Makine Öğrenmesi İle Ürün Sınıflandırma İncelemesi

Can Razbonyalı, canrazbonyali@gmail.com, Yrd. Doç. Dr. Aslı Uyar Özkaya, asli.uyar@okan.edu.tr

Fen Bilimleri Enstitüsü, T.C. Okan Üniversitesi

Özet: Günümüzde internetten satış yapan web sayfalarının yaygınlığı, internet kullanımının artması ile doğru oranda artmaktadır. Bununla birlikte internette sunulan ürünlerin çeşitliliği de artmaktadır. Bu ürünlerin sınıflandırılmasının da zaman alacağı kaçınılmazdır. Bu çalışmada, TeksoSA firmasının web sayfasında bulunan akıllı telefon, cep telefonu ve tablet bilgisayarların özellikleri kullanılarak oluşturulan eğitim seti yardımı ile yapılan bir sınıflandırmanın sonuçları sunulmaktadır. Bu çalışmanın sonucunda; elektronik ticaret firmalarına, ürünü az zaman harcayarak etkili şekilde sınıflandırmanın yolları hakkında fikir vermiş olmak amaçlanmaktadır.

Anahtar Sözcükler: Makine öğrenmesi, ürün sınıflandırması, makine öğrenmesi ile ürün sınıflandırması.

A Survey of Product Catagorization With Machine Learning

Abstract: Nowadays online comerce web sites are getting popular day by day. Besides, the products in these web sites are increasing in variety. It is inevitable that catagorization of these products takes time. In this study, smart phones, mobile phones and tablet computers, that are retrieved from TeknoSA company product catalogue on the internet, are created using the properties of a classification is made through the training set is presented. As a conclusion, giving an idea to e – commerce companies that are looking for a way to categorize products effectively by spending less time is aimed.

Keywords: Machine learning, product classification, product classification with machine learning.

1. GİRİŞ

İnternette bulunan bilginin büyüklüğünü kesin şekilde belirtmek mümkün değildir; ancak 2003 yılında yapılan araştırmaya göre, internette bulunan sayısal bilgilerin her yıl %69 oranında arttığı ortaya konmuştur[1]. 2005 yılının Ocak ayında yapılan bir araştırmaya göre ise internet üzerinde, 11,5 milyar adet indekslenebilir web sayfasının olduğu ortaya konmuştur[2]. Yine aynı araştırmada internette mevcut web sayfalarına her gün 11 milyar adet yeni web sayfasının eklendiği de belirtilmiştir[2]. Günümüzde ise bir dakikada; 277.000 tweet atılmakta, Google yaklaşık iki milyon arama yapmakta, YouTube ye yetmiş iki saatlik video yüklenmekte, yüz milyondan fazla elektronik posta gönderilmekte Facebook üç yüz elli GB büyüklüğündeki veriyi işlemekte ve beş yüz yetmiş bir yeni web sayfası yaratılmaktadır[3]. Yapılan bu araştırmalar ışığında internete olan ilginin arttığını ve internetin ne kadar yaygınlaştığını söylemek mümkün olacaktır. İnternetin bu denli yaygınlaşması ve popüler olması, ürün satan firmaların ayrıca internet üzerinden de satış yapılabilecek sanal vitrinlere yoğunlaştırmaktadır. Durum böyle olunca firmalar, sabit mağaza ile ulaşabilecekleri müşteri sayısından daha fazla sayıdaki müşteri sayısına internet üzerinden, web sayfaları sayesinde ulaşabilmekte, satış yapabilmektedirler. Daha çeşitli ürün daha fazla müşteri ve buna oranla daha fazla gelir anlamına gelmektedir. Dolayısı ile internet üzerinden ne kadar çeşitli ürünü sanal vitrinlerine koyarlarsa o kadar çok ilgi çekip, acımasız pazar rekabetinde o kadar sağlam ayakta kalabileceklerdir. Yaşam şartlarının zorluğu da göz önünde tutulursa, firmaların ürün çeşitliliği ve bu ürünlerin internet vasıtası ile de daha büyük kitlelere ulaştırma isteklerin hırsla artmakta olduğunu tahmin etmek kolay olacaktır.

Yeni piyasaya çıkan bir ürünü sınıflandırıp müşteriye ulaştırma aşamasında harcanan her fazla zaman, firmaya olumsuz etki olarak geri dönmektedir. Bu yüzden, piyasaya sürülen yeni ürünlerin ait oldukları ve birçok farklı kategoride bulunan ürünler ile ortak özelliklere sahip olmaları açısından tam olarak kategorize edilmeleri, ürün sınıflaması açısından firmaların karşı karşıya kaldıkları en büyük zorluktur. Ürün sınıflaması, tüketicinin ürün hakkında değerlendirmesini ve beklentilerini etkilemeye başladığından beri büyük önem kazanmıştır[4]. Belirsiz ürünlerin yaygınlığı ve doğru sınıflandırmanın kritikliği göz önünde bulundurularak tüketicilerin bu ürünleri nasıl sınıflandırdıkları hakkında çalışmalar yapılmıştır[4]. Ayrıca firmalar bu konuda, çok yoğun şekilde çalışıp, zamanlarının büyük bölümünü harcamaktadırlar[5]. Ürün sınıflamasında bir de ürün önerisi, fiyat kıyaslaması gibi ek hizmetler işin içerisine girdiğinde, firmaların bu destekleri sağlayabilme yükü kat kat artmaktadır. Bu noktada, etkin ürün sınıflandırma algoritmaları firmaların imdatlarına yetişmekte olup metin madenciliği ve makine öğrenmesi yöntemlerini içeren yarı otomatik interaktif ürün sınıflandırma yaklaşımları yapılan çalışmalar ile ortaya konulmuştur[6].

Bu çalışmada TeknoSA firmasının ürün kataloğundan faydalanılarak, bu katalogda bulunan akıllı telefonlar, cep telefonları ve tablet bilgisayarların özelliklerinden eğitim seti oluşturulmuştur. Bu eğitim seti daha sonra veri seti olarak kullanılıp Naive Bayes, en yakın k komşu ve karar ağacı sınıflandırma algoritmaları ile sınıflandırılacaktır. Bu sınıflandırmalar WEKA kullanılarak gerçekleştirilmiş ve sonuçlar gene aynı yardımcı yazılım ile elde edilmiştir. Elde edilen sonuçlar

değerlendirilip, Google ürün sınıflandırma ağacına göre sınıflandırılan bu üç ürünün doğrulukları karşılaştırılmaktadır. Böylece, internet vasıtası ile müşterilerine hizmet veren firmanın stokuna yeni giren ve sınıflandırmakta güçlük çekilen bir ürünün, müşterileri ile sağlıklı etkileşime geçebilmesi için makine öğrenmesi yöntemleri ile yapılan sınıflandırmanın etkinliği ortaya konularak, bu firmalara yol göstermek amaçlanmıştır.

2. LİTERATÜR TARAMASI:

Ürün sınıflandırması konusunda, Amazon.com sitesinde bulunan çeşitli distribütörlerin ürün kataloglarında bulunan ürünler incelenerek, yapılmak istenen sınıflama için öznitelikler inşa edilerek, yeni gelen ürünün elektronik veya otomotiv sınıflarından hangisine girdiğinin kararının makine tarafından yapılması amaçlanan bir çalışma ile karşılaşılmaktadır[5]. Bu çalışmanın sonunda sınıflanan ürünün kategori sınıfı ne kadar büyürse sınıfın doğruluğu da o kadar arttığı ortaya konmuştur[5]. Bir ürünün sınıfını cep telefonu olarak sınıflandırmak, elektronik ürün olarak sınıflandırmaktan daha riskli olması bu duruma örnek olarak verilebilir.

Bu konuda yapılan başka bir çalışmada ise, ürünlerin sınıflandırılması, metin algoritmaları ve makine öğrenmesi teknikleri ile sağlanmıştır[6]. Bu çalışmada, özgün tanımlayıcı kod, manuel aramalar için çeşitli eş anlamlı kelimeler içeren ürün açıklaması ve uzunluğu, çapı gibi bazı ek(teknik) özellikler gibi ürün arama ve ürünleri kategorize etmenin standartları ortaya konmuştur[6].

Farklı bir çalışmada; ürün sınıflandırmasının, önerilen ürün arama ve karşılaştırma servisi sunan elektronik alışveriş sistemleri için merkezi bir rol oynadığı belirtilmiştir[7].

3. UYGULANAN YÖNTEM VE KULLANILAN VERİ KÜMESİ:

3.1. Kullanılan Veri Kümesi ve Öznitelikler:

Bu çalışmada, daha önce de belirtildiği üzere, veri kümesi olarak TeknoSA firmasını web sayfasında bulunan akıllı telefon, tablet bilgisayarlar ve bu iki sınıfa ait olmayan cep telefonu ürünlerinin, web sayfasında belirtilen ürün özelliklerinden faydalanılarak hazırlanmıştır. Bu öznitelikler, bütün ürünlerde olmasına rağmen özellikler sayfasında belirtilmemiştir. Ürünün özelliklerinin her biri öznitelikleri oluşturmaktır. Bu öz nitelikler; akıllı telefon için arttırılabilir hafıza, dijital kamera, işletim sistemi, 3G, GPS, görüntülü konuşma, dokunmatik ekran; tablet bilgisayarlar için bellek, işletim sistemi, 3G, 4G, güç, ekran kartı; cep telefonları için ise ekran türü, hesap makinesi / kronometre, Mp3 / Radyo, titreşim, video kayıt, 3G, görüntülü konuşma, dokunmatik ekran, renkli ekran, saat / alarm / ajanda, telefon tipi olarak belirtilmektedir. Web sayfasından elde edilen bu bilgiler ışığında eğitim seti ortaya çıkartılacaktır. Çizelge 1 de hangi ürünün hangi özniteliği içerdiği gösterilmektedir.

Eğitim seti oluşturulurken; dokuz adet Nokia ve dokuz adet Samsung marka cep telefonları, on iki adet Samsung, yedi adet HTC, yedi adet LG, altı adet Blackberry, bir adet Alcatel marka akıllı telefon ve bir adet Dell, bir adet Polypad, bir adet Sony, üç adet Probook ve iki adet Samsung marka tablet bilgisayar kullanılmıştır.

Çizelge 1: **Veri kümesi (özellikler)**

		Cep Telefonu	Tablet pc	Akıllı telefon
1	arttirilabilir_hafiza		x	x
2	dijital_kamera			x
3	isletim_sistemi		x	x
4	3G	x	x	x
5	GPS			x
6	goruntulu_konusma	x		x
7	4G		x	x
8	guc		x	
9	ekran_karti		x	
10	ekran_turu	x		
11	hesap_makinesi kronometre	x		
12	Mp3/Radyo	x		
13	titresim	x		
14	video_kayit	x		
15	dokunmatik_ekran	x		x
16	renkli_ekran	x		
17	saat/alarm/ajanda	x		
18	telefon_tipi	x		

3.2. Naive Bayes Sınıflaması

Bayes sınıflandırıcılar istatistiksel sınıflandırıcılar olup sınıf üyelik olasılıklarını tahmin etmektedirler. Naive Bayes sınıflandırıcıları, belirli bir sınıfın öznitelik değerinin etkisinin, diğer öznitelik değerlerinden bağımsız olduğu varsayımına dayanmaktadır[7]. Bu varsayım sınıf koşullu bağımsızlık olarak adlandırılmaktadır[8].

3.3. Karar Ağacı Sınıflaması

Gözetimli öğrenme için kullanılan karar ağaçları, öz çağrılı bir biçimde ilgili yerel bölgenin, bir dizi bölme ile az sayıda adım ile bulunduğu ağaç yapılı bir modeldir[9]. Karar ağaçları, iç karar düğümleri ve uç yapraklardan oluşmaktadır. Resim 2 de karar ağacına örnek şekil verilmiştir. Karar ağacı modelinde, her düğümde denetim işlevi uygulanır ve sonuca göre dallardan biri seçilir. Bu süreç kökte başlar ve bir yaprak düğümüne gelene kadar sürer. Yaprakta yazan değer de çıktıyı oluşturur[9].

3.4. En Yakın k Komşu Sınıflaması

Bu sınıflandırma türü veri madenciliğine ait olup bellek tabanlı bir sınıflandırma çeşididir. Bu sınıflama metodu, öğrenim kümesindeki hatayı ve saklanan alt kümenin büyüklüğü olarak ölçülen karmaşıklığı birlikte azaltan bir algoritmadır[9]. Bu sınıflama metodunda örnek kümesinin sınıfları belli olup, yöntemin amacı örneğe katılacak yeni gözlemin hangi sınıfa ait olduğunu belirlemektir. Örnekler

"n" boyutlu uzayda bir nokta olarak alınır ve verilen noktaya en yakın komşuların sayısı olan "k" parametresi belirlenir. Bu yöntem, uzaklık hesaplama üzerine kurulu olduğu için verilen noktaya diğer tüm noktaların uzaklıkları tek tek hesaplanır. Bu hesaplama işlemi öklid bağıntısı ile yapılır. Hesaplanan uzaklık değerlerine göre satırlar sıralanır ve en küçük "k" tanesi seçilir.

4. DENEYSEL SONUÇLAR

Bu bölümde sınıflandırma algoritmaları kullanılarak elde edilen veriler ortaya konmuş olup, sınıflandırmanın doğruluk oranları ve verilerin ne kadar öğrenilebilir örüntü olduğu ortaya konmuştur.

Çizelge 2: **Sınıflama algoritmaları karmaşıklık matrisleri**

a	b	c	<-	TAHMİN(NaiveBayes)
31	1	1	a	akilli_telefon
0	11	0	b	tablet_pc
1	0	17	c	cep_telefonu

a	b	c	<-	TAHMİN(IBK)
31	1	1	a	akilli_telefon
0	11	0	b	tablet_pc
1	1	16	c	cep_telefonu

a	b	c	<-	TAHMİN(J48)
32	1	0	a	akilli_telefon
0	11	0	b	tablet_pc
0	3	15	c	cep_telefonu

Çizelge 2 de belirtilmiş olan karmaşıklık matrisleri her sınıf için ayrı ayrı incelendiğinde üç algoritmada da tablet_pc sınıfını ayırmada bir problem ile karşılaşılmadığı görülmektedir. Bunun yanında, akilli_telefon sınıfı için en iyi başarımı J48 algoritması göstermiş olup, cep_telefonu sınıfı için ise düşük bir başarımı göstermiş olduğu ortaya konmaktadır.

5. SONUÇLAR

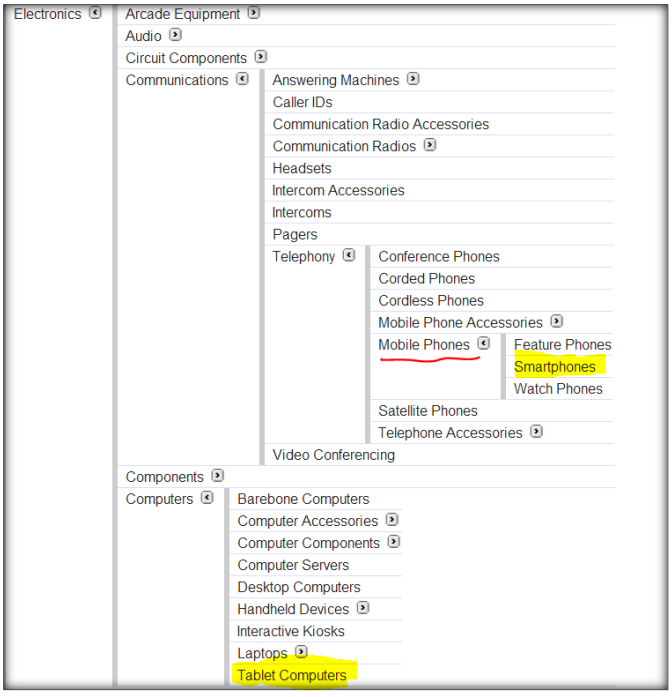
Bu çalışmada seçilmiş olan ürünlerin öznitelikleri ile sınıflandırma yapılmış olup, makine öğrenmesi ile ürün sınıflaması yapılmak istenen başka sistemeler için seçilecek özniteliklerin ve sınıflandırma algoritmalarının ne derecede etkin sonuç verdikleri ortaya konmuştur. Elde edilen sonuçlar ışığında, ürün sınıflandırmak için harcanan emeğin, farklı öznitelikler ile farklı ürünler için uygulandığında da etkin bir sınıflandırma ile sonuçlanacağı sonucuna varılmaktadır. Buna ek olarak, bu çalışma için seçilen ürünlerin Google ürün sınıflandırma ağacına göre hangi ana sınıfın altında konumlandığı araştırılmıştır.

Resim 1 de Google ürün sınıflandırma ağacına göre bu makalede incelenen ürünlerin yerleri gösterilmektedir. Bu sınıflandırmaya göre de cep telefonu ana başlığı altında akıllı telefon sınıfı yer aldığı ve tablet bilgisayar ürününün bilgisayarlar ana başlığı altında bulunduğu görülmektedir. Bu noktada sınıflandırılması zor olan cep telefonu ve akıllı telefon ürünlerinin Google ürün sınıflandırma ağacına göre etkin sınıflanamadıkları sonucu çıkarılmaktadır.

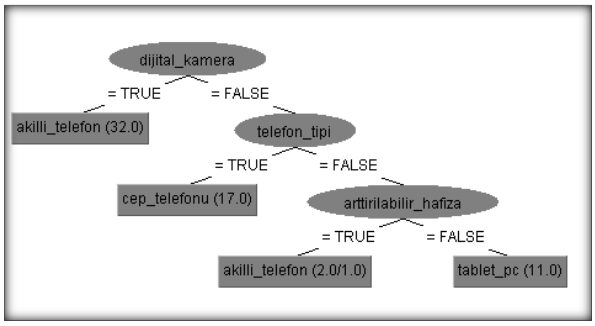
Çalışma kapsamında kullanılan ürünler, karar ağacı algoritması kullanılarak sınıflandırıldığında Resim 2 de görülen ağaç yapısı elde edilmektedir. Buna göre aşağıdaki kural Resim 2 den çıkartılabilmektedir.

```
if(dijital_kamera == true)
    {
     ürün_tipi = akilli_telefon
    }
else
    {
     if(telefon_tipi == true)
      {
       ürün_tipi = cep_telefonu
      }
     else
      {
        if(arttırılabilir_hafiza == true)
         {
           ürün_tipi = akilli_telefon
         }
        else
         {
           ürün_tipi = tablet_pc
         }
      }
    }
```

Resim 1: **Google ürün sınıflandırma ağacına göre cep telefonu, tablet bilgisayar ve akıllı telefon**



Resim 2: **Karar ağacı sınıflandırması sonucunda oluşan ağaç yapısı**



6. TEŞEKKÜR

Bu çalışmada yardımlarını eksik etmeyen Yrd. Doç. Dr. Aslı Uyar Özkaya' ya teşekkür ederim.

7. KAYNAKLAR

[1]http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_ repo rt.pdf
[2]http://hmi.ucsd.edu/howmuchinfo_research_report_consum.php
[3]http://removeandreplace.com/2013/03/13/how-much-data-is-on-the-internet-and-generated-online-every-minute/
[4] Priyali Rajagopal, Robert E. Burnkrant , "Consumer Categorization and Evaluation of Ambiguous Products", Southern Methodist University, Ohio State University
[5]"Applying Machine Learning to Product Categorization", Sushant Shankar and Irving Lin, Department of Computer Science, Stanford University
[6] Hans Friedrich Witschel, Fabian Schmidt , "Information Structuring and Product Classification", University of Leipzig
[7] Eli Cortez, Mauro Rojas Herrera, Altigran S. da Silva, and Edleno S. de Moura, "Lightweight Methods for Large-Scale Product Categorization", Department of Computer Science, Federal University of Amazonas
[8] K. Ming Leung, "Naive Bayesian Classifier", Department of Computer Science / Finance and Risk Engineering, Polytechnic University
[9] Alpaydın E., 2011. "Yapay Öğrenme". Boğaziçi Yayınları.