Matching de datos históricos usando algoritmos de clasificación

Big data and Machine learning for Economics

Alison Gissell Ruiz Ruiz - Código 202116230
John Daniel Delgado Vargas - Código 202225721
José Julián Parra Montoya - Código 202213144

Resumen

Seguir *outcomes* individuales durante un largo período de tiempo permite responder preguntas que son del interés actual de la disciplina. No obstante, la creación de estos datos para uso moderno, riñe con los objetivos con los cuales fueron creados: fuentes de información aisladas sin ninguna pretensión de estar vinculadas. Esto puede ser superado mediante el uso de un algoritmo de clasificación correctamente especificado. En este artículo se encuentra que el mejor algoritmo de clasificación en un contexto donde las fuentes de información tienen datos faltantes es un algoritmo tipo *Probit* o *Logit*, los cuales permite obtener un *True Positive Rate* del 85 % y un *Positive Predictive Value* del 75 %. También se postulan nuevas variables que permiten capturar las decisiones realizadas al momento de realizar la vinculación manual.

1. Introducción

La disponibilidad de microdatos durante un intervalo suficientemente largo de tiempo permite responder preguntas cuyas respuestas a menudo involucran un alto grado de especulación y que son de interés contemporáneo. Por ejemplo, gracias a datos de estas características se ha encontrado que, contrario a las concepciones comunes, las familias de migrantes estadounidenses no se quedan rezagadas respecto a los naturales, y que esto es cierto tanto para los migrantes modernos como migrantes de la época de las grandes migraciones (Abramitzky et al., 2021b). Este tipo de fuentes de información también permite entender los efectos de largo plazo de la gripe española en la educación infantil (Beach et al., 2022), y los efectos en el mercado laboral de la implementación de capital que ahorra mano de obra.

Construir microdatos que sigan individuos durante largos periodos de tiempo encierra varias dificultades. La principal de ellas es que, a menudo que las fuentes de información históricas son construidas sin la intención de ser enlazadas, de manera que los individuos no cuentan con identificadores únicos en las tablas, dejando de lado que muchas de ellas ni siquiera tienen formato tabular. Otras dificultades surgen de la pérdida de información, que hace difícil distinguir la representatividad de los datos y aumenta los falsos positivos al intentar establecer vínculos entre tablas. En la última década se ha desarrollado una agenda de investigación que busca responder al primer problema al generar métodos automáticos para vincular tablas de datos históricas.

Abramitzky et al. (2020) realiza un balance de las tres aproximaciones más comunes al problema de encontrar coincidencias de individuos entre tablas de datos históricas. La

primera de ellas consiste en algoritmos que no incorpora un modelo estadístico explícito en favor de la aplicación iterativa de una serie de reglas para definir el conjunto de posibles vínculos y finalmente el vinculo más probable. La segunda consiste en una aplicación del algoritmo de maximización de expectativas (EM) (Abramitzky et al., 2021a), y, finalmente, el uso de algoritmos de aprendizaje supervisado (Feigenbaum, 2016), (Bailey et al., 2020).

Sin embargo, las aplicaciones que motivan el uso de estos algoritmos son a menudo fuentes de datos históricas que han sido conservadas correctamente de manera que se minimice la pérdida de información. Es así como Feigenbaum (2016) ejemplifica el desempeño de su propuesta mediante el uso de un censo de 1915 de Iowa, Estados Unidos, y un censo federal de 1840, ambos transcritos a partir de archivos completos. En la misma dirección, Bailey et al. (2020)emplean el censo federal de 1880 de Estados Unidos y una muestra representativa del censo federal de 1850 para mostrar la eficacia de considerar variables de validación: variables adicionales que están correlacionadas con los vínculos, pero que no se usaron en su construcción y que tampoco los expliquen de forma determinística. (Abramitzky et al., 2020) también emplea censos estadounidenses para comparar los diferentes algoritmos evaluados, y Abramitzky et al. (2021a) utiliza censos estadounidenses y censos nuruegos del siglo XX para determinar la eficacia del uso del algoritmo de maximización de expectativas.

El desempeño de los diferentes métodos no ha sido probado en fuentes con algún grado de pérdida de información. Usando el enfoque de algoritmos de clasificación, el objetivo de este artículo es en primer lugar evaluar el comportamiento de los algoritmos de clasificación en situaciones donde las fuentes han sido sometidas a algún grado de pérdida de información, y seguidamente determinar el mejor algoritmo para realizar la vinculación de las fuentes. El uso de algoritmos de clasificación por encima de otros métodos es preferido por enfoque flexible al problema al permitir definir la función de pérdida y encontrar estrategia explícitas para lidiar con datos faltantes y escasez de variables explicativas que puede disminuir la probabilidad de que dos individuos sean identificados como el mismo.

Este artículo, pretende por tanto identificar el mejor algoritmo de clasificación en fuentes que no han sido conservadas buscando minimizar la fuente de información. Al hacer esto, se busca realizar los siguientes aportes a la literatura: en primer lugar, no existe una validación del comportamiento de los algoritmos para los censos históricos en los que se buscan aplicar; estos son censos sujetos a pérdidas de información, en idioma español y con las reglamentaciones propias de la recolección de censos en Colombia durante el siglo XIX. En segundo lugar, se prueban algoritmos basados en *boosting* que hasta ahora no han sido considerados por la literatura. Finalmente, se expande el conjunto de variables empleadas para establecer el posible vínculo entre individuos, por lo que se aporta también a la especificación del algoritmo.

A continuación se realizará un análisis descriptivo de los datos, y seguidamente se mostrarán los resultados de los diferentes algoritmos en términos de la función de pérdida.

2. Datos

Los datos iniciales consisten en dos censos de población del distrito de Medellín recolectados en los años 1859 y 1869. Estos censos cuentan con información como los nombres, los apellidos, la edad, el género, y, en algunos casos, información sobre un pariente masculino y uno femenino. Ambos censos cuentan con algún grado de pérdida de información: se estima que la población en Medellín es de aproximadamente 15.000 habitantes, mientras que el censos tiene al rededor de 11.000. Así mismo, se infiere que en 1869 la población es

claramente mayor a 15.000 habitantes pero el censos solo contiene 4.900. En la tabla 1 se muestran los nombres más comunes en ambos censos. Como puede observarse, para el año 1859, el nombre más común es un nombre tanto masculino como femenino, "María", mientras que los demás son masculinos a excepción del último. Para el censo de 1869 se tienen casi exactamente los mismos nombres.

Cuadro 1: Nombres más comunes en censos de 1859 y 1869

Censo	Nombre/Apellido	Porcentaje
1859	maria	0.12
	jose	0.05
	juan	0.03
	manuel	0.02
	dolores	0.02
1869	maria	0.09
	jose	0.05
	juan	0.03
	manuel	0.03
	francisco	0.02

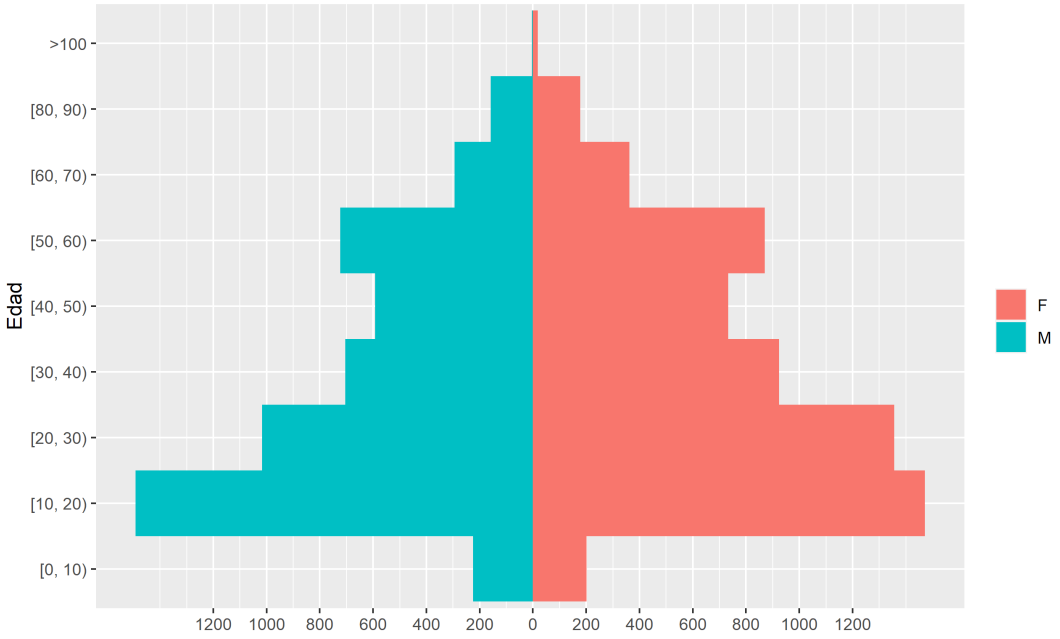
En la tabla 2 se tienen los apellidos más comunes en ambos censos. A diferencia de los nombres, en los apellidos se tiene algo más de variación, repitiéndose los primeros lugares, pero siendo los demás diferentes. Aquí se evidencia la dificultad añadida por la pérdida de información: es posible que las partes sobrevivientes de los censos puedan o no coincidir.

Cuadro 2: Apellidos más comunes en censos de 1859 y 1869

Censo	Nombre/Apellido	Porcentaje
1859	restrepo	0.04
	alvarez	0.03
	gomez	0.03
	arango	0.02
	zapata	0.02
1869	alvarez	0.07
	restrepo	0.06
	londono	0.05
	escobar	0.03
	posada	0.03

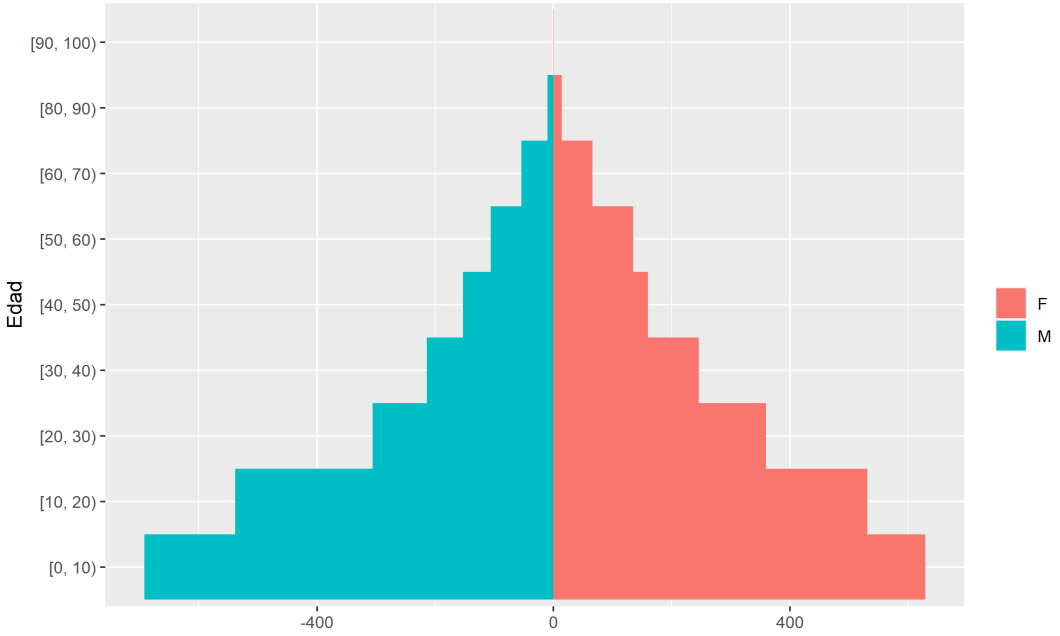
La estructura sociodemográfica de la población, implicada por los censos y caracterizada con ayuda de la edad y el género se presenta en las figuras 1 y 2. En ambos casos se observa una pirámide de forma tradicional lo que indica una población mayoritariamente joven. No obstante, en el año 1869 la base de la pirámide se encuentra en población una década más joven que en 1859; sin embargo esto también puede indicar que la pérdida de información en el segundo caso afectó a la población más joven.

Figura 1: Pirámide poblacional censo de 1859



Otro rasgo interesante de la pirámide, es que en 1859 se implica un mayor volumen de población femenina, lo cual se reversa para el año 1869. Así mismo, en el censos de 1859 se logra capturar población de mayor edad, mientras que en el censo de 1869, la información sobre esta población es escasa.

Figura 2: Pirámide poblacional censo de 1869



A partir de los censos explorados previamente, se procede a realizar la construcción de bloques. Estos bloques consisten en los conjuntos de individuos en el censo B que

podrían ser el mismo individuo en el censo A. En el extremo, este conjunto abarcaría todo el censo B para cada individuo del censo A. No obstante, por eficiencia computacional, es deseable descartar a los individuos para los que se tiene mayor certeza de no encontrar coincidencias. Una aproximación popular en la literatura es realizar la construcción de bloques por coincidencias en el primer nombre, el primer apellido y la edad. Del análisis realizado previamente, es notorio que los nombres y los apellidos, al menos en los casos más comunes, no están sujetos a un patrón regular de errores de digitación o de escritura, por lo que parece razonable emplearlos para realizar la construcción de los bloques. Así mismo, se evidencia que la población más joven o más vieja parece faltar en uno u otro censo, por lo que realizar la construcción de bloques por la edad también es razonable.

Después de realizar la construcción de los bloques, se establecen vínculos de forma manual para constituir una muestra de entrenamiento de la cual pueda aprender el algoritmo. La construcción de estos vinculos sigue la siguiente lógica:

Dos individuos del censo A y del censo B son el mismo, si se cumple que:

- Un único candidato:

    1. Sus géneros son iguales, y
    2. Primeros apellidos son suficientemente parecidos (idénticos o typos evidentes), y
    3. Sus segundos nombres y apellidos son suficientemente parecidos (idénticos o typos evidentes), y
    4. el nombre de sus padres del mismo género es similar (idénticos o typos evidentes) (requiere ser offspring, dependent puede implicar que no son los padres), ó
    5. hay nombre de padre en uno y nombre de madre vacío, y lo inverso sucede en el otro.

- Varios candidatos:

    1. se selecciona mismo género, y
    2. se selecciona segundo nombre o apellido común, o vacío en uno o ambos, y
    3. se reportan padres, se asegura que primer nombre y apellido sean coincidentes o typos evidentes, y
    4. más de una opción está en empate no se elige ninguna.

Usando esta heurística, se realizan 541 vínculos manuales y se realiza la construcción de una serie de variables explicativas que logren capturar las reglas implícitamente seguidas en el proceso de vinculación manual. Estas variables se muestran en la tabla 3:

Cuadro 3: Variables explicativas para el entrenamiento del algoritmo

Variables	Fuente	Resumen
strdistFN	Feigenbaum (2016)	Distancia de edición del primer nombre
strdistLN	Feigenbaum (2016)	Distancia de edición del primer apellido
hits	Feigenbaum (2016)	Tamaño del bloque
$hits^2$	Feigenbaum (2016)	
exact	Feigenbaum (2016)	Hay al menos un individuo con el mismo nombre y apellido
exactall	Feigenbaum (2016)	Número de individuos con el mismo nombre y apellido
genderexact	Propia	Coincidencia de género
lsoundex	Feigenbaum (2016)	Distancia fonética del primer nombre
fsoundex	Feigenbaum (2016)	Distancia fonética del primer apellido
exactmult	Feigenbaum (2016)	Existe al menos un individuo con el mismo nombre, apellido y género
fnamesubset	Propia	El primer nombre es un subconjunto
middlenamenotempty*strdistmn2	Propia	Distancia de edición entre los segundos nombres si ambos los tienen
secondlastnamenotempty*strdistsln2	Propia	Distancia de edición entre los segundos apellidos si ambos los tienen
mstart	Propia	Coincidencia en primera letra del segundo nombre
relationshipoffspring	Propia	Ambos individuos son hijos de alguien
fatherlnnotempty*fatherfnnotempty	Propia	Ambos individuos tienen información del primer nombre y apellido del padre
motherfnnotempty*motherlnnotempty	Propia	Ambos individuos tienen información del primer nombre y apellido de la madre
relationshipoffspring*fatherfnnotempty*strdistfatherfn	Propia	Distancia de edición entre el primer nombre de los padres si tienen información del padre y son hijos de alguien
relationshipoffspring*fatherlnnotempty*strdistfatherln	Propia	Distancia de edición entre el primer apellido de los padres si tienen información del padre y son hijos de alguien
relationshipoffspring*motherfnnotempty*strdistmotherfn	Propia	Distancia de edición entre el primer nombre de las madres si tienen información del padre y son hijos de alguien
relationshipoffspring*motherlnnotempty*strdistmotherln	Propia	Distancia de edición entre el primer apellido de las madres si tienen información del padre y son hijos de alguien
uniquegender	Propia	Es el único individuo del bloque para el cual coincide el género
uniquefnexact	Propia	Es el único individuo del bloque para el cual coincide el primer nombre de forma exacta

3. Resultados

Con base en las variables explicativas presentadas en la sección anterior, se procede a emplear tres algoritmos de clasificación, combinados con diferentes esquemas de remuestreo o no para solucionar lo que es esencialmente un problema de clasificación desbalanceado. Se consideran tres algoritmos: *Probit*, *Logit*, *XGBoost*, y se consideran dos formas generales de tratar el desbalanceo de clases: encontrar el punto de corte óptimo, e implementar diferentes técnicas de remuestreo: *ROSE*, *SMOTE*, *Upsampling* y *Downsampling*.

Para evaluar estos algoritmos, Feigenbaum (2016) propone la siguiente función de pérdida:

$$L = TPR + \gamma\, PPV$$

Donde:

$$TPR = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos + Falsos Negativos}}$$

$$PPV = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos + Falsos Positivos}}$$

y γ es un número mayor que cero.

Esta función de pérdida se preocupa entonces por cuántos vínculos que realiza el investigador logra identificar el modelo (TPR) y cuántos vínculos identificados por el modelo, son efectivamente vínculos reales (PPV), tomando los realizados por el investigador como los vínculos reales.

La tabla 4 muestra el comportamiento de los mejores algoritmos junto con la técnica empleada para tratar el problema de remuestreo, así como los hiperparámetros que permiten el mejor desempeño. Como se observa, de forma adicional a los hiperparámetros usuales y al peso γ ya mencionado, se tienen los hiperparámetros a y b. El primero consiste en el umbral de probabilidad usado para asignar la etiqueta de que los dos individuos consisten en el mismo, y el segundo la razón de probabilidad del primero respecto al segundo.

Cuadro 4: Resultados de los algoritmos

Estrategia de rem.	γ	a	b	L	TPR	PPV	Algoritmo
Umbral de prob.	0	0.6	1.0	0.8571429	0.8571429	0.7500000	Logit
Umbral de prob.	0	0.6	1.5	0.8571429	0.8571429	0.7500000	Logit
Umbral de prob.	0	0.6	2.0	0.8571429	0.8571429	0.7500000	Logit
Umbral de prob.	0	0.6	1.0	0.8571429	0.8571429	0.7500000	Probit
Umbral de prob.	0	0.6	1.5	0.8571429	0.8571429	0.7500000	Probit
Umbral de prob.	0	0.6	2.0	0.8571429	0.8571429	0.7500000	Probit
SMOTE	0	0.5	1.5	0.9523810	0.9523810	0.6451613	XGBoost
SMOTE	0	0.5	2.0	0.9523810	0.9523810	0.6451613	XGBoost

Como se puede observar en dicha tabla, aunque el algoritmo XGBoost permite maximizar la función de pérdida L, lo hace asumiendo el costo de una mayor cantidad de falsos positivos. Los algoritmos Probit y Logit usando el método de búsqueda del umbral de probabilidad óptimo entregan resultados más balanceados, sin importar el valor que se de a la razón de probabilidad b. Así mismo, los valores arrojados de las métricas de desempeño son ligeramente inferiores a los presentados en (Feigenbaum, 2016), para problemas con fuentes de información completas. Lo que permite sustentar que este método puede aplicarse aún en situaciones donde existe pérdida de información.

4. Conclusiones

Los algoritmos de clasificación de *Machine Learning* son aplicables para problemas de vinculación de tablas históricas incluso en situaciones donde estas presentan algún grado de pérdida de información. Debido a la naturaleza desbalanceada del problema, múltiples técnicas como las de remuestreo y la búsqueda del umbral de probabilidad óptimo son empleadas. Se encuentra que esta última es la mejor estrategia para tratar el problema del desbalanceo de clase, en combinación con algoritmos de clasificación tipo Logit y Probit. Los algoritmos basados en Boosting generan una mayor cantidad de falsos positivos en los vínculos.

Vías de investigación posteriores pueden incluir el uso de una función de pérdida distinta, que permita disminuir la presencia de falsos positivos, la cual es la consecuencia principal del uso de fuentes de información incompletas. Así mismo, el uso de validación cruzada que de cuenta de la autocorrelación de los vínculos al interior de los bloques, puede ser una forma deseable de reducir la posible sobreestimación del error de generalización al momento de entrenar el algoritmo.

Referencias

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., and Pérez, S. (2021a). Automated linking of historical data. *Journal of Economic Literature*, 59(3):865–918.

Abramitzky, R., Boustan, L., Jacome, E., and Perez, S. (2021b). Intergenerational mobility of immigrants in the united states over two centuries. *American Economic Review*, 111(2):580–608.

Abramitzky, R., Mill, R., and Pérez, S. (2020). Linking individuals across historical sources: A fully automated approach. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2):94–111.

Bailey, M., Cole, C., and Massey, C. (2020). Simple strategies for improving inference with linked data: A case study of the 1850–1930 ipums linked representative historical samples. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2):80–93.

Beach, B., Clay, K., and Saavedra, M. (2022). The 1918 influenza pandemic and its lessons for covid-19. *Journal of Economic Literature*, 60(1):41–84.

Feigenbaum, J. J. (2016). Automated census record linking: A machine learning approach.