

Visual Post Analysis of Association Rules

H. Hofmann
AT&T Statistics Research
Florham Park, NJ, 09750

Abstract

Association Rules are a widely used tool in data mining. Major problems originate from the mass of output as well as from the restriction to support and confidence for measures of quality. These are two very different problems - the first is unfortunately a fact, the second can be worked on. Due to their mass rules are often judged only by their support and confidence. We show, that this causes certain problems and introduce graphical techniques for examining association rules. These allow us not only to assess the quality of a single rule visually, but they also provide an overview of the structure among the rules, laying the basis for an interpretation and extraction of “real” results.

1 Introduction

What are Association rules and why do we need graphics to help the analysts?

Association rules (Agrawal et al., 1993) have their origin in the market basket analysis. A typical data set consists of a number of transactions - a set of items being purchased together by a customer. Rules have the form $X \cap Y \rightarrow Z(s, c)$, where X, Y and Z usually can be thought of as sets of available items. On a first glance, an association rule is fairly easy to interpret: “ $c\%$ of all customers who have bought items X and Y also have bought item Z ; $s\%$ of all customers did.” This statement holds with a certain amount of likelihood, called the rule’s *confidence* c ; and the number of actual purchases of all items together gives the rule a measure of marketable interestingness: its *support* s . Association rules are relatively easy to compute (Agrawal and Srikant, 1994) - however, the algorithm will produce depending on the setting of the two parameters huge quantities of output. The “real” problem therefore starts in the necessity of finding “real results” among all the rules.

One problem of association rules is their lopsidedness. Association rules only regard the positive event of an item “being bought” and make no statement about the negative event at all. Thereby annoying problems ap-

pear: if rule $X \rightarrow Y$ has confidence c , it can happen that the negative rule $\neg X \rightarrow Y$ has higher confidence. Any set of automatically generated rules will contain these problematic rules. Pruning mechanisms (e.g. Liu et al. (1999)) are used to eliminate these rules from the output. For the present paper we will assume that the set of association rules is already pruned, thereby getting rid of the annoying problems mentioned above. Each rule is thought to have a validity of its own.

Another class of problems shows up when dealing with several of these association rules at the same time. For a set of different association rules with the same response it is of interest to examine, whether and how far the populations described by the explanatory variables (left hand side of a rule) differ from each other. It could well be one and the same population, and different rules providing only different descriptions for it. - The goal therefore is, to examine the impact of rules on one another.

We will suggest graphical methods as a solution at this point. There have been various approaches of visualizing association rules, some of which are implemented in standard datamining software. They often have in common, though, that for visualising 2-D and 3-D objects three and four dimensions are used, respectively (We have to count colour as an additional dimension). This, however, is not necessary and not good practice either (Tuft, 1983). We want to show, how we may use the available dimensions more efficiently.

Doubledecker plots (Hofmann et al., 2000) concentrate on single association rules and show a rule in the background of its corresponding contingency table. In a different approach, we focussed on the two measurements of *confidence* and *support* and visualised large numbers of association rules in one plot, the so-called *Two-Key-Plot* (Unwin et al., 2001). Figures 1 and 2 show examples for each of these plots.

The present paper can be found somewhere in between these two extremes regarding the number of association rules involved: We assume, that we are dealing with a set of 30 - 100 association rules; enough to have trouble in dealing with them manually, and too few to collide with computer space or memory.

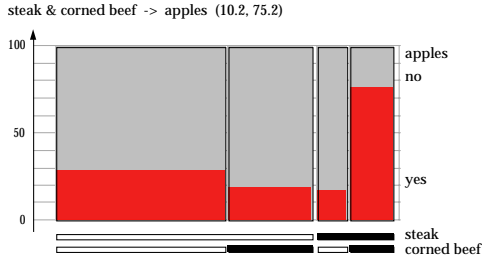


Figure 1: *Doubledecker plot corresponding to rule steak & corned beef \rightarrow apples. All four possibilities of buying/not buying steak and corned beef are drawn next to each other, the rectangles' widths representing the number of customers acting in this way. Highlighting (the dark marked areas) corresponds to the apple buyers among the customers, the height of highlighting areas corresponds to the conditional probability of buying apples (confidence). The only rectangle with high proportion of highlighting is the rightmost, corresponding to steak and corned beef purchases. This validates the rule steak & corned beef \rightarrow apples.*

Getting started . . .

For the further analysis of association rules it is important to be aware of the duality between an item or a set of items and a binary variable: each set of items X implies a binary variable \mathcal{X} . The i th element of this variable \mathcal{X}_i is given as:

$$\mathcal{X}_i = \begin{cases} 1 & \text{if transaction } i \text{ contains all items } X \\ 0 & \text{otherwise.} \end{cases}$$

After making this distinction, we will denote both the itemset and the variable by X . The exact use of it will always be clear from the context. There's room for another shortcut: instead of writing $X = 1$ and $X = 0$ we will distinguish these events by writing X and $\neg X$.

Now we are able to express both *confidence* and *support* of a rule in a more statistical manner:

$$c = \text{conf}(X \rightarrow Y) = P(Y | X),$$

the conditional probability of Y given X , and,

$$s = \text{supp}(X \rightarrow Y) = P(X \cap Y),$$

the joint probability of X and Y .

Using this notation the introductory problem can also be described. Due to the lopsidedness of association rules only two (out of 4) cells can be reconstructed for the corresponding contingency table.

	X	$\neg X$
Y	s	?
$\neg Y$	$s/c - s$?

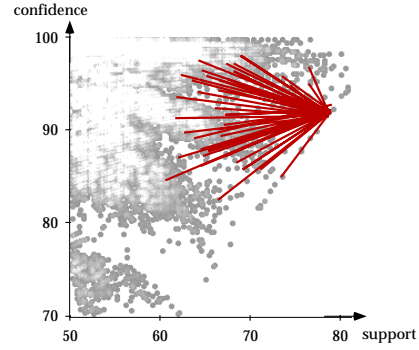


Figure 2: *TwoKey Plot: confidence and support of several thousand rules are plotted vs each other. The white and light gray shaded areas mark places of high density. Lines are drawn from one rule to all of its successors (all rules of the form $X \cap Y \rightarrow Z$ for rule $X \rightarrow Z$). This allows to exploit the underlying meta-structure among rules.*

This may be the reason for the uncomfortable feeling when analyzing association rules.

2 Distance between rules

In the following section we will assume that all association rules we deal with have the same right hand side. Toivonen et al. (1995) defined the distance between each pair of association rules by the number of transactions they have in common:

The *distance* between two rules $X \rightarrow Z$ and $Y \rightarrow Z$ is given by (cf. figure 3)

$$\begin{aligned} d(X \rightarrow Z, Y \rightarrow Z) &:= \\ &= n \cdot P(((X \cap Z) \cup (Y \cap Z)) \setminus (X \cap Y \cap Z)) \\ &= nP(X \cap Z) + nP(Y \cap Z) - 2nP(X \cap Y \cap Z). \end{aligned}$$

Since we only regard association rules with the same right-hand-side, we can write $d_Z(X, Y)$ for short.

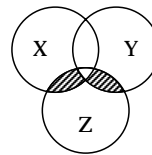


Figure 3: *Venn diagram of the basic sets of the rules $X \rightarrow Z$ and $Y \rightarrow Z$. The hatched area gives the distance as defined by Toivonen et al. (1995).*

In order to analyze the relationship among a set of rules with the same right hand side we set up the matrix of distances among them. Tools from Multidimensional

Scaling (MDS) can now be used to get more information out of these distance matrices.

The scatterplots in figures 4 and 5 display the results of a 2d metric MDS of a distance matrix corresponding to 75 association rules. Figure 4 shows a scatterplot of the coordinates of a 2d MDS. Several clusters of points indicate that the association rules can be classified into different groups. Further querying of this display reveals, that one of the clusters consists of a group of 9 rules, assembled from only 4 items altogether. Highlighting the five rules with highest confidence (100%) demonstrates that these rules belong to four different clusters, covering 71% of the response in total.

some concluding remarks ...

3 Fluctuation diagrams

What are Fluctuation Diagrams? What do they show and how are they constructed? *Fluctuation diagrams* are one of the variations of *mosaic plots* (Hartigan and Kleiner, 1981; Hofmann, 2000) used for displaying data matrices. Their main use is to visually show the numbers of an $r \times c$ spread sheet. Instead of the number a box is drawn, the size of which is proportional to the number it represents.

The display is useful for emphasising large bins, while small or empty bins simply vanish.

Construction of a fluctuation diagram: Starting from an $r \times c$ grid of equal sized bins, the information on the cell sizes is included by shrinking both height and width of each bin, such that the remaining area is proportional to the cell size while preserving the aspect ratio. If the original display is quadratic, the height and the width of a bin end up being proportional to the square root of the cell size. This makes comparisons of widths and heights of the bins possible at the same time.

Remark: the data does not need to be two dimensional, there are extensions for higher dimensions.

Clearly, the fluctuation diagrams in this context shows the distance matrix for a set of association rules.

Fluctuation diagrams & Association rules Again we restrict ourselves to a set of association rules with the same right hand side. The *intersection* of two rules is a similarity measure, given by the number of transactions that fulfill both left hand sides:

$$i(X \rightarrow Z, Y \rightarrow Z) = i(X, Y) = |X \cap Y|.$$

We are going to display the *intersection* each pair of association rules has in common, this gives a symmetric

matrix. To this matrix we add another row and column containing the intersection between the outcome variable Z and each of the right hand sides of the set. The resulting matrix M looks like this:

$$M = \left(\begin{array}{c|c} i(Z, Z) & i(X_j, Z) \\ \hline i(X_j, Z) & i(X_j, X_k) \end{array} \right)$$

Displaying this matrix in a fluctuation diagram (see figure 6) gives an overview of the relationships among the rules. Of course, the order of the rules has an important impact. In figure 6 the order is given by a greedy 1-step sorting algorithm according to the Euclidean distance between the rows. Nevertheless, the diagram shows the strong relationships among these rules: five distinct clusters of rules become apparent. The second cluster of rules includes all nine rules of figure 4 examined more closely above.

Why did we choose the matrix in this way? By regarding the amount of intersection between two rules rather than the distance, we focus on the *similarity* between two objects. This has two advantages: clusters of rules belonging together can be seen easily (see fig. 6). More important, though, is that we can draw conclusions both about support and confidence of rules and still have some information about the distance between rules from the same display.

Closer look at the boxes Besides the general overview, a fluctuation diagram displaying association rules in the way described above contains a further source of interpretation. Basically, there are three different types of boxes in a fluctuation diagram we have to distinguish. Figure 7 shows all of them: boxes on the diagonal (1), off-diagonal boxes in the first row or column (2) and the rest (3).

type	location	meaning
(1)	diagonal	$P(X_j)$ or $P(Z)$
(2)	1st row or column	$P(X_j \cap Z)$ = support of rule $X_j \rightarrow Z$
(3)	off-diagonal	$P(X_j \cap X_k)$ intersection of rules $X_j \rightarrow Z$ and $X_k \rightarrow Z$

Comparing the size of a box on the diagonal with its counterpart on the first row of the same column therefore gives an estimate of a rule's confidence. This estimation is, of course, a very crude one and only noticeable, if the differences between the boxes are big. The larger a box on the diagonal is compared to the first row box, the smaller is the confidence of the corresponding rule.

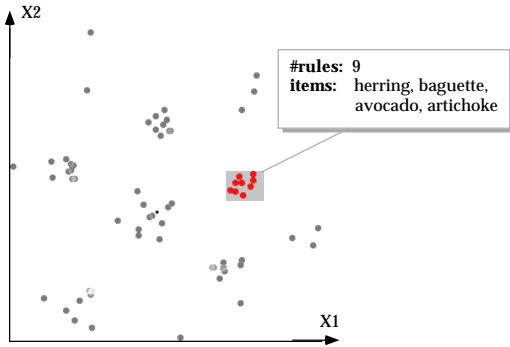


Figure 4: Scatterplot of the coordinates of a 2d metric MDS of the distances between each pair of 75 association rules with the same right hand side. Interactive querying reveals a cluster of 9 association rules consisting of only 4 highly correlated items.

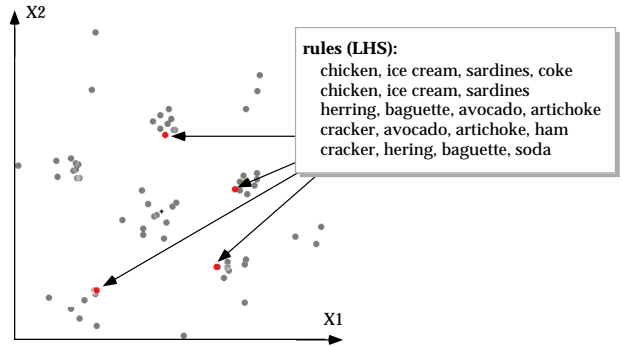


Figure 5: MDS scatterplot: highlighted are the five rules with confidence 100%. These rules belong to four different clusters.

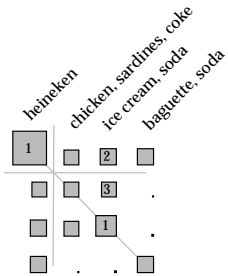


Figure 7: Closer look at the boxes of a fluctuation diagram. Three different types of boxes have to be distinguished.

From boxes of type (3) - off-diagonal boxes - we can get an also rather crude estimate of the distance between two rules. The box represents $|X_j \cap X_k|$. If this box is large compared to the two boxes on the diagonal corresponding to $|X_j|$ and $|X_k|$, we can assume, that the distance between the two rules $X_j \rightarrow Z$ and $X_k \rightarrow Z$ is not big (especially, if both rules have reasonable high confidence).

Figure 7 shows a fluctuation diagram corresponding to the intersections between these three association rules:

no	LHS of $X \rightarrow heineken$	supp	conf
1	chicken, sardines, coke	11.6	99.0
2	ice cream, soda	13.9	63.2
3	baguette, soda	13.8	90.2

The boxes in the first row (off-diagonal) are all approximately of the same size, indicating that all rules have about the same support. In the diagonal, however, the box corresponding to rule *ice cream, coke* \rightarrow *heineken* is the largest, indicating, that this rule has the lowest confidence of all. Among the three association rules a distinct pattern appears: while the transactions of the first and second rule coincide to a large amount, the third rule is notably different from the first two rules.

4 Summary and Conclusions

The two approaches of visualising the interactions among rules use their own definition of how similar/dissimilar each pair of rules is. There is an arbitrary number of definitions for association measures even for 2×2 tables (see e.g. (Goodman and Kruskal, 1954)). Generally, though, the first approach of plotting a 2d MDS result works better for dissimilarity matrices, whereas Fluctuation Diagrams give a better display for similarity matrices.

Both methods have in common that they only try to visualize association rules and their relationships between each other. It's a different, though related, problem to pick specific association rules out of the set. The plots are supposed to help choosing or validating a choice of rules.

References

- Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining associations between sets of items in massive databases. In *Proc. of the Int'l Conference on Management of Data*, pages 207 – 216, Washington D.C. ACM-SIGMOD.
- Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. Research report rj9839, IBM.
- Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49:732–764.
- Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In *13th Symposium on the Interface*, pages 268–273, New York. Springer Verlag.

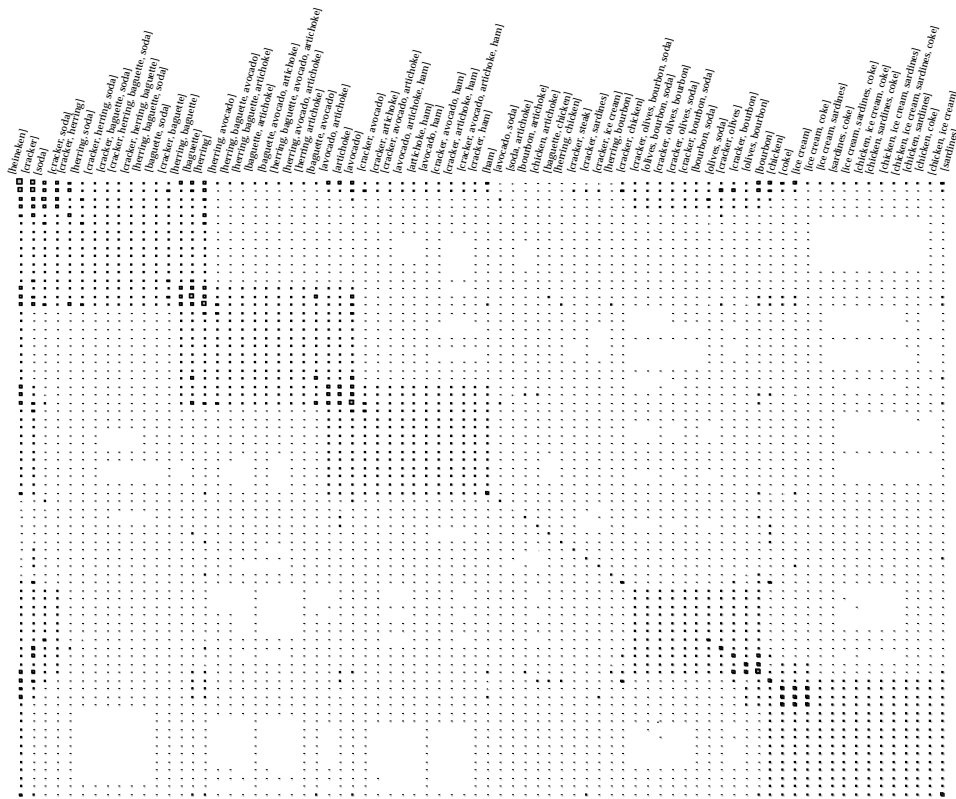


Figure 6: Fluctuation diagram of 79 association rules with the same right hand side. Five distinctive clusters appear, several single association rules are on the diagonal.

Hofmann, H. (2000). Exploring categorical data: interactive mosaic plots. *Metrika*, 51(1):11–26.

Hofmann, H., Siebes, A., and Wilhelm, A. F. (2000). Visualizing association rules with interactive mosaic plots. In *Proc. of the 6th Int'l conf. on Knowledge Discovery and data mining*, pages 227–235, Boston, MA. ACM-SIGKDD.

Liu, B., Hsu, W., and Ma, Y. (1999). Pruning and summarizing the discovered association rules. In *Proc. of the 5th Int'l Conf. on Knowledge Discovery and Data Mining*, pages 125–134. ACM SIGKDD.

Toivonen, H., Klemettinen, M., Ronkainen, P., Hätonen, K., and Mannila, H. (1995). Pruning and grouping discovered association rules. In *MLnet Workshop on Statistics, Machine Learning, and Discovery in Databases*, pages 47–52, Heraklion, Crete, Greece.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire Connecticut.

Unwin, A. R., Hofmann, H., and Bernt, K. (2001). Multiple association rules control. In *Proc. of the 5th Eu-*

ropean Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Freiburg, Germany.