SAKARYA ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ DERGİSİ
http://dergipark.gov.tr/saufenbilder

Karaciğer mikrodizi kanser verisinin sınıflandırılması için genetik algoritma kullanarak ANFIS'in eğitilmesi

Bülent Haznedar[1*], Mustafa Turan Arslan[2], Adem Kalınlı[3]

21.06.2016 Geliş/Received, 30.11.2016 Kabul/Accepted

doi: 10.16984/saufenbilder.41925

ÖZ

Sınıflandırma, verilerin analiz edilmesi için önemli bir veri madenciliği tekniği olup tıp, genetik ve biyomedikal mühendisliği başta olmak üzere birçok alanda kullanılmaktadır. Özellikle tıp alanında DNA mikrodizi gen ekspresyon verilerini sınıflandırmaya yönelik yapılan çalışmalarda artış görülmektedir. Ancak, mikrodizi gen ekspresyon (ifade) verilerinde bulunan gen sayılarının çokluğu ve bu veriler arasında doğrusal olmayan bağıntılar bulunması gibi problemlerden dolayı geleneksel sınıflandırma algoritmalarının başarımları sınırlı kalabilmektedir. Bu sebeplerden dolayı son yıllarda sınıflandırma probleminin çözümü için yapay zekâ tekniklerine dayalı sınıflandırma yöntemlerine olan ilgi giderek artmaya başlamıştır. Bu çalışmada, karaciğer mikrodizi kanser veri setinin sınıflandırılması için Uyarlamalı Ağ Tabanlı Bulanık Mantık Çıkarım Sistemi (ANFIS) ve Genetik Algoritmaya (GA) dayalı hibrid bir yaklaşım önerilmiştir. Simülasyon sonuçları, diğer bazı yöntemlere ait sonuçlarla karşılaştırılmıştır. Elde edilen sonuçlardan, önerilen yöntemin diğer yöntemlere göre daha başarılı olduğu görülmüştür.

Anahtar Kelimeler: Neuro-fuzzy, ANFIS, genetik algoritma, sınıflandırma, mikrodizi gen ifade

Training ANFIS structure using genetic algorithm for liver cancer classification based on microarray gene expression data

ABSTRACT

Classification is an important data mining technique, which is used in many fields mostly exemplified as medicine, genetics and biomedical engineering. The number of studies about classification of the datum on DNA microarray gene expression is specifically increased in recent years. However, because of the reasons as the abundance of gene numbers in the datum as microarray gene expressions and the nonlinear relations mostly across those datum, the success of conventional classification algorithms can be limited. Because of these reasons, the interest on classification methods which are based on artificial intelligence to solve the problem on classification has been gradually increased in recent times. In this study, a hybrid approach which is based on Adaptive Neuro-Fuzzy Inference System (ANFIS) and Genetic Algorithm (GA) are suggested in order to classify liver microarray cancer data set. Simulation results are compared with the results of other methods. According to the results obtained, it is seen that the recommended method is better than the other methods.

Keywords: Neuro-fuzzy, ANFIS, genetic algorithm, classification, microarray gene expression

* Sorumlu Yazar / Corresponding Author
1 Hasan Kalyoncu Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, Gaziantep – bulent.haznedar@hku.edu.tr
2 Mustafa Kemal Üniversitesi, Kırıkhan Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Hatay - mtarslan@mku.edu.tr
3 Erciyes Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği, Kayseri - kalinlia@erciyes.edu.tr

1. GİRİŞ (INTRODUCTION)

Günümüzde teknolojik gelişmeler ve gelişmiş donanımlar sayesinde artık yüksek boyutlu verileri depolamak mümkün hale gelmiştir. Bu yüzden yüksek boyutlu verilerden anlamlı bilgiler çıkarmak için veri madenciliği fonksiyonlarının kullanımında son yıllarda artış görülmüştür. Yüksek boyutlu verilere genel olarak genetik çalışmalar sonucu elde edilen biyolojik veriler gösterilebilir.

İnsan yapısındaki dokuların hepsi aynı genetik materyali içermektedir. Fakat her hücrede aynı genler aktif değildir. Hangi genlerin aktif olup hangilerinin aktif olmadığı bilgisi, hücrelerin nasıl bir fonksiyona sahip olduğu ve bazı genler normal şekilde çalışmadığında hücrenin bu olaydan nasıl etkileneceği bilgisini vermektedir. Hücrenin gen ifadesi işte bu aktiflik bilgisine eşittir [1]. Bütün organlarımız aynı genetik materyali içermesine rağmen genlerin farklı hücrelerde farklı ifade edilmelerinden dolayı böbrek, beyin, karaciğer gibi hücreler birbirinden farklı fonksiyonlara sahiptir [2].

Bilgisayar teknolojisinin moleküler biyoloji ile birlikte gelişmesi sonucunda biyoteknolojinin kavramsal olarak ulaşabileceği noktalardan biri olan mikrodizi çipler (gen çip) ortaya çıkmıştır [3]. Mikrodizi çip teknolojisi sayesinde aynı anda birden fazla genin incelenmesi ve genlerin birlikte araştırılması fırsatı doğmuştur [4].

DNA mikrodizi teknolojisi sayesinde yüzlerce genin ifade düzeyleri eş zamanlı olarak incelenebilmektedir [5]. Bu teknoloji sayesinde hasta ve sağlıklı hücrelerin gen ifadelerinin karşılaştırılması neticesinde hastalığa neden olan genlerin belirlenmesi mümkün olabilmektedir [6]. Mikrodizi yöntemi kullanarak gen ifade verilerinin analizi sayesinde psikolojik, kalp ve bulaşıcı hastalıklar gibi birçok hastalık hakkında bilgi vermektedir. Elde edilen bu bilgiler sayesinde hastalıkların önlemesi yönünde tedbirler alınabilmektedir. Bu teknik sayesinde ayrıca gen üzerinde taşınan kalıtsal hastalıklar, alt türlerine kadar incelenebilmektedir [7].

DNA mikrodizi teknolojisi ile bir çalışmada çok fazla veri analizi yapıldığından sonuçlar için çok fazla zaman gerekmektedir. Elde edilen sonuçlar ise yüzlerce genin ifadesini barındırdığından dolayı sonuçları yorumlamanın bir hayli zor olduğu ve çok pahalı bir yöntem olduğundan dolayı çok fazla deneylerin tekrarlanamaması ise dezavantajlarını oluşturur [7].

Mikrodizi gen ifade verilerini sınıflandırmak için literatürde Bayes Ağları, Destek Vektör Makineleri ve Karar Ağaçları gibi istatistiksel yöntemler sıklıkla kullanılmaktadır. Diaz-Uriarte ve Andres [8], rastgele orman algoritmasını kullanarak 9 farklı mikrodizi veri seti üzerinde hem gen seçimi hem de sınıflandırma yapmışlardır. Dudoit ve arkadaşları [9], gen ifade verisi kullanarak tümör sınıflandırması için fisher lineer diskriminant analizi, k-en yakın komşu algoritması ve torbalama yöntemlerini kullanmışlardır. Furey ve arkadaşları [10], mikrodizi ifade verilerini kullanarak kanser dokusu örneklerini sınıflandırmak için destek vektör makinelerini kullanmışlardır. Lee ve arkadaşı [11], kan kanseri gen ifade verisini kullanarak destek vektör makineleri yardımıyla çok sınıflı sınıflandırma yapmışlardır. Liu ve arkadaşları [12], gen ifade profillerini kullanarak k-en yakın komşu, naive bayes, destek vektör makineleri ve C4.5 karar ağacı gibi sınıflandırma yöntemlerinin karşılaştırmalı bir performans analizini yapmışlardır. Statnikov ve arkadaşları [13] ise mikrodizi gen ifade kanser teşhisinde destek vektör makineleri, k-en yakın komşu ve yapay sinir ağları gibi çok sınıflı yöntemler kullanılarak bu sınıflandırma yöntemlerinin karşılaştırmalı bir değerlendirmesini yapmışlardır. Mikrodizi verilerinde yüzlerce gen bulunması ve bu genler arasında doğrusal olmayan ilişkiler bulunması sebebiyle geleneksel yöntemlerin başarımları sınırlı kalabilmektedir. Bu zorluklar bilim adamlarını daha güçlü ve modern yöntemler keşfetmeye yöneltmiştir. Bu amaçla, araştırmacıların ilgisi zor problemlerin çözümünde yapay zekâ tabanlı yöntemlerin kullanılmasına yönelik olarak artmaya başlamıştır. Khan ve arkadaşları [14], yapay sinir ağları ve gen ifade profillerini kullanarak kanser hastalıkların tahmin edilmesi ve sınıflandırılması üzerine bir çalışma yapmışlardır. Ringner ve Peterson [15], yapay sinir ağları yardımıyla mikrodizi tabanlı kanser teşhisi üzerine bir çalışma gerçekleştirmişlerdir.

Son yıllarda yüksek boyutlu problemlerin çözümünde yapay zekâ tabanlı yöntemlerin kullanılması yaygınlaşmış ve öznitelik seçme ve sınıflandırma problemlerine karşılık algoritmalar geliştirilmiştir. Pirooznia ve arkadaşları [16], mikrodizi gen ifade (ekspresyon) verilerini kullanarak sınıflandırma çalışması yapmışlardır. Bu çalışmada istatistiksel yöntemler ile yapay zekâ yöntemleri karşılaştırılmış ve yapay zekâ yöntemlerinin daha iyi sonuçlar verdiği görülmüştür. Loganathan ve Girijia [17], ANFIS ağırlık parametrelerini Runge Kutta Learning algoritması ile eğitmişler ve oluşturulan modelin performansını, mikrodizi gen ekspresyon kanser verilerini sınıflandırarak ölçmüşlerdir. Kumar ve Punithavalli [18], ANFIS ile kanser verilerini sınıflandırmış ve elde edilen sonuçları istatistiksel yöntemler ile karşılaştırmışlardır.

Bu çalışmada karaciğer kanseri mikrodizi gen ekspresyon verilerini doğru sınıflara atayabilmek için popülasyon tabanlı bir algoritma olan genetik algoritma kullanılarak ANFIS modelinin parametrelerinin

eğitilmesine yönelik bir yaklaşım önerilmiştir. Önerilen yaklaşımın performansı k-katlı çapraz doğrulama yöntemi ile test edilmiş ve k parametresi 5 seçilmiştir. Önerilen yöntem ile bulunan sonuçların performansı diğer yöntemlerin performansları ile karşılaştırılmıştır. Bir sonraki bölümde ise kullanılan yöntemler hakkında bilgi verilmiş, 3. bölümde algoritmanın eğitildiği karaciğer kanser veri seti tanıtılmıştır. 4. Bölümde ise önerilen yöntem hakkında bilgi verilmiş ve elde edilen sonuçlar ise bölüm 5'te sunulmuştur.

2. METOTLAR (METHODS)

Bir sınıflandırma fonksiyonu temelde iki bölümden oluşur. Bunlar, öznitelik seçimi ve verilerin doğru sınıflara atanmasıdır. Bu çalışmada kullanılan sistem bileşenleri aşağıda açıklanmaktadır.

2.1. Öznitelik Seçme (Feature Selection)

Öznitelik seçme olayı, öznitelik uzayı içerisinden optimal bir öznitelik alt kümesinin bulunmasıdır. Ayrıca optimal öznitelik alt kümesinin bulunması ile problemin boyutunun indirgenmesi ve öznitelik kümesini en uygun alt kümenin temsil etmesidir. Mikrodizi verileri yüksek boyutlu veriler olduklarından dolayı sınıflandırma yöntemlerinin performanslarını arttırmak için öznitelik seçme yönteminin kullanılmasının bir gereklilik olduğundan söz edilebilir.

2.1.1. Korelasyon Tabanlı Öznitelik Seçimi (KTÖS-Correlation Based Feature Selection)

KTÖS, Sınıf ile yüksek korelasyona sahip ve birbiriyle düşük korelasyona sahip özniteliklerin bir arada bulunduğu alt kümeleri oluşturur.

$$G_S = \frac{k\overline{r}_{ci}}{\sqrt{k + k(k-1)\overline{r}_{ii'}}} \quad (1)$$

Eşitlik (1)'de yer alan *k* parametresi veri alt kümesinin öznitelik sayısı, *rci* parametresi ortalama öznitelik korelasyonu ve *rii*' parametresi ise ortalama öznitelik iç korelasyonudur [19].

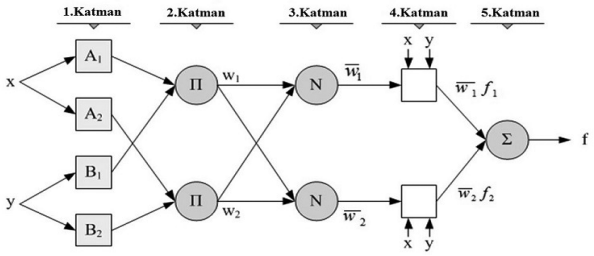
2.2. ANFIS ile Sınıflandırma (Classification with ANFIS)

Sınıflandırma işlemi, ağ tabanlı bulanık mantık çıkarım sisteminde (ANFIS) bulunan başlangıç ve sonuç parametrelerinin genetik algoritması (GA) ile optimize edilmesi ile oluşan model kullanılarak yapılmıştır.

2.2.1. Adaptif Ağ Tabanlı Bulanık Mantık Çıkarım Sistemi (Adaptive Neuro-Fuzzy Inference System, ANFIS)

Adaptif ağ tabanlı bulanık mantık çıkarım sistemi (ANFIS), Takagi Sugeno bulanık modeli dikkate alınarak geliştirilmiş bir yapay sistemdir [20]. Sinir ağlarının geriye yayılımlı öğrenme yeteneği ile bulanık mantığın sonuç çıkarma özellikleri ANFIS yapısında birleştirilmiştir. Üyelik fonksiyonları ile bulanıklaştırdığı giriş verilerini bulanık kurallar ile ağ üzerinde dağıtarak çıkış hesaplamaktadır. Bu süreç ANFIS modeline çıkarım yeteneği sağladığı için tahmin problemlerinde başarımı oldukça yüksektir.

ANFIS başlangıç ve sonuç parametreleri olmak üzere iki tür parametreye sahiptir. Bu iki parametre türü bulanık kuralları birbirine bağlar. Modelin eğitimi ise bu parametrelerin optimizasyonu ile sağlanır. ANFIS temel olarak beş katmandan oluşmaktadır. Şekil 1'de iki giriş ve bir çıkıştan oluşan temel bir ANFIS yapısı verilmiştir [21].



Şekil 1. Temel bir ANFIS yapısı [22] (A basic structure of ANFIS)

1.Katman

Bulanıklaştırma katmanı olarak adlandırılan bu katmandaki her düğümden alınan sinyal diğer katmana aktarılır. Her düğümden alınan sinyal giriş değerlerine ve kullanılan üyelik fonksiyonunun türüne bağlı olarak oluşmaktadır. Bu katmandaki düğümlerin çıkışları(O_{1i}) için Eşitlik (2) ve Eşitlik (3) aşağıda verilmiştir [23].

$$O_{1i} = \mu A_i(x) \qquad i = 1,2 \quad (2)$$

$$O_{1i} = \mu B_{i-2}(x) \qquad i = 3,4 \quad (3)$$

Burada A_i ve B_i girişlere ait herhangi bir üyelik fonksiyonu olup μA_i ve μB_i ise bu fonksiyon için hesaplanmış üyelik dereceleridir. Çan eğrisi türünde üyelik fonksiyonu için μA_i aşağıdaki eşitlik ile hesaplanmaktadır

.

$$\mu A_i = e^{-\frac{1}{2}(\frac{x-c}{a})^2} \qquad i = 1,2 \tag{4}$$

Burada a_i ve c_i sırasıyla, üyelik fonksiyonunun sigma ve merkez parametreleridir.

2.Katman

Kural katmanı olarak adlandırılan bu katmanda bir önceki katmandan gelen üyelik dereceleri kullanılarak her kuralın ateşleme seviyesi hesaplanmaktadır.

$$O_{2i} = w_i = \mu A_i(x).\mu B_i(y) \qquad i = 1,2 \tag{5}$$

3.Katman

Normalizasyon katmanı olarak adlandırılmaktadır. Bu katmanda her kuralın normalleştirilmiş ateşleme seviyesi hesaplanmaktadır.

$$O_{3i} = \overline{w_i} = \frac{w_i}{w_1 + w_2} \qquad i = 1,2 \tag{6}$$

4.Katman

Durulaştırma katmanı olarak adlandırılan bu katmanda her bir kural için çıkış değeri hesaplanmaktadır. Kuralların çıkışı, bir önceki katmandan gelen normalize edilmiş ateşleme seviyesi değerinin, birinci dereceden polinom ile çarpılmasıyla elde edilmektedir.

$$O_{4i} = \overline{w_i}.f_i = \overline{w_i}.(p_i x + q_i y + r_i) \quad i = 1,2 \tag{7}$$

5.Katman

Çıkış katmanıdır. Durulaştırma katmanında elde edilen her kurala ait çıkış değerleri toplanarak ANFIS'in çıkışı elde edilir.

$$O_{5i} = f = \sum \overline{w_i}.f_i = \frac{\sum w_i.f_i}{\sum w_i} \qquad i = 1,2 \tag{8}$$

Bulanık sistemlerin, öğrenme ve adaptasyon sürecini kolaylaştırması, sinirsel ağların ise kontrol parametrelerini ağ üzerinde dağıtarak doğrusal olmayan problemlerde başarılı olması, sinirsel bulanık bir ağ modeli olan ANFIS mimarisine büyük üstünlük kazandırmaktadır.

ANFIS modelinin eğitilmesindeki amaç giriş ve çıkış değerlerine bağlı olarak ağırlık değerleri için optimal değerlerin üretilmesidir. ANFIS'in ağırlık parametrelerinin eğitiminde türeve dayalı algoritmalar yaygın olarak kullanılmaktadır. Ancak türev tabanlı algoritmalarda eğim hesaplaması gibi zorluklar bulunmasının yanında yerel minimuma takılma gibi problemlere sebep olmaktadır. Bu sebepledir ki türev tabanlı algoritmalar ile ANFIS in eğitilmesi ve parametrelerin güncellenmesi temel sorunlardan biridir. Araştırmalar ANFIS'in parametrelerini eğitmek için son zamanlarda farklı algoritmalar önermiştir. Bu algoritmalardan bazıları ise türeve dayalı olmayan Genetik algoritma, PSO ve Diferansiyel Gelişim Algoritması gibi sezgisel algoritmalardır [24].

2.2.2. Genetik Algoritma (Genetic Algorithm)

1970'li yıllarda temel ilkeleri John Holland tarafından ortaya konulan Genetik Algoritma (GA) pek çok problem türünde başarı ile uygulanmaktadır [25]. GA, optimizasyon veya arama probleminde tam yada yaklaşık sonuçlar bulabilmek için kullanılan sezgisel bir algoritmadır. Bu algoritma kalıtım, mutasyon, seçim ve çaprazlama gibi evrimsel biyolojideki tekniklerden esinlenerek geliştirilmiştir. GA arama uzayı büyük olan ve ayrıca değişken sayısı çok fazla olan, çok boyutlu problemlere bile oldukça rahat uygulanabilmektedir. Problemlerin arama uzaylarında, uzayın tamamını aramak yerine daha iyi olabilecek değerleri deneme eğilimi sonucu makul sayılabilecek süreler içerisinde optimal sonuçlar üretebilme kabiliyetine sahiptir. GA'nın temel adımları Şekil 2'de verilmiştir.

Adım 1. Rastgele değerlerden başlangıç popülasyonunu üret
Adım 2. Kromozomların uygunluk değerlerini hesapla
Adım 3. Şartlar sağlanıncaya kadar devam et (Uygunluk değeri, işlem süresi vb)
- **(i)** Popülasyonda kötü uygunluk değerine sahip kromozomları belirle
- **(ii)** Çocuk kromozomlar için ebeveyn kromozomları tayin et
- **(iii)** Çaprazlama işlemi ile ebeveyn kromozomlardan çocuk kromozomlar üret
- **(iv)** Çocuk kromozomları mutasyona uğrat.
- **(v)** Popülasyona yeni dahil olan kromozomların uygunluk değerlerini hesapla

Şekil 2. Genetik Algoritmanın Temel Adımları (Main steps of the basic GA algorithm)

GA, popülasyon temelli bir optimizasyon algoritmasıdır. Popülasyonu oluşturan aday çözümlerin algoritmadaki karşılığı kromozomlardır. Bu kromozomlar çeşitli evrim işlemleri sayesinde daha iyi sonuçları temsil eden çözüm

adaylarına dönüşürler. Bu işlem kabul edilebilir bir uygunluk değerine ulaşıncaya kadar veya önceden belirlenen bir işlem süresi veya nesil sayısı gibi kriterler karşılanıncaya kadar sürdürülür.

GA'da kromozomlar (aday çözümler) temsil ettiği çözüme ait ayrık veya sürekli değerlere sahip değişkenleri tutan katarlardır. Uygunluk fonksiyonu ise kromozomların kalitesini ölçen amaç fonksiyonudur.

a) Başlangıç Popülasyonunun Oluşturulması (Creating Initial Populations)

Başlangıç popülasyonu rastgele aday çözümlerden oluşur. Popülasyonun büyüklüğü için belirli bir kriter yoktur. Ancak popülasyondaki birey sayısının çok fazla olmasının problemin çözüm zamanına veya elde edilecek çözümün iyiliğine etkisi olmadığı değerlendirilmektedir. Bu yüzden genellikle popülasyon 20-50 arasında alınır [26].

b) Uygunluk Değerinin Hesaplanması (Determination of Eligibility Value)

Uygunluk fonksiyonu yardımıyla hesaplanabilen bir değerdir. Uygunluk değeri, her bir kromozomun uyguluk fonksiyonuna tabi tutulması ile elde edilir. Uygunluk fonksiyonu değeri bireyin çözüm kalitesini gösterir. Bir sonraki nesle aktarılacak bireyler, uygunluk fonksiyonu değerlerine göre belirlenir.

c) Seçim (Selection)

Seçim stratejisi için sıklıkla Rulet Tekerlek yöntemi ve Turnuva yöntemi kullanılmaktadır.

Rulet Tekerleği Yöntemi

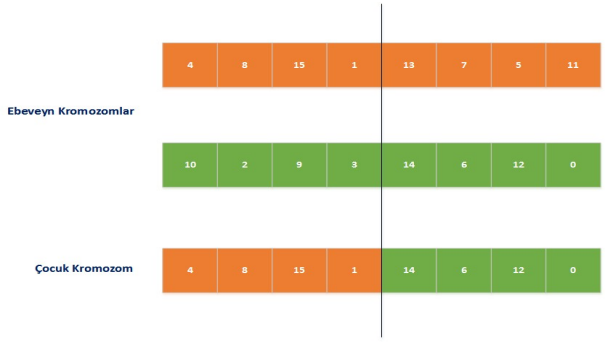
Bu yöntemin temel mantığı uygunluk değeri yüksek olan kromozomların ebeveyn kromozom olarak seçilebilme olasılığının, uygunluk değeri nispeten düşük olan kromozomlardan daha fazla olmasıdır [26].

Turnuva Yöntemi

Bu yöntemde popülasyondan rastgele seçilecek n adet kromozom içinden en iyi kromozomlar ebeveyn olarak seçilir [26].

d) Çaprazlama (Crossover)

Çaprazlama, genellikle iki ebeveyn kromozomdan alınan genlerle oluşturulur. Kromozomların başlangıç noktasından belirlenecek bir noktasına kadar olan genler ilk ebeveyn kromozomdan alınırken söz konusu noktadan sonraki genler ise ikinci ebeveynden alınarak yapılır. Farklı şekillerde çaprazlama yapılabilmektedir. Tek noktadan çaprazlama Şekil 3'de görülmektedir.



Şekil 3. Tek Noktadan Çaprazlama (One-Point Crossover)

e) Mutasyon (Mutation)

Mutasyon işlemi, popülasyondaki kromozomların çeşitliliğini arttırarak yeni çözüm adaylarının oluşmasını sağlar**.** Bir bireyin bir veya birkaç geninin değişerek farklı bir birey haline gelmesidir. Mutasyon popülasyonda çeşitliliğin oluşmasını sağlayan operatörlerden biridir. Mutasyon oranı genellikle %0,5 - %5 arasında bir sayıdır.

3. DENEYSEL VERİ SETİ (EXPERIMENTAL DATA SET)

Bu çalışmada veri kaynağı olarak Karaciğer kanseri mikrodizi gen ifade veri seti kullanılmıştır. Bu veri seti Rutgers Üniversitesinin Biyoinformatik Laboratuvarında bulunan veri tabanından elde edilmiştir [27].

Karaciğer Kanseri (Liver Cancer Chen-2002)

Chen ve arkadaşları [28] tarafından 2002 yılında çalışılan karaciğer kanser veri seti, hastalıklı ve normal kişilerin karaciğer dokularından alınmış 179 adet örnekten oluşmaktadır. Bu veri setindeki her örnek 85 mikrodizi gen dizilimi ile temsil edilmektedir. Veri seti HCC adında tümörlü doku ve normal doku olmak üzere 2 farklı sınıf ile temsil edilmektedir. Toplamda 179 örnek bulunan veri setinde 104 adet tümörlü doku bulunurken 75 adet de sağlıklı karaciğer dokusu bulunmaktadır.

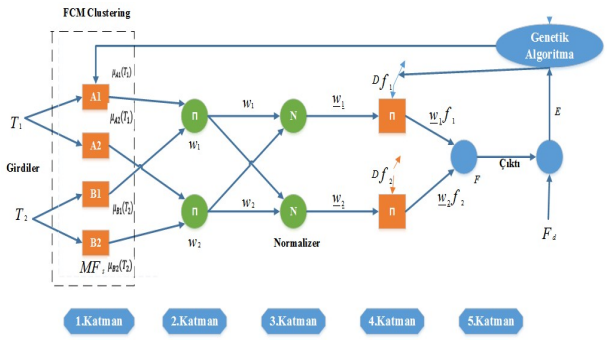
4. DENEYSEL ÇALIŞMA (EXPERIMENTAL STUDY)

Bu çalışmada karaciğer kanseri mikrodizi gen ifade verisi kullanılarak sınıflandırma yapılmıştır. Veri seti sınıflandırma işleminden önce [0-1] aralığına normalize

edilmiş daha sonra veri setlerine öznitelik seçme yöntemi uygulanmıştır. Bu çalışmada birçok öznitelik seçme yöntemi incelenmiş ve başarısından dolayı veri setlerinin bir alt setini oluşturmak için Korelasyon tabanlı öznitelik seçimi kullanılmıştır. Böylece yüksek boyutlu veri setleri yerine bu veri setlerini en iyi şekilde ifade eden alt veri setleri ile sınıflandırma yöntemlerinin performanslarını arttırmak ve sınıflandırmada daha başarılı sonuçlar almak mümkün hale gelmiştir.

Bu çalışmada, ANFIS modelinin GA kullanılarak eğitilmesine yönelik karaciğer kanseri mikrodizi gen ifade verisi kullanılarak uygulama geliştirilmiştir. Toplamda 179 örnek bulunan karaciğer verisini sınıflandırmada örneklerin az olması sebebiyle, modelin genelleştirebilme yeteneğini doğru bir şekilde değerlendirebilmek için *çapraz doğrulama (cross validation)* yöntemlerinden *K parçalı* yöntemi kullanılmıştır. *K parçalı çapraz geçerlilik* yönteminde amaç bir deneyi bağımsız koşullarda yineleyerek sonuçlarının geçerliliğini sınamadır. Örneğin, sınıflandırma problemlerinde, bir veri kümesini aşağı yukarı eşit k tane kümeye bölüp, her bir k için, sınıflandırıcıyı oluşturmak için, k-1 kümeyi kullanıp ve arta kalan k'ıncı küme üzerinde test işlemini yapmaktır. Bu çalışmada, 5-kat çapraz doğrulama yöntemi kullanılmıştır.

Kullanılan veri setlerindeki parametre sayılarının çokluğundan dolayı, bulanık kural ve üyelik fonksiyonu sayılarının az sayıda olduğu ANFIS modelleri oluşturabilmek için, ANFIS ile FCM kümelemenin entegre olduğu bir bulanık kural oluşturma tekniği kullanılmıştır. FCM uygulanarak oluşan ANFIS modellerindeki giriş sayısı her bir veri setindeki gen sayısına eşit olup her bir giriş için üyelik fonksiyonu türü *gauss,* üyelik fonksiyonu sayısı 10 ve bulanık kural sayıları 10 olarak belirlenmiştir. Yapılan simülasyon çalışmalarında kullanılan modelin oluşturulmasını ve oluşturulan modelin parametrelerinin optimize edilmesini gösteren blok diyagram Şekil 4'de verilmiştir. ANFIS'in güncellenmesi gereken iki parametre türü vardır. Bunlar başlangıç parametreleri ve sonuç parametreleridir. Başlangıç parametreleri Eşitlik (4)'de $\{a_i, c_i\}$ olarak verilen *gauss* üyelik fonksiyonlarına aittir. Buradaki a_i üyelik fonksiyonlarının varyansı, c_i üyelik fonksiyonlarının merkezidir. Başlangıç parametrelerinin toplam sayısı tüm üyelik fonksiyonlarındaki parametrelerinin toplamına eşittir. Sonuç parametreleri ise durulaştırma katmanında kullanılan Eşitlik (7)'de $\{p_i, q_i, r_i\}$ olarak gösterilen parametrelerdir. Tasarlanan ANFIS modelinin tüm parametrelerinin optimize edilmesi için GA algoritması kullanılmıştır.



Şekil 4. Önerilen yöntemle oluşturulan ANFIS'in temel yapısı (Basic Structure of ANFIS with FCM Clustering)

Bu çalışmada önerilen yöntem karaciğer kanseri problemine uygulanmıştır. Kanser verisinin doğru bir şekilde sınıflandırılabilmesi Şekil 4'de verilen ANFIS modelinin 1. ve 4. katmanındaki başlangıç ve sonuç parametrelerinin başlangıç değerleri için FCM kümeleme yöntemi ile bir çözüm uzayı oluşturulmuş ve sonrasında parametreler her bir iterasyonda GA kullanılarak güncellenmiştir. Eşitlik (9) ile tanımlanan *RMSE* hata fonksiyonu çözümün hata değerinin hesaplanmasında kullanılmıştır. Bu hata fonksiyonunda kullanılan F ve F_d parametreleri sırasıyla ANFIS tarafından elde edilen çıkış ve veri setinin gerçek çıkışıdır. GA ile ANFIS parametrelerinin eğitimi sürecinde, MATLAB ortamında oluşturulan özgün kodlar kullanılmıştır.

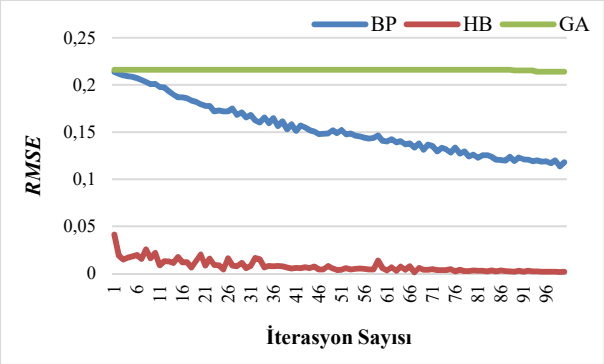
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(F(i) - F_d(i)\right)^2}{N}} \tag{9}$$

Yapay zeka optimizasyon algoritmalarının performansı algoritmaların kontrol parametrelerine fazlaca bağlıdır ve genellikle kontrol parametreleri için hangi değerlerin kullanılması gerektiğine yönelik kesin kural ve yöntemler bulunmamaktadır. Bu kapsamda, parametreler için araştırmacıların yaygın olarak önerdiği aralıklarda değerler tayin edilmesi veya çok sayıda deneme yapılarak optimal parametre değerlerinin belirlenmesi yaklaşımı kabul görmektedir. Çalışmada, GA algoritmasının parametre değerlerine karar vermek için birçok deneme gerçekleştirilmiş ve denemeler sonucunda GA için popülasyon sayısı 25, çaprazlama oranı 0,7 ve mutasyon oranı 0,15 olarak alınmıştır.

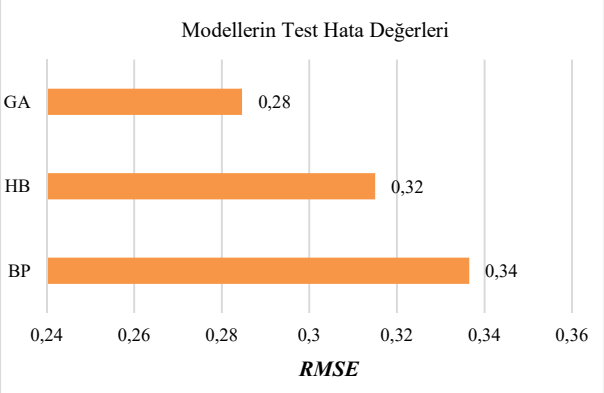
Önerilen yaklaşımın performansının farklı yöntemlerle karşılaştırılması amacıyla ANFIS ağı karaciğer mikrodizi gen ekspresyon kanser verisinin sınıflandırılmak için BP ve Hibrid algoritmalarla da eğitilmiştir. Bu amaçla, kullanılan BP algoritması için öğrenme katsayısı 0.2 ve momentum oranı 0.4 olarak seçilmiştir. Hibrid algoritma olarak da en küçük kareler tahmini ve BP algoritmasının birleştirilmesin oluşan bir yöntem kullanılmıştır. Ayrıca

bu iki algoritma için iterasyon sayısı 100 olarak belirlenmiştir.

Oluşturulan ANFIS modelleri BP, Hibrid ve GA algoritmaları ile eğitilerek sınıflandırmadaki performansları karşılaştırılmıştır. Her algoritma için ANFIS 15 kez eğitilmiştir. Ayrıca her model için elde edilen *RMSE* hata değerlerinin ortalaması alınarak, ortalama *RMSE* ($RMSE_{AVG}$) değeri bulunmuştur. Şekil 5'te yöntemlere ait 100 iterasyon için eğitim hata değerlerinin (*RMSE*) değişim grafiği verilmiştir. Şekil 6'da ise karaciğer verisini sınıflandırmak için farklı algoritmalar ile eğitilen ANFIS modellerine giriş olarak verilen test veri setinin gerçek çıkışı ile ANFIS tarafından elde edilen çıkış arasındaki farkı gösteren test hata değerleri (*RMSE*) verilmiştir.

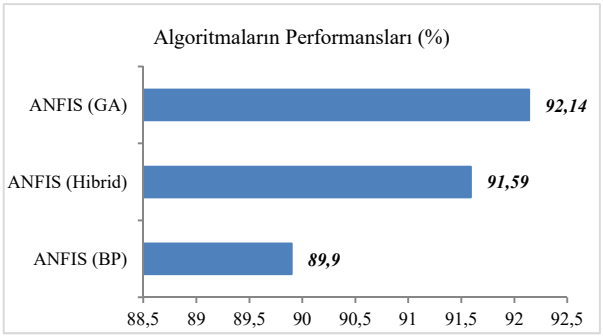


Şekil 5. ANFIS modellerine ait *RMSE* eğitim hata değerleri (Training *RMSE* error values of the ANFIS models)



Şekil 6. ANFIS modellerine ait *RMSE* test hata değerleri (Testing *RMSE* error values of the ANFIS models)

Şekil 5'te görüldüğü üzere en az eğitim hata değeri Hibrid algoritma ile bulunmuş olmasına rağmen, Şekil 6'da en az test hata değeri GA ile elde edilmiştir. Eğitim ve test hata değerleri arasındaki farkın çok olması aşırı öğrenme olduğunu ve eğitilen modelin uygulanabilir olmadığını ortaya koymaktadır. Eğitim ve test hata değerlerinin ise birbirine yakın olması daha güçlü ve başarımı yüksek modeller ortaya koymaktadır. Yapılan simülasyon çalışmaları neticesinde optimal parametre değerleri ile elde edilen sınıflandırma performanslarına ait bilgiler Şekil 7'de verilmiştir. Görüldüğü üzere en başarılı sınıflandırma performansı ANFIS-GA modeline aittir.



Şekil 7. Algoritmaların Karaciğer Kanserini Sınıflandırma Performansları (Classification performance of the algorithms for Liver Cancer)

Şekil 7' de görüldüğü gibi ANFIS - GA yönteminin ortalama % 92.14 oranıyla karaciğer kanser veri setini sınıflandırmada en yüksek performansı sergilemektedir. Ayrıca sınıflandırma yöntemlerinin ortalama başarımları ve birbirlerine göre gelişim yüzdeleri de aşamalı olarak Tablo 1'de verilmiştir. ANFIS ağının Hibrid algoritma ile eğitilmesi durumunda geleneksel BP algoritmasına göre sınıflandırma başarımının % 1.88 iyileştiği, GA ile eğitilmesi durumunda ise Hibrid algoritmasına göre % 0.60 daha başarılı olduğu görülmektedir. Gelişim yüzdeleri genetik algoritmanın sınıflandırma probleminde diğer yöntemlere göre daha başarılı olduğunu göstermektedir.

Tablo 1. Metotların ortalama başarımları ve gelişim yüzdeleri (The average performances and the percentages of development of the methods)

Metot	Başarım (%)	Gelişim Yüzdesi (%)
ANFIS – BP	89.90	-
ANFIS – Hibrid	91.59	1.88
ANFIS -GA	92.14	0,60

5. SONUÇLAR VE TARTIŞMA (CONCLUSIONS AND DISCUSSION)

Bu makalede, karaciğer mikrodizi gen ekspresyon kanser verisinin sınıflandırılması amacıyla ANFIS modelinin BP, Hibrid algoritmaları ve GA optimizasyon algoritmasıyla eğitilerek performansları karşılaştırılmıştır. Şekil 7'de verilen sonuçlardan, GA algoritması kullanılarak ANFIS'in eğitilmesine dayalı

yaklaşımın daha başarılı olduğu görülmüştür. Ayrıca, GA algoritmasının türeve dayalı algoritmalar gibi kısıtlamalar içermemesi ve problemlere kolaylıkla uygulanabilir olması nedeniyle, ANFIS ağının farklı problemlere yönelik uygulamalarında da kullanılabileceği değerlendirilmiştir.

TEŞEKKÜR (ACKNOWLEDGEMENT)

Bu çalışma Erciyes Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon birimi tarafından desteklenmiştir. (Proje Numarası: FDK-2016-6371) Ayrıca, bu çalışmanın sorumlu yazarını Yurt İçi Lisansüstü Burs Programı kapsamında yürütülen 2211-A Genel Yurt İçi Doktora Bursu ile destekleyen TÜBİTAK Bilim İnsanı Destekleme Daire Başkanlığı birimine teşekkür ederiz.

KAYNAKLAR (REFERENCES)

[1] G. S. Özcan, "Bütünleştirici Modül Ağlarıyla Gen Düzenleme Analizi," Başkent Üniversitesi, 2014.

[2] H. Ü. Lüleyap, *Moleküler Genetiğin Esasları*. İzmir: Nobel Kitabevi, 2008.

[3] H. S. BAL and F. Budak, "Mikroarray Teknolojisi," *Uludağ Üniversitesi Tıp Fakültesi Derg*, vol. 38, no. 3, pp. 227–233, 2012.

[4] Ö. Şimşek, "Mikroarray Teknolojisi ve Diş Hekimliğinde Kullanımı," *Atatürk Üniversitesi Diş Hekim Fakültesi Derg*, vol. 7, pp. 55–62, 2013.

[5] K. Shakya, H. J. Ruskin, G. Kerr, M. Crane, and J. Becker, "Comparison of Microarray Preprocessing Methods," Springer New York, 2010, pp. 139–147.

[6] K. Ipekdal, "Microarray Technology," 2011. [Online]. Available: http://yunus.hacettepe.edu.tr/~mergen/sunu/s_mikroarrayandecology.pdf. [Accessed: 05-Jul-2016].

[7] H. Liu, I. Bebu, and X. Li, "Microarray probes and probe sets.," *Front Biosci (Elite Ed)*, vol. 2, pp. 325–38, 2010.

[8] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest.," *BMC Bioinformatics*, vol. 7, p. 3, 2006.

[9] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data Comparison of Discrimination Methods for the Classi' cation of Tumors Using Gene Expression Data," *J Am Stat Assoc*, vol. 97457, no. July, pp. 77–87, 2002.

[10] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, Oct. 2000.

[11] Y. Lee and C.-K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data.," *Bioinformatics*, vol. 19, no. 9, pp. 1132–1139, 2003.

[12] H. Liu, J. Li, and L. Wong, "Classification and Study on Feature Gene Patterns Selection Expression and Profiles Methods Using Proteomic," *Genome Informatics*, vol. 13, pp. 51–60, 2002.

[13] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[14] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.," *Nat Med*, vol. 7, no. 6, pp. 673–9, 2001.

[15] M. Ringnér and C. Peterson, "Microarray-based cancer diagnosis with artificial neural networks," *Biotechniques*, vol. 34, no. 3 SUPPL., 2003.

[16] M. Pirooznia, J. Y. Yang, M. Q. M. Yang, and Y. Deng, "A comparative study of different machine learning methods on microarray gene expression data.," *BMC Genomics*, vol. 9 Suppl 1, p. S13, 2008.

[17] C. Loganathan and K. V Girija, "Cancer Classification using Adaptive Neuro Fuzzy Inference System with Runge Kutta Learning," *Int J Comput Appl*, vol. 79, no. 4, 2013.

[18] K. Anandakumar and M. Punithavalli, "Efficient Cancer Classification using Fast Adaptive Neuro-Fuzzy Inference System (FANFIS) based on Statistical Techniques," *IJACSA) Int J Adv Comput Sci Appl Spec Issue Artif Intell*, pp. 132–137, 2011.

[19] M. a. Hall and L. a. Smith, "Practical feature subset selection for machine learning," *Comput Sci*, vol. 98, pp. 181–191, 1998.

[20] J. S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System," *IEEE Trans Syst Man Cybern*, vol. 23, no. 3, pp. 665–685, 1993.

[21] D. Karaboga and E. Kaya, "Training ANFIS using artificial bee colony algorithm for nonlinear dynamic systems identification," in *2014 22nd Signal Processing and*

Communications Applications Conference (SIU), 2014, pp. 493–496.

[22] S. Uzundurukan, "Determining and modelling of principal parameters affecting swelling properties of soils," Suleyman Demirel , 2006.

[23] J. S. R. Jang and C. T. Sun, "Neuro-Fuzzy Modeling and Control," *Proc IEEE*, vol. 83, no. 3, pp. 378–406, 1995.

[24] D. Simon, "Training fuzzy systems with the extended Kalman filter," *Fuzzy Sets Syst*, vol. 132, no. 2, pp. 189–199, 2002.

[25] M. Mitchell, *An introduction to genetic algorithms*. MIT press, 1998.

[26] M. M. Kervan, "Çoklu Sensör Konumlandırma Probleminin Genetik Algoritmalar Ve Gen Havuzu Tabanlı Genetik Agoritmalar İle Çözülmesi," Hava Harp Okulu Komutanlığı, 2009.

[27] "Karaciğer Kanseri Mikroarray Gen İfade Profili Veri Seti (Chen-2002)." [Online]. Available: http://bioinformatics.rutgers.edu/Static/Supplements/CompCancer/CDNA/chen-2002/chen-2002_database.txt.

[28] X. Chen, S. T. Cheung, S. So, S. T. Fan, C. Barry, J. Higgins, K.-M. Lai, J. Ji, S. Dudoit, I. O. L. Ng, M. van de Rijn, D. Botstein, and P. O. Brown, "Gene Expression Patterns in Human Liver Cancers," *Mol Biol Cell*, vol. 13, no. 6, pp. 1929–1939, 2002.