INDUCCIÓN DE ÁRBOLES DE DECISION

Introducción

El aprendizaje de los árboles de decisión es uno de los más sencillos de implementar, pero a su vez de los más poderosos. Un árbol de decisión toma de entrada un objeto o instancia del dominio, la cual esta descrita por un conjunto de atributos y da como salida una *clase* del objeto.

Cada nodo interno del árbol corresponde a una prueba en el valor de uno de los atributos, y las ramas están etiquetadas con los posibles valores de la prueba. Los árboles de clasificación se construyen a partir de un conjunto de entrenamiento. Una vez construido, es posible verificar la clase de nuevas instancias probando cada uno de los valores de los atributos de la instancia con respecto a los atributos y valores asociados a los nodos y ramas que conforman al árbol.

A continuación se introducirán algunos conceptos básicos así como notación, para posteriormente describir a detalle el algoritmo ID3 (Induction of Decisión Trees)

Conceptos y Notación

Entenderemos por un *dominio* Ð a un conjunto de datos, donde cada objeto o elemento $\varphi \in$ Ð esta definido sobre un conjunto finito de atributos $<a_1, \ldots, a_m>$, donde $\forall\, a_i \in \mathbf{A}$, $\mathbf{A} = \{A_1, \ldots, A_m\}$ un universo de atributos. Cada atributo A_x esta definido sobre un conjunto de valores denominado *dominio* del atributo, denotado por $D(A_x)$.

Sea **A** un conjunto de atributos y $\Re$ una relación cualquiera sobre **A**. Un conjunto de entrenamiento *S* es una relación no vacía cualquiera, donde $S \subseteq \Re$.

En un conjunto de entrenamiento *S* suele presentarse que grupos de registros tienen características comunes entre si, en los cuales los valores de los respectivos atributos coinciden. A estos grupos se les conoce como *clases*. Formalmente, una clase *C* es un subconjunto de *S*, donde todas las instancias en *S* satisfacen alguna prueba específica sobre sus atributos, es decir:

$$C = \{\varphi \in S, \varphi = <a_1, \ldots, a_m> \text{ y } a_x = s, s \in A_x\}$$

Arboles de Clasificación

Los árboles de clasificación son construidos de tal forma que en cada rama del árbol queden asociadas instancias de una sola clase. La forma de construirlos depende del algoritmo seleccionado, los cuales finalmente determinan la estructura de los árboles. Un algoritmo muy representativo de los árboles de clasificación es ID3, desarrollado por Quinlan durante la década de los 70's [1], el cual opera utilizando la *entropía* de los datos.

Los árboles de clasificación buscan agrupar en los nodos de mayor profundidad instancias o registros de una misma clase, los cuales deben de cumplir todas las

condiciones especificadas por el camino desde la raíz del árbol hasta el nodo donde se encuentran asociados. Los nodos que no tienen un subárbol asociado (los de mayor profundidad) son conocidos como *hojas del árbol*.

A continuación se describen los pasos en el algoritmo ID3.

Primer paso

Consiste en determinar las clases que se usarán como base para la construcción. Para ello, se selecciona un atributo $A_c \in \mathbf{A}$, llamado *atributo clasificador*. Este atributo es seleccionado por el usuario. Notemos que la cardinalidad de A_c determina el número de clases a trabajar, esto es, existen $|A_c|$ posibles clases.

Una vez establecido este atributo, se debe de determinar al conjunto de atributos bajo los cuales se va a construir el árbol. A este conjunto de atributos lo representamos por **A'**, donde $\mathbf{A'} \subseteq \mathbf{A}$, donde $A_c \notin \mathbf{A'}$.

Asumamos que el conjunto de entrenamiento S esta definido sobre el conjunto de atributos $\mathbf{A'} \cup \{A_c\}$.

Segundo Paso

El proceso de construcción del árbol inicia determinando un atributo que, comparado con el resto de los atributos que se trabajan, demuestre que reduce en mayor grado la entropía del conjunto de instancias S. A este atributo lo llamaremos *atributo ganador*, denotado por A_g (notemos que $A_g \in \mathbf{A'}$). En ID3 para calcular el atributo ganador se debe elegir aquél que presente la mayor ganancia. La ganancia es un cálculo que depende directamente de la entropía que cada atributo presenta en relación al atributo clasificador.

Cada vez que se determina un atributo ganador, este se asociará a un nodo dentro del árbol, y se derivarán del mismo $|A_g|$ ramas, una por cada valor distinto en el dominio del atributo ganador. Cada vez que se determina un atributo ganador, las instancias en S son divididas entre las ramas del nuevo nodo creado, de tal forma que en cada rama solo queden registros que coincidan con el valor del atributo A_g asociado a la rama correspondiente.

Denotaremos por v_i^j al i-ésimo valor de un atributo A_j. Claramente $v_i^j \in D(A_j)$. Además, denotaremos por $S = \{ S_1^j, \ldots, S_t^j \}$ a las particiones sobre S, donde S_i^j representa al conjunto de instancias que tienen el valor v_i^j correspondiente al atributo A_j.

Establezcamos como se calcula el atributo ganador. Consideremos que S es el conjunto de registros o instancias sobre las cuales se trabaja.

1. Se selecciona un atributo $A_x \in \mathbf{A'}$. Sea $S = \{ S_1^x, \ldots, S_q^x \}$ el número de conjuntos diferentes que se definen a partir de los diferentes valores del domino del atributo A_x.

2. Cada instancia o registro perteneciente a un conjunto $S_j^x \in S$ tiene una clase asociada. Sea $S_j^x = \{S_1^c(S_j^x), \ldots, S_t^c(S_j^x)\}$ una partición sobre S_j^x, donde cada $S_i^c(S_j^x) \in S_j^x$ representa al conjunto de instancias que tienen el valor v_j^x (correspondiente al atributo A_x) con la clase v_i^c, $v_i^c \in A_c$. Con la información anterior, se puede calcular la probabilidad de cada clase con base en el conjunto S_j^x, relacionando la cantidad de elementos de cada conjunto $S_i^c(S_j^x)$ con respecto al número de registros del conjunto S_j^x. Esta probabilidad queda expresada como:

$$\Pr\left(S_i^c(S_j^x)\right) = \frac{\left|S_i^c\left(S_j^x\right)\right|}{\left|S_j^x\right|}$$

3. Se calcula la probabilidad del conjunto S_j^x con respecto al conjunto *S*.

$$\Pr\left(S_j^x\right) = \frac{\left|S_j^x\right|}{|S|}$$

4. Se calcula la entropía del atributo A_x:

$$E(A_x) = \sum_{m=1}^{|D(A_x)|} \Pr(S_m^x) \bullet I(\Pr(S_1^c(S_m^x)), \Pr(S_2^c(S_m^x)), \ldots, \Pr(S_t^c(S_m^x)))$$

donde $t = |D(A_c)|$ y:

$$I(\Pr(S_1^c(S_m^x)), \Pr(S_2^c(S_m^x)), \ldots, \Pr(S_t^c(S_m^x))) = -\sum_{i=1}^{|D(A_c)|} \Pr(S_i^c(S_m^x)) \bullet \log_2 \Pr(S_i^c(S_m^x))$$

La última expresión es conocida como el grado de información que un atributo A_x aporta con respecto al atributo A_c.

5. Se calcula la ganancia de seleccionar el atributo A_x:

$$G(A_x) = I(\Pr(S_1^c), \Pr(S_2^c), \ldots, \Pr(S_t^c)) - E(A_x)$$

6. Para cada atributo $A_x \in$ **A'**, se calcula $G(A_x)$. A partir de este cálculo, se obtendrá el atributo clasificador de la siguiente forma:

$$A_g = \max\{G(A_x): \forall A_x \in \mathbf{A'}\}$$

Tercer Paso

Una vez calculado A_g, ID3 añade un nuevo nodo al árbol que se construye (en caso de ser el primer nodo del árbol, éste pasa a formar la raíz). De forma recursiva, para cada rama derivada del nodo asociado a A_g, se actualiza el conjunto de atributos y de registros, esto es, **A'** ← **A'** – $\{A_g\}$ y $S \leftarrow S_i^g$ (notemos que *S* es diferente para cada una de las ramas correspondientes a los diferentes valores v_i^g derivados del atributo ganador).

Si una rama es analizada y se determina que el conjunto de instancias asociadas a ella es de una sola clase, el proceso de expansión termina, rematándose a la misma con la clase

encontrada y continuando el crecimiento de otra rama hermana. Por tanto, ID3 no termina hasta que no haya desarrollado cada una de las ramas y todas hayan concluido en un nodo donde los registros son de una sola clase ó, cuando **A'** = ∅.

En algunos casos, no es necesario que la totalidad de registros de una rama sea pura (una sola clase) para detener el proceso de crecimiento de la misma, ya que se pueden implementar ciertos criterios de tolerancia, es decir, se remata a una rama con la clase que presente la mayor probabilidad, la cuál sobrepasa un valor establecido (umbral). Lo anterior contribuye a evitar problemas de crecimiento en el árbol, conocidos como *sobre – ajuste*, en donde por más que se expanda una rama, ya no es posible obtener grupos de atributos con una sola clase.

Un aspecto sobre la ganancia que se ha definido es el hecho de favorecer a aquellos atributos donde la cardinalidad de su dominio es cercana al número de atributos totales en *S*. Una solución a este problema es dar una proporción a cada ganancia que se obtiene con respecto al número de atributos que se están manejando en ese momento. Por ello, en trabajos posteriores Quinlan sugirió usar la expresión:

$$G'(A_x) = \frac{I(\Pr(S_1^c), \Pr(S_2^c), ..., \Pr(S_t^c)) - E(A_x)}{I(\Pr(S_1^x), \Pr(S_2^x), ..., \Pr(S_{|A_x|}^x))}$$

REFERENCIAS

[1]. Quinlan, J. R. *Induction of Decision Trees.* Machine Learning. 1:81-106. Kluwer Academic Publishers (1986)