Práctica 204. Clasificación (I): árboles de decisión.

Un modelo del razonamiento de clasificación

Todos utilizamos procesos mentales de clasificación de forma mas o menos consciente. De forma natural tendemos a agrupar en *clases* o *categorías* sucesos y objetos con características observables distintas. Clasificar algo (un objeto, patrón, medida, etc.) es identificarlo como miembro de una clase conocida (figura 1.)

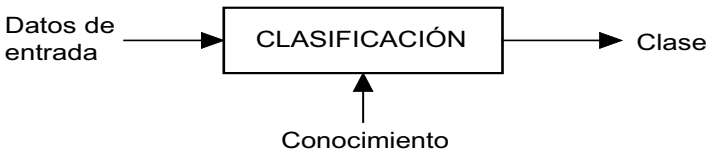

Figura 1: El proceso de clasificación.

Las clases expresan regularidades de modo que todos los miembros de una clase comparten ciertas propiedades, o lo que es lo mismo, las clases se definen en términos de condiciones necesarias. Consideraremos que en un sistema de clasificación existe siempre un conjunto finito y predefinido de soluciones o clases. La determinación cuáles son estas clases y su definición es un problema distinto y, en general, más complicado que la tarea de clasificación.

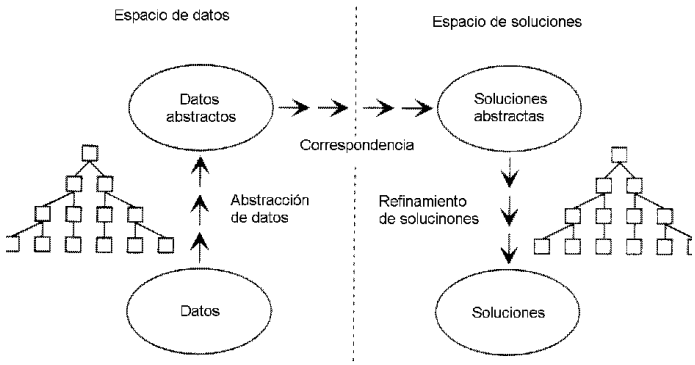

Figura 2: Fases en la clasificación.

Las tareas de clasificación, tal como se han definido aquí, comprenden muchos procesos que llevamos a cabo cotidianamente:

- la *selección* de entre un conjunto prefijado de opciones, a partir de una descripción de las preferencias del agente.
- el *diagnóstico* de una enfermedad o avería a partir de un conjunto de síntomas.

- la *clasificación* propiamente dicha, es decir, la determinación del valor de un atributo distinguido llamado *clase* a partir de los valores de otros atributos.

En un sistema basado en el conocimiento, el proceso de clasificación puede caracterizarse mediante las siguientes fases: (1) abstracción de datos, (2) correspondencia con una taxonomía de soluciones preenumeradas, y (3) refinado dentro de la taxonomía de soluciones, donde cada fase puede descomponerse a su vez en varias etapas de abstracción, correspondencia y refinado (véase la figura 2).

Por ejemplo, consideremos el problema de diagnostico planteado por el sistema pionero MYCIN. Se trata de identificar una enfermedad infecciosa (si la hay), a partir de los datos médicos de un paciente. Esta tarea puede modelarse de forma natural como una tarea de clasificación, donde cada clase o solución corresponde a una determinada enfermedad. Su implementación se realizó mendiante reglas de producción con encadenamiento hacia atrás.

Recordemos qué conocimiento utilizaba MYCIN para la clasificación. Un ejemplo lo tenemos en la siguiente regla, que establece una relación directa entre el número de glóbulos blancos de la muestra de sangre de un paciente y una posible causa de infección:

```
SI existe un análisis de sangre Y
    el número de glóbulos blancos es inferior a 2500
ENTONCES las siguientes bacterias pueden estar causando infección:
    E. coli (0.75),
    Pseudomonas-aeruginosa (0.5),
    Klebsiella-pneumoniae (0.5).
```

Pero, ¿qué significa en realidad esta regla? ¿Cuál es su justificación? La asociación heurística subyacente puede resumirse de la siguiente forma: un tipo de *condición de riesgo* de infección (aunque no la única) es la *inmunosupresión*, en la que el grado de respuesta del sistema inmunológico se encuentra inhibido por efecto de medicamentos, radiaciones o algún trastorno del sistema inmunológico. La *leucopenia* es una clase particular de inmunosupresión en la que hay una deficiencia de glóbulos blancos. Suele considerarse que un umbral inferior en el número de glóbulos blancos es 2500. Dicho en orden inverso, si el número de glóbulos blancos es inferior a 2500, entoces el paciente tiene leucopenia, que es una situación de inmunosupresión. Los pacientes con condición inmunosupresora estan en condición de *riesgo de infección*.

Cuando un paciente se encuentra en una condición de riesgo de infección, las bacterias que habitualmente se encuentran en las zonas no estériles del cuerpo pueden provocar una infección. Algunas de estas zonas son la piel, las vias altas respiratorias y el tracto digestivo. La zona más propensa a una infección es el tracto gastrointestinal donde se encuentran habitualmente *enterobacterias* como *E. coli*, *Pseudomonas* o *Klebsiella*. Resumiendo, cuando un paciente está en condición de riesgo de infección, las infecciones provocadas por este tipo de bacterias son probables. Para indicar la confianza en la certeza de esa relación, MYCIN emplea factores de certeza. Otras asociaciones heurísticas pueden confirmar o negar la hipótesis anterior, y deberán ser valoradas adecuadamente según el peso de su factor de certeza.

El análisis anterior revela algunas de las categorías de relaciones utilizadas con frecuencia en problemas de clasificación:

1. Relaciones para la abstracción de datos. En general, el nivel de certeza de

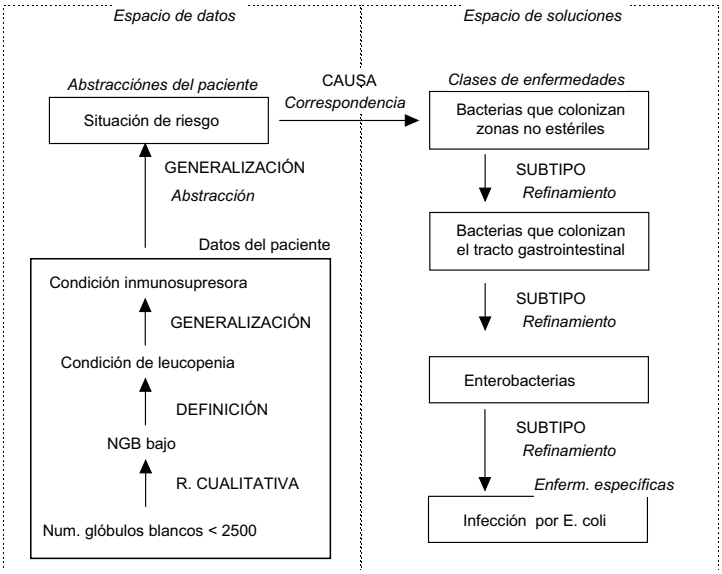

Figura 3: El conocimiento de MYCIN

estas relaciones suele ser muy elevado:

- Abstracción por definición: se centra en características esenciales y necesarias de los datos. Por ejemplo: "Si una bacteria puede vivir en un lugar sin oxígeno, entonces es una bacteria anaeróbica".
- Abstracción cualitativa: se utiliza para simplificar el manejo de datos cuantitativos. Por ejemplo: "Si el número de glóbulos blancos es menor de 2500, entonces el nivel de glóbulos blancos es bajo".
- Abstracción por generalización: utiliza una jerarquía de subtipos. Por ejemplo: "Si una persona es padre, entonces es un hombre".

2. Relaciones de correspondencia de clases. En general se trata de relaciones heurísticas, menos fiables que las anteriores, que se establecen entre los datos observables y las clases de soluciones. Pueden utilizarse en ambos sentidos, por ejemplo, bien para obtener un diagnóstico a partir de unos datos o bien para validar hipótesis. Algunos ejemplos son:

   - Asociaciones causales: es decir, relaciones causa-efecto más o menos fundadas.
   - Correlaciones empíricas: cuando se sabe que dos factores aparecen juntos en muchos casos, sin que exista un mecanismo causa-efecto conocido. Por ejemplo, grupos de riesgo establecidos estadísticamente para una enfermedad.

3. Relaciones de refinamiento. Conectan generalizaciones de soluciones con casos concretos. Son relaciones simétricas a las de abstracción. Pueden seguir, por ejemplo, una jerarquía de tipos o de componentes.

Cuando se construye un sistema de clasificación es interesante representar explícitamente todos estos tipos de relaciones. El resultado es que el conocimiento empleado queda más claro que si se utiliza una relación directa como la empleada originalmente en MYCIN (vd. figura 3)

En numerosos casos la relación entre datos y soluciones no se establece de forma directa, sino por medio de un espacio intermedio. Por ejemplo, la conceptualización más exacta del sistema PROSPECTOR es la que considera la existencia de varios espacios en los que se realizan clasificaciones heurísticas acerca de la edad de las formaciones geológicas consideradas, el entorno regional de las formaciones, etc. En estos casos, el diagrama "en Π"de la figura 2 se convierte en un diagrama "en M".

El caso más simple: árboles de decisión.

El procedimiento más sencillo para llevar a cabo la clasificación es el del *árbol de decisión*. El árbol está constituido por nodos; cada nodo n está etiquetado por un atributo $a(n)$. Cada nodo n tiene tantos hijos n_i como valores posibles tiene el atributo $a(n)$. Es decir, cada posible valor de $a(n)$ etiqueta un arco que sale de n y llega a uno de sus hijos. De esta manera, partiendo de la raíz, se va transitando en cada paso al hijo que corresponda al valor hallado para el atributo del último nodo considerado. Finalmente, se llegará a un nodo hoja que estará etiquetado con el valor de la clase correspondiente a la combinación de valores hallados para los atributos considerados. Un ejemplo se muestra en la figura 4.

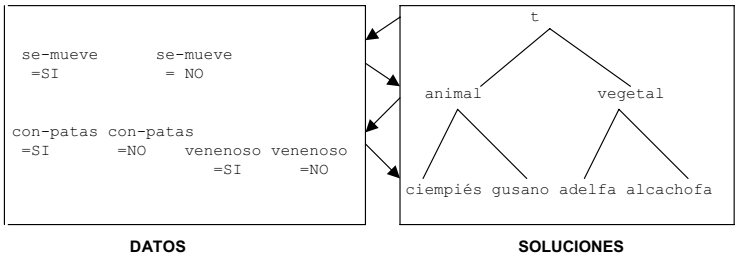

Figura 4: Un árbol de decisión

En este caso, la relación entre datos y soluciones es muy directa: cada combinación considerada de valores de atributos lleva a una solución parcialmente refinada. Los nodos hoja corresponderían a la solución con el máximo nivel de refinamiento.

Diversas variantes de los árboles de decisión –más o menos sofisticadas– aparecen ya compiladas en numerosos casos reales. Piénsese en las "claves taxonómicas", en las guías de campo (del tipo *¿Cómo identificar los pájaros de España?*), en las claves de diagnóstico de los libros de medicina o psicología, en los manuales de detección de averías, etc.

Sin embargo, como cualquiera que haya usado este material sabe, consultar un árbol de decisión no suele ser suficiente para resolver un problema de clasificación; pues puede ocurrir a) que el usuario sea incapaz de discernir el valor del atributo, pese a que éste sea evidente para el experto (ejemplo: *¿aparecen*

pequeñas pintas parduzcas en el lomo? El usuario inexperto no sabe exactamente qué se quiere decir con los difusos identificadores "pequeño" y "parduzco"); b) que el valor del atributo no sea de ninguna manera determinable en la situación considerada, en cuyo caso será necesario buscar una vía alternativa de razonamiento (por ejemplo: *¿Cuántos pétalos tiene la flor?* Si no estamos en la época de floración, difícilmente se podrá averiguar); c) que el auténtico experto detecte que la casificación dada por el árbol no es la correcta por tratarse de un çaso excepcional". Por todo ello, es un claro abuso de lenguaje decir que un árbol de decisión simple es un procedimiento "inteligente"de clasificación.

En los ficheros ARBOL-DEC1.CLP y ARBOL-DEC2.CLP se muestran dos implementaciones de árboles de decisión en CLIPS. La primera emplea reglas CLIPS para representar los arcos del árbol de decisión; la segunda emplea las reglas para definir el algoritmo de recorrido del árbol, cuyos arcos se describen mediante hechos desordenados.

Práctica 205. Algoritmos de recubrimiento.

Sea S_k una posible solución al problema de averiguar el valor de *Clase*. Sea un atributo D_i, y sea V_{ij} uno de sus posibles valores. Se pueden dar las siguientes relaciones abstractas entre la solución y el valor:

1. La solución S_k implica el valor V_{ij}, o más claramente, $Clase = S_k \rightarrow D_i = V_{ij}$ o, equivalentemente, $D_i \neq V_{ij} \rightarrow Clase \neq S_k$.

   Este tipo de conocimiento es el típicamente presente en los sistemas de clasificación. A veces tenemos un consecuente disyuntivo, es decir, la solución S_k implica uno de los valores $V_{ij_1} \ldots V_{ij_n}$, o más claramente, $Clase = S_k \rightarrow (D_i = V_{ij_1} \vee \ldots \vee D_i = V_{ij_n})$.

   Supongamos que en el caso que estamos clasificando está presente el dato D_i con el valor V_{ij}. Si se da esta relación, se dice que la solución *explica* el dato. Si hacemos uso del razonamiento abductivo, podemos decir que el valor V_{ij} *aporta evidencia a favor* de la solución S_k. Por el contrario, supongamos que en el caso que estamos clasificando está presente el dato D_i con un valor v distinto de V_{ij}. Entonces se dice que la solución *es incompatible* con el dato. Si hacemos uso del razonamiento deductivo, podemos decir que el dato *excluye* la solución.

2. El valor V_{ij} implica la solución S_k, o más claramente, $D_i = V_{ij} \rightarrow Clase = S_k$ o, equivalentemente, $Clase \neq S_k \rightarrow D_i \neq V_{ij}$.

   Este tipo de conocimiento no es frecuente en los sistemas de clasificación, pero lo mencionamos para completar el cuadro de todas las posibilidades. Supongamos que en el caso que estamos clasificando está presente el dato D_i con el valor V_{ij}. Si se da esta relación, se dice que el dato *determina* la solución. Si hacemos uso del razonamiento deductivo, podemos decir que el dato *implica* la solución. Por el contrario, supongamos que en el caso que estamos clasificando está presente el dato D_i con un valor v distinto de V_{ij}. Entonces se podría decir haciendo uso del razonamiento abductivo que el dato *aporta evidencia en contra* de la solución.

3. En cualquier otro caso, el dato es irrelevante para la solución. Si para todos sus posibles valores el dato es irrelevante, entonces no vale la pena investigar el valor concreto que ha tomado en el caso analizado.

Se suele adoptar el término "recubrimiento" como descripción común de los anteriores procesos de razonamiento. Dentro del razonamiento por recubrimiento, cabe seguir diversos criterios y objetivos para definir el concepto de solución. Por ejemplo:

- Criterio de completitud: buscar *una* solución; o buscar *todas* las soluciones; o buscar "un número razonable" de soluciones.

- Criterio de inclusión: puede ser inclusión *conservadora* cuando la solución no es incompatible con ningún dato; *recubrimiento positivo* cuando la solución explica algún dato y no es incompatible con ninguno; o *explicación completa* cuando la solución explica todos los datos y no es incompatible con ninguno.

- Criterio del umbral de confianza: se define una medida de la confianza en la solución y se exige que exceda de cierto valor.

Puede que las relaciones (1) y (2) de más arriba sean seguras, pero más habitualmente expresarán un conocimiento incierto. Para razonar con estas relaciones inciertas, el enfoque tradicional en los sistemas expertos ha sido asignar heurísticamente un factor numérico a cada una de las relaciones o reglas, y definir métodos *ad hoc* para manipular estos *factores de incertidumbre* de reglas y hechos (sistema EMYCIN, por ejemplo). Sin embargo, un tratamiento más fundamentado de la cuestión nos lleva al formalismo de las *redes bayesianas*.

El caso más simple de razonamiento por recubrimiento lo tendremos cuando se cumplan las siguientes hipótesis:

1. Todos los datos posiblemente relevantes están presentes al comienzo del proceso de razonamiento (ello implica que son relativamente pocos y relativamente fáciles de conseguir).

2. No hay jerarquía de soluciones; todas ellas pueden considerarse ya completamente refinadas.

3. Hay un predicado `supera-prueba-p(S, D)` que comprueba si la posible solución S es admisible en presencia de los datos D. Para ello se seguirá alguno de los criterios expuestos más arriba (inclusión conservadora, recubrimiento positivo, etc.)

4. Hay un procedimiento `ordenar(L)` que ordena las soluciones restantes según algún criterio (mayor certidumbre heurística, mayor simplicidad, etc.)

En estas condiciones, el algoritmo puede ser el que muestra el cuadro 1.

```
Algoritmo recubrimiento-1
1.    L <- NIL.
2.    Obtener el conjunto total de datos D.
3.    Abstraer D y obtener el conjunto de datos abstractos D'.
4.    Para cada posible solución S
5.        Si supera-prueba-p(S, D') entonces L <- cons(S, L).
6.    L <- ordenar(L).
```

Cuadro 1: Algoritmo simple de recubrimiento.

La condición (1) puede resultar irreal: ¿qué ocurre si los atributos posibles son varios miles? ¿Los solicitaremos todos antes de empezar a razonar? Por otra parte, ¿qué ocurre si algún atributo es tal que la determinación de su valor resulta costosa? Lo lógico será pues que al comienzo estén presentes tan solo ciertos datos, y que a lo largo del proceso de razonamiento se requieran únicamente los valores que vayan resultando necesarios.

Por otra parte, la condición (2) también puede resultar irreal. En muchos casos (lo vimos en los árboles de decisión) las soluciones forman de manera natural una jerarquía de refinamientos. Un algoritmo más apropiado será entonces el de la tabla 2.

```
Algoritmo recubrimiento-2
1.    L <- NIL.
2.    Obtener los datos iniciales D_0
3.    D'_0 <- abstraer(D_0)
4.    Para cada posible solución S del nivel inicial J_0
5.        Si supera-prueba-p(S, D'_0) entonces L <- cons(S, L)
6.    Para cada solución S de L
7.        Si S es refinable,
8.            L <- L -{S}
9.            Para cada S_j refinamiento inmediato de S
10.                Obtener los datos útiles D_j para el nivel de S_j
11.                D'_j <- abstraer(D_j)
12.                Si supera-prueba-p(S_j, D'_j)
13.                    L <- cons(S_j, L)
14.    L <- ordenar(L)
```

Cuadro 2: Un algoritmo más realista de recubrimiento.

Referencias