Mezotelyoma Hastalığı Verilerinin Sınıflandırılmasında Genetik Algoritma ile Özellik Seçimi

Feature Selection with Genetic Algorithm in Classification of Mesothelioma Disease Data

Muammer ALBAYRAK[1], Ahmet ALBAYRAK[2]
[1]Biyoistatistik ve Tıp Bilişimi A.B.D., Karadeniz Teknik Üniversitesi, Trabzon, Türkiye
m.albayrak@ktu.edu.tr
[2]Bilgisayar Teknolojileri Bölümü, Karadeniz Teknik Üniversitesi, Trabzon, Türkiye
ahmetalbayrak@ktu.edu.tr

***Özetçe*—Bu çalışmada hasta verilerinden Mesothelioma hastalığının teşhisi için hazırlanan k-nearest neighbors sınıflayıcının performansını ve başarımını arttırmak amacıyla genetik algoritma tabanlı özellik seçimi yöntemi önerilmiştir. Sonuçlar k-nearest neighbors sınıflayıcının özellik seçimi uygulanmamış durumdaki performansı ile karşılaştırılmıştır. Genetik algoritma ile özellik seçimi sonucunda 34 özellik 7 özelliğe indirilmiş ve k-nearest neighbors sınıflayıcının performansı %96'dan %100'e çıkarılmıştır.**

Anahtar Kelimeler — Genetik Algoritma; k-nearest neighbors; özellik seçimi, optimizasyon

***Abstract*—This study proposes a genetic algorithm based feature selection method that improves the performance and effectiveness of k-nearest neighbors classifier which prepared to diagnose Mesothelioma disease from patient data. The results were compared with the performance of k-nearest neighbors classifier without feature selection. As a result of feature selection with genetic algortihm 7 of the 34 features were selected while increasing the performance of the k-nearest neighbors classifier from %96 to %100.**

Keywords — Genetic algorithms; k-nearest neighbors; feature selection, optimization.

I. GİRİŞ

Malignan mezotelyoma (MM) plevranın çok agresif tümörlerindendir. Bu tümörler asbest maruziyetine bağlı olmakla birlikte simian 40 virüsü enfeksiyonu ve genetik yatkınlıkla da ilişkilidir. Moleküler mekanizmalar ve kırsal yaşam da mezotelyoma gelişiminde etkili olabilir. Asbest içeren toprak karışımları Anadolu'da, Türkiye'de ve Yunanistan'da bulunmaktadır [9].

Genetik algoritma, (GA) optimizasyon yöntemi olarak kullanılan sezgisel bir algoritmadır. DNA kopyalanması ve doğal seçim mekanizmalarını taklit eder. Rastgele bireylerle başlatılır ve her bireyin performansı bir uygunluk fonksiyonu ile hesaplanır [1, 2]. Başlangıçtaki rastgele seçilmiş bireylerin en uygun olması ihtimali çok düşüktür. Bu nedenle en uygun bireylerden oluşan kararlı bir popülasyona ulaşmak için iteratif bir doğal seçim süreci kullanılır [3, 4]. Her iterasyonda seçim ve çoğalma işlemlerinin uygulanacağı yeni bir jenerasyon üretmek için belli bir yüzde oranında en iyi bireyler yetiştirilir. En iyi bireyler yeni jenerasyona aktarılırken en kötü bireyler ise jenerasyondan çıkartılır [5, 6]. Sonuçta bu evrimsel süreç verilen uygunluk fonksiyonuna göre en iyi adapte olmuş değişken alt kümesini seçer. Bu yöntem biyomedikal ve klinik veri setlerinde özellik seçimi amacıyla başarı ile uygulanmıştır [1, 7, 8].

Bu çalışmada Mesothelioma hastalığı ile ilgili hasta verilerinin sınıflandırılmasında k-NN sınıflayıcı kullanılmış ve başarımı %96 olarak belirlenmiştir. K-NN sınıflayıcının başarımını arttırmak amacıyla veri setindeki özellik sayısının azaltılması için GA tabanlı özellik seçimi önerilmiştir. Özellik sayısının fazla olması hasta verilerinin sınıflandırılmasını zorlaştırmakta, modeli karmaşık hale getirmekte ve sınıflayıcının başarımını olumsuz etkilemektedir. GA, en az sayıda seçilmiş özellik kullanılarak k-NN sınıflayıcının performansını ve başarımını en üst düzeye çıkaracak en uygun çözümü sağlayan bireyi aramak amacıyla kullanılmıştır.

II. MATERYAL VE YÖNTEM

A. Veri Seti

Bu çalışmada Dicle Üniversitesi Tıp Fakültesi'nde hazırlanarak UCI (University of California, Irvine) Machine

978-1-5090-2386-8/16/$31.00 ©2016 IEEE

Learning Repository veri tabanına yüklenmiş olan "Mesothelioma disease data set" isimli veri seti kullanılmıştır. Bu veri seti 324 hastaya ait kayıtları içermektedir. Her kayıt 34 özelliğe sahiptir. Bunlar; yaş (age), cinsiyet (gender), şehir (city), asbest maruziyeti (asbestos exposure), Malignan mezotelyoma tipi (type of MM), asbeste maruz kalma süresi (duration of asbestos exposure), tanı yöntemi (diagnosis method), taraf (keep side), sitoloji (cytology), semptomların süresi (duration of symptoms), solunum güçlüğü (dyspnea), göğüs ağrısı (ache on chest), halsizlik (weakness), sigara içme alışkanlığı (habit of cigarette), performans durumu (performance status), beyaz kan hücre sayısı (white blood cell count WBC), hemoglobin (HGB), trombosit sayımı (platelet count PLT), sedimantasyon (sedimentation), kan laktik dehidrogenaz (blood lactic dehydrogenise LDH), alkalin fosfataz (alkaline phosphatise ALP), toplam protein (total protein), albümin (albümin), glikoz (glucose), plevral laktik dehidrogenaz (pleural lactic dehydrogenise), plevral protein (pleural protein), plevral albümin (pleural albümin), plevral glikoz (pleural glucose), ölüm durumu (dead or not), plevral efüzyon (pleural effusion), tomografide plevral kalınlığı (pleural thickness on tomography), plevral asidite seviyesi (pleural level of acidity pH), C-reaktif protein (C-reactive protein CRP). Ayrıca her kayıt için tanı sınıfını ifade eden bir değişken de mevcuttur. 324 kayıttan 228 tanesi sağlıklı, 96 tanesi ise hasta olarak belirtilmiştir [9]. Özellik seçiminde k-nearest algoritması sınıflandırma amaçlı kullanılmaktadır.

B. k-nearest neighbors (KNN) Sınıflayıcı

k-NN algoritması sınıflandırma ve regresyonda kullanılan parametrik olmayan bir yöntemdir. k-NN sınıflamada çıktı, bir sınıf üyeliğidir. Bir nesne komşularının çoğunluğunun oyuyla bir sınıfa atanır [10, 11]. k-NN örnek tabanlı bir öğrenme algoritmasıdır ve tüm makine öğrenmesi algoritmaları arasında en basit olanlarından biridir [12, 13].

Bu çalışmada Mesothelioma hastalığı verilerini sınıflandırmak için k-nearest neighbor (k-NN) sınıflayıcı kullanılmıştır. Sınıflayıcının performansını ve başarısını artırmak için GA tabanlı özellik seçimi uygulanmıştır. Veri setinde bulunan 34 özellik içinden sınıflandırmaya etkisi bakımından en önemli olanların seçilmesi ve bu şekilde k-NN sınıflayıcının başarı düzeyinin en yükseğe ve giriş sayısının en aza çekilmesi amaçlanmıştır.k-NN sınıflayıcı fonksiyonu GA uygunluk fonksiyonu tarafından gönderilen veri setini giriş olarak almakta, sınıflayıcıyı eğittikten sonra cross-validation (çapraz doğrulama) işlemine tabi tutmakta ve son olarak sınıflayıcının hata oranını hesaplayarak döndürmektedir.

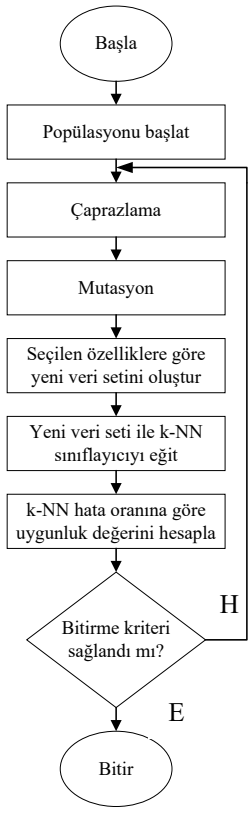
C. GA ile Özellik Seçimi

Makine öğrenmesi ve istatistikte özellik seçimi bir model inşa edilirken ilgili özelliklere ait bir alt küme seçilmesi işlemi olarak ifade edilir. Özellik seçimi ayrıca değişken seçimi olarak da adlandırılır. Özellik seçimi birçok farklı amaçla kullanılmaktadır. Bunlardan bazıları; 1- Araştırmacı ve kullanıcılar tarafından yorumlanmasını kolaylaştırmak amacıyla modellerin basitleştirilmesi, 2- Eğitim sürelerini kısaltmak ve 3- Aşırı öğrenmeyi azaltarak genelliği artırmak olarak sıralanabilir [2, 5].

Özellik seçimi tekniklerinin uygulanmasının temel dayanağı, veri seti içinde bulunan gereksiz yâda ilgisiz özelliklerin herhangi bir bilgi kaybına sebep olmadan kaldırılmasıdır [3]. Özellik seçimi yöntemleri özellik çıkarımı ile karıştırılmamalıdır. Özellik seçimi bir özellik alt kümesi verirken, özellik çıkarımı ise var olan özelliklerden yeni özellikler üretir [4, 6].

GA bir optimizasyon tekniği olarak özellik seçiminde sıklıkla kullanılmaktadır. Farklı özellik vektörlerini deneyerek en uygun olanı bulmak GA'nın çalışma prensibine son derece uygun bir problemdir [7, 8]. GA başlangıç popülasyonu özellik sayısı yüksek olduğundan 200 birey olarak seçilmiştir. Her birey 34 özelliği temsil eden 34 adet bitten oluşmaktadır. Bu nedenle bireyler ikili (binary, bit string) olarak kodlanmıştır. Bit değeri 0 ise ilgili özellik seçilmemiş, eğer 1 ise ilgili özellik seçilmiştir.

Uygunluk fonksiyonu ilgili bireyde seçilmiş olan özelliklere ait vektörlerden oluşan yeni veri setini KNN sınıflayıcıya göndermekte ve sınıflayıcıdan dönen hata oranını uygunluk değeri olarak almaktadır. GA'nın işlevi sınıflayıcıdan dönen hata değerini minimize etmek olarak ifade edilebilir. GA akış diyagramı Şekil 1'de verilmiştir.



Şekil 1. GA akış diyagramı

MATLAB optimizasyon araç kutusunda uygulama için girilen bazı önemli parametreler ve değerleri Tablo 1'de verilmiştir. Popülasyon tipi probleme uygun olarak 34 özelliği temsil edecek şekilde 34 bitten oluşan bit dizisi (bit string) olarak seçilmiştir. Popülasyon büyüklüğü özellik sayısının yüksek olması nedeniyle 200 olarak seçilmiştir. Seçim fonksiyonu MATLAB tarafından varsayılan olarak verilen Stochastic Uniform ve mutasyon oranı da yine varsayılan değer olan 0.01 olarak seçilmiştir. Çaprazlama yöntemi bit dizisi popülasyonu için uygun olan tek noktalı (single point) seçilirken elit sayısı ve uygunluk ölçekleme yöntemleri sırasıyla 2 ve sıra ölçekleme seçilmiştir.

Parametre	**Değer**
Popülasyon Tipi	Bit String
Popülasyon Büyüklüğü	200
Seçim Fonksiyonu	Stochastic Uniform
Mutasyon Oranı	0.01
Çaprazlama Şekli	Tek Noktalı
Elit sayısı	2
Uygunluk ölçekleme	Rank

Tablo 1. Uygulama parametreleri

III. SONUÇLAR VE TARTIŞMA

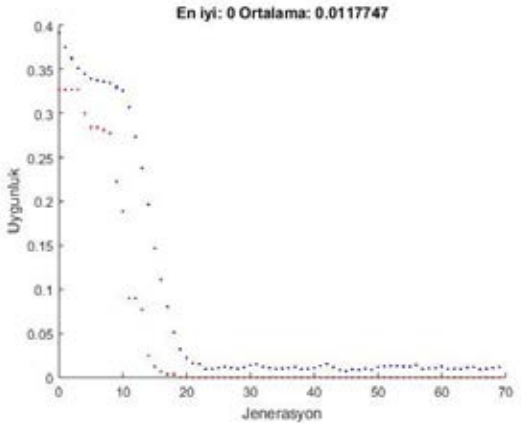
Verilen parametreler ile GA, yaklaşık 35 dakikalık işlem ve 69 iterasyon sonucu en iyi uygunluk değerine sahip bireyi belirlemiştir. En iyi birey 34 özellikten sadece yedi tanesinin seçilmiş olduğu durumda 0.0 hata değerine sahip olarak elde edilmiştir. Bu değer k-NN sınıflayıcının başarısının %100 olarak gerçekleştiğini göstermektedir. K-NN sınıflayıcı GA ile özellik seçimi yapılmadan yani 34 özelliğin tamamını dikkate alarak çalıştırıldığında %96 başarı ile çalışmaktadır.

k-NN sınıflayıcının sonuca ulaşma süresi bakımından ise kayda değer bir farklılık gözlenmemiştir. Özellik ve veri sayısının daha fazla olduğu veri setlerinde bu farklılığın daha net bir şekilde gözlenmesi muhtemeldir.

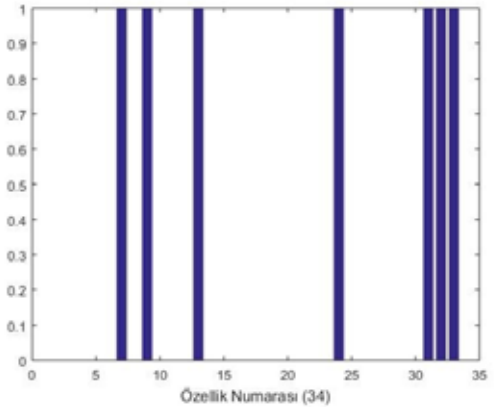
Çalışmanın tamamlanma süresi kullanılan bilgisayarın performans özellikleri ve seçilen GA parametrelerine bağlı olarak değişebilir. Bu çalışma Windows 10 64 bit işletim sistemi koşturan, 3.20 Ghz dört çekirdek işlemci ve 8Gb ram özelliklerine sahip bir bilgisayar kullanılarak gerçekleştirilmiştir.

Seçilen özellikler; asbeste maruz kalma durumu, tanı yöntemi, halsizlik, solunum güçlüğü, albümin, plevral efüzyon ve tomografide plevral kalınlık olarak gerçekleşmiştir.

Şekil 2'de elde edilen en iyi ve ortalama uygunluk değerlerini gösteren grafik verilmiştir. Burada başlangıçta 0.4 olan hata değerinin 69 iterasyon sonucunda 0.0'a düştüğü görülmektedir. Şekil 3'te ise seçilen özellikler görülmektedir.

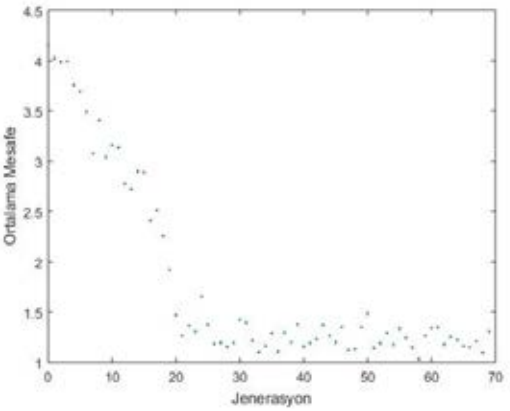


Şekil 2. En iyi ve ortalama uygunluk

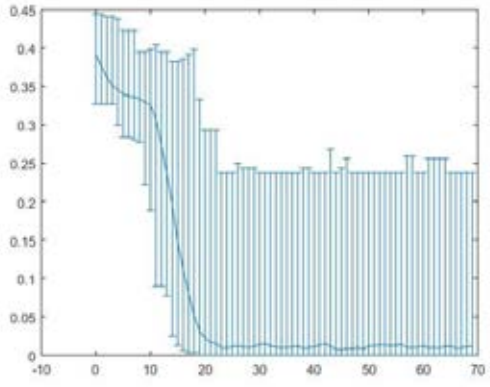


Şekil 3. Seçilen özellikler

Şekil 4'te ise bireyler arasındaki ortalama mesafe 4 birimden 1 birime indiği görülmektedir. Şekil 5'te en iyi, en kötü ve ortalama skorlardaki iyileşme açıkça görülebilmektedir.

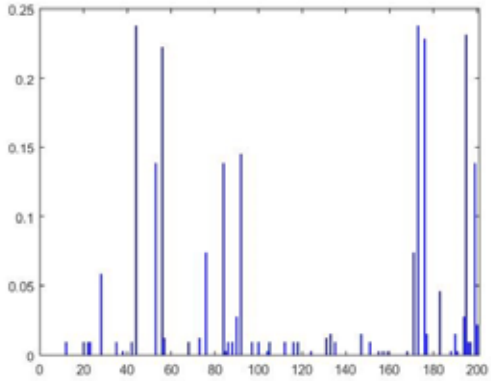


Şekil 4. Bireyler arası ortalama mesafe

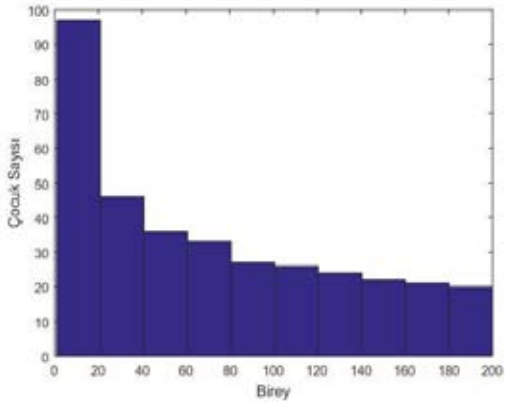


Şekil 5. En iyi, en kötü ve ortalama skorlar

Şekil 6'da 200 bireyden oluşan popülasyonun her bir bireyinin uygunluk değerleri, Şekil 7'de ise seçim fonksiyonunun bireylerden yaptığı seçimler görülmektedir.



Şekil 6. Her bir bireyin uygunluğu



Şekil 7. Seçim fonksiyonu

Bu çalışma tekrarlandığında birbirine yakın olmakla birlikte farklı sonuçlar üretebilmektedir. Bu durum GA'nın rastgele çalışma mantığının bir sonucudur. K-NN sınıflayıcının eğitiminde daha büyük bir veri seti kullanıldığı takdirde daha kararlı sonuçlar elde edilebilir.

Ayrıca çalışmaya k-NN sınıflayıcının dışında sınıflama algoritmaları da dâhil edilerek sınıflama performanslarının karşılaştırılması sınıflama algoritmalarının problem üzerindeki etkinliğinin analizi için karşılaştırmalı sonuçlar elde edilebilmesi açısından faydalı olacaktır.

KAYNAKÇA

[1] Ghaheri, A., Shoar, S., Naderan, M., Hoseini, S. S., "The Applications of Genetic Algorithms in Medicine"', *Oman Medical Journal, Vol. 30, No. 6: 406–416, 2015.*

[2] Welikala, R. A., Fraz, M. M., Dehmeshki, J., Hoppe, A., Tah, V., Mann, S., Williamson, T. H., Barman, S. A., "Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy", *Computerized Medical Imaging and Graphics 43: 64–77, 2015.*

[3] Khan, A. and Baig, A. R., "Multi-Objective Feature Subset Selection using Non-dominated Sorting Genetic Algorithm", *Journal of Applied Research and Technology, Vol.13, 2015.*

[4] Xu, L., Redman, C. W. G., Payne, S. J., Georgieva, A., "Feature selection using genetic algorithms for fetal heart rate analysis", *Physiological Measurement 35: 1357, 2014.*

[5] Singh, D. A. A. G., Leavline, E. J., Priyanka, R., Priya, P. P., "Dimensionality Reduction using Genetic Algorithm for Improving Accuracy in Medical Diagnosis", *I.J. Intelligent Systems and Applications, 1, 67-73, 2016.*

[6] Erguzel, T. T., Ozekes, S., Tan, O., Gultekin, S., "Feature Selection and Classification of Electroencephalographic Signals: An Artificial Neural Network and Genetic Algorithm Based Approach", *Clinical EEG and Neuroscience, Vol. 46(4) 321–326, 2015.*

[7] Cerrada, M., Sánchez, R. V., Cabrera, D., Zurita, G., Li, C., "Multi-Stage Feature Selection by Using Genetic Algorithms for Fault Diagnosis in Gearboxes Based on Vibration Signal", *Sensors, 15, 23903-23926, 2015.*

[8] Hsu, W., "Improving Classification Accuracy of Motor Imagery EEG Using Genetic Feature Selection", *Clinical EEG and Neuroscience, Vol. 45(3) 163–168, 2014.*

[9] Er, O., Tanrikulu, A. C., Abakay, A., "An approach based on probabilistic neural network for diagnosis of Mesothelioma's disease", *Computers & Electrıcal Engıneerıng Volume: 38 Issue: 1 Pages: 75-81, 2012.*

[10] Hilda, T., Rajalaxmi, R. R., "Effective Feature Selection for Supervised Learning Using Genetic Algorithm", *IEEE 2'nd Internatıonal Conference on Electronıcs and Communıcatıon Systems, 2015.*

[11] Baur, B., Bozdag, S., "A Feature Selection Algorithm to Compute Gene Centric Methylation from Probe Level Methylation Data", *PLoS ONE 11(2): e0148977, 2016.*

[12] Kumar, M., Kumar Rath, N., Kumar Rath, S., "Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier", *Journal of Biomedical Informatics 60: 395–409, 2016.*

[13] Liang, S., Ning, Y., Li, H., Wang, L., Mei, Z., Ma, Y., Zhao, G., "Feature Selection and Predictors of Falls with Foot Force Sensors Using KNN-Based Algorithms", *Sensors, 15, 29393-29407, 2015.*