BENEMÉRITA UNIVERSIDAD
AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO
MATEMÁTICAS

MÉTODOS MATEMÁTICOS PARA DETECCIÓN Y CORRECCIÓN DE DATOS ERRÓNEOS

Tesis para obtener el título de:
Licenciado en Matemáticas Aplicadas

Presenta:
Ozkar Hernández Montero

Directores de tesis:
Dr. Gilberto Calvillo Vives
Dr. Guillermo López Mayo

Agosto 2011

Agradecimientos

Gracias a mi familia por apoyarme siempre, a mi asesor Gilberto Calvillo por su paciencia y hospitalidad, a mi tutora Lidia Aurora Hernández por su apoyo y consejos, a Arnoldo Bezanilla por su excelente disposición y buen humor, a Carlos Guillén por su excelente disposición y solidaridad, a mis amigos por el fut, el fron, el baile, los viajes, el cŕculo de estudio y la simpleza.

Índice general

Introducción

En la actualidad la recolección de datos es una actividad frecuente e importante para obtener información en diversas disciplinas y actividades, tanto económicas como académicas. Este trabajo está dirigido al análisis de datos provenientes de un ejercicio censal tal como un censo de población. Antes de utilizar estos datos como fuente de información es necesario llevar a cabo un control de calidad sobre ellos, ya que tales datos son susceptibles de contener errores. El objetivo principal será utilizar la modelación y programación matemática para mejorar la calidad estadística contenida en los datos.
En México el responsable de realizar censos de población es el Instituto Nacional de Estadística y Geografía (INEGI)[13], el último se levantó en 2010, a continuación se cita la motivación y los objetivos que el INEGI expresa para realizar este censo.

> "Los censos de población y vivienda constituyen la fuente de información estadística más completa sobre la cual se apoya el conocimiento de la realidad nacional, y el Censo 2010 no es la excepción; con los datos que genera, además de responder a las preguntas de ¿cuántos somos? ¿cómo somos? y ¿dónde y cómo vivimos?, permite a los diversos sectores sociales identificar el rezago social, los grupos vulnerables; las necesidades de la población en materia de vivienda, educación, salud, servicios de agua potable, electricidad y drenaje, entre otras, y, a partir de ello, elaborar planes y programas que tiendan a mejorar las condiciones de vida de los habitantes.
>
> A la sociedad le provee de datos básicos sobre el volumen y las características de su localidad, municipio, estado y del país en general; a los empresarios les proporciona información útil para fundamentar la toma de decisiones referente a sus negocios; a los estudiantes e investigadores les suministra estadísticas que les permiten conocer el perfil demográfico, económico y social de la población y del parque habitacional, que favorecen la planeación de proyectos, estudios de los habitantes y diagnósticos,

entre otros; a los diferentes órdenes de gobierno e instituciones les brinda insumos básicos para la planeación, programación, toma de decisiones, seguimiento y evaluación de los planes y programas que elaboran.

En el plano internacional, es indispensable en el monitoreo de las metas establecidas en la Cumbre del Milenio de las Naciones Unidas, orientadas a combatir pobreza, enfermedades, analfabetismo, discriminación contra la mujer y degradación del medio ambiente. Además, permite la actualización de información que los organismos internacionales requieren con fines comparativos (ONU, OIT, CEPAL, CELADE, etcétera).

El propósito fundamental del Censo de Población y Vivienda 2010 es contar a la población residente del país, actualizar la información sobre sus principales características demográficas y socioeconómicas, y ubicar su distribución en el territorio nacional; así como enumerar a las viviendas y captar datos sobre sus características básicas.

Además, se busca enriquecer la serie histórica de información demográfica y socioeconómica, manteniendo en general la comparabilidad con los censos efectuados en México y en otros países; generar insumos para la elaboración de las proyecciones de población, y aportar información para la construcción de los marcos muestrales sobre los cuales se levantan las encuestas en hogares.

Junto a estos objetivos se establecieron las siguientes metas:

- Realizar una enumeración exhaustiva de la población y las viviendas existentes en el país.
- Ahondar en el conocimiento de algunos temas prioritarios, mediante la aplicación de un cuestionario ampliado en una muestra de las viviendas.
- **Obtener información de óptima calidad.**
- Entregar los resultados de manera oportuna"[14].

Los objetos que estudiaremos son los cuestionarios del censo, más concretamente, las respuestas a todo un cuestionario como unidad de información. Precisemos esto. Se entenderá por *registro* al conjunto de respuestas al cuestionario del censo obtenidas de una misma fuente, puede pensarse en el conjunto de respuestas de un individuo. La primera abstracción que necesitamos es identificar cada registro con un objeto matemático ubicado en un espacio matemático. La forma más natural de hacer esto, para nuestros fines, es considerar puntos en un producto cartesiano de conjuntos. La forma de entender esta identificación es pensar en cada pregunta como el conjunto de

las posibles respuestas a ella, así, un registro se identifica con un punto en el *espacio producto* de los conjuntos de respuestas a cada pregunta de la encuesta.

Un problema frecuente es que algunos de los datos obtenidos son incongruentes. Una parte de este trabajo es explicar que significa que un registro contenga errores. En este contexto se consideran dos tipos de errores:

1. El primero tiene que ver con las respuestas consideradas aceptables para cada pregunta, es decir, cuando la respuesta a cierta pregunta de la encuesta no corresponde a los valores que tienen sentido, como por ejemplo:

   Edad=-3, Edad=130.

2. El segundo tipo de error tiene que ver con la concordancia que debe existir entre las respuestas a distintas preguntas, como por ejemplo:

   (Edad=10 años) y (Estado civil= casado).

Estos errores pueden ser introducidos por las personas que recolectan los datos o en el proceso de transferir los datos de los cuestionarios a los sistemas de cómputo. La ocurrencia de errores en los datos observados hace necesario llevar a cabo un proceso de revisión de los datos recolectados y en caso de ser necesario corregirlos. Este proceso de revisión y corrección de datos es llamado *edición estadística de datos* (EED de aquí en adelante)[5]. Su objetivo es mejorar la calidad de la información estadística contenida en los datos.

En este trabajo se describen dos metodologías para llevar a cabo la EED de forma automática, es decir mediante la implementación de algoritmos computacionales. Realizar la EED de esta manera es necesario y al mismo tiempo presenta diversos problemas[6], veremos que los problemas surgen por la cantidad de datos que deben procesarse, un censo de población involucra millones de datos.
La EED involucra varias etapas, que a grandes rasgos son:

- Debe definirse las *reglas* o *ediciones* que determinarán la validez de cada dato, generalmente un grupo de expertos en el tema de la encuesta proporciona estas reglas, también veremos que este conjunto de reglas como sistema puede ser inconsistente; detectar esta situación es uno de los problemas que abordaremos.

- Las reglas se usan para detectar errores en los registros; analizaremos dos modelos diferentes para atacar este problema.

- Por último, una vez identificados los datos erróneos, es necesario un criterio para corregirlos o imputarlos.

Cada una de estas etapas supone problemas que deben ser resueltos obedeciendo ciertos objetivos.
En el capítulo 1 se describe el problema y se definen los principales conceptos que se usarán en el resto del trabajo. La parte I está dedicada a uno de los modelos más influyentes en el tema, presentado por Fellegi y Holt en 1976. En la parte II se presenta otro modelo, propuesto por Renato Bruni en 2005. Ambas partes se desglosan de la misma forma; comienzan con una descripción general del modelo, luego se describe la metodología referente a la identificación de los datos erróneos y por último se describe la metodología para corregir estos datos.
En el apéndice A se proporcionan definiciones y resultados básicos de programación lineal como marco de referencia, se evitan las demostraciones pues éstas son parte de los textos clásicos citados en la bibliografía. El apéndice B está dedicado a presentar el lema de Farkas, debido a que uno de los principales resultados de este trabajo (el usado para identificar inconsistencias en las reglas con los vértices del poliedro apropiado en el capítulo 6) es consecuencia de este lema y a que su demostración rigurosa no es común en los textos básicos, se presenta una demostración detallada.
Detrás de algunas ideas centrales en la edición estadística de datos se encuentra el problema denominado *set-covering*, el apéndice C esta dedicado a dar un panorama general de este problema.

Capítulo 1

Modelación

Este trabajo trata de plantear en un contexto matemático un problema concreto, así que lo primero que debemos hacer es delimitar muy bien el problema que necesitamos resolver, para después, construir el lenguaje y los objetos matemáticos que lo representen. El *problema principal* (PP de aquí en adelante) es el siguiente:

> Supongamos que tenemos los datos que se obtienen de un censo de población. Un grupo de expertos en el tema de la encuesta proporciona un conjunto de reglas, las cuales determinan la validez de cada dato. Queremos desarrollar un algoritmo que determine si un dato particular satisface todas las reglas. Si se determina que un dato no satisface todas las reglas queremos desarrollar un algoritmo que genere un dato correcto a partir del erróneo.

El objetivo de este capítulo es transformar el planteamiento del problema anterior en un problema puramente matemático, es decir, expresado en un lenguaje estrictamente matemático, con el objetivo de delimitarlo y evitar ambigüedades. Lo primero es precisar el significado matemático de los objetos que intervienen en el problema:

- datos que se obtienen de un censo de población,
- un conjunto de reglas,
- un algoritmo de detección de errores y
- un algoritmo de corrección de errores

Una parte fundamental en el trabajo de modelación que se presenta es la equivalencia que existe entre el lenguaje de la lógica y el lenguaje de los conjuntos. El

objetivo de este capítulo es tener acuerdo en el uso del lenguaje y precisar notación que pudiera causar confusión.
Sean $D_1, D_2, \ldots, D_m$ conjuntos. Usaremos la siguiente notación para el *espacio producto*

$$x = (x_1, x_2, \ldots, x_m) \in D = D_1 \times D_2 \times \cdots \times D_m \quad \Leftrightarrow \quad \forall i \in \{1, \ldots, m\} : x_i \in D_i.$$

Si U denota el conjunto universo y A es un subconjunto, su *complemento* se expresa como

$$A^c = U \setminus A = \{a \in U : a \notin A\}$$

Teorema 1.1. *Sean $A_1, \ldots, A_m$ y $B_1, \ldots, B_m$ conjuntos. Denotemos*

$$A = A_1 \times A_2 \times \cdots \times A_m.$$
$$B = B_1 \times B_2 \times \cdots \times B_m.$$

Entonces,

$$A \cap B = A_1 \cap B_1 \times A_2 \cap B_2 \times \cdots \times A_m \cap B_m$$

Definición 1.2. Sea $\bar{A}$ un conjunto, definimos la unión amalgamada de $\bar{A}$ como

$$\bigcup \bar{A} = \{a : \exists A \in \bar{A} : a \in A\}$$

1.1. Los datos

Los datos a los que nos referiremos en este trabajo son de tipo categórico. Al respecto de los datos categóricos basta decir que se trata de conjuntos finitos en los que sólo podemos hacer alguna clasificación mediante agrupación en subconjuntos, o de forma más precisa, particiones. Incluso en los campos que de forma natural tienen significado numérico, como edad e ingresos, son solo considerados los rangos a los que pertenecen, es decir, el subconjunto específico en que se encuentran.
Los datos provienen de una encuesta del tipo de un censo de población, así, cuando nos referimos a un *dato* o *registro* en particular se trata de las respuestas obtenidas de una misma fuente, o individuo, a todo el cuestionario. Cada una de las entradas que conforman el cuestionario se denomina *campo*, es decir, la respuesta a cada pregunta. Matemáticamente modelaremos un dato como un punto m-dimensional $y = (y_1, \ldots, y_m)$ cuya componente y_j es una *entrada* en el *campo* j, donde m es el número de campos que hay en la encuesta. Definimos el conjunto D_j como el conjunto de todas las entradas posibles para el campo j, lo llamaremos *dominio del campo j*. Definimos el *espacio de datos* como:

$$D = D_1 \times \cdots \times D_m = \prod_{j=1}^{m} D_j$$

el producto cartesiano $\prod$ de las posibles entradas de los campos. Usaremos la letra F (posiblemente acompañada de un subíndice y un supraíndice) para denotar subconjuntos de algún D_j.

Ejemplo 1.3.

dato : (Edad = 24, Estado Civil = soltero, Ocupación= estudiante).
Campos : Edad, Estado Civil, Ocupación.

Espacio de datos:

$D_1 = \{0, 1, \ldots, 110\}$,
$D_2 = \{soltero, casado, viudo, divorciado\}$,
$D_3 = \{estudiante, trabajador, desempleado, trabajador\ en\ el\ hogar\}$.

Otro aspecto relevante de los registros es su distribución de frecuencia. Esta característica será fundamental durante el proceso de imputación, o corrección, y por ello conviene comprender su significado. Los registros que tenemos son un muestreo de una determinada población, la cual suponemos tiene una distribución en el sentido estadístico y en general el objetivo final es hacer alguna inferencia estadística sobre ella. Cada registro es un punto multidimensional, en el que cada entrada corresponde a la respuesta de una pregunta específica. Cada entrada en un registro tiene una distribución de frecuencia particular, es decir, hay valores que aparecen más que otros, o son más probables que otros, en este sentido tenemos una *distribución marginal de frecuencia* para cada campo. Pero también existe una *distribución conjunta de frecuencia* cuando pensamos en cada punto como un solo registro, es decir, existen combinaciones de valores que ocurren con mayor frecuencia que otras, o son más probables que otras.

En lo sucesivo entenderemos el término *registro* simplemente como un punto en D, y viceversa. Denotaremos $Dat \subset D$ al conjunto de registros que deseamos editar.

1.2. Las reglas

Las reglas referidas en el planteamiento de PP son proposiciones lógicas acerca de los puntos en D, esto quiere decir que dado un registro y y una regla R siempre

podemos decidir si y hace verdadera a R, por lo tanto R identifica a un conjunto de puntos, un registro particular y es declarado correcto respecto a un conjunto de reglas si y sólo si satisface todas las reglas.

De aquí en adelante utilizaremos el término *regla* para referirnos a un subconjunto del espacio de datos.

Supondremos que inicialmente contamos con un conjunto de reglas, que acordaremos en llamar *reglas explícitas* y que denotaremos con $\bar{R}_e = \{R^1, \ldots, R^r\}$. Usaremos $\bar{R}$ (acompañado posiblemente por un subíndice) para denotar conjuntos de reglas, mientras que R (acompañado posiblemente por un superíndice) denotará siempre una regla. Constantemente haremos referencia al conjunto de puntos determinado por un conjunto de reglas $\bar{R}$ (su unión amalgamada), para facilitar la escritura usaremos la siguiente notación,

$$D\bar{R} = \bigcup \bar{R}$$

con la intención de enfatizar que se trata de un subconjunto de puntos en D.

1.3. Errores

Lo primero que debemos establecer acerca de lo que entenderemos como error en este trabajo es que **suponemos** que los errores son no intencionales y por lo tanto son aleatorios.
Antes de usar los registros para cualquier tipo de análisis o inferencia es necesario algún tipo de detección de errores. Por detección de errores nos referimos a:

I revisar cada campo de cada registro de la encuesta para cerciorarse de que contiene una entrada válida; y

II revisar las entradas en ciertas combinaciones predeterminadas de campos para asegurarse de que las entradas son consistentes unas con otras.

Verificar el tipo I incluye determinar si la entrada de cualquier campo es un blanco inválido (para algunos campos entradas en blanco pueden ser válidas, por ejemplo, los ingresos reportados de menores de edad), o si las entradas se encuentran en un conjunto válido de códigos para ese campo. Algunos ejemplos son: edad no debe ser blanco o negativo, estado civil no debe ser blanco y el número de hijos debe ser menos que 20. Mientras que las reglas del tipo I son consecuencia directa del cuestionario, las reglas del tipo II usualmente se establecen en base a un amplio conocimiento del tema de la encuesta. Conceptualmente, las reglas del tipo II especifican de alguna

manera conjuntos de valores para combinaciones específicas de campos que son conjuntamente inaceptables, por ejemplo, si la edad es menor que 15 años, el estado civil debe ser soltero.
Los valores a los que se refiere I, aquellos que no son aceptables sin tomar en cuenta los valores de todos los otros campos, son llamados valores *fuera de rango*. Removiendo los valores fuera de rango de un dominio D'_j, obtenemos el *dominio factible* $D_j \subseteq D'_j$. Los *dominios factibles* son delimitados usando reglas muy simples, pues solo intervienen en ellas valores de un solo campo, así, para facilitar la escritura y dar una presentación más clara de los modelos de EED, asumiremos que

> los valores fuera de rango han sido ya detectados, sin que esto afecte la generalidad de los métodos. En resumen a partir de ahora asumimos que de entrada **ya tenemos dados los dominios factibles para cada campo.**

La anterior suposición nos permitirá concentrarnos en los errores descritos en II, que es donde verdaderamente está la complicación del problema.

1.4. Edición estadística de datos

EL termino *edición estadística de datos* (EED de aquí en adelante) se utiliza para englobar una amplia gama de metodologías dirigidas a mejorar la calidad estadística de datos. En general consta de dos etapas: la de detección (edición) y la de corrección. Debido a su relevancia y proliferación, el problema anterior (detección y corrección de errores en registros provenientes de encuestas) ha sido extensamente estudiado en distintas comunidades científicas: estadística, ciencias de la computación y teoría de bases de datos (ver [5]). El enfoque que aquí se presenta es el estadístico.

1.4.1. Detección

La idea básica de la EED es muy simple; en el espacio de datos D hay dos tipos de puntos, los correctos y los erróneos, la detección es la etapa en la que se determina cuales son los registros correctos y cuales son los erróneos.

1.4.2. Corrección

Cuando se observa un registro erróneo la idea es modificarlo para obtener un registro correcto. Hacer esto implica responder ciertas preguntas; ¿que campos se deben modificar? (a este problema se le llama *localización del error*), una vez seleccionados

los campos que se modificarán surge la pregunta ¿que valores específicos se deben poner al modificar cada campo seleccionado? (a esto se le llama *imputación*). Para responderlas se usan dos criterios: modificar la menor cantidad posible de campos y respetar lo más posible la distribución de los registros.

Imputación

Por imputación de un registro dado entenderemos cambiar los valores en algunos de sus campos por alternativas posibles con el objetivo de asegurarse que el registro resultante satisfaga todas las reglas y conserve tanta información contenida en el registro erróneo como sea posible. Se espera que esto produzca un registro que debería estar tan cerca como sea posible del (desconocido) *registro original* (el que debería estar presente en ausencia de error). Los valores que se usan para hacer tal sustitución suelen tomarse de otros registros que ya han sido procesados y se consideran satisfactorios.

1.4.3. Resumen

Si acordamos que las reglas son conjuntos de puntos erróneos podemos plantear el PP en términos completamente matemáticos;

> Sea D el conjunto universo y supongamos que tenemos dados unos subconjuntos Dat y $D\bar{R}_e$. Queremos desarrollar un algoritmo que determine $Dat \cap D\bar{R}_e$. Si $Dat \cap D\bar{R}_e \neq \emptyset$, queremos desarrollar un algoritmo que asocie a cada registro en $Dat \cap D\bar{R}_e$ un registro en $(D\bar{R}_e)^c$,ver figura 1.1.

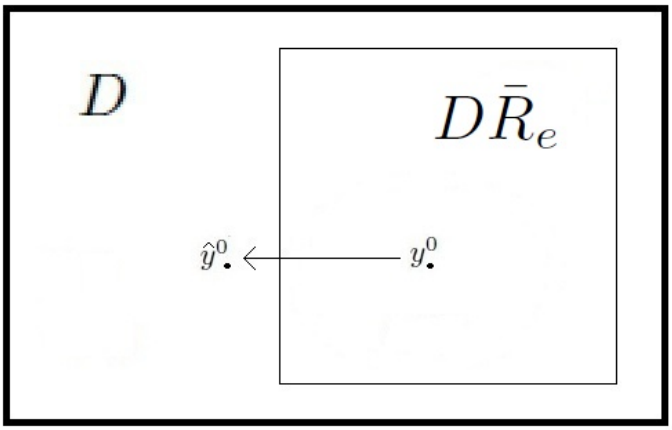

Figura 1.1: Idea básica de EED

Parte I

El método de Fellegi y Holt

Capítulo 2

Características generales

En 1976 Fellegi y Holt (FH de aquí en adelante) publicaron un artículo en el que presentaron un modelo para (un tipo de EED) la EED, que en la literatura se conoce como *el problema de la localización automática del error para errores aleatorios*. En la actualidad este sigue siendo el enfoque más usado para abordar este problema. En este capítulo desarrollaremos el marco teórico de este modelo sin profundizar en la implementación de él, esto debido a que a partir de entonces se han hecho muchas mejoras sobre el algoritmo que FH propusieron, sin embargo, debido a que la idea central de FH sigue siendo parte del desarroyo actual de la EED consideramos importante incluirlo en este trabajo. El modelo de FH se basa en tres criterios, el primero de los cuales es el más importante debido a que representa el paradigma que le dio trascendencia.

1. La información en cada registro debe corregirse para satisfacer todas las reglas cambiando la menor cantidad posible de campos. De esta manera se evita lo más posible manufacturar registros. Al mismo tiempo, si los errores son comparativamente raros, parece más verosímil que de esta manera se identifiquen los registros verdaderamente erróneos. Este criterio parece razonable, particularmente para registros categóricos, porque proporciona la única medida posible de la magnitud de los cambios debidos a la imputación.

2. Cuando se realice imputación, es deseable mantener, lo más posible, la distribución marginal o incluso la conjunta, preferentemente, como se manifiesta en los registros correctos.

3. Las reglas de imputación deben derivarse de las correspondientes reglas de detección sin especificaciones explícitas. Esto asegura que los valores imputados

> no continúen fallando reglas, simplifica la tarea de especificar reglas e imputaciones, simplifica su implementación usando computadoras, facilita la implementación de cambios posteriores en las especificaciones y en general induce un proceso más controlado.

A continuación una breve sección dedicada a contextualizar este modelo y justificar su desarrollo, tómese en cuenta que los argumentos siguientes se pensaron en el contexto histórico de 1976 .

2.1. Consideraciones sobre el modelo de FH

En el marco del problema que se ha descrito en el capítulo anterior, cuando un registro falla alguna de las reglas se tiene una gama de opciones, FH expresan una fuerte preferencia por la siguiente,

> usar la computadora para corregir el registro, usando ciertas reglas para remover las inconsistencias.

A continuación se presentan los argumentos que justifican esta elección.
Otra alternativa sería desechar los registros que fallen cualquiera de las reglas o al menos eliminarlos de los análisis que tomen en cuenta los campos involucrados en las reglas falladas. Hacer esto implicaría asumir, implícitamente, que las inferencias estadísticas no son afectadas por tal eliminación. Esto equivale a suponer que los registros eliminados tienen la misma distribución que los registros satisfactorios. Si suponemos esto tendríamos las hipótesis necesarias para proceder, justificádamnte, por métodos de imputación; que es una postura mucho más acorde con la filosofía básica de la estadística de que a mayor información mejores resultados.En resumen , si suponemos que justificadamente podemos simplemente eliminar los registros erróneos también podríamos en tal caso imputarlos y conservar parte de la información contenida en ellos.
Otra opción es que empleados corrijan los registros manualmente. Sin embargo, si los registros serán corregido usando reglas predeterminadas, usualmente es preferible dejar que las computadoras apliquen las reglas en lugar de los empleados. Al contrario de los empleados, las computadoras aplican las reglas de forma consistente y rápida; más aún, si la detección es realizada por computadora pero la corrección por empleados, usualmente los registros necesitan ser revisados otra vez para asegurarse de que todas las inconsistencias han sido removidas, lo que consume tiempo y dinero.

La mayor desventaja de usar un enfoque que involucra la computadora para ambas etapas de la EED, la detección y la corrección, es la posible complejidad de la programación y la relativa rigidez de los programas computacionales, es decir, la complejidad de los programas es tal que los cambios consumen mucho tiempo, cuestan mucho dinero y son propensos a errores. Frecuentemente las reglas para detección y corrección se establecen de forma más o menos independiente sin que se tenga la certeza de que las correcciones generarán un registro consistente con las reglas de detección.
Por lo anterior el modelo de FH aspira a simplificar la pogramación mediante la caracterización de las reglas de forma uniforme y particularmente simple, de forma que sea fácil incluir reglas adicionales o remover reglas previamente especificadas, de hecho, esta es la aportación principal de este enfoque(ver [1]); la simplicidad con que especificaciones sobre las reglas pueden ser cambiadas sin la necesidad de reprogramar.

Capítulo 3

Detección

Diremos que un registro y *viola* R si $y \in R$, en el caso contrario diremos que y *satisface* R. Con este acuerdo, $D\bar{R}_e$ es el conjunto de todos los puntos en el espacio de datos que violan alguna de las reglas del conjunto de reglas $\bar{R}_e$, es decir,

$$y \text{ es correcto} \Leftrightarrow y \in (D\bar{R}_e)^c = (R^1)^c \cap (R^2)^c \cap \cdots \cap (R^r)^c$$

3.1. Forma normal de las reglas

En esta sección definiremos el tipo de reglas que serán consideradas para poder asumir que se pueden caracterizar de una forma especialmente simple. Esta sección está relacionada con la primera pregunta ; ¿qué campos se deben modificar de forma tal que se modifique lo menos posible el registro y se obtenga un registro correcto? Se usará un ejemplo para introducir las ideas de esta sección.

Ejemplo 3.1. Supóngase que en una encuesta demográfica una de las reglas especificadas por los expertos en la materia es,

> "Si la edad de una persona es $\leq$ 15 años o si él (ella) es estudiante de educación básica, entonces su parentesco con el jefe, o jefa, del hogar no puede ser jefe y su estado civil debe ser soltero ".

Esta regla identifica un subconjunto del espacio de datos D, el objetivo de esta sección es caracterizar este conjunto de una manera particularmente simple. Usemos la notación que hemos desarrollado. Por simplicidad, supongamos que los campos en el espacio de datos D son sólo los que aparecen en esta regla, es decir,

$$D = D_1 \times D_2 \times D_3 \times D_4,$$

donde

- D_1 es el conjunto de posibles edades.
- D_2 es el conjunto de las posibles respuestas en el campo "nivel actual de estudios".
- D_3 es el conjunto de las posibles respuestas en el campo "parentesco con el jefe del hogar".
- D_4 es el conjunto de las posibles respuestas en el campo "estado civil".

Para facilitar la escritura denotaremos como

- $PF_1 = \{y_1 \in D_1 : y_1 \leq 15\} \times D_2 \times D_3 \times D_4$.
- $PF_2 = D_1 \times \{y_2 \in D_2 : y_2 = \text{educación básica}\} \times D_3 \times D_4$.
- $PF_3 = D_1 \times D_2 \times \{y_3 \in D_3 : y_3 \neq \text{jefe del hogar}\} \times D_4$.
- $PF_4 = D_1 \times D_2 \times D_3 \times \{y_4 \in D_4 : y_4 = \text{soltero}\}$.

Ahora podemos expresar el enunciado de manera formal,

> Para que un registro $y \in Dat$ se considere correcto es necesario que satisfaga
> $$y \in PF_1 \cup PF_2 \Rightarrow y \in PF_3 \cap PF_4 \tag{3.1}$$

Recuérdese que en el capítulo anterior se acordó que un registro $y \in Dat$ viola la regla R si $y \in R$, entonces, negando (3.1), obtenemos que

$$y \in R \text{ si } (y \in PF_1 \cup PF_2) \wedge (y \notin PF_3 \vee y \notin PF_4),$$

es decir,

$$R = (PF_1 \cup PF_2) \cap (PF_3^c \cup PF_4^c) \tag{3.2}$$

Aplicando repetidamente la ley distributiva que relaciona las operaciones de unión e intersección de conjuntos, podemos transformar (3.2) en una forma consistente de uniones de intersecciones de conjuntos:

$$\begin{aligned} R = \quad & (PF_1 \cup PF_2) \cap (PF_3^c \cup PF_4^c) \\ = \quad & (PF_1 \cap (PF_3^c \cup PF_4^c)) \cup (PF_2 \cap (PF_3^c \cup PF_4^c)) \\ = \quad & (PF_1 \cap PF_3^c) \cup (PF_1 \cap PF_4^c) \cup (PF_2 \cap PF_3^c) \cup (PF_2 \cap PF_4^c) \end{aligned}$$

si definimos

$$R^1 = PF_1 \cap PF_3^c,\ R^2 = PF_1 \cap PF_4^c,\ R^3 = PF_2 \cap PF_3^c,\ R^4 = PF_2 \cap PF_4^c,$$

podemos escribir

$$R = \bigcup_{i=1}^{4} R^i,$$

Para generalizar el ejemplo anterior usaremos la siguiente notación.

$$\begin{aligned} PF_i &= \prod_{j=1}^{m} F_{ij}, \\ F_{ij} &= D_j \quad \text{si } j \neq i \\ F_{ii} &\subset D_i \end{aligned} \tag{3.3}$$

Así, para expresar que un registro particular y tiene en el campo i un código o valor perteneciente a un subconjunto $F_i \subset D_i$, simplemente decimos que $y \in PF_i \subset D$.

Usaremos reglas como (3.2). De manera precisa, las reglas deben ser tales que

$$R = f(PF_1, PF_2, \ldots, PF_m), \tag{3.4}$$

donde la función f conecta estos subconjuntos mediante las operaciones $\cap$ (intersección de conjuntos) y $\cup$ (unión de conjuntos). Es importante remarcar que (3.4) es una hipótesis que supondremos durante el resto del trabajo y todos los resultados que se obtengan serán aplicables sólo a reglas que cumplan esta condición.

Aplicando repetidamente a $f(PF_1, PF_2, \ldots, PF_m)$ la ley distributiva que relaciona las operaciones de unión e intersección de conjuntos, podemos transformar el lado derecho de (3.4) en una forma consistente de uniones de intersecciones de conjuntos:

$$f(PF_1, PF_2, \ldots, PF_m) = \bigcup_{l=1}^{k} \left(\bigcap_{i \in S_l} PF_i \right) \tag{3.5}$$

Donde $S_l \subset \{1, 2, \ldots, m\}$ es un conjunto apropiado de índices. Claramente $y \in f(PF_1, PF_2, \ldots, PF_m)$ ocurrirá si y sólo si y está en alguno de los conjuntos definidos por los paréntesis en el lado derecho de (3.5). De esta manera obtenemos que la regla definida por (3.4) es equivalente a un conjunto de reglas de la forma:

$$R^l = \bigcap_{i \in S_l} PF_i. \tag{3.6}$$

En lo futuro sera de utilidad expresar las reglas explícitamente como productos de conjuntos, para ello haremos uso del resultado 1.1,

$$R^l = \bigcap_{i \in S_l} PF_i = \bigcap_{i \in S_l} \left(\prod_{j=1}^{m} F_{ij} \right) = \prod_{j=1}^{m} \left(\bigcap_{i \in S_l} F_{ij} \right) \tag{3.7}$$

Nótese que

$$\forall j \in \{1, \ldots, m\} : \quad F_j^l = \bigcap_{i \in S_l} F_{ij} = \begin{cases} D_j \text{ si } j \notin S_l \\ F_{jj} \text{ si } j \in S_l \end{cases}$$

En resumen, podemos escribir (3.6) de la siguiente manera,

$$R^l = \prod_{j=1}^{m} F_j^l,$$

Nótese que con esta convención todas las reglas tienen m factores. Esta última forma de escribir una regla es llamada *forma normal*[1]. De aquí en adelante asumiremos que las reglas explícitas $R \in \bar{R}_e$ están expresadas en forma normal, es decir,

$$R = \prod_{j=1}^{m} F_j \tag{3.8}$$

donde $F_j \subseteq D_j$. Diremos que el campo j *entra* en R si $F_j \neq D_j$ y diremos que es *ajeno* a R en el caso contrario.

Resumen

Una regla R es la intersección de subconjuntos de D, ver (3.6), uno por cada campo entrante. La importancia de esto radica en notar que

> **para sacar a y^0 de R es suficiente cambiar el valor de uno solo de los campos entrantes en R.**

Con las reglas expresadas en forma normal el problema de detección es simplemente determinar el conjunto $Dat \cap D\bar{R}_e$. El resto de la metodología de FH se centra en el problema de localización del error.

Capítulo 4

Corrección

Ahora veremos como resuelve FH el problema de corregir los datos que se determinaron erróneos en la etapa anterior.

4.1. El problema: *localización del error*

Dado un registro que ha sido declarado erróneo, es decir, que viola alguna (algunas) de las reglas, el problema de localización del error (LE de aquí en adelante) es:

> encontrar un conjunto de campos tal que cambiando todos y sólo los valores en este conjunto de campos el registro dado satisfaga todas las reglas, además tal conjunto debe ser minimal, es decir, tal que no exista otro conjunto con esta propiedad y que tenga menos elementos (o menor costo).

Nótese que no se trata de encontrar los valores específicos que se imputarán, únicamente se busca identificar los campos que serán modificados, sin alusión a cual sera el valor que será imputado en él.
Cuando se observa un registro $y^0 \in D\bar{R}_e$, el objetivo es producir un nuevo registro en $(D\bar{R}_e)^c$, cambiando en y^0 la menor cantidad posible de campos. Es decir, para y^0 deseamos asegurar la existencia de un registro sintético $y \in (D\bar{R}_e)^c$ que sea solución del siguiente problema de optimización (ver [3]):

$$\text{mín} \sum_{j=1}^{m} c_j x_j \tag{4.1}$$

sujeto a

$$y \in (D\bar{R}_e)^c \tag{4.2}$$

$$x_j = \begin{cases} 1 & si\ y_j \neq y_j^0 \text{ o si } y_j^0 \text{ es faltante} \\ 0 & \text{cualquier otro caso} \end{cases} \tag{4.3}$$

donde $x = (x_1, x_2, \ldots, x_m)$ es un vector cuyo soporte indica los campos que deben ser modificados en y^0; si $x_j = 1$ el campo j debe ser modificado y permanece como está si $x_j = 0$. El coeficiente c_j en (4.1) es una medida de la confianza en el campo j; c_j es grande cuando es poco probable que el campo j contenga error.

Nota 4.1. De hecho, el paradigma que plantea FH en el criterio número 1 (descrito al comienzo de este capítulo) no contempla la asignación de pesos a los campos (variables), esto correspondería a fijar $c_j = 1$, $j = 1, 2, \ldots, m$. Sin embargo, con el paso del tiempo el paradigma ha sido generalizado a "localizar el conjunto de campos de costo mínimo" en lugar de "localizar el mínimo conjunto de campos".

En efecto, si $y \in (D\bar{R}_e)^c$ entonces y satisface todas las reglas en $\bar{R}_e$, es decir, es un registro correcto. Si además y satisface (4.1), de acuerdo con (4.3), entonces y es diferente a y^0 en un conjunto de campos de costo mínimo. El modelo (4.1)-(4.3) es Llamado *Minimum Weighted Fields to Impute* (MWFI) y abreviado MFI si $c_j = 1, j = 1, \ldots, m$. En [7] se muestra que el modelo (4.1)-(4.3) maximiza la probabilidad de reconstituir el registro correcto.

En [3] se procede como sigue. Un primer paso para resolver MWFI es identificar todas las reglas que viola el registro en cuestión, denotemos tal conjunto como

$$\bar{R}_{y^0} = \{R \in \bar{R}_e : y^0 \in R\}.$$

Para editar y^0 necesariamente debemos modificarlo para que satisfaga todas las reglas de $\bar{R}_{y^0}$, además tenemos la intención de hacerlo modificando un conjunto de campos de costo mínimo. Esto lo podemos modelar con el siguiente problema de set-covering (ver figura 4.1):

$$\text{mín} \sum_{j=1}^{m} c_j x_j \tag{4.4}$$

sujeto a

$$\sum_{j=1}^{m} a_{ij} x_j \geq 1, \quad R^i \in \bar{R}_{y^0} \subset \bar{R}_e \tag{4.5}$$

$$x_j \in \{0, 1\}, \quad j \in \{1, 2, \ldots, m\} \tag{4.6}$$

donde

$$a_{ij} = \begin{cases} 1 & \text{si el campo } j \text{ entra en } R^i \\ 0 & \text{cualquier otro caso} \end{cases} \tag{4.7}$$

En efecto, primero notemos que conocemos y^0 que determina el conjunto $\bar{R}_{y^0}$ que a su vez determina los valores de a_{ij}. Entonces la única variable en el modelo (4.4)-(4.7) es el vector $x = (x_1, x_2, \ldots, x_m)$. Luego, tenemos una matriz $A = (a_{ij})$ en la que cada renglón representa a una regla, que y^0 viola, y cada columna representa un campo. Ahora recordemos que cada regla está en forma normal, es decir, el conjunto de puntos en D que la caracteriza es una intersección, de forma que para sacar a y^0 de tal conjunto basta con cambiar uno solo de los campos entrantes, este razonamiento está expresado en (4.5). En resumen, si x es solución de (4.4)-(4.7) su soporte indica un conjunto de campos tal que cambiándolos todos se obtiene un registro $\bar{y}^0 \in (D\bar{R}_{y^0})^c$ (recuérdese que el objetivo es obtener $\bar{y}^0 \in (D\bar{R}_e)^c$, no sólo $\bar{y}^0 \in (D\bar{R}_{y^0})^c$, ver figura 4.3).

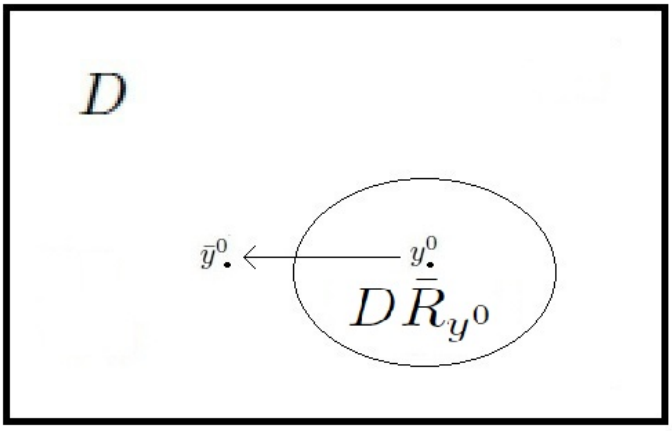

Figura 4.1: Diagrama de (4.4)-(4.7)

Proposición 4.2. *Si x satisface* (4.2) *y* (4.3) *entonces también debe satisfacer* (4.5) *y* (4.6) *(porque $D\bar{R}_{y^0} \subset D\bar{R}_e$, ver figura4.2).*

Demostración. En efecto, sea $y \in (D\bar{R}_e)^c$ y x seg'un (4.3). Si $R^k \in \bar{R}_{y^0}$ entonces y es diferente de y^0 en al menos uno de los campos entrantes en R^k, supongamos que es el campo $q \in \{1 \ldots, m\}$, entonces el valor de $a_{kq}x_q = 1$ y se cumple que:

$$\sum_{j=1}^{m} a_{kj}x_j \geq 1.$$

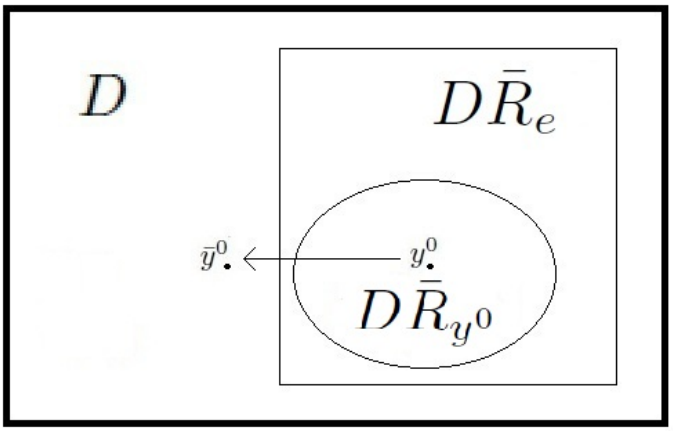

Figura 4.2: (4.2) y (4.3) $\Rightarrow$ (4.5) y (4.6)

□

Sin embargo, satisfacer (4.5) y (4.6) no es condición suficiente para satisfacer (4.2) y (4.3) (ver figura 4.3).

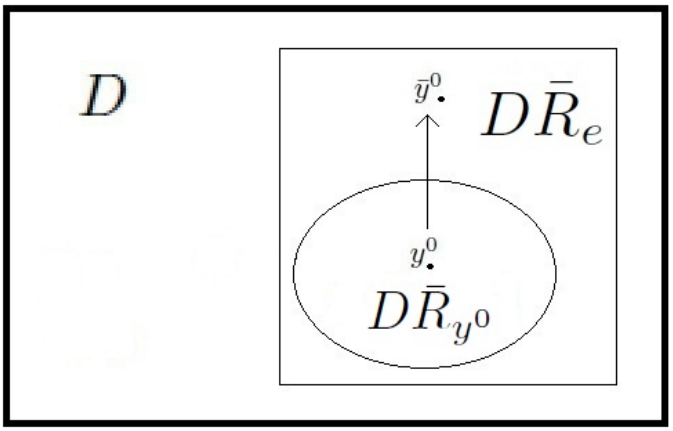

Figura 4.3: (4.2) y (4.3) $\not\Leftarrow$ (4.5) y (4.6)

El siguiente ejemplo servirá para clarificar lo anterior.

Ejemplo 4.3. Supóngase que un cuestionario contiene tres campos: Edad, Estado Civil y Relación con el Jefe del Hogar (en este orden se toman las componentes en el espacio de datos). Se denotará el espacio de datos con

$$D = D_1 \times D_2 \times D_3,$$

donde,

$$\begin{array}{lll} D_1 & = & \{\text{0-14}, 15+\} \\ D_2 & = & \{\text{soltero, casado, divorciado, viudo, separado}\} \\ D_3 & = & \{\text{jefe, esposo del jefe(a), otro}\} \end{array}$$

Supongamos que hay dos reglas,

$$\begin{array}{llllllll} I & R^1 = & \{\text{0-14}\} & \times & D_2 \setminus \{soltero\} & \times & D_3. \\ II & R^2 = & D_1 & \times & D_2 \setminus \{casado\} & \times & \{esposo\}. \end{array}$$

Supongamos que se observa el siguiente registro,

$$y = (y_1 = \text{0-14}, y_2 = casado, y_3 = esposo).$$

Este registro viola la regla I ($y \in R^1$) y satisface la regla II ($y \notin R^2$). En un intento por corregir este registro mediante imputación podríamos considerar cambiar el campo Estado Civil, esto corresponde al modelo (4.4)-(4.7) (pues se altera la menor cantidad posible de campos). Sin embargo, es fácil verificar que dejando Edad y Relación con el Jefe del Hogar como están, cualquier valor posible para Estado Civil resultará en un registro que viole alguna de las dos reglas.

$$\begin{array}{lll} (\text{0-14}, soltero, esposo) & \in & R^2. \\ (\text{0-14}, casado, esposo) & \in & R^1. \\ (\text{0-14}, divorciado, esposo) & \in & R^2. \\ (\text{0-14}, viudo, esposo) & \in & R^1. \\ (\text{0-14}, separado, esposo) & \in & R^1. \end{array}$$

Entonces se debería sospechar que existe algún conflicto escondido entre los actuales valores de Edad y Relación con el Jefe(a) del Hogar, independientemente del Estado Civil. En este simple ejemplo es intuitivamente claro que existe conflicto entre la edad de cero a catorce años y ser jefe del hogar. La existencia de tal conflicto puede ser establecida formalmente como sigue.
Usando las reglas R^1 y R^2 podemos hacer la siguiente inferencia. Supongamos que $y \notin R^1 \wedge y \notin R^2$, entonces,

$$\begin{array}{rl} y_1 = \text{0-14} \Rightarrow & y_2 = Soltero \\ \Rightarrow & y_2 \in D_2 \setminus \{casado\} \\ \Rightarrow & y_3 \in D_3 \setminus \{esposo\}. \end{array}$$

Es decir, si sólo consideramos las reglas R^1 y R^2,

$$(y \text{ es correcto }) \Longrightarrow (y_1 = \text{0-14} \Rightarrow y_3 \in D_3 \setminus \{esposo\})$$

$$\Updownarrow$$

$$\neg(y \text{ es correcto }) \Longleftarrow \neg(y_1 = \text{0-14} \Rightarrow y_3 \in D_3 \setminus \{esposo\})$$

$$\Updownarrow$$

$$y \in D\{R^1, R^2\} \Longleftarrow (y_1 = \text{0-14}) \wedge (y_3 \in \{esposo\})$$

Finalmente, la proposición anterior determina una nueva regla implicada

$$R^3 = \{\text{0-14}\} \times D_2 \times \{Esposo\},$$

implicada significa que

$$y \in R^3 \Rightarrow y \in D\{R^1, R^2\}.$$

Fellegi y Holt [1] propusieron el primer modelo para la EED que garantiza corregir los registros erróneos cumpliendo el criterio de modificar la menor cantidad posible de campos, es decir, establecen condiciones necesarias y suficientes para que modificando el problema de set covering (4.4)-(4.7) se obtengan soluciones al problema de optimización (4.1)-(4.3). FH mostraron que además de las reglas originalmente definidas (explícitas), se deben conocer de manera precisa las reglas implícitas.

4.1.1. Reglas implicadas

El conjunto de reglas explícitas originales $\bar{R}_e$ puede implicar otras reglas que contienen información útil, la cual no aparece en ninguna de las reglas originales.

Definición 4.4. Un subconjunto $R \subset D$ es llamado **regla implicada, o implícita**, por el conjunto de reglas $\bar{R}$ si satisface las dos condiciones siguientes,

1. R es una regla en forma normal,
2. $R \subseteq D\bar{R}$

Nota 4.5. Las reglas implícitas serán redundantes, es decir, añadirlas no cambiará en nada el conjunto de los registros erróneos.

El modelo de FH requiere que se conozcan todas las reglas implícitas relevantes (ver 4.16). El siguiente lema proporciona un método para derivar reglas implicadas por un conjunto de reglas.

Lema 4.6. *Sea* $\bar{R} = \{R^1, R^2, \ldots, R^r\}$ *un conjunto de reglas en forma normal, es decir,*

$$R^l = \prod_{j=1}^{m} F_j^l, \quad l = 1, \ldots, r. \tag{4.8}$$

Para cualquier elección arbitraria de i $(1 \leq i \leq m)$ *considérence los siguientes conjuntos:*

$$\begin{aligned} F_j &= \bigcap_{l=1}^{r} F_j^l, \quad j \neq i \\ F_i &= \bigcup_{l=1}^{r} F_i^l \end{aligned} \tag{4.9}$$

Si los m *conjuntos definidos en* (4.9) *son distintos del conjunto vacío, entonces*

$$R(i, \bar{R}) = \prod_{j=1}^{m} F_j \tag{4.10}$$

define un regla implicada y al campo i *lo llamaremos* ***generador****.*

Demostración. Si los conjuntos F_j son no vacíos, entonces (4.10) tiene la forma que hemos llamado normal. Lo siguiente que debemos verificar es que cualquier registro que viole (4.10) violará también alguna de las reglas en $\bar{R}$. Sea $y = (y_1, \ldots, y_m) \in R$, entonces $y_i \in F_i = \bigcup_{l=1}^{r} F_i^l$ (i es el campo generador), supongamos (sin pérdida de generalidad) que $y_i \in F_i^1$. Pero también, $y_j \in F_j = \bigcap_{l=1}^{r} F_j^l$, $j \neq i$, es decir, $y_j \in F_j^1$, $j \neq i$, por lo tanto $y \in R^1$, lo que completa la demostración. □

Ejemplo 4.7. Supongamos que $\bar{R}$ es el conjunto de las dos reglas del ejemplo 4.3, es decir:

$$\bar{R} = \{R^1, R^2\}$$

donde

$$R^1 = \{Edad = \text{0-14}\} \times \{\text{Casado, Divorciado, Viudo, Separado}\} \times D_3$$

$$R^2 = D_1 \times \{Soltero, Divorciado, Viudo, Separado\} \times \{Esposo\}$$

Usaremos el campo 2, Estado Civil, como generador.

$F_1 = \{\text{0-14}\} \cap D_1 = \{\text{0-14}\}$
$F_3 = D_3 \cap \{Esposo\} = \{Esposo\}$
$F_2 = \{Casado, Divorciado, Viudo, Separado\} \cup \{Soltero, Divorciado, Viudo, Separado\} = D_2.$

Entonces el conjunto

$$R(2, \bar{R}) = \prod_{j=1}^{3} F_j = \{\text{0-14}\} \times D_2 \times \{Esposo\},$$

define una regla implicada, la misma que se obtuvo antes de manera intuitiva.

También tenemos que cualquier regla implicada puede obtenerse por este método.

Teorema 4.8. *Si*

$$R = \prod_{j=1}^{m} F_j \tag{4.11}$$

es una regla que es lógicamente implicada por las reglas explícitas, entonces puede ser generada por el procedimiento del Lema 4.6.

DEMOSTRACIÓN. La afirmación de que la regla (4.11) es implicada lógicamente por las reglas explícitas significa que cualquier registro que viole esta regla también violará al menos una de las reglas explícitas. Considérense unos valores:

$$y_j^0 \in F_j, \quad j = 1, \ldots, m.$$

(Esto es posible porque los conjuntos F_j son no vacíos). El registro y^0 que tiene los valores y_j^0 debe violar al menos una de las reglas explícitas, pues estamos suponiendo que (4.11) es una regla implicada por las reglas explícitas. Selecciónese una regla explícita violada correspondiente a cada posible valor de $y_m^0 \in F_m$. Supóngase que hay K_m de tales reglas correspondientes a todos los valores posibles de $y_m^0 \in F_m$:

$$R^{k_m} = \prod_{j=1}^{m} F_j^{k_m}; \quad k_m = 1, \ldots, K_m$$

Ahora considérese la siguiente regla implicada por las reglas $\bar{R}_m = \{R^{k_m} : k_m = 1, 2, \ldots, K_m\}$, usando el lema 4.9, con el campo m como generador:

$$R(m, \bar{R}_m) = \prod_{j=1}^{m} F_j^q,$$

donde

$$F_j^q = \bigcap_{k_m=1}^{Km} F_j^{k_m}; \quad j = 1, \ldots, m-1 \ ,$$

y

$$F_m^q = \bigcup_{k_m=1}^{Km} F_j^{k_m}.$$

Esta es una regla implicada, porque ninguna de las intersecciones es vacía, ya que F_j^q contiene al menos el valor $y_j^0,\ j = 1, \ldots, m$. Tenemos que $F_m^q \supset F_m$, pues cada $y_m \in F_m$ corresponde a una de las K_m reglas y, por lo tanto, está incluido en uno de los conjuntos $F_m^{k_m}$. Como el valor y_{m-1}^0 fue elegido arbitrariamente del conjunto de valores F_{m-1}, es posible derivar una regla como $R(m, \bar{R}_m)$ para cada valor de $y_{m-1} \in F_{m-1}$. Considérese una de tales reglas generadas para cada valor de $y_{m-1} \in F_{m-1}$:

$$R^{k_{m-1}} = \prod_{j=1}^{m} F_j^{q_{k_{m-1}}}, \quad k_{m-1} = 1, \ldots, K_{m-1} \tag{4.12}$$

donde, por la forma de construir las reglas $R^{k_{m-1}}$, se tiene que

$$F_m^{q_{k_{m-1}}} \supset F_m; \quad k_{m-1} = 1, \ldots, K_{m-1} \tag{4.13}$$

Denotemos $\bar{R}_{m-1} = \{R^{k_{m-1}} : k_{m-1} = 1, \ldots, K_{m-1}\}$. Considérese la regla implicada por las reglas (4.12) usando el campo $m-1$ como generador:

$$R(m-1, \bar{R}_{m-1}) = \prod_{j=1}^{m} F_j^s$$

donde

$$F_j^s = \bigcap_{k_{m-1}=1}^{K_{m-1}} F_j^{q_{k_{m-1}}}; \quad i = 1, 2, \ldots, m-2, m,$$

$$F_{m-1}^s = \bigcup_{k_{m-1}=1}^{K_{m-1}} F_{m-1}^{q_{k_{m-1}}}.$$

Por la forma de construir esta regla

$$F_{m-1}^s \supset F_{m-1}$$

y debido a (4.13) tenemos

$$F_m^s \supset F_m.$$

Continuando de esta forma con $m-2, m-3, \ldots, 1$ podemos generar una regla

$$\prod_{j=1}^{m} F_j^x, \tag{4.14}$$

donde $F_j^x \supset F_j,\ j = 1, \ldots, m$. Finalmente, cada registro que viole (4.11) también violará (4.14) y en consecuencia violará alguna de las reglas explícitas.

□

4.1.2. Caracterización del problema de LE como un problema de set covering

Ahora veremos como modificar (4.4)-(4.7) para que sea equivalente a (4.1)-(4.3), para ello debemos definir algunos conceptos nuevos y obtener algunos resultados previos.

Definición 4.9. Diremos que la regla implicada (4.10) es **esencialmente nueva** si todos los conjuntos F_j son subconjuntos propios de D_j, pero

$$F_i = D_i,$$

donde i es el campo generador.

El conjunto de las reglas explícitas (inicialmente especificadas) junto con todas las reglas esencialmente nuevas es llamado un *conjunto completo de reglas* y es denotado por Ω.
Definimos Ω_K como el subconjunto de Ω que consta de todas las reglas en las que sólo se involucran los primeros K campos, es decir,

$$\Omega_K = \{R \in \Omega : R = \prod_{i=1}^{m} F_i \text{ y } F_i = D_i, i = K+1, K+2, \ldots, m\}.$$

Nota 4.10. En lo que sigue haremos un abuso de notación diciendo que un registro particular $(y_1^0, y_2^0, \ldots, y_K^0) \in D\Omega_K$ en lugar de la expresión formal $\forall y_{K+1}, \ldots, y_m : (y_1^0, \ldots, y_K^0, y_{K+1}, \ldots, y_m) \in D\Omega_K$.

El primer resultado que tenemos es el siguiente, ver [1].

Teorema 4.11. *Si $\bar{y}_i$ ($i = 1, 2, \ldots, K-1$) son, respectivamente, algunos posible valores para los primeros $K-1$ campos y si estos valores satisfacen todas las reglas en Ω_{K-1}, entonces existe algún valor $\bar{y}_K$ de forma que los valores $\bar{y}_i$ ($i = 1, 2, \ldots, K$) satisfacen todas las reglas en Ω_K. Es decir,*

$$(D\Omega_{K-1})^c \neq \emptyset$$

$$\Downarrow$$

$$\forall(\bar{y}_1, \ldots, \bar{y}_{K-1}) \in (D\Omega_{K-1})^c : \exists \bar{y}_K \in D_K : (\bar{y}_1, \ldots, \bar{y}_K) \in (D\Omega_K)^c$$

DEMOSTRACIÓN. Supongamos que el teorema es falso. Es decir,

> **Supongamos que** existen valores para los primeros $K-1$ campos, digamos $\bar{y}_1, \ldots, \bar{y}_{K-1}$, que satisfacen todas las reglas en Ω_{K-1} pero con la propiedad de que, para cada posible valor y_K del campo K, alguna de las reglas en Ω_K es violada.

Supongamos que hay W de tales reglas violadas en Ω_K, una por cada posible valor $y_K \in D_K$,

$$R^w = \prod_{i=1}^{K} F_i^w, \qquad w = 1, 2, \ldots, W.^1 \tag{4.15}$$

Donde

$$\begin{aligned} \bar{y}_i &\in F_i^w, \qquad i = 1, 2, \ldots, K-1, \qquad w = 1, 2, \ldots, W \\ y_K^w &\in F_K^w, \qquad w = 1, 2, \ldots, W. \end{aligned}$$

Nótese que $F_K^w \neq D_K$, porque de lo contrario tendríamos que[2] $R^w \in \Omega_{K-1}$ y $(\bar{y}_1, \ldots, \bar{y}_{K-1}) \in R^w$ en contradicción con la hipótesis original $(\bar{y}_1, \ldots, \bar{y}_{K-1}) \notin D\Omega_{k-1}$. Considérese la siguiente regla implicada por las reglas (4.15), generada por el método del lema 4.6 y usando el campo K como generador[3].

$$R(K, \{R^w : w = 1, \ldots, W\}) = \bigcap_{w=1}^{W} F_1^w \times \cdots \times \bigcap_{w=1}^{W} F_{K-1}^w \times \bigcup_{w=1}^{W} F_K^w. \tag{4.16}$$

Como $\bar{y}_i \in F_i^w$ $(i = 1, 2, \ldots, K-1)$ para toda w, las intersecciones $\cap_{w=1}^{W} F_i^w$ no son vacías. Por lo tanto, (4.16) es la expresión de una regla válida. También, de acuerdo con lo supuesto, cada posible valor y_K del campo K esta incluido en alguno de los conjuntos F_K^w, entonces,

$$\bigcup_{w=1}^{W} F_K^w = D_K.$$

Por lo tanto, (4.16) es una regla en Ω_{K-1}, es decir,

$$R(K, \{R^w : w = 1, \ldots, W\}) = \prod_{i=1}^{K-1} \left(\bigcap_{w=1}^{W} F_i^w \right). \tag{4.17}$$

[1]En $R^w = \prod_{i=1}^{K} F_i^w$ el subíndice i se tomo entre 1 y K porque en este contexto $(R^w \in \Omega_K)$ $F_i^w = D_i$ para $i = K+1, \ldots, m$, .

[2]Se deduce directamente de la definición.

[3]Una vez más, omitimos los factores $k+1, \ldots, m$ porque en este caso $\bigcap_{w=1}^{W} F_i^w = D_i$, $i = k+1, \ldots, m$.

Por la forma en que construimos esta regla tenemos que $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{K-1}$ es una combinación inválida; $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{K-1}) \in R(K, \{R^w : w = 1, \ldots, W\})$. Lo cual contradice la forma en la que elegimos los valores $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{K-1}$. Por lo tanto, la suposición debe ser falsa, y debe existir al menos un valor del campo K, digamos $\bar{y}_K$, que junto con $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{K-1}$, satisface todas las ediciones en Ω_K □

Nótese que implícitamente supusimos que la regla implicada (4.17) esta en Ω, nótese tambien que esta suposición es valida porque (4.17) es una regla esencialmente nueva de acuerdo a nuestra definición. Entonces, la validez del teorema depende de que el conjunto de reglas sea completo con respecto a las reglas esencialmente nuevas, y no necesariamente respecto a todas las reglas implicadas.

Corolario 4.12. *Supóngase que un cuestionario tiene m campos, y asúmase que los campos $1, 2, \ldots, K-1$ tienen valores $\bar{y}_j$ $(j = 1, 2, \ldots, K-1)$ tales que todas las reglas en Ω_{K-1} son satisfechas; entonces existen valores $\bar{y}_j$ $(j = K, \ldots, m)$ tales que los valores $\bar{y}_j$ $(j = 1, 2, \ldots, m)$ satisfacen todas las reglas en Ω. Es decir,*

$$(D\Omega_{K-1})^c \neq \emptyset$$

$$\Downarrow$$

$$\forall(\bar{y}_1, \ldots, \bar{y}_{K-1}) \in (D\Omega_{K-1})^c : \exists \bar{y}_j \in D_j (j = K, \ldots, m) : (\bar{y}_1, \ldots, \bar{y}_m) \in (D\Omega)^c$$

Demostración. La prueba es inmediata mediante aplicaciones repetidas del teorema 4.11. □

Definición 4.13. Sea A un conjunto y $\bar{B}$ una familia de conjuntos. Diremos que A **cubre** a la familia $\bar{B}$ si se cumple que

$$\forall B \in \bar{B} : \quad B \cap A \neq \emptyset.$$

El siguiente es el resultado final que estamos buscando.

Nota 4.14. intercambiar y y y^0 en el siguiente corolario

Corolario 4.15. *Sean Ω un conjunto completo de reglas, $y = (y_1, \ldots, y_m) \in D\Omega$,*

$$\begin{aligned} \bar{R}_y &= \{R \in \Omega : y \in R\}, \\ S_R &= \{i \in \{1, \ldots, m\} : \textit{el campo } i \textit{ entra en } R \in \Omega\}, \\ S_{\bar{R}_y} &= \{S_R : R \in \bar{R}_y\}. \end{aligned}$$

Supóngase que $S \subset \{1, 2, \ldots, m\}$ es un subconjunto de los campos que cubre a la familia $S_{\bar{R}_y}$. Entonces existen valores y_i^0 $(i \in S)$ tales que el registro imputado formado por los valores y_i $(i \notin S)$ junto con los valores y_i^0 $(i \in S)$, satisfacen todas las reglas en Ω.

DEMOSTRACIÓN. Supongamos, sin pérdida de generalidad, que S consiste de los campos $K, K+1, \ldots, m$. Entonces al menos uno de los campos $K, K+1, \ldots, m$ entra en cada regla violada. Por lo tanto no existen reglas violadas que involucren únicamente los campos $1, 2, \ldots, K-1$, así que $(y_1, \ldots, y_{K-1}) \notin D\Omega_{K-1}$, es decir, $(y_1, \ldots, y_{K-1})$ satisface todas las reglas en Ω_{K-1}. Entonces, por el corolario 4.12, existen valores y_i^0 $(i = K, \ldots, m)$ tales que y_i $(i = 1, 2, \ldots, K-1)$ junto con y_i^0 $(i = K, \ldots, m)$ satisfacen todas las reglas en Ω. □

El siguiente corolario es el principal resultado de FH, ya que proporciona una caracterización del problema de localización del error como un problema de set covering.

Teorema 4.16. *Sea $y^0 \in D\bar{R}_e$. Si se consideran todas las reglas de Ω que viola y^0 , es decir, si*

$$\bar{\Omega}_{y^0} = \{R \in \Omega : y^0 \in R\}.$$

Entonces, el problema de optimización MWFI (4.1)-(4.3) *es equivalente al problema de set covering*

$$\text{mín} \sum_{j=1}^{m} c_j x_j \tag{4.18}$$

sujeto a

$$\sum_{j=1}^{m} a_{ij} x_j \geq 1, \quad R^i \in \bar{\Omega}_{y^0} \tag{4.19}$$

$$x_j \in \{0, 1\}, \quad j \in \{1, 2, \ldots, m\} \tag{4.20}$$

donde

$$a_{ij} = \begin{cases} 1 & \textit{si el campo } j \textit{ entra en } R^i \\ 0 & \textit{cualquier otro caso} \end{cases} \tag{4.21}$$

es decir, el conjunto solución de ambos problemas es el mismo.

DEMOSTRACIÓN. Supongamos que $x^0 = (x_1^0, \ldots, x_m^0)$ es solución de MWFI, es decir, mín $\sum_{j=1}^{m} c_j x_j = \sum_{j=1}^{m} c_j x_j^0$ y

$$\exists y \in (D\bar{R}_e)^c : \quad x_j^0 = \begin{cases} 1 & \textit{si } y_j \neq y_j^0 \text{ o si } y_j^0 \text{ es faltante} \\ 0 & \text{cualquier otro caso} \end{cases}$$

Debemos demostrar que

$$\forall R^i \in \bar{\Omega}_{y^0} : \quad \sum_{j=1}^{m} a_{ij} x_j^0 \geq 1.$$

Para simplificar la escritura será de utilidad definir el siguiente conjunto,

$$S_{x^0} = \{j : x_j^0 = 1\}.$$

Sea $R^i \in \bar{\Omega}_{y^0}$. Supongamos que $S_{x^0} \cap S_{R^i} = \emptyset$, entonces

$$(\forall j \in S_{R^i} : \quad y_j = y_j^0) \Rightarrow (y \in R^i),$$

lo cual contradice la hipótesis $y \in (D\bar{R}_e)^c$. Por lo tanto, tenemos que

$$\begin{aligned} \exists l \in S_{x^0} \cap S_{R^i} \quad &\Rightarrow \quad a_{il} = 1 \wedge x_l^0 = 1 \\ &\Rightarrow \quad a_{il} x_l^0 = 1 \\ &\Rightarrow \quad \sum_{j=1}^{m} a_{ij} x_j^0 \geq 1. \end{aligned} \tag{4.22}$$

Por lo tanto, x^0 es solución de (4.18)-(4.21).

Ahora supongamos que x^0 es solución de (4.18)-(4.21). **Debemos demostrar que**

$$\exists y \in (D\bar{R}_e)^c : \quad x_j^0 = \begin{cases} 1 & si\ y_j \neq y_j^0 \text{ o si } y_j^0 \text{ es faltante} \\ 0 & \text{cualquier otro caso} \end{cases} \tag{4.23}$$

Esta parte de la demostración será más artificiosa y para evitar perderse en los detalles se describe el camino que seguiremos, antes de llevarlo a cabo propiamente:

I) Comenzaremos verificando que $(D\bar{R}_e)^c \neq \emptyset$.

II) Después verificaremos que x^0 cumple la condición expresada en (4.23). Esto último se logrará basándose en la siguiente propiedad (que será demostrada)

$$l \in S_{x^0} \Rightarrow \exists R \in \bar{\Omega}_{y^0} : S_{x^0} \cap S_R = \{l\}.$$

Ahora procedemos con la demostración.

I)

Comenzamos verificando que S_{x^0} cubre a la familia $S_{\bar{\Omega}_{y^0}}$, es decir, debemos verificar que se cumple lo siguiente,

$$\forall R^i \in \bar{\Omega}_{y^0} : \quad S_{R^i} \cap S_{x^0} \neq \emptyset.$$

Sea $S_{R^i} \in S_{\bar{\Omega}_{y^0}}$. Como x^0 es solución de (4.18)-(4.21) tenemos que $\sum_{j=1}^{m} a_{ij}x_j \geq 1$, entonces

$$\exists l \in \{1, \ldots, m\} : a_{il}x_l = 1$$

es decir, $a_{il} = 1$ y $x_l^0 = 1$, pero esto significa que $l \in S_{R^i} \cap S_{x^0}$. Por lo tanto, S_{x^0} cubre a la familia $S_{\bar{R}_{y^0}}$. El corolario 4.15 asegura la existencia de valores $z_i (i \in S_{x^0})$ tales que si

$$y_j = \begin{cases} z_j & \text{si} \quad j \in S_{x^0} \\ y_j^0 & \text{si} \quad j \notin S_{x^0} \end{cases}$$

entonces, $y = (y_1, \ldots, y_m) \in (D\bar{R}_e)^c$.

II)

Debemos verificar que si $l \in S_{x^0}$,entonces, $y_j \neq y_j^0$. Por si solo, el corolario 4.15 no garantiza que se cumpla esta última condición, para poder concluir usaremos que x^0 minimiza la función objetivo (4.18).
El razonamiento que usaremos será el siguiente,

> si existiera $R = \prod_{i=1}^{m} F_i \in \bar{R}_{y^0}$ tal que $S_{x^0} \cap S_R = \{l\}$, entonces, $y_l \notin F_l$ (porque de lo contrario se tendria $y \in R$ en contradicción con $y \in (D\bar{R}_e)^c$) y por lo tanto $y_l \neq y_l^0$ (porque $y_l^0 \in F_l$).

En principio podrían darse dos situaciones

1. $\exists R \in \bar{\Omega}_{y^0} : l \in S_R$.
2. $\forall R \in \bar{\Omega}_{y^0} : l \notin S_R$.

Comenzaremos observando que siempre ocurre 1, es decir que

$$S_{x^0} = \bigcup \{S_{x^0} \cap S_R : R \in \bar{\Omega}_{y^0}\}.$$

Para verificarlo, supongamos lo contrario, es decir, supongamos que

$$S_{x^0} \supset \bigcup \{S_{x^0} \cap S_R : R \in \bar{\Omega}_{y^0}\},$$

y definamos $x^1 = (x_1^1, \ldots, x_m^1)$ de la siguiente manera

$$x_j^1 = \begin{cases} 1 & \text{si} \quad j \in \bigcup \{S_{x^0} \cap S_R : R \in \bar{\Omega}_{y^0}\} \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Nótese que $S_{x^1} = \bigcup\{S_{x^0} \cap S_R : R \in \bar{\Omega}_{y^0}\}$, y por lo tanto, $S_{x^0} \supset S_{x^1}$ de lo que se sigue inmediatamente que

$$\sum_{j=1}^{m} c_j x_j^0 > \sum_{j=1}^{m} c_j x_j^1,$$

lo cual contradice la hipótesis original de que mín $\sum_{j=1}^{m} c_j x_j = \sum_{j=1}^{m} c_j x_j^0$. Entonces hemos mostrado que

$$S_{x^0} = \bigcup\{S_{x^0} \cap S_R : R \in \bar{\Omega}_{y^0}\}.$$

Para poder concluir la demostración mostraremos que

$$\exists R \in \bar{\Omega}_{y^0} : S_{x^0} \cap S_R = \{l\}. \tag{4.24}$$

Procederemos por contadicción. Supongamos que

$$l \in S_R \Rightarrow \exists j \neq l : j \in S_R. \tag{4.25}$$

Para llegar a una contradicción usaremos el mismo camino que antes; definiremos un x^2 de tal forma que nos lleve a concluir que x^0 no es solución de (4.4)-(4.7). Definamos

$$x_j^2 = \begin{cases} 1 & \text{si} \quad j \in S_{x^0} \setminus \{l\} \\ 0 & \text{en cualquier otro caso} \end{cases}$$

La suposición (4.25) garantiza que x^2 satisface (4.3), además, claramente

$$\sum_{j=1}^{m} c_j x_j^2 < \sum_{j=1}^{m} c_j x_j^0,$$

en contradicción con la suposición original de que x^0 es solución de (4.4)-(4.7). □

4.1.3. Inconsistencias

Un resultado útil del método de generar reglas implícitas es que cualquier inconsistencia interna en las reglas explícitas es identificada durante el proceso de generación de reglas implicadas.
Recuerdese que supusimos que D esta delimitado en base a los dominios factibles (ver sección 1.3) , es decir, cualquier $y \in D$ es un registro con valores validos.

Definición 4.17. Diremos que un conjunto de reglas es *inconsistente* si de ellas se implica que existen valores permisibles para un campo particular tales que causen automáticamente violaciones a reglas, sin importar que valores tomen los otros

campos, es decir, se tendrá un conjunto inconsistente de reglas si existe una regla implicada R que tenga la forma[4]:

$$R = PF_i$$

para algún i.

4.1.4. Resumen

El objetivo es resolver el problema de localización del error, para ello se debe resolver el siguiente problema de optimización (set covering),

$$\text{mín} \sum_{j=1}^{m} c_j x_j$$

sujeto a

$$\sum_{j=1}^{m} a_{ij} x_j \geq 1, \quad R^i \in \bar{R}_{y^0} \subset \Omega$$

$$x_j \in \{0, 1\}, \quad j \in \{1, 2, \ldots, m\}$$

donde

$$a_{ij} = \begin{cases} 1 & \text{si el campo } j \text{ entra en } R^i \\ 0 & \text{cualquier otro caso} \end{cases}$$

Es importante percatarse que para detectar datos erróneos no hace falta calcular reglas implicadas, basta con verificar el conjunto de reglas explícitas originales. Sin embargo, y este es el aporte de FH, si se desea corregir los datos erróneos cumpliendo el criterio n'umero 1 es necesario obtener las reglas implícitas.

4.2. Imputación

La imputación consiste en reemplazar el valor en un campo del registro dado por el valor para el mismo campo de otro registro que satisfaga todas las reglas de Ω. El método de FH procura mantener la distribución de los datos como se representa en los datos que satisfacen todas las reglas. Supongamos que tenemos un registro $y^0 = (y_1^0, y_2^0, \ldots, y_m^0)$ en el que los primeros K campos serán imputados, siendo este

[4]Ver la definición de PF_i en 3.3

el conjunto mínimo de campos que resuelve el problema de localización del error. Asumiremos de aquí en adelante que se tiene un conjunto completo de reglas. En el artículo original de FH [1] se proponen dos formas de llevar a cabo la imputación: imputación secuencial e imputación conjunta.

4.2.1. Imputación secuencial

Comencemos por imputar el campo K y después imputar sistemáticamente los campos $K-1, K-2, \ldots, 1$.
Considérense todas las reglas en las que el campo K entra pero los campos $1, 2, \ldots, K-1$ son ajenos (si es que existe alguna), denotemos el número de tales reglas como W,

$$R^w = \prod_{i=k}^{m} F_i^w, \qquad w = 1, 2, \ldots, W \tag{4.26}$$

Como el campo K entra en cada regla, F_K^w siempre es distinto del conjunto completo D_K.
De las reglas (4.26) ignoraremos aquellas que y^0 satisfaga en función de sus valores en los campos $K+1, K+2, \ldots, n$, porque estas reglas permanecerán satisfechas sin importar que valor se impute en el campo K. En otras palabras ignoraremos aquellas reglas en (4.26) para las que

$$y_i^0 \notin F_i^w, \quad \text{para al menos un } i = K+1, K+2, \ldots, m,$$

y consideraremos solo aquellas W' reglas tales que

$$y_i^0 \in F_i^w, \quad \text{para todo } i = K+1, K+2.$$

Es decir, de entre las reglas (4.26) consideraremos únicamente aquellas que el registro dado viole dependiendo del valor en el campo K. Si queremos satisfacer todas estas reglas imputando un valor en el campo K, el valor imputado debe satisfacer

$$y_K \in \bigcap_{w=1}^{W'} (F_K^w)^c, \tag{4.27}$$

Esto siempre es posible, porque si el conjunto en (4.27) fuera vacío entonces, según el contrarrecíproco del teorema[5] 4.11, existiría una regla que y^0 viola y en la que sólo

[5]Observemos que el orden en que se toman los campos en la definición de Ω_K es arbitraria, es decir, siempre podemos suponer que los campos que queremos imputar son los primeros K. Así que

entran los campos $K+1, K+2, \ldots, m$, lo cual contradice la elección de los campos $1, 2, \ldots, K$ como solución al problema de localización del error.
Después de imputar y_K, tenemos satisfechas todas las reglas en las que entra el campo K y cualquiera de los campos $K+1, K+2, \ldots, m$ y además los campos $1, 2, \ldots, K-1$ son ajenos. Lo siguiente es considerar todas las reglas en las que el campo $K-1$ entre pero no los campos $1, 2, \ldots, K-2$ y repetimos el procedimiento anterior. Continuamos de esta manera hasta que los campos $1, 2, \ldots, K$ sean imputados, y por la construcción todas las reglas sean satisfechas.

4.2.2. Imputación conjunta

Considérense todas las reglas. Asúmase que se imputarán los campos 1 al K y considérese, como en el método anterior, solo las W'' que el registro dado pueda violar potencialmente dependiendo de los valores imputados en los campos $1, \ldots, K$, es decir,

$$y_i^0 \in F_i^w, \quad \text{para todo } i = K+1, K+2, \ldots, m; \quad w = 1, 2, \ldots, W''.$$

Considérense los conjuntos

$$F_i = \bigcap_{w=1}^{W''} F_i^w, \quad i = k+1, \ldots, m. \tag{4.28}$$

Tenemos que F_i no puede ser vacío, por que $y_i^0 \in F_i^w$, $i = 1, \ldots, W''$. Si elegimos cualquier cualquier registro previamente procesado, es decir, editado e imputado, cuyos valores en los campos $K+1, \ldots, m$ estén en los correspondientes conjuntos (4.28), entonces, como tal registro satisface todas las reglas, sus valores en los campos $1, \ldots, K$ pueden ser usados para la imputación actual y automáticamente se cumplirán todas las reglas.

---

el teorema 4.11 sigue siendo válido si en lugar de tomar los primeros K campos en la definición de Ω_K, tomamos los últimos $m-(K-1)$. Es decir, si

$$\Omega_K = \{R \in \Omega : R = \prod_{i=1}^{m} F_i \text{ y } F_i = D_i, i = 1, 2, \ldots, K-1\}.$$

Entonces, el contrarrecíproco del teorema 4.11, en este caso, se lee

$$\forall y_K \in D_K : (y_K, y_{K+1}^0, \ldots, y_m^0) \in D\Omega_K \Rightarrow (y_{K+1}^0, \ldots, y_m^0) \in D\Omega_{K+1}$$

Es decir, $\exists R \in \Omega_{K+1} : (y_1^0, \ldots, y_{K-1}^0) \in R$.

Nota 4.18. Para poder aplicar esta forma de imputación es necesario contar con un número grande de registros correctos para poder encontrar uno que se ajuste adecuadamente al registro que se está imputando, afortunadamente esta condición se cumple en el caso de datos provenientes de un censo de población.

Parte II

El método de Bruni

Capítulo 5

Características generales

En este capítulo analizaremos otro modelo para la EED presentado en 2005 por Renato Bruni [8]. Como antes, comenzaremos exponiendo algunas de las situaciones que motivaron el desarrollo de este nuevo modelo.
En casos prácticos los métodos basados en las ideas de FH sufren de severas limitaciones computacionales [15, 16], cuya consecuencia es una fuerte limitación en el número de reglas y registros que pueden considerarse.
En particular, se han usado técnicas de programación matemática en propuestas anteriores para los problemas de corrección de datos. Muchos autores han considerado un modelo de *set covering* (principalmente) para el problema de localización del error, requiriendo que se cambie al menos uno de los valores involucrados en cada regla violada. Tal modelo ha sido resuelto por medio de algoritmos llamados *cutting plane* en [3]. Sin embargo, este modelo no representa todas las características del problema, en el sentido de que sus soluciones pueden fallar en ser soluciones al problema de localización.
Por lo tanto, Bruni presenta un procedimiento automático para la detección y corrección de errores aleatorios en datos genéricos usando nuevos modelos matemáticos.

5.1. Descripción del modelo

Las reglas se considerarán de acuerdo con una sintaxis específica y serán automáticamente codificadas como desigualdades lineales. El conjunto de reglas es verificado en busca de inconsistencias y redundancias usando técnicas matemáticas de poliedros. Las inconsistencias son detectadas seleccionando *sistemas inviables irreducibles*, usando una variante del Lema de Farkas, mientras que las redundancias son detectadas encontrando desigualdades implicadas. Después de esta etapa de validación,

las reglas son usadas para detectar registros erróneos. Tales registros son corregidos de forma que satisfagan todas las reglas. El problema de corrección se modela como el problema de minimizar la suma ponderada de los cambios sujeta a limitaciones impuestas por las reglas usando una formulación de programción entera.

Nota 5.1. Bruni evita las reglas redundantes, a diferencia del modelo de FH en el que las reglas redundantes (implícitas) son fundamentales.

Capítulo 6

Detección

6.1. Codificando reglas en desigualdades lineales

Como antes, las reglas sirven para discernir cuando un registro es erróneo. Sin embargo, en el modelo de Bruni las reglas se expresan de modo que identifiquen el conjunto de puntos que las respeta, veamos un ejemplo.

Ejemplo 6.1.

$$R = \{y \in D : \neg(\text{Estado Civil} = \text{casado}) \vee \neg(\text{Edad} < 14)\}.$$

Diremos que y satisface R si $y \in R$. Como hipótesis se pide que cada regla debe expresarse como el conjunto de puntos que satisface una cláusula (una disyunción de proposiciones posiblemente negadas),

$$R = \{y \in D : \bigvee_{j \in \pi} \alpha_j \vee \bigvee_{j \in \nu} \neg\alpha_j \text{ es verdadera respecto a } y\}. \tag{6.1}$$

donde α_j son proposiciones lógicas que únicamente involucran valores del campo j, llamadas *condiciones*, π es el conjunto de índices de sus sus condiciones positivas y ν es el conjunto de índices de sus condiciones negativas. Para facilitar la escritura, usaremos la notación siguiente,

$$R(y) = \begin{cases} 1 \text{ si } y \in R \\ 0 \text{ si } y \notin R \end{cases}$$

Decimos que dos valores y'_j y y''_j son *equivalentes* con respecto a $\bar{R}$ y escribimos $y'_j \cong y''_j$ cuando, para cada par de registros y' y y'' tales que tengan todos sus valores

idénticos excepto por el campo j-ésimo, $R(y') = R(y'')$ para toda $R \in \bar{R}$, es decir,

$$y'_j \cong y''_j \Leftrightarrow \forall R \in \bar{R}\ ;\ R(y_1, \ldots, y'_j, \ldots, y_m) = R(y_1, \ldots, y''_j, \ldots, y_m)$$

Es fácil ver que la anterior es una relación de equivalencia, ya que es reflexiva ($y'_j \cong y'_j$), simétrica ($y'_j \cong y''_j \Rightarrow y''_j \cong y'_j$) y transitiva ($y'_j \cong y''_j$ y $y''_j \cong y'''_j \Rightarrow y'_j \cong y'''_j$). Toda relación de equivalencia en un conjunto genera una partición en él. Lo que tenemos es lo siguiente.

Lema 6.2. *Cada dominio D_j siempre puede ser particionado en n_j subconjuntos*

$$D_j = S_{j1} \cup \ldots \cup S_{jn_j}$$

de tal forma que todos los valores del mismo S_{jw} son equivalentes con respecto a $\bar{R}$. Tales subconjuntos son las clases para la relación de equivalencia introducida.

Demostración. Fijémonos en el campo 1 y consideremos el conjunto de todas las condiciones que tienen que ver con él,

$$\bar{\alpha}_1 = \{\alpha_1\ ; \exists R \in \bar{R} : \alpha_1 \text{ es una condición de } R\}.$$

Para facilitar la escritura, usaremos la notación siguiente,

$$\alpha_1(y_1) = \begin{cases} 1 \text{ si } \alpha_1 \text{ respecto de } y_1 \text{ es verdadera.} \\ 0 \text{ si } \alpha_1 \text{ respecto de } y_1 \text{ es falsa.} \end{cases}$$

Cada $\alpha_1 \in \bar{\alpha}_1$ genera una partición en D_1, denotémosla como $\bar{A} = \{A_0, A_1\}$[1], de tal forma que

$$\forall y_1 \in A_0 : \alpha_1(y_1) = 0$$
$$\forall y_1 \in A_1 : \alpha_1(y_1) = 1$$

Podemos hacer una partición que considere a todas las particiones anteriores, en el sentido de que cada condición $\alpha_1 \in \bar{\alpha}_1$ tenga el mismo valor en todos los elementos de cada clase de tal partición. Para construir tal partición procederemos como sigue. Denotemos la cantidad de elementos de $\bar{\alpha}_1$ como r. Sea $z = (z_1, \ldots, z_r) \in \{0, 1\}^r$ y definimos el siguiente conjunto,

$$S_{1z} = \bigcap_{i=1}^{r} A^i_{z_i}. \tag{6.2}$$

[1] $A_0 \cup A_1 = D_1$ y $A_0 \cap A_1 = \emptyset$

El subíndice 1 denota que S_{1z} es un subconjunto de D_1, el subíndice z denota que a cada $z \in \{0,1\}^r$ le corresponde uno y sólo uno de los conjuntos S_{1z}, el subíndice i enumera los elementos de $\bar{\alpha}_1$ y el subíndice z_i indica cual de los dos conjuntos de $\bar{A}^i$ se considera para la intersección. Claramente de cada $\bar{A}^i$ se elige sólo un conjunto, de manera que, si $S_{1z} \neq \emptyset$, entonces

$$\left(y_1 \in S_{1z} \Rightarrow y_1 \in A^i_{z_i}\right) \Rightarrow \alpha^i_1(y_1) = z_i.$$

Es decir, cada condición $\alpha^i_1 \in \bar{\alpha}_1$ $(i = 1, \ldots, r)$ tiene el mismo valor en todos los elementos de S_{1z}. Para obtener la partición que buscamos tenemos que considerar todos los conjuntos S_{1z} con $z \in \{0,1\}^r$, quitando los que resulten vacíos. Como $\{0,1\}^r$ es finito, podemos enumerar todos los z para los que $S_{1z} \neq \emptyset$, digamos $\{z^1, \ldots, z^{n_1}\} \subset \{0,1\}^r$. Sólo resta verificar que

$$\{S_{1z^w} : w = 1, 2, \ldots, n_1\}$$

es una partición. Veamos primero que

$$\bigcup\{S_{1z^w} : w = 1, 2, \ldots, n_1\} = D_1.$$

Sea $y_1 \in D_1$. Como $\bar{A}^i$ $(i = 1, \ldots, r)$ es una partición se cumple que

$$\forall i \in \{1, \ldots, r\} : \exists g_i \in \{0,1\} : y_1 \in A^i_{g_i},$$

de forma que si definimos $g = (g_1, g_2, \ldots, g_r)$ tenemos que $y_1 \in S_{1g}$. Entonces, $g \in \{z^1, \ldots, z^{n_1}\}$ y por lo tanto

$$y_1 \in \bigcup\{S_{1z^w} : w = 1, 2, \ldots, n_1\}.$$

Entonces, $D_1 \subset \bigcup\{S_{1z^w} : w = 1, 2, \ldots, n_1\}$. La contención

$$\bigcup\{S_{1z^w} : w = 1, 2, \ldots, n_1\} \subset D_1$$

es trivial. Tenemos que

$$\bigcup\{S_{1z^w} : w = 1, 2, \ldots, n_1\} = D_1.$$

Por último debemos verificar que $S_{1z^w} \cap S_{1z^l} = \emptyset$ si $w \neq l$. Si $w \neq l$, entonces

$$\exists i \in \{1, \ldots, r\} : z^w_i \neq z^l_i.$$

Supongamos, sin pérdida de generalidad, que $z_i^w = 0$ y $z_i^l = 1$ Si existiera $y_1 \in S_{1z^w} \cap S_{1z^l}$, entonces, $y_1 \in A_0^i \cap A_1^i$ en contradición con que $\bar{A}^i = \{A_0^i, A_1^i\}$ es una partición. Entonces $S_{1z^w} \cap S_{1z^l} = \emptyset$.

Para tener la misma expresión que en el enunciado del lema simplemente acordamos que $S_{1w} = S_{1z^w}$. Podemos hacer lo mismo para cada campo. Así, obtenemos que los conjuntos S_{jw} son las clases de equivalencia para la relación introducida. □

Ahora podemos introducir las variables para las mencionadas desigualdades lineales; un conjunto de $n = n_1 + \ldots + n_m$ variables binarias $x_{jw} \in \{0, 1\}$, una por cada subconjunto S_{jw}.

La pertenencia de un valor y_j al subconjunto S_{jw} se codifica usando las variables binarias x_{jw},

$$x_{jw} = \begin{cases} 1 & \Leftrightarrow y_j \in S_{jw} \\ 0 & \Leftrightarrow y_j \notin S_{jw} \end{cases}$$

Nótese que cuando la variable x_{jw} es igual a 1, todas las demás variables

$$\{x_{j1}, \ldots, x_{jw-1}, x_{jw+1}, \ldots, x_{jn_i}\}$$

del campo j son iguales a 0. Denotaremos con x el vector formado por todas las componentes x_{jw} como sigue,

$$x = (x_{11}, \ldots, x_{1n_1}, \ldots, x_{m1}, \ldots, x_{mn_m})^T$$

Antes de seguir es conveniente demostrar el siguiente resultado que facilitará la escritura.

Proposición 6.3. *Si $\bar{A}^i$ tiene el significado que tiene en la demostración del lema anterior, entonces*

$$A_1^i = \bigcup \{S_{1w}; A_1^i \cap S_{1w} \neq \emptyset\}$$

Demostración. Sea $y_1 \in A_1^i \subseteq D_1$. Como $\{S_{1w}\}_{w=1}^{n_1}$ es una partición de D_1 debe existir un w' tal que $y_1 \in S_{1w'}$, entonces

$$S_{1w'} \in \{S_{1w}; A_1^i \cap S_{1w} \neq \emptyset\}$$

y por lo tanto

$$y_1 \in \bigcup \{S_{1w}; A_1^i \cap S_{1w} \neq \emptyset\}.$$

Es decir,

$$A_1^i \subseteq \bigcup \{S_{1w}; A_1^i \cap S_{1w} \neq \emptyset\}$$

Sea $S_{1w} \in \{S_{1w}; A_1^i \cap S_{1w} \neq \emptyset\}$. Sea $y_1 \in S_{1w}$. Sabemos que existe $y_1' \in A_1^i \cap S_{1w}$, y por lo tanto $\alpha_1^i(y_1') = 1$ (porque $y_1' \in A_1^i$) y sabemos que

$$\forall y_1, y_1' \in A_1^i : \alpha_1^i(y_1) = \alpha_1^i(y_1'),$$

entonces $y_1 \in A_1^i$, es decir

$$A_1^i \supseteq \bigcup \{S_{1w}; A_1^i \cap S_{1w} \neq \emptyset\}$$

□

Claramente lo mismo se puede hacer en cualquier campo. Lo que nos permite este resultado es lo siguiente. Si α_j es una condición, existe un conjunto apropiado de indices Δ_j tal que

$$\{y_j : \alpha_j(y_j) = 1\} = \bigcup_{w \in \Delta_j} S_{jw}$$

Proposición 6.4. *Cada regla $R \in \bar{R}$ puede asociarse con una desigualdad lineal $R_{\geq}$ sobre la variable x de forma tal que*

$$R = \{y \in D; x \text{ satisface } R_{\geq}\}$$

Demostración. Sea $R \in \bar{R}$. De acuerdo con la hipótesis del modelo, R se puede expresar de la siguiente forma,

$$R = \{y \in D : \bigvee_{j \in \pi} \alpha_j \vee \bigvee_{j \in \nu} \neg \alpha_j \text{ es verdadera respecto a } y\}.$$

Veremos como podemos expresar que la cláusula $\bigvee_{j\in\pi} \alpha_j \vee \bigvee_{j\in\nu} \neg\alpha_j$ es verdadera en otros términos. Usando la proposición (6.3) podemos sustituir "α_j es verdadera respecto a y_j" por $y_j \in \bigcup_{w\in\Delta_j} S_{jw}$, y obtenemos

$$\bigvee_{j \in \pi} \left(y_j \in \bigcup_{w \in \Delta_j} S_{jw} \right) \vee \bigvee_{j \in \nu} \neg \left(y_j \in \bigcup_{w \in \Delta_j} S_{jw} \right).$$

Si ocurre

$$\begin{aligned} \bigvee_{j \in \pi} \left(y_j \in \bigcup_{w \in \Delta_j} S_{jw} \right) \quad &\Leftrightarrow \quad \exists j \in \pi : \exists w \in \Delta_j : y_j \in S_{jw} \\ &\Leftrightarrow \quad \exists j \in \pi : \exists w \in \Delta_j : x_{jw} = 1 \\ &\Leftrightarrow \quad \sum_{j \in \pi} \sum_{w \in \Delta_i} x_{jw} \geq 1. \end{aligned}$$

Si ocurre

$$\begin{aligned}\bigvee_{j\in\nu} \neg\left(y_j \in \bigcup_{w\in\Delta_j} S_{jw}\right) \quad &\Leftrightarrow \quad \exists j \in \nu : \forall w \in \Delta_j : y_j \notin S_{jw} \\ &\Leftrightarrow \quad \exists j \in \nu : \forall w \in \Delta_j : x_{jw} = 0 \\ &\Leftrightarrow \quad \sum_{j\in\nu}\sum_{w\in\Delta_i}(1 - x_{jw}) \geq 1.\end{aligned}$$

En resumen,

$$\bigvee_{j\in\pi} \alpha_j \vee \bigvee_{j\in\nu} \neg\alpha_j \text{ es verdadera respecto a } y$$

$$\Updownarrow$$

$$\sum_{j\in\pi}\sum_{w\in\Delta_i} x_{jw} + \sum_{j\in\nu}\sum_{w\in\Delta_i}(1 - x_{jw}) \geq 1. \tag{6.3}$$

La desigualdad $R_{\geq}$ en el enunciado de la proposición es (6.3). □

Nota 6.5. Para facilitar la escritura, aceptaremos que al escribir x_{jw} se sustituya j por el nombre del campo y w por el conjunto S_{jw}, por ejemplo,

$$x_{\text{estado civil = casado}} \quad \text{ó} \quad x_{edad\in\{0,\ldots,13\}}$$

También aceptaremos escribir la regla

$$R = \{y \in D : \bigvee_{j\in\pi} \alpha_j \vee \bigvee_{j\in\nu} \neg\alpha_j \text{ es verdadera respecto a } y\}$$

simplemente como

$$R = \bigvee_{j\in\pi} \alpha_j \vee \bigvee_{j\in\nu} \neg\alpha_j$$

bajo el entendido de que nos referimos al conjunto de puntos en D que hace la clausula verdadera.

Ejemplo 6.6. Considere la siguiente regla,

$$\neg(\text{Estado Civil} = \text{casado}) \vee \neg(\text{Edad} < 14).$$

Sustituyendo las condiciones, se convierte en la desigualdad lineal:

$$(1 - x_{\text{estado civil = casado}}) + (1 - x_{edad\in\{0,\ldots,13\}}) \geq 1$$

Por lo tanto, del conjunto de reglas $\bar{R}$ se obtiene un conjunto de desigualdades lineales. Cada registro y determina una asignación de valores para las variables introducidas x_{jw}. Por construcción, únicamente las asignaciones de variables correspondientes a registros correctos satisfacen todas las desigualdades lineales. El conjunto de reglas $\bar{R}$ se convierte en un sistema de desigualdades lineales, expresadas en forma compacta como sigue,

$$\begin{cases} Bx \geq b \\ x \in \{0,1\}^n \end{cases} \tag{6.4}$$

si denotamos como l el número total de desigualdades, B es en general una matriz real $l \times n$ y b es un l-vector real. En resumen, aunque un poco informal,[2] un registro y debe satisfacer (6.4) para ser correcto.

Nota 6.7. En lo sucesivo se empleará frecuentemente el producto vector-matriz y matriz-vector alternadamente, acordemos ahora que los vectores, como y y x, se consideran como columnas, es decir, $y \in D^T$ y $x \in (\{0,1\}^n)^T$. Por lo tanto, $y^T \in D$.

6.2. Validación del conjunto de reglas

El siguiente paso es revisar el conjunto de reglas $\bar{R}$ en busca de inconsistencias y redundancias.

Definición 6.8. Una **inconsistencia completa** ocurre cuando cualquier registro es declarado erróneo. Una **inconsistencia parcial** ocurre cuando la inconsistencia se da sólo para valores particulares de campos particulares.

Tales inconsistencias corresponden a propiedades estructurales del sistema (6.4).

Ejemplo 6.9. Una inconsistencia completa muy simple, con las reglas:

(a) Todos deben tener una casa junto al mar.

(b) Todos deben tener una casa junto de campo.

(c) No está permitido tener casa junto al mar y casa de campo.

[2]En realidad no es y sino x la que deben satisfacer el sistema

Inconsistencias más complejas no son tan fáciles de detectar.

$$\begin{cases} x_{\text{casa junto al mar = si}} & \geq 1 \quad (a) \\ x_{\text{casa de campo = si}} & \geq 1 \quad (b) \\ (1 - x_{\text{casa junto al mar = si}}) + (1 - x_{\text{casa de campo = si}}) & \geq 1 \quad (c) \end{cases}$$

Ejemplo 6.10. Una inconsistencia parcial simple, con las reglas:

(a) Se debe tener casa junto al mar si y sólo si el ingreso anual es mayor o igual a 1000.

(b) Se debe tener casa de campo si y sólo si el ingreso anual es mayor o igual a 2000.

(c) No está permitido tener ambas casas, junto al mar y de campo.

Para "ingreso anual" < 2000, esta inconsistencia parcial no causa ningún efecto, pero cada persona con un "ingreso anual" ≥ 2000 es declarada errónea, aunque no lo sea. Tenemos inconsistencia parcial respecto al subconjunto "ingreso anual" ≥ 2000.

$$\begin{cases} (1 - x_{\text{ingreso anual} \geq 1000}) + (1 - x_{\text{casa junto al mar = si}}) & \geq 1 \quad (a) \\ (1 - x_{\text{ingreso anual} \geq 2000}) + (1 - x_{\text{casa de campo = si}}) & \geq 1 \quad (b) \\ (1 - x_{\text{casa junto al mar = si}}) + (1 - x_{\text{casa de campo = si}}) & \geq 1 \quad (c) \end{cases}$$

Teorema 6.11. *Codificando el sistema de reglas como el sistema de desigualdades lineales* (6.4)*, una inconsistencia completa ocurre si y sólo si* (6.4) *es inviable, es decir, no tiene soluciones enteras. Una inconsistencia parcial con respecto al subconjunto S_{jw} ocurre si y sólo si el sistema obtenido añadiendo $x_{jw} = 1$ a* (6.4) *es inviable, es decir, no tiene soluciones enteras.*

Demostración. En el primer caso, por definición de inconsistencia total, el conjunto de reglas no admite ningún registro correcto. Por lo tanto, el correspondiente sistema de desigualdades (6.4) no admite solución, y entonces es inviable. En el segundo caso, por definición de inconsistencia parcial, cualquier registro con valores pertenecientes a cierto subconjunto S_{jw} es declarado erróneo. Entonces, el correspondiente sistema de desigualdades (6.4) no admite ninguna solución con $x_{jw} = 1$. Entonces, añadiendo tal restricción a (6.4), obtenemos un sistema inviable.

□

Cuando existe algún tipo de inconsistencia todas las *reglas conflictivas* deben ser localizadas, esto equivale a seleccionar una parte del sistema (6.4).

Definición 6.12. Un **subsistema inviable irreducible** (SII) es un subconjunto de las desigualdades de un sistema inviable que es él mismo inviable, pero para el que cualquier subconjunto propio es viable.

En este caso estamos interesados en SII *enteros*, es decir, un subconjunto irreducible de ecuaciones que no tenga soluciones enteras. Con este lenguaje, el problema de revisar el conjunto de reglas en busca de inconsistencias es un problema de selecciónde SII enteros, es decir, debemos encontrar todos los SII contenidos en (6.4).

6.2.1. Identificación de Sistemas Inviables Minimales

La estrategia para encontrar los SII en (6.4) será asociarlos con los vértices de un poliedro, para ello necesitaremos algunos resultados previos. La primera etapa es conseguir este resultado para SII reales, y es el contenido de la referencia [2], luego determinaremos condiciones suficientes para aplicarlo a SII enteros.

Lema 6.13. *Sea* $K = \{y \in \mathcal{Q}^n | y^T A = 0, y \geq 0\}$. *Sea* S *el conjunto de los rayos extremos y de* K *que satisfacen* $y^T b < 0$. *Sin pérdida de generalidad se escalan los elementos de* S *de forma que:*

$$\forall y \in S : y^T b = -1.$$

Entonces, S *es igual al conjunto de los puntos extremos de*

$$P = \{y \in \mathcal{Q}^n | y^T A = 0, y^T b \leq -1, y \geq 0\}.$$

Lema 6.14. *(Lema de Farkas)Sean* A *una matriz racional* $m \times n$ y b *un* m*-vector racional. Entonces exactamente una de las siguientes proposiciones se cumple:*

1. *existe* $x \in \mathcal{Q}^n$ *tal que* $Ax \leq b$;
2. *existe* $y \in \mathcal{Q}^m$ *con* $y \geq 0, y^T A = 0,$ y $y^T b < 0$.

Dado un vector racional y, el *soporte* de y denotará los índices de sus componentes no negativas.

Teorema 6.15. *Sean* A *una matriz racional* $m \times n$ y b *un* m*-vector racional. Entonces, los índices de los subsistemas inviables minimales del sistema* $Ax \leq b$ *son exactamente los soportes de los vértices del poliedro*

$$P = \{y \in \mathcal{Q}^n | y^T A = 0, y^T b \leq -1, y \geq 0\}.$$

DEMOSTRACIÓN. Sea K definido como en el Lema 6.13, es suficiente demostrar que los soportes de los rayos extremos de K que satisfacen $y^T b < 0$ indexan los subsistemas inviables minimales del sistema. Sea $A_1 x \leq b_1$ un subsistema de $Ax \leq b$ que es inviable minimal y sea m_1 el número de renglones de A_1. Supongamos que los renglones del sistema $Ax \leq b$ se reordenan de forma tal que los renglones de $A_1 x \leq b_1$ son los primeros m_1. Entonces por el Lema 6.14, existe $w \in \mathcal{Q}^{m_1}$ con $y \equiv (w, \bar{0}) \in K$ y $y^T b < 0$. Nótese que $w > 0$, porque si no fuera así, aplicando el Lema 6.14 al subconjunto de renglones de A_1 correspondientes a los componentes no negativos de w obtendríamos que A_1 no es inviable minimal. Entonces $w > 0$ y el soporte de y indexa los renglones del subsistema inviable minimal $A_1 x \leq b_1$. Ahora debemos mostrar que existe un rayo extremo de K que tiene la misma propiedad. Supongamos que y no es un rayo extremo. Entonces existen $y_i \in K, i = 1, \ldots, k$, y $\lambda_i \geq 0, i = 1, \ldots, k$, con $y = \sum_{i=1}^{k} \lambda_i y_i$. Sin pérdida de generalidad podemos suponer que todos los y_i son rayos extremos de K. Además necesariamente al menos uno de los y_i, digamos y_1, cumple que $y_1 < 0$. Note que como cada $\lambda_i \geq 0$ y cada $y_i \geq 0$ debemos tener que el soporte de y_1 está contenido en el soporte de y. Sin embargo, el soporte de y_1 no puede estar contenido propiamente en el soporte de y porque, como antes, $A_1 x \leq b_1$ no sería inviable minimal. Por lo tanto, el soporte de y_1 indexa los renglones del sistema inviable minimal $A_1 x \leq b_1$. Ahora supongamos que y es un rayo extremo de K y que $y^T b < 0$. Sea $A_1 x \leq b_1$ el subsistema de $Ax \leq b$ indexado por el soporte de y, y sea m_1 el número de renglones de A_1. Sea $w \in \mathcal{Q}^{m_1}$ el vector consistente de los componentes de y distintos de cero. Entonces $w \geq 0, w^T b_1 < 0$, por lo tanto $A_1 x \leq b_1$ es inviable, por el Lema 6.14. Supongamos que el sistema $A_1 x \leq b_1$ no es inviable minimal. Entonces, aplicando nuevamente el Lema 6.14, existe un vector u tal que $u \geq 0, u^T A = 0$, y $u^T b < 0$ cuyo soporte está contenido propiamente en el soporte de y. Sin pérdida de generalidad escalemos u de manera que $y - u \geq 0$. Notemos que $y - u$ tiene al menos un componente distinto de cero. Más aun, $y - u \in K$. Entonces $y = u + (y - u)$, y por lo tanto es una combinación lineal no trivial de elementos de k. Esto contradice la supoción de que y era extremo en K y por lo tanto el sistema $A_1 x \leq b_1$ es inviable minimal. □

Se denotan $\mathbf{0}$, $\mathbf{1}$ y los vectores de ceros y unos de las dimensiones apropiadas. Comencemos por el siguiente resultado [2], el cual es consecuencia del conocido lema de Farkas, en el que se buscan soluciones reales. En resumen, podemos reescribir el teorema anterior como sigue.

Corolario 6.16. *Sea A una matriz real $s \times t$ y a un s-vector. Considere dos sistemas*

de desigualdades lineales de la siguiente forma

$$\begin{cases} Ax \leq a \\ x \in \mathbb{R}^t \end{cases} \tag{6.5}$$

$$\begin{cases} y^T A = \boldsymbol{0} \\ y^T a \leq -1 \\ y \geq \boldsymbol{0} \\ y \in \mathbb{R}^s \end{cases} \tag{6.6}$$

Si (6.6) *es inviable, entonces* (6.5) *es viable. Por el contrario, si* (6.6) *es viable, entonces* (6.5) *es inviable, y además, cada SII de* (6.5) *está dado por el soporte de cada vértice del poliedro* (6.6)*.*

Para poder utilizar el corolario 6.16, consideremos la relajación lineal del sistema (6.4).

$$\begin{cases} -Bx \leq -b \\ x \leq \mathbf{1} \\ -x \leq \mathbf{0} \\ x \in \mathbb{R}^n \end{cases} \tag{6.7}$$

El sistema (6.7) es de la forma (6.5). Las l desigualdades del primer grupo serán llamadas *desigualdades de reglas*. Si definimos

$$A' = \begin{bmatrix} -B \\ I \\ -I \end{bmatrix} \begin{matrix} l \\ n \\ n \end{matrix} \qquad a' = \begin{bmatrix} -b \\ \mathbf{1} \\ \mathbf{0} \end{bmatrix} \begin{matrix} l \\ n \\ n \end{matrix}$$

entonces podemos aplicar el corolario 6.16 al par de sistemas

$$\begin{cases} A'x \leq a' \\ x \in \mathbb{R}^n \end{cases} \tag{6.8}$$

$$\begin{cases} y^T A' = \mathbf{0} \\ y^T a' \leq -1 \\ y \geq \mathbf{0} \\ y \in \mathbb{R}^{l+2n} \end{cases} \tag{6.9}$$

La *restricción* del soporte de un vértice a las desigualdades de reglas denotará los índices de sus componentes distintos de cero dentro de aquellos correspondientes a las desigualdades de reglas.

Teorema 6.17. *Considere dos sistemas de desigualdades lineales en la forma* (6.4) *y* (6.9)*. En este caso, si* (6.9) *es viable, entonces* (6.4) *es inviable y la restricción de cada vértice del poliedro* (6.9) *a desigualdades de reglas contiene un SII entero de* (6.4)*. Por el contrario, si* (6.9) *es inviable, entonces* (6.8) *es viable, pero no se puede decir nada sobre la viabilidad de* (6.4)*.*

Demostración. Primero demostramos que la restricción del soporte de un vértice (6.9) a las desigualdades de reglas contiene un SII entero de (6.4). Supongamos que (6.9) es viable y sea v_1 el vértice encontrado. Entonces, (6.7) es inviable, por el corolario 6.16, y un SII en (6.7), que llamaremos SII_1, está dado por el soporte de v_1. Tal SII_1 en general está compuesto por un conjunto DE_1 de *desigualdades de reglas* y un conjunto RC_1 (posiblemente vacio) de *restricciones de caja* (las que imponen que $0 \leq x_{jw} \leq 1,\ 0 \leq z_i \leq U$). El conjunto de desigualdades DE_1 no tiene soluciones enteras, ya que removiendo RC_1 de SII_1 y sustituyéndolas por las *restricciones enteras* RE_1 más estrictas (las que imponen que $x_{jw} \in \{0,1\},\ z_i \in \{0,1,\ldots,U\}$), SII_1 se mantiene inviable. Entonces, un SII entero está contenido en DE_1. El SII entero puede ser un subconjunto propio de las desigualdades de DE_1, porque, como $SII_1 = DE_1 \cup RC_1$ es inviable minimal, $DE_1 \cup RE_1$ puede no ser minimal: estamos imponiendo las restricciones enteras, más estrictas, en lugar de las restricciones de caja. Por lo tanto, el procedimiento produce un subsistema inviable entero que contiene un SII entero de (6.4), es decir, puede suceder que bajo estas condiciones (6.4) sea inviable y entonces tendríamos al menos un SII entero en (6.4) que no corresponde a ningún vértice de (6.9).

Por otro lado, no todos los SII enteros en (6.4) pueden obtenerse por este método. Esto porque, si (6.9) es inviable entonces (6.7) es viable (corolario 6.16). Cuando imponemos las restricciones enteras, más estrictas, en lugar de las restricciones de caja, nada se puede decir sobre la viabilidad de (6.4). □

El teorema anterior nos dice que si sabemos que el sistema asociado, (6.9), por 6.16 a la relajación lineal (6.7) de nuestro problema original, (6.4),tiene solución real, entonces (6.8) no tiene soluciones reales, y por consiguiente (6.4) no tiene soluciones enteras. Además, aunque no nos dice exactamente que subconjunto de las desigualdades en (6.4) son un SII entero, nos da un subconjunto en el que con certeza esta contenido un SII entero.

Ejemplo 6.18. Supongamos que tenemos un espacio de datos $D = E_1 \times E_2$. Considere un conjunto de reglas $\bar{R}$ sobre dos condiciones lógicas, α_1 y α_2, sobre E_1 y E_2

respectivamente, como sigue,

$$R^1 = (\alpha_1),\ R^2 = (\alpha_2),\ R^3 = (\neg\alpha_1 \vee \neg\alpha_2),\ R^4 = (\alpha_1 \vee \neg\alpha_2).$$

Es posible no percatarse de que $\bar{R}$ contiene una inconsistencia completa. La partición a la que se refiere el lema 6.2 queda de la siguiente forma:

$$E_1 = S_{11} \cup S_{12},$$

donde

$$\forall y_1 \in S_{11} : \alpha_1(y_1) = 1 \quad y \quad \forall y_1 \in S_{12} : \alpha_1(y_1) = 0$$

Análogamente $E_2 = S_{21} \cup S_{22}$. Por lo tanto, $x = (x_{11}, x_{12}, x_{21}, x_{22})^T$. El sistema de ecuaciones (6.4) para este ejemplo es el siguiente,

$$\left\{\begin{array}{rl} x_{11} \geq & 1 \\ x_{21} \geq & 1 \\ (1 - x_{11}) + (1 - x_{21}) \geq & 1 \\ x_{11} + (1 - x_{21}) \geq & 1 \end{array}\right. \Rightarrow \left\{\begin{array}{rrrr} x_{11} & & \geq & 1 \\ & x_{21} & \geq & 1 \\ -x_{11} & -x_{21} & \geq & -1 \\ x_{11} & -x_{21} & \geq & 0 \end{array}\right.$$

Para construir los sistemas (6.8) y (6.9) tenemos

$$A' = \begin{bmatrix} -B \\ I \\ -I \end{bmatrix} \begin{array}{l} l = 4 \\ n = 4 \\ n = 4 \end{array} = \left(\begin{array}{rrrr} -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 \\ -1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{array}\right)$$

$$a' = \begin{bmatrix} -b \\ \mathbf{1} \\ \mathbf{0} \end{bmatrix} \begin{matrix} l = 4 \\ n = 4 \\ n = 4 \end{matrix} = \left(\begin{array}{c} -1 \\ -1 \\ 1 \\ 0 \\ \hline 1 \\ 1 \\ 1 \\ 1 \\ \hline 0 \\ 0 \\ 0 \\ 0 \end{array}\right)$$

De antemano sabemos que (6.8) es inviable asi que debemos resolver (6.9)

$$\begin{cases} y^T A' = \mathbf{0} \\ y^T a' \leq -1 \\ y \geq \mathbf{0} \\ y \in \mathbb{R}^{l+2n} \end{cases}$$

$$\downarrow$$

$$\begin{cases} -y_1 + y_3 - y_4 + y_5 - y_9 = 0 \\ + y_6 - y_{10} = 0 \\ -y_2 + y_3 + y_4 + y_7 - y_{11} = 0 \\ + y_8 - y_{12} = 0 \\ -y_1 - y_2 + y_3 + y_5 + y_6 + y_7 + y_8 \leq -1 \\ y_1,\ y_2,\ y_3,\ y_4,\ y_5,\ y_6,\ y_7,\ y_8,\ y_9,\ y_{10},\ y_{11},\ y_{12} \geq 0 \\ y \in \mathbb{R}^{12} \end{cases}$$

Resolviendo tal sistema se llega al vértice $(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$. Entonces, $\bar{R}$ contiene una inconsistencia, y el conjunto de reglas conflictivas es $\{E_1, E_2, E_3\}$ (esto se verifica fácilmente).

Propiedad del punto entero . Un poliedro, no vacío, que contiene al menos un punto entero, tiene la propiedad del punto entero (PE).

Teorema 6.19. *Si el poliedro* (6.7)*, que es la relajación lineal de* (6.4) *tiene la propiedad del punto entero, entonces satisface lo siguiente. Si* (6.9) *es inviable, entonces* (6.4) *es viable. Por el contrario, si* (6.9) *es viable,* (6.4) *es inviable y cada*

SII entero está dado por la restricción del soporte de cada vértice del poliedro (6.9) *a desigualdades de reglas.*

DEMOSTRACIÓN. Si (6.9) es inviable, (6.7) es viable por corolario 6.16. Como asumimos que (6.7) tiene la propiedad del PE, contiene al menos un punto entero. Como las restricciones de caja se satisfacen en (6.7), este punto entero debe ser tal que $x \in \{0,1\}^n$ y $z \in \{0,1,\ldots,U\}^m$, entonces (6.4) es viable. Por el contrario, si (6.9) es viable, la restricción del soporte de un vértice en (6.9) a las desigualdades de reglas, es decir un conjunto de desigualdades denotado como DE_1, no tiene soluciones enteras por el teorema 6.17. Lo siguiente es probar, por contradicción, que DE_1 es inviable minimal, y por lo tanto es un SII entero. Supongamos que DE_1 no es minimal; entonces existe un conjunto más pequeño DE_1' tal que $DE_1' \cup RE_1$ no tiene soluciones enteras. Por otra parte, por el corolario 6.16, $DE_1' \cup RE_1$ es viable, y como tiene la propiedad del PE, tiene una solución entera, que es una contradicción.

□

Con lo anterior tenemos que, cuando se cumple la propiedad del punto entero podemos resolver el problema de selección de inconsistencias resolviendo un problema de programación lineal.

Suponiendo que estamos en condiciones de aplicar el teorema 6.19 para localizar todos los SII en (6.4) (si es que existe alguno), aún tenemos un problema por resolver. Debemos decidir que reglas modificar o eliminar, de forma que se obtenga un sistema consistente y ademas quisiéramos hacer esto de la forma mas eficiente. Con eficiente queremos decir, minimizando el número de reglas que repararemos. Este enfoque nos lleva a un problema de set covering.
Consideremos una matriz $A = (a_{ij})$ de dimención $n \times r$, donde n es el número de SII contenidos en (6.4) y r es el número de reglas en $\bar{R}$. Claramente cada renglón representa un SII y cada columna una regla en $\bar{R}$, relacionados de la siguiente forma

$$a_{ij} = \begin{cases} 1 & \text{si la regla j forma parte del SII representado por el renglón } i \\ 0 & \text{cualquier otro caso} \end{cases} \tag{6.10}$$

Denotemos

$$\bar{B} = \{bx \geq \beta : bx \geq \beta \text{ es una de las desigualdades en el sistema (6.4)}\}.$$

Ahora podemos establecer formalmente el problema de set covering mencionado,

$$\text{mín} \sum_{j=1}^{r} \delta_j \tag{6.11}$$

sujeto a

$$\sum_{j=1}^{r} a_{ij}\delta_j \geq 1, \quad SII_i \subset \bar{B} \tag{6.12}$$

$$\delta_j \in \{0,1\}, \quad j \in \{1,2,\ldots,r\} \tag{6.13}$$

Sin embargo, desde un punto de vista práctico, son de interés los SII compuestos por un número pequeño de desigualdades de reglas y es posible que no todas ellas sean igualmente preferibles para la composición del SII que se esté seleccionando. Por esto es posible asignar un peso c_k para incluir a cada una de las $l + 2n$ desigualdades de (6.7) en el SII en selección. Lo anterior equivale a asignar un peso a las variables del sistema (6.9). La solución del siguiente programa lineal produce un SII con la compocición de desigualdades deseada.

$$\begin{cases} min\ c^T y \\ y^T A' = \mathbf{0} \\ y^T a' \leq -1 \\ y \geq \mathbf{0} \\ y \in \mathbb{R}^{l+2n} \end{cases} \tag{6.14}$$

Nota 6.20. Recurérdese, que bajo las hipótesis del teorema 6.19, cada vértice del poliedro definido por las restricciones en 6.14 identifica un SII en 6.4 y se sabe por teoría básica de programación lineal que el óptimo de la función objetivo en el problema 6.14 se alcanza en uno de los vértices del poliedro asociado.

Se dice que una regla es *redundante* cuando es lógicamente implicada por otras reglas. Es mejor eliminar las reglas redundantes porque se acelera el proceso y se mantiene el mismo poder de detección.

Lema 6.21. *Una regla es redundante si y sólo si su representación como desigualdad es implicada por la representación en desigualdades de todas las demás reglas.*

Dado un sistema lineal de desigualdades viable S y una sola desigualdad $s^{\geq}$, $s^{\geq}$ es implicada (es decir, el poliedro descrito por S y el poliedro descrito por $S \backslash s^{\geq}$ son iguales) si y solo si añadiendo su negación $s^{<}$ a S produce un sistema inviable.

Una vez se ha verificado el conjunto de reglas el siguiente paso es verificar los registros en busca de *registros erróneos* y^e. Esto se hace verificando que la asignación de valores para x determinada por cada registro y satisfaga el sistema (6.4).

Capítulo 7

Corrección

Cuando se detecta un registro erróneo y^e la corrección consiste, como antes, en cambiar algunos de sus valores para obtener un registro correcto y^c que satisface el sistema (6.4) y esta lo más cerca posible del *registro original* y^0 (desconocido), es decir, el que debería tenerse en ausencia de errores. Los criterios que guían el proceso de imputación son dos:

- modificar lo menos posible los registros erróneos y
- modificar lo menos posible la distribución de frecuencia de los datos.

Para lograr esto se asigna un costo a cada cambio que se introduce en y^e y se asume que (bajo el supuesto de que la presencia de un error es no intencional y menos probable que el valor correcto) la información errónea es el conjunto de valores con costo mínimo tales que al ser modificados permiten satisfacer el sistema (6.4). Para precisar lo anterior se introduce la siguiente notación; cada registro y^e corresponde a una asignación de n valores binarios e_{ij} para las variables x_{ij}. Se tendrá un costo $\hat{c}_{ij} \in \mathbb{R}_+$ por cambiar e^{ij}.

Para llevar a cabo el proceso de imputación se proponen dos metodologías; *localización del error* e *imputación mediante un donante*.

7.1. Localización del error

El problema de localización del error al que nos referimos es el mismo que el descrito en la sección 4.1, sin embargo lo enunciaremos de nuevo con la notación propia de este modelo, el problema es,

encontrar un conjunto H de campos de costo total mínimo tal que y^c puede obtenerse de y^e cambiando, todos y únicamente, los valores de H.

La imputación de los valores específicos para H puede hacerse de forma determinista o probabilística. Este procedimiento causa, en efecto, cambios mínimos en el registro erróneo, sin embargo, por si solo, no ofrece ninguna garantía de que el registro obtenido sea coherente con la distribución conjunta de los datos.

Para modelar el problema se introducen n variables binarias $g_{ij} \in \{0, 1\}$, para representar los cambios que es necesario introducir en e_{ij}.

$$g_{ij} = \begin{cases} 1 \ si \ cambiamos \ e_{ij} \\ 0 \ si \ conservamos \ e_{ij} \end{cases}$$

La minimización del costo total de los cambios puede ser expresada mediante la siguiente función objetivo.

$$\min_{g \in \{0,1\}^n} c^T g, \tag{7.1}$$

donde

$$g = (g_{11}, \ldots, g_{1n_1}, \ldots, g_{m1}, \ldots, g_{mn_m}).$$
$$c = (c_{11}, \ldots, c_{1n_1}, \ldots, c_{m1}, \ldots, c_{mn_m}).$$

Las restricciones en (6.4) están expresadas por medio de la variable x, por lo que usaremos la siguiente relación entre g y x para poder acoplar (6.4) y (7.1).

$$g_{ij} = \begin{cases} x_{ij} & si \ e_{ij} = 0. \\ 1 - x_{ij} & si \ e_{ij} = 1. \end{cases}$$

Usando esto, podemos reescribir la función objetivo (7.1) separando el caso $e_{ij} = 0$ del caso $e_{ij} = 1$. Denotemos con A la siguiente matriz de tamaño $n \times n$ con valores en $\{0, 1\}$.

$$A = diag\{e_{11}, \ldots, e_{1n_1}, \ldots, e_{m1}, \ldots, e_{mn_n}\}$$

El problema de la localización del error puede ahora modelarse como el siguiente problema de optimización.

$$\begin{cases} \min((I - A)c)^T x + (Ac)^T (1 - x) \\ Bx \geq b \\ x \in \{0, 1\}^n \end{cases} \tag{7.2}$$

7.2. Imputación mediante un donante

Un *registro donante* y^d es un registro correcto que deberá ser lo más similar posible a y^0. Esto se logra seleccionando un p^d lo más cercano posible a y^e de acuerdo a una función apropiada δ que asigna un valor v llamado la *distancia* entre y^e y y^d.

$$\delta : D \times D \to \mathbb{R}_+$$

$$(y^e, y^d) \mapsto v$$

También y^d corresponde a una asignación de variables; n valores binarios d_{ij} para las variables x_{ij}. El problema de la imputación mediante un donante es,

> encontrar un conjunto K de campos de costo total mínimo tal que y^c pueda ser obtenido de y^e copiando del donante y^d, todos y únicamente, los valores de de los campos en K.

En general se sabe que esto causa pocas alteraciones a la distribución de frecuencias original, aunque es posible que las alteraciones causadas a los registros no sean mínimas.

La minimización del costo total de los cambios puede ser expresada mediante la siguiente función objetivo.

$$\min_{g \in \{o,1\}^n} c^T g \tag{7.3}$$

Para acoplar (6.4) y (7.3) usaremos la siguiente relación entre g_{ij} y x_{ij}, que esta vez depende de e_{ij} y d_{ij}.

$$g_{ij} = \begin{cases} x_{ij} & \text{si} \quad e_{ij} = 0 \ y \ d_{ij} = 1. \\ 1 - x_{ij} & \text{si} \quad e_{ij} = 1 \ y \ d_{ij} = 0. \\ 0 & \text{si} \quad e_{ij} = d_{ij}. \end{cases} \tag{7.4}$$

Definamos ahora la siguiente matriz D de tamaño $n \times n$ con valores en $\{0, 1\}$.

$$D = diag\{d_{11}, \ldots, d_{1n_1}, \ldots, d_{m1}, \ldots, d_{mn_n}\}$$

Usando esto, podemos reescribir la función objetivo (7.3) y el problema de imputación mediante un donante puede ahora modelarse como sigue.

$$\begin{cases} \text{mín}((I-A)Dc)^T x + (A(I-D)c)^T(1-x) \\ Bx \geq b \\ x \in \{0,1\}^n \end{cases} \tag{7.5}$$

Aclaremos un poco la construcción de (7.5). Queremos encontrar un donante para el registro y^e que hemos observado. Como y^e esta fijo, los valores de la matriz A también están fijos. Buscaremos tal donante entre los registros que respetan todas las reglas en $\bar{R}$, podemos comenzar seleccionando un $y^d \in D \setminus P(\bar{R})$, el que sea, y así dejamos fijos los valores de la matriz D. Una vez que estamos calculando el valor de la función objetivo para el donante particular y^d, el vector x tiene los valores $x = (d_{11}, \ldots, d_{1n_1}, \ldots, d_{m1}, \ldots, d_{mn_n})$. Lo siguiente que haremos es cerciorarnos de que la relación (7.4) establecida entre g y x es cierta. Para ello fijémonos en un solo campo, digamos el primero, lo mismo pasará en todos. Por simplicidad tomemos $n_1 = 3$ (el número de conjuntos en la partición de E_1 inducida por las reglas). Supongamos que

$$\begin{array}{ccc} e_{11} = 1 & e_{12} = 0 & e_{13} = 0 \\ d_{11} = x_{11} = 0 & d_{12} = x_{12} = 0 & d_{23} = x_{13} = 1 \end{array}$$

en este caso $g = (g_{11} = 1, g_{12} = 0, g_{13} = 1, g_{21}, \ldots)$. De acuerdo con (7.4)

$$\begin{array}{rrl} g_{11} = & 1 - x_{11} = 1 - 0 = 1 & (e_{11} = 1 \ y \ d_{11} = 0) \\ g_{12} = & 0 & (e_{12} = d_{12}) \\ g_{13} = & x_{13} = 1 & (e_{13} = 0 \ y \ d_{13} = 1) \end{array}$$

El ejemplo anterior vale como demostración de (7.4) por que considera todos los casos posibles.

Ejemplo 7.1. Supongamos que el siguiente registro es declarado erróneo usando las dos reglas siguientes:

(i) Es imposible que alguien que no tiene auto viva en la ciudad A y trabaje en la ciudad B.

(ii) La edad mínima para conducir es 18 ańos.

(. . . , edad=17, auto=no, ciudad de residencia=A, ciudad en la que trabaja=B, . . .).

Los campos involucrados en las reglas falladas son: auto, ciudad de residencia, ciudad en la que trabaja. Sin embargo, el registro puede corregirse cambiando el valor del

campo ciudad de residencia, o cambiando el valor de ciudad en la que trabaja, o cambiando el valor en los campos edad y auto. Asumiendo que todos estos campos tienen el mismo costo, en este caso la solución de localización del error seria $H_1 = \{ciudad\ de\ residencia\}$ o $H_2 = \{ciudad\ en\ la\ que\ trabaja\}$. Sin embargo suponga que el mejor donador p^d disponible es el siguiente.

(. . . , edad=18, auto=si, ciudad de residencia=A, ciudad en la que trabaja=B, . . .).

La solución de imputación mediante un donador sería en este caso el conjunto $K = \{edad, auto\}$, con un costo mayor que el de H_1 y H_2.

Conclusiones

Hemos revisado dos metodologías para el problema básico de EED. La propuesta de FH es teóricamente sólida y consistente para editar datos categóricos y es la base de muchas investigaciones posteriores; sin embargo, su implementación práctica en datos masivos tiene limitaciones de cómputo.
La alternativa de Bruni ofrece mejoras significativas en su implementación computacional, haciendo uso de las teorías de programación lineal y entera.
Las metodologías para mejorar la calidad de datos estadísticos masivos se han diversificado en los últimos años, gracias al desarrollo tecnológico y en respuesta a necesidades particulares cada ves más complejas. Aún resta adaptar estos métodos al contexto actual de México.
Existe la base teórica que justifica llevar a cabo EED en los datos de la estadística oficial que maneja el INEGI, sin embargo la implementación de estos modelos aun presenta retos importantes; procesar la mayor cantidad de datos en el menor tiempo posible. Para avanzar en esta dirección se requiere mejorar los algoritmos, diseñar mejor la captura de los datos, hacer una buena elección de la metodología adecuada y finalmente hacer las gestiones necesarias en las instituciones pertinentes.

Parte III

Apéndices

Apéndice A

Elementos de programación lineal

A.1. Definiciones básicas

Definición A.1. Un conjunto no vacío $C \subset \mathcal{R}^n$ es llamado un **cono** (convexo) si:

$$x, y \in C \text{ y } \lambda, \mu \geq 0 \Rightarrow \lambda x + \mu y \in C$$

un cono C es *poliédrico* si:

$$C = \{x | Ax \leq 0\}.$$

El cono *generado* por los vectores $x_1, \ldots, x_m$ es el conjunto:

$$\text{cono}\{x_1, \ldots, x_m\} := \{\lambda_1 x_1 + \cdots + \lambda_m x_m | \lambda_1, \ldots, \lambda_m \geq 0\}.$$

Definición A.2. Un conjunto $P \subset \mathcal{R}^n$ es llamado un **poliedro** (convexo) si:

$$P = \{x | Ax \leq 0\}$$

Definición A.3. Un punto x en un conjunto convexo X se llama un **punto extremo** de X, si x no se puede representar como una combinación convexa estricta de dos puntos distintos en X, es decir:

$$\lambda \in (0, 1) : x_1, x_2 \in X : x = \lambda x_1 + (1 - \lambda) x_2 \Rightarrow x = x_1 = x_2$$

Definición A.4. Un **rayo** es una colección de puntos de la forma $\{x_0 + \lambda d | \lambda \geq 0\}$, en donde d es un vector distinto de cero.

En este caso, x_0 se llama el *vértice* del rayo y d es la *dirección* del rayo.

Definición A.5. Dado un conjunto convexo, un vector d, distinto de cero, se llama **dirección del conjunto** si, para cada x_0 en el conjunto, el rayo $\{x_0 + \lambda d | \lambda \geq 0\}$ también pertenece al conjunto.

Definición A.6. Una **dirección extrema** de un conjunto convexo es una dirección del conjunto que no se puede representar como una combinación lineal positiva de dos direcciones distintas del conjunto.

Se dice que dos vectores d_1 y d_2 son *distintos*, o no equivalentes, si d_1 no se puede representar como un múltiplo positivo de d_2.

Definición A.7. Cualquier rayo que está contenido en el conjunto convexo y cuya dirección es una dirección extrema se llama **rayo extremo**.

A.2. Programación lineal

El término *programación lineal* (de aquí en Adelante PL) se refiere al problema de maximizar o minimizar un funcional lineal sobre un poliedro, por ejemplo:

$$\text{máx}\{cx \mid Ax \leq b\},\ \text{mín}\{cx \mid x \geq 0;\ Ax \leq b\}.$$

La idea de programación lineal está presente en el trabajo de Fourier, sin embargo, su creación como disciplina así como el reconocimientó de su importancia llego en la década de 1940 con el trabajo de Dantzing, Kantorovich, Koopmans y von Neumann. El siguiente es un resultado importante conocido como *teorema de dualidad de la programación lineal*, debido a von Neumann (1947), Gale, Kuhn y Tucker (1951) como corolario del teorema fundamental de las desigualdades lineales.

Corolario A.8. *Sean A una matriz y b,c vectores. Entonces*

$$\text{máx}\{cx \mid Ax \leq b\} = \ \text{mín}\{yb \mid y \geq 0;\ yA = c\} \tag{A.1}$$

siempre que ambos conjuntos en (A.1) *no sean vacíos.*

Existen varias formas equivalentes de expresar un problema de programación lineal. Por ejemplo, todos los siguientes tipos de problemas son equivalentes en el sentido de que pueden reducirse unos a otros.

(i) $\text{máx}\{cx | Ax \leq b\}$

(ii) $\text{máx}\{cx | x \geq 0,\ Ax \leq b\}$

(iii) máx$\{cx|x \geq 0,\ Ax = b\}$

(iv) mín$\{cx|Ax \geq b\}$

(v) máx$\{cx|x \geq 0,\ ax \leq b\}$

(vi) máx$\{cx|x \geq 0,\ ax = b\}$

A.3. El método simplex

La idea del método simplex es recorrer el poliedro subyacente a un programa lineal, ir de vértice en vértice a trevés de las aristas hasta que se alcance el vértice óptimo. Esta idea se debe a Fourier y fue algebraicamente mecanizada por Dantzing. A continuación se describe el método simplex, incluyendo la regla de Bland.

El método simplex si se conoce un vértice

Sea A una matriz $m \times n$ con entradas racionales y $b \in \mathcal{Q}^m$. Primero supongamos que deseamos resolver:

$$\text{máx}\{cx \mid Ax \leq b\} \tag{A.2}$$

y que conocemos un vértice x_0 de la región factible $P := \{x \mid Ax \leq b\}$. Asumimos que las desigualdades en $Ax \leq b$ están ordenadas:

$$a_1x \leq \beta_1, \ldots, a_mx \leq \beta_m. \tag{A.3}$$

Elijamos un subsistema $A_0x \leq b_0$ de $Ax \leq b$ tal que $A_0x_0 = b_0$ y A_0 es no singular, podemos hacer esto porque si x_0 es un vértice debe satisfacer al menos n de las restricciones con igualdad. Determine u tal que $c = uA$ y u es 0 en las componentes fuera de A_0, es decir u se obtiene a'nadiendo ceros a cA_0^{-1}.
Cso 1. $u \geq 0$. Entonces x_0 es óptimo, porque:

$$cx_0 = uAx_0 = ub \geq \text{mín}\{yb \parallel y \geq 0;\ yA = c\} = \text{máx}\{cx \mid Ax \leq b\}. \tag{A.4}$$

Asi que al mismo tiempo, u es una solución óptima para el problema dual de (A.2).
Caso 2. $u < 0$. Elíjase el menor de los índices i^* para los cuales u tiene componente negativa v_{i^*}. Sea y el vector tal que $ay = 0$ para cada renglón de A_0 si $a \neq a_{i^*}$ y $a_{i^*}y = -1$, es decir, y es la columna apropiada de $-A_0^{-1}$. Obsérvece que para $\lambda \geq 0,\ x_0 + \lambda y$ atraviesa una arista de P, atraviesa un rayo de P ó está fuera de P para toda $\lambda > 0$. En efecto, supongamos sin perdida de generalidad que b_0 corresponde a las primeras n componentes de b

$$A(x_0 + \lambda y) = (\beta_1, \ldots, \beta_{i^*} - 1, \ldots, \beta_n, a_{n+1}x_0 + \lambda a_{n+1}y, \ldots, a_m x_0 + \lambda a_m y)$$

y sabemos que

$$a_i x_0 \leq \beta_i,\ i = n+1, \ldots, m.$$

Por lo tanto, depende de como sea $a_i y$ el que se respete $a_i x_0 + \lambda a_i y \leq \beta_i$.

Más a'un:

$$cy = uAy = -v_{i^*} > 0. \tag{A.5}$$

En consecuencia, el caso 2 se divide en dos sub casos:

Caso 2a. $ay \leq 0$ para cada renglón a de A. En este caso $x_0 + \lambda y$ atraviesa un rayo de P. En efecto:

$$a_i x_0 + \lambda a_i y < \beta_i,\ i = n+1, \ldots, m;\ \forall \lambda > 0.$$

Entonces $x_0 + \lambda y$ está en P para toda $\lambda \geq 0$ y por lo tanto el máximo (A.2) no está acotado, porque usando (A.5) tenemos:

$$c(x_0 + \lambda y) = cx_0 - \lambda v_{i^*} \to \infty,\ \text{si } \lambda \to \infty$$

Caso 2b. $ay > 0$ para alguna columna a de A. En este caso $x_0 + \lambda y$ atraviesa una arista de P o está fuera de P para toda $\lambda > 0$.

Sea λ_0 la mayor λ tal que $x_0 + \lambda y$ pertenece a P, es decir:

$$\lambda_0 := \text{mín} \left\{ \frac{\beta_j - a_j x_0}{a_j y} \mid j = 1, \ldots, m;\ a_j y > 0 \right\}.$$

Sea j^* el menor índice en que se alcanza este mínimo. Sea A_1 la matriz que se obtiene de A_0 reemplazando el renglón a_{i^*} por a_{j^*}, y sea $x_1 := x_0 + \lambda_0 y$. Así, $A_1 = b_1$, donde b_1 es la parte de b correspondiente a A_1. Comience el proceso de nuevo con $A_0,\ x_0$ reemplazados por $A_1,\ x_1$.

Teorema A.9. *El método anterior termina.*

Apéndice B

Lema de Farkas

Este capítulo está dedicado a demostrar el conocido Lema de Farkas, por lo que se desarrollan algunos conceptos y resultados previos.

Definición B.1. B.1. Resultados previos

Un subconjunto $\mathcal{C} \subset \mathbb{R}^n$ se dice convexo, si y sólo si:

$$\forall x, y \in \mathcal{C},\ \forall \lambda \in [0,1],\ \lambda x + (1-\lambda)y \in \mathcal{C}.$$

Lema B.2. *(Ley del paralelogramo) Para todos $a, b \in \mathbb{R}^n$, se cumple:*

$$\|a+b\|^2 + \|a-b\|^2 = 2\|a\|^2 + 2\|b\|^2.$$

Teorema B.3. *(de la Proyección) Sea $\mathcal{C} \subset \mathbb{R}^n$ un conjunto convexo, cerrado, no vacío y supongamos que $z \notin \mathcal{C}$. Entonces existe un 'unico punto $\bar{x} \in \mathcal{C}$,que llamaremos la proyección (ortogonal) de z en $\mathcal{C}$, y que tiene la mínima distancia entre los puntos de $\mathcal{C}$ y z, es decir:*

$$\|\bar{x} - z\| = \min_{x \in \mathcal{C}} \|x - z\|.$$

Además, $\bar{x}$ es el vector que cumple que la mínima distancia si y sólo si:

$$(x - \bar{x}, \bar{x} - z) \geq 0,\ \forall x \in \mathcal{C}$$

Demostración. Primero probaremos la existencia. Denotemos por $\gamma = \inf_{x \in \mathcal{C}} \|x - z\| \geq 0$, el cual existe por ser un conjunto acotado inferiormente. Existe también una sucesión $\{x_n\} \subset \mathcal{C}$, tal que;

$$\lim_{n \to \infty} \|x_n - z\| = \gamma.$$

Vamos a demostrar que $\{x_n\}$ es una sucesión de Cauchy. Sean $m,\ p$ enteros cualquiera y apliquemos la ley del paralelogramo a los vectores:

$$a = x_m - z,$$
$$b = x_{m+p} - z.$$

Tendremos:

$$\|x_m + x_{m+p} - 2z\|^2 + \|x_m - x_{m+p}\|^2 = 2\|x_m - z\|^2 + 2\|x_{m+p} - z\|^2,$$

es decir:

$$\|x_m - x_{m+p}\|^2 = 2\|x_m - z\|^2 + 2\|x_{m+p} - z\|^2 - 4\left\|\frac{x_m + x_{m+p}}{2} - z\right\|^2.$$

Por la convexidad:

$$\frac{x_m + x_{m+p}}{2} \in \mathcal{C},$$

y por definición del ínfimo:

$$\left\|\frac{x_m + x_{m+p}}{2} - z\right\| \geq \gamma,$$

con lo cual:

$$\|x_m - x_{m+p}\|^2 \leq 2\|x_m - z\|^2 + 2\|x_{m+p} - 2\|^2 - 4\gamma^2.$$

El miembro derecho de la 'ultima desigualdad tiende a cero si $m \to +\infty$. Esto implica que la sucesión $\{x_n\}$ es de Cauchy y por lo tanto converge a un punto $\bar{x} \in \mathcal{C}$, porque $\mathcal{C}$ es cerrado. Por la continuidad de la norma tenemos:

$$\lim_{n\to\infty} \|x_n - z\| = \|\bar{x} - z\| = \gamma.$$

Lo que significa que en $\bar{x}$ se alcanza la mínima distancia de z a $\mathcal{C}$.
Ahora ocupémonos de la unicidad. Supongamos que hay dos puntos $x_1, x_2 \in \mathcal{C}$ donde se alcanza el mínimo γ. Entonces:

$$\gamma \leq \left\|\frac{x_1 + x_2}{2} - z\right\| = \frac{1}{2}\|(x_1 - z) + (x_2 - z)\| \leq$$
$$\leq \frac{1}{2}\|x_1 - z\| + \frac{1}{2}\|x_2 - z\| = \gamma.$$

Tenemos la igualdad en todas partes y por lo tanto:

$$\|(x_1 - z) + (x_2 - z)\| = \|x_1 - z\| + \|x_2 - z\|,$$

de donde se deduce(por la condición de igualdad de la desigualdad triangular) que existe $\mu \in \mathbb{R}$ tal que:

$$(x_1 - z) = \mu(x_2 - z).$$

Como por hipótesis $\|x_1 - z\| = \|x_2 - z\| = \gamma$ se tiene que $|\mu| = 1$. No puede ser que $\mu = -1$ por que tendríamos:

$$x_1 - z = z - x_2 \Rightarrow z = \frac{x_1 + x_2}{2} \in \mathcal{C},$$

en contradicción con la hipótesis de que $z \notin \mathcal{C}$. Por lo tanto $\mu = 1 \Rightarrow x_1 = x_2$.

Por último demostremos la caracterización. Supongamos primero que $\bar{x} \in \mathcal{C}$ es el punto de mínima distancia a z. Entonces.

$$\|\bar{x} - z\| \leq \|x - z\|, \ \forall x \in \mathcal{C}.$$

Si $x \in \mathcal{C}$ y $\lambda \in (0, 1]$, tenemos por la convexidad que $\bar{x} + \lambda(x - \bar{x}) \in \mathcal{C}$, de donde:

$$\|\bar{x} + \lambda(x - \bar{x}) - z\|^2 \geq \|\bar{x} - z\|^2,$$

además:

$$\|\bar{x} + \lambda(x - \bar{x}) - z\|^2 = \|\bar{x} - z\|^2 + 2\lambda(\bar{x} - z, x - \bar{x}) + \lambda^2 \|x - \bar{x}\|^2,$$

Con lo que se obtiene

$$\begin{gathered} 0 \leq \lambda^2 \|x - \bar{x}\|^2 + 2\lambda(\bar{x} - z, x - \bar{x}) \Rightarrow \\ (x - \bar{x}, \bar{x} - z) \geq -\tfrac{\lambda}{2} \|x - \bar{x}\|^2 \to 0, \text{ si } \lambda \to 0. \end{gathered}$$

Para la segunda parte, supongamos que se cumple:

$$(x - \bar{x}, \bar{x} - z) \geq 0, \ \forall x \in \mathcal{C}.$$

Entonces, para todo $x \in \mathcal{C}$, tendremos:

$$\|x - z\|^2 = \|x - \bar{x} + \bar{x} - z\|^2 = \|x - \bar{x}\|^2 + \|\bar{x} - z\|^2 + 2(x - \bar{x}, \bar{x} - z).$$

Porque $\|x - \bar{x}\|^2 \geq 0$ y $2(x - \bar{x}, \bar{x} - z) \geq 0$ se obtiene la desigualdad:

$$\|x - z\|^2 \geq \|\bar{x} - z\|^2, \ \forall x \in \mathcal{C}$$

□

Definición B.4. Un hiperplano $\mathcal{H} = (p, \alpha)$ es un subconjunto de $\mathbb{R}^n$, definido por el vector $p \in \mathbb{R}^n,\ p \neq 0$, y un escalar $\alpha \in \mathbb{R}$, en la forma siguiente:

$$\mathcal{H} = \{x \in \mathbb{R}^n \ : \ p^T x = \alpha\}.^1$$

Definición B.5. Un hiperplano $\mathcal{H} = (p, \alpha)$ se dice que separa dos conjuntos $S_1, S_2 \subset \mathbb{R}^n$ si:

$$p^T x_1 \geq \alpha,\ \forall x \in S_1,$$
$$p^T x_2 \leq \alpha,\ \forall x \in S_2.$$

Se dice que $\mathcal{H}$ los separa estrictamente si alguna de las desigualdades es estricta.

Teorema B.6. *Sea $\mathcal{C} \subset \mathbb{R}^n$ un conjunto convexo, cerrado, no vacío y supongamos que $z \notin \mathcal{C}$. Entonces existe un hiperplano $\mathcal{H} = (p, \alpha)$, donde puede tomarse p con $\|p\| = 1$, tal que $\mathcal{H}$ separa estrictamente $\mathcal{C}$ de z, es decir:*

$$p^T x \leq \alpha,\ \forall x \in \mathcal{C},$$
$$p^T z > \alpha.$$

Demostración. Se cumple la hipótesis del teorema de la proyección. Existe $\bar{x} \in \mathcal{C}$, 'unico, tal que:

$$(x - \bar{x}, \bar{x} - z) \geq 0,\ \forall x \in \mathcal{C},$$

lo que es equivalente a :

$$(x, \bar{x} - z) \geq (\bar{x}, \bar{x} - z) \Leftrightarrow$$
$$-(x, z - \bar{x}) \geq -(\bar{x}, z - \bar{x}),\ \forall x \in \mathcal{C}.$$

Por otro lado, tenemos que:

$$\|z - \bar{x}\|^2 = (z - \bar{x}, z - \bar{x}) = (z, z - \bar{x}) - (\bar{x}, z - \bar{x}),$$

y por la desigualdad anterior se obtiene:

$$\|z - \bar{x}\|^2 \leq (z, z - \bar{x}) - (x, z - \bar{x}) = (z - x, x - \bar{x}),\ \forall x \in \mathcal{C}.$$

Definamos $p = (z - \bar{x})$. Entonces, podemos escribir la 'ultima desigualdad como:

$$\|p\|^2 \leq p^T z - p^T x,\ \forall x \in \mathcal{C},$$

[1] Se utilizara $y^T x$ para denotar el producto escalar de y con x, en otros casos se utiliza la notación con paréntesis $(y, x - z)$ para denotar producto escalar

y por lo tanto, la función lineal $x \longmapsto p^T x$ está acotada superiormente sobre el conjunto $\mathcal{C}$. Definiendo:

$$\alpha = \sup\{p^T x,\ x \in \mathcal{C}\},$$

tendremos:

$$p^T x \leq \alpha, \forall x \in \mathcal{C},$$
$$p^T z \geq \alpha + \|p\|^2 > \alpha.$$

Como $\|p\| \neq 0$, si redefinimos $\hat{p} = \frac{p}{\|p\|},\ \hat{\alpha} = \frac{\alpha}{\|p\|}$, tendremos

$$\hat{p}^T x \leq \hat{\alpha}, \forall x \in \mathcal{C},$$
$$\hat{p}^T z \geq \hat{\alpha} + \frac{1}{\|p\|} > \hat{\alpha},$$

lo que indica que podemos tomar siempre $\|p\| = 1$. □

Lema B.7. *Sea A una matriz $m \times n$. El conjunto:*

$$M = A^T(\mathbb{R}^m_+) = \{z \in \mathbb{R}^n\ :\ \exists \lambda \in \mathbb{R}^m_+,\ z = A^T\lambda\}$$

es un conjunto cerrado en $\mathbb{R}^n$.

Demostración. Si $A_{i\bullet} \in (\mathbb{R}^n)^T$ designa la i-ésima fila de la matriz A, tenemos que:

$$M = \left\{ \sum_{i=1}^{m} \lambda_i A_{i\bullet}^T\ :\ \lambda_i \geq 0,\ i = 1, \ldots, m \right\}.$$

Primero supongamos que los vectores $\{A_{i\bullet}^T\}$ $i = 1, \ldots, m$ son linealmente independientes, es decir, el rango de la matriz A es m y en consecuencia la $m \times m$ matriz AA^T es no singular. Para probar que es cerrado sea $x \in \bar{M}$, existe una sucesión $\{x_n\} \subset M$ tal que $x_n \longrightarrow x$. También existiría una sucesión $\{\lambda_n\} \subset \mathbb{R}^m_+$ tal que $x_n = A^T\lambda_n$, para todo $n \in \mathbb{N}$. Entonces:

$$\lambda_n = (AA^T)^{-1}Ax_n, \forall n \in \mathbb{N}$$

y por la convergencia de $\{x_n\}$ se tiene que $\{\lambda_n\}$ es convergente a $\lambda = (AA^T)^{-1}Ax$, porque $(AA^T)^{-1}A$ es una función continua. Además:

$$\lambda_n \geq 0,\ \forall n \in \mathbb{N} \Rightarrow \lambda \geq 0,$$

y tomando límite en la igualdad $x_n = A^T\lambda_n$ tenemos que $x = A^T\lambda$ lo que significa que $x \in M$ y M es cerrado. Ahora supongamos que los vectores $\{A^T_{i\bullet}\}$ $i = 1, \ldots, m$ son linealmente dependientes. Designemos por $\mathbf{A}$ el conjunto formado por estos vectores y si $\mathcal{A} \subset \mathbf{A}$ contiene $k \leq m$ vectores de $\mathbf{A}$, designemos por $M(\mathcal{A})$ el conjunto:

$$M(\mathcal{A}) = \left\{ \sum_{i=1}^{k} \lambda_i A^T_{i\bullet} \ : \ A^T_{i\bullet} \in \mathcal{A}, \lambda_i \geq 0, \ i = 1, \ldots, k \right\},$$

es decir, son todas las posibles combinaciones lineales positivas que se forman con los vectores del conjunto $\mathcal{A}$. Con estas notaciones se tiene $M(\mathbf{A}) = M = A^T(\mathbb{R}^m_+)$. Denotemos ahora por Φ la clase de todos los subconjuntos $\mathcal{A}$ de $\mathbf{A}$ que están formados por vectores linealmente independientes:

$$\Phi = \{\mathcal{A} \subset \mathbf{A} \ : \ \text{Todos los vectores de } \mathcal{A} \text{ son } l.i\}.$$

Es fácil ver que Φ es un conjunto finito y que tiene a lo sumo 2^m elementos, porque es subconjunto del conjunto potencia de $\mathbf{A}$. Vamos a demostrar que:

$$M = M(\mathbf{A}) = \bigcup_{\mathcal{A} \in \Phi} M(\mathcal{A}).$$

Cada uno de los conjuntos $M(\mathcal{A})$ es cerrado en virtud de la demostración anterior, ya que los vectores de $\mathcal{A}$ son $l.i$. De ese modo, se puede expresar M como unión finita de conjuntos cerrados y es por tanto cerrado. La inclusión $\bigcup_{\mathcal{A}\in\Phi} M(\mathcal{A}) \subset M$ es evidente pues, por definición, cada $M(\mathcal{A}) \subset M$. Por otro lado, sea $z \in M$, entonces existen $\lambda_i \geq 0, \ i = 1, \ldots, m$ tales que:

$$z = \sum_{i=1}^{m} \lambda_i A^T_{i\bullet}.$$

Eliminando los $\lambda_i = 0$ nos queda expresado x, en general, como la suma de k vectores $A^T_{i\bullet}$ de $\mathbf{A}$ con coeficientes estrictamente positivos. Sin pérdida de generalidad, podemos suponer (reordenando) que son los primeros k vectores. Si estos son $l.i$ entonces $z \in M(\mathcal{A}_k)$, donde:

$$\mathcal{A}_k = \{A^T_{1\bullet}, \ldots, A^T_{k\bullet}\}$$

y por tanto $z \in \bigcup_{\mathcal{A}\in\Phi} M(\mathcal{A})$ pues $\mathcal{A}_k \in \Phi$.

Si estos vectores son $l.d$ existen escalares $\mu_i \in \mathbb{R}$, no todos nulos, tales que:

$$\sum_{i=1}^{k} \mu_i A^T_{i\bullet} = 0_n$$

incluso podemos suponer que al menos uno de ellos es positivo (multiplicando por (-1) si hace falta). Mostraremos que el vector z puede expresarse como combinación lineal positiva de menos vectores que los de $\mathcal{A}_k$. Designemos por:

$$\mu = \text{máx}\left\{\frac{\mu_i}{\lambda_i},\ i = 1, \ldots, k\right\} > 0,$$

entonces tendremos:

$$z = \sum_{i=1}^{k} \lambda_i A_{i\bullet}^T - 0 = \sum_{i=1}^{k} \lambda_i A_{i\bullet}^T - \sum_{i=1}^{k} \frac{\mu_i}{\mu} A_{i\bullet}^T = \sum_{i=1}^{k} (\lambda_i - \frac{\mu_i}{\mu}) A_{i\bullet}^T = \sum_{i=1}^{k} \frac{\lambda_i}{\mu} (\mu - \frac{\mu_i}{\lambda_i}) A_{i\bullet}^T,$$

y porque $\lambda_i > 0,\ i = 1, \ldots, k$ y la manera como se ha escogido μ se tiene:

$$\frac{\lambda_i}{\mu}(\mu - \frac{\mu_i}{\lambda_i}) \geq 0,\ i = 1, \ldots, k.$$

Además, para el menos un índice $i_0 \in \{1, \ldots, k\}$ se tiene la igualdad:

$$\mu - \frac{\mu_i}{\lambda_i} = 0,$$

por consiguiente hemos expresado z como combinación lineal positiva de a lo sumo $k-1$ vectores, es decir, los vectores $A_{i\bullet}^T$ para $i \in \{1, \ldots, k\} \setminus \{i_0\}$. Si $\mathcal{A}_{k-1}$ designa este conjunto de vectores y si ellos son $l.i$ se obtiene que $z \in M(\mathcal{A}_{k-1})$ y la tesis. Si son $l.d$ se vuelve a repetir el proceso para eliminar otro vector, hasta que encontremos un conjunto $\mathcal{A}_r$ de vectores $l.i$ y z se exprese como combinación lineal positiva de estos. $z \in M(\mathcal{A}_r)$ y se tiene la tesis. □

B.2. Lema de Farkas

Teorema B.8. *(Lema de Farkas) Sea A una matriz $m \times n$ y $c \in \mathbb{R}^n$. Entonces se tiene la siguiente implicación:*

$$\forall x \in \mathbb{R}^n\ :\ Ax \geq 0\ \Rightarrow\ c^T x \geq 0,$$

si y sólo si

$$\exists \lambda \in \mathbb{R}^m,\ \lambda \geq 0 :\ c = A^T \lambda.$$

DEMOSTRACIÓN. ($\Leftarrow$) Si $c = A^T\lambda$ para algún $\lambda \geq 0$, entonces:

$$\forall x \in \mathbb{R}^n \ : \ Ax \geq 0 \ \Rightarrow \ c^T x = \lambda^T Ax \geq 0.$$

($\Rightarrow$) Consideremos el conjunto:

$$M = \{z \in \mathbb{R}^n \ : \ \exists \lambda \in \mathbb{R}^m_+, \ z = A^T\lambda\}$$

que por lo demostrado antes sabemos que es convexo y cerrado. Supongamos que $c \notin M$, entonces por el primer teorema de separación existe un hiperplano $H = (p, \alpha)$ con $p \neq 0$ tal que separa estrictamente M de c, es decir,

$$\begin{aligned} p^T z &\geq \alpha, \ \forall z \in M, \\ p^T c &< \alpha. \end{aligned}$$

por lo tanto, $\forall \lambda \in \mathbb{R}^m, \ \lambda \geq 0$ se tiene $p^T A^T \lambda \geq \alpha$. Por una parte esto implica que $\alpha \leq 0$ pues $0 \in M$. Por otra parte se tiene necesariamente que $p^T A^T \geq 0^T_m$ porque si no, la función $\lambda \longmapsto p^T A^T \lambda$ no podria ser acotada inferiormente sobre $\mathbb{R}^n_+$. En conclusio'n hemos demostrado que existe $p \in \mathbb{R}^n$ tal que:

$$\begin{aligned} p^T A^T \geq 0 \Leftrightarrow Ap \geq 0, \ &\text{y} \\ p^T c = c^T p < 0, & \end{aligned}$$

lo que contradice la hipótesis. □

Corolario B.9. *(Variante del Lema de Farkas) Sea A una matriz $m \times n$ y $b \in \mathcal{R}^n$. Entonces el sistema $Ax \leq b$ tiene una solución x si y sólo si $yb \geq 0$ para cada vector renglón $y \geq 0$ tal que $yA = 0$. Es decir, se cumple exactamente una de las siguientes proposiciones:*

1. *existe $x \in \mathcal{R}^n$ tal que $Ax \leq b$;*
2. *existe $y \in \mathcal{R}^m$ tal que $y \geq 0, y^T A = 0$ y $y^T b < 0$.*

Apéndice C

Set Covering Problem

El *Set Covering Problem* (SCP) es un modelo central en muchas aplicaciones importantes. El SCP puede definirse formalmente como sigue. Sea $A = (a_{ij})$ una matriz $m \times n$ de ceros y unos, $c = (c_j)$ un vector entero n-dimencional. En lo sucesivo nos referiremos a los renglones y columnas de A simplemente como renglones y columnas. Sea $M = \{1, \ldots, m\}$ y $N = \{1, \ldots, n\}$. El valor c_j $(j \in N)$ representa el costo de la columna j, asumimos que $c_j \geq 0$ para $j \in N$. Decimos que la columna $j \in N$ *cubre* un renglón $i \in M$ si $a_{ij} = 1$. El SCP busca un subconjunto $S \subseteq N$ de columnas de costo mínimo, de forma tal que cada renglón $i \in M$ sea cubierto por al menos una columna $j \in S$. El modelo matemático natural para el SCP es

$$v(\text{SCP}) = \text{mín} \sum_{j \in N} c_j x_j \tag{C.1}$$

sujeto a

$$\sum_{j \in N} a_{ij} x_j \geq 1 \qquad i \in M \tag{C.2}$$

$$x_j \in \{0, 1\} \qquad j \in N \tag{C.3}$$

donde $x_j = 1$ si $j \in S$, $x_j = 0$ de otra forma. Para facilitar, para cada $i \in M$ sea

$$J_i = \{j \in N : a_{ij} = 1\}$$

el conjunto de columnas que cubren el renglón i. Análogamente, para cada columna $j \in N$ sea

$$I_j = \{i \in M : a_{ij} = 1\}$$

el subconjunto de renglones que cubiertos por la columna j. EL SCP es NP-duro [9].

En la literatura se describen muchos procedimientos para reducir el tamaño de ejemplos particulares de SCP, mediante la eliminación de columnas y renglones redundantes. Los procedimientos más comunmente utilizados consideran:

- remover una columna j tal que exista una columna $k \neq j$ tal que $I_j \subseteq I_k$ y $c_j \geq c_k$,
- remover una columna j tal que $c_j \geq \sum_{i \in I_j} \text{mín}\{c_k : k \in J_i\}$,
- remover un renglón i tal que exista un renglón $h \neq i$ tal que $J_h \subseteq J_i$ y
- la inclusión en la solución de una columna j siempre que $J_i = \{j\}$ para algún renglón i.

C.1. Programación lineal y otras relajaciones

La Programación Lineal (PL) relajada del SCP se define como (C.1) y (C.2) sujeto a

$$0 \leq x_j \leq 1 \qquad j \in N. \tag{C.4}$$

Sin embargo, debido a que la solución exacta de esta relajación suele requerir un tiempo considerable, muchos algoritmos para el SCP recurren a la relajación Lagrangiana combinándola con técnicas de optimización de subgradiente con el fin de encontrar soluciones u cercanas al óptimo del problema dual de la PL relajada, dado por

$$\text{máx} \left\{ \sum_{i \in M} u_i : \sum_{i \in I_j} u_i \leq c_j \ (j \in N), u_i \geq 0 \ (i \in M) \right\}. \tag{C.5}$$

Ea relajación Lagrangiana del modelo (C.1)-(C.3) se define de la manera siguiente. Para cada vector $u \in \mathbb{R}^m_+$ multiplicadores Lagrangianos asociados con la restricción (C.2), el subproblema Lagrangiano es:

$$L(u) = \text{mín} \sum_{j \in N} c_j(u) x_j + \sum_{i \in M} u_i \tag{C.6}$$

sujeto a

$$x_j \in \{0, 2\} \qquad j \in N, \tag{C.7}$$

donde $c_j(u) = c_j - \sum_{i \in I_j} u_i$ es el *costo Lagrangiano* asociado a la columna $j \in N$. La determinación de multiplicadores Lagrangianos cercanos al óptimo corresponde en cierto sentido a una solución heurística del problema dual de PL relajado.

Un metódo bien conocido para encontrar vectores multiplicadores cercanos al óptimo en poco tiempo de computo hace uso del *vector subgradinte* $s(u) \in \mathbb{R}^m$, asociado con un vector multiplicador dado u y la correspondiente solución relajada $x(u)$, definida por:

$$s_i(U) = 1 - \sum_{j \in J_i} x_j(u), \qquad i \in M. \tag{C.8}$$

El método genera una secuencia $u^0, u^1, \ldots$ de vectores multiplicadores Lagrangianos no negativos, donde u^0 se define arbitrariamente. En cuanto a la definición de u^k, $k \geq 1$, una posible elección [10] consiste en usar la siguiente formula:

$$u_i^{k+1} = \text{máx}\left\{u_i^k + \lambda \frac{CS - L(u^k)}{\|s(u^k)\|^2} s_i(u^k), 0\right\} \qquad i \in M, \tag{C.9}$$

donde CS es una cota superior de $v(SCP)$ y $\lambda > 0$ es un parémetro que controla el tamao de paso a lo largo de la dirección del subgradiente $s(u^k)$.

Una técnica comúnmente usada para reducir el tamao de un SCP particular se basa en el hecho de que el costo Lagrangiano $c_j(u)$ proporciona una cota inferior en el incremento de la cota inferior $L(u)$ si x_j se fija en 1. Por consiguiente, se puede fijar x_j en 0 siempre que $L(u) - c_j(u) \geq CS$ (recuérdese que el costo Lagrangiano puede ser negativo). Esta técnica es llamada *fijación del costo Lagrangiano*.

El procedimiento de optimización por subgradiente recién descrito puede mejorarse, ver [11].

C.2. Algoritmos Heurísticos

A continuación ilustraremos los algoritmos heurísticos más efectivos para el SCP que han sido experimentalmente evaluados sobre los problemas de prueba de la Librería OR [12]. Muchos de estos algoritmos se basan en la siguiente observación. Para un vector multiplicador Lagrangiano u cercano al óptimo, el costo Lagrangiano $c_j(u)$ ofrece información confiable sobre la útilidad general de seleccionar la columna j. En base a esta propiedad, el costo Lagrangiano (más que el original) es usado para calcular, para cada $j \in N$, un *puntaje* σ_j posicionando las columnas de acuerdo a su verosimilitud para ser seleccionadas en una solución óptima. Estos puntajes son dados como la entrada para un procedimiento heurístico simple, que encuentra un buena solución para el SCP.

El enfoque común a todos los algoritmos considerados en [11] es el siguiente. Una solución S es inicializada como vacía y el conjunto M' de los renglones no cubiertos se establece igual a M. Luego, de forma iterativa, la columna j con el mejor puntaje

σ_j es ańadido a S y M' es actualizado. El puntaje σ_j típicamente es una función del costo original c_j, del número de renglones en M' cubiertos por la columna j y de los multiplicadores asociados con estos renglones. Al final de este procedimiento, el conjunto S típicamente contiene un conjunto W de columnas redundantes, es decir, columnas j tales que $S \setminus \{j\}$ sigue siendo una solución factible del SCP. La eliminación óptima de las columnas redundantes significa resolver un SCP definido por las columnas en W y los renglones en $M \setminus (\bigcup_{j \in S \setminus R} I_j)$.

Los enfoques exactos más efectivos de SCP son algoritmos de ramificación y acotación en los que las cotas inferiores son calculadas resolviendo la PL relajada del SCP.

Bibliografía

[1] P. Fellegi and D. Holt. A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71(353), 17-35, 1976.

[2] J. Gleeson and J. Ryan. Identifying Minimally Infeasible Subsystems of Inequalities. *ORSA Journal on Computing* 2(1), 61-63, 1990.

[3] R.S. Garfinkel, A. S. Kunnathur and G. E. Liepins, (1986). Optimal Imputation of Erroneous Data: Categorical Data, General Edits. *Operations Research*, **34**,744-751.

[4] A. Schrijver. *Theory of Linear and Integer Programming.* Wiley, New York, 1986.

[5] United Nations. *EVALUATING EFFICIENCY OF STATISTICAL DATA EDITING: GENERAL FRAMEWORK.* UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE, CONFERENCE OF EUROPEAN STATISTICIANS METHODOLOGICAL MATERIAL, Geneva, 2000.

[6] William E. Winkleer and Bor-Chung Chen. Extending the Fellegi-Holt Model of Statistical Data Editing. Statistical Research Division, U.S. Bureau of the Census, Washington D.C 20233. RESEARCH REPORT SERIES (*num.2002-02*).

[7] Naus, J. T., T. G. Johnson, and R. Montalvo. 1970. A probabilistic Model for Identifying Errors in Data Editing. *J.Am. Statist. Assoc.* 67,943-950.

[8] R. Bruni, Error Correction for Massive Data Sets, Optimization Methods and Software Vol. 20(2-3), 295-314, 2005.

[9] M.R. Garey and D.S Johnson. Computers and Intractability: A Guide to the theory of NP-completeness. Freeman (1979)

[10] M. Held and R.M. Krap. The Traveling Salesman Problem and Minimum Spanning Trees: Part II. *Mathematical Programing* 1 (1917) 6-25.

[11] Alberto Caprara, Matteo Fischetti and Paolo Toth. Algorithms for the Set Covering Problem.

[12] J.E. Beasley. OR-Library: Distributing Test Problems by Electronic Mail. *Journal of the Operational Research Society* 41 (1990) 1069-1072.

[13] www.inegi.org.mx

[14] http://www.censo2010.org.mx/Presentacion.aspx

[15] C.Poirier. A Functional Evaluation of edit and Imputation Tools. UN/ECE *Work Session on Statistical Data Editing*, Working Paper n.12, Rome, Italy, 1999.

[16] W.E Winkler. State of Statistical Data Editing and current Research Problems. UN/ECE *Work Session on Statistical Data Editing*, Working Paper n.29, Rome, Italy, 1999.

[17] Ton de Waal. An overview of statistical data editing. Statistics Netherlands, 2008.