Association Rule Analysis: Alzheimer Symptoms

Ralph Schlosser, ralph.schlosser@gmail.com

Introduction

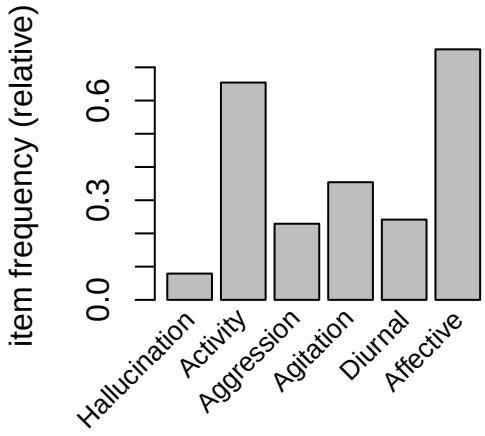
The goal of this homework is to perform an association rule analysis using the `arules` package on the presence / absence of six symptoms found in patients suffering from early onset Alzheimer's disease. The underlying data set is called `Alzheimer` and it is contained in the `BayesLCA` R package.

Analysis

Initial observations Using the `summary` command we can learn that the data set records 240 transactions for the 6 items (symptoms). Also, the symptoms Affective and Activity are the most frequent ones with empirical probabilities $p_{\text{Affective}} = 181/240 \approx 0.75$ and $p_{\text{Activity}} = 157/240 \approx 0.65$. The other probabilities (or proportions) can be reviewed by restricting the `apriori` algorithm to only generate rules of length 1:

```
fit1 <- apriori(Alzheimer.trans,
                parameter = list(confidence = 0, support = 0.1, minlen = 1, maxlen = 1))
```

A visual representation of these proportions can be obtained using the `itemFrequencyPlot` command:



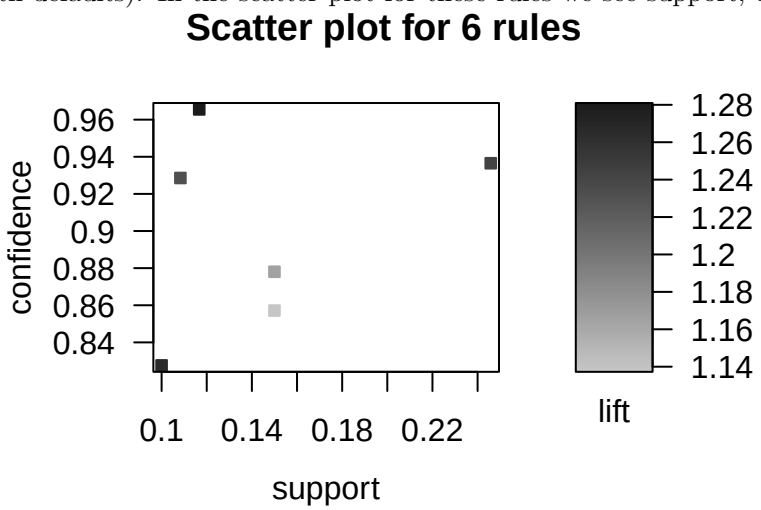
Association rules mining Moving on to rules of length two, we can also add further measures of interest, such as lower and upper *Frechet* bounds for the lift which allow us to interpret the lift value found in relation to it's theoretical boundaries.

When using the default settings (i.e. min. confidence = 0.8 and min. support = 0.1) we get a total of three interesting looking rules:

```
##   lhs             rhs           support confidence     lift        LB       UB
## 1 {Agitation}  => {Affective} 0.3208333  0.9058824 1.201170 0.4055899 1.325967
## 2 {Aggression} => {Affective} 0.1958333  0.8545455 1.133099 0.0000000 1.325967
## 3 {Diurnal}    => {Affective} 0.2041667  0.8448276 1.120213 0.0000000 1.325967
```

The presence of any of the symptoms Agitation, Agression, Diurnal seems to imply that Affective is also present as a symptom. This is supported by the fact that the respective lifts are larger than one and reasonably close to the theoretical upper boundary.

For rules of length three we get six rules with two items on the left hand side and one item on the right hand side (using `apriori` with defaults). In the scatter plot for these rules we see support, confidence and lift



(shading) for each rule.

The top three rules sorted by lift are:

```
##   lhs             rhs           support confidence     lift         LB       UB     chiSq
## 1 {Aggression,
##    Agitation}  => {Affective} 0.1166667  0.9655172 1.280244 0.00000000 1.325967  7.947409
## 2 {Aggression,
##    Agitation}  => {Activity}  0.1000000  0.8275862 1.265100 0.00000000 1.528662  4.384980
## 3 {Activity,
##    Agitation}  => {Affective} 0.2458333  0.9365079 1.241778 0.08418837 1.325967 15.319351
```

Just as before the algorithm yields rules involving Agression, Agitation and Affective, which is unsurprising given that these items (individually) have the highest support in the data set. A new addition to this is a rule involving Activity.

Again we find that the rules produced exhibit a lift higher than 1 which, together with the high confidence, suggests that these rules are interesting ones.

The printout above also has a χ^2 test column (`chiSq`). These are the results of a statistical test of independence of the left and right hand side of the mined rules. At $\alpha = 0.05$ the critical value is $\chi^2 = 3.84$. Higher values, as above, indicate that left and right hand side are *not* independent (for the given α), i.e. the rule is not just a random one. This is another measure of interestingness. Furthermore, it seems there are no rules of length 4 and higher that can be mined from the data set using `apriori`.

Conclusion and discussion

Results We have found that the presence of Agitation, Agression and Diurnal also implies that Affective is present as a symptom. Furthermore, when both Agression and Agitation are found it is very likely to also find symptoms of Affective or Agility. Rules involving four and more items could not be found.

Weaknesses **(1)** More emphasis could have been placed on exploring how lift and other measures of interest vary with changing minimum support and confidence. This information could then have been used in a subsequent step to prune the initially derived rules.

(2) Defining a minimum support leads to rules being omitted, but they could be interesting, e.g. from a medical point of view. For example, there are only 19 entries in the data set involving Hallucination, but we did not find any rules that are "interesting"" according to the given thresholds, *and* contain Hallucination as a symptom.