VERİ MADENCİLİĞİ DERSİ SUNU-3

Doç.Dr.Mehmet Akif ŞAHMAN

Veri Madenciliği Modelleri

- Veri madenciliği modelleri kullanıldıkları alanlara göre değişik modellere ayrılmaktadır. Bunlar;
  - Değer Tahmin Modeli
  - Kümeleme Modeli
  - Bağlantı Analizi
  - Fark Sapmaları

  Bu modeller iki yaklaşım olarak ifade edilebilir.
  - Tahminleyici modeller (Bu işlemde dolandırıcılık var mı?)
  - Tanımlayıcı modeller (Çocuk bezi alan bir müşterinin, mama alma olasılığı diğerlerinden 3 kat daha fazladır.)

Veri Madenciliği Modelleri

- **Değer Tahmin Modeli :** Bu modelde kendisine verilen veritabanını ve verileri inceleyerek verilerdeki temel unsurları birbirine benzeterek tanımlamaya, onları isimlendirmeye ve sınıflandırmaya çalışmaktadır. Örn. Cinsiyet kavramını zamanla kavrayıp, dış görünüşten çıkarımda bulunulması buna örnektir. Bir emlakçının bir mülke ait tahmini fiyat belirlemesi de buna örnek olarak verilebilir.
  - Denetimli Öğrenme : Bu öğrenmede türünde bir denetmen yardımıyla (veriler daha önceden verilerek) öğrenme gerçekleşir.
  - Denetimsiz Öğrenme : Bu öğrenmede veriler daha önceden verilmeden, mevcut veriler üzerinden kümeleme vb. yapılır.

Veri Madenciliği Modelleri

- **Bağlantı Analizi :** Bu modelde veriler bir bütün halinde değil de, her bir kayıt ve kayıt edilmiş veri grupları arasındaki bir bağlantı ve ilişki kurulmaya çalışılır. En çok kullanılan alanlar, çapraz satış, stok fiyat hareketleri ve hedef müşteri kitlesinin belirlenmesidir.
- **Birliktelik Kuralları :** Bu modelde, belirli türlerdeki veri ilişkilerini tanımlayan bir modeldir. Bir ürün alındığında başka bir ürünün alınma ihtimalinin bulunması bir birliktelik kuralı ile sağlanır. Örn. Bir markette hangi ürünlerin birlikte satıldığının bulunması birliktelik kuralları ile bulunabilir. Makarna içeren %40 alış verişin, %2'si aynı zamanda çocuk bezi içermektedir. Buradaki %40 güven seviyesini, %2 ise bu güven seviyesine olan desteği belirtmektedir.

Veri Madenciliği Modelleri

- **Örüntü Tanıma :** Daha önce veri tabanında tanımlanmış olan çok boyutlu veriler kullanılarak, yeni girilen verinin aynısını veya benzerini bulmak örüntü tanıma olarak tanımlanabilir. Parmak izi, yüz veya iris tanımlama buna örnek olarak verilebilir.
- **Ardışık Zaman Örüntüleri :** Zamana bağlı olarak elde edilen veriler kullanılarak sonrası için çıkarımda bulunması ardışık zaman örüntülerinin konusudur. Örn. Müşterinin birinci gün A ürününü alan daha sonraki gün B ürününü alan ve ondan sonraki gün C ürününü alması bir örüntü oluşturmaktır.

Veri Madenciliği Modelleri

Müşteri No	İşlem Zamanı	Ürün No
1	15 Mayıs 2003	27,21
2	15 Mayıs 2003	21
1	16Mayıs 2003	22,28,23
4	16 Mayıs 2003	56
4	17 Mayıs 2003	25,89,98,45
1	18 Mayıs 2003	25
2	18 Mayıs 2003	22
3	19 Mayıs 2003	36
2	20 Mayıs 2003	23

Müşteri No	Ürün No
1	27,21,22,28,23,25
2	21,22,23
3	36
4	56,25,89,98,45

4 numaralı müşterinin ürün satın alması başka bir müşteride yoktur. Eğer destek seviyesi %0 olarak alınırsa bu bir örüntü olabilir.

Minimum destek seviyesi %25 olarak alınırsa yani en az iki müşteri olarak belirlenirse 21,22,23 anlamlı ardışık bir örüntü olarak belirlenebilir. Bu ürünler 1 ve 2 nolu müşteriler tarafından desteklenmektedir.

Veri Madenciliği Modelleri

- **Dolandırıcılık Tespiti :** Dolandırıcılık tespiti fark sapmaları olarak da anılmaktadır ve aslında bir örüntü tanıma problemidir. Bilgisayar daha önceki kayıtları inceleyerek dolandırıcılık tespitini yapmaya çalışır. Burada veri ambarı oluşturulurken, daha önce dolandırıcılık yapan kayıtlar ile dolandırıcı olmayan kullanıcıların verileri birlikte yer alması önemlidir. Aksi durumda sistem dolandırıcılık yapanlar ile yapmayanların tespitini iyi bir şekilde yapamayabilir.

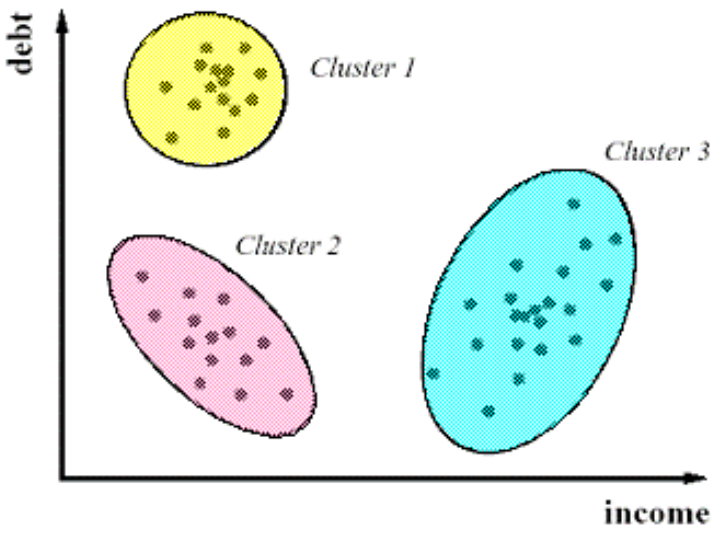
Veri Madenciliği Modelleri

Arayan Numara	Aranan Numara	Tarih	Saat	Gün	Süre (Dakika)
212021	654064	11/01/06	01:30	Salı	8.02
212021	090056	12/01/06	02:35	Çarş	45.00
313031	252525	01/01/06	14:13	Salı	7.32

Numara	Cinsiyet	Kayıt Tarihi	Ödeme Durumu	Ort. Konuşma	Dolandırıcılık
212021	E	…	İyi	230.12	Hayır
313031	E	…	Kötü	255.23	Evet
568695	K	…	Kötü	265.22	Hayır
456547	E	…	Orta	355.12	Evet

Veri Madenciliği Modelleri

- **Kümeleme Analizi :** Veri madenciliğinin en önemli alanlarından biridir. Amacı, nesnelerin birbirlerine olan benzerliklerine göre gruplara ayırmaktır. Birbirine benzeyen veriler bir kümeye, benzemeyen veriler başka bir kümeye yerleştirilir.Kümeleme, "gizli örüntülerin ortaya çıkartılması için uygulanan bir denetimsiz öğrenme yaklaşımı" olarak da tanımlanabilir.



Veri Madenciliği Modelleri

- **Kümeleme Analizi :** Veri madenciliğinde, eğer değişken sayısı artarsa burada kümelerin oluşturulması için;

1. $mes(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
2. $mes(x,y) = \sum_{i=1}^{n} | x_i - y_i |$
3. $mes(x,y) = \max_{i=1}^{n} | x_i - y_i |$

formüller kullanılabilir.

Veri Madenciliği Modelleri

Örn.-1) A={1,1,2,2,5} dizisine, B={1,2,3,4,2} dizisi mi, yoksa C={1,3,5,1,3} dizisi mi yakındır?

$$mes(A,B) = \sqrt{(1-1)^2 + (1-2)^2 + (2-3)^2 + (2-4)^2 + (5-2)^2} = \sqrt{15}$$

$$mes(A,C) = \sqrt{(1-1)^2 + (1-3)^2 + (2-5)^2 + (2-1)^2 + (5-3)^2} = \sqrt{18}$$

$$mes(A,B) < mes(A,C)$$

dolayısıyla A, B'ye daha yakındır.

Veri Madenciliği Modelleri

Örn.2-) A={1,1,2,2,5} B={1,2,3,4,2} C={1,3,5,1,3} web sitelerinin içerik dizisidir. A_1=1, A web sitesinde 1 nolu kelime 1 kere, B_5=2 B web sitesinde 5 numaralı sözcük 2 kere tekrarlanmış demektir. Buna göre hangi web sitelerinin içeriği birbirine daha yakındır.

$$mes(X,Y) = 1 - \frac{X * Y(skaler\ çarpımı)}{X(vektörü) * Y(vektörü)}$$

$$A, B\ skaler\ çarpımı = (1*1) + (1*2) + (2*3) + (2*4) + (5*2) = 27$$

$$A\ vektörel\ büyüklüğü = \sqrt{1^2 + 1^2 + 2^2 + 2^2 + 5^2} = \sqrt{35}$$

$$B\ vektörel\ büyüklüğü = \sqrt{1^2 + 2^2 + 3^2 + 4^2 + 2^2} = \sqrt{34}$$

$$mes(A,B) = 1 - \frac{27}{\sqrt{35} * \sqrt{34}} = 0,217 \qquad mes(A,C) = 1 - \frac{31}{\sqrt{35} * \sqrt{45}} = 0,218$$

$$mes(A,B) < mes(A,C)$$

dolayısıyla A, B'ye daha yakındır.