Journal of Machine Learning Research 12 (2011) 2021-2025 Submitted 1/11; Published 6/11

The arules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Data Sets

Michael Hahsler MHAHSLER@LYLE.SMU.EDU
Sudheer Chelluboina SCHELLUBOI@LYLE.SMU.EDU
Department of Computer Science and Engineering
Southern Methodist University
Dallas, Texas 75275-0122, USA

Kurt Hornik KURT.HORNIK@WU.AC.AT
Department of Finance, Accounting and Statistics
Wirtschaftsuniversität Wien, Augasse 2-6, A-1090 Wien, Austria

Christian Buchta CHRISTIAN.BUCHTA@WU.AC.AT
Department of Cross-Border Business
Wirtschaftsuniversität Wien, Augasse 2-6, A-1090 Wien, Austria

Editor: Mikio Braun

Abstract

This paper describes the ecosystem of R add-on packages developed around the infrastructure provided by the package **arules**. The packages provide comprehensive functionality for analyzing interesting patterns including frequent itemsets, association rules, frequent sequences and for building applications like associative classification. After discussing the ecosystem's design we illustrate the ease of mining and visualizing rules with a short example.

Keywords: frequent itemsets, association rules, frequent sequences, visualization

1. Overview

Mining frequent itemsets and association rules is a popular and well researched method for discovering interesting relations between variables in large databases. Association rules are used in many applications and have become prominent as an important exploratory method for uncovering cross-selling opportunities in large retail databases.

Agrawal et al. (1993) introduced the problem of mining association rules from transaction data as follows:

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 a unique transaction ID and contains a subset of the items in I. A *rule* is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$ are called *itemsets*. On itemsets and rules several quality measures can be defined. The most important measures are support and confidence. The *support* $\text{supp}(X)$ of an itemset X is defined as the proportion of transactions in the data set which contain the itemset. Itemsets with a support which surpasses a user defined threshold σ are called *frequent itemsets*. The *confidence* of a rule is defined as $\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$. *Association rules* are rules with $\text{supp}(X \cup Y) \geq \sigma$ and $\text{conf}(X) \geq \delta$ where σ and δ are user defined thresholds.

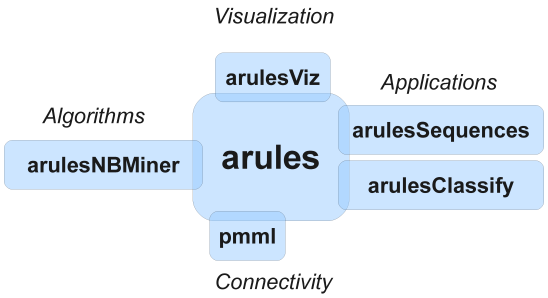


Figure 1: The **arules** ecosystem.

The R package **arules** (Hahsler et al., 2005, 2010) implements the basic infrastructure for creating and manipulating transaction databases and basic algorithms to efficiently find and analyze association rules. Over the last five years several packages were built around the **arules** infrastructure to create the ecosystem shown in Figure 1. Compared to other tools, the **arules** ecosystem is fully integrated, implements the latest approaches and has the vast functionality of R for further analysis of found patterns at its disposal.

2. Design and Implementation

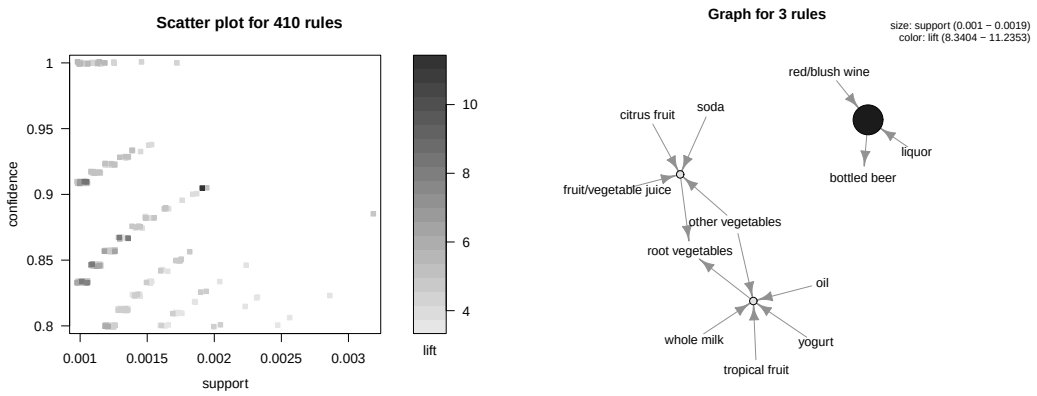
The core package **arules** provides an object-oriented framework to represent transaction databases and patterns. To facilitate extensibility, patterns are implemented as an abstract superclass *associations* and then concrete subclasses implement individual types of patterns. In **arules** the associations *itemsets* and *rules* are provided. Databases and associations both use a sparse matrix representation for efficient storage and basic operations like sorting, subsetting and matching are supported. Different aspects of arules were discussed in previous publications (Hahsler et al., 2005; Hahsler and Hornik, 2007b,a; Hahsler et al., 2008).

In this paper we focus on the ecosystem of several R-packages which are built on top of the arules infrastructure. While arules provides *Apriori* and *Eclat* (implementations by Borgelt, 2003), two of the most important frequent itemset/association rule mining algorithms, additional algorithms can easily be added as new packages. For example, package **arulesNBMiner** (Hahsler, 2010) implements an algorithm to find NB-frequent itemsets (Hahsler, 2006). A collection of further implementations which could be interfaced by arules in the future and a comparison of state-of-the-art algorithms can be found at the Frequent Itemset Mining Implementations Repository.[1]

arulesSequences (Buchta and Hahsler, 2010) implements mining frequent sequences in transaction databases. It implements additional association classes called *sequences* and *sequencerules* and provides the algorithm *cSpade* (Zaki, 2001) to efficiently mine frequent sequences. Another application currently under development is **arulesClassify** which uses the arules infrastructure to implement rule-based classifiers, including *Classification Based on Association rules* (CBA, Liu et al., 1998) and general associative classification techniques (Jalali-Heravi and Zaïane, 2010).

A known drawback of mining for frequent patterns such as association rules is that typically the algorithm returns a very large set of results where only a small fraction of patterns is of interest to the analysts. Many researchers introduced visualization techniques including scatter plots, matrix

1. The Frequent Itemset Mining Implementations Repository can be found at `http://fimi.ua.ac.be/`.



(a) (b)

Figure 2: Visualization of all 410 rules as (a) a scatter plot and (b) shows the top 3 rules according to lift as a graph.

visualizations, graphs, mosaic plots and parallel coordinates plots to analyze large sets of association rules (see Bruzzese and Davino, 2008, for a recent overview paper). **arulesViz** (Hahsler and Chelluboina, 2010) implements most of these methods for arules while also providing improvements using color shading, reordering and interactive features.

Finally, arules provides a *Predictive Model Markup Language (PMML)* interface to import and export rules via package **pmml** (Williams et al., 2010). PMML is the leading standard for exchanging statistical and data mining models and is supported by all major solution providers. Although **pmml** provides interfaces for different packages it is still considered part of the arules ecosystem.

The packages in the described ecosystem are available for Linux, OS X and Windows. All packages are distributed via the Comprehensive R Archive Network[2] under GPL-2, along with comprehensive manuals, documentation, regression tests and source code. Development versions of most packages are available from R-Forge.[3]

3. User Interface

We illustrate the user interface and the interaction between the packages in the **arules** ecosystem with a small example using a retail data set called *Groceries* which contains 9835 transactions with items aggregated to 169 categories. We mine association rules and then present the rules found as well as the top 3 rules according to the interest measure *lift* (deviation from independence) in two visualizations.

```
> library("arules")                    ### attach package 'arules'
> library("arulesViz")                 ### attach package 'arulesViz'
> data("Groceries")                    ### load data set
> ### mine association rules
```

2. The Comprehensive R Archive Network can be found at `http://CRAN.R-project.org`.
3. R-Forge can be found at `http://R-Forge.R-project.org`.

```
> rules <- apriori(Groceries, parameter = list(supp = 0.001, conf = 0.8))
> rules
set of 410 rules

> ### visualize rules as a scatter plot (with jitter to reduce occlusion)
> plot(rules, control=list(jitter=2))
> ### select and inspect rules with highest lift
> rules_high_lift <- head(sort(rules, by="lift"), 3)
> inspect(rules_high_lift)
  lhs                         rhs            support   confidence   lift
1 {liquor, red/blush wine}
                    => {bottled beer}    0.001931876  0.9047619 11.235269
2 {citrus fruit, other vegetables, soda, fruit/vegetable juice}
                    => {root vegetables} 0.001016777  0.9090909  8.340400
3 {tropical fruit, other vegetables, whole milk, yogurt, oil}
                    => {root vegetables} 0.001016777  0.9090909  8.340400

> ### plot selected rules as graph
> plot(rules_high_lift, method="graph", control=list(type="items"))
```

Figure 2 shows the visualizations produced by the example code. Both visualizations clearly show that there exists a rule (`{liquor, red/blush wine} => {bottled beer}`) with high support, confidence and lift. With the additionally available interactive features for the scatter plot and other available plots like the grouped matrix visualization, the rule set can be further explored.

References

Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216. ACM Press, 1993.

Christian Borgelt. Efficient implementations of Apriori and Eclat. In *FIMI'03: Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, November 2003.

Dario Bruzzese and Cristina Davino. Visual mining of association rules. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, pages 103–122. Springer-Verlag, 2008.

Christian Buchta and Michael Hahsler. *arulesSequences: Mining Frequent Sequences*, 2010. URL `http://CRAN.R-project.org/package=arulesSequences`. R package version 0.1-11.

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

Michael Hahsler. *arulesNBMiner: Mining NB-Frequent Itemsets and NB-Precise Rules*, 2010. URL `http://CRAN.R-project.org/package=arulesNBMiner`. R package version 0.1-1.

Michael Hahsler and Sudheer Chelluboina. *arulesViz: Visualizing Association Rules*, 2010. URL `http://CRAN.R-Project.org/package=arulesViz`. R package version 0.1-0.

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

Michael Hahsler and Kurt Hornik. Building on the arules infrastructure for analyzing transaction data with R. In R. Decker and H.-J. Lenz, editors, *Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8–10, 2006*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 449–456. Springer-Verlag, 2007b.

Michael Hahsler, Bettina Grün, and Kurt Hornik. arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1–25, October 2005.

Michael Hahsler, Christian Buchta, and Kurt Hornik. Selective association rule generation. *Computational Statistics*, 23(2):303–315, April 2008.

Michael Hahsler, Christian Buchta, Bettina Grün, and Kurt Hornik. *arules: Mining Association Rules and Frequent Itemsets*, 2010. URL `http://CRAN.R-project.org/package=arules`. R package version 1.0-3.

Mojdeh Jalali-Heravi and Osmar R. Zaïane. A study on interestingness measures for associative classifiers. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1039–1046. ACM, 2010.

Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In *Proceedings of the 4rd International Conference Knowledge Discovery and Data Mining (KDD-98)*, pages 80–86. AAAI Press, 1998.

Graham Williams, Michael Hahsler, Hemant Ishwaran, Udaya B. Kogalur, and Rajarshi Guha. *pmml: Generate PMML for various models*, 2010. URL `http://CRAN.R-project.org/package=pmml`. R package version 1.2.22.

Mohammed J. Zaki. SPADE: an efficient algorithm for mining frequent sequences. *Machine Learning*, 42:31–60, January–February 2001.