KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ
ANABİLİM DALI

DOKTORA TEZİ

TOPLULUK SINIFLANDIRICILARI VE ÖZELLİK SEÇME METOTLARIYLA GELİŞTİRİLEN UZAY ORMANLARI

ZEYNEP HİLAL KİLİMCİ

KOCAELİ 2018

KOCAELİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ

ANABİLİM DALI

DOKTORA TEZİ

TOPLULUK SINIFLANDIRICILARI VE ÖZELLİK SEÇME METOTLARIYLA GELİŞTİRİLEN UZAY ORMANLARI

ZEYNEP HİLAL KİLİMCİ

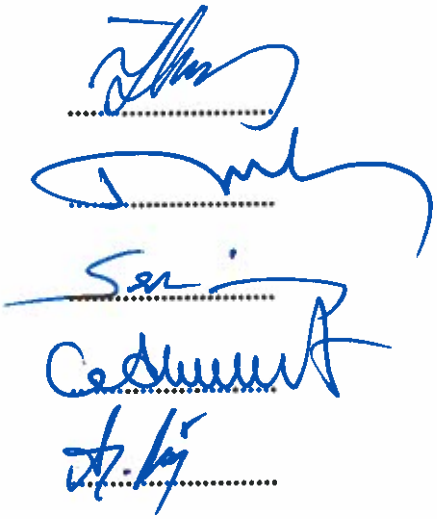
Doç.Dr. Sevinç İLHAN OMURCA
Danışman, Kocaeli Üniversitesi

Prof.Dr. Nevcihan DURU
Jüri Üyesi, Kocaeli Üniversitesi

Prof.Dr. Selim AKYOKUŞ
Jüri Üyesi, Doğuş Üniversitesi

Dr.Öğr.Üyesi Orhan AKBULUT
Jüri Üyesi, Kocaeli Üniversitesi

Dr.Öğr.Üyesi Aysun GÜRAN
Jüri Üyesi, Doğuş Üniversitesi



Tezin Savunulduğu Tarih: 18.05.2018

ÖNSÖZ VE TEŞEKKÜR

Farklı alanlardaki sınıflandırma performasına sağladığı katkıdan dolayı, topluluk algoritmaları literatürde sıkça kullanılmaya başlanmıştır. Topluluk algoritmalarıyla harmanlanarak özellik uzayında yapılan değişiklikler güncel çalışmaların ilgi odağı haline gelmiştir. Böylelikle, sınıflandırma performansının daha da iyileştirilebileceği öngörülmektedir.

Özellik uzaylarının genişletilerek sınıflandırma performasının iyileştirilmesi konusunda bana çalışma fırsatı veren değerli hocam Sevinç İlhan OMURCAYA' ya ve katkılarından dolayı değerli hocam Selim AKYOKUŞ' a teşekkür ederim. Ayrıca, hayatım boyunca benden desteğini esirgemeyen çok değerli anne ve babama, kıymetli eşime ve biricik kızıma da sonsuz minnet duygularımı sunarım.

Mayıs- 2018 Zeynep Hilal KİLİMCİ

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

α : Feromon değerinin nispi önemini belirleyen global bilgi
β : Sezgisel yerel bilgi
i : T zamanında karıncanın başlangıç noktası
j : Karıncanın gezinme sırasında seçeceği özelliği
J_k^i : Karınca k' nın ziyaret etmediği özellik kümesi
k : Karınca sayısı
n_{ij} : i. karıncanın j. özelliği seçmedeki sezgisel tercih edilebilirliği
ρ : Feromon buharlaşma katsayısı
$\tau_{ij}(t)$: (i, j) kenarındaki sanal feromon miktarı
$\Delta\tau_{ij}(t)$: Her bir karınca tarafından biriktirilen feromon miktarı

Kısaltmalar

ACO : Ant Colony Optimization (Karınca Kolonisi Optimizasyonu)
BG : Bagging (Torbalama)
BS : Boosting (Artırma)
CHI : Chi-sqaure (Ki-kare)
DT : Decision Trees (Karar Ağaçları)
GR : Gain Ratio (Kazanım Oranı)
IG : Information Gain (Bilgi Kazanımı)
k-NN : k Nearest Neighbour (k-En Yakın Komşu)
MNB : Multinomial Naïve Bayes (Çok Terimli Saf Bayes)
MVNB : Multivariate Bernoulli Naïve Bayes (Çok Değişkenli Saf Bayes)
NB : Naïve Bayes (Saf Bayes)
RF : Random Forest (Rastgele Orman)
RND : Random (Rastgele)
RS : Random Subspace (Rastgele Altuzay)
SVM : Support Vector Machine (Destek Vektör Makinesi)
WE : Word Embeddings (Kelime Yerleştirmeleri/Gömülmeleri)

TOPLULUK SINIFLANDIRICILARI VE ÖZELLİK SEÇME METOTLARIYLA GELİŞTİRİLEN UZAY ORMANLARI

ÖZET

Sınıflandırıcı toplulukların arkasındaki temel fikir, genel doğruluğu geliştirmeyi bekleyerek birden fazla sınıflandırıcı kullanmaktır. Sınıflandırıcı toplulukların, temel öğrenicilerin bireysel başarısı ve çeşitlilik olmak üzere iki faktöre bağlı olarak genel sınıflandırma performansını artırdığı bilinmektedir. Genişletilmiş uzay ormanları da sınıflandırma problemlerinde iyileştirmeler sağlamak için kullanılan ortak bir konudur. Daha zengin özellik uzayı sağlarlar ve orijinal özellik uzay tabanlı ormanlardan daha iyi performans sunarlar. Güncel literatür çalışmaların çoğu, genişletilmiş uzay orman yaklaşımı için giriş vektörleri olarak orijinal özelliklerin yanı sıra bunların çeşitli kombinasyonlarını da kullanmaktadır.

Bu amaçla tez kapsamında, genişletilmiş uzay ormanlarının homojen ve heterojen sınıflayıcı topluluklarla kombinasyonlarının sınıflandırma başarısını, bilgi kazanımı, ki-kare, karınca kolonisi optimizasyonu, derin öğrenmeye dayalı kelime göbekleri gibi özellik geliştirme yöntemleri ile incelenilmesine odaklanılmıştır. Topluluk sisteminin temel öğrenicileri, saf Bayes' in iki varyantı, destek vektör makineleri ve karar ağaçları gibi sınıflandırma algoritmalarına dayanmaktadır. Torbalama, artırma, rastgele alt uzaylar, rastgele ormanlar, çoğunluk oyu ve istifleme, veri çeşitliliğini sağlamak ve sistemin son kararını birleştirmek için bir araya getirme stratejileridir. Yaygın olarak kullanılan biyomedikal veri kümeleri, Türkçe ve İngilizce metinleri içeren veri kümeleri önerilen çalışmanın ilerlemesine katkıda bulunmak için geniş bir yelpazede gerçekleştirilen karşılaştırmalı deneylerin yürütülmesinde kullanılmıştır. Son olarak, önerilen yöntem ile genişletilmiş uzay ormanı yaklaşımı, güncel literatür çalışmaların orijinal versiyonuna ve çeşitli genişletilmiş versiyonlarına kıyasla performans ölçeklerinde dikkate değer deneysel sonuçları ortaya çıkarmaktadır.

Anahtar Kelimeler: Derin Öğrenme, Genişletilmiş Uzaylar, Metin Sınıflandırma, Sınıflandırıcı Toplulukları, Topluluk Öğrenmesi.

IMPROVED SPACE FORESTS WITH AN ENSEMBLE OF CLASSIFIERS AND FEATURE SELECTION METHODS

ABSTRACT

The basic idea behind the classifier ensembles is to use more than one classifier by expecting to improve the overall accuracy. It is known that the classifier ensembles boost the overall classification performance by depending on two factors namely, individual success of the base learners and diversity. Extended space forests are also a matter of common knowledge for ensuring improvements on classification problems. They provide richer feature space and present better performance than the original feature space based forests. Most of the contemporary studies employs original features as well as various combinations of them as input vectors for extended space forest approach.

For this purpose, we focus on to observe the classification success of the combination of extended space forests with homogeneous and heterogeneous classiifier ensembles by using feature enhancement methods such as information gain, chi-square, ant colony optimization, deep learning based word embeddings. The base learners of ensemble system are based on classification algorithms such as two variants of naïve Bayes, support vector machine, and decision tress. Bagging, boosting, random subspaces, random forests, majority voting, and stacking are the ensemble strategies to ensure the data diversity and combine the final of system. We conduct a wide range of comparative experiments on widely used biomedicine datasets, Turkish and English texts to contribute to the advancement of proposed study. Finally, extended space forest approach with our proposed technique turns out remarkable experimental results compared to the original version and various extended versions of recent state-of-art studies.

Keywords: Deep Learning, Extended Spaces, Text Classification, Classifier Ensembles, Ensemble Learning.

GİRİŞ

Günümüzde internetin günlük hayatımızda kullanımı tartışılmaz bir gerçektir. İnternette depolanan verilerin çoğunluğunun metin verileri olduğu açıktır. Metin sınıflandırması, internette depolanan metin belgelerinin miktarındaki üstel artış nedeniyle makine öğrenmesi alanında önemli birkonu halini almıştır. Dahası, belge sınıflandırması için otomatik çözümler için giderek artan bir ihtiyaç söz konusudur. Belge sınıflandırma problemleri, doğal dillerin karmaşıklığı ve belgelerin özellik uzayının çok yüksek boyutlu olmasından kaynaklı olarak makine öğrenimindeki zorlu görevler arasında kabul edilmektedir [1].

Metin sınıflandırması, günümüzde farklı uygulama alanlarındaki çok sayıda metin belgesi göz önüne alındığında her zaman önemli bir araştırma konusu olmuştur. Metin sınıflandırmasının amacı, belirli bir belgeyi, makine öğrenme teknikleri kullanılarak önceden tanımlanmış kategorilerden birine sınıflandırmaktır. Metin kategorizasyonu için, denetlenen öğrenme teknikleri genellikle bir dizi eğitim belgesinden sınıflandırıcılar oluşturmak için kullanılmaktadır. Eğitim kümesinden bir sınıflandırıcı, özellikler ve sınıf etiketleri (kategoriler) arasında bir ilişki modeli öğrenir ve oluşturur. Eğitim aşamasından sonra, sınıflandırıcı test veri kümesinden yeni bir belgenin kategorisini belirlemek için kullanılabilir. Metin kategorizasyon süreci genellikle belgelerin ayrıştırılmasını, kaldırılmasını, özelliklerin azaltılmasını, sözcüklerin kaldırılmasını, kaynakların kaldırılmasını, uygun formatlarda ağırlıklarla temsil edilmesini, sınıflandırıcıların seçimini (öğrenme algoritmalarını), eğitim ve test sürecini içermektedir.

Belgeleri temsil etmek için, kelime torbalama tekniği belge kategorizasyonunda yaygın olarak kullanılmaktadır. Torbalama modelinde, belge kümesi her satırın bir belgeyi tanımladığı ve her sütunun bir terime (kelimeye) karşılık geldiği belge-kelime matrisi olarak temsil edilmektedir. Matristeki her bir giriş, tüm doküman yığınına göre bir terimin önemini yansıtan bir ağırlık içerir. Terim frekansı ve TF-IDF gibi farklı terim yaklaşımları, her bir ağırlığı temsil etmek için kullanılan

yöntemlerdendir. Metin kategorizasyon algoritmalarının ayrıntılı bir incelemesi makalelerde [2, 3] verilmiştir. Saf Bayes, k-en yakın komşular, karar ağaçları, yapay sinir ağları ve destek vektör makineleri gibi sınıflandırma algoritmaları, tahmine dayalı performansları nedeniyle doküman sınıflandırmasında yaygın olarak kullanılmaktadır. Bu yöntemler arasında, topluluk sınıflandırıcı modeller tek bir sınıflandırıcı modelini kullansa dahi sistemin performansını artırmaktadır. Bu ilkenin arkasında yatan fikir, birden fazla sınıflandırıcıdan yararlanmaktır. Bir topluluk modeli, bir veya birden fazla makine öğrenme yöntemlerinden oluşan ve ismine temel öğreniciler denilen bir yapı üzerine inşa edilmiştir. Böylece, sınıflandırma görevinin daha sağlam ve doğru bir şekilde gerçekleştirileceği beklenir [4-8].

Topluluk yöntemlerini kullanan sistemler, aynı zamanda, çok sayıda sınıflandırıcı sistemleri, topluluk tabanlı sınıflandırıcılar, öğrenme toplulukları, uzmanların karışımı, sınıflandırıcılar topluluğu, topluluk algoritmaları veya yalnızca topluluk sistemleri olarak adlandırımaktadır [9-12]. Saf Bayes sınıflandırıcılar, karar ağaçları, destek vektör makineleri, yapay sinir ağları, k-en yakın komşuluk gibi denetimli makine öğrenme teknikleri, topluluk stratejileri için yaygın olarak kullanılmaktadır. Özellikle, karar ağacının, diğer sınıflandırma yöntemlerine kıyasla topluluk sınıflandırıcıları için literatürde daha yaygın olarak kullanıldığı görülmektedir [10, 11, 13-17]. Birden fazla karar ağacının kullanılması, sınıflandırıcı topluluklar için karar ormanlarını ortaya çıkarmaktadır. Eğitim sırasında, her bir temel sınıflandırıcı, belirli bir eğitim veri kümesinde ayrı ayrı eğitilir. Bir topluluk yaklaşımı genellikle topluluk oluşturma ve bütünleştirme (toplama, kombinasyon veya füzyon) adımlarından oluşur. Topluluk oluşturma aşamasında, eğitim veri kümesinden çeşitli temel sınıflandırıcılar kümesi oluşturulur. Entegrasyon aşamasında, eğitimli temel sınıflandırıcıların çıktıları nihai bir karar almak için entegre edilir. Topluluk yaklaşımındaki ana strateji bu nedenle birçok sınıflandırıcı üretmek ve sınıflandırıcıların çıktılarını tek tek sınıflandırıcıların performansını geliştirecek şekilde sınıflandırmaktır [4, 6-8].

Bir topluluk sisteminin başarısı, topluluğu oluşturan temel sınıflandırıcıların çeşitliliğine bağlıdır ve her temel sınıflandırıcı kendi aralarında çeşitlilik sergilemelidirler. Çeşitlilik veri çeşitliliği, parametre çeşitliliği ve yapısal çeşitlilik olarak üç yaklaşımla sağlanabilmektedir [8]. Veri çeşitliliğinde, yeniden örnekleme

teknikleriyle her bir temel sınıflandırıcı için orijinal veri kümesinden farklı eğitim verileri alt kümeleri oluşturulur. Parametre çeşitliliği yaklaşımı, farklı sınıflandırıcılar için farklı eğitim parametrelerinin kullanılmasıyla sağlanır. Örneğin, bir sinir ağı farklı katmanlar, başlangıç ağırlıkları ve öğrenme oranları ile eğitilebilir. Farklı öğrenme algoritmaları kullanılarak yapısal çeşitlilik elde edilebilir. Tüm temel sınıflandırıcılar aynı öğrenme algoritması kullanılarak oluşturuluyorsa, bu topluluk sistemine homojen denir, aksi halde heterojen olarak adlandırılmaktadır. Heterojen topluluk sistemleri, çeşitliliği gerçekleştirmek için birden fazla farklı öğrenme algoritması kullanırlar.

Metinsel veri madenciliğine olanak tanıyan bir diğer ortam ise sosyal medyadır. Sosyal medya da büyük miktarda bilgiyi analiz etmek ve birçok konuda fikirleri tespit etmek için çok popüler bir kaynak haline gelmiştir. Bilinen sosyal medya platformlarından biri olan Twitter, 100 milyona kadar aktif kullanıcının fikirlerini ifade etmesi için tercih edilen bir ortam olmuştur. Bu, Twitter' ın pazar dinamikleri için etkili olabilecek değerli bilgiler içerdiği anlamına gelir. Bu nedenle, duyarlılık analizi, kullanıcı taleplerini olumlu ve olumsuz yönler açısından anlamak için önemli bir yer tutmaktadır. Duygu analizi, geniş kapsamlı bir araştırma alanıdır ve kullanıcıların fikirlerinin metinden çıkarılması olarak özetlenebilir. Bu alandaki negatif, pozitif veya nötr gibi duygu polaritesini belirlemek için saf Bayes, destek vektör makineleri ve benzeri geleneksel makine öğrenme teknikleri kullanılmaktadır. En popüler ve en son kullanılanı, geleneksel makine öğrenimi algoritmalarına kıyasla daha yüksek sınıflandırma performansı elde eden derin öğrenme modelleridir.

Derin öğrenme ise yapay sinir ağları olarak adlandırılan ve beynin yapısı ve işlevinden esinlenilerek ortaya atılan makine öğrenmesinin bir alt alanıdır. Derin öğrenme modellerinin temel yaklaşımı, karmaşık özelliklerin minimum dış destekle eğitilmesiyle otomatik özellik çıkarımı sağlamak ve duygu analizi için derin sinir ağları aracılığıyla verilerin anlamlı sunumunu elde etmektir. Daha ayrıntılı olarak, derin öğrenme, özellik çıkarma işlemi için çok sayıda doğrusal olmayan bileşen katmanını kullanan geleneksel makine öğrenimi algoritmalarının bir parçasıdır. Çıkış, önceki katmandan ardışık olarak bir girdi olarak elde edilir. Öğrenme prosedürü, makine öğrenimi algoritmalarının eğitim aşaması gibi denetimli (örn., sınıflandırma), yarı denetimli veya denetimsiz (ör., desen analizi) olarak

gerçekleştirilebilir. Bu yapı ayrıca girdilerin çoklu düzeylerinin temsillerini öğrenir. Derin sinir ağları, derin düşünme ağları, tekrarlayan sinir ağları, konvolüsyonel sinir ağları ve sığ sinir ağları (word2vec) gibi derin öğrenme mimarileri, görüntü analizi, bilgisayarla görme, konuşma tanıma, doğal dil işleme, ses tanıma, sosyal ağ filtreleme, makine çevirisi ve biyoinformatik gibi alanlara uygulanmıştır.

Diğer sınıflandırma problemleriyle karşılaştırıldığında, metin kategorizasyon problemi, girdi uzayının yüksek boyutlu olması, belge vektörlerinin kısıtlılığı ve ilgisiz özelliklerin azlığı gibi birçok farklı özelliğe sahiptir [18]. Diğer bir taraftan literatürde yapılan çalışmalara bakıldığında metin kategorizasyonu alanında topluluk sistemleri ve genişletilmiş özellik uzayları ile derin öğrenme modellerinin kullanımı konusunda sınırlı araştırma yapıldığı gözlenmektedir. Bu çalışmada, öncelikli olarak homojen ve heterojen topluluk sınıflandırıcılarının sınıflandırma başarısı araştırılmaya çalışılıp sonrasında topluluk stratejisi yaklaşımının ve gelişmiş özellik uzaylarının konsolidasyonunun etkinliğini gözlemlemek için gelişmiş özellik uzaylarına odaklanıldı.

Bu amaçla rastgele, kazanım oranı, bilgi kazanımı, ki-kare ve karınca kolonisi optimizasyonu gibi teknikler özellik seçme metotları olarak kullanıldığında kelime gömülmelerini elde edebilmek için sığ derin öğrenme uygulamalarından biri olan word2vec yöntemi, öznitelik çıkarma yöntemi olarak kullanıldı. Torbalama, artırma, rastgele alt uzay, rastgele ormanlar, çoğunluk oyu ve istifleme ise deneylerde homojen ve heterojen topluluk stratejileri olarak kullanıldı.

Tez kapsamında kullanılan veri kümeleri, literatürde yaygın olarak kullanılan UCI makine öğrenme veri havuzundan alınan veri kümelerinden, haber ajanslarından toplanılarak elde edilen Türkçe ve İngilizce haber metinlerinden ve İngilizce Twitter metinlerinden oluşmaktadır. Önerilen çalışmanın ilerlemesine katkıda bulunmak için geniş kapsamlı ve karşılaştırmalı deneyler yaptık. Yapılan bu kapsamlı deney sonuçları, sınıflandırıcı topluluklar ile geliştirilmiş uzay ormanlarının sınıflandırma performansını, literatürde yapılan güncel çalışmalar ile karşılaştırıldığında etkili bir şekilde arttığını gösterdi.

Tezin geri kalanı şu şekilde düzenlendi: Topluluk sistemlerinin ve genişletilmiş uzayların kullanımına, literatür incelemesi kısmında değinildi. Bir sonraki bölümde, önerilen ve deneylerde kullanılan modelin detayları verildi. Daha sonraki bölümlerde ise sırasıyla deney kurulumu, deney sonuçları, sonuçlar ve önerilerden bahsedildi.

1. LİTERATÜR İNCELEMESİ

Topluluk öğrenmesi, bir dizi sınıflandırıcıdan oluşan ve sınıflandırma tahminlerini çoğunluk oyu kullanarak birleştiren yöntemlerin toplanması olarak literatürde tanımlanmış [19-20]. Önceki çalışmalar, topluluk öğrenmesinin topluluk içindeki tek sınıflandırıcılardan daha doğru ve sağlam olduğunu belirtmiş [19-29].

Yazarlar, özellik kümeleri ve sınıflandırma algoritmaları topluluğu üzerine gerçekleştirdikleri ilginç bir çalışmada [24], duygu sınıflandırma için topluluk yöntemlerinin etkinliğine odaklanmışlar. İlk olarak daha doğru bir sınıflandırma performansı elde etmek için konuşma bölümü (POS) ve kelime-ilişki (WR) tabanlı özellik kümelerinden oluşan iki tür özellik kümesi tanımlamışlar. Daha sonra, saf Bayes, maksimum entropi ve destek vektör makinelerini temel sınıflandırıcılar olarak kullanmışlar. Son olarak, sabit kombinasyon, ağırlıklı kombinasyon ve meta sınıflayıcı kombinasyonu topluluk prosedürü için kullanmışlar. Deneyler, Cornell film inceleme şirketi tarafından kullanılan beş yaygın veri kümesi üzerinde gerçekleştirilmiş ve POS tabanlı WR tabanlı topluluk olarak uygulanmış. Böylece, bireysel sınıflandırıcı ve üç topluluk yönteminin deney sonuçlarını yorumlayabilmişler. Deneysel sonuçlar, hem farklı özellik kümelerini hem de farklı sınıflandırma algoritmalarını birleştirmek için topluluk yöntemlerinin kullanımının, sınıflandırma performansını artırmak için etkili bir yöntem olduğunu göstermiş.

Topluluk öğrenimine dair bir başka önemli çalışma [25], dengesiz veri dağılımlarında destek vektör makinesi (SVM) algoritmasının başarısını artırmayı önermiş. Dengesiz dağılımı olan veriler için önyargılı karar sınırı probleminin üstesinden gelmek amacıyla tamamlayıcı bir yaklaşım benimsemişler ve SVM’ nin sınıflama başarısını iyileştirmek amacıyla topluluk tekniklerine odaklanmışlar. Bir diğer deyişle, SVM' nin dengesiz veriler üzerindeki sınıflandırma başarısını iyileştirmek için SVM topluluğu adında yeni bir SVM tekniğini önermişler. Bu çalışmada sekiz adet veri kümesi kullanılmış, bunlardan dördü UCI Makine Öğrenim Deposu'ndan toplanmış ve kalan dört veri kümesi de Klinik Değerlendirme Bilim Enstitüsü'nün (Kanada'dan

ICES) ve Ulusal Kanser Enstitüsü'nün (Amerika Birleşik Devletleri'nden NCI) klinik verileri olarak belirlenmiş. Kapsamlı deneyler, önerilen tekniklerin rekabetçi, etkili ve çeşitli veri örnekleme tekniklerinden üstün olduğunu göstermiş. Diğer bir çalışmada [26], topluluk sınıflayıcılarının çevrimdışı el yazısı karakter tanıma için etkinliğinin araştırılması önerilmiş. Homojen temel öğreniciler, heterojen temel öğreniciler, kararların hiyerarşik birleşimi, homojen temel öğreniciler ile eşleştirilen benzersiz özelliklerin kullanılması gibi dört tip farklı mimaride deneyler gerçekleştirilmiş. Deney sonuçları, topluluk algoritmaları kullanarak karakter tanıma başarısının, çevrimdışı el yazısı karakter tanıma başarısından daha iyi olduğunu göstermiş.

Yakın zamanda yapılan bir çalışmada [29], topluluk öğrenmesi tekniklerinin anahtar kelimelerle gösterilen metin belgeleri üzerindeki performansı deneysel olarak ölçülmüş. İlk olarak, anahtar kelime çıkarımı, terim sıklıklı cümle tabanlı anahtar kelime çıkarma, eşzamanlılık istatistiksel bilgi tabanlı anahtar kelime çıkarma, eksantriklik tabanlı anahtar kelime çıkarma ve veri kümesini test etmek için metin sıralaması algoritması olmak üzere farklı anahtar kelime çıkarma algoritmaları gerçekleştirmişler. Daha sonra, çeşitli öğrenme algoritmalarını (saf Bayes, destek vektör makineleri, lojistik regresyon ve rastgele ormanlar), adaboost, torbalama, dagging, rastgele altuzay, çoğunluk oyu gibi yaygın olarak kullanılan topluluk teknikleriyle kullanmışlar. Araştırmalarının sonucunda, metin belgelerinin topluluk öğrenmesiyle anahtar kelime temelli metin temsilinin, tahmine dayalı performansı arttırabileceği sonucuna varmışlar.

Makine öğrenmesi modellerinde, özellik uzayına orijinal uzayda var olmayan yeni özellikler eklenmesi fikri yeni değildir. Örneğin, çalışma [30], özelliklerin doğrusal kombinasyonlarını kullanmayı önermiş fakat sadece yeni özellikler kullanmış ve özellik uzayını yeni özellikler ile genişletmemiş. Breiman, yaklaşımının sınıflandırmada geçerli sonuçlar ortaya koyduğunu bildirmiş.

Diğer bir çalışma [10], yeni özellikleri rastgele seçerek ve orijinal özellik uzayına ekleyerek genişletilmiş özellik uzayını önermiş. Yeni özellikler üretmek için toplam, fark, bölme ve çarpma gibi yeni özellikler üretmek için birkaç özellik üretme operatörü kullanmışlar. En iyi operatörü seçmek için, temel operatörlerin ortalama

doğruluk düzeyleri, ortalama doğruluk dereceleri ve tüm operatörler için temel öğrenicilerin ortalama kappa değerleri ölçmüşler. Her üç metriğin de ilişkili olduğu durumlarda, fark operatörünün en iyisi olduğunu bildirmişler. Orijinal uzaya d sayıda yeni özellik eklemeye karar vermişler. Böylece, genişletilmiş özellik uzayını deneylerinde orijinal d sayıda özelliğin ve yeni elde edilen d sayıda özelliğin toplamı olarak ayarlamışlar. Temel öğrenicilerin sayısı 100' e ayarlanmış ve her veri kümesi ve topluluk algoritması için 10 kat çapraz doğrulama uygulanmış. 36 UCI veri kümesinden elde edilen deney sonuçlarında, genişletilmiş uzay versiyonları ve dört topluluk algoritmasının orijinal versiyonları, topluluk sınıflandırma doğrulukları açısından karşılaştırılmış. Tüm genişletilmiş versiyonların, tüm topluluk algoritmaları için orijinal versiyonlardan daha iyi performans gösterdiği gözlemlenmiş. Yazarlar, ayrıca diğer topluluk algoritmaları ile yürütme süreleri açısından da bir karşılaştırma yapmışlar. Bu karşılaştırma için eğitim ve test sürelerine ve ayrıca her bir temel öğrenicideki düğüm sayısına odaklanmışlar. Eğitim sürelerine bakıldığında genişletilmiş versiyonun daha fazla özellik kullanması nedeniyle orijinal algoritmalardan iki kat daha fazla eğitim süresi (daha az test süresi) gerektirdiği ancak daha az bir karmaşık ağaç ürettiği vurgulanmış. Toplulukların daha yüksek sınıflandırma performansı elde etmesi için genişletilmiş uzay yöntemlerinin kullanılması önerilmiş.

Genişletilmiş uzay karar ağaçları üzerine yapılan son çalışmalar [11, 16], topluluk doğruluğunu arttırmayı önermişler. Özellikleri rastgele üretmek yerine, her bir farklı aday özelliğin kazanım oranı hesaplanarak yüksek sınıflandırma kapasitesine sahip yeni özellikler üretilmiş. Bundan sonraki aşamada ise, özellik uzayını genişletmek için yeni oluşturulan özellikleri ve mevcut özellikleri bir araya getirmişler. Ardından, genişletilmiş uzay veri kümesinden bir karar ormanı oluşturulmuş. UCI Makine Öğrenim Deposu' ndan herkese açık olan erişilebilir veri kümeleri üzerinde deneyler yürütülmüş ve her veri kümesi için 10 kat çapraz doğrulama uygulanmış. Ayrıca, özellik uzayı için kullanılan farklı uzay uzantısı parametrelerinin etkisi de ölçülmüş. En iyi d ve d/2 sayıda özellikler, özelliklerin sayısı d olan aday özellik kümesinden seçilmiş. Özellik uzayının d/2 sayıda özellikle genişletilmesinin d sayıda özellikle genişletilmesinden daha uygun olduğunu gözlemlenmiş. Deney sonuçları, bu yaklaşımın hem orijinal özellik uzayının performansını hem de rastgele oluşturulmuş

genişletilmiş uzay versiyonunun başarısını geride bıraktığını göstermiş. Sonuç olarak yazarlar, uzatılmış uzay ormanların kullanımının tahmin doğruluğunu arttırmak için etkili bir yöntem olduğu sonucuna varmış, ancak rastgele seçilen özelliklerin yerine önemli özelliklerin kullanılarak geliştirilebildiğini vurgulamışlar.

Yakın zamanda yapılan bir başka çalışma [28], duygu kategorizasyon alanı için konuşma bölümlerini (POS-RS) temel alarak geliştirilmiş rastgele alt uzay yönteminin etkinliğini araştırmış. Temel öğrenicilerin topluluk öğrenmesinde çeşitliliğini oluşturmak için tek bir alt uzay kullanmak yerine, yazarlar POS-RS tekniği aracılığıyla iki önemli parametreyi, yani içerik temelli sözlük alt uzayını ve fonksiyonel sözlük alt uzayını kullanmışlar. Deneyler, önerilen tekniklerinin etkinliğini temsil etmek için on ayrı kamuya açık veri kümesi üzerinde yürütülmüş. POS-RS' in, sınıflandırma başarısını mükemmelleştirmek ve diğer metin sınıflandırma problemlerine uygulanmak için tercih edilebilir bir yöntem olduğu sonucuna varılmış.

Bu tez kapsamında, gelişmiş uzay ormanlarında özellik seçimi ve özellik çıkarma tekniklerinin etkinliği üzerine de karşılaştırmalı bir çalışma gerçekleştirildi. Bu alanda yapılmış olan çalışmalar, bizim çalışmamızın ortaya çıkmasında bize rehberlik etmiş ve motivasyon sağlamıştır.

Yang ve Pedersen çalışmalarında [31], metin kategorizasyonu için özellik seçim yöntemlerini değerlendirmişler. Bu amaçla, doküman sıklığı (DF), bilgi kazanımı (IG), karşılıklı bilgi (MI), ki-kare (CHI) ve kelime gücü (TS) gibi beş farklı özellik seçim yöntemine odaklanmışlar. Bu yöntemlerin katkısını değerlendirmek için birisi k-en yakın komşu sınıflandırıcı ve diğeri doğrusal en küçük karelere uygun haritalama olmak üzere iki farklı sınıflandırma algoritması kullanmışlar. Yazarlar, Reuters-22173 ve OHSUMED olmak üzere kamuya açık iki veri kümesi üzerinde sınıflandırma deneyleri gerçekleştirmişler. Deney sonuçları IG ve CHI' nin en iyi özellik seçim yöntemleri olduğunu ve sınıflandırma doğruluğunu artırdığını göstermiş.

Benzer bir çalışmada [32], kapsamlı deneysel çalışmalar uygulanarak özellik seçim metodlarının verimliliğini araştırılmış. Forman, ki-kare, bilgi kazancı, olasılık oranı, belge frekansı, iki-normal ayırma (BNS) gibi bilinen özellik seçim yöntemlerini

vurgulamış. Ayrıca, BNS' nin yeni bir özellik seçim metriği olduğu ve F-1 (tpr) - F-1 (fpr) olarak tanımlandığını vurgulanmış. Burada, z-skoru ya da tpr olarak da bilinen F-1, standart normal dağılımın ters kümülatif olasılık fonksiyonu, tpr ve fpr ise sırasıyla gerçek pozitif oran ve yanlış pozitif oran olarak belirtilmiş. BNS metriğinin, pozitif ve negatif sınıfın iki eşiği arasındaki ayrımı değerlendirdiği söylenmiş. Yazar, bilgisayar bilimi bildiri özetlerinden toplanan Cora veri kümesi üzerinde deneyler yapmış ve saf Bayes, C4.5, lojistik regresyon ve doğrusal kernelli SVM' yi sınıflandırma algoritmaları olarak kullanmışlar. Forman, tüm özellikleri kullanarak SVM' nin üstün performansını yenmenin zor olduğunu ve BNS' yi kullanmanın sınıflandırma performansını artırabildiğinin altını çizmiş. Sonuç olarak bu çalışmada, binormal ayırma işleminin geleneksel özellik seçme yöntemlerinden daha iyi olduğu iddia edilmiş.

Diğer bir çalışmada [33], Arapça dil makalelerinde bir özellik seçim tekniği olarak chi-square kullanarak sınıflandırma performansını analiz etmeyi önerilmiş. Bu makaleler üzerinde kelime kaldırma, filtreleme, sık kullanılan kelimeleri kaldırma gibi yöntemlerle ön işleme sürecini uygulamışlar. Ardından SVM bir sınıflandırıcı olarak kullanılmış ve sınıflandırma performansı hassasiyet, geri çağırma ve F-ölçüsü açısından değerlendirilmiş. Sonuçta, önerdikleri ki-kare tabanlı SVM sınıflandırıcısı performansının, saf Bayes ve k-NN yöntemlerinin sınıflandırma başarısından üstün olduğunu vurgulamışlar.

Başka bir çalışmada [34], özellik seçim yöntemlerinin sınıflara orantılı dağılımı olmayan veriler üzerindeki etkinliği araştırılmış. Metin kategorizasyon alanında özellik seçimi için bilgi kazanımı (IG), ki-kare (CHI), korelasyon katsayısı (CC) ve olasılık oranları (OR) teknikleri ve bunların geliştirilmiş sürümleri üzerinde yoğunlaşılmış. Yazarlar, bunlardan bazılarının (IG, CHI) iki taraflı metrikler denilen pozitif ve negatif özellikleri birleştirdiğinden ve kalanların (CC, OR) tek taraflı metriklere sahip olup sadece olumlu özellikleri seçtiğinden bahsetmişler. Özellik seçme yöntemlerinin geliştirilmiş sürümleri, standart sürümlerinin optimize edilmesiyle elde edilmiş. Özellik seçme yöntemlerinin başarısını ölçmek için, veri kümesi olarak Reuters-21578 ve sınıflandırıcı olarak saf Bayes ve lojistik regresyon seçilmiş. Yazarlar, pozitif ve negatif özelliklerin (iki taraflı metrikler) kullanımının dengesiz dağılıma sahip veriler üzerinde etkili olmadığını gözlemlemişler.

Geliştirilmiş özellik seçim tekniklerinin sınıflandırma performansını arttırmak için büyük potansiyele sahip olduğunu belirtmişler.

Karınca kolonisi optimizasyonu üzerine de bir özellik seçim tekniği olarak birçok çalışma bulunmaktadır. Bunlardan bir tanesinde [35], yüz tanıma sistemindeki karınca kolonisi optimizasyonu (ACO) tabanlı özellik seçme tekniğine odaklanılmış. Bu yaklaşımda, en kısa özellik uzunluğu ve sınıflandırma başarısı açısından en uygun özellik alt kümesi, ACO ve sezgisel bilgi kullanılarak seçilmiş. Önerilen algoritmanın katkısını göstermek için, yazarlar genetik algoritma tabanlı ve karınca kolonisi optimizasyon tabanlı özellik seçim yöntemlerini karşılaştırmışlar. Deney sonuçları, ACO tabanlı özellik seçim tekniğinin sınıflandırma başarısını artırdığını göstermiş.

ACO üzerine bir başka çalışma [36], metin kategorizasyon alanında ACO' ya dayanan yeni bir optimum özellik seçim tekniğini önermiş. Reuters veri kümesinde bilgi kazanımı, ki-kare ve genetik algoritma gibi çeşitli özellik seçim teknikleri arasında karşılaştırmalar yapılmış. Önerilen algoritmanın diğer özellik seçim yöntemleriyle kıyaslandığında üstün bir sınıflandırma performansına sahip olduğu sonucuna varılmış. Çalışma [37], ayrıca bir özellik seçimi arama prosedürü olarak ACO kullanmış. Özelliklerin yerel önemi ve alt kümelerin genel performansı önerilen algoritma tarafından ele alınmış. Yazarlar, konuşma segmentine ve doku sınıflandırma problemlerine odaklanmışlar ve ACO' nun genel sınıflandırma başarısını diğer özellik seçim yöntemiyle yani genetik algoritma (GA) ile karşılaştırmışlar. Yazarlar, ACO ile önerilen algoritmanın GA tabanlı özellik seçme tekniğinden daha iyi sonuç verdiğini bildirmişler.

Karınca kolonisi optimizasyonu, başka bir çalışmada bir özellik seçimi ve model geliştirme yöntemi olarak kullanılmış [38]. Urasil türevlerinin anti-HIV-1 aktiviteleri için kantitatif bir yapı aktivite ilişkisi modellemesi gerçekleştirilmiş. Moleküler tanımlayıcılar ve pEC50 verileri üzerinde deneyler doğrusal (çoklu doğrusal regresyon ve kısmi en küçük kareler regresyonu) ve doğrusal olmayan modeller (destek vektör makineleri regresyonu) ile gerçekleştirilmiş. Özellikle SVM regresyonu için, lineer ve nonlineer tekniklerin, doğru tahminler açısından ileri adımlı seçim kullanarak kısmi en küçük kareler regresyon temelli bir yöntemden

daha iyi olduğunu bildirmişler. Yazarlar, çalışmayı ACO tabanlı özellik seçim yönteminin MLR, PLS ve SVMR modelleriyle elde edilen önemli sonuçlar sağladığı sonucuna varmışlar.

Yukarıda bahsi geçen özellik seçme yöntemlerinin yanı sıra pek çok araştırmacı özellikle duygu analizinde daha doğru sınıflandırma modelleri sağlamak için hem özellik seçiminde hem de sınıflandırmada derin öğrenme yaklaşımına odaklanmışlar. Liao ve diğ. [39], çalışmalarında derin öğrenim modellerini kullanan Twitter verilerinin duygu analizini gerçekleştirmeyi hedeflemişler. Bu amaçla, basit bir konvolüsyonel nöral ağ modeli oluşturmuş ve SVM, saf Bayes sınıflandırıcılar gibi geleneksel öğrenme algoritmalarına kıyasla daha iyi sınıflandırma performansı sunulmuş. Kısa metinler üzerinde duygu analizi yapabilmek için karakterden cümle düzey bilgisine kadar kullanılan yeni bir derin konvolüsyonel nöral ağ, [40]' taki çalışma tarafından önerilmiş. Yaklaşımlarının, güncel çalışmaların sonuçlarından daha iyi performans gösterdiğini vurgulamış ve STS veri kümesi üzerinde %86,4 sınıflandırma doğruluğuna ulaştığını bildirmişler.

Başka bir çalışma [41], kelimelerin anlamlarını yorumlamak için anahtar kelimelerin önemini vurgulamış. Uzun kısa bellek ve kapılı tekrarlayan ünite, IMDB ve SemEval-2016 veri kümelerinde anahtar kelime sözlük kullanılarak gerçekleştirilmiş. Deney sonuçları, önerilen modelin verimliliğinin %1-2 doğruluk iyileşmesi ile doğrulandığını göstermiş. Çin mikro bloglarının duygu sınıflandırması, [42]' de geliştirilmiş tekrarlayan sinir ağı modeli kullanılarak yapılmış. Uzun süreli bağımlılığı çözmek için tekrarlayan sinir ağının gizli katmanını uzun süreli kısa süreli bellek yapısıyla değiştirilerek bir çıkış yolu bulunmuş. Sistemin sınıflandırma başarısının, geleneksel makine öğrenimi algoritması olan, %3,17 hassasiyet oranına sahip SVM' den daha iyi olduğu vurgulanmış. Duygu sınıflandırmasına ilişkin bir başka çalışmada [43], kopyalanmış tweetleri ve heterojen mikroblog duyarlılık sınıflandırması (MSC) olarak adlandırılan sosyal ilişkileri kullanarak yeni bir tekrarlayan rastgele yürüyüş ağı hedeflenmiş. Önerilen model, eğitim aşamasındaki geri yayılım yöntemini uygulayarak rastgele yürüme katmanına sahip derin sinir ağlarına dayandırılmış. Modelin başarısını göstermek için Twitter' dan bilinen ve yaygın olarak kullanılan veri kümeleri üzerinde deneyler yapılmış. Önerilen tekniğin, diğer güncel çalışmalardan daha iyi sınıflandırma performansı sergilediği

gözlemlenmiş. [44]' te etkili çeviriden bağımsız bir derin sinir ağ mimarisinden Twitter veri kümesinde çok dilli duygu analizi uygulamak için bahsedilmiş. Önerilen modelin önemli bir kısmı, sırasıyla, uzun kısa süreli bellek ve konvolüsyon ağları kullanılarak kelime ve karakter düzeyindeki yerleşimlere dayandırılmış. Karakter tabanlı mimariyi, uzun süreli kısa süreli bellek yerleştirme, konvolüsyonel gömülme, iç içe gömme donma, konvolüsyonel karakter seviyesi gömme ve geleneksel destek vektörü makinesi algoritmasını değerlendirme metrikleri olan doğruluk ve f1-skoru açısından karşılaştırmışlar. Kapsamlı deney sonuçları, önerilen tekniğin (konvolüsyon karakterli mimari), çok dilli duygu analizinde, güncel derin nöral modellere kıyasla etkili olduğunu göstermişler.

Yapılan başka bir çalışmada [45], geleneksel özellik seçim modellerinin karşılaştırılmasına ve belge düzeyi duygu sınıflandırması için derin öğrenme yaklaşımlarına odaklanılmış. Bu karşılaştırmalı çalışmada iki tip öznitelik modeli kullanılmış. Birincisinde, kelimelerin sırasını hesaba katmadan kelime frekansı, ikincisinde ise kelimelerinin gömülmesini kullanarak bağımlılık kavramı değerlendirilmiş. Lineer çekirdekli SVM sınıflandırıcı, geleneksel yaklaşımların sınıflandırma performansını göstermek için kullanılmış. Tek-gösterimli vektörler veya ince ayarlı semantik kelime göbekleri ile önerilen derin öğrenme temelli modellerin, ayarlama tekniğine gömülmeyen kelimeden daha iyi sonuçlar verdiğini bildirmişler.

Duygu sınıflandırma görevi için topluluk stratejileri ve derin öğrenme metodolojileri nin kombinasyonu üzerine sınırlı çalışma yer almaktadır. [46]' da önerilen çok katmanlı perceptron temelli topluluk modeli, metinlerinde iyimser ya da kötümser olarak finansal metinlerin duygu puanı tahmininde kullanılmış. Bu amaçla yazarlar, özellik oluşturma aşamasında yeni bir özellik vektörü oluşturarak özellik vektörünün çeşitliliğini elde etmek için konvolüsyonel sinir ağı, uzun süreli kısa süreli bellek yerleştirme, vektör ortalama ve özellik temelli dört model kullanmışlar. Birleştirme adımı uygulandıktan sonra, çok katmanlı perceptron ağı bir sınıflandırıcı olarak kullanılmış. Deneysel sonuçlar, derin öğrenme ve özellik tabanlı modellerin performansının olağanüstü sonuçlar verdiğini göstermiş. [47], duygu sınıflandırması için derin öğrenme ve topluluk tekniklerini değerlendirerek alan adaptasyonu probleminin ele alınması önerilmiş. Saf Bayes, destek vektör makinesi, oylama

perseptron, karar ağacı, lojistik regresyon, k-en yakın komşu ve rastgele orman temel öğreniciler olarak belirlenmiş. Torbalama, artırma, rastgele altuzay ve basit oylama, topluluk metotları olarak kullanılmış. Derin öğrenme kısmı, belirli bir yapay sinir ağı sınıfı olan otokodlayıcıdan oluşmuş. Yazarlar, çalışmayı güncel literatür çalışmalar ile karşılaştırıldığında önerilen yaklaşımın doğruluk sonuçlarının önemli ölçüde arttığını raporlamışlar.

Son yıllarda derin öğrenme teknikleri ile duygu analizi üzerine yapılan bir başka çalışmada [48] ise, derin öğrenme tekniklerinin başarısının, geleneksel yüzey modelleriyle birleştirilerek arttırılması önerilmiş. Bu amaçla, derin öğrenmeye dayalı kelime gömülmelerini ve doğrusal bir makine öğrenme algoritmasını topluluk sisteminin temel öğrenicisi olarak kullanan bir sınıflandırıcıya odaklanılmış. Daha sonra, temel öğreniciyi ve diğer yüzey sınıflandırıcılarını birleştirmek için topluluk stratejisi uygulanmış. Kapsamlı karşılaştırmalı deneyler, önerilen tekniklerin başarısının orijinal versiyonları F1-skoru açısından geride bıraktığını göstermiş.

2. ÖNERİLEN YÖNTEMLER

Özellik uzayını genişletmek, sınıflandırma doğruluğunu arttırmak için etkili bir yöntemdir. Özellik uzayını genişletmek için orijinal özellikleri giriş vektörleri olarak kullanmak yerine, özelliklerin çeşitli kombinasyonları üretilir ve orijinal özellik uzayıyla birleştirilir. Gelişmiş özellik uzayı oluşturmak için ana fikir, orijinal özellik uzayını genişletmektir. Şimdiye kadar geliştirilmiş özellik uzayı üzerine yapılan çalışmalar, rastgele seçilen özellikleri [10] ya da yeni aday özellikleri belirlemek için kazanç oranı [11, 16] gibi belirli bir özellik seçme yöntemiyle seçilmiş özellikleri kullanmışlar. Önceki çalışmalarda belirtildiği gibi [10-11, 16] özellik uzayının geliştirilmesi, sınıflandırma performansına önemli bir katkı sağlamış.

Bu çalışma kapsamında, yukarıda bahsedilen çalışmalardan esinlenerek topluluk sisteminin sınıflandırma başarısını genişletilmiş uzay ormanlarıyla iyileştirmeyi hedefledik. Bu amaçla, şimdiye kadar uygulanmış olan rastgele seçilen özelliklerle ve kazanım oranıyla seçilen özelliklerle genişletilen uzayların yanı sıra daha önce özellik uzayını genişletmek için uygulanmamış yöntemler olan bilgi kazanımı, ki-kare, karınca kolonisi optimizasyonu, ve kelime gömülmeleri üzerine yoğunlaşıldı.

2.1. Özellik Seçim/Çıkarım Yöntemleri

Genel olarak, özellik seçim süreci belirli bir özellik seçim yöntemine göre her bir özelliği puanlamak ve en iyi k sayıda özelliği belirlemek üzerinedir. Bu bölümde, orijinal özellik uzayıyla birleştirmek için yüksek sınıflandırma başarısına sahip en önemli özellik kümesinin oluşturulmasına çalışıldı.

Çalışmamızın ilk aşamasında, literatürdeki özellik seçme üzerine yapılan çalışmalarından esinlenilerek [49-56], rastgele özellik seçimi yöntemi, bilgi kazanımı (IG), kazanım oranı (GR) ve ki-kare (CHI) özellik seçimi yöntemleri üzerinde yoğunlaşıldı. Sonrasında, çok bilinen ve uygulanan, sezgisel optimizasyon yöntemlerden biri olan karınca kolonisi optimizasyon tekniğini özellik seçme yöntemi olarak uygulandı. Son olarak, kullanılan metin içerikli veri kümelerinden

veri kümesini en iyi ifade edebilecek anlamlı özellikler oluşturabilmek için kelime gömülmeleri, özellik çıkarım yöntemi olarak kullanıldı. Özellik seçim/çıkarım yöntemleri detaylı olarak aşağıdaki bölümlerde ele alındı.

2.1.1. Rastgele özellik seçimi yöntemi

Özellik uzayını genişletmek amacıyla veri kümesindeki özellik sayısının yarısı adedince rastgele olarak seçilen özellikleri içermektedir.

2.1.2. Bilgi kazanımı yöntemi

Bilgi kazanımı, bir veri kümesinin bir öznitelik üzerine bölünmesinden sonra entropi azalmasına dayanmaktadır. Bir karar ağacının inşası, en yüksek bilgi kazancını veren öznitelik bulmakla ilgilidir. Bir karar ağacı, bir kök düğümden yukarıdan aşağıya inşa edilmektedir. Verilerin, benzer değerlere sahip örnekler içeren alt kümelere ayrılmasını içermektedir. Karar ağacı inşasında kullanılan ID3 algoritması, bir özniteliğin homojenliğini hesaplamak için entropi kullanmaktadır. Öznitelikler, aynı sınıfa aitse (homojen) entropi 0, sınıflar arasında eşit dağılmışsa entropi 1, sınıflar arasında rastgele dağılmışsa da 0 ile 1 arasında bir değeri olmaktadır. D öğrenme kümesindeki bir özniteliği sınıflandırmak için gerekli bilgi olan D' nin entropisi:

$$\text{Bilgi}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{2.1}$$

şeklinde hesaplanmaktadır. Burada p_i, D öğrenme kümesindeki bir örneğin C_i sınıfına ait olma olasılığını ifade etmektedir. İkinci adımda, veri kümesi farklı özniteliklere bölünmektedir ve ağacın her dalı için entropi hesaplanmaktadır. D kümesi A özniteliğine göre v parçaya bölündükten sonra D' yi sınıflandırmak için gerekli olan bilgi aşağıdaki gibi formülize edilmektedir:

$$\text{Bilgi}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Bilgi}(D_j) \tag{2.2}$$

Bölünmede kullanmak adına toplam entropi elde etmek için bu entropi orantılı olarak eklenmektedir. Ortaya çıkan entropi, bölünmeden önceki entropiden çıkarılır. Elde edilen sonuç, bilgi kazanımı veya entropi azalması olarak adlandırılmaktadır:

$$\text{Kazanım}(A)=\text{Bilgi}(D)-\text{Bilgi}_A(D) \quad (2.3)$$

Böylelikle, karar düğümü olarak en büyük bilgi kazancı olan özniteliği seçilmekte, ve tüm veriler sınıflandırılıncaya kadar bu süreç tekrarlanmaktadır.

2.1.3. Kazanım oranı yöntemi

Bilgi kazanımı yöntemi, hangi özelliğin en büyük bilgi kazancı sağladığına bağlı olarak bir bölünmeyi seçmektedir. Kazanç, bitler cinsinden ölçülmektedir. Bu yöntem iyi sonuçlar vermesine rağmen, çok sayıda özniteliğe sahip değişkenlere ayırmayı kolaylaştırmaktadır. Başka bir deyişle, bilgi kazanım metodu çok çeşitli değerlere sahip özellikleri seçme eğilimindedir. Bu problemi çözmek için bilgi kazanım oranı yöntemi kullanılmaktadır. Bilgi kazanım oranı yöntemi ise, bilgi kazanımının hangi oranının bu bölme için gerçekten değerli olduğunu belirlemek üzere bir bölünmenin değerini içermektedir. En yüksek bilgi kazanım oranına sahip özellik seçilmektedir. Bölünme bilgisi aşağıdaki gibi formülize edilmektedir:

$$\text{BölünmeBilgisi}_A(D)= -\sum_{j=1}^{v}\frac{|D_j|}{|D|}\times\log_2\left(\frac{|D_j|}{|D|}\right) \quad (2.4)$$

Bölünme bilgisi elde edildikten sonra A özniteliğinin kazanım oranı aşağıdaki gibi hesaplanmaktadır:

$$\text{KazanımOranı}(A)=\text{Kazanım}(A)/\text{BölünmeBilgisi}(A) \quad (2.5)$$

2.1.4. Ki-kare özellik seçimi yöntemi

Bilgi kazanımı, sınıf tahmini için elde edilen bilgi bitlerinin sayısını, bir özelliğin ortaya çıkmasını veya çıkmamasını bilerek değerlendirirken ki-kare, özellik ve sınıf arasındaki bağımsızlık eksikliğini yorumlamaktadır [49-52].

Ki-kare testi, iki değişken arasında bulunan ilişkinin bağımsız ya da bağımlı olduğunu belirlemeye yarayan ve ayrık veriler için kullanılan bir hipotez test metodudur. Ki-kare istatistiğine dayanan özellik seçimi yöntemi iki aşamayı içermektedir. Yöntemin ilk aşamasında özelliklerin sınıflara göre ki-kare istatistikleri hesaplanmaktadır. İkinci aşamasında ise serbestlik derecesi ve belirlenen önemlilik

seviyesine göre ki-kaynaşımı prensibiyle ki-kare değerlerine bakılarak veri kümesi içerisindeki tutarsız özellikler bulunana kadar art arda özelliklerin ayrıştırılması gerçekleştirilmektedir.

Böylelikle, veri kümesi içindeki herhangi bir özellik için hesaplanan ki-kare değeri, o özelliğin sınıf içerisindeki bağımlılığını ölçmektedir. Sıfır değerine sahip bir özellik, o veri kümesi içinde bağımsız olduğunu göstermektedir. Yüksek bir ki-kare değerine sahip olan özellik ise veri kümesini daha iyi ifade eden, daha tanımlayıcı özellik anlamına gelmektedir. Ki-kare değerinin hesaplanmasında kullanılan denklemler aşağıda verilmektedir:

$$X^2=\sum_{i=1}^{2}\sum_{j=1}^{k}\frac{\left(A_{ij}-E_{ij}\right)^2}{E_{ij}} \tag{2.6}$$

Burada, k sınıf sayısını A_{ij} gözlenen frekans değerini, E_{ij} beklenen frekans değerini ifade etmektedir.

$$E_{ij}=\frac{\left(R_i\times C_j\right)}{N} \tag{2.7}$$

Burada ise R_i, i' nci aralıktaki veri sayısını, C_j j' nci sınıftaki gözlemlerin sayısını, N ise sınıflardaki toplam gözlem sayısını simgelemektedir.

2.1.5. Karınca kolonisi optimizasyonu özellik seçimi yöntemi

Karınca kolonisi optimizasyonu, çeşitli alanlarda özellik seçimi için de kullanılabilen bir optimizasyon tekniğidir. Yuvadan besin kaynağına kokulu bir madde olan ve karıncalar tarafından salgılanan feromon maddesi aracılığıyla en kısa yolların bulunması ilkesine dayanmaktadır. Bu nedenle, feromon birikimi belli bir süre boyunca en kısa yolları bulmak için temel faktördür. Salgılanan feromon yolu, daha fazla karınca ve feromon patikası tarafından kullanılır ve her izole karınca için daha önce işaretlenmiş yolu seçmeyi olasılıksal olarak zorlar. Daha az tercih edilen yollarda, feromon zamanla buharlaşır ve en kısa yol, karınca geçişlerinin daha yüksek oranı ile keşfedilir. Bu nedenle, karşılık gelen yolun seçilme olasılığını belirlemek için her karınca için bir geçiş olasılık kuralı bulunmaktadır. Bu nedenle, ACO tekniği, her seferinde optimum alt kümeye aramayı yönlendirebilen özellik

seçim süreci için caziptir [38]. Kolay uygulanabilmesi ve üstün performansından dolayı [57], özellik uzayını zenginleştirmede çalışmaları motive eden bir yöntemdir. Topluluk stratejisi için 100 temel öğrenicinin kullanılması düşünüldüğünde, her bir temel öğrenici için (her izole edilmiş karınca için) başlangıçta rastgele özellikler kullanılması beklenir. Ayrıca, feromon yoğunluğu, durgunluktan kaçınmak için her bir karınca için çizilen yol üzerinde güncellenir ve daha sonra izole edilen karıncalar farklı yollar (özellikler) seçebilirler. Bu nedenle, özellik uzayının genişletilmesi, her bir temel öğrenici için farklı özellikler ile sağlanabilmektedir. Olasılık geçiş kuralı, Denklem (2.8)' deki gibi formülüze edilir:

$$p_{ij}^{k}(t)=\begin{pmatrix} \sum_{l\in J_i^k} \frac{\left[\tau_{ij}(t)^{\alpha}\right]\left[\eta_{ij}{}^{\beta}\right]}{\left[\tau_{il}(t)^{\alpha}\right]\left[\eta_{il}{}^{\beta}\right]} & \text{if } j\in J_i^k, \\ 0 & \text{otherwise} \end{pmatrix} \tag{2.8}$$

Burada i, t zamanında karıncanın başlangıç noktası, j ise gezinme sırasında seçeceği özelliği, k karınca sayısını, η_{ij} i özelliğindeyken j özelliğini seçmedeki sezgisel tercih edilebilirliğini, J_k^i karınca k' nın ziyaret etmediği özellik kümesini, $\tau_{ij}(t)$ (i, j) kenarındaki sanal feromon miktarını göstermektedir. Ayrıca, α global bilgi sağlamakta ve feromon değerinin nispi önemini belirlemektedir, β ise sezgisel yerel bilgidir. ACO özellik seçim sürecinin ilk adımı, bir dizi k karınca üretmektir. Bu çalışmada karınca sayısı, veri kümesi içindeki özelliklerin sayısına ayarlandı. Böylece, her karınca rastgele bir özellik ile başladı ve durma göstergesi yerine getirilinceye kadar kenarları olasılıksal olarak gezdiler. Sonrasında, alt kümeler toplanıp değerlendirildi. Algoritma belirli bir sayıyı gerçekleştirdikten veya optimal bir alt kümeye ulaştıktan sonra, genel özellik seçim süreci en iyi özellik çıktısı elde edilerek sona erdi. Her iki durum da sağlanamadığında, feromonun yoğunluğunu güncellemek kaçınılmaz olduğundan yeni karıncalar üretilip özellik seçim süreci bir kez daha tekrar etti. Feromon güncellemesi, her kenarda Denklem (2.9)' da belirtilen kural tarafından gerçekleştirildi:

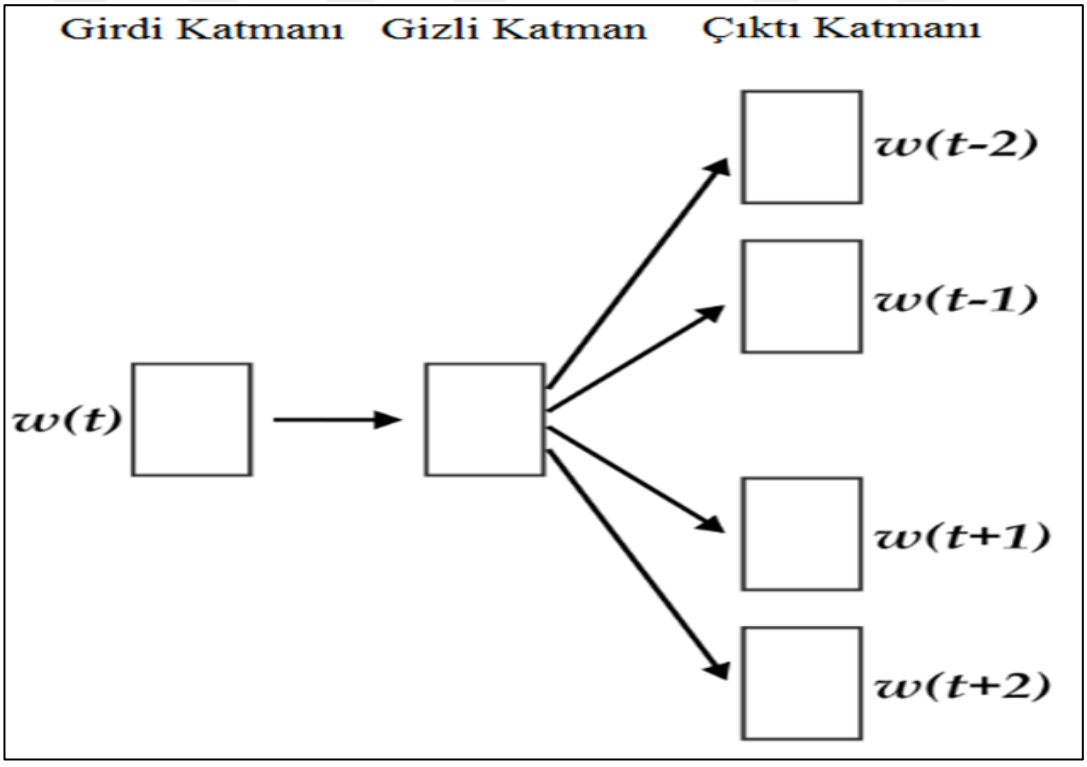
$$\tau_{ij}(t+1)=(1-\rho)\,\tau_{ij}(t)+\rho\Delta\tau_{ij}(t) \tag{2.9}$$

Burada, ρ feromon buharlaşma/güncelleme katsayısını, $\Delta\tau_{ij}(t)$ ise her bir karınca tarafından biriktirilen feromon miktarını belirtmektedir.

2.1.6. Kelime gömülmeleri özellik çıkarımı yöntemi

Bu çalışmada, ilk kez geleneksel özellik seçimi teknikleri yerine word2vec aracı kullanılarak sınıflandırıcı topluluklarla orijinal özellik uzayını genişletmek için kelime gömülmeleri/yerleşimleri kullanıldı. Böylelikle, sayısal veri kümelerinin yanında metin içerikli veri kümelerini de kullanarak önerdiğimiz yaklaşımın geçerliliği test edildi.

Word2vec, bir grup model kullanarak kelime yerleşimleri oluşturmak için kullanılan bir araçtır. Bu modeller, eğitimli sığ, iki katmanlı sinir ağları kullanarak sözcüklerin dilsel bağlamlarını yeniden yapılandırmayı önermektedirler. Büyük bir metin veri kümesi, word2vec tarafından girdi olarak değerlendirilmekte ve veri kümesindeki her benzersiz sözcükle bir vektör uzayı oluşturulmaktadır. Kelime vektörleri, vektör uzayında veri kümesindeki ortak içerikleri paylaşan sözcüklerin vektör uzayında birbirine yakın olarak konumlandırılmasıyla oluşmaktadır. Yani, word2vec, sözcükleri vektörler olarak modellemeye izin veren en yaygın kullanılan yaklaşımlardan biridir.



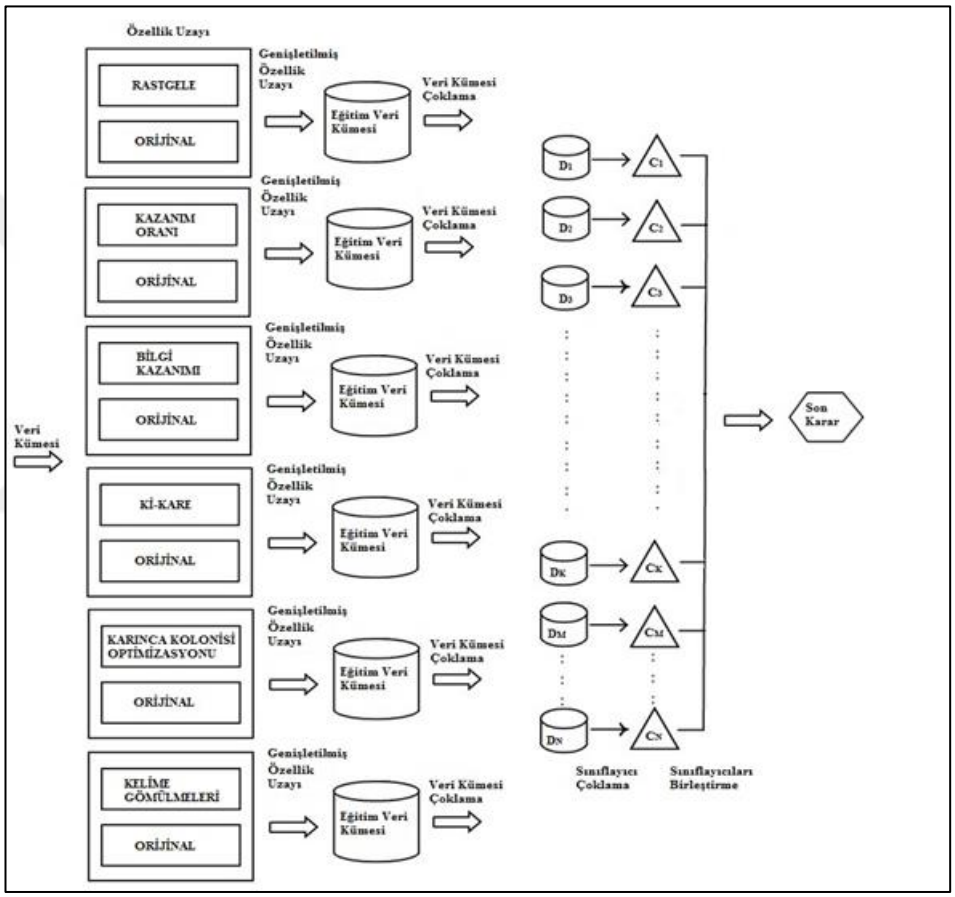
Şekil 2.1. Kelime gömülmelerini elde etmek için kullanılan sürekli atlama gramı modeli

Word2vec, kelimelerin dağıtılmış bir temsilini gerçekleştirmek için iki modelli mimariye, yani sürekli kelime torbalamaya (CBOW) ve sürekli atlama gramına (Skip Gram) dayanmaktadır. CBOW modeli, kelime-anlam yaklaşımı gibi bağlam sırasını görmezden gelerek, kelimeyi çevreleyen bağlam sözcükleri verilen bir kelimeyi tahmin eder. Öte yandan, sürekli atlama-gram modeli, verilen kelimeden o kelimeyi çevreleyen kelimeleri tahmin etmeyi amaçlamaktadır. Word2vec modeli, hiyerarşik softmax veya negatif örnekleme ile eğitilmiştir. Negatif örnekleme tekniği, örneklenmiş negatif örneklerin log-olabilirliğini en aza indirerek maksimizasyon problemini tahmin ederken, hesaplamayı azaltmak için bir Huffman ağacını kullanan hiyerarşik softmax yöntemi, bir modelin maksimize etmeyi amaçladığı koşullu log-olabilirliğe yaklaşır. Negatif örnekleme, frekansı fazla olan sözcükler için düşük boyutlu vektörler ile daha iyi sonuçlar sunarken hiyerarşik softmax, geçme sıklığı az olan kelimeler için dikkate değer sonuçlar vermektedir. Eğitim yüzdesi arttığında hiyerarşik softmax yönteminin yararlı olmadığını bildirmek önemlidir. Bu çalışmada, CBOW modeline kıyasla, nadir kelimeler için kayda değer performansından dolayı Şekil 2.1' de şematize edilen sürekli atlama modeline odaklanıldı.

2.2. Genişletilmiş Özellik Uzayı

Yukarıda belirtilen tekniklerle en anlamlı özellikler ve kelime gömülmeleri semantik olarak elde edildikten sonra, özellik uzayını bu yöntemlerle zenginleştirildi. Sonuçta olarak, temelde altı tip ve bunun türevleri olan genişletilmiş özellik uzayları elde edildi. Bunların ilk dört tanesi geleneksel özellik seçim teknikleriyle oluşturuldu. İlk genişletilmiş özellik uzayı, orijinal özelliklerin ve rastgele seçilen özelliklerin birleşiminden (orijinal+RND), ikinci genişletilmiş özellik uzayı orijinal ve bilgi kazanım tekniğiyle (orijinal + IG) toplanan özelliklerin kombinasyonundan, üçüncü genişletilmiş özellik uzayı ki-kare yöntemiyle elde edilen özelliklerin orijinal özelliklerle birleştirilmesinden (orijinal + CHI), dördüncü genişletilmiş özellik uzayı orijinal ve kazanım oranı yöntemiyle seçilen özelliklerin kombinasyonundan (original + GR), beşinci genişletilmiş özellik uzayı ise orijinal özelliklerin karınca kolonisi optimizasyonu metodu ile elde edilen özelliklerin birleşiminden (orijinal + ACO) oluştu. Sonuncusu ise, kelime gömülmeleri ve orijinal özelliklerin (orijinal + WE) birleştirilmesi yoluyla oluşturuldu. Uzay genişletme parametresi üstün performansından ötürü d/2 olarak ayarlandı. Yani, veri kümesi d sayıda özelliğe sahip

ise özellik seçim/çıkarım yöntemleriyle elde edilen d/2 sayıda özellik, orijinal özellik uzayına ilave edildi. Bu durumda, özelliklerin ilk kısmı orijinal özelliklerden oluşurken kalan kısmı ise ayrı ayrı olmak üzere rastgele, IG tabanlı, CHI tabanlı, GR tabanlı, ACO tabanlı, veya derin öğrenmeye dayalı genişletilmiş özellik uzayı için sırasıyla rastgele, IG, CHI, GR, ACO veya WE ile seçilen özelliklerden oluştu. Tez kapsamında önerilen yaklaşımımız, Şekil 2.2' de şematize edilip Şekil 2.3' te ayrıntılı olarak açıklandı.



Şekil 2.2. Önerilen yöntemlerle genişletilmiş uzay ormanları süreci

Çalışmamızda, zenginleştirilmiş özellik uzayı oluşturulduktan sonra, önerilen topluluk sistemi için temel sınıflandırıcıyı seçmek üzere çok merkezli saf Bayes (MNB), çok değişkenli saf Bayes (MVNB), destek vektör makinesi (SVM) ve rastgele orman gibi geleneksel makine öğrenme algoritmaları uygulandı. Bir sonraki adımda, çeşitliliği korumak ve sistemin nihai kararını almak için topluluk stratejisi yürütüldü.

Verilen: E= $\{x_p, y_p\}_{p=1...N}$ = [X Y]. Burada X, eğitim kümesi dahil bir N*d boyutlu matrisi, Y sınıf etiketlerini içeren bir N boyutlu sütun vektörünü, d özellik sayısını, N eğitim örneklerinin sayısını, T temel öğrenicilerin sayısını, BL_i temel öğreniciyi, EA topluluk algoritmasını, E_i ise BL_i için genişletilmiş eğitim kümesini simgelemektedir.

Başlangıç: Topluluk boyutunu T, temel öğrenici modeli BL_i ve topluluk algoritmasını EA olarak seçilmektedir.

Eğitim:
i=1' den T' ye kadar

1. Özellik seçme tekniklerini (RND, IG, CHI, GR, ACO) veya kelime gömülmelerini (WE) kullanarak yeni özellikler (EXi) oluşturulur.

   d/2 sayıda RND ile rastgele özellikler üretilir ve Ri' de saklanır veya
   d/2 sayıda IG ile önemli özellikler üretilir ve Ii' de saklanır veya
   d/2 sayıda CHI ile önemli özellikler üretilir ve Ci' de saklanır veya
   d/2 sayıda GR ile önemli özellikler üretilir ve Gi' de saklanır veya
   d/2 sayıda ACO ile önemli özellikler üretilir ve Ai' de saklanır veya
   d/2 sayıda WE ile önemli özellikler oluşturun ve Wi' de saklanır.

   j=1
   z=1' den d' ye 2. maddeye kadar
   $X_i(z)^{th}$ ve $Ri(z)^{th}$ veya $I_i(z)^{th}$ veya $C_i(z)^{th}$ veya $G_i(z)^{th}$ veya $A_i(z)^{th}$ veya $W_i(z)^{th}$ 'nin özelliklerine fark operatörü uygulanarak X matrisinin j. yeni özelliği oluşturulur.
   j=j+1
   iç döngü sonu

2. X matrisini (orijinal özellikler) sırasıyla R_i, I_i, C_i, G_i, A_i ve W_i (yeni özellikler) ile ayrı ayrı birleştirerek sırasıyla Ei= [X R_iY], Ei= [X I_iY], Ei= [X C_iY], Ei= [X G_iY], Ei= [X A_iY] ve Ei= [X W_iY] yeni eğitim kümesi (E_i) oluşturulur.
3. EA' ya göre temel öğrenici BL_i' yi E_i ile eğitilir.

dış döngü sonu

Test:
i = 1' den T' ye kadar
1. Test örneğinin özellik uzayı genişletilir.
2. Genişletilmiş test örneği BL_i ile sınıflandırılır.
döngü sonu
Temel öğrenicilerin kararları, seçilen topluluk algoritması EA' nın çoğunluk oylaması kuralıyla birleştirilir.

Şekil 2.3. Genişletilmiş uzay algoritması

2.3. Topluluk Stratejileri

Topluluk öğrenmesinde çeşitlilik, öğrenme algoritmalarının farklı belirlenmesiyle sağlanabilirken topluluk algoritmasının aynı belirlendiği durumlarda ise veri kümesinin farklı versiyonlarını kullanarak sağlanmaktadır. Bu bölümde, bu amaç doğrultusunda veri çeşitliliği sağlayan farklı yöntemler ele alınmaktadır.

2.3.1. Torbalama yöntemi

Torbalama en popüler ve en eski topluluk tabanlı algoritmalardan bir tanesidir. Farklı eğitim veri alt kümelerinin tüm eğitim veri kümesinin değiştirilmesiyle rastgele çizildiği yeniden örnekleme yoluyla çeşitlilik elde edilmektedir. Her bir veri alt kümesi, topluluk öğrenicileri grubunda farklı bir öğrenici yetiştirmek için kullanılmaktadır. Bireysel öğrenicilerin kararlarını çoğunluk oylamasıyla birleştirip nihai bir karara varılmaktadır.

2.3.2. Artırma yöntemi

Artırma algoritması, yakın zamanlardaki makine öğrenimindeki en önemli ilerlemelerden biri olarak kabul edilmektedir. Buradaki ana fikir, her bir örneğin bir ağırlıkla ilişkilendirildiği bir veri alt kümesini kullanan bir grup öğrenici oluşturmaktır. Zayıf öğreniciler eğitim verilerinin üzerine çeşitli dağılımlarda tekrar tekrar çalışır. Başlangıçta tüm örneklerin eşit ağırlığı bulunmaktadır. Her bir yinelemede, önceki sınıflandırıcıların eğitim hatalarına bağlı olarak, yanlış sınıflandırılmış örneklerin ağırlıkları güncellenir. Her bir sınıflandırıcı, eğitim veri kümesinin güncellenmiş bir dağıtımından alınan örneklerin bir alt kümesini kullanmaktadır. Her adımda, önceki sınıflandırıcılar tarafından yanlış tahmin edilen örnekler, doğru olarak tahmin edilen örneklerden daha sık seçilmektedir. Son karar, bireysel sınıflandırıcı tarafından tahmin edilen sınıfların ağırlıklı çoğunluk oyu ile elde edilmektedir. AdaBoost, AdaBoost.M1, AdaBoost.M2, AdaBoost.R, Arcing ve Real Adaboost [4, 6-8] gibi artırma algoritmasının birçok çeşidi bulunmaktadır. Deneylerimizde AdaBoost.M1 algoritması kullanıldı.

2.3.3. Rastgele altuzay yöntemi

Rastgele altuzay (RS) topluluğu torbalama işlemine benzemektedir ancak tüm örneklerin yerine veri kümesinden rastgele bir özellik kümesi seçmektedir. D özellikleri (boyutlar) olan bir veri kümesi verildiğinde, RS rastgele d' özelliklerini d'<d olmak koşuluyla seçer. Orijinal veri kümesindeki özelliklerin büyük bir kısmını kapsayacak şekilde S farklı özellik alt kümelerini almak için S kez tekrarlanır. Daha sonra S temel sınıflandırıcılar S özellik alt kümeleri ile eğitilir. Son karar, S temel sınıflandırıcılarının kararlarını bir oylama şemasıyla birleştirerek elde edilir. Bazı çalışmalarda [14, 58] özelliklerin sayısı eğitim nesnelerinin sayısından çok daha büyük olduğunda RS' nin iyi performans göstermesinin beklendiği belirtilmiş.

2.3.4. Rastgele orman yöntemi

Breiman [30] tarafından tanıtılan rastgele ormanlar, karar ağacı sınıflandırıcılarının bir koleksiyonu olduğu belirtilmiş. Rastgele ormanlar için her bir temel sınıflandırıcının bir karar ağacı olduğu torbalamanın özel bir uygulaması şeklinde tanımlanabilir. Torbalama, her bir karar ağacı için eğitim alt kümelerini seçmek için kullanılmaktadır. Rastgele ormanlarda kullanılan bölme kriteri, her bir düğümün diğer tüm özellikler arasında en iyi özellik tarafından ayrıldığı standart karar ağaçlarından farklıdır. Rastgele ormanlarda, önce rastgele bir özellik kümesi seçilerek en iyi bölünmeye, özelliklerin rastgele alt kümesiyle karar verilmektedir. Bu strateji iyi çalışmakla beraber torbalamaya ek olarak algoritmaya ekstra rastsallık da sağlar. Rastgele ormanlar, hem örnek hem de özellik uzaylarında uygulanan rastlantısallık nedeniyle ezberleme sürecine dayanıklıdır.

3. DENEY KURULUMU

Tez kapsamında gerçekleştirilen deneylerde, genişletilmiş uzay ormanlarının topluluk stratejileriyle olan başarısını ölçmek için sayısal ve metin içerikli olmak üzere iki farklı türde veri kümeleri kullanıldı. Sayısal veri kümeleri, UCI makine öğrenmesi deposundan elde edilen farklı boyut ve özellikteki 36 veri kümesinden oluşmaktadır. Metin içerikli veri kümeleri ise dört tanesi İngilizce ve dört tanesi de Türkçe olmak üzere haber sitelerinden yayınlanan haberleri ve sosyal medyada yer alan beş farklı veri kümesine ait tweetleri içermektedir. Ön işlem yapılmadığında veri kümelerinin karakteristik özellikleri Tablo 3.1, Tablo 3.2 ve Tablo 3.3' de gösterilmiştir. Tablo 3.1' de F özellik sayısını, C sınıf sayısını, S ise örnek sayısını gösterir.

Metin içerikli İngilizce veri kümelerinden olan 20News-18828, 20News-19997, ve Mininews, 20 haber grubu olarak adlandırılan veri kümesinin üç versiyonudur. 20News-19997, 20 haber grubu veri kümesinin orijinal versiyonu iken 20News-18828, aynı gönderileri tekrarlamadığından orijinale göre daha az doküman içermektedir. Ayrıca her gönderinin içerisinde yer alan gönderinin kimden geldiğini belirten "Kimden" ve gönderinin konusunun yer aldığı "Konu" başlıkları veri kümesinin içeriğinden kaldırılmıştır. 20 haber grubunun son versiyonu olan Mininews ise orijinal veri kümesinin küçük bir alt kümesi olup sınıf başına 100 dokümandan oluşan bir veri kümesidir. Bu veri kümelerinin üç versiyonu da yirmi farklı kategoriye sahiptir. İngilizce veri kümesi olarak bu veri kümelerini kullanan birçok çalışma [59-67] bulunmaktadır. Kullanılan son İngilizce veri kümesi olan WebKB4, farklı üniversitelerin bilgisayar bilimleri bölümlerinden toplanan web sayfalarını içermektedir. Öğrenci, fakülte, personel, kurs, proje, bölüm ve diğer yedi kategoriye daha sahiptir. Fakat, bazı çalışmalarda [59-60, 68] kullanılan WebKB veri kümesinin dört sınıflı sürümü kullanıldığından bu çalışmada da dört sınıflı versiyonu üzerinde deneyler gerçekleştirildi. Bu sebeple, veri kümesi WebKB4 olarak adlandırıldı.

Tablo 3.1. Sayısal veri kümelerin karakteristik özellikleri

Veri Kümesi No	Veri Kümesi	|F|	|C|	|S|
1	Abalone	10	19	4153
2	Anneal	62	4	890
3	audiology	69	5	169
4	Autos	71	5	202
5	balance-scale	4	3	625
6	breast-cancer	38	2	286
7	breast-w	9	2	699
8	col10	7	10	2019
9	Colic	60	2	368
10	credit-a	42	2	690
11	credit-g	59	2	1000
12	d159	32	2	7182
13	Diabetes	8	2	768
14	Glass	9	5	205
15	heart-statlog	13	2	270
16	hepatisis	19	2	155
17	Hypothyroid	31	3	3770
18	Ionosphere	33	2	351
19	iris	4	3	150
20	kr-vs-kp	39	2	3196
21	labor	26	2	57
22	letter	16	26	20000
23	lymph	37	2	142
24	mushroom	112	2	8124
25	primary-tumor	23	11	302
26	ringnorm	20	2	7400
27	segment	18	7	2310
28	sick	31	2	3772
29	sonar	60	2	208
30	soybean	83	18	675
31	splice	287	3	3190
32	vehicle	18	4	846
33	vote	16	2	435
34	vowel	11	11	990
35	waveform	40	3	5000
36	zoo	16	4	84

Tablo 3.2’ de yer alan diğer dört metin içerikli veri kümesi ise Türkçe’ dir. Milliyet veri kümesi, Milliyet gazetesinin 2002-2011 yılları arasındaki haberleri içermektedir. Bu veri kümesinin kategorileri kafe, dünya, ege, ekonomi, güncel, siyaset, spor, Türkiye, yaşam olmak üzere dokuz tanedir ve her bir kategoride 1000 doküman bulunmaktadır. Hürriyet veri kümesi, Hürriyet gazetesinde 2010' dan 2011' e kadar olan haberleri içermektedir. Bu veri kümesindeki kategoriler dünya, ekonomi, güncel, spor, siyaset, yaşam olmak üzere 6 tanedir ve her bir kategoride 1000

doküman bulunmaktadır. 1150haber veri kümesi ise yapılan bir çalışmadan [69] elde edilmiştir. Beş sınıfı olan (ekonomi, dergi, sağlık, siyaset, spor) 1150haber veri kümesinin her kategorisi için 230 doküman bulunmaktadır. Son Türkçe veri kümesi Aahaber [70], Türkiye Ulusal Haber Ajansı olan Anadolu Ajansı tarafından yayınlanan gazete makalelerinden oluşan bir veri kümesidir. Bu veri kümesinde sekiz kategori ve her kategoriye ait 2500 doküman bulunmaktadır. Kategoriler Türkiye, dünya, politika, ekonomi, spor, eğitim bilimi, kültür sanatı ve çevre sağlığıdır. Milliyet ve 1150haber köşe yazarlarının yazılarını içermektedir, bu yüzden diğer veri kümelerindeki haber metinlerine göre daha uzun ve resmidir. Öte yandan Hurriyet veri kümesi haber makalelerini içermektedir. Bu sebeple, daha düzensiz ve diğer veri kümelerinin dokümanlarından çok daha kısadırlar. Ön işlem yapılmadığında, veri kümelerinin açıklamaları Tablo 3.2' de sınıfların sayısı (|C|), doküman sayısı (|D|) ve kelime büyüklüğü (|V|) dahil olmak üzere verilmektedir. Bu veri kümeleri için sadece doküman sıklığı üçten az olan, sık olmayan kelimeleri filtrelendi.

Tablo 3.2. İngilizce ve Türkçe haber metinlerinin karakteristik özellikleri

Veri Kümesi	|C|	|D|	|V|
20News-18828	20	18828	50570
20News-19997	20	199997	43553
Mininews	20	2000	13943
WebKB4	4	4199	16116
1150Haber	5	1150	11040
Milliyet	9	9000	63371
Hurriyet	6	6000	18280
Aahaber	8	20000	14395

Kelimeyi köklerine ayıran algoritmalar veya sık kullanılan kelimeler listesiyle ortaya çıkabilecek sapmaları önlemek adına veri kümelerini herhangi bir ön işlemeye tabi tutulmadı. Tablo 3.3’ deki ilk iki veri kümesi, çok bilinen bir çalışmada [71] kullanılan yöntemlerle elde edildi. Son üç veri kümesi ise halka açık olup Twitter' dan 2014 yılında toplanan İngilizce metinlerden oluşmaktadır. Veri kümelerinin sınıf dağılımı ve ana teması Tablo 3.3' te özetlenmektedir. Deneyler, eğitim kümelerinin seviyelerini değiştirerek ve bir eğitim verisi olarak %5, %10, %30, %50, %80, %90 oranlarını kullanarak gerçekleştirildi. Doğruluk yüzdesi seviyeleri, karışıklığı önlemek adına “ts” eki ile kısaltıldı. Algoritmalar, her bir eğitim kümesi seviyesinde, rastgele 10 bölümün bölümlenmesiyle başlatıldı ve bu aşamada katmanlı örnekleme

kullanıldı. İstatistiksel anlamlılık testleri de birçok yerde kullanılırken öğrenci t-testinin değerlendirilmesinde farklı tekniklerin doğruluk sonuçlarının yakınlığı gözlendi. Anlamlılık düzeyi 0,05 olarak ayarlandı ve olasılık ile öğrenci t-testi' nin daha düşük olması durumunda fark, istatistiksel olarak anlamlı bulundu. Temel öğrenicilerin sayısı [10, 16] 'de belirtilen üstün performansından ötürü 100' e ayarlandı. Deneylerimizde temel öğrenici sayısını 10 ile 150 arasında değiştirirek sistemin performansını gözlemlediğimizde de temel öğrenici sayısını 100’ e sabitlenmesine karar verildi. Temel öğrenicilerin kararlarını birleştirmek için homojen topluluk öğrenmesi yapılan deneylerde tüm topluluklar için çoğunluk oylaması kullanılırken heterojen topluluk öğrenmesi yapılan deneylerde çoğunluk oylamasına ek olarak istifleme yöntemi de kullanıldı.

Tablo 3.3. Twitter veri kümelerinin karakteristik özellikleri

Veri Kümesi	Pozitif	Negatif	Toplam	Ana Tema
Sts-Gold	632	1402	2034	Genel
Sts-Test	181	177	358	Genel
Iphone6	371	161	532	Akıllı Telefon
Archeage	724	994	1718	Oyun
Hobbit	354	168	522	Film

Ayrıca, ACO özellik seçim süreci için bazı parametreleri belirtmek için önemlidir. İlk olarak, karınca sayısı, her veri kümesi için özelliklerin sayısına eşitlendi. Bu nedenle karınca sayısı veri kümesine göre değişiklik gösterdi. ACO algoritmasının belli sayıda gerçekleşmesi gerektiğinden temel öğrenicilerin sayısı ile aynı olarak 100’ e ayarlandı. Algoritma 100 kez uygulandıktan sonra feromon yoğunluğu güncellendi ve yeni bir karınca kümesi oluşturuldu ve işlem bir kez daha yinelendi. Her bir özelliğin ilk feromon yoğunluğu ilk başta 1'e ayarlandı. İki önemli bilgi olan yerel ve küresel bilgiler, karıncaların gezinimi ile ilgili olarak $\alpha=1$ ve $\beta=0,1$ şeklinde belirlendi. Feromon izi buharlaşma katsayısı, feromon yollarını güncelleyen ve 0 ile 1 arasında bulunan bir parametre olup deneylerde 0,2 değeriyle kullanıldı.

Toplulukların performansını değerlendirmek ve önerilen yöntemlerle geliştirilen sistemin performansını ölçmek amacıyla çeşitli başarı dinamikleri kullanıldı. Bu başarı dinamikleri topluluk doğruluğu (EA), temel öğrenicilerin bireysel doğruluğu (IA), temel öğrenicilerin kappa değeri (KP), F-ölçümü, eğri altındaki alan (AUC) şeklinde sıralanabilir. Kappa, bir çift yönlü çeşitlilik ölçümüdür ve iki sınıflandırıcı

çıktısı arasındaki anlaşma düzeyini ölçer [72]. Bizim çalışmamızda, sınıflandırıcılardan biri, bir temel öğrenici olarak istihdam edildi. Diğeri ise temel öğrenici olarak kullanılan sınıflandırıcı dışındaki tüm temel öğrenicilerin oy çokluğu kararını verdi. Topluluğun KP değeri, her bir temel öğrenicinin ortalama kappa değerine işaret edildi. Bu arada, KP değeri bir grubun çeşitliliği ile dolaylı olarak orantılıdır. Daha düşük KP değerleri, Kappa ölçümü ile değerlendirilen sınıflandırıcı çıktıları arasındaki anlaşma düzeyinden ötürü daha yüksek çeşitlilik gösterir.

Tüm bunlara ek olarak kelime gömülmelerinde Gensim tema modelindeki word2vec'in Python versiyonu kullanıldı. Deneylerimizde, bu modelin hiyerarşik softmax yöntemiyle eğittiği sürekli atlama modeli kullanıldı. Bu modelde, kelimeleri göstermek için 200 boyutlu bir vektör uzayı ve eğitim penceresi de 5 olarak ayarlandı.

4. DENEY SONUÇLARI

Deney sonuçları, sayısal ve metin içerikli veriler üzerinden olmak üzere ayrı ayrı elde edildi.

4.1. Sayısal Veriler Kullanılarak Elde Edilen Sonuçlar

Öncelikle sayısal veriler üzerinde rastgele özelliklerle, IG, CHI, ve ACO yöntemleriyle elde edilen özelliklerle özellik uzayını genişletip topluluk yöntemlerinin performanslarına yoğunlaşıldı. İlk aşamada, özellik uzayını IG ve CHI ile genişletildi. Orijinal versiyonlar ve geliştirilmiş uzay ormanları, toplulukların doğruluğu (EA), temel öğrenicilerin bireysel doğruluğu (IA) ve temel öğrenicilerin (KP) kappa değeri açısından karşılaştırıldı. Kısaltmalar aşağıdaki gibi kullanıldı: BG: Torbalama, RS: Rastgele Alt Uzay, RF: Rastgele Orman, X_0: X topluluk algoritması için veri kümesinin orijinal özellik uzayı, X_{IG}: X topluluk algoritması için bilgi kazanımlı geliştirilmiş uzay ormanları, X_{CHI}: X topluluk algoritması için ki-kare tabanlı geliştirilmiş uzay ormanları, Ts: Eğitim kümesi yüzdesi.

Ortalama topluluk doğruluğu, 36 veri kümesindeki eğitim kümesi yüzdeleri cinsinden analiz edildi. Tüm eğitim kümesi yüzdeleri için genel perspektiften bakıldığında, topluluk algoritmalarının geliştirilmiş sürümlerinin orijinal versiyona kıyasla üstün sınıflandırma performansına sahip olduğu gözlendi. Ayrıca, IG tabanlı geliştirilmiş uzay ormanları genel olarak 36 veri kümesinin hem orijinal versiyonunu hem de CHI tabanlı geliştirilmiş uzay ormanlarını performans açısından geride bıraktı. Ts80' deki topluluk doğruluklarını karşılaştırdığımızda, performans sırası RF_{IG}>RS_{IG}>RF_{CHI}> RS_{CHI}>BG_{IG}>BG_{CHI}>RF_{O}>RS_{O}>BG_{O} olarak elde edildi. Ts5 ve ts50 hariç, en iyi sınıflandırma performansı tüm eğitim kümesi boyutlarında RF_{IG} tarafından gerçekleştirildi. Bu nedenle, IG tabanlı gelişmiş uzay ormanlarının genellikle 36 veri kümesi için önemli ölçüde sınıflandırma performansına katkıda bulunduğu sonucuna varıldı. Ts10' dan ts80' e kadar, orijinal topluluk algoritmalarının başarı şu şekilde sıralandı: RF>RS>BG. Daha küçük eğitim kümesi boyutlarında, orijinal topluluk algoritmalarının performans sırası farklı oldu ancak

doğruluk sonuçlarının yakınlığı nedeniyle istatistiksel olarak anlamlı olduğunu iddia etmek için yeterli olmadığı sonucuna varıldı. Ayrıca, RF_{IG}' nin ts50 ve ts5 seviyeleri dışındaki tüm eğitim kümesi yüzdelerinde diğerlerinden daha iyi performans gösterdiği de gözlendi. Ts50 ve ts5 için, RS_{IG} diğer yöntemlere göre %2 daha iyi performans sergiledi. Rastgele alt uzayın bir topluluk algoritması olarak ve bir özellik uzayı geliştirme tekniği olarak bilgi kazanımının kullanılması, bahsettiğimiz eğitim kümesi seviyelerinde en yüksek doğruluk sonuçlarına ulaşmaktadır. Tablo 4.1' in sonunda ortalama doğruluk sonuçları verildi. Torbalama algoritması için IG tabanlı gelişmiş uzay ormanları, diğerlerine kıyasla ts80' de %87,4 doğruluk oranıyla en iyi sınıflandırma başarısına sahip oldu. Torbalama algoritmasının orijinal ve geliştirilmiş sürümlerini karşılaştırıldığında, performans sırası şu şekilde gözlendi: BG_{IG}> BG_{CHI}> BG_{o}. Benzer şekilde, IG tabanlı gelişmiş uzay ormanları, sırasıyla %88,0 ve %88,2 doğruluk sonuçları ile rastgele alt uzay ve rastgele orman algoritmaları için diğerlerinden daha iyi performans gösterdi. Torbalama algoritması gibi, rastgele alt uzay ve rastgele orman algoritmalarının orijinal ve geliştirilmiş sürümlerinin performans sırasında da aynı sonuçlar elde edildi. Bu nedenle, IG tabanlı geliştirilmiş uzay ormanları, ortalama doğruluk sonuçları değerlendirildiğinde her bir topluluk algoritması için sınıflandırma performansını iyileştiren en iyi teknik oldu.

Orijinal özellik uzayı için rastgele orman, %87,0 sınıflandırma başarısı ile en iyi topluluk algoritması oldu ve bunu sırasıyla %86,9 ile rastgele altuzay ve %86,2 doğrulukla torbalama yöntemi izledi. IG tabanlı gelişmiş uzay ormanları için sınıflandırma performansları, orijinale benzer bir şekilde sıralandı: RF> RS> BG ve sınıflandırma performansı, bu sonuçlar güncel literatür çalışmaların sonuçlarıyla tutarlılık gösterdi [10]. Bu sıralama aynı zamanda CHI tabanlı geliştirilmiş uzay ormanları için de geçerli olup sınıflandırma doğrulukları %87,8 (RF), %87,6 (RS), %87,2 (BG) olarak sonuçlandı. Rastgele orman ve rastgele alt uzay algoritmalarının sınıflandırma sonuçlarının birbirine yakın olduğu ancak Tablo 4.1 değerlendirildiğinde rastgele orman algoritmasının genellikle diğerlerine göre daha iyi performans gösterdiği gözlemlendi. Dolayısıyla, bir özellik uzayı geliştirme

tekniği olarak bilgi kazanımının kullanılmasının ideal olduğu sonucuna varıldı. Yukarıda bahsettiğimiz gibi, orijinal ve geliştirilmiş uzay ormanlarının tüm sürümlerini karşılaştırıldığında sınıflandırma başarı sırası şu şekilde gözlendi: RF_{IG}>RS_{IG}> RF_{CHI}> RS_{CHI}> BG_{IG}> BG_{CHI}>RF_{O}>RS_{O}>BG_{O}. Deneysel sonuçları, IG temelli geliştirilmiş özellik uzaylarının ve topluluk algoritması olarak da rastgele ormanların kombinasyonunun üstün sınıflandırma performansına sahip olduğunu gösterdi.

Daha küçük eğitim kümesi yüzdesi seviyelerinden farklı olarak, IG tabanlı rastgele orman algoritması ts10' dan ts80' e kadarki aralıkta maksimum değerlere ulaştı. IG tabanlı rastgele orman algoritması ve diğerleri arasındaki doğruluk farkı özellikle ts30 ve ts10'da %3' e kadar gözlemlendi. Rastgele alt uzay algoritmasının tüm sürümleri ikinci en iyi sınıflandırma performansına sahip oldu. Eğitim kümesi yüzdeleri arttıkça, tüm geliştirilmiş uzay ormanlarının başarısı da arttı ve bunun tersi durum da geçerli oldu. Tüm topluluk algoritmalarının orijinal versiyonları, yüksek eğitim kümesi yüzdelerinde bile en düşük sınıflandırma doğruluğuna ulaştı. Bu eğitim kümesi seviyelerinde, topluluk algoritmalarının orijinal versiyonlarının seçiminin, sınıflandırma problemleri için iyi bir tercih olmayacağı kanısına varıldı. Bunun yanında, performans artışları ile özellik sayısı ve veri kümelerinin sınıfları arasında açık bir ilişki olmadığı gözlenmektedir. Bu durum, özellikle düşük sınıflandırma performansına sahip olan abalone veri kümesinde gözlenmektedir. Bu veri kümesinin örnek sayısına yakın örenk sayısı olan veri kümelerinde çok daha iyi sınıflandırma performansı elde edilmektedir. Dolayısıyla, genişletilmiş alan ormanlarının, veri kümesi özelliklerinden bağımsız olarak daha iyi bir performansa sahiptir.

Tablo 4.1. Topluluk algoritmalarının geliştirilmiş ve orijinal versiyonlarının ts80' de sınıflandırılması

Veri Kümesi Numarası	BG_0	BG_{IG}	BG_{CHI}	RS_0	RS_{IG}	RS_{CHI}	RF_0	RF_{IG}	RF_{CHI}
1	27,3	29,1	28,5	27,2	28,1	27,5	28,1	28,4	28,3
2	99,1	99,8	99,5	99,2	99,3	98,8	99,6	99,7	99,5
3	89,3	91,5	89,3	87,4	92,1	90,2	87,2	88,9	87,2
4	73,1	75,9	75,2	72,6	74,1	73,8	72,1	75,8	75,2
5	86,1	98,2	96,4	87,0	94,5	92,1	88,1	99,3	97,6
6	73,4	74,6	74,1	75,3	75,3	75,3	75,4	74,1	73,5
7	97,3	98,1	97,5	97,8	97,7	97,7	97,8	97,8	97,8
8	81,6	81,9	80,8	81,7	81,6	81,5	81,7	81,7	81,7
9	85,3	86,7	86,4	85,9	88,1	87,5	84,2	87,3	86,7
10	88,1	89,1	88,8	89,2	89,2	88,7	88,6	88,4	87,9
11	77,9	79,2	78,8	77,6	79,2	78,6	78,1	79,5	78,7
12	99,1	99,8	99,7	99,0	99,6	99,3	99,9	99,2	99,5
13	76,9	77,3	76,6	76,2	77,4	76,5	77,1	78,4	77,9
14	74,1	75,8	75,4	75,1	77,9	77,2	74,3	73,5	73,0
15	82,2	82,7	82,2	83,6	83,4	83,0	84,2	83,6	83,0
16	82,9	86,4	85,7	85,7	87,9	87,5	86,0	87,2	86,9
17	99,6	99,7	99,6	97,7	99,9	99,9	99,7	99,8	99,8
18	93,9	95,4	94,9	94,9	95,8	95,1	94,0	95,7	95,2
19	97,1	97,0	96,8	96,5	97,6	97,3	96,9	96,4	96,8
20	99,1	99,3	99,2	98,4	99,5	99,0	98,8	99,3	99,2
21	90,7	90,4	89,0	93,9	93,2	92,7	92,2	96,7	96,2
22	93,7	97,8	97,0	96,0	97,2	96,9	96,1	97,4	96,8
23	85,7	87,4	87,1	86,8	86,8	86,7	86,2	88,3	87,9
24	98,7	99,0	98,5	99,2	99,0	98,7	99,7	99,3	99,1
25	51,3	51,0	50,6	51,8	51,9	51,4	51,7	53,8	53,3
26	95,7	97,2	96,4	97,8	97,8	97,7	96,4	97,5	97,0
27	97,3	97,0	97,6	97,6	98,0	98,1	98,2	97,8	98,3
28	99,1	89,7	99,0	98,3	99,5	99,2	98,7	98,1	98,8
29	79,4	79,5	79,1	80,8	82,3	81,9	81,7	82,9	82,5
30	93,2	92,7	92,2	93,5	93,1	93,2	92,5	93,6	93,4
31	96,0	96,5	96,2	97,1	96,9	96,7	96,5	97,8	97,4
32	76,0	80,6	80,2	76,4	79,8	79,2	77,3	80,8	80,1
33	97,1	98,0	98,4	97,3	98,2	98,1	97,8	98,5	98,4
34	83,4	87,7	87,1	88,1	89,9	91,0	88,5	90,4	90,2
35	86,1	87,8	87,6	87,5	88,7	88,3	88,2	88,9	88,6
36	96,5	97,5	97,1	99,1	99,0	99,0	99,7	99,3	99,2
ortalama	86,2	**87,4**	87,2	86,9	**88,0**	87,6	87,0	**88,2**	87,8

Tablo 4.2 yorumlandığında elde edilen sonuçlar şu şekilde özetlenebilir: BG_{IG}, 36 veri kümesinden 31' inde BG_O' dan daha yüksek doğruluğa sahip olup 31 sonuçtan 13' ünde anlamlı bir kazanım elde etti. RS_{IG}, 36 veri kümesinden 28' nde BG_{CHI}' den daha yüksek doğruluğa sahip olup 28 sonuçtan 8' inde anlamlı bir kazanım elde etti. RF_O, 36 veri kümesinden 23' ünde RS_O' dan daha yüksek doğruluğa sahip olup 23 sonuçtan 4'ünde anlamlı bir kazanım elde etti. RF_{CHI}, 36 veri kümesinden 24' ünde RS_{CHI}' dan daha yüksek doğruluğa sahip olup 24 sonuçtan 6' sında anlamlı bir kazanım elde etti.

Tablo 4.2. Algoritma çiftleri arasındaki karşılaştırma: "kazanım (anlamlı kazanım)/kayıp (anlamlı kayıp)" satır ve sütunlar

Yöntem	BG_{IG}	BG_{CHI}	BG_0	RS_{IG}	RS_{CHI}	RS_0	RF_{IG}	RF_{CHI}	RF_0
BG_{IG}	0/0	33(6)/3(0)	31(13)/5(2)	11(3)/25(8)	18(5)/18(4)	22(7)/14(3)	12(2)/24(6)	19(3)/17(6)	24(5)/12(4)
BG_{CHI}	3(0)/33(6)	0/0	27(7)/9(1)	8(2)/28(8)	13(1)/23(5)	19(3)/17(3)	7(1)/29(9)	13(2)/23(7)	15(5)/21(3)
BG_0	5(2)/31(13)	9(1)/27(7)	0/0	2(0)/34(16)	5(0)/31(13)	9(1)/27(8)	4(0)/32(18)	6(1)/30(14)	7(1)/29(10)
RS_{IG}	25(8)/11(3)	28(8)/8(2)	34(16)/2(0)	0/0	32(3)/4(0)	26(10)/10(0)	14(2)/22(4)	17(5)/19(3)	26(10)/10(0)
RS_{CHI}	18(4)/18(5)	23(5)/13(1)	31(13)/5(0)	4(0)/32(3)	0/0	21(6)/15(2)	13(2)/23(6)	12(3)/24(6)	24(8)/12(0)
RS_0	14(3)/22(7)	17(3)/19(3)	27(8)/9(1)	10(0)/26(10)	15(2)/21(6)	0/0	8(2)/28(17)	9(3)/27(8)	13(2)/23(4)
RF_{IG}	24(6)/12(2)	29(9)/7(1)	32(18)/4(0)	22(4)/14(2)	23(6)/13(2)	28(17)/8(2)	0/0	31(5)/5(0)	24(11)/12(1)
RF_{CHI}	17(6)/19(3)	23(7)/13(2)	30(14)/6(1)	19(3)/17(5)	24(6)/12(3)	27(8)/9(3)	5(0)/31(5)	0/0	25(9)/11(2)
RF_0	12(4)/24(5)	21(3)/15(5)	29(10)/7(1)	10(0)/26(10)	12(0)/24(8)	23(4)/13(2)	12(1)/24(11)	11(2)/25(9)	0/0

Tablo 4.3' te, geliştirilmiş uzay ormanları, orijinallerine kıyasla üstün performans sergiledi. Kazanım/kayıp numarası sırası, daha önce bahsedilen topluluk hassasiyeti (EA) sonuçları ile uyumlu çıktı: RF_{IG}> RS_{IG}> RF_{CHI}> RS_{CHI}> BG_{IG}> BG_{CHI}> RF_O> RS_O> BG_O. Orijinal topluluk algoritmalarının performansı da literatürdeki sonuçlarla uyumlu oldu: RF_O> RS_O> BG_O. Bireysel doğruluk (IA) ve temel öğrenicilerin çeşitliliği, topluluk başarısı için önemli parametrelerdir. Tek tek temel öğrenicilerin doğruluğu karşılaştırıldığında, başarı sırası şöyle gözlendi: BG_{IG}> BG_O> BG_{CHI}> RF_{IG}> RF_{CHI}> RS_{IG}> RS_{CHI}> RF_O> RS_O. Çeşitlilik ölçümü (1-KP) şu şekilde sıralandı: RS_O> RF_O> RF_{CHI}> RS_{CHI}> RF_{IG}> RS_{IG}> BG_{CHI}> BG_{IG}> BG_O. Bu sonuçlar, topluluk algoritmasının performansının, temel öğrenicilerin hem bireysel doğruluğuna hem de çeşitliliğine dayandığını göstermektedir. Dahası, beklendiği gibi ters orantılı sonuçlar elde edilmiştir.

Tablo 4.3' te, geliştirilmiş uzay ormanları, orijinallerine kıyasla üstün performans sergiledi. Kazanım/kayıp numarası sırası, daha önce bahsedilen topluluk hassasiyeti (EA) sonuçları ile uyumlu çıktı: RF_{IG}> RS_{IG}> RF_{CHI}> RS_{CHI}> BG_{IG}> BG_{CHI}> RF_{O}> RS_{O}> BG_{O}. Orijinal topluluk algoritmalarının performansı da literatürdeki sonuçlarla uyumlu oldu: RF_{O}> RS_{O}> BG_{O}. Bireysel doğruluk (IA) ve temel öğrenicilerin çeşitliliği, topluluk başarısı için önemli parametrelerdir. Tek tek temel öğrenicilerin doğruluğu karşılaştırıldığında, başarı sırası şöyle gözlendi: BG_{IG}> BG_{O}> BG_{CHI}> RF_{IG}> RF_{CHI}> RS_{IG}> RS_{CHI}> RF_{O}> RS_{O}. Çeşitlilik ölçümü (1-KP) şu şekilde sıralandı: RS_{O}> RF_{O}> RF_{CHI}> RS_{CHI}> RF_{IG}> RS_{IG}> BG_{CHI}> BG_{IG}> BG_{O}. Bu sonuçlar, topluluk algoritmasının performansının, temel öğrenicilerin hem bireysel doğruluğuna hem de çeşitliliğine dayandığını göstermektedir. Dahası, beklendiği gibi ters orantılı sonuçlar elde edilmiştir.

Önerilen yaklaşımımızın etkinliğini doğrulamak için, deney sonuçlarının karşılaştırılması güncel bir çalışma ile [10] değerlendirildi. Referans edien bu çalışmada da, bizim çalışmamızda olduğu gibi UCI veri deposundan 36 veri kümesi [73] ortak olarak kullanmıştır. Tablo 4.4’ te X_{RND}, X' in topluluk algoritması olduğu [10]' da öne sürülen rastgele seçilmiş özellikleri ekleyerek geliştirilmiş uzay ormanlarını belirtmektedir.

Tablo 4.3. Algoritmaların orijinal ve geliştirilmiş uzay versiyonlarının ts80' deki başarı dinamikleri: kazanım/kayıpsayıları, ortalama EA, IA doğrulukları ve KP değeri

Yöntem	Anlamlı Kazanım-Anlamlı Kayıp	EA ortalama doğruluk	IA ortalama doğruluk	KP
BG_{IG}	11	87,47	80,93	73,42
BG_{CHI}	-21	87,23	80,57	73,00
BG_{0}	-93	86,24	80,75	74,55
RS_{IG}	50	88,08	78,96	69,57
RS_{CHI}	18	87,68	78,57	68,63
RS_{0}	-33	86,96	75,73	61,48
RF_{IG}	66	88,24	79,94	68,92
RF_{CHI}	29	87,89	79,21	68,24
RF_{0}	-27	87,07	77,36	64,57

Ortalama doğruluk sonuçları göz önünde bulundurulduğunda, rastgele orman algoritmasının tüm geliştirilmiş uzay ormanları arasında en iyi topluluk algoritması olduğu gözlendi. Böylece, rastgele oluşturulan gelişmiş özellik uzayının sınıflandırma başarısının, deney sonuçları açısından yaklaşımlarımızla tutarlı olduğu açıktır. Topluluk algoritmalarının sınıflandırma performansı, tüm gelişmiş teknikler için şu şekilde gözlendi: RF> RS> BG. Ayrıca, tez kapsamında önerilen yaklaşımlarımız, topluluk sisteminin rastgele genişletilmiş uzay ormanlarıyla karşılaştırıldığında sistemin sınıflandırma başarısını artırdığı görüldü. Diğer bir deyişle, kapsamlı deney sonuçları, önerilen geliştirilmiş uzay ormanı modellerinin rastgele geliştirilmiş uzay ormanlarından önemli ölçüde daha iyi performans gösterebileceğini kanıtladı. Ortalama doğruluk sonuçlarından gözlendiği üzere X_{IG}, X_{RND}' ye kıyasla yaklaşık %1 iyileşme sağladı. Ayrıca, deneylerimizde X_{CHI} iddialı olup X_{RND} ile orantılı olarak rekabetçi bir performans sergiledi. Geliştirilmiş uzayların sınıflandırma performansı göz önüne alındığında, sıralama şu şekilde gerçekleşti: X_{IG}> X_{CHI}> X_{RND}.

Sayısal veriler üzerinde gerçekleştirilen deneylerimizin ikinci kısmı ise özellik uzayının ACO tabanlı özelliklerle geliştirilmesidir. Çalışmamızın bu kısmında topluluk doğruluğu (EA) değerlendirme ölçütü olarak kullanıldı.

Tablo 4.4. Güncel literatür çalışmayla ts50' de önerilen yöntemlerimizin karşılaştırılması

Yöntem	Ortalama doğruluk
BG_{RND} [10]	85,3
BG_{IG}	**85,8**
BG_{CHI}	**85,4**
RS_{RND} [10]	85,8
RS_{IG}	**86,7**
RS_{CHI}	**86,0**
RF_{RND} [10]	85,9
RF_{IG}	**86,9**
RF_{CHI}	**86,4**

Kısaltmalar, topluluk algoritmaları ve özellik seçim teknikleri için önceki deneyden farklı olarak şu şekilde kullanıldı: X_{RD}: X topluluk algoritması için rastgele seçilen özellikleri ekleyerek genişletilmiş uzay ormanı, X_{ACO+RD}: X topluluk algoritması için rastgele seçilmiş ve ACO tabanlı özellikler eklenerek genişletilmiş uzay ormanları.

Tablo 4.5. Topluluk algoritmalarının genişletilmiş ve orijinal versiyonlarının ts80'de sınıflandırma doğrulukları

Veri K. #	BG_0	BG_{RD}	BG_{ACO+RD}	RS_0	RS_{RD}	RS_{ACO+RD}	RF_0	RF_{RD}	RF_{ACO+RD}	Veri K. #	BG_0	BG_{RD}	BG_{ACO+RD}	RS_0	RS_{RD}	RS_{ACO+RD}	RF_0	RF_{RD}	RF_{ACO+RD}
1	27,3	28,4	29,2	27,2	27,2	27,5	28,1	27,8	27,8	19	97,1	96,8	97,0	96,5	97,3	97,5	96,9	96,8	96,5
2	99,1	99,3	99,5	99,2	99,0	99,4	99,6	99,6	99,6	20	99,1	99,3	99,5	98,4	99,5	99,0	98,8	99,9	99,9
3	89,3	89,5	90,4	87,4	90,6	92,1	87,2	87,3	87,2	21	90,7	89,2	90,4	93,9	92,4	92,8	92,2	96,1	98,3
4	73,1	75,0	76,5	72,6	74,7	75,9	72,1	75,3	76,2	22	93,7	97,1	98,5	96,0	97,0	97,5	96,1	96,8	97,1
5	86,1	97,2	98,7	87,0	93,2	95,2	88,1	98,6	99,1	23	85,7	87,8	88,3	86,8	86,6	86,4	86,2	85,8	86,0
6	73,4	74,1	74,9	75,3	75,1	75,0	75,4	73,1	74,4	24	98,7	98,9	98,7	99,2	99,7	99,9	99,7	99,9	99,9
7	97,3	97,5	97,5	97,8	97,3	97,1	97,8	97,9	97,5	25	51,3	50,7	51,6	51,8	51,5	51,0	51,7	53,6	55,4
8	81,6	81,0	81,4	81,7	81,2	81,4	81,7	81,5	81,7	26	95,7	96,5	97,6	97,8	97,6	97,5	96,4	97,1	97,8
9	85,3	86,5	87,9	85,9	87,2	87,8	84,2	86,8	89,5	27	97,3	97,7	97,7	97,6	98,2	98,7	98,2	98,0	98,5
10	88,1	88,8	89,1	89,2	89,0	89,2	88,6	87,7	88,3	28	99,1	99,0	99,9	98,3	99,2	99,6	98,7	98,6	98,3
11	77,9	78,8	79,3	77,6	78,7	79,3	78,1	78,6	79,0	29	79,4	79,1	79,5	80,8	80,7	80,5	81,7	82,4	83,7
12	99,1	99,7	99,7	99,0	99,6	99,9	99,9	99,9	99,9	30	93,2	92,8	93,0	93,5	93,2	93,0	92,5	93,4	94,6
13	76,9	76,6	76,3	76,2	76,9	77,1	77,1	78,0	78,7	31	96,0	99,9	99,9	97,1	96,5	96,8	96,5	97,6	97,6
14	74,1	75,8	76,2	75,1	77,3	77,9	74,3	74,0	74,5	32	76,0	77,2	77,7	76,4	79,8	80,9	77,3	80,2	84,1
15	82,2	82,3	82,1	83,6	83,2	83,0	84,2	82,1	83,7	33	97,1	98,0	98,6	97,3	98,2	98,6	97,8	98,2	98,4
16	82,9	85,7	86,7	85,7	87,4	88,0	86,0	86,9	87,4	34	83,4	87,6	89,5	88,1	91,0	92,5	88,5	90,4	91,3
17	99,6	99,5	99,4	97,7	99,9	99,9	99,7	99,7	99,5	35	86,1	87,7	88,2	87,5	88,4	88,9	88,2	88,7	88,0
18	93,9	94,9	95,6	94,9	95,1	95,4	94,0	94,3	94,0	36	96,5	97,1	97,8	99,1	99,0	99,0	99,7	99,7	99,5
								Ortalama doğruluk		86,2	87,3	**87,9**	86,9	87,7	**88,1**	87,0	87,8	**88,4**

Sınıflandırma başarısı, Tablo 4.5' te görüldüğü gibi, topluluk doğrulukları açısından, ts80'de RF_{ACO+RD}> RS_{ACO+RD}> BG_{ACO+RD}> RF_{RD}> RS_{RD}> BG_{RD}> RF_O> RS_O> BG_O olarak sıralandı. Ts30 dışında, en iyi sınıflandırma performansı tüm eğitim kümesi boyutlarında RF_{ACO+RD} tarafından gerçekleştirildi. Bu nedenle, ACO' ya dayanan geliştirilmiş uzay ormanlarının ve rastgele özelliklerin 36 veri kümesi için önemli ölçüde sınıflandırma performansına katkıda bulunduğu iddia edilebilir. Ts30 dışında, topluluk algoritmalarının orijinal sürümlerinin başarı sırası RF> RS> BG oldu. Ts30 gibi daha küçük eğitim kümesi seviyelerinde, topluluk algoritmalarının orijinal versiyonlarının performans sırası farklı olup doğruluk sonuçlarının yakınlığı nedeniyle istatistiksel olarak kayda değer bir iddia talep etmek için bu deneylerde de yeterli olmadığı gözlemlendi. Ayrıca, RF_{ACO+RD}, ts30 seviyesi dışındaki tüm eğitim kümesi yüzdelerinde diğerlerinden daha iyi performans gösterdi. Ts30'da, RS_{ACO+RD} rekabetçi olup diğer teknikleri %1 ile geride bıraktı. Topluluk algoritması olarak rastgele alt uzayın ve özellik seçim tekniği olarak ACO tabanlı ve rastgele seçilmiş

özelliklerin kombinasyonu, bu eğitim kümesi seviyesinde en yüksek doğruluk oranlarını sağladı.

Tablo 4.5' in son kısmında, ortalama doğruluk sonuçları sunuldu. Torbalama algoritması için ortalama doğruluk sonuçları, rastgele seçilmiş ve ACO tabanlı özellikler ekleyerek genişletilmiş uzay ormanlarının diğerlerine kıyasla ts80' de %87,9 doğruluk değeriyle kazanan olduğunu gösterdi. Torbalama algoritmasının orijinal ve genişletilmiş sürümlerinin başarı sırası aşağıdaki şekilde elde edildi: BG_{ACO+RD}> BG_{RD}> BG_{O}. Benzer şekilde, ACO+RD ile genişletilmiş versiyon, sırasıyla %88,1 ve %88,4 doğruluk sonuçları ile rastgele alt uzay ve rastgele orman algoritmaları için diğerlerinden daha iyi performans gösterdi. Torbalama algoritması gibi, rastgele alt uzay ve rastgele orman performans sıralaması aynı oldu. Bu nedenle sonuçlarımız, önerilen yöntem tabanlı uzatılmış uzay ormanı (ACO+RD), her topluluk algoritmasının sınıflandırma doğruluğunu ortalama doğruluk sonuçları açısından geliştiren en iyi model olduğunu kanıtladı.

Orijinal uzay ormanları için rastgele orman %87,0 sınıflandırma başarısı ile en iyi topluluk algoritması olup bunu sırasıyla %86,9 ile rastgele altuzay ve %86,2 doğrulukla torbalama izledi. ACO+RD tabanlı genişletilmiş uzay ormanlarının sınıflandırma performansı, RF> RS> BG gibi bir sıralamada sıralanıp sınıflandırma performansı güncel literatür çalışmalarıyla tutarlılık gösterdi [10]. Bu sıralama, rastgele seçilmiş özelliklere sahip genişletilmiş uzay ormanları için de geçerli olup sınıflandırma doğrulukları açısından %87,8 (RF), %87,7 (RS), %87,3 (BG) farklılık gösterdi. Rastgele orman ve rastgele alt uzay algoritmalarının sınıflandırma sonuçlarının birbirine yakın olduğu ancak rastgele orman algoritmasının genellikle Tablo 4.5' i değerlendirerek ts80' de diğerlerini geçtiği gözlendi. Bir özellik seçim tekniği olarak rastgele özelliklerin ve ACO temelli özelliklerin modifiye edilerek seçimi ve orijinal özellik uzayıyla bu özelliklerin konsolidasyonu özellik uzayını zenginleştirmek için en iyi performans sunan yöntem olduğu görüldü. Yukarıda belirttiğimiz gibi, uzay ormanlarının tüm sürümlerini topluluk algoritmalarına göre karşılaştırıldığında, sınıflandırma başarı sırası şu şekilde gözlendi: RF_{ACO+RD}>RS_{ACO+RD}>BG_{ACO+RD}>RF_{RD}>RS_{RD}>BG_{RD}>RF_{O}>RS_{O}> BG_{O}.

Deney sonuçları, rastgele ve ACO tabanlı özelliklerin modifiye edilmesiyle oluşturulan özellik uzayının, topluluk algoritması olan rastgele orman ile kombinasyonunun dikkate değer bir sınıflandırma performansına sahip olduğunu sergiledi. Daha küçük eğitim kümesi yüzdesi düzeylerinden farklı olarak, ACO+RD ile rastgele orman algoritması ts80' de maksimum değere ulaştı. Rastgele alt uzay algoritmasının tüm sürümleri, rastgele orman algoritmasını takip eden en iyi ikinci sınıflandırma performanslarına sahip oldu. Eğitim kümesi yüzdeleri arttıkça, tüm geliştirilmiş uzay ormanı sürümlerinin başarıları da arttı. Tüm topluluk algoritmalarının orijinal versiyonları, yüksek eğitim seti yüzdelerinde dahi en düşük sınıflandırma doğruluğu sundu. Tablo 4.6, algoritma çiftleri arasında yapılan karşılaşmaları göstermektedir. Bu sonuçlara göre, BG_{ACO+RD}, 36 veri kümesi içinden 28’ nde BG_O' dan daha yüksek doğruluğa sahip olup bunların 15’ i anlamlı çıktı. RS_{ACO+RD}, 36 veri kümesi içinden 21’i, BG_{ACO+RD}' den daha yüksek doğruluğa sahip olup ve bunların 11’ i anlamlı çıktı. RF_O, 36 veri kümesi içinden 22’ si RS_O' dan daha yüksek doğruluğa sahip olup bunların 4’ü anlamlı çıktı. RF_{RD}, RS_{RD}'den, 36 veri kümesi içinden 19 veri kümesinden daha yüksek doğruluğa sahip olup bunların 5’i anlamlı çıktı.

Tablo 4.6. Algoritma çiftleri arasındaki karşılaştırma: “kazanım (anlamlı kazanım)/kayıp (anlamlı kayıp)” satır ve sütunlar

Yöntem	BG_{ACO+RD}	BG_{RD}	BG_O	RS_{ACO+RD}	RS_{RD}	RS_O	RF_{ACO+RD}	RF_{RD}	RF_O
BG_{ACO+RD}	0/0	28(6)/8(0)	28(15)/8(0)	15(5)/21(11)	17(7)/19(6)	24(12)/12(3)	16(8)/20(9)	18(11)/18(6)	21(9)/15(5)
BG_{RD}	8(0)/28(6)	0/0	27(9)/9(1)	6(2)/30(12)	11(3)/25(8)	21(7)/15(5)	11(4)/25(14)	13(6)/23(9)	17(6)/19(5)
BG_O	8(0)/28(15)	9(1)/27(9)	0/0	5(0)/31(14)	6(0)/30(15)	10(2)/26(10)	4(1)/32(17)	8(1)/28(14)	6(2)/30(10)
RS_{ACO+RD}	21(11)/15(5)	30(12)/6(2)	31(14)/5(0)	0/0	25(3)/11(0)	23(12)/13(1)	17(4)/19(8)	21(7)/15(4)	25(10)/11(1)
RS_{RD}	19(6)/17(7)	25(8)/11(3)	30(15)/6(0)	11(0)/25(3)	0/0	20(9)/16(1)	10(2)/26(9)	17(4)/19(5)	22(8)/14(0)
RS_O	12(3)/24(12)	15(5)/21(7)	26(10)/10(2)	13(1)/23(12)	16(1)/20(9)	0/0	10(0)/26(14)	10(3)/26(8)	14(2)/22(4)
RF_{ACO+RD}	20(9)/16(8)	25(14)/11(4)	32(17)/4(1)	19(8)/17(4)	26(9)/10(2)	26(14)/10(0)	0/0	24(9)/12(0)	20(12)/16(1)
RF_{RD}	18(6)/18(11)	23(9)/13(6)	28(14)/8(1)	15(4)/21(7)	19(5)/17(4)	26(8)/10(3)	12(0)/24(9)	0/0	23(7)/13(2)
RF_O	15(5)/21(9)	19(5)/17(6)	30(10)/6(2)	11(1)/25(10)	14(0)/22(8)	22(4)/14(2)	16(1)/20(12)	13(2)/23(7)	0/0

4.2. İngilizce ve Türkçe Metinler Kullanılarak Elde Edilen Sonuçlar

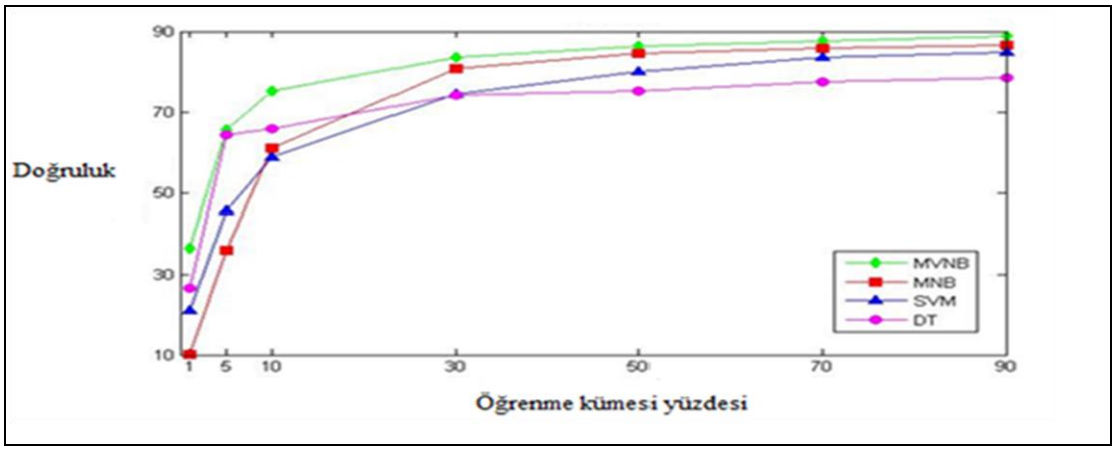
Deneylerin metin içerikli veriler üzerinde gerçekleştirilen kısmında ise ilk aşamada özellik uzayını genişletmeden topluluk algoritmalarının sınıflandırma başarılarını homojen ve heterojen olarak ayrı ayrı ölçülmesi hedeflendi. Bunları elde ettikten sonra, özellik uzayımızı rastgele, GR, IG, ACO, WE gibi özellik seçim/çıkarım teknikleriyle genişletip sınıflandırma performansı iyileştirilmeye çalışıldı.

İlk olarak, topluluk öğreniminde en iyi temel sınıflandırıcıyı belirlemek adına çeşitli makine öğrenme tekniklerinin sınıflandırma başarıları analiz edildi. Bunun için kullanılan sınıflandırıcılar, MVNB, MNB, SVM ve DT sınıflandırıcıları oldu.

Tablo 4.7. Bireysel sınıflandırıcıların doğrulukları

Veri Kümesi	MVNB	MNB	SVM	DT
20News-18828	88,52±0,42	86,26±0,36	84,18±0,79	79,82±0,72
WebKB4	84,10±1,12	85,64±1,17	89,85±0,96	77,45±1,07
Hurriyet	81,16±0,96	79,78±0,78	76,58±1,31	76,92±0,56
Aahaber	82,70±1,07	82,80±0,93	79,62±1,12	79,13±1,43
ortalama	**84,12±0,89**	83,62±0,81	82,56±1,05	78,33±0,95

Tablo 4.7, daha önce bahsedilen bireysel sınıflandırıcıların değerlendirme sonuçlarını göstermektedir: Bireysel sınıflandırıcıların sınıflandırma sonuçları, Tablo 4.7' de ts80 için elde edildi. MVNB' nin daha rekabetçi olduğu ve diğer sınıflandırma tekniklerini geçtiği açıkça görüldü. Böylelikle MVNB, topluluk sınıflandırıcılarını uygulamak için temel bir öğrenci olarak tercih edildi. Sonuçlara göre, bireysel sınıflandırıcıların sınıflandırma başarı sırasını şu şekilde gözlendi: MVNB> MNB> SVM> DT. Temel sınıflayıcıların farklı eğitim kümesi yüzdelerindeki başarısını da araştırmak amacıyla 20News-18828 veri kümesi Şekil 4.1’ de analiz edildi. MVNB ve diğerleri arasındaki doğruluk değerleri genellikle küçük eğitim kümesi seviyelerinde daha yakınken, MVNB' nin sınıflama başarısı, özellikle ts30' dan ts90' a kadar olan eğitim kümesi yüzdelerinde daha yüksek ve daha belirgin oldu. Böylece, MVNB’ nin, topluluk sistemini homojen olarak oluşturmak için seçilebilecek en iyi temel öğrenici olduğu sonucuna varıldı.



Şekil 4.1. 20News-18828 veri kümesi için temel sınıflandırıcıların eğitim kümesi yüzdelerine göre doğrulukları

Bir temel öğrenici belirlendikten sonraki aşamada eğitim veri kümelerini çeşitlendirmek amacıyla çeşitli topluluk algoritmaları (torbalama, artırma, rastgele alt uzay, rastgele orman) kullanıldı. Tablo 4.8, çok değişkenli Bernoulli saf Bayes modelini, ts80 için bir temel sınıflandırıcı olarak kullanan dört homojen topluluk yönteminin sınıflandırma doğruluğu sonuçlarını göstermektedir. Bir topluluk algoritması olarak torbalama, 20News-18828 ve WebKB4 veri kümeleri için sınıflandırma performansını artırmak için etkili sonuçlar vermediği saptandı.

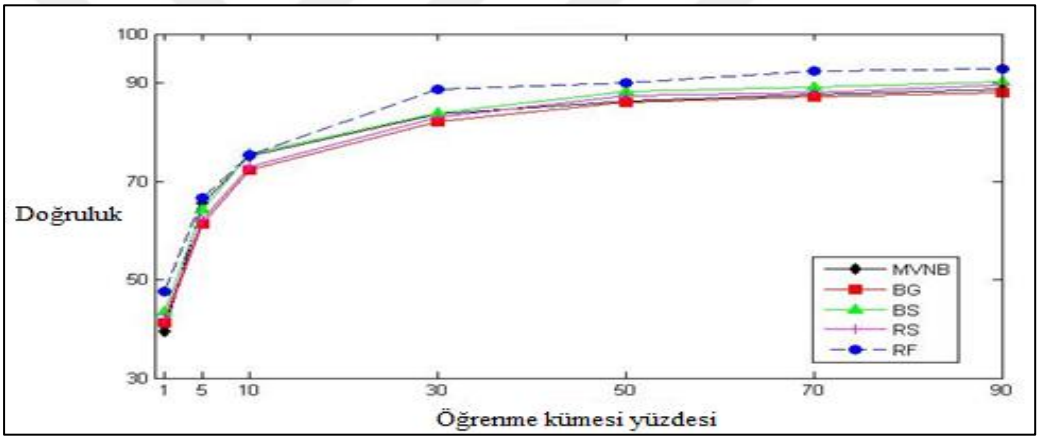
Tablo 4.8. Topluluk algoritmalarının sınıflandırma doğrulukları

Veri Kümesi	BG	BS	RS	RF
20News-18828	87,63±0,97	90,27±0,84	88,63±0,77	92,02±0,73
WebKB4	84,75±0,86	86,24±0,95	85,32±0,83	88,75±1,19
Hurriyet	82,24±1,02	81,77±0,79	83,62±0,85	84,13±0,96
Aahaber	83,16±0,88	85,91±1,05	84,69±0,96	87,35±1,37
ortalama	84,45±0,93	86,05±0,91	85,57±0,85	**88,06±1,06**

Torbalama yöntemi, 20News-18828 için yaklaşık %1 doğruluk sonuçlarını düşürdü ve ts80' de WebKB4 için MVNB sınıflandırıcısının başarısını çok az artırdı. Hürriyet ve Aahaber veri kümelerinde yaklaşık %1'lik bir sınıflandırma artışı sağladığından, Türkçe veri kümeleri için de önemli bir performans artışı elde edilemedi. Rastgele altuzay, sınıflandırma başarılarını iyileştirmek için torbalama ile benzer performans gösterdi. İngilizce veri kümelerinde anlamlı bir iyileştirme sağlamamakla beraber Türkçe veri kümelerinde %2' ye kadar doğrulukları artırdı. Bu nedenle, torbalama ve rastgele alt uzay algoritmalarının seçiminin çoğu durumda İngilizce metinlerde çok

ta anlamlı olmadığı gözlemlendi. Tercih edilecekse de, bu seçimin sınıflandırma başarısını arttırmak için özellikle Türkçe gibi sondan eklemeli dillerden yana kullanılmasının daha belirgin sonuçlar vereceği görüldü.

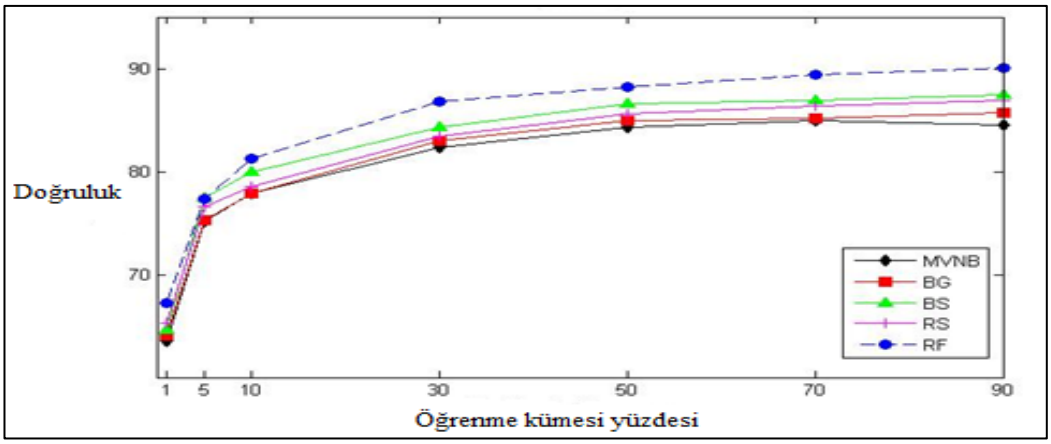
Diğer taraftan, artırma ve rastgele ormanlar, torbalama ve rastgele alt uzay algoritmalarına kıyasla daha belirgin sınıflandırma performansı gösterdi. Artırma, tek bir sınıflandırıcının performansına kıyasla yaklaşık %2-3 oranında iyileşme sağlarken sınıflandırma başarısı, torbalama ve rastgele altuzaya göre %1 ila %3 arasında değişti. Hürriyet veri kümesinde, artırma algoritması çok hafif bir iyileşme sağladı. Genel olarak, tüm veri kümeleri için topluluk tekniklerini belirlerken torbalama ve rastgele alt uzay yerine artırma algoritmasına öncelik verilebileceğini gözlemledik.



Şekil 4.2. 20News-18828 veri kümesi için topluluk algoritmalarının performans karşılaştırması

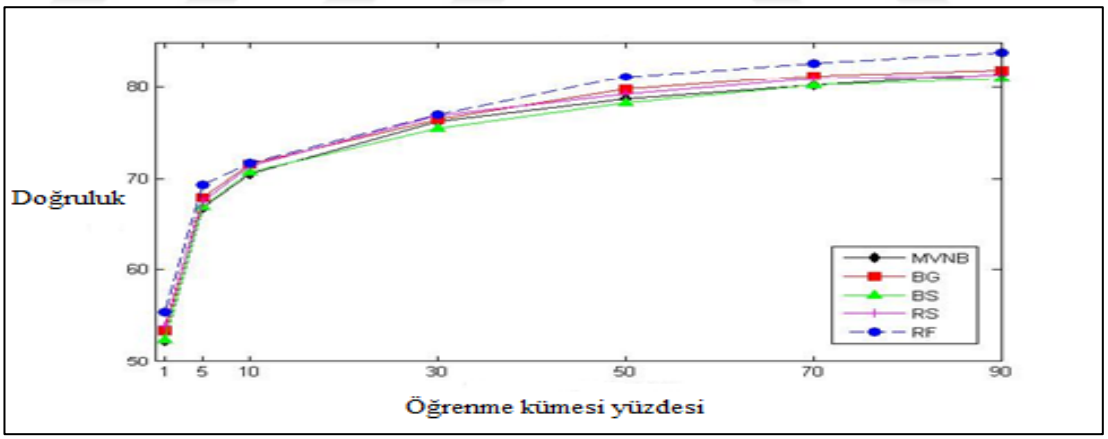
Şekil 4.2.2, 4.2.3, 4.2.4, 4.2.5, farklı eğitim kümesi yüzdelerinde tüm veri kümeleri için topluluk algoritmalarının ve temel öğrenicilerin performans karşılaştırmalarını göstermektedir. Şekil 4.2.2 ve 4.2.3' te, rastgele orman algoritmasının 20News-18828 ve WebKB4 ve temel sınıflandırıcı MVNB için tüm eğitim kümesi yüzdelerinde diğerlerini aştığı açıkça görüldü.

MVNB, her iki veri kümesi için torbalama algoritması ile benzer sınıflandırma başarı oranını gösterdi. Doğruluk değerleri, daha küçük eğitim kümesi yüzdeleriyle birbirine yakın olsa da, aralarındaki fark daha yüksek eğitim kümesi seviyelerinde arttı.

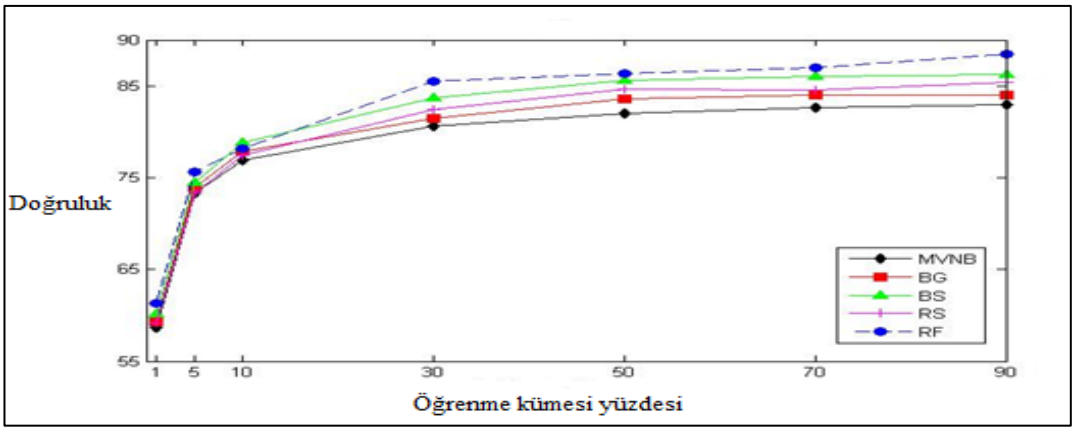


Şekil 4.3. WebKB4 veri kümesi için topluluk algoritmalarının performans karşılaştırması

Daha küçük eğitim kümesi boyutlarında, rastgele ormanın baskınlığı, Hurriyet için özellikle ts1' den ts10' a ve Aahaber için ts1'den ts30'a kadar gözlemlendi. Geriye kalan eğitim kümesi yüzdelerinde, rastgele ormanın başarı örüntüsü diğerlerine benzerlik gösterdi. Ayrıca, sınıflandırmaların başarı sıraları, veri kümelerine bağlı olarak değişmekle birlikte en iyi sınıflandırma performasının rastgele orman topluluk algoritmasına en kötü sınıflandırma performansının ise temel sınıflandırıcı MVNB' ye ait olduğu görüldü.



Şekil 4.4. Hürriyet veri kümesi için topluluk algoritmalarının performans karşılaştırması



Şekil 4.5. Aahaber veri kümesi için topluluk algoritmalarının performans karşılaştırması

4.3.Türkçe Metinler Kullanılarak Elde Edilen Sonuçlar

Diğer bir deneysel çalışmamızda, heterojen toplulukların sınıflandırma performansını sadece Türkçe veri kümeleri üzerinde araştırıldı. İlk önce, bireysel sınıflandırıcıların sınıflandırma performanslarını birbirleriyle karşılaştırmak için analiz edildi. Çok değişkenli Bernoulli saf Bayes (MVNB) ve çok terimli saf Bayes (MNB), destek vektör makineleri (SVM) ve rastgele orman (RF) öğrenme algoritmaları sınıflandırma için kullanıldı.

Tablo 4.9. Bireysel sınıflandırıcıların doğrulukları

Veri Kümesi	MVNB	MNB	SVM	RF
1150haber	93,74±1,35	94,00±1,64	89,65±2,62	94,32±1,02
Milliyet	84,64±0,95	81,48±1,03	89,41±0,53	90,18±0,87
Hurriyet	81,16±0,96	79,78±0,78	76,58±1,31	84,13±0,96
Aahaber	82,70±1,07	82,80±0,93	79,62±1,12	87,35±1,37
ortalama	85,56±1,08	84,52±1,09	83,82±1,39	**88,99±1,06**

Tablo 4.9, dört veri kümesinde daha önce açıklanan bireysel sınıflandırıcıların değerlendirme sonuçlarını ortalama doğruluk ve standart sapma açısından göstermektedir ve bu sonuçlar tutma yönteminin 10 tekrarı uygulanarak elde edildi. RF' nin daha rekabetçi ve diğer sınıflandırma tekniklerinden daha iyi performans gösterdiği görüldü. Bireysel sınıflandırıcıların sınıflandırma başarı sırası şu şekilde elde edildi: RF> MVNB> MNB> SVM. Heterojen topluluk sistemini kurmak için, MVNB, MNB, SVM ve RF algoritmalarını temel sınıflayıcılar olarak kullanıldı.

Temel sınıflandırıcıların her birinin kararları, çoğunluk oylama (MV) ve yığınlama (STCK) entegrasyon yöntemleri ile birleştirildi. Tablo 4.10' da, tek sınıflandırıcıların her birinin sınıflandırılma doğrulukları, çoğunluk oyuyla (Heter-MV) birleştirilen heterojen toplulukla ve istifleme ile birleştirilen heterojen toplulukla (Heter-STCK) karşılaştırıldı. Tablo 4.10' da, Heter-MV ve Heter-STCK topluluklarının tek bir sınıflandırıcıdan daha iyi doğruluklar ürettiği gözlendi. Heter-STCK topluluğu, tüm temel sınıflandırıcılar ve Heter-MV ile kıyaslandığında üstün sınıflandırma performansını her veri kümesi için gösterdi. Heter-STCK, çoğunluk oylama modeline göre maksimum %2 iyileştirme sağladı ve bu sınıflandırma başarısı, farklı veri kümeleri için tek sınıflandırıcılar ile karşılaştırıldığında %2 ile %13 arasında değişti.

En düşük ve en yüksek doğruluk artışları, veri kümelerine göre değişmekle beraber %3-8 1150haber için, %4-13 Milliyet için, %1-9 Hürriyet için ve Aahaber için maksimum %8 artış sergiledi.

Tablo 4.10. Bireysel sınıflandırıcıların ve heterojen topluluk algoritmalarının ts80' de sınıflandırma doğrulukları

Yöntem	1150haber	Milliyet	Hurriyet	Aahaber
MVNB	93,74±1,35	84,64±0,95	81,16±0,96	82,70±1,07
MNB	94,00±1,64	81,48±1,03	79,78±0,78	82,80±0,93
SVM	89,65±2,62	89,41±0,53	76,58±1,31	79,62±1,12
RF	94,32±1,02	90,18±0,87	84,13±0,96	87,35±1,37
Heter-MV	**95,71±0,83**	**92,27±0,76**	**84,08±0,51**	**87,06±0,97**
Heter-Stck	**97,16±1,23**	**94,57±2,06**	**85,44±0,92**	**87,73±1,17**

Özellikle Milliyet veri kümesinde %13' lük artış, topluluk stratejilerinin kullanımı açısından sınıflandırma performansına önemli katkı sağladı. Bu deneyin bir başka gözlemi Heter-MV ve Heter-STCK heterojen toplulukların performans sonuçlarının random forest algoritmasından daha iyi olması oldu. RF aslında sınıflandırıcı olarak karar ağacı koleksiyonu kullanan homojen bir topluluk sistemidir.

Ayrıca, her bir topluluk sınıflandırıcısının performans sonuçları, dört veri kümesinde de daha ayrıntılı olarak araştırıldı. Doğruluk, F-ölçümü ve AUC (Eğri Altındaki Alan) deneylerde değerlendirme kriteri olarak kullanıldı.

Tablo 4.11. Heter-MV topluluk algoritmasının sınıflandırma sonuçları

Veri Kümesi	Doğruluk	F-Ölçümü	AUC
1150haber	95,71±0,83	95,82±1,03	99,18±0,22
Milliyet	92,27±0,76	91,62±0,53	98,76±0,84
Hurriyet	84,08±0,51	83,75±0,96	99,82±0,09
Aahaber	87,06±0,97	85,90±0,41	97,66±0,34

Tablo 4.11, Heter-MV topluluk sınıflandırıcısının performanslarını göstermektedir. Çoğunluk oylama modeli %95 (1150haber), %91 (Milliyet), %82 (Hürriyet) ve %84 (Aahaber) F-puanlarına ulaştı ve bu sonuçlar, yığın oluşturma yaklaşımıyla karşılaştırıldığında sistem performansını önemli ölçüde artırmadığı görüldü. Tablo 4.12, Heter-STCK istifleme topluluk yaklaşımının performanslarını göstermektedir. İstifleme topluluk yaklaşımı %97 (1150 haber), %94 (Milliyet), %84 (Hürriyet), %85 (Aahaber) F puanlarına ulaştı. AUC' nin ölçümleri birbirine çok yakın değerler çıktı.

Tablo 4.12. Heter-STCK topluluk algoritmasının sınıflandırma sonuçları

Veri Kümesi	Doğruluk	F-Ölçümü	AUC
1150haber	97,16±1,23	97,32±0,82	99,23±0,05
Milliyet	94,57±2,06	94,03±1,25	99,21±0,18
Hurriyet	85,44±0,92	85,57±1,10	99,73±0,27
Aahaber	87,73±1,17	86,81±0,72	98,45±0,71

AUC, sınıflandırıcıların dengesiz veri kümeleri üzerindeki performansını değerlendirmek için uygun olup yüksek AUC değerlerine sahip sınıflandırıcı, yanlış pozitif oranı en aza indirirken gerçek pozitif oranı en üst düzeye çıkarmaktadır. Deneylerimizde, Heter-STCK' nin diğer tüm yöntemler arasında üstün performansa sahip olduğu açıkça gözlendi.

4.4. İngilizce Metinler Kullanılarak Elde Edilen Sonuçlar

Tez kapsamında yapılan bir sonraki deneyde ise topluluk sistemlerinin performansını orijinal veri kümesi üzerinden yorumlamak yerine orijinal veri kümesini kelime yerleştirmeleri/gömülmeleri üzerinden değerlendirmeye yoğunlaşıldı. Bu amaçla, literatürde sıkça bahsedilen İngilizce veri kümeleri kullanıldı. İlk başta, kelime yerleştirmeleri, orijinal veri kümelerinden word2vec ile çıkarıldı. Daha sonra, bireysel sınıflandırıcıların, hem orijinal veri kümeleri hem de kelime

yerleştirmelerinde, topluluk sisteminin temel sınıflandırıcısını belirlemek için sınıflandırma başarısı gözlendi. Temel öğreniciyi belirlemek için öğrenme yöntemlerinden çok değişkenli Bernoulli saf Bayes (MVNB), çok terimli saf Bayes (MNB), destek vektör makineleri (SVM) ve karar ağaçlarını (DTs) kullanıldı. MVNB, veri kümelerinin her iki sürümünde de diğer makine öğrenme algoritmaları arasında üstün sınıflandırma performansını sergiledi.

Tablo 4.13 ve Tablo 4.14' de görüldüğü gibi, MVNB, topluluk algoritmasının temel öğrenicisi olarak orijinal veri kümelerinde %83,89 ortalama doğruluk ve kelime yerleştirmelerinde %85,41 ortalama doğruluk sergiledi. Bireysel sınıflandırma başarı sırasını ile şu şekilde gerçekleşti: MVNB> MNB> SVM> DT. Ayrıca, kelime yerleştirmelerinin kullanımı, orijinal veri kümeleriyle karşılaştırıldığında MVNB' nin ortalama doğruluk sonuçları açısından yaklaşık %2' lik bir artışla sınıflandırma başarısına katkıda bulunduğu gözlemlendi.

Tablo 4.13. Orijinal veri kümelerinde temel sınıflandırıcıların doğrulukları

Veri Kümesi	MVNB	MNB	SVM	DT
20News-19997	75,68±1,02	76,06±0,84	61,84±1,13	54,15±0,77
20News-18828	88,52±0,42	86,26±0,36	84,18±0,96	79,82±0,72
Mininews	87,25±1,10	77,90±0,61	83,00±0,42	77,12±0,56
WebKB4	84,10±1,12	85,64±1,17	89,85±0,96	77,45±1,07
ortalama	**83,89±0,92**	81,47±0,75	79,72±0,87	72,14±0,78

Sistemin temel öğrenicisini belirledikten sonra topluluk algoritmaları olarak torbalama, rastgele alt uzay ve rastgele ormanlar, Tablo 4.13 ve Tablo 4.14' de gösterildiği gibi hem orijinal veri kümelerinden hem de kelime yerleştirmelerinde eğitim veri kümelerini çeşitlendirmek için kullanıldı. Torbalamanın, tüm veri kümeleri için sınıflandırma performansını, hem orijinal veri kümeleri hem de kelime yerleştirmeleri üzerinde yükseltmek için etkili olmadığı gözlendi.

Torbalama yöntemi, 20News-18828 ve Mininews için doğruluk değerini yaklaşık %1 azaltırken 20News-19997 ve WebKB4 için Tablo 4.13' deki MVNB sınıflandırıcısının başarısını düşük oranda artırdı. Torbalamanın sonuçları, Tablo 4.14’ de tüm veri kümelerinin kelime yerleştirmeleri versiyonları için temel öğrenici MVNB'nin sınıflandırma sonuçlarına yakın çıktı. Diğer bir deyişle, topluluk

algoritması olarak torbalamanın kullanılması sınıflandırma performansında önemli iyileşme sağlayamadı. Rastgele alt uzay, doğruluk oranlarını yaklaşık %2 artırarak ortalama hem torbalama algoritması hem de sistemin temel öğrenicisine (MVNB) kıyasla daha iyi sonuç verdi. Rastgele alt uzay, tek bir sınıflandırıcının performansına kıyasla yaklaşık %1-2' lik bir iyileşme sağlarken sınıflandırma başarısı, Tablo 4.4.1' de görüldüğü gibi, orijinal veri kümelerindeki torbalama yöntemine kıyasla %1 ila %2 arasında iyileşme gözlemlendi. WebKB4 veri kümesi hariç, rastgele alt uzay algoritması, Tablo 4.4.2' de görüldüğü gibi temel öğrenicisi ve torbalama yöntemine kıyasla, kelime yerleştirmelerinde yaklaşık %2' lik bir artış sağladı. Öte yandan, rastgele alt uzay ve rastgele orman, torbalama algoritmasına ve MVNB' ye kıyasla belirgin sınıflandırma performansı sergiledi.

Tablo 4.14. Kelime yerleştirmelerinde temel sınıflandırıcıların doğrulukları

Veri Kümesi	MVNB	MNB	SVM	DT
20News-19997	77,13±0,78	76,80±0,55	66,13±0,45	60,32±0,63
20News-18828	89,45±0,34	86,71±0,60	85,44±0,87	80,20±0,80
Mininews	88,70±0,94	80,44±0,92	86,23±1,00	78,14±1,06
WebKB4	86,35±0,47	87,36±1,02	90,55±0,74	79,51±0,52
ortalama	**85,41±0,63**	82,83±0,77	82,09±0,77	74,54±0,75

Tablo 4.13 ve Tablo 4.14' e detaylıca baktığımızda rastgele orman algoritmasının, sınıflandırma başarısı açısından diğerlerinden daha iyi olduğu ve tüm veri kümeleri için diğerlerinden çok daha üstün performans gösterdiği gözlemlendi. Rastgele orman, sınıflandırma başarısına önemli bir katkı sağlamakla beraber temel sınıflandırıcı, torbalama ve rastgele alt uzay algoritmalarının başarısına kıyasla en az %2, en fazla %5' lik iyileştirme sağladı. En düşük ve en yüksek doğrulukta yapılan iyileştirmeler, 20News-19997 için %1-6, 20News-18828 için %2-6, Mininews için %2-4 ve WebKB4’ deki kelime yerleşimleri için maksimum %3 artışla farklılık gösterdi. Ayrıca, doğruluk sonuçları üzerindeki geliştirmeler, Tablo 4.4.1' de gösterildiği gibi 20News-19997 için %1-5, 20News-18828 için maksimum %4, Mininews için maksimum %3 artış, WebKB4 için %1-4 artışla değişmektedir. Özellikle 20News-19997 ve 20News-18828 veri kümeleri için, %6' lık artış, topluluk stratejilerinin kullanımı açısından sınıflandırma başarısına önemli katkı sağladığını gösterdi. Doğrulukların ortalama değerleri değerlendirilirken üstün performansa

sahip algoirtmalar açıkça görülmekle beraber bunların sınıflandırma başarısı şu şekilde elde edildi: RF>RS>BG. Sonuç olarak, rastgele ormanın metinsel verilerin sınıflandırılmasında sınıflandırma başarısını iyileştirmek için diğerleri arasından bir topluluk algoritması olarak seçilmesinin anlamlı olduğunu gözlemledik.

Bu aşamaya kadar metin içerikli veri kümelerinde topluluk algoritmalarının performansını inceledik. Topluluk algoritmaları kullanımının sınıflandırma başarısını iyileştirdiğini gözlemledik. Bundan sonraki aşamada ise metin içerikli özellik uzayımızı daha önce bahsettiğimiz özellik seçim yöntemleriyle genişletmeyi, bu özellik uzayını homojen topluluklarla sınıflamayı ve sınıflandırma performansını incelemeyi hedefledik. Deneyimizde, Türkçe bir veri kümesi üzerinde yoğunlaştık.

4.5. Türkçe Metinlerden Önerilen Yöntemlerle Elde Edilen Sonuçlar

İlk olarak, çok değişkenli Bernoulli saf Bayes (MVNB), çok terimli saf Bayes (MNB), destek vektör makineleri (SVM), rastgele orman (RF) gibi temel sınıflandırıcıyı belirlemek için denetlenen makine öğrenimi algoritmalarının sınıflandırma performansı araştırıldı.

Tablo 4.15. Temel öğrenicilerin tüm eğitim kümesi yüzdelerinde sınıflandırma doğrulukları

TS	MVNB	MNB	SVM	DT
80	82,70±1,07	**82,80±0,93**	79,62±1,12	79,13±1,43
50	81,10±0,96	**81,96±0,76**	78,83±1,24	78,21±1,20
30	79,96±1,24	**80,45±1,04**	77,50±0,88	77,42±0,92
10	74,60±1,36	**76,18±0,87**	73,46±1,10	72,80±0,98
5	69,12±0,95	**72,40±1,25**	69,50±1,38	67,91±1,40

Tablo 4.15, temel öğrenicilerin farklı ts yüzdelerindeki sınıflandırma doğruluklarını göstermektedir. MNB' nin sınıflandırma performansı, diğer temel öğrencileri belirgin bir şekilde geride bıraktı. MNB' nin sınıflandırma başarısının, küçük eğitim kümesi yüzdelerinde daha iyi olduğunu ve doğruluk değerlerinin daha yüksek eğitim kümesi seviyelerinde daha yakın olduğunu tespit ettik. Saf Bayes' in özellikle ts80 ve ts50' deki iki modeli arasında küçük bir fark olup diğer algoritmalara nazaran %3-4 daha üstün doğruluk değerlerine sahip olduğunu gözlemledik. Eğitim prosedürü, daha düşük eğitim kümesi yüzdelerinde gerçekleştirildiğinden saf Bayes' in iki modelinin

sınıflandırma başarısı arasında küçük fark arttı ve MNB, diğer temel öğrenicileri ezici üstünlükle geçti. Ayrıca, karar ağacının bir temel öğrenici olarak belirlenmesi, en düşük doğruluk değerleri nedeniyle sınıflandırma performansı açısından iyi bir seçenek olmayacağını gösterdi. Son olarak, temel öğrenicilerin sınıflandırma başarı sırası, MNB> MVNB> SVM> DT olarak sıralandı ve MNB, topluluk sisteminin diğer aşamalarına geçebilmek için bir temel öğrenci olarak seçildi.

Bir sonraki aşamada, metin kategorizasyonu için genişletilmiş uzay ormanlarının iki farklı versiyonuna odaklandık. Özellik uzayını genişletmek için birincisinde, orijinal özellik uzayının rastgele özelliklerine, ikincisi ise [11, 16]' da önerilen kazanım oranı tekniğinin kullanımıyla yüksek sınıflandırma kapasitesine sahip olan özelliklerin seçimine odaklandık. Sonrasında, özellik seçimi yöntemleriyle elde edilen yeni özellikler orijinal özelliklerle birleştirilerek yeni genişletilmiş özellik uzayı oluşturuldu. Yeni özellik uzayı oluşturulduktan sonra, eğitim veri kümesini çeşitlendirmek için torbalama, rastgele alt uzay ve rasgtele orman topluluk algoritmaları olarak kullanıldı. Tüm eğitim kümesi büyüklükleri için bir temel sınıflandırıcı olarak çok terimli saf Bayes modelini kullanan sınıflandırıcı topluluklarla genişletilmiş uzay ormanları elde edildi.

Tablo 4.16. Tüm eğitim kümesi boyutlarında topluluk algoritmaları açısından genişletilmiş uzay ormanlarının sınıflandırma doğrulukları

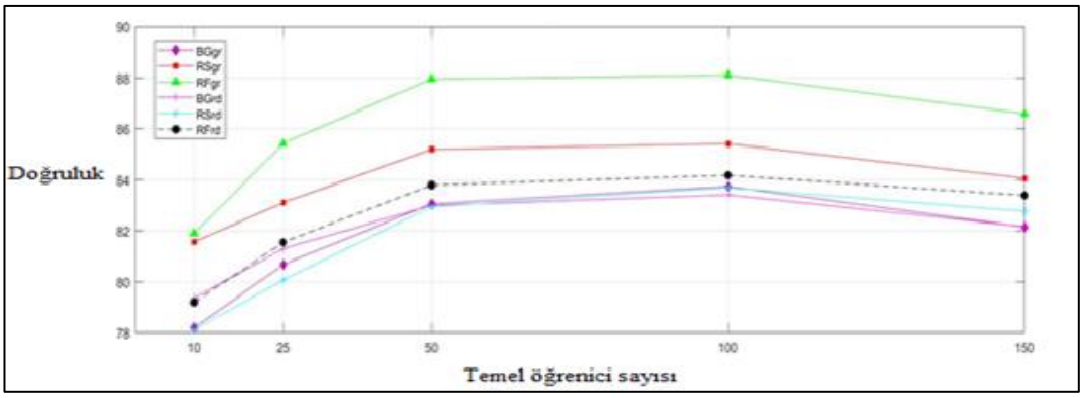
TS	MNB	BG_{GR}	RS_{GR}	RF_{GR}	BG_{RD}	RS_{RD}	RF_{RD}
80	82,80±0,93	83,75±0,72	85,42±0,85	**88,12±0,92**	83,42±0,43	83,70±1,40	84,18±0,73
50	81,96±0,76	82,37±1,26	84,16±0,99	**87,55±1,20**	82,14±0,78	82,40±0,95	83,57±1,09
30	80,45±1,04	81,50±0,94	83,47±1,24	**86,06±0,77**	81,33±1,66	81,52±1,13	82,90±1,48
10	76,18±0,87	76,84±1,06	79,32±1,12	**81,24±1,01**	76,42±1,21	77,50±2,14	78,63±2,27
5	72,40±1,25	72,88±1,36	76,59±2,44	**77,91±1,33**	72,57±1,90	74,10±1,36	75,29±1,77

Tablo 4.16, tüm eğitim kümesi yüzdelerinde tutma yönteminin on tekrarını yaparak, genişletilmiş uzay ormanlarının topluluk algoritmaları açısından ortalama sınıflandırma doğruluklarını ve ortalamadan sapmalarını göstermektedir. Kısaltmalar, topluluk algoritmaları ve özellik seçim teknikleri için şu şekilde kullanıldı: BG: Torbalama, RS: Rastgele Altuzay, RF: Rastgele Orman, X_{GR}: Veri kümesinin, X topluluğu algoritması için kazanç oranı ile genişletilmiş sürümü, X_{RD}: Veri kümesinin X topluluk algoritması için rastgele özelliklerle genişletilmiş sürümü, Ts:

Eğitim kümesi yüzdesi ve MNB, orijinal özellik uzayı versiyonuyla temel öğrenicinin sınıflandırma başarısını belirtti. Topluluk algoritmalarının performans sıralaması, hem rastgele olarak hem de kazanç oranı ile genişletilmiş uzay ormanları için diğer deneylerimizde gözlemlediğimiz sıralamayla aynı çıktı: RF> RS> BG. Tüm eğitim kümesi boyutlarında RF_{GR} en başarılı topluluk algoritması iken BG_{RD} genişletilmiş uzay orman teknikleri arasında en düşük doğruluk değerlerine sahip oldu. Bununla birlikte, BG_{RD}, temel öğrenicinin (MNB) sınıflandırma performansına kıyasla daha iyi olduğu görüldü. Bu sonuçlar, genişletilmiş uzay ormanlarının sınıflandırma başarısına katkı sağladığının kanıtı oldu. Sınıflandırma başarısı, özellik uzayı uzatma teknikleri açısından değerlendirildiğinde, en iyi sınıflandırma başarısı, çoğunlukla, kazanma oranı tekniği ile gerçekleştirildi ve bunu rastgele özellikler uzatılmış özellik uzayı izledi. Bu durumda, özellik uzatma teknikleri ve orijinal özellik uzayı sınıflandırma başarısı şu sırayla elde edildi: GR> RD> Orijinal.

Genişletilmiş tüm uzay teknikleri, topluluk algoritmaları ve sistemin temel öğrenicisine geniş bir perspektiften bakıldığında, sınıflandırma başarısı şu şekilde gerçekleşti: RF_{GR}> RS_{GR}> RF_{RD}> RS_{RD} ~ BG_{GR}> BG_{RD}> MNB. Daha önce de belirttiğimiz gibi, çoğunlukla kazanım oranıyla genişletilmiş uzay ormanlarının sınıflandırma performansı, diğerlerinden daha iyi performans gösterdi. RS_{RD} ve BG_{GR}' nin sadece ts10 ve ts5' deki doğruluk değerleri açısından farklılık gösterdiği ve RS_{RD}' nin daha küçük eğitim kümesi yüzdelerinde BG_{GR}' den daha iyi olduğu açıkça görüldü.

Şekil 4.6' da, 10' dan 150' ye değişen temel öğrenicilerin sayısının, genişletilmiş uzay ormanlarının sınıflandırma doğruluklarını gözlemlemek için önemli bir ölçüt olduğu gözlemlendi. Temel öğrenicilerin sayısı, 100' e yükseltildiğinde, doğruluk değerleri her genişletilmiş uzay ormanı yöntemi için de artış sağladı. Bununla birlikte, temel öğrenicilerin sayısı 100' den sonra da artmaya devam ettikçe sınıflandırma başarısı ters orantılı olarak önemli ölçüde azaldı. Bu nedenle, deneylerde temel öğrenici sayısı 100 olarak ayarlandı ve bu sonuçlar, literatürle uyumlu bulundu [10-11].



Şekil 4.6. Genişletilmiş uzay ormanlarının Aahaber veri kümesinde temel öğrenicilerin sayılarına göre sınıflandırma performansları

Gerçekleştirilen bir sonraki deneyde ise, metin içerikli özellik uzayı özellik seçim yöntemleriyle genişletilerek heterojen topluluklarla sınıflandırma performansının incelenmesi amaçlandı. Deneyimizde, üç Türkçe veri kümesi üzerine odaklandık. Öncelikle, deneylerimizde ilk yaptığımız iş olarak birbirleriyle karşılaştırmak amacıyla bireysel sınıflandırıcıların sınıflandırma başarısını araştırdık. Çok değişkenli Bernoulli saf Bayes (MVNB), çok terimli saf Bayes (MNB), destek vektör makinesi (SVM) ve Random Forest (RF) temel sınıflayıcılar olarak kullanıldı. Tablo 4.17' de de görüldüğü üzere, üç Türkçe veri kümesi üzerinde sınıflandırma sonuçları elde edildi. Tabloya ait her hücre, tutma yönteminin 10 tekrarının uygulanmasıyla elde edilen sınıflandırma doğruluklarını ve bu doğrulukların ortalamadan sapmasını içermektedir. RF' nin diğer sınıflandırma tekniklerine kıyasla daha üstün bir sınıflandırma sonucu gösterdiği açıkça görülmektedir. Temel sınıflandırıcıların sınıflandırma başarı sırasının şu şekilde gözlendi: RF> MVNB> MNB> SVM. Temel sınıflandırıcılarının kararlarını birleştirmek amacıyla çoğunluk oylama (MV) ve yığınlama (STCK) entegrasyon yöntemleri kullanıldı.

Tablo 4.17. Temel sınıflandırıcıların ts80' de sınıflandırma doğrulukları

Veri Kümesi	MVNB	MNB	SVM	RF
1150haber	93,74±1,35	94,00±1,64	89,65±2,62	94,32±1,02
Hurriyet	81,16±0,96	79,78±0,78	76,58±1,31	84,13±0,96
Aahaber	82,70±1,07	82,80±0,93	79,62±1,12	87,35±1,37
ortalama	85,56±1,08	84,52±1,09	83,82±1,39	**88,99±1,06**

Tablo 4.18’ de, her bir bireysel sınıflandırıcının doğrulukları, çoğunluk oyuyla (Heter-MV) elde edilen ve istifleme (Heter-STCK) ile elde edilen heterojen topluluk sistemleriyle karşılaştırıldı.

Tablo 4.18. Temel sınıflandırıcıların ve heterojen toplulukların ts80’ de sınıflandırma doğrulukları

Yöntem	1150haber	Hurriyet	Aahaber
MVNB	93,74±1,35	81,16±0,96	82,70±1,07
MNB	94,00±1,64	79,78±0,78	82,80±0,93
SVM	89,65±2,62	76,58±1,31	79,62±1,12
RF	94,32±1,02	84,13±0,96	87,35±1,37
Heter-MV	**95,71±0,83**	**84,08±0,51**	**87,06±0,97**
Heter-Stck	**97,16±1,23**	**85,44±0,92**	**87,73±1,17**

Tablo 4.18' den de görüldüğü gibi, Heter-MV ve Heter-STCK toplulukları ile, bireysel sınıflandırıcılardan daha iyi doğruluk sonuçları elde edildi. Heter-STCK, hem tüm temel sınıflandırıcılar arasından sıyrılarak hem de Heter-MV çoğunluk oylama topluluğu modelininin performansını geçerek üstün sınıflandırma başarısı gösterdi. Heter-MV (çoğunluk oylama topluluk modeli), topluluk doğrulukları açısından rekabetçi olduğu ve ayrıca bireysel sınıflandırıcılar ile karşılaştırıldığında daha iyi sınıflandırma başarısına sahip olduğu görüldü. Heter-STCK, çoğunluk oylama topluluğu modeline kıyasla sınıflandırma başarısının en fazla %2 oranında iyileşmesini sağladığında, sınıflandırma iyileştirmesi, farklı veri kümeleri için temel sınıflandırıcılara bağlı olarak %2 ila %9 arasında değişiklik gösterdi. Diğer bir önemli gözlem, heterojen topluluk sisteminin istifleme yöntemiyle sınıflandırma performansının, rastgele orman (RF) algoritmasından daha iyi olması oldu. Heter-Stck, homojen topluluk sistemine (RF) göre üstün sınıflandırma başarısı gösterirken, RF yöntemiyle homojen topluluk sistemi ve çoğunluk oylama modeline sahip heterojen topluluk sistemi benzer sınıflandırma performansları sergiledi. Bu nedenle, istifleme yöntemiyle elde edilen heterojen topluluk sisteminin performansı, önerilen sistemin sınıflandırma başarısı açısından önemli olduğunu kanıtladı.

1150haber veri kümesi için Tablo 4.19, Hürriyet veri kümesi için Tablo 4.20, ve Tablo 4.21 Aahaber veri kümesi için heterojen topluluk sisteminin orijinal ve genişletilmiş uzay versiyonlarının ortalama sınıflandırma doğruluklarını ts80 için göstermektedir.

Tablo 4.19. Bireysel sınıflandırıcıların ve heterojen topluluk algoritmalarının genişletilmiş ve orijinal versiyonlarının 1150haber veri kümesinde ts80'deki sınıflandırma doğrulukları

Yöntem	Orijinal	RND-ES	GR-ES
MVNB	93,74±1,35	92,56±1,05	94,15±0,94
MNB	94,00±1,64	93,67±0,82	94,88±1,12
SVM	89,65±2,62	89,40±1,56	90,23±1,09
RF	94,32±1,02	94,15±0,93	95,56±0,73
Heter-MV	**95,71±0,83**	**95,07±0,77**	**96,05±0,81**
Heter-Stck	**97,16±1,23**	**96,92±0,96**	**98,12±0,65**

Rastgele özelliklerle (RND-ES) genişletilmiş uzay ormanları, 1150haber ve Aahaber veri kümelerindeki orijinal sürümlerle orantılı olarak daha kötü veya hemen hemen aynı sınıflandırma sonuçları sundu. Diğer veri kümelerinden farklı olarak, Hürriyet, orijinal versiyona kıyasla rastgele genişletilmiş uzay ormanlarında %1' lik iyileştirme ile genellikle daha iyi doğruluk sonuçları gösterdi. Ayrıca, rastgele özelliklerle genişletilmiş homojen (RF) ve heterojen topluluk sistemlerinin (Heter-MV ve Heter-Stck) sınıflandırma başarısı, tüm veri kümeleri için tek sınıflandırıcıların genişletilmiş versiyonlarından daha yüksek doğruluklar sergiledi.

Tablo 4.20. Bireysel sınıflandırıcıların ve heterojen topluluk algoritmalarının genişletilmiş ve orijinal versiyonlarının Hürriyet veri kümesinde ts80'deki sınıflandırma doğrulukları

Yöntem	Orijinal	RND-ES	GR-ES
MVNB	81,16±0,96	81,75±1,30	82,04±0,74
MNB	79,78±0,78	79,13±1,27	79,95±0,81
SVM	76,58±1,31	76,40±0,85	76,81±1,06
RF	84,13±0,96	85,36±1,00	86,47±0,95
Heter-MV	**84,08±0,51**	**85,57±0,80**	**86,93±0,64**
Heter-Stck	**85,44±0,92**	**86,80±1,04**	**88,21±0,86**

GR tabanlı genişletilmiş uzay ormanları, topluluk sisteminin sınıflandırma performansını önemli ölçüde etkiledi. GR tabanlı genişletilmiş uzay versiyonları, orijinal versiyonlara kıyasla 1150haber için %1, Hürriyet için %2-3 iyileşme ve Aahaber veri kümesi için %1-2 artış sergiledi.

Tablo 4.21. Bireysel sınıflandırıcıların ve heterojen topluluk algoritmalarının genişletilmiş ve orijinal versiyonlarının Aahaber veri kümesinde ts80'deki sınıflandırma doğrulukları

Yöntem	Orijinal	RND-ES	GR-ES
MVNB	82,70±1,07	80,92±1,35	83,27±0,84
MNB	82,80±0,93	81,03±1,24	83,45±1,07
SVM	79,62±1,12	77,45±0,97	80,30±1,16
RF	87,35±1,37	84,18±0,73	88,12±0,92
Heter-MV	**87,06±0,97**	**84,36±1,59**	**88,72±0,88**
Heter-Stck	**87,73±1,17**	**85,80±1,10**	**89,25±0,63**

Ayrıca, GR tabanlı genişletilmiş uzay ormanları, diğer genişletilmiş uzay ormanı sürümleri arasında üstün başarıya sahip oldu. GR tabanlı model ile genişletilmiş uzay ormanları, rastgele genişletilmiş uzay ormanları ile 1150haber ve Hurriyet için %1-2 iyileşme, Aahaber için %2-4 artış göstermektedir. Üstelik, GR tabanlı homojen (RF) ve heterojen topluluk sistemlerinin (Heter-MV ve Heter-Stck) performansı, hem rastgele genişletilmiş uzay ormanların hem de tüm veri kümeleri için bireysel sınıflandırıcıların genişletilmiş sürümlerinin sınıflandırma performanslarını geçti. Son olarak, genişletilmiş uzay ormanlarının sınıflandırma başarısı, özellik uzayı geliştirme tekniklerine göre şu sırayla elde edildi: GR> RND> Orijinal.

4.6. İngilizce Twitter Verilerinden Elde Edilen Sonuçlar

Metin içerikli yapılan deneylerin sonuncusunda ise yine özellik uzayını genişletmeye ve yeni özellik uzayını topluluk algoritmalarıyla harmanlamaya odaklanıldı. Bu aşamaya kadar kullanılan uzay genişletme algoritmalarına ek olarak derin öğrenme temelli kelime yerleştirmelerini elde eden yöntemler kullanıldı. Son çalışmanın diğerlerinden bir başka farkı ise sosyal medyada yayınlanan paylaşımların duygu analizine yoğunlaşması oldu.

Sonuç olarak, sosyal medya kullanıcılarının paylaşımlarından faydalanarak önerilen genişletilmiş özellik uzayı ve topluluk algoritmaları kombinasyonunun duygu analizi

sınıflandırmasına katkı sağlaması amaçlandı. Bu amaçla, Twitter' dan toplanmış ve yaygın olarak kullanılan beş İngilizce veri kümesi üzerinde deneyler gerçekleştirildi.

Gerçekleştirilen deneyler, Tablo 4.22' de her temel sınıflandırıcının duygu analizi sınıflandırma başarısını göstermektedir. Kalın değerlerle, en iyi puanlar ifade edilmiştir. F-ölçüm ve doğruluk sonuçları, çalışmamızın katkısını göstermek için değerlendirme ölçütü olarak kullanıldı. Kısaltmalar ise şu şekilde kullanılmıştır: BG: Torbalama, BS: Artırma, RS: Rastgele alt uzay, RF: Rastgele orman, X_{IG}: X topluluk algoritması için IG tabanlı özelliklerle genişletilmiş özellik uzayı, X_{ACO}: X topluluk algoritması için ACO tabanlı özellikler ile genişletilmiş özellik uzayı ve X_{WE}: X topluluk algoritması için WE tabanlı özellikler ile genişletilmiş özellik uzayını ifade etmektedir.

Tablo 4.22. Temel sınıf sınıflandırıcıların ts80'de ortalama F-ölçümü sonuçları

Veri Kümesi	MNB	MVNB	SVM	RF
Sts-Gold	82,15±0,07	81,36±0,04	83,44±0,02	82,90±0,06
Sts-Test	81,30±0,05	80,12±0,02	82,96±0,01	81,75±0,04
Iphone6	70,42±0,03	74,48±0,05	73,66±0,03	72,15±0,09
Archeage	85,13±0,02	85,91±0,05	86,20±0,03	84,30±0,04
Hobbit	87,10±0,04	84,36±0,02	90,45±0,02	88,23±0,08
ortalama	81,22±0,04	81,25±0,03	**83,34±0,02**	81,87±0,06

Tablo 4.22' de görüldüğü gibi, her temel sınıf sınıflandırıcısının ortalama F-skor değerleri dikkate alındığında en iyi F-skoru performansı SVM tarafından elde edildi. RF, MNB ve MVNB' den biraz daha iyi bir performansa sahipken, MNB ve MVNB neredeyse aynı sınıflandırma başarısı sergilediler. Bu nedenle, bir temel öğrenicisi olarak SVM, en yüksek F-ölçüm değerleri nedeniyle sınıflandırma performansı açısından iyi bir seçimdir. Sonuçta, temel öğrenicilerin sınıflandırma başarı sırası SVM> RF> MVNB> MNB şeklinde gerçekleşti.

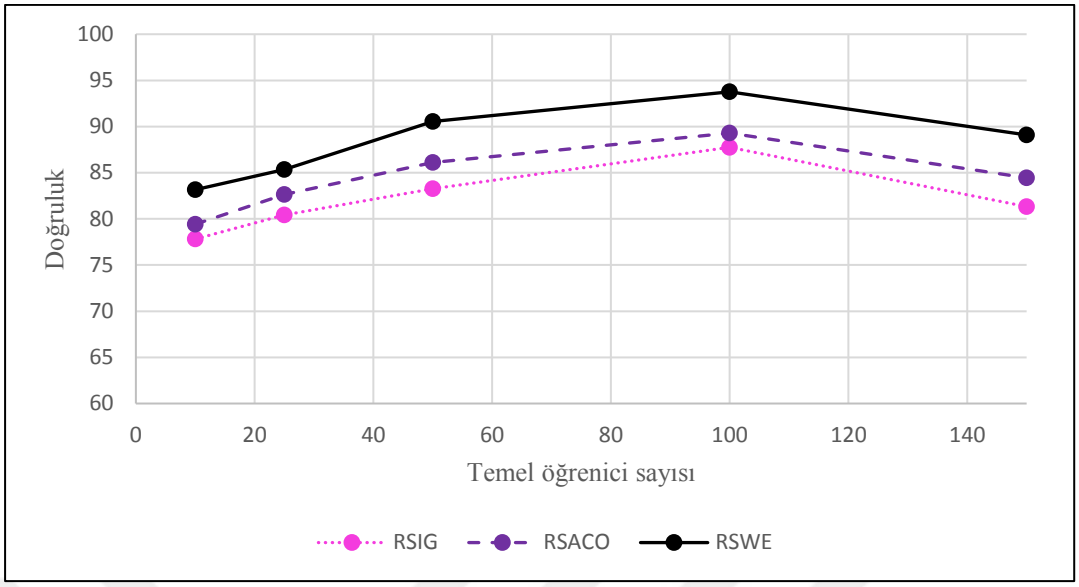
Sonuçlar, önerilen WE temelli topluluk sistemlerinin, diğer genişletilmiş özellik uzay tabanlı topluluk sistemlerinden herhangi birine göre daha üstün sınıflandırma performansı sunduğunu gösterdi. Sınıflandırma başarısı, ts80' de Tablo 4.6.2' de görüldüğü gibi, RS_{WE}> BS_{WE}> BG_{WE}> RS_{ACO}> BS_{ACO}> BG_{ACO}> RS_{IG}> BS_{IG}> BG_{IG}> SVM olarak sıralandı. Geliştirilmiş uzay tabanlı topluluk sistemlerinin tüm

sürümleri, temel sınıflandırıcılar ile karşılaştırıldığında %5'e kadar iyileşme sağlayarak sınıflandırma performansına önemli ölçüde katkıda bulundu. Topluluk algoritmalarının performans sırası, ortalama F-ölçümü sonuçları açısından tüm genişletilmiş uzay versiyonları için RS> BS> BG şeklinde sıralandı. Dahası, üstün başarısından ötürü özellik uzayı derin öğrenme temelli özelliklerle genişletildi. Topluluk algoritması RS olarak ayarlandığında, tüm veri kümeleri için uzay genişletme tekniklerinin sınıflandırma başarısı WE> ACO> IG şeklinde oldu.

Tablo 4.23. Önerilen yöntemin ts80'de ortalama F-ölçümü sonuçları

Yöntem	SVM	BG_{IG}	BS_{IG}	RS_{IG}	BG_{ACO}	BS_{ACO}	RS_{ACO}	BG_{WE}	BS_{WE}	RS_{WE}
Sts-Gold	83,44	83,40	83,45	83,88	83,72	83,91	84,20	86,46	86,95	**88,44**
Sts-Test	82,96	82,80	82,91	83,14	82,95	83,12	83,77	85,90	86,53	**87,45**
Iphone6	73,66	74,10	74,25	74,40	74,75	74,82	75,10	77,23	78,67	**79,95**
Archeage	86,20	86,53	86,70	86,80	86,75	86,90	87,05	89,21	90,23	**91,55**
Hobbit	90,45	90,12	90,20	90,55	90,44	90,73	91,33	94,22	95,88	**96,41**
ortalama	83,34	83,39	83,50	83,75	83,72	83,90	84,29	86,60	87,65	**88,76**

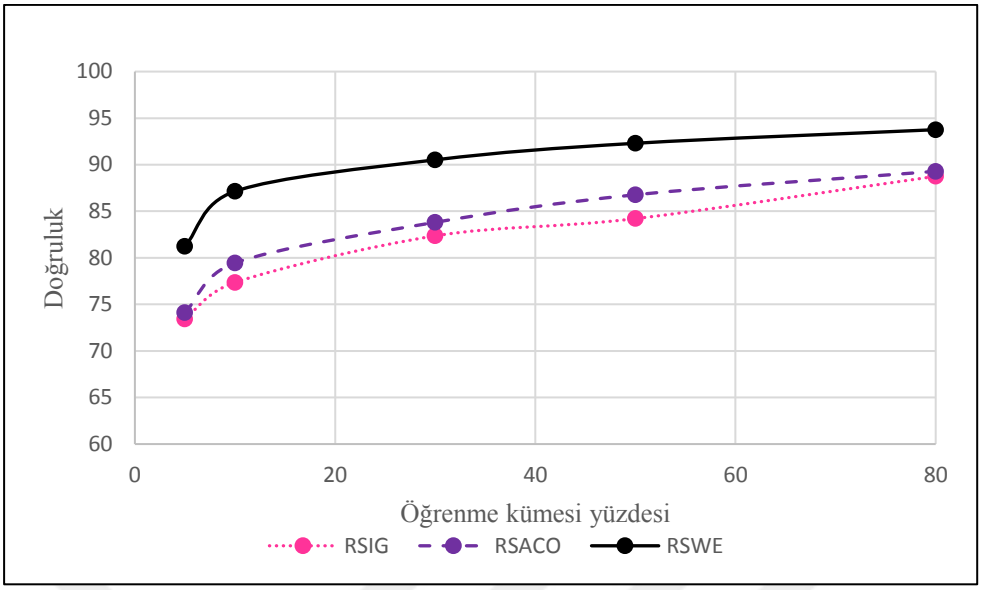
IG ve ACO tabanlı genişletilmiş özellik uzayının sınıflandırma performansı rekabetçi olup topluluk algoritmaları açısından sonuçların yakınlığı nedeniyle istatistiksel olarak kayda değer bir iddiada bulunmak için yeterli olmadı. Bu nedenle, bir topluluk algoritması rastgele alt uzayın ve WE tabanlı genişletilmiş özellik uzayının kombinasyonu ts80'de en yüksek sonuçların elde edilmesini sağladı. Diğer bir deyişle, %88,76 (RS_{WE}) sonuçla önerilen yöntemimiz, ortalama F ölçümü sonuçları açısından bakıldığında tüm veri kümelerinin sınıflandırma performansını artıran en iyi model oldu.



Şekil 4.7. Özellik uzayı genişletme tekniklerinin doğruluk sonuçları

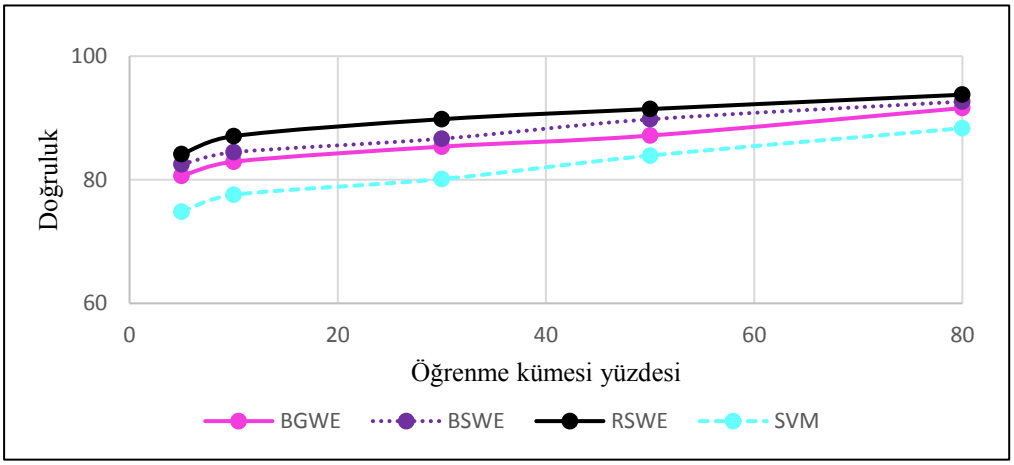
Şekil 4.7, uzay genişletme tekniklerinin sınıflandırma performansını temel öğrenicilerin sayısı açısından göstermektedir. Görüldüğü gibi, 10 'dan 150' ye değişen temel öğrenicilerin sayısı, genişletilmiş uzay tabanlı tekniklerin sınıflandırma başarısını gözlemlemek için önemli bir ölçüttür. Temel öğrenicilerin sayısı 100' e kadar yükseltildiğinde, doğruluk sonuçları her bir genişletilmiş özellik uzayı yöntemi için de artış gösterdi. Bununla birlikte, temel öğrenicilerin sayısı 100' den sonra da artmaya devam ettikçe, sınıflandırma başarısının ters orantılı olarak önemli ölçüde azaldığı gözlendi. Bu nedenle, deneylerde temel öğrenici sayısı, bu deney için de 100 olarak belirlendi.

Şekil 4.8' de, rastgele alt uzay algoritmasının sınıflandırma performansı önerilen genişletilmiş uzay yöntemleri açısından değerlendirildi. WE tabanlı genişletilmiş uzay modeli, tüm eğitim kümesi yüzdelerinde geleneksel özellik seçim tekniklerini geride bıraktı. WE tabanlı ve IG tabanlı modeller arasındaki doğruluk sonuçları farkı, daha küçük eğitim kümesi boyutlarında %10' a kadar çıktı. Öte yandan, ACO tabanlı genişletilmiş uzay modeli, ortalama doğruluk sonuçları açısından IG bazlı modelden daha rekabetçi oldu ve IG tabanlı modelden %2 daha fazla bir iyileştirme gerçekleştirdi. Böylece, ACO tabanlı model ile WE tabanlı model arasındaki fark, daha küçük eğitim seti yüzdeleriyle maksimum %8'e ulaşmış oldu.



Şekil 4.8. Önerilen genişletilmiş uzay teknikleri açısından öğrenme kümesi yüzdelerine göre RS topluluk algoritmasının doğruluk sonuçları

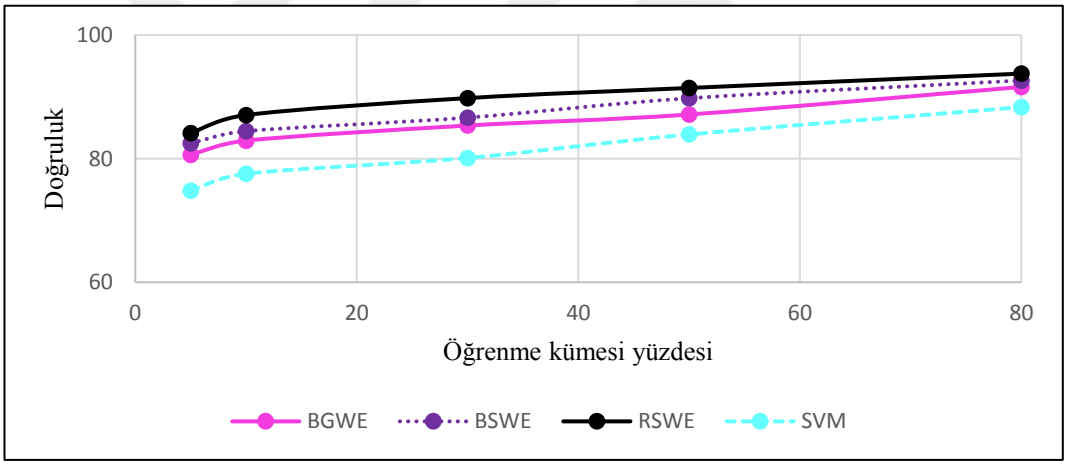
ACO tabanlı model, ts80' deki IG bazlı modele kıyasla yaklaşık %1' lik bir geliştirme sergilemesine rağmen, iki modelin sınıflandırma başarısının, tüm eğitim kümesi seviyelerinde genellikle birbirine çok yakın olduğu gözlendi. Sonuçta, özellik uzayını genişletmek için en iyi tekniğin, derin öğrenme temelli özellikleri (WE) çıkarmak ve orijinalleri ile entegre etmek olduğu gözlendi.



Şekil 4.9. WE tabanlı genişletilmiş uzayların topluluk algoritmaları açısından doğruluk sonuçları

Şekil 4.9' da, temel öğrenici SVM' in ve WE tabanlı genişletilmiş uzayların sınıflandırma başarısı, topluluk algoritmaları açısından analiz edildi. WE tabanlı model ile özellik uzayını genişletmek ve topluluk algoritmalarını kullanmak, ts80 ve ts5' te temel öğrenicinin sınıflandırma performansına kıyasla en az %3, en fazla %10

iyileşme sağladı. Yani, önerilen modeller ile temel öğrenici arasındaki fark, daha küçük yüzdelerde artarken daha büyük yüzdelerde azalmış oldu. Eğitim kümesi yüzdesi azaldıkça, her topluluk algoritmasının sınıflandırma başarısı arasındaki fark da daha belirgin hale geldi. Örneğin, BS yöntemi BG' ye göre %1 daha iyi performans gösterirken RS modeli ts80' de BS' ye kıyasla %1 daha iyi bir geliştirme sundu. RS ve BS, sistemin sınıflandırma performansını, daha küçük yüzdelerde BG' ye kıyasla sırasıyla yaklaşık %4 ve %2 oranında artırdı. Sonuç olarak, derin öğrenme temelli tekniklerle özellik uzayını genişletmek ve bu özellik uzayını topluluk algoritmalarıyla birleştirmek, tek bir sınıflandırıcıya kıyasla çok daha iyi sınıflandırma performansı sundu. Şekil 4.9 ve 4.10, kısa metin sınıflandırmasında derin öğrenme temelli özelliklerin kullanıldığı genişletilmiş özellik uzayının, bir topluluk algoritması olarak ta rastgele alt uzayların kullanımının en iyi sınıflandırma performansını sağlayacağını belirgin bir şekilde görüldü.



Şekil 4.10. WE tabanlı genişletilmiş uzayların topluluk algoritmaları açısından doğruluk sonuçları

Önerilen tekniğimizin katkısını göstermek için deney sonuçlarının genişletilmiş uzaylar üzerindeki güncel literatür çalışmaları [48, 74] ile karşılaştırılması önemlidir. Tablo 4.24' te görüldüğü gibi, kıyaslama yapılan çalışma [74], dokuz veri kümesi kullanmaktadır ve bunlardan dördü deneylerimizde kullandığımız veri kümeleriyle ortaktır. Aynı deneysel ayarlar yapıldığında tüm veri kümeleri için önerilen yöntemimizin üstünlüğü çok belirgin bir şekilde gözlendi.

Tablo 4.24. Önerilen yöntemin sınıflandırma başarısının F ölçümü sonuçları açısından 1. çalışma ile karşılaştırılması

Yöntem	Sts-Test	Iphone6	Archeage	Hobbit
RS_{WE}	**87,4**	**79,9**	**91,5**	**96,4**
1. çalışma	86,3[74]	73,8[74]	86,9[74]	92,1[74]

Diğer çalışma [48], topluluk sistemi için dokuz farklı yöntem önermiş. Yazarlar, altı veri kümesi kullanmıştır ve bunlardan biri Tablo 4.25' te görüldüğü gibi bizimkilerle ortaktır. Önerdiğimiz yöntem, MSG + bg tekniği dışındaki tüm yöntemleri geride bıraktı. Bizim %88,4 ve onların %89,2 F-ölçümü sonuçları arasındaki küçük fark, deneysel ortamlardaki farklılıklardan kaynaklanabilir ve ortalama F-skorlarının yakınlığı nedeniyle istatistiksel olarak elde edilen sonuçların anlamlı olduğunu iddia etmek yeterli değildir. Bu nedenle, önerilen tekniğimizin (RS_{WE}), son güncel çalışmaların sınıflandırma performanslarından çok daha üstün olduğu kanıtlandı.

Tablo 4.25. Önerilen yöntemin sınıflandırma başarısının F ölçümü sonuçları açısından 2. çalışma ile karşılaştırılması

Yöntem	Sts-Gold
RS_{WE}	***88,4***
M_G[48]	83,4
CEM_{SG}^{Vo}[48]	83,5
CEM_{SG}^{ME}[48]	84,5
M_{SG}[48]	84,7
M_{SG+bg}[48]	***89,2***
M_{GA}[48]	85,2
M_{SGA+bg}[48]	85,2
CEM_{SGA}^{Vo}[48]	87,0
CEM_{SGA}^{ME}[48]	85,5

5. SONUÇLAR VE ÖNERİLER

Topluluk sistemlerinin üstünlüğü, daha önce de belirtildiği gibi makine öğrenimi alanında yaygın kabul gören bir varsayımdır. Bu yaklaşıma göre daha doğru ve sağlam modeller üretilmesi tavsiye edilmektedir. Tez kapsamında, genişletilmiş uzay ormanlarının topluluk algoritmalarını kullanarak sınıflandırma performansına katkısının araştırılması önerilmektedir. Bu amaçla, özellik uzayı üzerinde daha önce denenmemiş özellik seçim/çıkarım teknikleri kullanılarak genişletilmiş uzay ormanları kavramının bir adım daha ileriye taşınması hedeflenmektedir.

Dahası, bu çalışma, karınca kolonisi optimizasyonun, bilgi kazanımının, ki-karenin ve kelime yerleştirmelerinin kullanımı açısından sınıflandırıcı toplulukları ile genişletilmiş uzay ormanları üzerine yapılan ilk araştırmadır. Belirtilen tekniklerle seçilen özellikler, yeni bir genişletilmiş özellik uzayı oluşturmak için orijinal özelliklerle harmanlanmaktadır. Sonrasında topluluk algoritmalarıyla, topluluk sisteminin homojen ve heterojen olmasına bağlı olarak sırasıyla veri kümesi çeşitliliği ve sınıflandırıcı çeşitliliğiyle sınıflandırma performansının artırılması hedeflenmektedir. Kapsamlı deneysel çalışmalar, önerilen yöntemlerle geliştirilen genişletilmiş uzay ormanlarının orijinal özellik uzaylarına ve güncel literatür çalışmaların çeşitli genişletilmiş sürümlerine kıyasla; sınıflandırma performansında kayda değer bir artış sağlamaktadır. Genel sınıflandırma performansları göz önünde bulundurulduğunda, orijinal özellik uzayına sahip uzay ormanları, tüm eğitim kümesi seviyelerinde en düşük doğruluk oranlarına sahiptir. Gerçekleştirilen deneyler ile saptanan bu durum, orijinal özellik uzayına sahip uzay ormanlarının geliştirilmeye ihtiyacının olduğunu göstermektedir.

Topluluk algoritmalarının sınıflandırma performanslarından elde ettiğimiz sonuçlar, veri kümelerinin özelliklerine ve kullanılan topluluk öğrenme yöntemlerinin çeşitliliğine göre değişiklik göstermektedir. Gerçekleştirilen genel bir performans değerlendirmesi sonucunda, homojen topluluk sistemleri kullanıldığında RF> RS> BS> BG şeklinde bir performans sıralamasının geçerli olduğu gözlenmektedir. Buna

karşın, heterojen topluluk sistemi kullandığında Heter-STCK' nin Heter-MV' ye kıyasla temel sınıflandırıcıları entegre eden en iyi yöntem olduğu gözlenmektedir. Önemli bir detay olarak heterojen topluluk sistemlerinin, homojen sistemlere göre çok daha iyi sınıflandırma performansı sergilediği görülmektedir. Genişletilen özellik uzaylarının topluluk yöntemleriyle harmanlamasından çıkan sonuçlar, orijinal özellik uzayı ile edilene göre çok daha üstün bir performans sergilemektedir. Buradan hareketle, çalışmamızın yapı taşını oluşturan özellik uzayı genişletme tekniklerinin performans değerlendirmesine odaklanılmaktadır.

Orijinal özellik uzayı ile rastgele genişletilen ve ACO ile genişletilen özellik uzayının sonuçları kıyaslandığında en iyi performansı ACO' nun sergilediği görülürken; rastgele özelliklerle genişletilen özellik uzayının onu takip ettiği ve en kötü sınıflandırma başarısının orijinal özellik uzayına ait olduğu gözlenmektedir. IG ve CHI ile genişletilen özellik uzayları ile orijinal özellik uzaylarının başarısı kıyaslandığında ise şu sıralama elde edilmektedir: IG> CHI> Orijinal. GR ile ve rastgele genişletilen özelliklerle elde edilen genişletilmiş özellik uzaylarının performansı yine orijinal özellik uzayı ile kıyaslandığında ise sıralama şu şekilde gerçekleşmektedir: GR> Rastgele> Orijinal.

Son olarak, IG, ACO ve WE ile genişletilen özellik uzayları orijinal özellik uzayı ile kıyaslandığında, en iyi performansın WE' ye ait olduğu; onu sırasıyla ACO, IG ve orijinal özellik uzaylarının sınıflandırma başarılarının takip ettiği gözlenmektedir. Sonuç olarak, metin ve sayısal veriler dahil olmak üzere, özellik uzayının genişletilmesinin ve bu genişletilmiş yeni özellik uzayının topluluk algoritmalarıyla homojen ve heterojen olarak ayrı ayrı harmanlamasının sunduğu başarılı sonuçlar, önerilen tekniklerin literatüre katkı sağladığını göstermektedir.

Genişletilmiş uzay ormanlarının sınıflandırma başarısının yanı sıra, yürütme zamanı analizi de test ve eğitim süreleri açısından değerlendirilmektedir. Daha fazla özelliğe sahip olması ve özelliklerin arama süresiyle doğrudan orantılı olması nedeniyle, orijinal uzaylara kıyasla genişletilmiş uzay ormanları için daha fazla eğitim süresi gerekmektedir. Test süresine dair ipucu veren temel öğrenicilerin karmaşıklığı, bir ağaçtaki düğüm sayısıyla orantılıdır. Bu nedenle, en karmaşık temel öğreniciler en büyük ağaçlara sahip olmalarından dolayı rastgele orman algoritması tarafından

oluşturulmaktadır. Geliştirilmiş uzay ormanlarının, orijinal uzay ormanlarına kıyasla daha küçük ağaçlara sahip olmalarından dolayı daha az test süresi gerektirdiğini vurgulamak önemlidir.

Twitter veri kümelerindeki eğitim süresi, Intel® Xeon® E5-2643 3,30 GHz makinede 12 iş parçacığı kullanılarak yaklaşık 1 saat 25 dakika, Türkçe ve İngilizce veri kümelerindeki eğitim süresi yaklaşık 3 saat 35 dakika ve UCI makine öğrenmesi deposundan elde edilen 36 veri kümesindeki toplam eğitim süresi yaklaşık 1 saat 5 dakika olarak gerçekleşmektedir. Önerilen modellerin eğitiminin, grafik işlem birimi (GPU) ile gerçekleştirilmesinin, eğitim süresi performansı üzerinde büyük bir etkisinin olabileceği düşünülmektedir. Gelecekte, önerdiğimiz bu teknikleri derin öğrenme temelli yapay sinir ağlarıyla harmanlayarak topluluk algoritmalarının sınıflandırma başarısının daha da artırılması hedeflenmektedir.

KAYNAKLAR

[1] Darwish S. M., Adel A. Z., Ebaid D. B., A Novel System for Document Classification Using Genetic Programming, *Journal of Advances in Information Technology*, 2015, **6**(4), 194-200.

[2] Sebastiani F., Machine learning in Automated Text Categorization, *ACM Computing Surveys*, 2002, **34**(1), 1-47.

[3] Aggarwal C. C., Zhai C. X., A Survey of Text Classification Algorithms, Editors: Aggarwal C. C., Zhai C. X., *Mining Text Data*, Springer-Verlag, New York, 163-222, 2012.

[4] Polikar R., Ensemble Based Systems in Decision Making, *IEEE Circ. Syst. Mag.*, 2006, **6**, 21-45.

[5] Rokach L., Taxonomy for Characterizing Ensemble Methods in Classification Tasks: A Review and Annotated Bibliography, *Comput. Stat. Data. An.*, 2009, **53**, 4046-4072.

[6] Rokach L., Ensemble Based Classifiers, *Artif. Intell. Rev.*, 2010, **33**, 1-39.

[7] Gopika D., Azhagusundari B., An Analysis on Ensemble Methods in Classification Tasks, *Int. J. Adv. Res. Comp. Com.*, 2014, **3**, 7423-7427.

[8] Ren Y., Zhang L., Suganthan P. N., Ensemble Classification and Regression-Recent Developments, Applications and Future Directions, *IEEE Comp. Intell. Mag.*, 2016, **11**, 41-53.

[9] Peralta B., Soto A., Embedded Local Feature Selection within Mixture of Experts, *Information Sciences*, 2014, **269**, 176-187.

[10] Amasyalı M. F., Ersoy O. K., Classifier Ensembles with the Extended Space Forest, *J. IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**(3), 549-562.

[11] Adnan M. N., Islam M. Z., Kwan P. W. H., Extended Space Decision Tree, *Machine Learning and Cybernetics Conference*, Lanzhou, China, 13-16 Temmuz 2014.

[12] Koutanaei F. M., Sajedi H., Khanbabaei M., A Hybrid Data Mining Model of Feature Selection Algorithms and Ensemble Learning Classifiers for Credit Scoring, *Journal of Retailing and Consumer Services*, 2015, **27**, 11-23.

[13] Breiman L., Friedman J., Olshen R., Stone C., *Classification and Regression Trees*, Chapman and Hall, New York, USA, 1984.

[14] Ho T. K., The Random Subspace Method for Constructing Decision Forests, *IEEE T. Pattern Anal.*, 1998, **20**, 832-844.

[15] Yıldız O. T., Alpaydın E., Omnivariate Decision Trees, *IEEE T. Neural Networ.*, 2001, **12**, 1539-1546.

[16] Adnan M. N., Islam M. Z. A., Comprehensive Method for Attribute Space Extension for Random Forest, *International Conference on Computer and Information Technology*, Dhaka, Bangladesh, 22-23 Aralık 2014.

[17] Ahmed A., Brown G., Random Projection Random Discretization Ensembles-Ensembles of Linear Multivariate Decision Trees, *J. IEEE T. Knowl. Data En.*, 2014, **26**, 1225-1239.

[18] Joachims T., Text Categorization with Suport Vector Machines: Learning with Many Relevant Features, *European Conference on Machine* Learning, Springer-Verlag, London, 21-23 Nisan 1998.

[19] Singh V., Pradhan M. A., Advanced Methodologies Employed in Ensemble of Classifiers: A Survey, *International Journal of Science and Research*, 2014, **3**(12), 591-595.

[20] Dietterich T. G., Ensemble Methods in Machine Learning, *Multiple Classifier Systems*, 2001, **1857**, 1-15.

[21] Tan A. C., Gilbert D, Ensemble Machine Learning on Gene Expression Data for Cancer Classification, *Appl. Bioinformatics*, 2003, **2**, 75-83.

[22] Dong Y. S., Han K. S., A Comparison of Several Ensemble Methods for Text Categorization, *IEEE International Conference on Service Computing*, Shanghai, China, 15-18 Eylül 2004.

[23] Koprinska I., Poon J., Clark J., Chan J., Learning to Classify E-mail, *Information Sciences*, 2007, **177**(10), 2167-2187.

[24] Xia R., Zong C., Li S., Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification, *Information Sciences*, 2011, **181**, 1138-1152.

[25] Liu Y., Yu X., Huang J. X., An A., Combining Integrated Sampling with SVM Ensembles for Learning from Imbalanced Datasets, *Information Processing and Management*, 2011, **47**, 617-631.

[26] Rahman A., Verma B., Effect of Ensemble Classifier Composition on Offline Cursive Character Recognition, *Information Processing and Management*, 2013, **49**(4), 852-864.

[27] Rooney N., Wang H., Taylor P. S., An Investigation into the Application of Ensemble Learning for Entailment Classification, *Information Processing and Management*, 2014, **50**(1), 87-103.

[28] Wang G., Zhang Z., Sun J., Yang S., Larson C. A., POS-RS: A Random Subspace Method for Sentiment Classification Based on Part-of-Speech Analysis, *Information Processing and Management*, 2015, **51**(4), 458-479.

[29] Onan A., Korukoglu S., Bulut H., Ensemble of Keyword Extraction Methods and Classifiers in Text Classification, *Expert Systems with Applications*, 2016, **57**, 232-247.

[30] Breiman L., Random Forests, *Machine Learning*, 2001, **45**(1), 5-32.

[31] Yang Y., Pedersen J. O., A Comparative Study on Feature Selection in Text Categorization, *International Conference on Machine Learning*, Nashville, Tennessee, 8-12 Temmuz 1997.

[32] Forman G., An Extensive Empirical Study of Feature Selection Metrics for Text Classification, *Journal of Machine Learning Research*, 2003, **3**, 1289-1306.

[33] Mesleh A. A., Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System, *Journal of Computer* Science, 2007, **3**(6), 430-435.

[34] Zheng Z., Wu X., Srihari R., Feature Selection for Text Categorization on Imbalanced Data, *SIGKDD Explorations*, 2004, **6**(1), 80-89.

[35] Kanan H. R., Faez K., An Improved Feature Selection Method Based On Ant Colony Optimization (ACO) Evaluated On Face Recognition System, *Appl. Math. Comput.*, 2008, **205**(2), 716-725.

[36] Aghdam M. H., Ghasem-Aghaee N., Basiri M. E., Text Feature Selection using Ant Colony Optimization, *Expert Syst. Appl.*, 2009, **36**(3), 6843-6853.

[37] Al-Ani A., Feature Subset Selection using Ant Colony Optimization, *International Journal of Computational Intelligence*, 2005, **2**(1), 53-58.

[38] Goodarzi M., Freitas M. P., Jensen R., Ant Colony Optimization as a Feature Selection Method in the QSAR Modeling of Anti-HIV-1 Activities of 3-(3,5-Dimethylbenzyl) Uracil Derivatives using MLR, PLS and SVM Regressions, *Chemometr. Intell. Lab.*, 2009, **98**(2), 123-129.

[39] Liao S., Wang J., Yu R., Sato K., Cheng Z., CNN for Situations Understanding Based on Sentiment Analysis of Twitter Data, *Procedia Comput. Sci.*, 2017, **111,** 376–381.

[40] Santos C. N., Gatti M., Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts, *International Conference on Computational Linguistics*, Dublin, Ireland, 23-29 Ağustos 2014.

[41] Hu F., Li L., Zhang Z., Wang J., Xu X., Emphasizing Essential Words for Sentiment Classification Based on Recurrent Neural Networks, *J. Comput. Sci. Technol.*, 2017, **32**(4), 785–795.

[42] Chen Q., Guo Z., Sun C. , Li W., Research on Chinese Micro-Blog Sentiment Classification Based on Recurrent Neural Network, *Internatinal Conference on Computer Science and Technology*, Guilin, China, 26-28 Mayıs 2017.

[43] Zhao Z., Lu H., Cai D., He X., Zhuang Y., Microblog Sentiment Classification via Recurrent Random Walk Network Learning, *Internatinal Conference on Artificial Intelligence*, Melbourne, Australia, 19-20 Ağustos 2017.

[44] Becker W., Wehrmann J., Cagnini H. E. L., Barros R. C., An Efficient Deep Neural Architecture for Multilingual Sentiment Analysis in Twitter *Internatinal Conference on Florida Artificial Intelligence Research Society*, Marco Island, Florida, 22-24 Mayıs 2017.

[45] Uysal A. K., Murphey Y. L., Sentiment Classification: Feature Selection Based Approaches Versus Deep Learning, *IEEE International Conference on Computer and Information Technology*, Helsinki, Finland, 21-23 Ağustos 2017.

[46] Ghosal D., Bhatnagar S., Akhtar M. S., Ekbal A., Bhattacharyya P., IITP at Semeval-2017 Task 5: An Ensemble of Deep Learning and Feature Based Models for Financial Sentiment Analysis, *International Workshop on Semantic Evaluations*, Vancouver, Canada, 3-4 Ağustos 2017.

[47] Nozza D., Fersini E., Messina E., Deep Learning and Ensemble Methods for Domain Adaptation, *International Conference on Tools with Artificial Intelligence*, California, USA, 25-28 Temmuz 2016.

[48] Araque O., Corcuera-Platas I., Sánchez-Rada J. F., Iglesias C. A., Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications, *Expert Syst. Appl.*, 2017, **77**, 236-246.

[49] Onan A., Classifier and Feature Set Ensembles for Web Page Classification, *J. Inf. Sci*., 2015, **42**(2), 150-165.

[50] Abu-Errub A., Arabic Text Classification Algorithm using Tfidf and Chi Square Measurements, *Int. J. Comput. Appl.*, 2014, **93**(6), 40–45.

[51] Chauraisa V., Pal S., Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability, *International Journal of Computer Science and Mobile Computing*, 2014, **3**(1), 10-22.

[52] Neha A. G., A Novel Clustering Approach Based Sentiment Analysis of Social Media Data, *Int. J. Eng. Dev. Res.*, 2015, **3**(4), 1099-1107.

[53] Uysal A. K., Gunal S., The Impact of Preprocessing on Text Classification, *Inf. Process. Manag.*, 2014, **50**(1), 104–112.

[54] Rachburee N., Punlumjeak W., A Comparison of Feature Selection Approach Between Greedy, Ig-Ratio, Chi-Square, and Mrmr in Educational Mining,

International Conference on Information Technology and Electrical Engineering, Chiang Mai, Thailand, 8-10 Temmuz 2015.

[55] Siddiqui M. A., An Empirical Evaluation of Text Classification and Feature Selection Methods, *Artif. Intel. Res.*, 2016, **5**(2), 70–81.

[56] Zorarpacı E., Özel S. A., A Hybrid Approach of Differential Evolution and Artificial Bee Colony for Feature Selection, *Expert Syst. Appl.*, 2016, **62**, 91-103.

[57] Sudholt D., Thyssen C., Running Time Analysis of Ant Colony Optimization for Shortest Path Problems, *J. Discrete Algorithms*, 2012, **10**, 165-180.

[58] Panov P., Džeroski S., Combining Bagging and Random Subspaces to Create Better Ensembles, *International Conference on Intelligent Data Analysis*, Berlin, Heidelberg, 3-7 Aralık 2007.

[59] McCallum A., Nigam K., A Comparison of Event Models for Naive Bayes Text Classification, *AAAI-98 Workshop on Learning for Text Categorization*, Winconsin, USA, 26-27 Temmuz 1998.

[60] Schneider K. M., On Word Frequency Information and Negative Evidence in Naive Bayes Text Classification, *International Conference on Advances in Natural Language Processing*, Alacant, Spain, 20-22 Ekim 2004.

[61] Rennie J. D. M., Shih L., Teevan J., Karger D. R., Tackling the Poor Assumptions of Naive Bayes Text Classifiers, *International Conference on Machine Learning*, Washington, USA, 21-24 Ağustos 2003.

[62] Juan A., Ney H., Reversing and Smoothing the Multinomial Naive Bayes Text Classifier, *International Workshop on Pattern Recognition in Information Systems*, Alacant, Spain, 2-3 Nisan 2002.

[63] Eyheramendy S., Lewis D. D., Madigan D., On the Naive Bayes Model for Text Categorization, *International Workshop on Artificial Intelligence and Statistics*, Key West, Florida, 3-6 Ocak 2003.

[64] Kim S-B., Han K-S., Rim H-C., Myaeng S. H., Some Effective Techniques for Naïve Bayes Text Classification, *IEEE Transactions on Knowledge and Data Engineering*, 2006, **8**(11), 1457-1465.

[65] Peng F., Schuurmans D., Wang S., Language and Task Independent Text Categorization with Simple Language Models, *Human Language Technology Conference*, Edmonton, Canada, 27 Mayıs- 1 Haziran 2003.

[66] Peng F., Schurmans D., Combining Naïve Bayes and n-Gram Language Models for Text Classification, *European Conference on Information Retrieval Research*, Pisa, Italy, 14-16 Nisan 2003.

[67] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., The WEKA Data Mining Software An Update, *ACM SIGKDD Explorations Newsletter*, 2009, **11**(1), 10-18.

[68] Vilar D., Ney H., Juan A., Vidal E., Effect of Feature Smoothing Methods in Text Classification Tasks, *International Workshop Pattern Recognition in Information Systems*, Porto, Portugal, 14-17 Nisan 2004.

[69] Amasyalı M. F., Beken A., Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması, *IEEE 17th Signal Processing and Communications Applications Conference*, Antalya, Turkey, 9-11 Nisan 2009.

[70] Tantuğ A. C., Document Categorization with Modified Statistical Language Models for Agglutinative Languages, *International Journal of Computational Intelligence Systems*, 2010, **3**(5), 632-645.

[71] Saif H., Fernandez M., He Y., Alani H., Evaluation Datasets for Twitter Sentiment Analysis: A Survey and a New Dataset, The Sts-Gold, 2013.

[72] Dietterich T. G., Margineantu D. D., Dietterich Pruning Adaptive Boosting, *International Conference on Machine Learning*, San Francisco, USA, 26-29 Ağustos 1997.

[73] Lichman M., UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Science, http://archieve.ics.uci.edu/ml, (Ziyaret Tarihi: 15 Mayıs 2014).

[74] Lochter J. V., Zanetti R. F., Reller D., Short Text Opinion Detection using Ensemble of Classifiers and Semantic Indexing, *Expert Syst. Appl.*, 2016, **62,** 243–249.

KİŞİSEL YAYIN VE ESERLER

[1] Poyraz M., Ganiz M. C., Akyokus S., Gorener B., **Kilimci Z. H.**, Exploiting Turkish Wikipedia as a Semantic Resource for Text Classification, *International Symposium on Innovations in Intelligent Systems and Applications*, Trabzon, Türkiye, 2-4 Temmuz 2012.

[2] Poyraz M., **Kilimci Z. H.**, Ganiz M. C., A Novel Semantic Smoothing Method Based on Higher Order Paths for Text Classification, *IEEE International Conference on Data Mining*, Brussels, Belçika, 10-13 Aralık 2012.

[3] Poyraz M., **Kilimci Z. H.**, Ganiz M. C., Higher-Order Smoothing: A Novel Semantic Smoothing Method for Text Classification, *Journal of Computer Science and Technology*, 2014, **29**(3), 376-391.

[4] **Kilimci Z. H.**, Ganiz M. C., Evaluation of Classification Models for Language Processing, *International Symposium on Innovations in Intelligent Systems and Applications*, Madrid, İspanya, 2-4 Eylül 2015.

[5] **Kilimci Z. H.**, Akyokus S., N-Gram Pattern Recognition using Multivariate-Bernoulli Model with Smoothing Methods for Text Classification, *Signal Processing and Communication Application Conference*, Zonguldak, Türkiye, 16-19 Mayıs 2016.

[6] **Kilimci Z. H.**, Akyokus S., Omurca S. I., The Effectiveness of Homogenous Ensemble Classifiers for Turkish and English Texts, *International Symposium on Innovations in Intelligent Systems and Applications*, Sinaia, Romanya, 2-5 Ağustos 2016.

[7] **Kilimci Z. H.**, Akyokus S., Omurca S. I., The Evaluation of Heterogeneous Classifier Ensembles for Turkish Texts, *International Symposium on Innovations in Intelligent Systems and Applications*, Gdynia, Polonya, 3-5 Temmuz 2017.

[8] **Kilimci Z. H.**, Omurca S. I., A Comparison of Extended Space Forests for Classifier Ensembles on Short Turkish Texts, *Academic Conference on Engineering IT and Artificial Intelligence*, Prag, Çek Cumhuriyeti, 11-14 Ağustos 2017.

[9] **Kilimci Z. H.**, Omurca S. I., Enhancement of the Heuristic Optimization Based Extended Space Forests with Classifier Ensembles, *The International Arab Journal of Information Technology*, 2020, **17**(2), 160-168.

[10] **Kilimci Z. H.**, Omurca S. I., The Impact of Enhanced Space Forests with Homogeneous Classifier Ensembles, *International Journal of Intelligent Sytems and Applications in Engineering*, 2018, **6**(2), 25-32.

[11] **Kilimci Z. H.**, Omurca S. I., Akyokuş S., The Impact of Extended Space Forests with Heterogeneous Classifier Ensembles on Turkish Texts, International Conference on Artificial Intelligence and Data Processing, Malatya, Türkiye, 28-30 Eylül 2018. (Değerlendirme aşamasında)

ÖZGEÇMİŞ

1985 yılında Erzurum' da doğdu. İlk ve orta öğrenimini Kocaeli' nde, lise öğrenimini İzmir' de tamamladı. 2005 yılında girdiği Doğuş Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü' nden 2008 yılında Bilgisayar Mühendisi olarak mezun oldu. 2009-2011 yılları arasında, Denizbank A.Ş.' nin Veri Ambarı bölümünde Yazılım Mühendisi olarak çalıştı. 2011-2013 yılları arasında, Doğuş Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı' nda Yüksek Lisans öğrenimini tamamladı. 2011 yılından beri Doğuş Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği bölümünde Araştırma Görevlisi olarak görev yapmaktadır.