their
 - molecular functions [mf]
 - biological processes [bp]
 - cellular components [cc]
- "Annotations" contain terms from a controlled vocabulary (e.g. the Gene ontology)
 - gene X: protein biosynthesis [mf]
ribosome biogenesis [bp]
ribosome [cc]

33

continued

Biological example

- Databases of gene annotations
 - e.g. GOA database; done manually
- Mining association rules
 - discover possible links among annotations
 - find reliable rules that can be applied automatically
- Example
 - {protein biosynthesis [mf], ribosome biogenesis [bp]}
 - ribosome [cc]
 - support: 0.2% (88 annotations); confidence: 93.2%

34

continued

Biological example

membrane [cc]
 ← oligopeptide transport [bp];
 transporter activity [mf]
 (0.106%/44, 100.0%)

binding [mf]
 ← mitochondrial transport [bp];
 mitochondrial inner membrane [cc]
 (0.106%/44, 100.0%)

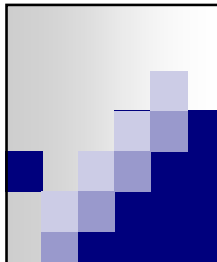
protein biosynthesis [bp]
 ← ribosome [cc];
 structural constituent of ribosome [mf]
 (0.504%/209, 90.4%)

35

continued

Biological example

36



Problems and challenges

37

Some problems: values

- If an attribute has potentially infinite (or large enough) set of values, it is very unlikely we would come up with a rule
 - association rule mining will not work with numerical (i.e. quantitative) attributes
 - it needs **categorical** attributes
 - e.g. mapping numerical values to categorical values
 - income from 0-10k → *low*
 - income above 40k → *high*

38

continued

Some problems: values

- Even with categorical attributes, there could be numerous values
 - aggregation and/or generalisation could help
 - e.g. mapping values using product hierarchies, taxonomies and ontologies

```

graph TD
    beverage --> noncarbonated
    beverage --> carbonated
    noncarbonated --> juice
    noncarbonated --> other1[ ]
    juice --> apple
    juice --> peach
  
```

```

graph TD
    dairy_product[dairy product] --> yogurt
    dairy_product --> milk
    yogurt --> healthy
    yogurt --> full_fat[full-fat]
  
```

39

continued

Some problems: values

- Use both specialisation and generalisation to mine more useful rules
 - generally
 - generalisation can improve support
 - specialisation can improve confidence

40

continued

Some problems: numbers

- Problems with huge numbers of itemsets
- Solutions
 - use filters applied to data warehouse
 - e.g. mine association rules for certain regions, dates
 - do some approximation
 - find all frequent k-sets from a sample
 - must loose something – accuracy or coverage

41

Sampling algorithm

- Step 1
 - load a random sample of baskets
 - run Apriori to get frequent itemsets from the sample
 - scale-down support threshold (e.g., if 1% sample, use $minsup/100$ as support threshold)
- Step 2
 - get exact support values for candidates to validate
- Errors
 - false negatives (X can be frequent, but not in sample)
 - no false positives (pass 2)

42

Sampling algorithm

- Improvement for false negatives
 - define *negative border* (NB):
 - X belongs to the negative border if X is not frequent (in the sample), but all his subsets are frequent
 - negative boarder contains the 'closest' items that could be frequent (in the whole database)
 - need to calculate support for each NB set as well
 - however, if X from NB is found to be frequent, then there is a potential that a superset of X is also frequent – needs a new run over the DB

43

Partitioning

- Divide the DB into non-overlapping subsets
 - find 'local' frequent itemsets, calculating the support only in each partition
 - local minimal support is different from the original one
 - may contain **false positives** (false global frequent itemsets)
 - but **no false negatives** (all global frequent itemsets)
 - these itemsets are now candidates – their support checked in pass 2 (as in sampling)
- Good for parallel and distributed computing

44

Multidimensional associations

- Previous examples focus on one dimension
 - e.g. product type
- Multiple dimensions can be included
 - e.g. Time(7:00-8:00) → {bread, milk}
 - Time(21:00-22:00) → {cheese, wine}
- Again, for quantitative attributes (e.g. time, income) partitioning into non-overlapping **intervals** (e.g. morning, afternoon) can be performed

45

Negative associations

- “customers who buy X do not buy Y”
- These are more difficult to mine
 - there are millions of items with low support, e.g.
 - 10,000 items → ~2.5 billion pairwise combinations
 - a DB with 100 million transactions misses most of these!!
- Find only **interesting** negative rules
- Use some background knowledge
 - X and Y to be somehow related

46

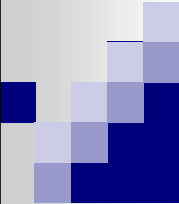
continued

Negative associations

- Hierarchies (ontologies) that represent knowledge about the domain
- $A \rightarrow B$ may be an interesting negative rule if its support is low and there is significantly higher support for $A_1 \rightarrow B_1$, where A_1 and B_1 are more general classes of A and B (respectively) in the hierarchy
- Use heuristics

47

48




Examples and applications

49

Application support

- Many database vendors include specialised data mining capabilities
- Also, specialised data mining environments
 - e.g. WEKA
 - <http://www.cs.waikato.ac.nz/ml/weka/>
 - decisions trees, association rules, clustering, etc.



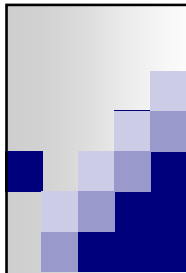
Software

50

Weka - example

age	spectacle-prescrip	astigmatism	tear-prod-rate	contact-lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetope	no	reduced	none
young	hypermetope	no	normal	soft
young	hypermetope	yes	reduced	none
young	hypermetope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetope	no	reduced	none
pre-presbyopic	hypermetope	no	normal	soft
pre-presbyopic	hypermetope	yes	reduced	none
pre-presbyopic	hypermetope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetope	no	reduced	none
presbyopic	hypermetope	no	normal	soft
presbyopic	hypermetope	yes	reduced	none
presbyopic	hypermetope	yes	normal	none

51



Wrapping up

61

Criticism of association rules

- Spurious associations
 - increases as the size of the data set increases
 - e.g. one project reports having found a high-confidence association between the purchase of mattresses and the purchase of chainsaws
- Reasons for associations are not explicit
 - e.g. one hardware chain found that toilet rings sell better at new stores - however, they may not use this 'knowledge' as they can't explain it

62

continued

Criticism of association rules

- Not all discovered rules are 'interesting'
 - e.g. an association between coffee and milk purchases is probably not interesting as it is unsurprising and is not *new* knowledge
 - Nor is the tendency of people to buy TV licences with TVs, because such purchases are *legislated*
 - Similarly, buying TV warranties leading to buying TV's is unsurprising since TV warranties are useless without TV's.
(Of course, the reverse - buying TV's leading to buying TV warranties may or may not be surprising).
 - Good/useful rules should therefore be unexpected and actionable

63

Criticism of association rules

- Traditional association rule algorithms do not take into account the value of items
 - sale of an expensive jar of caviar accounts for the same as the sale of a cheap loaf of bread.
 - use a "support metric" that allows us to ignore associations where relatively low margin items are involved; but capture rare transactions involving profitable high-value items

Summary

- Associations/links among itemsets based on frequent evidence in databases
- Describe some type of causality
 - Categorical attributes
- Use them to acquire patterns that can improve understanding of behaviour
- Finding rules can be computationally expensive
 - finding all frequent itemsets is the hard part
- Finding interesting rules

Reading and tutorials

- Chapter 27 (27.2) in [Elmasri & Navathe]
- Also: on-line materials, tutorials and software
<http://personalpages.manchester.ac.uk/staff/G.Nenadic/COMP37332/>
- Tutorial questions are already available
 - answer the question related to association rule mining
- WEKA tutorials
 - complete the WEKA association rule mining tutorial
 - go through the tutorial at
<http://maya.cs.depaul.edu/~classes/ect584/WEKA/associate.html>
