Pala, O., Aksaraylı, M. / Journal of Yasar University, 2020, 15/58, 236-245

# Kümeleme İçin Değiştirilmiş Dunn İndeksi İle Bir Parçacık Sürü Optimizasyon Yaklaşımı[1]

## A Particle Swarm Optimization Approach For Clustering With A Modified Dunn Index

Osman PALA, Karamanoğlu Mehmetbey Üniversitesi, Türkiye, osman.pala@deu.edu.tr
Orcid No: 0000-0002-2634-2653
Mehmet AKSARAYLI, Dokuz Eylül Üniversitesi, Türkiye, mehmet.aksarayli@deu.edu.tr
Orcid No: 0000-0003-1590-4582

*Öz: Kümeleme analizi, gözlem gruplarını ortak özelliklerine göre kümelere bölümlemek olarak tanımlanmaktadır. Sağlık alanında hastalıkların teşhisi, mühendislikte ürün kusur tespiti ve işletmelerde müşteri segmentasyonu kümelemenin gerçek hayatta uygulama alanlarından birkaçıdır. Kümelemede ön bilgi olmadığı takdirde problem sıklıkla sezgisel algoritmalar kullanılarak çözülmektedir. Çalışmada, önerilen yeni bir uygunluk fonksiyonu ile Parçacık Sürü Optimizasyonu kümeleme probleminin çözümünde kullanılmıştır. Önerilen Değiştirilmiş Dunn İndeksi, literatürde yer alan kümeleme uygunluk fonksiyonları ile kümeleme doğruluğu açısından karşılaştırılmıştır. Öte yandan kullanılan Parçacık Sürü Optimizasyonu yöntemi, Genetik Algoritma ve Rassal Arama yöntemleri ile kümeleme analizinde kıyaslanmıştır. Kümeleme analizi alanında kullanılan beş adet veri seti üzerinde analizler gerçekleştirilmiştir. Elde edilen analiz sonuçları ve yapılan istatistiki testler, önerilen DDI uygunluk fonksiyonunun kümeleme doğruluğu açısından başarılı olduğunu göstermektedir.*

*Anahtar Sözcükler: Parçacık Sürü Optimizasyonu, Kümeleme, Uygunluk Fonksiyonu*
*JEL Sınıflandırması: C38, C44, C61*

*Abstract: Cluster analysis is defined as the division of observation groups into clusters according to their common characteristics. Diagnosis of diseases in the field of health, product defect detection in engineering and customer segmentation in enterprises are some of the real life applications of clustering. If there is no prior knowledge in clustering, the problem is often solved by using heuristic algorithms. In this study, Particle Swarm Optimization with a proposed new fitness function is used in the solution of clustering problem. The proposed Modified Dunn Index was compared with the clustering fitness functions in the literature for clustering accuracy. On the other hand, the Particle Swarm Optimization method was compared with the Genetic Algorithm and Random Search methods in clustering analysis. Analysis was performed on five data sets used in the field of cluster analysis. The results of the analysis and the conducted statistical tests indicate that the proposed DDI fitness function is successful in terms of clustering accuracy.*

*Keywords: Particle Swarm Optimization, Clustering, Fitness Function*
*JEL Classification: C38, C44, C61*

## 1. Giriş

Kümeleme problemi bir çok disiplinde önemlidir. Örneğin bir işletmenin ürünlerine yönelik hedef müşteri segmentasyonun belirlenmesi ve buna göre planlama ve pazarlama önemli bir husustur (Fidan, 2009: 2153). Kümeleme analizi ile hem geçerli müşterilere göre segmentasyon hem de yeni müşterileri özelliklerine göre var olan kümelere yerleştirme işlemi yapılabilmektedir.

Kümeleme bir veri setinde bulunan elemanları veya verileri değişkenlik içeren özellik değerlerini dikkate alarak kümelere atama işlemidir. Aynı kümedeki elemanlar arasında özellik değerlerine göre yakınlık bulunurken, farklı kümedekiler birbirlerinden özellik değerleri bakımından farklılaşmaktadır. Çok sayıda kümeleme yöntemi bulunmaktadır. Bunlardan bir tanesi olan birleştirici hiyerarşik kümelemede, başlangıçta her bir veri kendi başına bir kümeyi oluşturmakta ve aşamalar boyunca kümelerin uzaklık değerlerine göre en yakın komşu kümeler birleşerek, sonunda tüm verilerin bulunduğu tek bir küme oluşmaktadır. Küme ve veriler arası uzaklık değerleri ise Öklid, Manhattan ve Minkowski gibi mesafe ölçütlerine göre hesaplanabilmektedir (Özekes, 2003: 75).

Kümeleme probleminde küme sayısının bilindiğini veya küme farklılık uzaklık değerlerinin belirli olduğunu varsayan çok sayıda klasik kümeleme algoritmaları bulunmaktadır. Halkidi ve diğerleri (2001) çalışmalarında klasik kümeleme algoritmalarını dört grupta incelemiştir. Bunlardan birincisi bölümleme kümeleme teknikleri olup bu sınıftaki yöntemler ön bilgi olarak küme sayısını kullanmaktadır. Bu sınıfta en bilenen yöntem olan k-ortalamalarda belirli K adet sayıda kümeye veriler ayrıştırılmaktadır. Amaç, verilere en yakın K adet küme merkezi belirlemektir. İkinci kümeleme tekniği sınıfı ise hiyerarşik kümeleme teknikleridir. Ön bilgi olarak verileri belirli bir uzaklık limit değerine göre, bu

[1] Bu çalışma, 17 -20 Ekim 2018 tarihleri arasında Antalya'da düzenlenen 19. Ekonometri, Yöneylem Araştırması ve İstatistik Sempozyumu'nda sunulan sözlü bildirinin gözden geçirilmiş, düzenlenmiş ve genişletilmiş halidir.

**Makale Geçmişi / Article History**
Başvuru Tarihi / Date of Application : 18 Şubat / February 2019
Kabul Tarihi / Acceptance Date : 14 Kasım / November 2019

değerden büyük uzaklıklara sahip olanları farklı kümelerde toplayan yaklaşımın en bilinen uygulama tekniği Aglomeratif kümeleme algoritmasıdır. Üçüncü grup kümeleme yaklaşımı ise yoğunluk temelli yaklaşımlardır. Bu sınıftaki algoritmalar ön bilgi olarak kümede bulunacak küme sayısı ve küme çap değerlerini kullanarak yoğunluk fonksiyonlarına göre kümeleme yapmaktadır. Son grupta ise ızgara tabanlı kümeleme teknikleri bulunmaktadır. Bu gruptaki yaklaşımlar veri setinin bulunduğu uzayı sonlu sayıda hücreye bölerek hücreler üzerinden işlemler yapmaktadır. Bu tip yaklaşımlar ise özellik boyutunda yer alacak ızgara hücre sayısı gibi ön bilgilere ihtiyaç duymaktadır.

Klasik kümeleme yaklaşımlarının iki tip önemli sorunu bulunmaktadır. Bunlardan ilki ön bilgiye ihtiyaç duymalarıdır. Fakat çoğu gerçek kümeleme probleminin başlangıcında herhangi bir kümeleme bilgisi mevcut değildir. Küme sayısının, fark değerlerinin belirlenemediği ve kümelemeye dair hiçbir ön bilgi bulunmadığı durumlarda doğru küme sayısına göre verileri kümelere ayırabilecek algoritmalara ihtiyaç vardır.

Klasik kümeleme yaklaşımlarının ikinci önemli problemi ise yerel minimumlara takılmalarıdır. Verilerin dağılış biçimine bağlı olarak kümeleme probleminin çok sayıda yerel optimumu bulunabilmektedir. Çok boyutlu ve büyük veriler ile yapılan çalışmalarda daha sık gözlenen bu durum istenmeyen sonuçlar üretmektedir (Rana ve diğerleri, 2011: 212).

Klasik kümeleme yaklaşımlarının sorunlarını aşabilmek için yapılan ve klasik kümeleme yaklaşımlarına dayanan çalışmaların bazılarına bakıldığında, Pelleg ve Moore (2000) küme sayısı ön bilgisi olmadan kümeleme yapabilen X-ortalamalar yöntemini geliştirmişlerdir. Buna göre çok sayıda küme adetine göre k-ortalamalar çalıştırılmakta ve bir uygunluk fonksiyonuna göre en iyi sonuç ve küme sayısı belirlenmektedir. Hamerly ve Elkan (2004) ise çalışmalarında ön bilgiye dayanmayan G-ortalamalar yöntemini ortaya atmışlardır. Başlangıçta bulunan kümelerin her birinin küme içi elemanları normal dağılıma uyup uymadıklarına göre k-ortalamalar kullanılarak yeni kümelere ayrılmasına dayanan iteratif bir yöntem ile görece orta derece büyüklükte veri setlerinin kümelemesini gerçekleştirmişlerdir. Fakat büyük boyutlu veri setleri için yeterli performansı gösteremediğini ifade etmişlerdir.

Klasik kümeleme yaklaşımlarına dayanan yaklaşımlar ön bilgiye ihtiyaç duyma problemini aşabilse de yerel optimumlara takılma problemine çözüm getirememektedirler. Problemin np-zor olması nedeniyle sezgisel algoritmalar yerel optimumları aşabilen yapıları ile kümeleme probleminde daha çok tercih edilmektedir (Pakrashi ve Chaudhuri, 2016: 705).

Bir sezgisel algoritma olan Parçacık Sürü Optimizasyonu (PSO) kümeleme için sıklıkla kullanılan bir yaklaşım olup ayrıca kullanılan uygunluk fonksiyonu sayesinde küme sayısını doğru bir şekilde belirleyebilen bir yöntemdir. PSO ile kümelemeye dair literatürde oldukça fazla sayıda çalışma mevcuttur. Kümeleme analizinde ilk kez PSO, Van der Merwe ve Engelhrecht (2003) tarafından kullanılmıştır. Çalışmada küme sayısı ön bilgisini kullanan iki farklı yaklaşım sunulmuştur. Bunlardan bir tanesi klasik PSO iken, diğeri K-ortalamalar kümeleme yolu ile bulunan kümeleme yapısını başlangıç popülasyonu olarak kullanan PSO algoritmasıdır. Kendilerinin ürettikleri iki yapay veri seti ve popüler veri setleri olan zambak, şarap, göğüs kanseri ve otomotiv veri setleri üzerinde yaptıkları karşılaştırmalar ile önerdikleri her iki yöntemin de K-ortalamalar kümeleme yöntemine göre daha iyi sonuç ürettiğini ifade etmişlerdir. Gelecek çalışmalarda PSO ile küme sayısı bilinmeden kümeleme yapılabileceğini öngördükleri çalışmalarında ayrıca kümelme uygunluk fonksiyonlarının geliştirilmesi gerektiğini belirtmişlerdir. Chen ve Ye (2004) ise çalışmalarında küme sayısı ön bilgisi ile kümeleme problemine özgü yeni bir PSO yaklaşımı geliştirmişlerdir. Çalışmada amaç fonksiyonu olarak verilerin bulundukları kümelerin merkezlerine olan uzaklıklarının toplamını kullanarak kendilerinin ürettiği dört yapay veri seti için PSO ile küme merkezlerini belirlemişlerdir. K-ortalamalar ve bulanık c-ortalamalar kümeleme yöntemlerine göre yerel optimumlara takılmadan amaç fonksiyonuna göre daha iyi sonuçlar elde ettiklerini ifade etmişlerdir.

Literatürde küme sayısı veya bir başka ön bilgi kullanmadan PSO ile kümeleme problemine çözüm getiren çalışmalara bakıldığında; Omran ve diğerleri (2006) çalışmalarında önerdikleri hibrit bir yöntem olan dinamik PSO ile PSO yöntemi ve K-ortalamalar algoritmasını çözümde birlikte kullanmışlardır. İlk önce ikili PSO ile küme sayısını belirlemişler ve sonrasında K-ortalamalar ile verileri kümelemişlerdir. Küme sayısını belirlemede Dunn İndeksi ve benzeri kümeleme geçerlilik indekslerini uygunluk fonksiyonu olarak tanımlayıp, test veri setlerinde sonuçları karşılaştırmışlardır. Karşılaştırmalar geçerlilik indekslerinin uygunluk fonksiyonu olarak kümelemede performansını görmek için yapılmış ve Turi (2001) tarafından önerilen kümeleme geçerlilik indeksinin daha iyi sonuç verdiği ifade edilmiştir. Öte yandan önerdikleri yaklaşımı Rassal Arama ve Genetik Algoritma yöntemleri ile kıyaslamışlar ve Rassal Arama'ya göre diğer iki yöntemin daha başarılı olduğunu ifade etmişlerdir. Das ve diğerleri (2008) çalışmalarında bir çekirdek fonksiyonunu uygunluk fonksiyonu olarak ele alıp çoklu elitist PSO ile kümeleme gerçekleştirmişlerdir. Parçacık gösteriminde kümelemeye özgü bir yeni öneri getirdikleri çalışmada, şarap ve göğüs kanseri gibi veri setlerini kullanarak önerdikleri çoklu elitist PSO'yu klasik PSO ve Genetik Algoritma ile kıyaslamışlardır. Önerilen yaklaşımın doğru küme sayısı ve verileri doğru kümelere ayırmayı ifade eden kümeleme doğruluk oranında daha iyi sonuçlar elde ettiğini ifade etmişlerdir. Cura (2012) çalışmasında, önerdiği PSO ile hem küme sayısını hem de küme merkezlerini uygunluk fonksiyonuna göre belirlemiştir. İki yapay veri seti ve zambak, tiroid bezi ve şarap veri setlerinin kümelenmesi yapılan çalışmada PSO, Karınca Kolonisi Optimizasyon Algoritması ve Yapay Arı Kolonisi Algoritmaları ile kıyaslanmış ve önerilen yöntemle daha iyi sonuçlar elde edildiği ifade edilmiştir. Ortakçı ve Göloğlu (2012) önerdikleri PSO yaklaşımı ile uygunluk fonksiyonu olarak kümeleme doğruluk indeksi adını verdikleri fonksiyonu kullanarak zambak veri seti ile bir adet kendilerinin ürettiği yapay veri seti için doğru küme sayısı ve küme merkezlerini tespit etmeye çalışmışlardır. Küme sayısı tespiti için belirli eşik değere göre kümelerin aktif veya pasif yapıda kalmasını önermişlerdir. Zambak veri setine göre yapay veriyi daha iyi kümeledikleri çalışmada buna neden olarak zambak veri setinin iç içe geçmiş yapısını öne sürmüşlerdir. Armano ve Framani (2016) çalışmalarında kümeleme problemine özgü çok amaçlı PSO yapısı önermişler ve tanımladıkları iki farklı fonksiyonu birlikte optimize etmişlerdir. İlk fonksiyon küme ve alt kümeler arası

bağlılık değerini verirken diğeri ise küme içi elemanların benzerlik değerlerinin hesabına dayanan yüksek benzeşme olarak ifade edilmiştir. Çalışmada bilinen klasik yöntemleri belirli kural yapılarıyla adapte ederek önerdikleri yöntem ile kıyaslamışlardır. Çok sayıda veri setini kullandıkları çalışmada PSO ile klasik yöntemlere göre daha iyi sonuç bulduklarını vurgulamışlardır. Ali (2016) çalışmasında kümeleme analizine uygun olarak geliştirdiği adaptif PSO ile parçacıkları iki boyutta tanımlamıştır. İlk boyutta optimum küme sayısı bilgisi bulunurken diğerinde küme merkez bilgileri bulunmaktadır. Parçacık konumları önceden tanımlanmış kurallara göre adaptif bir şekilde belirlenmektedir. Önerilen yöntem, yapay ve gerçek veri setleri kullanılarak literatürde yer alan yöntemlerle kıyaslanmış ve kümeleme doğruluğu açısından daha iyi sonuçlar elde edilmiştir. Esmin ve diğerleri (2015) çalışmalarında PSO ile kümeleme analizi konusunda yapılan çalışmaları incelemişlerdir. Çalışmaların genelde konu olarak; kümeleme analizi, veri akışı kümeleme ve internet madenciliği, belge kümeleme, özellik seçimi, görüntü işleme, endüstriyel uygulamalar, anomali saptama ve metin kümeleme gibi alanlarda yapıldığını ifade etmişlerdir. Çok sayıda hibrit yöntem ve farklı yaklaşımlarla PSO ile kümelemede yüksek performans sağlandığını ve gelecekte yapılacak iyileştirme çalışmalarıyla özellikle doğrusal şekilde ayrılamayan veri setlerinde daha da iyi sonuçlar elde edileceğini ifade etmişlerdir.

Genel olarak PSO veya diğer sezgisel algortimalar ile yapılan kümeleme çalışmalarında algoritmanın arama yöntemine odaklandığı gözlenilmektedir. Fakat ön bilgiye ihtiyaç duymadan verileri kümelere doğru şekilde ayrıştıran uygunluk fonksiyonlarının geliştirlmesi de en az arama yöntemi kadar önemli bir problemdir.

Bu çalışmada, küme sayısı ön bilgisi olmadan özellikle iç içe geçmiş verilere sahip veri setlerini doğru bir şekilde kümeleyebilen, yeni bir uygunluk fonksiyonu önerisi getirilerek literatüre bu yönden katkı yapılmıştır. Önerilen uygunluk fonksiyonu ile klasik PSO yöntemi kullanılarak küme sayısı önceden belirli olmadığı halde doğru küme sayısı ve elemanların doğru kümelere yerleşimi amaçlanmıştır. Bu nedenle doğru şekilde kümelere ayrımı zor olan üç adet yapay veri seti ve literatürde sıklıkla kullanılan zambak, şarap ve göğüs kanseri veri setleri ile önerilen uygunluk fonksiyonu literatürde en çok kullanılan uygunluk fonksiyonları ile küme sayısı bulma ve kümeleme doğruluğu açısından karşılaştırılmıştır.

## 2. Parçacık Sürü Optimizasyonu

Grup halinde yaşayan ve birbirleri ile etkileşimleri yüksek hayvanların davranış biçimlerinden yola çıkan Eberhart ve Kennedy (1995) PSO algoritmasını geliştirmişlerdir. PSO ortaya atıldığından itibaren yöntem üzerinde iyileştirmeler sıklıkla gerçekleştirilmiştir. Bu iyileştirmelerden, kabul görmüş biri ise Shi ve Eberhart (1999) tarafından sunulan, $W_{IN}$ eylemsizlik ağırlığının modele eklenmesidir. Bu sayede algoritma arama yönünü belirleme mümkün kılınmıştır. Aladağ ve diğerleri (2012) ise PSO parametrelerinin tamamının optimizasyon sürecinde değişimine bağlı yeni bir PSO algoritması önermişlerdir. Bu yaklaşımda genel olarak başlangıçta parçacıkların hızı yüksek iken sonlarda yavaşlayabilmektedir. Bu durumda ise arama genişten dar alanlara doğru kaymaktadır. Bu çalışmada da kullanılan PSO'yu Aladağ ve diğerleri (2012) aşağıda olduğu gibi ifade etmektedir;

Adım 1: Her bir j. (j=1,2,…,pn) parçacığı rassal olarak Xj vektörüne ve n adet pozisyona yerleştirip sakla.

$$X_j = \left(x_{j,1}, x_{j,2}, ..., x_{j,n}\right), \quad (j = 1,2,...,pn), \quad (i = 1,2,...,n)$$

Adım 2: Hız vektörünü rassal olarak oluştur ve Vj'de sakla.

$$V_j = \left(v_{j,1}, v_{j,2}, ..., v_{j,n}\right), \quad j = 1,2,...,pn$$

Adım 3: Her bir parçacığın en iyi performansını belirle ve Pbest'de sakla. Tüm parçacıklar içerisinde o ana kadar elde edilen en iyi performansı belirle ve Gbest'de sakla.

Adım 4: Parametrelerin güncellenmesi. Burada bilişsel parametre $c_1 = (c_{1i}, c_{1f})$, sosyal parametre $c_2 = (c_{2i}, c_{2f})$ ve global ile yerel arama gücünü etkileyen $W_{IN} = (W_{IN1}, W_{IN2})$ aralığında değer almaktadır. Maksimum iterasyon sayısı $\max t$ ile gösterilirken $t$ iterasyon numarasıdır.

$$c_1 = (c_{1f} - c_{1i}) \frac{t}{\max t} + c_{1i}$$

$$c_2 = (c_{2f} - c_{2i}) \frac{t}{\max t} + c_{2i}$$

$$W_{IN} = (W_{IN2} - W_{IN1}) \frac{\max t - t}{\max t} + W_{IN1}$$

Adım 5: Hızın ve konumun güncellenmesi ise aşağıdaki gibidir.

$$v_{i,n}^{t+1} = W_{IN} \times v_{i,n}^{t} + c_1 \times rand_1 \times (P_{i,n} - x_{i,n}) + c_2 \times rand_2 \times (P_{g,n} - x_{i,n})$$

$$x_{i,n}^{t+1} = x_{i,n}^{t} + v_{i,n}^{t+1}$$

PSO algoritması önceden belirlenmiş olan iterasyon sayısına ulaşana kadar çalışmakta ve en iyi çözüm değerine sahip çözüm en iyi olarak kabul edilmektedir. Çalışmada, Aladağ ve diğerleri (2012) tarafından önerilen PSO kullanıldığından dolayı sezgisel parametre değerleri c1 ve c2 (1, 2), ω ise (0.4, 0.9) aralığında Aladağ ve diğerleri (2012) ile aynı değerler seçilmiştir.

## 3. Parçaçık Sürü Optimizasyonu ile Kümeleme

PSO'da bulunan her bir parçacık kümeleme problemi açısından birer olası çözümü sunmaktadır. Buna göre parçacıklar küme merkez değerlerini ve küme sayı bilgilerini içermektedir. Küme merkez değerlerine olan uzaklıklar gözetilerek veri setindeki her bir eleman kendisine en yakın kümeye dahil edilmektedir. Sonrasında kümelerdeki elemanların değerlerine göre küme merkez değerleri güncellenmektedir. Burada amaçlanan, parçacıkların belirli bir uygunluk fonksiyonu dikkate alınarak uygunluk fonksiyonunu en iyileyecek küme merkezlerinin tespitinin yapılmasıdır. PSO ile kümeleme adımları aşağıdaki gibi ifade edilebilir.

Sürünün başlangıç popülasyonu X ile ifade edilirken kümeleme işlemleri sırasında da her bir parçacık, oluşturulan kümelerin merkezini ifade eden birer vektördür;

$X=\{X_1, \dots, X_l, \dots, X_{pop}\}$ . Pop = sürüdeki parçacık sayısı

Her bir parçacık, $X_i=\{\overrightarrow{O_1}, \dots, \overrightarrow{O_2}, \dots, \overrightarrow{O_J}, \dots \overrightarrow{O_k}\}$. k = oluşturulan küme sayısı

$O_J$= J. kümenin ağırlık merkezi

Veri setinde kümelenecek olan her elemanın özelliği veya boyutu ile küme merkezinin boyut sayısı aynıdır. Küme sayısının önceden bilinmediği durumlarda doğru kümelemeyi bulma ve veri setinin kaç kümeden oluşacağı önemli bir optimizasyon konusunu oluşturmaktadır (Ortakçı ve Göloğlu, 2012: 4). Kümeleme için iki nesne arasındaki benzer uzaklıklar Öklid uzaklığı ile tanımlanabilmektedir.

$$D(o_i, o_l) = \left(\sum_{j=1}^{m}(o_{ij} - o_{lj})^r\right)^{1/r} \Rightarrow D(o_i, o_l) = \sqrt{\sum_{j=1}^{m}(o_{ij} - o_{lj})^2}$$

$D(o_i, o_l)$ : i. ve j. Nesneler arasındaki benzemezlik ölçüsü

$o_{ij}$: i. nesnenin j. özelliğinin değeri (i = 1,. . . ,n ve j = 1, . . . ,m)

Küme sayısı bilindiği durumda m boyutlu uzaydaki n adet nesnelere ait veri seti K kümeye bölünür. Kümeleme probleminin matematiksel modeli bu durumda aşağıdaki gibidir (Shelokar ve diğerleri, 2004: 190) :

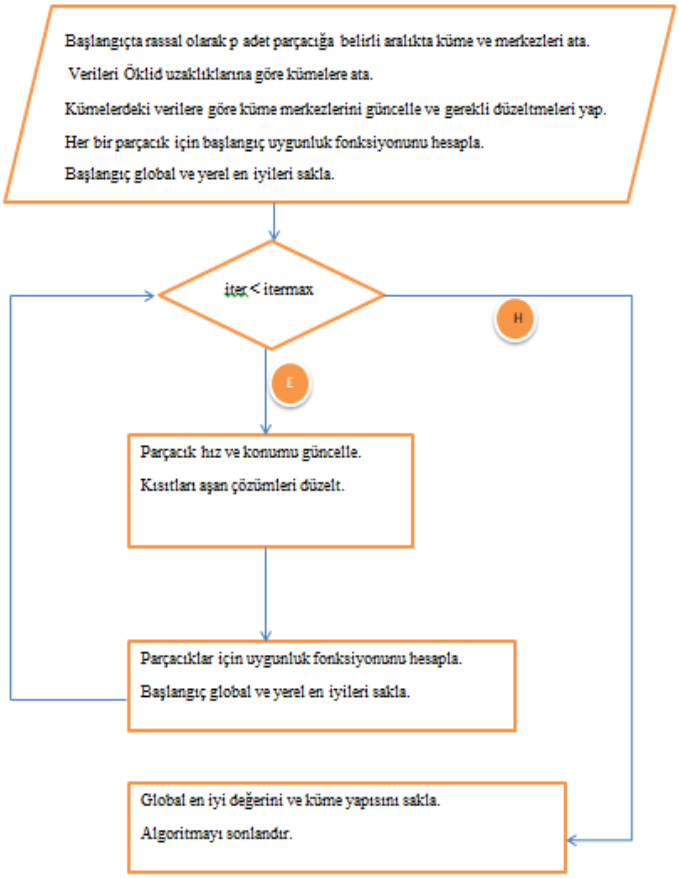
$$\text{Min} \sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik}\, D(o_i, z_k)$$

$$\sum_{k=1}^{K} w_{ik} = 1 \qquad i = 1, \dots, n$$

$$\sum_{i=1}^{n} w_{ik} \geq 1 \qquad k = 1, \dots, K$$

$z_k$, k kümesinin merkezini, $w_{ik}$, eğer $w_{ik} = 1$ ise k. kümeye ait i. nesneyi gösterir.

Küme sayısı bilinmediği durumdaki PSO ile kümeleme akış şeması ise Şekil 1'deki gibidir.




Şekil 1. PSO ile Kümeleme Akış Şeması

Kümelemede Dunn İndeksi gibi geçerlilik indeksleri küme sayısı bilinmediğinde çoğu zaman kullanılan önemli fonksiyonlardır (Pakhira ve diğerleri, 2004: 500). Çalışmada incelenen ve sıklıkla kullanılan Dunn indeksi (DI) Dunn (1973) tarafından aşağıdaki gibi tanımlanmıştır.

$$DI = \min_{k=1,\dots,K}\left\{\min_{kk=k+1,\dots,K}\left(\frac{dist(C_k, C_{kk})}{\max_{a=1,\dots,K} d(C_a)}\right)\right\}$$

$dist(C_k, C_{kk})$, $C_k$ ve $C_{kk}$ kümeleri arasındaki uzaklık fonksiyonudur.

$$dist(C_k, C_{kk}) = \min_{u\in C_k, w\in C_{kk}} d(u, w)$$

$d(u, w)$, u ve w elemanları arasındaki Öklid uzaklığını ifade etmektedir.

$$d(C) = \max_{u,v\in C} diam(u, v)$$

$d(C)$, herhangi kümedeki birbirine en uzak iki elemanın (diam(u,v)) Öklid uzaklığını verir ve küme çapı olarak adlandırılmaktadır. Dunn indeksi kısaca kümeleri ayrıştırmak için en kısa uzaklığa sahip küme merkezlerini maksimize etmeye ve aynı zamanda verileri yoğunlaştırmak için en büyük küme çapını minimize etmeye çalışmaktadır.

Çalışmada uygunluk fonksiyonu olarak önerilen ve kullanılan Değiştirilmiş Dunn indeksi (DDI) ise Dunn indeksinin küme çapını minimize etme hedefi, küme eleman sayısı oranlı küme çapını minimize etme hedefi ile yer değiştirmiş ve aşağıdaki şekilde ifade edilmiştir;

$$DDI = \min_{k=1,\dots,K}\left\{\min_{kk=k+1,\dots,K}\left(\frac{dist(C_k, C_{kk})}{\max_{a=1,\dots,k}[(c(C_a)/n)\, d(C_a)]}\right)\right\}$$

Kullanılan algoritmada küme merkezleri arasındaki uzaklık küme uzaklıkları olarak ele alınmıştır. Toplam gözlem sayısı n ile gösterilirken, herhangi bir kümedeki eleman sayısı ise $c(C)$ ile ifade edilmektedir. DI ve benzeri geçerlilik indekslerinin veri setinde yer alan gürültüden etkilenmesi ve örtüşen veri setlerinde iyi sonuç vermedikleri bilindiğinden (Halkidi ve diğerleri, 2001:131; Zhao ve diğerleri, 2009: 319), kümelemede küme çapını eleman sayısı ile ağırlıklandıran yeni bir uygunluk fonksiyonu DDI çalışmada önerilmiştir.

Çalışmada ayrıca Davies-Bouldin (DBI) ve Silhouette (SI) indeksleri de önerilen uygunluk fonksiyonu ile karşılaştırılarak DDI indeksinin uygunluk fonksiyonu olarak etkinliği gözler önüne serilmiştir.

DBI aşağıdaki gibi tanımlanabilmektedir (Selvi ve Çağlar, 2017: 420);

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} max\left\{\left(\frac{Sn(Qi) + Sn(Qj)}{Sn(Qi,Qj)}\right)\right\}$$

DBI'de Sn(Qi) ve Sn(Qj) sırasıyla i. ve j. kümelerin verilerinin küme merkezine olan uzaklıklarının ortalaması olarak hesaplanırken, Sn(Qi,Qj) i. ve j. küme merkezlerinin aralarındaki mesafeyi belirtmektedir. DBI değeri ne kadar küçük çıkarsa o kadar iyi yerleşim olduğu düşünülmektedir.

SI ise aşağıdaki gibi hesaplanmaktadır (Akgül ve Başkır, 2013: 56);

$$\text{SI} = \frac{b(Xi) - a(Xi)}{\max(a(Xi), b(Xi)}$$

SI'da b(Xi) i. elemanın diğer kümelerdeki tüm elemanlara ortalama uzaklığının minimum iken a(Xi) i. elemanın kendi kümesindeki tüm elemanlara olan uzaklıklarının ortalamasıdır. Tüm birimlerin SI değerlerinin ortalaması ise veri setine dair kümeleme SI değeri olarak adlandırılmakta ve büyük olması istenmektedir.

## 4. Uygulama

Çalışmada, önerilen uygunluk fonksiyonu DDI'ın PSO ile kümeleme performansını klasik uygunluk fonksiyonlarına göre değerlendirmek için toplam beş adet veri setinden faydalanılmıştır. İlk ikisi Kao ve diğerleri (2008) tarafından önerilen ve Cura (2012) tarafından da kullanılan yapay veri setleri olmuştur. Diğerleri ise Blake ve Merz (1998) tarafından oluşturulan veritabanından alınmış popüler kümeleme veri setleri olan zambak, şarap ve göğüs kanseri verileridir. Çalışmada kullanılan veri setleri aşağıdaki gibi tanımlanabilmektedir;

*Veri Seti 1*: Yapay veri seti 1 (Yapay1). 600 elemandan oluşan veri setinde iki özellik ve dört küme bulunmaktadır. Örnekler dört farklı bağımsız iki değişkenli normal dağılımdan $\mu = \begin{pmatrix} Qi \\ Qi \end{pmatrix}$ ve $Cov = \begin{pmatrix} 0.5 & 0.05 \\ 0.05 & 0.5 \end{pmatrix}$ olmak üzere Qi={-3, 0, 3, 6} değerlerine göre çekilmiştir.

*Veri Seti 2:* Yapay veri seti 2 (Yapay2). Veri setinde 250 eleman, üç özellik ve beş küme bulunmaktadır. Örnekler, aralık değerleri [85,100], [70, 85], [55, 70], [40,55] ve [25, 40] olan beş adet bağımsız uniform dağılıştan çekilmiştir.

*Veri Seti 3*: Zambak veri seti. Zambak verisi 150 veri noktasından oluşan dört boyutu bulunan bir veri setidir. Zambak veri seti boyutları sırasıyla, çiçeğinin santimetre cinsinden çanak yaprak uzunluğu, çanak yaprak genişliği, ayakçak yaprak uzunluğu ve ayakçak yaprak genişliği değerleridir. Üç kümeden oluşan zambak veri setinde her bir kümede eşit sayıda eleman bulunmakta ve veri seti tipi olarak iç içe geçmiş örtüşen kümelere sahiptir.

*Veri Seti 4*: Şarap veri seti. Şarap veri seti 178 veri noktasından oluşan 13 boyutu olan bir veri setidir. Şarap veri seti boyutları sırasıyla, alkol oranı, malik asit derecesi, sodyum karbonat oranı, sodyum karbonat alkanilitesi, magnezyum değeri, fenoller, flavanoid değeri, flavonoid içermeyen fenoller, proantosiyanidinler, renk yoğunluğu, renk tonu, seyreltme ve prolin değeri olarak adlandırılmaktadır. Üç kümeden oluşan şarap veri setinde birinci kümede 59, ikinci kümede 71, üçüncü kümede ise 48 gözlem bulunmaktadır.

*Veri Seti 5*: Göğüs kanseri veri seti. Göğüs kanseri veri seti 699 veri noktasından oluşan 9 boyutu olan bir veri setidir. Göğüs kanseri veri seti boyutları sırasıyla, kütle kalınlığı, hücre büyüklüğü tekdüzeliği, hücre şekli tekdüzeliği, marjinal yapışma, tek epitel hücre büyüklüğü, yalın çekirdek, hafif kromatin, normal nükleol ve mitozlar olarak adlandırılmaktadır. İki kümeden oluşan göğüs kanseri veri setinde iyi huylu birinci kümede 458, kötü huylu ikinci kümede ise 241 gözlem bulunmaktadır.

PSO kümeleme algoritması MATLAB programında kodlanmıştır. Yapılan denemeler sonucu PSO iterasyon sayısı 100 ve parçacık sayısı 20 seçilmiştir. PSO yöntemi dört ayrı uygunluk fonksiyonu ve beş veri setine göre 30 kez birbirinden bağımsız çalıştırılarak kümeleme sonuçları elde edilmiştir. Sonuçlar Tablo 1'deki gibidir.

Tablo 1. Veri Setleri ve Uygunluk Fonksiyonlarına Göre PSO ile Kümeleme Sonuçları

| PSO | ORTALAMA KÜMELEME DOĞRULUĞU | | | | Gerçek Küme Sayısı | ORTALAMA KÜME SAYISI | | | |
|---|---|---|---|---|---|---|---|---|---|
| | DDI | DI | DBI | SI | | DDI | DI | DBI | SI |
| Yapay1 | **0.978055** | 0.49833 | 0.962055 | 0.794888 | 4 | **4** | 2 | 3.9667 | 3.2 |
| Yapay2 | **0.727067** | 0.312 | 0.3996 | 0.396133 | 5 | **4.3** | 2 | 2 | 2 |
| Zambak | **0.960222** | 0.592445 | 0.66667 | 0.66667 | 3 | **3** | 2 | 2 | 2 |
| Şarap | **0.762922** | 0.523784 | 0.537078 | 0.59551 | 3 | **2.9333** | 3.4667 | 3.1 | 2 |
| Göğüs Kanseri | 0.919693 | 0.883166 | 0.93295 | **0.935716** | 2 | 2.6333 | 2.3667 | 2.4333 | **2.3667** |

Tablo 1'deki sonuçlara göre önerilen uygunluk fonksiyonu olan DDI her iki yapay veride ve zambak ile şarap veri setlerinde gerçek küme sayısına en yaklaşık ortalama küme sayısı değerine sahipken sadece göğüs kanseri veri setinde diğer uygunluk fonksiyonları daha iyi sonuç vermektedir. Cura (2012) tarafından yapılan çalışmada da kullanılmış olan yapay1, yapay2, zambak ve şarap veri setleri için bulunan ortalama küme sayısına göre önerilen DDI ile Cura'ya (2012) göre daha iyi sonuçlar elde edilmiştir. Öte yandan DDI ile her iki yapay veride ve zambak ile şarap veri setlerinde daha yüksek şekilde veriler ait oldukları kümelere yerleştirebilirken sadece göğüs kanseri veri setinde DBI ve SI küçük farklarla kümeleme doğruluğu bazında daha iyi sonuç vermektedir. Ortalama kümeleme doğruluğu açısından uygunluk fonksiyonlarının arasında istatistiki olarak anlamlı farklılıklar olup olmadığını analiz etmek için Alswaitti ve diğerleri (2018) tarafından önerilen ve Hodges ve Lehmann (1968) tarafından ortaya atılan Friedmann hizalı sıra (FHS) testi ve sonrasında post-hoc olarak Holm (1979) tarafından ortaya atılan Holm testi uygulanmıştır. En iyiden en kötüye göre sıralanan uygunluk fonksiyonlarının test sonuçları Tablo 2'deki gibidir.

Tablo 2. FHS ve Holm Testleri ile Kümeleme Doğruluğuna Göre Uygunluk Fonksiyonları Ortalama Sıralamaları

| Uygunluk Fonksiyonları | FHS Ortalama Sıralama | p-değerleri | Holm Düzeltilmiş p-değerleri |
|---|---|---|---|
| DDI | 3.6 | | |
| SI | 10.5 | 0.00028 | 0.05 |
| DBI | 10.9 | 0.000131 | 0.025 |
| DI | 17 | 1.04E-11 | 0.01667 |

FHS test istatistiği değeri 10.93 elde edilmiş olup FHS testinin anlamlılık değeri ise 0.01209 olarak bulunmuştur. Bu durumda gözlemler arasında farklılık olduğu söylenebilmektedir. Önerilen DDI en iyi sonuç veren fonksiyon olduğu için kontrol fonksiyonu olarak ele alınmış ve post-hoc olarak Holm testi uygulanmıştır. Sonuçlara göre DDI uygunluk fonksiyonunun, diğer uygunluk fonksiyonlarına göre fonksiyon p-değerlerinin kendilerine karşılık gelen Holm düzeltilmiş p-değerlerinden daha küçük olduğu için istatistiki olarak kümeleme doğruluğu açısından daha iyi sonuç verdiği görülmektedir.

Çalışmada ayrıca önerilen DDI uygunluk fonksiyonu ile kullanılan PSO, Omran ve diğerleri (2006) tarafından da yapılan şekilde Genetik Algoritma ve Rassal Arama Algoritmaları ile karşılaştırılmıştır. Maulik ve Bandyopadhyay (2000) tarafından kümeleme için önerilen Genetik Algoritma (GA), PSO ile benzer şekilde 20 populasyon büyüklüğü ve 100 iterasyon sayısına göre çalıştırılmıştır. Rassal Arama (RA) ise rassal olarak atanan küme merkezlerinin kendi kümelerine yerleşen elemanların değerleri ile güncellenmesine dayanmakta ve diğer sezgiseller ile benzer parametre değerlerine sahiptir. Her üç algoritmanın da benzer çalışma sürelerine sahip olduğu ve uygun sürelerde sonuç verdikleri gözlenmiştir. Her üç yöntem de önerilen ve diğer klasik uygunluk fonksiyonlarına göre daha iyi sonuç veren DDI uygunluk fonksiyonuna göre 30'ar kez ayrı ayrı çalıştırılmış ve Tablo 3'de sonuçları verilmiştir.

Tablo 3. Veri Setleri ve Yöntemlere Göre DDI ile Kümeleme Sonuçları

| DDI | ORTALAMA KÜMELEME DOĞRULUĞU | | | Gerçek Küme Sayısı | ORTALAMA KÜME SAYISI | | |
|---|---|---|---|---|---|---|---|
| | PSO | GA | RA | | PSO | GA | RA |
| Yapay1 | **0.978055** | 0.617445 | 0.494267 | 4 | **4** | 4.5333 | 3.7333 |
| Yapay2 | **0.727067** | 0.6224 | 0.6165 | 5 | 4.3 | **4.3333** | 4.1667 |
| Zambak | **0.960222** | 0.676 | 0.685111 | 3 | **3** | 3.8667 | 3.0333 |
| Şarap | **0.762922** | 0.59176 | 0.421161 | 3 | **2.9333** | 4.8 | 6.8333 |
| Göğüs Kanseri | **0.919693** | 0.622795 | 0.719266 | 2 | 2.6333 | 4.8 | **2** |

Tablo 3'deki sonuçlara göre PSO, yapay1, zambak ile şarap veri setlerinde gerçek küme sayısına en yaklaşık ortalama küme sayısı değerine sahip olmuştur. Ortalama kümeleme doğruluğu açısından ise tüm veri setlerinde PSO daha yüksek oranlara sahip olmuştur. Ortalama kümeleme doğruluğu açısından uygunluk fonksiyonlarının arasında istatistiki olarak anlamlı farklılıklar olup olmadığını analiz etmek için sırasıyla FHS ve Holm testleri uygulanmıştır. En iyiden en kötüye göre sıralanan yöntemlerin test sonuçları Tablo 4'deki gibidir.

Tablo 4. FHS ve Holm Testleri ile Kümeleme Doğruluğuna Göre Yöntemlerin Ortalama sıralamaları

| Yöntemler | FHS Ortalama Sıralama | p-değerleri | Holm Düzeltilmiş p-değerleri |
|---|---|---|---|
| PSO | 3 | | |
| GA | 9.6 | 6.93E-05 | 0.05 |
| RA | 11.4 | 6.18E-07 | 0.025 |

Tablo 4'de yer alan testlerden FHS test istatistiği değeri 7.12 elde edilmiş olup FHS testinin anlamlılık değeri ise 0.02842 olarak bulunmuştur. Bu durumda gözlemler arasında farklılık olduğu söylenebilmektedir. PSO en iyi sonuç veren yöntem olduğu için kontrol yöntemi olarak ele alınmış ve post-hoc olarak Holm testi uygulanmıştır. Sonuçlara göre yöntemlerin p-değerleri kendilerine karşılık gelen Holm düzeltilmiş p-değerlerinin altında çıkmıştır. Bu durumda PSO'nun kümeleme doğruluğu açısından diğer yöntemlere göre daha yüksek başarı sağladığı istatistiki olarak görülmektedir.

## 5. Sonuç

Kümeleme analizi, sağlık, mühendislik ve üretim gibi birbirinden bağımsız birçok sektörde kullanılmaktadır. Günümüzde oldukça popüler olan yapay zeka alanında kendine önemli bir yer edinmiş bir analiz olup aynı zamanda veri madenciliğinin de temellerini oluşturmaktadır. Kümeleme sonuçları yöntemden yönteme değişebilmektedir. Gerçek hayatta karşılaşılan problemlerde olduğu gibi kümelemeye dair herhangi bir ön bilgi yoksa etkin sonuçların alınması daha da zorlaşmaktadır.

Çalışmada önerilen uygunluk fonksiyonu DDI, PSO yönteminde kullanılarak, ön bilgi olmadan literatürde yer alan beş veri seti için kümeleme gerçekleştirilmiştir. Verilerin ait oldukları kümelere yerleşme oranını ifade eden kümeleme doğruluğu değeri açısından kümeleme analiznde popüler uygunluk fonksiyonları olan DI, DBI, ve SI'ya göre istatistiki olarak daha iyi sonuç elde edilmiştir. Ayrıca önerilen DDI ile bulunan ortalama küme sayısı gerçek küme sayılarına diğer uygunluk fonksiyonlarına göre daha yakın çıkmaktadır.

Öte yandan kullanılan yöntem olan PSO'nun etkinliğini görebilmek adına problemde GA ve RA karşılaştırmalar yapılmıştır. Elde edilen sonuçlara göre PSO'nun etkinliği ortaya çıkarken Omran ve diğerleri (2006) tarafından bulunan sonuçlara paralel olarak RA en kötü performansı sergilemiştir. Gelecekteki kümeleme analizi çalışmaları için veri seti yapısına adapte olabilen uygunluk fonksiyonlarının hibrit optimizasyon teknikleri ile uygulanmasının kümeleme doğruluğu ve doğru küme sayısını bulmada yararlı olması öngörülmektedir.

## KAYNAKÇA


Akgül, F. G., ve Başkır, M. B. (2013). Bankaların 2008-2012 Yılları Arasında Aktif Büyüklüklerini Etkileyen Kriterler Bakımından Hiyerarşik Kümeleme ve PAM Algoritması ile Sınıflandırılması. Bankacılık ve Sigortacılık Araştırmaları Dergisi, 1(5), 48-63.

Aladağ, C. H., Yolcu, U., Egrioğlu, E., ve Dalar, A. Z. (2012). A new time invariant fuzzy time series forecasting method based on particle swarm optimization. *Applied Soft Computing*, 12(10), 3291-3299.

Ali, Y. M. B. (2016). Unsupervised clustering based an adaptive particle swarm optimization algorithm. Neural Processing Letters, 44(1), 221-244.

Alswaitti, M., Albughdadi, M. ve Mat Isa, N. A. (2018). Density-Based Particle Swarm Optimization Algorithm For Data Clustering. *Expert Systems With Applications*, 91: 170-186.

Armano, G. ve Framani, M. R. (2016), Multiobjective Clustering Analysis Using Particle Swarm Optimization. *Expert Systems With Applications*, 55, 184–193.

Blake C. L. ve Merz C. J. (1998). UCI repository of machine learning databases.<http://www.ics.uci.edu/-mlearn/MLRepository.html>.

Chen, C.-Y., ve Ye, F. (2004). Particle swarm optimization algorithm and its application to clustering analysis. In Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control, Taipei, Taiwan (pp. 789–794).

Cura, T. (2012). A particle swarm optimization approach to clustering. *Expert Systems with Applications*, 39(1), 1582-1588.

Das, S., Abraham, A., ve Konar, A. (2008). Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm. *Pattern recognition letters*, 29(5), 688-699.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3). 32-57.

Eberhart, R., ve Kennedy, J. (1995). A new optimizer using particle swarm theory. In MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science (pp. 39-43). Ieee.

Esmin, A. A., Coelho, R. A., ve Matwin, S. (2015). A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. *Artificial Intelligence Review*, 44(1), 23-45.

Fidan, H. (2009). Pazarlama Bilgi Sistemi (Pbs) Ve Coğrafi Bilgi Sistemi (Cbs) Nin Pazarlamada Kullanimi. *Journal of Yaşar University*, *4*(14), 2151-2171.

Halkidi, M., Batistakis, Y., ve Vazirgiannis, M. (2001). On clustering validation techniques. Journal of intelligent information systems, 17(2-3), 107-145.

Hamerly, G., ve Elkan, C. (2004). Learning the k in k-means. In Advances in neural information processing systems (pp. 281-288).

Hodges, J. L., ve Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. The Annals of Mathematical Statistics, 33(2), 482-497.

Kao, Y. T., Zahara, E., ve Kao, I. W. (2008). A hybridized approach to data clustering. Expert Systems with Applications, 34(3), 1754-1762.

Maulik, U., ve Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. Pattern recognition, 33(9), 1455-1465.

Omran, M. G., Salman, A., ve Engelbrecht, A. P. (2006). Dynamic clustering using particle swarm optimization with application in image segmentation. *Pattern Analysis and Applications*, 8(4), 332-344.

Ortakçı, Y. ve Göloğlu, C. (2012). Parçacık Sürü Optimizasyonu İle Küme Sayısının Belirlenmesi. *Akademik Bilişim Akademik Bilişim'12 - XIV. Akademik Bilişim Konferansı Bildirileri 1 - 3 Şubat 2012 Uşak Üniversitesi*, 335–341.

Özekes, S. (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 2(3), 65-82.

Pakhira, M. K., Bandyopadhyay, S., ve Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern recognition*, 37(3), 487-501.

Pakrashi, A., ve Chaudhuri, B. B. (2016). A Kalman filtering induced heuristic optimization based partitional data clustering. Information Sciences, 369, 704-717.

Pelleg, D., ve Moore, A. W. (2000, June). X-means: Extending k-means with efficient estimation of the number of clusters. In Icml (Vol. 1, pp. 727-734).

Rana, S., Jasola, S., ve Kumar, R. (2011). A review on particle swarm optimization algorithms and their applications to data clustering. Artificial Intelligence Review, 35(3), 211-222.

Selvi, H. Z., ve Çağlar, B. (2017). Çok Değişkenli Haritalama İçin Kümeleme Yöntemlerinin Kullanilmasi. Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi, 6(2), 415-429.

Shelokar, P. S., Jayaraman, V. K., ve Kulkarni, B. D. (2004). An ant colony approach for clustering. *Analytica Chimica Acta*, 509, 187–195.

Shi, Y., ve Eberhart, R. C. (1999). Empirical study of particle swarm optimization. In Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on (Vol. 3, pp. 1945-1950). IEEE.

Turi, R. H. (2001). Clustering-based colour image segmentation (p. 446). PhD thesis: Monash University.

Van der Merwe, D. W., ve Engelbrecht, A. P. (2003, December). Data clustering using particle swarm optimization. In Evolutionary Computation, 2003. CEC'03. The 2003 Congress on (Vol. 1, pp. 215-220). IEEE.

Zhao, Q., Xu, M., ve Fränti, P. (2009). Sum-of-squares based cluster validity index and significance analysis. In International Conference on Adaptive and Natural Computing Algorithms (pp. 313-322). Springer, Berlin, Heidelberg.