ANÁLISIS DE DATOS SIMBÓLICOS

Edwin Diday

Colaboradores nacionales por el área en IRCE

IRICE

INSTITUTO ROSARIO DE INVESTIGACIONES EN CIENCIAS DE LA EDUCACIÓN
CONICET - UNR

ANÁLISIS DE DATOS SIMBÓLICOS

Edwin Diday

Conferencias pronunciadas por el autor en IRICE los días 15 y 16 de Julio de 1993

Edición a cargo de Nora Moscoloni

I S B N En trámite

Impreso en Argentina / Printed in Argentine

ÍNDICE

PRIMERA CONFERENCIA

Presentación:

Soy Presidente de la Sociedad Francófona de Clasificación hasta fin de año, es un cargo que se renueva todos los años.

Existe una Federación de Sociedades de Clasificación que se creó en 1979 y que agrupa a varias sociedades nacionales. La más activas son algunas sociedades norteamericanas, la sociedad alemana, la francesa, la inglesa y la italiana.

Como profesor de la Universidad de París IX y Director de Investigación en el INRIA, dirijo dos equipos de investigación. Mi trabajo es como el del marino que debe llevar el timón, consiste en dar direcciones. Estas dos conferencias se referirán sobre todo, a las directrices en las cuales posiblemente converjan las investigaciones en el Análisis de Datos. (AD en adelante).

La Ciencia de los Objetos. Una revisión histórica.

Cabe destacar que uno de los aportes fundamentales del AD, con respecto a la estadística clásica es darle más importancia a los individuos. Esta última se interesa más en las variables, buscando leyes de probabilidad y los individuos en sí mismos no aparecen tanto.

En forma general, más que en individuos centraré mi atención en los objetos. Por ejemplo, cuando se hace una encuesta en marketing, los individuos que son encuestados pueden ser llamados de una forma más general objetos, o cuando se efectúa una encuesta médica, los individuos de los que se toman muestras de sangre pueden ser llamados también objetos. En un estudio geográfico donde se observa el comportamiento demográfico o económico de los departamentos, éstos pueden ser considerados también como objetos.

El objetivo en el AD es estudiar una matriz de datos, con variables en las columnas e individuos en las filas.

En cambio en la estadística clásica si por ejemplo hacemos un histograma de la variable y_i se puede inducir una ley de probabilidad sobre y_i, elegir el o los modelos. Por lo tanto los individuos no son tan interesantes salvo si no se ajustan al modelo.

Por el contrario cuando se efectúa un Análisis Factorial de Correspondencias, por ejemplo, se representan sobre un mismo plano las variables y los individuos

En el Multidimensional Scaling también se le da importancia a los individuos, hay una tendencia a partir de una intención de modelización de variables a darle una mayor importancia a los individuos.

El punto fundamental es mirar más de cerca cuál es el sentido de esos individuos que están allí. Es en este momento cuando interviene la noción de objeto. Cuando se dice una silla,

que es un nombre común, se designa a un objeto y cuando se expresa esta silla, que es un silla individual, también se designa a un objeto.

En el primer caso, un objeto es una clase, cuando se dice "silla" se está nombrando un sustantivo común que representa el conjunto de todas las sillas individuales, en cambio cuando se dice esta silla se está designando a un individuo.

Si se mira más de cerca esta noción de objeto, descripta de esta manera, se puede llegar muy atrás en la historia de la humanidad. Aristóteles en el siglo IV A.C. introduce la noción de "Ciencia de los Objetos" y se encontró en uno de sus libros exactamente esta expresión. Es el libro "De Partibus Animalis" donde se interesa por las especies de animales, en los objetos considerados como especies o sea como clases.

La noción de clase apareció seguramente en el primer organismo viviente porque tenía que reconocer lo que podía de lo que no podía comer. O sea que es una noción muy antigua. Cuando el hombre prehistórico dibujó el mamut en su caverna o en su gruta, él representó más bien a una clase, la clase de los mamuts, más que al que se acababa de comer.

Es decir que ya 30.000 años antes de Jesucristo la gente pensaba en términos de clase. En el Génesis de la Biblia, 2000 A.C., Dios le dice a Adán que le dé un nombre a los animales de la tierra y a los pájaros del cielo. Esto también de darle un nombre a los animales, es ya conocer una noción de clase. Por lo tanto es una noción muy antigua.

Es necesario distinguir la noción de "clase" de la noción de "descripción de la clase". Aristóteles en "De Partibus Animalis" cuando habla de ciencia de los objetos muestra claramente la diferencia entre un objeto y su descripción. Esto parece simple decirlo de esta manera pero es una noción fundamental que los hombres tardaron mucho tiempo en descubrir.

Cuando se habla de clase se piensa en un conjunto en el sentido matemático, una colección de objetos, pero hay una segunda noción de clase que no es solamente la noción matemática de conjunto, sino que es la descripción de la clase.

Por ejemplo la silla que está allí o el individuo que está representado en el pizarrón se pueden describir por los valores que toman unas variables, o se puede describir un conjunto de mesas, de sillas, empresas que son productivas o que no lo son. Éstas son clases de empresas que forman una colección de objetos pero que también están descriptas por sus propiedades.

Intención y extensión de una clase.

Aquí aparece la noción de intención y extensión de una clase. Esta noción de intención y extensión se halla muy claramente en los autores Arnault y Nicole de la Escuela de los Lógicos de Port Royal en el siglo XVIII.[3]

Es un texto que está de moda actualmente y se los encuentra citados en muchos artículos. Allí se expresa que "la intención de una idea son los atributos o las variables que la describen y que no pueden ser suprimidos sin destruirla ...; la extensión de una idea es el conjunto de los individuos u objetos a los cuales estas propiedades se aplican".

Por ejemplo, si se piensa en la noción de auto esta idea no se encuentra en el espíritu, (los psicólogos han reflexionado mucho sobre esto). Hay un modo de descripción de la noción de auto que nos permite reconocerlo tal que cuando se ve pasar a uno de ellos, se dice esto es un auto.

La clase de los autos, de todos los autos, se representa en mi espíritu, no como un conjunto de autos sino como un conjunto de propiedades que describen a los autos. No es por lo tanto la clase como el conjunto en el sentido matemático sino la descripción de la clase.

No podríamos funcionar si tuviéramos todos los objetos del mundo en la cabeza, no es la teoría de los conjuntos lo que tenemos en la cabeza sino más bien la teoría de las intenciones. No son las extensiones lo que nos permite funcionar sino las intenciones. Esto es la base del Análisis de Datos Simbólico (en adelante el ADS). El ADS va a consistir en trabajar no sobre las extensiones es decir sobre los individuos, sino en reemplazar los individuos por las intenciones.

Por ejemplo, en una encuesta en Francia sobre la droga se entrevistó a 2000 adolescentes, bien elegidos en diferentes categorías sociales, se les preguntó sobre antecedentes familiares, consumo de cigarrillos, de bebidas alcohólicas, eventualmente de drogas . El problema era encontrar clases de adolescentes que tuvieran el mismo comportamiento, había que encontrar grupos de riesgo.

En términos del AD clásico el objetivo era encontrar, a partir de la matriz de datos, (olvidando todos los conocimientos de los expertos: psicólogos, sociólogos, psiquiatras, que habían trabajado en la encuesta) clases de comportamiento de manera lo más objetiva posible y modelizando lo menos posible.

Si nos ubicamos en el lugar del ADS podríamos apoyarnos más en el conocimiento de los expertos. Es decir, preguntar por ejemplo a los expertos que digan a partir de su experiencia sobre el terreno, si tienen una idea de comportamientos posibles de la población. Estos comportamientos serán descriptos en intención porque en la cabeza de los expertos hay una idea de clases de comportamientos de adolescentes, pero esta idea de clase no está expresada solamente en términos de individuos porque no son capaces por su experiencia de nombrarlos: Pablo, Juan, etc. Estas son clases descriptas en intención por una colección de propiedades.

Si un experto epidemiólogo que ha trabajado 10 años sobre terreno, provee esta descripción en intención de esta clase, es necesario utilizar esta noción.

El ADS va a poder hacerlo porque va a permitir escribir, bajo forma matemática, esta clase que le interesa al experto. Y una vez que se tiene la descripción matemática en intención se va a poder ver cuál es su extensión en la base de datos y aceptar o rechazar esta hipótesis de clase.

Si los expertos son capaces de describir varias clases en intención vamos a poder crear un nuevo tipo de matriz de datos donde los objetos ya no son individuos sino que son objetos definidos en intención, por lo tanto son objetos de primer nivel.

El ADS tiene entonces por finalidad hacer estudios de AD sobre este tipo de objetos definidos en intención. Este tipo de análisis contiene un caso particular, el caso donde las clases son individuos. Es entonces una extensión y una generalización del AD clásico.

Podemos ver la evolución del análisis de la siguiente manera: la estadística clásica se interesa sobre todo por la modelización de una población vista globalmente; el AD clásico comienza a interesarse por los individuos; el ADS generaliza la noción de individuo interesándose también en los objetos en sentido general, más en cuanto a objetos, que en cuanto a individuos.

Esta evolución que parece natural no aparece solamente en AD sino también de forma general en informática, en inteligencia artificial y en las ciencias del conocimiento.

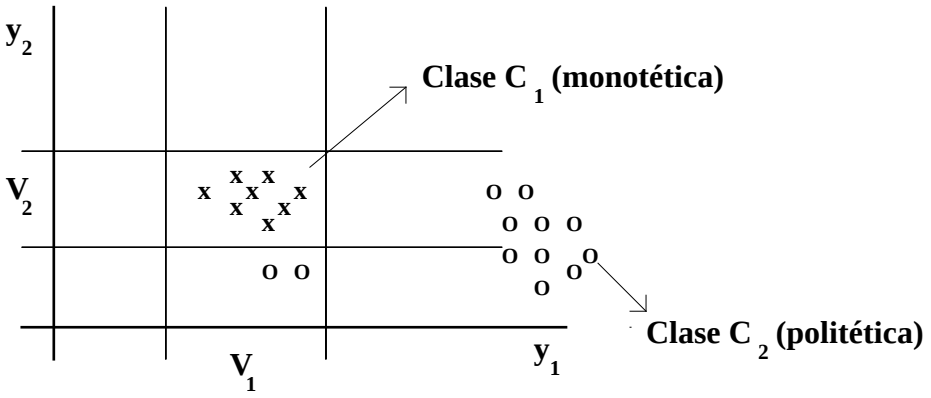
Representación y obtención de una clase. Clases monotética y politética

El primer texto que hemos leído sobre este tema es de Jevons de 1877 [46]. Beckner en 1959, [6] define exactamente la diferencia entre clase monotética y clase politética.

Una clase es monotética si existe un conjunto de propiedades necesarias y suficientes satisfechas por los individuos de la clase. Una clase es politética si no hay una condición necesaria y suficiente que haga que un individuo sea miembro de una clase.

Por ejemplo si se toma esta población de cruces caracterizada por las variables y_1 e y_2 se observa que todos los individuos de la clase C_1 están en el intervalo V_1 y en el intervalo V_2.

Fig.1



Un individuo a_1 pertenece a la clase C_1 si

$$a_1 = [y_1 = V_1] \wedge [y_2 = V_2]$$

O sea: para pertenecer a la clase C_1 debe cumplir estas condiciones y si cumple estas condiciones pertenece a la clase (condiciones necesarias y suficientes). Por lo tanto es una clase monotética.

Si el criterio se hubiera restringido a V_1 solamente no sería una condición suficiente porque hay otros individuos de otras clases, si se hubiera tomado V_2 solamente también habría otros individuos de otras clases por lo tanto ni V_1 ni V_2 son, cada una, condiciones necesarias y suficientes. La condición de monoteticidad es que exista una condición necesaria y suficiente, al menos, que haga que todos los individuos que la satisfagan estén en la clase.

Por el contrario si se ve la otra clase (C_2) hay individuos que son redondeles. No hay condiciones necesarias y suficientes utilizando y_1 e y_2 que hagan que esta población de la clase de los redondeles pueda ser descripta por propiedades. No hay, por lo tanto, condiciones necesarias y suficientes satisfechas por la clase de los redondeles.

¿Cómo describir una clase, cómo representarla y cuál es la intención de una clase?. Existen tres tradiciones:

Una llamada aristotélica que da clases monotéticas y que describe las clases por una conjunción lógica de propiedades.

La tradición adansoniana [Adanson, [1] y [2]] en donde simplemente una clase no está caracterizada por una conjunción de propiedades como lo decía Aristóteles sino por un alto grado de semejanza.

Una tercera tendencia que proviene de los psicólogos es la de Rosch en 1978, que dice que la tradición aristotélica es mala, que no se puede representar un concepto, por ejemplo el concepto de manzana, mediante una conjunción de propiedades. En nuestro espíritu el concepto de manzana se efectúa, a través de uno o varios prototipos. Por ejemplo el concepto de manzana puede estar representado por el de una manzana roja. O sea una clase puede representarse por ejemplos muy representativos que son llamados prototipos.

Éstas son, resumiendo, las tres tendencias que he encontrado en la literatura. Es decir la representación lógica, la noción de semejanza o de prototipo.

Ya se ha indicado cómo se representan las clases. Ahora se debe definir cómo obtenerlas.

Aquí también podemos retomar las tres tendencias precedentes. En la tradición aristotélica podemos definir una clase mediante un proceso de arriba hacia abajo que va a permitir elegir las propiedades que caracterizan a cada una de las clases, de la más general a la más específica.

Por ejemplo: si se tienen tres variables explicativas y una variable para explicar yo puedo buscar la variable más explicativa de **y**.

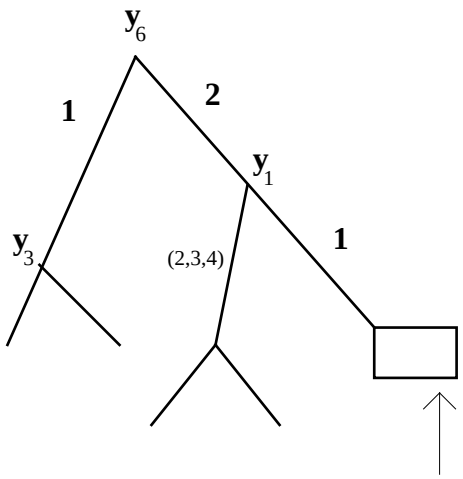
Fig.2

Individuos	Variables			
	y_1	y_2	y_3	y
w_1				
w_2				
...				
w_u				

Pero **y** no es forzosamente una variable exterior, puede ser simplemente una variable artificial que da su número a cada individuo. Entonces se busca la variable explicativa que

explique mejor las variables a explicar y encontramos por ejemplo la y_6 que toma dos modalidades y se obtienen dos ramas, una para cada lado. Entre la gente que tiene $y_6 = 1$ se busca cuál es la variable que explique mejor la **y** y así sucesivamente.

Fig.3



$C_1 : a_1 = [\, y_6 = 2 \,] \wedge [\, y_1 = 1 \,]$ [1]

$C_j : a_j = \hat{\imath}\, [\, y_i = v_i \,]$

Si se toma la rama derecha, se describe la población que toma como valor $y_6 = 2$ e $y_1 = 1$. Esta clase (C_1) se hallará descripta por [1]

Esto se halla dentro de la tradición aristotélica ya que es una clase monotética y en Francia se llama método de segmentación. Este método se remonta a Jussieu en 1774 que da el principio llamado de subordinación de los caracteres por el cual una clase está caracterizada por una conjunción de propiedades heredadas de sub-clases.

Varios autores trabajan en esta línea: Morgan y Sonquist (1963) sobre programas para Marketing; Breiman y Friedman et al. (1984) [10] en Estadística; Quinlan (1986) [66] en Inteligencia Artificial. Todas estas personas construyen árboles y encuentran clases monotéticas es decir de arriba hacia abajo.

El segundo método para obtener clases es otro tipo de algoritmo de abajo hacia arriba. Se puede nombrar a Adanson (1757) que da el primer algoritmo de clasificación ascendente jerárquico que es utilizado hasta nuestros días.

Adanson era realmente un precursor pero en su época la gente no se dio cuenta, tenía ideas revolucionarias pero fue necesario esperar doscientos años para utilizarlas.

Sus contemporáneos decían que perdía mucho tiempo porque él mismo deseaba hacer los cálculos a mano y no estaban equivocados. Quería utilizar este enfoque para organizar todo el mundo animal y vegetal

Hay que remarcar que a menudo los biólogos van por delante de las cosas, se plantean problemas de clasificación en biología sin darse cuenta a menudo, desgraciadamente, que los

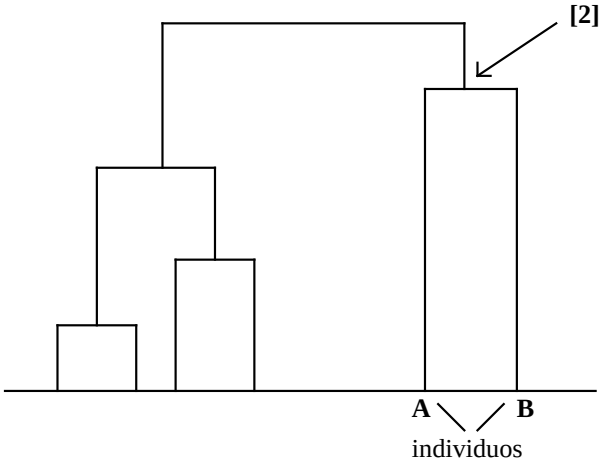
problemas que se plantean en biología son problemas del conocimiento en general, son problemas epistemológicos que tienen aplicaciones en los otros campos y están en la raíz del conocimiento.

Cuando se quiere hacer clasificación lo que se aplica en biología se aplica luego en análisis de mercado, en economía, en epistemología, en geografía, en demografía, y se puede hacer una larga lista.

Los biólogos han sido siempre de avanzada y nuestro trabajo ha sido entender lo que ellos hacían en biología para comprender lo que estaba afuera de la biología y generalizarlo.

Este método consiste en un algoritmo ascendente (de abajo hacia arriba), se comienza con clases reducidas de individuos, se reúne los que son más parecidos y se efectúa nuevamente el procedimiento.

Fig.4



[2] C_j:

$$a_J = \hat{i}\left[y_i = q_i\right]$$

$$q_i = \text{prob}$$

$$C_1 : a_1 = \left[\text{color} = \text{a menudo rojo, rara vez blanco}\right]$$

Se toman los dos puntos más cercanos, se reúnen y se forma una nueva clase. Así se continúa sucesivamente.

O sea, como se han utilizado semejanzas, las clases obtenidas son politéticas, pero por supuesto teniendo una clase puedo dar de ella una descripción monotética. Esta descripción tal vez va a cubrirla, es decir, que es la intención de esta clase, pero no es una condición necesaria y suficiente para los individuos que la forman. Se pueden encontrar individuos exteriores que la satisfagan.

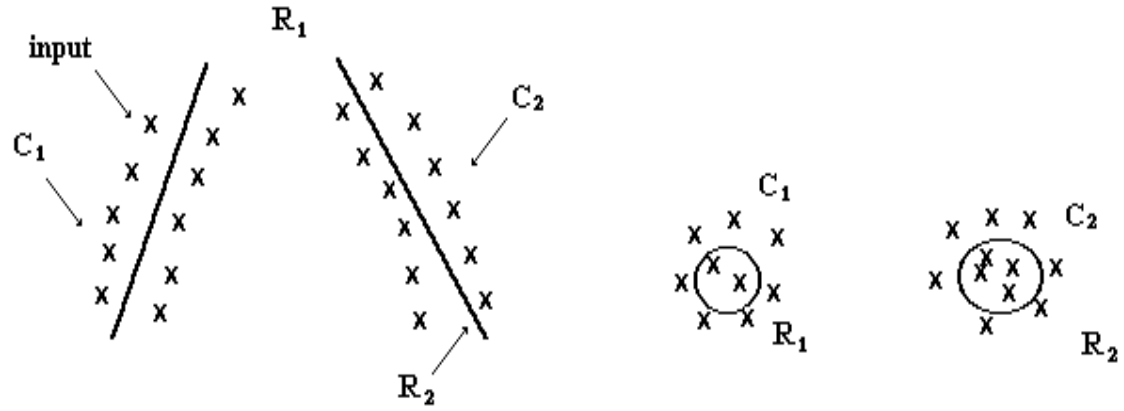
Por ejemplo: los individuos (A y B de la Fig.4) son rojos o blancos, pero hay individuos rojos o blancos que no están en esta clase, entonces no es una clase monotética.

Otra manera de representar a una clase politética, consiste en decir: en esta clase están los individuos que son a menudo rojos y rara vez blancos. Aquí tenemos otra noción que es una descripción seudológica que describe sin embargo la clase pero que no es monotética. Consiste en buscar directamente clases y su representación.

En síntesis: en el primer método se buscaban primero las descripciones y al cabo de una rama se deducía una clase, es decir, la representación eran las ramas. En el segundo, primero se efectúa una agregación para encontrar clases, se encuentran los niveles de la jerarquía y luego se hace la representación por una descripción.

El tercer método, consiste en hacer las dos cosas al mismo tiempo, se pueden buscar simultáneamente las clases y su representación. Aplicando estadística clásica, se calculan dos regresiones, o bien dos ejes factoriales, esto equivale a buscar simultáneamente dos clases, tal que la adecuación entre cada clase y cada representación sea la mejor posible.

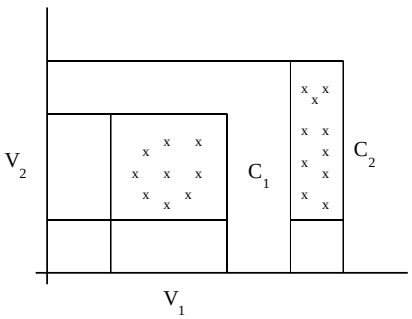
Fig.5



El input del programa son las cruces. El objetivo es encontrar que hay dos clases (C_1 y C_2) y dos representaciones (R_1 y R_2) formadas por los ejemplos más típicos: dos cruces R_1 y cuatro cruces R_2. Por lo tanto se busca simultáneamente dos clases y la representación de las clases.

Por ejemplo aquí se trata de encontrar clases y la representación lógica según la tradición aristotélica de cada una de las clases. Se buscan las clases y la conjunción de propiedades que caracterizan las clases.

Fig.6



$$c_j : a_j = \hat{i}\left[y_i = v_1\right]$$

$$c_1 : a_1 = \left[y_1 = v_1\right] \wedge \left[y_2 = v_2\right]$$

La clase C_1 está caracterizada por y_1 que toma el valor V_1 e y_2 que toma el valor V_2. En esta tercera orientación se encuentra el método de nubes dinámicas [26] [28] [39] y [11] [61]. Estos últimos trabajos describen asimismo parte del ADS.

Análisis de Datos Clásico

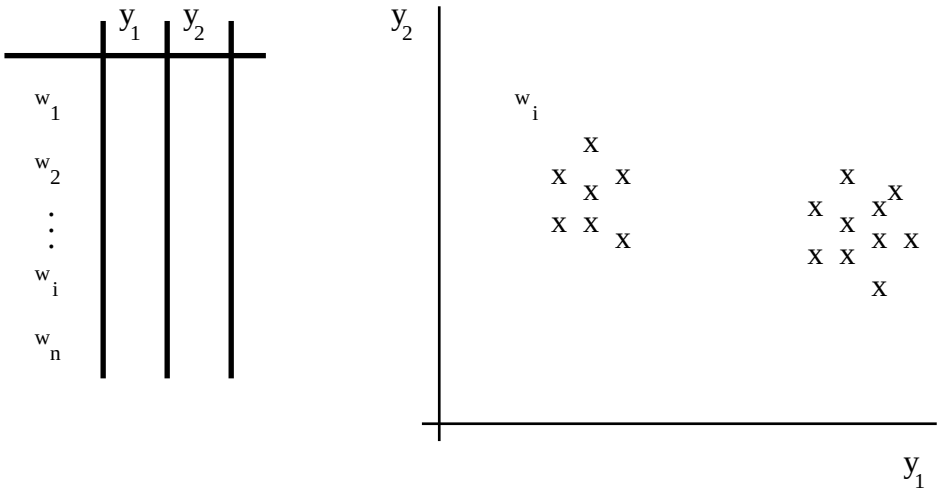
Algoritmos de clasificación.

Clasificación por particiones

Para desarrollar el ADS es mejor antes analizar algunos algoritmos clásicos de clasificación que les harán intuir lo que es el AD.

La entrada del programa son individuos caracterizados por dos variables, en este ejemplo, pero puede haber muchas más. Estas variables se pueden representar en el plano.

Fig. 7

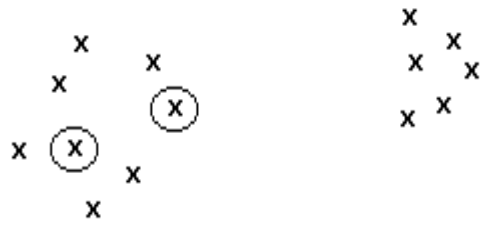


Cada punto representa una fila de la matriz, esto es la entrada. Puede ser una encuesta, por ejemplo, donde se tienen los individuos y las variables que describen a estos individuos.

Las variables pueden ser preguntas para encontrar índices o indicadores de votos para las próximas elecciones presidenciales. El problema es encontrar dos clases.

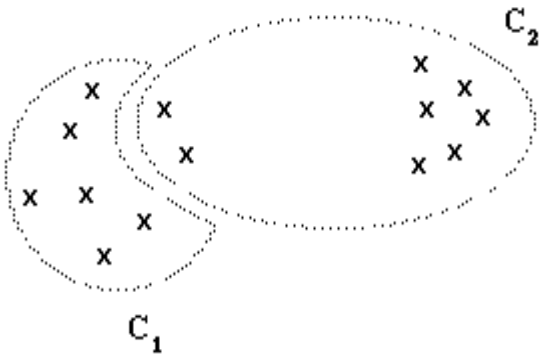
El algoritmo es el siguiente: se eligen dos puntos al azar que sean núcleos, o semillas (seeds).

Fig.8



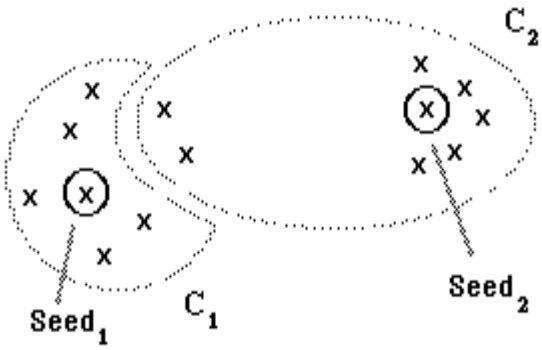
La etapa siguiente consiste en asociar cada punto al núcleo más cercano, cuando se dice el núcleo más cercano se sobrentiende que se está definiendo una distancia. Esta distancia puede ser la distancia euclidiana del plano si se toma la regla y se mide la distancia. Se asocia cada punto al núcleo más cercano y se van a obtener dos clases: la clase de los puntos que están más cercanos a $Seed_1$ y la clase de los puntos que están más cercanos a $Seed_2$.

Fig.9



En la tercera etapa dos nuevos puntos o seeds son extraídos de la siguiente manera: minimizando la distancia del punto que se busca a todos los otros puntos de su misma clase.

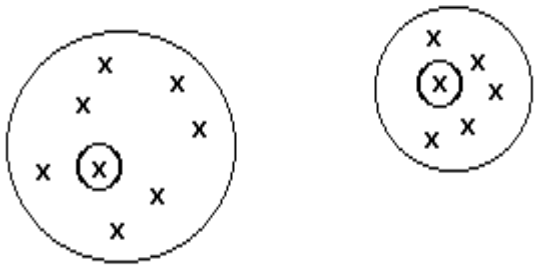
Fig.10



Es decir, en cada clase, para cada punto, se calculan las distancias a todos los otros puntos de la clase y se retiene el que da la menor suma de las distancias a todos los otros puntos.

Las clases C_1 y C_2 fueron obtenidas en la etapa precedente. Se vuelve a la etapa 2, posteriormente a la etapa 3 y así sucesivamente hasta la convergencia. La etapa 2 era la etapa que permitía: conociendo los centros construir clases, por lo tanto se encontraban dos nuevos núcleos y se asociaba cada punto al núcleo más cercano. Se va a ver entonces que un punto que está en una clase puede moverse y pasar a otra clase porque está más cercano de ese núcleo que del otro y así se encuentran dos nuevas clases.

Fig.11



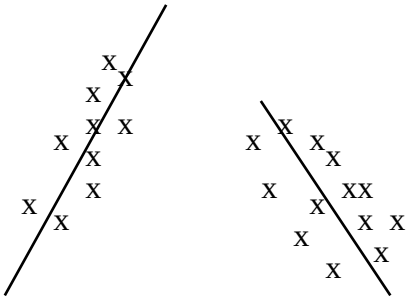
Se vuelven a calcular los centros, otras clases y así sucesivamente.

Este es un ejemplo simple de algoritmo de partición que busca clases y sus representaciones. Se puede decir que es una búsqueda de prototipos en la tradición de Rosch.

Se pueden tomar otros tipos de núcleos. Por ejemplo como se había enunciado antes se pueden tomar núcleos que no son puntos, considerar que el núcleo es una recta y en este momento se buscan dos clases y dos rectas tal que la adecuación entre las clases y las rectas sea la mejor posible.

Lo mismo en lugar de rectas se puede tomar un análisis factorial de rectas o de rectas distancias adaptivas o de prototipos que son puntos de la población.

Fig.12



Ej.:

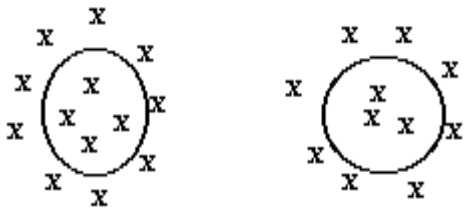
. Ejes factoriales

. Regresión local

. Distancias adaptativas

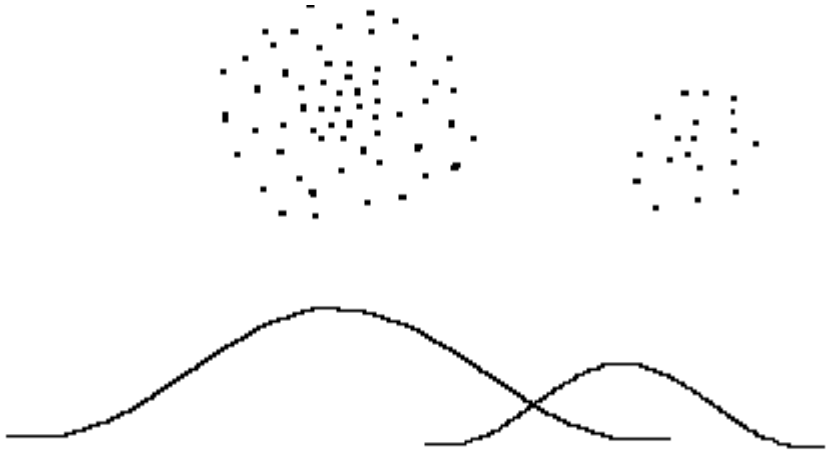
Hay clases y hay prototipos expresando estas clases.

Fig.13



En términos más estadísticos se puede tomar un input que son puntos y se buscan dos clases, dos leyes de probabilidad, por ejemplo la ley normal, tal que las dos clases que se encuentren y las dos leyes tengan la mejor adecuación posible; este problema es el que se llama mixture decomposition o descomposición mixta.

Fig.14

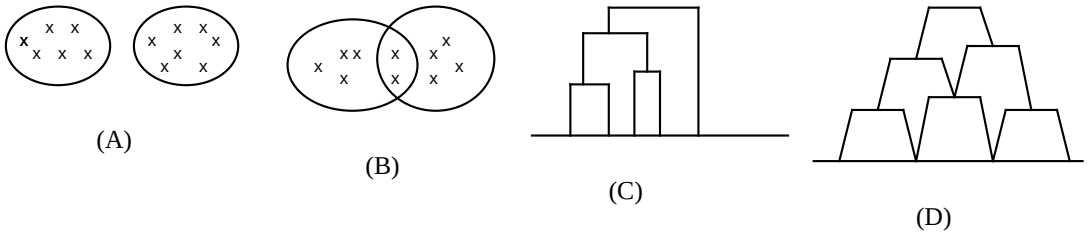


Método de Clasificación Dinámica

¿Qué es el método de las nubes dinámicas?

La primera etapaconsiste en elegir una estructura interclase C ya sean: una o varias particiones (A), particiones o clases que se superponen en parte (B), jerarquías (C) o pirámides (D).

Fig.15

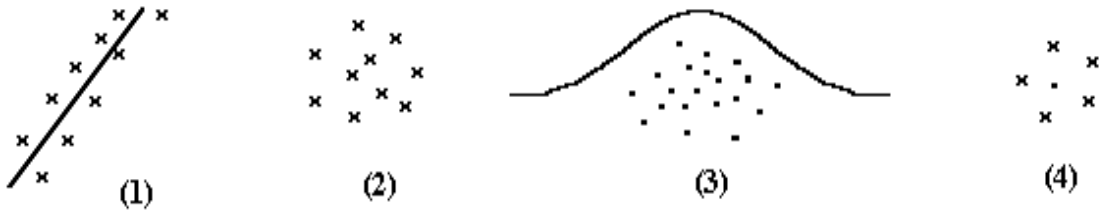


Las pirámides son las clases que pueden tener puntos en común, las jerarquías son hermafroditas porque cada clase no tiene más que un padre o una madre. En las pirámides cada clase puede tener dos ascendientes no son forzosamente hermafroditas, se obtienen clases que se superponen.

En la segunda etapa se elige un espacio de representación L

. Representar las clases por ejes factoriales(1), donde el espacio de representación son ejes factoriales, prototipos (sub-clases) (2), leyes de probabilidad (3), centros de gravedad (4), puntos de R^p.

Fig.16



En la tercera etapa hay que elegir un criterio matemático que mida la adecuación entre la representación y las clases.

Este criterio es de la forma:

$$W(S,P) = \sum_{i=1,R} D(S_{i,} P_i) \text{ donde } S_i \in L \text{ y } P_i \in C$$

D es una medida de adecuación entre la representación de la clase Pi que se llama Si., esta ecuación es la que hay que optimizar. Hay que encontrar S y P que optimicen W.

$S = \{S_1,... S_R\}$, $P = \{P_1, ..., P_R\}$

¿Cómo hacer para optimizar W?.

Es la quinta etapa del método de las nubes dinámicas . Se calcula iterativamente como en el algoritmo precedente una representación, las clases, una nueva representación, nuevas clases, etc.

$S^0 \rightarrow P^1 \rightarrow S^1 \rightarrow P^2 \rightarrow ...S^n \rightarrow P^n$, $W(S^n, P^n)$ decrece hasta converger

Las representaciones son elegidas en el espacio L y las clases son elegidas en uno de los espacios de representación seleccionados.

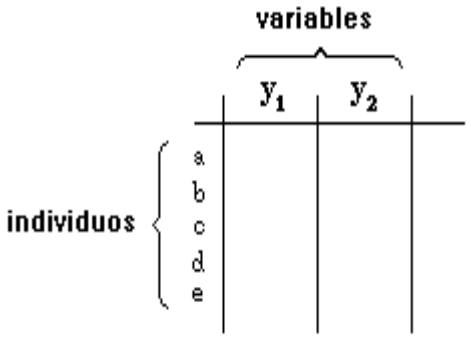
Se ha descripto el método de clasificación por búsqueda de adecuación entre una representación y las clases, a continuación se desarrollará un método de clasificación jerárquica según la segunda tendencia, la de Adanson.

Método de Clasificación Jerárquica

Su principio básico es el siguiente: en cada etapa, se reúnen las dos clases que más se parecen, al inicio cada individuo es considerado como clase y se van agregando luego tantas clases hasta que todos los individuos sean incluidos.

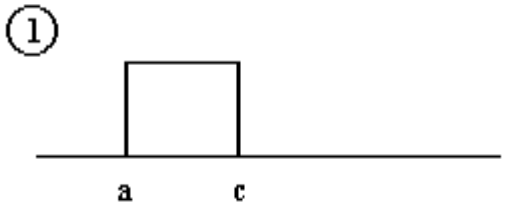
Por ejemplo:

Fig.17



De entrada hay cinco individuos y dos variables, se va a tratar de encontrar el par de individuos que estén más cercanos. La primera etapa consiste en buscar los dos puntos que estén más cercanos para formar la primera clase.

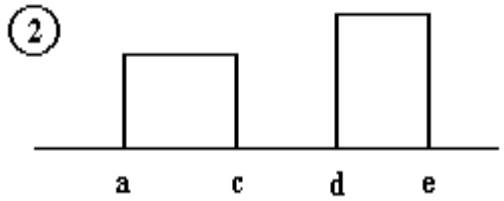
Fig.18



Se determinan los dos puntos más cercanos, y se representa la altura que es la distancia entre estos dos puntos.

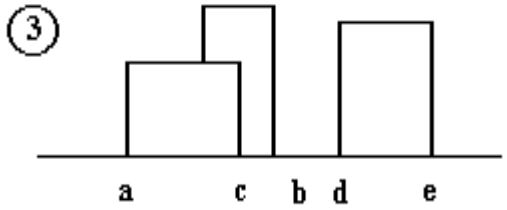
Luego se pregunta. ¿Cuáles son las otras dos clases más cercanas? Los puntos d y e resultan ser los más cercanos.

Fig.19



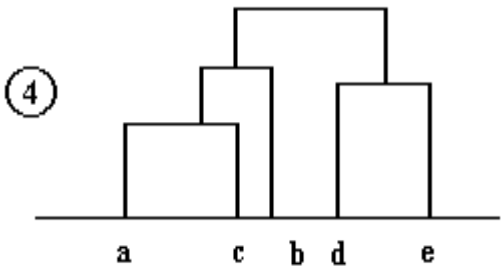
Posteriormente, lo más cercano es b con c y a.

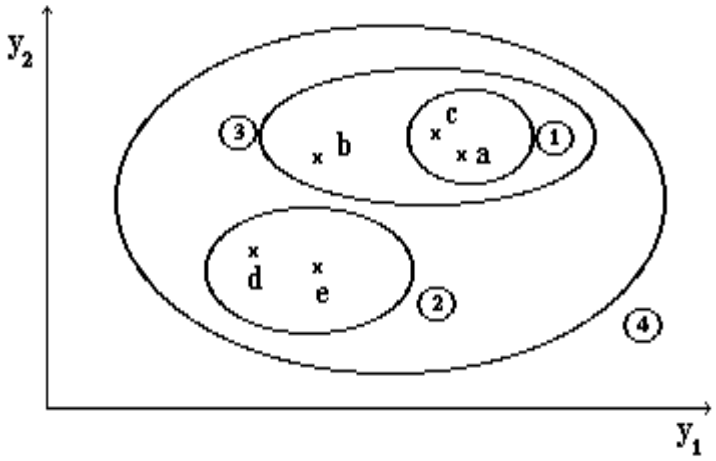
Fig.20



Se calcula una distancia entre b y ca, es decir la suma de las distancias d (b,c) y d (b,a) lo que resulta el tercer nivel. El último nivel se obtiene agregando las otras dos clases.

Fig.21

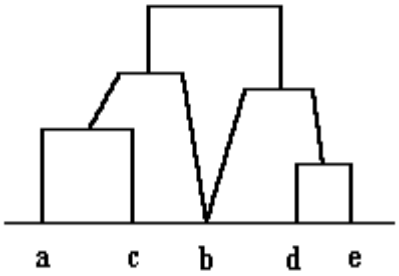




De esta forma las clases están encastradas, este método fue ideado por Adanson en 1757.

Actualmente existen variantes más eficaces pero la idea de base es ésta. Se pueden también hacer pirámides con el mismo procedimiento. Se va agregar primero a y c, luego d y e, luego b se va agregar tanto a a y c como a d y e.

Fig.23



Se observan dos clases que tienen un punto en común y se obtuvo otra representación: (a,c,b), (b,d,e)

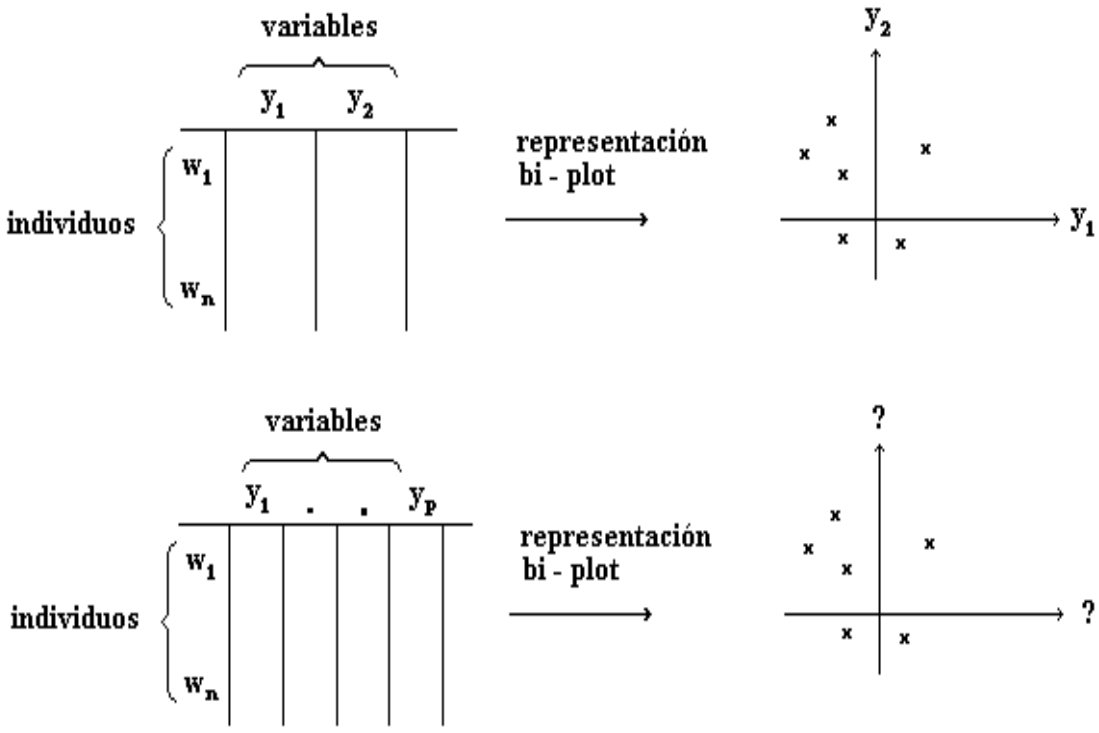
Escalamiento Multidimensional y Análisis Factorial.

Cuando se tienen individuos y dos variables siempre se pueden representar en un plano. Los métodos de Escalamiento Multidimensional y el Análisis Factorial responden al problema de representar individuos de espacios de p dimensiones, en lugar de representar individuos de espacios de dos dimensiones.

¿Cómo hacer una representación a dos dimensiones cuando se tienen p variables?. En el caso de dos variables solamente, es fácil, pero en el caso de p variables el problema se complica. Sin embargo es interesante poder representar una nube de puntos que existe en el espacio de p dimensiones.

Existen dos maneras de realizarlo. Se puede calcular una distancia entre los objetos, buscar el gráfico bidimensional, la representación plana en dos dimensiones que tenga la mejor adecuación. Es decir que si dos puntos son cercanos según la distancia, tendrán que estar cercanos en la representación plana. Esta es la idea de base del Escalamiento Multidimensional. Los ejes en este caso no tienen un sentido bien preciso, mientras que en el gráfico bidimensional cuando se tenían dos variables, cada eje correspondía a una variable, con este método, los ejes tienen una representación que no es tan simple.

Fig.24



El segundo enfoque es el enfoque factorial, consiste en buscar una combinación lineal de las variables que tenga la mejor adecuación posible con las variables iniciales y los individuos.

Si la variable expresa la edad, y es una variable que está descompuesta en muchos intervalos: joven, menos joven, se pueden representar los más jóvenes y los más viejos y luego si se tiene otra variable, se puede representar y analizar si hay una correlación entre las dos variables, si evolucionan de la misma manera hasta un cierto punto. Se pueden ver los niveles de salario y las edades, y esta relación se observa en el plano y da ideas.

En el Análisis Factorial en Componentes Principales se representan individuos, los ejes son combinaciones lineales de las variables originales como en el Análisis de Correspondencias y las variables son direcciones orientadas. Esto es un panorama muy global sobre el AD.

Por supuesto hay muchos métodos, pero se puede decir para concluir, con esta introducción, que el objetivo del AD es una ayuda al descubrimiento de regularidades, de leyes, de tendencias, de modelos.

Con el progreso de la computadora la gente construye matrices de datos cada vez más grandes, en las grandes empresas y ahora también en las pequeñas porque todo el mundo tiene

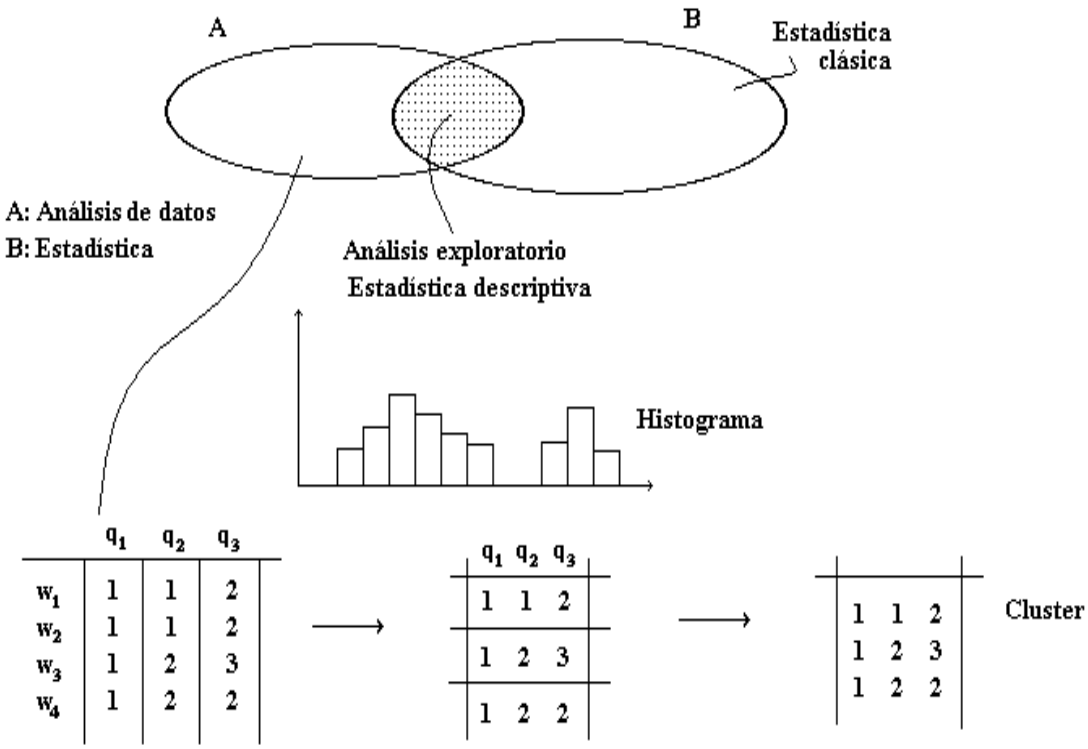
computadoras y cada uno quiere estudiar su clientela, sus regiones, sus productos, sus pacientes, sus especies de plantas, sus medicamentos, las evoluciones epidemiológicas, sus posiciones geográficas, la climatología, psicología, etc.

En todos estos campos en que las leyes no son bien precisas se requieren herramientas para ayudar a descubrir regularidades. El AD es un conjunto de herramientas que están para ayudar a los expertos a descubrir. Tukey, fue el primero que escribió un libro sobre este tema, separa la Estadística en dos vías: el Análisis Exploratorio y el Análisis Confirmatorio. El AD visto desde el punto de vista francés está incluído en el aspecto exploratorio.

¿Se puede decir que el AD está incluido en la Estadística? Está muy discutido. Existe una fuerte intersección pero no hay una inclusión completa.

Si se piensa en un esquema, donde se pueda representar la Estadística Clásica, la Estadística Exploratoria, la Estadística Descriptiva, éste podría ser el siguiente:

Fig.25



Por ejemplo un histograma permite descubrir la existencia de leyes, se pueden detectar uno o varios modos, por lo tanto es una herramienta que es usada tanto por la Estadística Descriptiva Clásica, como por el AD clásico.

¿Que encontramos en la zona de intersección del diagrama?. Se dice a menudo que en el AD no se hace modelo probabilístico o no hay probabilidades, esto no es exacto. Es verdad que en la medida que se quiere hacer análisis exploratorio se trata de reducir al máximo la modelización a priori.

Sin embargo, hay métodos de AD que se combinan completamente con la modelización. Por ejemplo, cuando se busca hacer la descomposición mixta de leyes de probabilidad. Si una población está formada por varias leyes se pueden buscar las clases y las leyes que hagan una mejor adecuación con cada una de las clases. Hay una mezcla de clasificación y de Estadística clásica, es Exploratorio y al mismo tiempo AD porque se construyen clases. No se considera a la población como un todo único sino formada por varias sub-poblaciones, y a menudo éste es el caso que ocurre en la práctica.

Cuando se considera que hay una regresión única sobre toda la población se han puesto a punto métodos basados en las nubes dinámicas que permiten buscar al mismo tiempo clases y regresiones que tienen localmente la mejor adecuación posible con la variable a explicar. Introducir el punto de vista exploratorio en la modelización estadística es más eficaz.

¿Qué se encuentra en el espacio que no es de la Estadística?

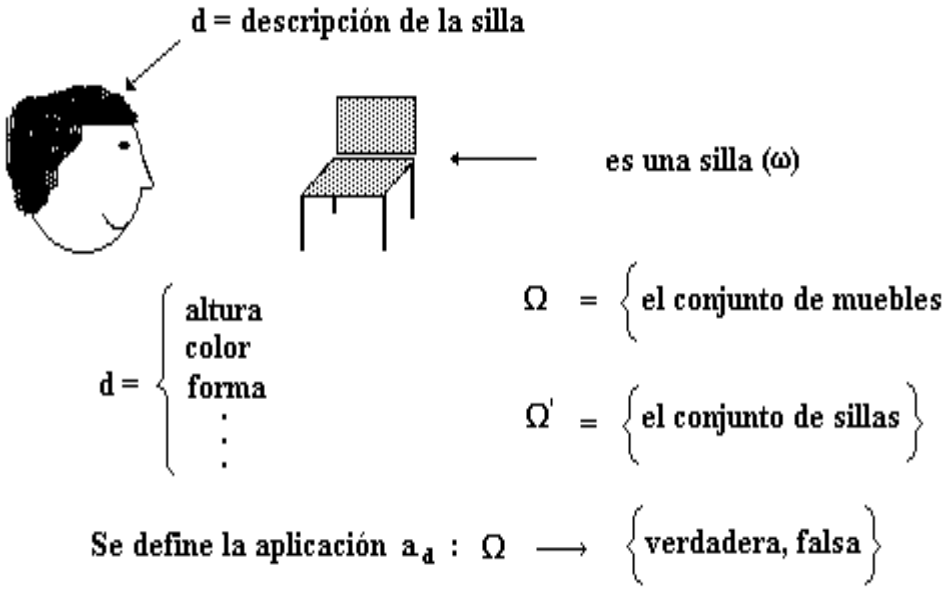
Si se observa la parte del AD que creo que está fuera de la preocupación de la Estadística clásica y se considera, por ejemplo una tabla de datos donde se hallan 4 individuos y preguntas de una encuesta, se puede suponer que dos individuos han respondido de la misma manera. Por lo tanto esos dos individuos los podemos juntar en una sola fila y decir que hay dos. Las otras dos respuestas son únicas. Hacer Estadística es considerar estas frecuencias, pero uno puede también interesarse por el problema que es el AD puro que consiste en olvidar las probabilidades, las frecuencias y no considerar más que los valores, las respuestas obtenidas, y hacer una clasificación olvidando totalmente las frecuencias, simplemente para estudiar las estructuras de respuestas. Esto nos aleja de los problemas de la Estadística clásica ya que no hemos tenido en cuenta la Estadística, no hemos hecho estadística, conteo, sino simplemente la estructura de respuesta, y allí se puede hacer la clasificación o el retículo o calcular las distancias que es muy frecuente en AD y esto interesa mucho menos a la estadística.

Análisis de Datos Simbólico

Una introducción intuitiva de "objeto simbólico"

Si se considera a un hombre que mira una silla para darle un nombre a ese objeto, él utiliza en su cerebro lo que se llama la descripción de la noción de silla. Esta descripción utiliza la altura, el color, la forma de la noción de silla.

Fig.26



O sea a_d toma el objeto la mesa, la mesa es un mueble entonces ### es el conjunto de muebles y Ω' es el conjunto de sillas. Cada mueble, podría haber tomado todos los objetos posibles del mundo, a_d está asociado al valor verdadero o falso para cada objeto del mundo, según este objeto sea una silla o no.

Cómo representar a_d .Hay varias representaciones simbólicas.

a_d= [altura = V_1] ^[color = amarilla, marrón, ...}^...

a_d (ω) = verdadera si altura $_{(\omega)} \in V_1$

color$_{(\omega)} \in$ {amarilla, marrón,...}

Donde V_1 es un intervalo de valores, color puede ser amarillo, marrón, etc., el conjunto de los colores posibles $a_d(\omega)$ es verdadero si y sólo si la altura de ω está en el intervalo V_1 y el color de ω es amarillo, marrón, etc. En general diremos que a_d es un objeto simbólico booleano.

Este es el tipo de objetos que se quiere estudiar, tratar de definir, de construir, que va a formar una base de objetos sobre la que vamos a hacer el AD.

Más generalmente un objeto simbólico es un marco matemático que permite a los especialistas expresar sus conocimientos. Se hablaba antes respecto a la Epidemiología, por ejemplo que los epidemiólogos sabían que los adolescentes que hacían deporte, que hacen actividades culturales en general no se drogan, este perfil, esta clase de adolescentes será descripta como una conjunción de propiedades por lo tanto como una aplicación de Ω en verdadero o falso que será verdadero si el adolescente satisface las propiedades descriptas por el experto. Cada conocimiento del experto se podrá expresar bajo la forma de dato simbólico.

Otro ejemplo: en Brasil, hay un gran programa inglés, de un gran instituto de investigación biológica de Londres, con la Universidad de Recife para construir una base de conocimientos de todas las especies de plantas del nordeste. El objetivo es aplicar esto para encontrar nuevos medicamentos. Cuando se estudian las especies de plantas, se obtiene una gran matriz de datos que no está formada por líneas representando individuos sino por filas representando clases y cuando se representan clases es necesario describirlas en intención. Es bien evidente que este grupo no va a representar las clases con todas las plantas que van a encontrar en el nordeste, no por el conjunto de los individuos sino por la descripción de esos individuos. En otras palabras las clases serán definidas en intención y se obtendrá una matriz de datos muy grande en el que cada línea no será un objeto individual, como en la estadística clásica, sino un objeto simbólico.

Ocurre que este equipo inglés donde uno de sus directores es el profesor Ponkerst considera que es suficiente describir las especies de plantas por conjunciones de propiedades lógicas. Por lo tanto hay aquí un gran dominio de aplicación del enfoque puramente lógico, aristotélico.

Sin embargo si se quiere describir a los objetos de una manera más precisa, por ejemplo incluso en biología, si se interesan por las flores o las plantas van a ver que se describe un cierto tipo de rosa diciendo que los pétalos son de tal especie, más bien rojos, a menudo rojos. En la descripción se siente la necesidad de moderar los valores no se dice siempre rojo, se siente la necesidad de decir más bien rojo, más bien rosa, es posible que sea rojo, o bien es a menudo rojo y si se quiere representar conocimientos se recurre a diferentes tipos de semántica: probabilista o no forzosamente de tipo probabilístico.

En resumen, se puede representar también, con una noción probabilística. En ADS cuando se dice color: amarillo o marrón se puede querer decir a menudo amarillo y raramente marrón. Hay que apoyarse en nociones que son clásicas en Estadística, a pesar de que esta ponderación no es siempre de orden estadístico, como ya se expresó antes: más bien, casi, es posible.

a_d= [color: a menudo amarillo, raramente marrón] ^ ....

a_d: $\Omega \rightarrow$ [0,1]

Esta generalización tiene valores que no son simplemente verdadero o falso sino más bien con un grado de verdadero o falso, pueden también expresarse bajo forma de aplicaciones, ahora no en {verdadero, falso} sino en el intervalo 0,1. El problema va a ser saber cómo calcular a_d (ω). Es necesario que si un individuo satisface completamente esto $a_d(\omega) = 1$; si un individuo no satisface esto $a_d(\omega)= 0$. Con una tendencia hacia 1 si más bien lo satisface y con una tendencia hacia 0 si en general no lo satisface.

Estos objetos se llaman objetos modales porque ellos moderan los valores y tienen varias ventajas.

Cuando se describe una clase por una conjunción de propiedades es muy explicativo para los usuarios, según la tradición aristotélica.

Este tipo de representación permite representar clases monotéticas pero también clases politéticas en la tradición adansoniana, lo que no podía hacerse en la tradición aristotélica porque no daba la posibilidad de representar clases politéticas. Es politética porque un individuo tiene un grado de pertenencia a la clase pero no está completamente fuera o dentro. Por lo tanto esta descripción no es una condición necesaria y suficiente sino que son grados de adecuación.

Finalmente tiene esta tercera ventaja. Este tipo de objetos puede dar prototipos utilizando unos modos asociados a cada suceso. Por ejemplo el modo asociado al color es amarillo, por lo tanto puedo escribir el prototipo como color = amarillo

a_{prot}= [color=amarillo] ^ ...

Entonces aquí se presentó una herramienta que da más facilidades a los expertos para expresar sus conocimientos.

Objetivo del Análisis de Datos Simbólico.

¿Qué problema resuelve el ADS?. Nuevos objetos aparecen en muchos campos: éstos son objetos más complejos, no aptos para ser representados simplemente en una tabla de datos y que están definidos en intención con el objetivo de representar el conocimiento.

¿Es posible extender entonces la Estadística y el AD a este tipo de objetos?.

El objetivo del ADS no es extender el AD clásico a objetos definidos en intención solamente, sino también dar la posibilidad de tratar individuos, no solamente intenciones, sino individuos complejos, más complejos de los que existen en un análisis de datos habitual, tan complejos que no los podemos colocar en un cuadro.

Por ejemplo, si se toman individuos que tienen una determinada imagen, hay una enorme cantidad de elementos en la imagen. Está por supuesto la descripción de cada uno de los objetos que están dentro, que se puede poner bajo forma de filas de un cuadro. Pero además están todas las relaciones entre los objetos. Si hay dos árboles sin hojas que están al norte de árboles con hojas, hay un río que se encuentra entre los árboles y las casas, las casas con techo están al norte de las casas sin techo. Se pueden establecer muchas relaciones que no pueden representarse simplemente en una matriz de datos. Hay que introducir nociones de lógica y esto impide representar este tipo de objetos en una matriz de datos, por lo tanto, el modelo tabular de la estadística clásica, sobre el cual está basado todo el método del AD, no funciona.

Por ejemplo, si se quiere representar a una compañía, a la enfermedad de Tom, una flor, un insecto, un accidente, etc.

Con respecto a los insectos se puede exponer una aplicación que es muy importante. Para estudiar la tercera enfermedad endémica que existe en el mundo que es la lesmaniasis, la cual se transmite a través de un mosquito que se llama flebotoma, presente en el sur de Francia, en Brasil, en América Latina. Existen 70 especies de flebotoma y se trata de distinguir las que son vectores de la enfermedad de las que no lo son. Cabe destacar que la descripción de cada uno de estos insectos es muy compleja.

Se puede hacer la extensión a individuos más complejos, o sea la extensión del AD a objetos definidos en intención.

Un objeto puede ser un conjunto de documentos que proviene de una compañía. Existe una empresa que quisiera clasificar automáticamente todos los documentos que vienen de la IBM o de Bull, etc. Hay que tener entonces una descripción de todos los documentos que vienen de IBM, por ejemplo, por lo tanto el objeto simbólico aquí representa una clase de documentos y será definido en intención.

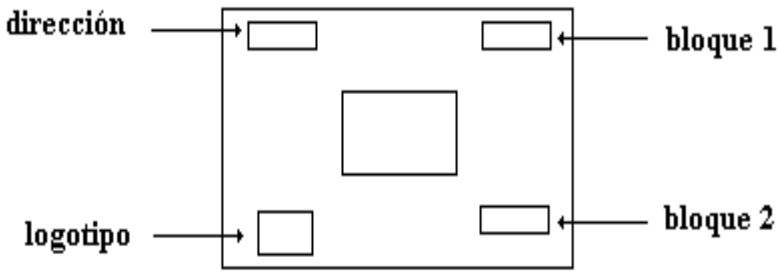
Otro ejemplo sería: en un gran supermercado se requiere estudiar a los clientes, los que vienen los domingos y los miércoles, los que compran tal producto y/o tal otro producto, cada clase de cliente será descripta en intención.

En Medicina, se ha realizado un estudio sobre el conjunto de enfermedades de la sordera, cada enfermedad se describe en intención. Se puede también describir una especie de insecto, de flores, o un guión de accidente.

Volviendo al problema de la clasificación de documentos, es uno de los campos más prometedores, el que tiene más perspectivas de aplicación. Es el problema de la documentación dirigida a población de interés. Por ejemplo, se quiere distribuir publicidad solamente a personas que tienen determinado perfil, los que poseen e-mail, podrían interesarles que estos mensajes estuvieran clasificados y recibir sólo los que corresponden a ese perfil; la biblioteca ha recibido un documento que puede interesar y automáticamente que le avisen al usuario. Este es un dominio que es muy importante y que va a tener más importancia porque el mundo está abarrotado de información.

El problema de la clasificación de los documentos es un campo de aplicación del ADS.

Por ejemplo: si se toma un documento en bloques se puede representar este documento por un objeto definido en intención.



En inteligencia artificial se llama un objeto estructurado. Este documento se describe por las variables: el tipo del bloque 1 que es una fecha, el tipo del bloque 2 que es una firma, y la posición del bloque 1 con respecto al bloque 2 que es al norte.

ω_i = [tipo ($bloc_1$) = fecha] ^ [tipo ($bloc_2$) = firma]

^ [posición ($bloc_1$, $bloc_2$) = Norte]

Aquí el AD no hubiera podido trabajar. Es muy difícil para representar en un matriz de datos. Teniendo 200 objetos así, ¿cómo hacer para representarlos en una matriz de datos?.

Esto es para representar un individuo, se observa que ya para un individuo se da algo que es mucho más general que para un objeto clásico. Si ahora hubiera que describir una clase, también se plantearía el problema del AD de objetos así. ¿Qué es lo que lo diferencia del objeto anterior? Es que los valores relevados pueden ser valores múltiples porque tenemos una clase y no un sólo documento. Por ejemplo, si ω_i representa la clase de los documentos que vienen de IBM puede ser que algunos tengan el bloque 1 que represente ya sea la fecha o un logo, el bloque 2 que represente una firma, la posición del bloque 1 con respecto al bloque 2 será el norte o el este.

ω_1 = [tipo ($bloc_1$) = fecha, logotipo] ^

[tipo ($bloc_2$) = firma] ^ [posición ($bloc_1$, $bloc_2$) = Norte o Este]

En el caso anterior se presentó la representación aristotélica, clases monotéticas, predicado lógico, objetos booleanos. A continuación se expresará en una tradición más adansoniana, clases politéticas. En lugar de decir o bien una fecha o bien un logo, se debe expresar a menudo una fecha y rara vez un logo. La posición del bloque 1 con respecto al bloque 2 es ya sea el norte, o el este y la longitud del bloque 1 es una ley normal de desvío standard sigma.

ω_I = [tipo ($bloc_1$) = a menudo la fecha, rara vez el logo]

^ [posición ($bloc_1$, $bloc_2$) = Norte, Este] ^ [longitud ($bloc_1$) = q_1]

De esta manera se presentaron los tres tipos de representación de conocimientos.

Del Análisis de Datos al Análisis de Datos Simbólico.

En el AD clásico se estudian objetos individuales representando datos que toman un sólo valor en cada casilla del cuadro. En ADS se estudian conjuntos de objetos de más alto nivel desde el punto de vista de las variables, donde los objetos simbólicos son los individuos a estudiar.

He aquí un ejemplo de datos a procesar: las propiedades y los problemas para los objetos simbólicos difieren de los que están propuestos para los objetos individuales.

¿Por qué este nuevo campo?, ¿de dónde viene?. En el INRIA, que es un Instituto de Informática, se encuentran en auge los lenguajes orientados a objetos, las bases de datos orientadas a objetos, las bases de conocimientos. Los usuarios que antes tenían bases de datos relacionales hacían estadística clásica sobre esto, ahora tienen bases de objetos complejos, bases de objetos definidos en intención y en consecuencia va tomando cada vez más importancia la necesidad de tratar de entrada datos más complejos conteniendo conocimientos en lugar de simples observaciones.

En Estadística aparece una nueva tendencia, la gente habla de los meta-datos o sea datos sobre los datos, si se calculan medias, desvíos standard, se tienen nuevos datos sobre los datos.

Más que tener millones de datos individuales puede ser interesante conservar sólo los datos sobre los datos. Es lo que hacen los investigadores, y después van a estudiar lo que han guardado, que son los datos sobre los datos. Datos definidos en intención que describen clases.

Un experto que describe los frutos que se cultivan en un pueblo dice que el peso de ellos se encuentra entre 300 y 400 gramos, el color es blanco o rojo y si el color es blanco entonces el peso es menor de 350 gramos.

Se ve que representar esta frase en una matriz de datos no es fácil, en este caso las filas representarían los pueblos y las columnas las frutas. Esto sería difícil porque en cada casilla del cuadro para el peso tendríamos varios valores, entre 300 y 400g. y también porque no se representarían las reglas. Es mucho más simple representar esta información mediante objetos.

Si los frutos producidos por este pueblo están representados por la aplicación Ω verdadero o falso, esto es un objeto simbólico cuya descripción es: el peso está comprendido entre 300 y 400, el color es rojo o blanco y si el color es blanco, entonces el peso es menor o igual a 350.

a_i = [peso = [300, 400]] ^ [color = { rojo, blanco}] ^

[si [color = blanco] entonces [peso $\leq$ 350]]

Si se tiene un conjunto de 1000 pueblos representados por 1000 objetos simbólicos a_1 ... a_{1000}. ¿Cómo aplicar los métodos estadísticos? ¿cómo hacer un histograma? y ¿cómo aplicar un método probabilístico?.

Existen cinco tesis docorales que han sido defendidas sobre este tema: De Carvalho es un profesor de la Universidad de Recife, Brito es una profesora de la Universidad de Porto (son dos extranjeros que han pasado cuatro años en Francia), Jacques Lebbe y Vignes son profesores en la Universidad de París VI y Sebbagh es profesor en la Escuela Politécnica de París.

También hay cuatro tesis doctorales que están a punto de ser defendidas que son más industriales: Catherine Jacq de la Thomson que es una fábrica que trabaja sobre diferentes tipos de radares, tratando de solucionar problemas militares; Niquil que trabaja en Alcathel-Alstom sobre problemas de calidad de materiales, se hacen árboles y se trata de describir zonas sobre las que no se sabe nada; N. Carruyt trabaja en el Museo de Historia Natural sobre el problema de las esponjas marinas; E.Auriel trabaja en Thomson sobre problemas militares y hay cinco tesis más que acaban de comenzar.

SEGUNDA CONFERENCIA

Introducción. (Síntesis de la 1ra. Conferencia)

El Análisis de Datos (A.D.) comienza por una matriz de datos en la cual hay muchos números, individuos, objetos en las filas caracterizados por variables. Extraer información de una matriz de este tipo es el objetivo del AD. y se sintetiza en reemplazar números por "conocimientos" nuevos.

Los dos métodos ya expuestos, son los métodos de Clasificación y el Método de Análisis Factorial. Cuando existe una gran matriz de números y cifras, el Análisis Factorial permite organizar individuos en un plano y ver de una vez cuáles son los puntos que se parecen, los que se oponen, extraer ejes factoriales que tienen sentido para expresar el tiempo, los perfiles, la distinta gravedad de una enfermedad, etc., conceptos que no estaban visibles, a priori, por sí mismos.

Si los ejes tienen una variancia suficiente, se puede llevar sobre el plano factorial clases de objetos que se obtienen por métodos de clasificación. Por ejemplo, por jerarquía, pirámides, nubes dinámicas, o por métodos de K-medias que dan particiones.

Una partición puede ser llevada sobre el plano factorial para ayudar a interpretarlo. No hubo tiempo de desarrollar "árboles" pero también es una técnica que permite ver cómo los individuos se sitúan unos en relación a otros.

En el caso de la jerarquía los individuos se hallan en la base, como también en el caso de las pirámides. Las clases de individuos están representadas por niveles, por el contrario en el caso de un árbol los individuos son los nudos y allí vemos racimos que son individuos que se parecen.

La Clasificación y el Análisis Factorial constituyen la base de todo conocimiento antes del análisis estadístico.

Benzècri que es uno de los fundadores del AD. siempre dice que el modelo debe provenir de los datos y no a la inversa. El AD. ayuda a encontrar el modelo.

El Análisis de Datos Simbólicos (ADS) tiene como objetivo reemplazar los individuos del Análisis de Datos tradicional por individuos de más alto nivel, más complejos y más aptos para representar conocimientos, porque están definidos en intención, utilizando el poder de la lógica, son los objetos simbólicos (OS).

¿Cómo hacer un Análisis Factorial cuando los puntos son objetos simbólicos y no solamente puntos del espacio de p-dimensiones como es clásico en el A.D. y en la estadística?

También se puede decir que las variables son de más alto nivel en el análisis de datos simbólicos, porque las variables no van a tomar un sólo valor por cada celda, sino que pueden tomar varios valores.

Por ejemplo: cuando se describe una clase, los individuos de la clase pueden tomar distintos colores. Si se describe una clase de empresas, que tienen beneficios de distinto orden, se puede tomar el beneficio en intervalos para esta clase de empresas. Si un individuo $###_1$ pertenece a una clase de empresa, en la variable beneficio tendré un intervalo de valores correspondiente al conjunto de beneficios hechos por los individuos de esta clase.

Uno puede decir que los individuos y las variables son de mayor nivel que en la estadística y el A.D. clásicos. Es muy importante porque va a plantear todos los problemas teóricos también a un segundo nivel o a mayor nivel, se va a subir un nivel toda la teoría del A.D. clásico.

En la estadística clásica, cuando se define una variable aleatoria, por ejemplo la variable aleatoria x, toma valores en tal intervalo y en este caso decimos que cuando se escribe:

$x = [3, 5]$

se dice que este suceso correspondiente a esta variable aleatoria expresa un conjunto de individuos en el espacio Ω.

El objetivo es estudiar las leyes asociadas a estas variables aleatorias. Estas leyes rigen sobre los individuos Ω. Donde Ω es el espacio muestral de los valores obtenidos.

En el Análisis de Datos Simbólicos Ω no es un conjunto de valores tomados por los individuos sino es el conjunto de eventos o sucesos. Así es como se sube un grado en la teoría.

El objetivo de esta conferencia es mostrar que los axiomas de Kolmogorov que son la base de la teoría de la probabilidad van a tener que subir un grado. Porque en lugar de aplicarse sobre conjuntos de individuos del espacio muestral, el Teorema de Kolmogorov se aplicará a los eventos o sucesos y no sobre las clases de individuos.

En el caso de la clasificación ocurre lo mismo, en lugar de tener individuos se tendrán objetos simbólicos que podrán ser considerados como eventos. Las clases podrán representar dos cosas como en clasificación clásica: simplemente un conjunto, pero también una intención. Cuando se habla de clase hay dos cosas: el conjunto y lo que permite describirla.

Por supuesto los objetos simbólicos contienen como caso particular individuos, puntos de R_p.

Tenemos por una parte objetos que son O.S. pero además las clases también están representadas por objetos simbólicos.

Por lo tanto tenemos que poder explicitarlo y cada nivel estará descripto por un conjunto de propiedades, de manera que el usuario, apoyándose en los nudos o nodos podrá ver las propiedades asociadas a este nodo.

Estas propiedades asociadas a cada nodo son propiedades para esta tesis de tipo monotético. Es decir que la extensión de las propiedades asociadas a un nodo son el conjunto de puntos que están allí y la intención de estos puntos son las propiedades asociadas a dicho nodo.

Por lo tanto las propiedades asociadas a cada nodo son las condiciones necesarias y suficientes que caracterizan las clases que lo sostienen.

La necesidad de la incertidumbre

En ciertos campos la representación booleana del conocimiento es suficiente. Una variable aleatoria en estadística clásica es una información de orden booleano. Cuando se escribe x = [3,5] son todos los individuos que satisfacen las condiciones del intervalo.

Es booleano, se está o no dentro del intervalo. Pero en muchos casos es necesario hacer intervenir la incertidumbre, por ejemplo, uno puede decir en el pueblo i el color de las frutas es a menudo roja y rara vez blanca. Un evento que se observa en estadística clásica donde dice evento color es igual rojo o blanco ahora es expresado de manera incierta.

Donde se tenía la aplicación de Ω, verdadero o falso. (Ω = [0,1]), la aseveración "color: igual a rojo o blanco", ahora pasa a ser:

a_i = [color = 0.9 rojo, 0.1 blanco]

Se puede generalizar en el caso booleano y en el caso incierto por:

a_i = [color = q_i]

donde q_i es la función característica de los valores tomados por el color por ejemplo, en el caso de un booleano y una medida de probabilidad en el segundo caso.

En el caso incierto o probabilístico por ejemplo color igual a q_i será una medida de probabilidad. Si en el caso booleano tengo:

a_i = [color = {rojo, blanco}] se tiene que

q_i = (rojo) = q_i (blanco) = 1 y q_i = 0 para los otros colores.

Por el contrario en el caso probabilístico tendré

q_i (rojo) = 0,9 y q_i (blanco) = 0,1. Pero en los dos casos puedo escribir

a = î [y_i = q_i]

En el caso booleano, q_i es una función característica y en el caso probabilístico q_i es una medida de probabilidad y de ahí la idea de ver si existen otros tipos de funciones que puedan expresar otro tipo de conocimiento para utilizar

Porque el objetivo del ADS es permitir que los conocimientos de los expertos sean expresados en los datos mismos y por lo tanto encontrar una expresión matemática que permita transformar las frases que expresan experiencia en forma de datos. Pero estos datos son de más alto nivel.

En el momento en que los conocimientos se expresan, es evidente que no son probabilidades, cuando uno habla utiliza otras nociones, como "a menudo", "frecuentemente", "rara vez".

Se da la posibilidad a los expertos de transmitir otras nociones que las de las funciones q_i y salir del caso booleano, otra cosa diferente de las funciones que expresan solamente frecuencia de probabilidades objetivas en forma de frecuencia, o incluso de probabilidades subjetivas, en los dos casos satisfaciendo los axiomas de Kolmogorov, que son los axiomas básicos de probabilidad.

Por lo tanto se van a usar funciones características y también funciones probabilísticas. Además existen otros tipos de conocimiento posibilístico que están unidos a la teoría de conjuntos borrosos de Zadeh. Asimismo hay otras teorías que son la teoría de las creencias o de las evidencias, la teoría de la posibilidad que dan otros axiomas diferentes. Podremos tener objetos probabilísticos, booleanos, posibilísticos y objetos credibilistas o de creencias. [82]

La probabilidad en el caso de la creencia o de las evidencias se escribe de esta forma :

a_i = [color = 0.4 { rojo, rosa}, 0.3 { blanco, gris}, 0.3 {0}]

La ignorancia es un caso particular de la teoría de la evidencia.

Otros tipos de semántica

Más precisamente estas teorías expresan diferentes tipos de semántica.

Por ejemplo en el caso de los documentos. Éstos están caracterizados por palabras claves que se hallan presentes en cada texto. Se pueden calcular las tres medidas siguientes:

1) Probabilidades:

Se tendría una tabla de datos con las palabras claves del texto 1, 2, etc. y el número de ellos puede representar la frecuencia. Si uno utiliza la frecuencia, ésta debe cumplir el axioma de Kolmogorov pudiéndose usar métodos basados en este axioma.

$Pr (E_1 \cup E_2) = Pr (E_1) + Pr (E_2) - Pr (E_1 \cap pr\ E_2)$

2) Posibilidades:

Supongamos ahora un texto que hable de matemáticas y no use nunca la palabra matemática. En el caso de la probabilidad frecuentista, la palabra matemática tendría un valor nulo. Es falso, pues un experto sabe muy bien que el texto se refiere a matemáticas o cosas que se hallan cercanas a matemáticas.

Por lo tanto se puede hacer intervenir la noción de posibilidad, de un experto que dice que tal palabra no apareció pero tiene una alta posibilidad de poder aparecer.

Si se observa la noción de posibilidad, el axioma de Kolmogorov es reemplazado por este axioma que ha sido propuesto por Zadeh que reemplaza el axioma de Kolmogorov y en el que se basa o constituye la base de la teoría de la posibilidad. La teoría de los conjuntos borrosos.

$Pos (E_1 \cup E_2) = Max (Pos (E_1), Pos (E_2))$

3) Creencias:

Una tercera noción, puede intervenir de la siguiente manera, un experto puede decir "pienso que esta palabra podría haber aparecido y el argumento que tengo para pensarlo me hace decir que este punto tiene la probabilidad p_1 de aparecer y p_2 de no aparecer".

Si $p_1 + p_2 < 1$ entonces

$1 - (p_1 + p_2)$ expresa su ignorancia

Y en este momento tienen estos axiomas que reemplazar al axioma de Kolmogorov.

$Bel (E_1 \cup E_2) \geq Bel (E_1) + Bel (E_2) - Bel (E_1 \cap E_2)$

La regla de Dempster permite la posibilidad de combinar la creencia de varios expertos.

Para la teoría de la creencia se puede consultar Schafer (72).

Una manera de intuir la noción de creencia en un ejemplo simple es cuando uno piensa que va a llover, por ejemplo 0.3 posibilidad, pero también hay sol y tengo 0.4 posibilidad de que no va a llover. El resto se expresa en ignorancia.

¿ Qué es un objeto simbólico?

En ADS en lugar de tener un conjunto de individuos, tenemos un conjunto de objetos simbólicos que están expresados por un conjunto de propiedades, donde cada propiedad puede ser del tipo probabilístico, booleano, posibilístico o de otra noción. De esta forma un experto con todo esto puede decir muchas cosas.

Objetos booleanos:

$a(w) = \hat{i} [y_i (w) \in V_i]$

$a: \Omega \rightarrow \{0,1\}: a(w) = \hat{i}[y_i (W) \in V_i$

$a(.) = \hat{i} [y_i (.) \in V_i]$ llamado $a = \hat{i} [y_i = V_i]$

Objetos de creencia:

$a_{Bel} = \hat{i}[y_i = q_i]$; donde q_i es una función de creencia

$a_{Bel} : \Omega \rightarrow [0,1]$

Resumiendo. Dependiendo de la elección de la función q_i utilizada a_i será un objeto booleano, probabilístico, posibilístico o de creencia. En todos estos casos a_i es una aplicación de Ω (el conjunto de frutas, por ejemplo) en [0, 1] y el problema crucial es el siguiente. ¿Cómo se calcula a_i (ω) ?

a_i es un objeto simbólico, por ejemplo una especie de flor y cuando tengo una flor ω ¿cuál es el valor de a (ω)?

Si a(ω) es grande, se dirá que la flor pertenece a la clase que la describe, si a(ω) es pequeño no es el caso.

Si uno dice

a = [color = q_1] y [altura = q_2]

y si tomo un individuo

ω = [color = rojo] y [altura = 10]

¿Cómo se calcula a(ω) ?

Este problema está basado en la elección de dos funciones, una función f_x y g_x.

En este caso me referiré a f_{bel} y g_{bel} (funciones f y g en el entorno credibilístico).

Por ejemplo : si poseo un objeto de creencia que es igual a una conjunción de funciones de creencias, y si tengo un individuo (ω), que también es igual a una conjunción de funciones de creencias. El valor de ω para a va a ser calculado con dos funciones: una función f y otra g. La función g mide la adecuación entre r y q y la función f combina los diferentes eventos:

$\omega = \hat{i}_{bel} [yi = ri]$

entonces $a_{bel}(\omega) = f_{bel}(\{g_{bel}(qi, ri)\}_i)$

Herramientas del Análisis Simbólico

Para trabajar con objetos simbólicos necesitamos un número de herramientas,ellas son:

a) la manera de calcular o expresar las extensiones.

Extensión: $Ext_\alpha (a/\Omega) = |a|_\Omega$

Por ejemplo: el análisis del problema de búsqueda de regiones por contigüidad que satisfagan ciertas condiciones: el tamaño, el número de personas, las variancias y sus parámetros demográficos. Se trataba de buscar regiones en Argentina de manera de reagrupar departamentos

que posean un mínimo de personas, que éstos sean contiguos y que ciertos parámetros demográficos sean de variancia mínima en estas clases. Esto es una conjunción de propiedades. El problema luego es encontrar la extensión de esta intención.

En el ejemplo de relatos de tipos de accidentes, habiendo descripto un relato de accidente a y conociendo un accidente ω, cómo se calcula a (ω) que mide en cuánto un accidente satisface a ese relato de accidentes.

Ese tipo o relato de accidentes es un objeto simbólico .

b) La generalización*:* $a_1 \cup_x a_2$

Si se realiza una jerarquía y tengo dos objetos simbólicos, cómo voy a pasar a la clase. Esta clase debería ser un objeto simbólico que generaliza de la mejor manera a estos dos objetos.

Como hablamos de conocimientos, de conceptos, la generalización es una noción que se puede comprender muy bien intuitivamente.

Si decimos mueble, el mismo generaliza mesa y silla. Uno puede decir que mamífero generaliza a especie de animales que tienen mama.

Se trata de definir matemáticamente esta noción teniendo un conjunto de objetos simbólicos, por ejemplo, que describen la producción de frutas de los pueblos de un departamento. Cómo describir las frutas producidas por este departamento.

c) La especialización*:* $a_1 \cap_x a_2$

De manera complementaria se necesita definir especializaciones, por lo tanto operadores de especificación, cómo describir para un objeto simbólico lo que es común a la producción de frutas de varios departamentos descriptos por objetos simbólicos.

Los operadores se deberán definir absolutamente en los cuatro casos: booleano, probabilístico, posibilístico y de creencias.

Propiedades de los objetos simbólicos

La *noción de extensión.* La extensión en Ω de nivel α de una aseveración es el conjunto de ω que pertenece a Ω tal que $a(\omega) \geq \alpha$

$$\text{Ext.}\ (a / \alpha) = \{\ \omega \in \Omega / a\ (\omega) \geq \alpha\ \}$$

A partir del momento en que se define esta extensión, se puede definir un orden parcial diciendo que $a_1 \leq_\alpha a_2$ si la extensión a_1 de nivel α está contenida en la extensión de a_2 nivel α :

$$a_1 \leq_\alpha a_2 \text{ si } \text{Ext}_\alpha\ (\ a_1 / \Omega\ ) \subseteq \text{Ext}\ (\ a_2 / \Omega\ )$$

A partir de aquí tenemos un teorema que dice que "Con este orden al nivel α y con una semántica dada por x , el conjunto de objetos simbólicos es un retículo"

Existen ciertas condiciones que son demasiado técnicas y no se explicitarán aquí.

Para las cuatro teorías de las que se habla en el artículo [82] se puede decir que el conjunto de los objetos simbólicos está organizado según un retículo.

De manera que subiendo en el retículo se generaliza y descendiendo se especializa.

Casos particulares de retículos son justamente las jerarquías y las pirámides. Son semi-retículos por otra parte.

Los tres niveles de datos

Voy a volver a la noción de dato. Resumiendo hay tres niveles de datos.

El nivel 0: por ejemplo en el caso que tenemos n manzanas que están caracterizadas por su altura, color y tenemos n peras caracterizadas también por su altura, color, etc.

Nivel 1: ahora, a estos datos clásicos se los pasa al nivel de los datos sobre los datos, es decir meta-datos. Y teniendo 500 manzanas y 300 peras se puede definir meta-datos sobre esto. Por ejemplo puedo hacer la suma. Se puede decir cuántas frutas promedio hay por clase.

Utilicé la palabra fruta que es una palabra que no está ni en manzana ni en pera. A partir del momento en que se hace una media aquí necesitamos una palabra que generalice a manzana y a pera. La palabra que encontramos es fruta.

Pero esta palabra fruta no tiene existencia en estadística. Mientras que en el ADS las 500 manzanas no es solamente un conjunto, además es una descripción. La clase de 500 manzanas puede ser un objeto simbólico llamado manzana cuya altura seguiría una ley de probabilidad, el color otra ley de probabilidad, etc.

Se hace entonces una especie de descomposición simbólica de una ley de probabilidad multidimensional.

Las 300 peras también pueden ser descritas por una ley de probabilidad para la altura, el color, etc.

La palabra fruta no tiene un sentido particular, yo sé solamente que tengo 400 frutas en promedio. Frutas está representado como un ser matemático real que generaliza la noción de manzana y pera de esta clase de manzanas y de esta clase de peras. Generaliza las dos leyes de probabilidad por medio de un operador de unión generalizador.

¿Cómo definir la noción de las dos leyes de probabilidad?

Esto no es nuevo. Bernoulli ya lo propuso en el siglo XVII.

En el nivel 2: uno puede definir y asociar un peso a cada clase definida aquí.

$\text{frutas}_{mp}(\text{manzana}_i)$ = especie de probabilidad

Si se calculó la fruta de la clase de manzana i esto será una especie de probabilidad, si manzana i es una manzana que puedo tomar en mi mano. Esta fruta es una aplicación de:

$\omega = [0, 1]$

Si ahora tomo varias manzanas para tener las "frutas" de estas manzanas voy a hacer una unión de otras manzanas.

$\text{frutas}_{mp}(\text{manzana}_i \cup_{pr} \text{manzana}_j \cup_{pr} \text{pera}_i \ldots)$

Aquí aparece otra noción: fruta no va a estar definido en Ω. Va a estar definido sobre la clase Ω.

Si yo defino fruta* que está definido en una clase de objetos simbólicos fruta* ¿podrá ser también definida con una ley de probabilidad? ¿Esto se cumplirá?

$\text{fruta}^*_{mp}(A_1 \cup_{pr} A_2) = \text{fruta}^*_{mp}(A_1) + \text{fruta}^*_{mp}(A_2) - \text{fruta}^*_{mp}(A_1 \cap_{pr} A_2)$

En otras palabras. Tenemos el axioma de Kolmogorov y será un principio de la elección de las funciones f y g de las que hablé anteriormente.

Nos interesa elegirla. De manera de que en el nivel 2 el axioma de Kolmogorov resista, se mantenga.

Se puede decir lo mismo de otra manera para que se entienda mejor.

. En el nivel 0: tenemos puntos que son individuos en un conjunto clásico (Ω). En este conjunto clásico soy capaz de definir unión, intersección, complemento, conjunción, disyunción, etc.

. En el nivel 1: mis puntos no son individuos, sino objetos simbólicos. El conjunto a será el conjunto de las q_i (probabilidades, posibilidades, creencias).

Los objetos simbólicos expresan clases y estas clases satisfacen los axiomas clásicos por ejemplo el axioma de Kolmogorov:

$pr(A \cup B) = pr(A) + pr(B) - pr(A \cap B)$

El axioma de Zadeh:

$pos(A \cup B) = Max(pos(A), pos(B))$

y el axioma de Dempster-Schafer:

$bel(A \cup B) \geq bel(A) + bel(B) - bel(A \cap B)$

Son capaces de definir ponderación de las clases sobre el conjunto clásico.

En el nivel 2: meta-teoría.

Si mis individuos son objetos simbólicos que son definidos por funciones, $q_{i,}$ cada individuo, cada punto, será una conjunción de leyes de probabilidad, por ejemplo leyes de probabilidad o de creencias y podré definir una especie de meta-teoría donde las clases serían objetos simbólicos. La unión son uniones de objetos simbólicos. No uniones entre conjuntos clásicos y voy a tener que elegir las funciones f y g de manera que los tres axiomas, igualdades o desigualdades sean satisfechos.

$pr^* (A \cup_{pr} B) = pr^* (A) + pr^* (B) - pr^* (A \cap_{pr} B)$

$pos^* (A \cup_{pos} B) = Max (pos^* (A), pos^* (B))$

$bel^* (A \cup_{bel} B) \geq bel^* (A) + bel^* (B) - bel^* (A \cap_{bel} B)$

Generalización

Hay muchas técnicas para la generalización. La gente que se dedica a inteligencia artificial conoce muy bien este tema porque han trabajado mucho sobre estas nociones.

Por ejemplo si tengo una conjunción de propiedades como éstas:

$a = [y_1 = V_1] \wedge [y_2 = V_2]$

Suprimiendo una de ellas obtengo un objeto simbólico más general.

Si digo la silla que es marrón que tiene un tamaño superior a tres y que tiene por lo menos tres patas si quito el color marrón tendré inmediatamente más posibilidades. Tendré un objeto más general.

Podemos también generalizar agregando valores, por ejemplo si agrego el valor dos doy más posibilidades y generalizo también:

$a = [y_1 = 1, 2, 3] \wedge [y_2 = 1, 2]$

Otra manera de generalizar es utilizando una taxonomía. Por ejemplo, al decir que el color gris, rosa o blanco, puede ser reemplazado por un color claro y si tengo entonces el objeto rosa o blanco, pasando a la palabra *claro* tendré un objeto más general. Porque existirá la posibilidad de que sea gris.

Esto se ve en el Análisis de Datos Textuales que permite pasar de una palabra a otra palabra, realizar clases y generalizar.

Por ejemplo se puede pasar de *física, química* o *matemáticas* a otra palabra *científica* y *psicología, letras* por *ciencias humanas*. Tenemos entonces palabras claves y después una especie de árbol que se generaliza con otras palabras.

Cuando hacemos clases podemos reemplazar ciertas palabras por otras más generales que simplifican la lectura.

También se puede generalizar pasando de una constante a una variable:

a = [edad (Pablo) = 20] → edad (x)

Asimismo, en lugar de generalizar globalmente se puede generalizar de manera más fina reagrupando simplemente los puntos que se acercan más.

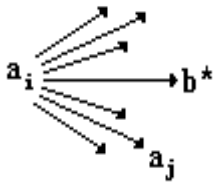
Ajuste a un conjunto de objetos simbólicos:

El siguiente es un problema interesante que se parece un poco a los problemas de estadística clásica.

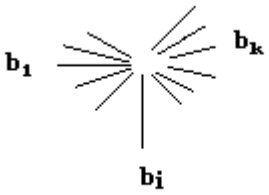
Si uno representa varios objetos simbólicos por medio de vectores, uno puede buscar el vector b^* tal que $b^*(a)$ sea mínimo y tratamos de maximizar el mínimo de $b^*(a)$.

$$A = \{a_1, \ldots a_m\}\ b_1^*, \ldots, b_k^*$$
$$b^* = \text{Arg}\ \underset{b}{\text{Max}}\ \underset{a \in A}{\text{Min}}\ b^*(a)$$

Podemos plantear esto globalmente y también localmente. Tenemos un problema de este tipo.



b* = Σ pi bi*



Teoría de la Evidencia

La teoría de la evidencia fue creada por Schafer en 1976 [71].

Se puede hablar también de Dempster [22] que ha escrito un artículo que da las primeras ideas.

Choquet [18] que desarrolló la teoría de la capacidad.

Matheron (1975) que hizo muy buen trabajo sobre la teoría de los conjuntos aleatorios. Ya en este título se siente el espíritu del Objeto Simbólico ya que trabaja sobre conjuntos que considera como aleatorios.

En él se ve que hay un grado superior con respecto a la teoría de la Estadística clásica.

Nociones de base

Función de Probabilidad:

Se define una función m de P (O) en [0,1] tal que asigna valor nulo al conjunto vacío y tal que la sumatoria de m sobre las partes de O es igual a la unidad.

$m(\varnothing) = 0$

$\Sigma \{ m(A) / A \in P(O)\} = 1$

Función de creencia:

Se define una asignación básica de probabilidad m de P(O) en [0,1],

$F = \{A_i \in P(O) / m(A_i) = 1\}$ (finito), se dice que el par (F,m) es un cuerpo de evidencia.

Sea (F,m) un cuerpo de evidencia, una función de creencia se define de P(O) en [0,1] como: bel: $P(O) \rightarrow [0,1]$

$bel(A) = \Sigma_i \{ m(A_i) / A_i \in F, A_i \subseteq A$

Sea una parte de O llamada A, la creencia de A será la sumatoria de las ponderaciones asociadas a cada parte de P(O) que está en A

Función de plausibilidad:

La plausibilidad es la suma de todas las partes que están en la intersección no vacía de A, que define plausibilidad de A.

Si hay intersección no vacía de A, lo tendremos en cuenta para efectuar la suma.

Pl: $P(O) \rightarrow [0,1]$

$Pl(A) = \Sigma_i \{ m(A_i) / A_i \in F, A_i \cap A \neq \varnothing\} \qquad A \in P(O)$

$bel(A) = 1 - Pl(A^c)$

$$a = \hat{i}[y_i = bel_i]$$

Si se definió la plausibilidad y la creencia se puede definir un objeto simbólico, donde aquí hay una función de creencia o de plausibilidad.

Semántica de la creencia

Tenemos una teoría que explica ciertos tipos de conocimientos. (Podría haber hecho lo mismo con las probabilidades por ejemplo). ¿Cuál es el sentido de esta teoría de la creencia?.

En primer lugar las probabilidades pueden ser conocidas solamente sobre partes.

En la teoría de probabilidades uno conoce la probabilidad sobre los individuos y se deducen las probabilidades sobre las partes, pero hay campos en los que no conocemos las probabilidades más que sobre las partes. Y es en este campo donde se permite que la teoría de las creencias funcione.

No tenemos la obligación de conocer la probabilidad sobre cada individuo.

Es interesante sobre los objetos simbólicos que trabajan ellos sobre las partes.

Segundo aspecto: se puede tener en cuenta la ignorancia.

Tercero, se pueden combinar creencias de varios testigos. Por ejemplo si alguno dice que esto es naranja, yo digo que es amarillo y otro dice rojo, vamos a combinar todas estas creencias y un juez va a deducir de allí ciertas creencias.

Por ejemplo: si dos expertos creen en el mismo escrito A, la fórmula de Schafer da una creencia en A que va a crecer.

Pero si un experto cree en A y otro cree en B y la intersección del evento A y del evento B es vacía, la creencia en el evento A y en el evento B va a decrecer.

Si dos personas no piensan exactamente lo mismo, por ejemplo en que este lápiz puede ser amarillo, el juez va a creer menos a una y a la otra.

Si A y B tienen intersección vacía la creencia en A y en B va a decrecer más, y la creencia en A decrecrá cuando la creencia en B sea más fuerte.

Si yo digo que esto es amarillo y Ud dice que esto es rojo la intersección de los dos eventos es vacía.

Cuanto más fuerte sea su creencia de que esto es rojo, menos creerá otra persona en lo que yo digo, es decir en que es amarillo.

Esto es lo que queremos obtener como resultado y es lo que da la fórmula de Dempster-Schafer.

Hemos definido una teoría en base a axiomas, vimos cuál es el sentido de esa teoría. Porque siempre es posible definir los axiomas pero es necesario que estos axiomas tengan sentido en la realidad. Que represente cierto tipo de conocimiento.

Efectivamente tenemos la impresión de que esto representa bien este tipo de creencias.

Objetos de creencia:

Los axiomas están dados, la semántica está bien definida y ahora vamos a poder definir "objetos de creencia"

De la misma manera se van a poder definir los objetos de probabilidad, objetos posibilísticos, etc.

Una aserción de creencia se escribe como una conjunción de propiedades

$$a_{bel} = \hat{i}_{bel}\left[y_i = \left\{q_i^j\right\}_j\right]$$

$$a_{YQ} = f_{bel}\left(\left\{g_{bel}\left(\cup_{bel} q_i^j,\ \cup_{j\ bel} r_i^j\right)\right\} i\right)$$

Donde cada función aquí es una función llamada de creencia. Es una aplicación a $_{YQ}$ de Ω en [0, 1] tal que si cada individuo ω pertenece a Ω, puedo escribir una función

$$y(\omega) = \left\{r_i^j\right\}$$

que es una función de creencia.

a(ω) está calculada por:

$$a_{YQ}(w) = f_{bel}\left(\left\{g_{bel}\left(\cup_{bel} q_i^j,\ \cup_{j\ bel} r_i^j\right)\right\} i\right)$$

En el caso más simple pide uno sacar las uniones.

g_{bel} mide la adecuación entre q_i y r_i^j, q_i^j. Este último representa el objeto a y r_i^j el individuo ω.

Cómo se definen g y f.

g se define así:

$$g_{bel:}g_{bel}\left(q_i^1, q_i^2\right) = \sum\left\{m_i^1 \cap bel\ m_i^2\left(V_2\right) \middle/ V_2 \subseteq V_1, \left(V_1, V_2\right) \in \ F_1 x\ F_2\right.$$

f toma solamente la media, pero se pueden afectar otras selecciones posibles.

Una pregunta natural se plantea: ¿es posible decir que $a_{bel}(\omega)$ mide una creencia de que ω pertenece a la clase representada por a_{bel}?

a_{bel} define una conjunción de funciones de creencia y nada prueba a priori que a_{bel} sea una función de creencia.

Para poder responder a esta pregunta, debemos extender a_{bel} a a_{bel^*} definida sobre A_{bel} un conjunto de objetos simbólicos, que son objetos credibilísticos, y definir operadores de intersección, unión, complemento, en el conjunto de las funciones de creencia.

De manera general en el ADS, se define primero el objeto a que tiene una función de creencia, luego se define la unión de dos objetos simbólicos haciendo la unión de las funciones que están aquí:

$$a_j = \hat{1}\,[y_i = q_i]$$

$$a_1 \cup_{bel} a_2 = \hat{i}_{bel}\left[y_i = q_i^1 \cup_{bel} q_i^2\right]$$

Luego se define la intersección de las dos afirmaciones o aserciones, $a_1 \cap_{bel} a_2$

El complemento creencia y la conjunción de las dos aserciones. Y también se puede definir una lógica de creencias.(conjunción, disyunción, negación, etc.)

$c_{bel}(a)$

$a_1 \wedge_{bel} a_2$

Hay que también definir $a_{bel}(\omega)$, luego a^*_{bel} definida sobre el conjunto de las aserciones credibilísticas y luego hay que verificar que a^*_{bel} es realmente creencia de creencias.

$$a(\omega) = f_{bel}(\{ g_{bel}(q_{i,} a, q_i, \omega)\})$$

$$a^*(a_i)$$

Vemos que tenemos en primer lugar objetos , en un nivel cero: los conjuntos clásicos que definen eventos. Luego tenemos creencias sobre estos eventos, nivel uno, y luego en el nivel dos tenemos creencias de creencias.

Por lo tanto si uno dice que los conjuntos clásicos representan un nivel de conocimientos de nivel cero, la aserción credibilística o probabilística nivel de conocimientos uno, y ahora la pregunta es saber si a^*_x que representa un nivel dos de conocimientos satisface los axiomas X. Donde X representa ya sean las probabilidades ó posibilidades

En otras palabras si a^*_x es una probabilidad o una creencia respectivamente asociada a los correspondientes operadores OP_x.

El teorema antes citado muestra de que éste es el caso si OP_x y las funciones g y f fueron bien elegidas.

De creencia a convicción

El teorema en el caso de creencia en pocas palabras demuestra de que a^* =1

i) $a^* (a_{bel}) = 1, a^* (\varnothing) = 0$

Y generaliza el axioma de base de la teoría de la evidencia.

Hay otros resultados pero no se desarrollarán.

ii) Siendo a_{bel} el conjunto de las aserciones de creencia. Si $\forall_i A_i \subseteq a_{bel}$ el cuerpo de evidencia de los subconjuntos A_i del a_{bel} son independientes, entonces:

$$a^*\left(\cup_{\substack{Bel \\ i\leq\{1\ldots n\}}} A_i\right) \geq \sum_{i\subseteq\{1\ldots n\}} (-1)^{|I|+1} a^*\left(\cap_{\substack{Bel \\ i\in I}} A_i\right)$$

iii) Si $\forall A \subseteq a_{bel}$ $m^*(A) = \dfrac{a^*_{bel}(A)}{a^*_{bel}(h(A))} \sum_{B\subseteq A} (-1)^{|A-B|}\ a^*_{bel}(h(B))$

donde $h(B) = \cap_{bel} \{ A_i/A_i = A - \{ai\}, ai \in A\backslash B, B\neq A\}$

$h(A) = \cup_{bel}\{ A_i/A_i = A - \{ai\}, a_i \in A\}$

luego m* es una función de asignación de probabilidad sobre a_{bel}

En otras palabras: $m^*: P(a_{bel}) \rightarrow [0,1]$ es tal que $m^*(\varnothing) = 0$, $\Sigma_{A\subseteq abel} m^*(A) = 1$ y

$\forall A \subseteq a_{bel}$ $a^*(A) = \Sigma_{B\subseteq A} m^*(B)$

Usando m* es entonces posible extender la regla y condicionamiento de Dempster al conjunto de aserciones de creencia.

La función m que se definió antes va a ser definida sobre un conjunto de objetos simbólicos y se encontrará que $a^*_1 (a_2)$ puede ser interpretada como una creencia de creencias, que en realidad es una especie de convicción de alguien llamado E_1, que cree en la creencia representada por a_1, de la creencia de algún otro llamado E_2 cuya creencia está representada por a_2

a^*_1 de (a_2) es la creencia de uno en la creencia del otro.

La semántica de a^* en el caso de objetos credibilísticos es entonces la creencia de creencias o la convicción de alguien llamado E_1 con la creencia a_1, en alguien llamado E_2 cuya creencia es a_2

Esto se escribe así:

Para i = 1,2 sea $a_i = [y = q_i]$ donde q_i es una función de creencia $O \rightarrow [0,1]$ con cuerpo de evidencia (F_i, m_i) y

$F_1 = F_2 = \{A, B, O\}$ con $A \cap B = \varnothing$, entonces tenemos que:

$$a_1 * (a_2) = g_{bel}(q_1, q_2) = \sum_{V \in F_1} m_1(V) q_2(V) \Rightarrow$$

$$\Rightarrow a_1 * (a_2) = m_1(A)m_2(A) + m_1(B)m_2(B) + m_1(0) \qquad (1)$$

$m_1(A)$ es la creencia de E_1 en a_1

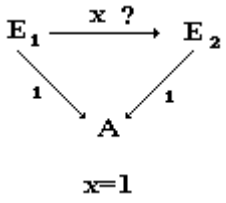
$m_2(A)$ es la creencia de E_2 en a_2

Por lo tanto tenemos 2 expertos: uno cree en A y el otro en B. La creencia de uno en la creencia del otro está dada por la fórmula anterior (1).

Este tema resulta de interés para los militares, y espero que posea otras aplicaciones, en el procesamiento de imágenes, por ejemplo, cuando hay varios sensores que están en un lugar donde no se puede intervenir y cada sensor puede tener una confianza más ó menos grande en el otro.

En las aplicaciones militares, es el problema de la fusión de datos. El comandante recibe la información de todos lados y debe hacer la fusión de esta información. Examinemos los casos siguientes.

Caso 1:



x=1

Si E_1 tiene una confianza completa en el evento A entonces p = 1 (probabilidad

igual a uno)

Si el experto E_2 tiene una confianza completa en el evento A igualmente.

¿Cuál será la creencia del experto uno en el experto dos?

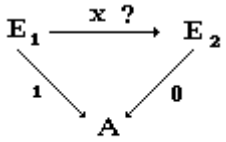
Será 1 por supuesto porque la creencia de:

E_1 en (A)=1

E_2 en (A)=1

La ignorancia es cero, B =0

Caso 2:

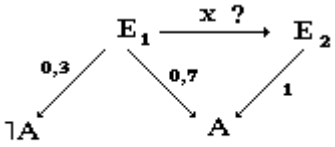


x= 0

Ahora si la creencia del experto uno en A es 1 y la creencia de E_2 en A es 0 .¿Cuál será la creencia de E_1 en E_2 ?

Cree exactamente lo contrario. E_1 no cree en lo que cree E_2 por lo tanto es cero.

Caso 3:



x= 0,7

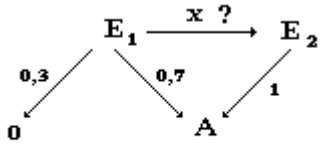
Pero si E_1 cree en A =0.7 y en no A =0.3.

E_2 cree en A = 1

Encontramos que la creencia de E_1 en E_2 será 0.7 según la fórmula (1)

Porque tiene algunas razones para creer que A no es verdad.

Caso 4:



x= 1

$m_1 (o) = 1 \Rightarrow a_1^* (a_2) = 1$

E_1 cree en A = 0.7 e ignora otra cosa es decir, no tiene razones para creer en otra cosa distinta de A.

Si por el contrario E_2 está seguro de A entonces E_1 creerá completamente en A. Es la noción de ignorancia. Nótese que si un experto (o un sensor) es completamente ignorante creerá completamente en otro sea lo que sea lo que este último crea, por lo tanto podemos rechazar el juicio de expertos (o sensores) que sean muy ignorantes.

De la misma manera se puede hacer con la plausibilidad.

Esquema del ADS

Para situar al Análisis de Datos con respecto al Análisis de Datos Simbólicos, podemos decir que hay 4 tipos de A.D. Según las entradas que se posean y según las salidas que se obtengan.

INPUT OUTPUT	OBJETOS - INDIVIDUOS (DATOS STANDARD)	OBJETOS SIMBÓLICOS
NUMÉRICO	(a)	(c)
SIMBÓLICO	(b)	(d)

En (a) tenemos datos clásicos en la entrada y tenemos resultados numéricos en la salida.

Este es el AD Clásico en la Estadística Clásica.

En (b) tenemos una entrada de datos clásicos y una salida en que obtenemos resultados de orden simbólico u objetos simbólicos. Por ejemplo si tenemos datos clásicos en la entrada se puede efectuar un Análisis Factorial clásico, una clasificación clásica y deseamos objetos simbólicos en la salida.

Las clases serán interpretadas como objetos simbólicos y los ejes factoriales serán interpretados por los objetos simbólicos.

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

Hay muchas formas de hacerlo pero una simple es tomar una unión de los valores alcanzados en la clase para cada variable. O bien en términos probabilísticos podemos tomar la frecuencia de cada uno de los valores alcanzados en cada una de las clases.

Y explicar esto como una conjunción de propiedades. De manera general podemos inducir según un modelo de ley normal una ley de probabilidad asociada a una clase y luego hacer de ella una descomposición simbólica para tener una conjunción de propiedades, de tipo booleano o de tipo probabilístico por ejemplo.

En el caso del análisis Factorial se pueden obtener también objetos simbólicos buscando para un eje los individuos más constitutivos. Por ejemplo programas SPAD de Análisis Factorial de Correspondencias proveen valores que permiten saber cuáles son los individuos más contributivos, cuáles son las variables más contributivas de este eje y teniendo este conjunto de individuos tenemos una clase. Teniendo una clase podemos construir un objeto simbólico, tomando la unión de los valores alcanzados por la clase o calculando las frecuencias en la clase por lo tanto podemos obtener una interpretación simbólica de un Análisis Factorial.

¿Cuál es el interés de tener una interpretación simbólica de un análisis clásico?. El interés primordial es el grado de explicación. Allí es donde el experto tiene el objeto en su mente y dice, yo tengo un conocimiento nuevo. Este eje expresa los individuos más bien de cierta edad,

que habitan en tal lugar o barrio de la ciudad. Esto el experto se lo dice para sí mismo. Para el A.D.S. este conocimiento es expresado por un objeto simbólico cuya calidad puede ser medida.

Así tenemos una nueva forma de extracción de conocimientos que se agrega a la cantidad que provee el análisis clásico, lo mismo ocurre con la clasificación. En lugar de decir solamente "esto es una clase", tenemos además una conjunción de propiedades que `provee el Objeto Simbólico que tiene también un poder explicativo mayor, interesante, que no reemplaza a los resultados habituales o a los indicadores de la estadística clásica, pero adopta una manera nueva e interesante de interpretar los resultados.

En (c) la entrada son objetos simbólicos y se extraen informaciones de orden numérico.

Es muy simple, basta definir y analizar las distancias entre objetos simbólicos. Luego con estas distancias se pueden efectuar los cálculos de escalamiento. El análisis de distancia es un caso clásico.

El último caso es el que tenemos en la entrada objetos simbólicos y en la salida también objetos simbólicos. Se puede hacer un Análisis Factorial de objetos simbólicos y efectuar una interpretación simbólica de los ejes.

En la actualidad muchos investigadores están haciendo esto ahora en lo que concierne a las pirámides.

Asimismo existe una sociedad en la que se va a desarrollar un software comercial (CISIA).

Se obtienen cosas así, mediante computadoras, apuntando a cada uno de los niveles que nos interesan tenemos objetos simbólicos asociados.

Algunas aplicaciones en Curso:

. THOMSON-CSF: La primera aplicación tiene lugar en un Programa Militar. Se trata de un avión que sobrevuela un país enemigo y los radares que lo quieren detectar, el problema del avión es tratar de confundir al radar que lo quiere detectar.

Se necesita conocer el tipo de radar que lo vigila, los tipos de radar son objetos simbólicos. Cuando un radar vigila a un avión se calcula $a_{(\omega)}$ para saber si lo detecta.

Un tipo de radar no es un punto R_p es algo más complejo.

. INRETS: Aplicación al problema de la construcción de tipos de accidentes.

. EDF: La Compañía de Electricidad de Francia, tiene 1000 proyectos de investigación por año, y cada proyecto está representado por un conjunto de palabras claves. Se describe cada uno de los proyectos como un objeto simbólico. Estos proyectos están organizados de manera que los directores de la compañía puedan tener rápidamente un panorama de lo que ocurre. Para saber cuáles proyectos son redundantes o si existen otros que se parezcan mucho a los de esta compañía.

Otro proyecto tiene que ver con las centrales nucleares, se quieren evitar accidentes como el de Chernovyl. Se forman expertos para pilotear la Central y las personas que comandan están frente a pantallas funcionando por equipos. Se hacen simulaciones en la pantalla, se colocan elementos como si ocurriera una cosa grave o como si no ocurriera nada y se observa el comportamiento de la persona frente a la pantalla. En realidad es un equipo de personas de las cuales se observa el comportamiento en un intervalo de tiempo

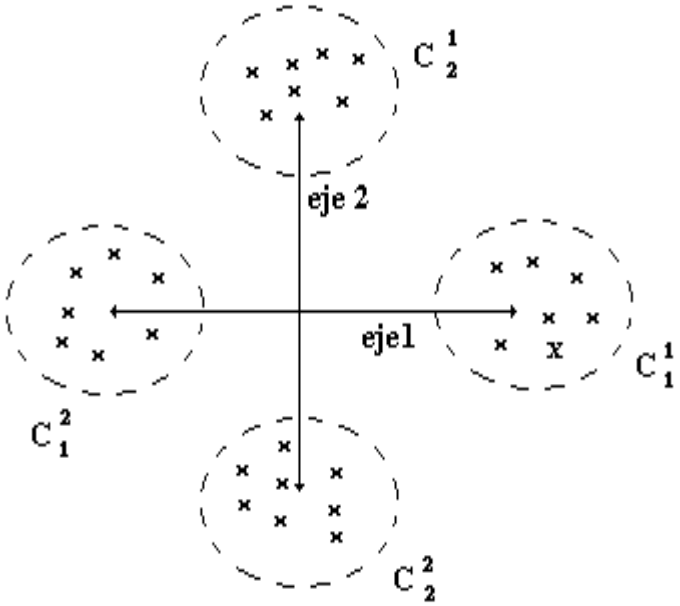
Se manipula una pantalla u otra y se ve la actitud del individuo, si intercambia conversación con su vecino, etc. Cada intervalo de tiempo es caracterizado por un objeto simbólico, es la conjunción de los eventos que se produjeron. Esto se hace para cada individuo y para cada equipo y luego se quiere estudiar este conjunto de datos.

. MUSEO: Otra aplicación se realiza en el Museo de París, consiste en investigar las especies de esponjas. Las esponjas son muy difíciles de describir, porque tienen dientes, cola, diferentes partes bien particionadas y hace a un objeto simbólico individual muy complejo. Se desea tener un sistema de identificación de esponjas.

. INSERM: Con un organismo de Investigación Médica se lleva a cabo un estudio sobre el reconocimiento de imágenes donde hay varios observadores. Con objetos simbólicos es posible organizar, visualizar y clasificar las imágenes. En el artículo [82] se proponen varios métodos para hacer esto y se pueden inducir reglas de objetos simbólicos para ser utilizadas especialmente como base de conocimiento para un sistema experto.

Interpretación Simbólica del Análisis Factorial

Mediante el Análisis de Datos es posible visualizar, organizar, clasificar objetos simbólicos, asimismo es posible inducir reglas entre ellos. En A.D. Clásico se pueden buscar los individuos que más contribuyen a cada eje y cada clase estará representada por un conjunto de propiedades booleanas o probabilísticas.



Cada clase de los objetos más contributivos está representada por un objeto simbólico

en el caso booleano $C_i^j = \hat{i}\ [y_i = v_i]$

en el caso probabilístico $C_i^j = \hat{i}\ [y_i = q_i]$

Hemos puesto a punto técnicas de cálculo de histogramas, caso simple de la estadística clásica y muy complejo para objetos simbólicos.

Primero porque para cada valor poseemos intervalos, y hay restricciones. El tamaño está comprendido entre dos valores, el color toma tal o cual valor, se pueden tener restricciones entre el tamaño y el color.

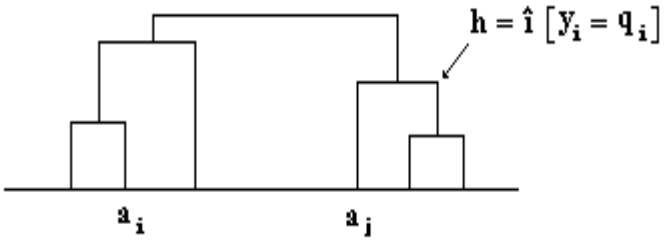
Por ejemplo cuando el tamaño es pequeño el color es claro, todo esto hay que tener en cuenta para la construcción de un histograma. Hay una tesis en curso sobre este tema (De Carvallo et al.) [21] y todavía hay muchas cosas por hacer.

A partir del histograma se pueden inferir leyes y aplicar la estadística clásica.

En clasificación tenemos O.S. de entrada y de salida.

A partir del momento de que hay una noción herencia del O.S., este O.S. tiene una extensión que está contenida en la intención de este Objeto.

Cada nivel generaliza el nivel que está debajo, lo menos posible. A partir del momento en que hay una generalización, se pueden inducir reglas.

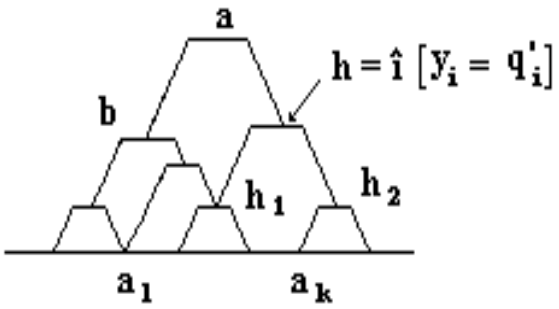


En la siguiente pirámide el objeto a, implica el O.S. b o h. La extensión de a contiene la extensión de h y la extensión de b, la unión de estas dos extensiones es la extensión a.

Como la extensión de b está incluida en la de a, ya que a es más general que b, tenemos la regla de b que implica a. Y también la regla h implica a. Por lo tanto podemos inducir reglas entre O.S.

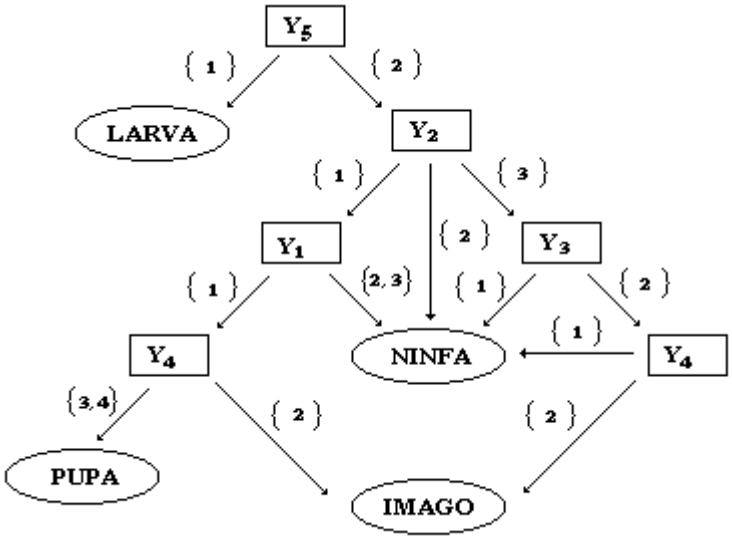
Gráfico pirámide (P.Bertrand, P.Brito) [11], [12], [13]

Cada nivel h es un objeto simbólico o una disyunción de objetos simbólicos.



El método de Árboles de Decisión.

Este ejemplo se trata de la Leismaniasis cuyo agente transmisor es el mosquito.



Cada especie de insecto está descrita por un conjunto de variables larva, pupa, etc. Cada especie es un O.S. y describe una clase. Se describe una clase de mosquito, no un individuo, cada objeto expresa a una clase. Buscamos la variable más explicativa. Es la variable que separa mejor las especies y encontramos que es Y_5, para la modalidad 1 es larva, para la modalidad 2 se busca entre todas las variables que quedan.

Cuál es la variable que discrimina mejor las especies y encontramos que es Y_2 que tiene dos modalidades, la modalidad 3 y la modalidad 1.

Se comienza el sistema y encontramos la variable y así sucesivamente. Cada especie está descrita por condiciones suficientes y aquí obtenemos reglas que dan condiciones necesarias para satisfacer la especie.

Podemos hacer esto de manera global o de forma adaptativa.

Cuando se dice global por ejemplo para reconocer una especie de planta, se puede hacer un árbol general que permite describir todas las plantas y reconocer bajando en este árbol cada planta. Este es un método global. Estamos obligados a descender de la misma manera para buscar la planta que necesitamos.

En el método adaptativo no se necesita pasar siempre de la misma manera en el árbol, por ejemplo observo una flor, y digo es blanca. El programa elimina todas las especies que tienen flores que no son blancas y se pueden producir dos casos O bien doy una nueva información o bien el programa, entre las flores que son blancas, busca la variable más discriminante de las especies que quedan y pregunta ¿la flor es grande?. En este momento respondemos a la pregunta si es que la flor es grande y conserva solamente a las flores que son blancas, por lo tanto es un sistema adaptativo.

Características esenciales de los objetos simbólicos

Los objetos simbólicos que hemos estudiado se expresan bajo forma de "conjunciones" (a veces puramente lógicas) de propiedades expresadas por variables clásicas (llamadas también "descriptores") del AD ya sean ellas cuantitativas o cualitativas (nominales u ordinales). Se distinguen de los objetos clásicos del AD en que ellos deben satisfacer las siguientes propiedades:

1) .*Los valores tomados por los descriptores pueden ser multivaluados ya sea para expresar clases definidas en intención o para expresar individuos que expresen una duda*.

Más formalmente, un ejemplo clásico de objeto simbólico utilizado en entrada o en salida es la "aserción". Una aserción booleana se define como la aplicación

a: $\Omega \rightarrow$ {verdadero, falso} tal que a (w) = $\hat{\imath}$ [y_i (w) $\in V_i$] donde las y_i son funciones

$\Omega \rightarrow O_i$ que describen los elementos de Ω y $V_i \subseteq O_i$. Por convención de notación se escribirá a = $\hat{\imath}$ [$y_i = V_i$]. De esta forma se puede expresar en intención, una clase de elementos llamados también instancias o individuos de Ω, por las propiedades que los caracterizan. Si a (w) = verdadero se dirá que w es parte de "la extensión" de a.

Ejemplo:

cepa = [color = { amarillo, marrón}] $\wedge$ [tamaño = [0,15]]

Si w es una cepa que yo he encontrado en el bosque, cepa (w) = verdadero si w es ya sea amarillo o marrón y si su tamaño está comprendido entre 0 y 15 centímetros.

Si ahora yo voy a describir un champiñon w del cual tengo dudas en lo que concierne a su color, entre blanco y amarillo, yo lo puedo escribir bajo la forma:

w_s = [color = {blanco, amarillo}] $\wedge$ [tamaño = 7.8]

Se da este caso cada vez que, luego de una observación, una persona expresa una duda entre varias modalidades de respuesta.

2) . *Los lazos conocidos entre los valores tomados por los descriptores deben poder ser expresados.*

Si existen lazos entre las V_i (es decir, si las V_i son funciones de las V_j) ello debe aparecer en la descripción del objeto.

Ejemplo: el tamaño de las cepas depende del color: ellas son claras cuando son pequeñas. Se escribe entonces:

a = [color = {amarillo, marrón}] $\wedge$ [tamaño = [0,7] si amarilla,] 7,15] si marrón]

Algunos descriptores pueden no tener sentido cuando otros toman ciertos valores.

Ejemplo: cuando el descriptor y_1 = "existencia de un sombrero" toma el valor "no", el descriptor y_2 = "color del sombrero" no tiene sentido. Se escribe entonces:

$a = [\ y_1 = \{ \text{sí, no}\}] \wedge [\ y_2 = \text{amarillo, marrón, } \varnothing \text{ si } [\ y_1 = \text{no }]]$

Por el contrario ciertos descriptores pueden tomar un valor cualquiera en una descripción.

Ejemplo: en un desperfecto de un cierto tipo que ha tenido lugar en un intervalo de temperatura [t = [20,25]] la velocidad del vehículo V: $\Omega \rightarrow O$ no interviene. Se escribirá entonces:

desperfecto = [t = [20,25]] $\wedge$ [V = O] para expresar el hecho de que V puede tomar no importa qué valor en O, el conjunto de velocidades posibles.

3) . *Los lazos conocidos entre partes de un objeto deben poder expresarse.*

En este caso, se ha extendido la noción de aserción a la de "horda"; una horda es un caso particular de objeto "estructurado" en IA; es una aplicación h: $\Omega p \rightarrow$ { verdadero, falso} tal que si $u = (u_1, ..., u_p) \in \Omega p$ entonces $h(u) = \hat{\imath}\ [\ y_1 (\ u_i\ ) \in V_i\ ]$; por convención se denotará:

$h = \hat{\imath}\ [\ y_i\ (\ u_i\ ) = V_i\ ]$

Ejemplo:

Descripción de un relato de accidente, haciendo intervenir un descriptor y_1 = "tipo de ruta", y_2 = "en", y_3 "posición", las u_i son las rutas y $u = (u_1, ..., u_p)$:

guión $h(u) = [\ y_1 (u_1) = A,B] \wedge [\ y_1 (\ u_2) = B, C\ ] \wedge [\ y_2 (u_1) = \text{auto}] \wedge [\ y_2 (u_2) = \text{moto}\ ] \wedge [y_3 (u_1, u_2) = \text{cruce múltiple}]$.

Dicho de otro modo, este relato describe los accidentes que se producen entre un auto en las rutas de tipo A o B y una moto en las rutas de tipo B o C en un cruce donde ellos se chocan en un cruce múltiple.

4) . *Se debe poder expresar las propiedades concernientes a las clases de individuos con la ayuda de expresiones relevantes de la lógica de 1er. orden: los "objetos de clases".*

En este caso, el elemento genérico que se describe es una parte llamada C de Ω; contrariamente a los casos precedentes donde los objetos simbólicos están definidos sobre Ω y describen las propiedades de un individuo genérico de una clase, aquí es necesario utilizar los cuantificadores $\forall$ y $\exists$ de la lógica de primer orden. Se denota $P(\Omega)$ al conjunto de las partes de Ω.

Ejemplo: Un experto en producción (en este caso, De Guio del ENSAIS en Estrasburgo) nos indica que las clases entre un conjunto de piezas de una usina Ω, deben ser por una parte constantes para una variable y_1 y por otra parte, que los valores 1 ó 4 de y_2 y 3 para y_2 no pueden cohabitar juntos; se describirá entonces este tipo de clase bajo la forma del objeto simbólico de clase siguiente:

a: P (Ω) →[verdadero, falso] tal que

a (C) = [$\forall$ $w_1w_2 \in C$, $y_1(w_1) = y_1(w_2)$] $\wedge$ [$\exists$ w_1w_2 : $y_2(w_2)$ = [1 ó 4] y $y_1(w_1) = 3$]

5) . *La semántica y la sintaxis subyacente en los datos y conocimientos de entrada debe poder expresarse: los "objetos modales".*

Los objetos simbólicos estudiados hasta ahora se dicen "booleanos" cuando no pueden tomar otro valor más que verdadero o falso (ej. a (w) ∈ {verdadero, falso}); en muchas aplicaciones esto será suficiente, pero a menudo el usuario se encuentra confrontado a objetos donde es indispensable una mayor suavidad en el conocimiento que él desea expresar. Para ello, hemos introducido los objetos llamados "modales" ya que ellos "moderan" los valores tomados por las variables; por ejemplo, un experto podrá decir que el color de un objeto de una clase es a menudo rojo y raramente amarillo bajo la forma del objeto simbólico a = [color = a *menudo* rojo, *raramente* amarillo] a es entonces una aplicación de Ω en el intervalo [0,1] y a (w) expresa un grado de "certidumbre" en cuanto a la pertenencia de w a la clase descripta por a. Más generalmente, se pueden definir los objetos modales bajo la forma:

a = î x [$y_ı = q_ı$] donde q_i puede ser una medida de probabilidad, de "posibilidad" o de "credibilidad" que satisface respectivamente los axiomas de la teoría de las probabilidades, posibilidades o de las credibilidades.

Adecuación de un objeto simbólico a un conjunto de objetos simbólicos por descomposición mixta de ley de leyes

Supongamos que estudiamos los pesos de bebés al nacer, de un país. Y que este peso varía según la misma ley en las localidades y luego en los departamentos, que son clases de localidades. Luego en las regiones, que son clases de departamentos y luego en los países, etc. Toda divergencia con respecto a tal modelo es interesante para efectuar el repertorio. Es cierto que en un país que se conduce bien la media de los pesos de los bebés debe ser la misma si se amplían las zonas de cada región.

Al principio tengo O.S. que pueden dar una probabilidad de peso y de tamaño, un conjunto de leyes de probabilidad sobre los pesos de los bebés en una localidad.

Ahora si una serie de localidades forman un departamento, si se observa la ley de probabilidad de cada localidad, cada punto no es un bebé sino una localidad. Busco una ley de probabilidad sobre la localidad que es un conjunto de bebés. Entonces paso a a*, subo un nivel. Y ahora si considero el conjunto de departamentos que forman la región, los departamentos son una clase en la cual cada individuo es una clase. Es una clase de clases. Puedo subir esto en distintos niveles y obtener objetos simbólicos de más alto nivel.

¿Cómo se estudia esto, en el caso en que paso de un nivel a otro? Se hace siempre de la misma manera y se pasa de a a a* y de a* a a**, etc.

O de q a q*, de q* a q**, etc.

Con una misma función h se pasa de q a q*, de q* a q**, y así sucesivamente. Esto da una idea de lo que es un fractal.

La ley de probabilidad es siempre la misma o si varía siempre varía de un nivel a otro.

Representación gráfica de objetos simbólicos por categorías y fractales

Podemos representar esta serie de O.S. en forma de fractal.

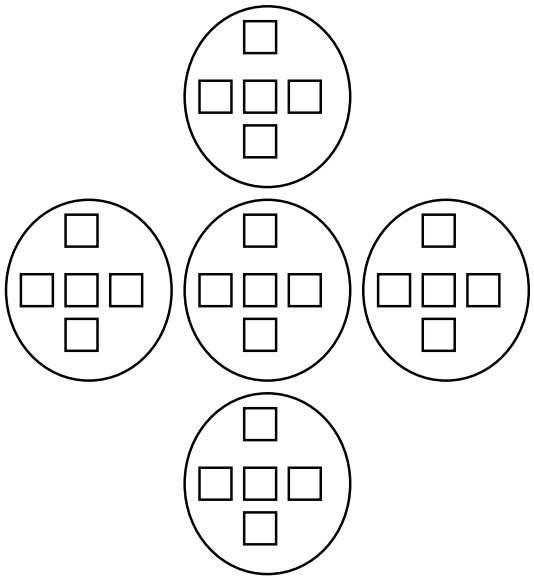


Figura A

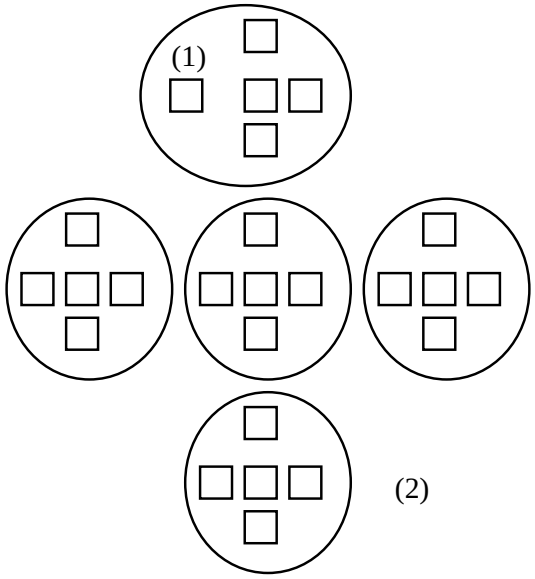


Figura B

Por ejemplo en la figura anterior si las localidades son los cuadrados, los departamentos los círculos. En el caso ideal tengo un fractal, (Fig.A) es decir, cuando el país se conduce bien y los pesos son los mismos de una localidad a otra, a pesar de la región.

Este es el modelo, pero si miro lo que ocurre en el terreno (Fig.B), si en el departamento me doy cuenta que hay una localidad que se aleja del modelo fractal, esto debería aparecer en la representación gráfica, de la misma forma si hay un departamento que se aparta del modelo fractal. De esta manera podemos detectar anomalías.

Otro tipo de aplicación sería lo del agujero de ozono, podemos recortar el cielo en pedazos, cada vez más grandes. Tener un fractal de diferentes proporciones de oxígeno, ozono, ozono libre, etc. y si existe alguna anomalía la podemos ver aparecer respecto al fractal.

Distintas etapas del Análisis Simbólico

¿Cómo tratar el problema?

Una estrategia que yo propongo consiste en partir de un conjunto de objetos individuales más o menos complejos.

Se extraen clases por Análisis Factorial, Clasificación, Arboles de Decisión, retículos, etc. Hasta aquí es análisis clásico.

Se representan estas clases a fin de obtener objetos definidos en intención en forma de descriptores. Esta intención puede ser obtenida directamente por los expertos, lo que hace un corte entre las dos etapas 1 y 2.

En la etapa 1 tenemos descripciones de clase, luego defino O.S.

Teniendo O.S. efectuamos el AD Simbólico, es decir podemos analizar, sintetizar, clasificar, discriminar, organizar por diferentes métodos de AD Simbólico.

En la etapa 4, podemos luego extraer de estos análisis, meta-conocimientos, como por ejemplo con la ayuda de la pirámide de herencia, reglas entre O.S., etc.

Por ejemplo con un Instituto de Investigaciones francés para trabajar sobre estrategias de pesca, se efectuaron entrevistas a pescadores para saber cómo van a pescar. Se consideraba el viento, la corriente, la humedad del aire y otros parámetros que se tenían en cuenta para pescar. Es decir que tienen una estrategia de pesca. Si se equivocan de estrategia puede ser muy peligroso para la gente que vive de la producción pesquera.

Se han considerado entonces que estas estrategias son objetos simbólicos.

Otro ejemplo de aplicación se refiere a la comparación de relatos de accidentes para obtener así una herramienta para los diagnósticos de seguridad en la ruta.

Si uno es capaz de definir tipos de accidentes se pueden entonces construir nuevas rutas, cortar árboles, etc. Se pueden enviar directivas al personal que se ocupa de las rutas. Esto permite establecer una estrategia para mejorar las rutas.

Existían expertos como Fleury y Pline que tenían varios años de experiencia y habían descripto diversos tipos de accidentes. Estos tipos de accidentes fueron transformados en O.S., por ejemplo, un tipo de accidente era: hombre de 30 a 50 años, que pierde el control de su vehículo, experimentado, a menudo alcoholizado y que ocurre durante el día. Otros tipos de accidentes eran previstos por otros expertos.

Por otra parte se proveía una base de datos aportada por los gendarmes que van al lugar del accidente y que describen todo. Es así como se expresaron los relatos como Objetos Simbólicos. Elegimos el modelo probabilístico.

Uno de los tipos era: de día, el 70 % los lunes, el 30 % los domingos, etc.

Luego que se ha definido el relato a partir del experto calculamos su extensión en la base de los datos de los gendarmes. En este momento nos damos cuenta que un tipo era de extensión vacía y otro de extensión demasiado grande. Por lo tanto el otro experto podría mejorar su tipo y una vez que se pusieron de acuerdo sobre los tipos, se recalcula su extensión.

Estos tipos cubrían sólo el 50 % de la población. Por lo tanto su experiencia podría ser buena pero no consideraba todos los relatos posibles. Sobre el resto de la población se aplicó un método de clasificación clásica, para obtener clases y luego O.S.

Por medio de discusiones con los expertos teníamos que ponernos de acuerdo sobre estos relatos. Y una vez que obtuvimos el conjunto de los tipos, realizamos una base que cubría todo el espectro. Después observamos cuál era la calidad de los relatos.

Es importante ver la calidad o cualidades de los objetos, completos, simples, fuertes. Si a pesar de tener muchas propiedades tiene una gran extensión, esto es contradictorio pues muchas propiedades reducen la extensión. Ellos querían tener estos objetos porque los consideraban muy interesantes. Los objetos fuentes era los tipos que ellos habían construido al principio.

Los conocimientos de los expertos habían permitido detectar clases, que los algoritmos de clasificación no hubieran podido obtener.

Habiendo obtenido guiones de accidentes se pueden definir "prototipos" tomando las probabilidades mayores en cada uno de los eventos y los eventos más frecuentes. Por ejemplo, el prototipo de un guión era: que se produjese el 70 % de los accidentes el lunes, el 30 % los domingos y entre las 7 y 9 horas.

Cuando se obtuvieron los distintos tipos de accidentes hicimos un tipo de pirámide. Conocimientos sobre los conocimientos. Partimos de tipos y se obtuvo una especie de árbol que permitió tener una visión sintética de los accidentes.

Si hay choque se puede producir fuera de la intersección o en una intersección. Si es fuera de la intersección es por cuestión de velocidad o porque es un usuario local, si es una cuestión de velocidad o es por acción de adelantarse o por la acción de disminuir la velocidad. Esto describiría cada uno de estos tipos.

Conclusiones

El Análisis de Datos Simbólico llena un vacío importante que se sitúa entre la inteligencia artificial y la estadística. Entre los enfoques lógicos, simbólicos y numéricos, abre la vía a un gran campo de aplicación: el del procesamiento de objetos complejos teniendo en cuenta conocimientos no necesariamente de orden puramente numérico.

Los objetos simbólicos son especies de átomos de conocimiento, comprenden un campo tan vasto como los conocimientos mismos.

Las aplicaciones han otorgado resultados interesantes en campos tan diversos como las señales de radares, las estrategias de pesca, los tipos de accidentes, de enfermedades o especies de flores.

El programa existe ya, generadores automáticos de reglas, de árboles de segmentación sobre objetos simbólicos, realizados en varias tesis que ya han sido defendidas. Se han realizado aplicaciones en diversas industrias con estudiantes que han desarrollado programas. No hay desempleo en este campo. Mis estudiantes son rápidamente contratados incluso antes de finalizar la tesis.

Existe un desafío, concretar todos estos avances por medio de un gran programa, en forma de biblioteca de programas. Por ejemplo una especie de Modulad o SPAD Simbólico.

¿Cómo adquirir conocimiento a partir de los expertos?

Debería existir un nexo entre la inteligencia artificial, la estadística y la lógica. Es un campo completamente interdisciplinario, trabajar sobre conocimientos y no sobre datos. Si Uds. expresan sus conocimientos en forma de frases los estadísticos y matemáticos luego lo deben poder expresar en términos matemáticos. A partir de allí muchas cosas se pueden hacer.

El punto crucial es la interdisciplinariedad de este campo. La dificultad de encontrar una enseñanza porque cada uno se ocupa de plantear ciencias al margen. Hay que interesarse por los problemas de la vida.

Referencias

[1] ADANSON, M. (1757), "Historire Naturelle du Sénégal-Coquillages", Bauche París.

[2] ADANSON, M. (1763), "Famille des plants", Vol.1, Vincent, París.

[3] ARNAULT, A. and NICOLE, P. (1662), La logique ou l'art de penser", reprinted by Froman, Stuttgart (1965).

[4] BACKHOFF, G. (1967), "Lattice Theory", Amer.Math.Soc.Providence (ed.3).

[5] BARBUT, M. MONJARDET, B. (1971), "Ordre et classification", T.2 Hachette, Paris.

[6] BECKNER (1959), "The Biological Way of thought", Columbia University Press, New York, 220 p.

[7] BELSON (1959), "Matching and prediction on the principle of biological classifications", Applied Statistics, vol VIII.

[8] BÉNZECRI, J.P. et al (1973), "L ánalyse des données", Dunod, París.

[9] BOCHENSKI, I.M. (1970),"A history of formal logic", I.Thomas, trans., New York: Chelsea Publishing Co.

[10] BREIMAN, L., FRIEDMAN, J.H., OLSKEN, R.A., STONE, C.S. (1984), "Classification and regression trees", Belmont, Wadsworth

[11] BRITO, P. and DIDAY, E.,"Pyramidal representation of symbolic objects", in Knowledge, Data and Computer Assisted Decisions, Schader M. and Gaul W. (ed.) NATO ASI serie F: Computer and System Sciences Vol.61.

[12] BRITO, P., (1993), "Symbolic objects: order structure and pyramidal clus- tering", in this issue.

[13] BRITO, P., DIDAY,E. (1990),"Pyramidal representation of symbolic objects", in NATO ASI Series, Vol. F.61, Knowledge Data and computer-assisted Decisions edited by Schader and W. Gaul. Springer Verlag.

[14] CARNAP, R. (1947), "Maning and necessity: a study in Semantic and Modal Logic", The University of Chicago Press, Chicago.

[15] CELEUX, G., DIDAY, E., GOVAERT, G. LECHEVALLIERY.,RALAMBONDRAINY, H. (1989),"Classification automatique: environement Statistique et Informatique", Dunod.

[16] CELEUX, G., DIEBOLT, J., (1985), "The SEM algorithm: A probabilist teacher algorithm derived fron the EM algorithm for the mixture problem", Computational Statistics Quarterly 2 p. 73-82.

[17] CHOMSKY, N. (1966), "Cartesian linguistics: a chapter in the history of rationalist thought", Harper & Row, New York, French transl. Seuil, Paris 1969.

[18] CHOQUET, G., (1953), "Théorie des capacités", Ann. Inst. Fourrier 5. 131-295.

[19] DALE, M.B. and ANDERSON, D.J. (1973), "Inosculate analysis of vegetation data", Aust. J.Bot. vol 21, pp.253-276.

[20] DALLWITZ (1974), "A flexible computer program for generating diagnostic keys", Syst. Zoology, 23 (1), pp. 50-57.

[21]DE CARVALHO, F.A.T. (1991), "Histogramme en Analyse des Données Symboliques", Dissertation Univ.París 9 Dauphine.

[22]DEMPSTER, A.P., (1967), "Upper and Lower Probabilites Induced by a Multivalued Mapping", Annals of Mathematical Statistics 38, 325-339.

[23]DESCLES, J.P. (1986), "Travaux de linguistique et de littérature", XXIV,1, Strasbourg, Klimcksieck.

[24]DESCLES, J.P. (1991), "La notion de typicalité: une approche formelle", in Sémantique et Cognition, pp. 225-244, Editions du CNRS París.

[25]DESCLES, J.P. and KANELLOS, I. (1991), "La notion de typicalité: une approche formelle", in Sémantique et Cognition, CNRS París, D.Dubois editor.

[26]DIDAY, E. (1971), "La méthode des nuées dynamiques", Rev. Stat. Appliquée, vol.XIX, N° 2, pp. 19-34.

[27]DIDAY, E. (1976), "Sélection typologique de variables", Rapport INRIA.

[28]DIDAY, E. and SIMON, J.C. (1976), "Cluster Analysis", in K.S.Fu (ed.), Digital Pattern Recognition, Springer Verlag, pp. 47-94.

[29]DIDAY, E. et al. (1979), "Optimisation en classification automatique", 800 p., INRIA (ed.), Rocquencourt 78150 Le Chesnay, France.

[30]DIDAY, E. LEMAIRE, J. POUGET, J. TESTU, F. (1984), "Eléments d'analyse des données", Dunod.

[31]DIDAY, E., (1990), "Knowledge representation and symbolic data analysis", in NATO ASI Series, Vol. F 61, Knowledge Data and computer-assisted Decisions edited by Schader and W. Gaul. Springer Verlag.

[32]DIDAY, E., (1991), "Des objets de l'analyse des données à ceux de l'analyse des connaissances", in "Induction symbolique-numérique à partir de données", Y.Kodratoff, E.Diday, editors, CEPADUES (Toulouse).

[33]DIDAY, E., (1992), "From data to Knowledge: new objects for a statistical analysis", in New Techniques and Technologies for statistics conference, GMD, Bonn.

[34]DIDAY, E., GOVAERT, G. LECHEVALLIER, Y., SIDI, J. (1980),

[35]DUBOIS, D. (1992), "Representations catégorielles, prototypes et typicalité", Le Courrier du CNRS "Sciences Cognitives", N° 79, Octobre p. 68

[36]DUBOIS, D., PRADE, H., (1986), "A set-theoretic view of belief functions", International Journal General Systems, Vol. 12, pp 193-226

[37]DUBOIS, D., PRADE, H., (1988), "Possibility theory", Plenum New York.

[38]DUQUENNE, V. (1986), "Contextual Implications between attributes and some representation properties for finite lattices", Beitrage zur Begriffsanalysis Ganter,Wille, Wolf (ed.) Wissensthafts Verlag Mannheim.

[39]FISHER, D. and LANGLEY, P. (1986), "Conceptual Clustering and its relation to Nemerical Taxonomy", Workshop on Artificial Intelligence & Statistics. W. Gale (ed.) Addison-Wesley.

[40] GANASCIA, J.G. (1991), "Charade: apprentissage de bases de connaissances", Y. Kodratoff and E.Diday (eds) Cepadues.

[41] GEACH, P. and BLACK, M., "Traslation from the philosophical writtings of Gottlob Frege", Oxford: Blackwell.

[42] GOWER, J.C. (1974), "Maximal predicative classification", Biomet. vol.30, pp. 643-654.

[43] GOWER, J.C. (1975), "Relating Classification to identification", in JR.J. Pankhurst (editor). Biological Identification with computer. pp. 65-72, London, Academic Press.

[44] HEIDEGGER, M. (1662), "Die Frage nach dem Ding", Max Niemeyer Verlag.Tübinger. In french "Quést-ce qu'une chose?" Gallimard (1971).

[45] JAMBU (1978), "Classification automatique pour l'analyse des données", Dunod, París.

[46] JEVONS, W.S. (1877), "The principles of Science: A treatise on Logic and Scientific Method", 2nd ed. rev. Macmillan London and New York, 786 p.

[47] KANT, E. (1785), "Fondement de la métaphysique des moerus", p. 71, Delagrave, Paris.

[48] KAPLAN, A. and SCHOTT, H.F. (1951), "A calculus for emprical classes", Méthodes 3, 165-190.

[49] KRUSKAL, J.B., WISH, M. (1978), "Multidimensional scaling", Sagr., Beverly Hills, Calif.

[50] LATTA, R. and McBeath, A. (1956), "The elements of logic", London Macmillan (Original work published in 1929).

[51] LAURITZEN, S.L. and SPIEGELHALTER, D.J. (1988), "Local computation with probabilities on graphical structures and their application to expert system". In Readings in uncertain reasoning (1990) edited by G. Shafer and J. Pearl. Morgan Kaufman Publishers.

[52] LE GUYADER, H. (1988), "Théorie et Histoire en biologie", J.Vrin, París.

[53] LEBBE, J. and VIGNES, R., "Génération de graphes d'identification à partir de descriptions de concepts", in Induction symbolique numérique Y. Kodratoff and E. Diday (editors) Cepadues-éditions.

[54] LEBBE, J., VIGNES, R., DARMONI, S., (1990), "Symbolic numeric approach for biological knowledge representations: a medical example with creation of identification graphs", in: Proc. of Conf. on Data Analysis, Learning Symbolic and Numerical Knowledge, Antibes ed. E.Diday, Nova Science Publishers, Inc. New York.

[55] LEBOWITZ, M. (1983), "Concept learning in a rich input domain" Proc.of. the Machine Learning Workshop, pp. 177-182.

[56] LEBOWITZ, M. "Generalisation from natural language yext", Cognitive Science 7, 1, pp. 1-40.

[57] LERMAN, I.C. (1981), "Classification et analyse ordinale des données", Dunod, Paris.

[58] LOUIS, P. (1956), "Aristote, les parties des animaux", reprinted by "Les belles lettres".

[59] MICHALSKI, R. (1980), "Knoeledge Acquisition Though conceptual clustering: a theoretical framework and an algorithm for partitioning data into conjunctive

concepts", Int.Jour. of Policy Analysis and Information Systems, Vol. 4, N° 3.

[60] MICHALSKI, R. and STEPP, R.E. (1983), "Automated Construction of Calssifications Conceptual Clustering versus Numerical Taxonomy", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.5, N° 4, July.

[61] MICHALSKI, R.S.; DIDAY, E.; STEPP, R.E., (1982) "A recent advances in data analysis: clustering objects into classes characterized by conjonctive concepts" Progress in Pattern Recognition vol.1. L.Kanal and A. Rosenfeld Eds.

[62] MICHALSKI, R.S.; CARBONELL, J.G.; MITCHELL, T.M., "Machine learning, and Artificial Intelligence Approach", Springer Verlag.

[63] MINSKY, M. (1975), "A framework for representing knowledge "the psychology of computer vision", New York, Mac Graw-Hill.

[64] PANHURST, R.J. (1978), "Biological identification. The principle and practice of identification methods in biology", London: Edward Arnold.

[65] PEARL, J. (1988), "Probabilist reasoning in intelligents systems" Morgan Kaufman, San Mateo.

[66] QUINLEN, J.R. (1986), "Induction of decision trees", Machine Learning 1: pp. 81-106, Kluwer Academic Publishers, Boston.

[67] RALAMBONDRAINY, H. (1991), "Apprentissage dans le contexte dún schéma de bases de données", Y. Kodratoff and E.Diday (editors), Cepadues.

[68] ROSCH, E. (1978), "Principle of categorization" in E. Rosch and B. Lloyd (eds), Cognition and Categorisation pp. 27-48, Hillsdale, N.J.: Erlbaum.

[69] ROSCH, E. and MERVIS, C.B. (1975), "Family resemblances: studies in the internal structure of categories" Cognitive Psychology, 7, pp. 573-605.

[70] ROUX, M. (1985), "Algorithmes de classification", Masson.

[71] SCHAFER, G. (1976), "A Mathematical theory of evidence" Princeton University Press.

[72] SCHAFER, G. (1990), "Perspectives on the Theory and Practice of Belief functions", International Journal of Approximate Reasoning, Vol.4, Numbers 5/6.

[73] SEBAGH, M.; DIDAY, E.; SCHOENAUER, M. (1980), "Incremental learning from Symbolic Objects", in Knowledge, Data and Computer Assisted Decisions, Schader M. and Gaul W. (ed) NATO ASI serie F: Computer and System Sciences Vol.61.

[74] SHWEIZER, B.; SKLAR, A. (1960), "Statistical matric spaces" Pacific J. Math 10: 313-334

[75] SOKAL, R.R. and SNEATH, P.H.A. (1963), "Principle of Numerical Taxonomy", San Francisco W.H. Freeman (sec.edit.1973).

[76] SUTCLIFFE, J.P. (1992), "Concept, class and category in the tradition of Aristotle", in Categories and concepts theoretical views and inductive data analysis. Academic Press.

[77] TUKEY, J. (1977), "Exploratory data analysis", Addison-Wesley, Reading, Mass.

[78] WAGNER, H. (1973), "Begriff", in Handbuch philosophischer Grundbegriffe, eds. H.Krungs, H.M.Baumgartner and C.Wild, Kösel, München 191-209.

[79] WARD, J.H. (1963), "Hierarchical grouping to optimize an objective function", J. Amer. Stat. Assoc. 58, pp. 236-244.

[80] WILLE, R. (1989), "Knowledge acquisition by methods of formal concept analysis", in: Data Analysis, Learning symbolic and numeric knowledge, E.Diday (Ed.), Nova Sciences Publishers.

[81] ZADEH, L.A. (1971), "Quantitative fuzzy semantics". Informations Sciences, 159-176)

[82] DIDAY, E. (1993) "An Introduction to Symbolic Data Analysis"

[83] DIDAY, E. (1993), "Quelques aspects de l'Analyse des Données Symboliques", presentado en las Jornadas MODULAD de Lannion de 1993.

Esta publicación se terminó de imprimir en el mes de junio de 1997, en CERIDER-IRICE, Bv. 27 de Febrero 210 bis – 2000 – Rosario- Provincia de Santa Fe – República Argentina

En el marco del convenio de cooperación entre el INRIA y organismos científicos argentinos el Prof. Edwin Diday expuso a lo largo de dos conferencias en IRICE los principios fundamentales de sus últimos desarrollos, en Análisis de Datos Simbólicos.

El tema, presentado por primera vez en español, es de interés para investigadores de distintas áreas del conocimiento por el enfoque multidisciplinario de la exposición, donde se revaloriza la utilización de sus juicios y experiencias en la construcción de clasificaciones y tipologías.

En efecto, ¿por qué desaprovechar los valiosos conocimientos de los expertos en pos de cuantificaciones reduccionistas? Saber representarlos por expresiones a la vez simbólicas y numéricas, saber manipular y utilizar estas expresiones a los fines de ayudar a decidir, de mejorar el análisis, de sintetizar y organizar nuestra experiencia y nuestras observaciones, es el objetivo de Análisis de Datos Simbólicos.

Esta nueva metodología se basa en el Análisis de Datos, grupo de técnicas aún no muy difundidas en nuestro medio que, tomando distancia de la estadística clásica, se proponen ayudar a descubrir regularidades o estructuras de respuestas de grandes conjuntos multidimensionales de unidades.

Si el Análisis de Datos clásico se viene planteando como la metodología más adecuada para el análisis cuantitativo en Ciencias Sociales, sobre todo porque no son necesarios supuestos ni modelos a priori, con el Análisis de Datos Simbólicos es posible otra vuelta de tuerca hacia el conocimiento de una realidad que en la mayoría de los casos se nos presenta difusa, caracterizada por percepciones, creencias, juicios y no siempre por categorías mutuamente excluyentes.

IRICE
27 de Febrero 210 Bis
2000 - Rosario
República Argentina