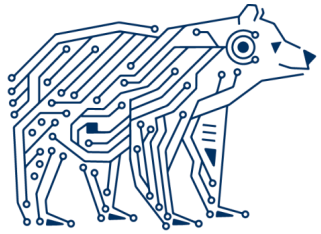


데이터마이닝 과제 1

Bank Service Data에서의 Association Rule 도출

고려대학교 통계학과 17학번

이재승

CONTENTS

데이터마이닝 과제 1

Bank Service Data 소개

Service	Description
ATM	Automated teller machine debit card
AUTO	Automobile installment loan
CCRD	Credit card
CD	Certificate of deposit
CKCRD	Check/debit card
CKING	Checking account
HMEQLC	Home equity line of credit
IRA	Individual retirement account
MMAD	Money market deposit account
MTG	Mortgage
PLOAN	Personal/consumer installment
SVG	Saving account
TRUST	Personal trust account

Bank Service Data

- To identify services that customers have at the same time

- 은행에서는 예금이체, 통장 확인, 신용카드, 대출 등 여러 서비스들을 이용할 수 있음.

→ 본 데이터에서는 크게 13종류의 서비스가 있음.

- '고객들이 은행을 이용하는데 있어 동시에 이용하는 서비스가 무엇인가?'를 파악하기 위한 데이터

Bank Service Data 소개 – BNKSERV

	A	B
1	ACCT	SERVICE
2	500026	CKING
3	500026	SVG
4	500026	ATM
5	500075	CKING
6	500075	MMDA
7	500075	SVG
8	500075	ATM
9	500075	TRUST
10	500129	CKING

24370	999881	IRA
24371	999881	AUTO
24372	999938	CKING
24373	999938	ATM
24374	999949	CKING
24375	999949	SVG
24376	999949	CD

Bank Service Data - BNKSERV

- **데이터의 Column : ACCT 및 SERVICE**

→ ACCT : 계좌 정보를 의미 (사람)

→ SERV : 해당 사람이 어떠한 서비스를 이용하였는가를 의미

Ex) 500026이라는 Account를 가진 사람은 CKING, SVG, ATM 서비스를 이용하였다.

- 총 데이터 : 24,375개

- Excel .csv 형태로 데이터가 존재

R을 이용한 데이터 탐색

arules 및 arulesviz 라이브러이 이용

```
> library(arules)
> library(arulesViz)
> # Reading Data
> BNKSERV = read.transactions("C:/Users/jason/바탕 화면/coding1/data_mining/Assignment/assignment1/
BNKSERV.csv", format = "single", cols = c(1,2), sep=",", skip=1, rm.duplicate=TRUE)
> inspect(BNKSERV)
      items                                                transactionID
[1]   {ATM, CKING, SVG}                                    500026
[2]   {ATM, CKING, MMDA, SVG, TRUST}                       500075
[3]   {ATM, CKING, IRA, SVG}                               500129
[4]   {CKCRD, CKING, SVG}                                  500256
[5]   {CKCRD, CKING, SVG}                                  500341
[6]   {CD, CKING}                                          500350
[7]   {ATM, SVG}                                           500458
[8]   {CD, CKING, SVG, TRUST}                              500595
[9]   {CCRD, CKCRD, CKING, HMEQLC, MTG, SVG}               500743
[10]  {CD, CKING}                                          500744
```

- Arules에서 inspect 함수를 이용하면 데이터 확인 가능

- Excel .csv 형태의 데이터를 읽어오고 데이터 확인

- **Excel .csv 형태에서 데이터를 읽었을 때와 다르게, 어떤 사람이 어떠한 서비스를 이용하였는지 한 눈에 확인 가능**

Ex) 50026이라는 Transaction ID를 가진 사람은 {ATM, CKING, SVG}라는 서비스를 이용하였다.

R을 이용한 데이터 탐색

```
> str(BNKSERV)
Formal class 'transactions' [package "arules"] with 3 slots
  ..@ data       :Formal class 'ngCMatrix' [package "Matrix"] with 5 slots
  .. .. ..@ i       : int [1:24375] 0 5 11 0 5 8 11 12 0 5 ...
  .. .. ..@ p       : int [1:7992] 0 3 8 12 15 18 20 22 26 32 ...
  .. .. ..@ Dim     : int [1:2] 13 7991
  .. .. ..@ Dimnames:List of 2
  .. .. .. ..$ : NULL
  .. .. .. ..$ : NULL
  .. .. ..@ factors : list()
  ..@ itemInfo   :'data.frame': 13 obs. of  1 variable:
  .. ..$ labels: chr [1:13] "ATM" "AUTO" "CCRD" "CD" ...
  ..@ itemsetInfo:'data.frame': 7991 obs. of  1 variable:
  .. ..$ transactionID: chr [1:7991] "500026" "500075" "500129" "500256" ...
```

```
> as(BNKSERV, "data.frame")[1:10,]
                                items transactionID
1                      {ATM,CKING,SVG}        500026
2           {ATM,CKING,MMDA,SVG,TRUST}        500075
3                  {ATM,CKING,IRA,SVG}        500129
4                    {CKCRD,CKING,SVG}        500256
5                    {CKCRD,CKING,SVG}        500341
6                           {CD,CKING}        500350
7                            {ATM,SVG}        500458
8                  {CD,CKING,SVG,TRUST}        500595
9  {CCRD,CKCRD,CKING,HMEQLC,MTG,SVG}        500743
10                          {CD,CKING}        500744
```

Bank Service Data - BNKSERV

- R의 str 및 as를 이용한 데이터 탐색

- Bank Service인 Label은 13개가 있음.

- 전체 Transaction ID의 수는 7,991개임.

→ 즉, 전체 관측치 개수는 24,375개 중 똑같은 사람이 여러 서비스를 이용한 것을 모두 처리해주면 Bank Service를 이용한 전체 사람 수는 7,991명이라는 것을 의미함.

- Data에서 Column의 수는 2개임.

→ items 및 transactionID

특정 조건을 만족하는 Association Rule 구축

```
> rules = apriori(BNKSERV, parameter=list(support=0.1, confidence=0.7, minlen=2), control=list(verbose=F))
> rules.sorted = sort(rules, by=c("support","lift")) #sorting data
> inspect(rules.sorted)
     lhs              rhs     support   confidence coverage  lift     count
[1]  {SVG}         => {CKING} 0.5417345 0.8756068  0.6186960 1.020711 4329
[2]  {ATM}         => {CKING} 0.3619071 0.9410999  0.3845576 1.097058 2892
[3]  {ATM, SVG}    => {CKING} 0.2485296 0.9673648  0.2569140 1.127675 1986
[4]  {CD}          => {CKING} 0.2098611 0.8556122  0.2452759 0.997403 1677
[5]  {HMEQLC}      => {CKING} 0.1646853 1.0000000  0.1646853 1.165718 1316
[6]  {MMDA}        => {CKING} 0.1558003 0.8931133  0.1744463 1.041119 1245
[7]  {CCRD}        => {CKING} 0.1485421 0.9595796  0.1547991 1.118600 1187
[8]  {CD, SVG}     => {CKING} 0.1425354 0.9068471  0.1571768 1.057128 1139
[9]  {CKCRD}       => {CKING} 0.1130021 1.0000000  0.1130021 1.165718  903
[10] {HMEQLC, SVG} => {CKING} 0.1115004 1.0000000  0.1115004 1.165718  891
```

- **Support 0.1 이상, Confidence 0.7 이상, 최소 길이 2 이상을 만족하도록 Association Rule을 구축**

- Association Rule을 구축한 이후 Support → Lift 순으로 높은 순서대로 내림차순 정렬

→ Inspect를 통해 Top 10 데이터를 확인할 수 있음.

Association Rule 해석 (1)

```
> rules = apriori(BNKSERV, parameter=list(support=0.1, confidence=0.7, minlen=2), control=list(verbose=F))
> rules.sorted = sort(rules, by=c("support","lift")) #sorting data
> inspect(rules.sorted)
     lhs                rhs      support   confidence coverage  lift     count
[1]  {SVG}           => {CKING} 0.5417345 0.8756068  0.6186960 1.020711 4329
[2]  {ATM}           => {CKING} 0.3619071 0.9410999  0.3845576 1.097058 2892
[3]  {ATM, SVG}      => {CKING} 0.2485296 0.9673648  0.2569140 1.127675 1986
[4]  {CD}            => {CKING} 0.2098611 0.8556122  0.2452759 0.997403 1677
[5]  {HMEQLC}        => {CKING} 0.1646853 1.0000000  0.1646853 1.165718 1316
[6]  {MMDA}          => {CKING} 0.1558003 0.8931133  0.1744463 1.041119 1245
[7]  {CCRD}          => {CKING} 0.1485421 0.9595796  0.1547991 1.118600 1187
[8]  {CD, SVG}       => {CKING} 0.1425354 0.9068471  0.1571768 1.057128 1139
[9]  {CKCRD}         => {CKING} 0.1130021 1.0000000  0.1130021 1.165718  903
[10] {HMEQLC, SVG}   => {CKING} 0.1115004 1.0000000  0.1115004 1.165718  891
```

[1] {SVG} => {CKING}에 대한 결과 해석

- **Support = 0.541** → 전체 사람 중 절반 이상이 SVG와 CKING을 같이 이용했다.
- **Confidence = 0.875** → SVG를 이용하였을 때, 이후 CKING을 이용할 확률은 87.5%이다.
- **Coverage = 0.618** → 전체 데이터 중 SVG를 이용한 경우는 전체의 61.8%다.
- **Lift = 1.020** → Lift가 1보다 크므로 Positive Association임. 즉, SVG를 이용하면 CKING을 이용하는 경우가 많다.
- **Count = 4,329** → SVG와 CKING을 같이 이용한 경우의 수. 즉, (4,329 / 7,991) = 0.541이라서 Support가 0.541이 나온 것이다.

Association Rule 해석 (2)

```
> rules = apriori(BNKSERV, parameter=list(support=0.1, confidence=0.7, minlen=2), control=list(verbose=F))
> rules.sorted = sort(rules, by=c("support","lift")) #sorting data
> inspect(rules.sorted)
     lhs                rhs      support   confidence coverage  lift     count
[1]  {SVG}           => {CKING} 0.5417345 0.8756068  0.6186960 1.020711 4329
[2]  {ATM}           => {CKING} 0.3619071 0.9410999  0.3845576 1.097058 2892
[3]  {ATM, SVG}      => {CKING} 0.2485296 0.9673648  0.2569140 1.127675 1986
[4]  {CD}            => {CKING} 0.2098611 0.8556122  0.2452759 0.997403 1677
[5]  {HMEQLC}        => {CKING} 0.1646853 1.0000000  0.1646853 1.165718 1316
[6]  {MMDA}          => {CKING} 0.1558003 0.8931133  0.1744463 1.041119 1245
[7]  {CCRD}          => {CKING} 0.1485421 0.9595796  0.1547991 1.118600 1187
[8]  {CD, SVG}       => {CKING} 0.1425354 0.9068471  0.1571768 1.057128 1139
[9]  {CKCRD}         => {CKING} 0.1130021 1.0000000  0.1130021 1.165718  903
[10] {HMEQLC, SVG}   => {CKING} 0.1115004 1.0000000  0.1115004 1.165718  891
```

[3] {ATM, SVG} => {CKING}에 대한 결과 해석

- **Support = 0.248** → ATM, SVG와 CKING을 같이 이용한 사람은 전체 중 24.8%이다.
- **Confidence = 0.967** → ATM과 SVG를 이용하였을 때, 이후 CKING을 이용할 확률은 96.7%에 이른다.
- **Coverage = 0.256** → 전체 데이터 중 ATM과 SVG를 이용한 경우는 전체의 25.6%다.
- **Lift = 1.127** → Lift가 1보다 크므로 Positive Association임. 즉, ATM과 SVG를 이용하면 CKING을 이용하는 경우가 많다.
- **Count = 1,986** → ATM, SVG, CKING을 같이 이용한 경우의 수. 즉, (1,996 / 7,991) = 0.248이라서 Support가 0.248이 나온 것이다.

Association Rule 해석 (3)

```
> rules = apriori(BNKSERV, parameter=list(support=0.1, confidence=0.7, minlen=2), control=list(verbose=F))
> rules.sorted = sort(rules, by=c("support","lift")) #sorting data
> inspect(rules.sorted)
     lhs                rhs      support   confidence coverage  lift     count
[1]  {SVG}           => {CKING} 0.5417345 0.8756068  0.6186960 1.020711 4329
[2]  {ATM}           => {CKING} 0.3619071 0.9410999  0.3845576 1.097058 2892
[3]  {ATM, SVG}      => {CKING} 0.2485296 0.9673648  0.2569140 1.127675 1986
[4]  {CD}            => {CKING} 0.2098611 0.8556122  0.2452759 0.997403 1677
[5]  {HMEQLC}        => {CKING} 0.1646853 1.0000000  0.1646853 1.165718 1316
[6]  {MMDA}          => {CKING} 0.1558003 0.8931133  0.1744463 1.041119 1245
[7]  {CCRD}          => {CKING} 0.1485421 0.9595796  0.1547991 1.118600 1187
[8]  {CD, SVG}       => {CKING} 0.1425354 0.9068471  0.1571768 1.057128 1139
[9]  {CKCRD}         => {CKING} 0.1130021 1.0000000  0.1130021 1.165718  903
[10] {HMEQLC, SVG}   => {CKING} 0.1115004 1.0000000  0.1115004 1.165718  891
```

[4] {CD} => {CKING}에 대한 결과 해석

- **Support = 0.209** → CD와 CKING을 같이 이용한 사람은 전체 중 24.8%이다.
- **Confidence = 0.855** → CD를 이용하였을 때, 이후 CKING을 이용할 확률은 85.5%이다.
- **Coverage = 0.245** → 전체 데이터 중 CD를 이용한 경우는 전체의 24.5%다.
- **Lift = 0.997** → 엄밀히 말하면 Lift가 1보다 작기에 Negative Association이기에 CD를 이용하면 CKING을 이용하지 않는다. 하지만, 0.997의 경우 1에 굉장히 근사하는 숫자이기 때문에 No Association이라고도 볼 수 있다. Confidence가 0.855로 높은데 비해 Lift가 1에 굉장히 근사하기 때문에 Association이 Random하다고도 볼 수 있다.
- **Count = 1,677** → CD와 CKING을 같이 이용한 경우의 수. 즉, (1,677 / 7,991) = 0.209라서 Support가 0.209가 나온 것이다.

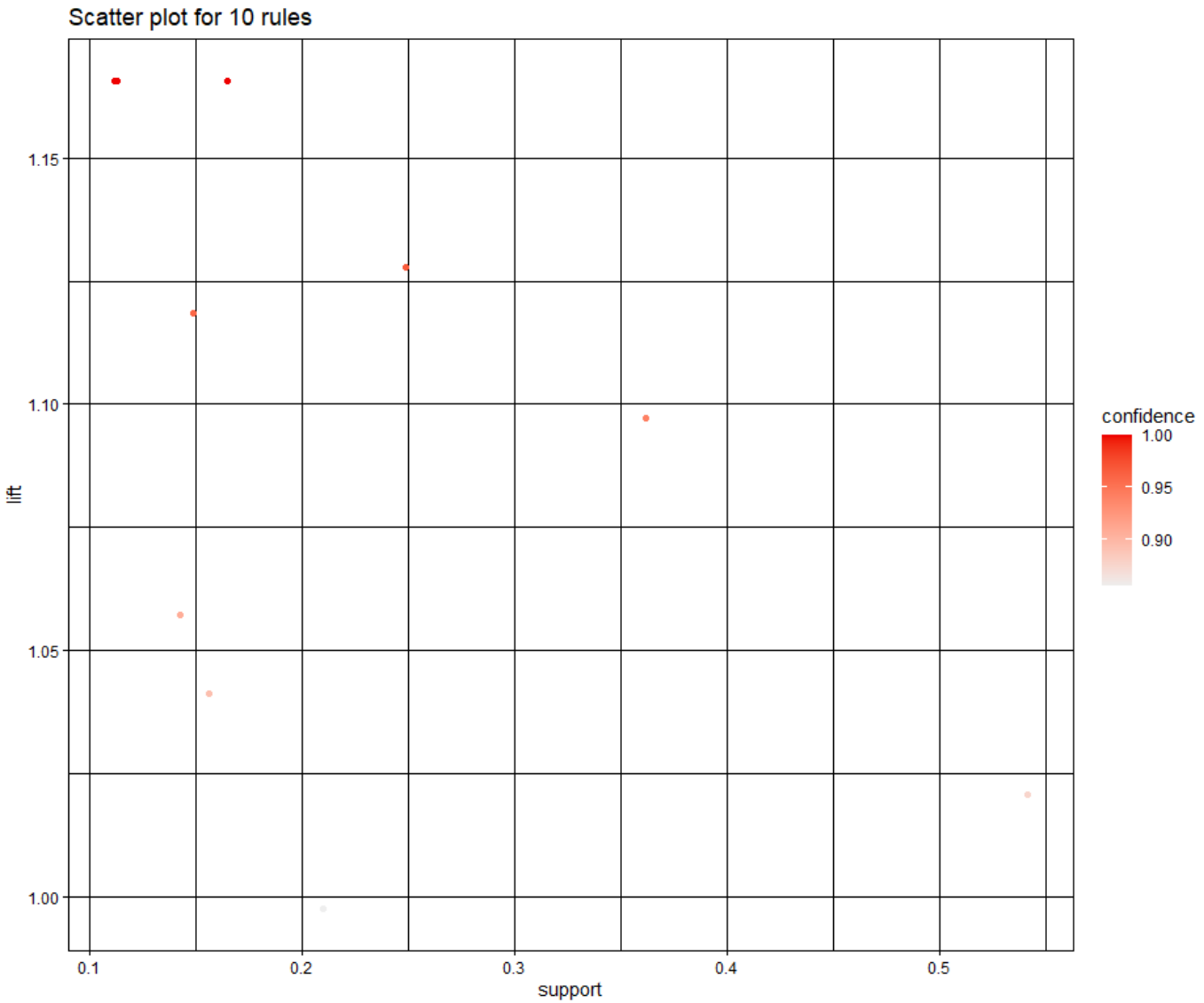
Association Rule 개수

```
> rules = apriori(BNKSERV, parameter=list(support=0.1, confidence=0.7, minlen=2), control=list(verbose=F))
> rules.sorted = sort(rules, by=c("support","lift")) #sorting data
> inspect(rules.sorted)
     lhs              rhs      support   confidence coverage  lift     count
[1]  {SVG}         => {CKING} 0.5417345 0.8756068  0.6186960 1.020711 4329
[2]  {ATM}         => {CKING} 0.3619071 0.9410999  0.3845576 1.097058 2892
[3]  {ATM, SVG}    => {CKING} 0.2485296 0.9673648  0.2569140 1.127675 1986
[4]  {CD}          => {CKING} 0.2098611 0.8556122  0.2452759 0.997403 1677
[5]  {HMEQLC}      => {CKING} 0.1646853 1.0000000  0.1646853 1.165718 1316
[6]  {MMDA}        => {CKING} 0.1558003 0.8931133  0.1744463 1.041119 1245
[7]  {CCRD}        => {CKING} 0.1485421 0.9595796  0.1547991 1.118600 1187
[8]  {CD, SVG}     => {CKING} 0.1425354 0.9068471  0.1571768 1.057128 1139
[9]  {CKCRD}       => {CKING} 0.1130021 1.0000000  0.1130021 1.165718  903
[10] {HMEQLC, SVG} => {CKING} 0.1115004 1.0000000  0.1115004 1.165718  891
> rules.sorted
set of 10 rules
```

지정한 조건을 만족하는 Association Rule의 개수는 10개

- 지정한 조건 : Support 0.1 이상, Confidence 0.7 이상, 최소 길이 2 이상
- 지정한 조건을 만족하는 전체 Association Rule의 개수는 10개로 도출되었기 때문에, 위와 같은 Association Rule을 도출할 수 있음.

Association Rule을 Plot으로 시각화



Plot으로 시각화

- 10개의 Association Rule들에 대하여 Scatter Plot으로 시각화

- X축 : Support / Y축 : Lift / 색깔의 진함 정도 : Confidence

→ Association Rule을 통해 도출되는 Support, Confidence, Coverage, Lift, Count – 5개 중 3개를 한눈에 확인해볼 수 있음.

- 좌상단에 점들이 꽤 있음을 확인할 수 있음.

→ Support는 낮고 Lift는 높은 경우

감사합니다