NIA 한국정보화진흥원

개인정보 비 식별화
(De-identification)

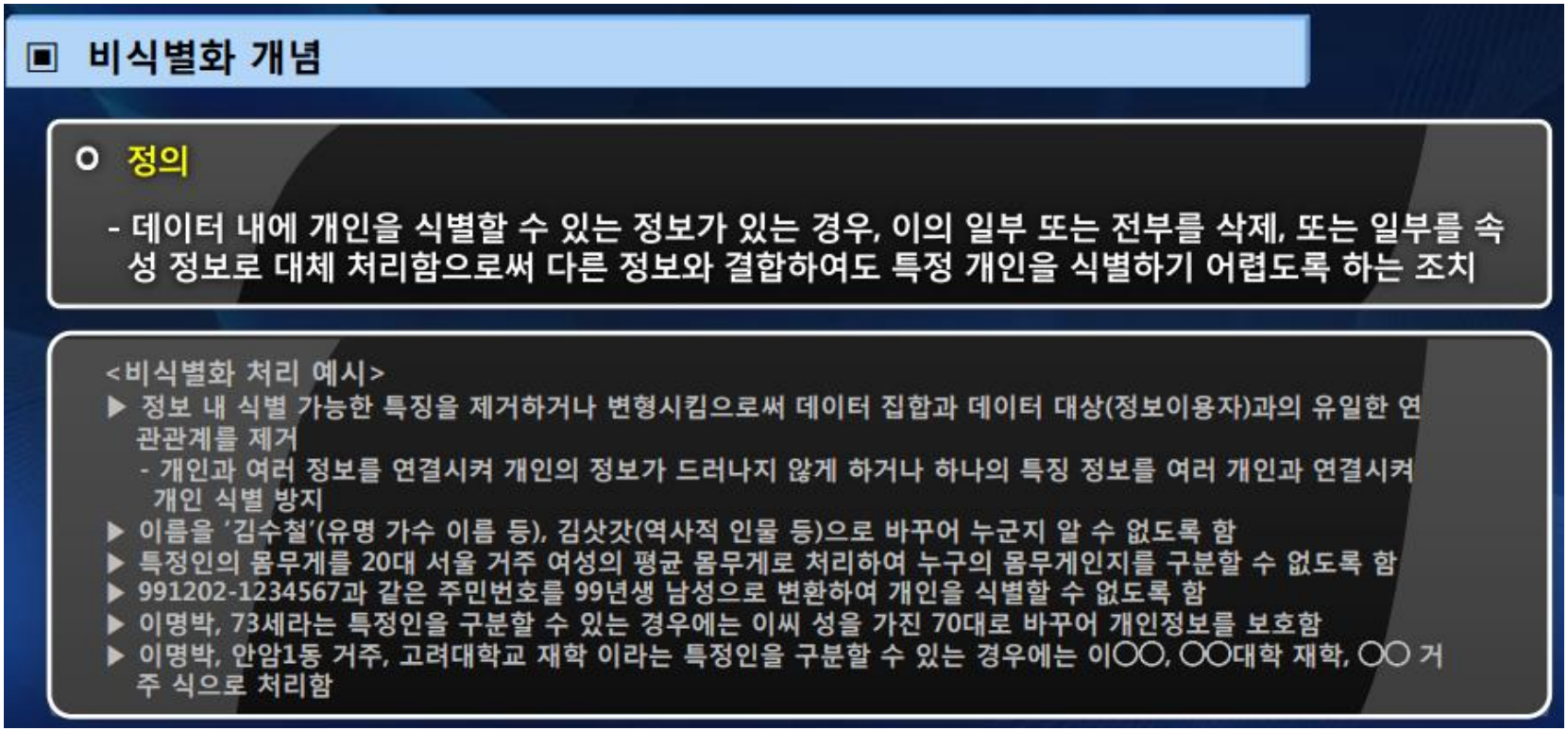
-ARX 실습-

인하대학교

김승환

swkim4610@inha.ac.kr

비 식별화 개념



비 식별화(De-identification)는 보유한 데이터가 가지고 있는 정보를 일부분 삭제, 대체 처리하여 특정 개인을 식별할 수 없도록 하는 방법론이다. 비 식별화의 이슈는 익명성 수준과 정보가치의 조화 문제이다.

잘된 비 식별화 처리는 재 식별 가능성을 일정 수준 이하로 낮추면서 정보로서의 가치가 유지 되도록 하는 것이다.

비 식별화 방법

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
2	[0, 5[	[0, 10[	[0, 20[	[0, 40[	*
3	[0, 5[	[0, 10[	[0, 20[	[0, 40[	*
4	[0, 5[	[0, 10[	[0, 20[	[0, 40[	*
5	[5, 10[	[0, 10[	[0, 20[	[0, 40[	*
6	[5, 10[	[0, 10[	[0, 20[	[0, 40[	*
7	[5, 10[	[0, 10[	[0, 20[	[0, 40[	*
8	[5, 10[	[0, 10[	[0, 20[	[0, 40[	*
9	[5, 10[	[0, 10[	[0, 20[	[0, 40[	*
10	[10, 15[	[10, 20[	[0, 20[	[0, 40[	*
11	[10, 15[	[10, 20[	[0, 20[	[0, 40[	*
12	[10, 15[	[10, 20[	[0, 20[	[0, 40[	*
13	[10, 15[	[10, 20[	[0, 20[	[0, 40[	*
14	[10, 15[	[10, 20[	[0, 20[	[0, 40[	*
15	[15, 20[	[10, 20[	[0, 20[	[0, 40[	*
16	[15, 20[	[10, 20[	[0, 20[	[0, 40[	*
17	[15, 20[	[10, 20[	[0, 20[	[0, 40[	*
18	[15, 20[	[10, 20[	[0, 20[	[0, 40[	*
19	[15, 20[	[10, 20[	[0, 20[	[0, 40[	*
20	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*
21	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*
22	[20, 25[	[20, 30[	[20, 40[	[0, 40[	*

위의 그림은 Level-0의 나이를 변환하는 과정이다. Level이 높아지면서 재식별 가능성은 낮아지고, 나이의 정보의 가치는 떨어진다. Level-1은 정보가치는 좋지만 재식별 가능성은 높을 수 있다. 반면, Level-4는 정보 가치가 낮고 재식별 가능성은 낮다.

비 식별화 방법

일반적 기법 : 개인 식별요소 삭제 방법

처리기법	예시	세부기술
가명처리 (Pseudonymization)	○ 홍길동, 35세, 서울 거주, 한국대 재학 → 임꺽정, 30대, 서울 거주, 국제대 재학	①휴리스틱 가명화 ②암호화 ③교환 방법
총계처리 (Aggregation)	○ 임꺽정 180cm, 홍길동 170cm, 이콩쥐 160cm, 김팥쥐 150cm → 물리학과학생키합:660cm, 평균키 165cm	④총계처리 ⑤부분총계 ⑥라운딩 ⑦재배열
데이터 삭제 (Data Reduction)	○ 주민등록번호 901206-1234567 → 90년대 생, 남자 ○ 개인과 관련된 날짜정보(합격일 등)는 연단위로 처리	⑧식별자 삭제 ⑨식별자 부분삭제 ⑩레코드 삭제 ⑪식별요소 전부삭제
데이터 범주화 (Data Suppression)	○ 홍길동, 35세 → 홍씨, 30 ~ 40세	⑫감추기 ⑬랜덤 라운딩 ⑭범위 방법 ⑮제어 라운딩
데이터 마스킹 (Data Masking)	○ 홍길동, 35세, 서울 거주, 한국대 재학 → 홍○○, 35세, 서울 거주, ○○대학 재학	⑯임의 잡음 추가 ⑰공백과 대체

주민번호, 회원번호 등 식별자는 제거가 가장 좋다. 제거할 수 없다면 가명처리 혹은 마스킹을 많이 사용한다.

나이, 지역, 회사명 등 준 식별자는 분석 목적에 크게 방해되지 않는 범위 내에서 범주화하거나 총계처리한다.

비 식별화 방법

기법	의미	적용례
k-익명성	특정인임을 추론할 수 있는지 여부를 검토, 일정 확률수준 이상 비식별 되도록 함	동일한 값을 가진 레코드를 k개 이상으로 함. 이 경우 특정 개인을 식별할 확률은 1/k임
l-다양성	특정인 추론이 안된다고 해도 민감한 정보의 다양성을 높여 추론 가능성을 낮추는 기법	각 레코드는 최소 l개 이상의 다양성을 가지도록 하여 동질성 또는 배경지식 등에 의한 추론 방지
t-근접성	l-다양성 뿐만 아니라, 민감한 정보의 분포를 낮추어 추론 가능성을 더욱 낮추는 기법	전체 데이터 집합의 정보 분포와 특정 정보의 분포 차이를 t이하로 하여 추론 방지

* k, l, t값은 전문가 등이 검토하여 마련

비 식별화에서 가장 기본적인 기법은 k-익명성(k-Anonymity)이다.

만약, **민감정보가 포함된다면 l-다양성(l-Diversity), t-근접성(t-Closeness)의 방법론도 고려해야 한다.**

비 식별화 방법

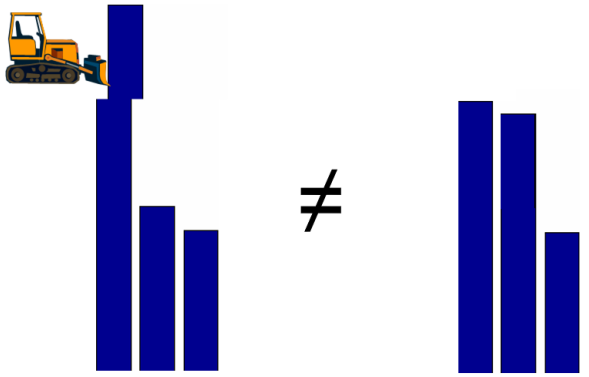
k-익명성, l-다양성은 단순해서 이해가 쉽지만 t-근접성은 통계적 개념으로 직관적으로 설명할 수 없다.

t 근접성은 두개의 분포에 대한 유사성을 거리로 측정하는 아이디어이다.

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t.
A table is said to have t-closeness if all equivalence classes have t-closeness.

즉, k-익명성, l-다양성을 만족하더라도 해당 조합의 민감정보 분포가 다른 조합의 분포와 차이가 크면 분포의 차이로 인해 민감정보가 유출되는 것이므로 모든 가능한 조합 n개에 대해 구한 n(n-1)/2 개의 거리가 임계값 t 보다 작게 만들자는 개념이다. 이 때, 두 분포의 거리를 계산하는 방법론으로 통계학에서 사용하는 EMD(Earth Mover's Distance)를 사용한다. 여기서, Earth Mover는 토사를 굴삭하여 목적장소로 압출하는 도구로 트랙터에 장치하여 사용라는 장비를 말한다.

Definition:
The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other.



비 식별화 방법

아래의 자료에서 Salary, Disease를 민감정보라고 할 때, 원자료 Table 3을 Table 4로 비 식별화 했다고 하자.

아래 자료는 3-익명성, 3-다양성을 만족한다.

그런데, ZipCode 476** 의 Salary <= 5K로 상대적으로 급여액이 작다는 정보가 존재한다.

이유는 Salary {3K, 4K, 5K}와 {6K, 11K, 8K} 그리고 {7K, 9K, 10K}와의 거리를 보면 직관적으로 {3K, 4K, 5K}와 다른 조합과의 거리가 큼을 알 수 있다.

Table 4에서 구한 EMD의 최대값이 비식별화 기준으로 제시된 임계값 t 보다 크면 이 자료는 t 근접성을 만족하지 않는다고 판단한다.

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

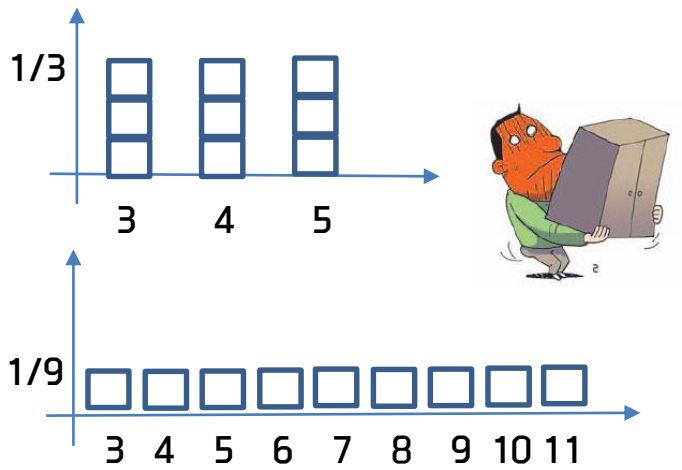
Table 4. A 3-diverse version of Table 3

비 식별화 방법

참고

{3K, 4K, 5K}와 {3K, 4K, 5K, 6K, 7K, 8K, 9K, 10K, 11K}의 EMD를 계산해 보자.

EMD는 위의 히스토그램을 아래의 히스토그램으로 바꾸는데 필요한 일의 량이다.



1/9 무게를 {5→11}, {5→10}, {5→9}, {4→8}, {4→7}, {4→6}, {3→5}, {3→4}로 6+5+4+4+3+2+2+1=27칸 이동하였다. 이는 평균 27/8= 3.375칸 이동한 것이다. 각 이동에 1/9 무게이므로 일의 량은 1/9 * 3.375 = 0.375 이다.

같은 방법으로 ZipCode 4790* 의 Salary {6K, 11K, 8K}와 {3K, 4K, 5K, 6K, 7K, 8K, 9K, 10K, 11K}의 거리를 계산하면 1/9 * 12/8 칸= 0.167이다.

비 식별화 방법

Table 4. 에서 Disease는 민감정보이지만 거리를 계산할 수 없는 형태이다. 하지만, {위궤양, 위병, 위암}은 모두 위에 관련된 것이란 정보가 존재한다. 이 경우, t 근접성은 거리를 계산하여야 하는데 문제는 명목척도이기 때문에 거리 계산이 불가능하다. 이에 대한 해결방법으로 병에 대한 위계도(Hierarchy plot)를 이용하는 방법을 고려할 수 있다.

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

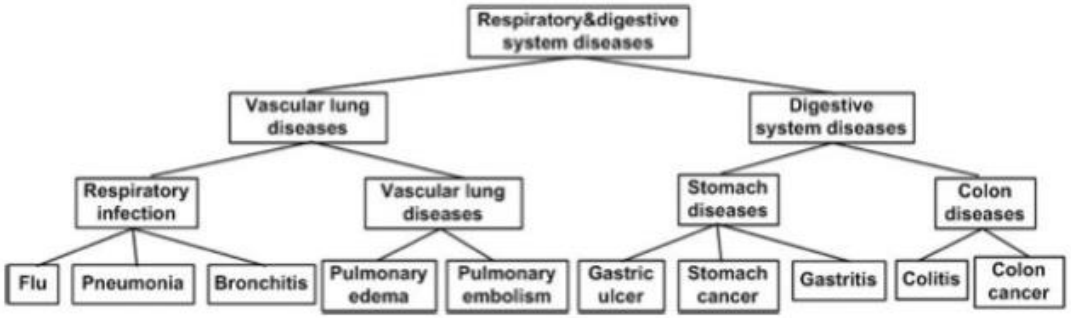


Figure 1. Hierarchy for categorical attributes *Disease*.

$$hierarchical - dist(v_i, v_j) = \frac{level(v_i, v_j)}{H}$$

예를 들어, Gastric Ulcer와 Gastritis는 거리가 "1/3"이고, Colitis는 거리가 "2/3"이다. 또한, Flu와의 거리는 "3/3"이다. 즉, 자기의 직계의 level수를 총 level 수로 나눈 값이다.

비 식별화의 어려움

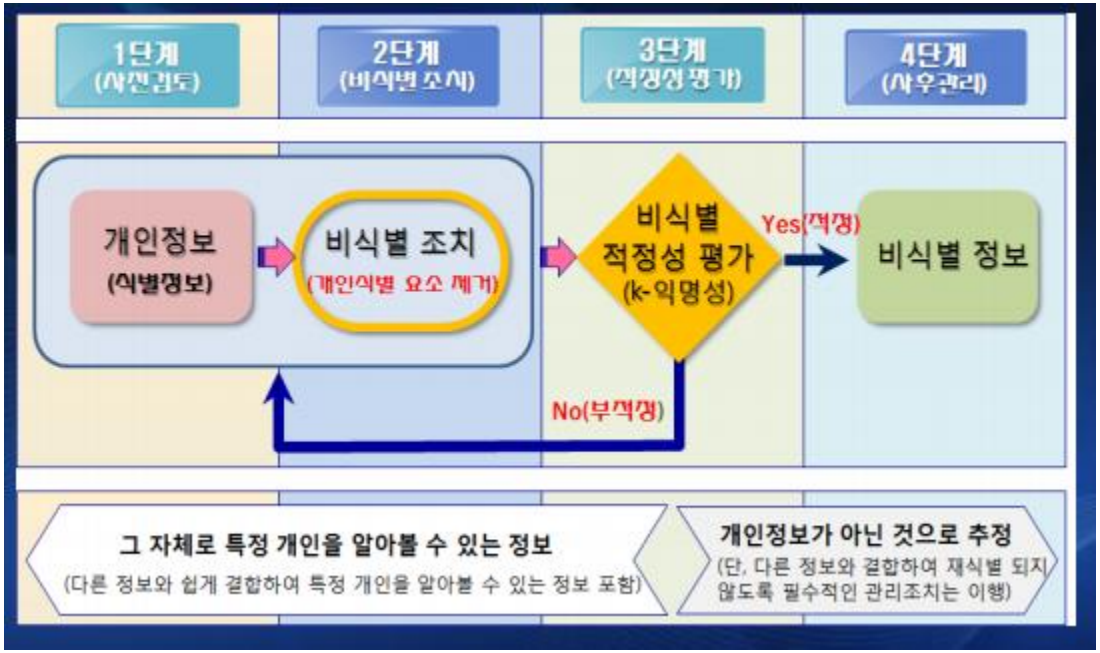
만약, 보유 정보가 다음과 같은 경우에 해당한다면 k-익명성 충족이 어려울 수 있으므로 적절한 조치를 취해야 한다.

- 준 식별자의 수가 너무 많은 경우 → 분석에 불필요한 준 식별자는 삭제한다.
  준 식별자가 증가하면 Population Uniqueness Issue가 발생한다.
- 특정 준 식별자 조합에 해당하는 Record 수가 k 미만인 경우(1명만 제주도 거주하는 경우 등)
  → 빈도분석을 통해 해당 관측값을 찾고 분석에 크게 문제가 없다면 이를 삭제한다.
- 심한 비대칭분포 → 연속변수의 경우는 로그변환 등을 통해 균등분포 혹은 정규분포 모양을 만드는 것이 좋다.
  명목변수인 경우, 빈도가 작은 Cell을 Pooling하여 일정 수 이상의 빈도를 확보하는 것이 좋다.

민감 정보가 포함되어 있는 경우, l-다양성이나 t-근접성을 만족해야 하는데 보유 정보가 다음과 같은 경우에 해당한다면 l-다양성이나 t-근접성 충족이 어려울 수 있으므로 적절한 조치를 취해야 한다.

- 준식별자와 민감정보의 상관 관계가 큰 경우 → 관측치의 개수를 늘려 같은 준식별자 수준에서 다양성을 만족할 수 있도록 해야 한다.
- 민감정보가 여러 개 존재하는 경우→ 민감정보는 각각에 대해 , l-다양성이나 t-근접성 기준을 설정하므로 여러 개를 동시에 설정할 경우, 특정 정보 하나가 기준을 만족하지 않아 모두 만족하지 않는 것처럼 변환될 수 있다.
  이 경우, 그 특정 정보를 삭제할 수 있다면 좋은 변환을 할 수 있다.

비 식별화 프로세스



3단계에서 비 식별 적정성 평가를 만족하는 여러 변환방법 중 분석에 적합하고 정보손실이 가장 작은 변환 방법을 택하여 4단계로 진행한다.

비 식별화 연습

예를 들어 아래와 같이 성별, 나이, 지역, 연봉 자료가 있다고 가정하자.

아래의 자료에서 성별, 나이, 지역이 준 식별자이고, 연봉은 민감정보라고 할때, 비 식별화는 아래의 정보로 부터 특정 개인의 연봉을 추정할 수 있는 가능성을 낮추는 작업이다.

만약, (m, 21, 강원) 조합의 직원이 사내에 1명이라고 가정하면 이 정보로 부터 특정인을 찾을 수 있어 민감정보가 노출된 것이다.

Raw Data			
성별	나이	지역	연봉
M	21	강원	1000
M	22	강원	1000
F	23	강원	1500
F	24	강원	1200
M	31	충청	2000
M	32	충청	2300
F	33	충청	2400
F	34	충청	2100
M	41	경기	3000
M	42	경기	3200
F	43	경기	3200
F	44	경기	3300

비 식별화 연습

좌측 정보를 가운데와 같이 변환하게 되면 일단, 모든 준 식별자 조합에 대해 최소 2개씩의 중복된 자료가 존재하므로 2-익명성을 만족한다. 하지만, (M, 20대, 강원) 조합의 직원이 2명이고 이들의 연봉은 모두 1000만원으로 같아 사실 상, 식별이 가능한 상황이다. 만약, l=2를 최소한으로 권고한다면 이 변환은 2-다양성을 만족하지 않는다.

우측의 변환은 평균연봉을 사용하여 비식별화를 시도하였다.

우측의 형태는 개인의 연봉을 식별할 수는 없으나 나이에 따라 연봉의 차이가 크다는 정보가 남아 있다.

연봉이 민감정보이므로 이 역시, 비 식별화가 잘된 경우는 아니다.

Raw Data

성별	나이	지역	연봉
M	21	강원	1000
M	22	강원	1000
F	23	강원	1500
F	24	강원	1200
M	31	충청	2000
M	32	충청	2300
F	33	충청	2400
F	34	충청	2100
M	41	경기	3000
M	42	경기	3200
F	43	경기	3200
F	44	경기	3300

Transformed Data

성별	나이	지역	연봉
M	20대	강원	1000
M	20대	강원	1000
F	20대	강원	1500
F	20대	강원	1200
M	30대	충청	2000
M	30대	충청	2300
F	30대	충청	2400
F	30대	충청	2100
M	40대	경기	3000
M	40대	경기	3200
F	40대	경기	3200
F	40대	경기	3300

Transformed Data

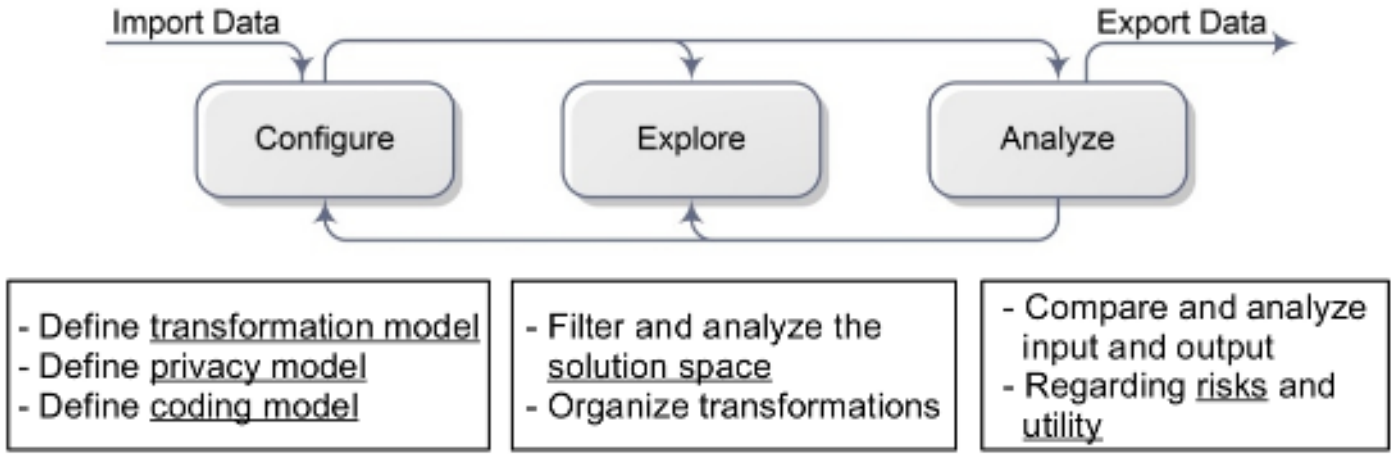
성별	나이	지역	평균연봉
M	20대	강원	1000
F	20대	강원	1350
M	30대	충청	2150
F	30대	충청	2250
M	40대	경기	3100
F	40대	경기	3250

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.

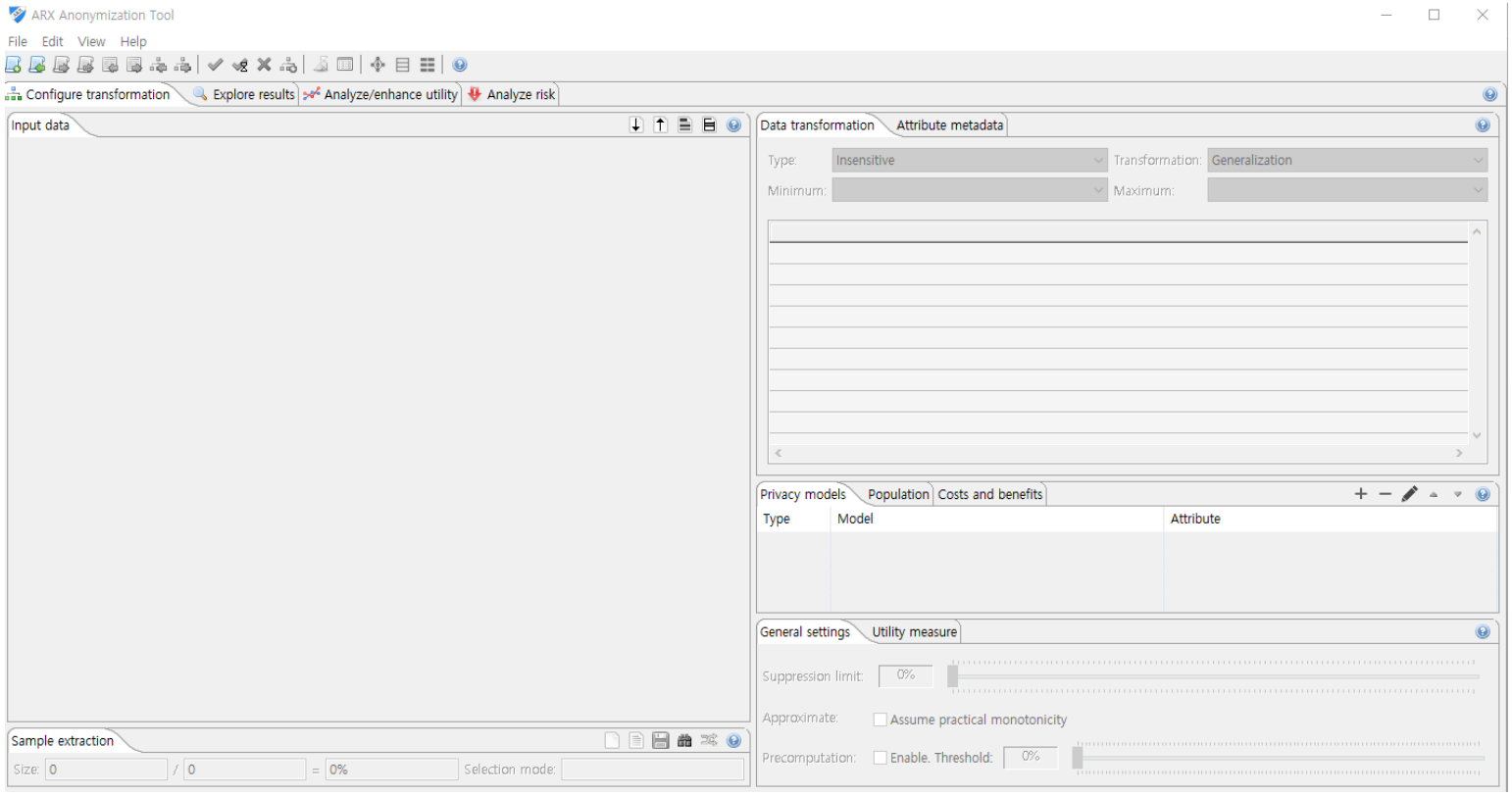
Analyze 단계에서는 재식별화 가능성 등 위험수준을 분석하여 최종 Export 여부를 결정하게 된다.

ARX 설치

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

윈도우 O.S는 arx-3.5.1-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. 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.

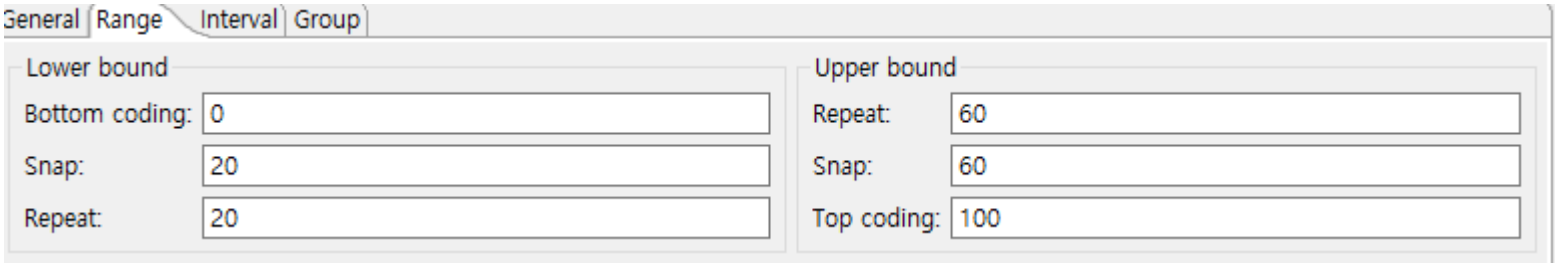
같은 방식으로 age는 Interval, loc는 ordering으로 해보자.

ARX-Configure

age interval은 아래와 같이 만든다.

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

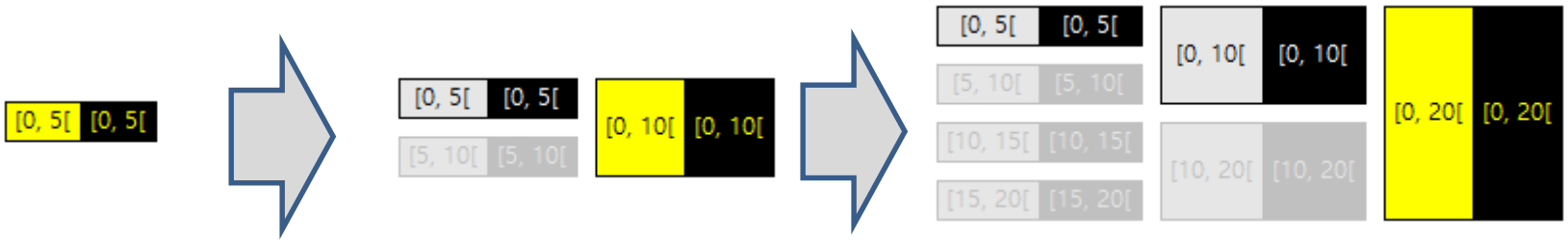
아래의 창에서 Lower, Upper Bound는 아래와 같이 입력한다. 0~100세 사이를 <=20, >=60은 나누지 않고 나머지 구간에 대해 처리한다는 의미이다.



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[	*

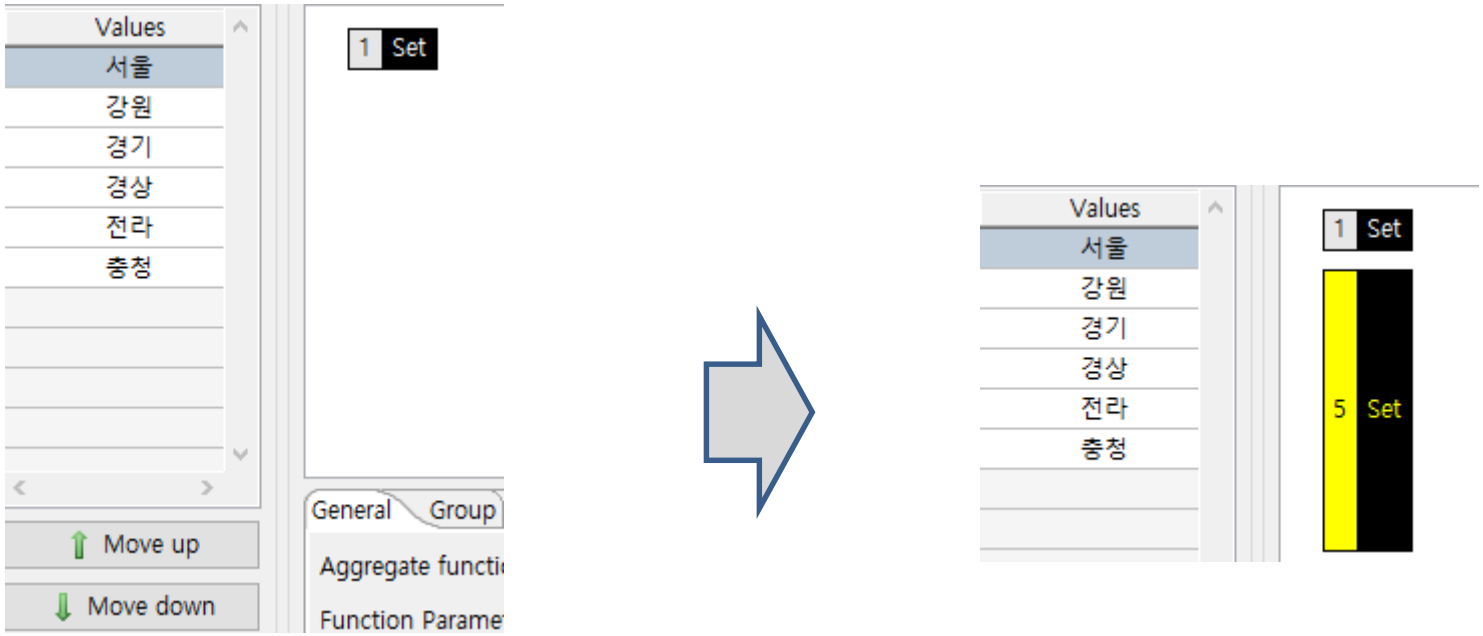
ARX-Configure

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.



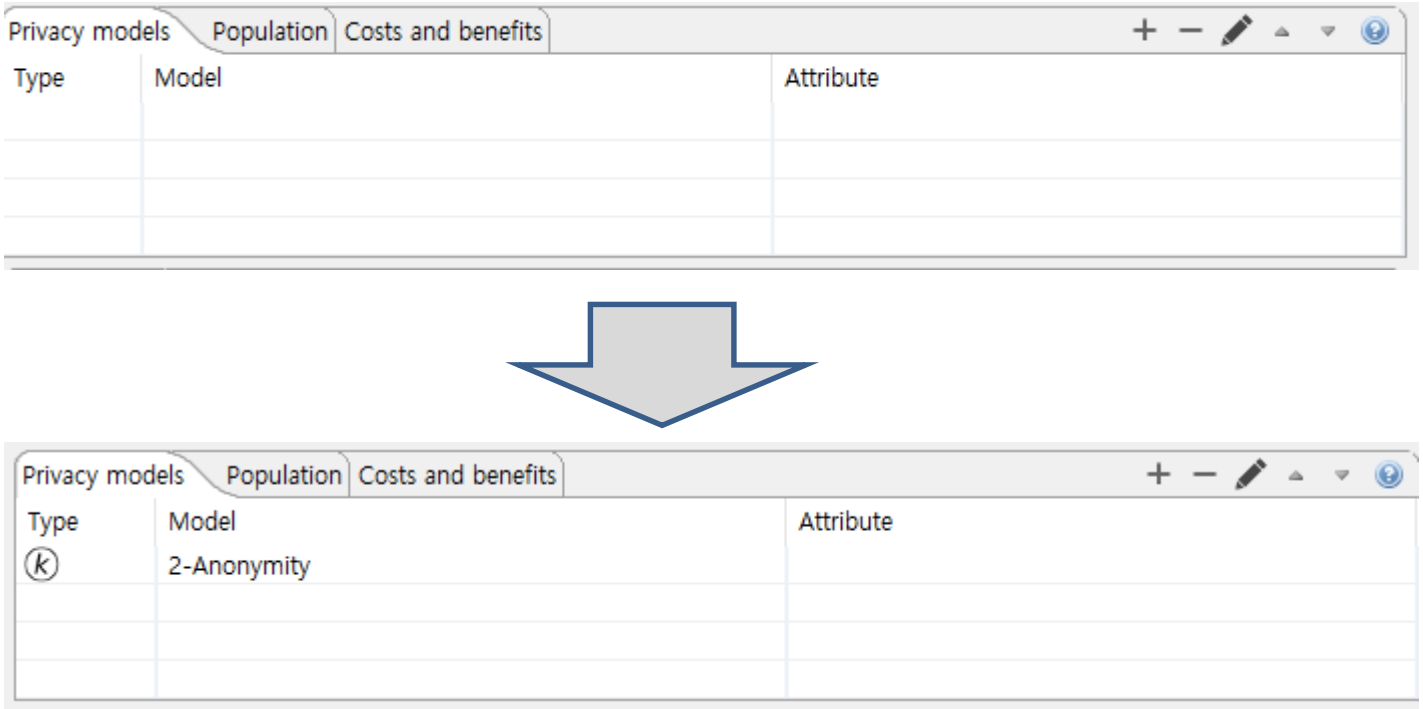
ARX-Anonymize

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 비 식별화 모형을 만족하는 All possible level 조합의 모형이 explore 화면에 나타난다.

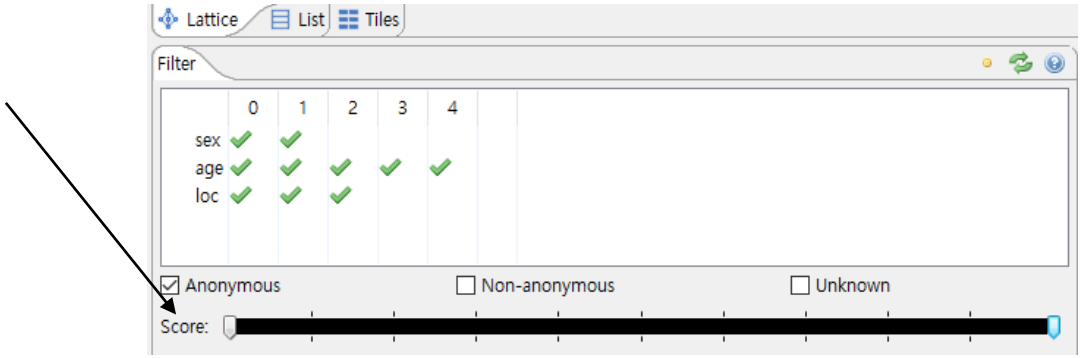
아래 그림에서 밑부분 {0,1,0}이 가장 낮은 level의 조합이고 위의 {1,4,2}가 가장 높은 level의 조합이다.

{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 선택하는가이다.
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 큰 모형

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

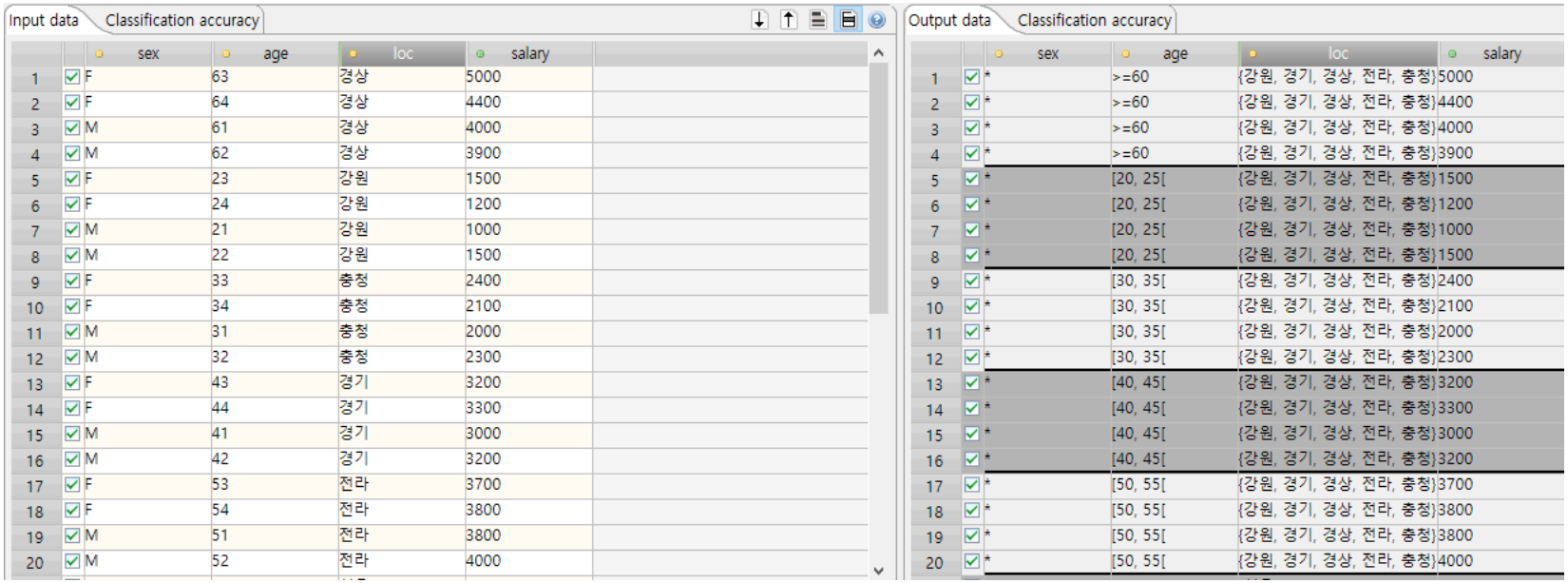
ARX-Transformation

{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.



Input data | Classification accuracy

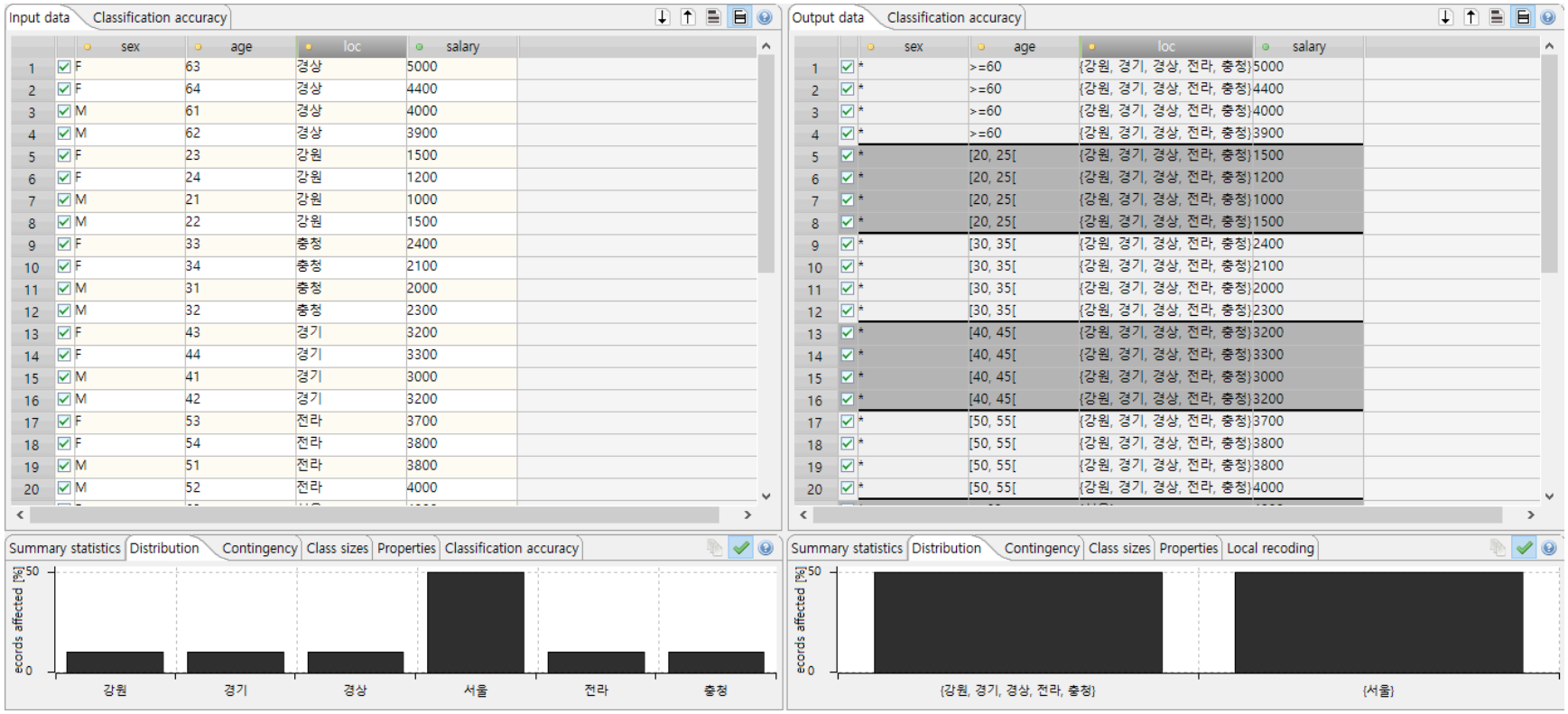
	sex	age	loc	salary
1	F	63	경상	5000
2	F	64	경상	4400
3	M	61	경상	4000
4	M	62	경상	3900
5	F	23	강원	1500
6	F	24	강원	1200
7	M	21	강원	1000
8	M	22	강원	1500
9	F	33	충청	2400
10	F	34	충청	2100
11	M	31	충청	2000
12	M	32	충청	2300
13	F	43	경기	3200
14	F	44	경기	3300
15	M	41	경기	3000
16	M	42	경기	3200
17	F	53	전라	3700
18	F	54	전라	3800
19	M	51	전라	3800
20	M	52	전라	4000

Output data | Classification accuracy

	sex	age	loc	salary
1	*	>=60	{강원, 경기, 경상, 전라, 충청}	5000
2	*	>=60	{강원, 경기, 경상, 전라, 충청}	4400
3	*	>=60	{강원, 경기, 경상, 전라, 충청}	4000
4	*	>=60	{강원, 경기, 경상, 전라, 충청}	3900
5	*	[20, 25[	{강원, 경기, 경상, 전라, 충청}	1500
6	*	[20, 25[	{강원, 경기, 경상, 전라, 충청}	1200
7	*	[20, 25[	{강원, 경기, 경상, 전라, 충청}	1000
8	*	[20, 25[	{강원, 경기, 경상, 전라, 충청}	1500
9	*	[30, 35[	{강원, 경기, 경상, 전라, 충청}	2400
10	*	[30, 35[	{강원, 경기, 경상, 전라, 충청}	2100
11	*	[30, 35[	{강원, 경기, 경상, 전라, 충청}	2000
12	*	[30, 35[	{강원, 경기, 경상, 전라, 충청}	2300
13	*	[40, 45[	{강원, 경기, 경상, 전라, 충청}	3200
14	*	[40, 45[	{강원, 경기, 경상, 전라, 충청}	3300
15	*	[40, 45[	{강원, 경기, 경상, 전라, 충청}	3000
16	*	[40, 45[	{강원, 경기, 경상, 전라, 충청}	3200
17	*	[50, 55[	{강원, 경기, 경상, 전라, 충청}	3700
18	*	[50, 55[	{강원, 경기, 경상, 전라, 충청}	3800
19	*	[50, 55[	{강원, 경기, 경상, 전라, 충청}	3800
20	*	[50, 55[	{강원, 경기, 경상, 전라, 충청}	4000

ARX-Transformation

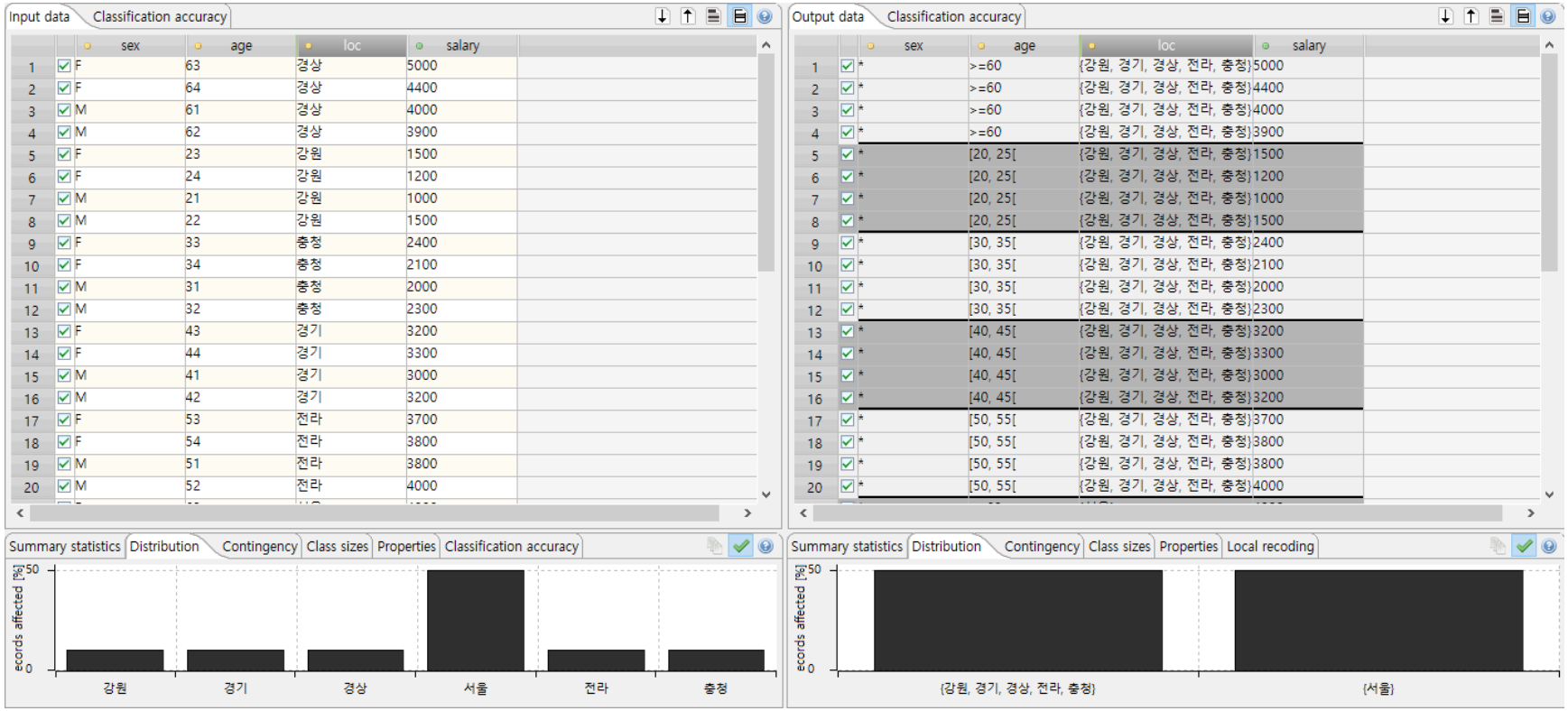
이 자료에서 salary를 sensitive로 바꾸고 2-다양성을 추가해도 결과는 달라지지 않는다. 아래의 그림에서 알 수 있듯이 모든 준 식별자 조합에서 2개 이상의 다양한 salary 값이 존재한다.



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.

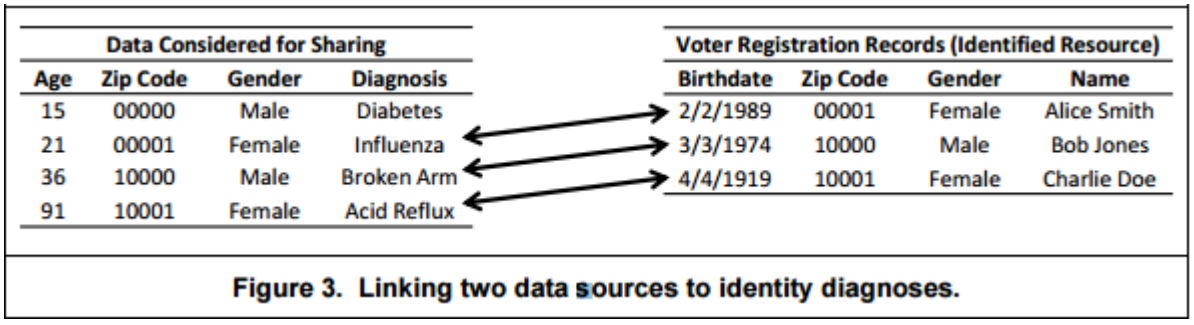
1) Prosecutor Model: 이는 특정 개인이 이 Dataset에 포함되어 있음을 알고 있다는 가정하에 그 개인을 찾을 수 있는 가능성을 분석하는 것이다.

2) Journalist model은 아래의 그림처럼 다른 데이터 베이스와 결합하여 개인 식별 가능성을 분석하는 것이다.

3) 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			
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

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 | »2

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 | »1

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

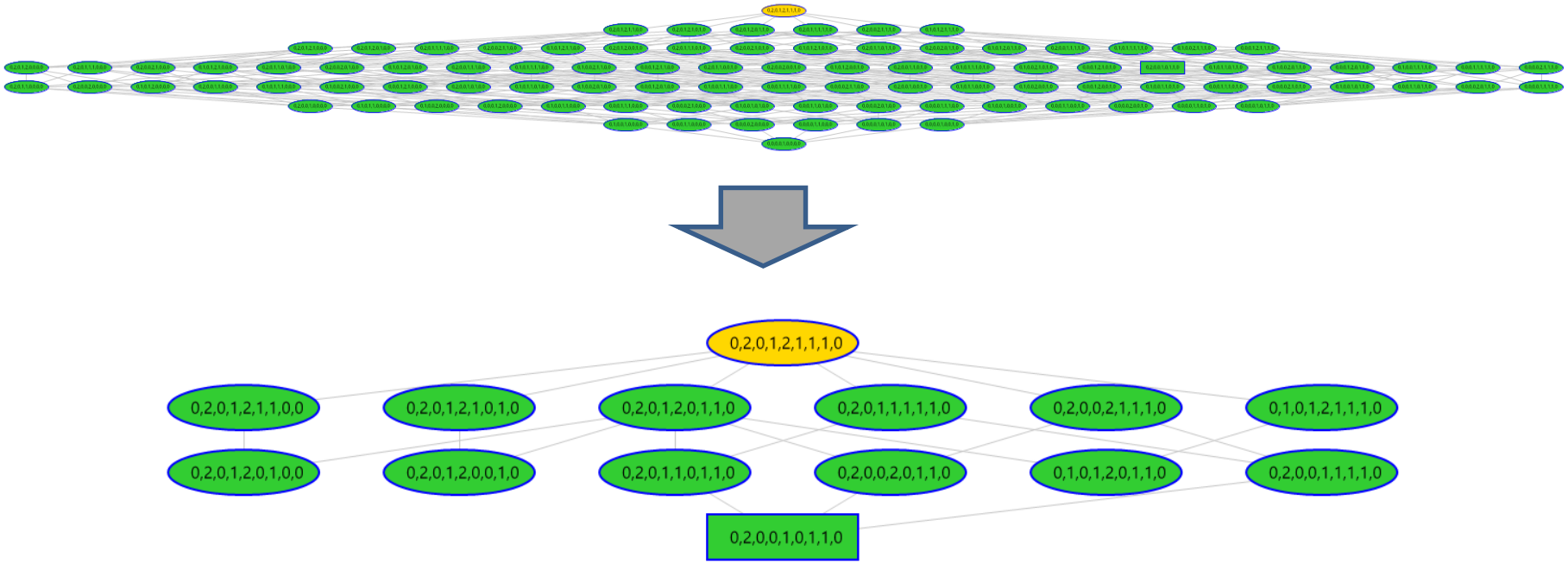
ARX-Example

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

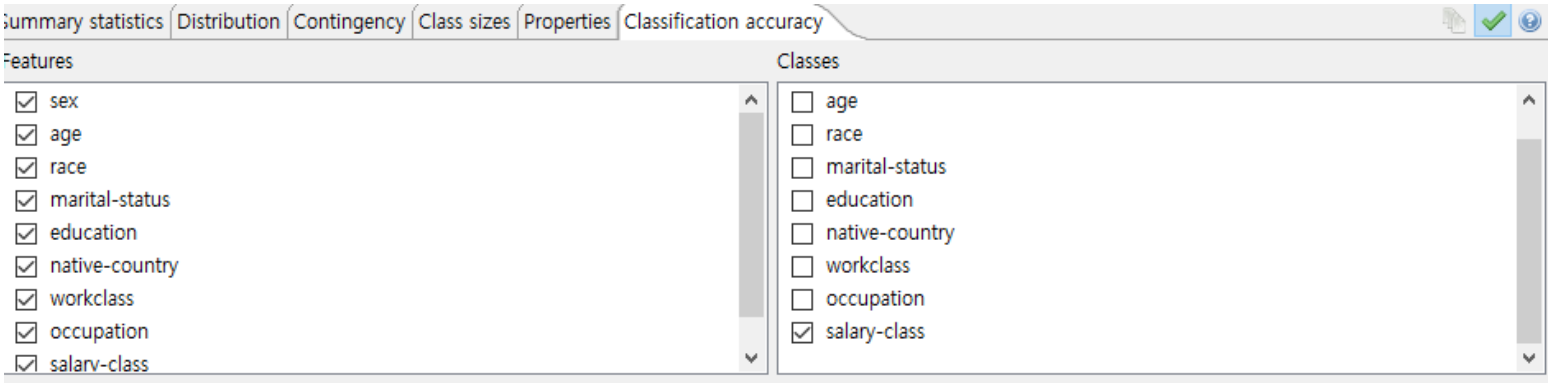
{0,2,0,1,2,1,1,1,0}는 {0,2,0,0,1,0,1,1,0}의 모형에서 marital status, education, native-country를 한 수준씩 더 가공한 것이다. 가공을 더 하면 Information Loss가 커지는데 얼마나 더 커질까 가늠하기는 어렵다.

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 Data의 정확도를 비교하여 손실이 작은 모형을 선택할 수 있도록 도와주고 있다.

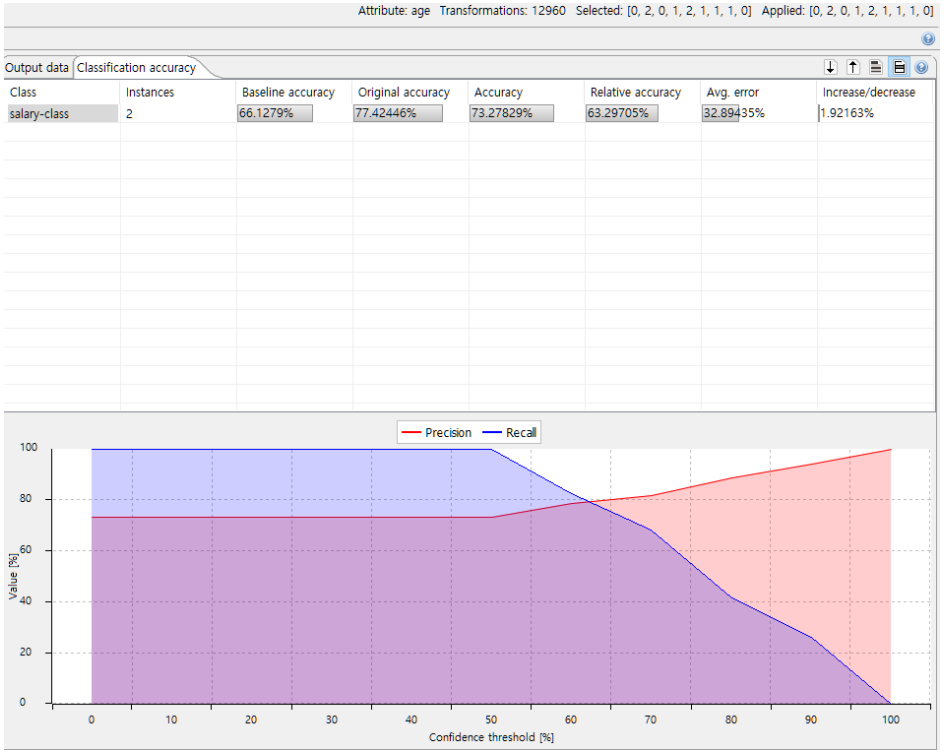
아래의 그림은 Feature 즉, 독립변수를 아래와 같이 선택하고 종속변수를 salary-class로 지정하여 Logistic Regression을 통해 예측 정확도를 검사하는 화면이다.



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	
4947	Male	61	White	Married-civ-spouse	
4948	Male	61	White	Married-civ-spouse	
4949	Male	61	White	Married-civ-spouse	
4950	Female	74	White	Divorced	
4951	Female	71	White	Never-married	
4952	Female	75	White	Widowed	
4953	Female	73	White	Married-civ-spouse	

4946	Male	[61, 70]	White	Spouse present	Se
4947	Male	[61, 70]	White	Spouse present	Se
4948	Male	[61, 70]	White	Spouse present	Se
4949	Male	[61, 70]	White	Spouse present	Se
4950	*	*	*	*	*
4951	*	*	*	*	*
4952	*	*	*	*	*
4953	*	*	*	*	*

ARX-Example

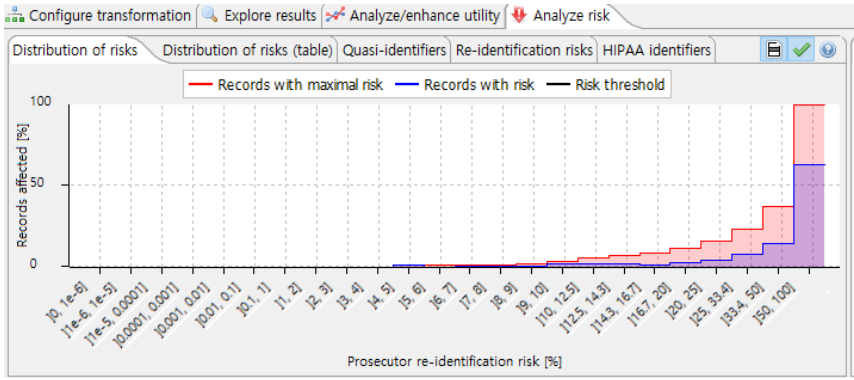
아래의 그림은 Prosecutor Re-identification 위험의 크기에 따른 레코드 비율을 그래프로 보여주고 있다.

잘 된 비 식별화는 우측처럼 파란색이 그래프 좌측에 많이 있어야 한다.

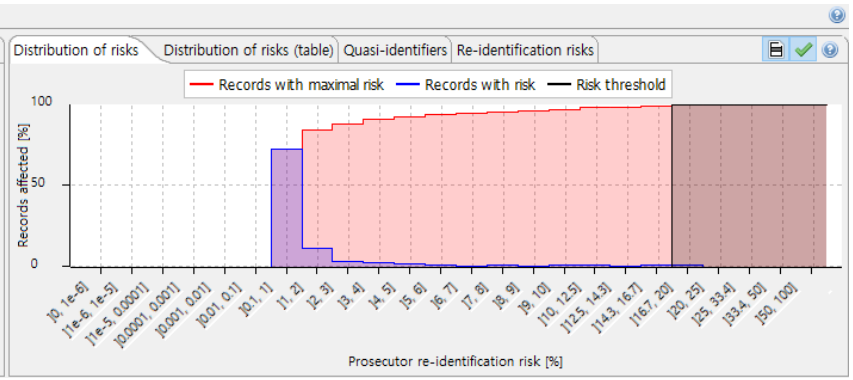
변환 전 Data는 파란색 그래프 즉 Records with risk가 그래프 우측에 많이 분포되어 있는 반면

변환 후 Data는 파란색 그래프가 좌측에 분포되어 이 변환을 통해 재 식별 가능성이 많이 낮아졌음을 알 수 있다.

변환 전 Data



변환 후 Data



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

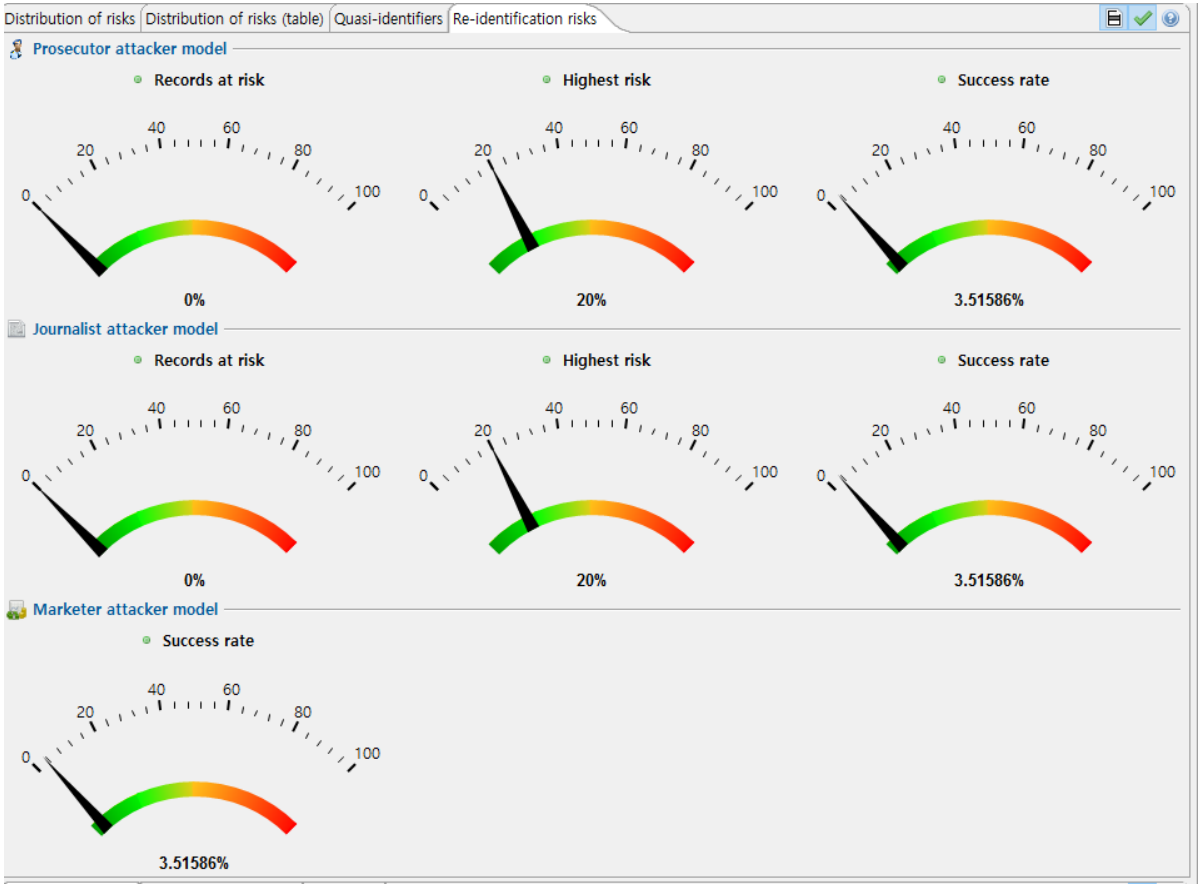
Quasi-identifier	Distinction	Separation
sex, age, race, marital-status, native-country, occupation, salary-class	31.07553%	99.954%
sex, age, race, native-country, workclass, occupation, salary-class	31.08879%	99.95183%
age, race, marital-status, education, native-country, workclass, salary-class	35.57125%	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

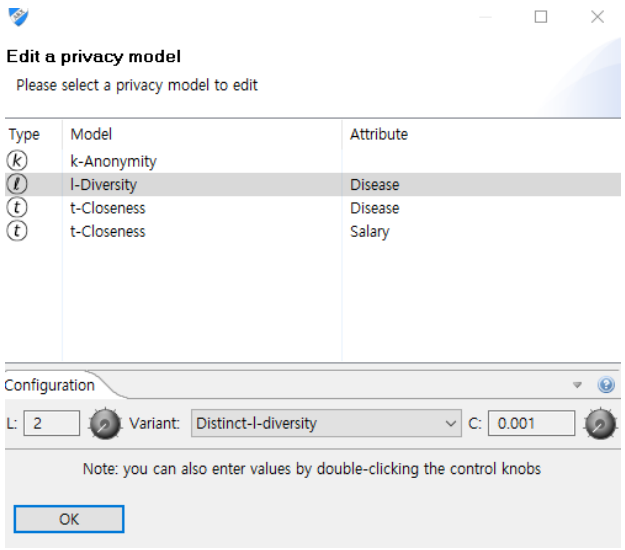
Level-0	Level-1	Level-2
l궤양	{1.위궤양, 1.위암, 1.위염}	*
l암	{1.위궤양, 1.위암, 1.위염}	*
l염	{1.위궤양, 1.위암, 1.위염}	*
감기	{2.감기, 2.기관지염, 2.폐렴}	*
l관지염	{2.감기, 2.기관지염, 2.폐렴}	*
l렴	{2.감기, 2.기관지염, 2.폐렴}	*

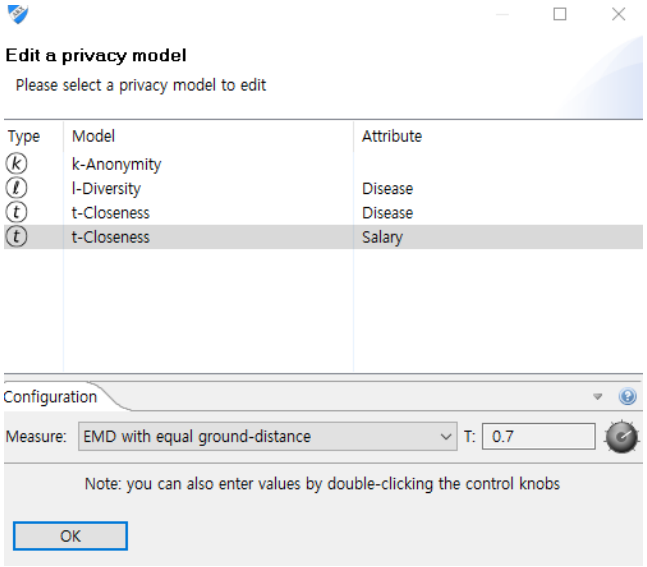
ARX-Sensitive Data

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.





통신자료 Review

자료는 1만개의 레코드로 이루어진 개인고객의 통신관련정보이다. (http://naver.me/xLqvMoHM)

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?

감 사 합 니 다

Q & A