Araştırma Makalesi

BMIJ

ISSN: 2148-2586

BUSINESS & MANAGEMENT STUDIES:
AN INTERNATIONAL JOURNAL
Vol.:7 Issue:2 Year:2019, pp. 1117-1135

Citation: Özarı, Ç., Eren, Ö. & Alıcı, A. (2019), K-Ortalamalar Yönteminin Başlangıç Merkez Seçim Sorunsalı Üzerine Bir Çalışma, BMIJ, (2019), 7(2): 1117-1135 doi: http://dx.doi.org/10.15295/bmij.v7i2.1124

K-ORTALAMALAR YÖNTEMİNİN BAŞLANGIÇ MERKEZ SEÇİM SORUNSALI ÜZERİNE BİR ÇALIŞMA

Çiğdem ÖZARI[1]
Özge EREN [2]
Agah ALICI[3]

Received (Başvuru Tarihi): 07/05/2019
Accepted (Kabul Tarihi): 21/06/2019
Published Date (Yayın Tarihi): 26/06/2019

ÖZ

K-ortalamalar kümeleme yöntemi, belirli bir veri kümesindeki birimleri önceden belirlenmiş sayıda kümeye ayıran en basit, denetimsiz öğrenme algoritmalarından biridir. Bu yöntem diğer iteratif yöntemler gibi başlangıç olarak seçilen ve küme merkezi olarak atanan değer veya değerlere bağlı kalarak bir kümeleme gerçekleştirir. K-ortalamalar yönteminde; ilk adımı rastlantısal olarak seçilen küme merkezleri yardımıyla, veri kümesindeki tüm birimlerin bu merkez noktalara olan uzaklıkları dikkate alınarak, birimlerin ait olduğu kümeler belirlenir. Bu rastlantısal olarak seçilen küme merkezleri farklı küme yapıları oluşturabilmektedir. Bu çalışma da başlangıç küme merkezi seçim sorunsalının varlığının daha detaylı anlaşılması adına, sorunsalın gözlemlendiği bir kurgu çalışma oluşturulmuştur. Kurgu çalışmada birimlerin iki ve üç kümeye ayrışmak istendiği durum için, veriler öncelikli olarak veri setinde yer alan tüm olası başlangıç merkez verilerle k-ortalamalar kümeleme yöntemi uygulanarak ayrıştırılmış ve farklı küme yapılarının farklı sıklıklarla elde edildiği gözlemlenmiştir. Ayrıca sorunsalın varlığını daha detaylı incelemek adına, veri setine yakın ve uzak konumlarda olacak şekilde veri setinde yer almayan yeni birimler oluşturabilmek için bir yöntem geliştirilmiştir. Daha sonra yöntemle elde edilen yeni birimler, başlangıç merkez veri olarak ele alınarak, veri seti kümelere ayrılmış ve daha önce elde edilmeyen yeni küme yapıları gözlemlenmiştir. Çalışmanın son kısmında ise başka bir kurgu çalışma ile veri seti içinden veya veri seti dışından seçilen başlangıç merkez birimlerle farklı sonuçlar elde edilebileceği gösterilmiştir.

Anahtar kelimeler: *K-ortalamalar Kümeleme Yöntemi, Kümeleme Analizi, Başlangıç Merkez Noktası.*

Jel Kodları: *C38, C15*

A CRITICAL OVERVIEW OF THE INITIAL CENTER SELECTION OF K-MEAN CLUSTERING ALGORITHM

ABSTRACT

The K-means clustering method is one of the simplest, unsupervised learning algorithms that divides the units of a given data set into a predetermined number of distinct clusters. This method, like other iterative methods, performs a cluster analysis based on initial center points which are randomly chosen. With the help of these initial center points, clusters belonging to similar data sets are determined and these randomly selected initial points may lead biased results. In addition, determining which of the results obtained from different initial centers is

[1] İstanbul Aydın Üniversitesi, İİBF, Ekonomi ve Finans Bölümü, cigdemozari@aydin.edu.tr https://orcid.org/0000-0002-2948-8957
[2] İstanbul Aydın Üniversitesi, Sağlık Kurumları İşletmeciliği, ozgeeren@aydin.edu.tr http://orcid.org/0000-0002-5421-363X
[3] İstanbul Aydın Üniversitesi Bilgi İşlem Daire Başkanlığı, agah@aydin.edu.tr https://orcid.org/0000-0002-3151-6814

more valid is another main and important problem of K-mean cluster algorithm. To understand the existence of the initial center problem of K-mean clustering method, a fictitious study has been created. In the fictitious study, to determine and show the existence of the problem, we decided to partition the data set into two and three clusters with all possible initial centers from the data set. Since initial centers can get values from anywhere, we developed a simple algorithm to construct new initial centers, which are out of the data set. The new initial centers constructed are so near to units, which belongs to the data set, and the others are far away. In the second part of the fictitious study, we cluster the same data set with new (progressed) initial centers and examine the results from this analysis and we found different and new cluster sets which we could not construct with initial centers from the data set. In addition, we aimed to show there will be some different cluster groups, when we start the method with initial centers from the data-set and with initial centers from outside the data-set or with initial center points combining inside and outside.

Keywords: *K-means Clustering, Cluster Analysis, Initial Cluster Centers*

Jel Codes: *C38, C15*

1. GİRİŞ

K-ortalamalar kümeleme yöntemi, 1967 yılında Mac Queen tarafından önerilen belirli bir veri kümesindeki birimleri (gözlemleri veya nesneleri) önceden belirlenmiş sayıda (k tane) kümeye ayıran en basit, denetimsiz öğrenme algoritmalarından biridir. Bu yöntem diğer iteratif yöntemler gibi başlangıç olarak seçilen ve küme merkezi olarak atanan değer veya değerlere bağlı kalarak bir kümeleme gerçekleştirir. Esas problem de bu nokta da başlar. Çünkü K-ortalamalar yöntemindeki ilk adım rastlantısal olarak seçilen küme merkezleri yardımıyla, veri kümesindeki tüm birimlerin bu merkez noktalara olan uzaklıklarını dikkate alarak birimlerin ait olduğu kümeleri belirler. Bu rastlantısal olarak seçilen küme merkezleri farklı bir açıdan taraflı sonuçlar doğurabilir. Bu değişen sonuçların hangisinin daha geçerli olduğunun belirlenmesi ise başka bir sorunsal olarak ortaya çıkar.

Araştırmacının başlangıçta karar vermesi gereken ayrıştırmak istediği küme sayısı aslında rastlantısal olarak seçilen küme merkezi olarak düşünülen veri sayısı ile aynıdır. Başlangıç küme merkez seçim sorunsalının varlığının daha detaylı anlaşılması adına, bu çalışmada sorunsalın gözlemlendiği bir kurgu çalışma oluşturulmuştur. Bunlara ek olarak, başlangıç merkez seçiminde kullanılmak üzere veri setinde yer almayan birimler oluşturmak için kolay uygulanabilir bir yöntem geliştirilmiştir. Bu yöntem ile veri seti dışında oluşan birimler, diğer birimlere yakın (yoğunlaştığı yerlerde) ve uzak bölgelerde konumlanmaktadır. Ayrıca çalışmada veri seti içinden veya önerilen yöntemle en az veri sayısı kadar elde edilen veri seti dışından seçilen başlangıç merkez birimlerle, farklı sonuçlar elde edilebileceği gösterilmiştir.

Kümeleme analizi benzer özellik gösteren birimlerin (verilerin) kendi aralarında kümelere ayrılmasıdır. Böylelikle benzemeyen birimler de belirlenmiş olur. Bu analiz ile k adet özelliğe sahip n sayıda birimin, benzerliklerine göre türdeş yapının sağladığı ayrık kümelerde toplanması amaçlanır (Duran ve Odel, 1974). Kümeleme analizi bir açıdan da alt grupların sayısı bilinmediği ve birimler hakkında önemli bilgilerin olmadığı durumlarda bile birimleri anlamlı alt gruplara ayırma yöntemi olarak da düşünülebilir (Fraley ve Raftery, 1998). Kümeleme analizini gerçekleştirmek için birçok kümeleme yöntemi (algoritması) geliştirilmiştir (Han ve Kamber, 2006; Jain ve Dubes, 1988; Mercer, 2003; Witten ve Frank, 1999). Bu analizde diğer çok değişkenli istatistik analizlerinde olduğu gibi verilerin normalliği varsayımı fazla önemli olmayıp uzaklık değerlerinin normalliği yeterli görülür (Tatlıdil, 1992). Bu çalışmada önemli bir kümeleme tekniği olan K-ortalamalar kümeleme yöntemine ait merkez seçim sorunsalı ele alınacaktır.

1967 yılında Mac Queen tarafından geliştirilen K-ortalamalar kümeleme algoritması; basit uygulanması ve hızlı çalışması nedeniyle yaygın olarak bilinen ve sıklıkla tercih edilen bir algoritmadır. Algoritmanın genel mantığı; n adet veri nesnesinden oluşan bir veri kümesini, giriş değeri olarak verilen k adet kümeye bölümlemektir (Çalışkan ve Soğukpınar, 2008). Bu yöntem diğer iteratif yöntemler gibi başlangıç olarak seçilen ve küme merkezi olarak atanan değer veya değerlere bağlı kalarak bir kümeleme gerçekleştirir. Yöntemin ana amacı gerçekleştirilen analiz sonucunda elde edilen kümelerin, küme içi benzerliklerinin en yüksek ve kümeler arası benzerliklerinin de en düşük olmasını sağlamaktır. Bu analizin diğer analizlerden farkı, kümelenmenin önceden tanımlanmış sınıflara (koşullara) dayanmamasıdır (Hajizadeh ve diğerleri, 2010).

Meila ve Heckerman'ın (2013) çalışmalarında belirttiği üzere K-ortalamalar yönteminin basamakları herkes tarafından genel kabul görmüş bir yöntem değildir. Genel kabul görmemiş olmasının sebeplerinden birinin, farklı küme merkez seçimlerinin yarattığı farklı sonuçlar (farklı kümeler) olabileceğidir. Steinley ve diğerleri (2007) tarafından özellikle yöntemin başlangıç küme merkezi seçiminin kümelenme şekline etkisi vurgulanıp, literatürde konu ile ilgili 12 çalışmanın bulguları karşılaştırılarak, bazı uygulamalar için öneri getirilmiştir.

Higgs ve diğerleri (1997) çalışmalarında orijinal veri tabanının bir alt kümesini oluşturarak, ilk merkezleri seçmek için bir "MaxMin" algoritmasının kullanılmasını

önermişlerdir. Bir başka çalışmada ise K-ortalamalar yönteminde tekrar eden iterasyonların bilgi depolaması yapılarak, bu depolanan bilgilerin bir sonraki iterasyonda kullanılmasını içeren bir yöntem önerilmiştir (Na ve diğerleri, 2010). Yapılan deneysel çalışmalarla da önerilen yöntemin kümeleme işleminin hızını arttırdığı ve bir açıdan da karmaşık hesaplama süreçlerinin atlanarak etkinliğin arttırıldığını göstermiştir. Khan ve Ahmad da çalışmalarında (2013) küme merkez seçim sorunsalına yeni bir yöntem önermişlerdir. Bu çalışmanın temel vurgusu K-ortalamalar yönteminde rastlantısal olarak seçilen ilk küme merkezlerinin rastlantısal olarak seçilmemesi gerektiği üzerinedir. İlgili çalışmada önerilen yöntemde ilk olarak veri kümesindeki birimlerinin uç değerlerinin belirlenmesi, var ise verinin bu uç değerlerden ayrıştırılmasıdır. Sürekli veri setleri için önerilen algoritmada, her bir boyut için merkez seçilmesi gerektiğinden, öncelikle veri setinin her bir boyut için normal dağılıp dağılmadığı kontrol edilerek, normal dağılım yoksa veri seti normalize edilir. Veri setindeki benzer yapıları her bir boyut için inceleyerek en çok tekrar eden yapılar yardımıyla kümelerin merkezleri belirlenir. İlgili çalışmada önerilen algoritmayı farklı veri çeşitleri için değerlendirerek daha uygun ve efektif olduğu sonucuna ulaşmışlardır.

Ayrıca bu analiz yöntemi ile biyoinformatik, örüntü tanıma, görüntü işleme, pazarlama, veri madenciliği, ekonomi gibi birçok alanda araştırma gerçekleştirilebilir (Yedla ve diğerleri, 2010). Örneğin Yaraş (2005) çalışmasında tüketicilerin pazarlama karması kararlarını algılamaları ve marka değeri algılamaları bakımından farklı pazar bölümleri oluşturup oluşturmadıklarını görebilmek adına kümeleme analizi, kümeleme analizinden elde ettiği bulguların tutarlılığını ölçmek için ise varyans analizi uygulamıştır. Zırhlıoğlu ve Karaca (2006) K-ortalamalar kümeleme yöntemini kullanarak 20 kişiden oluşan farklı 4 voleybol takım sporcularını fiziksel ve teknik özelliklere göre 3 farklı kümeye ayrıştırarak, kümeler arası farklılıkları belirlemeyi hedeflemişlerdir. Durucasu ve diğerleri (2006) kümeleme analizi uygulayarak Anadolu Üniversitesi İktisadi İdari Bilimler Fakültesi İşletme bölümünde açılan yaz okuluna katılan öğrencilerin genel profilini belirlemişlerdir. Atalay ve Tortum (2010) çalışmalarında; öncelikle 1997-2006 dönemi için Türkiye'deki illerde meydana gelen şehir dışı trafik kaza verilerini kullanarak, her il için ölüm ve yaralanma oranlarını hesaplamışlar, elde ettikleri oranları dikkate alarak K-ortalamalar ve Bulanık C-ortalamalar yöntemlerini kullanarak kümeleme analizi gerçekleştirmiştir. Analizlerden elde edilen en önemli sonuçlardan biri Bulanık C-ortalamalar yönteminin en az K-ortalamalar yöntemi kadar tutarlı sonuçlar verdiğidir. Fırat ve diğerleri (2012) çalışmalarında; K-ortalamalar yöntemini

uygulayarak yıllık toplam yağışların kümelenmesi ve homojen olan bölgelerin belirlenmesini amaçlayan bir uygulama çalışması gerçekleştirmişlerdir. Özdemir ve Orçanlı (2012) çalışmalarında; İki Aşamalı kümeleme algoritması kullanılarak demografik ve sosyokültürel özelliklere ait değişkenlerin verileri ile pazarların bölümlere ayrılması ve hedef pazar olabilecek pazar nişlerinin tespit edilmesine yönelik bir örnek uygulama gerçekleştirmişlerdir. Aydın ve Seven (2015) çalışmalarında; Türkiye'deki İl Nüfus ve Vatandaşlık Müdürlüklerini iş yoğunluklarına göre Hibrid Hiyerarşik K-ortalamalar kümeleme analizi ile sınıflandırılmışlardır. Akçapınar ve diğerleri (2016) ise K-ortalamalar yöntemini uygulayarak, eğitim alanında çevrimiçi öğrenme ortamında benzer davranış sergileyen öğrencileri, kümelere ayrıştırmayı hedeflemişlerdir. Ayrıştırılan kümeleri isimlendirmek bir başka ifade ile özelliklerini belirleyebilmek için öğrencilerin akademik performansları da incelenmiştir. Selvi ve Çağlar (2017) çalışmalarında kümeleme analiz yöntemlerinden K-ortalamalar, K-temsilci ve Birleştirici Hiyerarşik kümeleme yöntemlerini ele alarak, Türkiye'deki üç ayrı yıla ait trafik kaza verilerinden sınıflar oluşturarak üretilen çok değişkenli haritalar yardımıyla bu yöntemlerin karşılaştırılmasını gerçekleştirmişlerdir. Yalçın ve Ayyıldız (2018) Türkiye'de faaliyet gösteren 55 havalimanını K-ortalamalar yöntemiyle kullanan yolcu, havalimanlarından taşınan yük, uçuş sayısı gibi temel özelliklerini dikkate alarak kümelere ayrıştırmıştır. Bülbül ve Camkıran (2018) çalışmalarında; 2015 yılına ait sermaye yeterliliği oranlarına göre, Ward, K-ortalamalar ve Bulanık C-ortalamalar yöntemlerini kullanarak 46 bankayı kümelere ayırmışlardır. Elde edilen bulgular ise her yöntem için benzer yapıda 3 kümedir. Bulgularda yer alan küme yapıları incelendiğinde ise kümelerin sermaye kaynağı bakımından heterojen bir yapıda olduğu söylenir. Akay (2019) çalışmasının ilk kısmında panel veri kümeleme analizi yaparak, 2008-2017 yılları arasında Türkiye illeri için kütüphane kullanımını etkileyen faktörleri ele alarak kütüphane kullanımında benzer özellik gösteren illeri belirlemiştir. İkinci kısmında ise bu illerin insani gelişmişlikleri arasındaki ilişkiyi incelemiştir.

Literatür incelendiğinde çoğunlukla sosyal bilim alanlarında uygulanan K-ortalamalar yönteminin kullanıldığı uygulamalı çalışmalarda, küme merkezlerinin başlangıç seçimi ile ilgili sorunsal incelenmeden, yöntemden elde edilen bulguların değerlendirildiği görülmüştür. Oysa küme merkezlerinin başlangıç seçimi için çeşitli yöntemler önerilmiştir. Forgy yaklaşımı bu yöntemlerden ilki olup, veri setinden rastgele başlangıç merkezi seçimi yapmayı önerir.

Çınaroğlu ve Bulut (2018) çalışmalarında verileri kümelemek için standart K-ortalamalar ve Parçacık Sürü Optimizasyonu tabanlı kümeleme algoritmaları için başlangıç küme merkezlerinin seçimine yönelik iki yeni yöntem önermişlerdir. Yöntemlerden birinde; küme içindeki benzerliğin maksimum ve aynı zamanda kümeler arasındaki farklılığın maksimum olması amaçlandığından başlangıçta küme merkezlerinin birbirinden olabildiğince uzak seçilmesini hedeflemişlerdir ve bu hedef doğrultusunda geliştirdikleri yöntemlerdeki deneysel çalışmalarda, geliştirilen yaklaşımların öznitelik seçimi yapılmış normalize veri setleri üzerinde başarılı sonuçlar verdiği gözlemlenmiştir.

2. K-ORTALAMALAR KÜMELEME YÖNTEMİ VE ALGORİTMANIN YAPISI

K-ortalamalar yöntemi, n adet veriyi k adet kümeye ayrıştırmak için, ilk önce n adet veriden rastlantısal olarak k adet veri seçer (k<n). Bu verilerin her biri, bir kümenin merkezini veya orta noktasını temsil eder. Tüm veriler, seçilen k adet merkeze olan uzaklıkları dikkate alınarak en yakın olan kümeye dağılır. Ardından her küme için ortalama hesaplanır ve hesaplanan bu değer o kümenin yeni merkezi olur (Han ve Kamber, 2006). Bu yöntem merkez noktanın kümeyi temsil etmesi ana fikrine dayalıdır (Han ve Kamber, 2006) ve aslında aynı zamanda eşit büyüklükte küresel kümeleri bulmaya eğilimlidir (Işık ve Çamurcu, 2007). Kümeler içi benzerlikleri en yüksekte tutup, kümeler arası benzerlikleri en düşükte tutmayı hedefler. Küme benzerliği ise kümedeki verilerin ortalama değeri ile ölçülür, bu da kümenin ağırlık merkezidir (Xu ve Wunsch, 2005). Yeni merkezler belirlendikten sonra aynı işlem tüm veriler kümelere yerleşinceye kadar, bir başka ifade ile maksimum yineleme sayısına ulaşılana kadar veya bir yineleme ile bir sonraki arasında herhangi bir değişim olmayıncaya kadar devam eder. Bir başka ifade ile algoritmanın adımları aşağıdaki gibidir (Karypis ve diğerleri, 2000).

Adım 1: k değerini belirle (ayrıştırılmak istenen küme sayısı)

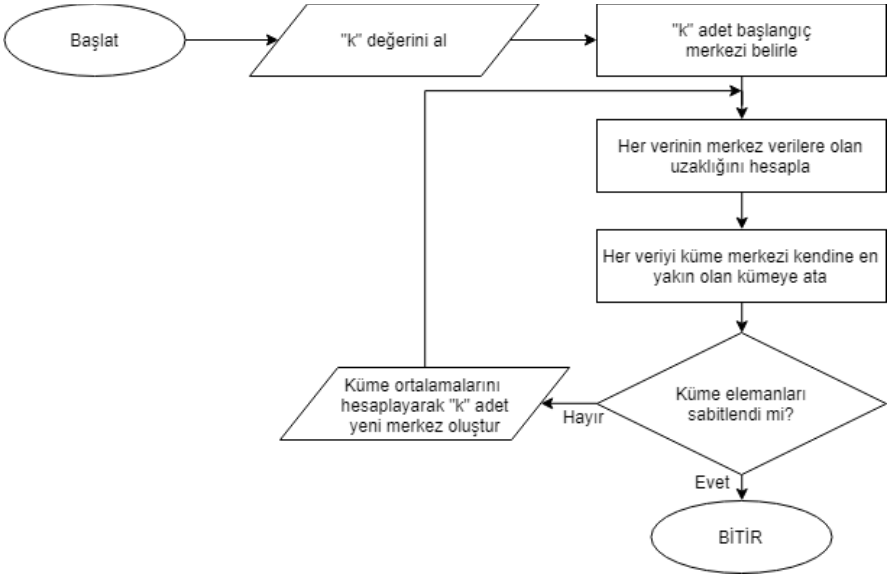
Adım 2: Başlangıç küme merkezleri olarak k veri seç,

Adım 3: Her veriyi, küme merkezi kendine en yakın olan kümeye ata,

Adım 4: Tüm veriler atandığında, küme ortalamalarını hesaplayarak, k adet merkezi yeniden hesapla,

Adım 5: Adım 3 ve Adım 4'ü küme elemanları sabitleninceye kadar tekrarla.

K-ortalamalar kümeleme yönteminin uygulama akış diyagramı ise Şekil 1'de sunulmuştur.



Şekil 1. K-Ortalamalar Kümeleme Yöntemi Akış Diyagramı

3. KURGU ÇALIŞMA

K-ortalamalar kümeleme yönteminin merkez seçim sorunsalının önemini vurgulamak ve varlığını göstermek üzere rastlantısal olarak oluşturulan ve 4 farklı boyutta (X, Y, Z, T) incelenen 10 veri Tablo 1'de sunulmuştur.

Tablo 1. İncelenen Veriler

Gözlem	X	Y	Z	T	Gözlem	X	Y	Z	T
0	63	43	62	30	5	16	25	97	98
1	62	22	96	15	6	59	77	72	7
2	18	5	46	67	7	37	40	28	85
3	24	59	62	81	8	60	88	31	97
4	98	30	29	99	9	18	69	83	2

Başlangıç merkez seçim sorunsalının varlığını göstermek adına, öncelikle K-ortalamalar yönteminin ilk adımı olan k değerinin belirlendiğini ve bu değerin 2 olduğunu varsayalım. Bir başka ifade ile veri setinin 2 kümeye ayrıştırılmak istendiği durumu inceleyelim. 2 kümeye ayrıştırılmak istendiği için 2 merkez verinin seçilmesi gerekir. Bu merkez veriler rastlantısal olarak seçildiğinden, ikisinin de veri seti içinden, ikisinin de veri seti dışından, birinin veri seti içinden diğerinin de veri seti dışından seçilmiş olduğu 3 farklı durumun da incelenmesi gerekir. Kurgu çalışmanın ilk kısmında farklı kümeler oluşabileceğinin gözlemlenmesi adına, öncelikle gözlemlenen verilerden seçilen başlangıç merkez verileriyle kümeleme analizi gerçekleştirilerek olası tüm durumlar değerlendirildi.

Tablo 2'de veri seti içinden seçilen başlangıç merkez verileriyle elde edilen tüm kümeler ve bu kümelerin görülme sıklıkları yer almaktadır.

Tablo 2. Veri Seti İçinden Seçilen Merkezlere Göre Oluşan Kümeler

	Küme 1	Küme 2	Başlangıç Merkez Veriler	Sıklık
A	0, 1, 6, 9	2, 3, 4, 5, 7, 8	(0,1) (0,3) (0,6) (0,7) (0,9) (1,2) (1,3) (1,5) (1,6) (1,7) (1,9) (2,3) (2,6) (2,9) (3,6) (3,7) (3,9) (4,6) (4,9) (5,6) (5,9) (6,7) (6,8) (7,9) (8,9)	25
B	0, 1, 2, 3, 5, 6, 9	4, 7, 8	(0,4) (2,4) (3,4) (3,8) (4,5) (4,7) (7,8)	7
C	0, 1, 2, 5, 6, 9	3, 4, 7, 8	(0,8) (1,8) (1,4) (2,8)	4
D	0, 1, 4, 6, 8, 9	2, 3, 5, 7	(0,2) (0,5) (2,5)	3
E	0, 3, 4, 6, 7, 8	1, 2, 5, 9	(2,7)	1
F	0, 1, 2, 3, 4, 6, 7, 8, 9	5	(3,5)	1
G	0, 2, 3, 4, 6, 7, 8, 9	1, 5	(5,7)	1
H	0, 4, 6, 7, 8	1, 2, 3, 5, 9	(5,8)	1
I	0, 1, 4, 6, 7, 8	2, 3, 5, 9	(6,9)	1
J	0, 1, 2, 4	3, 5, 6, 7, 8, 9	(4,8)	1

Tablo 2'den de görüldüğü üzere, veri seti içinden seçilen başlangıç merkez noktalarıyla 10 farklı küme yapısı elde edilmiştir. Oluşan bu farklı küme yapıları sırasıyla A, B, …, J harfleri kullanılarak isimlendirilmiştir. Burada önemli olan sorunsallardan biri, farklı küme yapılarından hangisinin daha benzer olduğu, bir diğer önemli ve çalışmanın da temel konusu olan sorunsal ise dışardan seçilen başlangıç merkez noktalarının farklı küme yapıları oluşturup oluşturmayacağıdır.

Veri seti içinden seçilen herhangi 2 merkez veri için farklı 45 olası durumdan 25 tanesinde 0, 1, 6 ve 9. gözlemler aynı kümede yer almaktadır. Diğer oluşan kümelerin görülme sıklığı ise çok azdır. En sık karşılaşılan kümenin diğer kümelere oranla daha kararlı olduğu varsayılarak, bu kümelerin kullanılması gerektiğinin bir nedeni olarak düşünülebilir. Ancak bu varsayım, kurgulanan çalışmadaki gibi sıklığı çok yüksek olan bir küme sonucuyla karşılaşılmama olasılığının var olması, merkez verilerin çalışılan veri dışından rastlantısal olarak seçilmesi sonucunda farklı kümeleme sıklıkları oluşabileceği ihtimali nedeniyle çok güç bir varsayımdır.

Şimdi aynı veri seti için, veri seti dışında yer alan aynı boyutta nesneler yardımıyla 2 farklı küme oluşturulduğu durumu inceleyelim. Veri seti dışından nesneler belirlemek için de bir yöntem belirlenmesi gerekir. Bu çalışmada veri seti dışından merkez seçmek için ilk adım olarak tüm verilerin birbirlerine olan uzaklıkları hesaplanmıştır. Uzaklıkların hesaplanmasındaki ana amaç birbirine en uzak ve en yakın olan veriler yardımıyla, veri setine yakın ve uzak olan yeni merkez nesneler belirlemektir. Veri seti dışında yer alan ve basit bir mantıkla elde edilen yeni merkez nesnelerin farklı kümeler oluşturup oluşturmadığını gözlemlemek sorunsalın varlığını gösterebilmek adına önemli bir adımdır. Bununla birlikte içerden seçilen başlangıç merkez noktaları ile elde edilmiş ve en çok gözlemlenen kümelerden daha fazla gözlemlenen yeni kümelerin elde edilip edilmemesi durumu da sorunsalın varlığı ortaya çıkarır.

Bu çalışmada başlangıç merkezi olarak kullanılmak üzere oluşturulan yeni nesneler aşağıdaki dört adım yardımıyla oluşturulmuştur.

Adım 1: Veri setinde yer alan tüm nesnelerin birbirine olan uzaklıkları hesaplanır.

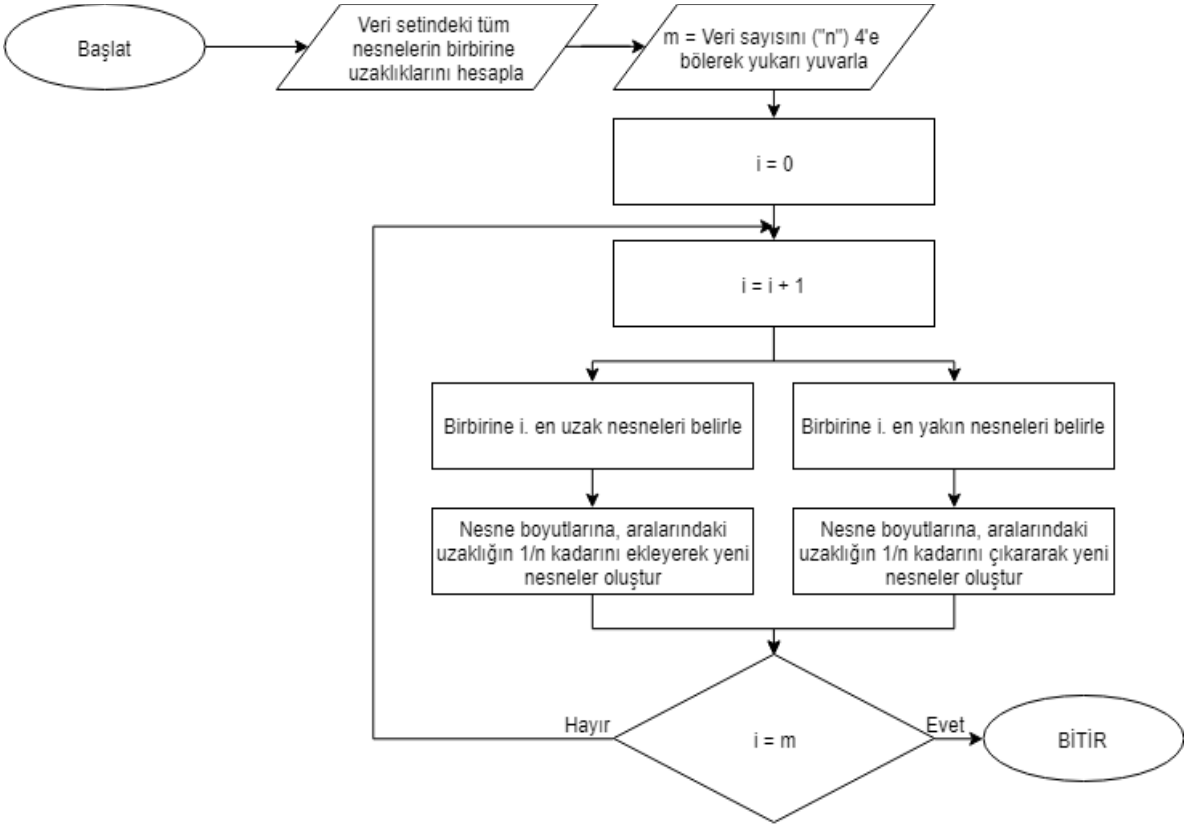
Adım 2: Veri sayısı (n) dörde bölünerek elde edilen değer yukarı yuvarlanır.

Adım 3: Adım 2'den elde edilen sayının m olduğu durum için;

Birbirine en uzak nesneler belirlenir. Bu iki nesneye uzaklık değerinin n'de biri eklenerek veri setinde yer almayan ve yoğunluğun olmadığı düzlemde yeni nesneler bulunur. Aynı işlem ikinci, üçüncü, … ve m. en uzak olan nesnelere de uygulanır. Böylelikle 2m tane veri seti dışından yeni nesne bulunur.

Adım 4: Adım 3'ün aynısı birbirine en yakın olan nesneler için yapılır fakat burada yeni nesnelere uzaklık değerinin n'de 1'i çıkartılarak bulunur.

Veri seti dışında yer alan aynı boyutta nesneler oluşturmak için geliştirilen yöntemin uygulama akış diyagramı Şekil 2'de sunulmuştur.



Şekil 2. Veri Seti Dışından Merkez Noktalar Oluşturma

Geliştirilen matematiksel yöntem yardımıyla elde edilen veri seti dışından kullanılacak merkez noktalar Tablo 3'de yer almaktadır. Elde edilen yeni gözlemlerin yarısı birbirine yakın olan gözlemler referans alınarak her bir boyutun gözlemler arası uzaklık değerinin 1/n'i kadar negatif yönde ötelenmesi ile diğer yarısı ise birbirine uzak olan gözlemler referans alınarak her bir boyutun gözlemler arası uzaklık değerinin 1/n'i kadar pozitif ötelenmesi ile oluşturulmuştur.

Tablo 3. Veri Seti Dışında Oluşturulan Merkez Noktalar

Gözlem	X	Y	Z	T	Gözlem	X	Y	Z	T
10	112,23	44,23	43,23	113,23	**16**	19,87	54,87	57,87	76,87
11	32,23	83,23	97,23	16,23	**17**	32,87	35,87	23,87	80,87
12	74,37	34,37	108,37	27,37	**18**	58,76	38,76	57,76	25,76
13	72,37	100,37	43,37	109,37	**19**	54,76	72,76	67,76	2,76
14	109,85	41,85	40,85	110,85	**20**	58,73	38,73	57,73	25,73
15	70,85	88,85	83,85	18,85	**21**	57,73	17,73	91,73	10,73

Tablo 4'de veri seti dışından yukarıdaki adımlarla elde edilen başlangıç merkez verileriyle oluşan tüm kümeler (bu kümelerin elde edilmesi için kullanılan başlangıç merkezler) ve bu kümelerin görülme sıklıkları yer alır.

Tablo 4. Veri Seti Dışından Seçilen Merkezlere Göre Oluşan Kümeler

	Küme 1	**Küme 2**	**Başlangıç Merkez Veriler (Dışardan)**	**Sıklık**
A	0,1,6,9	2,3,4,5,7,8	(10,15), (10,19), (11,12), (11,16), (11,17), (11,18), (11,20), (11,21), (12,16), (12,17), (12,19), (12,21), (13,15), (13,19), (14,15), (14,19), (15,16), (15,17), (15,18), (15,19), (15,20), (15,21), (16,17), (16,18), (16,19), (16,20), (16,21), (17,18), (17,19), (17,20), (17,21), (18,19), (18,20), (18,21), (19,20), (19,21), (20,21)	37
B	0,1,2,3,5,6,9	4,7,8	(10,11), (10,12), (10,16), (10,17), (10,18), (10,20), (10,21), (11,14), (12,14), (13,16), (13,17), (13,18), (13,20), (14,16), (14,17), (14,18), (14,20), (14,21)	18
C	0,1,2,5,6,9	3,4,7,8	(11,13), (12,13), (13,21)	3
G	0,2,3,4,6,7,8,9	1,5	(12,18), (12,20)	2
H	0,4,6,7,8	1, 2, 3, 5, 9	(11, 15)	1
J	**0,1,2,4**	**3,5,6,7,8,9**	**(10,13), (13,14)**	**2**
K	0,1,2,4,6,7,8	3,5,9	(11,19)	1
L	0,1,2,4,5	3,6,7,8,9	(12, 15)	1
M	0,1,2,3,45,6,7,8,9		(10, 14)	1

Veri seti dışından başlangıç merkez noktalarla analiz yapıldığında, veri seti içinden başlangıç merkez noktalarla analiz yapıldığında elde edilen bazı küme gruplarının (D, E, F ve I) elde edilmediği gözlemlenmiştir. Bunun yanında daha önce elde edilmeyen yeni küme gruplarının da (K, L, M) elde edildiği gözlemlenmiştir. Tablo 5'de ise veri seti içinden ve dışından yukarıdaki adımlarla elde edilen başlangıç merkez noktalarla oluşan tüm kümeler ve bu kümelerin görülme sıklıkları yer alır.

Tablo 5. Veri Seti İçinden ve Dışından Seçilen Merkezlere Göre Oluşan Kümeler

	Küme 1	**Küme 2**	**Başlangıç Merkez Veriler (Karışık)**	**Sıklık**
A	0,1,6,9	2,3,4,5,7,8	(0,11), (0,12), (0,15), (0,16), (0,17), (0,18), (0,19), (0,20), (0,21), (1,11), (1,12), (1,15), (1,16), (1,17), (1,18), (1,19), (1,20), (1,21), (2,11), (2,12), (2,15), (2,16), (2,19), (2,21), (3,11), (3,12), (3,15), (3,16), (3,17), (3,18), (3,19), (3,20), (3,21), (4,11), (4,12), (4,13), (4,15), (4,19), (5,11), (5,12), (5,15), (5,19), (5,21), (6,10), (6,12), (6,13), (6,14), (6,16), (6,17), (6,18), (6,19), (6,20), (6,21), (7,11), (7,12), (7,15), (7,16), (7,17), (7,18), (7,19), (7,20), (7,21), (8,11), (8,14), (8,15), (8,19), (9,11), (9,16), (9,17), (9,18), (9,20), (9,21)	72
B	0,1,2,3,5,6,9	4,7,8	(0,10), (0,13), (0,14), (1,10), (1,14), (2,10), (2,13), (2,14), (3,10), (3,13), (3,14), (4,16), (4,17), (4,18), (4,20), (5,10), (5,13), (5,14), (6,15), (7,10), (7,13), (7,14), (8,10), (8,16), (9,10), (9,14)	26
C	0,1,2,5,6,9	3,4,7,8	(1,13), (4,21), (8,12), (8,18), (8,20), (8,21), (9,13)	7
D	0,1,4,6,8,9	2,3,5,7	(2,18), (2,20), (5,18), (5,20)	4
E	0,3,4,6,7,8	1,2,5,9	(2,17)	1
F	0,1,2,3,4,6,7,8,9	5	(5,16)	1
G	0,2,3,4,6,7,8,9	1,5	(5,17)	1
H	0,4,6,7,8	1, 2, 3, 5, 9	(9,15)	1
I	0,1,2,4,6,7,8	3,5,9	(9,19)	1
K	0,1,2,4,6,7,8	3,5,9	(6,11)	1
L	0,1,2,4,5	3,6,7,8,9	(9,12)	1
M	0,1,2,3,4,5,6,7,8,9		(4,10), (4,14), (8,10)	3
N	**0,1,2,3,4,5,7,9**	**6,8**	**(8,17)**	**1**

Veri seti içinden ve dışından başlangıç merkez noktalarla analiz yapıldığında daha önce elde edilmeyen tek bir yeni küme grubunun (N) elde edildiği gözlemlenmiştir ve aynı zamanda yine daha önce veri seti içinden noktalarla elde edilen tek bir küme grubunun da (J) gözlemlenmediği görülmüştür.

Kurgu çalışmanın ikinci kısmında Tablo 1'de yer alan veri setinin üç kümeye ayrışmak istendiği durumu inceleyelim. Bir başka ifade ile aynı veri setini K-ortalamalar yöntemi ile ve k değerinin 3 olduğu durumda farklı başlangıç merkezlerden oluşabilecek tüm küme gruplarını inceleyelim. Kurgu çalışmanın bu kısmında veri seti içinden seçilebilecek olası tüm başlangıç merkezlerle, çalışmada kurgulanan dışardan elde edilen olası tüm başlangıç merkezlerle ve hem içerden hem de dışardan elde edilebilecek tüm başlangıç merkezlerle K-ortalamalar yöntemi Tablo 1'de yer alan veri setine uygulanmıştır. Elde edilen bulgularda 53 farklı küme yapısı

gözlemlenmiş ve oluşan küme grupları ve bu grupların görülme sıklığı Tablo 6 ve Tablo 7'de sunulmuştur. Bulgularda dışardan başlangıç merkez verileriyle elde edilen küme gruplarının bir kısmının (Örneğin H, N, R, S, T, AA, AJ, AJ, AL, AM, AU) içerden başlangıç merkez verileriyle elde edilmediği gözlemlenmiştir.

Tablo 6. K-ortalamalar Yöntemi Sonuçları (3 Küme)

	Küme 1	**Küme 2**	**Küme 3**	**İçerden**	**Dışardan**	**İçerden- Dışardan**
A	0,1,6,9	2,3,5,7	4,8	13	27	176
B	0,1,6,9	2,3,5	4,7,8	30	31	153
C	0,1,6,9	2,3,5,7,8	4	9	32	120
D	0,1	2,3,4,5,7,8	6,9	8	23	113
E	0,1,6,9	2,5	3,4,7,8	7	9	57
F	0,6,9	1	2,3,4,5,7,8	6	12	54
G	0,1,6	2,3,4,5,7,8	9	6	7	52
H	**0,1,6,9**	**2,3,4,5,7,8**	**-**	**0**	**9**	**52**
I	0,1,2,5,6,9	3,7,8	4	5	14	41
J	0,1,6,9	2,3,4,7,8	5	9	2	41
K	0,1,6,9	2,3,4, 5, 7	8	4	7	38
L	0,1,2,3,5	4,7,8	6,9	3	8	31
M	0,1,2,5	3,4,7,8	6,9	3	5	28
N	**0,1,2,3,5,6, 9**	**4,7,8**	**-**	**0**	**7**	**29**
O	0,1,6,9	2	3,4,5,7,8	4	0	29
P	0,1,4	2,3,5,7,8	6,9	1	6	11
R	**0,1,9**	**2,3,4,5,7,8**	**6**	**0**	**2**	**16**
S	**0,1,3,5,6**	**2,9**	**4,7,8**	**0**	**3**	**13**
T	**0,6,9**	**1,5**	**2,3,4,7,8**	**0**	**6**	**7**
U	**0,1,6,9**	**2,4,7**	**3,5,8**	**0**	**0**	**13**
V	0,1,6	2,9	3,4,5,7,8	1	1	9
Y	**0,1,2,5,6,9**	**3,4,7,8**	**-**	**0**	**0**	**11**
Z	**0,1,6,9**	**2,7**	**3,4,5,8**	**0**	**0**	**10**
AA	**0, 6**	**1,9**	**2,3,4,5,7,8**	**0**	**2**	**7**
AB	**0,1,6,9**	**2,4,5,7**	**3,8**	**0**	**0**	**9**
AC	0,1,3,5, 6, 9	2,4,7	8	1	1	6
AD	0,1,6,9	2,4	3,5,7,8	2	2	4

Tablo 6 ve Tablo 7'den de görüldüğü gibi, oluşan bazı küme gruplarının (U, Y, Z, AB, AK, AO, AP, AR, AS, AV, AY, AZ, BB, BC, BD, BE, BF, BG) ise sadece hem içerden hem dışardan (karışık) başlangıç merkez verileriyle elde edildiği gözlemlenmiştir. Y küme yapısı incelendiğinde ise sadece içerden ve dışardan küme merkez verileriyle elde edilmiş olup, 11 farklı başlangıç merkez veriyle elde edildiği ancak 3 kümeye ayrıştırmak istenilse dahi analizden elde edilen sonuçta 2 kümeye ayrıştığı gözlemlenmiştir.

Tablo 7. K-ortalamalar Yöntemi Sonuçları (3 Küme)

No	Küme 1	Küme 2	Küme 3	İçerden	Dışardan	İçerden- Dışardan
AE	0,2,3,5,6, 9	1	4,7,8	1	0	6
AF	0,2,5,6,9	1	3,4,7,8	1	0	6
AG	0,3,5,6,7,8,9	1,2	4	1	0	5
AH	0,3,6,7,8	1,2,5,9	4	1	0	5
AI	0,2,3,6,7,8,9	1,5	4	1	0	5
AJ	**0,6,9**	**1,2**	**3, 4,5,7,8**	**0**	**1**	**4**
AK	**0,1,2,3,5,7,9**	**4**	**6,8**	**0**	**0**	**5**
AL	**0,1,2,4**	**3,5,6,7,8,9**		**0**	**1**	**3**
AM	**0,1,2,4,5,7**	**3,9**	**6,8**	**0**	**1**	**3**
AN	0,3,4,6,7	1,2,5,9	8	1	0	2
AO	**0,1,2,3,5,6,9**	**4,7**	**8**	**0**	**0**	**3**
AP	**0,1,6,9**	**2,4,5**	**(3, 7, 8)**	**0**	**0**	**3**
AR	**0,1,4,5,7**	**-2**	**(3, 6, 8, 9)**	**0**	**0**	**3**
AS	**0,1,5**	**2,3,4,7,8**	**6,9**	**0**	**0**	**3**
AT	0,3,4, 6, 7,8,9	1,2	5	1	0	1
AU	**0,1,2,6**	**3,4,7,8**	**5,9**	**0**	**1**	**1**
AV	**0,1,2,5,6,9**	**3,8**	**4,7**	**0**	**0**	**2**
AY	**0,3, 6, 8, 9**	**1,4,5**	**2,7**	**0**	**0**	**2**
AZ	**0,6,7,8**	**1,2,3,5,9**	**4**	**0**	**0**	**2**
BA	0,4	1, 2, 5, 9	3,6,7,8	1	0	0
BB	**0,3,4,6,7,8**	**1,2,5,9**		**0**	**0**	**1**
BC	**0,1,2, 3, 5, 9**	**4,7**	**6,8**	**0**	**0**	**1**
BD	**0,1,2,3,4,5,6,7,8,9**	**-**	**-**	**0**	**0**	**1**
BE	**0,4,6,7,8**	**1,2,3,5,9**		**0**	**0**	**1**
BF	**0,6**	**1,2,9**	**3,4,5,7,8**	**0**	**0**	**1**
BG	**0,1,2,3,4,5,7,9**	**6,8**		**0**	**0**	**1**

Çalışmanın dördüncü kısmında ise farklı veri setleri için küme içinden merkezlerle elde edilmeyen yeni bir küme grubunun küme dışı merkezlerle elde edilebileceği ve bu küme grubunun görülme sıklığının diğer tüm içerden küme merkez noktalarıyla elde edilen küme gruplarının görülme sıklıklarından daha fazla olabileceği örnek veri setlerinin varlığı incelenmiştir. Tablo 8'de yer alan veri seti bu duruma örnektir.

Tablo 8. İncelenen Veriler 2

Gözlem	X	Y	Z	T	Gözlem	X	Y	Z	T
0	81	59	69	83	5	93	39	56	73
1	60	27	11	46	6	52	0	67	59
2	62	71	60	79	7	46	68	45	14
3	66	100	10	60	8	6	61	77	37
4	0	23	1	85	9	6	62	17	5

Tablo 8'de yer alan veri setini hem içerden seçilerek oluşabilecek tüm merkez noktalarla, hem çalışmada geliştirilen yöntemle elde edilen dışardan merkez noktalarla, hem de içerden ve dışardan oluşabilecek tüm merkez noktalarla K-ortalamalar kümeleme yöntemi

uygulanarak oluşan tüm küme grupları ve bu küme gruplarının görülme sıklığı Tablo 9'da sunulmuştur.

Tablo 9. İncelenen Veriler 2

Küme 1	Küme 2	İçerden	Dışardan	İçerden-Dışardan
0,2,5,6,8	1,3,4,7,9	4	1	6
0,2,5,6	1,3,4,7,8,9	4	6	17
0,2,3,4,5,6,7	1,8,9	3	5	6
0,2,3,7,8,9	1,4,5,6	1	5	1
0,2,3,5,7	1,4,6,8,9	2	5	6
0,1,2,3,5,6	4,7,8,9	4	0	7
0,2,3,5	1,4,6,7,8,9	4	2	8
0,1,2,3,5,6,7,8	4,9	3	4	8
0,1,4,5,6	2,3,7,8,9	3	1	2
0,1,2,3,4,5,6	7,8,9	1	0	2
0,1,2,5,6,7	3,4,8,9	1	0	0
0,1,2,5,6,7,8	3,4,9	2	4	4
0,1,2,5,6	3,4,7,8,9	4	5	12
0,2,3,5,7,8,9	1,4,6	1	0	2
0,2,3,5,6	1,4,7,8,9	1	0	1
0,1,2,3,5	4,6,7,8,9	1	0	1
0,1,2,3,5,6,7	4,8,9	1	4	4
0,1,2,3,5,7,8,9	4,6	1	0	0
0,1,2,3,5,7	4,6,8,9	2	3	3
0,1,2,4,5,6	3,7,8,9	1	0	3
0,2,5,6,7,8	1,3,4,9	1	2	3
0,1,2,3,5,6,7,8,9,	4	0	2	3
0,1,2,3,4,5,7,8,9	6	0	2	0
0,1,2,3,4,5,6,7,8,9	-	0	2	1
0,2,5	1,3,4,6,7,8,9	0	11	14
0,2,3,5,6,7	1,4,8,9	0	2	3
0,2,3,5,7,8	1,4,6,9	0	0	2
0,1,2,4,5,6,8	3,7,9	0	0	1

Tablo 9'dan da görüldüğü üzere {0,2,5} ve {1,3,4,6,7,8,9} olarak oluşan küme grubu veri seti içinden seçilen tüm ikili verilerle elde edilmemiş olup, veri seti dışından ve veri seti içinden ve dışından seçilen merkez noktalarla oluşmuştur. Bu küme grubunun da görülme sıklığı (11) diğerlerinden daha fazladır. Bunlara ek olarak, Tablo 9'dan da görüldüğü üzere farklı başlangıç merkezlerinden elde edilen çok farklı sayıda küme grubu olduğundan, bu küme gruplarının hangisinin daha iyi olduğu sorunsalı ortaya çıkar. Veri setinin yapısının araştırmacı tarafından çok iyi bilindiği durumlarda bile bilimsel olarak hangi küme grubunun bulgularının değerlendirileceği tartışmalıdır.

Tablo 8'de oluşan senaryonun görülme sıklığının araştırılması için kurgu çalışmadaki veri setine benzer yapıda 4 farklı boyutta (X, Y, Z, T) incelenen ve her bir boyutu [0,100] arasında rastlantısal olarak belirlenen 10.000 farklı veri seti oluşturulmuştur. Tüm bu veri setleri K-ortalamalar kümeleme yöntemi ile içerden, dışardan ve hem içerden hem de dışardan elde edilen başlangıç merkez noktalarla analiz edilmiş oluşan farklı küme grupları ve bu grupların görülme sıklığı hesaplanmıştır. %1,27 veri setinde küme içinden merkez noktalarla elde edilmeyen küme grupları gözlemlenmiştir. Bu grupların görülme sıklığı ise küme içinden merkezlerle elde edilen herhangi bir küme grubunun görülme sıklığından fazladır.

4. SONUÇ VE TARTIŞMA

Bu çalışmanın ana amacı K-ortalamalar yönteminin kullanıldığı durumlarda karşılaşılan en önemli sorunsallardan biri olan başlangıç merkez seçim sorunsalının varlığını vurgulamak ve yapılacak uygulamalı çalışmalarda bu sorunsalın varlığının araştırmacılar tarafından dikkate alınarak, bulguların değerlendirilmesi ve yorumlanmasını sağlamaktır.

K-ortalamalar yöntemi diğer iteratif yöntemler gibi başlangıç olarak seçilen ve küme merkezi olarak atanan değerlere bağlı kalarak bir kümeleme gerçekleştirir. Bu yöntemde, veri kümesindeki tüm gözlemlerin bu merkez noktalara olan uzaklıkları yardımıyla tüm verilerin ait olduğu kümeler belirlenir. Kümelere yerleşen veriler yardımıyla ikinci merkezler seçilerek tekrar uzaklıklar hesaplanıp, kümelere yerleşen yeni veriler bulunur. Bu işlemler de kümelerde bir önceki adım ile aynı veriler olduğu zaman sonlanır.

Algoritmanın ilk adımı rastlantısal olarak seçilen merkez noktalardan oluştuğundan, bu noktaların değişmesi nedeniyle aynı veri setinde farklı sonuçlar elde edildiği kurgu çalışmada örneklenmiştir. Bu sorunsal için literatürde önerilen yöntemlerden birinin n adet veri için k-ortalamalar yönteminin $\binom{n}{k}$ kere gerçekleştirilip, en çok karşılaşılan kümelerin bulgu olarak kabul edilebileceğidir. $\binom{n}{k}$ gerçekleştirilmek istenmesinin nedeni, yöntemde rastlantısal olarak seçilen ilk küme merkezlerinin veri seti içerisi ile sınırlandırılmasıdır. Ancak bu çalışmadaki kurgusal örnekte ortaya konulduğu gibi başlangıç merkezlerinin veri seti dışından ya da hem veri seti içinden hem veri seti dışından olabilme durumu da mevcuttur. Bu dışarıdan ya da hem içeriden hem dışarıdan seçilen başlangıç merkez verilerle elde edilen küme yapısının farklılaştığı, içerden başlangıç merkez verileriyle elde edilmeyen yeni küme yapıları olduğu kurgu çalışmanın üçüncü kısmında örneklendirilmiştir.

Sorunsalın varlığı ile ilgili literatürde çalışmalara rastlansa da uygulamalı çalışmalarda hangi merkez noktalarla küme yapılarının elde edildiği, başka merkez noktalarla başka küme yapılarının elde edilip edilmediği gibi bilgilere rastlanmamıştır. Uygulama araştırmalarında daha çok hangi kriterlerin dikkate alınarak K-ortalamalar kümeleme yönteminin uygulandığı, yöntemde hangi benzerlik ölçüsünün kullanıldığı ve analizden elde edilen bulgular ve bulguların değerlendirilmesi yer almaktadır. Bu durumda da başka başlangıç merkez verilerle elde edilecek küme yapılarının hangisinin daha benzer yapıda olduğu gibi karşılaştırmalı analizlerden daha çok, literatürde farklı yöntemlerle kaç kümeye ayrıştırmanın daha uygun olduğu hangi veri setleri için hangi yöntemin daha geçerli olabileceği yönünde çalışmalar mevcuttur. Ancak aynı sorunsal bu yöntemler için de geçerlidir.

Bu çalışmaya dair elde edilen en önemli bulgulardan biri; en yüksek sıklıkla elde edilen küme grubunun veri seti dışından ve veri seti içinden ve dışından seçilen merkez noktalarla da az bir yüzde ile olsa da elde edilebileceğidir. Bu bulgunun varlığı aynı zamanda, veri seti içinden elde edilen tüm başlangıç merkez seçimi ile oluşacak küme gruplarından hangisinin daha etkin ya da daha iyi ayrılmış olduğunu görülme sıklığına bakılarak karar verilmemesi gerektiğini de ifade eder.

Kısaca bu çalışmayla birlikte başka başlangıç merkez verilerle başka sonuçlar elde edilebilme ihtimalinin varlığı gösterilmiştir. Gelecek dönem çalışmalarda K-ortalamalar kümeleme yöntemi kullanacak olan araştırmacıların, bu ihtimalin varlığını da dikkate alarak araştırma bulgularını değerlendirilmesi önerilir.

KAYNAKÇA

Akay, Ö. (2019). "Türkiye'de Halk Kütüphanesi Kullanımının Panel Veri Kümeleme Analizi İle İncelenmesi", *Uluslararası Toplum Araştırmalar Dergisi,* 10(17), 1076-1099.

Akçapınar, G., Altun, A., & Aşkar, P. (2016). "Çevrimiçi Öğrenme Ortamındaki Benzer Öğrenci Gruplarının Kümeleme Yöntemi İle Belirlenmesi", *Eğitim Teknolojisi Kuram ve Uygulama*, *6*(2), 46-64.

Aydın, N. & Seven, A.N. (2015). "İl Nüfus Ve Vatandaşlık Müdürlüklerinin İş Yoğunluğuna Göre Hibrid Kümeleme İle Sınıflandırılması". *Yönetim ve Ekonomik Araştırmalar Dergisi*, 13(2), 181-201.

Atalay, A., & Tortum, A. (2010). "Türkiye'deki İllerin 1997-2006 Yılları Arası Trafik Kazalarına Göre Kümeleme Analizi". *Pamukkale University Journal of Engineering Sciences*, 16(3), 335-345.

Bülbül, Ş., & Camkıran, C. (2018). "Bankaların Klasik ve Bulanık Yaklaşımlarla Sınıflandırılması". *Trakya University Journal of Social Science*, 20(2), 367-385.

Çalışkan, S. K., & Soğukpınar, İ. (2008). "KxKNN: K-Means ve K En Yakın Komşu Yöntemleri İle Ağlarda Nüfuz Tespiti". *EMO Yayınları*, 120-24.

Çınaroğlu, S., & Bulut, H. (2018). "K-Ortalamalar ve Parçacık Sürü Optimizasyonu Tabanlı Kümeleme Algoritmaları İçin Yeni İlklendirme Yaklaşımları". *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 33(2), 413-424.

Duran, B. S. and P. L. Odel (1974). "Cluster Analysis (Lecture Notes in Economics and Mathematical Systems", Econometrics; Managing Editors: M. Beckmann and H. P. Kunzf. Springer Verlag: NewYork.

Durucasu, H., Aşan, Z., & Er, F. (2006). "Öğrencilerin Yaz Okulu Hakkındaki Görüşleri İçin Kümeleme Analizi". *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi,* 7(1), 97-101.

Fırat, M., Dikbaş, F., Koç, A. C., & Güngör, M. (2012). "K-Ortalamalar Yöntemi İle Yıllık Yağışların Sınıflandırılması Ve Homojen Bölgelerin Belirlenmesi". *İMO Teknik Dergi*, 383, 6037-6050.

Forgy E.W. (1965). "Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications, Biometrics, 21 (3), 768-769.

Fraley, C., & Raftery, A. E. (1998). "How Many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis". *The Computer journal*, 41(8), 578-588.

Khan, S. S., & Ahmad, A. (2013). "Cluster Center İnitialization Algorithm For K-Modes Clustering". Expert Systems with Applications, 40(18), 7444-7456.

Hajizadeh, E., Ardakani, H. D., ve Shahrabi, J. (2010). "Application of Data Mining Techniques ın Stock Markets: A Survey". *Journal of Economics and International Finance,* 2(7), 109.

Han, J., and Kamber, M., (2006), Data Mining Concepts and Techniques, Morgan Kauffmann Publishers Inc.

Işık, M., & Çamurcu, A. Y. (2007). K-means, K-medoids ve bulanık C-means algoritmalarının uygulamalı olarak performanslarının tespiti.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Vol:6. Englewood Cliffs: Prentice hall

Karypis, M. S. G., Kumar, V., & Steinbach, M. (2000, August). "A comparison of document clustering techniques". In TextMining Workshop at KDD2000 (2000).

Mac Queen, J.B., (1967). "Some Methods for Classification and Analysis of Multivariate Observations". In: Proceedings of the Symposium on Mathematics and Probability, 5th, Berkely.

Meila, M., & Heckerman, D. (2013). "An experimental comparison of several clustering and initialization methods". arXiv preprint arXiv:1301.7401.

Mercer D. P., (2003). "Clustering Large Datasets", http://www.stats.ox.ac.uk/∼ mercer/documents/Transfer.pdf (date accessed: 03.21.2011).

Na, S., Xumin, L., & Yong, G. (2010, April). "Research on k-means clustering algorithm: An improved k-means clustering algorithm". In 2010 Third International Symposium on intelligent information technology and security informatics (pp. 63-67). IEEE.

Özdemir, A., & Orçanlı, K. (2012). "İki Aşamalı Kümeleme Algoritması İle Pazar Bölümlemesi, Müşteri Profillerinin Belirlenmesi ve Niş Pazarların Tespiti". *Uşak Üniversitesi Sosyal Bilimler Dergisi*, (11).

Higgs, R. E., Bemis, K. G., Watson, I. A., & Wikel, J. H. (1997). "Experimental designs for selecting molecules from large chemical databases". Journal of Chemical Information and Computer Sciences, 37(5), 861-870.

Selvi, H. Z., Çağlar, B. (2016). "Using K-Means and K-Medoids Methods for Multivariate Mapping", *International Journal of Applied Mathematics,* Electronics and Computers, 4, 342-345.

Steinley, D., & Brusco, M. J. (2007). "Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques". *Journal of Classification*, 24(1), 99-121.

Tatlıdil, H. (1992). "Uygulamalı Çok Değişkenli İstatistiksel Analiz", H.Ü. Fen Fakültesi İstatistik Bölümü, Ankara.

Witten I. H., Frank E., (1999), "Data Mining: Practical machine learning tools with Java implementations", San Francisco, Morgan Kaufmann.

Yalçın, S., & Ayyıldız, E (2018). "Analysis of Airports Using Clustering Methods: Case Study In Turkey". *Journal of Management Marketing and Logistics, 5*(3), 194-205.

Yaraş, E. (2005). "Tüketicilerin Pazarlama Karması Kararları Ve Marka Değeri Algılamaların Göre Kümeler Halinde İncelenmesi". *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 19(2), 349-372.

Yedla, M., Pathakota, S. R., & Srinivasa, T. M. (2010). "Enhancing K-Means Clustering Algorithm with Improved Initial Center". *International Journal of computer science and information technologies, 1*(2), 121-125.

Zırhlıoğlu, G ve Karaca, S., (2006). "Genç Bayanlar Dünya Voleybol Şampiyonasına Katılan Sporcuların Kümeleme Analizi İle İncelenmesi". *Hacettepe J. Of Sport Sciences,* 17(1): 20-25.