APRENDIZAJE AUTOMATICO POR CLASIFICACION

LUIS EDUARDO MUNERA
Matemático de la Universidad del Valle, Máster y Doctor en Informática de la Universidad Politécnica de Madrid. Ex profesor de la Facultad de Informática de la Universidad Politécnica de Madrid. Profesor ICESI

I. INTRODUCCION

Una condición previa a la realización de sistemas expertos, es decir, de sistemas con un nivel de competencia semejante al de un ser humano, experto en un dominio específico del saber, es la identificación, recolección y formalización de conocimientos heurísticos.

Como lo señalamos en [2], el proceso previo a la creación de un sistema experto, que atañe a la adquisición de conocimiento heurístico, constituye hoy en día, el "cuello de botella" del desarrollo de los sistemas expertos. En [2] presentamos una herramienta de adquisición de dicho conocimiento, basada en técnicas de emparrillado que le permite a los expertos humanos realizar un proceso de clasificación bipolar, que conduce a la inferencia de reglas heurísticas.

Existen otras aproximaciones para atacar este problema. Las más prometedoras proceden de una joven disciplina llamada "Aprendizaje automático", que pretende elaborar técnicas de adquisición automática de los conocimientos simbólicos. Dentro de esta disciplina, existen a su vez, varias aproximaciones a la solución del problema. Una de las aproximaciones es la que se denomina "Aprendizaje de procedimientos de clasificación" y que parte del supuesto de que todo proceso cognitivo presupone la posibilidad de "nombrar" las cosas. Y nombrar supone definir categorías y recurrir a procedimientos de clasificación.

Los procedimientos de clasificación, básicamente, consisten en la identificación, a partir de una base de datos, de los conceptos y sus valores que permiten definir una categórica conclusión.

Este proceso de aprendizaje puede ser visto como un procedimiento de inferencia inductiva que, partiendo de un conjunto de casos o ejemplos, almacenados en la base de datos, genera un conjunto de reglas de clasificación.

Se han desarrollado varios sistemas de aprendizaje de este estilo. Pero sin duda, el pionero, el que abre el paradigma, es el sistema ID3 de Quinlan [3].

En este artículo se propone una definición formal de clasificación y un

algoritmo de clasificación consistente con la definición.

Partimos del supuesto de que la base de casos o ejemplos se encuentra sobre una estructura relacional (tabular). Por lo tanto en la siguiente sección presentamos algunas definiciones básicas sobre bases de datos relacionales, que son fundamentales para entender la formalización matemática que sobre clasificación exponemos en la sección posterior.

II. DEFINICIONES BASICAS

En el modelo relacional de datos, las relaciones se representan mediante tablas, que tienen un número fijo de columnas y un número variable de filas (llamadas tuplas).

Las columnas de las tablas son denotadas por un nombre de atributo. Un atributo es un nombre que designa a un dominio de una relación. Cualquier valor asociado con un atributo se le denomina valor del atributo.

Mientras que los dominios de una relación no necesitan ser distintos, los nombres de atributos asignados a ellos deben serlo. Esto significa que atributos diferentes de una misma relación pueden tener el mismo dominio, pero el nombre de atributo tiene que ser distinto para cada diferente columna (Atributo) de la tabla (relación).

Una definición normal es la siguiente:

Definición:

Dados los conjuntos $D_1, D_2 ..., D_n$ que llamamos dominios, no necesariamente distintos, y dado un conjunto de atributos $T = \{A_1, ... A_n\}$ tales que cada atributo A_i está asociado a un dominio D_i; una relación r sobre T (que simbolizamos por r (T)) es un conjunto finito de funciones t tales que:

$t: \{A_1, ..., A_n\} \longrightarrow D$, donde $D = U_i D_i$

La función debe enviar a cada atributo en T a un miembro de su correspondiente dominio.

Una base de datos relacional es un conjunto de tablas (relaciones) como la que acabamos de definir.

Sobre una relación r(T) se pueden ejecutar varias operaciones matemáticas que son la base de los lenguajes de consulta que utilizan los manejadores de bases de datos relacionales.

De esas operaciones nos interesan solamente dos: proyección y selección.

La proyección es una operación que aplicada sobre una tabla produce el efecto de fragmentar o partir verticalmente la tabla. Es decir, que si tenemos una tabla con "n" columnas, podemos partir la tabla, generando una nueva tabla que posea una cantidad menor de columnas que la original. Esta operación se expresa simbólicamente como, πA1,..., An (r), en donde "π" es el símbolo que denota la operación de proyección; "r" es la tabla que vamos a partir y "A_1,...,A_n" es la lista de columnas de r que nos interesa tomar.

La selección es una operación que aplicada sobre una tabla produce el efecto de fragmentar o partir horizontalmente la tabla. Es decir, que si tenemos una tabla con "m" filas, podemos partir la tabla, generando una nueva tabla que posea una cantidad menor de filas y el mismo número de columnas que la tabla original. Esta operación se expresa simbólicamente como:

$\Im_f(r)$ en donde "$\Im$" es el símbolo que denota la operación de selección; "r" es la tabla que vamos a partir y "f" es la fórmula que especifica las filas de r que nos interesan.

En este caso en particular, las fórmulas que consideramos son de la forma $f=(A_i=C_i)$, que significa seleccionar las filas que bajo la columna A_i tienen el valor constante C_i.

III. CLASIFICACION

En general podemos considerar el proceso de clasificación como consistente en encontrar reglas que particionan los datos dados de una relación, en grupos disyuntos.

Más correctamente, fijamos un atributo de la relación que va a ser el consecuente de las reglas, es decir, vamos a formular grupos disyuntos alrededor de sus valores con base en el resto de atributos de la relación que actúan como discriminadores y que vienen a ser los antecedentes de las reglas.

Sea r(T) una relación definida sobre $T=\{A_1, ..., A_n\}$, de grado n y cardinalidad m.

Cada tupla de la relación la vamos a identificar con un entero positivo. Sea $W = \{1, 2, ..., m\}$ el conjunto de identificadores. Ampliemos T para cobijar un nuevo atributo que es el identificador de fila y cuyo dominio es W.

Sea $T^* = \{A_0, A_1, A_2, ..., A_n\}$ en donde A_0 es el identificador de fila.

Sea $\wp(w)$ el conjunto de partes de W. Denominamos partición de W a cualquier subconjunto Q de $\wp(w)$, $Q=\{X_i\}$ $i\in I$, $Q \neq \phi$, tal que:

1. $i\neq j \Rightarrow X_i \cap X_j = \Phi$
2. $\bigcup_{i\in I} X_i = W$

Para cada atributo $A_i \neq A_o$ perteneciente a T*, sea $R(A_i \subseteq DOM(A_i)$ el rango de A_i que consiste en los valores del dominio presentes en la relación (puede haber valores nulos).

Sea $A_i \neq A_o \in T^*$ y sea $C \in R(A_i)$, consideremos la restricción $(A_i = C) \subseteq W$, definida como,

$(A_i=C)=\{X \in W \mid X \in \pi_{A_0}(\Im_{A_i=c}(r))\}$.

Estas restricciones inducen una participación sobre W, asociada a cada atributo A_i.

LEMA: $P(A_i)=\{(A_i=C) \mid C \in R(A_i)\}$ es una partición sobre W.

D/ Sean $(A_i=C_1)$ y $(A_i=C_2)$ con $C_1 \neq C_2$, pertenecientes a $P(A_i)$. Supongamos que $(A_i= C_1) \cap (A_i = C_2) \neq \Phi$, entonces existe un $X \in (A_i = C_1)$ y $X \in (A_i = C_2)$ lo que significa *que* $X \in \pi_{A_0}(\Im_{A_i=C_1}(r))$ y $X \in \pi_{A_0}(\Im_{A_i=C_2}(r))$, lo que se puede expresar como, $X \in \pi_{A_0}(\Im_{A_i=C_1 \wedge A_i=C_1}(r))$ lo cual es imposible para $C_1 \neq C_2$, pues C_1 y C_2 son valores constantes, por lo tanto si $C_1 \neq C_2$, necesariamente

$(A_i = C_1) \cap (A_i = C_2)$ es Φ.

Si $X \in \bigcup_C (A_i = C)$ entonces $X \in (A_i = C)$, $\exists C, C \in R$ (A_i) pero por definición $(A_i = C) \subseteq W$, por lo tanto si

$X \in \bigcup_C (A_i=C) \Rightarrow X \in W$, es decir,

$\bigcup_C (A_i = C) \subseteq W$.

Sea $X \in W$ entonces X es un identificador de fila, es decir,

$x \in \pi_{A_0}(\Im_{A_i=C}(r)), \exists i, i=1...n$ y $\exists C$, $C \in R(A_i)$, entonces $X \in (A_i = C), \exists i, i = 1...n$ y por lo tanto, $X \in \bigcup_C (Ai=C)$.

En consecuencia $\bigcup_C (A_i = C) = W$.

Queda demostrado que $\wp(A_i)$ es una partición sobre W.

Aunque $\wp(A_i)$ es una partición de W, es general, $\bigcup_{i=1}^{n} \wp(A_i)$ no es una partición de W.

Ahora, sin pérdida de generalidad asumimos que el atributo objetivo en la clasificación es A_n.

COROLARIO: $\wp(A_n) = \{(A_n = C) \mid C \in R(A_n)\}$ es una partición sobre W.

Las restricciones del tipo, $(A_i = C) \subseteq W$, pueden ser generalizadas a una conjunción de restricciones,

$(A_1=C_1)\wedge\ldots\wedge(A_{p=}C_p) \subseteq W =$

$\{X \in W / X \in \pi A_0(\Im_{A1=C1\wedge\ldots\wedge Ap=Cp}(r))\}$

con $C_1 \in R(A_1),\ldots,C_p \in R(A_p)$.

Formalmente, el proceso de clasificación, es un proceso de obtención de un conjunto de reglas, $I=\{R_1, R_2,\ldots, R_s\}$, en donde cada regla es de forma:

R_i: IF<condición i> THEN <CLASIFICACION i> con i=1 ...s, en donde:

1. <condición i> = $(A_1=C_1) \wedge\ldots\wedge (A_p=C_p)$,
   <clasificación i> = $(A_n=C_o)$
2. $\{A_1, A_2, \ldots, A_p\} \subseteq T^* - \{A_0, A_n\}$

DEFINICION

Dada una relación r(T) con $T=\{A_1, \ldots, A_n\}$. Decimos que $I=\{R_1, \ldots, R_s\}$ es una clasificación sobre r(T) respecto a A_n'si y sólo si:

C1) $p(r)=\{$<condición i>/ i=1...s$\}$ es una partición sobre W.

C2) $(\forall C), C\in R(A_n), P(A_n=C)=\{$<condición i>/<condición i>$\cap$(An=C)$\neq\emptyset$,

$\exists i$, i=1...s$\}$ es una partición sobre (An=C).

Ahora estamos en condiciones de establecer el siguiente resultado:

PROPOSICION

Dada una relación r(T) con $T=\{A_1, \ldots, A_n\}$. Decimos que $I=\{R_1,\ldots R_s\}$ es una clasificación sobre r(T) respecto a A_n, si y sólo si,

$P(r) = \bigcup_C P(A_n = C)$

D/: Si I es una clasificación entonces se cumplen C_1 y C_2.

$X \in P(r) \Rightarrow X =$ <condición i>, $\exists y$, i=1...s

Dado que $P(A_n)=\{(A_n=C/C)\in R(A_n)\}$ es una partición de W (por el corolario al Lema), entonces $X\subseteq(A_n=C)$, $\exists C \in R(A_n)$ y por lo tanto, $X\cap(A_n=C)\neq\phi$, es decir, <condición i> $\cap(A_n=C)\neq\phi$, $\exists i$, $\exists C$. Por lo tanto, $X\in P(A_n=C)$, $\exists C$, $C\in R(A_n)$. Es decir que $X\in \bigcup_C P(A_n = C)$.

En consecuencia, $P(r) \subseteq \bigcup_C P(A_n = C)$

Si $X\in \bigcup_C P(A_n=C) \Rightarrow X =$ <condición i>, $\exists i$, y en consecuencia $X\in P(r)$, por lo tanto, $\bigcup_C P(A_n=C) = P(r)$.

Queda entonces demostrado que si I es una clasificación,

$P(r)=\bigcup_C P(A_n = C)$

Supongamos que $P(r)= \bigcup_C P(A_n = C)$ necesitamos probar que C_1 y C_2 se cumplen.

Demostremos C_1 (P(r) es una partición de W).

Sean $K_i, K_j \in P(r)$ con $i\neq j$. Supongamos que $K_i\cap K_j=\phi$. Entonces $\exists X_0$ tal que $X_0\in K_i$ y $X_0\in K_j$, entonces,

$X_0\in \pi_{A_0}(\Im_{A_1=C_1\wedge\ldots\wedge A_p=C_p}(r))$ y

$X_0 \in \pi_{A_0}(\Im_{A_1=C_1\wedge\ldots\wedge A_p=C_p}(r))$ tal que $C_i\neq C_i', \exists i$, i=1...s

Entonces,

$X_0\in \pi_{A_0}(\Im_{A_1=C_1=C_1'\wedge\ldots\wedge A_i=C_i\wedge A_i=C_i'\wedge\ldots\wedge A_p=C_p=C_p'}(r))$ lo cual es imposible pues Ci y C_i' son constantes, por lo tanto $K_i\cap K_j=\phi$.

Probemos que: $\bigcup_{i\in I}$ <condición i>=W.

Si $X\in \bigcup_{i\in I}$ <condición i> entonces $X\in$ <condición i>, $\exists i\in I$. Entonces $X\in W$ y $\bigcup_{i\in I}$ <condición i>$\subseteq$ W.

Si $X\in W$ entonces $X\in (A_n=C)$, $\exists C\in R(A_n)$ esto por el corolario.

Supongamos que $X\notin$ <condición i> $\forall_i\in I$, entonces $X\in P(r)$ y por la igualdad $P(r) = \bigcup_C P(A_n = C)$ establecemos que, $X\in P(A_n=C)$ para todo C, lo cual está en contradicción con $X\in (A_n=C)$, $\exists C$. Por lo tanto, $X\in$ <condición i>, $\exists i\in I$. Es decir $W\subseteq\bigcup_{i\in I}$ <condición i> y en consecuencia, $\bigcup_{i\in I}$ <condición i>=W.

Por lo tanto queda demostrado que P(r) es una partición de $W(C_1)$.

Análogamente podemos probar $(C_2)\big(P(A_n=C)$ es una partición de $(A_n=C)$, $\forall C\in R(A_n)\big)$.

IV. ALGORITMO DE CLASIFICACION

Teniendo en cuenta el resultado obtenido en la proposición anterior, podemos generar un algoritmo de clasificación.

El algoritmo se deriva de la proposición en la medida en que a partir de ésta podemos obtener las siguientes características:

1.) Si $(A_i=C_i)\subseteq A_n=C_o)$, $\forall C_i\in R(A_i)$, $\exists C_o\in R(A_n)$. Entonces podemos eliminar la columna A_i. Esto debido a que todos los valores no nulos de A_i en la relación están asociados a un único valor de A_n y por lo tanto no sirven como discriminadores, es decir, $P(A_i)$ no es una partición sobre W y no se cumple el lema.

2.) Si <condición i>$\subseteq(A_n=C_o)$, $\exists C_o \in R(A_n)$ entonces podemos generar una regla del tipo.

R_i: IF <condición i> THEN $(A_n=C_o)$.

Esto por el hecho de que <condición i> $\in P(r)$ por definición de P(r) y dado que <condición i>$\subseteq (A_n=C_o)$ entonces pertenece a $P(A_n=C_o)$, y por ende a $\bigcup_C P(A_n = C)$.

El algoritmo propuesto es el siguiente:

Entrada: Una relación r(T), con $T=\{A_o, A_1, \ldots, A_n\}$, no vacía, sin filas repetidas y sin columnas con un único valor.

Salida: Un conjunto de reglas, $\mathfrak{R}$

Método:

1. Hacer $\mathfrak{R}=\phi$
2. Aplicar reducción vertical.

IF r(T) posee más de una fila y se cumple una de las siguientes condiciones:

a) $\exists A_i\notin \mathfrak{R}, \mathfrak{R}(A_i)=C_o$, fijo.

b) $\exists A_i\notin \mathfrak{R}$, $\exists K\in \mathfrak{R}(A_n), (A_i=C_i)\subseteq (A_n=C_o)$, $\forall C_i\in \mathfrak{R}(A_i)$ no nulo.

THEN Eliminar la columna de r(T).

ELSE

IF r(T) posee una única fila.

THEN: Eliminar la fila de r(T) y generar una regla con esa fila. Incluir esa regla en $\mathfrak{R}$.

3. Aplicar reducción horizontal.

For n=1 hasta n-2 hacer:

IF existe un conjunto de restricciones $\{(A_j=C_j)_k\}_{k=1}^{n-1}$ tal que $(A_i=C_i)_{1\wedge\ldots\wedge}(A_i=C_i)_{n-1}\subseteq (A_n=C_o), \exists C_o\in \mathfrak{R}(A_n)$.

THEN

IF existe un conjunto de restricciones $\{(A_j=C_j)_k\}_{k=1}^{n-1}$, $j\neq i$ tal que $(A_j=C_j)_1\wedge\ldots\wedge (A_j=C_j)_{n-1}\subseteq(A_n=C_o)$ y $(A_i=C_i)_1\wedge\ldots\wedge(A_i=C_i)_{n-1}=(A_j=C_j)_1\wedge\ldots\wedge (A_j=C_j)_{n-1}$

THEN

IF $(A_j=C_j)\subseteq(A_n=C_o)$, $\forall C_j\in \mathfrak{R}(A_j)$

THEN Eliminamos los valores de r(T) que se encuentran en la intersección de la columna A_j y las filas, $(A_j=C_j)\wedge(A_n=C_o)$.

Generamos la regla, "IF $(A_j=C_i)_1 \wedge ... \wedge (A_i=C_i)_{n-1}$ THEN $(A_n=C_o)$". La incluimos en $\Re$.

ELSE Generamos una de las siguientes reglas:

"IF $(A_k=C_k)_1 \wedge ... \wedge (A_k=C_k)_m \wedge ... \wedge (A_i=C_i)_1 \wedge ... \wedge (A_i=C_i)_{n-1}$ THEN $(A_n=C_o)$" ó "IF $(A_k=C_k)_1 \wedge ... \wedge (A_k=C_k)_m \wedge (A_j=C_j)_1 \wedge ... \wedge (A_j=C_j)_{n-1}$ THEN $(A_n=C_o)$". En donde $\{(A_k=C_k)_\alpha\}^m{}_{\alpha=1}$, m<n es tal que $\exists k$, $\exists C_k, k=i=j$, $C_k=C_i=C_j$. Incluirla en $\Re$.

ELSE

IF $(A_j=C_j)_1 \wedge ... \wedge (A_j=C_j)_{n-1} \subset (A_i=C_i)_1 \wedge ... \wedge (A_i=C_i)_{n-1}$

THEN Eliminar de r(T) las filas que contengan a $(A_i=C_i)_1 \wedge ... \wedge (A_i=C_i)_{n-1}$ y generar las reglas del tipo:

"IF $(A_i=C_i)_1 \wedge ... \wedge (A_i=C_i)_{n-1}$ THEN $(A_n=C_o)$" Incluirlas en $\Re$.

IF r(T)=ϕ THEN Fin

ELSE Aplicar reducción vertical

Hacer n=n+1

End For

4. Generar reglas con las filas restantes de r(T).

Incluirlas en $\Re$.

V. APLICACION

Consideramos la siguiente tabla de Decisión **[1]**,

Caso	Lluvia	Suelo	Topografía	Problema
1	Intensa	Empapado	Escarpada	Grave
2	Intensa	Empapado	Suave	Grave
3	Intensa	Húmedo	Escarpada	Grave
4	Intensa	Húmedo	Suave	Medio
5	Importante	Empapado	Escarpada	Grave
6	Importante	Húmedo	Escarpada	Medio
7	Importante	Húmedo	Suave	Medio
8	Baja	Empapado	Escarpada	Nulo
9	Baja	Húmedo	Escarpada	Nulo
10	Baja	Húmedo	Suave	Nulo

Aplicando el algoritmo, obtenemos dos conjuntos de reglas:

$\Re_1$: $\{R_1, R_2, R_3, R_4, R_5, R_6\}$ y

$\Re_2$: $\{R_1, R_2, R_3, R_4, R'_5, R'_6\}$, en donde:

R_1: IF Lluvia = Baja THEN problema = NULO

R_2: IF Lluvia = Importante y Suelo = Empapado

THEN problema = Grave

R_3: IF Lluvia = Importante y Suelo = Húmedo

THEN problema = medio

R_4: IF Lluvia = Intensa y Suelo = Húmedo y Topografía = Suave

THEN problema = medio

R_5: IF Lluvia = Intensa y Suelo = Empapado

THEN problema = Grave

R_6: IF Lluvia = Intensa y Suelo = Húmedo y Topografía = Escarpada

THEN problema = Grave

R'_5: IF Lluvia = Intensa y Topografía = Escarpada

THEN problema = Grave

R'_6: IF Lluvia = Intensa y Topografía = Suave y Suelo = Empapado

THEN problema = Grave

VI. BIBLIOGRAFIA

[1] CUENA, J. y otros. *Inteligencia Artificial: Sistemas Expertos.* Ed. Alianza Editorial, Serie Alianza Informática, Madrid, 1986.

[2] MÚNERA, L.E., RITHER, G.A.; MADRID, J.M. *Kelly: una herramienta para la adquisición de conocimiento heurístico. Publicaciones Icesi* Nº 47, Cali - Colombia, 1993.

[3] QUINLAN, J.R. *Machine Learning: An Artificial Intelligence Approach.* Eds. R. Michalski, J. Carbonell, and T. Mitchell. Los Altos, California. Morgan Kaufmann, 1986.