Metin Belgesi Kümelemede Metasezgisel Yöntemlere Dayalı Kümeleme Algoritmaları

Metaheuristics Based Clustering Algorithms on Document Clustering

Aytuğ Onan[1]
[1]Yazılım Mühendisliği Bölümü, Manisa Celal Bayar Üniversitesi, Manisa, Türkiye
aytug.onan@cbu.edu.tr

***Özetçe*—Kümeleme analizi, verileri benzerliklerine göre gruplarına ayıran önemli bir veri analizi tekniğidir. Belge kümeleme, kümeleme algoritmalarının metin belgeleri üzerinde uygulanması ile belgelerin etkin bir biçimde geri getiriminin, organizasyonunun, erişiminin ve özetlenmesinin olanaklı hale gelmesini sağlar. Belge kümeleme, metin belgelerinin organizasyonu, özetlenmesi ve sınıflandırılmasında kullanılabilir. Metasezgisel algoritmalar, aralarında kümeleme analizinin de yer aldığı birçok karmaşık eniyileme probleminin çözümünde uygulanmaktadır. Bu çalışmada, beş metasezgisel kümeleme algoritmasının (parçacık sürüsü eniyilemesi, genetik algoritma, guguk kuşu algoritması, ateşböceği algoritması ve yarasa algoritması) on beş metin veri seti üzerinde F-ölçütü aracılığı ile değerlendirilmiştir. Deneysel analizlerde, iki geleneksel kümeleme algoritması (K-ortalama ve ikiye ayırma K-ortalama) da dikkate alınmıştır. Deneysel analiz sonuçları, sürü zekâsına dayalı kümeleme algoritmalarının daha yüksek başarım elde ettiğini göstermektedir.**

Anahtar Kelimeler*—*belge kümeleme; sürü zekası; metasezgisel algoritmalar.

***Abstract*—Cluster analysis is an important exploratory data analysis technique which divides data into groups based on their similarity. Document clustering is the process of employing clustering algorithms on textual data so that text documents can be retrieved, organized, navigated and summarized in an efficient way. Document clustering can be utilized in the organization, summarization and classification of text documents. Metaheuristic algorithms have been successfully utilized to deal with complex optimization problems, including cluster analysis. In this paper, we analyze the clustering quality of five metaheuristic clustering algorithms (namely, particle swarm optimization, genetic algorithm, cuckoo search, firefly algorithm and yarasa algorithm) on fifteen text collections in term of F-measure. In the empirical analysis, two conventional clustering algorithms (K-means and bi-secting k-means) are also considered. The experimental analysis indicates that swarm-based clustering algorithms outperform conventional clustering algorithms on text document clustering.**

Keywords*—*document clustering; swarm intelligence; metaheuristic algorithms.

I. GIRIŞ

Kümeleme analizi, etiketsiz veri örüntülerinin benzerliklerine göre gruplara atanmasına yönelik öğreticisiz bir öğrenme yöntemidir. Kümeleme analizinde, her bir kümede yer alan veri nesnelerinin birbirlerine olabildiğince benzer ve diğer kümelerde yer alan veri nesnelerine olabildiğince farklı olmaları beklenmektedir. Veri nesnelerinin, herhangi bir etikete gereksinim duyulmaksızın kümelere atanmasına, birçok farklı alanda ihtiyaç duyulabilmektedir. Kümeleme analizi, aralarında, pazarlama, sosyoloji ve biyolojinin de yer aldığı birçok alanda uygulanmaktadır.

Kümeleme analizi, çizge teorisi, istatistik, bilgisayar bilimleri, bulanık mantık, örüntü tanıma gibi birçok farklı alandan teknik ve yöntemlerin kullanıldığı disiplinler arası bir araştırma ve uygulama alanıdır. Temelde, kümeleme, veri örüntülerinin belirli bir ölçütü en aza indirgeyecek ya da en çok yapacak biçimde parçalara ayrılmasına dayalı bir eniyileme problemi olarak modellenebilir [1]. Kümeleme probleminde amaç fonksiyonu, veri setinde yer alan nesnelere arasındaki temel istatistiksel ilişkilere dayalı olarak oluşturulur. Veri nesnelerinin, kümelere atanması işlemi, katı ya da bulanık olarak gerçekleştirilebilir. Kümeleme analizinde en yaygın kullanıma sahip, k-ortalama ya da bulanık c-ortalama gibi bölümleyici kümeleme yöntemleri, kümeleme problemi amaç fonksiyonunun eniyilenmesi işlemini yerel aramaya dayalı olarak gerçekleştirir. Yerel aramaya dayalı kümeleme algoritmalarının kümeleme başarımları, başlangıç durumundan oldukça etkilenmektedir. Bunun yanı sıra, k-ortalama gibi yöntemlerin yerel en iyiye takılmaları, algoritmaların başarımlarını düşürmektedir [2].

Belge kümeleme, metin belgeleri üzerinde, kümeleme algoritmalarının uygulanması ile metin belgelerinin etkin

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

bir biçimde organize edilebilmesini, özetlenebilmesini ve geri getirimini amaçlayan bir araştırma alanıdır [3, 4]. Belge kümeleme, belge organizasyonu, derlem özetleme ve belge sınıflandırma gibi alanlarda başarıyla uygulanmaktadır [5]. Bilgi ve iletişim teknolojilerindeki ilerlemeler ile birlikte, elektronik metin belgesi miktarı önemli ölçüde artmıştır. Metin belgelerinin etkin bir şekilde işlenmesi, önemli bir metin madenciliği problemi halini almıştır. K-ortalama gibi geleneksel kümeleme algoritmaları, metin belgeleri üzerinde başarıyla uygulanmaktadır.

Daha önce de değinildiği gibi, kümeleme analizinin bir eniyileme problemi olarak modellenmesi mümkündür [6]. Metasezgisel algoritmalar, karmaşık eniyileme problemlerinin çözümünde kullanılan etkin yöntemlerdir. Genetik algoritmalar [7], parçacık sürüsü eniyilemesi [8] ve arı kolonisi eniyilemesi [9] gibi metasezgisel yöntemler, kümeleme analizinde başarıyla uygulanmıştır. Bu çalışmada, sürü zekâsına dayalı beş kümeleme algoritmasının (parçacık sürüsü eniyilemesi, genetik algoritma, guguk kuşu algoritması, ateşböceği algoritması ve yarasa algoritması) on beş metin veri seti üzerinde F-ölçütü aracılığı ile değerlendirilmiştir. Deneysel analizlerde, iki geleneksel kümeleme algoritması (K-ortalama ve ikiye ayırma K-ortalama yöntemi) da dikkate alınmıştır.

Çalışmanın ikinci bölümünde belge kümelemede metasezgisel yöntemlerin uygulanmasına yönelik önceki çalışmalar tanıtılmaktadır. Üçüncü bölümde, geleneksel ve metasezgisel kümeleme algoritmalarına, dördüncü bölümde deneysel süreç ve sonuçlara yer verilmektedir. Beşinci bölüm çalışmanın temel sonuçlarını sunmaktadır.

II. İLGILI ÇALIŞMALAR

Bu bölümde, belge kümelemede metasezgisel kümeleme algoritmalarının uygulamalarına yönelik önceki çalışmalar tanıtılmaktadır. Song and Park [7] çalışmalarında, genetik algoritmaya dayalı kümeleme algoritması kullanmıştır. Metin belgelerinin temsilinde gizli anlamsal indeksleme yöntemi kullanılmıştır. Bir başka çalışmada, Hasanzadeh et al. [8] parçacık sürüsü eniyilemesi algoritması kullanarak metin belgelerini kümelere ayırmıştır. Bu çalışmada, metin belgelerinin temsilinde gizli anlamsal indeksleme yöntemi kullanılmıştır. Deneysel sonuçlar, parçacık sürüsü eniyilemesi algoritmasının K-ortalama algoritmasına kıyasla daha iyi sonuçlar verdiğini göstermektedir.

Vaijayanthi et al. [10] çalışmalarında, metin belgeleri üzerinde kümeleme için, karınca kolonisi eniyilemesi, tabu arama ve K-ortalama algoritmalarına dayalı melez bir kümeleme algoritması geliştirmiştir. Çalışma, iki aşamadan oluşmaktadır. Öncelikli olarak, K-ortalama algoritması metin verisi üzerinde kullanılmakta ardından, K-ortalama algoritması ile elde edilen bölümlemeler, karıncalar için başlangıç konumları olarak alınmaktadır. Azaryuon and Fakhar [11] tarafından gerçekleştirilen çalışmada, metin belgeleme için, iyileştirilmiş karınca kolonisi eniyilemesine dayalı bir kümeleme algoritması geliştirilmiştir. Geliştirilen yöntemde, sezgisel bir yaklaşım aracılığıyla, karınca hareketleri yönlendirilmiştir. Geliştirilen yöntemin başarımı, F-ölçütü aracılığıyla, K-ortalama algoritması ve geleneksel karınca kolonisi eniyilemesine dayalı kümeleme algoritması ile karşılaştırılmıştır. Bir başka çalışmada, Avanija and Ramar [12] tarafından, semantik benzerlik ölçütü ve parçacık sürüsü eniyilemesine dayalı bir metin belgesi kümeleme yöntemi geliştirilmiştir. Forsati et al. [13] tarafından geliştirilen çalışmada ise, K-ortalama algoritması ve harmoni arama algoritmasına dayalı melez bir kümeleme algoritması geliştirilerek metin belgesi kümeleme üzerinde uygulanmıştır.

Forsati et al. [9] tarafından geliştirilen bir başka çalışmada ise, iyileştirilmiş arı kolonisi eniyilemesi algoritması, metin belgesi kümeleme üzerinde uygulanmıştır. Onan et al. [4] tarafından geliştirilen çalışmada, iyileştirilmiş karınca kolonisi eniyilemesi kümeleme algoritması ve gizli Dirichlet tahsisi yöntemlerine dayalı bir algoritma geliştirilerek, metin belgesi kümeleme üzerinde uygulanmıştır.

III. KÜMELEME ALGORITMALARI

Bu bölümde, çalışmada kullanılan temel kümeleme algoritmaları ve sürü zekâsına dayalı ve metasezgisel kümeleme algoritmaları tanıtılmaktadır.

A. K-ortalama algoritması

K-ortalama algoritması (KM), en bilinen ve kullanılan kümeleme yöntemlerinden biridir. K-ortalama algoritması, küme sayısını (k) girdi parametresi olarak alır. Veri setini, küme içerisindeki benzerlikler yüksek ve kümeler arası benzerlikler düşük olacak şekilde bölümlere ayırır. Algoritma, rastgele olarak k tane nesnenin rastgele seçimi ile başlar. Geriye kalan her bir nesne, kümelerdeki nesnelerin ortalama değerlerine göre kendilerine en yakın kümelere atanır. Ardından, her bir küme için nesnelerin ortalama değeri hesaplanarak küme ortalamaları güncellenir. Süreç, değişiklik olduğu sürece yinelenir. K-ortalama algoritması, basit yapıya sahip, ölçeklenebilir ve etkin bir yöntemdir. Büyük veri setleri üzerinde etkin bir biçimde çalışabilmektedir [14]. Algoritma, yerel minimuma takılabilmektedir. Bunun yanı sıra, algoritmanın etkin bir biçimde işleyebilmesi için başlangıçta rastgele seçilen merkezler önem taşımaktadır [15].

B. İkiye ayırma k-ortalama algoritması

İkiye ayırma k-ortalama algoritması (bisecting K-means, BKM), K-ortalama algoritmasına dayalı, bölümleyici ve hiyerarşik bir kümeleme yöntemidir [16]. İkiye ayırma k-ortalama algoritması, metin belgesi kümelemede sıklıkla kullanılan yöntemler arasındandır. Algoritmada, veri setini en uygun şekilde bölümlemek için seçilecek iki alt düğüm C_1 ve C_2 belirlenirken, ata küme C üzerinde, K-ortalama algoritması yinelemeli olarak işletilir. Veri setinin en uygun şekilde bölümlenebilmesi için birbirleriyle eş büyüklükte alt kümeler oluşturulur. Algoritma, ata küme C'nin seçilmesi ile başlar. Ata küme üzerinde iki düğüm rastgele olarak küme merkezleri olarak

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

belirlenir. Ardından, geriye kalan veri nesneleri alt kümelere belirli bir uzaklık ölçütü aracılığıyla ölçümlenen uzaklığa göre atanır. Süreç, küme merkezleri ve atanan nesneler değiştiği sürece yinelenir. Bu şekilde, veri seti *k* adet parçaya ayrılır [17].

C. Parçacık sürüsü eniyilemesi algoritması

Parçacık sürüsü eniyilemesi algoritması (PSO), kuş ve balık gibi organizmaların sosyal davranışlarından esinlenen toplum tabanlı bir metasezgisel yöntemdir. PSO algoritmasında problem, parçacık adı verilen rastgele etmenler aracılığıyla çözülür. Burada, her bir parçacık kendisi ile ilişkili bir hız parametresine sahiptir. Parçacıklar, arama uzayında, geçmiş davranışlarına ve sürüdeki diğer parçacıkların bilgisine dayalı olarak dinamik bir biçimde ayarlanan hızlar aracılığıyla hareket eder. PSO algoritmasında, her bir parçacığın konumu, ilgili parçacığın o ana kadar ziyaret etmiş olduğu en iyi konum ve komşuluğundaki diğer parçacıkların en iyi konumlarına göre belirlenmektedir [18]. PSO algoritmasının kümelemede uygulanabilmesi için *N*, veri setinde yer alan *n* adet veri nesnesini ve P={P_1, P_2,…,P_k} parçacıkları temsil etmek üzere, her bir parçacığa rastgele olarak ilk değer atanmaktadır. Her bir parçacığın uygunluk değeri F={F_1, F_2,…,F_k}, her bir parçacığın hızı V={v_1^i, v_2^i,…,v_k^i} ile ve her bir v_j^i={x_1, x_2,…,x_3} şeklinde veri noktalarını temsil etmektedir. Herhangi bir parçacığın konumu, Denklem (1) ve Denklem (2)'de belirtilen eşitliklere göre belirlenir:

$$a_j^i = a_j^i + v_j^i \quad (1)$$

$$v_j^i = w * v_j^i + c_1 * r * \left(pa_j^i - a_j^i\right) + c_2 * r * (ga^i - a_j^i) \quad (2)$$

Burada, *w*, ağırlık parametresini, *r* ise rastgele Uniform dağılışa göre seçilen parametre değerini temsil etmektedir.

Algoritma 1. PSO kümeleme adımları [25]

Girdi: Veri nesneleri: $P=\{p_1, p_2,\ldots,p_l\}$, *w*, c_1 ve c_2 parametreleri.
Çıktı: En yüksek uygunluk değerine sahip etmen *A*

$A=\{A_1, A_2, \ldots, A_k\}$ etmenlerin ilk değerlerinin atanması,
Her bir etmenin uygunluk değeri $F=\{F_1, F_2, \ldots, F_k\}$'nin hesaplanması:

$$F(A) = \sum_{i=1}^{l} Distance(A, p_i)$$

$PA=\{PA_1, PA_2, \ldots, PA_k\}$ etmenlerin ilk değerlerinin atanması,
$PF=\{PF_1, PF_2, \ldots, PF_k\}$ değerlerinin hesaplanması,
Küresel en iyi etmenin atanması (*GA*),
Küresel en iyi etmenin uygunluk değerinin hesaplanması (*GF*),
$V=\{V_1, V_2, \ldots, V_k\}$ ilk değerlerinin atanması,

Sonlandırma ölçütü sağlanmadığı sürece:
- Her bir A_i etmeninin Denklem (2)'ye göre güncellenmesi,
- İlgili etmenin uygunluk değerinin hesaplanması (F_i),
- Eğer $F_i < PF_i$ ise, $PA_i=A_i$ ve $PF_i=F_i$ olarak atanması,
- Eğer $PF_i<GF$ ise, $GA=PA_i$ ve $GF=PF_i$ olarak atanması,

Algoritma 1'de parçacık sürüsü eniyilemesi algoritmasının kümelemede uygulanmasına ilişkin temel adımlar sunulmuştur.

D. Ateş böceği algoritması

Ateş böceği algoritması (FA), arama uzayında en iyi çözümün belirlenebilmesi için, ateş böceği algoritmasının davranışlarından esinlenen toplum tabanlı bir metasezgisel yöntemdir [19]. Ateş böceği algoritması, temelde α, δ ve γ olmak üzere, üç parametreye dayalı olarak çalışmaktadır. α parametresi, her bir etmenin rastgele olması durumunu, δ parametresi, indirgeme oranını ve γ parametresi soğrulma katsayısını temsil etmektedir. Algoritma, her bir etmene rastgele ilk değer atanması ile başlar. Herhangi bir A_i etmeninin konumu, daha iyi konuma sahip tüm etmenlere dayalı olarak güncellenir. Herhangi bir A_i etmeninin, A_x yönündeki hareketi, Denklem (3) ve Denklem (4)'te belirtilen eşitliklere göre belirlenir:

$$a_j^i = a_j^i + d * e^{-\gamma * d^2} + \alpha * r \quad (3)$$

$$d = (a_j^x - a_j^i) \quad (4)$$

Algoritma 2'de ateş böceği algoritmasının kümelemede uygulanmasına ilişkin temel adımlar sunulmuştur:

Algoritma 2. FA kümeleme adımları [25]

Girdi: Veri nesneleri: $P=\{p_1, p_2,\ldots,p_l\}$, α, δ ve γ parametreleri.
Çıktı: En yüksek uygunluk değerine sahip etmen *A*

$A=\{A_1, A_2, \ldots, A_k\}$ etmenlerin ilk değerlerinin atanması,

Sonlandırma ölçütü sağlanmadığı sürece:
- Her bir A_i etmeninin uygunluk değeri F_i'nin hesaplanması:

$$F(A) = \sum_{i=1}^{l} Distance(A, p_i)$$

- Her bir A_i etmeninin uygunluk değerini her bir A_j etmeni ile karşılaştır. Eğer $F_i>F_j$ ise, A_i etmenini Denkem (4)'e göre güncelle.
- $\alpha=\alpha*\delta$ olacak şekilde güncelle.

E. Guguk kuşu algoritması

Guguk kuşu algoritması (CSA), guguk kuşlarının yumurta konumlandırmasından esinlenen metasezgisel bir yöntemdir [20]. Olgun guguk kuşları yumurtalarını diğer kuş türlerinin yuvalarına yakın yerleştirir. CSA algoritması, temelde üç ilkeye dayalı olarak işletilir. Öncelikle, her bir guguk kuşu belirli bir zaman aralığında bir yumurta hazırlar ve bu yumurtayı rastgele olarak belirlediği bir yuvaya yerleştirir. Algoritmada, yüksek kaliteli yumurtalar içeren yuvalar, gelecek kuşaklara aktarılırken, diğer yuvalar elenir. Her bir konuk yuva sayısı sabit bir parametre değeri olarak alınır. Yerleştirilen herhangi bir yumurtanın kuş tarafından keşfedilme olasılığı $p_a \in [0,1]$ şeklinde alınan bir olasılık değerine göre belirlenir. Değinilen temel ilkelere dayalı olarak, belirli bir kuş, herhangi bir yumurtayı kaldırabilir,

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

herhangi bir yuvayı terk edebilir ve tamamıyla yeni bir yuva oluşturabilir [20, 21].

CSA algoritmasında, her bir yeni çözüm, Levy uçuşlarına dayalı olarak belirlenir. Herhangi bir $x_i^{(t+1)}$ çözümü için oluşturulacak Levy uçuşu, Denklem (5)'e göre belirlenir:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus Levy(\lambda) \quad (5)$$

Burada, $\alpha>0$ parametresi, problem ölçeğine ilişkin adım boyutunu temsil etmektedir ve genellikle bire eşittir. Algoritma 3'te CSA algoritmasının kümelemede uygulanmasına ilişkin temel adımlar özetlenmektedir:

Algoritma 3. CSA kümeleme adımları [25]

Girdi: Veri nesneleri: $P=\{p_1, p_2,...,p_l\}$, p_a parametreleri.
Çıktı: En yüksek uygunluk değerine sahip etmen A

$A=\{A_1, A_2, ...,A_k\}$ etmenlerin ilk değerlerinin atanması,
Her bir etmenin uygunluk değeri $F=\{F_1, F_2, ...,F_k\}$'nin hesaplanması:

$$F(A) = \sum_{i=1}^{l} Distance(A, p_i)$$

Tüm uygunluk değerleri içerisinden en düşük değeri (F_{min}) ve karşılık gelen etmeni belirle (A_{min}),
TA'yı geçici A etmeni olarak ilke,
TF'yi geçici F uygunluk değeri olarak ilkle,

Sonlandırma ölçütü sağlanmadığı sürece:
- Her bir A_i etmeninin Denklem (5)'e göre güncellenmesi,
- İlgili etmenin uygunluk değerinin hesaplanması (F_i),
- Eğer $TF<F_i$ ise, $A_i=TA$ ve $F_i=TF$ olarak atanması,

Tüm uygunluk değerleri içerisinden en düşük değeri (F_{min}) ve karşılık gelen etmeni belirle (A_{min}),
- Her bir A_i etmeni için rastgele bir (r) değeri al,
- Rastgele seçilmiş bir A_{rnd} belirle ve TA'yı buraya yerleştir,
- TF uygunluk değerini hesapla,
- Eğer $TF<F_i$ ise, $A_i=TA$ ve $F_i=TF$ olarak atanması,

Tüm uygunluk değerleri içerisinden en düşük değeri (F_{min}) ve karşılık gelen etmeni belirle (A_{min}),

F. Yarasa algoritması

Yarasa algoritması (BA), yarasa algoritmasından esinlenen metasezgisel bir yöntemdir [22]. Yarasa algoritması, ld, pr, fq_{min} ve fq_{max} olmak üzere dört temel parametreye dayalı olarak çalışır. Algoritmada, ld parametresi, gürültü şiddetini, pr parametresi titreşim oranını, fq_{min} ve fq_{max} parametreleri sırasıyla minimum ve maksimum alan sıklığını temsil eder. Öncelikli olarak her bir etmen ve karşılık gelen uygunluk değerlerine ilk değerler atanır. Bunun yanı sıra, minimum uygunluk değeri ve karşılık gelen etmen kaydedilir. Her bir etmenin hız parametresi (V) için ilk değer atanır. Herhangi bir A_i etmeninin konumu, Denklem (6) ve Denklem (7)'de belirtilen eşitliklere göre belirlenir:

$$a_j^i = a_j^i + v_j^i \quad (6)$$

$$v_j^i = v_j^i + fq * (a_j^{min} - a_j^i) \quad (7)$$

Burada, fq parametresi rastgele olarak seçilir. Algoritma 4'te yarasa algoritmasının kümelemede uygulanmasına ilişkin temel adımlar sunulmaktadır:

Algoritma 4. BA kümeleme adımları [25]

Girdi: Veri nesneleri: $P=\{p_1, p_2,...,p_l\}$, ld, pr, fq_{min} ve fq_{max} parametreleri.
Çıktı: En yüksek uygunluk değerine sahip etmen A

$A=\{A_1, A_2, ...,A_k\}$ etmenlerin ilk değerlerinin atanması,
Her bir etmenin uygunluk değeri $F=\{F_1, F_2, ...,F_k\}$'nin hesaplanması:

$$F(A) = \sum_{i=1}^{l} Distance(A, p_i)$$

Tüm uygunluk değerleri içerisinden en düşük değeri (F_{min}) ve karşılık gelen etmeni belirle (A_{min}),
$V=\{V_1, V_2, ...,V_k\}$ ilk değerlerinin atanması,
TA'yı geçici A etmeni olarak ilke,
TF'yi geçici F uygunluk değeri olarak ilkle,

Sonlandırma ölçütü sağlanmadığı sürece:
- Her bir A_i etmeni için rastgele bir sayı belirle,
- A_i etmenini Denklem (7)'ye göre güncelle ve TA'ya ata,
- Rastgele bir $r1$ sayısı oluştur ($r1\sim N(0,1)$ olacak şekilde),
- Eğer $r1<pr$ ise, TA_i'yi belirtilen formüle göre güncelle:

$$a_j^i = a_j^{min} + 0.001 * r$$

- TF uygunluk değerini hesapla,
- Rastgele bir $r2$ sayısı oluştur ($r2\sim U(0,1)$ olacak şekilde),
- Eğer $TF<F_i$ ve r2<ld ise, $A_i=TA$ ve $F_i=TF$ olarak ata,

Tüm uygunluk değerleri içerisinden en düşük değeri (F_{min}) ve karşılık gelen etmeni belirle (A_{min}),

G. Genetik algoritma

Genetik algoritma (GA), evrim teorisinden esinlenen metasezgisel bir yöntemdir. Genetik algoritma ile belirli bir problemin çözülebilmesi için her bir çözüm bir kromozom ile temsil edilir. Her bir kromozomun yararlılığı, uygunluk fonksiyonu adı verilen fonksiyon aracılığıyla değerlendirilir [23]. Genetik algoritma, rastgele olarak bir toplum oluşturulması ile başlar. Genetik algoritma, temelde çaprazlama, mutasyon ve seçilim operatörlerine dayalı olarak işlemektedir.

IV. Deneysel Süreç Ve Deneysel Sonuçlar

Bu bölümde, metasezgisel yöntemlere dayalı kümeleme algoritmalarının başarımlarının değerlendirilmesinde kullanılan metin veri setleri, deneysel süreç ve algoritmalara ilişkin parametre bilgileri, değerlendirme ölçütü ve deneysel sonuçlar sunulmaktadır.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

A. Veri setleri

Kümeleme algoritmalarının başarımlarının değerlendirilmesi için, e-posta, bilimsel metin, web sayfası, haber metni, duygu analizi, biyomedikal veri gibi farklı alanlardan on beş temel metin belgesi veri seti kullanılmıştır. Çalışmada kullanılan veri setleri, metin belgeleri üzerinde sınıflandırma ve kümeleme algoritmalarının başarımlarının değerlendirilmesinde kullanılan temel setler arasındadır [24]. Tablo 1'de deneysel analizlerde kullanılan veri setlerine ilişkin temel özellikler (belge sayısı, öznitelik sayısı ve küme sayısı) sunulmaktadır. Burada, öznitelik sayısı, veri temsilinde vektör uzay modeli kullanıldığında elde edilen öznitelik sayısını belirtmektedir.

Veri seti	**Özellikler**			
	Alan	**Belge sayısı**	**Öznitelik Sayısı**	**Küme Sayısı**
20ng	E-posta	18808	45434	20
ACM	Bilimsel	3493	60768	40
Classic4	Bilimsel	7095	7749	4
CSTR	Bilimsel	299	1726	4
Dmoz-Business-500	Web sayfası	18500	8303	37
Enron-Top-20	E-posta	13199	18194	20
FBIS	Haber	2463	2001	17
Irish-sentiment	Duygu analizi	1660	8659	3
La1s	Haber	3204	13196	6
La2s	Haber	3075	12433	6
Multi-Domain-Sentiment	Duygu analizi	8000	13360	2
Pubmed-Cancer	Biyomedikal	65991	28329	12
Tr11	TREC Belgesi	414	6430	9
Tr12	TREC Belgesi	313	5805	8
Tr21	TREC Belgesi	336	7903	6

Tablo I. Veri setine ilişkin temel özellikler [24]

B. Deneysel süreç

Deneysel analizlerde karşılaştırılan kümeleme algoritmalarının gerçekleştirimi Java dilinde yapılmıştır. Deneysel analizler, Intel Core i7 CPU 3.40 GHz özelliklerine sahip bir bilgisayarda gerçekleştirilmiştir. K-ortalama algoritması için, her bir veri setinde yer alan küme sayısı girdi parametresi olarak verilmiştir. Sürü zekâsına dayalı kümeleme algoritmalarından sonuçların elde edilmesinde kullanılan parametre değerleri, Tablo 2'de sunulmuştur. Metasezgisel kümeleme algoritmalarının başarımları doğrudan algoritmanın parametre değerlerine dayalı olarak değişmektedir. Bu doğrultuda, her bir algoritma ile deneysel analizler yapılmış, metin veri setleri üzerinde en yüksek başarımın elde edildiği değerler alınarak, deneysel sonuçlar bölümünde listelenmiştir. Bunun yanı sıra, deneysel sonuçlar, her bir metasezgisel algoritmanın 50 kez çalıştırılması sonucu elde edilen ortalama değerleri belirtmektedir.

Algoritma	**Parametre Değerleri**
PSO	w: 0.75, c_1: 1.5, c_2: 1.5, k: 15
FA	α: 0.5, γ: 0.3, δ: 0.95, k: 15
CSA	p_a: 0.25, k: 15
BA	ld: 0.5, pr: 0.5, fq_{min}: 0, fq_{max}: 2, k: 15
GA	Çaprazlama oranı: 0.5, mutasyon oranı: 0.04, Toplum büyüklüğü: 50

Tablo II. Kümeleme algoritmalarının parametre değerleri

C. Değerlendirme ölçütü

Kümeleme algoritmalarının başarımlarının değerlendirilmesinde F-ölçütü kullanılmıştır. F-ölçütü, duyarlılık ve geri çağırma ölçütlerine dayalı bir küme geçerleme ölçütüdür. Veri setindeki her bir i sınıfı n_i adet, kümeleme sonucu elde edilen her bir j kümesi n_i adet veri nesnesi içermek üzere, duyarlılık $(p(i,j))$ ve geri çağırma$(r(i,j))$ ile temsil edilmek üzere, F-ölçütü Denklem (10)'a göre hesaplanmaktadır:

$$p(i,j) = \frac{n_{ij}}{n_j} \tag{8}$$

$$r(i,j) = \frac{n_{ij}}{n_i} \tag{9}$$

$$F(i,j) = \frac{(b^2+1).p(i,j).r(i,j)}{b^2.p(i,j)+r(i,j)} \tag{10}$$

F-ölçütü [0, 1] aralığında değer almaktadır ve F-ölçütünün yüksek değer alması, kümeleme kalitesinin daha yüksek olduğunu göstermektedir. Burada, b, parametresi duyarlılık ve geri çağırmanın ağırlığına ilişkin bir değerdir ve genellikle b=1 olarak alınır.

D. Deneysel sonuçlar

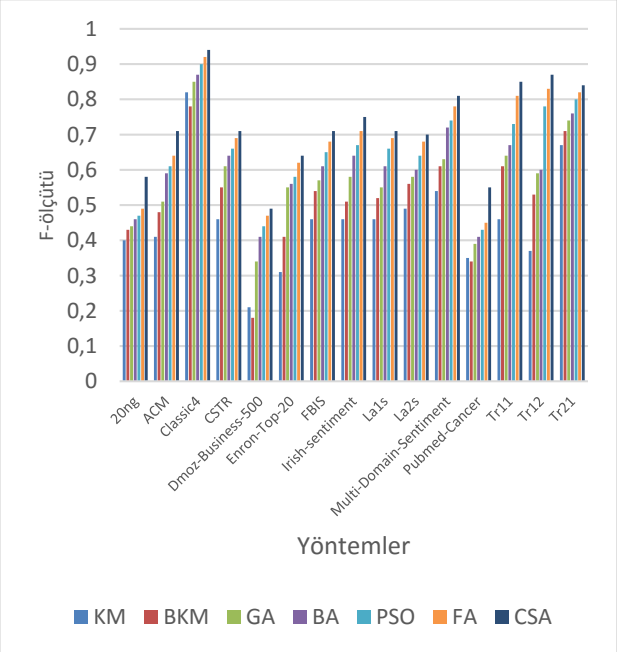
Tablo 3'te deneysel analizlerde kullanılan kümeleme algoritmaları ile elde edilen F-ölçütü değerleri sunulmaktadır. Deneysel sonuçlar incelendiğinde, sürü zekâsına dayalı kümeleme algoritmalarının F-ölçütü bakımından, deneysel analizlerde kullanılan metin belgesi veri setleri üzerinde, geleneksel kümeleme algoritmalarına kıyasla (K-ortalama ve ikiye ayırma k-ortalama algoritması) daha yüksek başarım elde ettiği gözlenmektedir. İkiye ayırma k-ortalama algoritması (BKM) ve K-ortalama (KM) algoritmalarının veri setleri üzerindeki başarımları incelendiğinde, BKM algoritmasının genellikle daha yüksek (daha iyi) F-ölçütü değerleri elde ettiği görülmektedir. Ancak, deneysel analizlerde kullanılan on beş veri seti

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

içerisinden yalnızca üç tanesi için (Classic4, Dmoz-Business-500 ve Pubmed-cancer veri setleri), K-ortalama algoritması, ikiye ayırma k-ortalama algoritmasına kıyasla daha iyi sonuçlar vermiştir.

Veri seti	KM	BKM	GA	BA	PSO	FA	CSA
20ng	0.4	0.43	0.44	0.46	0.47	0.49	**0.58**
ACM	0.41	0.48	0.51	0.59	0.61	0.64	**0.71**
Classic4	0.82	0.78	0.85	0.87	0.9	0.92	**0.94**
CSTR	0.46	0.55	0.61	0.64	0.66	0.69	**0.71**
Dmoz-Business-500	0.21	0.18	0.34	0.41	0.44	0.47	**0.49**
Enron-Top-20	0.31	0.41	0.55	0.56	0.58	0.62	**0.64**
FBIS	0.46	0.54	0.57	0.61	0.65	0.68	**0.71**
Irish-sentiment	0.46	0.51	0.58	0.64	0.67	0.71	**0.75**
La1s	0.46	0.52	0.55	0.61	0.66	0.69	**0.71**
La2s	0.49	0.56	0.58	0.6	0.64	0.68	**0.7**
Multi-Domain-Sentiment	0.54	0.61	0.63	0.72	0.74	0.78	**0.81**
Pubmed-Cancer	0.35	0.34	0.39	0.41	0.43	0.45	**0.55**
Tr11	0.46	0.61	0.64	0.67	0.73	0.81	**0.85**
Tr12	0.37	0.53	0.59	0.6	0.78	0.83	**0.87**
Tr21	0.67	0.71	0.74	0.76	0.8	0.82	**0.84**

Tablo III. Kümeleme algoritmalarından elde edilen F-ölçütü değerleri



Şekil 1. Algoritmaların farklı veri setleri üzerindeki başarımları

Deneysel analizlerde karşılaştırılan beş metasezgisel kümeleme yöntemi içerisinde, en düşük F-ölçütü değeri genellikle genetik algoritma (GA) ile elde edilmiştir. Genetik algoritmayı, yarasa algoritması (BA) ve parçacık sürüsü eniyilemesi (PSO) algoritmaları takip etmektedir. Karşılaştırılan farklı veri setleri için en yüksek başarım, guguk kuşu algoritmasına dayalı kümeleme yöntemi ile (CSA) elde edilmektedir. F-ölçütü bakımından ikinci en iyi sonuçlar ise ateş böceği algoritması (FA) ile elde edilmektedir. Şekil 1'de deneysel analizlerde kullanılan algoritmalar ile veri setlerinde elde edilen başarım sonuçları özetlenmiştir.

V. SONUÇ

Kümeleme analizi, verileri benzerliklerine göre gruplarına ayıran önemli bir veri analizi tekniğidir. Belge kümeleme, kümeleme algoritmalarının metin belgeleri üzerinde uygulanması ile belgelerin etkin bir biçimde geri getiriminin, organizasyonunun, erişiminin ve özetlenmesinin olanaklı hale gelmesini sağlar. Belge kümeleme, metin belgelerinin organizasyonu, özetlenmesi ve sınıflandırılmasında kullanılabilir. Metasezgisel algoritmalar, aralarında kümeleme analizinin de yer aldığı birçok karmaşık eniyileme probleminin çözümünde uygulanmaktadır. Bu çalışmada, beş metasezgisel kümeleme algoritmasının (parçacık sürüsü eniyilemesi, genetik algoritma, guguk kuşu algoritması, ateşböceği algoritması ve yarasa algoritması) on beş metin veri seti üzerinde F-ölçütü aracılığı ile değerlendirilmiştir. Deneysel analizlerde, iki geleneksel kümeleme algoritması (K-ortalama ve ikiye ayırma K-ortalama) da dikkate alınmıştır. Deneysel analiz sonuçları, sürü zekâsına dayalı kümeleme algoritmalarının daha yüksek başarım elde ettiğini göstermektedir.

KAYNAKÇA

[1] S. Das, A. Abraham, A. Konar, *Metaheuristic clustering*, Springer, 2009.

[2] M.J.A. Hasan, S. Ramakrishnan, "A survey: hybrid evolutionary algorithms for cluster analysis", *Artificial Intelligence Review*, 36, 179-204, 2011.

[3] L.Alsumait, C.Domeniconi, "Chapter 5: Text clustering with local semantic kernels", *Survey of Text Mining* içinde kitap bölümü, Springer-Verlag, 87-105, 2008.

[4] A.Onan, H.Bulut, S.Korukoğlu, "An improved ant algorithm with LDA-based representation for text document clustering", *Journal of Information Science*, 43(2), 275-292, 2017.

[5] C.C. Aggarwal, C.X.Zhai, *Mining text data*, Springer, 2012.

[6] E.R. Hruschka, R.J.G.B. Campello, A.A. Freitas, A.C. Carvalho, "A survey of evolutionary algorithms for clustering", *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39, 133-155, 2009.

[7] W.Song, S.C.Park, "Genetic algorithm for text clustering based on latent semantic indexing", *Computers and Mathematics with Applications*, 57, 1901-1907, 2009.

[8] E.Hasanzadeh, M.Poyanrad, H.A.Rokny, "Text clustering on latent semantic indexing with particle swarm optimization algorithm", *International Journal of the Physical Sciences*, 7(1), 116-120, 2012.

[9] R.Forsati, A.Keikha, M.Shamsfard, "An improved bee colony optimization algorithm with an application to document clustering", *Neurocomputing*, 159, 9-26, 2015.

[10] P.Vaijayanti, A.M. Natarajan, R.Murugadoss, "Ants for document clustering", *International Journal of Computer Science*, 9(2), 493-499, 2012.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.

[11] K.Azaryuon, B.Fakhar, "A novel document clustering algorithm based on ant colony optimization algorithm", *Journal of Mathematics and Computer Sciences,* 7, 171-180, 2013.

[12] J.Avanija, K.Ramar, "Semantic similarity-based clustering of web document using fuzzy c-means", *International Journal of Computational Intelligence and Applications,* 14, 2015.

[13] R.Forsati, M.Mahdavi, M.Shamsfard, M.R. Meybod, "Efficient stochastic algorithms for document clustering", *Information Science,* 220, 269-291, 2013.

[14] J.Han, M.Kamber, *Data mining: concepts and techniques,* Morgan Kaufmann, 2006.

[15] S.Theordoridis, K.Koutroumbas, *Pattern recognition,* Academic Press, 1999.

[16] M.Steinbach, G.Karypis, V.Kumar, "A comparison of document clustering techniques", *KDD Workshop on Text Mining,* August 20, Boston/USA, 2000.

[17] C.K.Reddy, B.Vinzamuri, "Chapter 4: A survey of partitional and hierarchical clustering algorithms", *Data Clustering: Algorithms and Applications* içinde kitap bölümü CRC Press, 87-107, 2013.

[18] E.G.Talbi, *Metaheuristics: from design to implementation,* Wiley, 2009.

[19] X.S.Yang, *Nature-inspired metaheuristic algorithms,* Luniver Press, 2008.

[20] X.S. Yang, S.Deb, "Cuckoo search via Levy flights", *NABIC 2009 Congress,* December 9-11, Coimbatore/India, 2009.

[21] Onan, A., "Hybrid supervised clustering based ensemble scheme for text classification", *Kybernetes,* 46(2), 330-348, 2017.

[22] X.S. Yang, "A new metaheuristic bat-inspired algorithm", *In the Proceedings of the Nature inspired cooperative strategies for optimization (NICSO),* May 12-14, Granada/Spain, 2010.

[23] Obitko, Marek. "Introduction to genetic algorithms." URL http://www. obitko. com/tutorials/genetic-algorithms, 1998.

[24] R.G.Rossi, R.M.Marcacini, S.O.Rezende, "Benchmarking text collections for classification and clustering tasks", *Technical Report,* University of Sao Paulo, 2013.

[25] X.Min, L.Liu, Y.He, G.Fong, Q.Xu, K.Wong, "Benchmarking swarm intelligence clustering algorithms with case study of medical data", *Computerized Medical Imaging and Graphics,* 2016.

Bu çalışma ASYU-2017 Konferansında sunulmuş ve sadece özet olarak
Bildiri Özetleri Kitabı (ISBN: 978-605-84722-8-0)'nda yayımlanmıştır.