KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

YÖNSEL VERİLERİN KÜMELENMESİNDE BULANIK C-ORTALAMALAR ALGORİTMASI

YÜKSEK LİSANS TEZİ

Özge TEZEL

ARALIK 2014
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK VE BİLGİSAYAR BİLİMLERİ ANABİLİM DALI

YÖNSEL VERİLERİN KÜMELENMESİNDE BULANIK C-ORTALAMALAR ALGORİTMASI

Özge TEZEL

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"YÜKSEK LİSANS (İSTATİSTİK)"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih : 05.12.2014
Tezin Savunma Tarihi : 26.12.2014

Tez Danışmanı : Yrd. Doç. Dr. Orhan KESEMEN

Trabzon 2014

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü

İstatistik ve Bilgisayar Bilimleri Anabilim Dalında

Özge TEZEL tarafından hazırlanan

YÖNSEL VERİLERİN KÜMELENMESİNDE BULANIK C-ORTALAMALAR ALGORİTMASI

başlıklı bu çalışma, Enstitü Yönetim Kurulunun 09 / 12 / 2014 gün ve 1580 sayılı kararıyla oluşturulan jüri tarafından yapılan sınavda

YÜKSEK LİSANS TEZİ

olarak kabul edilmiştir.

Jüri Üyeleri

Başkan : Prof. Dr. Mualla YALÇINKAYA …...……

Üye : Yrd. Doç. Dr. Orhan KESEMEN …...……

Üye : Yrd. Doç. Dr. Tolga BERBER ……...……

Prof. Dr. Sadettin KORKMAZ

Enstitü Müdürü

ÖNSÖZ

"Yönsel Verilerin Kümelenmesinde Bulanık C-Ortalamalar Algoritması" isimli bu tez Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik ve Bilgisayar Bilimleri Anabilim Dalı, Yüksek Lisans Programı'nda hazırlanmıştır.

Tez çalışma süresinde değerli yardım ve katkılarıyla beni yönlendiren benden yardımlarını, desteğini, sabrını ve bilgisini esirgemeyen değerli danışman hocam Yrd. Doç. Dr. Orhan KESEMEN'e, eğitimimde emeği geçen tüm hocalarıma, tez sürecinde hiçbir yardımdan kaçınmayan arkadaşlarım Buğra Kaan TİRYAKİ'ye, Merve KESİM'e ve Eda ÖZKUL'a teşekkürü bir borç bilirim.

Son olarak, tüm hayatım boyunca maddi ve manevi her zaman beni destekleyen, her adımımda arkamda duran aileme sonsuz teşekkürlerimi sunarım.

Bu tezin, bundan sonraki çalışmalara katkı sağlamasını temenni ederim.

Özge TEZEL

Trabzon 2014

TEZ BEYANNAMESİ

Yüksek Lisans Tezi olarak sunduğum "Yönsel Verilerin Kümelenmesinde Bulanık C-Ortalamalar Algoritması" başlıklı bu çalışmayı baştan sona kadar danışmanım Yrd. Doç. Dr. Orhan KESEMEN'in sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuvarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 05/12/2014

Özge TEZEL

İÇİNDEKİLER

Sayfa No

ÖNSÖZ III

TEZ BEYANNAMESİ IV

İÇİNDEKİLER V

ÖZET VII

SUMMARY VIII

ŞEKİLLER DİZİNİ IX

TABLOLAR DİZİNİ XI

SEMBOLLER DİZİNİ XII

1. GENEL BİLGİLER 1

1.1. Giriş 1

1.2. Gruplandırılmamış Veriler 2

1.3. Gruplandırılmış Veriler 3

1.3.1. Dairesel Histogramlar 3

1.3.2. Doğrusal Histogramlar 4

1.3.3. Rose Diyagramları 5

1.4. Dairesel Verilerin Betimsel İstatistikleri 6

1.4.1. Önbilgiler ve Gösterimler 6

1.4.2. Dairesel Ortalama 8

1.4.3. Dairesel Ortanca 11

1.4.4. Dairesel Mod 11

1.4.5. Dairesel Varyans 12

1.4.6. Dairesel Standart Sapma 13

1.4.7. Dairesel Saçılım 13

1.4.8. Dairesel Standart Hata 14

1.4.9. Yoğunlaşma Parametresi 14

1.5. Trigonometrik Momentler 15

1.5.1. Dairesel Basıklık ve Çarpıklık Katsayısı 16
1.6. Dairesel Dağılımlar 16
1.6.1. Dairesel Düzgün Dağılım 18
1.6.2. Von Mises Dağılımı 18
1.6.3. Sarmal Normal Dağılım 19
1.6.4. Üçgen Dağılım 20
1.7. Kümeleme Analizi 20
1.8. Dağılımdan Bağımsız Kümeleme Yöntemleri 22
1.8.1. K-Ortalamalar Kümeleme Algoritması 22
1.8.2. Bulanık C-Ortalamalar (FCM) Kümeleme Algoritması 25
1.8.3. Bulanık Kümeleme Algoritmalarında Optimal Küme Sayısı 28
1.8.3.1. Bölünme Katsayısı (PC) ve Bölünme Entropisi (CE) 29
1.8.3.2. Xie-Beni İndeksi(XB) 30
1.8.3.3. Fukuyama and Sugeno İndeksi(FS) 30
1.8.3.4. K İndeksi 30
1.8.3.5. KT İndeksi 31
1.9. Dağılıma Bağımlı Kümeleme Yöntemleri 32
1.9.1. En Büyük Beklenti (EM) Kümeleme Algoritması 32
1.9.2. Bulanık C-Yönsel (FCD) Kümeleme Algoritması 34
2. YAPILAN ÇALIŞMALAR 37
2.1. Yönsel Veriler İçin Bulanık C-Ortalamalar Kümeleme Algoritması 37
2.2. Örnek Uygulamalar 39
2.2.1. Örnek Uygulama 1 39
2.2.2. Örnek Uygulama 2 42
2.2.3. Örnek Uygulama 3 44
2.2.4. Örnek Uygulama 4 49
2.2.5. Örnek Uygulama 5 52
2.2.6. Örnek Uygulama 6 54
3. BULGULAR VE SONUÇLAR 56
4. ÖNERİLER 57
5. KAYNAKLAR 58
ÖZGEÇMİŞ

Yüksek Lisans

ÖZET

YÖNSEL VERİLERİN KÜMELENMESİNDE BULANIK C-ORTALAMALAR ALGORİTMASI

Özge TEZEL

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik ve Bilgisayar Bilimleri Anabilim Dalı
Danışman: Yrd. Doç. Dr. Orhan KESEMEN
2014, 60 Sayfa

Kümeleme analizi veri madenciliğinde önemli bir role sahiptir. Kümeleme analizinin amacı bir veri kümesini benzerliklerine ve farklılıklarını göre alt kümelere ayırmaktır. Bu çalışmada bulanık c-ortalamalar kümeleme algoritması yönsel veriler için uyarlanmıştır. Literatürde yönsel verilerin kümelenmesi için birçok yöntem geliştirilmiştir. Ancak yapılan kümeleme işlemlerinde yaklaşık sonuçlar elde edilmektedir. Dolayısıyla bu yaklaşım çok duyarlı problemlerde istenmeyen sonuçların çıkmasına neden olmaktadır. Literatürdeki yöntemlerde kümeleme, trigonometrik fonksiyonlar kullanılarak hesaplanan yaklaşık uzaklıklar ile yapılmaktadır. Bu çalışmada Yönsel Veriler için Bulanık C-Ortalamalar (FCM4DD) algoritması, doğrudan açısal uzaklığı kullanmaktadır. Böylece FCM4DD algoritması ile daha tutarlı sonuçlar elde edilmiştir. FCM4DD algoritması dairesel verilerin yanı sıra N boyutlu yönsel veriler için de kullanılabilen bir kümeleme algoritmasıdır. Bu çalışmada bazı mevcut kümeleme algoritmaları ile FCM4DD algoritması çeşitli sayısal örnekler üzerinde uygulanarak elde edilen sonuçlar karşılaştırılmıştır. Karşılaştırma sonuçları FCM4DD algoritmasının daha tutarlı, daha doğru ve daha hızlı olduğunu göstermiştir.

Anahtar Kelimeler: Kümeleme Algoritması, Yönsel Veriler, Bulanık c-Ortalamalar Algoritması, Yönsel Veriler için Bulanık c-Ortalamalar Algoritması, Açısal Uzaklık.

Master Thesis

SUMMARY

FUZZY C-MEANS CLUSTERING ALGORITHM FOR DIRECTIONAL DATA

Özge TEZEL

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Statistical and Computer Sciences Graduate Program
Supervisor: Assist. Prof. Dr. Orhan KESEMEN
2014, 60 Pages

Cluster analysis has an important role in data mining. The objective of clustering analysis is to partition the data set into subsets using their similarities or dissimilarities. In this study, fuzzy c-means clustering algorithm is adapted for directional data. Several methods have been developed for clustering of directional data in the literature. But, approximate results are obtained in those clustering methods. Therefore, these methods lead to undesirable results for very sensitive problems. In the methods of literature, clustering is performed with approximate distances which are calculated by using trigonometric functions. In this study, fuzzy c-means algorithm for directional data (FCM4DD) uses directly angular distance. Thus more consistent results are obtained with FCM4DD clustering algorithm. FCM4DD algorithm is a clustering algorithm which can be used for N dimensional data as well as circular data. In this study, some existing clustering algorithms and FCM4DD algorithm is applied on various numerical examples and obtained results are compared. The results show that FCM4DD algorithm is more consistent, more accuracy and faster.

Key Words: Clustering algorithms, Directional data, Fuzzy c-means algorithm, Fuzzy c-means for directional data, Angular distance.

ŞEKİLLER DİZİNİ

Sayfa No

Şekil 1. Gruplandırılmamış verilerin şematik gösterimi 2

Şekil 2. Gruplandırılmış verilerin dairesel histogramla gösterimi 3

Şekil 3. Gruplandırılmış verilerin dönel piramit histogramla gösterimi 4

Şekil 4. Gruplandırılmış verilerin doğrusal histogramla gösterimi 5

Şekil 5. Gruplandırılmış verilerin rose diyagramı ile gösterimi 5

Şekil 6. x dairesel verisinin z karmaşık sayısıyla gösterimi 7

Şekil 7. Açısal ölçümleri 60°, 180° ve 300° olan dairesel verilerin birim çember üzerinde gösterimi 9

Şekil 8. K-ortalamalar algoritmasıyla iki kümeye ayrılmış veri 25

Şekil 9. Bulanık c-ortalamalar algoritmasıyla üç kümeye ayrılan veri 28

Şekil 10. KT indeksinin yüzde doğruluk değeri açısından verilen diğer indekslerle karşılaştırılması 32

Şekil 11. Örnek uygulama 1'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 40

Şekil 12. Örnek uygulama 1'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 41

Şekil 13. Örnek uygulama 1'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 41

Şekil 14. Örnek uygulama 2'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 43

Şekil 15. Örnek uygulama 2'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 43

Şekil 16. Örnek uygulama 2'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 44

Şekil 17. 76 kaplumbağanın yönelimlerinin rose diyagramı 45

Şekil 18. 76 kaplumbağanın yönelimlerinin EM algoritması sonucunda elde edilen üyelik değerleri 46

Şekil 19. 76 kaplumbağanın yönelimlerinin FCD algoritması sonucunda elde edilen üyelik değerleri 46

Şekil 20. 76 kaplumbağanın yönelimlerinin FCM4DD algoritması sonucunda elde edilen üyelik değerleri 47

Şekil 21. Örnek uygulama 3'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 48

Şekil 22. Örnek uygulama 3'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 48

Şekil 23. Örnek uygulama 3'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 49

Şekil 24. Örnek uygulama 4'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 50

Şekil 25. Örnek uygulama 4'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 51

Şekil 26. Örnek uygulama 4'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 51

Şekil 27. Örnek uygulama 5'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 53

Şekil 28. Örnek uygulama 5'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 53

Şekil 29. Örnek uygulama 5'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi 54

Şekil 30. Küresel verilerin benzetimlerindeki modellerden bir tanesinin gösterimi (a) Küresel verilerin gösterimi; (b) FCM4DD algoritması ile küresel verilerin kümelere ayrılmasının gösterimi 55

TABLOLAR DİZİNİ

Sayfa No

Tablo 1. Örnek Uygulama 1 için kümeleme yöntemlerinin karşılaştırılması 39

Tablo 2. Örnek uygulama 2 için kümeleme yöntemlerinin karşılaştırılması 42

Tablo 3. 76 kaplumbağanın yumurtladıktan sonraki hareket yönleri 44

Tablo 4. Örnek uygulama 3 için kümeleme yöntemlerinin karşılaştırılması 45

Tablo 5. Örnek uygulama 4 için kümeleme yöntemlerinin karşılaştırılması 50

Tablo 6. Örnek uygulama 5 için kümeleme yöntemlerinin karşılaştırılması 52

Tablo 7. Örnek Uygulama 6'nın sonuçları 55

SEMBOLLER DİZİNİ

α_{rad}	:Radyan cinsinden tanımlanan açı
α_{deg}	:Derece cinsinden tanımlanan açı
$\bar{R}$	:Ortalama Bileşke Vektör Uzunluğu
$\bar{\theta}$	:Dairesel Ortalama
$\tilde{\theta}$	:Dairesel Ortanca
$\breve{\theta}$	:Dairesel Mod
V	:Dairesel Varyans
s	:Dairesel Standart Sapma
ϑ	:Dairesel Saçılım
σ	:Dairesel Standart Hata
κ	:Yoğunlaşma Parametresi
m_p	:p. Trigonometrik Moment
$\hat{s}$	:Dairesel Çarpıklık Katsayısı
$\hat{k}$	:Dairesel Basıklık Katsayısı
μ	:Üyelik Değerleri
EM	:En Büyük Beklenti Algoritması
FCD	:Bulanık c-Yönsel Kümeleme Algoritması
FCM	:Bulanık c-Ortalamalar Kümeleme Algoritması
FCM4DD	:Yönsel Veriler için Bulanık c-Ortalamalar Kümeleme Algoritması
$X \sim U(0,1)$	:Düzgün dağılımdan [0,1) aralığında rastgele bir sayı
$X \sim N(0,1)$	:Normal dağılımdan 0 ortalamalı 1 standart sapmalı bir rastgele sayı

1. GENEL BİLGİLER

1.1. Giriş

Rastgele örneklenmiş verilerin istatistiksel analizinde, verilerin bir rastgele değişkenden geldiği kabul edilir. Bu rastgele değişken değişik ölçü uzaylarında bulunabileceği gibi açısal bir uzayda da bulunabilmektedir. Tek değişkenli açısal değişim gösteren veriler, dairesel veriler olarak isimlendirilmektedir. Rüzgârların yönleri, kuşların veya diğer hayvanların göç yönleri (Chang-Chien, Yang, & Hung, 2010), salgın hastalıkların bir bölgede yayılım yönleri, cisimlerin düzlemdeki yönelimleri dairesel verilere örnek olarak verilebilir. Bu gözlemlerin elde edilmesinde kullanılan iki temel dairesel ölçüm aracı pusula ve saattir. Pusula kullanılarak yapılabilecek gözlemlere örnek olarak göçmen kuşların göç esnasındaki yönelimleri gösterilebilir. Saatle yapılabilecek gözlemlere örnek olarak da, bir hastanedeki acil servis birimine gelen hastaların 24 saat içerisindeki servise geliş zamanlarının dağılımı verilebilir (Mardia & Jupp, 2000). Bu türdeki veriler ay ya da yıl cinsinden de elde edilebilir. Dairesel bir gözlem, birim yarıçaplı bir daire üzerinde bir nokta ya da düzlemde bir birim vektör olarak kabul edilebilir. İlk olarak dairenin bir başlangıç yönü ve yönelimi seçilir, her dairesel gözlem çember üzerindeki noktanın başlangıç yönünden açısı ile belirtilebilir. Öte yandan yönelim içermeyen ancak periyodik bir süreçte gerçekleşen verilerde aynı sınıfta incelenebilir. Periyodik veriler belli zaman dilimlerinde aynı karakteristiği gösteren verilerdir. Bunlara örnek olarak, bir öğrencinin haftalık çalışma programı, bir canlının tükettiği günlük su miktarının yıllık bazda değişimi gösterilebilir. Sıklıkları periyodik olarak değişen veriler genel anlamda dairesel bir veri olmadığı halde dairesel verilere dönüştürülebilir. Bunun için öncelikle periyodik zaman aralığı birim çember üzerine göreceli olarak yerleştirilerek dairesel veri konumları elde edilebilir. Açısal değişimli veriler iki değişkenli olursa (θ_i, φ_i) küresel, ikiden fazla olursa hiperküresel veriler olarak isimlendirilmektedir (Fisher, 1993). Açı tabanlı veriler ise genel olarak yönsel veriler olarak isimlendirilmektedir.

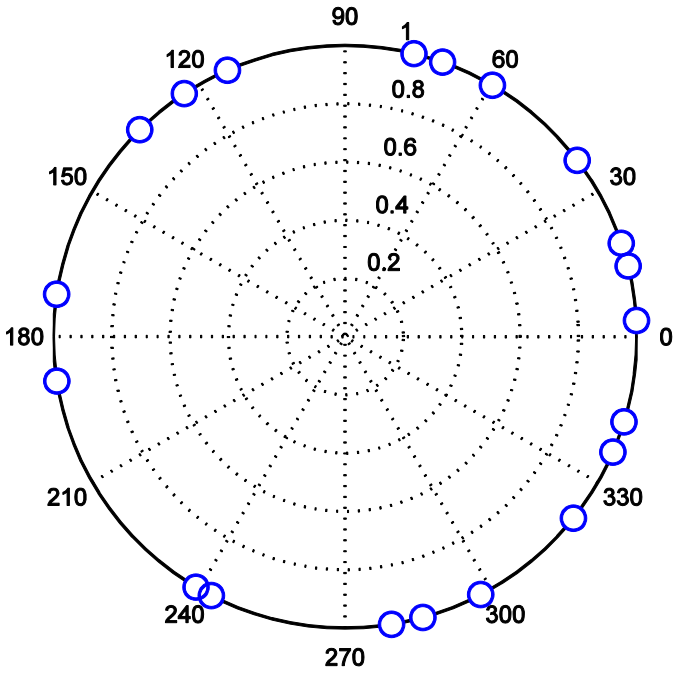
Dairesel veriler genellikle derece olarak ölçülür. Bununla birlikte, dairesel verileri radyan cinsinden ölçmek bazen yararlı olabilmektedir. Açısal ölçümler dereceden radyana $\frac{\pi}{180}$ ile çarpılarak dönüştürülür. Dairesel veriler eksenel verilerle yakından alakalıdır ve genellikle çember üzerinde gözlemler olarak verilirler. Bu gözlemlerin her birinin yönü

ters yöne eşdeğer olarak kabul edilir. Böylece θ ile $180 + \theta$ eşdeğerdir. Eksenel verileri kullanmanın standart yolu verileri ikiye katlayarak onları dairesel verilere dönüştürmektir (Mardia & Jupp, 2000).

Dairesel verilerin dağılımı ilk kez 1918'de Von Mises tarafından incelenmiştir (Von Mises, 1918). Daha sonra 1956'da Watson ve Wiliams tarafından dairesel verilerden istatistiksel sonuç çıkarımı üzerine çalışmalar yapılmıştır (Watson & Williams, 1956). Bu çalışmalardan sonra birçok araştırmacının bu alana olan ilgisi arttı. Dairesel verilerin istatistik uygulamaları yer bilimleri, meteoroloji, biyoloji, fizik, psikoloji, görüntü çözümleme, tıp, astronomi gibi alanlarda kullanılmıştır (Mardia & Jupp, 2000) (Fisher, 1993).

1.2. Gruplandırılmamış Veriler

Dairesel verilerin en basit gösterimi dairesel ham verilerin birim çember üzerinde bir nokta olarak verilmiş olduğu gösterimdir.

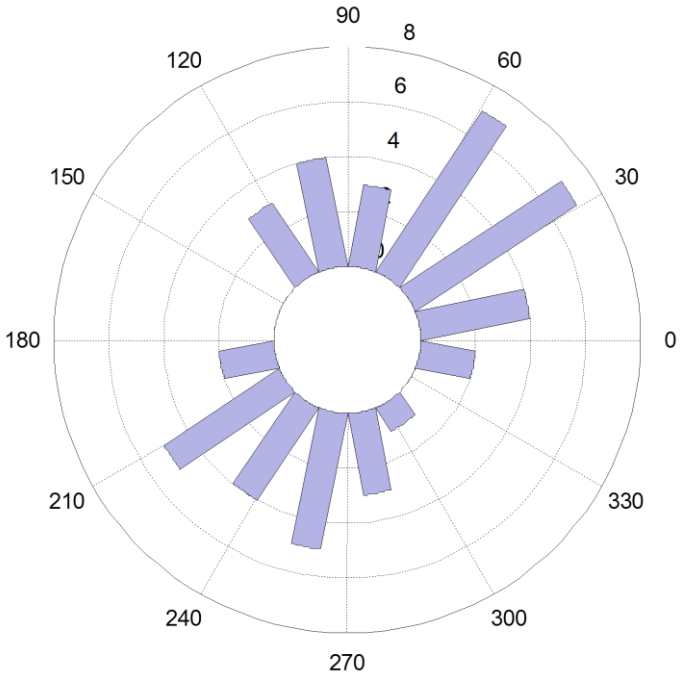


Şekil 1. Gruplandırılmamış verilerin şematik gösterimi

1.3. Gruplandırılmış Veriler

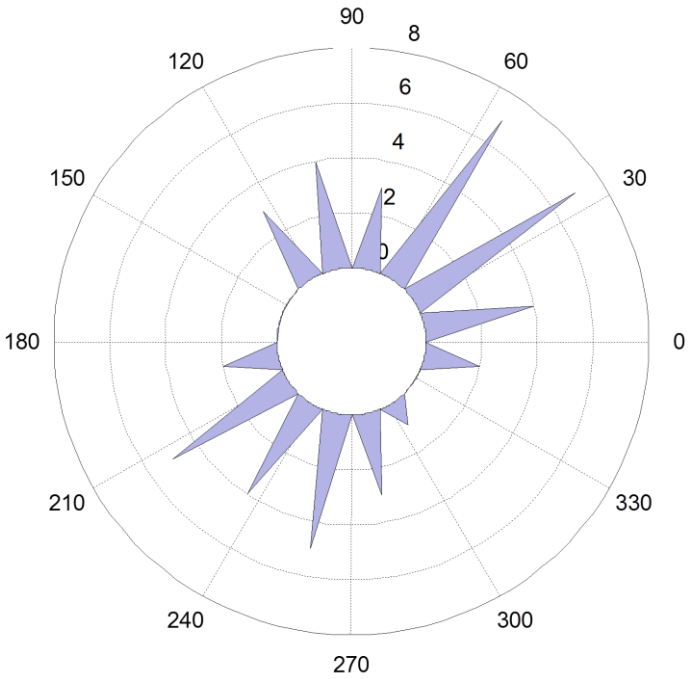
1.3.1. Dairesel Histogramlar

Gruplandırılmış veriler eksen üzerindeki histogramlara benzer olarak dairesel histogramlar ile temsil edilebilirler. Dairesel histogramlardaki her bir çubuk, açı grubuna karşılık gelen noktayı ortalar ve çubuğun alanı açı grubunun frekansı ile orantılıdır.



Şekil 2. Gruplandırılmış verilerin dairesel histogramla gösterimi

Gruplandırılmış verilerin dairesel histogram dışında diğer bir gösterimi ise dönel piramit histogram kullanılarak gösterimidir.

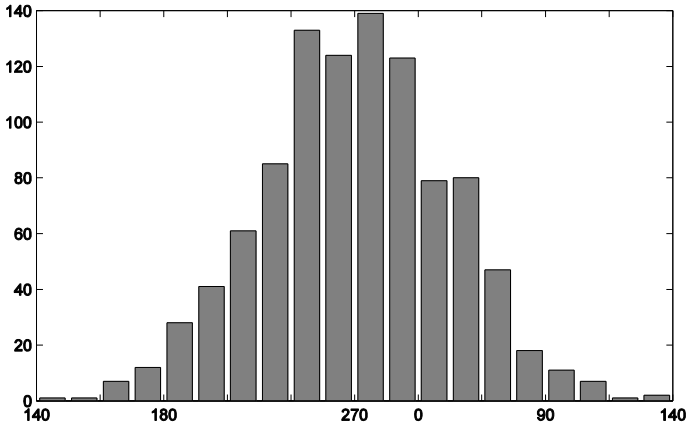


Şekil 3. Gruplandırılmış verilerin dönel piramit histogramla gösterimi

1.3.2. Doğrusal Histogramlar

Veri analizi yapılırken daha çok doğrusal histogramlar yorumlandığından, dairesel verileri yorumlarken dairesel histogramı doğrusal histograma dönüştürmek faydalı olabilmektedir. Bu dönüşüm, daire üzerinde uygun biçimde seçilen noktadan dairesel histogramın kesilmesi ve daha sonra dairesel histogramın 360° genişliğinde açılması ile yapılır.

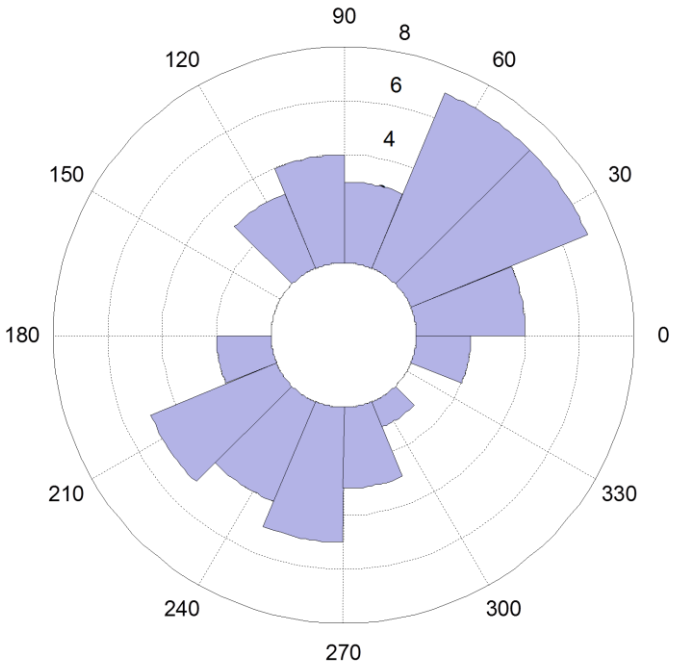
Doğrusal bir histogram olarak verilen görsel gösterim, dairenin kesildiği noktaya karşı duyarlı olabilir. Verilerin tek bir tepe değeri (tercih yönü) varsa o zaman hemen bu tepe değerinin karşı noktasında bir kesim kullanmak akıllıca olacaktır. Böylece doğrusal histogramın merkezi, tepe değerine yakın olacaktır. Kesim tepe değerine yakınsa verinin iki tepe değeri olduğuna dair yanıltıcı gösterim verecektir.



Şekil 4. Gruplandırılmış verilerin doğrusal histogramla gösterimi

1.3.3. Rose Diyagramları

Dairesel histogramın bir diğer kullanışlı biçimi de dairesel histogramdaki çubukların yerini daire dilimlerinin aldığı rose diyagramıdır. Her daire diliminin alanı karşılık geldiği grubun frekansı ile orantılıdır. Gruplar eşit genişlikte ise her bir daire diliminin yarıçapı ilgili frekansın karekökü ile orantılı olmalıdır.



Şekil 5. Gruplandırılmış verilerin rose diyagramı ile gösterimi

1.4. Dairesel Verilerin Betimsel İstatistikleri

Verilerin şematik gösterimi verildikten sonra, uygun tanımlayıcı istatistikler ile verileri özetlemek araştırmacı açısından faydalı olacaktır.

Dairesel veriler, uygun biçimde seçilmesi gereken bir sıfır yönü ve dönüş doğrultusuna göre ölçülen açılar biçiminde gösterilebilirler. Burada, sıfır yönü başlangıç noktasını, dönüş doğrultusu da pozitif yön olarak saat yönünün mü yoksa saat yönünün tersinin mi kullanılacağını belirtmektedir. Yön kavramında herhangi bir büyüklük söz konusu olmadığından, açısal bir gözlem değeri, merkezi orijin olan bir birim çemberin çevresi üzerinde noktalarla ya da orijini bu noktalarla birleştiren birim vektörlerle gösterilebilir. Buna göre derece cinsinden ölçülmüş tek bir gözlem olan $\theta°$ $(0° < \theta° < 360°)$, birim vektör olacaktır. Burada $\theta°$; vektör ile pozitif x-ekseninin, saat yönünün tersi yönünde yaptığı açıyı gösterir (Peker & Bacanlı, 2004).

İki boyutlu bir yönün açı ya da vektör şeklindeki birim çember üzerindeki gösterimi tek bir tane değildir. Çünkü dairesel gözlemin değeri, sıfır yönüne ve dönüş doğrultusunun saat yönünde olup olmaması seçimine göre değişebilmektedir. Elde edilen sonuçlar verilen gözlem değerlerinin bir fonksiyonudur ve fonksiyona verilen keyfi değerlere bağlı değildir. Bu özelliklerinden dolayı dairesel veri analizi, bilinen istatistiksel analizden oldukça farklıdır. Keyfi sıfır yönü ve dönüş doğrultusuna duyulan ihtiyaç, bilinen birçok istatistiksel tekniği ve ölçüleri tamamen olmasa da anlamsız ve çoğu zaman hatalı kılar (Peker & Bacanlı, 2004).

Verilerimizin üzerinde bulunduğu daire, uygun bir noktadan kesilip ve bu hat üzerinde çıkan gözlemler kullanılarak bu verilere geleneksel betimsel istatistikler uygulanabilir. Bu yaklaşımın dezavantajı, elde edilen betimsel istatistiklerin dairenin kesildiği noktaya bağlı olmasıdır. Bunu anlamak için çember üzerinde $\{1^{\circ}, 359^{\circ}\}$ açılarından oluşan iki boyutlu örnek düşünelim. Daireyi 0°den kesmek örnek ortalamasını 180° ve standart sapmasını 179° yapacaktır. Oysaki daireyi 180°den kesmek $(180^{\circ} = -180^{\circ})$ örnek ortalamasını 0° ve standart sapmasını 1° yapacaktır.

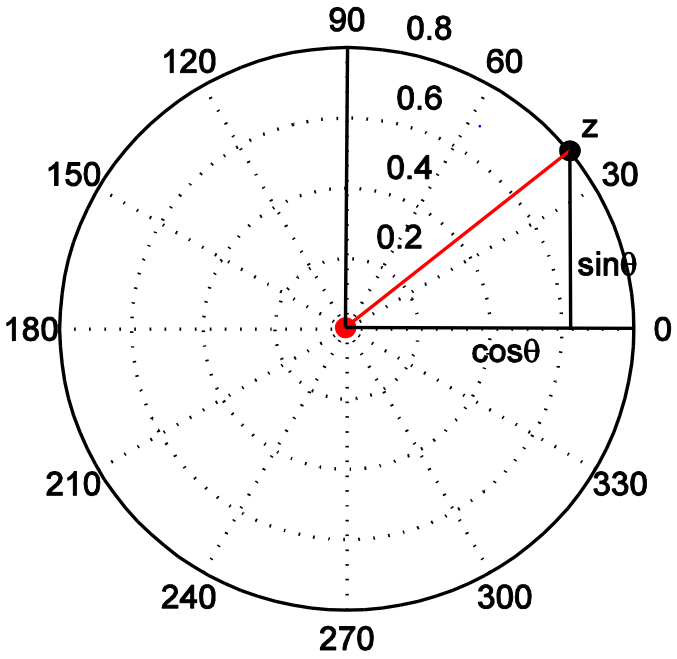
1.4.1. Önbilgiler ve Gösterimler

Düzlemdeki dairesel veriler x birim vektörü ya da buna eşdeğer olarak birim çember üzerindeki noktalar olarak kabul edilebilirler. (Örneğin; merkezi orijin olan birim yarıçaplı

çember) Dairesel verileri ifade etmenin açısal ve birim karmaşık sayı olmak üzere iki kullanışlı yöntemi vardır. Her iki yöntem içinde öncelikle birim çember için başlangıç yönü ve doğrultusu seçilir. Bu düzlemde koordinat sisteminde bir dik seçmekle eşdeğerdir. Daha sonra daire üzerinde her bir x noktası θ açısı veya buna eşdeğer z birim karmaşık sayısı ile temsil edilebilir.

$$x = (\cos\theta\,, \sin\theta)^T \tag{1}$$

$$z = e^{i\theta} = \cos\theta + i\sin\theta \tag{2}$$



Şekil 6. x dairesel verisinin z karmaşık sayısıyla gösterimi
$z = e^{i\theta} = \cos\theta + i\sin\theta$

Dairesel verilere ilişkin ölçümler derece cinsinden yapılır. Fakat bazı durumlarda derecelerin radyana dönüştürülmesi gerekebilir. Derece cinsinden verilmiş bir açısal ölçümü radyana çevirmek için (3) eşitliğinde verilen formül kullanılmaktadır.

$$\alpha_{rad} = \alpha_{deg} * \left(\frac{\pi}{180}\right) \tag{3}$$

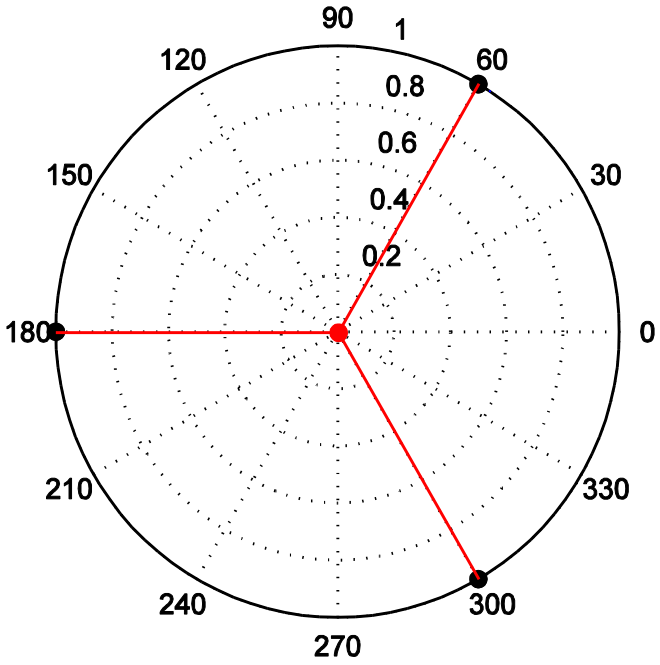
Radyan cinsinden verilmiş bir açısal ölçümü dereceye çevirmek için de (4) eşitliğinde verilen formül kullanılmaktadır.

$$\alpha_{deg} = 360 * \left(\frac{\alpha_{rad}}{2\pi}\right) \quad (4)$$

Teoride dairesel veriler kullanılırken, derece cinsinden ölçülen bir $\theta°$ gözleminin radyan cinsinden θ açısına dönüştürülmesi tercih edilir. Birim çember üzerinde θ açısı ile $\theta + 2\pi$ açısının aynı noktayı gösterdiği unutulmamalıdır. Bu nedenle daire üzerinde yapılacak tüm aritmetik işlemlerde 2π modu kullanılır (Örneğin; $(\theta_1 + \theta_2)\ mod\ 2\pi$ yerine $\theta_1 + \theta_2$ yazılabilir.) (Mardia & Jupp, 2000).

1.4.2. Dairesel Ortalama

Dairesel verilerin ortalaması, doğrusal verilerin ortalama formülüyle hesaplanamamaktadır. Dairesel verilerin ortalamasının hesaplanabilmesi için gözlem değerleri birim vektörler olarak düşünülür ve hesaplama işlemi sırasında bu vektörlerin bileşkelerinin yönü kullanılır. Dairesel veri kümesinin $60°, 180°$ ve $300°$'den oluştuğunu varsayalım. Bu verilerin ortalaması alınırken doğrusal verilerdeki ortalama formülü kullanıldığında ortalamanın $180°$ olduğu açıktır. Oysaki veriler birim çember üzerinde birer vektör olarak gösterildiğinde bunların bileşkesinin $0°$ olduğu görülmektedir.



Şekil 7. Açısal ölçümleri 60°, 180° ve 300° olan dairesel verilerin birim çember üzerinde gösterimi

$i = 1, \dots, n$ olmak üzere, θ_i açılarına karşılık gelen $x_1, \dots, x_n$ birim vektörlerin verildiğini varsayalım. θ_i'lerin dairesel ortalaması $\bar{\theta}$, aynı zamanda x_i'lerin ağırlık merkezlerinin yönü $\bar{x}$ olsun. Bu durumda ağırlık merkezlerinin kartezyen koordinatları $\bar{C}$ ve $\bar{S}$, (5) ve (6) eşitliklerindeki gibi tanımlanır.

$$\bar{C} = \frac{1}{n}\sum_{j=1}^{n} cos\theta_j \tag{5}$$

$$\bar{S} = \frac{1}{n}\sum_{j=1}^{n} sin\theta_j \tag{6}$$

Birim çember üzerinde vektör şeklinde gösterilen dairesel verilerin, ortalama bileşke vektör uzunluğu $(\bar{R})$, (7) eşitliğindeki formül yardımıyla bulunabilir.

$$\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2}, \qquad 0 \leq \bar{R} \leq 1 \tag{7}$$

Ayrıca dairesel ortalama biliniyorsa ortalama bileşke vektör uzunluğu,

$$\bar{R} = \frac{\bar{C}}{\cos\bar{\theta}} = \frac{\bar{S}}{\sin\bar{\theta}} \tag{8}$$

biçiminde de hesaplanabilmektedir. $\bar{R} = 0$ olduğunda $\bar{\theta}$ tanımlı değildir. $\bar{R} > 0$ olduğunda ise $\bar{\theta}$, (9) eşitliğindeki gibi tanımlanır ve dairesel ortalama olarak adlandırılır.

$$\bar{\theta} = \begin{cases} \tan^{-1}(\bar{S}/\bar{C}) & \bar{C} \geq 0 \\ \tan^{-1}(\bar{S}/\bar{C}) + \pi & \bar{C} < 0 \end{cases} \tag{9}$$

Buradaki $\tan^{-1}$ fonksiyonu $[-\pi/2\,, \pi/2]$ arasında değer almaktadır.

Bileşke vektör uzunluğunun sıfır olması verinin çember üzerinde herhangi bir yoğunlaşma göstermediğinin yani düzgün dağılıma sahip olduğunun göstergesidir.

Gruplandırılmış dairesel verilerde ise ortalama hesaplanırken genel bir yaklaşım olarak bir aralıktaki tüm gözlem değerlerinin o aralığın orta noktası olduğu varsayımı kullanılır (Peker & Bacanlı, 2004). Örneğin n sayıda başlangıç gözlem değeri k tane sınıfa göre gruplandığında, i. sınıfın orta noktası θ_i ve sıklığı f_i'dir. Buna bağlı olarak gruplandırılmış dairesel verilerin ortalamasını bulmak için,

$$\bar{C}_{\mathrm{n}} = \frac{1}{n}\sum_{i=1}^{k} f_i \cos\theta_i \tag{10}$$

$$\bar{S}_{\mathrm{n}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{k}} f_i \sin\theta_i \tag{11}$$

$$\bar{R} = \sqrt{\bar{C}_n{}^2 + \bar{S}_n{}^2} \tag{12}$$

değerleri hesaplanır.

1.4.3. Dairesel Ortanca

Birim çember üzerinde noktalar şeklinde verilen dairesel verileri iki eş parçaya ayıran değere dairesel ortanca denir. Eğer dairesel verilerin sayısı (n) çift ise verileri iki eş parçaya ayıran değer en yakın iki veri noktası arasındaki uzaklığın yarısında yer alır (Berens, 2009). Eğer n tek ise ortanca veri noktalarından birine karşılık gelecektir. Ortanca $\tilde{\theta}$ ile gösterilir ve

$$\int_{\tilde{\theta}}^{\tilde{\theta}+\pi} f(\theta)d\theta = \int_{\tilde{\theta}+\pi}^{\tilde{\theta}+2\pi} f(\theta)d\theta = \frac{1}{2} \tag{13}$$

biçiminde tanımlanan integralin çözümü ile bulunur. Ortanca tek bir tane olmayabilir ama bütün tek tepe değerine sahip dağılımlar tek ortancaya sahiptir (Mardia K. V., 1972).

Gruplandırılmış dairesel veriler için ortanca hesaplanırken (14) eşitliği kullanılır.

$$\tilde{\theta} = l + \frac{\left(\frac{\mathrm{n}}{2} - \mathrm{f}_0\right)}{\mathrm{f}_{+1} - \mathrm{f}_0} * \mathrm{h} \tag{14}$$

Burada; l medyan sınıfının alt sınırını, f_0 medyan sınıfının $(\theta_i - 180°, \theta_i)$ aralığındaki sıklığını, f_{+1} medyan sınıfından bir sonraki sınıfın $(\theta_i - 180°, \theta_i)$ aralığındaki sıklığını, h sınıf aralığının uzunluğunu göstermektedir (Mardia K. V., 1972).

1.4.4. Dairesel Mod

Dairesel mod, dairesel veri setinin en fazla yoğunlaştığı yön anlamına gelir. Mod, ortanca gibi en büyük ve en küçük sayıları dikkate almadığı için uç değerlerden etkilenmez. Buna karşılık gözlem sayısının küçük olduğu durumlarda mod değerinin fazla bir açıklayıcılığı yoktur. Daire üzerinde noktalar şeklinde verilen bir dairesel veri setinin modu dairenin kesim noktasının seçiminden sonra bilinen yöntemle hesaplanır.

Gruplandırılmış bir dairesel veri seti için mod hesaplanırken (15) eşitliği kullanılır.

$$\breve{\theta} = l + \frac{f_0 - f_{-1}}{2f_0 - \mathrm{f}_{-1} - \mathrm{f}_{+1}} * \mathrm{h} \tag{15}$$

Burada; l mod sınıfının alt sınırını, f_0 mod sınıfının sıklığını, f_{-1} mod sınıfından bir önceki sınıfın sıklığını, f_{+1} mod sınıfından bir sonraki sınıfın sıklığını, h sınıf aralığının uzunluğunu gösterir (Mardia K. V., 1972).

1.4.5. Dairesel Varyans

Dairesel varyans doğrudan ortalama bileşke vektör uzunluğu ile ilgilidir. Dairesel varyans bir veri kümesinin yayılımını göstermektedir. Eğer dairesel verilerin hepsinde bir yöne doğru yoğunlaşma oluyorsa ortalama bileşke vektör uzunluğu bire yakın olacaktır. Eğer dairesel veriler birim çember üzerinde geniş bir yayılım gösteriyorsa yani düzgün bir dağılım gösteriyorsa ortalama bileşke vektör uzunluğu sıfıra yakın veya eşit olacaktır. Dairesel varyans (Mardia K. V., 1972) de (16) eşitliğindeki gibi tanımlanmıştır.

$$\begin{aligned} V &= 1 - \frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i - \bar{\theta}) \\ &= 1 - \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\cos\theta_i\cos\bar{\theta} + \sin\theta_i\sin\bar{\theta}) \\ &= 1 - \frac{1}{n}(\bar{C}\cos\bar{\theta} + \bar{S}\sin\bar{\theta}) \\ &= 1 - \frac{1}{\mathrm{n}}(\mathrm{R}\cos\bar{\theta}\cos\bar{\theta} + \mathrm{R}\sin\bar{\theta}\sin\bar{\theta}) \\ &= 1 - \frac{1}{\mathrm{n}}\mathrm{R} \end{aligned} \tag{16}$$

Sonuç olarak dairesel varyans kısaca (17) eşitliğinde verilen formül yardımıyla bulunur.

$$V = 1 - \bar{R} \quad , \quad 0 \leq V \leq 1 \tag{17}$$

Eğer ortalama bileşke vektör uzunluğu 1'e yakın olduğunda dairesel varyans değeri 0'a yaklaşır. Eğer ortalama bileşke vektör uzunluğu 0'a yakınsa dairesel varyans değeri 1'e yaklaşır.

1.4.6. Dairesel Standart Sapma

Dairesel verilerinin standart sapmasının hesaplanması için birden çok standart sapma formülü tanımlanmıştır. Standart sapma için (Zar, 1999) tarafından önerilen formüllerden biri,

$$s = \sqrt{2(1-\bar{R})}, \qquad s \in [0, \sqrt{2}] \tag{18}$$

biçimindedir. Bu formüle alternatif olarak geliştirilmiş diğer standart sapma formülü ise (19) eşitliğindeki gibi tanımlanmıştır.

$$s_0 = \sqrt{-2(\ln \bar{R})}, \qquad s_0 \in (0, \infty) \tag{19}$$

Bu iki eşitlikten tanım aralığı sınırlı olduğu için genellikle (18) eşitliği kullanılır. Dairesel varyansın küçük değerler alması durumunda standart sapma, (20) eşitliği kullanılarak hesaplanabilir (Mardia K. V., 1972).

$$s = \sqrt{2V} \tag{20}$$

1.4.7. Dairesel Saçılım

Dairesel saçılım,

$$\vartheta = \frac{1-p_2}{2\bar{R}^2} \tag{21}$$

biçiminde tanımlanır. Eşitlikteki p_2 değeri ikinci trigonometrik momenti göstermektedir (Fisher, 1993) ve ikinci trigonometrik moment,

$$p_2 = \frac{1}{n}\sum_{i=1}^{n} \cos 2(\theta_i - \bar{\theta}) \tag{22}$$

biçiminde tanımlanır. Dairesel saçılım daha çok dairesel ortalamanın güven aralığının hesaplanmasında kullanılır (Peker & Bacanlı, 2004).

1.4.8. Dairesel Standart Hata

Veri analizinde standart hata örnek ortalamalarının standart sapmasıdır. Dairesel veri analizinde standart hata,

$$\sigma = \sqrt{\frac{\vartheta}{n}} \tag{23}$$

biçiminde tanımlanır. Dairesel standart hata, dairesel ortalama için güven aralıklarının belirlenmesinde kullanılır (Mardia & Jupp, 2000).

1.4.9. Yoğunlaşma Parametresi

Olasılık teorisi ve istatistikte yoğunlaşma parametresi olasılık dağılımlarının parametrik ailesinin sayısal parametrelerinin özel bir türüdür. Yoğunlaşma parametresi en çok kullanıldığı dağılım von Mises dağılımıdır. Yoğunlaşma parametresinin büyük değer alması dağılımın düzgün dağılıma daha çok eğilimi olduğunu gösterir. Yoğunlaşma parametresinin küçük değer alması ise dağılımın sadece bir nokta etrafında yoğunlaştığını göstermektedir.

Yoğunlaşma parametresi κ'nın en çok olabilirlik tahmini $\hat{\kappa}$;

$$A_1(\hat{\kappa}) = \frac{\mathrm{R}}{\mathrm{n}} = \bar{R} \tag{24}$$

biçimindedir. Burada $\bar{R}$;

$$A_1(x) = I_1(x)/I_0(x) \tag{25}$$

biçiminde dönüştürülmüş iki Bessel fonksiyonunun oranıdır. κ'nın en çok olabilirlik tahmin edicisi olan $\hat{\kappa}$'nın çözümü,

$$\hat{\kappa} = \begin{cases} 2\bar{R} + \bar{R}^3 + \dfrac{5\bar{R}^3}{6}, & \bar{R} < 0.53 \\ -0.4 + 1{,}39\bar{R} + 0.43(1 - \bar{R}), & 0.53 \leq \bar{R} < 0.85 \\ \dfrac{1}{\bar{R}^3 - 4\bar{R}^2 + 3\bar{R}}, & \bar{R} \geq 0.85 \end{cases} \tag{26}$$

biçiminde verilmektedir (Fisher, 1993).

1.5. Trigonometrik Momentler

Verilerin trigonometrik momentleri genellikle karmaşık sayılar şeklinde ifade edilirler. Dairesel verilerin sıfır yönü etrafındaki birinci trigonometrik momenti,

$$m_1' = \bar{C} + i\bar{S} = \bar{R}e^{i\bar{\theta}} \tag{27}$$

biçiminde tanımlanır. Eşitlik (27)'deki formülasyonu genişletecek olursak dairesel verilerin sıfır yönü etrafındaki p-inci momenti,

$$m_p' = a_p + ib_p = \bar{R}_{\mathrm{p}}e^{i\bar{\theta}_p} \tag{28}$$

biçiminde ve (28) eşitliğindeki a_p ve b_p terimleri,

$$a_p = \frac{1}{n}\sum_{j=1}^{n} cos\, p\theta_j\,, \qquad b_p = \frac{1}{n}\sum_{j=1}^{n} sin\, p\theta_j \tag{29}$$

şeklinde tanımlanır. Eşitlikteki $\bar{R}_p$ ve $\bar{\theta}_p$ ifadeleri $\{p\theta_1, p\theta_2, \dots, p\theta_n\}$ dairesel verileri için dairesel ortalama ve ortalama bileşke vektör uzunluğu anlamına gelmektedir. Ortalama etrafındaki p-inci trigonometrik moment,

$$m_p = \bar{a}_{\mathrm{p}} + i\bar{b}_p \tag{30}$$

biçiminde ve (30) eşitliğindeki $\bar{a}_p$ ve $\bar{b}_p$ terimleri,

$$\bar{a}_p = \frac{1}{n}\sum_{j=1}^{n} \cos p(\theta_j - \bar{\theta})\,, \qquad \bar{b}_p = \frac{1}{n}\sum_{j=1}^{n} \sin p(\theta_j - \bar{\theta}) \tag{31}$$

şeklinde tanımlanır. Bu durumda ortalama etrafındaki birinci moment $m_1 = \bar{R}$'ye eşit olur. Dairesel verilerde trigonometrik momentler dairesel dağılımlar teorisinde önemli bir rol oynamaktadır.

1.5.1. Dairesel Basıklık ve Çarpıklık Katsayısı

Dairesel verilerin simetri ölçüsü olan çarpıklık katsayısı ($\hat{s}$) ve verilerin tepe noktasının ölçüsü olan basıklık katsayısı ($\hat{k}$) (Mardia & Jupp, 2000) tarafından,

$$\hat{s} = \frac{\bar{R}_2\, sin(\bar{\theta}_2 - 2\,\bar{\theta})}{(1-\bar{R})^{\frac{3}{2}}} \tag{32}$$

$$\hat{k} = \frac{\bar{R}_2\, cos(\bar{\theta}_2 - 2\,\bar{\theta}) - \bar{R}_4}{(1-\bar{R})^2} \tag{33}$$

biçiminde tanımlanmıştır.

1.6. Dairesel Dağılımlar

Olasılık dağılımları istatistiksel veri analizinde önemli bir yer tutar. Doğrusal verilerdeki dağılımlara karşılık olarak dairesel veriler için de çeşitli dağılımlar vardır. Dairesel dağılım, bütün olasılıkları birim çemberin çevresinde yoğunlaşan bir olasılık dağılımıdır (Jammalamadaka & Gupta, 2001). Bu dağılımları tanımlamadan önce dairesel yoğunluk kavramı verilmelidir. Dairesel dağılımlar genellikle dairesel bir yoğunluk olarak

tanımlanırlar. Dairesel bir olasılık yoğunluk fonksiyonu aşağıdaki temel özelliklere sahiptir.

1. $\forall\theta$ için $f(\theta) \geq 0$'dır.
2. f fonksiyonu 2π periyoduna göre peryodiktir. Yani $f(\theta) = f(\theta + k.2\pi)$'dir.
3. $\forall\theta_0$ için $\int_{\theta_0}^{\theta_0+2\pi} f(\theta)d\theta = 1$ olmalıdır.

Birim çember üzerinde verilerin dağılımını belirlemenin bir yolu dağılım fonksiyonu yardımıyla olur. Dairesel dağılımlar kesikli veya mutlak sürekli olmak üzere ikiye ayrılırlar. Dairesel dağılımlarda θ dairesel rastgele değişkenleri radyan cinsinden ölçülür. Bu rastgele değişkenler $[0,2\pi)$ veya $[-\pi,\pi)$ aralığında değer alır. İlk olarak bir başlangıç yönü ve yönelimi seçildiği varsayılır. Daha sonra rastgele θ açısının dağılım fonksiyonu F,

$$F(x) = \Pr(0 < \theta < x), \qquad 0 \leq x \leq 2\pi \tag{34}$$

$$F(x + 2\pi) - F(x) = 1, \qquad -\infty < x < \infty \tag{35}$$

biçiminde tanımlanır. Eşitlik (35) herhangi bir yay uzunluğu 2π olan dairesel rastgele değişkenin olasılığının bire eşit olduğunu göstermektedir. Eğer $\alpha \leq \beta \leq \alpha + 2\pi$ ise $\Pr(\alpha < \theta \leq \beta)$ olasılığı,

$$\Pr(\alpha < \theta \leq \beta) = F(\beta) - F(\alpha) = \int_{\alpha}^{\beta} dF(x) \tag{36}$$

biçiminde hesaplanmaktadır. Dağılım fonksiyonu sağdan sürekli bir fonksiyondur ve doğrusal dağılım fonksiyonun tersine dağılım fonksiyonu,

$$\lim_{x\to\infty} F(x) = \infty \qquad \lim_{x\to-\infty} F(x) = -\infty \tag{37}$$

özelliklerini sağlamaktadır. F dağılım fonksiyonu sıfır yönünün seçimine bağlı olmasına rağmen $F(\beta) - F(\alpha)$ bu seçime bağlı değildir (Mardia & Jupp, 2000). Eğer dağılım fonksiyonu mutlak sürekli ise yani $-\infty < \alpha \leq \beta < \infty$ ise olasılık yoğunluk fonksiyonu,

$$\int_{\alpha}^{\beta} f(\theta)d\theta = F(\beta) - F(\alpha) \tag{38}$$

biçiminde tanımlanır. Birim çember üzerindeki en temel dağılım dairesel düzgün dağılımdır. Diğer önemli dağılımlar; "Von Mises Dağılımı", "Sarmal Normal Dağılım" ve "Üçgen Dağılım" dır. Von Mises dağılımı bilinen normal dağılıma benzer olarak, dairesel veri analizi teorisinin oluşturulmasında önemli bir rol oynamaktadır (Mardia & Jupp, 2000).

1.6.1. Dairesel Düzgün Dağılım

Elde edilen dairesel veriler çemberin çevresi üzerinde herhangi bir yöne doğru yoğunlaşma göstermeden düzgün olarak yayılım gösteriyorsa veriler dairesel düzgün dağılıma sahip demektir. Dairesel düzgün dağılım her rastgele değişken için sabit olasılık değerini veren tek dağılımdır (Mardia & Jupp, 2000). Dairesel düzgün dağılımın olasılık yoğunluk fonksiyonu,

$$f(\theta) = \frac{1}{2\pi}, \qquad 0 \leq \theta < 2\pi \tag{39}$$

biçiminde tanımlanmaktadır. Dairesel düzgün dağılımda birim çember üzerinde bulunan tüm veriler eşit olasılığa sahiptir. Bu nedenle dairesel düzgün dağılım izotopik dağılım veya rastgele dağılım olarak da bilinmektedir (Jammalamadaka & Gupta, 2001). Dağılım düzgün dağılım olmadığı zaman birim çember üzerinde temsil edilen verilerde farklı yönlerde yoğunlaşmalar görülmektedir.

1.6.2. Von Mises Dağılımı

Von Mises dağılımı ilk kez bir istatistik modeli olarak von Mises tarafından 1918 yılında ortaya atılmıştır (Von Mises, 1918). Gumbel ve diğerleri tarafından 1953 yılında dairesel normal dağılım olarak adlandırılmıştır, normal dağılımla benzerlikleri ve dağılımın önemi incelenmiştir (Gumbel, Greenwood, & Durand, 1953). Dairesel verilerin uygulama

problemlerinde genellikle von Mises dağılımı kullanılmaktadır. Von Mises dağılımına sahip bir θ rassal değişkeninin olasılık yoğunluk fonksiyonu, (40) eşitliğindeki gibi verilir.

$$f(\theta;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)}\quad,0\le\theta<2\pi \tag{40}$$

Bu eşitlikteki $0\le\mu<2\pi,\ \kappa\ge0$ von Mises dağılımının parametreleridir. Bu fonksiyon $VM(\mu,\kappa)$ ile de ifade edilebilir. Burada μ parametresi dairesel ortalama, κ parametresi ise yoğunlaşma parametresi olarak tanımlanır. Von Mises dağılımında κ değerinin büyük olması anakütle ortalaması ve modu etrafında daha büyük bir kümelenme olduğunu gösterir. Buna göre, κ dairesel ortalamaya ilişkin yoğunlaşmayı ölçen bir parametredir (Jammalamadaka & Gupta, 2001). $I_0(\kappa)$, birinci tür ve sıfır sırasında dönüştürülmüş Bessel fonksiyonudur ve Eşitlik (41)'deki gibi tanımlanır (Mardia K. V., 1972).

$$I_0(\kappa)=\frac{1}{2\pi}\int_0^{2\pi}e^{\kappa\cos(\phi-\mu)}d\phi=\sum_{r=0}^{\infty}\left(\frac{\kappa}{2}\right)^{2r}(r!)^{-2} \tag{41}$$

Daha önce de belirtildiği gibi, von Mises dağılımı dairesel veri analizinde en sık kullanılan dağılımdır. Ancak, çember üzerindeki normal dağılım olarak kabul edilebilmesine karşın, normal dağılımın tüm özelliklerine sahip değildir (Upton & Fingleton, 1989).

1.6.3. Sarmal Normal Dağılım

Sarmal normal dağılım (μ,σ^2) parametreli normal dağılımın birim çember üzerinde sarmal şekilde tekrarlanmasıyla elde edilmiştir. Bu dağılım fonksiyonu WN(μ,p) ile ifade edilir ve sarmal normal dağılımın olasılık yoğunluk fonksiyonu,

$$f(\theta;\mu,p)=\frac{1}{\sigma\sqrt{2\pi}}\sum_{-\infty}^{\infty}\exp\left\{\frac{-(\theta-\mu+2\pi\kappa)^2}{2\sigma^2}\right\} \tag{42}$$

biçiminde tanımlanır. Olasılık yoğunluk fonksiyonundaki $p = e^{-\frac{\sigma^2}{2}}$'dir. Sarmal normal dağılımın olasılık yoğunluk fonksiyonu normal dağılımın karakteristikleri kullanılarak elde edilmiştir (Jammalamadaka & Gupta, 2001). $\theta = \mu$ olduğu durumda sarmal normal dağılım simetrik ve tek modlu bir dağılım olur.

1.6.4. Üçgen Dağılım

Olasılık yoğunluk fonksiyonunun grafiği üçgen şeklinde olan üçgen dağılımın olasılık yoğunluk fonksiyonu,

$$f(\theta; p) = \frac{1}{8\pi}\{4 - \pi^2 p + 2\pi p|\pi - \theta|\}, \qquad 0 \leq \theta < 2\pi \tag{43}$$

biçimindedir. p üçgen dağılımın parametresidir ve $0 \leq p \leq \frac{4}{\pi^2}$ aralığında değer alır.

1.7. Kümeleme Analizi

İnsanoğlu var olduğundan bu yana etrafında bulunan nesneleri bazı özelliklerine göre sınıflara ayırma eğiliminde olmuştur. Birimlerin sayısı arttıkça birimleri sınıflandırmak daha da zorlaşmış ve yeni teknikler bulmayı gerektirmiştir. Bu gereksinim sonucu kümeleme analizi kavramı ortaya çıkmıştır (Günay Atbas, 2008). Kümeleme analizi çok değişkenli veri analiz yöntemlerinden biridir. Kümeleme analizi, bir araştırmada incelenen birimleri aralarındaki benzerliklerine ve farklılıklarına göre belirli gruplar içinde toplayarak sınıflandırma yapmayı, birimlerin ortak özelliklerini ortaya koymayı ve bu sınıflar ile ilgili genel tanımlamalar yapmayı sağlayan bir yöntemdir. Kümeleme işleminde küme içindeki elemanların benzerliği fazla, kümeler arası benzerlik ise az olmalıdır. Burada amaç; gruplanmamış verileri benzerliklerine göre sınıflandırmak ve araştırmacıya uygun, işe yarar özetleyici bilgiler elde etmede yardımcı olmaktır. Diğer bir amaç ise, benzer elemanların gruplanmasıyla veri setini küçültmektir (Işık & Çamurcu, 2010). Kümeleme analizi günümüzde veri madenciliği, bankacılık, pazarlama, tıp, sosyoloji, kriptoloji gibi çeşitli alanlarda kullanılmaktadır.

Çeşitli nesnelerden oluşan topluluğa klasik mantıkta küme denir. Küme teorisinin temeli bir elemanın kümeye ait olması veya olmamasına dayanır. Bir veri topluluğunu

kümelere ayırırken birçok farklı kümeleme seçeneği oluşturulabilir. Bu kümeleme seçenekleri bazı kriterlere göre değerlendirilip en iyi kümeleme seçeneğini bulmak bir optimizasyon problemi olarak ele alınabilir. Verilerin gruplandırılması sırasında üç temel aşama vardır.

1. Verilerin gruplandırılacağı optimum küme sayısı belirlenmelidir. Verilerin doğru gruplandırılmasının yanı sıra grup sayısının belirlemesi de kümeleme işleminde değerlendirmeye alınır.
2. Verilerin gruplandırılması sırasında kullanılacak benzerlik ölçütü belirlenmelidir. Bu benzerlik ölçütü aynı küme içerisindeki verilerin maksimum oranda benzerlik göstermesini sağlarken diğer gruptaki verilerle maksimum oranda farklılık göstermesini sağlamalıdır. En yaygın olarak kullanılan benzerlik ölçütü mesafeye dayalı benzerliktir.
3. Kümeleme işlemine en hızlı şekilde gerçekleştirecek yöntem belirlenmelidir (Ortakcı & Göloğlu, 2012).

Yönsel verilerin kümelenmesine istatistiksel açıdan bakılırsa, kümeleme yöntemleri genel olarak iki kategoriye ayrılabilir. Bunlar dağılıma bağımlı kümeleme yaklaşımları ile dağılımdan bağımsız kümeleme yaklaşımlarıdır. Dağılıma bağımlı yaklaşımlardan en yangın yöntemler en büyük beklenti algoritması (EM) (Dempster, Laird, & Rubin, 1977) (McLachlan & Basford, 1988) ile bulanık c-yönsel (FCD) (Yang & Pan, 1997) algoritmalarıdır. Dağılımdan bağımsız yaklaşımlardan ise en yaygın kullanılanı parçalı kümeleme yöntemleridir. Parçalı kümelemenin en bilinen algoritmaları k-ortalamalar (MacQueen, 1967) algoritması ve bulanık c-ortalamalar (FCM) kümeleme algoritmasıdır.

Günümüzde dairesel veriler için kümeleme analizi birçok uygulama alanında giderek önem kazanmaktadır. Dairesel istatistikte başlıca kaynaklar olarak Batschelet (1981), Fisher (1993) ve Mardia (2000) kitapları verilebilir. Bu kitaplarda dairesel verilerin analizi için değerli istatistiksel yaklaşımlar verilmiştir.

Genel olarak istatistiksel analiz, sayı doğrusu üzerinde değişim gösteren veriler için kullanılmaktadır (Kaufman & Rousseeuw, 1990). Ancak sayı doğrusu üzerindeki verilerde çalışan geleneksel istatistik yöntemleri doğası gereği yönsel veriler üzerinde çalışmamaktadır. Bunun en büyük nedeni yönsel verilerin modüler bir yapıya sahip olmasından kaynaklanmaktadır. Yönsel verilerde açısal değişim, $[-\pi, \pi)$ aralığında tanımlanmışsa $(\pi), (-\pi)$ noktaları arasında, $[0,2\pi)$ aralığında tanımlanmışsa $(2\pi), (0)$ noktaları arasında süreklidir. Ancak sayısal değerler açısından bu sınırlarda süreksizlik

göstermesi geleneksel yöntemlerin kullanımını geçersiz kılmaktadır. Bunun en net örneği, 359° ile 1° arasında 2° lik bir uzaklık varken sayısal farkta 358° lik bir uzaklık söz konusu olmaktadır.

Kümeleme algoritmaları benzerliklerin ölçüsünü veriler arasındaki uzaklıkları temel alarak hesaplamaktadır. Çember üzerinde verilen iki açı θ_a ve θ_b arasında açısal olarak saat yönünde ve saatin ters yönünde olmak üzere iki uzaklık bulunmaktadır. Genelde bu iki uzaklığın kısa olanını tercih edilmektedir (Mardia & Jupp, 2000).

Uzaklık kavramındaki bu karmaşanın çözümü için (Ackermann, 1997) tarafından önerilen eşitlikte dairesel uzaklık,

$$\delta = \pi - \left|\pi - |\theta_a - \theta_b|\right| \tag{44}$$

biçiminde verilmiştir. δ_{ij} ölçüsü θ_a ve θ_b arasında iki yayın küçük olanının uzunluğunu vermektedir.

İki açı arasındaki uzaklık için verilen diğer bir yaklaşım ise,

$$d = 1 - \cos(\theta_a - \theta_b) \tag{45}$$

biçiminde verilmektedir (Lund, 1999). (45) eşitliğindeki d uzaklığı $[0,2]$ aralığında değer almaktadır. Gerçekte (44) eşitliğinde verilen uzaklık ölçümü gerçek açısal uzaklığı vermesine rağmen birçok dairesel işlemde yetersiz kalmaktadır. Dolayısıyla dairesel verilerle ilgili işlemlerde çoğunlukla (45) eşitliği kullanılmaktadır.

1.8. Dağılımdan Bağımsız Kümeleme Yöntemleri

1.8.1. K-Ortalamalar Kümeleme Algoritması

En eski kümeleme algoritmalarından biri olan k-ortalamalar, 1967 yılında J.B. MacQueen tarafından geliştirilmiştir (MacQueen, 1967). En yaygın kullanılan gözetimsiz öğrenme yöntemlerinden birisi olan k-ortalamalar kümeleme algoritmasının atama mekanizması, her verinin sadece bir kümeye ait olabilmesine izin verir. Bu nedenle, keskin bir kümeleme algoritmasıdır. Merkez noktanın kümeyi temsil etmesi ana fikrine dayalı bir

metottur (Han & Kamber, 2006). K-ortalamalar kümeleme algoritması n tane nesneyi c tane kümeye bölmeyi amaçlar. Küme sayısı olan c parametresi giriş parametresi olarak önceden belirlenir. Küme içi benzerliğin yüksek fakat kümeler arası benzerliğin düşük olması amaçlanır. Benzerlik kavramı doğal olarak vektörler arasındaki uzaklık temeline dayanır. x ve y gibi iki vektör olduğunu varsayarsak bu vektörler arasındaki mesafe öklid uzaklığı olarak bilinir ve $d(x, y)$ ile gösterilir.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{46}$$

K-ortalamalar kümeleme algoritması, gözlemleri kümelerin önceden belirlenen sayısına göre gruplandırır. Böylece her biri tek gözlemden oluşan c tane küme ile işleme başlanır ve her bir yeni gözlem en yakın ortalamalı gruba eklenir. Gruba yeni bir gözlem eklendikten sonra küme ortalaması yeniden hesaplanır. Bu süreç tüm gözlemler gruplara atanıncaya kadar devam eder. Tüm gözlemler gruplara atandıktan sonra atandıkları küme ortalamasından daha yakın küme ortalaması varsa, gözlemlerin yerleri değiştirilmektedir. Amaç diğer kümeleme yöntemlerinde olduğu gibi, gerçekleştirilen kümeleme işlemi sonucunda elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin ise minimum olmasını sağlamaktır (Günay Atbas, 2008). Küme benzerliği, kümenin ağırlık merkezi kabul edilen bir birim ile kümedeki diğer birimler arasındaki uzaklıkların ortalama değeri ile ölçülmektedir (Han & Kamber, 2006).

Algoritma 1. K-Ortalamalar Kümeleme Algoritması

Adım 1:n tane veriden c tane veriyi rastgele seç. Seçilen c tane veri küme merkezlerini temsil eder $(v_1, v_2, \dots, v_c)$. Küme ortalamalarını (47) eşitliğini kullanarak hesapla.

$$v_c = \frac{1}{n_c} \sum_{\mathrm{i}=1}^{\mathrm{n_c}} x_{ic} \tag{47}$$

Adım 2: Tüm verilerin küme ortalamalarına olan uzaklıklarını hesapla.

Adım 3: Geriye kalan $(n - c)$ veriyi kendisine en yakın kümeye ata.

Adım 4: Verilerin hepsi en yakın kümelere atandığı zaman tekrar c tane küme için merkezleri hesapla.

Adım 5: Küme merkezlerinde bir değişiklik olmayıncaya kadar Adım 2 ve Adım 3'ü tekrarla

---

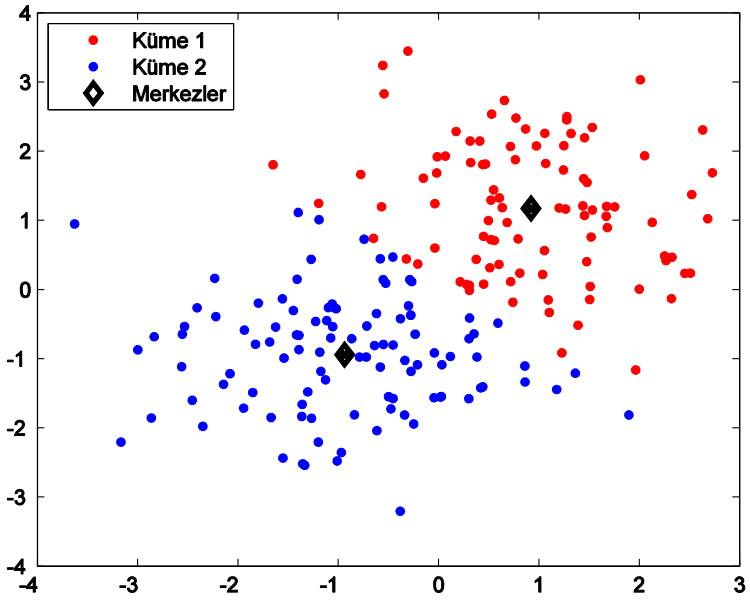
Kümeleme işleminin başarısını ölçmek için değişik performans göstergeleri oluşturulabilir. K-ortalamalar kümeleme algoritmasının değerlendirilmesinde en yaygın olarak karesel hata kriteri olan SSE kullanılır ve kümelemenin başarısı hakkında önemli bir bilgi verir. Karesel hataların toplamı,

$$J = \sum_{j=1}^{c} \sum_{i=1}^{n} \left\| x_i^j - v_j \right\|^2 \tag{48}$$

biçiminde hesaplanır. (48) eşitliğinde c merkez sayısını, n veri sayısını, $\left\| x_i^j - v_j \right\|$ ise x_i^j verisinin j. merkeze olan uzaklığını göstermektedir. j. kümedeki örneklerin ortalama vektörü (49) eşitliğindeki gibi hesaplanır (Pang-Ning, Steinbach, & Kumar, 2006).

$$c_j = \frac{1}{m_j} \sum_{\mathrm{x} \in \mathrm{c_j}} x \tag{49}$$

Bu kriter sonucu c tane kümenin olabildiğince yoğun ve birbirinden ayrı sonuçlanması hedeflenmeye çalışılır. Algoritma, karesel-hata fonksiyonunu azaltacak c parçayı belirlemeye gayret eder (Miller & Han, 2009). En düşük SSE değerine sahip kümeleme sonucu en iyi sonucu verir. Karesel hata kriteri verilerin küme merkezlerine olan uzaklıklarının karelerinin toplamıdır (Işık, Meltem; Çamurcu, A. Yılmaz, 2007). Bu algoritmanın üstünlüğü, büyük veri setlerinde çalışıldığında basit ve hızlı olmasıdır. Sakıncası ise her iterasyonda aynı sonucu üretmemesidir. Çünkü ortaya çıkan kümeler başlangıç rastgele atamasına bağlıdır.



Şekil 8. K-ortalamalar algoritmasıyla iki kümeye ayrılmış veri

1.8.2. Bulanık C-Ortalamalar (FCM) Kümeleme Algoritması

Bazı kümeleme problemlerinde kümeler birbirinden k-ortalamalar kümeleme algoritmasında olduğu gibi belirgin bir şekilde ayrılmıyorsa ya da bazı birimler küme üyeliğinde kararsızsa, klasik kümeleme yöntemleri yerine bulanık kümeleme yöntemleri tercih edilmelidir. Bulanık c-ortalamalar (FCM) kümeleme algoritması, bulanık kümeleme tekniklerinden en iyi bilinen ve en yaygın kullanılan yöntemdir. Bulanık c-ortalamalar kümeleme algoritması 1973 yılında Dunn tarafından ortaya atılmış ve 1981'de Bezdek tarafından geliştirilmiştir (Höppner, Klawonn, Kruse, & Runkler, 2000). FCM algoritması amaç fonksiyonu temelli bir metottur ve verilerin birden fazla kümeye farklı üyelik dereceleriyle ait olabilmesi prensibine dayanır. Bulanık mantık prensibi gereği bu üyelik dereceleri [0,1] arasında değişen değerler almaktadır ve bir verinin tüm kümelere ait üyelik derecelerinin toplamı 1 olmalıdır. Nesne hangi küme merkezine yakın ise o kümeye ait olma üyeliği diğer kümelere ait olma üyeliğinden daha büyük olacaktır.

D-boyutlu bir Euclidean uzayında N örnekten oluşan bir $X = \{x_1, x_2, \dots, x_N\}$ veri kümesinin verildiğini varsayalım ($x_i \in R^D$). Kümeleme, bu veri kümesinin, küme merkezleri $\{v_1, v_2, \dots, v_j, \dots, v_C\}$ olan c tane alt kümeye ayrılması işlemidir. Veri kümesini alt kümelere ayırırken istenen optimal kriter amaç fonksiyonunu minimize etmektir.

Algoritma, en küçük kareler yönteminin genellemesi olan (50) eşitliğindeki amaç fonksiyonunu öteleyerek minimize etmek için çalışır (Höppner, Klawonn, Kruse, & Runkler, 2000).

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} \mu_{ij}^{m} \left|\left|x_i - v_j\right|\right|^2 \tag{50}$$

Burada m değeri bulanıklaştırma parametresidir $1 < m < \infty$ aralığında değerler alabilmektedir ama genelde değeri 2 olarak seçilmektedir. μ_{ij} ise i. elemanın j. kümedeki üyelik değeridir ve (51-53) eşitliklerindeki koşulları sağlamak zorundadır (Bezdek J. C., 1981).

$$\mu_{ij} \in [0,1], \quad \forall\, i,j \tag{51}$$

$$\sum_{j=1}^{C} \mu_{ij} = 1\,, \qquad \forall i \tag{52}$$

$$0 < \sum_{i=1}^{N} \mu_{ij} < N, \qquad \forall N \tag{53}$$

Bulanık c-ortalamalar algoritması basit bir yöntem olmasının yanı sıra, tüm bulanık kümeleme yöntemlerin içerisinde hala en yaygın kullanıma sahip olan bir yöntemdir (Bezdek J. C., 1981). Temel olarak k-ortalamalar algoritmasına çok benzemekle beraber FCM algoritmasının, k-ortalamalar algoritmasından en önemli farkı verilerin her birinin sadece bir sınıfa dahil edilme zorunluluğunun olmamasıdır (Işık & Çamurcu, 2010). FCM kümeleme algoritması aşağıdaki gibi verilebilmektedir.

Algoritma 2. Bulanık C-Ortalamalar Kümeleme Algoritması

Adım 1: Verileri (x_i), önceden belirlenen küme sayısını $c \in [2, N)$, bulanıklaştırma parametresini $(m > 0)$ ve durdurma kriterini $(\varepsilon > 0)$ algoritmaya giriş verileri olarak ver.

Adım 2: Başlangıç üyelik değerlerini $\mu^{(0)} = [\mu_{ij}]$, $[0,1)$ aralığında düzgün dağılımdan rastgele olarak belirle, çevrim sayacını bire $(\mathrm{t} = 1)$ eşitle.

Adım 3: t. çevrimini başlat,

Adım 4: Yeni üyelik değerlerini kullanarak küme merkezlerini (54) eşitliği yardımıyla hesapla,

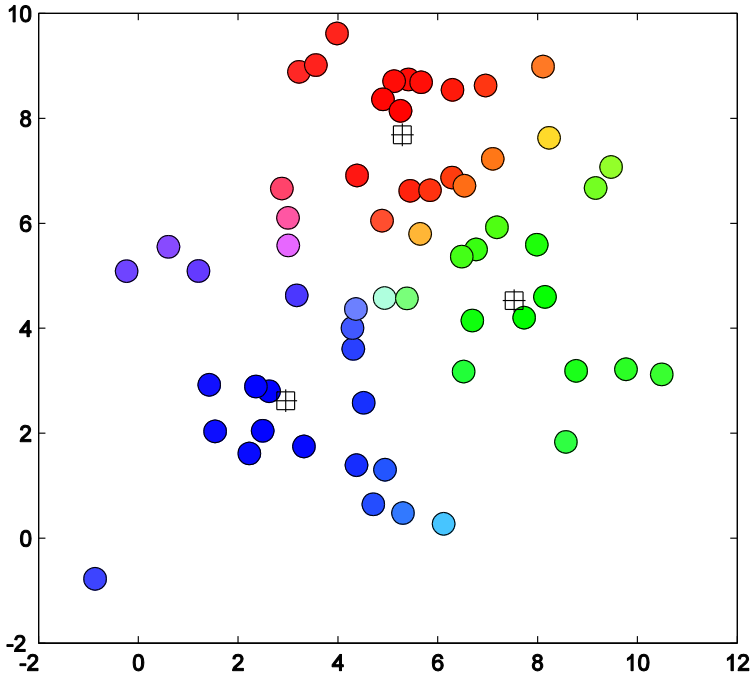
$$v_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}}, \quad (j = 1,2, \dots, c) \tag{54}$$

Adım 5: Üyelik değerlerini (55) eşitliği yardımıyla güncelleştir,

$$\mu_{ij} = \left(\sum_{k=1}^{C} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}} \right)^{-1} \tag{55}$$

Adım 6: (56) eşitsizliği sağlanıyorsa dur, yoksa t değerini bir artır ve Adım 3.'e giderek yeni çevrimi başlat,

$$\left\| \mu^{(\mathrm{t})} - \mu^{(\mathrm{t}-1)} \right\| < \varepsilon \tag{56}$$



Şekil 9. Bulanık c-ortalamalar algoritmasıyla üç kümeye ayrılan veri

1.8.3. Bulanık Kümeleme Algoritmalarında Optimal Küme Sayısı

Kümeleme analizinde, anlamlı ve sağlıklı sonuçlara ulaşabilmek için en uygun küme sayısının belirlenmesi önemli bir problemdir. Çoğu kümeleme algoritmasında küme sayısı algoritmaya giriş parametresi olarak verilmektedir. Birçok çalışmada, araştırmacının küme sayısı hakkında ön bilgisinin olmaması, bulunan küme sayısının gerçek küme sayısından az ya da çok olup olmadığının bilinmemesine yol açmaktadır. Eğer bulunan küme sayısı gerçek küme sayısından az çıkarsa, mevcut kümelerden bir veya birkaçı birleşmek durumunda olacaktır, çok çıkarsa mevcut kümelerden bir veya birkaçı bölünmelere uğrayacaktır. Farklı başlangıç küme sayısı seçimleri, farklı kümelenmelerin ortaya çıkmasına sebep olur. Onun için, kümelenme analizinden sonra her bir bulanık kümenin doğrulamasının yapılması gereklidir (Murat & Şekerler, 2009). Optimal küme sayısının belirlenme işlemlerine genel olarak Küme Geçerliliği (Cluster Validity) adı verilmektedir. Bazı karmaşık yapılar içeren verilerde, küme üyeliklerindeki karasızlıklar nedeniyle, küme geçerlilik indeksleri en uygun küme sayısını belirlemede birbirleri ile çelişen sonuçlar verebilmektedir. Ayrıca hangi indeksin en uygun küme sayısını belirlediğini ortaya koyan bir ölçüt de bulunmamaktadır (Alpaslan, Erilli, Yolcu, Eğrioğlu, & Aladağ, 2011).

Genel olarak, en uygun küme sayısının belirlenebilmesi için iki kriterden bahsedilir. Bu kriterlerden ilki küme elemanlarının birbiriyle olan yakınlıklarını ölçen yoğunluk kriteri diğeri de iki kümenin birbirinden ne kadar ayrıldıklarını gösteren yani iki küme arasındaki mesafeyi ölçen ayrılma kriteridir.

İki boyutlu verileri kümelere ayırırken bu verilerin dağılımını görsel olarak yorumlayıp küme sayısına karar verebiliriz. Ancak verilerin boyut sayısı arttıkça küme sayısına karar vermekte zorlandığımızdan ve araştırmacının küme sayısına karar vermedeki öznelliğini minimize etmek için küme geçerlilik indekslerine ihtiyaç duyulmaktadır. Bu küme geçerlilik indekslerinden literatürdeki en çok kullanılanları şunlardır:

1.8.3.1. Bölünme Katsayısı (PC) ve Bölünme Entropisi (CE)

Bölünme katsayısı ve bölünme entropisi indeksi Bezdek tarafından önerilmiştir (Bezdek J. C., 1974). Bu küme geçerlilik indekslerinin dezavantajı sadece üyelik derecelerini kullanmasıdır ve kümelerin veri yapılarını göz önünde bulundurmamasıdır (Kim, Lee, & Lee, 2004). Bölünme katsayısı indeksi (PC) iki bulanık kümenin üst üste gelme miktarını ölçer ve (57) eşitliği yardımıyla hesaplanır. Optimum küme sayısı, bu indeksin maksimum değerine karşılık gelen küme sayısıdır.

$$v_{PC} = \frac{\sum_{j=1}^{N}\sum_{i=1}^{c}\mu_{ij}^{2}}{n} \tag{57}$$

Bölünme entropisi indeksi küme ayrımlarının bulanıklığını ölçer (Murat & Şekerler, 2009) ve (58) eşitliği yardımıyla hesaplanır. Optimum küme sayısı, bu indeksin minimum değerine karşılık gelen küme sayısıdır.

$$v_{PE} = -\frac{1}{n}\sum_{j=1}^{N}\sum_{i=1}^{c}\left[\mu_{ij} log_a(\mu_{ij})\right] \tag{58}$$

1.8.3.2. Xie-Beni İndeksi(XB)

Xie ve Beni tarafından geliştirilen bu indeks, yoğunluk ve ayrılma geçerlilik fonksiyonu olarak da bilinir (Xie & Beni, 1991) ve (59) eşitliği ile hesaplanır. Optimum küme sayısı, bu indeksin minimum değerine karşılık gelen küme sayısıdır.

$$v_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}^{2}\left\|x_j - v_i\right\|^2}{n(\underset{i\neq k}{min}\|v_i - v_k\|^2)} \tag{59}$$

1.8.3.3. Fukuyama and Sugeno İndeksi(FS)

Fukuyama ve Sugeno tarafından geliştirilen bu indekste de yoğunluk ve ayrılma kavramlarından faydalanılmıştır (Fukuyama & Sugeno, 1989). Bu formüldeki J_m yoğunluk ölçüsünü, K_m ise kümeler arasındaki ayrılmanın derecesini gösterir ve (60-61) eşitlikleri kullanılarak hesaplanırlar. Optimum küme sayısı, bu indeksin minimum değerine karşılık gelen küme sayısıdır.

$$v_{FS} = J_m(U, V{:}X) - K_m(U, V{:}X) \tag{60}$$

$$v_{FS} = \sum_{j=1}^{N}\sum_{i=1}^{c}\mu_{ij}^{m}\,\|x_k - v_i\|^2 - \sum_{j=1}^{N}\sum_{i=1}^{c}\mu_{ij}^{m}\|v_i - \bar{v}\|^2 \tag{61}$$

1.8.3.4. K İndeksi

Kwon tarafından önerilen bu indekste Xie ve Beni indeksindeki küme sayısının veri sayısına yakın olması eğilimi azaltılmaya çalışılmıştır ve (62) eşitliği kullanılarak hesaplanmaktadır (Kwon, 1998). Optimum küme sayısı, bu indeksin minimum değerine karşılık gelen küme sayısıdır.

$$v_{VK} = \frac{\sum_{j=1}^{N}\sum_{i=1}^{c}\mu_{ij}^{2}\left\|x_j - v_i\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|v_i - \bar{v}\|^2}{\min\limits_{i\neq k}\|v_i - v_k\|^2} \tag{62}$$

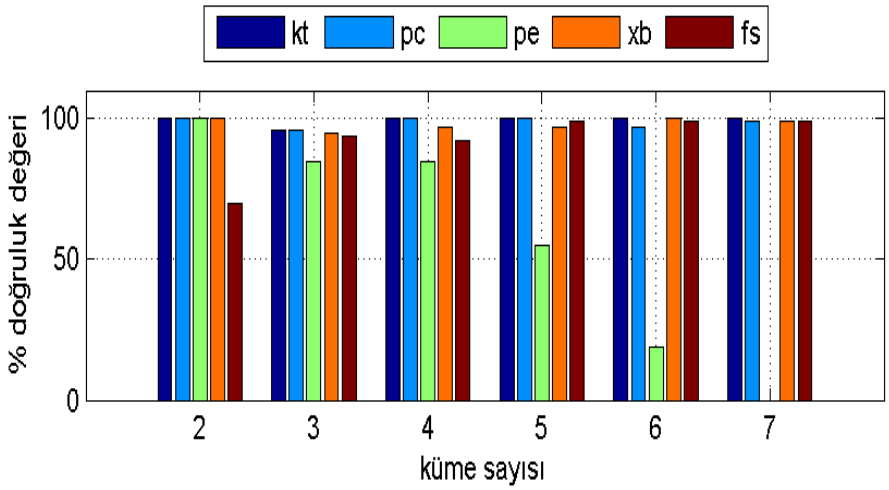
1.8.3.5. KT İndeksi

Kesemen ve Tezel tarafından önerilen bu indekste sadece küme içi benzerlik değil küme dışı farklılıklarda dikkate alınmıştır (Kesemen & Tezel, 2013). Bir elemanın tüm kümelere üyelik değerinin toplamı bire eşit olduğundan, elamanın ait olduğu kümeye üyelik değeri (μ_{ij}) olarak alınırsa, üye olmama değeri ise $(1-\mu_{ij})$ olacaktır. Bu iki değerin oranlanmasıyla elde edilen formülasyon (63) eşitliğinde verilmiştir. Optimum küme sayısı, bu indeksin maksimum değerine karşılık gelen küme sayısıdır.

$$v_{KT} = \sum_{j=1}^{N}\sum_{i=1}^{c}\left(\frac{\mu_{ij}^{2}}{\left(\Gamma - \mu_{ij}\right)^2}\right) \tag{63}$$

KT indeksindeki Γ teriminin en iyi değerini bulabilmek için bu terim [1,6] aralığındaki değerler (63) eşitliğinde yerine koyularak denenmiştir ve en iyi sonucun 2 değeri ile elde edildiği görülmüştür.

Bu indeks diğer küme geçerlilik indeksleriyle karşılaştırılarak optimal küme sayısının belirlenmesinde etkili bir yöntem olduğu yüzde doğruluk değeri ile gösterilmiştir.



Şekil 10. KT indeksinin yüzde doğruluk değeri açısından verilen diğer indekslerle karşılaştırılması

1.9. Dağılıma Bağımlı Kümeleme Yöntemleri

1.9.1. En Büyük Beklenti (EM) Kümeleme Algoritması

En büyük beklenti (EM) algoritması bir objenin hangi kümeye ait olduğunu belirlemede kesin mesafe ölçütlerini kullanmak yerine tahminsel ölçütleri kullanmayı tercih eder. EM algoritması ilk olarak Dempster, Laird ve Rubin tarafından 1977 yılında ortaya atılmıştır (Dempster, Laird, & Rubin, 1977). EM algoritması verinin eksik veri olması durumunda en çok olabilirlik tahmini için genel bir istatistiksel yöntemdir. EM algoritması döngüsel (iteratif) bir algoritmadır. EM algoritmasının her iterasyonu iki adımda gerçekleşir. EM algoritması beklenti (E) adımı ve maksimizasyon (M) adımlarından oluşur. Bu adımlar belirli bir yakınsama kriteri sağlanana kadar ardışık olarak gerçekleştirilir (Servi, 2009). E-adımında gözlenen verilerin parametrelerine ait kestirimler kullanılarak bilinmeyen (kayıp) veri ile ilgili en iyi olasılıklar tahmin edilirken, M-adımında ise tahmin edilen kayıp veri yerine konulup bütün veri üzerinden maksimum olabilirlik hesaplanarak parametrelerin yeni kestirimleri elde edilir (Bruzzone & Prieto, 2002). EM algoritması için alınan parametre başlangıç değerlerinin seçimi oldukça önemlidir. EM algoritmasının ardışık bir algoritma olması nedeniyle yakınsama bazı durumlarda oldukça yavaş olabilmektedir. Bazı durumlarda da EM algoritması parametrenin global bir maksimum değeri yerine yerel bir maksimum değerine yakınsayabilmektedir. EM algoritmasının bu tür dezavantajlarının algoritmada yapılacak

bazı yeniden düzenlemeler ile parametre başlangıç değerlerinin seçimi içinde ayrı bir algoritma veya yöntem kullanılması durumunda yerel maksimuma yakınsama tehlikesinin ortadan kalkabileceği ve hibrit algoritmalar kullanılırsa algoritmanın yakınsama hızının artabileceği belirtilmiştir (Servi, 2009). EM algoritması doğrusal verilere uygulanabildiği gibi dairesel verilere de uygulanabilmektedir. Dairesel verilerin kümelenmesinde en yaygın kullanılan yöntemlerden birisi EM (Dempster, Laird, & Rubin, 1977)- (Chang-Chien, Hung, & Yang, 2012) algoritmasıdır. Burada, $\Theta = \{\theta_1, \theta_2, \dots, \theta_N\}$ veri noktalarının karışık von Mises dağılımından elde edilmiş rastgele örnekler olduğu kabul edilmektedir. Dairesel veriler için kullanılan EM kümeleme algoritması aşağıdaki gibi verilebilmektedir.

Algoritma 3. EM Kümeleme Algoritması

Adım 1: Küme sayısını $c \in [2, N)$ ve tolerans değerini $(\varepsilon > 0)$ belirle,

Adım 2: Aşağıdaki koşullar altında başlangıç gösterge $z^{(0)}$ değerlerini rastgele belirle,

$$z_{ij} \sim U_d(0,1) \ , \qquad \left(\sum_{j=1}^{C} z_{ij} = 1 \right) \tag{64}$$

Burada $U_d(0,1)$, $\{0,1\}$ değerlerine sahip ayrık düzgün dağılımı göstermektedir. Çevrim sayacını bire $(t = 1)$ eşitle.

Adım 3: (t). çevrimini başlat,

Adım 4: $\alpha^{(t)}$ değerlerini, $z^{(t-1)}$ değerlerini ile (65) eşitliğini kullanarak hesapla,

$$\alpha_j = \frac{1}{N} \sum_{i=1}^{N} z_{ij} \, , \qquad (j = 1,2, \dots, c) \tag{65}$$

Adım 5: $v^{(t)}$ değerlerini $z^{(t-1)}$ değerleri ile (66) eşitliğini kullanarak hesapla,

$$v_j = atan2 \left(\sum_{i=1}^{N} z_{ij} \sin(\theta_i) \, , \sum_{i=1}^{N} z_{ij} \cos(\theta_i) \right), \qquad (j = 1,2, \dots, c) \tag{66}$$

Burada $atan2$ fonksiyonu $[-\pi, \pi)$ aralığında tanımlı ters tanjant fonksiyondur.

Adım 6: $\kappa^{(t)}$ değerlerini $z^{(t-1)}$, $v^{(t)}$ ve θ değerleri ile (67) eşitliğini kullanarak hesapla,

$$\kappa_j = A^{-1}\left(\frac{\sum_{i=1}^{N} z_{ij}\cos\left(\theta_i - v_j\right)}{\sum_{i=1}^{N} z_{ij}}\right), \qquad (j = 1,2,\dots,c) \tag{67}$$

Burada $A^{-1}(x)$ Batschelet tablosundan hesaplanabilir (Fisher, 1993).

Adım 7: $z^{(t)}$ değerlerini $\alpha^{(t)}$, $v^{(t)}$, $\kappa^{(t)}$ ve θ değerleri ile (68) eşitliğini kullanarak hesapla,

$$z_{ij} = \frac{\alpha_j f\left(\theta_i; v_j, \kappa_j\right)}{\sum_{k=1}^{C} \alpha_k f(\theta_i; v_k, \kappa_k)}, \quad \begin{matrix}(i = 1,2,\dots,N \\ j = 1,2,\dots,c)\end{matrix} \tag{68}$$

Burada $f(\theta; v, \kappa)$ fonksiyonu von Mises dağılımının olasılık yoğunluk fonksiyonunu göstermektedir.

Adım 8: (69) eşitsizliği sağlanıyorsa dur, yoksa t değerini bir artır ve Adım 3.'e giderek yeni çevrimi başlat,

$$\left\|z^{(t)} - z^{(t-1)}\right\| < \varepsilon \tag{69}$$

1.9.2. Bulanık C-Yönsel (FCD) Kümeleme Algoritması

Bulanık c-yönsel (FCD) kümeleme algoritması Yang ve Pan (1997) tarafından geliştirilmiştir (Yang & Pan, 1997). Bu algoritma verilerin birden fazla kümeye [0,1] arasında değişen değerler alan farklı üyelik dereceleriyle ait olabilmesi sağlayan bir kümeleme algoritmasıdır. FCD ve EM kümeleme algoritmaları esas yapısı itibari ile birbirine çok benzemektedir. Bu nedenle FCD kümeleme algoritması EM algoritmasıyla aynı dezavantajlara sahiptir. Fakat EM algoritmasından farklı olarak FCD algoritmasında, her veri noktasının tüm kümelere olan üyelik fonksiyonları bulanık olarak hesaplanmaktadır. Buradaki genel kabul, $\Theta = \{\theta_1, \theta_2, \dots, \theta_N\}$ veri noktalarının karışık von Mises dağılımından elde edilmiş rastgele örnekler olduğudur. FCD kümeleme algoritması aşağıdaki gibi verilebilmektedir.

Algoritma 4. FCD Kümeleme Algoritması

Adım 1: İstenen küme sayısını $c \in [2, N)$, bulanıklaştırma parametresini $(m > 1)$, durdurma kriterini $(\varepsilon > 0)$ ve $w > 0$ değerini belirle, Çevrim sayacını bire $(t = 1)$ eşitle.

Adım 2: Aşağıdaki koşullar altında başlangıç üyelik değerlerini rastgele $\mu^{(0)}$ belirle,

$$\mu_{ij} \sim U(0,1), \qquad \left(\sum_{j=1}^{C} \mu_{ij} = 1 \right) \tag{70}$$

Burada $U(0,1)$, $[0,1]$ aralığında sürekli düzgün dağılımdır.

Adım 3: (t). çevrimini başlat,

Adım 4: $\alpha^{(t)}$ değerlerini, $\mu^{(t-1)}$ değerlerini kullanarak hesapla,

$$\alpha_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m}}{\sum_{j=1}^{c} \sum_{i=1}^{N} \mu_{ij}^{m}}, \qquad (j = 1,2, \dots, c) \tag{71}$$

Adım 5: $v^{(t)}$ değerlerini $\mu^{(t-1)}$ değerleri ile (72) eşitliğini kullanarak hesapla

$$v_j = atan2\left(\sum_{i=1}^{N} \mu_{ij}^{m} \sin(\theta_i), \sum_{i=1}^{N} \mu_{ij}^{m} \cos(\theta_i) \right), \qquad (j = 1,2, \dots, c) \tag{72}$$

Burada atan2 fonksiyonu $[-\pi,\pi)$ aralığında tanımlı ters tanjant fonksiyondur.

Adım 6: $\kappa^{(t)}$ değerlerini $\mu^{(t-1)}$, $v^{(t)}$ ve θ değerleri ile (73) eşitliğini kullanarak hesapla,

$$\kappa_j = A^{-1}\left(\frac{\left(\sum_{i=1}^{N} \mu_{ij}^{m} \cos(\theta_i - v_j) \right)}{\sum_{i=1}^{N} \mu_{ij}^{m}} \right), \qquad (j = 1,2, \dots, c) \tag{73}$$

Burada $A^{-1}(x)$ Batschelet tablosundan hesaplanabilir (Fisher, 1993).

Adım 7: $\mu^{(t)}$ değerlerini $\alpha^{(t)}$, $v^{(t)}$, $\kappa^{(t)}$ ve θ değerleri ile (74) eşitliğini kullanarak hesapla,

$$\mu_{ij} = \left(\sum_{k=1}^{C} \frac{\left(\log\left(2\pi I_0(\kappa_j)\right) - \kappa_j \cos(\theta_i - v_j) - w \log(\alpha_j) \right)^{\frac{1}{m-1}}}{\left(\log(2\pi I_0(\kappa_k)) - \kappa_k \cos(\theta_i - v_k) - w \log(\alpha_k) \right)^{\frac{1}{m-1}}} \right)^{-1} \quad (74)$$

$$(i = 1,2,\dots,N;\ j = 1,2,\dots,c)$$

Adım 8: (75) eşitsizliği sağlanıyorsa dur, yoksa t değerini bir artır ve Adım 3.'e giderek yeni çevrimi başlat,

$$\left\| \mu^{(t)} - \mu^{(t-1)} \right\| < \varepsilon \quad (75)$$

2. YAPILAN ÇALIŞMALAR

Önceleri uzun zaman alan birçok iş günümüzde bilgisayar teknolojilerinin gelişmesiyle birlikte otomatik olarak yapılmaktadır. Bu işlemleri gerçekleştirmek için sıklıkla yapay öğrenme teknikleri kullanılmaktadır. Yapay öğrenme tekniklerinde en temel yaklaşım karar verme sürecinde veri analizi yapılmasıdır. Yönsel verilerin gruplandırılmasında geliştirilen birçok kümeleme algoritması von Mises dağılımını temel alarak geliştirilmiştir. Bu yöntemlerin en önemli eksikliği trigonometrik fonksiyonlar yardımıyla hesaplanan yaklaşık uzaklıkları kullandıkları için gerçek uzaklıktan sapmaların yaşanmasıdır.

2.1. Yönsel Veriler İçin Bulanık C-Ortalamalar Kümeleme Algoritması

Bu çalışmada kümeleme analizinde kullanılan FCM kümeleme algoritmasının esas yapısı dikkate alınarak yönsel veriler için uyarlanmıştır. Yönsel veriler için bulanık c-ortalamalar (FCM4DD) kümeleme algoritması yönsel veriler için uyarlanan hızlı ve basit bir yöntemdir. Bu yöntem, yönsel veriler için (45) eşitliğinde verilen uzaklık ölçüsü yerine (44) eşitliğinde verilen iki açı arasındaki farkı temel alan bir yöntemdir. Bu yöntem kullandığı uzaklık ölçüsü bakımından FCM algoritmasına benzerlik göstermektedir.

Yönsel veriler $\Theta = \{\theta_1, \theta_2, \dots, \theta_N\}$ biçiminde verildiği gibi periyodik $X = \{x_1, x_2, \dots, x_i, \dots, x_N\}$ biçiminde de verilebilir. Eğer veriler periyodik değerler biçiminde verilmişse (76) eşitliği yardımıyla yönsel verilere dönüştürülebilir.

$$\theta_i = 2\pi\left(\frac{x_i}{T} - \frac{1}{2}\right) \tag{76}$$

burada T, x_i değişkeninin periyodunu vermektedir.

Bu çalışmada yönsel veriler, $\theta_i \in [-\pi, \pi)$ aralığında tanımlanabileceği gibi $\theta_i \in [0, 2\pi)$ aralığında da tanımlanabilmektedir.

FCM4DD kümeleme algoritması FCM kümeleme algoritması gibi amaç fonksiyonu temelli bir metottur ve verilerin birden fazla kümeye farklı üyelik dereceleriyle ait olabilmesi prensibine dayanır. Önerilen kümeleme algoritması, en küçük kareler

yönteminin genellemesi olan (77) eşitliğindeki amaç fonksiyonunu minimize etmek için çalışır.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{c} \mu_{ij}^{m} \left\| \theta_i - \Phi_j \right\|^2, \quad 1 < m < \infty \tag{77}$$

Burada m değeri bulanıklaştırma parametresidir ve genellikle değeri iki olarak seçilir. Φ_j j. küme merkezini, μ_{ij} ise i. verinin j. kümeye olan üyelik derecesini göstermektedir ve μ_{ij} (51-53) eşitliklerindeki koşulları sağlamalıdır. Yönsel veriler için önerilen FCM4DD kümeleme algoritması aşağıdaki gibi verilebilmektedir.

Algoritma 5. FCM4DD Kümeleme Algoritması

Adım 1: Yönsel verileri (θ_i), önceden belirlenen küme sayısını (c), bulanıklaştırma (m) parametresini ve durdurma kriterini (ε) algoritmaya giriş verileri olarak ver.

Adım 2: Başlangıçtaki küme merkezlerini $\phi_j^{(0)}$ $[-\pi, \pi)$ aralığında düzgün dağılımdan rastgele olarak belirle ve (t) çevrim indisini 1'e eşitle.

Adım 3: (t). çevrimi başlat,

Adım 4: Her bir verinin geçici küme merkezine göre iki açı arasındaki farkı (78) eşitliğiyle hesapla ve bu farkı $[-\pi, \pi)$ aralığına indirge,

$$\psi_{ij} = \left(\left| \left(\theta_i - \phi_j \right) + 3\pi \right| \, mod \, 2\pi \right) - \pi \tag{78}$$

Adım 5: Bu fark yardımıyla üyelik değerlerini (79) eşitliğini kullanarak hesapla,

$$\mu_{ij} = \left(\sum_{k=1}^{c} \left(\frac{\|\psi_{ij}\|}{\|\psi_{ik}\|} \right)^{\frac{2}{m-1}} \right)^{-1} \quad \begin{matrix} i = 1,2,\ldots,N \\ , j = 1,2,\ldots,c \end{matrix} \tag{79}$$

Adım 6: İki açı arasındaki fark ve yeni üyelik değerleri kullanılarak küme merkezleri (80) eşitliğini kullanarak güncelle,

$$\phi_j^{(t+1)} = \left(\phi_j^{(t)} + \frac{\sum_{i=1}^{N} \mu_{ij}^{m} \psi_{ij}}{\sum_{i=1}^{N} \mu_{ij}^{m}} \right) mod\ 2\pi, \quad (j = 1,2, \dots, c) \tag{80}$$

Adım 7: (81) eşitsizliği sağlanıyorsa dur, yoksa Adım 3.'e git ve (t+1). çevrimi başlat.

$$\left\| \mu^{(t)} - \mu^{(t-1)} \right\| < \varepsilon \tag{81}$$

2.2. Örnek Uygulamalar

Yönsel verilerin kümelenmesinde kullanılan en yaygın iki yöntem olan EM ve FCD kümeleme algoritmaları ile önerilen FCM4DD kümeleme algoritmasının başarımları ve FCM4DD algoritmasının küresel verilere uygulanabilirliği 6 örnek uygulamayla gösterilmiştir.

2.2.1. Örnek Uygulama 1

Yöntemlerin başarımlarını karşılaştırmak için $0.7VM(\pi/3, 2.3) + 0.3VM(4\pi/3, 3.8)$ biçiminde verilmiş karışık von Mises dağılımından Best ve Fisher'in benzetim yöntemi (Best & Fisher, 1979) kullanılarak rastgele 100 örnek üretilmiştir. Bu örnekler EM, FCD ve FCM4DD algoritmaları ile tek tek kümelenmiştir. Bu işlem 1000 kez tekrarlanarak her yöntem için bulunan küme merkezlerinin ortalamaları ($\bar{v}_1; \bar{v}_2$), Tablo 1'de verilmiştir.

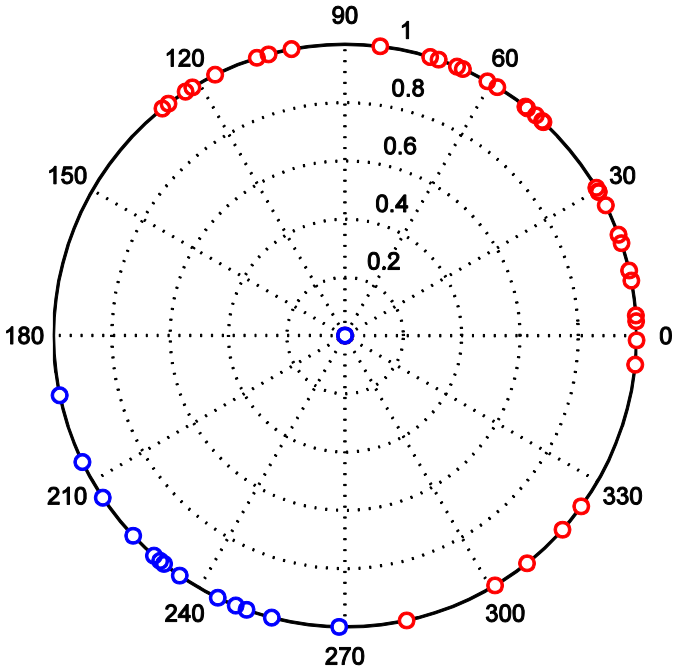
Tablo 1. Örnek Uygulama 1 için kümeleme yöntemlerinin karşılaştırılması

Algoritmalar	EM	FCD	FCM4DD
$\bar{v}_1$(deg)	59.7129	60.4142	59.9552
$\bar{v}_2$(deg)	239.6176	239.2764	240.2741
MSE	0.00006966	0.00021177	0.00002350
ACT (sec)	0.20307981	0.10858994	0.00094319

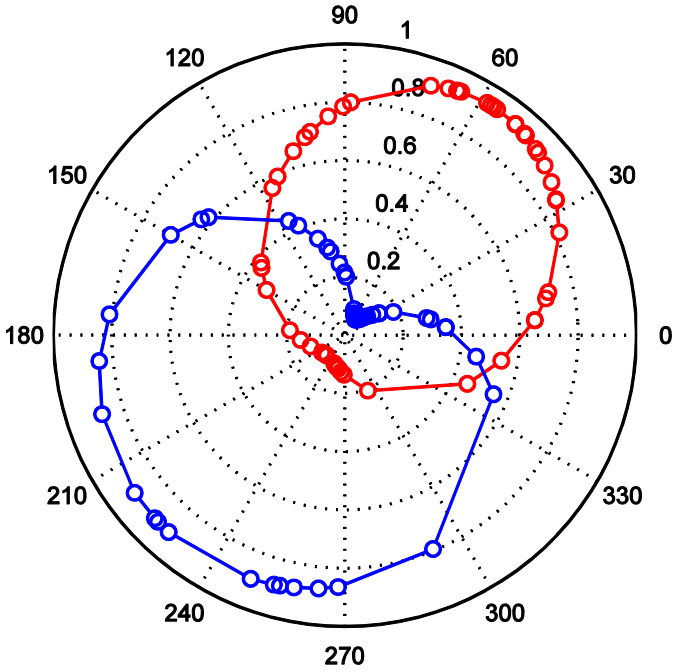
Burada ACT, ortalama hesaplama süresini saniye cinsinden göstermektedir. MSE ise ortalama hata karelerini göstermekte ve aşağıdaki eşitlik yardımıyla hesaplanmaktadır.

$$MSE = \frac{1}{2}[(\bar{v}_1 - \pi/3)^2 + (\bar{v}_2 - 4\pi/3)^2] \tag{82}$$

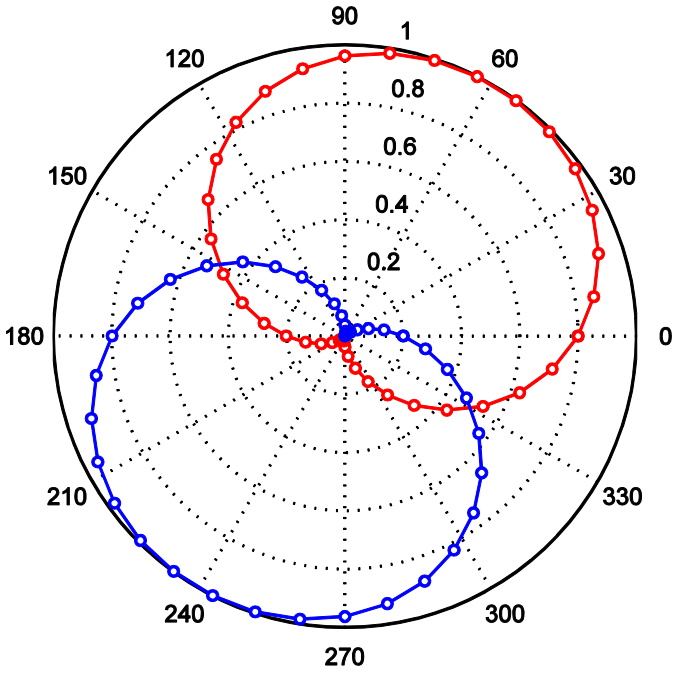
MSE sonuçlarına göre en iyi sonucu FCM4DD algoritması vermektedir. Hesaplama zamanı açısından FCM4DD algoritması diğer iki algoritmaya göre oldukça üstün bir başarı göstermektedir.



Şekil 11. Örnek uygulama 1'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 12. Örnek uygulama 1'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 13. Örnek uygulama 1'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi

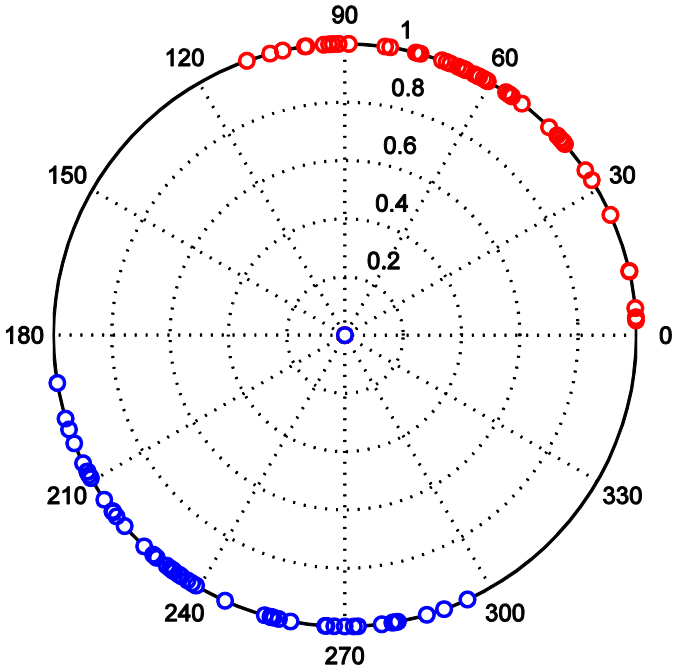
2.2.2. Örnek Uygulama 2

Bu örnekte ortalamaları $(\pi/3,4\pi/3)$ olan iki üçgen dağılımın karışık olasılık yoğunluk fonksiyonu $0.5\Lambda\left(0,\frac{\pi}{3},\frac{2\pi}{3}\right)+0.5\Lambda\left(\pi,\frac{4\pi}{3},\frac{5\pi}{3}\right)$ biçiminde verilmiş olsun. Bu dağılımdan rastgele 100 örnek oluşturulmuştur. Bu örnekler EM, FCD ve FCM4DD algoritmaları ile tek tek kümelenmiştir. Bu işlem 1000 kez tekrarlanarak her yöntem için bulunan küme merkezlerinin ortalamaları $(\bar{v}_1;\bar{v}_2)$, Tablo 2'de verilmiştir.

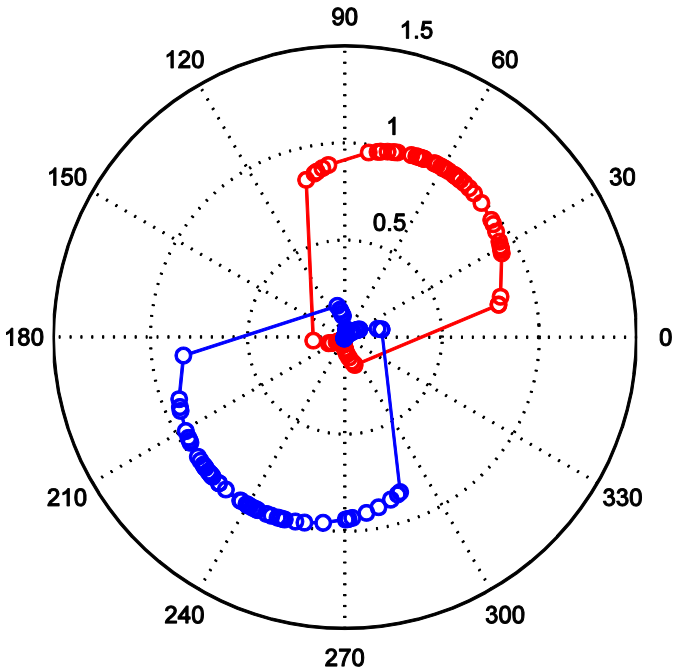
Tablo 2. Örnek uygulama 2 için kümeleme yöntemlerinin karşılaştırılması

Algoritmalar	EM	FCD	FCM4DD
$\bar{v}_1$(deg)	60.1391	62.2927	60.0469
$\bar{v}_2$(deg)	240.0490	240.17491	240.0006
MSE	0.00000662	0.00161054	0.00000067
ACT (sec)	0.03157753	0.09850940	0.00098954

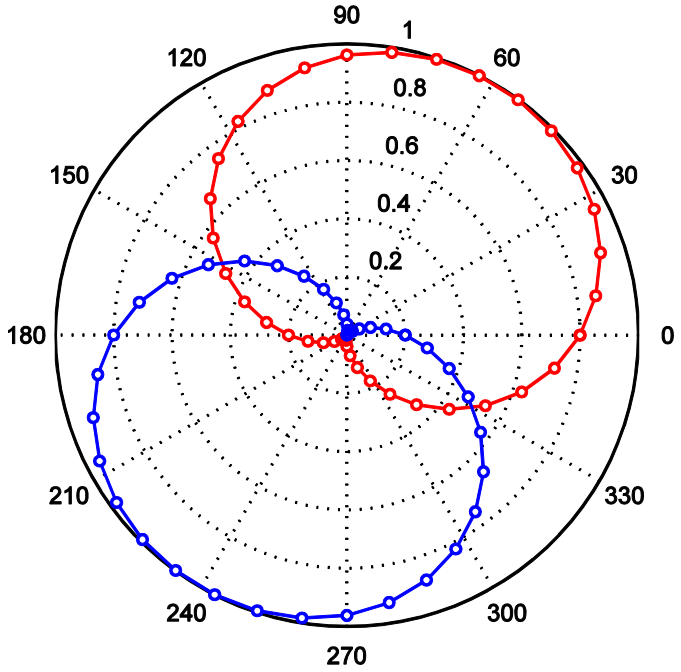
MSE değerlerinin karşılaştırılmasında FCM4DD algoritması diğer algoritmalara göre oldukça iyi başarı göstermesine rağmen birçok denemede EM algoritması ile yakın MSE değerlerinin elde edildiği gözlenmiştir. Önceki örnekte olduğu gibi hesaplama zamanı açısından FCM4DD algoritması üstünlüğünü korunmaktadır.



Şekil 14. Örnek uygulama 2'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 15. Örnek uygulama 2'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 16. Örnek uygulama 2'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi

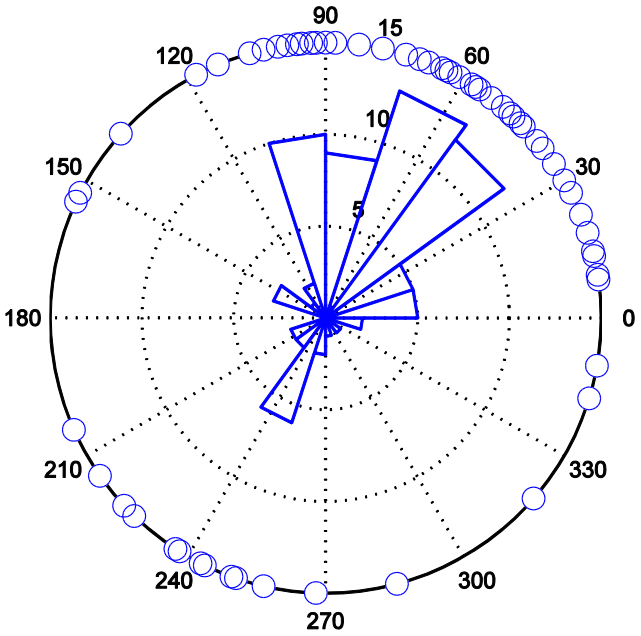
2.2.3. Örnek Uygulama 3

Bu örnekte, Stephens (Stephens, 1969) tarafından verilen 76 kaplumbağanın yumurtladıktan sonra gittikleri yönü gösteren gerçek veriler kullanılmıştır ve bu veriler Tablo 3'de verilmiştir.

Tablo 3. 76 kaplumbağanın yumurtladıktan sonraki hareket yönleri

8	9	13	13	14	18	22	27	30	34
38	38	40	44	45	47	48	48	48	48
50	53	56	57	58	58	61	63	64	64
64	65	65	68	70	73	78	78	78	83
83	88	88	88	90	92	92	93	95	96
98	100	103	106	113	118	138	153	153	155
204	215	223	226	237	238	243	244	250	251
257	268	285	319	343	350				

Örnek uygulama 3'deki verilerin rose diyagramı Şekil 17'de verilmiştir. Bu diyagrama göre verilerin iki küme olduğu ve küme merkezlerinin (60°, 240°) olduğu görülmektedir (Chang-Chien, Hung, & Yang, 2012).



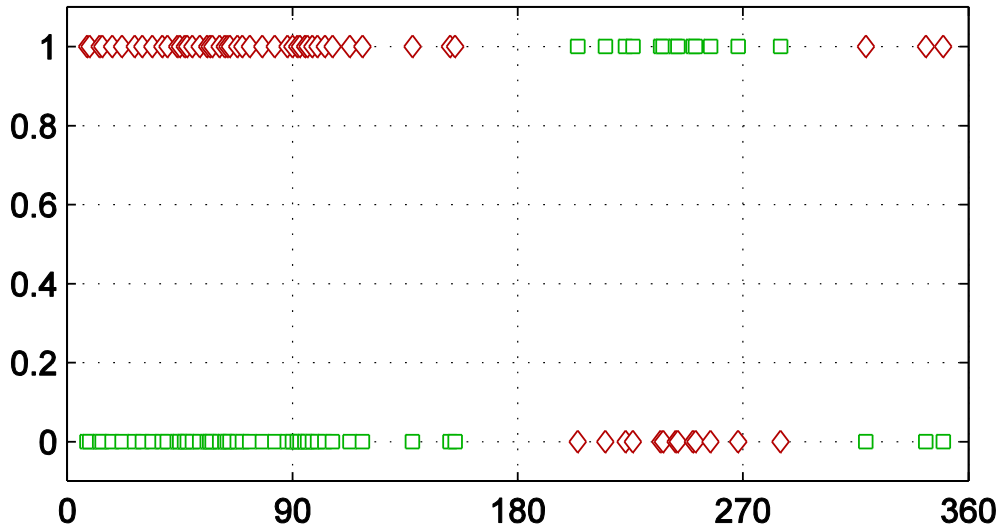
Şekil 17. 76 kaplumbağanın yönelimlerinin rose diyagramı

76 kaplumbağanın yönelimleri üç kümeleme yöntemi ile tek tek kümelenmiştir. Bu kümeleme işlemi 10 kez tekrarlanarak bulunan küme merkezleri ile yukarıda verilen küme merkezleri karşılaştırılarak yöntemlerin başarımları Tablo 4'de verilmiştir.

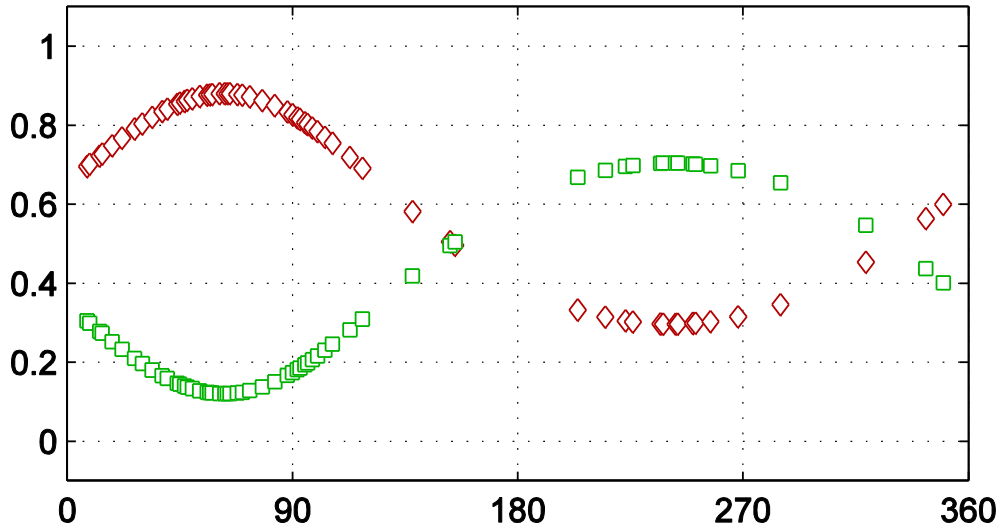
Tablo 4. Örnek uygulama 3 için kümeleme yöntemlerinin karşılaştırılması

Algoritmalar	EM	FCD	FCM4DD
$\bar{v}_1$(deg)	62.1523	63.1797	62.5729
$\bar{v}_2$(deg)	236.0323	238.5114	239.1145
MSE	0.00620652	0.00375486	0.00225540
ACT (sec)	0.69558950	0.15944514	0.00269273

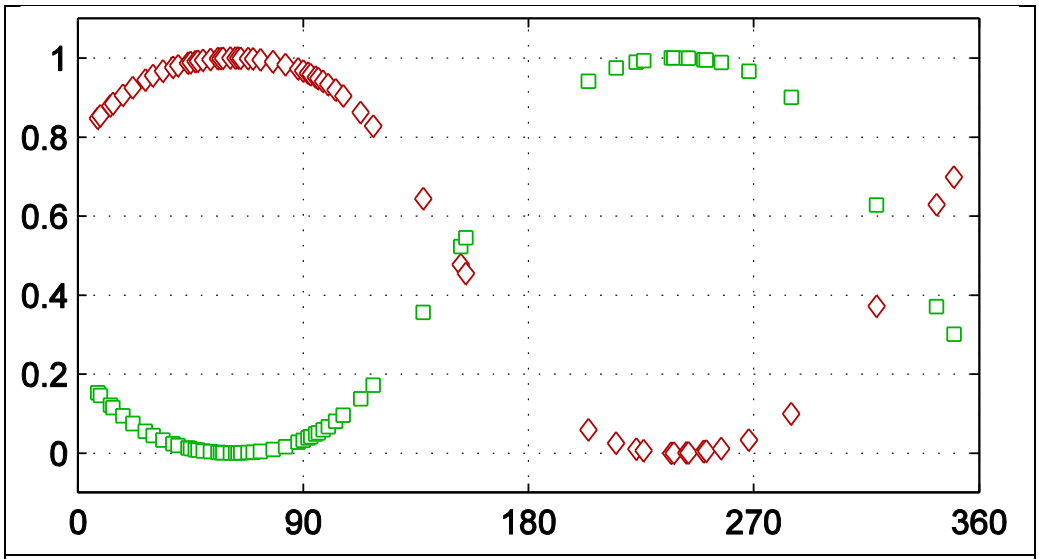
Verilen küme merkezlerine göre MSE ve ACT değerlerinde FCM4DD algoritması en iyi sonucu vermektedir. Kümeleme işleminin üyelik değerlerine göre karşılaştırılması Şekil (18-20)'de verilmiştir.



Şekil 18. 76 kaplumbağanın yönelimlerinin EM algoritması sonucunda elde edilen üyelik değerleri

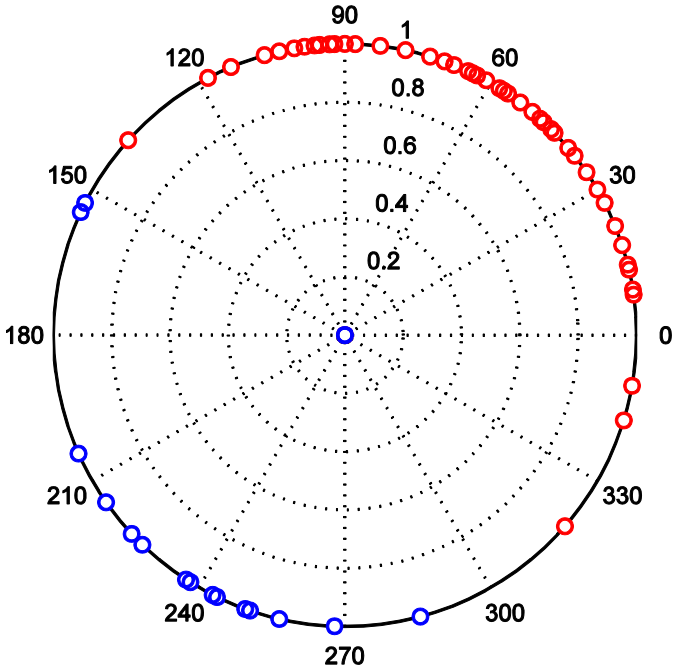


Şekil 19. 76 kaplumbağanın yönelimlerinin FCD algoritması sonucunda elde edilen üyelik değerleri

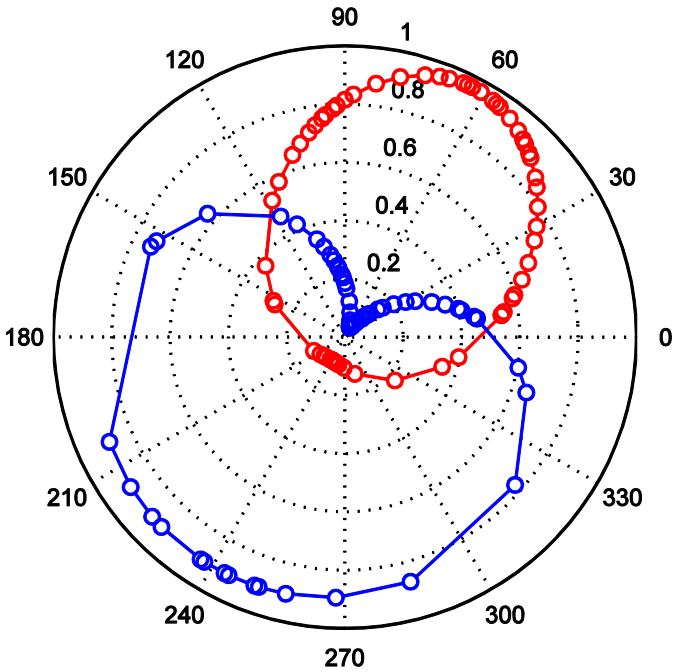


Şekil 20. 76 kaplumbağanın yönelimlerinin FCM4DD algoritması sonucunda elde edilen üyelik değerleri

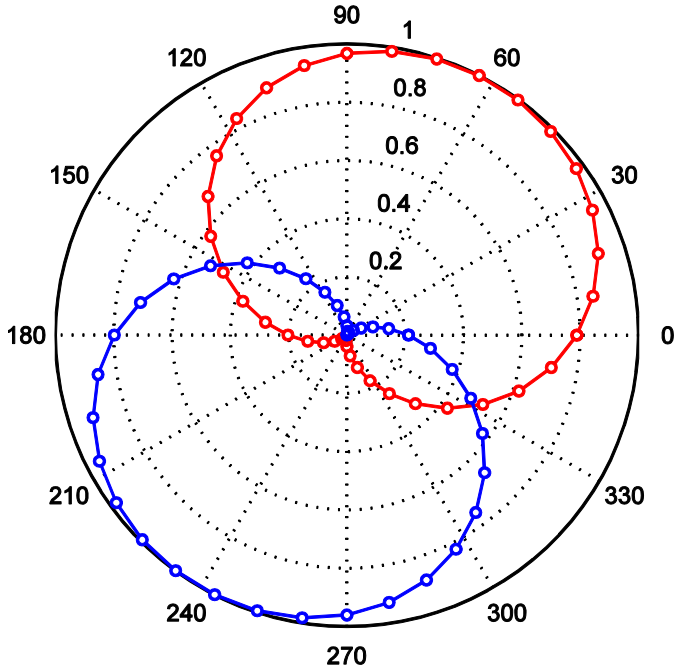
Maksimum üyelik değerlerine göre keskinleştirme yapılarak kaplumbağaların yönleri değerlendirilirken, FCD algoritması ile tüm denemelerde 61 kaplumbağa 60° yönünde 15 kaplumbağa 240° yönünde gittiği bulunmuştur. FCM4DD algoritması ile tüm denemelerde 59 kaplumbağa 60° yönünde 17 kaplumbağa 240° yönünde gittiği bulunmuştur. EM algoritmasında ise 4 denemede 60 kaplumbağa 60° yönünde 16 kaplumbağa 240° yönünde, 6 denemede ise 63 kaplumbağa 60° yönünde 13 kaplumbağa 240° yönünde gittiği bulunmuştur. Bu sonuçlara göre EM algoritmasının tutarlı olmadığı görülmektedir.



Şekil 21. Örnek uygulama 3'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 22. Örnek uygulama 3'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 23. Örnek uygulama 3'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi

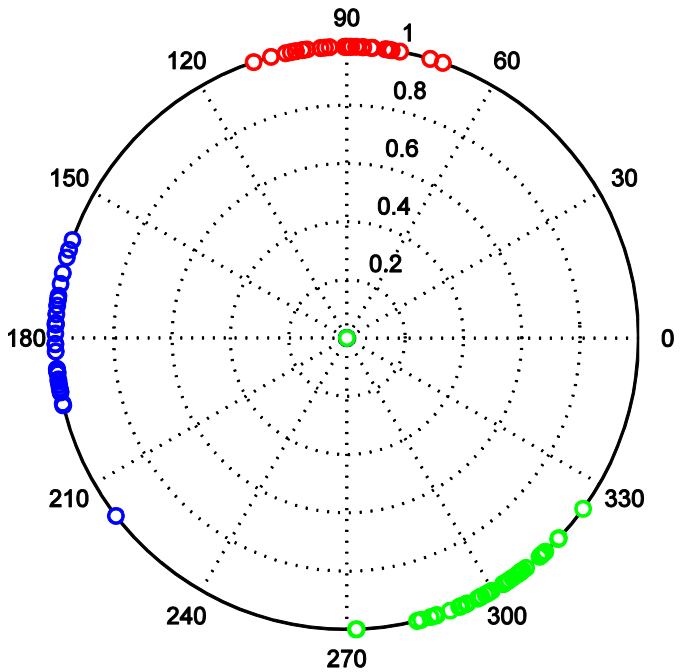
2.2.4. Örnek Uygulama 4

Yöntemlerin başarımlarını karşılaştırmak için $0.3VM(\pi/2, 25) + 0.3VM(\pi, 25) + 0.4VM(5\pi/3, 25)$ biçiminde verilmiş karışık von Mises dağılımından Best ve Fisher'in benzetim yöntemi (Best & Fisher, 1979) kullanılarak rastgele 100 örnek üretilmiştir. Bu örnekler EM, FCD ve FCM4DD algoritmaları ile tek tek kümelenmiştir. Bu işlem 1000 kez tekrarlanarak her yöntem için bulunan küme merkezlerinin ortalamaları $(\bar{v}_1; \bar{v}_2; \bar{v}_3)$, Tablo 5'de verilmiştir.

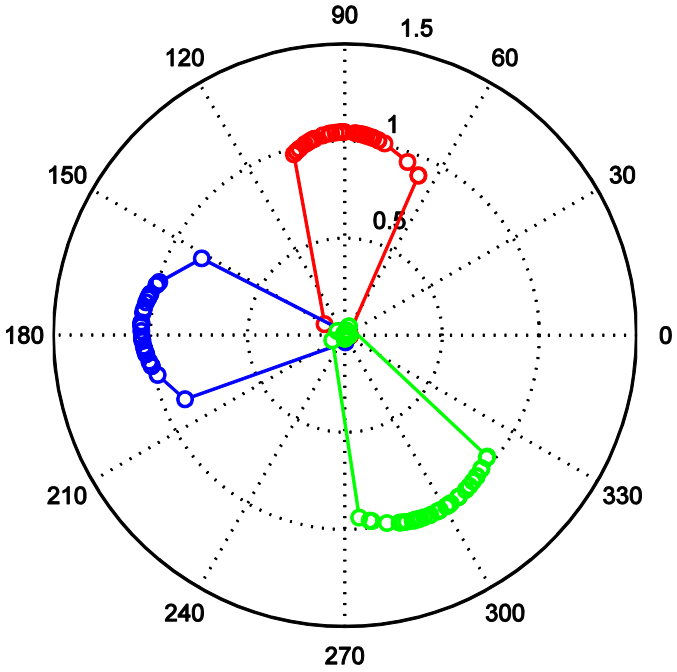
Tablo 5. Örnek uygulama 4 için kümeleme yöntemlerinin karşılaştırılması

Algoritmalar	EM	FCD	FCM4DD
$\bar{v}_1$(deg)	90.2071	89.0628	89.8338
$\bar{v}_2$(deg)	179.8445	180.3385	179.8775
$\bar{v}_3$(deg)	299.2960	300.8880	299.5070
MSE	0.00017139	0.00054269	0.00008702
ACT (sec)	0.45024185	0.71844227	0.00169275

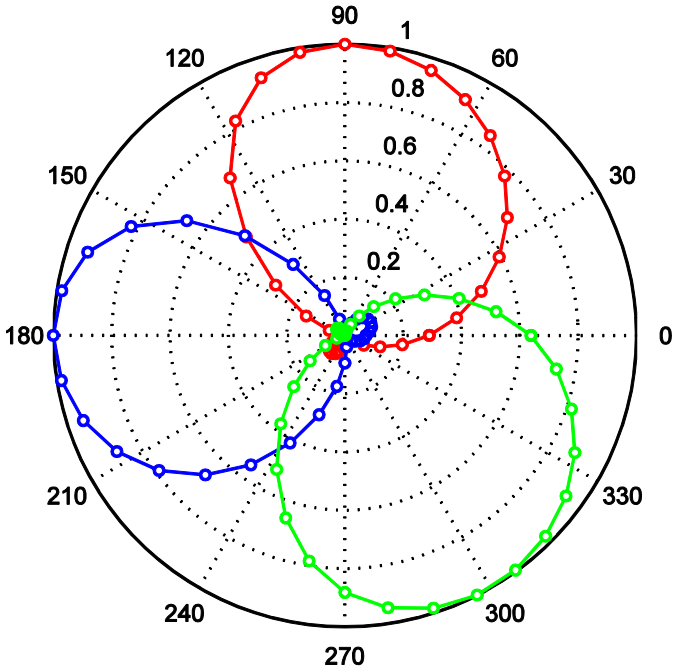
Bu örnekte, MSE değerlerinin karşılaştırılmasında FCM4DD algoritması diğer algoritmalara göre oldukça iyi başarı göstermektedir. Hesaplama zamanı açısından FCM4DD algoritmasının üstünlüğü korunmaktadır. Öte yandan EM algoritmasının ortalama hesaplama zamanına göre oldukça kötü sonuç vermesinin başlıca nedeni EM algoritmasının stabil bir yöntem olmamasından kaynaklanmaktadır.



Şekil 24. Örnek uygulama 4'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 25. Örnek uygulama 4'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 26. Örnek uygulama 4'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi

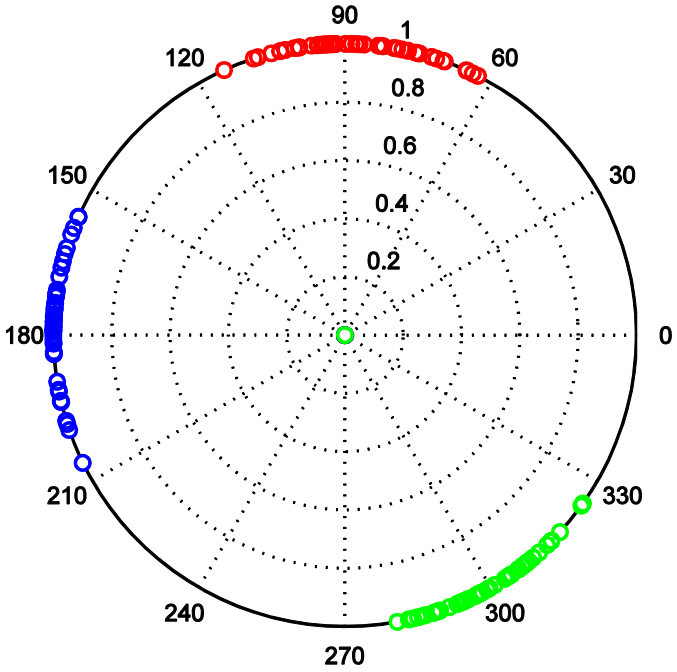
2.2.5. Örnek Uygulama 5

Bu örnekte ortalamaları $(\pi/2, \pi, 5\pi/3)$ olan üç üçgen dağılımın karışık olasılık yoğunluk fonksiyonu $0.34\Lambda\left(\frac{\pi}{3}, \frac{\pi}{2}, \frac{2\pi}{3}\right) + 0.33\Lambda\left(\frac{5\pi}{6}, \pi, \frac{7\pi}{6}\right) + 0.33\Lambda\left(\frac{9\pi}{6}, \frac{5\pi}{3}, \frac{11\pi}{6}\right)$ biçiminde verilmiş olsun. Bu dağılımdan rastgele 150 örnek oluşturulmuştur. Bu örnekler EM, FCD ve FCM4DD algoritmaları ile tek tek kümelenmiştir. Bu işlem 1000 kez tekrarlanarak her yöntem için bulunan küme merkezlerinin ortalamaları $(\bar{v}_1; \bar{v}_2; \bar{v}_3)$, Tablo 6'da verilmiştir.

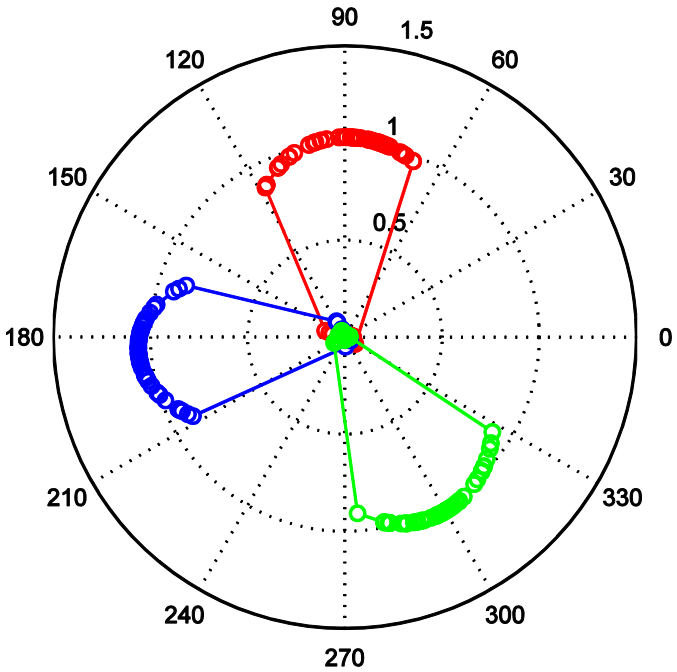
Tablo 6. Örnek uygulama 5 için kümeleme yöntemlerinin karşılaştırılması

Algoritmalar	EM	FCD	FCM4DD
$\bar{v}_1$(deg)	89.2385	90.0818	89.2115
$\bar{v}_2$(deg)	179.1986	179.1258	179.6384
$\bar{v}_3$(deg)	298.8556	298.6702	298.9577
MSE	0.00077120	0.00077349	0.00056014
ACT (sec)	1.01572538	0.63729468	0.00260909

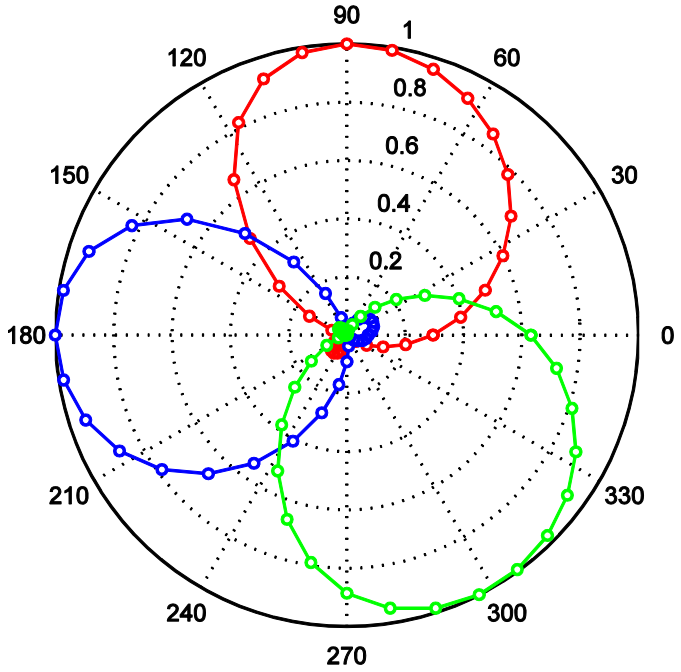
Bu örnekte, MSE değerlerinin karşılaştırılmasında FCD algoritması EM algoritmasıyla tüm denemelerde hemen hemen aynı sonuçları vermektedir. Hesaplama zamanı açısından FCM4DD algoritması üstünlüğünü korunmaktadır. Öte yandan EM algoritmasının ortalama hesaplama zamanına göre oldukça kötü sonuç vermesinin başlıca nedeni EM algoritmasının kararlı bir yöntem olmamasından kaynaklanmaktadır.



Şekil 27. Örnek uygulama 5'deki verilerin EM algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 28. Örnek uygulama 5'deki verilerin FCD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi



Şekil 29. Örnek uygulama 5'deki verilerin FCM4DD algoritması sonucunda elde edilen üyelik değerlerinin kutupsal koordinatta gösterimi

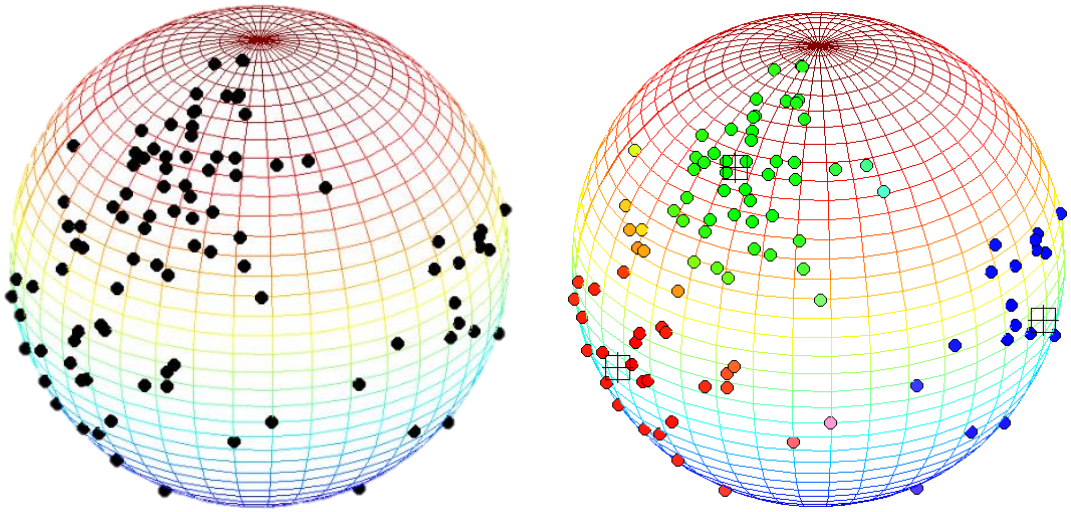
2.2.6. Örnek Uygulama 6

Bu örnekte, FCM4DD algoritmasının küresel veriler üzerindeki başarımını göstermek için ortalamaları ve standart sapmaları sırasıyla,

$$\mu = \left[\left(\frac{\pi}{4},\frac{\pi}{3}\right);\left(\frac{\pi}{2},\frac{\pi}{6}\right);\left(\frac{\pi}{2},\frac{3\pi}{4}\right)\right]^{T}$$

$$\sum = \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0.3 \end{bmatrix}$$

olan iki değişkenli normal dağılımdan $(0.5, 0.2, 0.3)$ oranlarında rastgele 100 örnek üretilmiştir ve Şekil. 30(a)'da gösterilmiştir. Şekil. 30(b)'de ise verilerin önerilen FCM4DD algoritması kullanılarak kümelere ayrılmış hali ve küme merkezleri gösterilmektedir. Ayrıca bulunan küme merkezlerinin ortalamaları Tablo 7'de verilmiştir.



(a) (b)

Şekil 30. Küresel verilerin benzetimlerindeki modellerden bir tanesinin gösterimi (a) Küresel verilerin gösterimi; (b) FCM4DD algoritması ile küresel verilerin kümelere ayrılmasının gösterimi

Tablo 7. Örnek Uygulama 6'nın sonuçları

Algoritma	FCM4DD
$\bar{v}_{1\theta}$, $\bar{v}_{1\varphi}$(deg)	(30.4065, 88.9690)
$\bar{v}_{2\theta}$, $\bar{v}_{2\varphi}$(deg)	(60.5684, 42.5454)
$\bar{v}_{3\theta}$, $\bar{v}_{3\varphi}$(deg)	(134.9254, 89.6836)
MSE	0.00039001
ACT (sec)	0.1630

Bu örnekte, Tablo 7'deki MSE değerine bakıldığında FCM4DD algoritmasının küresel veriler üzerinde de oldukça iyi sonuçlar verdiği görülmektedir.

3. BULGULAR VE SONUÇLAR

EM, FCD ve önerilen yöntem olan FCM4DD kümeleme algoritmaları başarımları çeşitli örneklerle test edilmiştir. EM algoritması yönsel veriler üzerinde oldukça iyi sonuçlar vermesine rağmen çoğu zaman kararlı olmamaktadır. Bu çalışmada EM algoritmasını kararlı olması için birçok kez çalıştırılarak çözüme gidilmiştir. Ancak bu durumda algoritmanın çalışma süresini artırmaktadır.

FCD algoritması, EM algoritmasına göre daha kararlı bir algoritmadır. Öte yandan FCD algoritması tarafından üretilen üyelik değerleri, FCM üyelik değerlerinin karakteristiğine göre farklılık göstermektedir. Örneğin FCM algoritmasında küme merkezindeki bir veri noktasının o kümeye olan üyelik değeri 1 iken diğer kümelere olan üyelik değerleri sıfır olmaktadır. Ancak FCD algoritmasında bu durum değişkenlik göstermektedir. Bir veri noktasının kümelere olan üyelik değerleri arasındaki farklılık küme merkezine yakın olan noktaların sayısıyla ilişkili olduğu görülmektedir.

Tüm yöntemlerin verilen örnekler üzerindeki uygulamalarında amaç fonksiyonları yerine hata kareleri ortalamasının karşılaştırmaları verilmiştir. Bunun birinci nedeni mevcut yöntemlerin amaç fonksiyonlarının verilmemiş olmaması, ikinci nedeni ise amaç değerleri hesaplanırken mevcut yöntemlere göre verilerin ortalamasının tam olarak hesaplanamamasıdır.

Bu çalışmada önerilen FCM4DD kümeleme algoritması klasik FCM algoritmasının yönsel verilere uygulanması için uyarlanmıştır. Önerilen yeni yöntem yönsel verilerin kümelenmesi için önerilen EM ve FCD yöntemlerin aksine dağılımdan bağımsız çalışmaktadır. Önerilen yöntem diğerlerine göre algoritma açısından basitlik, hesaplama zamanı açısından hızlılık ve doğrusal uzaklık açısından doğruluk gösteren bir yöntem olduğu uygulama örneklerinde gösterilmiştir. Yönsel veriler için bulanık c-ortalamalar kümeleme algoritması dairesel ve küresel verilerin yanı sıra N-boyutlu verilere de kolayca uygulanabileceği görülmektedir.

4. ÖNERİLER

Yönsel kümeleme algoritmaları değişik uygulama alanlarında sıklıkla kullanılmaktadır. Ancak doğruluk konusunda hassas olan birçok uygulamada mevcut yöntemler belli bir oranda hata yaptıkları için istenmeyen sonuçlar verebilmektedir. Oysa önerilen yöntemle oldukça yüksek orandaki doğruluk payının yanında kısa sürede hesaplama üstünlüğü sağlaması birçok alanda tercih edilen bir algoritma haline geleceği düşünülmektedir. Önerilen yöntem birçok farklı alana uygulanarak gerçek hayatta bir probleme sağlıklı çözümler üretebilir.

Literatürde sıklıkla dairesel ve küresel veriler üzerinde uygulanan kümeleme algoritmaları farklı biçimde uygulanmakta bu ise uygulamacının bir karmaşaya düşmesine neden olabilmektedir. Oysa önerilen yöntem aynı algoritmayla hem dairesel, hem de küresel verilere uygulanabilmektedir. Öte yandan hiperküresel (N-boyutlu) veriler içinde başka bir algoritma geliştirmeye gerek kalmaksızın uygulanabilmesi değişik alanlardaki uygulamacılara kolaylık sağlayabilmektedir.

Örnek uygulama 6'da önerilen FCM4DD algoritmasının küresel verilere uygulanabilirliği gösterilmiştir. Ancak küresel verilerin kümelenmesi için literatürde kullanılan yöntemlerle karşılaştırılması yapılmamıştır. FCM4DD algoritmasıyla literatürdeki yöntemlerin karşılaştırılması yapılarak başarımları test edilebilir.

5. KAYNAKLAR

Ackermann, H., 1997. A Note on Circular Nonparametrical Classification. *Biometrical Journal*, *39*,5, 577-587.

Alpaslan, F., Erilli, N. A., Yolcu, U., Eğrioğlu, E., ve Aladağ, Ç. H., 2011. Bulanık Kümelemede En Uygun Küme Sayısının Yapay Sinir Ağları ve Diskriminant Analizi İle Belirlenmesi. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 475-488.

Berens, P., 2009. CircStat: A MATLAB Toolbox for Circular Statistics. *Journal of Statistical Software*, *31*,10.

Best, D. J., ve Fisher, N. I., 1979. Efficient Simulation of the von Mises Distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*, 2, 152-157.

Bezdek, J. C., 1974. Numerical Taxonomy with Fuzzy Sets. *Journal of Mathematical Biology*, 57-71.

Bezdek, J. C., 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York: Plenum Press.

Bezdek, J. C., Ehrlich, R., ve Full, W., 1984. FCM: The Fuzzy c-Means Clustering Algorithm. *Computer & Geosciences*, *10*, 2-3, 191-203.

Bruzzone, L., ve Prieto, D. F., 2002. An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *11*, 452-466.

Chang-Chien, S. J., Hung, W. L., ve Yang, M. S., 2012. On Mean Shift-Based Clustering for Circular Data. *Soft Computing*, 1043-1060.

Chang-Chien, S. J., Yang, M. S., ve Hung, W. L., 2010. Mean shift-based clustering for directional data. *Proceedings of third international workshop on advanced computational intelligence*, (s.367–372).

Dempster, A. P., Laird, N. M., ve Rubin, D. B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1, 1-38.

Fisher, N. I., 1993. *Statistical Analysis of Circular Data.* Cambridge: Cambridge University Press.

Fukuyama, Y., ve Sugeno, M., 1989. A new method of choosing the number of clusters for the fuzzy c-means method. *Proceedings of the Fifth Fuzzy Systems Symposium*, (s. 247-250).

Gumbel, E. J., Greenwood, J. A., ve Durand, D., 1953. The Circular Normal Distribution: Theory and Tables. *Journal of the American Statistical Association*, *48*, 261, 131-152.

Günay Atbas, A. C., 2008. *Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma.* Yüksek Lisans Tezi, Ankara Üniversitesi, İstatistik Anabilim Dalı, Ankara.

Han, J., ve Kamber, M., 2006. *Data Mining Concepts and Techniques Second Edition.* Morgan Kaufmann Publishers Inc.

Höppner, F., Klawonn, F., Kruse, R., ve Runkler, T., 2000. *Fuzzy Cluster Analysis.* Chichester: John Wiley&Sons.

Işık, M., ve Çamurcu, A. Y., 2010. K-means ve Aşırı Küresel C-means Algoritmaları ile Belge Madenciliği. *Marmara Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 1-18.

Işık, Meltem; Çamurcu, A. Yılmaz., 2007. K-means, K-medoids ve Bulanık C-means Algoritmalarının Uygulamalı Olarak Performanslarının Tespiti. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 31-45.

Jammalamadaka, S. R., ve Gupta, A. S., 2001. *Topics in Circular Statistics.* London: World Scientific Publishing Co. Pte. Ltd.

Kaufman, L., ve Rousseeuw, P. J., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: John Wiley & Sons.

Kesemen, O., ve Tezel, Ö., 2013. Küme Sayısının Optimal Olarak Belirlenmesinde Yeni Bir Yaklaşım. *Uluslararası 8. İstatistik Kongresi.* Antalya.

Kim, D.-W., Lee, K. H., ve Lee, D., 2004. On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition*, 2009-2025.

Kwon, S. H., 1998. Cluster validity index for fuzzy clustering. *Electronics Letters*, 2176-2177.

Lund, U., 1999. Cluster analysis for directional data. *Communications in Statistics - Simulation and Computation*, *28*, 4, 1001-1009.

MacQueen, J., 1967. Some Methods for Clssification and Analysis of Multivariate Observations. *Proceeding of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1*, s. 281-297. University of California Press.

Mardia, K. V., 1972. *Statistics of Directional Data.* London, England: Academic Press.

Mardia, K. V., ve Jupp, P. E., 2000. *Directional Statistics.* New York: John Wiley & Sons, Inc.

McLachlan, G. J., ve Basford, K. E., 1988. Mixture Model: Inference and Applications to clustering. *Statistics: Textbooks and Monographs*.

Miller, H. J., ve Han, J., 2009. *Geographic Data Mining and Knowledge Discovery Second Edition.* Taylor and Francis.

Murat, Y. Ş., ve Şekerler, A., 2009. Trafik Kaza Verilerinin Kümeleme Analizi Yöntemi İle Modellenmesi. *Teknik Dergi/Technical Journal of Turkish Chamber of Civil Engineers*, 4759-4777.

Ortakcı, Y., ve Göloğlu, C., 2012. Parçacık Sürü Optimizasyonu İle Küme Sayısının Belirlenmesi. *XIV. Akademik Bilişim Konferansı Bildirileri.* Uşak.

Pang-Ning, T., Steinbach, M., ve Kumar, V., 2006. *Introduction to Data Mining.* Library of Congress.

Peker , K. Ö., ve Bacanlı, S., 2004. Dairesel Verilere Uygulanan Tanımlayıcı istatistiksel Yöntemler ve Meteorolojik Bir Uygulama. *Anadolu University Journal of Science and Technology, 5*, 1, s. 115-122.

Servi, T., 2009. *Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi.* Doktora Tezi, Çukurova Üniversitesi, İstatistik Bölümü, Adana.

Stephens, M. A., 1969. *Techniques for directional data Tech. Report #150.* Stanford: Department of Statistics, Stanford University.

Upton, G. J., ve Fingleton, B., 1989. *Spatial data analysis by example vol 2.* Chicester, England: John Wiley & Sons Ltd.

Von Mises, R., 1918. Uber die die "Ganzzahligkeit" der Atomgewicht und verwandte Fragen,. *Physikal, 19*, 490-500.

Watson, G. S., ve Williams, E. j., 1956. On the Construction of Significance Tests on the Circle and the Sphere. *Biometrika*, *43*, 344-352.

Xie, X. L., ve Beni, G., 1991. A validity measure for fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 841-847.

Yang, M. S., ve Pan, J. A., 1997. On fuzzy clustering of directional data. *Fuzzy Set and Systems*, *91*, 319-326.

Zar, J. H., 1999. *Biostatistical Analysis 4th edition.* Prentice Hill.

ÖZGEÇMİŞ

Özge TEZEL, 9 Ağustos 1990 tarihinde Trabzon'da doğdu. İlköğrenimini Yavuz Selim İlköğretim Okulu'nda, ortaöğrenimini ise Trabzon (YDA) Lisesi'nde tamamladı. 2008 yılında Karadeniz Teknik Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik ve Bilgisayar Bilimleri Bölümüne yerleşti ve 2012 yılında bu bölümden mezun oldu. Aynı yıl Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik ve Bilgisayar Bilimleri Anabilim dalında tezli yüksek lisans programına başladı. Ulusal ve uluslararası birçok sempozyum ve kongrede yayınlanmış bildirileri bulunmaktadır. Kasım 2014 tarihinde Karadeniz Teknik Üniversitesi İstatistik ve Bilgisayar Bilimleri Bölümü'ne Araştırma Görevlisi olarak atandı ve halen bu görevine devam etmektedir.