**arules package**

Support, confidence and lift are three common measures for selecting interesting association rules. Besides them, there are many other interestingness measures, such as chi-square, conviction, gini and leverage [Tan et al., 2002]. More than twenty measures can be calculated with function interestMeasure() in the arules package.

**rules.sorted**

lhs rhs support confidence lift

1 {Class=2nd,Age=Child} => {Survived=Yes} 0.011 1.000 3.096

7 {Class=2nd,Sex=Female,Age=Child} => {Survived=Yes} 0.006 1.000 3.096

4 {Class=1st,Sex=Female} => {Survived=Yes} 0.064 0.972 3.010

10 {Class=1st,Sex=Female,Age=Adult} => {Survived=Yes} 0.064 0.972 3.010

2 {Class=2nd,Sex=Female} => {Survived=Yes} 0.042 0.877 2.716

5 {Class=Crew,Sex=Female} => {Survived=Yes} 0.009 0.870 2.692

11 {Class=Crew,Sex=Female,Age=Adult} => {Survived=Yes} 0.009 0.870 2.692

8 {Class=2nd,Sex=Female,Age=Adult} => {Survived=Yes} 0.036 0.860 2.663

9 {Class=2nd,Sex=Male,Age=Adult} => {Survived=No} 0.070 0.917 1.354

3 {Class=2nd,Sex=Male} => {Survived=No} 0.070 0.860 1.271

12 {Class=3rd,Sex=Male,Age=Adult} => {Survived=No} 0.176 0.838 1.237

6 {Class=3rd,Sex=Male} => {Survived=No} 0.192 0.827 1.222

Some rules generated in the previous section (see rules.sorted, page 93) provide little or no extra information when some other rules are in the result. For example, the above rule 2 provides no extra knowledge in addition to rule 1, since rules 1 tells us that all 2nd-class children survived.

Generally speaking, when a rule (such as rule 2) is a super rule of another rule (such as rule 1) and the former has the same or a lower lift, the former rule (rule 2) is considered to be redundant. Other redundant rules in the above result are rules 4, 7 and 8, compared respectively with rules 3, 6 and 5.

We prune redundant rules with code below. Note that the rules (in rules.sorted) have already been sorted descendingly by lift.

> # find redundant rules

> subset.matrix <- is.subset(rules.sorted, rules.sorted)

> subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA

> redundant <- colSums(subset.matrix, na.rm=T) >= 1

> which(redundant)

{Class=2nd,Sex=Female,Age=Child,Survived=Yes}

{Class=1st,Sex=Female,Age=Adult,Survived=Yes}

{Class=Crew,Sex=Female,Age=Adult,Survived=Yes}

{Class=2nd,Sex=Female,Age=Adult,Survived=Yes}

> # remove redundant rules

> rules.pruned <- rules.sorted[!redundant]

> inspect(rules.pruned)

lhs rhs support confidence lift

1 {Class=2nd,Age=Child} => {Survived=Yes} 0.011 1.000 3.096

4 {Class=1st,Sex=Female} => {Survived=Yes} 0.064 0.972 3.010

2 {Class=2nd,Sex=Female} => {Survived=Yes} 0.042 0.877 2.716

5 {Class=Crew,Sex=Female} => {Survived=Yes} 0.009 0.870 2.692

9 {Class=2nd,Sex=Male,Age=Adult} => {Survived=No} 0.070 0.917 1.354

3 {Class=2nd,Sex=Male} => {Survived=No} 0.070 0.860 1.271

12 {Class=3rd,Sex=Male,Age=Adult} => {Survived=No} 0.176 0.838 1.237

6 {Class=3rd,Sex=Male} => {Survived=No} 0.192 0.827 1.222

In the code above, function is.subset(r1, r2) checks whether r1 is a subset of r2 (i.e., whether r2 is a superset of r1). Function lower.tri() returns a logical matrix with TURE in lower triangle. From the above results, we can see that rules 2, 4, 7 and 8 (before redundancy removal) are successfully pruned.

**9.5 Interpreting Rules**

While it is easy to find high-lift rules from data, it is not an easy job to understand the identified rules. It is not uncommon that the association rules are misinterpreted to find their business meanings. For instance, in the above rule list rules.pruned, the first rule "{Class=2nd, Age=Child} => {Survived=Yes}" has a confidence of one and a lift of three and there are no rules on children of the 1st or 3rd classes. Therefore, it might be interpreted by users as children of the 2nd class had a higher survival rate than other children. This is wrong! The rule states only that all children of class 2 survived, but provides no information at all to compare the survival rates of different classes. To investigate the above issue, we run the code below to find rules whose rhs is "Survived=Yes" and lhs contains "Class=1st", "Class=2nd", "Class=3rd", "Age=Child" and "Age=Adult" only, and which contains no other items (default="none"). We use lower thresholds for both support and confidence than before to find all rules for children of different classes.

> rules <- apriori(titanic.raw,

+ parameter = list(minlen=3, supp=0.002, conf=0.2),

+ appearance = list(rhs=c("Survived=Yes"),

+ lhs=c("Class=1st", "Class=2nd", "Class=3rd",

+ "Age=Child", "Age=Adult"),

+ default="none"),

+ control = list(verbose=F))

> rules.sorted <- sort(rules, by="confidence")

> inspect(rules.sorted)

lhs rhs support confidence lift

1 {Class=2nd,Age=Child} => {Survived=Yes} 0.010904134 1.0000000 3.0956399

2 {Class=1st,Age=Child} => {Survived=Yes} 0.002726034 1.0000000 3.0956399

5 {Class=1st,Age=Adult} => {Survived=Yes} 0.089504771 0.6175549 1.9117275

4 {Class=2nd,Age=Adult} => {Survived=Yes} 0.042707860 0.3601533 1.1149048

3 {Class=3rd,Age=Child} => {Survived=Yes} 0.012267151 0.3417722 1.0580035

6 {Class=3rd,Age=Adult} => {Survived=Yes} 0.068605179 0.2408293 0.7455209

In the above result, the first two rules show that children of the 1st class are of the same survival rate as children of the 2nd class and that all of them survived. The rule of 1st-class children didn’t appear before, simply because of its support was below the threshold specified in Section 9.3. Rule 5 presents a sad fact that children of class 3 had a low survival rate of 34%, which is comparable with that of 2nd-class adults (see rule 4) and much lower than 1st-class adults (see rule 3).