연관규칙의 탐색

Jinseog Kim
Dongguk University
jskim1986@gmail.com

2018-04-09

Contents

1 연관규칙분석이란 3

2 연관규칙분석의 적용 사례 4

3 연관규칙의 생성방법 예시 5

4 연관규칙분석의 측도 6

5 예제 7

6 R 예제 8

7 연습문제 12

1 연관규칙분석이란

- 연관규칙분석은 비지도학습법의 일종으로 자료에 있는 항목(item)들간의 연관규칙을 찾는 방법 - 항목은 상품, 서비스 등을 의미함
- 마케팅에서는 고객의 장바구니에 들어있는 품목간의 관계를 알아본다는 의미에서 장바구니분석(market basket analysis)이라고도 부름
- 활용
  - 효율적인 상품의 진열
  - 패키지 상품의 개발
  - 교차판매 전략
  - 기획상품의 결정

2 연관규칙분석의 적용 사례

1. 백화점, 호텔 등 서비스업에서는 연관규칙을 통하여 고객들이 특정 서비스를 받은 후 다음에 어떤 서비스를 원하는지 미리 알 수 있고,
2. 신용카드사나 은행에서는 고객들의 기존 금융서비스 내역으로부터 대출과 같은 특정한 서비스를 받을 가능성이 높은 고객을 찾을 수 있다.
3. 의료보험이나 손해보험에서는 고객의 보험금 청구가 기존의 정상적인 청구와 다른 패턴을 보이는 경우 보험사기의 징조로 간주하여 추가적인 조사를 하게 된다.
4. 인터넷 쇼핑몰에서 상품의 추천
5. 텍스트마이닝에서 연관 키워드 또는 유사 문서의 추출
6. 웹사이트의 접속자의 페이지간의 이동

3 연관규칙의 생성방법 예시

 예제 데이터

거래번호	품목
1	오렌지쥬스,사이다
2	우유, 오렌지쥬스, 식기세척제
3	오렌지쥬스, 세제
4	오렌지쥬스, 세제, 사이다
5	식기 세척제, 사이다

 동시구매표

	오렌지쥬스	식기세척제	우유	사이다	세제
오렌지쥬스	4	1	1	2	2
식기세척제	1	2	1	1	0
우유	1	1	1	0	0
사이다	2	1	0	3	1
세제	2	0	0	1	2

4 연관규칙분석의 측도

 연관규칙 $X \Rightarrow Y$에 대한 측도

1. 지지도(support): 품목 X의 지지도는 전체 거래에서 X를 포함한 거래 비율로 정의

$$\text{지지도}(X) = P(X) = \frac{\text{X가 포함된 거래수}}{\text{전체 거래수}}$$

2. 신뢰도(confidence): X를 구입한 거래 중 Y를 같이 구입한 비율

$$\text{지지도}(X \Rightarrow Y) = P(Y|X) = \frac{P(X,Y)}{P(Y)} = \frac{\text{품목 X와 Y를 동시에 포함하는 거래수}}{\text{품목 X를 포함하는 거래수}}$$

3. 향상도(lift): $X \Rightarrow Y$의 신뢰도와 Y의 지지도의 비율, 즉 X를 구매했을 때 Y의 구매 비율이 그러한 조건이 없는 경우에 비해 얼마나 커지는가를 나타내는 지표

$$\text{향상도}(X \Rightarrow Y) = \frac{P(Y|X)}{P(Y)} = \frac{conf(X \Rightarrow Y)}{supp(Y)}$$

5 예제

1. 지지도(support)

    - supp(오렌지쥬스) = 4/5
    - supp(사이다) = 3/5
    - supp(식기세척제) = 2/5
    - supp(세제) = 2/5
    - supp(오렌지쥬스, 사이다) = 2/5
    - supp(오렌지쥬스, 세제) = 2/5

2. 신뢰도(confidence): X를 구입한 거래 중 Y를 같이 구입한 비율

    - conf(오렌지쥬스 => 사이다) = (2/5)/(2/5) = 1
    - conf(사이다 => 오렌지쥬스) = (2/5)/(3/5) = 2/3

3. 향상도(lift): X를 구매했을 때 Y의 구매 비율이 그러한 조건이 없는 경우에 비해 얼마나 커지는가?

    - lift(오렌지쥬스 => 사이다) = 1/(3/5) = 5/3 > 1
    - lift(사이다 => 오렌지쥬스) = (2/3)/(4/5) = 5/6 < 1

6 R 예제

- Adult 자료 : 미국 Census Bureau의 Census Income 데이터베이스에서 추출된 설문조사 자료
  - 48,842개의 관측값과 나이, 직업군, 교육정도 등의 15개의 변수(대부분의 변수들이 범주형)
- 지지도, 신뢰도, 향상도 등에 대한 적절한 최소값을 넘는 연관규칙 탐색

```
# install and load arules library
# install.packages("arules")
library(arules)
# adult data
data(Adult)
```

1. association rules with support >= 0.5 &, confidence >= 0.9

```
rules <- apriori(Adult, parameter = list(supp = 0.5, conf = 0.9, target = "rules"))
```

```
## Apriori
##
## Parameter specification:
```

```
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.9    0.1    1 none FALSE            TRUE       5     0.5      1
##  maxlen target   ext
##      10  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 24421
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[115 item(s), 48842 transaction(s)] done [0.02s].
## sorting and recoding items ... [9 item(s)] done [0.00s].
## creating transaction tree ... done [0.02s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [52 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
summary(rules)
```

```
## set of 52 rules
```

```
##
## rule length distribution (lhs + rhs):sizes
##  1  2  3  4
##  2 13 24 13
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.923   3.250   4.000
##
## summary of quality measures:
##     support         confidence          lift            count
##  Min.   :0.5084   Min.   :0.9031   Min.   :0.9844   Min.   :24832
##  1st Qu.:0.5415   1st Qu.:0.9155   1st Qu.:0.9937   1st Qu.:26447
##  Median :0.5974   Median :0.9229   Median :0.9997   Median :29178
##  Mean   :0.6436   Mean   :0.9308   Mean   :1.0036   Mean   :31433
##  3rd Qu.:0.7426   3rd Qu.:0.9494   3rd Qu.:1.0057   3rd Qu.:36269
##  Max.   :0.9533   Max.   :0.9583   Max.   :1.0586   Max.   :46560
##
## mining info:
##   data ntransactions support confidence
##  Adult         48842     0.5        0.9
```

2. association rules having “sex” on LHS with support >=0.4 & lift >= 1

```r
rules.sub = subset(rules, subset = lhs %pin% "sex" & lift > 1)
inspect(sort(rules.sub)[1:3])
```

```
##     lhs                                rhs                                support confidence     l
## [1] {race=White,
##      sex=Male}                      => {native-country=United-States} 0.5415421  0.9204803 1.025
## [2] {sex=Male,
##      native-country=United-States} => {race=White}                    0.5415421  0.9051090 1.058
## [3] {race=White,
##      sex=Male,
##      capital-loss=None}             => {native-country=United-States} 0.5113632  0.9190124 1.024
```

7 연습문제

arules 패키지의 Income 자료에 대하여 연관규칙을 탐색하라