VERİ MADENCİLİĞİ

(Veri Önişleme-1)

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

VERİ ÖNİŞLEME

Veri Nedir?

- Nesneler ve nesnelerin niteliklerinden oluşan küme
  - kayıt (record), varlık (entity), örnek (sample, instance) nesne için kullanılabilir.
- Nitelik (attribute) bir nesnenin (object) bir özelliğidir
  - bir insanın yaşı, ortamın sıcaklığı…
  - boyut (dimension), özellik (feature, characteristic) olarak da kullanılır.
- Nitelikler ve bu niteliklere ait değerler bir nesneyi oluşturur.

Tid	Geri Ödeme	Medeni Durum	Gelir	Dolandırıcı
1	Evet	Bekar	125K	-1
2	Hayır	Evli	100K	-1
3	Hayır	Bekar	70K	-1
4	Evet	Evli	120K	-1
5	Hayır	Boşanmış	95K	1
6	Hayır	Evli	60K	-1
7	Evet	Boşanmış	220K	-1
8	Hayır	Bekar	85K	1
9	Hayır	Evli	75K	-1
10	Hayır	Bekar	90K	1

Değer Kümeleri

- Nitelik için saptanmış sayılar veya semboller
- Nitelik & Değer Kümeleri
  - aynı nitelik farklı değer kümelerinden değer alabilir
    - ağırlık: kg, lb(libre, ağırlık ölçüsü)
  - farklı nitelikler aynı değer kümesinden değer alabilirler
    - ID, yaş: her ikisi de sayısal

İstatistiksel Veri Türleri

- 1- Nümerik Veriler : Sayısal-Nümerik-Nicel Veriler de denmektedir. Boy,Yaş gibi süreklilik arzeden değerler Nümerik verilerdir. "Daha fazla" ifadesi ile kullanılabilirler. Sürekli ve süreksiz olarak iki başlıkta ele alınabilir:
  - a) Sürekli Nümerik Veriler: Yaş, Sıcaklık
  - b) Aralıklı Nümerik Veriler (Interval): Çocuk Sayısı, Kaza Sayısı
- 2-Nominal Veriler : Kategorik bir veri çeşididir. "Daha fazla" ifadesi ile kullanılmazlar. İkiye ayrılır:
  - a)Binary Veriler: Var-Yok, Kadın-Erkek, Hasta-Sağlıklı
  - b)İkiden Çok Kategorili: Medeni Durum-Renk-Irk-Şehir, İsim, Forma Numarası
  - Örneğin forma numarası oyuncunun seviyesi ile ilgili bir bilgi içermez.
- 3-Ordinal Veriler : Ordinal veriler de yine kategorik veri türündendir. Fakat değerleri arasında sıralı bir ilişki bulunmaktadır. "Daha fazla" ifadesi ile kullanılabilirler ancak nekadar daha fazla olduğunun ölçüsünü veremezler. Örneğim: Eğitim Düzeyi, Sosyoekonomik ölçek skorları gibi. Nominal veriler, ordinal verilere göre daha az bilgi taşırlar.
- 4-Ratio Veriler : Nümerik verilere benzerler. 100 santigrat derece, 50 santrigat derecenin iki katı denilemez ama derece kelvine çevrilirse 60 kelvin 30 kelvinin 2 misli sıcak denilebilir. Oran verilebilir veri türlerine Ratio veriler denir. Burada kelvin derece ratio türünden bir değişken iken, santigrat ise nümerik veri türüne örnek olarak verilebilir.

Nitelik Türleri

- Belli aralıkta yeralan değişkenler (interval)
  - sıcaklık, tarih
- İkili değişkenler (binary)
  - cinsiyet
- Ayrık ve sıralı değişkenler (nominal, ordinal, ratio scaled)
  - göz rengi, posta kodu

Problem

- Gerçek uygulamalarda toplanan veri kirli
  - eksik: bazı nitelik değerleri bazı nesneler için girilmemiş, veri madenciliği uygulaması için gerekli bir nitelik kaydedilmemiş
    - meslek = " "
  - gürültülü: hatalar var
    - maaş= "-10"
  - tutarsız: nitelik değerleri veya nitelik isimleri uyumsuz
    - yaş= "35", d.tarihi: "03/10/2004"
    - önceki oylama değerleri: "1,2,3", yeni oylama değerleri: "A,B,C"
    - bir kaynakta nitelik değeri 'ad', diğerinde 'isim'

Verinin Gürültülü Olma Nedenleri

- Eksik veri kayıtlarının nedenleri
  - Veri toplandığı sırada bir nitelik değerinin elde edilememesi, bilinmemesi
  - Veri toplandığı sırada bazı niteliklerin gerekliliğinin görülememesi
  - İnsan, yazılım ya da donanım problemleri
- Gürültülü (hatalı) veri kayıtlarının nedenleri
  - Hatalı veri toplama gereçleri
  - İnsan, yazılım ya da donanım problemleri
  - Veri iletimi sırasında problemler
- Tutarsız veri kayıtlarının nedenleri
  - Verinin farklı veri kaynaklarında tutulması
  - İşlevsel bağımlılık kurallarına uyulmaması

Sonuç

- Veri güvenilmez
  - Veri madenciliği sonuçlarına güvenilebilir mi?
  - Kullanılabilir veri madenciliği sonuçları kaliteli veri ile elde edilebilir.
- Veri kaliteli ise veri madenciliği uygulamaları ile yararlı bilgi bulma şansı daha fazla.

Veri Önişleme

- Veri temizleme
  - Eksik nitelik değerlerini tamamlama, hatalı veriyi düzeltme, aykırılıkları saptama ve temizleme, tutarsızlıkları giderme
- Veri birleştirme
  - Farklı veri kaynağındaki verileri birleştirme
- Veri dönüşümü
  - Normalizasyon ve biriktirme
- Veri azaltma
  - Aynı veri madenciliği sonuçları elde edilecek şekilde veri miktarını azaltma

Veriyi Tanıma

Veriyi Tanımlayıcı Özellikler

- Amaç: Veriyi daha iyi anlamak
  - Merkezi eğilim (central tendency), varyasyon, yayılma, dağılım
- Verinin dağılım özellikleri
  - Ortanca, en büyük, en küçük, sıklık derecesi, aykırılık, varyans
- Sayısal nitelikler -> sıralanabilir değerler
  - verinin dağılımı
  - kutu grafiği çizimi ve sıklık derecesi incelemesi

Merkezi Eğilimi Ölçme

Ortalama: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$

- ağırlıklı ortalama $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- kırpılmış ortalama: Uç değerleri kullanmadan hesaplama

Ortanca (median): Verinin tümü kullanılarak hesaplanır

- veri sayısı tek ise ortadaki değer, çift sayı ise ortadaki iki değerin ortalaması

$$median = L_1 + \left(\frac{n/2 - (\sum f)_l}{f_{median}}\right)c$$

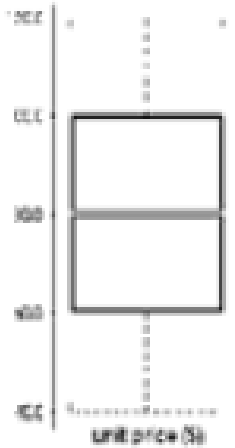
Mod

- Veri içinde en sıklıkla görülen değer
- Unimodal, bimodal, trimodal
- deneysel formül: $mean - mode = 3 \times (mean - median)$

Verinin Dağılımını Ölçme

Çeyrek, aykırılıklar, kutu grafiği çizimi

- Çeyrek (quartile) : nitelik değerleri küçükten büyüğe doğru sıralanır.
  - Q1: ilk %25, Q3: ilk %75
- Dörtlü aralık (Inter-quartile Range): IQR= Q3-Q1
- Five Number Summary: min, Q1, median, Q3, max
- Kutu Grafiği Çizimi:
  - Q1 ve Q3 aralığında bir kutu
  - kutu içinde ortanca noktayı gösteren bir çizgi
  - kutudan min ve max değerlere birer uzantı
- Aykırılıklar: 1,5xIQR değerinden küçük/büyük olan değerler



Varyans ve standart sapma

$$=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}[\sum_{i=1}^{n}x_i^2-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2] \qquad \sigma^2=\frac{1}{N}\sum_{i=1}^{n}(x_i-\mu)^2=\frac{1}{N}\sum_{i=1}^{n}x_i^2-\mu^2$$

Veri Temizleme

- Gerçek uygulamalarda veri eksik, gürültülü veya tutarsız olabilir.
- Veri temizleme işlemleri
  - Eksik nitelik değerlerini tamamlama
  - Aykırılıkların bulunması ve gürültülü verinin düzeltilmesi
  - Tutarsızlıkların giderilmesi

Eksik Veri

- Veri için bazı niteliklerin değerleri her zaman bilinemeyebilir.
- Eksik veri
  - diğer veri kayıtlarıyla tutarsızlığı nedeniyle silinmesi
  - bazı nitelik değerleri hatalı olması dolayısıyla silinmesi
  - yanlış anlama sonucu kaydedilmeme
  - veri girişi sırasında bazı nitelikleri önemsiz görme

Eksik Veriler nasıl Tamamlanır?

- Eksik nitelik değerleri olan veri kayıtlarını kullanma
- Eksik nitelik değerlerini elle doldur
- Eksik nitelik değerleri için global bir değişken kullan (Null, bilinmiyor,...)
- Eksik nitelik değerlerini o niteliğin ortalama değeri ile doldur
- Aynı sınıfa ait kayıtların nitelik değerlerinin ortalaması ile doldur
- Olasılığı en fazla olan nitelik değeriyle doldur

Gürültülü Veri

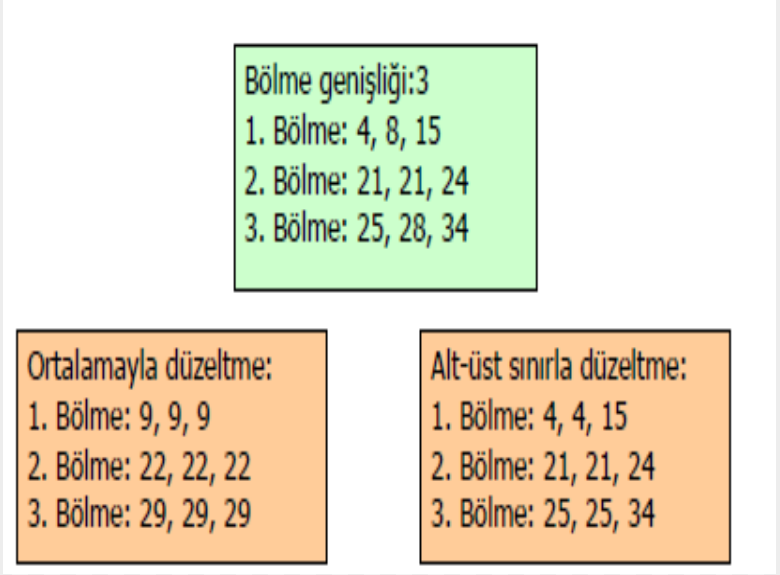
- Ölçülen bir değerdeki hata
- Yanlış nitelik değerleri
  - hatalı veri toplama gereçleri
  - veri girişi problemleri
  - veri iletimi problemleri
  - teknolojik kısıtlar
  - nitelik isimlerinde tutarsızlık

Gürültülü Veri nasıl düzeltilir?

- Gürültüyü yok etme
  - Bölmeleme
    - veri sıralanır, eşit genişlik veya eşit derinlik ile bölünür
  - Kümeleme
    - aykırılıkları belirler
  - Eğri uydurma
    - veriyi bir fonksiyona uydurarak gürültüyü düzeltir.

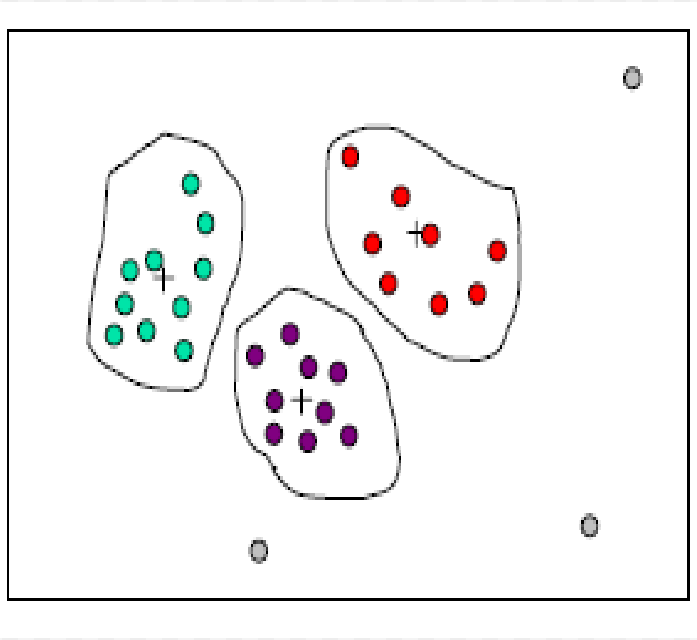
Bölmeleme

- Veri sıralanır: 4, 8, 15, 21, 21, 24, 25, 28, 34
  - Eşit genişlik: Bölme sayısı belirlenir. Eşit aralıklarla bölünür
  - Eşit derinlik: Her bölmede eşit sayıda örnek kalacak şekilde bölünür.
    - her bölme ortalamayla ya da bölmenin en alt ve üst sınırlarıyla temsil edilir.



Kümeleme

- Benzer veriler aynı kümede olacak şekilde gruplanır
- Bu kümelerin dışında kalan veriler aykırılık olarak belirlenir ve silinir.



Eğri Uydurma

- Veri bir fonksiyona uydurulur. Doğrusal eğri uydurmada, bir değişkenin değeri diğer bir değişken kullanılarak bulunabilir.

Veri Birleştirme

Veri Birleştirme

- Farklı kaynaklardan verilerin tutarlı olarak birleştirilmesi
- Şema birleştirilmesi
  - Aynı varlıkların saptanması
  - meta veri kullanılır
- Nitelik değerlerinin tutarsızlığının saptanması
  - Aynı nitelik için farklı kaynaklarda farklı değerler olması
  - Farklı metrikler kullanılması

Gereksiz Veri

Farklı veri kaynaklarından veriler birleştirilince gereksiz (fazla) veri oluşabilir

- aynı nitelik farklı kaynaklarda farklı isimle
- bir niteliğin değeri başka bir nitelik kullanılarak hesaplanabilir
  - korelasyon hesaplaması: sayısal nitelikler
  - =0: nitelikler bağımsız, >0: pozitif korelasyon, <0: negatif korelasyon

$$r_{A,B} = \frac{\sum (A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A \sigma_B} \qquad \overline{A} = \frac{\sum A}{n} \qquad \sigma_A = \sqrt{\frac{\sum (A-\overline{A})^2}{n-1}}$$

  - korelasyon hesaplaması: ayrık nitelikler (chi-square test)

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

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
- z-score normalizasyon
- ondalık normalizasyon