arules: Association Rule Mining with R

A Tutorial

Michael Hahsler

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

R User Group Dallas Meeting
February, 2015
Table of Contents

1 Motivation

2 Transaction Data

3 Introduction to Association Rules

4 Probabilistic Interpretation, Weaknesses and Enhancements

5 The arules Infrastructure

6 Life Demo
Motivation

We live 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
We live 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: 200+ million products (2013)
- 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

1. Motivation

2. Transaction Data

3. Introduction to Association Rules

4. Probabilistic Interpretation, Weaknesses and Enhancements

5. The arules Infrastructure

6. Life Demo
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>

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 The arules Infrastructure

6 Life Demo
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 $X$. 

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

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

→ called the support-confidence framework.
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 $X$.

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

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

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

→ 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>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</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>

Support count: 0

Episode: 'Frequent Itemsets'

- Basis for efficient algorithms (Apriori, Eclat).

---

Michael Hahsler (IDA@SMU)  R – Association Rules  RUG Dallas  10 / 25
**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.
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 $\geq 2$)

$\rightarrow$ 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.

- $(\{\text{eggs}\} \rightarrow \{\text{flour}\}) \quad \frac{3}{4} = 0.75$
- $(\{\text{flour}\} \rightarrow \{\text{eggs}\}) \quad \frac{3}{3} = 1$
- $(\{\text{eggs}\} \rightarrow \{\text{milk}\}) \quad \frac{2}{4} = 0.5$
- $(\{\text{milk}\} \rightarrow \{\text{eggs}\}) \quad \frac{2}{4} = 0.5$
- $(\{\text{flour}\} \rightarrow \{\text{milk}\}) \quad \frac{2}{3} = 0.67$
- $(\{\text{milk}\} \rightarrow \{\text{flour}\}) \quad \frac{2}{4} = 0.5$
- $(\{\text{eggs, flour}\} \rightarrow \{\text{milk}\}) \quad \frac{2}{3} = 0.67$
- $(\{\text{eggs, milk}\} \rightarrow \{\text{flour}\}) \quad \frac{2}{2} = 1$
- $(\{\text{flour, milk}\} \rightarrow \{\text{eggs}\}) \quad \frac{2}{2} = 1$
- $(\{\text{eggs}\} \rightarrow \{\text{flour, milk}\}) \quad \frac{2}{4} = 0.5$
- $(\{\text{flour}\} \rightarrow \{\text{eggs, milk}\}) \quad \frac{2}{3} = 0.67$
- $(\{\text{milk}\} \rightarrow \{\text{eggs, flour}\}) \quad \frac{2}{4} = 0.5$

'Frequent itemsets'
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>Rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>${\text{eggs}} \rightarrow {\text{flour}}$</td>
<td>$3/5 = 0.6$</td>
<td>$3/4 = 0.75$</td>
</tr>
<tr>
<td>${\text{flour}} \rightarrow {\text{eggs}}$</td>
<td>$3/5 = 0.6$</td>
<td>$3/3 = 1$</td>
</tr>
<tr>
<td>${\text{eggs, milk}} \rightarrow {\text{flour}}$</td>
<td>$2/5 = 0.4$</td>
<td>$2/2 = 1$</td>
</tr>
<tr>
<td>${\text{flour, milk}} \rightarrow {\text{eggs}}$</td>
<td>$2/5 = 0.4$</td>
<td>$2/2 = 1$</td>
</tr>
<tr>
<td>${\text{eggs, flour}} \rightarrow {\text{milk}}$</td>
<td>$2/3 = 0.67$</td>
<td>$2/3 = 0.67$</td>
</tr>
<tr>
<td>${\text{eggs, milk}} \rightarrow {\text{flour}}$</td>
<td>$2/4 = 0.5$</td>
<td>$2/2 = 1$</td>
</tr>
<tr>
<td>${\text{flour}} \rightarrow {\text{eggs, milk}}$</td>
<td>$2/4 = 0.5$</td>
<td>$2/3 = 0.67$</td>
</tr>
<tr>
<td>${\text{milk}} \rightarrow {\text{eggs, flour}}$</td>
<td>$2/4 = 0.5$</td>
<td>$2/4 = 0.5$</td>
</tr>
</tbody>
</table>
# Table of Contents

1. Motivation

2. Transaction Data

3. Introduction to Association Rules

4. Probabilistic Interpretation, Weaknesses and Enhancements

5. The `arules` Infrastructure

6. Life Demo
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):

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>$\Sigma$</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>

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:

\[
P(E_X \cap E_Y) = P(E_X) \cdot P(E_Y)
\]

In marketing values of lift are interpreted as:

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

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

**Problem:** 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>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>

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

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

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.
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))

<table>
<thead>
<tr>
<th>lhs</th>
<th>rhs</th>
<th>support</th>
<th>confidence</th>
<th>lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>{liquor, red/blush wine}</td>
<td>{bottled beer}</td>
<td>0.001931876</td>
<td>0.9047619</td>
<td>11.23527</td>
</tr>
<tr>
<td>{citrus fruit, other vegetables, soda, fruit}</td>
<td>{root vegetables}</td>
<td>0.001016777</td>
<td>0.9090909</td>
<td>8.34040</td>
</tr>
<tr>
<td>{tropical fruit, other vegetables, whole milk, yogurt, oil}</td>
<td>{root vegetables}</td>
<td>0.001016777</td>
<td>0.9090909</td>
<td>8.34040</td>
</tr>
</tbody>
</table>

Michael Hahsler (IDA@SMU)
R – Association Rules
RUG Dallas
1 Motivation

2 Transaction Data

3 Introduction to Association Rules

4 Probabilistic Interpretation, Weaknesses and Enhancements

5 The arules Infrastructure

6 Life Demo
Live Demo!

http://michael.hahsler.net/research/arules_RUG_2015/demo/
References


