Association Rules (Market Basket Analysis)

Association rules are of the form *if X then Y*. For example: *60% of those who buy comprehensive motor insurance also buy health insurance*; *80% of those who buy books on-line also buy music on-line*; *50% of those who have high blood pressure and are overweight have high cholesterol*. These rules are actionable in that they can be used to target customers for marketing, or for product placing, or more generally to inform decision making. Examples of areas in which association rules have been used include

- Credit card transactions: items purchased by credit card give insight into other products the customer is likely to purchase.
- Supermarket purchases: common combinations of products can be used to inform product placement on supermarket shelves.
- Telecommunication product purchases: commonly associated options (call waiting, caller display, etc) help determine how to structure product bundles which maximise revenue
- Banking services: the patterns of services used by retail customers are used to identify other services they may wish to purchase.
- Insurance claims: unusual combinations of insurance claims can be a sign of fraud.
- Medical patient histories: certain combinations of conditions can indicate increased risk of various complications.

We consider how to derive association rules directly from historical data, as opposed to via customer surveys or other means. Such data is characterised by being readily available in large quantities (and thus cheap), though it is often of poor quality or incomplete.

Theoretical Framework

We will discuss the problem in the context of supermarket purchases, which is where the terminology "market basket analysis" comes from. The data consists of a number of transaction records, each containing a set of items purchased by that customer. For example

Customer	Purchases
1	Tiling Cement; Tiles
2	Paint; White Spirit
3	Paint; Wallpaper; Plaster
4	Paint; Plaster; Tiling Cement; Tiles

We can visualise the data as an array

Customer	Tiling Cement	Tiles	Paint	White Spirit	Wallpaper	Plaster
1	Yes	Yes	No	No	No	No
2	No	No	Yes	Yes	No	No
3	No	No	Yes	No	Yes	Yes
4	Yes	Yes	Yes	No	No	Yes

Let S be the set of all possible purchases and let n be the number of transactions. Each transaction record is a *subset* of S. We consider rules of the form "(x1, x2, …, xj) implies (y1, y2, …, yk)" where

x1, x2, …, y1, y2, … are elements of S. The collection (x1, x2, …, xj) is called an *itemset*; read this as "x1 and x2 and … and xj". The *support* of the rule is defined as

$$\text{Supp}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots)) = \frac{\text{No. transactions containing x1, x2, … and y1, y2, …}}{n}$$

More generally we define the support of an itemset as

$$\text{Supp}(x1, x2, \ldots) = \frac{\text{No. transactions containing x1, x2, …}}{n}$$

The *confidence* of the rule is

$$\text{Conf}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots)) = \frac{\text{Supp}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots))}{\text{Supp}(x1, x2, \ldots)}$$

To consider a rule, we impose a minimum support, indicating a reasonable amount of data about the rule. The confidence measures how good a predictor the rule is. If we specify a minimum support s0 and a minimum confidence c0, then a *strong rule* is one which has Supp((x1, x2, …) implies (y1, y2, …)) > s0 and Conf((x1, x2, …) implies (y1, y2, …)) > c0.

Support and a high confidence do not necessarily mean that a rule is interesting. The *lift* or *improvement* of the rule is

$$\text{Lift}((x1, x2, \ldots) \text{ implies } (y1, y2, \ldots)) = \frac{\text{Supp(x1, x2, … and y1, y2, …)}}{\text{Supp}(x1, x2, \ldots)\ \text{Supp}(y1, y2, \ldots)}$$

The lift is > 1 if the association between (x1, x2, …) and (y1, y2, …) is due to more than just chance. It corresponds to *positive correlation* between the events "purchased x1, x2, …" and "purchased y1, y2, …".

For the data above we have, for example

Supp(Paint implies White Spirit) = 1/4
Conf(Paint implies White Spirit) = 1/3
Lift(Paint implies White Spirit) = 4/3

Supp(Paint and Plaster implies Wallpaper) = 1/4
Conf(Paint and Plaster implies Wallpaper) = 1/2
Lift(Paint and Plaster implies Wallpaper) = 2

Note that this can all be rephrased in terms of conditional probability using counting measure. Let Ω be the set of records and for any single record ω we put $P(\omega) = 1/|\Omega| = 1/n$. Define event E_A to be the set of records containing item set A, then Supp(A) = $P(E_A)$ and Conf(A implies B) = $P(E_B \mid E_A)$. We see that Lift(A implies B) is > 1 if and only if knowing that A is in a record increases the probability that B is in the record.

We would like to find all rules with good lift. In practice there are too many rules to search through for this to be practical, however it turns out that if we restrict ourselves to strong rules then the problem becomes tractable (see the A Priori algorithm below).

An alternative to using the lift to measure the interest of a rule is to use the *significance*. The significance is calculated using a 2*2 contingency table. This gives the observed frequencies of all possible combinations of transactions containing itemsets (x1, x2, …) and transactions containing itemsets (y1, y2, …).

	(x1, x2, …)	Not (x1, x2, …)	Total
(y1, y2, …)	Supp(x1, x2, … and y1, y2, …)	Supp(y1, y2, …) – Supp(x1, x2, … and y1, y2, …)	Supp(y1, y2, …)
Not (y1, y2, …)	Supp(x1, x2, …) – Supp(x1, x2, … and y1, y2, …)	1 – Supp(y1, y2, …) – Supp(x1, x2, …) + Supp(x1, x2, … and y1, y2, …)	1-Supp(y1, y2, …)
Total	Supp(x1, x2, …)	1 – Supp(x1, x2, …)	1

If the occurrence of (x1, x2, …) and (y1, y2, …) is independent, then we would expect Supp(x1, x2, … and y1, y2, …) = Supp(x1, x2, …)* Supp(y1, y2, …). The following chi-squared statistic measures how closely this is achieved

$$T = \frac{n*(Supp(x1, x2, \ldots \text{ and } y1, y2, \ldots) - Supp(x1, x2, \ldots)*Supp(y1, y2, \ldots))^2}{Supp(x1, x2, \ldots)*Supp(y1, y2, \ldots)*(1 - Supp(x1, x2, \ldots))*(1 - Supp(y1, y2, \ldots))}$$

T is a measure of how interesting the rule (x1, x2, …) implies (y1, y2, …) is. Under the null hypothesis that (x1, x2, …) and (y1, y2, …) occur independently, T is approximately chi-squared with 1 degree of freedom. Thus T is significant at the 95% level if its observed value is greater than 3.84146. We take this as an indication that the rule is unlikely to be just a chance effect.

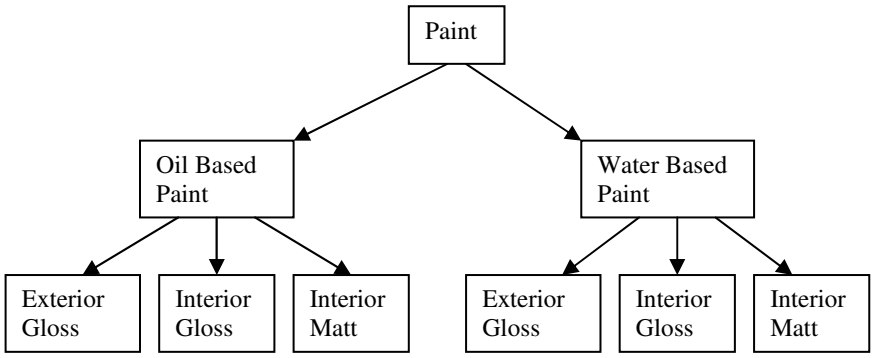
Note that we would expect 5% of all strong rules to be significant through chance alone. So if we are considering a large number of rules we should always expect some of them to be statistical anomalies.

Practical Considerations

Association rules are not always useful, even if they have high support, confidence and lift > 1. For example the rule “Customers who purchase maintenance agreements also purchase large appliances” might have good confidence and lift, but is still not useful. We can classify rules as useful, trivial and inexplicable. Useful rules are the ones we want, with high quality actionable information. Trivial rules will already be known by anyone familiar with the business. Inexplicable rules are those which have no apparent explanation and do not suggest a course of action. An example of the latter is the famous “Men who buy nappies on Thursdays also buy beer” rule.

There is no automatic way of identifying trivial or inexplicable rules. In practice one needs to experiment with the choice of the minimum levels of support and confidence, s0 and c0, to find all the interesting rules without including too many others. Typically rules where the “consequent” (y1, y2, …) consists of a single item y are the most useful.

Association rules can also be improved by combining purchase items. Items often fall into natural hierarchies. For example



In many cases better rules can be obtained by grouping items together according to this taxonomy. That is, rather than consider "red oil based exterior gloss" and "blue oil based exterior gloss" as separate items we combine them as "oil based exterior gloss", or even as "oil based paint" or just "paint". As a rule of thumb, market basket analysis tends to work better when individual items have roughly the same level of support.

Another way of extracting good rules from bad is to consider negations. If the rule "(x1, x2, …) implies (y1, y2, …)" has lift < 1 then the rule "(x1, x2, …) implies not (y1, y2, …)" has lift > 1. One should note however that such a rule is often not actionable, in that it does not lead to a useful course of action.

The "A Priori" Algorithm

Suppose there are a total of m items in S. The number of subsets of S is 2^m, thus to check every transition record to see which sets it belongs to requires $n2^m$ checks. Thus is computationally infeasible when m is even of moderate size. This is an instance of the "curse of dimensionality". However, if we restrict ourselves to sets with support greater than s0 the search becomes feasible. We call these the *frequent itemsets*. This is because most sets have very small support, and because of the fact that for any y

Supp(x1, x2, …, xk and y) is no greater than Supp(x1, x2, …, xk)

This means that when you find a set with small support, you do not need to check any other sets containing all of those items.

The first efficient algorithm for finding all sets with a given level of support was given by Agrawal and Srikant 1994, and was subsequently improved by these authors and others. Once all the sets with support greater than s0 have been found and their supports recorded, it is then a straight forward matter to calculate the confidence, lift and significance of all strong rules of the form "(x1, x2, …) implies (y1, y2, …)", since all of these measures are calculated using the supports of various itemsets.

References

R. Agrawal & R. Srikant, 1994. Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings of the 20th International Conference on Very Large Databases*, pages 487-499.

R. Agrawal, H. Mannila, R. Srikant, H. Toivonen & A.I. Verkamo, 1996. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining* (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy, ed.), American Association for Artificial Intelligence.