11ème Colloque Africain sur la Recherche en Informatique et en Mathématiques

Blocage des canaux d'inférences dans les données *k*-anonymes

Ousseynou Sané* — Fodé Camara* — Samba Ndiaye* — Yahya Slimani**

* Département mathématiques-informatique, Faculté des Sciences et Techniques
Université Cheikh Anta Diop de Dakar
SENEGAL
{fode.camara, samba.ndiaye}@ucad.edu.sn

** Département d'informatique, Faculté des Sciences
Université Tunis
TUNISIE
yahya.slimani@fst.rnu.tn

RÉSUMÉ. Le modèle de protection *k*-anonymity a été proposé pour protéger l'anonymat des individus dans les microdata à publier. Toutefois, à partir d'un microdata *k*-anonyme, il est possible d'inférer directement les données privées. Cette inférence directe est appelée "attribute linkage". En outre, le modèle *k*-anonymity souffre d'une autre forme d'attaque (inférence indirecte) basée sur les résultats de datamining. En effet, les modèles et règles de datamining constituent une menace de violation de vie privée même dans les microdata *k*-anonymes. Dans ce papier, nous illustrons d'abord le bien fondé de cette menace, puis proposons une approche qui élimine ces brèches de confidentialité. Nous avons validé expérimentalement notre proposition en utilisant le jeu de données classique "*Adult Data Set*" de l'*UCI*. Les résultats expérimentaux obtenus ont montré son efficacité.

ABSTRACT. The concept of *k*-anonymity protection model has been proposed as an effective way to protect the identities of subjects in a disclosed database. However, from a *k*-anonymous dataset it may be possible to directly infer private data. This direct disclosure is called attribute linkage. *k*-anonymity also suffer from another form of attack based on data mining results. In fact, data mining models and patterns pose a privacy threat even if the *k*-anonymity is satisfied. In this paper, we discuss how the privacy requirements characterized by *k*-anonymity can be violated by data mining results and introduce an approach to limit privacy breaches. We experiment it by using the adult dataset from the UCI KDD archive. We report the experimental results which show its effectiveness.

MOTS-CLÉS : Datamining, Protection d'anonymat, Vie privée, Anonymisation des données, Attaques de vie privée

KEYWORDS : Data mining, Protection of anonymity, Privacy, Data anonymization, Privacy attacks

1. Introduction

Devenu entre autres une composante clé des systèmes de sécurité de beaucoup de gouvernements, un outil de marketing pour les entreprises pour cibler des offres commerciales, le Datamining est aujourd'hui perçu par la plupart des personnes comme une technologie qui viole leur vie privée. Cette impression négative est alors devenue un obstacle à son avancement. Par exemple, un projet de recherche de datamining potentiellement bénéfique, appelé Terrorism Information Awareness (TIA), a été résilié par le Congrès américain principalement en raison de ses styles controversés de la collecte des données. En outre, de nombreuses études et expériences réelles ont montré comment des données apparemment anonymes pouvaient être analysées ou recoupées avec d'autres données disponibles pour retrouver les personnes concernées. De fait, de simples résultats statistiques établissent par exemple que 87 % des citoyens des Etats-Unis peuvent être identifiés de manière unique à partir de la seule connaissance de leur code postal, date de naissance et sexe [1].

Le concept de k-anonymity a été proposé comme une manière efficace de préserver l'anonymat dans les données personnelles collectées et publiées. A cause de sa simplicité et surtout des nombreux algorithmes [1] pour créer une version k-anonyme d'un microdata, le modèle k-anonymity a gagné une grande popularité. Cependant, il est vulnérable aux attaques de type "attribute linkage" (e.g. attaques basées sur l'homogénéité, attaques basées sur la connaissance à priori). Comme montré dans [2], le modèle k-anonymity souffre aussi d'une autre attaque basée sur les résultats de datamining. En effet, les modèles et règles de datamining constituent une menace de violation de vie privée même si le modèle k-anonymity est satisfait.

1.1. Comment les résultats de datamining créent des canaux d'inférences sur les données k-anonymes ?

L'application d'un processus de datamining sur des données anonymes peut révéler des connaissances sensibles. Pour illustrer cette menace, nous présentons d'abord l'algorithme C4.5 [3], ensuite nous montrons que celui-ci crée des canaux d'inférences.

1.1.1. L'algorithme C4.5

L'algorithme C4.5 [3] est une extension de l'algorithme ID3 de Quinlan pour la construction d'arbre de décision. Un des aspects les plus attirants des arbres de décision de C4.5 réside dans leur interprétation, surtout en ce qui concerne la construction des règles de décision. Les règles de décision peuvent être construites à partir d'un arbre de décision très simplement en traversant tous les chemins donnés de la racine vers n'importe quelle feuille. Elles peuvent également être représentées sous la forme antécédent $\Rightarrow$ conséquence. La conséquence est formée par la valeur de l'attribut de classe (i.e. nœud feuille en question). Le support de la règle de décision porte sur la proportion d'enregistrements dans l'ensemble de données qui restent dans un nœud feuille donné. La confiance de la règle porte sur la proportion d'enregistrements du nœud feuille pour lequel la règle de décision est vraie. Si la confiance est égale à 100% (=1), le nœud feuille est pur et la règle de décision est parfaite.

	Non sensible			Sensible
	Code postal	**Age**	**Nationalité**	**Maladie**
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Tableau 1: Un microdata k-anonyme

1.1.2. Canaux d'inférences créés par les règles de décision sur des données k-anonymes

Après le processus d'anonymization, il est possible d'inférer les données sensibles. Pour illustrer ceci, nous donnons l'exemple dans la Table 1 qui représente une version 4-anonyme d'une base de données médicale. Après la construction de l'arbre de décision C4.5, nous avons certaines règles de décision sensibles du genre 3* $\Rightarrow$ Cancer avec une confiance=1. De telles règles sont très utiles pour des recherches médicales, toutefois elles permettent d'inférer la maladie de certains individus parce qu'elles expriment clairement que tous les patients de la base de données initiale, âgés entre 30 et 39 ans sont cancéreux. A travers cet exemple, nous voyons bien que les résultats de datamining peuvent violer la vie privée des individus même à partir de données k-anonymes.

1.2. Contribution et Organisation du papier

Dans ce papier, nous avons montré que les résultats de datamining peuvent causer des brèches de confidentialité. Plus spécifiquement, nous avons montré que les règles de décision de C4.5 créent des canaux d'inférence qu'un adversaire peut utiliser pour retrouver les données privées d'un individu. Notre contribution est double : (i) Nous avons défini le concept de règle de décision sensible qui est potentiellement une menace de violation d'anonymat des données sources ; (ii) Nous avons développé un algorithme efficace qui a pour objectif d'éliminer les menaces d'anonymat en réduisant la confiance des règles de décision sensibles en dessous d'un seuil arbitrairement choisi.

Le reste du papier est structuré comme suit. Dans le chapitre 2, nous présentons les travaux relatifs. Notre proposition est présentée à la Section 3. La Section 4 évalue notre proposition. La Section 5 présente et discute les résultats obtenus. Enfin, la Section 6 résume l'ensemble de nos travaux et donne quelques perspectives pour leur prolongement.

2. Travaux relatifs

Il y a risque de violation de vie privée quand l'identité d'un individu est liée à un enregistrement ou quand elle est liée à une valeur d'un attribut sensible. Ces brèches d'anonymat sont respectivement appelées "record linkage" et "attribute linkage".

2.1. L'attaque par "Record linkage"

Cette attaque est possible quand certaines valeurs q de quasi-identifiants Q identifient un petit nombre d'enregistrements dans T, le microdata à révéler. Dans ce cas, l'individu possédant la valeur q est susceptible d'être lié à un petit nombre de d'enregistrements dans T. Le modèle k-anonymity [1] a été proposé pour combattre les attaques par "record linkage". La garantie obtenue avec celui-ci est qu'aucune information ne pourrait être liée à un groupe d'au moins k individus. Ainsi, le degré d'incertitude de l'attribut sensible est au moins égal à $1/k$. Toutefois le principal inconvénient de ce modèle est sa vulnérabilité aux attaques de type "attribute linkage".

2.2. L'attaque par "Attribute linkage"

Si certaines valeurs de l'attribut sensible sont prédominantes dans une classe d'équivalence (i.e. un groupe d'enregistrements ayant les mêmes valeurs de quasi-identifiants), un adversaire n'aurait pas de difficultés à les relier aux individus en question dans ce groupe. De telles attaques sont appelées "attribute linkage". Vu cette vulnérabilité, plusieurs modèles ont été définis pour combattre les attaques par "attribute linkage". Parmi ces modèles, nous pouvons citer l-diversity, (α,β) - Anonymity et (X,Y) - privacy [1]. Cependant, ces derniers sont souvent difficiles à satisfaire et compromettent généralement l'utilité et la valeur des résultats [5]. En effet, trouver un juste équilibre entre le niveau de protection de vie privée et la perte d'utilité résultant du processus d'anonymization est une importante direction de recherche.

---

3. Approche proposée

3.1. Définition du problème

Avant de présenter notre problème de protection d'anonymat, nous définissons quelques concepts de base.

Définition 1. Soit T un microdata et $A = a_1, a_2, ..., a_m$ un ensemble d'attributs. Une règle de décision est une implication de la forme $X \Rightarrow Y$, où $X \subseteq A$, $X \subseteq A$, et $X \cap Y = \Phi$. La règle de décision $X \Rightarrow Y$ présente dans T avec une confiance c égale à 1 est communément appelée *règle de décision parfaite*. Dans notre contexte, nous dénotons une telle *règle de décision sensible*.

Formellement, nous appelons canal d'inférences toute collection de règles de décision sensibles à partir de la quelle on pourrait inférer les données privées d'un individu.

Définition 2. Soit S un ensemble de règles de décision sensibles extraites à partir du microdata T et k un seuil d'anonymat, notre problème consiste à réduire la confiance de chaque règle $s \in S : \ 0 < conf(s) < k$.

3.2. Algorithme

En bloquant les canaux d'inférences, il est nécessaire de chercher un bon compromis entre le niveau de protection de vie privée et la perte d'utilité résultant de ce processus. Afin de mieux contrôler la perte d'utilité, nous utilisons les métriques suivantes : *Lost Rules Ratio (LR)* et *Ghost Rules Ratio (GR)* [6]. La première se rapporte au pourcentage de règles non-sensibles dans le microdata anonymisé par rapport au total des règles non-sensibles dans les données initiales. Et la seconde se rapporte au pourcentage des règles fantômes dans *D'*, le microdata anonymisé par rapport au total des règles dans *D'*. De fait, si *LR* ou *GR* est plus grande que *h*, un seuil pour apprécier les effets du processus de blocage des canaux d'inférences, le processus de réduction de la confiance est arrêté, et le microdata correspondant est retourné. A noter que la métrique *h* est également choisi arbitrairement.

Algorithm 1 Blocking Inference Channels (BIC)

Require: D, S, k et h ; où D est le jeu de données initial, S l'ensemble des règles de décision sensibles, k est le seuil d'anonymat et h, un seuil pour apprécier les effets de bord de l'algorithme.

Ensure: Diminue la confiance de toutes les règles de décision $s \in S$

1: Soit $s \leftarrow \{X, Y\}$ une règle sensible, où X est l'antécédent et Y la conséquence.
2: Soit $GR \leftarrow 0$ et $LR \leftarrow 0$
3: Répéter jusqu'à (confiance(s) <=k) ou ($GR\ ratio > h$) ou ($LR\ ratio > h$) :
(a) $D' \leftarrow D \uplus \{X, Y'\}$ où Y' représente la valeur manquante et l'opérateur $\uplus$ signifie qu'on ajoute $\{X, Y'\}$ à D.
(b) Calculer $GR \leftarrow (\sim R(D) - \sim R(D'))/\sim R(D)$ où $\sim R(D)$ (respectivement $\sim R(D')$) représente les règles de décision sensibles dans D (respectivement dans D')
(c) Calculer $GR \leftarrow (|R'| - |R \cap R'|)/|R'|$, où $|R'|$ représente le nombre de règles de décision dans D' et $|R \cap R'|$ représente le nombre de règles de décision contenues à fois dans D et D'.
4: **return** D'

4. Validation expérimentale

Nous avons lancé une série d'expériences sur le jeu de données Adult de l'UCI [7]. Nous avons évalué notre proposition avec les dix attributs suivants : education, race, sex, work-class, marital-status, age, sex, relationship, native-country et occupation. L'attribut salary est considéré comme un attribut sensible. Afin de garder l'utilité des données pour une tâche de datamining, nous ne considérons que les attributs age et sexe pour composer les quasi-identifiants. Nous avons supprimé du jeu de données Adult les enregistrements avec des valeurs manquantes et le jeu de données résultant comporte 45222 enregistrements. Nous pouvons résumer notre validation expérimentale en trois étapes :

1) **Le processus de k-anonymization.** Pour anonymiser le jeu de données *Adult*, nous avons particulièrement utilisé la méthode Datafly de l'outil d'anonymisation de l'UT Dallas [8]. L'algorithme Datafly consiste en une généralisation de domaine complet jusqu'à ce que toutes les combinaisons des valeurs de quasi-identifiants apparaissent au moins k fois. Ainsi, le degré d'incertitude de l'attribut sensible dans le jeu de données

anonymisé est au moins égal à $1/k$.

2) **Détection de canaux d'inférences.** Pour construire un arbre de décision sur le jeu de données 10-anonyme obtenu à la première étape, nous utilisons, $J48$ qui est une implémentation de l'algorithme C4.5 dans Weka [9]. Nous déterminons tous les canaux d'inférences comme décrit dans *Définition 1*.

3) **Blocage des canaux d'inférences.** Pour bloquer les canaux d'inférences créés par les règles de decision sensible générées à l'étape précendente, nous avons utilisé l'Algorithm 1. Pour évaluer, la quantité d'information perdue avec cette stratégie de blocage des canaux d'inférences, nous avons lancé *'J48'* à la fois sur le jeu de données initial, le jeu de données 10-anonyme obtenu à l'étape 1, et celui résultant du blocage des canaux d'inférences. Nous avons utilisé 70% des données comme ensemble d'apprentissage et 30% des données comme ensemble test. La Table 2 présente les resultants obtenus.

5. Résultats et Discussion

Les Tables 2, 3 comparent la version originale de k-anonymity avec la version que nous avons proposée en l'occurrence le BIN k-anonymity (Blocking Inference Channels in k-anonymity) suivant deux métriques : la qualité des données (i.e. data quality en anglais) et le niveau de protection de la vie privée. La Table 2 montre que la qualité des données résultant de notre approche est acceptable vue qu'elle est légèrement dégradée. Comme analysé dans [10], la protection de la vie privée à travers l'insertion de fausse information est souvent à l'origine de la dégradation de la qualité des données. Il est évident que plus des changements sont faits dans le jeu de données initial, plus celui-ci reflète moins le domaine d'intérêt. La précision du classifieur est fortement liée à la perte d'information encourue avec le processus de préservation d'anonymat. Notons que peu est la perte d'information, meilleure est la qualité des données.

Jeu de données	Taux de précision
Adult initial	83,23%
Adult 10-anonyme	82,54%
Adult BIC 10-anonyme	82,09%

Figure 1. *Qualité des données*

Bien que notre approche (BIN k-anonymity) occasionne une perte d'utilité, elle améliore le niveau de protection d'anonymat. La Table 3 compare la version originale de k-anonymity avec la notre (i.e. BIC k-anonymity) suivant trois forme d'attaques que nous avons présentées dans la Section 2. Contrairement à version originale, BIC k-anonymity contrôle les inférences basées les règles sensibles. Nous pouvons conclure que notre approche améliore la protection de k-anonymity parce qu'il fournit une meilleure protection d'anonymat avec une légère dégradation de la qualité des données. Comme discuté dans la littérature, maximiser à la fois la protection de la vie privée et la qualité des données n'est pas possible : meilleure est la protection de vie privée est, plus la perte d'utilité est grande, et vis-versa [10].

Modèle	"Record Linkage"	"Attribute Linkage"	Attaques basées sur les Règles Sensibles
k-anonymity	NV	V	V
BIC k-anonymity	NV	V	NV

Figure 2. *Tableau comparatif entre k-anonymity et BIC k-anonymity suivant leur vulnérabilité.*

6. Conclusion

Dans ce papier, nous avons étudié le problème de l'anonymisation des données qui a deux objectifs orthogonaux : la protection de l'anonymat et la préservation de la qualité des données. Afin de trouver un bon compromis entre le niveau de protection d'anonymat et la qualité des données, nous avons proposé une nouvelle approche qui améliore la protection offerte par le modèle k-anonymity tout en gardant l'utilité des données. Nous envisageons dans un futur proche d'étendre notre proposition en limitant les attaques probabilistes et celles de types "attribute linkage".

7. Bibliographie

[1] CIRIANI V. , DE CAPITANI DI VIMERCATI S., FORESTI S., SAMARATI P. : « k-Anonymous Data Mining : A Survey, in Privacy-Preserving Data Mining : Models and Algorithms », Charu C. Aggarwal and Philip S. Yu (eds), Springer-Verlag, 2008.

[2] ATZORI, M., BONCHI, F., GIANNOTTI, F, PEDRESCHI, D, « k-anonymous patterns », 9th European Conference On Principles And Practice Of Knowledge Discovery In Databases (PKDD 2005).

[3] QUINLAN, J.R., « C4.5 : Programs for Machine Learning, Morgan Kaufmann », 1993.

[4] LAROSE, D. T., « Discovering Knowledge in Data ». An Introduction to Data Mining, John Wiley & Sons, Inc., 2005.

[5] LI, N., LI, T., VENKATASUBRAMANIAN, S., « t-Closeness : Privacy Beyond k-Anonymity and l-Diversity ». 2007 IEEE 23rd International Conference on Data Enginering, Istanbul, Turkey, 15-20 April 2007, pp.106-115.

[6] JOHNSTEN, T. , RAGHAVAN, V. V. « A methodology for hiding knowledge in databases ». Proceedings of the IEEE International Conference on Privacy, Security and Data Mining, Maebashi City, Japan, Volume 14, pp. 9-17.

[7] U.C.IRVINE MACHINE LEARNING REPOSITORY, http ://www.ics.uci.edu/mlearn/mlrepository.html.

[8] UTD ANONYMIZATION TOOLBOX, http ://cs.utdallas.edu/dspl/cgi-bin/toolbox/

[9] IAN, H. W., EIBE, F., « Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations ». Morgan Kaufmann, October 1999.

[10] BERTINO, E., FOVINO, I., PROVENZA, I., « A Framework for Evaluating Privacy-Preserving Data Mining Algorithms ».Data Mining and Knowledge Discovery Journal, 11(2), 2005.