ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

İsmet BİRBİÇER

VERİ MADENCİLİĞİNDE MODELE DAYALI KÜMELEME ANALİZİ

İSTATİSTİK ANABİLİM DALI

ADANA, 2017

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE MODELE DAYALI KÜMELEME ANALİZİ

İsmet BİRBİÇER

YÜKSEK LİSANS TEZİ

İSTATİSTİK ANABİLİM DALI

Bu Tez 15/12/2017 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından Oybirliği/Oyçokluğu İle Kabul Edilmiştir.

……..	.....…….	……
Prof. Dr. Ali İhsan GENÇ	Doç. Dr. Deniz ÜNAL	Doç. Dr. Gülin TABAKAN
DANIŞMAN	ÜYE	ÜYE

Bu Tez Enstitümüz İstatistik Anabilim Dalında Hazırlanmıştır.
Kod No:

Prof. Dr. Mustafa GÖK
Enstitü Müdürü

ÖZ

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİNDE MODELE DAYALI KÜMELEME ANALİZİ

İsmet BİRBİÇER

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK

Danışman : Prof. Dr. Ali İhsan GENÇ
Yıl: 2017, Sayfa: 97
Jüri : Prof. Dr. Ali İhsan GENÇ
: Doç. Dr. Deniz ÜNAL
: Doç. Dr. Gülin TABAKAN

Klasik analiz yöntemleri, büyük ölçekli verilere uygulandığında verimli sonuçlar elde edilememektedir. Bu sebeple veri madenciliği yöntemleri, çeşitli disiplinlerin ortak çalışmaları ile geliştirilmiştir. Veri madenciliğinin temel amaçlarından biri de verideki bilinmeyen grup yapısını belirlemektir. Kümeleme analizi verideki gruplanmayı belirlemeyi amaçlayan yöntemlerin genel adıdır. Bu yöntemlerden biri olan modele dayalı kümeleme analizi, sonlu karma modeller kullanarak veriyi kümelemektedir. Sonlu karma modellerin kümeleme analizinde kullanımı ile küme belirleme problemi istatistiksel modelleme problemine dönüştürülmüş olur. Bu tez çalışmasında modele dayalı kümeleme analizi incelenmiş. Bu yöntemin veri madenciliği uygulamaları gerçekleştirilmiş ve diğer yöntemlerle karşılaştırılması yapılmıştır.

Anahtar Kelimeler: Kümeleme analizi, Kümeleme, Veri madenciliği, Sonlu karma modeller, Modele dayalı kümeleme

ABSTRACT

MASTER THESIS

MODEL BASED CLUSTER ANALYSIS IN DATA MINING

İsmet BİRBİÇER

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Prof. Dr. Ali İhsan GENÇ
Year: 2017, Pages: 97
Jury : Prof. Dr. Ali İhsan GENÇ
: Assoc. Prof. Dr. Deniz ÜNAL
: Assoc. Prof. Dr. Gülin TABAKAN

Classical analysis methods do not yield efficient results when applied to large scale datasets. For this reason, data mining methods have been developed through collective studies of various disciplines. One of the main purposes of data mining is to identify the unknown group structure on the dataset. Cluster analysis is the general name of methods that aim to determine the grouping on the dataset. One of these methods, which is model-based cluster analysis, is clustering dataset using finite mixture models. With the use of finite mixture models in the cluster analysis, the problem of cluster determination is transformed into statistical modeling problem. In this thesis, model-based clustering analysis is examined. Data mining applications of this method have been implemented and compared with other methods.

Keywords: Cluster analysis, Clustering, Data mining, Finite mixture models, Model-based clustering

GENİŞLETİLMİŞ ÖZET

Her geçen gün veri saklama becerimiz gelişmekte ve bu saklanan verilerden anlam çıkarma ihtiyacımız artmaktadır. Bu amaç için makine öğrenme ve istatistik disiplinlerinin ortak çalışması ile veri madenciliği yöntemleri geliştirilmektedir. Kümeleme analizi, veri madenciliğinde gözetimsiz öğrenme metotları altında yer alan ve verinin doğal grup yapısını bir önbilgi olmadan belirleyen yöntemlerdir.

Kümeleme analizi yöntemleri kabaca hiyerarşik ve bölümleyici yöntemler olmak üzere iki alt başlıkta incelenmektedir. Hiyerarşik yöntemlerde kümeler bir ağaç yapısıyla inşa edilmekteyken, bölümleyici yöntemlerde ise iteratif algoritmalar yardımıyla bir kriter fonksiyonu minimize edilerek gerçekleştirilmektedir. Bu iki tip yönteme sezgisel yöntemler de denilmektedir. Bundan farklı olarak verinin,

$$f(x;\pi) = \sum_{k=1}^{K} \pi_k f_k(x)$$

sonlu karma dağılımından geldiğini varsayan modele dayalı kümeleme yöntemleri istatistiksel çerçevede verideki grup yapısını belirleyen yöntemlerdir. Burada π_k, sıfır ve bir arasında değer alan karma oranı $f_k(x)$ k-ıncı bileşen dağılımın ne ağırlıkta modelde bulunduğunu göstermektedir. Modele dayalı kümeleme analizinde, sonlu karma modellerden geldiği varsayılan veri istatistiksel parametre tahmin metotları ile bileşen dağılımlarının parametreleri tahmin edilmektedir. Belirlenen her bir bileşen dağılımının elemanları bir küme olarak belirlenerek kümeleme analizi tamamlanmaktadır.

Sonlu karma dağılımların kümeleme analizinde kullanımı verideki alt grupları belirlemede büyük bir özgürlük tanımaktadır. Kümeleme analizi için sıklıkla normal dağılım karması,

$$f(x;\Theta) = \sum_{k=1}^{K} \pi_k f_k(x;\mu_k,\Sigma_k)$$

modeli yardımıyla belirlenmekte ve π_k, μ_k ve Σ_k için parametre tahmini EM (Expectation-Maximization) algoritması ile iteratif olarak belirlenmektedir. Bununla beraber bu modele Poisson bileşeni eklenerek sapan gözlemler otomatik olarak belirlenebilmektedir (Fraley ve Raftery, 2002). Ayrıca BIC (Bayesian information criteria) veya AIC (Akaike information Criteria) gibi kriterler yardımıyla küme sayısı otomatik olarak belirlenebilmektedir.

Kümeleme analizindeki en önemli zorluklardan bir yüksek değişkenli verilerde uygulamada ortaya çıkmaktadır. Bu durumda kümelemeyle ilgili olmayan değişkenlerin belirlenip kümeleme analizinden çıkarılması büyük bir önem taşımaktadır. Normal dağılım karma modelleri ile modele dayalı kümeleme analizi için değişken seçimi yapılabilmektedir.

Bu çalışmada kümeleme analizi yöntemlerinin performanslarının çeşitli senaryolar altında üretilen veriler üzerinden değerlendirmesi sağlanmıştır. Modele dayalı kümeleme analizinin veri madenciliğinde uygulanışı seeds verisi üzerinde detaylı olarak incelenmiştir ve değişken seçim yönteminin kümeleme performansını iyileştiridiği gözlenmiştir. Ayıca modele dayalı kümeleme analizinin diğer yöntemlerden üstünlükleri dışsal kriterler kullanılarak gerçek sınıflama verileri üzerinde gösterilmiştir.

TEŞEKKÜR

Bu tez çalışması sürecinde konunun belirlenmesinde ve gelişmesinde destek olan ikinci danışmanım Prof.Dr. Hamza EROL'a ve tez oluşma sürecinde yardımlarını esirgemeyen danışmanım Prof.Dr. Ali İhsan GENÇ'e teşekkürlerimi sunarım. Son olarak hocalarım ve çalışma arkadaşlarım, Çukurova Üniversitesi İstatistik Bölümü öğretim elemanlarına teşekkür ederim.

Motivasyon kaynağım olan ve her zaman desteklerini yanımda hissettiğim aileme teşekkür ederim.

İÇİNDEKİLER **SAYFA**

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

X	Veri matrisi
n	Gözlem sayısı
n_i	i-inci küme için gözlem sayısı
p	Değişken sayısı
K	Küme sayısı
x_i	i-inci gözlem
m	Küme merkezleri vektörü
m_i	i-inci küme merkezi
C	Kümeleme sonucu elde edilen kümeler vektörü
C_i	i-inci küme
π_k	k-ıncı karma oranı
Θ	Sonlu karma modelin parametre vektörü
θ_k	Bileşen dağılımı için parametre vektörü
$\overline{w}$	Kümeler için ortalama karmaşıklık

1. GİRİŞ

Bilgisayar sistemlerindeki gelişmeler ile veri üretimi ve bu verileri saklama becerimiz her geçen gün önemli ölçüde artmaktadır. Öyle ki, önemli olmayan ve eskiden kaydı tutulmayan bilgilerin dahi artık rahatlıkla kaydı tutulabilmektedir. Elde edilen bu veriler terabaytlarca yer kaplayabilir, yüksek sayıda değişken içerebilir ve ayrıca heterojen bir yapıda bulunabilir. Bu tipteki veriler için klasik yöntemler yetersiz kalmaktadır. Bu nedenle büyük verilerden bilgi sağlanması için yeni yöntemler geliştirilmesi gerekmiştir. Bu gereksinim veri madenciliğinin ortaya çıkmasını sağlamıştır (Tan ve Steinbach, 2005). Veri madenciliği, bilgisayar bilimleri ve istatistiğin ortak kullanımıyla büyük veri yapılarından bilgiye ulaşmak için özel olarak geliştirilmiş yöntemlerle ilgilenmektedir.

Veri madenciliği yöntemleri tahmin ve açıklama olmak üzere iki genel amaç için uygulanmaktadır. Tahmin amaçlı yöntemler, açıklayıcı değişkenler yardımıyla bir hedef değişkeninin tahmin edilmesi için bir model kurularak gerçekleştirilmektedir. Açıklayıcı yöntemler ise eğilim, sapan gözlemler, gözlemler arası ilişki ve küme yapısı gibi veride gizli olarak bulunan özellikleri ortaya çıkarıp özetlemekte kullanılmaktadır. Açıklayıcı veri madenciliği yöntemlerinden biri olan ve gözetimsiz öğrenme (unsupervised learning) yöntemleri altında incelenen kümeleme analizi, verideki heterojen yapıyı inceleyen ve verinin içerdiği homojen grupları belirleyen yöntemlerdir. Bir kümeleme analizi yöntemi veriye uygulandığında aynı kümede olduğu belirlenen gözlemler başka kümedekilere göre birbirlerine daha fazla benzeyeceklerdir.

Kümeleme analizi; biyolojide canlıların gruplanmasında, çeşitli yazılı dokümanların sınıflanmasında, gen ifade verilerinin analizi ve veri indirgeme gibi birçok farklı alanda uygulanmaktadır. Bu uygulama çeşitliliği sebebiyle kümeleme analizi yöntemlerinin çeşitli tipteki veriler için uygulanabilir olmalıdır. Berkhin (2006) veri

madenciliğinde kullanılacak olan bir kümeleme algoritmasında bulunması beklenen özellikleri aşağıdaki şekilde vermiştir.

- Farklı veri tiplerine uygulanabilmesi,
- Büyük boyutlardaki verilere uygulanabilir olması,
- Yüksek değişkenli verilere uygulanabilir olması,
- Karmaşık tipteki kümeleri belirleyebilmesi,
- Sapan gözlemleri belirleyebilmesi,
- Hızlı olması,
- Gözlemlerin etiketlenme türü (Sert veya bulanık),
- Yorumlanabilir sonuçlar üretmesi.

Modele dayalı kümeleme analizi, sonlu karma dağılım modellerinden faydalanarak kümeleme problemine çözüm arayan yöntemlerin genel adıdır. Sonlu karma dağılımlar verinin birden fazla alt gruptan oluştuğunu ve bu alt grupların farklı dağılımlardan ya da aynı dağılımın farklı parametrelere sahip modellerinden oluştuğunu varsaymaktadır. Literatürdeki ilk sonlu karma dağılım çalışması Pearson (1894), tarafından; yengeçlerin alın genişliklerinin vücut uzunluklarına oranına, iki bileşenli normal dağılım karma modelinin uydurulması ile gerçekleştirilmiştir. Sonlu karma modellerin kümeleme analizinde kullanılması ise ilk olarak Wolfe (1965), çalışmasında normal dağılım karması ile gerçekleştirmiştir. Bu çalışmada Wolfe normal dağılım karmalarının maksimum olabilirlik tahminlerini hesaplamış ve bu işlem için 4 farklı yöntemi içinde barındıran bir yazılım oluşturmuştur. Fakat hazırlanan bu yazılımın 5 değişken veya 6 bileşenden büyük veri gruplarına uygulandığında yöntemin etkinliğinde azalma olduğu görülmüştür. Day (1969), çalışmasında ise normal dağılım karmaları için kovaryans matrislerinin bileşenler arasında farklılık göstermediği durumlar için iteratif bir çözüm yöntemi geliştirilmiş ve bu yöntemin kümeleme analizi ile olan ilişkisi üzerinde durulmuştur. Wolfe (1970), çalışmasında kovaryans

matrislerinin sırasıyla sabit ve değişkenlik gösteren çok değişkenli normal dağılım karmalarının maksimum olabilirlik tahminini iteratif bir yaklaşımla belirlemiştir. Bu yöntemin kümeleme analizine uygunluğunu göstermek için Fischer'in süsen çiçekleri ile ilgili verisine ve farklı kovaryans matrisine sahip olarak üretilmiş simülasyon verisine geliştirdiği yöntemi uygulamıştır.

Dempster ve arkadaşları (1977), kayıp veri durumunda maksimum olabilirlik tahminini elde etmek için iteratif bir yöntem olan EM (Expectation-Maximization) algoritmasını geliştirmişlerdir. EM algoritması zaman içerisinde karma modellerin parametre tahminini hesabında kullanılmaya başlamış ve günümüzde bu amaç için kullanılan temel yöntem olmuştur. EM algoritmasının popülerleşmesi ile yöntemin birçok farklı varyasyonu türetilmiş ve karma dağılımlara uygulanmıştır. Bu varyasyonlardan bazıları stokastik EM (Celeux ve Diebolt, 1985), Monte Carlo EM (Wei ve Tanner,1990), sınıflama EM (Celeux ve Govaert, 1992), Hızlandırılmış EM (Jamshidian ve Jennrich, 1993), emEM (Biernacki ve ark, 2003), rndEM (Maitra, 2009) şeklindedir.

Normal dağılımın karması için bileşenlerinin kovaryans matrislerine özdeğer ayrışımı uygulanarak kümelerin hacim, şekil ve yön gibi geometrik özelliklerinin kısıtlanabileceği üzerinde durmuşlardır. Celeux ve Govaert (1995), bu geometrik özelliklerin sadece birini, hiçbirini veya hepsini kısıtlayarak 14 farklı model oluşturmuş ve her bir model için parametre tahminini elde etmişlerdir. Fraley ve Raftery (1999), S-PLUS programı için bu 14 modelin 8'ini içeren MCLUST paketini oluşturmuşlardır. MCLUST 2-inci versiyonuyla birlikte açık kaynak kodlu bir yazılım olan R programı (R Core Team, 2017) için bir paket olarak yayınlanmıştır. Bu paketin R'da yayımlanması modele dayalı kümeleme analizinin popülerleşmesine önemli bir katkı sağlamıştır. Yöntemin popülerleşmesinde bir diğer önemli katkı ise Fraley ve Raftery (2002) modele dayalı kümeleme, ayırma analizi ve yoğunluk tahmini isimli çalışmadır. Bu çalışmada

modele dayalı kümeleme analizi konusunda yapılan önemli çalışmaları derleyip yöntemin kolay uygulanması için bir çerçeve oluşturulmuştur.

Modele dayalı kümeleme yöntemleri fazla sayıda değişkene sahip verilerle başa çıkmada yetersiz kalmaktadır. Yüksek sayıda değişkene sahip verilere kümeleme analizi değişken seçimi yöntemleri ile gerçekleştirilmektedir. Raftery ve Dean (2006), çalışmasında değişken seçimi problemini adımsal olarak model karşılaştırma problemine dönüştürerek değişken seçimi yapılmıştır. Maugis ve arkadaşları (2009), ise bu yöntemi genelleştirmiştir. Ayrıca, yöntemin geliştiğinin kanıtlanması için simülasyon ve gerçek veri uygulamaları yapılmıştır.

Wehrens ve arkadaşları (2004), veride çok fazla sayıda gözlem bulunması durumunda modele dayalı kümelemenin çok yavaş işlemesinden söz etmişlerdir. Bu sebeple büyük veriden örnekleme seçmeye dayalı bir yöntem geliştirmiş ve imge bölütleme uygulamasında bu yöntemi kullanmışlardır. Fraley ve Raftery (2007), çalışmalarında modele dayalı kümelemenin kemometride Mclust paketi yardımıyla nasıl uygulanacağını göstermişlerdir. Vu ve arkadaşları (2013), sonlu karma modellerden faydalanarak ağ verilerinin kümelenmesi için bir çerçeve belirlemişlerdir ve bu çerçevenin işe yararlığını 175 milyar ayrıtı ve 135 bin boğumu olan bir ağ verisi üzerinde göstermişlerdir. Jacques ve Preda (2014), çok değişkenli fonksiyonel veri için modele dayalı kümeleme uygulayabilen bir yöntem önermiş ve bu yöntemin etkinliği simülasyon ve gerçek veri çalışması ile incelenmiştir. Melnykov (2016) tıklama davranışı verisine birinci sıra Markov modelleri üzerinden modele dayalı kümeleme uygulamıştır.

Bu tez çalışması beş bölümden oluşmaktadır. İkinci bölümde küme kavramı, kümeleme analizi, kümeleme analizinin uygulanması için gereken ön işlemleri ve kümeleme yöntemlerini karşılaştırma kriterler incelenmiştir. Üçüncü bölümde

kümeleme analizi yöntemleri alt başlıklara parçalanmış ve bu alt başlıklar altındaki yöntemler incelenmiştir. Dördüncü bölümde modele dayalı kümeleme analizi ve veri madenciliğindeki yeri incelenmiştir. Beşinci bölümde modele dayalı kümeleme analizinin ve diğer yöntemlerin, gerçek veri ve simülasyon üzerinden performansları değerlendirilmiştir. Altıncı bölümde ise elde edilen sonuçlar açıklanmıştır.

2. KÜMELEME ANALİZİ

Kümeleme analizi verideki benzer gözlemleri kendi aralarında kümelere atayan yöntemlerin genel adıdır. Günümüze kadar kümeleme analizi için birçok farklı tanım yapılmıştır. Bu tanımlardan bazıları aşağıda belirtildiği gibidir:

- Kategori yapısı bilinmeyen verinin doğal grup yapısını bulmaya çalışan yöntemdir (Anderberg, 1973).
- Gözlemlerin değişkenleri arasındaki benzerliklere göre oluşturulabilecek, görece olarak az sayıdaki homojen gruplar arayan yöntemdir (Bailey, 1975).
- Kümeleme, benzerlikleri temel alarak örüntülerin kümelere atanma organizasyonudur (Jain ve arkadaşları, 1999).
- Veriyi anlamlı, kullanışlı veya hem anlamlı hem de kullanışlı olacak şekilde gruplara ayıran yöntemlerdir (Tan ve arkadaşları, 2005).
- Kümeleme analizi, aynı küme içindeki gözlem çiftlerinin karşılıklı uzaklığı farklı kümelerdeki gözlem çiftlerinin karşılıklı uzaklığından daha az olacak şekilde gözlemleri gruplara atayan yöntemlerdir (James ve arkadaşları, 2013).

Temel gösterimlerle ifade edilirse, kümeleme analizi $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ şeklinde olan $\mathbf{X}$ verisindeki *p* değişkenli *n* gözlemi birbirlerine olan uzaklıklarına veya benzerliklerine göre *K* sayıdaki $\mathbf{C} = \{C_1, C_2, ..., C_K\}$ şeklinde farklı gruplara atayan yöntemdir. x_i gözleminin C_k kümesinin bir elemanı olduğu bu gözlemin sınıf değişkenine *k* ataması yapılarak belirtilir. Veriye Kümeleme analizinin uygulanması sonucunda belirlenen $\mathbf{C}$ kümeleri uygulanan yönteme göre farklılık göstermektedir.

Elde edilen bu kümelenme yapıları arasında hangisinin daha iyi ve/veya daha doğru olduğunu belirlemek kümeleme analizinin en zor problemlerinden biridir. Çünkü en iyi küme kavramının tek bir tanımı yoktur ve genellikle elde edilmek istenen kümenin kalitesi uygulamanın amacına göre değişmektedir. Henning (2016), kümeleme analizinin kullanım alanlarından bazılarını aşağıdaki gibi vermiştir.

- Biyolojide canlı türlerinin gruplanması,
- Hastalıkların sınıflanması,
- Arkeolojide zaman ve bölgelerin ayrıştırılarak keşfedilmesi,
- Nesne tanıma ve imge bölütleme,
- Piyasa bölümlendirme,
- Veri tabanlarının etkin bir şekilde düzenlenmesi.

Bu örneklerin her birinde küme olarak belirlenen gruplar, birbirinden farklı kavramlar ifade etmektedir. Bu sebeple küme kavramını için tek bir tanım vermek oldukça zordur. Buna bağlı olarak elde edilmek istenen kümenin yapısı, amaca göre farklılık gösterdiğinden kümenin aşağıdaki özelliklerden bir veya birkaçına sahip olması gerekmektedir (Henning, 2015).

1. Küme içerisindeki gözlemler birbirine yakın olmalı.
2. Farklı kümelerdeki gözlemler arasındaki uzaklık fazla olmalı.
3. Kümeler çeşitli olasılık modellerine uygun olmalı.
4. Kümelerin merkezleri (centroid) kendi kümesindeki elemanları temsil edebilmeli.

5. Verinin uzaklık matrisi uygulanan kümeleme ile iyi temsil edilebilmelidir.
6. Kümeler durağan olmalı.
7. Kümeler verideki yüksek yoğunluklu ve bağlantılı bölgeler olmalı.
8. Kümelerin belirli karakteristikleri olmalı.
9. Kümeler az sayıda değişkenle karakterize edilebilmeli.
10. Önceden bilinen sınıf değişkenleri varsa oluşturulan kümeler bu değerlerle uyumlu olmalı.
11. Değişkenler kümeden kümeye yaklaşık olarak bağımsız olmalı.
12. Tüm kümeler yaklaşık olarak aynı boyutta olmalı.
13. Küme sayısı az olmalı.

Kümelerin yukarıdaki özelliklerden hangilerini sağlayacağına karar vermek, çoğunlukla uygulanmasını gereken kümeleme yöntemini belirlemede önemlidir.

Kümeleme ve sınıflama analizi yöntemlerinin ikisi de veriyi gruplara ayırmakta kullanılmaktadır fakat bu yöntemlerin kullanım amacı ve çalışma biçimleri birbirlerinden oldukça farklıdır. Kümeleme analizinde amaç verinin homojen parçalanışını belirlemek iken sınıflama analizinde veriye yeni eklenen gözlemlerin hangi gruplara atanacağını tahmin etmektir. Kümeleme analizi uygulanacak veride önceden bilinen gözlemlerin hangi gruplara atandığını gösteren bir sınıf değişkeni bulunmamaktadır. Sınıflama analizinde ise verideki gözlemlerin hangi grupların elemanı olduğunu gösteren bir sınıf değişkeni mevcuttur. Bu sebeple kümeleme analizi literatürde gözetimsiz öğrenme (unsupervised learning), sınıflama analizi ise gözetimli öğrenme (supervised learning) ismini almıştır.

Milligan (1996), kümeleme analizinin uygulanabilmesi için yapılması gereken tercihleri aşağıdaki gibi vermiştir.

1. Kümelenecek gözlemlerin seçimi: Seçilen gözlemler var olduğuna inanılan bir küme yapısını temsil etmelidir.

2. Değişken seçimi: Kümeleme analizinde yalnızca doğal bir küme yapısı ilgili bilgi içerdiğine inanılan değişkenleri içermelidir. Alakasız değişkenler içeren bir veriye kümeleme analizi uygulanması hem zaman kaybına yol açmakta hem de elde edilen kümelerin kalitesini de olumsuz etkilemektedir.

3. Kayıp gözlemler: Veri kümesinin düşük oranda kayıp veri içermesi durumunda kayıp veri içeren gözlemler çıkarılarak kümeleme uygulanabilmektedir.

4. Değişkenlerin standartlaştırılması: Veri kümesinin standartlaştırılması her zaman gerekli değildir ve bazı durumlarda değişkenlerdeki kümeleme bilgisinin yok olmasına sebep olabilmektedir. Milligan ve Cooper (1988), kümeleme analizi uygulamalarında standartlaştırma gerekliyse standart sapma yerine açıklık değeri kullanılarak verinin standartlaştırılması gerektiğinin üzerinde durmuşlardır.

5. Yakınlık ölçüsü: Kullanılacak yakınlık ölçüsünün belirlenmesi Bölüm 2.1.'de verilmiştir.

6. Kümeleme yöntemi: Hangi kümeleme yönteminin kullanılacağına karar verirken elde edilmek istenen küme yapısına göre davranmak gerekmektedir. Elde edilmek istenen küme yapısı biliniyor ise o küme yapısını elde eden yöntemler arasında yapılan kıyaslamalar incelenerek bir yöntem seçilebilmektedir.

7. Küme sayısının belirlenmesi: Kümeleme analizinde genellikle grup yapısı önceden bilinmez. Bu sebeple küme sayısının otomatik olarak belirleyen yöntemler önemli bir avantaja sahiptir.

8. Küme geçerliliği: Elde edilen kümelerin anlamlı olup olmadığını belirlemek ya da uygulanan kümeleme yöntemlerinin hangisinin daha iyi sonuç verdiğine karar vermek için küme geçerliliği yöntemlerinin uygulanması gerekmektedir.

9. Yorumlama: Kümeleme analizi sonucu belirlenen kümeler arasındaki farklar betimleyici istatistikler ile yorumlanabilir. Diğer bir yöntem ise verinin grafiksel olarak incelenip yorumlanmasıdır. Veri kümesinin 3 değişkenden fazla değişken içermesi durumunda verideki değişkenliği en çok açıklayan iki temel bileşenin grafiğini çizdirerek kümeleme analizi görsel olarak yorumlanabilmektedir.

Bunlara ek olarak Everitt ve arkadaşları (2011), kümeleme analizi uygulanmadan önce verinin gerçekten bir küme yapısı içerip içermediğinin incelenmesi gerektiği üstünde durmuşlardır.

2.1. Yakınlık Ölçüleri

Kümeleme analizi yöntemlerinin çoğunluğu veriyi gruplarken gözlemlerin arasındaki yakınlıklardan faydalanmaktadır. Bu sebeple kümeleme performansını geliştirmek için doğru uzaklık ölçüsünün belirlenmesi oldukça önemlidir. Yakınlık kavramı ile uzaklık ve benzerlik kavramlarının birleşimi temsil edilmektedir. Kümeleme analizinde verideki kümelerin belirlenmesi diğer bir deyişle birbirlerine yakın olan gözlemlerin gruplanması için gözlemler arası benzemezlik veya uzaklık ölçülerinin belirlenmesi gerekmektedir.

x_i ve x_k gözlemleri arasındaki benzemezlik ya da uzaklık fonksiyonu $d(x_i, x_k)$ şeklinde gösterilir ve aşağıdaki özellikleri sağlar.

- $d(x_i, x_k) = d(x_k, x_i)$ (Simetri)
- $d(x_i, x_k) \geq 0$ (Pozitiflik)
- $d(x_i, x_k) \leq d(x_i, x_l) + d(x_l, x_k)$ (Üçgen eşitsizliği)
- $d(x_i, x_i) = 0$

Bu koşulların hepsini sağlayan uzaklık ölçüsü metrik adını alırken, yalnızca üçgen eşitsizliğini sağlamayanlar yarı metrik adını alır.

x_i ve x_k gözlemleri arasındaki benzerlik ise $s(x_i, x_k)$ fonksiyonu ile gösterilir ve aşağıdaki koşulları sağlar.

- $s(x_i, x_k) = s(x_k, x_i)$ (Simetri)
- $0 \leq s(x_i, x_k) \leq 1$ (Pozitiflik)
- $s(x_i, x_l)s(x_l, x_k) \leq [s(x_i, x_l)s(x_l, x_k)]s(x_i, x_k)$
- $s(x_i, x_i) = 1$

Yakınlık matrisi, *n* gözlem içeren bir veri için *n*x*n* boyutlu ve gözlem ikilileri arasındaki farklılık veya benzerliklerini gösteren simetrik matristir.

2.1.1. Nicel Değişkenler İçin Uzaklık Ölçüleri

Nicel ölçüm tipindeki değişkenler gerçek sayı değerleri almakta ve bu değişkenlerin yakınlık ölçüleri farklılık ya da uzaklık ölçüsü olarak hesaplanabilmektedir (Everitt, 2011). Nicel değişken seçimi için en sık kullanılan

uzaklık ölçüsü öklidyen uzaklıktır. Ayrıca, bu uzaklığa L_2 norm da denmektedir. x_{ij} i-inci gözlemin j-inci değişkeni olmak üzere,

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2} \tag{2.1.}$$

formülü ile hesaplanmaktadır. Öklidyen uzaklık metrik olmanın tüm koşullarını sağlamaktadır. Bu ölçümün kümeleme analizinde kullanılması ile küresel şekillere sahip kümeler elde edilmektedir (Xu ve Wunsch, 2008). Değişkenlerin farklı birimlerde ifade edilmesi durumunda öklidyen uzaklığın kullanılması büyük birimlere sahip değişkenlerin kümeleme sonucunu fazla etkilemesine sebep olmaktadır. Bu nedenle, farklı birimli değişkenler içeren verilerde bu uzaklık ölçümü kullanılmadan önce değişkenlerin standartlaştırılması gerekmektedir.

Öklidyen uzaklığın genelleştirilmiş hali Minkowski uzaklığıdır. L_r normu olarak da ifade edilen bu metrik,

$$d(x_i, x_k) = \sqrt[r]{\sum_{j=1}^{p} (x_{ij} - x_{kj})^r} \tag{2.2.}$$

formülü ile hesaplanır. Minkowski uzaklığında $r = 2$ alınması durumunda öklidyen uzaklık elde edilirken, $r = 1$ olması durumunda Manhattan uzaklığı (L_1 norm),

$$d(x_i, x_k) = \sum_{j=1}^{p} \left| x_{ij} - x_{kj} \right| \tag{2.3.}$$

elde edilir. $r = \infty$ alınması ile belirlenen sup uzaklığı (L_∞ normu) ise aşağıdaki gibidir.

$$d(x_i, x_k) = \underset{1 \le j \le p}{maks} \left| x_{ij} - x_{kj} \right|$$

Gözlemler arası uzaklık ayrıca korelasyon katsayıları ile de belirlenebilmektedir. En yaygın kullanılan Pearson korelasyon katsayısı,

$$r_{ik} = \frac{\sum_{j=1}^{p}(x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)}{\sqrt{\sum_{j=1}^{p}(x_{ij} - \bar{x}_j)^2 \sum_{j=1}^{p}(x_{ij} - \bar{x}_j)^2}}$$

formülü ile hesaplanmaktadır. Korelasyon katsayısı -1 ve 1 arasında değer almaktadır. Bu değerin bir uzaklık ölçüsüne dönüşümü,

$$d(x_i, x_k) = \frac{1 - r_{ik}}{2}$$

eşitliği ile sağlanmaktadır. İki gözlemin uzaklığı hesaplanırken korelasyon katsayısı kullanılması iki gözlem arasındaki uzaklığın büyüklüğünü değil, şekilsel bir farklılığı göstermektedir (Xu ve Wunsch, 2008).

Nicel değişkenlerde kullanılan bir diğer benzerlik ölçüsü ise Kosinüs Benzerliğidir.

$$s(x_i, x_k) = \cos\alpha = \frac{x_i^T x_k}{\|x_i\| \|x_k\|}$$

Bu benzerlik ölçüsünde büyük kosinüs değerleri iki gözlem arasında yapısal bir benzerliği temsil etmektedir. Yani kosinüs benzerliği, Pearson korelasyonuna benzer şekilde gözlemler arasındaki farkın büyüklüğü ile ilgili bilgi vermemektedir.

2.1.2. İkili Değişkenler İçin Benzerlik Ölçüleri

p sayıda yalnız iki değer alabilen değişkenler içeren $x_i = \{x_{i1}, x_{i2}, ..., x_{ip}\}$ ve $x_j = \{x_{j1}, x_{j2}, ..., x_{jp}\}$ gözlemleri arasındaki benzerliklerin incelenmesi için iki gözlemin aynı değişkenlerinin bulunabilecekleri dört farklı durum bulunur. Bunların her ikisi de 0 olabilir, her ikisi de 1 olabilir ya da biri 0 biri 1 olabilmektedir. I indikatör fonksiyonu göstermek üzere bu dört farklı durumun toplam sayıları a_1, a_2, a_3 ve a_4 gösterilir ve,

$$a_1 = \sum_{l=1}^{p} I(x_{il} = x_{jl} = 1),$$

$$a_2 = \sum_{l=1}^{p} I(x_{il} = 0, x_{jl} = 1),$$

$$a_3 = \sum_{l=1}^{p} I(x_{il} = 1, x_{jl} = 0),$$

$$a_4 = \sum_{l=1}^{p} I(x_{il} = x_{jl} = 0)$$

formülü ile hesaplanırlar. Bu hesaplanan değerler ikili değişkenler için benzerlik ölçülerinin hesaplanmasında kullanılmaktadır. İkili değişkenler için benzerlik ölçüleri Çizelge 2.1.'deki gibidir. Bu ölçüler gözlemlerin uyuşma veya uyuşmama durumlarına verdiği ağırlıkla birbirinden ayrılmaktadır. Bu yöntemlerin seçimi uygulanacak verideki ikili değişkenlerin çeşidine göre belirlenmektedir. İkili değişkenler arası benzerliğin hesaplanması için farklı benzerlik ölçüleri geliştirilmesinin en önemli sebebi, sıfır-sıfır eşleşmesinin (a_4 durumu) benzerliğe olan etkisine karar vermektir. İkili değişkenlerde 0 genelde bir yokluk (mevcut olmama) durumunu temsil etmektedir. İki gözlem arası uzaklık hesaplanırken, bir özelliğin iki gözlemde de mevcut olmamasının onları daha fazla benzer yapıp yapmadığı tamamen uygulamaya ve değişken tanımına göre

değişmektedir (Everitt ve ark, 2011). Örneğin verimizde sıfır-sıfır eşleşmeleri sayısının benzerliğe etkisi olmadığı düşünülüyorsa Çizelge 2.1.'deki (2), (3) veya (6) benzerlik ölçülerinden biri kullanılmalıdır.

Çizelge 2.1 İkili değişkenler için benzerlik ölçüleri (Everitt ve ark,2011)

Sıra	Ölçü	Formül
1	Basit Eşleme Katsayısı	$s(x_i, x_k) = \frac{(a_1 + a_4)}{a_1 + a_2 + a_3 + a_4}$
2	Dice	$s(x_i, x_k) = \frac{2a_1}{2a_1 + a_2 + a_3}$
3	Jaccard Katsayısı	$s(x_i, x_k) = \frac{a_1}{a_1 + a_2 + a_3}$
4	Russel ve Rao	$s(x_i, x_k) = \frac{a_1}{a_1 + a_2 + a_3 + a_4}$
5	Gower ve Legandre	$s(x_i, x_k) = \frac{a_1 + a_4}{a_1 + \frac{1}{2}(a_2 + a_3) + a_4}$
6	Sneath ve Sokal (1973)	$s(x_i, x_k) = \frac{a_1}{a_1 + 2(a_2 + a_3)}$

2.1.3. İkiden Fazla Değer Alan Kesikli Değişkenler

İkiden fazla farklı değerler alabilen kesikli değişkenlerin benzerliği için genellikle basit eşleşme kriteri kullanılmaktadır. x_i ve x_k p değişkenli bir uzaydaki iki gözlem olsun. Bu gözlemler arasındaki uzaklık,

$$s(x_i, x_k) = \frac{1}{p}\sum_{j=1}^{p} S_{ikj}$$

şekilde hesaplanır. Burada,

$$S_{ikj} = \begin{cases} 1, & x_{ij} = x_{kj} \\ 0, & x_{ij} \neq x_{kj} \end{cases}$$

şeklinde hesaplanmaktadır.

Kategorik değişkenlerin, sıralı ölçeklere sahip değerler alması durumunda iki gözlemin arasındaki uzaklık, Öklidyen veya Manhattan uzaklığı kullanılarak hesaplanabilir (Kaufman ve Rousseeuw, 1990). Değişkenlerin alabileceği sıralı değerlerin sayısı birbirlerinden farklı olabilir. Bu durumda M_k k-ıncı değişkenin alabileceği değerlerin sayısı olmak üzere, i-inci gözlemin j-inci değişkeninin aldığı değer r_{ij}, [0,1] arasındaki değerler alan r_{ij}^* değerine aşağıdaki gibi dönüştürülür (Xu ve Wunsch, 2008).

$$r_{ij} = \frac{r_{ij}^* - 1}{M_k - 1}$$

2.1.4. Karma Tipteki Veriler İçin Uzaklık Ölçüleri

Hem kategorik hem de sürekli değişkenler içeren verilere karma veri denilmektedir. Gerçek veriler, genellikle bu tanıma uymaktadır. Bu sebeple bu tipteki veriler için uzaklık ölçülerinin belirlenmesi oldukça önemlidir. Bu tipteki verilerin uzaklık ölçülerinin belirlenmesi için ilk akla gelen çözüm tüm değişkenlere bir dönüşüm uygulanarak [0, 1] aralığına dönüştürmek sürekli değişkenler için kullanılan yöntemler kullanılarak gözlemler arası yakınlığı belirlemektir. Ancak bu durumda dönüştürülen kategorik veriler anlamlı değerler almayacağından bu yöntem uygun olmamaktadır. Diğer bir alternatif ise sürekli verilerin kategorik verilere dönüştürülmesidir. Bu yöntem ise bilgi kaybına sebep olmaktadır (Xu ve Wunsch, 2008).

Gower (1971), daha güçlü bir yöntemi aşağıdaki gibi önermiştir.

$$s(x_i, x_k) = \frac{\sum_{j=1}^{p} w_{ikj} s_{ikj}}{\sum_{j=1}^{p} w_{ikj}}$$

Burada s_{ikj} i-inci ve k-ıncı gözlem arasındaki benzerliği göstermektedir. w_{ikj} ise iki gözlem arasındaki karşılaştırma geçerli ise 1 değerini geçerli değilse 0 değerini almaktadır. Benzerlik değerleri değişken sürekli ise

$$s_{ikj} = 1 - \frac{|x_{ij} - x_{kj}|}{R_j}$$

şeklinde hesaplanmaktadır. Burada R_j j-inci değişken için değişim aralığını göstermektedir. Değişkenin kesikli olması durumunda ise

$$s_{ikj} = \begin{cases} 0, & x_{ij} \neq x_{kj} \\ 1, & x_{ij} = x_{kj} \end{cases}$$

olarak hesaplanır.

2.2. Kümeleme İçin Geçerlilik Kriterleri

Kümeleme analizinde tek başına her veri için en iyi sonucu veren tek bir yöntem önermek mümkün değildir. Bu sebeple bir veriye kümeleme analizi uygulamadan önce o veri için en uygun olan yönteminin belirlenmesi gerekmektedir. Bu belirleme işlemi için veriye çeşitli kümeleme yöntemleri uygulanmakta ve elde edilen sonuçlar çeşitli geçerlilik kriterleri yardımıyla karşılaştırılarak gerçekleştirilir. Geçerlilik kriterleri, kümeleme analizinin yaptığı gruplamaların kalitesini araştıran yöntemlerdir (Halkidi ve ark, 2016). Geçerlilik kriterleri;

1. Verinin kümeleme analizi uygulanması için uygun bir yapıya sahip olup olmaması.
2. Küme sayısının belirlenmesi
3. Gerçek sınıf değişkeni bilinmeyen veri için uygulanan yöntemin veriye uyumunun test edilmesi
4. Gerçek sınıf değişkeni bilinen veriler için kümeleme algoritmalarının karşılaştırılması
5. Aynı veriye uygulanan çeşitli kümeleme yöntemlerinin karşılaştırılması

amaçları için kullanılmaktadır (Tan, 2006). Geçerlilik kriterleri dışsal ve içsel kriterler olmak üzere iki alt kategoride incelenmektedir.

2.2.1 Dışsal Kriterler

Bu kriterlerde veriye kümeleme yöntemi uygulayarak elde ettiğimiz kümeler verinin önceden bilinen gerçek sınıf değişkeni ile kıyaslanmaktadır. Sınıf değişkeni önceden bilindiği için özellikle belirli veri setlerinin kümelenmesinde gerçek sınıf değerlerine olan uyumluluğu en yüksek kümeleme yönteminin araştırılmasında kullanılmaktadır. Ayrıca bu kriterler uygulanan iki kümeleme arasındaki benzerlik incelenirken de kullanılmaktadır. Bu kriterlerin çoğunluğu çapraz tablo (Çizelge 2.3.) yardımıyla hesaplanmaktadır.

X, *n* gözlem ve *p* değişken içeren verisi $C = \{C_1, C_2, ..., C_K\}$ şeklinde birbirinden ayrık olan *K* kümeye bir kümeleme analizi yöntemiyle ayrıştırılmış olsun. $G = \{G_1, G_2, ..., G_K\}$ ise kümelerin önceden belirlenmiş gerçek sınıf değerleri olsun. Oluşabilecek tüm gözlem çiftleri düşünüldüğünde,

- Hem *C* hem de *G*'de aynı kümede bulunan gözlem çiftleri sayısı *a*
- *C*'de aynı kümede ve *G*'de farklı kümede bulunan gözlem çiftleri sayısına *b*
- *C*'de farklı kümede ve *G*'de aynı kümede bulunan gözlem çiftleri sayısına *c*
- Hem *C* hem de *G*'de farklı kümede bulunan gözlem çiftleri sayısına *d*

denilsin. Bu durumda *a* ve *d* iki gruplama arasındaki uyumlu sonuç sayısını gösterirken *b* ve *c* uyumsuz sonuçların sayısını gösterir ve *M* oluşabilecek toplam çiftlilerin sayısı olan $\binom{n}{2}$'ye eşittir.

R, ARI, J ve FM kriterleri (Çizelge 2.2.), gözlem çiftlerinin uyum yapılarına dayalı olarak belirlenmektedir. Bu kriterlerden Rand indeksi (R) Çizelge 2.2.'de görüldüğü gibi uyumlu ikililer sayısının toplam ikililer sayısına oranını göstermektedir. Yani yapılan gruplama ile gerçek küme değerlerinin tam uyumlu olması durumunda Rand indeksi bir değerini alacaktır. Rand indeksi (Çizelge 2.2.), kümeleme yöntemi ile gerçek küme değerlerinin aynı sonucu belirleme oranını vermektedir. Rand indeksinin beklenen değeri gözlem ve küme sayısına bağlı olarak artmaktadır. Bu sebeple fazla gözlem olan verilerde Rand indeksi ile kıyaslama zorlaşmaktadır. Bu sorunu ortadan kaldırmak için düzeltilmiş rand indeksi (ARI) geliştirilmiştir. Rand indeksin beklenen değerini hipergeometrik dağılım varsayımı altında kabul etmekte ve,

$$\frac{R - E(R)}{\max(R) - E(R)}$$

genel formülü ile hesaplanmaktadır.

Bu kriterlerden F ölçüsü kesinlik (precision) ve duyarlık (recall) ile hesaplanmaktdır. F ölçüsü [0,1] aralında değer almakta ve 1'e yakın değerler iyi kümeleme analizinin gerçek kümelerle uyumlu olduğunu göstermektedir.

Çizelge 2.2.Kümeleme analizi için dışsal geçerlilik kriterleri

	İndeks	**Formül**
R	Rand (1971)	$\frac{a+d}{M}$
ARI	Düzeltilmiş Rand (Hubert ve Arabie, 1985)	$\frac{\sum_{i,j}\binom{n_{ij}}{2}-\left[\sum_i\binom{s_i}{2}\sum_j\binom{t_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{s_i}{2}+\sum_j\binom{t_j}{2}\right]-\left[\sum_i\binom{s_i}{2}\sum_j\binom{t_j}{2}\right]\Big/\binom{n}{2}}$
J	Jaccard (1963)	$\frac{a}{a+b+c}$
FM	Fowkles ve Mallows (1983)	$\sqrt{\frac{a^2}{(a+b)(a+c)}}$
Γ	Γ istatistiği (Hubert ve Arabie, 1985)	$\frac{M(a-(a+b)(a+c))}{\sqrt{(a+b)(a+c)(M-(a+c))}}$
F	F-Ölçüsü	$\frac{1}{n}\sum_{i=1}^{K} n_i \max_j\{2n_{ij}/m_{j.}+s_{.i}\}$

Çizelge 2.3. C ve G kümelemeleri için çapraz tablo

	C_1	C_2	…	C_K	Toplam
G_1	n_{11}	n_{12}	.	n_{1K}	$s_{.1}$
G_2	n_{21}	n_{22}		n_{2K}	$s_{.2}$
⋮	⋮	⋮	⋱	⋮	⋮
G_K	n_{K1}	n_{K2}	…	n_{KK}	$s_{K.}$
Toplam.	$t_{.1}$	$t_{.2}$	…	$t_{.K}$	

2.2.2. İçsel Kriterler

Kümeleme analizi bir gözetimsiz öğrenme metodu olması sebebiyle çoğu zaman elde edilen kümenin doğru olup olmadığını belirlememiz için önceden bilinen bir sınıf

değişkeni bulunmamaktadır (Halkidi ve ark, 2002). Bu durumda küme performansının kalitesinin belirlenmesi için sadece veride bulunan bilgiler kullanılarak hesaplanan içsel kriterler geliştirilmiştir. Ayrıca bu yöntemler yardımıyla küme sayısı belirlenebilir veya farklı yöntemlerin performanslarını kıyaslanabilir. Bu bilgiler küme içi gözlemlerin yoğunluğu, kümeler arası ayrıklık ve gözlemler arası bağlantılardır.

İyi bir küme yapısı genellikle küme içi gözlemlerin yakınlığı, kümeler arası gözlemlerin ayrıklığı ve yakın gözlemlerin bağıntılı olması ile tarif edilmektedir. Dn, DB ve CH kriterleri bu tariften yola çıkarak hesaplanmaktadır. Dunn indeksi sapan değerlere karşı hassas olmakla birlikte küme içi uzaklık ve küme hiperküresinin çapı ile hesaplanmaktadır. Bir diğer çeşit içsel kriter ise kareler toplamlarına dayalı olarak hesaplanmaktadır. Bunlar ise BH ve Xu kriterleridir (Çizelge 2.4.). Ayrıca modele dayalı kümeleme içersinde model seçiminde kullanılan BIC kriteri de klasik yöntemlerin geçerliliği için kullanılabilmektedir. BIC kriteri detaylı olarak Bölüm 4.2.3.'te incelenecektir.

2.3. Yüksek Boyutlu Verilerde Kümeleme Analizi

Çok sayıda değişken tarafından açıklanan verilere yüksek boyutlu veri denmektedir. İlk bakışta gözlemlerin çok sayıda değişken içermesi veri ile ilgili daha çok bilgimiz olmasını sağlamaktadır ve bu iyi bir şeydir. Ancak uygulamada durum bunun tam tersidir. Verinin yüksek sayıda değişken içermesi birbiri ile ilişkili değişkenlerin bulunmasına sebep olmaktadır. Bu da aynı bilgilerin analize tekrar katılmasını sağlar ve analiz sonuçlarının karmaşıklaşmasına sebep olur. Not olarak büyük boyutlu veri kavramı için kesin bir değişken sayısından söz edemeyiz bazı çalışmalar sadece 10 değişkeni yüksek boyut olarak kategorilerken özellikle gen ifade verilerinde boyut sayısı binlere ulaşabilmektedir (Assent, 2012).

Çizelge 2.4. Kümeleme analizi için içsel değerlendirme kriterleri

Kısaltma	**Kriter**	**Formül**
Dn	Dunn (1974)	$\dfrac{\min\limits_{c_i,c_j\in C}\{\min\limits_{x\in C_i,x'\in C_j}(x-x')^2\}}{\max\limits_{k=1}^{K}\{\max\limits_{x,x'\in C_k}(x-x')^2\}}$
DB	Davies ve Bouldin (1979)	$R_{ij}=\dfrac{\frac{1}{n_i}\sum_{r=1}^{n_i}(x_r-m_i)^2+\frac{1}{n_j}\sum_{r=1}^{n_j}(x_r-m_j)^2}{(m_i-m_j)^2}\quad ,i\neq j$ $DB=\dfrac{1}{K}\sum_{i=1}^{K}\max\limits_{j=1,2,\ldots K}\{R_{ij}\}$
CH	Calinski ve Harabazs (1974)	$\dfrac{\sum_{i=1}^{K}n_i(m_i-\bar{X})^2/(K-1)}{\sum_{i=1}^{N}(x_i-m_{c_i})^2/(N-K)}$
BH	Ball ve Hall (1965)	$\sum_{i=1}^{N}(x_i-m_{c_i})^2/K$
Xu	Xu (1997)	$p\log\left(\sqrt{\sum_{i=1}^{N}(x_i-m_{c_i})^2/(pN^2)}\right)+\log(K)$

Kümeleme analizi yöntemlerinin çoğunlukla uzaklık ölçüleriyle uygulandığı Bölüm 2.1'de bahsedilmişti. Uzaklık ölçüleri hesaplamada değişkenlerin sayısındaki artış gözlemler arasındaki uzaklık farklarının önemsizleşmesine neden olmaktadır (Beyer ve ark, 1999). Kümeleme analizinde boyutsallık problemine çözüm olarak değişken indirgeme veya değişken seçimi yöntemleri uygulanmaktadır. Değişken indirgeme yöntemleri verideki değişkenlerin anlamlı bilgilerini taşıyan daha az sayıda

değişkene dönüştürerek gerçekleştirilir. Değişken seçiminde ise amaç verideki anlamlı bilgi kaybı minimum olması koşuluyla verideki en iyi alt değişkenler kümesini belirlemektir (Bouveyron ve ark, 2007).

3. KÜMELEME ANALİZİ YÖNTEMLERİ

Kümeleme analizi disiplinler arası bir yöntem olması sebebiyle ilk kullanımından günümüze kadar geçen zaman içerisinde kullanılan yöntemler birçok farklı şekilde sınıflandırılmıştır. 1990'lı yıllara kadar basit bir biçimde hiyerarşik ve bölümleyici şeklinde iki grup altında incelenen kümeleme yöntemlerinin günümüzde bu kadar basit biçimde gruplandırmak mümkün değildir. Verilerin artan boyutu ve karmaşıklaşan yapısı klasik kümeleme yöntemlerinin yetersiz kalmalarına sebep olmuştur. Bu sebeple eski yöntemler güncellenerek yeni yapıya uyum sağlaması amaçlanmış ve tamamen yeni yapıdaki yöntemler geliştirilmiştir. Bu bölümde kümeleme analizi yöntemlerinden olan bölümleyici, hiyerarşik ve yoğunluk tabanlı yöntemler incelenecektir. Tez çalışmasının temel konusu olan modele dayalı kümeleme analizi ise dördüncü bölümde detaylı olarak incelenecektir.

3.1. Bölümleyici Yöntemler

Bölümleyici kümeleme, belirli bir amaç fonksiyonunu optimize ederek verinin küme yapısını bulmayı sağlayan yöntemlerdir (Aggarwal, 2013). Bölümleyici yöntemlerde ilk olarak veri için uygun uzaklık ölçüsü belirlenmeli sonrasında ise algoritmanın belirli bir kriter fonksiyonu optimize etmesi için uygulanacak iteratif yöntemin ilk adımı için verinin bir başlangıç parçalanışı seçilmelidir (Han ve ark, 2011). Bu grup altında incelenen yöntemler kümeleme analizinin en popüler yöntemleridir. Bölümleyici yöntemler aşağıdaki yapıda işlemektedir.

1. *K* tane başlangıç merkezi (centroid) belirlenir.
2. Belirlenen küme merkezleri yardımıyla gözlemler kümelere atanır
3. Kümeler içerisindeki gözlemler kullanılarak küme merkezleri güncellenir

4. 2 ve 3. adımlar kümelerdeki gözlemlerde önemli değişiklikler olmayana kadar tekrarlanır.

Bölümleyici yöntemlerin 3 değiştirilebilir faktörü vardır. Bu faktörler değiştirilmesiyle farklı kümeleme problemlerine daha iyi çözüm sağlayan yöntemler elde edilmektedir. Bu faktörler başlangıç merkezlerinin nasıl belirleneceği, küme merkezlerinin nasıl belirleneceği ve gözlemlerin kümelere nasıl atanacağıdır.

3.1.1. K-means

K-means (Lloyd 1957; Macqueen 1967) yöntemi en eski, en çok bilinen ve en sık kullanılan kümeleme algoritmalarından biridir. Wu ve arkadaşlarının (2008)'deki çalışmasında en iyi 10 veri madenciliği algoritması arasında gösterilmiştir. Bu algoritmanın uygulama ve yorumlamasının kolaylığı en sık kullanılan algoritmalardan biri olmasına sebep olmuştur. Kısaca K-means yöntemi Öklid uzayında n gözlemi olan **X** verisini (3.1.)'deki hata kareler ortalaması kriterini iteratif olarak minimum yapan k sayıda kümeyi arayan yöntemdir. Hata kareler toplamını, C_j kümesinin merkezi m_j'ye göre türevini alıp sıfıra eşitleyerek hata kareler toplamının minimumu kümedeki gözlemlerin ortalaması alınarak hesaplanabileceği,

$$W_K = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - m_k)^2 \tag{3.1.}$$

$$\frac{\partial W_k}{\partial m_k} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \frac{\partial}{\partial m_k} (x_i - m_k)^2$$

$$= \sum_{x_i \in C_k} 2(m_k - x_i) = 0$$

$$= m_k = \frac{\sum_{x_i \in C_k} x_i}{n_k}$$

gösterilebilmektedir. Klasik bir K-means kümeleme algoritması aşağıdaki adımlarla çalışmaktadır (Steinley, 2006).

1. Başlangıç küme merkezleri olan $\mathbf{m}^k = \{m_1^k, m_2^k, ..., m_p^k\}$ $1 \le k \le K$ olacak şekilde tüm k kümenin p değişkeni için belirlenir. Küme merkezleri ile gözlemler arası uzaklık,

$$d(x_i, m_j) = \sqrt{\sum_{j=1}^{p} (x_{ij} - m_j^k)^2} \qquad (3.2.)$$

ile tüm gözlemler için belirlenir. Gözlemler bu uzaklığı minimum yapan $\mathbf{m}^k$ kümelerine atanır.

2. Her bir değişken için yeni küme merkezleri $m_j^k = \sum_{i \in C_k} x_{ij} / n_k$ şeklinde hesaplanır ve yeni $\mathbf{m}^k = \{m_1^k, m_2^k, ..., m_p^k\}$, $1 \le k \le K$ küme merkezleri belirlenir.
3. Gözlemlerin her bir küme merkeziyle arasındaki öklidyen uzaklık hesaplanır ve en yakınında bulunan kümeye atanır.
4. (2) ve (3) adımları kümeler arası gözlem değişimleri bitene kadar tekrarlanır.

(3.1.) fonksiyonu matris gösterimleri kullanılarak,

$$F(\mathbf{M}, \mathbf{U}) = tr[(\mathbf{X} - \mathbf{UM})'(\mathbf{X} - \mathbf{UM})] \qquad (3.3.)$$

şeklinde **M** ve **U**'nun bir fonksiyonu olarak yazılabilmektedir. Burada **X** veri matrisi, $\mathbf{U}_{nxK}$ i-inci gözlem *k*-ıncı kümenin bir elemanıysa 1 değilse 0 elemanı alan u_{ixk} değerlerinden oluşan bir matris. $\mathbf{M}_{Kxp}$ ise elemanları *k*-ıncı kümenin tüm değişkenler için merkez noktalarını gösteren $\mathbf{m}_k = [m_{1k}, m_{2k}, ..., m_{pK}]$ vektörlerinden oluşan $\mathbf{M}' = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_K]$ matrisidir. En küçük kareler algoritması ile (3.3.) eşitliği bilinen U değeri ile F(**M**,**U**)'yu minimize eden **M** matrisi hesaplanarak **U** güncellenir. Bu işlemler $F(\mathbf{M}, \mathbf{U})$'daki değişim sabitlenene kadar tekrarlanır.

K-means basitliği sayesinde hızlı çalışmaktadır. Bu sebeple büyük verilerde uygulamaya elverişli bir yöntemdir (Xu ve Wunsch, 2008). Yöntemin hızlı çalışması sebebiyle, hesap karmaşıklığı fazla olan modele dayalı kümeleme ve DBSCAN gibi birçok yöntemin başlangıç adımlarında kullanılmaktadır.

K-means algoritması tüm popülerliğine rağmen birçok dezavantajı da bulunmaktadır. Bunlar:

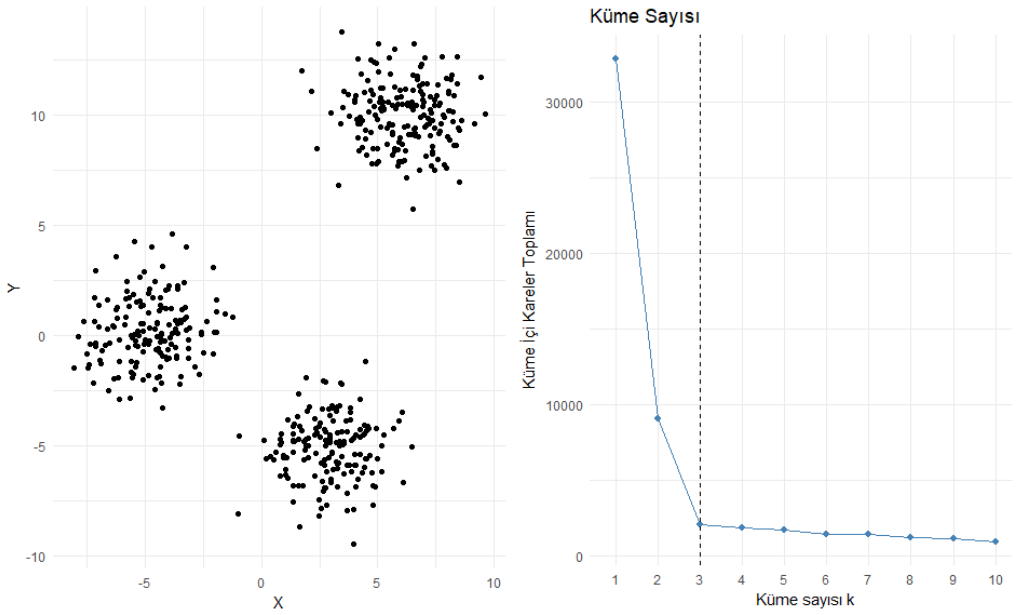
- Küme sayısı *K*-nın önceden bilinmesi gerekmektedir
- Genellikle yerel minimum çözüme yakınsar
- Kategorik verilere uygulanamaz
- Yalnızca küresel şekildeki kümeleri daha iyi belirler
- Sapan değerlere karşı dayanıklı değildir

şeklindedir.

3.1.1.1. Küme Sayısı K'nın Belirlenmesi

K-means yönteminin uygulanması için *K*, küme sayısının önceden bilinmesi gerekmektedir. Ancak kümeleme analizi uygulamalarında genellikle verinin sınıf

yapısıyla ilgili herhangi bir önbilgi bulunmamaktadır. Bu nedenle çoğu durumda küme sayısının araştırmacı tarafından belirlenmesi gerekmektedir. K'nın belirlenmesi için kullanılan temel yöntem, küme içi uzaklığının (W_K), K'nın bir fonksiyonu olduğu varsayılır ve her bir küme sayısı $\mathbf{K}=\{K_1, K_2, \ldots, K_{maks}\}$ için küme içi uzaklıkları olan $\{W_1, W_2, \ldots, W_{maks}\}$ değerleri hesaplanır. Elde edilen küme içi uzaklıklar K arttıkça azalmaktadır. Görsel olarak küme sayısını belirlemek için, Şekil 3.1.(a)'daki iki boyutlu verinin küme sayısını belirlerken, Şekil 3.1.(b)'deki gibi W_K'nın K değerlerine karşı grafiği çizdirilir ve W_K-W_{K+1} farkının küçülmeye başladığı K=3 dirsek noktasını optimum küme sayımız olarak belirleriz (Hastie ve ark, 2009).



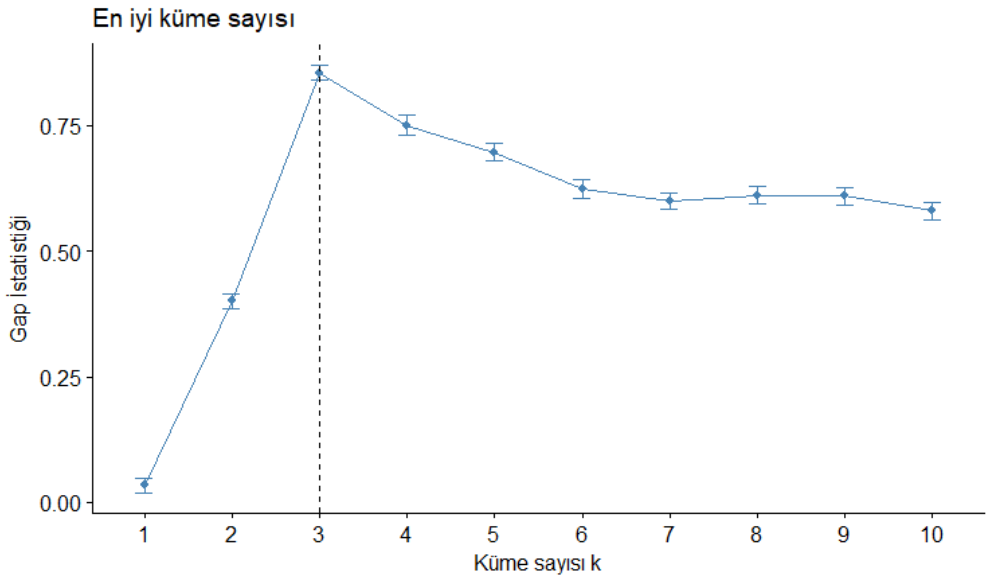
a) b)

Şekil 3.1. Üç küme bulunan 2 değişkenli bir veri kümesi için (a) saçılım grafiği (b) dirsek grafiği

Bir diğer popüler *K* belirleme yöntemi ise gap istatistiğidir (Tibshirani ve ark, 2001). Bu yaklaşımda $\log(W_K)$ ile $\log(W_K)$'nın uygun bir dağılım altındaki beklenen değeri karşılaştırılarak elde edilir. Gap istatistiği beklenen eğri ile gözlenen eğri arasındaki farktır ve,

$$\text{gap}(k) = E(\log(W_k)) - \log(W_k) \tag{3.4.}$$

şeklinde hesaplanarak elde edilir. $gap(k)$ değerini maksimum yapan *K* küme sayısı en iyi küme sayısı olarak belirlenir. Genellikle diğer yöntemler veride tek küme olması durumunda doğru sonucu belirleyememektedir fakat gap istatistiği verinin tek bir kümeden oluşması durumunda dahi doğru sonucu belirleyebilmektedir (Hastie ve ark, 2009).

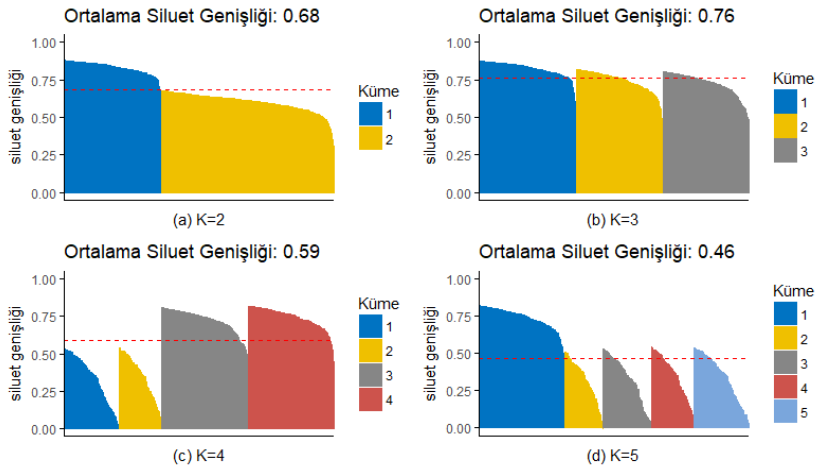


Şekil 3.2. Şekil 3.1.(a) verisi için gap istatistiği ile küme sayısının belirlenmesi

Kaufman ve Rousseeuw (1990), hem elde edilen kümelerin kalitesini belirlemek hem de *K* küme sayısını belirlemede kullanılan bir yöntem olan siluet istatistiğini

$$s(i) = \frac{b(i) - a(i)}{mak\{a(i), b(i)\}} \tag{3.5.}$$

şeklinde tanımlamışlardır. Burada $a(i)$, i-inci gözlemin bulunduğu kümedeki diğer gözlemler ile arasındaki ortalama uzaklık, $b(i)$ i-inci gözleme en yakın kümedeki (komşu küme) gözlemlere ortalama uzaklığını göstermektedir. $s(i)$ katsayısı -1 ile 1 arasında bir değer alacaktır. Eğer $s(i)$ 1'e yakınsa i gözlemi doğru kümelenmiş, 0'a yakınsa kararsız ve -1'e yakınsa komşu kümede olmasının daha uygun olduğu sonucuna varılmaktadır. Bu yöntemle küme sayısı belirlenirken olabilecek tüm küme sayıları için siluet grafiği çizdirilir ve kümeleri en uygun olan (ortalama siluet genişliği en büyük olan) küme sayısı gerçek küme sayısı olarak kabul edilir. Şekil 3.3.'te küme sayısının 3 olarak belirlendiği bir örnek verilmiştir.



Şekil 3.3. Siluet istatistiği ile Şekil 3.1 (a) verisi için küme sayısının belirlenmesi

3.1.1.2. Yerel Minimum Çözüm

K-means algoritmasında 2. ve 3. adımların tekrarlanması küme içi amaç fonksiyonunu sürekli olarak minimize eder. Fakat amaç fonksiyonunun konveks olmayan yapısı, çözümün çoğunlukla genel minimum yerine yerel minimum bir değer belirlenmesine neden olur. Bu sorunun üstesinden gelmek algoritma için doğru başlangıç merkezlerinin belirlenmesi ile mümkündür (Çelebi ve ark, 2012). Bu problem için en yaygın kullanılan sezgisel çözümde, K-means yöntemi faklı başlangıç merkezleri ile tekrarlı olarak veri uygulanır ve elde edilen tüm çözümler arasında kriter fonksiyonu minimum yapan çözüm genel minimum çözüm olarak kabul edilir. Bu yöntem özellikle büyük verilerde, yöntemin uygulanış süresini oldukça artırır. Bradley ve Fayyad (1998) ise başlangıç merkezlerinin belirlenmesi için bootstrap benzeri bir yöntem kullanmışlardır. Bu yöntemde gözlem sayısından daha az elemana sahip S gözlemli örneklem tekrar yerine koyularak seçilmekte ve her birine K-means yöntemi uygulanmaktadır.

Veri için çok sayıda yerel optimum çözüm olması K-means algoritması için başlangıç merkezlerinin belirlenmesini daha önemli kılmaktadır. Bu amaçla başlangıç merkezinin belirlenmesi için çeşitli deterministik yöntemler geliştirilmiştir. Örneğin Milligan (1980) hiyerarşik eklemeli kümeleme uygulayarak elde ettiği kümelerin merkez noktalarını K-means için başlangıç merkezleri olarak kullanmıştır.

3.1.1.3. K-means Algoritması Türevleri

K-medians: Küme merkezleri ortalama yerine medyan kullanılarak hesaplanan algoritmadır. Yöntemin K-means algoritmasından diğer bir farkı ise gözlemlerin küme

medyanına olan benzerliği öklid uzaklığı yerine manhattan uzaklığı ile belirlenmektedir. K-medians algoritması aşağıdaki kriter fonksiyonunu minimum yapmayı amaçlar.

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} \left| x_i - medyan_k \right| \tag{3.6.}$$

Kümelerin belirlenmesinde ortalama yerine medyan kullanımı yöntemin sapan değerlere karşı daha dayanaklı olmasını sağlamaktadır.

K-modes: K-means yönteminde küme merkezleri ortalama ile belirlendiğinden bu yöntem ile kategorik değişkenlere kümeleme uygulanamaz. Huang (1998) bu eksikliğin giderilmesi için K-modes algoritması önermiştir. Bu algoritma K-means farklı olarak kategorik gözlemler için basit eşleşme (simple matching) benzerlik ölçüsünü kullanmakta, kümelerin ortalama yerine modunu hesaplayarak kümeleri belirlemektedir.

Aşağıdaki kriter fonksiyonunu minimum yapmak K-modes algoritmasının amacıdır.

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \gamma_{ki} D(x_j, \mathrm{mod}_k) \tag{3.7.}$$

K-means gibi K-modes de çoğunlukla yerel optimum sonuç bulmaktadır bulunan sonucun kalitesi seçilen başlangıç merkezine bağlıdır.

X-means (Pelleg ve Moore, 2000): *K*, küme sayısını otomatik olarak belirleyerek K-means uygulayan bir algoritmadır. Bu yöntem için önbilgi olarak yalnızca küme sayısı K için bir aralık belirlemek gerekmektedir. Verilen aralıktaki tüm K değerleri için K-means algoritması veriye uygulanmaktadır ve her belirlenen kümeleme için modelin iyiliği AIC veya BIC kriteri ile değerlendirilerek en yüksek model skoru olan kümeleme en iyi sonuç olarak belirlenir.

K-means++ (Arthur ve Vassilvitskii, 2007): Sıradan K-means algoritmasından farklı olarak bu algoritma başlangıç merkezlerini seçmek için olasılığa dayalı bir prosedür kullanmaktadır. Bu yöntemde ilk küme merkezi rasgele olarak seçilir. İkinci küme merkezi ise ilk küme merkezine en uzak gözlem olarak belirlenir. Bu seçim işlemleri *K* adet centroid elde edilene kadar tekrarlanır. Bu elde edilen centroidlere K-means algoritması uygulanır. Bu sayede K-means'in yerel minimum çözüm bulmasının önüne geçilmesi sağlanır.

3.1.2. K-medoids

K-medoids algoritması, veri setinin özelliklerini temsil eden *K* tane temsilci nesne aramaktadır. Kümeleme analizinde bu temsilci nesnelere centrotype denilmektedir (Kaufman ve Rousseeuw, 2005). Temsilci gözlem, küme içerisindeki diğer gözlemlere olan ortalama uzaklığı minimum yapan kümenin merkezine en yakın gözlemdir. K-medoids algoritmasında, temsilci gözlemler herhangi bir gözlemle değiştirilir. Elemanların değişim maliyeti ile K-medoids algoritması için mutlak hata kriteri elde edilir. Her yeniden atama işlemi için bu değişim maliyeti hesaplanır ve toplam maliyet fonksiyonu minimize edilir (Reddy, 2013).

K-medoids algoritmasının adımları,

1. *K* tane başlangıç temsilci gözlemi seçilir.
2. Her gözlemi en yakın temsilci gözlemin kümesine atanır
3. Rasgele olarak temsilci gözlem olmayan bir x_i gözlemi seçilir
4. Temsilci gözlemin m'den x_i'ye değişiminin maliyetini hesaplanır
5. Maliyet sıfırdan küçükse m yerine x_i temsilci gözlem olarak alınır

6. Yakınsama gerçekleşene kadara 2, 3, 4 ve 5 adımları tekrarlanır

şeklindedir.

Kaufman ve Rousseeuw (2005), K-medoids yönteminin bir türevi olan PAM (Partitioing Aroun Medoids) algoritmasını geliştirmiştir. Uzaklık matrisi kullanarak kümeleme analizini uygulayan bu yöntem, minimum amaç fonksiyonuna yakınsayana kadar tüm medoid olmayan ve olan gözlemlerin iteratif olarak değiştirilmesiyle işlemektedir. K-means'e göre sapan değerlere karşı daha güçlüdür ve daha hızlı çalışmaktadır. Bu sebeple gözlem sayısı fazla olan verilerde tercih edilmektedir (Reddy, 2013).

3.1.3. Bulanık C-Means

Verideki gözlemlerin heterojen yapı göstermediği, kümeler arası gözlemlerin iç içe geçtiği durumlarda her bir gözlemi bir kümeye atanması yanlış yorumlamaya sebep olabilir. Bu durumlarda gözlemlerin hangi kümelerde olduğu sınıf etiketleri yerine olasılık değerleri ile belirtilmesi için C-ortalamalar algoritması kullanılmaktadır (Bezdek ve Ehrlic, 1984). C-ortalamalar yöntem için hata kareler ortalaması,

$$W_k = \sum_{k=1}^{K} \sum_{x_i \in C_k} w_{x_{ik}}^{\beta} \left(x_i - c_k \right)^2 \tag{3.8.}$$

$$w_{x_{ik}} = \frac{1}{\sum_{k=1}^{K} \left(\frac{x_j - c_k}{x_j - c_j} \right)^{\frac{2}{\beta - 1}}} \tag{3.9.}$$

$$c_k = \frac{\sum_{x_i \in C_k} w_{x_{ik}}^{\beta} x_i}{\sum_{x_i \in C_k} w_{x_{ik}}} \tag{3.10.}$$

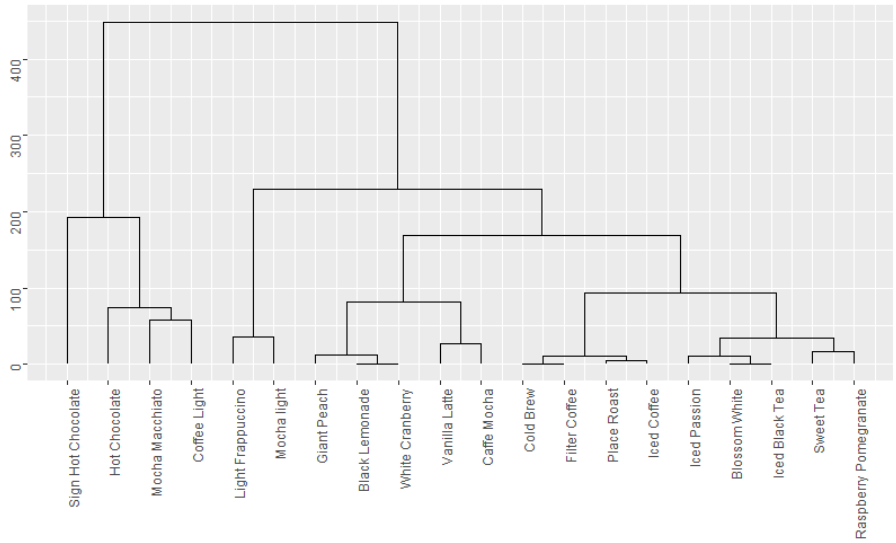
şeklinde verilir. Burada $w_{x_{ik}}$ değeri x_i 'nin c_k kümesine ait olma ağırlığıdır. Bu ağırlık C-ortalamalar algoritmasında kümeler güncellenirken kullanılmaktadır. c_k için bulanık ağırlıklar için c_k ağırlıklandırılmış merkezler hesaplanır. Algoritma K-means ile aynı mantıkla $w_{x_{ik}}$ ve c_k değerlerini güncelleyerek hata kareler toplamını minimum yapmaktadır.

3.2. Hiyerarşik Kümeleme Yöntemleri

Hiyerarşik kümeleme analizi, gözlemlerin birbiri arasında aşamalı bir biçimde bağlanarak kümeler oluşturulmasını sağlayan yöntemlerin genel adıdır. Verideki grup yapılanmasının hiyerarşik bir biçimde elde edilmesi verinin özetlenmesi ve yorumlanmasını oldukça kolaylaştırmaktadır. Bu yöntem sonucunda elde edilen iç içe küme yapıları dendogram (Şekil 3.4.) yardımıyla görselleştirilmektedir. Dendogram sayesinde küme sayısı ve veride herhangi bir sapan gözlem olup olmadığı görsel olarak belirlenebilmektir. Hiyerarşik kümeleme algoritmalarında kümeleme işlemine başlanması için öncelikle verideki gözlemler arası uzaklıkların ve kümeler arası uzaklıkların nasıl tanımlanacağı belirlenmelidir (Martinez ve ark, 2010). Bu iki faktörün değiştirilmesi elde edilen kümeleme sonuçlarını değiştirmektedir. Hiyerarşik yöntemlerin en büyük dezavantajı ise bir grup birleştirildiğinde veya ayrıştırıldığında bu adımın geriye alınamamasıdır. Kümelerin hiyerarşik olarak gösterilmesi, biyolojide

türler arasındaki evrimsel bağları gösteren ağaçlardan esinlenilmesi sebebiyledir (Henning ve Meila, 2016).

Hiyerarşik yöntemler birleştirici (agglomerative) ve ayrıştırıcı (divisive) olmak üzere iki ayrı alt başlıkta incelenmektedir. Birleştirici yöntemler, verideki her bir gözlemi bir küme olarak kabul ederek işleme başlar ve her adımda aralarındaki benzerliği en yüksek olan kümeler tek bir küme oluşturuluncaya kadar birleştirilir. Ayrıştırıcı yöntemlerde ise başlangıçta tüm veri bir küme olarak kabul edilir ve sonrasındaki her adımda gözlemler birbirinden ayrı kümeler oluşturuluncaya kadar en az benzerlik gösteren kümeler ayrıştırılır. Kümeleme analizi uygulamalarında, ayrıştırıcı yöntemlerin hesaplama karmaşıklığının fazla olması sebebiyle genellikle eklemeli yöntemler tercih edilir. Birleştirici hiyerarşik kümeleme algoritmaları kullanılan gruplar arası uzaklık ölçüleri ile isimlendirilmektedir.



Şekil 3.4. Starbucks'ta satılan bazı ürünler için dendogram grafiği

Hiyerarşik kümelemede veriyi belirli bir sayıda kümeye parçalama işlemi iki yolla gerçekleştirilir (Mirkin, 2011). Birinci çeşit yöntem verinin parçalanma veya birleşme işleminin belirli bir kriter ile durdurulmasıyla gerçekleştirilirken. Diğer yöntem tüm hiyerarşik yapı incelenip istenen küme sayısını veren bir kesim noktası belirlenerek gerçekleştirilir.

3.2.1. Birleştirici Hiyerarşik Kümeleme

Birleştirici hiyerarşik kümeleme algoritması aşağıdaki gibidir.

1. Tüm gözlemler bir küme olarak ele alınır.
2. Uzaklık/Benzerlik matrisi hesaplanır.
3. En benzer iki küme birleştirilir.
4. Tek bir küme elde edilene kadar 2. ve 3. adımlar tekrarlanır.

Birleştirici hiyerarşik kümelemenin uygulanabilmesi için birbirlerine en yakın olan kümelerin her adımda tespit edilmesi gerekmektedir. Kümeler arası yakınlıkların belirlenmesi için aşağıdaki kümeleme bağıntıları kullanılmaktadır.

Tekli Bağıntı (En Yakın Komşuluk):

Eklemeli kümelemede en sık kullanılan uzaklık bağıntısıdır. Bu uzaklığa göre iki küme arasındaki mesafe birbirlerine en yakın olan gözlemlerinin aralarındaki uzaklık kadardır. Bu sebeple tekil bağıntıya en yakın komşuluk bağıntısı da denmektedir ve aşağıdaki formdadır.

$$d(c_k, c_l) = \min\{d(x_{c_k i}, x_{c_l j})\} \qquad ,i = 1,2,...,n_{c_k} \; ; \; j = 1,2,...,n_{c_l} \qquad (3.11.)$$

Tekil bağıntı yöntemi zincirleme (chaining) denilen bir dezavantajı bulunmaktadır. Zincirleme problemi kümelerdeki gözlemler iyi ayrışmadıkları durumda ortaya çıkmaktadır. Hiyerarşik kümelemede bir zincir oluştuğu takdirde zincirin başındaki gözlem ile sonundaki gözlemler aynı kümede olmalarına rağmen birbirlerine benzememektedirler (Martinez ve ark, 2011). Tekil bağıntı yönteminin kullanımındaki bir diğer problem ise sapan gözlemlere karşı oldukça hassas olmasıdır.

Tam Bağıntı (En Uzak Komşuluk):

En uzak komşuluk bağıntısı olarak geçen bu yöntemde kümeler arası uzaklık, farklı kümelerdeki birbirine uzaklığı en fazla olan iki gözlemin uzaklığı kadardır. Bu uzaklık bağıntısı aşağıdaki gibi hesaplanabilir.

$$d(c_k, c_l) = \max\{d(x_{c_k i}, x_{c_l j})\},\ i = 1,2,...,n_{c_k};\ j = 1,2,...,n_{c_l} \qquad (3.12.)$$

Zincirleme problemi bu bağıntı kullanıldığında ortaya çıkmamaktadır. Ayrıca hiyerarşik kümeleme uygulanırken bu bağıntı seçildiğinde, oluşan küme yapısının küresel olduğu görülmektedir. Bu yüzden, bu yöntem farklı yapıdaki kümeleri belirlemekte zorlanmaktadır.

Ortalama Bağıntı:

Bu bağıntı yönteminde kümeler arası uzaklık,

$$d(c_k, c_l) = \frac{1}{n_{c_k} n_{c_l}} \sum_{1}^{n_{c_k}} \sum_{1}^{n_{c_l}} d(x_{c_k i}, x_{c_l j}) \qquad (3.13.)$$

şeklinde bir kümedeki tüm gözlemlerin diğer kümedeki tüm gözlemlere olan ortalama uzaklığı ile hesaplanır. Bu yöntemde varyansı küçük olan kümeler birleştirilmektedir ve

oluşan kümeler genellikle aynı varyanslıdır. Tekil bağıntı ve tam bağıntı yöntemlerine göre sapan değerlere karşı daha dayanıklıdır.

Ward Bağıntısı:

Ward (1963), tarafından geliştirilen bu yöntem, K-means algoritmasının kriter fonksiyonu olan hata kareler toplamı (3.1.) kullanılarak uygulanmaktadır. Bu yöntem, iki küme arasındaki hata kareler toplamını en az artıran kümeleri birbirleri ile birleştirmektedir (Murtagh ve Legendre, 2014).

$$\sum_{k=1}^{K} \sum_{x_i \in C_k} \left(x_i - m_k \right)^2 \tag{3.14.}$$

Ward yöntemi K-means ile benzer olarak veriyi küresel kümelere ayrıştırma eğilimindedir ve sapan gözlemlere karşı duyarlıdır.

3.2.2. Ayrıştırıcı Hiyerarşik Kümeleme

Ayrıştırıcı kümelemede başlangıçta tüm veriyi bir küme kabul eder ve her adımda kümeleri birbiri ile en uzak kümeleri birbirinden ayırmaya çalışırız. *n* gözlemi olan bir veriyi iki farklı boş olmayan kümeye ayırma işlemi 2^{n-1} farklı yolda yapılabilmektedir (Xu ve Wunsch, 2008). Bu durumda ayrıştırıcı yöntemler fazla gözleme sahip verilerde uygulanması durumunda hesaplanama süresi oldukça artacaktır. Algoritmayı hızlandırmak için genellikle veri belirli bir düzeyde ayrıştırıldıktan sonra işlemler sonlandırılmaktadır.

Ayrıştırıcı yöntemde kümeleme performansını etkileyen dört faktör Reddy (2013), tarafından aşağıdaki gibi verilmiştir.

1. Ayrıştırma kriteri

2. Ayrıştırma yöntemi
3. Ayrıştırılacak kümenin belirlenmesi
4. Sapan gözlemlerle başa çıkma

Ayrıştırıcı hiyerarşik kümeleme algoritması aşağıdaki adımlardan oluşmaktadır.

1. Tüm gözlemler tek bir küme olarak ele alınır.
2. Ağaçtaki bir önceki küme iki kümeye parçalanır.
3. Elde edilen kümelerin hata kareler toplamı hesaplanır.
4. Hata kareler toplamı en yüksek olan kümede 2.adım ve 3. adım tekrarlanır.
5. Tüm gözlemler birer küme oluşturuncaya kadar adımlar tekrarlanır.

3.3. Yoğunluk Tabanlı Kümeleme Analizi

Bu bölüme kadar işlenen kümeleme analizi yöntemleri veriyi parçalarken küresel şekiller oluşturma eğilimi göstermekteydi. Küresel olmayan kümeleri belirleme ihtiyacı, dünyadaki bir bölgeyi iki veya üç boyutlu olarak ifade eden mekânsal verilerde özellikle karşımıza çıkmaktadır. Mekânsal verilerin elde edilme tekniklerinin gelişmesiyle büyük miktarda verilerin analizi ihtiyacı ortaya çıkmıştır. Bu sebeple hem küresel olmayan yapıdaki kümelerin belirlenmesi hem de büyük verilerde etkin bir biçimde uygulanabilen yeni yöntemler geliştirilmeye başlanmıştır. Bu yöntemlerden bir kısmı verideki gözlemlerin yoğun bulundukları bölgelere odaklanarak verinin kümelerin belirlemeyi amaçlamaktadır. Bu yöntemlere yoğunluk tabanlı yöntemler denilmektedir.

Yoğunluk tabanlı yöntemler verinin dağılımı ya da küme sayısı ile ilgili herhangi bir varsayımda bulunmamaktadır. Bu sebeple yoğunluk tabanlı yöntemler

parametrik olmayan yöntemlerdir. Buna rağmen yoğunluk tabanlı yöntemlerin uygulanabilmesi için cevaplanması gereken üç kritik soru vardır (Ester ve ark, 1996).

1. Yoğunluk tahmini nasıl gerçekleşecek?
2. Gözlemlerin bağlantısı nasıl tanımlanacak?
3. Hangi veri yapısı algoritmanın etkin uygulamasını desteklemektedir.

Yoğunluk temelli her algoritma bu sorulara farklı yanıtlar vermektedir. Bu bölümde DBSCAN, OPTİCS ve DENCLUE yoğunluk temelli algoritmalardan bazılarıdır.

3.3.1. DBSCAN Algoritması

Karmaşık şekillere sahip kümeleri ayrıştırmak için geliştirilmiş bir algoritmadır. Yoğunluk tabanlı bir yöntem olduğu için küme sayısının önceden bilinmesine gerek yoktur. Bu algoritmada yoğunluk tahmini her bir gözlemin sabit bir yarı çap (ε) uzaklığı içerisindeki komşu gözlemler sayılarak gerçekleştirilir. Gözlem etrafındaki komşu gözlemlerin sayısı belirli bir *minpts* eşik değerini geçmesi beklenir. Bu eşik değerinin aşılması durumunda gözlem, merkez nokta olarak adlandırılır ve merkez noktanın etrafındaki bölge yoğun bölge adını alır. Algoritmanın daha iyi anlaşılması için üç kavramın tanımlanması gerekir.

Direkt yoğunluk erişilebilir: Bir *A* merkez noktasının komşusu olan *B* gözlemi *A*'dan direkt yoğunluk erişilebilirdir.

Yoğunluk Erişilebilir: Bir *A* noktası ile *B* noktası arasında çeşitli merkez noktalarla bağlantı kurulabiliyorsa *B* gözlemi *A* için yoğunluk erişilebilirdir.

Yoğunluk Bağlantılı: *A* ve *B* noktalarının her ikisi de bir *C* merkez noktasına yoğunluk erişilebilir ise *A* ve *B* noktaları yoğunluk bağlantılıdır.

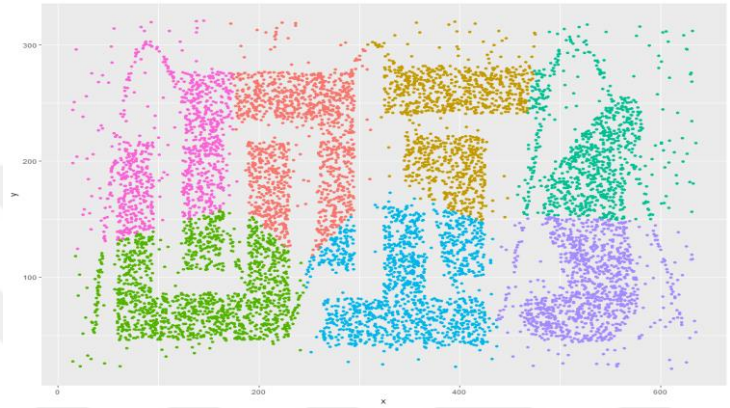
DBSCAN algoritmasının amacı, yoğunluk erişilebilirliği maksimum olan yoğunluk bağlantılı gözlem gruplarını belirlemektir. Bu belirlenen grupların birleşimi ise bir küme oluşturmaktadır. Bu süreç içerisinde hiçbir kümenin elemanı olmayan gözlemler ise sapan gözlem olarak belirlenmektedir (Ester ve ark, 1996).

1. Her bir gözlemin diğer gözlemlerle olan uzaklığı hesaplanır.
2. Bir başlangıç gözleminin tüm ε sınırı içerisindeki komşuları belirlenir ve komşu sayısı *minpts* eşik değerinden fazla olan tüm gözlemler merkez nokta olarak belirlenir.
3. (2) adımı tüm gözlemlerde tekrarlanır.
4. Herhangi bir kümeye ait olmayan gözlemler sapan değer olarak belirlenir.

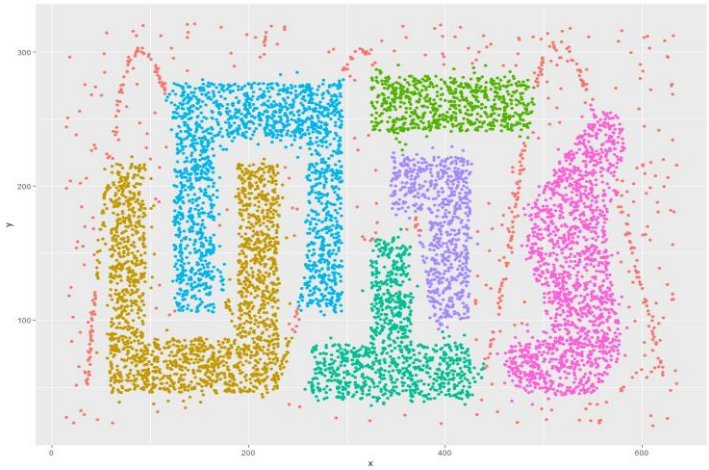
Aşamaları yukarıda özetlenen DBSCAN algoritmasında belirlenmesi gereken ve elde edilen kümeleri etkileyen iki önemli parametre vardır. Bunlar yarıçap (ε) ve eşik değerdir (*minpts*). Bu parametrelerin belirlenişi için k-uzaklık adı verilen sezgisel bir yöntemden faydalanılmaktadır. *k* değeri bir gözlemin en yakın *k*. komşusu ile arasındaki uzaklığı göstermektedir. Bu yöntemde tüm gözlemlerin birbirleri arasındaki uzaklıklar ve gözlemlerin *k*-ıncı en yakın komşuları hesaplanır. Bu bulanan değerler küçükten büyüğe doğru sıralanarak grafik üzerinde gösterilir ve oluşan dirsek noktası belirlenerek doğru yarıçap ve *k* değeri elde edilir.

DBSCAN algoritmasının diğer yöntemlere göre en önemli avantajı bağıntılı şekilleri bulunan ve karmaşık yapıdaki kümeleri oldukça iyi belirleyebilmektedir. Şekil 3.5.'te böyle bir yapıya sahip olan bir veriye K-means uygulanmış ve verinin grup yapısını doğru belirleyememiş ve kümeleri dairesel olarak belirlemeye çalışmıştır. Şekil 3.6.'da aynı veriye DBSCAN uygulanmış ve karmaşık verideki gruplar doğru

belirlenmiştir. Ayrıca şekil 3.6.'da herhangi bir kümeye ait olmayan sapan gözlemler bulunmaktadır. DBSCAN algoritması bu gözlemleri de doğru bir biçimde belirlemiştir.



Şekil 3.5. Karmaşık yapıdaki bir verinin K-means algoritması ile elde edilen küme yapısı (Karypis ve ark, 1999)



Şekil 3.6. Karmaşık yapıdaki bir verinin DBSCAN algoritması ile elde edilen küme yapısı (Karypis ve ark, 1999)

4. MODELE DAYALI KÜMELEME ANALİZİ

Modele dayalı kümeleme analizi, sonlu karma dağılımları kullanarak kümeleme analizine temel istatistiksel çerçevede çözüm arayan yöntemlerin genel adıdır. Bu yöntemlere aynı zamanda sonlu karma dağılımlar üzerinden kümeleme analizi de denilmektedir (Everit ve ark, 2011; Fraley ve Raftery, 2002). Bu yaklaşımda verideki her bir gözlemin bir karma dağılım bileşeninden geldiği ve her bir karma bileşeninin bir küme gösterdiği varsayılmaktadır (Fraley ve Raftery, 1998). Bu sebeple verinin küme yapısını belirlemek için istatistiksel model uydurma yöntemlerinden yararlanmak gerekmektedir.

Hem hiyerarşik yöntemler hem de K-means gibi sezgisel kümeleme yöntemlerinde sınıf yapıları ile ilgili herhangi bir varsayımda bulunulmamaktadır ve kümeleme adımlarının belirlenmesi, geleneksel, sonuca dayalı yaklaşımlar kullanılmaktadır. Sezgisel kümeleme yöntemlerinde sonuçları etkileyen uzaklık ölçüsünün ve küme sayısının belirlenmesi probleminde genellikle öznel kararlar verilir. Modele dayalı yöntemlerde ise veride küme sayısının belirlenmesi ve sapan gözlemlerle başa çıkmak için formal ve istatistiksel yöntemlere dayalı çözümler sunar (Fraley ve Raftery, 2002).

4.1. Sonlu Karma Dağılım

İstatistiksel modeller verideki tüm gözlemlerin aynı dağılımdan geldiğini varsaymaktadır fakat bu varsayım her zaman doğru değildir (Zhang ve Huang, 2015). Örneğin örneklemler birbirinden farklı gruplardan toplanmış olabilir. Bu gibi durumlarda veride homojenlik sağlanmayacağı için bu verinin klasik yöntemlerle modellenmesi doğru olmaz. Bu durumda modelleme işleminin sonlu karma modellerle gerçekletirilir. Sonlu karma dağılımların genel formu aşağıdaki gibidir.

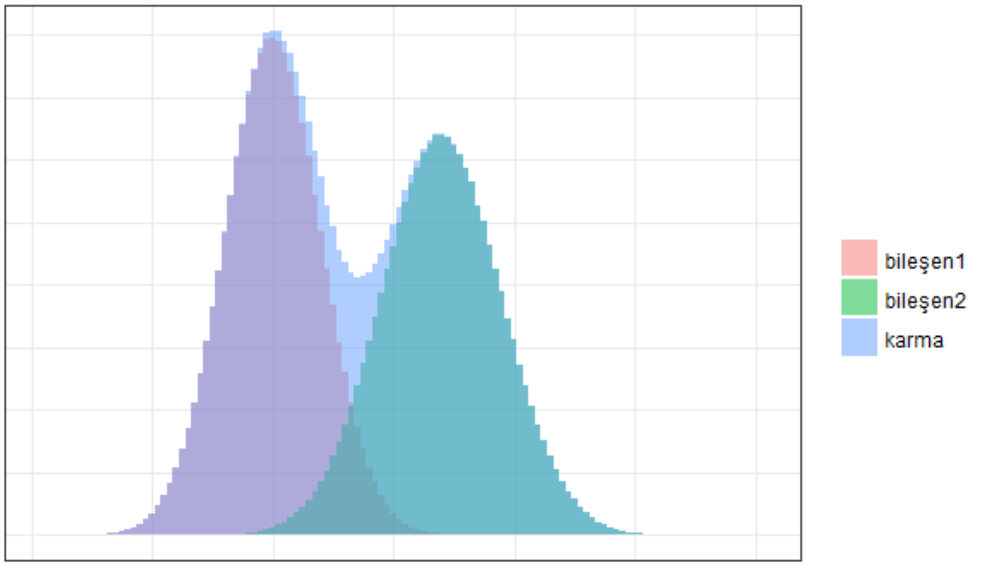
$$f(x;\pi) = \sum_{k=1}^{K} \pi_k f_k(x) \tag{4.1.}$$

Burada π_k karma oranı, [0,1] aralığında ve toplamları $\sum_{k=1}^{K} \pi_k = 1$ şeklindedir ve bir gözlemin *k*'ıncı karma bileşeninden gelme olasılığını temsil etmektedir. $f_k(x)$ ise *k*'ıncı bileşeninin yoğunluk fonksiyonunu göstermektedir. $f_k(x)$ bileşen yoğunluk fonksiyonu uygulamalarda genellikle, θ_k parametre vektörü ile $f_k(x;\theta_k)$ parametrik formunda gösterilir. Bu sayede (4.1.) eşitliği aşağıdaki şekilde yeniden düzenlenir.

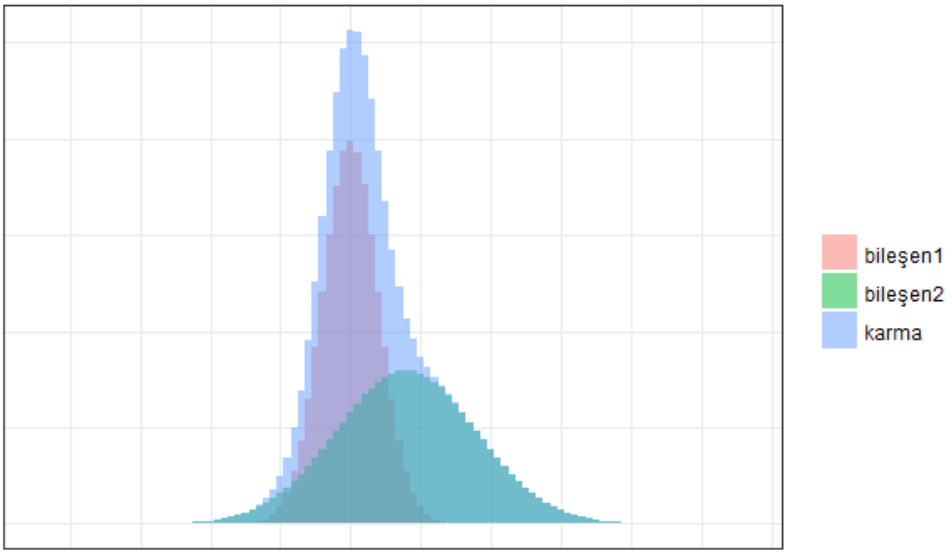
$$f(x;\Theta) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k) \tag{4.2.}$$

(4.2.) eşitliğinde Θ sonlu karma modeldeki tüm parametreleri içeren bir uzaydır ve $\Theta = \{\pi_1, \pi_2, ..., \pi_{K-1}; \theta_1, \theta_2, ..., \theta_K\}$ şeklinde bileşen dağılımlarının karma oranlarını ve parametrelerini içerir. (4.2.)'de verilen karma modelin bileşenlerinin yoğunluk fonksiyonları $f_k(x;\theta_k)$ farklı dağılımdan olabileceği gibi farklı parametrelere sahip aynı dağılımlı olabilir. Karma modellerde en sık kullanılan dağılım yapısı tüm bileşenlerin farklı parametrelere sahip olan normal dağılımlardan oluştuğu durumdur. Verinin uzun kuyruklu bir yapıda olması durumunda verinin normal dağılım karmaları ile modellenmesi sonucun yanlı çıkmasına sebep olabilmektedir (Zuang ve Huang, 2015). Bu tip durumlarda normal dağılım yerine genellikle t-dağılımı karmalarından faydalanılmaktadır. Ayrıca veri yapısının asimetrik olduğu durumlarda normal ve t-dağılımı gibi simetrik dağılımlar yetersiz kalmaktadır. Bu sorunun üstesinden gelebilmek için çarpık normal ve çarpık t-dağılımının karmalarından yararlanılmaktadır (Frühwirth-Schnatter ve Pyne, 2010).

Bir sonlu karma model dağılımın yoğunluk şekilleri oldukça esnektir. Örneğin dağılımın yoğunluğu çok modlu (Şekil 4.1.) ya da tek modlu ama önemli bir çarpıklığa sahip olabilir (Şekil 4.2.). Bu sebeple sonlu karma modeller dağılım yapıları belirgin olmayan verilerin modellenmesinde kullanılmaktadır (Frühwirth-Schnatter, 2006). Ayrıca sonlu karma dağılımlar istatistikte model kitle heterojenliğinin belirlenmesinde, genelleştirilmiş dağılım varsayımlarında ve son olarak ise kümeleme ve sınıflama analizinde kullanılmaktadır.



Şekil 4.1. Çift modlu sonlu karma dağılım



Şekil 4.2. Sağa çarpık sonlu karma dağılım

4.1.1. Karma Dağılımların Maksimum-Olabilirlik Tahmini:

$f(x;\theta)$, θ parametresine bağlı olasılık yoğunluk fonksiyonu ve $\mathbf{X}=\{x_1, x_2, \ldots, x_n\}$ rasgele örneklemi için olabilirlik fonksiyonu,

$$L(\theta \mid \mathbf{X}) = f(\mathbf{X} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{4.3.}$$

formülü ile elde edilir.

Maksimum olabilirlik tahmini, $L(\theta \mid \mathbf{X})$ olabilirlik fonksiyonunu maksimum yapan θ değerinin belirlenmesiyle elde edilir. Yani θ'nın maksimum olabilirlik tahmini $\hat{\theta} = \max_{\hat{\theta}}[L(\theta \mid \mathbf{X}), \theta]$ dır. Maksimum olabilirlik tahmini yapılırken işlem kolaylığı sağladığı için $L(\theta \mid \mathbf{X})$ yerine genellikle doğal logaritması $\ln\{L(\theta \mid \mathbf{X})\}$ tercih edilmektedir.

Bu maksimizasyon probleminin çözüm zorluğu $f(x_i;\theta)$ olasılık yoğunluk fonksiyonunun yapısına bağlıdır (Bilmes, 1998). Örneğin $f(x_i;\theta)$ parametreleri $\theta=(\mu,\sigma^2)$ olan Normal dağılım olasılık yoğunluk fonksiyonu olduğunda maksimum olabilirlik tahmini $\ln L(\theta\,|\,\mathbf{X})$ fonksiyonunun μ ve σ^2 parametrelerine göre türevleri alınarak kolaylıkla hesaplanabilir. Fakat dağılım yapısı çok daha karmaşık olan (4.2.) sonlu karma dağılımının parametrelerinin maksimum olabilirlik tahminlerini elde etmek bundan daha zor olacaktır.

Bir $x_1, x_2, ..., x_n$ rasgele örnekleminin $f(x_i;\Theta)$ olasılık yoğunluk fonksiyonu (4.2.)'deki karma dağılım modeline sahip olsun. Bu rasgele örneklem için olabilirlik fonksiyonu aşağıdaki gibidir.

$$L(\Theta\,|\,\mathbf{X}) = f(\mathbf{X}\,|\,\Theta) = \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k f_k(x_i\,|\,\theta_k) \tag{4.4.}$$

Bu fonksiyonun logaritmasının alınması ile elde edilen olabilirlik fonksiyonu ise,

$$\log(L(\Theta\,|\,\mathbf{X})) = \sum_{i=1}^{n}\log(\sum_{k=1}^{K}\pi_k f_k(x_i\,|\,\theta_k)) \tag{4.5.}$$

şeklinde ifade edilir. Sonlu karma modellerin maksimum olabilirlik fonksiyonunu, doğrusal olmayan yapısı sebebiyle klasik yöntemlerle maksimize etmek oldukça zordur. Bu nedenle $\log(L(\Theta))$ fonksiyonunu maksimum yapma işlemi doğrusal yöntemlerle gerçekleştirilememektedir (Everitt ve ark, 2011). Bu sorunla başa çıkmanın en popüler yolu en çok olabilirlik tahminine iteratif işlemlerle yakınsayan yöntemler kullanmaktır (Xu ve Wunsch, 2008). Bu yöntemler arasında en sık kullanılanı ise EM yani Expectation-Maximization (Dempster ve Laird, 1973) algoritmasıdır. EM algoritmasına alternatif olarak ise Bayesci tahmin metotları olan Gibbs Örneklemesi ya da diğer

Markov Zinciri Monte Carlo yöntemlerinin kullanımı günümüzde oldukça yaygındır (Everit ve ark, 2011).

4.1.2. EM Algoritması

EM algoritması veri kümesinin, tamamlanmamış yapıda olduğu ya da kayıp gözlem içerdiği durumlarında dağılım parametrelerinin maksimum olabilirlik tahminlerini elde etmek için kullanılan bir yöntemdir. İteratif şekilde işleyen ve tepe tırmanma mantığıyla fonksiyonun yerel maksimumunu belirleyen bir algoritmadır. EM algoritması E (expectation) adımı ve M (maksimization) adımı olmak üzere iki adımdan oluşmaktadır. Bu adımlar bir döngü halinde olabilirlik fonksiyonu bir maksimuma yakınsayıncaya kadar devam ettirilir.

X gözlenen veri matrisi ile gözlemin hangi bileşen yoğunluk fonksiyonundan üretildiğini bize gösteren $\mathbf{Z} = (z_1, z_2, ...z_n)$, $z_i = 1, 2, ..., K$ olmak üzere gözlenmemiş veri vektörüdür. **X** ve **Z**'nin birleştirilmesiyle **Y**=(**X**,**Z**) tamamlanmış veri matrisini oluşturur. Tamamlanmış verinin kullanımıyla (4.5.) modelindeki olabilirlik fonksiyonu EM ile maksimumu belirlenebilir hale getirilir (Melnykov ve Maitra, 2010). Bu **Z** vektörünün bilinmesi durumunda (4.5.) olabilirlik fonksiyonu,

$$\log(L(\Theta \mid \mathbf{Y})) = \log(f(\mathbf{X}, \mathbf{Z} \mid \Theta))$$

$$= \log(\prod_{i=1}^{n} f(x_i \mid z_i) f(z_i)) = \sum_{i=1}^{n} \log(\pi_{z_i} f_{z_i}(x_i \mid \theta_{z_i})) \qquad (4.6.)$$

şeklini almaktadır. Fakat kümeleme analizi uygulamalarında **Z** vektörü bilinmemektedir. Bu nedenle işleme devam edilebilmesi için **Z** genellikle rasgele bir vektör olarak ele alınması gerekir.

Gözlemlenmemiş **Z** vektörünü belirlemek için ilk olarak bilinmeyen karma dağılım parametrelerinin başlangıç değerleri belirlenmelidir. $\Theta^g = \{\pi_1^g, \pi_2^g, ..., \pi_{K-1}^g; \theta_1^g, \theta_2^g, ..., \theta_K^g\}$ parametre uzayı, olabilirlik fonksiyonunda kullanılabilmektedir. Bu parametreler verilmişken her i-inci gözlem ve k-ıncı karma bileşeni için bayes teoremi yardımıyla aşağıdaki şekilde hesaplanır.

$$f_k(z_i = k \mid x_i, \Theta^g) = \frac{\pi_k^g f(x_i \mid z_i = k, \theta_k^g)}{f(x_i \mid \Theta^g)} = \frac{\pi_k^g f(x_i \mid z_i = k, \theta_k^g)}{\sum_{k=1}^{K} \pi_k^g f_k(x_i \mid \theta_k^g)} \tag{4.7.}$$

Bu eşitlikte karma oranını gösteren π_k parametresi her bir karma bileşen için bir önsel olasılık olarak düşünülmektedir. $\mathbf{Z} = (z_1, z_2, ..., z_n)$ gözlenmemiş verisinden seçilen gözlemler olmak üzere

$$f_k(\mathbf{Z} \mid \mathbf{X}, \Theta^g) = \prod_{i=1}^{n} f(z_i \mid x_i, \Theta_k^g) \tag{4.8.}$$

formülü ile elde edilir.

Υ, **Z** değerlerinin uzayı ve $\delta_{k,z_i} = \begin{cases} 0, & z_i \neq k \\ 1, & z_i = k \end{cases}$ olmak üzere tamamlanmış veri log-olabilirlik fonksiyonunun beklenen değeri $Q(\Theta, \Theta^g)$'nın hesaplanışı Bilmes (1998), tarafından aşağıdaki şekilde verilmiştir.

$$Q(\Theta, \Theta^g) = E\left[\log f(\mathbf{X}, \mathbf{Z} \mid \Theta) \mid X, \Theta^{g-1}\right]$$

$$= \sum_{z \in \Upsilon} \log(L(\Theta \mid \mathbf{X}, z)) f(z \mid \mathbf{X}, \Theta^g)$$

$$= \sum_{z \in \Upsilon} \sum_{i=1}^{n} \log(\pi_{z_i} f_{z_i}(x_i \mid \vartheta_{z_i})) \prod_{j=1}^{n} f(z_j \mid x_j, \Theta^g)$$

$$= \sum_{z_1=1}^{K}\sum_{z_2=1}^{K}\dots\sum_{z_n=1}^{K}\sum_{i=1}^{n}\log(\pi_{z_i} f_{z_i}(x_i \mid \vartheta_{z_i}))\prod_{j=1}^{n} f(z_j \mid x_j, \Theta^g)$$

$$= \sum_{z_1=1}^{K}\sum_{z_2=1}^{K}\dots\sum_{z_n=1}^{K}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{k,z_i}\log(\pi_k f_k(x_i \mid \vartheta_k))\prod_{j=1}^{n} f(z_j \mid x_j, \Theta^g)$$

$$= \sum_{k=1}^{K}\sum_{i=1}^{n}\log(\pi_k f_k(x_k \mid \vartheta_k))\left[\sum_{z_1=1}^{K}\sum_{z_2=1}^{K}\dots\sum_{z_n=1}^{K}\delta_{k,z_i}\prod_{j=1}^{n} f(z_j \mid x_j, \Theta^g)\right] \qquad (4.9.)$$

(4.9.) eşitliğinde köşeli parantez içerisindeki kısım;

$$\sum_{z_1=1}^{K}\sum_{z_2=1}^{K}\dots\sum_{z_n=1}^{K}\delta_{k,z_i}\prod_{j=1}^{n} f(z_j \mid x_j, \Theta^g)$$

$$= \left(\sum_{z_1=1}^{K}\sum_{z_2=1}^{K}\dots\sum_{z_{i-1}=1}^{K}\sum_{z_{i+1}=1}^{K}\dots\sum_{z_n=1}^{K}\delta_{k,z_i}\prod_{j=1 j\neq i}^{n} f(z_j \mid x_j, \Theta^g)\right) f(k \mid x_i, \Theta^g)$$

$$= \prod_{j=1 j\neq i}^{n}\left(\sum_{z_j=1}^{K} f(z_j \mid x_j, \Theta^g)\right) f(k \mid x_i, \Theta^g)$$

$$= f(k \mid x_i, \Theta^g)$$

şeklinde sadeleştirilirse,

$$Q(\Theta, \Theta^g) = \sum_{k=1}^{K}\sum_{i=1}^{n}\log(\pi_k f_k(x_i \mid \vartheta_k)) f(k \mid x_i, \Theta^g)$$

$$= \left[\sum_{k=1}^{K}\sum_{i=1}^{n}\log(\pi_k) f(k \mid x_i, \Theta^g)\right] + \left[\sum_{k=1}^{K}\sum_{i=1}^{n}\log(f_k(x_i \mid \vartheta_k)) f(k \mid x_i, \Theta^g)\right] \qquad (4.10.)$$

eşitliği elde edilir. (4.10.) eşitliğini maksimize edebilmek için birinci ve ikinci parantez içerisindeki kısımlar ayrı ayrı maksimize edilmelidir. π_k değerini içeren birinci parantez içerisindeki formülün λ , lagrange çarpanı aşağıdaki gibi kullanılarak,

$$\frac{\partial}{\partial \pi_k}\left[\sum_{k=1}^{K}\sum_{i=1}^{n}\log(\pi_k) f(k \mid x_i, \Theta^g) + \lambda(\sum_k \pi_k - 1)\right] = 0$$

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} f(k \mid x_i, \Theta^g) \qquad (4.11.)$$

karma oranı için tahmin edici elde edilir.

θ_k parametreleri için yapılacak tahminler ise bileşenlerin olasılık yoğunluk fonksiyonlarına göre değişiklik göstermektedir. Modele dayalı kümeleme analizi için literatürde en sık kullanılan bileşen dağılımları çok değişkenli normal dağılımdır.

4.1.3. Çok Değişkenli Normal Dağılım Karmasının EM Algoritması ile Parametre Tahmini

Çok değişkenli normal dağılım karmaları, (4.2.) eşitliğindeki $f_k(x,\theta_k)$ bileşen fonksiyonları μ_k ortalamalı ve Σ_k varyans-kovaryans matris ile,

$$f_k(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_k\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)} \qquad (4.12.)$$

normal dağılmış olsunlar. Bu durumda (4.2.) eşitliği ,

$$f(x;\Theta) = \sum_{k=1}^{K} \pi_k f_k(x;\mu_k,\Sigma_k) \qquad (4.13.)$$

halini alır. Bu olasılık yoğunluk fonksiyonun logaritması alınıp (4.10.) eşitliğinin sağ tarafında yerine yazılırsa;

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\log(f_k(x_i \mid \mu_k, \Sigma_k)) f(k \mid x_i, \Theta^g)$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} \left(-\frac{p}{2} \log(2\pi) + \frac{-1}{2} \log(|\Sigma_k|) - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right) f(k \mid x_i, \Theta^g)$$

elde edilir. Bu fonksiyonun parametre tahmin edicilerinin belirlenmesi için matris cebir özelliklerinin kullanımı gerekir. Bunlardan ilki $iz(\mathbf{A})$, bir **A** kare matrisinin köşegen elemanlarının toplamını göstermektedir. $iz(\mathbf{A} + \mathbf{B}) = iz(\mathbf{A}) + iz(\mathbf{B})$ dır. Ayrıca $|\mathbf{A}|$, **A** matrisinin determinantını göstermektedir. Bir **A** matrisinin determinantının logaritmasının türevi,

$$\frac{\mathrm{dlog}|\mathbf{A}|}{da_{ij}} = 2\mathbf{A}^{-1} - diag(\mathbf{A}^{-1})$$

ile hesaplanmaktadır. Son olarak **A** ve **B** matrislerin çarpımının izinin **A** matrisine göre türevi

$$\frac{\mathrm{iz}(\mathbf{AB})}{dA} = 2\mathbf{B} + \mathrm{B}^T - diag(\mathbf{B})$$

formülü ile hesaplanır.

(4.14.) eşitliğinin μ 'ye göre türev alınıp aşağıdaki gibi 0'a eşitlenir.

$$\sum_{i=1}^{n} \Sigma_k^{-1} (x - \mu_k)^T \, f(k \mid x_i, \Theta^g) = 0 \qquad (4.15.)$$

Bu eşitlik çözüldüğünde μ 'nün tahmin edicisi,

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} x_i f(k \mid x_i, \Theta^g)}{\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)} \qquad (4.16.)$$

elde edilir. Σ 'nın tahmin edicisi için öncelikle (4.13.) eşitliği aşağıdaki gibi yazılır.

$$= \sum_{k=1}^{K}\left[\frac{1}{2}\log(\left|\Sigma_k^{-1}\right|)\sum_{i=1}^{n} f(k \mid x_i, \Theta^g) - \frac{1}{2}\sum_{i=1}^{n} f(k \mid x_i, \Theta^g) iz\left[\Sigma_k^{-1}(x-\mu_k)(x-\mu_k)^T\right]\right]$$

$$= \sum_{k=1}^{K}\left[\frac{1}{2}\log(\left|\Sigma_k^{-1}\right|)\sum_{i=1}^{n} f(k \mid x_i, \Theta^g) - \frac{1}{2}\sum_{i=1}^{n} f(k \mid x_i, \Theta^g) iz\left[\Sigma_k^{-1} N_{k,l}\right]\right] \qquad (4.17.)$$

Burada $N_{k,i} = (x-\mu_k)(x-\mu_k)^T$ dir ve (4.17.) eşitliğinin Σ_k^{-1}'e göre türev alınmasıyla,

$$\frac{1}{2}\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)(2\Sigma_l^{-1} - diag(\Sigma_l^{-1})) - \frac{1}{2}\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)(2N_{l,i} - diag(N_{l,i}))$$

$$= \frac{1}{2}\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)(2M_{k,i} - diag(M_{k,i}))$$

$$= 2S - diag(S) \qquad (4.18.)$$

elde edilir. Burada $M_{k,i} = \Sigma_k - N_{k,1}$ ve $S = \frac{1}{2}\sum_{i=1}^{n} f(k \mid x_i, \Theta^g) M_{k,i}$ dir. (4.18.) eşitliğinin 0'a eşitlenmesiyle

$$2S - diag(S) = 0$$

$$\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)(\Sigma_k - N_{k,i}) = 0 \qquad (4.19.)$$

elde edilir. (4.19.)'dan Σ_k'nın tahmin edicisi,

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} f(k \mid x_i, \Theta^g) N_{k,i}}{\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)}$$

$$= \frac{\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)} \tag{4.20.}$$

formülü elde edilir.

Bu sayede (4.13.) modeli için EM algoritması ile elde edilen parametre tahminleri aşağıdaki gibi verilebilir.

$$\hat{\pi}_k^{g+1} = \frac{1}{n} \sum_{i=1}^{n} f(k \mid x_i, \Theta^g) \tag{4.21.}$$

$$\hat{\mu}_k^{g+1} = \frac{\sum_{i=1}^{n} x_i f(k \mid x_i, \Theta^g)}{\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)} \tag{4.22.}$$

$$\Sigma_k^{g+1} = \frac{\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)(x_i - \mu_k^g)(x_i - \mu_k^g)^T}{\sum_{i=1}^{n} f(k \mid x_i, \Theta^g)} \tag{4.23.}$$

Normal dağılım karması için EM algoritması aşağıdaki adımlardan oluşmaktadır.

1. Karma dağılımdaki bileşen sayısı belirlenir.
2. Bileşen parametreleri için başlangıç tahminleri $\hat{\Theta}^0 = \{\hat{\pi}_1^0, \hat{\pi}_2^0, ..., \hat{\pi}_{K-1}^0; \hat{\mu}_1^0, \hat{\mu}_2^0, ..., \hat{\mu}_K^0, \hat{\Sigma}_1^0, \hat{\Sigma}_2^0, ..., \hat{\Sigma}_K^0\}$ belirlenir.
3. (4.21.), (4.22.) ve (4.23.) eşitlikleri kullanılarak parametre tahminleri bir değere yakınsayana kadar güncellenir. Bu adım aynı zamanda hem E hem de M adımını içermektedir (Bilmes, 1998).

4. Her gözlemin, hangi bileşen fonksiyonundan geldiğini gösteren sonsal olasılıklar (4.7.) eşitliği ile hesaplanır.

4.2. Modele Dayalı Kümeleme Analizi

Modele dayalı kümeleme analizi, kümeleme problemine sonlu karma modeller kullanarak çözüm arayan yöntemlerin genel adıdır. Bu yöntemlerde veri kümesinin sonlu karma dağılıma sahip olmakta ve her bir bileşen olasılık yoğunluk fonksiyonunun farklı bir kümeyi temsil ettiği varsayılmaktadır. Modele dayalı kümeleme analizi sonlu karma modellemeden türetilmiş olsa da iki yöntemin amaçları oldukça farklıdır. Sonlu karma modellemenin ilgi odağında model ve parametreden sonuç çıkarma bulunmaktayken modele dayalı kümelemede amaç verinin bir homojen parçalanışını elde etmektir. Bu amaca ulaşmak için ise sonlu karma modellemede model uydurulduktan sonra ek bir adım ile karmaların her bir bileşenindeki elemanların gruplara atanması gerçekleştirilmektedir (Melnykov ve Maitra, 2010). Kümeleme analizi için sonlu karma modeller kullanıldığında kümeleme problemi karma model için parametre tahmini problemine dönüşmekte ve bu parametre tahmini yardımıyla kümelerin elemanlarının sonsal olasılığı hesaplanabilmektedir (Everitt ve ark, 2011).

Modele dayalı kümeleme analizinin en açık şekilde görülebilen avantajı yöntemin istatistiksel dağılımlara dayanmasıdır. Bu da bilinen istatistik teorisi çerçevesinde kullanılan hipotez testlerini ve tahmin yöntemlerini uygulayabilmemizi sağlamaktadır (McLachlan ve Rathnayake, 2016). Yöntemin diğer bir avantajı ise bileşen dağılımını araştırmacının belirleyebilmesidir. Bu da uygulama amacına göre belirlemek istenen küme tiplerini elde etmede araştırmacıya esneklik sağlamaktadır.

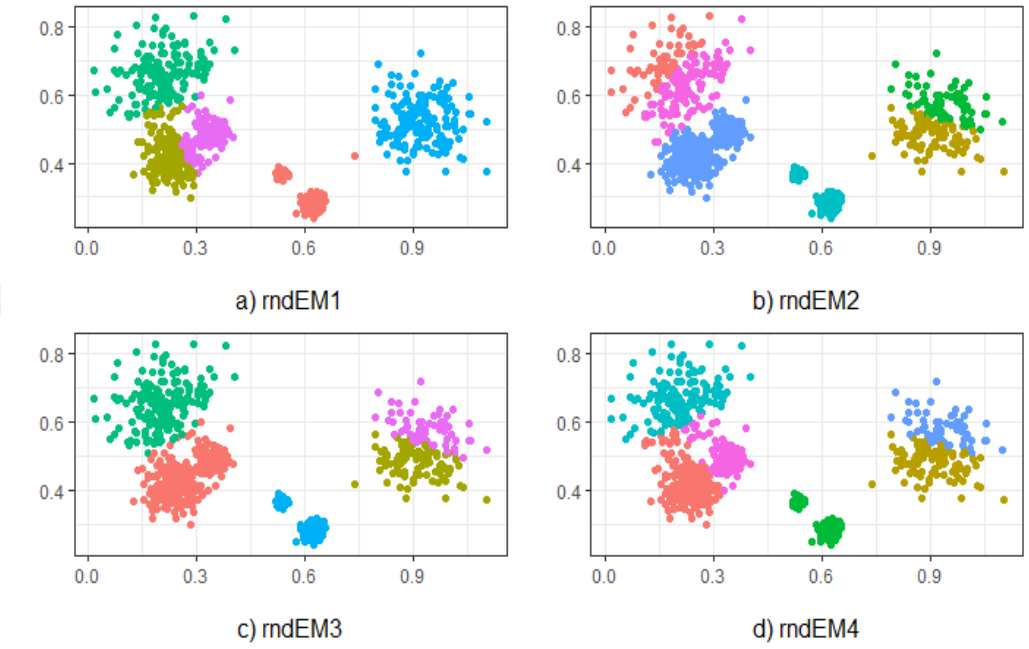
Modele dayalı kümeleme analizinin uygulanması için belirlenmesi gereken üç özellik vardır. Bunlar;

1. EM algoritması için başlangıç parametreleri
2. Parametre tahminleri
3. Uygun model ve küme sayısı

dır (Martinez ve ark, 2010).

4.2.1. EM İçin Başlangıç Değerlerinin Belirlenmesi

Sonlu karma modellerin çok modlu yapısı sebebiyle EM algoritması genellikle yerel maksimuma yakınsamaktadır. Bu da şekil 4.3.'te görüldüğü gibi yöntemin durağan olmayan farklı kümeler elde etmesine sebep olmaktadır. Global maksimum sonucun belirlenmesi için başlangıç parametrelerinin doğru belirlenmesi oldukça önemlidir. Başlangıç değeri belirlemek için çeşitli stokastik veya deterministik yöntemler kullanılmaktadır. Deterministik başlangıç yöntemler genellikle veriye belirli bir kümeleme yöntemi uygulamakta ve elde ettiği kümelerin parametrelerini EM algoritması için başlangıç parametreleri olarak kullanmaktadır. Mclust (Fraley ve ark, 2012) paketinde başlangıç değerleri hiyerarşik modele dayalı kümeleme uygulanarak deterministik olarak belirlenmektedir. Stokastik yöntemlerde ise yöntem birden fazla başlangıç parametresi kullanarak tekrarlanır ve bulunan en iyi çözüm genel çözüm olarak belirlenir. Stokastik yöntem olarak emEM (Biernacki ve ark, 2003) ve RndEM (Maitra, 2009) en sık kullanılan iki yöntemdir. İki yöntemin de farklı üstünlükleri ve dezavantajları vardır. Stokastik yöntemler, EM algoritmasının tekrarlı çalışmasını gerektirdiğinden zaman açısından verimsidir. Deterministik yöntemler ise zaman açısından verimlidir ama bu yöntemler her zaman en iyi başlangıç değerlerini belirleyememektedir. Bu da elde edilen küme yapılarının çoğunlukla yanlış olmasına yol açmaktadır.



Şekil 4.3. Farklı başlangıç parametreleri kullanarak EM algoritması ile kümeleme sonuçları

4.2.1.1. Modele Dayalı Birleştirici Hiyerarşik Kümeleme

Modele dayalı hiyerarşik kümeleme sınıflama olabilirlik modeli için yaklaşık maksimum hesaplamayı sağlayan bir yaklaşımdır (Fraley ve Raftery, 2002). Bu yöntem genellikle EM algoritması için başlangıç parametrelerinin belirlenmesinde kullanılmaktadır. Bu yöntemin birleştirici hiyerarşik kümeleme ile arasındaki en önemli fark, uzaklık ölçüsü tanımlamaya ihtiyaç duyulmamasıdır (Martinez ve ark, 2010). Bunun yerine aşağıda verilen sınıflama olabilirlik fonksiyonu kullanılmaktadır.

$$L_S(\theta_k, \gamma_i, x_i) = \prod_{i=1}^{n} f_{\gamma_i}(x_i, \theta_{\gamma_i}) \qquad (4.24.)$$

Burada γ_i i-inci gözlemin sınıf etiketini göstermektedir. Yani $\gamma_i = k$ olması x_i gözleminin *k*-ıncı bileşen fonksiyonundan üretildiğini göstermektedir.

Modele dayalı birleştirici hiyerarşik kümelemede amaç (4.24.) eşitliğini maksimum olacak şekilde kümeleri belirlemektir. Eklemeli kümeleme algoritmasındaki gibi her bir gözlemin bir küme olduğunu kabul ederek başlar ve her adımda sınıflama olabilirlikteki en büyük artışı sağlayan iki kümeyi birleştirir. Bu adımlar tüm gözlemler tek bir küme oluşturana kadar devam ettirilir.

Normal dağılım karma modeli ile birleştirici hiyerarşik kümelemede, farklı Σ_k yapıları için farklı kriterler türetilmiştir (Fraley ve Raftery, 1998). Örneğin kovaryansların kümeler arasında farklı olabilmesine izin verilen durumda aşağıdaki kriter her adımda

$$\omega_K = \sum_{k=1}^{K} n_k \log \left| \frac{\mathbf{W}_k}{n_k} \right| \qquad (4.25.)$$

Örneklem çapraz çarpım matrisi $\mathbf{W}_k$ ve *k*-ıncı kümedeki eleman sayısı n_k olmak üzere ω_K her adımda minimize edilmesi gereken kriterdir. Kovaryans matrisinin diğer yapıları için minimum edilmesi gereken kriterler Çizelge 4.1.'de verilmiştir.

Modele dayalı eklemeli kümeleme yöntemleri genellikle EM algoritması için başlangıç parçalaması belirlemede kullanılmaktadır. Bu kullanımda klasik birleştirici kümeleme yöntemleri gibi tüm kümelerin her birleştirilmesi gerekmemektedir. Sadece belirli bir iterasyondan sonra kümeleme işlemi tamamlanmaktadır. Bu sayede algoritma oldukça hızlı bir şekilde sonlanır. Bu da yöntemin, EM algoritmasının başlangıç parametrelerini hesaplamada kullanmak için elverişli olma sebebidir.

Çizelge 4.1. Kovaryans matrisinin dört farklı durumuna karşılık gelen kriterler (Gan ve ark, 2007)

Σ_k	**Kriter**
$\sigma^2 I$	$iz\left(\sum_{k=1}^{K} W_k\right)$
$\sigma_k^2 I$	$\sum_{k=1}^{K} n_k \log\left\|Tr\left(\frac{W_k}{n_k}\right)\right\|$
Σ	$\left\|\sum_{k=1}^{K} W_k\right\|$
Σ_k	$\sum_{k=1}^{K} n_k \log\left\|\frac{W_k}{n_k}\right\|$

4.2.2. Tutumlu Normal Dağılım Kümeleme Ailesi

Kısıtlanmamış kovaryansa sahip bir Normal dağılım karmaları üzerinden modele dayalı kümeleme yapılırken toplam tahmin edilmesi gereken parametre sayısı

$$K-1+(Kp)+\frac{(Kp)(p+1)}{2} \tag{4.26.}$$

şeklinde hesaplanabilir. Bunlardan $K-1$ karma oranlarının, Kp ortalamaların ve $(Kp)(p+1)/2$ ise kovaryans matrislerinin tahmin edilmesi gereken parametreleri sayısıdır. Tüm Σ_k'lar için toplam $(Kp)(p+1)/2$ parametre tahmini yapılması gerektiği için bunların kısıtlanması gerekmektedir. Bu kısıtlar bileşen kovaryans matrislerine aşağıdaki özdeğer ayrışımı uygulanarak elde edilmektedir.

$$\Sigma_k = \lambda_k D_k A_k D_k^T \tag{4.27.}$$

Bu ayrışımdaki $\lambda_k = |\Sigma_k|^{\frac{1}{p}}$ ile hesaplanır. D_k sütunları bileşen kovaryans matrisinin özvektörlerinden oluşan bir matristir ve A_k ise Σ_k'nın normalleştirilmiş özdeğerlerinden oluşan köşegen bir matristir. Bu ayrışımda elde edilen elemanlar kümelerin çeşitli geometrik özellikleri ile doğrudan ilişkilidir (Banfield ve Raftery, 1993). λ_k, D_k ve A_k sırasıyla *k*-ıncı kümenin hacmi, şekli ve yönünü belirlemektedir. bu ayrışımdaki elemanlara çeşitli kısıtlar konulmasıyla tutumlu normal dağılım modeli ailesi ortaya çıkarmıştır, (Çizelge 4.2.) (Celeux ve Govaert, 1995). Bu ailede bulunan 14 model küresel, diagonal ve genel olmak üzere üç alt aileye parçalanmışlardır.

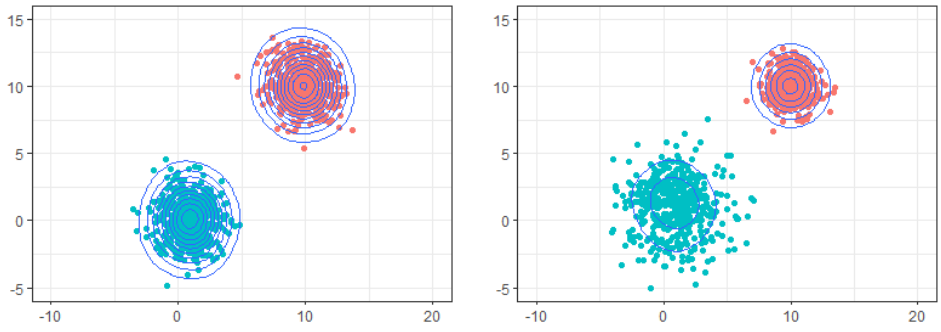
Küresel ailedeki kovaryans matrisleri en fazla kısıt uygulanmış ailedir. Bu ailedeki bir kovaryans modeline sahip kümeler yalnızca küresel şekillere sahiptir (Şekil 4.4.). Yalnızca λ_k değişken olabilmektedir.

Diagonal ailedeki kovaryans modellerinde kümeler hem farklı şekillerde hem de farklı hacimlerde (Şekil 4.5.) bulunabilmektedir.

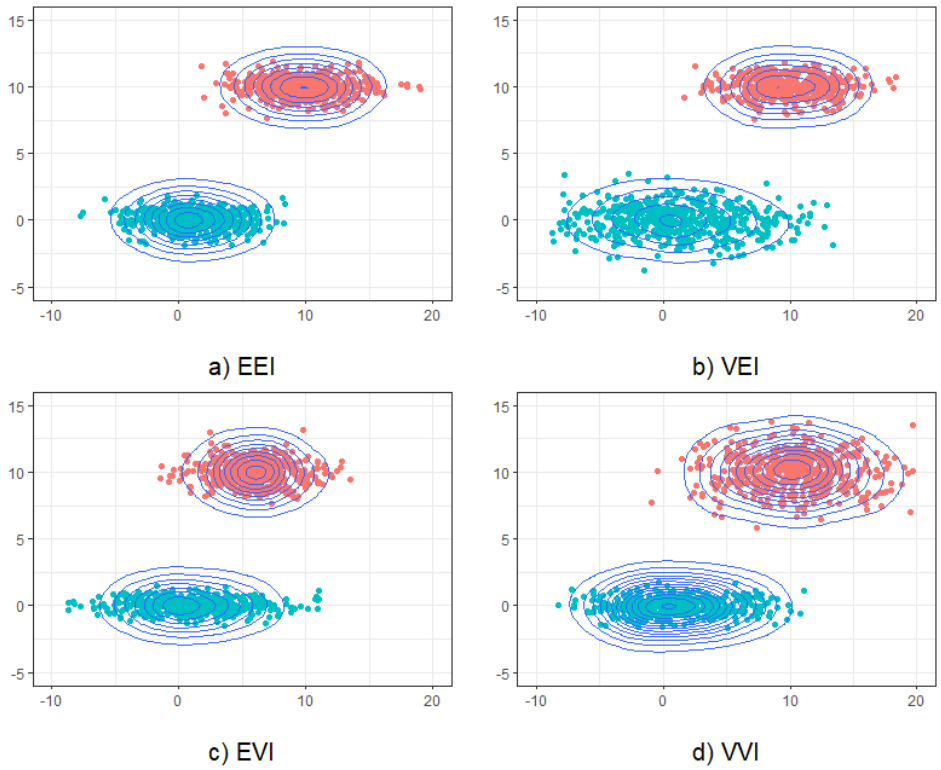
Bu üç kategori altındaki modellerden sadece genel ailedeki sekiz model (Şekil 4.6.) değişkenlerin bağımsız olmadıklarını öne sürer yani bu kategorideki modellerde değişkenlerin yönüne herhangi bir kısıt uygulamamıştır (McNicholas, 2016). Kovaryans ayrışımına uygulanan kısıtlar arttıkça tahmin edilmesi gereken parametre sayısı azalmaktadır ve bu parametre sayıları Çizelge 4.3.'te verilmiştir. Tutumlu normal dağılım modellerinin küme şekilleri ile ilgili önbilgi ile birlikte kullanılarak model kısıtlanabilir. Bu sayede modele dayalı kümeleme analizi için önemli bir hız kazancı sağlanır.

Çizelge 4.2. Tutumlu normal dağılım modelleri (McNicholas, 2016)

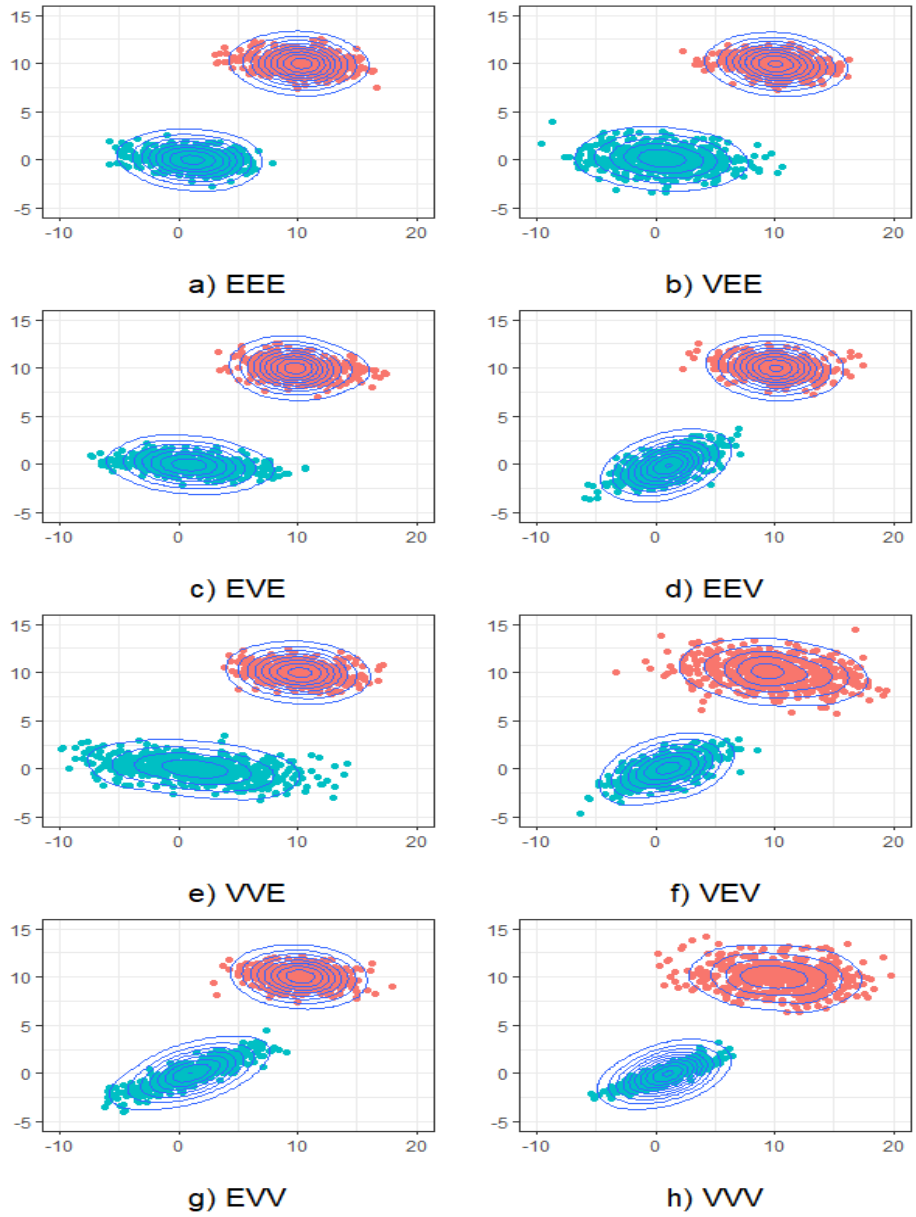
Aile	Kovaryans	Hacim	Şekil	Yön	Model
Küresel	$\Sigma_j = \lambda I$	Eşit	Küresel	-	EII
	$\Sigma_j = \lambda_j I$	Değişken	Küresel	-	VII
Diagonal	$\Sigma_j = \lambda A$	Eşit	Eşit	Eksen	EEI
	$\Sigma_j = \lambda_j A$	Değişken	Eşit	Eksen	VEI
	$\Sigma_j = \lambda A_j$	Eşit	Değişken	Eksen	EVI
	$\Sigma_j = \lambda_j A_j$	Değişken	Değişken	Eksen	VVI
Genel	$\Sigma_j = \lambda DAD^T$	Eşit	Eşit	Eşit	EEE
	$\Sigma_j = \lambda_j DAD^T$	Değişken	Eşit	Eşit	VEE
	$\Sigma_j = \lambda DA_j D^T$	Eşit	Değişken	Eşit	EVE
	$\Sigma_j = \lambda D_j A D_j^T$	Eşit	Eşit	Değişken	EEV
	$\Sigma_j = \lambda_j DA_j D^T$	Değişken	Değişken	Eşit	VVE
	$\Sigma_j = \lambda_j D_j A D_j^T$	Değişken	Eşit	Değişken	VEV
	$\Sigma_j = \lambda D_j A_j D_j^T$	Eşit	Değişken	Değişken	EVV
	$\Sigma_j = \lambda_j D_j A_j D_j^T$	Değişken	Değişken	Değişken	VVV



Şekil 4.4. Küresel aile için EII ve VII kovaryans modelli küme şekilleri



Şekil 4.5. Diagonal aile için EEI, VEI, EVI, VVI kovaryans modelli küme şekilleri



Şekil 4.6. Genel ailede bulunan kovaryans modelleri için küme şekilleri

Çizelge 4.3. Kısıtlanmış modellerinin serbest kovaryans parametreleri sayısı (McNicholas, 2016)

Model	Kovaryans	Kovaryans Parametreleri Sayısı
EII	$\Sigma_j = \lambda I$	1
VII	$\Sigma_j = \lambda_j I$	K
EEI	$\Sigma_j = \lambda B$	p
VEI	$\Sigma_j = \lambda_j B$	$K+p-1$
EVI	$\Sigma_j = \lambda B_j$	$1+K(p-1)$
VVI	$\Sigma_j = \lambda_j B_j$	Kp
EEE	$\Sigma_j = \lambda DAD^T$	$p(p+1)/2$
VEE	$\Sigma_j = \lambda_j DAD^T$	$K+p-1+p(p+1)/2$
EVE	$\Sigma_j = \lambda DA_j D^T$	$1+K(p-1)+p(p+1)/2$
EEV	$\Sigma_j = \lambda D_j AD_j^T$	$p+Kp(p-1)/2$
VVE	$\Sigma_j = \lambda_j DA_j D^T$	$Kp+p(p-1)/2$
VEV	$\Sigma_j = \lambda_j D_j AD_j^T$	$K+p-1+Kp(p-1))/2$
EVV	$\Sigma_j = \lambda D_j A_j D_j^T$	$1+K(p-1)+Kp(p-1)/2$
VVV	$\Sigma_j = \lambda_j D_j A_j D_j^T$	$Kp(p+1)/2$

4.2.3. Model Seçimi

Modele dayalı kümelemede model seçme işleminde modeller arasından veriye en iyi uyanı ve *K*, küme sayısını birlikte belirlenir. Normal dağılım karması üzerinden modele dayalı kümeleme analizi uygulamalarında bu kısıtlanmış ailelerdeki modellerin her birinin belirli bir aralıktaki her *K* değeri için parametre tahmini yapılır ve bu modeller arasından veriye en uygun olanı çeşitli kriterler yardımıyla belirlenir (Melnykov ve Maitra, 2010). Bu kriterler arasında en popülerleri Akaike Bilgi Kriteri (Akaike, 1973)

ve Bayesyan Bilgi Kriteridir (Schwarz, 1978). Gerçek model ile uydurulan model arasındaki göreceli uzaklığın yansız tahmin edicisi olan Akaike Bilgi Kriteri (AIC),

$$AIC = 2l(\hat{\theta}) - 2v \tag{4.28.}$$

şeklinde hesaplanmaktadır. Burada *v,* modeldeki tahmin edilmesi gereken parametre sayısını göstermektedir. Bu eşitlikteki ilk terim olabilirlik fonksiyonunu göstermekte iken ikinci terim model karmaşıklığını önlemek için ceza katsayısıdır. İstenen model veriye iyi uymakla beraber en az parametre içeren modeldir (Everitt ve ark, 2011). AIC'nin popülerliğine rağmen bu kriterin model seçiminde kullanılması bileşen sayısının aşırı tahminine yol açabilmektedir (Celeux ve Soromenho, 1996).

Bayesyan Bilgi Kriteri (BIC) ise,

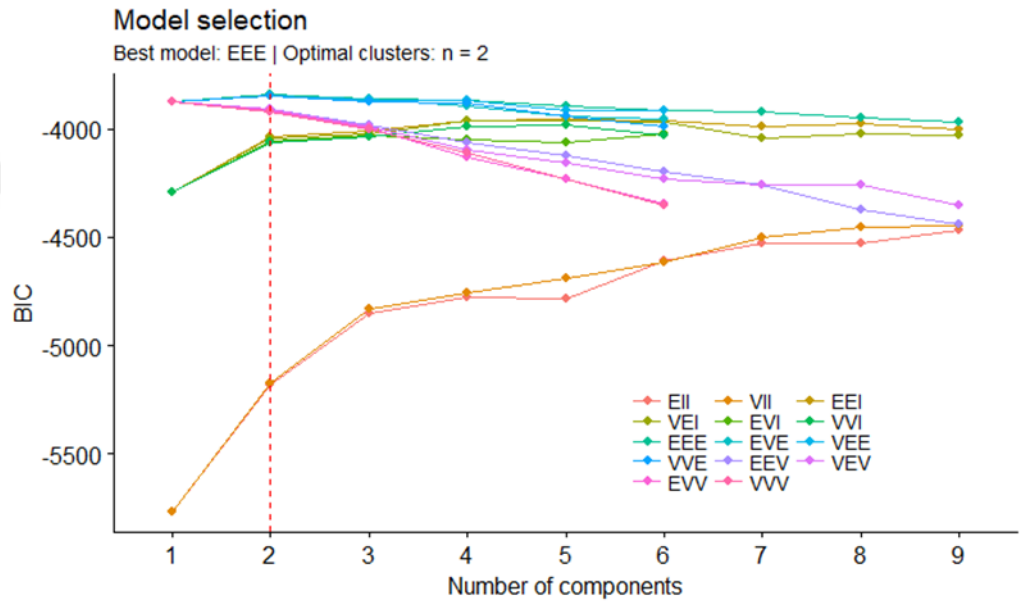
$$BIC = 2l(\hat{\theta}) - v\log(n) \tag{4.29.}$$

şeklinde hesaplanmaktadır. BIC hesaplanmasında örneklem büyüklüğü arttıkça model karmaşıklığının ceza katsayısı logaritmik artış göstermektedir. BIC doğru küme sayısı ve modeli belirlemede modele dayalı kümeleme analizi için en sık kullanılan yöntemdir (Dasgupta ve Raftery, 1993; Maitra ve Melnykov, 2010). Ancak gözlem sayısının küçük olduğu durumlarda küme sayısını normalde olduğundan daha az tahmin etmektedir (Melnykov ve Maitra, 2010). Her iki kriterde de büyük değerler elde eden modeller tercih edilmektedir. Şekil 4.7.'de bileşen sayısı iki olan EEE modeli maksimum BIC değerine sahiptir. Bu sebeple modele dayalı kümeleme analizi için tercih edilmektedir.

4.3. Yüksek Boyutlu Verilerde Modele Dayalı Kümeleme

Modele dayalı kümeleme yöntemleri tüm güçlü yönlerine rağmen değişken sayısının artması bu yöntemi nerdeyse kullanılmaz hale getirmektedir. Normal dağılım karmaları kullanılarak kümeleme yapıldığında toplam yapılması gereken parametre

tahmini sayısı, $(K\text{-}1)+Kp+Kp(p+1)/2$ olarak verilmişti. Yani değişken sayısı yapılması gereken parametre tahmini sayısının karesel bir fonksiyonudur (Bergé ve ark, 2012).



Şekil 4.7. BIC kriterlerinin karşılaştırılması ile veriye en uygun model ve küme sayısının belirlenmesi

Yani değişken sayısı arttıkça yöntemin uygulanış süresi de karesel bir biçimde artacaktır. Ayrıca normal dağılım karmalarına EM algoritması uygulanırken değişken sayısının fazla olması durumunda tahmin edilen kovaryans matirisi $\hat{\Sigma}_k$ uygun tanımlanmayacak ve durağan olmayan bir fonksiyona sebep olacaktır. Hatta gözlem sayısı değişken sayısından az olursa $\hat{\Sigma}_k$ tersi olmayan bir matris olacak ve modele dayalı kümeleme hiç uygulanamayacaktır (Bouveyron ve Brunet-Saumard, 2014). Veri madenciliği problemlerinde sıklıkla karşılaşılan bu durum özellikle gen ifade analizi uygulamalarında önemli problemlere sebep olmaktadır. Bu sebeple modele dayalı

kümeleme yöntemlerinin bunun gibi veri madenciliği problemlerine direkt olarak uygulanması önerilmez. Bu tipteki verilere modele dayalı kümeleme analizi uygulanabilmesi için öncelikle değişken seçimi yapılması gerekmektedir.

Kümeleme analizi için değişken seçimi verinin gerçek küme yapısı ile ilgili bilgi içeren değişkenleri seçip kümeleme ile ilgisiz (gürültü) değişkenlerin çıkarılması ile uygulanmaktadır. Raftery ve Dean (2006) modele dayalı kümeleme için değişken seçimini, değişken alt gruplarının yaklaşık bayes faktörü kullanılarak ikili karşılaştırılması ile model karşılaştırma problemine dönüştürmüşlerdir. Bu yöntemde **X** veri matrisinin değişkenleri,

X_1: Kümeleme değişkenleri

X_2: Kümeleme değişken olarak kullanılıp kullanılmaması henüz karar aşamasında olan değişkenler

X_3: Kalan değişkenler

şeklinde üç ayrık alt gruba ayrılır. X_2'deki değişkenlerin kümelemeye eklenip eklenmemesi

$$BIC_{Fark} = BIC_{küme}(X_1, X_2) - BIC_{küme^c}(X_1, X_2) \tag{4.30.}$$

farkı ile belirlenir. Burada $BIC_{küme}(X_1, X_2)$, $(X_1 \cup X_2)$ değişkenleri kullanılarak uydurulan en iyi karma model için BIC değeridir. $BIC_{küme^c}(X_1, X_2)$ ise aynı değişken grubuyla hiçbir kümeleme olmaması durumundaki BIC değeridir. Burada

$$BIC_{küme^c}(X_1, X_2) = BIC_{küme}(X_1) + BIC_{reg}(X_2 \mid X_{küme})$$

şeklinde açılabilir. Yani kümelemede kullanılacağı kesinleşmiş olan X_1 değişkenleri grubu için en iyi kümeleme modelinin BIC değeri ile kümeleme değişkeni olmaya aday

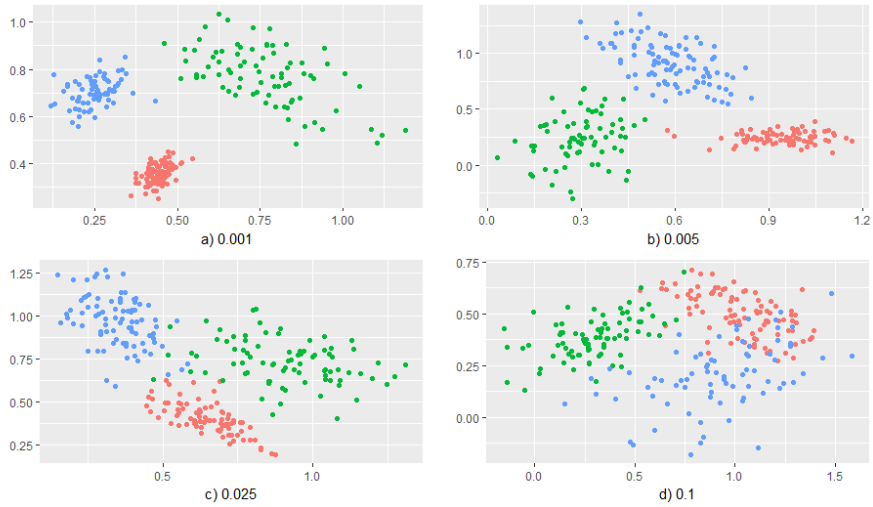
X_2 değişkenlerinin X_1 üzerindeki regresyonu için BIC değeridir. (4.29)'daki BIC_{Fark} eşitliği bayes faktörünün logaritmasının yaklaşık bir değeridir. Burada bayes faktörü X_2'in bir kümeleme değişkeni olduğu model ile X_2'nin koşullu olarak kümelemeden bağımsız olduğu modeli karşılaştırır. BIC_{Fark}'ın büyük değerleri X_i değişkenlerinin kümeleme bilgisi içerdiğini desteklemektedir.

5. UYGULAMALAR

5.1. Simülasyon

5.1.1. Veri Üretme Süreci

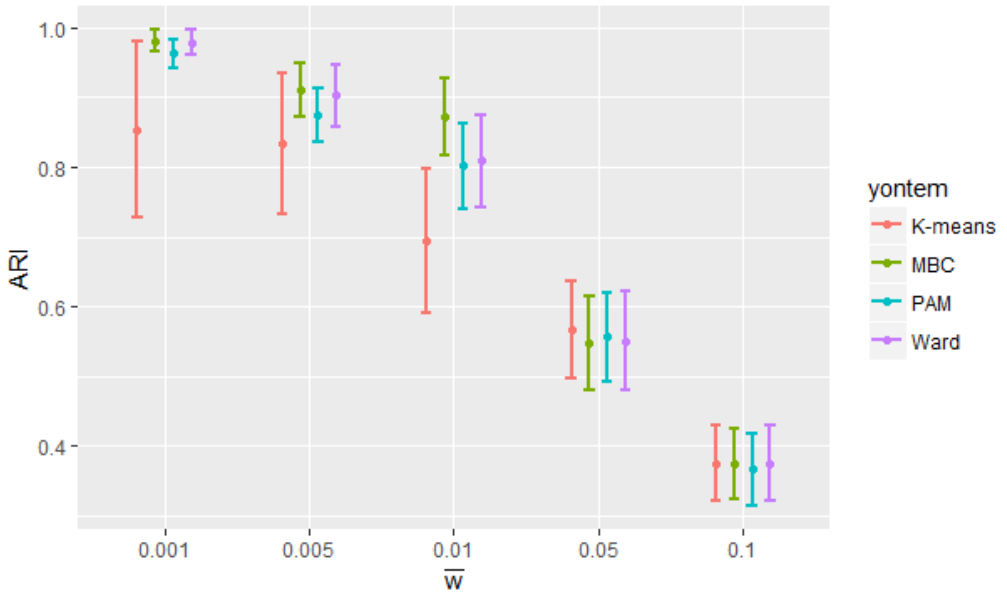
Kümeleme analizi sonuçlarını test etmek için geliştirilmiş bir R programı paketi olan MixSim (Melnykov ve ark, 2012) çok değişkenli normal dağılım karmaları kullanarak veri üretmeyi sağlamaktadır. Üretilecek verinin belirlenen girdileri olan, n gözlem sayısı, p değişken sayısı, K bileşen sayısı ve $\overline{w}$ ortalama ikişerli ayrıklık değerleri ile veriler üretilmiştir. Bu girdilerden $\overline{w}$ sayesinde veri üretilirken bileşenler arasındaki iç içe geçme durumunun şiddeti ayarlanabilmektedir (Maitra ve Melnykov, 2010). $\overline{w}$ ortalama karmaşıklık değeri arttıkça kümeler arası içiçeliğin arttığı Şekil 5.1.'de görülmektedir.



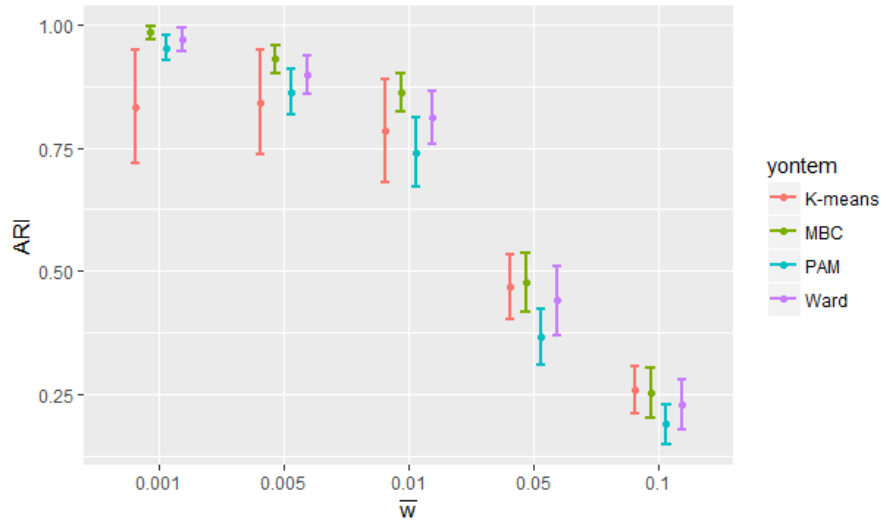
Şekil 5.1. Çeşitli $\overline{w}$ değerleri için üretilen iki boyutlu veriler ve bu verilerin küme yapıları

5.1.2. Yapay Veriler Üzerinden Karşılaştırma

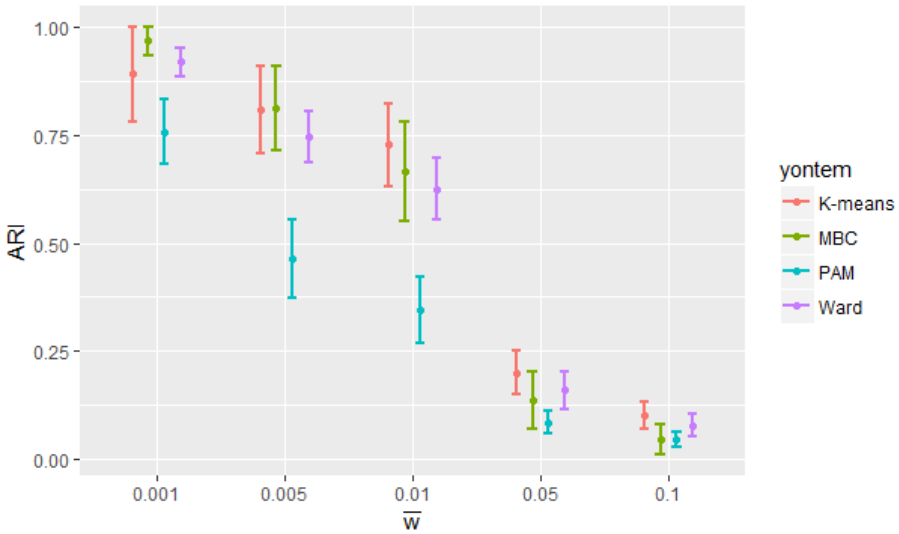
Örneklem büyüklüğü 250 ve küme sayısı 6 olacak şekilde 500 veri her bir $\overline{w} = (0.001, 0.005, 0.01, 0.05, 0.1)$ ve p= (5, 10, 25) değerleri için üretilmiştir. Bu üretilen verilere Modele dayalı kümeleme analizi (MBC), K-means, PAM ve Ward kümeleme yöntemleri küme sayısının bilindiği varsayımı altında uygulanmıştır. Ayrıca K-means, PAM ve Ward metotları uygulanmadan önce veriler standartlaştırılmıştır. Yöntemlerin performansları her bir veriye uygulanan yöntem sonunda elde edilen düzeltilmiş rand indekslerinin ortalama ve standart sapmaları üzerinden gösterilmiştir (Şekil 5.2-5.4). Düzeltilmiş rand indeksinin bire yakın değerleri yöntemin doğru kümeleme yaptığını göstermektedir. Bu sayede değişken sayısı ve ortalama karmaşıklığının kümeleme performansına etkisi gözlenmiştir.



Şekil 5.2. (p=5, K=6, n=250) için her bir $\overline{w}$ = (0.001, 0.005, 0.01, 0.05, 0.1) ortalama karmaşık değeri için oluşturulan 500'er veriye uygulanan dört farklı yöntemin sonuçları



Şekil 5.3. (p=10, K=6, n=250) için her bir $\overline{w}$ = (0.001, 0.005, 0.01, 0.05, 0.1) ortalama karmaşık değeri için oluşturulan 500'er veriye uygulanan dört farklı yöntemin sonuçları

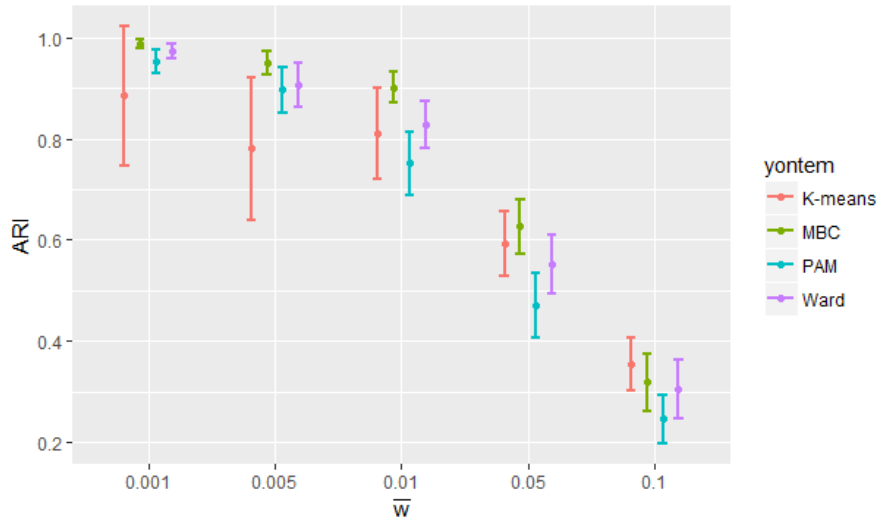


Şekil 5.4. (p=25, K=6, n=250) için her bir $\overline{w}$ = (0.001, 0.005, 0.01, 0.05, 0.1) ortalama karmaşık değeri için oluşturulan 500'er veriye uygulanan dört farklı yöntemin sonuçları

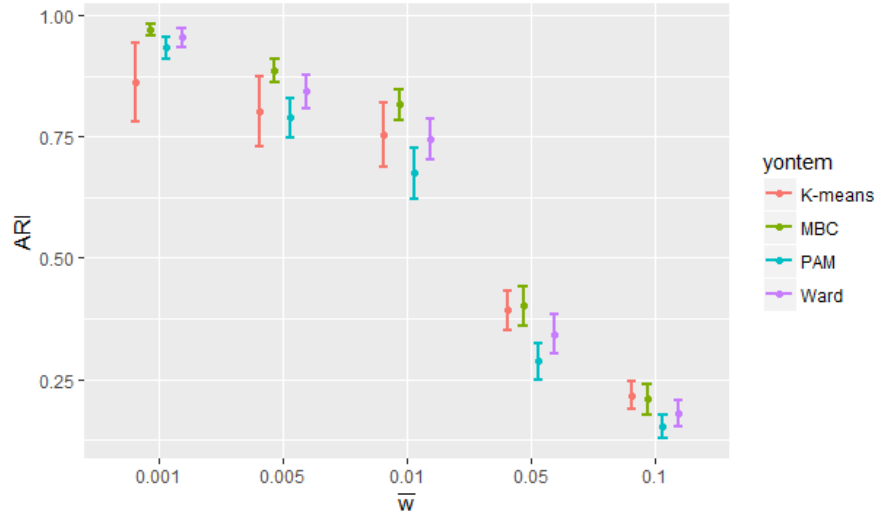
Şekil 5.2.-5.5.'te verinin ortalama karmaşıklığı $\bar{w}$, arttıkça değişken sayısı fark etmeksizin her bir kümeleme yönteminin performanslarının düştüğü gözlenmektedir. Değişken sayısındaki artış ise PAM ve Ward yöntemlerini etkilemiştir. Değişken sayısına bağlı en büyük performans kaybı PAM metodunda gözlemlenmektedir. Ayrıca yöntemler tek tek incelenip karşılaştırıldığında K-means'in diğer yöntemlere göre daha az durağan bir performans gösterdiği gözlenmektedir. Modele dayalı kümeleme analizi performansı incelendiğinde küçük $\bar{w}$ değerinde değişken sayısındaki artıştan en az etkilenen yöntem olduğu görülmektedir. Diğer $\bar{w}$ ve p değerleri için performanslar incelendiğin ise genel olarak modele dayalı kümelemenin diğer yöntemlerden daha üstün olduğu gözlenmektedir.

Küme sayısının K=(5, 10, 20) değerleri ve $\bar{w}$ = (0.001, 0.005, 0.01, 0.05, 0.1) değerleri için 250 gözlemli ve 10 değişkenli 500'er veri üretildiğinde ve öncekiyle aynı yöntemler uygulandığında Şekil 5.5.-5.8.'de verilen sonuçlar elde edilmiştir. Küme sayısındaki artış küme başına düşen gözlem sayısını azaltmıştır. Bu da özellikle düşük $\bar{w}$ değerlerinde K-means'in daha durağan sonuçlar elde etmesini sağlamıştır. Ayrıca küme sayısındaki artışla birlikte PAM yönteminin performansında önemli ölçüde azalma gerçekleşmiştir. Ayrıca yüksek ortalama karmaşıklık durumu yani $\bar{w} = 0.1$ olması durumunda küme sayısındaki artış her yöntemin ortalama düzeltilmiş rand indeksini 0.35 civarından 0.15 civarlarına geriletmiştir. Buradan yüksek karmaşıklık içeren ve küme sayısının fazla olduğu uygulamalar için uygulanan yöntemlerin hiçbirinin elverişli olmadığı görülmektedir.

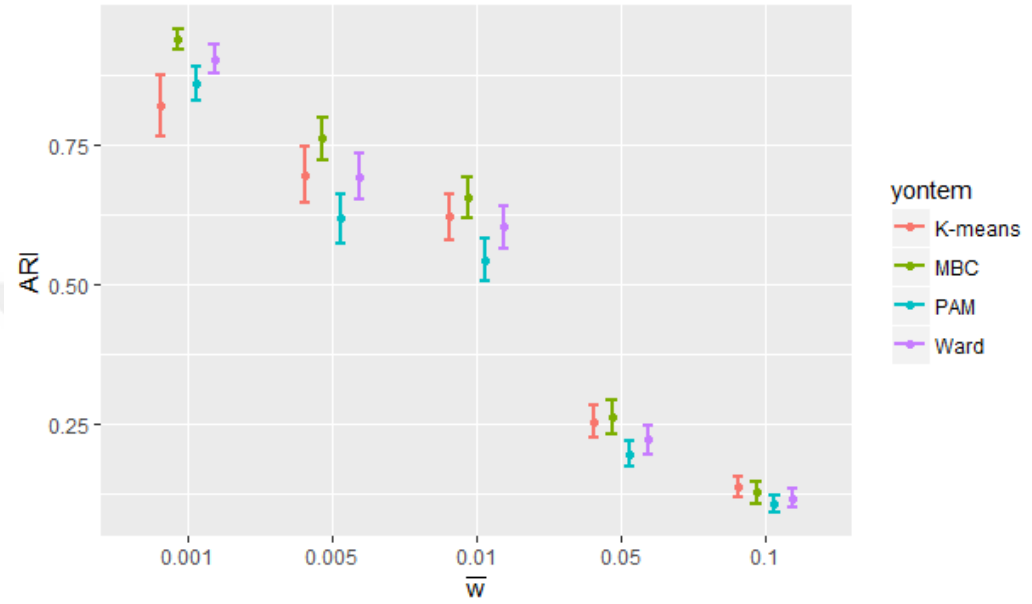
Bu uygulamadan PAM yönteminin değişken ve küme sayısının artışının kümeleme performansını düşürdüğü gözlenmektedir. K-means benzer tipteki verilere uygulandığında farklı sonuçların en çok gözlendiği yöntemdir. Bu yüzden K-means yöntemi kullanılırken en iyi sonucun bulunduğuna emin olunmalıdır. Modele dayalı kümeleme ve Ward yöntemi genel olarak en iyi sonuçları vermişlerdir.



Şekil 5. 5. (p=10, K=5, n=250) için her bir $\bar{w}$ = (0.001, 0.005, 0.01, 0.05, 0.1) ortalama karmaşık değeri için oluşturulan 500'er veriye uygulanan dört farklı yöntemin sonuçları



Şekil 5. 6. (p=10, K=10, n=250) için her bir $\bar{w}$ = (0.001, 0.005, 0.01, 0.05, 0.1) ortalama karmaşık değeri için oluşturulan 500'er veriye uygulanan dört farklı yöntemin sonuçları



Şekil 5. 7. (p=10, K=20, n=250) için her bir $\overline{w}$ = (0.001, 0.005, 0.01, 0.05, 0.1) ortalama karmaşık değeri için oluşturulan 500'er veriye uygulanan dört farklı yöntemin sonuçları

5.2. Gerçek Veri

Sınıflama verileri, küme etiketleri önceden bilinen verilerdir. Bu tip veriler kümeleme analizinde performans karşılaştırılması için kullanılabilmektedir. Kümeleme performansı değerlendirilmesi için sınıflama verisi kullanılacağı zaman dikkat edilmesi gereken nokta verideki sınıf etiketlerinin değişkenlerin doğal bir kümesini temsil etmesidir. Aksi halde kümeleme analizinin belirleyeceği kümelerin gerçek sınıf etiketleri ile uyuşma göstermesi beklenmemelidir. Bu çalışmada kullanılan sınıflama verileri ve bu verilerin özellikleri Çizelge 5.1.'de verilmiştir. Seeds verisi modele dayalı kümeleme ve değişken seçim süreci detaylı olarak incelenmiştir. Diğer veriler ise Bölüm

5.2.2.'de kümeleme yöntemlerinin performanslarının değerlendirilmesi için kullanılmıştır.

Çizelge 5.1. Sınıflama verileri

Veri	Gözlem Sayısı	Değişken	Küme Sayısı	Kaynak
Ais	202	12	2	Weisberg (2005)
Body	507	24	2	Heinz ve ark (2004)
İyonosfer	351	33	2	Sigillito ve ark (1989)
Olive	572	8	9	Forina ve ark (1986)
Seeds	210	8	3	Charytanowicz ve ark (2010)
Wine	178	13	3	Forina ve ark (1982)

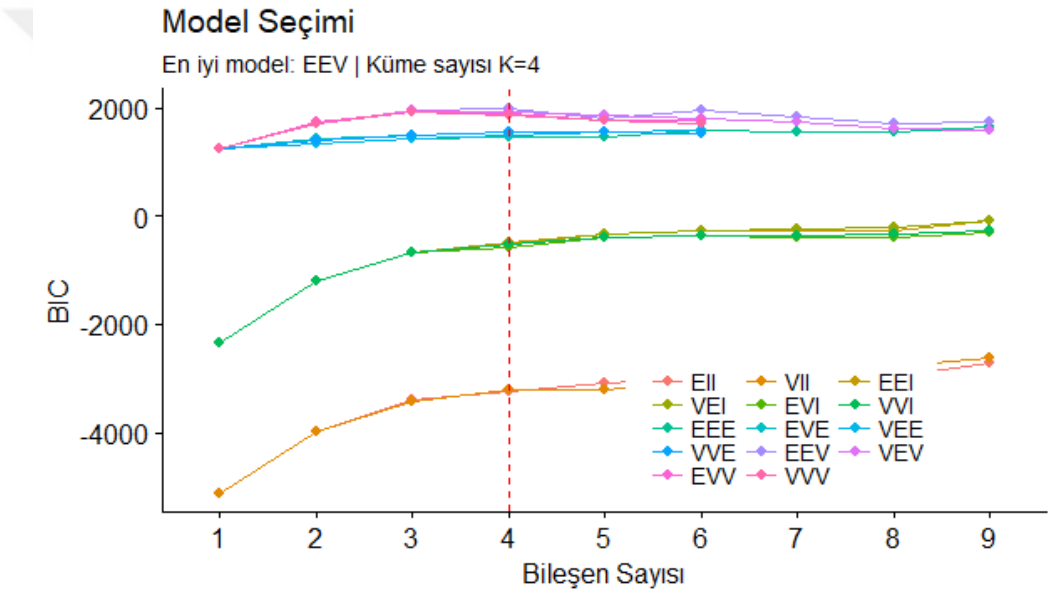
5.2.1. Seeds Verisi

UCI makine öğrenme deposundan (Lichman, 2013) alınan bu veri kümesi üç farklı türdeki 210 buğday tanesinin aşağıdaki yedi geometrik özelliğini içermektedir.

- V1: Alan
- V2: Çevre
- V3: Yoğunluk
- V4: Uzunluk
- V5: Genişlik
- V6: Asimetri katsayısı
- V7: Oluk uzunluğu

Bu veriye modele dayalı kümeleme analizi uygulanacak en iyi kovaryans modelinin ve küme sayısının belirlenmesi amacıyla birden dokuza kadar tüm küme sayısı değerleri için tüm tutumlu kovaryans modellerinin (Çizelge 4.2.) BIC değerleri

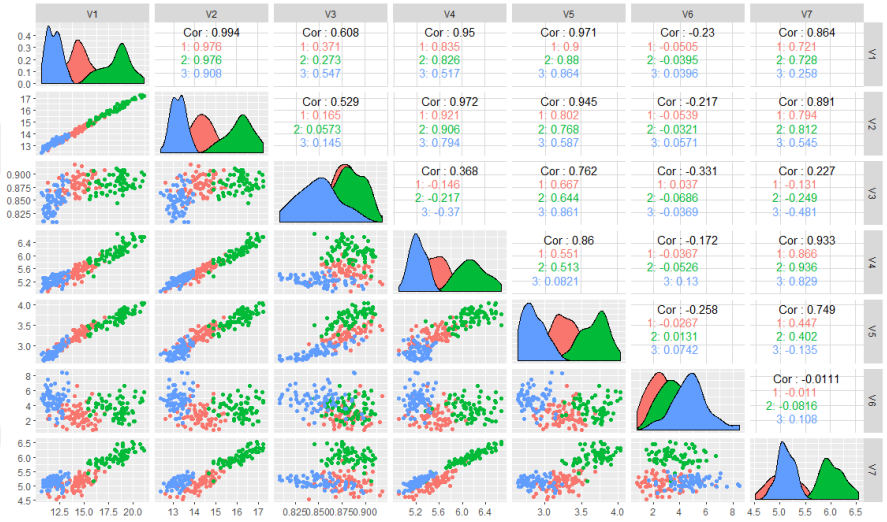
hesaplanmıştır. Sonuç olarak maksimum BIC değerine sahip 4 küme içeren EEV modeli en iyi model olarak belirlenmiştir (Şekil 5.8.). Bu belirlenen model ile yapılan modele dayalı kümeleme analizi sonucunda küme sayısını 3 yerine 4 olarak elde edilmiştir. Sonuç olarak modele dayalı kümeleme analizi Performansın derecesi doğru kümeleme yüzdesi %61 ve Düzeltilmiş rand indeksi 0.481 olarak belirlenmiştir.



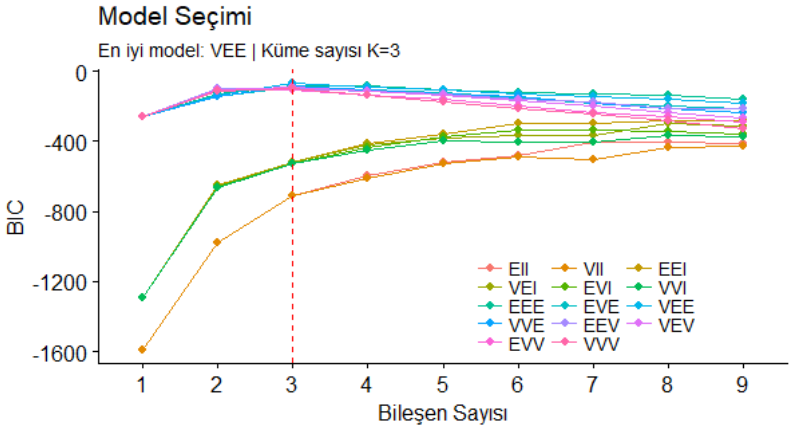
Şekil 5.8. Seeds verisi için en iyi modelin BIC değeri ile belirlenmesi

Şekil 5.9. incelendiğinde seeds verisinde bazı değişkenler arasında yüksek korelasyon olduğu ayrıca bazı değişken ikililerinin kümeleme bilgisi içermediği görülmektedir. Bu durumda kümeleme analizi için değişken seçimi uygulanırsa kümeleme performansının artacağından şüphelenilebilir. Clustvarsel (Scrucca ve ark, 2014), R paketi yardımıyla Seeds verisi için kümeleme analizi uygulandığında asimetri katsayısı, çevre ve uzunluk değişkenleri kümeleme analizi için önemli değişkenler olarak seçilmiştir. Bu üç değişken için model belirleme Şekil 5.10.'daki gibi gerçekleştirilmiştir. Bu sayede en iyi model maksimum BIC değerine sahip olan 3 küme içeren EEV modeli olarak belirlenmiştir. Yani değişken seçimi yardımıyla modele dayalı kümeleme Seeds verisi için doğru küme sayısını belirlemiştir. Bu model yardımıyla elde

edilen doğru kümeleme yüzdesi %94.7'ye ve düzeltilmiş rand indeksi 0.852'ye yükselmiştir. Yani değişken seçimi gerçek kümeleme bilgisi içeren değişkenlerin elde edilip içermeyen gürültü değişkenlerin analizin dışında tutarak kümeleme performansını önemli ölçüde geliştirmiştir.



Şekil 5.9. Seeds verisi için saçılım matrisi ve korelasyon katsayıları



Şekil 5.10. Seeds verisi için değişken seçimi sonrası en iyi modelin BIC değeri ile belirlenmesi

5.2.2. Sınıflama Verileri Kullanılarak Kümeleme Performansı Değerlendirmesi

Çizelge 5.1.'de bulunan veri setlerine küme sayısı bilindiği varsayılarak modele dayalı kümeleme (MBC), değişken seçimi sonrası modele dayalı kümeleme (DeğSeç), K-means, PAM ve DBSCAN algoritmaları uygulanmıştır. Elde edilen sonuçlar büyük değerleri daha iyi bir kümeleme performansını temsil eden düzeltilmiş rand indeki (ARI) ve F ölçüsü (F) ile birbirleri ile kıyaslanmıştır. Elde edilen sonuçlar Çizelge 5.2.'de özetlenmiştir

12 değişkenli ve küme sayısı 2 olan Ais verisi Avusturalya'daki 202 sporcunun çeşitli vücut özelliklerini içermektedir. Bu veride elde edilen sonuçlar incelendiğinde değişken seçiminin modele dayalı kümeleme analizi sonuçlarını geliştirdiği gözlenmektedir. Diğer yöntemler olan Ward, PAM ve DBSCAN ise doğru kümeleri belirlemede iyi bir performans gösterememişlerdir.

Çizelge 5.2. Sınıflama verilerinden elde edilen kümeleme performansları

Veri		MBC	DeğSeç	K-means	PAM	DBSCAN
Ais	ARI	0.464	0.674	0.131	0.327	0.125
	F	0.737	0.836	0.609	0.673	0.500
Body	ARI	0.915	0.756	0.467	0.529	0.308
	F	0.957	0.878	0.736	0.764	0.625
İyonosfer	ARI	0.339	0.543	0.178	0.173	0.683
	F	0.682	0.782	0.605	0.603	0.861
Olive	ARI	0.649	0.799	0.459	0.443	0.306
	F	0.712	0.835	0.549	0.536	0.525
Seeds	ARI	0.737	0.852	0.710	0.710	0.366
	F	0.824	0.901	0.807	0.807	0.614
Wine	ARI	0.967	0.813	0.371	0.371	0.291
	F	0.978	0.877	0.584	0.584	0.617

25 değişken içeren Body verisine modele dayalı kümeleme analizi uygulandığında 0.915 ARI ve 0.957 F-ölçüsü ile oldukça iyi ve uygulanan tüm diğer yöntemlerden üstün bir kümeleme performansı göstermiştir. Fakat değişken seçimi uygulanıp modele dayalı kümeleme analizi uygulandığında ARI ve F-ölçüsü

değerlerinin sırasıyla 0.756 ve 0.878'e gerilediği görülmektedir. Bunun sebebi olarak Body verisindeki 25 değişkenin de analiz için anlamlı bir kümeleme bilgisine sahip olmasıdır.

İyonosfer verisi 34 değişken içermekte ve 2 sınıf değeri bulunmaktadır. Bunlar, iyonosferde bir düzen belirleyen iyi radar sinyalleri ve belirleyemeyen kötü radar sinyallerdir. Elde edilen kriterler incelendiğinde bu veri için en iyi kümeleme performansını DBSCAN algoritması göstermektedir. DBSCAN algoritmasının iyi performans göstermesi, kümelerin karmaşık ve bağıntılı bir şekilde devam eden yapıda olması sebebiyledir. Bu sonuçtan yola çıkarak diğer yöntemlerin bu yapıdaki kümeleri belirlemekte başarısız olduğu gözlemlenmektedir. Modele dayalı kümeleme analizinin değişken seçimiyle uygulanması bu veride de performansın gelişmesini sağlamıştır. Fakat bu artışla bile iyi bir kümeleme elde edildiği söylenemez.

Olive verisi 572 gözlemli ve 8 değişkenlidir. Bu veride sınıf değişkeni olarak zeytinyağlarını yağasidi bileşenlerine göre 3 grubu yer almaktadır. Elde edilen sonuçlar incelendiğinde değişken seçimi modele dayalı kümeleme analizi performansını 0.80 ARI değerine yakınlaştırmış ve zeytinyağı gruplarının belirlenmesine önemli bir katkı sağlamıştır.

Wine verisinde 178 şarap 13 değişken üzerinden kalitelerine göre 3 farklı gruba atanmıştır. Bu grupları en iyi modele dayalı kümeleme analizi 0.967 ARI ve 0.978 F-ölçüsü ile belirlemiştir. Yani modele dayalı kümeleme analizi şarap kalitesini oldukça iyi bir biçimde gruplamıştır. Ayrıca bu veri için değişken seçimi modele dayalı kümeleme analizine faydalı katkı sağlayamamıştır ve kümeleme performansını düşürmüştür.

6. SONUÇ VE ÖNERİLER

Büyük ölçekli verilerden sonuç çıkarılmak istendiğinde klasik istatistiksel yöntemlerin uygulanması ile verimli sonuç elde edilememektedir. Özellikle yüksek değişkenli verilerde kümelerin belirlenmesi kümeleme analizinin önemli bir problemidir. Modele dayalı kümeleme analizi yöntemi ile değişken seçiminin birleştirilmesi bu problem için çözüm sağlamaktadır.

Bu çalışma, kümeleme analizinde kullanılan klasik yöntemler ve modele dayalı kümeleme analizinin, farklı senaryolar altında üretilmiş yapay veriler üzerinden yapılan performans değerlendirmesini içerir. Bu uygulama yardımıyla kümeleme yöntemlerinin performanslarının değişken sayısı, küme sayısı ve verinin karmaşıklığındaki değişimlerden ne ölçüde etkilendiği incelenmiştir. Ayrıca gerçek sınıflama verisi üzerinde modele dayalı kümeleme analizinin değişken seçimi adımları yardımıyla kümeleme performansının önemli ölçüde artırılabileceği gösterilmiştir. Son olarak ise altı sınıflama verisine çeşitli kümeleme analizi yöntemleri uygulanmış ve elde edilen sonuç düzeltilmiş Rand indeksi ve F-ölçüsü yardımıyla değerlendirilmiştir. Bu karşılaştırma sonucunda modele dayalı kümeleme analizinde değişken seçiminin kullanımının kümeleme performansına önemli katkı sağlayabileceği görülmüştür.

Sonraki çalışmalarda modele dayalı kümeleme ile yeni değişken seçimini sağlayacak yöntemler geliştirilebilir. Ayrıca küme sayısının fazla olduğu karmaşık yapıdaki kümeler bulunan verilerde kümeleme analizinin performansın gelişimi üzerinde durulabilirdir.

KAYNAKLAR

Aggarwal, C. C., 2013. An Introduction to Cluster Analysis. (C. C. Aggarwal, C. K. Reddy editör). Data Clustering, Taylor & Francis, London, 1-25.

Akaike, H., 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika, 60(2):255-265.

Anderberg, M. R., 1973. Cluster Analysis For Applications. Academic Press, London, 346.

Arthur, D., Vassilvitskii, S., 2007. K-means++: The Advantages of Careful Seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 1027-1035.

Assent, I., 2012. Clustering high dimensional data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(4):340–350.

Bailey, K. D., 1975. Cluster Analysis. (D. R. Hesie editör). Sociological methodology. San Francisco, Jossey-Bass:83-111.

Ball, G.H., Hall, D.J., 1965. ISODATA, a novel method of data analysis and pattern classification. Stanford CA.

Banfield, J. D., Raftery, A. E., 1993. Model-Based Gaussian and Non-Gaussian Clustering. Biometrics, 49(3):803–821.

Bergé, L., Bouveyron, C., Girard, S., 2012. HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. Journal of Statistical Software, 46(6):1-29.

Berkhin, P., 2006. A survey of clustering data mining techniques. (J. Cogan, C. Nicholas, M. Teboulle editör). Grouping multidimensional data, Springer, Berlin, 25-71.

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., When is nearest neighbor meaningful? In: Beeri C, Bune- man P, eds. Database Theory ICDT99. Lecture Notes in Computer Science, Springer, Berlin, 1999:217–235.

Bezdek, J. C., Ehrlich, R., Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2–3):191–203.

Biernacki, C., Celeux, G., Govaert, G., 2003. Choosing starting values for the EM algorithm for getting the highest likehood in multivariate Gaussian mixture models. Computational Statistics and Data Analysis, 41(3–4):561–575.

Bilmes, J. A., 1998. A Gentle Tutorial of the EM Algorithm. International Computer Science Institute, 4(510), 126.

Bouveyron, C., Girard, S., Schmid, C., 2007. High-dimensional data clustering. Computational Statistics and Data Analysis, 52(1):502–519.

Bouveyron, C., Brunet-Saumard, C., (2014). Model-based clustering of high dimensional data: A review. Computational Statistics & Data Analysis, 71:52-78.

Bradley, P. S., Fayyad, U.M., 1998. Refining Initial Points for K-Means Clustering. Appears in Proceedings of the 15th International Conference on Machine Learning (ICML98), Morgan Kaufman, San Francisco, 91–99.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics-Theory and Methods, 3(1):1-27.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Łukasik, S., Żak, S., 2010. Complete gradient clustering algorithm for features analysis of x-ray images. In Information technologies in biomedicine, Springer, Berlin, 231.

Celeux, G., Diebolt, J., 1985. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly, 2(1):73-82.

Celeux, G., Govaert, G., 1992. A classification EM algorithm for clustering and two stochastic versions. Computational Statistics & Data Analysis, 14(3):315-332.

Celeux, G., Govaert, G., 1995. Gaussian parsimonious mixture models. Pattern Recognition, 28(5):781–793.

Celeux, G., Soromenho, G., 1996. An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification, 13(2):195-212.

Çelebi, M. E., 2015. Partitional Clustering Algorithms. Springer International, Switzerland, 415.

Dasgupta, A., Raftery, A. E., 1998. Detecting Features İn Spatial Point Processes With Clutter via Model-Based Clustering. Journal of the American Statistical Association, 93(441):294-302.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2):224-227.

Day, N. E., 1969. Estimating the components of a mixture of normal distributions. Biometrika, 56(3):463-474.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (methodological), 39(1):1-38.

Dunn, J. C., 1974. Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4(1):95-104.

Ester, M., Kriegel, H. P., Sander, J., Xu. X., 1996. A Density-Based Algorithm For Discovering Clusters In Large Spatial Databases With Noise. In Kdd 96(34): 226-231.

Everitt, B. S., Landau, S., Morven, L., Daniel, S., 2011. Cluster Analysis. Wiley, London, 346.

Forina, M., Tiscornia, E., 1982. Pattern-recognition methods in the prediction of Italian olive oil origin by their fatty-acid content. Annali di Chimica, 72(3-4):143-155.

Fowlkes, E. B., Mallows, C. L., 1983. A method for comparing two hierarchical clusterings. Journal of the American statistical association, 78(383):553-569.

Fraley, C., Raftery, A. E., (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal, 41(8): 578–588.

Fraley, C., Raftery, A. E., 1999. MCLUST: Software for model-based cluster analysis. Journal of Classification, 16(2):297-306.

Fraley, C., Raftery, A. E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458):611–631.

Fraley, C., Raftery, A. E. 2007. Model-based methods of classification: using the mclust software in chemometrics. Journal of Statistical Software, 18(6):1-13.

Fraley, C., Raftery A. E., Murphy T. B., Scrucca L., 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering. Classification, and Density Estimation Technical Report, Department of Statistics, University of Washington

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer, New York, 494.

Frühwirth-Schnatter, S., Pyne, S., 2010. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. Biostatistics, 11(2):317–336.

Gan, G., Ma, C., Wu, J., 2007. Data clustering: theory, algorithms, and applications. American Statistical Assocation, Virginal, 466.

Gower, J. C., 1971. A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4):857-871.

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2002. Clustering validity checking methods: part II. ACM Sigmod Record, 31(3):19-27.

Halkidi, M, Vazirgiannis, M., Hennig, C., 2016. Method Independent Indices for Cluster Validation and Estimating the Number of Clusters. (C. Henning, M. Meila, F. Murtagh, R.Rocci editör). Handbook of Cluster Analysis, London, Chapman&Hall, 708-730.

Han, J., Kamber, M., Pei, Jian., 2012. Data Mining: Concepts and Techniques. Elsevier, Waltham, 744.

Hastie, T., Tibshiran, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 745.

Heinz, G., Peterson, L.J., Johnson, R.W., Kerk, C.J., 2003. Exploring Relationships in Body Dimensions. Journal of Statistics Education, 11(2).

Hennig, C., 2015. What are the true clusters? Pattern Recognition Letters, 64:53–62.

Hennig, C., Meila, M., 2016. Cluster Analysis: An Overview. (C. Henning, M. Meila, F. Murtagh, R.Rocci editör). Handbook of Cluster Analysis, London, Chapman&Hall:708-730.

Hennig, C., 2016. Clustering Strategy and Method Selection. (C. Henning, M. Meila, F. Murtagh, R. Rocci editör). Handbook of Cluster Analysis, London, Chapman&Hall:708-730.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3):283–304.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification, 2(1):193-218.

Jaccard, C., 1963. Thermoelectric effects in ice crystals. Zeitschrift für Physik B Condensed Matter, 1(2):143-151.

Jacques, J., Preda, C., 2014. Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis, 71:92-106.

Jain, A. K., Murty, M. N., Flynn, P. J., 1999. Data clustering: a review. ACM computing surveys (CSUR), 31(3):264-323.

James, G., Witten, D., Hastie, T., Tibshirani, R. 2013. An Introduction to Statistical Learning. New York, Springer, 426.

Jamshidian, M., Jennrich, R. I., 1993. Conjugate gradient acceleration of the EM algorithm. Journal of the American Statistical Association, 88(421):221-228.

Karypis, G., Han, E. H. S., Kumar, V. 1999. Chameleon: Hierarchical Clustering Using Dynamic Modeling. Computer, 32(8):68–75

Kaufman, L., Rousseeuw, P. J., 2005. Finding Groups in Data: An Introduction to Cluster Analysis. New Jersey, Wiley, 354.

Lloyd, S. P., 1957. Least squares quantization in PCM's Bell Telephone Labs.

Lichman, M., 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Maitra, R., & Melnykov, V. (2010). Simulating Data to Study Performance of Finite Mixture Modeling and Clustering Algorithms. Journal of Computational and Graphical Statistics, 19(2):354–376.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1(14):281-297.

Maitra, R., 2009. Initializing Partition-Optimization Algorithms. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 6(1):144-157.

Martinez, W. L., Martinez, A. R., Martinez, A., Solka, J., 2010. Exploratory Data Analysis with MATLAB. CRC Press, London:536.

Maugis, C., Celeux, G., & Martin-Magniette, M. L. (2009). Variable selection for clustering with Gaussian mixture models. Biometrics:65(3), 701-709.

McLachlan, G. J., Rathnayake, S.I., 2016. Mixture Models for Standart p-Dimensional Euclidean Data. (C. Henning, M. Meila, F. Murtagh, R. Rocci editör). Handbook of Cluster Analysis, London, Chapman&Hall:145-171.

McNicholas, P. D., 2016, Mixture Model Based Classification. Chapman and Hall, Boca Raton, 236.

Melnykov, V., 2016. Model-based biclustering of clickstream data. Computational Statistics & Data Analysis, 93:31-45.

Melnykov, V., & Maitra, R., 2010. Finite mixture models and model-based clustering. Statistics Surveys, 4(80):80-116.

Melnykov, V., Melnykov, I., 2012. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. Computational Statistics & Data Analysis, 56(6):1381-1395.

Milligan, G. W., Sokol, L. M., 1980. A Two-Stage Clustering Algorithm with Robust Recovery Characteristics. Educational and Psychological Measurement, 40(3):755–759.

Milligan, G. W., Cooper, M. C., 1988. A study of standardization of variables in cluster analysis. Journal of Classification, 5(2):181-204.

Milligan, G. W., 1996. Clustering Validation: Results And İmplications For Applied Analysis. (P. Arabie, L. J. Hubert, G. De Soete Editör). Clustering and Classification, World Scientific Publishing, Danvers:341–375.

Mirkin, B. 2011. Choosing the number of clusters. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3):252–260.

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification, 31(3): 274–295.

Pearson, K., 1894. Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London, 185:71-110.

Pelleg, D., Moore, A., 2000. X-means. CEUR Workshop Proceedings, 1542:33–36.

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Dean, N., (2006). Variable selection for model-based clustering. Journal of the American Statistical Association, 101(473):168-178.

Rand, W. M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336):846-850.

Reddy, C. K., 2013. A Survey of Partitional and Hierarchical Clustering Algorithms. (C. C. Aggarwal, C. K. Reddy editör). Data Clustering, Taylor & Francis, London, 87-109.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals Of Statistics, 6(2):461-464.

Scrucca, L., Raftery, A. E., (2014). clustvarsel: A package implementing variable selection for model-based clustering in R. arXiv preprint arXiv:1411.0606.

Sigillito, V.G., Wing, S.P., Hutton, L.V., Baker, K. B. 1989. Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest, 10(3):262-266.

Sneath, A., Sokal, R. R., 1963. Principles of Numerical Taxonomy. San Francisco, W.H. Freeman and Company, 963.

Steinley, D., 2006. K-means clustering: A half-century synthesis. British Journal of Mathematical and Statistical Psychology, 59(1), 1–34.

Tan, P. N., Steinbach, M., Kumar, V., 2005. Introduction to Data Mining. Boston, Pearson, 769.

Tibshirani, R., Walther, G., Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 63(2):411–423.

Vicari, D., Alfó, M., 2014. Model based clustering of customer choice data. Computational Statistics and Data Analysis, 71.

Vu, D. Q., Hunter, D. R., Schweinberger, M. 2013. Model-based clustering of large networks. The Annals of Applied Statistics, 7(2):1010.

Ward, J.H., 1963, Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58:236–244.

Wehrens, R., Buydens, L. M., Fraley, C., Raftery, A. E. 2004. Model-based clustering for image segmentation and large datasets via sampling. Journal of Classification, 21(2):231-253.

Wei, G. C., Tanner, M. A., 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association, 85(411):699-704.

Weisberg, S., 2005. Applied Linear Regression. John Wiley & Sons, 310.

Wolfe, J. H., 1965. A computer program for the maximum likelihood analysis of types. Naval Personnel Research Actıvıty, San Dıego.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, 5(3):329-350.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. 2008. Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1):1-37

Xu, L., 1997. Bayesian Ying–Yang machine, clustering and number of clusters. Pattern Recognition Letters, 18(11):1167-1178.

Xu, R., Wunsch, D., 2008., Clustering. Wiley, New Jersey, 368.

Zhang, H., Huang, Y., (2015). Finite Mixture Models and Their Applications: A Review. Austin Biom and Biostat, 2(2):1-6.

ÖZGEÇMİŞ

1989 yılında İstanbul'da doğdu. İlk, orta ve lise öğrenimini Adana'da tamamladı. 2010 yılında Çukurova Üniversitesi İstatistik Bölümünde başladığı lisans eğitimini 2014 yılında tamamladı. 2015 yılında Çukurova Üniversitesi İstatistik Bölümünde yüksek lisansa başladı