Análisis de catálogos robustos desde la perspectiva de la minería de reglas de asociación

Enrique Lazcorreta Puigmartí

Universidad Miguel Hernández de Elche

Tesis doctoral dirigida por
Dr. Federico Botella y Dr. Antonio Fernández-Caballero

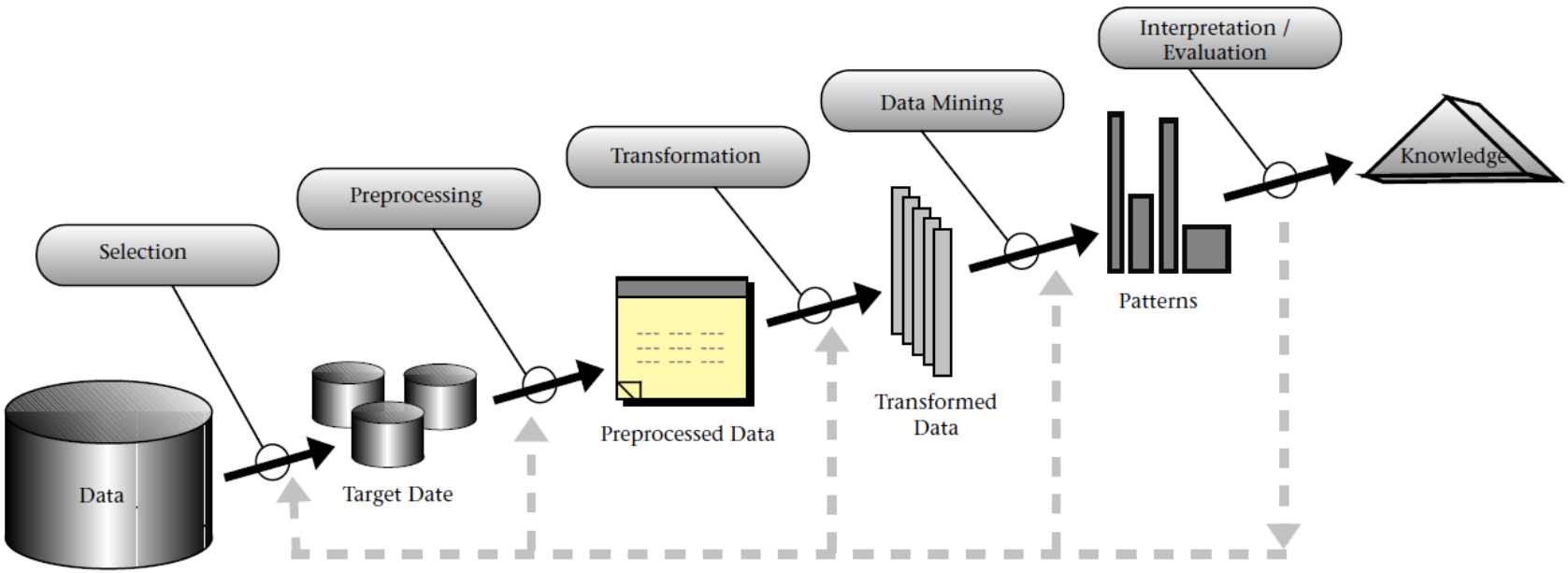
8 de septiembre de 2017

Índice general

1. Introducción
2. Catálogos
3. Experimentación
4. Conclusiones y trabajo futuro
5. Publicaciones

Los datos son la mejor fuente para obtener conocimiento.

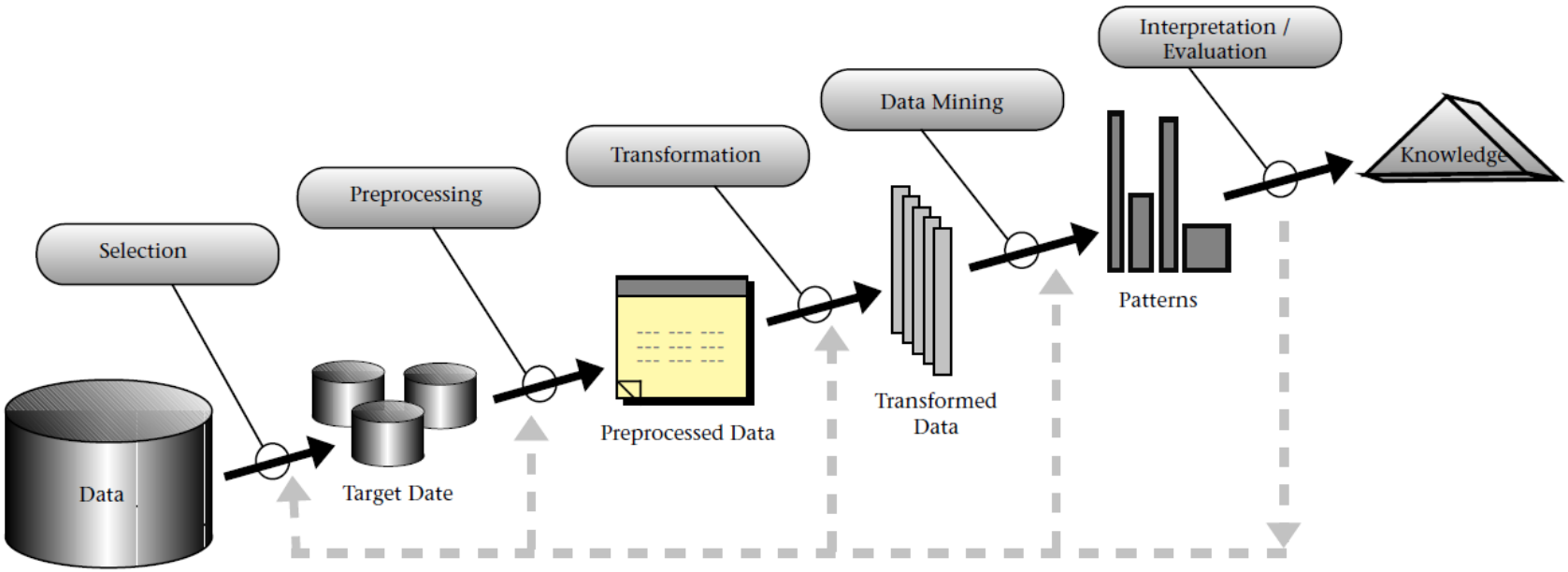
Las bases de datos son una representación digital de la realidad, representación que suele reducirse para poder ser analizada.



El Descubrimiento de Conocimento en Bases de Datos (KDD) surge en 1996 como respuesta al crecimiento exponencial de bases de datos existentes.

Los datos son la mejor fuente para obtener conocimiento.

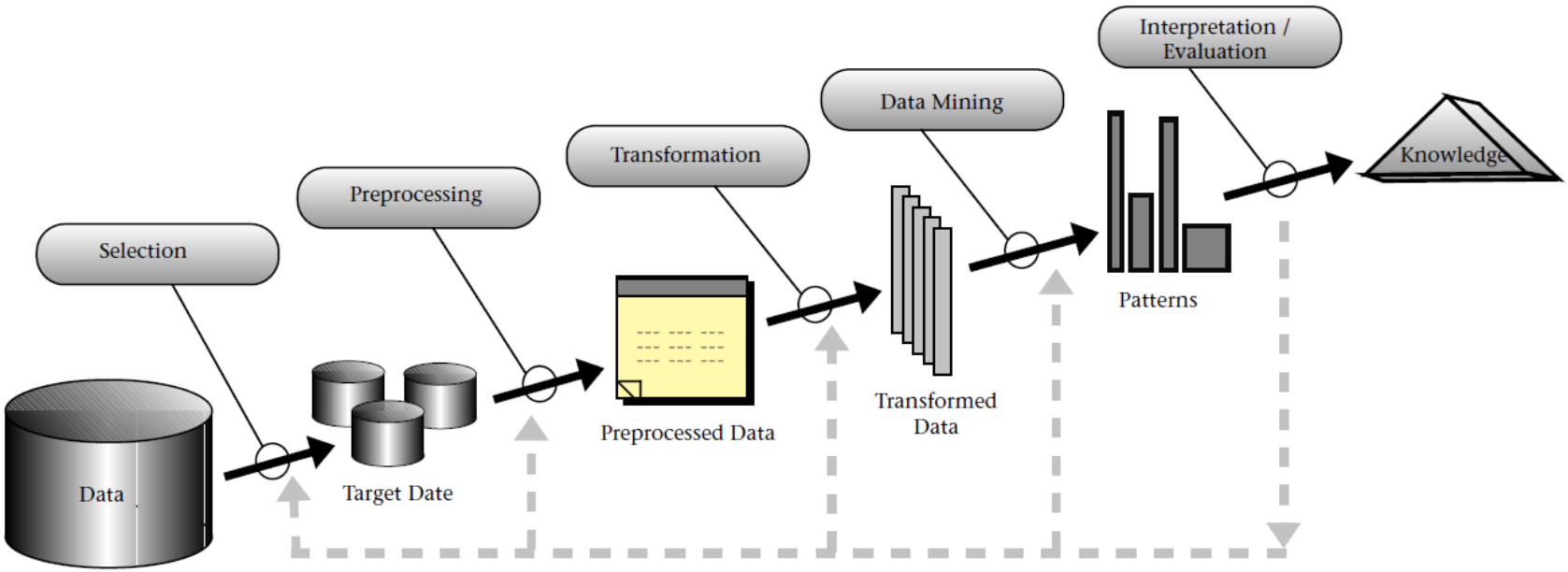
Las bases de datos son una representación digital de la realidad, representación que suele reducirse para poder ser analizada.



El Descubrimiento de Conocimento en Bases de Datos (KDD) surge en 1996 como respuesta al crecimiento exponencial de bases de datos existentes.

Los datos son la mejor fuente para obtener conocimiento.

Las bases de datos son una representación digital de la realidad, representación que suele reducirse para poder ser analizada.



El Descubrimiento de Conocimento en Bases de Datos (KDD) surge en 1996 como respuesta al crecimiento exponencial de bases de datos existentes.

Una base de datos de clasificación contiene evidencias formadas por la caracterización de un individuo y la clase a la que pertenece.

Las reglas de clasificación son patrones del tipo

$$\text{(sub)caracterización} \Rightarrow \text{clase}$$

Se utilizan modelos probabilísticos para determinar la utilidad de las reglas de clasificación descubiertas.

Si se reducen los datos recogidos de modo que no haya evidencias duplicadas ni datos desconocidos, se obtiene un catálogo.

Los catálogos contienen mucha información sobre la estructura de la población en estudio. Pero no contienen información sobre la distribución de frecuencias de individuos de la población.

Una base de datos de clasificación contiene evidencias formadas por la caracterización de un individuo y la clase a la que pertenece.

Las reglas de clasificación son patrones del tipo

$$\text{(sub)caracterización} \Rightarrow \text{clase}$$

Se utilizan modelos probabilísticos para determinar la utilidad de las reglas de clasificación descubiertas.

Si se reducen los datos recogidos de modo que no haya evidencias duplicadas ni datos desconocidos, se obtiene un catálogo.

Los catálogos contienen mucha información sobre la estructura de la población en estudio. Pero no contienen información sobre la distribución de frecuencias de individuos de la población.

Una base de datos de clasificación contiene evidencias formadas por la caracterización de un individuo y la clase a la que pertenece.

Las reglas de clasificación son patrones del tipo

$$\text{(sub)caracterización} \Rightarrow \text{clase}$$

Se utilizan modelos probabilísticos para determinar la utilidad de las reglas de clasificación descubiertas.

Si se reducen los datos recogidos de modo que no haya evidencias duplicadas ni datos desconocidos, se obtiene un catálogo.

Los catálogos contienen mucha información sobre la estructura de la población en estudio. Pero no contienen información sobre la distribución de frecuencias de individuos de la población.

Una base de datos de clasificación contiene evidencias formadas por la caracterización de un individuo y la clase a la que pertenece.

Las reglas de clasificación son patrones del tipo

$$\text{(sub)caracterización} \Rightarrow \text{clase}$$

Se utilizan modelos probabilísticos para determinar la utilidad de las reglas de clasificación descubiertas.

Si se reducen los datos recogidos de modo que no haya evidencias duplicadas ni datos desconocidos, se obtiene un catálogo.

Los catálogos contienen mucha información sobre la estructura de la población en estudio. Pero no contienen información sobre la distribución de frecuencias de individuos de la población.

Una base de datos de clasificación contiene evidencias formadas por la caracterización de un individuo y la clase a la que pertenece.

Las reglas de clasificación son patrones del tipo

$$\text{(sub)caracterización} \Rightarrow \text{clase}$$

Se utilizan modelos probabilísticos para determinar la utilidad de las reglas de clasificación descubiertas.

Si se reducen los datos recogidos de modo que no haya evidencias duplicadas ni datos desconocidos, se obtiene un catálogo.

Los catálogos contienen mucha información sobre la estructura de la población en estudio. Pero no contienen información sobre la distribución de frecuencias de individuos de la población.

Cuando dos o más datos aparecen simultáneamente en los registros de una base de datos, decimos que están asociados.

Las reglas de asociación son patrones del tipo

$$\text{conjunto-de-ítems} \Rightarrow \text{otro-conjunto-de-ítems}$$

Ejemplo

Supóngase que se descubre que, en un supermercado, el 30 % de las compras que contienen cerveza también contienen pañales, y el 2 % de todas las compras contienen ambos productos.

La regla de asociación

$$\text{cerveza} \Rightarrow \text{pañales}$$

tiene un soporte del 2 % y una confianza del 30 %.

Cuando dos o más datos aparecen simultáneamente en los registros de una base de datos, decimos que están asociados.

Las reglas de asociación son patrones del tipo

$$\text{conjunto-de-ítems} \Rightarrow \text{otro-conjunto-de-ítems}$$

Ejemplo

Supóngase que se descubre que, en un supermercado, el 30 % de las compras que contienen cerveza también contienen pañales, y el 2 % de todas las compras contienen ambos productos.

La regla de asociación

$$\text{cerveza} \Rightarrow \text{pañales}$$

tiene un soporte del 2 % y una confianza del 30 %.

Cuando dos o más datos aparecen simultáneamente en los registros de una base de datos, decimos que están asociados.

Las reglas de asociación son patrones del tipo

$$\text{conjunto-de-ítems} \Rightarrow \text{otro-conjunto-de-ítems}$$

Ejemplo

Supóngase que se descubre que, en un supermercado, el 30 % de las compras que contienen cerveza también contienen pañales, y el 2 % de todas las compras contienen ambos productos.

La regla de asociación

$$\text{cerveza} \Rightarrow \text{pañales}$$

tiene un soporte del 2 % y una confianza del 30 %.

Por pequeño que sea, cualquier dataset contiene un número exponencial de reglas de asociación. Teóricamente, $2^{N+1} - 1$ si el dataset contiene N datos diferentes ($2{,}5 \times 10^{30}$ si $N = 100$).

Se utilizan modelos probabilísticos para determinar la utilidad de las reglas de asociación descubiertas.

Los ítems poco frecuentes en el dataset pueden proporcionar el mismo número de reglas de asociación que los más frecuentes. Si hay desbordamiento de memoria durante el análisis es más útil obtener sólo las reglas que se basen en ítems frecuentes. Este problema se denomina *dilema del ítem raro*.

Por pequeño que sea, cualquier dataset contiene un número exponencial de reglas de asociación. Teóricamente, $2^{N+1} - 1$ si el dataset contiene N datos diferentes ($2{,}5 \times 10^{30}$ si $N = 100$).

Se utilizan modelos probabilísticos para determinar la utilidad de las reglas de asociación descubiertas.

Los ítems poco frecuentes en el dataset pueden proporcionar el mismo número de reglas de asociación que los más frecuentes. Si hay desbordamiento de memoria durante el análisis es más útil obtener sólo las reglas que se basen en ítems frecuentes. Este problema se denomina *dilema del ítem raro*.

Por pequeño que sea, cualquier dataset contiene un número exponencial de reglas de asociación. Teóricamente, $2^{N+1} - 1$ si el dataset contiene N datos diferentes ($2{,}5 \times 10^{30}$ si $N = 100$).

Se utilizan modelos probabilísticos para determinar la utilidad de las reglas de asociación descubiertas.

Los ítems poco frecuentes en el dataset pueden proporcionar el mismo número de reglas de asociación que los más frecuentes. Si hay desbordamiento de memoria durante el análisis es más útil obtener sólo las reglas que se basen en ítems frecuentes. Este problema se denomina *dilema del ítem raro*.

Si se filtran las reglas de asociación que se obtienen de un dataset de clasificación, de modo que el antecedente siempre sea una (sub)caracterización y el consecuente una clase se obtiene una regla de clasificación asociativa.

La Minería de Reglas de Clasificación Asociativa (CARM) está en auge debido a la eficiencia de los algoritmos de Minería de Reglas de Asociación (ARM).

El estado del arte sobre CARM muestra muchas investigaciones que utilizan algoritmos de ARM para analizar catálogos.

Si se filtran las reglas de asociación que se obtienen de un dataset de clasificación, de modo que el antecedente siempre sea una (sub)caracterización y el consecuente una clase se obtiene una regla de clasificación asociativa.

La Minería de Reglas de Clasificación Asociativa (CARM) está en auge debido a la eficiencia de los algoritmos de Minería de Reglas de Asociación (ARM).

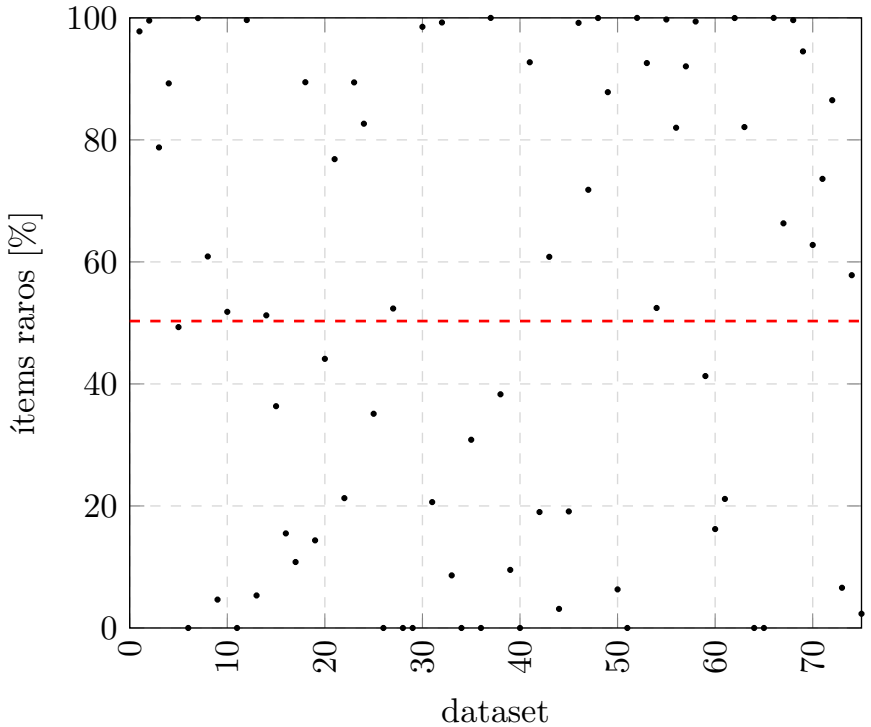
El estado del arte sobre CARM muestra muchas investigaciones que utilizan algoritmos de ARM para analizar catálogos.

Si se filtran las reglas de asociación que se obtienen de un dataset de clasificación, de modo que el antecedente siempre sea una (sub)caracterización y el consecuente una clase se obtiene una regla de clasificación asociativa.

La Minería de Reglas de Clasificación Asociativa (CARM) está en auge debido a la eficiencia de los algoritmos de Minería de Reglas de Asociación (ARM).

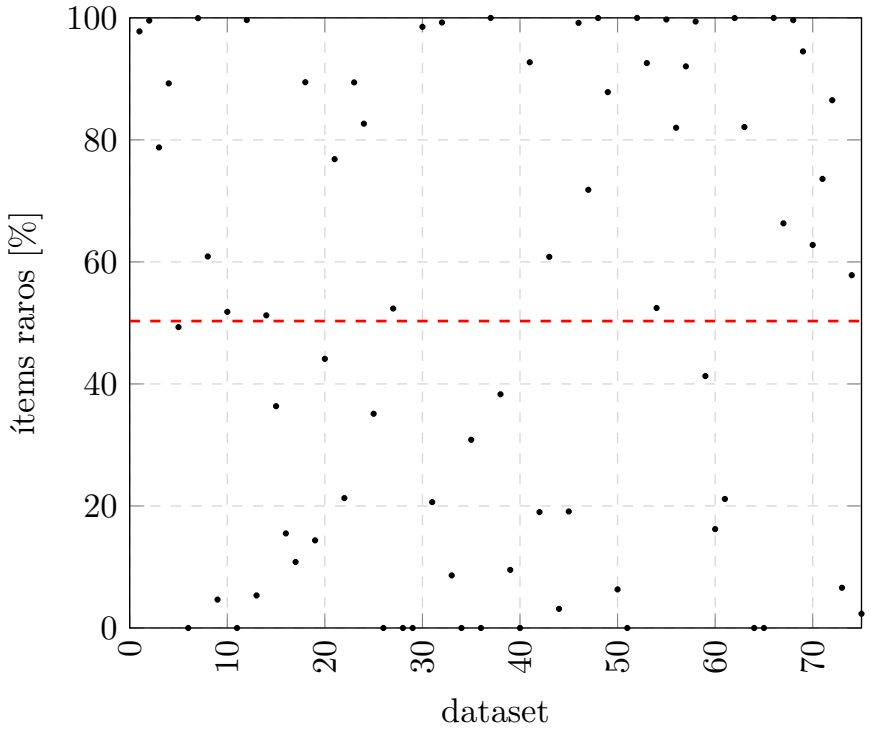
El estado del arte sobre CARM muestra muchas investigaciones que utilizan algoritmos de ARM para analizar catálogos.

Al aplicar soporte mínimo del 1 %, el porcentaje de ítems raros de los 75 datasets de clasificación de KEEL es, en general, alto.



En 40 casos se pierde información sobre más de la mitad de sus ítems. 37 de estos datasets son catálogos.

Al aplicar soporte mínimo del 1 %, el porcentaje de ítems raros de los 75 datasets de clasificación de KEEL es, en general, alto.



En 40 casos se pierde información sobre más de la mitad de sus ítems. 37 de estos datasets son catálogos.

Los 75 datasets de clasificación de KEEL presentan el dilema del ítem raro, debido a que todos sus ítems están fuertemente asociados entre sí y/o al gran número de ítems distintos que producen los atributos numéricos que contienen.

Ejemplo (Catálogo)

Si tenemos una muestra representativa formada por un millón de setas venenosas con la caracterización X, y mil setas comestibles con las caracterizaciones Y_1, Y_2 e Y_3, su catálogo sólo tendrá cuatro tipos de setas:

$$\begin{aligned} X &\Rightarrow \text{venenosa} \\ Y_1 &\Rightarrow \text{comestible} \\ Y_2 &\Rightarrow \text{comestible} \\ Y_3 &\Rightarrow \text{comestible} \end{aligned}$$

El analista deducirá que sólo el 25 % de las setas son venenosas.

Los 75 datasets de clasificación de KEEL presentan el dilema del ítem raro, debido a que todos sus ítems están fuertemente asociados entre sí y/o al gran número de ítems distintos que producen los atributos numéricos que contienen.

Ejemplo (Catálogo)

Si tenemos una muestra representativa formada por un millón de setas venenosas con la caracterización X, y mil setas comestibles con las caracterizaciones Y_1, Y_2 e Y_3, su catálogo sólo tendrá cuatro tipos de setas:

$$\begin{aligned} X &\Rightarrow \text{venenosa} \\ Y_1 &\Rightarrow \text{comestible} \\ Y_2 &\Rightarrow \text{comestible} \\ Y_3 &\Rightarrow \text{comestible} \end{aligned}$$

El analista deducirá que sólo el 25 % de las setas son venenosas.

Desarrollar un modelo teórico que permita analizar los catálogos sin recurrir a los fundamentos estadísticos de la ARM.

Desarrollar algoritmos que permita descubrir las mejores reglas de clasificación asociativa que contienen los catálogos aprovechando las capacidades de la tecnología actual.

Desarrollar un modelo teórico que permita analizar los catálogos sin recurrir a los fundamentos estadísticos de la ARM.

Desarrollar algoritmos que permita descubrir las mejores reglas de clasificación asociativa que contienen los catálogos aprovechando las capacidades de la tecnología actual.

Índice

1 Introducción

2 Catálogos
- Modelo teórico
- Algoritmo $\mathcal{ACDC}$

3 Experimentación

4 Conclusiones y trabajo futuro

5 Publicaciones

El dataset de clasificación `mushroom` tiene 5 versiones.

1. Jeff Schlimmer preparó un dataset de clasificación con 8 416 registros a partir de una guía sobre setas.
2. Donó una copia codificada a UCI, con 8 124 registros, reconociendo en un mensaje posterior que no sabía por qué habían desaparecido 292 registros del dataset original.
3. FIMI'03 utilizó una versión codificada del dataset de UCI. La codificación usada ocultaba el hecho de que el archivo origen tenía registros con datos desconocidos.
4. KEEL incorpora el dataset de UCI eliminando los 2 480 registros con datos desconocidos.
5. LUCS-KDD discretised/normalised ARM and CARM Data Library (2003) publica una versión discretizada del dataset de UCI.

Sea $\mathcal{C} = \{c_1, \ldots, c_\mathcal{Q}\}$ el conjunto de $\mathcal{Q}$ clases en que se particiona la población en estudio. Sea $\mathcal{A} = \{A_1, \ldots, A_\mathcal{N}\}$ el conjunto de $\mathcal{N}$ atributos medidos en los individuos de la población. Sea $\mathcal{B}$ el dataset recogido para llevar a cabo el experimento de clasificación $\mathcal{E}$. $\mathcal{B}$ contiene $\mathcal{M}_\mathcal{B}$ registros, compuestos por la caracterización y la clase de pertenencia de los individuos registrados.

Definición (Caracterización)

Una caracterización x es un conjunto ordenado de valores correspondientes al conjunto de atributos $\mathcal{A} = \{A_1, \ldots, A_\mathcal{N}\}$ de un experimento de clasificación $\mathcal{E}$.

$$x = (A_1 = x_1, \ldots, A_\mathcal{N} = x_\mathcal{N}) = (x_1, \ldots, x_\mathcal{N}) \quad (1)$$

donde x_j es un valor del rango de A_j.

Sea $\mathcal{C} = \{c_1, \ldots, c_\mathcal{Q}\}$ el conjunto de $\mathcal{Q}$ clases en que se particiona la población en estudio. Sea $\mathcal{A} = \{A_1, \ldots, A_\mathcal{N}\}$ el conjunto de $\mathcal{N}$ atributos medidos en los individuos de la población. Sea $\mathcal{B}$ el dataset recogido para llevar a cabo el experimento de clasificación $\mathcal{E}$. $\mathcal{B}$ contiene $\mathcal{M}_\mathcal{B}$ registros, compuestos por la caracterización y la clase de pertenencia de los individuos registrados.

Definición (Caracterización)

Una caracterización x es un conjunto ordenado de valores correspondientes al conjunto de atributos $\mathcal{A} = \{A_1, \ldots, A_\mathcal{N}\}$ de un experimento de clasificación $\mathcal{E}$.

$$x = (A_1 = x_1, \ldots, A_\mathcal{N} = x_\mathcal{N}) = (x_1, \ldots, x_\mathcal{N}) \tag{1}$$

donde x_j es un valor del rango de A_j.

Definición (Evidencia empírica)

Una evidencia empírica, e, es el par formado por la caracterización x de un individuo de la población en estudio y la clase c a la que pertenece.

$$e = (x, c) \tag{2}$$

Definición (Dataset de clasificación)

Un dataset de clasificación es un conjunto de evidencias empíricas.

Definición (Catálogo)

Un catálogo es un dataset de clasificación sin valores desconocidos ni evidencias duplicadas.

Definición (Evidencia empírica)

Una evidencia empírica, e, es el par formado por la caracterización x de un individuo de la población en estudio y la clase c a la que pertenece.

$$e = (x, c) \tag{2}$$

Definición (Dataset de clasificación)

Un dataset de clasificación es un conjunto de evidencias empíricas.

Definición (Catálogo)

Un catálogo es un dataset de clasificación sin valores desconocidos ni evidencias duplicadas.

Definición (Evidencia empírica)

Una evidencia empírica, e, es el par formado por la caracterización x de un individuo de la población en estudio y la clase c a la que pertenece.

$$e = (x, c) \tag{2}$$

Definición (Dataset de clasificación)

Un dataset de clasificación es un conjunto de evidencias empíricas.

Definición (Catálogo)

Un catálogo es un dataset de clasificación sin valores desconocidos ni evidencias duplicadas.

Lema (Existencia de catálogos en los datasets de clasificación)

Todo dataset de clasificación contiene al menos un catálogo.

Lema (Uso del soporte en el análisis de catálogos)

Los catálogos no contienen información sobre la distribución de frecuencias de los ítems de la población en estudio.

Definición (Clasificador ingenuo)

$\mathcal{C}_{\mathcal{D}_0}(x)$ es el clasificador ingenuo de la caracterización x en el catálogo $\mathcal{D}_0$.

$$\mathcal{C}_{\mathcal{D}_0}(x) = \{c \in \mathcal{C} \ /(x, c) \in \mathcal{D}_0\} \tag{3}$$

Lema (Confianza del clasificador ingenuo unitario)

Si $\mathcal{C}_{\mathcal{D}_0}(x) = \{c\} \Rightarrow$ la regla de asociación $\{x \rightarrow c\}$ tiene un 100 % de confianza.

Lema (Existencia de catálogos en los datasets de clasificación)

Todo dataset de clasificación contiene al menos un catálogo.

Lema (Uso del soporte en el análisis de catálogos)

Los catálogos no contienen información sobre la distribución de frecuencias de los ítems de la población en estudio.

Definición (Clasificador ingenuo)

$\mathcal{C}_{\mathcal{D}_0}(x)$ *es el clasificador ingenuo de la caracterización* x *en el catálogo* $\mathcal{D}_0$*.*

$$\mathcal{C}_{\mathcal{D}_0}(x) = \{c \in \mathcal{C} \ /(x, c) \in \mathcal{D}_0\} \tag{3}$$

Lema (Confianza del clasificador ingenuo unitario)

Si $\mathcal{C}_{\mathcal{D}_0}(x) = \{c\} \Rightarrow$ *la regla de asociación* $\{x \rightarrow c\}$ *tiene un 100 % de confianza.*

Lema (Existencia de catálogos en los datasets de clasificación)

Todo dataset de clasificación contiene al menos un catálogo.

Lema (Uso del soporte en el análisis de catálogos)

Los catálogos no contienen información sobre la distribución de frecuencias de los ítems de la población en estudio.

Definición (Clasificador ingenuo)

$\mathcal{C}_{\mathcal{D}_0}(x)$ *es el clasificador ingenuo de la caracterización* x *en el catálogo* $\mathcal{D}_0$.

$$\mathcal{C}_{\mathcal{D}_0}(x) = \{c \in \mathcal{C} \ /(x, c) \in \mathcal{D}_0\} \tag{3}$$

Lema (Confianza del clasificador ingenuo unitario)

Si $\mathcal{C}_{\mathcal{D}_0}(x) = \{c\} \Rightarrow$ *la regla de asociación* $\{x \rightarrow c\}$ *tiene un 100 % de confianza.*

Lema (Existencia de catálogos en los datasets de clasificación)

Todo dataset de clasificación contiene al menos un catálogo.

Lema (Uso del soporte en el análisis de catálogos)

Los catálogos no contienen información sobre la distribución de frecuencias de los ítems de la población en estudio.

Definición (Clasificador ingenuo)

$\mathcal{C}_{\mathcal{D}_0}(x)$ *es el clasificador ingenuo de la caracterización* x *en el catálogo* $\mathcal{D}_0$.

$$\mathcal{C}_{\mathcal{D}_0}(x) = \{c \in \mathcal{C} \ /(x, c) \in \mathcal{D}_0\} \tag{3}$$

Lema (Confianza del clasificador ingenuo unitario)

Si $\mathcal{C}_{\mathcal{D}_0}(x) = \{c\} \Rightarrow$ *la regla de asociación* $\{x \rightarrow c\}$ *tiene un 100 % de confianza.*

Definición (Tipos de caracterizaciones)

Las caracterizaciones del experimento de clasificación $\mathcal{E}$ pueden ser de tres tipos:

- *x es desconocida cuando no está registrada en el catálogo $\mathcal{D}_0$.*

$$x \text{ es desconocida } \Leftrightarrow |\mathcal{C}_{\mathcal{D}_0}(x)| = 0 \tag{4}$$

- *x es robusta cuando está relacionada con una única clase en el catálogo $\mathcal{D}_0$.*

$$x \text{ es robusta } \Leftrightarrow |\mathcal{C}_{\mathcal{D}_0}(x)| = 1 \tag{5}$$

- *x contiene incertidumbre cuando está relacionada con más de una clase en el catálogo $\mathcal{D}_0$.*

$$x \text{ contiene incertidumbre } \Leftrightarrow |\mathcal{C}_{\mathcal{D}_0}(x)| > 1 \tag{6}$$

Definición (Tipos de catálogos)

Hay dos tipos de catálogo, en función de la tipología de las caracterizaciones que contiene.

$$\mathcal{D}_0 \text{ es robusto} \Leftrightarrow |\mathcal{C}_{\mathcal{D}_0}(x)| = 1, \forall x \in \mathcal{D}_0 \quad (7)$$

$$\mathcal{D}_0 \text{ contiene incertidumbre} \Leftrightarrow \exists x \in \mathcal{D}_0 \ / \ |\mathcal{C}_{\mathcal{D}_0}(x)| > 1 \quad (8)$$

Obtención de catálogos robustos

Separando en $\mathcal{D}_\mathcal{E}$ todas las caracterizaciones con incertidumbre de $\mathcal{D}_0$ se obtiene el catálogo robusto $\mathcal{D}$.

$$\mathcal{D}_0 = \mathcal{D} \cup \mathcal{D}_\mathcal{E} \quad (9)$$

Definición (Tipos de catálogos)

Hay dos tipos de catálogo, en función de la tipología de las caracterizaciones que contiene.

$$\mathcal{D}_0 \text{ es robusto } \Leftrightarrow \ |\mathcal{C}_{\mathcal{D}_0}(x)| = 1, \forall x \in \mathcal{D}_0 \tag{7}$$

$$\mathcal{D}_0 \text{ contiene incertidumbre } \Leftrightarrow \ \exists x \in \mathcal{D}_0 \ / \ |\mathcal{C}_{\mathcal{D}_0}(x)| > 1 \tag{8}$$

Obtención de catálogos robustos

Separando en $\mathcal{D}_\mathcal{E}$ todas las caracterizaciones con incertidumbre de $\mathcal{D}_0$ se obtiene el catálogo robusto $\mathcal{D}$.

$$\mathcal{D}_0 = \mathcal{D} \cup \mathcal{D}_\mathcal{E} \tag{9}$$

Descomposición de datasets de clasificación

Cualquier dataset de clasificación $\mathcal{B}$ se puede dividir en cuatro matrices:

$$\mathcal{B} = \mathcal{D} \cup \mathcal{D}_{\mathcal{E}} \cup \mathcal{D}_{?} \cup \mathcal{D}_{d}$$

$\mathcal{D}$ es el mayor catálogo robusto contenido en $\mathcal{B}$.

$\mathcal{D}_{\mathcal{E}}$ contiene las caracterizaciones con incertidumbre registradas.

$\mathcal{D}_{?}$ contiene las caracterizaciones incompletas registradas.

$\mathcal{D}_{d}$ contiene caracterizaciones ya registradas en el resto de matrices.

Definición (Catálogo reducido)

Si se eliminan de $\mathcal{D}$ los datos correspondientes al atributo $A_i \in \mathcal{A}$ y se eliminan las evidencias duplicadas se obtiene la matriz $\mathcal{D}^{-i}$, catálogo reducido del experimento $\mathcal{E}$.

Generalización

Si se eliminan de $\mathcal{D}$ los datos correspondientes al conjunto de atributos $\mathcal{I} = \{A_i\} \subset \mathcal{A}$ y se eliminan las evidencias duplicadas se obtiene la matriz $\mathcal{D}^{-\mathcal{I}}$, catálogo reducido del experimento $\mathcal{E}$.

Definición (Catálogo reducido)

Si se eliminan de $\mathcal{D}$ los datos correspondientes al atributo $A_i \in \mathcal{A}$ y se eliminan las evidencias duplicadas se obtiene la matriz $\mathcal{D}^{-i}$, catálogo reducido del experimento $\mathcal{E}$.

Generalización

Si se eliminan de $\mathcal{D}$ los datos correspondientes al conjunto de atributos $\mathcal{I} = \{A_i\} \subset \mathcal{A}$ y se eliminan las evidencias duplicadas se obtiene la matriz $\mathcal{D}^{-\mathcal{I}}$, catálogo reducido del experimento $\mathcal{E}$.

Caracterización de atributos

Si $\mathcal{D}^{-i}$ es un catálogo robusto, A_i es un atributo redundante en el experimento $\mathcal{E}$. En otro caso es un atributo necesario. Si el atributo A_i sólo contiene un valor en el catálogo $\mathcal{D}_0$, pese a que teóricamente pueda tomar más valores se denominará atributo constante.

Definición (Conjunto de caracterizaciones)

$\mathcal{D}_0^{-\mathcal{C}}$ representa el conjunto de caracterizaciones que contiene $\mathcal{D}_0$. Se obtiene eliminando de $\mathcal{D}_0$ toda la información sobre la clase, y eliminando las caracterizaciones duplicadas.

$$\mathcal{D}_0^{-\mathcal{C}} = \{x \ / \ (x, c) \in \mathcal{D}_0\} \tag{10}$$

Generalización

$\mathcal{D}^{-\mathcal{I}-\mathcal{C}}$ representa el conjunto de caracterizaciones del catálogo reducido $\mathcal{D}^{-\mathcal{I}}$.

Caracterización de atributos

Si $\mathcal{D}^{-i}$ es un catálogo robusto, A_i es un atributo redundante en el experimento $\mathcal{E}$. En otro caso es un atributo necesario. Si el atributo A_i sólo contiene un valor en el catálogo $\mathcal{D}_0$, pese a que teóricamente pueda tomar más valores se denominará atributo constante.

Definición (Conjunto de caracterizaciones)

$\mathcal{D}_0^{-\mathcal{C}}$ representa el conjunto de caracterizaciones que contiene $\mathcal{D}_0$. Se obtiene eliminando de $\mathcal{D}_0$ toda la información sobre la clase, y eliminando las caracterizaciones duplicadas.

$$\mathcal{D}_0^{-\mathcal{C}} = \{x \ / \ (x, c) \in \mathcal{D}_0\} \tag{10}$$

Generalización

$\mathcal{D}^{-\mathcal{I}-\mathcal{C}}$ representa el conjunto de caracterizaciones del catálogo reducido $\mathcal{D}^{-\mathcal{I}}$.

Caracterización de atributos

Si $\mathcal{D}^{-i}$ es un catálogo robusto, A_i es un atributo redundante en el experimento $\mathcal{E}$. En otro caso es un atributo necesario. Si el atributo A_i sólo contiene un valor en el catálogo $\mathcal{D}_0$, pese a que teóricamente pueda tomar más valores se denominará atributo constante.

Definición (Conjunto de caracterizaciones)

$\mathcal{D}_0^{-\mathcal{C}}$ representa el conjunto de caracterizaciones que contiene $\mathcal{D}_0$. Se obtiene eliminando de $\mathcal{D}_0$ toda la información sobre la clase, y eliminando las caracterizaciones duplicadas.

$$\mathcal{D}_0^{-\mathcal{C}} = \{x \ / \ (x, c) \in \mathcal{D}_0\} \tag{10}$$

Generalización

$\mathcal{D}^{-\mathcal{I}-\mathcal{C}}$ representa el conjunto de caracterizaciones del catálogo reducido $\mathcal{D}^{-\mathcal{I}}$.

Teorema (Catálogo robusto)

Sea $\mathcal{D}_0^{-\mathcal{C}}$ el conjunto de caracterizaciones del catálogo $\mathcal{D}_0$.

$$\mathcal{D}_0 \text{ es robusto } \Leftrightarrow \ |\mathcal{D}_0^{-\mathcal{C}}| = |\mathcal{D}_0| \tag{11}$$

Si el experimento de clasificación se ha diseñado con algún atributo prescindible, A_i, entonces $\mathcal{D}^{-i}$ también es un catálogo robusto. En tal caso, contiene $\mathcal{M}^{-i} = |\mathcal{D}^{-i}|$ reglas de clasificación asociativa con un 100 % de confianza.

Teorema (Catálogo robusto)

Sea $\mathcal{D}_0^{-\mathcal{C}}$ el conjunto de caracterizaciones del catálogo $\mathcal{D}_0$.

$$\mathcal{D}_0 \text{ es robusto } \Leftrightarrow \ |\mathcal{D}_0^{-\mathcal{C}}| = |\mathcal{D}_0| \tag{11}$$

Si el experimento de clasificación se ha diseñado con algún atributo prescindible, A_i, entonces $\mathcal{D}^{-i}$ también es un catálogo robusto. En tal caso, contiene $\mathcal{M}^{-i} = |\mathcal{D}^{-i}|$ reglas de clasificación asociativa con un 100 % de confianza.

Si existen más atributos prescindibles en el experimento diseñado, aplicando recursivamente su eliminación se obtendrán nuevos conjuntos de reglas de clasificación asociativa con un 100 % de confianza, basadas en diferentes subconjuntos de $\mathcal{A}$.

Definición (Colección de catálogos robustos)

$\mathcal{CCR}_{\mathcal{D}}$ es la Colección de Catálogos Robustos que contiene el catálogo robusto $\mathcal{D}$.

$$\mathcal{CCR}_{\mathcal{D}} = \{\mathcal{D}^{-\mathcal{I}} \subset \mathcal{D} \ / \ \mathcal{D}^{-\mathcal{I}} \text{ es robusto}\} \tag{12}$$

$\mathcal{CCR}_{\mathcal{D}}$ contiene conjuntos de reglas de clasificación asociativa con un 100 % de confianza. También contiene información sobre qué conjuntos de atributos de $\mathcal{A}$ se pueden usar para clasificar a un nuevo individuo de la población (selección de características).

Si existen más atributos prescindibles en el experimento diseñado, aplicando recursivamente su eliminación se obtendrán nuevos conjuntos de reglas de clasificación asociativa con un 100 % de confianza, basadas en diferentes subconjuntos de $\mathcal{A}$.

Definición (Colección de catálogos robustos)

$\mathcal{CCR}_{\mathcal{D}}$ es la Colección de Catálogos Robustos que contiene el catálogo robusto $\mathcal{D}$.

$$\mathcal{CCR}_{\mathcal{D}} = \{\mathcal{D}^{-\mathcal{I}} \subset \mathcal{D} \ / \ \mathcal{D}^{-\mathcal{I}} \text{ es robusto}\} \tag{12}$$

$\mathcal{CCR}_{\mathcal{D}}$ contiene conjuntos de reglas de clasificación asociativa con un 100 % de confianza. También contiene información sobre qué conjuntos de atributos de $\mathcal{A}$ se pueden usar para clasificar a un nuevo individuo de la población (selección de características).

Si existen más atributos prescindibles en el experimento diseñado, aplicando recursivamente su eliminación se obtendrán nuevos conjuntos de reglas de clasificación asociativa con un 100 % de confianza, basadas en diferentes subconjuntos de $\mathcal{A}$.

Definición (Colección de catálogos robustos)

$\mathcal{CCR}_{\mathcal{D}}$ es la Colección de Catálogos Robustos que contiene el catálogo robusto $\mathcal{D}$.

$$\mathcal{CCR}_{\mathcal{D}} = \{\mathcal{D}^{-\mathcal{I}} \subset \mathcal{D} \ / \ \mathcal{D}^{-\mathcal{I}} \text{ es robusto}\} \tag{12}$$

$\mathcal{CCR}_{\mathcal{D}}$ contiene conjuntos de reglas de clasificación asociativa con un 100 % de confianza. También contiene información sobre qué conjuntos de atributos de $\mathcal{A}$ se pueden usar para clasificar a un nuevo individuo de la población (selección de características).

Análisis de Caracterizaciones en Datasets de Clasificación

El algoritmo $\mathcal{ACDC}$ obtiene la $\mathcal{CCR}_{\mathcal{D}}$ del experimento de clasificación $\mathcal{E}$.

Se basa en el manejo de matrices, por lo que puede ser implementado en cualquier lenguaje de programación o con cualquier sistema informático capaz de manipular eficientemente matrices.

Se presenta como una colección de funciones que, utilizando los conceptos definidos en el modelo teórico, devuelven una estructura similar a la utilizada en el algoritmo APRIORI, cuyos nodos son los atributos eliminados de cada catálogo robusto de $\mathcal{CCR}_{\mathcal{D}}$.

Análisis de Caracterizaciones en Datasets de Clasificación

El algoritmo $\mathcal{ACDC}$ obtiene la $\mathcal{CCR}_{\mathcal{D}}$ del experimento de clasificación $\mathcal{E}$.

Se basa en el manejo de matrices, por lo que puede ser implementado en cualquier lenguaje de programación o con cualquier sistema informático capaz de manipular eficientemente matrices.

Se presenta como una colección de funciones que, utilizando los conceptos definidos en el modelo teórico, devuelven una estructura similar a la utilizada en el algoritmo APRIORI, cuyos nodos son los atributos eliminados de cada catálogo robusto de $\mathcal{CCR}_{\mathcal{D}}$.

Análisis de Caracterizaciones en Datasets de Clasificación

El algoritmo $\mathcal{ACDC}$ obtiene la $\mathcal{CCR}_{\mathcal{D}}$ del experimento de clasificación $\mathcal{E}$.

Se basa en el manejo de matrices, por lo que puede ser implementado en cualquier lenguaje de programación o con cualquier sistema informático capaz de manipular eficientemente matrices.

Se presenta como una colección de funciones que, utilizando los conceptos definidos en el modelo teórico, devuelven una estructura similar a la utilizada en el algoritmo Apriori, cuyos nodos son los atributos eliminados de cada catálogo robusto de $\mathcal{CCR}_{\mathcal{D}}$.

Algoritmo $\mathcal{ACDC}$

El algoritmo $\mathcal{ACDC}$ está formado por 5 pasos.

1. Abrir el dataset y obtener el número de atributos en estudio. Inicializar el resto de variables a 0 o a $\emptyset$.
2. Obtener el catálogo $\mathcal{D}_0$ leyendo todas las evidencias del dataset.
3. Eliminar atributos constantes.
4. Aislar la incertidumbre del catálogo inicial $\mathcal{D}_0$.
5. Obtener la Colección de Catálogos Robustos $\mathcal{CCR}_{\mathcal{D}}$.

Algoritmo $\mathcal{ACDC}$

El algoritmo $\mathcal{ACDC}$ está formado por 5 pasos.

1. Abrir el dataset y obtener el número de atributos en estudio. Inicializar el resto de variables a 0 o a $\emptyset$.
2. Obtener el catálogo $\mathcal{D}_0$ leyendo todas las evidencias del dataset.
3. Eliminar atributos constantes.
4. Aislar la incertidumbre del catálogo inicial $\mathcal{D}_0$.
5. Obtener la Colección de Catálogos Robustos $\mathcal{CCR}_{\mathcal{D}}$.

Algoritmo $\mathcal{ACDC}$

El algoritmo $\mathcal{ACDC}$ está formado por 5 pasos.

1. Abrir el dataset y obtener el número de atributos en estudio. Inicializar el resto de variables a 0 o a $\emptyset$.
2. Obtener el catálogo $\mathcal{D}_0$ leyendo todas las evidencias del dataset.
3. Eliminar atributos constantes.
4. Aislar la incertidumbre del catálogo inicial $\mathcal{D}_0$.
5. Obtener la Colección de Catálogos Robustos $\mathcal{CCR}_{\mathcal{D}}$.

Algoritmo $\mathcal{ACDC}$

El algoritmo $\mathcal{ACDC}$ está formado por 5 pasos.

1. Abrir el dataset y obtener el número de atributos en estudio. Inicializar el resto de variables a 0 o a $\emptyset$.
2. Obtener el catálogo $\mathcal{D}_0$ leyendo todas las evidencias del dataset.
3. Eliminar atributos constantes.
4. Aislar la incertidumbre del catálogo inicial $\mathcal{D}_0$.
5. Obtener la Colección de Catálogos Robustos $\mathcal{CCR}_{\mathcal{D}}$.

Algoritmo $\mathcal{ACDC}$

El algoritmo $\mathcal{ACDC}$ está formado por 5 pasos.

1. Abrir el dataset y obtener el número de atributos en estudio. Inicializar el resto de variables a 0 o a $\emptyset$.
2. Obtener el catálogo $\mathcal{D}_0$ leyendo todas las evidencias del dataset.
3. Eliminar atributos constantes.
4. Aislar la incertidumbre del catálogo inicial $\mathcal{D}_0$.
5. Obtener la Colección de Catálogos Robustos $\mathcal{CCR}_{\mathcal{D}}$.

Algoritmo $\mathcal{ACDC}$

El algoritmo $\mathcal{ACDC}$ está formado por 5 pasos.

1. Abrir el dataset y obtener el número de atributos en estudio. Inicializar el resto de variables a 0 o a $\emptyset$.
2. Obtener el catálogo $\mathcal{D}_0$ leyendo todas las evidencias del dataset.
3. Eliminar atributos constantes.
4. Aislar la incertidumbre del catálogo inicial $\mathcal{D}_0$.
5. Obtener la Colección de Catálogos Robustos $\mathcal{CCR}_{\mathcal{D}}$.

Obtención del catálogo $\mathcal{D}_0$

```
ObtenerCatalogo(dataset)
{
    foreach (evidencia e en dataset)
        num_evidencias++

        if (e tiene valores desconocidos)
            D_?.add(e)
            continue;

        if (e no pertenece a D_0)
            D_0.add(e)

    return  num_evidencias, D_0, D_?
}
```

Eliminación de una columna

```
EliminaColumna(i, N, I, D_I)
{
    i = ObtenColumnaAEliminar(i, N, I)

    D_I_i = conjunto vacio

    foreach (evidencia e_I en D_I)
        e_I_i = e_I sin su i-esimo valor;

        if (e_I_i no esta en D_I_i)
            D_I_i.add(e_I_i})

    return  D_I_i
}
```

Obtener columna a eliminar

```
ObtenerColumnaAEliminar(i, N, I)
{
    for (columna_a_eliminar = 0...N)
        if (columna_a_eliminar esta en I)
            continue;
        if (columna_a_eliminar == i)
            break;

    return  columna_a_eliminar;
}
```

Comprobación sobre si un catálogo es robusto

```
EsRobusto(N, I, D_I)
{
    D_I_C = EliminaColumna(N, N, I)

    return  (|D_I| == |D_I_C|)
}
```

Aislar incertidumbre

```
AislarIncerticumbre(D_0)
{
    foreach (evidencia e en D_0)
        x = (e_1 ... e_N)
        c = e_N_1

        D_0.remove(e)
        D_incertidumbre.add(x,c)

    foreach (caracterizacion x en D_incertidumbre)
        if (|D_incertidumbre[x]| == 1)
            e = (x, c)
            D_incertidumbre.remove(x,c)
            D_0.add(e)
}
```

Obtener $\mathcal{CCR}_{\mathcal{D}}$

```
ObtenerCCR(ult_atributo_eliminado, N, I, D_I)
{
    for (a = (ult_atributo_eliminado + 1) ... N)
        if (a esta en I || D_I_a esta en CCR)
            continue;

        D_I_a = EliminarColumna(a, N, I, D_I)

        if (EsRobusto(D_I_a))
            CCR.add(D_I_a)
            ObtenerCCR(a, N, I + a, D_I_a)
}
```

Índice

1 Introducción

2 Catálogos

3 Experimentación
- Primera lectura de un dataset de clasificación
- Atributos constantes
- El algoritmo $\mathcal{ACDC}$
- Datasets de grandes dimensiones

4 Conclusiones y trabajo futuro

5 Publicaciones

Todos los experimentos realizados se han desarrollado y ejecutado sobre un ordenador de gama media, con procesador Intel Core i5 y 8GB de memoria RAM corriendo sobre el sistema operativo macOS.

Para implementar el algoritmo $\mathcal{ACDC}$ se ha usado la `C++ Standard Library (libc++, 5.0)` y `clang` con la intención de obtener un código estándar y compatible con la colección de compiladores `gcc`.

Las características del equipo y el código utilizado están a disposición de cualquier equipo de investigación, lo que garantiza que pueda ser probado y mejorado por la comunidad científica.

Todos los experimentos realizados se han desarrollado y ejecutado sobre un ordenador de gama media, con procesador Intel Core i5 y 8GB de memoria RAM corriendo sobre el sistema operativo macOS.

Para implementar el algoritmo $\mathcal{ACDC}$ se ha usado la `C++ Standard Library (libc++, 5.0)` y `clang` con la intención de obtener un código estándar y compatible con la colección de compiladores `gcc`.

Las características del equipo y el código utilizado están a disposición de cualquier equipo de investigación, lo que garantiza que pueda ser probado y mejorado por la comunidad científica.

Todos los experimentos realizados se han desarrollado y ejecutado sobre un ordenador de gama media, con procesador Intel Core i5 y 8GB de memoria RAM corriendo sobre el sistema operativo macOS.

Para implementar el algoritmo $\mathcal{ACDC}$ se ha usado la `C++ Standard Library (libc++, 5.0)` y `clang` con la intención de obtener un código estándar y compatible con la colección de compiladores `gcc`.

Las características del equipo y el código utilizado están a disposición de cualquier equipo de investigación, lo que garantiza que pueda ser probado y mejorado por la comunidad científica.

Esta fase de la investigación se ha desarrollado utilizando los 75 datasets de clasificación publicados en *KEEL Standard Dataset Repository*[a]. Y un dataset de gran tamaño publicado en *UCI Machine Learning Repository*[b].

[a] `http://sci2s.ugr.es/keel/category.php?cat=clas`
[b] http://archive.ics.uci.edu/ml/index.php

Algunos datasets podrían ser considerados muestras representativas, otros no tienen evidencias duplicadas y tienen el aspecto de catálogos.

Todos han sido tratados sin aplicar el criterio de soporte mínimo, y en ningún caso se ha detenido la ejecución del algoritmo debido a los problemas de desbordamiento típicos de ARM.

Esta fase de la investigación se ha desarrollado utilizando los 75 datasets de clasificación publicados en *KEEL Standard Dataset Repository*[a]. Y un dataset de gran tamaño publicado en *UCI Machine Learning Repository*[b].

[a] `http://sci2s.ugr.es/keel/category.php?cat=clas`
[b] http://archive.ics.uci.edu/ml/index.php

Algunos datasets podrían ser considerados muestras representativas, otros no tienen evidencias duplicadas y tienen el aspecto de catálogos.

Todos han sido tratados sin aplicar el criterio de soporte mínimo, y en ningún caso se ha detenido la ejecución del algoritmo debido a los problemas de desbordamiento típicos de ARM.

Esta fase de la investigación se ha desarrollado utilizando los 75 datasets de clasificación publicados en *KEEL Standard Dataset Repository*[a]. Y un dataset de gran tamaño publicado en *UCI Machine Learning Repository*[b].

[a] `http://sci2s.ugr.es/keel/category.php?cat=clas`
[b] http://archive.ics.uci.edu/ml/index.php

Algunos datasets podrían ser considerados muestras representativas, otros no tienen evidencias duplicadas y tienen el aspecto de catálogos.

Todos han sido tratados sin aplicar el criterio de soporte mínimo, y en ningún caso se ha detenido la ejecución del algoritmo debido a los problemas de desbordamiento típicos de ARM.

La minería de datos se basa en la gestión eficiente de gran cantidad de datos. Sin embargo, aunque las capacidades de los ordenadores son cada vez mayores, si se pide que gestionen más recursos de los que pueden aparecerán problemas de desbordamiento de memoria.

El planteamiento de este trabajo de investigación se basa en utilizar las capacidades de los ordenadores de esta época para resolver problemas que hace una década no se podían afrontar por falta de recursos informáticos.

Los datasets de clasificación utilizados por los investigadores se guardan generalmente en archivos de texto. Pueden tener tamaños diversos, desde unos pocos KB hasta varios GB, el algoritmo implementado obtiene el mayor catálogo que contienen y lo guarda como una matriz en memoria RAM.

La minería de datos se basa en la gestión eficiente de gran cantidad de datos. Sin embargo, aunque las capacidades de los ordenadores son cada vez mayores, si se pide que gestionen más recursos de los que pueden aparecerán problemas de desbordamiento de memoria.

El planteamiento de este trabajo de investigación se basa en utilizar las capacidades de los ordenadores de esta época para resolver problemas que hace una década no se podían afrontar por falta de recursos informáticos.

Los datasets de clasificación utilizados por los investigadores se guardan generalmente en archivos de texto. Pueden tener tamaños diversos, desde unos pocos KB hasta varios GB, el algoritmo implementado obtiene el mayor catálogo que contienen y lo guarda como una matriz en memoria RAM.

La minería de datos se basa en la gestión eficiente de gran cantidad de datos. Sin embargo, aunque las capacidades de los ordenadores son cada vez mayores, si se pide que gestionen más recursos de los que pueden aparecerán problemas de desbordamiento de memoria.

El planteamiento de este trabajo de investigación se basa en utilizar las capacidades de los ordenadores de esta época para resolver problemas que hace una década no se podían afrontar por falta de recursos informáticos.

Los datasets de clasificación utilizados por los investigadores se guardan generalmente en archivos de texto. Pueden tener tamaños diversos, desde unos pocos KB hasta varios GB, el algoritmo implementado obtiene el mayor catálogo que contienen y lo guarda como una matriz en memoria RAM.

Antes de proceder a aplicar el algoritmo $\mathcal{ACDC}$ conviene conocer las características del dataset que se quiere analizar. Su tamaño en disco, número de atributos y su tipología, rango de los atributos, número de clases, número de evidencias del dataset y de su mayor catálogo e incertidumbre son datos que se pueden obtener aplicando el algoritmo sin recursividad.

La tabla que se muestra a continuación contiene la información obtenida al analizar los 75 datasets de KEEL, en menos de un minuto y usando 170.2MB de memoria RAM.

Antes de proceder a aplicar el algoritmo $\mathcal{ACDC}$ conviene conocer las características del dataset que se quiere analizar. Su tamaño en disco, número de atributos y su tipología, rango de los atributos, número de clases, número de evidencias del dataset y de su mayor catálogo e incertidumbre son datos que se pueden obtener aplicando el algoritmo sin recursividad.

La tabla que se muestra a continuación contiene la información obtenida al analizar los 75 datasets de KEEL, en menos de un minuto y usando 170.2MB de memoria RAM.

Tabla 6.7: Características de los datasets de clasificación estándar de KEEL

#	dataset	Tamaño	$\mathcal{N}$(r)(c)	$rg(\mathcal{A})$	$\lvert\mathcal{C}\rvert$	$\lvert dataset\rvert$	$\lvert\mathcal{D}_0\rvert$	Inc.	(1)	(2)	t (sg)
1	abalone	224.58KB	8(8)(-)	6,047	28	4,174	4,174	-	A_1 (6,044 / 4,174)	A_5 (3,619 / 4,174)	0.0845
2	adult	5.23MB	14(7)(-)	27,243	2	45,222	45,170	5	A_{14} (27,202 / 45,160)	A_{12} (27,146 / 45,162)	0.9729
3	appendicitis	4.73KB	7(7)(-)	530	2	106	106	-	A_1 (456 / 106)	A_7 (431 / 106)	0.0024
4	australian	30.43KB	14(13)(-)	1,143	2	690	690	-	A_3 (942 / 689)	A_{14} (903 / 690)	0.0142
5	automobile	22.98KB	25(25)(1)	795	6	159	159	-	A_{25} (650 / 155)	A_{25} (650 / 155)	0.0060
6	balance	14.35KB	4(0)(-)	20	3	625	625	-			0.0009
7	banana	86.72KB	2(0)(-)	4,045	2	5,300	5,291	1			0.0079
8	bands	32.82KB	19(19)(-)	719	2	365	365	-	A_1 (692 / 365)	A_{11} (614 / 365)	0.0095
9	breast	18.38KB	9(2)(-)	41	2	277	257	6	A_4 (34 / 247)	A_4 (34 / 247)	0.0017
10	bupa	13.19KB	6(6)(-)	328	2	345	341	-	A_1 (302 / 341)	A_5 (234 / 341)	0.0029
11	car	51.00KB	6(0)(-)	21	4	1,728	1,728	-			0.0032
12	census	65.87MB	41(36)(2)	69,224	2	142,521	139,303	-	A_{32} (69,219 / 138,495)	A_6 (68,193 / 139,303)	16.8982
13	chess	241.32KB	36(9)(-)	73	2	3,196	3,196	-	A_{11} (71 / 3,091)	A_2 (71 / 3,160)	0.0815
14	cleveland	18.26KB	13(13)(-)	397	5	297	297	-	A_1 (356 / 297)	A_5 (245 / 297)	0.0055
15	coil2000	1.63MB	85(76)(-)	650	2	9,822	8,261	119	A_{54} (644 / 8,100)	A_5 (640 / 8,142)	1.3054
16	connect-4	5.76MB	42(42)(-)	126	3	67,557	67,557	-	A_1 (123 / 67,557)	A_1 (123 / 67,557)	5.7513
17	contraceptive	30.82KB	9(0)(-)	71	3	1,473	1,358	62			0.0034
18	crx	34.77KB	15(15)(-)	1,116	2	653	653	-	A_3 (917 / 652)	A_2 (776 / 653)	0.0121
19	dermatology	26.73KB	34(34)(-)	189	6	358	358	-	A_{34} (129 / 356)	A_{34} (129 / 356)	0.0144
20	ecoli	12.34KB	7(7)(-)	357	8	336	336	-	A_2 (294 / 335)	A_6 (277 / 336)	0.0032
21	fars	56.40MB	29(9)(-)	752	8	100,968	90,097	2,696	A_{29} (734 / 87,321)	A_{21} (679 / 87,401)	3.7714
22	flare	25.50KB	11(1)(-)	41	6	1,066	287	56	A_7 (39 / 213)	A_7 (39 / 213)	0.0034
23	german	99.63KB	20(20)(-)	1,075	2	1,000	1,000	-	A_5 (154 / 998)	A_5 (154 / 998)	0.0319
24	glass	16.80KB	9(9)(-)	905	6	214	213	-	A_5 (772 / 212)	A_1 (756 / 213)	0.0032
25	haberman	6.53KB	3(0)(-)	92	2	306	283	6			0.0007
26	hayes-roth	1.98KB	4(1)(-)	15	3	160	84	9	A_1 (12 / 55)	A_1 (12 / 55)	0.0003
27	heart	10.58KB	13(13)(-)	380	2	270	270	-	A_1 (339 / 270)	A_5 (236 / 270)	0.0049
28	hepatitis	4.71KB	19(19)(-)	272	2	80	80	-	A_1 (232 / 80)	A_{15} (213 / 80)	0.0021
29	housevotes	10.49KB	16(10)(-)	32	2	232	160	-	A_{10} (30 / 144)	A_4 (30 / 160)	0.0028
30	ionosphere	76.48KB	33(33)(-)	7,279	2	351	350	-	A_{27} (7,031 / 348)	A_{17} (7,020 / 350)	0.0289
31	iris	5.49KB	4(4)(-)	123	3	150	147	-	A_1 (88 / 142)	A_3 (80 / 143)	0.0[illegible]10
32	kddcup	71.89MB	41(37)(2)	19,445	23	494,020	145,583	2	A_{23} (18,955 / 113,564)	A_6 (8,720 / 138,690)	13.9[illegible]

Características de los datasets de clasificación estándar de KEEL (cont.)

#	dataset	Tamaño	$\mathcal{N}$(r)(c)	$rg(\mathcal{A})$	$\lvert\mathcal{C}\rvert$	$\lvert dataset\rvert$	$\lvert\mathcal{D}_0\rvert$	Inc.	(1)	(2)	t (sg)
32b	kddcup99[6]	708.18MB	41(37)(1)	41,894	23	4,898,431	1,074,974	17	A_{23} (41,382 / 914,464)	A_6 (20,401 / 950,794)	208.8805
33	kr-vs-k	547.33KB	6(0)(-)	40	18	28,056	28,056	-			0.0575
34	led7digit	32.57KB	7(0)(-)	14	10	500	85	37			0.0013
35	letter	716.18KB	16(13)(-)	256	26	20,000	18,668	-	A_2 (240 / 18,177)	A_1 (240 / 18,498)	0.5173
36	lymphography	15.43KB	18(18)(-)	59	4	148	148	-	A_{15} (56 / 147)	A_{14} (51 / 148)	0.0050
37	magic	1.43MB	10(10)(-)	147,107	2	19,020	18,905	-	A_1 (128,464 / 18,905)	A_6 (128,403 / 18,905)	0.5491
38	mammographic	10.81KB	5(0)(-)	92	2	830	519	45			0.0014
39	marketing	188.72KB	13(0)(-)	75	9	6,876	5,631	429			0.0173
40	monk-2	9.13KB	6(3)(-)	17	2	432	432	-	A_1 (14 / 144)	A_1 (14 / 144)	0.0017
41	movement_libras	257.76KB	90(90)(-)	18,180	15	360	330	-	A_1 (17,966 / 330)	A_{89} (17,945 / 330)	0.0771
42	mushroom	254.84KB	22(22)(1)	98	2	5,644	5,644	-	A_9 (89 / 1,662)	A_9 (89 / 1,662)	0.1699
43	newthyroid	5.07KB	5(5)(-)	334	3	215	215	-	A_1 (279 / 215)	A_2 (234 / 215)	0.0016
44	nursery	1.12MB	8(0)(-)	27	5	12,960	12,960	-			0.0365
45	optdigits	817.18KB	64(64)(2)	914	10	5,620	5,620	-	A_2 (905 / 5,620)	A_3 (897 / 5,620)	0.6831
46	page-blocks	241.03KB	10(10)(-)	9,094	5	5,472	5,393	13	A_1 (8,990 / 5,380)	A_9 (7,376 / 5,380)	0.0900
47	penbased	710.16KB	16(16)(-)	1,608	10	10,992	10,992	-	A_1 (1,507 / 10,992)	A_1 (1,507 / 10,992)	0.3276
48	phoneme	170.88KB	5(5)(-)	11,178	2	5,404	5,349	-	A_1 (9,109 / 5,326)	A_3 (8,659 / 5,341)	0.0588
49	pima	34.13KB	8(8)(-)	1,254	2	768	768	-	A_1 (1,237 / 768)	A_7 (737 / 768)	0.0099
50	poker	23.01MB	10(0)(-)	85	10	1,025,009	1,022,770	-			3.3446
51	post-operative	4.56KB	8(0)(-)	23	3	87	71	6			0.0010
52	ring	1.11MB	20(20)(-)	75,126	2	7,400	7,400	-	A_1 (71,386 / 7,400)	A_3 (71,319 / 7,400)	0.4051
53	saheart	21.04KB	9(9)(-)	1,767	2	462	462	-	A_1 (1,705 / 462)	A_4 (1,359 / 462)	0.0074
54	satimage	978.71KB	36(36)(-)	2,806	6	6,435	6,435	-	A_1 (2,755 / 6,435)	A_{12} (2,702 / 6,435)	0.4429
55	segment	410.58KB	19(17)(1)	14,929	7	2,310	2,086	-	A_2 (14,691 / 2,077)	A_{19} (12,992 / 2,086)	0.0793
56	shuttle	1.47MB	9(8)(-)	1,109	7	57,999	57,999	-	A_6 (810 / 27,480)	A_6 (810 / 27,480)	0.8551
57	sonar	75.87KB	60(60)(-)	8,208	2	208	208	-	A_1 (8,143 / 208)	A_{31} (8,018 / 208)	0.0246
58	spambase	1.33MB	57(50)(-)	15,091	2	4,597	4,203	3	A_{50} (14,450 / 4,189)	A_{55} (12,930 / 4,198)	0.5399
59	spectfheart	36.57KB	44(44)(-)	1,887	2	267	267	-	A_1 (1,848 / 267)	A_{42} (1,826 / 267)	0.0174
60	splice	574.01KB	60(60)(-)	287	3	3,190	3,005	1	A_1 (282 / 3,002)	A_{35} (281 / 3,003)	0.3150
61	tae	2.33KB	5(2)(-)	101	3	151	106	4	A_4 (99 / 101)	A_1 (99 / 102)	0.0007
62	texture	1.46MB	40(40)(-)	39,556	11	5,500	5,473	-	A_1 (38,695 / 5,473)	A_{33} (38,246 / 5,473)	0.5359

[6] Procedente de *UCI Machine Learning Repository* 2013.

Características de los datasets de clasificación estándar de KEEL (cont.)

#	dataset	Tamaño	$\mathcal{N}$ (r)(c)	$rg(\mathcal{A})$	$\lvert\mathcal{C}\rvert$	$\lvert dataset\rvert$	$\lvert\mathcal{D}_0\rvert$	Inc.	(1)	(2)	t (sg)
63	thyroid	635.77KB	21(21)(-)	1,438	3	7,200	7,129	-	A_1 (1,340 / 6,980)	A_{21} (972 / 7,129)	0.2798
64	tic-tac-toe	35.10KB	9(9)(-)	27	2	958	958	-	A_1 (24 / 958)	A_1 (24 / 958)	0.0114
65	titanic	51.86KB	3(3)(-)	8	2	2,201	14	10	A_1 (4 / 2)	A_1 (4 / 2)	0.0017
66	twonorm	1.23MB	20(20)(-)	135,085	2	7,400	7,400	-	A_1 (128,337 / 7,400)	A_{10} (128,295 / 7,400)	0.4535
67	vehicle	71.11KB	18(18)(-)	1,430	4	846	846	-	A_1 (1,386 / 846)	A_{12} (1,006 / 846)	0.0216
68	vowel	72.54KB	13(13)(-)	8,106	11	990	990	-	A_1 (8,104 / 990)	A_5 (7,229 / 990)	0.0209
69	wdbc	102.75KB	30(30)(-)	8,314	2	569	569	-	A_1 (7,858 / 569)	A_{24} (7,770 / 569)	0.0337
70	wine	13.70KB	13(13)(-)	1,276	3	178	178	-	A_1 (1,150 / 178)	A_2 (1,143 / 178)	0.0049
71	winequality-red	90.35KB	11(11)(-)	1,453	6	1,599	1,359	-	A_{11} (1,388 / 1,355)	A_8 (1,017 / 1,359)	0.0293
72	winequality-white	281.17KB	11(11)(-)	2,308	7	4,898	3,961	-	A_1 (2,240 / 3,943)	A_8 (1,418 / 3,961)	0.0723
73	wisconsin	14.37KB	9(8)(-)	89	2	683	449	-	A_1 (79 / 385)	A_1 (79 / 385)	0.0058
74	yeast	73.54KB	8(7)(-)	412	10	1,484	1,453	-	A_4 (334 / 1,451)	A_1 (331 / 1,452)	0.0178
75	zoo	4.06KB	16(14)(-)	36	7	101	59	-	A_7 (34 / 49)	A_1 (34 / 58)	0.0014
TOTAL			**(63)(6)**				**37**	**19**			**54.5209**

#	dataset	Tamaño	$\mathcal{N}$(r)(c)	$rg(\mathcal{A})$	$\lvert\mathcal{C}\rvert$	|dataset|	$\lvert\mathcal{D}_0\rvert$	Inc.
1	abalone	224.58KB	8(8)(-)	6,047	28	4,174	4,174	-
...								
32	kddcup	71.89MB	41(37)(2)	19,445	23	494,020	145,583	2
–	kddcup99	708.18MB	41(37)(1)	41,894	23	4,898,431	1,074,974	17
...								
42	mushroom	254.84KB	22(22)(1)	98	2	5,644	5,644	-
...								
50	poker	23.01MB	10(0)(-)	85	10	1,025,009	1,022,770	-
...								
75	zoo	4.06KB	16(14)(-)	36	7	101	59	-

Resumen

De los 75 datasets de clasificación publicados en KEEL:

- 63 tienen algún atributo redundante (en 43 casos todos son redundantes).
- 6 tienen atributos constantes.
- 37 no tienen evidencias duplicadas.
- 19 tienen incertidumbre.

#	dataset	Tamaño	$\mathcal{N}$(r)(c)	$rg(\mathcal{A})$	$\lvert\mathcal{C}\rvert$	$\lvert dataset\rvert$	$\lvert\mathcal{D}_0\rvert$	Inc.	(1)	(2)	t (sg)
1	abalone	224.58KB	8(8)(-)	6,047	28	4,174	4,174	-	A_1 (6,044 / 4,174)	A_5 (3,619 / 4,174)	0.0845
2	adult	5.23MB	14(7)(-)	27,243	2	45,222	45,170	5	A_{14} (27,202 / 45,160)	A_{12} (27,146 / 45,162)	0.9729
3	appendicitis	4.73KB	7(7)(-)	530	2	106	106	-	A_1 (456 / 106)	A_7 (431 / 106)	0.0024
4	australian	30.43KB	14(13)(-)	1,143	2	690	690	-	A_3 (942 / 689)	A_{14} (903 / 690)	0.0142
5	automobile	22.98KB	25(25)(1)	795	6	159	159	-	A_{25} (650 / 155)	A_{25} (650 / 155)	0.0060
6	balance	14.35KB	4(0)(-)	20	3	625	625	-			0.0009
25	haberman	6.53KB	3(0)(-)	92	2	306	283	6			0.0007
26	hayes-roth	1.98KB	4(1)(-)	15	3	160	84	9	A_1 (12 / 55)	A_1 (12 / 55)	0.0003
27	heart	10.58KB	13(13)(-)	380	2	270	270	-	A_1 (339 / 270)	A_5 (236 / 270)	0.0049
28	hepatitis	4.71KB	19(19)(-)	272	2	80	80	-	A_1 (232 / 80)	A_{15} (213 / 80)	0.0021
29	housevotes	10.49KB	16(10)(-)	32	2	232	160	-	A_{10} (30 / 144)	A_4 (30 / 160)	0.0028
30	ionosphere	76.48KB	33(33)(-)	7,279	2	351	350	-	A_{27} (7,031 / 348)	A_{17} (7,020 / 350)	0.0289
31	iris	5.49KB	4(4)(-)	123	3	150	147	-	A_1 (88 / 142)	A_3 (80 / 143)	0.0010
32	kddcup	71.89MB	41(37)(2)	19,445	23	494,020	145,583	2	A_{23} (18,955 / 113,564)	A_6 (8,720 / 138,690)	13.952

Resumen

Al eliminar tan solo un atributo:

- `kddcup` reduce sus evidencias de 145,583 a 11,564.
- `monk-2` reduce sus evidencias de 432 a 144.
- `mushroom` reduce sus evidencias de 5,644 a 1,662.
- `shuttle` reduce sus evidencias de 57,999 a 27,480.
- `zoo` reduce sus evidencias de 59 a 49.

Los datasets `automobile`, `census`, `kddcup`, `mushroom`, `optdigits` y `segment` tienen atributos constantes.

Cada atributo constante duplica el número de evidencias del $\mathcal{CCR}_{\mathcal{D}}$ sin aportar información al problema de clasificación. Si no se suprimen se corre el riesgo de sufrir el dilema del ítem raro, manifestándose como desbordamiento de memoria RAM.

Efecto de la supresión de atributos constantes

Dataset	$\lvert\mathcal{D}_0\rvert$	$\lvert\mathcal{A}\rvert$	Atributos redundantes	$\lvert\mathcal{CCR}_{\mathcal{D}}\rvert$	Evidencias robustas	Tiempo (sg.)
automobile	159	25	25	18,182,384	2,788,517,384	3,768
				9,091,192	1,394,258,692	1,839
mushroom	5,644	22	22	2,093,424	1,204,584,784	2,077
				1,046,712	602,292,392	1,026
segment	2,086	19	18	261,952	545,253,360	889
				130,976	272,626,680	413

Los datasets `automobile`, `census`, `kddcup`, `mushroom`, `optdigits` y `segment` tienen atributos constantes.

Cada atributo constante duplica el número de evidencias del $\mathcal{CCR}_{\mathcal{D}}$ sin aportar información al problema de clasificación. Si no se suprimen se corre el riesgo de sufrir el dilema del ítem raro, manifestándose como desbordamiento de memoria RAM.

Efecto de la supresión de atributos constantes

Dataset	$\lvert\mathcal{D}_0\rvert$	$\lvert\mathcal{A}\rvert$	Atributos redundantes	$\lvert\mathcal{CCR}_{\mathcal{D}}\rvert$	Evidencias robustas	Tiempo (sg.)
automobile	159	25	25	18,182,384	2,788,517,384	3,768
				9,091,192	1,394,258,692	1,839
mushroom	5,644	22	22	2,093,424	1,204,584,784	2,077
				1,046,712	602,292,392	1,026
segment	2,086	19	18	261,952	545,253,360	889
				130,976	272,626,680	413

Los datasets `automobile`, `census`, `kddcup`, `mushroom`, `optdigits` y `segment` tienen atributos constantes.

Cada atributo constante duplica el número de evidencias del $\mathcal{CCR}_{\mathcal{D}}$ sin aportar información al problema de clasificación. Si no se suprimen se corre el riesgo de sufrir el dilema del ítem raro, manifestándose como desbordamiento de memoria RAM.

Efecto de la supresión de atributos constantes

Dataset	$\lvert\mathcal{D}_0\rvert$	$\lvert\mathcal{A}\rvert$	**Atributos redundantes**	$\lvert\mathcal{CCR}_{\mathcal{D}}\rvert$	**Evidencias robustas**	**Tiempo (sg.)**
automobile	159	25	25	18,182,384	2,788,517,384	3,768
				9,091,192	1,394,258,692	1,839
mushroom	5,644	22	22	2,093,424	1,204,584,784	2,077
				1,046,712	602,292,392	1,026
segment	2,086	19	18	261,952	545,253,360	889
				130,976	272,626,680	413

Al aplicar el algoritmo $\mathcal{ACDC}$ sin supervisar a los datasets de clasificación con 25 o menos atributos redundantes se obtiene su $\mathcal{CCR}_{\mathcal{D}}$ completo, e información sobre el número mínimo de atributos (1) que, a partir de los datos recogidos, tienen el mismo poder de clasificación que el conjunto de todos los atributos del diseño original del experimento.

Dataset	$\lvert\mathcal{CCR}_{\mathcal{D}}\rvert$	**Evidencias robustas**	$\lvert\mathcal{D}\rvert$	mín $\lvert\mathcal{D}^{-\mathcal{I}}\rvert$	$\lvert\mathcal{A}\rvert$	**(1)**	**Tiempo (sg.)**
abalone	94	392,352	4,174	4,172	8	3	0.61
automobile	9,091,192	1,394,258,692	159	75	25	2	1,839.99
chess	288	882,560	3,196	2,826	36	29	2.80
fars	315	27,514,164	90,097	87,262	29	23	67.86
housevotes	332	43,340	160	78	16	8	0.12
monk-2	8	1,296	432	36	6	3	0.00
mushroom	1,046,712	602,292,392	5,644	22	22	3	1,085.09
ring	1,048,450	7,758,529,904	7,400	7,396	20	2	15,515.26
shuttle	160	5,008,378	57,999	6,455	9	4	12.03
thyroid	982,016	6,865,495,424	7,129	5,454	21	3	14,441.62
twonorm	1,048,554	7,759,299,600	7,400	7,400	20	2	14,004.66
zoo	8,912	361,576	59	20	16	5	0.57

Al aplicar el algoritmo $\mathcal{ACDC}$ sin supervisar a los datasets de clasificación con 25 o menos atributos redundantes se obtiene su $\mathcal{CCR}_{\mathcal{D}}$ completo, e información sobre el número mínimo de atributos (1) que, a partir de los datos recogidos, tienen el mismo poder de clasificación que el conjunto de todos los atributos del diseño original del experimento.

Dataset	$\lvert\mathcal{CCR}_{\mathcal{D}}\rvert$	**Evidencias robustas**	$\lvert\mathcal{D}\rvert$	mín $\lvert\mathcal{D}^{-\mathcal{I}}\rvert$	$\lvert\mathcal{A}\rvert$	**(1)**	**Tiempo (sg.)**
abalone	94	392,352	4,174	4,172	8	3	0.61
automobile	9,091,192	1,394,258,692	159	75	25	2	1,839.99
chess	288	882,560	3,196	2,826	36	29	2.80
fars	315	27,514,164	90,097	87,262	29	23	67.86
housevotes	332	43,340	160	78	16	8	0.12
monk-2	8	1,296	432	36	6	3	0.00
mushroom	1,046,712	602,292,392	5,644	22	22	3	1,085.09
ring	1,048,450	7,758,529,904	7,400	7,396	20	2	15,515.26
shuttle	160	5,008,378	57,999	6,455	9	4	12.03
thyroid	982,016	6,865,495,424	7,129	5,454	21	3	14,441.62
twonorm	1,048,554	7,759,299,600	7,400	7,400	20	2	14,004.66
zoo	8,912	361,576	59	20	16	5	0.57

Al aplicar el algoritmo $\mathcal{ACDC}$ sin supervisar a los datasets de clasificación con 25 o menos atributos redundantes se obtiene su $\mathcal{CCR}_{\mathcal{D}}$ completo, e información sobre el número mínimo de atributos (1) que, a partir de los datos recogidos, tienen el mismo poder de clasificación que el conjunto de todos los atributos del diseño original del experimento.

Dataset	$\lvert\mathcal{CCR}_{\mathcal{D}}\rvert$	**Evidencias robustas**	$\lvert\mathcal{D}\rvert$	mín $\lvert\mathcal{D}^{-\mathcal{I}}\rvert$	$\lvert\mathcal{A}\rvert$	**(1)**	**Tiempo (sg.)**
abalone	94	392,352	4,174	4,172	8	3	0.61
automobile	9,091,192	1,394,258,692	159	75	25	2	1,839.99
chess	288	882,560	3,196	2,826	36	29	2.80
fars	315	27,514,164	90,097	87,262	29	23	67.86
housevotes	332	43,340	160	78	16	8	0.12
monk-2	8	1,296	432	36	6	3	0.00
mushroom	1,046,712	602,292,392	5,644	22	22	3	1,085.09
ring	1,048,450	7,758,529,904	7,400	7,396	20	2	15,515.26
shuttle	160	5,008,378	57,999	6,455	9	4	12.03
thyroid	982,016	6,865,495,424	7,129	5,454	21	3	14,441.62
twonorm	1,048,554	7,759,299,600	7,400	7,400	20	2	14,004.66
zoo	8,912	361,576	59	20	16	5	0.57

Si se observan los datos obtenidos al aplicar el algoritmo $\mathcal{ACDC}$ sin supervisión y se considera que la fase de búsqueda de catálogos robustos es una función recursiva que trabaja en memoria RAM con matrices y submatrices, es fácil deducir que si se aplica el algoritmo sin supervisión sobre datasets de clasificación de grandes dimensiones aparecerán problemas de desbordamiento de memoria.

Los dos parámetros más determinantes para decidir si un dataset es de grandes dimensiones son el número de atributos redundantes que tiene, en primer lugar, y el tamaño del archivo que lo contiene.

[?] etiquetan como datasets de grandes dimensiones aquellos que tienen más de 35 atributos, sin embargo no hemos tenido problemas en analizar sin supervisión el dataset `chess` en menos de 3 segundos, ya que a pesar de sus 36 atributos sólo 9 son redundantes.

Si se observan los datos obtenidos al aplicar el algoritmo $\mathcal{ACDC}$ sin supervisión y se considera que la fase de búsqueda de catálogos robustos es una función recursiva que trabaja en memoria RAM con matrices y submatrices, es fácil deducir que si se aplica el algoritmo sin supervisión sobre datasets de clasificación de grandes dimensiones aparecerán problemas de desbordamiento de memoria.

Los dos parámetros más determinantes para decidir si un dataset es de grandes dimensiones son el número de atributos redundantes que tiene, en primer lugar, y el tamaño del archivo que lo contiene.

[?] etiquetan como datasets de grandes dimensiones aquellos que tienen más de 35 atributos, sin embargo no hemos tenido problemas en analizar sin supervisión el dataset `chess` en menos de 3 segundos, ya que a pesar de sus 36 atributos sólo 9 son redundantes.

Si se observan los datos obtenidos al aplicar el algoritmo $\mathcal{ACDC}$ sin supervisión y se considera que la fase de búsqueda de catálogos robustos es una función recursiva que trabaja en memoria RAM con matrices y submatrices, es fácil deducir que si se aplica el algoritmo sin supervisión sobre datasets de clasificación de grandes dimensiones aparecerán problemas de desbordamiento de memoria.

Los dos parámetros más determinantes para decidir si un dataset es de grandes dimensiones son el número de atributos redundantes que tiene, en primer lugar, y el tamaño del archivo que lo contiene.

[**?**] etiquetan como datasets de grandes dimensiones aquellos que tienen más de 35 atributos, sin embargo no hemos tenido problemas en analizar sin supervisión el dataset `chess` en menos de 3 segundos, ya que a pesar de sus 36 atributos sólo 9 son redundantes.

Aplicando hasta $\mathcal{K}$ iteraciones de la función recursiva del algoritmo, se obtiene $\mathcal{CCR}_{\mathcal{D},\mathcal{K}}$. Sus catálogos robustos contienen reglas de clasificación asociativa con $\mathcal{N}, (\mathcal{N}-1), \ldots, (\mathcal{N}-\mathcal{K})$ valores en sus antecedentes y un 100 % de confianza.

Dataset	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{A}\rvert$	$\mathcal{K}$	$\lvert\mathcal{CCR}_{\mathcal{D},\mathcal{K}}\rvert$	**Evidencias robustas**	$\lvert$ mín $\mathcal{D}^{-\mathcal{I}}\rvert$	**Tiempo (sg.)**
census	139,303	41(38)(2)	1	37	5,152,838	138,495	30
			2	664	92,447,127	138,283	271
			3	7,695	1,071,052,509	137,838	2,966
			4	64,695	9,002,142,142	137,655	23,343
kddcup	145,583	41(37)(2)	1	38	5,454,396	113,564	37
			2	701	99,168,860	104,179	300
			3	8,362	1,165,483,398	96,953	3,297
			4	72,488	9,950,832,287	91,926	27,464
kddcup99	1,074,974	41(37)(1)	1	37	39,243,116	914,464	208
			2	660	690,812,738	765,854	1,629
			3	7,564	7,814,485,719	700,146	17,156
mov_lib[1]	360	90(90)(-)	1	91	30,030	330	1
			2	4,096	1,351,680	330	2
			3	121,576	40,120,080	330	74
			4	2,676,766	883,332,780	330	1,616
			5	46,626,034	15,386,591,220	330	29,912

[1] movement_libras

Aplicando hasta $\mathcal{K}$ iteraciones de la función recursiva del algoritmo, se obtiene $\mathcal{CCR}_{\mathcal{D},\mathcal{K}}$. Sus catálogos robustos contienen reglas de clasificación asociativa con $\mathcal{N}, (\mathcal{N}-1), \ldots, (\mathcal{N}-\mathcal{K})$ valores en sus antecedentes y un 100 % de confianza.

Dataset	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{A}\rvert$	$\mathcal{K}$	$\lvert\mathcal{CCR}_{\mathcal{D},\mathcal{K}}\rvert$	**Evidencias robustas**	$\lvert$ mín $\mathcal{D}^{-\mathcal{I}}\rvert$	**Tiempo (sg.)**
census	139,303	41(38)(2)	1	37	5,152,838	138,495	30
			2	664	92,447,127	138,283	271
			3	7,695	1,071,052,509	137,838	2,966
			4	64,695	9,002,142,142	137,655	23,343
kddcup	145,583	41(37)(2)	1	38	5,454,396	113,564	37
			2	701	99,168,860	104,179	300
			3	8,362	1,165,483,398	96,953	3,297
			4	72,488	9,950,832,287	91,926	27,464
kddcup99	1,074,974	41(37)(1)	1	37	39,243,116	914,464	208
			2	660	690,812,738	765,854	1,629
			3	7,564	7,814,485,719	700,146	17,156
mov_lib[1]	360	90(90)(-)	1	91	30,030	330	1
			2	4,096	1,351,680	330	2
			3	121,576	40,120,080	330	74
			4	2,676,766	883,332,780	330	1,616
			5	46,626,034	15,386,591,220	330	29,912

[1] movement_libras

Aplicando hasta $\mathcal{K}$ iteraciones de la función recursiva del algoritmo, se obtiene $\mathcal{CCR}_{\mathcal{D},\mathcal{K}}$. Sus catálogos robustos contienen reglas de clasificación asociativa con $\mathcal{N}, (\mathcal{N}-1), \ldots, (\mathcal{N}-\mathcal{K})$ valores en sus antecedentes y un 100 % de confianza.

Dataset	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{A}\rvert$	$\mathcal{K}$	$\lvert\mathcal{CCR}_{\mathcal{D},\mathcal{K}}\rvert$	**Evidencias robustas**	$\lvert$ mín $\mathcal{D}^{-\mathcal{I}}\rvert$	**Tiempo (sg.)**
census	139,303	41(38)(2)	1	37	5,152,838	138,495	30
			2	664	92,447,127	138,283	271
			3	7,695	1,071,052,509	137,838	2,966
			4	64,695	9,002,142,142	137,655	23,343
kddcup	145,583	41(37)(2)	1	38	5,454,396	113,564	37
			2	701	99,168,860	104,179	300
			3	8,362	1,165,483,398	96,953	3,297
			4	72,488	9,950,832,287	91,926	27,464
kddcup99	1,074,974	41(37)(1)	1	37	39,243,116	914,464	208
			2	660	690,812,738	765,854	1,629
			3	7,564	7,814,485,719	700,146	17,156
mov_lib[1]	360	90(90)(-)	1	91	30,030	330	1
			2	4,096	1,351,680	330	2
			3	121,576	40,120,080	330	74
			4	2,676,766	883,332,780	330	1,616
			5	46,626,034	15,386,591,220	330	29,912

[1] movement_libras

Cada uno de los catálogos robustos $\mathcal{D}^{-\mathcal{I}}$ de la $\mathcal{CCR}_{\mathcal{D},\mathcal{K}}$ obtenida en cada experimento tiene su propia $\mathcal{CCR}_{\mathcal{D}^{-\mathcal{I}}}$.

Aplicando el algoritmo a cualquiera de ellos, de forma supervisada si aún contiene muchos atributos redundantes, se obtendrá una nueva colección de catálogos robustos con la misma información que el catálogo inicial pero con un número reducido de atributos usados para la clasificación.

Cada uno de los catálogos robustos $\mathcal{D}^{-\mathcal{I}}$ de la $\mathcal{CCR}_{\mathcal{D},\mathcal{K}}$ obtenida en cada experimento tiene su propia $\mathcal{CCR}_{\mathcal{D}^{-\mathcal{I}}}$.

Aplicando el algoritmo a cualquiera de ellos, de forma supervisada si aún contiene muchos atributos redundantes, se obtendrá una nueva colección de catálogos robustos con la misma información que el catálogo inicial pero con un número reducido de atributos usados para la clasificación.

El catálogo robusto $\mathcal{D}^{-\mathcal{I}}$ / $\mathcal{I} = \{16, 23, 24, 32, 33, 41\}$ del dataset census contiene 137,655 evidencias y 35 atributos. Aplicando el algoritmo en forma supervisada obtenemos nuevos catálogos robustos, cada vez con menos atributos en sus caracterizaciones.

$\lvert\mathcal{D}^{-\mathcal{I}}\rvert$	$\lvert\mathcal{A}^{-\mathcal{I}}\rvert$	$\mathcal{K}$	$\lvert\mathcal{CCR}_{\mathcal{D}^{-\mathcal{I}},\mathcal{K}}\rvert$	**Evidencias robustas**	mín $\lvert\mathcal{D}^{-\mathcal{I}-\mathcal{J}}\rvert$	**Tiempo (sg.)**
137,655	35(29)(-)	1	30	4,129,073	137,284	24
		2	434	59,725,266	136,988	168
		3	4,031	554,645,993	136,967	1,425
		4	27,005	3,715,177,551	136,761	9,243

$\mathcal{D}^{-12-16-23-24-32-33-35-36-37-41}_{\texttt{census}}$ contiene 136,761 evidencias y 31 atributos. Aún contiene demasiados atributos redundantes por lo que se aplica de nuevo el algoritmo en forma supervisada.

$\lvert\mathcal{D}^{-\mathcal{I}}\rvert$	$\lvert\mathcal{A}^{-\mathcal{I}}\rvert$	$\mathcal{K}$	$\lvert\mathcal{CCR}_{\mathcal{D}^{-\mathcal{I}},\mathcal{K}}\rvert$	**Evidencias robustas**	mín $\lvert\mathcal{D}^{-\mathcal{I}-\mathcal{J}}\rvert$	**Tiempo (sg.)**
136,761	31(25)(-)	1	26	3,555,699	136,735	19
		2	324	44,308,317	136,693	134
		3	2,575	352,131,960	136,679	865
		4	14,655	2,004,017,232	136,666	4,813
		5	63,596	8,696,257,039	136,652	21.571

$\mathcal{D}^{-7-12-16-23-24-27-28-31-32-33-35-36-37-40-41}_{\texttt{census}}$ contiene 136,652 evidencias y 26 atributos. A pesar de que aún tiene muchas evidencias, tiene 20 atributos redundantes como máximo, por lo que se puede aplicar el algoritmo $\mathcal{ACDC}$ en forma no supervisada.

$\lvert\mathcal{D}^{-\mathcal{I}}\rvert$	$\lvert\mathcal{A}^{-\mathcal{I}}\rvert$	$\mathcal{K}$	$\lvert\mathcal{CCR}_{\mathcal{D}^{-\mathcal{I}},\mathcal{K}}\rvert$	**Evidencias robustas**	mín $\lvert\mathcal{D}^{-\mathcal{I}-\mathcal{J}}\rvert$	**Tiempo (sg.)**
136,652	26(17)(-)	∞	91,264	12,467,899,776	136,524	37,563

Los resultados obtenidos no son necesariamente óptimos para analizar el dataset de clasificación census. Los catálogos robustos seleccionados para obtener su $\mathcal{CCR}_{\mathcal{D}^{-\mathcal{I}},\mathcal{K}}$ generan catálogos robustos que sólo necesitan 12 de los 41 atributos del experimento para clasificar correctamente cualquier caracterización del dataset original. Si se hubieran seleccionado otros catálogos robustos, el número de atributos a utilizar como mínimo podría ser menor.

$\mathcal{D}_{\texttt{census}}^{-7-12-16-23-24-27-28-31-32-33-35-36-37-40-41}$ contiene 136,652 evidencias y 26 atributos. A pesar de que aún tiene muchas evidencias, tiene 20 atributos redundantes como máximo, por lo que se puede aplicar el algoritmo $\mathcal{ACDC}$ en forma no supervisada.

$\lvert\mathcal{D}^{-\mathcal{I}}\rvert$	$\lvert\mathcal{A}^{-\mathcal{I}}\rvert$	$\mathcal{K}$	$\lvert\mathcal{CCR}_{\mathcal{D}^{-\mathcal{I}},\mathcal{K}}\rvert$	**Evidencias robustas**	mín $\lvert\mathcal{D}^{-\mathcal{I}-\mathcal{J}}\rvert$	**Tiempo (sg.)**
136,652	26(17)(-)	∞	91,264	12,467,899,776	136,524	37,563

Los resultados obtenidos no son necesariamente óptimos para analizar el dataset de clasificación census. Los catálogos robustos seleccionados para obtener su $\mathcal{CCR}_{\mathcal{D}^{-\mathcal{I}},\mathcal{K}}$ generan catálogos robustos que sólo necesitan 12 de los 41 atributos del experimento para clasificar correctamente cualquier caracterización del dataset original. Si se hubieran seleccionado otros catálogos robustos, el número de atributos a utilizar como mínimo podría ser menor.

Índice

1 Introducción

2 Catálogos

3 Experimentación

4 Conclusiones y trabajo futuro
- Conclusiones
- Trabajo futuro

5 Publicaciones

Tras la investigación mostrada en este trabajo, podemos afirmar que:

- La Minería de Reglas de Asociación es una potente herramienta informática. Descubre patrones y relaciones que pueden ser aprovechados con un enfoque estadístico para descubrir reglas que se verifican en muchas situaciones reales.
- El soporte de una regla, y su confianza cuando se usa soporte mínimo, no debe usarse si no se trabaja con muestras representativas.
- Los catálogos no son muestras representativas. Deben analizarse con el objetivo de descubrir información sobre el problema de clasificación en estudio, no sobre la distribución de los ítems en la población estudiada.

Tras la investigación mostrada en este trabajo, podemos afirmar que:

- La Minería de Reglas de Asociación es una potente herramienta informática. Descubre patrones y relaciones que pueden ser aprovechados con un enfoque estadístico para descubrir reglas que se verifican en muchas situaciones reales.
- El soporte de una regla, y su confianza cuando se usa soporte mínimo, no debe usarse si no se trabaja con muestras representativas.
- Los catálogos no son muestras representativas. Deben analizarse con el objetivo de descubrir información sobre el problema de clasificación en estudio, no sobre la distribución de los ítems en la población estudiada.

Tras la investigación mostrada en este trabajo, podemos afirmar que:

- La Minería de Reglas de Asociación es una potente herramienta informática. Descubre patrones y relaciones que pueden ser aprovechados con un enfoque estadístico para descubrir reglas que se verifican en muchas situaciones reales.
- El soporte de una regla, y su confianza cuando se usa soporte mínimo, no debe usarse si no se trabaja con muestras representativas.
- Los catálogos no son muestras representativas. Deben analizarse con el objetivo de descubrir información sobre el problema de clasificación en estudio, no sobre la distribución de los ítems en la población estudiada.

Tras la investigación mostrada en este trabajo, podemos afirmar que:

- La Minería de Reglas de Asociación es una potente herramienta informática. Descubre patrones y relaciones que pueden ser aprovechados con un enfoque estadístico para descubrir reglas que se verifican en muchas situaciones reales.
- El soporte de una regla, y su confianza cuando se usa soporte mínimo, no debe usarse si no se trabaja con muestras representativas.
- Los catálogos no son muestras representativas. Deben analizarse con el objetivo de descubrir información sobre el problema de clasificación en estudio, no sobre la distribución de los ítems en la población estudiada.

Todos los datasets de clasificación contienen reglas de asociación del tipo "*Si un individuo tiene ciertas características, entonces pertenece a cierta clase*" con un 100 % de confianza. Sin embargo, el número de reglas de este tipo puede ser muy grande y desbordar las capacidades de los equipos con que se lleva a cabo el análisis.

Cualquier dataset de clasificación se puede dividir en conjuntos homogéneos respecto a las caracterizaciones que contiene:

1. Se separan las evidencias duplicadas en $\mathcal{D}_d$.
2. Se separan las evidencias con datos desconocidos en $\mathcal{D}_?$.
3. Se separan las caracterizaciones con incertidumbre en $\mathcal{D}_\mathcal{E}$.
4. El conjunto resultante, $\mathcal{D}$, contiene las evidencias robustas del dataset original, sin duplicados.

Todos los datasets de clasificación contienen reglas de asociación del tipo "*Si un individuo tiene ciertas características, entonces pertenece a cierta clase*" con un 100 % de confianza. Sin embargo, el número de reglas de este tipo puede ser muy grande y desbordar las capacidades de los equipos con que se lleva a cabo el análisis.

Cualquier dataset de clasificación se puede dividir en conjuntos homogéneos respecto a las caracterizaciones que contiene:

1. Se separan las evidencias duplicadas en $\mathcal{D}_d$.
2. Se separan las evidencias con datos desconocidos en $\mathcal{D}_?$.
3. Se separan las caracterizaciones con incertidumbre en $\mathcal{D}_\mathcal{E}$.
4. El conjunto resultante, $\mathcal{D}$, contiene las evidencias robustas del dataset original, sin duplicados.

El catálogo robusto $\mathcal{D}$, cuyas evidencias son reglas de asociación con un 100 % de confianza y $\mathcal{N}$ ítems en sus antecedentes, puede ser analizado en bloque, como matriz homogénea que es, para descubrir si contiene catálogos robustos con las mismas capacidades de clasificación pero usando menos atributos, obteniendo la colección de catálogos robustos $\mathcal{CCR}_{\mathcal{D}}$.

El número de reglas de clasificación asociativa que contiene el dataset de clasificación original, $\mathcal{B}$, puede ser mucho mayor que el número de reglas contenidas en el $\mathcal{CCR}_{\mathcal{D}}$. Sin embargo, no hay criterios de soporte que ayuden a comparar las reglas.

El catálogo robusto $\mathcal{D}$, cuyas evidencias son reglas de asociación con un 100 % de confianza y $\mathcal{N}$ ítems en sus antecedentes, puede ser analizado en bloque, como matriz homogénea que es, para descubrir si contiene catálogos robustos con las mismas capacidades de clasificación pero usando menos atributos, obteniendo la colección de catálogos robustos $\mathcal{CCR}_{\mathcal{D}}$.

El número de reglas de clasificación asociativa que contiene el dataset de clasificación original, $\mathcal{B}$, puede ser mucho mayor que el número de reglas contenidas en el $\mathcal{CCR}_{\mathcal{D}}$. Sin embargo, no hay criterios de soporte que ayuden a comparar las reglas.

Divulgar esta investigación entre investigadores que lleven a cabo problemas reales de clasificación.

Utilizar computación en paralelo para mejorar la eficiencia del algoritmo $\mathcal{ACDC}$.

Añadir información de coste/oportunidad sobre los atributos en estudio para optimizar la selección del catálogo robusto a usar.

En experimentos con incertidumbre, encontrar el conjunto de atributos que maximice la fiabilidad de las clasificaciones obtenidas.

Divulgar esta investigación entre investigadores que lleven a cabo problemas reales de clasificación.

Utilizar computación en paralelo para mejorar la eficiencia del algoritmo $\mathcal{ACDC}$.

Añadir información de coste/oportunidad sobre los atributos en estudio para optimizar la selección del catálogo robusto a usar.

En experimentos con incertidumbre, encontrar el conjunto de atributos que maximice la fiabilidad de las clasificaciones obtenidas.

Divulgar esta investigación entre investigadores que lleven a cabo problemas reales de clasificación.

Utilizar computación en paralelo para mejorar la eficiencia del algoritmo $\mathcal{ACDC}$.

Añadir información de coste/oportunidad sobre los atributos en estudio para optimizar la selección del catálogo robusto a usar.

En experimentos con incertidumbre, encontrar el conjunto de atributos que maximice la fiabilidad de las clasificaciones obtenidas.

Divulgar esta investigación entre investigadores que lleven a cabo problemas reales de clasificación.

Utilizar computación en paralelo para mejorar la eficiencia del algoritmo $\mathcal{ACDC}$.

Añadir información de coste/oportunidad sobre los atributos en estudio para optimizar la selección del catálogo robusto a usar.

En experimentos con incertidumbre, encontrar el conjunto de atributos que maximice la fiabilidad de las clasificaciones obtenidas.

Incorporar el soporte para ordenar las reglas obtenidas por utilidad en los experimentos basados en muestras representativas de la población en estudio.

Crear un sistema experto capaz de gestionar la información descubierta en un dataset de clasificación en explotación, que incorpore las anteriores propuestas y actualice la información descubierta cuando se incorporen nuevas tomas de datos.

Crear aplicación para dispositivos móviles que permita el uso en el trabajo de campo de la información generada por el sistema experto anterior.

Incorporar el soporte para ordenar las reglas obtenidas por utilidad en los experimentos basados en muestras representativas de la población en estudio.

Crear un sistema experto capaz de gestionar la información descubierta en un dataset de clasificación en explotación, que incorpore las anteriores propuestas y actualice la información descubierta cuando se incorporen nuevas tomas de datos.

Crear aplicación para dispositivos móviles que permita el uso en el trabajo de campo de la información generada por el sistema experto anterior.

Incorporar el soporte para ordenar las reglas obtenidas por utilidad en los experimentos basados en muestras representativas de la población en estudio.

Crear un sistema experto capaz de gestionar la información descubierta en un dataset de clasificación en explotación, que incorpore las anteriores propuestas y actualice la información descubierta cuando se incorporen nuevas tomas de datos.

Crear aplicación para dispositivos móviles que permita el uso en el trabajo de campo de la información generada por el sistema experto anterior.

Índice

1 Introducción

2 Catálogos

3 Experimentación

4 Conclusiones y trabajo futuro

5 Publicaciones

Los resultados obtenidos en esta investigación se han publicado en diferentes ponencias y artículos.

Botella, F. and Lazcorreta, Enrique and Fernández-Caballero, Antonio and González, Pascual
Mejora de la usabilidad y la adaptabilidad mediante técnicas de minería de uso web
Actas del VI Congreso Internacional Interacción Persona-Ordenador (Interacción'05), 299–306, 2005.

Botella, F. and Lazcorreta, Enrique and Fernández-Caballero, Antonio and González, Pascual
Personalization through Inferring User Navigation Maps from Web Log Files
Proceedings of the 11th International Conference on Human-Computer Interaction (HCII'05), Las Vegas, Nevada (EEUU), 2005.

Botella, F. and Lazcorreta, Enrique and Fernández-Caballero, Antonio and González, Pascual and Gallud, José A. and Bia, Alejandro
Selecting the Best Tailored Algorithm for Personalizing a Web Site
Proceedings of the 12th International Conference on Human-Computer Interaction (HCII'07), Beijing (China), 2007.


Lazcorreta Puigmartí, Enrique and Botella, F. and Fernández-Caballero, Antonio
Towards personalized recommendation by two-step modified Apriori data mining algorithm
Expert Systems with Applications 35(3), 1422–1429, 2008.


Lazcorreta Puigmartí, Enrique and Botella, F. and Fernández-Caballero, Antonio
Reglas de Oportunidad: mejorando las recomendaciones web
Actas del X Congreso Internacional Interacción Persona-Ordenador (Interacción'09), 2009.


Lazcorreta Puigmartí, Enrique and Botella, F. and Fernández-Caballero, Antonio
Recomendaciones en sistemas web mediante el estudio de ítems raros en transacciones
Actas del XI Congreso Internacional Interacción Persona-Ordenador (Interacción'10), 385–388, 2010.


Lazcorreta Puigmartí, Enrique and Botella, F. and Fernández-Caballero, Antonio
Efficient Analysis of Transactions to Improve Web Recommendations
Actas del XIII Congreso Internacional Interacción Persona-Ordenador (Interacción'12), 2012.


Lazcorreta Puigmartí, Enrique and Botella, F. and Fernández-Caballero, Antonio
A novel characterisation-based algorithm to discover new knowledge from classication datasets without use of support
Expert Systems with Applications [En 2ª revisión], 2017.

¡ Gracias por vuestra atención !

Esta presentación, el manuscrito y código de la tesis defendida con su ayuda están disponibles en

`https://github.com/EnriqueLazcorreta/tesis-doctoral`