

Visual exploration of association rules ¹

Adalbert F.X. Wilhelm ²

School of Humanities and Social Sciences, Jacobs University Bremen, e-mail:
a.wilhelm@jacobs-university.de

Abstract: Association rules, which have their origin in market basket analysis, are a widely applied tool for mining large data bases containing many dichotomous variables. While generating association rules is a pretty straightforward task and easily achievable by modern computer technology, the evaluation and interpretation of association rules provides a challenge in the practice of data analysis and primarily it has to be achieved manually. This paper presents an overview on visualization techniques that are useful and have been successfully adapted to the association rule context.

Four different aspects are touched upon: first, providing an overview on all or a large group of rules, second, representing pruning techniques and their effect, third, exploring and comparing individual rules, and fourth visualizing sets of related rules.

Keywords: association rule mining, doubledecker plot, rule comparison, two-key plot, visual exploration

1. Association Rules

Association Rules have been proposed by Agrawal et al. (1993) in the context of market basket analysis. They were invented to provide an automated process, which could find connections among items, that were not known before, especially to answer questions like: “which items are likely to be bought together?”. Applications of association rules include market basket analysis, attached mailing in direct marketing, fraud detection and department store/shelf planning. Typically, the data to be examined consists of customer purchases, i.e. a set of items bought by a customer over a period of time.

An *association rule* is an implication of the form $X \rightarrow Y$, where X and Y are mutually exclusive itemsets. An association rule $X \rightarrow Y$ holds with *confidence* $c = c(X \rightarrow Y)$, if $c\%$ of transactions in D that contain X also contain Y . $X \rightarrow Y$ has *support* s in the database D , if $s\%$ of transactions in D contain $X \cup Y$.

Most data mining software offers a procedure to generate all association rules with confidence and support that exceed some user-specified minimum thresholds for support (*min-sup*) and confidence (*minconf*). The procedures are typically based on the a priori algorithm introduced by Agrawal & Srikant (1994). The key point of this algorithm lies in the fact that for any subset $X' \subseteq X$ of items its support is bounded from below by the support of the itemset X . In addition, the confidence for the rule $(X \setminus X') \rightarrow X'$ can be derived directly from the supports of the related itemsets using the equality

$$\text{confidence}((X \setminus X') \rightarrow X') = \frac{\text{support}(X)}{\text{support}(X \setminus X')}.$$

²Address of correspondence: Campus Ring 1, D-28759 Bremen, GERMANY

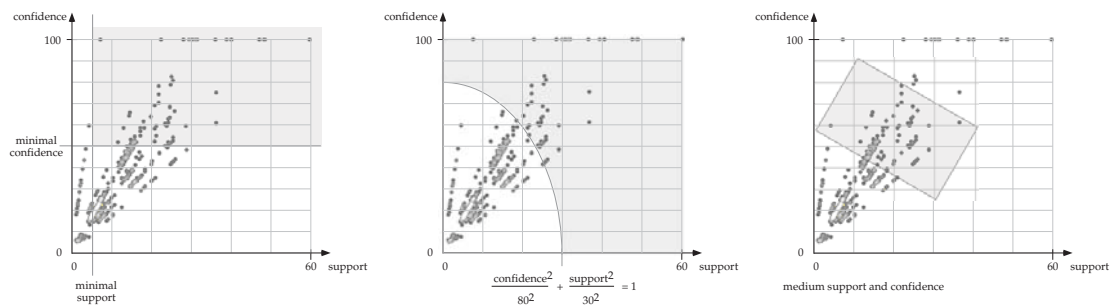
There are several problems related to this procedure: the setting of the thresholds of minimal support and confidence is crucial; choosing high support and confidence may lead to “uninteresting” results - insofar as they often are well known beforehand by domain experts. Lowering the minimal thresholds can lead to a vast increase of the number of results.

Another source of problems stems from the inability to estimate the quality of a rule merely from the two keys *confidence* and *support*. Graphical solutions showing a rule in the background of the corresponding contingency table have been proposed by Hofmann et al. (2000). Interactive methods further enhance these displays to powerful tools in the exploration of association rules.

2. Getting an overview on rules

To get an overview on all association rules, Unwin et al. (2001) have proposed the Two-Key-Plot that shows confidence and support of the rules in a scatter plot. Direct com-

Figure 1: *Three different principles of accepting rules. On the left, the traditional area of acceptance with specified minimal confidence and minimal support. The other two plots show potential alternatives.*



parisons between rules, which lie on vertical or horizontal lines are now possible: if two rules have the same confidence, the one with the higher support is better (the same is true for two rules with the same support). Accepted rules therefore (see sketch in Figure 1, left side) fall in a rectangle bounded by *minconf* and *minsupp*. Alternatively, one can also think of a different criterion for accepting rules. The plot in the middle of Figure 1 shows an area of acceptance, where a low support can be balanced by a high confidence and vice versa. Using interactive selection methods and linking opens a lot more choices for criteria of acceptance, all of which may be sensible in the background of a specific application. From a practitioner’s viewpoint rules with high confidence and support are not very interesting because they are already known to the domain experts and hence it is more desirable to focus on medium confidence and support levels, as is indicated by the acceptance area in the plot at the right hand side of Figure 1.

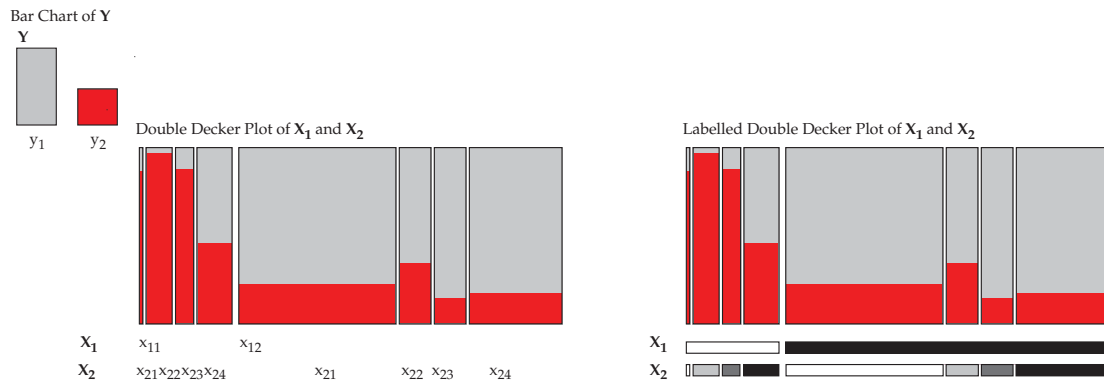
3. Comparing individual rules

Doubledecker plots have been introduced by Hofmann et al. (2000). They can be used to display an association rule $X \rightarrow Y$ (see figure 2). Usually, a left hand side X of a rule consists of a combination of several items being bought together - for a doubledecker plot we will consider all possible combination of these items. For all of these we will calculate support and confidence. This gives a list like:

heineken	coke	chicken	$\rightarrow$ sardines		$\rightarrow \neg$ sardines	
combination			supp	conf	supp	conf
no	no	no	0.07	0.39	0.11	0.61
no	no	yes	0.01	0.16	0.08	0.84
no	yes	no	0.03	0.26	0.07	0.74
no	yes	yes	0.00	0.05	0.02	0.95
yes	no	no	0.06	0.17	0.28	0.83
yes	no	yes	0.00	0.04	0.08	0.96
yes	yes	no	0.00	0.07	0.05	0.93
yes	yes	yes	0.12	0.98	0.00	0.02

This list is visualized by means of a doubledecker plot: each row is assigned a rectangle, the width of which corresponds to the sum of both supports. *Highlighted* are the cases for Y , i.e. 'sardines bought' in the example. Therefore the *area of highlighting* corresponds to the support of a rule, the *height of highlighting* corresponds to the rule's confidence.

Figure 2: Doubledecker plot of three variables. The rightmost bin corresponds to the rule heineken & coke & chicken $\rightarrow$ sardines.



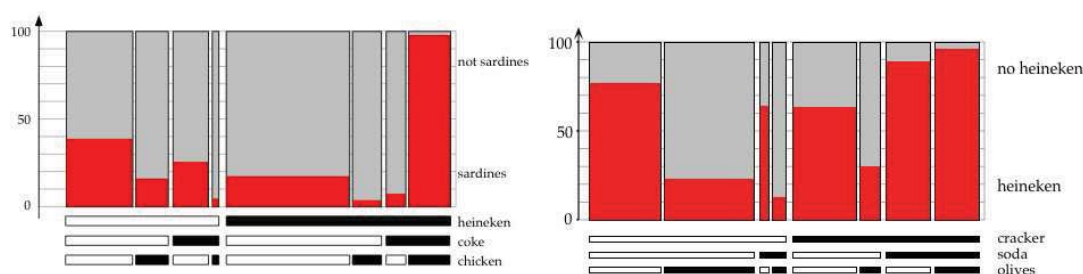
To determine the exact combination that a bin represents the labels below the plot in the right hand figure 2 are used: Each row of the labels corresponds to one variable, the white rectangles stand for "0"s, the blacks for "1"s. The first bin in fig. 2 therefore represents the combination of "no item bought" (all "0"s), the second bin contains all transactions, where only *chicken* has been bought, and so on.

Using a standard method for finding association rules, the following two rules are a typical part of the result set

rule	confidence	support
heineken & coke & chicken $\rightarrow$ sardines	98.31	11.59
cracker & soda & olives $\rightarrow$ heineken	96.83	12.18

From these numbers alone it is hard to decide on the quality and validity of the rules. One of the problems is that the support is quite incomparable between rules with different items, especially if the marginal probabilities differ as strongly as $P(\text{heineken}) = .6$ and $P(\text{sardines}) = .3$. Doubledecker plots provide us with the additional information we need. Figure 3 in the left hand plot shows the ideal situation for an association rule of high quality. Only the bin corresponding to $\text{heineken} \& \text{coke} \& \text{chicken} \rightarrow \text{sardines}$ shows high support and high confidence, whereas all other combinations of these variables have low confidence and support.

Figure 3: Doubledecker plots of three variables showing a strong and a weak association rule. In the left plot, the rightmost bin corresponds to the rule $\text{cracker} \& \text{soda} \& \text{olives} \rightarrow \text{heineken}$. This bin is filled almost entirely with highlighting, while none of the other bins is filled to more than a third. In the right hand plot the rightmost bin corresponds to the rule $\text{cracker} \& \text{soda} \& \text{olives} \rightarrow \text{heineken}$.



The right hand plot in Figure 3 on the other hand exposes the second rule to be rather weak. The plot is very different from the one on the left; five of the eight combinations show an increase in confidence ($> 60\%$ highlighting height), four of them also have a high amount of support. This indicates that the rule is very weak. From the two keys, confidence and support, however, both rules looked very similar.

References

- Agrawal R., Imielinski T. and Swami A. (1993) Mining associations between sets of items in massive databases, in: *Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data*, Washington D.C., 207 – 216.
- Agrawal R. and Srikant R. (1994) Fast algorithms for mining association rules, Technical report, IBM Research Report RJ9839.
- Hofmann H., Siebes A.P.J.M. and Wilhelm A.F.X. (2000) Visualizing association rules with interactive mosaic plots, in: *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, New York, NY, USA, 227–235.
- Hofmann H. and Wilhelm A. (2001) Visual comparison of association rules, *Computational Statistics*, 16, 399–415.
- Unwin A., Hofmann H. and Bernt K. (2001) The twokey plot for multiple association rules control, in: *Principles of Data Mining and Knowledge Discovery, PKDD 2001, Freiburg, Germany, September 3-5, 2001, Proceedings*, Raedt L.D. and Siebes A., eds., Springer, volume 2168 of *Lecture Notes in Computer Science*, 472–483.