Optimum Ekonomi ve Yönetim Bilimleri Dergisi, 2020, 7(1), 279-296
Optimum Journal of Economics and Management Sciences, 2020, 7(1), 279-296

Bulanık Parçacık Sürü Optimizasyon Yaklaşımı Temelli Kümeleme*

Mehmet AKSARAYLI**, Osman PALA***

ÖZ

Aynı özelliklere sahip gözlem noktalarını, özelliklerinin aldığı değerler açısından gruplara ayırma işlemine kümeleme analizi adı verilmektedir. Makine öğrenmesi tekniklerinden olan kümelemede amaç, gözlem noktalarını gruplayarak farklı gruptaki gözlem noktaları için farklı stratejiler uygulanmasının sağlanmasıdır. Birçok bilim dalında kullanılan kümelemede, küme sayısı gibi ön bilgilerin bulunmadığı durumlarda işlem zorlaşmaktadır. Küme sayısı önceden belirli olmadığı durumlarda, uygunluk fonksiyonuna göre küme sayısını belirleyebilen sezgisel algoritmalar kullanılabilmektedir. Çalışmada, önerilen Bulanık Parçacık Sürü Optimizasyonu sezgisel algoritmasının ve uygunluk fonksiyonunun kümelemedeki başarısını değerlendirebilmek adına kümeleme analizinde sıklıkla kullanılan veri setlerinden faydalanılmıştır. Analiz sonuçlarına göre önerilen algoritmanın doğru küme sayısını bulmada ve gözlemleri doğru gruplamada klasik yaklaşıma göre daha yüksek başarım gösterdiği gözlenmiştir.

Anahtar Kelimeler: Parçacık Sürü Optimizasyonu, Kümeleme Analizi, Bulanık Adaptif.

JEL Sınıflandırması: C610

Clustering Based on Fuzzy Adaptive Particle Swarm Optimization Approach

ABSTRACT

The process of grouping the observation points with the same characteristics in terms of the values of the characteristics is called clustering analysis. The aim of clustering, which is one of the techniques of machine learning, is to provide the implementation of different strategies for observation points in the different group by grouping the observation points. In clustering used in many disciplines, the process becomes difficult when there is no preliminary information such as the number of clusters. In cases where the number of sets is not predetermined, heuristic algorithms can be used, which can determine the number of sets according to the fitness function. In this study, in order to evaluate the success of the proposed Fuzzy Particle Swarm Optimization heuristic algorithm and fitness function in clustering, data sets frequently used in clustering analysis were utilized. According to the results of the analysis, it was observed that the proposed algorithm indicated a quite successful performance in finding the correct number of clusters and in the correct grouping of observations.

Keywords: Particle Swarm Optimization, Clustering Analysis, Fuzzy Adaptive.

JEL Classification: C610

Geliş Tarihi / Received: 11.06.2019 Kabul Tarihi / Accepted: 28.11.2019

* Bu çalışma, 05-07 Ekim 2018 tarihlerinde İzmir'de gerçekleştirilen 'Birinci Uluslararası Sosyal Bilimlerde Kritik Tartışmalar Kongresi'nde sunulan sözlü bildirinin gözden geçirilmiş, düzenlenmiş ve genişletilmiş halidir.

** Prof. Dr., Dokuz Eylül Üniversitesi, İİBF, Ekonometri Bölümü, mehmet.aksarayli@deu.edu.tr, ORCID: 0000-0003-1590-4582.

*** Arş. Gör., Karamanoğlu Mehmetbey Üniversitesi, İİBF, Ekonometri Bölümü, osmanpala@kmu.edu.tr, ORCID: 0000-0002-2634-2653.

1. GİRİŞ

Kümeleme analizi birçok alanda önemli problemlerin çözümünde kullanılan çok önemli bir istatistiksel yaklaşımdır. Punj ve Stewart (1983) kümeleme analizinin pazarlama alanında önemli bir yaklaşım olduğunu ve pazar segmentasyonu probleminin kümeleme analizinin en popüler uygulama alanı olduğunu belirtirken aynı zamanda yeni ürün tasarımı ve müşteri davranışlarını anlamada da kümelemeden faydalanıldığını ifade etmişlerdir. Belbin ve McDonald (1993) ekoloji alanında verilerin analizinde kümeleme analizinin oldukça faydalı ve açıklayıcı olduğunu belirtmişler ve bu sayede farklı kümedekilere farklı stratejiler geliştirilerek sorunların çözümünde önemli yol alındığını ifade etmişlerdir. Ketchen ve Shook (1996) stratejik yönetim alanında firmaların strateji, çevre, liderlik ve performans gibi değişkenler açısından kümeleme analizi ile değerlendirilmesinin ilgili değişkenler arasındaki ilişkiyi açıklamak ve stratejik yönetime etkisini incelemek için faydalı bir yaklaşım olduğunu ifade etmişlerdir. Wolfson vd. (2004) kümeleme analizi kullanarak ülkeleri ekonomik değişkenlere göre gruplamanın ilgili iktisadi değişkenler arası ilişkileri ve genel ekonomik başarıma etkilerini değerlendirmenin bir yolu olduğu ifade etmişlerdir. Haldar vd. (2008) tıp alanında yaptıkları kümeleme analizinde hastaların verileri ile astım tiplerini belirlemişler ve buna göre ilaç tedavi planları öngörmüşlerdir. Koh ve Tan (2011) sağlık sektöründe çok sayıda üretilen veriden anlamlı bilgi türetmek ve etkin tedavi hizmeti sunmak için kümeleme analizinden sıklıkla faydalanıldığını belirtmişlerdir.

Ölçülebilir ve aynı değişken özelliklerine sahip gözlemlerin özellik değerlerinin benzerlik derecelerine göre gruplama işlemine kümeleme adı verilmektedir. Aynı kümedeki gözlemler birbirleriyle değişken özellikleri bakımından yakınlık gösterirlerken, aynı kümede olmayanlar farklılaşmaktadır. Temel kümeleme metotlarından olan birleştirici hiyerarşik kümelemede, ilk başta tüm gözlemler ayrı ayrı birer küme iken her bir aşamada gözlemlerin özellik uzaklık değerlerine göre en yakın gözlemler ve onların bulundukları kümeler birleşir. Eğer küme birleştirme işlemini sonlandırmak için belirli bir küme ayrışma uzaklık limit değeri veya küme sayısı belirlenmediyse kümeleme işlemi bitiminde gözlemlerin tamamı bir kümede birleşmektedir. Gözlemlere dair mesafe metrikleri Öklid ve Mahalanobis gibi mesafe ölçen fonksiyonlar ile değerlendirilebilmektedir (Shirkhorshidi vd., 2015: 1-5).

Gözlemleri kümelere ayırmada faydalanılan algoritmaların çoğunluğunda gözlemlerin kaça ayrılacağı veyahut aynı kümede bulunacak gözlemlerin mesafe limit değerleri ön bilgi olarak gerekmektedir. Gerçek hayatta ise bu tip ön bilgiler çoğunlukla bulunmamaktadır. Kümelemede ön bilginin bulunmadığı ya da güvenilir olmadığı hallerde kümeleme işlemini yapabilecek algoritmalara gereksinim duyulmaktadır. Küme merkez değerleri ve kümelere atanacak gözlemlerin belirlenmesi gibi sorunlara sahip problem karma tam sayılı doğrusal olmayan programlama halini almakta ve problemin çözümünde sezgisel algoritmalar kullanılmaktadır. Popüler bir sezgisel algoritma olan Parçacık Sürü Optimizasyonu (PSO) kümeleme problemlerinde herhangi bir ön bilgiye gereksinim duymadan probleme özgü bir uygunluk fonksiyonunu en iyileyerek küme sayısını ve kümelere yerleşecek gözlemleri etkin bir yapıda belirleyebilen ve kümeleme analizi problemlerinde çoğunlukla kullanılan bir metottur.

Van der Merwe ve Engelbrecht (2003) yaptıkları çalışma ile PSO'yu kümeleme analizine adapte etmişler ve çalışmalarında altı farklı veri setinin kümelenmesinde sonuç olarak önerdikleri yaklaşım iyi başarım göstermiştir. Chen ve Ye (2004) gözlemlerin ait oldukları küme merkezlerine uzaklığı minimize etmeyi hedefleyen uygunluk fonksiyonunu ve belirli küme sayısını kısıt olarak ele aldıkları kümeleme modelini PSO ile çözmüşlerdir. Omran vd. (2006) kümeleme analizinde sonuçları değerlendirmek için faydalanılan geçerlik ölçütlerini uygunluk fonksiyonu yerine kullanarak kümeleme için PSO ve kümeleme yöntemlerini bir arada uygulamışlar ve uygunluk fonksiyonları açısından kümeleme sonuçlarını değerlendirmişlerdir. Das vd. (2008) benzerlik ölçütü olarak normal dağılışı baz alan bir kernel fonksiyonunu

oluşturmuşlar ve bunu bilinen veri setlerini kümelemede uygunluk fonksiyonu olarak kullanmışlardır. Önerdikleri Çoklu Elitist PSO ile kümeleme gerçekleştirmişler ve sonuçları klasik PSO, Genetik Algoritma, Dinamik PSO çözümleri ile kıyaslayarak önerdikleri metodun etkinliğini göstermişlerdir. Cura (2012) belirli ve belirli olmayan küme sayısına göre farklı şekilde modellediği kümeleme problemi için uyarladığı PSO'yu, hibrid K-ortamalar PSO, Yapay Arı Kolonisi Algoritması, Karınca Kolonisi Optimizasyon Algoritması gibi bilinen metotlarla kıyaslamıştır. Literatürde yer alan ve yapay ürettiği veri setlerini kümelemede kullanmış ve önerdiği metodun başarımının daha yüksek olduğu sonucuna ulaşmıştır. Ortakçı ve Göloğlu (2012) PSO ile önceden belirli olmayan küme sayısına göre kümeleme için önerdikleri metot ile bilinen bir veri seti ve yapay bir veri seti için sonuç elde etmişler ve sonuçlara göre önerdikleri yaklaşımın etkinliğini kümeleme doğruluk indeksi ile değerlendirmişlerdir. Ghorpade ve Metre (2014) kümelemede karşılaşılan zorlukları giderebilmek için gözlemlere ait değişken değerlerinin varyanslarının kullanılabilirliği üzerinde durmuşlar ve buna uygun geliştirdikleri uygunluk fonksiyonu ile ve kümeleme için geliştirdikleri PSO algoritması ile dokuz farklı tipte veri setini kümelemişlerdir. Sonuçlara göre önerilen fonksiyon ve metot çok sayıda boyuta sahip verileri kümelemede oldukça başarılı olmuştur. Esmin vd. (2015) yaptıkları çalışmada kümeleme analizinde PSO ve varyantlarından faydalanan çalışmaları derinlemesine incelemişlerdir. Uygulamalara bakıldığında kümeleme analizi en çok kullanılan alan olurken ayrıca veri akışı kümeleme, internet madenciliği, belge kümeleme, özellik seçimi, görüntü işleme, metin kümeleme ve çeşitli endüstriyel uygulama alanlarında PSO ile kümeleme yapıldığını ifade etmişlerdir. Armano ve Framani (2016) yaptıkları çalışmada kümeleme problemine uyarladıkları PSO algoritmasını yirmi yedi farklı test veri setini kümelemede kullanmışlar ve dört farklı klasikleşmiş kümeleme yöntemine göre daha iyi sonuç elde ettiklerini ifade etmişlerdir. Uygunluk fonksiyonu olarak önerdikleri bağlantı ve birleşmeye dayalı iki uygunluk fonksiyonunu çok amaçlı bir yapıda kullanmışlardır. Özmen vd. (2018) ele aldıkları telekomünikasyon sektöründe müşteri segmentasyonu probleminde bir küme geçerlilik indeksi olan Davies-Bouldin fonksiyonunu uygunluk fonksiyonu olarak PSO ile çözümde kullanmışlardır. Alswaitti vd. (2018) önerdikleri hibrit yöntem olan hiyerarşik kümeleme PSO'nun kümelemedeki başarımını on bir farklı veri seti üzerinde diğer klasik kümeleme yöntemleri ile karşılaştırmışlardır. Yöntemler isabet oranı, standart sapma ve Dunn İndeksi (Dİ) üzerinden kıyaslanmışlardır. Sonuç olarak önerilen yaklaşımın diğer klasik yöntemlere göre daha iyi sonuç verdiğini ifade etmişlerdir.

Sezgisel algoritmaların yapısı incelendiğinde, arama uzayında aramanın yönünü denetleyici olan parametrelerin alacağı değerleri belirlemek algoritmanın etkin çalışması için çok önemlidir. Çoğu zaman çok sayıda deneme yapılmasını gerektiren ve problemin matematiksel doğasından etkilenen parametre seçim süreci başlı başına zor bir problemdir. Sezgisel algoritmalarda, algoritma çalışmaya başlamadan önce belirlenmiş, sabit parametre değerleri veya algoritmanın her bir döngüsünde önceden belirli değişim oranları ile değişen parametreler sıklıkla kullanılmaktadır. Bu tip yaklaşımların eksik yanı ise arama uzayında döngüler boyunca bulunmuş sonuçlar göz ardı edilerek kullanılmamasıdır. Sezgisel algoritmanın parametre değerlerini döngülerdeki arama sonuç değerlerinin bilgisini kullanarak değiştirebilen adaptif yöntemler bu aşamada bir çözüm olarak öne çıkmaktadır. Adaptif yöntemlerden sıklıkla kullanılan bir tanesi ise sezgisel parametrelerin döngüler boyunca değişiminde bulanık küme teorisinden faydalanan bulanık adaptif yaklaşımdır.

Çalışmada kümeleme analizi için yeni bir Bulanık adaptif PSO (BPSO) önerilmiş ve önerilen yaklaşım ile PSO, kümelemedeki başarımları açısından karşılaştırılmıştır. BPSO ile kümeleme analizi ile ilgili yapılan çalışmalara bakıldığında; Niknam ve Amiri (2010) çalışmalarında girdi olarak en iyi amaç fonksiyon değeri ve en iyi değerin değişmeden devam ettiği iterasyon sayısı parametrelerini kullanarak PSO parametrelerini bulanık adaptif yaklaşımla kontrol altında tutmuşlardır. Önerdikleri hibrit yöntem ile küme sayısı bilindiği durumlar için

kümeleme problemini çözmüşler ve klasik yöntemlere göre daha başarılı olduklarını ifade etmişlerdir. Melin vd. (2013) çalışmalarında girdi parametreleri olarak döngü sayısı, parçacıkların arasındaki Öklid uzaklık ile hesaplanan parçacık çözüm çeşitliliği değeri ve parçacıkların en iyi parçacıktan uygunluk fonksiyonu ortalama farkının hesaplanması ile elde edilen hata değerini kullanmışlardır. Çıktı olarak ise PSO sezgisel parametrelerini Mamdani Tipi bulanık kurallar elde etmişlerdir. Önerdikleri BPSO yaklaşımında, sınıflamada kullanılmak üzere yeni bir bulanık sistem ve buna uygun kural tabanlarını tasarlamışlardır. Yaklaşımlarını zambak çiçeği veri setini sınıflamada test etmişlerdir. Ön bilgi olarak küme sayısı verisini kullandıkları çalışmada önerdikleri BPSO klasik PSO'ya göre daha iyi başarım göstermiştir. Duan vd. (2016) önerdikleri yaklaşım ile parçacık hızını parçacığın elde ettiği değerlerde iyileşme olmasına göre adaptif bir şekilde kontrol etmişler ve farklı veri setlerini küme sayısı ön bilgisi kullanarak kümelemişlerdir. Önerdikleri yöntemin klasik yöntemlere göre üstün geldiğini ifade etmişlerdir. Keerthana ve Akila (2016), Niknam ve Amiri (2010) tarafından önerilen yöntemde değişiklik yaparak önerdikleri hibrit BPSO ile zambak çiçeği veri setini küme sayısı ön bilgisini kullanarak kümelemişlerdir. Önerdikleri yöntem ile sapan verilerin kümelemede yarattığı problemi aştıklarını ifade etmişlerdir. Hasan vd. (2019) sosyal ağlardaki büyük verinin akışını kümeledikleri çalışmalarında adaptif PSO yaklaşımından faydalanarak küme sayısı bilinen durumlar için büyük veriyi kümelemişlerdir. Önerdikleri yaklaşım ile klasik yaklaşımlara göre kümeleme doğruluğu açısından daha yüksek başarım elde etmişlerdir. Genel olarak BPSO ile kümeleme çalışmalara bakıldığında küme sayısı ön bilgisi kullanılarak kümeleme gerçekleştirilmiştir.

Çalışmanın amacı, önerilen BPSO yaklaşımı ve yeni bir uygunluk fonksiyonu ile kümeleme problemlerinin çözümünde küme sayısı ön bilgisi kullanmadan etkin sonuç sağlamaktır. Gerçek hayatta çok sayıda veriye ve boyuta sahip olabilen kümeleme problemleri için önerilen bulanık adaptif yapısıyla daha hızlı sonuç alabilen bir yaklaşım geliştirilmesi hedeflenmiştir. Önerilen yöntem ve uygunluk fonksiyonunun başarımını test etmek için popüler kümeleme veri setlerinden faydalanılmıştır. Çalışmanın literatüre iki farklı katkısı bulunmaktadır. Bunlardan birincisi BPSO yönteminde parametre adaptasyonunda yeni bir yaklaşımdır. İkinci katkı ise küme sayısı ön bilgisi kullanılmadan veri setlerini kümelemeye yarayan yeni bir uygunluk fonksiyonu önerisidir.

2. PARÇACIK SÜRÜ OPTİMİZASYONU

Sürü halinde yaşayan hayvanların gıda arama sürecini taklit eden PSO algoritmasını Eberhart ve Kennedy (1995) geliştirmişlerdir. PSO ilk defa ortaya çıktıktan sonra üzerinde çokça çalışılan ve yenilik yapılan bir sezgisel algoritma olmuştur. Önemli yeniliklerden birini ise Shi ve Eberhart (1999) gerçekleştirmiş ve PSO döngülerinde parçacıkların arama hızını döngü sayısı arttıkça globalden yerele arama yönünü dönüştüren ve W_{IN} değişkeniyle sembolize edilen eylemsizlik ağırlığını parçacık hız vektörünün hesaplanmasına dahil etmişlerdir. Aladağ vd. (2012) ise PSO'da bulunan o anki en iyi parçacığa yakınsamayı sağlayan sosyal ve parçacığın kendi en iyisinin etrafını taramasına yarayan bilişsel parametrelerin döngü sayısı arttıkça artmasını temel alan bir yaklaşım geliştirmişlerdir. Aladağ vd. (2012) tarafından geliştirilen PSO alttaki adımlardaki şekilde ifade edilebilmektedir;

Adım 1: PSO popülasyonundaki bireyler olan parçacıkları (j=1, 2,…, pn) stokastik şekilde X_{ji} veri yapısına problemdeki boyut sayısı olan n adet noktayı gözeterek yerleştir.

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

Adım 2: Parçacık hız vektörünü her bir parçacık ve konumu için rassal oluştur ve V_{ji}'de muhafaza et.

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

Adım 3: Popülasyondaki parçacıkların her birisinin o ana kadar ki elde edilmiş en iyi uygunluk fonksiyon ve konum değerini saptayarak Pbest'de muhafaza et. Popülasyondaki parçacıkların tamamı içerisindeki tüm zamanlarda elde edilmiş en iyi uygunluk fonksiyon ve konum değerini saptayarak Gbest'de muhafaza et.

Adım 4: Maksimum döngü sayısı $\max t$, o anki döngü sayısı t iken, bilişsel en iyiye yaklaşma $c_1 = (c_{1i}, c_{1f})$, sosyal en iyiye yaklaşma $c_2 = (c_{2i}, c_{2f})$ ve çözüm uzayı keşfetme yönünü belirleyen $W_{IN} = (W_{IN1}, W_{IN2})$ parametrelerinin güncellenmesi ise aşağıdaki gibidir;

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

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

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

Adım 5: Popülasyondaki parçacıkların hız ve konumları ise aşağıdaki gibi değişmektedir;

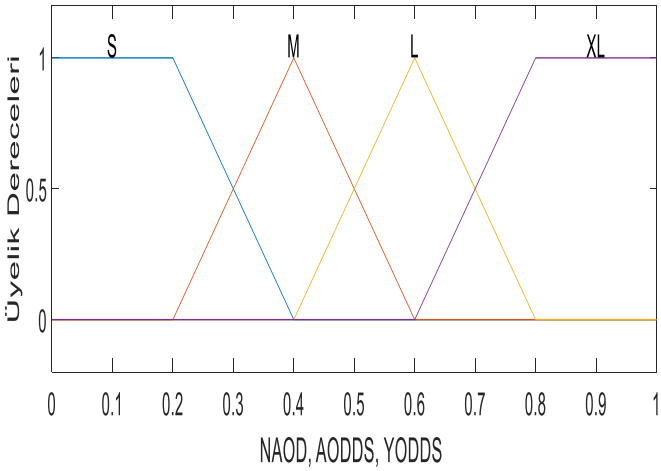
$$v_{i,n}^{t+1} = W_{IN} \times v_{i,n}^{t} + c_1 \times rand_1 \times (P_{i,n} - x_{i,n}) + c_2 \times rand_2 \times (P_{g,n} - x_{i,n})$$

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

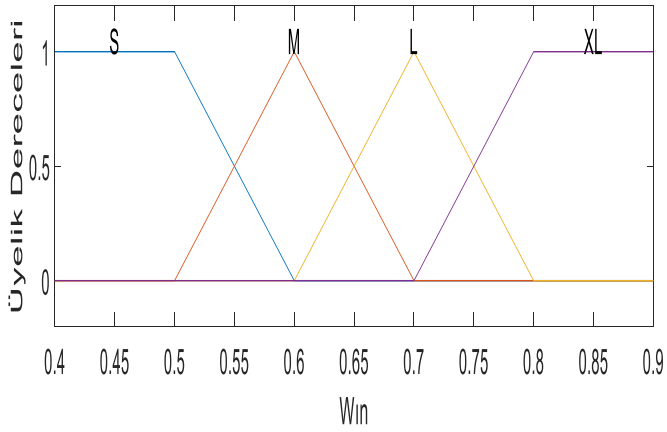
PSO algoritmasındaki adım 3, 4 ve 5 belirli bir maksimum döngü sayısına ulaşılana kadar sırayla tekrarlanarak çalıştırılır ve algoritma durduğunda en iyi uygunluk fonksiyon değeri ile ona ait konum, algoritmanın çıktısı ve sonucu olarak belirlenir.

3. BULANIK ADAPTİF PARÇACIK SÜRÜ OPTİMİZASYONU

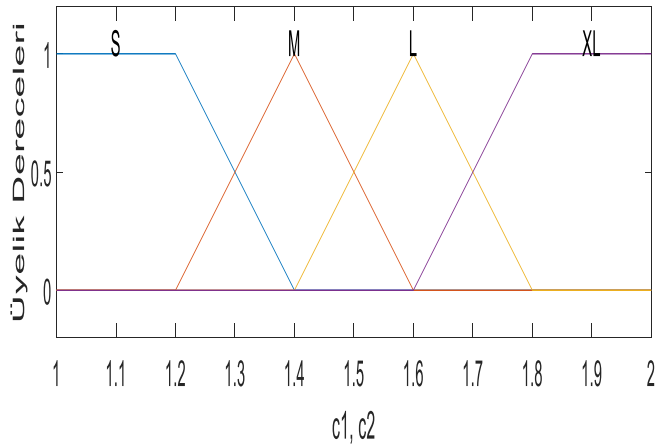
Niknam (2010), Niknam ve Amiri (2010) ve Melin vd. (2013) çalışmalarında PSO'da kullanılan W_{IN}, c_1 ve c_2 parametreleri değerlerinin optimizasyon sürecindeki bilgiler kullanılarak bulanık adaptif yapı ile kontrol altında tutulmasının algoritmaya katkıda bulunduğunu ifade etmektedirler. Çalışmada önerilen sezgisel parametrelerin bulanık kurallar ile değiştiği BPSO'da, Normalleştirilmiş Anlık Optimum Değer (NAOD) (Shi ve Eberhart, 2001), Anlık Optimumun Döngülerdeki Değişmeme Sayısı (AODDS) (Niknam, 2010) ve çalışmada BPSO için tarafımızca önerilen, Yerel Optimumun Döngülerdeki Değişmeme Sayısı (YODDS) parametrelerinden faydalanılmıştır. Tanımlanan başlangıç parametreleri ve bulanık kurallar ile W_{IN}, c_1 ve c_2 parametreleri elde edilmektedir. Şekil 1'de NAOD, AODDS ve YODDS parametreleri için geçerli olan bulanık kümeler verilmektedir. Şekil 2'de W_{IN} parametresi için bulanık kümeler, Şekil 3'te ise c_1 ile c_2 için geçerli olan bulanık kümeler bulunmaktadır. Şekillerde bulunan bulanık kümeler az (A-**S**), normal (N-**M**), çok (Ç-**L**) ve en çok (EÇ-**XL**) olarak ifade edilmiştir. NAOD ve YODDS girdileri c_1 üretmek için kullanılırken, NAOD ve AODDS girdileri ise WIN ile c_2 çıktı parametrelerinin üretiminde kullanılmıştır. Çalışmada kullanılan kural tabanları W_{IN} ve c_2 çıktıları için Niknam (2010) ile aynı olup, çalışmada NAOD ve tarafımızca önerilen YODDS girdileri kullanılarak c_1 çıktısı üretmek için yeni bir kural tabanı tarafımızca tasarlanmıştır. Bulanık kurallar Tablo 1-3'teki gibidir.



Şekil 1: NAOD, AODDS ve YODDS Parametreleri Bulanık Kümeleri



Şekil 2: W_{IN} Parametresi Bulanık Kümeleri



Şekil 3: c_1 ve c_2 Parametreleri Bulanık Kümeleri

Tablo 1: W_{IN} Parametresi Kural Tabanı

		AODDS			
		A	N	Ç	EÇ
NAOD	A	A	N	Ç	Ç
	N	M	N	Ç	EÇ
	Ç	Ç	Ç	Ç	EÇ
	EÇ	Ç	Ç	EÇ	EÇ

Tablo 2: c_1 Parametresi Kural Tabanı

		YODDS			
		A	N	Ç	EÇ
NAOD	A	EÇ	EÇ	N	N
	N	EÇ	Ç	A	A
	Ç	Ç	Ç	A	A
	EÇ	Ç	N	A	A

Tablo 3: c_2 Parametresi Kural Tabanı

		AODDS			
		A	N	Ç	EÇ
NAOD	A	EÇ	Ç	N	N
	N	Ç	N	A	A
	Ç	N	N	A	A
	EÇ	N	A	A	A

Kural tabanının ürettiği çıktılara bakıldığında; NAOD, AODDS, YODDS girdi değerleri sırasıyla EÇ, Ç ve A olsun. O halde Tablo 1'e göre W_{IN}, EÇ değerini, Tablo 2'ye göre c_1, Ç değerini, Tablo 3'e göre ise c_2, A değerini almaktadır.

BPSO'daki girdi değişkenlerinin elde edilişinde dikkat edilmesi gereken hususlar bulunmaktadır. NAOD değeri algoritmanın ilgili iterasyona kadar elde ettiği en iyi amaç fonksiyonu değeri ve Anlık Optimum Değer (AOD) olarak ifade edilen değerin normalleştirilmesi ile elde edilmektedir. Eşitlik 1'de ilgili problem için ulaşılabilecek teorik bir maksimum değer olan ve En İyi Değer (EİD) olarak adlandırılan değer ile istenmeyen ve optimumdan uzak bir değer olarak nitelenen En Kötü Değer (EKD) kullanılarak AOD normalize edilir ve NAOD elde edilmiş olur. Eşitlik 1 ile elde edilen NAOD, bu durumda algoritmanın başında AOD'nin EKD'ye yakın bir değer olması nedeniyle 1'e yakın bir değer alırken, algoritmada AOD'nin EİD'ye yaklaşması ile 0'a yakın bir değer almaktadır.

$$NAOD = \frac{E\dot{I}D - AOD}{E\dot{I}D - EKD} \tag{1}$$

AODDS parametresi kısaca iterasyonlar boyunca bulunmuş olan en iyi değer olan AOD'un ne kadar süreyle değişmediğine odaklanmaktadır. Örneğin algoritma 100. iterasyonda olsun. 100 iterasyon boyunca algoritmanın bu süreçte daha iyi sonuç elde edemeden devam ettiği en uzun iterasyon serisi 50 iterasyon olsun. En son AOD'da elde edilen iyileşme 80. iterasyonda olsun. O halde (AODDS= (100-80)/50) eşitliğinden faydalanılarak AODDS değeri 0.4 olarak elde edilir. YODDS'de AODDS'e benzer bir yapıya sahipken, onda farklı olarak her bir parçacık için ayrı ayrı kendi optimum değişimlerine odaklanılarak hesaplanmaktadır. BPSO'da ilgili girdi ve çıktı değişkenleri ve bulanık kuralların nasıl çalıştığına dair sayısal bir örnek ise aşağıda verilmiştir.

Bulanık küme mantığı gereği bir değer birden çok bulanık kümeye farklı bulanık üyelik değerlerine göre ait olabilmektedir. Fakat gösterim kolaylığı olması açısından sadece en büyük üyelikleri bulunan bulanık kümeler üzerinden işleyiş anlatılmaktadır. Çıktı parametrelerinin kesin olarak aldığı değerler girdi parametrelerinin tüm bulanık üyelik fonksiyonlarına aitliklerinden etkilendiği için çıktı değerleri için net ifadelerde bulunulmamıştır. 0 ile 1 aralığında değer alabilen NAOD, AODDS, YODDS herhangi bir iterasyonda sırasıyla 0.8, 0.6 ve 0.2 değerlerini almış olsun. O halde Şekil 1'e göre NAOD, AODDS ve YODDS girdi değerleri için en büyük bulanık üyelikleri sırasıyla EÇ, Ç ve A bulanık kümelerine olur. Bu girdi değerlerinden ilk önce W_{IN} değeri üretilmek istendiğini varsayalım. Tablo 1'deki bulanık kurallara göre NAOD EÇ iken ve AODDS ise Ç iken W_{IN} değişkeni EÇ değerini almaktadır. Şekil 2'de W_{IN} değişkeninin tanımlı aralığı 0.4 ile 0.9 arasındadır. Bu durumda W_{IN} değişkeni EÇ bulanık kümesinin tanımlı bulunduğu 0.8 ile 0.9 aralığında bir değer alacaktır. Benzer şekilde c_2, A bulanık kümesine daha fazla dâhil olacak ve 1 ile 1.2 değerleri arasında bir değer alacaktır. Parçacığa özel üretilen c_1 ise daha çok Ç bulanık kümesine yakın olacak ve 1.6 civarında bir değer alacaktır. Hesaplamalarda Mamdani tipi (Mamdani ve Assilian, 1975) bulanık çıkarım sistemi kullanılmaktadır.

Parametre c_1 için yeni bir elde ediliş önerisi ile geliştirdiğimiz BPSO algoritması, önemli bir PSO varyantı olan ve Aladağ vd. (2012) aracılığıyla ortaya çıkan PSO algoritması ile kümeleme analizi probleminde kıyaslanmıştır. Bu nedenle her iki algoritmada kullanılan sezgisel parametrelerin değer aralıkları Aladağ vd. (2012) tarafından yapılan çalışmada kullanılan aralıklarla aynı alınmış olup buna göre c_1 ile c_2 için (1, 2) arasındaki, W_{IN} için (0.4, 0.9) arasındaki sayısal değerler kullanılmıştır.

4. BPSO VE PSO İLE KÜMELEME

Küme sayısı bilinmediği durumlarda, BPSO ve PSO'da bulunan parçacıklar iki ayrı yapıdan oluşmaktadır. Bunlardan birincisi Aktif Küme Sayısı (AKS) bilgisini bulunduran skaler büyüklükte bir değişken iken küme merkezleri (KM) ise matris yapısında bulunmaktadır.

Parçacıkların sahip olduğu kümeler döngüler boyunca aktifleşmekte veya pasif hale gelerek algoritmada bulunmaktadır. Her bir döngüde parçacıkların sadece küme merkez değerleri değil ayrıca aktif küme sayısı değişebilmektedir. Çalışmada maksimum küme sayısı 10 ile sınırlanmış olup bir parçacığın en az 2 kümesi aktif olmalıdır. Veriler sadece aktif kümelere atanabilmektedir. Veriler kümelere atandıktan sonra verilere bağlı olarak yeni küme merkezleri hesaplanmaktadır.

Örneğin zambak çiçeği veri setini ele alalım. Bu veri setinde çiçekler dört adet nicel özelliğe göre kümelenmektedir. O halde her bir özellik ayrı satırlarda olacak şekilde 10 adet küme merkezi sütun vektörlerinin bileşiminden oluşan KM matrisi ve aktif küme sayısını belirten AKS için kümelemede kullanılan örnek parçacık kodlaması aşağıdaki gibidir.

AKS	3

Veri Seti Özellikleri	**Küme Merkezleri**									
Veri Seti Özelliği 1	0.43	**0.56**	0.76	**0.21**	0.3	**0.87**	0.11	0.09	0.65	0.91
Veri Seti Özelliği 2	0.47	**0.51**	0.36	**0.54**	0.6	**0.27**	0.71	0.29	0.55	0.92
Veri Seti Özelliği 3	0.63	**0.26**	0.46	**0.31**	0.35	**0.37**	0.18	0.39	0.61	0.51
Veri Seti Özelliği 4	0.37	**0.81**	0.31	**0.94**	0.67	**0.21**	0.61	0.27	0.85	0.72

Aktif 3 adet kümeye sahip parçacıkta yer alan değerler her bir özellik için küme merkez değerlerini ifade etmektedir. Aktif durumdaki 2, 4 ve 6. kümelerin altı çizilmiştir diğerleri ise pasif durumdadır. Parçacıktaki her iki yapı ayrı ayrı güncellenmektedir. Küme sayısı artması gerektiğinde rassal olarak yeni bir küme merkezi aktifleşirken, küme sayısı azalması gerektiğinde ise rassal olarak bir küme merkezi pasifleşmektedir.

Kümeleme problemlerinde gruplanacak verilerin kaç kümeye ayrılacağı bilinmediği durumlarda Dİ ve benzeri küme geçerlilik indeksleri kullanılmaktadır (Pakhira vd., 2004: 500). Çalışmada iç içe geçmiş verileri kümelemede yeni bir versiyonu önerilen ve kümeleme analizinde çokça kullanılan ve Dunn (1973) tarafından ortaya atılan Dİ Eşitlik 2'deki gibi ifade edilmektedir.

$$D\dot{I} = \min_{k=1,\ldots,K} \left\{ \min_{kk=1,\ldots,K} \left(\frac{dist(C_k, C_{kk})}{\max_{a=1,\ldots,K} diam(C_a)} \right) \right\} \tag{2}$$

Eşitlik 2'de C_k ve C_{kk} kümeleri arasındaki Öklid uzaklık değeri dist(C_k, C_{kk}) ile ifade edilmekte ve en yakın elemanlarının uzaklığı ya da küme merkezlerinin uzaklığı ile hesap edilebilmektedir. C_k ve C_{kk} kümelerinin birbirlerine en yakın elemanları u ve w olsun. O halde kümeler arası uzaklık ise Eşitlik 3'teki gibidir;

$$dist(C_k, C_{kk}) = \min_{u \in C_k, w \in C_{kk}} dist(u, w) \tag{3}$$

Bir kümede birbirlerine en uzak iki elemanın Öklid uzaklığı ise küme çapı olarak adlandırılmaktadır. Örneğin tüm kümeler içinde birbirine en uzak elemanlara sahip küme C_a

olsun ve en uzak elemanları da x ve y olsun. Buna göre C_a küme çapı Eşitlik 4'teki gibi ifade edilmektedir.

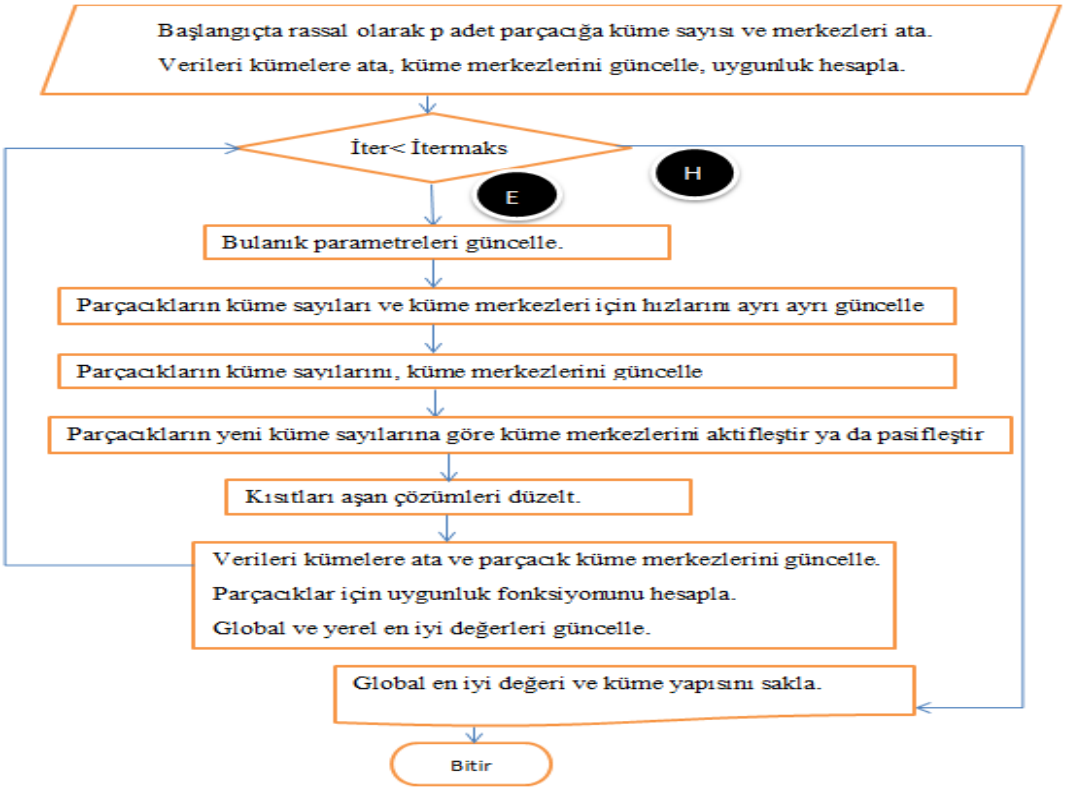
$$diam(C_a) = \max_{x,y \in C_a} dist(x,y) \tag{4}$$

Dİ'nin paydasında en büyük çapa sahip kümenin çap değeri bulunmaktadır. Dİ incelendiğinde payında küme merkezlerinin arasındaki mesafeyi açmak için en yakın küme merkezlerinin uzaklığını maksimize ederken, paydada ise küme içi yoğunluğu en büyük küme çapına sahip kümeyi minimize ederek sağlamaya çalışmaktadır.

Dİ'nin zambak çiçeği veri setini olması gerektiği gibi 3 değil de 2'ye bölen (Zhao vd., 2009: 319) ve şarap veri setini olması gerektiği gibi 3 kümeye ayıramadığından (Yeh vd., 2014: 169) tarafımızca Düzeltilmiş Dİ (DDİ) oluşturularak kullanılmıştır. Kümelemede PSO uygunluk fonksiyonu olarak kullanılan DDİ paydasında Dİ'den farklı olarak kümelerde bulunan eleman sayısını da dikkate alarak küme çapları arasındaki farkı minimize etmeye Eşitlik 5'teki gibi çalışmaktadır;

$$DD\dot{I} = \min_{k=1,\ldots,K} \left\{ \min_{kk=1,\ldots,K} \left(\frac{dist(C_k, C_{kk})}{\max_{a=1,\ldots,K} \left[(c(C_a)/n) * diam(C_a) \right] - \min_{b=1,\ldots,K} \left[(c(C_b)/n) * diam(C_b) \right]} \right) \right\} \tag{5}$$

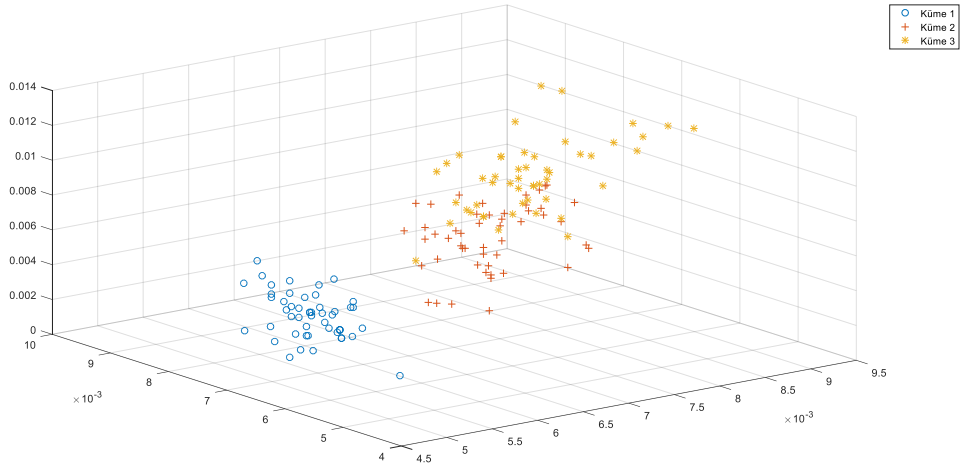
BPSO ile kümeleme yapılırken DDİ uygunluk fonksiyonunda küme merkezleri arasındaki Öklid uzaklık, kümeler arası uzaklıklar olarak ele alınmıştır. Eşitlik 5'te toplam veri sayısı n ile gösterilirken, a kümesine ait veri adeti c(C_a) olarak gösterilmektedir. Küme sayısı bilinmediği durumda kümeleme probleminin BPSO ile akış şeması Şekil 4'deki gibidir.



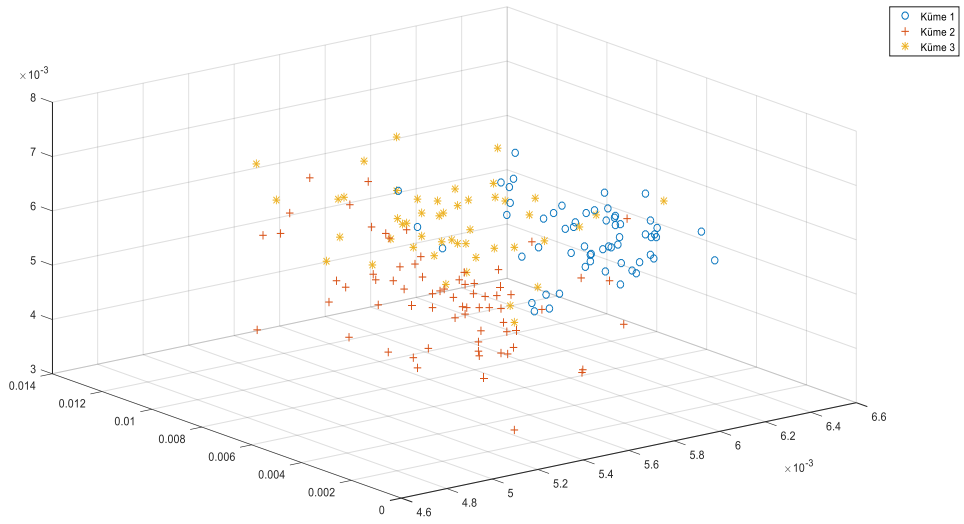
Şekil 4: BPSO ile Kümeleme Akış Şeması

5. UYGULAMA

Çalışmada, BPSO ve PSO sezgisel algoritmalarının bilinmeyen küme sayısına göre kümeleme analizinde başarımlarını karşılaştırmak için iki adet popüler veri öbeği kullanılmıştır. İlk veri öbeği olan zambak çiçeği veri seti 150 gözlem noktası ve 4 özellik boyutundan oluşmaktadır. Gerçekte üç kümeye ayrıldığı bilinen zambak çiçeği veri setinin gözlem noktaları kümelere eşit sayıda 50 şer adet dağılmaktadır. Kümeleme çalışması yapılan diğer gözlem grubu olan şarap veri setinde ise 178 gözlem noktası ve 13 boyutu bulunmaktadır. Gerçekte üç kümeye ayrıldığı bilinen şarap veri setinde sırasıyla kümelerde 59, 71 ve 48 gözlem noktası yer almaktadır. Zambak çiçeği ve şarap veri setlerinin gerçekte kümelere ayrılışı gösterimde ilk üç boyutları ile Şekil 5 ve Şekil 6'da gösterilmiştir.



Şekil 5: Zambak Çiceği Öbeği Gözlemlerinin Kümelere Ayrılışı



Şekil 6: Şarap Öbeği Gözlemlerinin Kümelere Ayrılışı

Çalışmada kümeleme için oluşturulan her iki algoritma BPSO ve PSO güncel bir yazılım geliştirme platformu olan MATLAB programlama dilinde kodlanmış ve çalıştırılmıştır. Önerilen BPSO'nun akıllı yapısıyla kümelemede normale göre düşük döngü ve parçacık sayısıyla çalıştığında da etkin sonuç alabildiğini gösterebilmek adına döngü ve parçacık sayısı sırasıyla 20 ve 5 olarak belirlenmiştir. BPSO ve PSO algoritmaları zambak çiçeği ve şarap veri setlerini kümelemek için DDİ uygunluk fonksiyonu ile 30'ar kez birbirinden bağımsız olarak çalıştırılmıştır. Kümeleme sonuçlarının tamamı Tablo 4'de verilmiştir.

Tablo 4'deki sonuçlara göre BPSO algoritmasına göre zambak çiçeği veri seti 26 kez 3 kümeye, 4 kez de 2 kümeye ayrılmış olarak elde edilirken, PSO algoritması ile 23 kez 3 kümeye ve 7 kez de 2 kümeye ayrılmıştır. Buna göre BPSO algoritması ile zambak çiçeği veri seti 30 kez ayrı ayrı kümelendiğinde denemelerin %87'si açısından literatürdeki gerçek değer olarak bilinen 3 kümeli yapı bulunmuşken optimum denemede gözlem noktalarını gerçek gruplarda eşleştirme yüzdesi %96 olarak elde edilmiştir. PSO algoritması ile zambak çiçeği veri seti için elde edilen 30 adet kümeleme sonucunun %77'si açısından literatürdeki gerçek değer olarak bilinen 3 kümeli yapı bulunmuşken optimum denemede gözlem noktalarını gerçek gruplarda eşleştirme yüzdesi %96 olarak elde edilmiştir. Şarap veri setinde ise çok az farkla PSO hem küme sayısını doğru bulmada hem de optimum sonuçta gerçek kümelere yerleştirme oranında BPSO'nun önünde yer almıştır. Genel olarak şarap veri seti için her iki algoritma da doğru küme sayısını bulmada düşük başarım gösterirken gerçek gruplara yerleştirme konusunda yüksek başarım sergilemişlerdir.

Tablo 4: DDİ ile Küme Analizi Çıktıları

	BPSO ile Kümeleme Çıktıları				**PSO ile Kümeleme Çıktıları**			
Veri Seti	**Tekrar Sayısı**	**Küme Sayısı**	**Küme Sayısını Doğru Bulma Oranı**	**Gerçek Gruplara Eşleştirme Yüzdesi**	**Tekrar Sayısı**	**Küme Sayısı**	**Küme Sayısını Doğru Bulma Oranı**	**Gerçek Gruplara Eşleştirme Yüzdesi**
Zambak Çiçeği Veri Seti	26	3	%87	%96	23	3	%77	%96
	4	2			7	2		
Şarap Veri Seti	9	3	%30	%89	10	3	%33	%91
	7	2			7	4		
	5	4			5	5		
	5	5			3	2		
	2	8			2	7		
	1	6			1	6		
	1	7			1	8		
					1	9		

Öte yandan önerilen DDİ uygunluk fonksiyonunun kümelemede etkinliğini göstermek için kümeleme analizinde en popüler geçerlilik indekslerinden olan Dİ ile de BPSO ve PSO ile kümeleme gerçekleştirilmiştir. Sonuçlar Tablo 5'teki gibidir. Sonuçlarda Dİ uygunluk fonksiyonu ile hem BPSO hem de PSO algoritmasına göre zambak çiçeği veri seti 30 kez 2 kümeye ayrılmış olarak elde edilirken, Buna göre Dİ uygunluk fonksiyonu kullanan her iki algoritma da zambak çiçeği veri seti 30 kez ayrı ayrı kümelendiğinde denemelerin hiçbirinde literatürdeki gerçek değer olarak bilinen 3 kümeli yapı bulunamamıştır. Optimum denemede gözlem noktalarını gerçek gruplarda eşleştirme yüzdesi BPSO ve PSO için sırasıyla %66 ve %62 olarak elde edilmiştir. Şarap veri setinde ise az farkla BPSO hem küme sayısını doğru bulmada hem de optimum sonuçta gerçek kümelere yerleştirme oranında PSO'nun önünde yer almıştır. Genel olarak şarap veri seti için her iki algoritma da doğru küme sayısını bulmada düşük başarım gösterirken gerçek gruplara yerleştirme konusunda yüksek başarım sergilemişlerdir.

Tablo 5: Dİ ile Küme Analizi Çıktıları

	BPSO ile Kümeleme Çıktıları				**PSO ile Kümeleme Çıktıları**			
Veri Seti	**Tekrar Sayısı**	**Küme Sayısı**	**Küme Sayısını Doğru Bulma Oranı**	**Gerçek Gruplara Eşleştirme Yüzdesi**	**Tekrar Sayısı**	**Küme Sayısı**	**Küme Sayısını Doğru Bulma Oranı**	**Gerçek Gruplara Eşleştirme Yüzdesi**
Zambak Çiçeği Veri Seti	30	2	%0	%66	30	2	%0	%62
Şarap Veri Seti	13	3	%43	%85	11	3	%36.	%83
	6	4			9	4		
	4	2			5	2		
	4	5			3	5		
	3	6			2	6		

Ayrıca Tablo 4 ve 5'teki sonuçlara göre DDİ ile Dİ karşılaştırıldığında zambak çiçeği veri seti açısından önerilen DDİ, özellikle örtüşen yapıdaki zambak çiçeği veri setini kümelemede çok daha üstün başarım sergilerken, şarap veri seti için her iki uygunluk fonksiyonu arasında büyük farklılıklar gözlenmemiştir. Önerilen BPSO yöntemi ise PSO'ya oranla genel olarak daha yüksek başarım sergilemiştir.

6. SONUÇ

Kümeleme analizi verinin toplanması ve depolanmasının kolaylaştığı ve büyük verinin ortaya çıktığı günümüzde önemli bir veri analizi yöntemidir. Çalışmada bilinmeyen küme sayısına göre veri setlerini kümelerken önerilen DDİ'nin, klasik Dİ ile doğru küme sayısına ayrılamayan örnek veri setlerini bilinen küme sayılarına göre başarılı bir şekilde ayırdığı ve hatta gerçek gruplara eşleştirme yüzde değerlerinin de oldukça yüksek çıktığı gözlenmiştir. Birçok disiplinde çok değişkenli yapıya sahip gözlem noktalarının özellik değerlerine göre gruplandırılmasına çözüm arayan kümeleme probleminde çalışmada önerilen BPSO algoritmasının bulanık adaptif yapısı ile akıllı hale gelmesiyle daha hızlı doğru sonuca ulaşmada başarılı olabileceği gözlenmiştir. Çalışmada yer alan veri setlerine göre önerilen BPSO'nun zambak çiçeği veri setini kümelerken bilinen küme sayısını elde etmede klasik yöntem olan PSO'ya nazaran daha isabetli olduğu gözlenirken şarap veri setinde ise iki algoritma birbirine

yakın fakat görece düşük isabet oranları elde etmiştir. Bu duruma şarap veri setinde 13 farklı özellik boyutunun olması ve algoritmaların görece az parçacık ve döngü sayılarında çalıştırılmasının neden olduğu görülmekle birlikte önerilen BPSO algoritmasının klasik PSO algoritmasına göre uygunluk fonksiyon değerine bakıldığında daha iyi sonuç elde ettiği görülmektedir. İleriki dönemlerde değişik özellikte veri öbekleri ve yüksek hacimli verileri kümelemede önerilen algoritmanın, kullanılan uygunluk fonksiyonu, bulanık kurallar ve girdi parametreleri geliştirilerek çok daha faydalı olacağı düşünülmektedir.

KAYNAKÇA

Aladağ, C. H., Yolcu, U., Egrioğlu, E., & 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.

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

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

Belbin, L., & McDonald, C. (1993). Comparing three classification strategies for use in ecology. *Journal of Vegetation Science*, 4(3), 341-348.

Chen, C.-Y., & 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., & Konar, A. (2008). Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm. *Pattern recognition letters*, 29(5), 688-699.

Duan, G., Hu, W., & Zhang, Z. (2016). A novel data clustering algorithm based on modified adaptive particle swarm optimization. International Journal of Signal Processing, Image Processing and Pattern Recognition, 9(3), 179-188.

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

Esmin, A. A., Coelho, R. A., & 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.

Ghorpade, J. A., & Metre, V. A. (2014). Clustering Multidimensional Data with PSO based Algorithm. *Soft Computing and Artificial Intelligence*, 87(6), 1-7.

Haldar, P., Pavord, I. D., Shaw, D. E., Berry, M. A., Thomas, M., Brightling, C. E., & Green, R. H. (2008). Cluster analysis and clinical asthma phenotypes. *American journal of respiratory and critical care medicine*, 178(3), 218-224.

Hasan, R. A., Alhayali, I., Royida, A., Zaki, N. D., & Ali, A. H. (2019). An adaptive clustering and classification algorithm for Twitter data streaming in Apache Spark. Telkomnika, 17(6). 3086-3099.

Keerthana, S., & Akila, M. S. (2016). An Efficıent Hybrıd Comparatıve Study Based On Aco, Pso, K-Means wıth K-Medoıds for Cluster Analysıs. International Research Journal of Engineering and Technology, 3(12). 905-911.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6), 441-458.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, 19(2), 64-72.

Mamdani, E.H. & Assilian, S. "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, 7(1), 1-13.

Melin, P., Olivas, F., Castillo, O., Valdez, F., Soria, J., & Valdez, M. (2013). Optimal design of fuzzy classification systems using PSO with dynamic parameter adaptation through fuzzy logic. *Expert Systems with Applications*, 40(8), 3196-3206.

Niknam, T. (2010). A new fuzzy adaptive hybrid particle swarm optimization algorithm for non-linear, non-smooth and non-convex economic dispatch problem. *Applied Energy*, 87(1), 327-339.

Niknam, T., & Amiri, B. (2010). An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. Applied soft computing, 10(1), 183-197.

Omran, M. G., Salman, A., & 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.

Özmen, M., Delice, Y., ve Aydoğan, E. K. (2018). Telekomünikasyon Sektöründe PSO ile Müşteri Bölümlenmesi. *Bilişim Teknolojileri Dergisi*, 11(2), 163-173.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, 20(2), 134-148.

Shi, Y., & 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.

Shi, Y., & Eberhart, R. C. (2001). Fuzzy adaptive particle swarm optimization. In Evolutionary Computation, 2001. Proceedings of the 2001 Congress on (Vol. 1, pp. 101-106). IEEE.

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, 10(12), e0144059. 1-20.

Van der Merwe, D. W., & 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.

Wolfson, M., Madjd-Sadjadi, Z., & James, P. (2004). Identifying national types: A cluster analysis of politics, economics, and conflict. *Journal of Peace Research*, 41(5), 607-623.

Yeh, J. H., Joung, F. J., & Lin, J. C. (2014). CDV index: a validity index for better clustering quality measurement. *Journal of Computer and Communications*, 2(04), 163-171.

Zhao, Q., Xu, M., & 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.

Extended Summary

Clustering Based on Fuzzy Adaptive Particle Swarm Optimization Approach

The aim of the study is to provide efficient results without using the preliminary information about number of clusters in the solution of clustering problems with the proposed Fuzzy Adaptive Particle Swarm Optimization approach and a new fitness function for clustering. It is aimed to develop a faster approach with the fuzzy adaptive structure proposed for clustering problems which can have many data and dimensions in real life. Popular clustering data sets were used to test the performance of the proposed method and the fitness function. The study has two different contributions to the literature. The first is a new approach to parameter adaptation in the BPSO method. The second contribution is to propose a new fitness function with high accuracy for clustering data sets without using the number of clusters.

Clustering offers different advantages such as; social and economic benefits and for future predictions. Clustering analysis is a process that grouping observations according to the degree of their similarity and dissimilarity. Observations in the same cluster show similarity with each other in terms of their variable characteristics, whereas those not in the same cluster differ. When there is no knowledge about the exact number of clusters there would be a fitness function which makes interior clusters more compact with minimizing the radius of the clusters while maximizing the separation levels for the clusters at the same time.

The majority of the algorithms used to divide the observations into clusters, or the distance limit values of the observations in the same cluster, are required as preliminary information. In real life, such information is often not available. Clustering algorithms are needed when preliminary information about clustering is not available or is not reliable. The problem with issues such as cluster center values and determination of observations to be assigned to clusters becomes mixed integer nonlinear programming and heuristic algorithms are used to solve the problem. Particle Swarm Optimization a popular heuristic algorithm, is a commonly used method in clustering analysis problems that can effectively determine the number of clusters and observations to be placed in clusters by optimizing a problem-specific fitness function without the need for any prior knowledge of clustering problems.

Particle Swarm Optimization algorithm that mimics the food search process of flocked animals is a heuristic research algorithm used widely in optimization problems. After the first appearance of Particle Swarm Optimization, it has become an important algorithm that has been widely studied and innovated. One of the important innovations and included the weight of inertia, which is symbolized by the W_{IN} variable, in the calculation of the particle velocity vector, which converts the search speed of the particles in the Particle Swarm Optimization iterations as the current number of the iteration increases then it changes the search direction from global to local. Another method is developed based on the social and convergence of the best particle in the Particle Swarm Optimization and the increase in the current number of iteration of the cognitive parameters used to scan the particle around its best.

When the structure of heuristic algorithms is examined, it is very important to determine the values of the parameters that control the direction of the search in solution space. The parameter selection process, which often requires a large number of experiments and is influenced by the mathematical nature of the problem, is a difficult problem in itself. In heuristic algorithms, fixed parameter values determined before the algorithm starts, or parameters that change with certain change rates in each iteration of the algorithm are frequently used. The shortcoming of such approaches is that they are not used in the search space by ignoring the results found during iterations. Adaptive methods that can control the parameter values of the heuristic algorithm by using the information of the search result values in iterations come to the

fore as a solution. One of the most commonly used adaptive methods is the fuzzy adaptive approach, which uses fuzzy set theory to control heuristic parameters across iterations.

In this study, popular clustering data sets were used to test the performance of the proposed method and the feasibility function. A new Fuzzy Adaptive Particle Swarm Optimization method is proposed and adopted for clustering analysis problem. There are considerations in obtaining input variables in the Fuzzy Adaptive Particle Swarm Optimization. Normalized Current Optimum Value and the normalized number of iterations that the global value remains unchanged are used for the input variables. New fuzzy logic rules also identified and proposed for obtaining heuristic parameters of Particle Swarm Optimization as output variables.

In this study, two popular data sets were used to compare the performance of the proposed Fuzzy Adaptive Particle Swarm Optimization and the classic Particle Swarm Optimization heuristic algorithms in clustering analysis where no preliminary information is available about the number of clusters. Both algorithms BPSO and PSO are run 30 times separately for each data set. With the intelligent structure of the proposed Fuzzy Adaptive Particle Swarm Optimization, the maximum number of iterations and particles are determined as 20 and 5, respectively, to show that it can achieve effective results even when working with low number of iterations and particles. According to the results of the clustering, the lily flower data set was obtained 26 times in 3 clusters, 4 times in 2 clusters with the proposed approach, while 23 times in 3 clusters and 7 times in 2 clusters with classical approach. While examining accuracy rate of the clustering it is seen that both approach have a maximum accuracy rate of %96. In the wine dataset, the PSO was slightly ahead of BPSO in finding the correct number of clusters as well as in the ratio of accuracy. In general, both algorithms for the wine data set showed low performance in finding the right number of clusters, while performing well in accuracy rate of clustering.

As a conclusion, it is observed that the BPSO algorithm proposed in the study in clustering problem can be successful in reaching the correct result faster with the fuzzy adaptive structure. According to the data set in the study, it is observed that BPSO is more accurate than PSO which is the classical method in obtaining the known number of clusters while clustering the lily flower data set while two algorithms have close but relatively low accuracy rates.