Sécurité Appliquée
Protection de la vie privée-PVP

Jean-François Couchot

Université de Franche-Comté, UFR-ST

Plan

Extensions du k-anonymat

Ajouter du bruit aux résultats ?

Plan

Extensions du k-anonymat

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

Ajouter du bruit aux résultats ?

Inspiré entre autre de [1]

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

Plan

Rappels de l'exemple

Les données

	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

Généralisation selon une hiérarchie

- CP : suppr. chiffre D→G : XXXX*, XXX**,XX***, X****, suppr.
- Age : par intervalles d'ampl. croissante : 10, 20, 40, suppr.
- Genre : suppr.
- Nationalité : par continent, suppr.

Homogénéité problématique

4-anonymat minimisant *Loss*

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

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 $\neq$

l-diversité vérifiée par entropie de Shanon

Définition pour chaque classe d'équivalence EQ

$$\forall \text{EQ} \qquad H_{\text{EQ}} = -\sum_{s \in S(\text{EQ})} p(s).\log_2 p(s) \geq \log_2 l$$

- $S(\text{EQ})$: domaine des attributs sensibles de EQ
- $p(s)$ estimée par $\frac{|s|}{\sum_{\sigma \in S(\text{EQ})} |\sigma|}$
- H_{EQ} : quantité d'information contenue dans $S(\text{EQ})$

4-anonymat et 3-diversité pas vérifiée par l'entropie de Shanon

	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**} : p(tr.card)= p(inf. vir.)= 0.25 et p(cancer)=0.5
- $H(\text{EQ}_{130**}) = -(2 \times 0.25 \log_2(0.25) + 0.5 \log_2(0.5)) = 1.5 \log_2(2) = 1.5 \not\geq \log_2(3) \approx 1.58$
- 3-diversité : aucune généralisation vérifiant l'entropie de Shanon

(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 [2]

- 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.

2. 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
  - d'entropie de Shanon,
  - $(c, 2)$ 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
  - d'entropie de Shanon,
- 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.
  - la 3-diversité vérifiée par entropie de Shanon

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[2]

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 [3] 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

3. 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).

Plan

Extensions du k-anonymat

Ajouter du bruit aux résultats ?

- Notations et motivation
- Quelle quantité α de bruit ajouter ?

Plan

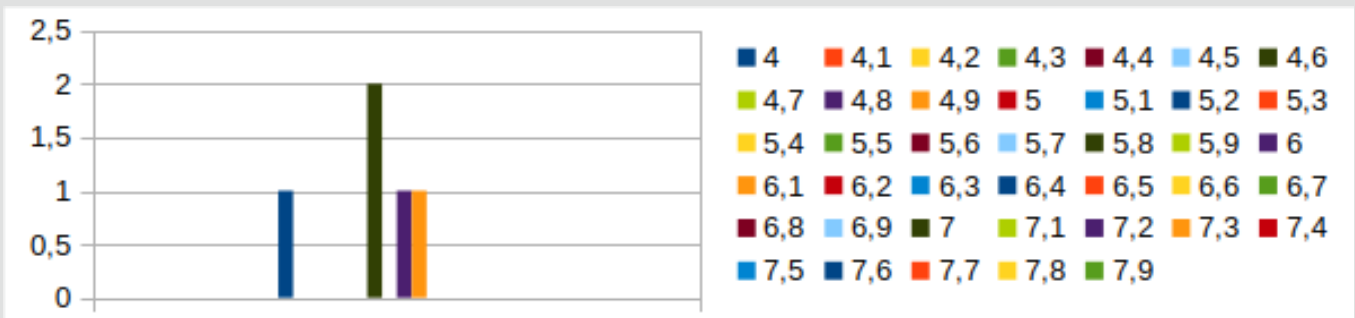
Des données ultra simples

Représentation

- $\mathcal{X}$: l'univers des valeurs possibles
- Base D : multi-ensemble d'éléments de $\mathcal{X}$
- Représentée par un histogramme : $D \in \mathbb{N}^{|\mathcal{X}|}$

Exemple sur une base de tailles en pieds

$D = 5'2,\ 6'1,\ 5'8,\ 5'8,\ 6'0$ et $\mathcal{X} = \{4'0, 4'1, \ldots, 7'9\}$ par ex.



Requête de comptage sur $D = \{\ldots, v_i : e_i, \ldots\}$

Définition d'une requête de comptage Q_S

- $S \subseteq \mathcal{X}$: valeurs de $\mathcal{X}$ à compter
- $Q_S(D) = \sum_{v_i \in S} e_i$

D_1=$\{5'2 : 1, 5'8 : 2, 6'0 : 1, 6'1 : 1\}$ D_2=$\{5'2 : 2, 5'3 : 1, 5'8 : 1, 6'0 : 1, 6'1 : 1\}$

- $S = \{5'2, 5'3, 5'8\}$
- $Q_S(D_1) = 1 + 0 + 2 = 3$ et $Q_S(D_2) = 2 + 1 + 1 = 4$

Comme un produit de vecteurs

- D_1=$(\ldots, 1, 0, \ldots, 2, 0, 1, 1, \ldots)$, D_2=$(\ldots, 2, 1, \ldots, 1, 0, 1, 1, \ldots)$
- S=$(\ldots, 1, 1, \ldots, 1, 0, 0, 0, \ldots)$
- $Q_S(D_1) = S.D_1^T = 3, Q_S(D_2) = S.D_2^T = 4$

Distance entre bases de données

Norme 1 pour un vecteur $x = (x_1, \ldots, x_k)$ de $\mathbb{R}^k$

la norme 1

$$\|x\|_1 = |x_1| + \ldots + |x_k|$$

induit la distance de déplacement à angle droit sur un damier (distance de Manhattan)

Norme 1 entre deux bases D_1 et D_2

- Exprimer D_1 et D_2 comme deux vecteurs
- Norme-1 du vecteur de différences entre les deux $\|D_1 - D_2\|_1$

D_1={5′2 : 1, 5′8 : 2, 6′0 : 1, 6′1 : 1} D_2={5′2 : 2, 5′3 : 1, 5′8 : 1, 6′0 : 1, 6′1 : 1}

- $D_1 = (\ldots, 1, 0, \ldots, 2, 0, 1, 1, \ldots) \rightsquigarrow \|D_1\|_1 = |1| + |2| + |1| + |1| = 5$
- $D_2 = (\ldots, 2, 1, \ldots, 1, 0, 1, 1, \ldots) \rightsquigarrow \|D_2\|_1 = |2| + |1| + |1| + |1| + |1| = 6$
- $\|D_1 - D_2\|_1 = \|(\ldots, 1, 1, \ldots, -1, 0, 0, 0, \ldots)\|_1 = |1| + |1| + |-1| = 3$

Borner le bruit du comptage

Motivation

- Peut-on trouver un algorithme respectueux qui retourne un comptage proche de l'original ?

Définition du bruit borné par α [4]

Un algorithme Q' ajoute un bruit borné par α si pour toute base D et toute requête S on a

$$|Q_S(D) - Q'_S(D)| \leq \alpha$$

Exemple de comptage bruité

$Q'_S(D) = Q_S(D) + b$ avec un bruit b t.q.

- $b \sim \mathcal{N}(0, 3\alpha)$ et $abs(b) < \alpha$
- $b \sim \mathcal{U}[-\alpha, \alpha]$
- ...

4. Dinur, I., & Nissim, K. (2003, June). Revealing information while preserving privacy. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 202-210).

Plan

Borner par α : D est $\approx$ reconstructible !

Théorème : attaque d'un algorithme Q' ajoutant un bruit borné par α

- si $Q'_S(D)$ connu pour tout S : on peut construire D' tq $\|D - D'\|_1 \leq 4\alpha$
- En pratique, D' est encore plus proche de D

Preuve

- Construction d'une solution D' :
  - par brute force sur l'ensemble des possibilités (infinies) sur $\mathcal{X}$
  - et qui vérifie $|Q_S(D') - Q'_S(D)| < \alpha$ pour chaque requête (finie) S
- $S_0 = D' \setminus D$ et $S_1 = D \setminus D'$
- Si $\|D - D'\|_1 = \|S_0\|_1 + \|S_1\|_1 > 4\alpha$
  - $\max(\|S_0\|_1, \|S_1\|_1) > 2\alpha \rightsquigarrow$ on suppose $\|S_0\|_1 > 2\alpha$
  - or $Q_{S_0}(D) = 0 \rightsquigarrow Q'_{S_0}(D) \leq \alpha$ (ajout de bruit par Q' borné par α)
  - or $Q_{S_0}(D') > 2\alpha \rightsquigarrow |Q_{S_0}(D') - Q'_{S_0}(D)| > |2\alpha - \alpha| = \alpha$ ↯

Une mauvaise nouvelle ?

Non démontré, mais evident

- $\alpha < \frac{n}{40}$: bruit ajouté faible mais 90% des données sont reconstructibles
- $\alpha = \frac{n}{2}$: données protégées mais utilité ?

Réalisme de la construction de D' ?

1. Générer une distribution sur $\mathcal{X}$
2. Générer toutes les requêtes S et tester, reprendre en 1 ev.
3. En $(\mathcal{O}(\exp(n)))$

En TP :

Ajouter un bruit $\alpha = o(\sqrt{n})$ est attaquable en $\mathcal{O}(n \log(n))$