Uludağ Üniversitesi
Eğitim Fakülteleri Dergisi
Cilt: VI, Sayı: 2, 1991

# DAĞILIMA BAĞLI SINIFLANDIRMA

Hasan SOYDAN*

## ÖZET

*İstatistiğin en önemli problemlerinden biri de, iki sınıftan birine ait olduğu bilinen bir nesnenin, en küçük risk ile hangi sınıfa ait olduğunu tayin etmektir.*

*Bu makalede, dağılıma bağlı sınıflandırma yöntemi ile kütle parametrelerinin bilindiği ve bilinmediği durumlarda ayırımın nasıl yapılması gerektiği ele alınmıştır.*

## SUMMARY

### Classification Depending on the Population the Distribution

*One of the most important problems of statistics is to determine with the minimum risk to which population an object belongs to when it is known as a member of two population.*

*In this paper, it is pointed out the necessity of how to make discriminations under sircumstances of known and unknown population characteristics by using the method of classifications depending on the population distribution.*

---

* *U.Ü. Necatibey Eğitim Fakültesi, Fen Bilimleri Eğitimi Bölümü, Öğretim Üyesi.*

Dağılıma bağlı diskriminasyon yöntemlerinin en kolay olanı, sınıflandırılması istenen nesnelerin geldiği populasyonun normal dağılımlı olduğunun bilinmesi veya kabul edilmesi ile ortaya çıkan yöntemdir.

Sınıflandırılması istenen birbirinden ayrı iki populasyon $A_1$ ve $A_2$ olarak adlandırılsın. Bu iki populasyonun p boyutlu normal dağılımlı olduğu kabul edilsin. $\mu_i$, p boyutlu beklenen diğer vektörü, S de pxp boyutlu bir varyans vektörünü göstermek üzere $A_i \subseteq N(\mu_i, s)$ $(i = 1, 2)$ olsun. Bu durumda p boyutlu uzaydan rastgele seçilecek bir x noktasının $A_1$ ve $A_2$ populasyonlarından hangisinden geldiğini tahmin etmek için şu şekilde bir yöntem düşünülür. $A_i$ populasyonlarının sıklık fonksiyonları $p_i(x)$ $(i=1, 2)$ ile tanımlanır. Bir R karar kavramı ile dizayn uzayı $R_1$ ve $R_2$ bölgelerine ayrılır. $x \varepsilon R_i$ olması halinde x'in $A_i$ populasyonuna ait olduğu kabul edilir. Bu halde sınıflandırmanın doğru olması olasılığı,

$$(1) \qquad p(i/i, R) = \int_{R_i} p_i(x)\, dx \qquad i = 1, 2$$

dır.

x'in yanlış sınıflandırılması iki farklı türde olur. Gerçekte $x \varepsilon A_2$ dir, ama R kuramı ile $x \varepsilon A_1$ kararı verilmiştir. Ya da bunun tersine gerçekte $x \varepsilon A_1$ dir ama $x \varepsilon A_2$ kararı verilmiştir. Bu iki karara zarar densin ve sıra ile bu zararlar C(1/2) ve C(2/1) ile gösterilsin. Bu durumda $x \varepsilon A_i$ olmanın riski R (i) ile gösterilirse,

$$(2) \qquad \begin{aligned} R(1) &= C(2/1) \cdot p(2/1, R) \\ R(2) &= C(1/2) \cdot p(1/2, R) \end{aligned}$$

olur.

Arka arkaya seçilen pek çok x noktası hakkında karar vermek düşünülebilir. Bu nedenle i inci defada seçilen x noktası $x_i$ ile gösterilsin. $x_i$ noktalarının $q_i$ apriori olasılıkları belli ise ve bu ihtimaller $q_i \varepsilon N(0, 1)$ olup $\Sigma q_i = 1$ eşitliği sağlanıyorsa R karar kuramının Bayes risk ölçüsü,

$$(3) \qquad B(R) = \sum_i q_i R(i)$$

olur. $i = 1, 2$ için bu risk

$$(4) \qquad B(R) = C(2/1)\, P(2/1)\, q_1 + C(1/2)\, P(1/2)\, q_2$$

olur.

Bu son eşitlikte C(1/2) ve C(2/1) zararları en büyük ve C(1/2) = C(2/1) = 1 kabul edilirse, hem iki olasılıklı seçme daha da basit hale gelir, hem de verilen bir x noktası için yanlış sınıflandırılma olasılığı, ancak daha büyük

şartlı olasılık veren sınıfa ait olduğu kabul edilerek azaltılabilir. Diğer bir deyimle,

$$(5) \qquad \frac{q_1 p_1(x)}{q_1 p_1(x) + q_2 p_2(x)} > \frac{q_2 p_2(x)}{q_1 p_1(x) + q_2 p_2(x)}$$

eşitsizliği doğru ise $x \varepsilon A_1$ olmasını kabul etmekle risk en aza indirilmiş olur. Yanlış sınıflandırma olasılığı her nokta için en aza indirildiğinden, iki sınıfı da kapsayan bütün uzay için de kendiliğinden minimize edilmiş olur. (5) eşitsizliği ve yukarıdaki ifade özetlenirse, x noktası,

$$(6) \qquad q_1 p_1 (x) > q_2 p_2 (x) \quad \text{ise } A_1 \text{ sınıfında}$$

$$(7) \qquad q_2 p_2 (x) > q_1 p_1 (x) \quad \text{ise } A_2 \text{ sınıfında}$$

olmaktadır.

$q_1 p_1 (x) = q_2 p_2 (x)$ ise seçim her türde olabilir.

$C(1/2) = 1$, $C(2/1) = 1$ varsayımları kullanılmadan (6) eşitsizliği,

$$(8) \qquad C(2/1)\, q_1 p_1 (x) > c(1/2)\, p_2 q_2 (x)$$

ve (7) de, tersine dönüşür.

$C(i\ j) \geq 0$ ve $q_i \geq 0$ olduklarından (8) eşitsizliği,

$$(9) \qquad \frac{p_1(x)}{p_2(x)} > \frac{C(1/2)\, q_2}{C(2/1)\, q_1}$$

şeklinde yazılabilir. (9) eşitsizliği doğru ise x, $A_1$ sınıfında aksi halde $A_2$ de olur.

Ayırımı yapılacak sınıfların dağılımları normal olduğundan,

$$(10) \qquad p_i (x) = (2\pi)^{-p/2}\, S^{-1/2} \exp \{-\frac{1}{2} (x - \mu_i)^T S^{-1} (x - \mu_i)\}$$

bulunur. $(x - \mu_i)^T$, $x - \mu_i$ 'nin transpozesi olan sütun (matrisi) vektörünü göstermektedir. $p_i (x)$'in (10) daki değeri $i=1$ ve $i=2$ için yazılıp (9)'da yerine yazılırsa,

$$(11) \qquad \frac{\exp\{-\frac{1}{2}(x-\mu_i)^T S^{-1}(x-\mu_1)\}}{\exp\{-\frac{1}{2}(x-\mu_2)^T S^{-1}(x-\mu_2)\}} > \frac{C(1/2)\, q_2}{C(2/1)\, q_1}$$

veya

$$(12) \quad \exp\{-\frac{1}{2}(x-\mu_1)^T S^{-1}(x-\mu_1) - (x-\mu_2)^T S^{-1}(x-\mu_2)\} > \frac{C(1/2)\, q_2}{C(2/1)\, q_1}$$

bulunur. (12) de $C(1/2)\, q_2 / C(2/1)\, q_1 = C'$ olsun. Ayrıca yine (12)'nin sol yanı daima pozitif ve monotonik olan bir fonksiyon olduğundan (12) eşitsizliğinin yanları yerine bu yanların logaritmalarını yazılabilir. Bu durumda

$$(13) \quad -\frac{1}{2}\{(x-\mu_1)^T S^{-1}(x-\mu_1) - (x-\mu_2)^T S^{-1}(x-\mu_2)\} > \ln C'$$

bulunur.

Burada da

$$(14) \qquad u(x) = -\frac{1}{2}\{(x-\mu_1)^T S^{-1}(x-\mu_1) - (x-\mu_2)^T S^{-1}(x-\mu_2)\}$$

denirse (13) eşitsizliği

$$(15) \qquad u(x) > \ln C'$$

olur. $u(x) > \ln C'$ eşitsizliğini sağlayan x noktası en az riskle $A_1$ de, $u(x) < \ln C'$ 'ü sağlayan x noktası da yine en az riskle $A_2$'de bulunur.

(14) eşitliği matris işlemleri ile,

$$U(x) = -\frac{1}{2}\{(x^T-\mu_1^T) S^{-1}(x-\mu_1) - (x^T-\mu_2^T) S^{-1}(x-\mu_2)\}$$

$$= -\frac{1}{2}\{x^T S^{-1} x - x^T S^{-1}\mu_1 - \mu_1^T S^{-1} x + \mu_1^T S^{-1}\mu_1$$

$$- x^T S^{-1} x + x^T S^{-1}\mu_2 + \mu_2^T S^{-1} x - \mu_2^T S^{-1}\mu_2\}$$

$$= \frac{1}{2}\{x^T S^{-1}(\mu_1-\mu_2) + (\mu_1-\mu_2)^T S^{-1} x - (\mu_1+\mu_2) S^{-1}(\mu_1-\mu_2)\}$$

veya

$$u(x) = x^T S^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2) S^{-1}(\mu_1-\mu_2) \quad (16)$$

şekline dönüşür. Bu eşitliğin ilk terimi x noktasının bir fonksiyonudur. x noktası p boyutlu bir vektörle ifade edildiğinde bu fonksiyonun doğrusal bir fonksiyon olduğu görülür.

$x \varepsilon A_1$ olduğunda,

$$E_1[u(x)] = \frac{1}{2}(\mu_1-\mu_2)^T S^{-1}(\mu_1-\mu_2)$$

$$V_1[u(x)] = (\mu_1-\mu_2) S^{-1}(\mu_1-\mu_2) \quad (17)$$

olurlar. Burada $(\mu_1-\mu_2)^T S^{-1}(\mu_1-\mu_2) = \propto$ denirse;

$$E_1[u(x)] = \frac{1}{2} \propto \; ; v_1[u(x)] = \propto$$

olur. u(x) normal dağılımlı elemanların bir birleşimi olduğundan kendisi de normal dağılımlıdır. O halde u(x), beklenen değeri $\propto/2$ varyansı $\propto$ olan bir normal dağılımdır.

$x \varepsilon A_2$ olduğunda da u(x) ortalaması $-\propto/2$ varyansı da $\propto$ dır ve yine normal dağılımlıdır.

Bu yaklaşımla $x \varepsilon A_1$ olması halinde yanlış sınıflandırma olasılığı,

$$P(1/2) = \int_{-\infty}^{C} \frac{1}{\sqrt{2\pi}\propto} e^{-\frac{(z-\propto/2)^2}{2\propto}}$$

veya $\frac{1}{1/2}(z - \frac{\propto}{2})$ = 't dönüşümü ile

(18) $$P(1/2) = \int_{-\infty}^{(C-\alpha/2)/\alpha^{1/2}} \frac{1}{\sqrt{2\pi}} e^{-1/2\, t^2} dt$$

x $\varepsilon A_2$ olması halinde yanlış sınıflandırma olasılığı da;

$$P(2/1) = \int_{C}^{\infty} \frac{1}{(2\pi\alpha)^{1/2}} e^{-\frac{(z-\alpha/2)^2}{2\alpha}} dz$$

veya yine yukarıdaki dönüşümle;

(19) $$P(2/1) = \int_{(C-\alpha/2)/\alpha^{1/2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-1/2\, t^2} dt$$

elde edilir.

(18) ve (19) eşitliklerinde kullanılmış olan C sayısı riski en aza indirmek için bulunması gereken bir değerdir.

Minimaks kuralına göre C sabiti en büyük riski en aza indirmek için,

(20) $$C(1/2)\, p\,(1/2) = C(2/1)\, p\,(2/1)$$

olacak şekilde seçilmelidir. Bu da ancak C = log C' ile sağlanır.

$\mu_1$, $\mu_2$ ve S populasyon parametrelerinin bilinmediği durumlarda, hangi sınıfa ait oldukları kesin olarak bilinen noktalardan oluşan örneklerin parametreleri $A_j$ örneklerin hacimleri $N_j$ (j = 1, 2) olmak üzere,

$$x_i = (x_{i1}, x_{i2}, ...., x_{ip}) \text{ için}$$

$$\mu_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x^j_{ik}$$

(21) $$\hat{\mu}_{jk} = \frac{1}{N_j} \sum_{i=1}^{N_j} x^j_{ik} \qquad h = 1, 2, ..., p$$

$$\hat{s}_{ij} = \frac{1}{N_1+N_2-2} \sum_{k=1}^{N_1} (x^1_{ki} - \hat{\mu}_{1i}) (x'_{kj} - \hat{\mu}_{1j})$$

$$(22) \qquad + \frac{1}{N_1+N_2-2} \sum_{k=1}^{N_2} (x^2_{ki} - \hat{\mu}_{2i}) (x^2_{kj} - \hat{\mu}_{2j})$$

ile hesaplanır. Çünkü bu parametrelerle,

$$\mu_j = (\mu_{j1}, \mu_{j2}, ..., \mu_{jp}) \qquad E(\mu_j) = \mu_j \qquad E(\hat{s}) = s$$

dir.

Görülüyor ki; parametrik ayırım yönteminde karar denklemi C katsayısına bağlıdır. C(1 2) $\neq$ C(2 1) halinde bu sayı normal dağılım tablolarından deneme ile bulunabilmektedir. Populasyon parametrelerinin bilinmediği durumlarda populasyon parametrelerinin, hangi sınıfa ait olduğu kesin olarak bilinen noktalardan elde edilen örneklerin parametrelerinin örnekleme dağılımından yararlanarak bulunması gerekmektedir.

## KAYNAKLAR


1. ANDERSON, T.W.: An Introduction to Multivariate Statistical Analysis, J. Wiley, Newyork, 1958.
2. JACKSON, E.C.: Missiry Values in Linear Multiple Discriminant Analysis. Biometrics, 1968.
3. SEN GUPTA, S.K.: Aplication of Discriminatory Analissis to an Engineering Problem, Statistician, 1970.