개인정보 비 식별화 (De-identification)

인하대학교

데이터사이언스 학과

김 승 환

swkim4610@inha.ac.kr

비 식별화 목표

구분	지역코드	연령	성별	질병
1	13053	28	남	전립선염
2	13068	21	남	전립선염
3	13068	29	여	고혈압
4	13053	23	남	고혈압
5	14853	50	여	위암
6	14853	47	남	전립선염
7	14850	55	여	고혈압
8	14850	49	남	고혈압
9	13053	31	남	위암
10	13053	37	여	위암
11	13068	36	남	위암
12	13068	35	여	위암

K-익명성 적용

데이터집합에 같은값이 K개(4) 이상 존재

구분	지역코드	연령	성별	질병	비고
1	130**	< 30	*	전립선염	다양한 질병이 혼재되어 안전
2	130**	< 30	*	전립선염	
3	130**	< 30	*	고혈압	
4	130**	< 30	*	고혈압	
5	1485*	> 40	*	위암	다양한 질병이 혼재되어 안전
6	1485*	> 40	*	전립선염	
7	1485*	> 40	*	고혈압	
8	1485*	> 40	*	고혈압	
9	130**	3*	*	위암	모두가 동일 질병(위암)으로 취약
10	130**	3*	*	위암	
11	130**	3*	*	위암	
12	130**	3*	*	위암	

information loss 증가, Data Quality 저하

privacy level 강화

비식별 처리는 익명화 수준과 분석에 필요한 정보량이 조화를 이루는 변환

information loss

Level-0	Level-1	Level-2	Level-3	Level-4
21	[20, 25[	[20, 30[	[20, 40[	*
22	[20, 25[	[20, 30[	[20, 40[	*
23	[20, 25[	[20, 30[	[20, 40[	*
24	[20, 25[	[20, 30[	[20, 40[	*
31	[30, 35[	[30, 40[	[20, 40[	*
32	[30, 35[	[30, 40[	[20, 40[	*
33	[30, 35[	[30, 40[	[20, 40[	*
34	[30, 35[	[30, 40[	[20, 40[	*
41	[40, 45[	[40, 50[	[40, 60[	*
42	[40, 45[	[40, 50[	[40, 60[	*
43	[40, 45[	[40, 50[	[40, 60[	*
44	[40, 45[	[40, 50[	[40, 60[	*
51	[50, 55[	[50, 60[	[40, 60[	*
52	[50, 55[	[50, 60[	[40, 60[	*
53	[50, 55[	[50, 60[	[40, 60[	*

위의 그림은 나이를 그룹화 변환하는 과정이다. 나이가 준식별자라고 가정할 때,

Level-0는 원 나이이고, Level-1은 5세, Level-2는 10세 단위로 범주화할 수 있다.

범주화는 정보 손실을 통해 정보를 은닉할 수 있어 프라이버시 보호에 도움이 되지만, 분석에는 부정적인 영향을 준다.

비 식별화 방법

구분	지역코드	연령	성별	질병
1	13053	28	남	전립선염
2	13068	21	남	전립선염
3	13068	29	여	고혈압
4	13053	23	남	고혈압
5	14853	50	여	위암
6	14853	47	남	전립선염
7	14850	55	여	고혈압
8	14850	49	남	고혈압
9	13053	31	남	위암
10	13053	37	여	위암
11	13068	36	남	위암
12	13068	35	여	위암

K-익명성 적용

데이터집합에 같은값이 K개(4) 이상 존재

구분	지역코드	연령	성별	질병	비고
1	130**	< 30	*	전립선염	다양한 질병이 혼재되어 안전
2	130**	< 30	*	전립선염	
3	130**	< 30	*	고혈압	
4	130**	< 30	*	고혈압	
5	1485*	> 40	*	위암	다양한 질병이 혼재되어 안전
6	1485*	> 40	*	전립선염	
7	1485*	> 40	*	고혈압	
8	1485*	> 40	*	고혈압	
9	130**	3*	*	위암	모두가 동일 질병(위암)으로 취약
10	130**	3*	*	위암	
11	130**	3*	*	위암	
12	130**	3*	*	위암	

위의 그림은 4-익명성을 만족하는 변환을 수행한 결과이다.

4-익명성이란 비식별화 된 자료에서 개인을 식별할 확률의 최대값이 25%라는 의미이다.

5-익명성은 20%, 10 익명성은 10%가 된다.

그렇다면 k는 어느 정도가 적당한가?

답은 없다. 하지만, 통상적으로 k는 5 ~ 10 사이 값을 많이 사용한다.

k=5라고 해서 모든 레코드의 식별 가능성이 20%라는 것은 아니다. 최악의 경우를 산정한 것이다.

비 식별화 절차

Step 1: Unique 한 식별자가 분석에 꼭 필요한 경우가 아니면 제거한다.

Step 2: 준 식별자 중에서 분석에 불필요한 것은 제거한다.

준 식별자가 증가하면 Population Uniqueness Issue가 발생한다.

Step 3: 준식별자에 대해 빈도분석을 수행하여 빈도가 너무 작은 그룹(<k)이 있다면 Pooling 혹은 삭제 등 적절한 조치를 취해야 한다. (준식별자 하나씩 빈도분석을 실시하고 다음으로는 준식별자 조합에서도 실시)

Step 4: 민감 정보가 포함되어 있는 경우, l-다양성이나 t-근접성 모형을 반드시 적용해야 한다.

처음에는 민감정보로 지정하지 않은 상태에서 적절한 k를 찾고 이후, 민감정보로 지정한 후, l 값이나 t 값을 만족시키는 변환을 수행하는 것이 수월하다.

Step 5: 원하는 수준의 변환이 이루어지지 않을 경우, 변환에 이슈가 되는 레코드를 삭제하는 것도 좋은 방법이다.

변수 변환이나 레코드 삭제나 모두 정보손실 이므로 삭제되는 레코드가 분석에 어떤 영향을 주는지 판단해야 한다.

Step 6: 변환이 완료되면 Risk Analysis를 통해 이 변환의 리스크를 정량적으로 계산하고 이에 따라 최종 변환 종료 여부를 결정한다.

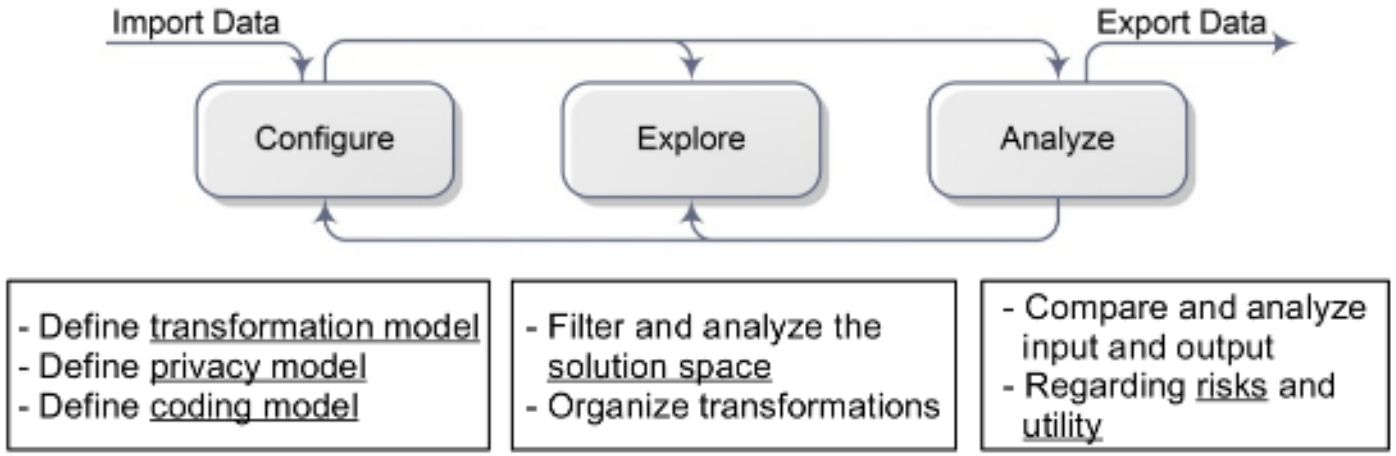
Population Uniqueness Issue 예
직업: 가수
거주지: 제주도 애월읍
나이: 40대
성별: 여자
결혼여부: 기혼

ARX 소개

ARX 는 비식별화를 수행하는 오픈소스 프로그램이다. (http://arx.deidentifier.org)

ARX외에도 많은 비식별화 프로그램이 존재한다.

ARX 비식별화 프로세스는 아래의 세가지 단계로 구성되어 있다.



첫 단계는 Raw Data를 Import하여 데이터 변환모형과 프라이버시 모형을 설정하는 Configure 단계이고

두번째 단계는 설정된 모형을 만족하는 모든 가능한 변환을 도식화하여 보여주는 Explore 기능이다.

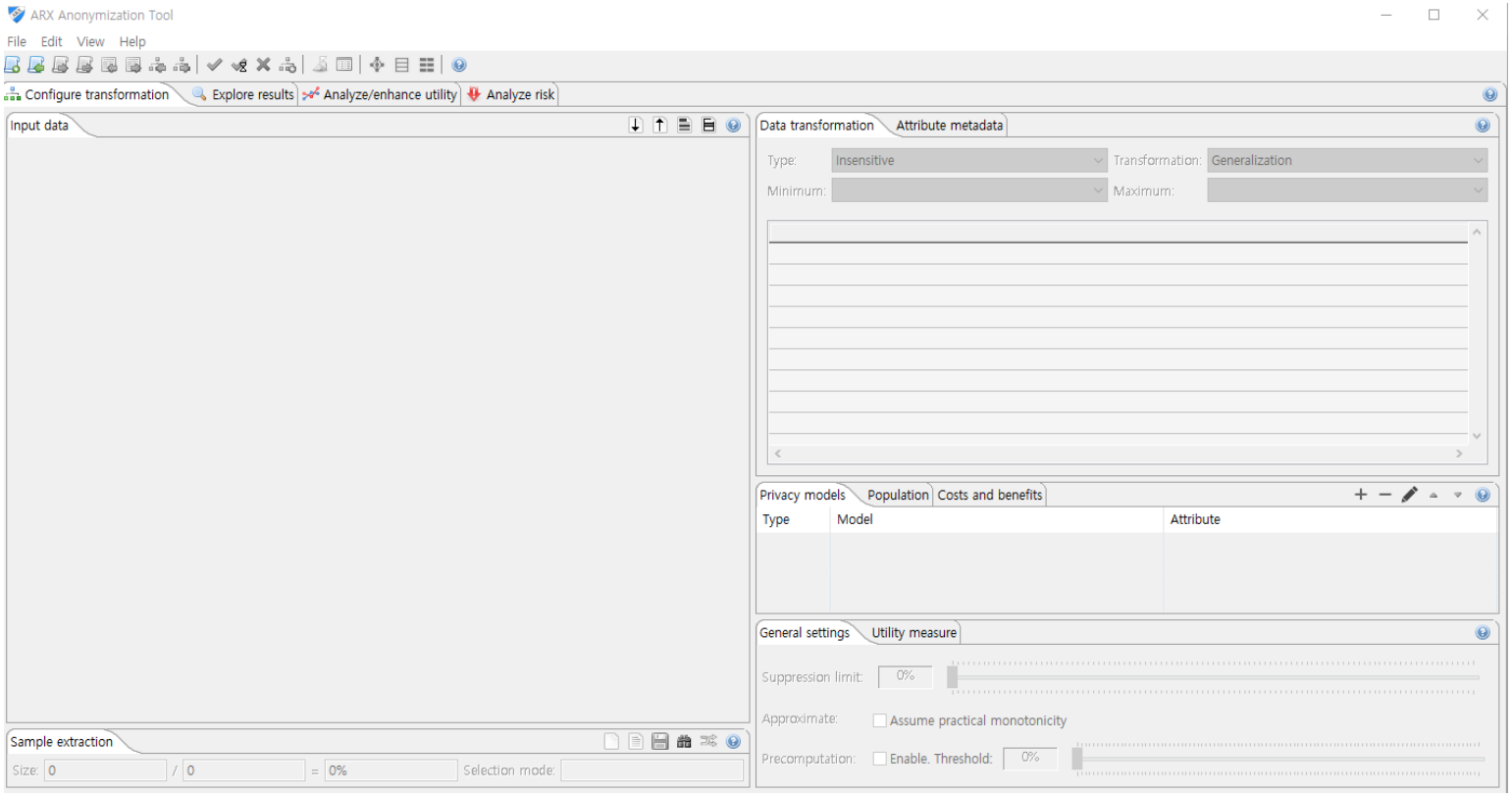
Explore 단계에서 적절한 변환모형을 선택하였다면 세번째 단계인 Analyze 단계로 넘어간다.

Analyze 단계에서는 재식별화 가능성 등 위험수준을 분석하여 최종 Export 여부를 결정하게 된다.

ARX 설치

http://arx.deidentifier.org/ 에서 Installer 중 자신의 O.S.에 맞는 파일을 다운로드한다.

설치가 성공하면 시작메뉴에 ARX 메뉴가 생긴다. 아래는 ARX를 실행한 초기 화면이다.



ARX-Configure

New Project 생성:

ARX의 시작은 새로운 프로젝트를 생성하는 것이다.

File → New Project 를 통해 새로운 프로젝트를 만들고 Raw Data를 Import한다.

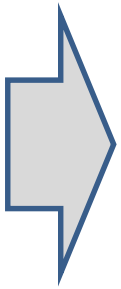
Data는 CSV, Excel, Database 형식으로 Import 할 수 있다.

CSV 파일은 아래와 같은 형식으로 만들고 Import 하면 된다.

Import하면 Input Data 창에 우측과 같이 Data가 나타난다.

Import 할 때에는 숫자형/문자형을 구분하여야 한다. 이 예에서 sex, loc는 문자, age, salary는 숫자이다.

```
sex,age,loc,salary
M,21,강원,1000
M,22,강원,1500
F,23,강원,1500
F,24,강원,1200
M,31,충청,2000
M,32,충청,2300
F,33,충청,2400
...
...
```



Configure transformation | Explore results | Analyze/enhance utility | An

Input data

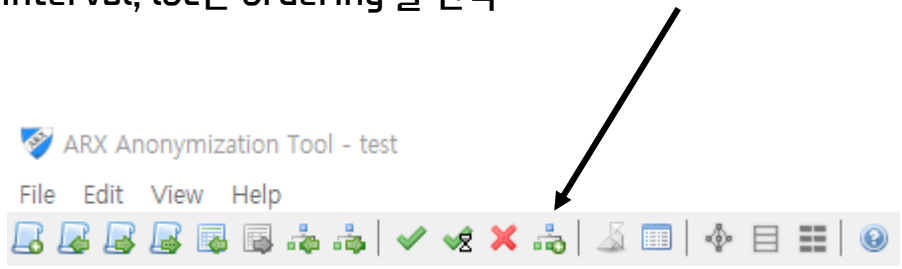
	sex	age	loc	salary
1	M	21	강원	1000
2	M	22	강원	1500
3	F	23	강원	1500
4	F	24	강원	1200
5	M	31	충청	2000
6	M	32	충청	2300
7	F	33	충청	2400
8	F	34	충청	2100
9	M	41	경기	3000
10	M	42	경기	3200
11	F	43	경기	3200
12	F	44	경기	3300

데이터 파일: https://drive.google.com/open?id=1q3hI9WC7qFcI6TlF-T6XhmxmlhcOJqYx

ARX-Configure

다음 단계는 Data Transformation 이다.

- 식별자(Identifier)는 "*" 로 처리
- sex, age, loc은 준식별자(Quasi-identifier) 로 지정
- salary는 Insensitive 혹은 Sensitive로 지정. 우선, salary는 Insensitive로 지정해보자.
- 방법: 좌측에서 sex를 선택하고 우측에서 type은 Quasi-identifying, transformation은 generalization을 선택 aggregation은 데이터를 평균이나 합 등으로 변환 하는 것이고 generalization은 15세→10대의 형식으로 변환하는 것임
- type과 transformation을 선택 후, Transformation Hierarchy를 만들기 위해 아래 화살표 방향 메뉴버튼 클릭 use interval, ordering, masking의 메뉴가 나타난다.
  sex의 경우는 masking을 선택한다.
  같은 방식으로 age는 Interval, loc는 ordering 을 선택



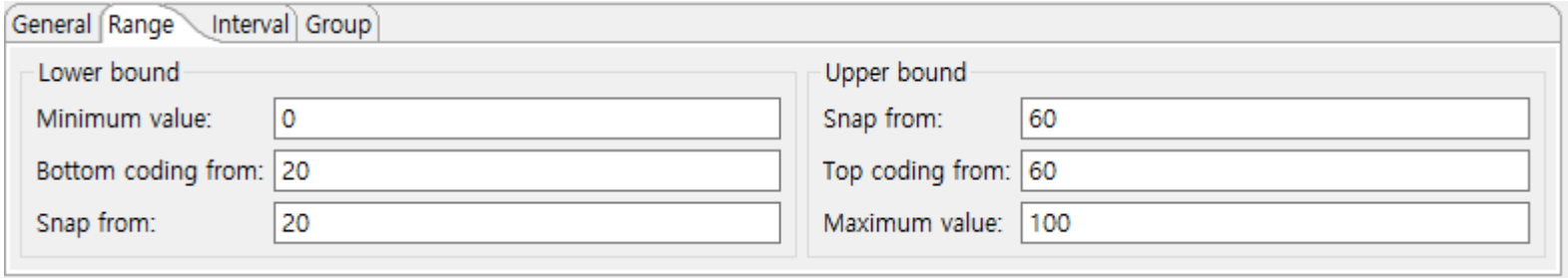
ARX-Configure

age interval은 아래와 같이 만든다.

Input Data 창에서 age를 선택하고 Create hierarchy 버튼을 클릭한 다음 Use Interval을 선택한다.

아래의 창에서 Lower, Upper Bound는 아래와 같이 입력한다.

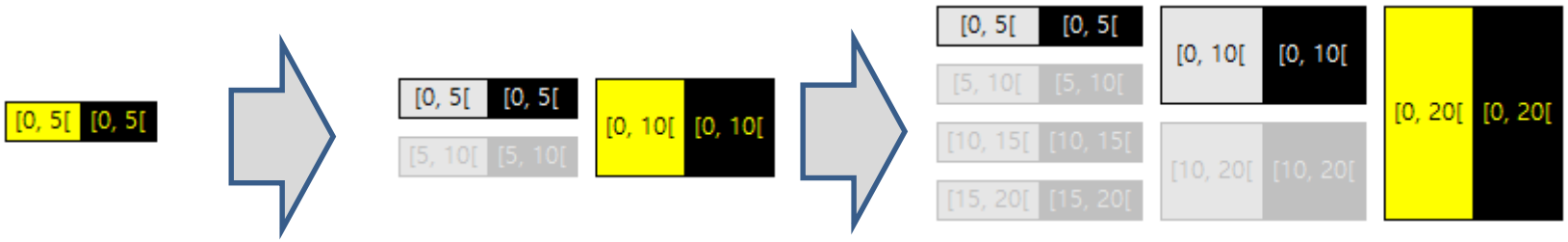
아래는 0~100세 사이를 <=20, >=60은 한 구간으로 처리하고 나머지 구간을 여러 구간으로 나눈다는 의미이다.



다음으로 interval 은 0~5 즉, 5세 간격으로 설정한다.

0~10세 간격을 추가하기 위해 [0~5[박스를 선택하고 우측 마우스 버튼을 눌러 Add New Label 을 선택한다.

다음으로 Group 에서 Size를 2로 입력하면 아래와 같이 만들어진다.



ARX-Configure

아래는 age interval을 최종적으로 완성한 모습이다.

Level-0	Level-1	Level-2	Level-3	Level-4
21	[20, 25[	[20, 30[	[20, 40[	*
22	[20, 25[	[20, 30[	[20, 40[	*
23	[20, 25[	[20, 30[	[20, 40[	*
24	[20, 25[	[20, 30[	[20, 40[	*
31	[30, 35[	[30, 40[	[20, 40[	*
32	[30, 35[	[30, 40[	[20, 40[	*
33	[30, 35[	[30, 40[	[20, 40[	*
34	[30, 35[	[30, 40[	[20, 40[	*
41	[40, 45[	[40, 50[	[40, 60[	*
42	[40, 45[	[40, 50[	[40, 60[	*
43	[40, 45[	[40, 50[	[40, 60[	*
44	[40, 45[	[40, 50[	[40, 60[	*
51	[50, 55[	[50, 60[	[40, 60[	*
52	[50, 55[	[50, 60[	[40, 60[	*
53	[50, 55[	[50, 60[	[40, 60[	*

다음에 비슷한 변환을 또 하려면 귀찮다. 이 때에는 이러한 변환방법으로 Export하여 다음에 재 사용할 수 있다.

ARX-Configure

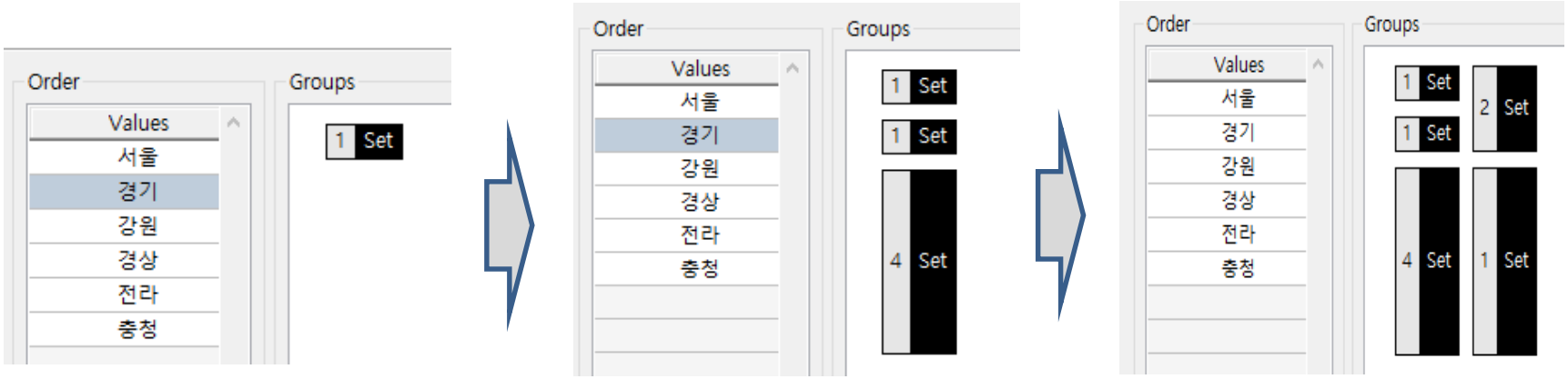
다음은 loc의 변환이다.

loc은 {서울, 경기, 충청, 강원, 전라, 경상}을 {서울, 경기, 지방}으로 분류한다고 하자.

Use Ordering을 선택한 다음 Move Up/Down을 이용하여 서울, 경기를 위로 올리면 좌측과 같은 모습이 된다.

다음은 1 Set을 선택하고 우측 마우스 버튼을 눌러 Add After를 선택한 다음 Group Size를 1로 지정, 또 하나를 Add After하면 우측의 모양이 된다.

만약, 서울, 경기를 수도권으로 통합하려면 Add New Level을 클릭한 다음 새로 만들어진 부분에서 다시 Add After를 선택한다. 그 후에 size를 2로 지정하면 우측의 모양을 만들 수 있다.



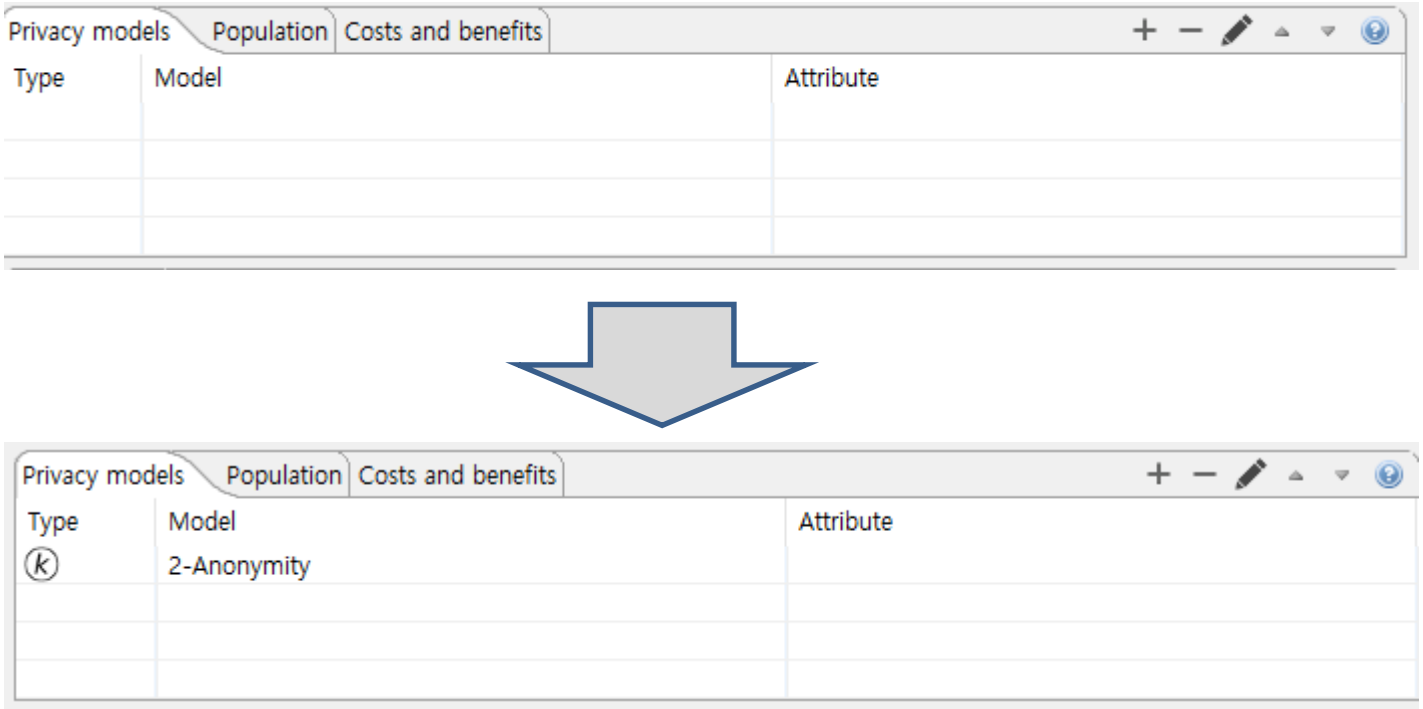
ARX-Anonymize

Privacy Model을 지정하는 절차이다.

2-익명성 모형을 지정해보자. 아래의 그림에서 + 버튼을 클릭한 다음 k-Anonymity를 선택하고 k=2를 지정한다.

Salary를 Sensitive로 지정했다면 l-Diversity, t-Closeness를 추가할 수도 있다.

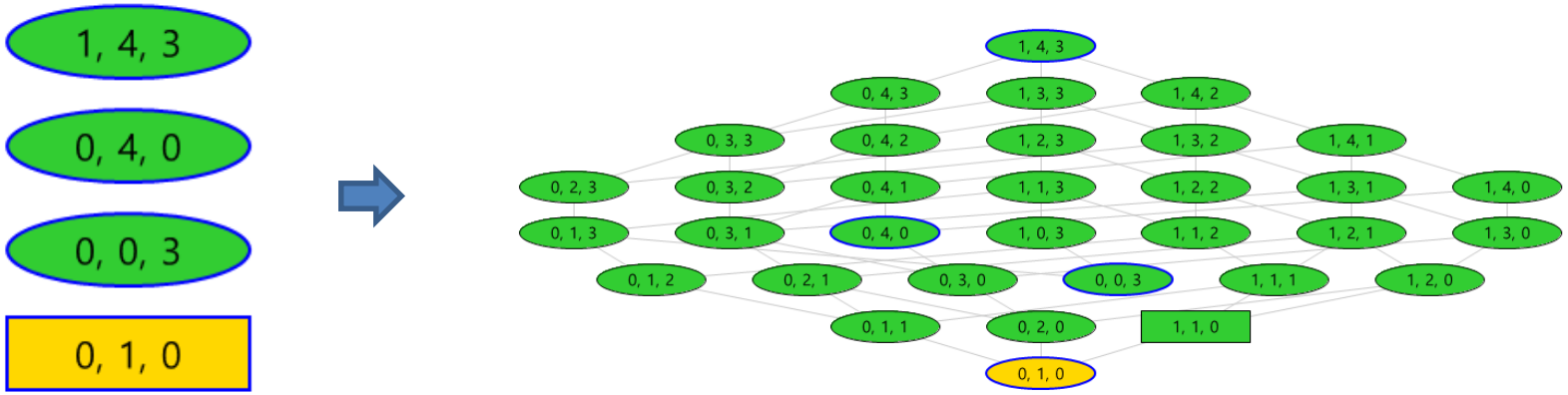
여기서는 간단하게 2-Anonymity 만 해보자.



ARX-Explore

Privacy Model을 지정한 후, 아래의 그림에서 Anonymize 버튼을 누르면 선택한 비 식별화 모형을 만족하는 여러 조합이 변환모형이 explore 화면에 나타난다.

아래 그림에서 밑부분 {0,1,0}이 가장 낮은 level의 조합이고 위의 {1,4,2}이 가장 높은 level의 조합이다. {0,0,2}의 의미는 좌측부터 sex는 level 0, age는 level 0, loc는 level 2의 변환을 의미한다.



좌측부터 그래프를 확장하면서 여러가지 변화모형을 볼 수 있다. 적절한 모형을 선택한다.

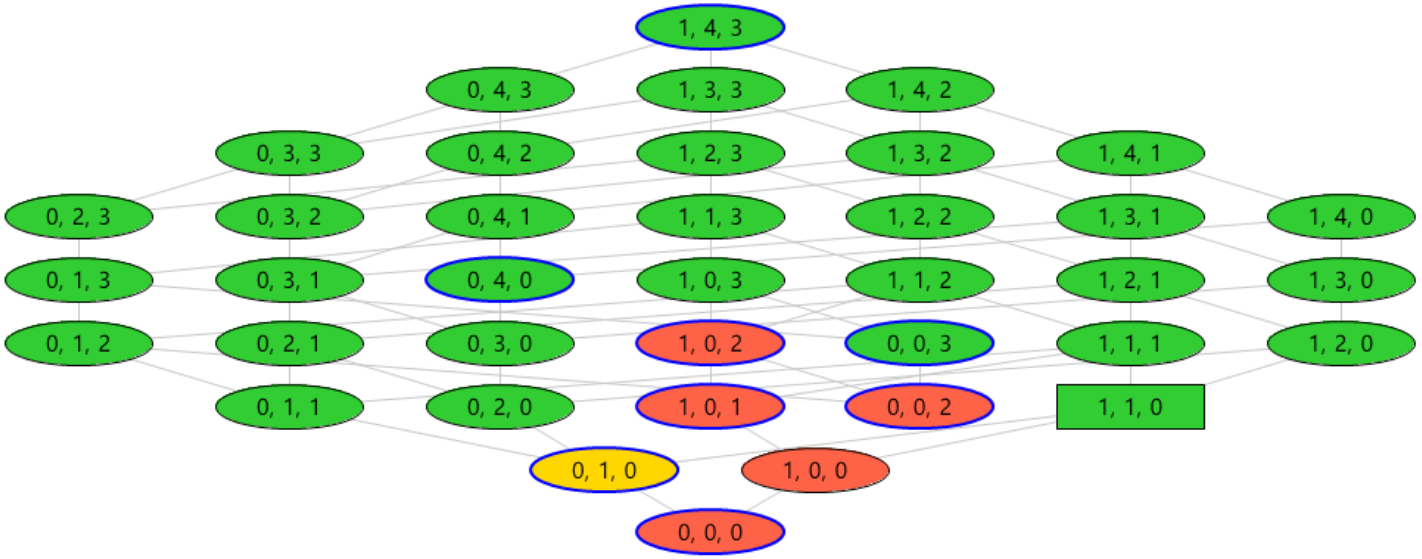
ARX-Explore

그림에서 녹색은 사용자가 지정한 프라이버시 모형에 부합되는 결과 집합이고, 빨강색은 그렇지 못한 결과집합이다.

노란색은 그 중에서도 Optimal Solution을 나타낸다.

여기서, 어떤 조합을 선택할 것인가를 결정하는 것이다.

변환모형의 선택은 이 변환이 프라이버시를 침해하지 않고 Information Loss도 어느 정도 감내할 수준인가에 대한 고민이다. (0,1,1)은 2-익명성을 만족하고 (1,0,1)은 만족하지 않는다.



ARX-Explore

그림에서 빨강색은 주어진 2-Anonymity를 만족하지 않는 경우이다. 아래의 R 코드로 확인할 수 있다.

```
setwd("D:/NaverCloud/Lecture/KICT/data")
df <- read.csv("arxTest.csv", header=TRUE)

library(sqldf)

# (0,0,0) 모형의 frequency
sqldf("select count(*) from df group by sex, age, loc")

df$sexL1 <- '*'

df$ageL1 <- NA
df$ageL1 <- ifelse(df$age < 20, 0, NA)
df$ageL1 <- ifelse(df$age >= 20 & df$age <25, 1, df$ageL1)
df$ageL1 <- ifelse(df$age >= 25 & df$age <30, 2, df$ageL1)
df$ageL1 <- ifelse(df$age >= 30 & df$age <35, 3, df$ageL1)
df$ageL1 <- ifelse(df$age >= 35 & df$age <40, 4, df$ageL1)
df$ageL1 <- ifelse(df$age >= 40 & df$age <45, 5, df$ageL1)
df$ageL1 <- ifelse(df$age >= 45 & df$age <50, 6, df$ageL1)
df$ageL1 <- ifelse(df$age >= 50 & df$age <55, 7, df$ageL1)
df$ageL1 <- ifelse(df$age >= 55 & df$age <60, 8, df$ageL1)
df$ageL1 <- ifelse(df$age >= 60, 9, df$ageL1)
```

ARX-Explore

모형의 frequency가 주어진 k 값보다 작은 것이 존재한다면 k 익명성을 만족하지 않는 것이다.

아래 결과에서 (0,1,0) 변환은 frequency 최소값이 2 이고, (1,0,1) 변환의 최소값은 1이다.

만약, k=2 라면 (1,0,1) 변환은 기준을 만족하지 않는다.

```
df$locL1 <- df$loc
df$locL1 <- ifelse(df$loc != "서울" & df$loc != "경기", "수도권외",df$locL1)
df$locL1 <- ifelse(df$loc == "서울", "서울",df$locL1)
df$locL1 <- ifelse(df$loc == "경기", "경기",df$locL1)

# (0,1,1) 모형의 frequency
sqldf("select sex, ageL1, locL1, count(*) from df group by sex, ageL1, locL1")

# (1,0,1) 모형의 frequency
sqldf("select sexL1, age, locL1, count(*) from df group by sexL1, age, locL1")
```

이 자료에서 모든 frequency가 같은 값이지만, 실제 자료에서는 모든 준식별자 조합의 frequency는 제각각 다른 값이 나오는데 이 중 최소값을 기준으로 판단한다. 문제는 대부분 k 이상인데 몇 개의 케이스만 k 미만인 경우다.

ARX-Explore

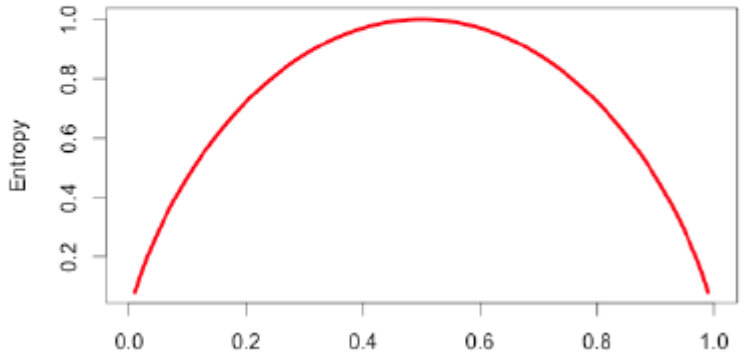
Information loss는 데이터에 대한 비식별기법을 적용하여 데이터를 변환하는 것에 대한 정보 손실량을 계산하는 것이다.

ARX에서는 여러가지 loss 계산을 제공한다. 여기서는 엔트로피를 예를 들어본다.

$$\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2(p_i)$$

```
> curve(-x * log2(x) - (1 - x) * log2(1 - x),
        col="red", xlab = "x", ylab = "Entropy", lwd=4)
```

This results in the following figure:



Entropy는 위의 그림처럼 확률이 0.5일 때, 최대가 된다. 즉, 확률이 0.5일 때, 가장 불확실성이 크다는 의미이다.

ARX-Explore

예를 들어, 나이를 좌에서 우로 변환할 때, Information loss는 아래와 같이 계산한다.

20	20-39
65	60-79
55	40-59
40	40-59
65	60-79
65	60-79

좌측 원 자료의 엔트로피는 20, 40, 55세가 각 1명이고, 65세가 3명이므로

$$-3 * \frac{1}{6} \cdot log\frac{1}{6} - \frac{3}{6} \cdot log\frac{3}{6} = 1.2424$$

우측 자료의 엔트로피는 20-29세 1명, 40-59세 2명, 60-79세가 3명이므로

$$-\frac{1}{6} \cdot log\frac{1}{6} - \frac{2}{6} \cdot log\frac{2}{6} - \frac{3}{6} \cdot log\frac{3}{6} = 1.011$$

엔트로피 변화량은 $\frac{1.2424-1.011}{1.2424} = 0.18$ 로 18%이다.

위의 계산을 비 식별화 변환을 수행한 모든 컬럼에 대해 계산하고 각 컬럼의 변화 량을 산술평균, 기하평균 등을 사용하여 하나의 값으로 변환하면 이 값이 Information loss가 된다.

ARX-Transformation

(0,1,1) 모형으로 변환한 결과, Input Data에는 총 40개의 관측값이 있고, 각 관측값은 40개의 클래스로 이루어져 있었는데 변환 후, 40개의 레코드는 그대로 있고, 클래스는 20개로 변했다. 즉, 클래스 당 2개의 관측 값이 만들어 졌다.

Summary statistics | Distribution | Contingency | Class sizes | Properties | Classification accuracy

Measure	Including outliers
Average class size	1 (2.5%)
Maximal class size	1 (2.5%)
Minimal class size	1 (2.5%)
Number of classes	40
Number of records	40
Suppressed records	0 (0%)

Summary statistics | Distribution | Contingency | Class sizes | Properties | Classification models

Measure	Value (incl. suppressed)	Value (excl. suppressed)
Average class size	2 (5%)	2 (5%)
Maximal class size	2 (5%)	2 (5%)
Minimal class size	2 (5%)	2 (5%)
Suppressed records	0 (0%)	0
Number of classes	20	20
Number of records	40	40

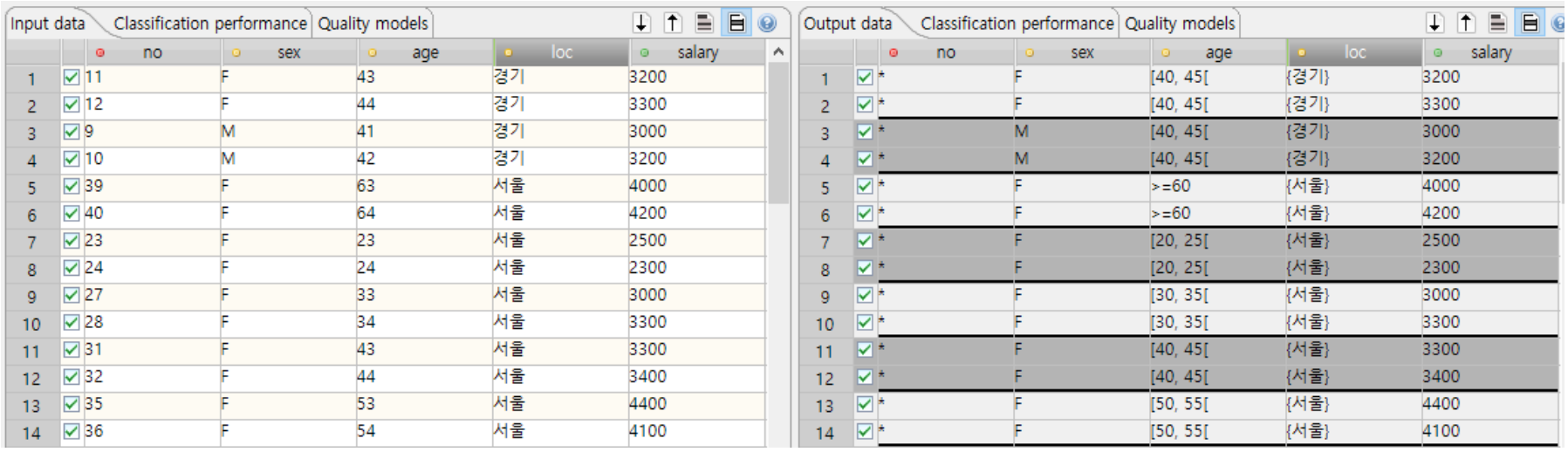
ARX-Transformation

{0,1,1} 모형에 의한 변환이 된 결과가 Analyze/enhance utility 탭에 나타난다.

이 상태에서 File → Export 를 누르면 우측의 결과를 저장할 수 있다.

Export된 파일이 최종적으로 비식별화된 결과 파일이다.

최종 Export 여부를 결정하기 전에 이 변환이 적절한지를 판단하기 위해 몇가지 추가 분석이 필요하다.



Input data | Classification performance | Quality models

	no	sex	age	loc	salary
1	11	F	43	경기	3200
2	12	F	44	경기	3300
3	9	M	41	경기	3000
4	10	M	42	경기	3200
5	39	F	63	서울	4000
6	40	F	64	서울	4200
7	23	F	23	서울	2500
8	24	F	24	서울	2300
9	27	F	33	서울	3000
10	28	F	34	서울	3300
11	31	F	43	서울	3300
12	32	F	44	서울	3400
13	35	F	53	서울	4400
14	36	F	54	서울	4100

Output data | Classification performance | Quality models

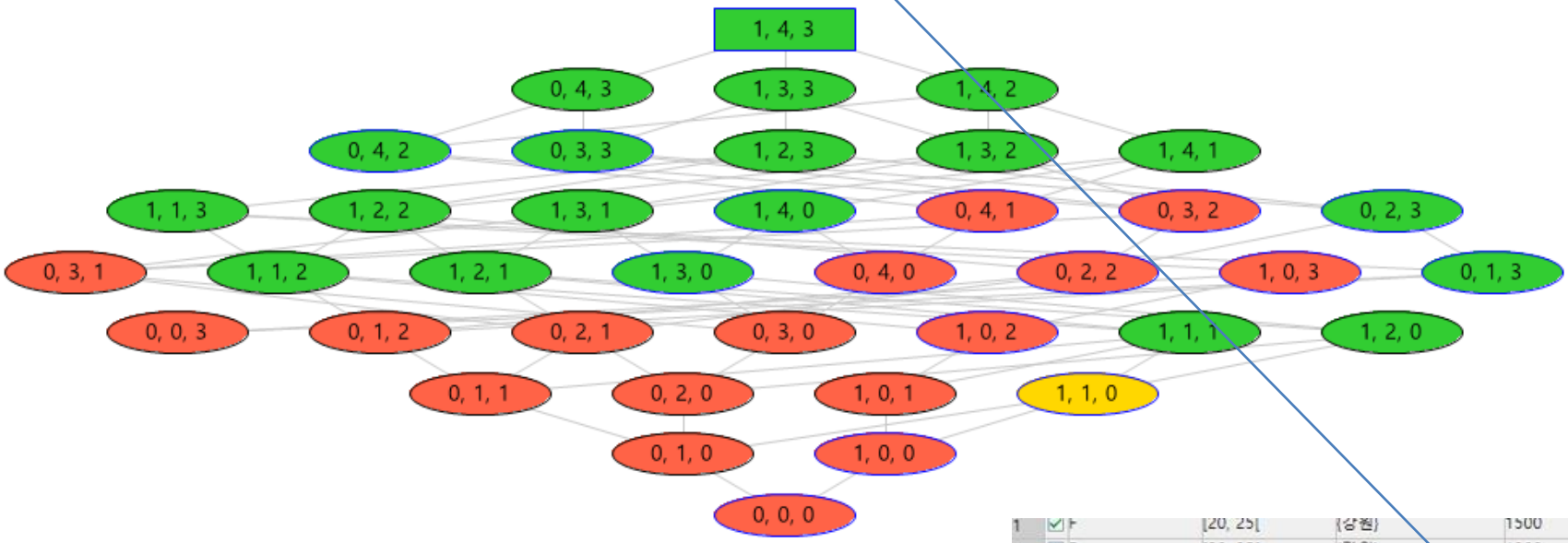
	no	sex	age	loc	salary
1	*	F	[40, 45[	{경기}	3200
2	*	F	[40, 45[	{경기}	3300
3	*	M	[40, 45[	{경기}	3000
4	*	M	[40, 45[	{경기}	3200
5	*	F	>=60	{서울}	4000
6	*	F	>=60	{서울}	4200
7	*	F	[20, 25[	{서울}	2500
8	*	F	[20, 25[	{서울}	2300
9	*	F	[30, 35[	{서울}	3000
10	*	F	[30, 35[	{서울}	3300
11	*	F	[40, 45[	{서울}	3300
12	*	F	[40, 45[	{서울}	3400
13	*	F	[50, 55[	{서울}	4400
14	*	F	[50, 55[	{서울}	4100

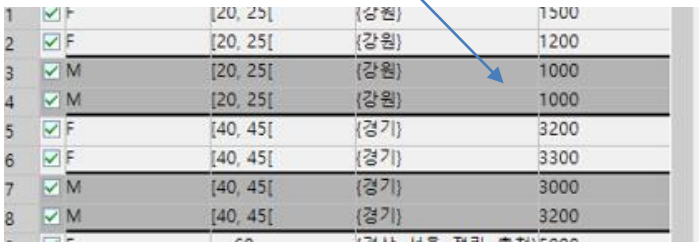
ARX-Transformation

이 자료에서 salary를 sensitive로 바꾸고 2-다양성을 추가해 보자.

그 결과 (0,1,1) 모형을 선택할 수 없다. 왜 일까? 이유는 3,4 번 레코드 때문이다.

최종적으로 (1,1,1) 모형을 선택하자.





1	F	[20, 25[	{강원}	1500
2	F	[20, 25[	{강원}	1200
3	M	[20, 25[	{강원}	1000
4	M	[20, 25[	{강원}	1000
5	F	[40, 45[	{경기}	3200
6	F	[40, 45[	{경기}	3300
7	M	[40, 45[	{경기}	3000
8	M	[40, 45[	{경기}	3200

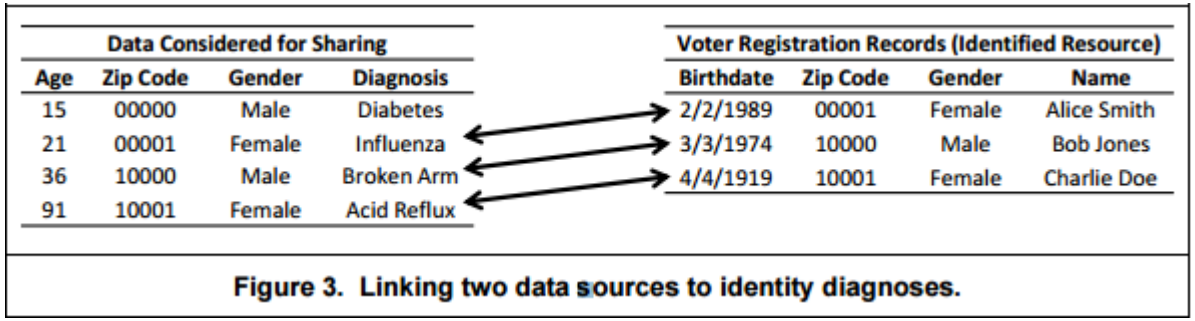
ARX-Risk Analysis

다음은 Risk Analysis이다.

재식별 가능성 분석은 특정 개인이 이 Dataset에 포함되어 있음을 알고 있다는 가정 하에 그 개인을 찾을 수 있는 가능성을 분석하는 것이다.

Latanya Sweeney는 성별, 생년월일, 우편번호를 가지고 전체 미국민의 87%의 개인을 식별할 수 있음을 보인 바 있다.

이처럼 비식별화를 했다고 해도 이를 재식별 할 수 있는 가능성이 "0"이 된다는 것이 아님을 이해해야 한다.



Data Considered for Sharing			
Age	Zip Code	Gender	Diagnosis
15	00000	Male	Diabetes
21	00001	Female	Influenza
36	10000	Male	Broken Arm
91	10001	Female	Acid Reflux

Voter Registration Records (Identified Resource)			
Birthdate	Zip Code	Gender	Name
2/2/1989	00001	Female	Alice Smith
3/3/1974	10000	Male	Bob Jones
4/4/1919	10001	Female	Charlie Doe

Figure 3. Linking two data sources to identity diagnoses.

ARX-Risk Analysis

예를 들어, 아래와 같이 데이터가 있고 USUBJID는 식별자, SEX, AGE가 준 식별자라고 하자.

이 데이터 set에 내가 아는 누군가 있다는 정보가 있다고 가정하면 재 식별 위험도는 아래와 같이 계산된다.

즉, 재 식별 위험도는 해당 속성을 가진 사람이 얼마나 중복되어 있는 가이다.

USUBJID	SEX	AGE	Equiv. Class (Size)	Re-Id risk
CT1/101	M	26	A (1)	1
CT1/102	F	28	B (2)	0.5
CT1/103	F	31	C (2)	0.5
CT1/104	M	29	D (3)	0.33
CT1/105	F	28	B (2)	0.5
CT1/106	M	30	E (1)	1
CT1/107	M	29	D (3)	0.33
CT1/108	F	32	F (1)	1
CT1/109	M	29	D (3)	0.33
CT1/110	F	31	C (2)	0.5

$$Pr_{max}(re\text{-}id \mid attempt) = max\ (1, 0.5, 0.5, 0.33, 0.5, 1, 0.33, 1, 0.33, 0.5) = 1$$

$$Pr_{avg}(re\text{-}id \mid attempt) = (1+0.5+0.5+0.33+0.5+1+0.33+1+0.33+0.5) / 10 = 0.6$$

위의 표에서 재 식별 Risk는 확률 “1”이고, 각 개인은 0.6의 확률로 식별가능성을 가진다.

이 개념이 Highest Prosecutor Risk와 Average Prosecutor Risk이다.

ARX-Risk Analysis

아래의 결과는

Lowest prosecutor risk=25% → 개인의 재 식별 리스크 값 중 최소값

Records affected by lowest risk =100% → 최저 리스크에 노출된 레코드 비율

Average prosecutor risk=25% → 개인의 재 식별 리스크 평균값

Highest prosecutor risk=25% → 개인의 재 식별 리스크 값 중 최대값

Records affected by highest risk = 100% → 최대 리스크에 노출된 레코드 비율

Estimated prosecutor risk=25% → Highest prosecutor risk 값과 같은 값이다.

Overview | Population uniques

Measure	Value [%]
Lowest prosecutor risk	25%
Records affected by lowest risk	100%
Average prosecutor risk	25%
Highest prosecutor risk	25%
Records affected by highest risk	100%
Estimated prosecutor risk	25%

ARX-Risk Analysis

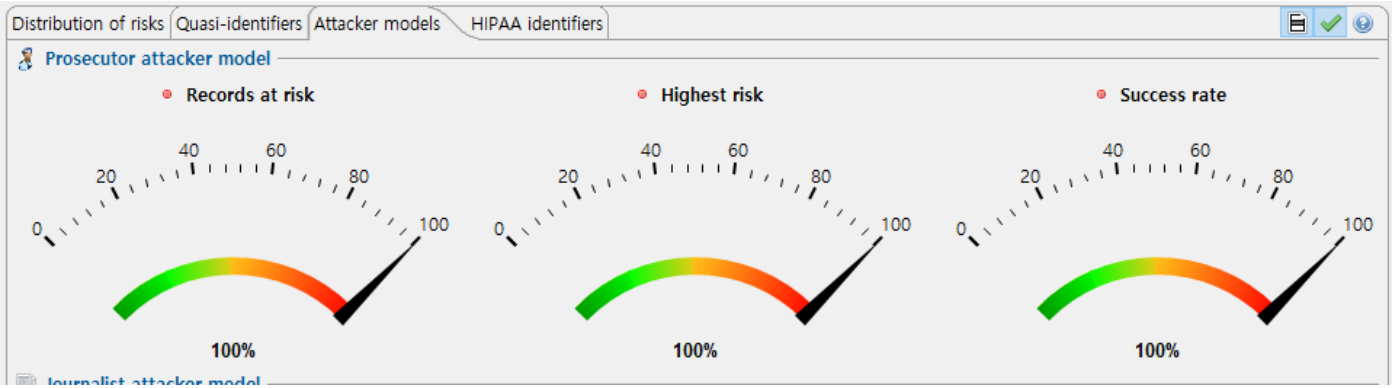
아래는 Re-Identification Risk 탭의 화면이다.

이 화면에는 Prosecutor, Journalist, Marketer Model에 대한 Records at risk, Highest risk, Success rate 값이 출력된다. 좌측 그림은 변환 전의 자료로 Risk가 100%이다. 우측 그림은 변환 후의 결과로 리스크가 낮아졌음을 알 수 있다.

Records at risk : Proportion of records with risk above the threshold 20%

highest risk: Highest risk of a single record

Success rate: Proportion of records that can be re-identified on average

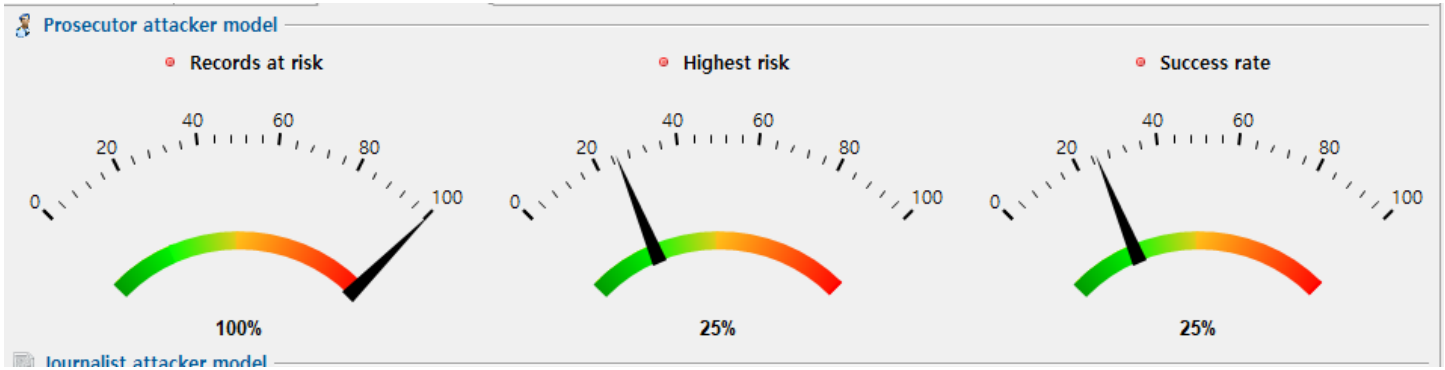


ARX-Risk Analysis

Records at risk : 이 변환의 위험확률은 25%로 임계값 20%를 모두 초과하므로 100% 레코드가 위험함

highest risk: 위험확률 중 가장 큰 값은 25%임

Success rate: 재식별 공격을 받았을 때, 25%의 확률로 재식별됨



비식별화 실전문제

실전 문제로 미국 성인의 소득자료를 이용해 비식별화를 해보자. Data는 아래에서 받을 수 있다.

https://drive.google.com/open?id=1N0Dz0i4sV7bRbKUfUrY42y0C0VtdlNHK

Data는 salary_class <=50k, >50k를 종속변수로 하고 나머지 변수로 급여수준을 예측하는 문제다.

주요 변수명	변수 타입	설명	구분
age	숫자	나이	17~90세
workclass	문자	직업 구분	?, Never-Worked, without-pay, private, gov, self-emp-not-inc, self-emp-inc
marital_status	문자	결혼 상태	Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Separated, Widowed, Never-married
race	문자	인종	Amer-Indian-Eskimo, Other, Black, Asian-Pac-Islander, White
sex	문자	성별	Male, Female
salary_class	문자	연봉	<=50K, >50K

비식별화 실전문제

1. 데이터 import 하기
2. 컬럼 구분하기

   식별자: 없음

   준식별자 지정 : age, workclass, marital_status, race, sex, native_country

   민감정보: salary_class

3. 준 식별자 변환 모형 만들기

   변환모형을 어떻게 만들 수 있는가?

   우선, 나이는 5세, 10세 단위로 변환하고, sex는 매스킹 한다.

   workclass, marital_status, race는 그룹화할 것이다.

   그룹화는 매우 중요한 단계다. 왜냐하면 그룹화를 잘못하면 데이터의 품질이 급속도로 나빠지기 때문이다.

   그룹화를 위해 R에서 아래와 같이 카이제곱 분석을 수행 해보자.

```
library(gmodels)
CrossTable(x = df$workclass, y = df$salary_class, prop.t=FALSE, expected=TRUE, chisq =TRUE)
CrossTable(x = df$marital_status, y = df$salary_class, prop.t=FALSE, expected=TRUE, chisq =TRUE)
CrossTable(x = df$race, y = df$salary_class, prop.t=FALSE, expected=TRUE, chisq =TRUE)
CrossTable(x = df$sex, y = df$salary_class, prop.t=FALSE, expected=TRUE, chisq =TRUE)
```

비식별화 실전문제

테이블 읽는 법

- 셀 단위로 위에서 차례대로 빈도, 기대 빈도, Cell Chi-square 값, Row Percent, Col Percent를 나타냄
- 기대 빈도는 예를 들어, 직업 구분과 연봉이 독립이라고 가정할 때, 기대되는 빈도로 실제 빈도와 기대 빈도의 차이가 크면 직업 구분과 연봉이 독립이 아니라는 주장이 설득력을 얻는다.
- 실제 빈도와 기대빈도의 차이를 표준화 한 값이 Cell Chi-square 로 아래의 식 중 Summation 안에 있는 부분이다.

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

- 우측 표에서 ?는 Cell Chi-square 값이 크고 빈도와 기대 빈도 차이가 '+'로 직업이 ? 인 경우는 연봉이 낮고 반대로 직업이 Federal-gov는 차이가 '-'로 연봉이 높다.
- 이 경우, ?와 Federal-gov를 그룹화 하면 정보 손실이 많이 발생하므로 그룹화하지 않는 것이 좋다.
- 반면, Federal-gov와 Local-gov는 차이의 부호가 같으므로 같은 그룹으로 묶어 비식별처리해도 무방하다.

df$workclass	df$salary_class <=50K	>50K	Row Total
?	2534	265	2799
	2129.250	669.750	
	76.939	244.602	
	0.905	0.095	0.057
	0.068	0.023	
Federal-gov	871	561	1432
	1089.349	342.651	
	43.766	139.139	
	0.608	0.392	0.029
	0.023	0.048	
Local-gov	2209	927	3136
	2385.612	750.388	
	13.075	41.568	
	0.704	0.296	0.064
	0.059	0.079	
Never-worked	10	0	10
	7.607	2.393	
	0.753	2.393	
	1.000	0.000	0.000
	0.000	0.000	
Private	26519	7387	33906
	25792.912	8113.088	
	20.440	64.982	
	0.782	0.218	0.694
	0.714	0.632	

비식별화 실전문제

df$workclass	df$salary_class <=50K	>50K	Row Total
?	2534	265	2799
	2129.250	669.750	
	76.939	244.602	
	0.905	0.095	0.057
	0.068	0.023	
Federal-gov	871	561	1432
	1089.349	342.651	
	43.766	139.139	
	0.608	0.392	0.029
	0.023	0.048	
Local-gov	2209	927	3136
	2385.612	750.388	
	13.075	41.568	
	0.704	0.296	0.064
	0.059	0.079	
Never-worked	10	0	10
	7.607	2.393	
	0.753	2.393	
	1.000	0.000	0.000
	0.000	0.000	
Private	26519	7387	33906
	25792.912	8113.088	
	20.440	64.982	
	0.782	0.218	0.694
	0.714	0.632	
Self-emp-inc	757	938	1695
	1289.417	405.583	
	219.842	698.916	
	0.447	0.553	0.035
	0.020	0.080	
Self-emp-not-inc	2785	1077	3862
	2937.894	924.106	
	7.957	25.296	
	0.721	0.279	0.079
	0.075	0.092	
State-gov	1451	530	1981
	1506.983	474.017	
	2.080	6.612	
	0.732	0.268	0.041
	0.039	0.045	
Without-pay	19	2	21
	15.975	5.025	
	0.573	1.821	
	0.905	0.095	0.000
	0.001	0.000	
Column Total	37155	11687	48842
	0.761	0.239	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 1610.752 d.f. = 8 p = 0

{?, Never-Worked, without-pay},
{private, self-emp-not-inc},
{gov, self-emp-inc} 로 변환해보자.

비식별화 실전문제

아래 결과로 다음과 같은 룰을 만들 수 있다.

married-civ-spouse > married-AF-spouse 는 한 그룹이고, 나머지가 한 그룹으로 나눌 수 있음을 알 수 있다.

df$marital_status	df$salary_class <=50K	>50K	Row Total
Divorced	5962	671	6633
	5045.844	1587.156	
	166.343	528.834	
	0.899	0.101	0.136
	0.160	0.057	
Married-AF-spouse	23	14	37
	28.147	8.853	
	0.941	2.992	
	0.622	0.378	0.001
	0.001	0.001	
Married-civ-spouse	12395	9984	22379
	17024.113	5354.887	
	1258.726	4001.708	
	0.554	0.446	0.458
	0.334	0.854	
Married-spouse-absent	570	58	628
	477.731	150.269	
	17.821	56.655	
	0.908	0.092	0.013
	0.015	0.005	
Never-married	15384	733	16117
	12260.496	3856.504	
	795.749	2529.824	
	0.955	0.045	0.330
	0.414	0.063	
Separated	1431	99	1530
	1163.899	366.101	
	61.297	194.872	
	0.935	0.065	0.031
	0.039	0.008	
Widowed	1390	128	1518
	1154.770	363.230	
	47.917	152.336	
	0.916	0.084	0.031
	0.037	0.011	
Column Total	37155	11687	48842
	0.761	0.239	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 9816.015 d.f. = 6 p = 0

비식별화 실전문제

df$race	df$salary_class <=50K	>50K	Row Total
Amer-Indian-Eskimo	415	55	470
	357.538	112.462	
	9.235	29.360	
	0.883	0.117	0.010
	0.011	0.005	
Asian-Pac-Islander	1110	409	1519
	1155.531	363.469	
	1.794	5.704	
	0.731	0.269	0.031
	0.030	0.035	
Black	4119	566	4685
	3563.965	1121.035	
	86.439	274.803	
	0.879	0.121	0.096
	0.111	0.048	
Other	356	50	406
	308.852	97.148	
	7.198	22.882	
	0.877	0.123	0.008
	0.010	0.004	
White	31155	10607	41762
	31769.115	9992.885	
	11.871	37.741	
	0.746	0.254	0.855
	0.839	0.908	
Column Total	37155	11687	48842
	0.761	0.239	

아래 결과로 다음과 같은 룰을 만들 수 있다.

White 그룹, Asian 그룹, (Indian, other) 그룹,

black 그룹으로 나눌 수 있다.

```
Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  487.0263     d.f. =  4     p =  4.284378e-104
```

비식별화 실전문제

	df$salary_class		
df$sex	<=50K	>50K	Row Total
Female	14423	1769	16192
	12317.550	3874.450	
	359.887	1144.142	
	0.891	0.109	0.332
	0.388	0.151	
Male	22732	9918	32650
	24837.450	7812.550	
	178.477	567.410	
	0.696	0.304	0.668
	0.612	0.849	
Column Total	37155	11687	48842
	0.761	0.239	

Statistics for All Table Factors

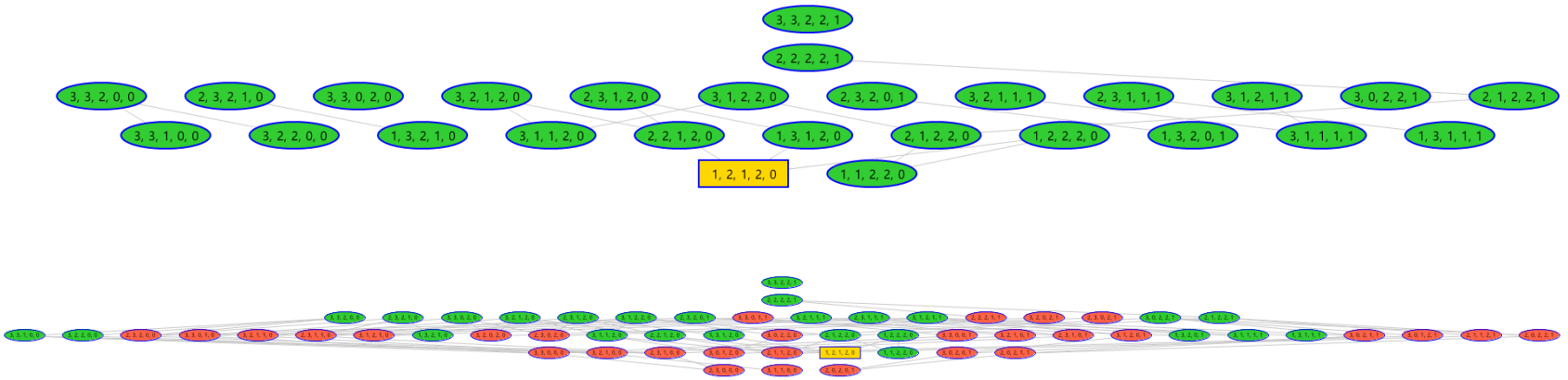
Pearson's Chi-squared test

Chi^2 = 2249.916 d.f. = 1 p = 0

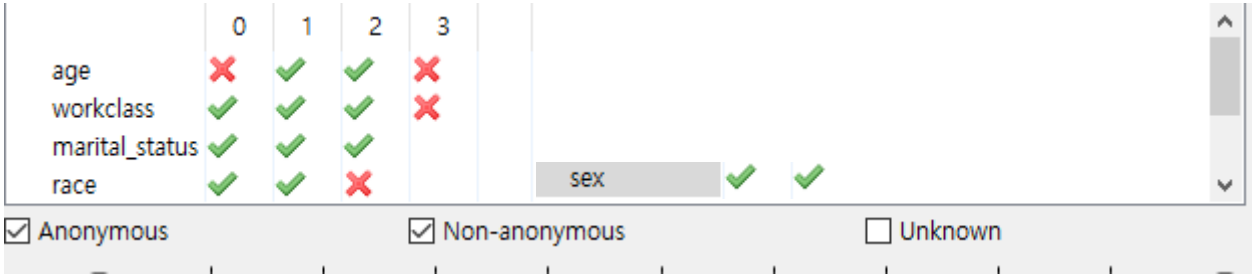
Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 2248.848 d.f. = 1 p = 0

비식별화 실전문제



ARX는 (1,2,1,2,0) 모형을 추천해주었다. 이 모형보다 좀 더 변환이 적게 되는 모형을 선택해보자.



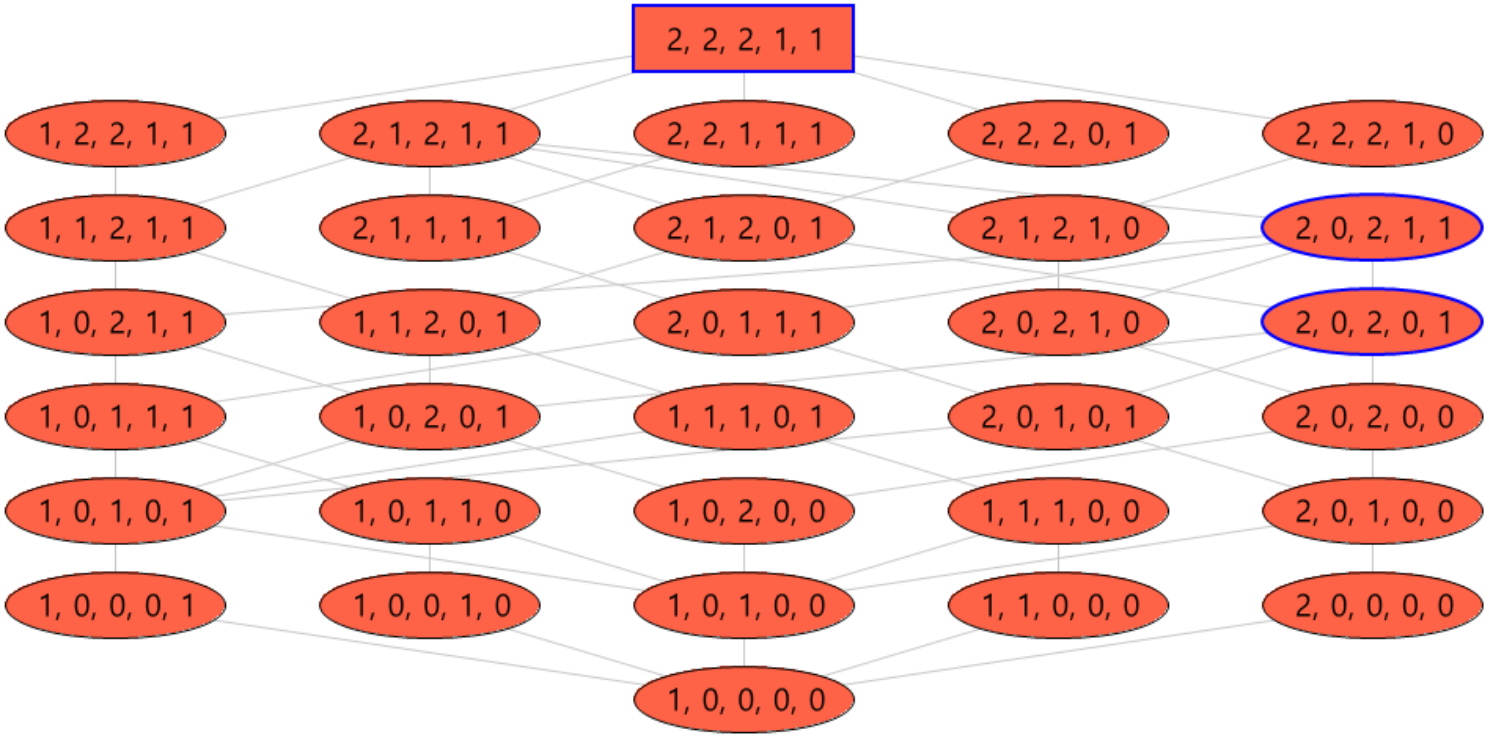
비식별화 실전문제

즉, 추천해주는 모형이 마음에 안들 경우, 10-익명성을 만족하지 않는 모형을 강제로 선택할 수 있다.

이 경우는 ARX는 선택된 모형을 만족하지 않는 케이스를 강제 삭제하는 변환을 수행한다.

일부, 케이스를 삭제하는 것도 비 식별 변환 중 하나로 볼 수 있다.

“어느 것을 선택하는가” 는 모형의 품질에 많은 영향을 준다. (1,1,1,0,1) 모형을 선택해보자.



비식별화 실전문제

(1,1,1,0,1) 모형을 선택한 결과, Suppressed records가 659케이스로 전체의 1.3% 발생했다.

(1,2,1,2,0)과 (1,1,1,0,1) 모형 중 어느 모형을 선택할 것인가?

이에 대한 의사결정은 쉽지 않다. 만약, 데이터가 충분히 많다면 케이스를 삭제하는 것도 긍정적으로 생각해 볼 수 있다.

(1,2,1,2,0) 모형 결과

Measure	Including outliers	Excluding outliers
Average class size	452.24074 (0.92593%)	452.24074 (0.92593%)
Maximal class size	3265 (6.68482%)	3265 (6.68482%)
Minimal class size	13 (0.02662%)	13 (0.02662%)
Number of classes	108	108
Number of records	48842	48842 (100%)
Suppressed records	0 (0%)	0

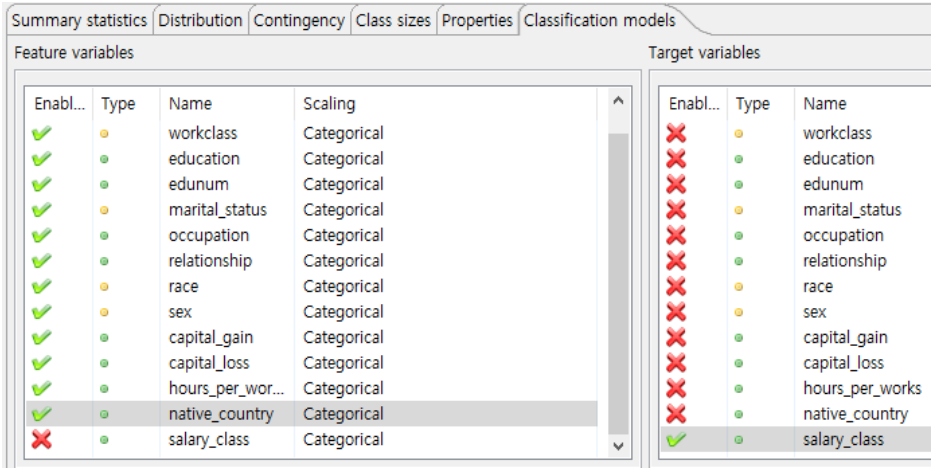
(1,1,1,0,1) 모형 결과

Measure	Including outliers	Excluding outliers
Average class size	241.79208 (0.49505%)	239.71642 (0.49751%)
Maximal class size	5158 (10.56058%)	5158 (10.70502%)
Minimal class size	10 (0.02047%)	10 (0.02075%)
Number of classes	202	201
Number of records	48842	48183 (98.65075%)
Suppressed records	659 (1.34925%)	0

ARX-Example

아래는 두 모형의 예측력을 비교하는 결과다.

입력 데이터를 변환하지 않고 모형화 한 결과, 정확도가 86.66% 인 데이터를 10-익명성 비식별 변환 후, (1,2,1,2,0) 모형의 예측 정확도는 86.11%로 소폭 낮아졌다. (1,1,1,0,1) 모형의 정확도는 86.68%로 오히려 높아졌다. 이는 많은 사람들이 비식별 변환 후, 데이터의 품질이 나빠질 것이라는 우려를 뒤집는 의미 있는 결과이다.



(0,0,0,0,0) 모형

Input data | Classification performance | Quality models

Target variable	Baseline accuracy	Accuracy
salary_class	76.07182%	86.66517%

(1,2,1,2,0) 모형

Output data | Classification performance | Quality models

Target variable	Baseline accuracy	Accuracy	Original accuracy	Relative accuracy
salary_class	76.07182%	86.1185%	86.26182%	98.59353%

(1,1,1,0,1) 모형

Output data | Classification performance | Quality models

Target variable	Baseline accuracy	Accuracy	Original accuracy	Relative accuracy
salary_class	76.07182%	86.6795%	86.66517%	100.13529%

통신자료 Review

자료는 1만개의 레코드로 이루어진 개인고객의 통신관련정보이다.

나이, 성별, 거주지, 가입년도, 멤버쉽, 태블릿PC 보유여부, 스마트워치 보유여부, 결합상품가입여부, 회선 상태, 납부방법, 통화량, 요금, 연체여부, 정지기간, 할부잔여금액, 할부잔여개월수로 이루어져 있다.

1. 이 자료에서 식별자 / 준식별자 / 민감정보 / 비 민감정보는 무엇인가?

ID	age	sex	Sido	SiGunGu	DongName	year	membership	tabletYN	watchYN	joinProductYN	status	payMethod	callAmt	pay	delayYN	pauseDuration	phonePrice	remainMoney	remainMonth
1	54	2	대구	달서구	상인동	2014	일반	N	N	N	사용중	은행자동납부	73	5	N	313	1816	3296	0
2	47	1	서울	노원구	중계동	2015	없음	N	N	N	사용중	은행자동납부	234	25	N	24	990000	248144	10
3	50	2	서울	노원구	하계동	2012	Silver	N	Y	Y	사용중	은행자동납부	23	10	N	44	5389	1273	0
4	47	1	경기	화성시	남양읍	2014	Silver	N	Y	N	사용중	은행자동납부	118	9	N	3	121000	238	0
5	61	1	경기	수원시 권선구	호매실동	2016	일반	N	N	N	사용중	은행자동납부	32	8	#	21	3180	4424	0
6	34	1	서울	성북구	정릉동	2011	Silver	N	N	N	정지	은행자동납부	90	3	N	24	8060	5417	0
7	43	2	서울	양천구	목동	2016	없음	N	Y	N	사용중	은행자동납부	49	16	N	21	348150	157900	22
8	61	1	경북	김천시	어모면	2013	Silver	N	Y	N	사용중	카드자동납부	114	26	N	21	238150	16578	2
9	38	2	서울	관악구	신림동	2013	VIP	N	N	N	사용중	은행자동납부	6	15	N	23	5053	421	0
10	35	2	서울	양천구	목동	2015	없음	N	N	N	사용중	은행자동납부	66	2	N	4	929612	331020	12
11	46	2	서울	금천구	시흥동	2015	일반	N	Y	N	사용중	은행자동납부	137	28	N	6	949850	862772	25

식별자와 준 식별자에 대한 변환방법을 결정해야 한다. 변환은 분석목적에서 허용가능한 형태의 변환을 해야 한다.

허용이 불가능한 변환이란 변환 후에 자료가치가 없어지는 변환을 말한다.

예: 나이는 5세 단위로 Aggregation 한다. 등 ...

시도, 시군구, 읍면동은 원래 행정동코드에 의해 작업하는 것이 깔끔한데 여기에서는 시간 관계상 3개를 합쳐 하나의 주소로 만들어 처리하자.(엑셀을 이용)

통신자료 Review

2. ARX에서 자료를 읽고 변환방법을 정의하고 확장자가 deid 가 되도록 프로젝트 파일을 저장하세요.

3. 적절한 비식별화 모형을 만들어 보세요.

   k 익명성, l 다양성, t 근접성 모형을 어떻게 구성하였는가? 그렇게 구성한 이유는 무엇인가?

4. 여러 개의 변환모형 중 여러분이 선택한 변환방법은 무엇인가? 그 이유는 무엇인가?

5. 최종적으로 선택된 모형에 만족하는가? 어떤 이슈가 존재하는가? 대안은 없는가?

ARX 실습

감 사 합 니 다

Q & A