VERİ MADENCİLİĞİ

(Karar Ağaçları ile Sınıflandırma)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Adımları
- Veri Madenciliği Yöntemleri
  - Sınıflandırma
  - Kümeleme
  - İlişkilendirme/birliktelik kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

İçerik

- Sınıflandırma yöntemleri
  - Karar ağaçları ile sınıflandırma
    - Entropi Kavramı
    - ID3 Algoritması
    - C4.5 Algoritması

    } Entropiye dayalı algoritmalar

    - Twoing Algoritması
    - Gini Algoritması

    } Sınıflandırma ve regresyon ağaçları (CART)

    - k-en yakın komşu algoritması } Bellek tabanlı algoritmalar

Karar Ağaçları ile Sınıflandırma

- Sınıflandırma problemleri için yaygın kullanılan yöntemdir.
- Sınıflandırma doğruluğu diğer öğrenme metotlarına göre çok etkindir.
- Öğrenmiş sınıflandırma modeli ağaç şeklinde gösterilir ve karar ağacı (decision tree) olarak adlandırılır.
- Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir. Dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı yaprak en üst yapı kök ve bunların arasında kalan yapılar dal olarak isimlendirilir.

Karar Ağaçlarında Dallanma Kriterleri

- Karar ağaçlarında en önemli sorunlardan birisi hangi kökten itibaren bölümlemenin veya dallanmanın hangi kritere göre yapılacağıdır. Aslında her farklı kriter için bir karar ağacı algoritması karşılık gelmektedir.
- Bu algoritmalar şu şekilde gruplandırılabilir.
  - ID3 ve C4.5, entropiye dayalı sınıflandırma algoritmalarıdır.
  - Twoing ve Gini, CART (Classification And Regression Trees) sınıflandırma ve regresyon ağaçlarına dayalı sınıflandırma algoritmalarıdır.
  - k-en yakın komşu algoritması bellek tabanlı sınıflandırma yöntemleri arasında yer almaktadır.

Entropi (1/3)

- Entropi, rastgele değere sahip bir değişken veya bir sistem için belirsizlik ölçütüdür.
- Enformasyon, rassal bir olayın gerçekleşmesi halinde ortaya çıkan bilgi ölçütüdür.
- Bir süreç için entropi, tüm örnekler tarafından içerilen enformasyonun beklenen değeridir.
- Eşit olasıklı durumlara sahip sistemler yüksek belirsizliğe sahiptirler.
- Shannon, bir sistemdeki durum değişikliğinde, entropideki değişimin enformasyon boyutunu tanımladığını öne sürmüştür.
- Buna göre bir sistemdeki belirsizlik arttıkça, bir durum gerçekleştiğinde elde edilecek enformasyon boyutu da artacaktır.

Entropi (2/3)

- Shannon bilgiyi bitlerle ifade ettiği için, logaritmayı 2 tabanında kullanmıştır.

- S bir kaynak olsun. Bu kaynağın $\{m_1, m_2, \dots m_n\}$ olmak üzere n mesaj üretildiğini varsayalım. Tüm mesajlar birbirinden bağımsız üretilmektedir ve m_i mesajlarının üretilme olasılıkları p_i'dir. $P = \{p_1, p_2, \dots p_n\}$ olasılık dağılımına sahip mesajları üreten S kaynağının entropisi $H(S)$ şu şekildedir.

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Entropi (3/3)

- Bir paranın havaya atılması olayı rassal X sürecini göstersin. Yazı ve tura gelme olasılıkları eşit olduğundan elde edilecek entropi,

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

Örnek

- Aşağıdaki 8 elemanlı S kümesi verilsin.
- S = {evet, hayır, evet, hayır, hayır, hayır, hayır, hayır}
- “evet “ ve “hayır” için olasılık,
- $p(evet) = \frac{2}{8}$, $p(hayır) = \frac{6}{8}$

$$H(S) = -\left(\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8}\right) = 0.81128$$

ID3 Algoritması (1/4)

- Karar ağaçları yardımıyla sınıflandırma işlemlerini yerine getirmek üzere Quinlan tarafından birçok algoritma geliştirilmiştir. Bunlar arasında ID3 ve C4.5 algoritması yer almaktadır.
- ID3(Iterative Dichotomiser 3) algoritması sadece *kategorik* verilerle çalışmaktadır.
- Karar ağaçları çok boyutlu veriyi belirlenmiş bir niteliğe göre parçalara böler.
- Her adımda verinin hangi özelliğine göre ne tür işlem yapılacağına karar verilir.
- Oluşturulabilecek tüm ağaçların kombinasyonu çok fazladır.
- Karar ağaçlarının en az düğüm ve yaprak ile oluşturulması için farklı algoritmalar kullanılarak bölme işlemi yapılır.

ID3 Algoritması (2/4)

- Karar Ağacında Entropi
  - Bir eğitim kümesindeki sınıf niteliğinin alacağı değerler kümesi T, her bir sınıf değeri C_i olsun.
  - T sınıf değerini içeren küme için P_T sınıfların olasılık dağılımı,

$$P_T = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \dots, \frac{|C_k|}{|T|} \right)$$

şeklinde ifade edilir.

  - T sınıf kümesi için ortalama entropi değeri ise

$$H(T) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

şeklinde ifade edilir.

ID3 Algoritması (3/4)

- Karar ağaçlarında bölümlemeye hangi düğümden başlanacağı çok önemlidir.
- Uygun düğümden başlanmazsa ağacın içerisindeki düğümlerin ve yaprakların sayısı çok fazla olacaktır.
- Bir risk kümesi aşağıdaki gibi tanımlansın. C_1 ="var", C_2="yok"
  - *RISK = {var, var, var, yok, var, yok, yok, var, var, yok}*

$$|C_1| = 6 \qquad |C_2| = 4 \qquad p_1 = 6/10 = 0{,}6 \qquad p_2 = 4/10 = 0{,}4$$

$$P_{RISK} = \left(\frac{6}{10}, \frac{4}{10}\right)$$

$$H(RISK) = -\sum_{i=1}^{n} p_i \log_2(p_i) = -\left(\frac{6}{10}\log_2\frac{6}{10} + \frac{4}{10}\log_2\frac{4}{10}\right) = 0{,}97$$

ID3 Algoritması (4/4)

- Dallanma için niteliklerin seçimi
  - Öncelikle **sınıf niteliğinin entropisi** hesaplanır.

$$H(T) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

  - Sonra **özellik vektörlerinin sınıfa bağımlı entropileri** hesaplanır.

$$H(X_k) = -\sum_{i=1}^{n} \frac{|T_i|}{|X_k|} \log \frac{|T_i|}{|X_k|} \qquad H(X,T) = \sum_{k=1}^{n} \frac{|X_k|}{|X|} H(X_k)$$

  - Son olarak sınıf niteliğinin entropisinden tüm özellik vektörlerinin entropisi çıkartılarak her özellik için **kazanç ölçütü hesaplanır.**

$$Kazanç(X,T) = H(T) - H(X,T)$$

  - **En büyük kazanca sahip özellik vektörü** o iterasyon için dallanma düğümü olarak seçilir.

Örnek

- Aşağıdaki tablo için karar ağacı oluşturulsun.

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ

$$H(T) = H(RISK) = -\sum_{i=1}^{n} p_i \log_2(p_i) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1$$

Örnek

$$H(BORÇ_{YÜKSEK}) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$$

$$H(BORÇ_{DUSUK}) = -\left(\frac{5}{7}\log_2\frac{5}{7} + \frac{2}{7}\log_2\frac{2}{7}\right) = 0{,}863$$

$$H(BORÇ, RISK) = \frac{3}{10}H(BORÇ_{YÜKSEK}) + \frac{7}{10}H(BORÇ_{DUSUK})$$

$$= \frac{3}{10}(0) + \frac{7}{10}(0{,}863) = 0{,}64$$

$$Kazanç(BORÇ, RISK) = 1 - 0{,}64 = 0{,}36$$

Örnek

$$H(GELIR_{YÜKSEK}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0{,}971$$

$$H(GELIR_{DUSUK}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

$$H(GELIR, RISK) = \frac{5}{10}H(GELIR_{YÜKSEK}) + \frac{5}{10}H(GELIR_{DUSUK})$$

$$= \frac{5}{10}(0{,}971) + \frac{5}{10}(0{,}971) = 0{,}971$$

$$Kazanç(GELIR, RISK) = 1 - 0{,}971 = 0{,}029$$

Örnek

$$H(STATU_{ISVEREN}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

$$H(STATU_{DUSUK}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

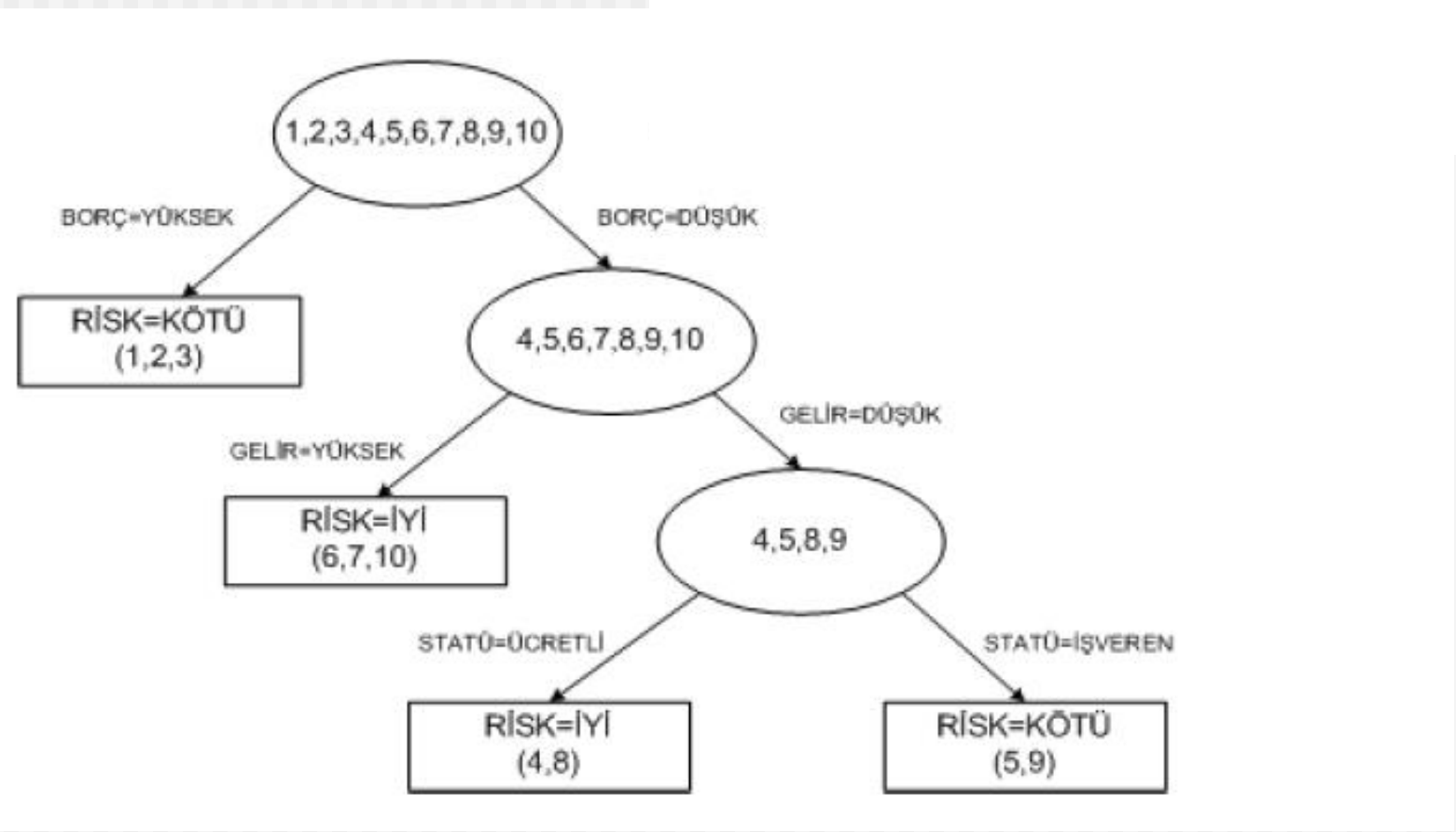
$$H(STATU, RISK) = \frac{5}{10}H(STATU_{YÜKSEK}) + \frac{5}{10}H(STATU_{DUSUK})$$

$$= \frac{5}{10}(0{,}971) + \frac{5}{10}(0{,}971) = 0{,}971$$

$$Kazanç(STATU, RISK) = 1 - 0{,}971 = 0{,}029$$

İlk dallanma için uygun seçim BORÇ niteliğidir.

Örnek


1,2,3,4,5,6,7,8,9,10
BORÇ=YÜKSEK
BORÇ=DÜŞÜK
RİSK=KÖTÜ
(1,2,3)
4,5,6,7,8,9,10
GELİR=YÜKSEK
GELİR=DÜŞÜK
RİSK=İYİ
(6,7,10)
4,5,8,9
STATÜ=ÜCRETLİ
STATÜ=İŞVEREN
RİSK=İYİ
(4,8)
RİSK=KÖTÜ
(5,9)

Örnek

- Karar ağacından elde edilen kurallar
- **1.EĞER**(BORÇ = YÜKSEK) **İSE** (RİSK = KÖTÜ)
- **2.EĞER**(BORÇ = DÜŞÜK) **VE** (GELİR = YÜKSEK) **İSE** (RİSK = İYİ)
- **3.EĞER**(BORÇ = DÜŞÜK) **VE** (GELİR = DÜŞÜK) **VE** (STATÜ = ÜCRETLİ) **İSE** (RİSK = İYİ)
- **4.EĞER**(BORÇ = DÜŞÜK) **VE** (GELİR = DÜŞÜK) **VE** (STATÜ = İŞVEREN) **İSE**(RİSK = KÖTÜ)

Uygulama: Hava problemi örneği

Eğitim kümesi				
HAVA	ISI	NEM	RÜZGAR	OYUN
güneşli	sıcak	yüksek	hafif	Hayır
güneşli	sıcak	yüksek	kuvvetli	Hayır
bulutlu	sıcak	yüksek	hafif	Evet
yağmurlu	ılık	yüksek	hafif	Evet
yağmurlu	soğuk	normal	hafif	Evet
yağmurlu	soğuk	normal	kuvvetli	Hayır
bulutlu	soğuk	normal	kuvvetli	Evet
güneş!i	ılık	yüksek	hafif	Hayır
güneşli	soğuk	normal	hafif	Evet
yağmurlu	ılık	normal	hafif	Evet
güneşli	ılık	normal	kuvvetli	Evet
bulutlu	ılık	yüksek	kuvvetli	Evet
bulutlu	sıcak	normal	hafif	Evet
yağmurlu	ılık	yüksek	kuvvetli	Hayır

Uygulama: Hava problemi

- *OYUN = {hayır, hayır, hayır, hayır, hayır, evet, evet, evet, evet, evet, evet, evet, evet, evet}*
- C1, sınıfı **"hayır",** C2, sınıfı ise **"evet"**
- P1=5/14, P2=9/14

$$H(OYUN) = -\left(\frac{5}{14}log_2\frac{5}{14} + \frac{9}{14}log_2\frac{9}{14}\right) = 0.940$$

Adım1: Birinci dallanma

ISI niteliği için kazanç ölçütü:

ISI	OYUN
soğuk	evet
soğuk	hayır
soğuk	evet
soğuk	evet
ılık	evet
ılık	hayır
ılık	evet
ılık	evet
ılık	evet
ılık	hayır
sıcak	hayır
sıcak	hayır
sıcak	evet
sıcak	evet

$$|ISI_{soğuk}| = 4$$

$$|ISI_{ılık}| = 6$$

$$|ISI_{sıcak}| = 4$$

$$H(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i)$$

$$H(ISI, OYUN) = \frac{4}{14} H(ISI_{soğuk}) + \frac{6}{14} H(ISI_{ılık}) + \frac{4}{14} H(ISI_{sıcak})$$

$$H(ISI_{soğuk}) = -\left(\frac{1}{4} log_2 \frac{1}{4} + \frac{3}{4} log_2 \frac{3}{4}\right) = 0.811$$

$$H(ISI_{ılık}) = -\left(\frac{2}{6} log_2 \frac{2}{6} + \frac{4}{6} log_2 \frac{4}{6}\right) = 0.918$$

$$H(ISI_{sıcak}) = -\left(\frac{2}{4} log_2 \frac{2}{4} + \frac{2}{4} log_2 \frac{2}{4}\right) = 1.00$$

$$H(ISI, OYUN) = \frac{4}{14}(0.811) + \frac{6}{14}(0.918) + \frac{4}{14}(1.00) = 0.911$$

$$Kazanç(ISI,OYUN) = H(OYUN) - H(ISI,OYUN)$$
$$= 0.940 - 0.911 = 0.029$$

Adım1: Birinci dallanma

HAVA niteliği için kazanç ölçütü:

$$|HAVA_{güneşli}| = 5 \qquad |HAVA_{yağmurlu}| = 5 \qquad |HAVA_{bulutlu}| = 4$$

$$H(HAVA, OYUN) = \frac{5}{14}H(HAVA_{güneşli}) + \frac{4}{14}H(HAVA_{bulutlu}) + \frac{5}{14}H(HAVA_{yağmurlu})$$

$$H(HAVA_{güneşli}) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.971$$

$$H(HAVA_{yağmurlu}) = -\left(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}\right) = 0.971$$

$$H(HAVA_{bulutlu}) = -\left(\frac{4}{4}log_2\frac{4}{4}\right) = 0$$

HAVA	OYUN
güneşli	hayır
güneşli	hayır
güneşli	hayır
güneşli	evet
güneşli	evet
yağmurlu	evet
yağmurlu	evet
yağmurlu	hayır
yağmurlu	evet
yağmurlu	hayır
bulutlu	evet
bulutlu	evet
bulutlu	evet
bulutlu	evet

$$H(HAVA, OYUN) = \frac{5}{14}H(HAVA_{güneşli}) + \frac{4}{14}H(HAVA_{bulutlu}) + \frac{5}{14}H(HAVA_{yağmurlu})$$

$$H(HAVA, OYUN) = \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971) = 0.694$$

$$Kazanç(HAVA,OYUN) = H(OYUN) - H(HAVA,OYUN)$$
$$= 0.940 - 0.693 = 0.247$$

Adım1: Birinci dallanma

NEM niteliği için kazanç ölçütü:

NEM	OYUN
yüksek	hayır
yüksek	hayır
yüksek	evet
y üksek	evet
yüksek	hayır
yüksek	evet
yüksek	hayır
normal	evet
normal	hayır
normal	evet
norma!	evet
normal	evet
normal	evet
normal	evet

$$|NEM_{yüksek}| = 7$$

$$|NEM_{normal}| = 7$$

$$H(NEM, OYUN) = \frac{7}{14}H(NEM_{yüksek}) + \frac{7}{14}H(NEM_{normal})$$

$$H(NEM_{yüksek}) = -\left(\frac{4}{7}log_2\frac{4}{7} + \frac{3}{7}log_2\frac{3}{7}\right) = 0.985$$

$$H(NEM_{normal}) = -\left(\frac{1}{7}log_2\frac{1}{7} + \frac{6}{7}log_2\frac{6}{7}\right) = 0.592$$

$$H(NEM, OYUN) = \frac{7}{14}H(NEM_{yüksek}) + \frac{7}{14}H(NEM_{normal})$$

$$H(NEM, OYUN) = \frac{7}{14}(0.985) + \frac{7}{14}(0.592) = 0.789$$

$$Kazanç(NEM,OYUN)=H(OYUN)-H(NEM,OYUN)$$

$$=0.940-0.789=0.151$$

Adım1: Birinci dallanma

RÜZGAR niteliği için kazanç ölçütü:

RÜZGAR	OYUN
hafif	hayır
hafif	evet
hafif	evet
hafif	evet
hafif	hayır
hafif	evet
hafif	evet
hafif	evet
kuvvetli	hayır
kuvvetli	hayır
kuvvetli	evet
kuvvetli	evet
kuvvetli	evet
kuvvetli	hayır

$$|RÜZGAR_{hafif}| = 8$$

$$|RÜZGAR_{kuvvetli}| = 6$$

$$H(RÜZGAR, OYUN) = \frac{8}{14}H(RÜZGAR_{hafif}) + \frac{6}{14}H(RÜZGAR_{kuvvetli})$$

$$H(RÜZGAR_{hafif}) = -\left(\frac{2}{8}log_2\frac{2}{8} + \frac{6}{8}log_2\frac{6}{8}\right) = 0.811$$

$$H(RÜZGAR_{kuvvetli}) = -\left(\frac{3}{6}log_2\frac{3}{6} + \frac{3}{6}log_2\frac{3}{6}\right) = 1.00$$

$$H(RÜZGAR, OYUN) = \frac{8}{14}H(RÜZGAR_{hafif}) + \frac{6}{14}H(RÜZGAR_{kuvvetli})$$

$$H(RÜZGAR, OYUN) = \frac{8}{14}(0.811) + \frac{6}{14}(1.00) = 0.892$$

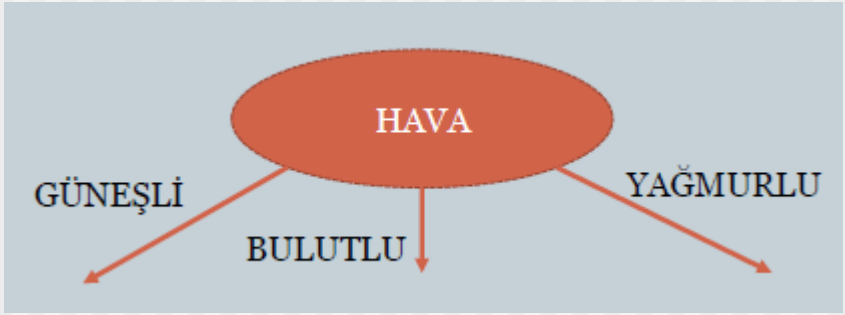
$$Kazanç(RÜZGAR,OYUN)=H(OYUN)-H(,OYUN)$$

$$=0.940-0.892=0.048$$

Nitelik	Kazanç
HAVA	0.246
ISI	0.029
NEM	0.151
RÜZGAR	0.048

Adım1: Birinci dallanma

- Birinci dallanma sonucu karar ağacı:



Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

HAVA=güneşli için gözlem değerleri				
HAVA	ISI	NEM	RÜZGAR	OYUN
güneşli	sıcak	yüksek	hafif	hayır
güneşli	sıcak	yüksek	kuvvetli	hayır
güneşli	ılık	yüksek	hafif	hayır
güneşli	soğuk	normal	hafif	evet
güneşli	ılık	normal	kuvvetli	evet

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

- Oyun için entropi:

$$H(OYUN) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.970$$

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

ISI niteliği için kazanç ölçütü:

$$|ISI_{soğuk}| = 1$$

$$H(ISI_{soğuk}) = -\left(\frac{1}{1}log_2\frac{1}{1}\right) = 0$$

$$H(ISI_{sıcak}) = -\left(\frac{2}{2}log_2\frac{2}{2}\right) = 0$$

$$H(ISI_{ılık}) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1$$

$$H(ISI, OYUN) = \frac{1}{5}(0) + \frac{1}{5}(0) + \frac{1}{5}(1) = 0.4$$

$$Kazanç(ISI, OYUN) = H(OYUN) - H(ISI, OYUN) = 0.970 - 0.4 = 0.570$$

ISI	OYUN
soğuk	evet
sıcak	hayır
sıcak	hayır
ılık	hayır
ılık	evet

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

NEM niteliği için kazanç ölçütü:

$$H(NEM_{yüksek}) = -\left(\frac{3}{3}log_2\frac{3}{3}\right) = 0$$

$$H(NEM_{normal}) = -\left(\frac{2}{2}log_2\frac{2}{2}\right) = 0$$

$$H(NEM, OYUN) = \frac{3}{5}(0) + \frac{2}{5}(0) = 0$$

NEM	OYUN
yüksek	hayır
yüksek	hayır
yüksek	hayır
normal	evet
normal	evet

$$Kazanç(NEM, OYUN) = H(OYUN) - H(NEM, OYUN) = 0.970 - 0 = 0.970$$

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

RÜZGAR niteliği için kazanç ölçütü:

$$H(RÜZGAR_{hafif}) = -\left(\frac{2}{3}log_2\frac{2}{3} + \frac{1}{3}log_2\frac{1}{3}\right) = 0.918$$

$$H(RÜZGAR_{kuvvetli}) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1$$

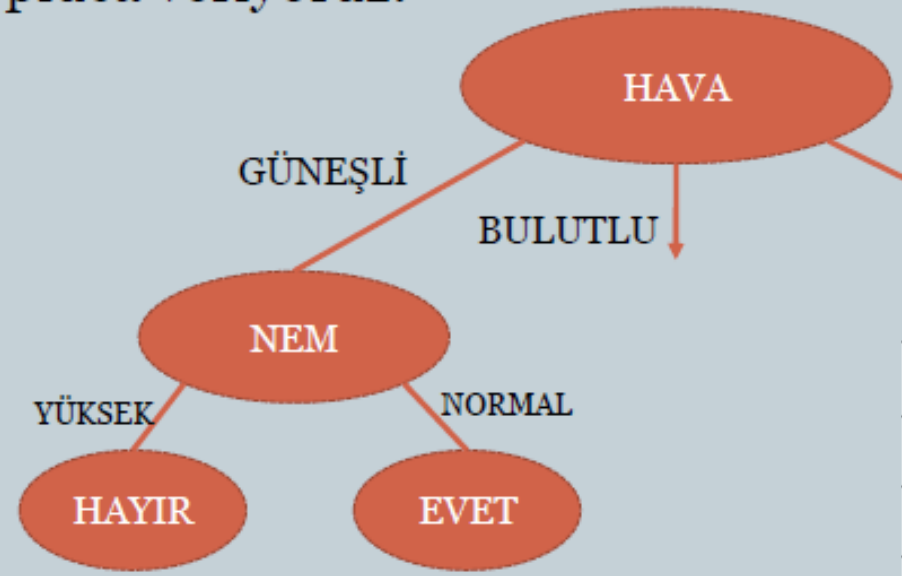
$$H(RÜZGAR, OYUN) = \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.951$$

RÜZGAR	OYUN
hafif	hayır
hafif	hayır
hafif	evet
kuvvetli	hayır
kuvvetli	evet

$$Kazanç(RÜZGAR, OYUN) = H(OYUN) - H(RÜZGAR, OYUN) = 0.970 - 0.951 = 0.019$$

Adım 2: HAVA niteliğinin "güneşli" değeri için dallanma

Elde edilen kazanç ölçütlerini aşağıdaki tabloda topluca veriyoruz:



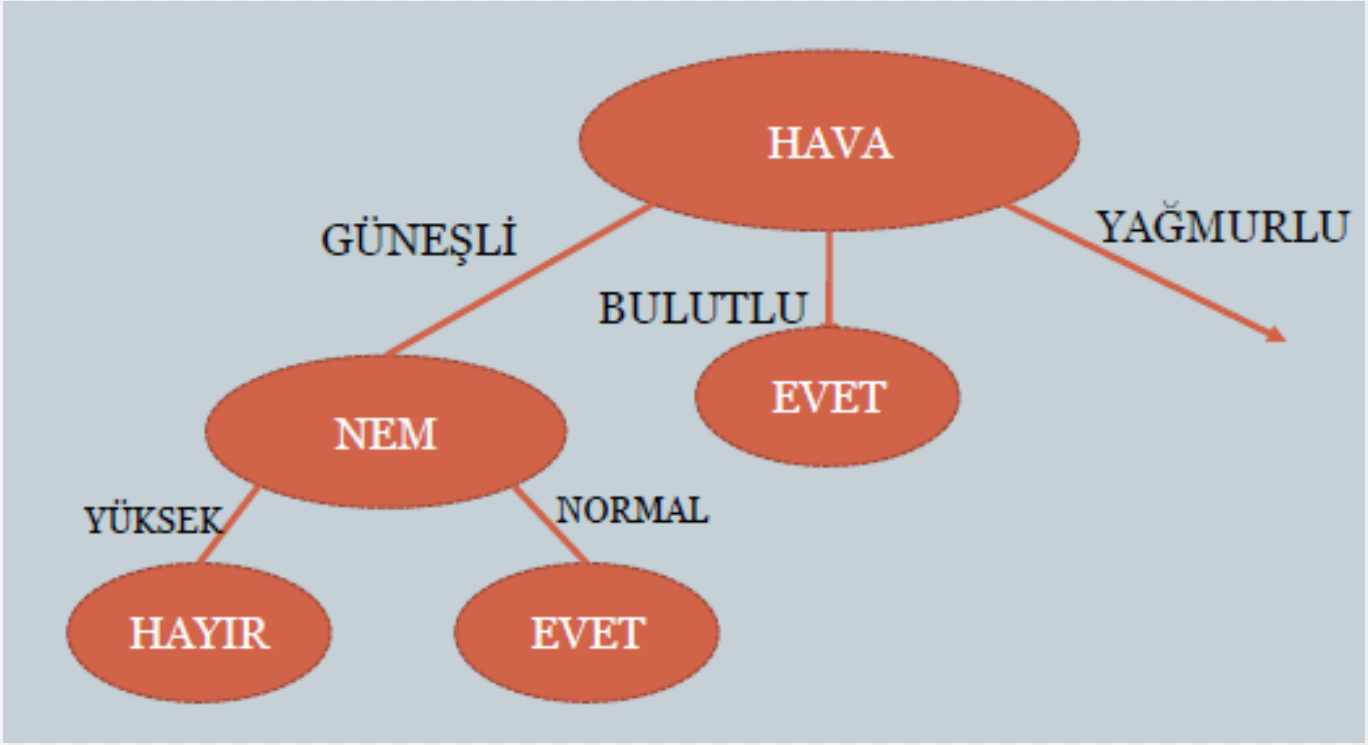
Nitelik	Kazanç
ISI	0.570
NEM	0.970
RÜZGAR	0.019

Adım 3: HAVA niteliğinin "bulutlu" değeri için dallanma:

Görüldüğü gibi tüm karar değerleri **"evet"** olduğu için herhangi bir analize gerek yoktur.

HAVA	ISI	NEM	RÜZGAR	OYUN
bulutlu	sıcak	yüksek	hafif	evet
bulutlu	soğuk	normal	kuvvetli	evet
bulutlu	ılık	yüksek	kuvvetli	evet
bulutlu	sıcak	normal	hafif	evet

Adım 3: HAVA niteliğinin "bulutlu" değeri için dallanma:



Adım 3:HAVA niteliğinin "yağmurlu" değeri için dallanma:

OYUN için entropi:

HAVA	ISI	NEM	RÜZGAR	OYUN
yağmurlu	ılık	yüksek	hafif	evet
yağmurlu	soğuk	normal	hafif	evet
yağmurlu	soğuk	normal	kuvvetli	hayır
yağmurlu	ılık	normal	hafif	evet
yağmurlu	ılık	yüksek	kuvvetli	hayır

$$H(OYUN) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.970$$

Adım 3:HAVA niteliğinin "yağmurlu" değeri için dallanma:

ISI niteliği için kazanç ölçütü:

ISI	OYUN
soğuk	evet
soğuk	hayır
ılık	evet
ılık	evet
Ilık	hayır

$$|ISI_{soğuk}| = 2 \qquad |ISI_{ılık}| = 3$$

$$H(ISI_{soğuk}) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1$$

$$H(ISI_{ılık}) = -\left(\frac{2}{3}log_2\frac{2}{3} + \frac{1}{3}log_2\frac{1}{3}\right) = 0.918$$

$$H(ISI, OYUN) = \frac{2}{5}(1) + \frac{3}{5}(0.918) = 0.951$$

$$Kazanç(ISI, OYUN) = H(OYUN) - H(ISI, OYUN) = 0.970 - 0.951 = 0.019$$

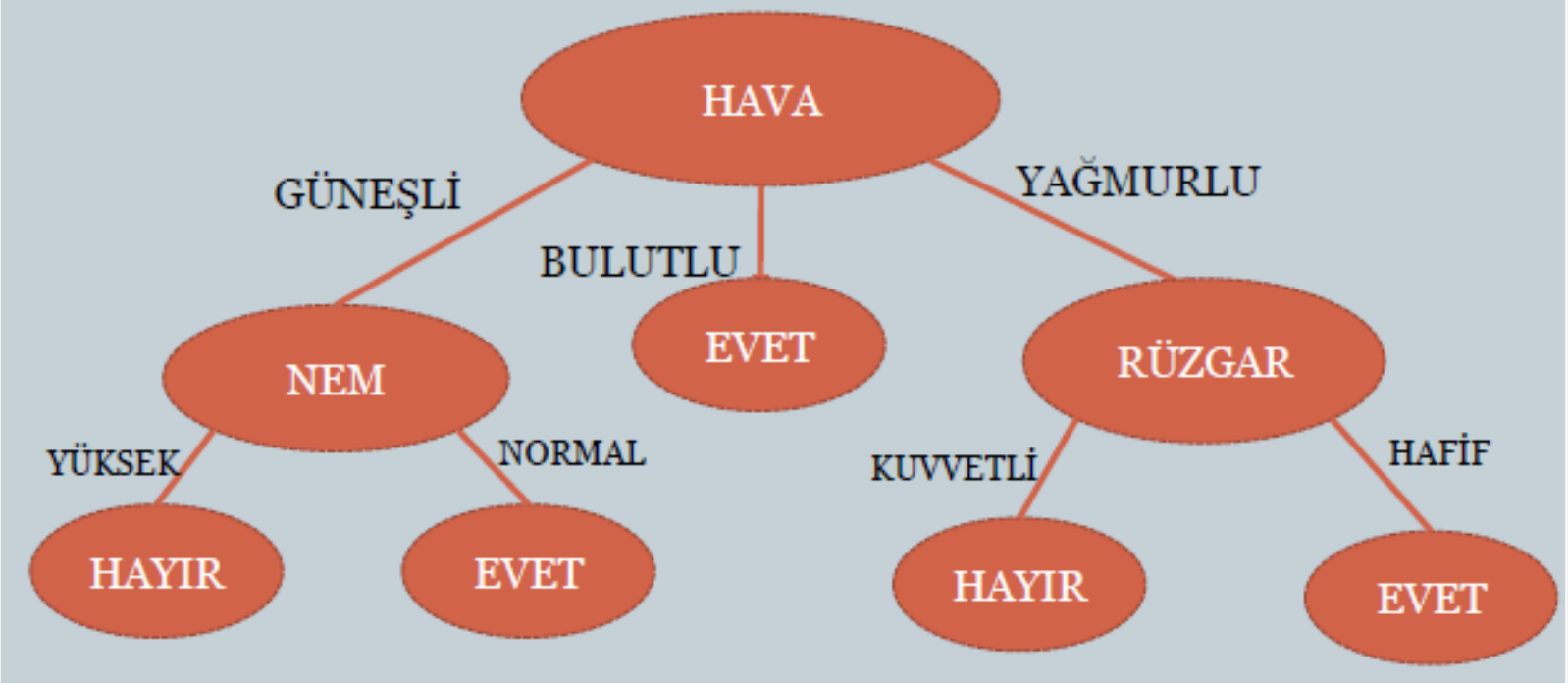
Adım 3:HAVA niteliğinin "yağmurlu" değeri için dallanma:

RÜZGAR niteliği için kazanç ölçütü:

$|RÜZGAR_{hafif}| = 3 \quad |RÜZGAR_{güçlü}| = 2$

RÜZGAR	OYUN
hafif	evet
hafif	evet
hafif	evet
kuvvetli	hayır
kuvvetli	hayır

Oluşturulan Karar Ağacı



C4.5 Algoritması

- C4.5 ile sayısal değerlere sahip nitelikler için karar ağacı oluşturmak için Quinlan tarafından geliştirilmiştir.
- ID3 algoritmasından tek farkı nümerik değerlerin kategorik değerler haline dönüştürülmesidir.
- En büyük bilgi kazancını sağlayacak biçimde bir eşik değer belirlenir.
- Eşik değeri belirlemek için tüm değerler sıralanır ve ikiye bölünür.
- Eşik değer için $[v_i, v_{i+1}]$aralığının orta noktası alınabilir.

$$t_i = \frac{v_i + v_{i+1}}{2}$$

- Nitelikteki değerler eşik değere göre iki kategoriye ayrılmış olur.

Örnek

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	eşit veya küçük	doğru	sınıf1
a	büyük	doğru	sınıf2
a	büyük	yanlış	sınıf2
a	büyük	yanlış	sınıf2
a	eşit veya küçük	yanlış	sınıf1
b	büyük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
b	eşit veya küçük	doğru	sınıf1
b	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	doğru	sınıf2
c	eşit veya küçük	yanlış	sınıf1
c	eşit veya küçük	yanlış	sınıf1
c	büyük	yanlış	sınıf1

Tabloda örneğe ait eğitim kümesi ele alındığında sayısal değerlere sahip olan NİTELİK2 niteliğinin seçilmesi durumunda bilgi kazancının bulunması istenmektedir.

Örnek

Eşik değerinin belirlenmesi

- Nitelik 2 = {65, 70, 75, **80, 85**, 90, 95, 96} için eşik değer **(80+85)/2 = 83** alınmıştır.

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	doğru	sınıf1
a	90	doğru	sınıf2
a	85	yanlış	sınıf2
a	95	yanlış	sınıf2
a	70	yanlış	sınıf1
b	90	doğru	sınıf1
b	78	yanlış	sınıf1
b	65	doğru	sınıf1
b	75	yanlış	sınıf1
c	80	doğru	sınıf2
c	70	doğru	sınıf2
c	80	yanlış	sınıf1
c	70	yanlış	sınıf1
c	96	yanlış	sınıf1

NİTELİK2≤ 83 veya NİTELİK2>83 testi uygulanarak düzenleme yapıldığında yandaki tablo elde edilir.

Örnek

$$H(SINIF) = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right) = 0{,}940$$

$$H(NITELIK1_a) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0{,}971$$

$$H(NITELIK1_b) = -\left(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right) = 0$$

$$H(NITELIK1_c) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0{,}971$$

Entropi değerleri ve Bilgi kazancı hesaplanır

$$H(NITELIK1, SINIF) = \frac{5}{14}H(NITELIK1_a) + \frac{4}{14}H(NITELIK1_b) + \frac{5}{14}H(NITELIK1_c)$$

$$= \frac{5}{14}0{,}971 + \frac{4}{14}0 + \frac{5}{14}0{,}971 = 0{,}694$$

$$Kazanç(NITELIK1, SINIF) = 0{,}940 - 0{,}694 = 0{,}246$$

Örnek

$$H(NITELIK2_{ek}) = -\left(\frac{7}{9}\log_2\frac{7}{9} + \frac{2}{9}\log_2\frac{2}{9}\right) = 0{,}765$$

$$H(NITELIK2_{b}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0{,}971$$

$$H(NITELIK2, SINIF) = \frac{9}{14}H(NITELIK2_{ek}) + \frac{5}{14}H(NITELIK1_{b})$$

$$= \frac{9}{14}0{,}765 + \frac{5}{14}0{,}971 = 0{,}836$$

$$Kazanc(NITELIK2, SINIF) = 0.940 - 0.836 = 0.104$$

Örnek

$$H(NITELIK3_d) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1$$

$$H(NITELIK3_y) = -\left(\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right) = 0{,}811$$

$$H(NITELIK3, SINIF) = \frac{6}{14}H(NITELIK3_d) + \frac{8}{14}H(NITELIK3_y)$$

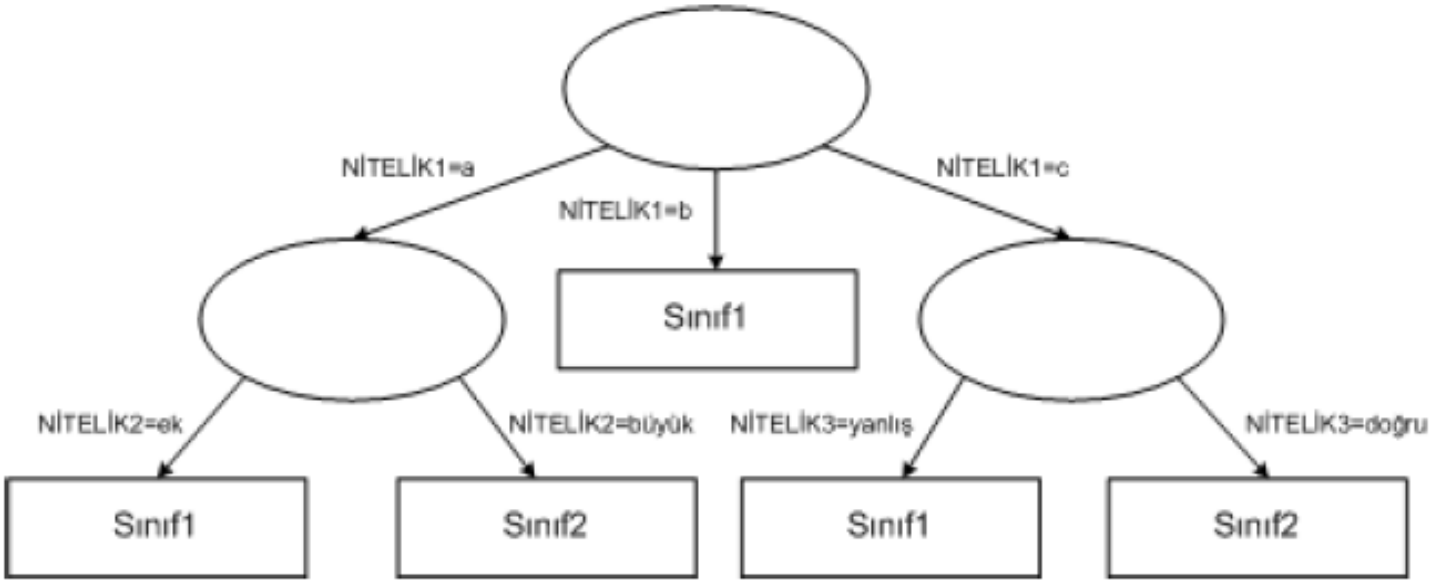
$$= \frac{6}{14}1 + \frac{8}{14}0{,}811 = 0{,}892$$

$$Kazanç(NITELIK\,3, SINIF) = 0{,}940 - 0{,}892 = 0{,}048$$

$$Kazanç(NITELIK\,3, SINIF) < Kazanç(NITELIK\,2, SINIF) < Kazanç(NITELIK1, SINIF)$$

Örnek

Oluşturulan karar ağacı



Örnek

- Karar ağacından elde edilen kurallar
- **1.EĞER**(NİTELİK1 = a) **VE**(NİTELİK2 = Eşit veya Küçük) **İSE**(SINIF = Sınıf1)
- **2.EĞER**(NİTELİK1 = a) **VE**(NİTELİK2 = Büyük) **İSE**(SINIF = Sınıf2)
- **3.EĞER**(NİTELİK1 = b) **İSE**(SINIF = Sınıf1)
- **4.EĞER**(NİTELİK1 = c) **VE**(NİTELİK3 = yanlış) **İSE**(SINIF = Sınıf1)
- **5.EĞER**(NİTELİK1 = c) **VE**(NİTELİK3 = doğru) **İSE**(SINIF = Sınıf2)