T.C.
HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

YÜKSEK BOYUTLU VERİ KÜMELERİNDEKİ ÖZNİTELİKLERİN HİBRİT YÖNTEMLE SEÇİLMESİ

Abdülkadir GÜMÜŞÇÜ

ELEKTRİK – ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

ŞANLIURFA
2016

Prof. Dr. Ramazan TAŞALTIN danışmanlığında Abdülkadir GÜMÜŞÇÜ'nün hazırladığı **'Yüksek Boyutlu Veri Kümelerindeki Özniteliklerin Hibrit Yöntemle Seçilmesi'** konulu bu çalışma 07/10/2016 tarihinde aşağıdaki jüri tarafından oy birliği ile Harran Üniversitesi Fen Bilimleri Enstitüsü Elektrik-Elektronik Mühendisliği Anabilim Dalı'nda DOKTORA TEZİ olarak kabul edilmiştir.

İmza

Danışman	: Prof.Dr. Ramazan TAŞALTIN	.....
Üye	: Doç.Dr. Mehmet Tahir GÜLLÜOĞLU	.....
Üye	: Yrd.Doç.Dr. İbrahim Berkan AYDİLEK	.....
Üye	: Yrd.Doç.Dr. Önder AYDEMİR	.....
Üye	: Yrd.Doç.Dr. Ahmet GÜNER	.....

Bu Tezin Elektrik-Elektronik Mühendisliği Anabilim Dalında Yapıldığını ve Enstitümüz Kurallarına Göre Düzenlendiğini Onaylarım.

Prof. Dr. Recep GÜNDOĞAN
Enstitü Müdürü

Bu çalışma HÜBAK Tarafından Desteklenmiştir.

Proje No: 15101

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge, şekil ve fotoğrafların kaynak gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

İÇİNDEKİLER

Sayfa No

ÖZET ii
ABSTRACT iii
TEŞEKKÜR iv
ŞEKİLLER DİZİNİ v
ÇİZELGELER DİZİNİ vi
SİMGELER DİZİNİ vii
1. GİRİŞ 1
2. ÖNCEKİ ÇALIŞMALAR 9
3. MATERYAL ve YÖNTEM 14
3.1. Materyal 14
3.1.1. Mikro-dizilim veri seti 14
3.2. Yöntem 16
3.2.1. Öznitelik seçme algoritmaları 17
3.2.1.1. Filtre modelli öznitelik seçme algoritmaları 17
3.2.1.2. Sarmal modelli öznitelik seçme algoritmaları 20
3.2.1.3. Gömülü modelli öznitelik seçme algoritmaları 22
3.2.2. Genetik algoritma 24
3.2.3. k-EK sınıflandırma algoritması 30
3.2.4. Birini dışarda bırak çapraz doğrulama (BDBÇD) 32
3.3. Önerilen Yöntem 33
4. ARAŞTIRMA BULGULARI ve TARTIŞMA 39
5. SONUÇLAR ve ÖNERİLER 56
KAYNAKLAR 59
ÖZGEÇMİŞ 63

ÖZET

Doktora Tezi

YÜKSEK BOYUTLU VERİ KÜMELERİNDEKİ ÖZNİTELİKLERİN HİBRİT YÖNTEMLE SEÇİLMESİ

Abdülkadir GÜMÜŞÇÜ

Harran Üniversitesi
Fen Bilimleri Enstitüsü
Elektrik-Elektronik Mühendisliği Anabilim Dalı

Danışman: Prof.Dr. Ramazan TAŞALTIN
Yıl: 2016, Sayfa: 64

Günümüzde mikro-dizilim veri setleri hastalık teşhisine önemli katkılar sağlamaktadır. Mikro-dizilim veri setlerini, makina öğrenme algoritmaları ile anlamlandırmak hasta sayısının azlığı gen sayısının fazlalığından ötürü oldukça çok zordur. Bu açıdan bakıldığında gen analizinde öznitelik seçme algoritmaları çok önemli bir işlem adımıdır. Literatürde genel olarak öznitelik seçme algoritmaları filtre, sarmal ve gömülü modeller olmak üzere 3 ana başlıkta incelenmektedir. Mikro-dizilim veri analizi için kullanılabilecek metotlar incelendiğinde; filtre modelli öznitelik seçme algoritmaları hızlı olmasına karşın her zaman istenilen başarı oranını sağlayamamaktadır, diğer taraftan sarmal modelli öznitelik seçme algoritmaları ise başarılı sonuçlar vermesine rağmen yavaş sonuçlar vermesi kullanım zorluğu yaşatmaktadır. Bahsedilen dezavantajları ortadan kaldırmak amacıyla bu tez çalışmasında filtre modelli öznitelik seçme algoritmalarının hızını, sarmal modelli öznitelik seçme algoritmalarının başarılı sonuçlarını harmanlayan hibrit bir öznitelik seçme algoritması önerilmiştir. Önerilen metodun filtre kısmında Ki-Kare, ReliefF ve F-Skor olmak üzere 3 farklı filtre modelli öznitelik seçme algoritması kullanıldı ve bu sonuçlar kombine edilerek genetik algoritmaya gönderildi. Genetik algoritma, filtrelenmiş veri setinin içinden en ideal veri setini seçmektedir. Seçilen final veri seti k-en yakın komşuluk (k-EK) sınıflandırma algoritması uygulanarak birini dışarıda bırak çapraz doğrulama (BDBÇD) ile değerlendirilmektedir. Önerilen metodun (CFR-GA) sınıflandırma başarı oranı, öznitelik seçme işlemi yapılmamış 7 sınıflandırma, 2 filtre modelli öznitelik seçme, 2 sarmal modelli öznitelik seçme ve 2 hibrit modelli öznitelik seçme algoritmaları ile kıyaslanmıştır. Deneysel sonuçlar, önerilen metodun kıyaslanan metotlara önemli ölçüde iyileştirmeler yaptığını göstermektedir.

ANAHTAR KELİMELER: Hibrit Öznitelik Seçme, Mikro-dizilim Sınıflandırma, Genetik Algoritma, Ki-Kare, ReliefF, F-Skor

ABSTRACT

PhD Thesis

HYBRID FEATURE SELECTION IN HIGH DIMENSIONAL DATA CLUSTERS

Abdülkadir GÜMÜŞÇÜ

Harran University
Graduate School of Natural and Applied Sciences
Department of Electrical and Electronic Engineering

Supervisor: Prof. Dr. Ramazan TAŞALTIN
Year: 2016, Page: 64

Nowadays, microarray data is an important contribution to the diagnosis of illness. In the way of machine learning classification, microarray data is a hard process task because of its large size of features and the low size of patient. In this regard feature selection is a pre-processing technique with great importance in microarray classification. In the literature, feature selection techniques in terms of classification can be examined under three titles. When we review the methods to be used for gene analysis; we can say that even though the filter methods are fast, they cannot reach to the required levels of success all the time and on the other hand, although wrapper models provide successful results, the slowness of this causes problems in utilization. The utilization of wrapper models in datasets with thousands of gene data such as microarray data is getting harder. In order to remove the disadvantages mentioned, a hybrid feature selection algorithm is suggested which combines the speed of filter models and the successful results of wrapper models. With the suggested method; three different filters of Chi-Squared, Fisher-Score and ReliefF (CFR) are implemented and this results are put together for genetic algorithm. Genetic algorithm chooses the most optimal feature set through filtered data set. This optimal feature set is evaluated with leave one out cross validation (LOOCV) by using k-nearest neighbor (k-NN) method. The classification success rate of the suggested method (CFR-GA) is compared with 7 classifications, 2 filter feature selection, 2 wrapper feature selection and 2 hybrid feature selection algorithms. The experimental results indicate that the suggested method has provided significant enhancements for the compared methods.

KEY WORDS: Hybrid feature selection, microarray classification, genetic algorithm, Chi-squared, ReliefF, F-Score

TEŞEKKÜR

Bu tezin hazırlanması aşamalarında bilgi ve birikimlerinden yararlandığım danışman hocam sayın Prof.Dr. Ramazan TAŞALTIN'a, tavsiyeleri ile bana yardım eden Doç.Dr. Burak ACAR'a, Prof.Dr. Fatih ALAGÖZ'e, Yrd.Doç.Dr. İbrahim Berkan AYDİLEK'e, Yrd.Doç.Dr. Önder AYDEMİR'e, Yrd.Doç.Dr. Ahmet GÜNER'e ve Doç.Dr. Mehmet Siraç ÖZERDEM'e ayrı ayrı teşekkür ederim.

Motivasyon ve teşviklerinden ötürü Doç.Dr. Mehmet Tahir GÜLLÜOĞLU'na, Yrd.Doç.Dr. Ümit ÜNCÜ'ye ve Arş.Gör. Mehmet Emin TENEKECİ'ye teşekkürü borç bilirim.

Bunların yanında manevi olarak beni hayatımın her anında destekleyen aile fertlerime ve eşime teşekkür ederim.

ŞEKİLLER DİZİNİ

Sayfa No

Şekil 1.1. Makina Öğrenme Algoritmalarının Genel Modeli 2
Şekil 1.2. Makina Öğrenme Algoritmalarının Sınıflandırılması 3
Şekil 1.3. Öznitelik Seçme Kullanılarak Sınıflandırma İşlem Adımları 5
Şekil 3.1. Filtre modelli öznitelik seçme işlem adımları 18
Şekil 3.2. Sarmal modelli öznitelik seçme işlem adımları 21
Şekil 3.3. Karar ağacı oluşturma işlem adımları 23
Şekil 3.4. Örnek bir karar ağacı yapısı 24
Şekil 3.5. Genetik Algoritma İşlem Adımları 26
Şekil 3.6. Tek Noktalı Çaprazlama İşlemi 27
Şekil 3.7. Çok Noktalı Çaprazlama İşlemi 28
Şekil 3.8. Çoklu Karışık Noktalı Çaprazlama İşlemi 28
Şekil 3.9. Noktasal Çaprazlama İşlemi 28
Şekil 3.10. Mutasyon İşlemi 29
Şekil 3.11. Eğitim veri setinin dağılımı 31
Şekil 3.12. k-EK modeline test verisi girişi 31
Şekil 3.13. Test verisinin komşuluk dağılımı 32
Şekil 3.14. CFR-GA öznitelik seçme algoritması temel işlem adımları 33
Şekil 3.15. Filtre modelli öznitelik seçme işlem adımı 34
Şekil 3.16. Eşikleme işlem adımı 35
Şekil 3.17. Birleştirme işlemi 35
Şekil 3.18. Sarmal modelli öznitelik seçme işlem adımı 36
Şekil 3.19. CFR-GA işlem adımları 37
Şekil 4.1. 9_Tumors veri seti GA işlem adımı – hata oranı grafiği 44
Şekil 4.2. 11_Tumors veri seti GA işlem adımı – hata oranı grafiği 45
Şekil 4.3. 14_Tumors veri seti GA işlem adımı – hata oranı grafiği 45
Şekil 4.4. Brain_Tumor1 veri seti GA işlem adımı – hata oranı grafiği 46
Şekil 4.5. Brain_Tumor2 veri seti GA işlem adımı – hata oranı grafiği 47
Şekil 4.6. Leukemia1 veri seti GA işlem adımı – hata oranı grafiği 47
Şekil 4.7. Leukemia2 veri seti GA işlem adımı – hata oranı grafiği 48
Şekil 4.8. Lung_Cancer veri seti GA işlem adımı – hata oranı grafiği 49
Şekil 4.9. Prostate_Tumor veri seti GA işlem adımı – hata oranı grafiği 49
Şekil 4.10. SRBCT veri seti GA işlem adımı – hata oranı grafiği 50
Şekil 4.11. DLBCL veri seti GA işlem adımı – hata oranı grafiği 50

ÇİZELGELER DİZİNİ

Sayfa No

Çizelge 1.1. Öznitelik Seçme Metotlarının Karşılaştırılması 6
Çizelge 3.1. Mikro-dizilim veri seti özellikleri..... 15
Çizelge 3.2. Mikro-dizilim veri setlerinin Gen Sayısı/Örnek Sayısı Oranları..... 15
Çizelge 3.3. Mikro-dizilim veri setlerinin açıklamaları..... 16
Çizelge 4.1. CFR filtre modelindeki filtrelerin sınıflandırma başarı oranı karşılaştırılması..... 40
Çizelge 4.2. CFR filtre modelindeki filtrelerin filtreleme süresi karşılaştırılması 41
Çizelge 4.3. CFR filtre modelindeki filtrelerin seçilen öznitelik sayısı karşılaştırılması 42
Çizelge 4.4. CFR-GA öznitelik seçme algoritmasının CFR ve GA ile sınıflandırma başarı oranları bakımından karşılaştırılması 42
Çizelge 4.5. CFR-GA öznitelik seçme algoritmasının CFR ve GA ile işlem süresi bakımından karşılaştırılması 43
Çizelge 4.6. CFR-GA hibrit modelli öznitelik seçme algoritmasının öznitelik seçilme yapılmadan sınıflandırma başarı oranları karşılaştırılması 51
Çizelge 4.7. CFR-GA hibrit modelli öznitelik seçme algoritmasının filtre modelli öznitelik seçme algoritmaları ile karşılaştırılması 52
Çizelge 4.8. CFR-GA hibrit modelli öznitelik seçme algoritmasının sarmal modelli öznitelik seçme algoritmaları ile karşılaştırılması 53
Çizelge 4.9. CFR-GA hibrit modelli öznitelik seçme algoritmasının hibrit modelli öznitelik seçme algoritmaları ile karşılaştırılması 54

SİMGELER DİZİNİ

DNA	Deoksirübo Nükleik asit
n	Veri setindeki örnek sayısı
p	Veri setindeki nitelik sayısı
DVM	Destek Vektör Makinası
YSA	Yapay Sinir Ağları
DVM – BKG	Destek Vektör Makinası – Bire Karşı Geri Kalanlar Yöntemi
DVM – BKB	Destek Vektör Makinası – Bire Karşı Bir Yöntemi
k-EK	k en yakın komşuluk Yöntemi
KÖS	Korelasyon-Tabanlı Öznitelik Seçme
İSS	İleri Sıralı Seçme Algoritması
CFR-GA	Önerilen Metod (Ki-Kare, F-Skor, ReliefF – Genetik Algoritma)
GSE	Geri Sıralı Eleme Algoritması
KKO	Karınca Kolonileri Optimizasyonu
GS	Gen Sayısı
ÖS	Örnek Sayısı
DVM – DZG	Destek Vektör Makinası – Direk Zincirli Grafik
DVM – WW	Destek Vektör Makinası – Weston ve Watkins
DVM – CS	Destek Vektör Makinası – Crammer ve Singer
BK	Bilgi Kazancı
OB	Ortak Bilgi
Gİ	Gini İndeksi
ÖS1	Ki-Kare Filtre Modelli Öznitelik Seçme Algoritması Sonuç Veri Seti
ÖS2	ReliefF Filtre Modelli Öznitelik Seçme Algoritması Sonuç Veri Seti
ÖS3	F-Skor Filtre Modelli Öznitelik Seçme Algoritması Sonuç Veri Seti
ÖS4	Birleştirme İşlemi Sonuç Veri Seti
SÖS	Seçilen Öznitelik Veri Seti
BK - GA	Bilgi Kazancı Öznitelik Seçme – Genetik Algoritma
KÖS – TGA	Korelasyon-Tabanlı Öznitelik Seçme – Taguchi Genetik Algoritma
İİPSO	İyileştirilmiş İkili Parçacık Sürü Optimizasyonu
BDBÇD	Birini Dışarıda Bırak Çapraz Doğrulama

1. GİRİŞ

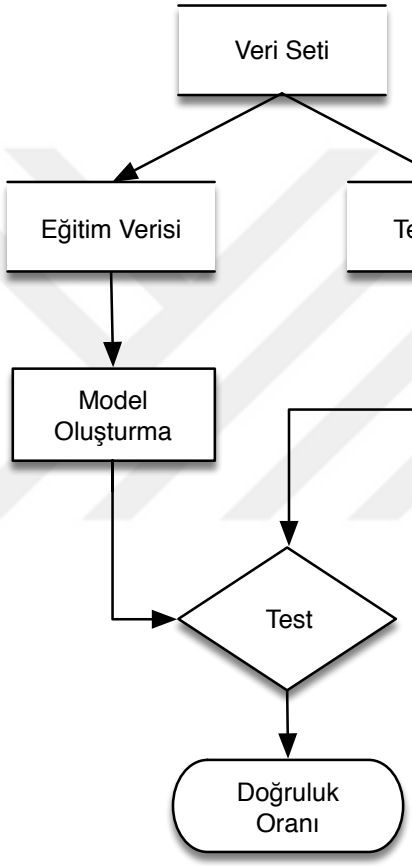
Günümüzde hastalık teşhisinde kalıtımsal araçlar çok yoğun kullanılmaya başlanmıştır. Kalıtımsal araçların arasında en çok kullanılan metotlarından biri de mikro-dizilimdir. Mikro-dizilim teknoloji ile genlerde meydana gelen değişiklikler belirlenebilmektedir. Deoksirübo nükleik asit (DNA) mikro-dizilim verileri ile kanserli hücrelerdeki gen değişimi karakterize edilebilmektedir. Bu değişimler kansere sebep olan genlerin belirlenmesinde önemli rol oynamaktadır. Kansere sebep olan genlerin belirlenmesi, veri işleme tekniklerine göre yapılmaktadır.

Bunun yanında patojenlerin tanılanmasında zahmetli, zaman alıcı ve masraflı olan geleneksel metotlar kullanılmaktadır. Buna ek olarak bazı patojenler pek çok seçici ortamda büyüyemediğinden tanılama için moleküler yaklaşımlar gerekmektedir (Yoltaş ve ark., 2010). Mikro-dizilim verileri bu metotların dezavantajlarının alt edilebilmesi için tanılayıcı olarak kullanılırlar (Kumar ve ark., 2005). Ayrıca moleküler tanıda artık biyosensör ve mikro-dizilim verilerinin birlikte kullanıldığı multidisipliner yaklaşımlar dikkat çekmeye başlamıştır (Cloarec ve ark., 2011). Mikro-dizilim veri setlerinin yaygınlaşması ile beraber patojenlerin tanımlanması hem daha kısa sürede hem de daha az zahmetle yapılabilecektir.

Mikro-dizilim verileri genellikle binlerce gen verisinin yanında düşük sayıda örnekten oluşmaktadır. Bunun nedeni mikro-dizilim veri setleri için gerekli olan deneysel şartların sağlanmasının zorluğu ve masraflı olmasıdır. Bunun yanında mikro-dizilim veri setleri nitelik bakımından çok fazla bilgi taşıması da makina öğrenme algoritmalarının başarı oranını önemli ölçüde olumsuz etkileyen durumlardan biri olan yüksek nitelik sayısı (p) düşük örnek sayısı (n) probleminin (Antoniadis ve ark., 2003) mikro-dizilim verileri içinde söz konusu olmasını sağlamaktadır. Genel olarak az sayıdaki ilişkili gen verisinin, tüm gen veri setine oranla daha yüksek korelasyon gösterdiği bilinmektedir (Chuang ve ark., 2011). Bunun anlamı binlerce gen arasından aslında sadece çok az sayıda istenilen probleme çözüm olacak bilgiye sahip olduğudur. Bu yüzden mikro-dizilim veri analizi

yaparken öznitelik seçme işleminin önemi çok büyük önem arz etmektedir. Öznitelik seçme algoritmalarının sınıflandırma başarı oranını önemli ölçüde olumlu etkilediği görülmektedir.

Makina öğrenmesinin konularından olan sınıflandırma algoritmaları genel olarak gözetimli ve gözetimsiz sınıflandırma olarak ikiye ayrılmaktadır. Şekil 1.1.'de makina öğrenme algoritmalarının genel kullanım modeli sunulmuştur.

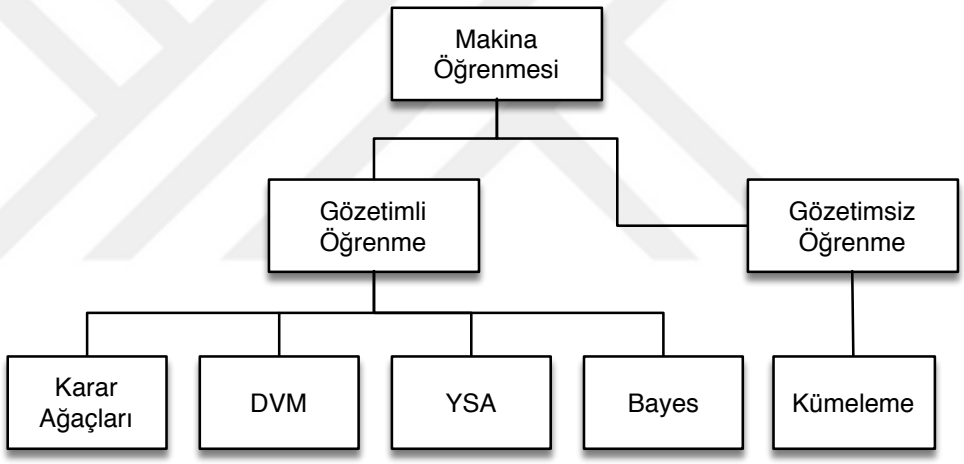


Şekil 1.1. Makina Öğrenme Algoritmalarının Genel Modeli

Makina öğrenme algoritmalarında Şekil 1.1.'de gösterildiği gibi öncelikle veri seti eğitim verisi ve test verisi olmak üzere ikiye ayrılır. Eğitim verisi ile bundan sonra girişine verilecek bilgilere cevap verecek bir model oluşturulur. Oluşturulan modelin sınıflandırma başarı oranını belirlemek amacıyla daha önce veri setinden ayrılan test verisi kullanılır. Sınıflandırma başarı oranı, test verisi modele

uygulanarak gerçek sonuçlar ile modelin tahmin ettiği sonuçlar kıyaslanarak hesaplanır.

Gözetimli sınıflandırma algoritmaları ile gözetimsiz sınıflandırma algoritmaları arasındaki en büyük fark gözetimli sınıflandırma algoritmalarında örneklerin sonuçları bilinerek model oluşturulmakta, fakat gözetimsiz sınıflandırma algoritmalarında örneklerin sonuçları bilinmeden sınıflandırma yapılmaktadır. Bu yüzden gözetimsiz sınıflandırma algoritmaları genelde kümeleme adı altında incelenmektedir. Genellikle gözetimli sınıflandırma algoritmaları, gözetimsiz sınıflandırma algoritmalarına göre daha iyi başarı oranı vermektedir. Makina öğrenme algoritmalarının nasıl sınıflara ayrıldığı Şekil 1.2.'de detaylı şekilde gösterilmiştir.



Şekil 1.2. Makina Öğrenme Algoritmalarının Sınıflandırılması

Şekil 1.2.'de de gösterildiği üzere makina öğrenme algoritmaları ilk aşamada gözetimli ve gözetimsiz diye iki gruba ayrılmıştır.

Gözetimsiz öğrenme algoritmaları literatürde kümeleme alt başlığı ile incelenebilmektedir.

Kümeleme algoritmaları, temel amacı veri setlerindeki örnekleri sahip oldukları istatistiksel özellikleri baz alarak gruplamak olan çok değişkenli teknikler grubudur. Kümeleme analizi, nesneleri küme içerisinde çok benzer biçimde, kümeler arasında farklı olacak biçimde kümelemeyi amaçlar (Xu ve ark., 2010).

Gözetimli öğrenme algoritmaları da kendi içerisinde Karar Ağaçları, Destek Vektör Makinaları (DVM), Yapay Sinir Ağları (YSA) ve Bayes algoritmaları olmak üzere dörde ayrılmaktadır.

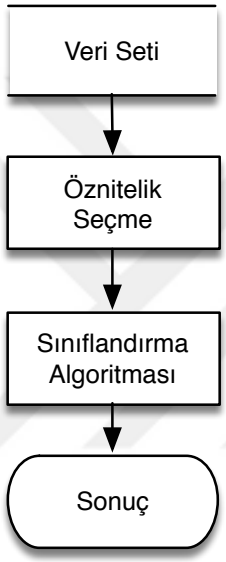
Karar ağaçları öğrenmesinde, eğitim kümesindeki bilgiler ile bir karar verme ağaç yapısı oluşturulur. Bu yapı yeni bir bilginin sahip olduğu niteliklere göre hangi sınıfa ayrıldığını bu ağaç yapısı ile karar verecektir (Lee ve ark., 2005). En büyük avantajı anlaşılması ve yorumlanması kolay olmasıdır. En çok kullanılan karar verme ağaçları algoritmaları: C4.5, RandomForest, ID3 (Pirooznia ve ark., 2008) şeklinde sıralanabilir.

Destek Vektör Makinaları (DVM) sınıflandırma konusunda kullanılan oldukça etkili yöntemlerden birisidir. Sınıflandırma için bir düzlemde bulunan iki grup arasında bir sınır çizilerek iki grubu ayırmak mümkündür. Bu sınırın çizileceği yer ise iki grubun da üyelerine en uzak olan yer olmalıdır. DVM bu sınırın nasıl çizileceğini belirleyen algoritmalardır (Molla ve ark., 2004). Literatürde en çok kullanılan DVM algoritmaları k-en yakın komşuluk (k-EK), Destek Vektör Makinası – Bire Karşı Geri Kalanlar Yöntemi (DVM-BKG) ve Destek Vektör Makinası – Bire Karşı Bir Yöntemi (DVM-BKB) algoritmalarıdır (Statnikov ve ark., 2005)

Yapay Sinir Ağları (YSA) ağırlıklandırılmış şekilde birbirlerine bağlanmış birçok işlem biriminden (nöronlar) oluşan matematiksel sistemlerdir. Bir işlem birimi, aslında sık sık transfer fonksiyonu olarak anılan bir denklemdir. Bu işlem birimi, diğer nöronlardan sinyalleri alır; bunları birleştirir, dönüştürür ve sayısal bir sonuç ortaya çıkartır. Genelde, işlem birimleri kabaca gerçek nöronlara karşılık gelirler ve bir ağ içinde birbirlerine bağlanırlar. Verilen eğitim verisi ile bu model oluşturulur ve test verileri ilgili modele uygulanarak sınıflandırma yapılabilir.

Bu tez çalışmasında kullanılan mikro-dizilim veri setindeki örneklerin sonuçlarına göre sınıflandırma modeli oluşturulduğundan gözetimli sınıflandırma kullanılmıştır. Literatürde mikro-dizilim veri setleri için kullanılan makina öğrenmesi k-EK algoritması olduğundan bu çalışmada da önerilen metodun diğer metotlarla kıyaslanması daha güvenilir olması amacı ile k-EK algoritması kullanılarak sınıflandırma başarı oranları hesaplanmıştır.

Makina öğrenme algoritmalarının uygulanacağı veri setinin özellikleri sınıflandırma başarı oranını önemli ölçüde etkilemektedir. Örnek sayısının az olması nitelik sayısının fazla olması veri seti üzerinden başarılı bir sınıflandırma modeli oluşturulmasını zorlaştırmaktadır. Bu yüzden veri setleri genellikle sınıflandırma algoritmalarına uygulanmadan önce öznitelik seçme işlemleri yapılarak, örnek sayısı ve nitelik sayısı arasındaki oranı daha ideal bir duruma getirilmesi planlanmaktadır. Şekil 1.3.'de öznitelik seçme işlemi kullanılarak sınıflandırma algoritmalarının işlem adımları belirtilmiştir.



Şekil 1.3. Öznitelik Seçme Kullanılarak Sınıflandırma İşlem Adımları

Sınıflandırma bağlamında öznitelik seçme teknikleri filtre, sarmal ve gömülü olmak üzere üç ana başlıkta incelenebilir. Bu üç temel öznitelik seçme metodu Çizelge 1.1.'de özetlenmiştir.

Çizelge 1.1. Öznitelik Seçme Metotlarının Karşılaştırılması

Seçme Modeli	Avantajı	Örnek Algoritmaları
Filtre Metotlar	Hızlı Sınıflandırma Algoritmasından Bağımsız	Kazanç Oranı (Ben-Bassat, 1982) Korelasyon-tabanlı Öznitelik Seçme (KÖS) (Hall, 1999)
Sarmal Metotlar	Başarı Oranı Yüksek Uygulanması Kolay Anlaşılır	İleri Sıralı Seçme (İSS) (Kittler, 1978) Rastgele Tırmanma Metodu (Skalak, 1994)
Gömülü Metotlar	Başarı Oranı Yüksek İşlem Yükü Az	Karar Ağaçları (Pirooznia ve ark., 2008)

Gömülü arama metotları, zaten sınıflandırma algoritmasının yürütülmesinde bulunan öznitelik seçme algoritmalarıdır. Karar ağaçları uygulanırken ilk hangi niteliğe göre ayrım yapılacağının belirlenmesi aslında en çok bilgiye sahip niteliği seçme işlemidir. Bu yüzden tek başına bir öznitelik seçme algoritması olarak değerlendirilmemektedir. Gömülü öznitelik seçme algoritmalarının bağımsız olarak öznitelik seçme işlemi yapamaması hibrit modellerde de kullanılmasını engellemektedir.

Filtre metotlarında, veri setinde bulunan nitelikler arasında istatistiksel metotlar kullanılarak nitelikler, sonuç sınıfı ile ilişkilendirilir. Bir niteliğin sınıf türü ile ne kadar ilişkisi kuvvetli ise ilgili nitelik taşıdığı bilgi açısından değerli bir bilgi olarak nitelendirilir. Filtre metotlarının en büyük avantajı hızlı olmasıdır. Özellikle mikro-dizilim veri seti gibi nitelik sayısı fazla olan veri setleri için hız çok önemli bir faktördür.

Sarmal metotlar ise rastgele veya belirli bir düzen ile seçilen öznitelikleri sınıflandırma algoritmasına uygulayarak başarı oranı hesaplamaktadır. Rastgele veya belli bir düzende seçilen özniteliklerin en iyi sınıflandırma başarı oranı veren öznitelik grubu seçilerek başarılı bir seçme işlemi yapılmaktadır. Bu tür arama modellerinin avantajı ise başarı oranının yüksek olmasıdır. Fakat bu çalışmada

kullanılacak veri setinin nitelik sayısının fazla olması sarmal metotlu öznitelik seçme algoritmalarının uygulanmasını zorlaştırmaktadır.

Özetle filtre öznitelik seçme algoritmaları hızlı olmasına karşın her zaman istenilen başarı oranını sağlayamamaktadır. Bunun yanında sarmal öznitelik seçme algoritmaları ise istenilen başarı oranını vermesine karşın işlem yoğunluğundan dolayı hızlı sonuç verememektedir. Yapılan bu tez çalışmasıyla, mikro-dizilim veri setlerinin sınıflandırılmasında kullanılabilecek öznitelik seçme metotlarından bahsedilen dezavantajları ortadan kaldırmak amacıyla filtre modellerinin hızını, sarmal modellerinin başarılı sonuçlarını harmanlayan hibrit bir öznitelik seçme algoritması önerilmiştir. Önerilen metodun filtre kısmında Ki-Kare (Liu ve ark., 1995), ReliefF (Robnik-Sikonja ve ark., 2003) ve F-Skor (Duda ve ark., 2001) olmak üzere 3 farklı filtre kullanıldı ve bu sonuçlar kombine edilerek Genetik algoritmaya gönderildi. Genetik algoritma, filtrelenmiş veri setinin içinden en ideal veri setini seçmektedir. Seçilen final veri seti k-EK sınıflandırma algoritması uygulanarak birini dışarıda bırak çapraz doğrulama (BDBÇD) ile değerlendirilmektedir. Önerilen metodun CFR-GA (Chi-Square, F-skor, ReliefF – Genetic Algorithm) sınıflandırma başarı oranı, öznitelik seçme işlemi yapılmamış 7 sınıflandırma, 2 filtre modelli öznitelik seçme, 2 sarmal modelli öznitelik seçme ve 2 hibrit modelli öznitelik seçme algoritmaları ile kıyaslanmıştır. Deneysel sonuçlar, önerilen metodun kıyaslanan metotlara önemli ölçüde iyileştirmeler yaptığını göstermektedir.

Bu bölümde; mikro-dizilim veri seti, makina öğrenme algoritmaları ve öznitelik seçme algoritma konuları özetlenerek tez çalışma konusu ve sunulan çözüm hakkında bilgi verilmiştir. Önceki Çalışmalar bölümünde literatürde geçen daha önceden önerilen öznitelik seçme algoritmaları ve özelleşmiş olarak hibrit algoritmalar incelenecektir. Materyal ve Yöntem bölümünde ise mikro-dizilim veri seti, öznitelik seçme algoritmaları, genetik algoritma, k-EK sınıflandırma algoritması ve birini dışarıda bırak çapraz doğrulama (BDBÇD) metotları detaylı şekilde ele alınacaktır ve tez çalışması kapsamında yüksek p düşük n parametresine sahip veri setleri için önerilen hibrit metodun algoritması açıklanacaktır. Araştırma Bulguları ve Tartışma bölümünde ise önerilen metodun mikro-dizilim veri setine uygulandığında

çıkan sonuçlar incelenerek, literatürdeki mevcut algoritmalara sağlanan iyileştirmeler belirtilecektir. Sonuçlar ve Öneriler kısmında ise çalışmaya eklenebilecek iyileştirmeler irdelenecektir.

2. ÖNCEKİ ÇALIŞMALAR

Sınıflandırma bağlamında öznitelik seçme teknikleri üç ana başlıkta incelenebilir. Bunlar filtre modelli öznitelik seçme algoritmaları, sarmal modelli öznitelik seçme algoritmaları ve gömülü modelli öznitelik seçme algoritmalarıdır. Filtre modelli öznitelik seçme algoritmalarında istatistiksel varsayımlar kullanılarak niteliklerin sonuca ulaşmaya katkısı hesaplanır. Filtre modelli öznitelik seçme algoritmaları sınıflandırma algoritmalarından bağımsız çalışır.

Filtre modelli öznitelik seçme algoritmalarının avantajları hızlı ve sınıflandırma algoritmalarından bağımsızlık şeklinde özetlenebilir (Saeys ve ark., 2007). En çok kullanılan filtre modelli öznitelik seçme algoritmaları Korelasyon-bazlı öznitelik seçme (KÖS) (Hall, 1999), bilgi kazancı (BK) (Quinlan, 1986), ortak bilgi algoritması (OB) (Battiti, 1994), Ki-kare (Liu ve ark., 1995), ReliefF (Robnik-Sikonja ve ark., 2003) ve F-Skor (Duda ve ark., 2001) şeklinde sıralanmaktadır. Korelasyon-bazlı öznitelik seçme algoritması nitelikler ile sınıf arasında korelasyonları hesaplayarak korelasyonları yüksek olan nitelikleri öznitelik olarak seçmektedir. Bilgi kazancı öznitelik seçme işleminde niteliğin ilgili niteliğin bilgiye ne kadar oranda katkı sağladığı stokastik temellerine dayandırarak öznitelikleri seçmektedir. Ki-kare metodu ise bulundukları sınıfa göre tüm niteliklerin Ki-kare'sini hesaplayarak tek tek değerlendirir. ReliefF öznitelik seçme algoritmasına göre incelenen nitelik değerli bir nitelik ise, niteliğe en yakın mesafesinde aynı sınıftan, niteliğe uzak kalan kısımda ise farklı sınıflardan örnekler bulunmalıdır. F-skor öznitelik seçme algoritması ise her niteliğin diğer niteliklerden farkını hesaplayan bir filtre modelidir.

Sarmal modelli öznitelik seçme algoritmaları ise sınıflandırma başarı oranını yükseltme amacıyla oluşturulan nitelik gruplarını sınıflandırma algoritmasında deneyerek en ideal öznitelikleri bulma yöntemleridir. Sarmal modelli öznitelik seçme algoritmalarının başarı oranının yüksek olmasının yanında özellikle nitelik sayısı büyük olan veri setlerinde işlem süresi çok uzamaktadır (Kohavi ve ark., 1997). İleri

Sıralı Seçme Algoritması (İSS) (Kittler, 1978), Geri Sıralı Eleme Algoritması (GSE) (Kittler, 1978) ve GA (Holland, 1975) kullanılan başlıca sarmal yöntemleridir. İSS sarmal modelli öznitelik seçme algoritmasında seçilen öznitelik kümesi başlangıçta boş küme olarak alınır. Daha sonra sıralı şekilde tüm niteliklerin oluşturduğu çözüm kümeleri oluşturulur. Bu kümelerin başarılı oranları karşılaştırılarak en başarılı çözüm kümesi seçilir. Daha sonra bu işlemler seçilen öznitelik sabit olmak koşuluyla tekrarlanır. İstenilen başarı oranı elde edilenince döngü sonlandırılır. İSS ise GSE'nin tersi işlemleri yapmaktadır. GSE yönteminde çözüm kümesi tüm nitelikler olarak alınıp eliminasyon yapılarak değersiz nitelikler çıkarılır. GA yönteminde ise genetik algoritma tabanlı seçme işlemleri yapılarak istenilen sınıflandırma başarı oranına ulaşılmaya çalışılır.

Gömülü modelli öznitelik seçme algoritmaları ise sınıflandırma algoritmasına dahil olan öznitelik seçme yöntemlerindendir. Bu algoritmalar zaten sınıflandırma yaparken öznitelik seçmek zorundadır. Bu metoda ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993) karar ağaçları örnek verilebilir. Karar ağacı oluşturulurken ilk önce hangi niteliğe göre karar ağacının dallanacağına karar verilmesi gerekmektedir. İlk dallanmasına karar verilen nitelik aslında diğer niteliklere göre daha fazla sonuçla ilişkilidir. Bu işlemde aslında bir öznitelik seçme işlemi olarak değerlendirilmektedir.

Sonuç olarak yüksek p düşük n parametrelerine sahip veri setleri analizi için kullanılabilecek yöntemler incelendiğinde; filtre modelli öznitelik seçme algoritmaları hızlı olmasına karşın her zaman istenilen başarı oranını sağlayamamaktadır, diğer taraftan sarmal modelli öznitelik seçme algoritmaları ise başarılı sonuçlar vermesine rağmen yavaş sonuçlar vermesi kullanım zorluğu yaşatmaktadır. Özellikle mikro-dizilim verileri gibi binlerce gen verisine sahip veri setlerinde sarmal modelli öznitelik seçme algoritmalarının kullanımı oldukça zorlaşmaktadır. Bahsedilen dezavantajları ortadan kaldırmak amacıyla bu tez çalışmasında filtre modelli öznitelik seçme algoritmalarının hızını, sarmal modelli öznitelik seçme algoritmalarının başarılı sonuçlarını harmanlayan hibrit bir öznitelik seçme algoritması önerilmiştir.

Veri madenciliği ve bu bağlamda öznitelik algoritmalarının klinik çalışmalarında en fazla kullanıldığı alanlardan biri de mikro-dizilim veri teknolojisidir. Bir insanın genomunda yaklaşık 40.000 gen olduğu düşünüldüğünde bu kadar fazla verinin anlamlandırılması çok zorlu bir işlem halini almaktadır. Literatürde uygulanan destek vektör makinaları ve karar ağaçları sınıflandırma çalışmaları yeterince başarılı sonuçlar vermemiştir (Parrish ve ark., 2009). Bu yüzden literatürde birçok gen seçme yöntemleri önerilmiştir (Chin ve ark., 2005) (Li ve ark., 2001). Sebban ve Nock (Sebban ve Nock, 2002) başlangıçta filtre modelli öznitelik seçme algoritmasını uygulayarak elde edilen öznitelikleri karar ağaçlarında sarmal modelli öznitelik seçme algoritması ile öznitelikleri seçme işlemini uygulamıştır. Alshamlan vd. (Alshamlan ve ark., 2015) genetik algoritma ve arı koloni algoritmalarını kombine ederek yeni bir hibrit öznitelik seçme algoritması önermişlerdir. (Alshamlan ve ark., 2015)' da önerilen metot zaten başarı oranı yüksek olan veri setlerinde test edilmiş olup mevcut başarı oranına katkı sağlamamıştır. Önerilen metot daha çok seçilen öznitelik sayısını azaltmayı hedeflemektedir. Bu tez kapsamında önerilen hibrit öznitelik seçme algoritması sınıflandırma başarı oranı düşük veri setlerinde de test edilmiştir. Sınıflandırma başarı oranı düşük veri setlerine önemli katkılar sağlamıştır. Li vd. (Li ve ark., 2013) karınca koloni optimizasyon yöntemi ile öznitelik seçme yöntemi önermiştir. Bu çalışmada beş farklı sınıflandırma algoritması ile önerilen yöntem hem sınıflandırma başarı oranı hem de seçilen öznitelik sayısı bakımından karşılaştırılmıştır. (Li ve ark., 2013)'daki önerilen yöntem sınıflandırma başarı oranı bakımından CFR-GA hibrit öznitelik seçme algoritması ile kıyaslandığında CFR-GA daha başarılı sonuçlar vermiştir. Muhammad Atif Tahir vd. (Tahir ve ark., 2007) k-EK sınıflandırma algoritması ile beraber çalışan Tabu Arama metodu ile sarmal bir öznitelik seçme yöntemi önermiştir. Önerilen yöntem prostat kanser görüntülerini tanıma probleminde test edilmiştir. Huang vd. (Huang ve ark., 2007) karşılıklı bilgi yöntemini genetik algoritmanın içine gömerek yeni bir sarmal modelli öznitelik seçme algoritması önermiştir. Önerilen yöntem Kaliforniya Üniversitesi veri setlerinden biri olan Wine veri seti ile test edilmiştir. Hsu vd (Hsu ve ark., 2011) ise sarmal modelli öznitelik seçme algoritmaları ile filtre modelli öznitelik seçme algoritmalarını kombine ederek hibrit modelli bir öznitelik seçme algoritması

önermiştir. Önerilen yöntemin başarı oranı DVM sınıflandırma algoritması ile test edilmiştir. Baison vd. (Baison ve ark., 2015) dalgacık dönüşümü tabanlı bir öznitelik modeli önermiştir. Bu model dalgacıklar metodu ile öznitelik çıkarımı yapıp, çıkarılan öznitelikleri k-EK, Bayes ve DVM'den oluşan hibrit bir sınıflandırma algoritmasında değerlendirmiştir. Tabakhi vd. (Tabakhi ve ark., 2015) ise karınca koloni optimizasyon (KKO) teknikleriyle öznitelik çıkarımı önermiştir. Yukarıda bahsedilen tüm hibrit öznitelik seçme algoritmaları tek bir filtre modelli ve tek bir sarmal modelli öznitelik seçme algoritmalarının kombine edilmesi ile gerçekleştirilmiştir. Bu tez kapsamında önerilen yöntemin diğer hibrit yöntemlerden farklı olarak filtreleme aşamasında tek bir filtre modelli değil de üç farklı filtre modelli öznitelik seçme algoritması kullanılmasıdır. Bu sayede tek bir filtre modelli öznitelik seçme algoritmasının baz aldığı tek bir parametreden farklı olarak üç farklı parametre filtre modelli öznitelik seçme aşamasında değerlendirilmiştir. Chuang vd. (Chuang ve ark., 2011), (Yang ve ark., 2009), (Chuang ve ark., 2008) hibrit modelli öznitelik seçme algoritmaları üzerine çalışmalar yürütmüşlerdir. Yang vd. (Yang ve ark., 2009) Bilgi kazancı ve genetik algoritmayı birleştirerek hibrit modelli bir öznitelik seçme algoritması önermiştir. Önerdiğimiz metot (Yang ve ark., 2009)'deki hibrit modelli öznitelik seçme algoritmasından daha başarılı sonuçlar vermiştir. Chuang vd. (Chuang ve ark., 2008) daha sonra İyileştirilmiş İkili Parçacık Sürü Optimizasyonu (İİPSO) temelli hibrit bir öznitelik seçme algoritması önermiş ve daha başarılı sonuçlar almıştır. Chuang vd. (Chuang ve ark., 2011) Korelasyon-bazlı öznitelik seçme (KÖS) ile Taguchi - GA'ı kombine ederek önerdiği hibrit modelli öznitelik seçme algoritması ile mikro-dizilim veri setlerinde kayda değer iyileştirmeler yapmıştır. (Chuang ve ark., 2011)'de önerilen hibrit modelli öznitelik seçme algoritması ile önerdiğimiz hibrit modelli öznitelik algoritması kıyaslandığında, önerdiğimiz algoritma mevcut başarı durumuna önemli katkılar yapmaktadır. Bolón-Canedo vd. (Bolón-Canedo ve ark., 2014) literatürde kullanılan öznitelik seçme algoritmalarının mikro-dizilim verilerinde uygulanmasını özetleyerek sınıflandırma başarı oranlarını araştırmacılara sunmaktadır. (Lee ve ark., 2004)'de ise 7 farklı mikro-dizilim veri setinin, 3 farklı gene seçme algoritması ile 21 sınıflandırma algoritmasındaki performansları ele alınmıştır.

Önerilen hibrit modellemiz üç aşamadan oluşmaktadır. İlk aşama filtreleme işlemidir. Önerilen metot ile 11 mikro-dizilim veri setine Ki-Kare, F-Skor ve ReliefF olmak üzere 3 farklı filtre uygulanmaktadır. Filtrelemenin sonucunda birbirinden farklı 3 adet öznitelik kümesi oluşmaktadır. İkinci aşama ise, bu 3 farklı öznitelik kümesini harmanlayarak bir sonraki işlem adımı için öznitelik kümesini hazırlayan birleştirme işlemidir. Bu işlem aslında 3 filtrenin de önemli bulduğu öznitelikleri kombine etmektedir. Sadece bir filtre tarafından seçilen öznitelikler bu işlem adımında elenmektedir. Üçüncü ve en son aşama ise birleştirme aşamasında elde edilen öznitelik kümesine Genetik algoritma (GA) modeli uygulanarak final öznitelik veri setinin elde edildiği sarmal öznitelik seçme işlem adımıdır. Genetik algoritma öznitelik uzayından rastgele öznitelikler seçerek başarı oranı en iyi optimal öznitelik setini seçmektedir. Çaprazlama ve mutasyon operatörleri optimal özniteliklerin daha geniş bir öznitelik seti uzayından seçilmesine yardımcı olmaktadır. Seçilen bu optimal öznitelik seti, k en yakın komşuluk (k-EK) sınıflandırma algoritması kullanılarak BDBÇD ile değerlendirilmiştir. Önerilen metodun (CFR-GA) sınıflandırma başarı oranı, öznitelik seçme işlemi yapılmamış 7 sınıflandırma, 2 filtre modelli öznitelik seçme, 2 sarmal modelli öznitelik seçme ve 2 hibrit modelli öznitelik seçme algoritmaları ile kıyaslanmıştır. Deneysel sonuçlar, önerilen metodun kıyaslanan yöntemlere önemli ölçüde iyileştirmeler yaptığını göstermektedir.

3. MATERYAL ve YÖNTEM

3.1. Materyal

Bu bölümde çalışmada kullanılan materyal olan mikro-dizilim veri seti detaylı şekilde ele alınacaktır.

3.1.1. Mikro-dizilim veri seti

DNA mikro-dizilim teknolojisi, dokulardaki gen ifade profillerindeki global değişikliklerin incelenmesinde kullanılan yeni ve güçlü bir teknolojidir (Bednar, 2000).

DNA mikroarray teknolojisinin, birçok genin aktivitesinin aynı zamanda izlenebilmesi; hızlı bir yöntem olması; hasta ve sağlıklı hücrelerdeki genlerin aktivitelerinin karşılaştırılmasını sağlaması gibi avantajları bulunmaktadır (Sassanfar ve ark., 2003).

Çok sayıda gen verisine ve az sayıda örneğe sahip olması sebebiyle mikro-dizilim veri setleri, halen bir çok araştırmacı için en ideal öznitelik veri setini seçme problemi adına bir çalışma konusudur. Bu tez çalışmasında http://www.gems-system.org (Statnikov ve ark., 2005) sitesinden indirilen veriler kullanılmıştır. Belirtilen sitede bulunan 9_Tumors, 11_Tumors, 14_Tumors, Brain_Tumor1, Brain_Tumor2, Leukemia1, Leukemia2, Lung_Cancer, SRBCT, Prostate_Tumors, DLBCL adında toplam 11 adet veri setine önerilen CFR-GA öznitelik seçme algoritması uygulanmıştır. Veri setlerine ait örnek sayısı, sınıf sayısı, gen sayısı Çizelge 3.1.'de özetlenmiştir.

Çizelge 3.1. Mikro-dizilim veri seti özellikleri

Veri Seti	Sınıf Sayısı	Örnek Sayısı	Gen Sayısı
9_Tumors	9	60	5726
11_Tumors	11	174	12533
14_Tumors	26	308	15009
Brain _Tumor1	5	90	5920
Brain _Tumor2	4	50	10367
Leukemia1	3	72	5327
Leukemia2	3	72	11225
Lung_Cancer	5	203	12600
SRBCT	4	83	2308
Prostate_Tumor	2	102	10509
DLBCL	2	77	5469

Çizelge 3.1.'de de görüldüğü üzere her mikro-dizilim veri seti farklı sınıf sayısına, farklı örnek sayısına ve farklı gen sayılarına sahiptir. Brain_Tumor2 veri seti, 10367 gen bilgisine sahip ve buna karşın sadece 50 örnek sayısına sahip olduğu göz önünde bulundurulduğunda yüksek gen sayısı düşük örnek sayısı probleminin en iyi örneği olarak kabul edilebilmektedir. Veri setlerinin örnek sayısı ve gen sayısı arasındaki oranlar Çizelge 3.2.'de verilmiştir.

Çizelge 3.2. Mikro-dizilim veri setlerinin Gen Sayısı/Örnek Sayısı Oranları

Veri Seti	Örnek Sayısı	Gen Sayısı	GS/ ÖS
9_Tumors	60	5726	95
11_Tumors	174	12533	72
14_Tumors	308	15009	49
Brain _Tumor1	90	5920	66
Brain _Tumor2	50	10367	207
Leukemia1	72	5327	74
Leukemia2	72	11225	156
Lung_Cancer	203	12600	62
SRBCT	83	2308	28
Prostate_Tumor	102	10509	103
DLBCL	77	5469	71

Çizelge 3.2.'de görüldüğü üzere GS/ÖS oranı en yüksek olan veri seti Brain_Tumor2 mikro-dizilim veri setidir. GS/ÖS oranı en düşük olan veri seti ise SRBCT mikro-dizilim veri setidir. GS/ÖS oranı sınıflandırma algoritmalarının başarı oranını ciddi oranda etkileyen faktörlerden biridir. Fakat bu oran tek başına bir ölçüt sayılmamaktadır. Veri seti içindeki sınıf sayısı, veriler arasındaki korelasyonlar gibi faktörler de başarı oranını etkilemektedir. Tez çalışmasında kullanılan veri setlerinin sınıflandırma algoritmalarındaki başarı oranları aşağıdaki grafiklerde gösterilmiştir.

Çizelge 3.3. Mikro-dizilim veri setlerinin açıklamaları

Veri Seti	Açıklaması
9_Tumors	9 Çeşit İnsan Tümör Tipi
11_Tumors	11 Çeşit İnsan Tümör Tipi
14_Tumors	14 Çeşit İnsan Tümör Tipi ve 12 Normal Doku
Brain _Tumor1	5 Çeşit Beyin Tümör Tipi
Brain _Tumor2	4 Çeşit Beyin Tümör Tipi
Leukemia1	AML, ALL B-cell ve ALL T-cell
Leukemia2	AML, ALL, MLL
Lung_Cancer	4 Çeşit Akciğer Tümör Tipi ve Normal Doku
SRBCT	4 Çeşit Hücre Tümörü
Prostate_Tumor	Prostat Tümörü ve Normal Doku
DLBCL	2 Çeşit Hücre Tümörü

Çizelge 3.3.'de bu tez kapsamında kullanılan mikro-dizilim veri setlerinin hangi tümör tiplerini belirlediği listelenmiştir. Bazı veri setleri bölgesel tümörlere yoğunlaşmış olmakla beraber 9_Tumors, 11_Tumors ve 14_Tumors veri setleri vücut genelinde tümör sınıflandırması yapmaktadır.

3.2. Yöntem

Bu kısımda tez kapsamında önerilen metodun daha iyi anlaşılabilmesi için bilinmesi gereken makina öğrenmesi algoritmalarından bahsedilecektir.

İlk kısımda öznitelik seçme algoritmaları anlatılacak. İkinci kısımda ise genetik algoritma ve k-EK algoritmasından bahsedilecektir. Üçüncü kısımda BDBÇD

doğrulama metodu hakkında bilgi verilecektir. En son olarak önerilen metot açıklanacaktır.

3.2.1. Öznitelik seçme algoritmaları

Makina öğrenme alanında, öznitelik seçme işlemi değişken seçme, özellik seçme şeklinde de isimlendirilmektedir. Öznitelik seçme, sonuca daha az işlem yükü ile sınıflandırılabilecek ve sınıf etiketleri ile daha ilişkili verileri bulmaya yarayan bir makina öğrenme aracı olarak tanımlanabilir. Öznitelik seçme algoritmaları genellikle makina öğrenme uygulamalarında ön işlem adımı olarak kullanılmaktadır.

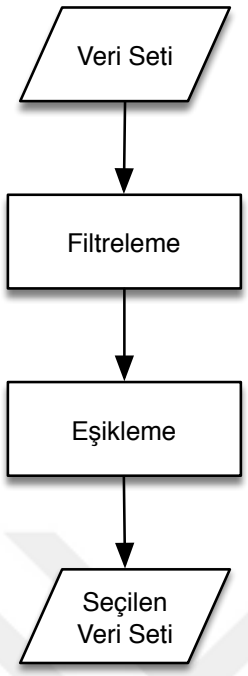
Sınıflandırılacak veri setleri her zaman etkili ve verimli verilerden oluşmamaktadır. Bazen veri setlerinde gereksiz veriler, bazen de sınıflandırma başarı oranını olumsuz yönde etkileyen veriler bulunabilmektedir. Bu veriler sınıflandırma algoritmasını hem işlem yükü olarak hem de sınıflandırma başarı oranı olarak olumsuz etkilemektedir. Bu problemi çözme adına literatürde öznitelik seçme algoritmaları önerilmiştir (Saeys ve ark., 2007) (Blum ve Langley., 1997). Öznitelik seçme algoritmaları sınıflandırma algoritmalarını hem işlem yükü açısından hem de sınıflandırma başarısı açısından desteklemektedir.

Öznitelik seçme algoritmaları, literatürde filtre, sarmal ve gömülü olmak üzere 3 ana başlıkta incelenmektedir (Saeys ve ark., 2007).

3.2.1.1. Filtre modelli öznitelik seçme algoritmaları

Filtre modelli öznitelik seçme algoritmaları, sınıflandırma algoritmasından bağımsız şekilde her bir niteliğin sınıf etiketi ile olan ilişkisini bir katsayı ile ifade etmektedir. Bu işlem filtre türüne göre farklı istatistiksel modeller kullanmakta ve dolaysıyla her filtre farklı katsayılar oluşturmaktadır.

Filtre modelli öznitelik seçme algoritmalarının en büyük avantajı hızlı olmasıdır. Bunun yanında filtre modelli öznitelik seçme algoritmalarının sınıflandırma algoritmasından bağımsız olarak çalışması da avantaj olarak sayılabilir.



Şekil 3.1. Filtre modelli öznitelik seçme işlem adımları

Filtre modelli öznitelik seçme algoritmalarının işlem adımları Şekil 3.1.'de gösterildiği gibidir. Veri seti filtreleme işlemine tabi tutularak her nitelik için bir ilişki katsayısı oluşturulur. Oluşturulan bu katsayılara kendi içerisinde eşikleme işlemi uygulanır. Eşikleme işleminden sonra ilişki katsayısı eşik değerinin üstünde kalan nitelikler öznitelik olarak seçilirler.

Bu tez kapsamında önerilen CFR-GA hibrit öznitelik seçme algoritmasında Ki-kare (Liu ve ark., 1995), ReliefF (Robnik-Sikonja ve ark., 2003) ve Fisher-Skor (Duda ve ark., 2001) filtre modelli öznitelik seçme işlemleri kullanılmıştır.

Ki-kare algoritması 1992 yılında Randy Kerber tarafından yazılmış ve 1995 yılında Huan Liu ve Rudy Setiono tarafından ise yeniden düzenlemiştir. Veri kümesi içinde yer alan bir faktör için hesaplanan ki-kare değeri, o faktörün sınıf içerisindeki bağımlılığını ölçmektedir. Sıfır değerine sahip bir faktör o küme içinde bağımsız olduğunu gösterir. Yüksek bir ki-kare değerine sahip olan faktör, veri kümesi için daha tanımlayıcıdır (Kavzaoğlu ve ark., 2014). Ki-kare istatistik temelli olup öznitelik seçme işlemlerinde yaygın şekilde kullanılmaktadır. Bu metot bulundukları

sınıfa göre tüm niteliklerin Ki-kare'sini hesaplayarak tek tek değerlendirir. Ki-kare istatistik değeri her nitelik değeri için formül (3.1) ile hesaplanır.

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(A_{ij} - \frac{R_i C_j}{N})^2}{\frac{R_i C_j}{N}} \quad (3.1)$$

Yukarıdaki formülde m verilen aralık sayısı, k sınıf sayısını, R_i i'ninci aralıktaki örnek sayısını, C_j j'ninci sınıftaki örnek sayısını, A_{ij} j'ninci sınıfta ve i'ninci aralıktaki örnek sayısını, N ise veri setinde bulunan toplam örnek sayısını temsil etmektedir.

F-Skor filtresi her niteliğin diğer niteliklerden farkını hesaplayan bir filtre modelidir. Dolaysıyla diğer niteliklerden farkı fazla olan nitelik sınıflandırma algoritmaları için daha önemlidir. Bundan dolayı F-skor filtresinin sonucunda yüksek değere sahip nitelikler daha ilişkili niteliklerdir (Hsu ve ark., 2011). İki sınıf olma durumunda her bir i'ninci nitelik için formül (3.2)'deki gibi F-skor hesaplanır.

$$F_i = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \quad (3.2)$$

ReliefF, Relief metodunun gelişmiş versiyonu olarak bilinmektedir. Relief metodunun uygulanmasında bir nitelik belirlenir. Relief metoduna göre incelenen nitelik değerli bir nitelik ise, niteliğe en yakın mesafesinde aynı sınıftan, niteliğe uzak kalan kısımda ise farklı sınıflardan örnekler bulunmalıdır (Robnik-Sikonja ve ark., 2003). Formül (3.3)'de Relief hesaplama denklemi verilmiştir.

$$S_i = \frac{\sum_{j=1}^{m} -fark(x_{ij}, enyakın_aynı_{ij}) + fark(x_{ij}, enyakın_farklı_{ij})}{m} \quad (3.3)$$

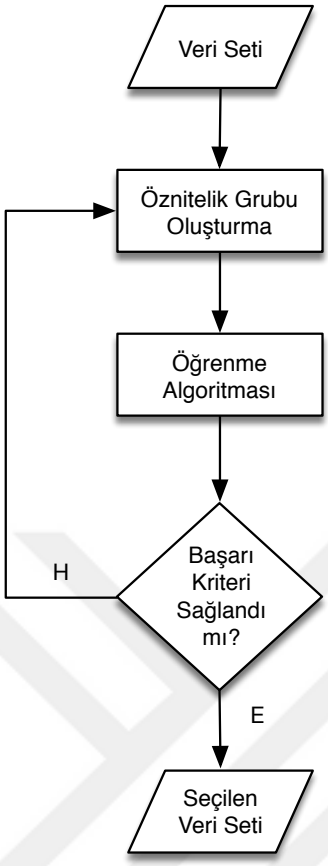
Formül (3.3)'de m veri setinde bulunan örnek sayısını, $fark(x_{ij}, enyakın_aynı_{ij})$ j'ninci örnekte bulunan i'ninci niteliğin aynı sınıfa sahip en yakın örneğe olan uzaklığını, $fark(x_{ij}, enyakın_farklı_{ij})$ ise j'ninci örnekte bulunan i'ninci niteliğin farklı sınıfa sahip en yakın örneğe olan uzaklığını, $S_i$ise i'ninci niteliğin Relief değerini belirtmektedir.

Dolaysıyla S_i Relief değerinin büyük olma şartı incelenen niteliğin, aynı sınıftaki örneklerle mesafesi az, farklı sınıftaki örneklerle mesafesi çok olmasıdır.

Bu tez çalışmasında yukarıda belirtilen Ki-kare, F-skor, ReliefF olmak üzere üç farklı filtre modelli öznitelik seçme algoritması kullanılmıştır. Ki-kare filtre modelli öznitelik seçme algoritması niteliğin sınıflara bağımlılığı, F-skor filtre modelli öznitelik seçme algoritması nitelikler arasındaki bağımlılığı, ReliefF filtre modelli öznitelik seçme algoritması ise niteliklerin aynı sınıfa yakınlık ve farklı sınıflara uzaklık bağımlılığı gözönünde bulundurularak diğer filtre modelli öznitelik seçme algoritmaları arasından seçilmiştir. Bu sayede yukarıda bahsedilen üç farklı parametrenin önerilecek hibrit modelli öznitelik seçme algoritmasında belirleyici bir faktör olması sağlanmıştır.

3.2.1.2. Sarmal modelli öznitelik seçme algoritmaları

Sarmal modelli öznitelik seçme algoritmaları en uygun veri setini sınıflandırma algoritmasında test ederek bulma temeline dayanmaktadır. Sarmal modelli öznitelik seçme algoritmaları, sürekli arama yaparak bulduğu veri setlerini sınıflandırma algoritmasında da test ettiğinden başarılı sonuçlar vermektedir. Fakat mikro-dizilim veri setleri gibi nitelik sayıları fazla olan veri setlerinde işlem süresi arttığından bu gibi veri setlerinde bu sınıftaki metotların kullanımı mümkün olmamaktadır.



Şekil 3.2. Sarmal modelli öznitelik seçme işlem adımları

Sarmal modelli öznitelik seçme algoritmalarının işlem adımları Şekil 3.2.'de verilmiştir. Veri setinden, uygulanacak metoda göre belirlenen kriterlere göre bir öznitelik grubu oluşturulur. Oluşturulan öznitelik grupları öğrenme algoritmasına tabii tutularak sınıflandırma başarı oranı hesaplanır. Hesaplanan sınıflandırma başarı oranı istenilen değerlere ulaşmamış ise iyileştirme yapılması amacıyla tekrar öznitelik grubu oluşturma adımına gönderilir. Bu işlem istenilen başarı oranı sağlanıncaya kadar sürdürülür. Sarmal modelli öznitelik seçme algoritmalarının işlem adımları da incelendiğinde nitelik sayısı fazla veri setleri için bu metotların kullanımının zorlaşacağı anlaşılmaktadır.

Literatürde sıklıkla kullanılan sarmal metotlar, İleri Sıralı Seçme (İSS) (Kittler, 1978), Geri Sıralı Eleme Algoritması (GSE) ve Genetik Algoritma (Holland, 1975) şeklinde listelenebilir.

İSS metodu başlangıçta öznitelik havuzunu boş küme olarak varsayıp tek tek tüm nitelikleri öznitelik havuzuna alarak tüm niteliklerin tek başlarına başarı oranını hesaplamaktadır. Tek elemana sahip öznitelik havuzuna en iyi sonucu veren nitelik seçilir. Eğer istenilen başarı oranı sağlanmamış ise öznitelik havuzuna yeni bir daha öznitelik eklenir ve bu işlem tüm ikili grup kombinasyonları gerçekleştirilinceye kadar yapılır. Bu şekilde istenilen başarı oranı sağlanıncaya kadar öznitelik havuzuna öznitelik ekleme işlemi sürdürülür.

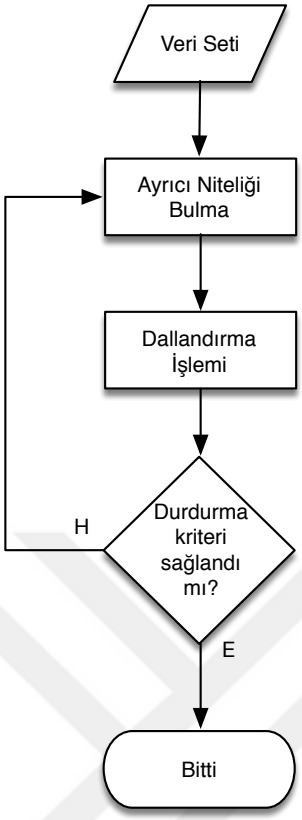
GSE metodu ise İSS metodunun tersi diye değerlendirilebilir. İSS başlangıçta öznitelik havuzunu boş olarak kabul edip öznitelik ekleme yapıyorsa, GSE metodu ise İSS'nin aksine başlangıçta öznitelik havuzunu tüm nitelikleri seçerek başlayıp dah sonra öznitelik havuzundan başarı oranına katkısı en az özniteliği elemektedir. İSS'de olduğu gibi istenilen başarı oranı sağlanana kadar eleme işlemi tekrarlanır.

Bu tez çalışmasında önerilen CFR-GA hibrit öznitelik seçme algoritmasında kullanılan Sarmal modelli öznitelik seçme metodu Genetik Algoritmadır. Genetik algoritma ile öznitelik seçme işlemi yaparken ikilik sistemde bir sayı belirlenip bu sayının 0 veya 1 olma durumuna göre gen seçilmiş veya seçilmemiş olarak kabul edilir.

3.2.1.3. Gömülü modelli öznitelik seçme algoritmaları

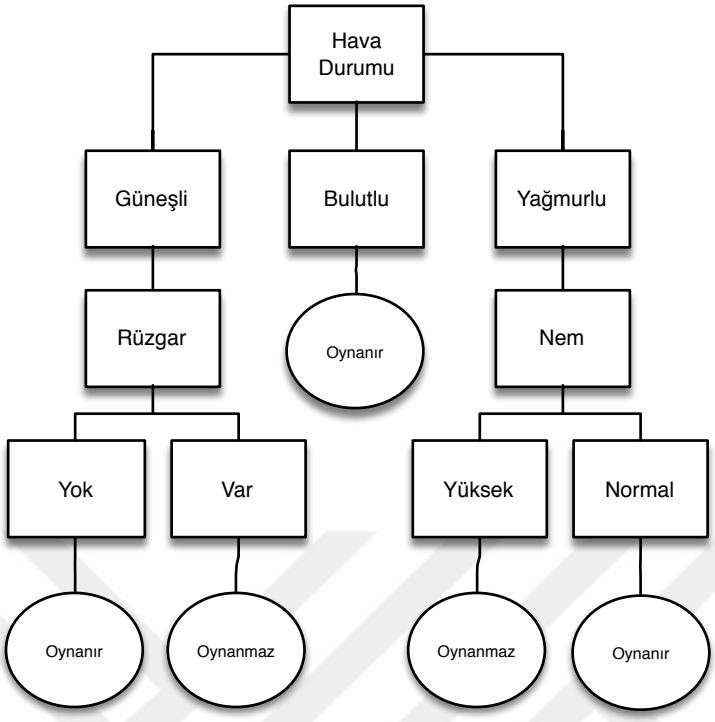
Gömülü modelli öznitelik seçme algoritmaları, adında anlaşılacağı üzere sınıflandırma algoritmasının içinde gömülü olarak bulunan öznitelik seçme algoritmalarıdır. Gömülü modelli öznitelik seçme algoritmalarını kullanan sınıflandırma algoritmaları, sınıflandırma yapabilmek için bir nitelik seçmek zorundadır. Bu seçme işlemi de öznitelik seçme işlemi olarak kabul edilmektedir.

Gömülü modelli öznitelik seçme algoritmalarına en iyi örnek karar ağaçları verilebilir.



Şekil 3.3. Karar ağacı oluşturma işlem adımları

Şekil 3.3.'te gömülü modelli öznitelik seçme algoritmalarına örnek olan karar ağaçları sınıflandırma algoritmasının ağaç oluşturma işlem adımları verilmiştir. Karar ağaçları sınıflandırma algoritması ilk önce eğitim verisi ile ağaç yapısını oluşturmakta ve daha sonra test verileri ile sınıflandırma başarı oranı belirlenmektedir. Karar ağacı oluşturulurken ilk önce hangi niteliğe göre karar ağacının dallanacağına karar verilmesi gerekmektedir. Bu kararı verirken Bilgi Kazancı (BK), Ortak Bilgi (OB) ve Gini İndeksi (Gİ) gibi istatistiksel parametreler kullanılmaktadır. Bu parametreler her bir niteliğin sonuçla olan korelasyonunu analiz ederek değerlendirmekte ve dallanacak nitelik seçilmektedir. İşte bu işlem karar ağacına gömülü olan öznitelik seçme işlemidir. İlk dallanmasına karar verilen nitelik aslında diğer niteliklere göre daha fazla sonuçla ilişkilidir.



Şekil 3.4. Örnek bir karar ağacı yapısı

Şekil 3.4.'te örnek bir karar ağacı yapısı gösterilmektedir. Eğitim seti ile futbol oynanıp oynanmayacağına karar veren bir karar ağacı yapısı oluşturulmuştur. Şekildeki karar ağacına göre futbol oynamak veya oynamamak havanın güneşli, bulutlu veya yağmurlu olmasına, rüzgarın durumuna ve nem durumuna göre belirlenmektedir. Ağaç yapısı oluşturulduktan sonra ise test verisine ağaç yapısının tepesinden itibaren sorular sorulur ve test verisinin sınıfı belirlenir.

3.2.2. Genetik algoritma

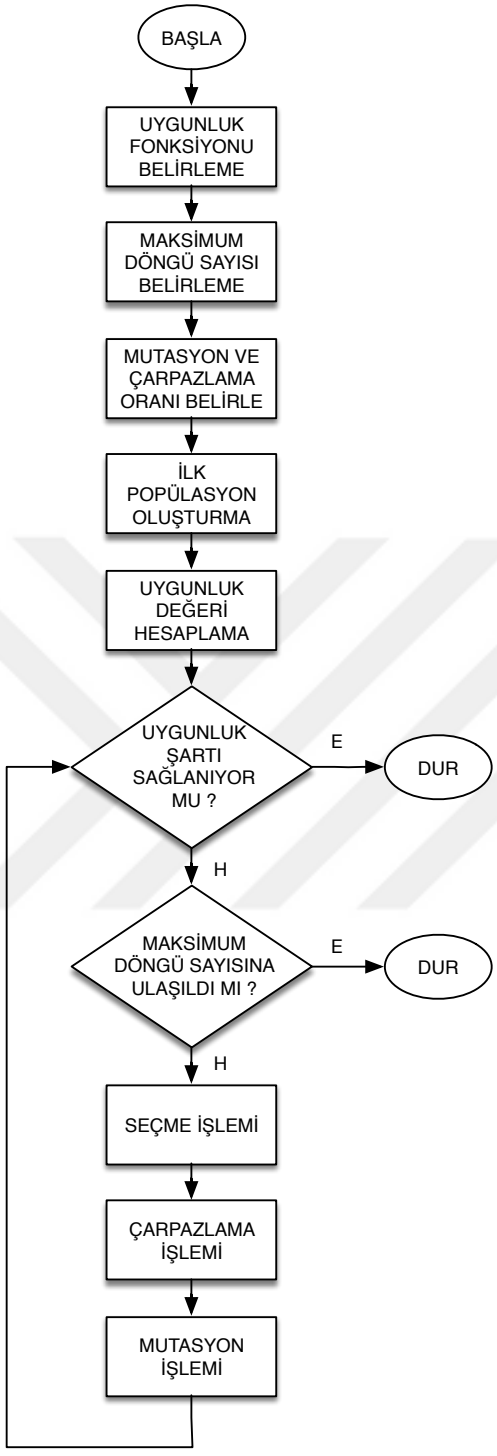
John Holland, en başarılı gen gruplarının hayatta kalacağı düşüncesinden yola çıkarak doğal seçim yöntemlerinden biri olan genetik algoritmayı önermiştir (Aydilek ve ark., 2013). GA'nın temeli aile bireylerinin sahip olduğu genleri çaprazlama, mutasyon vb. gibi işlemlere tabii tutarak daha iyi bir nesil üretmektir (Chang, 2009). GA birçok alanda yaygın olarak kullanılmıştır (Jones ve ark., 1995) (Walters ve ark., 1993) (Kazarlis ve ark., 1996) (Hengpraphrom ve ark., 2010) (Yang ve Xu, 2011). GA'ın amacı problemde verilen çözüm uzayında genetik operatörleri de kullanarak daha iyi çözümü bulmaktır (Li ve ark., 2009). Her oluşturulan yeni gen

havuzlarını nesil olarak adlandıracak olursak, nesiller sonra optimizasyon tamamlanmış ve en ideal çözüm kümesi bulunmuş olacaktır (Braik ve ark., 2008).

Genetik algoritma uygulanırken ilk adım olarak uygunluk fonksiyonu belirlenir. Uygunluk fonksiyonu optimize edilmesi düşünülen parametrelerin bulunduğu matematiksel bir fonksiyondur.

Maksimum döngü sayısı, mutasyon ve çaprazlama oranları belirlendikten sonra ilk popülasyon oluşturulur. İlk popülasyonun oluşturulması ile beraber artık uygunluk fonksiyonu ile her bireyin uygunluk katsayısı hesaplanabilir. Oluşturulan popülasyondaki en iyi uygunluk katsayısını veren bireyin sağladığı uygunluk katsayısı eğer istenilen uygunluk şartını sağlıyorsa genetik algoritma sonlandırılır. Çözüm kümesi olarak da en iyi uygunluk katsayısını veren birey seçilir. Eğer ilk oluşturulan popülasyonda uygunluk şartını sağlayan birey yok ise seçme, mutasyon ve çaprazlama gibi genetik algoritma operatörleri kullanılarak popülasyon değiştirilir.

Değiştirilen popülasyonda yukarıda belirtilen işlemlere tabii tutulur. Genetik algoritma uygunluk şartını sağlayan bireyi bulana kadar bu döngüye devam eder. Genetik algoritmanın sonlandırılması için uygunluk şartının sağlanması veya maksimum döngü sayısına ulaşılması gerekmektedir. Şekil 3.5.'de genetik algoritma işlem adımları gösterilmiştir.



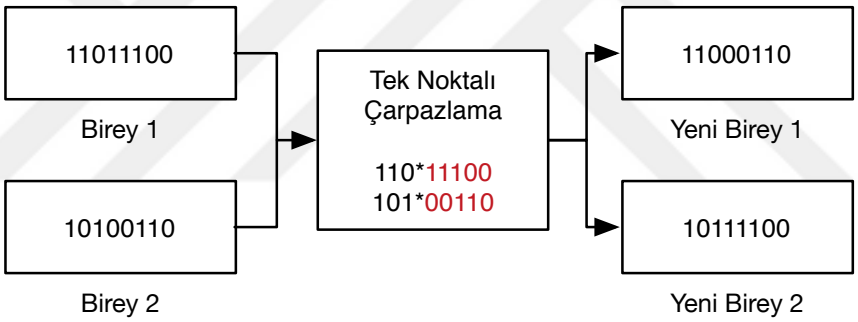
Şekil 3.5. Genetik Algoritma İşlem Adımları

Şekil 3.5.'den de görüldüğü üzere uygunluk fonksiyonu genetik algoritmanın araştırdığı ve optimize etmeye fonksiyondur. Problemin türüne göre uygunluk fonksiyon değeri küçültülerek veya büyültülerek optimize edilmektedir. Uygunluk

fonksiyonun belirlenmesi genetik algoritmanın amacına ulaşması bakımından çok önemlidir.

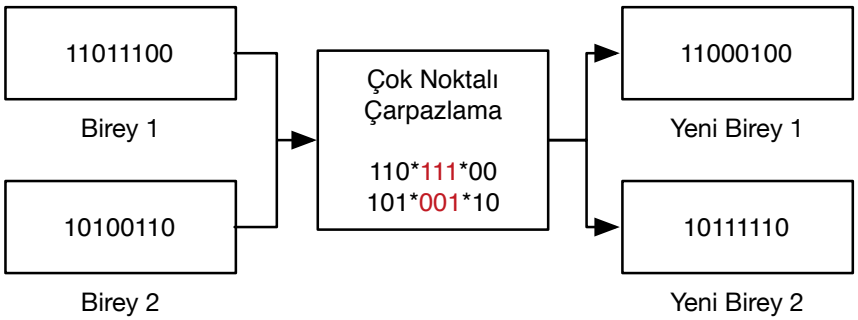
Oluşturulan her popülasyondaki her birey için bir uygunluk katsayısı hesaplanmaktadır. Seçme işleminde ise popülasyon içinde bulunan uygunluk şartından uzak olan uygunluk katsayısına sahip bireyler elenerek yerine yeni bireyler seçilmektedir. Bu şekilde jenerasyonlar arasında geçişte başarısız bireyler elenir. Bu da bundan sonraki işlem adımlarında daha başarılı sonuçlar oluşturmaya zemin hazırlamaktadır. Seçim için bazı kriterler belirlenmiştir. Bunlar rulet tekerleği, turnuva seçimi ve sıra seçimi şeklinde adlandırılmaktadır.

Çaprazlama işlemi ise popülasyonda bulunan bireylerin kromozomları arasında karşılıklı yapılan değişikliklerdir. Çaprazlama tek noktalı, iki noktalı, çoklu karışık noktalı ve noktasal olarak yapılabilmektedir.



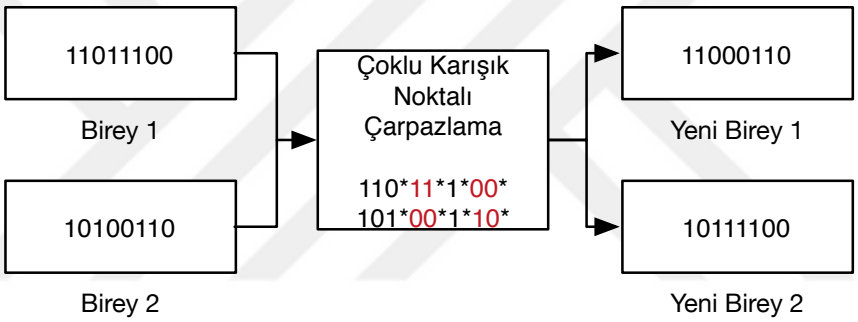
Şekil 3.6. Tek Noktalı Çaprazlama İşlemi

Şekil 3.6.'da iki birey arasında tek noktalı çaprazlama işlemi gösterilmiştir. Şekil 3.6.'da da görüldüğü üzere tek noktalı çaprazlama işlemi seçilen bir noktadan itibaren tüm kromozomların yer değiştirmesi ile gerçekleşmektedir.



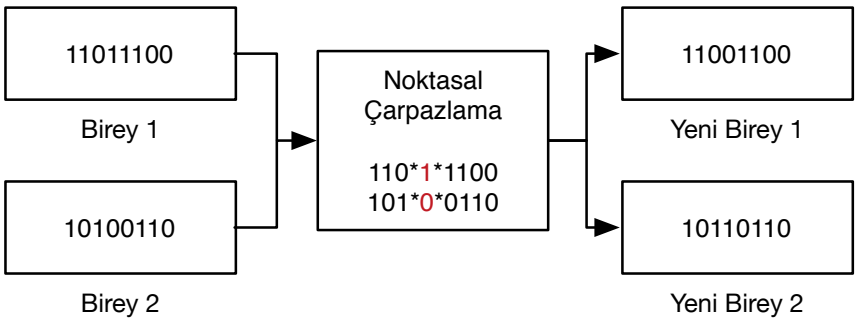
Şekil 3.7. Çok Noktalı Çaprazlama İşlemi

Şekil 3.7.'de de gösterildiği üzere çok noktalı çaprazlama işleminde iki nokta seçilir. Bu iki nokta arasında kalan kromozomlar karşılıklı olarak değiştirilerek yeni bireyler oluşturulur.



Şekil 3.8. Çoklu Karışık Noktalı Çaprazlama İşlemi

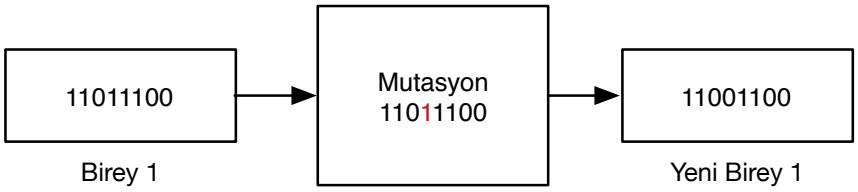
Şekil 3.8.'de çoklu karışık noktalı çaprazlama gösterilmiştir. Çoklu karışık noktalı çaprazlama işleminde birden fazla çok noktalı çaprazlama işlemi gerçekleştirilir. Birden fazla seçilen nokta grupları arasında kalan kromozomlar karşılıklı değiştirilerek yeni bireyler oluşturulur.



Şekil 3.9. Noktasal Çaprazlama İşlemi

Şekil 3.9.'da da gösterildiği üzere noktasal çaprazlama, bir nokta seçilerek sadece bir kromozomun karşılıklı değiştirilmesi ile yeni birey oluşturma temeline dayanmaktadır.

Mutasyon işlemi ise rastgele seçilen bir kromozomun değerinin değiştirilmesi ile gerçekleştirilmektedir.



Şekil 3.10. Mutasyon İşlemi

Şekil 3.10.'da mutasyon işlemi gösterilmiştir. Rastgele seçilen 4.kromozomun değeri 1 iken mutasyon sonucunda 0 olarak değişmiştir.

Genetik algoritma operatörlerinin uygulanma yoğunlukları kullanıcı tarafından ayarlanabilmektedir. Katsayılar büyük seçilirse değişimler çok çabuk gerçekleşmekte ve uygunluk fonksiyonu sonuçlarını hızlı bir şekilde etkilenmektedir. Katsayıların küçük seçilmesi ise uygunluk fonksiyonu sonuçlarını daha yavaş değiştirmekle beraber daha kararlı ideal sonuca yaklaşmasına sebep olmaktadır.

Bu tez çalışmasında öznitelik seçme işlemi yapılırken her nitelik numarası ikilik sayı sisteminde simüle edilmiştir. Her nitelik için 1 veya 0 sayıları verilmiştir. 1 ilgili niteliğin öznitelik diye seçildiğini, 0 ise ilgili bitin öznitelik olmadığını temsil etmektedir. Bu şekilde uygunluk fonksiyonu da formül (3.4)'de verilmiştir.

$$Uygunluk\ Fonksiyonu = k - EK_HataOranı() \quad (3.4)$$

Bu şekilde belirlenen bir genetik algoritmada genetik algoritma uygunluk fonksiyonunu yükseltme değil de düşürmeye çalışmaktadır. Çünkü hata oranı ne kadar düşük olursa o kadar başarılı sınıflandırma yapılmıştır.

3.2.3. k-EK sınıflandırma algoritması

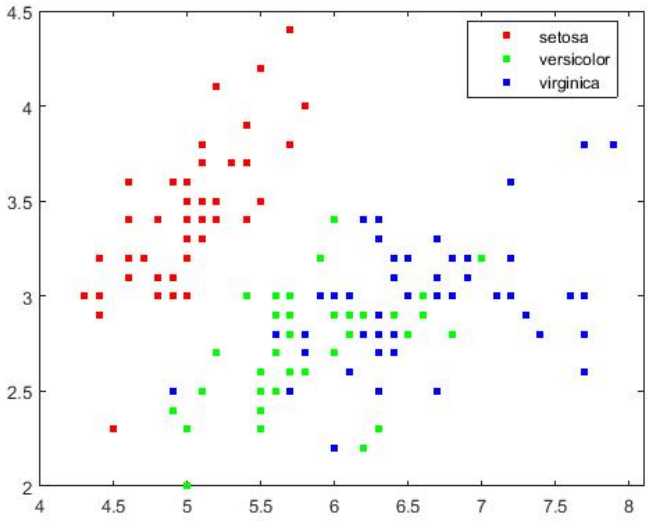
k-EK sınıflandırma algoritması birçok alanda sıklıkla kullanılan metotların başında gelmektedir (Cover ve ark., 1967). k-EK sınıflandırmasının önemli avantajları basitliği ve kullanım kolaylığı yanında büyük veri setlerinde sağladığı kararlılık da kullanımını yaygınlaştırmıştır.

k-EK sınıflandırma algoritması uygulanırken eğitim verisindeki örnekler düzleme yerleştirilerek sınıfları ile ilişkilendirilir. Kurulan düzleme test için kullanılacak örnekler düzleme yerleştirilir. k değeri kadar test verisinin en yakın komşularına bakılarak test verisinin sınıfı belirlenir. Test verisinin sınıfı, komşu örnekler en çok hangi sınıfa ait ise ilgili sınıfa ait olduğu tahmin edilir. Bu şekilde yapılan bir sınıflandırma da k sayısı olası eşit sayıda farklı sınıflar oluşmaması için tek sayı seçilmelidir. k değerine göre en yakın komşular seçileceğinden, k değeri sınıflandırma başarı oranını etkilemektedir. Farklı sınıf örnekleri arasında uzaklıkları az olan veri setlerinde k değeri küçük, farklı sınıf örnekleri arasında uzaklıkları çok olan veri setlerinde k değeri büyük seçilmelidir. En yakın komşular seçilirken uzaklık ölçüsü olarak birden fazla parametre kullanılabilir. Bunlar; Öklid uzaklığı, Manhattan uzaklığı, Minkowski uzaklığı şeklinde listelenebilir. Bu uzaklık ölçüleri arasında literatürde en yaygın kullanılan öklid uzaklığıdır.

Bu tez çalışmasında k değeri 1 olarak seçilip, komşuluk uzaklık hesabı öklid uzaklığa göre hesaplanmıştır. k değerinin 1 olarak seçilmesindeki amaç karşılaştırılacak olan literatürdeki öznitelik seçme algoritmalarının değerlendirmesinin de k=1 koşulunda yapılmış olmasıdır. Öklid uzaklık hesaplama yöntemi aşağıdaki formülde verilmiştir.

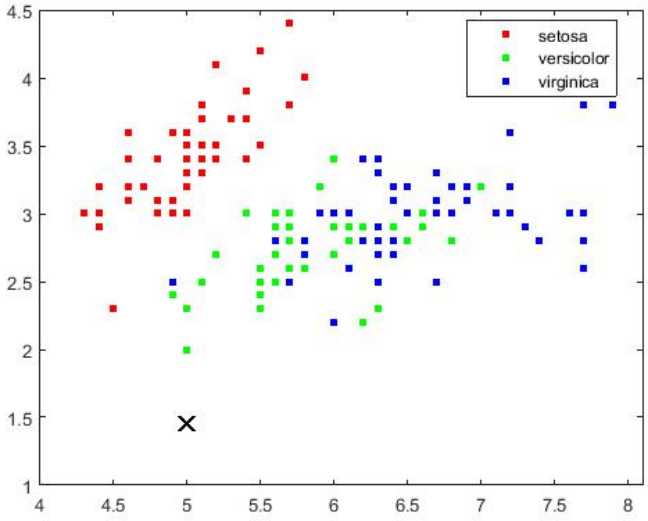
$$d_{\text{öklid}}(i,j) = \sqrt[2]{(x_i - x_j)^2 + (y_i - y_j)^2} \qquad (3.5)$$

Formül (3.5)'de i noktasının koordinatları (x_i, y_i) ve j noktasının koordinatları (x_j, y_j) olarak alınmıştır. $d_{\text{öklid}}(i,j)$ ise i noktasının j noktasına olan öklid uzaklığını temsil etmektedir.



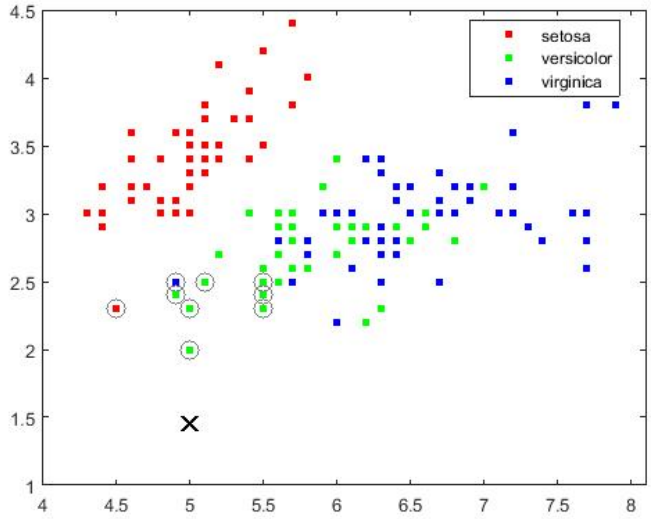
Şekil 3.11. Eğitim veri setinin dağılımı

Şekil 3.11.'de eğitim veri setinin örneklerinin dağılımı gösterilmektedir. Koordinat sisteminde dağıtılmış veri seti 3 adet sınıf bilgisine sahip olup bu sınıflar kırmızı, yeşil ve mavi renklendirilmiştir.



Şekil 3.12. k-EK modeline test verisi girişi

Şekil 3.12.'de oluşturulan k-EK modelinin girişine bir veri gönderilmiş ve bu verinin sınıflandırılması istenmiştir. X ile işaretlenen veri sınıflandırılmak üzere gönderilen veriyi temsil etmektedir.



Şekil 3.13. Test verisinin komşuluk dağılımı

k-EK modeline test edilmek üzere gönderilen veri için k değeri 9 seçilmiştir. Şekil 3.13.'de de görüldüğü üzere X ile işaretlenen test verisine en yakın 9 örnek işaretlenmiştir. İşaretlenen örnekler arasında en çok sınıf yeşil rengine ait olduğundan test verisi de yeşil renk olarak sınıflandırılır.

Bunun yanında BDBÇD çapraz doğrulama metodu ile sonuçlar doğrulanmıştır.

3.2.4. Birini dışarda bırak çapraz doğrulama (BDBÇD)

Veri madenciliği çalışmalarında, uygulanan yöntemin başarısının sınanması için, veri kümesini eğitim ve test kümeleri olarak ayırılmaktadır. Bu ayırma işlemi çeşitli şekillerde yapılabilir. Bu yöntemlerden en yaygın olarak kullanılanı Birini Dışarıda Bırak Çapraz Doğrulama (BDBÇD) yöntemidir.

BDBÇD yöntemi literatürde bulunan k-katlamalı çapraz doğrulama metoduna göre daha kararlıdır. Bunun sebebi BDBÇD yönteminin tüm örnekler için tek tek

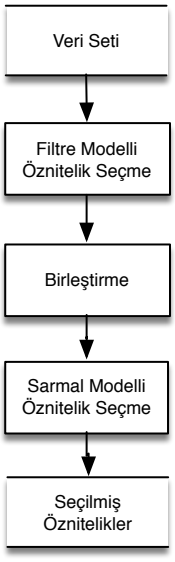
model oluşturmasıdır. Bunun yanında BDBÇD yöntemi k-katlamalı çaprazlama yöntemine göre daha uzun sürede doğrulama işlemini tamamlar.

BDBÇD yönteminde öncelikle veri setinde bulunan ilk örnek çıkartılarak geriye kalan örneklerle eğitim veri seti oluşturulur. Elde edilen eğitim seti ile model oluşturulduktan sonra çıkartılan ilk örnek test veri seti olarak kabul edilip sınıflandırma başarı oranı hesaplanır. Bu işlem tüm örnekler için tekrarlanır ve örnek sayısı kadar farklı sınıflandırma başarı oranları bulunur. Bulunan sınıflandırma başarı oranlarının aritmetik ortalaması alınarak veri setinin sınıflandırma başarı oranı hesaplanmış olur.. Her işlem adımında test veri seti değiştirilerek daha kararlı bir sınıflandırma işlemi yapılır.

Bu yöntemlerin asıl amacı veri setindeki eğitim ve test verilerinin kullanıcı tarafından değil de belirli bir algoritmaya göre seçilip başarı oranlarını daha evrensel kıyaslayabilmektir.

3.3. Önerilen Yöntem

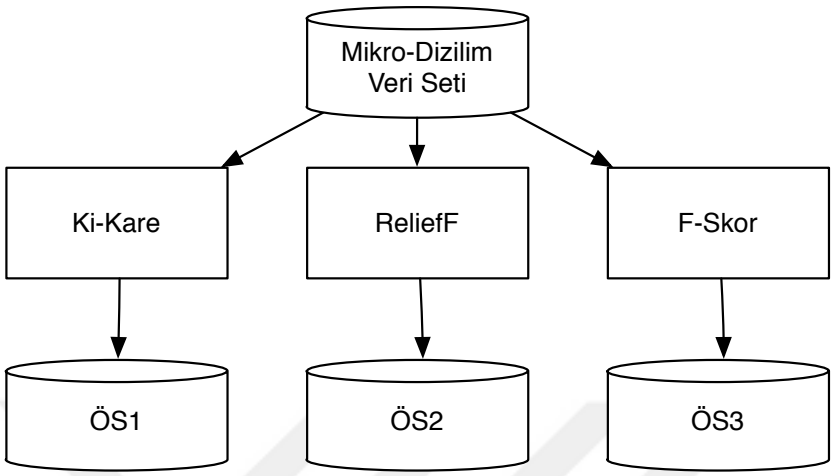
Bu bölümde tez kapsamında önerilen CFR-GA hibrit öznitelik seçme algoritması detaylı şekilde anlatılacaktır.



Şekil 3.14. CFR-GA öznitelik seçme algoritması temel işlem adımları

Şekil 3.14.'de de gösterildiği üzere önerilen metot 3 ana işlem adımından oluşmaktadır. Birincisi filtre modelli öznitelik seçme işlemi, ikincisi üç farklı

filtreleme modelli öznitelik seçme işleminin sonuçlarını birleştirme işlemi, üçüncüsü ise sarmal modelli öznitelik seçme işlemi olarak sıralanabilir.

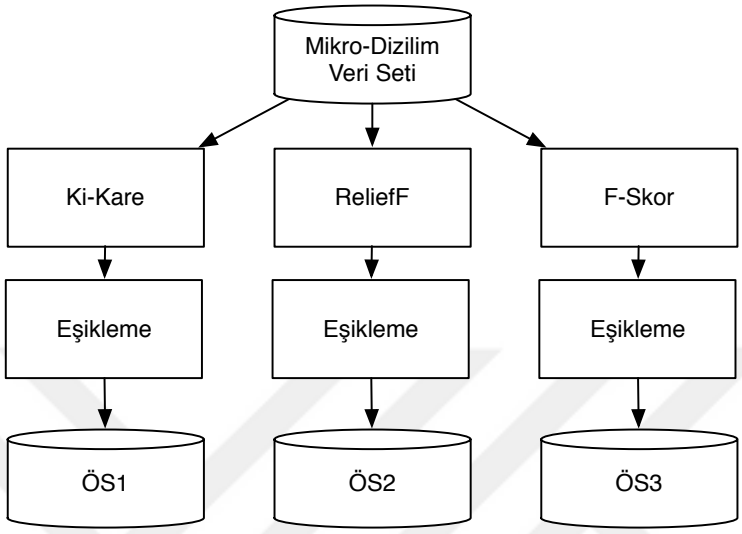


Şekil 3.15. Filtre modelli öznitelik seçme işlem adımı

Şekil 3.15.'de CFR-GA öznitelik seçme algoritmasının ilk işlem adımı olan filtre modelli öznitelik seçme işlemi gösterilmiştir. Mikro-dizilim veri seti sırasıyla Ki-Kare, ReliefF, F-Skor filtreleri ile öznitelik seçme işlemi yapılmakta ve çıkışta elde edilen öznitelik kümeleri ÖS1, ÖS2, ÖS3 şeklinde adlandırılmaktadır. Bu işlem adımı sonunda ÖS1, ÖS2 ve ÖS3 adında birbirinden farklı üç adet öznitelik kümesi oluşmaktadır. Filtre modelli öznitelik seçme işlemi, önerilen CFR-GA hibrit modeline işlem süresi bağlamında fayda sağlamaktadır. Filtre modelli öznitelik seçme işlem adımından sonra, elde edilen üç farklı öznitelik veri kümesi birleştirilerek tek veri setine dönüştürülecektir.

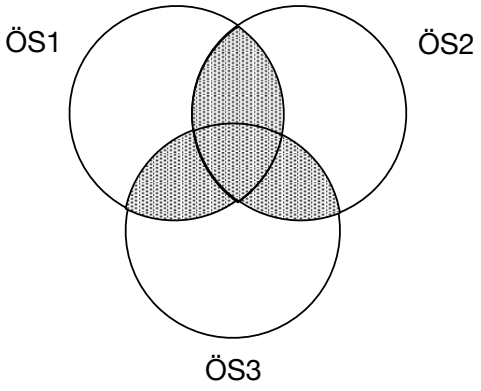
Bu tez çalışmasında kullanılan filtre modelli öznitelik seçme algoritmaları her bir nitelik için ayrı birer ilişki katsayısı üretmektedir. Filtre modelli öznitelik seçme algoritmaları sonucunda üretilen bu katsayılar, niteliklerin veri setinde bulunan sınıflar ile ilişkisini belirlemektedir. Katsayı büyüdükçe ilgili niteliğin sınıflar ile ilişkisi artmaktadır diye değerlendirilir.

Dolaysıyla filtre modelli öznitelik seçme işlemi sonucunda elde edilen katsayılar için bir kıstas belirlenip özniteliklerin seçimi yapılmalıdır. Bunun için eşikleme işlemi kullanılmıştır.



Şekil 3.16. Eşikleme işlem adımı

Şekil 3.16.'da gösterildiği gibi filtre modelli öznitelik seçme algoritmalarından çıkan katsayılar eşiklenerek ilgili nitelik hakkında öznitelik olup olmadığı konusunda karar vermemizi sağlamaktadır.



Şekil 3.17. Birleştirme işlemi

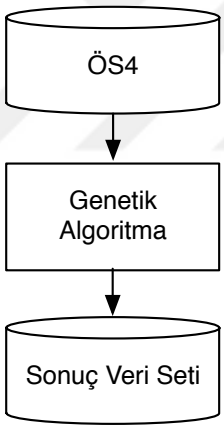
İkinci temel işlem adımı ise filtreleme işleminin sonucunda oluşan 3 farklı veri setini kombine etme işlemidir. Ki-Kare filtre modelli öznitelik seçme işleminde elde edilen ÖS1, ReliefF filtre modelli öznitelik seçme işleminde elde edilen ÖS2 ve

F-Skor filtre modelli öznitelik seçme işleminde elde edilen ÖS3 Şekil 3.17.'de görüldüğü gibi birleştirilerek tek bir öznitelik veri seti haline getirilmektedir.

Birleştirme işlem adımında bir referans belirleyip 3 veri setinden tek bir veri setine indirgeme yapılmaktadır. Bizim önerdiğimiz metotta ise en az iki veri setinde de bulunan öznitelikler oluşturulan yeni veri setine seçilmektedir. Buna göre sarmal modelli öznitelik seçme algoritmasına tabii tutulacak veri seti ÖS4 aşağıdaki formül ile seçilmektedir.

$$\text{ÖS}4 = (\text{ÖS}1 \cap \text{ÖS}2) \cup (\text{ÖS}1 \cap \text{ÖS}3) \cup (\text{ÖS}2 \cap \text{ÖS}3) \qquad (3.6)$$

(3.6)'da belirtildiği gibi bir özniteliğin ÖS4 öznitelik veri setine dahil olabilmesi için en az iki filtre modelli öznitelik seçme algoritması tarafından seçilmiş olması gerekmektedir. Bu işlem adımından sonra seçilen öznitelik veri seti ÖS4 sarmal modelli öznitelik seçme algoritmasına tabii tutulmaktadır.



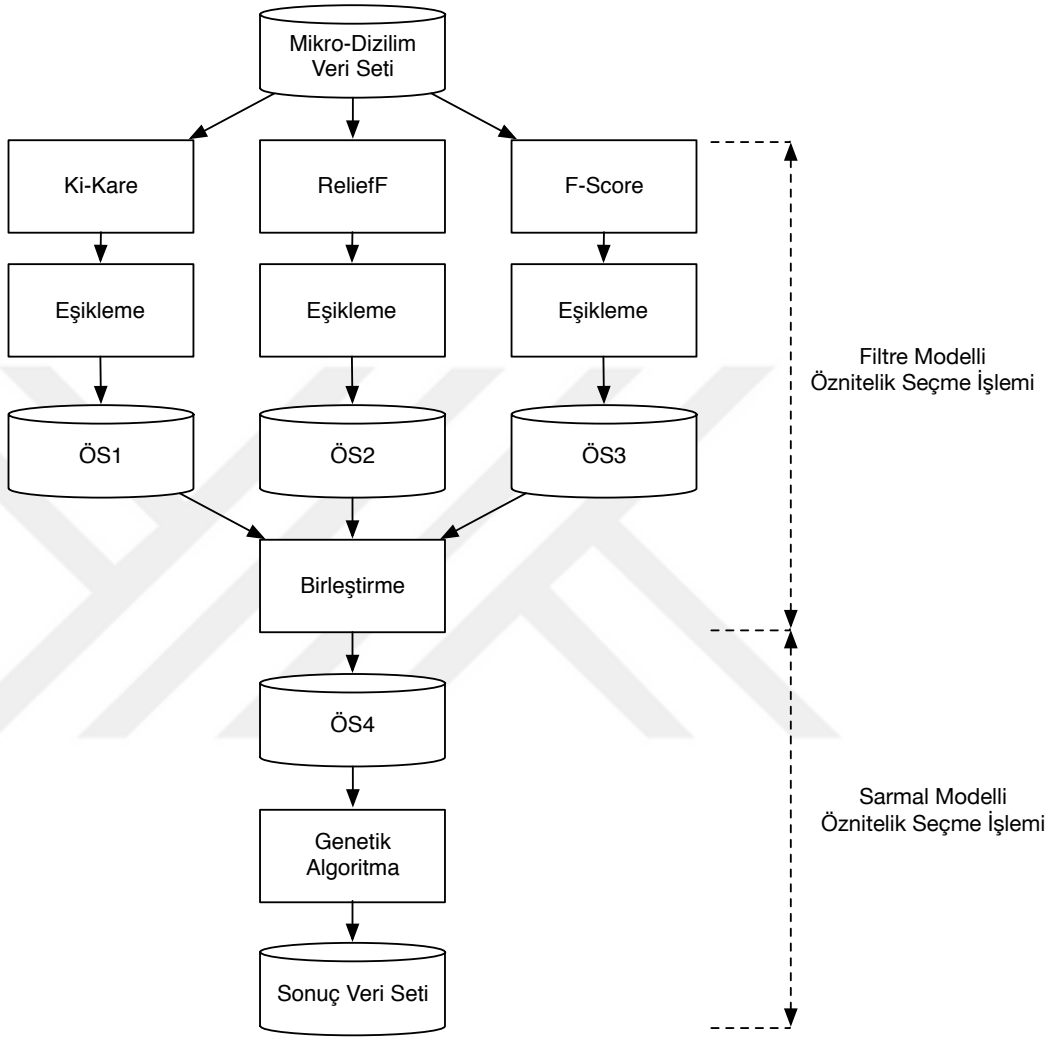
Şekil 3.18. Sarmal modelli öznitelik seçme işlem adımı

Şekil 3.18.'de görüldüğü gibi ÖS4 öznitelik veri setine GA sarmal modelli öznitelik seçme işlemi uygulanarak final öznitelik veri seti elde edilmektedir. Sarmal modelli öznitelik seçme işlemi, önerilen CFR-GA hibrit modeline sınıflandırma başarı oranı bağlamında fayda sağlaması amacıyla uygulanmaktadır.

Önerilen metotta Genetik algoritmanın uygunluk fonksiyonu k-EK sınıflandırma algoritmasının sınıflandırma hata oranı olarak alındığından ÖS4 içindeki hata oranını en düşük verecek veri seti sonuç olarak çıkacaktır.

$$Uygunluk\ Fonksiyonu = k - EK_HataOranı(S\ddot{O}S) \quad (3.7)$$

(3.7)'de belirtilen $S\ddot{O}S$ seçilen öznitelik veri kümesi manasına gelmektedir.



Şekil 3.19. CFR-GA işlem adımları

Özet olarak Şekil 3.19.'da da görüldüğü üzere önerilen metot, filtre modelli öznitelik seçme ve sarmal modelli öznitelik seçme işlemleri olmak üzere 2 farklı model öznitelik seçme işleminin kombine edildiği hibrit bir öznitelik seçme metodudur. Birincisi filtre modelli öznitelik seçme işlemi, ikincisi kombine etme işlemi, üçüncüsü ise genetik algoritma ile öznitelik seçme işlemi olarak sıralanabilir. Filtre modelli öznitelik seçme işleminde, farklı istatistiksel modellere sahip Ki-Kare, ReliefF, F-Skor olmak üzere 3 adet filtre uygulanmıştır. Filtre modelli öznitelik

seçme işlemleri sonucunda elde edilen üç ayrı öznitelik veri seti kombine edilerek tek bir öznitelik veri seti haline getirilmekte ve son işlem adımı olan sarmal modelli öznitelik seçme algoritmalarına gönderilmektedir. Sarmal modelli öznitelik seçme işlemi olan genetik algoritma ile öznitelik seçme işlemi k-EK sınıflandırma başarı oranını optimize ederek en ideal öznitelik veri setini seçmektedir.

CFR-GA algoritması ile elde edilen öznitelik veri seti k-EK sınıflandırma algoritması ile test edilmiştir. k-EK algoritmasında k değeri 1 olarak seçilmiş olup BDBÇD ile doğrulamaya tabii tutulmuştur.

Elde edilen test sonuçları öznitelik seçme işlemi yapılmamış 7 sınıflandırma, 3 filtre modelli öznitelik seçme, 2 sarmal modelli öznitelik seçme ve 2 hibrit modelli öznitelik seçme algoritmaları ile kıyaslanmıştır. Deneysel sonuçlar, önerilen metodun kıyaslanan metotlara önemli ölçüde iyileştirmeler yaptığını

4. ARAŞTIRMA BULGULARI ve TARTIŞMA

Bu tez çalışmasında önerilen CFR-GA hibrit modelli öznitelik seçme algoritması mikro-dizilim veri setlerinde 10 katlamalı doğrulama metodu ile k-EK sınıflandırma algoritmasında test edilmiştir. Önerilen CFR-GA hibrit modelli öznitelik seçme algoritması, 9_Tumors, 11_Tumors, 14_Tumors, Brain_Tumor1, Brain_Tumor2, Leukemia1, Leukemia2, Lung_Cancer, SRBCT, Prostate_Tumors, DLBCL adında toplam 11 adet farklı mikro-dizilim veri setine uygulanmıştır.

İşlem süreleri ile ilgili sonuçlar 2.5 GHz dört çekirdekli Intel Core i7 işlemcili, 16 GB 1600 MHz belleğe sahip bilgisayar ile hesaplanmıştır.

Bu kısımda öncelikle CFR-GA hibrit modelli öznitelik seçme modelinin işlem adımlarından olan 3 farklı filtre modelli öznitelik seçme algoritmasının birleştirme işlemini incelemek adına Ki-kare, F-Skor ve ReliefF filtre modelli öznitelik seçme algoritmaları işlem süreleri, sınıflandırma başarı oranı bakımından kıyaslanacaktır.

İkinci olarak CFR filtre modelli öznitelik seçme algoritması, GA sarmal modelli öznitelik seçme algoritması ve CFR-GA hibrit modelli öznitelik seçme algoritması arasında başarı oranı ve işlem süresi karşılaştırılması yapılacaktır. Bu karşılaştırma CFR ve GA öznitelik seçme algoritmalarının birbirlerine karşı avantajlarını ve dezavantajlarını belirlemek amacı ile yapılmaktadır.

Diğer bir karşılaştırma ise CFR-GA algoritmasının başarı oranının, öznitelik seçme yapılmadan elde edilen başarı oranı ile kıyaslanmasıdır. Bu karşılaştırma yapılan öznitelik seçme işleminin başarı oranına etkisini belirleyecektir.

Son karşılaştırma ise önerilen yöntem CFR-GA algoritmasının literatürde mikro-dizilim veri setlerinde kullanılan öznitelik seçme algoritmaları ile karşılaştırılması olacaktır. Bu aşamada yapılan bütün karşılaştırmalarda sınıflandırma algoritmaları k-EK algoritması olarak belirlenmiştir. Bunun amacı sınıflandırma algoritmalarını sabit tutup öznitelik seçme algoritmalarının başarı oranlarını kıyaslayabilmektir. Bu karşılaştırma literatürde önerilen yöntemleri modellerine göre

üçe ayırarak yapılacaktır. Bunlardan ilki filtre modelli öznitelik seçme yapılarak kullanılan veri setlerinin sınıflandırma başarı oranları ile önerilen CFR-GA öznitelik seçme algoritmasının karşılaştırılması olacaktır. İkinci karşılaştırma ise önerilen CFR-GA algoritması ile sarmal modelli öznitelik seçme algoritmaları başarı oranları arasında yapılacaktır. Üçüncü ve son karşılaştırma ise önerilen CFR-GA algoritması ile hibrit modelli öznitelik seçme algoritmaları başarı oranları arasında yapılacaktır. Bu karşılaştırmalar önerilen CFR-GA hibrit modelli öznitelik seçme algoritmasının literatürde önerilmiş diğer öznitelik seçme algoritmaları ile kıyaslanıp, önerilen yöntemin hangi veri setinde olumlu hangi veri setinde olumsuz etkilere sebep olduğunu belirlemektedir.

Çizelge 4.1. CFR filtre modelindeki filtrelerin sınıflandırma başarı oranı karşılaştırılması

Veri Seti	Ki-Kare k-EK	F-Skor k-EK	ReliefF k-EK	CFR k-EK
9_Tumors	%60.00	**%65.00**	%61.67	%61.67
11_Tumors	%89.08	%83.91	%86.88	**%89.66**
14_Tumors	**%56.82**	%56.17	%56.17	%53.57
Brain _Tumor1	%81.11	**%87.78**	%85.56	**%87.78**
Brain _Tumor2	%72.00	%72.00	%74.00	**%78.00**
Leukemia1	%94.44	%94.44	%94.44	**%94.44**
Leukemia2	**%95.83**	%93.06	**%95.83**	**%95.83**
Lung_Cancer	%91.13	**%94.58**	%89.66	%90.64
SRBCT	**%100.00**	**%100.00**	**%100.00**	**%100.00**
Prostate_Tumor	%81.37	%79.41	%81.37	**%82.35**
DLBCL	%93.51	%90.91	**%97.40**	%94.81
Ortalama	%83.20	%83.38	%83.90	**%84.22**

Çizelge 4.1.'de Ki-kare, F-Skor ve ReliefF filtre modelli öznitelik seçme algoritmalarının mikro-dizilim veri setlerine uygulandığında k-EK sınıflandırma başarı oranları verilmiştir. Koyu font ile yazılmış olan sınıflandırma başarı oranı ilgili veri seti için sağlanmış en iyi sınıflandırma başarı oranını belirtmektedir. Ki-kare 3, F-Skor 4, ReliefF 3 ve CFR 7 veri setinde en iyi sınıflandırma başarı oranını vermiştir. CFR filtre modelli öznitelik seçme algoritması sınıflandırma başarı oranı bağlamında genel olarak diğer tek filtre modelli öznitelik seçme algoritmalarına göre

daha başarılıdır. SRBCT mikro-dizilim veri seti tüm filtre modelli öznitelik seçme algoritmalarında %100 sınıflandırma başarı oranını sağlamıştır.

Çizelge 4.2. CFR filtre modelindeki filtrelerin filtreleme süresi karşılaştırılması

Veri Seti	Ki-Kare	F-Skor	ReliefF	CFR
9_Tumors	**1.21**	2.68	1.97	5.37
11_Tumors	**3.54**	6.45	12.60	25.12
14_Tumors	**7.34**	17.61	57.12	86.58
Brain _Tumor1	**1.32**	1.56	2.72	5.51
Brain _Tumor2	**1.74**	2.17	2.98	6.36
Leukemia1	1.13	**0.97**	1.96	3.87
Leukemia2	1.91	**1.86**	3.99	7.05
Lung_Cancer	3.34	**3.32**	12.11	18.52
SRBCT	0.72	**0.70**	1.19	2.09
Prostate_Tumor	1.77	**1.40**	5.38	7.89
DLBCL	1.28	**0.81**	1.91	3.55
Ortalama	**2.3**	3.59	9.44	15.62

Ki-kare, F-Skor ve ReliefF filtre modelli öznitelik seçme algoritmalarının mikro-dizilim veri setlerine uygulanmasında geçen işlem süreleri Çizelge 4.2.'de verilmiştir. Mikro-dizilim veri setleri için en hızlı filtre modelli öznitelik seçme algoritmaları Ki-kare ve F-Skor olarak belirlenmiştir. Çizelge 4.2.'de verilen süreler için birim saniye olarak alınmıştır. CFR filtre modelli öznitelik seçme algoritması zaten 3 filtrenin de uygulanması ile sonuç verdiğinden süre bakımından avantajlı olduğu söylenemez. Fakat burada CFR filtre modelli öznitelik seçme algoritmasını Genetik algoritmanın ön filtreleme aşaması gibi değerlendirmek gerekir. Bu yüzden Çizelge 4.2'de verilen süreler bakımından avantajlı olduğu görülmese de aslında CFR-GA algoritmasının devamında kullanılacak Genetik algoritma için süre bakımından çok fayda sağlayacaktır.

Çizelge 4.3. CFR filtre modelindeki filtrelerin seçilen öznitelik sayısı karşılaştırılması

Veri Seti	Ki-Kare	F-Skor	ReliefF	CFR
9_Tumors	**165**	762	1231	715
11_Tumors	3179	**396**	2752	2086
14_Tumors	**1986**	3393	4461	2819
Brain _Tumor1	1514	**692**	1954	1196
Brain _Tumor2	4465	**2226**	3140	2984
Leukemia1	837	**201**	722	509
Leukemia2	4548	**1570**	2221	2272
Lung_Cancer	2178	**536**	2833	1588
SRBCT	634	**122**	493	381
Prostate_Tumor	1984	**865**	954	1012
DLBCL	874	1102	868	**807**
Ortalama	2033	**1078**	1966	1488

Çizelge 4.3.'de Ki-kare, F-Skor ve ReliefF filtre modelli öznitelik seçme algoritmalarının mikro-dizilim veri setlerine uygulandığında seçilen öznitelik sayıları verilmiştir. CFR filtre modelli öznitelik seçme algoritması 3 filtrenin birleşiminden oluştuğundan seçilen öznitelik sayısı bakımından başarılı sonuçlar vermemektedir.

Çizelge 4.4. CFR-GA öznitelik seçme algoritmasının CFR ve GA ile sınıflandırma başarı oranları bakımından karşılaştırılması

Veri Seti	Ki-Kare k-EK	F-Skor k-EK	ReliefF k-EK	CFR k-EK	GA k-EK	CFR-GA k-EK
9_Tumors	%60.00	%65.00	%61.67	%61.67	%56.67	**%93.34**
11_Tumors	%89.08	%83.91	%86.88	%89.66	%83.33	**%98.28**
14_Tumors	%56.82	%56.17	%56.17	%53.57	%57.80	**%70.35**
Brain _Tumor1	%81.11	%87.78	%85.56	%87.78	%92.22	**%96.67**
Brain _Tumor2	%72.00	%72.00	%74.00	%78.00	%80.00	**%96.00**
Leukemia1	%94.44	%94.44	%94.44	%94.44	%97.22	**%100.0**
Leukemia2	%95.83	%93.06	%95.83	%95.83	%95.83	**%100.0**
Lung_Cancer	%91.13	%94.58	%89.66	%90.64	%94.09	**%98.53**
SRBCT	**%100.00**	**%100.00**	**%100.00**	**%100.00**	%96.39	**%100.0**
Prostate_Tumor	%81.37	%79.41	%81.37	%82.35	%91.18	**%95.10**
DLBCL	%93.51	%90.91	%97.40	%94.81	%93.51	**%100.0**
Ortalama	%83.20	%83.38	%83.90	%84.22	%85.29	**%95.29**

Çizelge 4.4.'de CFR filtre modelli öznitelik seçme algoritması, GA sarmal modelli öznitelik seçme algoritması ve tezde önerilen CFR-GA hibrit modelli öznitelik seçme algoritması sınıflandırma başarı oranları açısından karşılaştırılmıştır. Koyu font ile yazılmış olan sınıflandırma başarı oranları ilgili veri seti için en iyi sınıflandırma başarı oranını temsil etmektedir. Çizelge 4.4.'den de anlaşılacağı üzere, CFR-GA hibrit modelli öznitelik seçme algoritmasının sınıflandırma işlemi, filtre modelli öznitelik seçme algoritması CFR ve sarmal modelli öznitelik seçme algoritması GA'nın tek başlarına sağladıkları sınıflandırma işleminden daha başarılıdır.

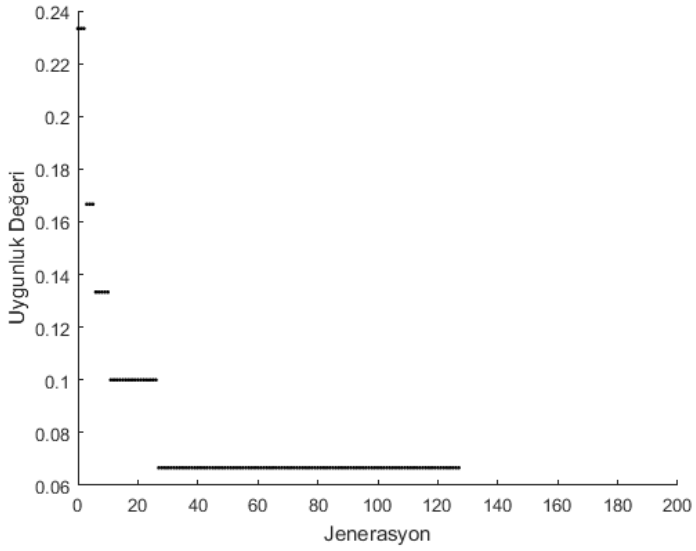
Çizelge 4.5. CFR-GA öznitelik seçme algoritmasının CFR ve GA ile işlem süresi bakımından karşılaştırılması

Veri Seti	CFR	GA	CFR-GA
9_Tumors	**5.37**	30977.33	17406.50
11_Tumors	**25.12**	12796.70	9584.51
14_Tumors	**86.58**	175479,23	148313.81
Brain _Tumor1	**5.51**	54495.31	49542.87
Brain _Tumor2	**6.36**	40945.48	38761.65
Leukemia1	**3.87**	2919.23	1255.21
Leukemia2	**7.05**	2215.29	2017.76
Lung_Cancer	**18.52**	197627.43	133073.74
SRBCT	**2.09**	1033.13	767.47
Prostate_Tumor	**7.89**	90292.24	69069.33
DLBCL	**3.55**	1518.92	1008.59
Ortalama	**15.62**	55481,84	42800,13

CFR filtre modelli öznitelik seçme algoritması, GA sarmal modelli öznitelik seçme algoritması ve tezde önerilen CFR-GA hibrit modelli öznitelik seçme algoritması işlem süreleri açısından karşılaştırılması Çizelge 4.5.'de verilmiştir. Çizelge'de verilen süreler için birim saniye olarak alınmıştır. Bu karşılaştırmada aslında dikkat edilmesi gereken GA'nın tüm mikro-dizilim veri setine uygulandığındaki işlem süresi ile CFR ile filtrelendikten sonraki GA'nın işlem süresi arasındaki farkdır. Çizelge 4.5.'de görüldüğü üzere CFR filtre modelli öznitelik seçme algoritması GA işlem süresine ciddi bir katkı sağlamaktadır.

Giriş kısmında da belirtildiği gibi bu tezde önerilen öznitelik seçme algoritması, filtre modelli öznitelik seçme algoritmalarının hızını ve sarmal modelli öznitelik seçme algoritmalarının sınıflandırma başarısını harmanlayan bir modeldir. Çizelge 4.4. ve Çizelge 4.5. incelenecek olursa, önerilen CFR-GA hibrit modelli öznitelik seçme algoritması, Çizelge 4.4.'de GA'nın sağladığı sınıflandırma başarı oranı ile ve Çizelge 4.5.'de ise GA'nın işlem süresi ile karşılaştırılmıştır. Sonuç olarak önerilen yöntem GA sarmal modelli öznitelik seçme algoritmasından daha başarılı ve hızlıdır. GA sarmal modelli öznitelik seçme algoritmasından hızlı olmasındaki faktör CFR filtrelemesidir. CFR filtreleme işlemi ile beraber nitelik sayısı ciddi şekilde düşürülmüş ve hem daha hızlı hem daha başarılı bir hibrit modelli öznitelik seçme algoritması önerilmiştir.

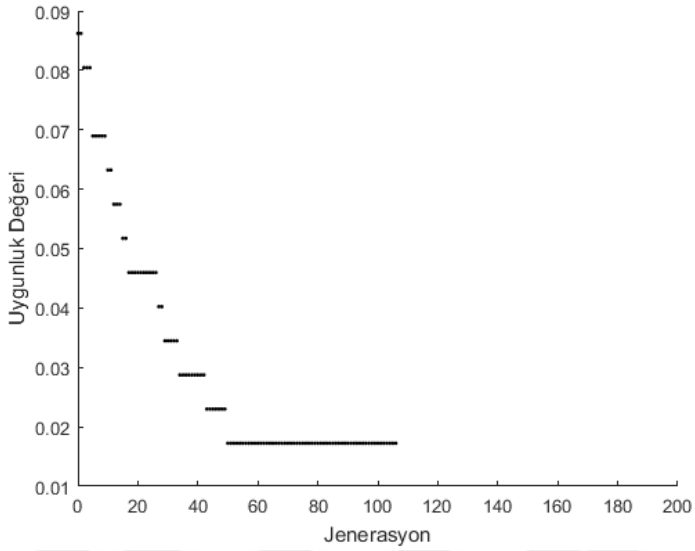
Genetik algoritma sarmal modelli öznitelik seçme algoritmasının önerilen yöntemin başarı oranına etkisini incelemek için gerçekleşen işlem adımları ile başarı oranları tüm veri setleri için aşağıdaki şekillerde verilmiştir.



Şekil 4.1. 9_Tumors veri seti GA işlem adımı – hata oranı grafiği

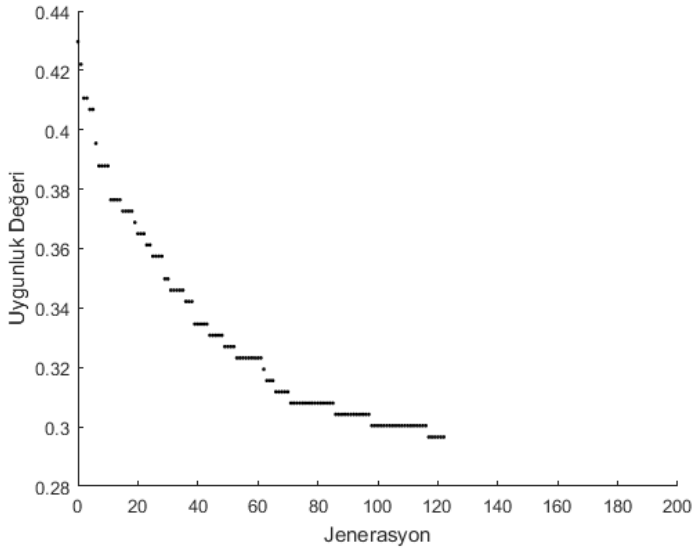
Şekil 4.1.'de 9_Tumors mikro-dizilim veri setinin k-EK sınıflandırma başarı oranının GA ile iyileştirme grafiği verilmiştir. Şekilde de görüldüğü üzere ilk işlem adımında yaklaşık 0.24 olan hata oranı, 125 işlem adımı gerçekleştikten sonra 0.066'ya kadar düşmüştür. Önerilen CFR-GA hibrit modelli öznitelik seçme

algoritmasında GA sarmal modelli öznitelik seçme aşaması görüldüğü üzere sınıflandırma başarı oranını iyileştirmektedir.



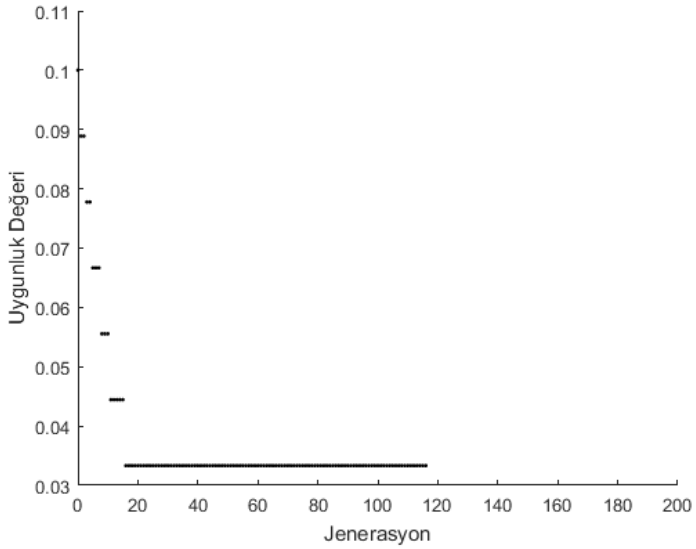
Şekil 4.2. 11_Tumors veri seti GA işlem adımı – hata oranı grafiği

CFR-GA algoritmasında 11_Tumors mikro-dizilim veri setinin sınıflandırma başarı oranına GA ciddi bir katkı sağlamaktadır. Şekil 4.2.'de de görüldüğü üzere yaklaşık 110 işlem adımdan sonra hata oranı 0.022'ye kadar düşürülmüştür.



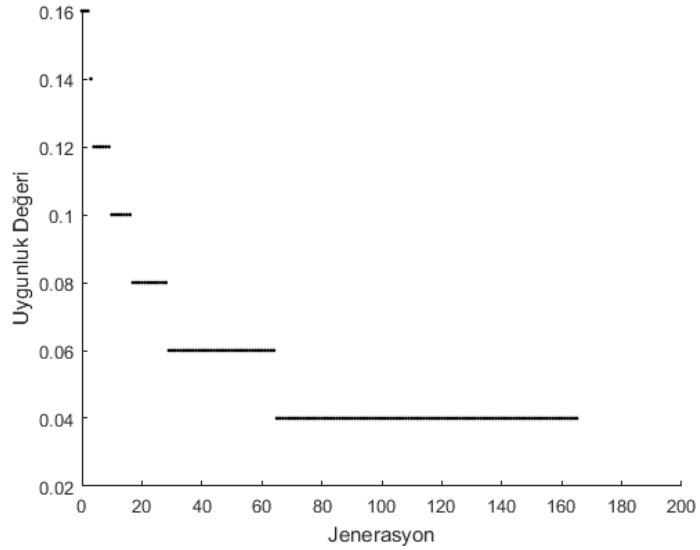
Şekil 4.3. 14_Tumors veri seti GA işlem adımı – hata oranı grafiği

Şekil 4.3.'de 14_Tumors mikro-dizilim veri seti için GA ile sağlanan iyileştirme gösterilmektedir. İlk işlem adımında yaklaşık %57 olan başarı oranı 120 işlem adımından sonra %70.35 seviyesine ulaşmaktadır. Bu iyileştirmeler GA'nın önerilen yöntemi sınıflandırma başarı oranına önemli katkılar sağladığını göstermektedir.



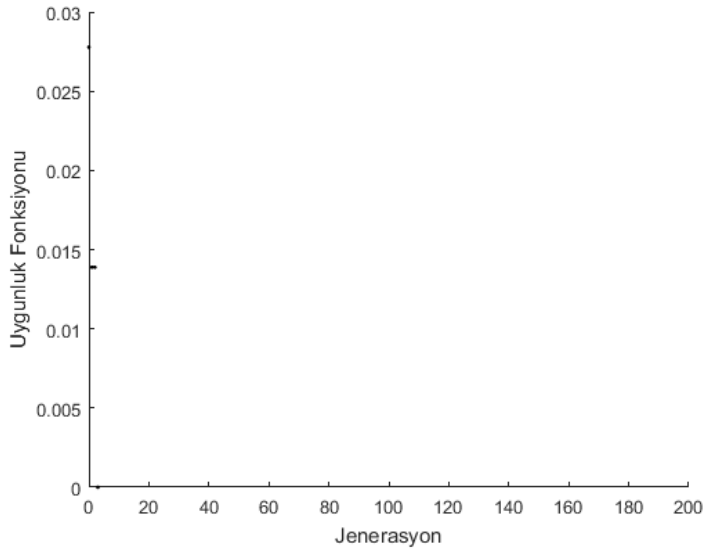
Şekil 4.4. Brain_Tumor1 veri seti GA işlem adımı – hata oranı grafiği

Şekil 4.4.'de ise Brain_Tumor1 mikro-dizilim veri seti için k-EK sınıflandırma başarı oranının GA ile iyileştirme grafiği verilmiştir. Şekilde de görüldüğü üzere ilk işlem adımında yaklaşık 0.09 olan hata oranı, 115 işlem adımı gerçekleştikten sonra 0.033'e kadar düşmüştür.



Şekil 4.5. Brain_Tumor2 veri seti GA işlem adımı – hata oranı grafiği

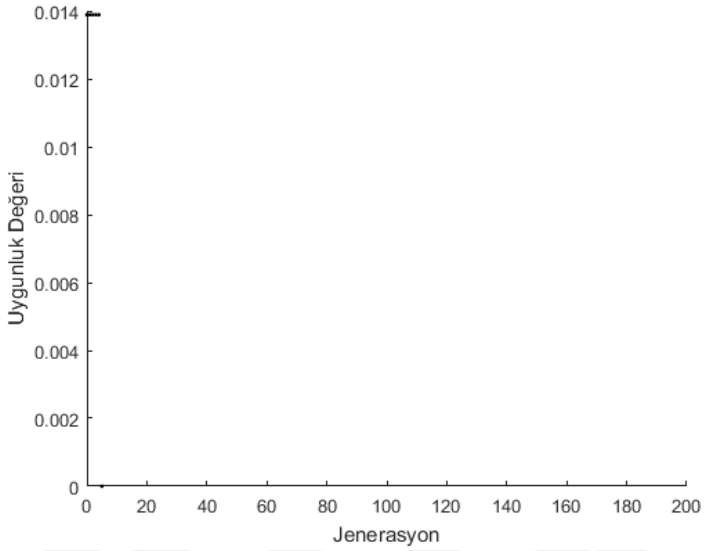
Önerilen yöntemde Brain_Tumor2 mikro-dizilim veri setinin sınıflandırma başarı oranına GA ciddi bir katkı sağlamaktadır. Şekil 4.5.'de de görüldüğü üzere yaklaşık 165 işlem adımdan sonra hata oranı 0.04'ye kadar düşürülmüştür.



Şekil 4.6. Leukemia1 veri seti GA işlem adımı – hata oranı grafiği

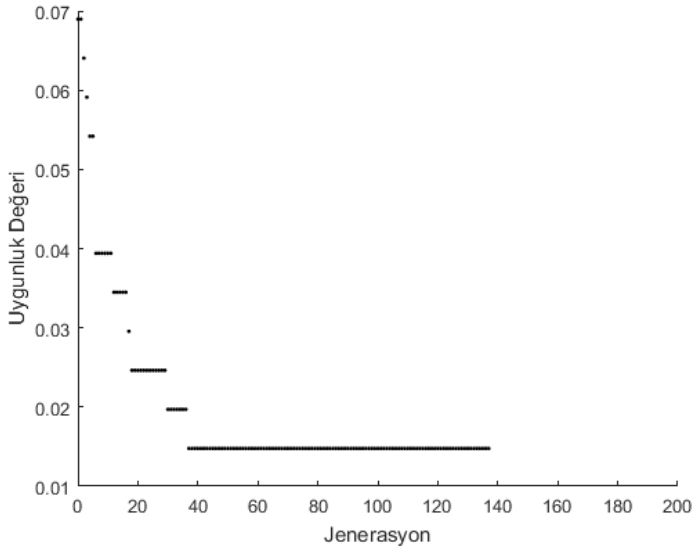
Şekil 4.6.'da Leukemia1 mikro-dizilim veri seti için GA ile sağlanan iyileştirme gösterilmektedir. İlk işlem adımında yaklaşık %98.6 olan başarı oranı 4

işlem adımından sonra %100 seviyesine ulaşmaktadır. Diğer veri setleri ile kıyaslandığında GA Leukemia1 veri setine çok fazla katkı sağlamamıştır.



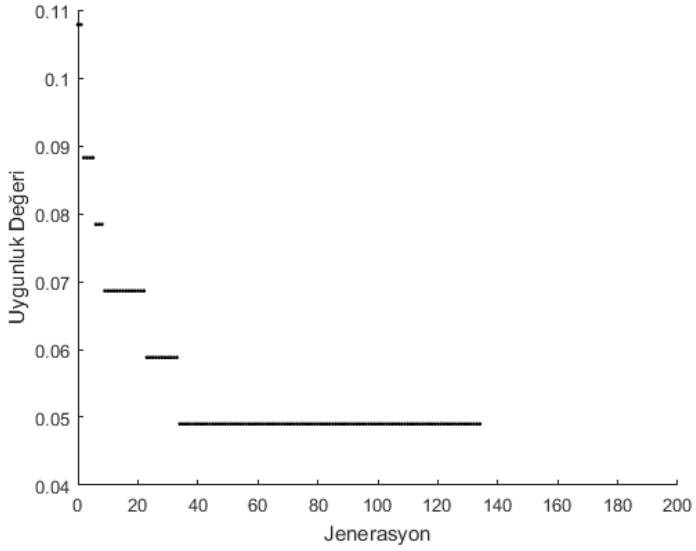
Şekil 4.7. Leukemia2 veri seti GA işlem adımı – hata oranı grafiği

Şekil 4.7.'de ise Leukemia2 mikro-dizilim veri seti için k-EK sınıflandırma başarı oranının GA ile iyileştirme grafiği verilmiştir. GA Leukemia2 mikro-dizilim veri setine çok fazla katkı sağlayamamıştır. Leukemia2 veri seti zaten CFR filtre modelli öznitelik seçme algoritmasında 95.83% sınıflandırma başarı oranını sağladığından yapılan 4 adet GA işleminden sonra sınıflandırma başarı oranı %100'ü sağlamıştır.



Şekil 4.8. Lung_Cancer veri seti GA işlem adımı – hata oranı grafiği

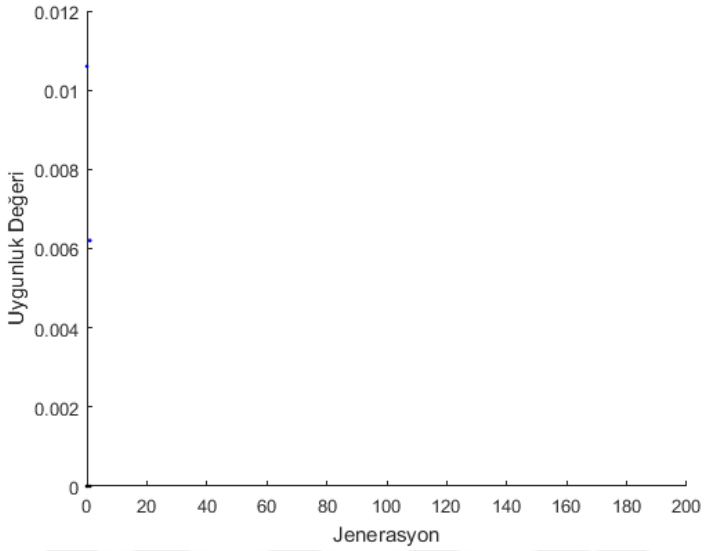
Önerilen yöntemde Lung_Cancer mikro-dizilim veri setinin sınıflandırma başarı oranına GA ciddi bir katkı sağlamaktadır. Şekil 4.8.'de de görüldüğü üzere yaklaşık 140 işlem adımdan sonra hata oranı 0.015'e kadar düşürülmüştür.



Şekil 4.9. Prostate_Tumor veri seti GA işlem adımı – hata oranı grafiği

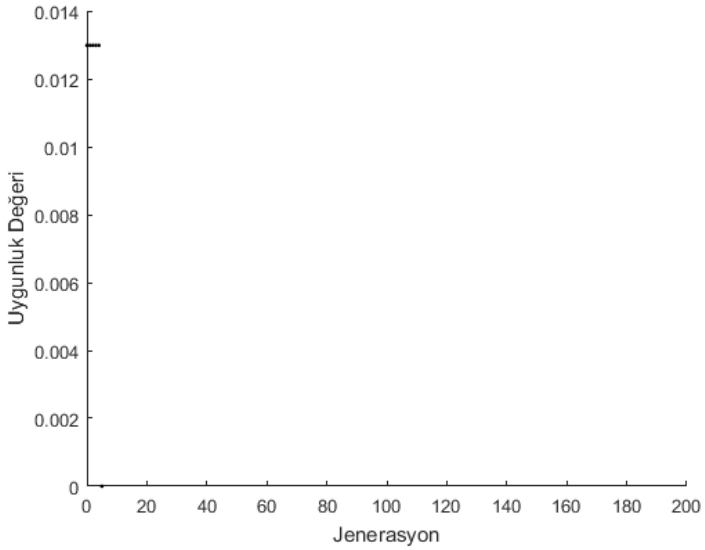
Şekil 4.9.'da ise Prostate_Tumor mikro-dizilim veri seti için k-EK sınıflandırma başarı oranının GA ile iyileştirme grafiği verilmiştir. Şekil 4.9.'da da

görüldüğü üzere ilk işlem adımında yaklaşık 0.11 olan hata oranı, 130 işlem adımı gerçekleştikten sonra 0.049'a kadar düşmüştür.



Şekil 4.10. SRBCT veri seti GA işlem adımı – hata oranı grafiği

Önerilen yöntemde SRBCT mikro-dizilim veri setinin sınıflandırma başarı oranına GA katkı sağlamamaktadır. Şekil 4.10.'da da görüldüğü üzere ilk işlem adımında bile %100 başarı oranı sağlanmıştır.



Şekil 4.11. DLBCL veri seti GA işlem adımı – hata oranı grafiği

DLBCL mikro-dizilim veri seti için GA'nın sağladığı iyileştirmeler Şekil 4.11.'de gösterilmiştir. İlk adımda 0.013 seviyesinde olan hata oranı iyileştirilerek %100 sınıflandırma başarı oranı sağlanmıştır.

Çizelge 4.6. CFR-GA hibrit modelli öznitelik seçme algoritmasının öznitelik seçilme yapılmadan sınıflandırma başarı oranları karşılaştırılması

Veri Seti	YSA	k-EK	DVM-BKB	DVM-DZG	DVM-WW	DVM-CS	DVM-BKG	CFR-GA k-EK
9_Tumors	%19.38	%43.90	%58.57	%60.24	%62.24	%65.33	%65.10	**%93.34**
11_Tumors	%54.14	%78.51	%90.36	%90.36	%94.68	%95.30	%94.68	**%98.28**
14_Tumors	%11.12	%50.40	%47.07	%47.35	%69.07	**%76.60**	%74.98	**%70.35**
Brain _Tumor1	%84.72	%87.94	%90.56	%90.56	%90.56	%90.56	%91.67	**%96.67**
Brain _Tumor2	%60.33	%68.67	%77.83	%77.83	%73.33	%72.83	%77.00	**%96.00**
Leukemia1	%76.61	%83.57	%97.32	%96.07	%97.50	%97.50	%97.50	**%100.0**
Leukemia2	%91.03	%87.14	%95.89	%95.89	%95.89	%95.89	%97.32	**%100.0**
Lung_Cancer	%87.80	%89.64	%95.59	%95.59	%95.55	%96.55	%96.05	**%98.53**
SRBCT	%91.03	%86.90	**%100.0**	**%100.0**	**%100.0**	**%100.0**	**%100.0**	**%100.0**
Prostate_Tumor	%79.18	%85.09	%92.00	%92.00	%92.00	%92.00	%92.00	**%95.10**
DLBCL	%89.64	%86.96	%97.50	%97.50	%97.50	%97.50	%97.50	**%100.0**
Ortalama	%67.73	%77.16	%85.69	%85.76	%88.02	%89.09	%89.43	**%95.29**

Çizelge 4.6.'da önerilen yöntemin sınıflandırma başarı oranı, öznitelik yapılmadan farklı sınıflandırma algoritma başarı oranları ile karşılaştırılmıştır. Koyu renkle ile belirtilen başarı oranı ilgili mikro-dizilim veri seti için elde edilen en iyi başarı oranını temsil etmektedir. Bu karşılaştırma yapılan öznitelik seçme işleminin özellikle öznitelik seçme olmadan k-EK sınıflandırma algoritmasını nasıl etkilediğini gözlemlemek amacı ile yapılmıştır. 14_Tumors mikro-dizilim veri setinde öznitelik seçme yapılmadan DVM-CS sınıflandırma algoritması en iyi başarı oranını vermiştir. Bunun yanında çizelgeden de anlaşılacağı üzere k-EK sınıflandırma algoritması önerilen CFR-GA hibrit modelli öznitelik seçme algoritması ile sınıflandırma başarı oranı bağlamında iyileştirilmiştir. Tüm mikro-dizilim veri setlerinde CFR-GA ile öznitelik seçme yapılarak uygulanan k-EK sınıflandırma algoritması, öznitelik seçme yapmadan uygulanan k-EK sınıflandırma algoritmasına göre daha başarılı sonuçlar vermiştir.

Çizelge 4.7. CFR-GA hibrit modelli öznitelik seçme algoritmasının filtre modelli öznitelik seçme algoritmaları ile karşılaştırılması

Veri Seti	KÖS k-EK	BK k-EK	CFR-GA k-EK
9_Tumors	%70.00	%66.67	**%93.34**
11_Tumors	%90.23	%83.33	**%98.28**
14_Tumors	%64.61	%56.82	**%70.35**
Brain _Tumor1	%92.22	%88.89	**%96.67**
Brain _Tumor2	%90.00	%78.00	**%96.00**
Leukemia1	%98.61	%93.06	**%100.0**
Leukemia2	**%100.0**	%91.67	**%100.0**
Lung_Cancer	%95.07	%90.15	**%98.53**
SRBCT	**%100.0**	%98.80	**%100.0**
Prostate_Tumor	%94.12	%89.22	**%95.10**
DLBCL	%96.10	%93.51	**%100.0**
Ortalama	%90.08	%84.55	**%95.29**

Önerilen yöntemin sınıflandırma başarı oranı ile filtre modelli öznitelik seçme algoritmalarının sınıflandırma başarı oranı karşılaştırılması Çizelge 4.7.'de verilmiştir. KÖS (Korelasyon Bazlı Öznitelik Seçme) korelasyon bazlı öznitelik seçme işlemi yapan bir algoritma olup nitelikler arasında korelasyonları hesaplayarak öznitelikleri seçmektedir (Hall, 1999). Karşılaştırılan diğer bir filtre modelli öznitelik seçme algoritması ise BK (Bilgi Kazancı) bilgi kazancıdır (Quinlan, 1986). Çizelge 4.7'de yapılan karşılaştırma öznitelik seçme algoritmaları arasında olduğundan sınıflandırma algoritması sabit tutulmuş olup sınıflandırma başarı oranındaki farklar öznitelik seçme algoritmalarından kaynaklanmaktadır. Bu bağlamda incelendiğinde önerilen yöntemin tüm mikro-dizilim veri setlerinde filtre modelli öznitelik seçme algoritmalarına göre daha başarılı olduğu söylenebilir. CFR filtre modelli öznitelik seçme algoritması ise Leukemia1 ve SRBCT mikro-dizilim veri setlerinde %100 sınıflandırma başarısını sağlamıştır. Belirtilen iki veri seti için eğer işlem süresi de önemli ise KÖS filtre modelli öznitelik seçme algoritması tercih edilebilir.

Çizelge 4.8. CFR-GA hibrit modelli öznitelik seçme algoritmasının sarmal modelli öznitelik seçme algoritmaları ile karşılaştırılması

Veri Seti	GA k-EK	İİPSO k-EK	CFR-GA k-EK
9_Tumors	%56.67	%78.33	**%93.34**
11_Tumors	%83.33	%93.10	**%98.28**
14_Tumors	%57.80	%66.56	**%70.35**
Brain _Tumor1	%92.22	%94.44	**%96.67**
Brain _Tumor2	%80.00	%94.00	**%96.00**
Leukemia1	%97.22	**%100.0**	**%100.0**
Leukemia2	%95.83	**%100.0**	**%100.0**
Lung_Cancer	%94.09	%96.55	**%98.53**
SRBCT	%96.39	**%100.0**	**%100.0**
Prostate_Tumor	%91.18	%92.16	**%95.10**
DLBCL	%93.51	**%100.0**	**%100.0**
Ortalama	%85.29	%92.28	**%95.29**

Çizelge 4.8.'de önerilen yöntemin sınıflandırma başarı oranı, sarmal modelli öznitelik seçme algoritmalarının sınıflandırma başarı oranları ile karşılaştırılmıştır. Bu bağlamda Genetik Algoritma (GA) ve İyileştirilmiş İkili Parçacık Sürü Optimizasyonu (İİPSO) sarmal modelli öznitelik seçme algoritmaları önerilen CFR-GA algoritması ile kıyaslanmaktadır. Chuang vd. (Chuang ve ark., 2008) önermiş olduğu İyileştirilmiş İkili Parçacık Sürü Optimizasyonu (İİPSO) temelli bir öznitelik seçme algoritması da bu kategoride önerilen CFR-GA hibrit modelli öznitelik seçme algoritması ile karşılaştırılmıştır. Çizelge 4.8.'den de anlaşılacağı üzere sınıflandırma başarı oranı bağlamında önerilen yöntem sarmal modelli öznitelik seçme algoritmalarına göre başarılıdır. Bunun yanında İİPSO sarmal modelli öznitelik seçme algoritması Leukemia1, Leukemia2, SRBCT ve DLBCL mikro-dizilim veri setlerinde %100 sınıflandırma başarı oranı ile başarılı sonuçlar vermiştir.

Çizelge 4.9. CFR-GA hibrit modelli öznitelik seçme algoritmasının hibrit modelli öznitelik seçme algoritmaları ile karşılaştırılması

Veri Seti	KÖS-TGA k-EK	BK-GA k-EK	CFR-GA k-EK
9_Tumors	%90.50	%85.00	**%93.34**
11_Tumors	**%100.0**	%92.53	%98.28
14_Tumors	**%74.39**	%65.26	%70.35
Brain _Tumor1	**%99.45**	%93.33	%96.67
Brain _Tumor2	**%100.0**	%88.00	%96.00
Leukemia1	**%100.0**	**%100.0**	**%100.0**
Leukemia2	**%100.0**	%98.61	**%100.0**
Lung_Cancer	%98.42	%95.57	**%98.53**
SRBCT	**%100.0**	**%100.0**	**%100.0**
Prostate_Tumor	**%99.22**	%96.08	%95.10
DLBCL	**%100.0**	**%100.0**	**%100.0**
Ortalama	**%96.54**	%92.21	%95.29

Önerilen yöntemin sınıflandırma başarı oranı ile hibrit modelli öznitelik seçme algoritmalarının sınıflandırma başarı oranı karşılaştırılması Çizelge 4.9.'da verilmiştir. Bu bağlamda KÖS-TGA ve BK-GA hibrit modelli öznitelik seçme algoritması ile önerilen CFR-GA algoritması karşılaştırılmıştır. Yang vd. (Yang ve ark., 2009) Bilgi kazancı ve genetik algoritmayı birleştirerek BK-GA adında bir öznitelik seçme algoritması önermiştir. Önerdiğimiz yöntem CFR-GA, (Yang ve ark., 2009)'deki yöntemden daha başarılı sonuçlar vermiştir. Chuang vd. (Chuang ve ark., 2011) Korelasyon-bazlı öznitelik seçme (KÖS) ile Taguchi - GA'ı kombine ederek önerdiği KÖS-TGA yöntemi ile mikro-dizilim veri setlerinde kayda değer iyileştirmeler yapmıştır. Önerilen yöntem CFR-GA hibrit modelli öznitelik seçme algoritması, 9_Tumors, Leukemia1, Leukemia2, Lung_Cancer, SRBCT, DLBCL mikro-dizilim veri setlerinde en iyi sınıflandırma başarı oranını sağlamıştır. Bunun yanında KÖS-TGA hibrit modelli öznitelik seçme algoritması Lung_Cancer ve 9_Tumors veri setleri dışında en iyi sınıflandırma başarı oranını sağlamıştır. Önerilen yöntem CFR-GA hibrit modelli öznitelik seçme algoritması özellikle 9_Tumors ve Lung_Cancer mikro-dizilim veri setlerine sınıflandırma başarı oranı bağlamında ciddi iyileştirmeler sağlamıştır. Bu çalışmada literatürde önerilen diğer hibrit öznitelik seçme algoritmalarından farklı olarak filtre modelli öznitelik seçme

algoritması, bir adet filtre modelli öznitelik seçme algoritmasından oluşmamakta ve 3 farklı filtre modelli öznitelik seçme algoritması kullanılmaktadır. KÖS gibi nitelikler arasında korelasyon hesaplayarak değerlilik katsayısı hesaplayan filtre modelli öznitelik seçme algoritmaları, bu çalışmada kullanılan filtre modelli öznitelik seçme algoritmalarının aksine çok hızlı değillerdir. Bu bağlamda tez çalışmasında önerilen CFR-GA hibrit modelli öznitelik seçme algoritması, filtre modelli öznitelik seçme algoritmalarının hızını ve sarmal modelli öznitelik seçme algoritmalarının yüksek başarı oranını harmanlamayı hedeflemektedir. Bu açıdan sonuçlar incelendiğinde gerek önerilen yöntemin amacına ulaşması, gerek de literatürde daha önce önerilmiş 3 farklı filtre modelli öznitelik seçme algoritmasının bu şekilde harmanlanarak hibrit modelli öznitelik seçme algoritması önerilmemiş olması çalışmaya değer kazandırmaktadır.

5. SONUÇLAR ve ÖNERİLER

Tez kapsamında önerilen CFR-GA hibrit modelli öznitelik seçme algoritmasının karşılaştırılmaları sonuç kısmında verilmiş olup bu bölümde elde edilen sonuçlar tartışılıp önerilen yöntemin olumlu olumsuz yönleri irdelenip bundan sonra bu konuda yapılabilecek çalışmalar belirtilecektir.

CFR filtre modelli öznitelik seçme ve GA sarmal modelli öznitelik seçme algoritmalarının tek tek sınıflandırma başarı oranı ve işlem süreleri incelendiğinde, tezin giriş kısmında belirtilen amaca uygun bir yöntemin önerildiği görülmektedir. Filtre modelli öznitelik seçme algoritmalarının hızını, sarmal modelli öznitelik seçme algoritmalarının sınıflandırma başarısını harmanlayan bir yöntem olarak sunulan hibrit modelli öznitelik seçme algoritması CFR-GA hibrit modelli öznitelik seçme algoritması tek başına uygulanan CFR filtre modeline ve GA sarmal modeline göre sınıflandırma başarı oranı bağlamında başarılı sonuçlar vermiştir. CFR filtre modelli öznitelik seçme algoritması CFR-GA hibrit modelli öznitelik seçme algoritmasına göre daha kısa işlem süresi ile işlem süresi bakımından daha başarılıdır. Fakat önerilen yöntemin işlem süresi bakımından incelenirken filtre modelli öznitelik seçme algoritmaları ile değil de filtreleme yapılmadan uygulanan sarmal modelli öznitelik seçme algoritmaları ile kıyaslanması gerekmektedir. Bu bağlamda bakıldığında önerilen yöntem CFR-GA hibrit modelli öznitelik seçme algoritması GA sarmal modelli öznitelik seçme algoritmasından daha kısa işlem süresi ile işlem süresi bakımından daha başarılıdır. Genel olarak önerilen CFR-GA hibrit modelli öznitelik seçme algoritması filtre modelli öznitelik seçme algoritmalarına sınıflandırma başarı oranı bakımından, sarmal modelli öznitelik seçme algoritmalarına da işlem süre bakımından ciddi olumlu katkılar sağlamıştır.

Sonuçlar incelendiğinde genel olarak önerilen yöntem CFR-GA hibrit modelli öznitelik seçme algoritması mevcut yöntemlere iyileştirmeler yaptığı görülmektedir. CFR-GA hibrit modelli öznitelik seçme algoritmasının CFR filtre modelli öznitelik seçme ve GA sarmal modelli öznitelik seçme şeklinde iki aşamadan oluştuğu

düşünüldüğünde, CFR filtre aşamasının SRBCT veri setinde yeterli olduğu görülmektedir. SRBCT mikro-dizilim veri setinde GA uygulanmadan da en başarılı sonuçlar alınabilmektedir. GA en büyük iyileştirmeyi sırasıyla 9_Tumors, Brain_Tumor2, 14_Tumors, Brain_Tumor1, Prostate_Tumor ve Lung_Cancer veri setlerinde sağlamıştır. Bu iyileştirmelerde hibrit modelli öznitelik seçme algoritmalarının gerekliliğini göstermektedir. Mikro-dizilim veri setleri arasında en çok sınıf sayısına sahip 14_Tumors veri seti en iyi sınıflandırma başarı oranını öznitelik seçme işlemi olmadan DVM-CS sınıflandırma metodunda vermiştir. Bu sonuçtan da anlaşılacağı gibi bu veri setinde uygulanan öznitelik seçme algoritmaları bilgi kaybına sebep olarak sınıflandırma başarı oranını düşürmektedir. CFR-GA metodu 9_Tumors ve Lung_Cancer veri setlerinde iyileştirmeler yaparak en iyi sınıflandırma başarı oranını yakalamıştır. Bunun yanında SRBCT, DLBCL, Leukemia1 ve Leukemia2 veri setlerinde de en iyi sınıflandırma başarı oranı olan %100 başarı oranını yakalamıştır. Mikro-dizilim veri setleri genel olarak ele alındığında 14_Tumors veri seti haricinde en iyi sonuçlar her zaman hibrit modelli öznitelik seçme algoritmaları ile sağlanmıştır. Bu da bize mikro-dizilim verilerinde hibrit modelli öznitelik seçme algoritmalarının kullanılmasının gereklilik olduğunu göstermektedir.

Bundan sonra bu çalışmaya ek olarak KÖS-TGA metoduna göre daha düşük sınıflandırma başarı oranı veren 11_Tumors, Brain_Tumor1, Brain_Tumor2 ve Prostate_Tumor veri setleri için farklı filtre kombinasyonları denenerek iyileştirilebilir. Birleştirme işlem adımı geri beslemeli yapılarak sınıflandırma başarı oranları açısından daha başarılı sonuçlar alınabilir.

Ayrıca GA dışında Parçacık sürü optimizasyonu, Karınca koloni optimizasyonu gibi sarmal modelli öznitelik seçme algoritmaları denenerek daha iyi sonuçlar alınabilir. Bunlara ek olarak öznitelik seçme algoritmaları kullanıldığında bilgi kaybına yol açan 14_Tumors veri setine uygun bir hibrit öznitelik seçme algoritması önerilebilir. Bu kısımda özellikle filtre modelli öznitelik seçme algoritmalarının öznitelik seçme kriterleri esnekleştirilip daha çok öznitelik ile sarmal modelli öznitelik seçme işlem adımına başlanabilir.

Mikro-dizilim veri setlerinin daha başarılı sınıflandırılması adına k-EK değil de DVM modelli bir sınıflandırma algoritması ile öznitelik seçme algoritmaları test edilebilir. Zaten sonuç kısmında verilen sınıflandırma başarı oranları incelendiğinde DVM modelli sınıflandırma algoritmalarının k-EK sınıflandırma algoritmasına göre daha iyi sonuçlar verdiği görülmektedir. Bu çalışmada ise literatürde mevcut önerilen hibrit öznitelik seçme algoritmaları ile önerilen öznitelik seçme algoritmasının öznitelik seçme bakımından karşılaştırma yapılabilmesi açısından sınıflandırma algoritması olarak k-EK sınıflandırma algoritması seçilmiştir. Mevcut çalışmalar k-EK ile değerlendirme yaptığından bu çalışmada da k-EK ile değerlendirme yapılmıştır.

KAYNAKLAR

ALSHAMLAN, HALA M., GHADA H. BADR, AND YOUSEF A. ALOHALI. 2015 "Genetic Bee Colony (GBC) Algorithm: A New Gene Selection Method for Microarray Cancer Classification." Computational Biology and Chemistry 56 (June 2015): 49–60. doi:10.1016/j.compbiolchem.2015.03.001.

ANTONIADIS, A., LAMBERT-LACROIX, S., LABLANC, F., 2003. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics, 5:563-570.

AYDILEK, I.B., ARSLAN, A., 2013 A hybrid method for imputation of missing values using optimized Fuzzy c-means with Support Vector Regression and a Genetic Algorithm, Information Sciences, 233;25–35.

BATTITI R., 1994 Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks 5:537–550.

BEDNAR, M., 2000, DNA Microarray Technology and Application, Med Sci Monid, 6: 796–800.

BEN-BASSAT,M., 1982 Pattern recognition and reduction of dimensionality. Handbook of Statistics II, Vol. 1.North-Holland, Amsterdam. pp. 773–791.

BOLÓN-CANEDO V., SÁNCHEZ-MAROÑO N., ALONSO-BETANZOS A., BENITEZ J.M., HERRERA F. 2014 A review of microarray datasets and applied feature selection methods Information Sciences, 282:111–135

BRAIK, M., SHETA, A., ARIEQAT, A., 2008 A Comparison between GAs and PSO in Training ANNto Model the TE Chemical Process Reactor, AISB 2008 Convention Communication, Interaction and Social Intelligence.

CHANG, J.F., 2009 A Performance Comparison Between Genetic Algorithms And Particle Swarm Optimization Applied In Constructing Equity Portfolios, International Journal of Innovative Computing, Information and Control ICIC International ISSN 1349-4198 5:12(B), Aralık 2009.

CHIN YL, DERIS S., 2005 A study on gene selection and classification algorithms for classification of microarray gene expression data. Journal Technology, 43:111-24.

CHUANG L., CHANG H., TU C., YANG C., 2008 IG-GA: Improved binary PSO for feature selection using gene expression data, Computational Biology and Chemistry, 32:29–38.

CHUANG L., YANG C., WU K., YANG C., 2011. A hybrid feature selection method for dna microarray data, Comput. Biol. Med., 41:228–237.

CLOAREC, J.P., CHEVOLOT, Y., LAURENCEAU, E., PHANER-GOUTORBE, M., SOUTEYRAND, E., 2008, A multidisciplinary approach for molecular diagnostics based on biosensors and microarrays, ITBM-RBM, 29:105–127.

COVER T. , HART P., 1967 Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13:21–27.

DUDA R.O., HART P.E., AND STORK D.G., 2001 Pattern Classification, John Wiley & Sons, New York.

HALL,M., 1999 Correlation-based feature selection for machine learning. Waikato University, Department of Computer Science, PhD Thesis., New Zealand, 169s.

HENGPRAPHROM, K., NA WICHIAN, S., MEESAD, P., 2010 Missing Value Imputation Using Genetic Algorithm, Third International Symposium on Intelligent Informatics (ISII2010) First International Symposium on Information and Knowledge Management (ISIKM2010) Ekim 5 – 7.

HOLLAND J., 1975 Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor.

HUANG, JINJIE, YUNZE CAI, AND XIAOMING XU. 2007 "A Hybrid Genetic Algorithm for Feature Selection Wrapper Based on Mutual Information." Pattern Recognition Letters 28, no. 13 (October 2007): 1825–44. doi:10.1016/j.patrec.2007.05.011.

HSU H., HSIEH C., LU M., 2011 Hybrid feature selection by combining filters and wrappers, Expert Systems with Applications, 38:8144–8150.

JONES, G., WILLETT, P., GLEN, RC., 1995 Molecular Recognition Of Receptor-Sites Using A Genetic Algorithm With A Description Of Desolvation, Journal Of Molecular Biology, 245:43-53.

KAVZOĞLU TAŞKIN, ŞAHİN EMREHAN KUTLUĞ, ÇÖLKESEN İSMAİL, Heyelan Duyarlılık Analizinde Ki-Kare Testine Dayalı Faktör Seçimi, V. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu (UZALCBS 2014), 14-17 Ekim 2014, İstanbul

KAZARLIS, SA., BAKIRTZIS, AG., PETRIDIS, V., 1996 A genetic algorithm solution to the unit commitment problem, IEEE Transactions On Power Systems, 11:83-90.

KITTLER,J. 1978 Pattern Recognition and Signal Processing, Chapter Feature Set Search Algorithms Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, pp. 41–60.

KUMAR, A., GOEL, G., FEHRENBACH, E., PUNIYA, A. K., SINGH, K., 2005, Microarrays: The Technology, Analysis and Application, Eng. Life Sci., 5:3.

KOHAVI R., JOHN G.H., 1997 Wrappers for feature subset selection, Artificial Intelligence, 97:273–324.

LEE J., LEE J.B., PARK M., SONG S.H., 2005. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis, 48: 869-885.

LI L, WEINBERG CR, DARDEN TA, PEDERSEN LG., 2001 Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics, 17:1131-1142.

LI, Y., YANG, Y., ZHOU, L., ZHU, R., Observations On Using Problem-Specific Genetic Algorithm For Multiprocessor Real-Time Task Scheduling, International Journal of InnovativeComputing, Information and Control ICIC International 2009 ISSN 1349-4198 Volume 5, Number 9, September 2009.

LI, YING, GANG WANG, HUILING CHEN, LIAN SHI, AND LEI QIN. 2013 "An Ant Colony Optimization Based Dimension Reduction Method for High-Dimensional Datasets." Journal of Bionic Engineering 10, no. 2 (April 2013): 231–41. doi:10.1016/S1672-6529(13)60219-X.

LIU H., SETIONO R., 1995 Chi2: Feature Selection and Discretization of Numeric Attributes. In: Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence, 338-391.

MOLLA, M., WADDELL M., PAGE D., SHAVLİK J., 2004. Using machine learning to design and interpret gene-expression microarrays. Ai Magazine, 25(1): 23-44.

PARRISH RS, SPENCER HJ, XU P., 2009 Distribution modelling and simulation of gene expression data. Computational Statistics and Data Analysis

PIROOZNIA M., YANG J.Y., YANG M.Q., DENG Y., 2008. A comparative study of different machine learning methods on microarray gene expression data. Bmc Genomics, 9:1-13.

ROBNİK-SIKONJA M., KONONENKO I., 2003 Theoretical and empirical analysis of ReliefF and RReliefF Mach. Learn. 53:23–69.

QUINLAN J.R., 1986 Induction of decision trees, Machine Learning 1:81–106.

QUINLAN, J. R., 1993 C4. 5: Programs for Machine Learning, Machine Learning, 16: 235-240.

SAEYS Y., INZA I., LARRANAGA P., 2007 A review of feature selection techniques in bioinformatics. Bioinformatics, 19: 2507–2517.

SASSANFAR, S., WALKER, G., 2003, DNA Microarray Technology. What Is It and How Is It Useful, MIT, Biology Science Outreach.

SEBBAN, MARC, RICHARD NOCK. 2002 "A Hybrid Filter/wrapper Approach of Feature Selection Using Information Theory." Pattern Recognition 35, no. 4 (April 2002): 835–46. doi:10.1016/S0031-3203(01)00084-X.

SKALAK,D., 1994 Prototype and feature selection by sampling and randommutation hill climbing algorithms. In Proceedings of the Eleventh International Conference on Machine Learning, pp. 293–301.

STATNIKOV A., TSAMARDINOS I., DOSBAYEV Y., ALIFERIS C.F., 2005. GEMS: A System for Automated Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data, International Journal of Medical Informatics, 74:491-503.

TABAKHI S., NAJAFI A., RANJBAR R., MORADI P., 2015 Gene selection for microarray data classification using a novel ant colony optimization, Neurocomputing, 168:1024–1036.

TAHİR, MUHAMMAD ATİF, AHMED BOURIDANE, AND FATİH KURUGOLLU. 2007 "Simultaneous Feature Selection and Feature Weighting Using Hybrid Tabu Search/K-Nearest Neighbor Classifier." Pattern Recognition Letters 28, no. 4 (March 2007): 438–46. doi:10.1016/j.patrec.2006.08.016.

XU R., DAMELİN S., NADLER B., WUNSCH D.C., 2010. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps. Artificial Intelligence in Medicine, 48(2-3): 91-98.

WALTERS DC., SHEBLE GB., 1993 Genetic Algorithm Solution Of Economic-Dispatch With Valve Point Loading, IEEE Transactions On Power Systems, 8:1325-1332.

YANG C., CHUANG L., YANG C., 2009 IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data, Journal of Medical and Biological Engineering, 30:23-28.

YANG, L.Y., XU, L.S., 2011 Topological properties of generalized approximation spaces, Inform Sciences, 181:3570-3580.

YOLTAŞ A., KARABOZ İ., 2010. DNA Mikroarray Teknolojisi ve Uygulama Alanları. Elektronik Mikrobiyoloji Dergisi, 8(1): 01-19.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Abdülkadir GÜMÜŞÇÜ
Uyruğu : T.C.
Doğum Yeri ve Tarihi : Şanlıurfa – 06/01/1985
Telefon : 0414 318 30 00 - 1041
Faks : -
e-mail : agumuscu@harran.edu.tr

EĞİTİM

Derece	Adı, İlçe, İl	Bitirme Yılı
Lise	: Şanlıurfa Anadolu Lisesi, Şanlıurfa	2002
Üniversite	: İnönü Üniversitesi, Malatya	2006
Yüksek Lisans	:Gaziantep Üniversitesi, Gaziantep	2012
Doktora	:	

İŞ DENEYİMLERİ

Yıl	Kurum	Görevi
2009-Halen	Harran Üniversitesi Mühendislik Fak.	Araştırma Görevlisi
2007-2009	SİSTEK A.Ş.	Otomasyon Müh.

UZMANLIK ALANI

Görüntü İşleme, Makina Öğrenmesi

YABANCI DİLLER

İngilizce – ÜDS:70
Almanca – Başlangıç düzeyinde

YAYINLAR

A.Gümüşçü, E.Erçelebi, "Gerçek Zamanlı Şüpheli Paket Tespiti", MASFOR 2012, İstanbul (Yüksek Lisans)

A.Gümüşçü, M.E.Tenekeci "Analysing effect of Dimension reduction algorithms on Spectroradiometer data classification", ICONSETE2015, Vienna (Doktora)

A.Gümüşçü, R.Taşaltın, İ.B. Aydilek "C4.5 Karar Ağaçlarında Genetik Algoritma ile Budama", DUFED (Doktora) (Kabul Edildi)

A.Gümüşçü, R.Taşaltın, İ.B. Aydilek "3 Farklı Filtre Modelli Öznitelik Seçme Algoritmalarının Kombine Edilerek İyileştirilmesi", 1st International Conference on Engineering Technology and Applied Sciences, Afyon Kocatepe University, Turkey 21-22 April 2016 (Doktora)

A.Gümüşçü, R.Taşaltın, İ.B. Aydilek "Mikro-dizilim Veri Sınıflandırmasında Öznitelik Seçme Algoritmalarının Karşılaştırılması", 1st International Conference on Engineering Technology and Applied Sciences, Afyon Kocatepe University, Turkey 21-22 April 2016 (Doktora)

A.Gümüşçü, R.Taşaltın, İ.B. Aydilek, M.E. Tenekeci "Comparing Filter Modelled Feature Selection Algorithms on Microarray Data Classification", ICONSETE2016, Barcelona (Doktora) (Kabul Edildi)

A.Gümüşçü, İ.B. Aydilek, R.Taşaltın "Mikro-dizilim Veri Sınıflandırmasında Öznitelik Seçme Algoritmalarının Karşılaştırılması", Harran Üniversitesi Mühendislik Dergisi, 2016(1):1-7 (Doktora)

A.Gümüşçü, İ.B. Aydilek, R.Taşaltın "3 Farklı Filtre Modelli Öznitelik Seçme Algoritmalarının Kombine Edilerek İyileştirilmesi", Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, 2016(Özel Sayı):31-35 (Doktora)