Veri Madenciliği

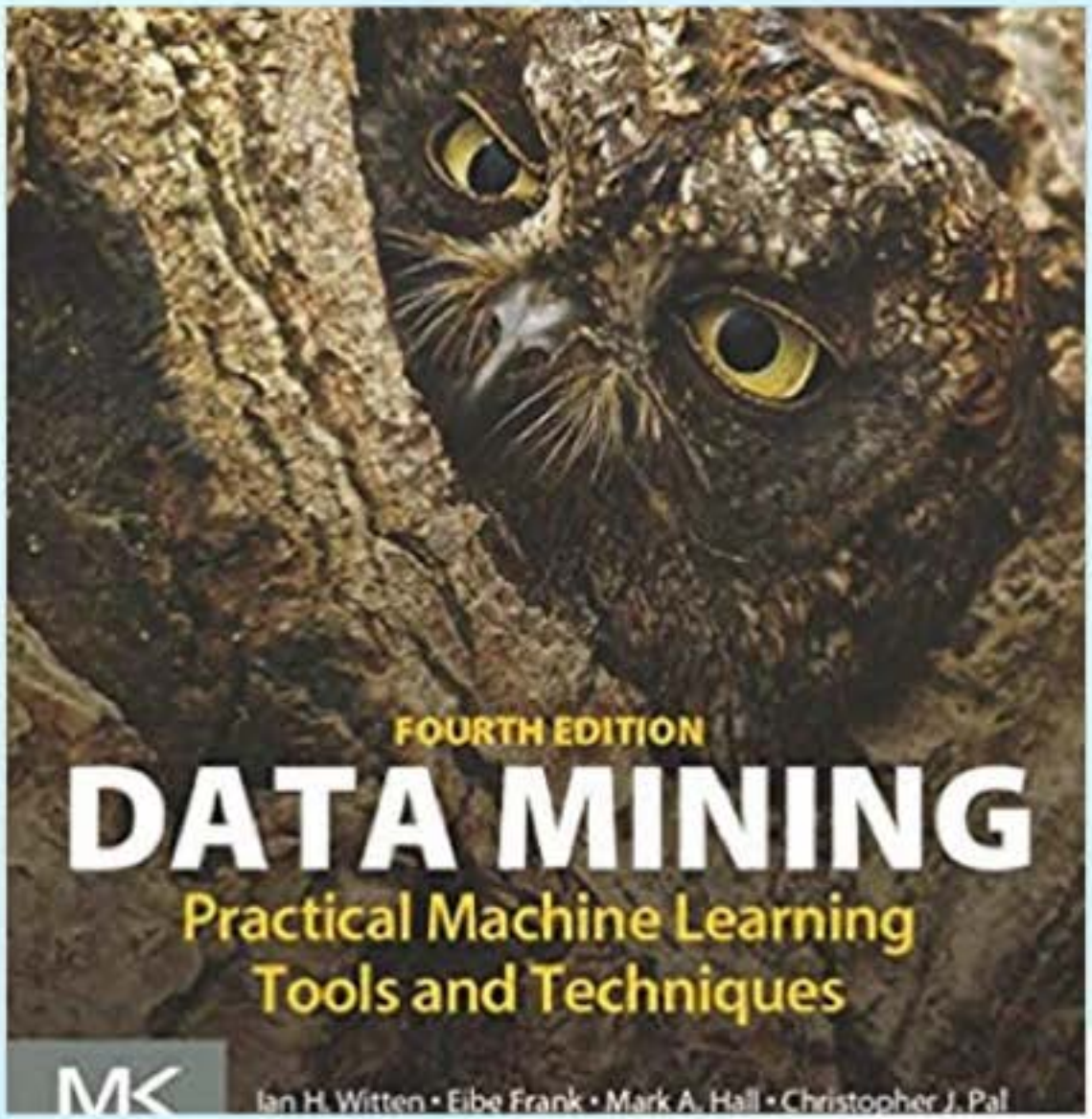
Güz 2023

Ders 5

- Öznitelik
- Girdi Hazırlama

Dersin Kitabı

- Data Mining: Practical Machine Learning Tools and Techniques, 4th Ed., by Ian Witten, Eibe Frank, Mark Hall, and Christopher Pal (Morgan Kaufmann Publishers. 2017. ISBN: 978-0-12-804291-5)



Girdi Hazırlama

Denormalizasyon tek sorun değil.....

Sorun: farklı veri kaynakları (örneğin satış departmanı, müşteri fatura departmanı, ...)

- Farklılıklar: kayıt tutma stilleri, kurallar, zaman dönemleri, veri toplama, birincil anahtarlar, hatalar
- Veriler bir araya getirilmeli, entegre edilmeli, temizlenmelidir
- "Veri ambarı": tutarlı erişim noktası

  Dış veriler gerekebilir ("yer paylaşımı verileri")

Kritik: veri toplama türü ve düzeyi, genellikle doğru almak için birçok yineleme gerekir.

ARFF Biçimi

Attribute-Relation File Format

```
%
Bazı sayısal özelliklere sahip hava durumu
 verileri için ARFF dosyası
%
@relation hava durumu

@attribute görünüm {güneşli, bulutlu, yağmurlu}
@attribute sıcaklık sayısal
@attribute nem sayısal
@attribute rüzgarlı {doğru, yanlış}
@attribute oyun? {evet, hayır}
@data
güneşli, 85, 85, yanlış, hayır
güneşli, 80, 90, doğru, hayır
bulutlu, 83, 86, yanlış, evet
...
```

Öznitelik açıklaması

Not: *sınıf ayırt edici değildir, bu dosyanın birden çok şekilde kullanılmasına izin verir*

Örnekler

Ek Öznitelik Türleri

ARFF dize özniteliklerini destekler:

```
@attribute açıklama dize
```

Nominal özniteliklere benzer, ancak değer listesi önceden belirtilmemiş

Ayrıca tarih özniteliklerini de destekler:

```
@attribute bugün tarih
```

ISO-8601 birleşik tarih ve saat biçimi yyyy-MM-dd-'T'HH:mm:ss kullanır

İlişkisel Öznitelikler

Çok örnekli sorunların ARFF biçiminde temsil edilmesine izin ver

İlişkisel özniteliğin değeri ayrı bir örnek kümesidir

```
@attribute çanta ilişkisel
        @attribute görünüm {güneşli, bulutlu, yağmurlu}
        @attribute sıcaklık sayısal
        @attribute nem sayısal
        @attribute rüzgarlı {doğru, yanlış}
@end çanta
```

İç içe öznitelik bloğu başvurulan örneklerin yapısını verir

Çok örnekli ARFF

```
Hava durumu verileri için birden çok örnek ARFF dosyası
@relation hava durumu
@attribute torba_ID { 1, 2, 3, 4, 5, 6, 7 }
@attribute torba ilişkisel
    @attribute görünüm {güneşli, bulutlu, yağmurlu}
    @attribute sıcaklık sayısal
    @attribute nem sayısal
    @attribute rüzgarlı {doğru, yanlış}
@end torba
@attribute oyun? {evet, hayır}
@data
% yedi "birden çok örnek" örneği
1, "güneşli, 85, 85, yanlış \ngüneşli, 80, 90, doğru", hayır
2, "bulutlu, 83, 86, yanlış \nyağmurlu, 70, 96, yanlış", evet
3, "yağmurlu, 68, 80, yanlış \nyağmurlu, 65, 70'ler, doğru", evet
.....
```

\n, iki gün süren oyunların koşullarını temsil eden iki örneği ayırır

Seyrek Veriler

- Bazı uygulamalarda çoğu öznitelik değerleri veri kümesi sıfırdır.
- Örneğin, bir metin kategorizasyonu problemindeki sözcük sayıları ARFF seyrek verileri destekler

```
0, 26, 0,  0, 0, 0, 63, 0, 0, 0, "sınıf A"
0,  0, 0, 42, 0, 0,  0, 0, 0, 0, "sınıf B"

{1, 26, 6, 63, 10, "sınıf A"}
{3, 42,          10, "sınıf B"}
```

Bu aynı zamanda WEKA'daki nominal öznitelikler için de çalışır, çünkü dahili nominal öznitelikler sayı olarak depolanır (numaralandırmalara benzer şekilde, ilk değer "sıfıra" karşılık gelir)

Not: Seyrek bir örnekte atlanan değerler "eksik" değil, 0'dır -- bir değer bilinmiyorsa, açıkça soru işaretiyle (?) temsil edilmelidir.

Öznitelik Türleri

ARFF'deki öznitelik türlerinin yorumlanması öğrenme şemasına bağlıdır

- Sayısal öznitelikler şöyle yorumlanır:
  - sıra ölçeklerinde if *less-than* ve *greater-than* kullanılır
  - mesafe hesaplamaları yapılıyorsa oran ölçekleri (normalleştirme/standardizasyon gerekebilir)
- Örnek tabanlı düzenler nominal değerler arasındaki mesafeyi tanımlar (değerler eşitse 0, aksi takdirde 1)

Belirli bir veri dosyasındaki tamsayılar: nominal, sıralı veya oran ölçeği?

Nominal vs. Sıralı

Öznitelik yaşı nominal

Eşitliği kontrol et!

```
If yaş = genç ve astigmatik = hayır
   ve gözyaşı üretim oranı = normal
   then öneri = yumuşak
If yaş = presbiyopik ve astigmatik = hayır
   ve gözyaşı üretim oranı = normal
   then öneri = yumuşak
```

Öznitelik yaş sırası

Sıralı!

(Örn. "genç" < "presbiyopik öncesi" < presbiyopik

```
If yaş ≤ presbiyopik_öncesi
   ve astigmatik = hayır
   ve gözyaşı üretim oranı = normal
   then öneri = yumuşak
```

Eksik Değerler

Sıklık aralık dışı girdiler tarafından belirtilir

- Tür: bilinmeyen, kayda alınamayan, ilgisiz
- Sebep:
  - arızalı ekipman
  - deneysel tasarımdaki değişiklikler
  - farklı veri kümelerinin harmanlanması
  - ölçümün mümkün olmaması
  - yanıtlayanların bilgi sağlamayı reddetmesi (örn. gelir)

Eksik değer kendi içinde önemi olabilir (örneğin tıbbi muayenede yapılmayan bir test)

Çoğu şema durumun böyle olmadığını varsayar:

"Eksik" değerin ek değer olarak kodlanmış olması gerekebilir

Soru: Eksik değerlerle nasıl başa çıkağız?

Yanlış Değerler

Sebep: madencilik için veri toplanamaması

Sonuç: verilerin orijinal amacını etkilemeyen hatalar ve eksiklikler (örn. müşteri yaşı)

Nominal özniteliklerdeki tipografik hatalar ==> değerlerin tutarlılık açısından denetlenmeleri gerekir

Sayısal özniteliklerde tipografik ve ölçüm hataları ==> aykırılıkların tespit edilmesi gerekir

Hatalar kasıtlı olabilir (örn yanlış posta kodları)

Diğer problemler: yinelenenler, eski veriler

Verileri Tanıma

Dengesiz Veriler: Eğer bir yanıt zamanın %99'unun doğruysa, neden her zaman bu yanıtı vermiyorsunuz?

Nadir sonucu tahmin edememe maliyetlerini tartmanız gerekir (örn. ölümcül hastalık tanısı)

Basit görselleştirme araçları yararlıdır

- **Nominal öznitelikler**: histogramlar (Arka plan bilgisi ile tutarlı dağılım mı?)
- **Sayısal öznitelikler**: grafikler (Belirgin aykırılıklar var mı?)

2-D ve 3-D çizimleri bağımlılıkları gösterir

Alan uzmanlarına danışmanız gerekiyor

İncelenemeyecek kadar çok veri var mı? **Örnek alın!**