VERİ MADENCİLİĞİ

(Veri Ön İşleme-2)

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

Veri Önişleme

- Veri
- Veri Önişleme
- Veriyi Tanıma
- Veri temizleme
- Veri birleştirme
- Veri dönüşümü
- Veri azaltma
- Benzerlik ve farklılık

Veri Dönüşümü

- Veri, veri madenciliği uygulamaları için uygun olmayabilir
- Seçilen algoritmaya uygun olmayabilir
  - Veri belirleyici değil
- Çözüm
  - Veri düzeltme
    - Bölmeleme
    - Kümeleme
    - Eğri Uydurma
  - Biriktirme
  - Genelleme
  - Normalizasyon
  - Nitelik oluşturma

Normalizasyon

- min-max normalizasyon
  - min-max normalleştirmesi ile orijinal veriler yeni veri aralığına doğrusal dönüşüm ile dönüştürülürler. Bu veri aralığı genellikle 0-1 aralığıdır.
- z-score normalizasyon
  - z Skor normalleştirmede (veya 0 ortalama normalleştirme) ise değişkenin her hangi bir y değeri, değişkenin ortalaması ve standart sapmasına bağlı olarak bilinen Z dönüşümü ile normalleştirilir.
- ondalık normalizasyon
  - Ondalık ölçekleme ile normalleştirmede ise, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme gerçekleştirilir. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Ondalık ölçeklemenin formülü aşağıdaki şekildedir:
  - Örneğin 900 maksimum değer ise, n=3 olacağından 900 sayısı 0,9 olarak normalleştirilir.

Normalizasyon

- min-max normalizasyon

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalizasyon

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- ondalık normalizasyon

$$v' = \frac{v}{10^j}$$

j: Max($|v'|$)<1 olacak şekildeki en küçük tam sayı

Nitelik Oluşturma

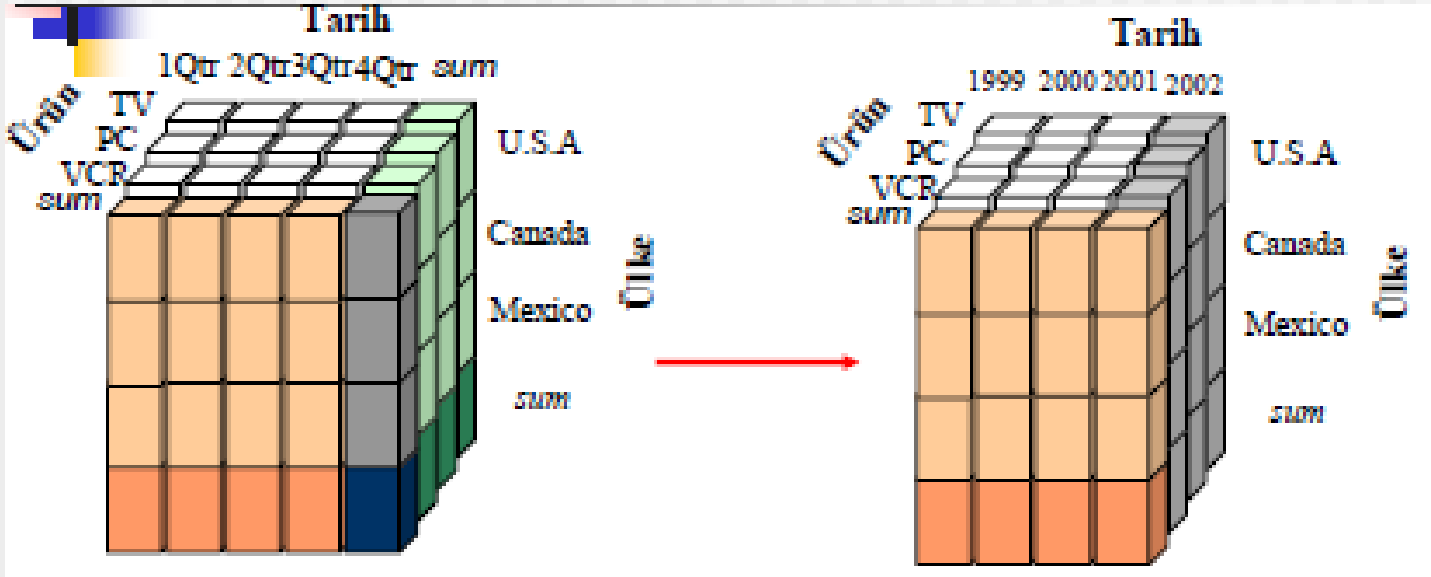
- Yeni nitelikler yarat
  - orjinal niteliklerden daha önemli bilgi içersin
    - alan=boy x en
  - veri madenciliği algoritmalarının başarımı daha iyi olsun

Veri Azaltma

Veri Azaltma

- Veri miktarı çok fazla olduğu zaman veri madenciliği algoritmalarının çalışması ve sonuç üretmesi çok uzun sürebilir
  - veriyi azaltma başarımı artırır
  - sonucun (nerdeyse) hiç değişmemesi gerekir
- Veri azaltma
  - nitelik birleştirme
  - nitelik azaltma
  - veri sıkıştırma
  - veri ayrıştırma ve kavram oluşturma
  - veri küçültme
    - eğri uydurma
    - kümeleme
    - histogram
    - örnekleme

Nitelik Birleştirme



- Sorgulama için gerekli olan boyutlar kullanılıyor.

Nitelik Seçme - Nitelik Azaltma

- Nitelik Seçme
  - Nitelikler kümesinin bir alt kümesi seçilerek veri madenciliği işlemi yapılır.
- Nitelik azaltma
  - d boyutlu veri kümesi k<d olacak şekilde k boyuta taşınır.

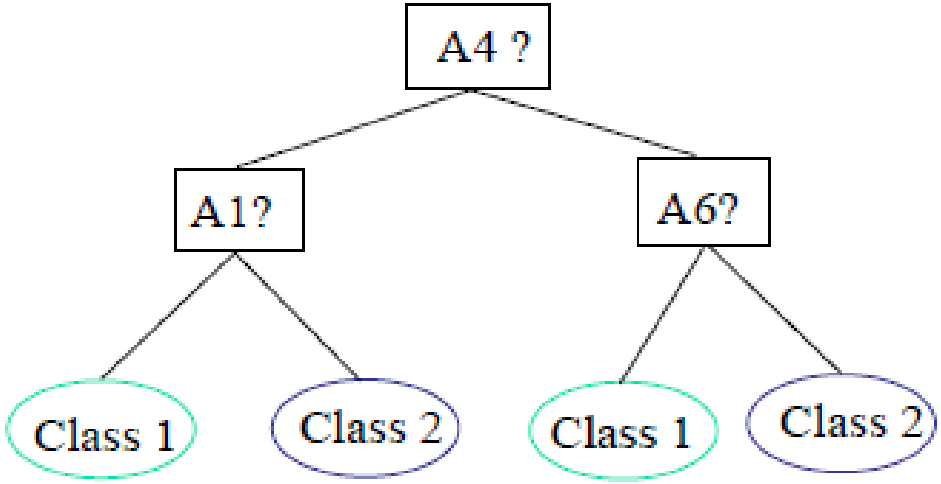
Nitelik Seçme

- Nitelik seçme
  - Veri madenciliği uygulaması için gerekli olan niteliklerin seçilmesi
  - Nitelikler altkümesi kullanılarak elde edilen sınıfların dağılımları gerçek dağılıma eşit ya da çok yakın olmalı
  - Veri madenciliği işlemi yer ve zaman karmaşıklığını azaltma
- Sistemin başarımını artırma
  - Sezgisel yöntemler kullanılarak nitelikler seçilebilir.
  - istatistiksel anlamlılık testi (statistical significance)
  - bilgi kazancı (information gain)
  - karar ağaçları

Örnek

Başlangıç nitelikler kümesi:

{A1, A2, A3, A4, A5, A6}



Seçilen nitelik kümesi: {A1, A4, A6}

Nitelik Azaltma

- Çok boyutlu veriyi daha küçük boyutlu uzaya taşıma
- d nitelikten oluşan n adet veri $D=\{x_1, x_2, ...,x_n\}$ k boyutlu uzaya taşınır:

$$\mathbf{x}_i \in \mathrm{R}^d \rightarrow \mathbf{y}_i \in \mathrm{R}^k (k << d)$$

- Veri kümesinde yer alan bütün nitelikler kullanılır
  - Niteliklerin doğrusal kombinasyonu
- Niteliklerin ayırıcılığına artırma

Veri Sıkıştırma

- Verinin boyutunu azaltır
  - daha az saklama ortamı
  - veriye ulaşmak daha çabuk
- Kayıplı ve kayıpsız veri sıkıştırma
  - bazı yöntemler bazı veri tiplerine uygun
  - her veri tipi için kullanılan yöntemler de var
- Eğer veri madenciliği yöntemi sıkıştırılmış veri üzerinde doğrudan çalışabiliyorsa elverişli

Veri Ayrıştırma

- Bazı veri madenciliği algoritmaları sadece ayrık veriler ile çalışır.
- Sürekli bir nitelik değerini bölerek her aralığı etiketler.
- Verinin değeri, bulunduğu aralığın etiketi ile değişir.
- Veri boyutu küçülür.
- Kavram oluşturmak için kullanılır.

Kavram Oluşturma

- Sayısal veriler
  - çok geniş aralıkta olabilir
  - değerleri çok sık değişebilir
- Sayısal veriler için kavram oluşturma
  - bölmeleme
  - histogram
  - kümeleme
  - entropi

Veri Küçültme

- Veriyi farklı şekillerde gösterme
  - parametrik
    - eğri uydurma
  - parametrik olmayan
    - histogram
    - kümeleme
    - örnekleme

Histogram ile Veri Küçültme

- Verinin dağılımı
- Veriyi bölerek her bölüm için veri değerini gösterir (toplam, ortalama)
  - eşit genişlik (equi-width): bölmelerin genişliği eşit
  - eşit yükseklik (equi-height): her bölmedeki veri sayısı eşit
  - v-optimal: en az varyansı olan histogram $\Sigma(count_b * value_b)$
  - MaxDiff: bölme genişliğini kullanıcı belirler

Kümeleme ile Veri Küçültme

- Veri kümelere ayrılır
- Veri kümeleri temsil eden örnekler (küme merkezleri) ve aykırılıklar ile temsil edilir
- Etkisi verinin dağılımına bağlı.
- Hiyerarşik kümeleme yöntemleri kullanılabilir.

Örnekleme ile Veri Küçültme

- Büyük veri kümesini daha küçük bir alt küme ile temsil etme
- Alt küme nasıl seçiliyor?
  - yerine koymadan örnekleme (SRSWOR)
  - yerine koyarak örnekleme (SRSWR)
  - küme örnekleme (yerine koymadan veya koyarak)
  - katman örnekleme (katman: nitelik değerine göre grup)

Benzerlik ve Farklılık

Benzerlik ve Farklılık

- Benzerlik
  - iki nesnenin benzerliğini ölçen sayısal değer
  - nesneler birbirine daha benzer ise daha büyük
  - genelde 0-1 aralığında değer alır
- Farklılık
  - iki nesnenin birbirinden ne kadar farklı olduğunu gösteren sayısal değer
  - nesneler birbirine daha benzer ise daha küçük
  - en küçük farklılık genelde 0
  - üst sınır değişebilir.

Uzaklık Çeşitleri

- Öklid(Euclid)
- Minkowski
- Manhattan

Öklid Uzaklığı

- Veri kümesi

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Uzaklık matrisi

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

- Öklid uzaklığı (Euclidean Distance) nesneler arası farklılığı bulmak için kullanılır.
  - p adet niteliği (boyutu) olan i ve j nesneleri arasındaki uzaklık

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

Minkowski Uzaklığı

- Öklid uzaklığının genelleştirilmiş hali

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+...+|x_{ip}-x_{jp}|^q)}$$

q: positif tam sayı

- q=1 → Manhattan uzaklığı

Uzaklık Özellikleri

- q=1⇒ Manhattan Uzaklığı
- q=2 ⇒ Öklid Uzaklığı
- Uzaklık ölçütünün sağlaması gereken özellikler:
  1. $d(i,j) \geq 0$
  2. $d(i,i) = 0$
  3. $d(i,j) = d(j,i)$
  4. $d(i,j) \leq d(i,h) + d(h,j)$
- Uzaklıklar ağırlıklı olarak da hesaplanabilir:

$$d(i,j) = \sqrt{w_1 \left| x_{i1} - x_{j1} \right|^2 + w_2 \left| x_{i2} - x_{j2} \right|^2 + \ldots + w_p \left| x_{ip} - x_{jp} \right|^2}$$

Benzerlik Özellikleri

- İki nesne arası benzerlik özellikleri
- 1. sim(i,j)>=0
- 2. sim(i,j)=sim(j,i)

İkili Değişkenler Arası Benzerlik

- İkili bir değişkenin 0 veya 1 olarak iki değeri olabilir.
- Bir olasılık tablosu oluşturulur:

		Nesne j	
		0	1
Nesne i	0	M_{00}	M_{01}
	1	M_{10}	M_{11}

M_{00}: i nesnesinin 0, j nesnesinin 0 olduğu niteliklerin sayısı
M_{10}: i nesnesinin 1, j nesnesinin 0 olduğu niteliklerin sayısı
M_{01}: i nesnesinin 0, j nesnesinin 1 olduğu niteliklerin sayısı
M_{11}: i nesnesinin 1, j nesnesinin 1 olduğu niteliklerin sayısı

- Yalın uyum katsayısı (simple matching coefficient): ikili değişkenin simetrik olduğu durumlarda

$$sim(i,j) = \frac{M_{11}+M_{00}}{M_{00}+M_{01}+M_{10}+M_{11}}$$

- Jaccard katsayısı (İkili değişkenin asimetrik olduğu durumlar):

$$d(i,j) = \frac{M_{11}}{M_{01}+M_{10}+M_{11}}$$

Kosinüs Benzerliği

- d_1 ve d_2 iki doküman. Kosinüs benzerliği

  $\cos(d_1,d_2)=d_1 \bullet d_2 / ||d_1|| \, ||d_2||$

  $d_i \bullet d_j$: iki dokümanın vektör çarpımı

  $||d_i||$: d_i dokümanının uzunluğu

- **Örnek**

  $d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$

  $d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

  $d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

  $||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

  $||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

  $\cos(d_1, d_2) = .3150$