Visualizing Association Rules in Hierarchical Groups

Michael Hahsler and Sudheer Chelluboina

Intelligent Data Analysis Group, Southern Methodist University

Interface 2011: Statistical, Machine Learning, and Visualization Algorithms
42th Symposium on the Interface

June 1–3, 2011

SMU | Bobby B. Lyle School of Engineering

Table of Contents

Association Rules

Mining association rules was fist introduced by Agrawal *et al.* (1993) and can formally be defined as:

- 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 $X \cap Y = \emptyset$.
- 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.

Association Rules II

- **Support:** $\text{supp}(X)$ is proportion of transactions which contain X
- **Confidence:** $\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$
- **Association rule** $X \Rightarrow Y$ will satisfy:

$$\text{supp}(X \cup Y) \geq \sigma, \text{conf}(X \Rightarrow Y) \geq \delta$$

- **Lift** (Brin *et al.*, 1997): $\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)\text{supp}(Y)}$

Example

$$\{\text{milk, bread}\} \Rightarrow \{\text{butter}\}$$

$$\text{support} = 0.2$$
$$\text{confidence} = 0.9$$
$$\text{lift} = 2$$

The AR Mining Process

Two-step process

1. Minimum support is used to generate the set of all **frequent itemsets.**
2. Each frequent itemsets is used to generate all possible rules which satisfy the minimum confidence constraint.

Worst case: $2^n - n - 1$ frequent itemsets (size ≥ 2 at n distinct items). Each frequent generates $2+$ rules $\Rightarrow O(2^n)$.

Practical Strategy: increase minimum support $\Rightarrow$ misses important rules.

Requirement for real setting

We need to be able to deal with large sets of association rules.

Example: Create Rules

```
R> library("arulesViz")
R> data("Groceries")
R> Groceries

transactions in sparse format with
 9835 transactions (rows) and
 169 items (columns)

R> rules <- apriori(Groceries, parameter = list(support = 0.001,
+     confidence = 0.5), control = list(verbose = FALSE))
R> rules

set of 5668 rules

R> inspect(head(sort(rules, by = "lift"), 3))

  lhs                          rhs                support confidence lift
1 {Instant food products,
   soda}                    => {hamburger meat} 0.00122      0.632 19.0
2 {soda,
   popcorn}                 => {salty snack}    0.00122      0.632 16.7
3 {flour,
   baking powder}           => {sugar}          0.00102      0.556 16.4
```

Visualization

Explore large sets of rules visually. Many researchers introduced visualization techniques

- Scatter plots
- Matrix visualizations
- Graphs
- Mosaic plots
- Parallel coordinates plots

Most existing visualization techniques are not suitable for displaying really large sets of rules!

"grouped matrix-based visualization"

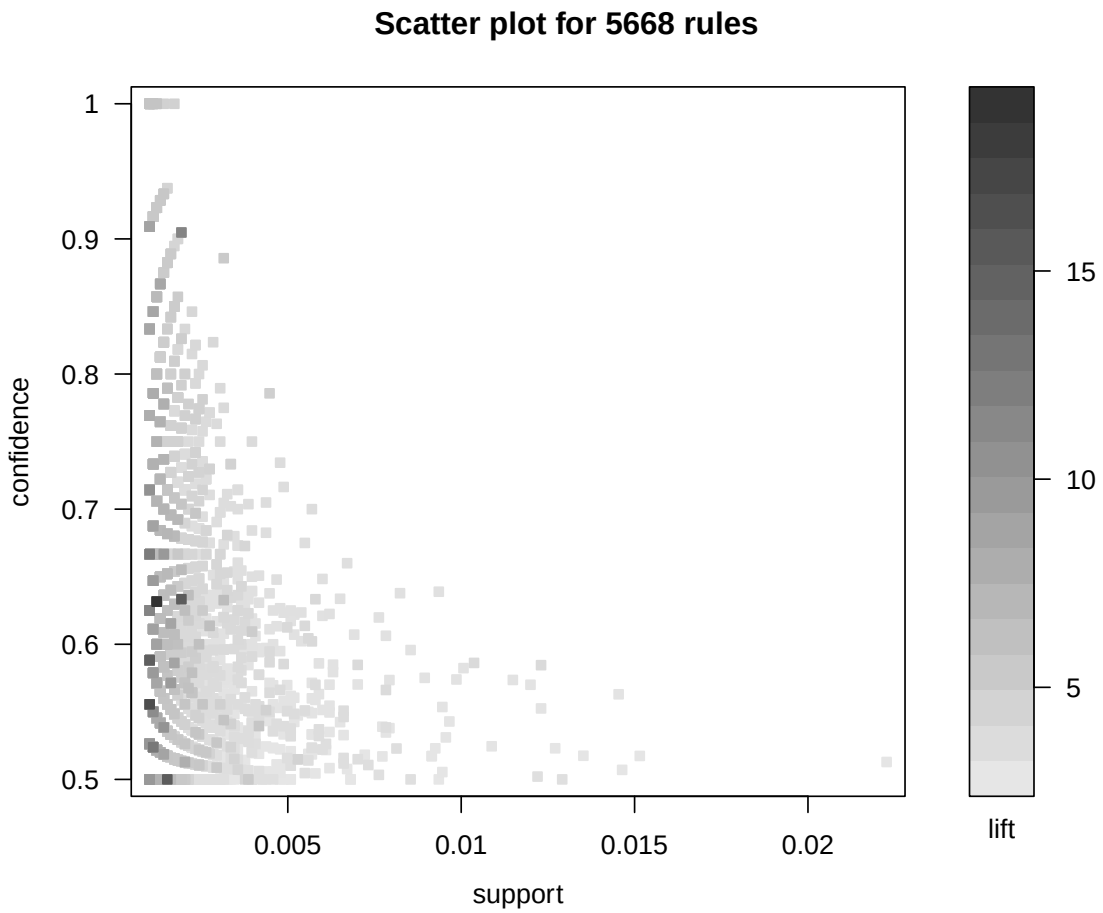
Visualize large sets of rules based on a novel way of creating nested groups of rules (more specifically antecedent itemsets) via clustering. The nested groups form a hierarchy which can be interactively explored down to the individual rule.

Table of Contents

Scatter plot

Scatter plot with two interest measures (e.g., support and confidence) on the axes (Bayardo, Jr. and Agrawal, 1999; Unwin *et al.*, 2001).

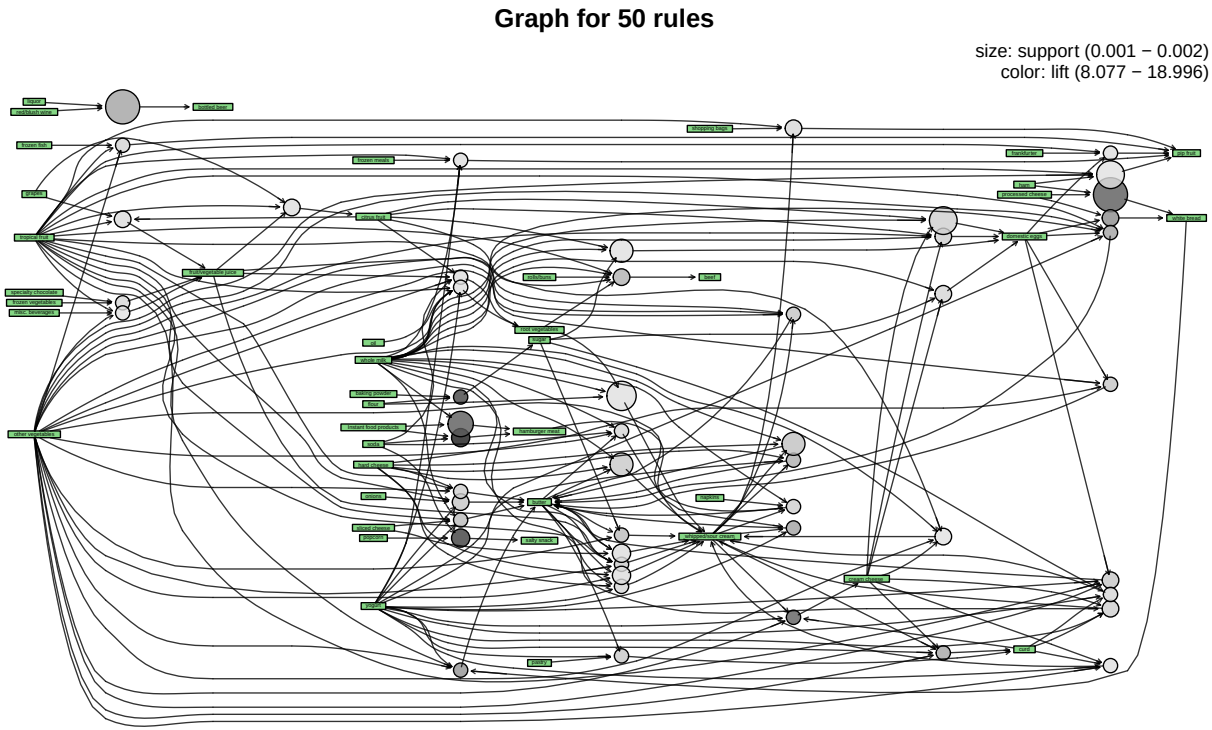
```
R> plot(rules)
```



Graph-based techniques

Visualize association rules using items/itemsets as vertices (Klemettinen *et al.*, 1994; Rainsford and Roddick, 2000; Buono and Costabile, 2005; Ertek and Demiriz, 2006).

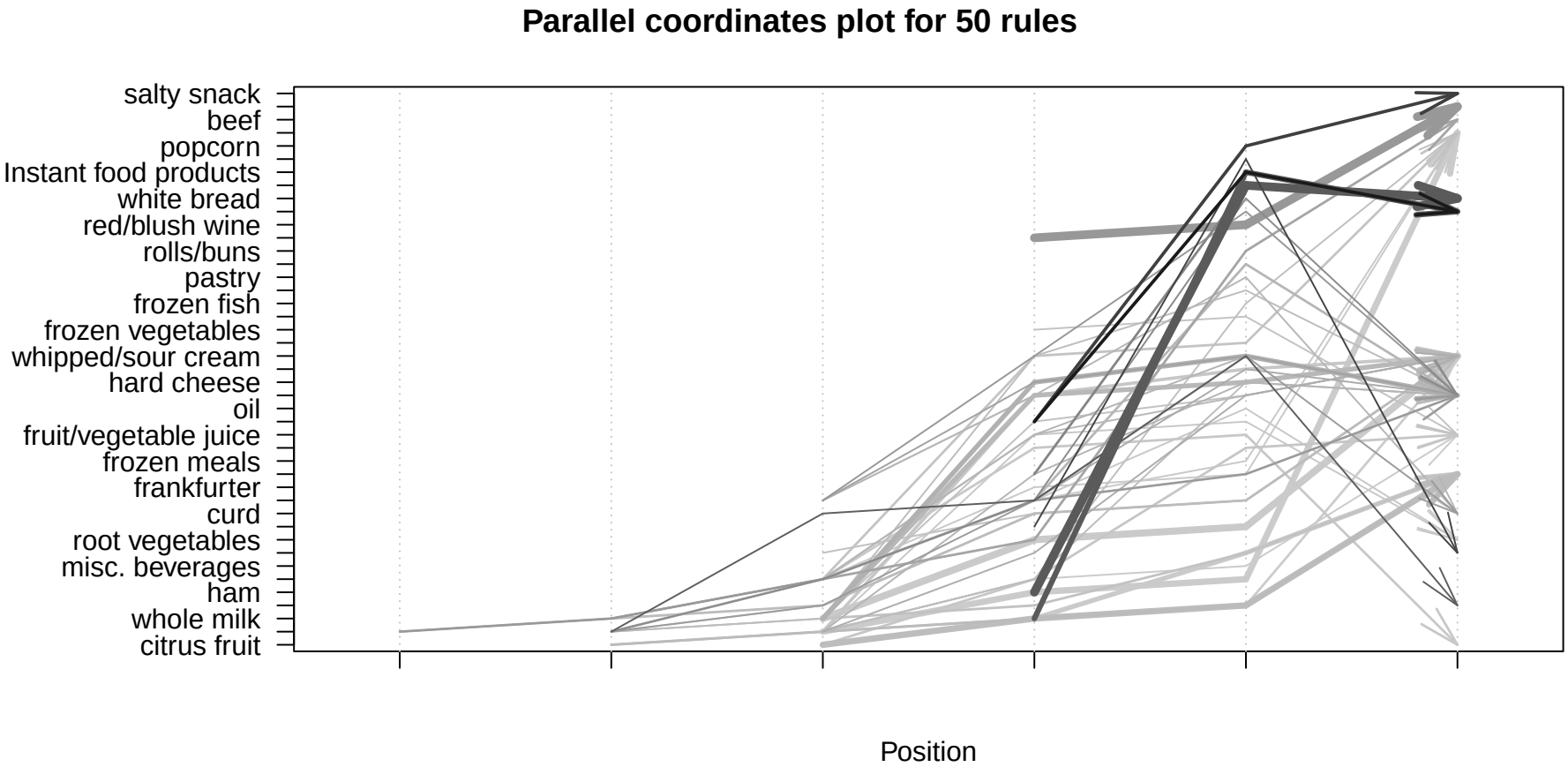
```
R> subrules <- head(sort(rules, by = "lift"), 50)
R> plot(subrules, method = "graph", control = list(type = "items",
+     engine = "graphviz"))
```



Parallel coordinates plots

Parallel coordinates plots were used previously to visualize discovered classification rules (Han *et al.*, 2000) and association rules (Yang, 2003).

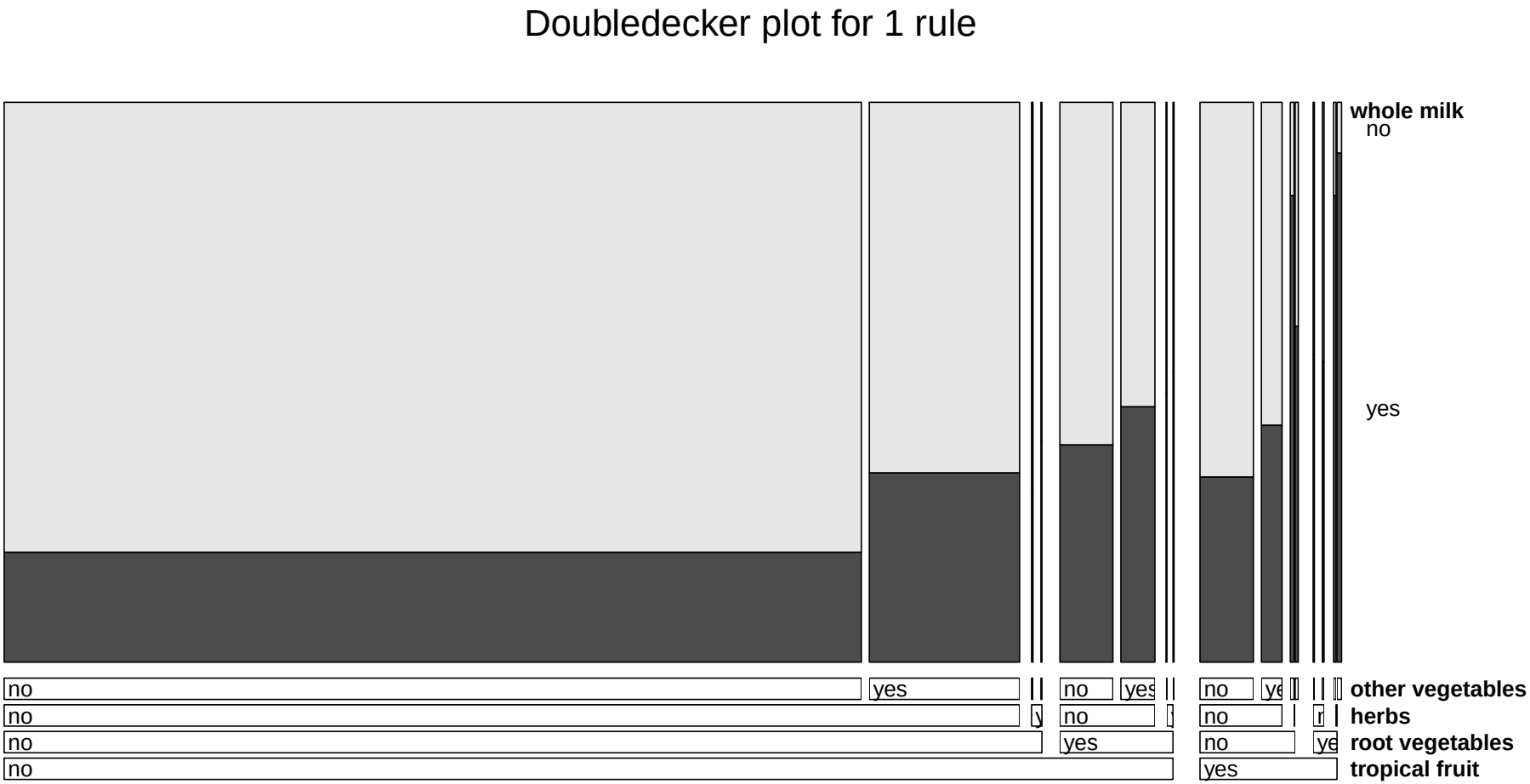
```
R> plot(subrules, method = "paracoord")
```



Double decker plots

Mosaic plot with a single horizontal split Hofmann *et al.* (2000).

```
R> oneRule <- sort(rules, by = "confidence")[100]
R> plot(oneRule, method = "doubledecker", data = Groceries)
```



Matrix-based Visualization

Organize LHS/RHS itemsets on x/y axis and display interest measure on the intersection (Wong *et al.*, 1999; Ong *et al.*, 2002).

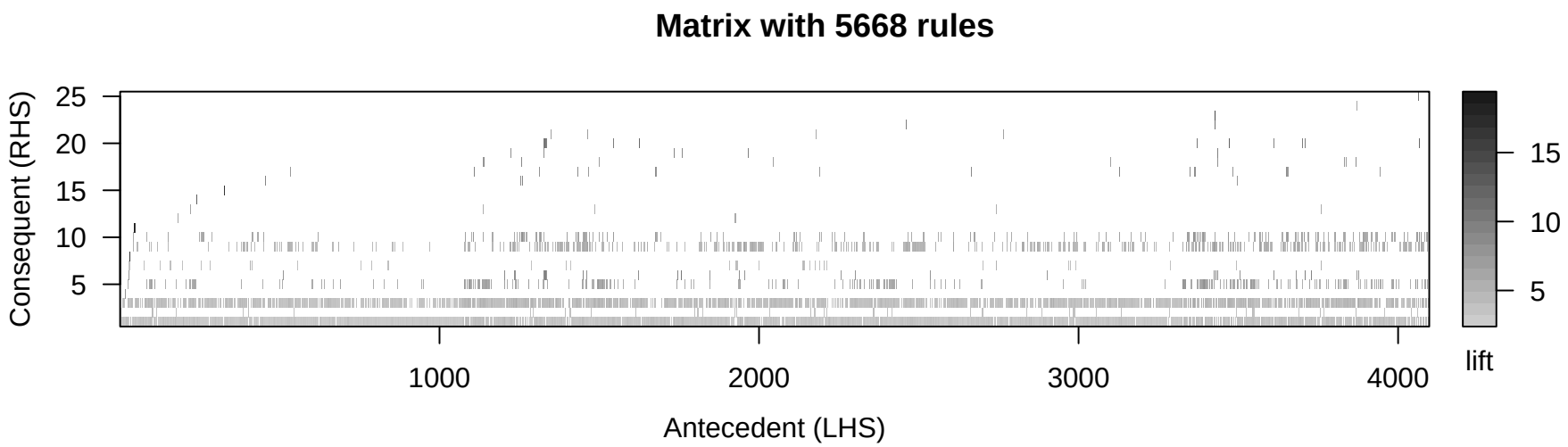
Construction

1. Create $\mathcal{R} = \{\langle X_1, Y_1, \theta_1\rangle, \ldots, \langle X_i, Y_i, \theta_i\rangle, \ldots, \langle X_n, Y_n, \theta_n\rangle\}$ where X_i is the antecedent, Y_i is the consequent and θ_i is the selected interest measure for the i-th rule, $i = 1, \ldots, n$.
2. Create a $A \times C$ matrix $\mathbf{M} = (m_{ac})$, $a = 1, \ldots, A$ and $c = 1, \ldots, C$, with one column for each unique antecedent and one row for each unique consequent in $\mathcal{R}$.
3. Populate $\mathbf{M}$ with $m_{ac} = \theta_i$ where $i = 1, \ldots, n$ is the rule index, and a and c correspond to the position of X_i and Y_i in the matrix.

Note: $\mathbf{M}$ will be sparse!

Matrix-based Visualization II

```
R> plot(rules, method = "matrix", measure = "lift")
```




```
Itemsets in Antecedent (lhs)
    [1] "{liquor,red/blush wine}"
    [2] "{curd,cereals}"
    [3] "{yogurt,cereals}"
    [4] "{butter,jam}"
    [5] "{soups,bottled beer}"
            (lines omitted)
    [343] "{tropical fruit,root vegetables,rolls/buns,bottled water}"
    [344] "{tropical fruit,root vegetables,yogurt,rolls/buns}"
Itemsets in Consequent (RHS)
    [1] "{bottled beer}"     "{whole milk}"       "{other vegetables}"
    [4] "{tropical fruit}"   "{yogurt}"           "{root vegetables}"
            (lines omitted)
```

Matrix-based Visualization II

```
R> plot(rules, method = "matrix3D", measure = "lift")
```

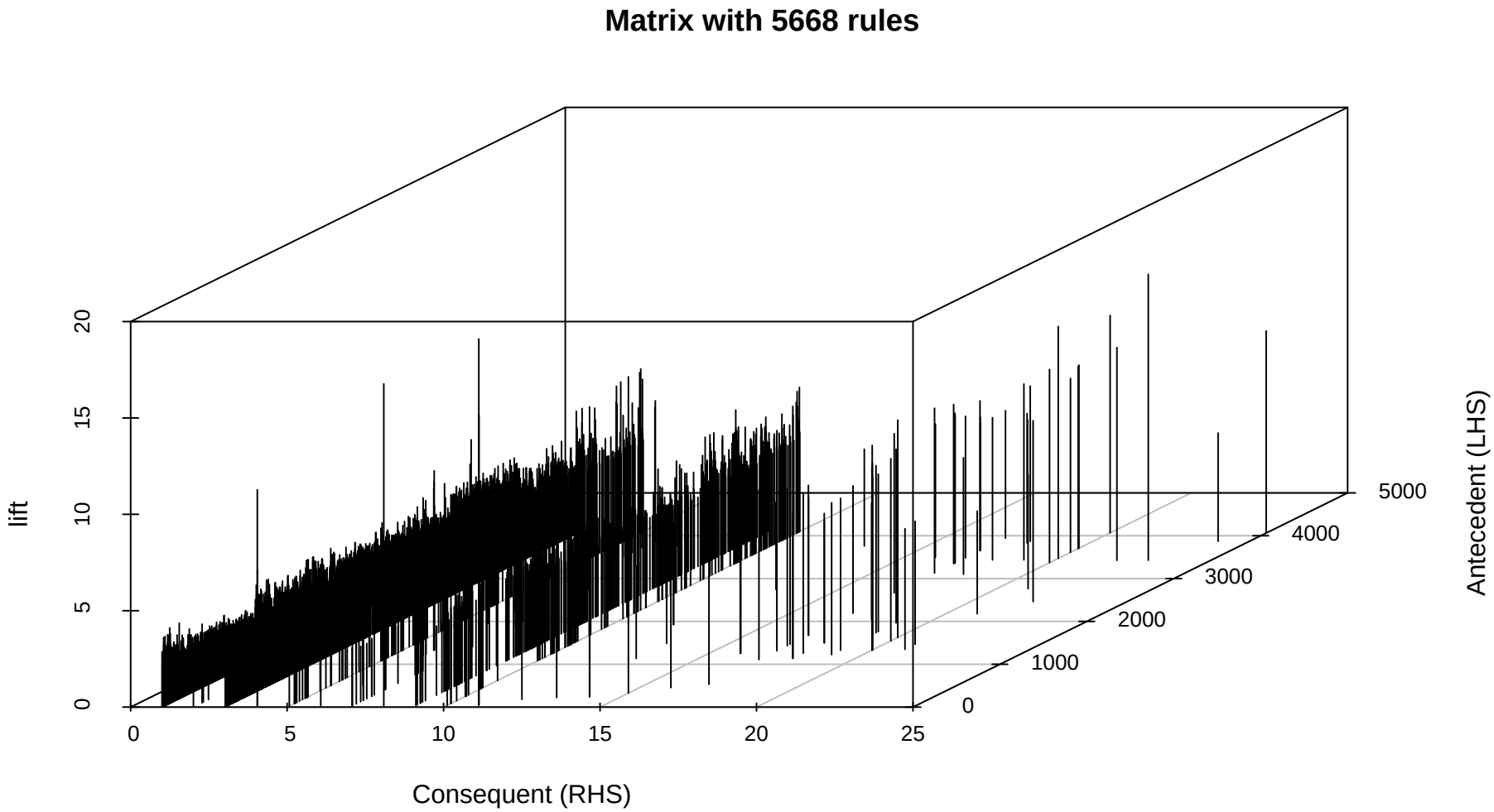


Table of Contents

Idea

Enhances matrix-based visualization by

- Group rules via clustering to handle large sets of rules.
- Groups of rules are presented by aggregating rows/columns of the matrix $\mathbf{M}$.
- Groups are nested and organized hierarchically allowing the user to explore groups interactively.
- Typically we only need to group LHS for rules with a single item in the RHS.

Difficulties:

- Clustering association rules
- Missing values in $\mathbf{M}$
- Visualization of groups

Grouping rules: Clustering

Distance-based clustering of rules/itemsets

$$d_{\text{Jaccard}}(X_i, X_j) = 1 - \frac{|X_i \cap X_j|}{|X_i \cup X_j|}.$$

Called **conditional market-basket probability** by Gupta *et al.* (1999).

Problems

- High dimensionality (many different items)
- Sparsity (high support rules are short)

Alternatives

Number of common covered transactions (Toivonen *et al.*, 1995; Berrado and Runger, 2007) $\Rightarrow$ strong bias towards clustering rules which are generated from the same frequent itemset.

Grouping rules II: Clustering

Group rules based on interest measure similarity

- Let $\mathbf{M}$ be a interest measure matrix item. Columns/rows in $\mathbf{M}$ are the unique LHS/RHS in $\mathcal{R}$
- Grouping rules means grouping columns/rows in $\mathbf{M}$.

Grouping LHS/RHS independently into k groups $\mathcal{S} = \{S_1, S_2, \ldots, S_k\}$ while minimizing the within-cluster sum of squares

$$\text{argmin}_{\mathcal{S}} \sum_{i=1}^{k} \sum_{\mathbf{m}_j \in S_i} ||\mathbf{m}_j - \boldsymbol{\mu}_i||^2,$$

where $\mathbf{m}_j$, $j = 1, \ldots, A$, is a column vector of $\mathbf{M}$ (all rules with the same antecedent) and $\boldsymbol{\mu}_i$ is the center (mean) of the vectors in S_i.

$\Rightarrow$ k-means algorithm by Hartigan and Wong (1979)

Grouping rules III: Missing values

Missing values

- $\mathbf{M}$ is very sparse due to minimum support and confidence.
- Values are not missing randomly! Values miss for "not interesting rules."

Aim: Group LHSs when they have many missing values with the same set of RHSs

- Replace all missing values with a "neutral" value. E.g., 1 for lift indicates that LHS and RHS are statistically independent
- Ensures that matching missing values will contribute positively for grouping.

Grouping rules IV: Aggregation

Aggregate interest measures to represent the whole group.

Several aggregation functions

- maximum
- minimum
- average
- median

In the examples in this paper we use the median to represent the group since it is robust against outliers.

Grouping rules V: Interest Measure

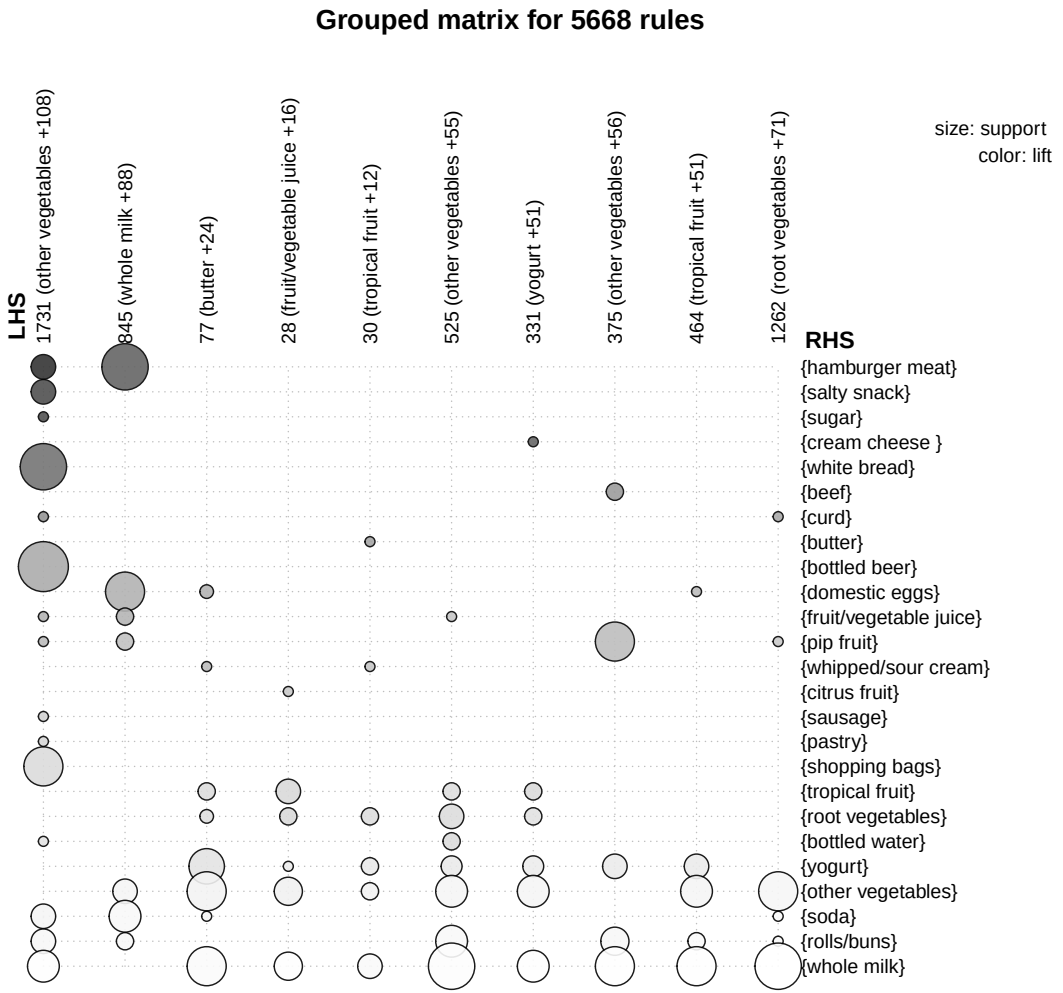
Interest measure of choice: **Lift**

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)\text{supp}(Y)}$$

- LHSs that are statistically dependent with the same RHSs (i.e., have a high lift value) are similar.
- Also groups LHSs containing **substitutes** (e.g., butter and margarine)! Substitutes are rarely purchased together but have a similar dependence relationship with the same RHSs (e.g., bread).

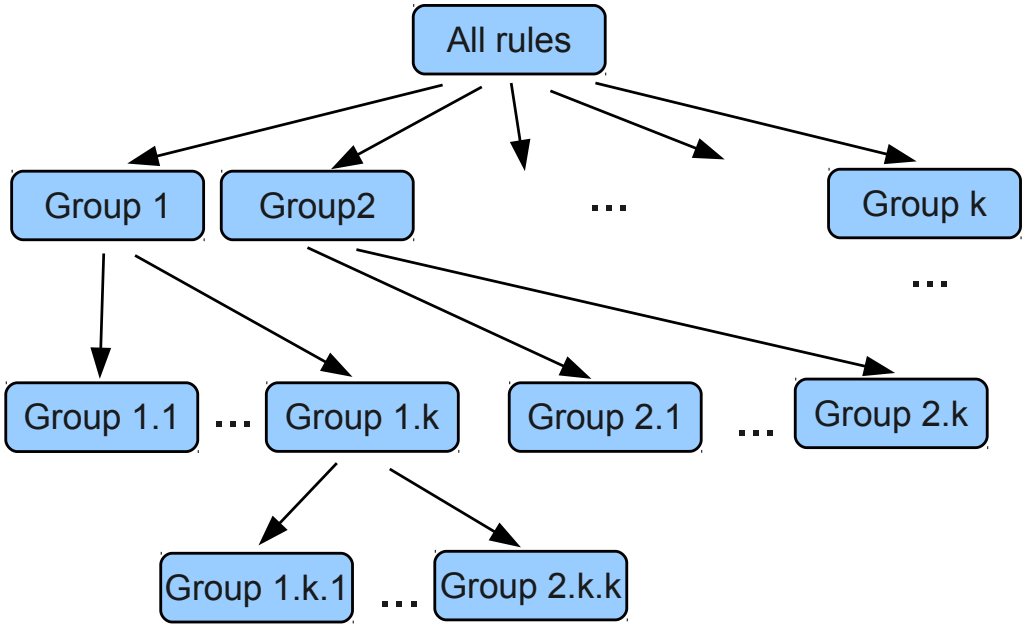
Visualization: Balloon plot

```
R> set.seed(8000)
R> plot(rules, method = "grouped", control = list(k = 10))
```



- Balloon represents a group (only LHSs are grouped)
- Balloon size represents aggregate support
- Balloon color represents the aggregated interest measure (lift)
- Order by aggregated interest measure (lift)

Visualization: Hierarchical Structure



Interactive "drill down" procedure

1. Create grouping for $\mathbf{M}$
2. Create for selected group $S_i, i = 1, \ldots, k$, a submatrix $\mathbf{M}_i$.
3. Apply grouping process again to submatrices $\mathbf{M}_i$.

Live Interactive Example

(paste into R)

```
### install package
install.packages("arulesViz")


library("arulesViz")
data("Groceries")
Groceries

rules <- apriori(Groceries, parameter=list(support=0.001,
    confidence=0.5))
rules

plot(rules, method="grouped", interactive=TRUE)
```

Table of Contents

Conclusion

Main features

- Visualization of large rule sets
- Handles complementary items
- Guide the user automatically to the most interesting groups/rules
- Easy to understand (similar to matrix-based visualization)

Future work

Explore different other ways to group antecedents and to look at grouping antecedents and consequents simultaneously (i.e., by co-clustering/two-mode clustering).

Code

Association rule visualizations are implemented in the R extension package **arulesViz** (Hahsler and Chelluboina, 2011) which is freely available via the Comprehensive R Archive Network at

`http://CRAN.R-project.org/package=arulesViz`.

References I

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.

Roberto J. Bayardo, Jr. and Rakesh Agrawal. Mining the most interesting rules. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 145–154. ACM, 1999.

Abdelaziz Berrado and George C. Runger. Using metarules to organize and group discovered association rules. *Data Mining and Knowledge Discovery*, 14(3):409–431, 2007.

Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 255–264, Tucson, Arizona, USA, May 1997.

Paolo Buono and Maria Francesca Costabile. Visualizing association rules in a framework for visual data mining. In *From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments*, pages 221–231, 2005.

Gürdal Ertek and Ayhan Demiriz. A framework for visualizing association mining results. In *ISCIS*, pages 593–602, 2006.

Gunjan Gupta, Alexander Strehl, and Joydeep Ghosh. Distance based clustering of association rules. In *Intelligent Engineering Systems Through Artificial Neural Networks (Proceedings of ANNIE 1999)*, pages 759–764. ASME Press, 1999.

Michael Hahsler and Sudheer Chelluboina. *arulesViz: Visualizing Association Rules*, 2011. R package version 0.1-1.

Jianchao Han, Aijun An, and Nick Cercone. *CViz: An Interactive Visualization System for Rule Induction*, pages 214–226. Springer Berlin / Heidelberg, 2000.

J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

Heike Hofmann, Arno Siebes, and Adalbert F. X. Wilhelm. Visualizing association rules with interactive mosaic plots. In *KDD*, pages 227–235, 2000.

Mika Klemettinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, and A. Inkeri Verkamo. Finding interesting rules from large sets of discovered association rules. In *CIKM*, pages 401–407, 1994.

Kian-Huat Ong, Kok leong Ong, Wee-Keong Ng, and Ee-Peng Lim. Crystalclear: Active visualization of association rules. In *In ICDM'02 International Workshop on Active Mining AM2002*, 2002.

References II

Chris P. Rainsford and John F. Roddick. Visualisation of temporal interval association rules. In *IDEAL '00: Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineering, and Intelligent Agents*, pages 91–96. Springer-Verlag, 2000.

H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hatonen, and H. Mannila. Pruning and grouping discovered association rules. In *Proceedings of KDD'95*, 1995.

Antony Unwin, Heike Hofmann, and Klaus Bernt. The twokey plot for multiple association rules control. In *PKDD '01: Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 472–483. Springer-Verlag, 2001.

Pak Chung Wong, Paul Whitney, and Jim Thomas. Visualizing association rules for text mining. In *INFOVIS '99: Proceedings of the 1999 IEEE Symposium on Information Visualization*, page 120, Washington, DC, USA, 1999. IEEE Computer Society.

Li Yang. Visualizing frequent itemsets, association rules, and sequential patterns in parallel coordinates. In *Computational Science and Its Applications – ICCSA 2003*, Lecture Notes in Computer Science, pages 21–30, 2003.