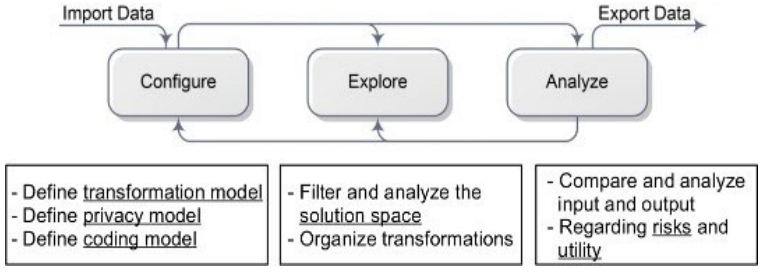
개인정보보호론

[13주차 참조. 비식별화 도구 ARX]

한국정보화진흥원(NIA) 공개자료 기반

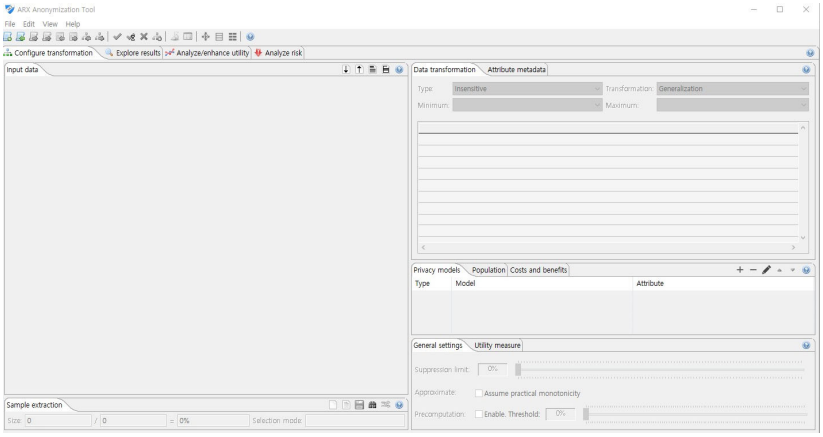
ARX 소개

- 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. 윈도우 O.S는 arx-3.8.0-windows-installer.exe 를 설치한다.
- 설치가 성공하면 시작메뉴에 ARX 메뉴가 생긴다. 아래는 ARX를 실행한 초기 화면이다.



ARX-Configure

- 비 식별화의 시작은 New Project를 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. (http://naver.me/GN7fs2ok)
- 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

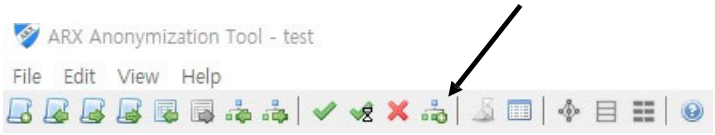
	sex	age	loc	salary
1	M	21	강원	1000
2	M	22	강원	1500
3	F	23	강원	1500
4	F	24	강원	1200
5	M	31	충청	2000
6	M	32	충청	2300
7	F	33	충청	2400
8	F	34	충청	2100
9	M	41	경기	3000
10	M	42	경기	3200
11	F	43	경기	3200
12	F	44	경기	3300

ARX-Configure

- 다음 단계는 Data Transformation 이다.
- 식별자(Identifier)는 "*" 로 처리하고 sex, age, loc은 준식별자(Quasi-identifier), salary는 Insensitive 혹은 Sensitive로 지정한다. Dataset 안에 Sensitive 변수가 없으면 l 다양성, t 근접성 모형을 사용할 수 없다.
- 우선, salary는 Insensitive로 해보자.
- 좌측에서 sex를 선택하고 우측에서 type은 Quasi-identifying, transformation은 generalization을 선택한다.
- transformation은 generalization과 aggregation이 있다.
- aggregation은 데이터를 평균이나 합 등으로 변환 하는 것이고 generalization은 15세→10대의 형식으로 변환하는 것이다.

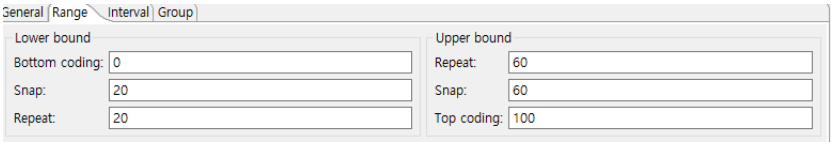
ARX-Configure

- type과 transformation을 선택하였다면 다음으로는 Transformation Hierarchy를 만들기 위해 아래의 그림과 같이 화살표 방향의 메뉴버튼을 클릭한다. use interval, ordering, masking의 메뉴가 나타난다.
- sex의 경우는 문자이므로 interval은 선택할 수 없고 ordering 혹은 masking인데 여기서는 masking을 선택한다.
- 같은 방식으로 age는 Interval, loc는 ordering으로 해보자.

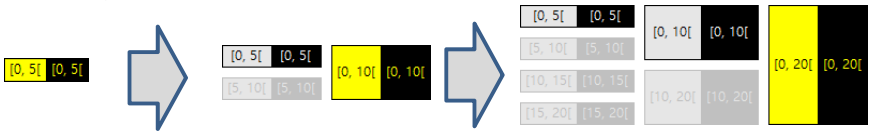


ARX-Configure

- 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. 0~100세 사이를 <=20, >=60은 나누지 않고 나머지 구간에 대해 처리한다는 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.



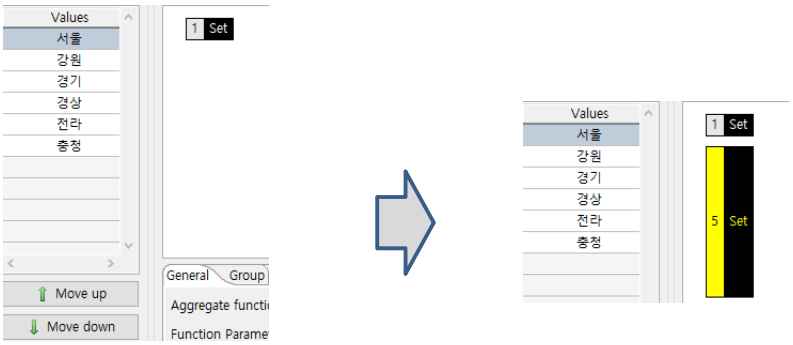
ARX-Configure

- 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[	*

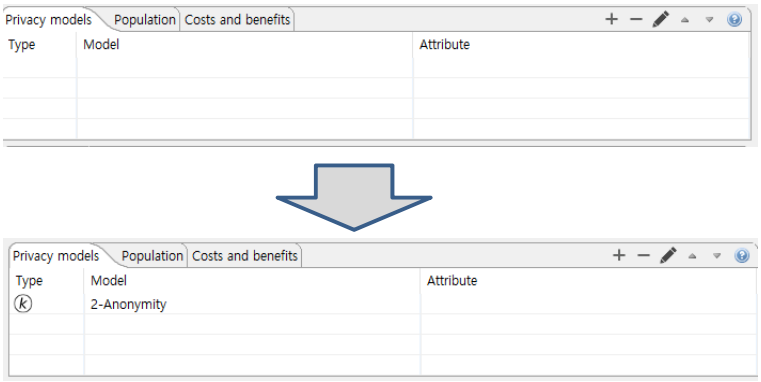
ARX-Configure

- 다음은 loc의 변환이다.
- loc은 {서울, 경기, 충청, 강원, 전라, 경상}을 {서울, 지방}으로 분류한다고 하자.
- Use Ordering을 선택한 다음 Move Up/Down을 이용하여 서울을 맨 위로 올리면 좌측과 같은 모습이 된다.
- 다음은 1 Set을 선택하고 우측 마우스 버튼을 눌러 Add After를 선택한 다음 Group Size를 5로 지정하면 우측의 모양이 된다.



ARX-Anonymize

- 다음은 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.



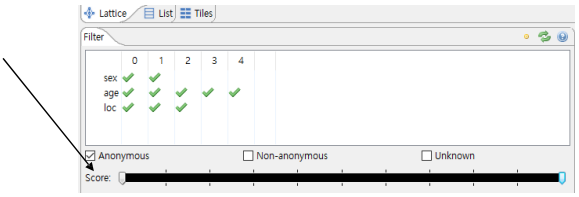
ARX-Explore

- Privacy Model을 지정한 후, 아래의 그림에서 Anonymize 버튼을 누르면 선택한 비식별화 모형을 만족하는 All possible level 조합의 모형이 explore 화면에 나타난다.

- 아래 그림에서 밑부분 {0,1,0}이 가장 낮은 level의 조합이고 위의 {1,4,2}가 가장 높은 level의 조합이다.
- {1,3,0}의 의미는 좌측부터 sex는 level 1, age는 level3, loc는 level 0의 변환을 의미한다.
- Tip: Dragging the mouse while holding the left button will move the current section of the solution space. Dragging the mouse while holding the right button will zoom in and out of the current view

ARX-Explore

- 그림에서 녹색은 사용자가 지정한 프라이버시 모형에 부합되는 결과 집합이고, 빨강색은 그렇지 못한 결과집합이다. 노란색은 그 중에서도 Optimal Solution을 나타낸다.
- 아래 그림에서 Score는 Information loss로 Solution Space의 결과가 너무 많을 때, 슬라이더 바를 움직여 Solution Space의 결과들의 수를 줄여 볼 수 있다.
- 만약, Information Loss를 최소화 한다면 낮은 Score 영역을 사용하고 재식별 가능성을 최소화하려면 높은 Score 영역을 사용하면 된다. 문제는 이 수 많은 조합 중에 어떤 Transformation을 선택하는가 이다.
- 이에 대한 기준은 Information Loss와 재 식별 Risk에 대한 고려이다.
- 여기서, 어떤 조합을 선택할 것인가를 결정하는 것이다. 여기서는 {1,1,1} 변환을 선택해 보자.



Information Loss가 큰모형

ARX-Transformation

- 그 결과, Input Data에는 총 40개의 관측값이 있고, 각 관측값은 40개의 클래스로 이루어져 있었는데 변환 후, 40개의 레코드는 그대로 있고, 클래스는 10개로 변했다. 즉, 클래스 당 4개의 관측값이 만들어 졌다.
- 여기서, 중요한 것은 변환 후, 레코드의 수가 얼마로 변하는 가이다. 레코드가 많이 줄어버리면 좋은 변환이 아니다.

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

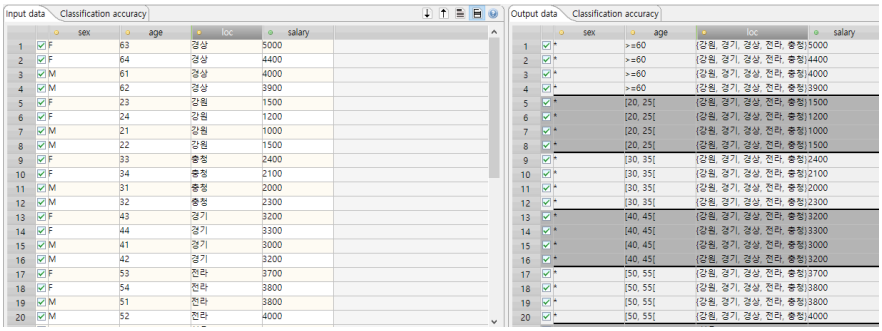
Measure	Including outliers
Average class size	1 (2.5%)
Maximal class size	1 (2.5%)
Minimal class size	1 (2.5%)
Number of classes	40
Number of records	40
Suppressed records	0 (0%)

Summary statistics | Distribution | Contingency | Class sizes | Properties | Local recoding

Measure	Including outliers	Excluding outliers
Average class size	4 (10%)	4 (10%)
Maximal class size	4 (10%)	4 (10%)
Minimal class size	4 (10%)	4 (10%)
Number of classes	10	10
Number of records	40	40 (100%)
Suppressed records	0 (0%)	0

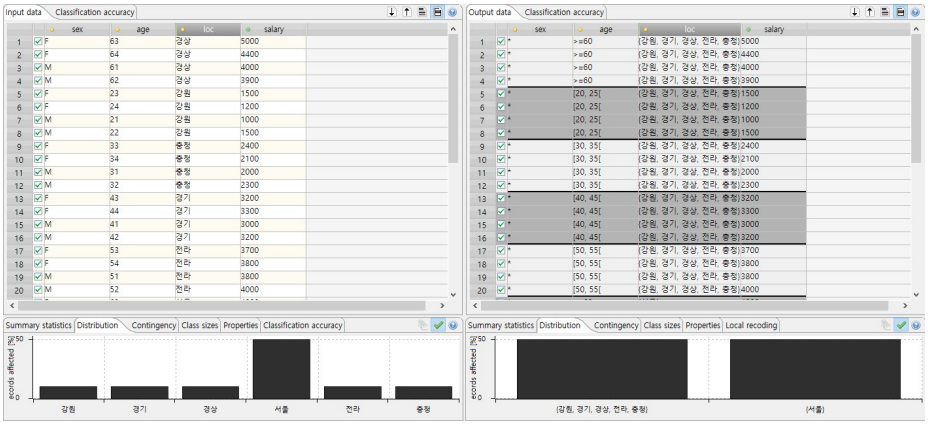
ARX-Transformation

- {1,1,1} 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.



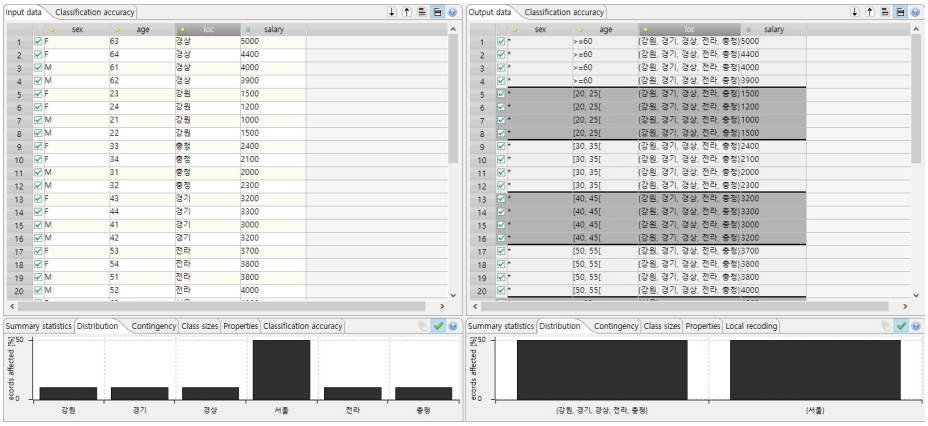
ARX-Transformation

- 이 자료에서 salary를 sensitive로 바꾸고 2-다양성을 추가해도 결과는 달라지지 않는다.
- 아래의 그림에서 알 수 있듯이 모든 준 식별자 조합에서 2개 이상의 다양한 salary 값이 존재한다.



ARX-Transformation

- 아래 그림은 {1,1,1} 변환에 대해 각 비 식별자가 어떻게 변하는가를 보여주는 결과이다.
- 지역 {서울, 경기, 충청, 강원, 전라, 경상}을 {서울, 지방}으로 단순화 됨을 알 수 있다.



ARX-Risk Analysis

- 다음은 Risk Analysis이다.
- ARX에서는 재식별 가능성에 대한 분석을 3가지 모형으로 수행한다.
  - Prosecutor Model: 이는 특정 개인이 이 Dataset에 포함되어 있음을 알고 있다는 가정하에 그 개인을 찾을 수 있는 가능성을 분석하는 것이다.
  - Journalist model은 아래의 그림처럼 다른 데이터 베이스와 결합하여 개인 식별 가능성을 분석하는 것이다.
  - marketer model은 개인을 식별하는 것에는 관심이 없고, 집단의 부분집합 전체가 식별되는 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 재식별할 수 있는 가능성이 "0"이 된다는 것이 아님을 이해해야 한다.

Data Considered for Sharing				Voter Registration Records (Identified Resource)			
Age	Zip Code	Gender	Diagnosis	Birthdate	Zip Code	Gender	Name
15	00000	Male	Diabetes	2/2/1989	00001	Female	Alice Smith
21	00001	Female	Influenza	3/3/1974	10000	Male	Bob Jones
36	10000	Male	Broken Arm	4/4/1919	10001	Female	Charlie Doe
91	10001	Female	Acid Reflux				

Figure 3. Linking two data sources to identity diagnoses.

ARX-Risk Analysis

- combinations of variables separate the records from each other and to which degree the variables make records distinct.
- 준 식별자 조합에 대해 Distinct 한 준 식별자는 결과적으로 De-identification Risk를 증가시키는 변수이므로 분석에 크게 중요하지 않으면 삭제하거나 변환단계에서 level을 증가시키는 것이 좋다.
- 아래의 결과에서 age의 Distinct 값이 50% 였는데 변환을 통해 12.5%로 감소 했고, sex는 5%→2.5%로 감소하여 Risk를 줄였음을 알 수 있다.

Input Data

Quasi-identifiers | Re-identification ri | HIPAA identifiers

Quasi-identifier	Distinct...	Separat...
sex	5%	51.28205%
loc	15%	71.79487%
age	50%	97.4359%
salary	57.5%	96.92308%
sex, loc	30%	87.17949%
sex, age	50%	97.4359%
sex, salary	77.5%	98.46154%
loc, salary	77.5%	98.71795%
age, salary	92.5%	99.61538%
age, loc	100%	100%
sex, age, salary	92.5%	99.61538%
sex, loc, salary	92.5%	99.61538%
age, loc, salary	100%	100%
sex, age, loc	100%	100%
sex, age, loc, salary	100%	100%

Transformed Data

Distribution of risk | Distribution of risk | Quasi-identifiers

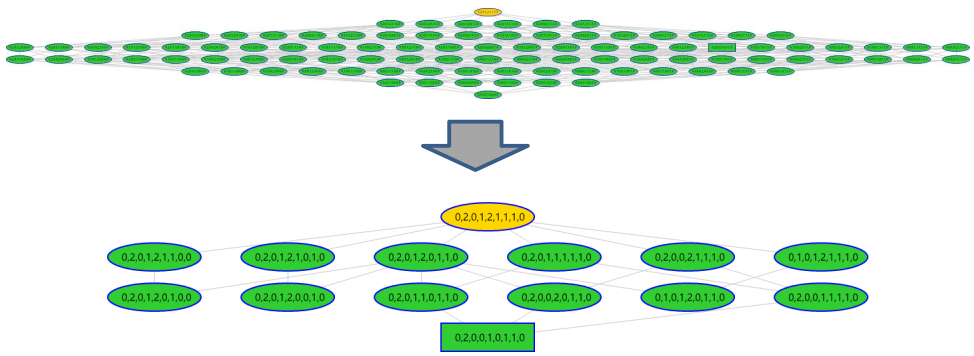
Quasi-identifier	Distincti...	Separati...
sex	2.5%	0%
age	12.5%	82.05128%
loc	15%	71.79487%
salary	57.5%	96.92308%
sex, age	12.5%	82.05128%
sex, loc	15%	71.79487%
age, loc	25%	92.30769%
sex, salary	57.5%	96.92308%
age, salary	75%	98.71795%
loc, salary	77.5%	98.71795%
sex, age, loc	25%	92.30769%
sex, age, salary	75%	98.71795%
sex, loc, salary	77.5%	98.71795%
age, loc, salary	90%	99.48718%
sex, age, loc, salary	90%	99.48718%

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
- highest risk: Highest risk of a single record
- Success rate: Proportion of records that can be re-identified on average

ARX-Example

- http://arx.deidentifier.org/downloads/ 에서 제공되는 예제 파일로 example.deid 파일이 있다.
- 총 30,162건의 자료이고, 변수는 sex, age, race, marital status, education, native-country, workclass, occupation, salary-class가 있다.
- 5-Anonymity 모형으로 식별화를 진행해보자.
- 5-Anonymity 모형을 만족하는 조합 중에 Information loss를 최소화하는 부분 집합을 선정해 보았다.



ARX-Example

- 여기서, 우리는 변환을 선택할 것인가?
- ARX는 {0,2,0,1,2,1,1,1,0}을 Optimal로 표시하고 있는데 밑 부분 {0,2,0,0,1,0,1,1,0}와 어떤 차이가 있는가?
- 두 모형의 차이는 아래와 같다.

marital status

Level-0	Level-1
Divorced	Spouse not pres...
Never-married	Spouse not pres...
Separated	Spouse not pres...
Widowed	Spouse not pres...
Married-spouse-...	Spouse not pres...
Married-AF-spo...	Spouse present

education

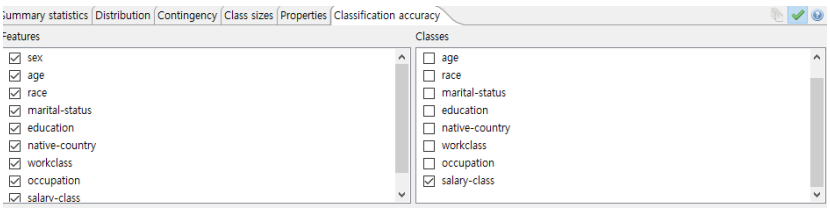
Level-0	Level-1	Level-2
Bachelors	Undergraduate	Higher education
Some-college	Undergraduate	Higher education
11th	High School	Secondary educ...
HS-grad	High School	Secondary educ...
Prof-school	Professional Edu...	Higher education
Assoc-acdm	Professional Edu...	Higher education
Assoc-voc	Professional Edu...	Higher education
9th	High School	Secondary educ...
7th-8th	High School	Secondary educ...
12th	High School	Secondary educ...
Masters	Graduate	Higher education
1st-4th	Primary School	Primary education
10th	High School	Secondary educ...
Doctorate	Graduate	Higher education
5th-6th	Primary School	Primary education
Preschool	Primary School	Primary education

native-country

Level-0	Level-1
Philippines	Asia
Italy	Europe
Poland	Europe
Jamaica	North America
Vietnam	Asia
Mexico	North America
Portugal	Europe
Ireland	Europe
France	Europe
Dominican-Rep...	North America
Laos	Asia
Ecuador	South America
Taiwan	Asia
Haiti	North America
Columbia	South America
Hungary	Europe
Guatemala	North America
Nicaragua	South America
Scotland	Europe
Thailand	Asia
Yugoslavia	Europe
El-Salvador	North America
Trinadad&Tobago	South America
Peru	South America
Hong	Asia

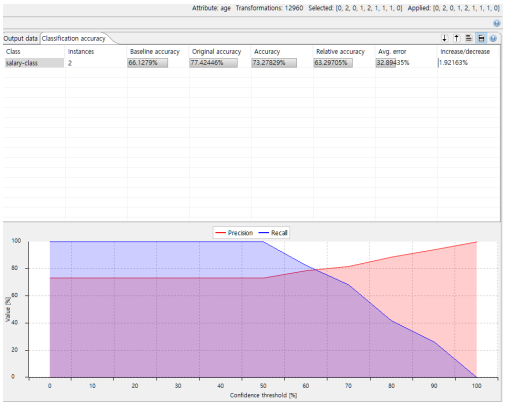
ARX-Example

- {0,2,0,1,2,1,1,1,0}는 {0,2,0,0,1,0,1,1,0}의 모형에서 marital status, education, native-country를 한 수준씩 더 가공한 것이다. 가공을 더 하면 Information Loss가 커지는데 얼마나 더 커질까 가늠하기는 어렵다.
- 이 데이터의 목적은 아마도 age, race, marital status, native-country, work-class, occupation이 salary-class와 어떤 관계가 있는지를 분석하는 것일 것이다.
- 통계학적으로 비 식별화 가공을 하면 이러한 관계 모형의 Power가 약해질 수 있다.
- ARX에서는 Logistic Regression, OneR의 방법론을 이용하여 Input Data의 예측 정확도와 변환 후, Output Data 의 정확도를 비교하여 손실이 작은 모형을 선택할 수 있도록 도와주고 있다.
- 아래의 그림은 Feature 즉, 독립변수를 아래와 같이 선택하고 종속변수를 salary-class로 지정하여 Logistic Regression을 통해 예측 정확도를 검사하는 화면이다.



ARX-Example

- Salary-Class는 <= 50k가 66.12%로 아무런 정보가 없어도 기본적으로 66.12%의 정확도를 가진다. 변환전 Input Data를 통한 정확도는 77.42%이고, {0,2,0,1,2,1,1,1,0} 변환 Output Data를 통한 정확도는 73.27%로 정확도가 낮아진 것을 확인할 수 있다.
- 정확도를 많이 낮추는 변환은 비 식별화 결과는 만족할 수 있으나 Data로서의 가치가 낮아져 의미없는 작업이 될 가능성이 크다.



ARX-Example

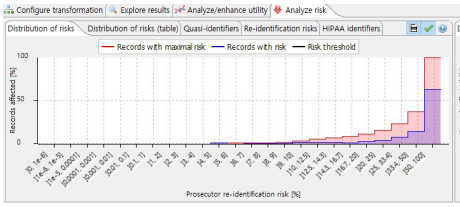
- {0,2,0,1,2,1,1,1,0}의 Information Loss는 28%이고, {0,2,0,0,1,0,1,1,0}는 34%이다.
- {0,2,0,1,2,1,1,1,0}의 Suppressed Records는 743이고 {0,2,0,0,1,0,1,1,0}는 1,603이다.
- {0,2,0,1,2,1,1,1,0} 은 준 식별자의 변환이 더 많지만 Suppressed Records의 수는 작고,
- {0,2,0,0,1,0,1,1,0} 는 준 식별자 변환은 작지만 Suppressed Records의 수가 크다.
- 5-Anonymity 모형은 식별자/준식별자 조합에서 최소 5개 이상의 관측값이 있어야 한다.
- 이 조건을 만족하지 못하는 Records는 모두 Suppressed Records로 처리된다.
- 이 자료에서 4950번 케이스 처럼 74세의 여자는 70세 이상의 나이를 >=70 으로 처리하는 것도 방법이다.

4946	Male	61	White	Married-civ-spouse		4946	Male	[61, 70]	White	Spouse present	Se
4947	Male	61	White	Married-civ-spouse		4947	Male	[61, 70]	White	Spouse present	Se
4948	Male	61	White	Married-civ-spouse		4948	Male	[61, 70]	White	Spouse present	Se
4949	Male	61	White	Married-civ-spouse		4949	Male	[61, 70]	White	Spouse present	Se
4950	Female	74	White	Divorced		4950	*	*	*	*	*
4951	Female	71	White	Never-married		4951	*	*	*	*	*
4952	Female	75	White	Widowed		4952	*	*	*	*	*
4953	Female	73	White	Married-civ-spouse		4953	*	*	*	*	*

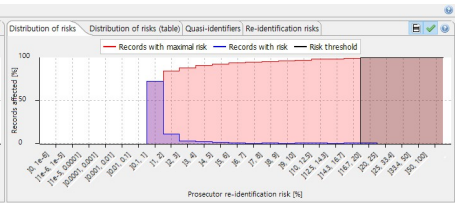
ARX-Example

- 아래의 그림은 Prosecutor Re-identification 위험의 크기에 따른 레코드 비율을 그래프로 보여주고 있다.
- 잘 된 비 식별화는 우측처럼 파란색이 그래프 좌측에 많이 있어야 한다.
- 변환 전 Data는 파란색 그래프 즉 Records with risk가 그래프 우측에 많이 분포되어 있는 반면
- 변환 후 Data는 파란색 그래프가 좌측에 분포되어 이 변환을 통해 재 식별 가능성이 많이 낮아졌음을 알 수 있다.

변환전Data



변환 후Data



ARX-Example

- 준 식별자 조합에 대한 유일성도 변환 후, 파격적으로 낮아짐을 알 수 있다.

Distribution of risk | Distribution of risk | Quasi-identifiers | Re-identification ri | HIPAA identifiers

Quasi-identifier	Distinction	Separation
sex, age, race, marital-status, native-country, occupation, salary-class	31.07553%	99.954%
sex, age, race, native-country, workclass, occupation, salary-class	31.08879%	99.95183%
age, race, marital-status, education, native-country, workclass, salary-class	35.57125%	99.91662%
sex, age, marital-status, education, native-country, workclass, salary-class	36.12824%	99.9378%
sex, age, race, marital-status, education, native-country, workclass	36.7648%	99.93571%
sex, age, race, marital-status, education, workclass, salary-class	37.05656%	99.94501%
sex, age, marital-status, native-country, workclass, occupation, salary-class	37.25549%	99.96374%
age, race, marital-status, native-country, workclass, occupation, salary-class	37.45441%	99.95673%
sex, age, race, marital-status, native-country, workclass, occupation	37.86553%	99.96093%
sex, age, race, marital-status, workclass, occupation, salary-class	38.36616%	99.96646%
sex, age, race, education, native-country, occupation, salary-class	41.82415%	99.97782%
age, race, education, native-country, workclass, occupation, salary-class	47.40733%	99.97764%
sex, age, race, education, native-country, workclass, occupation	48.024%	99.98045%
age, race, marital-status, education, native-country, occupation, salary-class	48.28924%	99.98007%
sex, age, education, native-country, workclass, occupation, salary-class	48.64067%	99.98143%
sex, age, marital-status, education, native-country, occupation, salary-class	48.75671%	99.98311%
sex, age, race, marital-status, education, native-country, occupation	48.97885%	99.98165%

Distribution of risks | Distribution of risks (table) | Quasi-identifiers | Re-identification risks

Quasi-identifier	Distinction	Separation
sex, age, race, marital-status, education, workclass, salary-class	1.22371%	91.24551%
sex, race, marital-status, education, native-country, occupation, salary-class	1.22371%	92.68319%
sex, race, education, native-country, workclass, occupation, salary-class	1.30019%	92.45559%
sex, marital-status, education, native-country, workclass, occupation, salary...	1.30019%	93.87126%
race, marital-status, education, native-country, workclass, occupation, salar...	1.33843%	93.09767%
age, race, education, native-country, workclass, occupation, salary-class	1.33843%	93.32348%
sex, age, race, marital-status, native-country, workclass, salary-class	1.37667%	91.56895%
sex, age, race, marital-status, education, workclass, occupation	1.41491%	94.29304%
sex, age, race, education, native-country, occupation, salary-class	1.41491%	94.30304%
age, race, marital-status, education, native-country, occupation, salary-class	1.4914%	94.69082%
sex, age, education, native-country, workclass, occupation, salary-class	1.4914%	95.15069%
sex, age, marital-status, education, native-country, occupation, salary-class	1.4914%	95.25909%
age, marital-status, education, native-country, workclass, occupation, salary...	1.52964%	95.55382%
sex, age, race, marital-status, native-country, workclass, occupation	1.54876%	94.51448%
sex, race, marital-status, education, workclass, occupation, salary-class	1.6826%	94.30345%
sex, age, race, education, workclass, occupation, salary-class	1.79732%	95.58996%
sex, age, race, marital-status, education, occupation, salary-class	1.83556%	95.67135%

ARX-Example

- {0,2,0,1,2,1,1,1,0} 변환에 대한 재식별 가능성은 거의 없음을 알 수 있다.

ARX-Sensitive Data

- ARX에서 민감정보를 다루는 방법을 알아보자.
- 아래의 자료에서 ZipCode, Age는 Quasi-Identifying이고, Salary, Disease는 Sensitive라고 하자. 여기서, Salary는 양적자료이고, Disease는 명목형 자료이다.
- Salary, Disease에 각각 l-다양성, t-근접성 모형을 추가할 수 있다.
- t-근접성 모형에서 Salary는 양적자료이므로 equal ground distance 를 사용하면 되고, Disease는 명목형 자료이므로 Hierarchical ground distance를 사용한다.
- {위염, 위궤양,위암}이 모두 위에 관련된 것이고, {기관지염, 폐렴, 감기} 폐에 관련된 것이므로 이들을 아래와 같이 그룹화한다.

	ZipCode	Age	Salary	Disease
1	47602	22	4	1.위염
2	47677	29	3	1.위궤양
3	47678	27	5	1.위암
4	47605	30	7	2.기관지염
5	47607	32	10	1.위암
6	47673	36	9	2.폐렴
7	47905	43	6	1.위염
8	47909	52	11	2.감기
9	47906	47	8	2.기관지염

Level-0	Level-1	Level-2
궤양	{1.위궤양, 1.위암, 1.위염}	*
암	{1.위궤양, 1.위암, 1.위염}	*
염	{1.위궤양, 1.위암, 1.위염}	*
기	{2.감기, 2.기관지염, 2.폐렴}	*
관지염	{2.감기, 2.기관지염, 2.폐렴}	*
렴	{2.감기, 2.기관지염, 2.폐렴}	*

ARX-Sensitive Data

- 아래는 l-다양성과 t-근접성을 설정하는 화면이다.
- Distinct-l-diversity를 Variant로 선택할 경우는 l 값만 결정하면 된다.
- t값은 0~1사이의 값으로 일단, 0.9와 같이 1에 가까운 값부터 출발하여 점차 줄여가면서 적정한 모형을 찾아야 한다.