Derinliğe Dayalı Diskriminasyon

Cemal ATAKAN[1], İhsan KARABULUT[1]

Özet: Bu çalışmada, çok değişkenli veri analizinde son yıllarda kullanılan derinlik kavramına dayalı olarak atama problemi üzerinde durulmuştur. Eşit varyans-kovaryans matrisine sahip iki değişkenli normal dağılımlı kitlelerden üretilen veri kümeleri kullanılarak, klasik diskriminant analizinde tanımlanan hata oranları ve simpleks, yarı-uzay derinlikleri ile yapılan atama kuralına göre elde edilen hata oranları için bazı değerlendirmeler ve gözlemler yapılmıştır.

Anahtar Kelimeler: Diskriminant analizi, derinlik ölçüsü, simpleks derinliği, yarı-düzlem derinliği, hata oranı.

Discrimination Based on Depth

Abstract: In this study, the allocation problem is concerned on the depth notion which has been used in the multivariate data analysis in recent years. Some evaluations and observation are made for the error rates those are defined in the clasical discriminant analysis and the allocate with simplicial, half-space depths by using two generated data sets from bivariate normal distributions with the same variance-covariance matrices.

Key Words: Discrimnant analysis, depth measure, simplicial depth, half-space depth, error rate.

Giriş

Diskriminant analizi, üzerinden ölçüm alınan bir birimin sonlu sayıda bilinen farklı kitlelerden birine atanmasını gerçekleştiren istatistiksel bir teknik olarak tanımlanır. Bu atama işlemi yapılırken birim aldığı gözlem değerine göre ait olduğu kitleden farklı bir kitleye atandığında, bir hata yapılmış olur. Diskriminant analizinde bu hataya, hata oranı ya da hatalı sınıflandırma olasılığı denmektedir. Diskriminant analizinde amaç, atama işlemini minumum hatayla yapmaktır.

Tek boyutta rasgele değişkenlerin sıralanması doğal bir şekilde olurken, rasgele vektörlerin sıralanması böyle bir doğallıkla yapılamamaktadır. Derinlik kavramı tek boyutlu rasgele değişkenlerin sıralanmasıyla da kısmen örtüşen, bir dağılımın merkezine göre değişken değerlerinin sıralanması olarak düşünülebilir. Çok değişkenli verilerin analizinde, derinlik ölçülerinin kullanımı son yıllarda yaygınlaşmaktadır. Liu, Parelius ve Singh'in çalışmalarında bu uygulamalardan özetle söz edilmektedir[1].

Derinlik sıra sayılarına göre de diskriminasyonun yapılabileceği Liu ile Liu ve Singh'in çalışmalarında söz edilmektedir[2,3]. Derinlik ölçüleri ve bu ölçülere dayalı atama konusunda özet bilgi takip eden alt bölümlerde verilecektir.

[1] Ankara Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 06100 Tandoğan, Ankara, TÜRKİYE.

Bu çalışmada, iki değişkenli aynı varyans – kovaryans matrisine sahip normal dağılımlı kitlelerden alınan iki özel örnekleme bağlı olarak lineer diskriminant analizi, yarı uzay ve simpleks örneklem derinlik değerlerine göre diskriminasyon yapılacaktır. Hatalı sınıflandırma olasılıkları konusunda bu özel örneklere dayalı birtakım gözlemler okuyucuyla paylaşılacaktır. Derinliklere bağlı atama tekniğinin potansiyeli konusunda ileride yapılması öngörülen çalışmalara zemin hazırlanacaktır. Yapılacak gözlemlerin henüz genellenemeye uygun olmadığını da yineleyelim.

Diskriminant Analizi ve Hata Oranları

Diskriminant analizi, üzerinde ölçüm yapılan bir bireyi sonlu sayıda bilinen farklı kitleden birine atanmasını gerçekleştiren istatistiksel bir tekniktir. Π_1 ve Π_2 birbirinden farklı iki kitle olmak üzere, $X=(X_1,X_2,...,X_p)'$ birey üzerinde ölçümlere karşılık gelen p - boyutlu rasgele vektörü Π_i kitlesinden ise X'in ortak olasılık yoğunluk fonksiyonu $f_i(x,\theta_i)$ biçiminde gösterilir. Burada, θ_i parametre vektörü ve $x\in\Re^p$ dir.

X'in aldığı değerler p - boyutlu $\Re^p$ örneklem uzayında olmak üzere, sınıflandırma işlemi bu uzayı $B_1\cup B_2=\Re^p$ ve $B_1\cap B_2=\varnothing$ olan B_1 ve B_2 bölgelerine ayırır. Eğer X'in gözlem değeri B_1 bölgesinde ise bu gözlemin yapıldığı gözlem birimi Π_1, aksi halde Π_2 kitlesine atanır.

Bu çalışmada Welch tarafından önerilen toplam hatalı sınıflandırma olasılıklarını minimize eden olabilirlik oran kriterine göre elde edilen lineer diskriminant fonksiyonu göz önüne alınmıştır[4]. Π_1 ve Π_2, çok değişkenli normal dağılımlı kitleler olmak üzere sırasıyla μ_1 ve μ_2 ortalama vektörleri ve aynı Σ varyans kovaryans matrisne sahip olduğu bilindiğinde, toplam hatalı sınıflandırma olasılığını minimize eden optimal sınıflandırma kuralı,

$$\xi:\begin{cases}U(x)>k \text{ ise } x, & \Pi_1\text{'e}\\ \text{aksi halde} & \Pi_2\text{'ye atanir.}\end{cases}$$

ile verilir. Burada,

$$U(x)=\left[x-\frac{1}{2}(\mu_1+\mu_2)\right]'\Sigma^{-1}(\mu_1-\mu_2)$$

iki kitleden birine ait olduğu ancak hangisine ait olduğu bilinmeyen rasgele gözlemin değeri *x* için parametrelerin bilindiği durumda $f_1(x;(\mu_1,\Sigma))$ ve $f_2(x;(\mu_2,\Sigma))$ ortak olasılık yoğunluk fonksiyonlarının birbirine oranlanmasıyla elde edilen kitle lineer diskriminant fonksiyonu, $k=\ln(\frac{q_2c_2}{q_1c_1})$ dır. q_i birimin *i* inci kitleye ait olması olasılığı (önsel olasılık) ve c_i bu birimin yanlışlıkla diğer kitleye atanmasının maliyetidir.

Parametreler bilinmediğinde Π_1 ve Π_2 kitlelerinden alınan n_1 ve n_2 hacimli rasgele örneklemlerden hesaplanan tahmin edicileri $\overline{X}_1$ ve $\overline{X}_2$ örneklem ortalama vektörleri ve S birleştirilmiş örneklem varyans-kovaryans matrisine bağlı optimal sınıflandırma kuralı

$$\hat{\xi}:\begin{cases}W(x)>k \text{ ise } x, & \Pi_1\text{'e}\\ \text{aksi halde} & \Pi_2\text{'ye atanir.}\end{cases}$$

ile verilir. Burada,

$$W(x)=\left[x-\frac{1}{2}(\bar{x}_1+\bar{x}_2)\right]'S^{-1}(\bar{x}_1-\bar{x}_2)$$

parametrelerin bilinmediği durumda, kitle lineer diskriminant fonksiyonunun elde edilmesine benzer biçimde elde edilen örneklem lineer diskriminant fonksiyonudur [5,6]. Bu çalışmada maliyetler ihmal edilip, önsel olasılıklar eşit alındığından $k=0$ olacaktır.

Atama işlemi yapılırken, bilinen kitlelerin hangisinden olduğu bilenmeyen bir birim, ait olduğu kitleye değil de diğer kitleye atanırsa, hata yapılmış olur. Diskriminant analizinde amaç, atama işlemini minumum hatayla yapmaktır. Bu optimizasyon ölçütüne göre elde edilen diskriminant fonksiyonlarının değerlendirilmesinde hata oranlarının ya da hatalı sınıflandırma olasılıklarının bilinmesi önemlidir. Önsel olasılıklar ile ağırlıklandırılmış hata oranına , toplam hata oranı veya toplam hatalı sınıflandırma olasılığı denilmektedir.

Hata oranı, diskriminant fonksiyonunun dağılımına bağlı olarak bulunur. Yukarıda verilen atama kurallarına ilişkin optimal, gerçek (koşullu) ve beklenen gerçek (koşulsuz) hata oranı gibi hata oranları tanımlanır. Optimal hata oranı, parametreler bilindiğinde elde edilen diskriminant fonksiyonu göz önüne alınarak bulunan hata oranı, gerçek hata oranı, parametreler bilinmediğinde örneklemlerden elde edilen tahminlere bağlı örneklem diskriminant fonksiyonuna göre bulunan hata oranı ve beklenen gerçek hata oranı, olası tüm örneklemler üzerinden gerçek hata oranının beklenen değeridir [5,6,7,8,9].

Çok değişkenli normal kitleler göz önüne alındığında, $X \sim N(\mu_i, \Sigma)$ ise $U(X)$'in dağılımı $(-1)^i(-\frac{\Delta}{2})$ ortalamalı ve Δ^2 varyanslı tek değişkenli normaldir. Burada,

$$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$$

iki kitle arasındaki Mahalanobis uzaklığıdır.

Parametrelerin bilinmediği durumda elde edilen $W(X)$ örneklem lineer diskriminant fonksiyonunun dağılımı kolaylıkla elde edilememektedir, ancak bazı şartlar altında elde edilebilmektedir [10,5]. $W(X)$ ifadesindeki tahmin ediciler yerine örneklemden elde edilen tahmin değerleri alındığında, $W(X)$'in bu tahmin değerlerine koşullandırılmış koşullu dağılımı elde edilebilir. Böylece $X \sim N(\mu_i, \Sigma)$ ise $W(X)$ 'in koşullu dağılımı, ortalaması $\left[\mu_i - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right]' s^{-1}(\bar{x}_1 - \bar{x}_2)$ ve varyansı $(\bar{x}_1 - \bar{x}_2)' s^{-1} \Sigma s^{-1} (\bar{x}_1 - \bar{x}_2)$ olan tek değişkenli normaldir [6]. Burada *s*, *S* tahmin edicisinin tahmin değeridir.

Diskriminant fonksiyonlarının dağılımları göz önüne alındığında, Π_1'den bir birimin hatalı atanması durumunda ξ kuralına göre optimal hata oranı,

$$\begin{aligned} \alpha_1(\xi) &= P(U(X) \le 0 / X \in \Pi_1) \\ &= \Phi(-\Delta/2) \end{aligned}$$

$\hat{\xi}$ kuralına göre gerçek (veya koşullu) hata oranı,

$$\begin{aligned} \alpha_1(\hat{\xi}) &= P(W(X) \le 0 / X \in \Pi_1, \bar{x}_1, \bar{x}_2, s) \\ &= \Phi\left(- \frac{\left[\mu_1 - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right]' s^{-1}(\bar{x}_1 - \bar{x}_2)}{\left[(\bar{x}_1 - \bar{x}_2)' s^{-1} \Sigma s^{-1} (\bar{x}_1 - \bar{x}_2)\right]} \right) \end{aligned}$$

ve beklenen gerçek hata oranı

$$E(\alpha_1(\hat{\xi})) = E[P(W(X) \le 0 / X \in \Pi_1)]$$

dır. Burada $\Phi(.)$ standart normal dağılım fonksiyonudur. Π_2'den bir birimin hatalı atanması durumunda da benzer ifadeler elde edilebilir. Buradan önsel olasılıkların eşit olması durumunda toplam hatalı sınıflandırma olasılığı optimal ve gerçek hata oranları için sırasıyla

$$\alpha(\xi) = \frac{1}{2}(\alpha_1(\xi) + \alpha_2(\xi))$$

ve

$$\alpha(\hat{\xi}) = \frac{1}{2}(\alpha_1(\hat{\xi}) + \alpha_2(\hat{\xi}))$$

biçiminde elde edilir.

Ayrıca gerçek hata oranı için parametrik olmayan bir tahmin edici olan tekrar yerine koyma tahmin edicisine ilişkin sonuçlar da verilecektir. Bu tahmin edicinin değeri, diskriminant fonksiyonu elde edildikten sonra örneklemlerdeki gözlemlerin bu fonksiyon kullanılarak kitlelere atanması sırasında her bir kitlede hatalı atanan gözlemlerin sayısının her bir kitledeki toplam gözlem sayısına oranlanmasıyla bulunur [11]. Bu tahmin edici

$$\hat{\alpha}(\hat{\xi}) = \frac{1}{2}(\hat{\alpha}_1(\hat{\xi}) + \hat{\alpha}_2(\hat{\xi}))$$

ile gösterilecektir.

Bu çalışmada optimal, gerçek hata oranı ve tekrar yerine koyma tahmin edicisine ilişkin sonuçlar verilecektir.

Derinlik Kavramı

N hacimli rasgele bir örneklemden gözlenen tek boyutlu verilerle doğal bir sıralama yapabilmek olanaklı iken, hepsi de *F* dağılımlı ve birbirlerinden bağımsız $X_i \in \Re^p$ rasgele vektörleri $X_1, X_2, ..., X_n$ için aynı doğallıkta bir sıralama söz konusu değildir. Rasgele vektörlerin sıralanmasında değişik sıralama yöntemleri vardır. Tukey'in 1975 de ortaya koyduğu derinlik ölçüleri kullanılarak, rasgele vektörlerin sıralanması da yapılabilmekte ve bu sıralama tek boyutlu rasgele değişkenlerin sıralanmasıyla aynıdır. Çok genel olmakla birlikte derinlik, bir $x \in \Re^p$ noktasının F dağılım fonksiyonunun *merkezine* göre yakınlığının bir ölçüsüdür. Bu ölçüyü $D(x)$ ile gösterilecektir. Bir $x \in \Re^p$ noktasının derinliği F dağılım fonksiyonuna veya $X_1, X_2, ..., X_n$ rasgele örneklemine göre tanımlanabilir. Rasgele örneklemin gözlenen değerleri derinlik değerlerine göre büyükten küçüğe doğru sıralanarak, gözlem değerlerinin herbirine ayrı ayrı sıra sayısı verilebilir. Örneklemden gözlenen değerler için belirlenen bir derinlik ölçüsüne göre elde edilen derinlik değerleri sıralandığında, ilk sırada yer alan gözlem en derin gözlem değeri olup, bu gözleme sıra sayısı olarak bir ve son sırada yer alan gözlem merkezden en uzak gözlem değeri olup, bu gözleme sıra sayısı olarak *n* atanır. Aynı derinliklere sahip olan gözlem değerlerine herhangi bir sıra gözetmeksizin, birbirini takip eden sıra sayıları atanır.

Son zamanlarda, derinlik ölçülerine dayalı olarak yapılan çok değişkenli parametrik olmayan istatistiksel analizler yaygınlaşmaktadır. Derinlik ölçülerinin bazı özellikleri, uygulama alanları ve değerlendirilmeleri konularında; çok değişkenli veri analizindeki yönleri için [1,12], kalite kontolü uygulamaları için [13], yüzdelik(quantile) süreçlerinde ise [14] çalışmaları örnek verilebilir.

Derinlik ölçülerinde aranması gereken veya istenilen özellikler; affin değişmezlik, dağılım merkezinde en büyük derinlik, en derin noktaya göre monotonluk, dağılım merkezinden uzaklaştıkça derinliğin sıfıra yaklaşması olarak özetlenebilir [15]. Bir çok durumda bu özelliklere sahip olduğu bilinen derinlik ölçülerinden ikisi aşağıda tanıtılmaktadır.

***Yarı-düzlem derinliği*:** Yukarıda belirtilen özelliklerin tümüne sahip olduğu bilinen bir derinlik ölçüsüdür. $x \in \Re^p$ noktasının F dağılımına göre yarı düzlem derinliği

$$HD(F;x) = \inf_H \left\{ P(H) : H, \Re^p \text{ 'de } x \in H \text{ olan kapalı bir yarı - düzlem} \right\}$$

olarak tanımlanır[15]. Yarı – düzlem derinliğinin örnek karşılığı

$$D_n(x) = \inf_H \left\{ \frac{n\{X_i; X_i \in H\}}{n}; H, \Re^p \text{ 'de } x \in H \text{ olan kapalı bir yarı - düzlem} \right\}$$

biçimindedir. Burada $n\{X_i; X_i \in H\}$ *H* yarı-düzlemi içinde üzerinde gözlem yapılan örneklem birimi sayısıdır. Bu derinlik ölçüsünü örneklem dağılım fonksiyonu F_n'ye göre yarı-düzlem derinliği olarak da tanımlanabilir. Yarı – düzlem derinliğinin bazı önemli özellikleri [16, 17]'

de incelenmiştir. $x \in \Re^p$ noktasının derinliğinin tahmin edicisi $D_n(x)$, $D(x)$'in tutarlı bir tahmin edicisidir.

Simpleks derinliği: $x \in \Re^p$ noktasının F dağılım fonksiyonuna göre simpleks derinliği

$$SD(F;x)=P\left(x \in \mathbf{S}[X_1,X_2,...,X_{p+1}]\right)$$

olarak tanımlanır. Burada $\mathbf{S}[X_1,X_2,...,X_{p+1}]$, köşeleri F dağılımından alınan $p+1$ tane, $X_1,X_2,...,X_{p+1}$, rasgele örneklem noktaları (vektörleri) olan kapalı bir simplekstir. $\Re^2$'de simpleksler üçgenler; daha yüksek boyutlarda çok yüzlülerdir(politoplardır). Simpleks derinliğinin örnek karşılığı

$$SD_n(F;x)=\binom{n}{p+1}^{-1}\sum \mathbf{I}_{\mathbf{S}[X_{j1},X_{j2},....X_{jp+1}]}(x)$$

şeklindedir. Burada $\mathbf{I}_{\mathbf{S}[X_{j1},X_{j2},....X_{jp+1}]}(x)$ gösterge fonksiyonu olup $x \in \Re^p$ noktası $\mathbf{S}\left[X_1,X_2,...,X_{p+1}\right]$ simpleksinin elemanı ise bir değilse sıfır değerini alır. Toplam $X_1,X_2,...,X_n$ örnekleminden $\binom{n}{p+1}$ sayıda oluşturulabilecek $\mathbf{S}[X_{j1},X_{j2},....,X_{jp+1}]$ gibi simpleksler üzerinden yapılmaktadır. Simpleks derinliği de *sürekli* dağılımlı rasgele vektörler için bütün özellikler sağlamakta ancak kesikli değerler alan rasgele vektörler için dağılım merkezinde en büyük derinlik, ve bazen de en derin noktaya göre monotonluk özelliklerini yitirmektedir[15].

Bu çalışmanın konusu olan atama ile ilgili uygulamalarda derinlik kavramının kullanımı için bilgi [2,3]' de kısaca verilmiştir. Bu konudaki yapılan ilk çalışma, yukarıdaki iki çalışmada kaynak olarak gösterilmiş olan Gross ve Liu(1988) yayımlanmamış çalışmasıdır. Bu konuda elde edilen bilgiler yukarıda verilen iki çalışma ile sınırlıdır. Bu nedenle, önerilen atama yönteminin özellikleri konusunda ve çalışmaların hangi derinlik ölçülerinin dikkate alınarak yapıldığı bilinmemektedir.

Liu ve Singhderinlik ölçülerinin atama problemlerinde kullanımını aşağıdaki şekilde tarif etmişlerdir[3]:

Π_1 ve Π_2 kitlelerinden alınan n_1 ve n_2 hacimli rasgele örneklemleri gözönüne alalım. Bu kitlelerin birine ait olduğu ancak hangisinden olduğu bilinmeyen yeni bir gözlemin Π_1 kitlesine göre hesaplanan örneklem derinlik sıra sayısı (rankı) r_1 ve Π_2 kitlesine göre hesaplanan örneklem derinlik sıra sayısı r_2 ile gösterilsin. Sıralama derinliklere göre yapılmaktadır; en derin noktaya sahip gözlemin sıra sayısı 1 ve en az derinliğe sahip gözlemin sıra sayısı n_i olacaktır. Burada dikkat edilmesi gereken husus, aynı derinliğe sahip olan noktalar gözlemin örnekleme girme sırasına göre sıralanacaktır. Böylece yeni bir gözlemin iki kitleden birine atanması için önerilen kural $r_1/n_1<r_2/n_2$ ise bu gözlem Π_1 kitlesine, aksi halde Π_2 kitlesine atanması biçiminde tanımlanır. Bu çalışmada atama işlemi sıra sayıları yerine, gözlemlerin örneklem derinlik değerlerine göre yapılacaktır. Buna göre bir $x \in \Re^p$ gözlemi Π_1 kitlesinden alınan örnekleme göre hesaplanan örneklem derinliği $D_{n_1}(x)$ ile Π_2 kitlesinden alınan örnekleme göre hesaplanan örneklem derinliği $D_{n_2}(x)$ olmak üzere, eğer $D_{n_1}(x) \geq D_{n_2}(x)$ ise *x* gözleminin yapıldığı birim Π_1 kitlesine aksi halde Π_2 kitlesine atanır. Bu değerlendirme işlemi yukarıda açıklanan örneklem derinliği sıra sayılarına göre yapılan değerlendirmeden daha kolay olacağı, karşılaştırma yapılırken sıra sayıları yerine gözlemlerin derinlik değerleri doğrudan kullanıldığından karşılaştırma daha hassas olacağı düşünülmektedir. Ayrıca sıra sayıları ile yapılan karşılaştırmada kitlelerden

alınan örneklem büyüklüklerinin farklı olması gözlemin derinlik sıra sayısını etkileyeceğinden, tahmin edilmeye çalışılan kitle derinliğinden ve sıra sayısından da sapma olacaktır.

Uygulama

Çalışmanın bu kısmında yukarıda ifade edilmeye çalışılan atama yöntemlerinin nasıl işletileceği dağılımı bilinen kitlelerden üretilen verilerle gösterilecek ve bu yöntemlere ilişkin hata oranları üretilen veriler için tahmin edilecektir. Burada ortalamaları farklı ancak varyans-kovaryans matrisleri eşit olan çok değişkenli normal dağılımlı kitleler göz önüne alınacaktır. Böyle bir uygulama için normal dağılımlı kitlelerden üretilmiş verilerin kullanılmasının birinci nedeni, diskriminant analizi uygulamalarında genellikle bu dağılımın ve söz konusu varsayımların yapılıyor olmasıdır. İkinci neden ise bir eliptik dağılım olarak normal dağılımın derinlik konturlarının(izdüşüm çizgilerinin) bu dağılıma ait olan ortak olasılık yoğunluk fonksiyonun konturlarıyla benzeşmesidir[1,2]. Bu çalışmada, basitlik için iki değişkenli normal dağılımlı kitleler göz önüne alınacaktır. Dağılımların varyans-kovaryans matrislerinin eşit olmaması durumu ayrıca değerlendirilebilir.

Tablo 1 de

$$N(\mu_1 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 9 & -7.8 \\ -7.8 & 16 \end{bmatrix}) \quad , \; N(\mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 9 & -7.8 \\ -7.8 & 16 \end{bmatrix})$$

dağılımlarından ve Tablo 2 de

$$N(\mu_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 4 & 3.9 \\ 3.9 & 9 \end{bmatrix}) \quad , \quad N(\mu_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 4 & 3.9 \\ 3.9 & 9 \end{bmatrix})$$

dağılımlarından üretilen $n_1 = n_2 = 50$ şer birimlik gözlemler yer almaktadır. Tablolarda atama kuralları sütunlarında SD(Simpleks Derinliği), HD(Yarı-Uzay Derinliği) ve LDF(Lineer Diskriminant Fonksiyonu) ifadelerine göre gözlemlerin atandıkları kitleler belirtilmiştir. Burada 1 gözlemin birinci kitleye, 2 gözlemin ikinci kitleye atandığını göstermektedir. Gözlemlere ait SD ve HD derinlik değerleri [18]' de verilen derinlik algoritmalarına göre hesaplanmıştır. Gözlemlere ait derinlik değerleri tablo sayısının artacağı düşüncesiyle verilmemiştir. LDF sütunu, gözlemlerin yeniden yerine koyma tahmin edicisine göre atandığı kitleyi göstermektedir.

Sonuç

Bu uygulama sonucu aşağıda verilecek değerlendirmeler, atama yöntemlerinin karşılaştırılması anlamında genellenmez. Her iki örneklemde de, derinliklere göre yapılan atamalarda hata oranları, klasik lineer diskriminant fonksiyonuna göre yapılan atamadaki hata oranlarına göre daha küçük olduğu gözlenmiştir. Bununla birlikte, kitleler birbirlerinden uzaklaştırıldığında, diskriminant analizinde olduğu gibi, derinliklere göre yapılan atamalarda da hata oranları azalmaktadır.

Tablo 1: İlk örneklem için gözlem değerleri ve atama kurallarına göre gözlemlerin atandığı kitleler.

Birinci Kitleden Gözlemler		Atama Kuralları			İkinci Kitleden Gözlemler		Atama Kuralları		
		SD	HD	LDA			SD	HD	LDA
-0.7307	-1.2595	2	2	2	-3.7932	-3.7352	2	2	2
1.8884	-8.2338	1	1	2	5.8090	-9.8613	2	2	1
1.1734	3.1310	1	1	1	4.1665	-2.3695	1	1	1
0.9187	5.2061	1	1	1	-2.1047	3.3030	2	2	2
-3.5678	8.0697	1	1	2	-3.8439	5.0575	2	2	2
-3.3012	5.8389	1	2	2	1.2977	-3.4512	2	2	2
1.0481	-2.1239	1	2	2	0.7161	-1.2897	2	2	2
6.9017	-5.6174	1	1	1	-3.0226	-0.9251	2	2	2
1.4402	7.7681	1	1	1	3.1819	-6.6277	2	2	1
1.0491	-1.2224	1	1	2	-0.5432	-4.9142	2	2	2
0.9788	-0.5901	1	1	1	-1.6788	-0.8717	2	2	2
-1.4810	6.2392	1	1	1	4.3558	-5.3207	2	2	1
6.2818	-0.7642	1	1	1	2.0524	-1.3405	1	1	1
-0.6161	-1.1204	2	2	2	0.0643	2.1578	2	2	1
0.6576	6.5167	1	1	1	2.5303	0.9496	1	1	1
-3.3556	6.1015	1	2	2	4.5681	-5.3498	2	1	1
-2.5724	1.9431	2	2	2	-0.7098	-2.9125	2	2	2
5.0392	-2.9703	1	1	1	-4.9031	6.5552	2	2	2
4.2051	1.5479	1	1	1	-1.7449	4.4073	2	2	2
3.7518	1.6333	1	1	1	2.3716	-6.4996	2	2	2
2.0426	-6.4987	1	1	2	-1.4542	3.3384	2	2	2
-0.6408	7.4862	1	1	1	-4.6313	3.6988	2	2	2
6.5501	-5.2915	1	1	1	-1.4124	0.1667	2	2	2
1.2010	2.0982	1	1	1	0.6512	0.4745	2	2	1
1.9558	-0.2080	1	1	1	-3.7691	7.5722	2	1	2
6.0907	-7.8146	1	1	1	-1.5456	-2.9287	2	2	2
1.3579	3.2018	1	1	1	1.2819	3.1575	1	1	1
1.0443	-0.6168	1	1	1	-0.6594	4.1693	2	1	1
4.0863	-4.5551	1	1	1	0.8128	5.5179	2	2	1
-0.9424	-3.4944	1	2	2	1.5970	0.3281	2	2	1
1.9897	2.3285	1	1	1	-7.1217	4.6105	2	2	2
5.1734	-4.0094	1	1	1	-0.7201	3.5803	2	2	2
-2.1545	1.0771	2	2	2	-1.0881	7.9022	1	1	1
1.1584	-1.9011	1	2	2	0.7119	-6.2985	2	2	2
-0.3389	-4.1890	1	2	2	0.3666	1.5771	2	2	1
1.0564	-0.6295	1	1	1	2.6126	-5.9097	2	2	2
-1.4174	-3.0028	1	2	2	-4.6864	1.6361	2	2	2
5.4423	-8.2168	1	1	1	-0.4866	0.2982	2	2	2
6.0640	-2.3849	1	1	1	0.8573	-0.3559	2	2	2
-1.6667	5.3338	1	1	2	1.2171	-1.9603	2	2	2
1.8990	-1.6257	1	1	1	4.1955	-7.7340	2	2	1
3.1994	0.1220	1	1	1	-3.4686	1.1875	2	2	2
2.0792	5.0507	1	1	1	1.6824	2.3027	1	1	1
3.2198	-0.8243	1	1	1	0.6779	-0.3888	2	2	2
2.3815	1.9181	1	1	1	0.1191	6.5012	2	1	1
1.0461	-3.2272	2	2	2	4.5596	-5.2924	2	1	1
4.4794	-5.2992	1	1	1	-0.4163	-3.1151	2	2	2
2.1982	-3.3117	1	1	1	-1.4065	-1.1540	2	2	2
2.1785	-0.8420	1	1	1	-4.4963	6.8715	2	2	2
2.9233	-0.0087	1	1	1	1.0040	3.8384	1	1	1

Tablo 2: İkinci örneklem için gözlem değerleri ve atama kurallarına göre gözlemlerin atandığı kitleler.

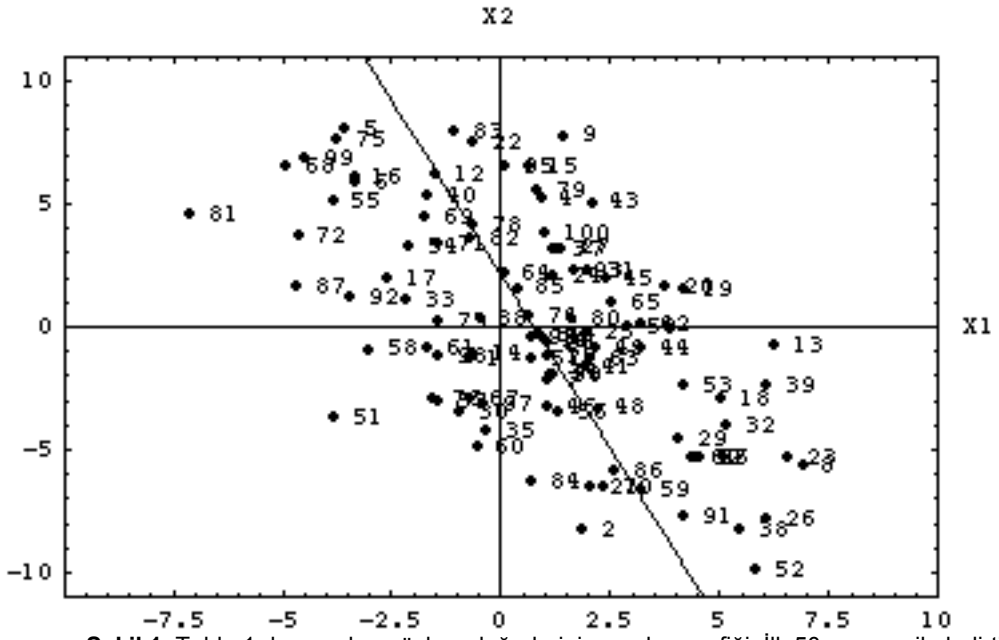
Birinci Kitleden Gözlemler		Atama Kuralları			İkinci Kitleden Gözlemler		Atama Kuralları		
		SD	HD	LDA			SD	HD	LDA
1.134	0.6909	1	1	1	-0.8760	0.0238	2	2	2
-0.4645	-2.7784	1	1	2	4.6468	-0.7379	2	2	1
2.3162	3.9511	1	1	1	0.9779	-1.3265	1	1	1
0.1253	1.0081	2	2	2	0.0733	-0.4935	2	2	2
-1.6251	-4.9915	1	1	2	-2.2918	-1.2898	2	2	2
1.7925	2.2600	1	1	1	-2.0556	-0.6007	2	2	2
-1.0465	-0.9614	2	2	2	1.2989	2.2337	2	2	2
2.1777	-1.0900	1	1	1	0.1431	1.6392	2	2	2
1.3208	-0.9927	1	1	1	-3.0897	-1.3468	2	2	2
3.3870	2.3254	1	1	1	0.5798	4.3926	2	2	2
0.8222	3.4577	1	2	2	0.8846	0.6838	2	2	1
3.0138	4.0553	1	1	1	0.4031	-0.1566	2	2	2
1.9398	0.0917	1	1	1	1.5195	1.7070	1	1	1
-0.3369	-0.5106	2	2	2	0.3071	0.3585	2	2	2
2.4169	1.8536	1	1	1	-1.8734	-0.3048	2	2	2
2.6577	5.8252	1	1	1	-3.9781	-2.9494	2	2	2
0.9082	-0.2010	1	1	1	-1.3610	0.4125	2	2	2
1.0819	-3.5142	1	1	1	-1.1275	-0.4687	2	2	2
0.3164	-4.0082	1	1	1	0.9733	0.5705	1	2	1
3.8829	4.0750	1	1	1	2.4442	4.8161	2	1	1
0.2933	-0.1749	2	2	2	1.4179	1.2183	1	1	1
1.9710	3.3710	1	1	1	-1.2301	2.5640	2	2	2
0.0093	-1.8210	1	1	2	-1.1178	-0.1959	2	2	2
0.4429	-3.8319	1	1	1	0.8023	-0.4573	1	1	1
-0.1233	-0.2960	2	2	2	0.4176	1.5335	2	2	2
4.3100	5.0721	1	1	1	1.2611	-1.3651	1	1	1
0.3537	-1.4731	1	1	1	-1.7038	-3.5307	2	2	2
4.1380	4.6653	1	1	1	1.4716	1.2647	1	1	1
0.9536	-1.0996	1	1	1	0.9170	4.8288	2	2	2
-0.2211	-1.6736	2	2	2	0.7941	-2.0762	1	1	1
0.8894	-2.0188	1	1	1	1.5367	2.8465	1	1	1
5.2639	8.0988	1	1	1	-0.4611	0.5477	2	2	2
2.0756	7.5233	1	1	2	-1.0317	0.3647	2	2	2
2.8763	1.1551	1	1	1	0.7909	-0.9398	1	1	1
3.3283	1.5615	1	1	1	-2.3534	-2.9345	2	2	2
2.8343	1.0905	1	1	1	-1.6156	-5.0212	2	2	2
1.9620	-0.3459	1	1	1	-0.0076	-1.8539	2	1	2
1.3760	3.9584	1	1	2	1.1183	0.8222	1	1	1
3.2387	0.7427	1	1	1	0.3852	1.1948	2	2	2
-0.0920	-1.9048	1	1	2	-0.0853	3.5854	2	2	2
1.2748	0.0688	1	1	1	-1.7820	-0.1207	2	2	2
2.1701	5.4523	1	1	2	-0.0266	-0.4367	2	2	2
4.5153	1.3886	1	1	1	1.7736	1.4802	1	1	1
3.0257	-0.1261	1	1	1	-0.8610	-1.8575	2	2	2
1.7833	5.6126	1	1	2	-0.3070	-1.8924	2	2	2
-0.5285	-3.9126	1	1	2	-0.6418	-3.6500	2	2	2
2.4824	-1.3428	1	1	1	3.0164	5.6079	2	1	1
-0.7217	-4.3052	1	1	2	0.8289	3.4657	2	2	2
1.6625	1.6384	1	1	1	1.8061	2.8609	1	1	1
0.9044	0.5314	1	1	1	0.2888	-0.1714	2	2	

Tablo 1 ve 2'den görülebileceği gibi, SD ve HD derinliklerine göre yanlış atanan gözlemleri, genelde LDF de yanlış atamaktadır. Ancak bu iki derinlik ölçütüne göre doğru atanan bazı gözlemleri LDF yanlış atayabilmektedir. Şekil 1 ve 2'den bu tür gözlemlerin örneklem lineer diskriminant fonksiyonundan elde edilen doğrunun etrafında yer alan gözlemler olduğu tesbit edilmiştir.

Tablo 3 : Atama yöntemlerine ilişkin elde edilen hata oranları

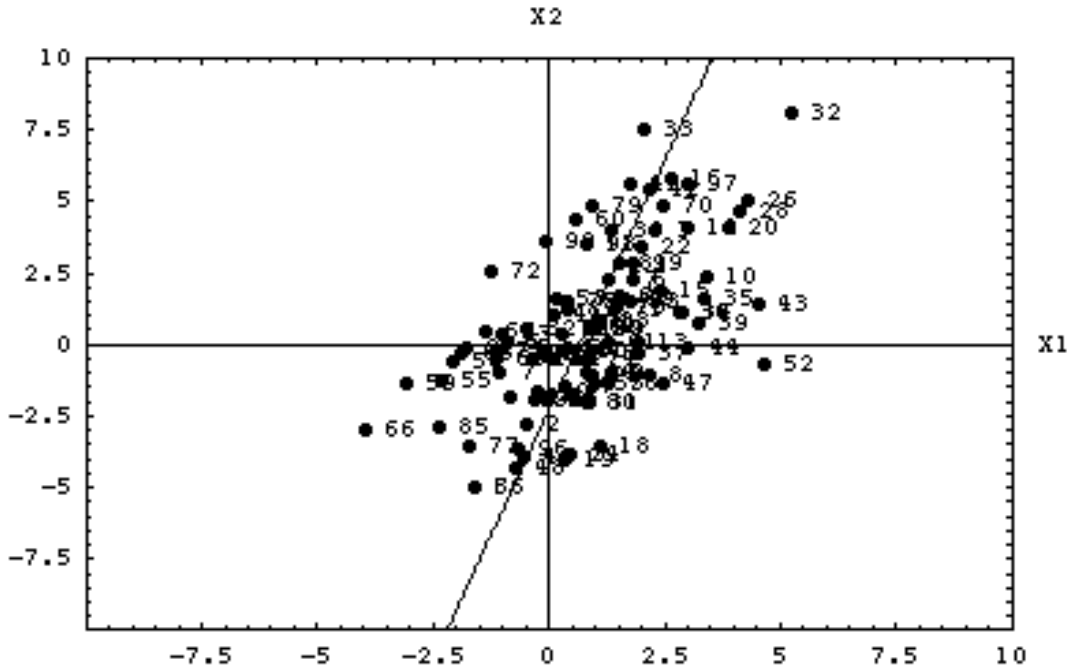
	Tablo 1 deki gözlemler için $\Delta^2 = 0.4329$, $\hat{\Delta}^2 = 1.1993$	Tablo 2 deki gözlemler için $\Delta^2 = 0.7555$, $\hat{\Delta}^2 = 0.6074$
$\alpha(\xi)$	0.3707	0.3300
$\alpha(\hat{\xi})$	0.3751	0.3345
$\hat{\alpha}(\hat{\xi})$	0.3400	0.3700
SD	0.1200	0.1900
HD	0.2400	0.2200

Tablo 3 de, 2. ve 3. alt bölümlerde verilen yöntemlere ilişkin sonuçlar her iki uygulama için verilmiştir. SD derinlik ölçüsüne göre elde edilen hata oranı tahmin değeri daha küçüktür. Bunu sırasıyla HD derinlik ölçütü ve LDF izlemektedir. SD'ye göre elde edilen hata oranı tahmin değerinin, HD derinliğine göre elde edilen değerden daha küçük olmasının nedenlerinden birinin HD derinliğinde aynı örneklem derinlik konturu(izdüşüm çizgileri) üzerinde yer alan eş derinlikli gözlemlerin çok sayıda olduğu düşünülmektedir. Örneklem büyüklükleri artıkça, ikisi arasındaki farklılaşmanın da azalacağı sonucu çıkarılabilir.



Şekil 1. Tablo 1 de yer alan gözlem değerlerinin saçılım grafiği. İlk 50 numara ile belirtilen

gözlemler birinci kitleye, son 50 numara ile belirtilen gözlemler ikinci kitleye aittir.



Şekil 2. Tablo 2 de yer alan gözlem değerlerinin saçılım grafiği. İlk 50 numara ile belirtilen gözlemler birinci kitleye, son 50 numara ile belirtilen gözlemler ikinci kitleye aittir.

Kaynaklar

[1] Liu, R. Y. , Parelius, J. M. and Singh, K., " Multivariate analysis by data depth: Descriptive statistics, graphics and inference", Ann. Statist.,Vol. 27, No.3.,783-858, (1999).
[2] Liu, R. Y., " On a notion of data depth based on random simplices", Ann. Statist., Vol.18, No.1, 405-414, (1990).
[3] Liu, R., Y. and Singh, K., " Ordering directional data : Concepts of data depth on circles and spheres", Ann. Statist. Vol.20, No.3, 1468-1484, (1992).
[4] Welch, B. L., " Note on the discriminant functions", Biometrica, 31, 218-220, (1939).
[5] Anderson, T. W,. An Introduction to multivariate satistical analysis, 2 nd ed., Wiley., (1984).
[6] Lachenbruch, P. A, Discriminant analysis., New York, Hafner Press, (1975).
[7] Hills, M., " Allocation rules and their error rates", Applied Statist. (JRSS-B), No. 28, 1-20, (1966).
[8] Lachenbruch, P. A. and Mickey, M. R., " Estimation of error rates in discriminant analysis", Technometrics, 10, 1-11, (1968).
[9] Seber, G. A. F., Multivariate observations, John Wiley& Sons. Inc., (1984).
[10] Wald, A., " On a statistical problem arising in the classification of an individual iInto one of two groups", Annals of Mathematical Statistics, 15, 145-162, (1944).
[11] Smith, C. A. B.," Some examples of discrimination", Annals of Eugenics, 18, 272-282, (1947).
[12] Rousseeuw, P. J., Ruts, I. and Tukey, J. W., "The bagplot: A bivariate boxplot", The American Statistician, Vol.53, No. 4, 382-387, (1999).
[13] Liu, R., Y. and Singh, K., " A quality index based on data depth and multivariate rank tests", J. Amer. Statist. Assoc. Vol. 88, No.421, 252-260, (1993).
[14] Serfling, R., "Generalized quantile processes based on multivariate depth functions, with applications in nonparametric multivariate Analysis", J. Multivariate Anal., Vol.83, No.1, 232-247, (2002).
[15] Zuo, Y. , Serfling, R., "General notions of statistical depth functions", Ann. Statist., Vol.28, No.2, 461-482, (2000a).
[16]Donoho, L. D., Gasko, M., "Breakdown properties of location estimates based on halfspace depth and projected outlyingness", Ann. Statist. Vol. 20, No. 4, 1803-1827, (1992).
[17] Zuo, Y. , Serfling, R., " Structural properties and convergence results for contours of sample statistical septh sunctions", Ann. Statist., Vol.28, No.2, 483-499, (2000b).

[18]Rousseeuw, P. J., Ruts, I. Algorithm AS 307: bivariate lacation depth. Applied Statist. (JRSS-C) 45, No. 4, 516-526. (1996).