Bases de Données Avancées
Extensions nécessaires au k-anonymat

Jean-François COUCHOT
couchot[arobase]femto-st[point]fr

16 mars 2021

1 Introduction

L'objectif du TP est de montrer que le k-anonymat doit être manipulé avec précaution pour protéger la vie privée de personnes ayant consenti à partager leurs données. On considère le jeu de données accessible

`https://www.kaggle.com/kingabzpro/heart-disease-from-cleveland`

et téléchargeable sur le site du cours où les attributs ont été francisés.

Voici une description des attributs non évidents :

- *genre* : 1 pour homme, 0 pour femme ;
- *doulthor* : type de douleur thoracique ressentie tel que : 1 = angine typique, 2 = angine atypique, 3 = douleur non angineuse, 4 = asymptotique ;
- *tensrepos* : valeur de la tension artérielle au repos en mmHg ;
- *cholesterol* : cholestérol sérique en mg/dl ;
- *gleucajeun* : 1 si la glycémie à jeun supérieure à avec 120mg/dl, 0 sinon ;
- *ecgrepos* : électrocardiogramme de repos tel que : 0 = normal, 1 = anomalie de l'onde ST-T, 2 = hypertrophie ventriculaire gauche ;
- *freqcardmax* : fréquence cardiaque maximale atteinte ;
- *angpoitrex* : 1 si angine de poitrine induite par l'exercice, 0 sinon ;
- *depexercice* : dépression ST induite par l'exercice par rapport au repos
- *penteexercice* : pente ST du pic d'exercice telle que : 1 = en pente ascendante, 2 = plat, 3 = en pente descendante ;
- *nbvaisscoloror* : nombre de vaisseaux majeurs (0-3) colorés par la flourosopie ;
- *thalassemie* : 3 = normal , 6 = défaut fixe, 7 = défaut réversible ;
- *classmaladiecard* : indique si l'individu souffre d'une maladie cardiaque ou non : 0 = absence, 3,4 = élevé.

On considère qu'une personne risque d'avoir une attaque cardiaque si elle est dans la classe 3 ou dans la classe 4.

Le service médical voudrait connaître *"quel est l'âge moyen des patients en classe 3 ou 4 ?"*, mais ne sait pas utiliser un tableur. Par contre des membres du service savent anonymiser les données.

Vous allez tout d'abord anonymiser ces données et exploiter ensuite ces données nettoyées pour répondre à cette question.

2 Mise en place du k-anonymat

Exercice 2.1 (Atributs : quasi-identifiants, sensibles).

1. *Quels attributs personnels :*
   (a) *pourraient être considérés comme quasi-identifiants ?*
   (b) *vous paraissent sensibles ?*
2. *Dans ARX, gérer les quasi-identifiants et leurs hiérarchies de généralisation. On sera le plus précis possible.*

Exercice 2.2 (6-anonymité).

1. *Demander à l'outil ARX de générer un jeu de données* 6*-anonyme en autorisant 5% de suppression.*
2. *Interpréter le score obtenu pour la généralisation avec la valeur de Loss la plus petite (qui devrait correspondre à la généralisation (3,0).*

3. *Ce jeu de données* 6*-anonyme contient les lignes reproduites ci-dessous :*

age	*genre*	*doul thor*	*tens repos*	*choles terol*	*gleuc ajeun*	*ecg repos*	*freq cardmax*	*angpoitr ex*	*dep exercice*	*pente exercice*	*nbvaiss coloror*	*thalassemie thalassemie*	*classmaladie card*
$[69, 77[$	1	4	145	174	0	0	125	1	2.6	3	0	7	4
$[69, 77[$	1	4	130	322	0	2	109	0	2.4	2	3	3	1
$[69, 77[$	1	3	160	269	0	0	112	1	2.9	2	1	7	3
$[69, 77[$	1	3	140	254	0	2	146	0	2.0	2	3	7	2
$[69, 77[$	1	1	160	234	1	2	131	0	0.1	2	1	3	0
$[69, 77[$	1	2	156	245	0	2	143	0	0.0	1	0	3	0

Vous connaissez un homme de 70 ans souffrant d'angine de poitrine due à l'exercice (attribut angpoitrex*). Montrer que la publication de cette base de données permet d'acquérir d'autres informations critiques sur ce patient, notamment qu'il souffre de problèmes cardiaques.*

3 Répondre au problème d'homogénéité

Exercice 3.1 (Une mauvaise réponse : incrémenter *k* sans réfléchir). *On décide d'incrémenter la valeur de* k *jusqu'à ne plus avoir ce problème d'homogénéité pour l'attribut* classmaladiecard*.*

1. *Montrer que dès que* $k = 7$*, la généralisation qui minimise le Loss est (4,0) qui aboutit à du 15-anonymat.*
2. *Quelle intuition avez-vous quant à l'utilité de telles données anonymisées ?*
3. *Même question avec* k*=8,...*

Exercice 3.2 (*l* diversité). *Ce que l'on cherche, c'est avoir au moins 3 valeurs distinces pour l'attribut* classmaladiecard *dans chaque regroupement.*

1. *Spécifier dans ARX que l'attribut* classmaladiecard *est sensible.*
2. *Ajouter le modèle de* 3*-diversité pour cet attribut.*
3. *Préciser que vous souhaitez du* 6*-anonymat.*
4. *Demander à ARX de tester les généralisations permettant de garantir ces deux propriétés et constater à nouveau que l'on a le choix entre la généralisation choisie par défaut est (5,0).*
5. *Quelle garantie sur le modèle généré a-t-on en plus dans cet exercice, par rapport au précédent.*
6. *Par rapport à la question initiale décrite dans l'introduction, pourquoi est-il finalement judicieux de choisir la généralisation (3,1) ? Appliquer la transformation pour ce niveau de généralisation.*
7. *Exporter le jeu de données sous le nom* `Cleveland_3_1.csv`*.*

4 Répondre à la question initiale

Exercice 4.1 (Age moyen des patients en classe 3 ou 4). *A l'aide d'un tableur, montrer qu'il est possible d'estimer l'âge moyen des patients dans les classes 3 ou 4.*

1. *Le mettre en place pour les données originales et pour celles de la généralisation (3,1).*
2. *Conclure.*