T.C.
NECMETTİN ERBAKAN
ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KÜMELEME ANALİZİ YÖNTEMLERİNİN KÜME GEÇERLİLİK İNDEKSLERİNE GÖRE KARŞILAŞTIRILMASI

Derya ALKIN

YÜKSEK LİSANS TEZİ

İstatistik Anabilim Dalı

Ocak-2019
KONYA

TEZ KABUL VE ONAYI

Derya ALKIN tarafından hazırlanan "**KÜMELEME ANALİZİ YÖNTEMLERİNİN KÜME GEÇERLİLİK İNDEKSLERİNE GÖRE KARŞILAŞTIRILMASI**" adlı tez çalışması 18/01/2019 tarihinde aşağıdaki jüri tarafından oy birliği ile Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü İSTATİSTİK Anabilim Dalı'nda YÜKSEK LİSANS tezi olarak kabul edilmiştir.

Jüri Üyeleri

Başkan
Doç. Dr. Murat ERİŞOĞLU

……..

Danışman
Dr. Öğr. Üyesi Aydın KARAKOCA

……..

Üye
Dr. Öğr. Üyesi Serkan AKOĞUL

……..

Yukarıdaki sonucu onaylarım.

Prof. Dr. Ahmet AVCI
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

İmza

Derya ALKIN

Tarih:

ÖZET

YÜKSEK LİSANS TEZİ

KÜMELEME ANALİZİ YÖNTEMLERİNİN KÜME GEÇERLİLİK İNDEKSLERİNE GÖRE KARŞILAŞTIRILMASI

Derya ALKIN

Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Aydın KARAKOCA

2019, 70 Sayfa

Jüri
Doç.Dr. Murat ERİŞOĞLU
Dr. Öğr. Üyesi Aydın KARAKOCA
Dr. Öğr. Üyesi Serkan AKOĞUL

Bu tez çalışmasında kümeleme analizinde kullanılan aşamalı ve aşamalı olmayan kümeleme yöntemleri içsel ve dışsal küme geçerlilik indeksleri kullanılarak küme sayılarına göre karşılaştırılmıştır. Farklı veri setleri üzerinde yapılan uygulamalar sonucunda ele alınan 26 içsel kriter ile aşamalı olmayan yöntemlerden k-ortalamalar tekniği ve aşamalı kümeleme yöntemlerinden tek bağlantı kümeleme yöntemi, tam bağlantı kümeleme yöntemi, ortalama bağlantı kümeleme yöntemi, ağırlıklı ortalama bağlantı kümeleme yöntemi, merkezi bağlantı kümeleme yöntemi, medyan bağlantı kümeleme yöntemi ve Ward bağlantı kümeleme yöntemi için elde edilen sonuçlar yorumlanmıştır.

Anahtar Kelimeler: Küme Değerlendirme Kriterleri, Kümeleme Analizi, Uzaklık Ölçüleri

ABSTRACT

MS THESIS

COMPARISON OF CLUSTERING ANALYSIS METHODS ACCORDING TO CLUSTER VALIDATION INDEXES

Derya ALKIN

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF NECMETTİN ERBAKAN UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE IN STATISTICS

Advisor: Dr. Associate Aydın KARAKOCA

2019, 70 Pages

Jury
Associate Prof. Dr. Murat ERİŞOĞLU
Dr. Associate Aydın KARAKOCA
Dr. Associate Serkan AKOĞUL

In this dissertation, the hierarchical and non-hierarchical clustering methods, which are used in clustering analysis, have been compared by using the clustering indices. As a result of the applications performed on different datasets, 26 internal criteria and non-hierarchical methods, k-means technique and hierarchical clustering methods, single link clustering method, full link clustering method, average link cluster method, weighted average link aggregation method, central link aggregation method, median link aggregation method and ward link aggregation method have been interpreted.

Keywords: Cluster Evaluation Criteria, Clustering Analysis, Distance Measures

ÖNSÖZ

Bu tez çalışmamda benden zaman ve emeğini esirgemeyen değerli hocam Dr. Öğr. Üyesi Aydın KARAKOCA'ya teşekkürlerimi sunarım. Ayrıca bu süreçte yanımda olan eşim Ruhi Can ALKIN ve her zaman manevi olarak yanımda olan annem Şerife ARSLAN'a teşekkürü bir borç bilirim.

Derya ALKIN
KONYA-2019

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Simgeler

k : Küme sayısı

d_{ij} : i. ve j. Gözlemin birbirine uzaklıkları

x_{ik} : i. gözlemin k. değişken değeri

p : Değişken sayısı

$\bar{x}_i$: i' inci gözlem üzerinden ölçülen tüm p değişken değerlerinin ortalaması

G : Gözlenen sıklık

B : Beklenen sıklık

D_{ij} : Mahalonobis uzaklığı

μ_i : i. grubun ortalama vektörü

μ_j : j. grubun ortalama vektörü

S^{-1} : Ortak varyans-kovaryans matrisinin tersi

μ_{ik} : i. gruptaki/kümedeki k. değişkenin ortalaması

S_k : k. değişkenin varyansı

n_i : i. kümedeki birim sayısı

$\bar{x}_{K_i}$: K_i küme merkezleridir

$\mu(K_i)$: $K_i$küme merkezini

$d(X, \mu(K_i))$: X veri matrisindeki verilerin K_i küme merkezlerinin ortalamalarına uzaklıkları

D : uzaklıkların matrisi

$\bar{d}$: D matrisindeki d_{ij} uzaklıklarının ortalaması

R : Bağlantı uzaklıklarını içeren matris

r_{ij} : R matrisindeki i. ve j. birimler arasındaki uzaklık

$\bar{r}$: R matrisindeki r_{ij} uzaklıklarının ortalaması

W : Kümeler içi çarpımlar ve kareler matrisi

|W| : Kümeler içi kareler ve çarpımlar toplamı matrisinin determinantı

|T| : Genel kareler ve çarpımlar toplamı matrisinin determinantı

T : Genel kareler matrisi

S_W : Her küme içindeki tüm nokta çiftleri arasındaki mesafenin toplamıdır.

S_{min} : Kümeler içindeki en küçük mesafelerin toplamıdır

S_{max} : Kümeler içindeki en büyük mesafelerin toplamıdır

N_W : Aynı küme içindeki ilgili noktalar arasındaki mesafelerin toplam sayısı

S_B : Kümeler arası uzaklıklar toplamı

N_B : Kümeler arası ilgili noktalar arasındaki mesafelerin toplamı

D_B : Küme merkezleri arasındaki en büyük uzaklıktır

E_W : Noktaların kendi küme merkezlerine uzaklıklarının toplamı

E_T : Tüm noktaların küme merkezine uzaklıklarının toplamı

M : Küme içinden bir nokta

R_i : i. ve diğer kümeler arasındaki maksimum karşılaştırma oranı

e_i : i. kümedeki gözlemlerin kendi küme merkezlerine olan uzaklıklarının ortalaması

S_n **:** Standart sapma

S^+ : Aynı kümede yer alan gözlemler arasındaki uzaklıkların sayısı

S^- : Aynı kümede yer almayan gözlemler arasındaki uzaklıkların sayısı

Kısaltmalar

WG : kümeler içi kareler matrisi

GAKT : Gruplar arası kareler toplamı

GİKT : Gruplar içi kareler toplamı

GİDM : Gruplar içi dağılım matrisi

GADM: Gruplar arası dağılım matrisi

$GAKT_j$: Gruplar arası kareler matrisinin j. köşegen terimidir

GKT_j : Genel kareler matrisinin j. köşegen terimidir

TABLOLAR DİZİNİ

1. GİRİŞ

Kümeleme analizi veri setindeki benzer özellikteki birimlerin yer aldığı yapılanmanın belirlenmeye çalışıldığı bir analiz türüdür. Benzer yapıdaki birimlerin bir kümede yer alması birçok alanda daha kullanışlı bir veri seti oluşmasını sağlar. Artan dünya nüfusu, gelişen teknolojiler, büyüyen ekonomiler gibi nedenlerle veri sayısında ciddi bir artış meydana gelmiştir. Artan veri sayısı ile birlikte bunların analizi, hesaplanması, verilerin etkin kullanımı, veri setlerinden hangi değişkenlerin kullanılacağına karar verilmesi gibi nedenlerle veri setlerinde boyut indirgemeye ihtiyaç duyulmuştur. Bu da kümeleme analizinin ortaya çıkmasına ve gelişimine katkı sağlamıştır.

Kümeleme analizi boyut indirgeme, doğal grupların tanımlanması, aşırı değerleri saptama, verileri alt gruplara ayırma gibi amaçlar için kullanılır. Kümeleme analizinin tıp, sosyoloji, psikoloji, eğitim bilimleri, biyoloji, veri madenciliği gibi birçok alanda uygulamaları mevcuttur. Tıpta; hastalıkların ya da semptomlara göre tedavilerin sınıflandırılması, hastalıkların sınıflandırılarak tanısının konması gibi alanlarda kullanılır. Örneğin, parkinson hastalığının erken dönemlerinde ve geç dönemlerinde bir takım bulguların olup olmadığı, bu hastalığa sahip hastalardan elde edilen veri seti üzerinde Ward kümeleme yöntemi kullanılarak test edilmiştir (Uribe ve ark., 2018). Astım hastaları üzerinde yapılan bir çalışmada ise 3 fenotip astım belirlenmiştir (Sendín-Hernández ve ark., 2017). Ekonomik gelişmelerle birlikte bu alanda da veri setlerinin daha kolay analiz edilmesi için kümeleme analizi kullanılmıştır. Örneğin bir şirketin beş yıllık tahvillerine kümeleme analizi uygulanarak gelecek yıllarla ilgili tahminlerde bulunulmuştur (Ramon-Gonen & Gelbard, 2017). Başka bir çalışmada da yakıt tüketiminde etkili özelliklerin seçimi için korelasyon analizi ile özellik boyutunu indirgemek amacıyla temel bileşenler analizi uygulanmıştır. Çalışmada araçlar değişik yollarda sürülerek test edilmiştir. Kümeleme analizi ile de yakıt tüketimine göre sürüş yapılan yollar kümelenmiştir (Xie ve ark., 2017). Veri madenciliğinde boyut indirgeme ve özellikle bilişim sistemlerinde yüz tanımlama gibi uygulamalarda, bilginin gruplanarak erişimi gibi birçok alanda kümeleme analizi kullanılmıştır. Örneğin; yapılan bir çalışma büyük veri setlerinin yerine veri setlerinde belli sınırlar belirleyerek küçük veri setleriyle belirlenen sınırlara göre kümeleme yapılmasını önermektedir (Xiu Li & Yuan, 2017). Veri tiplerine göre hangi kümeleme algoritmalarının uygulanması gerektiği önerilen başka bir çalışmada da küçük veri setleri kullanılarak büyük veri setlerinin veri

tipine göre kümeleme yönteminin seçimi kolaylaştırılmaya çalışılmıştır. (Dash & Misra, 2017). Fotoğrafçılık alanında yapılan bir çalışmada fotoğraf pikselleri kümelenerek bir filtreleme algoritmasıyla pikseller yok edilmiştir (Khan ve ark., 2015).

Kümeleme analizi ile küme içinde benzer, kümeler arasında farklı veri yapısına ulaşılmak amaçlanmaktadır. Bunun için uzaklık ölçülerinden faydalanılır. Kümeleme analizinin temelini uzaklık ölçüleri oluşturur. Veri tipine göre farklı uzaklık ölçüleri kullanılmaktadır. En çok kullanılan uzaklık ölçüsü Öklid uzaklık ölçüsüdür. Bu çalışmada verinin sürekli veya kesikli olması, sıklık sayıları olması, ikili sınıflı olması, küme, örneklem, grup olması yani belli özelliklere sahip olması durumunda kullanılacak benzerlik ölçüleri 3. bölümde 'Uzaklık ve Benzerlik Ölçülerinin Veri Tipine Göre Sınıflandırılması' başlığı altında verilmiştir.

Kümeleme analizi kendi içinde aşamalı ve aşamalı olmayan kümeleme yöntemleri olarak ikiye ayrılır. Aşamalı kümeleme yöntemleri küme sayısının önceden bilinmemesi durumunda kullanılan yöntemlerdir. Çalışmada aşamalı kümeleme tekniklerinden tek bağlantı kümeleme tekniği, tam bağlantı kümeleme tekniği, ortalama bağlantı kümeleme tekniği, ağırlıklı ortalama bağlantı kümeleme tekniği, merkezi bağlantı kümeleme tekniği, medyan bağlantı kümeleme tekniği, ward bağlantı kümeleme tekniği, esnek beta bağlantı tekniği 4. bölümde 'Kümeleme Yöntemleri' başlığı altında anlatılmıştır. Aşamalı olmayan kümeleme yöntemleri ise küme sayısı hakkında bilgi sahibi olunması durumunda kullanılan yöntemlerdir. En çok kullanılan aşamalı olmayan kümeleme tekniği olan k-ortalamalar tekniğine çalışmada yer verilmiştir.

Kümeleme analizinde kullanılan yönteme veya uzaklık ölçüsüne bağlı olarak değişik küme sayıları aynı veri seti için elde edilmektedir. Bu sebeple hangi veri setinde kaç küme olduğunun belirlenmesi kümeleme analizinin temel problemlerinden birisidir. Uygun küme sayısının belirlenmesinde küme geçerlilik indeksleri önemli rol oynamaktadır. Küme geçerlilik yöntemleri içsel ve dışsal kriterler olmak üzere iki ana başlıkta toplanmaktadır. Dışsal kriterler şunlardır: Rand indeksi, Jaccard katsayısı, Fowlkes ve Mallows indeksi, Γ istatistiği, düzeltilmiş Rand indeksi, Czekanowski-Dice indeksi, Kulczyns indeksi, McNemar indeksi, Phi indeksi, Rogers-Tanimoto indeksi, Russel-Rao indeksi ve Sokal-Sneat indeksi. İçsel kriterler ise şunlardır; kofenetik korelasyon katsayısı, Wilks'in lambda test istatistiği, Ball-Hall indeksi, Banfeld-Raftery indeksi, C indeksi, det-oran indeksi, Baker-Hubert gamma indeksi, ksq_detw indeksi, log_det_oran indeksi, log_ss_oran indeksi, McClain-Rao indeksi, PBM indeksi, nokta-iki serili indeks, Ratkowski-Lance indeksi, Ray-Turi indeksi, Scott-Symons indeksi, SD

indeksi, s_dbw indeksi, Tau indeksi, iz_w indeksi, iz_wib indeksi, Wemmert-Gançarski indeksi, Calinski ve Harabasz indeksi, Davies-Bouldin indeksi ve Dunn indeksidir. Çalışmada küme geçerlilik yöntemleri geniş kapsamlı olarak 5. bölümde ‘Kümelerin Değerlendirilmesi’ başlığı altında anlatılmaktadır.

Çalışmanın son bölümünde 10 gerçek veri seti üzerinde kümeleme analizi yöntemleri gerçekleştirilecek ve analizin sonuçları küme geçerlilik indekslerine göre değerlendirilecektir.

2. KAYNAK ARAŞTIRMASI

Veri kümeleme analizi kavram olarak ilk kez Tryon (1939) tarafından kullanılmıştır. Clements (1954), antropolojik verinin kümelenmesi çalışmasında kümeleme analizi kavramını ilk kez kullanmıştır. Cox (1957) ve Fisher (1958), kümeleme analizi kavramını gruplama adını verdikleri çalışmalarında kullanmışlardır. Sokal ve Sneath (1963), sayısal sınıflandırmanın prensipleri adını verdikleri kitaplarında kümeleme analizinden bahsetmişlerdir.

Kümeleme en temel anlamda benzer birimlerin belirlenmesi problemidir. Dolayısıyla kümeleme tekniği ne olursa olsun benzerliğin belirlenmesi oldukça önemlidir. Benzerliğin belirlenmesinde kullanılan uzaklık ölçülerinin seçimi kümelemenin başarısında önemli bir etkendir. Uzaklık ölçüleri ve uzaklık ölçülerinin seçimi ile ilgili literatürde yer alan çalışmalardan bazıları aşağıda verilmiştir.

Cronbach ve Gleser (1953) değişkenler arasında ilişki olması durumunda, analiz öncesi uygulanan temel bileşenler analizinden elde edilen birim varyanslı skorlar arasındaki karesel Öklid uzaklığı, Mahalonobis uzaklığı, Kendall'ın parametrik olmayan uzaklığı, Manhatan City-Block uzaklığının karesi ve Gower'ın logaritmik uzaklık ölçüsü arasındaki ilişkileri teorik olarak ortaya koymuşlardır.

Green ve Rao (1969), kümeleme analizinde kullanılan on farklı uzaklık ölçüsünü teorik ve uygulamalı olarak karşılaştırmışlardır. Çalışmalarında 15 farklı marka bilgisayarın 6 özelliğine göre oluşturulan veri setinde farklı uzaklık ölçüleri ile hesaplanan uzaklıklar arasındaki korelasyonları incelemişlerdir. Korelasyon değerleri incelendiğinde diğer uzaklık ölçülerinden en farklı uzaklık ölçüsünün parametrik olmayan Kendall uzaklık ölçüsü olduğu görülmüştür.

Roussos ve ark. (1998) çok boyutluluğun belirlenmesi için aşamalı kümeleme analizinde yeni bir benzerlik ölçüsünün kullanılması isimli çalışmalarında iki ve üç boyutlu olarak ürettikleri veri setlerinde koşullu kovaryans değerini temel alarak geliştirdikleri uzaklık ölçüsünün tek bağlantı, tam bağlantı, ortalama bağlantı ve ağırlıklı ortalama bağlantı aşamalı kümeleme tekniklerindeki performansı incelemişlerdir.

Finch ve Huynh (2000), iki değer alabilen değişkenlerle kümeleme analizi yapıldığında kullanılan Dice, Jaccard, eşleştirme ve Russel/Rao benzerlik ölçülerini karşılaştırmışlardır. Karşılaştırma sonucunda iki değer alan değişkenlerden oluşan veri setlerinde verinin iki kümeye ayrılması durumunda ele aldıkları benzerlik ölçülerinin benzer kümeleme performansı gösterdiğini ifade etmişlerdir.

Gunjaca ve ark (2000), bitki sınıflamasında kullanılan kümeleme analizinde farklı uzaklık ölçülerinin kullanılması durumunda farklı sonuçlar elde edileceğini göstererek analiz öncesi uygun uzaklık ölçüsünün seçiminin önemli olduğunu belirtmişlerdir. Çalışmalarında üç farklı türden 123 fasulyenin 17 nitel ve 9 nicel değişken bakımından gözlemlenmesi ile oluşan veri setini kullanmışlardır. Nitel ve nicel değişkenlerin bir arada bulunduğu bu veri seti için Gower (1971), Peeters-Martinelli (1989) ile Cole-Rodgers ve ark (1997) önerdikleri uzaklık ölçülerini kullanarak kümeleme analizi gerçekleştirmişlerdir. Uzaklık ölçülerine bağlı olarak kümeleme sonucunda farklı büyüklük ve yapıda kümelerin oluştuğunu göstererek, karma veri setleri için Cole-Rodgers ve ark (1997) tarafından önerilen uzaklık ölçüsünün en uygun uzaklık ölçüsü olduğunu vurgulamışlardır.

Mimmack ve ark (2001), iklim biliminde k-ortalamalar kümeleme algoritmasında yaygın bir kullanıma sahip olan Öklid uzaklığına alternatif olarak Mahalonobis uzaklığını önermişlerdir. Çalışmalarında 1961-1990 yılları arasında Güney Afrika ve Lesotho'da bulunan 517 istasyondan elde edilen metre kareye düşen aylık yağış miktarlarıyla ilgili veri setinin k -ortalamalar tekniği ile kümelenmesinde Mahalonobis uzaklığının Öklid uzaklığı kadar iyi sonuç verdiğini göstermişlerdir.

Jie ve ark (2006), k-ortalamalar tekniğinde birimlerin küme merkezlerine olan Öklid ve Manhattan uzaklıklarının hesaplanmasında aritmetik ortalama yerine geometrik ortalama, medyan ve harmonik ortalama gibi farklı ortalamaların kullanılması durumunda bu uzaklıkların kümeleme performansına olan etkilerini incelemişlerdir. Çalışma sonunda aritmetik ortalama yerine farklı ortalamalar kullanılarak elde edilen uzaklık ölçülerinin de klasik uzaklık ölçüleri kadar iyi performans gösterdiklerini vurgulamışlardır.

Pun ve Ali (2007), k-ortalamalar kümeleme algoritması için tek uzaklık ölçüsü yaklaşımı isimli çalışmalarında k-ortalamalar kümeleme algoritmasında "Hangi uzaklık ölçütü kullanılmalıdır ?" sorusuna cevap aramışlardır. Çalışmalarında farklı özelliklere sahip 112 veri seti bilinen gerçek küme sayılarına göre sırasıyla City-Block, Öklid ve karesel Öklid uzaklığı kullanılarak k-ortalamalar tekniğiyle kümelenmiş ve doğru sınıflandırma yüzdeleri hesaplanmıştır. Çalışma sonucunda entropiye dayanarak hesaplanan eşik değerinin 8.4595'den büyük olması durumunda k-ortalamalar algoritmasında uzaklık ölçüsü olarak City-Block uzaklığını, eşik değerinin 8.4595'e eşit veya daha küçük olması durumunda ise karesel Öklid uzaklığını kullanmayı önermişlerdir.

Vimal ve ark (2008), kümeleme analizinde en yaygın kullanılan Öklid uzaklık ölçüsünü, bilgisayar programlama algoritmalarında kullanılan bit-vektör uzaklığı, karşılaştırmalı kümeleme uzaklığı ve Huffman kod uzaklığı ile karşılaştırmışlardır. İlgili uzaklık ölçülerinin k-ortalamalar, yoğunluk tabanlı ve üstünlük-tabanlı kümeleme algoritmalarındaki etkinliklerini, sentetik veri setleri ve gerçek kriket verisini kullanarak incelemişlerdir. Çalışma sonucunda, tüm algoritmalarda Öklid uzaklığı en iyi kümeleme performansı gösteren uzaklık ölçüsü olarak belirlenmiştir.

Aşamalı olmayan kümeleme teknikleri veya başka bir deyişle parçalamaya dayalı kümeleme yöntemleri içerisinde en çok bilinen ve kullanılan k–ortalamalar tekniği farklı bilimsel alanlarda birbirlerinden bağımsız bir şekilde Steinhaus (1955), Lloyd (1957), Ball ve Hall (1965) ile MacQueen (1967)'in çalışmalarında önerilmiştir. MacQueen (1967)'in çalışmasında "k-ortalamalar" kavramını kullandığı için çoğu kaynakta algoritmayı ilk öneren olarak onun ismi geçmektedir. Ayrıca Hartigan ve Wong (1979)'un k−ortalamalar tekniği ile ilgili çalışması yöntemin gelişimine önemli katkılar sağlamıştır. Günümüzden yaklaşık 50-55 yıl öncesinde önerilen k-ortalamalar algoritması basitlik, uygulanabilirlik, deneysel çalışmalardaki başarısı ve etkinliği nedeniyle hala en çok tercih edilen kümeleme yöntemidir.

Küme geçerlilik indeksleri kümeleme sonuçlarını test etmek amacıyla geliştirilmiş yöntemlerdir. Son yıllarda yapılan çalışmalara bakacak olursak Haouas, F. (2017) yaptığı çalışmada uygun küme sayısını bulmak için bir küme geçerlilik indeksi sunmaktadır. Önerilen HF indeksi küme üyelik numarasına bağlıdır. Bu indeks genelleşmiş geçici bir cezalandırma teriminin küme merkezi ile ilişkilendirilir ve küme merkezlerinin birbirinden ayrılması için küme başına ortalama veri sayısı ile küme merkezlerinin çarpılmasıyla elde edilir. Buna göre en uygun küme sayısı HF indeksinin minimumuna karşılık gelir. Huang, K. Y., (2017) çalışmasında Mevcut Çoklu-Öznitelik Karar Verme (MADM) yönteminin performansını arttırmak için bir küme geçerlilik indeksi önermiştir. Önerilen dizin tabanlı indeks, Fuzzy Set (FS), Rough Set (RS) ve FRM-indeks yöntemi olarak adlandırılmaktadır. Bu yöntem sadece çoklu veri kümeleri için karar verme kurallarının çıkarılmasında daha güvenilir bir temel sağlamasını değil, aynı zamanda belirsizliği ortadan kaldırıp daha etkili bir Mevcut Çoklu-Öznitelik Karar Verme sisteminin inşa edilmesinin kolaylaşmasını sağladığını da göstermektedir. Kim, B. (2018) çalışmasında Charnes, Cooper & Rhodes kümelenme geçerliliği olarak adlandırılan yeni bir kümeleme geçerlilik indeksi olarak önerdiği indeks, kümeleme sonuçlarının kalitesini ölçmek için geliştirilmiştir. 12 sentetik ve 30 gerçek veri kümesi

üzerinde yapılan deneysel sonuçlara dayanarak, önerilen küme geçerlilik indeksi, bilinen geçerlilik indeksleri ile karşılaştırıldığında optimal ve makul kümelenme yapılarını belirleme konusunda üstün yetenek göstermektedir. Kumar, V., (2017) uygun küme sayısının tahmini için gerçek ve gerçek olmayan veriler üzerinde k-ortalamalar, bulanık-C ortalamalar ve modifiye uyum arama tabanlı kümeleme yöntemlerini uygulanarak elde edilen kümelenmeleri simetriye dayalı 7 (yaygın kullanılan) küme geçerlilik indeksi ile test etmiştir. Bu sonuçlar, simetriye dayalı mesafenin dahil edilmesiyle uygun sayıda kümenin bulunmasında, mevcut geçerlilik indekslerinin yeteneklerini geliştirdiğini ortaya koymaktadır. Lee, S. H., (2018) yaptığı çalışmada büyük boyutlu veri setlerinde gerçekleştirilen kümeleme analizi sonuçlarını test eden mevcut küme geçerlilik indekslerinin, rastgele kümelenmelere ve aykırı değerlere duyarlı olmasından dolayı 'destek vektör veri açıklama indeksini' önermiştir. Özellikle rastgele kümelenmeleri test etmede başarılı olduğunu yaptığı çalışmada belirtmiştir. Ben Said, A. Ve ark (2017). (2018) çalışmasında küme merkezleri arasındaki mesafeyi baz alarak elde edilen kümeleme analizi sonuçlarının doğruluğunun tartışılabilir olmasından dolayı yeni bir küme geçerlilik indeksi olarak Jeffrey indeksini önermiştir. Çalışmasında farklı tipteki veri setlerine, yaygın kullanılan küme geçerlilik indekslerini uygulamış ve Jeffrey indeksinin yüksek performans gösterdiğini gözlemlemiştir. Cheruku, R. Ve ark. (2017) diyabet sınıflandırması için radyal temelli fonksiyon sinir ağlarını kullanmıştır. Radyal temelli fonksiyon sinir ağları sırasıyla girdi katmanı, desen katmanı, toplama katmanı ve karar katmanı ile dört katmanlı ileri besleme ağıdır. Desen katmanındaki nöronların optimal sayısını belirlemek için sınıflama şekline göre küme geçerlilik indeksi kullandığını belirtmiştir. Toplama katmanı ve desen katmanı arasındaki ağırlıkları tanımlamak için yarasadan ilham alan eniyileme algoritması için yeni bir dışbükey uygunluk işlevi de tasarlanmıştır. Radyal temel fonksiyon sinir ağı için önerilen model Pima Indians Diabetes veri seti ve sentetik veri setleri üzerinde test edilmiştir. Deneysel sonuçlar, yaklaşımın doğruluk, duyarlılık, özgüllük, sınıflandırma süresi, eğitim süresi, ağ karmaşıklığı ve geleneksel radyal temel işlevli sinir ağına kıyasla hesaplama süresi açısından daha iyi olduğunu kanıtlamıştır. Lin, P. L. (2017) yaptığı çalışmada Gaussian dağıtılmış kümeler için dağılım ve örtüşme denilen iki ölçmeye dayanan yeni bir küme geçerlilik indeksi olarak dağılım ölçümünü önermiştir. Dağılım ölçümünü bir kümede yayılan verilerin durumunu tahmin etmek için kullandığını belirtmiştir. Bir küme için dağılım ölçüsü, veri noktalarının bu kümeye yakından nasıl dağıldığı ifade eder. Örtüşme ölçüsü ise veri kümesindeki herhangi bir çift küme arasındaki çakışma derecesini temsil

eder. Dağılma ve örtüşme ölçüsünü birleştirerek, çok etkili yeni bir küme geçerlilik indeksi elde edilmiştir. Bu çalışmada sekiz sentetik veri seti ve dört gerçek veri seti kullanarak geçerlilik indeksinin etkinliğini göstermek için çeşitli deneyler yapılmıştır. On veri seti için yeni indeksin doğruluk açısından en iyi performansı verdiği test sonucu görülmüştür. Luna-Romera, J. M. (2018)'e göre az hesaplama süresinde büyük miktarda veriyi işleyen iki küme geçerlilik indeksi önermişlerdir. İndeksler, küme içi mesafe hesaplamalarını basitleştirerek geleneksel indekslerin yeniden tanımlanmasına dayanmaktadır. Önerilen indekslerin performansını analiz etmek için 28 sentetik veri setinde iki tür test yapılmıştır. İlk olarak, indekslerin geleneksel olanlara benzer bir etkiye sahip olduğunu doğrulamak için indeksler, küçük ve orta büyüklükteki veri setleriyle test edilmiştir. Daha sonra, etkinliklerini kontrol etmek için 20 değişkenli 11 milyon girdili veri seti test edilmiştir. Sonuçta, her iki indeksin de kullanılan geleneksel indekslere benzer bir etki ile çok az bir hesaplama zamanında büyük veriyi doğru test edebildiği görülmüştür. Nawrin, S. (2017) çalışmasında trafik yönetim sisteminde yer alan verileri k-ortalamalar yöntemi ile sınıflandırmış, daha sonra küme geçerlilik indeksleriyle test etmiştir. Dunn indeksinin k-ortalamalar için en iyi sonucu verdiğini gözlemlemiştir. Singh, M. (2017) çalışmasında patolojideki radyolojik görüntülerin sınıflandırmasıyla ilgilenmiştir. Geliştirilen algoritma, bulanık c-ortalama algoritması ile elde edilen segmentasyon sonuçlarına uygulanmış ve Xie-Beni indeksi, klasik geçerlilik indeksleri ile karşılaştırılmıştır. Sonuçlar, önerilen yöntemin, yoğun görüntüler üzerinde uygun sayıda kümeyi bulma yeteneğinin daha iyi olduğunu belirtmiştir. Starczewski, A. (2017) çalışmasında STR adını verdiği yeni bir küme geçerlilik indeksi önermiştir. Önerdiği indeksi kümeleme sürecinde, kümelemelerin yoğunluk ve ayrılabilirlik değişikliklerini belirleyen iki bileşenin ürünü olarak tanımlamaktadır. Bu indeksin maksimum değeri, en iyi kümeleme şeklini tanımlar. Starczewski, A. ve A. Krzyzak (2017) çalışmalarını gerçek veri seti üzerinde tam bağlantı kümeleme yöntemi, beklenti maksimizasyonu ve k-algoritmalar uygulanarak elde edilen kümeleme sonuçlarını karşılaştırmışlar ve kümeleme sonuçlarını en iyi değerlendiren STR adını verdikleri kümeleme indeksini belirlemişlerdir. Wani, M. A. ve R. Riyaz (2017) çalışmalarında gen diziliminin kümelenme doğrulaması amacıyla yeni bir kümeleme geçerlilik indeksi olan AR (Puan) indeksini önermişlerdir. Kümelenmelerin yoğunluk ölçüsünü ve farklılık ölçüsünü belirlemek için yeni bir yaklaşım sunulmuştur. Yaygın olarak bilinen indeksleri yeniden gözden geçirmişler ve bu indekslerin önerilen indeksle tam bir karşılaştırmasını yapıp farklı indekslerin performans değerlendirmesinin bir özetini sunmuşlardır. Deneysel

sonuçlar, önerdikleri indeksin genel olarak bilinen kümelenme geçerlilik indekslerinden daha iyi performans gösterdiğini belirtmişlerdir.

3. UZAKLIK ÖLÇÜLERİNİN VERİ TİPİNE GÖRE SINIFLANDIRILMASI

Kümeleme analizinin temelini uzaklık ölçüleri oluşturmaktadır. Kümeleme analizinin amacı birbirine en yakın yani en benzer gözlemlerin oluşturduğu küme yapısını ortaya koymaktır. Küme içindeki birimlerin birbirine yakınlığı kümelemenin performansı hakkında bilgi vermektedir. Uygulama esnasında veri tipine göre farklı uzaklık ölçüleri kullanılmaktadır. Verinin kesikli ya da sürekli olması, sıklık sayısı olması, ikili sınıflı veri olması ya da küme-grup-örneklem olmasına göre farklı uzaklık ölçüsü tekniği kullanılmaktadır. Tablo 1'de n×p boyutlu veri seti için genel bir gösterim verilmiştir.

Tablo 1. n×p boyutlu veri matrisi

	Değişkenler			
Gözlem	X_1	X_2	. . . . .	X_p
1	x_{11}	x_{12}	. . . . .	x_{1p}
.	.	.	. . . . .	.
.	.	.	. . . . .	.
i	x_{i1}	x_{i2}	. . . . .	x_{ip}
j	x_{j1}	x_{j2}	. . . . .	x_{jp}
.	.	.	. . . . .	.
.	.	.	. . . . .	.
n	x_{n1}	x_{n2}	. . . . .	x_{np}

Tablo 1'de veri matrisinde gözlemlerin birbirlerine uzaklıkları *d* ile gösterilmek üzere

d_{ij}:*i.* ve *j*. gözlemin birbirine uzaklıkları olmak üzere uzaklıklar aşağıdaki özellikleri sağlar.

$\boldsymbol{d_{ij} = d_{ji}}$	Simetri özelliği. *i.* ve *j*. gözlemin arasındaki uzaklık ile *j.* ve *i.* gözlem arasındaki uzaklık birbirine eşittir.
Eğer $i \neq j$ ise $d_{ij} > 0$	Negatif olmama özelliği. *i.* ve *j*. gözlemin arasındaki uzaklık sıfırdan büyüktür.
Eğer $i = j$ ise $d_{ij} = 0$	Tanım özelliği. Bir gözlemin kendine uzaklığı sıfırdır. (Belirlilik)
$\boldsymbol{d_{ik} \leq d_{ij} + d_{jk}}$	Üç gözlemin (*i*,*j*,*k*) herhangi ikisi arasındaki uzaklık diğer iki çiftin arasındaki uzaklığın toplamını geçemez (Üçgen eşitsizliği) .

İki gözlem arasındaki büyük benzemezlik (uzaklık değeri) bu gözlemlerin birbirine uzak olduğunu, küçük benzemezlik değeri ise gözlemlerin birbirine yakın olduğunu gösterir (Alpar, 2017).

3.1. Sürekli ve Kesikli Sayısal Veriler İçin Uzaklık ve Benzerlik Ölçüleri

Sürekli veriler belli bir aralıkta mümkün tüm değerleri alabilen, kesikli veriler ise sayılabilir veya sayılabilir sonsuz değerler alabilen veri türüdür. Bu tür veriler için kullanılan uzaklık ölçülerine bu bölümde değinilmiştir.

3.1.1. Öklid uzaklık ölçüsü

Uzaklık ölçüleri arasında en çok kullanılanı öklid uzaklık ölçüsüdür. Öklid uzaklığı Pisagor bağıntısı yardımıyla eşitlik (1)'deki gibi hesaplanmaktadır. (Dudo ve ark.,2001)

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \qquad (1)$$

(1) Eşitliğindeki değişkenlerin karşılığı aşağıda verilmiştir.

3.1.2. Kare öklid uzaklık ölçüsü

Öklid uzaklığının karesi kare öklid uzaklığını vermektedir (Dudo ve ark, 2001) . Aşağıdaki (2) eşitliğinde kare öklid uzaklık ölçüsü verilmiştir.

$$d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \qquad (2)$$

3.1.3. Chebychev uzaklık ölçüsü

Veri matrisinde bulunan her bir gözlemin birbirleri arasındaki farklarının mutlak değerinin en büyüğü (maksimum) Chebychev uzaklık ölçüsünü vermektedir (Monev, 2004). (3) eşitliğinde Chebychev uzaklık ölçüsü verilmiştir.

$$d_{ij}=\max_k^p\left|x_{ik}-x_{jk}\right| \quad (3)$$

3.1.4. Manhattan City-Blok uzaklık ölçüsü

Veri matrisinde bulunan her bir gözlemin birbirleri arasındaki farklarının mutlak değerinin toplamı Manhattan City-Blok uzaklık ölçüsünü vermektedir. Daha çok kesikli sayısal veriler için önerilir (Deza & Deza, 2006). (4) eşitliğinde bu uzaklık ölçüsü verilmiştir.

$$d_{ij}=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right| \quad (4)$$

3.1.5. Minkowski uzaklık ölçüsü

Minkowski uzaklık ölçüsü (5) eşitliğinde görüldüğü gibi gözlemler arasındaki mutlak değerce farkların m. kuvvetinin toplamlarının $\frac{1}{m}$. kuvvetidir. Minkowski uzaklık ölçüsü uzaklık m=1 için Manhattan City-Blok uzaklık ölçüsünü, m=2 için öklid uzaklık ölçüsünü verir (Zezula ve ark., 2006).

$$d_{ij}=\left[\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{m}\right]^{\frac{1}{m}} \quad (5)$$

3.1.6. Karl Pearson uzaklık ölçüsü

Karl Pearson uzaklık ölçüsü öklid uzaklığının 1/ S_k^2 ile düzeltilmesi yani standartlaştırılması ile elde edilir. Bu nedenle standartlaştırılmış öklid uzaklığı da denir. (6) eşitliğinde Karl Pearson uzaklık ölçüsü verilmiştir. (Alpar, 2017)

$$d_{ij}=\sqrt{\sum_{k=1}^{p}\frac{1}{S_k^2}\left(x_{ik}-x_{jk}\right)^2} \quad (6)$$

3.1.7. Korelasyon uzaklık ölçüsü

Korelasyon iki değişken arasındaki doğrusal ilişkiyi gösterir. Korelasyon uzaklık ölçüsü ise (7), (8), (9) ve (10) eşitliklerinde görüldüğü gibi 4 şekilde hesaplanabilir.

$d_{ij}=\left(1-r_{ij}\right)^2$ (7)

$d_{ij}=1-\left|r_{ij}\right|$ (8)

$d_{ij}=1-r_{ij}^2$ (9)

$d_{ij}=1-r_{ij}$ (10)

Korelasyon katsayısı -1 ile +1 arasında değişirken bu uzaklıklar 0 ile 2 arasında değer alır (Alpar, 2017).

Pearson korelasyon katsayısı ise değişken çiftleri arasındaki ilişkiyi ölçer. Değişkenler çerçevesinde yapısal uyumluluk hakkında bilgi verir (Pearson, 1900). Korelasyon (11). eşitlik ile hesaplanır.

$$r_{ij}=\frac{\sum_{k=1}^{p}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{p}(x_{ik}-\bar{x}_i)^2\sum_{k=1}^{p}(x_{jk}-\bar{x}_j)^2}} \quad (11)$$

$\bar{x}_i$=i' inci gözlem üzerinden ölçülen tüm p değişken değerlerinin ortalaması

$$\bar{x}_i=\frac{1}{p}\sum_{k=1}^{p}x_{ik} \quad (12)$$

Korelasyon katsayısı kullanılarak iki gözlem vektörü arasındaki uzaklık (13). eşitlikteki gibidir. (Işık & Çamurcu, 2009)

$d_{ij}=1-r_{ij}$ (13)

3.1.8. Cosine uzaklık ölçüsü

Cosine uzaklık ölçüsü ile iki vektör arasındaki uzaklık bu iki vektör arasındaki açının kosinüs değeri hesaplanarak (14). eşitlikte olduğu gibi bulunur.

$$\text{Cosine uzaklığı}=\frac{\sum_{i,j}^{p}x_i x_j}{\sqrt{\sum_{i=1}^{p}x_i^2\sum_{j=1}^{p}x_j^2}} \quad (14)$$

x_i; i. gözlemin x değişkeni değeri x_j,; j. gözlemin x değişkeni değeri, p;değişken sayısı (Neyman, 1949) .

3.2. Sıklık Sayıları İçin Uzaklık Ölçüleri

Bir sınıfa düşen veri sayısı sıklık sayısı olarak tanımlanır. Bu veri türünde kullanılan uzaklık ölçüleri aşağıda verilmiştir.

3.2.1 Ki-kare uzaklık ölçüsü

İki gözlem arasındaki ki-kare istatistiğinin kare kökü uzaklık ölçüsü olarak kullanılır(Alpar, 2017).

$$d_{ij} = \sqrt{(G-B)^2/B} \quad (15)$$

Eşitlik (15)'de G değeri gözlenen sıklığı, B değeri beklenen sıklığı gösterir.

3.2.2 Phi-kare uzaklık ölçüsü

İki gözlem arasındaki ki-kare uzaklık ölçüsü toplam gözlem sayısına bölünüp karekökünün alınması ile elde edilir. Phi-kare uzaklık ölçüsü eşitlik 16'da verilmiştir (Alpar, 2017).

$$d_{ij} = \sqrt{(\sqrt{(G-B)^2/B})/(n_i + n_j)} \quad (16)$$

3.3. İkili Sınıflı (Binary) Veriler İçin Uzaklık ve Benzerlik Ölçüleri

İkili veri adı üstünde iki sınıfı bulunan veri türüdür. Bu veri türü genel olarak 0 (incelenen özelliğin bulunmaması) ya da 1 (incelenen özelliğin bulunması) olarak tanımlanır. İkili sınıflı veriler için uzaklık ve benzerlik ölçülerinin hesaplanmasında 2x2 boyutlarında çapraz tablolardan yararlanılır (Alpar, 2017).

Tablo 2. 2×2 boyutlu sınıflandırma matrisi

		*i.*gözlem		
		1	0	Toplam
j. gözlem	1	a	b	a+b
	0	c	d	c+d
	Toplam	a+c	b+d	a+b+c+d

3.3.1. Öklid uzaklık ölçüsü

Öklid uzaklık ölçüsü ikili sınıflı veriler için eşitlik (17)'deki gibidir (Krause, 1986).

$$d_{ij}= \sqrt{b + c} \tag{17}$$

3.3.2. Büyüklük farkları uzaklık ölçüsü

Büyüklük farkları uzaklık ölçüsü eşitlik (18)'deki gibidir ve en küçük değeri 0, en büyük değeri sonsuzdur (Alpar, 2017).

$$d_{ij}=\frac{(b-c)^2}{(a+b+c+d)^2} \tag{18}$$

3.3.3. Biçim farkları uzaklık ölçüsü

$$d_{ij}=\frac{bc}{(a+b+c+d)^2} \tag{19}$$

Biçim farkları uzaklık ölçüsü eşitlik (19)'daki gibidir ve en küçük değeri 0, en büyük değeri sonsuzdur (Alpar, 2017).

3.3.4. Russel ve Rao benzerlik ölçüsü

Russel ve Rao benzerlik ölçüsü sadece 1-1 (tam benzerlik) olan çiftlerin toplam içindeki payını verir ve 0 ile 1 arasında değişir. Bu benzerlik ölçüsü eşitlik (20)'deki gibi ifade edilir: (Russel & Rao, 1940)

$$d_{ij}=\frac{a}{a+b+c+d} \quad (20)$$

3.3.5. Basit benzerlik ölçüsü (eşleştirme katsayısı)

Basit benzerlik ölçüsü Zubin tarafından 1938 yılında önerilmiştir. 1-1 (tam benzerlik) ve 0-0 (benzerlik yok) olan çiftlerin toplam içindeki payını verir ve eşitlik (21)'deki gibi ifade edilir:

$$\text{Basit benzerlik ölçüsü}=\frac{a+d}{a+b+c+d} \quad (21)$$

3.3.6. Jaccard benzerlik ölçüsü

Jaccard benzerlik ölçüsü 0-0 olan çiftleri dikkate almaz. Bu benzerlik ölçüsü eşitlik (22)'deki gibi gösterilir: (Jaccard, 1901)

$$d_{ij}=\frac{a}{a+b+c} \quad (22)$$

3.3.7. Parçalı benzerlik ölçüsü

Parçalı benzerlik ölçüsü Dice tarafından 1945 yılında önerilmiştir. Bu benzerlik ölçüsü eşitlik (23)'deki gibidir:

$$d_{ij}=\frac{2a}{2a+b+c} \quad (23)$$

3.3.8. Rogers ve Tanimoto benzerlik ölçüsü

Czekanowski ya da Sorensen ölçüsü olarak da bilinir.1-1 olan çiftlere iki kat ağırlık verir. 0-0 olan çiftleri dikkate almaz. 0 ile 1 arasında değer alır. Eşitlik (24)'de bu benzerlik ölçüsü gösterilmiştir: (Rogers & Tanimoto, 1960)

$$d_{ij}=\frac{a+d}{a+d+2(b+c)} \quad (24)$$

3.3.9. Sokal ve Sneath benzerlik ölçüsü 1

Sokal ve Sneath benzerlik ölçüsü 0-0 ve 1-1 olan çiftlere iki kat ağırlık verir. 0 ile 1 arasında değer alır. Bu benzerlik ölçüsü eşitlik (25)'deki gibidir: (Sneath & Sokal, 1963)

$$d_{ij}=\frac{2(a+d)}{2(a+d)+b+c} \quad (25)$$

3.3.10. Sokal ve Sneath benzerlik ölçüsü 2

Sokal ve Sneath benzerlik ölçüsünün 2. si 0-0 olan çiftleri dikkate almazken 0-1 olan çiftlere paydada iki kat ağırlık verilir ve eşitlik (26)'daki gibi gösterilir (Sneath & Sokal, 1973)

$$d_{ij}=\frac{a}{a+2(b+c)} \quad (26)$$

3.3.11. Sokal ve Sneath benzerlik ölçüsü 3

Sokal ve Sneath benzerlik elçüsünün 3.sünün en küçük değeri 0 ve en büyük değeri ise yoktur. Bu benzerlik ölçüsü eşitlik (27)'deki gibidir ve b=0 ve c=0 olduğunda tanımsızdır. (Sneath & Sokal, 1973)

$$d_{ij}=\frac{a+d}{b+c} \quad (27)$$

3.3.12. Kulczynski benzerlik ölçüsü

Kulczynski benzerkik ölçüsünün en küçük değeri 0dır. En büyük değeri yoktur. Kulczynski benzerlik ölçüsü eşitlik (28)'deki gibidir ve b=0 ve c=0 olduğunda tanımsızdır. (Kulczynski, 1927)

$$d_{ij}=\frac{a}{b+c} \quad (28)$$

3.3.13. Ochiai benzerlik ölçüsü

Ochiai benzerlik ölçüsü 0 ile 1 arasında değer alır ve eşitlik (29)'daki gibidir: (Ochiai, 1957)

$$d_{ij}=\frac{a}{\sqrt{(a+b)+(a+c)}} \quad (29)$$

3.3.14. Yule Q benzerlik ölçüsü

Yule Q benzerlik ölçüsü-1 ile 1 arasında değer alır ve eşitlik (30)'daki gibidir (Yule, 1900):

$$d_{ij}=\frac{ad-bc}{ad+bc} \quad (30)$$

3.3.15. Gower ve Legendre benzerlik ölçüsü 1 (1986)

Gower ve Legendre benzerlik ölçüsü eşitlik (31)'deki gibidir. (Legendre & Gower, 1986)

$$d_{ij}=\frac{a+d}{a+\frac{1}{2}(b+c)+d} \quad (31)$$

3.3.16. Gower ve Legendre benzerlik ölçüsü 2 (1986)

Gower ve Legendre'in önerdiği 2. benzerlik ölçüsü eşitlik (32)'deki gibidir. (Legendre & Gower, 1986)

$$d_{ij}=\frac{a+d}{a+\frac{1}{2}(b+c)} \quad (32)$$

3.4. Kümeler/Örneklemler/Gruplar Arası Uzaklık ve Benzerlik Ölçüleri

Verilere ilişkin ortalama, varyans ve kovaryansların bilinmesi durumunda çok değişkenli örneklemlerin aralarındaki uzaklıkların belirlenmesine yönelik geliştirilmiş birçok ölçü vardır. Bunlardan en bilinenleri Mahalonobis uzaklık ölçüsü, Hotelling T^2 uzaklık ölçüsü ve Penrose uzaklık ölçüsüdür (Alpar, 2017).

3.4.1. Mahalonobis uzaklık ölçüsü

Mahalonobis uzaklık ölçüsü m_i^2 çok değişkenli X veri matrisindeki herhangi bir gözlemin, verinin merkezinden (ortalama vektöründen) uzaklığının bir ölçüsüdür ve m_i^2 değeri büyüdükçe ilgili gözlemin veri merkezine olan uzaklığı da artar. m_i^2'lerin en önemli özelliği de evren dağılımının normal dağılım göstermesi durumunda m_i^2'lerin p serbestlik derecesi ile ki-kare dağılımı göstermesidir. Buna göre anlamlı derecede büyük bir m_i^2 değeri, ilgili gözlemin aşırı bir gözlem olabileceği ya da başka bir örnekleme ait olabileceği ya da yazım hatasından kaynaklanan bir sorun olabileceği gibi sonuçlara ulaşmada araştırmacıya yol gösterici olabilir.

Örneklem varyans-kovaryans matrislerinin homojen olduğu varsayımı altında iki grup arasındaki Mahalonobis uzaklığı eşitlik (33)'de görüldüğü gibidir (Mimmack, 2001)

$$D_{ij}=(\mu_i-\mu_j)' S^{-1}(\mu_i-\mu_j) \quad (33)$$

3.4.2. Penrose uzaklık ölçüsü

Penrose uzaklık ölçüsü, Mahalonobis uzaklık ölçüsü gibi değişkenler arasındaki ilişkiyi dikkate almaz ve eşitlik (34)'deki gibi ifade edilir (Alpar, 2017) (Johnson & Wichern, 2014)

$$P_{ij}=\sum_{k=1}^{p} \frac{(\mu_{ik}-\mu_{jk})^2}{pS_k^2} \quad (34)$$

3.4.3. Hotelling T^2 uzaklık ölçüsü

İki grup ya da kümenin ortalama vektörlerinin karşılaştırılmasında Hotelling T^2 uzaklık ölçüsü kullanılır (Alpar, 2017). Hotelling T^2 uzaklık ölçüsü eşitlik (35)'de verilmiştir:

$$T^2=(\frac{n_i n_j}{n_{i+}n_j})\sum_{k=1}^{p}(\mu_{ik}-\mu_{jk})' S^{-1}(\mu_{ik}-\mu_{kj}) \quad (35)$$

4. KÜMELEME YÖNTEMLERİ

Uzaklık ve benzerlik matrislerinin elde edilmesinden sonra gözlemleri ya da değişkenleri kümelemekte kullanılacak yönteme karar verilir. Bu yöntemler genel olarak aşamalı ve aşamalı olmayan kümeleme yöntemleri olmak üzere iki alt başlıkta toplanır.

4.1. Aşamalı Kümeleme Yöntemleri

Aşamalı kümeleme algoritmaları, ardışık bir işlem süreci içerir. Aşamalı kümeleme yöntemleri araştırmacının incelenen veri setinde kaç küme bulunduğunu bilmediği durumlarda kullanılan bir yöntemdir. Bu yöntem araştırmacıya incelediği veri setinde daha önce gözlemlenmemiş ilişkileri ve prensipleri keşfetme olanağı vermektedir (Erişoğlu, 2011).

Aşamalı kümeleme yöntemleri eklemeli ve bölünmeli olarak ikiye ayrılır. Eğer başlangıçta *n* sayıda gözlemin her biri bir küme olarak düşünülüp, bu kümeler her aşamada en benzer iki kümenin bağlanması ile tek küme kalıncaya kadar birleştiriliyorsa teknik, eklemeli aşamalı kümeleme tekniği olarak isimlendirilir (Hansen ve ark. 1997). Başlangıçta *n* sayıda gözlemi tek bir küme içerisinde kabul edip, *n* birimlik bu kümeden, küme bölme işlemlerinin uygulanmasıyla her bir gözlem farklı kümede olacak şekilde *n* sayıda küme oluşturuluyorsa, teknik bölünmeli aşamalı kümeleme tekniği olarak isimlendirilir (Gordon, 1999).

En çok kullanılan teknik eklemeli aşamalı kümeleme yöntemidir. Bu yöntemin algoritması şu şekilde işler:

1. Her bir gözlem bir küme olarak alınır.
2. Benzerlik matrisi hesaplanır.
3. En benzer iki küme birleştirilir.
4. 2. ve 3. adımlar tek bir küme kalıncaya kadar tekrarlanır.

Eklemeli aşamalı kümeleme yöntemleri alt bölümlerde verilmiştir.

4.1.1. Tek bağlantı kümeleme yöntemi

"En yakın komşuluk" olarak da bilinen bu yöntem en benzer birimlerin bir araya getirilmesi esasına dayanır. Örnek olarak *x*, *y* ve *z* şeklinde üç birim göz önüne

alındığında, ilk olarak birimler arasındaki uzaklıklar hesaplanır ve birbirine en yakın iki birim bir kümeye atanır eşitlik (36) kullanılarak:

$$d(x,y,z) = enküçük(d(x), d(y), d(z)) = d(y,z) \quad (36)$$

Daha sonra oluşan kümenin diğer birimlere olan en yakın uzaklığına bakılır eşitlik (37) yardımıyla:

$$d(x,\{y,z\}) = enküçük(d(x,y), d(x,z)) \quad (37)$$

Bu işlem her birim bir kümeye dahil olana kadar devam eder ve kümeleme işlemi sonuçlanır (Florek ve ark., 1951).

4.1.2. Tam bağlantı kümeleme yöntemi

"En uzak komşuluk" olarak da bilinir. Tam bağlantı kümeleme tekniğinde, ilk olarak birimler arasındaki uzaklıklar hesaplanır. İkinci aşamada birimler arasındaki en küçük uzaklığa sahip iki birim ilk kümeyi oluşturur. Üçüncü aşamada belirlenen kümenin diğer birimlere olan uzaklıklarından en büyük olan o kümeye atanır ve ataması yapılmayan hiçbir birim kalmayıncaya kadar aşamalar tekrarlanır. Tam bağlantı kümeleme yönteminin üçüncü aşaması ve benzer aşamalarında örnek olarak 3 birim için *x* biriminin *y* ve *z* birimlerinden oluşan yeni kümeye olan uzaklığı,

$$d(x,\{y,z\}) = enbüyük(d(x,y), d(x,z)) \quad (38)$$

(38) eşitliğiyle belirlenir (Sorensen, 1948).

4.1.3. Ortalama bağlantı kümeleme tekniği

Ortalama bağlantı kümeleme tekniğinde, iki küme arasındaki uzaklık bir kümedeki birimlerin diğer kümedeki birimlere olan uzaklıklarının ortalaması ile elde edilir. Bu yöntemde ilk aşamada birimler arasındaki uzaklıklar hesaplanır ve bu uzaklıklardan en küçüğünü temsil eden birbirine en yakın iki birim ilk kümeyi oluşturur. İkinci aşamada bu küme birimlerinin diğer birimlere olan ortalama uzaklığı hesaplanır ve

bu yeni uzaklıkları oluşturur. Bu aşamalar atanmamış hiçbir birim kalmayıncaya kadar tekrarlanır. Formüle edilecek olursa her biri sırasıyla *n* ve *m* adet birimden oluşan $K_i$ve K_j kümeleri arasındaki uzaklık:

$$d(K_i, K_j)=\frac{1}{n\times m}\sum_{x_i\epsilon K_i}^{n}\sum_{x_j\epsilon K_j}^{m} d(x_i, x_j) \quad (39)$$

eşitlik (39)'da görüldüğü gibi belirlenerek aşamalar tekrarlanır (Jain & Dubes, 1988).

4.1.4. Ağırlıklı ortalama bağlantı yöntemi

Ağırlıklı ortalama bağlantı kümeleme tekniği, aritmetik ortalama ile ağırlıklı çift grup yöntemi olarak da adlandırılır (Jain & Dubes 1988). Bu yöntemde ilk aşamada birimler arasındaki uzaklıklar hesaplanır ve bu uzaklıklardan birbirine en yakın iki birim ilk kümeyi oluşturur. İkinci aşamada bu küme birimlerinin diğer birimlere olan ağırlıklı ortalama uzaklığı hesaplanır ve yeni uzaklıklar oluşur. Bu uzaklıklardan yine en küçük olanı yeni kümeyi oluşturur ve bu aşamalar hiçbir boşta birim kalmayıncaya kadar tekrarlanır. Bu yöntem formüle edilecek olursa K_l kümesinin, K_i ve K_j kümelerinin bağlanması ile elde edilen K_i ve K_j birleşim kümesine olan uzaklığı ağırlıklı ortalama bağlantı yönteminde,

$$d(K_l,\{ K_i \cup K_j\})=\frac{1}{2} d(K_l, K_i)+ \frac{1}{2} d(K_l, K_j) \quad (40)$$

şeklinde eşitlik (40) ile hesaplanır ve aşamalar tekrarlanır.

4.1.5. Merkezi bağlantı kümeleme yöntemi

Merkezi bağlantı kümeleme tekniğinde iki küme arasındaki uzaklık, iki küme merkezi arasındaki uzaklık olarak tanımlanır. Bu yöntemde ilk aşamada birimler arasındaki uzaklıklar hesaplanır ve bu uzaklıklardan birbirine en yakın iki birim ilk kümeyi oluşturur. İkinci aşamada bu küme birimlerinin aritmetik ortalaması alınır ve bu o kümenin merkezini oluşturur. Küme merkezinin diğer birimlere olan uzaklığı hesaplanır ve bu yeni uzaklıklar olur. Yeni oluşan kümelerin merkezleri tekrar hesaplanarak diğer birim ve küme merkezlerine uzaklıkları hesaplanır ve en küçük

uzaklığa sahip birimler yeni kümeyi oluşturur. Bu aşamalar hiçbir birim kalmayıncaya kadar tekrarlanır.

K_i kümesinin merkezi $\bar{x}$ ve K_j kümesinin merkezi $\bar{y}$ olarak alındığında, K_i kümesinin K_j kümesine olan uzaklığı merkez bağlantı yöntemine göre;

$$d(K_i, K_j)=d(\bar{x}, \bar{y}) \quad (41)$$

eşitlik (41) ile hesaplanır(Johnson ve Wichern, 2014).

4.1.6. Medyan bağlantı kümeleme yöntemi

Merkezi bağlantı kümeleme yöntemiyle benzer özellikte olan iki kümenin birleşmesi ile oluşan kümelerin merkezlerinin belirlenmesinde birim sayısı fazla olan kümenin etkisini ortadan kaldırmak için Gower (1967) medyan bağlantı kümeleme tekniğini önermiştir.

Medyan bağlantı kümeleme yönteminde ilk aşamada birimler arasındaki uzaklıklar hesaplanır ve uzaklıklardan birbirine en yakın iki birim ilk kümeyi oluşturur. İkinci aşamada merkezi bağlantı kümeleme yönteminden farklı olarak medyan bağlantı kümeleme yönteminde K_l kümesinin, K_i ve K_j kümelerinin bağlanması ile elde edilen $\{K_i \cup K_i\}$ kümesine olan uzaklığı:

$$d(K_l,\{ K_i \cup K_j\})=\frac{1}{2} d(K_l, K_i)+\frac{1}{2} d(K_l, K_j)-\frac{1}{4} d(K_i, K_j) \quad (42)$$

eşitlik (42) ile hesaplanır ve yeni uzaklıklar oluşturulur. Üçüncü aşamada birbirine en yakın birimler yeni kümeyi oluşturur ve hiçbir birim boşta kalmayıncaya kadar bu aşamalar tekrarlanır (Gower, 1967).

4.1.7. Ward bağlantı kümeleme yöntemi

Ward (1963), iki kümenin birleşmesinde oluşacak bilgi kaybını minimize etmeyi amaçlayan genel bir aşamalı kümeleme tekniği önermiştir. K_i ve K_j kümeleri birleştirilerek oluşturulan yeni küme $\{K_i \cup K_i\}$ olsun. Bu durumda bilgi kaybındaki artış,

$$\boldsymbol{I}_{\{K_i \cup K_i\}} = \frac{n_i n_j}{n_i + n_j} (\bar{X}_{(K_i)} - \bar{X}_{(K_j)})' (\bar{X}_{(K_i)} - \bar{X}_{(K_j)}) \quad (43)$$

eşitlik (43) ile hesaplanır (Rencher, 2002).

Ward bağlantı kümeleme yönteminde ilk aşamada birimler arasındaki uzaklıklar hesaplanır ve birbirine en yakın birimler ilk kümeyi oluştururlar. İkinci aşamada oluşan yeni küme içindeki birimlerin diğer birimlere uzaklıkları kümedeki birim sayısı dikkate alınarak hesaplanır. Bu hesaplamada :

$$d(K_l, \{K_i \cup K_j\}) = \frac{(n_l n_i d(K_l, K_i) + n_l n_j d(K_l, K_j) - n_l d(K_i, K_j)}{n_i + n_j + n_l} \quad (44)$$

(44) eşitliğinde K_l kümesinin, K_i ve K_j kümelerinin bağlanması ile elde edilen $\{K_i \cup K_i\}$ kümesine olan uzaklığı gösterilmiştir. Uzaklıklar bu şekilde hesaplandıktan sonra üçüncü aşamada birbirine en yakın birimler yeni kümeyi oluşturur. Bu aşamalar açıkta birim kalmayıncaya kadar devam eder ve kümeleme işlemi tamamlanır (Ward, 1963).

4.1.8.Esnek beta yöntemi

K_l sayıda aşamalı kümeleme tekniği bulunmaktadır. Lance ve Williams (1967), K_l kümesini $\{K_i \cup K_i\}$ kümesine olan uzaklığını belirlemek için önerilen aşamalı kümeleme tekniklerine,

$$d(K_l, \{K_i \cup K_j\}) = \alpha_i\, d(K_l, K_i) + \alpha_j\, d(K_l, K_j) + \beta\, d(K_i, K_j) + \gamma\, d(K_l, K_i) - d(K_i, K_j) \quad (44)$$

eşitliğindeki α_i, α_j, β, γ, parametrelerine uygun değerler verilerek ulaşılacağını göstermişlerdir. Lance ve Williams (1967), iyi bir kümeleme elde edebilmek için eşitlik (44)'deki parametreler için aşağıdaki (45), (46), (47), (48) kısıtlamaları önermişlerdir.

$\alpha_i + \alpha_j + \beta = 1$ (45)

$\alpha_i = \alpha_j$ (46)

$\gamma = 0$ (47)

$\beta < 1$ (48)

Tablo 3'de α_i, α_j, β, γ parametrelerinin aldığı değerlere göre hangi bağlantı tekniğine dönüşeceği görülmektedir.

Tablo 3: Aşamalı Kümeleme Yöntemleri

Aşamalı Kümeleme Yöntemleri	α_i	α_j	β	γ
Tek Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Tam Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Ortalama Bağlantı	$\frac{n_i}{n_i + n_j}$	$\frac{n_i}{n_i + n_j}$	0	0
Ağırlıklı Ortalama Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Merkezi Bağlantı	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$\frac{-n_i n_i}{(n_i + n_j)^2}$	0
Medyan Bağlantı	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward Bağlantı	$\frac{n_i + n_l}{n_i + n_j + n}$	$\frac{n_j + n_l}{n_i + n_j + n_l}$	$\frac{-n_l}{n_i + n_j + n_l}$	0
Esnek Beta Tekniği	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	$\beta(< 1)$	0

4.2. Aşamalı Olmayan Kümeleme Yöntemleri

Aşamalı olmayan kümeleme de denilmektedir. Yöntemlerin başlıca özelliği verinin kaç kümeye ayrılacağının önceden araştırmacı tarafından belirlenmesidir. Literatürde en çok kullanılan aşamalı olmayan kümeleme yöntemi k-ortalamalar yöntemi bu bölümde incelenecektir.

4.2.1. k-ortalamalar yöntemi

En yaygın kullanılan aşamalı olmayan kümeleme yöntemi k-ortalamalar yöntemidir. Araştırmacı tarafından belirlenen küme sayısına (k) göre veri setini kümelere ayıran kümeleme yöntemidir. En az küme sayısı 2, en fazla küme sayısı gözlem sayısını geçmeyecek şekilde araştırmacı tarafından belirlenir. k-ortalamalar yönteminde amaç gözlemlerin küme merkezlerine uzaklıklarını minimum, kümeler arasındaki uzaklıkların maksimum yapılmasıdır. Diğer bir ifadeyle amaç (50) eşitliği ile verilen hata fonksiyonunu minimize etmektir.

$$E = \sum_{i=1}^{k} \sum_{C_i} d(X, \mu(K_i)) \quad (49)$$

$K_1, K_2, \ldots, K_i$: X veri matrisindeki ayrık kümeler (Erişoğlu, 2011)

k-ortalamalar algoritması şu şekilde işlemektedir:

1. X veri setinin kaç kümeye ayrılacağı yani (k) küme sayısı belirlenir.

2. k sayıda kümenin küme merkezleri belirlenir.

3. Birimlerin her bir küme merkezine olan uzaklıkları hesaplanır.

4. Birimlerin hangi küme merkezine yakınsa o kümeye atamaları yapılır

5. Küme değiştiren birim kalmayana kadar bu adımlar 2. adımdan itibaren tekrarlanır.

5. KÜMELERİN DEĞERLENDİRİLMESİ

Aynı veri setine farklı kümeleme algoritmaları uygulanarak elde edilen sonuçların karşılaştırılması kümelerin değerlendirilmesi ya da küme geçerliliği olarak adlandırılmaktadır. Küme geçerliliğini incelemek için dışsal (external) kriterler, içsel (internal) kriterler olmak üzere iki yaklaşım bulunmaktadır. (Alpar, 2017)

Küme geçerliliğinin temel amaçları şöyledir:

1. Veri seti için en iyi küme sayısının belirlenmesi

2. Veri seti için alternatif kümeleme algoritmalarının sonuçlarının karşılaştırılması

3. Veri setinin herhangi bir doğal gruplama yapısının olup olmadığının belirlenmesi

5.1. Dışsal (External) Kriterler

Veri kümesi X ve bu veri kümesinden elde edilen kümeleme yapısı C ile gösterilsin. Dışsal kriter veri hakkında sahip olunan önbilgi (P) ile kümeleme algoritması sonunda elde edilen kümeleme sonucunun (C) karşılaştırılmasıdır.

P:X veri matrisi hakkında bilgi sahibi olunan kümelenmesi.

C: X veri matrisine kümeleme algoritması uygulanması sonucu elde edilen kümelenmesi.

X veri matrisinde x_i ve x_j gözlem çifti olsun. Bu gözlem çiftinin C ve P kümelenmelerine göre dört durumu vardır.

a : x_i ve x_j gözlem çiftinin hem C hem de P kümelenmesinde aynı kümelerde yer alması.

b : x_i ve x_j gözlem çiftinin C kümelenmesinde aynı küme içerisinde, P kümelenmesinde farklı küme içerisinde yer alması.

c : x_i ve x_j gözlem çiftinin C kümelenmesinde farklı küme içerisinde, P kümelenmesinde aynı küme içerisinde yer alması.

d : x_i ve x_j gözlem çiftinin hem C hem de P kümelenmesinde farklı kümelerde yer alması.

Bu dört durum göz önüne alındığında a ve d iki kümenin benzerliği ile ilgiliyken, b ve c iki kümenin farklılığı ile ilgilidir. Dışsal kriterlerde kullanılan başlıca indeksler şöyledir aşağıdaki tabloda verilmiştir.

Tablo 4: Dışsal Kriterler

İndeks Adı	Değerlendirme Kriteri	Öneren Kişi
Rand indeksi	$C_R=\frac{a+d}{a+b+c+d}$	Rand tarafından 1971'de önerilmiştir.
Jaccard katsayısı	$C_J=\frac{a}{a+b+c}$	Jaccard tarafından 1908'de önerilmiştir.
Fowlkes ve Mallows indeksi	$C_{FM}=\sqrt{\frac{a}{a+b}\frac{a}{a+c}}$	Fowlkes ve Mallows tarafından 1983'de önerilmiştir.
Γ istatistiği	$C_\Gamma=\frac{\mathrm{Ma}-\mathrm{m}_1\mathrm{m}_2}{\sqrt{\mathrm{m}_1\mathrm{m}_2(\mathrm{M}-\mathrm{m}_1)(\mathrm{M}-\mathrm{m}_2)}}$ M= a+b+c+d ; $m_1 = a + b$; m_2 =c+d	Xu ve Wunsch tarafından 2009'da önerilmiştir.yok
Czekanowski-Dice indeksi	$C_{CD} = \frac{2a}{2a+b+c}$	Czekanowski-Dice tarafından 1945'de önerilmiştir.
Kulczyns indeksi	$C_K = \frac{1}{2}\left(\frac{a}{a+c}+\frac{a}{a+b}\right)$	Kulczyns tarafından 1927'de önerilmiştir.
McNemar indeksi	$C_{MN} = \frac{a-c}{\sqrt{a+c}}$	McNemar tarafından 1947 yılında önerilmiştir.yok
Phi indeksi	$\mathrm{C_P} = \frac{\mathrm{a}\times\mathrm{d}-\mathrm{b}\times\mathrm{c}}{(\mathrm{a}+\mathrm{b})(\mathrm{a}+\mathrm{c})(\mathrm{b}+\mathrm{d})(\mathrm{c}+\mathrm{d})}$	Guilford tarafından 1936'da önerilmiştir.
Rogers-Tanimoto indeksi	$C_{RT} = \frac{a+d}{a+d+2(b+c)}$	Rogers-Tanimoto tarafından 1960 yılında önerilmiştir.
Russel-Rao indeksi	$C_{RR} = \frac{a}{N_T}$	Russel ve Rao 1940 yılında önermiştir.
Sokal-Sneath indeksi	$C_1 = \frac{a}{a+2(b+c)}$ $C_2 = \frac{a+d}{a+d+\frac{1}{2}(b+c)}$	Sokal-Sneath tarafından 1963 yılında önerilmiştir.

5.2. İçsel (İnternal) Kriterler

Veri setinin genel yapısı ile kümeleme algoritması sonucu elde edilen kümelemenin karşılaştırılmasını sağlayan kriterler içsel kriterlerdir. Genel olarak aşamalı kümeleme yapılarının karşılaştırılmasında kullanılır.

5.2.1. Kofenetik (Cophenetic) korelasyon katsayısı

Kofenetik korelasyon katsayısı başlangıçta hesaplanan uzaklık matrisi ile kümeleme algoritması uygulanarak oluşturulan ağaç grafiğindeki bağlantı uzaklıklar matrisinin karşılıklı elemanları arasındaki ilişkiyi ölçen bir katsayıdır. *CCC* ile gösterilir eşitlik (50)'de verilmiştir.

$$CCC=\frac{\sum_{i<j}(d_{ij}-\bar{d})(r_{ij}-\bar{r})}{\sqrt{\sum_{i<j}(d_{ij}-\bar{d})^2 \sum_{i<j}(r_{ij}-\bar{r})^2}} \quad (50)$$

Kofenetik korelasyon katsayısı [-1,1] arasında değer alır ve 1'e ne kadar yaklaşırsa kümeleme o kadar başarılıdır. (Erişoğlu, 2011)

5.2.2. Wilks'in lambda test istatistiği

Freidman ve Rubin tarafından 1907 yılında önerilmiştir. Aslında çok değişkenli varyans analizinde gruplar arası farklılıkların ölçülmesinde kullanılan bir test istatistiğidir.

|W|= Kümeler içi kareler ve çarpımlar toplamı matrisinin determinantı

|T|= Genel kareler ve çarpımlar toplamı matrisinin determinantı

$$\lambda = \frac{|W|}{|T|} \quad (51)$$

olmak üzere eşitlik (51) ile ifade edilen λ değeri ne kadar küçükse gerçekleştirilen test istatistiği o kadar başarılıdır (Erişoğlu, 2011)

5.2.3. Ball-Hall indeksi

Ball–Hall indeksi, kümeler içi kareler toplamının küme merkezlerine uzaklıklarının ortalamasıdır. Eşitlik (52)'deki gibi ifade edilen Ball-Hall indeksi ne kadar küçük değer alırsa kümeleme o kadar başarılıdır. (Ball & Hall, 1965)

$$C = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in I_k}\left\|M_i^{\{k\}} - G^{\{k\}}\right\|^2 \quad (52)$$

5.2.4. Banfeld-Raftery indeksi

Bu indeks, her kümenin varyans-kovaryans matrisinin köşegen elemanlarının toplamlarının logaritmalarının ağırlıklı toplamıdır. (Banfeld & Raftery, 1993)

$$\mathrm{C} = \sum_{k=1}^{K} n_k log\left(\frac{Tr(WG^{\{K\}})}{n_k}\right) \quad (53)$$

Bir küme tek bir nokta içeriyorsa, tek noktanın izi 0'a eşittir ve logaritma tanımsızdır. Eşitlik (53)'deki gibi ifade edilen Banfeld Raftery indeksi ne kadar küçük değer alırsa kümeleme o kadar başarılıdır. (Banfeld & Raftery, 1993)

5.2.5. C indeksi

Her küme içi nokta çiftleri arasındaki mesafeler dikkate alınır C indeksinde. İndeks ne kadar küçük değer alırsa kümeleme o kadar başarılıdır. Buna göre C indeksi eşitlik (54)'deki gibidir : (Hubert & Schultz, 1976)

$$C = \frac{S_W - S_{min}}{S_{max} - S_{min}} \quad (54)$$

5.2.6. Det-Oran indeksi

Det-Oran indeksi genel kareler matrisinin determinantının grup içi kareler matrisinin determinantına oranı şeklinde eşitlik (55)'deki gibi gösterilmiştir.

$$C = \frac{\det(T)}{\det(WG)} \quad (55)$$

Det-Oran indeksi ne kadar büyük değer alırsa kümeleme o kadar başarılıdır. (Scott & Symons, 1971)

5.2.7. Baker-Hubert Gamma indeksi

Baker-Hubert Gamma indeksi A ve B gibi aynı boyutlu vektörün korelasyonun uyarlamasıdır. Genel olarak i ve j gibi iki indis için $a_i < a_j$ ve $b_i < b_j$ ise yani her iki

vektör için sınıflandırma benzer sıralamada ise pozitif yani + uyumlu vektör çiftleridir. Aksi durumda ise negatif yani uyumsuz vektör çiftleridir. Buna göre Γ indeksi eşitlik (56)'da olduğu gibi tanımlanır:

$$C = \Gamma = \frac{s^+ - s^-}{s^+ + s^-} \quad (56)$$

Γ indeksi -1 ile +1 arasında değer alır. A vektöründeki iki nokta (M_i,M_j) arasındaki mesafe d_{ij}'dir (i<j). B vektörü de aynı koordinatlara karşılık gelen nokta çiftleri (M_i,M_j) barındırıyor ise Γ indeksi 0'dır ve aynı kümede yer alırlar bu vektörler. Bu durumda iki vektörün uzunluğu N(N-1)/2 dir. (Baker & Hubert, 1975)

$s^+ = \sum_{(r,s)\in I_B} \sum_{(u,v)\in I_W} 1_{\{d_{uv} < d_{rs}\}}$

$s^- = \sum_{(r,s)\in I_B} \sum_{(u,v)\in I_W} 1_{\{d_{uv} > d_{rs}\}}$

$s^+ - s^- = \sum_{(r,s)\in I_B} \sum_{(u,v)\in I_W} sgn(d_{rs} - d_{uv})$

5.2.8. G_Artı indeksi

Baker-Hubert Gamma indeksi kullanılarak G+ indeksi eşitlik (57)'deki gibi bulunur. Tüm nokta çiftleri arasında uyumsuz çiftlerin oranını ifade eder (Rohlf, 1974)

$$C = \frac{s^-}{N_T(N_T-1)/2} = \frac{2s^-}{N_T(N_T-1)} \quad (57)$$

Baker-Hubert indeksinin küçük değer alması kümelemenin başarılı olduğunu gösterir.

5.2.9. Ksq_detW indeksi

Ksq-detW indeksi grup içi kareler matrisinin determinantının, küme sayısının karesiyle çarpımıyla elde edilir . Eşitlik (58)'görülmektedir:

$$C = k^2|WG| \quad (58)$$

Ksq-detW indeksi ne kadar küçük değer alırsa kümeleme o kadar başarılıdır. (Marriot., 1971)

5.2.10. Log_det_oran indeksi

Log-Det-Oran indeksi genel kareler matrisinin determinantının, kümeler içi kareler matrisinin determinantına bölünerek logaritmasının alınarak gözlem sayısı ile çarpılması sonucu elde edilir. Eşitlik (59)'da bu indeks gösterilmiştir:

$$C=N \log \frac{\det(T)}{\det(WG)} \tag{59}$$

Log_det_oran indeksi ne kadar büyük değer alırsa kümeleme o kadar başarılıdır. (Scott & Symons, 1971)

5.2.11. Log_ss_oran indeksi

Log_SS_Oran indeksi gruplar kareler toplamının gruplar içi kareler toplamına bölümünün logaritması alınarak elde edilir. Eşitlik (60)'da bu indeks gösterilmiştir:

$$C = log\left(\frac{GAKT}{G\dot{I}KT}\right) \tag{60}$$

Log_ss_oran indeksi ne kadar büyük değer alırsa kümeleme o kadar başarılıdır. (Hartigan J. A., 1975)

5.2.12. McClain-Rao indeksi

McClain-Rao indeksi küme içi uzaklıklar toplamının kümeler arası uzaklıklar toplamına oranlamasına dayanır. Küme içi ve kümeler arası uzaklıklar toplamı sırasıyla S_W ve S_B olmak üzere ;

$$S_W=\sum_{(i,j)\in I_W} d(M_i, M_j) = \sum_{k=1}^{K} \sum_{\substack{(i,j)\in I_k \\ i<j}} d(M_i, M_j)$$

$$S_B=\sum_{(i,j)\in I_B} d(M_i, M_j) = \sum_{k<k'} \sum_{\substack{i\in I_k, j\in I_{k'} \\ i<j}} d(M_i, M_j)$$

N_B=N(N-1)/2-N_W

Olmak üzere Clain-Rao indeksi eşitlik (61)'de gösterilmiştir (McClain & Rao, 1975):

$$C=\frac{S_W/N_W}{S_B/N_B}=\frac{N_B}{N_W}\frac{S_W}{S_B} \quad (61)$$

5.2.13. Pbm indeksi

PBM indeksi yazarların Pakhira, Bandyopadhyay ve Maluki isimlerinin baş harflerinden oluşmaktadır. Bu indeks noktalar arası mesafeler, küme merkezleri ve küme merkezleri arasındaki mesafeler dikkate alınarak hesaplanır.

$D_B=max_{k<k'}d\left(G^{\{k\}},G^{\{k'\}}\right)$

$E_W=\sum_{k=1}^{K}\sum_{i\in I_k}d\left(M_i,G^{\{k\}}\right)$

$E_T=\sum_{i=1}^{N}d(M_i,G)$

Tanımlamalara göre PBM indeksi eşitlik (62)'deki gibidir. PBM indeksi ne kadar büyük değer alırsa kümeleme o kadar başarılıdır. (Pakhira ve ark, 2004)

$$C=\left(\frac{1}{K}\times\frac{E_T}{E_W}\times D_B\right)^2 \quad (62)$$

5.2.14. İki serili-nokta indeksi

Nokta iki serili indeks sürekli değişken olan A ve ikili değişken B (0 ya da 1 gibi) arasındaki korelasyondur. A ve B aynı sayıda birim içeren setlerdir. A'nın değerleri A0 ve A1 olarak ikiye ayrılır. B ise 0 ya da 1 'dir. M_{A_0} ve M_{A_1} , A_0 ve A_1 gruplarının ortalaması, $n_{A_0} ve\ n_{A_1}$, A_0 ve A_1 gruplarının eleman sayılarıdır. Buradan nokta iki serili korelasyon katsayısı şu şekilde hesaplanır:

$$r_{pb}(A,B)=\frac{M_{A_1}-M_{A_0}}{S_n}\sqrt{\frac{n_{A_0}n_{A_1}}{n^2}}$$

$A_{ij}=d\left(M_i,M_j\right)$

$$B_{ij}=\begin{cases}1 & (i,j)\in I_W\\ 0 & diğer\ yerlerde\end{cases}$$

$$C=S_{\boldsymbol{n}}\times r_{pb}(A,B)=(S_W/N_W-S_B/N_B)\frac{\sqrt{N_W N_B}}{N_T} \quad (63)$$

Eşitlik (63)'de gösterilen nokta iki serili indeks ne kadar küçük değer alırsa kümeleme o kadar başarılıdır (Baker & Hubert, 1975).

5.2.15.Ratkowsky-Lance indeksi

Ratkowski-Lance indeksi gruplar arası kareler matrisi ile genel kareler matrisi dikkate alınarak hesaplanır,

$$\bar{c}^2=\bar{R}=\frac{1}{p}\sum_{j=1}^{p}\frac{GAKT_j}{GKT_j}$$

Buna göre Ratkowsky-Lance indeksi eşitlik (64)ile gösterilmiştir (Ratkowsky & Lance, 1978):

$$C=\sqrt{\frac{\bar{R}}{K}}=\frac{\bar{c}}{\sqrt{K}} \quad (64)$$

4.2.16. Ray-Turi indeksi

Ray-Turi indeksi grup içi kareler toplamının farklı küme merkezleri arasındaki uzaklıkların minimumuna oranlanarak gözlem sayısına bölünmesi ile elde edilir. Eşitlik (65)'de Ray-Turi indeksi gösterilmektedir:

$$\min_{k<k'}\Delta_{kk'}^2=\min_{k<k'}d\left(G^{\{k\}},G^{\{k'\}}\right)^2=\min_{k<k'}\left\|G^{\{k\}}-G^{\{k'\}}\right\|^2$$

$$C=\frac{1}{N}\frac{G\dot{I}KT}{\min_{k<k'}\Delta_{kk'}^2} \quad (65)$$

Ray-Turi indeksi ne kadar küçük değer alırsa kümeleme o kadar başarılıdır. (Ray & Turi, 1999)

5.2.17. Scott-Symons indeksi

Scott-Symons indeksi her kümenin varyans-kovaryans matrisinin determinantının logaritma toplamlarının ağırlıklandırılmasıdır. Eşitlik (66)'da gösterilmektedir:

$$C = \sum_{k=1}^{K} n_k \log det\left(\frac{G\dot{I}DM^{\{k\}}}{n_k}\right) \quad (66)$$

$G\dot{I}DM^{\{k\}}$, 0'a eşittir ya da büyüktür 0'dan. Çünkü matrisler pozitiftir. Eğer sıfıra eşitse indeks tanımsızdır. İndeks ne kadar küçük değer alırsa kümeleme o kadar başarılıdır. (Scott & Symons, 1971)

5.2.18. Sd indeksi

S ve D sırasıyla kümeler için ortalama dağılım ve kümeler arası toplam dağılım olarak tanımlanır.

Veri kümesindeki her değişken için varyans vektörü şöyledir:

$$V = \left(Var(V_1), \dots, Var(V_p)\right)$$

Benzer şekilde her küme için bir varyans vektörü tanımlanırsa;

$$V^{\{k\}} = \left(Var\left(V_1^{\{k\}}\right), \dots, Var\left(V_p^{\{k\}}\right)\right)$$

Bu verilere dayanarak S yani kümeler için ortalama dağılım şu şekilde tanımlanır:

$$S = \frac{\frac{1}{K}\sum_{k=1}^{K}\left\|V^{\{k\}}\right\|}{\|V\|}$$

D yani kümeler arası toplam dağılım için aşağıdaki eşitlikler tanımlanmıştır:

$$D_{max} = \max_{k \neq k'} \left\| G^{\{k\}} - G^{\{k'\}} \right\|$$

$$D_{min} = \min_{k \neq k'} \left\| G^{\{k\}} - G^{\{k'\}} \right\|$$

$$D = \frac{D_{max}}{D_{min}} \sum_{k=1}^{K} \frac{1}{\sum_{\substack{k'=1 \\ k' \neq k}}^{K} \left\| G^{\{k\}} - G^{\{k'\}} \right\|}$$

Yukarıdaki S ve D eşitlikler kullanılarak SD indeksi eşitlik (67)'de tanımlanmıştır:

$$C = \alpha S + D \tag{67}$$

SD indeksi ne kadar küçük değer alırsa kümeleme o kadar başarılıdır. (Halkidi & Vazirgiannis, 2001)

5.2.19. S_Dbw indeksi

S_Dbw indeksinde ilk olarak farklı küme merkezleri arasındaki mesafelerin orta noktalarının, küme merkezleri arasındaki uzaklıkların maksimumuna oranlanır. Daha sonra bu oranların toplamının 2 ile çarpılarak, küme sayısı ve küme sayısının bir eksiğinin çarpımına oranı elde edilir. Kümelerin varyanslarının toplamının küme sayısına bölümünün genel varyansa oranı elde ediir. Son olarak eşitlik (67)'de görüldüğü gibi bu sonuçlar toplanır ve indeks sonucu elde edilir.

$$\mathrm{R_{kk'}} = \frac{\gamma_{\mathrm{kk'}}(\mathrm{H_{kk'}})}{\max\left(\gamma_{\mathrm{kk'}}(G^{\{k\}}), \gamma_{\mathrm{kk'}}(G^{\{k'\}})\right)}$$

$$G = \frac{2}{K(K-1)} \sum_{k<k'} \mathrm{R_{kk'}}$$

$$S = \frac{\frac{1}{K} \sum_{k=1}^{K} \left\| V^{\{k\}} \right\|}{\|V\|}$$

$$C = S + G \tag{68}$$

İndeks ne kadar küçük değer alırsa kümeleme o kadar başarılıdır. (Halkidi & Vazirgiannis, 2001)

5.2.20. Tau indeksi

Tau indeksi birimler arasındaki uzaklıkların sayısı dikkate alınarak oluşturulmuştur. (Friedman & Rubin, 1967)

$$C=\frac{S^{+}-S^{-}}{\sqrt{N_B N_W\left(\frac{N_T(N_T-1)}{2}\right)}} \quad (69)$$

Eşitlik (68)'deki Tau indeksinin büyük değer olması istenir.

5.2.21. İz_W indeksi

İz_W indeksi grup içi dağılım matrisinin köşegen elemanlarının toplamını ifade eder. İz_W indeksinin küçük değer alması istenir (Edwards & Cavalli-Sforza, 1965)

$$C=\text{İz}(\text{GİDM})=\text{GİKT} \quad (70)$$

5.2.22. İz_wib indeksi

İz_WiB indeksi grup içi dağılım matrisinin tersi ile gruplar arası dağılım matrisinin çarpımı sonucu elde edilen matrisin köşegen elemanlarının toplamıdır. Eşitlik (70)'de indeks formülize edilmiştir. İz_WiB indeksinin küçük değer alması istenir. (Friedman & Rubin, 1967)

$$C=\text{İz}(G\text{İ}DM^{-1}.GADM) \quad (71)$$

5.2.23. Wemmert-Gançarski indeksi

Wemmert-Gançarski indeksi tüm gözlem noktaları ile küme merkezleri arasındaki mesafelerin dikkate alınarak oluşturulmuştur.

R(M)=$\frac{\|M-G^{\{k\}}\|}{\min_{k'\neq k}\|M-G^{\{k'\}}\|}$

$J_k = max\left\{0{,}1 - \frac{1}{n_k}\sum_{i\in I_k} R(M_i)\right\}$

J_k, her noktanın küme merkezlerine uzaklıklarının ortalamasını kullanır.1'den büyükleri göz ardı eder.

C=$\frac{1}{N}\sum_{k=1}^{K} n_k J_k$

C=$\frac{1}{N}\sum_{k=1}^{K} max\{0, n_k - \sum_{i\in I_k} R(M_i)\}$ (72)

Wemmert-Gancarski indeksinin büyük değer alması itenir (wemmert&Gancarski, 2007).

5.2.24. Calinski ve Harabasz indeksi

Calinskİ ve Harabasz tarafından 1974 yılında önerilmiştir. Eşitlik (73), kümeler arası kareler matrisi ile küme içi kareler matrisi baz alınarak hesaplanır.

CH_k =$\frac{iz(\mathbf{B})(n-k)}{iz(\mathbf{W})(k-1)}$ (73)

Küme sayısı sabit tutularak farklı iki algoritmadan elde edilen kümelemeden CH_k değerini büyük yapan kümeleme en iyi kümelemedir (Calinski & Harabasz, 1974)

5.2.25. Davies-Bouldin indeksi

Davies-Bouldin tarafından 1979 yılında önerilmiştir. Küme içindeki gözlemlerin küme merkezine uzaklıklarını minimum, kümeler arası uzaklıkları maksimum yapmayı amaçlar. *DB(k)* ile gösterilir.

i=1,2,…,k

j=1,2,…,k

R_i=$\max_{i\neq j}\left(\frac{e_i+e_j}{d_{ij}}\right)$

$$DB(k)=\frac{1}{k}\sum_{i=1}^{k} R_i \quad (74)$$

DB(k) değerini minimum yapan k sayısı uygun küme sayısıdır. (Davies & Bouldin, 1979)

5.2.26. Dunn indeksi

Dunn tarafından 1974 yılında önerilmiştir. Yüksek yoğunluklu ve iyi bölünmüş kümelerin belirlenmesinde etkilidir. G_i ve G_j ile gösterilen $D(G_iG_j)$ uzaklığı sırasıyla G_i ve G_j kümelerinde yer alan **x** ve **y** gözlem çiftleri arasındaki minimum uzaklık olarak

$$D(G_iG_j)=\min_{\mathbf{x}\in G_i,\mathbf{y}\in G_j} d(\boldsymbol{x},\mathbf{y})$$

G_i, i. kümenin içerisinde yer alan gözlem çiftleri arasındaki uzaklıklar dikkate alınarak i. kümenin çapı $çap(G_i)=\max_{\mathbf{x},\mathbf{y}\in G_i} d(\mathbf{x},\mathbf{y})$ olmak üzere Dunn indeksi ise eşitlik (74)'de verilmiştir. Dunn indeksi büyük olan kümeleme geçerli kümelemedir (Dunn, 1974).

$$DN(K)=\min_{i=1,2,\dots,K}\left(\min_{J=i+1,\dots,K}\left(\frac{D(G_iG_j)}{\max_{l=1,2,\dots,K} çap(G_l)}\right)\right) \quad (75)$$

5.2.27. Xie-Beni indeksi

Xie-Beni indeksi bulanık kümeleme için kullanılan bir indeksdir. Ancak aynı zamanda net kümeleme için de geçerlidir. Farklı kümelerdeki gözlemlerin birbiri arasındaki en küçük mesafeleri baz alan bir indeksdir. Xie-Beni indeksinin minimum değer alması istenir. Xie-Beni indeksi eşitlik (75)'de formülize edilmiştir: (Xie & Beni., 1991)

$$\delta_1(C_k,C_{k'})=\min_{\substack{i\in I_k\\ j\in I_{k'}}} d(M_i,M_j) \quad (76)$$

6. UYGULAMA

Çalışmanın bu kısmında 5. bölümde verilen küme geçerlilik indekslerinden 26 içsel kriter, aşamalı ve aşamalı olmayan kümeleme yöntemleri kullanılarak Tablo 5'de özellikleri verilen veri setleri üzerinde uygulanmıştır. Veri setlerinden farklı küme sayıları için küme geçerlilik indekslerinin değerleri dikkate alınarak uygun küme sayıları belirlenmiştir. Kullanılan veri setleri https://archive.ics.uci.edu/ml/machine-learning-databases adresinden temin edilmiştir.

Tablo 5: Veri Setlerinin Özellikleri

Veri seti	Küme Sayısı	Değişken Sayısı	Gözlem Sayısı
Abalone (Deniz canlısı tür ismi)	3	8	4177
Balance (Denge ölçeği)	3	4	625
Breast-cancer-wisconsin (ABD'nin Wisconsin şehrinde bir klinikte meme kanseri araştırması)	2	10	699
Glass (Cam)	7	9	213
Haberman (Haberman'ın kanser araştırması)	2	3	304
İmagesegmentation(Görüntü pikselleri)	7	19	210
İris (Bitki Türü)	3	4	150
İonosphere(Sinyal Alma Aleti)	2	34	351
Sonar all (Deniz Radarı)	2	60	208
Statlogausturilan (kredi kartı Avusturalya'daaraştırması)	2	4	690
Statlogheart (kalp hastaları araştırması)	2	13	270
Teaching (Öğretme)	3	5	151

Uygulama bölümünde öncelikle veri setlerinde değişkenler arası ilişkinin yüksek olması durumunda küme geçerlilik indekslerinin k-ortalamalar yöntemiyle elde edilen kümeleme sonuçları simulasyon çalışması ve uygulama verisi üzerinde değerlendirildi.

Matlab programı üzerinden 26 adet küme değerlendirme kriteri, uzaklık ölçüleri, k- ortalamalar yöntemi kodları yazılarak analizler gerçekleştirilmiştir. Tablo 5'de bilgisi verilen veri setleri öncelikle kümeleme yöntemlerinden aşamalı kümeleme yöntemleri ile aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yöntemi ile ayrı ayrı test edildi ve burada veri setleri kümelere ayrıldı. Daha sonra elde edilen kümeleme sonuçları küme geçerlilik indekslerinden içsel kriterler ile test edip karşılaştırılması gerçekleştirildi.

6.1. Yüksek İlişkili Verilerde Küme Geçerlilik İndekslerinin Karşılaştırılması

Bu bölümde veri setinde değişkenler arasında yüksek ilişki olması durumunda k-ortalamalar yöntemi ile gerçekleştirilen kümeleme sonuçları kullanılarak küme geçerlilik indekslerinin performansları karşılaştırılmıştır.

Matlab programı kullanarak üretilen sentetik veri setleri üzerinde ve gerçek veri seti olan abalone veri setine uygulanan kümeleme analizi sonucu elde edilen sonuçların içsel kriterlerle değerlendirilerek uygun küme sayısı tahmin edilmeye çalışılmıştır. Simülasyon çalışmasında ilişkili bir veri seti üretilip farklı küme sayıları ile analiz yapılmıştır. Simülasyon çalışmasında çok değişkenli normal dağılımdan değişkenler arasında yüksek ilişki($r_{ij} \geq 0.90$) olacak şekilde sentetik veri setleri belirli küme ve değişken sayılarında üretilmiştir. Üretilen veri setleri için işlemler 1000 kez tekrarlanarak içsel kriterlerin doğru küme sayılarını belirtme sayıları Tablo 6'da verilmiştir. Tablo 6'da %70'den büyük değerler , (1000 denemede 700 ve üzeri doğru sayıları) farklı renkte belirtilmiştir.

Tablo 6: Yüksek İlişkili Sentetik Veriler için İçsel Kriterlerin Performansları

Küme sayısı	3				4			5	
Değişken sayısı	2	3	4	2	3	4	2	3	4
Calinski-Harabasz	631	682	725	555	584	566	504	495	495
Ball-Hall	990	878	902	2	2	2	0	0	0
Banfeld-Raftery	448	145	25	427	231	143	709	820	897
C indeksi	986	846	883	604	651	633	479	483	489
Davies-Bouldin	944	808	819	609	630	615	553	549	569
Dunn	986	847	883	604	648	631	461	465	460
Det_Ratio	19	159	135	295	282	306	163	109	88
Baker-Hubert	0	4	1	0	0	0	0	0	0
Mc-Clain	18	88	100	203	200	215	618	606	606
Ksq_DetW	987	843	873	293	59	67	269	87	96
Log_Det_Ratio	8	75	55	348	300	321	417	431	415
G Plus	0	4	1	0	0	0	0	0	0
Ray_Turi	987	850	886	604	649	630	460	464	458
Scott	18	0	0	74	27	13	924	869	755
Log_SS_Ratio	12	121	85	369	305	340	469	489	477
Ratkovsky-Lance	218	165	89	0	0	0	0	0	0
PBM	767	741	796	587	625	611	555	539	546
POBİ	10	88	66	197	137	123	972	978	988
SD	859	679	604	602	636	594	523	542	576
SDBW	397	167	75	413	302	236	570	552	578
Silhoutte	987	850	886	603	606	456	460	449	366
Tau	0	4	1	0	0	0	0	0	0
Trace_W	987	835	873	214	199	193	0	0	0
Trace_Wib	784	659	668	608	660	635	715	676	686
Wemgan	984	845	872	603	645	622	550	533	515
Xie_Beni	985	851	886	603	649	632	459	466	463

Tablo (6)'da üretilen sayılar için elde edilen sonuçlar verilmiştir. Buna göre küme sayısının 3 olması durumunda Ball-Hall indeksi, C indeksi, Davies-Bouldin indeksi, Dun indeksi, Ksq_DetW indeksi, Ray-Turi indeksi, PBM indeksi, Silhoutte indeksi, Trace_W indeksi, Trace_Wib indeksi, Wemmert-Gancarski indeksi ve Xie-Beni indeksinin küme sayısını tahmin etmede daha başarılı olduğu görülmektedir.

Küme sayısının 5 olması durumunda ise Banfeld-Raftery indeksi, Scott indeksi, POBİ indeksinin diğerlerine göre küme sayısını tahmin etmede başarılı olduğu görülmektedir.

Tablo 7: Sentetik ve Abalone veri seti için elde edilen sonuçların karşılaştırılması

Değişken sayısı	Simulasyon			Uygulama Abalone verisi		
	2	3	4	2	3	4
Baker-Hubert	0	4	1	5	5	5
Ball-Hall	990	878	902	3	3	3
Banfeld-Raftery	448	145	25	2	2	2
c indeksi	986	846	883	5	5	5
Calinski-Harabasz	631	682	725	3	2	5
Davies-Bouldin	944	808	819	5	4	4
Det_Ratio	19	159	135	3	3	3
Dunn	986	847	883	2	2	2
G Plus	0	4	1	5	5	5
Ksq_DetW	987	843	873	4	4	4
Log_Det_Ratio	8	75	55	5	5	5
Log_SS_Ratio	12	121	85	2	2	2
Mc-Clain	18	88	100	2	2	3
PBM	767	741	796	5	5	5
POBİ	10	88	66	5	5	5
Ratkovsky-Lance	218	165	89	2	2	2
Ray_Turi	987	850	886	5	5	5
Scott	18	0	0	5	5	5
SD	859	679	604	5	5	5
SDBW	397	167	75	5	5	5
Silhoutte	987	850	886	3	3	3
Tau	0	4	1	2	2	2
Trace_W	987	835	873	3	3	3
Trace_Wib	784	659	668	5	5	5
Wemgan	984	845	872	2	2	2
Xie_Beni	985	851	886	5	5	4

Tablo 7'de üretilen ve abalone veri seti için elde edilen sonuçlar verilmiştir. Buna göre ilk 3 sütunda k=3 ve p=2,3,4 için üretilen sayıların sonuçları, son 3 sütunda abalone veri seti için elde edilen sonuçlar yer almaktadır. Abalone veri seti için Ball-Hall indeksi, Det_Ratio indeksi, Silhouette indeksi ve Trace_w indeksi başarılıdır. Üretilen veri seti ve abalone veri seti için ortak bir değerlendirme yapılacak olursa Ball-Hall indeksi ve Silhouette indeksi başarılıdır denilebilir.

6.2. Aşamalı Olmayan Kümeleme Analizinde Küme Geçerlilik İndekslerinin Karşılaştırılması

Bu bölümde uygulama verileri için elde edilen sonuçlar doğrultusunda küme geçerlilik indekslerinin her bir veri seti için aşamalı olmayan kümeleme algoritması k-ortalamalar yöntemiyle elde edilen kümeleme yapısına göre indekslerin küme sayısını doğru tahmin etme performansları Tablo 8'de verilmiştir.

Tablo (8)'e göre k-ortalamalar yöntemiyle; Bak-Hub indeksi, G+ indeksi, Silhouette indeksi, Tau indeksi, Trace_W indeksi küme sayısını en çok doğru tahmin eden indeksler olmuşlardır. Yani en iyi performansı bu indeksler göstermişdir.

Tablo 8: k-ortalamalar yöntemine göre içsel kriterlerin performansları

İndis/Veri Seti	Abalone	Balance	(BCW)	Glass	Haberman	İris	Sonar all	Statlogaua	Statloghe	Teaching	Toplam
Call-Har			X			X	X		X		4
Ball-Hall	X	X				X				X	4
Ban-Raf											
C			X					X			2
Dav-Boul			X								
Dunn			X					X			2
Det-Ratio	X										
Bak-Hub			X		X		X	X	X		5
Mc Clain								X			1
Ksq				X							1
Log-DetRatio											
G+			X		X		X	X	X		5
Ray-Turi	X		X					X	X		4
Scott							X				1
Log-SS-Ratio											
Rat-Lance		X	X					X			3
Pbm			X			X				X	3
Pobi							X				1
Sd			X		X						2
SdbW											
Silhouette	X		X		X	X	X				5
Tau			X		X		X	X	X		5
TraceW	X	X		X		X				X	5
TraceWiB											
Wem-Gan			X				X	X	X		4
Xie-Beni	X		X					X			3

6.3. Aşamalı Kümeleme Analizinde Küme Geçerlilik İndekslerinin Karşılaştırılması

Bu bölümde aşamalı kümeleme algoritmalarına göre indekslerin küme sayısını doğru tahmin etme sayılarına göre performansları Tablo 9-15'de verilmiştir.

Tablo 9'da tek bağlantı yöntemine göre içsel kriterlerin performansları verilmiştir. Tablo 9'a göre PBM indeksi, Dunn indeksi, G+ indeksi, Ray-Turi indeksi, Tau indeksi, Wemmert-Gan. indeksi küme sayısını en çok doğru tahmin eden indeksler olmuşlardır.

Tablo 9: Tek bağlantı yöntemine göre içsel kriterlerin performansları

İndis/Veri Seti	Abalone	Balance	(BCW)	Glass	Haberman	İris	Sonar all	Statlogaua	Statloghe	Teaching	Toplam
Call-Har					X				X		2
Ball-Hall	X	X				X				X	4
Ban-Raf											0
C					X			X	X		3
Dav-Boul					X			X	X	X	4
Dunn			X		X		X	X	X	X	6
Det-Ratio		X								X	2
Bak-Hub			X		X			X	X		4
Mc Clain					X			X	X	X	4
Ksq										X	1
Log-DetRatio		X									1
G+			X		X		X	X	X		5
Ray-Turi	X		X		X			X	X		5
Scott							X				1
Log-SS-Ratio											0
Rat-Lance								X			1
Pbm	X		X		X	X	X	X	X		7
Pobi			X		X		X				3
Sd			X								1
SdbW											
Silhouette	X										1
Tau			X		X		X	X	X		5
TraceW				X						X	2
TraceWiB				X							1
Wem-Gan			X		X		X	X	X		5
Xie-Beni			X		X			X	X		4

Tablo 10'da tam bağlantı yöntemine göre içsel kriterlerin performansları verilmiştir. Tablo 10'a göre tam bağlantı yöntemiyle; Baker-Hubert indeksi, G+ indeksi, Tau indeksi küme sayısını en çok doğru tahmin eden indeksler olmuşlardır.

Tablo 10: Tam bağlantı yöntemine göre içsel kriterlerin performansları

İndis/Veri Seti	Abalone	Balance	(BCW)	Glass	Haberman	İris	Sonar all	Statlogaua	Statloghe	Teaching	Toplam
Call-Har					X					X	2
Ball-Hall		X				X				X	3
Ban-Raf											0
C									X		1
Dav-Boul			X			X			X		3
Dunn							X		X		2
Det-Ratio	X	X				X				X	4
Bak-Hub			X		X		X	X	X		5
Mc Clain									X		1
Ksq	X	X				X				X	4
Log-DetRatio				X							1
G+			X		X		X	X	X		5
Ray-Turi			X					X	X		3
Scott							X				1
Log-SS-Ratio											0
Rat-Lance								X			1
Pbm							X				1
Pobi			X		X		X				3
Sd			X		X						2
SdbW											0
Silhouette					X	X		X		X	4
Tau			X		X		X	X	X		5
TraceW						X				X	2
TraceWiB				X							1
Wem-Gan						X	X	X	X		4
Xie-Beni									X		1

Tablo 11'de ortalama bağlantı yöntemine göre içsel kriterlerin performansı verilmiştir. Tablo 11'e göre ortalama bağlantı yöntemiyle; Ball-Hall indeksi, Ray-Turi indeksi, Tau indeksi küme sayısını en çok doğru tahmin eden indeksler olmuşlardır.

Tablo 11: Ortalama bağlantı yöntemine göre içsel kriterlerin performansları

İndis/Veri Seti	Abalone	Balance	(BCW)	Glass	Haberman	İris	Sonar all	Statlogaua	Statloghe	Teaching	Toplam
Call-Har						X				X	2
Ball-Hall	X	X	X			X				X	5
Ban-Raf											0
C					X			X			2
Dav-Boul	X		X		X			X			4
Dunn					X		X	X	X		4
Det-Ratio		X								X	2
Bak-Hub					X		X	X	X		4
Mc Clain			X		X			X	X		4
Ksq			X							X	2
Log-DetRatio				X							1
G+					X		X	X	X		4
Ray-Turi			X		X		X	X	X		5
Scott				X			X				2
Log-SS-Ratio			X								1
Rat-Lance		X	X								2
Pbm					X	X	X	X			4
Pobi			X		X		X				3
Sd			X				X				2
SdbW											0
Silhouette		X				X			X	X	4
Tau			X		X		X	X	X		5
TraceW		X		X		X				X	4
TraceWiB	X										1
Wem-Gan					X		X	X	X		4
Xie-Beni					X		X	X			3

Tablo 12'de ağırlıklı ortalama bağlantı yöntemine göre içsel kriterlerin performansları verilmiştir. Tablo 12'ye göre ağırlıklı ortalama bağlantı yöntemiyle; PBM indeksi, Calinski-Harabasz indeksi, Buker-Hubert indeksi, G+ indeksi, Tau indeksi, Wemmert-Gancarski indeksi küme sayısını en çok tahmin eden indeksler olmuştur.

Tablo 12: Ağırlıklı ortalama bağlantı yöntemine göre içsel kriterlerin performansları

İndis/Veri Seti	Abalone	Balance	(BCW)	Glass	Haberman	İris	Sonar all	Statlogaua	Statloghe	Teaching	Toplam
Call-Har			X	X		X	X			X	5
Ball-Hall	X	X				X				X	4
Ban-Raf											
C					X			X	X		3
Dav-Boul			X		X			X	X		4
Dunn					X			X	X		3
Det-Ratio		X								X	2
Bak-Hub			X		X		X	X	X		5
Mc Clain					X			X	X		3
Ksq		X								X	2
Log-DetRatio											
G+			X		X		X	X	X		5
Ray-Turi			X		X			X			3
Scott							X				1
Log-SS-Ratio											
Rat-Lance		X	X								2
Pbm			X		X	X	X	X		X	6
Pobi			X		X		X				3
Sd			X				X				2
SdbW											
Silhouette			X			X	X			X	4
Tau			X		X		X	X	X		5
TraceW		X		X		X				X	4
TraceWiB	X										1
Wem-Gan			X		X		X	X	X		5
Xie-Beni	X				X			X	X		4

Tablo 13’de merkezi bağlantı yöntemine göre içsel kriterlerin performansları verilmiştir. Tablo 13’e göre merkezi bağlantı yöntemiyle; PBM indeksi, Ray-Turi indeksi, Buker-Hubert indeksi, G+ indeksi, Tau indeksi, Silhouette indeksi, Wemmert-Gancarski indeksi küme sayısını en çok doğru tahmin eden indeksler olmuştur.

Tablo 13: Merkezi bağlantı yöntemine göre içsel kriterlerin performansları

İndis/Veri Seti	Abalone	Balance	(BCW)	Glass	Haberman	İris	Sonar all	Statlogaua	Statloghe	Teaching	Toplam
Call-Har			X			X					2
Ball-Hall	X					X				X	3
Ban-Raf											0
C					X			X			2
Dav-Boul					X			X			2
Dunn			X		X		X	X			4
Det-Ratio										X	1
Bak-Hub			X		X		X	X	X		5
Mc Clain					X			X	X		3
Ksq				X						X	2
Log-DetRatio											0
G+			X		X		X	X	X		5
Ray-Turi			X		X		X	X	X		5
Scott							X				1
Log-SS-Ratio											0
Rat-Lance			X								1
Pbm			X		X	X	X	X			5
Pobi	X		X								2
Sd											0
SdbW											0
Silhouette		X	X			X			X	X	5
Tau			X		X		X	X	X		5
TraceW										X	1
TraceWiB	X										1
Wem-Gan			X		X		X	X	X		5
Xie-Beni	X				X		X	X			4

Tablo 14'de medyan bağlantı yöntemine göre içsel kriterlerin performansları verilmiştir. Tablo 14'e göre medyan bağlantı yöntemiyle; PBM , Davies-Bouldin, Buker-Hubert indeksi, Dunn indeksi, SD indeksi, Tau indeksi, Wemmert-Gancarski indeksi, Xie-Beni indeksi küme sayısını en çok tahmin eden indeksler olmuştur.

Tablo 14: Medyan bağlantı yöntemine göre içsel kriterlerin performansları

İndis/Veri Seti	Abalone	Balance	(BCW)	Glass	Haberman	İris	Sonar all	Statlogaua	Statloghe	Teaching	Toplam
Call-Har						X					1
Ball-Hall	X	X				X				X	4
Ban-Raf						X					1
C								X	X		2
Dav-Boul			X		X	X	X	X	X		6
Dunn			X			X	X	X	X		5
Det-Ratio		X								X	2
Bak-Hub			X		X		X	X	X		5
Mc Clain								X	X		2
Ksq						X					1
Log-DetRatio		X									1
G+						X	X	X	X		4
Ray-Turi							X				1
Scott							X				1
Log-SS-Ratio											0
Rat-Lance	X					X				X	3
Pbm	X	X				X	X	X	X		6
Pobi			X		X		X				3
Sd			X	X	X		X		X		5
SdbW											0
Silhouette	X					X				X	3
Tau			X		X		X	X	X		5
TraceW	X									X	2
TraceWiB											0
Wem-Gan					X	X	X	X	X		5
Xie-Beni			X			X	X	X	X		5

Tablo 15'de Ward bağlantı yöntemine göre içsel kriterlerin performansları verilmiştir. Tablo 15'e göre Ward bağlantı yöntemiyle; Calinski-Harabasz indeksi, Ball-Hall indeksi, Dunn indeksi, Baker-Hubert indeksi, G+ indeksi, Tau indeksi küme sayısını en çok tahmin eden indeksler olmuştur.

Tablo 15: Ward bağlantı yöntemine göre içsel kriterlerin performansları

İndis/Veri Seti	Abalone	Balance	(BCW)	Glass	Haberman	İris	Sonar all	Statlogaua	Statloghe	Teaching	Toplam
Call-Har			X			X	X		X	X	5
Ball-Hall	X	X		X		X				X	5
Ban-Raf											0
C											0
Dav-Boul			X				X		X		3
Dunn											5
Det-Ratio	X	X				X				X	4
Bak-Hub			X		X		X	X	X		5
Mc Clain											0
Ksq		X				X					2
Log-DetRatio											0
G+			X		X		X	X	X		5
Ray-Turi	X		X				X		X		4
Scott											0
Log-SS-Ratio											0
Rat-Lance			X					X		X	3
Pbm			X			X				X	3
Pobi					X		X				2
Sd			X		X		X				3
SdbW											0
Silhouette	X		X							X	3
Tau			X		X		X	X	X		5
TraceW	X	X				X				X	4
TraceWiB				X							1
Wem-Gan			X				X	X	X		4
Xie-Beni			X								1

6.4. Uzaklık Ölçülerine Göre Aşamalı Kümeleme Yöntemlerinde İçsel Kriterlerin Performanslarının Karşılaştırılması

Aşamalı kümeleme yöntemleri kullanılarak 10 gerçek veri seti üzerinde 8 farklı uzaklık ölçüsü kullanılarak elde edilen küme yapıları için içsel kriterlerin doğru küme sayılarını belirtme sayıları Tablo16'de verilmiştir. Tabloda her satır için en yüksek değer farklı renkte işaretlenmiştir. Tablonun son sütununda kriterlerin, aşamalı yöntemlerin kaç tanesinde en iyi sonucu verdiği belirtilmiştir. Böylece kümeleme yöntemlerine göre indekslerin performansları değerlendirilmiştir. Benzer şeklide en alt satırda her bir yöntemde kaç indeks değerinin en iyi değeri verdiği belirtilmiştir.

Tablo 16: Aşamalı kümeleme yöntemlerine göre içsel kriterlerin performansı

İndis/Yöntem	Tek Bağlantı	Tam Bağlantı	Ortalama Bağ.	Ağ.lı Ort. Bağ.	Merkezi Bağ.	Medyan Bağ.	Ward Bağlantı	
Call-Har	21	17	20	24	17	19	34	
Ball-Hall	32	22	26	24	24	25	26	
Ban-Raf	4	3	7	5	6	6	3	
C	20	12	17	19	11	17	2	
Dav-Boul	28	28	25	28	18	31	26	
Dunn	36	22	25	29	21	28	9	
Det-Ratio	12	20	14	16	11	13	26	
Bak-Hub	39	39	39	39	31	39	47	7
Mc Clain	24	12	21	20	14	20	5	
Ksq	9	14	9	10	13	9	17	
Log-DetRatio	2	3	2	1	0	1	1	
G+	39	39	39	39	31	39	47	7
Ray-Turi	36	32	35	34	25	27	30	
Scott	0	1	0	0	2	1	1	
Log-SS-Ratio	1	0	2	0	0	1	0	
Rat-Lance	13	18	11	13	15	20	27	
Pbm	40	28	41	39	28	35	37	3
Pobi	24	23	26	23	18	24	26	
Sd	11	26	27	23	16	22	40	
SdbW	0	4	4	2	2	2	2	
Silhouette	24	30	32	27	28	26	40	
Tau	39	39	39	39	31	39	47	7
TraceW	12	21	15	15	13	15	27	
TraceWiB	5	1	6	4	5	5	4	
Wem-Gan	35	38	32	32	28	38	40	2
Xie-Beni	24	18	20	30	13	25	9	
	4	4	4	4	3	4	3	

Tablo (16)'de aşamalı kümeleme yöntemlerinden tek bağlantı yöntemiyle elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla PBM indeksi, Baker-Hubert indeksi, G+ indeksi ve TAU indeksi olduğu görülmektedir.

Tablo (16)'de aşamalı kümeleme yöntemlerinden tam bağlantı yöntemiyle elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, TAU indeksi ve Wemmert-Gancarski indeksi olduğu görülmektedir.

Tablo (16)'de aşamalı kümeleme yöntemlerinden ortalama bağlantı yöntemiyle elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, PBM indeksi, G+ indeksi ve TAU indeksi olduğu görülmektedir.

Tablo (16)'de aşamalı kümeleme yöntemlerinden ağırlıklı ortalama bağlantı yöntemiyle elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla PBM indeksi, Baker-Hubert indeksi, G+ indeksi ve TAU indeksi olduğu görülmektedir.

Tablo (16)'de aşamalı kümeleme yöntemlerinden merkezi bağlantı yöntemiyle elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi ve TAU indeksi olduğu görülmektedir.

Tablo (16)'de aşamalı kümeleme yöntemlerinden medyan bağlantı yöntemiyle elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, TAU indeksi ve Wemmert-Gancarski indeksi olduğu görülmektedir.

Tablo (16)'de aşamalı kümeleme yöntemlerinden Ward bağlantı yöntemiyle elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi ve TAU indeksi olduğu görülmektedir.

Genel olarak tablo (16)'ü değerlendirecek olursak; kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indeksler G+ indeksi, TAU indeksi, Baker-Hubert indeksi olduğu görülmektedir. Tablo (16) aşamalı kümeleme yöntemleri açısından incelendiğinde gerçek küme sayısını en çok doğru tahmin eden kümeleme yöntemleri tek bağlantı ve ağırlıklı ortalama bağlantı yöntemidir.

Analiz sonuçlarına göre öneride bulunacak olursak kümeleme analizi yapılırken aşamalı kümeleme yöntemlerinden tek bağlantı veya ağırlıklı bağlantı yöntemi seçilebilir. Elde edilen sonuçlar test edilirken G+ indeksi, TAU indeksi ve Baker-Hubert indeksi kullanılabilir.

Aşamalı kümeleme yöntemleriyle kümeleme işlemi yapılırken kullanılan uzaklık ölçülerinin kümeleme analizine etkisi içsel kriterlerle test edilmiştir ve elde edilen kümeleme sonuçları incelendiğinde uzaklık ölçülerine göre gerçek küme sayısını en çok doğru tahmin eden indeksler 17. tabloda verilmiştir.

Tablo 17: Aşamalı kümeleme yöntemlerinde içsel kriterlerin uzaklık ölçülerine göre performansları

İndis/uzaklık Ölçüsü	öklid	seuclidean	cityblok	minkowski	chebychev	mahalanobis	cosine	correlation	
Call-Har	24	22	22	24	15	18	12	14	
Ball-Hall	24	16	25	23	24	17	24	26	
Ban-Raf	2	5	2	1	6	6	5	7	
C	12	16	18	13	12	16	9	8	
Dav-Boul	24	25	23	26	19	24	22	23	
Dunn	17	22	24	21	23	23	28	18	
Det-Ratio	18	16	11	18	12	14	14	9	
Bak-Hub	35	35	35	35	35	35	35	28	8
Mc Clain	13	16	18	17	18	19	10	14	
Ksq	15	10	10	15	7	9	8	7	
Log-DetRatio	1	2	1	0	0	4	2	0	
G+	35	35	35	35	35	35	35	28	8
Ray-Turi	27	26	28	29	23	31	29	33	
Scott	0	3	0	0	1	1	0	0	
Log-SS-Ratio	1	0	0	1	1	0	1	0	
Rat-Lance	18	15	15	18	14	12	12	14	
Pbm	33	30	39	35	33	36	25	18	2
Pobi	18	19	21	20	20	25	21	14	
Sd	19	22	17	16	12	30	25	19	
SdbW	0	6	0	0	0	10	0	1	
Silhouette	35	34	25	31	23	22	18	20	
Tau	35	35	35	35	35	35	35	28	8
TraceW	23	8	20	21	14	4	12	17	
TraceWiB	6	3	3	6	8	4	0	1	
Wem-Gan	32	32	34	33	29	29	28	29	2
Xie-Beni	19	16	26	21	18	19	14	10	
	4	4	5	3	3	4	3	4	

Tablo 17’de aşamalı kümeleme yöntemleriyle kümeleme işlemi yapılırken kullanılan Öklid uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, Silhouette indeksi ve TAU indeksi olduğu görülmektedir.

Tablo 17’de aşamalı kümeleme yöntemleriyle kümeleme işlemi yapılırken kullanılan Karesel Öklid uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, Silhouette indeksi ve TAU indeksi olduğu görülmektedir.

Tablo 17’de aşamalı kümeleme yöntemleriyle kümeleme işlemi yapılırken kullanılan City-Blok uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, PBM indeksi, TAU indeksi ve Wemmert-Gancarski indeksi olduğu görülmektedir.

Tablo 17’de aşamalı kümeleme yöntemleriyle kümeleme işlemi yapılırken kullanılan Minkowski uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi ve TAU indeksi olduğu görülmektedir.

Tablo 17’de aşamalı kümeleme yöntemleriyle kümeleme işlemi yapılırken kullanılan Chebychev uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi ve TAU indeksi olduğu görülmektedir.

Tablo 17’de aşamalı kümeleme yöntemleriyle kümeleme işlemi yapılırken kullanılan Mahalonobis uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, PBM indeksi ve TAU indeksi olduğu görülmektedir.

Tablo 17’de aşamalı kümeleme yöntemleriyle kümeleme işlemi yapılırken kullanılan Cosine uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi ve TAU indeksi olduğu görülmektedir.

Tablo 17’de aşamalı kümeleme yöntemleriyle kümeleme işlemi yapılırken kullanılan correlation uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, TAU indeksi ve Wemmert Gancarski indeksi olduğu görülmektedir

Genel olarak Tablo 17’i değerlendirecek olursak; kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indeksler G+ indeksi, TAU indeksi, Baker-Hubert indeksi olduğu görülmektedir. Tablo 17’de aşamalı kümeleme yöntemlerinde kullanılan uzaklık ölçüleri açısından gerçek küme sayısını en

çok doğru tahmin eden uzaklık ölçüleri Öklid, karesel Öklid, city-blok, minkowski, mahalonobis ve korelasyon uzaklık ölçüsüdür.

Analiz sonucuna göre öneride bulunulacak olursak kümeleme analizi yapılırken aşamalı kümeleme yöntemlerinde Öklid, karesel Öklid, City-Blok, Minkowski, Mahalonobis ve Korelasyon uzaklık ölçüleri kullanılarak analiz yapılmalı ve elde edilen sonuçlar test edilirken G+ indeksi, TAU indeksi ve Baker-Hubert indeksi kullanılmalıdır.

Tablo 14 ve Tablo 15 incelendiğinde indeksler açısından en iyi sonucu G+, TAU ve Baker-Hubert indeksleri vermiştir. Kümeleme analizi sonucu elde edilen kümelemeler bu indeksler kullanılarak test edilmelidir.

6.5. Uzaklık Ölçülerine Göre k-Ortalamalar Yönteminde İçsel Kriterlerin Performanslarının Karşılaştırılması

Aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yöntemi kullanılarak 10 gerçek veri seti, 8 farklı uzaklık ölçüsü kullanılarak içsel kriterlerin doğru küme sayılarını belirtme sayıları Tablo18'de verilmiştir.

Tablo 18: k-ortalamalar yönteminde içsel kriterlerin uzaklık ölçülerine göre performansları

İndis/Uzaklık ölçüsü	oklid	seuclidean	cityblok	minkowski	chebychev	mahalanobis	cosine	correlation	
Call-Har	3	3	4	2	4	2	4	4	
Ball-Hall	2	2	4	5	4	3	4	3	1
Ban-Raf	0	1	0	0	1	0	1	0	
C	1	0	2	1	2	0	0	1	
Dav-Boul	4	3	4	3	3	0	5	5	1
Dunn	2	2	2	2	3	2	4	2	
Det-Ratio	4	2	2	4	1	1	1	2	
Bak-Hub	5	4	5	4	5	5	5	4	4
Mc Clain	1	0	2	2	1	0	0	2	
Ksq	1	2	2	2	1	3	2	2	
Log-DetRatio	0	0	0	1	0	0	0	0	
G+	5	4	5	4	5	5	5	4	4
Ray-Turi	5	3	4	5	5	0	5	5	4
Scott	0	1	0	1	0	2	1	0	
Log-SS-Ratio	0	0	0	1	0	0	0	0	
Rat-Lance	4	5	4	2	5	2	4	3	2
Pbm	3	3	4	4	4	2	3	5	1
Pobi	1	3	1	3	1	4	2	1	
Sd	4	4	5	3	4	5	7	5	4
SdbW	0	0	0	1	0	1	0	0	
Silhouette	3	4	5	2	6	5	5	4	2
Tau	5	4	5	5	5	5	5	4	4
TraceW	3	4	4	5	4	1	4	4	1
TraceWiB	0	0	0	0	0	1	0	1	
Wem-Gan	4	3	4	4	5	0	5	3	1
Xie-Beni	3	2	2	2	2	2	2	2	
	3	1	5	3	6	5	1	4	

Tablo 18'de k-ortalamalar yöntemi ile kümeleme işlemi yapılırken kullanılan Öklid uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, Ray-Turi indeksi olduğu görülmektedir.

Tablo 18'de aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yöntemi ile kümeleme işlemi yapılırken kullanılan karesel öklid uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Ratkowsky-Lance indeksi olduğu görülmektedir.

Tablo 18'de aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yöntemi ile kümeleme işlemi yapılırken kullanılan City-Blok uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, SD indeksi, Silhouette indeksi, Tau indeksi, olduğu görülmektedir.

Tablo 18'de aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yöntemi ile kümeleme işlemi yapılırken kullanılan Minkowski uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Ball-Hall indeksi, Ray-Turi indeksi, Tau indeksi, TraceW indeksi olduğu görülmektedir.

Tablo 18'de aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yöntemi ile kümeleme işlemi yapılırken kullanılan Chebychev uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, Ray-Turi indeksi, Ratkowsky-Lance indeksi, Tau indeksi, Wemmert-Gancarsky indeksi olduğu görülmektedir.

Tablo 18'de aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yöntemi ile kümeleme işlemi yapılırken kullanılan Mahalonobis uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Baker-Hubert indeksi, G+ indeksi, SD indeksi, Silhouette indeksi, Tau indeksi olduğu görülmektedir.

Tablo 18'de aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yöntemi ile kümeleme işlemi yapılırken kullanılan Cosine uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indeks SD indeksi olduğu görülmektedir.

Tablo 18'de aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yöntemi ile kümeleme işlemi yapılırken kullanılan Correlation uzaklık ölçüsü ile elde edilen kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indekslerin sırasıyla Davies-Bouldin indeksi, Ray-Turi indeksi, PBM indeksi, SD indeksi olduğu görülmektedir.

Genel olarak Tablo 18'i değerlendirecek olursak; kümeleme sonuçları incelendiğinde gerçek küme sayısını en çok doğru tahmin eden indeksler G+ indeksi, TAU indeksi, Baker-Hubert indeksi, Ray-Turi indeksi ve SD indeksi olduğu görülmektedir. Tablo 18'de aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yönteminde kullanılan uzaklık ölçüleri açısından gerçek küme sayısını en çok doğru tahmin eden uzaklık ölçüleri city-blok, chebychev, mahalonobis uzaklık ölçüsüdür.

Analiz sonucuna göre öneride bulunulacak olursak kümeleme analizi yapılırken aşamalı olmayan kümeleme yöntemi olan k-ortalamalar yönteminde City-Blok, Chebychev, Mahalonobis uzaklık ölçüleri kullanılarak analiz yapılmalı ve elde edilen sonuçlar test edilirken G+ indeksi, Ray-Turi indeksi, Baker-Hubert indeksi, SD indeksi ve TAU indeksi kullanılmalıdır.

7. SONUÇLAR VE ÖNERİLER

Tez çalışmasında 3. bölümde uzaklık ölçüleri tanıtılmış, 4. bölümde aşamalı ve aşamalı olmayan kümeleme analizi yöntemleri üzerinde durulmuştur. 5. bölümde ise kümeleme analizinin temel problemlerinden olan küme sayısının belirlenmesi problemi için geliştirilen küme geçerlilik indeksleri tanıtılmıştır.

Tezin uygulama bölümünde Tablo 5'de belirtilen veri setleri üzerinde farklı uzaklık ölçüleri kullanılarak, aşamalı ve aşamalı olmayan kümeleme yöntemleri ile veri setleri için kümeleme yapıları oluşturulmuş ve oluşturulan bu kümeleme yapıları küme geçerlilik indekslerinden içsel kriterler kullanılarak küme sayıları belirlenmiştir. Veri setleri incelendiğinde biyoloji, sağlık, endüstri, bilgisayar bilimleri, havacılık ve uzay bilimleri, sosyal bilimler ve eğitim gibi farklı alanlardan veri setleri seçildiği görülmektedir.

Kümeleme analizinin günümüzde birçok araştırmacı tarafından sıklıkla kullanılan bir analiz olduğu ve analiz içerisinde farklı yöntemlerin seçilme olasılığı göz önüne alınmıştır. Tezin uygulama bölümünde öncelikle araştırıcı veri setleri ile ilgili alan bilgileri göz ardı edildiğinde, ham veri setinden elde edilecek ön bilgiye göre küme sayısının belirlenmesinde içsel kriterlerden acaba hangisini kullanmalıdır? sorusuna cevap aranmıştır. Bu amaçla değişkenler arasındaki korelasyonlar dikkate alınarak bu konu k-ortalamalar yöntemi kullanılarak incelenmiştir. Hem simülasyon çalışması hem de gerçek veri seti üzerindeki sonuçlar incelendiğinde değişkenler arasında yüksek ilişki olması durumunda içsel kriterlerden Ball-Hall ve Silhoutte indeksinin simülasyon ve uyulama sonuçlarının benzer olduğu görülmüştür.

Aşamalı kümeleme analizi yöntemlerinin ve k-ortalamalar yöntemi sonuçlarının içsel kriterlerle değerlendirildiği 6.2 ve 6.3 bölümleri incelendiğinde yöntemlerin 10 veri setinden en çok 7'sinda doğru küme sayısını belirttiği görülmüştür. Her kümeleme yönteminde de en iyi sonucu veren bir geçerlilik indeksi gözlenmemiştir. Ancak Pbm, Tau, G+ ve Bak-Hub indeksleri diğerlerinden daha iyi sonuç vermiştir.

Farklı uzaklık ölçüleri kullanılarak aşamalı ve k-ortalamalar için elde edilen kümeleme sonuçlarına göre içsel kriterlerin değerlendirildiği 6.4. ve 6.5. bölümleri incelendiğinde;

Aşamalı kümeleme yöntemlerinin tüm uzaklık ölçüleri için genel değerlendirilmesinde ve özel olarak her bir uzaklık ölçüsü için Bak-Hub, G+ ve Tau indeksleri en iyi performansı göstermiştir.

Aşamalı kümeleme yöntemlerinde correlation uzaklık ölçüsü için içsel kriterlerin en kötü performansı gösterdiği bulunmuştur.

Aşamalı olmayan k-ortalamalar yöntemi için tüm uzaklık ölçüleri için genel değerlendirildiğinde Bak-hub, G+, Tau, Ray-Turi ve Sd indeksleri en iyi performansı göstermiştir.

Aşamalı olmayan k-ortalamalar yöntemi için Chebychev ve cosine uzaklık ölçüleri ile içsel kriterler en iyi performansı sergilerken, seuclidean (standartlaştırılmış öklid) uzaklı ölçüsü için en kötü performansı sergilemiştir.

Sonuç olarak kümeleme analizinde küme sayısını belirlemede kullanılan kümeleme yönteminin ve uzaklık ölçüsünün içsel kriterler üzerinde etkileri ortaya konmakla beraber veri setinin hangi alandan geldiğinin de önem arz ettiği her veri seti için aynı kümeleme yöntemi ve uzaklık ölçüsünün kullanılmasının uygun olmayacağı sonucuna varılmıştır. Uygulamaların çoğunda k-ortalamalar yöntemi ve standart olarak Öklid uzaklığı kullanıldığı literatür araştırmamızda belirtilmiştir. Bununla beraber uygulama bölümünde öklid uzaklık ölçüsü ve k-ortalamalar yöntemi kullanılarak elde edilen sonuçlarda Öklid uzaklık ölçüsü için içsel kriterlerin iyi sonucu vermediği gözlenmiştir. Araştırmacıların kümeleme analizini kullanırken kendi alanları ile ilgili çalışmaları baz alarak alana ilişkin verilerin özel olarak değerlendirilip alanla ilgili uygun kümeleme yöntemini ve uzaklık ölçüsünü kullanarak çalışmalarını gerçekleştirmelerinin daha uygun olacağı sonucuna varılmıştır.

KAYNAKLAR

Alpar, R. (2017). *Çok Değişkenli İstatistiksel Yöntemler.*

Artur Starczewski, A. K. (2017). A Study of Cluster Validity Indices for Real-Life Data. *Artificial Intelligence and Soft Computing* , 148-158.

Baker, F. B., & Hubert, L. J. (1975). Measuring the power of hierarchical clusteranalysis. *Journal of the American Statistical Association*, 31-38.

Ball, G., & Hall, D. (1965). A novel method of data analysis and pattern classification. *Menlo Park: Stanford Research Institute*.

Banfeld, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering.

Ben Said, A., Hadjidj, R., & Foufou, S. (2017). Cluster validity index based on Jeffrey divergence. *Pattern Analysis ve Applications*, s. 21-31.

Boseop Kim, H. L. (2018). Integrating cluster validity indices based on data envelopment analysis. *Applied Soft Computing*, 94-108.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 1-27.

Clements, F. (1954). Use of Cluster Analysis with Anthropological Data.

Cole-Rodgers, P., Smith, D., & Bosland, P. (1997).

Cox, D. R. (1957). Note on grouping. Journal of the American Statistical.

D. J. Rogers, T. T. (1960). A computer program for classifying.

Dash, R., & Misra, B. B. (2017). Performance analysis of clustering techniques over microarray data: A case study. India.

Davies, D., & Bouldin, D. w. (1979). A cluster separation measure. *Transactions on Pattern Analysis and Machine Intelligence*, 224-227.

Deza, E., & Deza, M. (2006). Dictionary of Distances.

Dice, L. (1945). Measures of the amount of ecologic association between species. *Ecology*.

Dudo, R., Hart, P.E., Storg, & D.G. (2001). Pattern Classification.

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

Edwards, A. W., & Cavalli-Sforza, L. (1965). A method for cluster analysis. *Biometrika*, 362-375.

Erişoğlu, M. (2011). Uzaklık Ölçülerinin Kümeleme Analizine Olan Etkilerinin İncelenmesi ve Geliştirilmesi. Adana.

Fatma Haouas, Z. B. (2017). A new efficient fuzzy cluster validity index: Application to images clustering.

Finch, H., & Huyn, H. (2000).

Fisher, W. (1958). On grouping for maximum homogeneity. Journal of the.

Florek, K., Lukaszewicz, J., Steinhaus, H., & Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, 282-285.

Fowlkes, E., & Mallows, C. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, s. 553-584.

Friedman, H. P., & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 1159–1178.

Gordon, A. (1999). Classification. 2nd edn. NewYork, NY.

Gower, J. (1967). A comparison of some methods of cluster analysis. *Biometrics*, 623-637.

Gower, J. (1971). A general coefficient of similarity and some of its properties. s. 857-871.

Green, P., & Rao, V. (1969). A Note on Proximity Mesaures and Cluster.

Guilford, J. (1936). Psychometric Methods.

Gunjaca, J., Satavic, Z., & Kolak, I. (2000). Combining qualitative and quantitative TR4IT data in classification of gene bank accessions. *22 nd Int. Conf. information Technology Interfaces ITI-2000*, (s. 311-315).

Halkidi, M., & Vazirgiannis, M. (2001). Clustering validity assessment: Findingthe optimal partitioning of a data set. *Proceedings IEEE InternationalConference on Data Mining*, 187–194.

Hartigan, J. A. (1975). Clustering algorithms. *New York: Wiley*.

Hartigan, J., & Wong, M. (1979). A k-means clustering algorithm. *Applied Statistics*, 100-108.

Hongyan Cui, K. Z. (2017). A Clustering Validity Index Based on Pairing Frequency.

Hubert, L., & Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychologie*, 190-241.

Işık, M., & Çamurcu, A. (2009). Kümelemede Benzerlik ve Uzaklık Ölçütleri Başarılarının Karşılaştırılması. *Marmara Üniversitesi Fen Bilimleri Enstitüsü Dergisi* .

Jaccard, P. (1901). Étude comparative de la distribuition florale dans une portion. *Bull Soc Vandoise Sci Nat.*, 547-579.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci*, s. 223–270.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice Hall.

Jie, Y., Jaume, A., Nicu, S., & Qi, T. (2006). A New Study On Distance Metrics As Similarity Measurement. *IEEE International Conference on Multimedia and Expo(ICME)*, (s. 533-536).

Johnson, R., & Wichern, D. (2014). Applied Multivariate Statistical Analysis. Sixth Edition. *Pearson Prentice Hall*. New Jersey.

José María Luna-Romera, J. G.-G.-B. (2018). An approach to validity indices for clustering techniques in Big Data. *Progress in Artificial Intelligence* , 81-94.

Khan, A., Waqas, M., Ali, M. R., Altalhi, A., Alshomrani, S., & Shim, S.-O. (2015). Imagede-noisingusingnoiseratioestimation,K-meansclusteringandnon-localmeans-basedestimator.

Kim, B. (2018). Integrating cluster validity indices based on data envelopment analysis. *Applied Soft Computing*, s. 94-108.

Krause, E. (1986). Taxicab Geometry An Adventure in Non-Euclidean Geometry. *Courier Dover Publications*.

Kuang Yu Huang, S.-B. C.-D. (2017). The Approach to Classifying Multi-Output Datasets based on Cluster Validity Index Method. *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, 552-556.

Kulczynski, S. (1927). Classedes Sciences Mathmatiques et Naturelles. *Bulletin International de l'Acadamie Polonaise des Sciences et des Lettres Serie B (Sciences Naturelles) (Supplement II)*, 57-203.

L. A. Russos, E. S. (1998). Using New Proximity Measures With Hierarchical Cluster Analysis to Detect. *Journal of Educational Measurement*, 1-30.

L. J. Croabach, G. C. (1953). Assessing Similarity Between.

Lance, G., & Williams, W. (1967). A general theory of classificatory sorting strategies.

Legendre, P., & Gower, J. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of classification. New York: Springer.*, 5-48.

Lloyd, S. (1957). Least square quantization in PCM. Bell Telephone Laboratories Paper. *IEEE Transactions on Information Theory*, s. 129-137.

M. Arif Wani, R. R. (2017). A novel point density based validity index for clustering gene expression datasets. *International Journal of Data Mining and Bioinformatics* , 66-84.

Mac Queen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (s. 281-297). California: L. Cam and J. Neyman, Berkeley, CA: University of California Pres.

Marriot., F. H. (1971). Practical problems in a method of cluster analysis. *Biometrics*, 456-460.

McClain, J. O., & Rao, V. R. (1975). Clustisz: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research*, 456-460.

McNemar, Q. (1947). Note on the sapling error of difference between correlate proportions or percentages. *Psychometrika*, 153-157.

Mimmack, G., Mason, S., & Galpin, J. (2001). Choice of Distance Matrices in Cluster Analysis: Defining Regions. Journal of Climate.

Monev, V. (2004). Introduction to Similarity Searching in Chemistry.

Munendra Singh, R. B. (2017). An improved xie-beni index for cluster validity measure. *2017 Fourth International Conference on Image Information Processing* , 95-99.

Neyman , J. (1949). Contributions to the theory of the χ^2 test. In.

Nicolas Durand, S. D. (2007). Ontology-Based Object Recognition for Remote Sensing Image Interpretation. *Tools with Artificial Intelligence.* Patras.

Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan andits neighbouring regions. *Bulletin of the Japanese Society for Sci Fish*, 526-530.

Pakhira, M. K., Bandyopadhyay, S., & Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 487-501.

Pearson, K. (1900). On the Criterion that a given system of deviations from the.

Peeters, J., & Martinelli, J. (1989). Hierarchical Cluster Analysis As A Tool To Manage Variation in Germplasm Collections. *Theory Appl. Genet*, 42-48.

Phen-Lan Lin, P.-W. H.-H. (2017). Validity index for Gaussian-distributed clusters of different sizes with various degrees of dispersion and overlap. *Proceedings of 2017 International Conference on Machine Learning and Cybernetics* , 455-460.

Pun, W., & Ali, A. S. (2007). Unique Distance Measure Approach for Kmeans (UDMA-Km) Clustering Algorithm. *2007 IEEE Region 10 Conference*, (s. 1-4). Tencon.

Ramon-Gonen, R., & Gelbard, R. (2017). Cluster evolution analysis: Identification and detection of similar clusters and migration patterns.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 846–850.

Ratkowsky, D. A., & Lance, G. N. (1978). A criterion for determining the numberof groups in a classification. *Australian Computer Journal*, 115–117.

Ray, S., & Turi, R. H. (1999). Determination of number of clusters in k-meansclustering and application in colour image segmentation. *in Proceedings ofthe 4th International Conference on Advances in Pattern Recognition andDigital Techniques*, 137–143.

Rencher, A. (2002). Methods of Multivariate Analysis. J. W. Publication. içinde

Rogers, J., & Tanimoto, T. (1960). A computer program for classifying plants. *Scince*, 1115-1118.

Rohlf, F. J. (1974). Methods of comparing classifications. *Annual Review of Ecologyand Systematics*, 101-113.

Russel, P., & Rao, T. (1940). On habitat and association of species of anopheline larvae in south-eastern Madras. *Journal of Malaria India*, 153-178.

Sadia Nawrin, M. R. (2017). Exploreing K-Means with Internal Validity Indexes. *International Journal of Advanced Computer Science and Applications* , 264-272.

Scott, A. J., & Symons, M. J. (1971). clustering methods based on likelihoodratio criteria. *Biometrics*, 387–397.

Sendín-Hernández, M. P., Ávila-Zarza, ,. C., Sanz, C., García-Sánchez, A., Marcos-Vadillo, ,. E., Muñoz-Bellido, ,. F., . . . Dávila, I. (2017). Cluster Analysis Identifies 3 Phenotypes within.

Sneath , P., & Sokal, R. (1973). Numerical Taxonomy: The Principles and Practice of Numerical Classification. *W.H. Freeman and Company*. San Francisco.

Sokal, R., & Sneath, P. (1963). Principles of Numerical Taxonomy.

Soo-Hyun Lee, Y.-S. J.-Y. (2018). A new clustering validity index for arbitrary shape of clusters. *Pattern Recognition Letters*, 263-269.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, 1-34.

Starczewski, A. (2017). A new validity index for crisp clusters. *Pattern Analysis and Applications* , 687-700.

Steinhaus, H. (1955). Sur la division des corp materiels en parties. *Bulletin of acad. polon. sci., IV(C1. III)*, 801–804.

Ting-Cheng Chang, C.-J. J. (2018). Cluster Validity Indexes to Uncertain Data for Multi-Attribute Decision-Making Datasets. *Journal of Internet Technology* , 533-538.

Tryon, R. (1939). Cluster Analysis. A. A. Edwards Brothers. içinde

Uribe, C., Segura, B., Baggio, H. C., Abos, A., Garcia-Diaz, A. İ., Campabadal, A., . . . Junque, C. (2018). Cortical atrophy patterns in early Parkinson's disease patients using hierarchical cluster analysis. *Parkinsonism and Related Disorders*.

Vijay Kuma, J. K. (2017). Performance Evaluation of Line Symmetry-Based Validity Indices on Clustering Algorithms. *Journal of Intelligent Systems*, 483-503.

Vimal, A., Valluri, S., & Karlapalem, K. (2008).

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 234-244.

Xie, H., Tian, G., Chen, H., Wang, J., & Huang, Y. (2017). A distribution density-based methodology for driving data cluster analysis: a case study for an extended-range electric city bus. China.

Xie, X., & Beni., G. (1991). A validity measure for fuzzy clustering. *Transactions on Pattern Analysis and Machine Intelligence*, 841-846.

Xiu Li, Q. T., & Yuan, B. (2017). Efficient distributed clustering using boundary information. China.

Xu, R., & Wunsch , D. (2009). Clustering. J. W. Sons. içinde New Jersey.

Yule, G. (1900). On the Association of Attributes in Statistics: With Illustrats from the Material of the Childhood Society, Philosophical Transactions of the Royal Society of London. Series A. *Containing Papers of a Mathematical or Physical Character*, 257-319.

Zezula, P., Amato, G., Dohnal, V., & Batko, M. (2006). Similarity Search The.

8. ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	**:**	Derya Alkın
Uyruğu	**:**	T.C.
Doğum Yeri ve Tarihi	**:**	Akşehir/23.06.1986
Telefon	**:**	0 554 761 34 88
Faks	**:**	
e-mail	**:**	derkonizm@gmail.com

EĞİTİM

Derece	**Adı, İlçe, İl**	**Bitirme Yılı**
Lise :	**Eşrefpaşa Lisesi, Karabağlar, İZMİR**	**2004**
Üniversite :	**Selçuk Üniversitesi, Selçuklu, KONYA**	**2009**
Yüksek Lisans :	**Necmettin Erbakan Üniversitesi, Meram, KONYA**	
Doktora :		

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2018	**Ticaret Bakanlığı**	**Memur**

UZMANLIK ALANI

YABANCI DİLLER İngilizce

BELİRTMEK İSTEĞİNİZ DİĞER ÖZELLİKLER

YAYINLAR

Uluslararası bilimsel toplantılarda sunulan ve bildiri kitaplarında (proceedings) basılan bildiriler

1- Alkın Derya,KARAKOCA AYDIN (2018). A Comparative Study of Internal Validity Indices for Correlated Data with k-means Algorithm. ISDC2018 11. INTERNATIONAL STATISTICS DAYS CONFERENCE, 63-63. (Özet Bildiri/Sözlü Sunum)(Yayın No:4811311)