TÜBİTAK–2209-A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI

ARAŞTIRMA ÖNERİSİ FORMU

2019

Ekim (Güz) Dönem Başvurusu

2209/A ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJELERİ DESTEĞİ PROGRAMI ARAŞTIRMA ÖNERİSİ FORMU

A. GENEL BİLGİLER

Başvuru Sahibinin Adı Soyadı: Ömer KÖROĞLU
Araştırma Önerisinin Başlığı: SOS-Tabanlı Meta-Sezgisel Kümeleme Algoritmasının Tasarımı ve Geliştirilmesi
Danışmanın Adı Soyadı: Hamdi Tolga KAHRAMAN
Araştırmanın Yürütüleceği Kurum/Kuruluş: Karadeniz Teknik Üniversitesi

ÖZET

Nesnelerin interneti alanında yaşanan gelişmeler, elektronik ortamların/cihazların, bu ortamlardaki uygulamaların ve bu uygulamaları kullananların sayısındaki artış, verinin de beklenmedik bir hızla büyümesine yol açmaktadır. Bu gelişmelere bağlı olarak ortaya çıkan büyük veriyi bilgiye dönüştürme çabaları her geçen gün daha da artmakta ve stratejik öneme sahip bir konu haline gelmektedir. Veri madenciliği çalışmaları, ekonomiden siyasete günlük yaşantının her alanını etkilemektedir. Bu açılardan ele alındığında veri madenciliğinde kullanılan yöntemlerin önemi daha net anlaşılmaktadır. Kümeleme konusu, veri madenciliğinin en önemli çalışma alanlarından biri olmasının da ötesinde gözetimsiz (denetimsiz/danışmansız) çalışan yapay zekâ tabanlı uygulamaları geliştirmenin de başlıca yoludur. Günümüzde verinin akışındaki süreklilik ve dinamik yapısı dikkate alındığında gözetimsiz öğrenmenin ve nesneleri bu yolla gruplandırmanın önemi de anlaşılmaktadır. Sektörde akan veri olarak da adlandırılan dinamik veri bankacılıktan, elektronik ticarete, sosyal medya uygulamalarından endüstriyel otomasyon sistemlerine kadar birçok alanda kümelenmeye çalışılmaktadır. Kümeleme sürecinde en çok ihtiyaç duyulan yöntem ise nesnelerin benzerliklerine göre sınırlı ve belirli bir sayıda gruba ayrılmasıdır. Bu amaçla en sık kullanılan algoritmaların başında k-ortalamalar tekniği gelmektedir. k-ortalamalar algoritmasının kolay anlaşılabilir ve basit uygulama adımlarına sahip olması rakiplerine karşın üstünlük kurmasını sağlarken, veri sayısındaki ve verinin karmaşıklık düzeyindeki artış algoritmanın performansını olumsuz yönde etkilemektedir. Bu handikapları ortadan kaldırmak ve kümeleme başarısını artırmak amacıyla k-ortalamalar tekniğinin farklı uzaklık metrikleri, çeşitli ağırlıklandırma yöntemleri ve çeşitli teknikler ile birlikte kullanılarak melezleştirildiği görülmektedir. Bu proje çalışmasının amacı k-ortalamalar yönteminin kümeleme performansını artırmak ve kararlı hale getirmektir. Özellikle, veri sayısındaki ve verinin karmaşıklık düzeyindeki artışa bağlı olarak ortaya çıkan performans düşüşünü azaltmak ve genel olarak algoritmanın daha başarılı bir kümeleme hassasiyeti göstermesini sağlamak amaçlanmaktadır. Bu amaçla, k-ortalamalar algoritmasında halihazırda kullanılmakta olan nesneleri uzaklık esaslı olarak (küme merkezlerine olan uzaklıklarına bağlı olarak) gruplandırma yöntemi terk edilmektedir. Bunun yerine proje çalışmasında, meta-sezgisel arama algoritması ile nesnelerin gruplara atanması önerilmektedir. Meta-sezgisel arama (MSA) algoritmalarının en önemli özellikleri çok boyutlu ve karmaşık arama uzaylarında geleneksel matematik yöntemlerine kıyasla çok daha üstün bir performans sergilemeleridir. MSA algoritmalarının bu özelliğinden faydalanılarak, k-ortalamalar yönteminin çok boyutlu ve karmaşık arama uzaylarındaki handikapları ve performans kararsızlığı da ortadan kaldırılmaya ya da azaltılmaya çalışılacaktır. Geliştirilecek uygulama web tabanlı bir platform üzerinde çalışacaktır. Uygulamaya ait web sayfası hazırlanacak ve araştırmacıların erişimine açık olacaktır.

1. ÖZGÜN DEĞER

1.1. Konunun Önemi, Araştırma Önerisinin Özgün Değeri ve Araştırma Sorusu/Hipotezi

Konunun önemi: Veri, özellikle teknolojideki gelişmelerden kaynaklı olarak dünyanın en güncel ve stratejik öğesi haline gelmiştir. İnternet üzerinden ya da farklı ağlar üzerinden birbirine bağlı, birbirleriyle hızlı ve etkili bir şekilde haberleşebilen, veri ve bilgi paylaşımı sağlayan elektronik cihazların ve bu cihazlarda çalışan uygulamaların artması ile birlikte bireylerden devletlere, küçük ölçekli işletmelerden uluslararası ölçekli firmalara kadar birçok aktörü yakından ilgilendiren fırsatlar ve tehditler ortaya çıkmıştır. Üstelik her geçen gün teknoloji gelişmekte, elektronik cihazlar üzerinden gerçekleştirilen işler çeşitlenmekte ve buna bağlı olarak yaşamın her alanıyla ilgili olarak daha fazla veri toplanmaktadır. Veri sayısındaki ve niteliğindeki artış, veriye dayalı olarak karar veren ve çalışan sistemlerin de performanslarını iyileştirmektedir. Bu sayede daha hassas, ideale çok yakın hatta kesin cevaplar veren ve kararlı çalışan uygulamalar/sistemler geliştirilmektedir. Veri toplama ve işleme konusu ticaretten savunmaya sanayiine kadar geniş bir yelpazede şirketler ve devletler için stratejik/hayati derecede önemli bir konu haline gelmiştir. Örneğin internet üzerinden alışverişin yapıldığı alibaba.com, amazon.com gibi dünyanın en büyük firmaları daha fazla müşteriye ulaşabilmek, ulaştığı müşteriye daha fazla satış yapabilmek ve müşterilerini kaybetmeden artırabilmek için yapay zekâ tabanlı uygulamalardan faydalanmaktadırlar. Bu uygulamalar müşterilerden toplanan verilere göre kişiselleştirme sağlamakta ve müşterilerin ihtiyaçlarını öngörerek kişiye özgü bir alışveriş deneyimi sağlamaktadır. Benzer şekilde devletlerin silahlı kuvvetleri tehdit yaratabilecek füzeleri, roketleri, uçakları, gemileri kısaca tüm silahları onların elektronik ortamda bıraktıkları izler üzerinden teşhis ve takip edebilmektedirler. Bahsedilen tüm bu konular, bugün hayati derecede önem kazanan "Veri'nin" gelecekte stratejik önemini artırarak koruyacağını açıkça göstermektedir. Dolayısıyla verinin yaygın bir şekilde işlendiği yapay zekâ alanlarından biri olan kümeleme konusunda yapılacak algoritma geliştirme çalışmasının önemi ortadadır. Literatürde yaygın bir şekilde kullanılan k-ortalamalar yönteminin performansının artırılması/iyileştirilmesi

amaçlamaktadır. Bu amaca yönelik yapılacak çalışma ise yapay zekanın en önemli ve güncel konularından olan meta-sezgisel arama algoritmalarına dayanmaktadır. Güncel bir MSA tekniği olan SOS (symbiotic organisms search) [1] algoritması k-ortalamalar yöntemiyle melezlenerek güçlü ve etkili bir kümeleme algoritmasının geliştirilmesi amaçlanmaktadır. Bu sayede veri işleme konularından biri olan kümeleme alanında önemli bir çalışma gerçekleştirilecektir.

Araştırma önerisinin özgün değeri: Literatürde k-ortalamalar yöntemi ile ya da sezgisel algoritmalarla çözümlenmiş çok sayıda çalışmaya rastlanılmaktadır [2-10]. Problem bazlı çözümler şeklinde geliştirilmiş olan bu uygulamaların farklı problemler için yeniden kullanılabilirliği bulunmamaktadır. K-ortalamalar yönteminin kümeleme doğruluğunu iyileştirmeye yönelik çalışmalara da rastlanılmaktadır [11-14]. Ancak bu yöntemler arasında güçlü ve etkili bir MSA tekniği olan SOS [1] algoritmasının tatbik edildiği bir çalışmaya rastlanılmamaktadır. Üstelik literatürdeki çalışmalar arasında problemden bağımsız uygulanabilen bir çalışmaya da rastlanılmamaktadır. Bu projede yürütülmesi planlanan iki faaliyet, literatür açısından özgünlüğe sahiptir. Bunların ilki, güçlü ve modern bir MSA tekniği olan SOS algoritması ile k-ortalamalar yöntemi ilk defa melezlenerek yeni bir kümeleme algoritmasının geliştirilecek olmasıdır. Bu özellik, kümeleme algoritmasına teknik özgünlük kazandırmaktadır. İkincisi ise, kümeleme yaklaşımı açısından sağlanacak özgünlüktür. Problemin nitelik sayısından (boyutundan) ve veri sayısından bağımsız olarak çalışabilen bir algoritma tasarlanacaktır. SOS algoritmasıyla kazandırılacak olan sezgisel kümeleme yaklaşımı veri sayısından bağımsız bir çalışma sağlayacaktır. Geliştirilecek algoritmada, yazılım tasarım prensiplerinin tatbik edilmesiyle de problemden bağımsız tatbik edilebilen bir mimari çatı sağlanmış olacaktır.

Araştırma sorusu: k-ortalamalar yönteminin kümeleme performansını iyileştirmek için meta-sezgisel arama algoritmaları kullanılabilir mi? K-ortalamalar yönteminin temel handikabı olan büyük arama uzaylarında yerel çözüm tuzaklarına takılma problemi meta-sezgisel yöntemler kullanılarak aşılabilir mi?

Hipotezin dayandığı temeller: Meta-sezgisel arama algoritmaları, geleneksel matematik yöntemlerini ve kesin kuralları esas alan arama yöntemlerinin etkisiz ya da yetersiz kaldıkları problemlerin çözümlenmesinde başarılı olmaktadırlar. Çok boyutlu, konveks olmayan ve karmaşıklık düzeyi yüksek arama uzaylarında etkili olan bu yöntemler, k-ortalamalar algoritmasının zayıf yönlerini güçlendirebilirler. k-ortalamalar yöntemi özellikle büyük veri sayısına sahip ve çok boyutlu nesnelerden oluşan kümeleme problemlerinin çözümlenmesinde yetersiz kalmaktadır. Dolayısıyla meta-sezgisel arama algoritmalarının gücü ve k-ortalamalar yönteminin zayıf yönü birbirini tamamlar niteliktedir.

Hipotez: Meta-sezgisel arama algoritması ile k-ortalamalar yönteminden melezlenecek bir kümeleme algoritmasının karmaşık arama uzaylarındaki performansının geleneksel k-ortalamalar algoritmasına kıyasla iyileşmesi mümkündür.

1.2. Amaç ve Hedefler

Bu proje çalışmasındaki amaç, yapay zekanın önemli bir uygulama alanı olan kümeleme problemleri için problemden bağımsız olarak uygulanabilen güçlü ve etkili bir algoritma geliştirmektir. Bu amaca ulaşılma durumu ise aşağıdaki somut hedeflere varılması ile ölçülebilir. Proje sonucunda elde edilmesi beklenen somut çıktılar (hedefler):

i) *Geliştirilecek olan SOS-tabanlı k-ortalamalar yönteminin klasik k-ortalamalar algoritmasından daha üstün bir performans sergilemesi:* En az 6 farklı probleme ait veri seti (veri setleri uluslararası veri havuzundan elde edilecektir [15]) üzerinde yürütülecek olan deneysel çalışmalar yoluyla bu başarının ispatlanması sağlanacaktır. Geliştirilecek olan algoritma gerek doğruluk oranı (ya da ortalama % hata değeri) gerekse de hesaplama süresi açılarından k-ortalamalar yönteminden daha üstün bir performans sergileyecektir.

ii) *Geliştirilecek olan SOS-tabanlı k-ortalamalar yöntemi problemin nitelik sayısından (bağımsız değişken sayısı/problem boyutu) bağımsız şekilde tatbik edilebilmesi:* bu hedefe ulaşılma durumu, nitelik sayısı birbirinden farklı en az 6 kümeleme problemi üzerinden ispat edilecektir. Ayrıca, web ortamından çevrim içi zamanda uygulamanın farklı problemler ile çalışabilirliği sağlanacak ve gösterilecektir. Yani araştırmacılar kendi kümeleme problemlerini internet ortamında uygulamanın web sitesine giriş yaparak test edebileceklerdir.

iii) *Kullanım kolaylığı:* geliştirilecek olan uygulama kümeleme ya da algoritma konularında bilgisi olmayan araştırmacılar tarafından kolaylıkla kullanılabilecektir.

2. YÖNTEM

2.1. k-ortalamalar yöntemi

Hiyerarşik olmayan kümeleme yöntemleri arasında k-ortalamalar (k-means) yöntemi önem taşır ve yaygın biçimde kullanılır. Bu yöntemde, başlangıçta belirlenen küme sayısı için algoritmanın başarı oranını gösteren toplam kare hatayı minimize etmek amaçlanır [16].

K-ortalamalar algoritması aşağıdaki adımları tatbik ederek gerçekleştirilir:

1. Küme sayısı olan k değeri seçilir. (Bu değer kullanıcıdan alınan keyfi bir değer veya optimum değer bulunup belirlenebilir.)
2. Gözlemler tesadüfi şekilde en az bir ve en fazla bir kümede olmak koşuluyla dağıtılır.
3. Küme merkezleri hesaplanır.

4. Kümedeki her gözlemin küme merkezlerine olan uzaklıkları (küme içi değişmeler) hesaplanır.
5. Küme içi değişmeler toplanarak kare-hata elde edilir. Kare-hatanın amacı, kare-hatayı minimize eden k değerini bulmaktır.
6. Küme merkez değerleri ile gözlemler arası uzaklıklar hesaplanır. Gözlem hangi kümeye yakınsa o kümeye dahil edilir ve kümeler güncellenir.
7. Kümelerde herhangi bir değişiklik olmayana kadar 4., 5. ve 6. Adımlar tekrar edilir.

2.2. SOS algoritması

Çoğu doğa olaylarından esinlenerek geliştirilen meta-sezgisel algoritmalar arasında basit ve güçlü yapısı ile dikkat çeken SOS (ortak yaşayan organizmalar algoritması/symbiotic organism search) bir ekosistemde birlikte yaşayan simbiyotik organizmalar arasında gerçekleşen simbiyotik etkileşimleri taklit etmektedir. SOS'u diğer meta-sezgisellerden ayıran en önemli özellik algoritmanın performansı açısından öneme sahip olan algoritmik parametrelere ihtiyaç göstermemesidir. Gerek doğayı üstün taklit yeteneğinden gerekse de parametrik yapısının kullanıcı müdahalesine ihtiyaç bırakmamasından dolayı SOS algoritması performans açısından farklı problemlere karşı kararlılık gösterir [1, 17].

SOS popülasyon tabanlı bir algoritma olup, arama işlemine ekosistem adı verilen ve genellikle rastgele oluşturulan başlangıç popülasyonu ile başlar. Ekosistemdeki her bir çözüme organizma adı verilir ve organizmalardan her biri probleme belli derecedeki çözümü ifade eder. SOS algoritması doğada en yaygın biçimde görülen üç simbiyotik ilişkiden yararlanılarak geliştirilmiştir. Bu ilişkiler sırasıyla mutualizm, kommensalizm ve parazitizmdir. Her bir faz organizmanın hareketini ve başka bir organizmanın yerine geçip geçmeyeceğine karar verir. Mutualizm iki ayrı türün karşılıklı yarar sağladıkları simbiyotik ilişkiyi canlandırır. Kommensalizm iki organizmadan birinin yarar gördüğü diğerinin ise ne yarar ne de zarar gördüğü simbiyotik ilişkidir. Parazitizm ise iki organizmadan birinin yarar görürken diğerinin zarar gördüğü simbiyotik ilişkidir.

SOS algoritması aşağıdaki adımları tatbik ederek gerçekleştirilir:

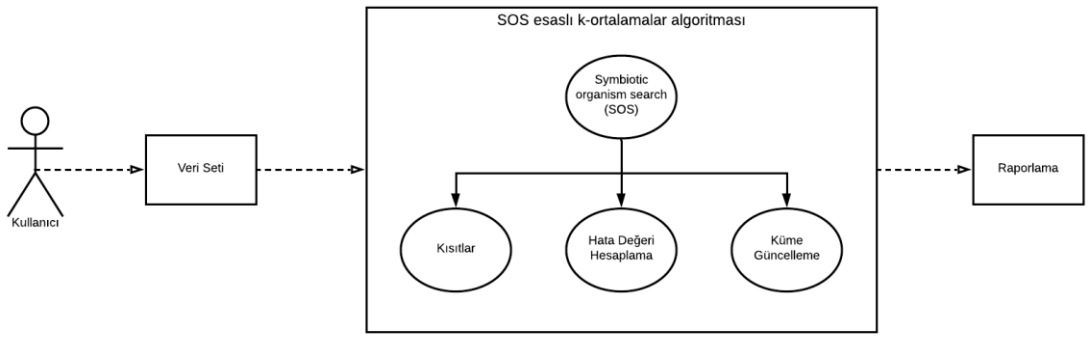
1. Ekosistem oluşturulur.
2. Ekosistemden en iyi sonucu veren organizma belirlenir.
3. Ekosistemden sırayla organizma seçilir ve sırayla üç evre uygulanır.
    i. Mutualizm Evresi
    ii. Kommensalizm Evresi
    iii. Parazitizm Evresi
4. Seçilen organizma ekosistemin son üyesi değilse Adım 2'ye dönülür. Aksi halde bir sonraki adıma geçilir.
5. Sonlandırma kriteri sağlanmışsa algoritma durdurulur. Aksi halde 2. Adıma dönülür ve sonraki iterasyona başlanılır.

2.3. Önerilen kümeleme yöntemi: SOS-tabanlı k-ortalamalar algoritması

SOS-tabanlı k-ortalamalar algoritması, sezgisel arama algoritması ve kümeleme algoritmasının melezleştirilmesi ile oluşmaktadır. İki algoritma melezleştirilirken kümeleme problemi, sezgisel arama algoritmasının öğeleri kullanılarak bir optimizasyon problemine dönüştürülmelidir. Optimizasyon probleminde tasarım parametreleri değerleri kullanılarak amaç fonksiyon üzerinden en uygun sonucu üretmek amaçlanır. Kümeleme problemi optimizasyon problemine çevrilirken sezgisel arama algoritması öğelerinin hangi kümeleme öğesine karşılık geldiği:

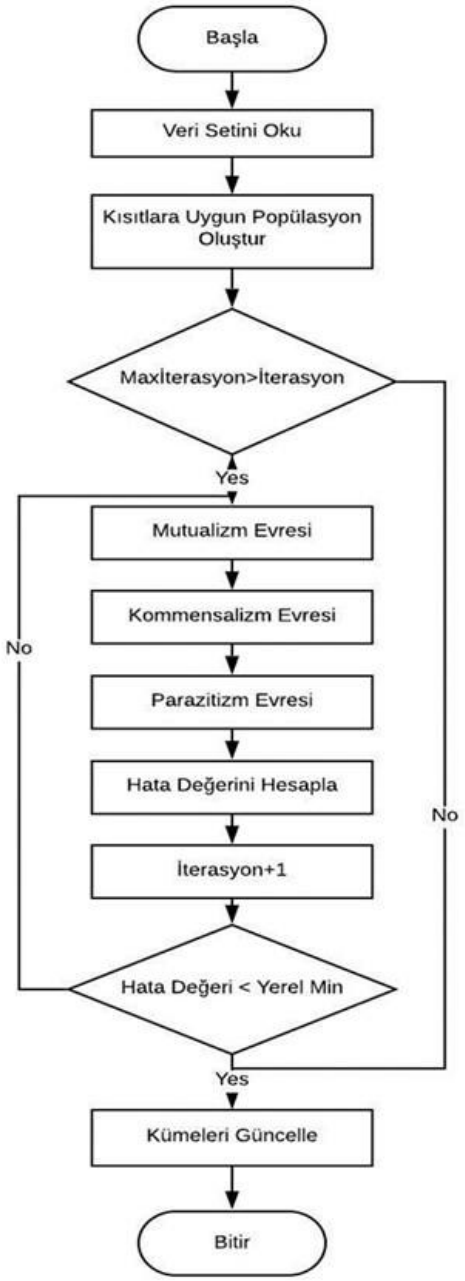
- Amaç fonksiyon → Kare hata
- Tasarım parametreleri → Gözlemler
- Kısıtlar → Bir gözlemin en az ve en çok bir kümede bulunması

şeklinde belirlenmiştir.



Şekil 1: Meta-sezgisel kümeleme algoritması temel ögeleri

Algoritmanın temel ögelerinden biri olan veri seti ayrık değerli olmalı ve k-ortalamalar algoritmasının veri setlerindeki gürültülere duyarlı olmasından dolayı veri setinin ön hazırlıktan geçirilmesi gerekir.



Şekil2: SOS-tabanlı k-ortalamalar algoritması akış diyagramı

SOS-TABANLI K-ORTALAMALAR ALGORİTMASI SÖZDE KODU

1. Veri seti okunur.
2. Küme sayısı belirlenir.
3. Amaç fonksiyon olarak gözlemlerin ait oldukları küme merkezleri ile aralarındaki uzaklıklardan yola çıkılarak küme içi değişimler sonra da bunların toplamı olarak kare hata hesabı yapılır.
4. Çözüm adayı tasarımı yapılır. Çözüm adayı kullanıcının tanımladığı sayıda kümeden oluşur. Bu kümeler birleştiğinde her bir gözlemden en az ve en çok bir adet bulunmalıdır. Rastgele yöntemle çözüm adayları topluluğu oluşturulur.

%Her bir birey probleme ait gözlemleri içermelidir.

5. Çözüm adaylarının uygunluk değerleri hesaplanır.

% Amaç fonksiyon toplam kare hata hesabı yaptığı için ve amaç bu hatayı minimize etmek olduğu için bu bir minimizasyon problemidir.

6. **While** (*maxiterasyonSayısı* > *iterasyonSayısı*)
7. **For** birey sayısı kadar
8. Hata değeri en iyi olan birey seçilir. *% Daha uygun bir değer oluşması halinde kıyaslamak için*

% Mutualizm evresi

9. $Birey_i \neq Birey_j$ olmak koşuluyla iki birey seçilir.
10. İki bireye aynı oranda komşu birey (*mutualVector*) oluşturulur. *% İki bireyin ortalamasından yeni bir birey elde edilir.*
11. İki adet yarar vektörü (*BF1, BF2*) oluşturulur. *%Faydalanma katsayıları rastgele olarak 1 veya 2 değerini alır.*
12. Komşu birey vektörü ve yarar vektörü kullanılarak bireyler mutasyona uğratılır.
13. Yeni oluşan bireyler uygunluk değeri daha iyiyse eskisi ile değiştirilir. Kümeler güncellenir.

% Kommensalizm Evresi

14. $Birey_i \neq Birey_j$ olmak koşuluyla iki birey seçilir.
15. $Birey_i$, $Birey_j$ kullanılarak mutasyona uğratılır.
16. Yeni oluşan bireyler uygunluk değeri daha iyiyse eskisi ile değiştirilir. Kümeler güncellenir.

% Parazitizm Evresi

17. $Birey_i \neq Birey_j$ olmak koşuluyla iki birey seçilir.
18. $Birey_i$ ile parazit vektör oluştur. Eğer $Birey_i$, $Birey_j$*'den* daha uygun değere sahipse $Birey_j$'yi *öldür.*
19. **End for**
20. *iterasyonSayısı*=*iterasyonSayısı*+1
21. **End while**

2.4. Deneysel çalışma

Bu bölümde, SOS-tabanlı k-ortalamalar algoritmasının test edilmesi ve doğrulanması için gerçekleştirilecek olan deneysel çalışmalar hakkında bilgi verilmektedir. Kümeleme algoritmalarının performanslarının ölçülmesinde sınıflandırma problemlerine ait veri setleri kullanılmaktadır. Bunun nedeni sınıflandırma problemlerine ait veri setlerinde nesnelerin ait oldukları sınıfların bilinmesidir. Bu sayede bir taraftan veri setindeki toplam sınıf sayısı bilinirken diğer taraftan hangi nesnenin (veri örneğinin) hangi sınıfa ait olduğu da bilinmektedir. Toplam sınıf sayısı bilgisi, kümeleme probleminde küme sayısına karşılık gelirken, nesnelerin ait oldukları sınıflar ise aynı kümede bulunması gereken nesnelerin bilinmesini sağlamaktadır. Bu iki bilgi, kümeleme algoritmasının performansını belirlemek için kullanılacaktır. Deneysel çalışmalarda kullanılacak olan sınıflandırma veri setleri ise uluslararası bir veri havuzu olan “uci machine learning datasets” üzerinden temin edilecektir [15]. Uluslararası araştırmacıların tezlerinden, araştırmalarından ve deneysel çalışmalarından elde ettikleri verileri paylaştıkları bu havuza erişim kısıtı bulunmamaktadır. Binlerce veri setinin bulunduğu bu veri havuzuna web ortamında https://archive.ics.uci.edu/ml/index.php internet adresinden erişim sağlanabilmektedir. Deneysel çalışmalarda k-ortalamalar yöntemi ile SOS-tabanlı k-ortamalar yöntemi arasında performans karşılaştırmaları yapılacaktır. Karşılaştırmalarda kullanılacak olan veri setleri bahsi geçen havuzdan temin edilecektir.

Deneysel çalışmalarda algoritmaların performanslarını ölçmek için izlenecek yol maddeler halinde aşağıda özetlenmektedir.

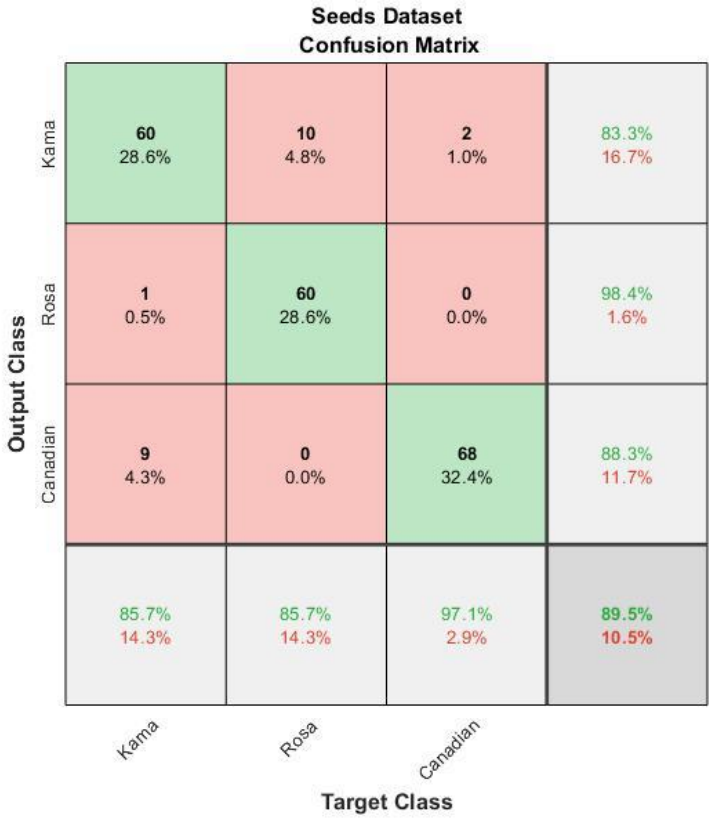
Veri seti seçiminde dikkat edilecek hususlar:

i) Az boyutludan çok boyutluya, farklı boyutlardaki problemlere ait olmak üzere altı (6) veri seti kullanılacaktır.
ii) Veri setlerinde bulunan veri örneklerinin sayılarının birbirlerinden farklı olmasına yani az, orta ve çok sayıda veri örneğini temsil edecek şekilde olmasına dikkat edilecektir.
iii) Veri setleri belirlenirken küme sayısının da “az”, “orta” ve “çok” şeklinde nitelendirilebilecek olmasına dikkat edilecektir.
iv) Altı adet veri setinden ikisi, çok boyutlu, çok sayıda veri örneği içeren ve küme sayısı fazla olacaktır. İki veri seti ise az boyutlu, çok sayıda veri örneği içeren ve küme sayısı da az olacaktır. Son olarak iki veri seti de az boyutlu, az sayıda veri örneği içeren ve küme sayısı da az olacaktır.

Buna göre deneysel çalışmada k-ortalamalar yöntemi ve SOS-tabanlı k-ortalamalar yöntemi arasındaki performans karşılaştırması yapılacaktır. Her iki algoritmanın farklı zorluk düzeylerindeki problemleri kümeleme doğrulukları (doğru sınıflandırılan veri örneklerinin yüzde olarak oranı), kümeleme hassasiyeti ve hata yüzdesi sabitken referans hata oranına ulaşmak için harcanan süre ölçülecektir. Bu üç değere bağlı olarak algoritmaların kümeleme performansları belirlenecektir. Böylelikle meta-sezgisel arama yönteminin (bir örneği olarak SOS algoritmasının) k-ortalamalar algoritmasının kümeleme performansına olan etkisi farklı şartlar altında incelenmiş olunacaktır. Hipotezin geçerlilik durumu araştırılacak ve deneysel çalışmalardan elde edilen verilerle ortaya koyulacaktır.

2.5. Test ve karşılaştırma

Deneysel çalışma sonuçlarının sunumunda karşılaştırma matrisleri (confusion matrix) kullanılacaktır. Sınıflandırma problemlerinin modellenmesinde ve modellerin performanslarının anlaşılabilir bir şekilde sunulmasında en yaygın kullanılan yöntem karşılaştırma matrisleridir [18]. Gerek karmaşayı önlemesi, gerek yüksek bir özetleme yeteneğine sahip olması, gerekse de kolay anlaşılabilir ve yorumlanabilir olması, karşılaştırma matrislerini sınıflandırma ve kümeleme sonuçlarının sunulmasında ideal bir seçenek haline getirmektedir.



Şekil 3: Karşılaştırma matrisi örneği [18]

Karşılaştırma matrisi, verideki var olan durum ile kümeleme modelinin doğru ve yanlış tahminlerinin sayısını gösterir. Örneğin Şekil 2'de tohumların özelliklerine göre kümelenmesi sonucunda oluşan karşılaştırma matrisi verilmiştir. Burada Kama, Rosa ve Canadian türü tohumlar incelenmiştir. Oluşan matrise göre 1 adet tohum Kama türü olması beklenirken Rosa türü olarak hesaplanmıştır ve aynı şekilde 9 adet tohum Kama olması beklenirken Canadian türü olarak hesaplanmıştır. Bu hatalar toplam veride %4,8'lik hataya sebep olurken Kama türü tohum açısından %14,3'lük hata oranına sebep olmuştur.
Karşılaştırma matrisi üzerinden çıkarılan ve literatürde sıklıkla kullanılan formülasyonlar bulunmaktadır.

Accuracy (Doğruluk Oranı): Sistemde doğru olarak yapılan tahminlerin tüm tahminlere oranıdır. (Örnekteki veride %89,5 oranında Accuracy değeri elde edilmiştir.)

Recall (Hassasiyet): Pozitif durumların ne kadar başarılı tahmin edildiğini gösterir. Kümeler içerisinde doğru olduğu bilinen gözlemlerin doğru olarak tahmin edilenlerinin bütün doğru olduğu bilinen gözlemlere oranıdır. (Örneğin Kama türü tohumlar için yapılan hesaplamanın sonucunda %85,7 oranında Recall değeri elde edilmiştir.)

Precision: Pozitif olarak tahmin edilen bir durumdaki başarıyı gösteren durum. Doğru olduğu bilinen gözlemlerin doğru olarak tahmin edilmişlerinin bütün doğru olarak tahmin edilmişlere oranıdır. (Örnek matriste Kama türü tohumlar için %83,3 oranında Precision değeri elde edilmiştir.)

F-Score: Recall ve Precision'ın harmonik ortalaması. (Örnekteki verilere göre bu kümelemenin F-score'u %84,5 oranındadır.)

3. PROJE YÖNETİMİ

3.1 İş- Zaman Çizelgesi

İŞ-ZAMAN ÇİZELGESİ

İP No	İş Paketlerinin Adı ve Hedefleri	Kim(ler) Tarafından Gerçekleştirileceği	Zaman Aralığı (..-.. Ay)	Başarı Ölçütü ve Projenin Başarısına Katkısı
1	Problem Tanımı ve Planlama	Ömer KÖROĞLU	01.09.2019-30.09.2019	Geliştirilecek sistemin fonksiyonel ve fonksiyonel olmayan gereksinimlerine kaynaklık teşkil edecek bilgilerin tanımlanması: k-ortalamalar yönteminin tatbik edilebildiği ve farklı boyutlardaki kümeleme problemlerinin çözümlenebileceği bir modelin fonksiyonel ve bu sistemin fonksiyonel olmayan gereksinimleriyle birlikte tanımlanması. Katkı: %20
2	Analiz ve Tasarım Çalışmaları	Ömer KÖROĞLU	02.10.2019-30.10.2019	Kullanım senaryolarının ve sınıf diyagramlarının oluşturulması: 4+1 UML diyagramlarının hazırlanması. Katkı: %15
4	Veri Setlerinin Elde Edilmesi	Ömer KÖROĞLU	03.02.2020-10.02.2020	UCI machine learning veri havuzundan 6 adet kümeleme problemine ait veri setlerinin elde edilmesi. Katkı: %5
5	Kodlama (Gerçekleştirim)	Ömer KÖROĞLU	22.02.2019-30.04.2020	Farklı boyutlardaki kümeleme problemlerinin SOS-tabanlı k-ortalamalar yöntemiyle kümelendiği uygulamanın gerçekleştirilmesi. Katkı: %25
6	Test, Doğrulama ve Bakım	Ömer KÖROĞLU	01.05.2019-30.06.2020	Uygulamanın web ortamından erişime açılması: Araştırmacıların problemlerine ait veri setlerini sisteme yükleyerek SOS-tabanlı k-ortalamalar yöntemiyle kümeleyebilmesi ve sonuçları k-ortalalamalar yöntemiyle kıyaslayabilmesi. Katkı: %35

3.2 Risk Yönetimi

RİSK YÖNETİMİ TABLOSU

İP No	En Önemli Riskler	Risk Yönetimi (B Planı)
1	Kısıtlı zamandan dolayı projeyi yetiştirememe	Öncelikle k-ortalamalar yöntemi geliştirilecek ve web ortamından kullanıma açılacaktır. Önerilen algoritma olmasa da araştırmacıların kullanımına açık ve problem boyutundan bağımsız çalışan bir kümeleme algoritması web ortamından paylaşıma açılmış olacaktır. Bunun yanında projeye bir ortak daha eklenerek kodlama aşamasında iş yükü azaltılarak projenin tüm fonksiyonlarıyla çalışması sağlanacaktır.
2	Önerilen algoritmanın beklendiği kadar başarılı olamaması	Bu durumda farklı meta-sezgisel arama algoritmaları ile k-ortalamalar yöntemi melezlenecektir. Ayrıca halihazırda klasik k-ortalamalar yönteminin tatbik edilebileceği web tabanlı ücretsiz bir uygulama da bulunmamaktadır. Projede geliştirilecek uygulamanın sadece bu fonksiyonu taşıması bile değerlidir.

3.3. Araştırma Olanakları

ARAŞTIRMA OLANAKLARI TABLOSU

Kuruluşta Bulunan Altyapı/Ekipman Türü, Modeli (Laboratuvar, Araç, Makine-Teçhizat, vb.)	**Projede Kullanım Amacı**
Bilgisayar laboratuvarları	Geliştirilecek olan algoritmanın test ve doğrulaması için kullanılacaktır.

4. YAYGIN ETKİ

ARAŞTIRMA ÖNERİSİNDEN BEKLENEN YAYGIN ETKİ TABLOSU

Yaygın Etki Türleri	Önerilen Araştırmadan Beklenen Çıktı, Sonuç ve Etkiler
Bilimsel/Akademik (Makale, Bildiri, Kitap Bölümü, Kitap)	Proje çerçevesinde 1 (bir) adet bildiri ve 1 adet makale çalışması yapılacaktır.
Ekonomik/Ticari/Sosyal (Ürün, Prototip, Patent, Faydalı Model, Üretim İzni, Çeşit Tescili, Spin-off/Start- up Şirket, Görsel/İşitsel Arşiv, Envanter/Veri Tabanı/Belgeleme Üretimi, Telife Konu Olan Eser, Medyada Yer Alma, Fuar, Proje Pazarı, Çalıştay, Eğitim vb. Bilimsel Etkinlik, Proje Sonuçlarını Kullanacak Kurum/Kuruluş, vb. diğer yaygın etkiler)	Araştırmacılar, proje çerçevesinde geliştirilecek olan ürüne web ortamından erişim sağlayabilecekler ve kümeleme problemlerini ücretsiz bir şekilde modelleyebilecekler/çözümleyebileceklerdir.
Araştırmacı Yetiştirilmesi ve Yeni Proje(ler) Oluşturma (Yüksek Lisans/Doktora Tezi, Ulusal/Uluslararası Yeni Proje)	Bu proje konusunu araştırırken çok sayıda uluslararası makaleye rastladım. Dolaysıyla proje konusu lisansüstü çalışmaların yapılabileceği niteliktedir. Lisans sonrası için lisansüstü eğitimime devam etme planım bulunmaktadır. Sezgisel kümeleme konusundaki bu projeyi başarıyla tamamlamam ve bu konuda danışmanımla birlikte makale hazırlamam durumunda

	lisansüstü eğitim ve yeni projeler konusunda da ciddi bir yol almam söz konusu olabilecektir. Halihazırda Sabancı Holding de yazılım biriminde iş yeri eğitimim devam etmektedir. Mezuniyet sonrası için Sabancı Holding'de yazılım mühendisi olarak işe başlamam ve yapay zekâ biriminde veri madenciliği uygulamalarında görev alabilmem söz konusudur. Bu birimde sanayi projeleri hazırlamam konusunda proje tecrübesinin önemli katkı sağlayacağını düşünmekteyim.

5. BELİRTMEK İSTEDİĞİNİZ DİĞER KONULAR

Proje çerçevesinde geliştirilecek olan uygulamanın test ve doğrulanmasında uluslararası bir veri havuzu paylaşım uygulaması olan "UCI Machine Learning Data Repository" kullanılacaktır. Bu uygulamaya ve 488 adet veri setine aşağıdaki bağlantıdan erişilebilmektedir:

https://archive.ics.uci.edu/ml/index.php

6. EKLER

EK-1: KAYNAKLAR

[1] MY. Cheng, D. Prayogo, "Symbiotic organisms search: A new metaheuristic optimization algorithm", Computers and Structures, 139, 98 – 112, 2014.

[2] Yeşilbudak, M., Kahraman, H., & Karacan, H. (2011). Veri madenciliğinde nesne yönelimli birleştirici hiyerarşik kümeleme modeli. Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 26(1).

[3] Banerjee, S., & Chattopadhyay, S. (2017). Power optimization of three dimensional turbo code using a novel modified symbiotic organism search (MSOS) algorithm. Wireless Personal Communications, 92(3), 941-968.

[4] Dosoglu, M. K., Guvenc, U., Duman, S., Sonmez, Y., & Kahraman, H. T. (2018). Symbiotic organisms search optimization algorithm for economic/emission dispatch problem in power systems. Neural Computing and Applications, 29(3), 721-737.

[5] Yilmaz, C., Kahraman, H. T., & Söyler, S. (2018). Passive mine detection and classification method based on hybrid model. IEEE Access, 6, 47870-47888.

[6] Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science, 54, 764-771.

[7] Jamal, A., Handayani, A., Septiandri, A. A., Ripmiatin, E., & Effendi, Y. (2018). Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction. LONTAR KOMPUTER: Jurnal Ilmiah Teknologi Informasi, 192-201.

[8] Capó, M., Pérez, A., & Lozano, J. A. (2017). An efficient approximation to the K-means clustering for massive data. Knowledge-Based Systems, 117, 56-69.

[9] Yu, S. S., Chu, S. W., Wang, C. M., Chan, Y. K., & Chang, T. C. (2018). Two improved k-means algorithms. Applied Soft Computing, 68, 747-755.

[10] Bai, L., Cheng, X., Liang, J., Shen, H., & Guo, Y. (2017). Fast density clustering strategies based on the k-means algorithm. Pattern Recognition, 71, 375-386.

[11] Ding, Y., Zhao, Y., Shen, X., Musuvathi, M., & Mytkowicz, T. (2015, June). Yinyang k-means: A drop-in replacement of the classic k-means with consistent speedup. In International Conference on Machine Learning (pp. 579-587).

[12] Ferrandez, S. M., Harbison, T., Weber, T., Sturges, R., & Rich, R. (2016). Optimization of a truck-drone in tandem delivery network using k-means and genetic algorithm. Journal of Industrial Engineering and Management (JIEM), 9(2), 374-388.

[13] Mustafi, D., & Sahoo, G. (2019). A hybrid approach using genetic algorithm and the differential evolution heuristic for enhanced initialization of the k-means algorithm with applications in text clustering. Soft Computing, 23(15), 6361-6378.

[14] Costa, L. R., Aloise, D., & Mladenović, N. (2017). Less is more: basic variable neighborhood search heuristic for balanced minimum sum-of-squares clustering. Information Sciences, 415, 247-253.

[15] https://archive.ics.uci.edu/ml/index.php (son erişim tarihi: 10 Ekim 2019)

[16] Lee, E., Schmidt, M., & Wright, J. (2017). Improved and simplified inapproximability for k-means. Information Processing Letters, 120, 40-43.

[17] Kahraman, H. T., Aras, S., Sönmez, Y., Güvenç, U., & Gedikli, E. Analysis, Test and Management of the Meta-Heuristic Searching Process: An Experimental Study on SOS. Politeknik Dergisi.

[18] Kahraman, H. T. (2016). A novel and powerful hybrid classifier method: Development and testing of heuristic k-nn algorithm with fuzzy distance metric. Data & Knowledge Engineering, 103, 44-59.