Fusión de registros

Gonzalo Rivero

Big Data en investigación social y opinión pública

22 a 26 de julio, 2019

Contexto del problema

- Las organizaciones almacen cada vez más datos que pueden ser útiles si son puestos en relación
- Fusionar las fuentes es fácil si existe un identificador único como Pasaporte, ISBN, o códigos de producto
- Si los identificadores únicos no existen, podemos usar características de la entidad como apellidos, fecha de nacimiento o género.
- Necesitamos herramientas para estimar probabilidad de que dos registros se refieran a la misma entidad.

Fusión de registros

- Varias formas de hacerlo mediante reglas de decisión y métodos de comparación
- Metodos deterministicos no permiten cuantificar probabilidad de error
- Usado para enlazar, enriquecer o limpiar/deduplicar bases de datos.
  - Deduplicar datos del censo de población
  - Detectar fraude bancario
  - Fusionar bases de datos para estudios de mercado

El marco Fellegi-Sunter

- Marco estadístico proveído por Fellegi-Sunter.
- Plantea el problema como uno de datos perdidos: la comparación es conocida pero no si el par pertenece a la misma entidad.
- Clasifica pares de registros en *links*, *non-links* y *possible links*.
  - Genera reglas de asociación que relacionan pares con probabilidad de tipo de asociación.
  - Permite estimar número de errores en la clasificación
- Puede aprovechar que algunos registros son menos frecuentes para extraer información. Apellidos poco frecuentes son más útiles para decidir si es un link que apellidos más frecuentes.

Proceso de generación de registros

- Si todos los atributos de dos registros son idénticos es probable que pertenezcan a la misma entidad.
- El proceso de generación de registros está sujeto a errores. Los atributos pueden no ser idénticos y referirse a la misma persona.
  - Marie puede representarse como Mary
  - Pueden ser la misma persona si encajan en los otros atributos (edad, dirección)
- El objetivo es decidir cuántos atributos y en qué medida tienen que coincidir para que un par se refiera a la misma entidad

Secuencia de tareas

1. Preparar las bases de datos para que estén en un formato comparable
   - Las dos bases son la misma en una tarea de desduplicación
2. Crear todos los pares de registros viables
   - Todos los pares posibles pueden ser demasiados
   - Indexado limita el número de comparaciones
3. Comparar los pares de registros para comprobar la similitud de la información entre atributos
4. Decidir si los pares son links, non-links, o possible links
5. Evaluar manualmente los manual links

Preparación de los datos

- El objetivo es simplificar la comparación entre las dos bases de datos.
- Eliminar caracteres especiales como signos de puntuación
  - J. A. Domínguez ⇒ J A Domínguez
- Transformar de mayúscula a minúscula
  - J A Domínguez ⇒ j a domínguez
- (Transliterar a ASCII)
  - j a domínguez ⇒ j a dominguez
- Expandir convenciones o abreviaciones obvias
  - C/ Segunda ⇒ Calle Segunda

Una precaución sobre valores perdidos

- No es obvio que hacer con casos en los que un campo no existe
- Borrar no es buena idea si la inexistencia del campo puede ser informativa. Por ejemplo, número de planta en una dirección
- Imputar es factible pero no siempre sencillo
- Es posible modificar el proceso para que la ausencia sea informativa

Indexado

- Número de pares a crear crece de forma cuadrática y comparar todos puede no ser factible
- Podemos reducir el número de pares mediante índices que limitan el número de candidatos. Equivalente a asumir que no son la misma entidad
- Objetivo es reducir número de pares manteniendo los que contienen los auténticos enlaces
- Crear bloques usando campos idénticos es la opción más habitual
  - Año de nacimiento para emparejar personas
  - Código postal para emparejar direcciones

Método de vecinos ordenados

- Si la variable de bloque no es idéntica, crear bloque basado en cercanía
- Dividir cada cadena de texto en k subcadenas de longitud q llamadas *q-grams*. Es habitual coger un número pequeño
  - Jonathan $\Rightarrow$ ['jo','on', 'na', 'at', 'th', 'ha', 'an'] con $q = 2$
- Proximidad es proporción de q-grams en común. Podemos fijar un umbral que defina si es cercano o no.

Creación de bloques

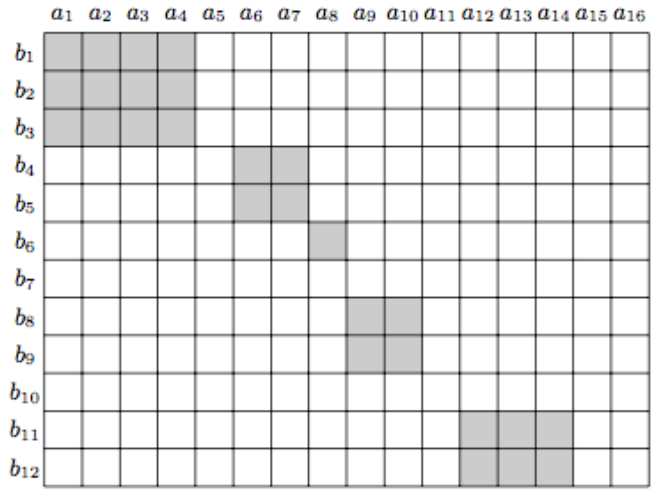

Figure 2-2: Standard indexing (or blocking) on sorted data

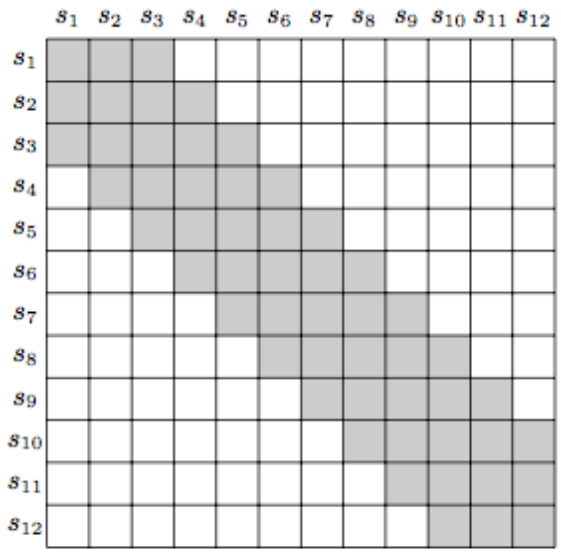

Figure 2-3: Sorted Neighbourhood Indexing with $w = 3$. The sorting key values are $s_1, \ldots, s_{12}$.

Comparar registros

- Comparar números o variables categóricas es sencillo aunque es específico al problema.
  - Imponer máxima diferencia entre variables numéricas
  - Determinar como acuerdo solo si las dos variables son idénticas
- Podemos tratar fechas como variable categórica o numérica
- Compara cadenas de texto tiene más flexibilidad
  - Métodos basados en fonética como Soundex
    - Marie y Mary en inglés tienen código Soundex M600
    - Específicos a cada lenguaje
    - Asumen un tipo de proceso de generación de registros
  - Métodos basados en métricas entre cadenas de texto
    - Medir distancia entre “mary” y “marie”

Métricas en cadenas de texto

Truncación	Comparar si las dos cadenas son idénticas hasta un determinado carácter o a partir de un determinado carácter
Mayor subcadena común	Longitud de la mayor subcadena que las dos cadenas tienen en común
Levenshtein	Distancia es el número mínimo de cambios necesarios para ir de una cadena a la otra. Transformaciones admisibles son sustituciones, inserciones y eliminaciones
Q-gram	Proporción de q-grams en común entre las dos cadenas, divido por el número de q-grams de la cadena de menor longitud.
Jaro-Winkler	Métrica comparación entre q-gram y Levenshtein desarollada para comparar nombres y apellidos.

Fusión determinística

- Procedimiento más intuitivo. Sumar similitudes y usar suma como peso.
- Definir umbral de disimilitud.
- Un par es clasificado como link si la similitud es superior al umbral
- Uso de reglas para comparar dos combinaciones de similitudes

El modelo Fellegi-Sunter

- Cada par de registros puede ser clasificado en un *linked set* o en un *non-linked set* según si representan la misma entidad o no. El auténtico estado de cada par lo representaremos con M.
- Para cada par crearemos un vector de comparación y en K campos que pueden estar en concordancia o no.
- Para cada par tomaremos una *acción*: *I* si fusionamos los dos registros, *III* si no lo hacemos y *II* si no podemos decidir que generan una regla $d(y)$.

Errores de clasificación

- Probabilidad de encontrar un vector de comparación dado que el par corresponde a una entidad $m = \Pr(Y = y|M = 1)$. Si el par no corresponde a una entidad, la probabilidad es $u = \Pr(Y = y|M = 0)$.
- Una regla de asociación entre pares tiene niveles de error $E[d_1(Y)|M = 0] = \mu$ and $E[d_3(Y)|M = 1] = \lambda$.
- Regla óptima es la que minimiza la probabilidad de acción // dados unos niveles de error.

Reglas óptimas de clasificación

- Ordenamos las comparaciones con la razón de verosimilitudes $m(y)/u(y)$ de mayor a menor. La razón m/y es muestra de la evidencia en favor de fusionar dos registros.
- Usamos los niveles de error μ y λ para dividir la sequencia de m/y en tres regiones

$$\sum_{i=1}^{n-1} u(y_i) < \quad \mu \quad \leq \sum_{i=1}^{n} u(y_i) \tag{1}$$

$$\sum_{i=n'}^{N} m(y_i) \geq \quad \lambda \quad > \sum_{i=n'+1}^{N} m(y_i) \tag{2}$$

  siempre que $1 < n' < n - 1 < N$

- De 1 a $n-1$ aplicamos acción *I* de $n'+1$ a N aplicamos acción *III*. Los demás son asignados a *II*.

Simplificaciones para la estimación

- *Independencia condicional* de los componentes de la comparación nos permite factorizar $m(y) = \prod_{i=1}^{K} m_i(y^i)$ y $u(y) = \prod_{i=1}^{K} u_i(y^i)$
  - Reduce el número de parámetros. Con 10 atributos $\Rightarrow$ $2^{10} = 1024$ comparaciones. Independencia condicional $\Rightarrow$ tenemos 20.
- *Comparaciones binarias* en cada componente y^i
- Nos permiten expresar

$$m(y) = \prod_{i=1}^{K} m_i(1)^{y^i} m_i(0)^{1-y^i} \tag{3}$$

$$u(y) = \prod_{i=1}^{K} u_i(1)^{y^i} u_i(0)^{1-y^i} \tag{4}$$

Estimación de los parámetros

- Es habitual definir el *peso* en logaritmos

$$w_i(y^i) = \log\left(\frac{m_i(y^i)}{u_i(y^i)}\right) = \log m_i(y^i) - \log u_i(y^i) \qquad (5)$$

  de tal manera que pesos positivos *y grandes* son evidencia a favor de fusionar y $w(y) = \sum w_i(y^i)$.
- Si $u(y) = 0$ entonces $w(y) = \infty$ y la comparación y solo ocurre en los registros enlazados. Si $m(y) = 0$ entonces $w(y) = \infty$ y y solo ocurre en registros no enlazados.

Estimación de los parámetros

- Estamos interesados en la razón

$$\frac{\Pr(M=1|Y=y)}{\Pr(M=0|Y=y)} = \frac{m(y)}{u(y)}\frac{\pi}{1-\pi} \tag{6}$$

  con $\pi = \Pr(M=1)$ siendo prevalencia de enlaces.
- m, u y π son desconocidos.
- Estimación mediante algoritmo EM

Cómo usar los parámetros

- Ordenamos según w
- Fijar λ y μ