



BABD

Masters in Business Analytics and Big Data

Association Rules

The “diapers and beer” story

“A number of convenience store clerks noticed that men often bought beer at the same time they bought diapers.

The store mined its receipts and proved the clerks' observations correct.

So the store began stocking diapers next to the beer coolers, and sales skyrocketed”.



Association Rules

Find interesting **REGULAR PATTERNS** and **RECURRENCES** within a large set of transactions

Some application domains:

Market basket analysis

Identify associations between items in shopping baskets (i.e. which items are frequently purchased together).
Can be used to understand customer shopping, optimize shelf location, catalog-design, develop co-marketing initiatives.

Credit Card Risk Analysis

Highlight novel strange patterns or find the characteristics of customers who are likely to default on credit card (notice that product and services are virtually infinite).
Can be used to reduce risks when assessing new credit card applications.

Web Mining

Understand the patterns of navigation paths. Can be used to improve the navigation experience.

Association Rules

Some application domains:

Fraud Detection

Discover characteristics of individuals who perform potential fraudulent actions from a set of transactions consisting of incident reports and applications for compensation.

Can be used to prevent the payment of compensation based on unlawful conduct.

Stock Market Analysis

Find relationships between individual stocks, or between stocks and economic factors.

Can help stock traders to select interesting stocks and improve trading strategies.

Medical Diagnosis

Identify relationships between symptoms, test results and illness.

Can be used for assisting doctors on illness diagnosis or even on treatment.

Association Rules: Some definitions

Consider the following set of n objects (items):

$$\mathcal{O} = \{o_1, o_2, \dots, o_n\}$$

A **K-ITEMSET** is a generic subset $L \subseteq \mathcal{O}$ containing k objects

A **TRANSACTION** can be represented as an itemset that has been recorded in a database in a single activity

The data set is composed by a list of m transactions T_i , each being associated to a unique identifier t_i

Data set of transactions

List of transactions

| identifier t_i | transaction T_i |
|------------------|-------------------|
| 001 | $\{a, c\}$ |
| 002 | $\{a, b, d\}$ |
| 003 | $\{b, d\}$ |
| 004 | $\{b, d\}$ |
| 005 | $\{a, b, c\}$ |
| 006 | $\{b, c\}$ |
| 007 | $\{a, c\}$ |
| 008 | $\{a, b, e\}$ |
| 009 | $\{a, b, c, e\}$ |
| 010 | $\{a, e\}$ |

$$\mathcal{O} = \{a, b, c, d, e\} = \{\text{bread, milk, cereals, coffee, tea}\}$$

Data set of transactions

The list of m transactions is expressed by means of a bi-dimensional matrix:

- Columns correspond to the objects in the set \mathcal{O}
- Rows correspond to the different transactions
- The generic elements of the matrix is given by $x_{ij} = \begin{cases} 1 & \text{if } o_j \text{ belongs to } T_i \\ 0 & \text{otherwise} \end{cases}$

Matrix of transactions

| identifier t_i | a | b | c | d | e |
|------------------|-----|-----|-----|-----|-----|
| 001 | 1 | 0 | 1 | 0 | 0 |
| 002 | 1 | 1 | 0 | 1 | 0 |
| 003 | 0 | 1 | 0 | 1 | 0 |
| 004 | 0 | 1 | 0 | 1 | 0 |
| 005 | 1 | 1 | 1 | 0 | 0 |
| 006 | 0 | 1 | 1 | 0 | 0 |
| 007 | 1 | 0 | 1 | 0 | 0 |
| 008 | 1 | 1 | 0 | 0 | 1 |
| 009 | 1 | 1 | 1 | 0 | 1 |
| 010 | 1 | 0 | 0 | 0 | 1 |

$$\mathcal{O} = \{a, b, c, d, e\} = \{\text{bread, milk, cereals, coffee, tea}\}$$

Association Rules

The **EMPIRICAL FREQUENCY** of the itemset L is defined as the number of transactions in the data set containing L

When dealing with a large sample, the ratio $f(L)/m$ (relative frequency) approximates the probability of occurrence of the itemset L

$\mathcal{O} = \{a, b, c, d, e\} = \{\text{bread, milk, cereals, coffee, tea}\}$

| identifier t_i | a | b | c | d | e |
|------------------|-----|-----|-----|-----|-----|
| 001 | 1 | 0 | 1 | 0 | 0 |
| 002 | 1 | 1 | 0 | 1 | 0 |
| 003 | 0 | 1 | 0 | 1 | 0 |
| 004 | 0 | 1 | 0 | 1 | 0 |
| 005 | 1 | 1 | 1 | 0 | 0 |
| 006 | 0 | 1 | 1 | 0 | 0 |
| 007 | 1 | 0 | 1 | 0 | 0 |
| 008 | 1 | 1 | 0 | 0 | 1 |
| 009 | 1 | 1 | 1 | 0 | 1 |
| 010 | 1 | 0 | 0 | 0 | 1 |

$$L = \{a, c\} \Rightarrow f(L) = 4 \Rightarrow \Pr(L) \approx 4/10 = 0.4$$

What is a rule?

Let Y and Z be two propositions which may be true or false.

A **RULE** is an implication in the form $Y \Rightarrow Z$: “ if Y is true, then Z is also true ”

A rule is called **PROBABILISTIC** if the validity of Z is associated with a certain probability p :
“ if Y is true, then Z is also true with probability p ”

Consider two disjoint itemsets, $L \subset \mathcal{O}$ and $H \subset \mathcal{O}$, and the transaction T

An **ASSOCIATION RULE** is a probabilistic implication denoted as $L \Rightarrow H$ with the following meaning:
“ if L is contained in T , then H is also contained in T with probability p ”

↓
body

↓
head

Confidence and Support

➤ CONFIDENCE $p = \text{conf}\{L \Rightarrow H\}$

- proportion of transactions containing H among those including L (reliability of the rule)
- it approximates the conditional probability that H belongs to T given that L belongs to T

➤ SUPPORT $s = \text{supp}\{L \Rightarrow H\}$

- proportion of transactions containing both H and L (frequency of the rule)
- it approximates the probability that L and H are both contained in a future transaction

$\mathcal{O} = \{a, b, c, d, e\} = \{\text{bread, milk, cereals, coffee, tea}\}$

| identifier t_i | a | b | c | d | e |
|------------------|-----|-----|-----|-----|-----|
| 001 | 1 | 0 | 1 | 0 | 0 |
| 002 | 1 | 1 | 0 | 1 | 0 |
| 003 | 0 | 1 | 0 | 1 | 0 |
| 004 | 0 | 1 | 0 | 1 | 0 |
| 005 | 1 | 1 | 1 | 0 | 0 |
| 006 | 0 | 1 | 1 | 0 | 0 |
| 007 | 1 | 0 | 1 | 0 | 0 |
| 008 | 1 | 1 | 0 | 0 | 1 |
| 009 | 1 | 1 | 1 | 0 | 1 |
| 010 | 1 | 0 | 0 | 0 | 1 |

$L = \{a, c\}$
 $H = \{b\}$

• (rows 005, 009)

$$p = \text{conf}\{L \Rightarrow H\} =$$

$$s = \text{supp}\{L \Rightarrow H\} =$$

Confidence and Support

What about rules with very high confidence and very low support?

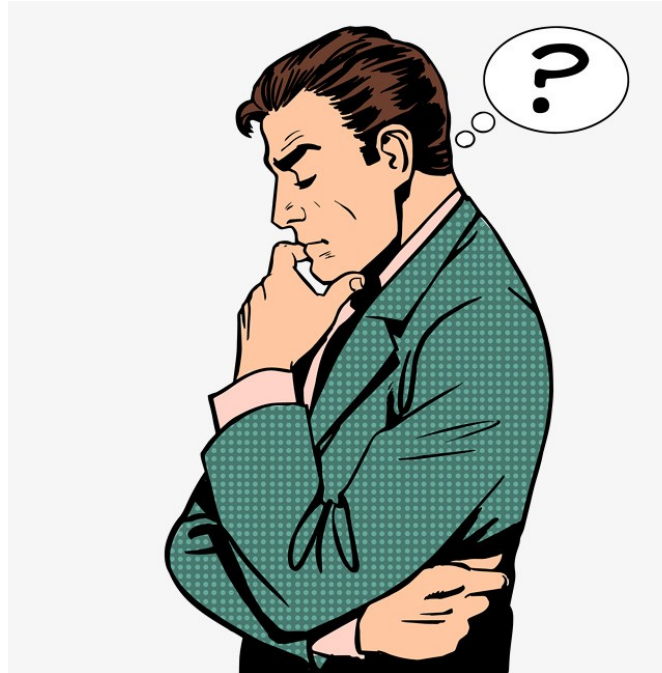
They may occur simply by chance.

They are likely to be uninteresting from a business perspective (items that customers seldom buy together).

But...they might be incredibly relevant!*

What about items with low empirical frequency?

They may correspond to expensive products (jewelry?) that are seldom purchased but whose patterns are still interesting.



What about rules with very high support and very high confidence?

Interesting but...

...they may correspond to something "already known".

The inference made by a rule does not necessarily imply causality...

...which typically involves relationships occurring over time.

*{Down's syndrome} \rightarrow { trisomy 21} (a chromosomal defect). Patients with trisomy 21 suffer from Down's syndrome with almost 100% confidence. But Down's syndrome is found in about one in thousand cases.

Strong Association Rules

Objective \Rightarrow Find the **STRONG** association rules:

$$s \geq s_{min}$$

$$p \geq p_{min}$$

what about extracting all the rules?...

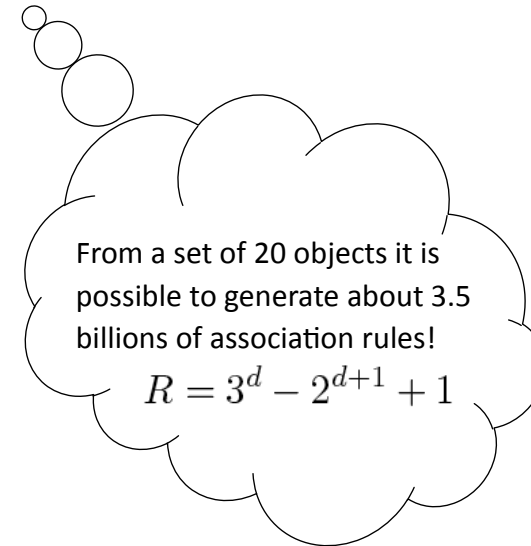
The number of possible rules increases exponentially as the number of objects increases ...



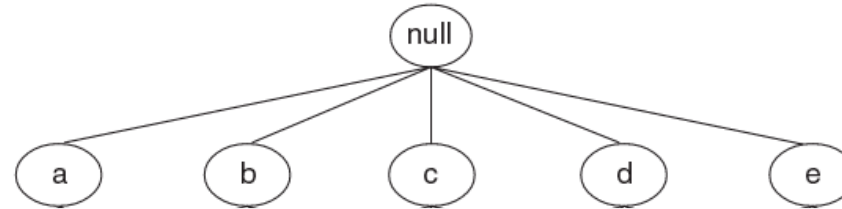
... a two-phase procedure is applied:

🌐 Generation of **FREQUENT ITEMSETS**

🌐 Generation of *strong* rules



Itemset Lattice



Association Rules

FIRST REMARK

If the itemset $L \cup H$ is not frequent, we can exclude from the analysis all the rules obtained by using all the proper subsets of $L \cup H$...

$$\begin{array}{ll} \{a, b\} \Rightarrow \{c\} & \{a, c\} \Rightarrow \{b\} \\ \{b, c\} \Rightarrow \{a\} & \{a\} \Rightarrow \{b, c\} \\ \{b\} \Rightarrow \{a, c\} & \{c\} \Rightarrow \{a, b\} \end{array}$$

... the real problem, however, is how to generate all the frequent itemsets !

Association Rules

A data set of m transactions defined over a set of n objects may contain up to $2^n - 1$ frequent itemsets (excluding the empty set)
...exhaustive enumeration is impracticable...

The **APRIORI ALGORITHM** is an effective method for extracting strong rules.

It relies on the following property (Apriori principle):

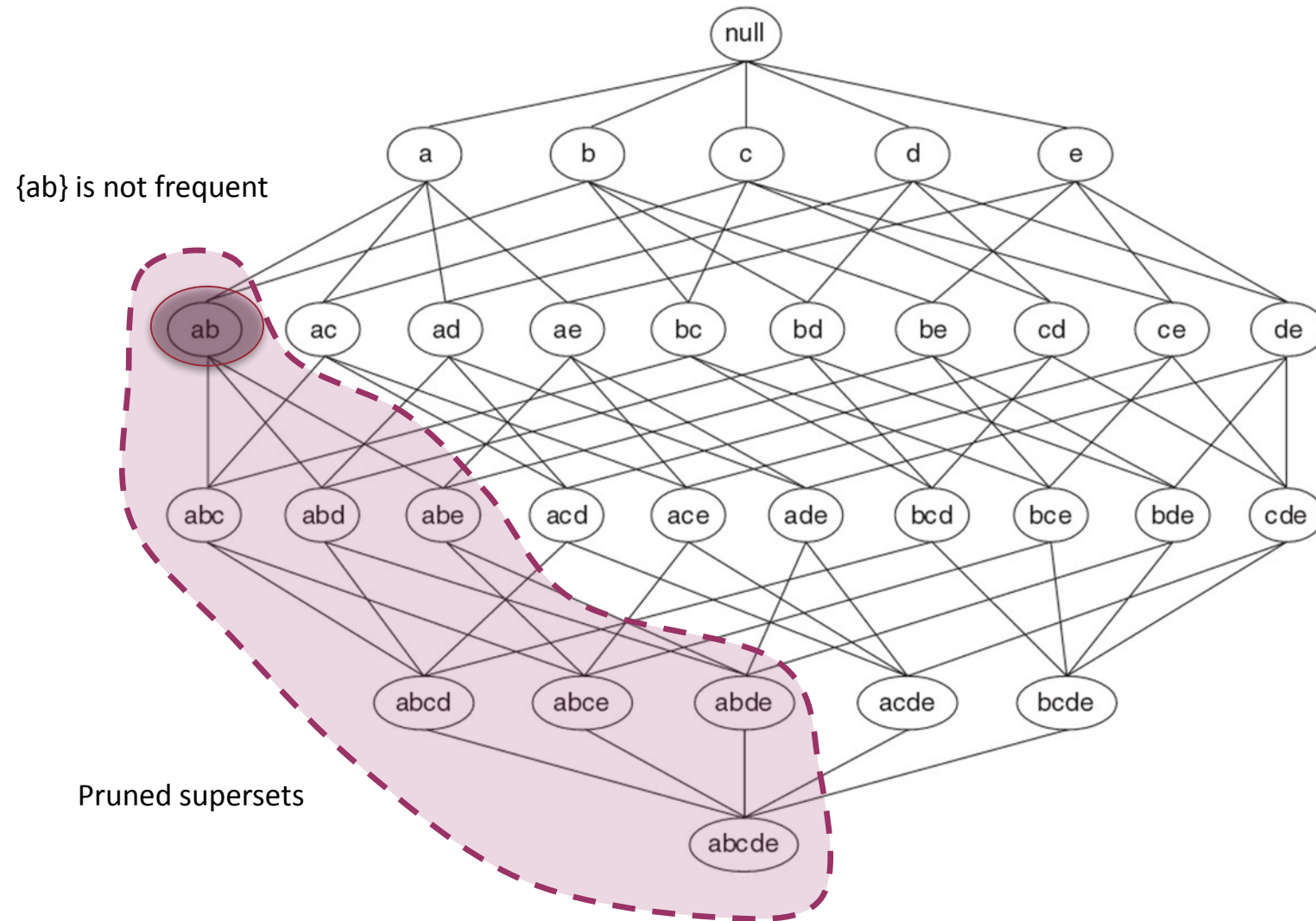
“if an itemset is frequent, then all its subsets are also frequent”



if an itemset is not frequent, then each of the itemset containing
it are not frequent as well!

Once a non-frequent itemset is identified in the course of the algorithm, all its supersets are implicitly eliminated and excluded from the analysis

Itemset Lattice



Apriori Algorithm: Phase 1

The Apriori algorithm generates the frequent itemsets iteratively, starting from the frequent 1 – itemset to the frequent k – itemsets

Phase 1 - initialization

Compute the empirical frequency of each object.

Discard the objects with a frequency smaller than the given threshold (support = 0.2).

$$\mathcal{O} = \{a, b, c, d, e\} = \{\text{bread, milk, cereals, coffee, tea}\}$$


| itemset | relative frequency | status |
|---------|--------------------|--------|
| $\{a\}$ | | |
| $\{b\}$ | | |
| $\{c\}$ | | |
| $\{d\}$ | | |
| $\{e\}$ | | |

Apriori Algorithm: Phase 1

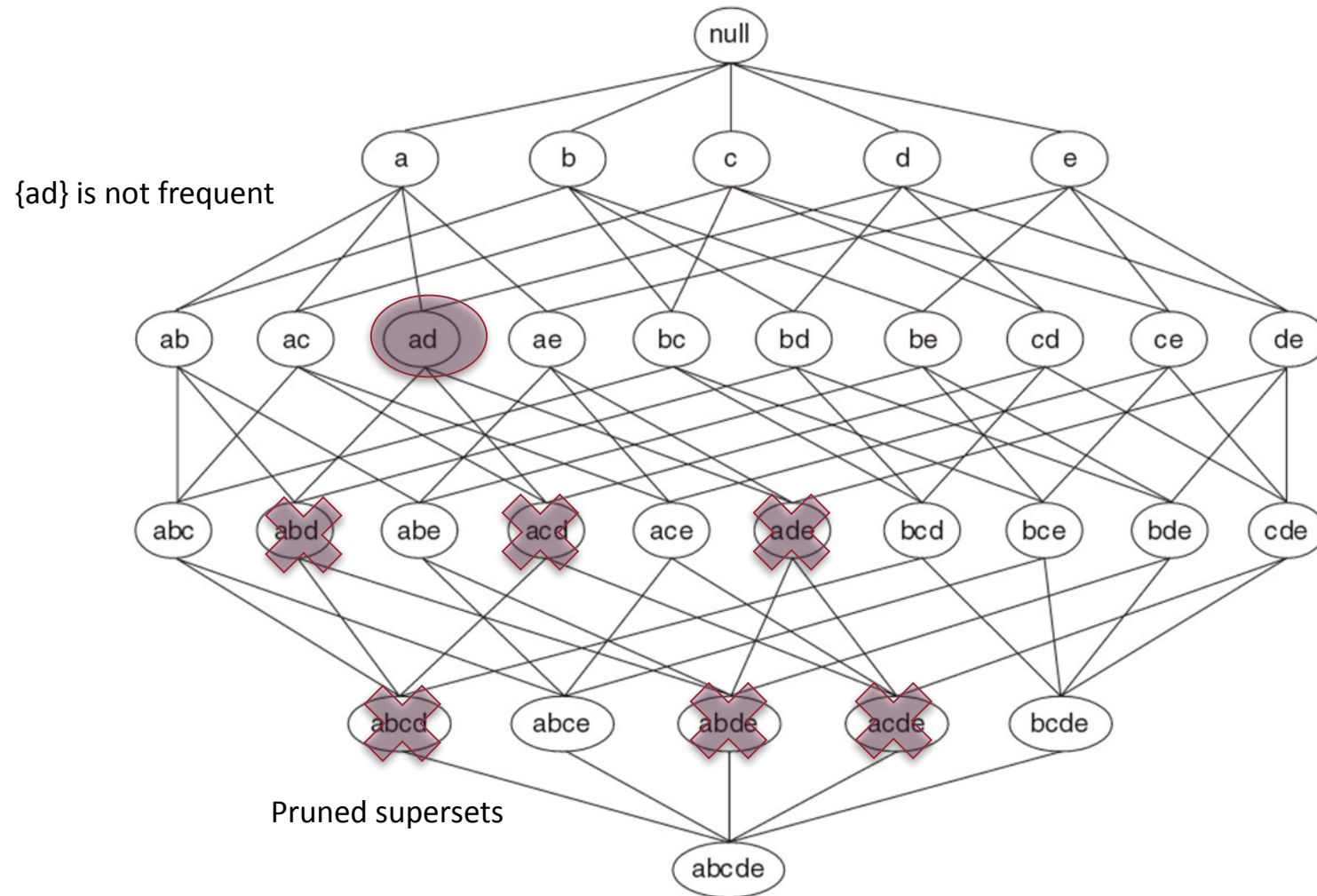
Phase 1 - iteration 1

The candidate 2 – itemsets are obtained from the 1 – itemset found during the former iteration.
Non-frequent itemsets are discarded.

$\mathcal{O} = \{a, b, c, d, e\} = \{\text{bread, milk, cereals, coffee, tea}\}$

| itemset | relative frequency | status |
|------------|--------------------|--|
| $\{a, b\}$ | $4/10 = 0.4$ | frequent |
| $\{a, c\}$ | $4/10 = 0.4$ | frequent |
| $\{a, d\}$ | $1/10 = 0.1$ | not frequent  |
| $\{a, e\}$ | $3/10 = 0.3$ | frequent |
| $\{b, c\}$ | $3/10 = 0.3$ | frequent |
| $\{b, d\}$ | $3/10 = 0.3$ | frequent |
| $\{b, e\}$ | $2/10 = 0.2$ | frequent |
| $\{c, d\}$ | $0/10 = 0.0$ | not frequent |
| $\{c, e\}$ | $1/10 = 0.1$ | not frequent |
| $\{d, e\}$ | $0/10 = 0.0$ | not frequent |

Itemset Lattice



Apriori Algorithm: Phase 1

Phase 1 - iteration 2

The candidate 3 – itemsets are obtained from the 2 – itemsets found during the former iteration.
Non-frequent itemsets are discarded.

| $\mathcal{O} = \{a, b, c, d, e\} = \{\text{bread, milk, cereals, coffee, tea}\}$ | | |
|--|--------------------|----------|
| itemset | relative frequency | status |
| $\{a, b, c\}$ | $2/10 = 0.2$ | frequent |
| $\{a, b, e\}$ | $2/10 = 0.2$ | frequent |

The first phase of the algorithm terminates since no candidate 4 – itemsets can be generated!

Apriori Algorithm: Phase 2

Phase 2 - step 1

Split each frequent itemset B into two disjoint non-empty subsets L and $H = B - L$, according to each possible combination.

Generate all the corresponding candidate rules.



Given a frequent k -itemsets it is possible to generate $2^k - 2$ candidate rules which may be strong rules (excluding those with empty body or head)

Phase 2 - step 2

For each candidate rule $L \Rightarrow H$ compute the confidence.



Very fast step: frequencies of $(L \cup H)$ and L are known at the end of the first phase

Phase 2 - step 3

Take just the rules with a confidence greater than the prefixed threshold (i.e. $\text{confidence} \geq 0.55$) \Rightarrow strong rules

Apriori Algorithm: Phase 2

Strong rules generation

| itemset | rule | confidence | status |
|---------------|--------------------------|------------------|------------|
| $\{a, b\}$ | $\{a \Rightarrow b\}$ | $p = 4/7 = 0.57$ | strong |
| $\{a, b\}$ | $\{b \Rightarrow a\}$ | $p = 4/7 = 0.57$ | strong |
| $\{a, c\}$ | $\{a \Rightarrow c\}$ | $p = 4/7 = 0.57$ | strong |
| $\{a, c\}$ | $\{c \Rightarrow a\}$ | $p = 4/5 = 0.80$ | strong |
| $\{a, e\}$ | $\{a \Rightarrow e\}$ | $p = 3/7 = 0.43$ | not strong |
| $\{a, e\}$ | $\{e \Rightarrow a\}$ | $p = 3/3 = 1.00$ | strong |
| $\{b, c\}$ | $\{b \Rightarrow c\}$ | $p = 3/7 = 0.43$ | not strong |
| $\{b, c\}$ | $\{c \Rightarrow b\}$ | $p = 3/5 = 0.60$ | strong |
| $\{b, d\}$ | $\{b \Rightarrow d\}$ | $p = 3/7 = 0.43$ | not strong |
| $\{b, d\}$ | $\{d \Rightarrow b\}$ | $p = 3/3 = 1.00$ | strong |
| $\{b, e\}$ | $\{b \Rightarrow e\}$ | $p = 2/7 = 0.29$ | not strong |
| $\{b, e\}$ | $\{e \Rightarrow b\}$ | $p = 2/3 = 0.67$ | strong |
| $\{a, b, c\}$ | $\{a, b \Rightarrow c\}$ | $p = 2/4 = 0.50$ | not strong |
| $\{a, b, c\}$ | $\{c \Rightarrow a, b\}$ | $p = 2/5 = 0.40$ | not strong |
| $\{a, b, c\}$ | $\{a, c \Rightarrow b\}$ | $p = 2/4 = 0.50$ | not strong |
| $\{a, b, c\}$ | $\{b \Rightarrow a, c\}$ | $p = 2/7 = 0.29$ | not strong |
| $\{a, b, c\}$ | $\{b, c \Rightarrow a\}$ | $p = 2/3 = 0.67$ | strong |
| $\{a, b, c\}$ | $\{a \Rightarrow b, c\}$ | $p = 2/7 = 0.29$ | not strong |
| $\{a, b, e\}$ | $\{a, b \Rightarrow e\}$ | $p = 2/4 = 0.50$ | not strong |
| $\{a, b, e\}$ | $\{e \Rightarrow a, b\}$ | $p = 2/3 = 0.67$ | strong |
| $\{a, b, e\}$ | $\{a, e \Rightarrow b\}$ | $p = 2/3 = 0.67$ | strong |
| $\{a, b, e\}$ | $\{b \Rightarrow a, e\}$ | $p = 2/7 = 0.29$ | not strong |
| $\{a, b, e\}$ | $\{b, e \Rightarrow a\}$ | $p = 2/2 = 1.00$ | strong |
| $\{a, b, e\}$ | $\{a \Rightarrow b, e\}$ | $p = 2/7 = 0.29$ | not strong |

Strong Association Rules

Strong rules are always meaningful?

A retailer wishes to analyze a set of transactions to identify associations between sales of milk and sales of biscuits:

- 1000 transactions
- 600 include 'milk'
- 750 include 'biscuits'
- 400 include both products

Probability of purchasing biscuits
conditioned to the purchase of milk



The rule { milk } \Rightarrow { biscuits } has support 0.4 and confidence 0.67

What is the probability of purchasing biscuits?...0.75, which is greater than 0.67!



Biscuits and milk show a negative correlation...
...the purchase of milk reduces the probability of buying biscuits!

Lift Index

A third measure of significance is the **LIFT** :

$$l = \text{lift}\{L \Rightarrow H\} = \frac{\text{conf}\{L \Rightarrow H\}}{\text{supp}(H)}$$

🔵 The lift is greater than 1: body and head are *positively associated*

The rule is effective in assuming the presence of the head in a given transaction

🔵 The lift is lower than 1: body and head are *negatively associated*

The rule is less effective than the estimate obtained by the frequency of the head

For the former example:

$$\begin{aligned}\text{conf}\{L \Rightarrow H\} &= 400/600 = 0.67 \\ \text{supp}(H) &= 750/1000 = 0.75\end{aligned}$$

$$\text{lift}\{L \Rightarrow H\} = 0.67/0.75 \approx 0.89$$

Association Rules

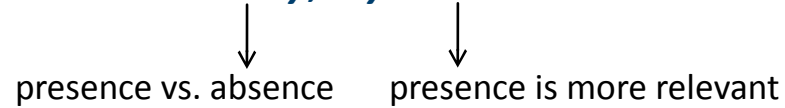
The computational effort required grows exponentially as the number n of objects increases

To improve the efficiency:

- Split the data set into disjoint subsets of transactions and apply the algorithm to each subset (local frequent itemsets)
- Use the algorithm on a given sample of transactions
- Resorting to hierarchies of objects
- Discard in the preprocessing phase less interesting objects

Association Rules

The aforementioned rules are *binary*, *asymmetric* and *one-dimensional* rules



More general association rules:

- **Rules for binary symmetric variables** ({male,female}, {interested,non-interested} ...)
→ two asymmetric binary variables are introduced for each symmetric variable
- **Rules for categorical variables** (education, profession,...)
→ a set of asymmetric binary variables is introduced for each categorical variable
- **Rules for continuous variables** (age, income,...)
→ discretization is first applied to obtain a categorical variable
- **Multi-dimensional rules**
- **Sequential rules** (when transactions are recorded according to a specific temporal sequence)

Hospital infections control

Hospital-associated infections (HAI) are infections whose development is favored by the hospital environment (are those acquired by a patient during his hospital staying).

Very relevant issue (Italy, 2008-2010):

- Recoveries / Year ~ 11 Mln
- Hospital infections / Year ~ 700 – 800 thousands
- Death / Year ~ 7 – 7.5 thousands
- Costs for hospitals and medical centers ~ 1.5 – 3.5 Billion Euros
- Preventable infections ~ 30% (21.000 per year)
- Preventable deaths ~ 30% (2300 per year)

Index of the quality of
hospital assistance



Number of hospital infections every
1000 patients releases

Hospital infections control

Hospital infections are classified according to their localization

Most prominent infections and isolated microorganisms:

- 📌 Urinary tract infections (42% , *Escherichia coli* ~ 30.7%)
- 📌 Surgical site infections (24% , *Staphylococcus aureus* ~ 18.6%)
- 📌 Pneumonia and respiratory tract infections (11% , *Pseudomonas aeruginosa* ~ 16.9%)
- 📌 Other infections (5%)

Hospital infections control

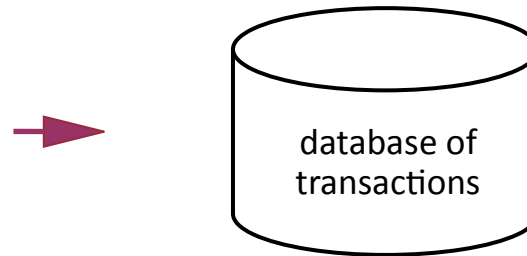
Machine learning-based surveillance system that uses association rules to identify new and interesting patterns in surveillance data

Microorganism investigated: *Pseudomonas aeruginosa*

Objective ⇒ Control the susceptibility of the microorganism to a set of antibiotic agents over time and identify the so-called *events**

Each transaction is composed by the following items:

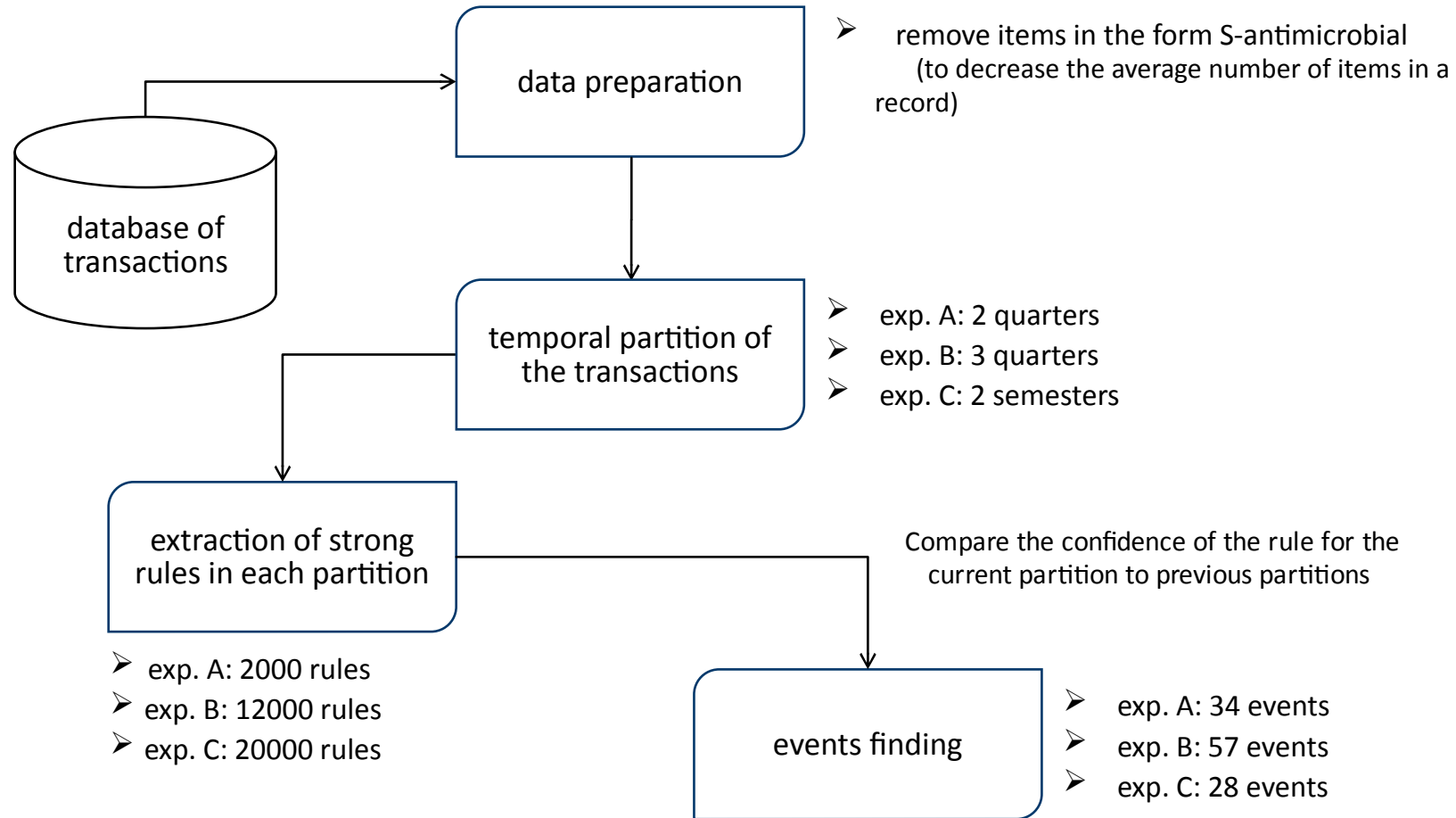
- date reported
- source of isolate (sputum, blood, urine)
- location of patient in hospital
- patient's geographical living area
- antibiogram (results of the testing for the sensitivity to different antibiotics):
R (resistant), I (intermediate resistance), S (susceptible)



*An event describes a significant change in the confidence of an association rule over time.

Hospital infections control

Machine learning-based surveillance system



Hospital infections control

Rule / Event

Suggested action

R~Ciprofloxacin → R~Imipenem

An increase from 7% (3/43) in Q3 to 24% (8/33) in Q4 in the probability that a *Pa* isolate is resistant to imipenem given that it is resistant to ciprofloxacin.

Review of antimicrobial selection in the treatment of patients with *Pa* infection.

SourceUrine → I~Imipenem

An increase from 2% (2/96) in S1 to 9% (12/142) in S2 in the probability that a *Pa* isolate is intermediate to imipenem given that it is from urine.

Utilization review of imipenem

SourceSP → LocNICU*

An increase from 4% (2/54) in S1 to 23% (12/53) in S2 in the probability that a *Pa* isolate is from a patient in the NICU given that the isolate is from sputum.

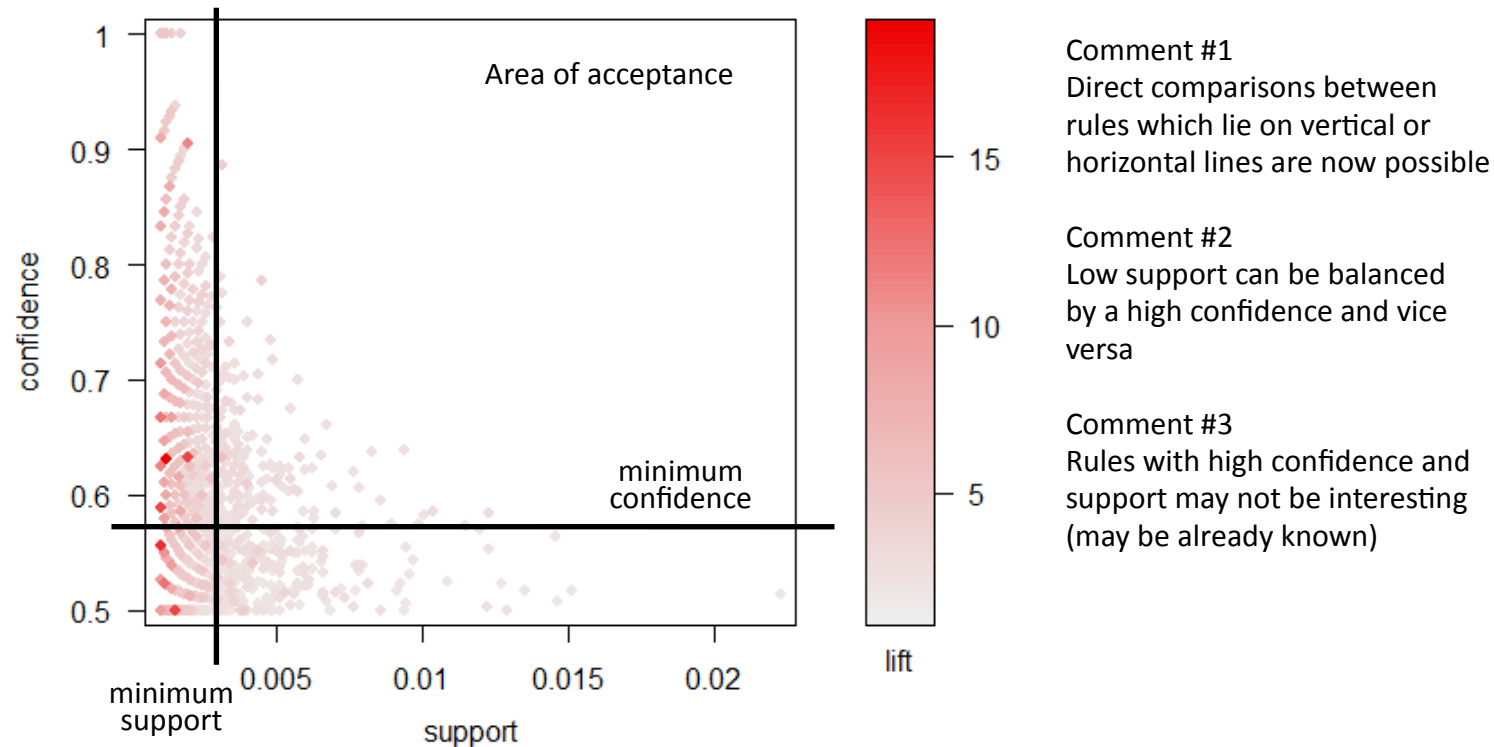
Review the intensive care procedures in the NICU and/or antimicrobial usage.

*Neonatal Intensive Care Unit

Visual Comparison of Association Rules

SCATTER PLOTS

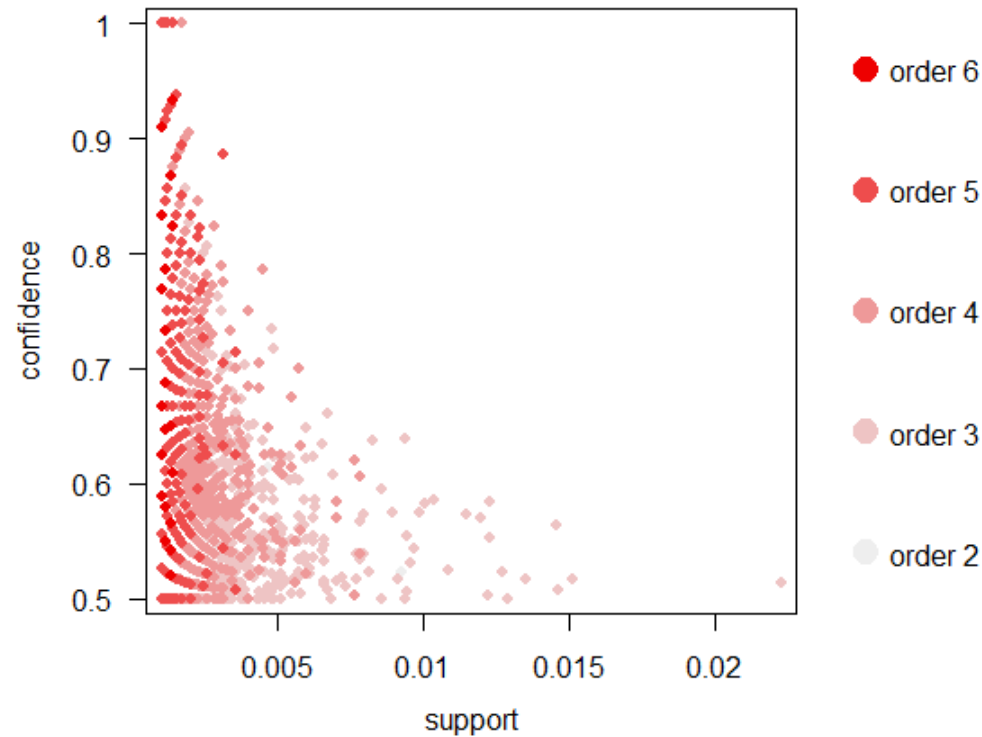
- Display two criteria of interest (i.e. confidence and support) in a single plot.
A third measure (i.e. *lift*) can be used as the color of the points.



Visual Comparison of Association Rules

SCATTER PLOTS: TWO-KEY PLOT

- Special version of a scatter plot.
The color indicates the number of items contained in the rule.

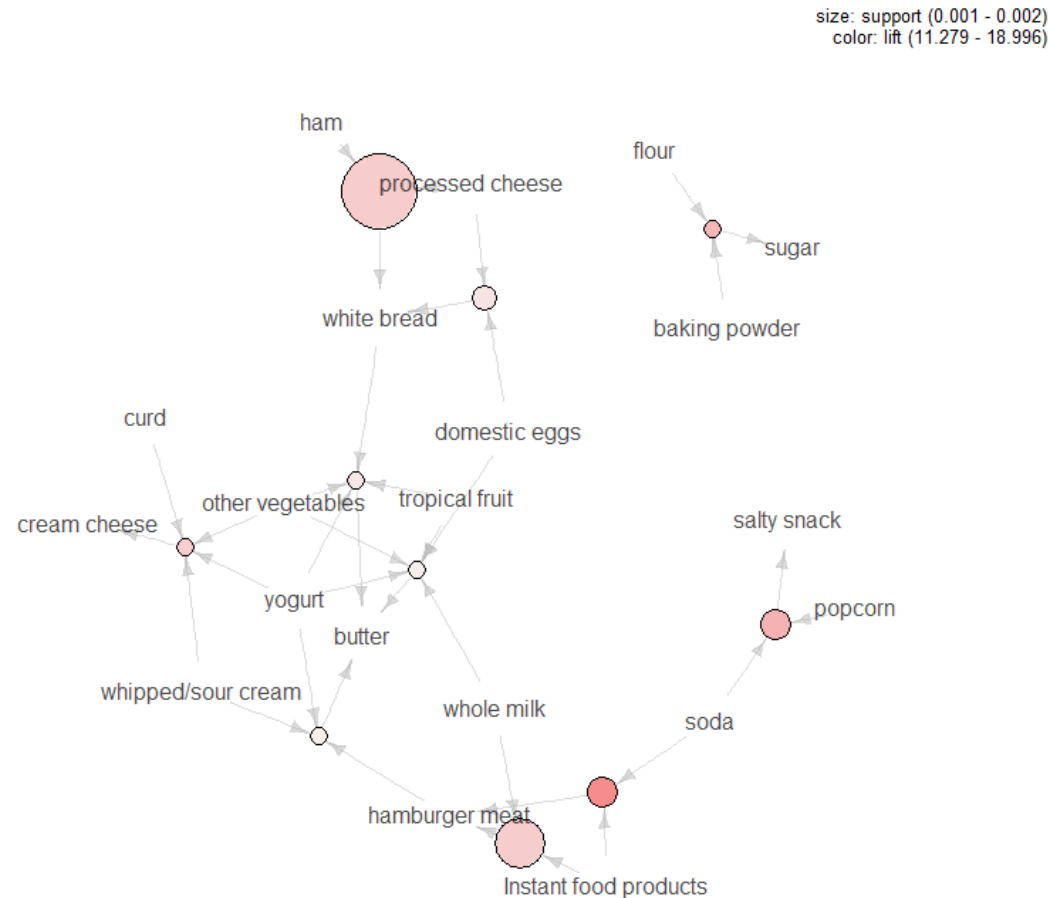


Comment #1
Order and support have a very strong inverse relationship

Visual Comparison of Association Rules

GRAPH-BASED VISUALIZATION

- Visualize association rules using arrows and nodes.
 - Arrows usually represent items and their direction the relationship in the rule.

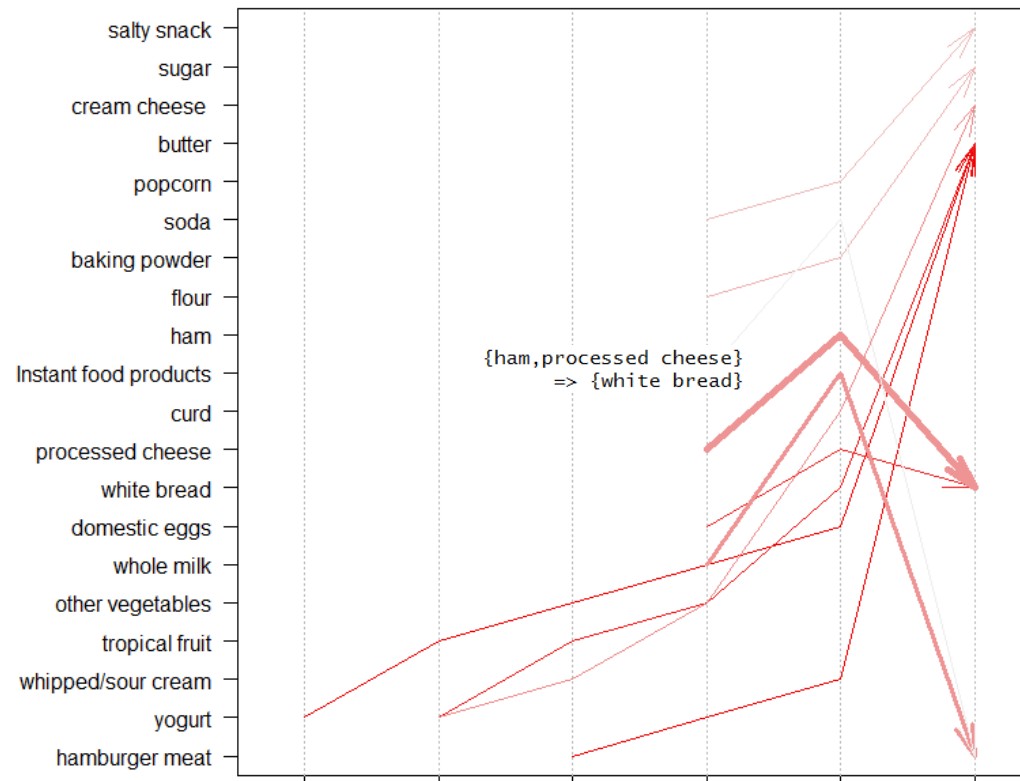


Visual Comparison of Association Rules

PARALLEL COORDINATES PLOT

- Displays the items on the y-axis as nominal values and the x-axis represents the positions in a rule (i.e. first item, second item and so on).

Arrows indicate consequent items (width \rightarrow support, color intensity \rightarrow confidence).



Comment #1
With an increasing number of rules also the number of crossovers between the lines increases.

Visual Comparison of Association Rules

DOUBLE DECKER PLOT

- 🌐 To analyze a single rule and/or comparing the “strength” of different rules.
It displays a contingency table obtained by evaluating the *support* and *confidence* of all the rules generated for each antecedent’s subset.

`{other vegetables,hard cheese,domestic eggs}`
`=> {butter}`

| other veg. | hard cheese | dom. eggs | → | butter | → | ¬ | butter |
|-------------|-------------|-----------|------|--------|------|------|--------|
| COMBINATION | | | SUPP | CONF | SUPP | CONF | |
| no | no | no | 0.07 | 0.39 | 0.11 | 0.61 | |
| no | no | yes | 0.01 | 0.16 | 0.08 | 0.84 | |
| no | yes | no | 0.03 | 0.26 | 0.07 | 0.74 | |
| no | yes | yes | 0.00 | 0.05 | 0.02 | 0.95 | |
| yes | no | no | 0.06 | 0.17 | 0.28 | 0.83 | |
| yes | no | yes | 0.00 | 0.04 | 0.08 | 0.96 | |
| yes | yes | no | 0.00 | 0.07 | 0.05 | 0.93 | |
| yes | yes | yes | 0.12 | 0.98 | 0.00 | 0.02 | |

Visual Comparison of Association Rules

DOUBLE DECKER PLOT

The *support* of an association rule is represented by the highlighted area in the corresponding bin. The *confidence* is given by the height of the highlighted area.



Visual Comparison of Association Rules

DOUBLE DECKER PLOT

Comparison with other rules with similar support and confidence (0.96 and 0.12).



Visual Comparison of Association Rules

DIFFERENCE OF CONFIDENCE

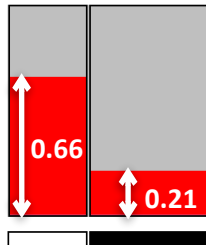
🌐 **Measure how much the body L of the rule contributes to the prediction of the head H**

(we are interested in detecting associations between the presence of itemsets)

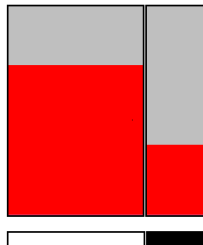
$$doc(L \Rightarrow H) = conf(L \Rightarrow H) - conf(\text{not } L \Rightarrow H)$$

| Rule | | | Support | Confidence | Lift | Doc |
|----------|---|-------------|---------|------------|------|-------|
| heineken | → | corned beef | 0.12 | 0.21 | 0.54 | -0.45 |
| apples | → | heineken | 0.10 | 0.34 | 0.56 | -0.39 |
| herring | → | avocado | 0.17 | 0.34 | 0.97 | -0.02 |
| herring | → | ham | 0.16 | 0.32 | 1.09 | +0.05 |
| cracker | → | heineken | 0.37 | 0.75 | 1.25 | +0.34 |
| coke | → | ice cream | 0.22 | 0.74 | 2.40 | +0.61 |

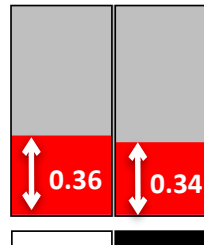
heineken ⇒
corned beef



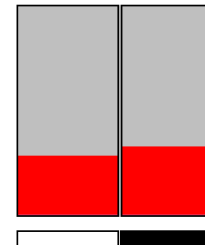
apples ⇒
heineken



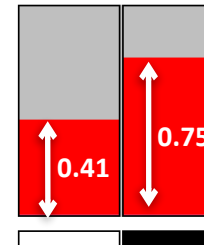
herring ⇒
avocado



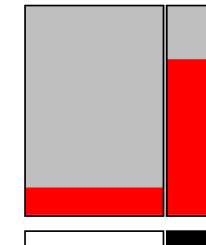
herring ⇒
ham



cracker ⇒
heineken



coke ⇒ ice
cream



Case Study #1: Census Bureau

INCOME ANALYSIS ON CENSUS DATA

- The data set (Adult data set) contains a sample of data (48842 examples) extracted in 1994 from the Census Bureau database.
- Available fields are:
 - age – Age of the individual
 - workclass – Type of employment (Government, Private, Military, etc.)
 - education – Highest level of education achieved for that individual
 - education-num – Highest level of education in numerical form
 - marital-status – Never-married, Separated, Widowed, etc.
 - occupation – Exec-managerial, Farming-fishing, etc.
 - relationship – Family relationship value (Husband, Father, Unmarried, etc.)
 - race – Amer-Indian-Eskimo, Asian-Pac-Islander, Black, etc.
 - sex – Male, Female
 - capital-gain – Capital gains recorded (Income from investment sources, apart from salary)
 - capital-loss – Capital loss recorded (Losses from investment sources)
 - fnlwgt – The # of units in the target population that the responding unit represents
 - hours-per-week – Hours worked per week
 - native-country – Country of origin of the individual
 - income – Annual income of the individual (small: $\leq 50K$ USD, large: otherwise)

Case Study #1: Census Bureau

INCOME ANALYSIS ON CENSUS DATA

- 🎯 **Objective:** Analyze the elements promoting high annual incomes (above 50K USD) based on census data by means of association rules mining
- 🎯 **Numeric fields must be converted into ordinal attributes.**
Let's consider the following mapping:
 - 🎯 **age** →
Levels: Young (0-25), Middle-aged (26-45), Senior (46-65) and Old (66+)
 - 🎯 **hours-per-week** →
Levels: Part-time (0-25), Full-time (25-40), Over-time (40-60) and Workaholic (60+)
 - 🎯 **capital-gain and capital-loss** →
Levels: None (0), Low ($0 < \text{median of the values greater zero}$) and High ($\geq \text{median of the values greater zero}$)

Case Study #2: Medical Appointments

INVESTIGATING SHOWING-UP AND NO-SHOW

- The data set contains medical appointments for 300k individuals. Each appointment is described in terms of 14 attributes.
- Available fields are:
 - Age – Age of the individual (numeric)
 - Gender – Male or Female (factor)
 - Appointment Registration – Date and time of registration (datetime)
 - Appointment Date – Date of the appointment (datetime)
 - Day Of The Week (factor)
 - Diabetes, Alcoholism, Hypertension, Smokes, Tuberculosis (binary)
 - Handicap – Four categories (factor)
 - Sms_Reminder – Number of SMS sent (0,1,2)
 - Awaiting Time – (in days, numeric)
 - Status – ‘Show-Up’ or ‘No-Show’ (factor) (30% are no shows)

Case Study #2: Medical Appointments

INVESTIGATING SHOWING-UP AND NO-SHOWS

Prepare the data set for association rules mining

- Apply the Apriori algorithm to find associations (generate rules must be composed by at most 4 items each).
- Find the best rules (rules with the highest Lift) in the form:
 - { list of items } → { No-Show }
 - { list of items } → { Show-Up }
- Attempt to provide a meaningful interpretation of No-Show rules
- Identify the most important factors in Show-Up rules

