Veri Madenciliği Yöntemleri

Dr. Yalçın ÖZKAN

Veri Madenciliği Yöntemleri

Dr. Yalçın ÖZKAN

PAPATYA

© **PAPATYA YAYINCILIK EĞİTİM**
Bilgisayar Sis. San. ve Tic. A.Ş.
Ankara Caddesi, Prof. Fahreddin Kerim Gökay Vakfı İşhanı Girişi,
No: 11/3, Cağaloğlu (Fatih)/İstanbul

Tel : (+90 212) 527 52 96 - (+90 532) 311 31 10
Faks : (+90 212) 527 52 97
e-Posta : admin@papatya.gen.tr
Web : www.papatya.gen.tr www.papatyabilim.com.tr

Veri Madenciliği Yöntemleri – Dr. Yalçın ÖZKAN

2. Basım Ocak 2013
3. Basım Eylül 2016

Editör : Dr. Rifat ÇÖLKESEN
Yardımcı Editör : Dr. Cengiz UĞURKAYA (Post-Edu Institute)
Üretim : Necdet AVCI
Satış : Mustafa DEMİR
Sayfa Düzenleme : Papatya - Kelebek Tasarım
Kapak Tasarım : Papatya - Kelebek Tasarım
Basım ve Ciltleme : Özkaracan Matbaacılık (Sertifika No: 12228)
Evren Mah. Gülbahar Cad. No:62 Güneşli/İstanbul

© Bu kitabın her türlü yayın hakkı **Papatya Yayıncılık Eğitim A.Ş.**'ye aittir. Yayınevinden yazılı izin alınmaksızın alıntı yapılamaz, kısmen veya tamamen hiçbir şekil ve teknikle ÇOĞALTILAMAZ, BASILAMAZ, YAYIMLANAMAZ. Kitabın, tamamı veya bir kısmının fotokopi makinesi, ofset gibi teknikle çoğaltılması, hem çoğaltan hem de bulunduranlar için yasadışı bir davranıştır.

Lütfen kitabımızın fotokopi yöntemiyle çoğaltılmasına engel olunuz. Fotokopi hızsızlıktır.

Yalçın, Özkan.
Veri Madenciliği Yöntemleri / Yalçın Özkan - İstanbul: Papatya Yayıncılık Eğitim, 2013.
x, 240 s.; 24 cm.
Kaynakça ve dizin var.
Sertifika No: 11218
ISBN 978-975-6797-82-2
1. Veri Ambarı 2. Bayes Teoremi 3. Kümeleme 4. K-Komşu Algoritması 5. Gini Algoritması
I. Title

Çok değerli hocam,
bilim adamı ve büyük insan
Prof. Dr. Kenan Ural'a...

İçindekiler

Önsöz **9**

Bölüm 1. **VERİ MADENCİLİĞİ VE VERİ AMBARI** **11**

1.1. Veriyi Bilgiye Dönüştürmenin Yolu 11
1.2. Veri Madenciliği 12
1.3. Veri Madenciliği için Veri Kaynakları 12
1.3.1. Veritabanı Sistemleri 13
1.3.2. Veri Ambarı 15
1.3.3. Veri Kümeleri 16
1.3.4. Veri Ambarının Tasarlanması 15
1.3.4.1. Konuya Yöneliktir 15
1.3.4.2. Bütünleşiktir 17
1.3.4.3. Zaman Boyutu Vardır 17
1.3.4.4. Sadece Okunabilir 18
1.3.5. Veri Ambarının Temel Özellikleri 19
1.3.6. Veri Ambarının İçerdiği Veri 20
1.3.6.1. Metedata 20
1.3.6.2. Ayrıntı Veri 21
1.3.6.3. Eski Ayrıntı Veri 21
1.3.6.4. Düşük Düzeyde Örneklenmiş Veri 21
1.3.6.5. Yüksek Düzeyde Örneklenmiş Veri 21
1.3.7. Veri Ambarı Veri Modeli 21
1.4. Özet 22
1.5. Sorular 23

Bölüm 2. **VERİ MADENCİLİĞİ SÜRECİ VE UYGULAMALARI** **25**

2.1. Uygulama Alanları 25
2.2. Veri Madenciliği Süreci 27
2.2.1. Veri Temizleme 27
2.2.2. Veri Bütünleştirme 28
2.2.3. Veri İndirgeme 28
2.2.4. Veri Dönüştürme 29
2.2.4.1. *Min-Max* Normalleştirilmesi 29
2.2.4.2. *Z-score* Standartlaştırma 30
2.2.5. Veri Madenciliği Algoritmasını Uygulama 31
2.2.6. Sonuçları Sunum ve Değerlendirme 32
2.3. Veri Madenciliği Yöntemleri 32
2.3.1. Sınıflandırma 32

2.3.2. Kümeleme 34
2.3.3. Birliktelik Kuralları 36
2.4. Özet 50
2.5. Sorular 50

Bölüm 3. **KARAR AĞAÇLARI İLE SINIFLANDIRMA** **39**

3.1. Sınıflandırma 39
3.2. Sınıflandırma Süreci 40
3.3. Karar Ağaçları ile Sınıflandırma 41
3.4. Karar Ağaçlarında Dallanma Kriterleri 42
3.5. ID3 Algoritması 42
3.5.1. Entropi 43
3.5.2. Karar Ağacında Entropi 44
3.5.3. Dallanma İçin Niteliklerin Seçilmesi ve Kazanç Ölçütü 45
3.5.4. ID3 Algoritması 48
3.5.5. Uygulama 48
3.5.6. Kazanç Oranı 63
3.6. C4.5 Algoritması 65
3.6.1. Sayısal Değerlere Sahip Nitelikler 65
3.6.2. Uygulama 65
3.6.3. Bilinmeyen Öznitelik Değerleri 76
3.7. Karar Ağaçlarının Budanması 78
3.7.1. C4.5'de Budama 79
3.8. Karar Kuralları Oluşturmak 80
3.9. Sınıflandırma Modelinin Doğruluğu 81
3.9.1. Öngörülerin Elde Edilmesi 81
3.9.2. Karışıklık Matrisi 83
3.9.3. İki Sınıflı Modeller için Doğruluk Ölçütleri 84
3.9.4. Aşırı Öğrenme Durumu 86
3.9.5. Doğrulama Süreci 87
3.9.5.1. Veri Kümesinin Eğitim ve Test için Bölünmesi 87
3.9.5.2. *Holdout* Yöntemi 88
3.9.5.3. *k-Katlı* Çapraz Doğrulama 89
3.9.5.4. Birini Dışarıda Bırakan Çapraz Doğrulama 89
3.9.5.5. *Bootstarp* 90
3.10. Özet 90
3.11. Sorular 91

Bölüm 4. **SINIFLANDIRMA VE REGRESYON AĞAÇLARI** **93**

4.1. *Twoing* Algoritması 93
4.1.1. Uygulama 97

4.1.2. Modelin Başarımı 110
4.2. *Gini* Algoritması 111
4.2.1. Uygulama 113
4.2.2. Sürekli Değerlerin Kullanım 120
4.2.3. Uygulama: Sayısal Değerler için *Gini* Algoritması 120
4.3. Regresyon Ağaçları 132
4.4. Özet 138
4.4. Sorular 139

Bölüm 5. **EN YAKIN K-KOMŞU ALGORİTMASI İLE SINIFLAMA** **141**

5.1. En Yakın k-Komşu Algoritması 142
5.1.1. Uygulama 1 143
5.1.2. Uygulama 2 146
5.2. Ağırlıklı Oylama 149
5.2.1. Uygulama 3 149
5.3. Özet 153
5.4. Sorular 154

Bölüm 6. **BAYES SINIFLANDIRICILAR** **155**

6.1. Koşullu olasılık 155
6.2. *Bayes* Teoremi 156
6.3. *Bayes* Sınıflandırıcısı 157
6.3.1. Sade *Bayes* Sınıflandırıcısı 157
6.3.2. Uygulama 1 158
6.3.3. Sınıflandırma Modelinin Performansı 161
6.3.4. *Bayes* Sınıflandırıcılarda Sıfır Değer Sorunu 163
6.3.5. Sayısal Nitelik Değerleri 164
6.3.6. Uygulama 2 164
6.4. Özet 182
6.5. Sorular 183

Bölüm 7. **DESTEK VEKTÖR MAKİNESİ İLE SINIFLANDIRMA** **167**

7.1. Doğrusal Olarak Ayrılabilme Durumu 167
7.1.1. Primal Çözüm 172
7.1.2. Lagrange Çarpanları 173
7.1.3. *Karush-Kuhn-Tucker* Koşulları 173
7.1.4. Dual Çözüm 174
7.2. Verilerin Doğrusal Olarak Ayrılamama Durumu 178
7.3. Doğrusal Olmayan Sınıflandırıcılar 182
7.3.1. Doğrusal Olmayan Özellik Uzayı 182

7.3.2. Çekirdek Fonksiyonlar 184
7.3.3. Destek Vektör Makinası ve Çekirdek Fonksiyonlar 185
7.4. Özet 189
7.5. Sorular 190

Bölüm 8. **KÜMELEME** **191**

8.1. Kümeleme Çözümlemesi 191
8.2. Uzaklık Ölçüleri 132
8.3. Hiyerarşik Kümeleme 196
8.3.1. Birleştirici Hiyerarşik Yöntemler 196
8.3.2. En Yakın Komşu Algoritması 196
8.3.2.1. Uygulama 198
8.3.3. En Uzak Komşu Algoritması 203
8.3.3.1. Uygulama 204
8.4. Hiyerarşik Olmayan Kümeleme 209
8.4.1. k-Ortalamalar Yöntemi 209
8.4.1.1. Uygulama 210
8.5. Özet 215
8.6. Sorular 216

Bölüm 9. **BİRLİKTELİK KURALLARI** **217**

9.1. Destek ve Güven Ölçütleri 217
9.2. *Apriori* Algoritması 219
9.3. Uygulama 219
9.4. Özet 224
9.5. Sorular 224

Kaynakça **227**

Dizin **233**

Önsöz

Günümüzde kurumlar büyük boyutlarda veri üretmekte ve bu veri içinde anlamlı ve yararlı bilgiyi ortaya çıkarmakta zorluklar yaşamaktadır. Geleneksel istatistik yöntemlerle büyük boyuttaki veriyi çözümlemek kolay değildir. Bu nedenle verileri işlemek ve çözümlemek için özel yöntemlere gereksinim duyulmuştur. Veri madenciliği yöntemleri bu gereksinimi karşılamak üzere ortaya çıkmıştır.

Elinizdeki bu kitap veri madenciliği yöntemlerini ele alan bir giriş kitabıdır. Sınıflama, kümeleme ve birliktelik kuralları ile ilgili birçok algoritmaya en basit biçimiyle yer vermektedir.

İşletmelerin kendi ürettiği veri yanısıra dış çevreden elde ettiği veriyi de saklayarak karar destek sistemlerini hazırlamaları söz konusu olmaktadır. Üretilen veriyi veri ambarı biçiminde tasarlayarak, karar destek sistemleri için bir alt yapı hazırlamak gerekebilecektir. Kitabımızın birinci bölümünde veri ambarı kavramını ana hatlarıyla ele alarak inceliyoruz. İkinci bölümde ise veri madenciliğinin ne olduğu ve nerelerde kullanıldığı konusu işlenmektedir.

Verinin bazı ortak özelliklerinden yararlanılarak sınıflandırılması yoluna gidilebilir. Sınıflandırma işlemi belirli bir süreç izlenerek yerine getirilir. Öncelikle bir eğitim kümesi ele alınarak onun üzerinde karar kuralları oluşturulur ve bu karar kuralları kullanılarak yeni durumlar yorumlanır. Veri madenciliğinin çok sayıda sınıflandırma algoritması bulunmaktadır. Üçüncü bölümde söz konusu yöntemlerden *ID3* ve *C4.5* algoritmaları ele alınarak incelenmektedir. Söz konusu algoritmalar, karar ağacında bilginin ölçülmesi ve entropiye dayalı olarak sınıflandırmanın başlatılması esasına dayanmaktadır.

Sınıflandırma işlemlerinde kullanılan bir diğer yöntem, "Sınıflandırma ve Karar Ağaçları" adıyla bilinmektedir. Bu yöntemde, karar ağacının ikili dallanmasına izin verilmektedir. Kitabın dördüncü bölümünde *Twoing, Gini* ve *regresyon ağaçları* algoritmaları ele alınarak incelenmektedir. Beşinci bölümde ise, veriyi sınıflandırmak *en yakın k-komşu algoritması* anlatılmaktadır. Bu yöntem, verilen bir noktaya en yakın k komşunun belirlenmesi ve yeni gözlemin buna göre değerlendirilmesi esasına dayanmaktadır.

Altıncı bölümde *Bayes* sınıflandırma modeli ele alınarak sınıflandırma işlemi farklı bir açıdan incelenmektedir. Bu bölümde ayrıca "Bayes ağları" ele alınmaktadır. Yedinci bölümde ise yine sınıflandırma işleminde kullanılmak üzere "Destek Vektör Makineleri" ele alınmakta ve matematiksel özellikleri üzerinde durulmaktadır.

Sekizinci bölümde kümeleme konusu işlenmektedir. Verinin birbirine benzeyen kısımlarının gruplandırılmasına "kümeleme" adını veriyoruz. Kümeleme çözümlemeleri veri madenciliğinde geniş bir uygulama alanı bulmuştur. Bu ders kapsamında verilerin birbirine olan uzaklıklarını esas alarak *hiyerarşik ve hiyerarşik olmayan* kümeleme biçiminde geliştirilmiş iki algoritma türü anlatılmaktadır.

Bir arada gerçekleşen olayları ele alarak çözümleyen veri madenciliği yöntemlerine "birliktelik kuralları" adı verilmektedir. Bu yöntem özellikle pazarlama alanında "sepet çözümlememeleri" adıyla bilinmektedir. Bir satış mağazasında müşterilerin aynı anda satın aldıkları ürün sepeti birliktelik kuralları algoritmaları ile çözümlenerek müşterilerin harcama eğilimleri ortaya konulmaktadır. Dokuzuncu bölümde birliktelik kuralları ele alınarak incelenmektedir.

Yalçın ÖZKAN