DEĞİŞKEN MALİYETLİ ZÜMRELERE GÖRE ÖRNEKLEMEDE GENETİK ALGORİTMA YAKLAŞIMI

Timur KESKİNTÜRK

İstanbul Üniversitesi, İşletme Fakültesi,
Sayısal Yöntemler Anabilim Dalı, İSTANBUL
tkturk@istanbul.edu.tr

Şebnem ER

İstanbul Üniversitesi, İşletme Fakültesi,
Sayısal Yöntemler Anabilim Dalı, İSTANBUL
sebnemer@istanbul.edu.tr

ÖZET

Zümrelere göre örnekleme, heterojen yapıdaki bir anakütlenin homojen alt gruplara ayrılarak incelendiği bir örnekleme türüdür. Bu çalışmada maliyetli zümrelere göre örnekleme problemi incelenmektedir. Örnekleme maliyetinin değişken olduğu durumda zümre sınırlarının ve örnek büyüklüğünün belirlenmesi iki farklı şekilde ele alınabilir. Bunlardan ilki, tahmin varyansını minimum yapacak şekilde diğeri ise örnekleme maliyetini minimum yapacak şekilde problemin çözülmesini amaçlamaktadır. Bu problemlerin çözümünde sezgisel bir optimizasyon tekniği olan genetik algoritma kullanılmış olup çeşitli test problemleri üzerinde denenmiş ve sonuçlar yorumlanmıştır.

Anahtar Sözcükler: *Zümrelere göre örnekleme; zümre sınırları; değişken maliyet; genetik algoritma.*

1. GİRİŞ

Zümrelere göre örnekleme özellikle farklı değerlere sahip anakütlelerin elemanlarının (N), satışlar, çalışan sayısı gibi önemli bir ya da birkaç özelliğe dayanarak, daha homojen alt gruplara (zümre) (N_1, ..., N_h) ayrıldığı bir yöntemdir (Cyert ve Davidson, 1962; Cochran, 1963; Hess ve diğerleri, 1966; Bretthauer ve diğerleri, 1999; Rao, 2000). Zümrelere göre örneklemede en önemli hedef tahmin varyansını minimize ederek, basit rassal örneklemeye kıyasla istatistiki doğruluğu arttırmaktır (Cochran, 1963). Bu hedefe ancak her bir zümrenin kendi içindeki değişkenliğinin minimum olması ile ulaşılabilmektedir (Cyert ve Davidson, 1962). Sonuç olarak, zümre sınırlarının belirlenmesi zümrelere göre örneklemenin uygulanması aşamasında karşılaşılan en önemli problemlerin başında gelmektedir. Diğer önemli husus ise örnek büyüklüğünün zümreler arasında paylaştırılması ile ilgilidir. Bu çalışmada Keskintürk ve Er (2007) çalışması temel alınarak örnekleme maliyetinin değişken olduğu durumda zümre sınırlarının belirlenmesi ve örnek büyüklüğünün dağıtılması problemlerinin genetik algoritma ile çözülmesi araştırılmaktadır.

2. ÖRNEKLEME BÜTÇESİ KISITI ALTINDA ZÜMRE SINIRLARININ BELİRLENMESİ VE ÖRNEK BÜYÜKLÜĞÜNÜN DAĞITIMI

Literatürde zümre sınırlarının belirli bir örnekleme maliyeti kısıtı altında belirlenmesi konusunda çeşitli algoritmalar (Benedetti ve diğerleri, 2005; Bretthauer ve diğerleri, 1999; Ericson, 1965) geliştirilmiştir. Bu çalışmada Keskintürk ve Er (2007)'in çalışmasında önerilen GA yaklaşımı benimsenmiş ve bu algoritma örnekleme maliyeti kısıtı altında geliştirilmiştir.

Çalışmada şu notasyonlara yer verilmektedir:

Y	Zümrelere ayrılacak anakütle
N	Anakütle büyüklüğü
n	Örnek büyüklüğü
H	Zümre sayısı

N_h	h. (h=1,…,H) zümredeki eleman sayısı
n_h	h. zümreden çekilecek örnek büyüklüğü
σ_{yh}^2	h. zümrenin varyansı
$\bar{Y}_h$	h. zümrenin ortalaması
$\bar{y}_{strat}$	Zümrelere göre örneklemede ortalamanın tahmini
C_T	Toplam örnekleme maliyeti
C_h	h'nci zümrenin örnekleme maliyetinin ortalaması
B	Örnekleme bütçesi
ξ	Tahmin varyansının üst limiti

Ortalamanın tahmini ve tahmin ortalamasının $\bar{y}_{strat}$ varyansı Cochran (1977)'da aşağıdaki gibi verilmektedir:

$$\bar{y}_{strat} = \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{N}, \tag{1}$$

$$S^2_{\bar{y}_{strat}} = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \frac{\sigma_{yh}^2}{n_h} \left(1 - \frac{n_h}{N_h}\right), \tag{2}$$

Bu formülde her bir zümrenin varyansının bilindiği ve aşağıdaki gibi hesaplandığı varsayılmaktadır:

$$\sigma_{yh}^2 = \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 / (N_h - 1), \tag{3}$$

Burada Y_{hi} h'nci zümredeki i'nci elemanın değerini temsil etmektedir. Toplam örnekleme büyüklüğü ise aşağıdaki gibi tanımlanmaktadır:

$$C_T = \sum_{h=1}^{H} C_h n_h \tag{4}$$

3 no'lu denklemde Nh>1 olduğu varsayılmaktadır, dolayısıyla Nh=1 olduğu durumda sapma sıfır olmaktadır. Bu çalışmada örnek büyüklüğü dağıtımı problemi için GA (Keskintürk, Er, 2007) yaklaşımı benimsenmiştir. Örnekleme maliyeti kısıtı altında zümre sınırlarının belirlenmesinde ve örnek büyüklüğünün zümrelere dağıtımında GA'nın nasıl uygulandığı bir sonraki bölümde detaylı olarak açıklanmaktadır.

3. ZÜMRELERE GÖRE ÖRNEKLEMEDE GENETİK ALGORİTMA YAKLAŞIMI

İlk olarak J. Holland tarafından geliştirilen GA sezgisel bir optimizasyon yöntemidir (Holland, 1975; Goldberg, 1989; Michalewicz, 1992). Problem değişkenleri kromozom adı verilen yapılarla temsil edilmektedir. Başlangıç populasyonu oluşturulduktan sonra, her kromozomun değeri uygunluk fonksiyonu ile belirlenmektedir ve çeşitli GA operatörleri

(seçim, çaprazlama ve mutasyon) uygulanmaktadır. GA'daki bu süreç önceden belirlenmiş olan iterasyon sayısına ulaşılana kadar tekrarlanmaktadır. GA'nın temel işleyiş mantığı aşağıdaki gibi özetlenebilmektedir:

Başla *Başlangıç populasyonu rassal olarak oluşturulur.*
Uygunluk Fonksiyonu: *Her bir kromozomun uygunluğu değerlendirilir.*
Seçim: *Uygunluk değeri daha iyi olan bireyler bir sonraki nesle seçilir.*
Çaprazlama: *Yeni bireyler oluşturulması için çaprazlama olasılığı ile bireylerin özellikleri değiştirilir.*
Mutasyon: *Yeni bireyler bir mutasyon olasılığı ile mutasyona uğrar.*
Döngü: *Durdurma kriterine ulaşılmadığı sürece uygunluk fonksiyonu adımına gidilir.*
Dur *ve mevcut nesildeki en iyi çözümü göster.*

Değişken örnekleme maliyetli zümreleme probleminin çözülebilmesi için zümreleme değerlerinin kromozomlar şeklinde kodlanması gerekmektedir. Bu çalışmada örnek büyüklüğü dağıtım yöntemlerinden m4 için zümre sınırlarının belirlenmesinde ikili ve reel değerli kodlama yöntemleri kullanılmıştır. Bu iki kodlama yöntemi basit bir örnek üzerinde aşağıda Şekil 1'deki gibi gösterilebilmektedir.

Değer	1.2	2.0	3.2	3.8	4.0	4.9	5.2	5.3	5.8	6.0		
İkili & reel-değerli kodlama (m4)	0	0	0	1	0	0	1	0	0	1	3	2	3

Şekil 1. Zümrelere göre örneklemede ikili ve reel-değerli kodlama örneği

Birinci "0"dan birinci "1"e kadar olan genlerin sayısı birinci zümrenin büyüklüğünü (N_1), birinci "1"den sonra gelen "0"dan ikinci "1"e kadar olan genlerin sayısı ikinci zümrenin büyüklüğünü (N_2) göstermektedir. Dolayısıyla "1" ile kodlanmış genler her bir zümrenin sınırını temsil etmektedir. Bu durumda 3.8, 5.2 ve 6.0 değerleri zümre sınırlarını göstermekte ve böylelikle zümre büyüklükleri sırasıyla 4, 3 ve 3 olmaktadır. Bu zümrelerden çekilecek örnek büyüklüklerini gösteren son 3 gene bakıldığında da örnek büyüklüklerinin 3, 2 ve 3 olduğu anlaşılmaktadır. İlk altdizinin son geni son zümrenin üst sınırını gösterdiğinden her zaman "1" olmak zorundadır.

Başlangıç populasyonu oluşturulduktan sonra, her bir kromozom bir kısıt altında uygunluk fonksiyonu da denen amaç fonksiyonu ile değerlendirilmektedir. Uygunluk değeri bireylerin bir sonraki nesle geçme olasılığını belirlemektedir. Bizim çalışmamızdaki algoritmada iki farklı kısıtlı uygunluk fonksiyonu kullanılmıştır. Bunlardan ilki tahmin varyansını bütçe kısıtı altında minimize eden ikincisi ise toplam örnekleme maliyetini varyans kısıtı altında minimize eden amaç fonksiyonlarıdır. Bu fonksiyonlar aşağıdaki gibi formülle ifade edilebilir:

Uygunluk Fonksiyonu I:

$$\min C_T,$$
$$\text{s.t. } S^2_{\bar{y}_{strat}} \le \xi,$$

Uygunluk Fonksiyonu II:

$$\min S^2_{\bar{y}_{strat}},$$
$$\text{s.t. } C_T \le B.$$

Bir sonraki adım seçim sürecidir. Seçim operatörü ile uygunluk değeri göz önüne alınarak kromozomların bir sonraki nesle geçip geçmeyeceklerine karar verilmektedir. Bu çalışmada rulet tekerleği seçim yöntemi kullanılmıştır. Çaprazlama kromozomlar arası bilgi değişimini sağlamaktadır ve bu çalışmada tek-nokta çaprazlama yöntemi kullanılmıştır. Çaprazlamadan sonra her bir kromozom düşük bir olasılıkla mutasyona tabi tutulmaktadır. Mutasyon çözüm uzayında yerel optimuma takılmayı önlemekte ve global optimumu bulma olasılığını

arttırmaktadır. Bu çalışmada rassal değişim mutasyonu kullanılmıştır. Kromozomda iki nokta rassal olarak seçilmiş ve bu noktalara karşılık gelen genler değiştirilmiştir. Kromozomdaki ikinci kısım için iki genin rassal olarak seçildiği ve bu genlerden birinden bir birimlik örnek büyüklüğü çıkartılarak diğerine eklendiği bir mutasyon türü geliştirilmiştir. Rassal seçilen genlerden hangisinden bir birimin çıkartılacağı kararı ise tamamen keyfi olarak yapılmaktadır.

4. UYGULAMA

Bu bölümde GA ile elde edilen sonuçlara yer verilmiştir. Çalışmada 2004 yılında İSO tarafından yayımlanan Türkiye'nin 500 büyük imalat sanayi firması içerisinden 487'si ile çalışılmış ve zümrelere göre örneklemede firma büyüklüğü göstergelerinden net satış değerleri kullanılmıştır. Anakütleler 2, 3, 4, 5 ve 6 zümreye ayrılmış olup, 2, 4 ve 5 zümre için toplam örnek büyüklüğü 80; 3 ve 6 zümre için 81 olarak belirlenmiştir. GA parametrelerinden populasyon büyüklüğü 50, çaprazlama oranı 0.99 ve mutasyon oranı 0.25 olarak kabul edilmiştir.

Tahmin varyansı değerleri belirlenirken Keskintürk ve Er (2007)'in çalışmasında elde edilen minimum varyans değerleri hedef olarak alınmıştır. Bu değerler 20 farklı kısıt değeri elde edebilmek için sabit bir miktarda arttırılmış ve bu doğrultuda elde edilen minimum ve maksimum tahmin varyansı kısıt değerleri Tablo 1'de verilmiştir.

Tablo 1. Tahmin Varyansı için Hedef Değerler (10^{14})

	min varyans hedefi	max varyans hedefi
2	2.133	11.633
3	0.459	6.159
4	0.197	4.187
5	0.125	2.975
6	0.061	2.436

Maliyet kısıtı belirlenirken de varyans kısıtlı problemlerin çözümünden elde edilen maliyet değerleri göz önüne alınarak 6500 ve 3150 arasında 20 farklı bütçe kısıtı hesaplanmıştır.

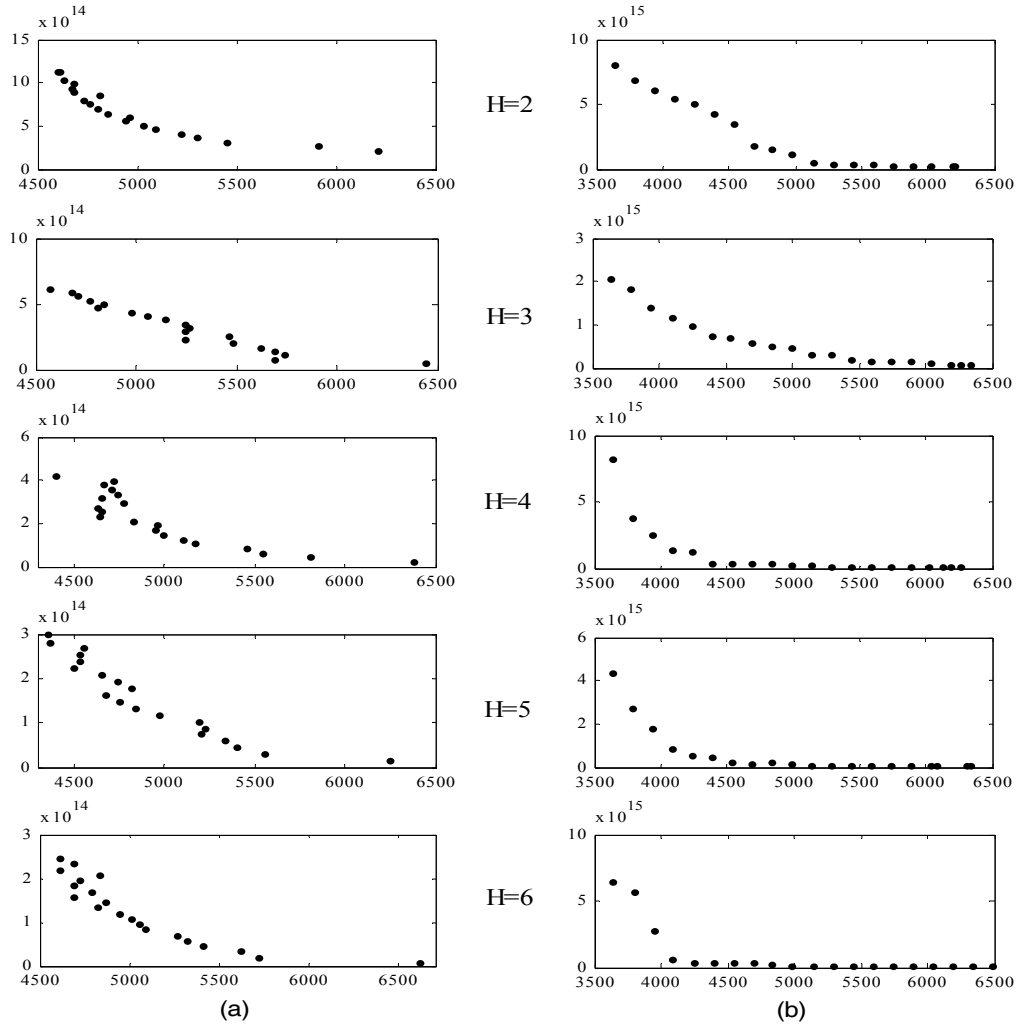
Tablo 2 her iki amaç fonksiyonu için bütün kısıtlar sağlandığında GA ile elde edilen değerleri göstermektedir.

Tablo 2. Maliyet ve Varyans (10^{14}) Minimizasyonu Uygunluk Fonksiyonlarının Sonuçları

	Uygunluk Fonksiyonu I'in maliyet sonuçları					*Uygunluk Fonksiyonu II*'nin varyans sonuçları				
Problem	H=2	H=3	H=4	H=5	H=6	H=2	H=3	H=4	H=5	H=6
1	6213	6448	6380	6253	6614	2.133	0.459	0.242	0.137	0.074
2	5912	5695	5814	5555	5717	2.133	0.500	0.211	0.135	0.079
3	5452	5739	5551	5407	5618	2.133	0.569	0.825	0.143	0.123
4	5304	5697	5457	5343	5410	2.181	0.916	0.251	0.161	0.134
5	5229	5619	5179	5205	5327	2.289	1.262	0.274	0.290	0.116
6	5090	5490	5111	5228	5264	2.459	1.507	0.338	0.348	0.171
7	5030	5248	4999	5198	5092	2.730	1.503	0.420	0.300	0.366
8	4945	5464	4953	4977	5057	3.158	1.867	0.781	0.324	0.550
9	4965	5245	4969	4849	5015	3.680	2.933	0.880	0.439	0.695
10	4850	5271	4838	4757	4952	4.304	2.856	1.386	0.644	1.105

11	4802	5250	4642	4679	4824	10.758	4.309	1.456	1.207	0.941
12	4765	5149	4662	4825	4871	14.392	4.910	2.916	1.611	2.163
13	4737	5060	4636	4742	4698	17.976	5.611	3.778	1.475	3.053
14	4813	4982	4782	4662	4796	34.942	6.977	2.863	2.155	2.992
15	4686	4812	4654	4506	4690	42.748	7.317	3.345	4.550	3.434
16	4673	4840	4748	4536	4730	49.771	9.565	11.706	5.165	3.170
17	4685	4773	4716	4535	4842	54.428	11.631	13.901	7.894	5.861
18	4634	4711	4669	4561	4616	60.355	13.985	24.700	17.656	26.763
19	4618	4688	4720	4369	4688	68.669	18.193	37.844	26.569	56.347
20	4605	4577	4400	4358	4620	79.969	20.538	81.724	43.278	63.637

Şekil 2'deki serpilme diyagramı her bir zümre büyüklüğü ve uygunluk fonksiyonu için GA ile elde edilen varyans ve maliyet sonuçlarını göstermektedir.



Şekil 2. Maliyet ve tahmin varyansı değerlerinin serpilme diyagramı

Tablo 2 ve Şekil 2'den de görülebileceği gibi örnekleme bütçesi arttıkça tahmin varyansını minimize etmek ve böylelikle istatistik doğruluğu arttırmak mümkün olmaktadır. Aynı şekilde varyans kısıtı arttırıldığında daha düşük örnekleme maliyeti ile çalışma sağlanabilmektedir.

5. SONUÇ

Zümrelere göre örnekleme özellikle dışa düşene sahip heterojen yapıdaki anakütleler için oldukça yaygın olarak kullanılan bir örnekleme türüdür. Yaygın olarak kullanılmasının temel nedeni bu yapıdaki anakütlelerde basit rassal örneklemeye kıyasla istatistik olarak daha etkin olmasıdır. Tahmin varyansının minimize edilebilmesi için zümre sınırlarının tespit edilmesi ve bu zümrelerden çekilecek örnek büyüklüklerinin ne olacağına karar verilmesi gerekmektedir. Bu çalışma Keskintürk ve Er (2007)'in çalışmasında zümre sınırlarının ve örnek büyüklüğünün belirlenmesi problemine önermiş oldukları GA yaklaşımını örnekleme maliyetinin değişken olduğu durum için ele almıştır. Sonuçlar GA ile elde edilen varyans ve maliyet değerleri arasında olması gereken negatif ilişkiyi desteklemektedir. Böylelikle GA'nın örnekleme maliyetinin değişken olduğu heterojen yapıdaki populasyonların incelenmesinde etkin bir şekilde kullanılabileceğini görülmektedir.

KAYNAKÇA

Benedetti, R., Giuseppe, E., Giovanni, L., 2005. A tree-based approach to forming strata in multipurpose business surveys. *Department of Economics Working Papers 0505*, Department of Economics, University of Trento, Italia.

Bretthauer, K. M., Ross, A., Shetty, B., 1999. Nonlinear integer programming for optimal allocation in stratified sampling. *European Journal of Operational Research* 116, 667-680.

Cochran, William G., 1977. *Sampling Techniques*, 3rd ed., John Wiley & Sons, Inc. USA.

Cyert, R.M., Davidson, H.J., 1962. *Statistical Sampling for Accounting Information*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 116-127.

Ericson, W. A., 1965. Optimum Stratified Sampling Using Prior Information. *Journal of the American Statistical Association* 60, 311, 750-771.

Goldberg, D.E., 1989. *Genetic Algorithms in Search Optimization and Machine Learning*, Addison Wesley Publishing Company, New York.

Hess, I., Sethi, V.K., Balakrishnan, T.R., 1966. Stratification: A Practical Investigation. *Journal of the American Statistical Association* 61, 313, 74-90.

Holland, J.H., 1975. *Adaptation in natural and artificial systems*, University of Michigan Press Ann Arbor.

Keskintürk, T., Er, Ş., 2007. A Genetic Algorithm Approach to Determine Stratum Boundaries and Sample Sizes of Each Stratum in Stratified Sampling. *Computational Statistics & Data Analysis* 52, 1, 53-67.

Orhunbilge, N., 2000. *Örnekleme Yöntemleri ve Hipotez Testleri*, 2. Baskı, Avcıol Basım Yayın, İstanbul, Türkiye.

Rao, P.S.R.S., 2000. *Sampling Methodologies with Applications*. Chapman & Hall/CRC, Washington.