Probabilistic Approach to Association Rule Mining

Michael Hahsler

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

Department of Statistics & Actuarial Science
University of Waterloo
October, 2018
<table>
<thead>
<tr>
<th>Table of Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Motivation</td>
</tr>
<tr>
<td>2. Transaction Data</td>
</tr>
<tr>
<td>3. Introduction to Association Rules</td>
</tr>
<tr>
<td>4. Probabilistic Interpretation, Weaknesses and Enhancements</td>
</tr>
<tr>
<td>5. A Probabilistic Independence Model</td>
</tr>
<tr>
<td>- Application: Evaluate Quality Measures</td>
</tr>
<tr>
<td>- Application: NB-Frequent Itemsets</td>
</tr>
<tr>
<td>- Application: Hyper-Confidence</td>
</tr>
<tr>
<td>6. Conclusion</td>
</tr>
<tr>
<td>7. Appendix: The arules Infrastructure</td>
</tr>
</tbody>
</table>
Motivation

We live in the era of big data. Examples:

- **Retail data**: POS systems, loyalty cards, credit cards and e-commerce.
- **Web navigation data**: Web analytics, search engines, Wikis, etc.
- **Social media data**: Facebook, Google+, Instagram, LinkedIn, Pinterest, Snapchat, Tumblr, Twitter, etc.
- **Internet of Things**: Mobile phones, vehicles, home appliances, etc.
- **Biological data**: Electronic health records, gene expression data, etc.
Motivation

We live in the era of big data. Examples:

- **Retail data**: POS systems, loyalty cards, credit cards and e-commerce.
- **Web navigation data**: Web analytics, search engines, Wikis, etc.
- **Social media data**: Facebook, Google+, Instagram, LinkedIn, Pinterest, Snapchat, Tumblr, Twitter, etc.
- **Internet of Things**: Mobile phones, vehicles, home appliances, etc.
- **Biological data**: Electronic health records, gene expression data, etc.

Typical size of data sets:

- **Typical Retailer**: 10–500 product groups and 500–10,000 products
- **Amazon**: 560+ million products in the US (2018)
- **Wikipedia**: almost 5.7+ million articles (2018)
- **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.

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, digital 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).
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

1. Motivation

2. Transaction Data

3. Introduction to Association Rules

4. Probabilistic Interpretation, Weaknesses and Enhancements

5. A Probabilistic Independence Model
   - Application: Evaluate Quality Measures
   - Application: NB-Frequent Itemsets
   - Application: Hyper-Confidence

6. Conclusion

7. Appendix: The *arules* Infrastructure
Transaction Data

Example of market basket data:

<table>
<thead>
<tr>
<th>transaction ID</th>
<th>items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>milk, bread</td>
</tr>
<tr>
<td>2</td>
<td>bread, butter</td>
</tr>
<tr>
<td>3</td>
<td>beer</td>
</tr>
<tr>
<td>4</td>
<td>milk, bread, butter</td>
</tr>
<tr>
<td>5</td>
<td>bread, butter</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>transactions</th>
<th>milk</th>
<th>bread</th>
<th>butter</th>
<th>beer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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

1 Motivation

2 Transaction Data

3 Introduction to Association Rules

4 Probabilistic Interpretation, Weaknesses and Enhancements

5 A Probabilistic Independence Model
   - Application: Evaluate Quality Measures
   - Application: NB-Frequent Itemsets
   - Application: Hyper-Confidence

6 Conclusion

7 Appendix: The arules Infrastructure
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 $\text{supp}(Z) = \frac{n_Z}{n}$.
   → share of transactions in the database that contains $Z$.

2. **Confidence** of a rule $X \rightarrow Y$ is defined as
   $\text{conf}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$
   → 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 \( \text{supp}(Z) = \frac{n_Z}{n} \).
   \( \rightarrow \) share of transactions in the database that contains \( Z \).

2. **Confidence** of a rule \( X \rightarrow Y \) is defined as
   \[ \text{conf}(X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \]
   \( \rightarrow \) 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.

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>beer</th>
<th>eggs</th>
<th>flour</th>
<th>milk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

'Basis for efficient algorithms (Apriori, Eclat).'
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 \( \geq 2 \))

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>beer</th>
<th>eggs</th>
<th>flour</th>
<th>milk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

→ Basis for efficient algorithms (Apriori, Eclat).
Minimum Confidence

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

\begin{itemize}
  \item \{eggs\} \rightarrow \{flour\} \quad \text{Confidence} \quad \frac{3}{4} = 0.75
  \item \{flour\} \rightarrow \{eggs\} \quad \frac{3}{3} = 1
  \item \{eggs\} \rightarrow \{milk\} \quad \frac{2}{4} = 0.5
  \item \{milk\} \rightarrow \{eggs\} \quad \frac{2}{4} = 0.5
  \item \{flour\} \rightarrow \{milk\} \quad \frac{2}{3} = 0.67
  \item \{milk\} \rightarrow \{flour\} \quad \frac{2}{4} = 0.5
  \item \{eggs, flour\} \rightarrow \{milk\} \quad \frac{2}{3} = 0.67
  \item \{eggs, milk\} \rightarrow \{flour\} \quad \frac{2}{2} = 1
  \item \{flour, milk\} \rightarrow \{eggs\} \quad \frac{2}{2} = 1
  \item \{eggs\} \rightarrow \{flour, milk\} \quad \frac{2}{4} = 0.5
  \item \{flour\} \rightarrow \{eggs, milk\} \quad \frac{2}{3} = 0.67
  \item \{milk\} \rightarrow \{eggs, flour\} \quad \frac{2}{4} = 0.5
\end{itemize}
Minimum Confidence

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

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

<table>
<thead>
<tr>
<th>Left Itemset</th>
<th>Right Itemset</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>{eggs}</td>
<td>{flour}</td>
<td>3/5 = 0.6</td>
<td>3/4 = 0.75</td>
</tr>
<tr>
<td>{flour}</td>
<td>{eggs}</td>
<td>3/5 = 0.6</td>
<td>3/3 = 1</td>
</tr>
<tr>
<td>{eggs, milk}</td>
<td>{flour}</td>
<td>2/5 = 0.4</td>
<td>2/2 = 1</td>
</tr>
<tr>
<td>{flour, milk}</td>
<td>{eggs}</td>
<td>2/5 = 0.4</td>
<td>2/2 = 1</td>
</tr>
</tbody>
</table>
Table of Contents

1 Motivation

2 Transaction Data

3 Introduction to Association Rules

4 Probabilistic Interpretation, Weaknesses and Enhancements
   - A Probabilistic Independence Model
     - Application: Evaluate Quality Measures
     - Application: NB-Frequent Itemsets
     - Application: Hyper-Confidence

5 Conclusion

6 Appendix: The arules Infrastructure
Probabilistic interpretation of Support and Confidence

Support

\[ \text{supp}(Z) = \frac{n_Z}{n} \]

corresponds to an estimate for \( \hat{P}(E_Z) = \frac{n_Z}{n} \), the probability for the event that itemset \( Z \) is contained in a transaction.
Probabilistic interpretation of Support and Confidence

Support

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

corresponds to an estimate for $\hat{P}(E_Z) = \frac{n_Z}{n}$, 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):

<table>
<thead>
<tr>
<th>Support</th>
<th>Number of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>0.05</td>
<td>20</td>
</tr>
<tr>
<td>0.10</td>
<td>40</td>
</tr>
<tr>
<td>0.15</td>
<td>60</td>
</tr>
<tr>
<td>0.20</td>
<td>80</td>
</tr>
</tbody>
</table>

Weaknesses of Support and Confidence

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

<table>
<thead>
<tr>
<th></th>
<th>X=0</th>
<th>X=1</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y=0</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Y=1</td>
<td>70</td>
<td>20</td>
<td>90</td>
</tr>
<tr>
<td>Σ</td>
<td>75</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

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

**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).

<table>
<thead>
<tr>
<th></th>
<th>X=0</th>
<th>X=1</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y=0</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Y=1</td>
<td>70</td>
<td>20</td>
<td>90</td>
</tr>
<tr>
<td>Σ</td>
<td>75</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

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

**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.

→ What is the risk and cost attached to using spurious rules or missing important in an application?
The measure lift (interest, Brin et al., 1997) is defined as

\[
\text{lift}(X \rightarrow Y) = \frac{\text{conf}(X \rightarrow Y)}{\text{supp}(Y)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \cdot \text{supp}(Y)}
\]

and can be interpreted as an estimate for \( P(E_X \cap E_Y) / (P(E_X) \cdot P(E_Y)) \).

Measure for the deviation from stochastic independence:

\[
P(E_X \cap E_Y) = P(E_X) \cdot P(E_Y)
\]
The measure **lift** (interest, Brin *et al.*, 1997) is defined as

\[
\text{lift}(X \rightarrow Y) = \frac{\text{conf}(X \rightarrow Y)}{\text{supp}(Y)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \cdot \text{supp}(Y)}
\]

and can be interpreted as an estimate for \( P(E_X \cap E_Y)/(P(E_X) \cdot P(E_Y)) \).

→ 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:

- lift\((X \rightarrow Y) = 1 \ldots X \text{ and } Y \text{ are independent}
- lift\((X \rightarrow Y) > 1 \ldots \text{complementary effects between } X \text{ and } Y
- lift\((X \rightarrow Y) < 1 \ldots \text{substitution effects between } X \text{ and } Y
The measure \textit{lift} (interest, Brin \textit{et al.}, 1997) is defined as

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

and can be interpreted as an estimate for $P(E_X \cap E_Y)/(P(E_X) \cdot P(E_Y))$.

→ 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:

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

**Example**

<table>
<thead>
<tr>
<th></th>
<th>$X=0$</th>
<th>$X=1$</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y=0$</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>$Y=1$</td>
<td>70</td>
<td>20</td>
<td>90</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>75</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

\[
\text{lift}(X \rightarrow Y) = \frac{.2}{.25 \cdot .9} = .89
\]

**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 $X \rightarrow Y$.

<table>
<thead>
<tr>
<th></th>
<th>$X=0$</th>
<th>$X=1$</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y=0$</td>
<td>5</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>$Y=1$</td>
<td>70</td>
<td>20</td>
<td>90</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>75</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

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

$$X^2 = \sum \sum \frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})}$$

with $E(n_{ij}) = \frac{n_i \cdot n_j}{n}$

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

The result of the test for the contingency table above:

$X^2 = 3.7037$, df = 1, p-value = 0.05429

→ 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

1 Motivation

2 Transaction Data

3 Introduction to Association Rules

4 Probabilistic Interpretation, Weaknesses and Enhancements

5 A Probabilistic Independence Model
   • Application: Evaluate Quality Measures
   • Application: NB-Frequent Itemsets
   • Application: Hyper-Confidence

6 Conclusion

7 Appendix: The arules Infrastructure
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.

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

<table>
<thead>
<tr>
<th>p</th>
<th>i_1</th>
<th>i_2</th>
<th>i_3</th>
<th>...</th>
<th>i_k</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>0.0250</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 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 |

Michael Hahsler (IDA@SMU)
Table of Contents

1 Motivation

2 Transaction Data

3 Introduction to Association Rules

4 Probabilistic Interpretation, Weaknesses and Enhancements

5 A Probabilistic Independence Model
   • Application: Evaluate Quality Measures
   • Application: NB-Frequent Itemsets
   • Application: Hyper-Confidence

6 Conclusion

7 Appendix: The arules Infrastructure
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

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.

Retail data
Comparison: Confidence

Simulated data

\[
\text{Retail data}
\]

\[
\text{conf} \left( \{i_i\} \rightarrow \{i_j\} \right) = \frac{\text{supp}(\{i_i, i_j\})}{\text{supp}(\{i_i\})}
\]
Comparison: Lift

Simulated data

\[ \text{lift}(\{i_i\} \rightarrow \{i_j\}) = \frac{\text{supp}(\{i_i, i_j\})}{\text{supp}(\{i_i\}) \cdot \text{supp}(\{i_j\})} \]

Retail data
Comparison: Lift + Minimum Support

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

- Considerably higher lift values in retail data (indicate the existence of associations).

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

   ![Global model graph]

   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 known patterns.

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

1 Motivation
2 Transaction Data
3 Introduction to Association Rules
4 Probabilistic Interpretation, Weaknesses and Enhancements
5 A Probabilistic Independence Model
   - Application: Evaluate Quality Measures
   - Application: NB-Frequent Itemsets
   - Application: Hyper-Confidence
6 Conclusion
7 Appendix: The arules Infrastructure
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

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

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:

\[
\text{hyper-confidence}(X \rightarrow Y) \geq \gamma
\]

**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 \(n_Y\)).
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

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

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

<table>
<thead>
<tr>
<th></th>
<th>$X = 0$</th>
<th>$X = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = 0$</td>
<td>$n - n_Y - n_X - N_{XY}$</td>
<td>$n_X - N_{XY}$</td>
</tr>
<tr>
<td>$Y = 1$</td>
<td>$n_Y - N_{XY}$</td>
<td>$N_{XY}$</td>
</tr>
<tr>
<td></td>
<td>$n - n_X$</td>
<td>$n_X$</td>
</tr>
<tr>
<td></td>
<td>$n$</td>
<td></td>
</tr>
</tbody>
</table>
Hyper-Confidence: Complementary Effects

Simulated data

\[ \gamma = .99 \]

Retail data

Expected spurious rules: \( \alpha\left(\frac{k}{2}\right) = 141.98 \)
Hyper-Confidence: Complementary Effects

Simulated data

Retail data

\[ \gamma = .9999993 \]

Bonferroni correction

\[ \alpha = \frac{\alpha_i}{k} \]
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

1 Motivation

2 Transaction Data

3 Introduction to Association Rules

4 Probabilistic Interpretation, Weaknesses and Enhancements

5 A Probabilistic Independence Model
   - Application: Evaluate Quality Measures
   - Application: NB-Frequent Itemsets
   - Application: Hyper-Confidence

6 Conclusion

7 Appendix: The arules Infrastructure
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

1 Motivation

2 Transaction Data

3 Introduction to Association Rules

4 Probabilistic Interpretation, Weaknesses and Enhancements

5 A Probabilistic Independence Model
   - Application: Evaluate Quality Measures
   - Application: NB-Frequent Itemsets
   - Application: Hyper-Confidence

6 Conclusion

7 Appendix: The arules Infrastructure
**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))
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


