

Biyoenformatik DNA Mikrodizi

Veri Madenciliđi



Biyoenformatik DNA Mikrodizi

Veri Madenciliđi

Dr. Yalçın ÖZKAN
Dr. Çiğdem SELÇUKCAN EROL

Papatya Bilim
İstanbul, Ankara, İzmir, Adana

© PAPATYA YAYINCILIK EĞİTİM
BİLGİSAYAR SİS. SAN. VE TİC. A.Ş.

İnönü Cad. Hacıhanım Sok. 10/6, 80090, Gümüşsuyu/İstanbul

Tel : (+90 212) 245 37 40 (+90 532) 311 31 10
Faks : (+90 212) 245 37 41
e-mail : admin@papatyabilim.com.tr
Web : www.papatyabilim.com.tr
Dağıtım : TDK Bilim

Biyoenformatik DNA Mikrodizi Veri Madenciligi – Yalçın ÖZKAN ve Çiğdem SELÇUKCAN EROL

1. Basım Haziran 2015
2. Basım Ocak 2017

Yayın Danışmanı : Dr. Toros Rifat ÇÖLKESEN (Post-Edu Institute)
Üretim : Necdet AVCI
Pazarlama : Mustafa DEMİR
Satış : Türkiye'nin İnternet Kitapçısı www.tdk.com.tr
Sayfa Düzenleme : Müge URCAN, Papatya ve 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 AŞ'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.

Yalçın, Özkan; Selçukcan Erol, Çiğdem.

Biyoenformatik DNA Mikrodizi Veri Madenciligi / Yalçın Özkan ve Çiğdem Selçukcan Erol - İstanbul: Papatya Yayıncılık Eğitim, 2015.

xviii, 432 s.; 24 cm.

Kaynakça ve Dizin var.

Sertifika No: 11218

ISBN 978-605-4220-89-2.

1. Gen Seçimi. 2. Sınıflandırma Algoritmaları. 3. Kümeleme Algoritmaları. 4. R Dili.
I. Title

Sevgili torunum Ata'ya..

Yalçın ÖZKAN

Sevgili eşim Osman'a ..

Çiğdem SELÇUKCAN EROL

*Bu kitabın yayınlanması
konusunda desteklerinden ötürü
Sayın Rifat ÇÖLKESEN'e ve
emeęi geçen tüm
Papatya Yayıncılık Eęitim
alıřanlarına ok teřekkür ediyoruz.*

İçindekiler

Önsöz

xvii

Bölüm 1: R DİLİNE GİRİŞ	19
1.1 Bu Bölümün Amacı	19
1.2 R'nin Kurulumu	19
1.3 R Arayüzü	19
1.4 Komut Satırı	21
1.5 Nesneler	22
1.6 R Veri Nesneleri	23
1.6.1 Skaler	23
1.6.2 Vektör	24
1.6.3 Faktör	26
1.6.4 Dizilim	27
1.6.5 Matris ve Dizi	27
1.6.6 Liste	31
1.6.7 Veri Çerçevesi	32
1.7 İşleçler	34
1.7.1 Aritmetik İşleçler	34
1.7.2 Mantıksal ve İlişkisel İşleçler	35
1.8 Döngüler	37
1.9 Fonksiyonlar	38
1.9.1 Hazır Fonksiyonlar	38
1.9.2 Kullanıcı Fonksiyonları	38
1.10 R'de İstatistiksel Veri İşleme	40
1.11 Grafikler	42
1.12 R'de Çizim İşlemleri	44
1.13 Dizin ve Dosya İşlemleri	45
1.13.1 Dizin İşlemleri	45
1.13.2 Dosyanın Okutulması	45
1.14 R Paketleri	47
1.15 Bioconductor	47
1.15.1 Bioconductor Yazılımlarının Yüklenmesi	47
1.15.2 Bioconductor Paketlerini Güncelleme ve Derleme	48
1.16 Özet	48
1.17 Sorular	49
1.18 Kaynaklar	50

Bölüm 2: TEMEL BİYOLOJİ VE MİKRODİZİ	51
2.1 Bu Bölümün Amacı	51
2.2 Temel Biyoloji Bilgileri	51
2.2.1 Genetik Materyal	53
2.2.2 Nükleik Asitler	53
2.3 DNA	55
2.4 RNA58	
2.5 Protein	60
2.6 Kromozom	64
2.7 Gen İfadesi	68
2.8 Mutasyon	68
2.9 Önemli Terimler	69
2.10 Dizileme Teknolojileri	70
2.11 Mikrodizi	70
2.12 Mikrodizi Veri Kümeleri	76
2.13 Özet	77
2.14 Sorular	78
2.15 Kaynaklar	78
 Bölüm 3: MİKRODİZİ VERİ YAPILARI	 81
3.1 Bu Bölümün Amacı	81
3.2 Mikrodizi Veri Dosyaları	81
3.2.1 Mikrodizi Veri Kümelerinin Yüklenmesi	82
3.2.2 Bioconductor Veri Kümelerinin Yüklenmesi	82
3.3 Affymetrix Veri	83
3.3.1 PM ve MM Problemleri	84
3.3.2 Affymetrix CEL Veri Dosyasının Görünümü	84
3.3.3 Bioconductor Affymetrix Paketleri	85
3.3.4 Bioconductor Affymetrix Dosyalarının Okunması	86
3.3.5 Bioconductor Dışındaki Affymetrix Veri Kümeleri	86
3.3.6 Dilution Veri Kümesi Bilgilerini Görüntüleme	87
3.3.7 Gen İfadeleri	87
3.3.8 Veri Kümesindeki Örnekler	88
3.3.9 PM ve MM Yoğunlukları	88
3.3.10 Gen İsimleri	89
3.3.11 Yoğunluk Grafiği	89
3.3.12 Prob Kümesi Dağılım Grafiği	90
3.3.13 RNA Degradasyon Grafiği	91
3.3.14 Görüntü	92
3.3.15 Kutu Grafiği	93
3.3.16 Histogram	93

3.3.17	MA Grafikleri	94
3.4	Agilent Veri	95
3.4.1	Paketin Tanıtılması	96
3.4.2	Swirl Veri Kümesi	96
3.4.3	Veri Kümesinin Okutulması	97
3.4.4	Gen İsimleri	97
3.4.5	Görüntünün Elde Edilmesi	98
3.4.6	Kutu Grafiği	98
3.4.7	Dağılma Grafiği	99
3.4.8	Tek Renk Agilent Veriler	100
3.5	Illumina Veri	101
3.5.1	Illumina Veri Kümesinin Yüklmesi	101
3.5.2	Örnek Illumina Veri Kümesi	101
3.5.3	Phenodata	102
3.5.4	Yoğunluk Grafiği	102
3.5.5	Örnek Korelasyonları	103
3.5.6	Kutu Grafiği	104
3.5.7	Birikimli Dağılım Fonksiyonu (CDF)	105
3.5.8	MA Grafiği	106
3.5.9	Varyans Katsayısı Grafiği	107
3.6	Özet	108
3.7	Sorular	109
3.8	Kaynaklar	109
Bölüm 4:	MİKRODİZİLERDE ÖN İŞLEME	111
4.1	Bu Bölümün Amacı	111
4.2	Ön İşleme Süreci	112
4.3	Arka Plan Düzeltme	112
4.3.1	MAS5 Yöntemi ile Arka Plan Düzeltme	113
	Örnek: MAS5 Arka Plan Düzeltme Uygulaması	113
4.3.2	RMA Yöntemi ile Arka Plan Düzeltme	114
	Örnek: RMA Yöntemi Uygulaması	115
4.4	İki Renk Dizilerde Arka Plan Düzeltme	116
	Örnek: Bioconductor ile İki Renk Dizilerde Arka Plan Düzeltmeler	117
4.5	Normalizasyon	117
4.5.1	Tek Renk Mikrodizi Verilerinde Normalizasyon	118
4.5.2	Kantil Normalizasyon	118
	Örnek: Kantil Normalizasyon Uygulaması	118
	Örnek: Kantil Normalizasyon için Bioconductor Uygulaması	121
4.5.3	LOWESS Normalizasyon	125
	Lowess Normalizasyon Yönteminin Esasları	126
	Güçlü Lowess Normalizasyon Yöntemi	127

Örnek: Lowess Yöntemiyle Normalizasyon İşlemi	129
Örnek: Lowess Normalizasyon Yöntemi için Bioconductor Uygulaması	134
4.6 İki Renk Mikrodizi Veri İçin Normalizasyon	136
4.6.1 Dizi İç Normalizasyon	136
Örnek: Dizi iç Normalizasyon için Bioconductor Uygulaması	136
4.6.2 Diziler Arası Normalizasyon	137
Ölçek Normalizasyonu	137
Örnek: Ölçek Normalizasyonu Uygulaması	138
Örnek: Bioconductor ile Ölçek Normalizasyonu Uygulaması	140
4.7 PM Düzeltme	143
4.7.1 Farkların Hesaplanması Yöntemi	143
4.7.2 İdeal Uyumsuz Değerleri Hesaplama Yöntemi	143
Tukey Bi-Weight Ortalaması	143
Örnek: Tukey Bi-Weight Uygulaması	144
MAS5 Algoritması	146
Örnek: MAS5 Algoritması ile İlgili Bioconductor Uygulaması	147
4.8 Özetleme	149
4.8.1 Tukey Bi-Weight Yöntemi ile Özetleme	149
4.8.2 Bioconductor ile Tukey Bi-Weight Yöntemi Uygulaması	150
4.8.3 Medyan Parlatma Yöntemi ile Özetleme	150
Örnek: Medyan Parlatma Yöntemi ile Özetleme Uygulaması	151
Örnek: Bioconductor ile Medyan Parlatma Yönteminin Uygulanması	153
4.9 Bioconductor ile Affymetrix Mikrodizi Ön İşlemleri İçin Pratik Bir Yol	153
4.10 Özet	154
4.11 Sorular	154
4.12 Kaynaklar	155

Bölüm 5: VERİ MADENCİLİĞİNDE ÖNİŞLEME VE GEN SEÇİM YÖNTEMLERİ

5.1 Bu Bölümün Amacı	157
5.2 Mikro Dizi Veri Kümeleri	157
5.2.1 Bioconductor Leukemia Veri Kümesi	157
Veri Kümesi Nitelikleri	158
R ile Leukemia Veri Kümesinin Kullanımı	158
Bioconductor Leukemia Veri Kümesinin CMA Sürümü	160
5.2.2 Bioconductor Acute Lymphoblastic Leukemia Veri Kümesi	161
Veri Kümesi Nitelikleri	161
R ile Acute Lymphoblastic Leukemia Veri Kümesinin Kullanımı	163
5.2.3 Küçük Yuvarlak Mavi Hücreli Tümör Veri Kümesi	164
R ile Küçük Yuvarlak Mavi Hücreli Tümör Veri Kümesinin Kullanımı	165
5.3 Veri Madenciliğinde Kullanılan Uzaklık Ölçüleri	165

5.3.1	Öklid Uzaklığı	166
5.3.2	Manhattan Uzaklığı	168
5.3.3	Minkowski Uzaklığı	168
5.3.4	Hamming Uzaklığı	169
5.4	Entropi Kavramı	170
5.5	Mikrodizi Verilerde Ayırıklaştırma	172
5.5.1	ChiMerge Ayırıklaştırma Yöntemi	172
	Örnek: ChiMerge Yöntemiyle Ayırıklaştırma	173
	Örnek: R ile ChiMerge Uygulaması	178
	Örnek: R ile Leukemia Veri Kümesi Üzerinde ChiMerge Uygulaması	179
5.5.2	Sınıf-Nitelik Bağımlılık Maksimizasyonu-CAIM Algoritması	181
	Örnek: CAIM Yöntemiyle Ayırıklaştırma	183
	Örnek: R ile CAIM Algoritmasının Uygulanması	188
	Örnek: R ile Leukemia Veri Kümesi CAIM Algoritmasının Uygulaması	189
5.5.3	Sınıf-Nitelik Kontenjans Katsayısı -CACC Algoritması	190
	Örnek: CACC Yöntemiyle Ayırıklaştırma	191
	Örnek: R ile CACC Algoritmasının Uygulanması	194
	Örnek: R ile Leukemia Veri Kümesi İçin CACC Algoritmasının Uygulaması	196
5.6	Mikrodizilerde Gen Seçim Yöntemleri	198
5.6.1	Entropiye Dayalı Filtreler	198
	Bilgi Kazancı	198
	Kazanç Oranı	199
	Simetrik Belirsizlik	199
	Örnek: Bilgi Kazancı Filtresinin Uygulanması	199
	Örnek: R ile Leukemia Veri Kümesi için Entropi Tabanlı Nitelik Seçimi	202
5.6.2	Korelasyon Tabanlı Nitelik Seçimi	206
	Örnek: Korelasyon Tabanlı Nitelik Seçimi	207
	R ile Korelasyon Tabanlı Nitelik Seçimi	210
	Örnek: R ile Leukemia Veri Kümesi için Korelasyon Tabanlı Nitelik Seçimi	210
5.6.3	Ki-Kare Filtresi	211
	Örnek: Ki-Kare Filtresiyle Nitelik Seçimi	212
	Örnek: R ile Leukemia Veri Kümesi İçin Ki-Kare Filtresi ile Nitelik Seçimi	215
5.6.4	Pearson Korelasyon Filtresi	216
	Örnek: Pearson Korelasyon Katsayısı filtresi ile Nitelik Seçimi	216
	Örnek: R ile Pearson Korelasyon Filtresi	218
	Örnek: R ile Leukemia Veri Kümesi için Pearson Korelasyon Filtresi ile Nitelik Seçimi	219

5.6.5	Spearman Sıra Korelasyon Filtresi	220
	Örnek: Spearman Sıra Korelasyon Filtresi ile Nitelik Seçimi	220
	Örnek: R ile Spearman Sıra Korelasyon Filtresi Uygulaması	222
	Örnek: R ile Leukemia Veri Kümesi için Spearman Sıra Korelasyon Filtresi ile Nitelik Seçimi	222
5.7	Özet	223
5.8	Sorular	224
5.9	Kaynakça	224
Bölüm 6: SINIFLANDIRMA		227
6.1	Bu Bölümün Amacı	227
6.1.1	Mikrodizi Verilerinde Sınıflandırma	227
6.1.2	Sınıflandırma Süreci	228
6.1.3	Öğrenme ve Test Veri Kümelerinin Elde Edilmesi	228
	Örnek: R ile Leukemia Veri Kümesi İçin Öğrenme Kümelerinin Oluşturulması	229
	Örnek: R ile Leukemia Veri Kümesine Bootstrap Yönteminin Uygulanması	230
6.2	Sınıflandırma Ağaçları	231
6.2.1	Mikrodizi Sınıflandırma Ağaçlarında Öğrenme Süreci	232
6.2.2	Sınıflandırma Kurallarının Elde Edilmesi	232
6.2.3	Kuralların Uygulanması ve Tahmin	234
6.2.4	C4.5 Algoritması	235
	Sınıflandırma Ağacında Entropi	235
	Sürekli Değerlere Sahip Nitelik Değerlerinin Kategorik Değerlere Dönüştürülmesi	236
	Dallanma İçin Niteliklerin Seçilmesi ve Bilgi Kazancı	236
	Örnek: C4.5 Algoritmasının Uygulanması	237
	Örnek: R ile C4.5 Algoritmasının Uygulanması	249
	Örnek: Küçük Yuvarlak Hücreli Tümör Veri Kümesi Üzerinde C4.5 ile Sınıflandırma	252
6.3	Sınıflandırma ve Regresyon Ağaçları	254
6.3.1	Sürekli Değerler İçin Gini Algoritması	254
	Örnek: Gini Algoritması ile Sınıflandırma	256
	Örnek: R ile Gini Algoritmasının Uygulanması	261
	Örnek: R ile Leukemia Veri Kümesi Üzerinde Gini Algoritması ile Sınıflandırma	263
6.3.2	Regresyon Ağaçları	265
	Örnek: Regresyon Ağaçlarının Elde Edilmesi	266
	Örnek: R ile Regresyon Ağacı Algoritmasının Uygulanması	271
	Örnek: R ile Leukemia Veri Kümesi Üzerinde Regresyon Ağaçları Algoritması ile Sınıflandırma	272

Örnek: Küçük Yuvarlak Hücreli Tümör Veri Kümesi Üzerinde Regresyon Ağaçları Algoritması İle Sınıflandırma	274
6.3.3 Rastgele Orman Algoritması	275
Örnek: R ile Leukemia Veri Kümesi Üzerinde Rastgele Orman Algoritması ile Sınıflandırma	277
Örnek: Küçük Yuvarlak Hücreli Tümör Veri Kümesi Üzerinde Rastgele Orman Algoritması İle Sınıflandırma	280
Örnek: Rastgele Orman Algoritması ile Gen Seçimi	282
6.4 En Yakın K-Komşu Algoritması ile Sınıflandırma	283
6.4.1 Öklid Uzaklıklarının Kullanımı	283
6.4.2 Ağırlıklı Oylama	284
6.4.3 En Yakın K-Komşu Algoritması	284
Örnek: En Yakın k-Komşu Algoritması ile Sınıflandırma	285
Örnek: R ile En Yakın k-komşu Algoritmasının Uygulanması	287
Örnek: R ile Leukemia Veri Kümesi Üzerinde En Yakın k-Komşu Algoritması ile Sınıflandırma ve Kestirim	289
6.5 Bayes Sınıflandırıcılar	290
6.5.1 Koşullu Olasılık	290
6.5.2 Bayes Bağıntısı	291
6.5.3 Sade Bayes Sınıflandırıcısı ve MAP	291
6.5.4 Sürekli Nitelik Değerleri	292
6.5.5 Sıfır Değer Sorunu	293
Örnek: Bayes Sınıflandırma	293
Örnek: e1071 Paketi ile Sade Bayes Sınıflandırması	299
Örnek: CMA Paketi ile Bayes Sınıflandırma	300
Örnek: Küçük Yuvarlak Hücreli Tümör Veri Kümesi Üzerinde Bayes Algoritması ile Sınıflandırma ve Kestirim	302
6.6 Destek Vektör Makinesi ile Sınıflandırma	303
6.6.1 Doğrusal Olarak Ayrılabilme Durumu	303
6.6.2 Primal Çözüm	307
6.6.3 Lagrange Çarpanları	307
6.6.4 Karush-Kuhn-Tucker Koşulları	308
6.6.5 Dual Çözüm	309
Örnek: Destek Vektör Makinesi ile Sınıflandırma	310
Örnek: R ile Destek Vektör Makinesi Uygulaması	313
Örnek: Küçük Yuvarlak Hücreli Tümör Veri Kümesi Üzerinde Destek Vektör Makinesi ile Sınıflandırma ve Kestirim	314
6.7 Özet	315
6.8 Sorular	316
6.9 Kaynaklar	317

Bölüm 7: KÜMELEME	319
7.1 Bu Bölümün Amacı	319
7.2 Kümeleme Çözümlemesi	319
7.3 Mikrodizi Deneylerinde Kümeleme	320
7.4 Hiyerarşik Kümeleme Yöntemleri	321
7.4.1 Birleştirici Kümeleme: AGNES	321
En Yakın Komşu Algoritması	321
En Uzak Komşu Algoritması	322
Örnek: En Yakın Komşu Algoritması ile Kümeleme	322
Örnek: R ile En Yakın Komşu ve En Uzak Komşu Algoritmasının Uygulaması	328
Örnek: R ile Bioconductor Leukemia Veri Kümesi İçin En Yakın Komşu Algoritması ile Kümeleme	331
7.4.2 Ortalama Link Yöntemi	333
Örnek: Ortalama Link Yöntemiyle Kümeleme	333
Örnek: R ile Ortalama Link Yönteminin Uygulanması	342
Örnek: R ile Bioconductor Leukemia Veri Kümesi İçin Ortalama Link Yöntemiyle Kümeleme	343
7.4.3 Ward Yöntemi	345
Örnek: Ward Yöntemini Kullanarak Kümeleme İşlemi	345
Örnek: R ile Ward Yönteminin Uygulanması	350
Örnek: R ile Bioconductor Leukemia Veri Kümesi İçin Ward Yöntemiyle Kümeleme	351
Örnek: R ile Bioconductor Leukemia Veri Kümesi İçin Cluster Paketini Kullanarak Kümeleme	353
7.4.4 Küme Merkezi Yöntemi	354
Örnek: Küme Merkezi Yöntemi ile Kümeleme	354
Örnek: R ile Küme Merkezi Yönteminin Uygulanması	357
Örnek: Bioconductor Leukemia Veri Kümesi için Küme Merkezi Yöntemiyle Kümeleme	358
7.4.5 Ayırıcı Kümeleme: DIANA	359
Örnek: DIANA Yöntemiyle Kümeleme	360
Örnek: R ile DIANA Kümeleme Yönteminin Uygulaması	365
Örnek: Bioconductor Leukemia Veri Kümesi İçin DIANA Yöntemiyle Kümeleme	368
7.5 Hiyerarşik Olmayan Kümeleme	369
7.5.1 K-Ortalamalar Yöntemi	369
Örnek: K-Ortalamalar Yöntemi ile Kümeleme	370
Örnek: K-Ortalamalar Yönteminin R Uygulaması	374
Örnek: Bioconductor Leukemia Veri Kümesi İçin K-Ortalamalar Yöntemiyle Kümeleme	377
7.5.2 PAM Kümeleme Algoritması	380

Örnek: PAM Algoritmasının Uygulanması	381
Örnek: R ile PAM Algoritmasının Uygulanması	384
Örnek: Bioconductor Leukemia Veri Kümesi İçin PAM	
Yöntemiyle Kümeleme	386
7.5.3 Büyük Veri Kümeleri ve CLARA	390
Örnek: Bioconductor Leukemia Veri Kümesi için CLARA	
Yöntemiyle Kümeleme	390
7.7 Özet	393
7.8 Sorular	393
7.8 Kaynaklar	393
Bölüm 8: BİRLİKTELİK KURALLARI	395
8.1 Bu Bölümün Amacı	395
8.2 Mikrodizi Verilerde Birliktelik Kuralları	395
8.3 Destek ve Güven Ölçütleri	396
8.4 Apriori Algoritması	396
8.4.1 Apriori Algoritması ile Birliktelik Kuralları	397
8.4.2 R ile Apriori Algoritması Uygulaması	403
8.4.3 R Ortamında Apriori Algoritması ile Leukemia Veri Kümesi için	
Birliktelik Kurallarının Elde Edilmesi	407
8.5 Eclat Algoritması	410
8.5.1 Eclat Algoritması ile Birliktelik Kuralları	411
8.5.2 R Ortamında Eclat Algoritması ile Birlikteliklerin Elde	
Edilmesi	415
8.5.3 R Ortamında Eclat Algoritması ile Leukemia Veri Kümesi için	
Birlikteliklerin Elde Edilmesi	419
8.6 Özet	420
8.7 Sorular	420
8.8 Kaynaklar	420
Dizin	425

Önsöz

DNA mikrodizi teknolojisi, genlerin bilinen ve bilinmeyen fonksiyonlarının tespit edilmesinde kullanılan çok yönlü bir teknoloji olarak karşımıza çıkmaktadır. Mikrodiziler, binlerce farklı DNA parçalarının sentez edildiği ya da yerleştirildiği binlerce noktayı içeren çipler olarak da değerlendirilebilir. Mikrodizi, genellikle sağlıklı ve hastalıklı dokulardaki gen farklılıklarını ortaya çıkarmak üzere tüm genlerin eşzamanlı ifadelerini belirlemek için kullanılmaktadır. Çok sayıda veriden oluşan bu gen ifadelerinin analiz edilmesi ve bu analizlere dayalı olarak yorumlanması önemli bir konu olarak karşımıza çıkmaktadır.

Bu kitap mikrodizi verisi üzerinde temel veri madenciliği analiz yöntemlerini uygulayarak, sınıflandırma ve kümeleme işlemlerinin nasıl yapılacağı ve birliktelik kurallarının nasıl elde edilebileceği konusunu ortaya koymak amacıyla hazırlanmıştır.

Mikrodizi analiz yöntemlerinin uygulanmasında R programlama dili yaygın biçimde kullanılmaktadır. Bu konuda araştırmacıların hazırladığı çok sayıda mikrodizi R paketi bulunmaktadır. Kitabımızda ele alınan veri madenciliği yöntemleri uygulamalarında R paketlerinden yararlanılmıştır. R diline alışkın olmayan okuyucular için kitabın birinci bölümü hazırlanmıştır. Bu bölümdeki bilgiler ele alınan uygulamaların anlaşılması açısından yeterli düzeydedir.

Kitabın ikinci bölümünde biyoloji konusunda eğitim almayan okuyucular için bazı temel bilgilere yer verilmiştir. Hücre bilgisi, DNA, RNA, protein ve kromozomlara ilişkin bilgilerin yanısıra bu kitapta yoğun biçimde kullanılan *gen ifadesi* kavramına açıklık kazandırılmaktadır. Bu bölümde mikrodizi teknolojilerinin neler olduğu hususuna değinilmektedir.

Üçüncü bölümde, mikrodizi veri yapıları üzerinde durularak yaygın biçimde kullanılan Affymetrix, Agilent ve Illumina platformlarında üretilen mikrodizi dosyaların, R ortamına nasıl yüklenebileceği ve kullanılabileceği konusu işlenmektedir.

Mikrodizi verilerinin analize tabi tutulabilmesi için bazı ön işlemlerden geçirilmesi gerekmektedir. Veri üzerindeki bozulmaların düzeltilmesi için arka plan düzeltme işlemleri yapılır. Bu işlemin ardından verinin normalize edilmesi söz konusudur. Affymetrix veri kümelerinde PM düzeltme aşaması gerçekleştirilir. Son olarak veri özetlenerek veri analizlerinde kullanılabilecek gen ifadeleri elde edilmektedir. Sözü edilen bu aşamalar dördüncü bölümde ayrıntılı biçimde incelenmektedir.

Beşinci bölümden itibaren veri madenciliği yöntemleri ele alınmaktadır. Veri madenciliği sürecinde verinin analizi için uygun hale getirilmesi gerekmektedir. Bu amaçla ön işleme yöntemleri uygulanır. Beşinci bölümde veri madenciliği ön işleme

yöntemleri arasında yer alan *ayrıklaştırma* konusu ayrıntılı biçimde incelenmektedir. Bunun dışında nitelik seçme konusu *gen seçimi* kavramı içinde ele alınarak analiz edilmektedir.

Kitap içinde yer alan yöntemleri ortaya koymak amacıyla iki tür uygulama yapılmaktadır. Birinci tür uygulamalarda ele alınan bir algoritmayı açıklamaya yönelik az sayıda veriden oluşan bir veri kümesi ele alınmıştır. Sözü edilen veri kümeleri tarafımızdan oluşturulmuş, gerçek olmayan bir tür sanal veri kümeleridir. İkinci tür veri kümeleri ise R ortamında yaygın biçimde kullanılan gerçek veri kümeleridir. Bu veri kümeleri beşinci bölümde tanıtılmaktadır. Veri kümelerini Bioconductor ortamından seçmeye özen gösterdik.

Altıncı bölümde veri madenciliğinin *sınıflandırma* ile ilgili konusu ele alınarak çeşitli sınıflandırma algoritmalarına yer verilmiştir. Sınıflandırma ağaçlarının oluşturulması, veri madenciliğinde önemli bir konu olarak karşımıza çıkmaktadır. Entropi tabanlı sınıflandırma ağaçları arasında yer alan C4.5 algoritması bu bölümde yer almaktadır. Bu yöntemin dışında sınıflandırma ağaçları kapsamında Gini algoritması, Regresyon ağaçları ve Rastgele Orman algoritmasına da yer verilmiştir. Sözü edilen yöntemler dışında En yakın k-komşu algoritması, Bayes sınıflandırıcılar ve Destek Vektör Makineleri ile sınıflandırma ayrıntılı biçimde incelenmiştir.

Mikrodizi verisini gen anlatım düzeylerine göre gruplandırma veya bir başka deyişle *kümeleme* yaygın biçimde uygulanmaktadır. Kitabın yedinci bölümü mikrodizi verisine, veri madenciliğinin kümeleme algoritmalarının nasıl uygulanabileceği konusuna ayrılmıştır. Bu kapsamda hiyerarşik olan ve hiyerarşik olmayan kümeleme yöntemleri işlenmiştir. Hiyerarşik kümeleme yöntemleri arasında AGNES, DIANA, Ortalama Link, Ward ve Küme Merkezi; hiyerarşik olmayan yöntemler arasında K-ortalamalar, PAM ve CLARA algoritması ayrıntılı biçimde incelenmiştir.

Kitabın son bölümünü Birliktelik Kuralları konusu oluşturmaktadır. Olayların birlikte gerçekleşme durumlarını çözümleyen veri madenciliği yöntemlerine *birliktelik kuralları* adını veriyoruz. Genlerin ifade düzeyleri göz önüne alındığında, birlikte hareket eden ve etmeyen genlerin belirlenmesi birliktelik kuralları algoritmaları ile sağlanabilir. Bu bölümde birliktelik kuralları konusu Apriori ve Eclat algoritmaları yardımıyla açıklanmaktadır.

Kitap içinde gözden kaçmış hatalar olabilir. Bu konuda okuyucularımızın eleştirilerini ve uyarılarını bekliyoruz.

Yalçın ÖZKAN

Çiğdem SELÇUKCAN EROL