WEB SAYFALARINA İLİŞKİN YAPAY SİNİR AĞLARI İLE SINIFLANDIRMA YÖNTEMİ

Doç.Dr Erhan Akyazı

Marmara Üniversitesi

Bilişim Bölümü

eakyazi@marmara.edu.tr

Şafak Kayıkçı

Marmara Üniversitesi

Bilişim Bölümü

safak@safakkayikci.com

ÖZET

Bu bildiride web sayfalarının kategorilere ayrılmasında, kelime içeriklerine göre yapay sinir ağları ile öğrenen bir sistem üzerinde durulmuştur. İnternet üzerinde belli kategorilerdeki site adresleri, Açık Dizin Projesi üzerinden alınmıştır. Bu siteler içerindeki kelimelerin kullanılma sayılarını ölçen bir program geliştirilmiş, frekans sayılarına ve kategorilere göre tablolar oluşturulmuştur. Bu tablolar daha sonra Matlab Yapay sinir ağları modülünde kullanılarak web sayfalarını sınıflandırmayı amaçlayan bir araç geliştirilmiştir.

ABSTRACT

In this paper, a learning atrificial neural network system used for classification of web pages into categories related to content mentioned. Internet addresses of sites in specified categories are supplied from Open Directory Project. A program which calculates the frequency of words in content of this pages used and tables are generated related to frequencies and categories. These tables are used in Matlab Neural Network toolbox in order to develop a tool that classify web pages.

Anahtar Kelimeler : Yapay Zeka, Yapay Sinir Ağları, Sınıflandırma, Web İçerik

1.GİRİŞ

Günümüzde internet aracılığıyla erişilebilen bilgi sistemlerinin sayısı hızla artmaktadır. Bu sistemler üzerindeki bilgi kaynakları da giderek çeşitlenmekte ve daha fazla yer kaplamaktadır. Bu çeşitlilik ve çokluk içerisinde, insan beyni her gün binlerce enformasyon ile karşı karşıya kalmaktadır.Aşırı bilgi yüküyle baş etmenin en iyi yolu sınıflandırma yapmaktır.

Sınıflandırma, nesneleri ya da insanları, belirli bir takım ortak niteliklerini temel alarak gruplara ya da sınıflara ayırma sürecidir[1]. Sınıflandırma, aşırı bilgi yükünden kurtulmanın bir yoludur [2].

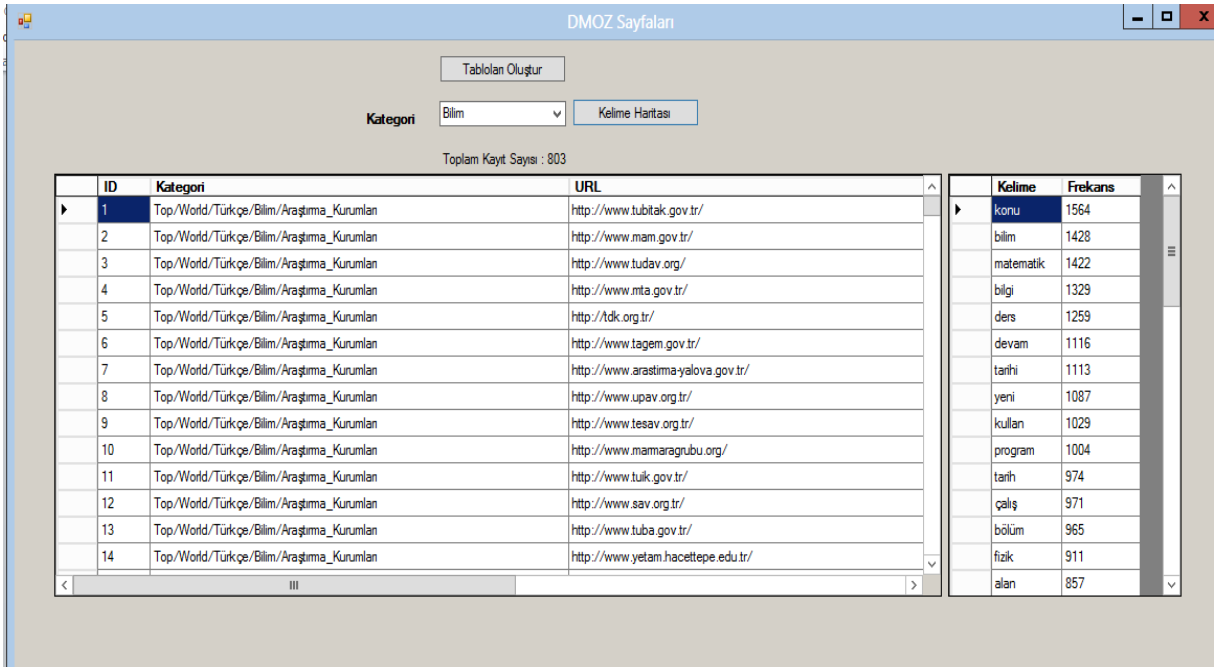
Sınıflandırma problemi, yeni karşılaşılan bir nesnenin, özelliklerinin incelenmesi ve önceden tanımlanmış sınıflardan birine atanmasından oluşur. Bir sınıflandırıcı, sınıf etiketleri bilinen kayıtlardan oluşan bir veritabanı verildiğinde, sonradan gelen kayıtları sınıflandırmak için kullanılabilecek kısa ve anlamlı açıklamalar türetir [3]. Belirli bir amaç doğrultusunda sınıflandırılacak olan nesneler, genellikle bir veritabanındaki kayıtlar ile gösterilir ve sınıflandırma, her bir kaydın sınıf kodu alanının doldurularak güncellenmesinden ibarettir. Sınıflandırma görevi, iyi tanımlanmış sınıf tanımları ve sınıf etiketi önceden bilinen bir eğitim kümesi ile tanımlanır. Görev, sınıflandırılmamış verileri sınıflandırmak üzere uygulanabilen bir çeşit model oluşturmaktır[4].

2.AÇIK DİZİN PROJESİ

Açık Dizin Projesi Web'deki en büyük ve en kapsamlı, insanlar tarafından düzenlenen dizindir. Dünyanın her tarafından katılımda bulunan geniş bir gönüllü editörler topluluğu tarafından inşa edilmiştir ve varlığı onlar tarafından sürdürülmektedir.[5]

Birçok arama motoru ODP'deki siteleri direkt olarak izler ve özgür bir lisansla yayınlanan ODP verilerini kullanılır. Oldukça güvenilir bir alt yapısı vardır. Yasa dışı, problemli ve kalitesiz sitelerin listelenmesine izin verilmez. Projenin sahibi Netscape'tir. İsminde Mozilla geçmesine rağmen, projenin Mozilla Vakfı ve Mozilla projeleriyle bir ilişkisi yoktur.

Bu projede kullanılmak üzere beş kategori seçilmiştir: Bilim, Kültür ve Sanat, Spor, Ekonomi ve İş Dünyası, Sağlık.



ID	Kategori	URL
1	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.tubitak.gov.tr/
2	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.mam.gov.tr/
3	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.tudav.org/
4	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.mta.gov.tr/
5	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://tdk.org.tr/
6	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.tagem.gov.tr/
7	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.arastirma-yalova.gov.tr/
8	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.upav.org.tr/
9	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.tesav.org.tr/
10	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.marmaragrubu.org/
11	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.tuik.gov.tr/
12	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.sav.org.tr/
13	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.tuba.gov.tr/
14	Top/World/Türkçe/Bilim/Araştırma_Kurumları	http://www.yetam.hacettepe.edu.tr/

Kelime	Frekans
konu	1564
bilim	1428
matematik	1422
bilgi	1329
ders	1259
devam	1116
tarihi	1113
yeni	1087
kullan	1029
program	1004
tarih	974
çalış	971
bölüm	965
fizik	911
alan	857

Şekil 1 Açık Kaynak Projesinde Site Adresleri

İlk aşamada, bu kategoriler içerisindeki siteler içerikleri incelenmiş ve her kategoride en fazla geçen elli kelime saptanmıştır. Toplamda beş kategoriye ait ikiyüzelli kelimelik bir kelime haritası oluşturulmuştur.

3) YAPAY SİNİR AĞLARI

Genel anlamda ypay sinir ağları (YSA), beynin bir işlevi yerine getirme yöntemini modellemek için tasarlanan bir sistem olarak tanımlanabilir. YSA, yapay sinir hücrelerinin birbirleri ile çeşitli sekilerde bağlanmasından oluşur ve genellikle katmanlar seklinde düzenlenir. Donanım olarak elektronik devrelerle ya da bilgisayarlarda yazılım olarak gerçeklenebilir. Beynin bilgi isleme yöntemine uygun olarak YSA, bir öğrenme sürecinden sonra bilgiyi toplama, hücreler arasındaki bağlantı ağırlıkları ile bu bilgiyi saklama ve genelleme yeteneğine sahip paralel dağılmış bir işlemcidir. Öğrenme süreci, arzu edilen amaca ulaşmak için YSA ağırlıklarının yenilenmesini sağlayan öğrenme algoritmalarını içerir.

Sinir hücreleri bir grup halinde işlev gördüklerinde ağ olarak adlandırılırlar ve böyle bir grupta binlerce nöron bulunur. Yapay nöronların birbirleriyle bağlantılar aracılığıyla bir araya gelmeleri yapay sinir ağını oluşturmaktadır. Yapay sinir ağıyla aslında biyolojik sinir ağının bir modeli oluşturulmak istenmektedir. Nöronların aynı doğrultu üzerinde bir araya gelmeleriyle katmanlar oluşmaktadır [6].

Katmanların değişik şekilde bir birleriyle bağlanmaları değişik ağ mimarilerini doğurur. YSA'lar üç katmadan oluşur. Bu katmanlar sırasıyla;

a) Girdi Katmanı: Bu katmandaki proses elemanları dış dünyadan bilgileri alarak ara katmanlara transfer ederler. Bazı ağlarda girdi katmanında herhangi bir bilgi isleme olmaz.

b) Ara Katman (Gizli Katman) : Girdi katmanından gelen bilgiler islenerek çıktı katmanına gönderilirler. Bu bilgilerin islenmesi ara katmanlarda gerçekleştirilir. Bir ağ içinde birden fazla ara katman olabilir.

c) Çıktı Katmanı: Bu katmandaki proses elemanları ara katmandan gelen bilgileri isleyerek ağın girdi katmanından sunulan girdi seti için üretmesi gereken çıktıyı üretirler. Üretilen çıktı dış dünyaya gönderilir.

Giriş katındaki nöronlar tampon gibi davranırlar ve giriş sinyalini ara kattaki nöronlara dağıtırlar. Ara kattaki her bir nöronun çıkışı, kendine gelen bütün giriş sinyallerini takip eden bağlantı ağırlıkları ile çarpımlarının toplanması ile elde edilir. Elde edilen bu toplam, çıkışın toplam bir fonksiyonu olarak hesaplanabilir. Buradaki fonksiyon, basit bir eşik fonksiyonu, bir sigmoid veya hiperbolik tanjant fonksiyonu olabilir. Diğer katlardaki nöronların çıkışları da aynı şekilde hesaplanır. Kullanılan eğitme algoritmasına göre, ağın çıkışı ile arzu edilen çıkış arasındaki hata tekrar geriye doğru yayılarak hata minimuma düşünceye kadar YSA'nın ağırlıkları değiştirilir. Yapay sinir ağlarında kullanılan çok sayıda öğrenme algoritması bulunmaktadır. Bu çalışmada Levenberg-Marquardt (LM) öğrenme algoritması kullanılmıştır.

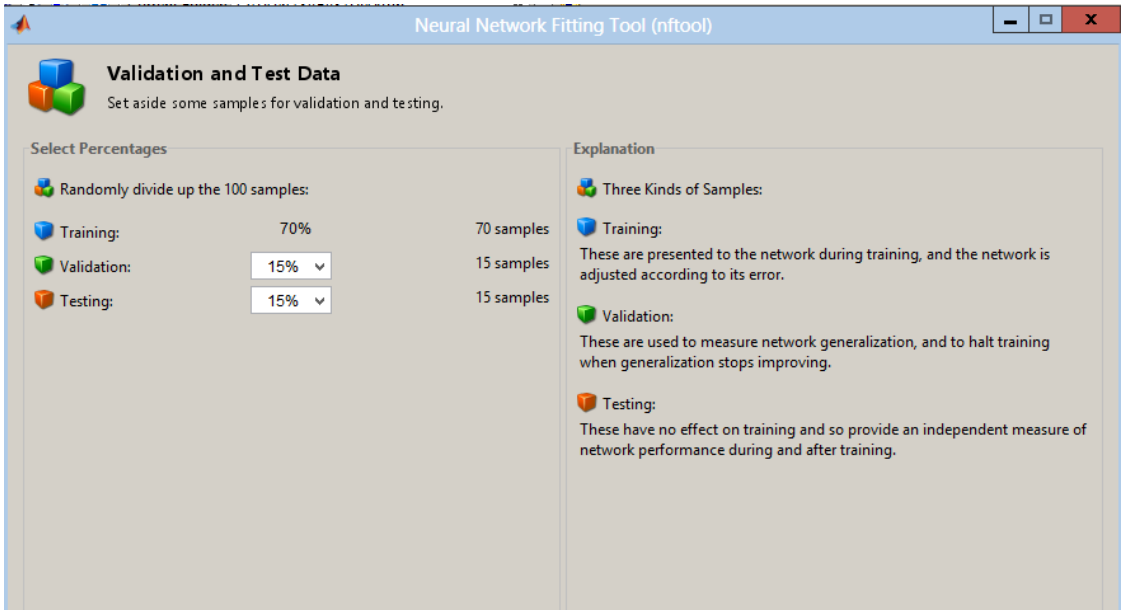
4) SİSTEMİN YAPISI VE UYGULAMASI

Daha önceden belirlenen 5 kategoriden 20'şer adet sayfa için toplamda 100 girişli bir eğitim kümesi oluşturulmuştur. Çıkış değerleri eğitim seti için sayfanın ait olduğu kategori için 1, diğer kategori değerleri için 0 verilmiştir. Yapay sinir ağının eğitiminde kullanılan verilere ilk örnek Şekil 2'de verilmiştir. Toplamda yüz adet örnek bulunmaktadır. Burdaki giriş verilerindeki değerler, kelime frekanslarının toplam kelime sayısına oranından hesaplanmıştır.Oluşturulan bu eğitim kümesi YSA'ya sunulmuş ve Levenberg-Marquardt öğrenme algoritması kullanılarak eğitim yapılmıştır. Eğitim sırasında nöronlar arasındaki ağırlıklara ilk değer olarak [-1,+1] arasında rastgele değerler atanmıştır ve ara katmanda 20 adet nöron kullanılmıştır.

Örnek	Giriş	Çıkış
1	0,002617801 0,039267016 0,002617801 0,007853403 0 0,028795812 0 0,007853403 0 0,060209424 0 0,002617801 0 0 0,010471204 0 0,005235602 0 0,007853403 0,007853403 0 0,04973822 0,031413613 0 0 0,002617801 0,002617801 0,005235602 0,002617801 0,007853403 0,007853403 0 0 0 0,005235602 0,002617801 0,002617801 0 0,002617801 0,002617801 0,002617801 0 0,007853403 0 0 0,010471204 0,005235602 0 0,002617801 0 0 0 0,007853403 0,028795812 0 0,007853403 0 0,002617801 0,002617801 0,002617801 0 0 0 0 0 0,005235602 0,002617801 0 0 0,007853403 0 0 0 0 0,002617801 0 0 0 0 0,002617801 0 0 0 0 0,010471204 0 0,002617801 0 0 0 0 0 0 0,002617801 0 0,007853403 0 0,007853403 0,002617801 0 0 0 0 0,007853403 0 0 0,028795812 0 0 0 0 0 0,007853403 0 0 0 0,015706806 0 0,015706806 0 0 0,002617801 0 0 0 0 0 0,020942408 0 0 0 0 0,002617801 0,002617801 0 0 0,005235602 0 0,002617801 0,010471204 0 0 0 0 0 0 0 0,002617801 0,002617801 0,002617801 0,002617801 0,007853403 0,028795812 0,007853403 0 0 0 0 0,007853403 0,010471204 0 0,005235602 0,031413613 0,007853403 0,005235602 0,002617801 0,002617801 0 0 0 0 0 0 0,002617801 0 0 0,007853403 0,002617801 0,005235602 0,002617801 0,002617801 0 0,023560209 0,007853403 0 0 0,013089005 0 0 0,007853403 0 0 0,013089005 0 0,002617801 0 0,002617801 0,020942408 0,005235602 0 0,028795812 0,007853403 0 0 0 0,002617801 0 0 0 0,002617801 0 0,002617801 0 0 0,007853403 0,007853403 0 0,002617801 0,007853403 0 0 0,002617801 0 0 0,007853403 0 0,002617801 0 0 0 0 0,005235602 0 0 0 0 0,002617801 0 0,010471204 0 0,060209424 0 0 0 0 0 0,020942408 0	1 0 0 0 0

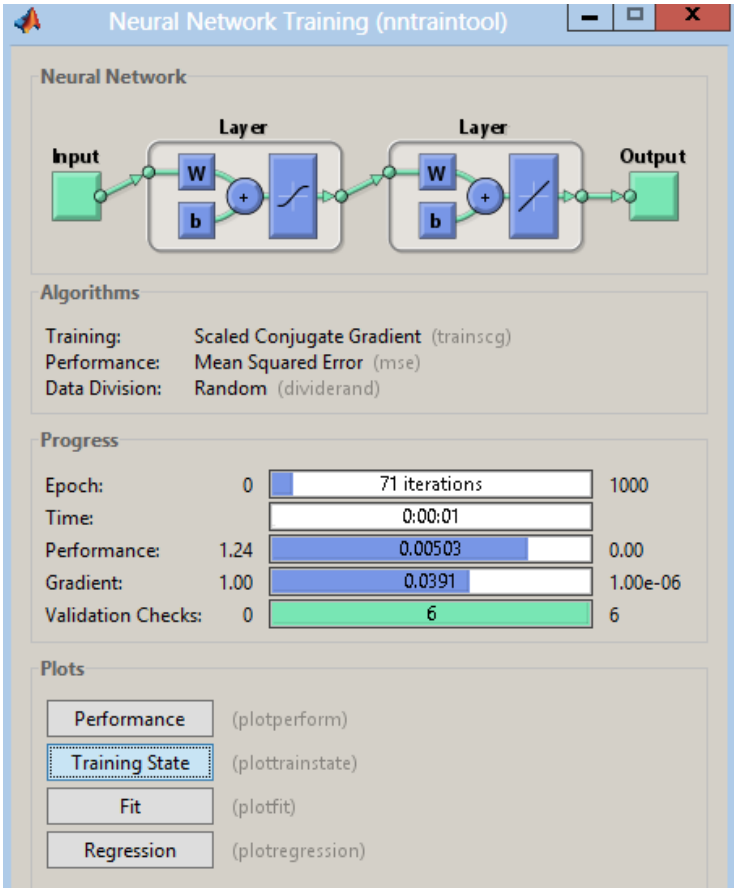
Şekil 2 İlk Örneğe Ait Giriş ve Çıkış Değerleri

Öngörülen 100 adet verinin %70'i eğitim, %15'i doğrulama, %15'i de test amaçlı kullanılmıştır.



Şekil 3 YSA'da Kullanılan Ölçüm Değeleri

Yaklaşık 71 iterasyondan sonra eğitim tamamlanmış, sonrasındaki performans ölçümleri Şekil 4'teki gibi gözlemlenmiştir.



Şekil 3 YSA'da Kullanılan Ölçüm Değeleri

4) DEĞERLENDİRME VE SONUÇ

Levenberg-Marquardt öğrenme algoritması kullanılarak eğitim sağlanmıştır. Daha sonra, test eğitim seti sisteme girilmiştir. Test sonuçlarında, yüksek oranda beklenen değerlere göre yüksek oranda başarı görülmüştür. Başarı oranlarının, kelime haritası oluşturulurken yapılan tarama miktarına, test sonucu alınan sayfadaki metnin semantik olarak aslında içerik olarak farklı kategoride olmasına, yapay sinir ağı oluşturulurken kelimeler üzerinde verilen ilk ağırlık değerlerine göre değişebileceği saptanmıştır**.**

Sürekli değişen dış çevreye uyum sağlamak zorunda olan bireyin, değişik kaynaklardan kendisine gelen tüm bilgiler dikkat edip, onları algılamak, çözümlemek, saklamak ve gerektiğinde kullanmak için yeterli kapasitesi yoktur. Birey yoğun, karmaşık ve hızlı bilgi akışı ile "sınıflandırma" yardımıyla baş etmektedir. Geliştirilen örnek uygulama ile yapay zeka tekniklerinin internet üzerinde zeki yazılımlar oluşturmak için kullanılabilir olduğu görülmüştür. Sayfaların yapay sinir ağları ile sınıflandırılması, doğru bilgiye hızlı erişimin sağlanabilmesini olanaklı kılacaktır.

KAYNAKLAR

[1] Tajfel, H. ve Forgas, J. P. (1981). "Social categorization: Cognitions, values and groups". J. P. Forgas (der.), Social cognition: Perspectives on everyday understanding. London: Academic Press. 113-41.

[2] Spears, R. ve Haslam, S. A. (1997). "Stereotyping and the burden of cognitive load". R. Spears, P. H. Oakes, N. Ellemers, ve S. A. Haslam (der.), The social psychology of stereotyping and group life, 171-207.

[3] R. Rastogi and K. Shim, PUBLIC: A decision tree classi¯er that integrates," Data mining and Knowledge Discovery, vol. 4, pp. 315{344, 2000.

[4] M. J. Berry and G. Lino®, Data Mining Techniques: For Marketing, Sales, and Customer Support. New York, NY, USA: John Wiley & Sons, Inc., 1997.

[5] DMOZ Web Sitesi http://www.dmoz.org/docs/tr/about.html

[6] Haykin, S. 1994. Neural Networks. Macmillan College Publishing Company, USA, 696P