Revue littéraire - Data Mining Ethics in Privacy Preservation

Thomas

September 2017

Présentation de règles d'association préservant la confidentialité des données dans le cadre du data minning.

1 Règles d'association

1.1 Notations

Soit $I = \{I_1, I_2, ..., I_m\}$ un ensemble d'éléments et D une base de données de transactions telles que :

- Chaque transaction T est un ensemble d'éléments où : $T \subseteq I$ i.e T est un sous-ensemble de I.
- Chaque transaction est associé à un indentificateur appelé TID.
- Une transaction contient A si et seulement si $A \subseteq T$.

1.2 Définition d'une règle

Une règle d'association est une implication de la forme : $A \Rightarrow B$ avec :

$$A \subseteq I,\ B \subseteq I,\ et\ A \cap B = \emptyset$$

1.3 Support de la règle

Le support de la règle représente la proportion des transactions de D contenant $(A \cup B)$, tel que :

$$Sup(A \Rightarrow B) = \mathbb{P}(A \cup B) = \frac{\mid A \cup B \mid}{\mid D \mid}$$

1.4 Confiance de la règle

La confiance de la règle $(A \Rightarrow B)$ dans D est définie par :

$$Conf(A \Rightarrow B) = \mathbb{P}(A \mid B) = \frac{\mid A \cup B \mid}{\mid A \mid}$$

où $\mid A \mid$ est le "compteur de support" (*support count*) des éléments de A dans D.

1.5 Propriétés

Les règles qui vérifient des valeurs minimales de support et de confiance sont qualifiées de *règles fortes*.

Un échantillon contenant k éléments est un *k-élément échantillon*.

Les échantillons qui vérifient un min_sup sont appelés *échantillons d'éléments fréquents*.

1.6 Algorithmes pour préserver l'anonymat

On compte trois catégories d'algorithmes pour préserver l'anonymat des données lors de l'établissement des règles d'association :

1. Techniques basées sur des heuristiques
2. Règles d'association basées sur la reconstruction
3. Cryptographie

2 Techniques basées sur des heuristiques

L'objectif ici est de modifier les données en altérant les valeurs d'attributs : en les remplaçant par une autre valeur, en y ajoutant un terme de bruit ou encore en remplaçant la valeur correspondante par une valeur inconnue.
Pour savoir quelle transaction ou quel élément d'un échantillon doit être modifié, il faut faire en sorte de réduire au maximum l'influence de la base de données originale.

3 Règles d'association basées sur la reconstruction

Cette partie est basée sur le travail de Agrawal et al. ***Privacy-preserving Data Minning****.*

L'objectif est de développer des modèles précis sans avoir accès à des informations individuelles sur les données. Notamment d'obtenir un classifieur sous forme d'arbre de décision basé sur un jeu d'apprentissage dont les valeurs des observations individuelles ont été perturbées.
Le jeu de données alors obtenu diffère fortement du jeu orignal de sorte que la distribution des valeurs des données est également très différente des données initiales. Alors qu'il est donc devenu impossible d'estimer correctement les valeurs des données au départ, l'idée ici est d'avoir recours à une procédure de reconstruction qui permette d'estimer la **distribution** initiale des données.
Grâce à ces distributions reconstruites il est alors possible de déterminer des estimateurs dont la précision est comparable à ceux élaborés à partir du jeu de donnés original.

3.1 Méthodes pour préserver l'anonymat

On considère deux types de méthodes afin de préserver l'anonymat "individuel" du jeu de donnés :

1. Restrictions sur les transactions/queries
2. Perturbation des données

3.1.1 Restriction sur les transactions

Cette méthode comprend :

- Restreindre la taille des queries.
- Contrôler "l'entrecoupage" (*overlap*) des transacations successives.
- Garder une trace de toutes les transactions afin de constamment pouvoir vérifier si l'on ne peut pas trouver de compromis[1] d'anonymie sur la base de données.
- Supprimer les données de petites tailles
- Regrouper certaines données

[1] Un compromis permet d'utiliser une séquence de transactions afin d'en déduire des informations privées censées demeurer secrètes

	Niveau de confiance		
	50%	95%	99.9%
Discrétisation	$0.5 \times W$	$0.95 \times W$	$0.999 \times W$
Uniforme	$0.5 \times 2\alpha$	$0.95 \times 2\alpha$	$0.999 \times 2\alpha$
Gaussien	$1.34 \times \sigma$	$3.92 \times \sigma$	$6.8 \times \sigma$

Table 1: Quantité d'anonymat par méthode en fonction du niveau de confiance

3.1.2 Perturbation des données

Cette méthode comprend :

- Intervertir les valeurs de certaines observations
- Remplacer la base de données originale par un échantillon de même distribution
- Ajouter du bruit aux résultats issus d'une transaction

Il existe par ailleurs plusieurs méthodes afin de modifier la valeur d'un champs.
Notamment en définissant son appartenance à une classe-valeur : Il suffit de partitionner les valeurs de l'attribut en des ensembles disjoints, mutuellement exhaustifs et ne retourner que la classe dans laquelle se trouve la "vraie" valeur de x_i lors de la transaction.
Il est également possible d'avoir recours à la distorsion de valeur, c'est à dire renvoyer une valeur $x_i + r$ au lieu de simplement x_i, où r est une V.A issue d'une distribution paramétrée connue (typiquement ici on prendra $r \sim \mathcal{N}(0, \sigma^2)$ ou encore $r \sim \mathcal{U}(-\alpha, \alpha)$), où r s'interprète comme un terme de bruit.
Enfin la méthode de dissociation permet d'avoir une vraie valeur dans un champs pour une observation tant que cette valeur provient du même champ d'une autre observation strictement distincte. (*Cette méthode n'est pas utilisée ici car elle nécessite la connaissance des valeurs des autres observations*).

3.2 Mesure de la qualité

L'objectif est de pouvoir obtenir des statistiques de bonnes qualités tout en s'assurant de ne pas avoir de fuites partielles ou totales de données individuelles privées.
La qualité statistique est donc mesuré en terme de :

Biais : Différence entre les statistiques non-perturbées et les espérances des estimations perturbées.

Précision : Variance des estimateurs obtenus résultant des modifications.

Divulgation exact : Obtenir à la suite d'une ou plusieurs requêtes la valeur exacte d'un attribut confidentiel d'un individu.

Divulgation partielle : Obtenir un estimateur suffisamment précis (Var $\ll \epsilon$) sur un attribut privé.

Afin de quantifier l'anonymat fournie par une méthode on utilise une mesure qui permet de déterminer la précision avec laquelle les valeurs orignales peuvent être estimées à partir des attributs modifiés.
S'il peut être estimé avec une confiance de $c\%$ que la valeur x appartient à l'interval $[x_1, x_2]$, alors la taille de l'intervalle $(x_2 - x_1)$ défini la quantité d'anonymat au niveau c offert par la méthode (cf. *Table 1.*).
On notera que W correspond à la largeur de l'intervalle dans le cadre d'une discrétisation, on comprend alors que pour obtenir un anonymat important avec cette méthode il est nécessaire d'augmenter la taille des intervalles (et donc réduire leurs nombres). Malheureusement les résultats obtenus seront donc très imprécis puisque toutes les valeurs d'un intervalle seront modifiées et redéfinie à la même valeur.

On en déduit donc que la meilleure méthode correspond à ajouter un bruit gaussien aux données (méthode de distorsion gaussienne) et dans une moindre mesure celle ajoutant un bruit uniforme. C'est pourquoi l'étude se focalise désormais uniquement sur ces deux méthodes.

3.3 Reconstruction de la distribution originale

Afin de pouvoir utiliser les méthodes ci-dessus, il est nécessaire de s'assurer pouvoir reconstruire la distribution initiale des données à partir de celles "bruitées".
On rappelle que l'on ne cherche pas à reconstruire les valeurs individuelles orignales, mais seulement leurs ***distributions****.*

Problème de reconstruction :
On a F_Y la distribution cumulative de Y et les réalisation de n V.A iid : $X_1+Y_1, X_2+Y_2, ..., X_n+Y_n$. On w_i la valeur de $X_i + Y_i$ connue.

On peut utiliser les formules de Bayes afin d'estimer la fonction de distribution a postériori F'_{X_1} de X_1, en supposant que l'on connaisse les fonctions de densité f_X et f_Y pour les variables X et Y (ce qui n'est pas le cas en pratique).

On a :

$$F'_{X_1}(a) = \int_{-\infty}^{a} f_{X_1}(z \mid X_1 + Y_1 = w_1)dz$$

$$\dots$$

$$F'_{X_1}(a) = \frac{\int_{-\infty}^{a} f_Y(w_1 - z) f_X(z)dz}{\int_{-\infty}^{\infty} f_Y(w_1 - z) f_X(z)dz}$$

Pour estimer la fonction de distribution à posteriori F'_X sachant $X_1 + Y_1$, $X_2 + Y_2$, ..., $X_n + Y_n$, il suffit de faire la moyenne des fonctions de distribution des F'_{X_i} :

$$F'_{X_1}(a) = \frac{1}{n} \sum_{i=1}^{n} F'_{X_i}$$

Pour obtenir la densité il suffit alors de différencier la fonction ainsi obtenue, tel que :

$$f'_{X_1}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - a) f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X(z)dz}$$

Avec un nombre suffisamment grands d'échantillons on estime que f'_X sera très proche de la densité originale f_X. Néanmoins ici seule la densité f_Y est connu, il faut donc initialiser la densité f_X en utilisant une distribution uniforme pour f^0_X puis redéfinir itérativement cette densité au cours de l'exécution de l'algorithme :

$f^0_X \leftarrow Distribution\, Uniforme$
$j \leftarrow 0$
while $Critère\, d'arrêt$ **do**
  $f^{j+1}_X(a) \leftarrow \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - a) f^j_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f^j_X(z)dz}$
  $j \leftarrow j + 1$
end while

Le *critère d'arrêt* proposé par l'article correspond à comparer successivement les estimations de la distribution original par l'algorithme, puis d'arrêter lorsque les différences entre ces distributions deviennent très petites.

3.4 Accélérer la reconstruction

On peut recourir au partitionnement des données au sein d'intervalles afin d'améliorer les performances de l'algorithme précédent grâce à deux approximations :

- Approximation de la distance entre a et w_i par la distance séparant les points médians des intervalles dans lesquels ils appartiennent.
- Approximation de $f_X(a)$ par la moyenne de la fonction de densité sur l'intervalle dans lequel a se trouve.

3.5 Construire un arbre de décision sur les données bruitées

Afin de pouvoir élaborer notre arbre de décision sur ces données, il faut changer deux éléments dans l'algorithme classique de construction d'arbre de décision :

- La manière de déterminer les "splits".
- Le partitionnement des données au sein des branches de l'arbre.

Par ailleurs, il faut également choisir une méthode de reconstruction de la distribution des données parmi plusieurs possibilités :

- Faut-il réaliser une reconstruction globale en utilisant toute les données, ou d'abord partitionner les données par classe et reconstruire séparément chacune des classes ?
- Est-il préférable d'effectuer une reconstruction à chaque noeud de l'arbre ou uniquement à sa racine ?

En raison du partitionnement des données lors de la reconstruction, les "splits" seront forcément les bornes des intervalles : pour chaque "split" on utilise les statistiques issues de la distribution reconstruite afin de calculer l'index de Gini.

On considère trois algorithmes différents afin de reconstruire les distributions au sein l'abre de décision :

- **Global :** Reconstruire la distribution pour chaque attribut une fois au départ, puis utiliser les données reconstruite pour réaliser l'arbre.
- **ByClass :** Séparer chaque attribut au sein du jeu d'apprentissage par classe, reconstruire les distributions séparément pour chaque classe, puis élaborer l'arbre à partir des données reconstruite.
- **Local :** Similaire au *ByClass*, sauf que la reconstruction n'a pas lieu qu'une seule fois, mais à chaque noeud. Cependant pour éviter le sur-apprentissage, la reconstruction est stoppée après qu'un nombre minimal d'éléments appartenant à un noeud a été atteint.

On remarque que l'algorithme *Local* est le plus coûteux en terme de temps d'exécution, alors que la méthode *Global* s'avère être la plus rapide. L'algorithme *ByClass* se trouve entre les deux, bien que plus proche du temps d'exécution de la méthode *Globale* puisqu'ici aussi la reconstruction n'a lieu qu'au niveau de la racine.

Les techniques présentées ici s'appliquent au cas supervisé. Dans le cas contraire, le modèle de classification doit être directement appliqué chez l'utilisateur.

3.6 Résultats obtenus

Les trois méthodes de reconstruction présentées ci-dessus ont été testées. Par ailleurs afin de déterminer la précision de ce classifieur, il a été testé sur le jeu de donnés originale et sa version bruitée.
Les tests sont réalisés avec différentes fonctions de classification ayant chacune une surface de décision différente et plus ou moins complexe. Par ailleurs les valeurs de confidentialité ont également été modifiées au cours de ces essais, en variant de 25% à 200%.

Exemple :
Si la variable "salaire" est uniformément distribuée entre 20K et 150K, alors avec un taux de confidentialité de 50%, il est impossible d'évaluer la valeur du salaire d'un individu (avec une confiance de 95%) autrement qu'au sein d'un intervalle de taille 65K ($= \frac{1}{2} * (150K - 20K)$*).*

Les algorithmes *Local* et *ByClass* se comportent particulièrement bien aux niveaux de confidentialité 25%, 50% et dans une moindre mesure 100%, en garantissant une bonne précision comparée au classifieur sur les données non bruitées.
L'algorithme *Global* de part son implémentation qui utilise la même distribution reconstruite pour toutes les classes, ne fournit pas d'aussi bons résultats.
Enfin lorsque l'on dépasse un niveau de confidentialité de 100%, la précision des méthodes en est fortement impactée et se détériore très vite.

Ces méthodes ont été élaborées sur des données quantitatives uniquement, une démarche possible consisterait à les transposer aux données catégorielles.

References

[1] S. M. Mahajan and A. K. Reshamwala, *Data Mining Ethics in Privacy Preservation.* , International Journal of Computer Theory and Engineering, Vol. 3, No. 4, August 2011.