Antidiscriminación en la detección de delitos e intrusiones

Sara Hajian, Josep Domingo-Ferrer and Antoni Martínez-Ballesté
Universitat Rovira i Virgili
Dept. d'Enginyeria Informàtica i Matemàtiques, Cátedra UNESCO de Privacidad de Datos
Av. Països Catalans 26 - E-43007 Tarragona
Correo-e {sara.hajian,josep.domingo,antoni.martinez}@urv.cat

***Resumen*—La recogida automática de datos ha impulsado el uso de la minería de datos para detectar delitos e intrusiones. Los bancos, las compañías de seguros, casinos, etc. exploran cada vez más los datos de sus clientes o empleados para detectar potenciales intrusiones, fraudes o incluso delitos. Los algoritmos de minería se entrenan a partir de conjuntos de datos que pueden estar sesgados en lo que respecta a género, raza, religión u otros atributos. Además, con frecuencia la minería se realiza totalmente o parcialmente por parte de terceras entidades. Por dichas razones, las posibles discriminaciones son un motivo de preocupación. Las intrusiones, los fraudes o los delitos potenciales deberían deducirse de males comportamientos objetivos, y no de atributos sensibles como género, raza o religión. Este artículo expone como limpiar los conjuntos de datos usados para entrenar algoritmos de minería y/o exportados a terceras partes. El objetivo de dicha limpieza es que puedan seguirse extrayendo reglas de clasificación legítimas pero no reglas discriminatorias basadas en atributos sensibles.**

I. Introducción

La discriminación puede definirse como tratar injustamente a las personas a causa de su pertenencia a un cierto grupo. Por ejemplo, los individuos pueden ser discriminados a causa de su raza, ideología, género, etc. En economía y en ciencias sociales se viene estudiando la discriminación desde hace más de medio siglo. Hay varias tareas de toma de decisiones que se prestan a la discriminación, por ejemplo la concesión de préstamos y la selección de personal. En las últimas décadas la mayoría de gobiernos democráticos han adoptado leyes antidiscriminatorias. Algunos ejemplos son la US Equal Pay Act [1], la UK Sex Discrimination Act [2], la UK Race Relations Act [3] y la Directiva europea 2000/43/EC sobre antidiscriminación [4].

Sorprendentemente, la detección de la discriminación en el tratamiento de la información no recibió mucha atención hasta 2008 [5], a pesar de que el uso de sistemas de información para la toma de decisiones está a la orden del día. En efecto, se crean modelos a partir de datos reales (datos de entrenamiento) para facilitar la toma de decisiones en varios ámbitos, tales como medicina, banca o seguridad de redes. En estos casos, si los datos de entrenamiento están sesgados a favor o en contra de una determinada comunidad (por ejemplo los extranjeros), el modelo que se aprende de los datos puede mostrar un comportamiento con prejuicios ilegales. Detectar dichos sesgos potenciales y depurar los datos de entrenamiento sin dañar su utilidad de cara a la toma de decisiones es pues altamente deseable. Las tecnologías de la información pueden jugar un papel importante en la detección y la evitación de la discriminación (esto es, la antidiscriminación [5], [6]). En este sentido, varias técnicas de minería de datos han sido adaptadas para detectar decisiones discriminatorias.

La antidiscriminación juega también un papel importante en ciberseguridad, en la medida en que en dicho campo se usan tecnologías de inteligencia computacional como la minería de datos para la toma de ciertas decisiones. Por lo que sabemos, somos los primeros en ocuparnos de la antidiscriminación en el contexto de la ciberseguridad. Claramente, el reto es evitar la discriminación a la vez que se mantiene la utilidad de los datos para las aplicaciones de seguridad que se apoyan en minería de datos, por ejemplo los sistemas de detección de intrusiones (IDS) o los predictores de delitos.

La contribuciones principales de este artículo son: (1) introducir la antidiscriminación en el contexto de la ciberseguridad; (2) proponer un nuevo sistema de evitación de la discriminación basado en la transformación de los datos que tiene en cuenta varios atributos discriminatorios y sus combinaciones; (3) proponer medidas para evaluar el nuevo método en términos de su éxito en evitar la discriminación y de su impacto en la calidad de los datos.

En este artículo, la Sección II pasa revista a la literatura relacionada; la Sección III introduce la antidiscriminación para aplicaciones de ciberseguridad basadas en minería de datos; la Sección IV repasa resultados de detección de la discriminación; la Sección V presenta el método de evitación de la discriminación y su evaluación; las conclusiones se resumen en la Sección VII.

II. Literatura relacionada

La literatura existente sobre antidiscriminación versa mayormente sobre modelos de minería de datos y técnicas relacionadas. Algunas propuestas se orientan a la detección y medida de la discriminación. Otras tratan de la evitación de la discriminación.

- La *detección de la discriminación* se basa en formalizar las definiciones legales de discriminación[1] y proponer

[1]Por ejemplo, la U.S. Equal Pay Act [1] establece que: "una tasa de selección para cualquier raza, sexo o grupo étnico que sea menos de cuatro quintos de la tasa para el grupo con mayor tasa será en general considerada como una evidencia de impacto adverso".

Género	Etnia	Edad	CP	PerfDesc	P2P	PortScan	Intruso
Mujer	Blanca	Joven	43799	Poco	Sí	Sí	No
Hombre	Negra	Joven	43700	Mucho	No	Sí	Sí
Hombre	Blanca	Mayor	84341	Normal	Sí	No	No
Hombre	Gitana	Joven	72424	Poco	No	Sí	Sí
Mujer	Blanca	Mayor	43743	Mucho	Sí	Sí	Sí
Mujer	Blanca	Joven	43251	Mucho	No	No	No

Figura 1. Ejemplo: información de clientes recogida por un operador de telecomunicaciones

medidas cuantitativas para dicha detección. Estas medidas fueron propuestas por Pedreschi a partir de 2008 [5], [7]. Este enfoque ha sido extendido para incluir la significación estadística de los patrones de discriminación extraídos en [8], y ha sido implementado tal como se describe en [9]. La detección de la discriminación utiliza técnicas propias de la minería de datos.

- *La prevención de la discriminación* consiste en inducir patrones que no lleven a decisiones discriminatorias aunque el sistema se haya entrenado a partir de un conjunto de datos que contenga patrones discriminatorios. Existen tres posibles aproximaciones: (1) adaptar las técnicas de transformación de datos y la generalización basada en jerarquías de la literatura de preservación de la privacidad [6], [11]; (2) cambiar los algoritmos de minería de datos integrando evaluaciones de la discriminación [12]; y (3) postprocesar el modelo de minería de datos para reducir la posibilidad de decisiones discriminatorias [8]. Aunque se hayan propuesto algunos métodos, lo cierto es que la prevención de la discriminación es un campo de investigación abierto y prometedor.

Claramente, una forma trivial de tratar la prevención de la discriminación consistiría en eliminar todos los atributos discriminatorios del conjunto de datos. Sin embargo, pueden haber otros atributos altamente correlacionados con el discriminatorio [5], [6]. La eliminación de estos claramente reduciría la utilidad final del conjunto de datos. Así pues, un tema interesante es encontrar el balance óptimo entre antidiscriminación y utilidad del conjunto de datos.

III. Antidiscriminación y seguridad

En este artículo, usamos como ejemplo el conjunto de datos de aprendizaje mostredo en la Figura 1. Corresponde a los datos recogidos por un proveedor de Internet para detectar posibles intrusos o atacantes de sistemas entre sus abonados. El último atributo (Intruso) es el atributo de clase. Además de atributos personales (Género, Edad, Código Postal (CP), Etnia), el conjunto incluye estos atributos:

- *PerfDesc*: significa perfil de descargas y mide la cantidad promedio de datos que el abonado descarga mensualmente. Sus posibles valores son *Alto*, *Normal*, *Bajo*, *Muy bajo*.
- *P2P*: indica si el abonado suele utilizar programas peer-to-peer.
- *PortScan*: indica si el abonado suele utilizar escaneadores de puertos.

Las técnicas antidiscriminación deberían usarse en el ejemplo anterior: si los datos de entrenamiento están sesgados hacia un determinado grupo de usuarios (por ejemplo, jóvenes), el modelo de aprendizaje mostrará un comportamiento discriminatorio hacia este colectivo, y la mayoría de abonados jóvenes podrían ser clasificados incorrectamente como potenciales intrusos.

Adicionalmente, las técnicas antidiscriminación podrían ser útiles en el contexto de compartición de datos entre IDS. Supongamos que varios IDS comparten sus informes (que contienen información sobre intrusos) para mejorar sus respectivos modelos de detección de intrusos. Antes de que un IDS comparta sus informes, estos se deben limpiar para evitar la propagación de decisiones discriminatorias hacia los otros IDS.

IV. Detección de la discriminación

La detección de la discriminación versa sobre hallar decisiones discriminatorias ocultas en un conjunto de datos sobre un historico de toma de decisiones. El problema básico en la detección y análisis de la discriminación es cuantificar el grado de discriminación que padece un determinado grupo (por ejemplo, grupo étnico) en un contexto dado con respecto al atributo de clase (intruso sí o no). La Figura 2 muestra el proceso de detección de la discriminación, basado en las técnicas que describimos en esta sección.

IV-A. Definiciones básicas

- Un *ítem* es un atributo con su valor, por ejemplo $\{$Género=Mujer$\}$.
- La *extracción de reglas de asociación* intenta, dado un conjunto de transacciones (registros), predecir la ocurrencia de un ítem basándose en las ocurrencias de otros elementos en la transacción.
- Un *itemset* es una colección de uno o más ítems, por ejemplo $\{$Género=Hombre, CP=54341$\}$.
- Una *regla de clasificación* es una expresión $X \rightarrow C$, donde X es un itemset, que no contiene ítems de clase, y C es un ítem de clase, por ejemplo $\{$Género=Mujer, CP=54341$\}$ $\rightarrow$ Intruso=Sí. X es la *premisa* de la regla.
- El *soporte* de un itemset, $supp(X)$, es la fracción de registros que contienen el itemset X. Decimos que una

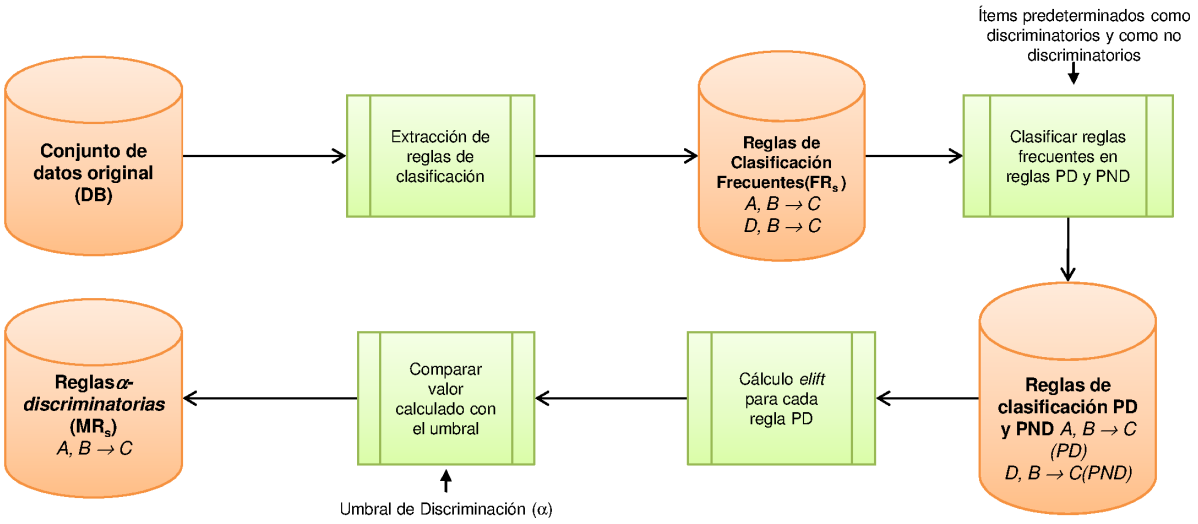

Figura 2. Proceso de detección de la discriminación

regla $X \rightarrow C$ está completamente soportada por un registro si ambos X y C aparecen en el registro.

- La *confianza* de una regla de clasificación, $conf(X \rightarrow C)$, mide la frecuencia en que el ítem de clase C aparece en los registros que contienen X. Entoces, si $supp(X) > 0$

$$conf(X \rightarrow C) = \frac{supp(X, C)}{supp(X)}$$

El rango del soporte y la confianza varia entre $[0, 1]$. La notación también se usa para itemsets negados, es decir $\neg X$.

- Una *regla de clasificación frecuente* es una regla de clasificación con un soporte o confianza mayor que un umbral específico. Sea $\mathcal{DB}$ un conjunto de datos original y sea $\mathcal{FR}_s$ el conjunto de datos de reglas de clasificación frecuentes.

IV-B. Reglas de clasificación potencialmente discriminatorias y no-discriminatorias

Suponiendo que los ítems discriminatorios que contiene $\mathcal{DB}$ están predeterminados (por ejemplo, Etnia=Negra, Edad=Joven), las reglas se dividen en las siguientes categorías:

1. Una regla de clasificación $X \rightarrow C$ es *potencialmente discriminatoria* (PD) si $X = A, B$ siendo A un itemset discriminatorio no vacío y B un itemset no-discriminatorio. Por ejemplo $\{\text{Etnia=Negra, CP=43700}\} \rightarrow \text{Intruso=Sí}$.
2. Una regla de clasificación $X \rightarrow C$ es *potencialmente no-discriminatoria* (PND) cuando X es un itemset no-discriminatorio. Por ejemplo $\{\text{PortScan=Sí, CP=43700}\} \rightarrow \text{Intruso=Sí}$.

La palabra "potencialmente" significa que una regla PD posiblemente lleve a decisiones discriminatorias, por lo que se deben aplicar técnicas para cuantificar este potencial discriminatorio. Además, una regla PND podría llevar a decisiones discriminatorias si se combina con "conocimientos previos", por ejemplo, en el caso anterior que un determinado CP fuera habitado mayoritariamente por personas de color. Este hecho se conoce como *discriminación indirecta*.

IV-C. Medidas de discriminación

Pedreschi *et al.*[5], [8] adaptaron los elementos existentes en la legislación al entorno de las reglas de clasificación, introduciendo una familia de medidas para el grado de discriminación de una regla PD. En nuestra propuesta, usamos su medida *extended lift* ($elift$):

Definición 1: Sea $A, B \rightarrow C$ una regla de clasificación con $conf(B \rightarrow C) > 0$. El extended lift de la regla es

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)}$$

Aquí la idea es evaluar la discriminación de una regla mediante la ganancia de confianza debido a la presencia de atributos discriminatorios (es decir A) en la premisa de la regla. Ciertamente, $elift$ se define como la relación de la confienza entre dos reglas: *con* y *sin* los ítems discriminatorios. El que la regla se considere discriminatoria se puede evaluar mediante aplicar un umbral[2] a $elift$ del siguiente modo[8].

Definición 2: Sea $\alpha \in R$ un umbral fijado. Una regla de clasificación PD $c = A, B \rightarrow C$ es α-protectora con respecto a $elift$ si $elift(c) < \alpha$. Si no es así, c es α-discriminatoria.

Considérese la regla

$$c = \{\text{Etnia=Negra, CP=43700}\} \rightarrow \text{Intruso=Sí}$$

Si α =1.4 y $elift(c)$ =1.46, la regla c es 1.4-discriminatoria.

[2]Nótese que α es un umbral fijo elegido de acuerdo con legislación sobre discriminación.

En términos de discriminación indirecta, la combinación de reglas PND con conocimiento previo podría generar, posiblemente, reglas α-discriminatorias. Si una regla PND c con respecto a conocimiento previo genera una regla α-discriminatoria, c es una regla PND α-discriminatoria; en caso contrario, c es una regla PND α-protectora. De todos modos, en nuestra propuesta sólo nos centramos en discriminación directa, considerando pues reglas α-discriminatorias y suponiendo que todas las reglas PND de $\mathcal{PR}_s$ son PND α-protectoras. De acuerdo con la Figura 2, $\mathcal{MR}_s$ es el conjunto de reglas α-discriminatorias extraídas del conjunto $\mathcal{DB}$.

V. Una propuesta para la prevención de la discriminación

En esta sección presentamos un nuevo sistema de prevención de la discriminación que sigue la aproximación de preprocesado mencionado en la Sección II anterior. El método transforma los datos de origen borrando sesgos discriminatorios, de modo que no se puedan extraer reglas injustas del conjunto de datos generado. La solución propuesta se basa en el hecho de que el conjunto de datos de reglas de decisión estará libre de acusaciones discriminatorias si por cada regla α-discriminatoria r' hay, como mínimo, una regla PND r que lleve a la misma clasifcación que r'.

Nuestro método usa el concepto de p-instancia, formalizado en[13] del modo siguiente:

Definición 3: Sea $p \in [0,1]$. Una regla de clasificación $r' : A, B \rightarrow C$ es una p-instancia de $r : D, B \rightarrow C$ si

1. $conf(r) \geq p \cdot conf(r')$ y
2. $conf(r'' : A, B \rightarrow D) \geq p$.

Si cada r' en $\mathcal{MR}_s$ fuese una p-instancia (donde p es 1 o cercano a 1) de una regla PDN r en $\mathcal{PR}_s$, el conjunto de datos estaría libre de acusaciones discriminatorias.

Considérense las reglas r y r' extraídas del conjunto de datos de la Figura 1:

$$r' = \{\text{Etnia=Negra}, \text{CP=43700}\} \rightarrow \text{Intruso=Sí}$$

$$r = \{\text{PortScan=Sí}, \text{CP=43700}\} \rightarrow \text{Intruso=Sí}$$

Con $p = 0{,}8$, r' es una 0.8-instancia de r si:

1. $conf(r) \geq 0.8 \cdot conf(r')$
2. $conf(r'') \geq 0.8$

donde r'' es:

$$r'' = \{\text{Etnia=Negra}, \text{CP=43700}\} \rightarrow \text{PortScan=Sí}$$

Aunque r' es α-discriminatoria usando $elift$, la existencia de una regla PND r que lleva al mismo resultado que r' y que satisface ambas condiciones (1) y (2) de la Definición 3 muestra que el abonado es clasificado como potencial intruso no a causa de la etnia sino porque usa escaneador de puertos. Así pues, r' está libre de acusaciones discriminatorias, porque el IDS podría argumentar que r' es una instancia de una regla no discriminatoria y más general r. Claramente, r es legítimo, porque el uso de un escaneador de puertos puede considerarse un indicador no sesgado de potencial intruso (no es discriminatorio).

Nuestra solución se basa, pues, en la idea anterior. Transformamos los datos eliminando pruebas de discriminación que aparezcan en forma de reglas α-discriminatorias. Estas reglas α-discriminatorias se dividen en dos grupos: reglas α-discriminatorias tales que hay como mínimo una regla PND que lleve al mismo resultado y reglas α-discriminatorias tales que no hay estas reglas PND. Para el primer grupo se pueden aplicar transformaciones con una pérdida de información mínima. Para el segundo grupo, se deben aplicar cambios que garantizen una mínima pérdida de información de modo que estas reglas α-discriminatorias se conviertan en reglas α-protectoras basándose en la definición de la medida de la discriminación (en nuestro caso $elift$). Nuestro proceso se puede describir mediante cuatro fases:

- *Fase 1*. Usar las medidas de Pedreschi en cada regla para descubrir patrones de discriminación a partir de los datos. La Figura 2 detalla los pasos de esta fase.
- *Fase 2*. Basándonos en la Definición 3, hallar la relación entre reglas α-discriminatorias y las reglas PND extraídas en la primera fase y determinar los requisitos de transformación para cada regla.
- *Fase 3*. Transformar los datos originales para proporcionar los requisitos de transformación para cada regla α-discriminatoria sin afectar considerablemente los datos u otras reglas.
- *Fase 4*. Evaluar el conjunto de datos transformados mediante medidas de prevención de discriminación y pérdida de información que describimos en la Sección V-B.

La primera fase, ilustrada en la Figura 2, consiste en los siguientes pasos. En el primero, las reglas de clasificación frecuentes son extraídas de $\mathcal{DB}$ usando algoritmos bien conocidos como Apriori [10]. En el segundo paso, con respecto a los ítems discriminatorios predeterminados, las reglas extraídas se dividen en dos categorías: PD y PND. En el tercer paso, para cada regla PD, se calcula la medida $elift$ para determinar el conjunto de reglas α-discriminatorias almacenadas en $\mathcal{MR}_s$.

La segunda fase se resume a continuación. En el primer paso de esta fase, para cada regla α-discriminatoria de $\mathcal{MR}_s$ del tipo $r' : A, B \rightarrow C$, se halla un conjunto de reglas PND en $\mathcal{PR}_s$ del tipo $r : D, B \rightarrow C$. Llámese D_{pn} al conjunto de estas reglas PND. Entonces, se evalúan las condiciones de la Definición 3, para un valor de p como mínimo de 0.8 para cada regla de D_{pn}. Pueden darse tres casos, en función de si se cumplen o no las condiciones (1) y (2):

1. Como mínimo, hay una regla en D_{pn} tal que se cumplen ambas condiciones (1) y (2);
2. No hay ninguna regla en D_{pn} que satisfazca ambas condiciones (1) and (2), pero hay como mínimo una regla que satisface una de estas dos condiciones;
3. Ninguna regla en D_{pn} satisface las condiciones (1) o (2).

En el primer caso, es obvio que hay como mínimo una regla r en D_{pn} tal que r' es una p-instancia de r para $p \geq 0{,}8$.

En este caso, no se precisa ninguna transformación. En el segundo caso, la regla PND r_b en D_{pn} se debería seleccionar de modo que el cumplir ambas condiciones requiera el mínimo de transformaciones. Una menor diferencia entre los valores de las dos partes de las condiciones de (1) o (2) para cada r en D_{pn} indica una menor transformación de datos requerida. En este caso, las condiciones (1) y (2) en r_b determinan los requisitos de transformación de r'.

Para el tercer caso, el requisito de transformación de r' determina que esta regla α-discriminatoria se podría transformar en una regla α-protectora basándose en la definición de la medida (es decir, $elift$).

La salida de la segunda fase es un conjunto de datos $\mathcal{TR}_s$ con todas las $r' \in \mathcal{MR}_s$, sus respectivas reglas transformadas r_b y sus respectivos requisitos de tranformación.

La siguente lista muestra el primer, segundo y tercer requisito de transformación que se pueden generar para cada $r' \in \mathcal{MR}_s$ de acuerdo con los casos anteriores:

1. $conf(r' : A, B \rightarrow C) \leq conf(r : D, B \rightarrow C)/p$
2. $conf(r" : A, B \rightarrow D) \geq p$
3. Si $f() = elift$, $conf(r' : A, B \rightarrow C) < \alpha \cdot conf(B \rightarrow C)$

Para las reglas α-discriminatorias con el primer y segundo requisito de transformación, es posible que el coste de satisfacer estos requisitos sea mayor que el coste de satisfacer el tercer requisito. En otras palabras, satisfacer el tercer requisito podría conllevar una menor transformación que satisfacer el primer o segundo requisito. Por lo tanto, para estas reglas el método debe también hacer esta comparación y seleccionar el requisito de transformación que precise un menor coste.

V-A. Método de transformación de los datos

El método de transformación de datos debería incrementar o decrementar la confianza de las reglas hacia un determinado valor objetivo, con el mínimo impacto en la calidad de los datos. Cabe decir que decrementar la confianza de determinadas reglas se ha usado con anterioridad para ocultación de conocimiento [14], [15], [16] y para minería de datos con preservación la privacidad (PPDM).

Suponemos que el ítem de clase C es un atributo binario. Los detalles del proceso de transformación se describen a continuación:

1. Para las reglas α-discriminatorias con el primer requisito de transformación ($conf(A, B \rightarrow C) \leq conf(D, B \rightarrow C)/p$), los valores de ambas partes de la desigualdad son independientes, por lo tanto el valor de la parte izquierda se podría decrementar sin impacto alguno en el valor de la parte derecha. Una solución posible para decrementar

$$conf(A, B \rightarrow C) = \frac{supp(A, B, C)}{supp(A, B)} \qquad (1)$$

   hasta un valor objetivo es perturbar el ítem de clase de C a $\neg C$ en el subconjunto $\mathcal{DB}_c$ de todos los registros del conjunto de datos original que soporten completamente la regla $A, B \rightarrow C$ y tenga impacto mínimo en otras reglas para decrementar el numerador de la expresión (1), manteniendo fijo el denominador[3].
2. Para las reglas α-discriminatorias con el segundo requisito de transformación ($conf(A, B \rightarrow D) \geq p$), el valor del lado derecho es fijo, con lo que el valor del lado izquierdo se puede incrementar independientemente. Una posible solución para incrementar

$$conf(A, B \rightarrow D) = \frac{supp(A, B, D)}{supp(A, B)} \qquad (2)$$

   por encima de p es perturbar el ítem D de $\neg D$ a D en el subconjunto $\mathcal{DB}_c$ de todos los registros en el conjunto de datos original que soporten completamente la regla $A, B \rightarrow \neg D$ y tengan un impacto mínimo sobre las otras reglas al incrementar el numerador de la expresión (2) fijando el denominador.
3. Para las reglas α-discriminatorias con el tercer requisito de transformación ($conf(A, B \rightarrow C) < \alpha \cdot conf(B \rightarrow C)$), a diferencia de los casos anteriores, los dos lados son dependientes; así pues, se requiere una transformación que decremente el lado izquierdo sin impacto en el lado derecho. Una solución posible para decrementar

$$conf(A, B \rightarrow C) = \frac{supp(A, B, C)}{supp(A, B)} \qquad (3)$$

   es perturbar el ítem A de $\neg A$ a A en el subconjunto $\mathcal{DB}_c$ de todos los registros del conjunto de datos original que soporten completamente la regla $\neg A, B \rightarrow \neg C$ y tenga mínimo impacto en las otras reglas para incrementar el denominador de la expresión (3) manteniendo el numerador y $conf(B \rightarrow C)$ fijados[4].

Los registros de $\mathcal{DB}_c$ deberían modificarse hasta que el requisito de transformación se alcance para cada regla α-discriminatoria. Entre los registros de $\mathcal{DB}_c$, se deberían cambiar aquellos con menor impacto en las otras reglas.

Así pues, para cada registro $db_c \in \mathcal{DB}_c$ se considera el impacto de db_c, esto es $impact(db_c)$; el razonamiento es que cambiando db_c se repercute en la confianza de estas reglas. Entonces, los registros db_c con mínimo $impact(db_c)$ se seleccionan para modificación, con el objetivo de puntuar bien en relación con las medidas de utilidad que proponemos a continuación.

Esto significa que transformar db_c con un impacto mínimo $impact(db_c)$ podría reducir el impacto de esta transformación cambiando las reglas α-protectoras a reglas α-discriminatorias y generando las reglas extraíbles del conjunto de datos original en en conjunto de datos transformado.

[3]Eliminar los registros que completamente soportan $A, B \rightarrow C$ no ayudaría porque decrementaría tanto el numerador como el denominador de la expresión (1).

[4]Borrar los registros del conjunto de datos originales que soportan completamente la regla $A, B \rightarrow C$ no ayudaría porque se decrementaría tanto el numerador como el denominador de la expresión (3) y también $conf(B \rightarrow C)$. Cambiar el ítem de clase C tampoco ayudaría porque tendría repercusión en $conf(B \rightarrow C)$.

V-B. Medidas de utilidad

La solución propuesta debe evaluarse en relación con dos aspectos:

- El éxito de la propuesta en borrar pruebas de discriminación del conjunto de datos original (grado de prevención de la discriminación).
- El impacto en la calidad de los datos (grado de pérdida de información).

Un método de prevención de la discriminación debería proveer un buen compromiso entre los dos aspectos anteriores. Con todo esto, proponemos las siguientes medidas:

- *Grado de Prevención de Discriminación* (GPD). Esta medida cuantifica el porcentaje de reglas α-discriminatorias que ya no lo son en el conjunto de datos transformados.
- *Preservación de la Protección Andi-discriminación* (PPD). Esta medida cuantifica el porcentaje de reglas α-protectoras del conjunto original que se mantienen como α-protectoras en el conjunto transformado (DPP puede no ser un 100 % como efecto colateral de los cambios).
- *Coste en Pérdidas* (CP). Esta medida cuantifica el porcentaje de reglas que eran extraíbles en el conjunto original y no lo son en el conjunto transformado.
- *Coste Fantasma* (CF). Esta medida cuantifica el porcentaje de reglas que se pueden extraer del conjunto transformado, pero que no se extraían del original.

Las medidas GPD y PPD se usan para evaluar el éxito del método propuesto, e idealmente deberían ser el 100 %. Las medidas CP y CF se usan para evaluar el grado de pérdida de información (es decir, impacto en la calidad de los datos); idealmente, deberían ser 0 %. Estas dos medidas fueron propuestas como medidas de pérdida de información en PPDM [17].

VI. Resultados experimentales

Esta sección presenta los resultados de la evaluación de nuestra propuesta. Usamos el conjunto de datos "German Credit"[5], el cual es conocido y habitualmente usado en el contexto que nos concierne. Contiene 1,000 registros y 20 atributos de titulares de cuentas bancarias. Para los experimentos, hemos usado DI_s= {Foreign worker=Yes, Personal Status=Female and not Single, Age=Old} (Age=Old: 50).

La Figura 3 muestra a la izquierda el grado de pérdida de información (promedios de CP y CF) y a la derecha el grado de eliminación de la discriminación (promedios de GPD y PPD), para valor del umbral α variando entre 1.2 y 1.7, soporte mínimo 5 % y confianza mínima 10 %.

El número de reglas discriminatorias directas extraídas del conjunto de datos es 991 para α =1.2, 415 para α =1.3, 207 para α =1.4, 120 para α =1.5, 63 para α =1.6 y 30 para α =1.7, respectivamente.

Tal y como se muestra en la Figura 3, el grado de eliminación de la discriminación por todos los métodos y distintos valores es también 100 %. Sin embargo, el grado de pérdida de información decrece substancialmente a medida que α incrementa; la razón es que, mientras α crece, el número de reglas discriminatorias con las que tratar decrece. Adicionalmente, tal y como se muestra en la Figura 3, los valores más bajos de pérdida de información para la mayoría de valores α se obtienen por el método de transformación vinculado a la Expresión (1) (cambio del valor de clase).

[5]http://archive.ics.uci.edu/ml

VII. Conclusiones

Hemos examinado el posible impacto de la discriminación en aplicaciones de ciberseguridad, especialmente en sistemas de detección de intrusiones (IDSs). Los IDSs usan tecnologías de inteligencia computacional tales como minería de datos. Es obvio que, si los datos para entrenar dichos sistemas fuesen discriminatorios, ello les conduciría a tomar decisiones discriminatorias a la hora de predecir intrusiones o, más generalmente, delitos. Nuestra contribución se centra en producir datos de entrenamiento que estén libres o casi libres de discriminación y que a la vez preserven su utilidad para detectar intrusiones o crímenes reales. Con el fin de controlar la discriminación en un conjunto de datos, un primer paso consiste en detectar si hay tal discriminación. Si la hay, el conjunto de datos es modificado hasta que la discriminación se reduce por debajo de un cierto umbral o es completamente eliminada. En [18] presentamos experimentos sobre datos reales con algunos de los métodos descritos. Como trabajo futuro, vamos a extender los experimentos a conjuntos de datos de distintos tamaños, con varias proporciones de atributos discriminatorios. Pretendemos asimismo mejorar los métodos de forma que se optimice el compromiso entre utilidad y supresión de discriminación.

Agradecimientos

Los autores están encuadrados en la Cátedra UNESCO de Privacidad de Datos, pero las ideas expresadas en este artículo lo son a título personal y no reflejan necesariamente la posición de UNESCO. Este trabajo ha sido financiado por el Gobierno de España bajo los proyectos TSI2007-65406-C03-01 "E-AEGIS", TSI2011-27076-C03-01 "CO-PRIVACY" y CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", por la Generalitat de Catalunya mediante la ayuda 2009 SGR 01135, y por la Comisión Europea mediante el proyecto del 7o PM "DwB". El segundo autor está financiado parcialmente como investigador ICREA Acadèmia por la Generalitat de Catalunya.

Referencias

[1] United States Congress, *US Equal Pay Act*, 1963. http://archive.eeoc.gov/epa/anniversary/epa-40.html

[2] Parliament of the United Kingdom, *Sex Discrimination Act*, 1975. http://www.opsi.gov.uk/acts/acts1975/PDF/ukpga_19750065_en.pdf

[3] Parliament of the United Kingdom, *Race Relations Act*, 1976. http://www.statutelaw.gov.uk/content.aspx?activeTextDocId=2059995

[4] European Commission, *EU Directive 2000/43/EC on Anti-discrimination*, 2000. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2000:180:0022:0026:EN:PDF

[5] D. Pedreschi, S. Ruggieri y F. Turini, "Discrimination-aware data mining". *Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pp. 560-568. ACM, 2008.

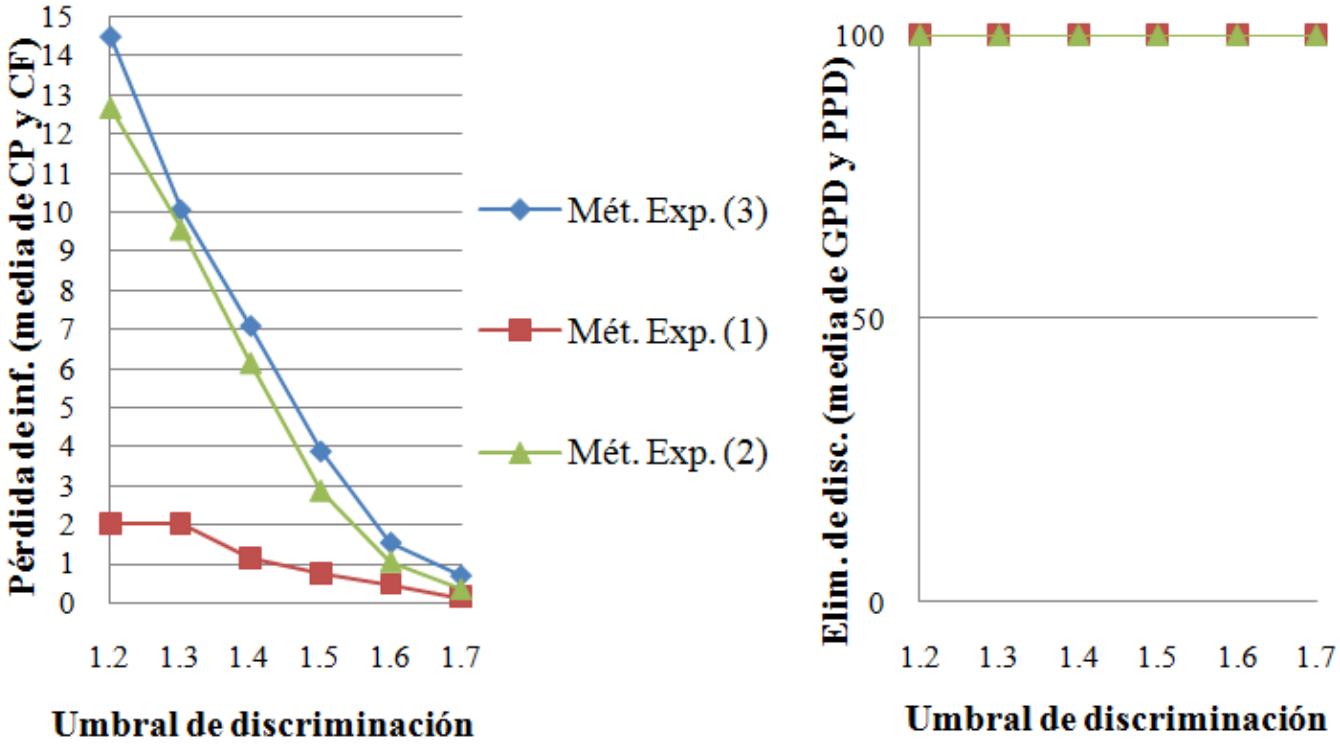

Figura 3. Pérdida de información (izquierda) y grado de eliminación de la discriminación (derecha) para $\alpha \in [1{,}2, 1{,}7]$.

[6] F. Kamiran y T. Calders, "Classification without discrimination". *Proc. of the 2nd IEEE International Conference on Computer, Control and Communication (IC4 2009)*. IEEE, 2009.

[7] S. Ruggieri, D. Pedreschi y F. Turini, "Data mining for discrimination discovery". *ACM Transactions on Knowledge Discovery from Data*, 4(2) Article 9, ACM, 2010.

[8] D. Pedreschi, S. Ruggieri y F. Turini, "Measuring discrimination in socially-sensitive decision records". *Proc. of the 9th SIAM Data Mining Conference (SDM 2009)*, pp. 581-592. SIAM, 2009.

[9] S. Ruggieri, D. Pedreschi y F. Turini, "DCUBE: Discrimination Discovery in Databases". *Proc. of the ACM International Conference on Management of Data (SIGMOD 2010)*, pp. 1127-1130. ACM, 2010.

[10] R. Agrawal y R. Srikant, "Fast algorithms for mining association rules in large databases". *Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499. VLDB, 1994.

[11] F. Kamiran y T. Calders, "Classification with No Discrimination by Preferential Sampling". *Proc. of the 19th Machine Learning conference of Belgium and The Netherlands*, 2010.

[12] T. Calders y S. Verwer, "Three naive Bayes approaches for discrimination-free classification", *Data Mining and Knowledge Discovery*, 21(2):277-292. 2010

[13] D. Pedreschi, S. Ruggieri y F. Turini, "Integrating induction and deduction for finding evidence of discrimination". *Proc. of the 12th ACM International Conference on Artificial Intelligence and Law (ICAIL 2009)*, pp. 157-166. ACM, 2009.

[14] V. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin y E. Dasseni, "Association rule hiding". *IEEE Trans. on Knowledge and Data Engineering*,16(4):434-447, 2004.

[15] Y. Saygin, V. Verykios y C. Clifton, "Using unknowns to prevent discovery of association rules". *ACM SIGMOD Record*, 30(4):45-54, 2001.

[16] J. Natwichai, M. E. Orlowska y X. Sun, "Hiding sensitive associative classification rule by data reduction". *Advanced Data Mining and Applications (ADMA 2007)*, LNCS 4632, pp: 310-322. 2007.

[17] S. R. M. Oliveira y O. R. Zaiane. "A unified framework for protecting sensitive association rules in business collaboration". *International Journal of Business Intelligence and Data Mining*, 1(3):247–287, 2006.

[18] S. Hajian, J. Domingo-Ferrer y A. Martínez-Ballesté, "Discrimination prevention in data mining for intrustion and crime detection", in *IEEE Symposium on Computational Intelligence in Cyber Security-CICS 2011*, Paris, France, pp. 47-54. IEEE Xplore Digital Library.