SAKARYA ÜNİVERSİTESİ

Veri Madenciliği Uygulamaları

Hafta 6

Yrd. Doç.Dr. Nilüfer YURTAY

Sınıflandırma- İstatistiğe dayalı algoritmalar (Regresyon Ağaçları)

6.1 Giriş

Verinin içerdiği ortak özelliklere göre ayrıştırılması işlemi sınıflandırma olarak anılır.

Karar ağaçları sınıflandırma yöntemlerinden biridir.

Bilimsel çalışmalardan elde edilen verilerin analizinde sınıflama ve regresyon modelleri sıkça kullanılmaktadır. Ancak bu tür modellerin gerektirdiği varsayım gerektirmemesi nedeniyle , sınıflama ve regresyon ağaçları (CART) bu tür istatistiksel sınıflama ve regresyon tekniklerine karşı güçlü bir alternatif olarak ortaya çıkmaktadır.

Veri setinin çok karmaşık olduğu durumlarda bile CART , bağımlı değişkeni etkileyen değişkenleri ve bu değişkenlerin modeldeki önemini basit bir ağaç yapısı ile görsel olarak sunabilmektedir.

Gerek tanımlayıcı, gerekse tahmin edici modellerde yoğun olarak kullanılan belli başlı istatistiksel yöntemler;

1) Sınıflama (Classification) ve regresyon (Regression) ,
2) Kümeleme (Clustering),
3) Birliktelik Kuralları (Association Rules)
4) Ardışık Zamanlı Örüntüler (Sequential Patterns) ,
5) Bellek tabanlı yöntemler ,
6) Yapay sinir ağları ve karar ağaçları

olmak üzere **altı** ana başlık altında incelemek mümkündür.

Sınıflama ve regresyon modelleri

1. **Tahmin edici ,**
2. **Kümeleme ,**
3. **Birliktelik kuralları** ,
4. **Ardışık zamanlı örüntü** tanımlayıcı modellerdir.

6.2 Sınıflama ve Regresyon Modelleri

Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve veri madenciliği teknikleri içerisinde en çok kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin **kategorik** veya **süreklilik gösteren** bir değere sahip olmasıdır.

Ancak çok terimli lojistik regresyon (multinominal logistic regression) gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerle, her iki model giderek birbirine yaklaşmakta ve bunun bir sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır.

Sınıflama kategorik değerleri tahmin ederken, **regresyon** süreklilik gösteren değerlerin tahmin edilmesinde kullanılır.Örneğin bir **sınıflama modeli** banka kredi uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, **regresyon modeli** geliri ve mesleği verilen

potansiyel müşterilerin bilgisayar ürünleri alırken yapacakları harcamaları tahmin etmek için kurulabilir.

Karar ağaçları, veri madenciliğinde kuruluşlarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir.

Karar ağacı, adından da anlaşılacağı gibi bir ağaç görünümünde, tahmin edici bir tekniktir

Ağaç yapısı ile, kolay anlaşılabilen kurallar oluşturan, bilgi teknolojileri işlemleri ile kolay entegre olabilen en popüler sınıflama tekniğidir.

Karar düğümü, gerçekleştirilecek testi belirtir. Bu testin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olur. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşir ve bu ayrılma işlemi üst seviyedeki ayrımlara bağımlıdır.

Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır.

Karar ağacı tekniğini kullanarak verinin sınıflanması iki basamaklı bir işlemdir**:**

İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması tarafından analiz edilir. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir.

İkinci basamak ise sınıflama basamağıdır. Sınıflama basamağında test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır.

Karar ağaçlarına ait bazı kullanım alanları;

- Hangi demografik grupların mektupla yapılan pazarlama uygulamalarında yüksek cevaplama oranına sahip olduğunun belirlenmesi (Direct Mail),
- Bireylerin kredi geçmişlerini kullanarak kredi kararlarının verilmesi (Credit Scoring),
- Geçmişte işletmeye en faydalı olan bireylerin özelliklerini kullanarak işe alma süreçlerinin belirlenmesi,
- Tıbbi gözlem verilerinden yararlanarak en etkin kararların verilmesi,
- Hangi değişkenlerin satışları etkilediğinin belirlenmesi, üretim verilerini inceleyerek ürün hatalarına yol açan değişkenlerin belirlenmesi .

olarak sıralanabilir.

6.3 Gini Algoritması

İkili bölünmeler şeklinde gerçekleşen bir sınıflandırma yöntemidir.

Gini , ikili yinelemeli ,bölümleme için en iyi bilinen kurallardandır. Çünkü her kural karar ağacı amacı olarak, farklı bir felsefeyi temsil eder.

Her bir ağaç farklı bir stil ile gelişir.
Algoritma nitelik değerlerinin sol ve sağda olmak üzere ikili bölünmeler şeklinde ayrılması temeline dayanır.

<u>Uygulama Şekli :</u>

- ✓ Nitelik değerlerinin her biri ikili bölünmeler olacak şekilde sınıflanır. Elde edilen sol ve sağ bölünmelere karşılık gelen sınıf değerleri gruplandırılır.

- ✓ Her bir düğümde ilgili sol ve sağ bölünmeler için ayrı ayrı hesaplamalar yapılır.

Herbir nitelikle ilgili sol ve sağ bölünmeler için $Gini_{sol}$ ve $Gini_{sağ}$ ifadeleri

L_i sol daldaki i grubundaki örnek(lerin) sayısı,

R_i sağ daldaki i grubundaki örnek(lerin) sayısı,

k sınıfların sayısı,

T düğümdeki örnekler

$|T_{sol}|$ Sol daldaki örnek(lerin) sayısı,

$|T_{sağ}|$ Sağ daldaki örnek(lerin) sayısı.

tanımlamaları ile aşağıdaki bağıntılar hesaplanabilecektir.

$$Gini_{Sol} = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{|T_{Sol}|} \right)^2 \qquad Gini_{Sağ} = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{|T_{Sağ}|} \right)^2$$

Her bir j niteliği için n eğitim kümesindeki kayıt sayısı olmak üzere aşağıdaki bağıntı hesaplanır.

$$Gini_j = \frac{1}{n}\left(|T_{Sol}|\,Gini_{Sol} + |T_{Sağ}|\,Gini_{Sağ}\right)$$

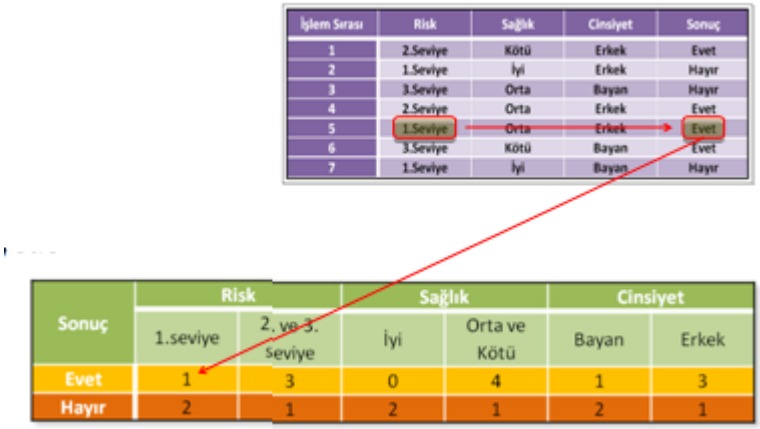
Her j niteliği için hesaplanan $\mathbf{Gini_j}$ ifadelerinden en küçük olanı seçilir ve bölünme bu nitelik üzerinden yapılır.

Sonraki aşamada işlemler tekrar edilir.

Örnek 6.1

Aşağıdaki öğrenme verilerine göre Gini algoritması yardımıyla sınıflandırma işlemi yapalım.

İşlem Sırası	Risk	Sağlık	Cinsiyet	Sonuç
1	2.Seviye	Kötü	Erkek	Evet
2	1.Seviye	İyi	Erkek	Hayır
3	3.Seviye	Orta	Bayan	Hayır
4	2.Seviye	Orta	Erkek	Evet
5	1.Seviye	Orta	Erkek	Evet
6	3.Seviye	Kötü	Bayan	Evet
7	1.Seviye	İyi	Bayan	Hayır



Sonuç	Risk		Sağlık		Cinsiyet	
	1.seviye	2. ve 3. seviye	İyi	Orta ve Kötü	Bayan	Erkek
Evet	1	3	0	4	1	3
Hayır	2	1	2	1	2	1

tablosu elde edilir.

Risk niteliği için

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.44$$

$$Gini_{sağ} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.37$$

Sağlık niteliği için;

Sonuç	Risk		Sağlık		Cinsiyet	
	1.seviye	2. ve 3. seviye	İyi	Orta ve Kötü	Bayan	Erkek
Evet	1	3	0	4	1	3
Hayır	2	1	2	1	2	1

$$Gini_{sol} = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$Gini_{sağ} = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0.32$$

Sonuç	Risk		Sağlık		Cinsiyet	
	1.seviye	2. ve 3. seviye	İyi	Orta ve Kötü	Bayan	Erkek
Evet	1	3	0	4	1	3
Hayır	2	1	2	1	2	1

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.44$$

$$Gini_{sağ} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.37$$

$$\mathbf{Gini_{risk}} = \frac{3(0.44)+4(0.37)}{7} = \mathbf{0.40}$$

$$\mathbf{Gini_{sağlık}} = \frac{2(0)+5(0.320)}{7} = \mathbf{0.22}$$

$$\mathbf{Gini_{cinsiyet}} = \frac{3(0.44)+4(0.37)}{7} = \mathbf{0.40}$$

değerleri bulunur. Bu değerler aşağıdaki tabloda özetlenmiştir.

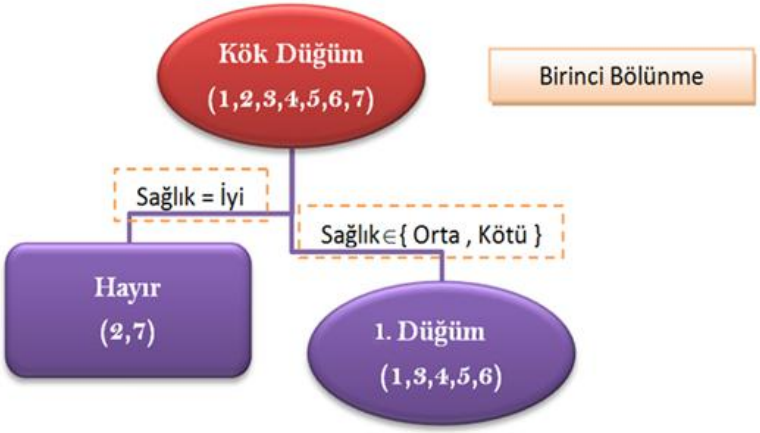
Kabul	Risk		Sağlık		Cinsiyet	
	1.Sev.	2. ve 3.Seviye	İyi	Orta ve Kötü	Bayan	Erkek
EVET	1	3	0	4	1	3
HAYIR	2	1	2	1	2	1
$Gini_{sol}$, $Gini_{sağ}$	0.44	0.37	0.00	0.32	0.44	0.37
$Gini_j$	0.40		0.22		0.40	

İlk bölünme en küçük gini değeri nedeni ile sağlık alanında olup (iyi) ve (orta-kötü) biçiminde olacaktır. Tablodan sağlık-iyi değerleri arandığında

İşlem Sırası	Risk	Sağlık	Cinsiyet	Sonuç
1	2.Seviye	Kötü	Erkek	Evet
2	1.Seviye	İyi	Erkek	Hayır
3	3.Seviye	Orta	Bayan	Hayır
4	2.Seviye	Orta	Erkek	Evet
5	1.Seviye	Orta	Erkek	Evet
6	3.Seviye	Kötü	Bayan	Evet
7	1.Seviye	İyi	Bayan	Hayır

olarak tespit edilir.

Dolayısıyla ilk bölünme 2-7 ve 1-3-4-5-6 biçiminde oalcaktır. Ağaç da



biçimde başlar.

İlk tablodan 2 ve 7.kayıtlar çıkarılırsa

İşlem Sırası	Risk	Sağlık	Cinsiyet	Sonuç
1	2.Seviye	Kötü	Erkek	Evet
3	3.Seviye	Orta	Bayan	Hayır
4	2.Seviye	Orta	Erkek	Evet
5	1.Seviye	Orta	Erkek	Evet
6	3.Seviye	Kötü	Bayan	Evet

tablosu oluşur. Yeni eğitim kümemiz oluşmuştur.

Sonuç	Risk		Sağlık		Cinsiyet	
	1.seviye	2. ve 3. seviye	Orta	Kötü	Bayan	Erkek
Evet	1	3	2	2	1	3
Hayır	0	1	1	0	1	0

Sol ve sağ gini değerleri tekrar hesaplanırsa:

Risk niteliği için

$$Gini_{sol} = 1 - \left[\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right] = 0$$

$$Gini_{sağ} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.37$$

Sağlık niteliği için;

$$Gini_{sol} = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.44$$

$$Gini_{sağ} = 1 - \left[\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right] = 0$$

Cinsiyet niteliği için;

$$Gini_{sol} = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.50$$

$$Gini_{s} = 1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right] = 0$$

Olarak elde edilir.

$$\mathbf{Gini_{risk} = \frac{1(0)+4(0.37)}{5} = 0.30}$$

$$\mathbf{Gini_{sağlık} = \frac{3(0.44)+2(0)}{5} = 0.26}$$

$$\mathbf{Gini_{cinsiyet} = \frac{2(0.50)+3(0)}{5} = 0.20}$$

Değerlerini tabloda toplarsak;

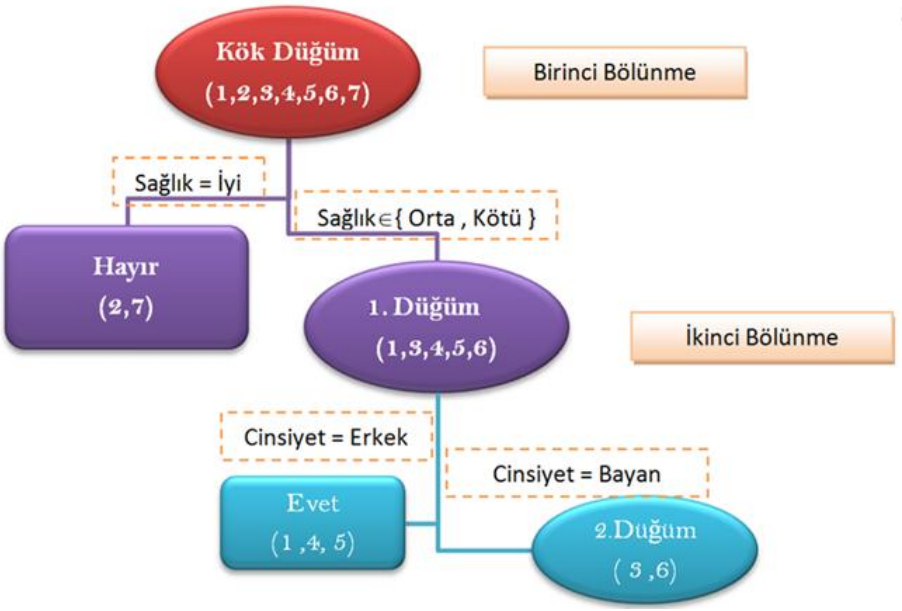
Kabul	Risk		Sağlık		Cinsiyet	
	1.Sev.	2. ve 3.Seviye	Orta	Kötü	Bayan	Erkek
EVET	1	3	2	2	1	3
HAYIR	0	1	1	0	1	0
$Gini_{sol}$, $Gini_{sağ}$	0.00	0.37	0.44	0.00	0.50	0.00
$Gini_j$	0.30		0.26		0.20	

elde edilir.

Cinsiyet en küçük gini değeridir. Bayan 3-6 ve erkek 1-4-5 kayıtlarındadır.

İşlem Sırası	Risk	Sağlık	Cinsiyet	Sonuç
1	2.Seviye	Kötü	Erkek	Evet
3	3.Seviye	Orta	Bayan	Hayır
4	2.Seviye	Orta	Erkek	Evet
5	1.Seviye	Orta	Erkek	Evet
6	3.Seviye	Kötü	Bayan	Evet

Karar ağacına uygularsak;

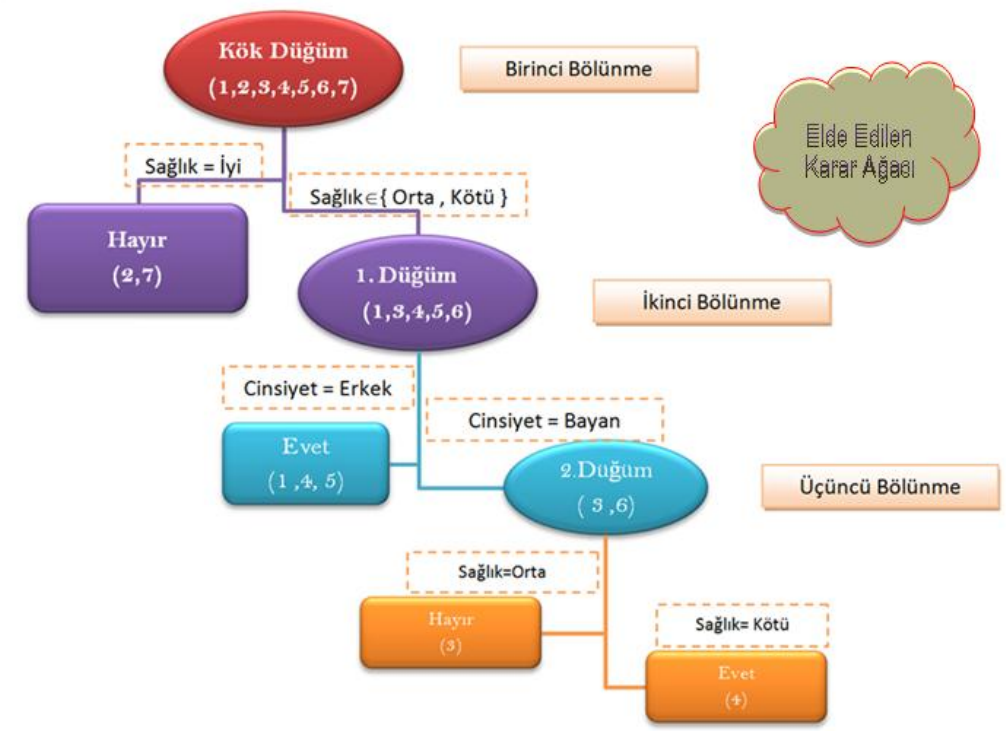


elde edilir.

Tablodan 1-4-5 kayıtları çıkarılıp işlemler tekrarlanırsa

İşlem Sırası	Risk	Sağlık	Cinsiyet	Sonuç
3	3.Seviye	Orta	Bayan	Hayır
6	3.Seviye	Kötü	Bayan	Evet

Tablosu kalmıştır. Hesaplamalar yapıldığında son karar ağacı



biçiminde oluşur ve tamamlanır.

Karar Ağacına bağlı olarak kural tablosu oluşturulursa ;

1.Kural

Eğer Sağlık=İyi ise Sonuç=Hayır;

2.Kural

Eğer Sağlık=Orta veya Sağlık=Kötü ise ve

Eğer Cinsiyet=Erkek ise Sonuç=Evet ;

3.Kural

Eğer Sağlık=Orta veya Sağlık=Kötü ise ve

Eğer Cinsiyet=Bayan ise ve

Eğer Sağlık=Orta ise Sonuç=Hayır;

4.Kural

Eğer Sağlık=Orta veya Sağlık=Kötü ise ve

Eğer Cinsiyet=Bayan ise ve

Eğer Sağlık=Kötü ise Sonuç=Evet;

Kaynaklar

- Veri Madenciliği Yöntemleri Dr. Yalçın Özkan 06’2008
- Veri Madenciliği DR .Gökhan Silahtaroğlu 06’2008
- İstanbul Ticaret Üniversitesi Dergisi Ver Madenciliği Modeller Ve uygulama Alanları (Serhat ÖZEKES)