Лабораторная работа 3 по дисциплине “Data Mining”

Ассоциативное моделирование

Используются данные о пассажирах «Титаника».

1. Введите данные в рабочее пространство. (в этом случае Данные «Титаник» содержатся в базовом наборе примеров R, и это делать необязательно)
2. Выведите общие сведения о данных. Укажите, какие имеются параметры, какого они вида, и какие возможные значения принимают. Сделайте несколько (4-5) предположений о том, какие закономерности можно найти в этих данных. Для каждого предположения сформулируйте, как должен быть поставлен исследовательский (аналитический) вопрос о его выяснении.

> str(Titanic)

3. Переведите данные в формат фрейма.

> df<- as.data.frame(Titanic)

4. Создайте чистую таблицу данных “titanic.raw”, создав заголовки из исходных данных, и затем переместив исходные данные в эту таблицу. Проверьте, сколько наблюдений и параметров содержит этот файл (dim())?

Рассматриваются ли категориальные переменные действительно как категориальные?

Обратите внимание, что в новых данных категориальным переменным также присвоены числовые кодировки для возможных значений.

Сделайте общие выводы о наблюдениях в данных (summary).

> titanic.raw <- NULL

> for(i in 1:4) {titanic.raw<-cbind(titanic.raw, rep(as.character(df[,i]), df$Freq))}

> titanic.raw<-as.data.frame(titanic.raw)

> names(titanic.raw)<-names(df)[1:4]

> dim(titanic.raw)

> summary(titanic.raw)

Присвоить имена колонкам можно командой (в данном случае колонки уже озаглавлены, и команда повторяет эти названия)

> names(titanic.raw)<-c(“Class”, “Sex”, “Age”, “Survived”)

5. Выполните поиск ассоциативных закономерностей. Сколько всего закономерностей вы получили? Приведите полученные закономерности в виде списка или таблицы. Приведите несколько ассоциативных правил (5-7), представляющих исследовательский интерес. Приведите несколько ассоциативных правил (3-4), которые не имеют содержательного смысла.

> library(arules)

> rules.all<-apriori(titanic.raw)

> rules.all

> inspect(rules.all)

6. Выполните ассоциативный анализ для поиска закономерностей, характеризующих только выживаемость, и отсортируйте их по убыванию критерия интересности (lift). Сколько было получено правил? Приведите полученные правила в виде списка или таблицы.

> rules <- apriori(titanic.raw, control = list(verbose=F), parameter = list(minlen=2, supp=0.005, conf=0.8), appearance = list(rhs=c("Survived=No", "Survived=Yes"), default="lhs"))

> quality(rules) <- round(quality(rules), digits=3)

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

> inspect(rules.sorted)

7. Из полученных правил ассоциации найдите три пары закономерностей, из которых одна является превосходящей по отношению к другой, и, как следствие, излишней, проведя математическую проверку для выявления таких правил (является ли одно множество подмножеством другого). Опишите эти правила, указав, что конкретно является не предоставляет дополнительной информации для каждого излишнего правила.

(Определение. Излишним признается правило, которое является более общим по отношению к другому, и имеет такой же либо более низкий критерий подъема (lift). Например, правило 2 излишне по отношению к правилу 1, поскольку, согласно правилу 1, все дети из второго класса выжили. Этот пример в отчете о лабораторной использовать уже нельзя ☺ )

> 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)

8. удалите излишние правила на основе этой проверки. Приведите, в виде списка или таблицы, оставшиеся правила.

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

> inspect(rules.pruned)

9. Интерпретируйте оставшиеся правила на основе их содержательного значения либо здравого смысла, с учетом достоверности.

(Например, правило 1 имеет уровень достоверности 1.0 и подъем 3.0, и интерпретируется как «все дети из класса 2 выжили». В то же время, правил по детям из 1 и 3 классов нет. Как следствие, можно неверно интерпретировать правило 1 как то, что дети 2 класса имели больший шанс на выживание, однако, правило всего лишь утверждает, что все дети 2 класса выжили, но не содержит информации о детях других классов, поэтому необходимо получить подобную информацию отдельно – см. задание 10. )

10. Проведите ассоциативный анализ ***выживших*** по 1-3 классам на основе возрастов «ребенок» и «взрослый», исключая прочую информацию:

> 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))

Ответьте на вопросы:

А) отличается ли выживаемость детей из 1 класса от ехавших 2-м?

Б) какой процент детей, ехавших первым классом, выжили?

В) какой процент детей, ехавших третьим классом, выжили?

Г) какой процент взрослых, ехавших вторым классом, выжили? Сопоставьте это значение с числом в вопросе (В)

Д) какой процент взрослых, ехавших первым классом, выжили? Сопоставьте это значение с числами в вопросах (В) и (Г)

11. Представьте данные в виде графика ассоциаций. Приведите 2-3 примера ассоциативных правил из графика. Можно ли утверждать обратное для этих правил? Для каких можно, и для каких – нельзя? Обоснуйте ваш ответ.

> plot(rules.all, method=”graph”)

или

> plot(rules.all, method=”graph”, control=list(type=”items”)

(для выполнения лабораторной приведите оба графика)

12. Напишите аналитическую записку (0,5-1 стр.) о том, какие результаты получены в результате ассоциативного анализа.

lhs rhs support confidence coverage lift

[1] {} => {Age=Adult} 0.9504771 0.9504771 1.0000000 1.0000000

[2] {Class=2nd} => {Age=Adult} 0.1185825 0.9157895 0.1294866 0.9635051

[3] {Class=1st} => {Age=Adult} 0.1449341 0.9815385 0.1476602 1.0326798

[4] {Sex=Female} => {Age=Adult} 0.1930940 0.9042553 0.2135393 0.9513700

[5] {Class=3rd} => {Age=Adult} 0.2848705 0.8881020 0.3207633 0.9343750

[6] {Survived=Yes} => {Age=Adult} 0.2971377 0.9198312 0.3230350 0.9677574

[7] {Class=Crew} => {Sex=Male} 0.3916402 0.9740113 0.4020900 1.2384742

[8] {Class=Crew} => {Age=Adult} 0.4020900 1.0000000 0.4020900 1.0521033

[9] {Survived=No} => {Sex=Male} 0.6197183 0.9154362 0.6769650 1.1639949

[10] {Survived=No} => {Age=Adult} 0.6533394 0.9651007 0.6769650 1.0153856

[11] {Sex=Male} => {Age=Adult} 0.7573830 0.9630272 0.7864607 1.0132040

[12] {Sex=Female, Survived=Yes} => {Age=Adult} 0.1435711 0.9186047 0.1562926 0.9664669

[13] {Class=3rd, Sex=Male} => {Survived=No} 0.1917310 0.8274510 0.2317129 1.2222950

[14] {Class=3rd, Survived=No} => {Age=Adult} 0.2162653 0.9015152 0.2398910 0.9484870

[15] {Class=3rd, Sex=Male} => {Age=Adult} 0.2099046 0.9058824 0.2317129 0.9530818

[16] {Sex=Male, Survived=Yes} => {Age=Adult} 0.1535666 0.9209809 0.1667424 0.9689670

[17] {Class=Crew, Survived=No} => {Sex=Male} 0.3044071 0.9955423 0.3057701 1.2658514

[18] {Class=Crew, Survived=No} => {Age=Adult} 0.3057701 1.0000000 0.3057701 1.0521033

[19] {Class=Crew, Sex=Male} => {Age=Adult} 0.3916402 1.0000000 0.3916402 1.0521033

[20] {Class=Crew, Age=Adult} => {Sex=Male} 0.3916402 0.9740113 0.4020900 1.2384742

[21] {Sex=Male, Survived=No} => {Age=Adult} 0.6038164 0.9743402 0.6197183 1.0251065

[22] {Age=Adult, Survived=No} => {Sex=Male} 0.6038164 0.9242003 0.6533394 1.1751385

[23] {Class=3rd, Sex=Male, Survived=No} => {Age=Adult} 0.1758292 0.9170616 0.1917310 0.9648435

[24] {Class=3rd, Age=Adult, Survived=No} => {Sex=Male} 0.1758292 0.8130252 0.2162653 1.0337773

[25] {Class=3rd, Sex=Male, Age=Adult} => {Survived=No} 0.1758292 0.8376623 0.2099046 1.2373791

[26] {Class=Crew, Sex=Male, Survived=No} => {Age=Adult} 0.3044071 1.0000000 0.3044071 1.0521033

[27] {Class=Crew, Age=Adult, Survived=No} => {Sex=Male} 0.3044071 0.9955423 0.3057701 1.2658514

count

[1] 2092

[2] 261

[3] 319

[4] 425

[5] 627

[6] 654

[7] 862

[8] 885

[9] 1364

[10] 1438

[11] 1667

[12] 316

[13] 422

[14] 476

[15] 462

[16] 338

[17] 670

[18] 673

[19] 862

[20] 862

[21] 1329

[22] 1329

[23] 387

[24] 387

[25] 387

[26] 670

[27] 670

>

lhs rhs support confidence coverage lift count

[1] {Class=2nd, Age=Child} => {Survived=Yes} 0.011 1.000 0.011 3.096 24

[2] {Class=2nd, Sex=Female, Age=Child} => {Survived=Yes} 0.006 1.000 0.006 3.096 13

[3] {Class=1st, Sex=Female} => {Survived=Yes} 0.064 0.972 0.066 3.010 141

[4] {Class=1st, Sex=Female, Age=Adult} => {Survived=Yes} 0.064 0.972 0.065 3.010 140

[5] {Class=2nd, Sex=Female} => {Survived=Yes} 0.042 0.877 0.048 2.716 93

[6] {Class=Crew, Sex=Female} => {Survived=Yes} 0.009 0.870 0.010 2.692 20

[7] {Class=Crew, Sex=Female, Age=Adult} => {Survived=Yes} 0.009 0.870 0.010 2.692 20

[8] {Class=2nd, Sex=Female, Age=Adult} => {Survived=Yes} 0.036 0.860 0.042 2.663 80

[9] {Class=2nd, Sex=Male, Age=Adult} => {Survived=No} 0.070 0.917 0.076 1.354 154

[10] {Class=2nd, Sex=Male} => {Survived=No} 0.070 0.860 0.081 1.271 154

[11] {Class=3rd, Sex=Male, Age=Adult} => {Survived=No} 0.176 0.838 0.210 1.237 387

[12] {Class=3rd, Sex=Male} => {Survived=No} 0.192 0.827 0.232 1.222 422

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

1 2

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

3 4

{Class=2nd,Sex=Female,Survived=Yes} {Class=Crew,Sex=Female,Survived=Yes}

5 6

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

7 8

{Class=2nd,Sex=Male,Age=Adult,Survived=No} {Class=2nd,Sex=Male,Survived=No}

9 10

{Class=3rd,Sex=Male,Age=Adult,Survived=No} {Class=3rd,Sex=Male,Survived=No}

11 12

После 8 пункта ничего не выходит

Пункт 10 – команду взять в круглые скобки, тогда выдаст содержимое

После 11:

