T.C.
YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ
ANABİLİM DALI

GENETİK ALGORİTMA VE K-EN YAKIN KOMŞU KULLANARAK METİN BELGELERİNİN SINIFLANDIRILMASI

YÜKSEK LİSANS TEZİ

HAZIRLAYAN : Parisa LARİBİ
DANIŞMAN : Doç. Dr. Rıdvan SARAÇOĞLU

VAN-2018

T.C.
YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ
ANABİLİM DALI

GENETİK ALGORİTMA VE K-EN YAKIN KOMŞU KULLANARAK METİN BELGELERİNİN SINIFLANDIRILMASI

YÜKSEK LİSANS TEZİ

HAZIRLAYAN: Parisa LARİBİ

VAN-2018

KABUL VE ONAY SAYFASI

Elektrik-Elektronik Mühendisliği Anabilim Dalı'nda Doç. Dr. . Rıdvan SARAÇOĞLU'nun danışmanlığında, Parisa LARİBİ tarafından sunulan "Genetik Algoritma ve K-En Yakın Komşu Kullanarak Metin Belgelerinin Sınıflandırılması" isimli bu çalışma Lisansüstü Eğitim–Öğretim Yönetmeliği'nin ilgili hükümleri gereğince 03/08/2018 tarihinde aşağıdaki jüri tarafından oy birliği ile başarılı bulunmuş ve Yüksek Lisans Tezi olarak kabul edilmiştir.

Başkan	: Doç. Dr. Halife KODAZ	İmza:
Üye	: Doç. Dr. Rıdvan SARAÇOĞLU	İmza:
Üye	: Dr. Öğr. Üyesi Özkan ATAN	İmza:

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ...16.../.08..../2018 tarih ve 2018/60-I..... sayılı kararı ile onaylanmıştır.

31/08/2018

Enstitü Müdürü
Doç. Dr. Harun AYDIN
Enst. Müdür Yrd.

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

(İmza)

Parisa LARİBİ

ÖZET

GENETİK ALGORİTMA VE K-EN YAKIN KOMŞU KULLANARAK METİN BELGELERİNİN SINIFLANDIRILMASI

LARİBİ, Parisa
Yüksek Lisans Tezi, Elektrik-Elektronik Mühendisliği Anabilim Dalı
Tez Danışmanı: Doç. Dr. Rıdvan SARAÇOĞLU
Temmuz 2018, 45 sayfa

Metin Madenciliği büyük miktardaki metinsel verilerden, önceden bilinmeyen bilgilerin elde edilmesini amaçlayan veri madenciliğinin bir dalıdır. Sınıflandırma, kümeleme ve tahmin, Metin Madenciliğinin önemli bir parçasıdır. Başarılı bir Metin Madenciliği yine başarılı bir sınıflandırma işlemine bağlıdır. Sınıflandırma sisteminin başarısını ve verimini artırmak için genellikle boyut azaltma işlemi gerçekleştirilir. Bu çalışmada metin belgelerinin sınıflandırılmasında boyut azaltma işlemi gerçekleştirilmiştir. Bunun için iki yöntem kullanılmıştır. Bunlardan ilki özellik çıkarımı, diğeri ise özellik seçimidir. Özellik çıkarımı için Temel Bileşen Analizi yöntemi kullanılmıştır. Özellik seçiminden sonra seçilen özellikleri için katsayı ile ağırlıklandırma kullanılmıştır. Özellik seçimi aşaması için ve özellik çıkarımından sonra en iyi kat sayıların seçimi için Genetik Algoritma kullanılmıştır. Deneysel sonuçlara göre özellik seçimi sınıflandırma başarısını kısmen azaltmıştır. Özellik çıkarımı ve bu aşamadan sonra eklenen katsayı ağırlıklandırma işlemi sınıflandırma başarısını önemli ölçüde artırmıştır.

Anahtar kelimler: Metin Sınıflandırma, Genetik Algoritma, KNN, Temel Bileşen Analizi.

ABSTRACT

CLASSIFICATION OF TEXT DOCUMENTS USING GENETIC ALGORITHM AND K- NEAREST NEIGHBORS

LARİBİ, Parisa
M. Sc., Electrical-Electronics Engineering
Supervisor: Assoc. Prof. Dr. Rıdvan SARAÇOĞLU
July 2018, 45 pages

Text Mining is a branch of data mining that aims to obtain previously unknown information from large quantities of textual data. Classification, clustering and estimation are some important piece of Text Mining. An important part of a successful Text Mining is the successful classification process. Dimension reduction is usually performed to improve the success and efficiency of the classification system. In this study, the dimension reduction process was performed in the classification of text documents. Two methods have been used for this. One of them is feature selection and the other is feature extraction. Principial Component Analysis method is used for feature extraction. Weighting with coefficients is used for selected features after feature selection. Genetic Algorithm is used for the feature selection phase and for the selection of the best coefficients after feature extraction. According to the experimental results, the feature selection partially reduced the classification success. Feature extraction and coefficient weighting added after this step significantly increased the classification success.

Key words: Text Classification, Genetic Algorithm, KNN, Principal Component Analysis.

ÖN SÖZ

Benim tezimin tüm aşamalarında hiç bir desteğini esirgemeyen danışmanım Sayın Doç. Dr. Rıdvan SARAÇOĞLU'na, yürekten şükranlarımı sunmak istiyorum. Ayrıca ebeveynlerime, özellikle ablama ve bir şekilde bana yardım eden tüm diğer arkadaşlarıma teşekkür etmek istiyorum.

2018

Parisa LARİBİ

İÇİNDEKİLER

Sayfa

ÖZET i
ABSTRACT iii
ÖN SÖZ v
İÇİNDEKİLER vii
ÇİZELGELER LİSTESİ ix
ŞEKİLLER LİSTESİ xi
SİMGELER VE KISALTMALAR xiii
1. GİRİŞ 1
1.1. Amaç ve Tezin Önemi 3
2. KAYNAK BİLDİRİŞLERİ 5
3. MATERYAL VE YÖNTEM 14
3.1. Metin Önişleme 14
3.2. K-En Yakın Komşu Sınıflandırma Yöntemi 14
3.2.1. K-NN Parametreleri 16
3.3. Genetik Algoritma 17
3.3.1. Genetik Algoritma Operatörleri 18
3.4. Temel Bileşenler Analizi 24
4. BULGULAR 26
4.1. Veri Kümesi 26
4.2. Classic 3 Veri Kümesi için Sınıflandırma Sonuçları 26
4.2.1. Classic 3 veri seti için özellik seçimi 27
4.2.2 Classic 3 veri seti için özellik çıkarımı 29
4.3. Reuters Veri Kümesi İçin Sınıflandırma Sonuçları 31
4.3.1. Reuters veri seti için özellik seçimi 31
4.3.2. Reuters veri seti için özellik çıkarımı 34
5. TARTIŞMA VE SONUÇ 39
KAYNAKLAR 43
ÖZGEÇMİŞ 46

ÇİZELGELER LİSTESİ

Çizelge **Sayfa**

Çizelge 4.1. Classic 3 veri seti ve GA1 programının parametreleri.....27

Çizelge 4.2. Classic 3 veri seti ve GA2 programının parametreleri.....28

Çizelge 4.3. Classic 3 veri seti için GATBA parametreleri.....29

Çizelge 4.4. Reuters veri seti ve GA1 programının parametreleri.....31

Çizelge 4.5. Reuters veri seti ve GA2 programının parametreleri.....32

Çizelge 4.6. Reuters veri seti için kullanılan GATBA parametreleri.....33

Çizelge 4.7. Classi 3 veri seti sonuçlar.....34

Çizelge 4.8. Reuters veri seti sonuçları.....34

Çizelge 4.9. Classic 3 Veri Seti için *k* parametresinin sınıflandırma başarıları.....35

Çizelge 4.10. Reuters Veri Seti için *k* parametresinin sınıflandırma başarıları…….....36

ŞEKİLLER LİSTESİ

Şekil **Sayfa**

Şekil 3.1. Örnek rulet çarkı 20

Şekil 4.1. Classic veri seti GA1 programı sonuçları.....27

Şekil 4.2. Classic 3 veri seti ve GA2 programı sonuçları.....28

Şekil 4.3. Classic 3 veri kümesi için GATBA sonuçları.....30

Şekil 4.4. Reuters veri seti ve GA1 programı sonuçları.....31

Şekil 4.5. Reuters veri seti ve GA2 programı sonuçları.....32

Şekil 4.6. Reuters veri kümesi için GATBA Sonuçları.....34

Şekil 4.7. Classic 3 Veri Kümesi için *K* Sonuçları.....35

Şekil 4.8. Reuters Veri Kümesi için *K* Sonuçları.....36

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılan bazı kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
GA	Genetik Algoritma
MATLAB	Matrix Laboratory
TBA	Temel Bileşen Analizi
ÇİAİ	Çevrim İçi Analitik İşleme
KNN	K Nearest Neighbors
BGA	Biyolojik Temelli GA
PSO	Parçacık Sürüsü Optimizasyonu
GA1	MATLAB ile Hazırlanan Genetik Algoritma
GA2	MATLAB Toolbox ile Hazırlanan Genetik Algoritma
GATBA	Genetik Algoritma Temel Bileşen Analizi
TAŞ	Terim Ağırlıklandırma Şeması

1. GİRİŞ

Gelişen teknolojiyle birlikte günümüzde yaptığımız bütün işlemler dijital ortamda kayıt altına alınmaktadır. Devlet kurumlarından özel sektöre kadar birçok alanda yaptığımız işlemlerin sayısal ortamda kayıtları tutulmaktadır. Tüm bu kayıt altında olan veriler, veri tabanlarında günyüzüne çıkarılmayı bekleyen çok değerli bilgiler içermektedir. Veri madenciliği önceden bilinmeyen geçerli ve uygulanabilir bilgilerin geniş veri tabanlarından elde edilmesi ve bu bilgilerin işletme kararları verilirken kullanılmasıdır. Veri madenciliği 1990'lardan bu yana veri depolama araçları, barkod ve RFID teknolojilerine bağlı olarak gelişip yaygınlaşmaktadır (Silahtaroğlu, 2013). Veri madenciliğinde önemli olan konulardan biri elde edilecek bilgilerin önceden bilinmiyor olması veya sonucunun tahmin edilebilir olmamasıdır.

Veri madenciliği uygulama alanları:

- pazarlama ve risk yöntemi
- işaret işleme
- biyoloji (DNA sıra analizi)
- tıp

Veri ambarları ve Çevrim İçi Analitik İşleme (ÇİAİ): Veri madenciliği çalışmalarında veri tabanı gerekmektedir. Ancak işlemsel veri tabanından (transactional database) alınan veriler direk uygulamalarda kullanılamamaktadır. Bu verilerin kullanabilir olması için yapılan çalışmaya göre konu odaklı olarak düzenlenmiş, birleştirilmiş ve sabitlenmiş olması gerekmektedir. Bu işlemlere ait veri tabanlarına veri ambarları denilir. Veri ambarları üzerinde, farklı taktikler ve stratejik konularda karar vermeye yardım eden veri analizi ve sorgulama işlemlerine ÇİAİ denir. Kısacası bilgisayar üzerinde akıl yürütecek işlem yapma olarak tanımlanabilir (Şeker, 2014).

Veri madenciliği çalışmalarında ilk başta yapılması gereken verilerin hazırlanmasıdır. Bu işlem verilerin temizlenmesi ve verilerin yeniden yapılandırılması olarak ikiye ayrılabileceği gibi, yapılacak işlemler beş ayrı başlıkta toplanabilir:

1. Kayıp verilere yapılacak işlemler (veri temizlenmesi)
2. Verilerdeki gürültülünün ortadan kaldırılması (veri temizlenmesi)
3. Verilerin bütünleştirilmesi (veri yapılandırılması)
4. Verilerin dönüştürülmesi (veri yapılandırılması)
5. Verilerin azaltılması (veri yapılandırılması)

Veri madenciliği modelleri değer tahmin modeli (predictive modeling), veri tabanı bölümleme modeli (database segmentation modeling), bağlantı analizi (link analysis) ve fark sapmaları (deviation detection) olarak dört ayrı ana başlık altında toplanmıştır (Cabena ve ark., 1998).

Metin madenciliği en kısa tabiriyle metinleri veri kaynağı olarak kabul eden veri madenciliğidir. Metin madenciliğinin çalışma alanları; Bilgi erişimi (information retrieval), doğal dil işleme (natural language processing), adlandırılmış varlık tanıma (named entity recognition), örüntüsü tanımlı varlıkların bulunması (pattern identified entities), eş atıf (coreference), varlik iliski-kural-olay çıkarımı (entity relationship modelling), duygu analizi (sentimental analysis) ve benzeri konulardır. Metin madenciliğinde bir metin veri tabanından alınan verilerden ilk olarak bir özellik çıkarımına tabi tutulur. Daha sonra sınıflandırma, kümeleme ve tahmin uygulanarak neticede yapılandırılmış veri elde edilir.

Metin madenciliğinin ilk aşaması, bir kişinin hesaplamalar yapabilmesi ve istatistiksel yöntemlerin kullanabilmesi için belge setindeki tüm belgeler için öznitelik çıkarımıdır. Metin belgeleri, kelimeler (öznitelikler) ve aralarındaki illişki ile nitelendirilir. Bu konudaki en önemli yaklaşım, vektör uzayı modelidir. Bu modelde her kelime sayısal ağırlıklı ayrı bir değişken olarak ele alınabilir. En popüler ağırlıklandırma yöntemleri terim sıklığı, normalize edilmiş şekilde, ters belge sıklığıdır. Özellik seçimi ise belgeyi göstermek üzere özelliklerin bir alt kümesinin seçilmesidir (Abizi, 2015).

Sınıflandırma, veri madenciliğinde en yaygın kullanılan tekniklerdendir. Sınıflandırma esas olarak bir tahminleyici modeldir. Verilerin içeriklerindeki ortak özelliklere göre ayrıştırılması sınıflandırma olarak adlandırılır. Sınıflandırma bir öğrenme algoritmasına dayalıdır. Bir çırpıda tüm veriler kullanılarak eğitme işi yapılamaz. Bu işlem verilerin tümünden bir örnek alarak gerçekleşmektedir. Amaç bir

sınıflandırma modelinin yaratılmasıdır. Başka bir tabirle hangi sınıfa ait olduğu bilinmeyen verilerin sınıf belirleme sürecidir (Abizi, 2015).

1.1. Amaç ve Tezin Önemi

Bu çalışmada veri madenciliğinin alt birimi olan metin madenciliği üzerinde durulmuştur. Metin madenciliğinin önemli bir konusu olan sınıflandırma bu çalışmanın odak noktasıdır. Metinsel belge sınıflandırması ile ilgili bir çerçeve oluşturulması ve mevcut sınıflandırma başarısının artırılması temel amaçtır. Bu doğrultuda seçilen sınıflandırıcı k-en Yakın Komşu (KNN) yöntemidir. Yöntemin başarısının artırılması için belge matrisi bir katsayı vektörü ile geliştirilmiştir. Bu katsayıların belirlenmesi için ise Genetik Algoritma (GA) yönteminden faydalanılmıştır. Sınıflandırma başarısının artırılması sınıflandırmanın kullanıldığı birçok çalışmada ve alanda önem taşımaktadır. Çünkü metin madenciliğinin birçok aşamasında sınıflandırma kullanılmaktadır. Daha başarılı bir sınıflandırıcı, daha başarılı metin madenciliği uygulamaları anlamına gelmektedir.

Çalışmada kullanılan veri kümeleri öncelikle önişleme aşamasından geçirilmiştir. Bu aşamada belgeler oluşturuldukları kelimelerin bir listesi olarak ele alınır. Bu kelimelerden anlam için önemsiz olan ve durma kelimeleri olarak adlandırılan kelimeler ayıklanır. Daha sonra kelimler gövdelerine indirgenir. Bu sayede belgelerin matematiksel bir modeli oluşturulmuş olur. Bu aşamada vektör uzay modeli kullanılmıştır. Sınıflandırma için seçilen yöntem KNN'dir. Bu algoritmada sınıflandırma yapılırken veriler içerisinden, aradığımız veriye en yakın mesafede bulunan verilerin uzaklığı hesaplanmaktadır. Bu yöntemin başarısını artırmak için mevcut veriler belirlenen bir katsayı ile çarpılır.

KNN algoritmasını kullanma sebebi, diğer alternatif yöntemlere göre daha hızlı ve kabul edilebilir bir başarı oranına sahip olmasıdır. Bu çalışmada katsayı değeri sıfır ile bir aralığından bir değer olarak belirlenmiştir.

.

2. KAYNAK BİLDİRİŞLERİ

Ağırlıklandırma metotlarının ve mesafe ölçümlerinin, KNN tabanlı sınıflandırma üzerinde önemli bir etkisi vardır. Niteliklerin en uygun ağırlık değerlerinin nasıl bulunacağı ve KNN sınıflandırma doğruluğunu etkileyen komşular arasındaki mesafelerin nasıl ölçüleceği önemli problemlerdir. Kahraman (2016), çalışmasında, bulanık mesafe metriği adı verilen güçlü bir benzerlik ölçüm yöntemini, test ve eğitim gözlemleri arasındaki mesafeleri ölçmek için açıklayıp genişletmiştir. Bulanık ölçüme dayalı olan benzerlik dizileri, klasik ve diğer ağırlıklı uzaklık ölçümlerinden daha verimli sonuçlar vermektedir. Son olarak, ağırlıklandırma yöntemleri bulanık metrik temelli benzerlik ölçümü ve KNN algoritması ile birleştirilerek önerilen algoritmanın sınıflandırma doğruluğu arttırılmıştır.

Metin kategorizasyonu, dijital biçimdeki belgelerin düzenlenmesinde yaygın bir şekilde kullanılır. Dijital formdaki belgelerin sayısının artmasından dolayı, otomatik metin sınıflandırması son on yıla bakılırsa gelecek vadetmektedir. Metin kategorizasyonunun en büyük sorunu, özellik sayısının çok fazla olmasıdır. Bunların çoğu sınıflandırıcıyı yanıltabilecek alakasız gürültülerdir. Bu nedenle, özellik alanının boyutunu azaltmak ve performansı artırmak için metin sınıflandırmasında sıklıkla, özellik seçimi kullanılır. Uğuz (2011), çalışmasında, metin kategorizasyonunun performansını artırmak için iki aşamalı özellik seçimi ve öznitelik çıkarımı kullanılmıştır. İlk aşamada, dokümandaki her terim, bilgi kazanımı yöntemini kullanarak sınıflandırma için önemlerine göre sıralanır. İkinci aşamada, GA ve Temel Bileşen Analizi (PCA) özellik seçimi ve öznitelik çıkarma yöntemleri, önemi azalan sırada sıralanan terimlere ayrı ayrı uygulanmakta ve bir boyut azaltılması gerçekleştirilmektedir.

Ghareb ve ark. (2015), çalışmalarında GA temelli karma özellik seçimi yaklaşımları önermektedir. Bu yaklaşım, filtre alan seçimi yöntemlerinin avantajlarını, yüksek boyutlu özellik değerlerini ele almak ve aynı anda kategorizasyon performansını geliştirmek için geliştirilmiş GA ile bir karma arama tekniğini kullanır. Mutasyon orijinal ebeveynlerin sınıflandırıcı performansına ve özellik önemine dayalı olarak

gerçekleştirilirken, Çaprazlama işlemi kromozom girdilerinin (özellikler) terim ve belge sıklığıyla, kromozom (özellik alt kümesi) bölümlemesi temel alınarak gerçekleştirilmiştir. Böylece, çaprazlama ve mutasyon işlemleri, olasılık ve rasgele seçim yerine, yararlı bilgilere dayanarak gerçekleştirilmiştir. Melez özellik seçimi yaklaşımları oluşturmak için epigenetik algoritma ile altı tanınmış filtre özelliği seçme yöntemi birleştirilmiştir.

KNN, metin sınıflandırması için çok popüler bir algoritmadır. Trstenjak ve ark. (2014), çalışmalarında, TF-IDF metodu ile KNN algoritmasının kullanılmasını ve metin sınıflandırması için bir çerçeve oluşturulmasını sunmaktadırlar. Çerçeve, çeşitli parametrelere göre sınıflamayı, sonuçların ölçülmesini ve analiz edilmesini sağlar. Çerçevenin değerlendirilmesi aşamasında sınıflamanın hızı ve kalitesi üzerine odaklanılmıştır. Test sonuçları benzer çerçevelerin daha da geliştirilmesi için rehberlik sağlayan algoritmanın iyi ve kötü özelliklerini göstermektedir.

Metinsel veriler katlanarak artmaktadır. Verilerin önceden tanımlanmış sınıflardan birine otomatik olarak atanması işlemi sınıflandırma olarak adlandırılır. Buna duyulan ihtiyaç giderek artmaktadır. Birçok pratik uygulamada, eğitim verilerinin tüm sınıflar arasında eşit olarak dağıldığı varsayılır, ancak dengesiz veri dağılımı bir problemdir. Kang ve ark. (2012), Ayrıştırılmış K-En Yakın Komşusu'nu (DCM-KNN) önermektedirler. Eğitim adımında, DCM-KNN, eğitim verilerinin, geleneksel KNN'nin sonucuna dayanan yanlış sınıflandırılmış ve doğru sınıflandırılmış veri kümesine ayrılır ve her küme için uygun KNN bulunur. Test aşamasında DCM-KNN, test verisinin yanlış sınıflandırılmış ve doğru sınıflandırılmış veri kümesine benzeyip benzemediğini tahmin eder ve uygun KNN'leri uygular. Deneysel sonuçlar, önerilen algoritmanın dengesiz bir durumda daha doğru sonuçlar elde edebildiğini göstermektedirler.

Onan ve ark. (2016), çalışmalarında beş istatistiksel anahtar kelime özütleme yönteminin (en sık ölçüye dayanan anahtar kelime çıkarma, terim sıklığı-ters cümle sıklığına dayalı anahtar kelime çıkarma, ortak oluşma istatistiği tabanlı anahtar kelime çıkarma, sıradışılığa dayalı anahtar kelime çıkarımı ve TextRank algoritması), tahminsel başarımını incelemişlerdir. Çalışmalarında, yaygın olarak kullanılan beş topluluk yöntemiyle (AdaBoost, Bagging, Dagging, Rastgele Altuzay ve Çoğunluk Oylayışı) temel öğrenme algoritmalarını (Naïve Bayes, destek vektör makineleri, lojistik

regresyon ve Random Forest) karşılaştırmak için kapsamlı bir çalışma yapılmıştır. Sınıflandırma şemaları, sınıflandırma doğruluğu, F-ölçümü ve eğri altındaki alanlar açısından karşılaştırılmıştır. ACM belge kolleksiyonu için, en yüksek ortalama öngörme başarısı (% 93.80), en sık rastlanan anahtar kelime çıkarma yönteminin Random Forest algoritmasının Bagging grubu kullanılarak elde edilmiştir.

Belli bir eğitim veri setinden Gürültülü verileri filtreleyerek veri boyutunu azaltmak için kullanılan çözüm yöntemlerinden birisi Örnek seçimidir. Tsai ve ark. (2014), çalışmalarında bu iş için biyolojik temelli GA (BGA) olarak adlandırılan yeni bir algoritma tanıtmışlardır. BGA biyolojik bir süreçtir. Bu süreç mantıklı kurallarla uyuşan en sade işlemleri içerir. Bu sayede algoritmanın doğal gelişim sürecine çok yakın benzetim yapılabilir bu ise algoritmanın daha verimli ve etkili bir sonuç vermesini sağlar. En gelişmiş beş algoritma üzerinden BGA'nın performansını gösteren deneysel sonuçlarda TechTC-100 ve Reuters-21578 veri setleri kullanılmıştır. BGA, k-NN ve destekçi vektör makinesi sınıflandırıcılarına benzer veya biraz daha iyi bir sınıflandırma sonucu sağlamıştır.

Metin sınıflandırma bağlamında terim ağırlıklandırma şemaları, öğrenmeye yeni bir yaklaşım sağlamaktadır. Metin madenciliğinde bir Terim Ağırlıklandırma Şeması (TAŞ), bir sınıflandırıcı uygulamadan önce belgelerin vektör uzayı modelinde temsil edilme şeklini belirler. Kabul edilebilir performans standart TAŞ ile elde edilebilmiştir (örneğin, Boolean ve term-frekans şemaları). TAŞ'lerin tanımı geleneksel bir yaklaşımdır. Dahası, belirli bir problem için en iyi TAŞ'nin ne olduğunu belirlemek hala zor bir iştir ve hali hazırda mevcut olanlardan daha iyi şemaların üretilip üretilemeyeceği net değildir. Escalante ve ark. (2015), çalışmalarında, metin sınıflandırma bağlamında öğrenme-terim ağırlıklandırma düzenleri (TAŞ) için yeni bir yaklaşım açıklamışlardır. Metin madenciliğinde bir TAŞ, bir sınıflandırıcı uygulamadan önce belgelerin bir vektör Uzay modelinde temsil edilme şeklini belirler. Halbuki standart TAŞ'ler ile kabul edilebilir bir performans elde edilmiştir (ör. Boolean ve terim sıklığı şemaları). Ayrıca, belirli bir problem için en iyi TAŞ'nin hangisi olduğunu belirlemek hala zor bir iştir ve henüz net değildir, mevcut şemalardan daha iyi şemalar olup olmadığı araştırılabilir ve bilinen TAŞ ile birleştirilerek yeni şemalar oluşturulabilir. Bu çalışmada etkili TAŞ'leri öğrenmeyi amaçlayan bir genetik program

önerilmiştir. Metin sınıflandırmasında mevcut planların performansını arttırmaktadır. Kapsamlı deneysel sonuçlar, tematik ve tematik dışı metin sınıflandırmasından ve görüntü sınıflandırmasından elde edilen veri setlerini kapsar.

Uysal ve Günal (2014), çalışmalarında, metin sınıflandırmasında belgelerin daha iyi gösterimini sağlamak için GA'ya dayalı gizli anlamsal özellikler önermişlerdir. Önerilen yaklaşım, özellik seçimi ve özellik dönüştürme aşamalarını içermektedir. Birinci aşama, son teknoloji ürünü filtre tabanlı yöntemler kullanılarak gerçekleştirilir. İkinci aşamada, standart Gizli Anlamsal İndeksleme (GAİ) yaklaşımının aksine, en büyük tekil değerlere karşılık gelen tekil vektörlerle sınırlı olmayan uygun tekil vektörler kullanılarak daha iyi bir izdüşüm elde edilebilmesi için, GA tarafından güçlendirilen GAİ kullanılmaktadır. Bu şekilde, küçük tekil değerlere sahip tekil vektörler projeksiyon için de kullanılabilirken, daha büyük ayrışım elde etmek için büyük tekil değerlere sahip vektörler de elenebilir. Deneysel sonuçlar, GA'ya dayalı gizli anlamsal özelliklerin çeşitli özellik boyutları için karşılaştırmalı veri kümelerindeki GAİ ve filtre temelli özellik seçme yöntemlerinden daha iyi performans sergilediğini göstermektedir.

Abualigah ve ark. (2017), metin belge kümeleme problemi için, özellik ağırlık şeması ve dinamik boyut azaltımı ile üç özellik seçimi algoritması önermişlerdir. Metin belge kümeleme, metin madenciliğinde yeni bir akımdır. Bu süreçte, metin dokümanları, terim frekansına bağlı olan uygun değerlendirme fonksiyonunu kullanarak, dikkatle seçilmiş bilgilendirici özelliklere göre birkaç tutarlı kümeye ayrılmıştır. Her dokümandaki bilgilendirici özellikler, özellik seçim yöntemleri kullanılarak seçilir. GA, harmoni arama algoritması ve parçacık sürüsü optimizasyonu (PSO) algoritması, en başarılı özellik seçimi yöntemleridir. Terim frekansına ve uzunluğa bağlı ve uzunluk özelliği ağırlığı olarak adlandırılan yeni bir ağırlıklandırma şeması da bu yöntemlerle kullanılmıştır.

Kümelemede kullanılan özelliklerin sayısını azaltmak ve algoritmaların performansını artırmak için yeni bir dinamik boyut azaltma yöntemi de sağlanmaktadır. Popülasyonel bir kümeleme yöntemi olan k-ortalaması, dinamik indirgeme ile elde edilen terimlere (veya özelliklere) dayalı metin belgelerinin kümelenmesinde kullanılır.

Silva ve ark. (2017), çalışmalarında en az tanımlama uzunluğu prensibine dayanan, etkin, basit, ölçeklenebilir ve hızlı bir multinominal metin sınıflandırıcısı olan MDLText'i önermişlerdir. MDLText, gerçek dünyadaki uygulamalarda, büyük ölçekli problemlerde ve çevrimiçi senaryolarda arzu edilen özellikler olan aşırı eklemeyi önlemek için yeterince sağlam olmanın yanı sıra hızlı artımlı öğrenme sergilemiştir. Denemeler, önerilen yaklaşımın öngörü gücü ve hesaplama verimliliği arasında iyi bir denge kurduğunu göstermiştir.

Lu ve ark. (2015), çalışmalarında PSO aracılığıyla metin özelliği seçiminin etkisini arttırmaya çalışmışlardır. Geliştirilmiş PSO'nun güncel başarılarını ve klasik özellik seçim yöntemlerinin karakteristik özelliklerini analiz ederek, bu çalışmada birçok araştırma yapmışlardır. Önce, ortak PSO modeli, özellik seçme yöntemlerini optimize etmek için sırasıyla fonksiyonel atalet ağırlıkları ve sabit sıkıştırma faktörü temel alan iki geliştirilmiş PSO modeli seçilmiştir. Daha sonra sabit sıkışma faktörüne göre yeni bir fonksiyonel daralma faktörü oluşturulmuş ve geleneksel PSO modeline eklenmiştir. Son olarak, fonksiyonel daralma faktörüne ve fonksiyonel atalet ağırlıklı iki gelişmiş PSO modeli önerilmiştir. Sırasıyla eşzamanlı olarak geliştirilmiş PSO modeli ve eş zamansız olarak geliştirilmiş PSO modeli olmuştur.

Song ve ark. (2009), çalışmalarında ontoloji yöntemine dayalı metin kümelemesi için kendi kendini organize eden bir GA önermişlerdir. Metin kümelemesi alanındaki ortak problem, belgenin bir kelime grubu olarak gösterilmesi ve kavramsal benzerliğin göz ardı edilmesidir. Bu sorunun üstesinden gelmek için sözlük tabanlı ve eş anlamlılar tabanlı ontolojiden yararlanmışlardır. Bununla birlikte, geleneksel sözlük (korpus) tabanlı yöntemin üstesinden gelinmesi oldukça zordur. Bu çalışmada, ilişkili semantik benzerliği uygun bir şekilde yakalayabilen, dönüştürülmüş bir GAİ modeli önerilmiş ve korpus tabanlı ontoloji olarak gösterilmiştir. Metin kümelemesinde ontoloji yöntemlerinin nasıl etkili bir şekilde kullanılabileceğini araştırmak için, çeşitli benzerlik önlemleri kullanan iki hibrid strateji uygulanmaktadır. Deneysel sonuçlar, ontoloji stratejisi ile bağlantılı olarak kullanılan GA yöntemlerinin, dönüştürülmüş GAİ tabanlı ölçüm ve eş anlamlılar tabanlı ölçümün kombinasyonundan daha iyi olduğunu göstermektedir. Kümeleme algoritmaları aynı benzerlik ortamlarında standart GA ve k-ortalamalarla karşılaştırıldığında performansı verimli bir şekilde arttırmıştır.

Song ve Park (2009), çalışmalarında metin kümelemesi için gizli bir semantik model (GAL) temelli bir GA yöntemi geliştirmişlerdir. Belge kümelemesi için GA'arın uygulanmasında karşılaşılan en büyük güçlük, metinsel veriler için tipik olan özellik alanındaki binlerce veya on binlerce boyuttur. Çünkü, en açık ve en popüler yaklaşım vektör uzayı modelidir. Bu modelde sözcük grubundaki benzersiz her bir terim tek bir boyutu temsil eder. GAİ, bir sorgunun veya bir belgenin ima ettiği gizli anlamı, boyut azaltılmış bir alanda temsil ederek keşfetmeye çalışan, bilgi alımında başarılı bir teknolojidir. Bu arada, GAİ, metinsel verilerde anlamsal bir yapı oluşturan eşanlamlılığın ve çok anlamlılığın etkilerini dikkate alır. GA, azaltılmış alanda en iyi çözümü etkili bir şekilde geliştirebilen arama tekniklerine sahiptir. Doğru küme sayısını otomatik olarak geliştirecek ve optimum veri seti kümeleşmesini sağlayan, değişken dize uzunluklu bir GA önermişlerdir.

Das ve ark. (2017), çalışmalarında veri madenciliğinde özellik seçme problemini bi-objektif GA'ya dayalı özellik seçimi yöntemi önermekle ele almışlardır. Veri kümesi, değiştirme stratejisi ile örneklendirilir ve yöntem, her örneklenmiş veri kümesinden hakim olmayan özellik alt kümelerini belirlemek için uygulanır. Son olarak, bu tür bi-nesnel GA'ya dayalı öznitelik seçicilerinin topluluğu, çok genelleştirilmiş özellik alt kümesi üretmek için paralel uygulamalar yardımıyla geliştirilmiştir. Bu çalışmada, özellikle özellik seçimi veri setleri için UCI makine öğrenme havuzu veri setleri için doğrulanmış ve deneysel sonuçlar, önerilen topluluk özelliği seçim yönteminin etkililiğini göstermek için ilgili teknik özellik seçim yöntemleri ile karşılaştırılmıştır.

Dağıtılmış bilgi erişimi ortamında bilgi kaynaklarının seçimi kritik bir konudur. Bu bağlamda, dağıtılmış bir bilgi alım sisteminin çok sayıda kaynaktan oluştuğu bilinmektedir. Verimli bir bilgi erişimi sağlamak, bir sorgu için yalnızca ilgili bilgileri içerecek olası kaynakları aramaktır. Lebib ve ark. (2017), çalışmalarında, bu problem için GA kullanmışlardır. Önerdikleri GA, potansiyel çözümlerin geniş alanındaki en iyi seçimi bulmaya çalışmaktadır. Bu seçim, bir dizi kaynaktan oluşan bir kombinasyon olarak temsil edilir. Seçim doğruluğunun geliştirilmesi, kullanıcıların kaynakları kullanımlarının izlenmesine, kaynakların etiketlenerek tarihsel etiketleme ile zenginleştirilmesine bağlıdır.

Büyük miktardaki metin kolleksiyonunda metin sınıflaması, sınıflandırma algoritmalarının uygulanması için zaman alıcıdır. Bu nedenle, metin sınıflandırma problemlerinin boyutunu azaltmak önemlidir. Gasanova ve ark. (2014), çalışmalarında, kooperatif evrimsel GA kullanarak, terimlerin hiyerarşik aglomeratif kümeleme ve küme ağırlığı optimizasyonuna dayanan boyut azaltma için bir yöntem önermektedirler. Yöntem, farklı metin önişlemlerine sahip çeşitli sınıflandırma yöntemleri kullanılarak 5 farklı veri seti üzerinde uygulanmıştır. Yöntem, metin sınıflaması probleminin boyutunu önemli ölçüde azaltır. Sınıflandırma etkinliği, kümelenme ağırlıklarının optimizasyonu ile kümelenmeden sonra önemli ölçüde artar veya azalır.

Özellik seçimi, özellik alanının yüksek boyutlamasına iyi bir çözüm olarak bilinir ve metin sınıflandırması için çoğunlukla tercih edilen özellik seçme yöntemleri, filtre tabanlı olanlardır. Ortak bir filtre temelli özellik seçim şemasında, ayırt edici güçlerine bağlı olarak özelliklere benzersiz puanlar atanır ve bu özellikler skorlara göre azalan sırada sıralanır. Daha sonra, son adım, N'nin genel deneysel olarak belirlenmiş bir sayı olduğu özellik kümesine üst N özellikleri eklemektir. Uysal (2016), çalışmasında, daha genel bir özellik seçim şemasındaki son adımın daha temsili bir özellik kümesi elde etmek için modifiye edildiği gelişmiş bir genel özellik seçim şeması önermiştir. Ortak bir özellik seçimi şeması tarafından oluşturulan özellik seti, bazı sınıfları başarıyla temsil etse de bir takım sınıflar gösterilemeyebilir. Sonuç olarak, Bu şema, tüm sınıfları neredeyse eşit olarak temsil eden bir özellik kümesi oluşturarak küresel özellik seçim yöntemlerinin sınıflandırma performansını geliştirmeyi amaçlamaktadır. Bu amaçla, şemadaki özellikleri sınıflar üzerindeki ayırt edici gücüne göre etiketlemek için yerel bir özellik seçme yöntemi kullanılır ve bu etiketler özellik kümelerini üretirken kullanılır. Çeşitli sınıflandırıcılarla tanınmış veri setleri üzerindeki deneysel sonuçlar, bu şemanın yaygın olarak bilinen iki metrik olan Micro-F1 ve Macro-F1 açısından sınıflandırma performansını geliştirdiğini göstermektedir.

Pietramala ve ark. (2008), çalışmalarında Olex-GA olarak adlandırılan bir GA sunmaktadırlar. Olex-GA, verimli bireysel-kurallar için ikili bir gösterim kullanır ve F-ölçüsünü uygunluk fonksiyonu olarak kullanır. Önerilen yaklaşım, Reuters-21578 ve Ohsumed standart test setleri üzerinde test edilmiştir ve çeşitli sınıflandırma algoritmalarıyla (Naive Bayes, Ripper, C4.5, SVM) karşılaştırılmıştır. Deneysel

sonuçlar, her iki veri koleksiyonunda da çok iyi bir performans elde ettiğini ve değerlendirilen sınıflandırıcılarla rekabet edebilecek sonuçlar (hatta bazı durumlarda daha iyi performans sergilediğini) göstermiştir.

3. MATERYAL VE YÖNTEM

Bu tez çalışmasında programlama ortamı olarak MATLAB'dan (Matrix Laboratory) yararlanılmıştır. Ayrıca sınıflandırma yapabilmek için belgelerin önişlenmesi gerekmektedir. Bu aşamada durma kelimelerinin çıkarılması ve kelimelerin gövdelere dönüştürülmesi gereklidir. Daha sonra sınıflandırıcı yardımıyla belgeler önceden belirli sınıflara ayrılırlar. Aşağıda bu tezde kullanılan yöntemler, aşamalar ve algoritmalar hakkında bilgiler verilmiştir.

3.1. Metin Önişleme

Belgeler, sınıflandırma sisteminin hem oluşturulmasında hem de kullanılmasında, bir başka deyişle sistemin hem eğitim hem de test aşamasında, bazı önişleme işlemlerine tabi tutulurlar. Bunlar, belgenin sisteme hazırlanması anlamına gelmektedir. Bu hazırlık işlemlerin bir kısmı yöntem ve araca bağlı olmayan, standart işlemlerdir. Bunlara örnek olarak; eğer belge bir HTML sayfası ise HTML taglarından, scriptlerden arındırılması, bir bilimsel makale ise yazar adı, yayımcısı gibi bilgilerden arındırılması olarak verilebilir. Daha sonra bu belgeler içerdikleri kelimelere parçalanacaktır. Sıradaki işlem bu kelimelerin terimlere dönüştürülmesidir.

Terim çıkarımında Murata ve ark. (2000) göre 4 yöntem mevcuttur (Weng and Lin, 2003). Bunlar, sadece en kısa terimleri kullanmak, bütün terim örüntülerini kullanmak, bir kafes (lattic) kullanmak ve aşağı ağırlıklandırma (down-weighting) metodudur. Bu çalışmada işlemi basitleştirmek amacıyla bu dört yöntemden ilk yöntem seçilmiştir, ilerleyen kısımlardaki açıklanan işlemler buna dayanmaktadır.

3.2. K-En Yakın Komşu Sınıflandırma Yöntemi

Veri madenciliği modellerinden Sınıflandırma Modeli ve algoritmalarına genel bir bakış aşağıda verilmiştir (Silahtaroğlu, 2013).

a. İstatistiğe Dayalı Algoritmalar:

i. Bayesyen sınıflandırma.
ii. Regresyon
iii. Chaid
b. Mesafeye Dayalı Algoritmalar:
i. En yakin komsu
ii. En kucuk mesafe sınıflandırıcısı
c. Karar Ağacları:
i. CART
ii. ID3
iii. C4.5
iv. Sprint
d. GA'lar
e. Yapay Sinir Ağları

Bu algoritma, Cover ve Hart (1967) tarafından önerilen ve denetimli öğrenme adı verilen örnek veri noktalarının mevcut olduğu bir yöntemdir. Bu yöntem en yakın komşuların k değerine göre belirlendiği bir karar aşaması içermektedir. Bu algoritma yaygın olarak kullanılmaktadır. Ayrıca iyi bilinen, çok eski, oldukça basit ve etkili sınıflandırma yöntemlerindendir. Nesnelerin sınıflandırılması, önemli bir araştırma alanı olup veri madenciliği, yapay zekâ çalışmaları, tıp, biyoinformatik gibi oldukça çeşitli alanlarda kullanılmaktadır. KNN algoritmasında eğitim süreci yoktur. Yöntemin gerçekleştirimi kolaydır ve analitik olarak izlenebilir. Çeşitli bilgi türlerine uyarlanabilir ve paralel gerçekleştirime oldukça uygun olup gürültülü eğitim verilerine karşı dayanıklıdır. Bu ve benzeri avantajları, sınıflandırma uygulamalarında çokça tercih edilmesini sağlamaktadır (Bhatia ve Vandana, 2010). Bu avantajların yanı sıra bazı dezavantajları da bulunmaktadır. Örneğin çok miktarda bellek olan alanlarda gereksinim duymaktadır. Veri seti ve öznitelik boyutu arttıkça işlem yükü ve maliyetde önemli ölçüde artar. Ayrıca performans k komşu sayısı, uzaklık ölçütü ve öznitelik sayısı gibi parametrelere bağldır. K sayısının seçimi için belirli bir kriter yoktur. Sezgisel olarak karar verilebilir (Bhatia ve Vandana, 2010).

KNN algoritması, öğrenme algoritmalarının en temel örnek tabanlı algoritmaların başında gelmetedir. Bu algoritmalarda, eğitim setinde tutulan veriler öğrenme işlemine dayalı olarak gerçekleştirilmektedir. Karşılaştığımız yeni örneklerin, elimizde olan eğitim setindeki örnekler ile benzerliği ölçülüp ona göre sınıflandırılmaktadırlar (Mitchell, 1997). KNN algoritmasında, eğitim setindeki örnekler *n* boyutlu sayısal nitelikler ile belirtilmektedir. Her örnek *n* boyutlu uzayda bir noktayı temsil edecek biçimde tüm eğitim örnekleri *n* boyutlu bir örnek uzayında tutulur. Bilinmeyen bir örnek ile karşılaşıldığında, eğitim setinden ilgili örneğe en yakın *k* tane örnek belirlenerek, yeni örneğin sınıf etiketi, *k* en yakın komşusunun sınıf etiketlerinin çoğunluk oylamasına göre atanır (Han ve Kamber, 2006).

3.2.1. K-NN Parametreleri

K-NN algoritmasının performansını etkileyen en önemli parametreler uzaklık ölçütü, komşu sayısı (*k*) ve ağırlıklandırmadır. Aşağıda bu parametreler açıklanmaktadır.

Öklid uzaklığı, sınıflandırma ve kümeleme algoritmalarında en sık kullanılan uzaklık ölçütüdür. Öklid uzaklığı, iki nokta arasındaki doğrusal uzaklık olup herhangi iki nokta, *P* ve *Q* arasındaki Öklid uzaklığı *P*=(*x1, x2,…, xn*) ve *Q*=(*y1, y2,…, yn*) olmak üzere, Eş. 3.1'e göre hesaplanır (Kresse ve Danko, 2012):

$$\left(\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\right) \tag{3.1}$$

Öklid uzaklığı, K-ortalama kümeleme algoritması, temel K-NN algoritması gibi sınıflandırma ve kümeleme algoritmalarında, mesafe hesabı için kullanılan temel uzaklık ölçütüdür.

K-NN algoritmasında, komşu sayısı (*k*) parametresinin değerine dayalı olarak sınıflandırma yapılmaktadır. Sınıflandırma sürecinde, *k*=1 için, sadece en yakın komşunun bulunduğu sınıfa atanırken, *k* sayısı örnek sayısına yaklaştıkça veri setinde yer alan tüm veriler dikkate alınmakta ve bu duruma göre seçim yapılmaktadır.

3.3. Genetik Algoritma

Canlı organizmaların doğasından ilham alınan bir optimizasyon metodudur. GA'nın uygulama alanlarını, sınıflandırma, sayısal yöntemler, doğrudan ve rasgele arama olarak sayabiliriz. Bu algoritma, tekrarlama temelli bir algoritma olup, orijinal ilkeleri daha önce de belirtildiği gibi genetik bilimden uyarlanmıştır ve doğal gelişimde gözlemlenen bir dizi işlemi taklit ederek çalışır. Genel amacı yeni ve geliştirilmiş çözümler üretmektir. Algoritma, sistem optimizasyonu, tanımlama ve kontrol, görüntü işleme ve hibrid problemler, yapay sinir ağlarının topoloji belirlenmesi ve eğitimi ve karar verme sistemleri gibi çeşitli konularda kullanılır. Genetik bilim, biyolojik özelliklerin nesilden nesile nasıl aktarıldığı hakkında çalışılan bir bilimdir. Canlı organizmalarda biyolojik özelliklerin iletilmesinin başlıca nedeni, kromozomlar ve genlerdir. Üstün ve güçlü olan genler ve kromozomlar yeni nesle aktarılır ve zayıf genler kaybolur. Başka bir deyişle, genlerin ve kromozomların etkileşiminin sonucu üstün organizmaların hayatta kalmasıdır. Bu algoritma, meknizmaları optimize etmek, araştırmak ve öğrenmek için kullanılır. Bu algoritmanın temeli, zayıf canlıların kaybolduğunu ve canlıların daha güçlü olduklarını belirten evrim kanunudur. Aslında, bir sürecin evrimi, o alanın varlıklarını temsil eden canlılar üzerinde değil dizeler üzerinde yapılır. Aslında, doğal seleksiyon yasasına göre mevcut adaptasyon olasılığı ne kadar yüksek olursa o kadar çok hayatta kalınabilir ve çoğalma ihtimali daha da artar. GA, doğanın taklitçiliğinden dolayı geleneksel arama yöntemleriyle bir takım temel farklılıklara sahiptir ve aşağıda bunlardan bazıların üzerinde durulmuştur. GA, her biri değişkenlerin tüm kümesini temsil eden bit dizeleri ile çalışır; çoğu yöntem özel olarak bağımsız değişkenlerle ilgilenir. GA, türev bilgileri gerektirmeyen rehberlik amacıyla rastgele seçilim gerçekleştirir. GA, arama yöntemleri seçim mekanizması ve doğal genetiğe dayanmaktadır. Bu algoritmalar rastgele organize edilen bilgiler arasından en uygun dizeleri seçerler. Her jenerasyonda, önceki dizinin en iyi bölümleri ve yeni rastgele bölüm uygun bir yanıt almak için yeni bir dizgi grubunu oluşturulur. Algoritmalar rastgele olmasına rağmen, basit rastgele algoritmalar değildir. Bir arama noktasından en iyi cevaba ilerlemek için arama alanında geçmiş bilgiler verimli bir

şekilde keşfedilir. GA, her tekrarda arama alanının birkaç noktasını hesaba katar, bu nedenle yerel bir maksimuma yaklaşma olasılığı azaltılmış olur. Çoğu geleneksel arama yönteminde (eğim yöntemi), karar veren mekanizmanın kuralları bir noktadan diğerine geçtikçe değişir. Bu yöntemler arama alanlarında birkaç yanıltıcı eğilime sahip olabilirler. Çünkü yerel maksimumlara takılabilirler. Fakat GA, tam popülasyonlar (noktalar) üretir, sonra her bir noktayı tek tek inceler ve içeriğini birleştirerek gelişmiş noktalar da dahil olmak üzere yeni bir popülasyon oluşturur. Bir aramadan bağımsız olarak, GA'daki birçok noktanın aynı anda gözlemlenmesi onların paralel hesaplanmasını gerektirir. Çünkü burada her noktanın evrimi bağımsız bir süreçtir. Bu nedenle, GA, her bir değişken kümesi tarafından oluşturulan çözümlerin kalitesiyle ilgili bilgiye ihtiyaç duyarken, bazı optimizasyon yöntemleri bilgi gerektirir, hatta problemi organize etmeli ve değişkenlerini tam olarak anlamalıdır. GA, sorunun bu gibi özel bilgilerine ihtiyaç duymadığından, çoğu arama yöntemlerinden daha esnektir. Ayrıca, GA, hata ve karar verme yöntemlerini tanımlama şansına sahiptir. Burada arama alanında rasgele ilerlemekle birlikte, arama yöntemleride değiştirilebilir. GA kesin kuralları değil, olasılık kurallarını takip eder.

3.3.1. Genetik Algoritma Operatörleri

Kodlama, algoritmanın en zor aşaması olabilir. GA, parametreler veya problem değişkenleri üzerinde çalışmak yerine kodlanmış formlarıyla ilgilenir. Örneğin kodlama yöntemlerinden biri olan ikili kodlama, problemi bir ikili sayı dizisine dönüştürecektir.

Kodlama türlerinden bazılarına aşağıda yer verilmiştir. Bununla birlikte, genellikle ikili kod kullanılır, ancak birçok durumda, sorunun doğası gereği başka kodlamalar gerekebilir.

- İkili kodlama
- Değer kodlaması
- Ağaç kodlama
- Permütasyon kodlama

İkili Kodlama, GA'lardaki standart dönüşümdür. İkili kodlama, genetik operatörler için en basit kodlama ve en iyi dönüştürmedir. Ancak karmaşık

problemlerde bu tür dönüşüm pek uygun değildir. Çünkü genellikle kromozomların uzunluğunun yanıt bilgisini yansıtabilecek kadar büyük olması gerekmektedir. İkili dönüşümde, popülasyonun üyeleri 0 ve 1 dizeleri haline gelir.

Örneğin algoritma, F (x, y, z) fonksiyonunun maksimum değerini bulmak istediği varsayılırsa 8 bitlik bir ikili kodlamada arama, 0'dan 255'e kadar olan pozitif tamsayılar arasında yapılmalıdır. Her cevap, üç sayıyı X, Y ve Z'yi içerir.

Aralıktaki herhangi bir sayının uzunluğu, ikili dönüştürmede 8 bite kadar çıkabilir. Her bir kromozomu XYZ olarak kabul eder ve olası tüm yanıtları kapsayacak olursa, kromozomun uzunluğu $3 \times 8 = 24$ olmalı ve bunun için Eş. 3.2'de gösterildiği gibi kromozom C'ye sahip olabilir.

C=11010010 11100011 00110111 (3.2)

Bu durumda, gerekirse negatif sayılar aranabilir. Her dizenin başına bir bit eklenebilir, örneğin 0 ise, sayı pozitif ve 1 ise sayı negatif kabul edilir Eş. 3.3.

000000001=1 (3.3)
100000001=-1

Bu gibi düzenlemeleri kullanarak ondalık sayıların dönüştürülmesi de yapılabilir.

Bir diğer kodlama türü olan Değer kodlama yönteminde, kromozomlar problemle ilgili her türlü veriyi kendi alanlarından alabilirler. Yani bu veriler, gerçek sayılar, mantıksal cümleler, kodlanmış veriler ve alfa sayısal dizgelerden oluşabilirler.

Popülasyon, veya GA popülasyonu kavramı, doğal yaşamda var olana benzer. Popülasyon, doğru veya yanlış olabilen problemin alternatif çözümleri olan yanıtlar olarak düşünülebilir. Bu yanıtlar olası cevaplardır. Örneğin, problem tamsayılar kümesinde bir fonksiyonun maksimumunu bulmaksa, tüm tamsayılar sorunun olası cevabı olarak düşünülebilir. GA'da, ilk adım olarak, başlangıç popülasyonu adı verilen bir dizi çözüm üretmek gereklidir. Bu koleksiyonun üyeleri genellikle rastgele seçilir, ancak iyi algoritmalarda popülasyonun aşırı dağılmasını önlemek için farklı yöntemler

kullanılabilir. Popülasyonun üye sayısı problemin türüne bağlıdır. Aslında üye sayısı, değiştirildiğinde yanıtların doğruluğunu ve arama yakınsamasının hızını artırabilen bir parametredir. Bazı durumlarda, 8 üyeli bir nüfus mükemmel bir şekilde uygun olabildiği gibi; bazılarında ise 100 bireylik bir populasyon bile yeterli değildir. Tecrübelere dayanarak, populasyonun birey sayısı 10 ile 160 arasında seçilebilir.

Değerlendirme veya Uygunluk fonksiyonu, hedef fonksiyondan yani optimize edilmesini istediğimiz fonksiyondan oluşur. Bu fonksiyon her çözüm alternatifini bu alternatifin kalitesini gösteren sayısal bir değere dönüştürür. Çözüm alternatifinin kalitesi ne kadar yüksek olursa, değerde o kadar yüksek olur; bu çözümün gelecek nesilde yer alama olasılığıda o kadar artar.

Seçim safhası, takip eden safhalar için kromozom çiftlerinin seçilmesinden sorumludur. Bu aşama iki kuşak arasında arayüzü oluşturur ve mevcut kuşaktan bazı üyeleri bir sonraki kuşağa aktarır. Seçildikten sonra genetik operatörler iki seçilen üyeye uygulanır, ancak seçim işlemi rasgele bir yapıya sahiptir. Belki de en iyi iki bireyin ilk etapta seçileceği şekilde doğrudan ve sıralı seçim iyi bir yaklaşım olabilir. Ancak GA'mızda neslin iyi bir üyesi olmasa da yani düşük uygunluğa sahip bir bireyle de eşleştirilebilir. Farklı seçme yöntemleri mevcut olmasına karşın, iyi bireylerin seçilme ihtimalinin fazla olması daha uygun çözümlerin bulunmasına katkıda bulunacaktır. Bir probleme en uygun seçim yöntemini belirlemek, daha uygun üyeleri seçmeyi mümkün kılacak şekilde yöntemin seçilmesine bağlıdır. Seçim mümkün olduğunca her neslin önceki nesilden daha iyi bir ortalamaya sahip olacağı şekilde yapılmalıdır.

Örnek bazı seçme yöntemleri şunlardır:

- Rulet çarkı
- Sıralı seçim
- Boltzmann'ın Seçimi
- Kararlı durum seçimi
- Rekabetçi seçim
- Brindle'ın kesin seçimi
- Trunuva Seçimi
- Rastgele eşleme seçimi

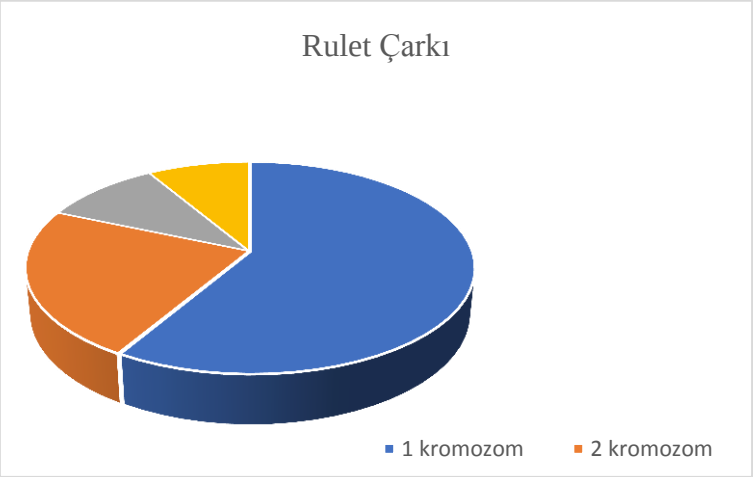
Rulet çarkı, en uygun rastgele seçim yöntemlerinden biridir. Karşılık gelen kromozomun uygunluğuna göre seçilme olasılığı (F_k) bu kromozoma karşılık gelen hayatta kalma ihtimali (P_k) Eş. 3.4'deki formülle hesaplanır.

$$P_k = \frac{F_k}{\sum_{i=1}^{n} F_i} \tag{3.4}$$

sonra, Eş. 3.5'deki formül yardımıyla P_k değerleri kümülatif q_k değerlerine dönüştürülür (Kurucu, 2009).

$$q_k = \sum_{i}^{k} P_i \tag{3.5}$$

Rulet çarkı, her kromozomu seçmek için sıfır ile bir arasında rastgele bir sayı kullanır. Bu sayı, ilgili kromozomun yer aldığı bölgenin seçimini sağlar. Elbette, rulet çarkının uygulanışı, gerçek hayattaki rulet çarkı örneklerinde olduğu gibidir. Bir daire vardır ve bu daire kromozomal uygunluğa tekabül eden bir dizi parçaya bölünür. Tekerlek çevirilir ve tekerleğin durduğu her yer seçilen kromozomu ifade edecektir. Örnek bir Rulet çarkı Şekil 3.1 de görülmektedir.



Şekil 3.1. Örnek rulet çarkı.

Bir rulet çarkını çevirmek, aslında eşleşen değerin oranını seçen bir yöntemdir. Bu yöntem, hangi üyelerin yeniden üretme şansı bulunduğunu belirlemek için bir ruleti simüle eder.

Rulet çarkı yönteminde, en iyi kromozomun uygunluk değeri çok yüksek ise sürekli yüksek olasılığa sahip kromozom seçilecek olması sıkıntı oluşturabilir. Bu nedenle, Rank seçim yöntemi uygulanabilir. Bu yöntemde popülasyon uygunluk değerine göre tersten sıralanır ve 1'den N'e kadar kromozomlara değer atanır. Yani en iyi kromozom N adetlik bir popülasyonda N değerini alır. Seçim bu değerlere göre yapılır.

GA'da en önemli operatör çaprazlamadır. Bu aşama, eski nesil kromozomların karıştırıldığı ve yeni bir kromozom kuşağı oluşturmak için birleştirildiği bir süreçtir. Seçim bölümünde ebeveyn olarak kabul edilen çiftler, genlerini bir araya getirir ve yeni üyeler oluşturur. GA'da çaprazlama, popülasyonun genetik çeşitliliği üzerinde çok olumlu etkisi oluşturmasa bile kromozomların iyi genlerinin bulunmasına olanak sağlar.

İkili kodlamada en sık kullanılan çaprazlama yöntemlerinden bir tek noktalı çaprazlamadır. Bu yöntemde tek noktaya göre bir yer değiştirme yapılır. Önce ebeveyn kromozomlarının çifti (ikili dize) dize belirli noktalarda kesilir ve kesme noktalarının parçaları karşılıklı değiştirilir ve böylece her biri kromozomlardan genler içeren iki yeni kromozom ortaya çıkar. Ebeveyn özellikleirni miras alırlar**.**

Bir diğer çaprazlama yöntemi ise artimetik çaprazlamadır. Bu yöntemde ebeveyn bireylerden oluşan yeni çocuk bireyler rasgele seçilen bir λ_1 değerine bağlı olarak Eş. 3.6'deki formüllerle hesaplanır (Erdoğan, 2013).

$$\lambda_1 + \lambda_2 = 1$$
$$h_i^1 = \lambda_1 C_i^2 + \lambda_2 C_i^1 \qquad (3.6)$$
$$h_i^2 = \lambda_1 C_i^1 + \lambda_2 C_i^2$$

Mutasyon, doğada ultraviyole ışınları gibi bazı faktörler kromozomlarda öngörülemeyen değişikliklere neden olur. GA'lar evrim kanununu takip ettiğinden, bu algoritmalarda düşük frekanslı bir sıçrama operatörü uygulanmaktadır.

Mutasyon, olası diğer çözüm alternatiflerini doğuran bir diğer önemli işlevdir. GA'da, yeni bir popülasyonda bir üye oluşturulduktan sonra, herhangi bir gen mutasyon yardımıyla değişebilir. Mutasyonda bir gen değişerek, populasyondan bir başka gene dönüşebileceği gibi populasyonda hiç bulunmayan bir gen de dönüşebilir. Bir genin mutasyonu, o genin değişimi anlamına gelir ve mutasyon yöntemleri kodlama türüne bağlıdır. Mutasyonun en önemli görevinin, yakınsamadan yerel olarak uzaklaşılmasını önlemek olduğu söylenebilir.

Heuristik Mutasyon Yönteminde, mutasyonun amacının kromozomları iyileştirmek ya da yeni bir çözüm bulmak olduğu göz önüne alınırsa, rastgele kromozom yerine kromozomlardaki değişiklikler hedeflenebilir. Bunu yapmak için, problemin türüne bağlı olarak, seçilen kromozom üzerinde, klasik problem çözme yöntemlerinden birini uygulamak ve oluşan cevabı yeni bir kromozom olarak değiştirmek gerekmektedir. "Mutasyon" olarak bilinen bu yöntemin kullanılması, nihai çözümü elde etmek için problemin türüne bağlı olarak yöntemi hızlandırabilir.

Mutasyonlar popülasyonda değişikliklere neden olur ve yeni bireylerin oluşmasını sağlar. Aslında, mutasyonun bu açıdan çaprazlama ile aynı avantaja sahiptir. Sadece mutasyon kullanılırsa, en iyi cevap bulunabilir, ancak sadece çaprazlama kullanılırsa optimal cevabın bulunması garanti edilemez.

Kod çözme, kodlama işleminin tersidir. Bu noktada, algoritma sorunun en iyi cevabını sunduktan sonra, cevabın gerçek versiyonunu net bir şekilde görebilmemiz için kod çözme sürecini uygulamak gereklidir.

Sonlandırma yöntemi, GA'lar üretim ve teste dayandığından, problemin cevabı bilinmemektedir. Oluşturulan yanıtlardan hangisinin en uygun çözüm olduğu genellikle bilinmez. Popülasyondaki nihai cevabı bulunması için sonlandırma koşulunu tanımlamak gerekmektedir. Bu nedenle, sonlandırma şartı için farklı kriterler düşünülebilir:

- Belirli sayıda nesil: Sonlandırma kriteri örneğin orijinal döngünün 100 turu olabilir.

- Ardışık nesiller boyunca popülasyonun en iyi bireyinde iyileşme olmaması

- Popülasyonun uygunluğu belirli bir değerin altında değişmesi veya birkaç ardışık nesilde boyunca değişmemesi

- Popülasyonun en iyi değeri için önceden belirlenmiş bir değer olması

Bunların kombinasyonları da sonlandırma kriteri olarak kullanılabilmektedir.

3.4. Temel Bileşenler Analizi

Temel Bileşenler Analizi (TBA) Karl Pearson tarafından 1901 yılında ortaya atılmıştır (Filiz, 2003). TBA, birbiri ile ilişkili değişkenler içeren verinin boyutlarını veri içerisinde varolan değişimlerin mümkün olduğunca korunarak daha az boyuta indirgenmesini sağlar (Çilli, 2007). Analiz, eldeki veriyi daha az değişken kullanarak ifade edebilecek en iyi dönüşümü belirlemeyi amaçlamaktadır. Dönüşüm sonrasında elde edilen değişkenler ilk değişkenlerin temel bileşenleri olarak adlandırılır. İlk temel bileşen varyans değeri en büyük olandır ve diğer temel bileşenler varyans değerleri azalacak şekilde sıralanır. Gürültüye karşı düşük hassasiyeti, hafıza ihtiyaçlarını azaltması, az boyutlu uzaylarda daha etkin çalışması TBA'nın temel avantajları arasında sıralanabilir (Sütçüler, 2006).

4. BULGULAR

4.1. Veri Kümesi

Bu çalışmada kullanılan metin belge koleksiyonlarından ilki Classic 3 veri setidir. Kullanılan bu veri setinde CACM, CRAN ve MED olarak adlandırılan 3 kategoriye ait belgeler bulunmaktadır. Bu kategorilere ait belge sayıları ise sırasıyla 1589, 1398 ve 1033'tür. Bu belge koleksiyonundan eşit dağılımlı 600 belge seçilmiş, bu belgelerin yarısı eğitim yarısı da test verisi olarak kullanılmıştır.

Bu çalışmadaki ikinci metin belge kolleksiyonu ise metin madenciliği araştırmalarında sıkça kullanılan Reuters-21578 dağıtım 1.0'dır. Bu koleksiyonun özelliği kategorilerin hiyerarşik olmamasıdır. Koleksiyondaki bazı belgeler aynı anda birden fazla kategoriye ait olabilmektedir. Bu koleksiyon 135'in üzerinde kategoriye sahip 21578 belge içermektedir. Mevcut olan 135 konudan bazıları çok az sayıda belgede bulunmaktadır. Bu yüzden bu çalışmada kullanılmak üzere 135 kategoriden en çok yer alan 3'üne ait belgeler seçilmiştir. Seçilen bu konulara ait toplam 600 belge mevcut olup, bunlardan 300 belge eğitim verisi olarak, 300 belge ise test verisi olarak kullanılmıştır.

Kullanılan veri kümeleri öncelikle ön işlemeye tabi tutulmuştur. İlk olarak 350 kelimeden oluşan stop-words kelimeleri bu belgelerden çıkarılmıştır. Kelimelerin gövdeleme işlemi için, metin madenciliğinde yaygın bir şekilde kullanılan Porter Stemmer algoritması seçilmiştir. Bunun sonucunda belgeler içerdikleri gövde kelimelere, yani terimlere göre modellenmiştir.

4.2. Classic 3 Veri Kümesi için Sınıflandırma Sonuçları

Veri madenciliğinde olduğu gibi metin madenciliği çalışmalarında da veri miktarı oldukça fazladır. Ayrıca Metin madenciliği çalışmaları için kullanılacak olan verilerin yüksek boyutlu olma durumlarından kaynaklı üzerinde çalışma yapılması

zordur. Bu sebeple genellikle metinsel verilerdeki bu yüksek boyutluluk problemi nitelik azaltma ile aşılmaya çalışılmaktadır.

Öncelikle mevcut veri setinin sınıflandırma başarısına bakılmıştır. *K* paramteresinin 1 olarak seçildiği KNN yönteminin kullanılmasıyla gerçekleştirilen sınıflandırmanın başarısı % 64.66 olarak bulunmuştur. Bu sınflandırmada mevcut 1027 niteliğin tamamı kullanılmıştır. *k* parametresini değiştirerek farklı sonuçlar elde edilir, ancak ortalama olarak aynı yüzdeler elde edilir.

Nitelik azaltma özellik çıkarma veya özellik seçme yaklaşımları ile gerçekleştirilmektedir. Bu alandaki yaygın olarak kullanılan yöntemlerden bazıları şunlardır: GA, Temel Bileşen analizi (TBA), Bilgi kazancı, vb.

Bu çalışmada kullanılan ilk veri kümesindeki veriler, Classic 3 veri kümesinden alınan 600 adet belgeden oluşmaktadır. Bu belge kolleksiyonu içerisinden 9 ve daha az sayıda yer alan terimler çıkarılmıştır. Bu aşamadan sonra her bir belge 1027 adet terim (sütun) ile ifade edilmiştir. Yani veri boyutumuz 600 x 1027 adettir.

4.2.1. Classic 3 veri seti için özellik seçimi

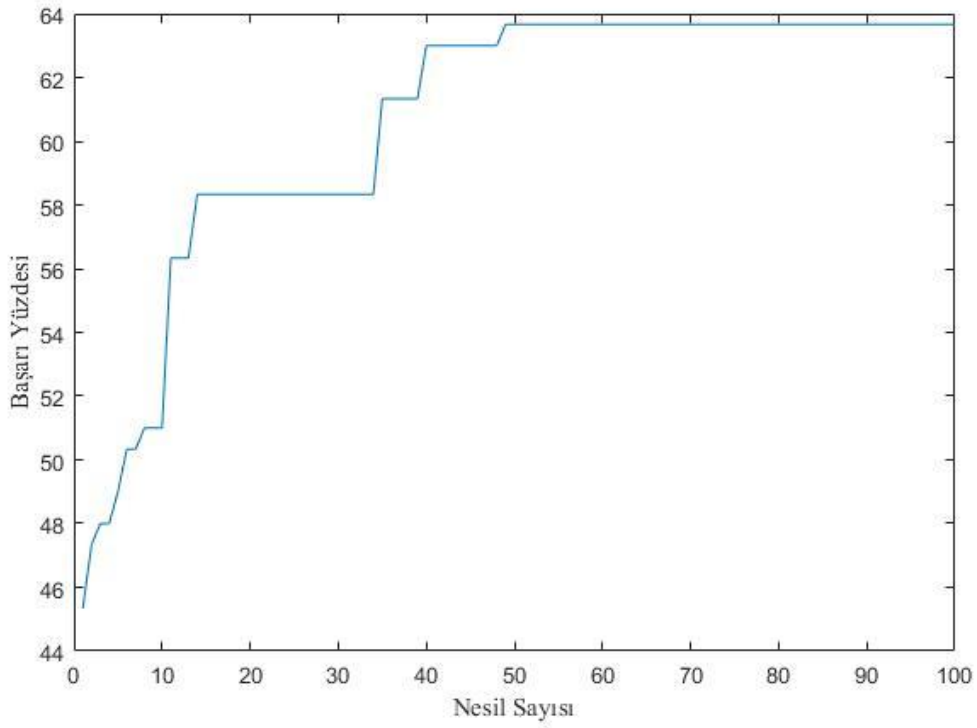
Bu aşamada iki farklı nitelik azaltma yöntemi kulanılmıştır. Bunlardan ilki özellik seçme yaklaşımıdır. Özellik seçme işlemi için seçilen yöntem ise GA'dır. Bu uygulamada nitelikler terimler (kelimeler) anlamına gelmektedir. Özellik seçme bir diğer deyişle terim seçimidir. Sınıflandırma başarısını en iyi yapan terimlerin seçimi amaçlanır. Bu özellik seçimi için iki ayrı GA programı kullanılmıştır. Bunlardan ilki MATLAB programı kullanılarak hazırlanan GA (GA1), diğer ise MATLAB toolbox kullanılarak hazırlanan GA (GA2) programlarıdır.

Şekil 4.1'de GA1 programının örnek bir çalıştırma sonucu görülmektedir. Başlangıç popülasyonda en iyi bireyin başarısı ortalama % 43.34 olup 250 nesil sonra ortalama % 58.69 başarılı çözümlere ulaşılmıştır. Bu başarı hesaplanırken kullanılan KNN sınıflandırıcısında *k* parametresi 3 olarak belirlenmiştir.

GA1 programı için kullanılan parametre bilgileri Çizelge 4.1'de verilmiştir.

Çizelge 4.1. classic 3 veri seti ve GA1 programının parametreleri

Parametre	*Değer*
Gen Sayısı (Terim Sayısı)	10
Kodlama	Gerçek-Değer kodlama
Popülasyon Boyutu	10
Seçme	Rulet Çarkı – Rank
Çaprazlama	Aritmetik Çaprazlama
Çaprazlama Oranı	1.0
Mutasyon	Sezgisel Mutasyon
Mutasyon Oranı	0.1
Sonlandırma Ölçütü	İterasyon Sayısı (250 iterasyon)



Şekil 4.1. classic 3 veri seti GA1 programı sonuçları.

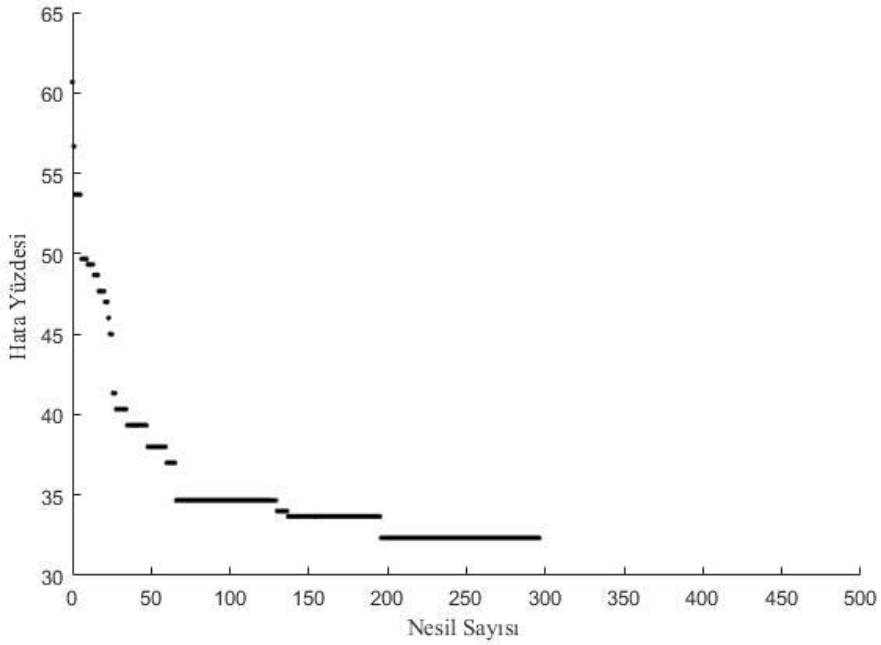
Şekil 4.2'de GA2 programının örnek bir çalıştırma sonucu görülmektedir. Başlangıç popülasyonda en iyi bireylerin başarısı ortalama % 40.11 olup 500 nesil sonra

ortalama % 71.47 başarılı çözümlere ulaşılmıştır. Bu başarı hesaplanırken kullanılan KNN sınıflandırıcısında K parametresi 1 olarak belirlenmiştir.

Çizelge 4.2. classic 3 veri seti ve GA2 programının parametreleri

Parametre	*Değer*
Gen Sayısı (Terim Sayısı)	10
Kodlama	Gerçek-Değer kodlama
Popülasyon Boyutu	10
Seçme	Rulet Çarkı - Rank
Çaprazlama	Aritmetik Çaprazlama
Çaprazlama Oranı	0.8
Mutasyon	Sezgisel Mutasyon
Mutasyon Oranı	0.1
Sonlandırma Ölçütü	İterasyon Sayısı (500 iterasyon)

GA2 programı için kullanılan parametre bilgileri Çizelge 4.2'de verilmiştir.



Şekil 4.2. classic 3 veri seti ve GA2 programı sonuçları.

4.2.2 Classic 3 veri seti için özellik çıkarımı

Nitelik çıkarımı kullanılarak yapılacak olan nitelik azaltma yöntemi TBA dir. TBA yöntemi kullanılarak veri setinin boyutu azaltılmıştır. Veri kümesinin % 30.55 özelliğini gösteren ilk on sütun seçilmiştir. Bu sebeple veri seti 600 x10 boyutuna indirgenmiştir.

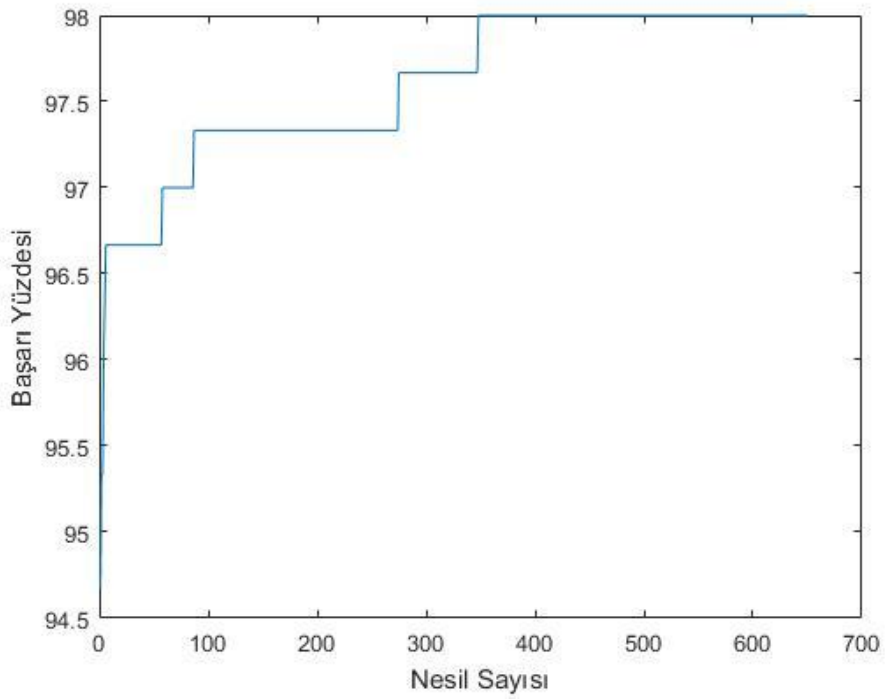
Bu çalışmada amaç sınıflandırma başarısını artırmaktır. Bunun için önce verilerin sınıflandırlması yapılmış ve çalışmanın başlangıcındaki başarı belirlenmiştir. Sınıflandırmada kullanılan yöntem k en Yakın komşu algoritmasıdır. İlk veri kümesi için sınıflandırma başarısı % 94.66 olarak bulunmuştur. Bu aşamada sınıflandırma yönteminin *k* parametresi 5 olarak seçilmiştir. Bu değerin seçimi yapılırken deneme yanılma yöntemi kullanılmıştır. 1-20 arası değerler denenmiş ve en iyi sonucu veren k seçilmiştir. Sınıflandırmanın başarısını artırmak için herbir öz nitelik değeri birer katsayı ile çarpılır. Bu katsayıların en iyi değerini seçmek için bir optimizasyon uygulamak gereklidir. Bunun için seçilen yöntem GATBA 'dır.

Şekil 4.3'de örnek bir program çıktısı görülmektedir. Bu veri setinin sınıflandırma başarısı GATBA 'nın başlangıcında ortalama % 94.660 değerinden başlamıştır. Algoritma 300 nesil çalışmış ve başarı değeri ortalama % 97.717değerine yükselmiştir. Bu değer sınıflandırmada katsayı kullanılmasının kullanılmaması alternatifine göre başarıyı yaklaşık % 3.057 artırdığı anlamına gelmektedir. Bu aşamadaki KNN sınıflandırıcısının *k* parametresi 5 olarak seçilmiştir.

GATBA'da kullanılan parametreler Çizelge 4.3'de verilmiştir. GATBA sonucunda elde edilen başarı değerleri Şekil 4.3'de gösterilmiştir.

Çizelge 4.3. classic 3 veri seti için GATBA parametreleri

Parametre	*Değer*
Gen Sayısı (Terim Sayısı)	10
Kodlama	Gerçek-Değer kodlama
Popülasyon Boyutu	10
Seçme	Rulet Çarkı – Rank
Çaprazlama	Aritmetik Çaprazlama
Çaprazlama Oranı	1.0
Mutasyon	Sezgisel Mutasyon
Mutasyon Oranı	0.1
Sonlandırma Ölçütü	İterasyon Sayısı (300 iterasyon)



Şekil 4.3. classic 3 veri kümesi için GATBA sonuçları.

4.3. Reuters Veri Kümesi İçin Sınıflandırma Sonuçları

Reuters veri kümesinden en fazla belgeye sahip olan ilk üç sınıfa ait belgelerden seçilen 600 adet belge mevcuttur. Bu kolleksiyon içerisinde 10'dan daha az sayıda yer alan terimler çıkarılmıştır. Bu aşamadan sonra her bir belge 4377 adet terim (sütun) ile ifade edilmiştir. Yani veri boyutumuz 600x4377 adettir.

İlk olarak mevcut veri setinin sınıflandırma başarısı araştırılmıştır. *K* parametre değeri 1 olarak kullanılan KNN yöntemiyle yapılan sıflandırmanın başarısı % 81 olarak bulunmuştur. Bu sınflandırmada mevcut 4377 niteliğin tamamı kullanılmıştır.

4.3.1. Reuters veri seti için özellik seçimi

Bu aşamada Reutres veri seti iki farklı nitelik azaltma yöntemi kulanılmıştır. Bunlardan ilki özellik seçme yaklaşımıdır. Özellik seçme işlemi için seçilen yöntem ise yine GA'dır. Bu uygulamada nitelikler terimler (kelimeler) anlamına gelmektedir. Sınıflandırma başarısını en iyi yapan terimlerin seçimi amaçlanır. Bu özellik seçimi için

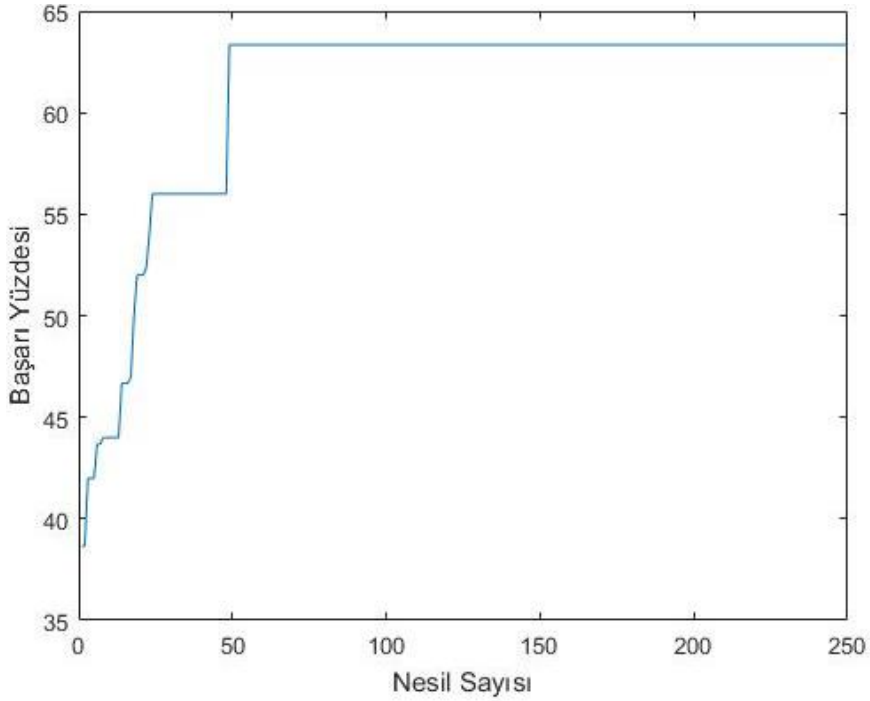
iki ayrı GA programı kullanılmıştır. Bunlardan ilki MATLAB programı kullanılarak hazırlanan GA (GA1), diğer ise MATLAB toolbox kullanılarak hazırlanan GA (GA2) programlarıdır.

Şekil 4.4'de GA1 programının örnek bir çalıştırma sonucu görülmektedir. Başlangıç popülasyonda en iyi bireyin başarısı ortalama % 45.36 olup 250 nesil sonra ortalama % 58.93 başarılı çözümlere ulaşılmıştır. Bu başarı hesaplanırken kullanılan KNN sınıflandırıcısında *k* parametresi 7 olarak belirlenmiştir.

GA1 programı için kullanılan parametre bilgileri Çizelge 4.4'de verilmiştir.

Çizelge 4.4. reuters veri seti ve GA1 programının parametreleri

Parametre	*Değer*
Gen Sayısı (Terim Sayısı)	10
Kodlama	Gerçek-Değer kodlama
Popülasyon Boyutu	10
Seçme	Rulet Çarkı – Rank
Çaprazlama	Aritmetik Çaprazlama
Çaprazlama Oranı	1.0
Mutasyon	Sezgisel Mutasyon
Mutasyon Oranı	0.1
Sonlandırma Ölçütü	İterasyon Sayısı (200 iterasyon)



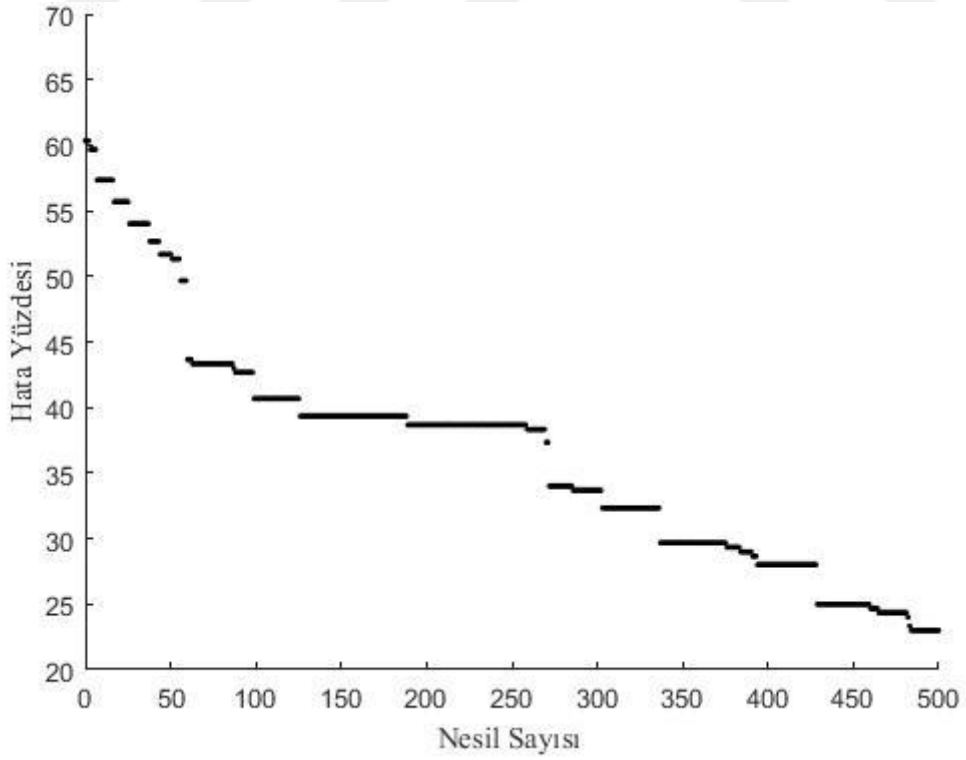
Şekil 4.4. reuters veri seti ve GA1 programı sonuçları.

Şekil 4.5'de GA2 programının örnek bir çalıştırma sonucu görülmektedir. Başlangıç popülasyonda en iyi bireylerin başarısı ortalama % 35.35 olup 300 nesil sonra ortalama % 69.57 başarılı çözümlere ulaşılmıştır. Bu başarı hesaplanırken kullanılan KNN sınıflandırıcısında *k* parametresi 1 olarak belirlenmiştir.

Çizelge 4.5. reuters veri seti ve GA2 programının parametreleri

Parametre	*Değer*
Gen Sayısı (Terim Sayısı)	10
Kodlama	Gerçek-Değer kodlama
Popülasyon Boyutu	10
Seçme	Rulet Çarkı – Rank
Çaprazlama	Aritmetik Çaprazlama
Çaprazlama Oranı	0,8
Mutasyon	Sezgisel Mutasyon
Mutasyon Oranı	0,1
Sonlandırma Ölçütü	İterasyon Sayısı (300 iterasyon)

GA2 programı için kullanılan parametre bilgileri Çizelge 4.5'de verilmiştir.



Şekil 4.5. reuters veri seti ve GA2 programı sonuçları.

4.3.2. Reuters veri seti için özellik çıkarımı

Bu veri setinde de kullanılan ikinci nitelik azaltma yöntemi yine TBA dir. TBA yöntemi kullanılarak kullanılan veri kümesinin boyutu azaltılmıştır. Veri kümesinin % 37.35 özelliğini gösteren ilk on sütun seçilmiştir. Bu sayede veri seti 600 x10 boyutuna indirgenmiştir.

Sınıflandırma başarısını artırmak amaçlanmıştır. Bunun için önce verilerin sınıflandırlması yapılmış ve çalışmanın başlangıcındaki başarı belirlenmiştir. Sınıflandırmada kullanılan yöntem k en yakın komşu algoritmasıdır. İlk veri kümesi için sınıflandırma başarısı % 92.395 olarak bulunmuştur. Bu aşamada sınıflandırma yönteminin *k* parametresi 1 olarak seçilmiştir. Bu değerin seçimi yapılırken deneme yanılma yöntemi kullanılmıştır. 1-20 arası değerler denenmiş ve en iyi sonucu veren k seçilmiştir.

Sınıflandırmanın başarısını artırmak için öz nitelik değeri katsayılar ile çarpılmaktadır. En iyi sonucu veren katsayıları bulmak için yine GATBA yöntemi kullanılmıştır.

Şekil 4.6'dan anlaşılacağı üzere veri setinin sınıflandırma başarısı GATBA 'nın başlangıcında % 94.33 değerinden başlamıştır. Algoritma 350 nesil çalışmış ve başarı değeri % 96.607 değerine yükselmiştir. Bu değer sınıflandırmada katsayı kullanılmasının katsayı kullanılmaması alternatifine göre başarıyı % 2.277 artırdığı anlamına gelmektedir.

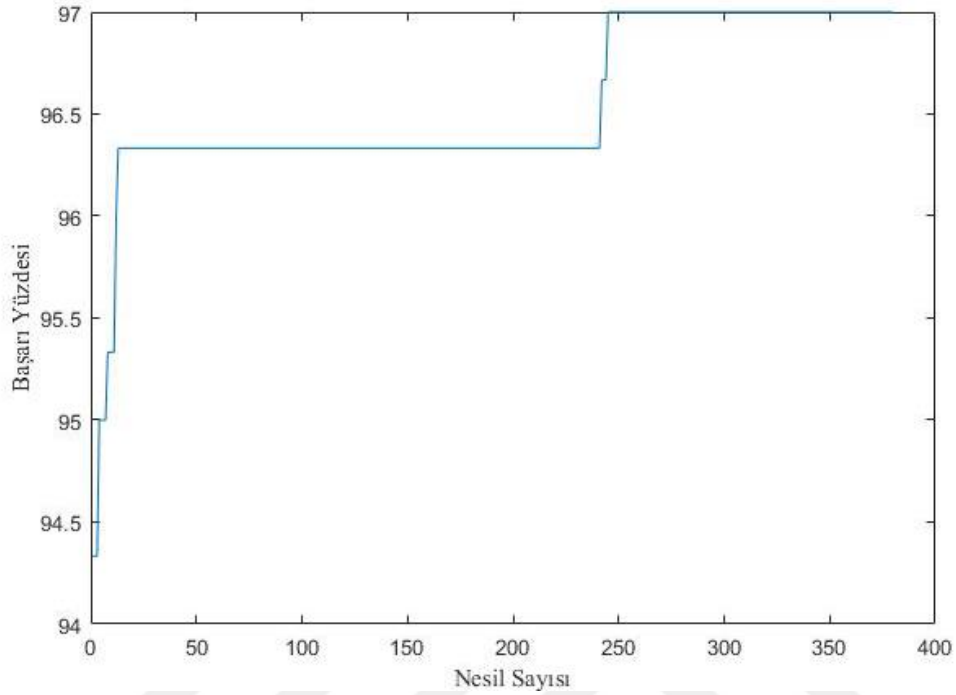
Oluşturulan bu GATBA'da için parametre değerleri Çizelge 4.6'da gösterilmiştir.

Çizelge 4.6. reuters veri seti için kullanılan GATBA parametreleri

Parametre	*Değer*
Gen Sayısı (Terim Sayısı)	10
Kodlama	Gerçek-Değer kodlama
Popülasyon Boyutu	10
Seçme	Rulet Çarkı – Rank
Çaprazlama	Aritmetik Çaprazlama
Çaprazlama Oranı	1.0
Mutasyon	Sezgisel Mutasyon

Mutasyon Oranı	0.1
Sonlandırma Ölçütü	İterasyon Sayısı (350 iterasyon)

GATBA sonucunda elde edilen başarı değerleri Şekil 4.6'da verilmiştir.



Şekil 4.6. reuters veri kümesi için GATBA sonuçları.

Her bir program 10 kez çalıştırılması sonucu elde edilen sonuçlardan hesaplanan ortalama başarı yüzdesi değerleri Classic 3 veri seti için çizelge 4.7'de, Reuters veri seti için ise 4.8'de verilmiştir.

Çizelge 4.7. classi 3 veri seti sonuçları

Classic 3	*Başarı Yüzdesi*
KNN	64.66
GA1 – KNN (özellik seçme)	58.697
GA2 – KNN (özellik seçme)	71.47
PCA – KNN	94.33
PCA - *k* – GATBA – KNN	97.717

Çizelge 4.8. reuters veri seti sonuçları

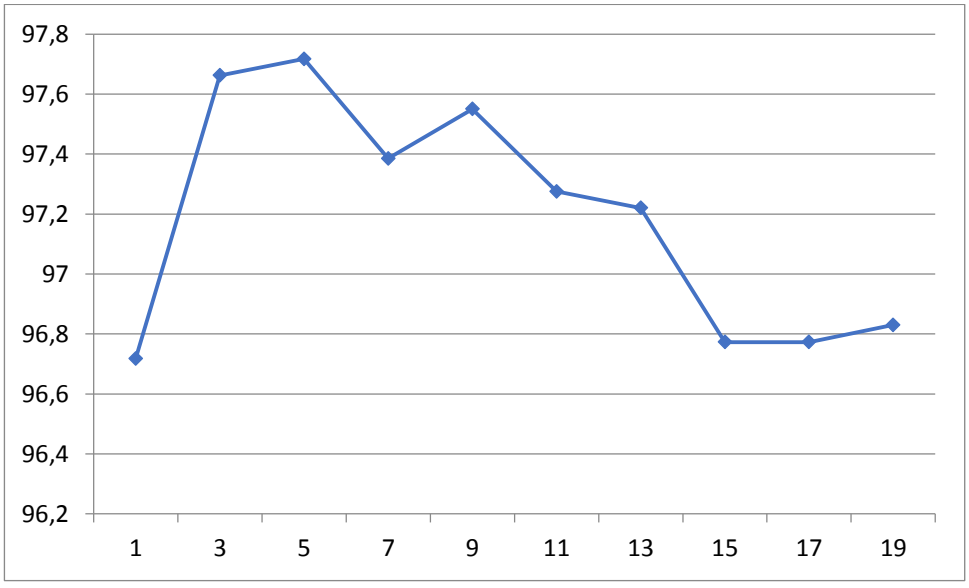
Reuters	*Başarı Yüzdesi*
KNN	81
GA1 – KNN (özellik seçme)	58.93
GA2 – KNN (özellik seçme)	69.57

PCA – KNN	94.33
PCA - *k* – GATBA – KNN	96.607

Bu çalışmada en önemli olan noktalardan biri ise en iyi sonucu veren *k* parametresinin tespit edilmesidir. Bunun için k parametrsi 1-19 aralığında denenerek en iyi değer seçilmiştir. Ayrıntılı olarak sonuçlar Classic 3 veri seti için çizelge 4.9'da ve Reuters veri seti için ise 4.10'da gösterilmiştir.

Çizelge 4.9. classic 3 veri seti için *k* parametresinin sınıflandırma başarıları

k Parametresi	*Ortalama*	*Standart Sapma*
1	96.717	0.139
3	97.662	0.212
5	97.717	0.328
7	97.385	0.135
9	97.550	0.170
11	97.275	0.248
13	97.220	0.170
15	96.773	0.176
17	96.773	0.176
19	96.830	0.186

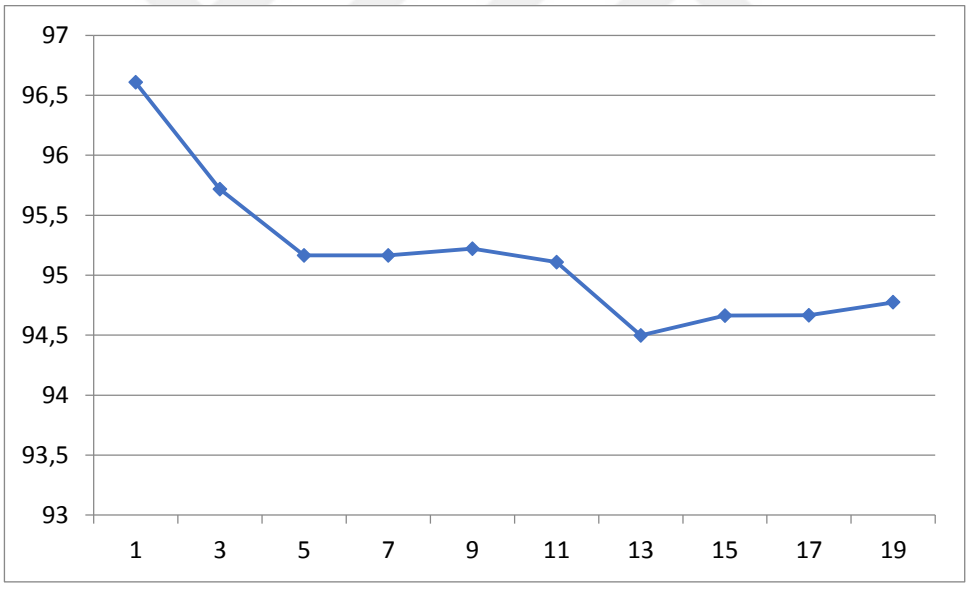


Şekil 4.7. classic 3 veri kümesi için *k* sonuçları.

Classic 3 veri seti için *k* değerinin 5 olduğu durumda sınıflandırma başarısı 97.717 değeri ile en iyi sonuç olmuştur.

Çizelge 4.10. reuters veri seti için *k* parametresinin sınıflandırma başarıları

k Parametresi	*Ortalama*	*Standart Sapma*
1	96.607	0.252
3	95.717	0.139
5	95.165	0.181
7	95.165	0.181
9	95.220	0.170
11	95.108	0.273
13	94.497	0.349
15	94.663	0.422
17	94.665	0.631
19	94.773	0.345



Şekil 4.8. reuters veri kümesi için *k* sonuçları.

Reuters veri seti için *k* değerinin 1 olduğu durumda sınıflandırma başarısı 96.607 değeri ile en iyi sonuç olarak görülmektedir.

5. TARTIŞMA VE SONUÇ

Genel olarak yapılmak istenen uygulama için mevcut iki veri seti 3 ayrı programa tabi tutulmuştur ve programların sonuçları karşılaştırılarak en iyi sonuca ulaşılması aşama aşama çizelgeler ve şekillerle gösterilmiştir.

İlk olarak metin belge koleksiyonlarından Classic 3 veri seti hazırlanan programlara tabi tutulmuştur. Kullanılan bu veri setinde CACM, CRAN ve MED olarak adlandırılan 3 kategoriye ait belgeler bulunmaktadır. Bu kategorilere ait belge sayıları ise sırasıyla 1589, 1398 ve 1033'tür. Bu belge koleksiyonundan eşit dağılımlı 600 belge seçilmiş, bu belgelerin yarısı eğitim yarısı da test verisi olarak kullanılmıştır.

Öncelikle mevcut veri setinin sınıflandırma başarısı *k* parametre değeri 1 olarak kullanılan KNN yöntemiyle % 64,66 olarak bulunmuştur. Bu sınflandırmada mevcut 1027 niteliğin tamamı kullanılmıştır. Classic 3 veri seti için uygulanan programların sonucunda elde edilen yüzdelik sınıflandırma başarısı değerlerinden anlaşıldığı gibi ilk olarak GA ile veri setini özellik seçimine tabi tutulmuştur. Sınıflandırma başarısının yüzdesi ortalama olarak 250 nesil sonra % 43.34'den % 58.69'e ulaşmıştır. Bu başarı hesaplanırken kullanılan KNN sınıflandırılmasındaki *k* parametresi 3 olarak belirlenmiştir. Daha sonra aynı verileri MATLAB'daki GA toolbox yazılımı kullanılarak hazırlana program ile sınıflandırma başarısı ortalama % 40.11 değerinden başlayıp 500 nesil sonra ortalama % 71.47'a ulaşmıştır. Başarı hesaplanırken kullanılan KNN sınıflandırıcısında *k* parametresi 1 olarak belirlenmiştir.

Bir sonraki aşamada, sınıflandırma uygulaması için özellik çıkarımı kullanılarak yapılan boyut azaltma yöntemi yani TBA kullanılmıştır. Sınıflandırmanın başarısını artırmak için herbir öz nitelik değerinin birer katsayı ile çarpılması amaçlanmıştır. Bu katsayıların en iyi değerinin seçilmesi için bir optimizasyon gereklidir. Bunun için seçilen yöntem GATBA'dır. Classic 3 veri seti için sınıflandırma başarı yüzdesi GATBA 'nın başlangıcında ortalama % 94.66'tür. Algoritma 300 nesil çalıştıktan sonra başarı değeri ortalama % 97.717 yükselmiştir. Bu değer sınıflandırmada katsayı kullanılmasının kullanılmaması alternatifine göre başarıyı yaklaşık % 3.057 artırdığı anlamına gelmektedir. Bu aşamadaki KNN sınıflandırıcısının *k* parametresi 5 olarak

seçilmiştir ve ortalma değerler yöntemin 10 kez çalıştırılması sonucu hesaplanan değerlerdir.

Aynı şekilde ikinci veri seti için de bu programlar hazırlanmış ve denedikten sonra başarı yüzdeleri değerlendirilmiştir. İkinci metin belge kolleksiyonu ise metin madenciliği araştırmalarında oldukca sık kullanılan Reuters-21578 dağıtım 1.0'dır. Koleksiyon 135'in üzerinde kategoriye sahip 21578 belge içermektedir. Mevcut olan 135 konudan bazıları çok az sayıda belgede bulunduğundan çalışmada kullanılmak üzere 135 kategoriden en çok yer belgede geçen ilk 3 kategoriye (konuya) ait belgeler seçilmiştir. Seçilen bu konulara ait toplam 600 belge mevcut olup, bunlardan 300 belge eğitim verisi olarak, 300 belge ise test verisi olarak kullanılmıştır. Bu kolleksiyon içerisinde 10'dan daha az sayıda yer alan terimler çıkarılmıştır. Yani veri boyutu 600x4377 adettir.

Öncelikle mevcut veri setinin sınıflandırma başarısı *k* parametre değeri 1 olarak kullanılan KNN yöntemiyle %81 olarak bulunmuştur. Bu sınıflandırmada mevcut 4377 niteliğin tamamı kullanılmıştır. Bu verileri ilk olarak GA'ile veri setini özellik seçimine tabi tutulmuş ve 200 nesil sonra başarı yüzdesin ortalama % 45.36'dan başlayıp % 58.93'ye ulaştığı görülmüştür. Bu başarı hesaplanırken kullanılan KNN sınıflandırılmasındaki *k* parametresi 7 olarak belirlenmiştir. Daha sonra aynı verileri MATLAB'daki GA toolbox programını kullanarak sınıflandırma başarısı ortalama % 35.35'den başlamış 300 nesil sonra % 69.57 değerine ulaşmıştır. Bu başarı hesaplanırken kullanılan KNN sınıflandırıcısında *k* parametresi 1 olarak belirlenmiştir. Ayrıca ortalama değerler hesaplanırken program 10 kez çalıştırılmıştır.

En son bu veri setinde de kullanılan ikinci boyut azaltma yöntemi yine TBA dir. TBA yöntemi kullanılarak kullanılan veri setinin boyutu azaltılmıştır. Veri kümesinin % 37.35 özelliğini gösteren ilk on sütun seçilmiştir. Bu verilere göre ellimizdeki belge sayısı 600x10'a inmiştir. Veri setinin sınıflandırma başarısı GATBA 'nın başlangıcında % 94.33 değerinden başlamıştır. Algoritma 350 nesil çalışmış ve başarı değeri % 96.607 değerine yükseldiği gözlenmiştir. Bu değer sınıflandırmada katsayı kullanılmasının katsayı kullanılmaması alternatifine göre başarıyı % 2.277 artırdığı anlamına gelmektedir.

Bu çalışmada nesil sayısı 1 ile veri 1000 arasındaki sayılarla denenmiş ve en makul nesil sayısı seçilmiştir. Aynı şekilde *k* parametresi ise 1 ile 20 arasındaki sayılarla denendikten sonra en iyi başarıya ulaştıran *K* değerleri seçilmiştir.

Son olarak her iki veri setinin sonuçları değerlendirildikten sonra, GATBA programındaki kullanılan boyut azaltma yönteminin yani TBA'nin en iyi başarıya ulaştırdığı açıkca görülmektedir. Verilerin GA uygulanıktan sonra katsayılar ile çarpıldığında verilerden elde edilen sınıflandırma başarısının daha da arttığı görülmüştür.

KAYNAKLAR

Abualigah,L., Khader,A., Al-Betar,M., Alomari,O., 2017. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering, ***Expert Systems With Applications***, 84: 24–36.

Abizi, H., 2015. Veri madenciliği ve Yapay Zeka. ***http://rayanehmag.net/magazine/255/730/%D9%85%D8%AA%D9%86-%DA%A9%D8%A7%D9%88%DB%8C-%DA%86%DB%8C%D8%B3%D8%AA%D8%9F#comment-84*** Erişim Tarihi: 25.10.2017.

Baker, J.E., 1985**.** Adaptive Selection Methods for Genetic Algorithms**.** ***In Proceedings of an International Conference of Genetic Algorithms and Their Applications***. Pittsburgh, July 24-26, 101-111.

Bhatia, N. and Vandana., 2010. Survey of nearest neighbor techniques**.** ***International Journal of Computer Science and Information Security***. 8(2):302-305.

Cover, T.M. and Hart, P.E., 1967.Nearest neighbor pattern classification. ***IEEE Transactions on Information Theory***. IT13(1):21–27.

Çalışkan, S.K., ve Soğukpınar, İ., 2008**.** KxKNN**:** K-Means ve K En Yakın Komşu Yöntemleri ile Ağlarda Nüfuz Tespiti. ***2. Ağ ve Bilgi Güvenliği Sempozyumu.*** 16-18 Mayıs, Girne, 120-124.

Çilli, M., 2007. ***İnsan Hareketlerinin Modellenmesi ve Benzeşiminde Temel Bileşenler Analizi Yönteminin Kullanılması***(doktora tezi). Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, 240.

Das, A., Das S., Ghosh A., 2017. Ensemble feature selection using bi-objective genetic algorithm**.** ***Knowle dge-Base d Systems***, 123:116–127.

Deb, K., 1999. Multi-Objective Genetic Algorithms: Problem Difficulties and Construction of Test Problems. ***Evolutionary Computation. 7***(3):205-230.

Deb, K., 2001. Multiobjective Optimization Using Evolutionary Algorithms**.** ***Wiley & Sons***. England. 1-5.

Erdoğan, Y.S., 2013. ***İstanbul Teknik Üniversitesi, İnşaat mühendisliği yapılarında akıllı hesaplama teknikleri ile yapısal tanılama ve parametre tahmini***(doktora tezi, basılmamış). Fen Bilimleri Enstitüsü, İstanbul.

Escalante, H.J., García-Limón M. A., Morales-Reyes A., Graff M., Montes-y-Gómez M., Morales E. F., Martínez-Carranza J., 2015. Term-weighting learning via genetic programming for text classification. ***Knowledge-Based Systems***. 83:176–189.

Filiz, Z., 2003**.** Güvenilirlik Çözümlemesi, Temel Bileşenler ve Faktör Çözümlemesi, ***Anadolu Üniversitesi Bilim ve Teknoloji Dergisi***, 4(2): 211–222.

Ghareb, A., Bakar A., Hamdan A., 2016. Hybrid feature selection based on enhanced genetic algorithm for text categorization. ***Expert Systems With Applications,*** 49:31–47.

Gasanova, T., Sergienko R., Semenkin E., Minker W., 2014. Dimension Reduction with Coevolutionary Genetic Algorithm for Text Classification . ***In Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics***. 215-222.

Goldberg, D.E., 1989. Genetic Algorithms for Search, Optimizations and Machine Learning. ***Addison Wesley***. 1-17.

Han, J. and Kamber, M., 2006. Data mining: concepts and techniques**.** ***Morgan Kaufmann Publishers***, Burlington.

Joines, J., Houck C., 1994. On the use of non-staionary penalty functions to solve nonlinear constrained optimization problems with Gas. ***the Proceedings of the First IEEE Conference on Evolutionary Computation***.98-108.

Kahraman, H. T., 2016. A novel and powerful hybrid classifier method: developmentand testing of heuristic k-nn algorithm with fuzzydistance metric. ***Data & Knowledge Engineering,***103:44-59.

Kang, H. S., K. Nam, and S.-i. Kim 2012. ***The Decomposed K-Nearest Neighbor Algorithm for Imbalanced Text Classification***. 87–94.

Kresse, W. and Danko, D.M. 2012. ***Springer Handbook of Geographic Information***. Springer-Verlag, Berlin.

Kuruca, E., 2009. ***Gezgin satıcı problemi tabanlı bir sistemin dinamik bulanık genetik algoritmalar ile optimizasyon***(yüksek lisans tezi, basılmamış). Yıldız Teknik Üniversitesi, FBE, Endüstri Mühendisligi Anabilim Dalı Sistem Mühendisligi Programında Hazırlanan, İstanbul.

Lebibi, F., Mellah H., Drias H., 2017**.** Enhancing information source selection using a genetic algorithm and social tagging. ***International Journal of Information Management,*** 37:741–749.

Lu, Y., Liang M., Ye M ., Cao L., 2015. Improved particle swarm optimization algorithm and its applicationin text feature selection. ***Applied Soft Computing***, 35: 629–636.

Mitchell, T., 1997. ***Machine Learning***, McGraw Hill, New York.

Murata, M., Ma, Q., Uchimoto K., Ozaku H., Utiyama M., Isahara H., Japanese 2000. Probabilistic Information Retrieval Using Location and Category Information. ***Proceedings of The Fifth International Workshop on Information Retrieval with Asian Language.***

Onan, A., Korukoglu S., Bulut H., 2016. Ensemble of keyword extraction methods and classifiers in text classification. **Expert Systems With Applications**, 57: 232–247.

Pietramala, A., Policicchio V., Rullo P., Sidhu I., 2008. ***A Genetic Algorithm for Text Classification Rule Induction.***188–203.

Reeves, C.R., 1993. Modern Heuristic Techniques for Combinatorial Problems. ***John Wiley & Sons***. Inc., New York, NY.1-3.

Saraçoğlu, R., 2007. ***Bulanık Kümeleme Kullanılarak Benzer Belge Aranması*** (doktora tezi) Selçuk Üniversitesi. Fen Bilimleri Enstitüsü,.

Silva, R., Almeida T., Yamakami A., 2017. MDLText: An efficient and lightweight text classifier. ***Knowle dge-Base d Systems***, 118: 152–164.

Silahtaroğlu, G., 2013. ***Veri Madenciliği Kavram ve Algoritmaları***. Papatya Yayıncılık, İstanbul, Türkiye.

Song, W., Cheol Park S., 2009. Genetic algorithm for text clustering based on latent semantic indexing. ***Computers and Mathematics with Applications***, 57:1901-1907.

Song, W., Hua Li Ch., Cheol Park S., 2009. Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. ***Expert Systems with Applications***, 36: 9095–9104.

Sütçüler, E., 2006. ***Gerçek Zamanlı Video Görüntülerinden Yüz Bulma ve Tanıma Sistemi(***yüksek lisans tezi). Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, 90.

Şeker Ş.E., 2014. Metin Madenciliği. ***http://bilgisayarkavramlari.sadievrenseker.com/2014/06/15/metin-madenciligi-text-mining/*** Erişim Tarihi: 20.10.2017 .

Tasi, Ch., Chen Z., Ke Sh., 2014. Evolutionary instance selection for text classification. ***The Journal of Systems and Software***, 90: 104–113.

Trstenjaka, B., Mikac, S., Donko, D., 2014. KNN with TF-IDF Based Framework for Text Categorization. ***Procedia Engineering,*** 69: 1356 – 1364.

Uğuz, H., 2011. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. ***Knowledge-Based Systems***, 24: 1024–1032.

Uysal, A.K., 2016. An improved global feature selection scheme for text classification. ***Expert Systems With Applications***, 43: 82–92.

Uysal, A. K., Gunal S., 2014. Text classification using genetic algorithm oriented latent semantic features. ***Expert Systems with Applications***, 41:5938–5947.

Yalçın, Ö., 2013. ***Veri Madenciliği Yöntemleri***. Papatya Yayıncılık, İstanbul, Türkiye.

ÖZ GEÇMİŞ

Iran-Urmia'da 1992 yılında doğmuştur. İlk ve orta öğrenimini Iran-Urmia'de tamamlamıştır. 2015 yılında, Urmia Payam Noor Üniversitesi, Bilişim Mühendisliği Bölümünde Lisans eğitimini tamamlamıştır. Yüksek Lisansa, Van Yüzüncü Yıl Üniversitesi, Fen Bilimleri Enistitüsü, Elektrik-Elektronik Mühendisliği Anabilim dalında, 2016 yılında başlamıştır.

UNIVERSITY OF VAN YUZUNCU YIL
THE ISTITUTE OF NATURAL AND APPLIED SCIENCES
THESIS ORIGINALITY REPORT

Date:31/08/2018

Thesis Title: CLASSIFICATION OF TEXT DOCUMENTS USING GENETIC ALGORITHM AND K-NEAREST NEIGHBORS

The title of the mentioned thesis above having total 47 pages with cover page, introduction, main parts and conclusion,has been checked for originality by Turnitin computer program on the date of 31/08/2018 and its detected similar rate was 1% according to the following specified filtering originality report rules:

- Excluding the Cover page,
- Excluding the Thanks,
- Excluding the Contents,
- Excluding the Symbols and Abbreviations,
- Excluding the Materials and Methods
- Excluding the Bibliography,
- Excluding the Citations,
- Excluding the publications obtained from the thesis,
- Excluding the text parts less than 7 words (Limit match size to 7 words)

I read the Thesis Originality Report Guidelines of Van Yuzuncu Yil University for Obtaining and Using Similarity Rate for the thesis, and I declarethe accuracy of the information I have given above and my thesisdoes not contain any plagiarism; otherwise I accept legal responsibility for any dispute arising in situations which are likely to be detected.

Sincerely yours,

31/08/2018
Date and signature

Name and Surname: Parisa LARİBI

Student ID# :159101188

Science : Department of Electrical and Electronics Engineering

Program: .....

Statute:M. Sc.☒ Ph.D.☐

APPROVAL OF SUPERVISOR
SUITABLE
Assoc. Prof. Dr. Rıdvan SARAÇOĞLU

(Title, Name-Surname, Signature)

APPROVAL OF THE INSTITUTE
SUITABLE

(Title, Name-Surname, Signature)
Doç. Dr. Harun AYDIN
Enst. Müdür Yrd.