

Çoklu Doğrusal Regresyonda Uygun Model Seçiminde Genetik Algoritma Yaklaşımının Kullanılması

Yılmaz KAYA¹ Abdullah YEŞİLOVA² M. Nuri ALMALI³

¹ Yüzüncü Yıl Üniversitesi, Van Meslek Yüksekokulu, Bilgisayar Teknolojileri ve Programcılığı, 65080 VAN

² Yüzüncü Yıl Üniversitesi, Ziraat Fakültesi, Zootekni, Biyometri Genetik Anabilimdalı, 65080 VAN

³ Yüzüncü Yıl Üniversitesi, Mühendislik Mimarlık Fakültesi, Elektrik ve Elektronik Anabilimdalı, 65080 VAN

Özet: Çoklu doğrusal regresyon modelinde bağımsız değişken sayısının fazla olması durumunda, bağımlı değişkeni en iyi açıklayan bağımsız değişkenlerin seçilmesi oldukça önemli bir aşamadır. Uygun modelin oluşturulmasında birçok yöntem kullanılmaktadır. Değişken sayısının çok fazla olması durumunda klasik yöntemler yetersiz kalabilmektedir. Klasik yöntemler yerine genetik algoritma yöntemi kullanılarak uygun model seçimi yapılabilir. Bu çalışmada, beden eğitim ve spor alanında elde edilen gerçek bir veri kümesinde, genetik algoritma yaklaşımı kullanılarak en iyi model seçimi yapılmıştır. Sınav mekik sayısını en iyi açıklayan bağımsız değişkenlerin sınav öncesi mekik sayısı ile sınava hazırlanma süresinin olduğu saptanmıştır.

Anahtar Kelimeler: Optimizasyon, Genetik Algoritma, En İyi Doğrusal Model

Using Genetic Algorithm Approach for Select Suitable Model in Multiple Linear Regression

Abstract: The selection of independent variables is an important stage for the best expression of dependent variable in case of more independent variable in multiple linear regression. A lot of methods have been used for the formation of suitable models. The conventional methods have been insufficient in case of much more variables. The suitable model selection can be made using genetic algorithm instead of conventional methods. In the present study, the best model selection was made using genetic algorithm approach in real data cluster provided from physical education and sport area. The best independent variables expressing the number of shuttle are determined to be the preexam shuttle number and the period of the preparation for exam.

Key Words: Optimization, Genetic Algorithm, Best linear model

Giriş

Regresyon yöntemleri, bağımlı değişken ile bağımsız değişkenler arasındaki fonksiyonel ilişkiyi modellemek için kullanılır. Çoklu doğrusal regresyon modeli, bağımlı değişkendir toplam değişimi birden fazla bağımsız değişken ile açıklamak için kurulur. Bağımsız değişken sayısının çok olması durumunda hangi değişkenlerin modele dahil edileceği çoğu zaman problem olmaktadır. Uygun modelin kurulmasında kaç değişkenin veya hangi değişkenlerin ölçülmesi gerektiğine karar vermek geleneksel yöntemlerle mümkün değildir. Ölçülen tüm değişkenlerle oluşturulan modeller çoğu zaman çeşitli problemlere neden olmaktadır. Bu yüzden bağımlı değişkeni en iyi açıklayan bağımsız değişkenlerin seçilmesi önem kazanmaktadır (Draper ve Smith, 1989).

Değişken sayısının artmasıyla oluşturulabilecek model sayısı üstel olarak artmaktadır. Örneğin değişken sayısının $n=20$ veya $n=200$ olması durumunda $2^{20} \sim 10^6$ ve $2^{200} \sim 10^{20}$ kadar model oluşturulabilmektedir (Andrew ve Dan, 2000). Elde edilen tüm değişkenler bizim için önemli olmayabilir. Bağımlı değişkeni en iyi açıklayan değişkenlerin seçilmesi gerekir.

Uygun değişkenlerin seçilmesi regresyon analizinin önemli aşamalarından biridir. En iyi regresyon yaklaşımı, veri setinde tüm bağımsız değişkenlerle bağımlı değişken arasında kurulabilecek tüm olası kombinasyonlara ilişkin regresyon denklemlerini belirlemektedir (Miller, 1990). Değişken seçimi ile bağımlı değişkene katkısı olan bağımsız değişkenler seçilerek daha doğru tahminler yapılır. Değişken seçiminde çok sayıda yöntem kullanılmaktadır. Genelde aşamalı

yöntemler (Forward, Backward, Stepwise) veya kriter bazlı ölçütler (AIC, BIC, Belirleme katsayısı= R^2 , Düzeltilmiş Belirleme Katsayısı= R_{Diz}^2 , CP Mallow İstatistiği= C_p ve Hata Kareler Ortalaması= HKO) kullanılmaktadır (Draper ve Smith, 1998). Aşamalı yöntemlerde tahmin edilen eşitlik bulunana kadar, her adımda modele bir değişken eklenir ya da modelden bir değişken çıkarılır. Aşamalı yöntemlerin bazılarındaki değişkenler arasındaki etkileşimler dikkate alınmamaktadır (Yeniay ve Göktaş, 2003). Oysa bağımsız değişkenlerin birlikte etkileşimleri ile bağımlı değişken üzerinde önemli bilgi vermesi olasıdır. Tek başına önemsiz olan değişkenler birlikte etkileşimleri önemli olması söz konusu olabilir. Değişken sayısının çok olması durumunda en iyi modelin oluşturulmasında olası tüm kombinasyonların oluşturulması oldukça zaman almaktadır. Bu durumda Genetik Algoritma (GA) en iyi modeli oluşturulmasında farklı bir yaklaşım olarak kullanılabilir (Wasserman ve Sudjianto, 1994; Waller ve ark., 1996; Trevino ve Falciani, 2006).

Materyal

Bu çalışmada kullanılan veri seti, 2005 yılında Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Beden Eğitimi ve Spor Bölümü Özel Yetenek Sınavına giren 156 öğrenciden elde edilen bazı fiziksel özelliklerden oluşmaktadır. Sınav öncesi yapılan ölçümlere göre deneklerin yaş, boy, ağırlık, cinsiyet, ayak büyüklüğü, spor branşı, spor geçmişi, el kullanımı, sigara kullanımı, sınav öncesi mekik sayısı, sınava hazırlanma süresi, sınava kaçınıcı kez girdiği ve sınav mekik sayıları gibi özellikler elde edilmiştir. Elde edilen özellikler ve kodlanmış sonuçları aşağıdaki gibidir.

Tablo 1: Çalışmada kullanılan değişkenler ve kodlanmış sonuçları.

KOD	Değişken
S1	Yaş
S2	Boy
S3	Kilo
S4	Ayak No
S5	Spor Branşı
S6	Spor Geçmişi
S7	Eİ Kullanımı
S8	Sigara Kullanımı
S9	Sınav Öncesi Mekik Sayısı
S10	Sınav Hazırlanma Süresi
S11	Sınav Giriş Sayısı
S12	Sınav Mekik Sayısı

Sınav mekik sayısı (S12) bağımlı değişken, diğer özellikler ise bağımsız değişken olarak modellenmiştir. Uygun modelin oluşturulması için gerekli analizler R istatistik yazılımı kullanılmıştır.

Yöntem

Genetik Algoritmalar: Evrim teorisine göre kaynakların kısıtlı olduğu değişken bir ortamda, farklı genetik bilgilere sahip canlılar bunlar için yarışlar ve kabaca güçlü (FITNESS) olanlar hayatta kalıp çoğalarak/üreyerek genetik bilgilerini bir sonraki kuşağa (GENERATION) geçirme şansını yakalarlar (Goldberg, 1989). Genetik Algoritma, evrimsel hesaplama yöntemi kullanan yapay zeka uygulamalarından biri olarak nitelendirilebilir. Adından da anlaşılacağı üzere Darwin'in evrim teorisinden ilham alınarak modellenmiş olup, doğal seleksiyon ve doğal genetik mekanizmaya dayanan araştırma metodu kullanır. En iyi çözümü elde edebilmek için, zayıf çözümleri evrimsel bir işleyişe göre eleme yoluna gider. En iyi çözüm, yapılan çevrimler (iteration/yineleme) sonucunda hala hayatta kalabilmeyi başaran en iyi çözümdür (Goldberg, 1989; Kivijarvi ve ark., 2003).

GA, ya da daha geniş kapsamıyla Evrimsel Algoritmalar (EA), doğadaki evrimsel süreçleri model olarak kullanan bilgisayara dayalı problem çözme teknikleridir. Geleneksel programlama teknikleriyle çözülmesi güç olan, özellikle sınıflandırma ve çok boyutlu optimizasyon problemleri, bunların yardımıyla daha kolay ve hızlı çözülebilmektedir (Franti ve ark., 1997; Cowgill ve ark., 1999).

Genetik Algoritma aşamaları: Oluşturulan GA'nın kaba kodu aşağıdaki gibi yazılabilir (Hsiang ve Chong, 2008; Trevino ve Falciani, 2006),

1. Aşama: $t=0$
2. Aşama: ilk popülasyonu oluştur $P(t)$
3. Aşama: $P(t)$ popülasyonun uygunluk değerini hesapla
4. Aşama: $t=t+1$
5. Aşama: eğer bitirme kriteri sağlanıyorsa 10. adıma git.
6. Aşama: önceki popülasyondan $P(t)$ 'den $P(t)$ 'yi seç.
7. Aşama: $P(t)$ için Çaprazlama işlemleri
8. Aşama: $P(t)$ için mutasyon işlemleri
9. Aşama: 3. adıma git.
10. Aşama En iyi sonucu ekrana göster ve bitir.

İkili sayı gösterimi regresyonda bağımsız değişkenlerin kodlanmasında uygun bir yöntemdir. Eğer X_i ($i=1, \dots, n$), değişkeni modele dahil edilmişse 1, modele dahil edilmemesi durumunda ise 0 (sıfır) ile gösterilir. N değişken için model,

$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon \quad (1)$$

ile gösterilir. Burada y cevap değişkeni, X_i ise bağımsız değişken vektörüdür. β_i regresyon katsayılarıdır. Bu modelin 2'li sayı gösterimi modelde değişkenin yer alması durumunda 1, modelde yer almaması ise 0 ile gösterilir ve aşağıdaki gibi yazılabilir.

$$\begin{matrix} 1 & 0 & 1 & 1 & 0 & 1 & 1 & \dots & 1 \\ \text{1.bit} & \text{2.bit} & \text{3.bit} & \text{4.bit} & \text{5.bit} & \text{6.bit} & \text{7.bit} & \text{11.bit} \end{matrix}$$

Yukarıdaki ikili dizinin, çalışmada kullanılan gösterimi,

$$y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_6 X_6 + \beta_7 X_7 + \dots + \beta_{11} X_{11} + \varepsilon \quad (2)$$

biçiminde yazılabilir.

GA'lar bir başlangıç çözümler (kromozomlar) popülasyonu ile başlar. İlk popülasyon bizim başlangıç neslimizi (initial population) ifade eder. Şayet en başta elimizde varolan bir başlangıç nesli yoksa, popülasyon değerleri (kromozomlar) şansa bağlı üretilebilir. Daha sonra doğal seleksiyon sürecini başlatabilmek için çaprazlama ve mutasyon operatörleri kullanılarak yeni nesiller (popülasyonlar) türetilir. Popülasyondaki üyelerin her biri kodlanmış (binary) birer genetik sayı dizisi ile temsil edilir. Bu sayı dizilerine kromozom denir. Dizideki her haneye(0,1) gen denilir.

Oluşturulan tüm modeller binary (ikili) kod ile gösterilmektedir. Her bir modele kromozom denilmekte ve 11 genden oluşmaktadır. . Daha iyi çözümler için GA'nın seçim ölçütlerinin, yani doğal seleksiyonun yeterince iyi olmasına bağlıdır (Golberg, 1989). En uygun kromozomlarının seçilebilmesi için "uygunluğun" ne olduğunu iyi tanımlanmalıdır. Diğer bir ifadeyle bizi en iyiye (optimal) götürecek olan uygunluk fonksiyonun zayıf çözümleri eleyecek, kuvvetli çözümleri ise yaşatacak şekilde modellenmesi gerekmektedir.

GA'nın her çevriminde (iteration), elde edilen kromozomlar, uygunluk fonksiyonuna tabi tutularak uygunluk değerleri hesaplanır. Daha sonra yüksek uyuma sahip olanlar seçilir. (Şen, 2004). Bu çalışmada uygunluk fonksiyonu olarak eşitlik 3'teki fonksiyon kullanılmıştır.

$$F(m) = - \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i(m)}{1 - (q+1)/n} \right) \quad (3)$$

eşitlik 3'te $\hat{y}_i(m)$: m modelindeki i. gözlemin tahmin değerini, q: kurulan modeldeki değişken sayısını belirtir. (Zhu ve Chipman, 2006)

Genetik Algoritmanın Sonlandırılması: GA ile yapılan optimizasyon çözümü, istenilen hata sınırları dahilinde kalması veya önceden tespit edilen iterasyon sayısına ulaşması ile elde edilir. Belirli bir iterasyondan sonra kromozomlar ve bunların uygunluk değerleri birbirine yakın olur. Sonlandırma ölçütü olarak kromozomların uzunluğuna bağlı olan "entropy" ölçütü (Zhu ve Chipman, 2006),

$$entropy = -\frac{1}{2} \sum_{j=1}^p r_j \log_2(r_j) + (1-r_j) \log_2(1-r_j) \quad (4)$$

biçiminde yazılabilir. Eşitlik 4'te p: kromozom büyüklüğünü (kromozom uzunluğundaki bit sayısı), r_j : kromozomdaki bitlerden 1'lerin sayısını belirtir. Kromozomdaki bitlerin tümünün 1 veya 0 olması entropi değerinin 0 çıkmasına neden olur (Zhu ve Chipman, 2006)

Genetik Operatörler

Çaprazlama: Mevcut neslin seçilmiş kromozomlarından yeni bireyler oluşturmak üzere çaprazlama işlevi yürütülür. Ebeveynlerin çaprazlama oranı farklı olabilir. Çaprazlama operatöründen kromozomların iyi özellikleri birleştirilerek daha iyi kromozomlar oluşturması beklenir. Böylece arzu edilen en iyi bireye yani çözüme ulaşılır (Şen, 2004; Rachel ve ark., 2008).

Mutasyon: Kromozom üzerindeki bazı dizilerin yerleri ile oynayarak değişiklikler yapılmasıdır.

Tablo 2: Paralel birbirinden bağımsız 10 farklı genetik algoritma sonuçları

GA.No	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
1	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	1.00	1.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
4	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.05
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.05
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.95	1.00	0.00
8	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00

Tablo 2'de birbirinden bağımsız 10 farklı genetik sürecin 100 kez çalıştırılması sonucu bağımsız değişkenlerin popülasyonda bulunma frekansları verilmiştir. Örneğin 1'inci genetik süreç 100 kez çalıştırılmış ve S1, S2, S3, S4, S5, S6, S8, S11 değişkenleri hiç gözlenmezken, S7 değişkeni %5 oranında, S9 ve S10 değişkenleri %100 oranında gözlenmiştir. Her iterasyonda modelde S9 ve S10 değişkenleri bulunmamaktadır. Modele alınmayan değişkenlerin, değerlerinin birbirine çok yakın olmalarından dolayı bu değişkenlerin bağımlı değişken "sınav mekik sayısı" üzerinde etkilerinin olmadığı saptanmıştır.

Tablo 2'de görüldüğü gibi paralel çalıştırılan 10 farklı genetik süreçlerin sonuçlarına göre, S9 (Sınav Öncesi Mekik Sayısı) ve S10 (Sınav hazırlanma süresi) değişkenlerin modelde bulunması gereken değişkenler olduğu saptanmıştır.

Özellikle, problemin çözüm uzayı dar ise mutasyon oldukça önem taşır. Zira dar topluluklarda oluşan bir mutasyonun sonraki nesiller üzerindeki etkisi de yüksek olacaktır. Nesiller arası çeşitliliğin azalması durumunda mutasyon iyi bir etki yaratabilir. GA içindeki temel işlevi, öngörüğü zorlaştırırsa da nesiller arasındaki farklılaşmaya katkıda bulunarak sonuca daha hızlı ulaşmayı sağlamaktır. Kromozomlar üzerindeki mutasyon olasılığı genelde düşük bir değere sahiptir. Mutasyonlar, kromozom üzerindeki 1 değerini 0; 0 değerini ise 1 yaparlar (Hsiang ve Chornng, 2008; Rachel ve ark., 2008).

GA modeli olarak popülasyon boyutu 20 kromozomla ile başlanmıştır. Mutasyon oranı $1/m = 1/20 = 0.5$ dir.

Mutasyon değeri GA için önemli bir parametredir. Bu değerin yüksek olması yakınsama (konverge) işlemi uzatır.

Bulgular

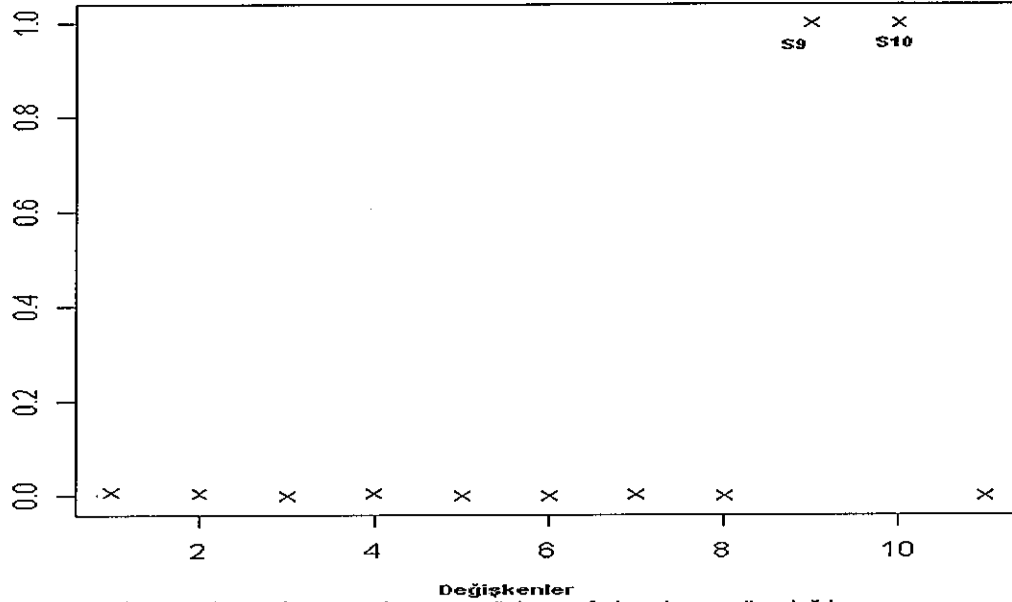
Genetik algoritmalar kullanılarak çözüme birbirinden bağımsız birden fazla yoldan ulaşılmaya çalışılmaktadır. Birden fazla başlangıç nesil ile başlanarak en iyi çözüm bulunur. Oysa çözüm uzayının tek genetik algoritma ile taranması durumunda, başlangıç parametrelerin (mutasyon, çaprazlama) dikkatli seçilmesi gerektirmektedir. Paralel genetik algoritmalar ile çözüme daha hızlı ulaşılır. Bu çalışmada birbirinden bağımsız paralel 10 farklı genetik algoritma ile çözüme gidilmiş ve elde edilen sonuçlar Tablo 2'de verilmiştir.

Tüm genetik algoritmalarda en iyi modelin kromozom yapısı 0 0 0 0 0 0 0 1 1 0 olarak bulunmuştur. Kromozomdaki genler regresyon katsayılarını belirtmektedir. Kromozom yapısına göre oluşturulabilecek en uygun model

$$y = \beta_9 X_9 + \beta_{10} X_{10} + \varepsilon$$

biçiminde yazılabilir. En iyi modelin kromozom yapısında görüldüğü gibi modelde "Sınav öncesi Mekik Sayısı" ve "Sınav hazırlanma süresi" değişkenlerin modelde olması beklenir.

Tüm değişkenlerin GA'nın 100 kez çalıştırılması sonucunda elde edilen değişkenlerin gözlenen frekans değerlerine göre grafiği Şekil 1'de verilmiştir.



Şekil 1. Değişkenlerin 100 iterasyon boyunca gözlenme frekanslarına göre dağılımı.

Şekil 1'de görüldüğü gibi tüm bağımsız değişkenler içerisinde, sınav öncesi mekik Sayısı (S9) ve Sınava hazırlanma süresi (S10) değişkenlerinin sınav mekik sayısı (bağımlı değişkeni) ile en iyi modeli oluşturmuştur. Dolayısıyla Sınav mekik sayısı üzerinde sınav öncesi mekik sayıları ve sınava hazırlanma süresi etkili olduğu saptanmıştır.

Stepwise regresyon yöntemine göre de en iyi modelin oluşturulması için analiz yapıldı. Oluşturulan her model için Akaike bilgi ölçütü değerleri Tablo 3'te verilmiştir. Regresyon modelleri için en küçük AIC değerine sahip olan model en uygun model olarak kabul edilmektedir.

Tablo 3: Stepwise yöntemine göre oluşturulan modeller ve AIC değerleri

Model	AIC
Y ~ S1 + S2 + S3 + S4 + S5 + S6 + S7 + S8 + S9 + S10 + S11	860.62
Y ~ S2 + S3 + S4 + S5 + S6 + S7 + S8 + S9 + S10 + S11	858.62
Y ~ S2 + S3 + S4 + S5 + S6 + S7 + S9 + S10 + S11	856.66
Y ~ S2 + S3 + S4 + S5 + S7 + S9 + S10 + S11	854.98
Y ~ S2 + S3 + S5 + S7 + S9 + S10 + S11	853.73
Y ~ S3 + S5 + S7 + S9 + S10 + S11	852.34
Y ~ S5 + S7 + S9 + S10 + S11	851.10
Y ~ S5 + S9 + S10 + S11	849.82
Y ~ S9 + S10 + S11	848.47
Y ~ S9 + S10	847.21

Tablo 3'te görüldüğü gibi en küçük AIC değeri 847.21 dir. Dolayısıyla en uygun modelin $Y \sim S9 + S10$ olduğu saptanmıştır.

Sonuç

Genetik algoritma geleneksel yöntemlerle çözümü zor veya imkansız olan problemlerin çözümünde kullanılmaktadır. Çok genel anlamda genetik algoritmanın üç uygulama alanı bulunmaktadır (Franti, 1997; Şen, 2004). Bunlar deneysel çalışmalarda optimizasyon, pratik endüstriyel uygulamalar ve sınıflandırma sistemleridir. Bu problemlerin hemen hemen hepsi çok geniş bir çözüm alanının taranmasını gerektirmektedir. Bu çözüm alanının geleneksel yöntemlerle taranması çok uzun sürmekte, genetik algoritmayla ise kısa bir sürede kabul edilebilir bir sonuç alınabilmektedir. Genetik Algoritma, problemin ele alındığı ortamda (doğada) yer alan uygun ve güçlü çözümlerin (bireylerin) yaşatılarak, iyiler arasından daha da iyileştirilmiş çözümler üretmeyi amaç eden bir yöntemdir (Golberg, 1989;

Şen, 2004). Ancak her problem için en optimum sonucu bulduğunu garanti edilemez (Şen, 2004). GA'nın işleyişinde en temel unsur "uygunluğun" (fitness) ne olduğunun tarifinde yatmaktadır. Matematiksel anlamda bu modellenmesi ve muhtemel her çözüm üzerinde uygulanması gereken bir fonksiyondur.

Biyolojik evrim teorisinden esinlenerek ortaya konulmuş GA yöntemi, çözüm alanını rasgele biçimde ele alarak en iyi çözümü arayan bir yöntemdir. GA, çözüme ulaşmak için önce karar değişkeni uzayında rasgele olarak noktalar topluluğu ele alır, daha sonra gösterilecek kuralların ışığı altında bu noktalar arasında eşleştirmeler yapılarak kitlenin bazı üyeleri yok olurken onların yerine yenileri gelir (Golberg, 1989). Yeni gelenlerin kitleye katılması ile daha sağlıklı ve çözüme yakın olması sağlanmış olur. Kitlenin üyeleri arasında gerekli genetik işlemler yapıldıktan sonra daha zinde ve hedefe yakın yeni bir model elde edilir.

Bu çalışmada çoklu doğrusal regresyonda, uygun modelin (değişken seçiminde) elde edilmesinde genetik algoritma yaklaşımı kullanılmıştır. GA bağımsız değişken sayısına bakmaksızın en iyi çözümü bulmaya çalışmaktadır. Değişken sayısında herhangi bir kısıtlama olmadığından dolayı GA'nın kullanılmasının uygun olduğu saptanmıştır.

Yapılan denemelerde, işlemin 100 iterasyon sonucunda yakınsadığı ve her denemede uygun bağımsız değişkenler olarak "Sınav Öncesi Mekik Sayısı (S9)" ve "Sınav Hazırlanma Süresi (S10)" yer almıştır. Yani sınav mekik sayısı üzerinde sınav öncesi mekik sayısı ve sınav hazırlanma sürelerinin etkili olduğu görülmüştür. Bunun yanı sıra Stepwise yöntemi kullanılarak en iyi modelin seçimi yapılmıştır. Elde edilen sonuçlar bakımından GA ve Stepwise yöntemlerinin benzer sonuçlar verdiği saptanmıştır.

Kaynaklar

- Andrew S., Dan K., 2002. Genetic Algorithm search for large logistic regression models with significant variables. 22. Int. Conf. information Technology Interfaces, June 13-16, 2000, Pula, Croatia
- Cowgill, M., 1999. A Genetic Algorithm Approach to Cluster Analysis. An international computer and mathematics with application. 37: 99-108.
- Draper, N. R., Smith, H., 1989. Applied regression analysis, John Wiley & Sons, New York.
- Franti, P., and at all, 1997. Genetic Algorithms for large scale clustering problems. The Computer Journal. 40:547-554.
- Goldberg, D., 1989. Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading MA
- Hsiang H.L., Chomg-S.O., 2008. Variable selection in clustering for marketing segmentation using genetic algorithms. Expert Systems with Applications 34: 502-510
- Joseph, M., 2004. A Useful Tool for statistical estimation: Genetic Algorithms. Journal of Statistical and Simulation, 75: 37-251.
- Kivijarvi, J., 2003. Self-Adaptive Genetic Algorithm for Clustering. Journal of Heuristics. 9: 113-129.
- Miller, A. J., 1990. Subset selection in regression, London: Chapman and Hall.
- Rachel C. ve ark., 2008. Genetic algorithms for simultaneous variable and sample selection in metabonomics. Bioinformatics. 25: 112-118.
- Şen, Z., 2004. Genetik Algoritmalar ve En iyileme yöntemleri. Su Vakfı yayınları. İstanbul.
- Trevino, V., Falciani, F., 2006. Galgo an R package for multivariate variable selection using genetic algorithms. Bioinformatics. 22: 1154-1156.
- Wallet, B. C., Marchette, D. J., Solka, J. L. And Wegman, E. J., 1996. A genetic algorithm for best subset selection in linear regression, Proceedings of the 28th Symposium on the Interface.
- Wasserman, G. S. and Sudjianto, A., 1994. All subsets regression using a genetic algorithm, Computers and Industrial Engineering, 27(1): 489-49
- Yeniay Ö., Göktaş A., 2003, Doğrusal Regresyonda En İyi Altküme Seçimine Genetik Algoritma Yaklaşımı . G.Ü. Fen Bilimleri Dergisi. 16: 37-45.

Zhu M, Chipman HA., 2006. Darwinian evolution in parallel universes: A parallel genetic algorithm for variable selection. Technometrics, 48: 491-502.