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

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: 200+ million products (2013)
- 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 $\mathrm{supp}(Z) = n_Z/n$.
   $\rightarrow$ share of transactions in the database that contains Z.
2. Confidence of a rule $X \rightarrow Y$ is defined as
   $$\mathrm{conf}(X \rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$
   $\rightarrow$ 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 $\mathrm{supp}(Z) = n_Z/n$.
   $\rightarrow$ share of transactions in the database that contains Z.

2. Confidence of a rule $X \rightarrow Y$ is defined as
   $$\mathrm{conf}(X \rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$
   $\rightarrow$ share of transactions containing Y in all the transactions containing X.

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

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

$\rightarrow$ 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.

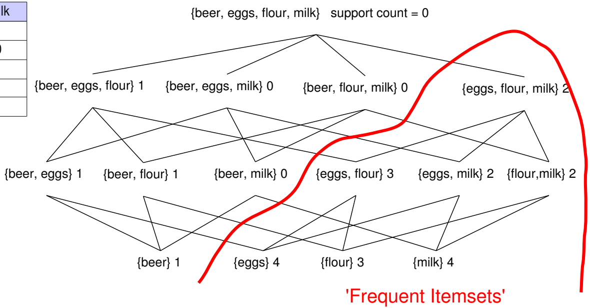
Minimum Support

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 ≥ 2)

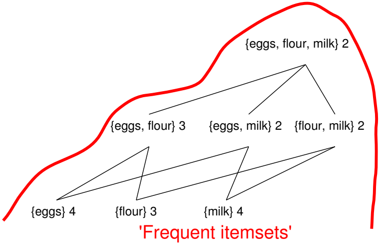
Transaction ID	beer	eggs	flour	milk
1	0	1	1	1
2	1	1	1	0
3	0	1	0	1
4	0	1	1	1
5	0	0	0	1



→ Basis for efficient algorithms (Apriori, Eclat).

Minimum Confidence

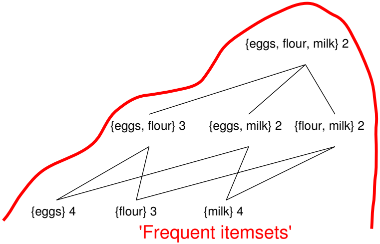
From the set of frequent itemsets all rules which satisfy the threshold for confidence $\mathrm{conf}(X \rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)} \geq \gamma$ are generated.



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

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

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.

Confidence can be interpreted as an estimate for the conditional probability

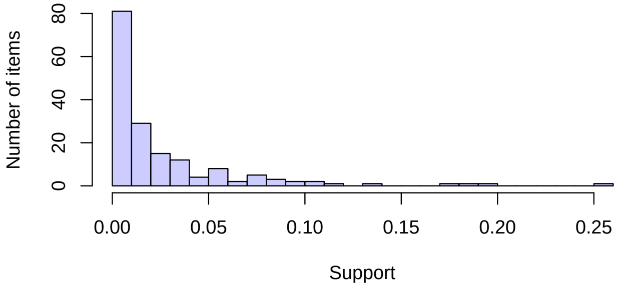
$$P(E_Y|E_X) = \frac{P(E_X \cap E_Y)}{P(E_X)}.$$

This directly follows the definition of confidence:

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

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

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

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

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

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

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

Example

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

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

Problem: small counts!

Chi-Square Test for Independence

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

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

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

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

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

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

The result of the test for the contingency table above:

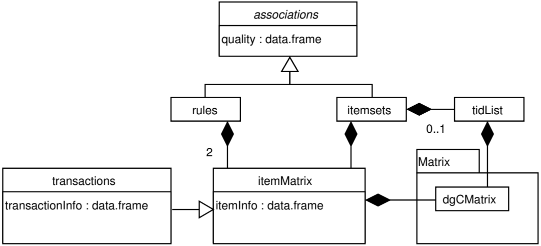
$X^2 = 3.7037, \mathrm{df} = 1, \text{p-value} = 0.05429$

$\to$ 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 **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
```

Table of Contents

Live Demo!

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

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.

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.

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.

References II

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.