# Probabilistic Approach to Association Rule Mining

Michael Hahsler

Intelligent Data Analysis Lab (IDA@SMU)
Dept. of Engineering Management, Information, and Systems, SMU
mhahsler@lyle.smu.edu

IESEG School of Management
May, 2016

SMU | Bobby B. Lyle School of Engineering

# Table of Contents

# Motivation

We life in the era of big data. Examples:

- Transaction data: Retailers (point-of-sale systems, loyalty card programs) and e-commerce
- Web navigation data: Web analytics, search engines, digital libraries, Wikis, etc.
- Gene expression data: DNA microarrays

# Motivation

We life in the era of big data. Examples:

- Transaction data: Retailers (point-of-sale systems, loyalty card programs) and e-commerce
- Web navigation data: Web analytics, search engines, digital libraries, Wikis, etc.
- Gene expression data: DNA microarrays

Typical size of data sets:

- Typical Retailer: 10–500 product groups and 500–10,000 products
- Amazon: 480+ million products in the US (2015)
- Wikipedia: almost 5 million articles (2015)
- Google: estimated 47+ billion pages in index (2015)
- Human Genome Project: approx. 20,000–25,000 genes in human DNA with 3 billion base pairs.

- Typically 10,000–10 million transactions (shopping baskets, user sessions, observations, patients, etc.)

# Motivation

The aim of association analysis is to find 'interesting' relationships between items (products, documents, etc.). Example: 'purchase relationship':

milk, flour and eggs are frequently bought together.

or

If someone purchases milk and flour then that person often also purchases eggs.

# Motivation

The aim of association analysis is to find 'interesting' relationships between items (products, documents, etc.). Example: 'purchase relationship':

milk, flour and eggs are frequently bought together.
or
If someone purchases milk and flour then that person often also purchases eggs.

Applications of found relationships:

- Retail: Product placement, promotion campaigns, product assortment decisions, etc.
  → exploratory market basket analysis (Russell *et al.*, 1997; Berry and Linoff, 1997; Schnedlitz *et al.*, 2001; Reutterer *et al.*, 2007).
- E-commerce, dig. libraries, search engines: Personalization, mass customization
  → recommender systems, item-based collaborative filtering (Sarwar *et al.*, 2001; Linden *et al.*, 2003; Geyer-Schulz and Hahsler, 2003).

# Table of Contents

# Transaction Data

Example of market basket data:

| transaction ID | items |
|---|---|
| 1 | milk, bread |
| 2 | bread, butter |
| 3 | beer |
| 4 | milk, bread, butter |
| 5 | bread, butter |

| | | items | | | |
|---|---|---|---|---|---|
| | | milk | bread | butter | beer |
| transactions | 1 | 1 | 1 | 0 | 0 |
| | 2 | 0 | 1 | 1 | 0 |
| | 3 | 0 | 0 | 0 | 1 |
| | 4 | 1 | 1 | 1 | 0 |
| | 5 | 0 | 1 | 1 | 0 |

Formally, let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of $n$ binary attributes called items. Let $\mathcal{D} = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the database. Each transaction in $\mathcal{D}$ has an unique transaction ID and contains a subset of the items in $I$.

Note: Non-transaction data can be made into transaction data using binarization.

# Table of Contents

# Association Rules

A rule takes the form $X \rightarrow Y$

- $X, Y \subseteq I$
- $X \cap Y = \emptyset$
- $X$ and $Y$ are called itemsets.
- $X$ is the rule's antecedent (left-hand side)
- $Y$ is the rule's consequent (right-hand side)

**Example**

$$\{\text{milk, flower, bread}\} \rightarrow \{\text{eggs}\}$$

# Association Rules

To select 'interesting' association rules from the set of all possible rules, two measures are used (Agrawal *et al.*, 1993):

1. Support of an itemset $Z$ is defined as $\text{supp}(Z) = n_Z/n$.
   $\rightarrow$ share of transactions in the database that contains $Z$.
2. Confidence of a rule $X \rightarrow Y$ is defined as
   $$\text{conf}(X \rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$$
   $\rightarrow$ share of transactions containing $Y$ in all the transactions containing $X$.

# Association Rules

To select 'interesting' association rules from the set of all possible rules, two measures are used (Agrawal *et al.*, 1993):

1. Support of an itemset $Z$ is defined as $\text{supp}(Z) = n_Z/n$.
   $\rightarrow$ share of transactions in the database that contains $Z$.
2. Confidence of a rule $X \rightarrow Y$ is defined as
   $$\text{conf}(X \rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$$
   $\rightarrow$ share of transactions containing $Y$ in all the transactions containing $X$.

Each association rule $X \rightarrow Y$ has to satisfy the following restrictions:

$$\text{supp}(X \cup Y) \geq \sigma$$
$$\text{conf}(X \rightarrow Y) \geq \gamma$$

$\rightarrow$ called the support-confidence framework.

# Minimum Support

**Idea:** Set a user-defined threshold for support since more frequent itemsets are typically more important. E.g., frequently purchased products generally generate more revenue.

# Minimum Support

**Idea:** Set a user-defined threshold for support since more frequent itemsets are typically more important. E.g., frequently purchased products generally generate more revenue.

**Problem:** For $k$ items (products) we have $2^k - k - 1$ possible relationships between items. Example: $k = 100$ leads to more than $10^{30}$ possible associations.

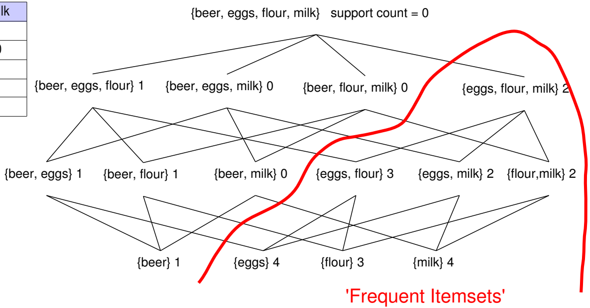
# Minimum Support

**Idea:** Set a user-defined threshold for support since more frequent itemsets are typically more important. E.g., frequently purchased products generally generate more revenue.

**Problem:** For $k$ items (products) we have $2^k - k - 1$ possible relationships between items. Example: $k = 100$ leads to more than $10^{30}$ possible associations.

**Apriori property** (Agrawal and Srikant, 1994): The support of an itemset cannot increase by adding an item. Example: $\sigma = .4$ (support count $\geq 2$)

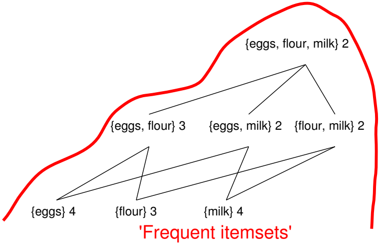
| Transaction ID | beer | eggs | flour | milk |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 0 |
| 3 | 0 | 1 | 0 | 1 |
| 4 | 0 | 1 | 1 | 1 |
| 5 | 0 | 0 | 0 | 1 |




→ Basis for efficient algorithms (Apriori, Eclat).

# Minimum Confidence

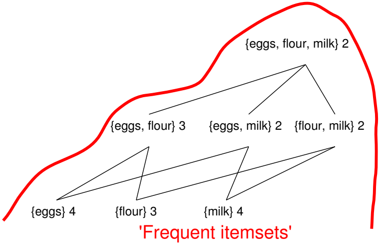
From the set of frequent itemsets all rules which satisfy the threshold for confidence $\mathrm{conf}(X \to Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)} \geq \gamma$ are generated.




| | | | Confidence |
|---|---|---|---|
| {eggs} | $\to$ | {flour} | $3/4 = 0.75$ |
| {flour} | $\to$ | {eggs} | $3/3 = 1$ |
| {eggs} | $\to$ | {milk} | $2/4 = 0.5$ |
| {milk} | $\to$ | {eggs} | $2/4 = 0.5$ |
| {flour} | $\to$ | {milk} | $2/3 = 0.67$ |
| {milk} | $\to$ | {flour} | $2/4 = 0.5$ |
| {eggs, flour} | $\to$ | {milk} | $2/3 = 0.67$ |
| {eggs, milk} | $\to$ | {flour} | $2/2 = 1$ |
| {flour, milk} | $\to$ | {eggs} | $2/2 = 1$ |
| {eggs} | $\to$ | {flour, milk} | $2/4 = 0.5$ |
| {flour} | $\to$ | {eggs, milk} | $2/3 = 0.67$ |
| {milk} | $\to$ | {eggs, flour} | $2/4 = 0.5$ |

# Minimum Confidence

From the set of frequent itemsets all rules which satisfy the threshold for confidence $\mathrm{conf}(X \to Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)} \geq \gamma$ are generated.




| | | | Confidence |
|---|---|---|---|
| {eggs} | $\to$ | {flour} | $3/4 = 0.75$ |
| {flour} | $\to$ | {eggs} | $3/3 = 1$ |
| {eggs} | $\to$ | {milk} | $2/4 = 0.5$ |
| {milk} | $\to$ | {eggs} | $2/4 = 0.5$ |
| {flour} | $\to$ | {milk} | $2/3 = 0.67$ |
| {milk} | $\to$ | {flour} | $2/4 = 0.5$ |
| {eggs, flour} | $\to$ | {milk} | $2/3 = 0.67$ |
| {eggs, milk} | $\to$ | {flour} | $2/2 = 1$ |
| {flour, milk} | $\to$ | {eggs} | $2/2 = 1$ |
| {eggs} | $\to$ | {flour, milk} | $2/4 = 0.5$ |
| {flour} | $\to$ | {eggs, milk} | $2/3 = 0.67$ |
| {milk} | $\to$ | {eggs, flour} | $2/4 = 0.5$ |

At $\gamma = 0.7$ the following set of rules is generated:

| | | | Support | Confidence |
|---|---|---|---|---|
| {eggs} | $\to$ | {flour} | $3/5 = 0.6$ | $3/4 = 0.75$ |
| {flour} | $\to$ | {eggs} | $3/5 = 0.6$ | $3/3 = 1$ |
| {eggs, milk} | $\to$ | {flour} | $2/5 = 0.4$ | $2/2 = 1$ |
| {flour, milk} | $\to$ | {eggs} | $2/5 = 0.4$ | $2/2 = 1$ |

# Table of Contents

# Probabilistic interpretation of Support and Confidence

## Support

$$\mathrm{supp}(Z) = n_Z/n$$

corresponds to an estimate for $\hat{P}(E_Z) = n_Z/n$, the probability for the event that itemset $Z$ is contained in a transaction.

# Probabilistic interpretation of Support and Confidence

Support

$$\text{supp}(Z) = n_Z/n$$

corresponds to an estimate for $\hat{P}(E_Z) = n_Z/n$, the probability for the event that itemset $Z$ is contained in a transaction.

Confidence can be interpreted as an estimate for the conditional probability

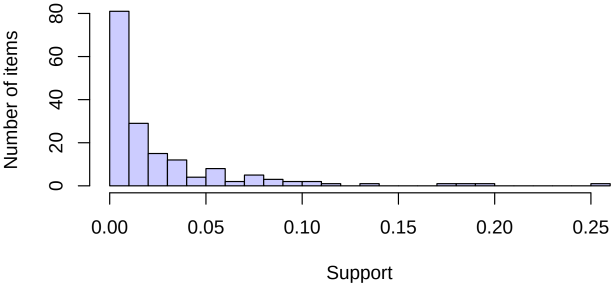
$$P(E_Y|E_X) = \frac{P(E_X \cap E_Y)}{P(E_X)}.$$

This directly follows the definition of confidence:

$$\text{conf}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = \frac{\hat{P}(E_X \cap E_Y)}{\hat{P}(E_X)}.$$

# Weaknesses of Support and Confidence

- Support suffers from the 'rare item problem' (Liu *et al.*, 1999a): Infrequent items not meeting minimum support are ignored which is problematic if rare items are important.
  E.g. rarely sold products which account for a large part of revenue or profit.
  Typical support distribution (retail point-of-sale data with 169 items):




- Support falls rapidly with itemset size. A threshold on support favors short itemsets (Seno and Karypis, 2005).

# Weaknesses of Support and Confidence

- Confidence ignores the frequency of $Y$ (Aggarwal and Yu, 1998; Silverstein *et al.*, 1998).

| | X=0 | X=1 | Σ |
|---|---|---|---|
| Y=0 | 5 | 5 | 10 |
| Y=1 | 70 | **20** | **90** |
| Σ | 75 | **25** | 100 |

$$\mathrm{conf}(X \to Y) = \frac{n_{X \cup Y}}{n_X} = \frac{20}{25} = .8$$

**Weakness:** Confidence of the rule is relatively high with $\hat{P}(E_Y|E_X) = .8$. But the unconditional probability $\hat{P}(E_Y) = n_Y/n = 90/100 = .9$ is higher!

# Weaknesses of Support and Confidence

- Confidence ignores the frequency of $Y$ (Aggarwal and Yu, 1998; Silverstein *et al.*, 1998).

|  | X=0 | X=1 | Σ |
|---|---|---|---|
| Y=0 | 5 | 5 | 10 |
| Y=1 | 70 | **20** | **90** |
| Σ | 75 | **25** | 100 |

$$\mathrm{conf}(X \to Y) = \frac{n_{X \cup Y}}{n_X} = \frac{20}{25} = .8$$

**Weakness:** Confidence of the rule is relatively high with $\hat{P}(E_Y|E_X) = .8$. But the unconditional probability $\hat{P}(E_Y) = n_Y/n = 90/100 = .9$ is higher!

- The thresholds for support and confidence are user-defined.
  In practice, the values are chosen to produce a 'manageable' number of frequent itemsets or rules.

  $\rightarrow$ What is the risk and cost attached to using spurious rules or missing important in an application?

# Lift

The measure lift (interest, Brin *et al.*, 1997) is defined as

$$\mathrm{lift}(X \rightarrow Y) = \frac{\mathrm{conf}(X \rightarrow Y)}{\mathrm{supp}(Y)} = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) \cdot \mathrm{supp}(Y)}$$

and can be interpreted as an estimate for $P(E_X \cap E_Y)/(P(E_X) \cdot P(E_Y))$.
$\rightarrow$ Measure for the deviation from stochastic independence:

$$P(E_X \cap E_Y) = P(E_X) \cdot P(E_Y)$$

# Lift

The measure lift (interest, Brin *et al.*, 1997) is defined as

$$\text{lift}(X \to Y) = \frac{\text{conf}(X \to Y)}{\text{supp}(Y)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \cdot \text{supp}(Y)}$$

and can be interpreted as an estimate for $P(E_X \cap E_Y)/(P(E_X) \cdot P(E_Y))$.
$\to$ Measure for the deviation from stochastic independence:

$$P(E_X \cap E_Y) = P(E_X) \cdot P(E_Y)$$

In marketing values of lift are interpreted as:

- $\text{lift}(X \to Y) = 1$ ... $X$ and $Y$ are independent
- $\text{lift}(X \to Y) > 1$ ... complementary effects between $X$ and $Y$
- $\text{lift}(X \to Y) < 1$ ... substitution effects between $X$ and $Y$

# Lift

The measure lift (interest, Brin *et al.*, 1997) is defined as

$$\text{lift}(X \to Y) = \frac{\text{conf}(X \to Y)}{\text{supp}(Y)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \cdot \text{supp}(Y)}$$

and can be interpreted as an estimate for $P(E_X \cap E_Y)/(P(E_X) \cdot P(E_Y))$.
$\to$ Measure for the deviation from stochastic independence:

$$P(E_X \cap E_Y) = P(E_X) \cdot P(E_Y)$$

In marketing values of lift are interpreted as:

- $\text{lift}(X \to Y) = 1 \ldots X$ and $Y$ are independent
- $\text{lift}(X \to Y) > 1 \ldots$ complementary effects between $X$ and $Y$
- $\text{lift}(X \to Y) < 1 \ldots$ substitution effects between $X$ and $Y$

**Example**

| | X=0 | X=1 | $\Sigma$ |
|---|---|---|---|
| Y=0 | 5 | 5 | 10 |
| Y=1 | 70 | **20** | **90** |
| $\Sigma$ | 75 | **25** | 100 |

$$\text{lift}(X \to Y) = \frac{.2}{.25 \cdot .9} = .89$$

**Weakness:** small counts!

# Chi-Square Test for Independence

Tests for significant deviations from stochastic independence (Silverstein *et al.*, 1998; Liu *et al.*, 1999b).

**Example:** $2 \times 2$ contingency table ($l = 2$ dimensions) for rule $X \to Y$.

| | X=0 | X=1 | $\Sigma$ |
|---|---|---|---|
| Y=0 | 5 | 5 | 10 |
| Y=1 | 70 | 20 | 90 |
| $\Sigma$ | 75 | 25 | 100 |

Null hypothesis: $P(E_X \cap E_Y) = P(E_X) \cdot P(E_Y)$ with test statistic

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})} \quad \text{with} \quad E(n_{ij}) = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$

asymptotically approaches a $\chi^2$ distribution with $2^l - l - 1$ degrees of freedom.

The result of the test for the contingency table above:

$X^2 = 3.7037, \mathrm{df} = 1, \text{p-value} = 0.05429$

$\to$ The null hypothesis (independence) can not be be rejected at $\alpha = 0.05$.

**Weakness:** Bad approximation for $E(n_{ij}) < 5$; multiple testing.

# Table of Contents

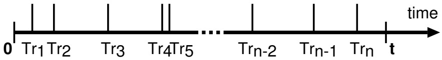
# The Independence Model

1. Transactions occur following a homogeneous Poisson process with parameter $\theta$ (intensity).


$$P(N = n) = \frac{e^{-\theta t}(\theta t)^n}{n!}$$

# The Independence Model

1. Transactions occur following a homogeneous Poisson process with parameter $\theta$ (intensity).




$$P(N = n) = \frac{e^{-\theta t}(\theta t)^n}{n!}$$

2. Each item has the occurrence probability $p_i$ and each transaction is the result of $k$ (number of items) independent Bernoulli trials.

| | $i_1$ | $i_2$ | $i_3$ | ... | $i_k$ |
|---|---|---|---|---|---|
| p | 0.0050 | 0.0100 | 0.0003 | ... | 0.0250 |
| $Tr_1$ | 0 | 1 | 0 | ... | 1 |
| $Tr_2$ | 0 | 1 | 0 | ... | 1 |
| $Tr_3$ | 0 | 1 | 0 | ... | 0 |
| $Tr_4$ | 0 | 0 | 0 | ... | 0 |
| ... | . | . | . | ... | . |
| $Tr_{n-1}$ | 1 | 0 | 0 | ... | 1 |
| $Tr_n$ | 0 | 0 | 1 | ... | 1 |
| $n_i$ | 99 | 201 | 7 | ... | 411 |

$$P(N_i = n_i) = \sum_{m=n_i}^{\infty} P(N_i = n_i | N = n) \cdot P(N = n) = \frac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!} \quad \text{with} \quad \lambda_i = p_i \theta t$$

# Table of Contents

# Application: Evaluate Quality Measures

Authors typically construct examples where support, confidence and lift have problems (see e.g., Brin *et al.*, 1997; Aggarwal and Yu, 1998; Silverstein *et al.*, 1998).

**Idea:** Compare the behavior of measures on real-world data and on data simulated using the independence model (Hahsler *et al.*, 2006; Hahsler and Hornik, 2007).

# Application: Evaluate Quality Measures

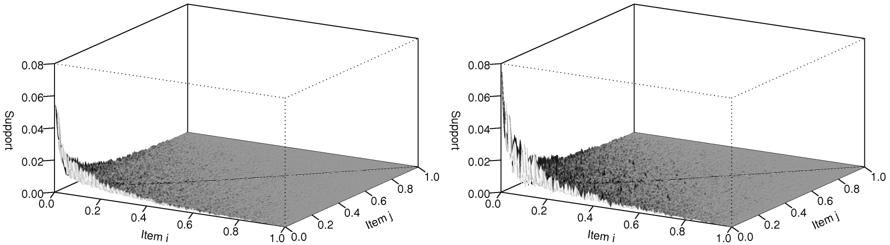
Authors typically construct examples where support, confidence and lift have problems (see e.g., Brin *et al.*, 1997; Aggarwal and Yu, 1998; Silverstein *et al.*, 1998).

**Idea:** Compare the behavior of measures on real-world data and on data simulated using the independence model (Hahsler *et al.*, 2006; Hahsler and Hornik, 2007).

Characteristics of used data set (typical retail data set).

- $t = 30$ days
- $k = 169$ product groups
- $n = 9835$ transactions
- Estimated $\theta = n/t = 327.2$ transactions per day.
- We estimate $p_i$ using the observed frequencies $n_i/n$.

# Comparison: Support

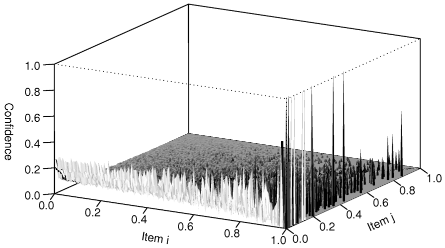



Simulated data

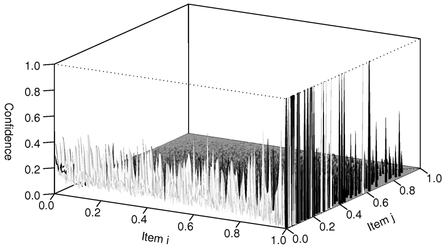
Retail data

Only rules of the form: $\{i_i\} \rightarrow \{i_j\}$
**X-axis:** Items $i_i$ sorted by decreasing support.
**Y-axis:** Items $i_j$ sorted by decreasing support.
**Z-axis:** Support of rule.

# Comparison: Confidence




Simulated data



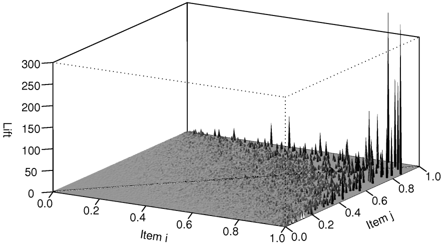

Retail data

$$\mathrm{conf}(\{i_i\} \rightarrow \{i_j\}) = \frac{\mathrm{supp}(\{i_i, i_j\})}{\mathrm{supp}(\{i_i\})}$$

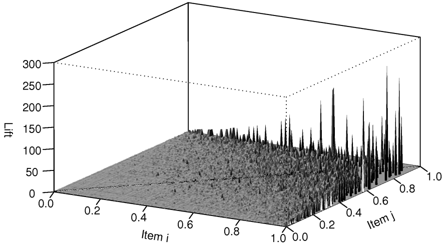
Systematic influence of support

- Confidence decreases with support of the right-hand side ($i_j$).
- Spikes with extremely low-support items in the left-hand side ($i_i$).

# Comparison: Lift




Simulated data

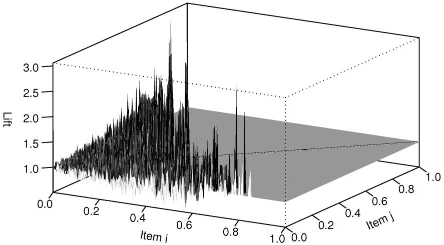



Retail data

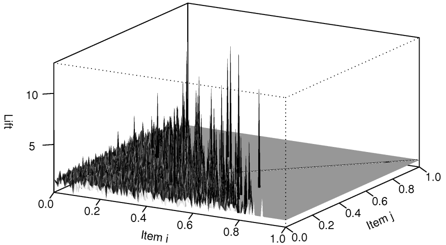
$$\mathrm{lift}(\{i_i\} \rightarrow \{i_j\}) = \frac{\mathrm{supp}(\{i_i, i_j\})}{\mathrm{supp}(\{i_i\}) \cdot \mathrm{supp}(\{i_j\})}$$

- Similar distribution with extreme values for items with low support.

# Comparison: Lift + Minimum Support




Simulated data
(min. support: $\sigma = .1\%$)




Retail data
(min. support: $\sigma = .1\%$)

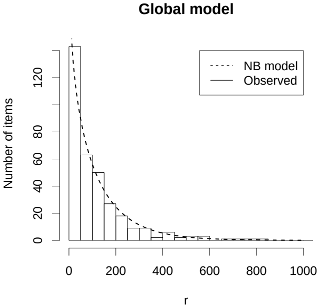
- Considerably higher lift values in retail data (indicate the existence of associations).
- Strong systematic influence of support.
- Highest lift values at the support-confidence border (Bayardo Jr. and Agrawal, 1999). If lift is used to sort found rules, small changes of minimum support/minimum confidence totally change the result.

# Table of Contents

# Application: NB-Frequent Itemsets

**Idea:** Identification of interesting associations as deviations from the independence model (Hahsler, 2006).

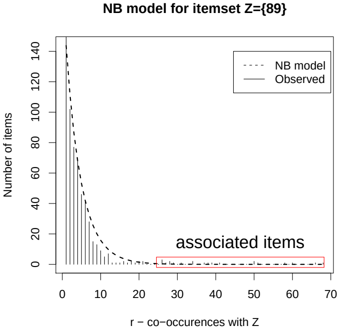
1. Estimation of a global independence model using the frequencies of items in the database.
   The independence model is a mixture of $k$ (number of items) independent homogeneous Poisson processes. Parameters $\lambda_i$ in the population are chosen from a $\Gamma$ distribution.




Number of items which occur in $r = \{0, 1, \ldots, r_{max}\}$ transactions
$\rightarrow$ Negative binomial distribution.

# NB-Frequent Itemsets

2. Select all transactions for itemset $Z$. We expect all items which are independent of $Z$ to occur in the selected transactions following the (rescaled) global independence model. Associated items co-occur too frequently with $Z$.

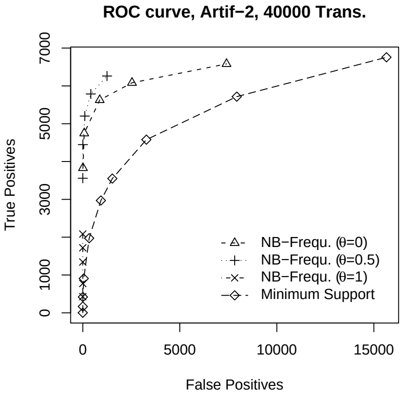


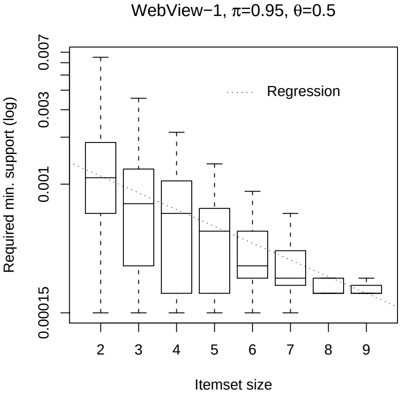

- Rescaling of the model for $Z$ by the number of incidences.
- Uses a user-defined threshold $1 - \pi$ for the number of accepted 'spurious associations'.
- Restriction of the search space by recursive definition of parameter $\theta$.

Details about the estimation procedure for the global model (EM), the mining algorithm and evaluation of effectiveness can be found in Hahsler (2006).

# NB-Frequent Itemsets

Mine NB-frequent itemsets from an artificial data set with know patterns.




- Performs better than support in filtering spurious itemsets.
- Automatically decreases the required support with itemset size.

# Table of Contents

# Hyper-Confidence

**Idea:** Develop a confidence-like measure based on the probabilistic model (Hahsler and Hornik, 2007).

**Informally:** How confident, 0–100%, are we that a rule is not just the result of random co-occurrences?

# Hyper-Confidence

**Idea:** Develop a confidence-like measure based on the probabilistic model (Hahsler and Hornik, 2007).

**Informally:** How confident, 0–100%, are we that a rule is not just the result of random co-occurrences?

Model the number of transactions which contain rule $X \rightarrow Y$ $(X \cup Y)$ as a random variable $N_{XY}$. Give the frequencies $n_X$ and $n_Y$ and independence, $N_{XY}$ has a hypergeometric distribution.

> *The hypergeometric distribution arises for the 'urn problem': An urn contains $w$ white and $b$ black balls. $k$ balls are randomly drawn from the urn without replacement. The number of white balls drawn is then a hypergeometric distributed random variable.*

# Hyper-Confidence

> *The hypergeometric distribution arises for the 'urn problem': An urn contains $w$ white and $b$ black balls. $k$ balls are randomly drawn from the urn without replacement. The number of white balls drawn is then a hypergeometric distributed random variable.*

Application: Under independence, the database can be seen as an urn with $n_X$ 'white' transactions (contain $X$) and $n - n_X$ 'black' transactions (do not contain $X$). We randomly assign $Y$ to $n_Y$ transactions in the database. The number of transactions that contain $Y$ and $X$ is a hypergeometric distributed random variable.

The probability that $X$ and $Y$ co-occur in exactly $r$ transactions given independence, $n$, $n_X$ and $n_Y$, is

$$P(N_{XY} = r) = \frac{\binom{n_Y}{r}\binom{n-n_Y}{n_X-r}}{\binom{n}{n_X}}.$$

# Hyper-Confidence

$$\text{hyper-confidence}(X \to Y) = P(N_{XY} < n_{XY}) = \sum_{i=0}^{n_{XY}-1} P(N_{XY} = i)$$

A hyper-confidence value close to $1$ indicates that the observed frequency $n_{XY}$ is too high for the assumption of independence and that between $X$ and $Y$ exists a complementary effect.

As for other measures of association, we can use a threshold:

$$\text{hyper-confidence}(X \to Y) \geq \gamma$$

**Interpretation:** At $\gamma = .99$ each accepted rule has a chance of less than $1\%$ that the large value of $n_{XY}$ is just a random deviation (given $n_X$ and $n_Y$) .

# Hyper-Confidence

$2 \times 2$ contingency table for rule $X \rightarrow Y$

| | $X = 0$ | $X = 1$ | |
|---|---|---|---|
| $Y = 0$ | $n - n_Y - n_X - N_{XY}$ | $n_X - N_{XY}$ | $n - n_Y$ |
| $Y = 1$ | $n_Y - N_{XY}$ | $N_{XY}$ | $n_Y$ |
| | $n - n_X$ | $n_X$ | $n$ |

Using minimum hyper-confidence ($\gamma$) is equivalent to Fisher's exact test.

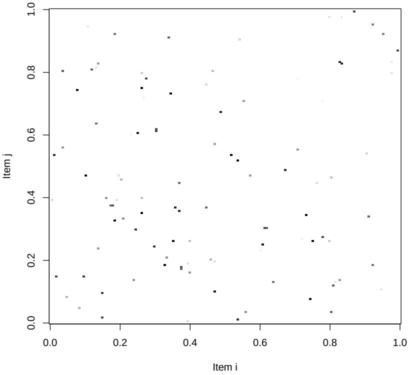
> *Fisher's exact test is a permutation test that calculates the probability of observing an even more extreme value for given fixed marginal frequencies (one-tailed test). Fisher showed that the probability of a certain configuration follows a hypergeometric distribution.*

The p-value of Fisher's exact test is

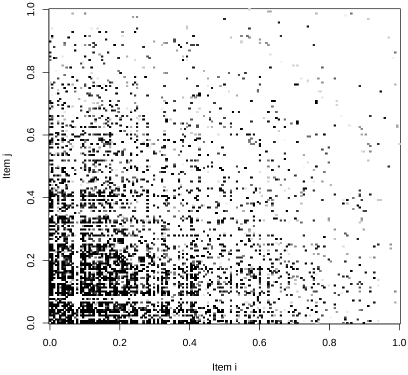
$$\text{p-value} = 1 - \text{hyper-confidence}(X \rightarrow Y)$$

and the significance level is $\alpha = 1 - \gamma$.

# Hyper-Confidence: Complementary Effects




Simulated data

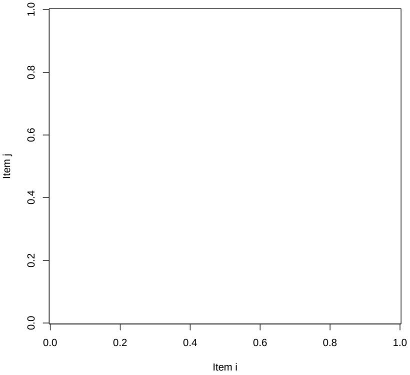



Retail data

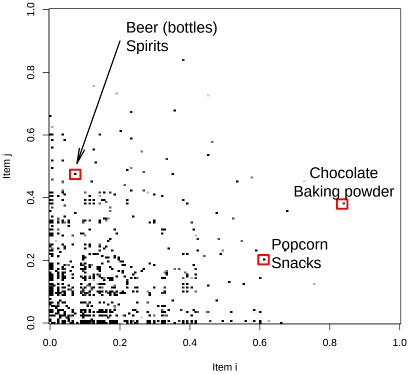
$$\gamma = .99$$

Expected spurious rules: $\alpha\binom{k}{2} = 141.98$

# Hyper-Confidence: Complementary Effects




Simulated data




Retail data

$$\gamma = .9999993$$

Bonferroni correction $\alpha = \frac{\alpha_i}{\binom{k}{2}}$

# Hyper-Confidence: Substitution Effects

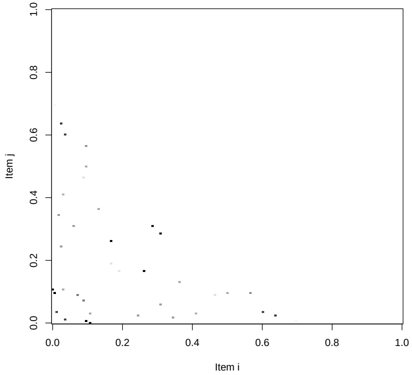
Hyper-confidence uncovers complementary effects between items.
To find substitution effects we have to adapt hyper-confidence as follows:

$$\text{hyper-confidence}^{\text{sub}}(X \rightarrow Y) = P(N_{XY} > n_{X,Y}) = 1 - \sum_{i=0}^{n_{XY}} P(N_{XY} = i)$$

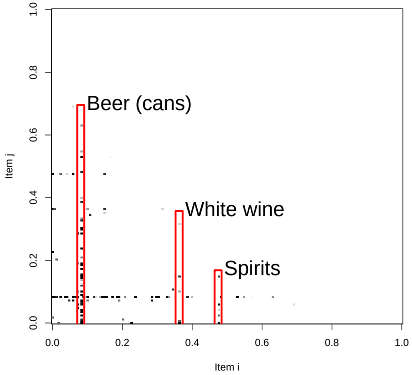
with

$$\text{hyper-confidence}^{\text{sub}}(X \rightarrow Y) \geq \gamma$$

# Hyper-Confidence: Substitution Effects




Simulated data

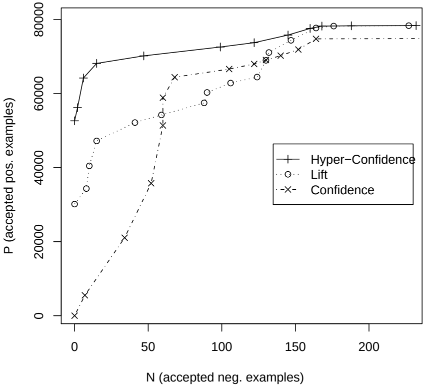



Retail data

$$\gamma = .99$$

# Hyper-Confidence: Simulated Data

PN-Graph for the synthetic data set *T10I4D100K*
with a *corruption rate* of .9 (Agrawal and Srikant, 1994).



# Table of Contents

# Conclusion

The support-confidence framework cannot answer some important questions sufficiently:

- What are sensible thresholds for different applications?
- What is the risk of accepting spurious rules?

# Conclusion

The support-confidence framework cannot answer some important questions sufficiently:

- What are sensible thresholds for different applications?
- What is the risk of accepting spurious rules?

Probabilistic models can help to:

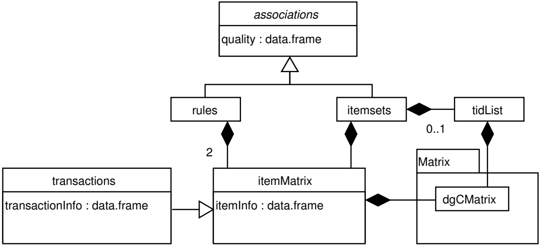
- Evaluate and compare measures of interestingness, data mining processes or complete data mining systems (with synthetic data from models with dependencies).
- Develop new mining strategies and measures (e.g., NB-frequent itemsets, hyper-confidence).
- Use statistical test theory as a solid basis to quantify risk and justify thresholds.

# Thank you for your attention!

- Contact information and full papers can be found at `http://michael.hahsler.net`
- The presented models and measures are implemented in **arules** (an extension package for R, a free software environment for statistical computing and graphics; see `http://www.r-project.org/`).

# Table of Contents

# The **arules** Infrastructure




Simplified UML class diagram implemented in R (S4)

- Uses the sparse matrix representation (from package **Matrix** by Bates & Maechler (2005)) for transactions and associations.
- Abstract associations class for extensibility.
- Interfaces for Apriori and Eclat (implemented by Borgelt (2003)) to mine association rules and frequent itemsets.
- Provides comprehensive analysis and manipulation capabilities for transactions and associations (subsetting, sampling, visual inspection, etc.).
- **arulesViz** provides visualizations.

# Simple Example

```
R> library("arules")
R> data("Groceries")

R> Groceries
transactions in sparse format with
 9835 transactions (rows) and
 169 items (columns)

R> rules <- apriori(Groceries, parameter = list(support = .001))

apriori - find association rules with the apriori algorithm
version 4.21 (2004.05.09)        (c) 1996-2004   Christian Borgelt
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
sorting and recoding items ... [157 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 5 6 done [0.05s].
writing ... [410 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

# Simple Example

```
R> rules
set of 410 rules

R> inspect(head(sort(rules, by = "lift"), 3))
  lhs                     rhs                    support confidence     lift
1 {liquor,
   red/blush wine}     => {bottled beer}     0.001931876  0.9047619 11.23527
2 {citrus fruit,
   other vegetables,
   soda,
   fruit}              => {root vegetables} 0.001016777  0.9090909  8.34040
3 {tropical fruit,
   other vegetables,
   whole milk,
   yogurt,
   oil}                => {root vegetables} 0.001016777  0.9090909  8.34040
```

# References I


C. C. Aggarwal and P. S. Yu. A new framework for itemset generation. In *PODS 98, Symposium on Principles of Database Systems*, pages 18–24, Seattle, WA, USA, 1998.

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, pages 487–499, Santiago, Chile, September 1994.

R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington D.C., May 1993.

Robert J. Bayardo Jr. and Rakesh Agrawal. Mining the most interesting rules. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 145–154. ACM Press, 1999.

M. J. Berry and G. Linoff. *Data Mining Techniques*. Wiley, New York, 1997.

Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255–264, Tucson, Arizona, USA, May 1997.

Andreas Geyer-Schulz and Michael Hahsler. Comparing two recommender algorithms with the help of recommendations by peers. In O.R. Zaiane, J. Srivastava, M. Spiliopoulou, and B. Masand, editors, *WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles 4th International Workshop, Edmonton, Canada, July 2002, Revised Papers*, Lecture Notes in Computer Science LNAI 2703, pages 137–158. Springer-Verlag, 2003.

Michael Hahsler and Kurt Hornik. New probabilistic interest measures for association rules. *Intelligent Data Analysis*, 11(5):437–455, 2007.

Michael Hahsler, Kurt Hornik, and Thomas Reutterer. Implications of probabilistic data modeling for mining association rules. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul, editors, *From Data and Information Analysis to Knowledge Engineering*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 598–605. Springer-Verlag, 2006.

Michael Hahsler. A model-based frequency constraint for mining associations from transaction data. *Data Mining and Knowledge Discovery*, 13(2):137–166, September 2006.

# References II


Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, Jan/Feb 2003.

Bing Liu, Wynne Hsu, and Yiming Ma. Mining association rules with multiple minimum supports. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 337–341. ACM Press, 1999.

Bing Liu, Wynne Hsu, and Yiming Ma. Pruning and summarizing the discovered associations. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 125–134. ACM Press, 1999.

Thomas Reutterer, Michael Hahsler, and Kurt Hornik. Data Mining und Marketing am Beispiel der explorativen Warenkorbanalyse. *Marketing ZFP*, 29(3):165–181, 2007.

Gary J. Russell, David Bell, Anand Bodapati, Christina Brown, Joengwen Chiang, Gary Gaeth, Sunil Gupta, and Puneet Manchanda. Perspectives on multiple category choice. *Marketing Letters*, 8(3):297–305, 1997.

B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the Tenth International World Wide Web Conference, Hong Kong, May 1-5*, 2001.

P. Schnedlitz, T. Reutterer, and W. Joos. Data-Mining und Sortimentsverbundanalyse im Einzelhandel. In H. Hippner, U. Müsters, M. Meyer, and K.D. Wilde, editors, *Handbuch Data Mining im Marketing. Knowledge Discovery in Marketing Databases*, pages 951–970. Vieweg Verlag, Wiesbaden, 2001.

Masakazu Seno and George Karypis. Finding frequent itemsets using length-decreasing support constraint. *Data Mining and Knowledge Discovery*, 10:197–228, 2005.

Craig Silverstein, Sergey Brin, and Rajeev Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2:39–68, 1998.