VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Boyut Azaltmanın Bulanık C-Ortalama Kümeleme Teknikleri Üzerindeki Etkisi

Nuran PEKER[1]*, Cemalettin KUBAT[1]

[1]*Sakarya Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği, Sakarya*

Makale Bilgisi

Başvuru:
14/07/2020
Kabul:
02/01/2021

Özet

Bulanık c-ortalama kümeleme, literatürde farklı alanlarda kullanılan yaygın kümeleme algoritmalarından biridir. Boyut küçültme, büyük veri kümelerini, en az bilgi kaybıyla eşdeğeri olan daha küçük boyutlu veri kümelerine dönüştüren bir tekniktir. Bu makalede, boyut azalmasının farklı bulanık kümeleme teknikleri üzerindeki etkisi incelenmektedir. Bu amaçla farklı dört bulanık kümeleme algoritması kullanıldı: Bulanık C-Ortalamalar (BCO), Tip-2 Bulanık C-Ortalamalar (BCO2), Olasılıksal Bulanık C-Ortalamalar (OBCO) ve Denetimsiz Olasılıksal Bulanık C-Ortalamalar (DOBC). Boyut küçültme için verilerdeki varyansı minimum %80 açıklayan bir dizi bileşen seçildi. Boyutsallığın azaltılması için Kesik Tekil Değer Ayrıştırma (KTDA) tekniği kullanıldı. Çalışmada, ilk olarak, orijinal gerçek dünya veri kümeleri, bahsedilen dört yöntemle kümelendi. Daha sonra, bu veri kümelerinin boyutu küçültülmüş hali de yine bu dört yöntemle kümelendi. Kümeleme performansı için dört dahili kümeleme değerlendirme metriği kullanıldı. Bunlar Silhouette İndeksi (SI), Bölme Katsayısı (BK), Bölme Entropisi (BE) ve Kök Ortalama Kare Hatası (KOKH). Yöntemlerin, orijinal ve boyutu azaltılmış veri kümeleri için kümeleme performansı, karşılaştırmalı olarak sunulmaktadır. Sonuçlara göre, indirgenmiş veriler üzerinde yöntemlerin performansı, orijinal verilerden daha başarılıdır. Boyut azaltımının kümeleme başarısına katkısı en çok BCO için, en az BCO2 için elde edilmektedir.

Anahtar Kelimeler: *Kümeleme, Bulanık C-ortalamalar, KTDA, Boyut azaltma*

Effectiveness of Dimension Reduction on Fuzzy C-Means Clustering Techniques

Abstract

Fuzzy c-mean clustering is one of the common clustering algorithms used in different fields in the literature. Dimension reduction is a technique that transforms large data sets into smaller size datasets, which is equivalent to it with minimal loss of information. This article examines the effect of dimension reduction on different fuzzy clustering techniques. For this purpose, four different fuzzy clustering algorithms were used: Fuzzy C-Means (FCM), Type-2 Fuzzy C-Means (FCM2), Possibilistic Fuzzy C- Means (PFCM) and Unsupervised Possibilistic Fuzzy C-Means (UPFC). For dimension reduction, a number of components were selected that explain the variance in the data by a minimum of 80%. Truncated Singular Value Decomposition (TSVD) technique was used for dimentionality reduction. In the study, first, original real-world datasets were clustered with the four methods mentioned, seperately. Then, these datasets that have reduced dimensions were clustered by these four methods, also. Four internal cluster evaluation metrics were used for cluster performance. These are Silhouette Index (SI), Partition Coefficient (PC), Partition Entropy (PE) and Root Mean Square Error (RMSE). The clustering performance of the original and reduced data sets of the methods is presented

* e-posta: nuran.peker@ogr.sakarya.edu.tr

** Bu çalışmanın bir kısmı III. International Conference on Data Science and Applications 2020'de sözlü olarak sunulmuştur.

comparatively. According to the results, the performance of the methods on the reduced data is more successful than the original data. The contribution of dimension reduction to clustering success was achieved the most for FCM and the least for FCM2.

Keywords: *Clustering, Fuzzy C-means, TSVD, Dimension reduction*

1 Giriş

Kümeleme, veri yapısı içindeki gizli örüntü ve benzer yapıları bulmayı amaçlayan denetimsiz bir veri madenciliği yaklaşımıdır. Katı kümeleme (hard clustering) yaklaşımında bir veri noktası sadece bir kümeye ait olabilmekte iken; bulanık kümeleme yaklaşımında, belirli bir üyelik derecesine göre birden fazla kümeye ait olabilir. Literatürde iyi bilinen bir bulanık kümeleme yöntemi olan BCO[1,2]'ya, yakın geçmişte yeni yöntemler eklenmiştir. BCO2[3] algoritması, tüm veri noktalarının küme oluşum hesaplanmasında aynı katkıyı vermesi yerine, daha yüksek üyeliğe sahip olanların daha fazla katkıda bulunması esasına dayanır [4]. OBCO[5], OBC'nin bulanık yaklaşımını ve Possibilistic C-Means[6] algoritmasının olasılıksal yaklaşımını birleştiren ve Kısım 2.4'te açıklanan hedef fonksiyonu minimize etmeye çalışan bir yöntemdir. Dolayısıyla OBCO'nun, veri noktasıyla ilişkili olasılıklı ve bulanık bir üyeliği mevcuttur. DOBC[7] algoritması ise, Possibilistic Clustering Algorithm(PCA)[8] yaklaşımını temel alan ve bu yaklaşımın başlangıç koşullarını iyileştirmeyi amaçlayan bir yöntemdir. [9] çalışması, yukarıda bahsi geçen yöntemleri de kapsayan önemli bulanık kümeleme algoritmalarının performansının kapsamlı ve deneysel analizini yapmaktadır.

Literatürde bulanık kümelemenin boyut azaltmada kullanıldığı [10,11,12] gibi çalışmalar vardır. [13] çalışması sekiz ayrı boyut azaltma tekniğinin performansını BCO üzerinde test etmektedir. Bizim çalışmamızda farklı olarak, boyut azaltmanın bulanık yaklaşımlar üzerindeki etkisi ele alınmaktadır. BCO ve türevleri, çoğu kümeleme probleminde iyi çalışmakla birlikte, bu algoritmaların yüksek boyutlu veri kümeleri ile ilgili sorunları olduğu bilinmektedir[14]. Bu çalışmada, boyut azaltmanın bulanık kümeleme yöntemleri üzerindeki başarısı ele alınmaktadır. Boyut azaltma, veri sıkıştırma kapasitesine sahip olduğu için depolama alanını, orijinal veriye göre daha az boyutlu olduğu için de hesaplama zamanını azaltmaya yardımcı bir tekniktir. Ayrıca verilerin görselleştirilmesine ve daha kolay anlaşılıp yorumlanmasına katkı yapmaktadır. Çalışmada, boyut indirgeme yöntemi olarak KTDA[15] kullanılmaktadır. KTDA, Tekil Değer Ayrıştırma(TDA)[16] yönteminin, ilk en büyük k tane tekil değerinin seçilip, gerisinin sıfıra eşitlendiği bir lineer boyut azaltma tekniğidir. Bu yönüyle, bellek kullanımı ve hesaplama zamanı açısından daha avantajlıdır. [17] çalışması, EKG verileri için öznitelik çıkarımı ve veriyi sıkıştırma da KTDA yöntemini kullanmaktadır. [18] çalışması, yine EKG verileri için, KTDA ve gömülü sıfır ağaç dalgacık (embedded zero tree wavelet) modelini birleştirerek büyük sistemler için EKG veri sinyallerini sıkıştırmaya dayalı bir model sunmaktadır.

2 Materyal ve yöntemler

Çalışmada dört farklı veri setinin orijinal ve boyutu azaltılmış versiyonları, bulanık kümelemeye dayalı bahsi geçen dört farklı yöntem ile kümelenmekte ve boyut azaltımının, bu kümeleme yöntemleri üzerindeki etkisi, dört değerlendirme metriği kullanılarak test edilmektedir.

2.1 Veri setleri

Çalışmada, UCI Repository'den[19] seçilen ve tamamı gerçek dünya verisinden oluşan, Tripadvisor, Parkinsons, Sales_Transaction_Weekly (Sales), Satatlog Vehicle veri setleri kullanılmıştır. Veri setlerine ait özellikler Tablo 1'de görülmektedir.

Tablo 1. Veri setleri

Veri Seti	#Örnek	#Öznitelik	#Sınıf
Tripadvisor	980	11	-
Parkinsons	197	23	2
Sales	811	53	-
Vehicle	946	18	4

2.2 Bulanık C-ortalamalar(BCO)

BCO yöntemi, literatürde en çok kullanılan bulanık kümeleme algoritmalarından biridir. Algoritmada bir veri noktası, değeri [0,1] arasında değişen bir üyelik derecesine göre birden fazla kümeye ait olabilmektedir. Algoritma, veri setini, grup içi kare hatalarının toplamını en aza indirecek şekilde, önceden tanımlı k adet kümeye ayıran özyinelemeli bir yöntemdir ve Denklem (1)'de verilen amaç fonksiyonunun, $1 \leq m < \infty$ kısıtı altında minimize edilmesi prensibine dayanır. Denklemdeki m, bulanıklık değerini ifade eder.

$$A_{\mathrm{BCO}} = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ij}^{m} \parallel x_i - v_j \parallel^2 \qquad (1)$$

Küme merkezlerini ifade eden v_j Denklem(2) ile hesaplanır.

$$v_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \qquad (2)$$

Üyelik matrisini gösteren u_{ij}, başlangıçta rastgele atanır ve Denklem(3)'e göre güncellenir.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\parallel x_i - v_i \parallel}{\parallel x_i - v_k \parallel} \right)^{2/(m-1)}} \qquad (3)$$

2.3 Tip-2 bulanık C-ortalamalar(BCO2)

Bu algoritmanın temelindeki ana fikir, tüm veri noktalarının eşit katkısı yerine, daha yüksek üyelik değerine sahip veri noktalarının küme merkezi hesaplamasında egemen olmasıdır. Veri kümeleri küresel olan veri setleri üzerinde başarılı, ancak veri kümeleri küresel olmayan ve karmaşık yapılar içeren veri setleri üzerinde başarısız olan bir algoritmadır[9]. Tip-2 üyeliği Denklem(4) ile hesaplanır:

$$z_{ij} = u_{ij} - \frac{1 - u_{ij}}{2} \qquad (4)$$

Burada u_{ij} ve z_{ij}, sırasıyla Tip-1 ve Tip-2 bulanık üyeliğini ifade eder. Küme merkezleri v_j Denklem(5)'teki eşitliğe göre güncellenir.

$$v_j = \frac{\sum_{i=1}^{n} z_{ij}^{m} x_i}{\sum_{i=1}^{n} z_{ij}^{m}} \quad (5)$$

2.4 Olasılıksal bulanık C- ortalamalar(OBCO)

OBCO, BCO'nun bulanık ve PCM'nin olasılıksal yaklaşımını entegre eden, böylelikle veri noktasıyla ilişkili olasılıksal bir o_{ij} ve bulanık bir u_{ij} üyeliği oluşturan bir yöntemdir. Algortima, Denklem(6)'da verilen amaç fonksiyonunu minimize etmeye çalışır.

$$A_{OBCO} = \sum_{j=1}^{c}\sum_{i=1}^{n}\left(au_{ij}^{m} + bo_{ij}^{\lambda}\right) \| x_i - v_j \| + \sum_{j=1}^{c} \omega_j \sum_{i=1}^{n}\left(1 - o_{ij}\right)^{\lambda} \quad (6)$$

Burada, $0 \le u_{ij} < 1$, $0 \le o_{ij} < 1$, m>1, λ>1, a>0 ve b>0 kısıtları sağlanmalıdır. Fonksiyondaki a ve b, bulanık üyelik ve tipiklik değerlerinin göreceli önemini tanımlayan birer sabit sayıdır. λ, kümeleme için tipiklik miktarını belirten tipiklik üssü; ω kümelerin varyansını kontrol etmek için olası ceza terimidir. Bulanık u_{ij} üyeliği, Denklem(3)'teki gibi hesaplanır. Olasılıksal o_{ij} üyeliği Denklem(7), küme merkezlerinin güncellenmesi olan v_j ise, Denklem(8)'deki gibi hesaplanır:

$$o_{ij} = \frac{1}{1 + \left(\frac{b \| x_i - v_j \|^2}{\omega_j}\right)^{\frac{1}{\lambda - 1}}} \quad (7)$$

$$v_j = \frac{\sum_{i=1}^{n} (au_{ij}^{m} + bo_{ij}^{\lambda}) x_i}{\sum_{i=1}^{n} au_{ij}^{m} + bo_{ij}^{\lambda}} \quad (8)$$

2.5 Denetimsiz olasılıksal bulanık C-ortalamalar(DOBC)

DOBC, PCA algoritmasının bir uzantısıdır. PCA, başlangıç koşullarına karşı çok duyarlıdır ve bazen çakışan kümeler oluşturabilir[7]. DOBC, OBCO yönteminden esinlenerek hem başlangıç koşullarından kaynaklı gürültü problemine hem de çakışan küme problemine çözüm önermektedir. Algoritma Denklem(9)'daki amaç fonksiyonunu minimize etmeye dayalıdır.

$$A_{DOBC} = \sum_{j=1}^{c}\sum_{i=1}^{n}\left(au_{ij}^{m} + bo_{ij}^{\lambda} \| x_i - v_j \|\right) + \frac{\beta}{n^2\sqrt{c}} \sum_{j=1}^{c}\sum_{i=1}^{n}\left(o_{ij}^{\lambda} \log o_{ij}^{\lambda} - o_{ij}^{\lambda}\right) \quad (9)$$

Fonksiyondaki u_{ij} ve v_j sırasıyla Denklem(3) ve Denklem(8)'deki gibi hesaplanır. Fonksiyondaki o_{ij} değeri ise $1 \le i \le n, 1 \le j \le c$ kısıtları altında Denklem(10)'daki gibi hesaplanır.

$$o_{ij} = exp\left(-\frac{bn\sqrt{c} \| x_i - v_j \|^2}{\beta}\right) \quad (10)$$

Eşitliklerde a ve b, a>0, b>0 koşulunu sağlayan birer sabit sayı; $\overline{x}$, *x*'e bağımlı verilerin ortalaması olmak üzere, β değeri, Denklem(11) ile hesaplanır:

$$\beta = \frac{\sum_{i=1}^{n} \| x_i - \overline{x} \|^2}{n} \quad (11)$$

2.6 Kesik tekil değer ayrıştırma(KTDA)

TDA, lineer cebirde reel veya kompleks matrisleri çarpanlarına ayırma yöntemlerinden biridir. M matrisi *mxn* boyutuna sahip olmak üzere, Denklem(12)'deki eşitlik ile çarpanlarına ayrılır.

$$M = U \sum V^T \quad (12)$$

Burada U, *mxm* boyutlu üniter bir matris; ∑, *mxn* boyutlu bir köşegen matrisi; V ise *nxn* boyutlu bir üniter matristir. M'nin rankı *r* olmak üzere köşegen matrisi, Denklem(13) ile gösterilir.

$$\sum = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix} \quad (13)$$

Burada $\sigma_1 \ge \sigma_2 \ge \sigma_3 \ldots \ge \sigma_r \ge 0$ M'nin tekil değerleri olarak adlandırılır. İlk *k* adet en büyük tekil değerin seçilip, geri kalanların tümünün sıfıra eşitlenmesi ve böylece U ve V'nin sadece ilk *k* sütununun kullanılması ile KTDA bulunmuş olur. Bu şekilde boyut indirgemek, sonraki adımlarda matris hesaplarını kolaylaştırmaktadır.

2.7 Değerlendirme metrikleri

SI: SI[20] değeri, bir veri noktasının, diğer kümelere kıyasla kendi kümesine ne kadar benzer olduğunun bir ölçüsüdür ve Denklem(14)'teki gibi hesaplanır:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad (14)$$

Burada *a(i)*, i. veri noktasının ait olduğu kümedeki diğer bütün veri noktalarına olan ortalama uzaklığıdır. *b(i)* ise, i. veri noktasının ait olmadığı diğer kümelerdeki verilerle olan minimum ortalama uzaklığıdır. SI değeri, [-1,+1] arasında bir değer alır. Burada +1'e yaklaşan yüksek bir değer o verinin kendi kümesine iyi uyduğunu ve komşu kümelerle iyi eşleşmediğini gösterir. 0 değeri, verinin, kümeler arasındaki karar sınırında olduğunu, negatif değerler ise bu verinin yanlış kümeye atanmış olabileceğini gösterir.

BK: BK[21], kümeler arasındaki örtüşmeyi ölçen bir yöntemdir. [0,1] arasında değer alır. 1'e yakın değerler bulanıklığın az olduğuna ve kümelemenin başarısına işaret eder. Denklem(15) ile hesaplanır:

$$BK = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \quad (15)$$

BE: BE[22], bulanık kümeler için önerilmiş diğer bir ölçüm yöntemidir. [0, 1] arasında değer alır. BE'nin düşük olması yani 0'a yaklaşan değerler, kümelemenin iyi yapıldığını gösterir. BE değeri, Denklem(16)'da gösterilen eşitlik ile hesaplanır:

$$BE = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} \log\left(u_{ij}\right) \quad (16)$$

Denklem(15) ve Denklem(16)'daki u_{ij} bulanıklık derecesini, *c* küme sayısını, *n* eleman sayısını göstermektedir.

KOKH: Bir model tarafından öngörülen değerler ile gerçek değerler arasındaki farkların sık kullanılan bir ölçüsüdür. Denklem(17)'ye göre hesaplanır:

$$KOKH = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x - \overline{x})^2} \qquad (17)$$

Denklem(17) eşitliğindeki x , gerçek değerleri; $\overline{x}$ ise öngörülen değerleri ifade etmektedir.

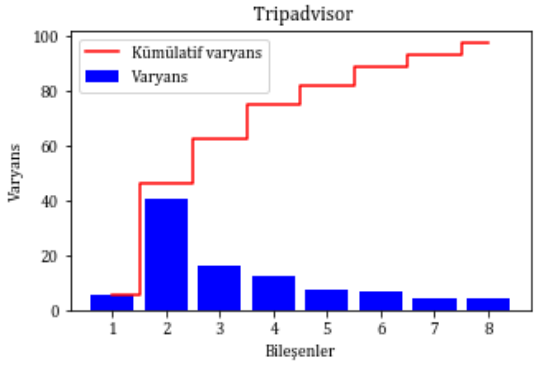
3 Deneysel Sonuçlar

Bu bölümde, yapılan çalışmanın detayları üzerinde durulmaktadır. Çalışmada, farklı alanlardan seçilen dört veri seti önce orijinal haliyle, BCO, BCO2, OBCO ve DOBC yöntemleriyle kümelenmektedir. Daha sonra KTDA yöntemi kullanılarak indirgenen bu veri setleri, yine bu yöntemler kullanılarak kümelenmektedir. KTDA'nın uygulanması için *scikit-learn*[23] kütüphanesi kullanılmaktadır. Bulanık yöntemlerin uygulanmasında ise, CRAN-R [24] yazılımından faydalanılmaktadır. KTDA ile boyut indirgeme yaklaşımında, her bir veri seti için, verideki minimum %80'lik varyansı açıklayacak şekilde bileşen seçilmektedir. Bu koşul altında, Tripadvisor veri seti için beş, Parkinsons ve Vehicle veri setleri için dört, Sales veri seti için ise iki bileşen seçilmektedir. Şekil 1'de Tripadvisor, Şekil 2'de Parkinsons, Şekil 3'te Sales ve Şekil 4'te Vehicle veri setleri üzerinde bileşenlerin varyansı ve kümülatif varyansı görülmektedir. Yöntemlerin kümeleme sonuçlarına örnek olması açısından, BCO'nun, veri setleri üzerindeki kümeleme sonuçları Şekil 5-8'de verilmektedir.

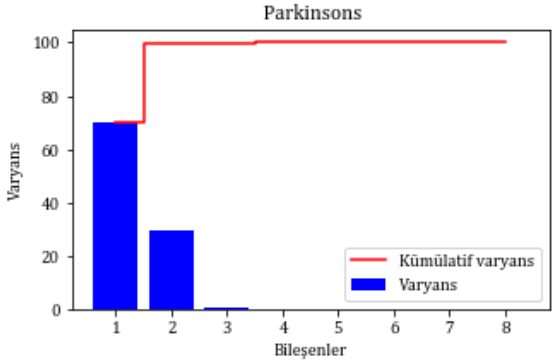
Yöntemlerin uygulanması esnasında optimum küme sayısının bulanabilmesi için her bir bulanık kümeleme yaklaşımı, farklı parametre ve farklı küme sayıları ile çalıştırılmıştır. Çalışmada baz alınan değerlendirme metriklerinin maksimum performansına dayalı olarak elde edilen sonuca göre, bütün kümeleme yaklaşımları veri setlerinin iki kümeye bölünmesi sonucunu vermiştir. Kümeleme esnasında bütün bulanık yaklaşımlar için bulanıklığı ifade eden m değeri 2 alınmaktadır. Denklem(6) ve Denklem(9)'daki amaç fonksiyonunun bulanık kısmına ait nispi önemi ifade eden a ve b değerleri 1; tipiklik miktarını ifade eden λ değeri ise 2 olarak alınmaktadır. Bütün yöntemler için uzaklık mesafesini hesaplamada *squared Euclidean* kullanılmaktadır.

Yöntemlerin, orijinal ve indirgenmiş veri setleri üzerindeki performansları Tablo 2'de görülmektedir. Tabloda *Orijinal* ve *İndirgenmiş* olarak belirtilen sütunlar, sırasıyla veri setlerinin orijinal ve indirgenmiş boyutları ile elde edilen sonuçları ifade etmektedir. Ayrıca tabloda verilen metriklerden SI ve BK değerlerinin yüksek, BE ve KOKH değerlerinin düşük olması, kümeleme başarısının daha yüksek olduğu anlamını taşımaktadır. Bu açıdan bakıldığında, Tripadvisor veri seti üzerinde boyut indirgemenin, genel olarak yöntemlerin başarısını arttırdığı söylenebilir. Ancak BCO2 için BK'nın düştüğü, BE'nin arttığı; OBCO için ise BK'nın düştüğü görülmektedir. Parkinsons veri seti için boyut indirgeme OBCO hariç diğer üç yöntemin KOKH değerinin yükselmesine yol açmaktadır. Ancak diğer metriklerin değerleri olumlu yönde etkilenmektedir. Sales veri seti için boyut indirgeme, BCO2 dışındaki diğer üç yaklaşım üzerinde olumlu sonuçlar üretmektedir. Vehicle veri seti üzerinde boyut azaltma, yöntemlerin performansını arttırmaktadır. Sadece KOKH değeri açısından BCO2 ve DOBC olumsuz etkilenmektedir.

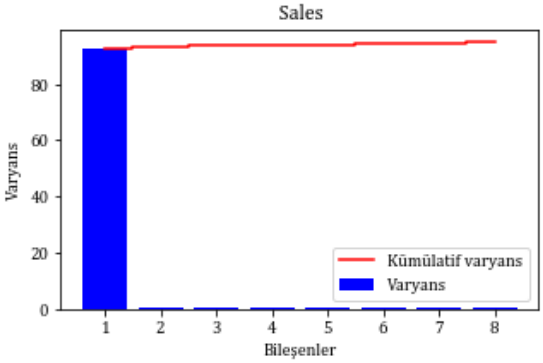
Yöntemlerin her bir metrik değeri için ayrı ayrı hesaplanan ortalama değerleri, Şekil 9-12'de görülmektedir.



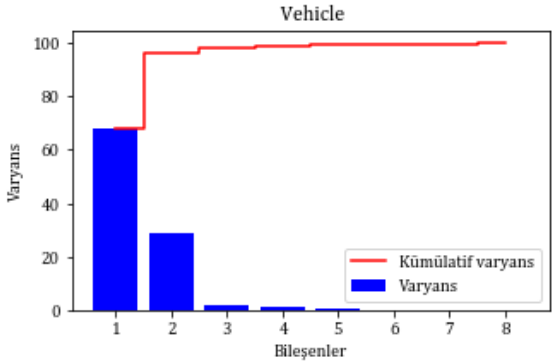
Şekil 1. Tripadvisor bileşen ve varyansları



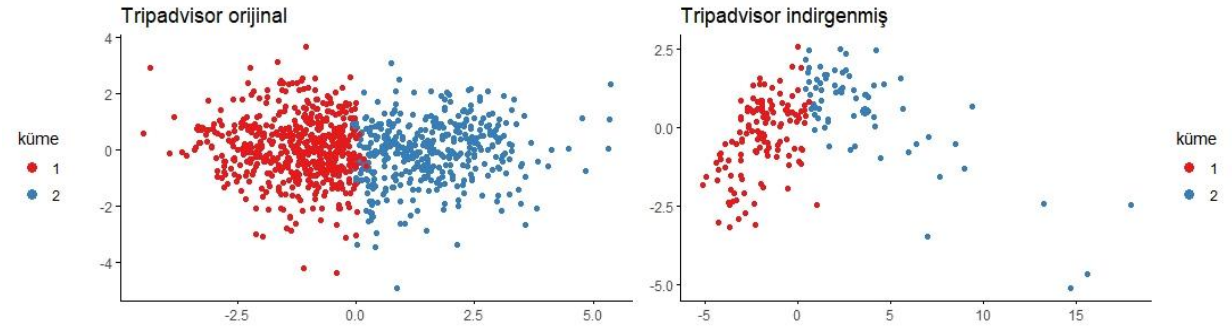
Şekil 2. Parkinsons bileşen ve varyansları



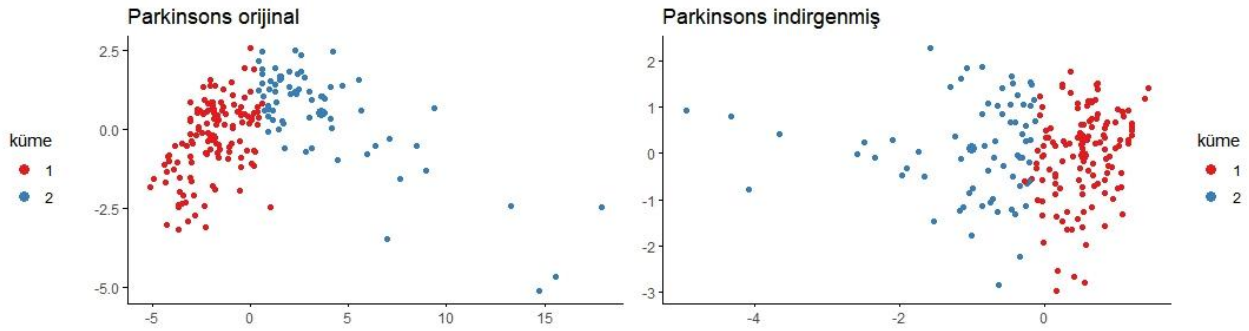
Şekil 3. Sales bileşen ve varyansları



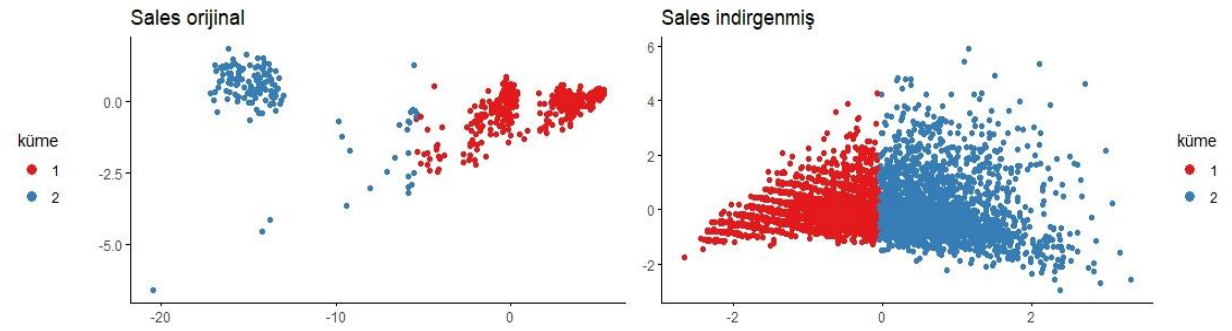
Şekil 4. Vehicle bileşen ve varyansları



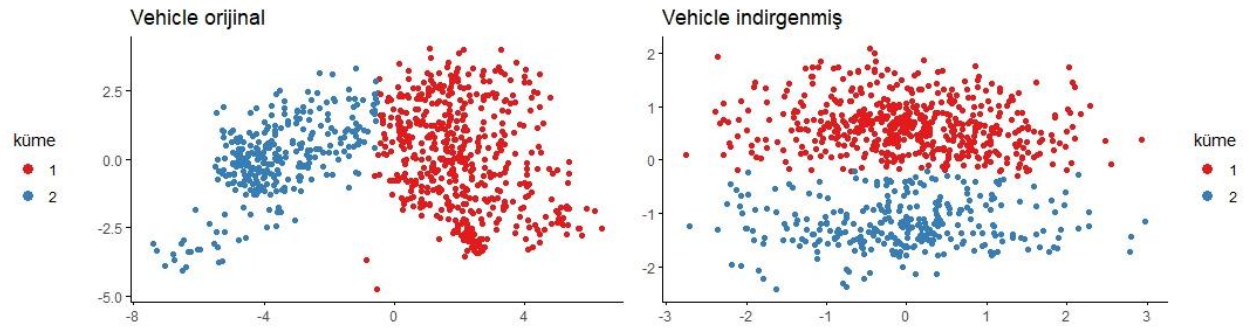
Şekil 5. BCO'nun orijinal ve indirgenmiş Tripadvisor veri seti üzerinde kümeleme sonuçları



Şekil 6. BCO'nun orijinal ve indirgenmiş Parkinsons veri seti üzerinde kümeleme sonuçları



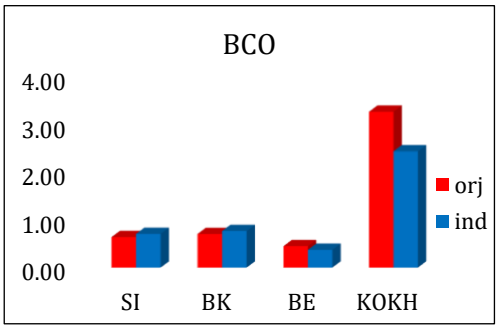
Şekil 7. BCO'nun orijinal ve indirgenmiş Sales veri seti üzerinde kümeleme sonuçları



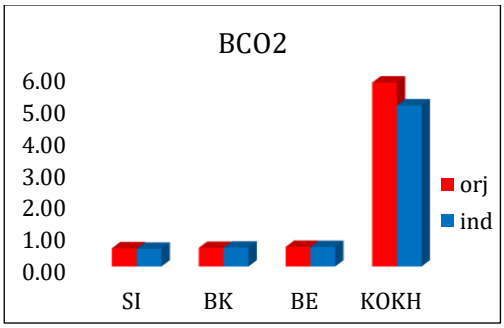
Şekil 8. BCO'nun orijinal ve indirgenmiş Vehicle veri seti üzerinde kümeleme sonuçları

Tablo 2 Bulanık kümeleme yöntemlerinin orijinal ve indirgenmiş veri setleri üzerindeki performansı

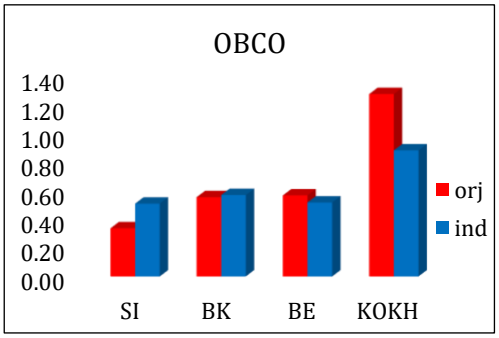
Veri Seti	Yöntem	Metrik	Orijinal	İndirgenmiş
Tripadvisor	BCO	SI	0.40	0.54
		BK	0.54	0.66
		BE	0.66	0.53
		KOKH	2.81	1.21
	BCO2	SI	0.41	0.50
		BK	0.58	0.51
		BE	0.61	0.69
		KOKH	4.09	2.30
	OBCO	SI	0.05	0.47
		BK	0.63	0.58
		BE	0.62	0.61
		KOKH	0.76	0.74
	DOBC	SI	0.40	0.54
		BK	0.09	0.17
		BE	0.42	0.37
		KOKH	4.28	2.41
Parkinsons	BCO	SI	0.59	0.65
		BK	0.66	0.71
		BE	0.52	0.45
		KOKH	3.56	3.94
	BCO2	SI	0.52	0.49
		BK	0.57	0.63
		BE	0.61	0.55
		KOKH	4.66	7.84
	OBCO	SI	-0.24	-0.08
		BK	0.70	0.74
		BE	0.57	0.54
		KOKH	3.29	2.18
	DOBC	SI	0.57	0.61
		BK	0.18	0.24
		BE	0.42	0.40
		KOKH	3.38	5.18
Sales	BCO	SI	0.93	0.94
		BK	0.92	0.94
		BE	0.15	0.12
		KOKH	3.91	2.70
	BCO2	SI	0.78	0.58
		BK	0.67	0.59
		BE	0.59	0.71
		KOKH	13.7	9.60
	OBCO	SI	0.93	0.95
		BK	0.44	0.51
		BE	0.44	0.32
		KOKH	0.27	0.23
	DOBC	SI	0.94	0.96
		BK	0.51	0.64
		BE	0.19	0.14
		KOKH	11.7	1.48
Vehicle	BCO	SI	0.64	0.70
		BK	0.70	0.76
		BE	0.47	0.39
		KOKH	2.78	1.91
	BCO2	SI	0.60	0.67
		BK	0.53	0.66
		BE	0.67	0.52
		KOKH	3.13	3.19
	OBCO	SI	0.61	0.70
		BK	0.45	0.45
		BE	0.64	0.60
		KOKH	0.79	0.38
	DOBC	SI	0.64	0.71
		BK	0.14	0.23
		BE	0.35	0.31
		KOKH	3.02	4.64



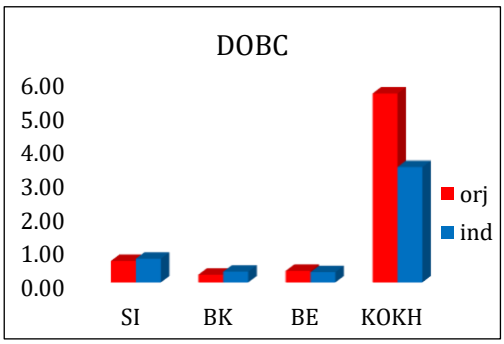
Şekil 9. BCO için ortalama metrik değerleri



Şekil 10. BCO2 için ortalama metrik değerleri



Şekil 11. OBCO için ortalama metrik değerleri



Şekil 12. DOBC için ortalama metrik değerleri

4 Sonuç

Bu çalışmada boyut indirgemenin, BCO, BCO2, OBCO ve DOBC kümeleme yöntemleri üzerindeki etkisi incelenmektedir. Elde edilen sonuçlara göre boyut indirgeme, genel olarak bu yaklaşımlar üzerinde kümeleme başarısını arttırmaktadır. Sadece BCO2 için SI değeri, azalma eğilimi göstermektedir. Dört veri seti üzerinde elde edilen ortalama değerlere göre, BCO için SI, 0.64'ten 0.71'e; BK, 0.71'den 0.77'ye yükselmekte, BE,

0.45'ten 0.37'ye; KOKH, 3.27'den 2.44'e düşmektedir. BCO2 için SI, 0.58'den 0.56'ya düşmekle beraber, BK değeri 0.59'dan 0.60'a yükselmektedir. BE değeri değişmemekte, KOKH değeri 6.40'tan 5.98'e düşmektedir. OBCO için SI, 0.34'ten 0.51'e; BK, 0.56'dan 0.57'ye yükselmekte, BE, 0.57'den 0.52'ye; KOKH 1.28'den 0.88'e düşmektedir. DOBC için SI, 0.64'ten 0.71'e; BK, 0.23'ten 0.32'ye yükselmekte, BE 0.35'ten 0.31'e, KOKH, 5.60'dan 3.43'e düşmektedir.

Kaynaklar

[1] Dunn JC. "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". *Journal of Cybernetics*, **3**(3), 32–57,1973.

[2] Bezdek JC. Pattern recognition with fuzzy objective function algorithms, Plenum, NY,1981.

[3] Rhee FCH, Hwang C. "A type-2 fuzzy C-means clustering algorithm". *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*. Vol. 4. IEEE, 2001.

[4] Gosain A, Dahiya S. "Performance analysis of various fuzzy clustering algorithms: a review". *Procedia Computer Science* 79, 100-111, 2016.

[5] Pal NR, Pal K, Keller JM, Bezdek JC. "A possibilistic fuzzy c-means clustering algorithm". *IEEE transactions on fuzzy systems*, *13*(4), 517-530, 2005.

[6] Krishnapuram R, Keller JM. "A possibilistic approach to clustering". *IEEE transactions on fuzzy systems*, *1*(2), 98-110, 1993.

[7] Wu X, Wu B, Sun J, Fu H. "Unsupervised possibilistic fuzzy clustering". *Journal of Information & Computational Sci.*, 7 (5), 1075-1080, 2010.

[8] Yang MS, Wu KL. "Unsupervised possibilistic clustering". *Pattern Recognition*, *39*(1), 5-21,2006.

[9] Gosain A, Dahiya S. "Performance analysis of various fuzzy clustering algorithms: a review". *Procedia Computer Science*, *79*, 100-111, 2016.

[10] Eschrich S, Ke J, Hall LO, Goldgof DB. "Fast accurate fuzzy clustering through data reduction". *IEEE transactions on fuzzy systems*, *11*(2), 262-270, 2003.

[11] Lee, KY. "Local fuzzy PCA based GMM with dimension reduction on speaker identification". *Pattern recognition letters*, *25*(16), 1811-1817,2004.

[12] Karami A. "Application of fuzzy clustering for text data dimensionality reduction". *arXiv preprint arXiv:1909.10881*, 2019.

[13] Yildiz K, Çamurcu AY, Dogan B. "Comparison of dimension reduction techniques on high dimensional datasets". *Int. Arab J. Inf. Technol.*, *15*(2), 256-262, 2018.

[14] Winkler R, Klawonn F, Kruse R. " Problems of fuzzy c-means clustering and similar algorithms with high dimensional data sets". *Challenges at the Interface of data analysis, computer science, and optimization* (pp. 79-87), Springer, Berlin, Heidelberg, 2012.

[15] Halko N, Martinsson PG, Tropp JA. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions,arXiv:909,https://arxiv.org/pdf/0909.4061.pdf, 2009.

[16] Golub *GH,* Reinsch C. *"*Singular value decomposition and least squares solutions*". Numerische Mathematik. **14** (5),403–42, 1970.*

[17] Wei JJ et al. "ECG data compression using truncated singular value decomposition". *IEEE Transactions on Information Technology in Biomedicine* 5(4), 290-299,2001.

[18] Kumar R, Kumar A, Singh GK. "Hybrid method based on singular value decomposition and embedded zero tree wavelet technique for ECG signal compression". *Computer methods and programs in biomedicine*, 129, 135-148,2016.

[19] Asuncion A, Newman D. UCI Machine Learning Repository,https://archive.ics.uci.edu/ml/index.php(10.05.2020)

[20] Rousseeuw PJ. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of computational and applied mathematics* 20, 53-65, 1987.

[21] Bezdek JC. "Numerical taxonomy with fuzzy sets". *Journal of Mathematical Biology*, *1*(1), 57-71, 1974

[22] Bezdek JC. "Cluster validity with fuzzy sets", J. Cybern, 3, 58–78, 1974.

[23] Pedregosa *et al.*, Scikit-learn: Machine Learning in Python, JMLR 12, 2825-2830, 2011.

[24] R Foundation for Statistical. https://www.R-project.org (15.05.2020).