숭실대학교

인공지능특론 1 보고서

Assignment #1 – Anonymization

1102040007 오혁진

1102340003 유자연

1102378004 서영재

1102378010 강민정

목차

1 서론 ..... - 1 -

2 익명화 ..... - 3 -

2.1 데이터 선정 ..... - 3 -

2.2 익명화 방법 ..... - 4 -

3 익명화 결과 ..... - 8 -

3.1 익명화 데이터 ..... - 8 -

3.2 익명화 검증 ..... - 11 -

4 결론 ..... - 12 -

1 서론

- 개념

K-anonymity : 데이터 세트의 각 그룹(또는 카테고리)이 동일한 속성 세트를 가진 k 개 이상의 개인을 갖도록 요구하는 프라이버시 개념

L-diversity : 데이터 세트의 각 그룹(또는 카테고리)이 민감한 속성에 대해 최소한 l 개의 고유한 값을 가져야 하는 개인 정보 보호

T-closeness : 그룹(또는 카테고리)의 민감한 속성 분포와 데이터 세트의 민감한 속성의 전체 분포 사이의 거리를 측정하는 프라이버시 개념

예시)

K-anonymity :

나이	성별	건강 상태
32	남	당뇨
32	남	암
24	여	암
24	여	당뇨

⇨ 남성 환자 모두와 여성 환자 모두는 암과 당뇨를 앓고 있음

⇨ 이 데이터 세트는 연령, 성별 및 질병 속성과 관련하여 $k = 2$ 를 충족

L-diversity :

거래 금액(만 원)	위치	카드 소유자
50	서울	김철수
50	부산	이영희
75	대전	김철수
75	대구	이영희

⇨ 50만 원, 75만 원 거래는 각 서로 다른 개인이 거래함

⇨ 이 데이터 세트는 거래 금액 속성에 대해 $l = 2$ 를 만족

T-closeness :

나이	성별	등급
22	남	4
24	여	2
38	남	5
40	여	3

나이	성별	등급
20-30	남	4
20-30	여	2
30-40	남	5
30-40	여	3

⇨ 각 그룹의 등급 데이터 세트의 전체 등급 분포와 크게 다르지 않도록 등급을 그룹화해야함

⇨ 위 예에서 각 그룹의 등급 분포는 크게 다르지 않음

2 익명화

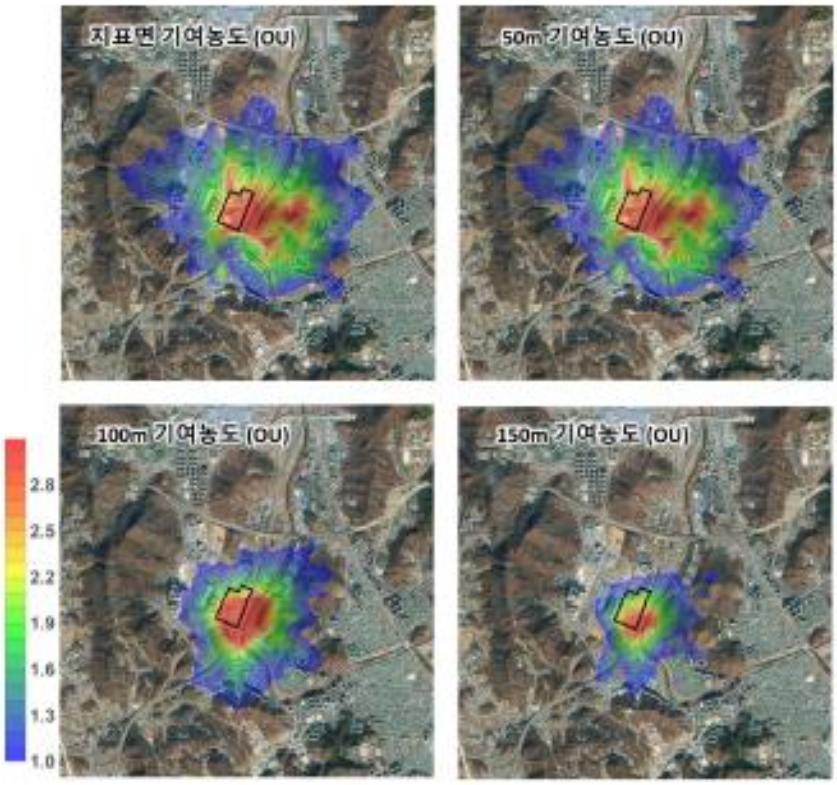
2.1 데이터 선정

데이터 : 종관기상관측(23년 2월 1일 24시간 분당 데이터) 중 7 개 지역 임의 선택

내용 : 측정 지점, 일시, 기온(℃), 풍향(deg), 풍속(m/s), 현지기압(hPa), 습도(%) 등의 데이터

데이터 수 : 약 1 만 개

데이터 사용 예시 : 시뮬레이션 프로그램을 이용한 특정 지역의 시간에 따른 대기오염물질 확산 추이 예상



[출처] 2023 악취 및 대기확산 모델링(CALPUFF) 교육, 안양대학교 기후·에너지·환경융합연구소, 한국냄새환경학회

커스터마이징 : 해당 데이터에 특정 민감 정보가 존재하지 않아 임의로 설정
(QI : “지점”, “일시”, “기온(℃)” / SA : “풍향(deg)”)

※ k, l, t 값에 임의의 값을 넣기 위해 칼럼별로 일 또는 십의 자리에서 반올림하여 데이터 수정

2.2 익명화 방법

익명화 값 :

k-anonymity ➔ k = 5

l-diversity ➔ l = 3

t-closeness ➔ t = 0.2

익명화 파이썬 코드(data 저장 코드는 보고서에서는 생략)

- **사전처리**

```
# pip install openpyxl
# pip install faker
# pip install pycanon
# pip install matplotlib
#!pip install pycanon

import pandas as pd
from pycanon import anonymity, report

df = pd.read_csv( "./original_data(2).csv")
df.head()
```

- **k-Anonymity**

```
import numpy as np

def generalize(data, columns, generalization_levels):
    generalized_data = data.copy()
    for col, level in zip(columns, generalization_levels):
        if level > 0 and np.issubdtype(data[col].dtype, np.number):
            col_min = data[col].min()
            col_max = data[col].max()
            step = (col_max - col_min) // level
            generalized_data[col] = (data[col] // step) * step
    return generalized_data


def check_k_anonymity(df, columns, k):
    grouped_data = df.groupby(columns).size().reset_index(name='count')
    return all(grouped_data['count'] >= k)

def k_anonymize(data, columns, k, generalization_levels):
    generalized_data = generalize(data, columns, generalization_levels)
    if check_k_anonymity(generalized_data, columns, k):
        return generalized_data
    else:
        print("The dataset doesn't satisfy k-anonymity.")
        return None

# Load the CSV file
file_path = "./original_data(2).csv"
data = pd.read_csv(file_path)

# Define the columns to anonymize and their generalization levels
# Here, we assume that you have 3 columns to anonymize, and the generaliz
ation levels are [2, 2, 2]
columns_to_anonymize = ["현지기압(hPa)"  ,"해면기압(hPa)"]
generalization_levels = [5]

# Apply k-anonymity
k = 5
anonymized_data = k_anonymize(data, columns_to_anonymize, k, generalizati
on_levels)

if anonymized_data is not None:
    print("Anonymized data:")
    print(anonymized_data)
```

- **l-diversity**

```
def check_l_diversity(df, sensitive_column, columns, l):
    grouped_data = df.groupby(columns)[sensitive_column].nunique().re
set_index(name='unique_values')
    return all(grouped_data['unique_values'] >= l)

def l_diversify(data, sensitive_column, columns, l):
    if check_l_diversity(data, sensitive_column, columns, l):
        return data
    else:
        print("The dataset doesn't satisfy l-diversity.")
        return None

# Load the CSV file
file_path_2="./Anonymized_data_output1.csv"
data_2 = pd.read_csv(file_path_2)

# Define the sensitive column and the columns to group by (e.g., anon
ymized columns)
sensitive_column = '지점'
columns_to_group_by = ['풍향(deg)']

# Apply l-diversity
l = 3
diversified_data = l_diversify(data_2, sensitive_column, columns_to_g
roup_by, l)

if diversified_data is not None:
    print("Diversified data:")
    print(diversified_data)
```

- **t-closeness**

```
from scipy.stats import wasserstein_distance

def get_distribution(df, sensitive_column):
    return df[sensitive_column].value_counts(normalize=True)

def apply_t_closeness(df, sensitive_column, columns, t):
    overall_dist = get_distribution(df, sensitive_column)
    grouped_data = df.groupby(columns)
    result_df = pd.DataFrame()

    for _, group in grouped_data:
        group_dist = get_distribution(group, sensitive_column)
        emd = wasserstein_distance(overall_dist, group_dist)

        if emd <= t:
            result_df = result_df.append(group)

    if not result_df.empty:
        return result_df.reset_index(drop=True)
    else:
        print("No groups satisfy t-closeness.")
        return None

# Load the CSV file
file_path = "./original_data(2).csv"
data = pd.read_csv(file_path)

# Define the sensitive column and the columns to group by (e.g., anonymized columns)
sensitive_column = '지점'
columns_to_group_by = ['풍향(deg)','풍속(m/s)']

# Apply k-anonymity and l-diversity (or any other anonymization techniques)
# ...

# Apply t-closeness
t = 0.2
tclosed_data = apply_t_closeness(data, sensitive_column, columns_to_group_by, t)

if tclosed_data is not None:
    # Print the DataFrame
    print("Data after applying t-closeness:")
    print(tclosed_data)
```

3 익명화 결과

3.1 익명화 데이터

- K-anonymity 데이터 비교

지점	풍향(deg)	풍속(m/s)	현지기압(hPa)	해면기압(hPa)	습도(%)	일사(MJ/m^2)
133	260	2	1007	1016	66	0
133	270	2	1007	1016	67	0
133	270	2	1007	1016	68	0
133	270	1	1008	1016	67	0
133	270	1	1008	1016	67	0
133	280	2	1008	1016	68	0
133	260	2	1008	1016	67	0
133	240	2	1008	1016	67	0
133	250	1	1008	1016	67	0
133	260	1	1007	1016	68	0
133	260	1	1007	1016	68	0
133	300	1	1007	1016	68	0
133	280	2	1007	1016	68	0
133	310	1	1007	1016	68	0
133	270	2	1007	1016	68	0
133	270	2	1007	1016	68	0
133	310	1	1007	1016	67	0
133	290	2	1007	1016	67	0
133	280	2	1007	1016	67	0
133	270	2	1007	1016	67	0
133	280	1	1007	1016	67	0
133	280	1	1007	1016	67	0
133	290	1	1007	1016	67	0
133	280	2	1007	1016	67	0
133	270	2	1007	1016	67	0
133	250	2	1007	1016	67	0
133	310	1	1007	1016	66	0
133	290	1	1007	1016	67	0
133	250	1	1007	1016	67	0

지점	풍향(deg)	풍속(m/s)	현지기압(hPa)	해면기압(hPa)	습도(%)	일사(MJ/m^2)
133	260	2	1004	1016	66	0
133	270	2	1004	1016	67	0
133	270	2	1004	1016	68	0
133	270	1	1008	1016	67	0
133	270	1	1008	1016	67	0
133	280	2	1008	1016	68	0
133	260	2	1008	1016	67	0
133	240	2	1008	1016	67	0
133	250	1	1008	1016	67	0
133	260	1	1004	1016	68	0
133	260	1	1004	1016	68	0
133	300	1	1004	1016	68	0
133	280	2	1004	1016	68	0
133	310	1	1004	1016	68	0
133	270	2	1004	1016	68	0
133	270	2	1004	1016	68	0
133	310	1	1004	1016	67	0
133	290	2	1004	1016	67	0
133	280	2	1004	1016	67	0
133	270	2	1004	1016	67	0
133	280	1	1004	1016	67	0
133	280	1	1004	1016	67	0
133	290	1	1004	1016	67	0
133	280	2	1004	1016	67	0
133	270	2	1004	1016	67	0
133	250	2	1004	1016	67	0
133	310	1	1004	1016	66	0
133	290	1	1004	1016	67	0
133	250	1	1004	1016	67	0

- K-anonymity + L-diversity

지점	풍향(deg)	풍속(m/s)	현지기압(hPa)	해면기압(hPa)	습도(%)	일사(MJ/m^2)
133	260	2	1007	1016	66	0
133	270	2	1007	1016	67	0
133	270	2	1007	1016	68	0
133	270	1	1008	1016	67	0
133	270	1	1008	1016	67	0
133	280	2	1008	1016	68	0
133	260	2	1008	1016	67	0
133	240	2	1008	1016	67	0
133	250	1	1008	1016	67	0
133	260	1	1007	1016	68	0
133	260	1	1007	1016	68	0
133	300	1	1007	1016	68	0
133	280	2	1007	1016	68	0
133	310	1	1007	1016	68	0
133	270	2	1007	1016	68	0
133	270	2	1007	1016	68	0
133	310	1	1007	1016	67	0
133	290	2	1007	1016	67	0
133	280	2	1007	1016	67	0
133	270	2	1007	1016	67	0
133	280	1	1007	1016	67	0
133	280	1	1007	1016	67	0
133	290	1	1007	1016	67	0
133	280	2	1007	1016	67	0
133	270	2	1007	1016	67	0
133	250	2	1007	1016	67	0
133	310	1	1007	1016	66	0
133	290	1	1007	1016	67	0
133	250	1	1007	1016	67	0

지점	풍향(deg)	풍속(m/s)	현지기압(hPa)	해면기압(hPa)	습도(%)	일사(MJ/m^2)
133	260	2	1004	1016	66	0
133	270	2	1004	1016	67	0
133	270	2	1004	1016	68	0
133	270	1	1008	1016	67	0
133	270	1	1008	1016	67	0
133	280	2	1008	1016	68	0
133	260	2	1008	1016	67	0
133	240	2	1008	1016	67	0
133	250	1	1008	1016	67	0
133	260	1	1004	1016	68	0
133	260	1	1004	1016	68	0
133	300	1	1004	1016	68	0
133	280	2	1004	1016	68	0
133	310	1	1004	1016	68	0
133	270	2	1004	1016	68	0
133	270	2	1004	1016	68	0
133	310	1	1004	1016	67	0
133	290	2	1004	1016	67	0
133	280	2	1004	1016	67	0
133	270	2	1004	1016	67	0
133	280	1	1004	1016	67	0
133	280	1	1004	1016	67	0
133	290	1	1004	1016	67	0
133	280	2	1004	1016	67	0
133	270	2	1004	1016	67	0
133	250	2	1004	1016	67	0
133	310	1	1004	1016	66	0
133	290	1	1004	1016	67	0
133	250	1	1004	1016	67	0

- T-closeness

지점	풍향(deg)	풍속(m/s)	현지기압(hPa)	해면기압(hPa)	습도(%)	일사(MJ/m^2)
133	260	2	1007	1016	66	0
133	270	2	1007	1016	67	0
133	270	2	1007	1016	68	0
133	270	1	1008	1016	67	0
133	270	1	1008	1016	67	0
133	280	2	1008	1016	68	0
133	260	2	1008	1016	67	0
133	240	2	1008	1016	67	0
133	250	1	1008	1016	67	0
133	260	1	1007	1016	68	0
133	260	1	1007	1016	68	0
133	300	1	1007	1016	68	0
133	280	2	1007	1016	68	0
133	310	1	1007	1016	68	0
133	270	2	1007	1016	68	0
133	270	2	1007	1016	68	0
133	310	1	1007	1016	67	0
133	290	2	1007	1016	67	0
133	280	2	1007	1016	67	0
133	270	2	1007	1016	67	0
133	280	1	1007	1016	67	0
133	280	1	1007	1016	67	0
133	290	1	1007	1016	67	0
133	280	2	1007	1016	67	0
133	270	2	1007	1016	67	0
133	250	2	1007	1016	67	0
133	310	1	1007	1016	66	0
133	290	1	1007	1016	67	0
133	250	1	1007	1016	67	0

지점	풍향(deg)	풍속(m/s)	현지기압(hPa)	해면기압(hPa)	습도(%)	일사(MJ/m^2)
133	0	1	1007	1016	71	0
133	0	1	1016	1025	53	13
133	0	1	1017	1026	54	13
90	0	1	1013	1015	42	0
90	0	1	1013	1015	43	0
90	0	1	1013	1015	42	0
102	0	1	1010	1015	96	0
106	0	1	1007	1012	50	0
106	0	1	1007	1012	52	0
106	0	1	1009	1014	38	0
106	0	1	1009	1014	38	0
106	0	1	1009	1014	38	0
106	0	1	1010	1014	36	0
106	0	1	1010	1015	36	0
106	0	1	1010	1015	37	0
106	0	1	1010	1015	37	0
106	0	1	1010	1015	37	0
152	0	1	1005	1015	69	0
152	0	1	1004	1014	65	0
133	0	2	1017	1026	55	13
90	0	2	1013	1015	36	0
159	0	2	1011	1020	28	13
184	0	2	1022	1025	63	11
184	0	2	1022	1024	67	11
102	0	2	1011	1016	97	0
102	0	2	1011	1016	97	0
102	0	2	1011	1016	98	0
102	0	2	1011	1016	98	0
102	0	2	1011	1016	98	0

3.2 익명화 검증

- K-anonymity + L-diversity

```
The dataset verifies:
                - k-anonymity with k = 21
                - (alpha,k)-anonymity with alpha = 0.9210526315789473 and k = 21
                - l-diversity with l = 3
                - entropy l-diversity with l = 1
                - (c,l)-diversity with c = 1 and l = 3
                - basic beta-likeness with beta = 5.447368421052632
                - enhanced beta-likeness with beta = 1.9459101490553135
                - t-closeness with t = 0.38411458333333326
                - delta-disclosure privacy with delta = 3.978345648359219
```

- T-closeness

```
The dataset verifies:
                - k-anonymity with k = 3
                - (alpha,k)-anonymity with alpha = 0.7560975609756098 and k = 3
                - l-diversity with l = 3
                - entropy l-diversity with l = 2
                - (c,l)-diversity with c = 2 and l = 3
                - basic beta-likeness with beta = 10.553884711779448
                - enhanced beta-likeness with beta = 2.852496827208995
                - t-closeness with t = 0.39598278094469386
                - delta-disclosure privacy with delta = 3.4527268496528607
```

4 결론

k-anonymity, l-diversity, t-closeness 의 서로 다른 k, l , t 값을 이용하여 익명화된 데이터 생성

- k, t 에 대한 값들은 정상적으로 적용하여 진행
- 처음에는 l 값은 임의로 다른 값을 넣는 것에 어려움이 있었음

  ⇨ 기존의 데이터를 그대로 쓰는 것이 아닌 불필요한 칼럼 제거 및 데이터 반올림(일의 자리 또는 십의 자리에서)하여 데이터 커스터마이징 후 적용

pyCanon을 사용하여 익명성 검증

- pyCanon을 사용하여 진행
- k, t, l에 대한 값들은 정상적으로 report가 생성됨을 확인

  ⇨ 모든 값들이 각 값들의 정확한 값이 report 된 것은 아님

  ⇨ 정확한 값 또는 근사치의 값이 report 되는 것을 확인