İstatistik Araştırma Dergisi Cilt 01 No 01 Sayfa 131-137 Nisan 2002

Tiroit Bezi Verilerinin Bayes ve En Yakın K-Komşu Gibi Eğiticili Yöntemlerle Sınıflanması

Songül ALBAYRAK*

ÖZET

Bu makalede tiroit bezi fonksiyonlarına ait verilerin eğiticili yöntemlerle kümelenmesi konusu karşılaştırmalı olarak incelenmiştir. Eğiticili yöntemler arasında Bayes sınıflama ve En yakın 1-Komşu (1-Nearest Neighbour) ile sınıflama incelenerek sonuçlar karşılaştırılmıştır. 215 hastaya ait veriler 5 boyutlu özellik vektörü şeklinde alınıp belirtilen yöntemler uygulanmış ve örnekler üç ayrı sınıfa ayrılmıştır.

Anahtar Kelimeler: *Bayes sınıflayıcı, En yakın k-komşu, Tiroit bezi verileri*

1. GİRİŞ

Bilimsel çalışmalarda yapılan deneyler ve ölçmeler sonucunda, işlenmesi ve yorumlanması gereken verinin miktarında büyük bir artış olmuştur. Teknolojik gelişmelere paralel olarak özellikle tıp alanında belli hastalıklara ait verilerin toplanması için kullanılan cihazlar hızla gelişmiş ve elde edilen verinin değerlendirilmesinde bilgisayar kullanımı kaçınılmaz olmuştur. Tiroit bezi fonksiyonlarını ölçmek üzere Dr. Coomans (1983) tarafından 215 hasta üzerinde laboratuar testleri yapılmış ve bu test sonuçları daha sonra da pek çok araştırmacı tarafından kullanılmıştır (Zhang, 2000).

Bu makalede tiroit bezi fonksiyonları ile ilgili olarak 215 hastadan toplanan verilerin eğiticili yöntemlerle kümelenmesi konusu incelenecek ve sonuçları karşılaştırılacaktır. Tiroit bezi fonksiyonlarını ölçmek için yapılan 5 ayrı test sonucunda hastalar **normal** ω_1, **hipotiroit** ω_2 ve **hipertiroit** ω_3 şeklinde üç sınıfta toplanacaktır.

Tiroit bezi fonksiyonlarını ölçmek amacıyla hastalara uygulanan 5 test sırasıyla aşağıda verilmiştir. Bu testler sonucunda her hastaya ait özellik vektörü $x=[x_1,x_2,x_3,x_4,x_5]$ şeklinde 5 boyutlu olarak gösterilmiştir.

1-T3-resin uptake testi (yüzde sonuç gösterir)
2-Isotopik gösterge metoduyla ölçülen toplam serum thyroxin
3-Radioimmuno deneyi ile ölçülen toplam serum triiodothyronine
4- Radioimmuno deneyi ile ölçülen basal thyroid-stimulating hormone (TSH)
5-200 mikro gram thyrotropin-releasing hormonu enjekte edildikten sonraki TSH değeri ile basal değer arasındaki mutlak farkdır.

* Arş.Gör.Dr., Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayara Mühendisliği Bölümü, Beşiktaş-İSTANBUL

215 hastaya uygulanan laboratuar testleri sonucunda bu hastaların 150'sinin normal, 35'inin hipertiroit ve 30'unun hipotiroit olduğu belirtilmiştir. Eğiticili sınıflama yöntemlerinde amaç bu ön bilgilerden de yararlanarak hastanın hangi sınıfa dahil olduğunu bulmaktır. Eğiticisiz kümeleme yöntemlerinde ise amaç sadece test sonuçlarına bakarak hastanın hangi sınıfa dahil olduğunu bulmaktır.

2. EĞİTİCİLİ SINIFLAMA

Örüntü tanıma uygulamalarının çoğunluğunda her sınıfı temsil eden tipik örneklerin olduğu kabul edilir. Bu durumda eğiticili örüntü tanıma algoritmaları uygulanabilir. Eğiticili bir ortamda, değişik uyarlanabilir yöntemlerle, sistem örüntüleri tanımak için eğitilir. Bu yaklaşımın temeli hangi sınıfa ait olduğu bilinen örneklerin bir kısmının eğitim seti olarak ayrılması ve öğrenme işleminin bu set üzerinde yapılması şeklinde özetlenebilir. Bu bölümde Bayes ve En yakın k-komşu yöntemleri incelenecektir.

2.1. Bayes Metodu ile Sınıflama

Bayes karar teorisi örüntü tanıma problemlerine istatistik yaklaşımla çözüm getiren temel yöntemlerdendir.

Tiroit bezi fonksiyonları ile ilgili problemde üç sınıf oluşturulmaktadır. Bayes sınıflamanın en temel kabulü $P(\omega_1)$, $P(\omega_2)$ ve $P(\omega_3)$ olasılıklarının bilinmesidir. Pratikte bu bilgi mevcut eğitim setinden de çıkarılabilir. N eğitim setinin toplam sayısı, N_1,N_2 ve N_3 ise ω_1,ω_2 ve ω_3 sınıflarına ait örneklerin sayısı göstermektedir. ω_1,ω_2 ve ω_3 sınıflarına ait olasılıklar sırasıyla aşağıdaki gibi yazılabilir.

$$P(\omega_1) = \frac{N_1}{N} = \frac{150}{215} = 0.698$$

$$P(\omega_2) = \frac{N_2}{N} = \frac{35}{215} = 0.163$$

$$P(\omega_3) = \frac{N_3}{N} = \frac{30}{215} = 0.140$$

Her sınıftaki özellik vektörlerinin dağılımından yararlanılarak, sınıfa ait sınıf-koşullu olasılık yoğunluk fonksiyonları $p(x|\omega_i)$, i=1,2,3 bulunabilir. Bu ifadelerden yararlanılarak Bayes kuralı şu şekilde tanımlanabilir;

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

Burada p(x), x'in olasılık yoğunluk fonksiyonudur. 5 boyutlu özellik uzayında ω_i sınıfının olasılık yoğunluk fonksiyonunun normal dağılım gösterdiği kabul edilirse, sınıf-koşullu olasılık yoğunluk fonksiyonu şöyle yazılabilir;

$$p(x|\omega_i) = \frac{1}{(2\pi 2^{5/2}|\Sigma_i|^{1/2}} \exp(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)), \quad i=1,2,3$$

Burada μ_i=E(x), ω_i sınıfının **ortalamasını** ve Σ_i, ω_i sınıfının 5x5 **kovaryans** matrisini göstermektedir.

$$\mu_i = E[x] = E\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix} \qquad \Sigma_i = E[(x-\mu_i)(x-\mu_i)] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{bmatrix}$$

Her ω_i i=1,2,3 sınıfı için 5 boyutlu bir ortalama vektörü ve 5x5 boyutunda bir kovaryans matrisi hesaplanır. Kovaryans matrisi simetrik bir matristir ve 5x5'lik (d=5; 5 boyutlu) matriste 15 adet değer hesaplanmalıdır. Kovaryans matrisinde hesaplanacak eleman sayısı $(d^2+d)/2$ bağıntısı ile bulunabilir.

Kovaryans matrisinin elemanları her sınıf için ayrı ayrı aşağıdaki formda hesaplanır.

$$\sigma_{jk} = Cov(x_j, x_k) = E[x_j.x_k] - \mu_j\mu_k$$

$$\sigma_{jj} = Cov(x_j, x_j) = E[x_j.x_j] - \mu_j\mu_j = \sigma_j^2$$

$$\Sigma_1 = \begin{bmatrix} 65.17 & 4.81 & 1.13 & -0.19 & 3.04 \\ 4.81 & 4.17 & 0.25 & -0.01 & -0.62 \\ 1.13 & 0.25 & 0.22 & 0 & 0.06 \\ -0.19 & -0.01 & 0 & 0.25 & 0.08 \\ 3.04 & -0.62 & 0.06 & 0.08 & 3.88 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} 342.0 & -20.96 & -22.08 & -0.77 & 1.38 \\ -20.96 & 16.82 & 4.09 & 0.32 & 0 \\ -22.08 & 4.09 & 4.94 & 0.07 & -0.03 \\ -0.77 & 0.32 & 0.07 & 0.16 & -0.02 \\ 1.38 & 0 & -0.03 & -0.02 & 0.07 \end{bmatrix}$$

$$\Sigma_3 = \begin{bmatrix} 118.20 & -6.78 & -1.02 & 22.24 & -16.69 \\ -6.78 & 2.98 & 0.56 & -11.84 & 8.9 \\ -1.02 & 0.56 & 0.30 & -2.73 & 2.51 \\ 22.24 & -11.84 & -2.73 & 148.30 & 18.45 \\ -16.69 & 8.9 & 2.51 & 18.45 & 232.40 \end{bmatrix}$$

Mevcut örneklerin hangi sınıfa ait olduğunu belirlemek için bir **diskriminant fonksiyonu** tanımlanmalıdır.

$$g_i(x) = \ln p(x|\omega_i) + \ln P(\omega_i) \quad i = 1,2,3$$

Her sınıfa ait ortak değişinti matrisi **birbirinden farklı** olduğuna göre diskriminant fonksiyonu aşağıdaki forma indirgenebilir;

$$g_i(x) = x^T W_i x + w_i^T x + \omega_{i0} \quad i = 1,2,3$$

$$W_i = -\frac{1}{2}\Sigma_i^{-1} \qquad w_i = \Sigma_i^{-1}\mu_i \qquad \omega_{i0} = -\frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\log|\Sigma_i| + \log P(\omega_i)$$

$$W_1 = \begin{bmatrix} -0.0094 & 0.0101 & 0.0340 & -0.0099 & 0.0087 \\ 0.0101 & -0.1436 & 0.1187 & 0.0121 & -0.033 \\ 0.0340 & 0.1187 & -2.5413 & 0.0415 & 0.0322 \\ -0.0099 & 0.0121 & 0.0415 & -2.0580 & 0.0524 \\ 0.0087 & -0.0330 & 0.0322 & 0.0524 & -1.1427 \end{bmatrix} \quad w_1 = \begin{bmatrix} 1.7658 \\ 0.1274 \\ -1.1613 \\ 6.9796 \\ -0.8447 \end{bmatrix}$$

$$w_{10} = -102.9148$$

$$W_2 = \begin{bmatrix} -0.0023 & -0.0005 & -0.0095 & 0.0006 & 0.0408 \\ -0.0005 & -0.0388 & 0.0290 & 0.0705 & 0.0467 \\ -0.0095 & 0.0290 & -0.1670 & -0.0109 & 0.1159 \\ 0.0006 & 0.0705 & -0.0109 & -3.5448 & -1.2459 \\ 0.0408 & 0.0467 & 0.1159 & -1.2459 & -8.2514 \end{bmatrix} \quad w_2 = \begin{bmatrix} 0.5332 \\ 1.0957 \\ 2.2306 \\ 4.3281 \\ -8.3279 \end{bmatrix}$$

$$w_{20} = -46.3740$$

$$W_3 = \begin{bmatrix} -0.0049 & -0.0135 & 0.0047 & -0.0003 & 0.0001 \\ -0.0135 & -0.4045 & 0.3755 & -0.0249 & 0.0124 \\ 0.0047 & 0.3755 & -2.7108 & -0.0227 & 0.0170 \\ -0.0003 & -0.0249 & -0.0227 & -0.0059 & 0.0017 \\ 0.0001 & 0.0124 & 0.0170 & 0.0017 & -0.0029 \end{bmatrix} \quad w_3 = \begin{bmatrix} 1.2803 \\ 5.6074 \\ 1.9027 \\ -0.3901 \\ -0.0989 \end{bmatrix}$$

$$w_{30} = -99.5425$$

Belirlenen **diskriminant fonksiyonu** ile Bayes sınıflayıcı şöyle çalışır;

Eğer $g_1(x)>g_2(x)$ ve $g_1(x)>g_3(x)$ ise x örneği ω_1 sınıfına aittir,
Eğer $g_2(x)>g_1(x)$ ve $g_2(x)>g_3(x)$ ise x örneği ω_2 sınıfına aittir,
Eğer $g_3(x)>g_1(x)$ ve $g_3(x)>g_2(x)$ ise x örneği ω_3 sınıfına aittir.

2.2. En Yakın K-Komşu (k-Nearest Neighbors: K-NN) ile Sınıflama

En yakın k-komşu (K-NN) tekniği parametrik olmayan bir kestirim metodudur ve doğrudan sınıflama için çok pratik bir yaklaşımdır (Tou ve Gonzalez,1974; Schalkoff,1992). K-NN yöntemi eğitim seti formunda bir ön bilgiye ihtiyaç duyar. Bu yöntemde $\{s_1,s_2,\ldots,s_N\}$ şeklinde bir grup örneğin eğitim seti olarak alınması ve bu örneklerin her birinin $\omega_1,\omega_2,\ldots,\omega_M$ sınıflarından hangisine ait olduğu ön bilgisinin verilmesi gerekir.

En yakın-bir-komşu yöntemiyle sınıflama, hangi sınıfa ait olduğu bilinmeyen bir x örneğinin, kendisine en yakın komşunun ait olduğu sınıfa dahil edilmesi şeklinde özetlenebilir. Burada x örneğine en yakın komşu $s_i \in \{s_1,s_2,\ldots,s_N\}$ olsun;

$$d(s_i,x) = \min_l \{d(s_l,x)\}, \quad l=1,2,\ldots,N$$

d iki vektör arasındaki mesafeyi göstermektedir ve bu uygulamada mesafe ölçüsü olarak Euclid mesafesi kullanılmıştır.

En yakın k-komşu yöntemiyle sınıflama aşağıdaki gibi özetlenebilir:

1- N adet örnekten oluşan bir eğitim seti (hangi sınıfa ait olduğu belli) oluşturulur. Eğitim seti dışındaki test örneğine en yakın k adet komşu belirlenir. Uygulamada k sayısı tek sayı olmalıdır.
2- k adet en yakın komşudan her (ω_i) sınıfına ait örnek sayısı (k_i) belirlenir.
3- Test örneği, en fazla elemanı olan sınıfa (ω_i) dahil edilir.

Bu araştırmada en yakın-bir-komşu yöntemi uygulanırken her test örneği için geri kalan örneklerin tamamı eğitim seti olarak alınarak sınıflama yapılmıştır **(leave-one-out method).**

3. SINIFLAMA SONUÇLARI

Bayes ve En yakın 1-komşu yöntemleri makale yazarı tarafından Object Pascal ve MatLab'de kodlanmış ve tiroit bezi fonksiyonlarına ait 215 veri üzerinde sınıflama işlemleri yapılmıştır. En yakın 1-komşu yönteminde, sınıflama yapılan her örnek için geri kalan 214 adet örnek eğitim seti olarak kullanılmıştır. Yapılan sınıflama işlemi sonucunda hatalı sınıflanan örnek sayısı ve hata oranları her iki yöntem için de Tablo 1'de gösterilmiştir.

Tablo 1. Tiroit bezi fonksiyonlarına ait veriler için sınıflama yöntemlerinin hata oranları

	Bayes Sınıflama	En Yakın 1-Komşu
Hatalı Sınıflanan Örnek Sayısı	7	11
Hata Oranı	%3.3	%5.1

4. SONUÇ

Tiroit bezi hastalıkları ile ilgili olarak 215 hasta üzerinde yapılan 5 ayrı ölçmeye bağlı olarak bu hastalar normal, hipotiroit ve hipertirot olmak üzere üç ayrı sınıfa ayrılmıştır. Hastalara ait veriler incelendiğinde sınıfların birbiri içine geçmiş olduğu ve sınıflamanın güç olduğu gözlenmiştir. Eğiticili sınıflama yöntemlerinden Bayes ve en yakın 1-komşu kullanılarak 215 hastaya ait veriler test edilmiş ve Bayes metodu ile sınıflamanın en yakın 1-komşu metoduna göre daha iyi sonuç verdiği görülmüştür.

KAYNAKLAR

COOMANS, D., BROECKAERT, I., JONCKHEER, M. and MASSART, D.L. (1983), *Comparison of Multivariate Discrimination Techniques for Clinical Data*, Application to the Thyroid Functional State, Methods of Information in Medicine, Vol.22, pp.93-101.

KUN Z. (2000), *Pattern Recognition Techniques, Applied to Thyroid Gland Data*, Erişim: [http://www.Sal.ufl.edu/kun/eel6825/report.html]. Erişim Tarihi : 23.01.2001.

TOU, J. T. and GONZALEZ R.C. (1974), *Pattern Recognition Principles*, Addison-Wesley Publishing Company.

SCHALKOFF, R. J. (1992), *Pattern Recognition*: Statistical, Structural and Neural Approaches, John Wiley &Sons, Inc.

Classification of Thyroid Gland Data by Supervised Methods Like Bayes and K-Nearest Neighbour

ABSTRACT

The purpose of this paper is to make a comparison of supervised classification methods based on a set of thyroid gland data. Among the supervised classification methods Bayes Classification and 1-Nearest Neighbour are examined and the results are compared. Data related to 215 patients is obtained as 5 dimensional feature vector and the stated methods are applied to this data and samples are classified into 3 class.

Key Words: *Bayes Classifier, K-nearest neighbour, Thyroid gland data*