G.Ü. Fen Bilimleri Dergisi,
16(1):37-45, 2003

ISSN 1303-9709

G.Ü. Journal of Science
16(1):37-45, 2003

DOĞRUSAL REGRESYONDA EN İYİ ALTKÜME SEÇİMİNE GENETİK ALGORİTMA YAKLAŞIMI

Özgür YENİAY*, Atilla GÖKTAŞ

Hacettepe Üniversitesi, İstatistik Bölümü, Ankara, TÜRKİYE

ÖZET

Çok sayıda bağımsız değişken ve bir bağımlı değişkenden oluşan veri seti verildiğinde, bağımlı değişkeni kestiren en iyi modelin seçilmesi "değişken seçimi" ya da "en iyi altküme modelinin seçimi" olarak bilinmektedir. Değişken seçimi için çok sayıda yöntem önerilmiştir. Maalesef, bağımsız değişkenler arasındaki ilişki yüksek olduğunda mevcut yöntemler çoğu kez başarısız olmaktadır. Ayrıca, bağımsız değişken sayısı arttığında olası altküme sayısı üstel olarak arttığından, tüm olası altküme yönteminin büyük boyutlu veri setlerini ele alma güçlüğü bulunmaktadır. Bu çalışmada, doğrusal regresyonda değişken seçimi için Genetik Algoritmaya (GA) dayalı yeni bir stokastik optimizasyon yöntemi önerilmektedir. Önerilen yöntemin ve klasik değişken seçim yöntemlerinin performansı literatürde yaygın olarak kullanılan veri setleri kullanılarak karşılaştırılmaktadır.

Anahtar Kelimeler:Doğrusal regresyon, değişken seçimi, stokastik optimizasyon, genetik algoritma.

A GENETIC ALGORITHM APPROACH FOR THE BEST SUBSET SELECTION IN LINEAR REGRESSION

ABSTRACT

When a data set including many explanatory variables and a response variable is given, the choice of best model which predicts the response variable is known as "variable selection" or "the selection of the best subset model". Many methods for variable selection have been suggested. Unfortunately, when the correlation between explanatory variables is high, currently used methods are mostly unsuccesful. Also, as the number of possible subsets grows exponentially when the number of explanatory variables increase, all possible subset methods have difficulty handling large dimensional data sets. In this study, a new stochastic optimization method based on Genetic Algorithm (GA) is proposed for variable selection in linear regression. The performance of the method proposed and that of classical variable selection methods are compared by using data sets commonly given in literature.

Key Words: Linear regression, variable selection, stochastic optimization, genetic algorithm.

1. GİRİŞ

İncelenen sistem hakkında önsel bilgi bulunmadığında, yeterli bir model kurabilmek için kaç değişkenin ölçülmesi gerektiğine karar verebilmek çoğu kez zordur. Modelleyici hangi değişkenlerin önemli olduğundan emin olmadığında, çoğu zaman çok sayıda değişken ölçülmektedir. Tüm mevcut değişkenleri içeren bir model kurmak çeşitli veri analizi problemleri yaratacağından, modelleme için yanlış bir yol olacaktır. Bu nedenle, bağımlı değişkeni en iyi şekilde açıklayabilecek model denklemini kullanım amacına uygun olarak elde etmek için mevcut bağımsız değişkenler arasında seçim yapmak regresyon analizinin en önemli aşamalarından biridir (1-3). Değişken seçimi yapmanın nedenleri,

a) Verinin toplanacağı değişken sayısını azaltarak daha düşük maliyetle tahmin ya da kestirim yapmak,

b) Modele katkısı önemsiz değişkenleri çıkararak daha doğru kestirim yapmak,

c) Bağımsız değişkenlerden bazıları yüksek derecede ilişkili olduklarında, regresyon katsayılarını daha küçük standart hata ile tahmin etmektir (4).

Değişken seçimi için çok sayıda yöntem önerilmiştir. Draper ve Smith (5); Montgomery ve Peck (6) bu yöntemleri ayrıntılı olarak ele almaktadır. Uygulamada en yaygın kullanılanları, adımsal yöntemler (ileriye doğru seçim, geriye doğru çıkarma ve adımsal regresyon) ve tüm olası altküme yöntemidir. Adımsal yöntemlerde, bir durdurma kuralı sağlanana kadar her adımda modele bir değişken eklenir ya da çıkartılır. Bu yöntemlerde her değişken bağımsız olarak incelenir ve değişken etkileşimleri hesaba katılmaz. Oysa x_i ve x_j değişkenlerin ayrı ayrı önemsiz iken, birlikte faydalı bilgi verebilmeleri olasıdır. Bu nedenle, bu yöntemlerle mevcut arama uzayının küçük bir bölümü incelenir ve optimal olmayan çözümlere ulaşılır. Berk (7), tek başına önemsiz olmasına karşın birlikte önemli duruma gelen değişkenlerin varlığında adımsal yöntemler ile tüm olası altküme yöntemi arasındaki farkın önemine dikkat çekmektedir. Berk (7), 4 bağımsız değişkenin olduğu ve 1, 2 ve 3 değişkenli en iyi altkümelerin sırasıyla (x_1), (x_2,x_3) ve (x_1,x_2,x_4) olduğu basit bir örnek ile durumu açıklamıştır. Bu örnekte adımsal regresyon optimal olmayan (x_1,x_2) çözümüne ulaştırmaktadır.

Tüm olası altküme yöntemi, belirlenen seçim ölçütüne göre en iyi altkümenin bulunmasını garanti eder. Olası her altkümeyi incelemenin kombinatoryal yapısının getireceği aşırı işlemsel maliyetten dolayı kullanımları kısıtlıdır. Örneğin k=100 bağımsız değişkenin bulunduğu bir veri setinde 10 bağımsız değişkenli tüm olası altkümeler incelenmek istenildiğinde, toplam 1,7x1013 altküme mevcuttur. Tüm olası altkümeler içinse, bir veya daha fazla bağımsız değişkenli toplam 2100-1 olası altküme olacaktır. Aday değişkenlerin sayısı arttığında olası altkümelerin sayısı üstel olarak arttığından tüm olası altküme yöntemini kullanarak büyük boyutlu veriler ile uğraşmak güçleşmektedir. Bu nedenle tüm olası altküme yöntemi için, istatistiksel paket programlardan SAS'ta 25, S-Plus'ta 31 ve MINITAB'ta 21 değişken sınırlaması bulunmaktadır.

2. GENETİK ALGORİTMALARA GENEL BİR BAKIŞ

Genetik Algoritmalar, doğadaki evrim mekanizmasını örnek alan arama yöntemidir. 1970'lerin başında John Holland tarafından ortaya atılmıştır. Doğada geçerli olan en iyinin yaşaması (survival of the fittest) kuralına dayanarak sürekli iyileşen çözümler üretirler (8). Herhangi bir problem GA ile çözülecekse, aşağıdaki temel adımlar izlenmelidir:

1. Problemin çözümleri uygun bir biçimde kodlanır. Kodlamanın çeşitli yolları olmasına karşın (Gray kodlama, kayan nokta vb.) ikili kodlama yaygın olarak kullanılmaktadır.

2. Problemin büyüklüğüne bağlı olarak N tane kodlanmış çözümden oluşan bir çözüm grubu oluşturulur (Çözüm grubu biyolojideki benzerliği nedeniyle popülasyon, çözüm kodları da kromozom ya da dizi olarak adlandırılır).

3. Popülasyondaki her bir dizinin uyum değeri hesaplanır. Hesaplanan uyum değeri, dizinin (çözüm) ne kadar iyi olduğunun bir göstergesidir. Bu hesaplamada kullanılan fonksiyona uyum fonksiyonu adı verilir. Uyum fonksiyonu, diziyi problemin parametreleri haline dönüştürerek onların bir bakıma

şifresini çözer (decoding). Sonra, bu parametrelere göre hesaplamayı yaparak dizilerin uygunluğunu bulur.

4. Popülasyona yeni diziler kazandırabilmek, diğer bir ifadeyle arama uzayındaki farklı çözümlere ulaşabilmek için seçme, çaprazlama ve mutasyon operatörleri kullanılır.

5. Yeni popülasyondaki dizilerin uyum değerleri yeniden hesaplanır ve 4. ve 5. adımlar belirli bir durdurma kriteri sağlanana kadar tekrarlanır. Örneğin, belirli bir uyum değerinin bulunması, algoritmanın belirlenen iterasyon sayısına ulaşması ya da popülasyondaki dizilerin benzer yapıya yakınsaması GA'da durdurma kriteri olarak kullanılabilir.

3. ALTKÜME SEÇİMİNDE GENETİK ALGORİTMA KULLANIMI

İkili dizi gösterimi regresyonda değişken altkümelerini kodlamanın uygun bir yoludur. Bu gösterimde i. değişkenin altkümede yer aldığını göstermek için dizinin i. konumuna 1 değeri, yer almadığını göstermek için 0 değeri kullanılır. k aday değişkene sahip regresyon problemi için tam model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1} + \beta_k x_k + \varepsilon \qquad [31]$$

biçiminde ifade edilebilir. Burada y bağımlı değişkeni, x1,x2,…,xk bağımsız değişkenleri, 0, 1,…, k regresyon katsayılarını ve normal dağılımlı, ortalaması sıfır ve varyansı 2 olan hata terimini göstermektedir. Bu modelin ikili dizi gösteriminde, k konumun herbirinde bit değeri 1'dir.

0	*1*	*1*	...	*1*	*0*	*1*
1.bit	2.bit	3.bit		(k-2).bit	(k-1).bit	k.bit

ikili dizi gösterimi ise,

$$y = \beta_0 + \beta_2 x_2 + \ldots + \beta_{k-2} x_{k-2} + \beta_k x_k + \varepsilon$$

indirgenmiş modelin gösterimidir (9,10).

Regresyon modellerini (ya da popülasyondaki dizileri) değerlendirebilmek için R2, düzeltilmiş R2, Cp, artık kareler ortalaması (AKO) ya da AIC gibi seçim kriterlerinden herhangi biri GA'da uyum fonksiyonu olarak kullanılabilir. Her altkümenin (dizinin) uyum değerinin hesaplanmasının ardından, orijinal popülasyonda sahip oldukları uyum değerlerine orantılı seçim işlemi yapılır. Bunun sonucunda daha iyi uyum değerine sahip diziler daha yüksek seçilme olasılığına sahiptir. Bu yolla iki dizi seçilir ve seçilen diziler [1,k-1] arasından rasgele belirlenen bir noktadan çaprazlanarak, diziler arasında bit bloklarının yer değiştirmesi sağlanır. Burada amaç, çaprazlanan dizilerin iyi özelliklerinin biraraya gelmesini kolaylaştırarak daha iyi dizilere ulaşmayı sağlamaktır. Çaprazlama operatörü yardımıyla elde edilen yeni diziler (yavrular) elde edildikleri diziler (ataları) ile aynı dizi gösterimine sahip iseler, mutasyon operatörü kullanılır. Bu operatör dizideki bitlerden bazılarının rasgele olarak değişimine yol açar (1 iken 0 ya da 0 iken 1). Bu aşamada elde edilen yeni diziler, popülasyondaki en kötü uyum değerine sahip iki dizinin yerini alır ve yeni bir popülasyon elde edilir. Bu nedenle popülasyon büyüklüğünde bir değişiklik meydana gelmez. Elde edilen yeni popülasyonda uyum değerlendirmesi, seçme, çaprazlama ve mutasyon işlemleri algoritmanın durdurma kriteri sağlanana kadar tekrarlanır.

GA'nın bir iterasyonunu açıklayabilmek amacıyla k=7 aday değişkene sahip regresyon problemini ele alalım. Dizi gösterimleri ve Cp değerleri verilen N=4 altküme, başlangıç popülasyonunu meydana getirsin.

1 0 1 0 1 0 1	$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \varepsilon$	C_p=42
1 0 0 0 1 1 0	$y = \beta_0 + \beta_1 x_1 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$	C_p=40
1 1 0 0 1 0 1	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \beta_7 x_7 + \varepsilon$	C_p=38
1 0 0 1 0 1 0	$y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \beta_6 x_6 + \varepsilon$	C_p=36

Seçim sürecinde en düşük Cp değerlerine (uyum değerleri) sahip olan

1 1 0 0 1 0 1 ve 1 0 0 1 0 1 0 dizilerinin seçildiğini ve dizilerin rasgele belirlenen 3. noktadan çaprazlandığını düşünelim:

1 1 0/0 1 0 1
1 0 0/1 0 1 0

Bu işlem sonucunda,

1 1 0 1 0 1 0	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_6 x_6 + \varepsilon$	C_p=39
1 0 0 0 1 0 1	$y = \beta_0 + \beta_1 x_1 + \beta_5 x_5 + \beta_7 x_7 + \varepsilon$	C_p=32

dizileri meydana gelir. Meydana gelen bu iki dizi, popülasyondaki en kötü uyum değerlerine sahip (Cp=42 ve Cp=40) iki dizinin yerini aldığında aşağıdaki yeni popülasyon elde edilir:

1 1 0 1 0 1 0	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_6 x_6 + \varepsilon$	C_p=39
1 0 0 0 1 0 1	$y = \beta_0 + \beta_1 x_1 + \beta_5 x_5 + \beta_7 x_7 + \varepsilon$	C_p=32
1 1 0 0 1 0 1	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \beta_7 x_7 + \varepsilon$	C_p=38
1 0 0 1 0 1 0	$y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \beta_6 x_6 + \varepsilon$	C_p=36

Algoritmanın ikinci iterasyonunda 1 0 0 0 1 0 1 ve 1 0 0 1 0 1 0 dizilerinin çaprazlama için seçilme şansı daha yüksektir. Eğer bu diziler seçilir ve rasgele olarak belirlenen 3. noktadan çaprazlanırsa aşağıdaki gibi bir sonuç ortaya çıkabilir:

1 0 0/0 1 0 1
1 0 0/1 0 1 0

Bu çaprazlama sonucunda yeni genetik kombinasyonların bulunamayacağı ve sonucunda popülasyon kalitesinde bir iyileşme olmayacağı açıktır. Böyle bir durumda mutasyon operatörü dizideki bazı bitleri rasgele değiştirerek sorun giderilir.

Seçme, çaprazlama, mutasyon, uyum hesaplama ve dizi değiştirme adımları GA'ya arama uzayında serbest hareket edebilme, optimal çözümün yer almadığı bölgelerden uzak dururken muhtemelen optimal çözümün bulunduğu bölgeleri arayabilme olanağı sağlar. GA optimali içeren bir bölge bulduğunda bu bölgedeki en iyi noktayı yerel olarak aramaktadır.

4. UYGULAMA

Çalışmamızda GA'ya dayalı seçim yöntemi, adımsal yöntemler ve en iyi altküme yönteminin en iyi altküme modellerine ulaşma performanslarını karşılaştırabilmek için, altküme seçim çalışmalarında karşılaştırma amaçlı yaygın olarak kullanılan 4 veri seti seçilmiştir. Kullanılan herbir veri setine ilişkin bilgiler Çizelge 1'de özetlenmiştir.

Çizelge 1. Karşılaştırmalarda kullanılan veri setleri

Veri Seti	Gözlem Sayısı	Bağımsız Değişken Sayısı	Kaynak
Steam	25	9	(5)
Longley	16	6	(11)
Detroit	13	11	(12)
Gasoline	32	11	(13)

GA ile değişken seçimi yaparken MATLAB 6.2'de çalışabilen ticari PLS-Toolbox 2.1 (14) kullanılmıştır. Her veri seti için N=30 altkümeden oluşan bir popülasyon rasgele yaratılmış ve GA 100 iterasyon işletilmiştir. Her model genişliği için GA, ileriye doğru seçim, geriye doğru çıkarma, adımsal regresyon ve tüm olası altküme yöntemleri ile elde edilen altkümeler ve bu altkümelere karşılık gelen Cp, düzeltilmiş R_2 ve AKO Çizelge 2, Çizelge 3, Çizelge 4 ve Çizelge 5'de verilmiştir.

Çizelge 2. Steam verisi için yöntemlerin preformansı

Değişken Sayısı	Yöntem	Seçilen Değişkenler	Cp	Düzeltilmiş R^2	AKO
1	İleriye doğru seçim	X7	35.1	0.702	0.792100
	Geriye doğru çıkarma	X7	35.1	0.702	0.792100
	Adımsal regresyon	X7	35.1	0.702	0.792100
	En iyi altküme	X7	35.1	0.702	0.792100
	Genetik algoritma	X7	35.1	0.702	0.792100
2	İleriye doğru seçim	X7 X1	8.5	0.847	0.405769
	Geriye doğru çıkarma	X7 X1	8.5	0.847	0.405769
	Adımsal regresyon	X7 X1	8.5	0.847	0.405769
	En iyi altküme	X1 X7	8.5	0.847	0.405769
	Genetik algoritma	X1 X7	8.5	0.847	0.405769
3	İleriye doğru seçim	X7 X1 X5	6.7	0.862	0.366025
	Geriye doğru çıkarma	X7 X1 X5	6.7	0.862	0.366025
	Adımsal regresyon	X7 X1 X5	6.7	0.862	0.366025
	En iyi altküme	X4 X5 X7	5.6	0.868	0.349281
	Genetik algoritma	X4 X5 X7	5.6	0.868	0.349281
4	İleriye doğru seçim	X7 X1 X5 X4	6	0.872	0.339889
	Geriye doğru çıkarma	X9 X5 X1 X7	6.4	0.870	0.346921
	Adımsal regresyon	X7 X1 X5 X4	6	0.872	0.339889
	En iyi altküme	X1 X4 X5 X7	6	0.872	0.339889
	Genetik algoritma	X1 X4 X5 X7	6	0.872	0.339889
5	İleriye doğru seçim	X7 X1 X5 X4 X9	6.9	0.872	0.339889
	Geriye doğru çıkarma	X9 X8 X5 X1 X7	7.2	0.870	0.344569
	Adımsal regresyon	X7 X1 X5 X4 X9	6.9	0.872	0.339889
	En iyi altküme	X1 X2 X5 X7 X9	6.7	0.873	0.337561
	Genetik algoritma	X1 X2 X5 X7 X9	6.7	0.873	0.337561
6	İleriye doğru seçim	X7 X1 X5 X4 X9 X2	7.9	0.872	0.341056
	Geriye doğru çıkarma	X9 X8 X5 X1 X7 X3	5.4	0.889	0.295936
	Adımsal regresyon	X7 X1 X5 X4 X9 X2	7.9	0.872	0.341056
	En iyi altküme	X1 X3 X5 X7 X8 X9	5.4	0.889	0.295936
	Genetik algoritma	X1 X3 X5 X7 X8 X9	5.4	0.889	0.295936
7	İleriye doğru seçim	X7 X1 X5 X4 X9 X2 X8	8.87	0.872	0.341056
	Geriye doğru çıkarma	X9 X8 X4 X5 X1 X7 X3	6.7	0.887	0.300304
	Adımsal regresyon	X7 X1 X5 X4 X9 X2 X8	8.87	0.872	0.341056
	En iyi altküme	X1 X3 X4 X5 X7 X8 X9	6.7	0.887	0.300304
	Genetik algoritma	X1 X3 X4 X5 X7 X8 X9	6.7	0.887	0.300304
8	İleriye doğru seçim	X7 X1 X5 X4 X9 X2 X8 X3	8.6	0.881	0.315844
	Geriye doğru çıkarma	X9 X8 X4 X6 X5 X1 X7 X3	8.2	0.884	0.308025
	Adımsal regresyon	X7 X1 X5 X4 X9 X2 X8 X3	8.6	0.881	0.315844
	En iyi altküme	X1 X3 X4 X5 X6 X7 X8 X9	8.2	0.884	0.308025
	Genetik algoritma	X1 X3 X4 X5 X6 X7 X8 X9	8.2	0.884	0.308025
9	İleriye doğru seçim	X7 X1 X5 X4 X9 X2 X8 X3 X6	10	0.878	0.324900
	Geriye doğru çıkarma	X7 X1 X5 X4 X9 X2 X8 X3 X6	10	0.878	0.324900
	Adımsal regresyon	X7 X1 X5 X4 X9 X2 X8 X3 X6	10	0.878	0.324900
	En iyi altküme	X1 X2 X3 X4 X5 X6 X7 X8 X9	10	0.878	0.324900
	Genetik algoritma	X1 X2 X3 X4 X5 X6 X7 X8 X9	10	0.878	0.324900

Çizelge 3. Longley verisi için yöntemlerin performansı

Değişken Sayısı	Yöntem	Seçilen Değişkenler	Cp	Düzeltilmiş R^2	AKO
1	İleriye doğru seçim	X2	52.9	0.965	431150
	Geriye doğru çıkarma	X2	52.9	0.965	431150
	Adımsal regresyon	X2	52.9	0.965	431150
	En iyi altküme	X2	52.9	0.965	431150
	Genetik algoritma	X2	52.9	0.965	431150
2	İleriye doğru seçim	X2 X3	28.5	0.978	275310
	Geriye doğru çıkarma	X3 X6	25.2	0.98	251703
	Adımsal regresyon	X2 X3	28.5	0.978	275310
	En iyi altküme	X3 X6	25.2	0.98	251703
	Genetik algoritma	X3 X6	25.2	0.98	251703
3	İleriye doğru seçim	X2 X3 X4	21.7	0.981	229728
	Geriye doğru çıkarma	X3 X4 X6	6.2	0.991	110277
	Adımsal regresyon	X2 X3 X4	21.7	0.981	229728
	En iyi altküme	X3 X4 X6	6.2	0.991	110277
	Genetik algoritma	X3 X4 X6	6.2	0.991	110277
4	İleriye doğru seçim	X2 X3 X4 X6	3.2	0.994	78064
	Geriye doğru çıkarma	X2 X3 X4 X6	3.2	0.994	78064
	Adımsal regresyon	X2 X3 X4 X6	3.2	0.994	78064
	En iyi altküme	X2 X3 X4 X6	3.2	0.994	78064
	Genetik algoritma	X2 X3 X4 X6	3.2	0.994	78064
5	İleriye doğru seçim	X2 X3 X4 X5 X6	5	0.993	83938
	Geriye doğru çıkarma	X2 X3 X4 X5 X6	5	0.993	83938
	Adımsal regresyon	X2 X3 X4 X5 X6	5	0.993	83938
	En iyi altküme	X2 X3 X4 X5 X6	5	0.993	83938
	Genetik algoritma	X2 X3 X4 X5 X6	5	0.993	83938
6	İleriye doğru seçim	X1 X2 X3 X4 X5 X6	7	0.992	92934
	Geriye doğru çıkarma	X1 X2 X3 X4 X5 X6	7	0.992	92934
	Adımsal regresyon	X1 X2 X3 X4 X5 X6	7	0.992	92934
	En iyi altküme	X1 X2 X3 X4 X5 X6	7	0.992	92934
	Genetik algoritma	X1 X2 X3 X4 X5 X6	7	0.992	92934

Çizelge 4. Detroit verisi için yöntemlerin performansı

Değişken Sayısı	Yöntem	Seçilen Değişkenler	Cp	Düzeltilmiş R^2	AKO
1	İleriye doğru seçim	X6	1889.7	0.931	18.619
	Geriye doğru çıkarma	X4	13734.6	0.488	137.429
	Adımsal regresyon	X6	1889.7	0.931	18.619
	En iyi altküme	X6	1889.7	0.931	18.619
	Genetik algoritma	X6	1889.7	0.931	18.619
2	İleriye doğru seçim	X4 X6	304	0.988	3.356
	Geriye doğru çıkarma	X4 X3	1395.79	0.441	13.958
	Adımsal regresyon	X4 X6	304	0.988	3.356
	En iyi altküme	X4 X6	304	0.988	3.356
	Genetik algoritma	X4 X6	304	0.988	3.356
3	İleriye doğru seçim	X6 X4 X10	196.6	0.991	2.415
	Geriye doğru çıkarma	X11 X4 X3	210.2	0.990	2.579
	Adımsal regresyon	X6 X4 X10	196.6	0.991	2.415
	En iyi altküme	X2 X4 X11	58.2	0.997	0.757
	Genetik algoritma	X2 X4 X11	58.2	0.997	0.757
4	İleriye doğru seçim	X6 X4 X10 X1	299.58.9	0.994	1.698
	Geriye doğru çıkarma	X11 X4 X3 X8	98.13	0.995	2.579
	Adımsal regresyon	X6 X4 X10 X1	299.58.9	0.994	1.698
	En iyi altküme	X2 X4 X6 X11	32.4	0.998	0.477
	Genetik algoritma	X2 X4 X6 X11	32.4	0.998	0.477
5	İleriye doğru seçim	X6 X4 X10 X1 X2	79.68	0.995	1.241
	Geriye doğru çıkarma	X11 X4 X3 X8 X7	83.77	0.995	1.302
	Adımsal regresyon	X6 X4 X10 X1 X2	79.68	0.995	1.241
	En iyi altküme	X1 X2 X4 X6 X11	22.8	0.999	0.366
	Genetik algoritma	X1 X2 X4 X6 X11	22.8	0.999	0.366
6	İleriye doğru seçim	X6 X4 X10 X1 X2 X11	23.59	0.998	0.406
	Geriye doğru çıkarma	X11 X4 X3 X9 X8 X7	69.1	0.995	1.228
	Adımsal regresyon	X6 X4 X10 X1 X2 X11	23.59	0.998	0.406
	En iyi altküme	X1 X2 X4 X6 X7 X11	14	0.999	0.233
	Genetik algoritma	X1 X2 X4 X6 X7 X11	14	0.999	0.233

7	**İleriye doğru seçim**	X6 X4 X10 X1 X2 X11 X7	15.5	0.999	0.269
	Geriye doğru çıkarma	X11 X4 X3 X9 X10 X8 X7	44.83	0.997	0.899
	Adımsal regresyon	X6 X4 X10 X1 X2 X11 X7	15.5	0.999	0.269
	En iyi altküme	X1 X2 X3 X4 X6 X8 X11	13.6	0.999	0.228
	Genetik algoritma	X1 X2 X3 X4 X6 X8 X11	13.6	0.999	0.228
8	**İleriye doğru seçim**	X6 X4 X10 X1 X2 X11 X7 X9	16.99	0.999	0.323
	Geriye doğru çıkarma	X11 X4 X5 X3 X9 X10 X8 X7	19.41	0.999	0.388
	Adımsal regresyon	X6 X4 X10 X1 X2 X11 X7 X9	16.99	0.999	0.323
	En iyi altküme	X1 X2 X3 X4 X6 X8 X10 X11	13.7	0.999	0.235
	Genetik algoritma	X1 X2 X3 X4 X6 X8 X10 X11	13.7	0.999	0.235
9	**İleriye doğru seçim**	X6 X4 X10 X1 X2 X11 X7 X9 X3	17.32	0.999	0.371
	Geriye doğru çıkarma	X11 X2 X4 X5 X3 X9 X10 X8 X7	14.4	0.999	0.266
	Adımsal regresyon	X6 X4 X10 X1 X2 X11 X7 X9 X3	17.32	0.999	0.371
	En iyi altküme	X1 X2 X3 X4 X6 X8 X9 X10 X11	14.3	0.999	0.261
	Genetik algoritma	X1 X2 X3 X4 X6 X8 X9 X10 X11	14.3	0.999	0.261
10	**İleriye doğru seçim**	X6 X4 X10 X1 X2 X11 X7 X9 X3 X8	15.4	0.999	0.345
	Geriye doğru çıkarma	X1 X2 X3 X4 X5 X7 X8 X9 X10 X11	11.6	0.999	0.141
	Adımsal regresyon	X6 X4 X10 X1 X2 X11 X7 X9 X3 X8	15.4	0.999	0.345
	En iyi altküme	X1 X2 X3 X4 X5 X7 X8 X9 X10 X11	11.6	0.999	0.141
	Genetik algoritma	X1 X2 X3 X4 X5 X7 X8 X9 X10 X11	11.6	0.999	0.141
11	**İleriye doğru seçim**	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	1	0.108
	Geriye doğru çıkarma	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	1	0.108
	Adımsal regresyon	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	1	0.108
	En iyi altküme	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	1	0.108
	Genetik algoritma	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	1	0.108

Çizelge 5. Gasoline verisi için yöntemlerin performansı

Değişken Sayısı	Yöntem	Seçilen Değişkenler	Cp	Düzeltilmiş R^2	AKO
1	**İleriye doğru seçim**	X1	0.2	0.752	9.7469
	Geriye doğru çıkarma	X10	3.8	0.718	11.0889
	Adımsal regresyon	X1	0.2	0.752	9.7469
	En iyi altküme	X1	0.2	0.752	9.7469
	Genetik algoritma	X1	0.2	0.752	9.7469
2	**İleriye doğru seçim**	X1 X4	0.5	0.760	9.4433
	Geriye doğru çıkarma	X8 X10	1.0	0.755	9.6348
	Adımsal regresyon	X1 X4	0.5	0.760	9.4433
	En iyi altküme	X1 X4	0.5	0.76	9.4433
	Genetik algoritma	X1 X4	0.5	0.76	9.4433
3	**İleriye doğru seçim**	X1 X4 X9	1.6	0.759	9.4679
	Geriye doğru çıkarma	X5 X8 X10	-0.5	0.781	8.6084
	Adımsal regresyon	X1 X4 X9	1.6	0.759	9.4679
	En iyi altküme	X5 X8 X10	-0.5	0.781	8.6084
	Genetik algoritma	X5 X8 X10	-0.5	0.781	8.6084
4	**İleriye doğru seçim**	X1 X4 X9 X8	2.37	0.757	9.5666
	Geriye doğru çıkarma	X5 X8 X10 X3	1.3	0.774	8.8864
	Adımsal regresyon	X1 X4 X9 X8	2.37	0.757	9.5666
	En iyi altküme	X5 X8 X9 X10	1.0	0.777	8.7498
	Genetik algoritma	X5 X8 X9 X10	1.0	0.777	8.7498
5	**İleriye doğru seçim**	X1 X4 X9 X8 X10	2.54	0.754	9.6721
	Geriye doğru çıkarma	X5 X8 X2 X10 X3	2.9	0.769	9.0721
	Adımsal regresyon	X1 X4 X9 X8 X10	2.54	0.754	9.6721
	En iyi altküme	X5 X8 X7 X9 X10	1.2	0.772	8.9580
	Genetik algoritma	X5 X8 X7 X9 X10	1.2	0.772	8.9580
6	**İleriye doğru seçim**	X1 X4 X9 X8 X10 X5	4.9	0.761	9.4065
	Geriye doğru çıkarma	X9 X5 X8 X2 X10 X3	4.3	0.766	9.1991
	Adımsal regresyon	X1 X4 X9 X8 X10 X5	4.9	0.761	9.4065
	En iyi altküme	X4 X5 X8 X7 X9 X10	4.3	0.766	9.1809
	Genetik algoritma	X4 X5 X8 X7 X9 X10	4.3	0.766	9.1809

7	İleriye doğru seçim	X1 X4 X9 X8 X10 X5 X7	5.9	0.760	9.4249
	Geriye doğru çıkarma	X9 X5 X8 X2 X1 X10 X3	5.5	0.765	9.2234
	Adımsal regresyon	X1 X4 X9 X8 X10 X5 X7	5.9	0.760	9.4249
	En iyi altküme	X1 X2 X3 X5 X8 X9 X10	5.5	0.765	9.2234
	Genetik algoritma	X1 X2 X3 X5 X8 X9 X10	5.5	0.765	9.2234
8	İleriye doğru seçim	X1 X4 X9 X8 X10 X5 X7 X3	6.8	0.762	9.3514
	Geriye doğru çıkarma	X9 X5 X8 X2 X7 X1 X10 X3	6.4	0.768	9.1144
	Adımsal regresyon	X1 X4 X9 X8 X10 X5 X7 X3	6.8	0.762	9.3514
	En iyi altküme	X1 X2 X3 X5 X7 X8 X9 X10	6.4	0.768	9.1144
	Genetik algoritma	X1 X2 X3 X5 X7 X8 X9 X10	6.4	0.768	9.1144
9	İleriye doğru seçim	X1 X4 X9 X8 X10 X5 X7 X3 X2	8.1	0.760	9.4249
	Geriye doğru çıkarma	X4 X9 X5 X8 X2 X7 X1 X10 X3	8.1	0.760	9.4249
	Adımsal regresyon	X1 X4 X9 X8 X10 X5 X7 X3 X2	8.1	0.760	9.4249
	En iyi altküme	X1 X2 X3 X4 X5 X7 X8 X9 X10	8.1	0.760	9.4249
	Genetik algoritma	X1 X2 X3 X4 X5 X7 X8 X9 X10	8.1	0.760	9.4249
10	İleriye doğru seçim	X1 X4 X9 X8 X10 X5 X7 X3 X2 X6	10	0.748	9.8973
	Geriye doğru çıkarma	X6 X4 X9 X5 X8 X2 X7 X1 X10 X3	10	0.748	9.8973
	Adımsal regresyon	X1 X4 X9 X8 X10 X5 X7 X3 X2 X6	10	0.748	9.8973
	En iyi altküme	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10	10	0.748	9.8973
	Genetik algoritma	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10	10	0.748	9.8973
11	İleriye doğru seçim	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	0.735	10.4200
	Geriye doğru çıkarma	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	0.735	10.4200
	Adımsal regresyon	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	0.735	10.4200
	En iyi altküme	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	0.735	10.4200
	Genetik algoritma	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11	12	0.735	10.4200

Çizelge 2, Çizelge 3, Çizelge 4 ve Çizelge 5 incelendiğinde her model genişliğinde de en iyi altküme yöntemi ile birlikte GA'nın da en küçük Cp, en küçük AKO ve en büyük düzeltilmiş R2 değerlerine sahip altkümelere ulaştığı görülmektedir. Buna karşın adımsal yöntemler genel olarak optimalden farklı altküme modellerine ulaşmışlardır. Çoğu model genişliğinde, adımsal yöntemler arasında da ulaşılan çözüm yönünden farklılık bulunmaktadır. Değişkenler arasındaki mevcut çoklu bağlantının adımsal yöntemlerin performansını olumsuz etkilediği açıktır. Örneğin steam verisi için 2, 3 ve 4 değişkenli en iyi altkümeler sırasıyla (x_1, x_7), (x_4, x_5, x_7) ve (x_1, x_4, x_5, x_7) dur. İleriye doğru seçim yöntemi ile (x_1, x_7) altkümesine ulaşıldığında (x_4, x_5, x_7) altkümesine ulaşmak imkansızdır. Benzer biçimde geriye doğru çıkarma yöntemi ile (x_1, x_5, x_7, x_9) altkümesinden (x_4, x_5, x_7) altkümesine ulaşmak mümkün olmamaktadır. Adımsal yöntemlerin her adımda modele girecek veya çıkarılacak tek bir değişkeni göz önüne alıyor olması bu sorunu yaratmaktadır. Diğer üç veri setinde de adımsal yöntemlerde aynı sorun etkisini göstermiştir.

GA'da aynı anda birden fazla çözümün popülasyonda yer alıyor olması, aramayı çözüm uzayının farklı bölgelerinde eş zamanlı olarak yapabilmeyi sağlamaktadır. Bunun sonucunda yukarıda sözü edilen sorun (yerel en iyilere takılma) GA ile yaşanmamaktadır.

5. SONUÇ

Bu çalışmada doğrusal regresyonda değişken seçimi için yeni bir yöntem önerilmiş ve performansı uygulamalarda yaygın olarak kullanılan çeşitli değişken seçim yöntemlerinin performansları ile karşılaştırılmıştır. Çeşitli veri setleri üzerinde yapılan karşılaştırmalar sonucunda, kullanılan tüm veri setlerinde de her model genişliği için GA ile elde edilen altkümelerin, tüm olası altküme yöntemi ile elde edilenler ile aynı oldukları ve her iki yöntemin aynı performansa sahip oldukları görülmüştür. Popüler istatistiksel paket programlarda bağımsız değişken sayısı üzerine konulan kısıt nedeniyle büyük boyutlu veri setlerinde tüm olası altküme yöntemini kullanmak mümkün olmamaktadır. Adımsal yöntemler ise bağımsız değişken sayısının artmasıyla ortaya çıkan çoklu bağlantıdan olumsuz etkilenmektedir. Buna karşın, GA'nın bağımsız değişken sayısı yönünden kullanıcıya bir kısıtlama getirmemesi ve dört farklı veri setinden elde ettiğimiz sonuçlar, en iyi altküme yönteminin kullanılamadığı problemlerde GA'nın kullanımını desteklemektedir. Ayrıca elde ettiğimiz sonuçlar büyük boyutlu veri setlerinde GA ile elde edilecek performans yönünden ümit verici olmuştur.

KAYNAKLAR

1. Hocking, R.R., "The analysis and selection of variables in linear regression", ***Biometrics,*** 32: 1-49 (1976).

2. Thompson, M.L., "Selection of variables in multiple regression, part I, a review and evaluation", ***International Statistical Review***, 46: 1-19 (1978a).

3. Thompson, M.L., "Selection of variables in multiple regression, part II, chosen procedures", ***Computations***

and Examples, International Statistical Review, 46: 129-146 (1978b).

4. Miller, A., Subset selection in regression, London, ***Chapman and Hall*** (1990).
5. Draper, N.R. and Smith, H., applied regression analysis, 3rd edition, ***John Wiley & Sons***, New York (1998).
6. Montgomery, D.G. and Peck, E.A., "Introduction to linear regression analysis", 2nd edition, ***John Wiley & Sons***, New York (1991).
7. Berk, K.N., "Comparing subset regression procedures", ***Technometrics***, 20(1): 1-6 (1978).
8. Goldberg, D.E., Genetic algorithms in search optimization and machine learning, ***Addison-Wesley*** (1989).
9. Wasserman, G.S. and Sudjianto, A., "All subsets regression using a genetic algorithm", ***Computers and Industrial Engineering***, 27(1): 489-492 (1994).
10. Wallet, B.C., Marchette, D.J., Solka, J.L. and Wegman, E.J., "A genetic algorithm for best subset selection in linear regression", ***Proceedings of the 28th Symposium on the Interface*** (1996).
11. Longley, J.W., "An appraisal of least-squares programs from the point of view of the user", ***JASA***, 62: 819-841 (1967).
12. Gunst, R.F. and Mason, R.L., "Regression analysis and its applications", ***Marcel Dekker***, New York (1980).
13. Chatterjee, S., Hadi, A.S., and Price, B., "Regression analysis by example, 3rd edition", ***John Wiley & Sons***, New York (2000).
14. PLS-Toolbox Version 2:1, ***Eigenvector Research Inc. Manso***n, WA (2000).

Geliş Tarihi:26.06.2002 *Kabul Tarihi:11.11.2002*