

Otomatik Doküman Sınıflandırma

Rumeysa YILMAZ, Rifat AŞLIYAN, Korhan GÜNEL

Adnan Menderes Üniversitesi, Fen Edebiyat Fakültesi Matematik Bölümü, Aydın

rumeysa2903@gmail.com, rasliyan@adu.edu.tr, kgunel@adu.edu.tr

Özet: İnternetin hızla gelişmesi elektronik ortamdaki bilgileri ve işlemleri hızlandırmış fakat bu ortamlarda depolanan ve işlenen bilgilerin boyutunun artması aranan bilgiye erişmekte problemler çıkarmıştır. Kullanıcıların istedikleri bilgiye daha doğru ve hızlı bir şekilde ulaşma ihtiyacı doğmuştur. Bu amaçla elektronik ortamdaki dokümanların sınıflandırılmasında yeni yaklaşımlar geliştirilmiştir. Bu çalışmada metin içerikli dokümanların sınıflandırılmasında Yapay Sinir Ağlarından Çok Katmanlı Algılayıcı metodu kullanılarak bir sistem geliştirilmiştir. Çalışmanın gerçekleştirilmesi için her biri 75'er doküman içeren eğitim, otomobil, sağlık, spor ve teknoloji sınıfları ele alınmıştır. Bu dokümanlardan 25'er tanesi sistemin eğitilmesi aşamasında 50'er tanesi ise sistemin test edilmesi aşamasında kullanılmıştır. Çalışmada sisteme verilen dokümanlar öncelikle önışlemeden geçirilmiştir. Önışlemeden geçirilen dokümanların frekansları hesaplanıp normalize edildikten sonra her bir sınıf için öznitelik sözcük ve hece vektör veritabanı oluşturulmuştur. Öznitelik vektör veritabanı oluşturulurken sözcüklerin ve hecelerin dokümanlarda karşılaştırılmasında belli bir eşik değeri kullanılmıştır. Sistemin test edilmesinde, test setindeki dokümanlar sisteme verilmiş ve her bir sınıf için oluşturulan öznitelik vektör veritabanındaki sözcükler ve heceler ile karşılaştırılarak dokümanın hangi sınıfa dahil olduğu belirlenmiştir. Sonuç olarak, bu yaklaşım ile en iyi sınıflandırma başarı oranı, sözcük tabanlı sistemde %87 ve hece tabanlı sistemde ise %93 olarak bulunmuştur.

Anahtar Sözcükler: Doküman Sınıflandırma, Yapay Sinir Ağları, Çok Katmanlı Algılayıcı, Veri Madenciliği.

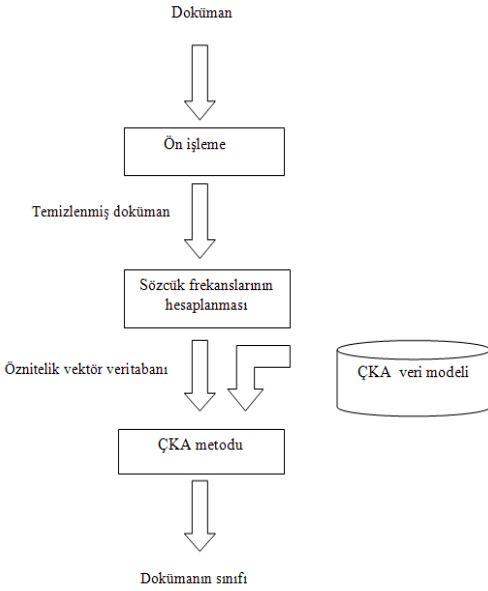
1. Giriş

Doküman sınıflandırma çalışmaları 1960'lı yıllarda başlamıştır. Gelişen teknolojiyle beraber elektronik ortamdaki dokümanların sayısı artmakta ve bunlara ulaşabilmek zorlaşmaktadır. Bu alanda yapılan çalışmalarla gereksiz bilgilerin kullanıcıya ulaşması engellenerek istenilen bilgiye daha hızlı ve daha doğru bir şekilde ulaşılması kolaylaşmıştır. Otomatik doküman sınıflandırma; bilgi alma, bilgi çıkarma, doküman indeksleme, doküman filtreleme, otomatik olarak meta-data elde etme ve web sayfalarını hiyerarşik olarak düzenleme gibi pek çok alanda kullanılır.

Doküman sınıflandırmanın amacı bir dokümanın özelliklerine bakarak önceden belirlenmiş belli sayıdaki kategorilerden hangisine dâhil olacağını belirlemektir. Bunun için çeşitli sınıflandırma yöntemleri geliştirilmiştir. Yaygın olarak kullanılan sınıflandırma yöntemleri; Naive Bayes [1], Karar Ağaçları [2], K-En Yakın Komşu Modeli (KNN) [3], Maksimum Entropi Modelleri [4][5], Bulanık Mantık Teorisi Yaklaşımları [6], Destek Vektör Makineleri [7][8] ve Yapay Sinir Ağlarıdır.

Bu çalışmada, doküman sınıflandırma işlemi yapılırken yapay sinir ağlarından Çok Katmanlı Algılayıcı Ağı kullanılmıştır. İlk

olarak yapay sinir ağıları tanıtilmiş, sistemin yapısı verilmiş, uygulamada takip edilen adımlar sunulmuştur.



Şekil 1.1 Doküman sınıflandırmanın genel yapısı

Eğitim, otomobil, sağlık, spor ve teknoloji kategorilerine ait 75'er doküman, toplamda 3755 doküman ele alınmıştır. Bunlardan 25'er tanesi sistemin eğitilmesi aşamasında 50'şer tanesi de test aşamasında kullanılmıştır.

Otomatik doküman sınıflandırmada sisteme verilen dokümanların sayısının önemli bir rolü vardır. Dokümanların gereğinden fazla olması sistemin öğrenmesini zorlaştırmakta, gereğinden az olması da yapay sinir ağı sonuçlarındaki hata oranını arttırmaktadır.

Metin dokümanları oldukça fazla sözcük içerirler. Bazı sözcükler vardır ki bunların bütün dokümanlardaki frekansı oldukça yüksektir. Bunlara Türkçede çok sık kullanılan; "gibi", "ise", "yani", "veya", "ama", "ne", "neden", "şey", "hiç" sözcükleri örnek verilebilir. Bundan dolayı bu sözcükler ayırt edici özelliğe sahip değildir ve bu

sözcükler dokümanlardan elenir. Eleme işlemi indeksleme işlemi olarak adlandırılır ve bunu takip eden adımlardan oluşur.

Doküman sınıflandırmada Şekil 1.1'de gösterildiği gibi; dokümanlar ilk önce sisteme alınarak ön işlemden geçirilir. Önleme safhasında dokümanlardaki boşluk, rakam ve noktalama işareti gibi herhangi bir anlam ifade etmeyen karakterler elenir, büyük harfler küçük harflere dönüştürülerek temizlenmiş doküman haline getirilir. Dokümanlardaki sözcükler, RASAT heceleme algoritmasıyla [9] hecelere ayrılır. Dokümanlardaki sözcüklerin ve hecelerin frekansları 0 ile 1 arasında normalize edildikten sonra her sınıf için oluşturulmuş olan öznitelik vektör veritabanındaki sözcükler ve hecelerle karşılaştırılarak dokümanın sınıfı belirlenir.

2. Yapay Sinir Ağları

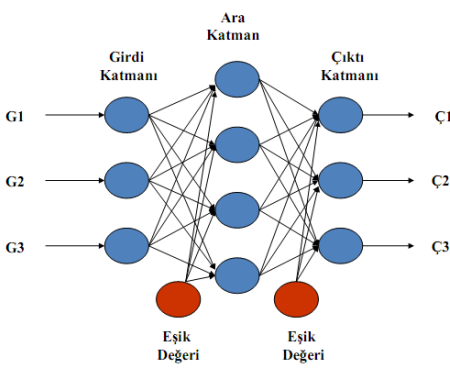
Yapay sinir ağları insan beyninin sinir sistemini model alan ve çalışma prensibine dayanan bir yöntemdir. İnsan beyninin öğrenme yolu ile yeni bilgiler üretebilme, keşfedebilme, mevcut bilgiler ile olaylar hakkında yorum yapabilme, karar verebilme, olaylar arasında ilişki kurabilme gibi özelliklerini yapabilmek için tasarlanmıştır.

Bir yapay sinir ağı belli bir amaç için oluşturulur ve insanlar gibi örnekler sayesinde öğrenir. İnsanlarda öğrenme sinir hücrelerinin arasındaki sinaptik boşluklarda yer alan elektriksel ayarlamalarla oluyorken, Yapay Sinir Ağlarında bu durum tekrarlanan girdiler sayesinde ağı kendi yapısını ve ağırlıklarını değiştirmesi ile olmaktadır. İnsanlardaki sinir hücresinin Yapay Sinir Ağlarındaki karşılığı proses elemanıdır ve Yapay Sinir Ağları birçok proses elemanının birleşmesi ile oluşur. Yapay Sinir Ağları öğretmenli öğrenme, öğretmensiz öğrenme ve destekleyici öğrenme olarak 3 farklı öğrenme tipine sahiptir.

Bu çalışmada öğretmenli öğrenme metotlarından olan Çok Katmanlı Algılayıcı Ağı kullanılmıştır. Çok Katmanlı Algılayıcı Modeli, 1 girdi katmanı, 1 veya daha fazla ara katman ve bir de çıktı katmanından oluşur. Şekil 2.1' de Çok Katmanlı Algılayıcı Modelinin yapısı verilmiştir.

Dış dünyadan alınan bilgiler hiçbir işleme tabi tutulmadan ara katmana iletilir. Dolayısıyla bu katmandaki k tane proses elemanının çıktısı 2.1 denkleminde görüldüğü üzere ζ_K^I olarak belirlenir.

$$\zeta_K^I = G_k \quad (2.1)$$



Şekil 2.1 Çok Katmanlı Algılayıcı Modeli

Ara katmandaki her bir proses elemanının çıktısı girdi katmanından gelen her bir çıktının ağırlıkları ile (A_1, A_2, \dots) çarpımlarının toplanması sonucu elde edilir.

$$NET_j^\alpha = \sum_{k=1}^n A_{kj} \zeta_k^j \quad (2.2)$$

Denklemler 2.2'de A_{kj} k . girdi katmanı elemanını j . ara katman elemanına bağlayan bağlantının ağırlık değerini gösterir. j . ara katman elemanının çıktısı NET girdinin aktivasyon fonksiyonundan geçirilmesi ile hesaplanır. Kullanılan aktivasyon fonksiyonu,

lineer fonksiyon, step fonksiyonu, sinüs fonksiyonu, eşik değer fonksiyonu, hiperbolik tanjant fonksiyonu veya sigmoid fonksiyonu olabilir. Bu çalışmada ağın bütün elemanları için aktivasyon fonksiyonu olarak sigmoid fonksiyon kullanılmıştır.

Sigmoid fonksiyona göre ara katmanın çıktısı denklem 2.3' deki gibidir.

$$\zeta_j^\alpha = \frac{1}{1 + e^{-(NET_j^\alpha + \beta_j^\alpha)}} \quad (2.3)$$

Ele alınan β_j değeri ara katmandaki j . elemana bağlanan eşik değer elemanının ağırlığıdır. Burada ağın çıktısı ile beklenen çıktı arasındaki fark hatayı verir. Bu hata tekrar geriye doğru yayılarak minimuma düşünceye kadar yapay sinir ağının ağırlıkları değiştirilir.

$(\beta_1, \beta_2, \dots)$ ağın beklenen çıktıları, $(\zeta_1, \zeta_2, \dots)$ ağın çıktısı olmak üzere çıktı katmanındaki m . proses elemanında oluşan hata denklem 2.4' de verilmiştir.

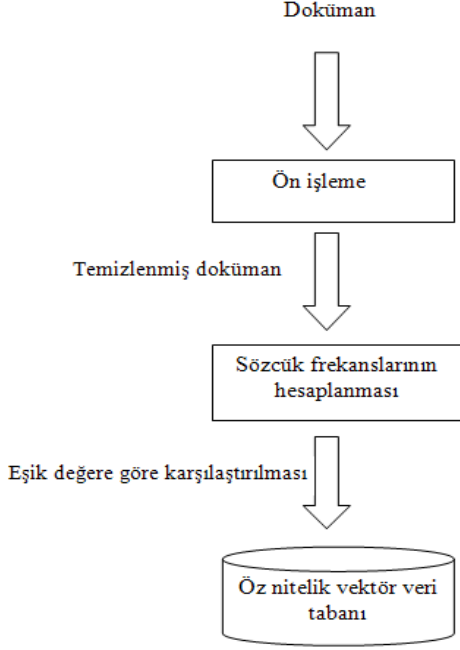
$$E_m = B_m - \zeta_m \quad (2.4)$$

3. Sistemin Yapısı Ve Uygulanması

Bu çalışmada eğitim ve test dokümanlarını toplanması, öznitelik vektörlerinin belirlenmesi, yapay sinir ağının eğitilmesi aşaması ve test aşamalarının izlendiği bir sistem geliştirilmiştir.

İlk olarak sisteme verilen dokümanlar ön işlemden geçirilir. Ön işleme safhasında dokümanlardaki rakam, boşluk, noktalama işareti gibi herhangi bir anlam ifade etmeyen karakterler elenir, bütün sözcükler küçük harflere dönüştürülür ve sadece sözcüklerden

oluşan temizlenmiş doküman elde edilir. Bu aşamadan sonra eğitim, otomobil, sağlık, spor ve teknoloji kategorileri için sözcük ve hece öznitelik vektör uzayının oluşturulması gerekir.



Şekil 3.1 Öznitelik vektör veritabanının oluşturulması

Her bir kategoriye ait test setindeki 25 doküman birleştirilerek tek bir doküman haline getirilir. Böylece 5 kategori için 5 doküman elde edilmiş olur. Bu dokümanlar sisteme verilerek ön işlemden geçirilir ve temizlenmiş doküman haline getirilir. Her sınıf için olasılıkları hesaplanan sözcüklerin ve hecelerin diğer dokümanlardaki olasılıkları belli bir eşik değerinden küçük ise bu sınıfın öznitelik vektör uzayına alınır. Bu çalışmada eşik değeri 0.5 olarak kullanılmış ve her kategori için 5 tane sözcük ve hece öznitelik vektör veritabanı oluşturulmuştur. Öznitelik vektör uzayı bu sınıfları en iyi temsil edecek olan sözcüklerden ve hecelerden oluşur. Şekil 3.1’de öznitelik vektör uzayını oluştururken izlenen adımlar verilmiştir.

Tablo 3.1’de eğitim sınıfı için oluşturulan sözcük öznitelik vektör veritabanları verilmiştir.

Sözcük Sırası	Eğitim Sınıfı Sözcükleri
1	Akademik
2	Bakanlık
3	Ders
4	Dil
5	Dönemi
6	Eğitimine
7	Hafta
8	Katsayı
9	Lisans
.	.
.	.
.	.
88	Yıllık

Tablo 3.1 Eğitim sınıfının sözcük öznitelik vektör veritabanı

3.1 Sistemin Eğitilmesi

Eğitim aşamasında eğitim, otomobil, sağlık, spor ve teknoloji kategorilerine ait 25’er dokümandan oluşan toplamda 125 doküman içeren eğitim kümesi oluşturulur. Bu dokümanlar sisteme verilerek teker teker ön işlemden geçirilirler. Temizlenmiş olan dokümanlardaki sözcüklerin ve hecelerin normalize edilmiş frekansları hesaplanır. Öznitelik vektör veritabanındaki sözcükler ve heceler, dokümanlardaki sözcükler ve hecelerle karşılaştırılarak hangi sınıfa ait olduğu belirlenir.

Bu çalışmada oluşturulan yapay sinir ağı 1 girdi katmanı, 2 ara katman ve bir de çıktı katmanından oluşur. Sistem 5 kategori için eğitildiğinden 5 farklı model oluşturulmuştur. Buna göre girdi katmanında; her sınıf için için öznitelik vektör boyutu kadar proses elemanı bulunmaktadır.

Çıktının belirlenmesinde aktivasyon fonksiyonu olarak sigmoid fonksiyonu kullanılmıştır. Sigmoid fonksiyonu 0 ile 1 arasında değişen bir değer olduğundan sistem çıktıyı doğru sınıflandırma için 1'e yanlış sınıflandırma için 0'a götürecektir. Yapay sinir ağının oluşturulmasında öğrenme kat sayısı olarak 0.3 değeri alınmıştır. Bu çalışmada hata oranı 10^{-3} olarak alındığında 300000 iterasyon sonucunda sistem eğitimi tamamlamıştır.

Böylece, sistemimiz eğitim, otomobil, sağlık, spor ve teknoloji sınıfları için eğitilmiş olur.

3.2 Sistemin Test Edilmesi ve Sonuçlar

Sistemin test edilmesinde 5 kategoriye ait 50 doküman toplamda 250 doküman sisteme verilir. Matlab programında Çok Katmanlı Algılayıcı Metodu kullanılarak sistem test edilmiş ve sonuçlar Tablo 3.2 ve 3.3' de verilmiştir.

Sınıflar	Sınıflandırma Başarı Yüzdeleri
Eğitim	84
Otomobil	93
Sağlık	80
Spor	93
Teknoloji	83
Ortalama	87

Tablo 3.2. ÇKA kullanarak sözcük tabanlı doküman sınıflandırma

Sınıflar	Sınıflandırma Başarı Yüzdeleri
Eğitim	96
Otomobil	94
Sağlık	93
Spor	94
Teknoloji	89
Ortalama	93

Tablo 3.3. ÇKA kullanarak hece tabanlı doküman sınıflandırma

Burada eşik değerini 0.5 olarak kullanılarak sözcük tabanlı sistemde %87 ve hece tabanlı sistemde ise %93 oranında başarı elde edilmiştir. Fakat daha farklı sonuçlar da elde edilebilir. Eğitim aşamasında sisteme verilen dokümanların sayısı, eşik değeri, Çok Katmanlı Algılayıcıdaki ara katmanların sayısı, proses elemanlarının sayısı, kullanılan aktivasyon fonksiyonu, iterasyon sayısı, sınıflandırma metodları gibi faktörler sistemin hata oranına etki eder.

4. Tartışma ve Sonuçlar

Bu çalışmada yapay sinir ağlarından ÇKA metodu kullanılarak bir sistem geliştirilmiştir. Bu sisteme göre dokümanlar 5 farklı kategori altında sınıflandırılmıştır. Her sınıfı temsil eden 5 farklı öznitelik vektör veritabanı oluşturulmuştur. Öznitelik vektör veritabanının oluşturulmasında bir dokümandaki olasılıkları hesaplanan sözcüklerin diğer dokümanlardaki olasılıklarının belli bir eşik değerinden küçük olması göz önünde bulundurulmuştur. Oluşturulan yapay sinir ağı modeline göre sözcük tabanlı ve hece tabanlı sistemler test edilmiş ve bütün sınıflar için sırasıyla %87 ve %93 oranında başarı elde edilmiştir. Görüldüğü üzere, sözcük tabanlı sistemin başarısı, hece tabanlı sistem oluşturulmak suretiyle %6 artırılmıştır.

Daha sonraki çalışmalarımızda daha farklı yöntemler (K-En Yakın Komşu, Destek Vektör Makinesi) kullanarak sistemlerin başarı oranlarını karşılaştıracağız. Aynı zamanda, sözcük ve hece n -gramlarını kullanarak sistemlemler geliştireceğiz.

5. Kaynaklar

[1] Kim, S.B., Rim, H.C., Yook, D., Lim, H.S., "Effective methods for improving Naive Bayes text classifiers", **The 7th Pacific rim international conference on artificial intelligence**, 414–423 (2002).

[2] Wu, M.C., Lin, S.Y., Lin, C.H., "An effective application of decision tree to stock trading", **Expert Syst Appl**, 31(2):270–274 (2006).

[3] Soucy, P., Mineau, G.W., "A simple K-NN algorithm for text categorization", **Proceeding of the first IEEE international conference on data mining (ICDM_01)**, 647–648 (2001).

[4] Li, R., Wang, J., Chen, X., Tao, X., Hu, Y., "Using maximum entropy model for Chinese text categorization", **J Comput Res Dev**, 42(1):94–101 (2005).

[5] Kazama, J., Tsujii, J., "Maximum entropy models with inequality constraints: A case study on text categorization", **Mach. Learn.**, 60(1–3):159–194 (2005).

[6] Liu, W.Y., Song, N., "A fuzzy approach to classification of text documents", **J. Comput. Sci. Technol.**, 18(5):640–647 (2003).

[7] Joachims, T., "Text categorization with support vector machines: Learning with many relevant features", **Nedellec C, Rouveirol C (eds) Proceedings of the 10th European**

conference on machine learning (ECML-98), Springer, Chemnitz, 137–142 (1998).

[8] Yang, Y., Liu, X., "A re-examination of text categorization methods", **Proceedings of SIGIR'99**, 42–49 (1999).

[9] Aşlıyan R., Günel K., "A Comparison of Syllabifying Algorithms for Turkish", **Journal of Advanced Research in Computer Science**, 3(1):58-78 (2011).