Erciyes Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 36, Sayı 3 , 2020

Erciyes University
Journal of Institue Of Science and
Technology
Volume 36, Issue 3, 2020

Çok Değişkenli Normal Dağılımların Karmasına Dayalı Kümelemede TOPSIS Yöntemi ile Küme Sayısının Belirlenmesi

Serkan AKOĞUL [1]*, Murat ERİŞOĞLU[2]*, Ülkü ERİŞOĞLU[3]*

*[1] Pamukkale Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, DENİZLİ
[2] Necmettin Erbakan Üniversitesi Fen Fakültesi, İstatistik Bölümü, KONYA
[3] Necmettin Erbakan Üniversitesi Fen Fakültesi, İstatistik Bölümü, KONYA

(Alınış / Received: 06.03.2020, Kabul / Accepted: 19.11.2020, Online Yayınlanma / Published Online: 31.12.2020)

Anahtar Kelimeler
Bilgi Kriterleri,
Çok kriterli Karar Verme,
Modele Dayalı Kümeleme,
TOPSIS

Öz: Çok kriterli karar verme yöntemleri, birden fazla kriterin optimizasyonu ile mümkün çözüm kümeleri içerisinden alternatifin seçimi, sıralanması ve sınıflanmasını sağlar. Bu çalışmanın amacı, Çok kriterli karar verme yöntemlerinden birisi olan TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) ile modele dayalı kümeleme analizinde küme sayısını belirlemektir. Çalışmada, veri setleri aday küme sayılarına göre modele dayalı kümeleme ile modellenmiş ve elde edilen her bir kümeleme için Akaike bilgi kriteri, kanıtların yaklaşık ağırlık kriteri, Bayesci bilgi kriteri, sınıflandırma olabilirlik kriteri ve Kullback bilgi kriteri birer karar kriteri olarak hesaplanmıştır. Kriterlerin ağırlıklandırılmasın da simülasyon sonuçları kullanılmış olup TOPSIS ile veri seti için en uygun küme sayısı belirlenmiştir. Önerilen yaklaşımın başarısı gerçek ve sentetik veri setleri üzerinde test edilmiştir. Uygulama sonucunda uygun küme sayısının belirlenmesinde önerilen yaklaşım ilgili bilgi kriterlerine göre daha başarılı bir performans göstermiştir.

Determining the Number of Clusters with the TOPSIS Method in Clustering Based on the Multivariate Mixtures of Normal Distributions

Keywords
Information Criteria,
Multi Criteria Decision Making,
Model Based Clustering,
TOPSIS

Abstract: Multiple criteria decision making methods provide the selection, ordering and classification of the alternative among the possible solution sets with the optimization of multiple criteria. The aim of this study is to determine the number of clusters in model-based cluster analysis with the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution), which is one of the multiple criteria decision making methods. In the study, the data sets were modeled with model-based clustering according to the number of candidate clusters, and for each cluster obtained, the Akaike information criterion, the approximate weight criterion of evidence, Bayesian information criterion, classification likelihood criterion and Kullback information criterion is calculated as a decision criterion. Simulation results were used in weighting the criteria and the most suitable number of clusters was determined with the TOPSIS. The success of the proposed approach was tested on the real and the synthetic datasets. As a result of the application, the proposed approach in determining the appropriate cluster number performed better than the relevant information criteria.

*İlgili Yazar, email: sakogul@pau.edu.tr

1. Giriş

Veri setindeki küme sayısının ve yapısının belirlenmesi kümeleme analizindeki en önemli problemlerden birisidir. Modele dayalı kümeleme, veri setini oluşturan kümelerin (bileşenlerin) her birinin tek ya da çok değişkenli dağılımlar ile modellenmesidir. Modele dayalı kümeleme analizi son zamanlarda kullanımı yaygınlaşmakta olup hem küme sayısı için yeni yöntemler önerilmekte hem de veri yapısının belirlenmesine ve verinin modellenmesine olanak sağlamaktadır. Bu nedenle tıp, mühendislik, zooloji, su ürünleri, ekonomi ve psikoloji gibi çok sayıda uygulama alanına sahiptir [1].

Karma dağılım fikrini ilk ortaya atan Pearson [2] çalışmasında, birbirinden farklı ortalama ve varyansa sahip tek değişkenli iki bileşenli karma dağılım modelini incelemiştir. Karma dağılımı kümeleme analizinde ilk kullananlar ise Wolfe [3,4], Day [5] ve Binder [6] olmuştur. Dempster ve ark. [7] tarafından önerilen EM (Expectation–Maximization) algoritması, karma dağılım modelindeki parametre tahminlerinde kolaylık sağlamış olup literatürde sıkça karşımıza çıkmaktadır. Modele dayalı kümeleme analizinde, sonlu normal dağılımların karmasına dayalı kümeleme ise yaygın olarak kullanılan bir yaklaşımdır [8-25]. Bu yaklaşım yaygın olarak kullanılmasına rağmen, veri setini yeterince temsil etmek için kaç tane bileşenin gerekli olduğu kararı, birçok araştırmacılara göre önemli bir sorundur ancak tatmin edici istatistiksel bir çözüm de mevcut değildir [26,27].
Modele dayalı kümelemede küme sayısını tahmin etmek için bilgi kriterleri yaygın olarak kullanılmıştır. Bu bilgi kriterleri bir veri setinin küme sayısı tahmininde farklı sonuçlar verebilmektedir. Bu sorunun üstesinden gelmek için, küme sayısı belirleme problemi çok kriterli karar verme (ÇKKV) problemi olarak ele alınmıştır. Modele dayalı kümeleme sonucunda hesaplanan bilgi kriterleri ile bir karar matrisi oluşturulmuştur. Akogul ve Erisoglu'nun [28] çalışmasındaki bilgi kriterlerinin gerçek ve simülasyon verilerine dayalı başarı karşılaştırması sonuçları kriterler için ağırlıklar olarak alınmıştır. Daha sonra ÇKKV yöntemlerinden biri olan TOPSIS ile küme sayısı tahmininde bulunulmuştur.

2. Materyal ve Metot

2.1. Modele dayalı kümeleme analizi

Modele dayalı kümeleme analizinde, veri setinin aynı ya da farklı dağılımlara sahip çeşitli alt kümelerden oluştuğu varsayılmaktadır. Tüm veri seti, bu dağılımların bir karması ile modellenir. p boyutlu n gözleme sahip bir rasgele örneklem $\boldsymbol{y} = (\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_n^T)^T$ olmak üzere sonlu karma dağılım modelinin olasılık yoğunluk fonksiyonu

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \Phi_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i) \tag{1}$$

şeklinde ifade edilir ($j = 1, \dots, n$ ve $i = 1, \dots, g$). Burada g modeldeki bileşen sayısı olmak üzere π_i, $0 < \pi_i < 1$ ve $\sum_{i=1}^{g} \pi_i = 1$ olan i'inci bileşenin karma oranı ve $\Phi_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$ de karma dağılımın i'inci olasılık yoğunluk fonksiyonunu ifade eder. $\boldsymbol{\Psi} = \left(\pi_1, \dots, \pi_{g-1}, \boldsymbol{\xi}^T\right)^T$, karma modeldeki tüm bilinmeyen parametrelerin bir vektörü olup $\boldsymbol{\xi}$ de $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_g$ bileşen olasılık yoğunluk fonksiyonlarının bilinmeyen parametrelerini içeren bir vektördür [12].
Çok boyutlu veri setlerinin kümelenmesinde bileşenlere ait olasılık yoğunluk fonksiyonu olarak çok değişkenli normal dağılımın olasılık yoğunluk fonksiyonu

$$\Phi_i(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{p}{2}} |\Sigma_i|^{-\frac{1}{2}} e^{\left\{-\frac{1}{2}(y_j-\mu_i)^T \Sigma_i^{-1} (y_j-\mu_i)\right\}} \tag{2}$$

yaygın olarak kullanılmaktadır. $\boldsymbol{\mu}_i$ 'inci bileşene ait ortalama vektörü ve $\boldsymbol{\Sigma}_i$ 'inci bileşene ait varyans-kovaryans matrisi olmak üzere $\boldsymbol{\Psi} = \left(\pi_1, , \dots, \pi_{g-1}, \mu_1, \mu_2, \dots, \mu_g, \Sigma_1, \Sigma_2, \dots, \Sigma_g\right)^T$ karma modeldeki tüm bilinmeyen parametre vektörüdür. O halde $\boldsymbol{\Psi}$ vektörünün log-olabilirlik fonksiyonu

$$logL(\boldsymbol{\Psi}) = \sum\nolimits_{j=1}^{n} log\left[\sum_{i=1}^{g} \pi_i \Phi_i(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right] \tag{3}$$

şeklindedir. Gözlenen $\boldsymbol{y}_j$'nin karma modelin i'inci bileşenine ait olma olasılığı

$$\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \pi_i \Phi_i(\boldsymbol{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \Big/ \sum\nolimits_{m=1}^{g} \pi_m \Phi_i(\boldsymbol{y}_j; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{4}$$

olup $\boldsymbol{\Psi}$ parametresinin EM algoritmasıyla hesaplanan en çok olabilirlik tahminleri

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)})}{n} \tag{5}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{1}{\sum_{j=1}^{n} \tau_i\left(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}\right)} \sum_{j=1}^{n} \tau_i\left(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)}\right) \boldsymbol{y}_j \tag{6}$$

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{1}{\sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)})} \sum_{j=1}^{n} \tau_i(\boldsymbol{y}_j; \boldsymbol{\Psi}^{(k)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\boldsymbol{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T \quad (7)$$

iterasyonları ile elde edilir [12, 28].
Modele dayalı kümeleme analizinde en önemli problemlerden birisi uygun bileşen yani küme sayısının belirlenmesidir. Uygun küme sayısının belirlenmesinde ise genellikle log-olabilirlik fonksiyonunu temel alan bilgi kriterleri kullanılmaktadır. Bilgi kriterleri genel olarak

$$-2logL(\widehat{\boldsymbol{\Psi}}) + C \quad (8)$$

şekilde ifade edilir. Burada C, ceza terimi olarak adlandırılır ve modeldeki serbest parametre sayısı (d) veya veri setindeki gözlem sayısına (n) bağlı olarak belirlenir. Eşitlik (8) verilen bilgi kriteri için i bileşen sayısını göstermek üzere $Bilgi\ Kriteri(i) \leq Bilgi\ Kriteri(i+1)$ koşulunu sağlayan ilk i sayısı, uygun küme sayısı (g) olarak seçilir [12, 28]. Bu çalışmada Akaike bilgi kriteri (AIC) [29], kanıtların yaklaşık ağırlık kriteri (AWE) [30], Bayesci bilgi kriteri (BIC) [31], sınıflandırma olabilirlik kriteri (CLC) [32] ve Kullback bilgi kriteri (KIC) [33] kullanılmıştır. Bu bilgi kriterleri

$$\text{AIC} = -2logL(\widehat{\boldsymbol{\Psi}}) + 2d \quad (9)$$

$$\text{AWE} = -2logL_c(\widehat{\boldsymbol{\Psi}}) + 2d(3/2 + log(n)) \quad (10)$$

$$\text{BIC} = -2logL(\widehat{\boldsymbol{\Psi}}) + dlog(n) \quad (11)$$

$$\text{CLC} = -2logL(\widehat{\boldsymbol{\Psi}}) + 2EN(\hat{\tau}) \quad (12)$$

$$\text{KIC} = -2logL(\widehat{\boldsymbol{\Psi}}) + 3(d+1) \quad (13)$$

eşitlikleri ile hesaplanır. Burada $EN(\tau) = -\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij} \log \tau_{ij}$ olup $logL_c(\Psi) = logL(\Psi) + EN(\tau)$ dir [28].

2.2. TOPSIS yöntemi

Çok kriterli karar verme (ÇKKV) problemleri, birden fazla kriterin optimizasyonu ile mümkün çözüm kümeleri içerisinden en iyi alternatifin belirlendiği problemlerdir [34-37]. ÇKKV yöntemlerinden birisi olan TOPSIS (Technique for Order Preference by Similarity to Ideal Solution), Yoon ve Hwang [38] tarafından geliştirilmiştir. TOPSIS yönteminde, seçilen bir alternatifin kriterlere göre ideal çözüme yakın, ideal olmayan çözüme ise uzak olması amaçlanmaktadır. TOPSIS altı adımdan oluşan bir çözüm sürecini içerir ve sırası ile bu adımlar aşağıda verilmiştir [38-42]:

Adım 1. *Karar matrisinin oluşturulması*: $D = \left[d_{ij}\right]_{mxk}$ karar matrisi, karar verici tarafından oluşturulan mxk boyutlu bir matristir. Matrisin sütunlarında karar vermede kullanılacak k adet değerlendirme kriterleri, satırlarında ise m adet alternatifin karar kriterlerine göre değerlendirmesi yer alır. Burada d_{ij} elemanı, i'inci alternatifin ($i = 1, \dots, m$) j'inci kritere ($j = 1, \dots, k$) göre önem seviyesini belirtir.

Adım 2. *Normalize matrisin elde edilmesi:* $N = \left[n_{ij}\right]_{mxk}$ normalize matrisin elemanları, D karar matrisi kullanılarak $n_{ij} = d_{ij} \Big/ \sqrt{\sum_{i=1}^{m} d_{ij}^2}$ şeklinde hesaplanır.

Adım 3. *Ağırlıklandırılmış normalize matrisin elde edilmesi:* Normalize edilmiş matrise ait her bir değer w_j gibi bir değerle ağırlıklandırılır. Ağırlıklandırma işlemi TOPSIS yönteminin tek sübjektif aşamasıdır. Burada $\sum_{j=1}^{k} w_j = 1$ olup N matrisinin her bir sütunundaki elemanlar ilgili w_j değeri ile çarpılarak ($v_{ij} = w_j n_{ij}$) ağırlıklandırılmış normalize matris $V = \left[v_{ij}\right]_{mxk}$ elde edilir.

Adım 4. *İdeal ve negatif ideal çözümlerin elde edilmesi:* Problemin amacı doğrultusunda, ağırlıklandırılmış normalize matrisin (V) sütun değerleri göz önünde bulundurulur. Eğer amacımız minimizasyon ise sütuna ait minimum değerler ideal çözümleri (A^+), maksimum değerler ise negatif ideal çözümleri (A^-) oluşturur. Bu durumda $A^+ = \{v_1^+, v_2^+, \dots, v_k^+\}$ her bir sütuna ait minimum değerler ve $A^- = \{v_1^-, v_2^-, \dots, v_k^-\}$ her bir sütuna ait

maksimum değerler şeklinde belirlenir. Eğer amacımız maksimizasyon ise elde edilen değerler tam tersi olarak alınacaktır.

Adım 5. *İdeal ve negatif ideal ayırım ölçülerinin hesaplanması:* İdeal ve negatif ideal noktalara olan uzaklık değerleri hesaplanırken Öklid uzaklığı kullanılmaktadır. İdeal ayırım $S_i^+ = \sqrt{\sum_{j=1}^{k}\left(v_{ij} - v_j^+\right)^2}$ ve negatif ideal ayırım $S_i^- = \sqrt{\sum_{j=1}^{k}\left(v_{ij} - v_j^-\right)^2}$ ölçüleri şeklinde hesaplanmaktadır. Burada hesaplanacak S_i^+ ve S_i^- ölçülerinin sayısı alternatiflerin sayısı kadar olacaktır.

Adım 6. *İdeal çözüme göreli yakınlığın hesaplanması:* Her alternatifin ideal çözüme göreli yakınlığının (C_i^+) hesaplanmasında ideal ve negatif ideal ayırım ölçülerinden yararlanılır. İdeal çözüme göreli yakınlık değeri $C_i^+ = S_i^-/(S_i^- + S_i^+)$ şeklinde hesaplanır. Burada C_i^+ değeri $0 \leq C_i^+ \leq 1$ aralığında değer alır ve $C_i^+ = 1$ ise ilgili alternatifin ideal çözüme, $C_i^+ = 0$ ise ilgili alternatifin negatif ideal çözüme mutlak yakınlığını gösterir.

2.3. Modele dayalı kümeleme analizinde küme sayısının belirlenmesi için yeni bir yaklaşım

Modele dayalı kümeleme analizinde veri setinin küme sayısının belirlenmesinde bilgi kriterleri yaygın olarak kullanılmaktadır. Literatürdeki yaygın olarak kullanılan kriterlerden bazıları AIC, AWE, BIC, CLC, KIC vb. Bu bilgi kriterleri, aynı veri setinin küme sayısını farklı tahmin edebilmektedirler. Bu sorunun üstesinden gelebilmek için, modele dayalı kümelemede bir veri setinin küme sayısının belirlenmesi problemi bir ÇKKV problemi olarak ele alınmış ve TOPSIS yönteminden yararlanılarak yeni bir yaklaşım önerilmiştir. Şekil 1'de önerilen yaklaşım açıklanmıştır. Amaç; küme sayısının belirlenmesi, kriterler; AIC, AWE, BIC, CLC ve KIC bilgi kriterleri, alternatifler ise küme sayısı alınarak model oluşturulmuştur. Önerilen yaklaşım küme sayısı bilinen gerçek ve sentetik veri setleri üzerinde test edilmiş ve bu beş bilgi kriterlerinden daha iyi sonuçlar elde edilmiştir.

Şekil 1. Küme sayısının belirlenmesi için önerilen yaklaşım

Önerilen yaklaşım aşağıdaki adımlarda özetlenmiştir:

Adım 1. *Problemin tanımı:* Problemimiz bir veri setinin küme sayısının belirlenmesidir. Problemi çözmek için ele alınan kriterler, küme sayısı tahmininde kullanılan bilgi kriterlerinden AIC, AWE, BIC, CLC ve KIC'dan oluşmaktadır. Kriterlerin ağırlık vektörünü oluşturmak için Akogul ve Erisoglu'nun[28] çalışmasının bir sonucu olan kriterlerin küme sayısı tahmini ortalama başarıları kullanılmıştır. Böylece AIC, AWE, BIC, CLC ve KIC kriterlerine ait sırasıyla 43.6, 21.2, 47.4, 17.3 ve 58.2 olan ortalama başarıları $W = (w_1, w_2, w_3, w_4, w_5) = \{0.2323,\ 0.1129,\ 0.2525,\ 0.0922,\ 0.3101\}$ şeklinde ağırlıklara dönüştürülmüştür. Veri setinin küme sayısı alternatifleri ise 2, 3, 4 ve 5 olarak seçilmiştir. Alternatiflerin değerleri bilgi kriterlerinin en düşük ve en yüksek tahmin değerleri göz önünde bulundurularak belirlenmiştir. Benzer şekilde alternatif sayısı arttırılabilir.

Adım 2. *Modele dayalı kümeleme:* Veri seti, modele dayalı kümeleme ile farklı küme sayılarına göre çok değişkenli normal dağılımlarının karması ile modellenmiştir. Model parametrelerinin en çok olabilirlik tahminleri EM algoritması ile elde edilmiştir.

Adım 3. *Bilgi kriterleri:* Tahmin edilen parametrelere göre log-olabilirlik fonksiyonları hesaplanmış ve her bir küme sayısı için bilgi kriterleri değerleri elde edilmiştir. Bu değerler kullanılarak bir karar matrisi oluşturulmuştur.

Adım 4. *TOPSIS yöntemi:* Oluşturulan karar matrisi kullanılarak problemimiz bir ÇKKV problemine dönüştürülmüştür. Böylece TOPSIS yöntemi kullanılarak her bir alternatifin ideal çözüme göreli yakınlığı hesaplanmıştır.

Adım 5. *Küme sayısı tahmini:* Son adım olarak da TOPSIS ile en uygun alternatif yani küme sayısı belirlenmiştir. Problemin amacı doğrultusunda ideal ve negatif ideal çözümler oluşturur. Bilgi kriterleri için minimum değeri veren model en iyi model olarak seçileceğinden amacımız minimizasyondur. Bu doğrultuda en yüksek ideal çözüme göreli yakınlığa sahip alternatif uygun küme sayısı olarak belirlenmiştir.

3. Bulgular

Önerilen yaklaşım, Tablo 1'de verilen farklı örneklem büyüklükleri (n), farklı değişken sayıları (p) ve farklı küme sayılarına (g) sahip gerçek ve sentetik veri setleri üzerinde test edilmiştir.

Tablo 1. İncelenen veri setlerinin örneklem büyüklükleri, değişken sayıları ve küme sayıları

Veri Setleri	Örneklem Büyüklüğü (n)	Değişken Sayısı (p)	Küme Sayısı (g)
Crab	200	5	2
Liver Disorders	345	6	2
Ionosphere	351	34	2
Chemical Diabetes	145	4	3
Iris	150	4	3
Wine	178	13	3
Ruspini	75	2	4
E.coli	336	7	4
Vehicle Silhouettes	846	18	4
Sentetik-1	1000	2	2
Sentetik-2	1000	2	3
Sentetik-3	1000	2	4

Gerçek veri setleri UCI Machine Learning Repository [43] web sitesinden alınmıştır. Sentetik veri setleri ise Tablo 2'de verilen karma oranlar, ortalama vektörlere ve kovaryans matrislere göre normal dağılımların karmasından üretilmiştir.

Tablo 2. Sentetik veri setlerinin karma oranları, ortalama vektörleri ve kovaryans matrisleri

	Karma Oranlar	Ortalama vektörleri	Kovaryans matrisleri
Sentetik-1	$\pi_1 = 1/2$	$\mu_1 = [2,4]^{\mathrm{T}}$	$\Sigma_1 = [1,0;\ 0,1]$
	$\pi_2 = 1/2$	$\mu_2 = [5,6]^{\mathrm{T}}$	$\Sigma_2 = [2,0;\ 0,0.5]$
Sentetik-2	$\pi_1 = 1/3$	$\mu_1 = [-1,2]^T$	$\Sigma_1 = [1,0;\ 0,1]$
	$\pi_2 = 1/3$	$\mu_2 = [1,1]^T$	$\Sigma_2 = [0.5,-0.7;\ -0.7,1.5]$
	$\pi_3 = 1/3$	$\mu_3 = [0,-4]^T$	$\Sigma_3 = [2,0;\ 0,2]$
Sentetik-3	$\pi_1 = 1/4$	$\mu_1 = [-2,-2]^T$	$\Sigma_1 = [0.2,0;\ 0,0.2]$
	$\pi_2 = 1/4$	$\mu_2 = [-2,-2]^T$	$\Sigma_2 = [3,2;\ 2,7]$
	$\pi_3 = 1/4$	$\mu_3 = [3,1]^T$	$\Sigma_3 = [1,0;\ 0,4]$
	$\pi_4 = 1/4$	$\mu_4 = [1,-3]^T$	$\Sigma_4 = [1,0;\ 0,1]$

Çalışmada Crab veri seti için önerilen yöntem detaylı olarak açıklanıp hesaplamalar verilmiş ve sonuçlar yorumlanmıştır. Diğer veri setleri için karar matrisi, ideal ve negatif ideal ayırımlar ile ideal çözüme göreli yakınlık değerleri verilmiştir. Karar matrisleri kullanılarak hesaplamalar benzer şekilde yapılabilir.
Crab veri setinin alternatif küme sayılarına göre modele dayalı kümeleme analizi sonucu hesaplanan bilgi kriterlerinin değerlerinden oluşan karar matrisi Tablo 3'de verilmiştir.

Tablo 3. Crab veri seti için karar matrisi

Alternatifler	AIC	AWE	BIC	CLC	KIC
2	2929.3	3474.2	3064.5	2916.8	2973.3
3	2920.6	3690.3	3125.1	2847.3	2985.6
4	2888.0	3915.7	3161.8	2787.2	2974.0
5	2816.7	4065.9	3159.7	2651.8	2923.7

Tablo 3'e göre 2 kümeye sahip Crab veri seti için AIC, AWE, BIC, CLC ve KIC bilgi kriterlerinin küme sayısı tahminleri sırasıyla 5, 2, 2, 5 ve 5 olduğu görülmektedir. Karar matrisi yardımıyla elde edilen normalize edilmiş matris Tablo 4'deki gibi hesaplanır.

Tablo 4. Crab veri seti için normalize edilmiş matris

Alternatifler	AIC	AWE	BIC	CLC	KIC
2	0.5070	0.4580	0.4898	0.5204	0.5015
3	0.5055	0.4864	0.4995	0.5080	0.5036
4	0.4998	0.5162	0.5054	0.4973	0.5016
5	0.4875	0.5359	0.5051	0.4731	0.4932

Normalize edilmiş matrisinin her bir sütunundaki elemanlar ilgili ağırlık vektörü $W = (w_1, w_2, w_3, w_4, w_5) = \{0.2323,\ 0.1129,\ 0.2525,\ 0.0922,\ 0.3101\}$ ile çarpılarak ağırlıklandırılmış normalize matris Tablo 5'deki gibi elde edilmiştir.

Tablo 5. Crab veri seti için ağırlıklandırılmış normalize matris

Alternatifler	AIC	AWE	BIC	CLC	KIC
2	0.1178	0.0517	0.1237	0.0480	0.1555
3	0.1174	0.0549	0.1261	0.0468	0.1562
4	0.1161	0.0583	0.1276	0.0458	0.1555
5	0.1132	0.0605	0.1275	0.0436	0.1529

Amacımız minimizasyon olup ağırlıklandırılmış normalize matrisin her bir sütuna ait minimum değerler ideal çözümleri $A^+ = \{0.1132,\ 0.0517,\ 0.1237,\ 0.0436,\ 0.1529\}$, maksimum değerler ise negatif ideal çözümleri $A^- = \{0.1178,\ 0.0605,\ 0.1276,\ 0.0480,\ 0.1562\}$ oluşturur. İdeal ayırım $S^+ = \{0.0068,\ 0.0074,\ 0.0089,\ 0.0096\}$ ve negatif ideal ayırım $S^- = \{0.0097,\ 0.0059,\ 0.0036,\ 0.0071\}$ ölçüleri şeklinde hesaplanmaktadır. Her alternatifin ideal çözüme göreli yakınlığı $C^+ = \{0.5871,\ 0.4442,\ 0.2863,\ 0.4238\}$ olup en yüksek C_i^+ değerine (0.5871) sahip olan 2, alternatifler içerisinde en uygun küme sayısı olarak belirlenmiştir. Böylece 2 kümeye sahip Crab veri setinin küme sayısı önerilen yöntem ile doğru olarak tespit edilmiştir. Tüm veri setleri için hesaplanan karar metrisleri Tablo 6'da verilmiştir.

Tablo 6. Her bir veri setinin karar matrisleri

Veri Setleri	Alternatifler	AIC	AWE	BIC	CLC	KIC
Liver Disorders	2	14752.4	15503.7	14963.8	14695.9	14810.4
	3	14693.7	15865.2	15012.8	14646.2	14779.7
	4	14605.7	16176.3	15032.4	14546.0	14719.7
	5	14643.7	16582.9	15177.9	14541.4	14785.7
Ionosphere	2	12260.8	28277.6	17121.6	9743.1	13522.8
	3	13487.6	37518.7	20780.7	9709.7	15379.6
	4	12065.5	44111.7	21790.9	7028.0	14587.5
	5	14236.6	54296.8	26394.2	7938.6	17388.6
Chemical Diabetes	2	6332.5	6663.0	6418.9	6287.3	6364.5
	3	6241.7	6744.5	6372.7	6174.6	6288.7
	4	6234.9	6924.2	6410.5	6160.0	6296.9
	5	6223.3	7065.4	6443.5	6106.8	6300.3
Iris	2	487.1	806.7	574.4	429.1	519.1
	3	449.1	944.1	581.6	371.2	496.1
	4	448.9	1126.5	626.5	358.3	510.9
	5	474.1	1378.8	696.9	415.2	551.1
Wine	2	5222.3	7597.3	5887.3	4804.3	5434.3
	3	4802.5	8382.5	5801.6	4186.3	5119.5
	4	4753.3	9517.9	6086.5	3918.5	5175.3
	5	4841.1	10797.3	6508.4	3794.8	5368.1
Ruspini	2	1409.2	1515.2	1434.7	1387.2	1423.2
	3	1369.9	1534.0	1409.3	1336.3	1389.9
	4	1329.9	1552.3	1383.2	1284.7	1355.9
	5	1322.5	1602.2	1389.7	1264.8	1354.5
E.coli	2	3377.5	4283.3	3648.6	3244.3	3451.5
	3	569.3	1929.7	977.7	363.9	679.3
	4	465.2	2279.0	1011.0	186.3	611.2
	5	514.4	2786.0	1197.7	166.5	696.4
Vehicle Silhouettes	2	80747.8	86249.1	82544.5	80002.8	81129.8
	3	78171.5	86431.4	80868.8	77053.7	78743.5
	4	76987.3	88003.1	80585.3	75493.9	77749.3
	5	77496.4	91282.8	81995.2	75642.2	78448.4
Sentetik-1	2	6849.0	7286.0	6903.0	7101.0	6863.0
	3	6846.5	7971.4	6929.9	7685.5	6866.5
	4	6862.9	8416.3	6975.8	8029.6	6888.9
	5	6860.6	8945.2	7002.9	8457.6	6892.6
Sentetik-2	2	5233.9	5634.1	5284.0	5457.0	5247.9
	3	5110.2	5711.6	5187.6	5437.9	5130.2
	4	5101.0	5945.1	5205.7	5574.8	5127.0
	5	5126.2	6478.0	5258.2	6011.1	5158.2
Sentetik-3	2	8298.5	8727.6	8352.4	8542.6	8312.5
	3	8268.1	9055.1	8351.5	8769.2	8288.1
	4	8082.1	8994.7	8195.0	8608.0	8108.1
	5	8084.3	9315.3	8226.6	8827.6	8116.3

Bilgi kriterleri için minimum değer en uygun kümesayısı olduğu için burada amacımız minimizasyondur. O halde ağırlıklandırılmış normalize matrisin her bir sütuna ait minimum değerler ideal çözümleri (A^+), maksimum değerler ise negatif ideal çözümleri (A^-) oluşturur. İdeal ve negatif ideal çözümleri kullanarak elde edilen ideal ayırım (S_i^+) ve negatif ideal ayırım (S_i^-) ölçüleri Tablo 7'de verilmiştir.

Tablo 7. Veri setlerinin her bir alternatife göre ideal ayırım ve negatif ideal ayırım değerleri

	İdeal ayırım (S_i^+)				**Negatif ideal ayırım (S_i^-)**			
Veri Setleri	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
L. Disorders	0.0016	0.0017	0.0024	0.0043	0.0042	0.0029	0.0025	0.0010
Ionosphere	0.0145	0.0363	0.0361	0.0777	0.0772	0.0450	0.0478	0.0096
C. Diabetes	0.0032	0.0009	0.0023	0.0036	0.0034	0.0040	0.0030	0.0029
Iris	0.0144	0.0075	0.0202	0.0431	0.0399	0.0383	0.0261	0.0036
Wine	0.0183	0.0066	0.0134	0.0258	0.0236	0.0258	0.0204	0.0144
Ruspini	0.0124	0.0067	0.0017	0.0032	0.0032	0.0063	0.0119	0.0122
E.coli	0.3644	0.0106	0.0070	0.0227	0.0000	0.3556	0.3626	0.3501
V.Silhouettes	0.0096	0.0028	0.0011	0.0042	0.0032	0.0075	0.0098	0.0076
Sentetik-1	0.0000	0.0059	0.0096	0.0141	0.0141	0.0082	0.0045	0.0000
Sentetik-2	0.0053	0.0008	0.0032	0.0095	0.0092	0.0100	0.0080	0.0037
Sentetik-3	0.0094	0.0024	0.0017	0.0030	0.0023	0.0076	0.0094	0.0089

Her veri seti için hesaplanan ideal ayırım ve negatif ideal ayırım ölçüleri kullanılarak elde edilen her bir alternatifin ideal çözüme göreli yakınlığı C_i^+, Tablo 8'de verilmiştir. En büyük C_i^+ değerine sahip olan i'inci alternatif kalın olarak yazılmış olup uygun küme sayısı olarak seçilmektedir.

Tablo 8. Veri setlerinin her bir alternatife göre ideal çözüme göreli yakınlık değerleri

	Alternatifler			
Veri Seti	**2**	**3**	**4**	**5**
L. Disorders	**0.7270**	0.6382	0.5017	0.1932
Ionosphere	**0.8417**	0.5532	0.5702	0.1097
C. Diabetes	0.5120	**0.8148**	0.5587	0.4455
Iris	0.7354	**0.8370**	0.5637	0.0776
Wine	0.5631	**0.7970**	0.6029	0.3591
Ruspini	0.2031	0.4861	**0.8778**	0.7915
E.coli	0.0000	0.9710	**0.9811**	0.9391
V. Silhouettes	0.2524	0.7263	**0.8970**	0.6452
Sentetik-1	**0.9969**	0.5842	0.3165	0.0028
Sentetik-2	0.6366	**0.9287**	0.7152	0.2792
Sentetik-3	0.1991	0.7597	**0.8494**	0.7482

Tablo 9'da tüm veri setlerinin bilgi kriterleri ve önerilen yaklaşımla belirlenen küme sayısı tahminleri verilmiştir. En alt satırda da, yaklaşımların doğru belirledikleri küme sayısı toplamları belirtilmiştir. Tablo 9'da önerilen yaklaşımın üstünlüğü görülmektedir. Böylece modele dayalı kümelemede önerilen bu yeni yaklaşım literatürde var olan ilgili beş bilgi kriterinden daha üstün bir performans sergilemiştir.

Tablo 9. Bilgi kriterleri ve önerilen yaklaşımın karşılaştırılması

Veri Seti	**Küme Sayısı**	**AIC**	**AWE**	**BIC**	**CLC**	**KIC**	**Önerilen Yaklaşım**
Crab	2	5	2	2	5	5	2
L. Disorders	2	4	2	2	5	4	2
Ionosphere	2	4	2	2	4	2	2
C. Diabetes	3	5	2	3	5	3	3
Iris	3	4	2	2	4	3	3
Wine	3	4	2	3	5	3	3
Ruspini	4	5	2	4	5	5	4
E.coli	4	4	3	3	5	4	4
V. Silhouettes	4	4	2	4	4	4	4
Sentetik-1	2	3	2	2	2	2	2
Sentetik-2	3	4	2	3	3	4	3
Sentetik-3	4	4	2	4	2	4	4
Toplam Doğru Sayısı		3	4	10	3	8	12

4. Tartışma ve Sonuç

Kümeleme analizinde küme sayısı tahmini için literatürde birçok yaklaşım önerilmiştir. Modele dayalı kümeleme analizinde yaygın olarak bilgi kriterleri kullanılmaktadır. Bu çalışmada, veri setleri modele dayalı kümeleme ile modellenmiş ve yaygın kullanılan AIC, AWE, BIC, CLC ve KIC bilgi kriterleri birer karar kriteri olarak alınmıştır. Böylece küme sayısı belirleme problemini çok kriterli karar verme yöntemlerinden TOPSIS ile çözülmesi amaçlanmıştır. Bu sayede çok değişkenli normal dağılımların karmasına dayalı kümelemede bir veri setinin küme sayısının belirlenmesi için yeni bir yaklaşım önerilmiştir. Önerilen yaklaşımın bu beş ilgili bilgi kriterlerinden daha doğru sonuçlar ürettiği görülmüştür.
Bu çalışma ile veri setinin küme sayısı belirleme problemi bir ÇKKV problemine dönüştürülmüş ve bazı bilgi kriterlerinin ortak kullanılması ile daha iyi sonuçlar elde edilebileceği gösterilmiştir. Daha sonraki çalışmalar için araştırmacılar bu yöntemi, farklı kriterler ve ağırlık matrisleri belirleyerek hem modele dayalı kümelemede hem de diğer kümeleme yöntemlerinde kolaylıkla uygulayabilirler.

Kaynakça

[1] Titterington, D.M., Smith, A. F., Makov, U. E. 1985. Statistical analysis of finite mixture distributions, Wiley.

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

[3] Wolfe, J. H. 1965. A computer program for the maximum likelihood analysis of types, Naval Personnel Research Activity San Diego Calif.

[4] Wolfe, J. H. 1967. NORMIX: Computational methods for estimating the parameters of multivariate normal mixtures of distributions, Naval Personnel Research.

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

[6] Binder, D. A. 1978. Bayesian cluster analysis, Biometrika, 65 (1), 31-38.

[7] Dempster, A. P., Laird, N. M., Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm, Journal of the royal statistical society. Series B (methodological), 1-38.

[8] Celeux, G., Govaert, G. 1993. Comparison of the mixture and the classification maximum likelihood in cluster analysis, Journal of Statistical Computation and simulation, 47 (3-4), 127-146.

[9] Celeux, G., Govaert, G. 1995. Gaussian parsimonious clustering models, Pattern recognition, 28 (5), 781-793.

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

[11] Biernacki, C., Celeux, G., Govaert, G. 2003. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics & Data Analysis, 41(3-4), 561-575.

[12] McLachlan, G., Peel, D. 2004. Finite mixture models, John Wiley & Sons.

[13] Pernkopf, F., Bouchaffra, D. (2005). Genetic-based EM algorithm for learning Gaussian mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1344-1348.

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

[15] Browne, R. P., McNicholas, P. D. 2012. Model-based clustering, classification, and discriminant analysis of data with mixed type. Journal of Statistical Planning and Inference, 142(11), 2976-2984.

[16] Yang, M. S., Lai, C. Y., Lin, C. Y. 2012. A robust EM clustering algorithm for Gaussian mixture models. Pattern Recognition, 45(11), 3950-3961.

[17] Lee, S. X., McLachlan, G. J. 2013. Model-based clustering and classification with non-normal mixture distributions. Statistical Methods & Applications, 22(4), 427-454.

[18] Kwedlo, W. 2015. A new random approach for initialization of the multiple restart EM algorithm for Gaussian model-based clustering. Pattern Analysis and Applications, 18(4), 757-770.

[19] Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B. 2016. Model-based clustering based on sparse finite Gaussian mixtures. Statistics and computing, 26(1-2), 303-324.

[20] Marbac, M., Biernacki, C., Vandewalle, V. 2017. Model-based clustering of Gaussian copulas for mixed data. Communications in Statistics-Theory and Methods, 46(23), 11635-11656.

[21] Fop, M., Murphy, T. B. 2018. Variable selection methods for model-based clustering. Statistics Surveys, 12, 18-65.

[22] Scrucca, L., Raftery, A. E. 2018. clustvarsel: a package implementing variable selection for Gaussian model-based clustering in R. Journal of Statistical Software, 84.

[23] Celeux, G., Maugis-Rabusseau, C., Sedki, M. 2019. Variable selection in model-based clustering and discriminant analysis with a regularization approach. Advances in Data Analysis and Classification, 13(1), 259-278.

[24] Wei, Y., Tang, Y., McNicholas, P. D. 2019. Mixtures of generalized hyperbolic distributions and mixtures of skew-t distributions for model-based clustering with incomplete data. Computational Statistics & Data Analysis, 130, 18-41.

[25] Gögebakan, M., Servi, T. 2019. Genetik Algoritma Kullanılarak Verilerin Karma Normal Modele Dayalı Kümelenmesi. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi, 35(3), 12-23.

[26] Bozdogan, H. 1994. Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity, Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach, 69-113.

[27] Oliveira-Brochado, A., Martins, F. V. 2005. Assessing the number of components in mixture models: a review, Universidade do Porto, Faculdade de Economia do Porto.

[28] Akogul, S., Erisoglu, M. 2016. A comparison of information criteria in clustering based on mixture of multivariate normal distributions. Mathematical and Computational Applications, 21(3), 34.

[29] Akaike, H. 1998. Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike (pp. 199-213). Springer, New York, NY.

[30] Banfield, J. D., Raftery, A. E. 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics, 803-821.

[31] Schwarz, G. 1978. Estimating the dimension of a model. The annals of statistics, 6(2), 461-464.

[32] Biernacki, C., Govaert, G. 1997. Using the classification likelihood to choose the number of clusters. Computing Science and Statistics, 451-457.

[33] Cavanaugh, J. E. 1999. A large-sample model selection criterion based on Kullback's symmetric divergence. Statistics & Probability Letters, 42(4), 333-343.

[34] Yıldırım, B. F., Önder, E., Turan, G. 2015. Operasyonel, Yönetsel ve Stratejik Problemlerin Çözümünde Çok Kriterli Karar Verme Yöntemleri. Dora Yayıncılık, 2, 15.

[35] Akogul, S., Erisoglu, M. 2017. An approach for determining the number of clusters in a model-based cluster analysis. Entropy, 19(9), 452.

[36] Özdemir, B., Özcan, B., Aladağ, Z. 2017. Güneş enerjisi santrali kuruluş yerinin AHS ve VIKOR yöntemlerine dayalı bütünleşik yaklaşım ile değerlendirilmesi. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi, 33(2), 16-34.

[37] Zararlı, F., Yazgan, H. R., Delice, Y. 2018. AHP ve VIKOR Bütünleşik yaklaşımıyla Lojistik Merkez Yer Seçimi: Kayseri ili örneği. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi, 34(3), 1-9.

[38] Hwang, C. L., Yoon, K. 1981. Methods for multiple attribute decision making. In Multiple attribute decision making (pp. 58-191). Springer, Berlin, Heidelberg.

[39] Lai, Y.-J., Liu, T.-Y., Hwang, C.-L. 1994. Topsis for MODM, European journal of operational research, 76 (3), 486-500.

[40] Wang, T. C., Lee, H. D. 2009. Developing a fuzzy TOPSIS approach based on subjective weights and objective weights. Expert systems with applications, 36(5), 8980-8985.

[41] Ishizaka, A., Nemery, P. 2013. Multi-criteria decision analysis: methods and software, John Wiley & Sons, p.

[42] Akman, G., Özcan, B., Başlı, H., Gündüz, E. B. 2018. Çok Kriterli Karar Vermede AHP ve TOPSIS Yöntemleriyle Uçuş Noktası Seçimi. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi, 34(3), 45-57.

[43] Lichman, M. 2013. UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml: (Erişim Tarihi: 06.06.2016).