Istanbul Commerce University Journal of Science
İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 15(30), Güz 2016
http://dergipark.gov.tr/ticaretfbd

Araştırma Makalesi

İKİ DÜZEYLİ OLASILIK MODELLERİNDE KLASİK VE META SEZGİSEL OPTİMİZASYON TEKNİKLERİNİN PERFORMANSI ÜZERİNE BİR ÇALIŞMA*

Özge AKKUŞ[1*] **Emre DEMİR[2]**

[1]Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla Türkiye
ozge.akkus@mu.edu.tr
[2]Hitit Üniversitesi, Tıp Fak., Biyoistatistik Böl., Çorum, Türkiye
emredemir@hitit.edu.tr

Öz

Bağımlı değişkenin kategorik olduğu durumda, model parametrelerinin tahmininde kullanılan geleneksel yöntem, En Çok Olabilirlik Tahmin Edicisi (EÇOTE)'dir. Bu yöntemde olabilirlik eşitliklerinin çözümünde, klasik Newton-Raphson (NR) algoritması kullanılmaktadır. Ancak bu algoritma olabilirlik fonksiyonunun diferansiyellenebilir özellikte olduğu durum için uygundur. Bu çalışmada, iki düzeyli bağımlı değişken modellerinde klasik optimizasyon yöntemlerinin uygulanabilmesi için gerekli olan varsayımların sağlandığı durumda optimal parametre tahminlerine ulaşabilmek için NR algoritmasına alternatif olan Genetik Algoritma (GA) yaklaşımının etkinliği araştırılmıştır. Bu amaçla, ilk olarak Alopesia hastalığı verisi kullanılmıştır. Gerçek veri uygulamasına ek olarak yapay bir veri kümesi üzerinden elde edilen sonuçlar da sunulmuştur. Son olarak, yöntemlerin Matlab program kodları ve açıklamaları ayrıntılı bir biçimde verilmiştir.

Anahtar Kelimeler: Genetik algoritma, lojistik regresyon, optimizasyon, parametre tahmini, Newton-Raphson.

Research Article

A STUDY ON THE PERFORMANCE OF THE CLASSICAL AND META HEURISTIC OPTIMIZATION TECHNIQUES IN PROBABILITY MODELS WITH TWO LEVELS

Abstract

The traditional method in the parameter estimation when we study with a categorical dependent variable is the Maximum Likelihood Estimator (MLE). In this method, the classical Newton-Raphson (NR) algorithm is used in the solution of the obtained likelihood equations. However, this algorithm is suitable when the likelihood function is defferentiable. In this study, the efficiency of the Genetic Algorithm approach (GA), alternative to the NR algorithm, is investigated for obtaining the optimal parameter values when the required assumptions for the classical optimization techniques are satisfied in the binary dependent variable models. For this purpose, first, data related to the Alepecia disease is used. In addition to the real data application, the simulated data results are also presented. Finally, the Matlab commands and their explanations are given in detail.

Keywords: Genetic algorithm, logistic regression, optimization, parameter estimation, Newton-Raphson.

*Received / Geliş tarihi: 12/10/2016 Accepted / Kabul tarihi: 16/12/2016
*Corresponding Author/ Sorumlu Yazar: ozge.akkus@mu.edu.tr

1. GİRİŞ

Doğrusal olmayan bir fonksiyonun optimum değerlerinin bulunabilmesi için iteratif olarak çözüme ulaşan yöntemlerin başında Newton Raphson, Regula-False, Secant Metodu, Bisection (yarılama) metodu, En Dik İnme yöntemleri gelmektedir. Bütün bu sayısal analiz yöntemlerinde az sayıda iterasyon ve en az hatayla doğru sonuca ulaşmak temel hedeftir. Bu yöntemler, bazı karmaşık problemlerin çözümlerinin içerdikleri işlemsel yük, tekrarlamalı hesaplamalar ve uzun zaman ihtiyacı nedeni ile çoğu zaman bilgisayar kullanılmasını gerektirirler. Literatürde yaygın olarak kullanılan türev kullanmayan sezgisel optimizasyon tekniklerinden bazıları ise; Yapay Sinir Ağları, Genetik Algoritma, Parçacık Sürü Optimizasyonu, Kuş Sürüsü Optimizasyonu ve Nelder-Mead Algoritmasıdır.

Genetik Algoritma (GA), evrimsel süreçleri kullanarak probleme ait en iyi çözümü bulan bir yapay zeka tekniğidir. Charles Darwin'in evrim teorisinden etkilenilerek geliştirilen GA, evrim alanlarındaki araştırmalara öncülük eden Rechenberg (1973) tarafından, "evrim stratejileri" adlı çalışmasında tanıtılmış, John Holland (1975) tarafından geliştirilmiş ve 1989 yılında Goldberg'in yayınladığı "Genetic Algorithms in Search, Optimization and Machine Learning" kitabı ile popülerlik kazanmış bir meta-sezgisel (meta-heuristic) algoritmadır. Yöntemin mühendislik, İstatistik, İşletme, Ekonomi, ve Kimya alanlarında yapılan uygulamalarında detaylıca incelendiği görülmektedir (Karr ve Freeman, 1999).

Bu alanda yapılan çalışmalar incelendiğinde, bağımlı değişkenin sürekli olduğu regresyon modellerinde parametre tahmininde GA'nın başarısının tartışıldığı görülmektedir (Altunkaynak ve Esin, 2004). Bu çalışmada ise, kategorik bir bağımlı değişkenin yer aldığı modellerde GA'nın parametre tahminindeki başarısı, klasik NR algoritmasına göre üstün ya da eksik yönleri ve GA'nın Matlab'da uygulama adımları ve açıklamalarına ayrıntılı bir biçimde yer verilmiştir.

2. LİTERATÜR ÖZETİ

Genetik algoritma ile ilgili yapılmış olan güncel çalışmalar kronolojik olarak aşağıda listelenmiştir.

Koh vd. (2008), yaptıkları medikal araştırmada ters ilaç reaksiyonu değerlendirme sistemi için bir olasılık skorlama sisteminin yapılandırılmasında genetik algoritma yaklaşımını kullanmışlardır. Yeni kurulan sistemin %83.8 oranında başarılı sinyaller verdiği gözlenmiştir.

Liu ve Ong (2008), pazarlama segmentasyonu için yapılması gerekli olan kümeleme analizinin başarılı sonuçlar verebilmesi için en etkin değişkenleri ve küme sayılarını belirlemek amacıyla genetik algoritma yaklaşımını kullanmışlardır. Sonuçlar incelendiğinde bulunan küme sayısı dikkate alındığında ve değişkenler belirlendiğinde genetik algoritmanın global optimum noktaları bulmada son derece başarılı olduğu görülmüştür.

Hadi ve Gonzalez-Andujar (2009), termal zamanın bir fonksiyonu olarak bitkilerin fidelenme zamanını doğrusal olmayan regresyon ve genetik algoritma ile tahmin etmeye çalışmışlardır. Varsayımlarının esnekliği ve diğer avantajları dikkate alındığında genetik algoritmanın daha etkili bir yöntem olduğu sonucuna varmışlardır.

Babaoğlu vd. (2010), parçacık sürü optimizasyonu ve genetik algoritmanın etkinliklerini karşılaştırdıkları çalışmalarında koroner arter hastalığı ile ilgili bir veri kümesi kullanmışlardır.

Meng ve Weng (2011) tarafından yapılan çalışmada, çalışma bölgesinin riskini değerlendirmek amacıyla genetik algoritma yaklaşımı ve lojistik regresyon analizi kullanılmıştır. Sonuçlar, tahminin doğruluğu bakımından genetik algoritma yaklaşımının iki düzeyli lojistik regresyona göre daha iyi performans gösterdiğini ortaya koymuştur.

Johnson vd. (2013) tarafından yapılan çalışmada Alzheimer hastalığının tanısında hastalığın süresinin tahmininde lojistik regresyon ve genetik algoritma yaklaşımlarını kullanmışlardır.

Gordini (2014), İtalya'da kobilerin iflas etme ihtimallerini tahmin etmeye çalıştıkları araştırmalarında genetik algoritma, lojistik regresyon ve destek vektör makineleri ile elde edilen sonuçları karşılaştırmıştır. Genetik algoritma yaklaşımının önemli ölçüde başarılı sonuçlar verdiği gözlenmiştir.

Lee ve Kim (2015) yaptıkları çalışmalarında ters yüzey radyasyon problemleri için parçacık sürü optimizasyonu ve genetik algoritma yaklaşımlarının performanslarını karşılaştırmışlardır.

Stylianou vd. (2015), yanık nedeniyle ölüm oranını araştırdıkları çalışmalarında geleneksel olarak kullanılan lojistik regresyon modeline alternatif olarak yapay sinir ağları, genetik algoritma, destek vektör makineleri gibi diğer birçok sezgisel optimizasyon yöntemlerinin etkinliğini tartışmışlardır.

Yuan ve Lee (2015), 2001-2010 periyodunda 88 Tayvan şirketinin finansal göstergelerini dikkate alan bir risk olasılığı hesaplamak amacıyla farklı yöntemlerin etkinliklerini karşılaştırmışlardır.

Pfeifer vd. (2015) tarafından proje yönetimi ile ilgili olarak yapılan çalışmada projenin tamamlanmasındaki gecikmeler üzerinde durulmuştur. Proje gecikmelerinin taşıyacağı maksimum riski tahmin etmek amacıyla genetik algoritma yaklaşımından yararlanılmıştır.

Hadji vd. (2015) tarafından yapılan çalışmada, genetik algoritmaya dayalı bir mühendislik uygulaması ile PV (photovoltaic) sistemleri üzerinden Maksimum Güç Noktası İzleme yönteminin teorik ve deneysel analizi sunulmuştur.

Aguilar-Rivera vd. (2015) tarafından yapılan çalışmada finansal problemlerin çözümü için genetik algoritma, genetik programlama, çok amaçlı evrimsel algoritma gibi birçok evrimsel hesaplama yöntemlerinin kullanımı tanıtılmıştır. Diğer

yöntemlere ilgi zamana bağlı olarak değişirken, genetik algoritmanın her dilimde en popüler yöntem olduğu sonucuna ulaşılmıştır.

3. GENETİK ALGORİTMA

GA'da evrim süreci potansiyel çözüm uzayı boyunca kromozom popülasyonlarının GA'nın operatörlerinin uygulanmasıyla gerçekleştirilir. GA'nın kullandığı operatörler, Kopyalama, Çaprazlama ve Mutasyon operatörleridir. Bu operatörler yardımıyla GA, araştırılan problemin çözümü için çözüm uzayının taranması ve en iyi çözümün bulunmasını amaçlar. Bir popülasyondan çözümler alınır. Bu çözümler daha sonra yeni bir popülasyon oluşturmak için kullanılır. Bu işlem yeni popülasyonun eski popülasyondan daha iyi olacağı varsayımından hareketle yapılır. Yeni çözümleri (nesilleri) oluşturmak için seçilen çözümler uygunluk değerlerine göre seçilir (Goldberg, 1989).

Popülasyon için hesaplanan uygunluk değerlerine göre kromozomlar iyi veya kötü olarak değerlendirilebilir. İyi olarak değerlendirilen bireylerin yaşamlarını devam ettirme olasılıkları ve kötü olarak değerlendirilen bireylerin ölme olasılıkları artırılarak iyi bireylerin çoğalmaları ve kötü bireylerin yok olmaları istenir. Evrim sürecinde GA'nın evrim operatörleri yardımıyla kromozomlar farklı bir popülasyon oluşturur ve bir iterasyon tamamlanır. Belirli iterasyon sayısına ulaşıldığında veya durdurma koşulu sağlandığında son popülasyondaki en iyi uygunluk değerine sahip kromozomun kodladığı çözüm, problemin çözümü olarak belirlenir (Mitchell, 1999).

GA'lar diğer optimizasyon yöntemlerinden farklı olarak, arama işlemini tek bir aday çözüm ile gerçekleştirmek yerine birden fazla aday çözümün oluşturduğu bir topluluk ile gerçekleştirir. Bundan dolayı çözüm uzayının birden çok başlangıç noktasıyla paralel olarak taranması sağlanmış olur. Bu özellik, GA'ların yerel optimum değerlerine takılmadan, global optimum değerlerini bulabilmesinde en büyük etken olmaktadır (Mitchell, 1999). Goldberg'e göre GA'ların geleneksel arama metotlarından farkı aşağıdaki şekilde verilebilir.

GA'lar parametrelerin kendisi yerine parametreler kümesinin kodlanması ile çalışır.
GA'lar arama aracı olarak tek bir noktayı değil, noktalar kümesini kullanır. Böylece yerel en iyiye takılma riskleri daha düşüktür.
GA'lar arama sırasında probleme ilişkin türevlenebilme koşulu veya diğer ek bilgilere ihtiyaç duymaz, genetik algoritmalar için amaç fonksiyonunun bilinmesi yeterlidir.
GA'lar belirli geçiş kuralları ile değil stokastik geçiş kuralları ile çalışır (Goldberg, 1989).
GA'da bir problemin optimum çözümü bazı temel işlemler ile gerçekleştirilir. Bunlar; amaç fonksiyonunun belirlenmesi, kodlama, başlangıç popülasyonunun oluşturulması, üreme, çaprazlama ve mutasyon işlemleridir.

Kodlama: Çözümlerin veya popülasyondaki bireylerin nasıl temsil edileceğine karar verilmesidir. Kodlama genellikle kromozomlar şeklinde gerçekleşir. Her kromozom

içindeki özel yapılar gen olarak adlandırılır. İkili kodlamada, her kromozom bit (0 veya 1) karakter dizilerinden oluşmaktadır (Pasia vd. 2005).

Başlangıç Popülasyonu: Bu adımda bir başlangıç seti olarak n kromozom rastgele seçilir. Bazı araştırmalar ideal popülasyon büyüklüğünün 20-30 civarı olması gerektiğini belirtirken bazı araştırmalarda 50-100 civarı popülasyon büyüklüğünün en ideal olduğunu söylemektedir (Reeves ve Rowe, 2002; Pasia vd. 2005).

Uygunluk Değeri: GA'nın her iterasyonunda popülasyonu oluşturan her bir bireyin uygunluk değeri hesaplanarak belleğe kaydedilir. Uygunluk değeri diğer bireylere göre daha iyi olan bireyin bir sonraki nesle aktarılma olasılığı fazladır.

Seçme Operatörü: Seçim işlemi, yeni nesillerde daha yüksek uygunluk değerlerine sahip bireylerin oluşması için, eski popülasyondaki bir kromozomun, uygunluk değerine bağlı olarak bazı yöntemlerle yeni oluşturulacak bir popülasyon içine seçme işlemidir. Seçim işlemi için Turnuva seçimi, Rulet tekerleği seçimi, Sıralama seçimi, Sabit durum seçimi ve Stokastik seçim yöntemleri kullanılabilir. Goldberg ve Deb (1991) herhangi bir seçim yönteminin diğerine karşı belirgin bir üstünlüğünün olmadığını belirtmektedir.

Çaprazlama Operatörü: Bu operatörün temel fonksiyonu, yeni nesillerin bir öncekinden farklı nesiller üreterek çeşitliliği arttırmak ve böylece daha geniş bir çözüm uzayında çalışarak arzu edilen sonuca ulaşma olasılığını arttırmaktır. Çaprazlama, yeni kromozomların eski kromozomların iyi genlerini alıp daha iyi olacakları düşüncesiyle yapılır. Çaprazlama operatöründen kromozomların iyi özelliklerini birleştirerek daha iyi kromozomlar oluşturması beklenir.

Mutasyon Operatörü: Mutasyon GA ile yapılan aramanın yerel optimumlarda takılmasının önüne geçer. Araştırmacı tarafından belirlenen mutasyon olasılığı parametresi bunu etkileyen parametredir. Mutasyon genellikle çaprazlamaya göre daha az olasılıkta kullanılır. Bunun nedeni çaprazlama sonucu elde edilen uyum değeri yüksek dizileri kaybetmemektir (Holland, 1992; Goldberg, 1989; Reeves ve Rowe, 2002).

4. İSTATİSTİKTE OPTİMİZASYON VE PARAMETRE TAHMİNİ

Doğrusal veya doğrusal olmayan bir fonksiyonu en büyük veya en küçük yapan noktaları bulmak için fonksiyonun optimize edilmesi gerekmektedir. Örneğin, Lojistik Regresyon (LR) Analizinin parametreleri olabilirlik fonksiyonunu en büyük yapan köklerdir. Dolayısıyla istatistikte optimizasyon önemli bir yere sahiptir. Literatürde bağımlı değişkenin sürekli olduğu doğrusal regresyon analizinde sezgisel yöntemlerden GA uygulamalarının yeterince tartışıldığı gözlenmektedir. Bağımlı değişkenin kategorik olduğu ve dolayısıyla açıklayıcı değişkenler ile doğrusal olmayan bir yapı ile bağlı olması durumunda GA'nın etkinliği bu çalışmanın temelini oluşturmaktadır. Özel olarak çalışmada bağımlı değişkenin iki düzeyli kategorik bir değişken olması durumu ele alınmakta ve LR model parametrelerinin tahmini üzerine yoğunlaşılmaktadır.

4.1. Lojistik Regresyon Analizi

İki düzeyli LR modellerinde kategorik bağımlı değişken (Y), bir olgunun gerçekleşmesi (Y=1) ve gerçekleşmemesi (Y=0) gibi iki durumludur ve Bernoulli dağılımı göstermektedir. Bağımsız değişken değerlerine bağlı olarak (x), i. gözlemin bağımlı değişkenin "1" olarak kodlanan düzeyine ait olması olasılığı için matematiksel eşitlik ($\pi(x_i)$) aşağıdaki biçimde tanımlanabilir.

$$\pi(x_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_p X_P)}} = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_P}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_P}} \quad (1)$$

Burada,

β_0: sabit terim; $\beta_1, \beta_2, \ldots, \beta_p$: regresyon katsayıları; $X_1, X_2 \ldots, X_p$: bağımsız değişkenler; p: bağımsız değişken sayısı; e = 2.718 doğal logaritma tabanı ve $\pi(x_i)$: i'inci gözlemin bağımlı değişkende "1" olarak kodlanan düzeye ait olma olasılığını, 1- $\pi(x_i)$ ise i'inci gözlemin bağımlı değişkende "0" olarak kodlanan düzeye ait olması olasılığını göstermektedir.

$$Z_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} = X_i'\beta \quad (2)$$

olarak tanımlanan açıklayıcı değişkenlerin doğrusal kombinasyonları Eş.(1)'de yerine yazıldığında, $\pi(x_i)$ olasılığı,

$$\pi(x_i) = \frac{1}{1+e^{-Z_i}} = \frac{e^{Z_i}}{1+e^{Z_i}} \quad (3)$$

biçimine dönüşür. Burada,

Eş.(3) bazı dönüşümler yapılarak, Eş.(4) ile verilen biçimde açıklayıcı değişkenlerin doğrusal kombinasyonları olarak ifade edilebilir.

$$\ln\frac{\pi(x_i)}{1-\pi(x_i)} = \ln e^{Z_i} = Z_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} = X_i'\beta \quad (4)$$

Eşitliğin sol tarafı "lojit" olarak adlandırılır. Aynı ifadenin "Lojistik Regresyon" olarak ifadesi Eş.(5) ile verilmektedir.

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_p X_{pi}}}{1+e^{\beta_0+\beta_1 X_{1i}+\cdots+\beta_p X_{pi}}} \quad (5)$$

biçiminde ifade edildiğinde ise bu modele "LR Modeli" adı verilir.

y_i, bir özelliğin olup olmamasına göre 0 ve 1 değerini alabilen bağımlı değişkeni (i=1,2,3,…,n) ve x_i, i.'nci gözlem için bağımsız değişkenin değerini göstermek üzere Eş.(1)'deki $\pi(x_i)$ ifadesi, verilen x değerleri için y'nin 1'e eşit olması koşullu olasılığını vermektedir. LR modeli iki seçenekli bağımlı değişkenin $\pi(x_i)$ olasılığını i. gözlenen değer için Eş.(1)'deki gibi varsayar. Burada (x_i, y_i) çifti için $y_i = 1$ olduğunda olabilirlik fonksiyonuna katkısı $\pi(x_i)$ iken $y_i = 0$ olduğunda olabilirlik fonksiyonuna katkısı $1 - \pi(x_i)$ kadardır. Buna göre (x_i, y_i) çiftinin olabilirlik

fonksiyonuna katkısı Eş.(6)'daki gibidir.

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad (6)$$

Gözlemlerin birbirlerinden bağımsız olduğu varsayıldığında, Eş.(6)'daki terimlerin çarpılmasıyla elde edilen olabilirlik fonksiyonu Eş.(7) ile verilmiştir.

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\,[1-\pi(x_i)]^{1-y_i} \quad (7)$$

En Çok Olabilirlik Tahmin Edicisi (EÇOTE), bu eşitliği maksimum yapan parametreleri tahmin etme temeline dayanır. Kolaylık olması bakımından, bu ifadenin logaritması alındığında çarpım durumu toplam durumuna getirilerek Eş.(8) ile verilen ifade elde edilir.

$$L(\beta) = \mathrm{In}[l(\beta)] = \sum_{i=1}^{n}\{y_i\,\mathrm{In}\,[\pi(x_i)] + (1-y_i)\mathrm{In}[1-\pi(x_i)]\} \quad (8)$$

Bu eşitliği maksimum yapan β değerlerini bulabilmek için eşitliğin β'ya göre türevi alınıp sıfıra eşitlenerek olabilirlik eşitlikleri elde edilir. Bu eşitliklerden elde edilen parametre tahminleri, en çok olabilirlik tahminidir. Eş.(8)'in LR modeli için açık ifadesi,

$$L(\beta) = \sum_{i=1}^{n}\left[y_i ln\left(\frac{e^{(x_i^T\beta)}}{1+e^{(x_i^T\beta)}}\right) + (1-y_i) ln\left(1-\frac{e^{(x_i^T\beta)}}{1+e^{(x_i^T\beta)}}\right)\right] \quad (9)$$

biçimindedir. LR modeli için olabilirlik eşitlikleri ise aşağıdadır.

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n}\left[y_i - ln\left(1-\frac{e^{(x_i^T\beta)}}{1+e^{(x_i^T\beta)}}\right)\right]x_{ij} \quad (10)$$

Sonuç olarak, L(β), doğrusal olmayan bir amaç fonksiyonu, β'lar ise karar değişkenleri olarak alındığında problem; kısıtsız, doğrusal olmayan bir optimizasyon problemi olarak düşünülebilir. Burada L(β)'nın maksimizasyonu amaçlanır. 1. dereceden kısmi türevler alındığında ulaşılan olabilirlik eşitlikleri parametrelerde doğrusal olmadığından dolayı iteratif ve amaç fonksiyonunun türevlenebilirlik özelliğine sahip olması gibi kısıtlayıcı bazı varsayımlar üzerine kurulu olan yöntemler geliştirilmiştir. NR algoritması yaygın olarak kullanılan bu yöntemlerin başında gelmektedir. Yöntem bir çok yerel maksimuma takılabilmekte ve global maksimum ya da minimum noktalara ulaşılamadan iterasyon tamamlanabilmektedir. Olabilirlik fonksiyonundaki değişim sonraki aşamalarda ihmal edilebilir duruma gelinceye kadar çözüme devam edilir Bu bir olumsuzluktur (Menard, 2002; Agresti, 2002).

İteratif yöntemlerin kısıtlayıcı varsayımları dikkate alındığında Eş.(10) ile verilen olabilirlik fonksiyonunu maksimize edecek en iyi parametre kümesini oluşturmak için GA yaklaşımını kullanmak, analitik yöntemlere iyi bir alternatif olarak görülmektedir. Sezgisel yöntemler kesin çözüme çok yakın çözümler veren ve hızla çözüme ulaşan yöntemlerdir. GA son yıllarda oldukça yaygın bir şekilde kullanılan

sezgisel yöntemlerden birisidir ve özellikle doğrusal olmayan fonksiyonların optimizasyonunda sıkça kullanılmaktadır.

5. GENETİK ALGORİTMA'NIN MATLAB KODLARI

Bu bölümde, GA'nın uygulanışı ile ilgili Matlab programındaki pratik kodlamalar tanıtılacaktır. GA, en genel şekli ile aşağıdaki biçimde kodlanmaktadır.

[x,fval,exitflag,output,population,scores]=ga(fitnessfcn,nvars,a,b,aeq,beq,lb,ub,nonlcon,options);

Burada eşitliğin sol tarafı çıktı değişkenlerini sağ tarafı ise girdi değişkenlerini oluşturmaktadır. Çıktı değişkenleri çözüm sonrasında yazdırılması gereken sonuçları belirtmektedir. 'x' fonksiyonu optimize edilen değişkenlerin sonucunu, 'fval' fonksiyonun optimum değerini yazdırmaktadır. Buraya yazılacak ek kodlar ile daha fazla çıktı değerinin yazdırılması sağlanabilmektedir.

Çıktı değişkenlerinin tamamının açıklamaları aşağıda verilmiştir.

[x,fval,exitflag,output,population,scores);

X :Fonksiyonu en küçük yapan değişken değerlerini yazdırır.

Fval :Amaç fonksiyonun x değişkeni için bulunan optimum değerini yazdırır.

Exitflag :Algoritmanın sonlandırma nedenini gösteren tamsayı değerini yazdırır. Bu tamsayı değeri doğrusal kısıt olmadığı durumlarda -5,-4,-2,-1, 0, 1, 5 değerlerini almaktadır. Örneğin çözüm sonucu ekranda '5' değeri yazdırılmışsa belirlenen zaman sınırının aşıldığı için iterasyonların durdurulduğu anlaşılmaktadır. Aşağıda tüm tamsayı değerlerinin kısa açıklaması verilmiştir.

(0) Nesillerin sayısı aşıldı;

(1) Optimum sonuç ve değişken değerleri bulundu;

(5) Belirlenen fonksiyon hassaslığı kısıtlama ihlali aşıldı;

(-1) Optimizasyon, çıktı veya çizim fonksiyonu tarafından durduruldu;

(-2) Uygun nokta bulunamadı;

(-4) Durma süresi sınırı aşıldı;

(-5) Zaman sınırı aşıldı.

Output :Her jenerasyonda algoritmanın performansı hakkında aşağıdaki bilgileri içeren bir çıktı verir.

Rngstate :Rasgele sayı üretecinin durumunu,

Generations	:Toplam nesil sayısını,
Funccount	:Toplam fonksiyon sayısını,
Maxconstraint	:Maximum kısıtlama ihlalini,
Message	:Sonlandırma mesajını,
Population	:Çözüm için oluşturulan popülasyonların çıktısını verir.
Scores	:Optimizasyon sonuç değerini yazdırır.

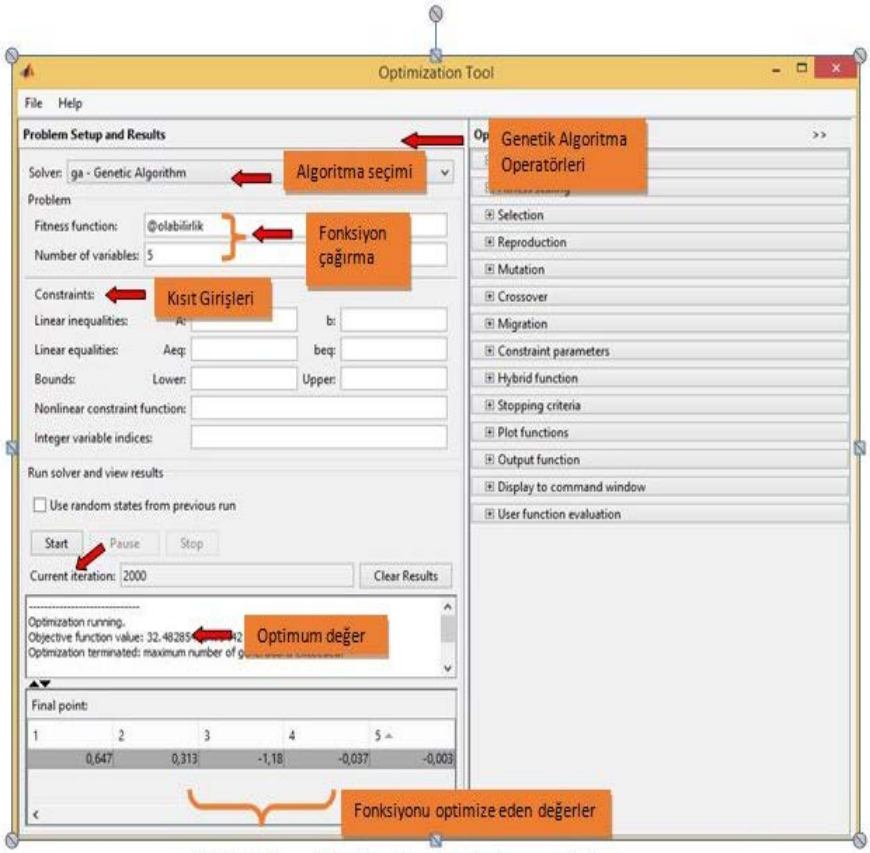
Girdi değişkenlerinin tamamının açıklamaları ise aşağıda verilmiştir.

ga(fitnessfcn,nvars,a,b,aeq,beq,lb,ub,nonlcon,options);

@fitnessfcn	: m dosyası olarak kaydedilen amaç fonksiyonunu çağırır.
Nvars	:Amaç fonksiyonundaki bağımsız değişken sayısı belirtilir.
A	:Eşitsizlik kısıtlamaları matrisi
B	:Eşitsizlik kısıtlamaları vektörü
Aeq	:Eşitlik kısıtlamaları matrisi
Beq	:Eşitlik kısıtlamaları matrisi
Lb	:Alt sınır
Ub	:Üst sınır
Nonlcon	:Doğrusal olmayan kısıt fonksiyonu
Options	:'gaoptimset' ile oluşturulan seçeneklerin yapısını belirler. Kodunun detaylı kullanım şekli aşağıdaki gibidir.

optıons = gaoptımset('param1',value1,'param2',value2,....).

'gaoptımset'ten sonra iki tırnak içerisinde yazılan tüm parametre değişkenleri için virgülden sonra yazılan değere göre çözüm aranması sağlanır. Optimum çözüm için belirleyeceğimiz mutasyon, seçim, çaprazlama gibi operatörlerin seçimleri, 'options' kodu ile yapılmaktadır.



Şekil 1. Genetik Algoritma Optimizasyon Ekranı

Şekil 1'de optimizasyon menüsü tanıtılmıştır. Sol tarafta,

"Solver" seçeneği ile, fonksiyonu optimize etmek için kullanılacak optimizasyon yöntemi seçilmektedir. Bu çalışmada model parametrelerinin optimum değerleri GA yöntemine göre belirlenmeye çalışıldığı için GA seçilmiştir.

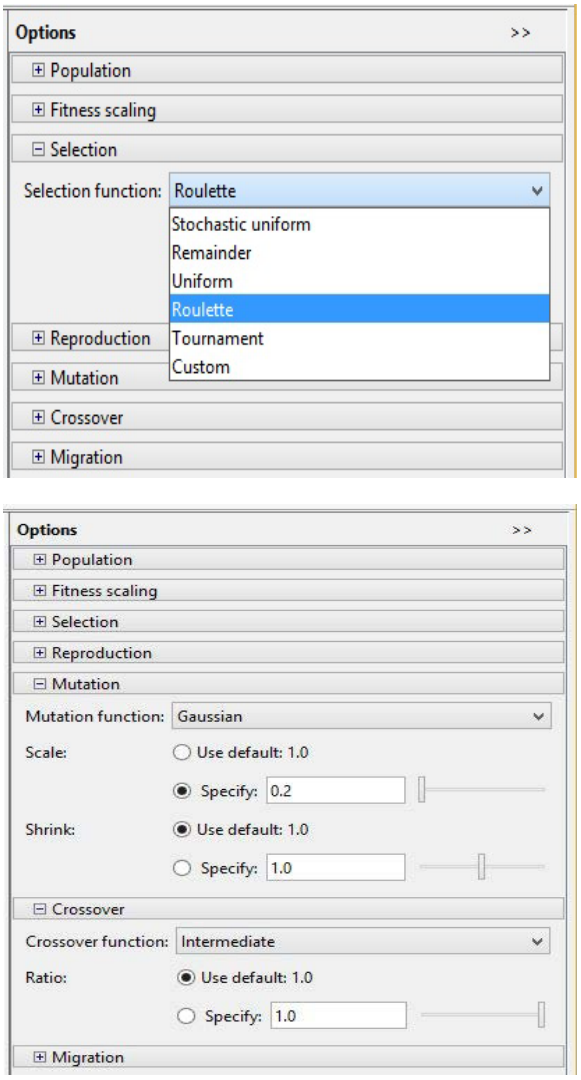
"Fitness function" bölümünde daha önce oluşturulan olabilirlik fonksiyonu çağrılmaktadır.

"Number of variable" bölümüne olabilirlik fonksiyonu için sabit parametre dahil olmak üzere modelde tahmin edilecek parametre sayısı girilmelidir.

"Constraints" bölümü kısıtlı optimizasyon problemleri için kullanılmaktadır.

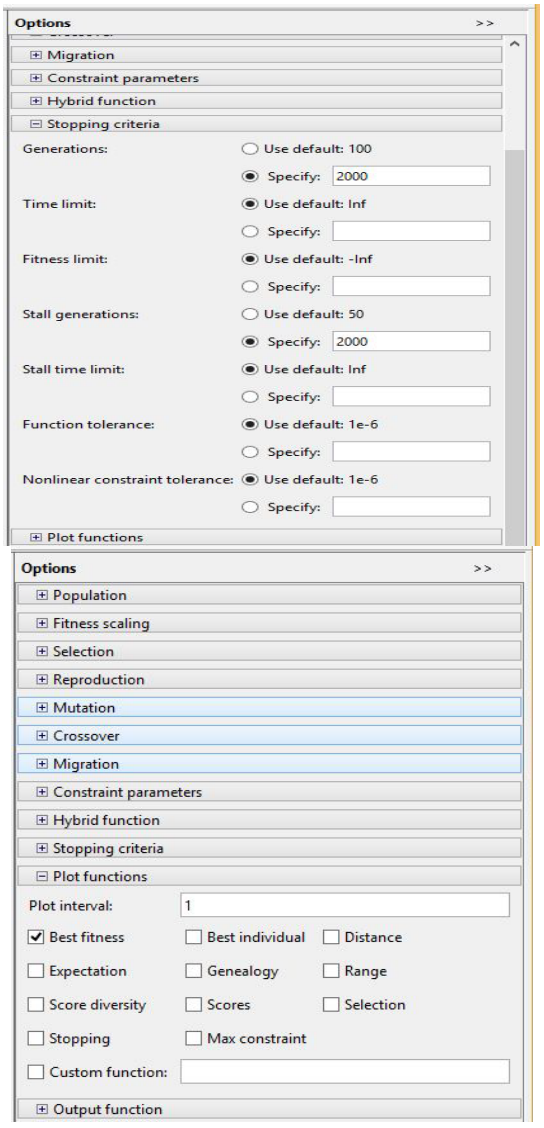
"Start" seçeneği seçildiğinde sağ tarafta "options" bölümünde seçilen tercihlere göre fonksiyon optimize edilmekte ve "Final point" kısmında parametre tahmin sonuçları bulunmaktadır.

"Objective function value" ile olabilirlik fonksiyonunun optimum değeri verilmektedir. Şekil 2.'nin sağ tarafında; GA operatörleri, durdurma kriterleri ve çıktı seçenekleri gibi birçok alternatif tercihin bulunduğu "options" bölümünün önemli operatörleri aşağıda tanıtılmıştır.



Şekil 2. Seçim Fonksiyonu, Mutasyon ve Çaprazlama Operatörleri

Şekil 2'nin sol tarafında Options menüsü altında bulunan seçim fonksiyonları gösterilmiştir. Seçim fonksiyonu için altı farklı seçenek bulunmaktadır. Şekil 2'nin sağ tarafında ise mutasyon ve çaprazlama fonksiyonları gösterilmiştir. Mutasyon fonksiyonu için bu bölümde beş, çaprazlama fonksiyonu için yedi farklı seçenek bulunmaktadır. Şekilde de görüldüğü üzere mutasyon oranı 0.2, mutasyon küçültme oranı 1.0 ve çaprazlama oranı 1.0 olarak belirlenmiştir. Daha hassas değerler için mutasyon oranı (scale) ve çaprazlama oranı (ratio) değerleri değiştirilebilmektedir.



Şekil 3. Durdurma Kriterleri ve Grafik Seçenekleri

Şekil 3'ün sol tarafındaki menüde optimizasyon için durdurma kriterleri gösterilmiştir. Bu menüde yedi farklı seçenek bulunmaktadır. Ayrıca daha hassas değerler için bu menüde fonksiyon ve doğrusal olmayan kısıt toleransı için 1e-7, 1e-8 veya daha fazla değerler tercihe bağlı olarak seçilebilir. Şekil 3'ün sağ tarafında ise grafik seçenekleri gösterilmiştir. "Best fitness" seçeneği ile optimum sonuca ulaşırken oluşturulan nesiller ve fonksiyon değerlerinin gösterildiği grafik çizdirilebilmektedir. Bu çalışmanın uygulama bölümü için alınan grafik çıktısı Şekil 4'de verilmiştir.

6. UYGULAMA

Bu çalışmada kullanılan gerçek veri kümesi, Hitit Üniversitesi Çorum Eğitim ve Araştırma Hastanesinin Cildiye bölümünden alınmıştır. Saç dökülmesi sorunu ile Cildiye bölümüne 2013 yılında başvuran hastalardan Alopesi (Saçkıran) tanısı konulan 52 hasta deney grubu olarak, alopesi hastalığı bulunmayan 43 hasta ise kontrol grubu olarak belirlenmiştir. Araştırma toplam 95 hastanın verilerinden oluşmaktadır. Çalışmaya katılan 95 hastanın %55'i deney grubunu %45'i kontrol grubunu oluşturmaktadır.

Çalışmada incelenen iki düzeyli bağımlı değişken, Alopesi Tanısı'dır. Tanı konulan hastalar "1", konulmayan hastalar ise "0" olarak kodlanmıştır. Alopesi tanısı konulmasını etkileyen değişkenler ve değişkenler ile ilgili yapılan kısaltmalar Tablo 1'de verilmiştir.

Tablo 1. Çalışmada Kullanılan Değişkenler

Değişkenler		**Özellikler**
Y	Alopesi (Saçkıran) Hastalığı	Nitel (Kategorik) 1 = Hastalık var 0 = Hastalık yok
X_1	Yaş	Nicel
X_2	Cinsiyet	Nitel (Kategorik) 1 = Erkek 0 = Kadın
X_3	ALT (karaciğer hücrelerinde bulunan enzimdir)	Nicel
X_4	AST (karaciğer hücrelerinde bulunan enzimdir)	Nicel
X_5	Glukoz	Nicel
X_6	HDL (High density lipoprotein) kolesterol (iyi huylu)	Nicel
X_7	LDL (Low-density lipoprotein) kolesterol (kötü huylu) (Karaciğerde üretilen ve kolestrolü kan yoluyla taşıyan moleküler proteinler)	Nicel
X_8	Kolesterol (dolaşım sisteminde bulunan bir lipiddir (yağlı bir madde)	Nicel
X_9	Serbest t3 (tiroit bezi hormonları)	Nicel
X_{10}	Serbest t4 (tiroit bezi hormonları)	Nicel
X_{11}	TSH (tiroid uyarıcı hormon)	Nicel
X_{12}	Trigliserit (Tg) (yağların ana bileşeni)	Nicel

"Cinsiyet" açıklayıcı değişkeni kategorik bir yapıya sahip olduğu için Erkek "1" ve Kadın "0" olarak kodlanmıştır. Erkek ve kadın hastaların oranı sırasıyla %60 (n=57) ve % 40 (n=38)'dır.

6.1. Lojistik Regresyon Analizi ve Newton-Raphson Yöntemi ile Parametre Tahmini

Daha önce de vurgulandığı gibi çalışmanın uygulama kısmının asıl amacı, kurulan LR modeli sonucu Alopesia hastalığına etki edebileceği düşünülen tüm aday açıklayıcı değişkenler için NR algoritması kullanılarak tahmin edilmiş olan parametreleri, GA ile de bulmaya çalışmak ve dolayısıyla GA'nın bağımlı değişkenin kategorik olduğu modellerde kullanılabilirliğini ölçmektir. Bunun için öncelikle hastalığa etki edebileceği düşünülen önemli risk faktörleri belirlenmiştir. Hastaların laboratuvar test sonuçları ve kategorik bir değişken olan cinsiyet, bağımsız değişkenler olarak alınmıştır. Parametreler öncelikle NR yöntemi ile tahmin edilmiştir. Parametre tahminleri ve diğer LR analiz sonuçları Tablo 2'de verilmiştir.

Tablo 2. LR Modelinde Newton-Raphson Algoritması İle Elde Edilen Sonuçlar

	$\widehat{\boldsymbol{\beta}}$	Standart Hata	Wald Testi	Serbestlik Derecesi	Sig.	$\text{Exp}(\widehat{\boldsymbol{\beta}})$
Sabit	2.985	3.165	0.889	1	0.346	19.782
Yaş	-0.025	0.024	1.029	1	0.310	0.976
Cinsiyet	1.618	0.619	6.840	1	0.009	5.044
ALT	-0.051	0.038	1.761	1	0.185	0.951
AST	0.135	0.059	5.237	1	0.022	1.145
Glukoz	-0.004	0.010	0.162	1	0.687	0.996
HDL	-0.012	0.045	0.070	1	0.791	0.988
Kolesterol	0.011	0.038	0.083	1	0.774	1.011
LDL	-0.032	0.037	0.750	1	0.387	0.968
Serbest t3	-1.611	0.777	4.306	1	0.038	0.200
Serbest t4	2.357	1.056	4.977	1	0.026	10.557
TSH	-0.039	0.240	0.026	1	0.871	0.962
Trigliserit	0.006	0.008	0.423	1	0.515	1.006

6.2. Genetik Algoritma ile Parametre Tahmini

GA ile parametre tahmin değerleri, Matlab paket programı kullanılarak elde edilmiştir. GA ile amaç fonksiyonunu optimize etmenin iki yolu vardır. Birincisi kod yazmayı gerektirmeyen, "Optimization Toolbox" menüsünü kullanmaktır. İkinci yol ise Matlab'de "command window" bölümüne gerekli kodları yazarak sonuca ulaşmaktır. Parametre tahmininde kullandığımız aşağıda tanımlanan

olabilirlik fonksiyonunu en çok yapan $\hat{\beta}$ parametreleri, optimizasyon sonuç değerleri, kodlar ve işlevleri özet olarak aşağıda verilmiştir.

Uygulamada kullanılan hastalara ait bağımsız değişken verileri için aşağıda verilen log-olabilirlik fonksiyonuna ulaşılmıştır.

f($\beta_0, \beta_1, \beta_2, \ldots \beta_{12}$) =52*x(1)+1481*x(2)+37*x(3)+1167*x(4)+1189*x(5)+4899*x(6)+2255.4*x(7)+8430*x(8)+4826.07*x(9)+178.45*x(10)+55.28*x(11)+103.67*x(12)+6695*x(13)-log(1+exp(x(1)+18*x(2)+x(3)+20*x(4)+17*x(5)+100*x(6)+39.1*x(7)+199*x(8)+76.7*x(9)+4.57*x(10)+1.31*x(11)+2.5*x(12)+416*x(13)))-log(1+exp(x(1)+23*x(2)+14*x(4)+30*x(5)+92*x(6)+39.2*x(7)+143*x(8)+93.19*x(9)+2.97*x(10)+0.79*x(11)+0.94*x(12)+53*x(13)))-.....-log(1+exp(x(1)+25*x(2)+10.6*x(4)+16.5*x(5)+88*x(6)+41.3*x(7)+140*x(8)+89.09*x(9)+3.19*x(10)+0.62*x(11)+2.52*x(12)+136.7*x(13)))-log(1+exp(x(1)+25*x(2)+13*x(4)+17*x(5)+81*x(6)+69.7*x(7)+202*x(8)+118.7*x(9)+3.01*x(10)+0.86*x(11)+0.19*x(12)+68*x(13)))

Ancak program maksimizasyon üzerine değil minimizasyon üzerine kurulu olduğundan dolayı $f(\beta_0, \beta_1, \beta_2, \ldots, \beta_{12})$ fonksiyonu (–1) ile çarpılarak bu fonksiyonu minimize eden $\beta_0, \beta_1, \beta_2, \ldots, \beta_{12}$ parametrelerinin tahminleri bulunacaktır.

Matlab Paket programında en basit haliyle GA optimizasyon kodu,

[x,fval] = ga(@olabilirlik,13,options)

şeklindedir. Burada eşitliğin sağ tarafı girdileri sol tarafı ise çıktıları oluşturmaktadır. Girdiler, fonksiyonun tanımlanması, değişken sayısının girilmesi, çözüm araması hangi operatörler ile yapılacaksa bunların tanımlanmasından oluşmaktadır. Ayrıca özel olarak başlangıç popülasyonu, seçim tekniği, mutasyon ve çaprazlama oranları, iterasyonların tüm durdurma kriterleri bu bölümdeki 'options' kodu ile yapılmaktadır.

Çıktı fonksiyonları ise çözüm sonrasında yazdırılması gereken sonuçları belirtmektedir. 'x' fonksiyonu optimize edilen bağımsız değişkenlerin sonucunu, 'fval' fonksiyonun optimum değerini yazdırmaktadır. Buraya yazılacak ek kodlar ile daha fazla çıktı değerinin yazdırılması sağlanabilmektedir.

Yukarıda açıklanan tüm bilgiler sonucu GA ve NR algoritması ile elde edilen $\beta_0, \beta_1, \beta_2, \ldots, \beta_{12}$ parametreleri için tahmin değerleri ve log-olabilirlik fonksiyonunun negatifinin minimum değerleri Tablo 3'de verilmiştir.

Tablo 3 incelendiğinde, optimum değerler arasındaki fark yaklaşık olarak 0.0498 bulunmuştur. GA ile parametre tahmininde tablodaki 52.833533819034635 optimizasyon sonuç değerini bulabilmek için; population size '50'; selection function 'roulette'; reproduction crossover function '0.95'; mutation function 'gaussian'; mutation scale '0.15'; mutation shrink '0.88'; crossover function 'intermediate'; crossover ratio '1.0'; durdurma kriterleri ise generations '50000'; stall generations '50000'; function tolerance '1e-6'; nonlinear constraint tolerance

'1e-6' olarak seçilmiştir. Diğer operatör ve oranlar ise varsayılan olarak işlem görmüştür.

Tablo 3. Parametre Tahminlerinin Karşılaştırılması

Parametre Tahminleri	**Lojistik Regresyon Analizi**	
	Newton-Raphson	**Genetik Algoritma**
$\hat{\beta}_0$ (Sabit)	2.985	1.984
$\hat{\beta}_1$	-0.025	-0.022
$\hat{\beta}_2$	1.618	1.548
$\hat{\beta}_3$	-0.051	-0.049
$\hat{\beta}_5$	-0.004	-0.004
$\hat{\beta}_6$	-0.012	-0.01
$\hat{\beta}_7$	0.011	0.012
$\hat{\beta}_8$	-0.032	-0.032
$\hat{\beta}_9$	-1.611	-1.424
$\hat{\beta}_{10}$	2.357	2.283
$\hat{\beta}_{11}$	-0.039	-0.024
$\hat{\beta}_{12}$	0.006	0.005
Log-olabilirlik fonksiyonun negatifinin minimize değeri	**52.8833**	**52.833533819034635**

Çalışmanın başında da belirttiğimiz gibi amaç, bağımlı değişkenin kategorik olduğu modellerde parametre tahmini için kısıtlayıcı varsayımlar gerektiren klasik NR algoritmasına alternatif bir algoritma önerebilmekti. Bu amaçla GA, daha esnek varsayımlar gerektirmesi ve yukarıda bahsedilen birçok avantajından dolayı iyi bir optimizasyon yöntemi olarak düşünülmüş ve klasik NR algoritması ile optimize edilen log-olabilirlik fonksiyonunun GA ile optimize edilmesi durumda nasıl bir sonuç vereceği araştırılmak istenmiştir.

Her iki yöntemle de fonksiyonun negatifinin minimum değeri birbirine çok yakındır. Ancak optimizasyon problemlerinde çok küçük bir fark bile çok önemli olabilmekte; bu durum dikkate alındığında ise GA'nın bu örnek için NR algoritmasına göre daha iyi sonuç verdiği gözlenmiştir. β parametrelerinin tahminleri incelendiğinde ise yine birbirine çok yakın değerler bulunmuştur. En fazla farklılığın β_0 (sabit) parametre tahmininde olduğu görülmektedir.

Bu sonuçlar ile araştırmalarda kullanılacak verilerin özelliklerine göre bazı analizlerde GA ile farklı operatör ve oranlar kullanılarak NR algoritmasının bulduğu parametre tahmin değerlerinden daha iyilerinin bulunabileceği ortaya çıkmıştır. Böylece NR algoritmasına çok yakın hatta daha iyi sonuç veren GA'nın da kategorik

bağımlı değişkenlerin parametre tahminlerinde kullanılmasının çok iyi bir alternatif olduğu söylenebilmektedir.

6.3. Yapay Veri Üzerinden Elde Edilen Sonuçlar

Çalışmanın bu bölümünde, iki düzeyli kategorik yapıdaki bir bağımlı değişken ve dört açıklayıcı değişkenin yer aldığı Matlab programı ile rastgele üretilen veriler kullanılmıştır. Bağımlı değişken Y, Bernoulli dağılımından, bağımsız değişkenlerden üç tanesi (X_1, X_2, X_3) yine Bernoulli dağılımından ve X_4 ise normal dağılımdan gelmektedir. Oluşturulan farklı senaryolar altında hem NR algoritma sonuçlarına hem de GA sonuçlarına yer verilmiştir.

6.3.1. Newton Raphson algoritması sonuçları

NR yöntemini kullanılarak elde edilen parametre tahminleri ve diğer LR analiz sonuçları aşağıda verilmiştir.

Tablo 4. Lojistik Regresyon Analizi Sonuçları

	$\hat{\beta}$	S.Hata	Wald	df	Sig.	Exp($\hat{\beta}$)
Sabit	0.744	0.817	0.829	1	0.363	2.103
X_1	0.333	0.620	0.288	1	0.591	1.395
X_2	-1.175	0.607	3.745	1	0.053	0.309
X_3	-0.060	0.616	0.010	1	0.922	0.941
X_4	-0.006	0.022	0.090	1	0.764	0.994
Min (-Log Olabilirlik Fonk.)	**32.4960**					

Tablo 4'de NR algoritmasına göre olabilirlik fonksiyonunun maksimum (-log olabilirlik fonksiyonunun minimum) değeri 32.4960 bulunmuştur. Tabloda ayrıca parametre tahminleri, tahminlerin standart hataları, Wald testi sonuçları ve odds oranlarını veren exp ($\hat{\beta}$) değerlerine yer verilmiştir.

Uygulamalı çalışmaların çoğunda model parametreleri olabilirlik fonksiyonunun diferansiyellenebilir özellikte olup olmadığına bakılmaksızın ve iterasyon sürecinde karşılaşılan tahmin edicilerin başlangıç değerlerinin belirlenmesi problemi dikkate alınmadan geleneksel olarak NR algoritması ile tahmin edilmektedir. Gerçek veriler ile yapılan çalışmalarda, sonuçların güvenilir olması ve gerçeği yansıtması için parametrelerin en iyi tahminlerine ulaşmak ve gerçeğe daha yakın istatistiksel sonuçlar elde edip yorumlamak gerekir. Ancak, parametreler için öngörülen başlangıç değerlerinin gerçek parametre değerlerinden önemli ölçüde uzak olması durumunda iterasyon sayısı artmakta; yakınsama hızı azalmakta ya da süreç yerel optimumlara takılıp global optimum noktalarda yakınsama sağlanamamaktadır. Bu olumsuzlukları da göz önünde bulundurarak bir sonraki bölümde NR algoritması ile karşılaştırıldığında daha esnek varsayımlara sahip olan GA algoritmasının başarısı

tartışılmaktadır. Bu tartışmanın anlamlı olabilmesi için NR algoritmasının optimum koşullarının sağlandığı varsayımı altında GA algoritmasının NR algoritma sonuçlarına yakınlığı gözlenmiştir.

6.3.2. Genetik algoritma sonuçları

Bu bölümde GA'nın parametre tahmininde yakınsama hızı ve optimum sonuç değerleri göz önüne alındığında hangi koşullarda en iyi sonuca ulaştığı üzerinde durulmuştur. Bu amaçla rastgele üretilen veriler ile GA için parametre tahmininde en uygun başlangıç popülasyonu, seçim, çaprazlama, mutasyon operatörleri ve çaprazlama ve mutasyon olasılıkları belirlenmiştir.

Tablolar incelendiğinde, GA ile çözüm yapılırken birden çok başlangıç popülasyonu, çaprazlama ve mutasyon oranları ve farklı seçim yöntemlerinin denendiği görülmektedir. Optimum sonuca ulaşmak için tek tek alternatifler tasarlanarak GA ile 100 den fazla farklı ihtimal üzerinde durulmuştur. Mutasyon fonksiyonu için dört (Use Constraint Dependent Default, Gaussian, Uniform, Adaptive Feasible); Mutasyon oranı için on; Crossover fonksiyonu için altı (Scattered, Sinle point, Two point, Intermediate, Heuristic, Arithmetic), Crossover oranı için on; Seçim tercihi için beş ve diğer değişkenler için farklı ihtimaller denenmiştir. Yapılan bu denemelerden üç tanesinin sonucu aşağıda verilen tablolarda yer almaktadır.

Tablo 5. Popülasyon Boyutuna Göre Yapılan Karşılaştırmalar

Population Size	**20**	**40**	**60**	**80**	**100**
Selection Function	Roulette	Roulette	Roulette	Roulette	Roulette
Mutation function	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian
Mutation ratio	0.2	0.2	0.2	0.2	0.2
Mutation shrink	1.0	1.0	1.0	1.0	1.0
CrossoverFcn	Intermediate	Intermediate	Intermediate	Intermediate	Intermediate
Crossover ratio	1.0	1.0	1.0	1.0	1.0
Generations	5000	5000	5000	5000	5000
StallGenLimit	5000	5000	5000	5000	5000
TolFun	1e-6	1e-6	1e-6	1e-6	1e-6
TolCon	1e-6	1e-6	1e-6	1e-6	1e-6
Others	Default	Default	Default	Default	Default
$\hat{\beta}_0$ (Sabit)	0.647	0.647	0.647	0.647	0.647
$\hat{\beta}_1$	0.313	0.313	0.313	0.313	0.313
$\hat{\beta}_2$	-1.18	-1.18	-1.18	-1.18	-1.18
$\hat{\beta}_3$	-0.036	-0.037	-0.037	-0.037	-0.037
$\hat{\beta}_4$	-0.003	-0.003	-0.003	-0.003	-0.003
Min (-Log-Olabilirlik Fonksiyon Değeri)	32.4828546652257	**32.482854664707**	32.4828546 64724	32.482854664709	32.4828546 64708

Tablo 5 incelendiğinde, diğer özellikler sabit tutularak popülasyon boyutu için beş farklı durum (20-40-60-80-100) incelenmiştir. Parametre tahminlerinin farklılık göstermediği bu tasarımda, popülasyon büyüklüğünün en az olduğu durum (20) için log-olabilirlik fonksiyonunun en küçük (-log olabilirlik fonksiyonunu en büyük) değeri aldığı (32.4828546652257) ve en kötü sonucun elde edildiği görülmektedir. Daha büyük popülasyon değerleri için daha iyi sonuçlar elde edilse de popülasyon büyüklüğü arttıkça her zaman daha iyi sonuçların elde edildiği söylenemez. Optimal popülasyon büyüklüğünü bulmak en doğru sonucu verecektir.

Tablo 6. Durdurma Kriterlerine Göre Yapılan Karşılaştırmalar

PopulationSize	60	60	60	60	60
Selection Function	Roulette	Roulette	Roulette	Roulette	Roulette
Mutation function	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian
Mutation ratio	0.2	0.2	0.2	0.2	0.2
Mutation shrink	1.0	1.0	1.0	1.0	1.0
CrossoverFcn	Intermedia	Intermediate	Intermedia	Intermediate	Intermedia
Crossover ratio	te	1.0	te	1.0	te
Generations	1.0	20000	1.0	20000	1.0
StallGenLimit	10000	20000	30000	20000	20000
TolFun	10000	1e-6	30000	1e-10	20000
TolCon	1e-6	1e-6	1e-6	1e-10	1e-15
Others	1e-6	Default	1e-6	Default	1e-15
	Default		Default		Default
$\hat{\beta}_0$ (Sabit)	0.647	0.647	0.647	0.647	0.647
$\hat{\beta}_1$	0.313	0.313	0.313	0.313	0.313
$\hat{\beta}_2$	-1.18	-1.18	-1.18	-1.18	-1.18
$\hat{\beta}_3$	-0.037	-0.037	-0.037	-0.037	-0.037
$\hat{\beta}_4$	-0.003	-0.003	-0.003	-0.003	-0.003
Min (-Log-Olabilirlik Fonksiyon Değeri)	32.48285466470744	**32.482854664706964**	32.48285466470703	32.48285466470710	32.48285466470703

"Durdurma kriteri" ve "Nesil sayısının" değişken olarak alındığı tasarımda, popülasyon büyüklüğü sabitlenerek "60" alınmıştır. Sonuçlar incelendiğinde, iterasyonu durdurma kriteri "1e-6" için en düşük fonksiyon değerinin (32.482854664706964) "20000" nesil sayısı için elde edildiği yani iyi bir model sonucu için optimal nesil sayısının da doğru belirlenmesi gerektiği gözlenmiştir. Ayrıca nesil sayısının 20000 olarak sabitlendiği planda, "1e-6", "1e-10" ve "1e-15" olarak belirlenen üç farklı durdurma kriterine göre sonuçlar incelendiğinde ise, tahmin edilen parametreler arasındaki farkın en küçük olduğu **"1e-6"** değeri için optimal fonksiyon değerine ulaşılmıştır.

Tablo 7. Seçim Yöntemine Göre Karşılaştırmalar

PopulationSize	60	60	60	60
Selection Function	**Stochastic Uniform**	**Roulette**	**Uniform**	**Tournament (size 4)**
Mutation function	Gaussian	Gaussian	Gaussian	Gaussian
Mutation ratio	0.2	0.2	0.2	0.2
Mutation shrink	1.0	1.0	1.0	1.0
CrossoverFcn	Intermediate	Intermediate	Intermediate	Intermediate
Crossover ratio	1.0	1.0	1.0	1.0
Generations	20000	20000	20000	20000
StallGenLimit	20000	20000	20000	20000
TolFun	1e-6	1e-6	1e-6	1e-6
TolCon	1e-6	1e-6	1e-6	1e-6
Others	Default	Default	Default	Default
$\hat{\beta}_0$ (Sabit)	0.647	0.647	0.647	0.646
$\hat{\beta}_1$	0.313	0.313	0.313	0.314
$\hat{\beta}_2$	-1.18	-1.18	-1.18	-1.18
$\hat{\beta}_3$	-0.037	-0.037	-0.037	-0.036
$\hat{\beta}_4$	-0.003	-0.003	-0.003	-0.003
Min (-Log-Olabilirlik Fonksiyon Değeri)	32.48285466470700	**32.482854664706964**	32.48285478452693	32.48285501285599

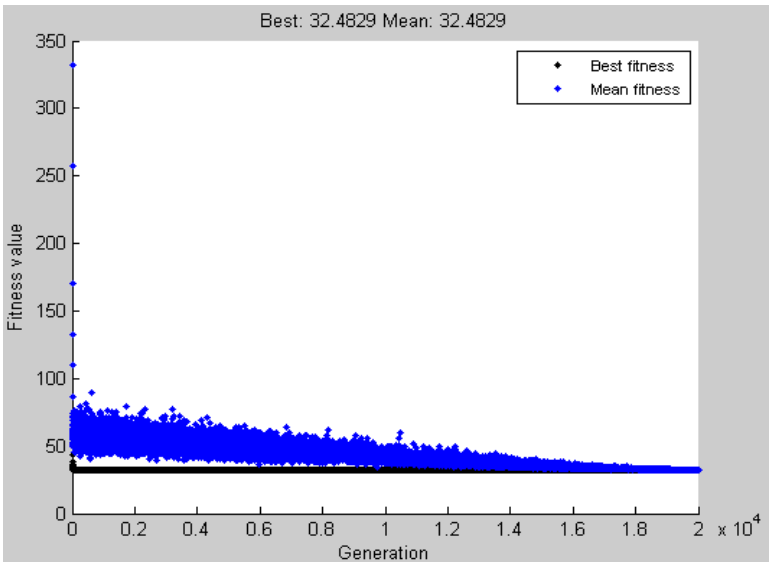
Tablo 7'de farklı seçim fonksiyonlarına göre elde edilen sonuçlara yer verilmiştir. "Stochastic Uniform", "Roulette", "Uniform" ve "Tournament" olarak belirlenen fonksiyonlara göre, en iyi sonucun "Roulette" ile en kötü sonucun ise "Tournament" ile elde edildiği görülmektedir.

Yapılan tüm denemeler dikkate alındığında optimum sonucun bulunduğu kombinasyon Tablo 8 ile ve iterasyon sürecinin grafiksel gösterimi ise Şekil 4 ile verilmiştir.

Tablo 8. Kodlamada Kullanılan Parametreler

GA parametresi	**Değer / Metod**
PopulationType	'doubleVector'
PopulationSize	60
Selection Function	Roulette
Mutation function	Gaussian
Mutation ratio	0.2
Mutation shrink	1.0
CrossoverFcn	Intermediate
Crossover ratio	1.0
Migration Direction	'forward'
Others (InitialPopulation,	Use default

InitialScores et al.)	
Stopping criteria	
Generations	20000
StallGenLimit	20000
TolFun	1e-6
TolCon	1e-6
Output	
PlotFcns	@gaplotbestf
Display	'iter','final','diagnose'
Others	Use default



Şekil 4. Uygunluk Değeri (20000 iterasyon)

Şekil 4 incelendiğinde, 20000 iterasyon sonucunda uygunluk fonksiyonunun hangi değerlerden başladığı, hangi sonuçlara ulaştığı ve arama noktalarının serpilmeleri daha açık görülmektedir. Uygunluk fonksiyonunun değerinin ilk iterasyonlarda 50 ve 100 civarında olduğu daha sonra ise 10000 iterasyon sonrasında genetik operatör işlemleri sonucunda fonksiyonun optimum değeri olan 30-35 aralığına yaklaştığı görülmektedir.

Tablo 9. GA'da İterasyon Sayıları ve Ulaşılan Optimum Değerler

Generation	En İyi f-sıklığı	Ortalama f(x)	Stall f(x)	Generation
19981	1198920	32.48	32.48	1927
19982	1198980	32.48	32.48	1928
19983	1199040	32.48	32.48	1929
19984	1199100	32.48	32.48	1930
19985	1199160	32.48	32.48	1931
19986	1199220	32.48	32.48	1932
19987	1199280	32.48	32.48	1933
19988	1199340	32.48	32.48	1934
19989	1199400	32.48	32.48	1935
19990	1199460	32.48	32.48	1936
19991	1199520	32.48	32.48	1937
19992	1199580	32.48	32.48	1938
19993	1199640	32.48	32.48	1939
19994	1199700	32.48	32.48	1940
19995	1199760	32.48	32.48	1941
19996	1199820	32.48	32.48	1942
19997	1199880	32.48	32.48	1943
19998	1199940	32.48	32.48	1944
19999	1200000	32.48	32.48	1945
20000	1200060	32.48	32.48	1946

LR modelinin parametre tahmininde kullanılan alternatif iki optimizasyon yönteminden elde edilen sonuçlar aşağıdaki tabloda karşılaştırmalı olarak verilmiştir.

Tablo 10. En İyi Parametre Tahminlerinin Karşılaştırılması

Parametre Tahminleri	Lojistik Regresyon	
	Newton-Raphson	**Genetik Algoritma**
$\hat{\beta}_0$ (Sabit)	0.744	0.647
$\hat{\beta}_1$	0.333	0.313
$\hat{\beta}_2$	-1.175	-1.18
$\hat{\beta}_3$	-0.060	-0.037
$\hat{\beta}_4$	-0.006	-0.003
Min(-Log Olabilirlik Fonksiyonu)	32.4960	**32.482854664706964**

Tablo 10 incelendiğinde, parametre tahminlerinin işaretlerinde ve büyüklüklerinde önemli farklılıklar gözlenmemiştir. Her iki algoritmadan elde edilen sonuçlara göre, olabilirlik fonksiyonunun negatifinin minimum değerlerine göre GA yaklaşımının NR algoritmasına göre daha başarılı olduğu sonucuna ulaşılmıştır.

Bu sonuç bize kategorik bağımlı değişken modellemesinde parametre tahmini için klasik optimizasyon tekniklerinin varsayımlarının sağlanamaması durumunda sezgisel optimizasyon tekniklerinin de kullanılabilir olduğunu ve daha başarılı sonuçların bile elde edilebileceğini göstermiştir.

7. SONUÇ VE ÖNERİLER

Bu çalışmada, GA ve iteratif kök bulma algoritmalarından NR algoritmasının etkinliği, bağımlı değişkenin kategorik yapıda olduğu LR modeli altında incelenmiştir. Ayrıca bu alanda çalışan uygulamacılar için yöntemin Matlab kodları ayrıntılı olarak tanıtılmıştır.

Parametre tahmininde kullanılan NR gibi iteratif yöntemler başlangıç noktası problemi ve optimize edilecek fonksiyonun sürekli ve türevlenebilir olması gibi kısıtlayıcı varsayımlar gerektirmektedir. GA'da ise arama işlemi tek nokta yerine potansiyel çözümlerin bir kümesi üzerinde gerçekleştirilir ve en iyi çözüme ulaşıncaya kadar çözümler değerlendirilir. GA'ların, problemlerin çözümü için türev veya diğer yardımcı bilgilere gereksinim duymaması ve yerel optimum noktalara takılmadan global optimum noktalarını bulabilmesi gibi bazı avantajlarından dolayı kategorik bağımlı değişken modellemesinde parametre tahmini için iteratif yöntemlere alternatif olabileceği düşünülmüştür. Bu amaçla yapılan çalışmada farklı senaryolar altında model parametreleri NR ve GA yaklaşımları ile tahmin edilmiş ve olabilirlik fonksiyonunu en büyük yapan parametre tahminlerinin birbirine çok yakın olduğu gözlenmiştir. Buna göre, klasik optimizasyon varsayımlarının sağlanması durumunda GA yaklaşımının NR algoritması kadar başarılı olduğu sonucuna ulaşılmıştır. Böylece, iki düzeyli bağımlı değişken modellerinde klasik optimizasyon varsayımlarının sağlanamaması durumunda kullanılması olanaksız olan NR algoritmasının yerine GA yaklaşımının başarılı bir şekilde uygulanabileceği sonucuna ulaşılmıştır.

KAYNAKLAR

Agresti, A., (2002). Categorical Data Analysis. 2th edt., New Jersey, John Wiley&Sons Inc.

Aguilar-Rivera, R., Valenzuela-Rendón, M., Rodríguez-Ortiz, J.J., (2015), "Genetic Algorithms and Darwinian Approaches in Financial Applications: A Survey", Expert Systems with Applications, 42(21), 7684-7697.

Altunkaynak, B., Esin, A., (2004), "Doğrusal Olmayan Regresyonda Parametre Tahmini İçin Genetik Algoritma Yöntemi". Gazi Üniversitesi Fen Bilimleri Dergisi, 17(2), 43-51.

Babaoğlu, İ., Findik, O., Ülker, E., (2010), "A Comparison of Feature Selection Models Utilizing Binary Particle Swarm Optimization and Genetic Algorithm in Determining Coronary Artery Disease Using Support Vector Machine", Expert Systems with Applications, 37(4), 3177-3183.

Goldberg, D.E., (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, Reading, MA.

Goldberg D.E., Deb, K., (1991), A Comparative Analysis of Selection Schemes Used in Genetic Algorithms, Foundations of Genetic Algorithms., San Francisco, CA: Morgan Kaufmann.

Gordini, N., (2014), "A Genetic Algorithm Approach for Smes Bankruptcy Prediction: Empirical Evidence From Italy", Expert Systems with Applications, 41(14), 6433-6445.

Hadi, H.S., J.L. Gonzalez-Andujar, (2009), "Comparison of Fitting Weed Seedling Emergence Models With Nonlinear Regression and Genetic Algorithm", Computers and Electronics in Agriculture, 65(1), 19-25.

Hadji, S., Gaubert, J.P., Krim, F., (2015). "Theoretical and Experimental Analysis of Genetic Algorithms Based MPPT for PV Systems", Energy Procedia, 74, 772-787.

Holland, J.H., (1975). Adaptation in Natural and Artificial Systems. USA, University of Michigan Press.

Holland, J.H., (1992). Adaptation in Natural and Artificial Systems. 2th edition, Cambridge, London., The MIT Press.

Karr, C.L., Freeman, M. L., (1999). Industrial Applications of Genetic Algorithms., USA, CRC Press.

Koh, Y., Yap, C.W., Li, S.C., (2008). "A Quantitative Approach of Using Genetic Algorithm in Designing A Probability Scoring System of an Adverse Drug Reaction Assessment System", International Journal of Medical Informatics, 77(6), 421-430.

Johnson, P., Graham, P., Wilson, P., Macaulay, L., Maruff, P., Savage, G., Ellis, K., Martins, R., Rowe, C., Masters, C., Ames, D., Zhang, P., (2013), "Genetic Algorithm with Logistic Regression for Alzheimer's Disease Diagnosis and Prognosis", Alzheimer's & Dementia, 9(4), P455-P456.

Lee, K.H., Kim, K.W., (2015), "Performance Comparison of Particle Swarm Optimization and Genetic Algorithm for Inverse Surface Radiation Problem", International Journal of Heat and Mass Transfer, 88, 330-337.

Liu, H.H., Ong, C.S., (2008), "Variable Selection in Clustering for Marketing Segmentation Using Genetic Algorithms", Expert Systems with Applications, 34(1), 502-510.

Menard, S., (2002). Applied Logistic Regression Analysis, 2th Edition, USA, Sage Publications.

Meng, Q., Weng, J., (2011), "A Genetic Algorithm Approach to Assessing Work Zone Casualty Risk". Safety Science, 49, 1283-1288.

Mitchell, M., (1999). An Introduction to Genetic Algorithms, 5th Edition, Cambridge, London, The Mit Press.

Pasia, J., Hermosilla, A., Ombao, H., (2005), "A Useful Tool for Statistical Estimation: Genetic Algorithm", Journal of Statistical Computation and Simulation, 75(4), 237-251.

Pfeifer, J., Barker, K., Ramirez-Marquez, J.E., Morshedlou, N., (2015), "Quantifying the Risk of Project Delays with a Genetic Algorithm", International Journal of Production Economics, 170(A), 34-44.

Reeves, C.R., Rowe, J.E., (2002). Genetic Algorithms Principles and Perspectives. A Guide to GA Theory., USA., Kluwer Academic Press.

Rechenberg, I., (1973), Evolutions Strategie–Optimierungtechnischersystemenach Prinzipien Der Biologischen Evolution. (PhD.Thesis). Fromman-Holzboog, Germany.

Stylianou, N., Akbarov, A., Kontopantelis, E., Buchan, I., W. Dunn, K., (2015), "Mortality Risk Prediction in Burn Injury: Comparison of Logistic Regression with Machine Learning Approaches", Burns, 41(5), 925-934.

Yuan, F.C., Lee, C.H., (2015), "Using Least Square Support Vector Regression with Genetic Algorithm to Forecast Beta Systematic Risk", Journal of Computational Science, 11, 26-33.