Mallea, Adriana – Gallardo Vanesa
Una introducción a datos simbólicos / Adriana Mallea. - 1a ed. - San Juan : Universidad Nacional de San Juan. Facultad de Filosofía, Humanidades y Artes, 2021.
Libro digital, PDF

Archivo Digital: online
ISBN 978-950-605-915-6

1. Análisis de Datos. I. Título.
CDD 510.1

Editor: effha
Jefe Departamento Publicaciones: Alfredo Ginbert
Publicación autorizada por el Consejo Editorial de la Facultad de Filosofía, Humanidades y Artes
Edición: primera
Impreso en San Juan Argentina – Printed in San Juan, Argentina
Hecho el depósito que determina la Ley 11.723

ISBN: 978-950-605-915-6

UNIVERSIDAD NACIONAL DE SAN JUAN

Facultad de Filosofía, Humanidades y Artes

Una Introducción a Datos Simbólicos

Vanesa Gallardo-Adriana Mallea

Año: 2015

Índice

Prólogo

Esta Monografía es resultado de las tareas de investigación desarrolladas por la alumna becaria Vanesa Gallardo en su Beca de Estímulo a Investigadores Jóvenes, otorgada por el Consejo Interuniversitario Nacional (CIN), bajo la dirección de la Dra. Adriana Mallea. La beca se enmarca en el proyecto Determinación y comparación de perfiles sociales y culturales de estudiantes universitarios a través de Técnicas Estadísticas Multivariadas (2014-2015).

Se encuadra en el Análisis de Datos Simbólicos, que permite el análisis de conocimientos. La extracción de conocimientos del Análisis de Datos clásico es la obtención de resultados por sí mismos, explicativos, que representan conceptos. En esta monografía se presenta la formalización de los conceptos mediante objetos simbólicos que constituyen la entrada y la salida de las técnicas de Análisis de Datos Simbólicos. Además, estos objetos simbólicos permiten consultas a una base de datos, facilitando la propagación de los conceptos.

El Análisis de Datos Simbólicos permite la extensión de la Estadística Clásica a la Estadística de las intenciones o conceptos y la extensión de problemas, métodos y algoritmos de Análisis de Datos a datos simbólicos. Según Diday [7] el Análisis de Datos Simbólicos crea un puente entre la Estadística y el Aprendizaje Automático.

Ya Aristóteles en su Organón (Aristotle, IV a.C.) distingue entre un individuo y la descripción del mismo. Si bien los conceptos de intención y extensión se deben a Arnauld y Nicole (Arnauld y Nicole, 1662) es Diday (Diday (1987,1988)) quien formaliza estos términos del Análisis de Datos Simbólicos. La intención de un concepto constituye su descripción, mientras que la extensión es el conjunto de individuos cuya descripción es acorde a la del concepto. La intención, que se representa por un objeto simbólico, se describe por los datos simbólicos y por un mecanismo de reconocimiento de los individuos de la extensión.

Diday introduce los objetos simbólicos y presenta una formalización que permite tratar conocimientos más ricos que los datos habituales, y establece una relación con el modelo clásico de Análisis de Datos (Diday, 1991). Un objeto simbólico representa una intención, un concepto y se define, en términos generales, como una conjunción de valores o conjuntos de valores, correspondientes a variables; que pueden ser ponderados. Constituye una descripción en intención de una clase de individuos que constituyen la extensión.

Según Diday [8], el Análisis de Datos Simbólicos nace influido por tres campos:

- el Análisis Exploratorio de Datos,
- la Inteligencia Artificial, donde se realiza un gran esfuerzo por el desarrollo de lenguajes de representación de conocimiento,
- la Taxonomía Numérica usada en las Ciencias Biológicas.

La formalización de los objetos simbólicos ya ha evolucionado desde sus inicios (Diday (1987, 1988, 1991, 1993a, 1993b)). Esta monografía presenta la adoptada en el libro editado por Bock y Diday (Bock y Diday, 2000a), que según sus editores es la primera monografía sistemática y completa del Análisis de Datos Simbólicos.

Las ventajas del Análisis de Datos Simbólicos son:

1. forma de representación del conocimiento en lenguaje fácilmente comprensible al usuario,
2. análisis estadístico de datos que representan intenciones o conceptos. Así, como el Análisis de Datos extrae conocimientos, el Análisis de Datos Simbólicos extrae nuevos conocimientos a partir de conocimientos previos,
3. extensión de las técnicas del Análisis de Datos a los datos simbólicos,

4. se representan y analizan datos de mayor complejidad que los datos tradicionales ya que contienen variación interna, como los intervalos, conjuntos de valores, distribuciones de probabilidad, etc. Y, además, son estructurados como las taxonomías y las dependencias jerárquicas y lógicas entre variables. Es decir, los datos simbólicos contienen además metadatos,

5. los resultados de los análisis se interpretan fácilmente en el lenguaje del usuario,

6. los objetos de entrada y salida de las técnicas de Análisis de Datos Simbólicos son representados por un único formalismo, comprensible al usuario,

7. los objetos simbólicos pueden venir dados por el conocimiento de un experto,

8. las intenciones se pueden extraer de bases de datos reagrupando o agregando datos individuales (Stéphan et al., 2000),

9. cada intención viene acompañada de una extensión que es el conjunto de individuos de una base de datos que se adecuan a la intención. Una misma intención puede aplicarse (es una consulta) a diversas bases de datos o a una misma base de datos en distintos momentos de tiempo, facilitando la propagación de conceptos,

10. desde un punto de vista formalista, se presenta una representación unificada de aproximaciones a la incertidumbre: datos expresados como distribuciones de probabilidad, de posibilidad, conjuntos difusos, creencias (Diday, 1995a),

11. se preserva la confidencialidad de los datos individuales, al analizar agrupaciones de los mismos,

12. se da una solución a la selección de información y almacenamiento de datos de un DataWarehouse, aportándose a las Oficinas de Estadística un medio de extracción de conocimientos de sus grandes bases de datos.

Michalski (Michalski, 1983) define la Inferencia Inductiva como el proceso de ir de un conocimiento derivado de la observación de algunos objetos a un conocimiento más estructurado en forma de complejos que son soportados con diversa intensidad por los datos observados. Una de las técnicas de inferencia inductiva es el Aprendizaje Automático a partir de ejemplos, en el cual el conocimiento observado se compone de unos objetos de clases conocidas y el conocimiento estructurado derivado del proceso se expresa por un conjunto de reglas que permiten la clasificación de éstos y de nuevos ejemplos de clases desconocidas.

Quinlan define el Aprendizaje como la adquisición de conocimiento estructurado en forma de conceptos, redes de discriminación o reglas de producción (Quinlan, 1986a, Wu, 1993).

La Sección 1 presenta una introducción al Análisis de Datos Simbólicos y sus ventajas sobre el Análisis Multidimensional de Datos clásico.

En la Sección 2 se hace una presentación de los objetos simbólicos (OS) de una manera informal, indicando y ejemplificando el tipo de variables simbólicas que se pueden definir y se introducen las tablas simbólicas. Además se clasifican los OS según su construcción; la obtención a partir de una matriz de datos clásica y posibilidades de visualización a partir del software SODAS (Symbolic Official Data Analysis System), un software realizado con la cooperación de 17 grupos de investigación europeos y tres Institutos Nacionales de Estadística: EUSTAS/ España – INE/ Portugal – ONS/ Londres.

En la Sección 3 se comienza a formalizar el concepto de OS a partir de su génesis y primeras definiciones de E. Diday.[2]

En la Sección 4 se realiza una introducción al Análisis de Datos Simbólicos y al software SODAS. [7]

La Sección 5 presenta formalmente los conceptos básicos del Análisis de Datos Simbólicos e introduce los datos y los objetos simbólicos que representan una formalización única para los datos complejos y la incertidumbre.

1. Introducción

En el desarrollo de esta monografía, trataremos el tema: "Análisis de Objetos Simbólicos", una nueva unidad estadística utilizada en el campo de la Estadística, que pretende resumir una gran cantidad de información almacenada en bases de datos, que describen tanto individuos como grupos de una población, a fin de realizar un análisis estadístico posterior.

A través del análisis de Datos Simbólicos, podemos obtener información en forma mas detallada y analizar múltiples variables. Si por ejemplo consideramos como una base de datos la información obtenida a partir de una encuesta, podemos agrupar individuos provenientes de diferentes categorías sociales. En términos del ADM clásico se pueden construir clases de manera mas objetiva y modelizar lo menos posible, pero en este caso se desperdician los conocimientos de expertos, tales como psicólogos, sociólogos, que no pueden ser plasmados en los cuestionarios. Si nos ubicamos por otro lado en el punto de vista de Análisis de Datos Simbólicos (ADS) sí podemos apoyarnos mas en el conocimiento de expertos. Podemos considerar experiencias, conocimientos y distintos comportamientos que serán descriptos en términos no solo de individuos sino por un conjunto de propiedades. Con el Análisis de Datos Simbólicos, podemos describir con una mayor intensidad los datos e información obtenida, nos permite observar cuál es la extensión de cada descripción en la base de datos.

La estadística clásica se interesa sobre todo por la modelización de una población vista globalmente, el ADM generaliza el análisis multivariado y comienza a interesarse por los individuos. El ADS se interesa también en los objetos en sentido general, mas en cuanto a objetos que en individuos, tiene como objetivo reemplazar los individuos del análisis de datos tradicional por individuos de mas alto nivel, mas complejos y aptos para representar estos conocimientos, porque están definidos en intensión, utilizando el poder de la lógica. Estos individuos de mas alto nivel, son los objetos simbólicos (OS). Las variables en este caso son de mas alto nivel ya que no toman solo un valor por celda, sino que pueden tomar varios valores. El ADS, permite que los conocimientos de los expertos sean expresados en los datos mismos, encontrar expresiones matemáticas que permitan transformar frases que expresan experiencia en forma de datos, siendo estos datos de mas alto nivel.

En el ADS, Ω no es simplemente el conjunto de individuos, ni un conjunto de valores tomados por los individuos sino que es el conjunto de eventos o sucesos. Los axiomas de la teoria de probabilidad, se aplican en este caso a los eventos. En lugar de tener individuos se tiene objetos simbólicos, que pueden considerarse eventos.

Este trabajo de investigación se enmarca en el proyecto "Determinación y comparación de perfiles sociales y culturales de estudiantes universitarios a través de Técnicas Estadísticas Multivariadas", el cual pretende emplear la metodología del ADS y su aplicación, con el objeto de caracterizar a grupos de estudiantes universitarios a partir de variables adecuadas. Las mismas son características sociales y culturales de interés, para la gestión académica e investigaciones, que tienen a la universidad como objeto de estudio. Se pretende de este modo caracterizar el perfil social y cultural tanto de los estudiantes ingresantes, como de los que permanecen y egresan de la universidad.

2. Definiciones y extensiones de objetos simbólicos

2.1. ¿Qué son los objetos simbólicos?

Los objetos simbólicos son especies de átomos de conocimientos, comprenden un campo tan vasto como los conocimientos mismos, se plantean como nuevas unidades de análisis que pretenden resumir una gran cantidad de información almacenada en bases de datos, describiendo tanto individuos como grupos.

Pueden verse como una representación de conceptos estadísticos que permiten el análisis de datos agregados a partir de la combinación de variables seleccionadas al analizar grandes matrices de datos. Cada objeto puede representar un grupo de individuos con caractarísticas comunes que resulten del cruce de variables.

Los objetos simbólicos que van a resolver problemas de respuesta no unitaria, es decir, los valores que toman las variables pueden ser no atómicos (un grupo de valores, un intervalo de valores o una distribución de probabilidad). Así pues, un Objeto Simbólico es un modo de representación de datos complejos que surge al analizar grandes ficheros de datos.

En el ADS en lugar de tener un conjunto de individuos tenemos un conjunto de objetos simbólicos que están expresados por un conjunto de propiedades, permitiendo expresar una mayor cantidad de conocimiento. Su objetivo es extender el análisis de datos clásico al estudio de objetos mas complejos que se expresan bajo forma de "conjunción" de propiedades aplicadas sobre las variables clásicas: continuas, nominales u ordinales.

Veamos los tipos de variables que los predican:

a) Cada variable puede tomar múltiples valores para un mismo objeto simbólico, por ejemplo: [opinión = {regular, mala, regular $\wedge$ indiferente}] para expresar el hecho de que una clase de individuos puede tener opinión regular, mala, o regular e indiferente. O bien, [edad = [17, 29]] para indicar que los individuos encuestados tienen entre 17 y 29 años. Estos valores no son una modalidad mutuamente excluyente, para no perder la información contenida.

b) Como consecuencia de a) se expresan diferentes tipos de relaciones entre las variables: cuando una variable toma una modalidad, la otra puede no tener sentido (no se describen las computadoras de una empresa que no las posee) o se debe restringir su campo de valores posibles. Se obtienen así, objetos simbólicos provistos de propiedades, variables llamadas madre - hija.

Los OS se distinguen también a nivel de su manipulación:

c) Un objeto simbólico es una descripción en intensión de una clase de objetos de la cual constituyen la extensión. El objeto [categoría = {obrero, empleado}] tiene por extensión todos los objetos en los cuales la categoría sea obrero o empleado.

d) Como consecuencia de c) se puede generalizar o especificar un objeto simbólico modificando sus propiedades de manera ya sea de extender o de restringir su extensión.

e) Para generalizar se utilizan las operaciones de unión, intersección y de complementación. Por ejemplo, para generalizar "cualquiera que beba whisky y agua" y "cualquiera que beba vino y agua" con: "cualquiera que beba alcohol y agua".

2.2. Necesidad de objetos simbólicos

Presentamos algunos ejemplos, no relacionados con la encuesta Kolla, que muestran la necesidad de utilizar objetos simbolicos:

Para un grupo de individuos, consideremos la variable Y = "Minutos dedicados a la práctica de deporte al día"; es una variable que permite una respuesta no unitaria, ya que varía de día a día.

Para un individuo k, esa variable puede expresarse de una forma no clásica:

$Y(k) = [20, 60]$

$Y(k) = \{20\text{minutos}(0{,}15), 30\text{minutos}(0{,}45), 45 \text{ minutos}(0{,}1), 60\text{minutos}(0{,}3)\}$

$Y(k) = \{\text{Participación Nula}(0{,}1), \text{Part. Escasa}(0{,}5), \text{Part. Media}(0{,}3), \text{Part. Alta}(0{,}1)\}$.

Para clases de individuos, si k denota la provincia 'San Juan', la variable Y = "Relación con la Actividad" puede ser especificada por:

$Y(k) = \{\text{Ocupado}(0{,}47), \text{Desocupado}(0{,}11), \text{Jubilado}(0{,}42)\}$, que indica que el 47 %de los individuos de San Juan están ocupados, el 11 % desocupados, etc; por ejemplo.

Además, podemos encontrarnos con estudios que no estén basados en resultados experimentales o de encuestación únicos, sino que tienen en cuenta alguna inexactitud. De aquí surgen otro tipo de objetos simbólicos basados en resultados imprecisos: los datos probabilísticos o posibilísticos, los datos difusos, o los intervalos que pueden resultar de dos fuentes: observaciones o de un conocimiento experto. Existen diversas situaciones en las que la asociación de un único valor a un único individuo o a una única clase de individuos resulta en extremo limitada y no representa satisfactoriamente los contextos reales, más complejos. En estas circunstancias, los objetos simbólicos se muestran necesarios.

2.3. Método

La utilización de Objetos Simbólicos fue propuesta por E. Diday, el proceso de creación de objetos simbólicos tiene como punto de partida consultas a una base de datos relacional. Por medio de estas consultas se extraen automáticamente grupos de individuos con características comunes, como, por ejemplo, familias, regiones, etc. Es decir, cada objeto simbólico puede describir un grupo o una clase de individuos.

Los objetos simbólicos creados son también almacenados en tablas, llamadas tablas simbólicas. Cada celda de estas tablas, con objetos simbólicos por filas y variables por columnas, puede contener datos de diferentes tipo, tales como:

Un valor cuantitativo: edad (w) = 23

Un valor cualitativo: sexo (w) = mujer

O bien, de varios valores, ya sean:

- Cuantitativos

Número de hijos (w) = {no responde, uno, dos, tres o mas}

- Cualitativos

Estado civil (w) = {no responde, soltero, casado, unido civilmente, unido de hecho, divorciado, separado, viudo};

- Intervalo

Edad (w) = [18, 25] que significa que la edad de w varía entre 20 y 25;

- Varios valores con pesos: Número de hijos (w) = [no responde (0,73) ; uno (0,15) ; dos (0,06) ; tres o mas (0,06)], que puede ser un histograma o una función de pertenencia; siendo edad, sexo, estado civil y peso variables y w unidades.

Veamos una tabla simbólica:

OS	Sexo	Busca trabajo	Estado Civil	Titulo
OS1	{masculino(0,19), femenino(0,81)}	{no responde(0,14), si(0,75), no(0,11)}	{soltero(0,74), casado(0,14), otro(0,12)}	{prof.letras(0,10), prof.matematica(0,12), ..., lic.turismo(0,08)}
OS2	Femenino	{no responde(0,16), si(0,78), no(0,06)}	{soltero(0,72), casado(0,13), otro(0,15)}	{prof.letras(0,10), prof.matematica(0,11), ..., lic.turismo(0,10)}
OS3	Masculino	{no responde(0,06), si(0,63), no(0,31)}	{soltero(0,75), casado(0,19), otro(0,06)}	{prof.letras(0,06), prof.matematica(0,13), ..., lic.turismo(0,13)}
OS4 (45e)	{masculino(0,22), femenino(0,78)}	{no responde(0,11), si(0,78), no(0,11)}	{soltero(0,73), casado(0,16), otro(0,11)}	{prof.letras(0,09), prof.matematica(0,13), ..., lic.turismo(0,11)}

En la tabla cada objeto simbólico (por filas) representa un grupo de individuos con características comunes descritas por 3 variables. Así, por ejemplo, el objeto simbólico OS 2 representa a una clase de individuos formada por un 35 % de mujeres, 50 % de varones y un 10 % de encuestados que no respondieron, cuyo número de hijos es uno, dos, tres o más, o bien que no respondieron y de nacionalidad Argentina en un 65 %, del Mercosur un 20 %, otra 0,05 % y que no respondieron un 0,10 %.

Las variables que describen los objetos simbólicos pueden ser a su vez:

Variables con dominio Taxonómico: Si ofrecen la posibilidad de definir una jerarquía en los valores que toma la variable. Esta taxonomía representa un conocimiento a priori de los datos.

Estado civil = soltero, no soltero (casado, viudo, divorciado/separado).

Variables Madre-Hija (o Dependencias Jerárquicas): Si ofrecen la posibilidad de definir variables que no son aplicables a todos los individuos, pero sí lo son a individuos que verifican algunas propiedades.

SI condición laboral = no trabaja ENTONCES Tipo de Contrato es no aplicable.

Variables con Dependencias Lógicas (o Reglas): Si ofrecen la posibilidad de definir conocimiento a priori de los datos en forma de restricción de las posibles combinaciones de valores para diferentes variables.

SI edad>65 ENTONCES Situación profesional = Jubilado

2.4. Tipos de objetos simbólicos según su construcción

- Según el nivel de agregación: los Objetos Simbólicos pueden describir individuos tanto como clases de individuos. Según esto, podemos distinguir objetos simbólicos de:

a) **Primer Orden**: Se dice que los objetos simbólicos son de primer orden cuando los datos se refieren a individuos. Por ejemplo, la variable Y = "Edad" para cada alumno k de la universidad: Y(k) = {20} o Y(k) = [18, 22]

b) **Segundo Orden**: Se dice que los objetos simbólicos son de segundo orden (objetos agregados) cuando los datos se refieren a clases de individuos más o menos homogéneos. Como no todos los individuos de la misma clase toman el mismo valor en cada variable, habrá varias categorías que se aplicarán simultáneamente a la clase, normalmente con porcentajes especificados. Ahora k denota una clase de individuos, consideramos los datos de 85 encuestados de la encuesta Kolla, la variable Y = Çantidad de hijos"puede ser especificada por: Y(k) = {No responde (0.73), Uno(0.15), Dos(0.06), Tres o mas(0.06)}, que indica que el 73 % de los encuestados no respondió, el 15 % tiene solo un hijo, el 0,06 % tiene dos hijos y otro 0,06 % tienen tres o mas hijos.

- Según el número de variables clasificadoras: las clases de individuos pueden crearse, aparte de con conocimiento experto o análisis previos para formar grupos, mediante una sola variable o combinación de varias. De esta forma, tenemos:

a) **Atributo de Grupo Simple**: la formación de los grupos se hace mediante una sola variable. Se obtendrán tantos objetos simbólicos como modalidades tenga esa variable. Consideremos por ejemplo, a partir de la encuesta Kolla, las variables .estado_civil", çondición_laboral", "la entidad donde trabaja es", "volvería a estudiar". Con esta consulta, se van a obtener 8 objetos simbólicos que describen el estado civil de los escuestados: "no responde", "soltero", çasado", ünido civilmente", ünido de hecho", "divorciado", "separado", "viudo". Veamos uno de ellos:

OS: çasado-[condición_laboral ={"no responde"(..), "trabaja en relación con la profesión"(..), "trabaja y no tiene relación con la profesión"(..), "no trabaja"(..)}] ∧ [la entidad donde trabaja es ={"no responde"(..), "pública"(..), "privada"(..), .ambas"(..)}] ∧ [volvería a estudiar ={ "no responde"(..), "si, la misma carrera"(..), "si, en otra carrera"(..), "no"(..)}]

Donde entre paréntesis, en los puntos suspensivos, se coloca el porcentaje observado de cada modalidad de la variable involucrada.

b) **Atributo de Grupo Compuesto**: Si el atributo de grupo se compone del cruce de dos o más variables nominales. Se obtendrán tantos objetos simbólicos como producto de modalidades de las variables. Consideremos en este caso, según variables de encuesta Kolla, estado_civil & sexo, çondición_laboralz "volvería a estudiar". Veamos uno de ellos:

- OS: çasado / masculino- [condición laboral ={"no responde"(..), "trabaja en relación con la profesión"(..), "trabaja y no tiene relación con la profesión"(..), "no trabaja"(..)}] ∧ [volvería a estudiar ={ "no responde"(..), "si, la misma carrera"(..), "si, en otra carrera"(..), "no"(..)}]

2.5. Esquema del análisis de datos simbólicos

Hay cuatro tipos de análisis según cuales sean las entradas que se posean y según las salidas que se obtengan:

1. Datos clásicos en la entrada y resultados numéricos en la salida, es el ADM clásico

2. Datos clásicos en la entrada y una salida con resultados de orden simbólico o objetos simbólicos

   ¿Cómo transformar una clase obtenida con un método de clasificación clásico en un objeto simbólico?

Una forma simple, es tomar una unión de los valores alcanzados en la clase para cada variable.

¿Cuál es el interés de tener una interpretación simbólica de un análisis clásico?

El interés principal es el grado de explicación. Así obtenemos una nueva forma de extracción de conocimientos que se agrega a la cantidad que provee el análisis clásico, vamos a tener una conjunción de propiedades que provee el OS, con un mayor poder explicativo, que no reemplaza a los resultados habituales, pero adopta otra manera de interpretar los resultados.

3. Objetos simbólicos en la entrada y se extraen informaciones de orden numérico

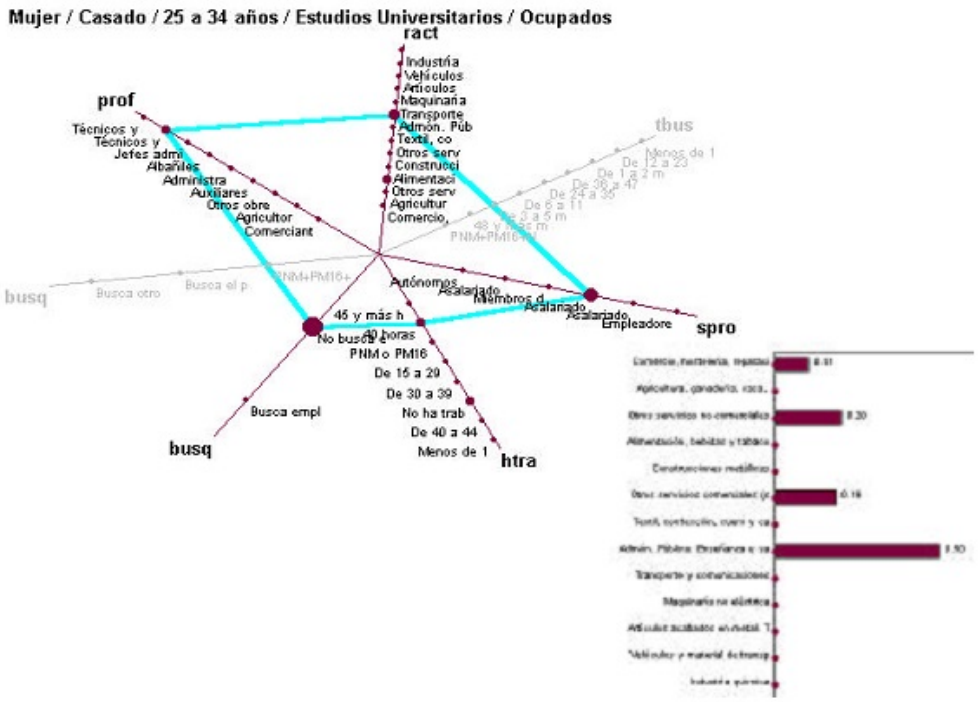
4. Objetos simbólicos en la entrada y en la salida.

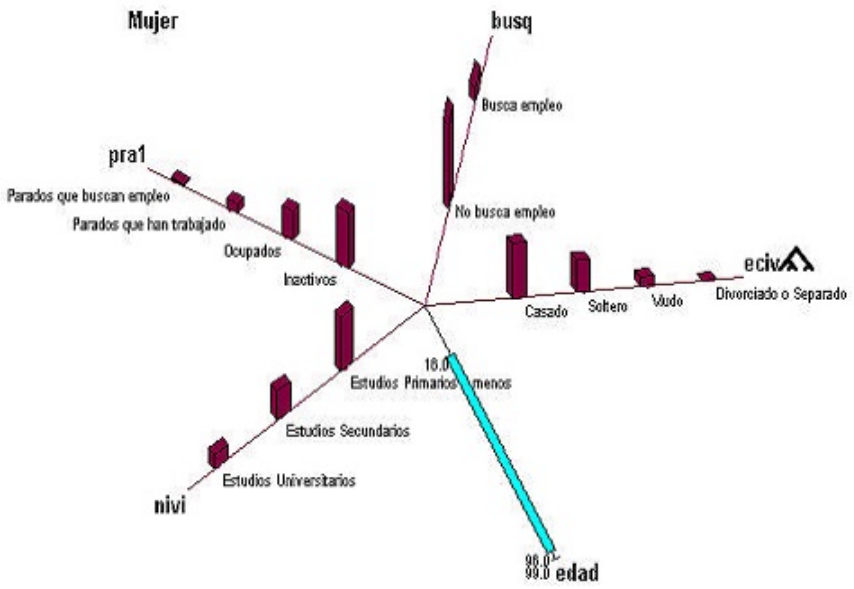
2.6. Posibilidades gráficas

La visualización de un objeto simbólico puede realizarse a través de un gráfico, esta representación esta basada en los diagramas de Kiviat, donde cada eje representa una variable. El software SODAS da la posibilidad de dos tipos de representación, en 2 dimensiones y en 3 dimensiones, la primera da una visión mas general mientras que la segunda brinda información con mas detalle al proveer simultaneamente las distribuciones de frecuencias de todas las variables.

En 2D los ejes están unidos por una línea que conecta los valores mas frecuentes de cada variable. Así se pueden comparar las distribuciones de frecuencias de dos OS, a partir de la forma que toma esta línea de conexión.

Veamos algunas imágenes de gráficos que podemos obtener:





3. Formalizando los objetos simbólicos

En la medida que los datos se hacen mas complejos, dejando de ser solo numéricos, pasando a mezclarse con datos de naturaleza cualitativa, con subjetividad, impresición y demas elementos, resulta mas dificultoso extraer información útil. Lo que interesa, son las similaridades y diferencias entre grupos poblacionales que atiendan a factores sociales, económicos, culturales, de salud y otros.

A finales de los años 70, R.S. Michalski introdujo un conjunto de ideas que dieron origen a lo que se denominó agrupamiento conceptual. Se pretendía arrojar mas información acerca de los agrupamientos, y así caracterizarlos a partir de las propiedades que cumplen. A partir de esto, E. Diday y un grupo de científicos han estado desarrollando trabajos en torno al concepto de objeto simbólico.

3.1. Génesis del concepto de objeto simbólico

El origen de los objetos simbólicos esta dado por los trabajos desarrollados por Michalski, cuya finalidad era: dada una colección de objetos, construir agrupamientos o subcategorías de los mismos, caracterizados por descripciones conjuntistas de propiedades formadas a partir de los atributos de los objetos.

Según Michalski, tenemos:

$X_1, X_2, ..., X_n$ variables discretas que describen objetos de la muestra. Para cada variable se define un conjunto de valores admisibles, el mismo contiene todos los posibles valores que esta variable puede tomar para cualquier objeto de la muestra. Se asume que estos conjuntos de valores son finitos y que pueden ser representados como:

$$D_i = \{0, 1, ..., d_{i-1}\} \ , \ i = 1, 2, ..., n$$

Estos conjuntos de valores pueden diferir en su tamaño y en la estructura relacionada con sus elementos.

En los trabajos desarrollados por Michalski y colaboradores, solo se distingue entre variables nominales y lineales, cuyos dominios son no ordenados y conjuntos linealmente ordenados respectivamente.

Un **evento**, se define como cualquier secuencia de valores de las variables $X_1, X_2, ..., X_n$ y se denota

$$e = (r_1, r_2, ..., r_n) \text{ donde } r_i \in D_i \ , \ i = 1, 2, ..., n$$

El conjunto de todos los posibles eventos, es llamado **espacio de eventos** denotado por

$$S = \{e_1, e_2, ..., e_d\}$$

donde $d = d_1$x d_2 x ... x d_n es el tamaño del espacio de eventos.

Una proposición relacional, del cálculo proposicional clásico $[X_i \# R_i]$ donde R_i es un conjunto de uno o mas elementos de la variable X_i y $\#$ simboliza los operadores relacionales $\geq, >, \leq, <, =, \in$ o sus negaciones, es llamada **selector**.

En el caso de las variables lineales, el operador " = " en $[X_i = R_i]$ puede ser reemplazado por los operadores relacionales $\geq, >, \leq, <$ para un R_i apropiado. Veamos algunos ejemplos de selectores:

- $[número\_hijos \geq 1]$ es decir, el número de hijos es mayor o igual a 1
- $[nacionalidad = argentina, mercosur, otra]$ es decir, la nacionalidad es argentina, del mercusur u otra
- $[estado\_civil \neq soltero]$ es decir, el estado civil no es soltero
- $[cantidad\_horas\ de\ trabajo\_semanales = 15..,45]$ es decir, la cantidad de horas de trabajo semanales varían entre 15 y 45,

A los selectores, E. Diday les llama eventos.

Un producto lógico de selectores es llamado un **complejo lógico** $(l - complejo)$, si los valores de las variables en e satisfacen todos los selectores en el complejo se dice que e satisface al $l - complejo$. Por ejemplo, el evento $e = (2, 7, 0, 1, 5, 4, 6)$ satisface al $l - complejo$ $[X_1 = 2, 3]$, $[X_3 \leq 3]$, $[X_5 = 3..,8]$. En este ejemplo estamos preservando las notaciones de Michalski.

Cualquier conjunto de eventos para los cuales exista un $l - complejo$ que estos satisfagan y sólo ellos, es llamado un **conjunto complejo** $(s - complejo)$.

3.2. Los primeros conceptos de objeto simbólico

En el análisis de datos clásico la agrupación de objetos se logró por la minimización de la disimilaridad dentro de cada agrupamiento y maximizando la disimilaridad entre agrupamientos, con los objetos simbólicos E. Diday pretende lograr extensiones, no individualizar los objetos como en conjuntos de datos clásicos, sino unificarlos por medio de vínculos, propiedades. Los objetos simbólicos son mucho mas complejos:

1. Todos los objetos de un conjunto de datos simbólicos pueden no estar definidos en términos de las mismas variables
2. Cada variable, en la descripción de un objeto simbólico puede tomar mas de un valor, incluso un conjunto infinito de valores

3. En objetos simbólicos mas complejos, los valores que toman las variables pueden incluir uno o más objetos simbólicos elementales

4. La descripción de un objeto simbólico puede depender de las relaciones que existan entre otros objetos simbólicos

5. Los valores que las variables toman pueden estar tipificados, indicando la frecuencia de ocurrencia, probabilidad relativa, nivel de importancia de valores, etc.

Diday, en sus trabajos nos brinda varias definiciones, que son necesarias recordar.

Un **evento** es una pareja $(variable - valor)$ que enlaza las variables y valores de las mismas en los objetos. Veamos algunos ejemplos de eventos:

$e_1 = [conocimiento\ de\ inglés = \{no\ responde, bueno, básico, no\ tiene\}]$

$e_2 = [título = profesor\ de\ matemática]$

$e_3 = [cantidad\_horas\ de\ trabajo\_semanales = [15..,45]]$.

En la terminología de Diday, los **objetos simbólicos** son definidos por una conjunción lógica de eventos. Así, en la formulación de Michalski, un $l - complejo$ es un objeto simbólico.

Diday introduce diferentes tipos de objetos simbólicos:

- Un **objeto aseverativo** es una conjunción de eventos pertenecientes a las descripciones individuales de objetos reales. Por ejemplo:

  $a = e_1 \wedge e_2 \wedge e_3 = [conocimiento\ de\ inglés = \{no\ responde, bueno, básico, no\ tiene\}] \wedge$
  $[título = profesor\ de\ matemática] \wedge [cantidad\_horas\ de\ trabajo\_semanales = [15..,45]]$

  Aquí, a es un objeto simbólico aseverativo, que verifica las propiedades que el conocimiento de inglés es bueno, básico, no tiene o no responde, el título es de profesor de matemática y la cantidad de horas semanales trabajadas está entre 15 y 45.

- Un **objeto simbólico acumulativo** es una conjunción de 2 o mas objetos simbólicos aseverativos y eventos. Por ejemplo:

  $h = [[título(enc1) = licenciado\_letras] \wedge$
  $[conocimiento\_inglés(enc1) = \{bueno, básico, no\_tiene\}]] \wedge$
  $[[título(enc2) = profesor\_física] \wedge$
  $[conocimiento\_inglés(enc2) = \{bueno, básico, no\_tiene\}]]$

  Esto significa que el objeto h consiste en la conjunción de los dos objetos aseverativos:

1. Encuestado 1, licenciado en letras y su nivel de inglés es bueno, básico o no tiene

2. Encuestado 2, profesor de física y su nivel de inglés es bueno, básico o no tiene

- Un **objeto simbólico tipo síntesis**, es una conjunción de 2 o más objetos simbólicos acumulativos y eventos.

3.3. Definición formal de E. Diday

A partir de estas ideas iniciales, Diday propone una formalización de objeto simbólico. Consideremos el siguiente diagrama:

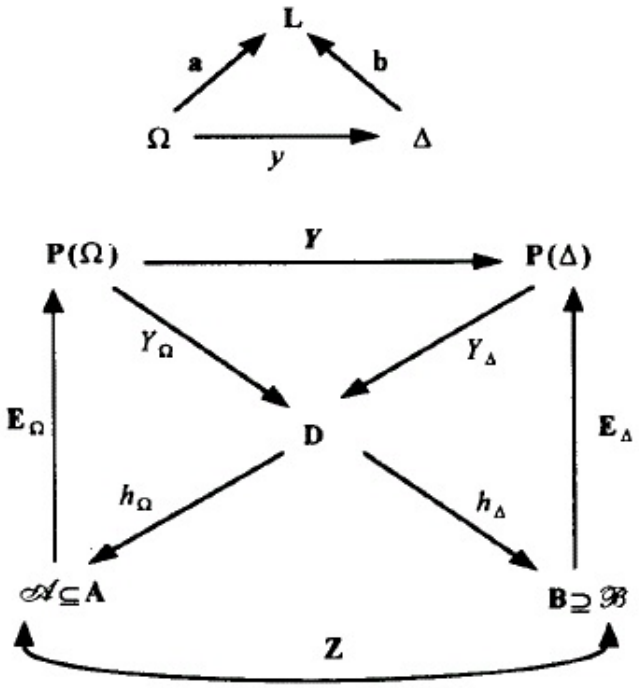


Figura 1. Diagrama de E. Diday

Donde:

Ω es el conjunto de entes elementales llamados .ºbjetos individuales"
Δ es un conjunto de posibles descripciones de los elementos de Ω
$L = \{verdadero, falso\}$ mas general $L = [0, 1]$
$y : \Omega \longrightarrow \Delta$ la cual asocia a cualquier $w \in \Omega$ su descripción $\delta = y(w)$
$a : \Omega \longrightarrow L$ tal que $a(w) = verdadero$ si y solo si, $y(w) = \delta \in d$
$b : \Delta \longrightarrow L$ tal que $b(\delta) = verdadero$ si y solo si $\delta \in d$
D es un conjunto de descripción, de subconjuntos de Ω $(D \subseteq \Delta)$
$Y_\Omega : P(\Omega) \longrightarrow D$ donde $P(\Omega)$ es el conjunto potencia de Ω el cual asocia a cualquier $\Omega' \subseteq \Omega$ su descripción $d \in D$
$Y : P(\Omega) \longrightarrow P(\Delta)$ tal que $Y(\Omega') = \Delta'$ si y solo si, $\Delta' = \{y(w) : w \in \Omega'\}$
$Y_\Delta : P(\Delta) \longrightarrow D$ la cual asocia a cualquier $\Delta' \subseteq \Delta$ una descripción $d \in D$ que satisface al menos la soguiente propiedad, $Y_\Delta(\Delta') \subseteq D$
A es un conjunto de funciones $\Omega \longrightarrow L$ donde $L = [0, 1]$
$h_\Omega : D \longrightarrow A$ tal que $h_\Omega(d) = a$
B es el conjunto de funciones $\Delta \longrightarrow L$ donde $L = [0, 1]$
$h_\Delta : D \longrightarrow B$ tal que $h_\Delta(d) = b$,
Ademas, $\breve{A} = h_\Omega(D)$ y ß$= h_\Delta(D)$
Z :ß$\longrightarrow \breve{A}$ tal que $Z(b) = a$ si y solo si, $a = b \circ y$

Una intensión de un conjunto de objetos individuales $\Omega' \subseteq \Omega$ puede ser definida por

$$d = Y_\Omega(\Omega') \ , \ a = h_\Omega(Y_\Omega(\Omega')) \text{ o bien, } b = h_\Delta(Y_\Delta(\Omega')).$$

La extensión de a en Ω es un subconjunto de Ω definido por

$$Ext(a/\Omega) = \{w \in \Omega : a(w) = verdadero\}$$

La extensión de b en Δ es un subconjunto de Δ definido por

$$Ext(b/\Delta) = \{\delta \in \Delta : b(\delta) = verdadero\}$$

La extensión de $d \in D$ en $Q = \Delta$ se denota por $Ext(d/Q)$, y por definición tomamos

$$Ext(d/\Omega) = Ext(a/\Omega) \text{ y } Ext(d/\Delta) = Ext(b/\Delta)$$

E_Δ :ß$\longrightarrow P(\Delta)$ es un mapeo, tal que $E_\Delta(b) = Ext(b/\Delta)$
$E_\Omega : \breve{A} \longrightarrow P(\Omega)$ es un mapeo, tal que $E_\Omega(a) = Ext(a/\Omega)$

De esta manera, para E. Diday, un .ºbjeto simbólico.ᵉˢ un conjunto de propiedades concernientes a un subcojunto de Ω. Cualquier elemento D, B, o A puede ser considerado como un .ºbjeto simbólico".

3.4. Elementos críticos

En la definición formal de objeto simbólico de Diday, con la notación que utiliza no hace una distinción precisa entre un objeto individual y un objeto simbólico. Además tiena algunas impresiciones:

1. Recordemos que Y_Δ esta definida por $Y_\Delta : P(\Delta) \longrightarrow D$ la cual asocia a cualquier $\Delta' \subseteq \Delta$ una descripción $d \in D$ que satisface al menos la soguiente propiedad, $Y_\Delta(\Delta') \subseteq D$. Entonces por definición $Y_\Delta(\Delta') = d \in D$, y sin embargo por la propiedad que satisface $Y_\Delta(\Delta') \subseteq D$, por lo que de esta forma $d \subseteq D$, lo cual es una inconsistencia.

2. En la práctica existen objetos individuales tales que sus variables admiten un conjunto de valores, a diferencia de los dados por Diday que solo aceptan un único valor por función y. Recordando que dicha función estaba dada por $y : \Omega \longrightarrow \Delta$ la cual asocia a cualquier $w \in \Omega$ su descripción $\delta = y(w)$. Por ejemplo en el caso de que estemos describiendo una persona y decimos que su estatura esta entre 1,70m y 1,80m, así nuestra impresión esta acotada y sabemos en definitiva que la persona tiene una única estatura.

3. La formalización del concepto de objeto simbólico no es general, ya que se va modificando el universo de objetos.

3.5. Redefinición de la formalización anterior

Teniendo en cuenta las anteriores irregularidades, se han analizado y discutido varias soluciones para mejorar lo hecho por Diday.

- Sea Ω un conjunto de objetos del universo de estudio E ($\Omega \subseteq E$)

- Δ un conjunto de descripciones admisibles de Ω, siendo F el conjunto de todas las descripciones posibles ($\Delta \subseteq F$)

- O_i el conjunto de valores admisibles de la variable. X_i con $i = 1, .., p$

- $\Delta = O_1$ x ... x O_p , $\Delta_1 = P(O_1)$ x ... x $P(O_p)$ con $P(O_i)$ el conjunto de partes de O_i , $i = 1, ..., p$. Y $D \subseteq \Delta_1$

- Sea B el conjunto de fórmulas bien formadas (fbf) en un cálculo proposicional C_p.

Consideremos el siguiente diagrama:

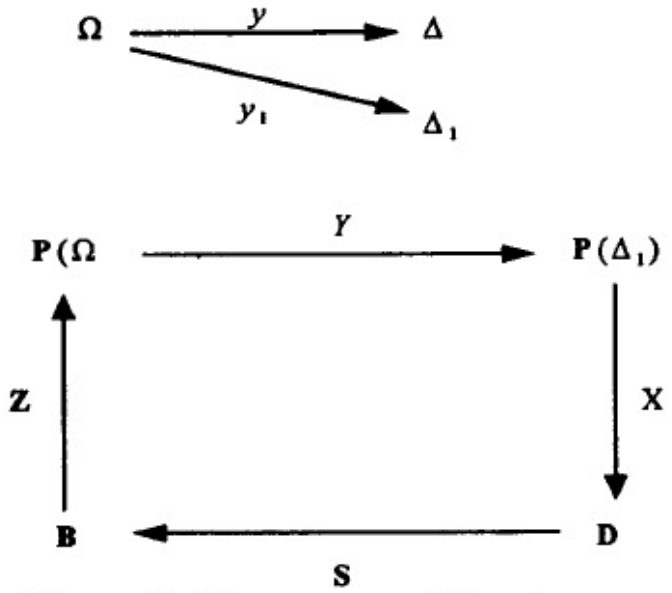


Figura 2. Diagrama modificado

Donde:

$y : \Omega \longrightarrow \Delta$ es una función de descripción de objetos de Ω en $p-uplas$ cuyas componentes tienen asignado un único valor, $y(w) = \delta$ con $w \in \Omega$, $\delta = (x_1(w), ..., x_p(w))$

$y_1 : \Omega \longrightarrow \Delta_1$ función de descripción de objetos de Ω en $p-uplos$ cuyas componentes tienen asignado un conjunto de valores. Esto permite a un rasgo x_i tomar uno o mas valores admisibles en O_i, esto se da comunmente cuando hay una impresición, es decir, que el dato se encuentra entre ciertos valores o en un intervalo de valores.

$Y : P(\Omega) \longrightarrow P(\Delta_1)$ función de descripción de subcojuntos de Ω en un conjunto de $p-uplas$ cuyas componentes tienen asignado un conjunto de valores, es decir:

$$\Omega' \subseteq \Omega \ , \ Y(\Omega') = \left\{ \left(V_1^1, ..., V_p^1\right), ..., \left(V_1^s, ..., V_p^s\right) \right\} \text{ con } 0 \leq s \leq p$$

$X : P(\Delta_1) \longrightarrow D$ la cual asigna a un conjunto de $p-uplos$ cuyas componentes son conjuntos de valores, un conjunto de una única $p-upla$ cuyas componentes son conjuntos de valores, al cual denominaremos objeto simbólico (O_s).

En símbolos:

$$X\left(\left\{\left(V_1^1, ..., V_p^1\right), ..., \left(V_1^s, ..., V_p^s\right)\right\}\right) = \{(V_1, ..., V_p)\} = O_s$$

En particular, si restringimos el dominio de X, $P(\Delta_1)$ a $Y(P(\Omega))$ que es la imagen de la función Y, tenemos que:

$$X(Y(\Omega')) = \left\{(V_1, ..., V_p) : V_i = \phi_i(Y(\Omega')_{[x_i]})\right\}$$

donde,

$$\phi_i : P(O_i)^s \longrightarrow P(O_i) \text{ con } 0 \leq s \leq p$$

$$Y_{[x_i]} : P(\Delta_1) \rightarrow P(O_i)^s$$

$$Y(\Omega')_{[x_i]} = \left(V_1^s, ..., V_p^s\right)$$

$$\phi_i\left(\left(V_1^s, ..., V_p^s\right)\right) = V_i$$

$S : D \longrightarrow B$ función de asignación de un O_s en una fórmula lógica bien formada, es decir: $S(O_s) = fbf\,(\{(V_1, ..., V_p)\})$, como por ejemplo la conjunción de proposiciones de Michalski que contituyen $l - complejos$.

$Z : P \longrightarrow P(\Omega)$ función de asiganción de una $fbf(O_s)$ a un subconjunto $\Omega'' \subseteq \Omega$. No necesariamente $\Omega'' = \Omega$, aunque si se verifica que $\Omega' \subseteq \Omega''$

En este diagrama modificado, se introducen una serie de elementos que salvan las deficiencias que habiamos señalado del diagrama de Diday, pero no se supera el inciso 3), el cual era:

"La formalización del concepto de objeto simbólico no es general, ya que al considerar objetos de tipo acumulativos o de tipo síntesis, se modifica su diagrama particularmente en lo que se refiere al universo de objetos Ω al considerar Ω^p y $\Omega_1, ..., \Omega_p$ respectivamente y por consiguiente todas las funciones que utilizan Ω"

Por esta razón, se propone una nueva formalización del concepto de objeto simbólico, la cual tiene como casos particulares la propuesta por Diday y la modificada.

3.6. Nueva formalización del concepto de objeto simbólico

Para la nueva formalización, daremos una serie de definiciones sobre la base de las cuales se construirá todo el andamiaje de la Teoría de Objetos Simbólicos, que tiene como objetivo central, la comparación entre conjuntos de datos.

Definición 1: Sean $U_1, U_2, ..., U_r$ universos, no necesariamente diferentes, de objetos reales, una variable simbólica sobre $U_1, U_2, ..., U_r$ es una función parcialmente definida sobre 2^{*U_1}x...x2^{*U_r} como sigue

$$X : 2^{*U_1}\text{x...x}2^{*U_r} \rightarrow 2^M \text{ tal que } X\,((A_1, ..., A_r)) = V$$

donde, $A_i \subseteq U_i$ para $i = 1, ..., r$; $V \subseteq M$; $X\,((A_1, ..., A_r))$ denota el valor de la variable X en el $r - uplo$ $(A_1, ..., A_r)$, $2^{*U_1}x...x2^{*U_r}$ es el producto cartesiano de los conjuntos potencia de los respectivos U_i con $i = 1, ..., r$ excluyendo al vacío. A M lo denominamos conjunto generador de los valores admisibles de la variable X y a 2^M el conjunto potencia de M, lo denominaremos conjunto de valores admisibles de la variable X. El valor $X\,((A_1, ..., A_r)) = \phi$ denota la ausencia de información en cuanto al valor que toma el $r - uplo$ $(A_1, ..., A_r)$ en la variable X.

Nota: el conjunto Potencia de un conjunto A, también llamado partes de A $(\mathcal{P}(A))$, está formado por todos los conjuntos que son subconjuntos de A. Se denota con 2^A. Si a este conjunto potencia le extraemos el conjunto vacío, lo denotamos2^{*A}.

Veamos un ejemplo: Sea $U = \{encuestado\_1, encuestado\_2\}$, $M = \mathbb{N}$, $X : 2^{*U} \rightarrow 2^M$ la variable simbólica $(r = 1)$ que denota la edad de un elemento de 2^{*U}. Así:

$X\,(\{encuestado\_1\}) = \{17, ..., 22\}$ esto es, la edad del encuestado_1 varía entre 17 y 22 años

$X\,(\{encuestado\_2\}) = \{25\}$ esto implica que, el encuestado_2 tiene 25 años

$X\,(\{Lucas, Pedro\})$= $\{17, ..., 26\}$ esto es, la edad del conjunto $\{encuestado\_1, encuestado\_2\}$ se encuentra entre 17 y 26.

Definición 2: Una variable simbólica se denomina relacional heterogénea si y solo si, $U_i \neq U_j$ *para* $i \neq j$ *con* $i, j = 1, ..., r$ *y* $r > 1$

Ejemplo: Sea X la variable que indica el porcentaje de encuestados provenientes de cierta provincia, considerando,

$U_1 = \{femenino, masculino\}$ y las provincias $U_2 = \{San\ Juan, Córdoba, Buenos\ Aires\}$ y $M = [0, 100]$.

$X(\{femenino\}, \{Córdoba\}) = \{3\}$

$X(\{femenino, masculino\}, \{Buenos\ Aires\}) = \{11\}$

$X(\{masculino\}, \{San\ Juan\}) = \{16\}$

$X(\{femenino, masculino\}, \{San\ Juan\}) = \{70\}$

Definición 3: Una variable simbólica se denomina relacional homogénea si y solo si, $U_i = U_j$ *para* $i \neq j$ *con* $i, j = 1, ..., r$ *y* $r > 1$

Ejemplo:

Sea X la variable "misma_profesión"$U_1 = \{e\_10, e\_26, e\_69\}$, $U_2 = \{e\_10, e\_26, e\_69\}$ y $M = \{si, no\}$.

$X(\{e\_10\}, \{e\_26\}) = \{si\}$

$X(\{e\_69\}, \{e\_10, e\_26\}) = \{no\}$

En este caso, se trata de una variable simbólica relacional homogénea binaria, pues $r = 2$

Definición 4: Una variable simbólica se denomina conjuntual si y solo si, $r = 1$

Por ejemplo, sea X la variable "profesión", donde

$U = \{e\_10, e\_26, e\_69\}$, $M = \{Prof.Letras, Prof.Geografía\}$ entonces:

$X(\{e\_10\}) = \{Prof.Letras\}$

$X(\{e\_26, e\_69\}) = \{Prof.Letras, Prof.Geografía\}$

$X(\{e\_69\}) = \{Prof.Geografía\}$

Las variables relacionales, según el cardinal de los conjuntos A_i pueden ser:

a) unitarias $|A_i| = 1$, $i = 1, ..., s$ siendo s el orden de la relación

b) parcialmente unitarias si $\exists j_1, ..., j_p$ con $|A_t| > 1$ para $t = j_1, ..., j_p$, $\exists i_1, ..., i_q$ con $|A_t| = 1$ para $t = i_1, ..., i_q$ y además $p + q = s$

De manera análoga las variables conjuntuales se denominan unitarias si $|A| = 1$

Tanto las variables relacionales como conjuntuales se denominan **univaluadas** si $|V| = 1$. Estos atributos no son excluyentes, es decir que pueden existir variables simbólicas univaluadas.

Veamos un ejemplo de una variable simbólica conjuntual unitaria univaluada

Ejemplo: Sea X la variable ."entidad donde trabaja", considerando

$U = \{e\_15, e\_75\}$ y $M = \{pública, privada\}$ entonces,

$X(\{e\_15\}) = \{pública\}$ en este caso, es unitaria pues $|A| = 1$ ya que solo considera a e_15, que es el encuestado 15 y es univaluada, ya que $|V| = 1$ pues, solo toma una sola opción y es en este caso la entidad *pública*.

Atendiendo a la naturaleza del conjunto M las variables simbólicas pueden ser reales, enteras, de intervalos, booleanas, k-valentes, difusas, linguísticas, etc. En la encuesta realizada, no se ha trabajado con variables difusas ni linguísticas.

Las variables simbólicas conjuntuales unitarias se pueden poner en correspondecia biunívoca con las variables tradicionales, esto es, variables cuantitativas (reales, enteras, de intervalos,

etc.) y cualitativas (booleanas, k-valentes, difusas, linguísticas, etc.). Por ejemplo, tenemos que $X(\{e\_15\}) = \{pública\}$ se corresponde con $X(e\_15) = pública$ siendo X : "entidad donde trabaja".

Definición 5: Una variable simbólica se denomina difusa si es una función parcialmente definida sobre $2^{*U_1} \times \cdots \times 2^{*U_r}$ *como sigue*

$$\widetilde{X} : 2^{*U_1} \times \cdots \times 2^{*U_r} \to 2^{\widetilde{M}} \text{ tal que } \widetilde{X}((A_1, ..., A_r)) = \widetilde{V}$$

siendo, $Img\ \widetilde{X} : 2^{*U_1}\text{x...x}2^{*U_r} = \phi + \sum_{i=1}^{n} X((A_1, ..., A_r))_{|\mu Img(X((A_1,...,A_r)))}$ *el conjunto de todos los subconjuntos difusos en* 2^M *por* X *, donde* $\mu Img(X((A_1, ..., A_r)))$ *describe el grado con el que la variable simbólica* X *toma el valor* V *en el* $t-uplo$ *de conjunto de objetos reales* $(A_1, ..., A_r)$ *donde* $A_i \subseteq U_i$ *para* $i = 1, ..., r$ *;* $\widetilde{V} \subseteq 2^M$*. Denotaremos con* $2^{\widetilde{M}}$ *al conjuno de todos los subconjuntos difusos de* 2^M *y la denominaremos potencia difusa de* 2^M*.*

Definición 6: Una variable simbólica se denomina ***linguística*** *si y solo si, toma como valores variables simbólicas difusas*

$$X : 2^{*U_1}\text{x...x}2^{*U_r} \to \left(2^{\widetilde{M}}\right)^{2^{*U_1}\text{x...x}2^{*U_r}} \text{ tal que } X((A_1, ..., A_r)) = \widetilde{X}((A_1, ..., A_r))$$

Definición 7: Sea dada una variable simbólica cualquiera X_i *con* $i = 1, ..., n$*. Un* **operador asociado a** X_i *es una función parcialmente definida sobre* $2^{*U_1}\text{x...x}2^{*U_r}$ *como sigue,*

$$\varphi\,|_{X_i}((A_1, ..., A_r)) = X_i((A_{i_1}, ..., A_{i_s})) = V_i$$

donde $V_i \subseteq M_i$ *,* $U_{i_1,...,}U_{i_s}$ *son los universos sobre los que está definida la variable* X_i *;* $i = 1, ..., n$ *;* $1 \leq s \leq r$

Definición 8: Sean dados $U_{1,...,}U_r$ *universos de objetos reales no necesariamente diferentes,* $r \geq 1$ *;* $X_1, ..., X_n$ *variables simbólicas,* $\varphi|_{X_i}$ *los operadores asociados a* X_i *;* $i = 1, ..., n$ *;* $A_j \subseteq U_j$ *con* $j = 1, ..., r$*. Diremos que un objeto simbólico* (O_s) *es un elemento cualquiera de la imagen del operador*

$$D_c : 2^{*U_1}\text{x...x}2^{*U_r} \to 2^{M_1}\text{x...x}2^{M_n}$$

$$(A_1, ..., A_r) \to\ D_c(A_1, ..., A_r) = (V_1, ..., V_r) = O_s$$

tal que $\varphi|_{X_i}(A_1, ..., A_r) = V_i$ *siendo* $V_{i_j} = \phi$*, para* $j = 1, ..., p$ *; tal que* $0 \leq p \leq n$*.*
Tomamos la convención,

$$O_s = (\phi, ..., \phi,\ \ V_{i_1}, ..., V_{i_j}, \phi, ...,\quad V_{i_{j+1}}, ...,\quad V_{i_p}, \phi, ..., \phi) = \left(V_{i_1}, ..., V_{i_p}\right)$$

Definición 9: Sea dado un cálculo proposicional C_p*. Sobre el mismo se considera el concepto de fórmula bien formada (fbf). Una* ***determinación intencional*** *de un* O_s *está dada por una imagen cualquiera del operador:*

$$D_I : 2^{M_1}\text{x...x}2^{M_n} \to C_p$$

$$D_c(A_1, ..., A_r)\ \ = (V_1, ..., V_n) = O_s \to D_I(O_s) = fbf\ (V_{i_p}, ..., V_{i_p})$$

Definición 10: una ***determinación extensional*** *de un objeto simbólico O_s es una imagen cualquiera del operador:*

$$D_B : Img\varphi \left|_{X_{i_r}} \text{x...x} Img\varphi \right|_{X_n} \rightarrow 2^M$$

$$D_B(O_s) = \left\{ \left(o, \mu_{D_B(O_s)}(o)\right) \middle| o \in U = \bigcup_{j=1}^{r} U_{i_j} \wedge \mu_{D_{B(O_s)}}(o) = v\left(fbf(V_{i_1}, ..., V_{i_p})\right) \right\}$$

donde $\mu_{D_B(O_s)}(o)$ describe el grado de pertenencia de o al subconjunto difuso $D_E(O_s) \subseteq 2^M$ y v es la función valor verativo definida sobre el cálculo proposicional seleccionado C_p

Corolario:

a) $SopD_E(O_s) = \{o \,|\{o\}\} \in 2^{*U}$ *objeto simbólico conjuntual unitario asociado a* $o \wedge$

$D_I(\{o\}) = fbf\left(X_1'(\{o\}), ..., X_n'(\{o\})\right) \wedge \forall i = 1, ..., n \; X_i'(\{o\}) \subseteq X_i(O_s)$

donde X_i' es la variable simbólica conjuntual asociada a X_i

b) $|SopD_E(O_s)| \geq \prod_{i=1}^{n} \left(2^{|X_i'(\{o\})|} - 1\right)$

Definición 11: De acuerdo a su descripción intencional los objetos simbólicos pueden ser booleanos, k-valentes, difusos, de creencia, probabilísticos, modales, etc. en dependencia del cálculo proposicional C_p que se emplee.

Un caso particular de objeto simbólico, es cuando $r = 1$, es decir, cuando trabajamos con objetos de un solo universo.

Definición 12: Un objeto simbólico individual está definido por

$$D_c(\{a\}) = \left(X_{i_1}(\{a\}), ..., X_{i_p}(\{a\})\right)$$

donde X_{i_j} , $j = 1, ..., p$ son variables simbólicas conjuntuales. Cuando las X_{i_j} sean variables simbólicas conjuntuales unitarias entonces $D_c(\{a\})$ será denominado ***objeto simbólico individual unitario****.*

Definición 13: Un objeto simbólico individual con impresición acotada en las variables X_{i_j} con $j = 1, ..., s \leq p$ es un objeto simbólico individual tal que $|V_{i_t}(w(a))| > 1$, $t = 1, ..., s$

Definición 14: Un evento es un objeto simbólico $(O_s) = D_c(A) = (V_1, ..., V_p)$ tal que

$$\exists i = 1, ..., p \; , \; \forall i \neq \phi.$$

Los eventos serán denotados $[X_i(O_s(A)) \in V_i]$, con $A \subseteq U$

De acuerdo a la descripción extensional, los objetos simbólicos pueden ser:

- duros, si $\mu_{D_B(O_s)}(o) \in \{0, 1\}$
- difusos, si $\mu_{D_B(O_s)}(o) \in [0, 1]$
- L-difuso si $\mu_{D_B(O_s)}(o) \in L$ siendo L un conjunto totalmente ordenado

4. Una Introducción al Análisis de Datos Simbólico y al Software Sodas

Las descripciones de los datos de las unidades se llaman simbólicas cuando son mas complejas que las estandar, debido a que contienen una variación interna y son estructuradas. Los objetos simbólicos constituyen una salida explicativa para el analisis de datos, además se pueden utilizar con el fin de definir consultas de una Base de Datos Relacional y propagar conceptos entre bases de datos. Definimos .ªnálisis Simbólico de Datos"(ADS) como la extensión del análisis de datos estándar para tablas de datos simbólicos como entrada, con el fin de encontrar los objetos simbólicos como salida.

Cualquier ADS se basa en cuatro espacios: el espacio de los individuos, el espacio de los conceptos, el espacio de modelado de descripciones de individuos o clases de individuos, el espacio de objetos simbólicos que modelan conceptos.

En base a estos cuatro espacios, aparecen nuevos problemas tales como la calidad, robustez y fiabilidad de la aproximación de un concepto por un objeto simbólico, la descripción simbólica de una clase y el concenso entre descripciones simbólicas. En esta sección damos una visión general sobre el desarrollo en ADS. Se presentan algunas de las herramientas y métodos de ADS,el prototipo de software SODAS (expedido a partir del trabajo de 17 equipos de nueve países que participan en un proyecto europeo de EUROSTAT).

4.1. Introducción

Cuando grandes conjuntos de datos se separan en otros de menor tamaño mas manejabales, se necesitan tablas de datos mas complejas denominadas "tablas de datos simbólicos"debido a que, una celda de tales tablas, no contiene necesariamente como de costumbre, un único valor cuantitativo o categórico.

En una tabla de datos simbólica, una celda puede contener una distribución, intervalos, o varios valores vinculados por una taxonomía y reglas lógicas. La necesidad de extender los métodos de análisis de datos estándar (exploratorio, cluster, análisis factorial, discriminación, etc) a las tablas de datos simbólicas es para obtener información más precisa y resumir grandes bases de datos.

4.1.1. La entrada a un análisis de datos simbólicos: una “Tabla de Objetos Simbólicos”

La "Tabla de datos simbólicosçonstituye el elemento principal de un análisis de datos simbólico. Las columnas de la tabla de datos de entrada son "variables simbólicas"que se utilizan para describir un conjunto de unidades llamadas "individuos ". Las filas se denominan "descripciones simbólicas"de esos individuos, ya que no son tan habituales, sólo enlaces de los valores cuantitativos o categóricos individuales.

Cada celda de esta "Tabla de datos simbólicaçontiene datos de diferentes tipos:

a) Valor individual cualitativo: por ejemplo, si "$edad$" es una variable y w un individuo: $edad(w) = 27$

b) Valor individual categórico: por ejemplo, $ciudad(w) = San\ Juan$

c) De varios valores (multievaluado): por ejemplo, altura(w) = {3,5 ; 2,1 ; 5}, significa que la altura de w puede ser de 3,5 o 2,1 o 5. (a) y (b) son casos especiales de (c).

d) Intervalo: por ejemplo, para la $edad(w) = [24, 32]$, lo que significa que la edad de w varía en el intervalo $[24, 32]$.

e) De varios valores con pesos: por ejemplo, un histograma o una función de pertenencia (notar que (a) y (b) son casos especiales de (e) cuando los pesos son iguales a 1 o 0).

A su vez, las variables pueden ser:

f) Taxonómica: por ejemplo, *el color* se considera "luz"si es amarillo, blanco o rosa.

g) Jerárquicamente dependientes: por ejemplo, podemos describir el tipo de computadora de una empresa sólo si ella tiene una computadora, por lo tanto, la variable **¿la empresa tiene computadoras?** y la variable **tipo de computadora** están vinculados.

h) Con dependencias lógicas: por ejemplo, "si la edad (w) es menos de 2 meses entonces altura (w) es de menos de 10"

Los datos simbólicos también pueden aparecer después de realizar un análisis de agrupación (cluster) usando las variables iniciales. También pueden ser *nativos* en el sentido que resultan del conocimiento de un experto. También se obtienen a partir de bases de datos relacionales, con el fin de estudiar un conjunto de unidades cuya descripción surge de varias relaciones como se mostrará en ejemplos posteriores.

4.1.2. Salida de Análisis de Datos Simbólico

La mayoría de los algoritmos en el ADS dan como salida la descripción simbólica d de una clase de individuos, usando un proceso de "generalización". Se comienza con su descripción, un modelo de objetos simbólicos que subyace el concepto y proporciona una manera de encontrar, al menos, los individuos de la clase.

Ejemplo: Las edades de dos individuos, $w1$, $w2$ que satisfacen un concepto dado (por ejemplo, que viven en la misma ciudad), son $edad(w1) = 30$, $edad(w2) = 35$, la descripción de la clase $C = \{w1, w2\}$ obtenidos por un proceso de generalización puede ser $[30, 35]$. El alcance de esta descripción contiene al menos $w1$ y $w2$, pero podría contener otros individuos.

En este caso simple, el objeto simbólico "s"se define por una terna: s = (a,R, d), donde $d = [30, 35]$, R = $\in$ y "a" es el mapeo: $\Omega \rightarrow \{verdadero, falso\}$ tal que $a(w) = verdadero$, de la "edad(w) R d" denotado con [edad (w) R d]. Un individuo w, esta en la extensión de s, si $a(w) = verdadero$.

Más formalmente, sea Ω un conjunto de individuos, D un conjunto que contiene descripciones de los individuos, d_w, o de una clase de individuos d_C, "y" una asignación definida de Ω en D, que asocia a cada $w \in \Omega$ una descripción $d_w \in D$ a partir de una tabla de datos simbólica.

Denotamos con R, una relación definida en D. Si $(x, y) \in \Omega$ decimos que x e y están conectados por R y se denota por xRy. Más en general podemos decir que xRy toma valor en un conjunto L. El mismo, puede ser $L = \{verdadero, falso\}$, en este caso $[d'Rd] = verdadero$ significa que hay una conexión entre el d y d'. También $L = [0, 1]$, si d se conecta en un cierto grado con d'. En estos casos, $[d'Rd]$ puede interpretarse como el "valor de verdad" de xRy o "el grado en que d' está en relación R con d ". Por ejemplo, $R \in \{=, \approx, \leq, \subseteq\}$ o R es una implicación, etc. R también puede ser una combinación lógica de tales operadores.

4.2. Tipos de objetos simbólicos

Un objeto simbólico se define a partir de una descripción "d", una relación "R"para comparar d, con la descripción d_w de un individuo y "a" una "función de pertenencia".

Más formalmente: "un objeto simbólico es una terna $s = (a, R, d)$ donde R es una relación entre descripciones, d es una descripción y a es una asignación definida de Ω en L que depende de R y d".

El análisis simbólico de datos, se interesa usualmente, en clases de objetos simbólicos en los que la relación R es fija, "d" "varía en un conjunto finito de descripciones y "a" es tal que, $a(w) = [y(w)Rd]$, que es por definición el resultado de comparar la descripción del individuo w con d. Aunque, de manera más general, pueden considerarse otros casos. Si, por ejemplo, la asignación de "a" es del tipo siguiente: $a(w) = [h_e\,(y(w))\,h_j(R)h_i(d)]$ donde las asignaciones h_e, h_jy h_i son "filtros" que se discutirán.

Luego, hay dos **tipos** de objetos simbólicos:

1. "**Objetos simbólicos booleanos**"si $[y(w)Rd] \in L = \{verdadero, falso\}$. En este caso, si $y(w) = (y_1, ..., y_p)$, los y_i son de tipo (a) a (d), definido en la sección anterior.
2. "**Objetos simbólicos modales**"si $[y(w)Rd] \in L = [0, 1]$.

4.2.1. Sintaxis de Objetos Simbólicos en el caso de "afirmaciones"

Si la tabla de datos inicial contiene p variables denotamos $y(w) = (y_1(w), ..., y_p(w))$, $D = (D_1, ..., D_p)$, $d \in D$ tal que $d = (d_1, ..., d_p)$ y $R' = (R_1, ..., R_p)$ donde R_i es una relación definida en D_i. Llamamos «**afirmación**» a un caso especial de un objeto simbólico definido por $s = (a, R, d)$ donde R se define por $[d'Rd] = {}_{\wedge i=1,...,p}[d'_i R_i d_i]$ donde $\wedge$ es la conjunción lógica y "a" está definido por: $a(w) = [y(w)Rd]$ en el caso de Boole.

Notemos que, la expresión $a(w) = \wedge_{i=1,...,p}[y_i(w)R_i d_i]$ permite definir el objeto simbólico $s = (a, R, d)$. Por lo tanto, podemos decir que esta expresión define un objeto simbólico llamado "afirmación".

Por ejemplo, una afirmación booleana es:
$a(w) = [edad(w) \subseteq \{12, 20, 28\}] \wedge [CSP(w) \subseteq \{empleado, trabajador\}]$.
Si el individuo u se describe en una tabla de datos simbólicos originales por
$edad(u) = \{12, 20\}$ y $CSP(u) = \{empleado\}$ entonces:
$a(u) = [\{12, 20\} \subseteq \{12, 20, 28\}] \wedge [\{empleado\} \subseteq \{empleado, trabajador\}] = verdadero.$
En el caso modal, las variables son multi-valuadas y ponderadas, un ejemplo es
$a(u) = [y(u)Rd]$ con $[d'Rd] = f(\{[y_i(w)R_i d_i]\}_{i=1,...,p})$ donde, por ejemplo
$f(\{[y_i(w)R_i d_i]\}_{i=1,...,p}) = \prod_{i=1,2} [d'_i R_i d_i]$.

Por analogía con el caso Booleano denotamos $[d'Rd] = \wedge^*_{i=1,2} p_i[d'_i R_i d_i]$ donde el significado de "$\wedge^*$"se da por la definición de la asignación "f".

Por ejemplo, una afirmación modal $I = (a, R, d)$ está completamente definida por la igualdad:
$a(w) = [edad(w)R_1\{(0{,}2)12, (0{,}8)[20, 28]\}]\wedge^*[CSP(w)R_2\{(0{,}4)empleados, (0{,}6)trabajador\}]$.

Extensión de un objeto simbólico s.
En el caso Booleano, la extensión de un objeto simbólico se denota por $Ext(I)$ y se define por la extensión de a, que es:

$$Extensión(a) = \{w \in \Omega / a(w) = verdadero\}.$$

En el caso modal, dado un umbral α, se define por:

$$Ext_\alpha(s) = Extensión_\alpha(a) = \{w \in \Omega / a(w) \geq \alpha\}.$$

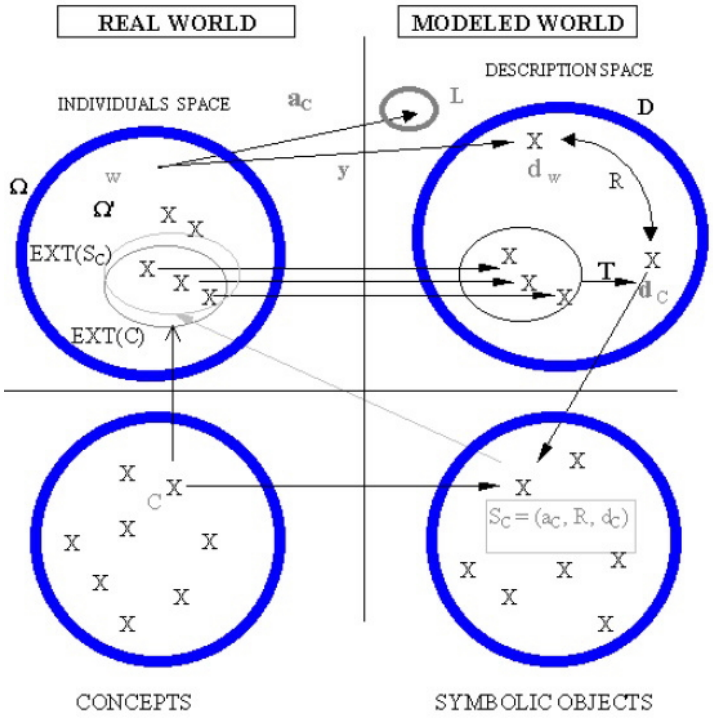


Figura: Modelado por un objeto simbólico de un concepto que se conoce por su extensión.

4.2.2. Estructuras subyacentes de Objetos Simbólicos: Latice conceptual generalizado

Bajo algunas hipótesis sobre la elección de R y T (por ejemplo, $T \equiv Max$ si $R \equiv \leq$ y $T \equiv Min$ si $R \equiv \geq$) se puede demostrar que la estructura subyacente de un conjunto de objetos simbólicos es un latice (Galois), donde los vértices son conjuntos cerrados definidos por «objetos simbólicos completos». Más precisamente, la correspondencia asociada de Galois se define por dos asignaciones F y G:

- F: donde $P(\Omega)$ (el conjunto de partez de Ω) en S (el conjunto de objetos simbólicos) tal que $F(C) = s$ donde $s = (a, R, d)$ se define por $d = T_{c \in C} y(C)$ y $a(w) = [y(w) R T_{c \in C} y(C)]$ para una determinada relación R. Por ejemplo, si $T_{c \in C} y(C) = \bigcup_{c \in C} y(C)$, $R \equiv «\subseteq»$, $y(u) = \{rosa, azul\}$, $C = \{c, c'\}$, $y(C) = \{rosa, rojo\}$, $y(c') = \{azul, rojo\}$, entonces $a(u) = [y(w) R T_{c \in C} y(c)] = [\{rosa, azul\} \subseteq \{rosa, rojo\} \cup \{azul, rojo\}\}) = \{rosa, rojo, azul\}]$ $= verdadero$ y $u \in Ext(s)$.
- G : de S en $P(\Omega)$ tal que: $G(s) = Ext(s)$.

Un «objeto simbólico completo» s es tal que $F(G(s)) = s$. Tales objetos pueden seleccionarse de un latice de Galois, sino también, a partir de una partición, una agrupación jerárquica o piramidal, a partir de los individuos más influyentes en un eje factorial, a partir de un decisión árbol, etc.

Con el fin de ver que cantidad de un objetos simbólicos son característicos de una clase A, puede usarse una distribución hipergeométrica. Sea N el tamaño de Ω', n el tamaño de A, $p = Ext(s/\Omega')/N$ de la proporción de los individuos que pertenecen a la extensión de s, una variable aleatoria X cuyo valor es la proporción de individuos de A que pertenece a la extensión de s, en una muestra de tamaño n. Enronces X sigue una distribución hipergeométrica con parámetros N, N_p y n. Siendo N_p el número de individuos de Ω' que pertenece a la extensión de s. Si el operador T produce k objetos simbólicos en A con tamaño $x_1, ..., x_k$ entonces mientras $Y = \sum_{i=1}^{k} Pr(X = x_i)/k$ es pequeña, el número de objetos simbólicos es mas chico.

4.2.3. Modelando individuos, clases de individuos y conceptos.

En la figura 1 el “conjunto de individuos” y el “conjunto de conceptos” se consideran en el “Mundo real”, el “mundo modelado” es el “conjunto de descripciones” que modelan individuos (o clases de individuos) y el “conjunto de objetos simbólicos” que modela conceptos. Comenzamos con un “concepto” C cuya extensión denotada por $Ext(C/\Omega')$ se conoce en una muestra Ω' de individuos. Por ejemplo, si el concepto es "las compañías de seguros", 30 compañías de seguros en una muestra Ω' de 1000 empresas. Cada individuo w de la extensión de C en Ω' se describe mediante el uso de la asignación de Y tal que $Y(w)$ describe el individuo w. Se generaliza el conjunto de descripciones de los individuos de $Ext(C/\Omega')$ con el operador T con el fin de producir la descripción d_C (que puede ser un conjunto producto cartesiano de intervalos y (o) las distribuciones).

i) La relación de comparación R se elige en relación con la elección T. Por ejemplo, si $T = \cup$ entonces $R =$ ” $\subseteq$ ”, que $T = \cap$, entonces $R =$ ” $\supseteq$ ”.

ii) La función de pertenencia se define entonces por $a_C(w) = [y(w)R_C d_C]$ y luego el objeto simbólico modelado por el concepto C es la terna $s = (a_C, R, d_C)$.

Cuando no tenemos conceptos como entrada, los conseguimos de la siguiente manera:

i) Un agrupamiento de Ω' usando la descripción de los individuos produce un conjunto de clases.

ii) Para cada clase de interés, denotada con A, asociamos un concepto C y un objeto simbólico $s_A = (a_A, R_A, d_A)$ con $a_A = [Y(w)R_A d_A]$ donde d_A se obtiene mediante el uso de un operador T sobre el conjunto de descripciones de los individuos de A, como en el caso anterior.

iii) El concepto C se considera modelado por S_A.

4.2.4. Algunas ventajas en el uso de Objetos Simbólicos

Observemos algunas ventajas en el uso de objetos simbólicos

1. Dan un resumen de la tabla de datos simbólicos original de una manera explicativa, expresando descripciones basadas en propiedades relativas a las variables iniciales o variables significativas (por ejemplo, indicadores obtenidos por regresión o ejes factoriales).

2. Los OS pueden transformarse fácilmente en términos de una consulta a una base de datos y usarse con el fin de propagar conceptos entre bases de datos (por ejemplo, de un país a otro país).

3. Por ser independientes de la tabla de datos inicial, son capaces de identificar cualquier coincidencia del individuo descrito en cualquier tabla de datos.

4. En el uso de su parte descriptiva, permiten dar una nueva tabla de datos simbólica de nivel superior, sobre la que se puede aplicar un análisis de datos simbólicos de segundo nivel.

5. Con el fin de caracterizar un concepto, pueden unirse fácilmente a varias propiedades basadas en distintas variables que provienen de diferentes relaciones en una base de datos de diferentes muestras de una población.

6. Con el fin de aplicar el análisis exploratorio de datos a varias bases de datos, en lugar de combinarlas en una enorme base de datos, una alternativa es resumir cada base por objetos simbólicos y después aplicar el análisis de datos simbólico a el conjunto total de objetos simbólicos obtenidos.

4.3. Algunos métodos de análisis de datos simbólicos

Los métodos de análisis de datos simbólicos se caracterizan principalmente por el siguiente principio:

i) Comienzan como entrada con una tabla de datos simbólica y dan como salida un conjunto de objetos simbólicos. Estos objetos simbólicos proporcionan una explicación de los resultados en un lenguaje cercano al del usuario y, tienen todas las ventajas mencionado en 5).

ii) Usan procesos de generalización eficientes durante los algoritmos con el fin de seleccionar las mejores variables e individuos.

iii) Proporcionan descripciones gráficas teniendo en cuenta la variación interna de los objetos simbólicos.

Los siguientes métodos desarrollados por Bock y Diday (2000) y en el software SODAS:

- Análisis de componentes principales y análisis factorial discriminante, de análisis de una tabla de datos simbólica. La salida de estos métodos preserva la variación interna de los datos de entrada en el sentido que los individuos no están representados en el plano factorial por un punto como es habitual, sino por un rectángulo que permite la definición de un objeto simbólico con ejes factoriales explicativos como variables.

- Extensión de la estadística descriptiva elemental a datos simbólicos (objeto central, histogramas, dispersión, co-dispersión, etc de una tabla de datos simbólicos).

- Extracción de objetos simbólicos a partir de respuestas a consultas de una base de datos relacional.
- Partición, agrupación jerárquica o piramidal de un conjunto de individuos descriptos por una tabla de datos simbólica, de tal manera que cada clase, se asocie con una objeto simbólico completo.
- Disimilaridades entre objetos simbólicos Booleanos o probabilísticos.
- Extensión de árboles de decisión sobre los objetos simbólicos probabilísticos.
- Generalización por una disyunción de objetos simbólicos de una clase de individuos descriptos en forma estándar.
- Representación gráfica interactiva y ergonómica de objetos simbólicos.

4.4. Análisis de datos simbólicos en el Software SODAS

El objetivo general de SODAS es la construcción simbólica de datos a fin de resumir conjuntos enormes de datos y luego, analizarlos mediante ADS. Por ejemplo, si un conjunto de hogares se caracteriza por su región, el número de dormitorios y de sala de estar, el grupo socio-económico, se obtiene un dato.

En la siguiente tabla, las unidades que se muestran corresponden a los departamentos de la provincia de San Juan, caracterizados por el total de viviendas, hombres y mujeres que poseen.

Tabla 1.

Departamento	Total de viviendas	Total de población	Varones	Mujeres
Total	**194.188**	**680.427**	**334.494**	345.933
Albardón	6.424	23.863	11.875	11.988
Angaco	2.686	8.178	4.128	4.050
Calingasta	2.975	8.453	4.507	3.946
Capital	39.908	108.720	50.501	58.219
Caucete	9.543	38.513	18.985	19.528
Chimbas	21.006	87.739	43.515	44.224
Iglesia	2.725	9.141	5.766	3.375
Jáchal	6.425	21.812	11.072	10.740
9 de Julio	2.436	9.314	4.655	4.659
Pocito	13.894	51.480	25.471	26.009
Rawson	30.925	114.946	56.122	58.824
Rivadavia	23.378	82.985	40.057	42.928
San Martín	2.860	10.969	5.546	5.423
Santa Lucía	13.482	48.137	23.451	24.686
Sarmiento	5.705	22.176	11.434	10.742
Ullúm	1.507	4.982	2.550	2.432
Valle Fértil	2.516	7.201	3.704	3.497
25 de Mayo	4.265	17.053	8.751	8.302
Zonda	1.537	4.765	2.404	2.361

En los datos hay una gran cantidad de hogares. A fin de comparar los departamentos, se puede resumir mediante la descripción de cada uno por sus habitantes, distinguiendo entre hombres y mujeres. Para hacerlo, borramos la primera columna de esta tabla y obtenemos la tabla 2.

La primera columna de la tabla, de acuerdo al número de hogares ha sido borrada:

Departamento	Total de población	Varones	Mujeres
Total	**680.427**	**334.494**	345.933
Albardón	23.863	11.875	11.988
Angaco	8.178	4.128	4.050
Calingasta	8.453	4.507	3.946
Capital	108.720	50.501	58.219
Caucete	38.513	18.985	19.528
Chimbas	87.739	43.515	44.224
Iglesia	9.141	5.766	3.375
Jáchal	21.812	11.072	10.740
9 de Julio	9.314	4.655	4.659
Pocito	51.480	25.471	26.009
Rawson	114.946	56.122	58.824
Rivadavia	82.985	40.057	42.928
San Martín	10.969	5.546	5.423
Santa Lucía	48.137	23.451	24.686
Sarmiento	22.176	11.434	10.742
Ullúm	4.982	2.550	2.432
Valle Fértil	7.201	3.704	3.497
25 de Mayo	17.053	8.751	8.302
Zonda	4.765	2.404	2.361

Ahora podemos describir cada departamento por el histograma de las categorías de cada variable. Esto se hace en la tabla 3, que es una tabla de datos simbólicos, ya que cada celda contiene un histograma y no un número o categoria como en las tablas de datos estándar. Es fácil ver que los métodos de análisis de datos estándar no se aplican de la misma manera con este tipo de datos simbólicos. Por ejemplo, un árbol de decisión no será el mismo si las variables son categorías y cada celda de la tabla de datos asociada contiene una frecuencia y si las variables son simbólicas y cada célula contiene un histograma. En el primer caso cada rama del árbol de decisión representa un intervalo de frecuencia (por ejemplo, "la frecuencia de la categoría $[20, 30]$ años de edad es menor que $0,3$ "), mientras que en el segundo caso representa un intervalo de valores (por ejemplo, "la edad es menor que 50 años").

Tabla 3. Una tabla de datos simbólica donde las unidades son ahora departamentos.

Departamento	Población
Albardón	{varones(0,5) ; mujeres(0,5)}
Iglesia	{varones(0,63) ; mujeres(0,37)}
Santa Lucía	{varones(0,47) ; mujeres(0,53)}

Los pasos principales para un análisis de datos simbólicos en SODAS, se pueden definir como sigue:

Si hay más de una tabla de datos, colocar los datos en una base de datos relacional (ORACLE, ACCESS, por ejemplo). A continuación, definir un contexto dando: las unidades (individuos, hogares, y así sucesivamente), las clases (regiones, grupos socio-economicos, ...), las variables descriptivas de las unidades. A continuación, crear una tabla de datos simbólica donde las unidades son las clases anteriores, las descripciones de cada clase se obtienen por un histograma, o por un proceso de generalización aplicado a sus elementos. Esto se hace por un programa informático de SODAS llamado "DB2SO"(de Bases de Datos Dos Objetos Simbólico). Por último, se aplican a esta tabla de datos simbólica, métodos de análisis de datos simbólicos (histograma de cada variable simbólica, disimilaridades entre las descripciones simbólicas, clustering, análisis factorial, discriminación de una tabla de datos simbólica, visualización gráfica de descripciones simbólicas, entre otros).

5. Fundamentos del Análisis de Objetos Simbólicos

5.1. Introducción

En el Análisis de Datos, los datos analizados proceden de observaciones únicas, de determinadas variables sobre individuos únicos. Debido a la gran cantidad de datos que se recogen en la actualidad, el inmenso tamaño de las bases de datos, la necesidad de sistemas de información y el uso cada vez más extendido de Internet y multimedia hacen necesario el procesamiento y el análisis de datos más complejas que los datos clásicos: los datos simbólicos.

Pueden provenir de la agregación de individuos considerando clases o grupos de los mismos y la descripción de las propiedades de estas clases por nuevos tipos de variables y datos, las variables y datos simbólicos. Estas agregaciones de individuos pueden establecerse "a priori" o provenir del resultado de otros análisis.

Los datos simbólicos, representan las propiedades o descripciones de un elemento genérico de la clase que describen. Desde otro punto de vista, los datos simbólicos pueden establecerse, así mismo, por el conocimiento del experto sin necesidad de datos individuales y, así mismo, venir dados con incertidumbre

También, los datos simbólicos permiten representar metadatos tales como: asociaciones entre las categorías de una variable, formando taxonomías; dependencias jerárquicas entre variables estableciendo aquellas que son no aplicables según los valores de otra variable; y dependencias lógicas entre variables.

Los objetos simbólicos representan conceptos, entendidos como la intención y extensión del mismo. La intención de un concepto representa las propiedades que lo definen y que lo hacen

distinto de los demás conceptos. La extensión de un concepto se compone de los individuos que se definen por el concepto o que cumplen las propiedades que definen el concepto.

Los objetos simbólicos se describen por variables y datos simbólicos y proporcionan un mecanismo de vuelta a bases de datos o conjuntos de individuos en el sentido de conocer aquéllos que se adecuan o relacionan con las descripciones simbólicas representadas por los objetos (las intenciones), según determinadas relaciones que también forman parte de las intenciones.

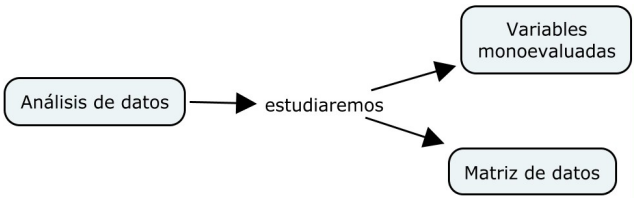
El Análisis de Datos Simbólicos es una extensión de las técnicas de Análisis de Datos aplicadas a matrices de datos simbólicos siendo el Análisis de Datos un caso particular del Análisis de Datos Simbólicos. Las definiciones y notación de variables simbólicas y objetos simbólicos han estado en constante evolución desde sus inicios (Diday (1987, 1988, 1991, 1993a, 1993b)).

En este libro de María del Carmen Bravo Llatas, se realiza por primera vez una distinción explícita entre variables y datos simbólicos de una parte, y objetos simbólicos de otra.

Esta Memoria se centra en el caso de las variables cualitativas o categóricas y por tanto, se presentan aquí las variables simbólicas relacionadas.

5.2. Análisis de datos

En esta parte, trabajaremos con:



5.2.1. Variables monoevaluadas

Sea $\Omega = \{\omega_1, ..., \omega_2\}$ un conjunto de individuos y sea $\mathcal{Y}$ un conjunto o dominio de posibles valores observados.

Definición: Se dice que X es una variable monoevaluada, definida en Ω, con dominio $\mathcal{Y}$ si es una aplicación $X : \Omega \to \mathcal{Y}$ tal que dado $\omega \in \Omega$ le asocia un único valor $X(\omega) \in \mathcal{Y}$ que es la descripción del individuo ω dada por la variable monoevaluada X.

Se dice que X es una **variable categórica monoevaluada**, si el dominio $\mathcal{Y}$ es un conjunto finito cuyos valores no permiten establecer una relación de orden entre ellos.

Sin pérdida de generalidad, se asume en lo sucesivo que la imagen de una variable X coincide con $\mathcal{Y}$.

Se puede extender la definición anterior al caso multivariante. Sean $X_1, ..., X_p$ p variables categóricas monoevaluadas definidas en Ω con dominios $\mathcal{Y}_1, ..., \mathcal{Y}_p$ respectivamente.

Sea $\mathcal{Y} := \mathcal{Y}_1 \times ... \times \mathcal{Y}_p$ el producto cartesiano de los dominios $\mathcal{Y}_j$. El **vector de variables categóricas monoevaluadas** $X = (X_1, ...X_p)$ definido en Ω es la aplicación:

$$X : \Omega \to \mathcal{Y}$$
$$\omega \to X(\omega) = (X_1(\omega), ...X_p(\omega))$$

que a un individuo $\omega \in \Omega$ asocia el vector $X(\omega) = (X_1(\omega), ...X_p(\omega))$. El vector $X(\omega) \in \mathcal{Y}$ es la descripción del individuo ω en $\mathcal{Y}$ definida por las variables X_j y el conjunto $\mathcal{Y}$ es el conjunto de las descripciones de los elementos de Ω.

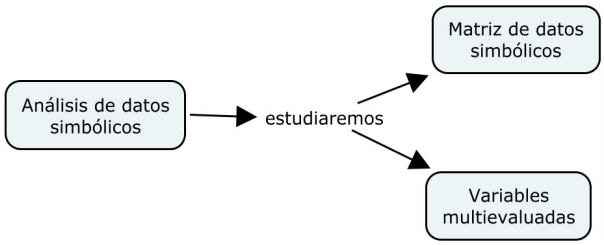
5.2.2. Matriz de datos

La matriz de datos en el Análisis de Datos es la matriz $[X]$ cuyas filas $(X(i))$, con$i = 1, .., n$ representan observaciones del vector X de variables monoevaluadas en n unidades que son los elementos de Ω.

El Análisis de datos categóricos trata del estudio de las relaciones del conjunto Ω de individuos con las variables descriptivas $X_1, ..., X_p$ de los individuos en dicho conjunto, donde cada variable X_j toma para cada individuo $\omega \in \Omega$ una única categoría de un dominio finito $\mathcal{Y}_j$. En el Análisis de Datos tradicional se estudia el par $(\Omega, \mathcal{Y}_1 \times ... \times \mathcal{Y}_p)$, con Ω conjunto de individuos e $\mathcal{Y}_1 \times ... \times \mathcal{Y}_p$ el conjunto de descripciones de elementos de Ω.

5.3. Análisis de datos Simbólicos

Trabajaremos con,



El principal objetivo del Análisis de Datos Simbólicos es extender las técnicas de Análisis de Datos a estructuras de datos más complejas. Si bien los datos simbólicos pueden describir individuos, un caso muy habitual es utilizarlos para describir información agregada de clases de individuos. En un nivel superior, pueden ser utilizados para describir información agregada de clases de clases de individuos y así, sucesivamente...

Algunas estructuras de datos más complejas a los datos clásicos, en relación con datos categóricos son: las variables y datos multievaluados y los modales probabilistas, posibilistas y difusos. Es decir:

1. Un conjunto de valores puede describir un individuo, objeto o clase de individuos. Por ejemplo, para la variable $Colorpetalo$, un dato representado como$\{amarillo, naranja\}$ puede corresponder a:

   - una familia de rosas que tienen como color de sus pétalos, o bien amarillo, o bien, naranja.
   - una rosa cuyos pétalos son de color amarillo y naranja.

   En esta representación, la variable $Colorpetalo$ es una variable simbólica multievaluada y el dato simbólico $\{amarillo, naranja\}$ es un dato multievaluado.

2. Los datos referentes a un individuo, objeto o clase de individuos pueden venir dados por una distribución de probabilidad: Por ejemplo, para la variable $Empleo$, la distribución $(sí(0{,}8), no(0{,}2))$ puede representar:

   - una subpoblación o clase de individuos de los cuales el 20 % es desempleado.
     Este dato representa la variación en la variable Empleo de una subpoblación.
   - un individuo que en la globalidad de su vida en activo, ha estado el 20 % del tiempo desempleado.
     Este dato representa una incertidumbre.

   En esta representación, la variable simbólica $Empleo$ es una variable modal probabilista y la distribución $(sí(0{,}8), no(0{,}2))$ es un dato modal probabilista.

3. Una distribución de posibilidad sobre un conjunto de valores puede describir individuos, objetos o clases de individuos.

4. Un conjunto de grados de pertenencia a varios conjuntos difusos pueden describir individuos, objetos o clases de individuos.

5.3.1. Matriz de datos simbólicos

Sea $E = \{e_1, ..., e_n\}$ un conjunto de objetos. Como casos particulares más frecuentes se tiene que E es un subconjunto de Ω o un subconjunto de las clases de Ω, es decir, $E \subseteq \Omega$ o $E \subseteq P(\Omega)$. En el segundo caso, los datos simbólicos correspondientes, describen clases de individuos de Ω.

Y sea $\mathcal{Y}$ un conjunto finito de elementos sin relación de orden. La descripción de un elemento $e \in E$ por una variable con dominio $\mathcal{Y}$ puede darse por:

- Un elemento del conjunto $\mathcal{Y}$. Este es el caso de una variable monoevaluada tratada en el Análisis de Datos clásico.
- Un subconjunto de elementos del conjunto $\mathcal{Y}$. Este es el caso de una variable simbólica multievaluada.
- Un subconjunto de elementos del conjunto $\mathcal{Y}$, donde cada uno de ellos es ponderado por un peso o modo. Este es el caso de una variable simbólica modal. Según el tipo de peso, se pueden distinguir varios tipos de variables

modales:

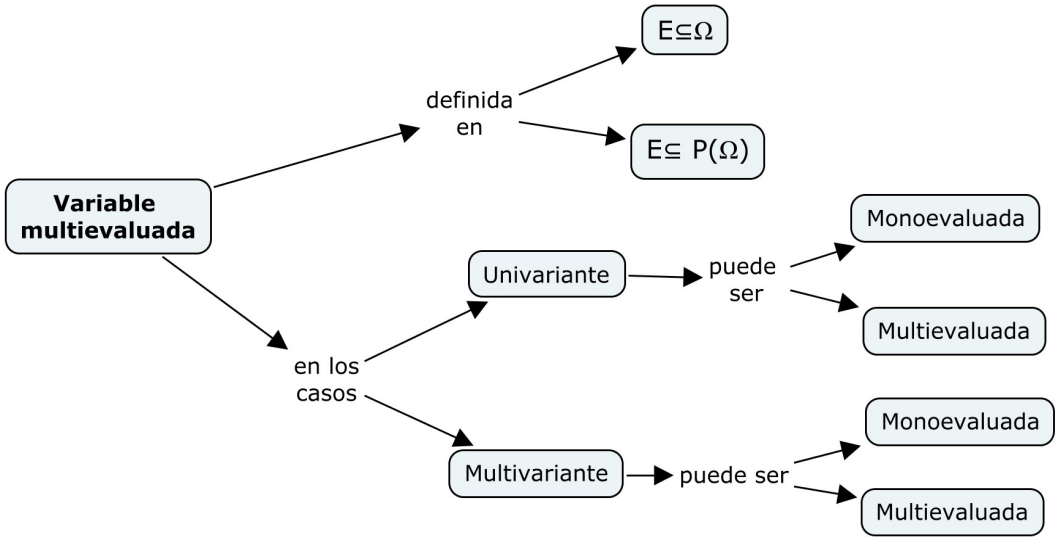
- Variables modales probabilistas, frecuentistas
- Variables modales posibilistas y difusas
- Variables modales de creencia

Las variables categóricas monoevaluadas son un caso particular de las variables simbólicas. La matriz de datos en el Análisis de Datos Simbólicos es la matriz $[X]$ cuyas filas $(X(i)), i = 1, .., n$ representan n unidades u objetos del conjunto E descritos por un vector de variables simbólicas $X = (X_1, ..., X_p)$. Es decir, las celdas de una fila se corresponden con los datos simbólicos descritos por el vector X aplicado a un elemento de E.

El Análisis de Datos Simbólicos estudia las relaciones del conjunto E con las variables simbólicas $X_1, ..., X_p$ sobre los elementos de E. En el Análisis de Datos Simbólicos se estudia el par $(E, \mathcal{D})$,con Eun conjunto de elementos y $\mathcal{D}$ un conjunto de descripciones simbólicas de elementos de E.

A continuación, se definen los distintos tipos de variables y datos simbólicos, así como los distintos tipos de conjuntos de descripciones simbólicas.

5.3.2. Variable multievaluada



Definición: Se dice que X es una variable categórica multievaluada si es una aplicación:

$$X : E \to P(\mathcal{Y})$$
$$e \to X(e)$$

- $X(e)$ es la descripción (multievaluada) de un elemento $e \in E$ en $P(\mathcal{Y})$ dada por la variable multievaluada X
- $P(\mathcal{Y})$ es el conjunto de descripciones (multievaluadas) de los elementos de E.

Se puede extender la definición anterior al caso **multivariante**. Sean $X_1, ..., X_p$, p variables categóricas multievaluadas definidas en E, con dominios respectivos $\mathcal{Y}_j$.

Sea $P(\mathcal{Y}) = P(\mathcal{Y}_1) \times ... \times P(\mathcal{Y}_p)$ el producto cartesiano de las partes de dichos dominios. El **vector de variables categóricas multievaluadas X** definido en E es:

$$X : E \to P(\mathcal{Y})$$
$$e \to X(e) = (X_1(e), ..., X_p(e))$$

- $X(e)$ es la descripción (multievaluada) de un elemento $e \in E$ en $P(\mathcal{Y})$ dada por el vector de variables multievaluadas X
- $P(\mathcal{Y})$ es el conjunto de descripciones (multievaluadas) de los elementos de E.

En el caso en que $E \subseteq P(\Omega)$, la variable X o el **vector X** se llama **descriptor** (multievaluado) de clases de individuos de Ω y $P(\mathcal{Y})$ conjunto de las descripciones (multievaluadas) de clases de Ω, o de los elementos de $P(\Omega)$

- **Descripción de clase de individuos a partir de descripciones de individuos**

Sea el conjunto $E = \{S_1, ..., S_m\} \subseteq P(\Omega)$ un subconjunto de $P(\Omega)$.Se presenta la forma más habitual de descripción de una clase por generalización de las descripciones de los individuos que la componen.

1. Caso univariante.

MONOEVALUADO

Sea $\widetilde{X}$ una variable categórica monoevaluada definida en Ω con dominio $\mathcal{Y}$

$$\begin{array}{c} \widetilde{X} : \Omega \to \mathcal{Y} \\ \omega \to \tilde{Y}(\omega) \end{array}$$

MULTIEVALUADO

A partir de la variable monoevaluada $\widetilde{X}$ de descripción de individuos, se define la variable multievaluada X en E de descripción de clase de individuos como:

$$\begin{array}{c} X : E \to P(\mathcal{Y}) \\ S_i \to X(S_i) = \left\{ \widetilde{X}(\omega) : \omega \in S_i \right\} \end{array}$$

$X(S_i)$ es la descripción de la clase S_i en $\mathcal{P}(\mathcal{Y})$ dada por la variable multievaluada X definida en $P(\Omega)$ inducida por la variable monoevaluada $\widetilde{X}$ definida en Ω.

2. Caso multivariante.

MONOEVALUADO

Sea $\widetilde{X} = \left(\widetilde{X}_1, ..., \widetilde{X}_p \right)$ un vector de variables categóricas monoevaluadas definidas en Ω con dominios respectivos $\mathcal{Y}_j$:

$$\begin{array}{c} \widetilde{X} : \Omega \to \mathcal{Y} = (\mathcal{Y}_1, ..., \mathcal{Y}_p) \\ \omega \to \widetilde{X}(\omega) = \left(\widetilde{X}_1(\omega), ..., \widetilde{X}_p(\omega) \right) \end{array}$$

MULTIEVALUADO

Sea $P(\mathcal{Y}) := P(\mathcal{Y}_1) \times ... \times P(\mathcal{Y}_p)$. A partir del vector de variables monoevaluadas $\widetilde{X}$ de descripción de individuos se define el vector de variables multievaluadas $X = (X_1, ..., X_p)$ en E de descripción de clase de individuos como:

$$\begin{array}{c} X : E \to P(\mathcal{Y}) \\ S_i \to X(S_i) = (X_1(S_1), ..., X_p(S_p)) = \left(\left\{ \widetilde{X}_1(\omega) : \omega \in S_i \right\}, ..., \left\{ \widetilde{X}_p(\omega) : \omega \in S_i \right\} \right) \end{array}$$

- $X(S_i)$ es la descripción de la clase S_i en $P(\mathcal{Y})$ dada por el vector de variables multievaluados X definido en $P(\Omega)$ inducido por el vector de variables monoevaluadas $\widetilde{X}$ definido en Ω. La descripción de la clase S_i de individuos se obtiene a partir de las categorías de Y que son observadas por el vector $\widetilde{X}$ en los individuos $\omega \in S_i$.

Veamos un ejemplo.

Sea el conjunto de individuos $\Omega = \{\omega_1, ..., \omega_7\}$ descrito por las variables categóricas monoevaluadas $\widetilde{X}_1 = \widetilde{sexo}$ e $\widetilde{X}_2 = \widetilde{estado\ civil}$ con dominios respectivos $\mathcal{Y}_1 = \{masculino, femenino\}$ e $\mathcal{Y}_2 = \{soltero, casado, viudo\}$. La matriz de datos se representa por:

$$\begin{pmatrix} id & \widetilde{sexo} & \widetilde{estado\ civil} \\ \omega_1 & femenino & casado \\ \omega_2 & femenino & soltero \\ \omega_3 & femenino & soltero \\ \omega_4 & masculino & casado \\ \omega_5 & masculino & viudo \\ \omega_6 & masculino & casado \\ \omega_7 & masculino & viudo \end{pmatrix}$$

Sea $E \subset P(\Omega)$, $E = \{S_1, S_2\} = \{\{\omega_1, \omega_2, \omega_3\}, \{\omega_4, \omega_5, \omega_6, \omega_7\}\}$. A partir de los descriptores de individuos $\widetilde{sexo}$ y $\widetilde{estado\ civil}$ se pueden definir los descriptores de clase de individuos sexo y profesión. Las variables multievaluadas sexo y profesión se definen como:

$$sexo : E \to P(\{masculino, femenino\})$$
$$S_i \to sexo(S_i) = \{\widetilde{sexo}(\omega) : \omega \in S_i\}$$

$$estado\ civil : E \to P(\{soltero, casado, viudo\})$$
$$S_i \to estado\ civil(S_i) = \left\{\widetilde{estado\ civil}(\omega) : \omega \in S_i\right\}$$

Así, por ejemplo para la clase de individuos S_1, se tiene que $sexo(S_1) = \{femenino\}$ y $profesión(S_1) = \{soltero, casado\}$ y el vector de descripciones multievaluadas para la clase S_1 es

$$(sexo, estado\ civil)(S_1) = (sexo(S_1), estado\ civil(S_1)) = (\{femenino\}, \{soltero, casado\})$$

En este caso, la matriz de datos simbólicos que representa E es:

$$\begin{pmatrix} ID & sexo & profesión \\ S_1 & \{femenino\} & \{soltero, casado\} \\ S_2 & \{masculino\} & \{casado, viudo\} \end{pmatrix}$$

Las variables categóricas monoevaluadas son un caso particular de las variables categóricas multievaluadas que describen los elementos de E por categorias únicas.

5.3.3. Variables modales probabilistas

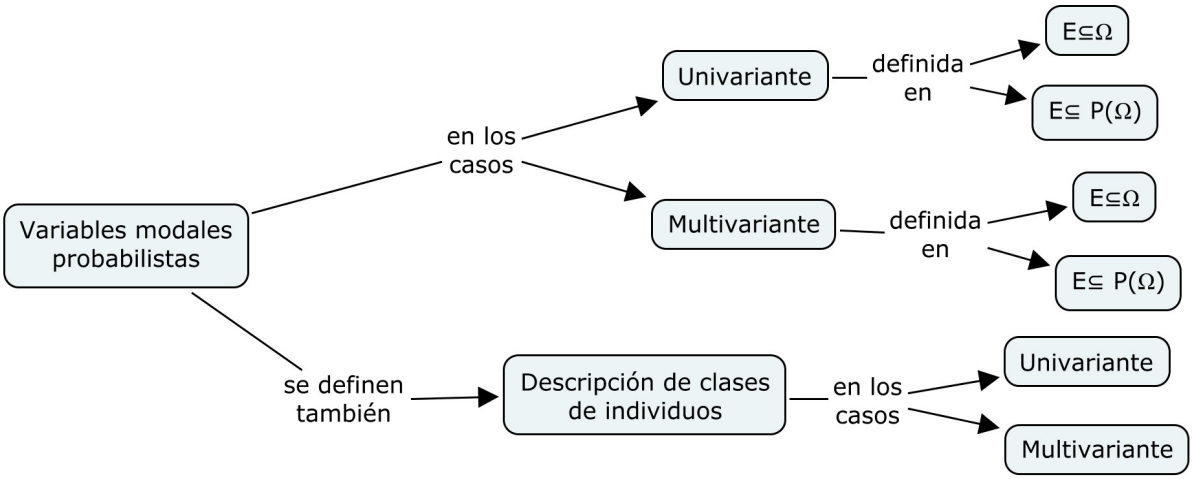
Una variable modal es aquella que describe un elemento del conjunto E no sólo por un subconjunto de elementos del conjunto Y sino también por unos modos o pesos de cada uno de ellos. Las variables modales probabilistas asocian a cada elemento de E una distribución de probabilidad o de frecuencias que puede ser:

- estimada de la observación de una variable monoevaluada sobre un individuo, en diversos instantes de tiempo

- derivada de la observación de una variable monoevaluada en una clase de individuos, estimadas las probabilidades como frecuencias relativas.
- no derivada de la observación directa, sino que es una distribución de probabilidad subjetiva derivada de un conocimiento 'a priori' o que tiene en cuenta la imprecisión o incertidumbre en la recogida de datos.

Las variables modales posibilistas y los conjuntos difusos, introducidos en la siguiente sección, también tienen en cuenta la imprecisión e incertidumbre.

También las variables modales pueden representar para cada una de las categorías una frecuencia, en lugar de una probabilidad o una posibilidad.



Sea $\mathcal{Y} = \{z_1, ..., z_x\}$ y sea $\mathcal{M}(\mathcal{Y}) = \{q : q \text{ es una distribución de probabilidad definida en } \mathcal{Y}\}$, el **conjunto de descripciones modales probabilistas** de elementos de E. Una descripción $q \in \mathcal{M}(\mathcal{Y})$ se define como:

$$\begin{gathered} q : \mathcal{Y} \to [0,1] \\ z_i \to q(z_i) \quad \text{con} \sum_{i=1,...,x} q(z_i) = 1 \end{gathered}$$

Se identifica el **dato simbólico** o descripción simbólica q con

$$q \equiv (z_1 q\,(z_1)\,, ..., z_x q\,(z_x))$$

También se identifica con esta misma expresión, pero en la que desaparecen los términos que no se encuentran en el soporte de q.

1. Caso univariante

Definición: se dice que X es una variable modal probabilista definida en E, si es una aplicación

$$\begin{gathered} X : E \to \mathcal{M}(\mathcal{Y}) \\ e \to X(e) = q_e \end{gathered}$$

tal que dado $e \in E$ le asocia $X(e) = q_e$ donde q_e es una distribución de probabilidad en el conjunto $\mathcal{Y}$ de posibles valores de observación completado por una $\sigma - álgebra$

$X(e)$ es la descripción modal probabilista (en $\mathcal{M}(\mathcal{Y})$) del elemento $e \in E$ dada por la variable modal probabilista X.

En el caso en que $E \subseteq \mathcal{P}(\Omega)$, la variable T es un **descriptor modal probabilista de clases de individuos** de Ω y $\mathcal{M}(\mathcal{Y})$ **el conjunto de las descripciones modales probabiliastas de clases** de Ω, o de los elementos de $\mathcal{P}(\Omega)$.

La definición de variable modal probabilista se puede extender a una variable modal cuyos modos asociados a las categorías de $\mathcal{Y}$ son frecuencias o pesos.

2. Caso multivariante

Extendemos ahora la definición anterior al caso multivariante. Sean $X_1, ..., X_p$, p variables modales probabilistas definidas en E, con dominios respectivos $\mathcal{Y}_j$ y conjuntos de descripciones respectivos $\mathcal{M}(\mathcal{Y}_j)$. Sea $\mathcal{M}(\mathcal{Y}) := \mathcal{M}(\mathcal{Y}_1) \times ... \times \mathcal{M}(\mathcal{Y}_p)$ el producto cartesiano de dichos conjuntos de descripciones. El vector de variables modales probabilistas $X = X_1, ..., X_p$ se define como:

$$X : E \to \mathcal{M}(\mathcal{Y})$$
$$e \to X(e) = (X_1(e), ..., X_p(e)) = (q_{e,1}, ..., q_{e,p})$$

donde $q_{e,j}$ es una distribución de probabilidad definida como:

$$q_{e,j} : \mathcal{Y}_j \to [0, 1]$$
$$y \to q_{e,j}(y) \text{ para } j \in \{1, ..., p\}$$

El conjunto $\mathcal{M}(\mathcal{Y})$ es el **conjunto de descripciones modales probabilistas** de los elementos de E. Dado $e \in E$, $X(e) \in \mathcal{M}(\mathcal{Y})$ es la **descripción modal probabilista** de e dada por el vector de variables modales X.

En el caso de que $E \subseteq \mathcal{P}(\Omega)$, el vector X se llama **descriptor modal probabilista de clases de individuos** de Ω y $\mathcal{M}(\mathcal{Y})$ **conjunto de las descripciones modales probabilistas de clases** de Ω o de los elementos de $\mathcal{P}(\Omega)$.

- **Descripción de clase de individuos a partir de descripciones de individuos**

Sea $E = \{S_1, ..., S_m\} \subseteq \mathcal{P}(\Omega)$. Se presenta la forma más habitual de descripción de la clase por generalización de las descripciones de los individuos de dicha clase.

1. Caso univariante

- Sea $\widetilde{X}$ una variable categórica monoevaluada definida en Ω con dominio $\mathcal{Y}$ definida por:

$$\widetilde{X} : \Omega \to \mathcal{Y}$$
$$\omega \to \widetilde{X}(\omega)$$

- Sea $\mathcal{M}(\mathcal{Y}) := \{q : \ q$ es una distribución de probabilidad definida en $\mathcal{Y}\}$. A partir de la variable monoevaluada $\widetilde{X}$ de descripción de individuoos se define la variable modal X de descripción de clase de individuos como:

$$X : E \to \mathcal{M}(\mathcal{Y})$$
$$S_i \to X(S_i) = q_{S_i}$$

donde la distribución de probabilidad q_{S_i} se define como:

$$q_{S_i} : \mathcal{Y} \to [0,1]$$

$$y \to q_{S_i}(y) = \frac{Card\left(\left\{\omega \in S_i : \widetilde{Y}(\omega) = y\right\}\right)}{Card(S_i)}$$

La distribución de probabilidad de una clase de individuos $S_i \in \mathcal{P}(\Omega)$ en $\mathcal{M}(\mathcal{Y})$ se obtiene a partir de las frecuencias relativas de las categorías de $\mathcal{Y}$ que son observadas por la variable $\widetilde{X}$ en los individuos $\omega \in S_i$, siguiendo la tendencia frecuentista de la probabilidad.

La distribución $X(S_i) = q_{S_i}$ es la **descripción de la clase de individuos** $S_i \in \mathcal{P}(\Omega)$ definida por una **variable modal probabilista** Y definida en $\mathcal{P}(\Omega)$ **inducida por la variable monoevaluada** $\widetilde{X}$ definida en Ω.

2. Caso multivariante

Sea $\widetilde{X} = \left(\widetilde{X}_1, ..., \widetilde{X}_p\right)$ un vector de variables categóricas monoevaluadas definidas en Ω con dominios respectivos $\mathcal{Y}_j$:

$$\widetilde{X} : \Omega \to \mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_p$$
$$\omega \to \widetilde{X}(\omega) = \left(\widetilde{X}_1(\omega), ..., \widetilde{X}_p(\omega)\right)$$

Sea $\mathcal{M}(\mathcal{Y}) = \mathcal{M}(\mathcal{Y}_1) \times ... \times \mathcal{M}(\mathcal{Y}_p)$. A partir del vector de variables monoevaluadas $\widetilde{X}$ de descripción de individuos se define el vector de variables modales probabilistas $X = (X_1, ..., X_p)$ en E de descripción de clase de individuos como:

$$X : E \to \mathcal{M}(\mathcal{Y})$$
$$S_i \to X(S_i) = (X_1(S_i), ..., X_p(S_i)) = (q_{S_i,1}, ..., q_{S_i,p})$$

donde la distribución de probabilidad $q_{S_i,j}$ para $j \in \{1, ..., p\}$ se define como:

$$q_{S_i,j} : \mathcal{Y}_j \to [0,1]$$

$$y_j \to q_{S_i,j}(y_j) = \frac{Card\left(\left\{\omega \in S_i : \widetilde{X}_j(\omega) = y_j\right\}\right)}{Card(S_i)}$$

$X(S_i)$ es la **descripción de la clase de individuos** $S_i \in \mathcal{P}(\Omega)$ el **vector de variables modales probabilistas** definido en $\mathcal{P}(\Omega)$, inducido por el vector $\widetilde{X}$ de **variables monoevaluadas** definido en Ω.

Veamos un ejemplo de variables y datos probabilistas.

Sea el conjunto de individuos $\Omega = \{\omega_1, ..., \omega_7\}$ dado en el ejemplo anterior, recordemos entonces sus datos. El conjunto de individuos descrito por las variables categóricas monoevaluadas $\widetilde{X}_1 = \widetilde{sexo}$ e $\widetilde{X}_2 = \widetilde{estado\ civil}$ con dominios respectivos $\mathcal{Y}_1 = \{masculino, femenino\}$ e $\mathcal{Y}_2 = \{soltero, casado, viudo\}$. La matriz de datos se representa por:

$$\begin{pmatrix} id & \widetilde{sexo} & \widetilde{estado\ civil} \\ \omega_1 & femenino & casado \\ \omega_2 & femenino & soltero \\ \omega_3 & femenino & soltero \\ \omega_4 & masculino & casado \\ \omega_5 & masculino & viudo \\ \omega_6 & masculino & casado \\ \omega_7 & masculino & viudo \end{pmatrix}$$

Sea $E \subset P(\Omega)$, $E = \{S_1, S_2\} = \{\{\omega_1, \omega_2, \omega_3\}, \{\omega_4, \omega_5, \omega_6, \omega_7\}\}$.

A partir de los descriptores de individuos $\widetilde{sexo}$ y $\widetilde{estado\ civil}$ se definen los descriptores de clase de individuos como variables probabilistas.

La variable modal probabilista sexo se define como:

$$sexo : E \to \mathcal{M}(\{masculino, femenino\})$$
$$S_i \to q_{sexo,i}$$

con $q_{sexo,i}$, definido por:

$$q_{sexo,i} : \{masculino, femenino\} \to [0,1]$$

$$masculino \to \frac{Card(\{\omega \in S_i : \widetilde{sexo}(\omega) = masculino\})}{Card(S_i)}$$

$$femenino \to \frac{Card(\{\omega \in S_i : \widetilde{sexo}(\omega) = femenino\})}{Card(S_i)}$$

La variable modal probabilista estado civil se define como:

$$estado\ civil : E \to \mathcal{M}(\{soltero, casado, viudo\})$$

$$S_i \to q_{estado\ civil,i}$$

con $q_{estado\ civil,i}$ definido por:

$$q_{estado\ civil,i} : \{soltero, casado, viudo\} \to [0,1]$$

$$soltero \to \frac{Card\left(\left\{\omega \in S_i : \widetilde{estado\ civil}(\omega) = soltero\right\}\right)}{Card(S_i)}$$

$$casado \to \frac{Card\left(\left\{\omega \in S_i : \widetilde{estado\ civil}(\omega) = casado\right\}\right)}{Card(S_i)}$$

$$viudo \to \frac{Card\left(\left\{\omega \in S_i : \widetilde{estado\ civil}(\omega) = viudo\right\}\right)}{Card(S_i)}$$

Así, por ejemplo para la clase de individuos S_1, se tiene que $sexo(S_1) = (femenino)$ y $estado\ civil(S_1) = (casado\frac{1}{3}, soltero\frac{2}{3})$. El vector de descripciones modales para la clase S_1 es:

$$(sexo, estado\ civil)(S_1) = (sexo(S_1), estado\ civil(S_1)) = \left((femenino), \left(casado\tfrac{1}{3}, soltero\tfrac{2}{3}\right)\right)$$

Haciendo el mismo procedimiento para S_2, resulta en este caso que, la matriz de datos simbólicos que representa E es:

$$\begin{pmatrix} ID & sexo & estado\ civil \\ S_1 & (femenino) & (casado\frac{1}{3}, soltero\frac{2}{3}) \\ S_2 & (masculino) & (casado\frac{1}{2}, viudo\frac{1}{2}) \end{pmatrix}$$

Las variables categóricas monoevaluadas son un caso particular de las variables categóricas modales probabilistas que describen los elementos de E por distribuciones de probabilidad degeneradas.

5.3.4. Variables modales posibilistas

Las distribuciones de posibilidad y los conjuntos difusos son otras formas de representación de la incertidumbre y pueden encuadrarse en el marco de las variables y datos simbólicos. Según Diday (Diday (1991,1995a)) las variables modales posibilistas asocian a cada elemento de E una distribución de posibilidad sobre el conjunto de categorías. En este caso, el peso asociado a cada categoría para un elemento de E representa el grado que tiene dicha categoría de ser real o relevante para ese elemento de E.

Desde otro punto de vista, las categorías de una variable se pueden definir cómo conjuntos difusos y los elementos de E tienen unos grados de pertenencia a estos conjuntos difusos. Se extienden las variables modales posibilistas de Diday (1991, 1995) a variables expresadas por conjuntos difusos.

En esta sección se presentan las variables modales posibilistas y las variables modales posibilistas definidas por conjuntos difusos. Para extensión de los conceptos de posibilidad y conjuntos difusos.

Distribuciones de posibilidad

Definición: una distribución de posibilidad q definida en $\mathcal{Y}$ es una función

$$\begin{gathered} q : \mathcal{Y} \to [0,1] \\ y \to q(y) \end{gathered}$$

Se dice **normalizada** si $\exists y \in \mathcal{Y}$ tal que $q(y) = 1$. Se dice que este elemento, es un elemento totalmente posible.

El **núcleo** de una distribución de posibilidad q es:

$$C(q) = \{y \in \mathcal{Y} :\ q(y) = 1\}$$

El **soporte** de una distribución de posibilidad q es:

$$S(q) = \{y \in \mathcal{Y} :\ q(y) > 0\}$$

Definición: P es una medida de posibilidad definida sobre $\mathcal{P}(\mathcal{Y})$ si es una función

$$P : \mathcal{P}(\mathcal{Y}) \to [0,1]$$

$$A \to P(A)$$

tal que verifica:

- $P(\mathcal{Y}) = 1$
- $P(\mathcal{Y}) = 0$
- $\forall A, B \in \mathcal{P}(\mathcal{Y})$, se tiene que $P(A \cup B) = \text{máx}\{P(A), P(B)\}$

Una medida de posibilidad, verifica las siguientes propiedades:

- $\text{máx}\left\{P(A), P(A^C)\right\} = 1$. Dos sucesos contrarios pueden ser totalmente posibles simultáneamente, pero al menos uno lo es.
- Si $A \subseteq B$ entonces $P(A) \leq P(B)$
- $P(A) + P(A^C) \geq 1$
- $P(A \cap B) \leq \text{mín}\{P(A), P(B)\}$

Definición: N es una medida de necesidad definida sobre $P(\mathcal{Y})$ si es una función

$$\begin{array}{c} N : P(\mathcal{Y}) \to [0,1] \\ A \to N(A) \end{array}$$

tal que

- $N(\mathcal{Y}) = 1$
- $N(\phi) = 0$
- $\forall A, B \in P(\mathcal{Y})$ se tiene que $N(A \cap B) = \text{mín}\{N(A), N(B)\}$

Además, una medida de necesidad N cumple las siguientes propiedades:

- $\text{mín}\left\{N(A), N(A^C)\right\} = 0$. Dos sucesos contrarios no pueden ser necesarios simultaneamente.

  Observación: $N(A) = 0$ significa ausencia total de certeza respecto a A pero no que sea A necesariamente falsa
- $N(A) + N(A^C) \leq 1$
- $N(A \cup B) \geq \text{máx}\{N(A), N(B)\}$

Una medida de posibilidad P lleva asociada una medida de necesidad N definida como

$$N : P(\mathcal{Y}) \to [0,1]$$

$$A \to 1 - P(A^C)$$

La necesidad de un suceso se corresponde con la imposibilidad de su complementario. Se cumplen las siguientes propiedades entre una medida de posibilidad P y una medida de necesidad N asociadas:

- $P(A) = 1 - N(A^C)$
- $N(A) \leq P(A)$
- $N(A) > 0 \Rightarrow P(A) = 1$
- $P(A) < 1 \Rightarrow N(A) = 0$, es decir, un suceso debe ser completamente posible para ser algo necesario.

A partir de una distribución de posibilidad normalizada q se pueden definir las medidas de posibilidad P y de necesidad N siguientes:

$$P : P(\mathcal{Y}) \rightarrow [0, 1]$$

$$A \rightarrow P(A) = \sup_{y \in A} q(z)$$

$$N : P(\mathcal{Y}) \rightarrow [0, 1]$$

$$A \rightarrow N(A) = \inf_{y \in A^C} \{1 - q(z)\}$$

Variables modales posibilistas

Sea $\mathcal{Y} = \{z_1, ..., z_x\}$ y sea $\mathcal{M}(\mathcal{Y}) = \{q : q$ es una distribución de posibilidad definida en $\mathcal{Y}\}$ el conjunto de descripciones modales posibilistas de elementos de E. Una descripción $q \in \mathcal{M}(\mathcal{Y})$ se define como:

$$\begin{array}{c} q : \mathcal{Y} \rightarrow [0, 1] \\ z_i \rightarrow q(z_i) \end{array}$$

Se identifica el dato simbólico o descripción simbólica q con,

$$q \equiv (z_1 q(z_1), ..., z_x q(z_x))$$

También se identifica esta expresión, con la misma donde desaparecen los términos que no se encuentran en el soporte de q.

Definición: Se dice que X es una variable modal posibilista si es una aplicación

$$\begin{array}{c} X : E \rightarrow \mathcal{M}(\mathcal{Y}) \\ e \rightarrow X(e) = q_e \end{array}$$

con q_e una distribución de posibilidad definida en el conjunto $\mathcal{Y}$.

$\mathcal{M}(\mathcal{Y})$ es el **conjunto de descripciones modales posibilistas** de los elemtos de E. Sea $e \in E$, $X(e)$ es la **descripción modal posibilista** de e en $\mathcal{M}(\mathcal{Y})$.

En el caso de que $E \subseteq \mathcal{P}(\Omega)$, la variable X se llama **descriptor modal posibilista de clases de individuos** de Ω y $\mathcal{M}(\mathcal{Y})$ **conjunto de las descripciones modales posibilistas de clases** de Ω, o de los elementos de $\mathcal{P}(\Omega)$.

Se extiende en esta monografía la definición de variable modal posibilista dada por Diday (1991, 1995) en la que imponía la condición de que las descripciones simbólicas q (en (1.35)) debían ser distribuciones de posibilidad normalizadas.

Por su parte, las variables modales posibilistas se pueden extender a variables cuyas categorías vienen definidas por conjuntos difusos. En el desarrollo de esta monografía, no se trabajará con variables difusas, ya que el proyecto en el cual se está trabajando, no aparecen este tipo de variables.

5.3.5. Conjunto de descripciones simbólicas

Como recapitulación a las variables y datos simbólicos, sea una variable definida en E como una aplicación:

$$X : E \rightarrow \mathcal{D}$$

Con $\mathcal{D}$ un conjunto de descripciones de elementos de E asociado al conjunto o dominio $\mathcal{Y}$. Este conjunto se denota por $\mathcal{D}(\mathcal{Y})$ y puede ser:

- $\mathcal{D} = \mathcal{Y}$ en el caso de que X sea una variable monoevaluada
- $\mathcal{D} = \mathcal{P}(\mathcal{Y})$ en el caso de que X sea una variable simbólica multievaluada
- $\mathcal{D} = \mathcal{M}(\mathcal{Y})$ en el caso de que X sea una variable simbólica modal con:
  - $\mathcal{M}(\mathcal{Y}) = \mathcal{M}^{\Pr ob}(\mathcal{Y}) = \{q : (\mathcal{Y}, \mathcal{P}(\mathcal{Y})) \rightarrow [0,1] \ / \ q$ es una distribución de probabilidad$\}$
  - o bien, $\mathcal{M}(\mathcal{Y}) = \mathcal{M}^{Pos}(\mathcal{Y}) = \{ \ q : \mathcal{Y} \rightarrow [0,1] \ / \ q$ es una distribución de posibilidad o son grados de pertenencia a categorías difusas$\}$

Se dice que una descripción $d \in \mathcal{D}$ está asociada al conjunto o dominio $\mathcal{Y}$. En el caso de que $E \subseteq P(\Omega)$, el conjunto $\mathcal{D}$ se llama conjunto de descripciones de clases de Ω.

La generalización a un conjunto de descripciones simbólicas dada por un vector de variables simbólicas es directa. Un vector de variables simbólicas asocia a un elemento de E un vector de descripciones de los conjuntos de descripciones asociados, es decir, un elemento del conjunto de descripciones.

$$\mathcal{D} = \mathcal{D}_1 \times ... \times \mathcal{D}_p$$

con $\mathcal{D}_j$ el conjunto de descripciones asociado al conjunto o dominio $\mathcal{Y}_j$. Este conjunto se denota por $\mathcal{D}(\mathcal{Y})$.

Por simplicidad en la notación se consideran los conjuntos de descripciones asociados a un vector de conjuntos o dominios $\mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_p$ como los conjuntos $\mathcal{Y}$, $\mathcal{P}(\mathcal{Y})$, $\mathcal{M}^{\Pr ob}(\mathcal{Y})$, $\mathcal{M}^{Pos}(\mathcal{Y})$ cuando los conjuntos de descripciones asociados a los conjuntos o dominios $\mathcal{Y}_j$ son todos del mismo tipo: monoevaluado, multievaluado, probabilista o posibilista, respectivamente. Es decir,

$$\mathcal{Y} : = \mathcal{Y}_1 \times ... \times \mathcal{Y}_p$$
$$\mathcal{P}(\mathcal{Y}) : = \mathcal{P}(\mathcal{Y}_1) \times ... \times \mathcal{P}(\mathcal{Y}_p)$$
$$\mathcal{M}^{\Pr ob}(\mathcal{Y}) : = \mathcal{M}^{\Pr ob}(\mathcal{Y}_1) \times ... \times \mathcal{M}^{\Pr ob}(\mathcal{Y}_p)$$
$$\mathcal{M}^{Pos}(\mathcal{Y}) : = \mathcal{M}^{Pos}(\mathcal{Y}_1) \times ... \times \mathcal{M}^{Pos}(\mathcal{Y}_p)$$

5.4. Objetos Simbólicos

En la sección anterior se han introducido los datos simbólicos, es decir, estructuras de datos de mayor complejidad que los datos clásicos. En esta sección, se introducen los objetos simbólicos. Un objeto simbólico se describe por variables monoevaluadas o simbólicas referidas a un conjunto E, por unos datos simbólicos y por unas relaciones de dominio que permiten la vuelta al conjunto E para obtener aquellos elementos de E cuyas descripciones dadas por dichas variables se relacionan con la descripción dada por los datos simbólicos.

Un objeto simbólico por una parte representa la intención de un concepto y por otra proporciona una herramienta para la obtención de la extensión de esa intención o concepto en un conjunto de individuos o en una base de datos.

Anteriormente, se definieron las relaciones de dominio que permiten relacionar o comparar pares de descripciones. Estas relaciones son necesarias para relacionar la descripción de un elemento de E dada por una variable con un dato simbólico. En las secciones siguientes se definirán los objetos simbólicos más habituales, llamados eventos y aserciones. Y por último, otros tipos de objetos simbólicos y antecedentes a la generalización por objetos simbólicos.

Sea el conjunto $E = \{e_1, ..., e_n\}$ de elementos descriptos por p variables monoevaluadas o simbólicas $X_1, ..., X_p$ definidas en E con dominios finitos $\mathcal{Y}_1, ..., \mathcal{Y}_p$. Se puede considerar que todas las variables son simbólicas, dado que una variable monoevaluada no es más que un caso particular de una variable simbólica sin más que considerar $E = \Omega$ y las variables X_j , monoevaluadas. Por otra parte, el conjunto de referencia E suele coincidir con el conjunto Ω, y es por esto que los elementos de E son individuos en esta sección.

5.4.1. Relaciones de dominio

Sean $(\mathcal{D}_1, ..., \mathcal{D}_p)$ y $\left(\mathcal{D}'_1, ..., \mathcal{D}'_p\right)$ dos colecciones de p conjuntos de descripciones asociados a los dominios $\mathcal{Y}_1, ..., \mathcal{Y}_p$. En lo sucesivo, estos conjuntos $\mathcal{D}_j$ y $\mathcal{D}'_j$ se denominan conjuntos de descripciones de clase, ya que es habitual que los datos sibólicos decriban clases de individuos de Ω. Sin embargo, pueden ser conjuntos de descripciones de elementos de E, simplemente.

Defincón: Sean $\mathcal{D}$ y $\mathcal{D}'$ dos conjuntos de descripciones de clase asociados a un mismo dominio, $\mathcal{D} \times \mathcal{D}'$ su producto cartesiano, una relación de dominio $\mathcal{R}$ definida en $\mathcal{D} \times \mathcal{D}'$ es una aplicación:

$$\mathcal{R} : \mathcal{D} \times \mathcal{D}' \to \mathcal{L}$$
$$(d, d') \to \mathcal{R}(d, d') := [d\mathcal{R}d']$$

que a cada par de descripciones $(d, d') \in \mathcal{D} \times \mathcal{D}'$ le asocia un valor, denotado por $[d\mathcal{R}d']$ que

mide el grado de adecuación o conexión de ambas descripciones.

$\mathcal{L}$ es el conjunto de comparación de descripciones. El valor $[d\mathcal{R}d']$ es el nivel de relación entre las descripciones d y d' , o nivel de relación de la descripción d con la descripción d'. En general, $\mathcal{L} = \{0, 1\}$ o $\mathcal{L} = [0, 1]$.

El nivel de relación entre dos descripciones $[d\mathcal{R}d']$ no es más que el resultado de la comparación entre ellas según la relación $\mathcal{R}$. n el caso de que tenga sentido definirlos, se establece que el nivel de relación de una descripción con el conjunto vacío es nulo, y con el conjunto $\mathcal{Y}$ en la unidad:

$$[d\mathcal{R}\phi] = [\phi\mathcal{R}d] = 0$$
$$[d\mathcal{R}\mathcal{Y}] = [\mathcal{Y}\mathcal{R}d] = 1$$

Aunque en principio las relaciones de dominio no se encuentran así definidas axiomáticamente, las relaciones propuestas en la literatura verifican estas dos propiedades.

Definición: $\mathcal{R}$ es una relación de dominio booleana si el conjunto de comparación de descripciones es $\mathcal{L} = \{0, 1\}$. Dadp un par de descripciones $(d, d') \in \mathcal{D} \times \mathcal{D}'$,

- cuando $[d\mathcal{R}d'] = 1$ entonces la relación entre d y d' es verdad, o d y d' se relacionan
- cuando $[d\mathcal{R}d'] = 0$ entonces la relación entre d y d' es falsa, o d y d' no se relacionan

En lo sucesivo, se identifica en una relación booleana el valor 1 con verdad o v y el valor 0 con falso o f.

En el caso de que el conjunto de comparación de descripciones sea $\mathcal{L} = [0, 1]$, se dice que $\mathcal{R}$ es una relación difusa, mientras el valor $[d\mathcal{R}d']$ se acerca más a 1, entonces la relación es mas fuerte para el par (d, d') y es más debil, mientras más se acerque a 0.

Aquí se distingue la denominación de relación probabilista cuando alguno de los conjuntos de descripciones en la relación sea un conjunto de descripciones modales probabilistas y cuando se aplique el cálculo de probabilidades o funciones entre distribuciones de probabilidad.

Las relaciones de dominio probabilistas o difusas se denotan por $\sim$.

Un tipo particular de relaciones de dominio que se establecen entre descripciones simbólicas son las de tipo **matching** que representan la comparación de dos descripciones dada una de ellas como patrón de referencia, siendo por lo general no simétricas. Una relación de este tipo puede establecerse entre una descripción genérica de clase tomada como patrón de referencia y la descripción de un individuo para establecer si el individuo puede ser considerado como un elemento de la clase descrita.

Definición: Sea una colección de relaciones de dominio $(\mathcal{R}_1, ..., \mathcal{R}_p)$, cada $\mathcal{R}_j$ definida en el producto cartesiano $\mathcal{D}_j \times \mathcal{D}'_j$.Sean $\mathcal{D} := \mathcal{D}_1 \times ... \times \mathcal{D}_p$ y $\mathcal{D}' := \mathcal{D}'_1 \times ... \times \mathcal{D}'_p$ los correspondientes productos cartesianos de los conjuntos de descripciones. La relación producto $\mathcal{R} = \mathcal{R}_1 \times ... \times \mathcal{R}_p$ definida en $\mathcal{D} \times \mathcal{D}'$ es la aplicación:

$$\mathcal{R} : \mathcal{D} \times \mathcal{D}' \to \mathcal{L}$$
$$(d, d') \to \mathcal{R}(d, d') := [d\mathcal{R}d'] := g\left(\left\{\left[d_j \mathcal{R}_j d'_j\right], j = 1, ..., p\right\}\right) = \bigwedge_{j=1,...,p} \left[d_j \mathcal{R}_j d'_j\right]$$

que a cada par de descripciones $(d, d') \in \mathcal{D} \times \mathcal{D}'$, $d = (d_1, ..., d_p)$, $d' = \left(d'_1, ..., d'_p\right)$ le asocia un valor, denotado por $[d\mathcal{R}d']$. El conjunto $\mathcal{L}'$ es el conjunto comparación de descripciones. La aplicación $g(.)$ es una aplicación simétrica.

La aplicación $g(.)$ se denota por $\wedge$ si bien este operador no es siempre el operador conjuntivo lógico estándar.

La aplicación g, llamada aplicación de combinación de niveles de relación (o de adecuación) está definida como:

$$g : \underbrace{\mathcal{L} \times ... \times \mathcal{L}}_{p \ - \ veces} \to \mathcal{L}'$$
$$(l_1, ..., l_p) \to g(l_1, ..., l_p)$$

Por lo general, $\mathcal{L}' \subseteq [0, 1]$.

Se considera que la función $g(.)$ verifica (salvo que se diga lo contrario)

$$g(1, v) = V, \forall v \in [0, 1]$$
$$g(0, v) = 0, \forall v \in [0, 1]$$

y de modo similar si la función $g(.)$ se aplica a mayor número de argumentos con alguno de ellos el valor unidad o el valor nulo:

$$g(1, l_2, ..., l_p) = g(l_2, ..., l_p), \forall l_2, ..., l_p \in [0, 1]$$
$$g(0, l_2, ..., l_p) = 0, \forall l_2, ..., l_p \in [0, 1]$$

El valor $[d\mathcal{R}d']$ es el nivel de relación entre las descripciones d y d', o nivel de relación de la descripción d con la descripción d'. El nivel de relación entre dos descripciones $[d\mathcal{R}d']$ por una relación $\mathcal{R}$, mide el grado de adecuación entre ellas como la aplicación de una función $g(.)$ a los niveles de relación $\left[d_j\mathcal{R}d'_j\right]$.

Definición: un **producto de relaciones de dominio es booleano** si el conjunto de comparación de descripciones $\mathcal{L}'$ es el conjunto $\{0, 1\}$.

- Proposición: Sea una colección de relaciones de dominio $(\mathcal{R}_1, ..., \mathcal{R}_p)$, cada $\mathcal{R}_j$ una relación de dominio booleana definida en el producto cartesiano $\mathcal{D}_j \times \mathcal{D}'_j$. La relación producto $\mathcal{R} = \mathcal{R}_1 \times ... \times \mathcal{R}_p$ tal que

$$\mathcal{R} : \mathcal{D} \times \mathcal{D}' \to \mathcal{L} = [0, 1]$$
$$(d, d') \to \mathcal{R}(d, d') := [d\mathcal{R}d'] := g\left(\left\{\left[d_j\mathcal{R}_j d'_j\right], j = 1, ..., p\right\}\right) = \bigwedge_{j=1,...,p}\left[d_j\mathcal{R}_j d'_j\right]$$

con la función g el operador conjuntivo lógico estándar, es una relación producto booleana.

5.4.2. Eventos

Los objetos simbólicos que toman en consideracion una única variable son los eventos
Definición: un objeto simbólico de tipo evento en E es una t-upla $(a, \mathcal{R}, d)$ donde:

- a es una función, denotada por $a = [X\mathcal{R}d]$, con X una variable monoevaluada o simbólica con dominio $\mathcal{Y}$ definida por $Y : E \to \mathcal{D}$ y $\mathcal{D}$ un conjunto de descripciones de elementos de E, asociado al conjunto $\mathcal{Y}$. La función a es:

$$a : E \to \mathcal{L}$$
$$e \to a(e) = [X(e)\mathcal{R}d]$$

- que asocia a cada elemento de E el nivel de relación de su descripción en $\mathcal{D}$ (dada por X) con la descripción d.
- $\mathcal{R}$ es una relación de dominio definida en $\mathcal{D} \times \{d\}$
- d es una descripción de un conjunto de descripciones asociado al conjunto $\mathcal{Y}$.

Se llaman indistintamente, la t-upla $(a, \mathcal{R}, d)$ y $a = [X\mathcal{R}d]$ **eventos**.

$a(e) = [X(e)\mathcal{R}d]$ es el **nivel de relación del individuo e con un evento a** o nivel de relación de la descripción de e en $\mathcal{D}$ (dada por X) con la descripción d.

Un objeto simbólico es en realidad, una t-upla $(a, \mathcal{R}, d)$ con $a = [X\mathcal{R}d]$, donde d es una descripción, $\mathcal{R}$ una relación entre descripciones y a es una función definida de E en $\mathcal{L}$ que mide

el nivel de relación de la descripción de un elemento $e \in E$ dada por X con la descripción d según la relación $\mathcal{R}$. Es una función que permite obtener la extensión del objeto simbólico en el conjunto de individuos E. Los individuos que pertenecen a esta extensión son aquellos cuya descripción se relaciona con d (en el caso booleano) o tiene un nivel de relación alto con d.

Definición: Un objeto simbólico de tipo evento es booleano si el conjunto de comparación de descripciónes es $\mathcal{L} = \{0, 1\}$.

En este caso, dado $e \in E$,

- si $a(e) = [X(e)\mathcal{R}d] = 1$, entonces la descripción de e en $\mathcal{D}$ (dada por X) se relaciona con la descripción d, o e se relaciona con el evento a
- si $a(e) = [X(e)\mathcal{R}d] = 0$, entonces la descripción de e en $\mathcal{D}$ (dada por X) no se relaciona con la descripción d, o e no se relaciona con el evento a

Definición: Sea $(a, \mathcal{R}, d)$ con $a = [X\mathcal{R}d]$, un evento booleano definido en E. Se llama extensión del evento booleano a en E y se denota por $Ext_E(a)$, al subconjunto de elementos de E cuya descripción en $\mathcal{D}$ (dada por X) se relaciona con el evento a:

$$Ext_E(a) = \{e \in E : a(e) = [X(e)\mathcal{R}d] = 1\}$$

Sea $(a, \mathcal{R}, d)$ con $a = [X\mathcal{R}d]$, un evento definido en E y $\delta \in [0, 1]$. Se llama extensión de nivel δ del evento a en E y se denota por $Ext_{E,\delta}(a)$, al subconjunto de elementos de E cuya descripción en $\mathcal{D}$ (dada por X) tiene un nivel de relación con el evento a igual o superior a δ :

$$Ext_{E,\delta}(a) = \{e \in E : a(e) = [X(e)\mathcal{R}d] \geq \delta\}$$

Ejemplos de eventos según distintos tipos de variable X, relación $\mathcal{R}$ y descripción d.

Sea $(a, \mathcal{R}, d)$ con $a = [X\mathcal{R}d]$, un evento. Sea X una variable monoevaluada o simbólica, con dominio finito $\mathcal{Y} = \{z_1, ..., z_x\}$ y d una descripción asociada al conjunto $\mathcal{Y}$.

A continuación se presentan algunos tipos de relación $\mathcal{R}$ que se pueden establecer dependiendo del tipo de variable X y de la descripción d.

Sea X una variable monoevaluada definida por:

$$\begin{array}{c} X : \Omega \to \mathcal{Y} \\ \omega \to z_\omega \end{array}$$

Se pueden definir los siguientes tipos de eventos:

- Si $d = z \in \mathcal{Y}$, el evento es $(a, \mathcal{R}, z)$ con $a = [X\mathcal{R}z]$. Ejemplos de relaciones booleanas $\mathcal{R}$ son la relación igualdad " = " y la relación desigualdad " $\neq$ ". Es decir, con la relación " = " por ejemplo, se tiene

$$a(\omega) = [X(\omega) = z] = [z_\omega = z] = \begin{cases} 1 \Leftrightarrow z_\omega = z \\ 0 \Leftrightarrow z_\omega \neq z \end{cases}$$

- Si $d = D \in \mathcal{P}(\mathcal{Y})$, el evento es $(a, \mathcal{R}, D)$ con $a = [X\mathcal{R}d]$. Ejemplos de relaciones booleanas $\mathcal{R}$ son la pertenencia " $\in$ ", etc... Con esta relación por ejemplo, se tiene:

$$a(\omega) = [X(\omega) \in D] = [z_\omega \in D] = \begin{cases} 1 \Leftrightarrow z_\omega \in D \\ 0 \Leftrightarrow z_\omega \notin D \end{cases}$$

- Si $d = q \equiv (z_1 q(z_1), ..., z_x q(z_x)) \in \mathcal{M}^{\Pr ob}(\mathcal{Y})$, el evento es $(a, \mathcal{R}, q)$ con $a = [X\mathcal{R}q]$.

  Ejemplos de relaciones probabilistas $\mathcal{R}$ son relaciones que se establecen entre una categoría de $\mathcal{Y}$ y una distribución de probabilidad en el conjunyo $\mathcal{Y}$. Esta relación puede ser:

$$a(\omega) = [X(\omega) \sim q] = [z_\omega \sim (z_1 q(z_1), ..., z_x q(z_x))] = q(z_\omega)$$

- que representa la probabilidad de la categoría z_ω, según la distriución de probabilidad q en $\mathcal{Y}$.

Sea $\underline{Y \text{ una variable multievaluada}}$ definida por:

$$X : E \to \mathcal{P}(\mathcal{Y})$$
$$e \to D_e$$

Se pueden definir los siguientes tipos de eventos:

- Si $d = z \in \mathcal{Y}$, el evento es $(a, \mathcal{R}, z)$ con $a = [X\mathcal{R}d]$. Ejemplo de relación booleana $\mathcal{R}$ es la relación de pertenencia contraria "$\ni$"

$$a(e) = [X(e) \ni z] = [D_e \ni z] = \begin{cases} 1 \Leftrightarrow z \in D_e \\ 0 \Leftrightarrow z \notin D_e \end{cases}$$

- Si $d = D \in \mathcal{P}(\mathcal{Y})$, el evento es $(a, \mathcal{R}, D)$ con $a = [X\mathcal{R}D]$. Ejemplos de relaciones booleanas $\mathcal{R}$ son el contenido " $\subseteq$ " y el continente " $\supseteq$ ". Otra relación que puede establecerse es si la intersección entre dos descripciones es el ϕ o no, etc...

  Por ejemplo, con la relación " $\subseteq$ " se tiene

$$a(e) = [X(e) \subseteq \mathcal{D}] = [D_e \subseteq \mathcal{D}] = \begin{cases} 1 \Leftrightarrow D_e \subseteq D \\ 0 \Leftrightarrow D_e \nsubseteq D \end{cases}$$

- Si $d = q \equiv (z_1 q(z_1), ..., z_x q(z_x)) \in \mathcal{M}^{\Pr ob}(\mathcal{Y})$, el evento es $(a, \mathcal{R}, q)$ con $a = [X\mathcal{R}q]$. Ejemplos de relaciones $\mathcal{R}$ son relaciones que se establecen entre un subconjunto de categorías de $\mathcal{Y}$ y una distribución de probabilidad sobre el conjunto $\mathcal{Y}$. Esta relación probabilista puede ser:

$$a(e) = [X(e) \sim q] = [D_e \sim (z_1 q(z_1), ..., z_x q(z_x))] = \sum_{z_i \in D_e} q(z_i)$$

- que representa la probabilidad del subconjunto $D_e \subseteq \mathcal{Y}$, dada la ley de probabilidad en $\mathcal{Y}$ expresada por $(z_1 q(z_1), ..., z_x q(z_x))$.

Sea $\underline{X \text{ una variable modal probabilista}}$ definida por:

$$X : E \to \mathcal{M}^{\Pr ob}(\mathcal{Y})$$
$$e \to q_e \equiv (z_1 q_e(z_1), ..., z_x q_e(z_x))$$

Se pueden definir los siguientes tipos de eventos relacionados con esta variable:

- Si $d = z \in \mathcal{Y}$, el evento es $(a, \mathcal{R}, z)$ con $a = [X\mathcal{R}z]$. Ejemplo de relación probabilista es:

$$a(\omega) = [X(e) \sim z] = [z \sim (z_1 q_e(z_1), ..., z_x q_e(z_x))] = q_e(z)$$

- que representa la probabilidad de la categoría z para el individuo descrito por la distribución de probabilidad q_e.

- Si $d = D \in \mathcal{P}(\mathcal{Y})$, el evento es $(a, \mathcal{R}, D)$ con $a = [X\mathcal{R}D]$. Ejemplo de relación probabilista es:

$$a(e) = [X(e) \sim D] = [D \sim (z_1 q_e(z_1), ..., z_x q_e(z_x))] = \sum_{z_i \in D} q_e(z_i)$$

- que representa la probabilidad del subconjunto D, según ley de probabilidad q_e que describe el individuo e.

- Si $d = q \equiv (z_1 q(z_1), ..., z_x q(z_x)) \in \mathcal{M}^{\Pr ob}(\mathcal{Y})$, el evento es $(a, \mathcal{R}, q)$ con $a = [X\mathcal{R}q]$. Ejemplos de relaciones probabilistas $\mathcal{R}$ son relaciones que se establecen entre dos distribuciones de probabilidad definidas sobre un mismo conjunto $\mathcal{Y}$. Se pueden definir diferentes tipos de relación entre q_e y q:

$$\mathcal{R} : \mathcal{M}^{\Pr ob}(\mathcal{Y}) \times \mathcal{M}^{\Pr ob}(\mathcal{Y}) \to \mathcal{L} \, (\subseteq [0, 1])$$

$$(q_e, q) \to [q_e \mathcal{R} q] := [q_e \sim q]$$

- Si la relación entre dos distribuciones es el producto escalar, entonces:

$$a(e) = [X(e) \sim q] =< (q_e(z_1), ..., q_e(z_x)) \, , (q(z_1), ..., q(z_x)) >$$

- representa la probabilidad de que la categoría observada en dos experiencias aleatorias de distribuciones q y q_e sea la misma.

Algunos autores (Titterington et al., 1985, Gil et al., 1993, Gower, 19 y Bock, 2000b,), presentan medidas de similaridad, disimilaridad, distancias y divergencias entre distribuciones de probabilidad. A partir de una medida de disimilaridad, divergencia o distancia se deriva fácilmente una medida de similaridad que representa la relación de dominio correspondiente.

Las relaciones de dominio que no son simétricas, son del tipo matching.

Si X una variable modal posibilista definida por:

$$X : E \to \mathcal{M}^{Pos}(\mathcal{Y})$$
$$e \to q_e \equiv (z_1 q_e(z_1), ..., z_x q_e(z_x))$$

en este caso, se presentan en la tesis que estamos estudiando, ejemplos de relaciones difusas, que no estudiaremos en este proyecto.

5.4.3. Aserciones

Una aserción es un objeto simbólico referido a varias variables. Se compone de varios eventos y está dotada de una función combinación de niveles de relación. Esta función combina los niveles de relación de cada uno de los eventos aplicados a un elemento del conjunto sobre el cual está definida la aserción.

Definición: Un objeto simbólico de tipo **aserción** definido en E es una t-upla $(a, \mathcal{R}, d)$ donde:

- a es una función, denotada por $a = [X\mathcal{R}d]$ con $X = (X_1, ..., X_p)$ un vector de variables monoevaluadas o simbólicas con dominios $\mathcal{Y}_1, ..., \mathcal{Y}_p$ definido por $X : E \to \mathcal{D}$, y $\mathcal{D} = \mathcal{D}_1 \times ... \times \mathcal{D}_p$ un conjunto de descripciones de elementos de E, $\mathcal{D}_j$ asociado al dominio $\mathcal{Y}_j$. La aserción a se define como

$$a : E \to \mathcal{L}$$
$$e \to a(e) = [X(e)\mathcal{R}d] = g\left(\left\{[X_j(e)\mathcal{R}_j d_j], j = 1, ..., p\right\}\right) = \bigwedge_{j=1,...,p} [X_j(e)\mathcal{R}_j d_j]$$

- que asocia a cada elemento de E el nivel de relación de su descripción dada por Y con la descripción $d = (d_1, ..., d_p)$, según la relación producto $\mathcal{R}$ con la función de combinación de niveles de relación g.

- $\mathcal{R} = \mathcal{R}_1 \times ... \times \mathcal{R}_p$ es un producto de relaciones de dominio definido en $\mathcal{D} \times \{d\}$ como $[d'\mathcal{R}d] = g\left(\left\{\left[d'_j \mathcal{R}_j d_j\right], j = 1, ..., p\right\}\right)$ para $d' = (d'_1, ..., d'_p) \in \mathcal{D}$.

- $d = (d_1, ..., d_p)$ es una descripción de un conjunto de descripciones asociado a los dominios $\mathcal{Y}_1, ..., \mathcal{Y}_p$

Se llama indistintamente a la t-upla $(a, \mathcal{R}, d)$ y a $a = [X\mathcal{R}d]$ aserción. Así mismo la aserción a se denota por $a = \bigwedge_{j=1,...,p} [Y_j \mathcal{R}_j d_j]$, si bien $\wedge$ no es la conjunción booleana necesariamente.

$a(e) = [X(e)\mathcal{R}d]$ es el nivel de relación del individuo e con la aserción a o nivel de relación de la descripción de e en $\mathcal{D}$ dada por X con la descripción d.

Una aserción de tipo individuo es $\bigwedge_{j=1,...,p} [X_j = z_j]$, con $X_j : \Omega \to \mathcal{Y}_j$ una variable monoevaluada, $z_j \in \mathcal{Y}_j$ y $\wedge$ el operador conjuntivo estándar.

Definición: Un objeto simbólico de tipo aserción es booleano si el conjunto de comparación de descripciones es $\mathcal{L} = \{0, 1\}$. En este caso, dado $e \in E$:

- si $a(e) = [X(e)\mathcal{R}d] = 1$ entonces la descripción de e en $\mathcal{D}$ (dada por el vector X) se relaciona con la descripción d, o e se relaciona con la aserción a

- si $a(e) = [X(e)\mathcal{R}d] = 0$ entonces la descripción de e en $\mathcal{D}$ (dada por el vector X) no se relaciona con la descripción d, o e no se relaciona con la aserción a

Definición: Sea $(a, \mathcal{R}, d)$ con $a = [X\mathcal{R}d]$, una aserción booleana definida en E. Se llama extensión de la aserción booleana a en E, $Ext_E(a)$, al subconjunto de elementos de E cuya descripción en $\mathcal{D}$ (dada por el vector X) se relaciona con la aserción a :

$$Ext_E(a) = \{e \in E : a(e) = [X(e)\mathcal{R}d] = 1\}$$

Sea $(a, \mathcal{R}, d)$ con $a = [X\mathcal{R}d]$, una aserción definida en E y $\delta \in [0,1]$. Se llama extensión de nivel δ de la aserción a en E y se denota por $Ext_{E,\delta}(a)$, al subconjunto de elementos de E cuya descripción en $\mathcal{D}$ (dada por X) tiene un nivel de relación con la aserción a igual o superior a δ :

$$Ext_{E,\delta}(a) = \{e \in E : a(e) = [X(e)\mathcal{R}d] \geq \delta\}$$

Se pueden definir las extenciones de aserciones booleanas sobre subconjuntos de E. Sea $S \in \mathcal{P}(E)$, la extensión de a en S es:

$$Ext_S(a) = \{e \in S : a(e) = [X(e)\mathcal{R}d] = 1\}$$

y extensiones de nivel de aserciones sobre subconjuntos de E. Sea $\delta \in [0,1]$, la extensión de nivel δ de a en S es:

$$Ext_{S,\delta}(a) = \{e \in S : a(e) = [X(e)\mathcal{R}d] \geq \delta\}$$

Diday (Diday, 1991) distingue entre las aserciones probabilistas, posibilistas y de creencia según las descripciones simbólicas y las relaciones de dominio que definen las aserciones.

La importancia de los objetos simbólicos es que éstos permiten la vuelta a bases de datos. Es decir, un objeto simbólico además de representar una intención, permite obtener la extensión de esta intención en una base de datos.

Definición: sean $(a, \mathcal{R}, d)$ y $(b, \mathcal{R}', d')$ con $a = [X\mathcal{R}d]$, $b = [X'\mathcal{R}'d']$ dos aserciones definidas en E. Las aserciones a y b son equivalentes, y se denota por $a \equiv b$ si:

$$a(\omega) = b(\omega), \forall \omega \in E$$

Ejemplos de aserciones según tipos de eventos y funciones de combinación de niveles de relación.

- **Ejemplo de aserción booleana**

Consideremos como base de datos, el ejemplo dado anteriormente. Recordemos sus datos:

Sea el conjunto de individuos $\Omega = \{\omega_1, ..., \omega_7\}$, descrito por las variables categóricas monoevaluadas $\widetilde{X}_1 = \widetilde{sexo}$ e $\widetilde{X}_2 = \widetilde{estado\ civil}$ con dominios respectivos $\mathcal{Y}_1 = \{masculino, femenino\}$ e $\mathcal{Y}_2 = \{casado, soltero, viudo\}$. La matriz de datos se representa por:

$$\begin{pmatrix} id & \widetilde{sexo} & \widetilde{estado\ civil} \\ \omega_1 & femenino & casado \\ \omega_2 & femenino & soltero \\ \omega_3 & femenino & soltero \\ \omega_4 & masculino & casado \\ \omega_5 & masculino & viudo \\ \omega_6 & masculino & casado \\ \omega_7 & masculino & viudo \end{pmatrix}$$

A partir de dicha matriz de datos simbólicos, se definen sobre Ω las aserciones siguientes:

$$a_1 = [\widetilde{sexo} \in \{femenino\}] \wedge \left[\widetilde{estado\ civil} \in \{casado, soltero\}\right]$$
$$a_2 = [\widetilde{sexo} \in \{masculino\}] \wedge \left[\widetilde{estado\ civil} \in \{casado, viudo\}\right]$$

con las variables $\widetilde{sexo}$ y $\widetilde{estado\ civil}$ definidas en Ω y $\wedge$ el operador lógico conjuntivo.

La aserción a_1 es de la siguiente forma:

$$a_1 : \Omega \to \{0, 1\}$$
$$\omega \to a_1(\omega) = [\widetilde{sexo}(\omega) \in \{femenino\}] \wedge \left[\widetilde{estado\ civil}(\omega) \in \{casado, soltero\}\right]$$

Se tiene que
$a_1(\omega) = 1 \Longleftrightarrow \widetilde{sexo}(\omega) = femenino$ y $\widetilde{estado\ civil}(\omega) \in \{casado, soltero\}$

La aserción a_2 se define en forma similar.

Las aserciones a_1 y a_2 representan la intención de dos subconjuntos de individuos S_1 y S_2, respectivamente. Cada una de las aserciones a_1 y a_2 son también un medio de obtención de individuos que verifican dicha intención.
Las extensiones en Ω de las aserciones a_1 y a_2 son:

$$Ext_\Omega(a_1) = \{\omega \in \Omega : a_1(\omega) = 1\} = \{\omega_1, \omega_2, \omega_3\}$$
$$Ext_\Omega(a_2) = \{\omega \in \Omega : a_2(\omega) = 1\} = \{\omega_4, \omega_5, \omega_6, \omega_7\}$$

La expresión de las aserciones a_1 y a_2 se puede simplificar sin más que considerar equivalentes $[\widetilde{sexo} \in \{mujer\}]$ con $[\widetilde{sexo} = mujer]$ y $[\widetilde{sexo} \in \{masculino\}]$ con $[\widetilde{sexo} = masculino]$ ya que respectivamente dan los mismos valores al aplicarlos sobre los individuos de Ω, por ser la variable $\widetilde{sexo}$ monoevaluada.

- **Ejemplo de aserciónes probabilistas**

De forma similar al ejemplo anterior, consideramos la misma matriz de datos simbólicos y se definen sobre Ω las aserciones siguientes:

$$a_3 = [\widetilde{sexo} \sim (femenino)] \wedge \left[\widetilde{estado\ civil} \sim \left(casado\tfrac{1}{3}, soltero\tfrac{2}{3}\right)\right]$$
$$a_4 = [\widetilde{sexo} \sim (masculino)] \wedge \left[\widetilde{estado\ civil} \sim \left(casado\tfrac{1}{2}, viudo\tfrac{1}{2}\right)\right]$$

con las variables $\widetilde{sexo}$ y $\widetilde{estado\ civil}$ definidas en Ω y $\wedge$ el operador producto. La relación de dominio $\sim$ definida en $\mathcal{Y} \times \mathcal{M}^{\mathrm{Pr}\,ob}(\mathcal{Y})$ es la definida anteriormente, dada por:

$$a(\omega) = [X(\omega) \sim q] = [z_\omega \sim (z_1 q(z_1), ..., z_x q(z_x))] = q(z_\omega)$$

siendo X una variable monoevaluada, $d = q \equiv (z_1 q(z_1), ..., z_x q(z_x)) \in \mathcal{M}^{\Pr ob}(\mathcal{Y})$ y $a = [X\mathcal{R}q]$

La aserción a_3 es de la siguiente forma:

$$a_3 : \Omega \to \{0,1\}$$
$$\omega \to a_3(\omega) = [\widetilde{sexo}(\omega) \sim (femenino)] \wedge \left[\widetilde{estado\ civil}(\omega) \sim \left(casado\tfrac{1}{3}, soltero\tfrac{2}{3}\right)\right]$$

La aserción a_4 se define de forma similar.

En este caso las extensiones en Ω de las aserciones a_3 y a_4 dependen de un umbral de adecuación ya que las relaciones de dominio $\sim$ no son booleanas. Los niveles de relación de los individuos $\omega \in \Omega$ de la matriz de individuos con a_3 y a_4 son:
$a_3(\omega_1) = \frac{1}{3},\ a_3(\omega_2) = \frac{2}{3},\ a_3(\omega_3) = \frac{2}{3},\ a_3(\omega_4) = 0,\ a_3(\omega_5) = 0,\ a_3(\omega_6) = 0,\ a_3(\omega_7) = 0$;
$a_4(\omega_1) = 0,\ a_4(\omega_2) = 0,\ a_4(\omega_3) = 0,\ a_4(\omega_4) = \frac{1}{2},\ a_4(\omega_5) = \frac{1}{2},\ a_4(\omega_6) = \frac{1}{2},\ a_4(\omega_7) = \frac{1}{2}$

Así mismo, dado un umbral $\delta \in [0,1]$ se obtienen las extensiones de nivel δ en Ω de las aserciones a_3 y a_4. Varios ejemplos son:

$$Ext_{0,33,\Omega}(a_3) = \{\omega \in \Omega : a_3(\omega) \geq 0{,}33\} = \{\omega_1, \omega_2, \omega_3\}$$
$$Ext_{0,5,\Omega}(a_3) = \{\omega \in \Omega : a_3(\omega) \geq 0{,}5\} = \{\omega_2, \omega_3\}$$
$$Ext_{0,5,\Omega}(a_4) = \{\omega \in \Omega : a_4(\omega) \geq 0{,}5\} = \{\omega_4, \omega_5, \omega_6, \omega_7\}$$
$$Ext_{0,66,\Omega}(a_4) = \{\omega \in \Omega : a_4(\omega) \geq 0{,}66\} = \phi$$

La expresión de las aserciones a_3 y a_4 se puede simplificar sin más que considerar equivalentes $[\widetilde{sexo} \sim (femenino)]$ con $[\widetilde{sexo} = femenino]$ y $[\widetilde{sexo} \sim (masculino)]$ con $[\widetilde{sexo} = masculino]$ ya que respectivamente dan los mismos valores al aplicarlos sobre los individuos de Ω, por ser la variable $\widetilde{sexo}$ monoevaluada.

A continuación se muestran otros dos ejemplos de aserción probabilista y aserción posibilista.

Sea la aserción probabilista $(a, \mathcal{R}, q)$ definida en E, con

$$a = \bigwedge_{j=1,...,p} [Y_j \sim q_j],\ q = (q_1, ..., q_p) \in \mathcal{M}^{\Pr ob}(\mathcal{Y})$$

y el vector $X = (X_1, ..., X_p)$ de variables modales probabilistas:

$$X : E \to \mathcal{M}^{\Pr ob}(\mathcal{Y})$$
$$e \to q_e = (q_{e,1}, ..., q_{e,p})$$

con $q_{e,j} \in \mathcal{M}^{\Pr ob}(\mathcal{Y}_j)$, $\mathcal{Y}_j = \left\{z_1, ..., z_{l_j}\right\},\ j \in \{1, ..., p\}$. La aserción a es:

$$a : E \to [0,1]$$
$$e \to a(e) = [X(e) \sim q] = \bigwedge_{j=1,...,p} [X_j(e) \sim q_j] = \bigwedge_{j=1,...,p} [q_{e,j} \sim q_j]$$

Sea la aserción posibilista $(a, \mathcal{R}, q)$ definida en E con

$$a = \bigwedge_{j=1,...,p} [X_j \sim q_j] \ , \ q = (q_1, ..., q_p) \in \mathcal{M}^{\Pr ob}(\mathcal{Y})$$

y sea el vector $X = (X_1, ..., X_p)$ de variables modales posibilistas

$$X : E \to \mathcal{M}^{Pos}(\mathcal{Y})$$
$$e \to q_e = (q_{e,1}, ..., q_{e,p})$$

con $q_{e,j} \in \mathcal{M}^{Pos}(\mathcal{Y}_j)$, $\mathcal{Y}_j = \left\{z_1, ..., z_{l_j}\right\}$, $j \in \{1, ..., p\}$. La aserción a es:

$$a : E \to [0, 1]$$
$$e \to a(e) = [X(e) \sim q] = \bigwedge_{j=1,...,p} [Y_j(e) \sim q_j] = \bigwedge_{j=1,...,p} [q_{e,j} \sim q_j]$$

5.4.4. Otros tipos de datos y objetos simbólicos

Existen otros tipos de variables, datos y objetos simbólicos que exceden el ámbito de la Memoria que estamos estudiando:

- Las variables y datos simbólicos que se refieren a variables con dominio en un continuo. Este es el caso de las variables de intervalo. Por ejemplo, un dato de intervalo puede estar representado por el intervalo [156, 170] para la variable simbólica de intervalo edad.
- Las variables y datos simbólicos de creencia y las correspondientes aserciones definidas con estas variables y/o descripciones.
- Las variables y objetos simbólicos asociados a una generalización de modos que expresen grados de certeza sobre las categorías de una variable (Diday, (1991, 1995b)). En este caso, el conjunto de modos es un conjunto ordenado.
- Objetos simbólicos modales exteriores
- Objetos simbólicos síntesis (Diday, 1991). Un objeto de síntesis es una conjunción de varios objetos simbólicos horda. Un objeto horda se define sobre una potencia del conjunto de individuos
- Objetos simbólicos regla (cuando se cumplen bajo ciertas condiciones)
- Objetos simbólicos resultantes de la aplicación de una aplicación de filtro a variables simbólicas y las correspondientes descripciones simbólicas.

Los datos y objetos simbólicos pueden representar información adicional acerca de los datos, es decir, **metadatos**. Los metadatos que se consideran en el Análisis de Datos Simbólicos son:

- Las variables taxonómicas representan taxonomías o estructuras jerárquicas entre categorías de una variable. Por ejemplo, la variable Alimentos de cinco categorías es verdura si es acelgas, espinacas o judías; y es legumbres, si es garbanzos y lentejas.

- Las dependencias jerárquicas entre variables representan variables que no son aplicables para determinados valores de otra variable. Por ejemplo, si la variable $fumador$ es igual a no, entonces la variable $marca\_de\_cigarrillos$ es no aplicable. La forma de representar este metadato es mediante el objeto simbólico:

$$\text{Si}\,[fumador = no] \Rightarrow [marca\_cigarrillos = No\_aplicable]$$

Este tipo de objeto simbólico recibe el nombre de $regladenoaplicabilidad$.

- Las dependencias lógicas entre variables representan valores posibles de una variable en función de los valores de otra. Por ejemplo, si la variable $animal$ es $ratón$, entonces la variable $longitud$ es menor o igual que 20 centímetros. Este metadato se puede representar por el objeto simbólico:

$$\text{Si}\,[animal = rata] \Rightarrow [longitud \leq 25]$$

5.4.5. Generalización

El proceso de generalización de un conjunto de individuos descritos por datos monoevaluados consiste en la obtención de datos y objetos simbólicos que los describan agregadamente.

Antecedentes de generalización pueden encontrarse en Michalski, 1973 inspirados en la lógica de primer orden que aplica técnicas de Inteligencia Artificial mediante la búsqueda heurística de complejos (Michalski, 1969, Michalski y Larson, 1983, Michalski et al., 1986, Clark y Nibblet, 1989), dada una clasificación inicial $\{c_1, ..., c_s\}$. Un complejo no es más que un predicado lógico definido en los predictores con los operadores conjuntivo y disyuntivo. Clark y Nibblet combinan esta búsqueda heurística con la entropía de Shannon para evaluar la calidad de los complejos y con el estadístico de la razón de verosimilitud (Kalbfleish, 1979) para evaluar su significatividad. Este estadístico es:

$$2 \sum_{i=1,...,s} n_i^k \log\left(\frac{n_i^k}{e_i^k}\right)$$

con n_i^k frecuencia de elementos que cumplen el complejo k y son de la clase c_i y e_i^k frecuencia esperada bajo la hipótesis de distribución aleatoria de los elementos que cumplen el complejo. En determinadas circunstancias, este estadístico se distribuye aproximadamente como un estadístico chi-cuadrado con $s-1$ grados de libertad.

La generalización por datos y objetos simbólicos puede aplicarse a consultas de una base de datos, es decir, a los conjuntos de individuos resultantes de las consultas, a clases obtenidas por una técnica de Análisis de Conglomerados o a cualquier otra clasificación.

Se ha presentado la descripción agregada de clases de individuos a partir de las descripciones monoevaluadas de los individuos de la clase. Proponen para cada una de las clases, realizar un proceso de generalización aplicado a cada variable para obtener descripciones simbólicas multievaluadas o modales probabilistas y objetos simbólicos de un conjunto o clase de individuos, seguido de un proceso de especificación que evite una sobregeneralización.

Este proceso de especificación, se realiza de forma univariante mediante la calidad de la descripción obtenida. La medida de calidad que proponen combina la homogeneidad de los individuos de la clase con la extensión del objeto simbólico que se obtiene en la fase de generalización.

Se propone una generalización de conglomerados de individuos con variables originales categóricas, por objetos multievaluados y modales probabilistas descritos por las probabilidades empíricas. Realizan la generalización de cada conglomerado por un proceso iterativo añadiendo sucesivamente eventos con mayor poder generalizante y discriminante frente a los demás conglomerados. Comparan los resultados de la clasificación original con las extensiones a un determinado umbral de los objetos simbólicos obtenidos.

Con esta misma idea, también se propone un proceso de marcaje, en un proceso iterativo que generaliza todas las variables originales. Este proceso comienza por eventos booleanos de descripciones monoevaluadas, combina criterios de homogeneidad de cada clase con criterios de discriminación con respecto a las demás clases y obtiene objetos simbólicos multievaluados y modales probabilistas.

El proceso de generalización puede aplicarse también a la obtención de objetos simbólicos que ayuden a la interpretación de ejes factoriales obtenidos en una técnica de Análisis de Datos a individuos de datos monoevaluados

Se dividen los ejes entre s clases según las proyecciones de los individuos. Se procede a la descripción simbólica de las tres clases por generalización o por aplicación de una técnica de Segmentación.

5.5. Operaciones sobre conjuntos de aserciones

Esta sección introduce algunas operaciones sobre conjuntos de aserciones.

5.5.1. Unión, intersección y complementariedad

Sea,

$$\mathcal{A}' = \{(a, \mathcal{R}, d) : a = [X\mathcal{R}d] \text{ con } a : \Omega \to [0,1], \ d \in \mathcal{D} \}$$

un conjunto de aserciones definidas en el conjunto Ω con $X : \Omega \to \mathcal{D}'$ una variable o vector de variables monoevaluadas o modales probabilistas, con $\mathcal{D}'$ y $\mathcal{D}$ dos conjuntos de descripciones (univariantes o multivariantes) relativos a un dominio $\mathcal{Y}$. Se asume que en $\mathcal{D}$ están definidas las operaciones de unión, intersección y complementariedad.

En particular, se considera que $\mathcal{D} = \mathcal{Y}$ o $\mathcal{D} = \mathcal{P}(\mathcal{Y})$ y que la unión, intersección y complementariedad definidas son las conjuntistas en $\mathcal{D}_i$ (para $\mathcal{D} = \mathcal{D}_1 \times ... \times \mathcal{D}_p$) o en $\mathcal{D}$ univariante.

Definiremos la unión, intersección y complementariedad de aserciones. Por simplificación en la notación, se denotan la unión, intersección y complementariedad de la misma manera que en el conjunto $\mathcal{D}$ y en el conjunto $\mathcal{A}'$.

Definición: Sean $a_1 = (X\mathcal{R}d) \in \mathcal{A}'$, $a_2 = (X\mathcal{R}d') \in \mathcal{A}'$ dos aserciones de $\mathcal{A}'$. Se define la unión de las aserciones a_1 y a_2 como la aserción

$$a_1 \cup a_2 = (X\mathcal{R}\,(d \cup d'))$$

La unión de a_1 y a_2 es la **aserción total** si $d \cup d' = \mathcal{Y}$. La aserción total se denota por $t^{\mathcal{A}'}$ y verifica que:

- $\forall \omega \in \Omega, t^{\mathcal{A}'}(\omega) = 1$

- $\forall S \subseteq \Omega,\ Ext_S(t^{\mathcal{A}'}) = S$

Además, se cumple la siguiente:

Proposición: Si uno de los eventos que componen una aserción es el evento total, entonces esta aserción es equivalente a la aserción que se compone de todos los eventos excluidos el evento total.

Definición: Sean $a_1 = (X\mathcal{R}d) \in \mathcal{A}'$, $a_2 = (X\mathcal{R}d') \in \mathcal{A}'$ dos aserciones de $\mathcal{A}'$. Se define la intersección de las aserciones a_1 y a_2 como la aserción:

$$a_1 \cup a_2 = (X\mathcal{R}\,(d \cap d'))$$

La intersección de a_1 y a_2 es la aserción vacía si $d \cap d' = \phi^p = (\phi, ..., \phi)\ (p - veces)$. La aserción vacía se denota por $\phi^{\mathcal{A}'}$.

La aserción vacía verifica que:

- $\forall \omega \in \Omega,\ \phi^{\mathcal{A}'}(\omega) = 0$
- $\forall S \subset \Omega,\ Ext_S(\phi^{\mathcal{A}'}) = \phi$

Se verifica la siguiente proposición:

Proposición: Si uno de los eventos que componen la aserción es el evento vacío, la serción es equivalente a la aserción vacía.

Definición: Sea $a = (X\mathcal{R}d) \in \mathcal{A}'$ una aserción de $\mathcal{A}'$. Se define la aserción complementaria de la aserción a como la aserción:

$$a^C = \left(X\mathcal{R}d^C\right)$$

Proposición: Sea la aserción $a = (X\mathcal{R}d)$ se tiene que:

$$a \cup a^C = t^{\mathcal{A}'}$$
$$a \cap a^C = \phi^{\mathcal{A}'}$$

La proposición siguiente demuestra, bajo determinadas circunstancias, la relación entre la unión, intersección y el complemento de aserciones con la extensión de las aserciones resultantes de dichas operaciones, respectivamente.

Proposición: Si $\mathcal{R}$ es una relación de dominio definida en $\mathcal{D}' \times \mathcal{D}$, que verifica $\forall d' \in \mathcal{D}'$, $\forall d_1$, $d_2 \in \mathcal{D}$,

$$d'\mathcal{R}\,(d_1 \cup d_2) = \text{máx}\,\{d'\mathcal{R}d_1, d'\mathcal{R}d_2\}$$
$$d'\mathcal{R}\,(d_1 \cap d_2) = \text{mín}\,\{d'\mathcal{R}d_1, d'\mathcal{R}d_2\}$$
$$d'\mathcal{R}d^C = 1 - d'\mathcal{R}d$$

entonces se cumple para $a_1 = [X\mathcal{R}d_1]$, $a_2 = [Y\mathcal{R}d_2]$, $a = [Y\mathcal{R}d] \in \mathcal{A}'$:

$$Ext_\Omega(a_1 \cup a_2) = Ext_\Omega(a_1) \cup Ext_\Omega(a_2)$$
$$Ext_\Omega(a^C) = \Omega - Ext_\Omega(a)$$

Esta proposición de verifica en el caso particular de que los conjuntos de descripciones sean $\mathcal{D}' = \mathcal{Y}$ y $\mathcal{D} = \mathcal{P}(\mathcal{Y})$ o $\mathcal{D} = \mathcal{Y}$, la relación de dominio $\mathcal{R}$ sea la de pertenencia ($\in$) y la función $g(.)$ de combinación de niveles de relación sea la conjunción lógica.

5.5.2. Conjunción

A continuación, se define la conjunción de aserciones distinguiéndose la conjunción de aserciones definidas sobre distintos vectores de variables de la conjunción de aserciones definidas sobre el mismo vector de variables.

Definición (Conjunción de aserciones definidas sobre distintos vectores de variables): Sean $\mathcal{A}_1$ y $\mathcal{A}_2$ dos conjuntos de objetos simbólicos de tipo aserción definidos sobre el mismo conjunto Ω y asociados a los respectivos vectores de variables monoevaluadas o simólicas X_1, X_2 definidos sobre Ω, tales que los vectores de variables X_1, X_2 no comparten ninguna variable. Y sea $g(.)$ una función de combinación de niveles de relación definida en $[0,1] \times [0,1]$.

Sean $a_1 \in \mathcal{A}_1$, $a_2 \in \mathcal{A}_2$ dos aserciones de $\mathcal{A}_1$ y $\mathcal{A}_2$ respectivamente, entonces la conjunción de a_1 y a_2, se denota por $a_1 \wedge a_2$ y se define como:

$$a_1 \wedge a_2 : \Omega \rightarrow [0,1]$$

$$\omega \rightarrow a_1 \wedge a_2(\omega) := a_1(\omega) \wedge a_2(\omega) = g\left(a_1(\omega), a_2(\omega)\right)$$

Se define el conjunto $\mathcal{A}_1 \hat{\wedge} \mathcal{A}_2$ como:

$$\mathcal{A}_1 \hat{\wedge} \mathcal{A}_2 := \{a_1 \wedge a_2 : a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2\}$$

el conjunto de las conjunciones de elementos de $\mathcal{A}_1$ y $\mathcal{A}_2$.

Esta definición hace referencia a vectores de variables distintos para los conjuntos $\mathcal{A}_1$ y $\mathcal{A}_2$. La función $g(.)$ de combinación de niveles de relación entre dos aserciones a_1 y a_2 puede coincidir o no con las funciones de combinaciones de niveles de relación definidas internamente en las aserciones a_1 y a_2.

En realidad, esta definición permite la construcción de aserciones a partir de eventos y aserciones, dada una función de combinación de niveles de relación. El caso más elemental es la construcción de una aserción a partir de dos eventos y una función de combinación de niveles de relación.

Se deduce la siguiente:

Proposición: Si $a_1 = [X_1 \mathcal{R}_1 d_1] \in \mathcal{A}_1$, $a_2 = [X_2 \mathcal{R}_2 d_2] \in \mathcal{A}_2$, son dos aserciones y $\wedge$ es una función de combinación de niveles de relación definida en $[0,1] \times [0,1]$, entonces $a_1 \wedge a_2$ es una aserción. Además, se identifica la conjunción de las dos aserciones con

$$a_1 \wedge a_2 = [X_1 \mathcal{R}_1 d_1] \wedge [X_2 \mathcal{R}_2 d_2] \text{ y con } \left(a_1 \wedge a_2, \mathcal{R}_1 \times \mathcal{R}_2, (d_1, d_2)\right).$$

Definición (Conjunción de aserciones definidas sobre el mismo vector de variables): Sea $\mathcal{A} = \{[X \mathcal{R} d] : \Omega \rightarrow [0,1],\ d \in \mathcal{D}(\mathcal{Y})\}$ un conjunto de objetos simbólicos de tipo aserción asociados al vector de variables monoevaluadas o simbólicas X definido sobre Ω y con descripciones en un conjunto $\mathcal{D}(\mathcal{Y})$ que tiene definida la intersección, $\cap$. Sean $a_1 = [X \mathcal{R} d_1] \in \mathcal{A}$, $a_2 = [X \mathcal{R} d_2] \in \mathcal{A}$, dos aserciones de $\mathcal{A}$, entonces la conjunción de a_1 y a_2 se denota por $a_1 \wedge a_2$ y se define como:

$$a_1 \wedge a_2 : \Omega \rightarrow [0,1]$$
$$\omega \rightarrow a_1 \wedge a_2(\omega) := [X\mathcal{R}d_1 \cap d_2](\omega)$$

Se define el conjunto $\mathcal{A}\hat{\wedge}\mathcal{A}$ como:

$$\mathcal{A}\hat{\wedge}\mathcal{A} := \{a_1, a_2 \in \mathcal{A}\}$$

el conjunto de las conjunciones de elementos de $\mathcal{A}$.
Se deduce la siguiente:

Proposición: Si $a_1 = [X\mathcal{R}d_1] \in \mathcal{A}$, $a_2 = [X\mathcal{R}d_2] \in \mathcal{A}$ son dos aserciones, entonces $a_1 \wedge a_2$ es una aserción. Además, se identifica la conjunción de las dos aserciones con $a_1 \wedge a_2 = [X\mathcal{R}d_1 \cap d_2]$ y con $(a_1 \wedge a_2, \mathcal{R}, d_1 \cap d_2)$

Ejemplo: Conjunción de aserciones definidas sobre el mismo vector de variables

Si

$$a_1 = [estado\ civil \in \{soltero, casado, viudo\}],$$
$$a_2 = [estado\ civil \in \{soltero, viudo\}]$$

entonces

$$a_1 \wedge a_2 = [estado\ civil \in \{soltero, viudo\}]$$

Si

$$a_1 = [estado\ civil \sim \{soltero, casado, viudo\}],$$
$$a_2 = [estado\ civil \sim \{soltero, viudo\}]$$

entonces

$$a_1 \wedge a_2 = [estado\ civil \sim \{soltero\}]$$

Si

$$a_1 = [profesión \in \{soltero, casado\}],$$
$$a_2 = [profesión \in \{viudo\}]$$

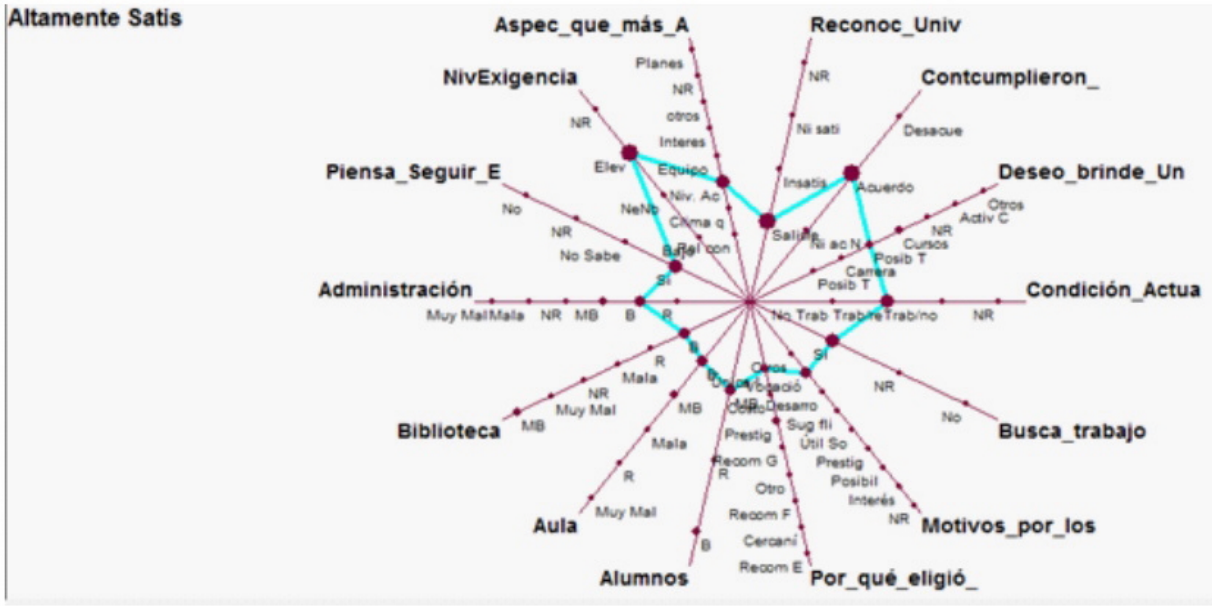
entonces

$$a_1 \wedge a_2 = [profesión \in \phi] = \phi^{\mathcal{A}}$$

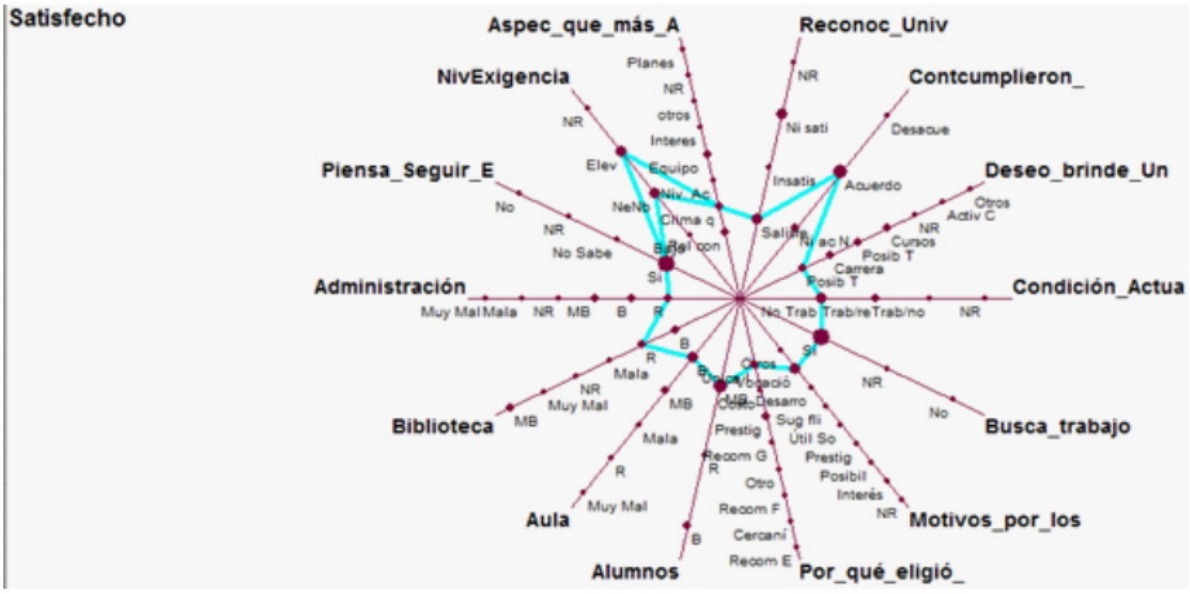
Proposición: Sea $a = [X\mathcal{R}\mathcal{D}]$ una aserción definida sobre Ω, relativa a un vector de variables X, $D = (D_1, ..., D_p) \in \mathcal{D}$ con $\mathcal{D} = \mathcal{Y}$ o $\mathcal{D} = \mathcal{P}(\mathcal{Y})$ y sea X_j una variable definida sobre Ω. Se tiene que las aserciones a y $a \wedge [X_j\mathcal{R}\mathcal{D}_j]$ son equivalentes.

5.6. Ejemplos de objetos simbólicos sobre encuesta Kolla

Con los datos de la encuesta Kolla al recién graduado, se han creado objetos simbólicos que representan a las tres clases de egresados 2014 de la FFHA-UNSJ que se encontraron utilizando clustering a partir de los factores principales de un análisis de correspondencias múltiples, obtenido usando el software estadístico SPAD_N. Estos objetos simbólicos se visualizan y describen a continuación:

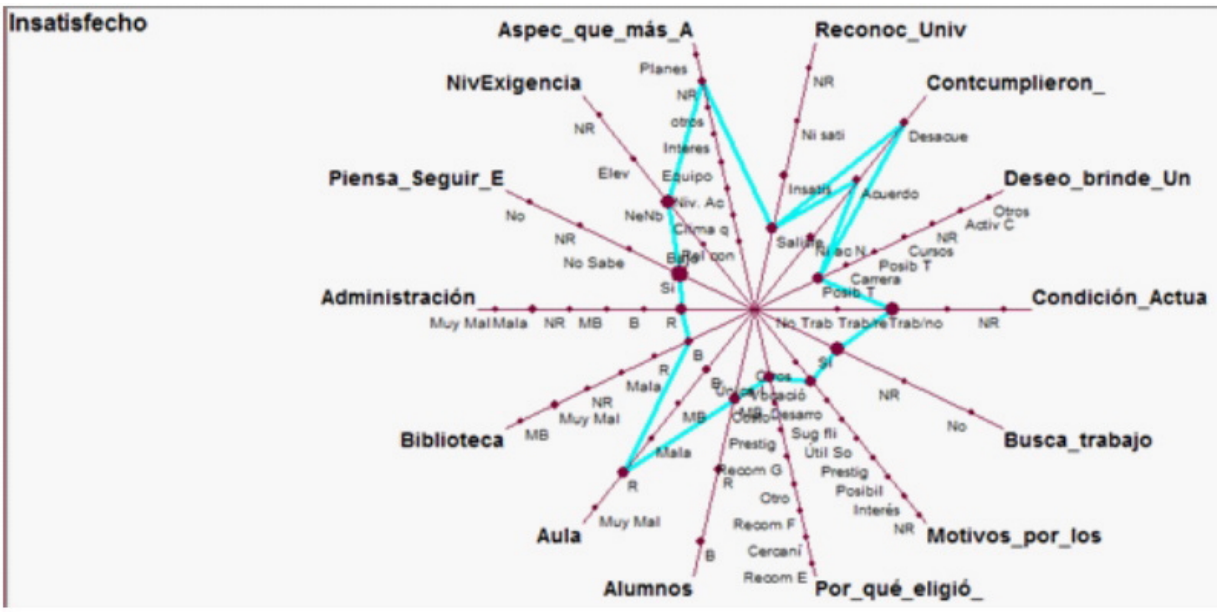


El grupo de alta conformidad se caracteriza porque la trabaja en relación con la carrera estudiada, busca trabajo, eligió su carrera por vocación y la universidad por ser única y por su prestigio académico. Su opinión acerca de sus pares (alumnos) es muy buena y con respecto a aulas, biblioteca y administración es buena. La mayoría piensa seguir estudiando, opina que el nivel de exigencia de la carrera estudiada es elevado, el aspecto que más aprecia es el nivel académico. Piensa que el reconocimiento de la sociedad hacia la universidad es satisfactorio, acuerda con que los contenidos estudiados cumplieron sus expectativas y desea que la universidad le brinde oportunidad de trabajo en investigación.



El grupo que en general está conforme o satisfecho se caracteriza porque la mayoría no trabaja pero buscan trabajo. Eligió su carrera por vocación y la universidad por ser única.

Opinan muy bien sobre sus pares, bien sobre las aulas, regular de biblioteca y administración. La mayoría piensa seguir estudiando. El nivel de exigencia de su carrera fue elevado, el aspecto que más aprecian es el clima que se vive. Acuerdan con que los contenidos estudiados cumplieron sus expectativas y desean que la universidad les brinde posibilidades de trabajo en docencia.



El grupo que hemos denominado insatisfecho se caracteriza porque trabaja en relación a su carrera, busca trabajo, eligió su carrera por vocación y la universidad por ser única. Su opinión acerca de sus pares (alumnos) es muy buena, con respecto a aulas y administración es regular y biblioteca buena. El nivel de exigencia no es alto ni elevado. Desea que la universidad le brinde posibilidad de trabajo en docencia.

Conclusión

Es importante destacar que los objetos simbólicos individuales unitarios pueden ponerse en correspondencia biunívoca con las descripciones de los objetos reales, con las que habitualmente trabajamos en el Análisis de Datos. Las operaciones entre descripciones de objetos reales y entre variables, pueden ser extensibles al caso de objetos simbólicos individuales y unitarios y de variables simbólicas conjuntuales unitarias respectivamente.

La necesidad de extender los métodos de análisis de datos estándar (entre otros: exploratorio, clustering, análisis factorial, discriminación) a las tablas de datos simbólicos para extraer nuevos conocimientos es cada vez mayor, debido a la expansión de la tecnología de la información, que ahora es capaz de almacenar una cantidad enorme de datos. Esto dio lugar a una nueva metodología llamada "Análisis de datos simbólicos", cuyo objetivo es extender estos métodos que tienen como entrada una matriz de datos, a un nuevo tipo de tabla de datos llamada "tabla simbólica de datos". El objetivo del proyecto EUROSTAT, Comunidad Europea, llamado SODAS era producir un primer software para el ADS. El mismo se desarrolló hasta el año 2004, aproximadamente. Los métodos recientes, han sido programados, generalmente en paquetes, bajo la plataforma R.

En este trabajo se han introducido los datos y objetos simbólicos y se han relacionado con los datos monoevaluados. Se han presentado antecedentes de generalización por datos simbólicos de estas clases. Se ha mostrado cómo a partir de estas descripciones simbólicas se pueden definir objetos simbólicos dotándolos de relaciones de dominio que permiten la comparación de las descripciones de los individuos con las descripciones simbólicas.

Las aserciones definidas, por variables, relaciones de dominio y descripciones simbólicas permiten obtener clases de individuos cuyas descripciones se relacionan con dichas descripciones simbólicas. Es decir, las aserciones son instrumentos que permiten la vuelta a la base de datos original mediante la obtención de sus extensiones. Las aserciones son independientes de las bases de datos a partir de las cuales fueron creadas. Esto significa que las extensiones de las mismas pueden aplicarse a bases de datos diferentes o a una misma base de datos en distintos instantes de tiempo. Por lo cual, permiten la propagación de conceptos.

Por otra parte, las intenciones de las aserciones pueden ser definidas por un experto sin necesidad de ser creadas a partir de bases de datos. También de esta manera se puede acceder a una base de datos y obtener mediante una consulta a la misma, los individuos que se relacionan con esas intenciones.

Los objetos simbólicos constituyen un nuevo sistema de representación del conocimiento que engloba, en un mismo formalismo, tanto conocimientos obtenidos de una base de datos como conocimientos aportados por un experto, siendo de mayor complejidad que los datos habituales. En este sentido, este sistema de representación es más rico que otros sistemas anteriores de representación del conocimiento.

Se ha destacado también la importancia de un marco común para los tres enfoques de expresión de incertidumbre: la probabilidad, la posibilidad y la creencia. Y se han formalizado, en este contexto, las dos primeras.

Antecedentes a la creación de un marco común de representación de la incertidumbre pueden verse en Ruspini, (1990) con un modelo unificado semántico que permite comparar las tres expresiones del razonamiento aproximado y en Dubois y Prade, (1989) que proponen un marco común de combinación de información de diversas fuentes y tipos.

El formalismo presentado en la última sección de esta monografía, incluye además otros tipos de datos como son los conjuntos de valores, los intervalos y la inclusión de metadatos, lo que enriquece aún más este tipo de representación. Se han formalizado los primeros e introducido los demás.

Por último, destacar también que este formalismo permite analizar conjuntamente todos estos tipos de datos, siempre que se establezcan las correspondientes relaciones producto de

las aserciones. Se ha introducido que este modo de representación admite formalizar datos y objetos aún de mayor complejidad como son los objetos simbólicos modales exteriores, horda, síntesis, disyunciones, etc, que exceden el ámbito de esta monografía.

Finalmente se han definido operaciones entre objetos simbólicos.

A modo de ejemplo se han creado objetos simbólicos que describen a las tres clases de egresados 2014 de la FFHA-UNSJ que se encontraron utilizando clustering a partir de los factores principales de un análisis de correspondencias múltiples, obtenido usando el software estadístico SPAD_N.

Bibliografía

[1] "Análisis de Segmentación en el análisis de datos Simbólicos", María del Carmen Bravo Llatas . Tesis de la Universidad Complutense de Madrid, Facultad de Ciencias Matemáticas, Departamento de Estadística e Investigación Operativa. (2001)

[2] "Bases conceptuales para una teoría de objetos Simbólicos", José Ruiz Shulcloper, Martín G. Chac Kantún, José F. Martínez Trinidad, Computación y sistemas. Vol 1. (1997)

[3] "La sociedad de la información analizada mediante objetos simbólicos", Patricia Calvo Garrido, Yolanda Pérez Díez, Instituto vasco de estadística. (2001)

[4] "Las nubes de datos" Métodos para analizar la complejidad, Nora Moscoloni. Primera edición. Editorial de la Universidad Nacional de Rosario. (2011)

[5] "Comparing Dissimilarity Measures for Symbolic Data Analysis", Donato Malerba, Floriana Esposito, Incenzo Gioviale, Valentina Tamma. Departamento de Informática, Universidad de Bari.

[6] "The state of the art in symbolic data analysis: overview and future", Edwin Diday. (2008)

[7] "An Introduction to Symbolic Data Analysis and the Sodas Software", Edwin Diday. University Paris, Dauphine.

[8] "Algoritmos para la clasificación piramidal simbólica", Oldemar Rodríguez, Maria Paula Brito, Edwin Diday. Revista de Matemática: Teoría y Aplicaciones. (2000)