Y. Y. Ü. Fen Bilimleri Enstitüsü Dergisi
Yıl : 2009, Cilt : 14, Sayı : 2, Sayfa : 112-116

Yapay Sinir Ağı ve Lojistik Regresyon Kullanılarak Kategorik Verilerin Modellenmesi*

Yılmaz KAYA[1] Abdullah YEŞİLOVA[2]

[1]Yüzüncü Yıl Üniversitesi, Van Meslek Yüksekokulu, Bilgisayar Teknolojileri ve Programcılığı Böl., 65080 Van
[2]Yüzüncü Yıl Üniversitesi, Ziraat Fakültesi, Zootekni Böl., 65080 Van

Özet: Yapay sinir ağları; insan beyninin özelliklerinden olan, öğrenme yolu ile yeni bilgiler türetebilme, oluşturabilme ve keşfedilme gibi yetenekleri herhangi bir yardım almadan otomatik olarak gerçekleştirmek amacı ile geliştirilen yapay zeka uygulamalarıdır. Bu çalışmada verilerin sınıflandırmak için Yapay sinir ağı modeli ile Lojistik regresyon yöntemleri karşılaştırılmıştır. Çalışma sonunda Yapay sinir ağının lojistik regresyona göre verilerin sınıflandırılmasında daha etkin olduğu görülmüştür.
Anahtar kelimeler : Lojistik regresyon, Yapay sinir ağı, Sınıflama.

Modeling Categorical Data By Using Neural Network and Logistic Regression

Abstract: Artificial Neural Network is an application of artificial intelligence developed for the aim of enabling the ability to carry out the features of human brain of deriving, forming and discovering new knowledge via learning without taking any support. In the present study, the artificial neural network and logistic regression methods were compared to classify the data. At the end of the study, the artificial neural network was found to be more effective than logistic regression for data classification.
Key words: Logistic regression, Artificial neural network, classification.

Giriş

Yapay sinir ağı (YSA), hiyerarşik olarak birbirine bağlı ve paralel olarak çalışan yapay sinir hücrelerinden oluşan yapılardır. Her hücreye neron, sinir, düğüm gibi isimler verilmektedir. Sinirleri birbirine bağlayan bağlantıların gücünü, etkisini belirten ağırlık değerleri bulunmaktadır. Sinirlerin birbirine bağlanması ile YSA oluşur (Kröse, 1996). Lojistik regresyon modeli, basit bir şekilde oluşturulabilen sigmoid (lojistik) aktivasyon fonksiyonuna sahip sinir hücresi modeli ile eşdeğerdir. Bu bilgi, lojistik regresyon modeli ile YSA'ı modelinin karşılaştırılması gereğini bazı uygulamalarda önemli kılmaktadır.

YSA uygulamalarda daha çok diskriminant analizi ve lojistik regresyona alternatif bir yöntem olarak kullanılmıştır (Warren, 1994; Claudia ve Foster, 2003; Rocha, 2007). Başka bir ifadeyle, YSA birçok alanda istatistiksel bir araç olarak kullanılmaktadır. Ayrıca, YSA lojistik regresyon ve diskriminant analizinde (Weinstein, 1992; Gallinari, 1993), regresyon ağacı analizinde (Breinman, 1984), genelleştirilmiş eklemeli modellerde (Generalized Addive Models=GAM) (Hastie ve Tibshirani, 1990) ve diğer parametrik olmayan regresyon modellerde yoğun bir kullanım alanına sahiptir (Poli ve Jones, 1994;Gene ve ark., 1997; Lai ve Shing, 2001).

YSA, biyolojik sinirlerden esinlenerek oluşturulmaktadır. YSA bugün birçok probleme çözüm üretebilecek düzeydedir. Tıp, mühendislik gibi birçok alanda başarılı YSA uygulamaları bulunmaktadır. (Elmas, 2003). Sağlık bilimlerinde geniş bir kullanım alanı bulan YSA, EGG sinyallerinin sınıflandırılmasında (Reddy ve ark. 1992; Subasi ve Erçelebi, 2005; Aklan ve ark., 2005), hipertansiyon parametre tahmininde (Türe ve ark., 2005), Ateroskleruz tahmininde (Çolak ve ark. 2005), PET taramasında (Kippenhan ve ark 1992) ve kanser araştırmalarında (Köküer, 2005; Bourdes ve ark., 2007; Navdeep ve ark., 2008) yoğun bir şekilde kullanılmaktadır.

Değişkenler arasındaki ilişkiler doğrusal olmadığında bu tür problemleri modellemek de, çözmek de zordur. Çözüm için bazı varsayımlar yapmak gerekir. Bu da modellenen sistem ile gerçek sistem arasında farklılık olmasına sebep olur. Oysa yapay sinir ağları doğrusal olmayan ilişkileri içinde geleneksel yöntemlerden daha iyi ve gerçekçi çözümler üretir. Yapay sinir ağları klasik istatistiksel yaklaşımlara göre daha karmaşık problemlerin modellenmesinde ve çözülmesinde kullanılabilecek matematiksel modellerdir. Veriler arasındaki doğrusal olmayan ilişkileri başarılı bir şekilde modelleyebilmektedir. Ayrıca YSA, klasik istatistiksel modellere göre herhangi bir varsayım gerektirmezler.

*Bu çalışma 18. İstatistik Araştırma Sempozyumunda sözlü bildiri olarak sunulmuştur

Bu çalışmada, veri kümesine sırasıyla lojistik regresyon ve yapay sinir ağları uygulanarak, bu yöntemlerin doğru sınıflandırma oranları tahmin edilmesi amaçlanmıştır. Bununla birlikte YSA ve lojistik regresyona ait teorik bilgilerde incelenmiştir.

Materyal ve Yöntem

Bu çalışmada kullanılan veri kümesi, 2005-2006 öğretim yılı için Yüzüncü Yıl Üniversitesi Eğitim fakültesi Beden Eğitimi ve Spor Öğretmenliği Bölümü için açılan özel yetenek sınavına katılan 467 erkek adaydan oluşturulmuştur. Adayların sınavı kazanıp kazanmaması bağımlı değişken olarak, öğrencilere ait ÖSS puanları ve Mekik sayıları, OÖBP(Orta Öğretim Başarı Puanı) değerleri ise bağımsız değişken olarak modele alınmıştır. Analiz sürecinde LR için Minitap YSA için R istatistiksel analiz programları kullanılmıştır.

Lojistik Regresyon

Lojistik regresyon (LR), cevap değişkenin ikili (binary) olarak gözlendiği durumlarda ikili bağımsız değişken ile bağımlı değişkenler arasındaki neden-sonuç ilişkisini belirlemede kullanılan bir yöntemdir. Bağımsız değişkenlere göre bağımlı değişkenin beklenen değerlerinin olasılık olarak elde edildiği sınıflama ve atama işlemi yapmaya yardımcı olan bir regresyon yöntemidir ((Molenbergs ve Goetghber, 1997; Palmgren ve Ripatti, 2001, Özdamar,2005).

LR birikimli olasılık fonksiyonu,

$$P_i = F(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}) = F(Z) \qquad (1)$$

biçiminde yazılabilir. Eşitlik 1'de verilen Z fonksiyonu,

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \quad (2)$$

olarak yazılabilir. Logit fonksiyon,

$$P(Y) = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-(Z)}} \qquad (3)$$

biçiminde yazılabilir. Eşitlik 2'de $\beta_0, \beta_1, \beta_2 \cdots, \beta_p$ regresyon katsayıları, $X=(X_1, X_2, ...X_p)$ açıklayıcı değişkenleri belirtir (Halekoh,2004). Böylece bilinmeyen regresyon katsayılar,

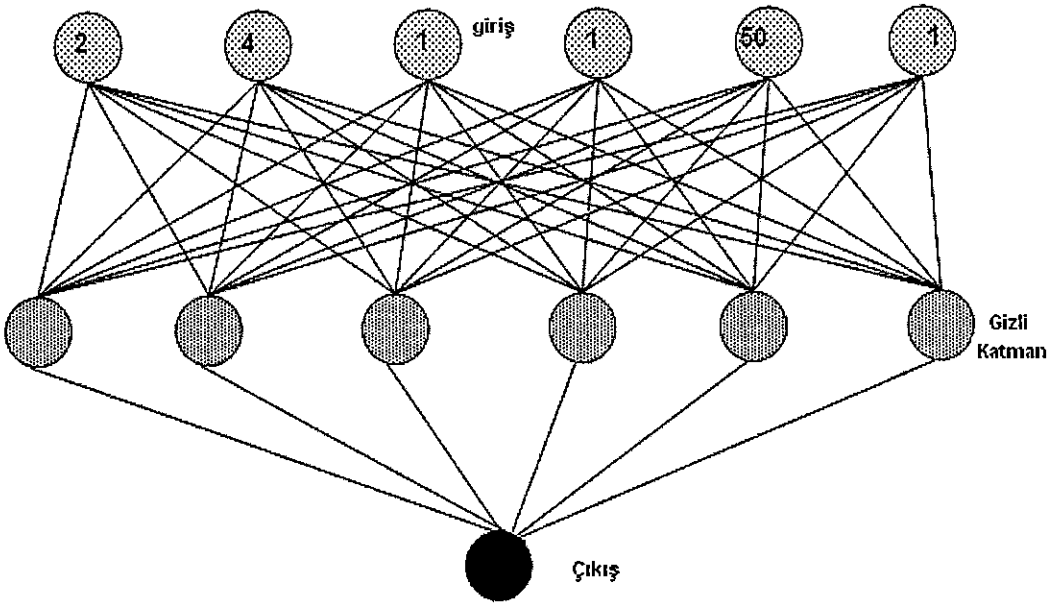
$$\ln\left(\frac{P(y)}{1-P(y)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \qquad (4)$$

$$\frac{P(y)}{1-P(y)} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+....+\beta_p x_p}$$

biçiminde hesaplanabilir.

Yapay Sinir Ağı

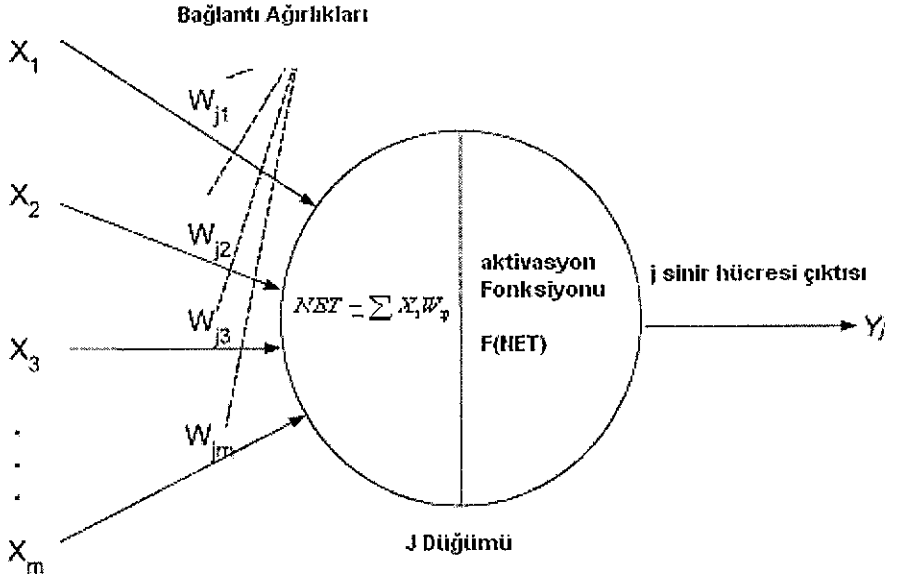
YSA'nın oluşturulması biyolojik sinir sistemi hakkındaki bulgulara dayanmaktadır. Bir yapay sinir ağın yapısında birbirleriyle bağlantılı sinirler yer almaktadır.



Şekil 1: Yapay Sinir Ağı.

Bir yapay sinir ağında birbirleri ile bağlantılı sinirlerin yer aldığı girdi katmanı, ara katman (gizli katman) ve çıktı katmanı bulunur. Girdi katmanı dış dünyadan verileri alır. Çıktı katmanı ise verileri kullanıcıya sunar. Girdi ve çıktı katmanları arasında kalan gizli katmanı ise verileri işleyen katmandır. Gizli katmanda bulunacak sinir hücrelerin sayısı oldukça önemlidir. İşlem hacmi ve ağın büyüklüğü açısından önemlidir.

Bu üç katmanın her birinde bulunan sinir hücreleri ve bunları birbirine bağlayan ağırlıklar Şekil 1'de gösterilmiştir. Şekildeki yuvarlaklar sinir hücrelerini, hücreleri birbirine bağlayan çizgiler ise ağırlıkları göstermektedir. Bir yapay sinir ağındaki en önemli unsurlardan biri de sinir hücrelerinin birbirlerine veri aktarmalarını sağlayan bağlantılardır. Herhangi bir hücreden diğer bir hücreye bilgi ileten bir bağlantı aynı zamanda bir ağırlık değerine sahiptir.



Şekil 2 :Yapay Sinir Hücresi.

Şekil 2'de, $X=\{X_1,X_2,X_3....X_m\}$ girdi değişkenleridir. Bir YSA hücresine dış dünyandan işlenmek üzere gelen bilgilerdir. Girdiler dış dünyadan geldiği gibi sinir hücrelerinden de gelebilirler. $W=\{W_0,W_1,W_2....W_m\}$ ise ağırlıklardır, bir sinir hücresine gelen bilginin önemini ve hücre üzerindeki etkisini gösterir. Ağırlıkların büyük veya küçük olması özellikle tahminleme (predict) açısından önemli. Bu nedenle mümkün olduğu kadar ağırlıklar sıfır etrafında tutularak, önemsiz olanlar ayıklanır ve sonuçta kalan ağırlıklarla tahmin, genelleştirme yapılır. Bu isleme prunnig denir ve kalan ağırlıklar ayni zamanda efektif parametre sayısını verir. Ağırlıkların negatif veya pozitif olması etkinin yönünü belirtir. Ağırlıkların öğrenme süresince değerleri değişebilir. NET, toplama fonksiyonudur, bir hücreye gelen net girdiyi hesaplar. Farklı toplama fonksiyonları bulunmaktadır. Yaygın kullanılan fonksiyon ağırlıklar toplamıdır. Bu fonksiyonda girdilerle ağırlıklar çarpılarak toplanır.
F(.) aktivasyon fonksiyonudur. Bu fonksiyona gelen NET girdiyi işleyerek çıktıyı üreten fonksiyondur. Farklı aktivasyon fonksiyonları bulunmaktadır. Bir YSA'daki hücrelerin tümü aynı veya farklı aktivasyon fonksiyonuna sahip olabilir. Özellikle çok katmanlı YSA modelleri, hesaplamaların daha kolay yapılması açısından kullanılacak aktivasyon fonksiyonun türevi alınabilir türden olmasını istemektedir. Hangi aktivasyon fonksiyonun kullanılacağına kullanıcının denemeleri sonucunda karar verilir.

Yapay sinir ağlarında veriler rasgele olarak eğitim, geçerlilik ve test seti olmak üzere üç bölüme ayrılmaktadır. . Eğitim seti bağımlı değişkenler ile bağımsız değişkenler arasındaki ilişkiyi ortaya koymaya yararken, geçerlilik set ağırlıkların düzeltilmesini sağlar. Bu nedenle NET bir değer ürettiği zaman o değer ile geçerlilik setindeki değer karşılaştırılır ondan sonra bu bilgi geri gönderilir. (Bu nedenle back propogation denilir). Bu bilgiye göre ağırlıklar yeniden güncelleştirilir. Başlangıçta geçerlilik hatası küçülme eğilimi gösterir. Bu hata eğitimin belirli bir aşamasında artış gösterir. Özellikle overfitting olduğu zaman artış olur. Test veri seti ise NET'in hic bir zaman görmediği veri setidir. Bu set genelleştirme için kullanılmaktadır. Veri seti 467 erkek adaydan oluşmakta. 235 rasgele aday eğitim

seti, 116 aday geçerlilik seti, 116 aday ise test seti olarak kullanılmıştır.

Bu çalışmada çok katmanlı ileri beslemeli yapay sinir ağı kullanılmıştır. Yapılan denemeler sonucunda çok katmanlı percetron (Multilayer Perceptron) MLP(3:25:1) şeklinde bir yapay sinir ağı ile veri kümesi sınıflandırılmıştır. Kullanılan yapay sinir ağı 3 katmandan oluşmaktadır. Giriş katmanında 3 giriş değişkeni için 3 sinir hücresi vardır. Gizli katman hiperbolik tanjant sigmoid aktivasyon fonksiyonlu 25 sinir hücresinden oluşmaktadır. Çıktı katmanı ikili (binary) çıktılı sigmoid aktivasyon fonksiyonuna sahip tek sinir hücresinden oluşmaktadır. Yapay sinir ağı çıktısı ile gerçek çıktı değerleri arasındaki farklar(hata) ölçülerek ağın ağırlıkları değiştirilir. Ağ yapısının performansını ölçmek için mutlak hata ortalaması (MHO) veya hata kareler ortalamasına (HKO) kullanılmaktadır. En küçük HKO veya MKO değeri uygun YSA'yı gösterir.

Bulgular

Lojistik Regresyon Analizi Sonuçları: Lojistik regresyon için elde edilen sınıflandırma sonuçları Çizelge 1'de verilmiştir.

Çizelge 1. Lojistik regresyon için Sınıflandırma sonuçları

	Gerçek Durum	LR (Doğru)	LR (Yanlış)	%
Başarılı	79	77	2	0,974
Başarısız	388	379	9	0,976
	467	456	11	0,976

Çizelge 1'de görüldüğü gibi, lojistik regresyon sınavı kazanan adayların %97,4'nü ve sınavı kazanmayan adayların %97,6'nı doğru sınıflanmıştır. LR, toplam 11 aday için yanlış sınıflandırma yapmıştır.

LR analizi sonucunda elde edilen parametre tahminleri, güven aralıkları ve odds ratio değerleri Çizelge 2'de verilmiştir.

Çizelge 2: Lojistik regresyon analizi bağımsız değişkenlere ilişkin parametre tahminleri

Bağımsız Değişkenler	P	OR (Odds Ratio)	OR'nin %95 Güven Aralığı Alt Sınır	Üst Sınır
ÖSS	0,406	0,99	0,95	1,02
MEKİK	0,000	1,70	1,46	1,97
ORTALAMA	0,760	1,01	0,93	1,10
INTERCEPT	0,000			

LR analizi sonuçlarına göre adayın ÖSS puanı arttığında sınavı kazanma oranının 0,99 kat, mekik sayısı sınavı kazanma oranının 1,7 kat, AÖBP'nın artması durumunda ise sınavı kazanma oranının 1,01 kat artığı saptanmıştır. Dolayısıyla sınavı kazanmada mekik değişkeninin diğer değişkenlere göre daha etkin olduğu saptanmıştır.

Yapay Sinir Ağları Analiz Sonuçları: Yapılan denemeler sonucunda verileri sınıflandırmak için MLP(3:25:1) şeklinde bir YSA uygun bulunmuştur. Bu modelin seçiminde her denemede YSA'ın yaptığı doğru sınıflandırma oranlarına bakılarak karar verilmiştir.Giriş katmanında 3, gizli katmanda 25 ve çıkış katmanında 1 sinir hücresi kullanılmıştır. Giriş değişkenlerine karşılık elde edilen 0.5'den büyük çıkış değeri "başarılı", küçük çıkış değerleri ise "başarısız" grubuna dahil edilmiştir.

Bağımsız değişkenler için YSA kullanılarak elde edilen önemlilik değerleri Çizelge 3'de verilmiştir.

Çizelge 3: YSA için Girdi değişkenlerinin önemlilik değerleri

Mekik sayısı	ÖSS	OÖBP
4,794390	1,100246	0,973477

Çizelge 2'de, sınavı kazanmada mekik sayısı değişkenin ÖSS ve OÖBP değişkenlerinden daha önemli olduğu saptanmıştır. Yapay sinir ağları kullanılarak elde edilen sınıflandırma sonuçları Çizelge 4'de verilmiştir.

Çizelge 4: Yapay sinir ağı için sınıflandırma sonuçları.

	BAŞARISIZ	BAŞARILI	TOPLAM
DOĞRU	386 (%99,48)	75 (%94,9)	461
YANLIŞ	2	4	6
TOPLAM	388	79	%98,7

Çizelge 4'de, MLP(3:25:1) modelindeki YSA'ı, sınavı kazanan adayların %94.9'nu ve sınavı kazanmayan adayların %99,48'ni doğru sınıflandırmıştır.

Yapay sinir ağları kullanılarak elde edilen Eğitim Performansı, geçerlilik Performansı ve test Performansı oranları Çizelge 5'de verilmiştir.

Çizelge 5. YSA performans değerleri.

Eğitim Performansı	%99,5745
Geçerlilik Performansı	%98,2759
Test Performansı	%97,4138

Yapay sinir ağı eğitim seti verilerinin %99'unu, geçerlilik seti verilerinin %98'ini ve test seti verilerinin %97'sini doğru sınıflandırdığı saptanmıştır.

Tartışma ve Sonuç

Bu çalışmada, veri kümesine sırasıyla lojistik regresyon ve yapay sinir ağları uygulanarak, bu yöntemlerin veri kümesini doğru sınıflandırma oranları karşılaştırılmıştır. Elde edilen bulgulara göre doğru sınıflandırma oranları bakımından, yapay sinir ağı veri kümesini %98,7, lojistik regresyon ise %97,6 olarak doğru sınıflandırdığı görülmüştür. Bununla birlikte, YSA başarılı olan adayları sınıflamada, lojistik regresyona göre daha etkili olduğu saptanmıştır. Zaten yapay sinir ağları kullanılarak elde edilen Eğitim Performansı, geçerlilik Performansı ve test Performansı oranlarının çok yüksek çıkması elde edilen sonuçlar ile paralellik göstermektedir.

Hatayı minimize etmek amacıyla geriye yayılma algoritması (Back Probagation) kullanılmıştır. Standart geriye yayılım olarak adlandırılan bu eğitim metodu hata kareler toplamının geriye yayılım yöntemiyle küçültülmesi fikrine dayanır ve genelleştirilmiş delta kuralını kullanır. Modelde öğrenme değeri 0.01 olarak seçilmiş ve ağın eğitimi için 500 iterasyon ile "Başarı" değişkeni ile diğer değişkenler arasında bir genelleme yaklaşımı gerçekleştirilmiştir.

Lojistik regresyonda hatalı tahmin edilen gözlem sayısı 11 iken, yapay sinir ağı modelinde 6 olarak elde edilmiştir. Sonuç olarak, yapay sinir ağının lojistik regresyona göre, veri kümesini sınıflandırmada daha iyi tahmin gücüne sahip olduğu saptanmıştır.

Kaynaklar

Aitkin, M., Titterington,D. M., 2000. *Statistics and Neural Network*. The Statistican, 49, 627-628.

Aklan, A.ve ark., (2005). *Automatic seizure detection in EGG using logistic regression and artificial neural network*. Journal of Neuroscience Methods, 148, 167- 176.

Andrew, R. B., 1994. *A Review from Statistical Perpective*. Statistical Science, 9, 33-35

Bing, C., Titterington, D.M., (1994). *A Review from Statistical Perpective*. Statistical Science, 9, 49-54

Bourdes, V., S.ve ark., 2007. *Breast Cancer Predictions by Neural Networks Analysis: A Comparison with Logistic Regression*. Processing of the 29th Annual International Conference of the IEEE EMBS, Lyon, France, August 23-26.

Claudia, P., Foster, P., 2003. *Tree view. Logistic Regression: A Learning- Curve Analysis*. Journal of Machine Learning Researc. 4, 211-255.

Çolak, C. ve ark., 2005. *Ateroskleroz'un tahmini için bir sinir ağı*. Ankara Üniversitesi Tıp Fakültesi Mecmuası. 58, 159-162

Elmas, Ç., 2003. *Yapay Sinir Ağları*. Seçkin Kitapevi, İstanbul.

Gene Hwang, J. T., Adam Ding, A. 1997. *Prediction Intervals for Artificial Neural Networks*. Journal of American Statistical Association, 92, 748-757.

Halekoh, U., (2004). *Logistic Regression*. http://genetics.agrsci.dk/biometry/courses/statmaster/course/module05/module.pdf

Kröse, B., 1996. *An Introduction to Neural Networks*. www.getpedia.com

Köküer, M., 2005. *A Comparison of Multi-Layer Neural Network and Logistic Regression in Hereditary Non-Polyposis Colorectal Cancer Risk Assessment*. Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4.

Lai, T. L., Shing, S. P., 2001. *Stochastic Neural Networks with Applications to Nonlinear Time series*. Journal of American Statistical Association, 96, 968-981.

Navdeep, T. ve ark., 2008. *Prediction technigue survival in peritoneal dialysis patients: comparing artificial neural Networks and logistic regression*. Nephrol Dial Transplant 1-10.

Özdamar K., 2005. *Paket Programlarla İstatiksel Veri Analizi –I*. Kaan Kitapevi, Eskişehir. 196.

Poli, I., Jones, R. D., 1994. *A Neural Model for Prediction*. Journal of American Statistical Association, 89, 117-121.

Rocha, M., 2007. *Evolution of neural network for classification and regression*. Neurocomputing, 70, 2809- 2816.

Subasi, A., Erçelebi, E., 2005. *Classificiation of EGG signals using neural network and logistic regression*. Computer Medhods and Programs in Biomedicine. 78, 87-99.

Türe, M. Ve ark., 2005. *Hipertansiyon tahmini için çoklu tahmin modellerinin karşılaştırılması*. Anadolu Kardiyoloji Dergisi, 5, 24-28

Warren, S., 1994. *Neural Networks and Statistical Models*. SAS.