Aprendizaje supervisado para el enlace de registros a través de la media ponderada

Daniel Abril
Instituto de Investigación
en Inteligencia Artificial (IIIA),
Consejo Superior de
Investigaciones Científicas (CSIC)
Email: dabril@iiia.csic.es

Guillermo Navarro-Arribas
Dep. Enginyeria de la Informació
i de les Comunicacions (DEIC),
Universitat Autònoma
de Barcelona (UAB)
Email: guillermo.navarro@uab.cat

Vicenç Torra
Instituto de Investigación
en Inteligencia Artificial (IIIA),
Consejo Superior de
Investigaciones Científicas (CSIC)
Email: vtorra@iiia.csic.es

***Resumen*—En el área de la privacidad de datos, las técnicas para el enlace de registros son utilizadas para evaluar el riesgo de revelación de un conjunto de datos protegido. La idea principal detrás de estas técnicas es enlazar registros que hacen referencia a un mismo individuo, entre diferentes bases de datos. En este trabajo se presenta una variación del enlace de registros basada en una media ponderada para calcular distancias entre registros. Mediante el uso de un método supervisado de aprendizaje nuestra propuesta permite determinar cuales son los pesos que maximizan el número de enlaces entre los registros de la base de datos original y su versión protegida. El resultado de este trabajo se aplica en la estimación del riesgo de revelación de datos protegidos.**

***Palabras clave*—enlace de registros (record linkage), privacidad de datos (data privacy), riesgo de divulgación (disclosure risk).**

I. Introducción

El enlace de registros consiste en el proceso de identificación de forma rápida y precisa de dos o más registros distribuidos en varias bases de datos (o fuentes de información en general) los cuales hacen referencia a la misma entidad o individuo. Esta técnica fue inicialmente introducida por Dunn [1] en el campo de la sanidad pública, con el fin de crear historiales médicos completos mediante el enlace de toda la información recopilada sobre un paciente, la cual estaba repartida por varias bases de datos. Estos enlaces fueron posibles gracias a la utilización de campos clave como el nombre o la fecha de nacimiento, entre otros. En los siguientes años, esta idea fue mejorada y matemáticamente formalizada [2], [3], [4]. Hoy en día se ha convertido en una popular técnica utilizada por agencias estadísticas, comunidades de investigación y otras instituciones, no solo para la integración de bases de datos [5], [6], sino también para la limpieza de datos [7] o el control de la calidad de los datos [8]. Un claro ejemplo de la utilización de estos métodos es para la detección de registros duplicados entre diferentes conjuntos de datos [9].

Debido a la necesidad de distintas agencias gubernamentales u otras instituciones de coleccionar y analizar grandes cantidades de datos confidenciales, las técnicas de enlace de registros fueron recientemente introducidas en el área de la privacidad de datos. Esta área de investigación proporciona métodos de seguridad para las bases de datos estadísticos con el fin de combatir la revelación de información confidencial contenida en dichas bases de datos. *Privacy Preserving Data Mining (PPDM)* [10] y *Statistical Disclosure Control (SDC)* [11] son dos disciplinas cuya función es la investigación de métodos y herramientas para asegurar la privacidad de estos datos. Dentro de estos campos el enlace de registros se utiliza para obtener una evaluación del riesgo de revelación de información confidencial sobre un conjunto de datos previamente protegido [12], [14]. Por lo tanto el riesgo de re-identificación de un individuo se evalúa mediante la identificación de enlaces de registros pertenecientes al mismo individuo entre los datos protegidos y originales. En [13] los autores definen un método general de evaluación de un conjunto de datos protegido basado en la combinación de diferentes medidas analíticas, cuyo objetivo es evaluar el riesgo de revelación de información confidencial y la evaluación de la cantidad de información perdida en el proceso de protección.

En este artículo se introduce un nuevo método de evaluación del riesgo de revelación con el fin de mejorar la precisión de las técnicas actualmente utilizadas. Este consiste en la utilización de una media ponderada como distancia en el proceso de enlace de registros y un algoritmo de aprendizaje supervisado, de manera que el algoritmo de aprendizaje aprenda cuales son los pesos que maximizan el número de enlaces entre el conjunto de datos original y el protegido.

La organización de este artículo es la siguiente. En la Sección II, se presentan algunos conceptos básicos necesarios para el resto de las secciones. En la Sección III se describe el método de aprendizaje supervisado para el enlace de registros cuando se utiliza la media ponderada. La evaluación del método presentado se introduce en la Sección IV. Finalmente, la Sección V presenta las conclusiones del trabajo realizado.

II. Enlace de registros en la privacidad de datos

En esta sección se presentan algunas ideas y definiciones básicas para comprender el uso de las técnicas de enlace de registros en el campo de la privacidad de datos.

PPDM y *SDC* están orientadas a trabajar sobre bases de datos, principalmente tablas o ficheros (*microdata*). Estas bases de datos pueden verse como una matriz, X, de N filas (registros) y n columnas (atributos), donde cada fila

ISBN: 978-84-9717-323-0

corresponde a un único individuo (o entidad). En este contexto se pueden diferenciar dos tipos diferentes de atributos:

- *Identificadores*: son aquellos atributos que pueden identificar directamente un individuo, como por ejemplo el número de identificación nacional (DNI) o el número de cuenta bancaria.
- *Casi-identificadores*: son aquellos atributos que por ellos mismos no son capaces de identificar un único individuo, sin embargo, cuando dos o más de ellos son combinados, pueden identificar inequívocamente un individuo. Estos atributos se pueden dividir a su vez en dos tipos, los *confidenciales* (X_c) y los *no confidenciales* (X_{nc}), dependiendo del tipo de información que representen. Un ejemplo de atributo no confidencial sería el código postal, mientras que un ejemplo de atributo confidencial podría ser el salario.

Previa publicación de un conjunto de datos X, es necesaria su protección mediante la aplicación de un método ρ, el cual dará lugar a un conjunto de datos protegidos Y. Estos métodos de protección solo protegerán los atributos considerados como casi-identificadores no confidenciales, $Y_{nc} = \rho(X_{nc})$, ya que para asegurar la privacidad de los individuos los identificadores son eliminados y/o cifrados. Los casi-identificadores confidenciales no son modificados ni eliminados debido a su interés de análisis, por lo tanto quedarán intactos. De este modo, podemos ver el conjunto de datos protegido como $Y = \rho(X_{nc})||X_c$. Este escenario fue presentado en [13] y posteriormente utilizado en otros trabajos como [14].

En el campo de la privacidad de datos, el enlace de registros es utilizado para re-identificar individuos entre la base de datos protegida y la base de datos original, es decir, se usa como una medida de evaluación del riesgo de revelación. Actualmente existen dos enfoques diferentes del enlace de registros para la evaluación del riesgo. El *enlace de registros probabilístico (PRL)* [16] y el *enlace de registros basado en distancias (DBRL)* [17].

En el trabajo realizado en este artículo se utiliza el segundo tipo, el enlace de registros basado en distancias, el cual es explicado con detalle a continuación.

II-A. Enlace de registros basado en distancias

La idea principal del enlace de registros basado en distancias es la definición de una función de distancia. Como es sabido, dependiendo de la distancia utilizada se pueden obtener resultados completamente diferentes. A continuación se revisan dos de las distancias más utilizadas y testeadas en la literatura del DBRL, la distancia Euclídea y la distancia de Mahalanobis.

En este artículo usaremos $V_1^X, \ldots, V_n^X$ y $V_1^Y, \ldots, V_n^Y$ para indicar el conjunto de atributos de los ficheros X e Y, respectivamente. Usando esta notación, podemos expresar los valores de cada atributo de un registro $a \in X$ como $a = (V_1^X(a), \ldots, V_n^X(a))$ y un registro $b \in Y$ como $b = (V_1^Y(b), \ldots, V_n^Y(b))$. Además, indicaremos la media de los valores de un atributo V_i^X como $\overline{V_i^X}$.

- La *distancia Euclídea* es usada para bases de datos con atributos estandarizados. La distancia entre dos registros a y b se define como:

$$d(a,b)^2 = \sum_{i=1}^{n} \left(\frac{V_i^X(a) - \overline{V_i^X}}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V_i^Y}}{\sigma(V_i^Y)} \right)^2 \quad (1)$$

- La *distancia de Mahalanobis* se define como:

$$dMD(a,b)^2 = (a-b)'\Sigma^{-1}(a-b)$$

donde, $\Sigma = [Var(V^X) + Var(V^Y) - 2Cov(V^X, V^Y]$ y $Var(V^X)$ es la varianza de los atributos V^X, $Var(V^Y)$ es la varianza de los atributos V^Y y $Cov(V^X, V^Y)$ es la covarianza entre los atributos V^X y V^Y. Si la matriz de covarianza es una matriz identidad, entonces la distancia de Mahalanobis se reduce a la distancia Euclídea.

III. Aprendizaje supervisado para el enlace de registros

En esta sección se presenta el método de aprendizaje supervisado, el cual se usa junto a una distancia ponderada, para determinar cuales son los pesos de esta distancia que maximizan el número de re-identificaciones entre los registros del fichero original y protegido. En la Sección III-A se introduce la distancia usada, la media ponderada, mientras que en la Sección III-B se introduce el problema de aprendizaje como un problema de optimización en base a la media ponderada.

III-A. Distancia ponderadas

Es bien conocido que la multiplicación de la distancia Euclídea por una constante no altera los resultados de ningún algoritmo de enlace de registros. De modo que se puede expresar la Ecuación 1 como una media ponderada. Esta se pude formalizar como:

$$d(a,b)^2 = \sum_{i=1}^{n} \frac{1}{n} \left(\frac{V_i^X(a) - \overline{V^X}_i(a)}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V^Y}_i(b)}{\sigma(V_i^Y)} \right)^2$$

La cual, si definimos para cada atributo:

$$d_i(a,b)^2 = \left(\frac{V_i^X(a) - \overline{V^X}_i(a)}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V^Y}_i(b)}{\sigma(V_i^Y)} \right)^2$$

Esta expresión se pude redefinir como,

$$d(a,b)^2 = AM(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

donde AM es la media aritmética

$$AM(c_1, \ldots, c_n) = \sum_i c_i/n$$

.

En general, cualquier operación de agregación $\mathbb{C}$ [18] puede utilizarse: $d(a,b)^2 = \mathbb{C}(d_1(a,b)^2, \ldots, d_n(a,b)^2)$.

De esta definición, es trivial la consideración de diferentes operadores de agregación, como por ejemplo la media ponderada.

Definición 1 *Considerando $p = (p_1, \ldots, p_n)$ como un vector de pesos, es decir, $p_i \geq 0$ y $\sum_i p_i = 1$. Entonces, la distancia ponderada se define como:*

$$d^2WM_p(a,b) = WM_p(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

donde $WM_p(c_1, \ldots, c_n) = \sum_i p_i \cdot c_i$.

III-B. Aprendizaje supervisado

Para simplificar la formalización del proceso, se asume que cada registro (fila) b_i de Y es la versión protegida de a_i de X. Es decir, los dos conjuntos de datos están alineados. Entonces, dos registros están correctamente enlazados usando un operador de agregación $\mathbb{C}$ cuando la agregación de los valores a_i y b_i es más pequeña que para a_i y b_j para todo $i \neq j$. Formalmente, a_i se considera correctamente enlazado con b_i cuando se satisface la siguiente ecuación para todo $i \neq j$.

$$\mathbb{C}(a_i, b_i) < \mathbb{C}(a_i, b_j) \tag{2}$$

En condiciones óptimas estas desiguales las satisfacen todos los registros a_i. No obstante, en general no se pude esperar que se cumpla debido a los errores introducidos en el conjunto de datos por el algoritmo de protección. Por lo tanto, el proceso de aprendizaje es formalizado como un problema de optimización.

La Ecuación (2) debe ser relajada de manera que la solución pueda violar alguna de las restricciones. Esta relajación es formalizada creando el concepto de bloque. Un bloque es el conjunto de ecuaciones referentes a un registro a_i, es decir, el conjunto de todas las distancias entre un registro original y todos los registros protegidos. Así, podemos asignar a cada bloque una variable K_i, teniendo tantas variables como número de registros. Además, hemos considerado para la formalización una constante C que multiplica K_i para superar las inconsistencias y satisfacer las restricciones. La idea de este enfoque es que cada variable K_i indique, por cada bloque, si todas las restricciones correspondientes se satisfacen ($K_i = 0$) o si por el contrario, no se satisfacen ($K_i = 1$). Así, si un registro a_i no cumple la Ecuación (2) por algún registro b_j, no importa que otro registro b_k ($k \neq j$) también viole la ecuación para el mismo a_i. Teniendo en cuenta esta asunción, el objetivo del problema será minimizar el número de bloques que no cumplen sus restricciones. De este modo, podremos encontrar los pesos que minimizan el número de violaciones, o en otras palabras, los pesos que maximizan el número de re-identificaciones entre los dos conjuntos de datos.

Utilizando esta notación, tenemos que la siguiente restricción tiene que satisfacerse para todos los pares $i \neq j$.

$$\mathbb{C}(a_i, b_j) - \mathbb{C}(a_i, b_i) + CK_i > 0.$$

Como $K_i = \{0, 1\}$, se pude usar C como una constante, la cual expresa la *mínima distancia* requerida entre el enlace correcto y el resto. Cuanto más grande es el valor, más enlaces correctos se distinguen de los incorrectos.

Utilizando estas restricciones anteriores y el operador de agregación presentado en la Definición 1, d^2WM_p, se puede definir el siguiente problema de optimización:

$$\text{Minimize: } \sum_{i=1}^{N} K_i \tag{3}$$

Subject to:

$$d^2WM_p(a_i, b_j) - d^2WM_p(a_i, b_i) + CK_i > 0, \ \forall i, j = 1, \ldots, N, i \neq j \tag{4}$$

$$K_i \in \{0, 1\} \tag{5}$$

$$\sum_{i=1}^{n} = 1 \tag{6}$$

$$p_i \geq 1 \tag{7}$$

Como se puede observar, este es un problema de optimización con una función objetivo (Ecuación (4)) y unas restricciones (Ecuación (5)) lineales. Debido al operador de agregación usado, se han tenido que añadir un par de restricciones al problema. Las Ecuaciones (6) y (7) hacen referencia a las restricciones introducidas por el uso de pesos de la media ponderada.

Teniendo en cuenta que N es el número de registros y n el número de variables de los dos conjuntos de datos X y Y, se puede calcular fácilmente el número total de restricciones del problema. N^2 restricciones de la Ecuación (4), N son las restricciones necesarias para la Ecuación (5), 1 restricción para la suma de todos los pesos, Ecuación (6), y finalmente n restricciones de la Ecuación (7). Por lo tanto el problema tiene un total de $N^2 + N + n + 1$ restricciones.

IV. Análisis experimental

El método presentado en la sección anterior ha sido evaluado utilizando diferentes conjuntos de datos protegidos. Para la protección de datos se ha utilizado la *Microaggregation* [5], un método muy conocido para la protección de microdatos. Este método proporciona privacidad mediante la agrupación de los datos en pequeños grupos de k elementos, y posteriormente reemplazando los datos originales de cada agrupación por su correspondiente centroide. El parámetro k determina el grado de protección aplicado: cuanto mayor es el valor de k, mayor es la protección aplicada y a su vez mayor es la información perdida en el proceso.

Se han considerando los conjuntos de datos con los siguientes parámetros de protección:

- *M4-33*: 4 atributos microaggregados en grupos de 2 con $k = 3$.
- *M4-28*: 4 atributos, los primeros 2 atributos con $k = 2$, y los últimos 2 con $k = 8$.
- *M4-82*: 4 atributos, los primeros 2 atributos con $k = 8$, y los últimos 2 con $k = 2$.
- *M5-38*: 5 atributos, los primeros 3 atributos con $k = 3$, y los últimos 2 con $k = 8$.
- *M6-385*: 6 atributos, los primeros 2 atributos con $k = 3$, los siguientes 2 atributos con $k = 8$, y los últimos 2 con $k = 5$.
- *M6-853*: 6 atributos, los primeros 2 atributos con $k = 8$, los siguientes 2 atributos con $k = 5$, y los últimos 2 con $k = 3$.

Para cada uno de los conjuntos de datos se han protegido 400 registros aleatoriamente extraídos del censo [19] del proyecto *European CASC* [20], el cual contiene 1080 registros y 13 atributos, y ha sido extensamente usado en otras investigaciones como [21], [22], [23]. Los grados de protección, k, aplicados, varían entre la mínima protección posible, $k = 2$ y un buen grado de protección, $k = 8$, según [13].

	d^2AM	d^2MD	d^2WM
M4-33	0,84	0,94	0,955
M4-28	0,685	0,9	0,93
M4-82	0,71	0,9275	0,9425
M5-38	0,3975	0,8825	0,905
M6-385	0,78	0,985	0,9925
M6-853	0,8475	0,98	0,9875

Tabla I
RESULTADOS EN EL ENLACE DE REGISTROS BASADO EN DISTANCIAS.

En la Tabla I se muestran los resultados de aplicar los métodos estándar de enlace de registros basado en distancias, d^2AM y d^2MD, y el nuevo método supervisado basado en la media ponderada, d^2WM, presentado en la Sección III. Los valores de dicha tabla son el ratio de registros correctamente re-identificados, de modo que 1 significa que el 100 % de las re-identificaciones fueron correctas.

Como se puede apreciar, el método presentado obtiene un incremento relevante en el número de re-identificaciones cuando es comparado con los métodos estándar actualmente utilizados. Este incremento es especialmente importante al comparar d^2WM con la distancia Euclídea, en el cual vemos un incremento de hasta un 50 % para el conjunto *M5-38*.

V. CONCLUSIÓN

En este artículo se ha introducido una variante de enlace de registros basado en distancias. Nuestra propuesta utiliza un algoritmo de aprendizaje supervisado el cual gracias a una media ponderada permite determinar los pesos de esta que maximizan el número de re-identificaciones entre el fichero original y el protegido. De este modo, se han mejorado los sistemas estándar de evaluación del riesgo de un fichero protegido.

Este y otros trabajos en la misma línea de investigación, el enlace de registros basado en distancias, se pueden encontrar en los siguientes artículos [24], [25].

AGRADECIMIENTOS

Esta investigación está parcialmente financiada por el MICINN (proyectos ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, TIN2010-15764 y TIN2011-27076-C03-03) y por by the EC (FP7/2007-2013) Data without Boundaries (número de subvención 262608). Algunos de los resultados presentados en este artículo han sido obtenidos gracias al Centro de Supercomputación de Galicia (CESGA). El trabajo contribuido por el primer autor ha sido parte de un programa de doctorado en Informática de la Universidad Autónoma de Barcelona (UAB).

REFERENCIAS

[1] Dunn, H.L. (1946). Record Linkage. American Journal of Public Health 36 (12), pp. 1412–1416.
[2] Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959) Automatic Linkage of Vital Records, Science, 130, pp. 954–959.
[3] Newcombe, H. B., J. M. Kennedy (1962). Record linkage: making maximum use of the discriminating power of identifying information. Commun. ACM 5, 11, pp. 563-566.
[4] Fellegi, I., Sunter, A., (1969). A Theory for Record Linkage. Journal of the American Statistical Association 64 (328), pp. 1183¿ –1210.
[5] Defays, D., Nanopoulos, P. (1993) Panels of enterprises and confidentiality: The small aggregates method, Proc. of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada, pp. 195–204.
[6] Statistics Canada. (2010). Record linkage at Statistics Canada. http://http://www.statcan.gc.ca/record-enregistrement/index-eng.htm
[7] Winkler, W. E. (2003) Data cleaning methods, Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[8] Batini, C., Scannapieco, M. (2006) Data Quality - Concepts, Methodologies and Techniques Series: Data-Centric Systems and Applications.
[9] Elmagarmid, A., Panagiotis G., Verykios, V., (2007). Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering 19 (1), pp. 1–16.
[10] Agrawal, R., Srikant, R. (2000) Privacy-preserving data mining. ACM Sigmod Record, vol. 29, issue 2, pp. 439-450.
[11] Willenborg, L., De Waal, T. (2001) Elements of Statistical Disclosure Control. Springer Verlag.
[12] Spruill, N.L., (1982) Measures of confidentiality. Proc. Survey Research Section American Statistical Association, pp. 260–265.
[13] Domingo-Ferrer, J., Torra, V., (2001) A quantitative comparison of disclosure control methods for microdata, pp. 111–133 of [15].
[14] Winkler, W.E, (2004) Re-identification methods for masked microdata, Privacy in Statistical Databases 2004, Lecture Notes in Computer Science 3050, pp. 216–230.
[15] Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (editors) (2001) Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies, Elsevier Science.
[16] Jaro, M. A. (1989) Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Society, 84:406, pp. 414–420.
[17] Pagliuca, D., Seri, G., (1999), Some results of individual ranking method on the system of enterprise accounts annual survey, Esprit SDC Project, Delivrable MI-3/D2.
[18] Torra, V., Narukawa, Y., (2007) Modeling decisions: information fusion and aggregation operators, Springer.
[19] U.S. Census Bureau. Data Extraction System, 2011, http://www.census.gov/.
[20] Brand, R., Domingo-Ferrer, J., Mateo-Sanz, J.M., (2002) Reference datasets to test and compare SDC methods for protection of numerical microdata. Technical report, European Project IST-2000-25069 CASC.
[21] Laszlo, M., Mukherjee, S., (2005) Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering 17(7), pp. 902–911.
[22] Domingo-Ferrer, J., Torra, V., (2005) Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation, Data Mining and Knowledge Discovery 11(2), pp. 195–212.
[23] Yancey, W., Winkler, W.E., Creecy, R., (2002) Disclosure risk assessment in perturbative microdata protection. In Inference Control in Statistical Databases, Lecture Notes in Computer Science 2316, pp. 135–152.
[24] Abril, D., Navarro-Arribas, G., Torra, V., (2012). Improving record linkage with supervised learning for disclosure risk assessment. Information Fusion, 13(4), 274–284.
[25] Abril, D., Navarro-Arribas, G., Torra, V., (2012). Choquet integral for record linkage. Annals of Operations Research, 195(1), 97–110.