Veri Madenciliği

Yapay Zekâ ve Veri Dersleri Rehberi · Kısım IV: Veri

Bilgi keşfi süreci | Birliktelik kuralları | Kümeleme | Aykırı değerler | İlginçlik ölçütleri | Komşu dersler

Sercan KÜLCÜ · Giresun Üniversitesi · Bilgisayar Mühendisliği

© Sercan Külcü. Tüm hakları saklıdır. · Sunumun etkileşimli HTML sürümü: https://sercankulcu.github.io/files/ai_courses/sunumlar/Bolum_25_Veri_Madenciligi.html

GİRİŞ

Ders Hangi Soruyu Sorar?

> **Büyük veri yığınlarındaki gizli ve işe yarar örüntüler nasıl bulunur?**

Veri madenciliği, büyük veri yığınlarında önceden bilinmeyen, geçerli ve işe yarar **örüntüleri keşfetmeyi** inceler. Makine öğrenmesiyle aynı algoritmaları kullanır; ama amaç yeni örnekte tahmin değil, verinin kendisi hakkında kural, küme ve aykırılık bulmaktır.

Bir bakışta

YAKLAŞIM	GİRDİ	ÇIKTI
keşfe dayalı çözümleme	büyük tablo ya da işlem verisi	kural, küme, aykırı değer

BAŞARI ÖLÇÜTÜ	MATEMATİK YÜKÜ	VERİ İHTİYACI
ilginçlik, destek, güven	orta	çok

Kısa Tarihçe

1967
MacQueen "k-ortalamalar" (k-means) adını verdiği kümeleme yöntemini yayımlar.

1989
İlk "veri tabanlarında bilgi keşfi" (KDD) çalıştayı düzenlenir; alan adını bulur.

1994
Agrawal ve Srikant Apriori algoritmasıyla birliktelik kuralı madenciliğini yaygınlaştırır.

1996
Fayyad ve arkadaşları KDD sürecini tanımlar; aynı yıl yoğunluğa dayalı DBSCAN yayımlanır.

2000
CRISP-DM 1.0 yayımlanır: veri madenciliği projeleri için sektörden bağımsız süreç modeli.

2006
Netflix Prize açılır: öneri doğruluğunu yüzde 10 artırana bir milyon dolar ödül.

DERS

Temel Kavramlar

Bilgi keşfi (KDD)
Seçme, ön işleme, dönüştürme, madencilik, yorum.

Sık öğe kümeleri
Birlikte sık görülen ürünler: Apriori, FP-growth.

Birliktelik kuralı
X ise Y; destek, güven ve kaldıraç (lift) ile ölçülür.

Kümeleme
k-ortalamalar, hiyerarşik, DBSCAN; etiketsiz gruplar.

Aykırı değer
Çoğunluğa benzemeyen kayıt: hata mı, sahtekârlık mı?

Tanımlayıcı model
Karar ağacı gibi okunabilir, açıklayan modeller.

Ön işleme
Eksik değer, gürültü, ayrıklaştırma, normalleştirme.

Dizisel örüntü
Zaman sırasıyla tekrarlanan olay dizileri.

DERS

Tipik İş Akışı



Örnek: Bir market bir yılın kasa fişlerini seçer, iadeleri ve hatalı kayıtları ayıklar, her fişi ürün kümesine dönüştürür, Apriori ile "makarna alan sos da alır" gibi kurallar bulur ve kaldıracı yüksek kuralları raf düzeni için yöneticiye sunar.

DERS

Derste Neler İşlenir?

1. Veri madenciliği süreci: KDD ve CRISP-DM
2. Veriyi tanımak: öznitelik türleri, benzerlik
3. Ön işleme: temizleme, bütünleştirme, indirgeme
4. Veri ambarı ve OLAP küpleri
5. Sık öğe kümeleri: Apriori ve FP-growth
6. Birliktelik kuralları ve ilginçlik ölçütleri
7. Sınıflandırma: karar ağaçları, kural tabanlı yöntemler
8. Sınıflandırıcı değerlendirme ve dengesiz sınıflar
9. Kümeleme: k-ortalamalar ve hiyerarşik yöntemler
10. Yoğunluğa dayalı kümeleme ve küme geçerliliği
11. Aykırı değer tespiti
12. Dizisel örüntü, metin ve çizge madenciliği

Kazandırdığı beceriler

pandas ve SQL ile veri hazırlamak | mlxtend ile birliktelik kuralı çıkarmak | Weka ya da Orange ile hızlı deney | kümeleme sonucunu ölçüp yorumlamak | bulguyu iş diline çevirmek

Tipik bir dönemin akışıdır; sıra ve kapsam üniversiteden üniversiteye değişir.

DENE

Market Sepetinde Gizli Kurallar

Mahalle marketi 14 kasa fişine bakıyor: kimse bir soru sormadı, hangi ürünlerin birlikte alındığını veri söylesin. Eşikleri değiştirin, Apriori kuralları yeniden bulsun; bir kurala tıklayıp sepetlerde görün.



En az destek: **3 (%21)** sepet

En az güven: **%60**

Apriori: 1-li: 7 aday → **6** sık; 2-li: 15 aday → **10** sık; 3-li: 7 aday → **4** sık. **18** kural güven eşiğini geçti, 4 tanesi:

{Simit} → {Çay}
destek %43 · güven %100 · kaldıraç 2,00

{Ekmek, Peynir} → {Zeytin}
destek %43 · güven %100 · kaldıraç **2,00**

{Ekmek, Simit} → {Çay}
destek %21 · güven %100 · kaldıraç **2,00**

{Zeytin} → {Ekmek}
destek %43 · güven %86 · kaldıraç **1,09**

Veri madenciliği **keşiftir**: soru sorulmadan veride sık tekrar eden kalıplar aranır. Kaldıraç 1 civarındaysa kural boştur (Ekmek zaten herkesin sepetinde); işe yarayan kural kaldıracı 1'in belirgin üstünde olandır.

KONUM

Ön Koşullar

Bu dersten önce bilinmesi yararlı olan dersler ve neden gerektikleri.

Ders	Neden gerekir
Olasılık ve İstatistik	destek, güven, anlamlılık
Veritabanı ve SQL	tablo, birleştirme ve gruplama: verinin kaynağı
Veri Yapıları	karma tablo ve ağaç: FP-growth gibi algoritmaların yapı taşları
Algoritmalar	karmaşıklık: aday kümelerin üstel büyümesini anlamak

KONUM

Bu Ders Sana Göre mi?

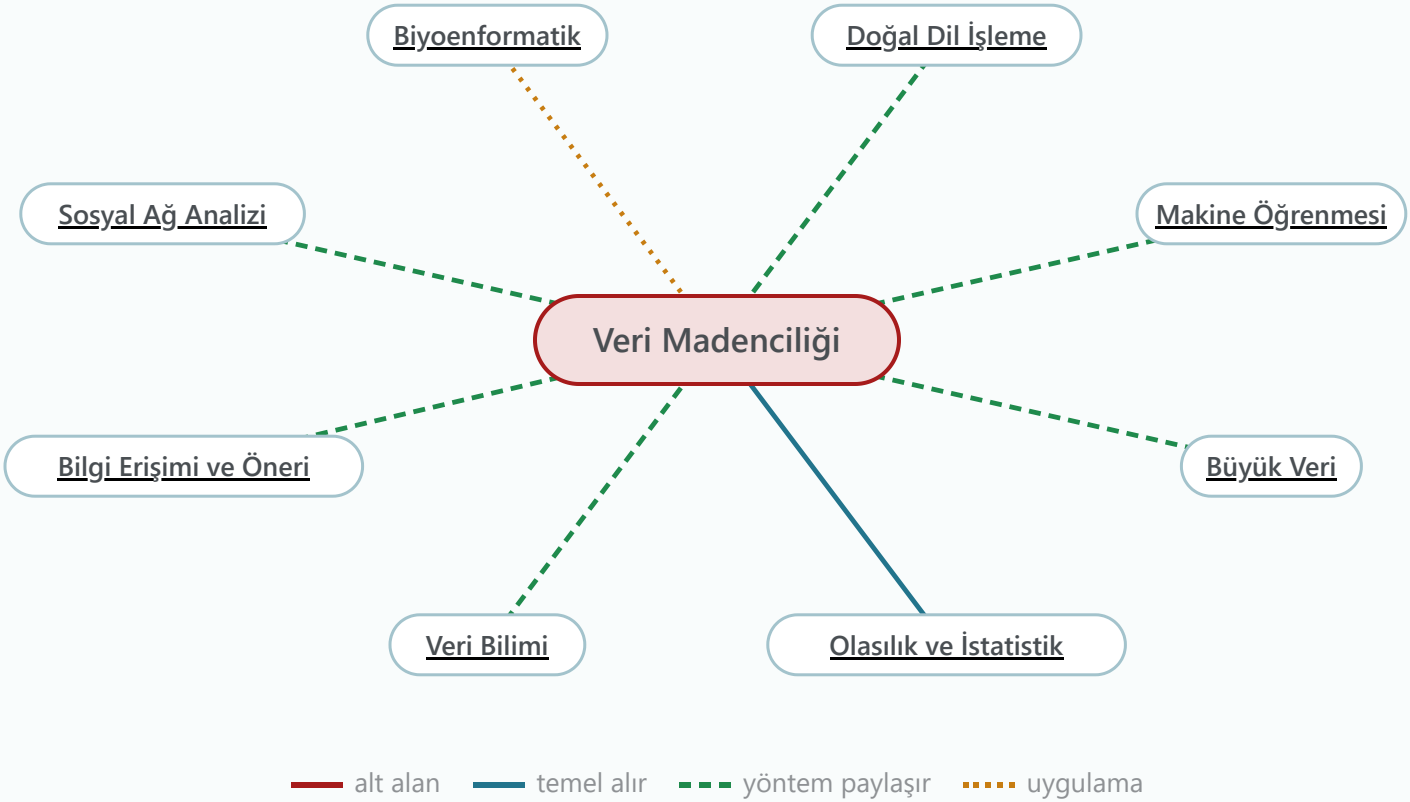
Kendine sor; "hayır" dediklerin için sağdaki sütun nereden başlayacağını gösterir.

Kendine sor	Hayırsa önce
İki tabloyu bir SQL sorgusuyla birleştirip gruplayabiliyor musun?	Bir veritabanı dersindeki SQL: JOIN ve GROUP BY
Koşullu olasılığı ve bağımsızlığı bir örnekle açıklayabiliyor musun?	Olasılık ve İstatistik
Karma tablosunun ne zaman ağaçtan iyi olduğunu açıklayabiliyor musun?	Veri Yapıları dersi: karma tabloları ve ağaçlar
Python'da bir CSV dosyasını okuyup eksik değerleri sayabiliyor musun?	Programlamaya giriş, ardından pandas'a kısa bir giriş

Çoğuna "evet" diyebiliyorsan derse hazırsın; bir ya da iki "hayır" dönem içinde kapanabilir.

KONUM

Dersin Komşuları



ÖRTÜŞME

Örtüştüğü Dersler

İlişki	Ders	Ortak konu
temel alır	Bu ders Olasılık ve İstatistik dersini temel alır	destek, güven, anlamlılık
uygulama	Biyoenformatik bu dersin yöntemlerini uygular	dizi motifleri, kümeleme
yöntem paylaşır	Makine Öğrenmesi ile yöntem paylaşır	aynı algoritmalar; amaç tahmin değil keşif
yöntem paylaşır	Doğal Dil İşleme ve Metin Madenciliği ile yöntem paylaşır	metin sınıflandırma, konu modelleme
yöntem paylaşır	Veri Bilimi ile yöntem paylaşır	keşifsel veri analizi
yöntem paylaşır	Büyük Veri Sistemleri ile yöntem paylaşır	ölçeklenebilir madencilik
yöntem paylaşır	Bilgi Erişimi ve Öneri Sistemleri ile yöntem paylaşır	birliktelik, işbirlikçi süzme
yöntem paylaşır	Sosyal Ağ Analizi ile yöntem paylaşır	topluluk bulma

Ayrıştığı Noktalar (1/2)

Ölçüt	Veri Madenciliği	Makine Öğrenmesi	Veri Bilimi
Temel soru	Büyük veri yığınlarındaki gizli ve işe yarar örüntüler nasıl bulunur?	Bir program örneklerden nasıl genelleme yapar?	Veriden doğru soruları sorup karara dönüşen sonuçlar nasıl çıkarılır?
Yaklaşım	keşfe dayalı çözümleme	istatistiksel öğrenme	soru, veri, analiz, anlatı
Girdi	büyük tablo ya da işlem verisi	etiketli ya da etiketsiz örnekler	ham, dağınık veri
Çıktı	kural, küme, aykırı değer	tahmin yapan model	rapor, görselleştirme, model
Başarı ölçütü	ilginçlik, destek, güven	görülmemiş veride hata	kararlara katkı, doğruluk
Matematik yükü	orta	yüksek: doğrusal cebir, olasılık, optimizasyon	orta: istatistik
Veri ihtiyacı	çok	orta ve çok	orta ve çok

- **Makine Öğrenmesi:** Aynı algoritmaları kullanırlar; MÖ yeni örnekte tahmini, veri madenciliği eldeki veride keşfi amaçlar.
- **Veri Bilimi:** Veri bilimi sorudan karara bütün süreci kapsar; veri madenciliği bu süreçteki örüntü bulma adımına odaklanır.

Ayrıştığı Noktalar (2/2)

Ölçüt	Veri Madenciliği	Büyük Veri	Olasılık ve İstatistik
Temel soru	Büyük veri yığınlarındaki gizli ve işe yarar örüntüler nasıl bulunur?	Tek bir makineye sığmayan veri nasıl saklanır ve işlenir?	Belirsizlik nasıl ölçülür ve sınırlı veriden nasıl güvenilir sonuç çıkarılır?
Yaklaşım	keşfe dayalı çözümleme	dağıtık depolama ve hesaplama	olasılık modeli ve çıkarım
Girdi	büyük tablo ya da işlem verisi	çok büyük ve hızlı akan veri	örneklem verisi
Çıktı	kural, küme, aykırı değer	işlenmiş veri, sorgu sonucu	kestirim, güven aralığı, test sonucu
Başarı ölçütü	ilginçlik, destek, güven	ölçeklenme, gecikme, maliyet	yansızlık, güç, anlamlılık
Matematik yükü	orta	düşük	kuramın kendisi
Veri ihtiyacı	çok	çok fazla	örneklem

- **Büyük Veri:** Büyük veri sistemleri veriyi saklar ve dağıtık işler; veri madenciliği o veride hangi örüntünün aranacağını sorar.
- **Olasılık ve İstatistik:** İstatistik önceden kurulmuş bir hipotezi sınar; veri madenciliği hipotezi veriden üretir, sonra sınanması gerekir.

UYGULAMA

Uygulama Alanları

Market sepeti
Birlikte alınan ürünler, raf ve kampanya.

Müşteri bölütleme
Benzer müşterileri gruplamak.

Sahtekârlık tespiti
Olağandışı işlemleri öne çıkarmak.

Web kullanım
Ziyaretçilerin gezinme örüntüleri.

Arıza kayıtları
Birlikte görülen hata kodları.

Tıbbi kayıtlar
Birlikte görülen tanı ve ilaçlar.

UYGULAMA

Riskler ve Etik

Hassas bilgiyi çıkarsamak

Alışveriş sepetinden hamilelik, hastalık ya da maddi sıkıntı çıkarılabilir; kişi bunu hiç paylaşmamış olsa bile.

Anonim veri anonim kalmaz

Netflix Prize verisi IMDb puanlarıyla eşleştirilip bazı kullanıcılar tanındı; ikinci yarışma bu yüzden iptal edildi.

Veri tarama

Yeterince çok örüntü aranırsa tesadüfi olanlar da anlamlı görünür; çoklu karşılaştırma düzeltmesi ve yeni veride sınama gerekir.

Ayrımcı profilleme

Posta koduna ya da alışveriş örüntüsüne göre kurulan müşteri kümeleri, bazı gruplara dolaylı olarak farklı fiyat ve hizmet getirir.

Ayrıntılı tartışma: Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği.

UYGULAMA

Araçlar ve Örnek Veriler

Araçlar ve kütüphaneler

Weka | KNIME | Orange | RapidMiner | scikit-learn | mlxtend | R (arules)

Tanınmış veri setleri

Groceries (arules) | Online Retail (UCI) | Instacart Market Basket | Adult (gelir) | KDD Cup 1999 | Mushroom

UYGULAMA

Sık Karıştırılanlar

Yaygın düşünce	Doğrusu
Veri madenciliği veri toplamaktır.	Ad yanıltıcıdır: kazılan veri değil bilgidir; veri zaten vardır, örüntü aranır.
Bulunan her kural işe yarar.	Binlerce kural çıkar; çoğu bariz ya da tesadüfidir. İlginçlik ölçütleri ve uzman süzgeci gerekir.
Güçlü birliktelik neden-sonuçtur.	Birlikte görülmek birinin ötekine yol açtığını göstermez; ortak bir neden olabilir.
Kümelemenin tek doğru sonucu vardır.	Sonucu küme sayısı, uzaklık ölçüsü ve ölçekleme belirler; farklı seçim, farklı küme.

SONRASI

Bu Dersten Sonra

Bu dersi temel alan ya da genişleten dersler

Biyoenformatik

İlgili meslekler

Veri madenciliği uzmanı

Veri analisti

İş zekâsı uzmanı

Müşteri analitiği uzmanı

Dolandırıcılık analisti

Veri madenciliği, veri kısmının keşif ayağıdır. Buradan kurallarla öneri sistemlerine, topluluk bulmayla sosyal ağ analizine, ölçeklenebilir madencilikle büyük veri sistemlerine ya da tahmine odaklanmak için makine öğrenmesine geçilebilir.

SONRASI

Kaynaklar

- Han, Kamber ve Pei, *Data Mining: Concepts and Techniques*, 3. baskı, Morgan Kaufmann, 2011.
- Tan, Steinbach, Karpatne ve Kumar, *Introduction to Data Mining*, 2. baskı, Pearson, 2019.
- Witten, Frank, Hall ve Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4. baskı, Morgan Kaufmann, 2017.
- Leskovec, Rajaraman ve Ullman, *Mining of Massive Datasets*, 3. baskı, Cambridge University Press, 2020.

Son

Bölüm 25: Veri Madenciliği

← Önceki bölüm · Alanın haritası · Sonraki bölüm →