

VERİ MADENCİLİĞİ YILIÇI SINAVI

Fen Bilimleri Enstitüsü

Yrd. Doç. Dr. Şule Öğüdücü

Çözümler

1. Veritabanı ve veri madenciliği işlemleri arasındaki farkları yazınız. Örnekle açıklayınız.

Cevap:

- Veritabanı sorgulamaları tanımlıdır. Veri madenciliği işlemlerinde aranan bilgi veri içinde gizli olduğundan tanımlı bir sorgulama yoktur. Veritabanı sorgulamaları için SQL sorgulama dili kullanılır. Veri madenciliği işlemlerinde kullanılan yaygın sorgulama dili yoktur.
- Veritabanı işlemleri canlı veri üzerinde yapılır. Veri madenciliği işlemleri, üzerinde işlem yapılmayan veri üzerinde yapılır.
- Veritabanı işlemlerinin çıkışı bellidir ve verinin bir altkümesidir. Veri madenciliği işlemlerinin sonucunda elde edilecek bilgi belirgin değildir.
- SQL sorgulamasına örnek: DVD satın alan tüm müşterileri bul.
- Veri madenciliği işlemlerine örnek: DVD birlikte sıkça satın alınan ürünü bul

2. Veri madenciliği uygulamalarında kullanılan modeller kaç'a ayrılır? Açıklayınız.

Cevap: İkiye ayrılır.

- kestirime dayalı: sınıflandırma, eğri uydurma, zaman eğrileri
- tanımlayıcı: demetleme, özetleme, bağıntı kuralları, sıralı diziler

3. Bir veri madenciliği uygulaması tasarlanacaktır. Çalışılacak veri kümesindeki niteliklerden biri *yaş* bilgisidir. Bu niteliğe ait değerler şu şekildedir: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 30, 45

- (a) *yaş* niteliğine ait veri için gürültüyü azaltmak amacıyla eşit aralıklarla bölme genişliği 3 olacak şekilde bölmeleme (*binning*) yapılacaktır. Bu işlemi yapmak için gerekli adımları gösteriniz. Yukarıdaki veri için bu işlemin etkilerini yorumlayınız.

Cevap:

Bölme No	Veri	Ortalama	Alt-Üst Sınır
1	13-15-16	14,7-14,7-14,7	13-16-16
2	16-19-20	18,3-18,3-18,3	16-20-20
3	20-21-22	21-21-21	20-22-22
4	22-25-25	24-24-24	22-25-25
5	25-30-45	33,3-33,3-33,3	25-25-45

Aynı değere sahip veriler farklı bölmelerde yer aldıkları için farklı değerler alıyorlar (Örnek: 16 hem 1. hem de 2. bölmede var. 1. bölmede ortalama kullanıldığında 14,7 değeri ile, 2. bölmede 18,3 değeri ile düzeltiliyor). Ancak alt ve üst sınır kullanılarak düzeltme yapılırsa bu problem ortadan kalkıyor.

- (b) Verideki aykırılıkları belirlemek için hangi işlemler yapılabilir?

Cevap: Demetleme

(c) Verideki gürültüyü düzeltmek için başka hangi yöntemler uygulanabilir?

Cevap: Eğri uydurma

4. Aşağıda verilen tablodaki verileri sınıflandırmak için karar ağacı oluşturulacaktır.

A	B	Sınıf
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

(a) A ve B nitelikleri için bilgi kazancını hesaplayın. Karar ağacı algoritması hangi niteliği kullanarak bölmeleme yapar?

Cevap:

A ve B niteliklerine ait sınıf dağılımları

	A=T	A=F
+	4	0
-	3	3

	B=T	B=F
+	3	1
-	1	5

Bölünmeden önce entropi:

$$E_{orig} = -0.4 \log 0.4 - 0.6 \log 0.6 = 0.9710$$

A niteliği kullanılarak bölündüğü durumda bilgi kazancı:

$$E_{A=T} = -\frac{4}{7} \log \frac{4}{7} - \frac{3}{7} \log \frac{3}{7} = 0.9852$$

$$E_{A=F} = -\frac{3}{3} \log \frac{3}{3} - \frac{0}{3} \log \frac{0}{3} = 0$$

$$Gain(S, A) = E_{orig} - 7/10 E_{A=T} - 3/10 E_{A=F} = 0.2813$$

B niteliği kullanılarak bölündüğü durumda bilgi kazancı:

$$E_{B=T} = -\frac{3}{4} \log \frac{3}{4} - \frac{1}{4} \log \frac{1}{4} = 0.8113$$

$$E_{B=F} = -\frac{1}{6} \log \frac{1}{6} - \frac{5}{6} \log \frac{5}{6} = 0.6500$$

$$Gain(S, B) = E_{orig} - 4/10 E_{B=T} - 6/10 E_{B=F} = 0.2565$$

Veriyi bölmek için A niteliği seçilir.

- (b) A ve B nitelikleri için Gini index değerini hesaplayınız. Karar ağacı algoritması hangi niteliği kullanarak bölmeleme yapar?

Cevap:

Bölünmeden önce Gini index

$$Gini_{orig} = 1 - 0.4^2 - 0.6^2 = 0.48$$

A niteliği kullanılarak bölündüğü gini göstergesi için kazanç:

$$G_{A=T} = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4898$$

$$E_{A=F} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$Gain(S, A) = G_{orig} - 7/10Gini_{A=T} - 3/10Gini_{A=F} = 0.1371$$

B niteliği kullanılarak bölündüğü gini göstergesi için kazanç:

$$G_{B=T} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.3750$$

$$E_{B=F} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.2778$$

$$Gain(S, B) = G_{orig} - 4/10Gini_{B=T} - 6/10Gini_{B=F} = 0.1633$$

Veriyi bölmek için B niteliği seçilir.

- (c) Bilgi kazancı ve gini index değerlerinin bölmeleme için farklı nitelikleri seçmesi mümkün müdür? Açıklayınız.

Cevap:Gini göstergesi ve entropi ölçütleri benzer özellikler gösterir, ancak bu ölçütleri kullanarak hesaplanan kazanç aynı yukarıdaki örnekte olduğu gibi aynı özelliği göstermeyebilir.