Introduction to arules – A computational environment for mining association rules and frequent item sets

Michael Hahsler  
Southern Methodist University

Bettina Grün  
Johannes Kepler Universität Linz

Kurt Hornik  
Wirtschaftsuniversität Wien

Christian Buchta  
Wirtschaftsuniversität Wien

Abstract

Mining frequent itemsets and association rules is a popular and well researched approach for discovering interesting relationships between variables in large databases. The R package arules presented in this paper provides a basic infrastructure for creating and manipulating input data sets and for analyzing the resulting itemsets and rules. The package also includes interfaces to two fast mining algorithms, the popular C implementations of Apriori and Eclat by Christian Borgelt. These algorithms can be used to mine frequent itemsets, maximal frequent itemsets, closed frequent itemsets and association rules.

Keywords: data mining, association rules, apriori, eclat.

1. Introduction

Mining frequent itemsets and association rules is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro (1991) describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal, Imielinski, and Swami (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 $D = \{t_1, t_2, \ldots, t_m\}$ 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 $I$. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short itemsets) $X$ and $Y$ are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule.

To illustrate the concepts, we use a small example from the supermarket domain. The set of items is $I = \{\text{milk, bread, butter, beer}\}$ and a small database containing the items is shown in Figure 1. An example rule for the supermarket could be $\{\text{milk, bread}\} \Rightarrow \{\text{butter}\}$ meaning that if milk and bread is bought, customers also buy butter.

To select interesting rules from the set of all possible rules, constraints on various measures of
significance and interest can be used. The best-known constraints are minimum thresholds on 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. In the example database in Figure 1, the itemset \{milk, bread\} has a support of $2/5 = 0.4$ since it occurs in 40% of all transactions (2 out of 5 transactions).

Finding frequent itemsets can be seen as a simplification of the unsupervised learning problem called “mode finding” or “bump hunting” (Hastie, Tibshirani, and Friedman 2001). For these problems each item is seen as a variable. The goal is to find prototype values so that the probability density evaluated at these values is sufficiently large. However, for practical applications with a large number of variables, probability estimation will be unreliable and computationally too expensive. This is why in practice frequent itemsets are used instead of probability estimation.

The confidence of a rule is defined $\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$. For example, the rule \{milk, bread\} $\Rightarrow$ \{butter\} has a confidence of $0.2/0.4 = 0.5$ in the database in Figure 1, which means that for 50% of the transactions containing milk and bread the rule is correct.

Confidence can be interpreted as an estimate of the probability $P(Y|X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS (see e.g., Hipp, Günther, and Nakhaeizadeh 2000).

Association rules are required to satisfy both a minimum support and a minimum confidence constraint at the same time. At medium to low support values, often a great number of frequent itemsets are found in a database. However, since the definition of support enforces that all subsets of a frequent itemset have to be also frequent, it is sufficient to only mine all maximal frequent itemsets, defined as frequent itemsets which are not proper subsets of any other frequent itemset (Zaki, Parthasarathy, Ogihara, and Li 1997b). Another approach to reduce the number of mined itemsets is to only mine frequent closed itemsets. An itemset is closed if no proper superset of the itemset is contained in each transaction in which the itemset is contained (Pasquier, Bastide, Taouil, and Lakhal 1999; Zaki 2004). Frequent closed itemsets are a superset of the maximal frequent itemsets. Their advantage over maximal frequent itemsets is that in addition to yielding all frequent itemsets, they also preserve the support information for all frequent itemsets which can be important for computing additional interest measures after the mining process is finished (e.g., confidence for rules generated from the found itemsets, or all-confidence (Omiecinski 2003)).

A practical solution to the problem of finding too many association rules satisfying the support and confidence constraints is to further filter or rank found rules using additional interest measures. A popular measure for this purpose is lift (Brin, Motwani, Ullman, and Tsur 1997). The lift of a rule is defined as $\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{(\text{supp}(X)\text{supp}(Y))}$, and

<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>

Figure 1: An example supermarket database with five transactions.
can be interpreted as the deviation of the support of the whole rule from the support expected under independence given the supports of the LHS and the RHS. Greater lift values indicate stronger associations.

In the last decade, research on algorithms to solve the frequent itemset problem has been abundant. Goethals and Zaki (2004) compare the currently fastest algorithms. Among these algorithms are the implementations of the Apriori and Eclat algorithms by Borgelt (2003) interfaced in the arules environment. The two algorithms use very different mining strategies. Apriori, developed by Agrawal and Srikant (1994), is a level-wise, breadth-first algorithm which counts transactions. In contrast, Eclat (Zaki et al. 1997b) employs equivalence classes, depth-first search and set intersection instead of counting. The algorithms can be used to mine frequent itemsets, maximal frequent itemsets and closed frequent itemsets. The implementation of Apriori can additionally be used to generate association rules.

This paper presents arules\footnote{The arules package can be obtained from \url{http://CRAN.R-Project.org} and the maintainer can be contacted at michael@hahsler.net.}, an extension package for R (R Development Core Team 2005) which provides the infrastructure needed to create and manipulate input data sets for the mining algorithms and for analyzing the resulting itemsets and rules. Since it is common to work with large sets of rules and itemsets, the package uses sparse matrix representations to minimize memory usage. The infrastructure provided by the package was also created to explicitly facilitate extensibility, both for interfacing new algorithms and for adding new types of interest measures and associations.

The rest of the paper is organized as follows: In the next section, we give an overview of the data structures implemented in the package arules. In Section 2 we introduce the functionality of the classes to handle transaction data and associations. In Section 3 we describe the way mining algorithms are interfaced in arules using the available interfaces to Apriori and Eclat as examples. In Section 4 we present some auxiliary methods for support counting, rule induction and sampling available in arules. We provide several examples in Section 5. The first two examples show typical R sessions for preparing, analyzing and manipulating a transaction data set, and for mining association rules. The third example demonstrates how arules can be extended to integrate a new interest measure. Finally, the fourth example shows how to use sampling in order to speed up the mining process. We conclude with a summary of the features and strengths of the package arules as a computational environment for mining association rules and frequent itemsets.

A previous version of this manuscript was published in the Journal of Statistical Software (Hahsler, Grün, and Hornik 2005a).

2. Data structure overview

To enable the user to represent and work with input and output data of association rule mining algorithms in R, a well-designed structure is necessary which can deal in an efficient way with large amounts of sparse binary data. The S4 class structure implemented in the package arules is presented in Figure 2.

For input data the classes transactions and tidLists (transaction ID lists, an alternative way to represent transaction data) are provided. The output of the mining algorithms comprises the classes itemsets and rules representing sets of itemsets or rules, respectively. Both classes
Introduction to arules

Figure 2: UML class diagram (see Fowler 2004) of the arules package.

directly extend a common virtual class called associations which provides a common interface. In this structure it is easy to add a new type of associations by adding a new class that extends associations.

Items in associations and transactions are implemented by the itemMatrix class which provides a facade for the sparse matrix implementation ngCMatrix from the R package Matrix (Bates and Maechler 2005).

To control the behavior of the mining algorithms, the two classes ASparameter and AScontrol are used. Since each algorithm can use additional algorithm-specific parameters, we implemented for each interfaced algorithm its own set of control classes. We used the prefix ‘AP’ for Apriori and ‘EC’ for Eclat. In this way, it is easy to extend the control classes when interfacing a new algorithm.

2.1. Representing collections of itemsets

From the definition of the association rule mining problem we see that transaction databases and sets of associations have in common that they contain sets of items (itemsets) together with additional information. For example, a transaction in the database contains a transaction ID and an itemset. A rule in a set of mined association rules contains two itemsets, one for the LHS and one for the RHS, and additional quality information, e.g., values for various interest measures.

Collections of itemsets used for transaction databases and sets of associations can be represented as binary incidence matrices with columns corresponding to the items and rows corresponding to the itemsets. The matrix entries represent the presence (1) or absence (0) of
Figure 3: Example of a collection of itemsets represented as a binary incidence matrix.

<table>
<thead>
<tr>
<th>itemsets</th>
<th>milk</th>
<th>bread</th>
<th>butter</th>
<th>beer</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$X_3$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$X_4$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 3: Example of a collection of itemsets represented as a binary incidence matrix.

an item in a particular itemset. An example of a binary incidence matrix containing itemsets for the example database in Figure 1 on Page 2 is shown in Figure 3. Note that we need to store collections of itemsets with possibly duplicated elements (identical rows), i.e., itemsets containing exactly the same items. This is necessary, since a transaction database can contain different transactions with the same items. Such a database is still a set of transactions since each transaction also contains a unique transaction ID.

Since a typical frequent itemset or a typical transaction (e.g., a supermarket transaction) only contains a small number of items compared to the total number of available items, the binary incidence matrix will in general be very sparse with many items and a very large number of rows. A natural representation for such data is a sparse matrix format. For our implementation we chose the \texttt{ngCMatrix} class defined in package \texttt{Matrix}. The \texttt{ngCMatrix} is a compressed, sparse, logical, column-oriented matrix which contains the indices of the \texttt{TRUE} rows and the pointers to the initial indices of elements in each column of the matrix. Despite the column orientation of the \texttt{ngCMatrix}, it is more convenient to work with incidence matrices which are row-oriented. This makes the most important manipulation, selecting a subset of transactions from a data set for mining, more comfortable and efficient. Therefore, we implemented the class \texttt{itemMatrix} providing a row-oriented facade to the \texttt{ngCMatrix} which stores a transposed incidence matrix\footnote{Note that the developers of package \texttt{Matrix} contains some support for a sparse row-oriented format, but the available functionality currently is rather minimal. Once all required functionality is implemented, \texttt{arules} will switch the internal representation to sparse row-oriented logical matrices.}. In sparse representation the following information needs to be stored for the collection of itemsets in Figure 3: A vector of indices of the non-zero elements (row-wise starting with the first row) 1, 2, 2, 4, 1, 2, 3, 3 and the pointers 1, 3, 5, 8 where each row starts in the index vector. The first two pointers indicate that the first row starts with element one in the index vector and ends with element 2 (since with element 3 already the next row starts). The first two elements of the index vector represent the items in the first row which are $i_1$ and $i_2$ or milk and bread, respectively. The two vectors are stored in the \texttt{ngCMatrix}. Note that indices for the \texttt{ngCMatrix} start with zero rather than with one and thus actually the vectors 0, 1, 1, 3, 0, 3, 3 and 0, 3, 4, 7 are stored. However, the data structure of the \texttt{ngCMatrix} class is not intended to be directly accessed by the end user of \texttt{arules}. The interfaces of \texttt{itemMatrix} can be used without knowledge of how the internal representation of the data works. However, if necessary, the \texttt{ngCMatrix} can be directly accessed by developers to add functionality to \texttt{arules} (e.g., to develop new types of associations or interest measures or to efficiently compute a distance matrix between itemsets for clustering). In this case, the \texttt{ngCMatrix} should be accessed using the coercion mechanism from \texttt{itemMatrix} to \texttt{ngCMatrix} via \texttt{as()}.
In addition to the sparse matrix, itemMatrix stores item labels (e.g., names of the items) and handles the necessary mapping between the item label and the corresponding column number in the incidence matrix. Optionally, itemMatrix can also store additional information on items. For example, the category hierarchy in a supermarket setting can be stored which enables the analyst to select only transactions (or as we later see also rules and itemsets) which contain items from a certain category (e.g., all dairy products).

For itemMatrix, basic matrix operations including `dim()` and subset selection (`[`) are available. The first element of `dim()` and `[` corresponds to itemsets or transactions (rows), the second element to items (columns). For example, on a transaction data set in variable `x` the subset selection `'x[1:10, 16:20]'` selects a matrix containing the first 10 transactions and items 16 to 20.

Since itemMatrix is used to represent sets or collections of itemsets additional functionality is provided. `length()` can be used to get the number of itemsets in an itemMatrix. Technically, `length()` returns the number of rows in the matrix which is equal to the first element returned by `dim()`. Identical itemsets can be found with `duplicated()`, and duplications can be removed with `unique()`. `match()` can be used to find matching elements in two collections of itemsets.

With `c()`, several itemMatrix objects can be combined by successively appending the rows of the objects, i.e., creating a collection of itemsets which contains the itemsets from all itemMatrix objects. This operation is only possible if the itemMatrix objects employed are “compatible,” i.e., if the matrices have the same number of columns and the items are in the same order. If two objects contain the same items (item labels), but the order in the matrix is different or one object is missing some items, `recode()` can be used to make them compatible by reordering and inserting columns.

To get the actual number of items in the itemsets stored in the itemMatrix, `size()` is used. It returns a vector with the number of items (ones) for each element in the set (row sum in the matrix). Obtaining the sizes from the sparse representations is a very efficient operation, since it can be calculated directly from the vector of column pointers in the ngCMatrix. For a purchase incidence matrix, `size()` will produce a vector as long as the number of transactions in the matrix with each element of the vector containing the number of items in the corresponding transaction. This information can be used, e.g., to select or filter unusually long or short transactions.

`itemFrequency()` calculates the frequency for each item in an itemMatrix. Conceptually, the item frequencies are the column sums of the binary matrix. Technically, column sums can be implemented for sparse representation efficiently by just tabulating the vector of row numbers of the non-zero elements in the ngCMatrix. Item frequencies can be used for many purposes. For example, they are needed to compute interest measures. `itemFrequency()` is also used by `itemFrequencyPlot()` to produce a bar plot of item count frequencies or support. Such a plot gives a quick overview of a set of itemsets and shows which are the most important items in terms of occurrence frequency.

Coercion from and to matrix and list primitives is provided where names and dimnames are used as item labels. For the coercion from itemMatrix to list there are two possibilities. The usual coercion via `as()` results in a list of vectors of character strings, each containing the item labels of the items in the corresponding row of the itemMatrix. The actual conversion is done by `LIST()` with its default behavior (argument `decode` set to `TRUE`). If in turn `LIST()`
is called with the argument \texttt{decode} set to \texttt{FALSE}, the result is a list of integer vectors with column numbers for items instead of the item labels. For many computations it is useful to work with such a list and later use the item column numbers to go back to the original \texttt{itemMatrix} for, e.g., subsetting columns. For subsequently decoding column numbers to item labels, \texttt{decode()} is also available.

Finally, \texttt{image()} can be used to produce a level plot of an \texttt{itemMatrix} which is useful for quick visual inspection. For transaction data sets (e.g., point-of-sale data) such a plot can be very helpful for checking whether the data set contains structural changes (e.g., items were not offered or out-of-stock during part of the observation period) or to find abnormal transactions (e.g., transactions which contain almost all items may point to recording problems). Spotting such problems in the data can be very helpful for data preparation.

\subsection*{2.2. Transaction data}

The main application of association rules is for market basket analysis where large transaction data sets are mined. In this setting each transaction contains the items which were purchased at one visit to a retail store (see e.g., \cite{Berry1997}). Transaction data are normally recorded by point-of-sale scanners and often consists of tuples of the form:

\begin{verbatim}
< transaction ID, item ID, ... >
\end{verbatim}

All tuples with the same transaction ID form a single transaction which contains all the items given by the item IDs in the tuples. Additional information denoted by the ellipsis dots might be available. For example, a customer ID might be provided via a loyalty program in a supermarket. Further information on transactions (e.g., time, location), on the items (e.g., category, price), or on the customers (socio-demographic variables such as age, gender, etc.) might also be available.

For mining, the transaction data is first transformed into a binary purchase incidence matrix with columns corresponding to the different items and rows corresponding to transactions. The matrix entries represent the presence (1) or absence (0) of an item in a particular transaction. This format is often called the \textit{horizontal} database layout (\cite{Zaki2000}). Alternatively, transaction data can be represented in a \textit{vertical} database layout in the form of transaction
ID lists (Zaki 2000). In this format for each item a list of IDs of the transactions the item is contained in is stored. In Figure 4 the example database in Figure 1 on Page 2 is depicted in horizontal and vertical layouts. Depending on the algorithm, one of the layouts is used for mining. In arules both layouts are implemented as the classes transactions and tidLists. Similar to transactions, class tidLists also uses a sparse representation to store its lists efficiently. Objects of classes transactions and tidLists can be directly converted into each other by coercion.

The class transactions directly extends itemMatrix and inherits its basic functionality (e.g., subset selection, getting itemset sizes, plotting item frequencies). In addition, transactions has a slot to store further information for each transaction in form of a data.frame. The slot can hold arbitrary named vectors with length equal to the number of stored transactions. In arules the slot is currently used to store transaction IDs, however, it can also be used to store user IDs, revenue or profit, or other information on each transaction. With this information subsets of transactions (e.g., only transactions of a certain user or exceeding a specified profit level) can be selected.

Objects of class transactions can be easily created by coercion from matrix or list. If names or dimnames are available in these data structures, they are used as item labels or transaction IDs, accordingly. To import data from a file, the read.transactions() function is provided. This function reads files structured as shown in Figure 4 and also the very common format with one line per transaction and the items separated by a predefined character. Finally, inspect() can be used to inspect transactions (e.g., “interesting” transactions obtained with subset selection).

Another important application of mining association rules has been proposed by Piatetsky-Shapiro (1991) and Srikant and Agrawal (1996) for discovering interesting relationships between the values of categorical and quantitative (metric) attributes. For mining associations rules, non-binary attributes have to be mapped to binary attributes. The straightforward mapping method is to transform the metric attributes into \( k \) ordinal attributes by building categories (e.g., an attribute income might be transformed into an ordinal attribute with the three categories: “low”, “medium” and “high”). Then, in a second step, each categorical attribute with \( k \) categories is represented by \( k \) binary dummy attributes which correspond to the items used for mining. An example application using questionnaire data can be found in Hastie et al. (2001) in the chapter about association rule mining.

The typical representation for data with categorical and quantitative attributes in R is a data.frame. First, a domain expert has to create useful categories for all metric attributes. This task is supported in arules by functions such as discretize() which implements equal interval length, equal frequency, clustering-based and custom interval discretization. After discretization all columns in the data.frame need to be either logical or factors. The second step, the generation of binary dummy items, is automated in package arules by coercing from data.frame to transactions. In this process, the original attribute names and categories are preserved as additional item information and can be used to select itemsets or rules which contain items referring to a certain original attributes. By default it is assumed that missing values do not carry information and thus all of the corresponding dummy items are set to zero. If the fact that the value of a specific attribute is missing provides information (e.g., a respondent in an interview refuses to answer a specific question), the domain expert can create for the attribute a category for missing values which then will be included in the transactions as its own dummy item.
The resulting transactions object can be mined and analyzed the same way as market basket data, see the example in Section 5.1.

2.3. Associations: itemsets and sets of rules

The result of mining transaction data in arules are associations. Conceptually, associations are sets of objects describing the relationship between some items (e.g., as an itemset or a rule) which have assigned values for different measures of quality. Such measures can be measures of significance (e.g., support), or measures of interestingness (e.g., confidence, lift), or other measures (e.g., revenue covered by the association).

All types of association have a common functionality in arules comprising the following methods:

- **summary()** to give a short overview of the set and **inspect()** to display individual associations,
- **length()** for getting the number of elements in the set,
- **items()** for getting for each association a set of items involved in the association (e.g., the union of the items in the LHS and the RHS for each rule),
- sorting the set using the values of different quality measures (**sort()**),
- subset extraction (**[]** and **subset()**),
- set operations (**union()**, **intersect()**, and **setequal()**), and
- matching elements from two sets (**match()**),
- **write()** for writing associations to disk in human readable form. To save and load associations in compact form, use **save()** and **load()** from the base package.
- **write.pmml()** and **read.pmml()** can be used to write and read associations using PMML (Predictive Model Markup Language) via package pmml (Williams 2008).

The associations currently implemented in package arules are sets of itemsets (e.g., used for frequent itemsets of their closed or maximal subset) and sets of rules (e.g., association rules). Both classes, itemsets and rules, directly extend the virtual class associations and provide the functionality described above.

Class itemsets contains one itemMatrix object to store the items as a binary matrix where each row in the matrix represents an itemset. In addition, it may contain transaction ID lists as an object of class tidLists. Note that when representing transactions, tidLists store for each item a transaction list, but here store for each itemset a list of transaction IDs in which the itemset appears. Such lists are currently only returned by eclat().

Class rules consists of two itemMatrix objects representing the left-hand-side (LHS) and the right-hand-side (RHS) of the rules, respectively.

The items in the associations and the quality measures can be accessed and manipulated in a safe way using accessor and replace methods for items, lhs, rhs, and quality. In addition
the association classes have built-in validity checking which ensures that all elements have compatible dimensions.

It is simple to add new quality measures to existing associations. Since the quality slot holds a data.frame, additional columns with new quality measures can be added. These new measures can then be used to sort or select associations using sort() or subset(). Adding a new type of associations to arules is straightforward as well. To do so, a developer has to create a new class extending the virtual associations class and implement the common functionality described above.

3. Mining algorithm interfaces

In package arules we interface free reference implementations of Apriori and Eclat by Christian Borgelt (Borgelt and Kruse 2002; Borgelt 2003). The code is called directly from R by the functions apriori() and eclat() and the data objects are directly passed from R to the C code and back without writing to external files. The implementations can mine frequent itemsets, and closed and maximal frequent itemsets. In addition, apriori() can also mine association rules.

The data given to the apriori() and eclat() functions have to be transactions or something which can be coerced to transactions (e.g., matrix or list). The algorithm parameters are divided into two groups represented by the arguments parameter and control. The mining parameters (parameter) change the characteristics of the mined itemsets or rules (e.g., the minimum support) and the control parameters (control) influence the performance of the algorithm (e.g., enable or disable initial sorting of the items with respect to their frequency). These arguments have to be instances of the classes APparameter and APcontrol for the function apriori() or ECparameter and ECcontrol for the function eclat(), respectively. Alternatively, data which can be coerced to these classes (e.g., NULL which will give the default values or a named list with names equal to slot names to change the default values) can be passed. In these classes, each slot specifies a different parameter and the values. The default values are equal to the defaults of the stand-alone C programs (Borgelt 2004) except that the standard definition of the support of a rule (Agrawal et al. 1993) is employed for the specified minimum support required (Borgelt defines the support of a rule as the support of its antecedent).

For apriori() the appearance feature implemented by Christian Borgelt can also be used. With argument appearance of function apriori() one can specify which items have to or must not appear in itemsets or rules. For more information on this feature we refer to the Apriori manual (Borgelt 2004).

The output of the functions apriori() and eclat() is an object of a class extending associations which contains the sets of mined associations and can be further analyzed using the functionality provided for these classes.

There exist many different algorithms which which use an incidence matrix or transaction ID list representation as input and solve the frequent and closed frequent itemset problems. Each algorithm has specific strengths which can be important for very large databases. Such algorithms, e.g. kDCI, LCM, FP-Growth or Patricia, are discussed in Goethals and Zaki (2003). The source code of most algorithms is available on the internet and, if a special

---

3Christian Borgelt provides the source code for Apriori and Eclat at http://www.borgelt.net/fpm.html.
algorithm is needed, interfacing the algorithms for \texttt{arules} is straightforward. The necessary steps are:

1. Adding interface code to the algorithm, preferably by directly calling into the native implementation language (rather than using files for communication), and an \texttt{R} function calling this interface.

2. Implementing extensions for \texttt{ASparameter} and \texttt{AScontrol}.

4. Auxiliary functions

In \texttt{arules} several helpful functions are implemented for support counting, rule induction, sampling, etc. In the following we will discuss some of these functions.

4.1. Counting support for itemsets

Normally, itemset support is counted during mining the database with a given minimum support constraint. During this process all frequent itemsets plus some infrequent candidate itemsets are counted (or support is determined by other means). Especially for databases with many items and for low minimum support values, this procedure can be extremely time consuming since in the worst case, the number of frequent itemsets grows exponentially in the number of items.

If only the support information for a single or a few itemsets is needed, we might not want to mine the database for all frequent itemsets. We also do not know in advance how high (or low) to set the minimum support to still get the support information for the itemsets in question. For this problem, \texttt{arules} contains \texttt{support()} which determines the support for a set of given sets of items (as an \texttt{itemMatrix}).

For counting, we use a prefix tree (Knuth 1997) to organize the counters. The used prefix tree is similar to the itemset tree described by Borgelt and Kruse (2002). However, we do not generate the tree level-wise, but we first generate a prefix tree which only contains the nodes necessary to hold the counters for all itemsets which need to be counted. Using the nodes in this tree only, we count the itemsets for each transaction recursively. After counting, the support for each itemset is contained in the node with the prefix equal to the itemset. The exact procedure is described in Hahsler, Buchta, and Hornik (2008).

In addition to determining the support of a few itemsets without mining all frequent itemsets, \texttt{support()} is also useful for finding the support of infrequent itemsets with a support so low that mining is infeasible due to combinatorial explosion.

4.2. Rule induction

For convenience we introduce $\mathcal{X} = \{X_1, X_2, \ldots, X_l\}$ for sets of itemsets with length $l$. Analogously, we write $\mathcal{R}$ for sets of rules. A part of the association rule mining problem is the generation (or induction) of a set of rules $\mathcal{R}$ from a set of frequent itemsets $\mathcal{X}$. The implementation of the Apriori algorithm used in \texttt{arules} already contains a rule induction engine and by default returns the set of association rules of the form $X \Rightarrow Y$ which satisfy given
minimum support and minimum confidence. Following the definition of Agrawal et al. (1993) Y is restricted to single items.

In some cases it is necessary to separate mining itemsets and generating rules from itemsets. For example, only rules stemming from a subset of all frequent itemsets might be of interest to the user. The Apriori implementation efficiently generates rules by reusing the data structures built during mining the frequent itemsets. However, if Apriori is used to return only itemsets or Eclat or some other algorithm is used to mine itemsets, the data structure needed for rule induction is no longer available for computing rule confidence.

If rules need to be induced from an arbitrary set of itemsets, support values required to calculate confidence are typically missing. For example, if all available information is an itemset containing five items and we want to induce rules, we need the support of the itemset (which we might know), but also the support of all subsets of length four. The missing support information has to be counted from the database. Finally, to induce rules efficiently for a given set of itemsets, we also have to store support values in a suitable data structure which allows fast look-ups for calculating rule confidence.

Function ruleInduction() provided in arules uses a prefix tree to induce rules for a given confidence from an arbitrary set of itemsets \( X \) in the following way:

1. Count the support values for each itemset \( X \in X \) and the subsets \( \{X \setminus \{x\} : x \in X\} \) needed for rule generation in a single pass over the database and store them in a suitable data structure.

2. Populate set \( R \) by selectively generating only rules for the itemsets in \( X \) using the support information from the data structure created in step 1.

Efficient support counting is done as described in Section 4.1 above. After counting, all necessary support counts are contained in the prefix tree. We can retrieve the needed support values and generating the rules is straight forward. The exact procedure is described in Hahsler et al. (2008).

4.3. Sampling from transactions

Taking samples from large databases for mining is a powerful technique which is especially useful if the original database does not fit into main memory, but the sample does. However, even if the database fits into main memory, sampling can provide an enormous speed-up for mining at the cost of only little degradation of accuracy.

Mannila, Toivonen, and Verkamo (1994) proposed sampling with replacement for association rule mining and quantify the estimation error due to sampling. Using Chernov bounds on the binomial distribution (the number of transactions which contains a given itemset in a sample), the authors argue that in theory even relatively small samples should provide good estimates for support.

Zaki, Parthasarathy, Li, and Ogiwara (1997a) built upon the theoretic work by Mannila et al. (1994) and show that for an itemset \( X \) with support \( \tau = \text{supp}(X) \) and for an acceptable relative error of support \( \epsilon \) (an accuracy of \( 1 - \epsilon \)) at a given confidence level \( 1 - c \), the needed sample size \( n \) can be computed by
\( n = \frac{-2\ln(c)}{\tau \epsilon^2} \)

Depending on its support, for each itemset a different sample size is appropriate. As a heuristic, the authors suggest to use the user specified minimum support threshold for \( \tau \). This means that for itemsets close to minimum support, the given error and confidence level hold while for more frequent itemsets the error rate will be less. However, with this heuristic the error rate for itemsets below minimum support can exceed \( \epsilon \) at the given confidence level and thus some infrequent itemsets might appear as frequent ones in the sample.

Zaki et al. (1997a) also evaluated sampling in practice on several data sets and conclude that sampling not only speeds mining up considerably, but also the errors are considerably smaller than those given by the Chernov bounds and thus samples with size smaller than obtained by Equation 1 are often sufficient.

Another way to obtain the required sample size for association rule mining is progressive sampling (Parthasarathy 2002). This approach starts with a small sample and uses progressively larger samples until model accuracy does not improve significantly anymore. Parthasarathy (2002) defines a proxy for model accuracy improvement by using a similarity measure between two sets of associations. The idea is that since larger samples will produce more accurate results, the similarity between two sets of associations of two consecutive samples is low if accuracy improvements are high and increases with decreasing accuracy improvements. Thus increasing sample size can be stopped if the similarity between consecutive samples reaches a “plateau.”

Toivonen (1996) presents an application of sampling to reduce the needed I/O overhead for very large databases which do not fit into main memory. The idea is to use a random sample from the data base to mine frequent itemsets at a support threshold below the set minimum support. The support of these itemsets is then counted in the whole database and the infrequent itemsets are discarded. If the support threshold to mine the sample is picked low enough, almost all frequent itemsets and their support will be found in one pass over the large database.

In arules sampling is implemented by `sample()` which provides all capabilities of the standard sampling function in R (e.g., sampling with or without replacement and probability weights).

### 4.4. Generating synthetic transaction data

Synthetic data can be used to evaluate and compare different mining algorithms and to study the behavior of measures of interestingness.

In arules the function `random.transactions()` can be used to create synthetic transaction data. Currently there are two methods available. The first method reimplements the well known generator for transaction data for mining association rules developed by Agrawal and Srikant (1994). The second method implements a simple probabilistic model where each transaction is the result of one independent Bernoulli trial for each item (see Hahsler, Hornik, and Reutterer 2005b)).

### 4.5. Sub-, super-, maximal and closed itemsets

For some calculations it is necessary to find all sub- or supersets for a specific itemset in a
set of itemsets. This functionality is implemented as `is.subset()` and `is.superset()`. For example, `is.subset(x, y, proper = TRUE)`, finds all proper subsets of the itemsets in `x` in the set `y`. The result is a logical matrix with `length(x)` rows and `length(y)` columns. Each logical row vector represents which elements in `y` are subsets of the corresponding element in `x`. If `y` is omitted, the sub- or superset structure within the set `x` is returned.

Similar methods, `is.maximal()` and `is.closed()`, can be used to find all maximal itemsets or closed itemsets in a set. An itemset is maximal in a set if no proper superset of the itemset is contained in the set (Zaki et al. 1997b). An itemset is closed, if it is its own closure (i.e., for an items no superset with the same support exits) (Pasquier et al. 1999).

Note that these methods can be extremely slow and have high memory usage if the set contains many itemsets.

4.6. Additional measures of interestingness

`arules` provides `interestMeasure()` which can be used to calculate a broad variety of interest measures for itemsets and rules. To speed up the calculation, we try to reuse the quality information available from the sets of itemsets or rules (i.e., support, confidence, lift) and, only if necessary, missing information is obtained from the transactions used to mine the associations.

For example, available measures for itemsets are:

- All-confidence (Omiecinski 2003)
- Cross-support ratio (Xiong, Tan, and Kumar 2003)
- Support

For rules the following measures are implemented:

- Chi square measure (Liu, Hsu, and Ma 1999)
- Conviction (Brin et al. 1997)
- Confidence
- Difference of Confidence (DOC) (Hofmann and Wilhelm 2001)
- Hyper-lift and hyper-confidence (Hahsler and Hornik 2007b)
- Leverage (Piatetsky-Shapiro 1991)
- Lift
- Improvement (Bayardo, Agrawal, and Gunopulos 2000)
- Support
- Several measures from Tan, Kumar, and Srivastava (2004) (e.g., cosine, Gini index, φ-coefficient, odds ratio)
4.7. Distance based clustering transactions and associations

To allow for distance based clustering (Gupta, Strehl, and Ghosh 1999), *arules* provides `dissimilarity()` which can be used to calculate dissimilarities and cross-dissimilarities between transactions or associations (i.e., itemsets and rules). Currently, the following standard measures for binary data are available: Jaccard coefficient, simple matching coefficient and dice coefficient. Additionally, dissimilarity between transactions can be calculated based on affinities between items (Aggarwal, Procopiuc, and Yu 2002).

The result of `dissimilarity()` is either a `dist` object, which can be directly used by most clustering methods in R (e.g., `hclust` for hierarchical clustering), or an object of class `ar_cross_dissimilarity`.

Since the number of transactions or associations is often too large to efficiently calculate a dissimilarity matrix and apply a clustering algorithm, `sample()` can be used to cluster only a subset of transactions (associations). To assign the remaining transactions (associations) to clusters, `predict()` implements the nearest neighbor approach for predicting memberships for new data.

A small example can be found in Hahsler and Hornik (2007a).

5. Examples

5.1. Example 1: Analyzing and preparing a transaction data set

In this example, we show how a data set can be analyzed and manipulated before associations are mined. This is important for finding problems in the data set which could make the mined associations useless or at least inferior to associations mined on a properly prepared data set. For the example, we look at the Epub transaction data contained in package *arules*. This data set contains downloads of documents from the Electronic Publication platform of the Vienna University of Economics and Business available via [http://epub.wu-wien.ac.at](http://epub.wu-wien.ac.at) from January 2003 to December 2008.

First, we load *arules* and the data set.

```r
> library("arules")
> data("Epub")
> Epub
```

transactions in sparse format with
15729 transactions (rows) and
936 items (columns)

We see that the data set consists of 15729 transactions and is represented as a sparse matrix with 15729 rows and 936 columns which represent the items. Next, we use the `summary()` to get more information about the data set.

```r
> summary(Epub)
```
transactions as itemMatrix in sparse format with
15729 rows (elements/itemsets/transactions) and
936 columns (items) and a density of 0.001758755

most frequent items:
doc_11d doc_813 doc_4c6 doc_955 doc_698 (Other)
356 329 288 282 245 24393

element (itemset/transaction) length distribution:
sizes
1 2 3 4 5 6 7 8 9 10 11 12
11615 2189 854 409 198 121 93 409 198 121 93 409
13 14 15 16 17 18 19 20 21 22 23 24
10 10 6 8 6 5 8 2 8 2 3 2 3
25 26 27 28 30 34 36 38 41 43 52 58
4 5 1 1 1 2 1 2 1 1 1 1

Min. 1st Qu. Median Mean 3rd Qu. Max.
1.000 1.000 1.000 1.646 2.000 58.000

includes extended item information - examples:
  labels
1 doc_11d
2 doc_13d
3 doc_14c

includes extended transaction information - examples:
  transactionID TimeStamp
10792 session_4795 2003-01-01 19:59:00
10793 session_4797 2003-01-02 06:46:01
10794 session_479a 2003-01-02 09:50:38

summary() displays the most frequent items in the data set, information about the transaction
length distribution and that the data set contains some extended transaction information.
We see that the data set contains transaction IDs and in addition time stamps (using class
POSIXct) for the transactions. This additional information can be used for analyzing the data
set.

> year <- strftime(as.POSIXlt(transactionInfo(Epub)[["TimeStamp"]]), ",%Y")
> table(year)

year
2003 2004 2005 2006 2007 2008
987 1375 1611 3015 4050 4691

For 2003, the first year in the data set, we have 987 transactions. We can select the corre-
sponding transactions and inspect the structure using a level-plot (see Figure 5).
The plot is a direct visualization of the binary incidence matrix where the dark dots represent the ones in the matrix. From the plot we see that the items in the data set are not evenly distributed. In fact, the mostly white area to the right side suggests, that in the beginning of 2003 only very few items were available (less than 50) and then during the year more items were added until it reached a number of around 300 items. Also, we can see that there are some transactions in the data set which contain a very high number of items (denser horizontal lines). These transactions need further investigation since they could originate from data collection problems (e.g., a web robot downloading many documents from the publication site). To find the very long transactions we can use the `size()` and select very long transactions (containing more than 20 items).

```r
> transactionInfo(Epub2003[size(Epub2003) > 20])

<table>
<thead>
<tr>
<th>transactionID</th>
<th>TimeStamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>11092</td>
<td>session_56e2 2003-04-29 12:30:38</td>
</tr>
<tr>
<td>11371</td>
<td>session_6308 2003-08-17 17:16:12</td>
</tr>
</tbody>
</table>
```
We found three long transactions and printed the corresponding transaction information. Of course, size can be used in a similar fashion to remove long or short transactions. Transactions can be inspected using `inspect()`. Since the long transactions identified above would result in a very long printout, we will inspect the first 5 transactions in the subset for 2003.

```r
> inspect(Epub2003[1:5])
```

<table>
<thead>
<tr>
<th>items</th>
<th>transactionID</th>
<th>TimeStamp</th>
</tr>
</thead>
</table>

Most transactions contain one item. Only transaction 4 contains three items. For further inspection transactions can be converted into a list with:

```r
> as(Epub2003[1:5], "list")
```

```r
$session_4795
[1] "doc_154"

$session_4797
[1] "doc_3d6"

$session_479a
[1] "doc_16f"

$session_47b7
[1] "doc_11d" "doc_1a7" "doc_f4"

$session_47bb
[1] "doc_83"
```

Finally, transaction data in horizontal layout can be converted to transaction ID lists in vertical layout using coercion.

```r
> EpubTidLists <- as(Epub, "tidLists")
> EpubTidLists
```

tidLists in sparse format with
936 items/itemsets (rows) and
15729 transactions (columns)

For performance reasons the transaction ID list is also stored in a sparse matrix. To get a list, coercion to list can be used.
$doc_11d

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&quot;session_47b7&quot;</td>
<td>&quot;session_47c2&quot;</td>
</tr>
<tr>
<td>4</td>
<td>&quot;session_4855&quot;</td>
<td>&quot;session_488d&quot;</td>
</tr>
<tr>
<td>7</td>
<td>&quot;session_489b&quot;</td>
<td>&quot;session_489c&quot;</td>
</tr>
<tr>
<td>10</td>
<td>&quot;session_4897&quot;</td>
<td>&quot;session_48a0&quot;</td>
</tr>
<tr>
<td>13</td>
<td>&quot;session_48a5&quot;</td>
<td>&quot;session_489a&quot;</td>
</tr>
<tr>
<td>16</td>
<td>&quot;session_48aa&quot;</td>
<td>&quot;session_48f0&quot;</td>
</tr>
<tr>
<td>19</td>
<td>&quot;session_4b35&quot;</td>
<td>&quot;session_4bac&quot;</td>
</tr>
<tr>
<td>22</td>
<td>&quot;session_4c9a&quot;</td>
<td>&quot;session_4d8c&quot;</td>
</tr>
<tr>
<td>25</td>
<td>&quot;session_4e89&quot;</td>
<td>&quot;session_5071&quot;</td>
</tr>
<tr>
<td>28</td>
<td>&quot;session_51e6&quot;</td>
<td>&quot;session_5227&quot;</td>
</tr>
<tr>
<td>31</td>
<td>&quot;session_5265&quot;</td>
<td>&quot;session_52e0&quot;</td>
</tr>
<tr>
<td>34</td>
<td>&quot;session_53e1&quot;</td>
<td>&quot;session_5522&quot;</td>
</tr>
<tr>
<td>37</td>
<td>&quot;session_558b&quot;</td>
<td>&quot;session_5714&quot;</td>
</tr>
<tr>
<td>40</td>
<td>&quot;session_57c5&quot;</td>
<td>&quot;session_5813&quot;</td>
</tr>
<tr>
<td>43</td>
<td>&quot;session_wu4852&quot;</td>
<td>&quot;session_5955&quot;</td>
</tr>
<tr>
<td>46</td>
<td>&quot;session_5aaa&quot;</td>
<td>&quot;session_5acd&quot;</td>
</tr>
<tr>
<td>49</td>
<td>&quot;session_5bfc&quot;</td>
<td>&quot;session_5f3d&quot;</td>
</tr>
<tr>
<td>52</td>
<td>&quot;session_5f69&quot;</td>
<td>&quot;session_5fcf&quot;</td>
</tr>
<tr>
<td>55</td>
<td>&quot;session_6053&quot;</td>
<td>&quot;session_6081&quot;</td>
</tr>
<tr>
<td>58</td>
<td>&quot;session_635b&quot;</td>
<td>&quot;session_64b4&quot;</td>
</tr>
<tr>
<td>61</td>
<td>&quot;session_65d2&quot;</td>
<td>&quot;session_67d1&quot;</td>
</tr>
<tr>
<td>64</td>
<td>&quot;session_68c4&quot;</td>
<td>&quot;session_68f8&quot;</td>
</tr>
<tr>
<td>67</td>
<td>&quot;session_6c95&quot;</td>
<td>&quot;session_6e19&quot;</td>
</tr>
<tr>
<td>70</td>
<td>&quot;session_6f88&quot;</td>
<td>&quot;session_718e&quot;</td>
</tr>
<tr>
<td>73</td>
<td>&quot;session_72d6&quot;</td>
<td>&quot;session_7303&quot;</td>
</tr>
<tr>
<td>76</td>
<td>&quot;session_782d&quot;</td>
<td>&quot;session_7856&quot;</td>
</tr>
<tr>
<td>79</td>
<td>&quot;session_799b&quot;</td>
<td>&quot;session_7b24&quot;</td>
</tr>
<tr>
<td>82</td>
<td>&quot;session_7cfd&quot;</td>
<td>&quot;session_7d5d&quot;</td>
</tr>
<tr>
<td>85</td>
<td>&quot;session_819b&quot;</td>
<td>&quot;session_8329&quot;</td>
</tr>
<tr>
<td>88</td>
<td>&quot;session_84d7&quot;</td>
<td>&quot;session_85b0&quot;</td>
</tr>
<tr>
<td>91</td>
<td>&quot;session_867f&quot;</td>
<td>&quot;session_8688&quot;</td>
</tr>
<tr>
<td>94</td>
<td>&quot;session_88fe&quot;</td>
<td>&quot;session_8730&quot;</td>
</tr>
<tr>
<td>97</td>
<td>&quot;session_87a9&quot;</td>
<td>&quot;session_880a&quot;</td>
</tr>
<tr>
<td>100</td>
<td>&quot;session_88b0&quot;</td>
<td>&quot;session_8986&quot;</td>
</tr>
<tr>
<td>103</td>
<td>&quot;session_8a73&quot;</td>
<td>&quot;session_8a87&quot;</td>
</tr>
<tr>
<td>106</td>
<td>&quot;session_8ae2&quot;</td>
<td>&quot;session_8db4&quot;</td>
</tr>
<tr>
<td>109</td>
<td>&quot;session_wu53a2&quot;</td>
<td>&quot;session_8fad&quot;</td>
</tr>
<tr>
<td>112</td>
<td>&quot;session_9083&quot;</td>
<td>&quot;session_90d8&quot;</td>
</tr>
<tr>
<td>115</td>
<td>&quot;session_9145&quot;</td>
<td>&quot;session_916e&quot;</td>
</tr>
<tr>
<td>118</td>
<td>&quot;session_919e&quot;</td>
<td>&quot;session_91df&quot;</td>
</tr>
<tr>
<td>121</td>
<td>&quot;session_9333&quot;</td>
<td>&quot;session_9376&quot;</td>
</tr>
<tr>
<td>124</td>
<td>&quot;session_94d5&quot;</td>
<td>&quot;session_9539&quot;</td>
</tr>
<tr>
<td>127</td>
<td>&quot;session_96a0&quot;</td>
<td>&quot;session_9745&quot;</td>
</tr>
<tr>
<td>130</td>
<td>&quot;session_985b&quot;</td>
<td>&quot;session_9873&quot;</td>
</tr>
</tbody>
</table>
Introduction to arules
[274] "session_1517f" "session_151b2" "session_15549"
[277] "session_155a9" "session_1571b" "session_15b18"
[280] "session_15b99" "session_15d2c" "session_15e0c"
[283] "session_15f75" "session_15fbf" "session_16621"
[286] "session_16691" "session_16f0d" "session_17027"
[289] "session_173fe" "session_17eaf" "session_17ecd"
[292] "session_180dd" "session_18641" "session_187ae"
[295] "session_18a0b" "session_18b18" "session_18db4"
[298] "session_19048" "session_19051" "session_19510"
[301] "session_19788" "session_197ee" "session_19c04"
[304] "session_19c7a" "session_19f0c" "session_1a557"
[307] "session_1ac3c" "session_1b733" "session_1b76a"
[310] "session_1b76b" "session_1ba83" "session_1c0a6"
[313] "session_1c1c1" "session_1c304" "session_1c43c"
[316] "session_1cea1" "session_1cfb9" "session_1db2a"
[319] "session_1db96" "session_1d6ea" "session_1dc94"
[322] "session_1e361" "session_1e36e" "session_1e91e"
[325] "session_wu6bf8f" "session_1f3a8" "session_1f56c"
[328] "session_1f61e" "session_1f6831" "session_1fced"
[331] "session_wu6c9e5" "session_1fuba7" "session_20074"
[334] "session_2019f" "session_201a1" "session_209f9"
[337] "session_20e87" "session_20f5b" "session_212a2"
[340] "session_2143b" "session_214def" "session_218ca"
[343] "session_21bea" "session_21fbd" "session_223e1"
[346] "session_2248d" "session_22ae6" "session_232d4"
[349] "session_23636" "session_23912" "session_23a70"
[352] "session_23b6d" "session_23c17" "session_240ea"
[355] "session_24256" "session_24484"

$doc_13d
[1] "session_4809" "session_5dbc" "session_8e0b" "session_cf4b"
[5] "session_d92a" "session_102bb" "session_10e9f" "session_11344"
[9] "session_11ca4" "session_12dc9" "session_155b5" "session_1b563"
[13] "session_1c411" "session_1f384" "session_22e97"

$doc_14c
[1] "session_53fb" "session_564b" "session_5697" "session_56e2"
[5] "session_630b" "session_680c" "session_6f7c" "session_7c8a"
[9] "session_8903" "session_890c" "session_89d2" "session_907e"
[13] "session_98b4" "session_c268" "session_c302" "session_cb86"
"session_d70a" "session_d854" "session_e4c7" "session_f220"
[21] "session_fd57" "session_fe31" "session_10278" "session_115b0"
[25] "session_11fb9" "session_11e26" "session_12185" "session_1414b"
[29] "session_14daa" "session_14e47" "session_15738" "session_15a38"
[33] "session_16305" "session_17b35" "session_19a1f" "session_1d074"
[37] "session_1fccc4" "session_2272e" "session_23a3e"
In this representation each item has an entry which is a vector of all transactions it occurs in. \texttt{tidLists} can be directly used as input for mining algorithms which use such a vertical database layout to mine associations.

In the next example, we will see how a data set is created and rules are mined.

5.2. Example 2: Preparing and mining a questionnaire data set

As a second example, we prepare and mine questionnaire data. We use the Adult data set from the UCI machine learning repository (Asuncion and Newman 2007) provided by package \texttt{arules}. This data set is similar to the marketing data set used by Hastie \textit{et al.} (2001) in their chapter about association rule mining. The data originates from the U.S. census bureau database and contains 48842 instances with 14 attributes like age, work class, education, etc. In the original applications of the data, the attributes were used to predict the income level of individuals. We added the attribute \texttt{income} with levels \texttt{small} and \texttt{large}, representing an income of $\leq$ USD 50,000 and $>$ USD 50,000, respectively. This data is included in \texttt{arules} as the data set \texttt{AdultUCI}.

```r
> data("AdultUCI")
> dim(AdultUCI)
[1] 48842   15
> AdultUCI[1:2,]

age workclass fnlwgt education education-num marital-status
1 39 State-gov 77516 Bachelors 13 Never-married
2 50 Self-emp-not-inc 83311 Bachelors 13 Married-civ-spouse
   occupation relationship race sex capital-gain capital-loss
1 Adm-clerical Not-in-family White Male 2174 0
2 Exec-managerial Husband White Male 0 0
   hours-per-week native-country income
1  40 United-States small
2  13 United-States small
```

\texttt{AdultUCI} contains a mixture of categorical and metric attributes and needs some preparations before it can be transformed into transaction data suitable for association mining. First, we remove the two attributes \texttt{fnlwgt} and \texttt{education-num}. The first attribute is a weight calculated by the creators of the data set from control data provided by the Population Division of the U.S. census bureau. The second removed attribute is just a numeric representation of the attribute \texttt{education} which is also part of the data set.

```r
> AdultUCI["fnlwgt"] <- NULL
> AdultUCI["education-num"] <- NULL
```

Next, we need to map the four remaining metric attributes (\texttt{age}, \texttt{hours-per-week}, \texttt{capital-gain} and \texttt{capital-loss}) to ordinal attributes by building suitable categories. We divide the attributes \texttt{age} and \texttt{hours-per-week} into suitable categories using knowledge about typical age
Michael Hahsler, Bettina Grün, Kurt Hornik, Christian Buchta

groups and working hours. For the two capital related attributes, we create a category called
\texttt{None} for cases which have no gains/losses. Then we further divide the group with gains/losses
at their median into the two categories \texttt{Low} and \texttt{High}.

\begin{verbatim}
> AdultUCI[["age"]]<- ordered(cut(AdultUCI[["age"]], c(15,25,45,65,100)),
+    labels = c("Young", "Middle-aged", "Senior", "Old"))
> AdultUCI[["hours-per-week"]]<- ordered(cut(AdultUCI[["hours-per-week"]],
+    c(0,25,40,60,168)),
+    labels = c("Part-time", "Full-time", "Over-time", "Workaholic"))
> AdultUCI[["capital-gain"]]<- ordered(cut(AdultUCI[["capital-gain"]],
+    c(-Inf,0,median(AdultUCI[["capital-gain"]][AdultUCI[["capital-gain"]]>0]),Inf)),
+    labels = c("None", "Low", "High"))
> AdultUCI[["capital-loss"]]<- ordered(cut(AdultUCI[["capital-loss"]],
+    c(-Inf,0,
+    median(AdultUCI[["capital-loss"]][AdultUCI[["capital-loss"]]>0]),Inf)),
+    labels = c("none", "low", "high"))
\end{verbatim}

Now, the data can be automatically recoded as a binary incidence matrix by coercing the
data set to \texttt{transactions}.

\begin{verbatim}
> Adult <- as(AdultUCI, "transactions")
> Adult

transactions in sparse format with
48842 transactions (rows) and
115 items (columns)

The remaining 115 categorical attributes were automatically recoded into 115 binary items. During encoding the item labels were generated in the form of \texttt{<variable name>=<category label>}. Note that for cases with missing values all items corresponding to the attributes with the missing values were set to zero.

\begin{verbatim}
> summary(Adult)

transactions as itemMatrix in sparse format with
48842 rows (elements/itemsets/transactions) and
115 columns (items) and a density of 0.1089939

most frequent items:
\begin{verbatim}
capital-loss=none   capital-gain=\texttt{None}  
46560   44807
native-country=United-States  race=White  
43832   41762
workclass=Private  (Other)  
33906   401333
\end{verbatim}
element (itemset/transaction) length distribution:
sizes
9 10 11 12 13
19 971 2067 15623 30162

Min. 1st Qu. Median Mean 3rd Qu. Max.
9.00 12.00 13.00 12.53 13.00 13.00

includes extended item information - examples:

<table>
<thead>
<tr>
<th>labels</th>
<th>variables</th>
<th>levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>age=Young</td>
<td>Young</td>
</tr>
<tr>
<td>2</td>
<td>age=Middle-aged</td>
<td>Middle-aged</td>
</tr>
<tr>
<td>3</td>
<td>age=Senior</td>
<td>Senior</td>
</tr>
</tbody>
</table>

includes extended transaction information - examples:

<table>
<thead>
<tr>
<th>transactionID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

The summary of the transaction data set gives a rough overview showing the most frequent items, the length distribution of the transactions and the extended item information which shows which variable and which value were used to create each binary item. In the first example we see that the item with label age=Middle-aged was generated by variable age and level middle-aged.

To see which items are important in the data set we can use the itemFrequencyPlot(). To reduce the number of items, we only plot the item frequency for items with a support greater than 10% (using the parameter support). For better readability of the labels, we reduce the label size with the parameter cex.names. The plot is shown in Figure 6.

```r
> itemFrequencyPlot(Adult, support = 0.1, cex.names=0.8)
```

Next, we call the function apriori() to find all rules (the default association type for apriori()) with a minimum support of 1% and a confidence of 0.6.

```r
> rules <- apriori(Adult, 
+ parameter = list(support = 0.01, confidence = 0.6))
```

Apriori

Parameter specification:

```
  confidence minval smax arem aval originalSupport maxtime support minlen
     0.6     0.1   1 none FALSE          TRUE      5  0.01    1
maxlen target   ext
     10    rules FALSE
```
Algorithmic control:
filter tree heap memopt load sort verbose
0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 488

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[115 item(s), 48842 transaction(s)] done [0.05s].
sorting and recoding items ... [67 item(s)] done [0.02s].
creating transaction tree ... done [0.05s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [1.61s].
writing ... [276443 rule(s)] done [0.06s].
creating S4 object ... done [0.17s].

> rules

set of 276443 rules

First, the function prints the used parameters. Apart from the specified minimum support and minimum confidence, all parameters have the default values. It is important to note that with parameter maxlen, the maximum size of mined frequent itemsets, is by default restricted to 5. Longer association rules are only mined if maxlen is set to a higher value. After the parameter settings, the output of the C implementation of the algorithm with timing information is displayed.
The result of the mining algorithm is a set of 276443 rules. For an overview of the mined rules, `summary()` can be used. It shows the number of rules, the most frequent items contained in the left-hand-side and the right-hand-side and their respective length distributions and summary statistics for the quality measures returned by the mining algorithm.

```r
> summary(rules)
```

**set of 276443 rules**

**rule length distribution (lhs + rhs):sizes**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
<td>432</td>
<td>4981</td>
<td>22127</td>
<td>52669</td>
<td>75104</td>
<td>67198</td>
<td>38094</td>
<td>13244</td>
<td>2588</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>5.000</td>
<td>6.000</td>
<td>6.289</td>
<td>7.000</td>
<td>10.000</td>
</tr>
</tbody>
</table>

**summary of quality measures:**

<table>
<thead>
<tr>
<th>support</th>
<th>confidence</th>
<th>lift</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. :0.01001</td>
<td>Min. :0.6000</td>
<td>Min. :0.7171</td>
<td>Min. : 489</td>
</tr>
<tr>
<td>1st Qu.:0.01253</td>
<td>1st Qu.:0.7691</td>
<td>1st Qu.: 1.0100</td>
<td>1st Qu.: 612</td>
</tr>
<tr>
<td>Median :0.01701</td>
<td>Median :0.9051</td>
<td>Median : 1.0554</td>
<td>Median : 831</td>
</tr>
<tr>
<td>Mean :0.02679</td>
<td>Mean :0.8600</td>
<td>Mean : 1.3109</td>
<td>Mean : 1308</td>
</tr>
<tr>
<td>3rd Qu.:0.02741</td>
<td>3rd Qu.:0.9542</td>
<td>3rd Qu.: 1.2980</td>
<td>3rd Qu.: 1339</td>
</tr>
<tr>
<td>Max. :0.95328</td>
<td>Max. :1.0000</td>
<td>Max. :20.6826</td>
<td>Max. :46560</td>
</tr>
</tbody>
</table>

**mining info:**

- data
  - ntransactions
  - support
  - confidence

```
Adult 48842 0.01 0.6
```

As typical for association rule mining, the number of rules found is huge. To analyze these rules, for example, `subset()` can be used to produce separate subsets of rules for each item which resulted form the variable `income` in the right-hand-side of the rule. At the same time we require that the `lift` measure exceeds 1.2.

```r
> rulesIncomeSmall <- subset(rules, subset = rhs %in% "income=small" & lift > 1.2)
> rulesIncomeLarge <- subset(rules, subset = rhs %in% "income=large" & lift > 1.2)
```

We now have a set with rules for persons with a small income and a set for persons with a large income. For comparison, we inspect for both sets the three rules with the highest confidence (using `head()`).

```r
> inspect(head(rulesIncomeSmall, n = 3, by = "confidence"))
```

<table>
<thead>
<tr>
<th>lhs</th>
<th>rhs</th>
<th>support</th>
<th>confidence</th>
<th>lift</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>{workclass=Private, marital-status=Never-married,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
relationship=Own-child,
sex=Male,
hours-per-week=Part-time,
native-country=United-States} => {income=small} 0.01074895 0.7104195 1.403653 525

[2] {workclass=Private,
marital-status=Never-married,
relationship=Own-child,
sex=Male,
hours-per-week=Part-time} => {income=small} 0.01144507 0.7102922 1.403402 559

[3] {workclass=Private,
marital-status=Never-married,
relationship=Own-child,
sex=Male,
capital-gain=None,
hours-per-week=Part-time,
native-country=United-States} => {income=small} 0.01046231 0.7097222 1.402276 511

> inspect(head(rulesIncomeLarge, n = 3, by = "confidence"))

lhs rhs support confidence lift
c
[1] {marital-status=Married-civ-spouse,
capital-gain=High,
native-country=United-States} => {income=large} 0.01562180 0.6849192 4.266398

[2] {marital-status=Married-civ-spouse,
capital-gain=High,
capital-loss=none,
native-country=United-States} => {income=large} 0.01562180 0.6849192 4.266398

[3] {relationship=Husband,
race=White,
capital-gain=High,
native-country=United-States} => {income=large} 0.01302158 0.6846071 4.264454

From the rules we see that workers in the private sector working part-time or in the service industry tend to have a small income while persons with high capital gain who are born in the US tend to have a large income. This example shows that using subset selection and sorting a set of mined associations can be analyzed even if it is huge.

Finally, the found rules can be written to disk to be shared with other applications. To save rules in plain text format the function `write()` is used. The following command saves a set of rules as the file named `data.csv` in comma separated value (CSV) format.

> write(rulesIncomeSmall, file = "data.csv", sep = ",", col.names = NA)

Alternatively, with package `pmml` (Williams 2008) the rules can be saved in PMML (Predictive Modelling Markup Language), a standardized XML-based representation used by many data mining tools. Note that `pmml` requires the package `XML` which might not be available for all operating systems.
> write.PMML(rulesIncomeSmall, file = "data.xml")

The saved data can now be easily shared and used by other applications. Itemsets (with `write()` also transactions) can be written to a file in the same way.

5.3. Example 3: Extending arules with a new interest measure

In this example, we show how easy it is to add a new interest measure, using \textit{all-confidence} as introduced by Omiecinski (2003). The all-confidence of an itemset $X$ is defined as

$$
\text{all-confidence}(X) = \frac{\text{supp}(X)}{\max_{I \subset X} \text{supp}(I)}
$$

(2)

This measure has the property $\text{conf}(I \Rightarrow X \setminus I) \geq \text{all-confidence}(X)$ for all $I \subset X$. This means that all possible rules generated from itemset $X$ must at least have a confidence given by the itemset’s all-confidence value. Omiecinski (2003) shows that the support in the denominator of equation 2 must stem from a single item and thus can be simplified to $\max_{i \in X} \text{supp}\{\{i\}\}$.

To obtain an itemset to calculate all-confidence for, we mine frequent itemsets from the previously used Adult data set using the Eclat algorithm.

> data("Adult")
> fsets <- eclat(Adult, parameter = list(support = 0.05), + control = list(verbose=FALSE))

For the denominator of all-confidence we need to find all mined single items and their corresponding support values. In the following we create a named vector where the names are the column numbers of the items and the values are their support.

> singleItems <- fsets[[size(items(fsets)) == 1]]
> ## Get the col numbers we have support for
> singleSupport <- quality(singleItems)$support
> names(singleSupport) <- unlist(LIST(items(singleItems), + decode = FALSE))
> head(singleSupport, n = 5)

66 63 111 60 8
0.9532779 0.9173867 0.8974243 0.8550428 0.6941976

Next, we can calculate the all-confidence using Equation 2 for all itemsets. The single item support needed for the denomination is looked up from the named vector `singleSupport` and the resulting measure is added to the set’s quality data frame.

> itemsetList <- LIST(items(fsets), decode = FALSE)
> allConfidence <- quality(fsets)$support / + sapply(itemsetList, function(x) + max(singleSupport[as.character(x)]))
> quality(fsets) <- cbind(quality(fsets), allConfidence)
The new quality measure is now part of the set of itemsets.

```r
> summary(fsets)

set of 8496 itemsets

most frequent items:
capital-loss=None native-country=United-States  
  4082 3973
  capital-gain=None race=White  
    3962 3781
  workclass=Private (Other)  
       3142 21931

element (itemset/transaction) length distribution:

                      sizes
length   1   2  3   4   5   6   7  8   9  10
36 303 1078 2103 2388 1689 706 171 21  1

Min. 1st Qu. Median Mean 3rd Qu. Max.
1.000 4.000 5.000 4.811 6.000 10.000

summary of quality measures:

  support  count   allConfidence
Min. :0.05002 Min. : 2443 Min. :0.05247
1st Qu.:0.06038 1st Qu.: 2949 1st Qu.:0.06597
Median :0.07546 Median : 3686 Median :0.08428
3rd Qu.:0.11279 3rd Qu.: 5509 3rd Qu.:0.12711
Max. :0.95328 Max. :46560 Max. :1.00000

includes transaction ID lists: FALSE

mining info:
  data ntransactions support
Adult     48842   0.05

It can be used to manipulate the set. For example, we can look at the itemsets which contain an item related to education (using partial match with `%pin%`) and sort them by all-confidence (we filter itemsets of length 1 first, since they have per definition an all-confidence of 1).

```r
> fsetsEducation <- subset(fsets, subset = items %pin% "education")
> inspect(sort(fsetsEducation[size(fsetsEducation)>1],
+       by = "allConfidence")[1 : 3])

                   items                  support  count
[1] {education=HS-grad,hours-per-week=Full-time}  0.2090209 10209
[2] {education=HS-grad,income=small}             0.1807051  8826
The resulting itemsets show that the item high school graduate (but no higher education) is highly associated with working full-time, a small income and working in the private sector. All-confidence is along with many other measures of interestingness already implemented in arules as the function `interestMeasure()`.

### 5.4. Example 4: Sampling

In this example, we show how sampling can be used in arules. We use again the Adult data set.

```r
> data("Adult")
> Adult
transactions in sparse format with
48842 transactions (rows) and
115 items (columns)

To calculate a reasonable sample size $n$, we use the formula developed by Zaki et al. (1997a) and presented in Section 4.3. We choose a minimum support of 5%. As an acceptable error rate for support $\epsilon$ we choose 10% and as the confidence level $(1 - c)$ we choose 90%.

```r
> supp <- 0.05
> epsilon <- 0.1
> c <- 0.1
> n <- -2 * log(c)/ (supp * epsilon^2)
> n
[1] 9210.34
```

The resulting sample size is considerably smaller than the size of the original database. With `sample()` we produce a sample of size $n$ with replacement from the database.

```r
> AdultSample <- sample(Adult, n, replace = TRUE)
```

The sample can be compared with the database (the population) using an item frequency plot. The item frequencies in the sample are displayed as bars and the item frequencies in the original database are represented by the line. For better readability of the labels, we only display frequent items in the plot and reduce the label size with the parameter `cex.names`. The plot is shown in Figure 7.
Figure 7: Item frequencies in a sample of the Adult data set (bars) compared to the complete data set (line).

> itemFrequencyPlot(AdultSample, population = Adult, support = supp, 
  + cex.names = 0.7)

Alternatively, a sample can be compared with the population using the lift ratio (with lift = TRUE). The lift ratio for each item \( i \) is \( P(i | \text{sample}) / P(i | \text{population}) \), where the probabilities are estimated by the item frequencies. A lift ratio of one indicates that the items occur in the sample in the same proportion as in the population. A lift ratio greater than one indicates that the item is over-represented in the sample and vice versa. With this plot, large relative deviations for less frequent items can be identified visually (see Figure 8).

> itemFrequencyPlot(AdultSample, population = Adult, 
  + support = supp, lift = TRUE, 
  + cex.names = 0.9)

To compare the speed-up reached by sampling we use the Eclat algorithm to mine frequent itemsets on both, the database and the sample and compare the system time (in seconds) used for mining.

> time <- system.time(itemsets <- eclat(Adult, 
  + parameter = list(support = supp), control = list(verbose = FALSE)))
> time

    user  system elapsed
  0.336  0.000  0.336
Introduction to arules

Figure 8: Deviations of the item frequencies in the sample from the complete Adult data set.

```r
> timeSample <- system.time(itemsetsSample <- eclat(AdultSample,
+   parameter = list(support = supp), control = list(VERBOSE = FALSE)))
> timeSample
    user  system elapsed
   0.060   0.004   0.064

The first element of the vector returned by `system.time()` gives the (user) CPU time needed
for the execution of the statement in its argument. Therefore, mining the sample instead of
the whole data base results in a speed-up factor of:

```r
> # speed up
> time[1] / timeSample[1]

user.self
 5.6
```

To evaluate the accuracy for the itemsets mined from the sample, we analyze the difference
between the two sets.

```r
> itemsets

set of 8496 itemsets
```
set of 8566 itemsets

The two sets have roughly the same size. To check if the sets contain similar itemsets, we match the sets and see what fraction of frequent itemsets found in the database were also found in the sample.

```r
> match <- match(itemsets, itemsetsSample, nomatch = 0)
> # remove no matches
> sum(match > 0) / length(itemsets)

[1] 0.9735169
```

Almost all frequent itemsets were found using the sample. The summaries of the support of the frequent itemsets which were not found in the sample and the itemsets which were frequent in the sample although they were infrequent in the database, respectively, give:

```r
> summary(quality(itemsets[match == 0])$support)

       Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     
0.05004 0.05047 0.05098 0.05128 0.05174 0.05604

> summary(quality(itemsetsSample[-match])$support)

       Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     
0.05005 0.05049 0.05114 0.05152 0.05223 0.05624
```

This shows that only itemsets with support very close to the minimum support were falsely missed or found.

For the frequent itemsets which were found in the database and in the sample, we can calculate accuracy from the the error rate.

```r
> supportItemsets <- quality(itemsets[which(match > 0)])$support
> supportSample <- quality(itemsetsSample[match])$support
> accuracy <- 1 - abs(supportSample - supportItemsets) / supportItemsets
> summary(accuracy)

       Min. 1st Qu.    Median    Mean 3rd Qu.    Max.     
0.8502 0.9642 0.9795 0.9742 0.9904 1.0000
```

The summary shows that sampling resulted in finding the support of itemsets with high accuracy. This small example illustrates that for extremely large databases or for application where mining time is important, sampling can be a powerful technique.
6. Summary and outlook

With package **arules** we provide the basic infrastructure which enables us to mine associations and analyze and manipulate the results. Previously, in R there was no such infrastructure available. The main features of **arules** are:

- Efficient implementation using sparse matrices.
- Simple and intuitive interface to manipulate and analyze transaction data, sets of itemsets and rules with subset selection and sorting.
- Interface to two fast mining algorithms.
- Flexibility in terms of adding new quality measures, and additional item and transaction descriptions which can be used for selecting transactions and analyzing resulting associations.
- Extensible data structure to allow for easy implementation of new types of associations and interfacing new algorithms.

There are several interesting possibilities to extend **arules**. For example, it would be very useful to interface algorithms which use statistical measures to find “interesting” itemsets (which are not necessarily frequent itemsets as used in an association rule context). Such algorithms include implementations of the $\chi^2$-test based algorithm by Silverstein, Brin, and Motwani (1998) or the baseline frequency approach by DuMouchel and Pregibon (2001). Also, the implementation of distance based clustering in **arules** could be used for visualization of associations (see e.g., Strehl and Ghosh 2003).

Acknowledgments

Part of **arules** was developed during the project “Statistical Computing with R” funded by of the “Jubiläumsstiftung der WU Wien.” The authors of **arules** would like to thank Christian Borgelt for the implementation of Apriori and Eclat.

References


Introduction to arules


**Affiliation:**

Michael Hahsler
Engineering Management, Information, and Systems
Lyle School of Engineering
Southern Methodist University
P.O. Box 750122
Dallas, TX 75275-0122
E-mail: mhahsler@lyle.smu.edu
URL: http://lyle.smu.edu/~mhahsler