Bases de Données Avancées
k-anonymat et extensions

Jean-François Couchot

Université de Franche-Comté, UFR-ST

Plan

Un premier modèle de PVP : le k-anonymat

Extensions du k-anonymat

Plan

Plan

Le k-anonymat[??] et les quasi-identifiants -1

Quasi-identifiants : QID

- QID, intuition [1] : "des éléments qui ne sont pas en eux-mêmes des identificateurs uniques, mais qui sont suffisamment bien corrélés avec une entité pour pouvoir être combinés avec d'autres quasi-identifiants afin de créer un identificateur unique"
- QID, définition [2] : Les attributs de $Q \subseteq \{A_1, \ldots, A_M\}$ sont quasi-identifiants de la relation T si la requête suivante retourne au moins un résultat

```
SELECT Q FROM T
GROUP BY Q
HAVING COUNT(*)=1
```

- (CP, genre, date de naissance) : triplets uniques dans 87% des cas $\rightsquigarrow$ quasi-identifiants

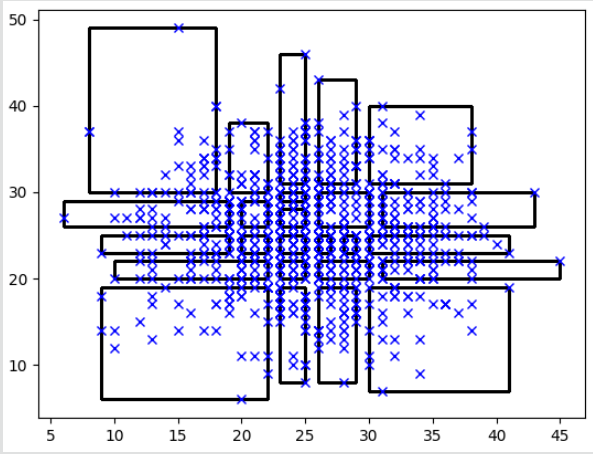
1. https://en.wikipedia.org/wiki/Quasi-identifier

2. Nguyen, B., & Castelluccia, C. (2020). Techniques d'anonymisation tabulaire : concepts et mise en oeuvre. arXiv preprint arXiv :2001.02650.

Le k-anonymat[??] et les quasi-identifiants -2

Intuition : regrouper les QID pour casser l'unicité

- Niveau de détail des valeurs des QID : à réduire pour qu'il y ait au moins k individus différents dont les QIDs sont égaux
- Individus avec mêmes QIDs : font partie de la même classe d'équivalence



Plan

Rappel de l'exemple pseudonymisé

	Non-sensibles				Sensibles
H	CP	Age	Genre	Nationalité	Pathologie
1	13053	28	H	russe	trouble cardiaque
2	13068	29	H	américaine	trouble cardiaque
3	13068	21	F	japonaise	infection virale
4	13053	23	H	américaine	infection virale
5	14853	49	H	indienne	cancer
6	14853	48	F	russe	trouble cardiaque
7	14850	47	H	américaine	infection virale
8	14850	49	F	américaine	infection virale
9	13053	31	H	américaine	cancer
10	13053	37	H	indienne	cancer
11	13068	36	F	japonaise	cancer
12	13068	35	F	américaine	cancer

k-anonymat par généralisation

Réduction des niveaux de détail

Sur l'exemple :

- CP : laisser, **regroupements par 2**, suppr.
- Age : laisser, par intervalles d'amplitudes 10, 20,**suppr.**
- Genre : laisser, **suppr.**
- Nationalité : laisser, par continent, **suppr.**
- $\rightsquigarrow 3 \times 4 \times 2 \times 3 = 72$ combinaisons de généralisation !

Regroupement par classes d'équivalence de card. $\geq$ à 4

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

4 individus

4 individus

4 individus

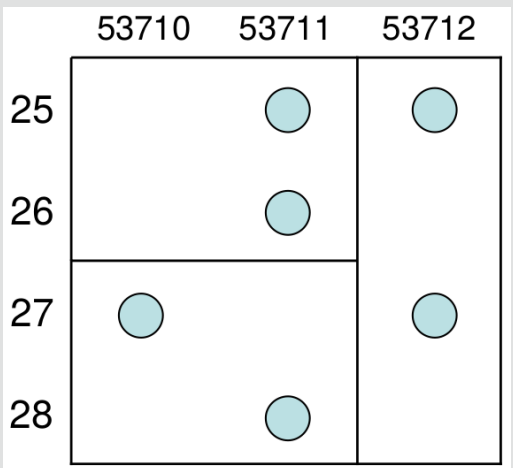
k-anonymat par Mondrian [3]

Algorithme général, en factorielle

1. Pour chaque séq. $[A_1, \ldots, A_f]$ de QID, telle qu'il reste plus de k éléments par groupe
   - partitionner intelligemment selon la médiane des données de A_i
2. généraliser les valeurs des attributs : un groupe $\equiv$ une généralisation
3. évaluer la perte d'information

2-anonymat, partitionnement selon [CP, Age, Genre]

Age	Genre	CP	Pathologie
25	H	53711	Grippe
25	F	53712	Hépatite
26	H	53711	Bronchite
27	H	53710	Bras cassé
27	F	53712	SIDA
28	H	53711	Ongle perdu



Age	Genre	CP	Pathologie
[25-26]	H	53711	Grippe
[25-27]	F	53712	Hépatite
[25-26]	H	53711	Bronchite
[27-28]	H	[53710-53711]	Bras cassé
[25-27]	F	53712	SIDA
[27-28]	H	[53710-53711]	Ongle perdu

3. LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006, April). Mondrian multidimensional k-anonymity. In 22nd International conference on data engineering (ICDE'06) (pp. 25-25). IEEE.

Plan

C_{AVG} : nbre. moy. d'élements par classe normalisé

Définition pour $|T|$ enregistrements et propriétés

$$C_{AVG} = \frac{|T|}{|EQs|} \times \frac{1}{k}$$

- et $|EQs|$: nbre. de classes d'équiv.
- $\frac{|T|}{|EQs|}$ nbre. moy. d'éléments par classe $(\geq k) \rightsquigarrow C_{AVG} \geq 1$
- Utilité décroissante à mesure que C_{AVG} croît.

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

- $C_{AVG} = \frac{12}{3} \times \frac{1}{4} = 1$
- Optimal pour cette métrique

Loss : Perte dûe à la généralisation

Définition pour $|T|$ enregistrements, n QID et propriétés

$$Loss = \frac{1}{n|T|} \sum_{j=1}^{|T|} \sum_{i=1}^{n} \frac{R_{i_j}}{R_i}$$

- R_{i_j}/R_i : rapport acquis/acquérables
  - discret : $(|\text{généralisation } i_j| - 1)/(|\text{Qid } i| - 1)$
  - continu : $(U_{i_j} - L_{i_j})/(U_i - L_i)$
- Moyenne des acquis/acquérables, utilité décroissante // *Loss* croît

Exemple avec 4-anonymat

CP	Age	Genre	Nationalité	Pathologie
{13053 13058}	[21; 31[	*	*	trouble cardiaque
	[21; 31[	*	*	trouble cardiaque
	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
{14850 14853}	[41; 50[	*	*	cancer
	[41; 50[	*	*	trouble cardiaque
	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
{13053 13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

- Nat. $= \{\text{russe}, \text{am.}, \text{jap.}, \text{ind.}\} \rightsquigarrow$ $R_{\text{Nat}} = 3$, $R_{\text{CP}} = 3$, $R_{\text{Genre}} = 1$,
- $R_{\text{Age}} = 50 - 21 = 29$
- $Loss = \frac{1}{4\times 12}(8 \times (\frac{1}{3} + \frac{10}{29} + \frac{1}{1} + \frac{3}{3}) + 4 \times (\frac{1}{3} + \frac{9}{29} + \frac{1}{1} + \frac{3}{3})) = \frac{2}{3}$

Plan

k-anonymat : attaques

Exemple

CP	Age	Genre	Nationalité	Pathologie
	[21; 31[	*	*	trouble cardiaque
{13053	[21; 31[	*	*	trouble cardiaque
13058}	[21; 31[	*	*	infection virale
	[21; 31[	*	*	infection virale
	[41; 50[	*	*	cancer
{14850	[41; 50[	*	*	trouble cardiaque
14853}	[41; 50[	*	*	infection virale
	[41; 50[	*	*	infection virale
	[31; 41[	*	*	cancer
{13053	[31; 41[	*	*	cancer
13058}	[31; 41[	*	*	cancer
	[31; 41[	*	*	cancer

4 individus

4 individus

4 individus

Attaques

- Homogénéité :
  - $\oplus$ Patient de 35 ans connu $\rightsquigarrow$ cancer.
  - $\ominus$ Patient de 29 ans connu $\rightsquigarrow$ ~~cancer~~.
- Connaissance supplémentaire : un japonais de 21 ans, P(trouble cardiaque|japonais)=faible $\rightsquigarrow$ infection virale.

Plan

Un premier modèle de PVP : le k-anonymat

Extensions du k-anonymat

- Casser l'homogénéité par l-diversité
- Préserver les distributions sensibles par t-proximité

Inspiré entre autre de [4]

4. Nguyen, B., & Castelluccia, C. (2020). Techniques d'anonymisation tabulaire : concepts et mise en oeuvre. arXiv preprint arXiv :2001.02650.

Plan

l-diversité pour contrer l'homogénéité

Définition et remarque

- Classe d'équivalence : l-diverse si contient au moins l valeurs "représentatives" par donnée sensible
- Base de données : l-diverse si toutes ses classes d'équivalence le sont
- Représentatives : valeurs distinctes $\subset$ répartition sensée...

4-anonymat et 3-diversité distinctes

	Quasi-IDentifiants				Sensibles
H	CP	Age	Genre	Nationalité	Pathologie
1	130**	[21; 41[	*	*	trouble cardiaque
2	130**	[21; 41[	*	*	trouble cardiaque
3	130**	[21; 41[	*	*	infection virale
4	130**	[21; 41[	*	*	infection virale
9	130**	[21; 41[	*	*	cancer
10	130**	[21; 41[	*	*	cancer
11	130**	[21; 41[	*	*	cancer
12	130**	[31; 41[	*	*	cancer
5	148**	[41; 50[	*	*	cancer
6	148**	[41; 50[	*	*	trouble cardiaque
7	148**	[41; 50[	*	*	infection virale
8	148**	[41; 50[	*	*	infection virale

} 3 val. sensibles $\neq$

} 3 val. sensibles $\neq$

(c, l)-diversité récursive

Définition pour chaque classe d'équivalence EQ

$$\forall \text{EQ}\ r_1 < c(r_l + r_{l+1} + \cdots + r_m)$$

- $r_1, r_2, \ldots, r_m$: fréq. décroissantes des valeurs des attributs sensibles
- c : facteur (réel, >1) de tolérance de variations des extrêmes

4-anonymat et $(2.0000000000000001, 3)$-diversité récursive

	Quasi-IDentifiants				Sensibles
H	CP	Age	Genre	Nationalité	Pathologie
1	130**	[21; 41[	*	*	trouble cardiaque
2	130**	[21; 41[	*	*	trouble cardiaque
3	130**	[21; 41[	*	*	infection virale
4	130**	[21; 41[	*	*	infection virale
9	130**	[21; 41[	*	*	cancer
10	130**	[21; 41[	*	*	cancer
11	130**	[21; 41[	*	*	cancer
12	130**	[31; 41[	*	*	cancer
5	148**	[41; 50[	*	*	cancer
6	148**	[41; 50[	*	*	trouble cardiaque
7	148**	[41; 50[	*	*	infection virale
8	148**	[41; 50[	*	*	infection virale

- Dans EQ_{130**} et EQ_{148**} : $r_1 = 0.5$, $r_2 = r_3 = 0.25$
- $r_1 = 2.r_3 \rightsquigarrow$ $r_1 < 2.0000000000000001 \times r_3$

/-diversité : mise en œuvre difficile

Exemple de sur-représentation d'une valeur [5]

- Données originales 10000 personnes :
  - Un seul attribut sensible : résultat de test viral.
  - Deux valeurs : positive (1%) et négative (99%).
- Valeurs avec des degrés de sensibilité très $\neq$:
  - Peu d'opposition que l'on sache que le test est négatif (comme 99 % de la population)
  - Forte réticence à être connu positif
- 2-diversité :
  - Au maximum $10000 \times 1\% = 100$ classes d'équivalence
  - $\rightsquigarrow$ perte d'informations importante.

5. Li, N., Li, T., & Venkatasubramanian, S. (2009). Closeness : A new privacy measure for data publishing. IEEE Transactions on Knowledge and Data Engineering, 22(7), 943-956.

/-diversité : statistiques mensongères

Exemple précédent avec une classe d'équivalence équilibrée

- Satisfait la 2-diversité :
  - distincte
  - $(c, 2)$ récursive, pour $c > 1$.
- Toute personne de cette classe : considérée 1 fois/2 positive

Exemple précédent avec une classe d'équivalence en 49/1

- 49 positifs / 1 négatif : satisfait la 2-diversité :
  - distincte
- Mais une personne de cette classe : considérée positive à 98% (vs. 1%).
- Mais cette classe a exactement la même diversité qu'une classe avec 1 positif / 49 négatifs

Avec des éléments $\neq$ mais sémantiqu^t proches ?

Exemple agrégé avec salaire et maladies sensibles

CP	Age	Salaire	Pathologie
476**	2*	3K	ulcère gastrique
476**	2*	4K	gastrite
476**	2*	5K	cancer de l'estomac
4790*	$\geq$ 40	6K	gastrite
4790*	$\geq$ 40	11K	grippe
4790*	$\geq$ 40	8K	bronchite
476**	3*	7K	bronchite
476**	3*	9K	pneumonie
476**	3*	10K	cancer de l'estomac

- Satisfait
  - la 3-diversité distincte.

Fuite due à la non prise en compte de la proximité de valeurs.

Déductions possibles de la connaissance que Bob est dans la classe 1 :

- son salaire ([3K-5K]) est relativement bas
- souffre de l'estomac (toutes les pathologies y sont liées)

Plan

t-proximité[5]

Définition : vérification de la $t-$ *proximité*

- dans chaque classe EQ si pour chaque attribut sensible, la *distance* entre sa distribution dans EQ et sa distribution dans la table complète est $\leq t$
- dans la base complète si toutes ses classes d'équivalence la respectent
- attribut sensible $\equiv$ pas de généralisation

EMD entre les distributions $P = (p_1, p_2, \ldots, p_n)$ et $Q = (q_1, q_2, \ldots, q_n)$

- Travail minimal à fournir pour modifier un tas de terre en un autre
- Pour un attribut numérique : $v_1 < v_2 < \cdots < v_m$

  $D(P, Q) = |p_1 - q_1| + |p_2 - q_2 + p_1 - q_1| + \ldots |p_m - q_m + \cdots + p_1 - q_1|$

- Pour un attribut catégoriel : $\{v_1, v_2, \ldots, v_m\}$ par distance au sol :

$$D(P, Q) = \frac{1}{2} \sum_{i=1}^{m} |p_i - q_i|$$

t-proximité sur un exemple

Données originales et généralisées

QID				Sensibles
CP	Age	Genre	Nationalité	Pathologie
13053	28	H	russe	trouble cardiaque
13068	29	H	américaine	trouble cardiaque
13068	21	F	japonaise	infection virale
13053	23	H	américaine	infection virale
14853	49	H	indienne	cancer
14853	48	F	russe	trouble cardiaque
14850	47	H	américaine	infection virale
14850	49	F	américaine	infection virale
13053	31	H	américaine	cancer
13053	37	H	indienne	cancer
13068	36	F	japonaise	cancer
13068	35	F	américaine	cancer

$q_{\text{tc}} = \frac{1}{4}$ $q_{\text{iv}} = \frac{1}{3}$ $q_{\text{c}} = \frac{5}{12}$

QID				Sensibles
CP	Age	Genre	Nationalité	Pathologie
*	*	F	*	infection virale
*	*	F	*	trouble cardiaque
*	*	F	*	infection virale
*	*	F	*	cancer
*	*	F	*	cancer
*	*	H	*	trouble cardiaque
*	*	H	*	trouble cardiaque
*	*	H	*	infection virale
*	*	H	*	cancer
*	*	H	*	infection virale
*	*	H	*	cancer
*	*	H	*	cancer

$p^F_{\text{tc}} = \frac{1}{5}$, $p^F_{\text{iv}} = \frac{2}{5}$, $p^F_{\text{c}} = \frac{2}{5}$ et $p^H_{\text{tc}} = \frac{2}{7}$, $p^H_{\text{iv}} = \frac{2}{7}$, $p^H_{\text{c}} = \frac{3}{7}$

Distances entre les distributions P^F et Q puis P^H et Q

- $D(P^F, Q) = \frac{1}{2}\left(\left|\frac{1}{5} - \frac{1}{4}\right| + \left|\frac{2}{5} - \frac{1}{3}\right| + \left|\frac{2}{5} - \frac{5}{12}\right|\right) = \frac{1}{15} \approx 0.06666666$
- $D(P^H, Q) = \frac{1}{2}\left(\left|\frac{2}{7} - \frac{1}{4}\right| + \left|\frac{2}{7} - \frac{1}{3}\right| + \left|\frac{3}{7} - \frac{5}{12}\right|\right) = \frac{1}{21} \approx 0.0476190$
- Données généralisées $\frac{1}{15}$-proches des originales

Mesure du gain d'information moyen $\mathcal{A}_{\text{know}}$

Définition [6] pour chaque classe EQ

$$\mathcal{A}_{\text{know}} = \frac{1}{|T|} \sum_{eq \in EQs} |eq| d(P^{eq}, Q)$$

- Un tuple de quasi-identifiants $\rightsquigarrow$ classe eq $\rightsquigarrow$ quantité moyenne d'informations apprise sur les attributs sensibles

$\mathcal{A}_{\text{know}}$ sur l'exemple

QID				Sensibles
CP	Age	Genre	Nationalité	Pathologie
*	*	F	*	infection virale
*	*	F	*	trouble cardiaque
*	*	F	*	infection virale
*	*	F	*	cancer
*	*	F	*	cancer
*	*	H	*	trouble cardiaque
*	*	H	*	trouble cardiaque
*	*	H	*	infection virale
*	*	H	*	cancer
*	*	H	*	infection virale
*	*	H	*	cancer
*	*	H	*	cancer

- $\mathcal{A}_{\text{know}} = \frac{1}{12}\left(5\frac{1}{15} + 7\frac{1}{21}\right) = \frac{1}{18} \approx 5.5\%$
- Femme ds la base connue et publication $\rightsquigarrow$ chance d'avoir une IV augmentée

6. Brickell, J., & Shmatikov, V. (2008, August). The cost of privacy : destruction of data-mining utility in anonymized data publishing. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 70-78).