Chapitre 1

(titre du chapitre)

1. *K*-ANONYMAT PAR MICRO-AGRÉGATION

Josep DOMINGO-FERRER[1]

1.1 Introduction

Il existe un grand répertoire de méthodes de contrôle statistique de la révélation (CSR) pour protéger des données individuelles (micro-données). La plupart des méthodes CSR sont paramétriques, donc l'utilisateur a affaire à un double choix :

1. Sélection de la méthode
2. Sélection des paramètres

Par conséquent, il faut des orientations pour diminuer l'embarras du choix. Après un survol de quelques alternatives, cette présentation se concentre sur

- Le *k*-anonymat et ses limitations
- Comment réparer le *k*-anonymat par moyen de la micro-agrégation

1.1.1 Plan de cette présentation

[1] Université Rovira i Virgili de Tarragone, Dép. de Génie Informatique et de Mathématiques, Av. Països Catalans 26, E-43007 Tarragona, Catalogne. Courriel josep.domingo@urv.net

On discutera d'abord de l'équilibrage entre la perte d'information et le risque de révélation Ensuite, on examinera le concept de *k*-anonymat et ses problèmes de calcul. Finalement, on décrira comment arriver au *k*-anonymat par le biais de la micro-agrégation.

1.2 Équilibrage entre la perte d'information et le risque de révélation

On trouve dans la littérature trois approches alternatives pour contrôler de façon conjointe la perte d'information et le risque de révélation :

- Construction d'un score
- Cartes R-U
- *k*-Anonymat

1.2.1 Construction d'un score

Le but du CSR est de modifier les données en vue d'obtenir une protection suffisante avec une perte d'information minimale. Donc, une bonne méthode CSR est celle qui donne un bon équilibre entre la perte d'information et le risque de révélation. Domingo-Ferrer et Torra (2001) ont proposé comme score la moyenne entre la perte d'information et le risque de révélation, c'est-à-dire,

$$Score(V,V') = 0.5*IL(V,V') + 0.5*DR(V,V')$$

Dans l'expression précédente, *V* est le jeu de données originales et *V'* le jeu de données protégées. *IL* est calculé comme une combinaison pondérée des variations moyennes des moyennes, variances, covariances et de l'erreur absolue moyenne des corrélations. *DR* est calculé comme une combinaison pondérée des mesures d'appariement d'enregistrements et de révélation par intervalles.

1.2.2 Cartes R-U

Ce type de cartes a été conçu par Duncan *et al.* (2001). Elles sont des ensembles de paires (R,U) du risque de révélation et d'utilité qui correspondent aux différentes stratégies CSR. Ces paires *(R,U)* sont normalement représentées dans un graphe 2D en sorte que l'utilisateur puisse appréhender facilement l'influence d'un certain choix de méthode ou de paramètre.

1.2.3 *k*-Anonymat

Le *k*-anonymat (Samarati et Sweeney, 1998 ; Sweeney, 2002) est une approche différente pour résoudre le conflit entre perte d'information et risque de révélation. Un jeu de données protégées satisfait au *k*-anonymat si, pour n'importe quelle combinaison des valeurs du quasi-identifiant (par exemple, adresse, âge, sexe, etc.), il y a au moins *k* enregistrements partageant cette combinaison. Si *V'* est un jeu *k*-anonyme, un assaillant essayant d'apparier *V'* avec un jeu *V* externe et non-anonyme trouve au moins *k* enregistrements correspondants dans *V'* quelle que soit la valeur utilisée du quasi-identifiant. Si, pour un certain *k*, on suppose que le *k*-anonymat donne assez de protection, on peut se concentrer à minimiser la perte d'information, avec la seule contrainte de satisfaire au *k*-anonymat. Donc celui-ci est une solution élégante et propre du conflit entre protection et utilité.

1.3 Problèmes de calcul du *k*-anonymat

La méthode de calcul suggérée pour atteindre le *k*-anonymat est la combinaison de la généralisation et de la suppression locale. Dans cette méthode, minimiser la perte d'information veut dire minimiser soit le nombre des suppressions, soit la perte de détail causée par les généralisations.

Or le *k*-anonymat avec généralisation et suppression locale minimales est NP-difficile (Meyerson et Williams, 2004 ; Aggarwal *et al.*, 2005). Même la combinaison optimale de généralisation et de suppression locale reste un problème ouvert (une combinaison mal faite peut nuire gravement a l'utilité). En plus, la généralisation pose quelques problèmes d'ordre pratique :

1. Coût de trouver le recodage optimal : pour un attribut avec *c* catégories, il y a 2^c-*c*-*1* généralisations possibles ;
2. Détermination du sous-ensemble des généralisations correctes : quelles sont les nouvelles catégories et quel est le recodage entre les catégories anciennes et nouvelles. Par exemple, si on généralise des codes postaux, le recodage de 08201 et 08205 comme 0820* n'a de sens que si 0820* est significatif du point de vue géographique. Pour la même raison, il est probable que le recodage de 08201 et 05201 comme 0*201 n'ait aucun sens. Donc, la généralisation automatique est loin d'être évidente.

En ce qui concerne spécifiquement la généralisation, pour une certaine règle de généralisation $c_i \rightarrow C$, la littérature n'est pas unanime sur quels

enregistrements contenant c_i faut-il recoder. Dans le logiciel µ-Argus (Hundepool *et al.*, 2005), toutes les occurrences de c_i sont recodées (recodage global). Sweeney (2002) et Samarati (2001) ne recodent que quelques occurrences (recodage local). Les inconvénients du recodage global sont qu'il cause plus de perte d'information et que la règle de généralisation convenant à un groupe d'enregistrements peut ne pas convenir à un autre groupe. D'autre part, les inconvénients du recodage local sont qu'il est difficile à automatiser et qu'il complique l'analyse des données (les catégories anciennes et nouvelles co-existent et, en plus, une catégorie ancienne peut être recodée comme plusieurs catégories nouvelles).

En ce qui concerne la suppression locale, il faut d'abord souligner qu'on ignore comment faut-il la combiner de façon optimale avec la généralisation. En outre, l'utilisation de la suppression locale n'est unanime non plus. Sweeney (2002) supprime des tuples entières, alors que µ-Argus ne supprime que certains attributs pour certains enregistrements. Il y a d'ailleurs un choix entre effacer une valeur supprimée ou bien la remplacer par une valeur neutre (quelque sorte de moyenne). Quel que soit le type de suppression, l'utilisateur a du mal à analyser des données partiellement supprimées ; il lui faut probablement du logiciel spécifique pour traiter des jeux avec des données manquantes.

Finalement, la généralisation/suppression s'adapte mal à quelques types de données. Pour des données catégorielles (nominales ou ordinales), la généralisation/suppression reste raisonnable, même si fort maladroite, comme on vient de le montrer. Pour des données numériques, elle est tout à fait inadéquate : elle transforme l'information numérique continue en information catégorielle, si bien qu'on perd la sémantique numérique sous-jacente.

1.4 *k*-Anonymat par micro-agrégation

La micro-agrégation multivariée (Domingo-Ferrer et Mateo-Sanz, 2002) se révèle une méthode de calcul naturelle pour atteindre le *k*-anonymat. Il s'agit d'une approche unifiée, face à la dualité généralisation-suppression. Même si la micro-agrégation multivariée elle aussi est NP-difficile (Oganian et Domingo-Ferrer, 2001), des heuristiques presque optimales existent, ce qui n'est pas le cas pour la généralisation-suppression. En plus, la micro-agrégation ne complique pas l'analyse des données par l'addition de nouvelles catégories ou par la suppression des données. Enfin, la micro-

agrégation est parfaitement adéquate à protéger tous les types de données : numériques, catégorielles ordinales et catégorielles nominales.

MDAV-générique (Domingo-Ferrer et Torra, 2005) est une variante générique de l'algorithme MDAV (Maximum Distance to Average Vector) disponible dans μ-Argus pour faire de la micro-agrégation. MDAV est à son tour une évolution de la micro-agrégation multivariée à taille de groupe fixée proposée dans Domingo-Ferrer et Mateo-Sanz (2002). En changeant l'opérateur moyenne et la distance utilisés, MDAV-générique peut micro-agréger, c'est-à-dire, *k*-anonymiser des données numériques, ordinales ou nominales. On donne ensuite des détails suivant le type des données.

Pour le cas des attributs numériques, on utilise comme opérateur moyenne la moyenne arithmétique et comme distance la distance euclidienne. Avec ces choix, MDAV-générique correspond a l'algorithme MDAV disponible dans μ-Argus. Avant d'appliquer MDAV-générique, on standardise les attributs du jeu de données, afin qu'ils aient tous le même poids quand on calcule des distances. Après MDAV-générique, on dé-standardise pour recupérer les échelles originales des attributs. Par construction, le jeu micro-agregé est *k*-anonyme et préserve les moyennes des attributs originaux. Il suffit d'un simple changement d'échelle décrit dans Domingo-Ferrer et Torra (2005) pour que les attributs du jeu micro-agregé et *k*-anonyme préservent en plus les variances des attributs originaux.

Pour le cas des attributs ordinaux, les opérateurs moyenne utilisés peuvent être soit la médiane, soit la médiane convexe. Prenons un exemple simple de catégories numérotées avec des entiers dans l'échelle 1-7 : la médiane de l'ensemble {1,2,7} est 2, tandis que la médiane convexe est 4, ce qui se trouve plus près de la moyenne arithmétique 3.3. La médiane convexe permet une certaine compensation : la catégorie « eloignée » 7 pèse plus qu'elle ne peserait avec l'opérateur médiane (avec ce dernier, 7 pèse comme 3). Comme distance ordinale pour MDAV-générique entre deux catégories a et b d'un attribut, telles que $a < b$, on peut prendre le nombre des catégories $\geq a$ et $< b$ divisé par le nombre total des catégories dans l'échelle ordinale pour cet attribut. Cela donne une distance dans l'intervalle [0,1).

Enfin, pour des attributs nominaux, l'opérateur moyenne à utiliser avec MDAV-générique est la règle de pluralité, c'est-à-dire la mode. Quant à la distance entre deux catégories d'un attribut nominal, on dit que la distance est 0 si les deux catégories sont identiques et 1 si elles sont différentes.

BIBLIOGRAPHIE

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D. et Zhu, A. (2005). «Anonymizing tables» dans T. Eiter et L. Libkin (eds.), Proceedings of ICDT'2005, vol. LNCS 3363, pp. 246-258, Springer-Verlag, Berlin.

Domingo-Ferrer, J., et Mateo-Sanz, J. M. (2002). «Practical data-oriented microaggregation for statistical disclosure control», IEEE Transactions on Knowledge and Data Engineering, vol. 14, No. 1, pp. 189-201.

Domingo-Ferrer, J., et Torra, V. (2001). «A quantitative comparison of disclosure control methods for microdata», dans *Confidentiality, Disclosure and Data Access*, pp. 111-134, North-Holland, Amsterdam.

Domingo-Ferrer, J., et Torra, V. (2005). «Ordinal, continuous and heterogeneous *k*-anonymity through microaggregation», Data Mining and Knowledge Discovery, à paraître.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R. et Roehrig, S. F. (2001). «Disclosure limitation methods and information loss for tabular data», dans *Confidentiality, Disclosure and Data Access*, pp. 135-166, North-Holland, Amsterdam.

Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R., et Giessing, S. (2005). *μ-ARGUS version 4.0.2. Software and User's Manual*, Statistics Netherlands, Voorburg NL. http://neon.vb.cbs.nl/casc

Meyerson, A. et Williams, R. (2004). «On the complexity of optimal *k*-anonymity», dans *PODS'2004*, pp. 223-228, Paris.

Oganian, A. et Domingo-Ferrer, J. (2001). «On the complexity of optimal microaggregation for statistical disclosure control», Statistical Journal of the UNECE, vol. 18, No. 4, pp. 345-354.

Samarati, P. (2001). «Protecting respondents' identities in microdata release», IEEE Transactions on Knowledge and Data Engineering, vol. 13, No. 6, pp. 1010-1027.

Samarati, P. et Sweeney, L. (1998). «Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression», SRI International Technical Report.

Sweeney, L. (2002). «Achieving k-anonymity privacy protection using generalization and suppression», International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, No. 5, pp. 571-588.