Biyoinformatik Veri Analizinde

R ile Hiyerarşik Kümeleme

Prof. Dr.
Zeynel CEBECİ

PAPATYA

EĞİTİM BİLGİSAYAR SİS. SAN. VE TİC. A.Ş.

Ankara Cad. Prof. F. Kerim Gökay Vakfı İşhanı Girişi
No: 11/6 Cağaloğlu/İstanbul

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

Biyoinformatik Veri Analizinde R ile Hiyerarşik Kümeleme – Prof. Dr. Zeynel CEBECİ

1. Basım Haziran 2019 Deneme basımı
2. Basım Şubat 2021 Gözden geçirilmiş basım

Yayına Hazırlayan : Cengiz UĞURKAYA (Ph. D)
Üretim : Necdet AVCI
Pazarlama Satış : Mustafa DEMİR
Satış : TDK Bilim ~ www.tdk.com.tr
Kapak Resmi : Zeynel CEBECİ
Kapak Tasarım : Papatya Kelebek Görsel Tasarım Atölyesi
Sayfa Düzenleme : Papatya Kelebek Görsel Tasarım Atölyesi
Basım : Özkaracan Matbaacılık (Sertifika No: 12228)
Evren Mah. Gülbahar Cad. No: 62 Güneşli/İstanbul

Cebeci, Zeynel.
Biyoinformatik Veri Analizinde R ile Hiyerarşik Kümeleme / Zeynel Cebeci – İstanbul: Papatya Yayıncılık Eğitim, 2021
xii, 300. ; 24 cm
Kaynakça ve dizin var.
ISBN 978-605-9594-44-8
Sertifika No: 11218
1. Veri Önişleme 2. Geçerlilik Testleri 3. Dendrogram Çizme 4. İleri Algoritmalar
I. Title
QA76.9 .D35 C64 2011

Nalan ve Çağatay'a

İçindekiler

Önsöz

Hiyerarşik kümeleme, veri madenciliğinde örüntü tanıma, makine öğrenmesi, pazarlama ve müşteri yönetimi, gen keşfi, ilaç tasarımı gibi birçok fen, sosyal ve yaşambilimleri alanında en yaygın kullanılan keşifçi ya da açınsal istatistiksel analizdir. Çoğu kez, analiz öncesi belli bir küme sayısı (k parametresi) ve küme merkezleri için başlatma değerleri istememesi gibi önemli avantajlar sunması nedeniyle K-ortalamalar ve K-ortancalar gibi yine yaygın olarak kullanılan hiyerarşik olmayan yöntemlere göre tercih edilmektedir. Analiz için ihtiyaç duyulan yegâne şey veri öğeleri arasındaki benzerlik ölçüsü olduğundan uygulanması kolay bulunmaktadır.

R, istatistik ve grafik analiz için güçlü bir hesaplama ortamı ve programlama dilidir. R projesi kapsamında GNU lisansı ile dağıtılan açık kaynak ve özgür yazılım olması, yeni algoritmaları geliştirmek ve test etmek için etkin ve kolay bir programlama ortamı sunması nedeniyle istatistik, matematik, bilgisayar bilimleri ve biyoinformatik gibi alanlarda hızla popüler olmasını sağlamıştır. Dünya genelinde hemen her kuramsal ve uygulamalı alanda çalışan araştırmacılar, veri analizcileri ve öğrenciler tarafından yaygın şekilde kullanılmaktadır. Günümüzde R için çok çeşitli alanlarda çalışan araştırmacılar tarafından geliştirilmiş 16000'e yaklaşan sayıda R paketi bulunmakta olup CRAN, BioConductor ve Github üzerinden dağıtılmaktadır.

Bu kitap veri madenciliğinde ya da bilgi keşfinde önemli bir açınsal istatistik aracı olarak hiyerarşik kümeleme analizine yöntemler, R ile uygulamalar ve bazı gerçek veri kümeleri üzerinde analiz örnekleriyle kapsamlı bir bakış sağlamaktadır.

Bu kitap, daha önce R ile çalışmamış olanların birkaç gün içinde ileri düzeyde kümeleme analizi yapmalarını sağlayacak bir yaklaşımla yazılmıştır. Kitap, hem veri madenciliği ve istatistik konulu dersler için bir uygulama rehberi hem de biyoteknoloji ve biyoinformatik bilim dallarında çalışan araştırmacılar için bir başvuru eseri olacak şekilde tasarlanmıştır.

Yararlı olması dileğiyle,

Zeynel Cebeci
Adana, 2020

Kitap Hakkında

Kitabımızın ilk bölümünde küme, kümeleme terminolojisi ve kümeleme yöntemleri taksonomisi anlatılmaktadır.

İkinci bölümde birleştirici kümeleme yöntemlerinin teorik temelleri küçük bir veri kümesinde uygulamalı olarak tüm ayrıntıları ile sunulmaktadır.

Üçüncü bölüm ise ayrıcı kümeleme yöntemlerini açıklamakta ve örneklemektedir. Monotetik ve politetik ayırıcı yöntemleri verilmiştir ve ileri algoritmalar ele alınmıştır.

Dördüncü bölümde kümeleme sonuçlarının görselleştirilmesi için kullanılan temel grafik türleri örneklerle açıklanmıştır.

Beşinci bölümde küme sayısının belirlenmesinde kullanılan ölçütler ve teknikler tanıtılmıştır. Çeşitli örnekler verilerek konunun daha kolay anlaşılmasına gayret edilmiştir.

Altıncı bölümde R istatistiksel hesaplama ve grafik analiz yazılımının indirilmesi ve kurulması, R'de çalışma, R dilinin tanıtımı ve R fonksiyonlarına giriş yapılmıştır.

Veri okuma ve önişleme örnekleri ise yedinci ve sekizinci bölümlerde sunulmuştur. Dokuzuncu bölüm ayrıntılı olarak R ile birleştirici kümeleme; onuncu bölüm ise R ile ayırıcı kümeleme analizini açıklamaktadır.

On birinci bölümde kümeleme geçerlilik testleri konusu örneklerle anlatılmıştır.

On ikinci bölümde ise ileri düzeyde görselleştirme araçları ve grafik türleri sunulmaktadır.

On üçüncü bölüm gen ifade profillerinin analizi ve gen keşfi çalışmalarında kullanılan teknik ve yöntemlerin üç gerçek veri kümesine uygulanması ve yorumlanmasını kapsamaktadır. Mikrodizilerde kümeleme analizi anlatılmıştır.

Kısaltmalar

Kısaltma	Terim
AU	Approximately Unbiased
BP	Bootstrap Probability
CA	Clustering Algorithm
CCC	Cubic Clustering Criterion
CPCC	Cophenetic Correlation Coefficient
HAC	Hierarchical Agglomerative Clustering
HDC	Hierarchical Divisive Clustering
MCAR	Missing Completely At Random
MNAR	Missing Not At Random
RMSSTD	Root Mean Square Standard Deviation
VRC	Variance Ratio Criterion
WPGMA	Weighted Pair-Group Method using Arithmetic Averages
WPGMC	Weighted Pair-Group Method using the Centroid Average
UPGMA	Unweighted Pair-Group Average

Kitaptaki Örneklerin Uygulanması

Kitapta herhangi bir konu ile ilgili örnekler aşağıda görüldüğü gibi bir kutu içinde Courier yazı tipinde verilmiştir.

```
> xdianak <- diana(x, metric = "euclidean",
+    stand = F)
```

Komut örneklerinde:

a) Satır başlarındaki > simgesi R tarafından otomatik olarak görüntülenen komut satırı göstergesidir. Örnekler çalıştırılırken yazılmaması gerekir.

b) Satır başlarındaki + simgesi R tarafından otomatik olarak görüntülenen komut satırı devam simgesidir. Girilen komutlar tek bir satıra sığmadığında R izleyen satıra geçer ve satırın başına önceki satırdaki komutun devamı olduğunu gösteren + işareti koyar. Kitaptaki örnekler çalıştırılırken yazılmaması gerekir.

c) Komutlar çalıştırıldıktan sonra sonuçları görüntülenir. Bazı sonuçlar [1], [7] gibi köşeli parantezler içinde sayılarla başlayan şekilde gösterilir. Bunlar sonuçlar bir vektör olduğunda satır başındaki elemanın indis numarasıdır. Sadece bilgi amaçlıdır. Dikkate alınmaması gerekir.

d) Kitapta gerek bazı komutlar ve gerekse bazı sonuçları vurgulamak amacıyla koyu/kalın yazılmıştır. Komut çalıştırırken dikkate alınmamalıdır.

e) Kitabın elektronik sürümünde (e-kitap) R komut örnekleri önce fare ile işaretlenip blok halinde kopyalanır. Daha sonra R çalışma ortamında sağ fare düğmesi tıklanarak *Sadece komutları yapıştır* seçeneği çalıştırılır. Bu işlem kopyalanan komut blokunda bulunan > ve + simgelerini ve diğer metinleri otomatik olarak temizleyerek sadece R komutlarının çalıştırılmasını sağlar.

f) Kitabın elektronik sürümünde (e-kitap) R komut örnekleri > ve + simgesi olmadan kopyalanıp çalıştırılmak istenirse R çalışma ortamında sağ fare düğmesi tıklanarak *Yapıştır* seçeneği seçilmelidir. Bu işlem kopyalanan komut blokundaki kodların hemen çalıştırılmasını sağlar.

Bölüm 1

Kümeleme Analizine Giriş

Küme ve Kümeleme

Kümeleme bireyler, nesneler, olaylar ve durumlar gibi soyut ve somut veri öğelerini benzerliklerine göre gruplara ayırmak işlemidir. Bir başka deyişle öğeleri özelliklerine göre gruplandırma işidir. Günlük yaşamda kümelemenin çok sayıda örneğine rastlamak mümkündür. Örneğin, süpermarketlerdeki tatlılar, baharatlar, süt ve süt ürünleri, et ve et ürünleri, temizlik ürünleri vb. ürünler reyon olarak adlandırılan yüzlerce farklı bölümdeki raflarda sergilenirler. Botanik bahçelerinde bitkiler kayalık bitkileri, güller, tıbbi bitkiler ve ibreliler gibi farklı alanlarda gruplanarak yetiştirilirler. Kütüphaneler veya kitapçılardaki kitaplar konularına göre farklı raflarda dizilirler. Belli bir coğrafya bitki örtüsü veya hayvan biyoçeşitliliği bakımından farklı alt bölgelere veya ekosistemlere bölünürler. Sosyologlar toplumu farklı demografik bölümlere ayırarak incelerler. Pazarlama faaliyetlerinde müşteriler bazı özellikler açısından gruplara ayrılarak değerlendirilir. Günlük yaşamın bir parçası olarak her gün kullanılan kara taşıtları segment olarak tanımlanan farklı gruplara ayrılırlar.

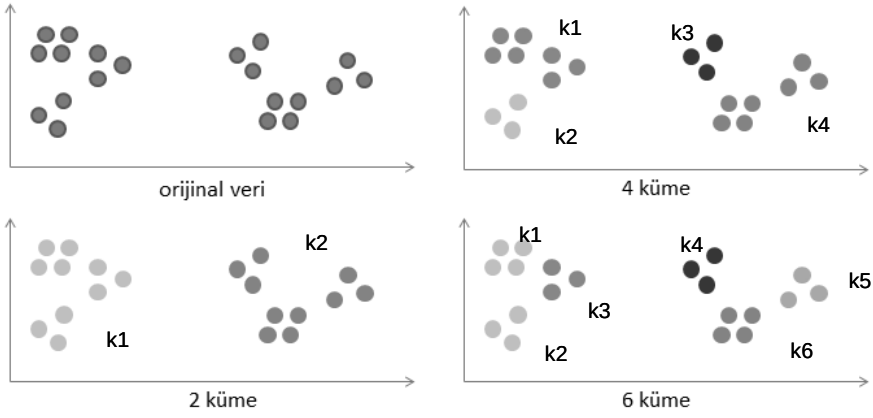
Yukarıdaki örneklerin hepsinde öğeler belli özelliklerine göre ait oldukları küme (sınıf veya grup) içine atanırlar. Öğelerin sınıfa veya gruba atanmasında yani kümelemede gerekli özellikler veya kriterler atama işlemi öncesinde bellidir. Yapılan işlem bir sınıflama işlemidir. Sınıflama analizi ile öğeler ait oldukları sınıf etiketleriyle etiketlenmekte ya da işaretlenmektedirler. Bu bir önbilgi dâhilinde gerçekleştirilir ve yani eğitimli bir sınıflama işlemidir. Bu nedenle makine öğrenmesi gibi bazı bilim dallarında *eğitimli öğrenme* olarak da adlandırılmaktadır.

Kümeleme analizi incelenen veri kümesindeki öğeleri gruplamak için kullanılan yöntem ve teknikleri kapsamaktadır. Ancak kümeleme analizinde öğelerin atanacağı sınıf üyelikleri hakkında daha önce elde edilmiş herhangi bir önbilgi yoktur. Öğeler ve içinde yer aldıkları gruplar ya da kümeler daha önceki önbilgi ile değil, doğrudan analiz ile saptanmaya çalışılmaktadır. Kümeleme analizi tipik şekilde verilen özellikleri kullanarak ve neden olduklarını araştırmaksızın sadece sınıfsal yapıları keşfetmeyi hedefler. Bu yüzden kümeleme analizi normalde büyük bir problemin küçük bir ön çözümünü amaçlar. Sonuç olarak verideki yapıları ya da örüntüyü görmek ve daha sonraki analizlerde kullanmak üzere yapılan keşifsel bir analizdir. Bu nedenle, bu tür bir sınıflama *eğitimsiz öğrenme* işlemidir. Bazı disiplinlerde, örneğin biyolojide *sınıflama analizi* veya *sayısal taksonomi* olarak da adlandırılmaktadır.

Yukarıdaki açıklamalardan da anlaşılacağı üzere, kümeleme analizi örüntü tanıma ve önceden atanmış etiketlerle sınıflama kurallarını araştıran ayırma analizi ve karar analizi gibi yöntemlerden farklıdır. Ayırma analizi, öğelerin ait olduğu küme (ya da grup veya sınıf) üyelikleri hakkında önbilgilere gereksinim duyar. Bu bilgiler sınıflama kuralı oluşturmak için kullanılır. Kümeleme analizi sınıfları bulmada bir ön sonuç sunmak için kullanışlı olabilirken diğer analizlerde çok daha fazlası söz konusudur. Örneğin, veri kümesi öğelerini temsil edecek özelliklerin neler olduğuna karar vermek örüntü tanıma gibi uygulama alanlarında temel amaçlardan biridir.

Kümeleme analizi tek bir yöntem veya teknik değildir. Aksine veri öğelerinin kendileri ve ilişkilerini açıklamak amacıyla, onları küme olarak adlandırılan gruplara ayırmak için kullanılan yüzlerce farklı algoritma veya yaklaşımları kapsayan çok değişkenli analiz yöntemlerinden oluşmaktadır. Aslında verilerdeki küme veya sınıfları bulmak sanatı olarak da tanımlanmaktadır (Kaufman ve Rousseeuw, 1990). Kümeleme birbirine benzeyen öğelerin yani birbirine yakın olanların aynı gruba; farklı olanların ise diğer gruplara yerleştirilmesiyle gerçekleştirilir. Amaç, kendi içlerinde mümkün olduğunca birbirine benzer (homojen) öğelere sahip olan ancak birbirinden mümkün olduğunca uzak grupların elde edilmesini sağlamaktır. Bir başka deyişle, aynı kümedeki öğelerin olabildiğince birbirine benzeyen öğeler olmaları; diğer kümelerdekilerden farklı olmaları kümeleme analizinde temel amaçtır.

Kümelerdeki öğelerin benzerliği yüksek ve kümeler arasında farklılık ne denli fazla ise oluşturulan kümeler belirgin biçimde birbirinden ayrık olurlar ki, bu kümeleme işleminin iyi ve kaliteli yapıldığını gösterir. Ancak elde edilecek küme sayısı ve bunların verideki hangi öğeleri kapsayacağı konusunda etkin bir istatistiksel test söz konusu değildir. Bir öğenin birden fazla kümeye üye olmasına imkan veren bulanık kümeleme bir istisna olmak üzere, kümeleme analizinde genel olarak öğelerin birbiriyle çakışmayan kümelere ayrılması araştırılır.



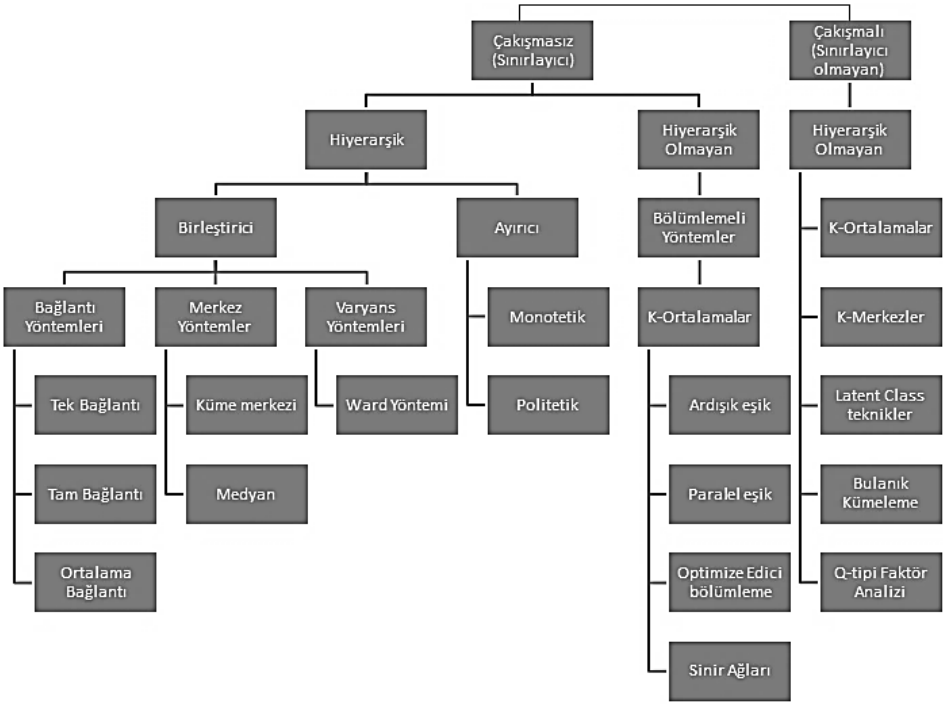
Şekil 1. Aynı veriden elde edilmesi muhtemel kümeler

Çok sayıda kümeleme yönteminin mevcudiyetine rağmen kümeleme işleminin geçerliliği, yani analiz sonucunda elde edilen küme sayıları ve kümelenme şekli bakımından çeşitli güçlükler söz konusu olabilmektedir. Şekil 1'de 20 öğe için muhtelif kümeleme sonuçları görülmektedir. Şeklin sol üstündeki grafiğe bakıldığında orijinal verinin iki kümeden oluştuğu/oluşacağı düşünülebilir. Ancak daha dikkatli bakıldığında bu iki kümeden herbirinin üçer alt kümeden oluştuğu düşünülerek toplam 6 küme olduğu sonucuna varılabilir (Şekil 1'in sağ alt köşesindeki gösterildiği gibi). Ancak bu değerlendirme insan görme sisteminin bir kusuru da sayılabilir. Orijinal veriyi dört kümeye ayırmak da akla gelebilir varılabilir (Şekil 1'in sağ üst köşesindeki gösterildiği gibi). Ancak bu kümelemeler için makul bir neden belirtmek ve yorumlamak fazlasıyla güçtür. Bu nedenle kümeleme analizinde, kümeleme işleminin geçerliliği üzerinde durulması gereken çok önemli bir husus olup yoğun çalışılan araştırma konuları arasında yer almaktadır. Diğer yandan kümeleme işleminin geçerliliği, yani kaç küme oluşturulacağına karar vermek için üzerinde çalışılan veriyi iyi incelemek ve anlamak gerekli olup alan uzmanlığı ve deneyim de gerektirir.

Kümeleme Yöntemleri

İlk olarak Tryon (1939) tarafından kullanıldığından beri üzerinde en çok çalışılan çok değişkenli istatistik analiz yöntemlerinden biri olarak kümeleme analizi için geliştirilmiş çok sayıda yöntem, teknik ve algoritma söz konusudur. Downs ve Barnard (1995) kümeleme yöntemlerinin en genel şekliyle *hiyerarşik* ve *hiyerarşik olmayan* yöntemler olmak üzere iki gruptan oluştuğunu belirterek çok basit bir sınıflama vermişlerdir. Kümeleme yöntemleri üzerinde birçok araştırıcı tarafından yapılan inceleme ve derlemelerde çeşitli taksonomiler önerilmiş olmasına karşın ideal bir taksonomi olmadığını söyleyebiliriz. Çünkü bir kümeleme yöntemi farklı ölçütlere göre farklı sayıda kümeleme yöntemi sınıfında yer alabilir. Bu nedenle kümeleme yöntemleri için belli bir sınıflama/taksonomi vermek yerine Sneath ve Sokal'ın yaptığı gibi herhangi bir kümeleme yöntemini aşağıdakilere uygunluk bakımından çeşitli kategorilere dâhil etmek mümkündür (Sneath & Sokal, 1973).

- *Hiyerarşi: Hiyerarşik yöntemler / Hiyerarşik olmayan*
- *Kümeleme yönü: Birleştirici / Ayırıcı*
- *Geçiş durumu: Tek geçişli / Çok geçişli*
- *Eşanlılık: Ardışık / Eşanlı, Tek değişkenli geçiş / Çok değişkenli geçiş*
- *Ölçek: Yerel / Genel*
- *Yineleme: Direkt / Yinelemeli (İteratif)*
- *Ağırlık: Ağırlıklı / Ağırlıksız*
- *Uyumsallık: Adaptif / Adaptif olmayan*
- *Katılık: Katı / Yumuşak*
- *Çakışma / Sınırlama: Çakışmalı / Çakışmasız, Sınırlayıcı / Sınırlayıcı olmayan*



Şekil 2. Kümeleme yöntemleri taksonimisi

Yukarıda anlatıldığı gibi kümeleme yöntemlerine ait belli bir taksonomi önermek zor olsa da başlangıç itibariyle kullanılabilecek basit taksonomilerden biri aşağıda sunulmaktadır:

- Hiyerarşik yöntemler
    - Birleştirici yöntemler
        - Bağlantı yöntemleri
            - Tek bağlantı
            - Tam bağlantı
            - Ortalama bağlantı
        - Merkezci yöntemler
            - Küme merkezleri
            - Ortanca
        - Varyans yöntemleri
            - Ward
        - Diğer Yöntemler
            - BIRCH, CURE, CHAMALEON, OPTICS

- Ayırıcı yöntemler
  - Monotetik ayırıcı yöntemler
    - MONA algoritması
  - Politetik ayırıcı yöntemler
    - DIANA algoritması

- Hiyerarşik Olmayan Yöntemler
  - Bölümleyici yöntemler
    - Tek geçişli yöntemler
      - Leader algoritması
    - Uzaklık Tabanlı Yöntemler
      - Katı algoritmalar
        - K-ortalamalar
        - K-medoid (PAM)
        - K-mod
        - SOM
        - CLARANS
      - Yumuşak algoritmalar
        - Bulanık C-ortalamalar (FCM)
        - Olabilirlikli C-ortalamalar (PCM)
        - Bulanık ve Olabilirlikli Karma algoritmalar (FPCM, PFCM)
  - Yoğunluk tabanlı yöntemler
    - En yakın komşu tabanlı yöntemler
      - DBSCAN, OPTICS, DENCLUE,CLIQUE…
    - Izgara tabanlı yöntemler
      - BANG, STING, CLIQUE, WaveCluster, PROCLUS, ORCLUS, OPTIGRID, …
  - Model tabanlı yöntemler
    - MLE
    - Beklenti çoklaması (EM)
  - Grafik tabanlı modeller

- Karma Yöntemler

Kümeleme Analizinin Aşamaları

Kümeleme analizi, veri öğelerini belli kümelere ayıran bir algoritmanın işletilmesi veya hesaplamanın yapılmasından ibaret tek aşamalı bir işlem değil aşağıda listelenen aşamalardan oluşan çok aşamalı bir analizdir.

1. Önişleme işlemleri
    a. Veri kümesinin okunması ve analize hazırlanması
    b. Kayıp değerlerin tahmin edilmesi
    c. Normalleştirme veya standartlaştırma işlemlerinin yapılması
2. Yakınlık ölçülerinin hesaplanması ve uzaklık matrisinin oluşturulması
3. Kümeleme yöntemi/algoritmasının belirlenmesi
4. Kümeleme analizinin yapılması
5. Kümeleme grafiklerinin çizilmesi
6. Kümeleme geçerliğinin kontrol edilmesi
7. Kümeleme profilinin oluşturulması
8. Kümeleme sonuçlarının yorumlanması

Kümeleme analizinde ilk aşama hemen her istatistik analiz veya veri madenciliği için de geçerli olan analiz öncesi işlemlerdir. Veri önişleme olarak adlandıran bu aşamada analiz edilecek veri kümesinin okunması ve analize hazır hale getirilmesi hedeflenmektedir. Kümeleme işlemi uygulamadan önce veri kümesinin incelenmesi varsa kayıp veya eksik öğelerin tamamlanması veya tahmin yoluyla veriye eklenmesi gereklidir. Araştırıcının veri türlerini ve boyutunu incelemesi uygun analiz yöntem veya tekniğine karar vermek açısından son derece kritik bir konudur. Ayrıca nicel sürekli değişkenlere bazı hallerde standartlaştırma ve/veya normalleştirme işlemlerine de analiz öncesi gerek duyulabilir. Verideki aykırı değerler veya gürültülerin saptanması ve elimine edilmesi de gerekli olabilir.

Kümeleme analizinde uygun yönteme karar vermek de önemlidir. Karar sırasında gerek çalışma disiplini, gerek veri türü ve boyutu ve gerekse işlem hızı ve maliyeti belirleyici olmaktadırlar. Örneğin, biyolojik ya da ekolojik taksonomi çalışmalarında sıkça hiyerarşik kümeleme kullanılmaktadır. Öte yandan bilgi sistemlerinde veri madenciliği veya sosyal medya için metin madenciliği gibi uygulamalarda hiyerarşik olmayan bir kümeleme yöntemi; mümkün olduğunca hızlı işlem imkânı sağlayan bir yöntem veya algoritma amaçlanabilmektedir. Günümüzde veri madenciliği uygulamalarında büyük-veri (big data) söz konusudur ve mutlaka çok hızlı çözüm sağlayan yöntemlere gereksinim duyulmaktadır.

Küme geçerliliği, yapılan kümelemenin ne kadar uygun olduğunu ya da kümelemenin uyum iyiliğini test etmektir. Bunun için çeşitli çapraz doğrulamalar ve istatistiksel ölçütlere başvurulur. Küme profilleme, ilgili boyutlarda nasıl farklılaştığını izah etmek için her bir kümenin karakteristiklerini açıklar. Bunun için ayırma analizi veya varyans analizi (ANOVA) gibi yöntemlere başvurulur.

Kümeleme Analizinin Uygulama Alanları

Kümeleme analizi, tarımdan astronomiye; gen biliminden bilgisayar bilimlerine hemen hemen tüm temel ve uygulamalı disiplinlerde kullanılan bir analizdir. Tablo 1'de bilim dalları itibariyle kümeleme analizinin kullanıldığı alanlardan bazıları listelenmektedir.

Tablo 1. Kümeleme analizi uygulama alanı örnekleri

Uygulama Alanı	Kullanım Örnekleri
Ekoloji	Farklı ortamlardaki organizma topluluklarının mekânsal ve zamansal karşılaştırmalarını yapmak.
Biyoloji Biyoinformatik Genetik	İlişkili ifade örüntülerinden gen gruplarını saptamak. Evrim biyolojisi ve biyoinformatik çalışmalarında önemli bir konu olan gen aileleri için homolog dizilişleri gruplandırmak.
Kimya	Çok sayıda bileşiği çok sayıda özelliğine göre gruplandırarak yapısal benzerliklerini ortaya koymak.
Eczacılık	İlaç keşif çalışmaları yapmak.
Tıp	PET taramalarında üç boyutlu görüntülerde kan ve doku türleri arasındaki farklılığı, farklı grupları ortaya koymak. Hastalıklar için tedavi yöntemleri ve belirtileri kümeleyerek tıbbi taksonomiler oluşturmak.
Sosyal Ağlar	Sosyal ağ kullanıcı gruplarını belirlemek.
Öneri Sistemleri	Kullanıcılara kendilerine benzeyenlerden üretilen grup özelliklerine göre ürün ve hizmet önerilerinde bulunmak.
Bilgi erişimi	Arama ve bilgiye erişimde Web siteleri ve belgeleri gruplandırmak.
Görüntü İşleme	Nesne tanıma, kenar bulma amaçlarıyla görüntüyü farklı bölümlere ayırmak (görüntü bölümleme).
Suç analizi	Farklı alanları suç özelliklerine göre gruplara ayırmak.
Eğitim	Okullar ve öğrencileri özelliklerine göre gruplandırmak.
İklim/Meteoroloji	İklim rejimleri, hava durum analizi için gruplandırmalar yapmak.
Tarım	Benzer arazi kullanım özelliklerine sahip tarımsal alanları tanımak ve sınıflandırmak.
Pazarlama	Ele alınan özelliklere göre farklı müşteri gruplarını belirlemek ve hedef pazarlama programları tasarlamak.
Kentsel planlama	Türü, değeri ve coğrafi konumuna göre konutları gruplandırmak ve gruplara ilişkin planlama çalışmaları yapmak.
Deprem çalışmaları	Fay hatları boyunca deprem merkez üslerini gözlemek.
Bilgisayar/Yazılım Mühendisliği	Görüntü analizi/işleme, Görüntü bölümleme, Makine öğrenmesi, Örüntü tanıma, Arama motorları, Bilgi erişimi, Veri madenciliği, Ses madenciliği, Metin madenciliği, Sunucu/Web kümeleme.

Hiyerarşik Kümeleme Yöntemleri

Kümeleme analizinde en çok kullanılan yöntemlerden olan hiyerarşik kümelemede, kümeler biri diğerinin içinde yer alacak şekilde yuvalandırılarak oluşturulurlar. Ebeveyn küme olarak adlandırılan her bir üst küme, çocuk küme olarak adlandırılan alt kümeleri içerir. Ebeveyn kümeler böylece çocuk kümelerdeki elemanları da kapsarlar. Küme oluşumunda böylece alt-üst ilişkisine dayanan bir hiyerarşi söz konusu olduğundan bu tür kümeleme yöntemleri *hiyerarşik kümeleme* olarak adlandırılırlar. Hiyerarşik kümeleme yöntemleri kümeleri ve aynı zamanda hiyerarşiyi ortaya koyarlar.

Kümeleme yöntemleri öğeler arasındaki uzaklıkları kullanarak birbirine en yakın olanları bir araya getirerek bir küme ya da grup içine almayı amaçlarlar. Hiyerarşik kümeleme yöntemleri uzaklıkları kullanmalarına göre farklılık gösterirler:

- En küçük uzaklık tabanlı algoritmalar
- En küçük ortalama uzaklık tabanlı algoritmalar
- En küçük uzaklık kareleri toplamı tabanlı algoritmalar

Literatürde çeşitli hiyerarşik kümeleme yöntemleri ve teknikleri önerilmiş olup bunları aşağıdaki şekilde sınıflandırmak olasıdır:

- *Birleştirici yöntemler*
    - *Bağlantı yöntemleri*
        - *Tek bağlantı yöntemi*
        - *Tam bağlantı yöntemi*
        - *Ortalama bağlantı yöntemi*
            - *Tartısız ortalama bağlantı yöntemi (McQuitty yöntemi)*
            - *Tartılı ortalama bağlantı yöntemi*
    - *Merkezci yöntemler*
        - *Kitle merkezi (centroid) yöntemi*
        - *Ortanca (median) yöntemi (Gower yöntemi)*
    - *Varyans tabanlı yöntemler*
        - *Ward yöntemi*
        - *Ward II yöntemi*
    - *Esnek yöntemler*
        - *Esnek beta yöntemi*
        - *Genelleştirilmiş yöntem (Lance-Williams yöntemi)*

- *Ayırıcı yöntemler*
    - *Monotetik ayırıcı yöntemler*
        - *MONA algoritması*
    - *Politetik ayırıcı yöntemler*
        - *DIANA algoritması*
- *İleri yöntemler*
    - *BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)*
    - *CURE (Clustering Using Representative)*
    - *CHAMALEON (A Hierarchical Clustering Algorithm Using Dynamic Modeling)*
    - *OPTICS (Ordering Points to Identify the Clustering Structure)*