VERİ MADENCİLİĞİ DERSİ SUNU-4

Doç.Dr.Mehmet Akif ŞAHMAN

Sınıflama Teknikleri ve Algoritmaları

Sınıflandırma, en çok bilinen veri madenciliği tekniklerinden birisidir. Sınıflandırma tahminleyici bir modeldir. Resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konuları sınıflandırmanın çokça kullanıldığı alanlardır. Sınıflandırma metotları genel olarak üç grupta toplanır, bunlar ;

- Karar Ağaçları
- İstatistiği Dayalı Algoritmalar
- Mesafeye Dayalı Algoritmalar

Sınıflama Teknikleri ve Algoritmaları

Karar Ağaçları

Sınıflandırma probleminde bir ağaç yapısı oluşturulur ve veritabanındaki her kayıt bu ağaca uygulanır, daha sonra çıkan sonuca göre kayıt sınıflandırılır.

Karar Ağaçları matematiksel olarak aşağıdaki gibi ifade edilebilir.

$D = \{t_1, ..., t_n\}$ bir veri tabanı olsun. Veri tabanındaki tablolar farklı özelliklerden(attribute) $A = \{A_1, ..., A_n\}$ oluşmaktadır. Bu tablolarda da $C = \{C_1, ..., C_n\}$ kadar sınıf olsun.

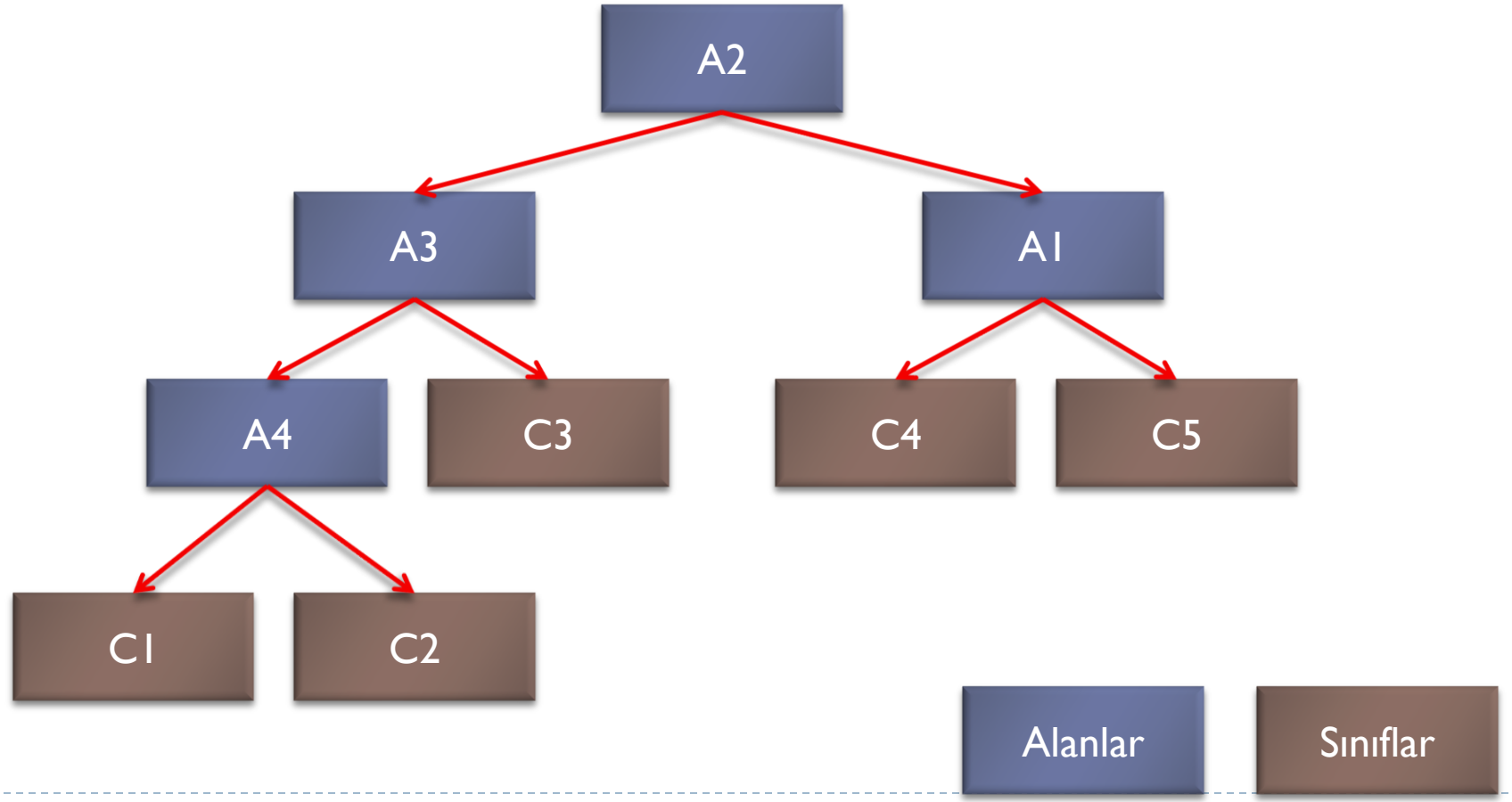
- Her bir düğüm A_i alanıyla isimlendirilmiştir.
- Her düğümden ayrılan kollar bu alanla ilgili soruya yanıt verir.
- Her yaprağın bir sınıf olduğu bir ağaçtır.

Sınıflama Teknikleri ve Algoritmaları

Cinsiyet	Kilo	Boy	Beden
K	48	170	Orta
K	49	151	Küçük
K	52	158	Orta
K	56	165	Orta
E	59	160	Küçük
K	61	159	Orta
E	62	162	Küçük
E	63	174	Orta
K	68	168	Orta
K	69	177	Büyük
E	72	170	Orta
E	74	165	Küçük
E	85	175	Orta
E	85	190	Büyük
E	98	190	Büyük

Sınıflama Teknikleri ve Algoritmaları

Karar Ağacının Yapısı



Sınıflama Teknikleri ve Algoritmaları

ID3 Algoritması

Bu algoritma ile sınıflandırma yapılırken ayırıcı özellik bulunmaya çalışılır ve entropi kullanılır. Entropi, eldeki bilgilerin sayısallaştırılması ve beklentisizliğin maksimumlaştırılmasıdır.

Örn. Herkes aynı futbol takımını tutsaydı, her hangi birine takımını sorduğunuzda alacağınız cevap bizi şaşırtmayacaktı ve entropi 0 (sıfır) olacaktı. Bütün olasılıklar eşit olduğu durumda entropi maksimum değerini alacaktır.

<p1,p2, … , pn> olasılıkları ifade ederse tüm olasılıkların toplamı 1 olmalıdır. $\sum_{i=1}^{n} p_i = 1$, bu durumda entropi aşağıdaki gibi olacaktır.

$$H(p_1, p_2, ..., p_n) = \sum (p_i \log(1/p_i))$$

Sınıflama Teknikleri ve Algoritmaları

ID3 (Iterative Dichotomizer 3) Algoritması

Verileri doğru bir şekilde sınıflandırabilmek için öncelikle tüm veriler ve anlamlı şekilde parçalanmış arasındaki farka bakılır. Bu veriler ışığında öncelikli düğüm ve dallanmalara karar verilir. İşlemler sonucunda kazanımlara bakılarak, en büyük kazanım değerine sahip alan kök düğüm olarak belirlenir. Kazanım aşağıdaki gibi hesaplanır.

$$Kazanım(D;S) = H(D) - \sum_{i=1}^{n} P(D_i)H(D_i)$$

Sınıflama Teknikleri ve Algoritmaları

Örn. Tabloda verilen veriler kullanılarak bir karar ağacı hesaplanmak istenirse kök düğümün ne olacağını hesaplayınız.

Öncelikle genel durum entropisi hesaplanması için; 4 adet küçük, 8 orta ve 3 tanesi büyük sınıfında verimiz vardır. Buna göre;

$$H(p_1, p_2, ..., p_n) = \sum (p_i \log(1/ p_i))$$

$$\frac{4}{15}\log(\frac{15}{4}) + \frac{8}{15}\log(\frac{15}{8}) + \frac{3}{15}\log(\frac{15}{3}) = 0,4384$$

olarak bulunur.

Sınıflama Teknikleri ve Algoritmaları

Cinsiyet için entropi hesaplanırsa;

$$Entropi\ Kadın = \frac{1}{7}\log(\frac{7}{1}) + \frac{5}{7}\log(\frac{7}{5}) + \frac{1}{7}\log(\frac{7}{1}) = 0,3458$$

$$Entropi\ Erkek = \frac{3}{8}\log(\frac{8}{3}) + \frac{3}{8}\log(\frac{8}{3}) + \frac{2}{8}\log(\frac{8}{2}) = 0,47$$

$$Ağırlıklı\ Toplam = (\frac{7}{15})x0,3458 + (\frac{8}{15})x0,47 = 0,4120$$

$$Cinsiyet\ Kazanımı = 0,4384 - 0,4120 = 0,0264$$

olarak bulunur.

Sınıflama Teknikleri ve Algoritmaları

Kilo için entropi hesaplanırsa;

Aralık	Grup No
48-55,99	1
56-64,99	2
65-75,99	3
76-85,99	4
86-100	5

$$1.Grup = \frac{1}{3}\log(\frac{3}{1}) + \frac{2}{3}\log(\frac{3}{2}) = 0,2764$$

$$2.Grup = \frac{2}{5}\log(\frac{5}{2}) + \frac{3}{5}\log(\frac{5}{3}) = 0,2923$$

$$3.Grup = \frac{1}{4}\log(\frac{4}{1}) + \frac{2}{4}\log(\frac{4}{2}) + \frac{1}{4}\log(\frac{4}{1}) = 0,4515$$

$$4.Grup = \frac{1}{2}\log(\frac{2}{1}) + \frac{1}{2}\log(\frac{2}{1}) = 0,3010$$

$$5.Grup = \frac{1}{1}\log(\frac{1}{1}) = 0$$

$$Ağırlıklı\ Toplam = 0,3132$$

$$Kilo\ Kazanımı = 0,4384 - 0,3132 = 0,1252$$

Sınıflama Teknikleri ve Algoritmaları

Boy için entropi hesaplanırsa;

$$Ağırlıklı\ Toplam = 0,4404$$

$$Boy\ Kazanımı = 0,4384 - 0,4404 = -0,002$$

$$Cinsiyet\ Kazanımı = 0,0264$$

$$Kilo\ Kazanımı = 0,1252$$

$$Boy\ Kazanımı = -0,002$$

Aralık	Grup No
151-159,99	1
160-164,99	2
165-174,99	3
175-184,99	4
185- ++++	5

En Yüksek kazanım **Kilo** da olduğu için kök düğüm Kilo olarak belirlenir. Bu aşamadan sonra kilo değerleri 48-55,99 arasındaki veriler ele alınarak boy ve cinsiyet kazanımları tekrar hesaplanacak ve buradaki düğümün ismi belirlenecektir. Daha sonra bu işlem diğer dallar için de uygulanacak ve ağaç yapısı oluşturulacaktır.

Sınıflama Teknikleri ve Algoritmaları

C 4.5 ve C 5 Algoritması

Bu algoritma, ID3 algoritmasından kayıp verileri de hesaba katması ile ayrılmaktadır. C 4.5 algoritması, kayıp verileri diğer veri ve değişkenler yardımıyla öngörerek kazanım oranın hesaplamasında kullanılır. Ayrıca bu algoritmada kazanım aşağıdaki formüller kullanılarak hesaplanır.

$$Kilo\ Kazanımı(D,S) = Kazanım\ (D,S) / Ayırma\ Bilgisi(D,S)$$

$$Ayırma\ Bilgisi(D,S) = H\left(\frac{|D_1|}{|D|},\ldots,\frac{|D_n|}{|D|}\right)$$

Sınıflama Teknikleri ve Algoritmaları

Örn. Tabloda verilen veriler kullanılarak bir karar ağacı hesaplanmak istenirse kök düğümün ne olacağını C4.5 algoritmasını kullanarak hesaplayınız.

Cinsiyet için $H\left(\frac{7}{15},\frac{8}{15}\right)=\frac{7}{15}\log(\frac{15}{7})+\frac{8}{15}\log(\frac{15}{8})=0,3001$

$$Kazanım\ Oranı\ (C)=0,4384/0,3001=1,46$$

Kilo için $H\left(\frac{3}{15},\frac{5}{15},\frac{4}{15},\frac{2}{15},\frac{1}{15}\right)=0,6470$

$$Kazanım\ Oranı\ (K)=0,4384/0,6470=0,6776$$

Boy için $H\left(\frac{3}{15},\frac{2}{15},\frac{6}{15},\frac{2}{15},\frac{2}{15}\right)=0,6490$

$$Kazanım\ Oranı\ (B)=0,4384/0,6490=0,6755$$

Kazanım oranlarından en küçük değerli değişken kök (ya da bir sonraki düğüm olarak belirlenir.)

Sınıflama Teknikleri ve Algoritmaları

SPRINT(Scalable Parallelizable Induction of Decision Trees) Algoritması

ID3 ve C 4.5 ile C 5 algoritmaları derinlik ilkesine göre çalışırken, bu algoritma listeler kullanarak sınıflandırma yapar. Her değişken için bir ayrı bir değişken listesi hazırlanır, listeler bölünse bile yeniden sıralama yapılmaz. Bu algoritmada, kök düğüm aşağıdaki formüller kullanılarak bulunur.

$$gini(K) = 1 - \sum p_j^2$$

$$gini_{bolünmüş}(K) = \frac{n_1}{n_2} gini(K_1) + \frac{n_2}{n_2} gini(K_2)$$

Sınıflama Teknikleri ve Algoritmaları

Örn. Tabloda verilen veriler kullanılarak bir karar ağacı hesaplanmak istenirse kök düğümün ne olacağını SPRINT algoritmasını kullanarak hesaplayınız.

Cinsiyet için,

$$Gini(Kadın) = 1 - \left[(1/7)^2 + (5/7)^2 + (1/7)^2 \right] = 0,4490$$

$$Gini(Erkek) = 1 - \left[(3/8)^2 + (3/8)^2 + (2/8)^2 \right] = 0,6563$$

$$Gini_{bolünmüş}(Cinsiyet) = (7/15) * 0,4490 + (8/15) * 0,6563 = 0,5596$$

Sınıflama Teknikleri ve Algoritmaları

Kilo için,

$$Gini(1) = 1 - \left[(1/3)^2 + (2/3)^2 + (0/3)^2 \right] = 0,4444$$

$$Gini(2) = 1 - \left[(2/5)^2 + (3/5)^2 + (0/5)^2 \right] = 0,4800$$

$$Gini(3) = 1 - \left[(1/4)^2 + (2/4)^2 + (1/4)^2 \right] = 0,6250$$

$$Gini(4) = 1 - \left[(0/2)^2 + (1/2)^2 + (1/2)^2 \right] = 0,5000$$

$$Gini(5) = 1 - \left[(0/1)^2 + (0/1)^2 + (1/1)^2 \right] = 0$$

$$Gini_{bolünmüş}\ (Kilo) = (3/15) * 0,4444 + (5/15) * 0,48 + (4/15) * 0,6250$$

$$+ (2/15) * 0,5 + (1/15) * 0 = 0,4822$$

Sınıflama Teknikleri ve Algoritmaları

Boy için (En küçük değere sahip olduğu için boy kök olarak seçilir),

$$Gini(1) = 1 - \left[(1/3)^2 + (2/3)^2 + (0/3)^2 \right] = 0,4444$$

$$Gini(2) = 1 - \left[(2/2)^2 + (0/2)^2 + (0/2)^2 \right] = 0$$

$$Gini(3) = 1 - \left[(1/6)^2 + (5/6)^2 + (0/6)^2 \right] = 0,2449$$

$$Gini(4) = 1 - \left[(0/2)^2 + (1/2)^2 + (1/2)^2 \right] = 0,5000$$

$$Gini(5) = 1 - \left[(0/2)^2 + (0/2)^2 + (2/2)^2 \right] = 0$$

$$Gini_{bolünmüş}\ (Boy) = (3/15) * 0,4444 + (2/15) * 0 + (6/15) * 0,2449$$

$$+ (2/15) * 0,5 + (1/15) * 0 = 0,2534$$

Ödev

14 günlük hava durumuna göre dışarıda tenis oynanması için karar verilmiştir. Buna göre, ID3 algoritmasına göre ağaç yapısını oluşturunuz.

Day	Outlook	Temp.	Humidity	Wind	Decision	Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No	8	Sunny	Mild	High	Weak	No
2	Sunny	Hot	High	Strong	No	9	Sunny	Cool	Normal	Weak	Yes
3	Overcast	Hot	High	Weak	Yes	10	Rain	Mild	Normal	Weak	Yes
4	Rain	Mild	High	Weak	Yes	11	Sunny	Mild	Normal	Strong	Yes
5	Rain	Cool	Normal	Weak	Yes	12	Overcast	Mild	High	Strong	Yes
6	Rain	Cool	Normal	Strong	No	13	Overcast	Hot	Normal	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes	14	Rain	Mild	High	Strong	No