

Döntési fák, FIND-S és ID3 algoritmus

Induktív tanulás

Cél, hogy pozitív és negatív példák alapján olyan logikai függvényt állítsunk elő, ami igaz értéket ad a pozitív és hamisat a negatív példákra és feltételezzük, hogy ha elég nagyszámú példánk van akkor helyes értéket fogunk kapni előre nem látott példák esetén is.

Fogalom tanulása pozitív példák alapján

Tegyük fel, hogy egy használtautó kereskedő a tapasztalatok alapján el szeretné dönteni melyek a jól eladható autók.

Példa	Gyártás helye	Kor	Motor	Szín	Cm ³	Jól eladható
1.	német	3-6	Diesel	fehér	1300-1600	igen
2.	német	3-6	Diesel	piros	1600 felett	igen
3.	japán	6-10	Benzin	fehér	1300-1600	nem
4.	német	3-6	Benzin	kék	1000-1300	igen

Tanulási példák

Gyártás helye	Kor	Motor	Szín	Cm ³
német	10 felett	Diesel	fehér	1000 alatt
japán	6-10	Benzin	piros	1000-1300
olasz	3-6		kék	1300-1600
francia			zöld	1600 felett
			fekete	

Attribútumok lehetséges értékei

Fontos kérdés: milyen alakban keressük a tanulandó hipotézist?

Válasszuk a legegyszerűbb alakot, az attribútumpéldányok konjunkcióját.

$h = (\text{Gyártás helye} = \text{német} \wedge \text{kor} = 3-6 \wedge \text{motor} = \text{diesel} \wedge \text{szín} = \text{piros} \wedge \text{cm}^3 = 1300-1600)$
vektorként megadva

$h = (\text{német}, 3-6, \text{diesel}, \text{piros}, 1300-1600)$

A vektor komponenseinek megengedjük a '?' és a 'Ø' szimbólumokat is. A '?' azt jelöli, hogy az adott komponens bármelyik értéket felveheti, a 'Ø' pedig azt, hogy az attribútum egyetlen értéke sem elfogadható.

Pl.: $h = (\text{német}, 3-6, ?, ?, ?)$

A fogalmi tanulás tulajdonképpen egy keresési feladat, amelyben a példákra legjobban illeszkedő hipotézist keressük.

Mivel nagy számú esetről van szó, fontos a hatékony keresési módszer, ezért célszerű rendezési relációt megadni a hipotézisek között. Tekintsük az alábbi példát:

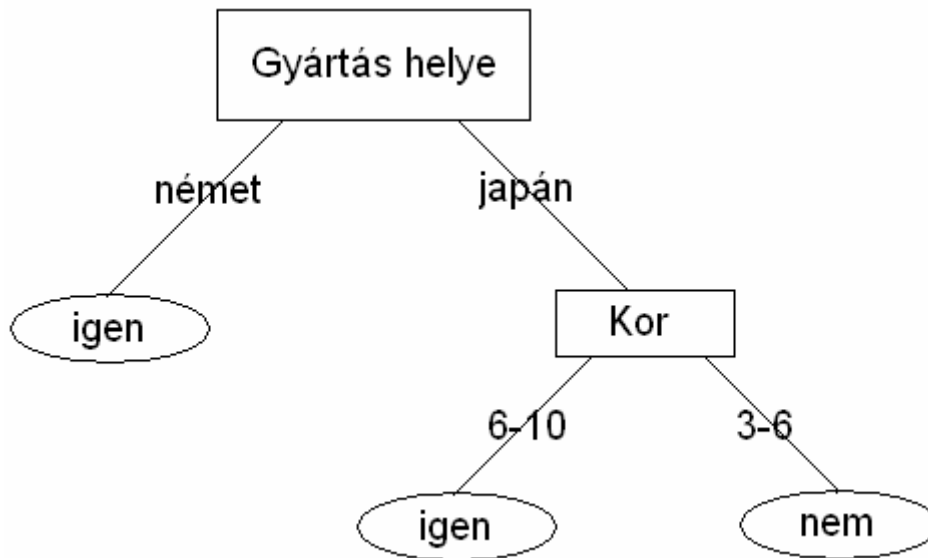
$h_1 = (?, 3-6, ?, ?, ?)$

$h_2 = (?, 3-6, ?, \text{fehér}, ?)$

Látható, hogy h_1 kevesebb korlátozást tartalmaz mint h_2 , tehát h_2 szigorúbb (h_1 általánosabb).

Ennél a példánál a FIND-S fog konzisztens hipotézist találni, mivel a az attribútumértékekre vonatkozó feltételek csak konjunkcióval kapcsolhatók össze. Ha a konjunkció mellett megengedjük a diszjunkció használatát is akkor az alábbi formula konzisztens lesz a táblázat tanulási példáival:

Gyártás helye = német $\vee$ (gyártás helye = japán $\wedge$ kor = 6-10)
 Ezt a formulát reprezentálhatjuk egy ún. döntési fával.



A 'jól eladható' probléma döntési fája

A döntési fa tanuló algoritmusok a legjobban elterjedt, gyakorlati problémák megoldásában is alkalmazott induktív tanulási módszerek. A döntési fa tanulási módszerek könnyen kiterjeszthetők úgy, hogy a megtanult függvénynek több mint két lehetséges értéke legyen. További előnye a módszernek, hogy nem érzékeny a tanulási példák esetleges hibáira.

ID3 algoritmus

A döntési fa minden csomópontjában egy tesztet hajtunk végre valamelyik attribútum értékeire. A döntési fa konstruálásának legfontosabb kérdése az, hogy melyik attribútum értékei szerint végezzük el a tesztet. Fontos, hogy azt az attribútumot válasszuk amelyik a leghasznosabb a példák értékelése szempontjából. Ezen hasznosság mérésére vezetjük be az ún. információs előnyt (information gain). Ez az érték mutatja meg, hogy ha csak egy attribútum szerinti teszt mennyire jól osztályozná az adott tanulási példákat. Az információs előny definiálásához először is szükség lesz az entrópia definíciójára.

Definíció: Legyen S pozitív és negatív példák gyűjteménye egy adott fogalomra vonatkozóan. Ekkor: $\text{Entrópia}(S) = -P_{\text{pos}} \cdot \log_2 P_{\text{pos}} - P_{\text{neg}} \cdot \log_2 P_{\text{neg}}$, ahol P_{pos} a pozitív P_{neg} pedig a negatív példák arányát jelöli S -ben.

Az ID3-ban használt információs előny azt mutatja meg, hogy egy attribútum szerinti osztályozás mennyivel csökkenti az S entrópiáját.

$$\text{Előny}(S, A) = \text{Entrópia}(S) - \sum_{v \in \text{Érték}(A)} (|S_v|/|S|) \cdot \text{Entrópia}(S_v),$$

ahol $\text{Érték}(A)$ az A attribútum lehetséges értékeit jelöli, és S_v az S azon részhalmaza, amelyekre az A attribútum értéke v .

Ezután már definiálható az ID3 algoritmus, amelyet rekurzív eljárással írunk le.

$\text{ID3}(S, A, \text{Gyökér})$

Paraméterek:

$S = S^+ \cup S^-$: Pozitív és negatív példák uniója. Feltesszük hogy $|S| > 0$

A: az attribútumnevek halmaza

Gyökér: egy részfa gyökér csomópontjának tárolására szolgál.

Algoritmus:

Begin

```
  If  $|S^-| = 0$  then
    begin
      Gyökér := Új Csomópont Igen címkével
      RETURN
    end
  If  $|S^+| = 0$  then
    begin
      Gyökér := Új Csomópont Nem címkével
      RETURN
    end
  If  $|A| = 0$  then
    begin
      If  $|S^+| \geq |S^-|$  then  $l = \text{Igen}$  else  $l = \text{Nem}$ 
      Gyökér := Új Csomópont  $l$  címkével
      RETURN
    end
  End
```

Legyen $a \in A$ attribútum, amelyre $\text{Előny}(S, a)$ maximális.

Gyökér := Új Csomópont az 'a' címkével

Minden $a = v_i$ értékre hozzunk létre a Gyökér csomópontból egy ágot. Az ágakat $a = v_i$ jelöléssel címkézzük.

Jelölje S_i azon példák halmazát, amelyekre $a = v_i$, és m a különböző v_i értékek számát.

For $i := 1$ to m do

```
  begin
    if  $|S_{v_i}| = 0$  then
      begin
        if  $|S^+| \geq |S^-|$  then  $l = \text{igen}$ 
        else  $l = \text{nem}$ ;
        Gyökér $a=v_i$  := Új levél csomópont  $l$ 
        címkével ( $a = v_i$  ágra)
      end
    else
      begin
        ID3( $S_{v_i}$ ,  $A - \{a\}$ , Csp);
        Gyökér $a=v_i$  := Csp
        (az  $a = v_i$  ághoz hozzáadjuk a Csp-ban
        visszaadott részfat)
      end
    end
  End
```

Az ID3 szintén nem képes megadni az összes konzisztens hipotézist, illetve másik hátránya, hogy nem alkalmaz visszalépést a keresés közben, azaz nincs lehetőség rá, hogy ha egy adott pontnál kiválasztottunk egy attribútumot, akkor utólag megváltoztassuk a döntést. Az algoritmus lokális optimumok alapján dolgozik, emiatt nem biztos, hogy megtalálja a globálisan optimális megoldás(oka)t, ugyanakkor nagy előnye, hogy minden lépésben statisztikán alapuló döntést hoz, ezáltal sokkal kevésbé érzékeny a tanulási példák egyedi hibáira (zajos tanulás).